กระบวนการแบบผสมผสานเพื่อการจำแนกความรู้สึกแบบข้อความสั้น

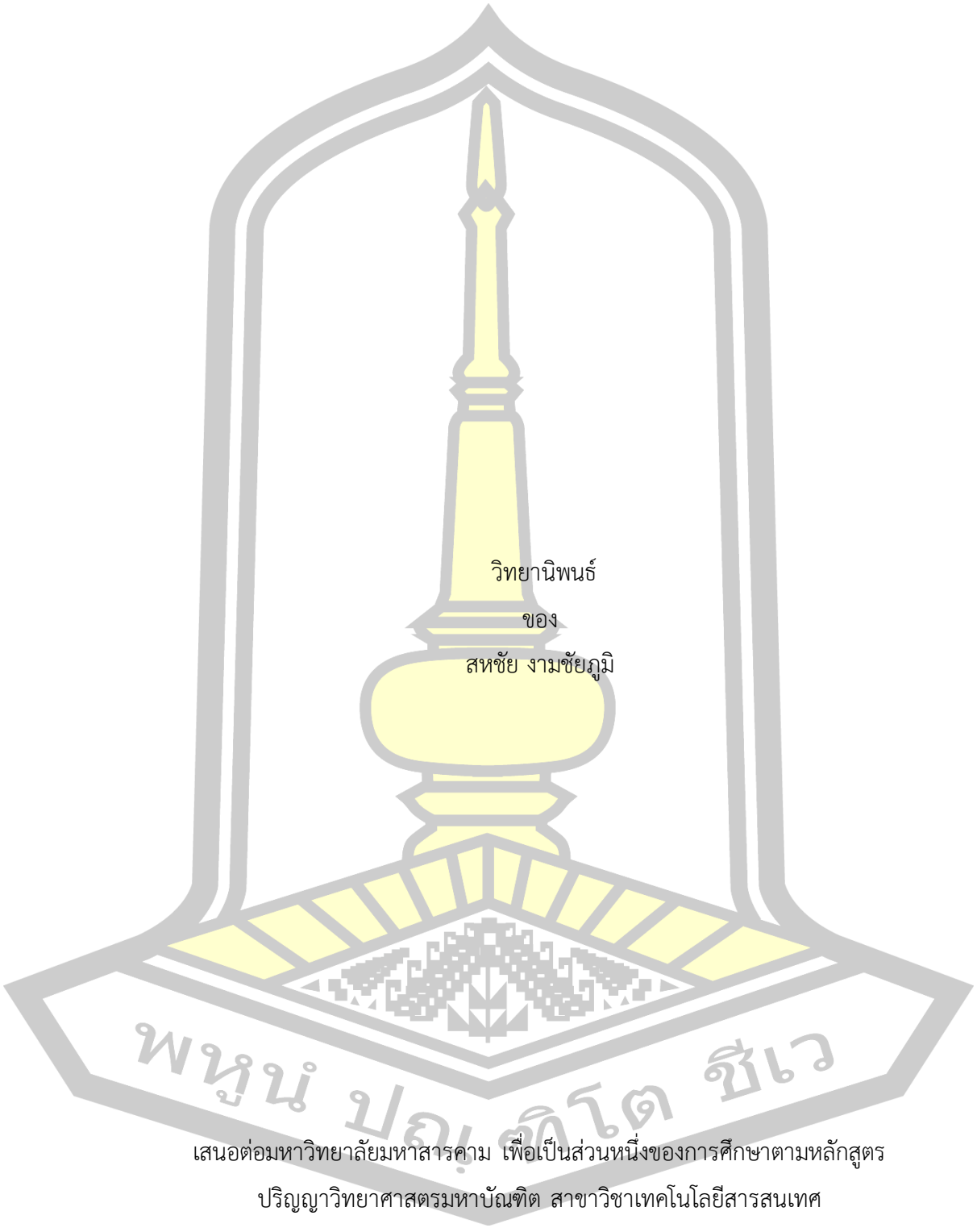
วิทยานิพนธ์

ของ

สหชัย งามชัยภูมิ

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
เมษายน 2563

กระบวนการแบบผสมผสานเพื่อการจำแนกความรู้สึกแบบข้อความสั้น



วิทยานิพนธ์

ของ

สหชัย งามชัยภูมิ

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

เมษายน 2563

A Hybrid Method for Sentiment Classification of Short Texts

Sahachai Ngamchaiyaphum

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

April 2020

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายสหชัย งามชัยภูมิ แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. สิทธิชัย บุษหมั่น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จันทิมา พลพินิจ)

.....กรรมการ
(ผศ. ดร. พนิดา ทรงรัมย์)

.....กรรมการ
(อ. ดร. สาธิต แสงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....

(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง กระบวนการแบบผสมผสานเพื่อการจำแนกความรู้สึกแบบข้อความสั้น
ผู้วิจัย สหชัย งามชัยภูมิ
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จันทิมา พลพินิจ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2563

บทคัดย่อ

บทวิจารณ์ความคิดเห็นของลูกค้าส่วนใหญ่มีรูปแบบเป็นข้อความสั้น ดังนั้นความยาวของข้อความที่มีอยู่จำกัด เป็นความท้าทายสำหรับการจำแนกบทวิจารณ์ของลูกค้า เพราะจำนวนคำที่แสดงในข้อความมีจำนวนน้อยทำให้ไม่สามารถคัดเลือกคุณลักษณะที่เหมาะสมและมีความหมาย หรืออาจจะสกัดได้น้อยเกินไปจนยากต่อการสร้างตัวจำแนกความรู้สึกจากข้อความที่มีคุณภาพต่อการใช้งานที่ดีได้ งานวิจัยนี้ได้นำเสนอกระบวนการจำแนกบทวิจารณ์ที่มีลักษณะข้อความสั้น ด้วยการสร้างโมเดลแบบผสมผสานด้วย 3 เทคนิค คือ Support Vector Machine, Naive Bayes และ K nearest neighbor เพื่อเพิ่มประสิทธิภาพ และความแม่นยำในการจำแนกความรู้สึกจากเอกสารของข้อความที่มีข้อความสั้น โดยการวัดประสิทธิภาพด้วย ค่าความถูกต้อง = 0.97 ค่าความแม่นยำ = 0.98 ความความระลึก = 0.97 และ ค่า F-measur = 0.97

คำสำคัญ : การวิเคราะห์ความรู้สึก, บทวิจารณ์, ข้อความสั้น, การประมวลผลภาษาธรรมชาติ, การจำแนกข้อความ

TITLE A Hybrid Method for Sentiment Classification of Short Texts

AUTHOR Sahachai Ngamchaiyaphum

ADVISORS Assistant Professor Jantima Polpinij , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2020

ABSTRACT

Customer reviews can be represented as short text, i.e. limited in length and usually spanning one sentence or less, but this may pose a challenge for sentiment analysis. When a customer review contains only a few words, this may present difficulty for traditional methods of analysis when dealing with short text classification. This is because a few words in a short text cannot represent the feature space and the relationship between words and documents. As a result, there is tremendous interest in sentiment analysis of customer reviews with short text. This study aims to presents a method for dealing with customer reviews with short text classification. Three weighting schemes and two machine learning algorithms are compared and used for modelling customer review classifiers. After testing by accuracy, recall, precision, and F1, the most satisfactory results are 0.97, 0.98, 0.97, and 0.97 respectively.

Keyword : Sentiment analysis, Customer reviews, Short text, Natural language processing, Text classification

กิตติกรรมประกาศ

วิทยานิพนธ์ เรื่อง กระบวนการแบบผสมผสานเพื่อการจำแนกความรู้สึกของข้อความสั้น ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผศ. ดร. จันทิมา พลพินิจ อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ. ดร. สิทธิชัย บุษหมั่น ประธานกรรมการสอบ ผศ. ดร. พนิดา ทรงรัมย์ กรรมการสอบ อ.ดร. สาธิต แสงประดิษฐ์ กรรมการสอบ

ขอขอบพระคุณ ครูบาอาจารย์ทุกท่าน ที่ประสิทธิ์ประสาทวิชาความรู้ในทุกด้าน ไม่ว่าจะในด้านวิชาการ งานวิจัย หรือ ด้านการดำรงค์ชีวิต

ขอขอบพระคุณ หัวหน้างาน เพื่อนร่วมงาน เจ้าหน้าที่ทุกท่าน ที่ให้ความช่วยเหลือ แนะนำ ในทุก ๆ ด้านจนวิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยดี

สหชัย งามชัยภูมิ

สารบัญ

หน้า

สารบัญภาพ

หน้า

สารบัญตาราง

หน้า

บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

การวิเคราะห์ความรู้สึก (Sentiment Analysis) [1-3] จัดเป็นงานวิจัยแขนงหนึ่งในสาขาการประมวลผลธรรมชาติ (Natural Language Processing: NLP) [4] โดยเป็นการศึกษาและวิเคราะห์เกี่ยวกับความรู้สึก (Feelings) ทัศนคติ (Attitude) อารมณ์ (Emotions) และ ความคิดเห็น (Opinion) ที่เกี่ยวข้องกับองค์กร สินค้า หรือ บริการ จากผู้คนที่ได้แสดงไว้ในรูปแบบเอกสาร (Documents) หรือข้อความ (Message) แบบอัตโนมัติ [5-7] ซึ่งงานวิจัยเรื่องนี้ยังสามารถเรียกชื่ออื่น ๆ ได้อย่างหลากหลาย เช่น การสกัดความคิดเห็น (Opinion Extraction) [8] เหมืองความคิดเห็น (Opinion Mining) [8, 9] เหมืองความรู้สึก (Sentiment Mining) [10] หรือ การวิเคราะห์อัตวิสัย (Subjectivity Analysis) [11] การวิเคราะห์ความรู้สึกสามารถทำได้หลายๆ ลักษณะ โดยขึ้นอยู่กับเป้าหมายของการประยุกต์ใช้งาน โดยการประยุกต์ใช้งานที่สำคัญคือ การวิเคราะห์ความรู้สึกในลักษณะของการจำแนกเอกสาร (Text Classification) การจำแนกความรู้สึก (Sentiment Classification) หรือการจัดอันดับความรู้สึกแบบอัตโนมัติ (Automatic Rating) [1, 2, 10, 12]

จากการศึกษาที่ผ่านมาพบว่าการวิเคราะห์ความรู้สึกในลักษณะของการจำแนกเอกสาร (Text Classification) สามารถใช้ได้กับหลายเทคนิควิธี เช่น นาอีฟเบย์ (Naïve Bayes) [13, 14], Support Vector Machines (SVM) [15, 16] , โครงข่ายประสาทเทียม (Neural Network) [17, 18], ต้นไม้ตัดสินใจ (Decision Tree) [19, 20], การหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) [21, 22] ซึ่งเทคนิควิธีทั้งหมดที่กล่าวมาข้างต้น แม้จะได้รับความนิยมในงานวิจัยเป็นอย่างมากและสามารถใช้กับงานจำแนกเอกสารได้ในระดับที่น่าพอใจในการจำแนกความรู้สึกหรือความคิดเห็น แต่อย่างไรก็ตาม ยังพบปัญหาในการจำแนกความรู้สึกในกลุ่มข้อมูลที่มีข้อความสั้น เนื่องจากจำนวนคำที่แสดงในข้อความมีจำนวนน้อยทำให้ไม่สามารถคัดเลือกคุณลักษณะ (features) ที่เหมาะสมและมีความหมาย [23, 24] หรืออาจจะสกัดได้น้อยเกินไปจนยากต่อการสร้างตัวจำแนกความรู้สึกจากข้อความที่มีคุณภาพต่อการใช้งานที่ดีได้ ข้อความแสดงความรู้สึกสั้นๆ แบบนี้พบได้ทั่วไป เช่น ข้อความการแสดงความรู้สึกทางโซเชียลมีเดีย การแสดงความรู้สึกต่อสินค้าหรือบริการตามเว็บไซต์ต่างๆ

จากปัญหาข้างต้น ในงานวิจัยนี้ผู้วิจัยจึงได้นำเสนอกระบวนใหม่ในสร้างตัวจำแนกความรู้สึกจากเอกสารข้อความขนาดสั้น โดยวิธีการที่เรียกว่ากระบวนการแบบผสมผสาน (Hybrid Method) ที่

พัฒนามาจากเทคนิคเหมืองข้อมูล (Data mining techniques) และการประมวลผลธรรมชาติ (Natural Language Processing: NLP) เพื่อเพิ่มประสิทธิภาพ และความแม่นยำในการจำแนกความรู้สึกจากเอกสารของข้อความที่มีข้อความสั้น สำหรับกรณีศึกษาจะเป็นข้อความแสดงความคิดเห็นต่อโรงแรมหรือที่พักบนอินเทอร์เน็ต

1.2 วัตถุประสงค์ของการวิจัย

1. นำเสนอกระบวนใหม่ในการสร้างตัวจำแนกความรู้สึกจากเอกสารข้อความขนาดสั้น โดยวิธีการที่เรียกว่ากระบวนการแบบผสมผสาน (Hybrid Method)

1.3 ความสำคัญของการวิจัย

1. ได้กระบวนใหม่ในการสร้างตัวจำแนกความรู้สึกจากเอกสารข้อความขนาดสั้น
2. การจำแนกเอกสารข้อความขนาดสั้นมีประสิทธิภาพที่ดีขึ้น

1.4 ขอบเขตของงานวิจัย

1. การสร้างตัวจำแนกความรู้สึกจากเอกสารข้อความขนาดสั้น โดยวิธีการที่เรียกว่ากระบวนการแบบผสมผสาน (Hybrid Method)
2. กระบวนการแบบผสมผสานจะเป็นการผสมผสานระหว่างอัลกอริทึมแบบมีผู้สอน 3 อัลกอริทึม คือ Naive Bayes, Support Vector Machine และ K-nearest neighbor
3. สร้างโมเดลสำหรับจำแนกความรู้สึกจากข้อความแสดงความคิดเห็นขนาดสั้นที่เกี่ยวกับโรงแรมแบบ 2 กลุ่ม คือ ความรู้สึกเชิงบวก (Positive) และความรู้สึกเชิงลบ (Negative)
4. การตัดคำภาษาอังกฤษจะใช้การตัดคำด้วย white space แต่เทียบความถูกต้องของคำด้วยพจนานุกรมของ SentiWordNet
5. ข้อมูลที่ใช้ในการสร้างโมเดลสำหรับจำแนกเอกสารในงานวิจัยนี้ เป็นข้อความแสดงความคิดเห็นเกี่ยวกับโรงแรมต่าง ๆ อย่างน้อย 8,000 เอกสาร โดยแบ่งเป็นเอกสารที่มีความรู้สึกเป็นบวก (Positive) 4,000 เอกสาร และเอกสารที่มีความรู้สึกเป็นลบ (Negative) จำนวน 4,000 เอกสาร โดยแต่ละเอกสารจะมีคำอยู่ระหว่าง 150-250 ตัวอักษรในแต่ละเอกสาร
6. ข้อมูลที่ใช้ในการทดสอบโมเดลสำหรับจำแนกคงามรู้สึกในงานวิจัยนี้ เป็นข้อความแสดงความคิดเห็นเกี่ยวกับโรงแรมต่าง ๆ อย่างน้อย 2,400 เอกสาร โดยแบ่งเป็นเอกสารที่มีความรู้สึกเป็นบวก (Positive) 1,200 เอกสาร และเอกสารที่มีความรู้สึกเป็นลบ (Negative) จำนวน 1,200 เอกสาร โดยแต่ละเอกสารจะมีคำอยู่ระหว่าง 150-250 ตัวอักษรในแต่ละเอกสาร
7. การวัดประสิทธิภาพของโมเดลการจำแนกความรู้สึกด้วยค่าความระลึก (Recall) ค่าความแม่นยำ (Precision) และการวัดค่าเอฟ (F-measure)

1.5 นิยามศัพท์เฉพาะ

1. การวิเคราะห์ความรู้สึก (Sentiment Analysis) [5, 7, 12, 25] คือ กระบวนการวิเคราะห์และสกัดความรู้สึกของมนุษย์จากข้อความ (Text) เพื่ออธิบายความหมายที่บ่งบอกถึงความรู้สึกจากข้อความหรือความคิดเห็นที่มีต่อสินค้าหรือบริหาร เช่น ความรู้สึกดี (Positive หรือ Good) หรือความรู้สึกที่ไม่ดีหรือไม่ชอบ (Negative หรือ Bad) โดยเป็นงานวิจัยที่อยู่ในกลุ่มของการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) [4]

2. การจำแนกเอกสาร (Text Classification) คือ กระบวนการจัดหมวดหมู่ของข้อความหรือกลุ่มคำ โดยการสร้างกฎเพื่อช่วยในการตัดสินใจ เพื่อจำแนกประเภทของข้อความให้อยู่ในหมวดหมู่หรือประเภทของข้อความที่กำหนดไว้ จัดเป็นลักษณะวิธีการหนึ่งของการวิเคราะห์ความรู้สึก

3. ข้อความสั้น (Shot Text) คือ กลุ่มของข้อความหนึ่งซึ่งมีคำน้อยกว่า 150 ตักอักษร [26] เช่น ข้อความที่ส่งถึงกันในโทรศัพท์มือถือ (SMS) การแสดงความรู้สึกหรือกิจกรรมที่ทำอยู่ในโซเชียลมีเดีย การแสดงความคิดเห็นวิพากษ์วิจารณ์ต่อสินค้าหรือบริการที่อยู่บนเว็บไซต์

4. กระบวนการแบบผสมผสาน (Hybrid Method) สำหรับงานวิจัยฉบับนี้จะหมายถึงกระบวนการในการวิเคราะห์เพื่อจำแนกเอกสารบทวิจารณ์ออกเป็น 2 กลุ่มคือ เอกสารในกลุ่มเชิงบวกและเอกสารในกลุ่มเชิงลบ โดยในการวิเคราะห์เพื่อการจำแนกเอกสารบทวิจารณ์นั้นจะเป็นการวิเคราะห์ของอัลกอริทึมการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised machine learning) อย่างอิสระอย่างน้อย 3 อัลกอริทึม ก่อนจะนำผลลัพธ์ที่ได้นั้นมาสรุปเพื่อให้ได้คำตอบสุดท้ายด้วย Voting Ensemble method

5. Voting Ensemble method เป็นเทคนิคที่ใช้โมเดลหลายๆ โมเดล มาช่วยในการหาคำตอบ ในขั้นตอนการสร้างโมเดลจะใช้ชุดข้อมูลชุดทดสอบ (Training Set) ชุดเดียวกัน สร้างโมเดลไม่ต่ำกว่า 3 โมเดลที่แตกต่างกัน ซึ่งทั้ง 3 โมเดลจะทำการจำแนกบทวิจารณ์ ว่าอยู่คลาสใด ซึ่งคลาสที่ได้ผลโหวตสูงสุด จะเป็นคำตอบของการทดสอบ

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้วิจัยได้ศึกษาหลักการของทฤษฎีต่าง ๆ ที่จะนำมาอ้างอิงและประยุกต์ใช้งาน รวมไปถึงการศึกษางานวิจัยที่เกี่ยวข้อง เพื่อนำมาเป็นแนวทางในการทำวิจัย โดยมีดังต่อไปนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 การวิเคราะห์ความรู้สึก (Sentiment Analysis)

การวิเคราะห์ความรู้สึก จัดเป็นงานวิจัยแขนงหนึ่งในสาขาการประมวลผลธรรมชาติ (Natural Language Processing: NLP) [4] โดยเป็นการศึกษาและวิเคราะห์เกี่ยวกับความรู้สึก (Feelings) ทัศนคติ (Attitude) อารมณ์ (Emotions) และ ความคิดเห็น (Opinion) ที่เกี่ยวข้องกับองค์กร สินค้า หรือ บริการ จากผู้คนที่ได้เขียนหรือแสดงไว้ในรูปแบบเอกสาร (Documents) หรือข้อความ (Message) แบบอัตโนมัติ [5-7] เพื่อบ่งบอกความรู้ของตนเองที่มีต่อบางสิ่งบางอย่าง เช่น ความรู้สึกดี (Positive หรือ Good) หรือความรู้สึกที่ไม่ดีหรือไม่ชอบ (Negative หรือ Bad)

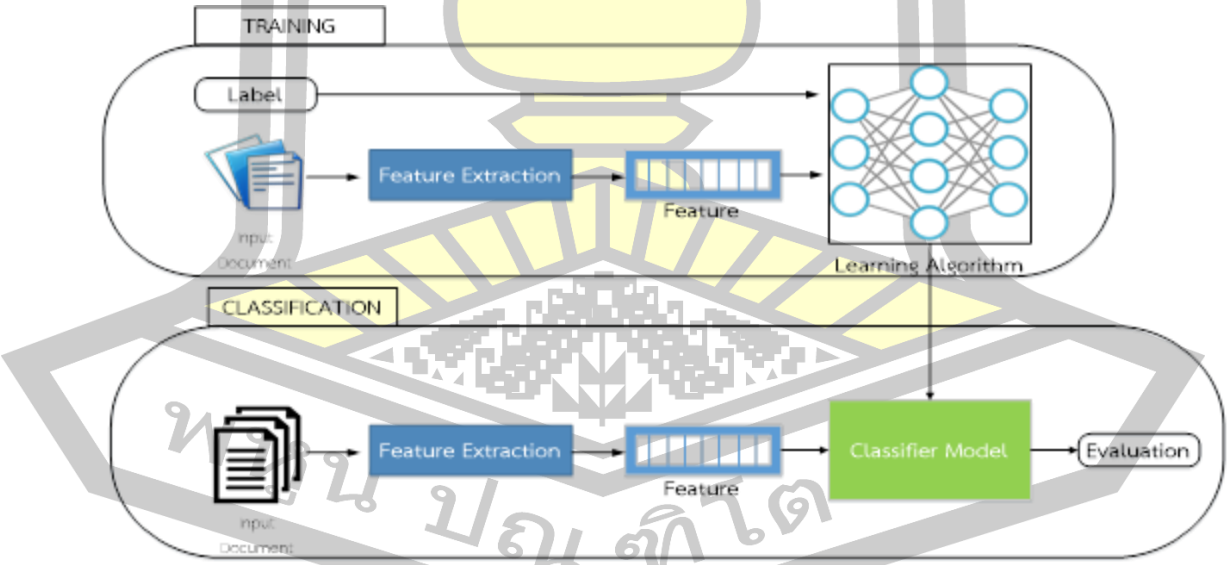
ซึ่งงานวิจัยเรื่องนี้ยังสามารถเรียกชื่ออื่น ๆ ได้อย่างหลากหลาย เช่น การสกัดความคิดเห็น (Opinion Extraction) [8] เหมืองความคิดเห็น (Opinion Mining) [8, 9] เหมืองความรู้สึก (Sentiment Mining) [10] หรือ การวิเคราะห์อัตวิสัย (Subjectivity Analysis) [11] การวิเคราะห์ความรู้สึกสามารถทำได้หลายๆ ลักษณะ โดยขึ้นอยู่กับเป้าหมายของการประยุกต์ใช้งาน โดยการประยุกต์ใช้งานที่สำคัญคือ การวิเคราะห์ความรู้สึกในลักษณะของการจำแนกเอกสาร (Text Classification) การจำแนกความรู้สึก (Sentiment Classification) หรือการจัดอันดับความรู้สึกแบบอัตโนมัติ (Automatic Rating) [1, 2, 10, 12]

