การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม
โดยใช้เทคนิคการตัดคำแบบผสมผสาน

สันติ สุขเกษม

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
บัณฑิตวิทยาลัย สาขาเทคโนโลยีสารสนเทศ
สถาบันเทคโนโลยีไทย-ญี่ปุ่น
ปีการศึกษา 2561

ANALYSIS OF SENTIMENT FOR HOTEL DATA USING THE HYBRID WORD WRAP TECHNIQUE

Santi Sukkasem

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science Program in Information Technology
Graduate School
Thai-Nichi Institute of Technology
Academic Year 2018

หัวข้อวิทยานิพนธ์	การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน
โดย	สันติ สุขเกษม
สาขาวิชา	เทคโนโลยีสารสนเทศ
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร. ประจักษ์ เฉิดโฉม
อาจารย์ที่ปรึกษาวิทยานิพนธ์ (ร่วม)	ดร. ธงชัย แก้วกิริยา

บัณฑิตวิทยาลัย สถาบันเทคโนโลยีไทย-ญี่ปุ่น อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญามหาบัณฑิต

.....คณบดีบัณฑิตวิทยาลัย
(รองศาสตราจารย์ ดร. พิชิต สุขเจริญพงษ์)
วันที่.....เดือน.....พ.ศ.....

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รองศาสตราจารย์ ดร. พยุง มีสัจ)

.....กรรมการ
(รองศาสตราจารย์ ดร. วรากร ศรีเชวงทรัพย์)

.....กรรมการ
(ดร. ณัฐกิตติ์ จิตรเอื้อตระกูล)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์
(ผู้ช่วยศาสตราจารย์ ดร. ประจักษ์ เฉิดโฉม)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ (ร่วม)
(ดร. ธงชัย แก้วกิริยา)

สันติ สุขเกษม : การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมโดยใช้เทคนิคการตัดคำแบบผสมผสาน. อาจารย์ที่ปรึกษา : ผู้ช่วยศาสตราจารย์ ดร. ประจักษ์ เฉิดโฉม อาจารย์ที่ปรึกษาร่วม : ดร. ธงชัย แก้วกิริยา, 87 หน้า.

การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผมผสาน มีวัตถุประสงค์ในการทำวิจัย ดังนี้ 1) เพื่อพัฒนาเทคนิคการตัดคำแบบผสมผสาน 2) เพื่อวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม 3) เพื่อประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็นโดยการจำแนกแบบตัดสินใจต้นไม้ แบบนาอีฟ เบย์ แบบเคเนียเรสเนเบอร์ และดำเนินการโดยรวบรวมข้อมูลจากเว็บไซต์อโกด้าเป็นข้อมูลการแสดงความคิดเห็นของผู้ใช้บริการที่เข้าพักโรงแรมนำมาวิเคราะห์ความรู้สึกโดยหลักการวิธีการตัดคำแบบผสมผสาน ซึ่งผลของการวิจัยมีดังนี้ ผลการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผมผสาน การวิเคราะห์จำแนกแบบนาอีฟ เบย์ ผลวิเคราะห์ความคิดเห็นเชิงบวก (Positive) 99.36%, ผลวิเคราะห์ความคิดเห็นเชิงลบ (Negative) 98.89%, ผลวิเคราะห์ความคิดเห็นเป็นกลาง (Neutral) 99.43%, ความถูกต้อง (Accuracy) 99.66% การจำแนกแบบต้นไม้ตัดสินใจ ผลวิเคราะห์ความคิดเห็นเชิงบวก (Positive) 100%, ผลวิเคราะห์ความคิดเห็นเชิงลบ (Negative) 0.00%, ผลวิเคราะห์ความคิดเห็นเป็นกลาง (Neutral) 0.00%, ความถูกต้อง (Accuracy) 65.49% และการจำแนกแบบเคเนียเรสเนเบอร์ ผลวิเคราะห์ความคิดเห็นเชิงบวก (Positive) 83.68%, ผลวิเคราะห์ความคิดเห็นเชิงลบ (Negative) 26.16%, ผลวิเคราะห์ความคิดเห็นเป็นกลาง (Neutral) 45.00%, ความถูกต้อง (Accuracy) 70.21%

บัณฑิตวิทยาลัย ลายมือชื่อนักศึกษา .....
สาขาวิชา เทคโนโลยีสารสนเทศ ลายมือชื่ออาจารย์ที่ปรึกษา .....
ปีการศึกษา 2561 ลายมือชื่ออาจารย์ที่ปรึกษา (ร่วม) .....

SANTI SUKKASEM : ANALYSIS OF SENTIMENT FOR HOTEL DATA USING THE HYBRID WORD WRAP TECHNIQUE. ADVISOR : ASST. PROF. DR. PRAJAK CHERTCHOM AND CO-ADVISOR : DR. THONGCHAI KAEWKIRIYA, 87 PP.

This research's objective is the analysis of sentiment for hotel data using the hybrid word wrap technique. The objectives of this research are as follows: 1) to develop hybrid word wrap technique. 2) to analysis the form of comments of hotel information. 3) to evaluate the efficiency of sentiment analysis by deciding on the decision tree, Naive Bayes, K-Nearest Neighbor and proceed by collecting information from the website Agoda is the information showing the sentiment of the guests who stay. The hotel is analysis by the hybrid word wrap technique. The results of the research are as follows: Analysis by Naïve Bayes Positive = 99.36%, Negative = 98.89%, Neutral 99.43%, Accuracy = 99.66% and Analysis by Decision Tree Positive = 100%, Negative = 0.00% Neutral = 0.00%, Accuracy 65.49% by K-Nearest Neighbor Positive = 83.68%, Negative = 26.16%, Neutral 45.00%, Accuracy = 70.21% respectively.

Graduate School Student's Signature.....

Field of Study Information Technology Advisor's Signature.....

Academic Year 2018 Co-Advisor's Signature.....

กิตติกรรมประกาศ

การทำวิทยานิพนธ์นี้จะสำเร็จลุล่วงไปได้ด้วยดีนั้น ผู้วิจัยขอกราบขอบพระคุณในความกรุณาของ ผู้ช่วยศาสตราจารย์ ดร. ประจักษ์ เฉิดโฉม และ ดร. ธงชัย แก้วกิริยา ซึ่งได้สละเวลาในการให้คำแนะนำ คำปรึกษา การสนับสนุน รวมถึงตรวจสอบข้อบกพร่องทุกขั้นตอนและกำลังใจเต็มเปี่ยมตลอดระยะเวลาในการทำวิทยานิพนธ์ฉบับนี้

ทั้งนี้ผู้วิจัยขอกราบขอบพระคุณคณะกรรมการสอบทุกท่าน ได้แก่ รองศาสตราจารย์ ดร. พยุง มีสัจ รองศาสตราจารย์ ดร. วรากร ศรีเชวงทรัพย์ และ ดร. ณัฐกิตติ์ จิตรเอื้อตระกูล ที่ได้กรุณาให้ข้อเสนอแนะที่เป็นประโยชน์และช่วยตรวจสอบชี้ข้อบกพร่องของวิทยานิพนธ์ฉบับนี้ด้วยความเอาใจใส่ จนทำให้วิทยานิพนธ์ฉบับนี้สำเร็จและมีความสมบูรณ์

เหนือสิ่งอื่นใดผู้วิจัยขอกราบขอบพระคุณบิดามารดาของผู้วิจัยที่คอยให้กำลังใจ และให้การสนับสนุนในทุกๆ ด้านอย่างที่สุดเสมอมา