ปัจจุบัน เทคนิคด้านการวิเคราะห์ความรู้สึก เริ่มเข้ามามีบทบาทเป็นอย่างมากในหลายๆ องค์กร [3, 8, 9] ทั้งธุรกิจที่เกี่ยวข้องกับสินค้าและบริการ การศึกษา และการให้บริการด้านการแพทย์ โดยเทคนิคการวิเคราะห์ความรู้สึกได้ถูกรวมเข้าไว้ในเว็บไซต์เชิงพานิชย์ (Commercial Website) หรือ ระบบบริหารความสัมพันธ์ลูกค้า (Customer Relationship Management: CRM) ของแต่ละบริษัทหรือองค์กร เพื่อให้ง่ายต่อการวิเคราะห์ความรู้สึกของลูกค้าหรือผู้ใช้บริการ จนนำไปสู่การแก้ปัญหาอย่างรวดเร็ว

2.1.2 การจำแนกความรู้สึก (Sentiment Classification)

การจำแนกความรู้สึก [27, 28] เป็นเทคนิคการสร้างโมเดลเพื่อจำแนกกลุ่มของความรู้สึกให้อยู่ในกลุ่มที่กำหนดขึ้นมา เพื่อแสดงให้เห็นความแตกต่างระหว่างคลาส หรือ กลุ่มของความรู้สึก โดยการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อความที่มีอยู่ ทั้งนี้จะขึ้นอยู่กับการวิเคราะห์เซ็ตของข้อมูลช่วยสอน (Training data) โดยนำข้อมูลช่วยสอน มาสอนให้ระบบเรียนรู้ว่ามีข้อมูลใดอยู่ในกลุ่มเดียวกันบ้าง ผลลัพธ์ที่ได้จากการเรียนรู้ คือ โมเดลการจำแนกประเภทข้อมูล (Classification model) และจะนำข้อมูลส่วนที่เหลือจากข้อมูลชุดสอนการเรียนรู้มาเป็นข้อมูลที่ใช้ทดสอบ (Testing data) ซึ่งเป็นกลุ่มที่แท้จริงของข้อมูลที่ใช้ทดสอบนี้จะถูกนำมาเปรียบเทียบกับกลุ่มที่หามาได้จากโมเดลเพื่อทดสอบความถูกต้อง โดยจะปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้องในระดับที่น่าพอใจ

หลังจากนั้นเมื่อมีข้อมูลใหม่เข้ามา จะนำข้อมูลผ่านโมเดล โดยโมเดลจะสามารถทำนายกลุ่มของข้อมูลใหม่นี้ได้ โดยเทคนิคที่นิยมใช้ในงานวิจัยด้านการจำแนกข้อความ เช่น นาอีฟเบย์ (Naïve Bayes) [13, 14] ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines : SVM) [15, 16] โครงข่ายประสาทเทียม (Neural Network) [17, 18] ต้นไม้ตัดสินใจ (Decision Tree) [19, 20] การหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) [21, 22] โดยกระบวนการโดยทั่วไปของการจำแนกหมวดหมู่ของเอกสาร [29] กระบวนการโดยทั่วไปของการจำแนกความรู้สึกแสดงได้ดังรูปที่ 2-1



รูปที่ 2-1 กระบวนการของการจำแนกความรู้สึก

จากรูปที่ 2-1 เป็นการแสดงกระบวนการของการจำแนกความรู้สึก โดยแยกเป็นสองส่วน คือ ส่วนที่เป็นการสอนการเรียนรู้ (Training Set) และส่วนของการจำแนกเอกสาร (Classification) โดยสามารถอธิบายขั้นตอนต่าง ๆ ของแต่ละส่วนดังต่อไปนี้

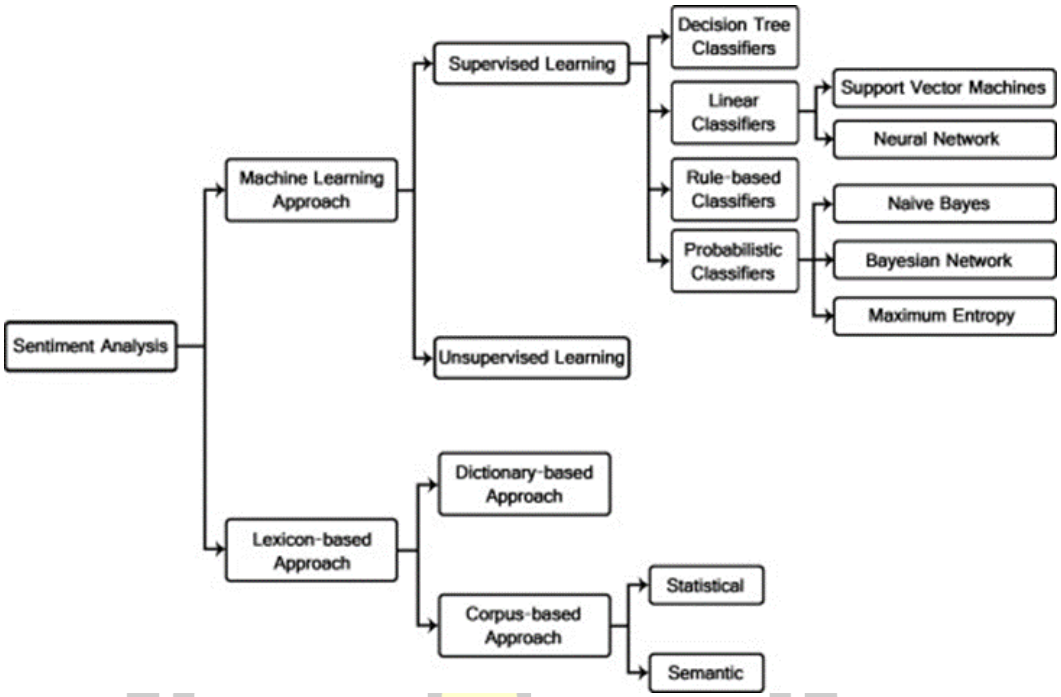
1) การสอนการเรียนรู้ (Training) เป็นขั้นตอนในการสร้างโมเดลโดยมีขั้นตอนการทำงานดังต่อไปนี้

1.1) นำข้อมูลชุดสอนการเรียนรู้เข้าสู่กระบวนการสกัดคุณลักษณะ (Feature Extraction) ซึ่งจำนวนของข้อมูลชุดสอนการเรียนรู้จะแบ่งกันกับข้อมูลชุดทดสอบโมเดล โดยทั่วไปแล้วข้อมูลชุดสอนการเรียนรู้จะเยอะกว่าข้อมูลชุดทดสอบโมเดล เช่น มีข้อมูล 100 ชุด แบ่งเป็นข้อมูลชุดสอนการเรียนรู้ 70 ชุด และที่เหลือ 30 ชุดจะเป็นข้อมูลชุดทดสอบโมเดล

1.2) เมื่อได้คุณลักษณะของเอกสารตามที่ต้องการแล้ว จะนำเข้าสู่เทคนิคการเรียนรู้ (Learning Algorithm) ซึ่งขึ้นอยู่กับแต่ละงานที่จะเลือกใช้วิธีใดที่เหมาะสมกับงานมากที่สุด เช่น นาอีฟเบย์ (Naïve Bayes), Support Vector Machines (SVM), โครงข่ายประสาทเทียม (Neural Network), ต้นไม้ตัดสินใจ (Decision Tree), การหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor)

2) การจำแนกเอกสาร (Classification) เป็นขั้นตอนในการนำข้อมูลที่เหลือจากข้อมูลชุดสอนการเรียนรู้มาสกัดคุณลักษณะ เมื่อได้คุณลักษณะของข้อความแล้วจะนำเข้าทดสอบประมวลผลในโมเดลที่ได้จากขั้นตอนการสอนการเรียนรู้ ซึ่งขั้นตอนนี้อาจมีการปรับปรุงโมเดลเพื่อให้ได้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุด เมื่อได้ผลลัพธ์อย่างไรแล้วก็จะนำไปวัดประสิทธิภาพในขั้นตอนต่อไป

เทคนิคที่สำคัญของการจำแนกความรู้สึก สามารถแบ่งออกได้ตามประเภทของการจำแนก [30] คือ การจำแนกโดยการเรียนรู้เครื่อง (Machine Learning Approach) และ การจำแนกโดยพจนานุกรมข้อมูล (Lexicon-based Approach) โดยสามารถอธิบายโดยสรุปได้ดังรูปที่ 2-2



รูปที่ 2-2 เทคนิคที่สำคัญของการจำแนกความรู้สึก

จากรูปที่ 2-2 เป็นการการอธิบายเทคนิคของการจำแนกความรู้สึกโดยใช้ประเภทของการจำแนกความรู้สึกเป็นตัวแบ่ง จะเห็นได้ว่าการจำแนกโดยการเรียนรู้เครื่องยังแบ่งย่อยไปอีกคือ การสอนการเรียนรู้แบบมีผู้สอน (Supervised Learning) และ การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ซึ่งในงานวิจัยฉบับนี้ ผู้วิจัยได้เลือกใช้เทคนิคในส่วนของการเรียนรู้แบบมีผู้สอน คือ naïveBayes และ Support vector machine ซึ่งจะอธิบายการทำงานของทั้งสองเทคนิคในหัวข้อทฤษฎีที่เกี่ยวข้องต่อไป

2.1.3 ปัญหาของการวิเคราะห์ความรู้สึก

จากการศึกษาที่ผ่านมาพบว่าการวิเคราะห์ความรู้สึกในลักษณะของการจำแนกเอกสาร แม้จะได้รับความนิยมในงานวิจัยเป็นอย่างมาก แต่อย่างไรก็ตาม ก็ยังพบว่าปัญหาในการจำแนกความรู้สึกในกลุ่มข้อมูลที่มีจำนวนคำในข้อความมีน้อย หรือ มีความถี่ของคำน้อย ซึ่งในกระบวนการเตรียมข้อมูลต้องมีการตัดคำที่ไม่มีนัยสำคัญในข้อความหรือมีผลต่อการวิเคราะห์ข้อมูลออกไป จึงทำให้จำนวนคำยิ่งน้อยลงไปอีก ทำให้การให้น้ำหนักคำมีน้อย ส่งผลให้ไม่สามารถคัดเลือกคุณลักษณะ (Features Selection) ที่เหมาะสมและมีความหมายได้ [23, 24] หรืออาจจะสกัดได้น้อยเกินไป จนไม่สามารถสร้างโมเดลที่มีประสิทธิภาพดีพอ

สมมุติให้มีเอกสารอยู่ 2 ชุดคือ

D1 : Good service and clean rooms

D2 : Hotel was located about 10mins walk from the Hakata Station subway. Staff was friendly & helpful

เมื่อเข้าสู่กระบวนการเตรียมข้อมูล คือ ตัดคำ และ ตัดคำหยุด เสร็จแล้วผลลัพธ์ที่ได้คือ

D1 : Good | clean

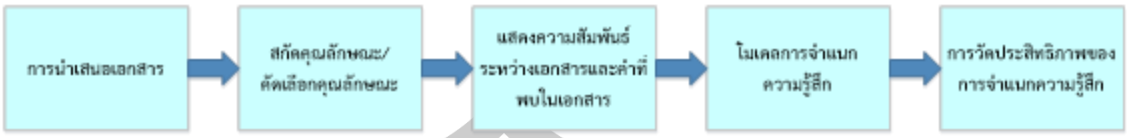
D2 : Hotel | located | Staff | friendly | helpful

จะเห็นว่าในเอกสารชุดแรกมีจำนวนของคำน้อย หรือ ข้อความสั้น เมื่อผ่านกระบวนการตัดคำ และ ตัดคำหยุด จะเหลือคำที่จะนำเข้าสู่กระบวนการจำแนกเอกสารน้อยมาก ดังนั้นจึงทำให้ยากต่อการสร้างโมเดลในการจำแนกความรู้สึกจากข้อความที่มีคุณภาพต่อการใช้งานที่ดีได้ ดังนั้นการวิเคราะห์ความเชื่อมต่อของประโยคเพื่อนำไปสู่ความหมายที่ถูกต้องจะทำได้ยากหรืออาจเป็นไปไม่ได้ [23, 24] ข้อความแสดงความรู้สึกที่มีจำนวนคำน้อยแบบนี้พบได้ทั่วไป เช่น ข้อความการแสดงความรู้สึกทางโซเชียลมีเดีย การแสดงความรู้สึกต่อสินค้าหรือบริการตามเว็บไซต์ต่าง ๆ

2.1.4 การจำแนกหมวดหมู่เอกสาร (Text Classification)

การจำแนกหมวดหมู่เอกสาร [27] เป็นการนำวิธีการเรียนรู้ด้วยคอมพิวเตอร์ (Machine Learning) มาประยุกต์ใช้ร่วมกับการประมวลผลภาษาธรรมชาติ เป็นการจัดแบ่งกลุ่มเอกสารแบบอัตโนมัติ โดยการแบ่งกลุ่มตามเนื้อหาของเอกสารที่มีการกำหนดกลุ่มหรือหมวดหมู่ของเอกสารไว้ก่อนหน้า โดยจะเปรียบเทียบเอกสารกับต้นแบบในแต่ละหมวดหมู่ เอกสารจะถูกจัดอยู่ในหมวดหมู่ที่มีต้นแบบลักษณะคล้ายกับตัวมันเองมากที่สุด

การจัดหมวดหมู่เอกสารในภาษาอังกฤษ ไม่สามารถนำมาใช้กับการจัดหมวดหมู่เอกสารภาษาไทยได้โดยตรง เนื่องจากมีปัญหาในการประมวลผลระดับคำ เช่น การตัดคำในภาษาอังกฤษจะสามารถทำได้ง่าย เนื่องจากว่ามีช่องว่างระหว่างคำ แต่ในภาษาไทยไม่มีช่องว่าง ทำให้การตัดคำมีความยากและซับซ้อนกว่าภาษาอังกฤษ เช่น คำว่า “ตากลม” จะสามารถตัดคำได้สองแบบคือ “ตา-กลม” กับ “ตาก-ลม” หรือปัญหาในกรณีที่เจอคำที่ไม่รู้จักเช่น คำแสลง คำอุทาน เป็นต้น กระบวนการโดยทั่วไปของการจำแนกหมวดหมู่ของเอกสาร [29] แสดงได้ดังรูปที่ 2-3



รูปที่ 2-3 กระบวนการโดยทั่วไปของการจำแนกหมวดหมู่ของเอกสาร

จากรูปที่ 2-3 สามารถอธิบายกระบวนการของการจำแนกหมวดหมู่ได้ ดังนี้

1) การนำเสนอเอกสาร (Document Representation) เป็นกระบวนการเตรียมข้อมูลก่อนเข้าสู่กระบวนการวิเคราะห์ เช่น การตัดคำ การตัดคำหยุด การให้น้ำหนักคำ การทำถุงคำ (Bag of Words)

2) การคัดเลือกคุณสมบัติและการแปลงคุณสมบัติ (Feature Selection / Feature Transformation) เป็นขั้นตอนการสกัดหาคุณสมบัติที่สนใจจากข้อมูลที่ผ่านกระบวนการที่ผ่านการเตรียมข้อมูลมาแล้ว หลังจากนั้นจะทำการแปลงคุณสมบัติให้อยู่ในรูปแบบที่กำหนด

3) การจัดการโครงสร้างของข้อมูล (Construction a Vector Space Model) ขั้นตอนนี้เป็นกระบวนการจัดการโครงสร้างของข้อมูลที่จะนำเข้ากระบวนการวิเคราะห์ข้อมูล ซึ่งวิธีที่นิยมใช้มากที่สุดคือการให้ค่าน้ำหนักคำ (TF-IDF)

4) การนำข้อมูลเข้ากระบวนการจำแนกข้อมูล (Application of Classification Algorithm) เป็นขั้นตอนการสร้างโมเดลสำหรับนำไปใช้ในการทดลอง ซึ่งเทคนิคที่นิยมใช้ในงานวิจัยด้านการจำแนกข้อความ เช่น นาอีฟเบย์ (Naïve Bayes) [13, 14], Support Vector Machines (SVM) [15, 16] , โครงข่ายประสาทเทียม (Neural Network) [17, 18], ต้นไม้ตัดสินใจ (Decision Tree) [19, 20], การหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) [21, 22]

5) การวัดผลการจำแนกข้อมูล (Evaluation of Text Classifier) เป็นกระบวนการที่จะวัดประสิทธิภาพและความแม่นยำของโมเดลที่สร้างขึ้นมาหลังจากนำข้อมูลเข้า ซึ่งการวัดประสิทธิภาพที่นิยมได้แก่ ค่าความระลึก (precision) ค่าความแม่นยำ (recall) และ ค่าวัดประสิทธิภาพ (F-measure)

2.1.5 การตัดคำ (Word Segmentation หรือ Tokenization)

เนื่องจากในงานวิจัยที่เป็นการจำแนกเอกสารจำเป็นอย่างยิ่งที่จะต้องหาขอบเขตของแต่ละคำ เพื่อที่จะสามารถทำการประมวลผลกับข้อความในเอกสารได้อย่างสะดวก การตัดคำจะมีวิธีการดังนี้

การตัดคำ ในงานวิจัยฉบับนี้ได้ใช้หลักการตัดคำภาษาอังกฤษโดยจะใช้ช่องว่างในการแบ่งขอบเขตของคำ ซึ่งทำให้ง่ายต่อการกำหนดขอบเขตของคำ ตัวอย่างการตัดคำภาษาอังกฤษแสดงได้ดังตัวอย่างในตารางที่ 2-1

ตารางที่ 2-1 แสดงตัวอย่างการตัดคำ

ประโยคเดิม	**ประโยคที่ผ่านการตัดคำ**
Pond in the morning time	Pond | in | the | morning | time
This is beautiful natural views	This | is | beautiful | natural | views

2.1.6 การตัดคำหยุด (Stop Word)

การตัดคำหยุด [5] คือ กระบวนการตัดคำหรือสัญลักษณ์ที่พบบ่อยมากในเอกสาร แต่คำหรือสัญลักษณ์เหล่านั้นไม่ได้ส่งผลต่อการจัดกลุ่มเอกสาร ดังนั้นเมื่อทำการตัดออกแล้วไม่ทำให้ใจความในเอกสารนั้นๆ เปลี่ยนไป เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญ ในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ความหมายของคำในเอกสารเปลี่ยนแปลง คือ

การตัดคำหยุด มีความจำเป็นอย่างมากในการจัดกลุ่มเอกสารแบบอัตโนมัติ เพราะจะช่วยลดระยะเวลาในการประมวลผลลงได้เป็นอย่างมาก เนื่องจากการทำงานจะไม่เสียเวลาในการประมวลผลคำเหล่านี้ ยกตัวอย่างการตัดคำหยุด สามารถแสดงได้ดังตารางที่ 2-2

ตารางที่ 2-2 แสดงการตัดคำหยุด

ประโยคที่ผ่านการตัดคำ	คำที่ได้จากประโยค
Pond | in | the | morning | time	Pond | morning
This | is | beautiful | natural | views	beautiful | natural | views

2.1.7 การสกัดคุณลักษณะและคัดเลือกคุณลักษณะ (Feature Extraction and Feature Selection)

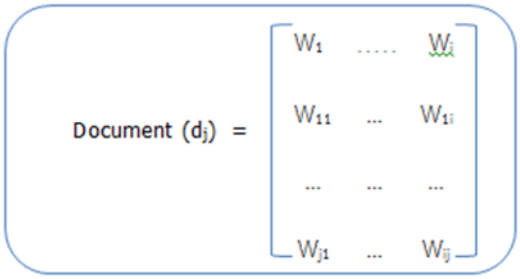
วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะเอกสาร คือ การดึงคุณลักษณะ [27] ของเอกสารออกมา กับการลดขนาดเอกสารลง ซึ่งการดึงคุณลักษณะออกมานั้น ต้องกำหนดก่อนว่า จะใช้อะไรเป็นตัวแทนคุณลักษณะของเอกสาร และใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการศึกษางานวิจัยที่ผ่านมาทั้งในประเทศและต่างประเทศ พบว่าส่วนใหญ่จะใช้คำเป็นตัวแทน คุณลักษณะของ

เอกสาร และใช้พื้นฐานค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้วลี หรือกลุ่มของคำประโยค แทนคุณลักษณะของเอกสารได้เช่นกัน ตัวแทนคุณลักษณะของเอกสารที่นิยมใช้ในการจัดหมวดหมู่เอกสารประเภทข้อความ คือ ถุงคำ (Bag of words) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์

โดยองค์ประกอบของเวกเตอร์อาจจะแทนด้วยคุณลักษณะของค่าความจริง (Boolean) แทนด้วยค่าความถี่ของคำ (Word Frequency) หรือแทนด้วยค่าน้ำหนักของคำแบบอื่นๆ ซึ่งในงานวิจัยฉบับนี้ นี้ใช้การเลือกคุณลักษณะแบบคำเดี่ยว (Single word) ซึ่งได้จากการตัดคำโดยใช้พจนานุกรมเรียบร้อยแล้ว ผลลัพธ์ที่ได้จากการตัดคำจะได้เป็นคำเดียวจำนวนมาก เพื่อมาใช้เป็นตัวแทนเอกสารในการเรียนรู้ การสร้างดัชนี (indexing) เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ ดังนั้น จึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสารเรียกว่า การทำดัชนี เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้

วัตถุประสงค์ของการสร้างดัชนี คือ การคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้วาการหาค่าน้ำหนัก (Term Weighting) การสร้างดัชนีโดยทั่วไปที่นิยมใช้กันจะเริ่มจากการสร้างเวกเตอร์ ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม [4] ซึ่งในงานวิจัยนี้ได้ใช้วิธีความถี่ของคำที่ปรากฏในเอกสารที่ผ่านจากการตัดคำมาเป็นค่าน้ำหนัก ถ้าคำใดที่ผ่านการตัดคำมีปริมาณมาก ก็จะมีค่าความถี่มาก ซึ่งจะส่งผลให้ได้คาน้ำหนักที่มีค่าสูงมาก

เมื่อถึงขั้นตอนนี้จะได้รูปแบบที่มีลักษณะของการแสดง ความสัมพันธ์ระหว่างคำ (Words: w) และเอกสารสารทั้งหมด (Documents: d) ด้วยเวกเตอร์ 2 มิติ ซึ่งคำที่ได้นั้นต้องผ่านทำดัชนีและตัดคำหยุด (Stop-words) ออกไป และเอกสารทั้งหมด รูปแบบ Vector Space Model หรือบางครั้งเรียกรูปแบบนี้ว่า Bag of Words โดยสามารถแสดง ได้ดังรูปที่ 2-4



รูปที่ 2-4 แสดง Bag of Word

2.1.8 การให้น้ำหนักคำ (Term Weighting)

การให้น้ำหนักคำเป็นการกำหนดค่าน้ำหนัก “คำ” หรือ “วลี” โดยเป็นหลักการให้ความสำคัญกับคำหรือวลีในด้านของค่าสถิติศาสตร์ เช่น การให้ค่าสถิติความถี่ (frequency) การเกิดขึ้นของคำ จากหลักไวยากรณ์ภาษาในการใช้คำหรือวลีเดิมซ้ำ ๆ เพื่อเป็นการเน้นย้ำถึงความสำคัญของคำหรือวลีนั้นในเอกสาร ความสันพันธ์ระหว่างเอกสารและคำที่พบในเอกสารนั้น ในรูปแบบของเมตริกซ์แบบ 2 มิติที่เรียกว่า Vector Space Model (VSM) หรือ Bag of Words (BOW) อย่างไรก็ตาม เพื่อให้ความสัมพันธ์ระหว่างเอกสารและคำมีความน่าเชื่อถือ ซึ่งรูปแบบนี้ เป็นรูปแบบที่พร้อมต่อการนำเอาเอกสารเหล่านี้เข้าสู่กระบวนการเรียนรู้เพื่อสร้างโมเดล ด้วยอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning)

การแทนเอกสารด้วยรูปแบบ BOW เป็นการกำหนด “คำ” ในเอกสารด้วย Wij ดังนั้น เอกสารลำดับที่ j ใดๆ สามารถเขียนแทนได้ด้วย dj = (w11, w12, w13, ..., wij)

ตัวอย่างการให้น้ำหนักคำ สามารถทำได้ดังนี้ สมมุติให้มีเอกสารเป็นเอกสาร ไทย 4 ฉบับ ดังนี้

D1 : The bathroom is very clean, the food is delicious.