ผู้วิจัยหวังเป็นอย่างยิ่งว่า วิทยานิพนธ์ฉบับนี้จะเกิดประโยชน์ต่อผู้ที่มีความสนใจ สามารถนำไปต่อยอด องค์ความรู้เพื่อประยุกต์ใช้งานกับระบบการแนะนำการศึกษาเพื่อให้เกิดคุณค่าและสามารถใช้งานได้จริง

สันติ สุขเกษม

สารบัญ

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

บทที่ 1
บทนำ

1.1 สภาวะความเป็นมา แนวทางเหตุผล และปัญหา

ปัจจุบันข้อมูลขนาดใหญ่ (Big Data) มีความสำคัญและเป็นเป้าหมายหลักของการวิเคราะห์ธุรกิจ (Business Analysis) ชุดข้อมูลในปัจจุบันมีความซับซ้อนมากเป็นอย่างยิ่งทำให้การวิเคราะห์และประมวลผลชุดข้อมูลจำเป็นต้องใช้แบบจำลองทางคณิตศาสตร์จากปัจจัยต่างๆ ที่นำมาวิเคราะห์ข้อมูลและสามารถนำมาใช้หาความหมายแสดงผลการวิเคราะห์ข้อมูลในด้านต่างๆ รวมถึงประกอบการตัดสินใจวางแผนต่างๆ เพื่อให้สามารถบรรลุเป้าหมายที่จะเกิดขึ้นได้อย่างมีประสิทธิภาพ [1] โดยในช่วงยุคปัจจุบันที่มีอินเตอร์เน็ตเข้ามามีบทบาทต่อการเชื่อมต่อสื่อสารระหว่างบุคคลรวมไปถึงการรับรู้ข่าวสารที่รวดเร็วผ่านโซเชียลเน็ตเวิร์ก (Social Network) จึงเกิดการวิเคราะห์ข้อมูลที่บ่งชี้สภาวะทางอารมณ์ของคน (Sentiment Analysis : SA) หรือที่ทั่วไปนั้นเรียกว่า “การวิเคราะห์ความรู้สึก” เข้ามาเกี่ยวข้องเป็นอย่างมากไม่ว่าจะเป็นการแสดงออกทางด้านความคิดความรู้สึกเพื่อทำการแยกค่าคะแนนที่เป็นเชิงบวก (Positive) เชิงลบ (Negative) [2] และข้อความคิดเห็นที่เป็นภาษาไทยจะมีการเขียนคำที่ติดกันไปทั้งประโยคไม่มีการเว้นว่างระหว่างคำเหมือนกับภาษาอื่นๆ เช่น ภาษาอังกฤษหรือให้อยู่ในรูปของอักษรตัวเดียวเหมือนอย่างภาษาจีน ดังนั้นการประมวลผลทางภาษาหรือการตัดคำ (Word Segmentation) จึงทำให้เกิดการแบ่งคำแต่ละคำที่อยู่ในประโยคนั้นออกจากกันเพื่อการนำไปวิเคราะห์ประโยคต่างๆ เช่น การวิเคราะห์หลักไวยกรณ์ การประมวลผลภาษาธรรมชาติ การเชื่อมโยงความหมายของคำ การสังเคราะห์และวิเคราะห์กฎของการสร้างประโยค เป็นต้น

การตัดคำที่เป็นภาษาไทยถูกพัฒนาขึ้นในรูปแบบต่างๆ โดยแบ่งออกเป็น 3 หลักวิธีที่ใหญ่ๆ คือ 1. การใช้กฎ (Rule-Based) ให้การประมวลผลที่มีความเร็วสูงใช้ทรัพยากรที่น้อยมีความถูกต้องหลังการตัดในระดับพยางค์ค่อนข้างสูง 2. การใช้พจนานุกรม (Dictionary) ให้ความแม่นยำในการตัดคำค่อนข้างสูงและยังสามารถเพิ่มคำที่ไม่รู้จักเข้าไปในพจนานุกรมได้ 3. การใช้คลังข้อความ (Text Corpus) สามารถสร้างกฎไวยากรณ์ได้จากการเรียนรู้จากคลังข้อมูล ซึ่งแต่ละวิธีมีข้อดีที่แตกต่างกันออกไปตามแต่ละคุณสมบัติที่กล่าวไปข้างต้น แต่เมื่อพิจารณาแล้วทุกๆ วิธียังคงมีการใช้กฎร่วมด้วยเป็นส่วนใหญ่ เพื่อที่จะจัดการกับคำที่ไม่รู้จักและคำที่ไม่อยู่ในพจนานุกรม เช่น คำเฉพาะบางกลุ่ม คำกำกวม คำทับศัพท์ที่มาจากภาษาต่างประเทศ คำที่อ่านออกเสียงได้หลายแบบแต่ยังคงรูปเดิม คำสแลง ซึ่งปัจจุบันจะพบคำพวกนี้มากขึ้นและมีการถอดความเป็นภาษาไทยไม่ตรงกับหลักไวยากรณ์ไทยตามหลักของบัณฑิตยสถานฉบับปีพุทธศักราช 2554 งานวิจัยส่วนใหญ่มุ่งเน้นในการพัฒนาการตัดคำตามหลักไวยากรณ์ของภาษาไทยเพราะเป็นปัจจัยหลักในกระบวนการตัดคำที่ผ่านมาในงานวิจัย

เล่มนี้จึงขอนำเสนอการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน เพิ่มประสิทธิภาพของการตัดคำให้มีความถูกต้องแม่นยำมากยิ่งขึ้นเพื่อลดปัญหาของคำที่เกิดจากความกำกวานในประโยค คำที่เกิดจากการทับศัพท์จากภาษาอังกฤษ และคำที่เกิดจากการอ่านออกเสียงได้หลายแบบแต่คงรูปเดิมรวมไปถึงการพัฒนาวิธีการและขั้นตอนในการตัดคำให้ถูกต้องมากยิ่งขึ้นและเข้าสู่กระบวนการของเหมืองข้อความต่อไป

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อพัฒนาเทคนิคการตัดคำแบบผสมผสาน

1.2.2 เพื่อวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม

1.2.3 เพื่อประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็นโดยการจำแนกนาอีฟ เบย์ (Naïve Bayes) แบบต้นไม้ตัดสินใจ (Decision Tree) และแบบเคเนียเรสเนเบอร์ (K-Nearest Neighbor)

1.3 ขอบเขตของการวิจัย

1.3.1 ขอบเขตด้านข้อมูล

การวิเคราะห์กระบวนการตัดคำโดยเทคนิคผสมผสานนั้นข้อมูลได้นำมาจากการแสดงความคิดเห็นของผู้ใช้โรงแรมผ่านเว็บไซต์อโกด้า (Agoda) ซึ่งเป็นข้อมูลการแสดงความคิดเห็นของผู้ใช้บริการที่ผ่านการล็อกอินจำนวน 2,040 ความคิดเห็น จาก 10 จังหวัด ที่นักท่องเที่ยวเที่ยวมากที่สุดในประเทศไทยจำนวน 40 โรงแรม

1.3.2 ขอบเขตด้านเทคนิคและวิธีการ

1.3.2.1 ส่วนของการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมจะทำการวิเคราะห์โดยใช้เทคนิคการตัดคำแบบผสมผสาน เพื่อให้ได้ผลลัพธ์ของการตัดคำที่เหมาะสมถูกต้องที่สุดและทำไปเข้าสู่กระบวนการวิเคราะห์ของขั้นตอนการทำเหมืองข้อความต่อไป