D2 : Good atmosphere Delicious breakfast

D3 : The parking is quite narrow.

D4 : The price is quite expensive.

จากทั้ง 4 เอกสาร เมื่อผ่านขั้นตอนการตัดคำและตัดคำหยุด สามารถแสดงได้ดังตารางที่ 2-3 แสดงการตัดคำและการตัดคำหยุด

ตารางที่ 2-3 แสดงการตัดคำและการตัดคำหยุด

เอกสารที่	ประโยคที่ผ่านการตัดคำ	คำสำคัญที่ได้ภายหลังการตัดคำหยุด
1	The | bathroom | is | very | clean | the | food | is | tasty	clean / tasty
2	Good | atmosphere | tasty | breakfast	good / tasty
3	The | parking | is | quite | narrow.	narrow
4	The | price | is | quite | expensive	expensive

เมื่อได้ผลลัพธ์ดังจะสามารถแสดงความสัมพันธ์ระหว่าง “คำสำคัญ” และ “เอกสาร” ในรูปแบบของ Vector Space Model (หรือ Bag of Words: BOW) ดังแสดงได้ในตารางที่ 2-4

ตารางที่ 2-4 แสดงความสัมพันธ์ระหว่างคำสำคัญและเอกสาร

	clean	tasty	good	narrow	expensive
D1	1	1	0	0	0
D2	0	1	1	0	0
D3	0	0	0	1	0
D4	0	0	0	0	1

ในเอกสารหนึ่งฉบับ จะพิจารณาจากความถี่ของคำ (Term Frequency) ที่ปรากฏในเอกสารนั้นและจำนวนที่ปรากฏมีทั้งหมดกี่คำ โดยในที่นี้จะใช้วิธีการให้น้ำหนักของคำด้วยวิธี *tf-idf* (Term Frequency – Inverted Document Frequency)

tf-idf เป็นวิธีการสร้างตัวแทนเอกสารในรูปแบบของเวกเตอร์เพื่อใช้ในการจัดกลุ่มของเอกสารให้ตรงกับหมวดหมู่ที่ถูกกำหนดไว้ โดย *tf* เป็นการหาความถี่ของคำที่ปรากฏในเอกสาร และ *idf* เป็นการหาส่วนกลับของเอกสารหรือที่เรียกว่าระบบน้ำหนักความถี่เอกสารผกผัน โดยสามารถหาได้จาก สมการ (2.1)

$$idf = 1 + log\,(N/df) \tag{2.1}$$

โดยที่ N คือจำนวนเอกสารทั้งหมดในกลุ่ม และ df คือจำนวนเอกสารที่มีคำๆ นั้นปรากฏอยู่ในสมการ (2.2)

$$tfidf = tf \times idf \tag{2.2}$$

จากสมการดังกล่าวเป็นวิธีการหาตัวแทนเวกเตอร์เพื่อนำไปค้นคืนสารสนเทศที่เป็นกลุ่มของเอกสาร ซึ่งวิธีนี้เป็นการให้น้ำหนักอย่างง่ายแต่ก็ได้รับการยอมรับว่ามีประสิทธิภาพที่น่าพอใจกับการจัดกลุ่มเอกสาร

แต่อย่างไรก็ตาม จากการศึกษาค้นคว้าในงานวิจัยอื่น ๆ เพิ่มเติมพบว่า ถึงแม้ *tf-idf* จะเป็นวิธีที่ได้รับความนิยมในการให้ค่าน้ำหนักของคำที่มีประสิทธิภาพที่ดี แต่เป็นการให้น้ำหนักคำทั้งคลังของเอกสาร ซึ่งในงานวิจัยด้านการจัดกลุ่มข้อความ (Text Classifier) การพิจารณน้ำหนักคำเฉพาะกลุ่มจะให้ผลลัพธ์ที่ดีกว่า ซึ่งในงานวิจัยนี้ ได้นำวิธีการให้น้ำหนักของคำที่เฉพาะเจาะจงไปที่กลุ่ม (class) คือ การให้น้ำหนักคำแบบ *tf-icd* (Term frequency – Inverse Corpus frequency)

วิธีการให้น้ำหนักคำแบบ *tf-icf* เป็นการให้น้ำหนักคำที่อยู่บนแนวคิดที่เรียกว่า “การให้น้ำหนักคำที่ให้ความสำคัญในแต่ละคลาส” โดยเป็นการให้น้ำหนักที่ทำการปรับมาจากวิธีการให้น้ำหนักคำแบบ *tf-idf* เพราะเมื่อพิจารณาแล้วจะพบว่า *tf-idf* จะเป็นการให้น้ำหนักคำที่สะท้อนความสำคัญของคำที่อยู่ในเอกสารหนึ่งๆ ที่อยู่ในคลังเอกสาร แต่ถ้าหากจะทำการจำแนกเอกสารออกเป็นกลุ่มหรือคลาส การให้น้ำหนักของคำก็ควรสะท้อนความสำคัญของคำในเอกสารที่เ (2.3) แต่ละคลาส ดังนั้นจึงมีการปรับวิธีการของ *tf-idf* ด้วยการแทนที่ *idf* ด้วย *icf* สามารถแสดงได้ตามสมการ (2.3)

$$w(t_k) = \log(1 + tf_k) \times \log\left(\frac{N+1}{df(t_k)+1}\right) \quad (2.3)$$

เมื่อ $w(t_k)$ คือ ความถี่ของคำ t_k ที่พบในเอกสาร

N คือ จำนวนเอกสารทั้งหมดในคลาสนั้น ๆ

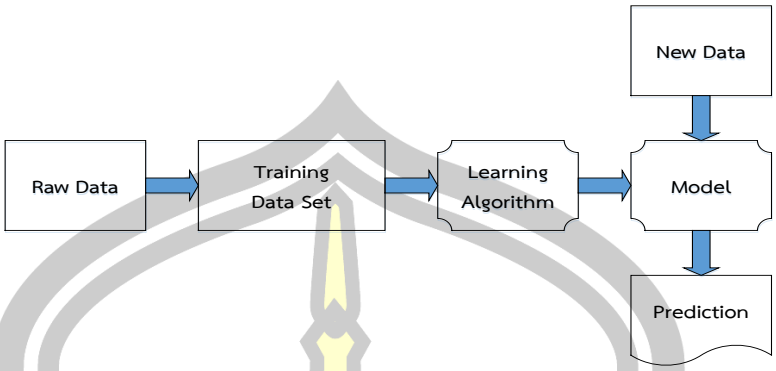
$df(t_k)$ คือ จำนวนเอกสารในคลาสที่พบคำ t_k

ซึ่ง tf-icf สามารถลดความซับซ้อนของการประมวลผลที่มีการให้น้ำหนักคำด้วย *tf-idf* จาก $O(N^2)$ มาเป็น $O(N)$

2.1.9 การเรียนรู้เครื่อง (Machine Learning)

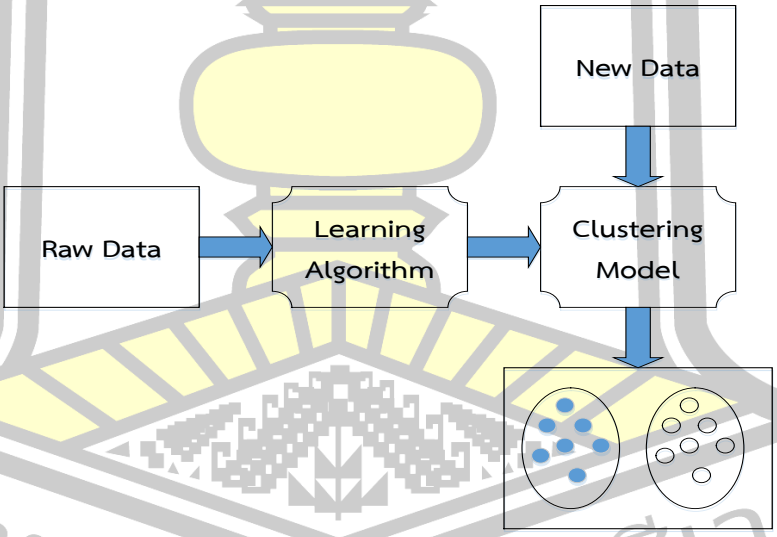
การเรียนรู้เครื่อง: (Machine Learning) คือ การสอนให้คอมพิวเตอร์มีความสามารถที่จะเรียนรู้ได้ด้วยตนเอง เมื่อมีข้อมูลเข้ามาสามารถทำนายหรือตัดสินใจได้เองโดยอัตโนมัติโดยปราศจากการทำงานตามลำดับคำสั่งโปรแกรม กระบวนการของการเรียนรู้เครื่องสามารถอธิบายได้ประเภทของการเรียนรู้เครื่องโดยทั่วไปแล้วสามารถแบ่งได้ออกเป็น 2 ประเภทหลักๆ ด้วยกันคือ

1) การเรียนรู้แบบมีผู้สอน (Supervised learning) คือ การเรียนรู้ประเภทที่ต้องมีการสอนการเรียนรู้ให้กับโปรแกรม หรือ Training Data ก่อนถึงจะสามารถประมวลผลข้อมูลได้ ยกตัวอย่างเช่นโปรแกรมจดจำลายนิ้วมือ หรือ หากจะเขียนโปรแกรมให้บอกว่าภาพถ่ายเป็นภาพแอปเปิล หรือส้ม จะต้องสอนให้โปรแกรมเรียนรู้ก่อนว่าแอปเปิลมีสีแดง ส้มมีสีส้ม เมื่ออินพุตภาพผลไม้ทรงกลมสีแดงเข้ามา โปรแกรมจะตัดสินในได้เองมารูปที่อินพุตเข้ามาคือแอปเปิล โดยสามารถแสดงภาพรวมของการเรียนรู้แบบมีผู้สอนได้ดังรูปที่ 2-5



รูปที่ 2-5 การเรียนรู้แบบมีผู้สอน

การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) คือ การเรียนรู้จะไม่มีการระบุผล (Target variable) ที่ต้องการไว้ก่อน ต้องให้คอมพิวเตอร์หาความสัมพันธ์จากข้อมูลที่อินพุตเข้ามาด้วยตนเอง จึงกล่าวได้ว่าการเรียนรู้ประเภทนี้เป็นการเรียนรู้แบบไม่มีผู้สอน ตัวอย่างของการเรียนรู้แบบไม่มีผู้สอนเช่น การจัดกลุ่ม (Clustering) การเรียนรู้แบบนี้จะเป็นการแบ่งกลุ่มของข้อมูลอินพุตโดยอาศัยการเรียนรู้จากข้อมูลที่อินพุตเข้ามาด้วยตัวเอง ว่าข้อมูลมีความสัมพันธ์หรือคล้ายคลึงกันแบบใด จะจัดข้อมูลที่มีความคล้ายกันนั้นไปอยู่กลุ่มเดียวกัน ดังแสดงในรูปที่ 2-6



รูปที่ 2-6 การเรียนรู้แบบไม่มีผู้สอน

2.1.10 เทคนิควิธีของการเรียนรู้เครื่อง

เทคนิคของการเรียนรู้เครื่องที่สำคัญและเลือกมาใช้ในงานวิจัยนี้มีดังต่อไปนี้

1) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines: SVM) เป็นเทคนิคหนึ่งที่จัดอยู่ในกลุ่มการจำแนกประเภทข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด

แนวคิดหลักของวิธีการนี้ใชเพื่อหาระนาบการตัดสินใจในการแบงข้อมูลออกเป็นสองส่วน โดยใชสมการเส้นตรงเพื่อแบงเขตข้อมูล 2 กลุมออกจากกัน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระระหว่างขอบเขตของทั้งสองกลุ่มมาก กำหนดให้ (Xi,Yi),..., (Xn,Yn) เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูลตัวอย่าง m คือ จำนวนมิติของข้อมูลนำเข้า และ Y คือผลลัพธ์ +1 หรือ -1 ดังสมการ (2.4)

(2.4)

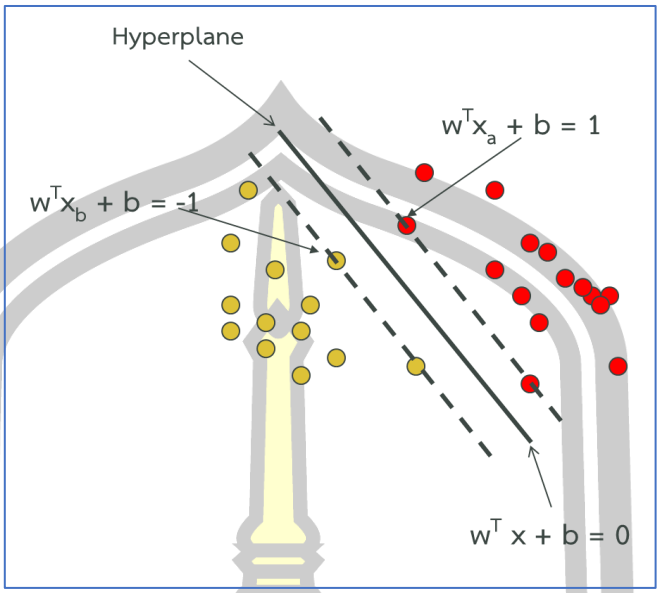
$$(Xi, Yi), \ldots, (Xn, Yn) \text{ เมื่อ } x \in R, y \in \{+1, -1\}$$

ตัวอย่างขั้นตอนการจัดกลุ่มเอกสารมีดังต่อไปนี้

1. นำเอกสารที่อินพุตเข้ามาหาค่า y ซึ่งค่าของ $y \in \{-1,1\}$ และ ค่า $x \in R^n$ โดย ถ้าค่าของ $w^T x + b > 0$ จะกำหนดให้ค่า $y = +1$ ซึ่งจะจัดอยู่ใน Class 1 และ ถ้าค่าของ $w^T x + b < 0$ จะกำหนดให้ค่า $y = -1$ ซึ่งจะจัดอยู่ใน Class 2

2. คำนวณหาค่าเส้นตรงที่แบ่งเอกสารซึ่งเรียกว่า เส้น Optimal Hyperplane

3. นำค่าที่ได้จากข้อ 1. และ 2. ไปเขียนบนเส้นตรงตามแนวแกนตั้งและแกนนอนดังตัวอย่างในรูปที่ 2-7



รูปที่ 2-7 ตัวอย่างการแบ่งกลุ่มข้อมูลโดยซัพพอร์ตเว็กเตอร์แมชชีน

2) นาอีฟเบย์ (Naïve Bayes) เป็นอัลกอริทึมที่ถูกนำมาใช้อย่างแพร่หลายในงานจำแนกเอกสาร และให้ผลที่ดี การหาจำนวนนาอีฟเบย์ เริ่มจากแต่ละ อินสแตนซ์ (Instance) x ซึ่งจัดอยู่ ในรูปเวกเตอร์ของค่าคุณลักษณะทุกคุณลักษณะดังนี้ $<a_1, a_2 ,..., a_n>$ โดยที่ ค่าเป้าหมายที่ต้องการของแต่ละอินสแตนซ์ เป็นค่าใดๆ ภายใน เซต V เมื่อ V มีสมาชิกเป็นค่าเป้าหมายที่ต้องการ ในที่นี้หมายถึงจำนวนกลุ่มของข้อมูล

นาอีฟเบย์เป็นการเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทของข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดีเหมาะกับกรณีของเซตตัวอย่างที่นำเสนอไปข้างต้น ที่มีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นตรงต่อกัน มีการจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในการจำแนกประเภทของข้อความ (Text Classification) การวิจัย (Diagnosis) และพบว่าสามารถใช้งานได้ดีไม่ต่างจากวิธีการจำแนกวิธีอื่น ๆ เป็นเหตุให้ผู้วิจัยเลือกใช้วิธีนี้ในงานวิจัยชิ้นนี้ เนื่องจากให้ประสิทธิภาพการทำงานที่ดี และวิธีการทำงานไม่ซับซ้อนเหมือนวิธีการอื่น ๆ

การกำหนดความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม V_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{a_1,a_2,...a_n\}$ หรือใช้สัญลักษณ์ว่า $P(a_1,a_2,...a_n)$ ตามสมการ (2.5)

$$P(V_j|a_1,a_2,\ldots,a_n) = \prod_{i=1}^{n} P(a_i|v_j) \quad (2.5)$$

โดยที่ $\prod$ หมายถึงผลคูณของค่า $P(a_i \mid v_j)$ เมื่อ I และ j มีค่าเท่ากับ $1, 2, 3, ..., n$

วิธีการเรียนรู้เบย์ อย่างง่ายไปใช้มีวิธีดังต่อไปนี้คือ

1) หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(v_j \mid a_1,a_2,\ldots,a_n)$ จากสมการมาคูณกับค่าความน่าจะเป็นของกลุ่มนั้นๆ คือ $P(v_j)$ ได้เท่ากับ V_{NB}

2) นำค่าที่ได้มาเปรียบเทียบกันกลุ่มที่มีความน่าจะเป็นสูงสุดคือกลุ่มที่ข้อมูลนั้นอยู่ และจะถูกจัดเข้าไป เขียนเป็นสมการได้ตามสมการ (2.6)

$$(v_j) \times = \prod_{i=1}^{n} P(a_i \mid v_j) : \; v_j \in V \qquad (2.6)$$

ตัวอย่างการใช้อัลกอริทึมการเรียนรู้เบย์อย่างง่าย แสดงตัวอย่างข้อมูลที่ต้องการจำแนกประเภทของข้อมูลในรูปที่ 2-8

Class

Attribute →	Like	Phukradung	Beautiful	Dirty	Noisy	Not	Result
Value	1	1	0	0	0	0	Yes
	0	1	1	0	0	0	Yes
	0	0	0	1	0	0	No
	0	0	0	0	1	1	No

รูปที่ 2-8 แสดงตัวจัดกลุ่มเอกสารในรูปแบบตาราง

การนำการเรียนรู้เบย์อย่างง่ายมาใช้ในการสร้างตัวจัดกลุ่มเอกสารโดยอาศัยการคำนวณหาค่าความน่าจะเป็นของคำสำคัญในแต่ละเอกสารดังตารางที่ 2-5

ตารางที่ 2-5 แสดงการหาค่าความน่าจะเป็นของคำสำคัญในแต่ละเอกสาร

	Word	P(Word)	*tf-idf*	P(Word) x *tf-idf*
Positive	Like	0.25	0.602	0.1505
	Phukradung	0.5	0.125	0.0625
	Beautiful	0.25	0.602	0.1505
Negative	Phukradung	0.25	0.125	0.0312
	Dirty	0.25	0.602	0.1505
	Noisy	0.25	0.602	0.1505
	Not	0.25	0.602	0.1505

3) k-nearest neighbor

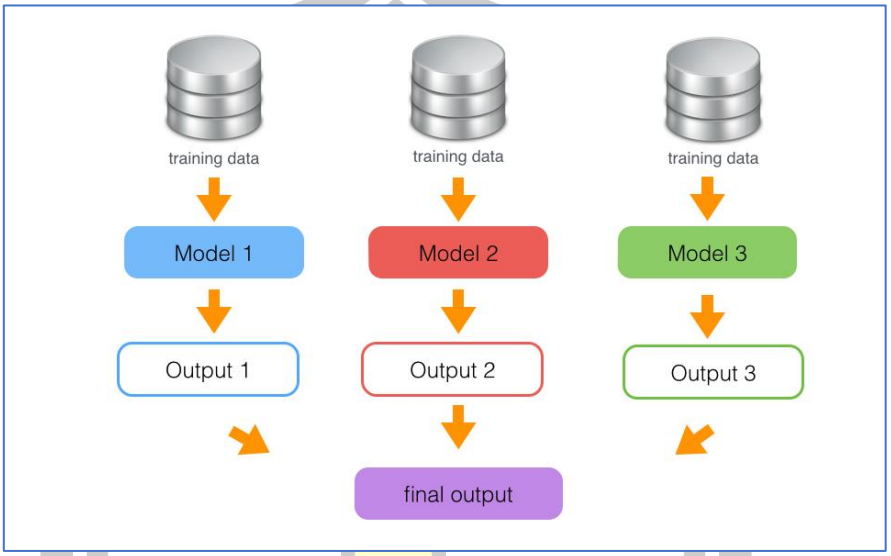
วิธีการ KNN [9] จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัว จากข้อมูลบนชุดข้อมูลตัวอย่าง จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วย แอททริบิวต์หลายตัวแปร Xi ซึ่งจะนำมาใช้ในการแบง กลุ่ม Yi โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอก จำนวนของกรณี (Case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ โดยบทความนี้กำหนด 1-KNN หมายถึง อัลกอริทึมนี้จะค้นหา 1 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (1 Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝน มาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทาง (Distance) ใกล้เคียงที่สุดออกมา K ตัวโดยใช้การวัดระยะทางแบบ Euclidean distance มีหลักการ คือ การวัดระยะทางระหว่างสองวัตถุ ถ้าวัตถุห่างกันมากแสดงว่า วัตถุนั้นมีความคล้ายกันน้อย ถ้ามีค่าน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก โดยที่ คา Pi แทน คุณสมบัติจากฐานข้อมูล qi แทนคุณสมบัติที่ผู้ใช้ระบุ [9] ดังแสดงในสมการ (2.7)

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (2.7)$$

2.1.11 การสร้างโมเดล Ensemble

เทคนิค Ensemble เป็นเทคนิคเป็นเทคนิคที่มีการนำมาใช้ใน learning model ใน machine learning โดยเฉพาะเป้าหมายการเพิ่มประสิทธิภาพการทำงานของโมเดล โดยการใช้

โมเดล classification หลายๆ โมเดล มาช่วยในการหาคำตอบ สามารถแสดง concept การทำงานของเทคนิค Ensemble ได้ดังรูปที่ 2-9



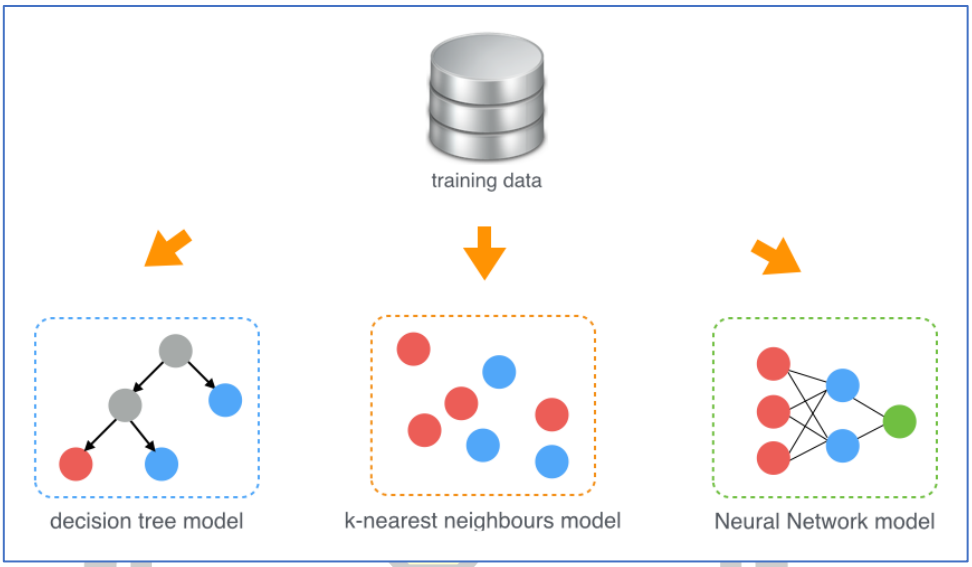
รูปที่ 2-9 แสดง concept การทำงานของ ensemble model

จากรูปที่ 2-9 จะเห็นว่าเป็นการนำข้อมูลชุดทดสอบ (training data) มาสร้างโมเดลต่าง ๆ โดยข้อมูลเทรนชุดทดสอบเหล่านี้จะเป็นข้อมูลชุดเดียวกันก็ได้ (เช่น วิธีการ Vote Ensemble) หรือจะเป็นข้อมูลที่ต่างกันก็ได้ (เช่น วิธี Bagging และ RandomForest) หลังจากได้โมเดลมาชุดหนึ่งแล้วจะนำไปทำนายข้อมูลที่ยังไม่รู้คำตอบ สำหรับการทำนายด้วยเทคนิค Ensemble ซึ่งมีหลายๆ โมเดลนี้ แต่ละโมเดลก็จะให้คำตอบออกมา ในขั้นตอนสุดท้ายเราจะต้องนำคำตอบเหล่านี้มารวมกันเพื่อดูว่าคำตอบไหนเหมาะสมที่สุด โดยอาจจะใช้วิธีการโหวต (vote) เลือกคำตอบที่ตอบตรงกันมากที่สุด

หลักการสร้างโมเดล Ensemble คือโมเดลที่สร้างควรจะมีความหลากหลายเพื่อให้ทำนายข้อมูลแบบต่าง ๆ กันได้มาก การสร้างโมเดลที่หลากหลายนี้ อาจจะทำได้โดยการใช้เทคนิค classification หลาย ๆ ประเภท หรือ การสร้างเทรนนิ่ง ดาต้า ที่มีลักษณะต่าง ๆ กัน เช่น มีตัวอย่างต่างกัน หรือมีแอตทริบิวต์ต่างกัน

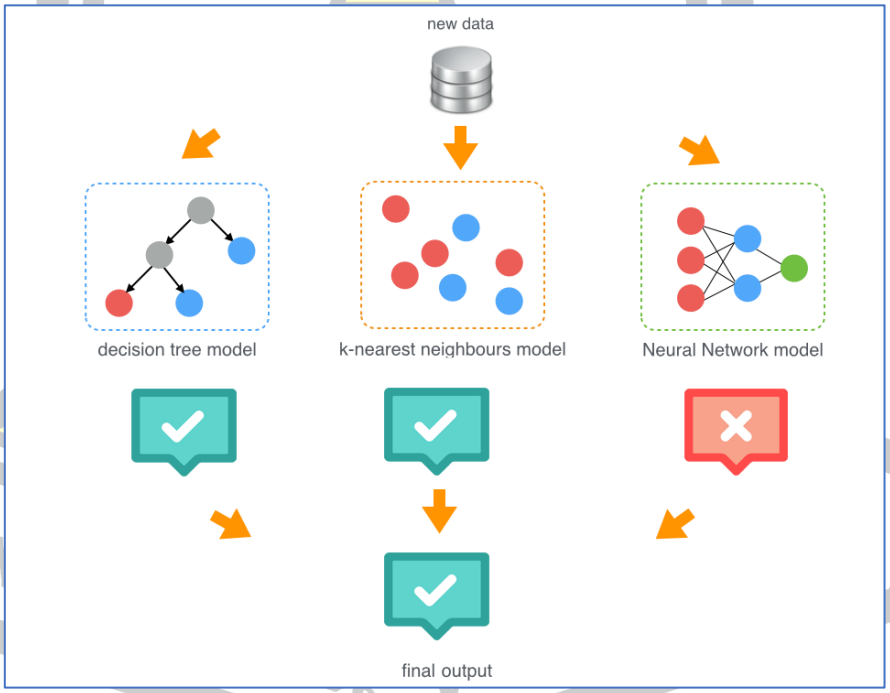
เทคนิค Ensemble มี 3 เทคนิค ดังต่อไปนี้

1) Vote Ensemble เป็นการใช้เทรนนิ่ง ดาต้า (training data) ชุดเดียวกันแต่สร้างโมเดลด้วยเทคนิคต่างๆ กัน แล้วนำผลลัพธ์ที่ได้ทั้ง 3 เทคนิค มาทำการโหวต เพื่อหาข้อสรุปของผลลัพธ์



รูปที่ 2-10 แสดงการทำงานของ Vote Model

หลังจากที่สร้างโมเดล Ensemble ด้วย 3 เทคนิคได้แล้ว ขั้นตอนถัดไป คือ การนำโมเดลที่สร้างได้ไปทำนายข้อมูลใหม่ ดังรูปที่ 2-11



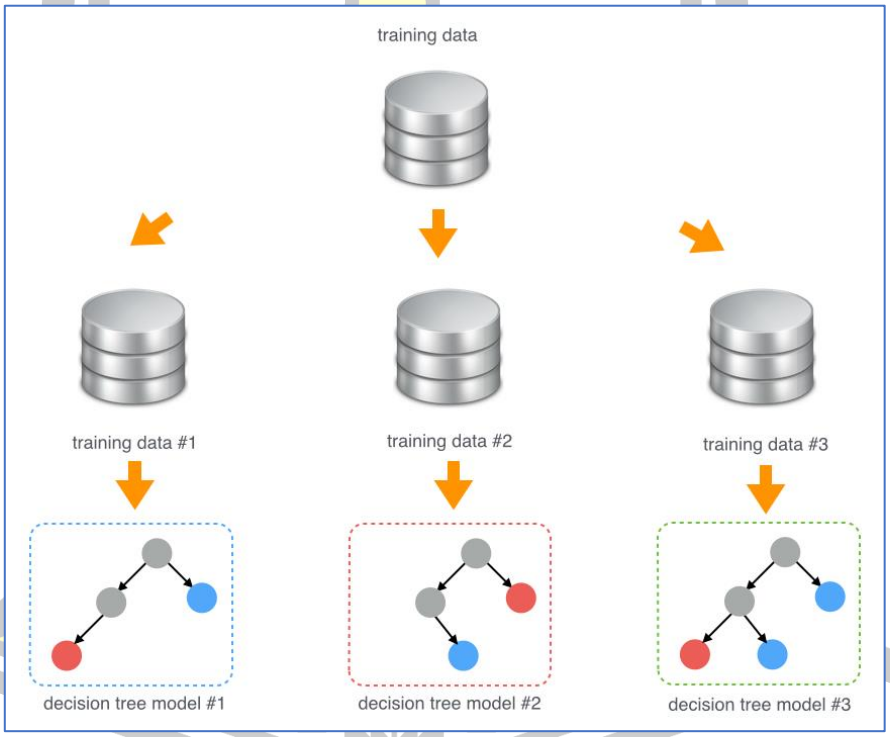
รูปที่ 2-11 การนำ Vote Model ไปใช้งาน

จากรูปที่ 2-11 จะมีข้อมูลใหม่ (new data) ที่ยังไม่รู้คลาส โมเดล Decision Tree (โมเดลที่ 1) ทำนายคำตอบออกมาว่าข้อมูลใหม่เป็น Positive โมเดล K-Nearest Neighbours (K-NN)

ทำนายคำตอบออกมาว่าข้อมูลใหม่เป็น Positive และ โมเดล Neural Network ทำนายเป็น Negative

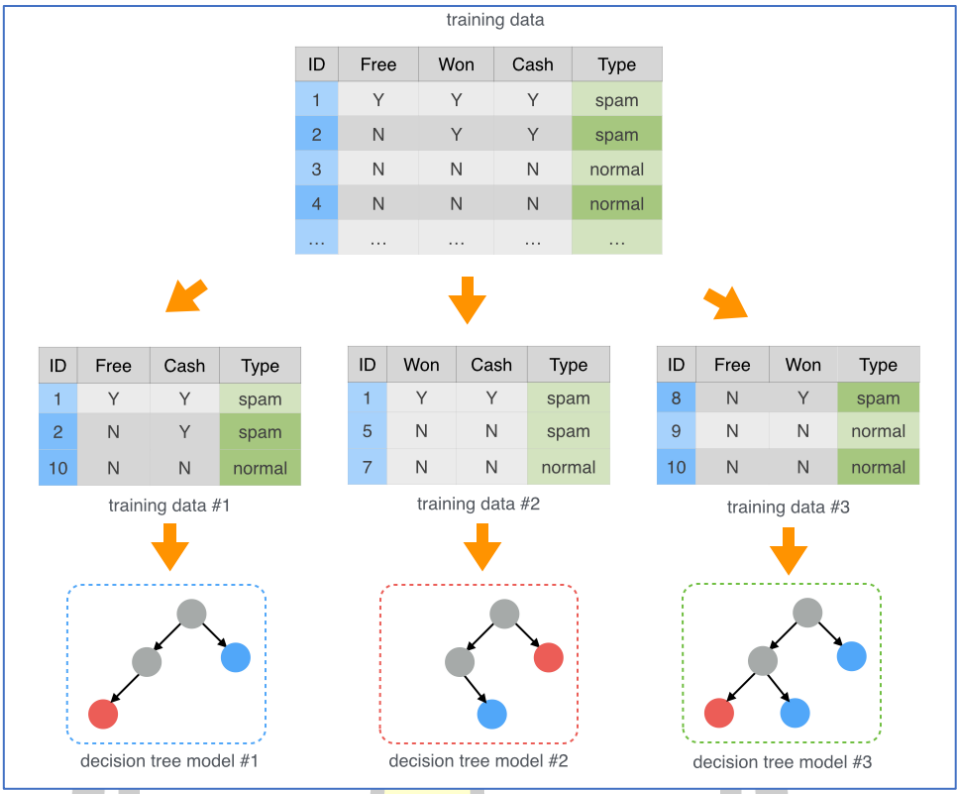
ดังนั้นจากผลการทำนายของทั้งสามโมเดลเราจะได้ว่าชุดข้อมูลเป็น Positive

2) Bootstrap Aggregating (Bagging) วิธีนี้แตกต่างจากวิธีการ Vote Ensemble โดยการสร้างโมเดลที่หลากหลายนั้นใช้การสุ่มข้อมูลตัวอย่างข้อมูลชุดทดสอบออกมาเป็นหลาย ๆ ชุด (แทนที่จะใช้ข้อมูลชุดทดสอบทั้งหมดแบบวิธีการ Vote Ensemble) แต่ใช้การสร้างโมเดลด้วยเทคนิค classification เดียวกัน เช่น ใช้เทคนิค Decision Tree หรือ เทคนิค Neural Networks ทั้งหมด ดังแสดงในรูปที่ 2-12 ซึ่งใช้เทคนิค Decision Tree ทั้งสามโมเดล แม้ว่าจะเป็นเทคนิค Decision Tree เหมือนกันแต่ข้อมูลที่ใช้ในการสร้างโมเดลต่างกันก็ทำให้โมเดลที่สร้างขึ้นมาได้มีลักษณะที่ต่าง



รูปที่ 2-12 แสดงการทำงานของ Bootstrap Aggregating

3) Random Forest เป็นวิธีการที่คล้ายกับ Bagging แต่เพิ่มการสร้างความหลากหลายของโมเดลด้วยการสุ่มแอตทริบิวต์ แทนที่จะเป็นการสุ่มเฉพาะข้อมูลตัวอย่างเพียงอย่างเดียวเหมือน Bagging และเทคนิคที่ใช้ในการสร้างโมเดลก็เป็นเพียงแค่ Decision Tree อย่างเดียวเท่านั้น ดังแสดงในรูป ซึ่งมีการสุ่มแอตทริบิวต์ต่าง ๆ กัน



รูปที่ 2-13 แสดงการทำงานของ Random Forest

2.1.12 การวัดประสิทธิภาพ (Evaluation)

เป็นขั้นตอนการประเมินโมเดลเพื่อใช้ในการจัดกลุ่มเอกสาร ก่อนการนำไปใช้งานจริง ซึ่งโดยทั่วไป จะใช้เทคนิคมาตรฐานที่เรียกว่า การวัดค่าความระลึก (Recall) 27], การวัดค่าความแม่นยำ (Precision) [27], และ การวัดค่าวัดประสิทธิภาพ (F-measure) [27] โดยใช้ Confusion matrix ดังนี้

	Prediction Positive	Prediction Negative	
Condition Positive	True Positive (TP)	False Negative (FN)	P_a
Condition Negative	False Positive (FP)	True Negative (TN)	N_a
	P_a	N_p	

ในงานวิจัยนี้ สามารถให้ความหมายของ TP, FP, TN และ FN ได้ดังนี้:

TP (true positive) หมายถึง จำนวนเอกสารที่จำแนกได้ว่าอยู่ในคลาส positive และถูกต้อง

TN (true Negative) หมายถึง จำนวนเอกสารที่จำแนกได้ว่าอยู่ในคลาส negative และถูกต้อง

FP (false positive) หมายถึง จำนวนเอกสารที่จำแนกได้ว่าอยู่ในคลาส positive แต่ไม่ถูกต้อง

FN (true Negative) หมายถึง จำนวนเอกสารที่จำแนกได้ว่าอยู่ในคลาส negative แต่ไม่ถูกต้อง

การวัดค่าความระลึก (Recall) คือ เป็นอัตราส่วนของเอกสารที่ทำนายได้ จากเอกสารทั้งหมดที่มีอยู่สามารถแสดงสมการได้ดังสมการ (2.8)

$$Recall = \frac{TP}{TP + FN} \tag{2.8}$$

การวัดค่าความแม่นยำ (Precision) คือ เป็นอัตราส่วนของเอกสารที่ทำนายได้และถูกต้องส่วนด้วยจำนวนของเอกสารที่ทำนายได้สามารถแสดงสมการได้ดังสมการ (2.9)

$$Precision = \frac{TN}{TN + FN} \tag{2.9}$$

การวัดค่าวัดประสิทธิภาพ (F-measure) คือการวัดค่าความถูกต้อง จากการพิจารณาค่าความสัมพันธ์ระหว่างค่าความระลึกและค่าความแม่นยำสามารถแสดงสมการได้ดังสมการ (2.10)

$$F - measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{2.10}$$

โดยที่ค่า *F* จะมีค่าระหว่าง 0 ถึง 1 ซึ่งถ้าหาก *F* มีค่าใกล้เคียง 1 มากเท่าไหร่ก็จะหมายถึงการจัดกลุ่มเอกสารนั้นมีประสิทธิภาพและมีความถึงต้องมากขึ้นเท่านั้น

2.2 งานวิจัยที่เกี่ยวข้อง

Chunyong Yin และคณะ [23] ได้นำเสนอกระบวนการจำแนกเอกสารข้อความสั้น ด้วยเทคนิค Semi-Supervised Learning และ SVM โดยคณะผู้วิจัยได้เล็งเห็นว่า ข้อความสั้นกำลังได้รับความนิยมและถูกใช้มากที่สุดในปัจจุบัน เช่น การส่งข้อความถึงกันในแชท การตั้งสถานะในสื่อสังคมออนไลน์ต่างๆ รวมไปถึงการแสดงความคิดเห็นต่อภาพยนตร์ที่รับชม ซึ่งข้อมูลทั้งหมดเป็นข้อมูลที่มีขนาดใหญ่ (Big Data) แต่การที่จะสกัดเอาคุณลักษณะของข้อความสั้นเป็นไปได้ยากเนื่องจากจำนวนคำที่มีอยู่น้อย ในงานวิจัยนี้ คณะผู้วิจัยจึงได้นำเสนอวิธีการสกัดคุณลักษณะของข้อความสั้น สำหรับการจำแนกเอกสารด้วยเทคนิค Semi-Supervised Learning และ SVM ผลการทดลองแสดงให้เห็น

ว่าการสกัดคุณลักษณะของข้อความโดยวิธีการแบบดังเดิมมีประสิทธิภาพที่ดี แต่มีปัญหากับข้อความที่มีจำนวนคำน้อยหรือข้อความสั้น แต่เมื่อนำเทคนิค Semi-Supervised Learning และ SVM มาปรับปรุงกระบวนการแบบดั้งเดิม แสดงให้เห็นว่าประสิทธิภาพในการจำแนกเอกสารดีขึ้นกว่าเดิม

Quan Yuan, Gao Cong และ Nadia M. Thalmann [31] ได้นำเสนอกระบวนการเพิ่มประสิทธิภาพของการจำแนกเอกสารข้อความสั้นให้รองรับข้อความรูปแบบใหม่หรือคำใหม่ๆ ที่มีเพิ่มขึ้นเรื่อยๆ ด้วยเทคนิคนาอีฟเบย์ (Naive Bayes) ผลการทดลองในข้อมูลที่มีขนาดใหญ่ และเป็นข้อมูลที่อยู่ในสถานการณ์ปัจจุบัน แสดงให้เห็นว่าการใช้ smoothing methods สามารถปรับปรุงประสิทธิภาพการทำงานของนาอีฟเบย์ได้เป็นอย่างดี และในงานวิจัยนี้ยังได้ศึกษาผลของการใช้ข้อมูลชุดเรียนรู้ที่มีขนาดใหญ่และเป็นข้อความขนาดยาวในการสอนการเรียนรู้อีกด้วย

Liangliang Li และ Shouning Qu [24] ได้นำเสนอปัญหาของการจำแนกเอกสารที่มีข้อความสั้น โดยมุ่งประเด็นไปที่ปัญหาในส่วนของการคัดเลือกคุณลักษณะของข้อความ โดยในงานวิจัยนี้คณะผู้วิจัยได้เลือกใช้เทคนิค ITC Algorithm ในการคัดเลือกคุณสมบัติของข้อความ แทนการคัดเลือกคุณลักษณะโดย TFIDF แบบทั่วไป โดยนำเสนอปัญหาของ ITC Algorithm และปรับปรุงกระบวนการของ ITC Algorithm ให้ดีกว่าเดิมตามลักษณะของการจำแนกข้อความสั้น ในขณะเดียวกันก็จะผสมผสานตามลักษณะแนวคิดของเอนโทรปี (Entropy) ด้วยตำแหน่งของการกระจายน้ำหนัก ผลการทดลองแสดงให้เห็นว่าการปรับปรุงวิธีการคัดเลือกคุณลักษณะของ ITC Algorithm สอดคล้องกับการคัดเลือกคุณลักษณะของข้อความสั้น และเป็นไปตามขั้นตอนการคัดเลือกคุณลักษณะของ TFIDF แบบเดิม

Chen Mengen และคณะ [32] ได้เล็งถึงปัญหาด้านความถี่ของคำในเอกสารประเภทข้อความสั้นที่กำลังได้รับความนิยมและมีข้อมูลเพิ่มขึ้นเรื่อยๆ ซึ่งหากข้อมูลมีความถี่น้อย จะทำให้เป็นอุปสรรคเป็นอย่างมากในกระบวนการของการเรียนรู้ของเครื่อง (Machine Learning) และการทำเหมืองข้อความ (Text Mining) เนื่องจากหากจำนวนความถี่ของข้อความมีน้อย การสกัดคุณลักษณะของข้อความก็จะได้ออกน้อยเช่นกัน ซึ่งจะเป็นอุปสรรคในขั้นตอนการสร้างโมเดล ทำให้ได้โมเดลที่มีประสิทธิภาพน้อย ในงานวิจัยนี้ ผู้วิจัยได้นำเสนอกระบวนการจำแนกเอกสารของข้อความสั้นด้วย Learning Multi-Granularity Topics ผลการทดลองพบว่า วิธีการของงานวิจัยนี้ลดข้อผิดพลาดจากวิธีการพื้นฐาน 20.25%

Bharath Sriram [33] ได้นำเสนอกระบวนการจำแนกเอกสารข้อความสั้นโดยใช้ข้อมูลจากทวิตเตอร์ (Twitter) เนื่องจากว่าผู้วิจัยเห็นว่าข้อมูลจากทวิตเตอร์นั้นล้วนแล้วแต่เป็นข้อมูลดิบ เป็นข้อความที่เป็นภาษาธรรมชาติเป็นส่วนใหญ่ ผู้วิจัยได้จำแนกประเภทของเอกสารโดยใช้เทคนิคถุงคำ

(Bag-Of-Word) ในการจำแนกประเภทของเอกสาร โดยตั้งหมวดหมู่ของประเภทไว้คือ ข่าว, กิจกรรม , ข้อเสนอ, ความคิดเห็น และ ข้อความส่วนตัว การจำแนกประเภทของเอกสารจะอ้างอิงจากข้อมูลภายนอก คือ วิกิพีเดีย และ เวิร์ดเน็ท (WordNet) ผลการทดลองอยู่ในระดับดีเมื่อในข้อมูลอ้างอิงมีข้อความสั้นเป็นจำนวนมาก

Chunyong Yin และคณะ [34] ได้ทำการนำเสนอวิธีการใหม่โดยใช้ SVM Method สำหรับการจำแนกข้อความสั้นโดยใช้การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) โดยคณะผู้วิจัยได้มองถึงข้อมูลข้อความสั้นที่เพิ่มขึ้นมาเรื่อยๆในปัจจุบัน เนื่องจากมีโปรแกรมประยุกต์ทางสังคม (Social Software) และเว็บไซต์ที่เปิดให้ผู้คนแสดงความคิดเห็นกันอย่างแพร่หลาย ทำให้เกิดข้อมูลขนาดใหญ่ ดังนั้นการสกัดหาคุณลักษณะสำคัญของข้อความสั้นในข้อมูลที่มีขนาดใหญ่จึงมีความสำคัญเป็นอย่างยิ่ง แต่อย่างไรก็ตามคณะผู้วิจัยได้ค้นพบว่า การที่จะสกัดคุณลักษณะที่สำคัญในข้อความสั้นนั้นเป็นไปได้ยาก เพราะมีปัญหาด้านจำนวนของคุณลักษณะของข้อความมีน้อย ดังนั้นในงานวิจัยนี้จึงนำเสนอการจำแนกข้อความสั้นโดยใช้การเรียนรู้แบบกึ่งมีผู้สอน เพื่อปรับปรุงวิธีการจำแนกข้อความสั้นเพื่อสกัดข้อความที่สำคัญออกมาจากข้อมูลที่มีขนาดใหญ่ (Big Data) ผลการทดลองแสดงให้เห็นว่าการใช้เทคนิคการเรียนรู้แบบกึ่งมีผู้สอน สามารถจำแนกเอกสารได้ดียิ่งขึ้น

Mike Thelwall และคณะ [35] ได้นำเสนอวิธีการใหม่ในการจำแนกข้อความที่เป็นความรู้สึก (Sentiment Strength Detection) โดยใช้เทคนิค SentiStrength เพื่อสกัดหาคุณสมบัติของคำ จากข้อความภาษาอังกฤษที่มีลักษณะเป็นข้อความสั้น ซึ่งอาศัยการใช้ไวยากรณ์ของคำและรูปแบบการสะกดคำ โดยนำไปประยุกต์ใช้กับข้อความการแสดงความคิดเห็น ผลการทดลองสามารถทำนายอารมณ์ที่เป็นเชิงบวกได้ความแม่นยำ 60.6% และ อารมณ์เชิงลบ 72.8%