1.3.2.2 การประเมินประสิทธิภาพการคำตัดของข้อมูลโรงแรมด้วยอัลกอริทึมตันสินใจต้นไม้ที่พิจารณาลักษณะของข้อมูลเพื่อเชื่อมโยงข้อมูลเข้าด้วยกันอัลกอริทึมนาอีฟ เบย์การที่คาดการณ์ผลลัพธ์โดยทำการวิเคราะห์ระหว่างความสัมพันธ์การสร้างเงื่อนไขของความน่าจะเป็น และอัลกอริทึมเคเนียเรสเนเบอร์ในการทดลองวิจัยครั้งนี้

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ได้ข้อมูลที่ถูกต้องเพื่อนำไปเป็นแนวทางในการพัฒนาและปรับปรุงงานในกระบวนการของการทำเหมืองข้อความ

1.4.2 ได้กระบวนการขั้นตอนวิธีในการตัดคำของภาษาไทยโดยการใช้เทคนิคผสมผสาน

1.4.3 ได้ฝึกฝนวิธีการตั้งปัญหาในด้านข้อมูลและแก้ไขปัญหา รวมถึงการนำศาสตร์ต่างๆ มาประยุกต์ใช้เพื่อประโยชน์ในการสร้างสรรค์งานวิจัย

1.4.4 ได้รู้จักแนวคิดกระบวนการสร้างพัฒนาแบบจำลอง เพื่อนำมาเป็นเครื่องมือในการเลือกและตัดสินใจหรือเป็นแนวทางนำวิธีดังกล่าวไปประยุกต์ในรูปแบบอื่น

1.5 นิยามศัพท์เฉพาะ

1.5.1 **การวิเคราะห์ความรู้สึก (Sentiment Analysis)** เป็นการวิเคราะห์ความคิดเห็นจากความรู้สึกข้อความบนโซเชียลมีเดีย (Social Media) ผ่านโครงสร้างและได้ผลลัพธ์ออกมาเป็นความรู้สึกเชิงบวก (Positive) และความรู้สึกเชิงลบ (Negative)

1.5.2 **การตัดคำ (Word Segmentation)** เป็นคำที่ตัดออกมาจากคำหรือประโยคเต็มที่เป็นภาษามาตรฐานถูกต้องตามหลักไวยากรณ์และวิธีการ เช่น คำว่า “อาจ” ตัดมาจากคำองอาจ หรืออาจหาญ

1.5.3 **โซเชียลเน็ตเวิร์ก (Social Network)** เป็นเว็บไซต์ที่เชื่อมโยงผู้คนไว้ด้วยกันผ่านระบบอินเตอร์ ซึ่งเป็นเว็บไซต์ที่สามารถช่วยสร้างพื้นที่ส่วนตัวขึ้นมาหรือแม้กระทั่งแชร์เรื่องราวให้กับบุคคลอื่นๆ ก็ได้ ตัวอย่าง Social Network เช่น Facebook, Twitter

1.5.4 **พจนานุกรม (Dictionary)** เป็นหนังสือสำหรับค้นความหมายของคำเรียงลำดับตามตัวอักษรพยัญชนะ สระ วรรณยุกต์ มักบ่งบอกที่มาของคำที่ค้นหาและความหมายของคำด้วย

1.5.5 **การตัดสินใจต้นไม้ (Decision Tree)** เป็นอัลกอริทึมที่ใช้หลังการทางคณิตศาสตร์มาพิจารณาลักษณะของข้อมูลเพื่อเชื่อมโยงข้อมูลเข้าด้วยกันการทำงานแบบต้นไม้ตัดสินใจจะทำการจัดกลุ่ม ชุดข้อมูลนำเข้าในแต่ละกรณีแต่ละบัพหรือเรียกว่า “โหนด” ของต้นไม้ตัดสินใจคือตัวแปรต่างๆ ของข้อมูล

1.5.6 **นาอีฟ-เบย์ (Naïve Bayes Classifier)** เป็นอัลกอริทึมที่ใช้หลังการความน่าจะเป็นที่สามารถให้การคาดการณ์ผลลัพธ์และยังสามารถอธิบายได้ โดยการทำการวิเคราะห์ระหว่างความสัมพันธ์ของตัวแปรเพื่อใช้ในการสร้างเงื่อนไขของความน่าจะเป็น

บทที่ 2
แนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้อง

การวิจัยครั้งนี้ผู้วิจัยได้ศึกษาค้นคว้าแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องจากเอกสาร บทความวิชาการและแหล่งสารสนเทศอินเทอร์เน็ต ดังหัวข้อต่อไปนี้

2.1 แนวคิดเกี่ยวกับความคิดเห็น

2.2 แนวคิดเกี่ยวกับความรู้สึก

2.3 ลักษณะของภาษาไทย

2.4 แนวคิดและทฤษฎีเกี่ยวกับการตัดคำภาษาไทย

2.5 วิธีที่ใช้ในการตัดคำ

2.6 อัลกอริทึม การแบ่งประเภท

2.7 การให้น้ำหนักคำ

2.8 งานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดเกี่ยวกับความคิดเห็น

2.1.1 ความหมายของความคิดเห็น

ราชบัณฑิตยสถานพจนานุกรมศัพท์สังคมวิทยาฉบับราชบัณฑิตยสถาน 2542 ได้ให้ความหมาย ความคิดเห็นว่าเป็นข้อพิจารณาว่าเป็นจริงจากการใช้ปัญญาความคิดประกอบถึงแม้จะไม่ได้อาศัยหลักฐานพิสูจน์ ยืนยันได้เสมอไปก็ตาม นพมาศ ธีรเวคิน [3] อธิบายว่า ความคิดเห็นนั้นถูกจัดว่าเป็นส่วนที่มนุษย์ได้แสดงออกมาโดยการพูดหรือเขียน มนุษย์นั้นจะพูดจากใจจริง พูดตามสังคมหรือพูดเพื่อเอาใจผู้ฟังก็ตามแต่เมื่อพูดหรือเขียนไปแล้วก็ทำให้เกิดผลได้ คนส่วนใหญ่มักจะถือว่าสิ่งที่มนุษย์แสดงออกมานั้นเป็นสิ่งที่สะท้อนถึงความในใจด้วยเหตุนี้ จึงเป็นที่นิยมกันมากที่จะสำรวจความคิดเห็นต่อสิ่งหนึ่งหรือเรียกกันว่าการสำรวจประชามติ (Polling) จึงอาจกล่าวได้ว่า การหยั่งประชามติเป็นเครื่องมือสำคัญทางวิชาการที่ศึกษาและสำรวจการแสดงออกทางความคิดเห็นในปัจจุบันความคิดเห็นว่าเป็นการแสดงออกทางถ้อยคำ (Verbal Expression) เกี่ยวกับทัศนคติความเชื่อหรือค่านิยม แต่ความคิดเห็นไม่ใช่สิ่งเดียวกับทัศนคติ เพราะในตัวของมันเองไม่จำเป็นต้องประกอบด้วยองค์ประกอบทางอารมณ์หรือพฤติกรรม อัมพาพร นพรัตน์ [4] สรุปไว้ว่า ความคิดเห็นเป็นการแสดงออกด้านความรู้สึกของบุคคล ต่อสิ่งใดสิ่งหนึ่ง ซึ่งมีพื้นฐานจากส่วนประกอบต่างๆ ได้แก่ ความรู้ ประสบการณ์ และสภาพแวดล้อม สภาพแวดล้อม สมรรถชัย คันธมาทน์ [5] สรุปไว้ว่า ความคิดเห็นเป็นการแสดงออกทางด้านความรู้สึกโดยอาศัยพื้นความรู้ ประสบการณ์ที่ผ่านมา หรือที่ได้พบเจอมา ซึ่งการแสดงความ

คิดเห็นนี้อาจจะได้รับการยอมรับหรือปฏิเสธจากคนอื่นๆ ก็ได้ มัณฑนา โพธิ์แก้ว [6] สรุปไว้ว่า ความคิดเห็นเป็นการแสดงออกทางด้านความรู้สึก ความคิด ความเชื่อ และการตัดสินใจเกี่ยวกับข้อเท็จจริงอย่างใดอย่างหนึ่ง หรือเป็นการประเมินจากสถานการณ์ต่างๆ การแสดงออกถึงความคิดเห็นจะมีผลต่อเนื่องมาจากอารมณ์ พื้นความรู้ ภูมิหลังของทางสังคม และสภาพความจริงในขณะนั้น ความคิดเห็นไม่อาจบอกได้ว่าถูกต้องหรือไม่อาจจะได้รับการยอมรับหรือปฏิเสธจากคนอื่นก็ได้ ซึ่งความคิดเห็นสามารถเปลี่ยนแปลงได้เมื่อเวลาเปลี่ยนไปในขณะที่ จิตตินันท์ เดชะคุปต์ [7] ให้ความหมายความคิดเห็นว่าหมายถึง การแสดงออกถึงความเชื่อบางอย่าง เจตคติบางอย่าง และค่านิยมบางอย่างของบุคคลต่อสิ่งหนึ่งสิ่งใด ซึ่งการกระทำอาจไม่สอดคล้องกับการแสดงออกทางความคิดเห็นต่อสิ่งนั้น และนพพร ไพบูลย์ [8] ได้กล่าวถึง ความคิดเห็นว่าเป็นการแสดงออกของบุคคลอาจเป็นการพูดหรือการเขียน ซึ่งบุคคลนั้นมีความรู้สึกจากประสบการณ์เดิม และสภาพแวดล้อมในขณะนั้น อาจนำไปสู่การคาดคะเนหรือแปลผลในพฤติกรรมของบุคคลนั้น แม้อาจไม่ถูกต้องหรือตรงกับความคิดเห็นของผู้อื่นก็ตาม

นอกจากนี้ความคิดเห็นว่าเกิดจากมูลเหตุ 2 ประการ คือ 1) ประสบการณ์ที่บุคคลมีต่อสิ่งของ บุคคล หมู่คณะ เรื่องราว หรือสถานการณ์ต่างๆ โดยความคิดเห็นจะเกิดขึ้นในตัวบุคคลจากการได้พบเห็นความคุ้นเคย ซึ่งเป็นประสบการณ์ตรงจากการได้ยินได้ฟังได้เห็นรูปถ่ายหรืออ่านจากหนังสือ โดยไม่ได้พบเห็นของจริงถือว่าเป็นประสบการณ์ทางอ้อม 2) ระบบค่านิยมและการตัดสินค่านิยมหากแต่ละกลุ่มที่นิยมและการตัดสินของค่านิยมไม่เหมือนกัน ความคิดเห็นในสิ่งต่างๆ ก็จะแตกต่างกันไปด้วย [9]

การสำรวจความคิดเห็นเป็นการศึกษาความรู้สึกของบุคคลกลุ่มคนที่มีต่อสิ่งใดสิ่งหนึ่ง แต่ละคนจะแสดงความเชื่อและความรู้สึกใดๆ ออกมาโดยการพูดและการเขียน เป็นต้น การสำรวจความคิดเห็นจะเป็นประโยชน์ต่อการวางนโยบายต่างๆ การเปลี่ยนแปลงนโยบาย หรือการเปลี่ยนแปลงระบบงาน รวมทั้งในการฝึกหัดทำงานด้วยเพราะว่าจะทำให้การดำเนินงานต่างๆ เป็นไปด้วยความเรียบร้อย และเป็นไปตามความพอใจของผู้ร่วมงาน [10]

ความคิดเห็นว่า หมายถึง การแสดงออกด้านความรู้สึกของบุคคลต่อสิ่งหนึ่งสิ่งใดด้วยการพูดและการเขียน โดยมีพื้นฐานความรู้เดิมประสบการณ์ที่บุคคลได้รับตลอดจนสภาพแวดล้อมของบุคคลนั้นเป็นหลักในการแสดงความคิดเห็น [11]

ปัจจัยที่มีอิทธิพลต่อการต่อการเกิดความคิดเห็น นอกจากประสบการณ์คือ 1) ปัจจัยทางพันธุกรรมและสรีระ ได้แก่ อวัยวะต่างๆ ของบุคคลที่ใช้รับรู้ ความผิดปกติของอวัยวะ ความบกพร่องของอวัยวะ สัมผัสมีผลต่อความคิดเห็นไม่ดีต่อบุคคลภายนอก 2) อิทธิพลของผู้ปกครองคือ เมื่อเป็นเด็กผู้ปกครองจะเป็นผู้ที่อยู่ใกล้ชิดและให้ข้อมูลแก่เด็กได้มาก ซึ่งจะมีผลต่อพฤติกรรมและความคิดเห็นของเด็กด้วย 3) ทัศนคติและความคิดเห็นของกลุ่มคือ เมื่อบุคลิกเจริญเติบโตย่อมจะมีกลุ่มและสังคม ดังนั้นความคิดของกลุ่มเพื่อน กลุ่มอ้างอิงหรือการอบรมสั่งสอนในโรงเรียน หน่วยงานที่มีความคิดเห็น

เหมือนหรือแตกต่างกันย่อมจะส่งผลต่อความคิดเห็นของบุคคลด้วย 4) สื่อมวลชน คือ สื่อต่างๆ ที่เข้ามามีบทบาทในชีวิตประจำวัน อันได้แก่ โทรทัศน์ วิทยุ หนังสือพิมพ์ และนิตยสารเป็นปัจจัยอันหนึ่งที่มีผลต่อความคิดเห็นของบุคคล [12] และความคิดเห็นเป็นการแสดงออกด้านความรู้สึกสิ่งหนึ่งสิ่งใดเป็นความรู้สึกเชื่อถือที่ไม่ได้อยู่บนความแน่นอนหรือความจริง แต่ขึ้นอยู่กับจิตใจบุคคลจะแสดงออกโดยมีข้ออ้างหรือการแสดงเหตุผลสนับสนุน หรือปกป้องความคิดเห็นนั้น ความคิดเห็นบางอย่างเป็นผลของการแปลความหมายของข้อเท็จจริงขึ้นอยู่กับคุณสมบัติเฉพาะตัวของแต่ละคน เช่น พื้นความรู้ ประสบการณ์การทำงาน สภาพแวดล้อมและมีอารมณ์เป็นส่วนประกอบที่สำคัญ การแสดงความคิดเห็นอาจจะได้รับการยอมรับหรือปฏิเสธจากคนอื่นๆ ก็ได้ [13]

จากความหมายของความคิดเห็นสรุปได้ว่า ความคิดเห็นเป็นการแสดงออกด้านความรู้สึกต่อเหตุการณ์ ความคิดเห็นไม่มีถูกหรือผิดการแสดงออกนั้นออกมาจากอารมณ์ในขณะนั้นและความคิดเห็นสามารถเปลี่ยนแปลงไปตามกาลเวลา ณ ขณะนั้นได้