บทที่ 3

วิธีดำเนินการวิจัย

งานวิจัยนี้ได้นำเสนอกระบวนใหม่ในสร้างตัวจำแนกความรู้สึกจากเอกสารข้อความขนาดสั้นที่ โดยกระบวนการแบบผสมผสาน (Hybrid Method) ที่พัฒนามาจากเทคนิคเหมืองข้อมูล (Data mining techniques) และการประมวลผลธรรมชาติ (Natural Language Processing: NLP) ซึ่งระเบียบวิธีการวิจัยมีขั้นตอนดังนี้

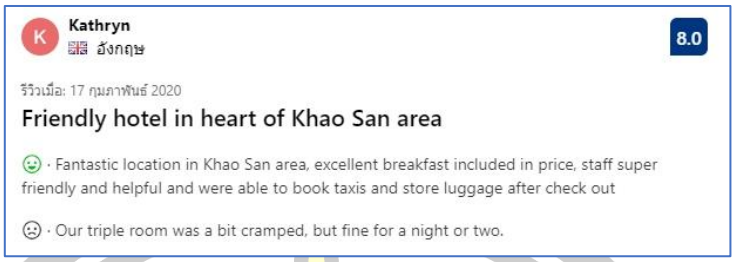
3.1 ชุดข้อมูลที่ใช้ (Dataset)

ในงานวิจัยนี้ ได้ใช้ชุดข้อมูลที่เป็นข้อความสั้นจาก ชุดข้อมูลที่เป็นบทวิจารณ์หรือการแสดงความคิดเห็นในเว็บไซต์ที่ให้บริการจองที่โรงแรมที่พักทั้งในและต่างประเทศ ซึ่งข้อมูลแต่ละชุดมีกระบวนการการเก็บรวบรวมดังนี้

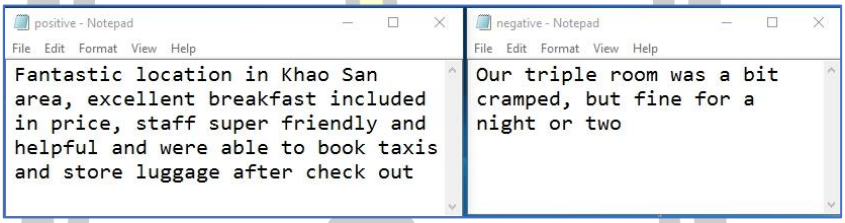
3.1.1 ชุดข้อมูลที่เป็นบทวิจารณ์หรือการแสดงความคิดเห็นในเว็บไซต์

ชุดข้อมูลชุดนี้จะเป็นบทวิจารณ์หรือข้อความการแสดงความคิดเห็นเกี่ยวกับโรงแรม จากผู้คนที่ทำการเขียนบทวิจารณ์ลงในเว็บไซต์บุ๊คกิ้ง (www.booking.com) ซึ่งเป็นเว็บไซต์เกี่ยวกับการให้บริการจองโรงแรมผ่านเว็บไซต์ โดยเก็บรวบรวมข้อมูลเป็นภาษาอังกฤษจำนวนไม่น้อยกว่า 8,000 เอกสาร แบ่งเป็นความคิดเห็นเชิงบวกไม่น้อยกว่า 4,000 เอกสาร และเป็นความคิดเห็นเชิงลบไม่น้อยกว่า 4,000 เอกสาร

การกำหนดกลุ่มของเอกสารว่าอยู่ในกลุ่มความคิดเห็นเชิงบวกหรือความคิดเห็นลบ ซึ่งจากการสำรวจในเว็บไซต์บุ๊คกิ้งแล้วพบว่าในการแสดงความคิดเห็นของผู้เข้าใช้บริกหาร 1 คน จะสามารถให้ความคิดเห็นด้านบวก และ ด้านลบ ดังนั้นงานวิจัยนี้จึงถือว่าความคิดเห็นนั้น ผู้ให้ความคิดเห็นได้แสดงความคิดเห็นทั้งสองกลุ่ม และถือว่าเป็นการจัดกลุ่มข้อมูลให้ทั้งสองกลุ่มแล้ว โดยที่เป็นความคิดเห็นของผู้ใช้บริการเอง ตัวอย่างข้อความการแสดงความคิดเห็นในเว็บไซต์บุ๊คกิ้งที่มีการกำหนดกลุ่มของความรู้สึกโดยผู้ใช้เองแสดงได้ดังรูปที่ 3-1 และถูกจัดเก็บลงใน text file ดังแสดงตัวอย่างในรูปที่ 3-2

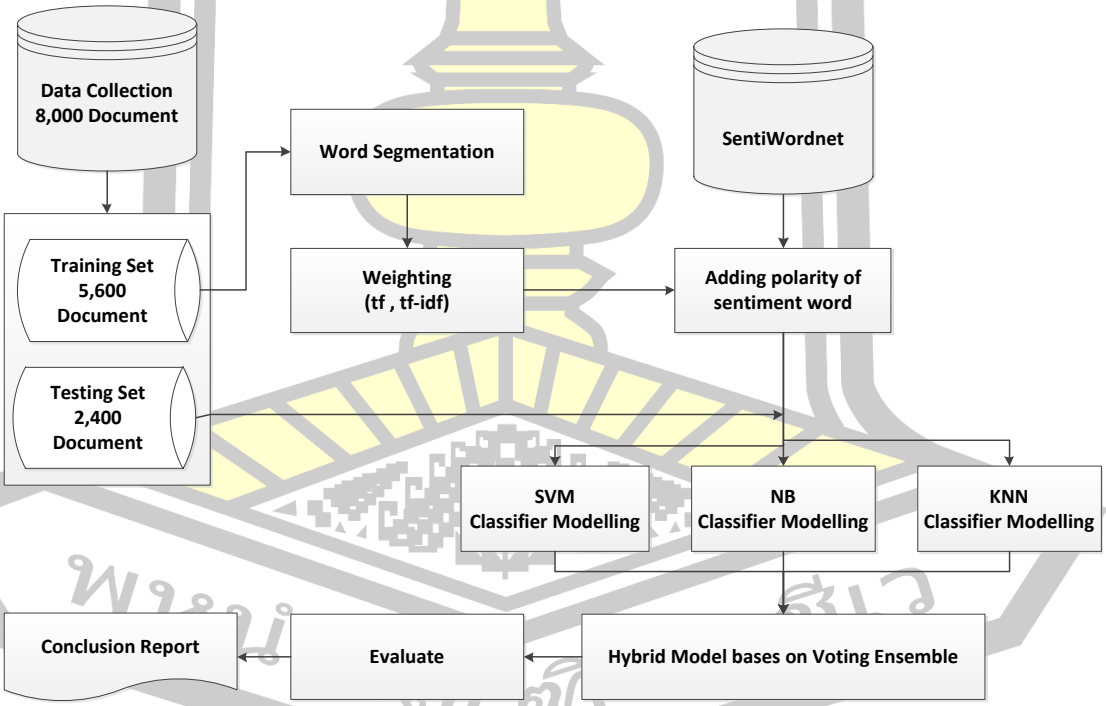


รูปที่ 3-1 แสดงตัวอย่างการแสดงความคิดเห็นของผู้ใช้บริการในเว็บไซต์ booking.com



รูปที่ 3-2 แสดงตัวอย่างการจัดเก็บความคิดเห็นลงใน text file

3.2 กระบวนการดำเนินงานวิจัยที่นำเสนอ (Research Methodology)



รูปที่ 3-3 แสดงกระบวนการดำเนินการวิจัย

พิจารณาจากรูปที่ 3-3 แสดงกระบวนการดำเนินงานวิจัย โดยประกอบด้วยขั้นตอนต่าง ๆ ดังต่อไปนี้

3.2.1 การเตรียมข้อมูลก่อนการประมวลผล (Data Pre-processing)

ในขั้นตอนนี้ เป็นขั้นตอนการเตรียมข้อมูลก่อนเข้าสู่กระบวนการประมวลผลสร้างโมเดล โดยเป็นขั้นตอนที่สำคัญมากขั้นตอนหนึ่ง เนื่องจากหากเตรียมข้อมูลได้ไม่ดี หรือ ข้อมูลไม่พร้อม อาจส่งผลกระทบไปถึงกระบวนการสร้างโมเดล ทำให้ประสิทธิภาพการทำงานของโมเดลไม่ดีตามไปด้วย ซึ่งขั้นตอนการเตรียมข้อมูลมี 5 ขั้นตอน สมมุติให้มีเอกสารทั้งหมด 5 เอกสารได้แก่

D1: Clean and safe.

D2: The staff service Terrible

D3: The pool is clean, the staff are all smiling.

D4: Bad internal equipment, terrible service

D5: Good service staff, always smiling

ขั้นตอนที่ 1 ตัดคำ เป็นขั้นตอนในการการแบ่งคำแต่ละคำออกจากประโยค โดยจะใช้ช่องว่างในการแบ่งขอบเขตของคำ โดยสามารถตัดคำจากเอกสารทั้งหมดได้ดังนี้

D1: Clean | and | safe

D2: The | staff | service Terrible

D3: The | pool | clean | staff | smiling

D4: Bad | internal | equipment | terrible | service

D5: Good | service | staff | always | smiling.

ขั้นตอนที่ 2 ตัดคำหยุด (Stop Word) เป็นขั้นตอนการกำจัดคำที่ไม่มีนัยสำคัญหรือไม่ส่งผลใด ๆ กับการประมวลผลออกไป ซึ่งจะทำให้การประมวลผลเร็วยิ่งขึ้น โดยเมื่อตัดคำหยุดออกแล้ว จะได้เอกสารทั้งหมดดังนี้

D1: clean | safe

D2: staff | service | terribl

D3: pool | clean | staff | smile

D4: bad | intern | equip | terribl | servic

D5: good | service | staff | always | smile

ในงานวิจัยนี้ เมื่อทำการตัดคำ และ ตัดคำหยุดออกเป็น จะได้คุณลักษณะของคำมาทั้งสิ้น 6,763 คำ ซึ่งถือว่าเพียงพอต่อการนำไปใช้งาน

ขั้นตอนที่ 3 การนำเสนอเอกสาร (Document Representation) เป็นขั้นตอนในการนำเสนอความสัมพันธ์ระหว่างคำและเอกสาร ให้อยู่ในรูปแบบเวกเตอร์ (Vector) โดยมีการให้น้ำหนักแบบ *tf* และ *tf-idf* ซึ่งสามารถแสดงการให้น้ำหนักแต่ละรูปแบบดังนี้

1) การให้น้ำหนักแบบ tf

tf เป็นการหาความถี่ของคำที่ปรากฏในเอกสารเป็นการหาความถี่ของคำแต่ละคำที่อยู่ในเอกสารนั้นๆ ว่าพบกี่ครั้ง แสดงตัวอย่างการหาค่า tf ดังตารางที่ 3-1

ตารางที่ 3-1 ตารางการให้น้ำหนักคำด้วย tf

alway	clean	equip	good	intern	pool	safe	servic	smile	staff	terribl
0	0.30103	0	0	0	0	0.30103	0	0	0	0
0	0	0	0	0	0	0	0.30103	0	0.30103	0.30103
0	0.30103	0	0	0	0.30103	0	0	0.30103	0.30103	0
0	0	0.30103	0	0.30103	0	0	0.30103	0	0	0.30103
0.30103	0	0	0.30103	0	0	0	0.30103	0.30103	0.30103	0

2) การให้นำหนักแบบ tf-idf

tf-idf เป็นวิธีการสร้างตัวแทนเอกสารในรูปแบบของเวกเตอร์เพื่อใช้ในการจัดกลุ่มของเอกสารให้ตรงกับหมวดหมู่ที่ถูกกำหนดไว้ โดย idf สามารถหาได้จาก สมการ (3.1)

$$idf = log\,(N/df) \tag{3.1}$$

โดยที่ N คือจำนวนเอกสารทั้งหมดในกลุ่ม และ df คือจำนวนเอกสารที่มีคำ ๆ นั้นปรากฏอยู่

และ tf-idf หาได้จากสมการ (3.2)

$$tfidf = tf \times idf \tag{3.2}$$

ต่อมาภายหลังจากการหาค่า tf และ idf เรียบร้อยแล้ว จะเข้าสู่ขั้นตอนของการให้นำหนักคำใน BOW หรือค่าของ tf ที่หาได้ในตารางที่ 3-2

ตารางที่ 3-2 ตารางแสดง BOW ของคำและน้ำหนักของคำในแต่ละเอกสารด้วย tf-idf

alway	clean	equip	good	intern	pool	safe	servic	smile	staff	terribl
0	0.11979	0	0	0	0	0.21041	0	0	0	0
0	0	0	0	0	0	0	0.06678	0	0.06678	0.11979
0	0.11979	0	0	0	0.21041	0	0	0.11979	0.06678	0
0	0	0.21041	0	0.21041	0	0	0.06678	0	0	0.11979
0.21041	0	0	0.21041	0	0	0	0.06678	0.11979	0.06678	0

3.2.2 ปรับค่าของคำด้วยค่าขั้วความรู้สึก (Polarity of Sentiment Word)

ในวิทยานิพนธ์ฉบับนี้ได้เลือกใช้คลังคำแสดงค่าขั้วความรู้สึกในภาษาอังกฤษที่มีชื่อว่า SentiWordNet โดยนำมาจากมหาวิทยาลัยในอิตาลีชื่อ Istituto di Scienza e Tecnologie dell' Informazione หรือ ISTI ซึ่งเป็นคลังคำที่พัฒนาโดย Andrea Esuli และ Fabrizio Sebastiani ตัวอย่างคำที่ถูกจัดเก็บในคลังคำนี้ สามารถแสดงได้ดัง

POS	ID	PosScore	NegScore	SynsetTerms	Gloss
a	00002098	0	0.75	unable#1	(usually followed by `
a	00003700	0.25	0	dissilient#1	bursting open with for
a	00003829	0.25	0	parturient#2	giving birth; "a partu
a	00005107	0.5	0	uncut#7 full-length#2	complete; "the
a	00005205	0.5	0	absolute#1	perfect or complete or
a	00005473	0.75	0	direct#10	lacking compromising o
a	00005599	0.5	0.5	unquestioning#2 implicit#2	being
a	00005718	0.125	0	infinite#4	total and all-embracin
a	00005839	0.5	0.125	living#3	(informal) absolute; "
a	00006032	0.25	0.5	relative#1 comparative#2	estima
a	00006777	0.375	0	sorbefacient#1 absorbefacient#1	induci
a	00006885	0	0.75	assimilatory#1 assimilative#2 assimilating#1	

รูปที่ 3-4 แสดงตัวอย่างการจัดเก็บของ SentiWordNet

จากค่าขั้วความรู้สึกที่แสดงตัวอย่างในรูปที่ 3-4 ในการปรับค่าของคำด้วยค่าขั้วความรู้สึกจะนำค่าน้ำหนักที่ได้จากการ weighting ด้วย tf และ tf-idf มาบวกกับค่าขั้วความรู้สึกของแต่ละคำ สามารถแสดงตัวอย่างค่าน้ำหนักของคำที่บวกกับค่าขั้วความรู้สึกได้ดังตารางที่ 3-3 และ ตารางที่ 3-4

ตารางที่ 3-3 แสดงการปรับค่าของคำน้ำหนักด้วย tf ด้วย sentiment polarity

alway	clean	equip	good	intern	pool	safe	servic	smile	staff	terribl
0	0.45235	0	0	0	0	0.56783	0	0	0	0
0	0	0	0	0	0	0	0.84245	0	0.74535	0.6401
0	0.45235	0	0	0	0.56709	0	0	0.85543	0.74535	0
0	0	0.53403	0	0.74353	0	0	0.84245	0	0	0.6401
0.50313	0	0	0.54353	0	0	0	0.84245	0.85543	0.74535	0

ตารางที่ 3-4 แสดงการปรับค่าของคำน้ำหนักด้วย tf-idf ด้วย sentiment polarity

alway	clean	equip	good	intern	pool	safe	servic	smile	staff	terribl
0	0.42679	0	0	0	0	0.71781	0	0	0	0
0	0	0	0	0	0	0	0.36678	0	0.34248	0.54219
0	0.42679	0	0	0	0.54421	0	0	0.45249	0.34248	0
0	0	0.45231	0	0.54351	0	0	0.36678	0	0	0.54219
0.54241	0	0	0.56231	0	0	0	0.36678	0.45249	0.34248	0

จากตารางที่ 3-3 และ ตารางที่ 3-4 จะเห็นว่าจะได้ค่าน้ำหนักของคำที่มีการเพิ่มน้ำหนักจาก sentiment polarity ที่จะช่วยบ่งบอกถึงความน่าจะเป็นของคำในแต่ละคลาส

3.2.3 การสร้างโมเดลจำแนกความรู้สึกของข้อความสั้นแบบผสมผสาน

จากการศึกษาที่ผ่านมาพบว่าการวิเคราะห์ความรู้สึกในลักษณะของการจำแนกเอกสาร (Text Classification) สามารถใช้ได้กับหลายเทคนิควิธี ในงานวิจัยนี้ ได้ใช้เทคนิคการจำแนกความรู้สึกแบบมีผู้สอน 3 วิธี คือ Support Vector Machines (SVM) [15, 16] นาอีฟเบย์ (Naïve Bayes) [13, 14] และ การหาเพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) [21, 22] เนื่องจากเทคนิควิธีทั้ง 3 วิธี ได้รับความนิยมในงานวิจัยเป็นอย่างมากและสามารถใช้กับงานจำแนกเอกสารได้ในระดับที่น่าพอใจในการจำแนกความรู้สึกหรือความคิดเห็น ซึ่งรายละเอียดแต่ละเทคนิควิธีมีดังนี้

3.2.3.1 การสร้างโมเดลจำแนกความรู้สึกของข้อความสั้นด้วยซัพพอร์ตเว็กเตอร์แมชชีน

เป็นขั้นตอนการสร้างโมเดลจำแนกความรู้สึก เพื่อจัดกลุ่มของเอกสารว่าอยู่ในกลุ่มใด โดยในงานวิจัยนี้ได้กำหนดกลุ่มของความรู้สึก 2 กลุ่ม คือ กลุ่มความรู้สึกเชิงบวก และ กลุ่มความรู้สึกเชิงลบ โดยจะใช้อัลกอริทึม ซัพพอร์ทเวกเตอร์แมชชีน (support Vector Machines) ซึ่งเป็นเทคนิคที่ได้รับความนิยมในการใช้จำแนกเอกสาร [15, 16, 38] ด้วยสมการ (3.3)

$$\sum_{x=1}^{i} w^T x_i + b \tag{3.3}$$

การจัดกลุ่มเอกสารด้วยซัพพอร์ทเวกเตอร์แมชชีนมีขั้นตอนดังต่อไปนี้

1) หาค่า y ของเอกสารที่นำเข้ามา โดยที่ค่าของ y∈{-1,1} หาได้จากสมการ (3.4)

$$w^T x + b \tag{3.4}$$

จาก wx + b = 0

C_q = - wx

= - [(1*1)+(0.031*0.031)+(0.031*0.031) +(0.031*0.031)
+(0.031*0.031) +(0.031*0.031) +(0.031*0.031)
+(0.031*0.031) +(0.031*0.031) +(0.031*0.031)
+(0.031*0.031) +(0.031*0.031)]

= -[1+0.00096+0.00096+0.00096+0.00096+0.00096+
0.00096+0.00096+0.00096+0.00096+0.00096+0.00096]
= -1.011

โดยที่ถ้าค่าของ $w^T x + b > 0$ จะกำหนดให้ค่า y= +1 จะจัดอยู่ในกลุ่มความคิดเห็นลบ

และถ้าค่าของ $w^T x + b < 0$ จะกำหนดให้ค่า y= -1 จะจัดอยู่ในกลุ่มความคิดเห็นเชิงบวก

2) คำนวณเส้นตรงที่ใช้ในการแบ่งกลุ่มของเอกสาร ที่เรียกว่า Optimal Hyperplane จากสมการ (3.6)

$$wx + b = 0$$

D1 = wx+b
= [(1*0)+(0.031*0.031)+(0.031*0)+(0.031*0)+(0.031*0)
+(0.031*0)+(0.031*0.031)+(0.031*0)+(0.031*0) +(0.031*0)+(0.031*0)]+(-1.011)
= -1.01 จะได้ y = -1

D2 = wx+b
= [(1*0)+(0.031*0)+(0.031*0)+(0.031*0)+(0.031*0)
+(0.031*0)+(0.031*0)+(0.031*0.031)+(0.031*0.031)
+(0.031*0.031)+(0.031*0)]+(-1.011)
= -1.09 จะได้ y = -1

D3 = wx+b
= [(1*0)+(0.031*0.031)+(0.031*0)+(0.031*0)+(0.031*0)
+(0.031*0.031)+(0.031*0)+(0.031*0)+(0.031*0.031)
+(0.031*0.031)+(0.031*0)]+(-1.011)
= -1.07 จะได้ y = -1

D4 = wx+b
= [(1*0)+(0.031*0)+(0.031*0.031)+(0.031*0)+(0.031*0.031)
+(0.031*0)+(0.031*0)+(0.031*0.031)+(0.031*0)
+(0.031*0.031)+(0.031*0.031)]+(-1.011)

= 1.07 จะได้ y = +1

D4 = wx+b

= [(1*0)+(0.031*0.031)+(0.031*0)+(0.031*0)+(0.031*0.031)
+(0.031*0)+(0.031*0)+(0.031*0)+(0.031*0.031)
+(0.031*0.031)+(0.031*0.031)]+(-1.011)

= 1.07 จะได้ y = +1

3) นำค่าที่ได้จาก 1) และ 2) ไปเขียนบนเส้นตรงตามแนวแกนตั้งและแกนนอน เพื่อที่จะหาจุดที่ใกล้เส้น Optimal Hyperplane ที่สุด

4) หาระยะห่างระหว่างเส้นขอบทั้งสองโดยจะเลือกเอาค่าระยะทางที่ห่างจากเส้น Optimal Hyperplane ที่น้อยที่สุดเป็นตัวแทนในการจำแนกเอกสาร (Support Vector) นั่นคือ เอกสาร D1 และ เอกสาร D3

3.2.3.2 การสร้างโมเดลจำแนกความรู้สึกของข้อความสั้นด้วยนาอีฟเบย์

การสร้างตัวจัดกลุ่มเอกสาร ซึ่งขั้นตอนนี้เป็นฟังก์ชันสำหรับการสร้างโมเดล (Modeling) เพื่อการวิเคราะห์ความรู้สึกจากข้อความ ด้วยอัลกอริทึม Naïve Bayes ซึ่งการจัดกลุ่มข้อความเป็นเทคนิคที่นำมาประยุกต์ใช้เพื่อการวิเคราะห์ความรู้สึกแบบสองกลุ่ม ด้วยสมการของเบย์ในสมการ (3.6)

$$P(v_j \mid a_1, a_2, \ldots, a_n) = \prod_{i=1}^{n} P(a_i \mid v_j) \quad (3.6)$$

จากเอกสารทั้งหมด 5 เอกสารมีเอกสารที่เป็น Positive จำนวน 3 เอกสารและ เอกสารที่เป็น Negative จำนวน 2 เอกสาร ดังนั้น กลุ่มที่เป็น Positive มีจำนวนเอกสารทั้งหมด 2 เอกสารจากทั้งหมด 4 เอกสาร และกลุ่มที่เป็น Negative มีจำนวนเอกสารทั้งหมด 2 เอกสารจากทั้งหมด 4 เอกสาร ตามลำดับ จากนั้นจะหาค่าความน่าจะเป็นของคำสำคัญที่อยู่ในแต่ละเอกสารที่แยกคลาสออกจากกันจะได้ความน่าจะเป็นออกมาด้วยสมการ (3.7)

$$P(a_i|v_j) = \frac{count(a_i, v_j)}{count(v_i)} \quad (3.7)$$

เมื่อ $count(a_i,v_j)$ คือ ค่าความถี่ของคำที่ i ในกลุ่มที่ j

และ $count(v_i)$ คือ ค่าความถี่รวมในกลุ่มที่ j

แต่ในบางครั้งการหาค่าความน่าจะเป็นของ Naïve Bayes นั้นอาจจะมีกรณีที่ความถี่ของคำที่เกิดขึ้นเป็น 0 หรือก็คือคำที่อยู่ใน ถุงของคำ ไม่ปรากฏอยู่ในเอกสารนั้นทำให้ค่าความน่าจะเป็นที่ได้มีค่าเป็น 0 ตามไปด้วย ซึ่งไม่เป็นที่ยอมรับในทางสถิติที่โอกาสในการพยากรณ์จะเป็นศูนย์ เพื่อหลีกเลี่ยงกรณีดังกล่าวสามารถใช้การปรับสมการด้วย Laplace Smoothing

สาเหตุการปรับสมการด้วย Laplace Smoothing เพราะถ้าสังเกตโมเดลการจำแนกเอกสารด้วย นาอีฟเบย์จะพบว่า อาจจะมีค่าความน่าจะเป็นของบาง “คำ” มีค่าเป็น 0 นั่นคือ ไม่มีรูปแบบของ “คำ” นี้เกิดขึ้นในชุดข้อมูลการเรียนรู้ (training data) ดังนั้นการใช้งานโมเดลที่มีค่าความน่าจะเป็นมีค่าเท่ากับ 0 จากเหตุนี้เองจะทำให้ค่าที่จะทำนายมีค่าเป็น 0 ไปด้วยดังนั้น จึงมีการเพิ่มค่าความถี่ของข้อมูลเข้าไปอีกครั้งละ 1 และบวกเพิ่มค่าความถี่รวมด้วยค่าคงที่ k จากคำทั้งหมด n คำ และ กลุ่มทั้งหมด m กลุ่ม ดังนั้น สมการของ Naïve Bayes จึงสามารถปรับสมการ (3.8)

$$P(a_i|v_j) = \frac{1 + count(a_i, v_j)}{k + count(v_i)} \quad (3.8)$$

เมื่อ count(ai,vj) คือ ค่าความถี่ของคำที่ i ในกลุ่มที่ j

count(vi) คือ ค่าความถี่รวมในกลุ่มที่ j

k คือ ค่าคงที่ที่นำมาบวกเข้า

i มีค่าเท่ากับ 1, 2, 3, ..., n

j มีค่าเท่ากับ 1, 2, 3, ..., m

3.2.3.3 การสร้างโมเดลจำแนกความรู้สึกของข้อความสั้นด้วย K-nearest neighbor

K-NN เป็นวิธีการที่ไม่ซับซ้อนและเข้าใจง่ายในการจำแนกประเภทข้อมูล โดยใช้หลักการเปรียบเทียบข้อมูลที่สนใจ (x) กับข้อมูลในคลังข้อมูลที่จัดกลุ่มเตรียมเอาไว้ เพื่อตรวจสอบข้อมูล x นั้นคล้ายคลึงกับข้อมูลกลุ่มใดที่อยู่ในคลังข้อมูล และหากข้อมูล x อยู่ใกล้ข้อมูลกลุ่มใดมากที่สุด ระบบก็จะจัดให้ข้อมูล x เป็นข้อมูลในกลุ่มที่อยู่ใกล้ที่สุดนั้น ซึ่งการตัดสินใจว่าข้อมูล x จะคล้ายกับข้อมูลกลุ่มใดในคลังข้อมูลนั้น จะขึ้นกับการกำหนดค่า k ซึ่ง ค่า k หมายถึงการเอาค่าที่ใกล้เคียงที่สุดจำนวน k ตัวมาพิจารณา ยกตัวอย่างเช่น สมมติว่ามีข้อมูลอยู่ 2 กลุ่ม (คือข้อมูลกลุ่ม A และ B) และกำหนด k = 3ภายหลังการประมวลผล หากมีข้อมูล 5 อันดับแรกที่อยู่ใกล้ข้อมูล x นั้นมาจากกลุ่ม A จำนวน 2 ตัว และมาจากกลุ่ม B จำนวน 1 ตัว ระบบฯ ก็จะพิจารณาให้ข้อมูล x อยู่ในกลุ่ม A

โดยขั้นตอนของ K-NN มีดังนี้

ขั้นที่ 1: กำหนดค่า k

เป็นการกำหนดค่า k เพื่อใช้เป็นเป้าหมายในการที่จะเลือกค่าที่ใกล้เคียงกับข้อมูลที่สนใจ โดยค่า k ที่กำหนดนั้นจำเป็นต้องเป็นเลขคี่ เพื่อให้โปรแกรมสามารถตัดสินใจในการวิเคราะห์ผลลัพธ์ออกมา

ขั้นที่ 2: คำนวณหาระยะห่างระหว่างข้อมูลที่สนใจ x และข้อมูลทุกตัวในคลังข้อมูล

ในที่นี้จะใช้วิธีการคำนวณหาระยะห่างระหว่างข้อมูลที่สนใจ x และข้อมูลทุกตัวในคลังข้อมูลด้วยวิธีการ Euclidian distance ครอบคลุมส่วนประกอบต่างๆ สามารถคำนวณได้จากสมการ (3.9)

$$E(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \tag{3.9}$$

โดยที่

E คือ ระยะห่างระหว่างข้อมูลที่สนใจ x กับข้อมูลที่คัดเลือกไว้ในคลังข้อมูล y

xi คือ คุณลักษณะที่ i ของข้อมูลที่สนใจ x

yi คือ คุณลักษณะที่ i ของข้อมูลที่คัดเลือกไว้ในคลังข้อมูล y

ซึ่งข้อมูล x จะถูกเปรียบเทียบกับข้อมูลในคลังข้อมูล y ทั้งหมด

ขั้นที่ 3: เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัว

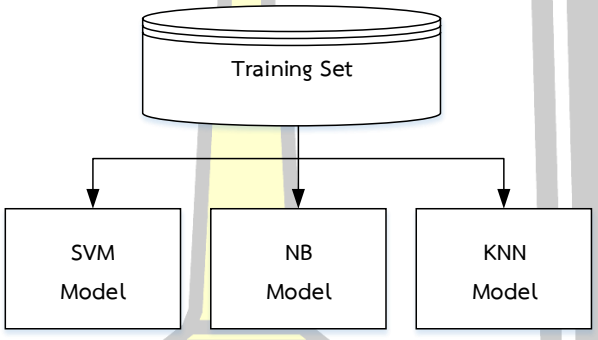
เลือกค่าข้อมูลที่มีค่าระยะห่างน้อยที่สุด k ตัวเพื่อนำมาพิจารณาหาคำตอบ สมมติว่าเราเลือกใช้ k = 3 ถ้าสมมติว่าเมื่อมีการเปรียบเทียบข้อมูล x กับข้อมูลในคลังข้อมูลทั้งหมดแล้วพบว่าข้อมูล x ที่มีค่าใกล้เคียงกับข้อมูลในกลุ่ม positive 2 ตัว และมีค่าใกล้เคียงกับข้อมูลในกลุ่ม negative 1 ตัว จากผลลัพธ์ที่ได้ ระบบฯ จะตัดสินให้ข้อมูล x อยู่ในกลุ่มข้อมูลที่เป็น positive

อย่างไรก็ตาม มีข้อสังเกตว่าถ้าเลือกค่า k น้อยเกินไปอาจทำให้เป็นความไวต่อสัญญาณรบกวนได้ และถ้าเลือกค่า k มากเกินไปอาจจะทำให้มีกลุ่มข้อมูลอื่นมาประปนกับข้อมูลที่กำลังสนใจได้เช่นกัน ดังนั้นวิธีนี้มีข้อดีเป็นวิธีที่ง่ายและมีมีประสิทธิภาพ แต่ข้อเสียคือแต่การประมวลผลค่อนข้างช้า เพราะทำนายข้อมูลใหม่โดยอาศัยการเปรียบเทียบกับข้อมูลเรียนรู้จำนวน k ตัวที่อยู่ใกล้ที่สุด

3.2.4 การสร้าง Ensemble Model ด้วยวิธีการ Voting

เทคนิค Voting Ensemble เป็นเทคนิคที่ใช้โมเดลหลายๆ โมเดล มาช่วยในการหาคำตอบ ซึ่งในงานวิจัยนี้ใช้ Voting Ensemble มาสร้างโมเดล ตัวอย่างการทำงานของ Voting Ensemble มีดังต่อไปนี้

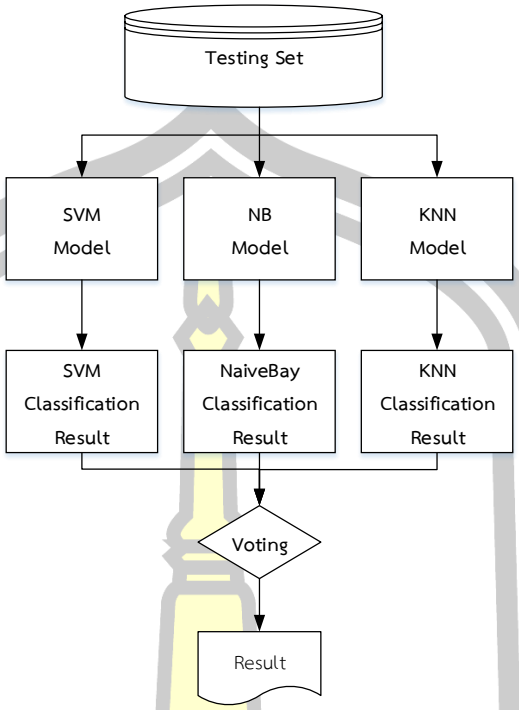
1) การสร้างโมเดล ในขั้นตอนนี้จะใช้ชุดข้อมูล Training Set ชุดเดียวกันสร้างชุดโมเดลจำแนกความรู้สึกที่ต่างกัน ซึ่งในงานวิจัยนี้ได้สร้างโมเดลจำแนกความรู้สึก 3 โมเดล คือ Support Vector Machine, NaïveBays และ k-nearest neighbor แสดงตัวอย่างดังรูปที่ 3-5



รูปที่ 3-5 ขั้นตอนการจำแนกเอกสารในแต่ละอัลกอริทึม

จากรูปที่ 3-5 เป็นการแสดงขั้นตอนการสร้างโมเดลจำแนกเอกสารของ 3 อัลกอริทึม คือ Support Vector Machine, NaïveBays และ k-nearest neighbor

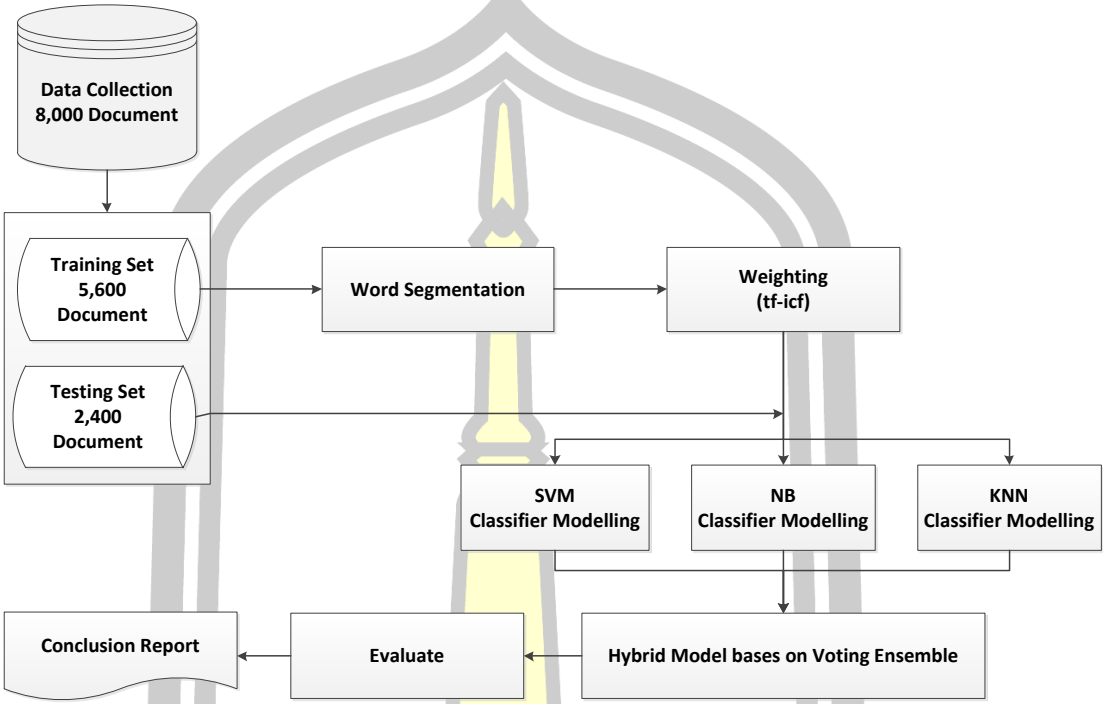
2) นำโมเดลไปใช้งาน ซึ่งหลังจากที่สร้างโมเดล Ensemble ด้วย 3 เทคนิคข้างต้นได้แล้ว ขั้นตอนถัดไป คือ การนำโมเดลที่สร้างได้ไปทำนายข้อมูลใหม่ โดยใช้วิธีโหวต แสดงตัวอย่างดังรูปที่ 3-6



รูปที่ 3-6 การนำ Voting Ensemble Model มาใช้งาน

จากรูปที่ 3-6 จะมีชุดข้อมูลทดสอบเข้ามา (Testing Set) ซึ่งยังไม่รู้คลาส ซึ่งทั้ง 3 โมเดล จะทำการจำแนกเอกสาร ว่าอยู่คลาสใด ซึ่งคลาสที่ได้ผลโหวตสูงสุด จะเป็นคำตอบของการทดสอบ

3.3 กระบวนการดำเนินงานวิจัยที่ปรับปรุง (Improved Research Methodology)



รูปที่ 3-7 กระบวนการดำเนินงานวิจัยที่ปรับปรุง

ในกระบวนการเดิมที่นำเสนอ พบว่าการสร้างคำในคลังขั้วความรู้สึกไม่ครอบคลุมและไม่สอดคล้องกับคำที่อยู่ในเอกสารที่เก็บรวบรวมมา เนื่องจากคำใน SentiWordnet ที่นำมาสร้างคลังขั้วความรู้สึกนั้น สร้างมาจากหลากหลายโดเมน ดังนั้นในงานวิจัยนี้จึงมีการปรับปรุงกระบวนการวิจัยใหม่ โดยกระบวนการวิจัยที่ถูกปรับปรุง ได้ปรับปรุงในส่วนของการให้น้ำหนักคำ โดยนำเสนอการให้น้ำหนักคำแบบ *tf-icf* ซึ่งเป็นการให้น้ำหนักคำที่อยู่บนแนวคิดที่เรียกว่า “การให้น้ำหนักคำที่ให้ความสำคัญในแต่ละคลาส” โดยเป็นการให้น้ำหนักที่ทำการปรับมาจากวิธีการให้น้ำหนักคำแบบ tf-idf เพราะเมื่อพิจารณาแล้วจะพบว่า tf-idf จะเป็นการให้น้ำหนักคำที่สะท้อนความสำคัญของคำที่อยู่ในเอกสารหนึ่งๆ ที่อยู่ในคลังเอกสาร สามารถแสดงได้ตามสมการ (3.10)

$$w(t_k) = \log(1 + tf_k) \times \log\left(\frac{N+1}{df(t_k)+1}\right) \quad (3.10)$$

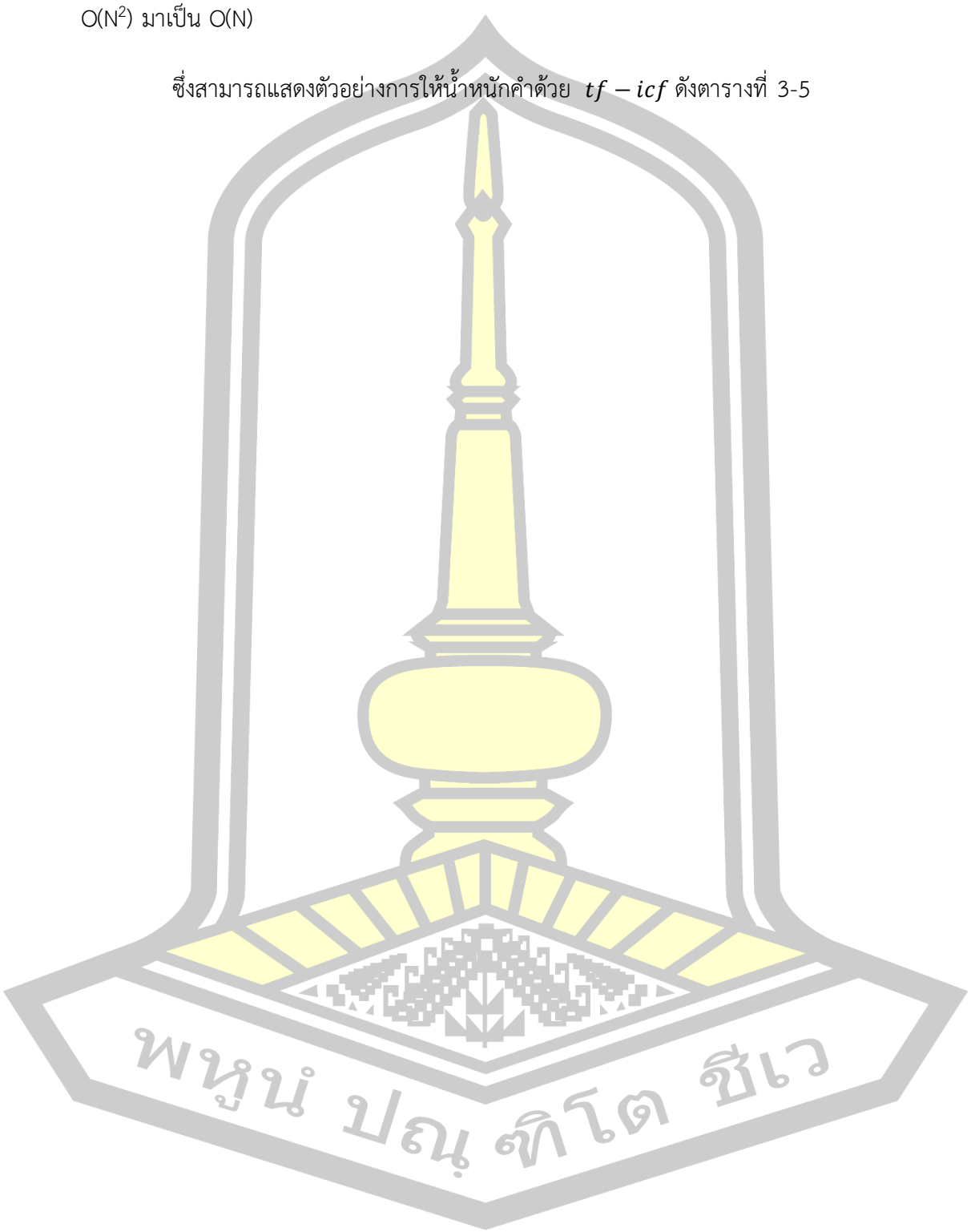
เมื่อ $w(t_k)$ คือ ความถี่ของคำ t_k ที่พบในเอกสาร

N คือ จำนวนเอกสารทั้งหมดในคลาสนั้น ๆ

$df(t_k)$ คือ จำนวนเอกสารในคลาสที่พบคำ t_k

ซึ่ง tf-icf สามารถลดความซับซ้อนของการประมวลผลที่มีการให้น้ำหนักคำด้วย *tf-idf* จาก $O(N^2)$ มาเป็น O(N)

ซึ่งสามารถแสดงตัวอย่างการให้น้ำหนักคำด้วย $tf - icf$ ดังตารางที่ 3-5



ตารางที่ 3-5 แสดงการแทนค่าน้ำหนักคำด้วย tf-icf

alway	clean	equip	good	intern	pool	safe	servic	smile	staff	terribl
0	1.08436	0	0	0	0	1.08436	0	0	0	0
0	0	0	0	0	0	0	1.08436	0	1.08436	1.08436
0	1.08436	0	0	0	1.08436	0	0	1.08436	1.08436	0
0	0	1.08436	0	1.08436	0	0	1.08436	0	0	1.08436
1.08436	0	0	1.08436	0	0	0	1.08436	1.08436	1.08436	0

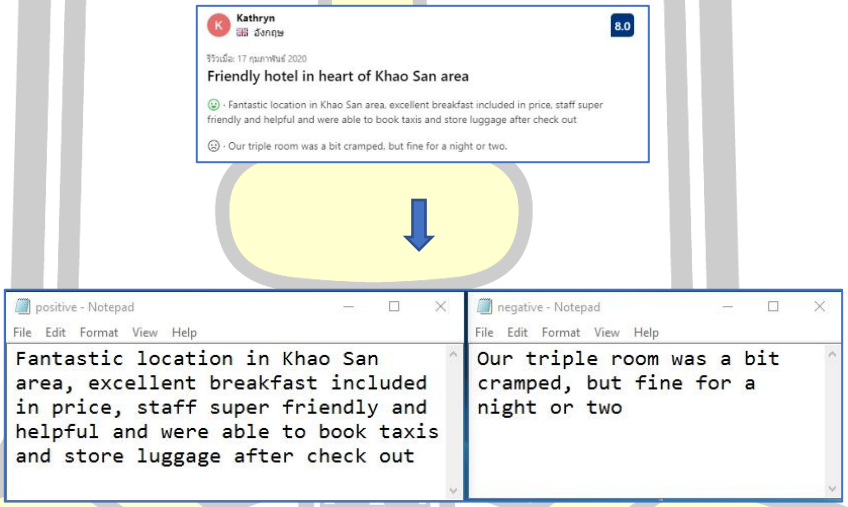
บทที่ 4

ผลการทดลอง

ในบทนี้ ผู้วิจัยจะกล่าวถึงการวัดประสิทธิภาพของการทดลองการจำแนกความรู้สึกข้อความสั้น โดยการให้น้ำหนักคำแบบต่าง ๆ และ การวัดประสิทธิภาพของการทดลองในการสร้างโมเดลแบบ 1 อัลกอริทึม กับ การสร้างโมเดลด้วยหลายอัลกอริทึม ในการจำแนกความรู้สึกข้อความสั้น

4.1 ชุดข้อมูลที่ใช้ในการทดสอบ

สำหรับในส่วนการทดสอบจะใช้ข้อมูลบทวิจารณ์จากเว็บไซต์ www.booking.com จำนวน 8,000 เอกสาร โดยเป็นบทวิจารณ์เชิงบวกจำนวน 4,000 เอกสาร และ บทวิจารณ์เชิงลบจำนวน 4,000 เอกสาร ซึ่งข้อมูลบทวิจารณ์เหล่านี้เป็นการเก็บข้อมูลในระหว่างวันที่ 1 กุมภาพันธ์ 2563 ถึงวันที่ 29 กุมภาพันธ์ 2563 โดยข้อมูลเหล่านี้ถูกจัดเก็บในรูปแบบ Text file ดังตัวอย่างในรูปที่ 4-1