2.1.2 ปัจจัยที่ทำให้เกิดความคิดเห็น

ความคิดเห็นเป็นสิ่งที่เกิดจากการเรียนรู้มากกว่าเป็นสิ่งกำเนิดเองสิ่งแวดล้อมต่างๆ จึงมีอิทธิพลต่อความคิดเห็น ซึ่งได้แก่ ศาสนา ความเชื่อในสังคม ขนบธรรมเนียมประเพณีของสังคม สื่อมวลชนต่างๆ ดังนั้น ปัจจัยที่กำหนดความคิดเห็นของบุคคลจึงได้แก่ 1) การเรียนรู้ซึ่ง ได้แก่ การอบรมสั่งสอน อันจะเป็นการสะสมและรวบรวมประสบการณ์เอาไว้เป็นจำนวนมาก เช่น เด็กเกิดในครอบครัวที่นับถือศาสนาพุทธก็จะมีความเลื่อมใสในพุทธศาสนา เพราะได้รับอิทธิพลจากการอบรมสั่งสอนประสบการณ์ต่างๆ ไว้ 2) ประสบการณ์ส่วนตัวของบุคคลโดยทางตรง เช่น บุคคลที่เคยรับประทานอาหารทะเลแล้วแพ้ก็จะยอมมีความคิดเห็นที่ไม่ดีต่ออาหารทะเล 3) เหตุการณ์ประทับใจใน 2 ข้อข้างต้นที่กล่าวมานั้น เป็นการสะสมประสบการณ์หลายๆ ครั้ง และเกิดความคิดเห็น แต่ความคิดเห็นก็สามารถเกิดขึ้นได้หากได้รับเหตุการณ์เพียงครั้งเดียวและรู้สึกประทับใจ ซึ่งอาจจะประทับใจในทางบวกและลบได้ 4) การรับเอาแบบของความคิดเห็นของผู้อื่นมาเป็นของตน โดยยอมรับเอาความคิดเห็นของผู้ที่เหนือกว่ามาปฏิบัติต่อ เช่น รุ่นน้องรับความคิดเห็นบางเรื่องมาจากรุ่นพี่ 5) เกิดจากลักษณะบุคลิกภาพของแต่ละคน เช่น การมองคนในแง่ร้ายก็จะมีแนวโน้มในทางที่มีความคิดเห็นที่ไม่ดีต่อสิ่งต่างๆ อยู่เสมอ 6) เกิดจากอิทธิพลจากสื่อมวลชน สื่อมวลชนเป็นแหล่งให้ข้อมูลที่ก่อให้เกิดทั้งความเข้าใจ และอารมณ์ชักจูงไปสู่การปฏิบัติได้ [14]

ความสำคัญของความคิดเห็น การศึกษาเชิงสำรวจความคิดเห็นนั้นเป็นการศึกษาจากองค์ประกอบของความคิดเห็นของบุคคล โดยทั่วไป 3 ประการ คือ บุคคลที่ถูกวัดสิ่งเร้าและการตอบสนอง ซึ่งจะแสดงออกมาในระดับที่สูง ต่ำ มากหรือน้อย โดยทั่วไปแล้ววิธีวัดความคิดเห็นนั้นจะใช้การตอบแบบสอบถามและการสัมภาษณ์ ทั้งนี้เพื่อจะค้นหาความรู้สึกการตัดสินใจจากการประเมิน

ค่าหรือทัศนะเรื่องใดเรื่องหนึ่ง โดยเฉพาะของบุคคล กลุ่มบุคคลที่มีต่อสิ่งต่างๆ ดังนี้ การศึกษาวิจัยด้านความคิดเห็นจึงเป็นประโยชน์ต่อการวางแนวทางกำหนดนโยบาย การพัฒนาหน่วยงานต่างๆ ทำให้เกิดการเปลี่ยนแปลงไปในทิศทางที่ดีขึ้นตรงเป้าหมายและยังสามารถตอบสนองความต้องการของหน่วยงานหรือบุคคลในหน่วยงานนั้น โดยเฉพาะอย่างยิ่งหน่วยงานที่มีหน้าที่รับผิดชอบในการปลูกฝังอบรมให้ความรู้ สร้างสรรค์พัฒนาสติปัญญาของบุคคลด้วยแล้ว ยิ่งต้องตระหนักในการสำรวจเก็บข้อมูลความคิดเห็นความรู้สึกเหล่านี้ เพื่อเป็นข้อมูลย้อนกลับมาปรับปรุงเตรียมความพร้อมของหน่วยงานให้ทันสมัยทันต่อสถานการณ์อยู่เสมอ [15]

การวัดความคิดเห็น (Opinion Measurement) ทัศนคติ แรงจูงใจ และค่านิยมได้มีการสร้างแบบทดสอบสำหรับวัดสิ่งต่างๆ ดังกล่าว แต่ยังไม่สามารถแยกจากกันได้อย่างเด็ดขาดเพราะมีบางส่วนที่ซ้ำซ้อนกันอยู่ การวัดความคิดเห็นส่วนใหญ่แล้วยังไม่มีการแบ่งแยกออกจากทัศนคติอย่างชัดเจน และมีบ่อยครั้งที่คำทั้งสองใช้สลับกัน อย่างไรก็ตามการสำรวจความคิดเห็นมักจะเป็นการถามสิ่งที่เฉพาะเจาะจง เช่น การสำรวจความคิดเห็นเกี่ยวกับการเลือกตั้งสมาชิกสภาผู้แทนราษฎร การสำรวจความคิดเห็นของประชาชนต่อโครงการสาธารณะต่างๆ เป็นต้น ซึ่งผลที่ได้จากการสอบถามความคิดเห็นเหล่านี้จะเป็นตัวชี้วัดความพึงพอใจ ไม่พอใจ เห็นด้วยหรือไม่เห็นด้วยของกลุ่มประชากรดังกล่าว อีกทั้งหน้าที่ของความคิดเห็น โดยทั่วไปหน้าที่ของความคิดเห็นมีดังต่อไปนี้ 1) ทำหน้าที่เป็นแรงจูงใจให้บุคคลปรับตัวเมื่อมีความคิดเห็นที่ดีต่อสิ่งใดย่อมเข้าหาสิ่งนั้น และย่อมหลีกหนีสิ่งที่มีความคิดเห็นที่ไม่ดี 2) ทำหน้าที่ให้ค่านิยมหรือให้ความชื่นชอบต่อเนื่องไปถึงสิ่งอื่นๆ เช่น มีความคิดเห็นว่าการเปลี่ยนแปลงทางสังคมยังต้องใช้วิธีการสันติ ถ้าพบบุคคลที่มีแนวความคิดนี้จะนิยมชมชอบไปด้วย 3) ทำหน้าที่ช่วยให้ตีความหมายของสถานการณ์ต่างๆ ได้ เช่น ถ้ามีความคิดเห็นที่ดีต่อพ่อแม่บางครั้งพ่อแม่อาจจะขัดแย้งกับตน ก็จะตีความหมายไปว่าพ่อแม่ทำไปเพราะความหวังดี

ประโยชน์ของความคิดเห็นมีดังต่อไปนี้ 1) ช่วยทำให้เข้าใจสิ่งแวดล้อมรอบ ๆ ตัว โดยการจัดรูปหรือการจัดระบบสิ่งของต่างๆ ที่อยู่รอบตัว 2) ช่วยให้มี Self-Esteem โดยจะช่วยให้บุคคลหลีกเลี่ยงยิ่งสิ่งที่ไม่ดีหรือปกปิดความจริงบางอย่าง ซึ่งนำความไม่พอใจมาสู่ตัว 3) ช่วยในการปรับตัวให้เข้ากับสิ่งแวดล้อมที่สลับซับซ้อน ซึ่งมีปฏิกิริยาตอบโต้หรือกระทำสิ่งใดสิ่งหนึ่งออกไปนั้นส่วนมากจะนำความพึงพอใจมาให้ 4) ช่วยให้บุคคลสามารถแสดงออกในด้านค่านิยม ความรู้สึกของตนเอง อันจะนำความพอใจมาสู่บุคคลนั้นๆ จากการศึกษาข้างต้น ผู้ศึกษาพอสรุปแนวความคิดที่เกี่ยวกับการเกิดความคิดเห็น ประการแรกคือ เกิดจากประสบการณ์ที่บุคคลมีกับสิ่งของ บุคคล หรือสถานการณ์ และประการที่สองเกิดจากค่านิยม และการตัดสินค่านิยม ซึ่งแต่ละบุคคลมีค่านิยมและการตัดสินค่านิยมที่แตกต่างกันได้ย่อมทำให้มีความคิดเห็นที่แตกต่างกัน

2.1.3 การประยุกต์ใช้การวิเคราะห์ความคิดเห็น

ข้อความบนเครือข่ายสังคมออนไลน์แบ่งเป็น 3 ประเภท ได้แก่ ข้อเท็จจริง ความคิดเห็น และข้อเสนอแนะ โดย (1) ข้อเท็จจริงหมายถึง ข้อความหรือเหตุการณ์ที่เป็นมา หรือที่เป็นอยู่ตามจริง (2) ความคิดเห็นหมายถึงความเห็นหรือข้อวินิจฉัยหรือความเชื่อที่แสดงออกตามที่เห็น รู้ คิด และ (3) ข้อเสนอแนะหมายถึงข้อคิดเห็นเชิงแนะนำที่เสนอเพื่อพิจารณา ชี้แจงให้ทำหรือปฏิบัติตามดังตัวอย่าง “เมื่อคืนได้ดูรายการพื้นที่ชีวิต ไม่ชอบพิธีกรเลย อยากให้ปรับปรุงเรื่องการใช้ภาษาหน่อย” จากประโยคตัวอย่าง สามารถจำแนกประเภทของบทวิจารณ์ได้ ดังนี้ ข้อเท็จจริง : “เมื่อคืนได้ดูรายการพื้นที่ชีวิต” การวิเคราะห์เหมืองข้อความ แบ่งข้อมูลออกเป็น 2 รูปแบบคือ (1) ข้อมูลที่เป็นโครงสร้าง (Structured Data) ซึ่งการประมวลผลข้อมูลที่เป็นโครงสร้าง เรียกว่า การวิเคราะห์เหมืองข้อมูล และ (2) ข้อมูลที่ไม่เป็นโครงสร้าง (Unstructured Data) หรือไม่มีโครงสร้างที่ชัดแจ้ง (Implicit Structured Data) ซึ่งส่วนใหญ่มักอยู่ในรูปแบบของข้อความ หรือภาษาธรรมชาติ และเรียกกระบวนการวิเคราะห์ข้อความนี้ว่าการวิเคราะห์เหมืองข้อความ (Text Mining) โดยเป็นการนำความรู้ด้านการวิเคราะห์เหมืองข้อมูลมาประยุกต์ใช้กับหน่วยงาน Cross Industry Standard Process for Data Mining (CRISP DM) ได้นำเสนอกระบวนการวิเคราะห์เหมืองข้อมูล อันประกอบด้วยขั้นตอนดังต่อไปนี้ (1) การทำความเข้าใจกับธุรกิจและระบุปัญหาของงาน (Business Understanding) (2) การรวบรวมความหมายเหมือนกันแต่ใช้คำต่างกันจะแก้ไขให้เป็นคำเดียว เช่น คำ “คนอ่านข่าว” กับ “ผู้ประกาศข่าว” มีความหมายเหมือนกัน จึงแก้ไขเป็น “ผู้ประกาศข่าว” เป็นต้น (3) การแทนข้อความเป็นกระบวนการแทนข้อความให้อยู่ในรูปแบบที่มีโครงสร้างแน่นอน มีเทคนิคที่นิยมใช้คือ วิคเตอร์สเปซโมเดล (Vector Space Model : VSM) เป็นวิธีการหนึ่งในการแทนข้อความให้มีโครงสร้างแบบฟอร์มฟิวเจอร์ วิคเตอร์ (Feature Vector) เพื่อให้คอมพิวเตอร์สามารถนำไปประมวลผลต่อได้ โดยมีวิธีการแทนข้อความให้มีโครงสร้าง ดังนี้ การแทนข้อความด้วยถุงคำ (Bag-of-Word) เป็นวิธีการที่ง่ายและสะดวกที่สุด โดยแทนข้อความด้วยคำที่รวบรวมไว้แล้วในคลังคำ หากมีคำในถุงคำเกิดขึ้นในข้อความที่นำมาวิเคราะห์จะแทนค่าด้วย 1 แต่หากในข้อความไม่มีคำที่กำหนดไว้ในถุงคำเกิดขึ้น จะแทนค่าด้วย 0

2.2 แนวคิดเกี่ยวกับความรู้สึก

การใช้ภาษาโดยทั่วไปไม่ได้มีการแสดงความรู้สึกในทุกๆ ข้อความ ดังนั้นการวิเคราะห์ความรู้สึกจึงจำแนกออกเป็นงานวิจัยแขนงย่อยๆ เริ่มจากการจำแนกข้อความแสดงอัตวิสัย (Subjectivity Classification) โดยแบ่งประเภทของข้อความออกเป็นข้อความที่แสดงอัตวิสัย (Subjective) และข้อความที่บอกข้อเท็จจริง (Objective) จากนั้นจึงนำข้อความแสดงอัตวิสัยมาวิเคราะห์ความรู้สึกต่อไป การวิเคราะห์ความรู้สึกของข้อความนี้สามารถทำได้ 2 แนวทางคือ การวิเคราะห์แบบอิงคลังศัพท์ (Lexicon-Based) และการวิเคราะห์โดยอาศัยการเรียนรู้ด้วยเครื่อง (Machine Learning-Based)