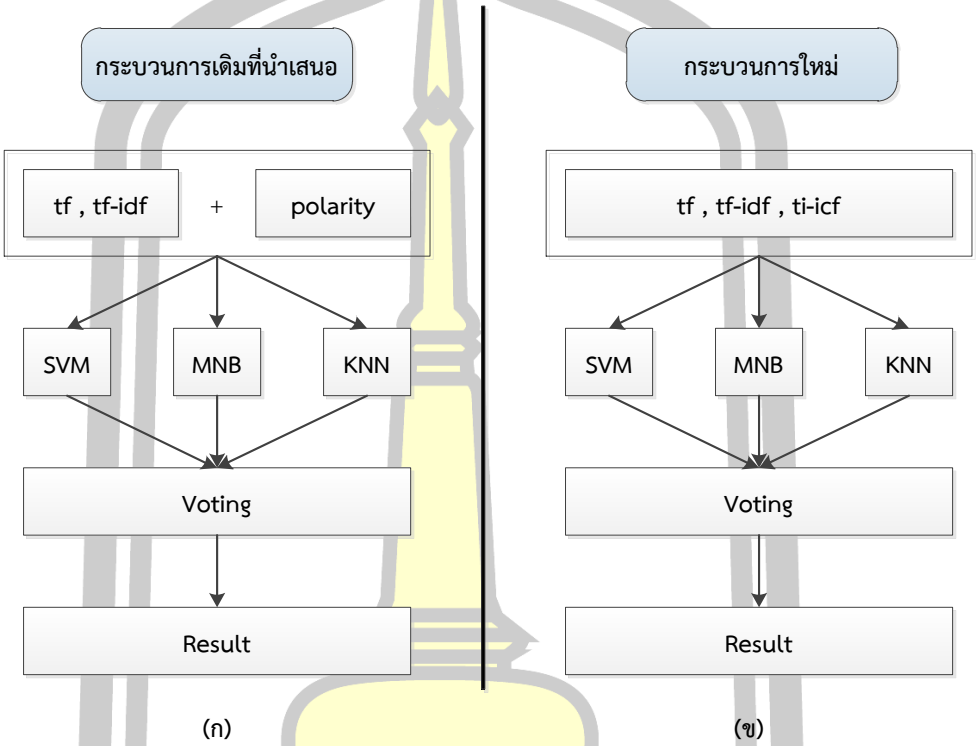
รูปที่ 4-1 การเก็บข้อมูลที่ใช้ในการทดสอบ

จากข้อมูลที่เก็บรวบรวมมา สามารถแบ่งเป็นข้อมูลชุดเรียนรู้ (Training Set) ด้วยวิธีการสุ่ม 70% และ ข้อมูลชุดทดสอบ 30% และเมื่อผ่านกระบวนการเตรียมข้อมูลก่อนการประมวลผล มีจำนวนคุณลักษณะของคำจำนวน 6,763 คำ

4.2 ผลการทดลอง

เดิมกระบวนในการจำแนกบทวิจารณ์เป็นออกเป็นกลุ่มที่เป็นเชิงบวกและเชิงลบ จะใช้ขั้วความรู้สึก (Sentiment polarity) เข้ามาช่วยในการให้น้ำหนักร่วมกับ tf และ tf-idf ด้วย แต่

ภายหลังมีการปรับมาใช้การให้น้ำหนักแบบ tf-icf ที่เป็นการให้น้ำหนักที่เน้นการให้ความสำคัญแต่ละ “คำ” ที่ปรากฏในคลาสที่แตกต่างกันอยู่แล้ว ดังนั้นจึงทำให้ไม่ต้องมีการใช้ค่า ขั้วความรู้สึกเข้ามาอีก ดังที่แสดงการเปรียบเทียบกระบวนการในรูปที่ 4-2



รูปที่ 4-2 แสดงการเปรียบเทียบกระบวนการในการสร้างโมเดลแบบเดิมและแบบใหม่

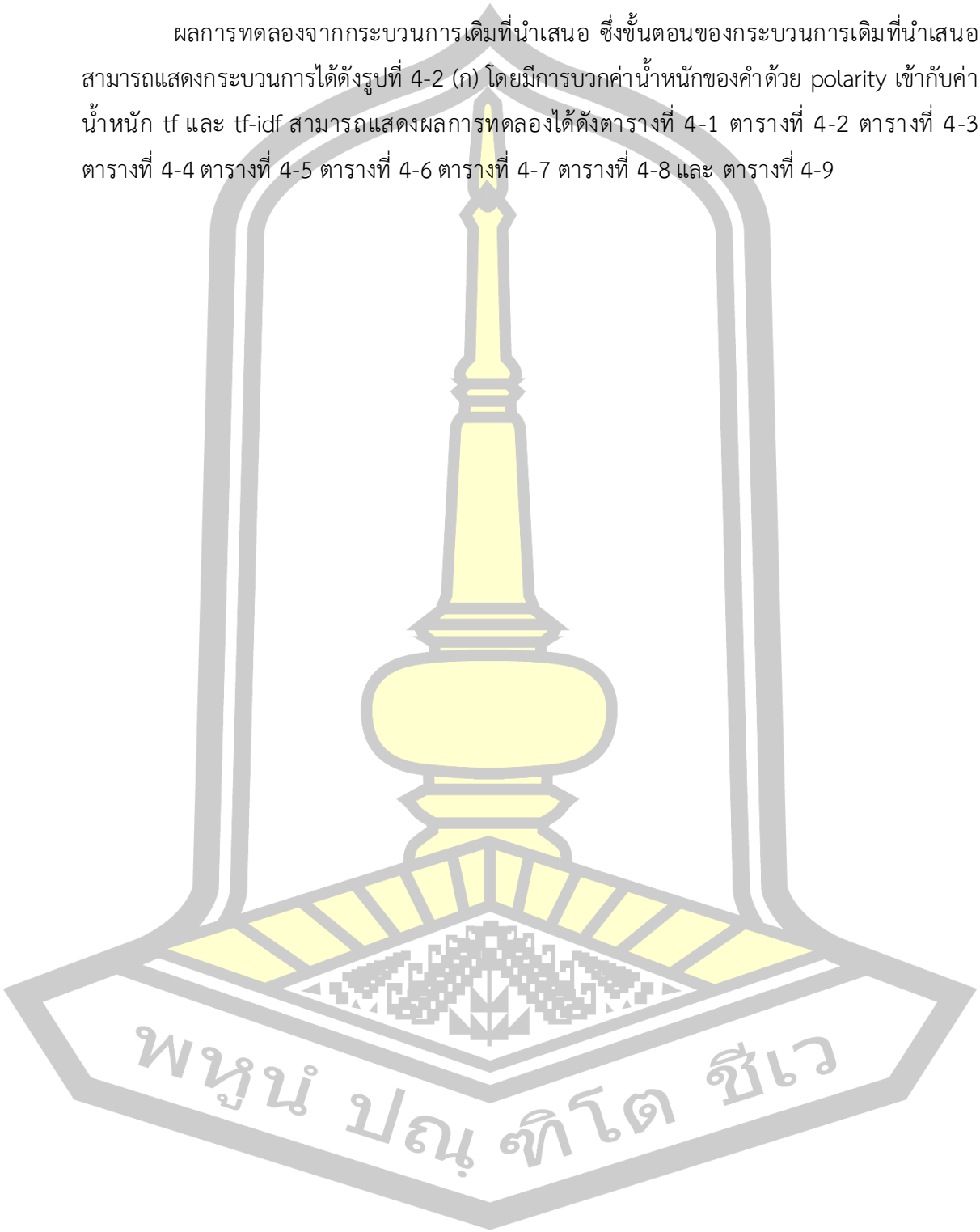
จาก รูปที่ 4-2 เป็นการเปรียบเทียบกระบวนการวิจัยเดิมที่นำเสนอ กับ กระบวนการวิจัยที่นำเสนอใหม่ ซึ่งมีข้อแตกต่างในกระบวนการให้น้ำหนักของคำ ซึ่งกระบวนการวิจัยที่นำเสนอเดิม จะมีการบวกค่า polarity ของคำแต่ละคำหลังจากหาค่า tf และ tf-idf เสร็จสิ้นแล้ว แต่กระบวนการที่นำเสนอใหม่มีการปรับมาใช้การให้น้ำหนักแบบ tf-icf ที่เป็นการให้น้ำหนักที่เน้นการให้ความสำคัญแต่ละ “คำ” ที่ปรากฏในคลาสที่แตกต่างกันอยู่แล้ว ดังนั้นจึงทำให้ไม่ต้องมีการใช้ค่าขั้วความรู้สึกเข้ามาอีก

โดยอัลกอริทึมการเรียนของเครื่องแบบมีผู้สอนที่ใช้ในการสร้างตัวจำแนกบทวิจารณ์มี 3 อัลกอริทึมคือ SVM, MNB, และ KNN จากนั้นจึงเอาผลลัพธ์ของการจำแนกจากทั้ง 3 อัลกอรอทึม มาวิเคราะห์ด้วยการโหวต ภายใต้กระบวนการแบบ Voting ensemble

ซึ่งผลการทดลองทั้งสองกระบวนการ สามารถแสดงรายละเอียดดังต่อไปนี้

4.2.1 การทดลองการจำแนกบทวิจารณ์ของแต่ละโมเดลตามกระบวนการเดิมที่นำเสนอ

ผลการทดลองจากกระบวนการเดิมที่นำเสนอ ซึ่งขั้นตอนของกระบวนการเดิมที่นำเสนอสามารถแสดงกระบวนการได้ดังรูปที่ 4-2 (ก) โดยมีการบวกค่าน้ำหนักของคำด้วย polarity เข้ากับค่าน้ำหนัก tf และ tf-idf สามารถแสดงผลการทดลองได้ดังตารางที่ 4-1 ตารางที่ 4-2 ตารางที่ 4-3 ตารางที่ 4-4 ตารางที่ 4-5 ตารางที่ 4-6 ตารางที่ 4-7 ตารางที่ 4-8 และ ตารางที่ 4-9



ตารางที่ 4-1 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล SVM แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf

SVM	tf + polarity										
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	2561.19	0.92	0.90	0.89	0.87	0.92	0.93	0.93	0.93	0.92	0.91	**0.91**
Recall		0.92	0.90	0.89	0.87	0.92	0.93	0.93	0.93	0.92	0.91	**0.91**
Precision		0.92	0.90	0.89	0.89	0.92	0.93	0.93	0.93	0.92	0.91	**0.91**
F1		0.92	0.90	0.89	0.87	0.92	0.93	0.93	0.92	0.92	0.91	**0.91**

ตารางที่ 4-2 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล NB แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf

NB	tf + polarity										
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	135.38	0.93	0.89	0.90	0.86	0.92	0.92	0.92	0.91	0.90	0.91	**0.91**
Recall		0.93	0.89	0.90	0.86	0.92	0.92	0.92	0.91	0.90	0.91	**0.91**
Precision		0.93	0.89	0.90	0.86	0.92	0.93	0.92	0.91	0.90	0.91	**0.91**
F1		0.92	0.89	0.90	0.86	0.91	0.92	0.92	0.91	0.90	0.91	**0.91**

ตารางที่ 4-3 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล KNN แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf

KNN	tf + polarity											
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	1413.94	0.75	0.74	0.76	0.75	0.78	0.81	0.79	0.81	0.79	0.75	**0.77**
Recall		0.75	0.74	0.76	0.75	0.78	0.81	0.79	0.81	0.79	0.75	**0.77**
Precision		0.76	0.75	0.76	0.75	0.79	0.82	0.80	0.81	0.79	0.76	**0.78**
F1		0.75	0.74	0.76	0.75	0.77	0.81	0.79	0.80	0.79	0.75	**0.77**

ตารางที่ 4-4 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล Ensemble แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf

Ensemble	tf + polarity											
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	89.18	0.91	0.97	0.88	0.78	0.89	0.90	0.91	0.94	0.92	0.82	**0.89**
Recall		0.91	0.97	0.88	0.78	0.89	0.90	0.91	0.94	0.92	0.82	**0.89**
Precision		0.91	0.97	0.88	0.78	0.89	0.91	0.91	0.94	0.92	0.82	**0.89**
F1		0.91	0.97	0.88	0.78	0.89	0.90	0.91	0.94	0.92	0.82	**0.89**

ตารางที่ 4-5 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล SVM แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-idf

SVM	tf-idf + polarity										
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	2910.06	0.91	0.89	0.88	0.87	0.91	0.92	0.92	0.90	0.90	0.89	**0.90**
Recall		0.91	0.89	0.88	0.87	0.91	0.92	0.92	0.90	0.90	0.89	**0.91**
Precision		0.91	0.89	0.89	0.88	0.91	0.93	0.93	0.91	0.91	0.90	**0.90**
F1		0.91	0.89	0.88	0.86	0.91	0.92	0.92	0.90	0.90	0.89	**0.90**

ตารางที่ 4-6 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล NB แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-idf

NB	tf-idf + polarity										
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	135.04	0.92	0.89	0.90	0.84	0.91	0.91	0.90	0.89	0.89	0.91	**0.90**
Recall		0.92	0.89	0.90	0.84	0.91	0.91	0.90	0.89	0.89	0.91	**0.90**
Precision		0.92	0.89	0.90	0.84	0.91	0.92	0.90	0.89	0.89	0.91	**0.90**
F1		0.92	0.89	0.90	0.84	0.91	0.91	0.90	0.89	0.89	0.91	**0.90**

ตารางที่ 4-7 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล KNN แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-idf

KNN	tf-idf + polarity										
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	1562.12	0.78	0.74	0.75	0.74	0.76	0.78	0.79	0.78	0.78	0.75	0.77
Recall		0.78	0.74	0.75	0.74	0.76	0.78	0.79	0.78	0.78	0.75	0.77
Precision		0.79	0.74	0.75	0.74	0.77	0.79	0.79	0.79	0.79	0.75	0.77
F1		0.78	0.74	0.74	0.74	0.76	0.78	0.78	0.78	0.78	0.75	0.76

ตารางที่ 4-8 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล Ensemble แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-idf

Ensemble	tf-idf + polarity										
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	3676.09	0.93	0.90	0.89	0.86	0.93	0.93	0.92	0.92	0.91	0.91	**0.91**
Recall		0.93	0.90	0.89	0.86	0.93	0.93	0.92	0.92	0.91	0.91	**0.91**
Precision		0.93	0.90	0.89	0.86	0.93	0.93	0.92	0.92	0.91	0.91	**0.91**
F1		0.92	0.90	0.89	0.85	0.93	0.93	0.92	0.92	0.91	0.91	**0.91**

ตารางที่ 4-9 สรุปผลการทดลองการจำแนกบทวิจารณ์ของแต่ละโมเดลตามกระบวนการเดิมที่นำเสนอ

Algorithm	tf + polarity								tf-idf + polarity							
	Acc		R		P		F1		Acc		R		P		F1	
	Acc	Time	Neg	Pos	Neg	Pos	Neg	Pos	Acc	Time	Neg	Pos	Neg	Pos	Neg	Pos
SVM	**0.91**	222.98	**0.95**	0.86	0.87	**0.94**	**0.91**	0.90	0.89	276.49	0.95	0.83	0.85	0.94	0.89	0.88
NB	**0.91**	**37.95**	0.89	**0.92**	**0.91**	0.90	0.90	**0.91**	0.89	**37.93**	0.88	**0.91**	**0.90**	0.89	0.89	0.90
KNN	0.76	202.50	0.66	0.87	0.84	0.71	0.74	0.78	0.75	202.91	0.67	0.83	0.81	0.70	0.73	0.76
Ensemble	**0.91**	387.77	0.92	0.89	0.90	0.92	**0.91**	0.90	**0.91**	445.20	**0.92**	0.90	**0.90**	**0.93**	**0.91**	**0.91**

จากผลการทดลองในตารางที่ 4-1 ตารางที่ 4-2 ตารางที่ 4-3 ตารางที่ 4-4 ตารางที่ 4-5 ตารางที่ 4-6 ตารางที่ 4-7 ตารางที่ 4-8 และ ตารางที่ 4-9 จะเห็นว่าการให้น้ำหนักคำด้วยวิธี tf และ tf-idf ให้ผลการทดลองในเกณฑ์ที่ดี ซึ่งโมเดลที่มีประสิทธิภาพดีที่สุดคือการสร้างโมเดลด้วยเทคนิค Support Vector Machine เนื่องจากว่าเป็นเทคนิคการสร้างโมเดลสำหรับชุดข้อมูลที่ไม่มีโครงสร้าง หรือกึ่งโครงสร้าง จึงทำให้ชุดข้อมูลข้อความสั้นจึงมีความเหมาะสมกับการทำงานของ Support Vector Machine มากที่สุด [23, 24] และ เทคนิคที่ทำเวลาได้ดีที่สุดคือเทคนิค NaiveBayes โดยใช้เวลาในกระบวนการเรียนรู้ 37 วินาที

4.2.2 การทดลองการจำแนกบทวิจารณ์ของแต่ละโมเดลตามกระบวนการใหม่ที่นำเสนอ

ผลการทดลองจากกระบวนการใหม่ที่นำซึ่งขั้นตอนของกระบวนการที่นำเสนอใหม่ สามารถแสดงกระบวนการได้ดังรูปที่ 4-1 (ข) ซึ่งจะไม่มีการบวกค่า polarity เข้าไป แต่จะมีการให้ค่าน้ำหนักคำด้วย tf-icf แทน สามารถแสดงผลการทดลองได้ดังตารางที่ 4-10 ตารางที่ 4-11 ตารางที่ 4-12 ตารางที่ 4-13 และ ตารางที่ 4-14

ตารางที่ 4-10 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล SVM แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-icf

SVM	tf-icf + polarity											
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	1794.63	0.96	0.97	0.96	0.94	0.98	0.97	0.98	0.98	0.97	0.96	**0.97**
Recall		0.96	0.97	0.96	0.94	0.98	0.97	0.98	0.98	0.97	0.96	**0.97**
Precision		0.96	0.97	0.96	0.95	0.98	0.97	0.98	0.98	0.97	0.97	**0.97**
F1		0.96	0.97	0.96	0.94	0.98	0.97	0.98	0.98	0.97	0.96	**0.97**

ตารางที่ 4-11 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล NB แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-icf

NB	tf-icf + polarity											
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	128.70	0.95	0.94	0.94	0.89	0.96	0.95	0.96	0.94	0.94	0.95	0.94
Recall		0.95	0.94	0.94	0.89	0.96	0.95	0.96	0.94	0.94	0.95	0.94
Precision		0.95	0.94	0.94	0.89	0.96	0.95	0.96	0.94	0.94	0.95	0.94
F1		0.95	0.93	0.94	0.89	0.95	0.95	0.96	0.94	0.94	0.95	0.94

ตารางที่ 4-12 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล KNN แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-icf

KNN	tf-icf + polarity											
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	900.53	0.96	0.94	0.92	0.91	0.95	0.95	0.95	0.96	0.95	0.93	0.94
Recall		0.96	0.94	0.92	0.91	0.95	0.95	0.95	0.96	0.95	0.93	0.94
Precision		0.96	0.94	0.92	0.91	0.95	0.95	0.95	0.96	0.95	0.93	0.94
F1		0.96	0.94	0.91	0.91	0.95	0.94	0.95	0.96	0.95	0.93	0.94

ตารางที่ 4-13 ผลการทดลองการจำแนกบทวิจารณ์ด้วยโมเดล Ensemble แบบ 10-fold cross validation โดยการให้น้ำหนักแบบ tf-icf

Ensemble	tf-icf + polarity											
	Time	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	2572.88	0.97	0.97	0.96	0.95	0.98	0.98	0.99	0.99	0.98	0.98	**0.97**
Recall		0.97	0.97	0.96	0.95	0.98	0.98	0.99	0.99	0.98	0.98	**0.97**
Precision		0.97	0.97	0.96	0.95	0.98	0.98	0.99	0.99	0.98	0.98	**0.97**
F1		0.97	0.97	0.96	0.95	0.98	0.98	0.99	0.98	0.97	0.98	**0.97**

ตารางที่ 4-14 ผลการทดลองการจำแนกบทวิจารณ์ของแต่ละโมเดลตามกระบวนการใหม่ที่นำเสนอ

Algorithm	tf-icf							
	Acc		R		P		F1	
	Acc	Time	Neg	Pos	Neg	Pos	Neg	Pos
SVM	0.97	180.35	0.99	0.94	0.94	0.99	0.97	0.97
NB	0.94	35.84	0.91	**0.97**	0.97	0.91	0.94	0.94
KNN	0.94	204.33	0.96	0.92	0.93	0.96	0.94	0.94
Ensemble	0.97	355.99	0.99	0.96	0.95	0.99	0.97	0.97

จากผลการทดลอง ตารางที่ 4-10 ตารางที่ 4-11 ตารางที่ 4-12 ตารางที่ 4-13 และ ตารางที่ 4-14 จะเห็นได้ว่าการให้น้ำหนักคำด้วย tf-icf จะให้ผลการทดลองได้ดีกว่า tf และ tf-idf ในการทดลองโดยใช้โมเดลเดียวกัน เนื่องจากว่า tf-icf เป็นการเป็นการให้น้ำหนักคำที่อยู่บนแนวคิดที่เรียกว่า “การให้น้ำหนักคำที่ให้ความสำคัญในแต่ละคลาส” ซึ่งส่งผลให้มีประสิทธิภาพโดยรวมที่ดีกว่า

จากตารางที่ 4-10 ตารางที่ 4-11 ตารางที่ 4-12 ตารางที่ 4-13 และ ตารางที่ 4-14 ยังแสดงให้เห็นว่า บางเทคนิควิธีก็ให้ประสิทธิภาพในผลการทดลองที่แตกต่างกัน เช่น ค่า accuracy เทคนิค SVM และ Ensemble ให้ประสิทธิภาพการทำงานที่ดีที่สุด ค่า Recall ผลปรากฏว่า SVM และ NB ให้ประสิทธิภาพที่ดีในคลาสที่แตกต่างกัน ค่า Precision SVM, NB และ Ensemble และค่า F1 เทคนิคที่ให้ประสิทธิภาพที่ดีที่สุดคือ SVM และ Ensemble ดังนั้น จึงสรุปได้ว่า การใช้โมเดลที่หลากหลายร่วมกันจำแนกเอกสารด้วยวิธีการโหวต จะสามารถสรุปผลของการจำแนกเอกสารโดยโมเดลใดโมเดลหนึ่ง

จากผลการทดลองจำแนกบทวิจารณ์แต่ละโมเดลโดยใช้การให้น้ำหนักคำด้วย tf-icf การสร้างโมเดลด้วย Voting Ensemble จะให้ผลการทดลองได้ดีที่สุด โดยผลการวัดประสิทธิภาพของโมเดลที่ได้ คือ Accuracy = 0.97, Recall = 0.98, Precision = 0.97 และ F1 = 0.97