การวิเคราะห์ความรู้สึกแบบอิงคลังศัพท์อาศัยการระบุคำที่บ่งบอกความรู้สึกในประโยค โดยอ้างอิงจากคลังศัพท์บอกความรู้สึก ซึ่งรวบรวมคำบอกความรู้สึกขั้วบวกและขั้วลบทั้งหมดเอาไว้ ส่วนการวิเคราะห์ความรู้สึกจากการเรียนรู้ด้วยเครื่องอาศัยการใช้ข้อมูลจำนวนมากเพื่อฝึก (Train) เครื่องจำแนก (Classifier) ให้เรียนรู้การจำแนกประเภทของข้อความ ออกเป็นข้อความที่แสดงความรู้สึกเชิงบวกและเชิงลบ วิธีการวิเคราะห์ความรู้สึกที่เป็นเป้าหมายของงานวิจัยนี้คือการวิเคราะห์แบบอิงคลังศัพท์ซึ่งมีแนวทางการการสร้างคลังศัพท์อยู่ 2 แบบคือ แบบใช้พจนานุกรม (Dictionary-Based) และแบบใช้คลังข้อมูลภาษา (Corpus-Based) การสร้างคลังศัพท์บอกความรู้สึก โดยใช้พจนานุกรมจะอาศัยการรวบรวมรายการคำบอกความรู้สึกจากความสัมพันธ์ของคำ ซึ่งระบุไว้ในพจนานุกรม เช่น กลุ่มคำไวพจน์ (Synonym) และปฏิพจน์ (Antonym) เป็นต้น ส่วนการสร้างคลังศัพท์บอกความรู้สึกโดยใช้คลังข้อมูลภาษาจะเก็บรวบรวมคำบอกความรู้สึก จากข้อมูลการใช้ภาษาจริงจำนวนมากที่เก็บรวบรวมไว้ในคลังข้อมูล ความแตกต่างระหว่างคลังศัพท์ที่ได้จากการสร้างคลังศัพท์ทั้ง 2 แนวทางคือ การใช้พจนานุกรมจะทำให้ได้คำบอกความรู้สึกครบถ้วนตามรายการคำในพจนานุกรม คำเหล่านี้จะมีความหมายเชิงบวกหรือเชิงลบ เช่น “ดี” และ “แย่” ซึ่งสามารถใช้วิเคราะห์ความรู้สึกของข้อความในโดเมน (Domain) ใดก็ได้ จึงนับเป็นประเภทคำบอกความรู้สึกแบบไม่เจาะจงโดเมน ส่วนการใช้คลังข้อมูลภาษาจะทำให้ได้คำบอกความรู้สึกแบบเจาะจงโดเมนเพิ่มเข้ามาในคลังศัพท์ด้วย โดยคำประเภทนี้จะมีขั้วความรู้สึกบวกหรือลบขึ้นอยู่กับโดเมนที่ปรากฏ แต่การสร้างคลังศัพท์จากคลังข้อมูลภาษาอาจได้รายการคำบอกความรู้สึกไม่ครบถ้วน ขึ้นอยู่กับการใช้ภาษาที่พบในคลังข้อมูลนั้นๆ คลังศัพท์บอกความรู้สึกเป็นหัวใจสำคัญของการวิเคราะห์ความรู้สึกแบบอิงคลังศัพท์ การสร้างคลังศัพท์นั้นมีได้ 2 วิธีการ คือ ใช้ข้อมูลความสัมพันธ์ระหว่างคำในพจนานุกรม เพื่อรวบรวมคำบอกความรู้สึกหรือสกัดเอาคำบอกความรู้สึกจากข้อมูลการใช้ภาษาจริงในคลังข้อมูลภาษา ข้อดีของการใช้พจนานุกรมคือได้คลังศัพท์ที่มีความครบถ้วนมากกว่า เนื่องจากคำศัพท์เกือบทั้งหมดในภาษาสามารถพบได้ในพจนานุกรม แต่มีข้อเสียคือกลุ่มคำบอกความรู้สึกที่ได้จะเป็นคำในประเภทไม่เจาะจงโดเมนเท่านั้น เนื่องจากคำในกลุ่มเจาะจงโดเมนจะมีความหมายเปลี่ยนไปตามบริบทของโดเมนที่คำ ๆ นั้นปรากฏ

ด้วยเหตุนี้ ความสามารถในการรวบรวมคำในกลุ่มเจาะจงโดเมน จึงเป็นข้อดีของการสร้างคลังศัพท์บอกความรู้สึกจากคลังข้อมูลภาษา ซึ่งในการใช้ภาษาจริงนั้นจะพบทั้งการใช้คำบอกความรู้สึกในกลุ่มเจาะจงโดเมนและไม่เจาะจงโดเมน แต่วิธีการนี้ก็มีข้อเสียคืออาจได้คลังศัพท์ที่ไม่ครบถ้วนขึ้นอยู่กับธรรมชาติของภาษาที่พบในคลังข้อมูลนั้นๆ

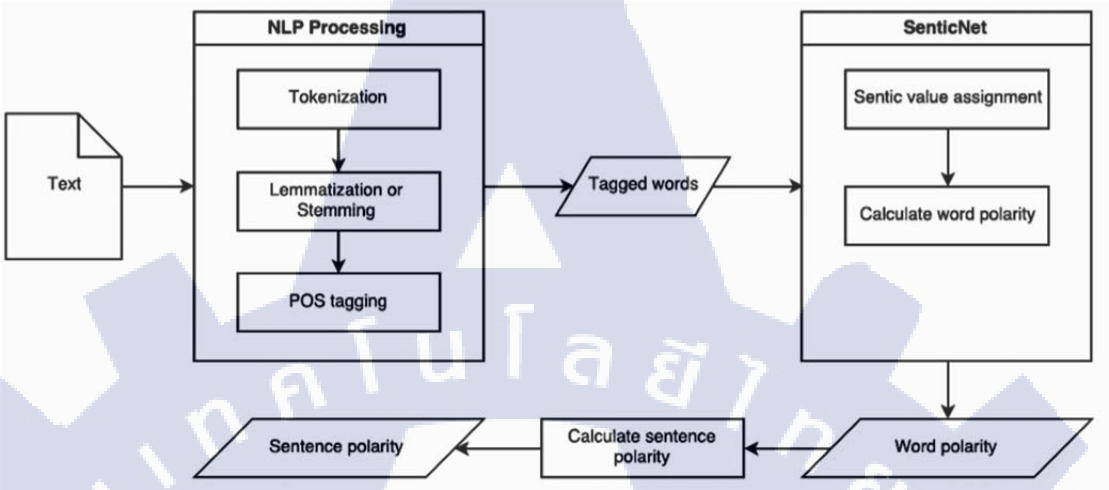
งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ความรู้สึกแบบอิงคลังศัพท์ในภาษาไทยนั้นมีจำนวนน้อย อาจเป็นเหตุมาจากการขาดแคลนแหล่งข้อมูลและเครื่องมือต่างๆ เช่น โปรแกรมแท็กชนิดของคำ เป็นต้น รวมไปถึงแนวโน้มของการเลือกใช้แบบจำลองในการวิเคราะห์ข้อมูล ซึ่งในปัจจุบันวิธีการเรียนรู้ด้วยเครื่องเป็นที่นิยมมากกว่า ตัวอย่างเช่น การจำแนกประโยคแสดงอัตวิสัยใน

งานวิจัยของ K. Sukhum et al. [16] และการวิเคราะห์ความรู้สึกของข้อความในงานวิจัย ซึ่งอาศัยการฝึกแบบจำลองการจำแนกด้วย Naive Bayes จากคลังข้อมูลที่ผ่านการระบุประเภทของข้อความแล้ว (Tagged Corpus) ส่วนงานวิจัยในภาษาไทยที่พบขั้นตอนการรวบรวมคำบอกความรู้สึกเพื่อสร้างคลังศัพท์ ได้แก่ ซึ่งทดลองรวบรวมคำบอกความรู้สึกและคำบอกลักษณะสินค้า ในบทวิจารณ์โรงแรมภาษาไทยจากเว็บ www.agoda.com โดยแบ่งข้อมูลออกเป็น 2 โดเมนย่อย ได้แก่ โดเมนอาหารเช้า จำนวน 301 บทวิจารณ์ และโดเมนการบริการ จำนวน 831 บทวิจารณ์ แล้วจึงนำมาระบุชนิดของคำบอกลักษณะสินค้า (FEA) และคำบอกความรู้สึก (POL) ด้วยตัวเอง เพื่อสร้างชุดข้อมูลฝึกฝนข้อมูลทั้งหมดในแต่ละโดเมนจะถูกแบ่งเป็นข้อมูลฝึกฝน 80% ข้อมูลทดสอบ 20% ผลที่ได้มีความถูกต้องในการดึง (คำบอกลักษณะสินค้า/คำบอกความรู้สึก) สำหรับโดเมนอาหารเช้าและการบริการเท่ากับ (80.00%/95.74%) และ (82.56%/89.29%) ตามลำดับ ด้วยเหตุนี้ การเลือกศึกษาแนวทางการวิเคราะห์ความรู้สึกแบบอิงคลังศัพท์ของงานวิจัยนี้จึงเป็นการสำรวจอีกแนวทางหนึ่ง ซึ่งอาจไม่เป็นที่นิยมแพร่หลายแต่ก็สามารถนำไปใช้วิเคราะห์ความรู้สึกของข้อความได้เช่นกัน และยังสามารถนำไปใช้พัฒนาเครื่องมือสำหรับการวิเคราะห์ความรู้สึกในภาษาไทยต่อไปในอนาคตได้อีกด้วย เนื่องจากผลที่ได้จากงานวิจัยนี้คือคลังศัพท์ที่มีรายการคำบอกความรู้สึก ทั้งแบบเจาะจงโดเมนและไม่เจาะจงโดเมน ซึ่งสามารถนำไปใช้ร่วมกับวิธีการเรียนรู้ด้วยเครื่องอื่นๆ เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลให้ดียิ่งขึ้นได้