4.3 วิจารณ์ผลการทดลอง

การจำแนกบทวิจารณ์ข้อความสั้น เนื่องจากจำนวนคำที่แสดงในข้อความมีจำนวนคำที่น้อย ทำให้ไม่สามารถคัดเลือกคุณลักษณะ (features) ที่เหมาะสมและมีความหมาย [23, 24] หรืออาจจะสกัดได้น้อยเกินไปจนยากต่อการสร้างตัวจำแนกความรู้สึกจากข้อความที่มีคุณภาพต่อการใช้งานที่ดีได้ ดังนั้นในงานวิจัยนี้ จึงนำเสนอการเพิ่มประสิทธิภาพของการจำแนกบทวิจารณ์ข้อความสั้นให้มีประสิทธิภาพดีขึ้นโดยให้ความสำคัญกับการให้น้ำหนักคำ (term weighting) และ สร้างโมเดลด้วยการสร้างโมเดลแบบผสมผสาน โดยใช้ 3 อัลกอริทึมคือ SVM, MNB, และ KNN จากนั้นจึงเอาผลลัพธ์ของการจำแนกจากทั้ง 3 อัลกอรอทึม มาวิเคราะห์ด้วยการโหวต ภายใต้กระบวนการแบบ Voting Ensemble Model

หากมองในมุมของการให้น้ำหนักในกระบวนการที่นำเสนอเดิม จากที่มีการให้น้ำหนักคำบวกด้วยค่า polarity ซึ่งผลการทดลองสามารถแสดงได้ในตารางที่ 4-1 และ การให้น้ำหนักของคำในกระบวนการใหม่ที่นำเสนอ จะไม่มีการบวกค่า polarity เข้าไป แต่จะใช้การให้ค่าน้ำหนักคำด้วย tf-icf แทนการให้ค่าน้ำหนักของคำด้วย tf-icf ซึ่งผลการทดลองในตารางที่ 4-2 จะเห็นได้ว่า การให้น้ำหนักคำด้วย tf-icf ให้ผลการทดลองที่มีประสิทธิภาพที่ดีกว่ากระบวนการเดิมที่นำเสนอ ทั้งนี้

เนื่องมาจาก tf-icf จะเน้นการให้ความสำคัญกับคำที่เจอในแต่ละคลาส ซึ่งต่างจาก tf และ tf-idf จะไม่พิจารณาเป็นคลาส แต่จะให้น้ำหนักของคำรวมทุกคลาส แต่ทั้งนี้ในแง่ของเวลาในการทำงาน tf และ tf-idf จะใช้เวลาในการทำงานน้อยกว่าเนื่องจากกระบวนการคำนวนไม่ซับซ้อน

นอกจากนี้ ในกระบวนการสร้างโมเดล ในงานวิจัยนี้ได้เลือกใช้วิธีการสร้างโมเดลด้วย Ensemble Model ด้วยวิธีการโหวต โดยใช้เทคนิคในการสร้างโมเดลด้วยกัน 3 เทคนิค คือ SVM, MNB และ KNN ซึ่งผลการทดลองในตารางที่ 4-2 จะเห็นว่า เมื่อเทียบระหว่างอัลกอริทึมทั้งสามตัว SVM จะให้ประสิทธิภาพการทำงานได้ดีที่สุด เนื่องจาก SVM เป็นโมเดลที่สามารถจำแนกประเภทเอกสารได้ดี และ สามารถทำงานได้ดีกับข้อมูลที่ไม่มีโครงสร้าง หรือ โครงสร้างไม่ชัดเจน ส่วนโมเดลที่ให้ประสิทธิภาพได้ดีรองลงมาคือ NB โดยค่าวัดประสิทธิภาพได้ใกล้เคียงกับ SVM เนื่องจาก NB ก็เป็นอีกเทคนิควิธีหนึ่งที่เหมาะสำหรับการจำแนกประเภทเอกสาร ซึงจุดเด่นของ NB คือ จะไม่ให้ความสำคัญกับคุณลักษณะของคำที่ไม่มีความเกี่ยวข้อกัน ส่วน KNN ให้ค่าการวัดประสิทธิภาพได้ต่ำที่สุด เนื่องจาก KNN จำเป็นต้องให้ความสำคัญในการปรับจูนคุณลักษณะของคำ (feature) ให้เหมาะสมที่สุด แต่การสร้างโมเดลด้วย KNN ใช้เวลาในการประมวลผลน้อย เนื่องจากไม่ต้องเสียเวลาในการสอนชุดข้อมูล

จากผลการทดลอง จะเห็นได้ว่า แต่ละโมเดลก็จะให้ประสิทธิภาพที่ดีแตกต่างกันออกไป ในแต่ละคลาส ในแต่ละวิธีการวัดประสิทธิภาพ ดังนั้น การเลือกใช้เทคนิคแบบผสมผสานโดยทั้งสามเทคนิคจะช่วยจำแนกเอกสารได้หลากหลายรูปแบบ จึงมีความเหมาะสมที่จะช่วยเพิ่มประสิทธิภาพในการจำแนกเอกสารที่มีโครงสร้างไม่ชัดเจน หรือ กึงโครงสร้างได้เป็นอย่างดี จึงทำให้การสร้างโมเดลด้วย Ensemble Model ภายใต้กระบวนการ Voting สามารถให้ประสิทธิภาพในการจำแนกบทวิจารณ์ได้ดีกว่าการสร้างโมเดลด้วยอัลกอริทึมเดียว เนื่องจากทั้ง 3 อัลกอริทึมจะช่วยกันจำแนกบทวิจารณ์ภายใต้ข้อมูลชุดเรียนรู้ชุดเดียวกัน แล้วนำมาเปรียบเทียบกัน คำตอบที่ได้รับการโหวตสูงสุดจะเป็นคำตอบสุดท้ายของข้อมูลชุดทดสอบชุดนั้น ในส่วนของผลการทดลองการจำแนกบทวิจารณ์ ด้วยวิธีการสร้างโมเดลด้วย Ensemble Model มีค่าการวัดประสิทธิภาพที่ Accuracy 0.97 Recall 0.98 Precision 0.97 F1 0.97

บทที่ 5

สรุปผลการทดลอง

ในบทนี้ ผู้วิจัยจะกล่าวถึงบทสรุปของการวิจัย ปัญหาอุปสรรคที่พบ และ แนวทางการพัฒนางานวิจัยทางด้านการจำแนกความรู้สึกข้อความสั้นในอนาคต ดังนี้

5.1 บทสรุปของการวิจัย

งานวิจัยนี้ เป็นงานวิจัยเกี่ยวกับการจำแนกเอกสารที่มีรูปแบบเป็นข้อความขนาดสั้น ที่มีข้อจำกัดในส่วนของจำนวนคำที่แสดงในข้อความมีจำนวนน้อย ทำให้ไม่สามารถคัดเลือกคุณลักษณะที่เหมาะสมและมีความหมาย หรือ อาจจะสกัดได้น้อยเกินไปจนยากต่อการสร้างตัวจำแนกความรู้สึกจากข้อความที่มีคุณภาพต่อการใช้งานที่ดีได้ โดยใช้เทคนิคการจำแนกเอกสารแบบผสมผสานกันได้แก่ Support Vector Machines, Naïve Bayes และ k-nearest neighbor ในส่วนของการวัดประสิทธิภาพการทดลองวัดประสิทธิภาพจาก ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Recall) = 0.98 ความความระลึก (Precision) = 0.97 และ ค่า F-measure = 0.97

ในขั้นตอนการจัดเก็บรวบรวมข้อมูล ทำการเก็บรวบรวมข้อมูลมาจากเว็บไซต์ให้บริการจองโรงแรมที่พักคือ www.booking.com ซึ่งเป็นข้อความบทวิจารณ์ห้องพักหลังจากที่ผู้ใช้บริการได้เข้าพักแล้ว ซึ่งมีทั้งความคิดเห็นเชิงลบ และ ความคิดเห็นเชิงบวก ในผู้ใช้คนเดียว โดยเก็บข้อมูลอยู่ในรูปแบบของ text file แบ่งเป็นความคิดเห็นเชิงบวกภาษาอังกฤษจำนวน 4,000 เอกสาร และ ความคิดเห็นเชิงลบภาษาอังกฤษจำนวน 4,000 เอกสาร เมื่อได้ข้อมูลเอกสารที่อยู่ในรูปแบบ text file จะนำเอกสารที่ได้เข้าสู่กระบวนการเตรียมข้อมูล ซึ่งมีขั้นตอนดังนี้

การตัดคำ ในงานวิจัยนี้ตัดคำโดยการแบ่งคำแต่ละคำออกจากประโยค โดยจะใช้ช่องว่างในการแบ่งขอบเขตของคำ

การตัดคำหยุด (Stop Word) เป็นขั้นตอนการกำจัดคำที่ไม่มีนัยสำคัญหรือไม่ส่งผลใด ๆ กับการประมวลผลออกไป ซึ่งจะทำให้การประมวลผลเร็วยิ่งขึ้น

การนำเสนอเอกสาร (Document Representation) เป็นขั้นตอนในการนำเสนอความสัมพันธ์ระหว่างคำและเอกสาร ให้อยู่ในรูปแบบเวกเตอร์ (Vector) โดยการให้ค่าน้ำหนักด้วย tf-tf-idf และ tf-icf เพื่อนำไปสร้างตัวแทนเว็กเตอร์ของเอกสารให้อยู่ในรูปแบบ Vector Space Model หรือ Bag of Word

เมื่อทำการให้ค่าน้ำหนักของคำและอยู่ในรูปแบบ Vector Space Model หรือ Bag of Word เรียบร้อยแล้ว ก็จะเข้าสู่กระบวนการเรียนรู้การจำแนกเอกสาร และ กระบวนการทดสอบการจำแนกเอกสาร แบบผสมผสาน โดยข้อมูลเอกสารในการเรียนรู้ เป็นข้อความความคิดเห็นเชิงบวก จำนวน 2,800 เอกสาร และ ข้อความความคิดเห็นเชิงลบจำนวน 2,800 เอกสาร เพื่อสร้างโมเดลแบบผสมผสานด้วย Ensemble Model ภายใต้กระบวนการ Voting ด้วย 3 เทคนิคคือ Support Vector Machine, Naive Bayes และ K nearest neighbour โดยข้อมูลชุดสอนการเรียนรู้ชุดเดียวกัน

หลังจากกระบนการเรียนรู้เสร็จสิ้น เข้าสู่กระบวนการทดสอบโมเดล ด้วยชุดข้อมูลทดสอบแยกเป็นความคิดเห็นเชิงบวกจำนวน 1,200 เอกสาร และ ข้อความเชิงลบจำนวน 1,200 เอกสาร การวัดประสิทธิภาพการทดลองโดยใช้วิธีการวัดค่าความถูกต้อง (Accuracy) ความความระลึก (Precision) ค่าความแม่นยำ (Recall) และ ค่า F-measure ซึ่งผลการวัดประสิทธิภาพการทดลองการจำแนกเอกสารด้วยวิธีผสมผสานเป็นดังนี้

ค่าความถูกต้อง (Accuracy) = 0.97
ค่าความแม่นยำ (Recall) = 0.98
ความความระลึก (Precision) = 0.97
ค่า F-measure = 0.97

ดังนั้นจากผลการทดสอบระบบมีความน่าเชื่อถือในระดับที่ดีมาก แสดงให้เห็นว่าเทคนิคการจำแนกเอกสารด้วยเทคนิควิธีแบบผสมผสานสามารถจำแนกเอกสารได้ดีกว่าการจำแนกเอกสารแบบเทคนิควิธีเดียว ทั้งนี้ประสิทธิภาพการจำแนกเอกสารจะขึ้นอยู่กับข้อความที่นำมาเรียนรู้ด้วย หากมีความชัดเจนด้านความรู้สึกและสื่อความหมายได้ดีประสิทธิภาพในการจำแนกเอกสารจะสูง

นอกจากนี้ ปัจจัยที่ส่งผลต่อประสิทธิภาพการทดลองปัจจัยหนึ่งก็คือ การให้ค่าน้ำหนักคำด้วยวิธีต่างๆ ซึ่งในงานวิจัยนี้ ได้มีการเปรียบเทียบผลการทดลองจากการให้ค่าน้ำหนักของคำด้วย 3 วิธี คือ tf, tf-idf และ tf-icf ผลการทดลองเมื่อเปรียบเทียบผลจากการให้น้ำหนักด้วยวิธีต่างๆ แล้วผลปรากฏว่าการให้น้ำหนักคำด้วย tf-icf ให้ผลการทดลองได้ดีกว่า การให้ค่าน้ำหนักคำด้วย tf และ tf-idf ค่อนข้างมาก เพราะ tf-icf จะให้ความสำคัญโดยพิจารณาค่าน้ำหนักของคำหนึ่งคำโดยแยกคลาส แต่ tf และ tf-idf จะไม่พิจารณาน้ำหนักของแต่ละคลาส แต่จะพิจารณาแบบรวมทุกคลาส

5.2 ปัญหาอุปสรรคที่พบ

ในการเก็บรวบรวมข้อมูลเพื่อเข้าสู่กระบวนการเรียนรู้ จะใช้เวลาค่อนข้างมาก เพราะ ในแต่ละเพจจะมีความคิดเห็นที่ใช้ได้ไม่เยอะ เนื่องจากในงานวิจัยต้องการข้อความที่จะประมวลผลเป็นข้อความสั้น ที่มีอักขระ 150-250 อักขระ จำต้องให้ความสำคัญกับส่วนนี้เป็นพิเศษ

ในข้อความแสดงความรู้สึกของผู้ใช้บริการบางท่านจะมีการใส่อีโมจิ (emoji) เข้ามาในข้อความด้วย จนทำให้บางครั้งโปรแกรมเก็บรวบรวมข้อมูล ที่ผู้วิจัยได้พัฒนาขึ้นอาจมีความผิดพลาดได้ เนื่องจากค้นพบอักขระบางตัวที่ไม่สามารถจัดเก็บ หรือ ประมวลผลได้

ในกระบวนการเตรียมข้อมูล การให้น้ำหนักคำด้วย tf-icf อาจจะใช้เวลามาก เนื่องจากการให้น้ำหนักด้วย tf-icf จะประมวลผลของคำทีละคลาส ทำให้ใช้เวลามากกว่า tf และ tf-idf ซึ่งจะให้ค่าน้ำหนักแบบรวมคลาส

ในกระบวนสร้างเดลในการเรียนรู้ เนื่องจากงานวิจัยนี้ ได้เลือกใช้เทคนิคแบบผสมผสานด้วย 3 เทคนิค ภายใต้กระบวนการ Ensemble Model ด้วยวิธีการโหวต กระบวนการนี้จะใช้เวลานาน เพราะต้องใช้เวลาสร้างโมเดลด้วยกัน 3 โมเดลก่อน และ หลังจากนั้นต้องทำการโหวต ดังนั้นจะต้องใช้เวลาในการประมวลผลเป็นอย่างมาก

5.3 แนวทางการพัฒนางานวิจัยทางด้านการจำแนกความรู้สึกข้อความสั้น

ความท้าทายในการจำแนกเอกสารข้อความสั้นในอนาคต สามารถขยายขอบเขตของกระบวนการวิจัยได้ดังนี้

การเพิ่มคลาสให้มากขึ้น นอกจาก การแสดงความคิดเห็นเชิงบวก และ ความคิดเห็นเชิงลบแล้ว อาจจะเพิ่มคลาสได้อีก เช่น ความคิดเห็นเป็นกลาง เพื่อความละเอียดในคำตอบภาพรวมได้ดียิ่งขึ้น

การจำแนกเอกสารที่มีอีโมจิ หรือ เครื่องหมายแสดงอารมณ์อื่น ๆ ที่ผสมผสานมากับข้อความ เนื่องจากปัจจุบัน การแสดงความคิดเห็นต่าง ๆ ตามเว็บไซต์ ผู้ใช้บริการจะใช้ สมาร์ทโฟนในการให้คำตอบ ดังนั้นจึงมีการใส่ไอค่อนแสดงอารมณ์เข้ามาด้วย ดั้งนั้นการที่จะประมวลผลด้วยสัญลักษณ์ หรือ ไอค่อนต่าง ๆ ที่แสดงอารมณ์ ก็เป็นอีกความท้าทายหนึ่งที่น่าสนใจ

การจำแนกบทวิจารณ์ข้อความสั้นแบบอัตโนมัติภาษาอื่น ๆ เช่น ภาษาไทย ที่โครงสร้างทางภาษาแตกต่างจากภาษาอังกฤษพอสมควร หรือ คำคำหนึ่งอาจมีความหมายที่แต่กต่าง เช่น “ตากลม” อาจอ่านได้ทั้ง ตาก-ลม และ ตา-กลม

การจำแนกเอกสารด้วย Ensemble model ภายใต้กระบวนการอื่น ๆ นอกจากกระบวนการโหวต เช่น Bagging หรือ Random forest หรือ แม้กระทั่งการเปลี่ยนเทคนิคในการสร้างโมเดลเป็นเทคนิคอื่นที่เหมาะสมกับ Ensemble แบบ Bagging หรือ Ensemble แบบ Random forest เช่น Neural Network หรือ Decision Tree

บรรณานุกรม

1. Pang, B. and L. Lee, *Opinion mining and sentiment analysis.* Foundations and trends in information retrieval, 2008. **2**(1-2): p. 1-135.
2. Liu, B., *Sentiment analysis and opinion mining.* Synthesis lectures on human language technologies, 2012. **5**(1): p. 1-167.
3. Esuli, A. and F. Sebastiani. *Sentiwordnet: A publicly available lexical resource for opinion mining*. in *Proceedings of LREC*. 2006. Citeseer.
4. Chowdhury, G.G., *Natural language processing.* Annual review of information science and technology, 2003. **37**(1): p. 51-89.
5. Aurchana.P, I.R., Periyasamy.P,, *Sentiment Analysis in Tourism.* IJISET - International Journal of Innovative Science, Engineering & Technology, 2014. **1**(9).
6. Shi, H.-X. and X.-J. Li. *A sentiment analysis model for hotel reviews based on supervised learning*. in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*. 2011. IEEE.
7. Elango, V. and G. Narayanan, *Sentiment Analysis for Hotel Reviews.* 2011.
8. Dave, K., S. Lawrence, and D.M. Pennock. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. in *Proceedings of the 12th international conference on World Wide Web*. 2003. ACM.
9. Pak, A. and P. Paroubek. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. in *LREc*. 2010.
10. Whitehead, M. and L. Yaeger, *Sentiment mining using ensemble classification models*, in *Innovations and advances in computer sciences and engineering*. 2010, Springer. p. 509-514.
11. Montoyo, A., P. MartíNez-Barco, and A. Balahur, *Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments.* Decision Support Systems, 2012. **53**(4): p. 675-679.
12. Gräbner, D., et al., *Classification of customer reviews based on sentiment analysis*. 2012: na.

13. McCallum, A. and K. Nigam. *A comparison of event models for naive bayes text classification*. in *AAAI-98 workshop on learning for text categorization*. 1998. Citeseer.
14. Kim, S.-B., et al., *Some effective techniques for naive bayes text classification.* IEEE transactions on knowledge and data engineering, 2006. **18**(11): p. 1457-1466.
15. Joachims, T. *Transductive inference for text classification using support vector machines*. in *ICML*. 1999.
16. Joachims, T. *Text categorization with support vector machines: Learning with many relevant features*. in *European conference on machine learning*. 1998. Springer.
17. Wermter, S., *Neural network agents for learning semantic text classification.* Information Retrieval, 2000. **3**(2): p. 87-103.
18. Lam, S.L. and D.L. Lee. *Feature reduction for neural network based text categorization*. in *Database Systems for Advanced Applications, 1999. Proceedings., 6th International Conference on*. 1999. IEEE.
19. Friedl, M.A. and C.E. Brodley, *Decision tree classification of land cover from remotely sensed data.* Remote sensing of environment, 1997. **61**(3): p. 399-409.
20. Yang, Y. and X. Liu. *A re-examination of text categorization methods*. in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 1999. ACM.
21. Zhang, M.-L. and Z.-H. Zhou. *A k-nearest neighbor based algorithm for multi-label classification*. in *2005 IEEE international conference on granular computing*. 2005. IEEE.
22. Han, E.-H.S., G. Karypis, and V. Kumar. *Text categorization using weight adjusted k-nearest neighbor classification*. in *Pacific-asia conference on knowledge discovery and data mining*. 2001. Springer.
23. Yin, C., et al., *Short Text Classification Algorithm Based on Semi-Supervised Learning and SVM.* International Journal of Multimedia and Ubiquitous Engineering, 2015. **10**(12): p. 195-206.
24. Li, L. and S. Qu, *Short Text Classification Based on Improved ITC.* Journal of

Computer and Communications, 2013. **2013**.

25. Kasper, W. and M. Vela. *Sentiment analysis for hotel reviews*. in *Computational linguistics-applications conference*. 2011.
26. Ge Song, Y.Y., Xiaolin Du, Xiaohui Huang, and Shifu Bie, *Short Text Classification: A Survey.* JOURNAL OF MULTIMEDIA, 2014. **9**(5): p. 10.
27. Abbasi, A., H. Chen, and A. Salem, *Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums.* ACM Transactions on Information Systems (TOIS), 2008. **26**(3): p. 12.
28. Melville, P., W. Gryc, and R.D. Lawrence. *Sentiment analysis of blogs by combining lexical knowledge with text classification*. in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2009. ACM.
29. Jindal, R., R. Malhotra, and A. Jain, *Techniques for text classification: Literature review and current trends.* Webology, 2015. **12**(2): p. 1.
30. Medhat, W., A. Hassan, and H. Korashy, *Sentiment analysis algorithms and applications: A survey.* Ain Shams Engineering Journal, 2014. **5**(4): p. 1093-1113.
31. Yuan, Q., G. Cong, and N.M. Thalmann. *Enhancing naive bayes with various smoothing methods for short text classification*. in *Proceedings of the 21st International Conference on World Wide Web*. 2012. ACM.
32. Chen, M., X. Jin, and D. Shen. *Short text classification improved by learning multi-granularity topics*. in *IJCAI*. 2011. Citeseer.
33. Sriram, B., et al. *Short text classification in twitter to improve information filtering*. in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010. ACM.
34. Yin, C., et al. *A new svm method for short text classification based on semi-supervised learning*. in *Advanced Information Technology and Sensor Application (AITS), 2015 4th International Conference on*. 2015. IEEE.
35. Thelwall, M., et al., *Sentiment strength detection in short informal text.* Journal of the American Society for Information Science and Technology, 2010. **61**(12): p. 2544-2558.

พหูน ปณ ฑิโต ชีเว

พหูน ปณฺ ฑิโต ชีเว

ประวัติผู้เขียน

ชื่อ	นายสหชัย งามชัยภูมิ
วันเกิด	วันที่ 3 ตุลาคม พ.ศ. 2531
สถานที่เกิด	อำเภอคอนสวรรค์ จังหวัดชัยภูมิ
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 390 หมู่ 10 ตำบลแวงน่าง อำเภอเมือง จังหวัดมหาสารคาม รหัสไปรษณีย์ 4400
ตำแหน่งหน้าที่การงาน	นักวิชาการคอมพิวเตอร์
สถานที่ทำงานปัจจุบัน	ศูนย์คอมพิวเตอร์ มหาวิทยาลัยราชภัฏมหาสารคาม
ประวัติการศึกษา	พ.ศ. 2546 มัธยมศึกษาตอนต้น โรงเรียนสามหมอวิทยา อำเภอคอนสวรรค์ จังหวัดชัยภูมิ พ.ศ. 2549 มัธยมศึกษาตอนปลาย โรงเรียนสามหมอวิทยา อำเภอคอนสวรรค์ จังหวัดชัยภูมิ พ.ศ. 2553 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาเทคโนโลยีสารสนเทศและการสื่อสาร คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม พ.ศ. 2563 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
ผลงานวิจัย	Sahachai Ngamchaiyaphum, Bancha Luaphol, Jantima Polpinij. (2020). Classification of Customer Reviews with Short Text based on Sentiment Analysis. International Symposium on Artificial Life and Robotics.