2.2.1 การวิเคราะห์ความรู้สึก (Sentiment Analysis)

การวิเคราะห์ความรู้สึกคือ กระบวนการเชิงคำนวณเพื่อกำกับข้อความหรือส่วนของข้อความที่แสดงความคิดเห็นด้วยทัศนคติของผู้เขียนข้อความนั้นๆ ต่อสินค้าหรือหัวข้อหนึ่งๆ ซึ่งมักจะเป็นทัศนคติเชิงบวก เชิงลบ หรือเป็นกลาง อย่างไรก็ตามในการกำกับข้อความระดับคำสามารถใช้ทรัพยากร SenticNet ในการแทนค่าอารมณ์ Sentic Value 4 มิติคือ ความพึงพอใจ (Pleasantness) ความสนใจ (Attention) ความอ่อนไหว (Sensitivity) และ ความฉลาด (Aptitude) และเมื่อต้องการวิเคราะห์ความรู้สึกต้องใช้ SenticNet ในการวิเคราะห์ SenticNet คือ ทรัพยากรสำหรับการทำเหมืองความคิดเห็นที่สร้างขึ้น โดยใช้เทคนิค AI และเว็บเชิงความหมายเพื่อหาขั้วความคิดเห็นของ Concept ต่างๆ โดยประมวลผลข้อความภาษาธรรมชาติในระดับเชิงความหมาย ด้วย SenticNet จะต้องนำข้อความที่ต้องการวิเคราะห์ผ่านกระบวนการ NLP (Natural Language Processing) โดยต้องทำ Tokenization เพื่อแยกข้อความออกเป็นคำ ซึ่งในภาษาไทยต้องมีการตัดคำ (Word Segmentation) หลังจากนั้นนำมาทำ Lemmatization หรือ Stemming ซึ่งโดยปกติกระบวนการ NLP จะเลือกใช้ Stemming เพราะสามารถทำงานได้รวดเร็วกว่า จากนั้นนำมากำกับชนิดของคำจากข้อความ (Part Of Speech Tagging) อาทิเช่น คำนาม คำกริยา เป็นต้น หลังจากนั้นนำคำที่ผ่านกระบวนการ NLP

แล้วมาหาขั้วอารมณ์จาก Senticnet ซึ่งสามารถดาวน์โหลดได้ฟรี ทั้งแบบสแตนด์อโลนหรือแบบออนไลน์ผ่านทาง API โดย SenticNet จะค้นหาค่าความรู้สึกและค่าขั้วอารมณ์ดังรูปที่ 2.1



รูปที่ 2.1 การนำ SenticNet ไปใช้งาน

ตารางที่ 2.1 การเปรียบเทียบระหว่าง การวิเคราะห์ความคิดเห็น และการวิเคราะห์ความรู้สึก

การวิเคราะห์ความคิดเห็น (Opinion Analysis)	**การวิเคราะห์ความรู้สึก (Sentiment Analysis**
1. การรวบรวมความคิดเห็นจากหลายๆ ข้อความ ในเรื่องใดเรื่องหนึ่งโดยเฉพาะ เพื่อนำมาวิเคราะห์ความคิดเห็น ซึ่งมักจะวิเคราะห์เป็นเชิงบวก เชิงลบ หรือเชิงความเป็นจริงเท่านั้น	1. กระบวนการเชิงคำนวณเพื่อกำกับข้อความหรือส่วนของข้อความที่แสดงความคิดเห็นด้วยทัศนคติของผู้เขียนข้อความนั้นๆ ต่อสินค้าหรือหัวข้อหนึ่งๆ ซึ่งมักจะวิเคราะห์เป็นทัศนคติเชิงบวก เชิงลบ หรือเป็นกลาง
2. ใช้เทคนิควิคเตอร์สเปซโมเดล (Vector Space Model : VSM) เป็นวิธีการหนึ่งในการแทนข้อความให้มีโครงสร้างแบบฟอร์มฟิวเจอร์วิคเตอร์ (Feature Vector) ประมวลผลแทนข้อความด้วยถุงคำ (Bag-of-Word) เป็นวิธีการที่ง่ายและสะดวกที่สุด โดยแทนข้อความด้วยคำที่รวบรวมไว้แล้วในคลังคำ	2. ใช้เทคนิค AI และเว็บเชิงความหมาย เพื่อหาขั้วความคิดเห็นของ Concept ต่างๆ โดยประมวลผลข้อความภาษาธรรมชาติในระดับเชิงความหมายด้วย SenticNet นำข้อความวิเคราะห์ผ่านกระบวนการ NLP (Natural Language Processing) โดยใช้ Tokenization เพื่อแยกข้อความออกเป็นคำ ซึ่งในภาษาไทยต้องมีการตัดคำ (Word Segmentation)

ตารางที่ 2.1 การเปรียบเทียบระหว่าง การวิเคราะห์ความคิดเห็น และการวิเคราะห์ความรู้สึก (ต่อ)

การวิเคราะห์ความคิดเห็น (Opinion Analysis)	**การวิเคราะห์ความรู้สึก (Sentiment Analysis**
3. การแสดงความคิดเห็นจึงเป็นประโยชน์ต่อการวางแนวทาง กำหนดนโยบาย พัฒนาหน่วยงานต่างๆ ทำให้เกิดการเปลี่ยนแปลงไปในทิศทางที่ดีขึ้น	3. การบอกความรู้สึกคือคำที่ใช้วิจารณ์ลักษณะสินค้าหนึ่งๆ ในเชิงบวกหรือเชิงลบ และนำคำวิจารณ์มาปรับปรุงพัฒนาประสิทธิภาพในอนาคต
4. เครื่องมือที่ใช้สรุปเพื่อจะค้นหาความรู้สึกการตัดสินใจจากการประเมินค่า หรือทัศนะเรื่องใดเรื่องหนึ่ง	4. เครื่องมือที่ใช้สรุปความคิดเห็นผู้บริโภคมีต่อสินค้าหรือบริการต่างๆ ได้โดยอัตโนมัติ ซึ่งจะเป็นประโยชน์ต่อทั้งผู้ซื้อและผู้ขายสินค้านั้นๆ
5. การวิเคราะห์ความคิดเห็นเกี่ยวกับโรงแรมที่ใช้นิพจน์ปกติ (Regular Expression) เพื่อสร้างชุดข้อมูลการเรียนรู้ (Training Set)	5. การวิเคราะห์คำบอกความรู้สึกและคำบอกลักษณะสินค้าในบทวิจารณ์โรงแรมไทย (www.agoda.com) เพื่อสร้างคลังศัพท์ แล้วนำมาระบุชนิดของคำ เพื่อสร้างชุดข้อมูลฝึกฝน (Practice Set)
6. อาศัยการจำแนกด้วยอัลกอริทึมต่างๆ เช่น ต้นไม้การตัดสินใจ นาอีฟเบย์ เป็นต้น	6. อาศัยการจำแนกด้วยอัลกอริทึมต่างๆ เช่น ต้นไม้การตัดสินใจ นาอีฟเบย์ เป็นต้น

2.3 ลักษณะของภาษาไทย

ภาษาไทยมีลักษณะที่เป็นเอกลักษณ์เฉพาะตัวแตกต่างจากภาษาอื่นๆ ภาษาไทยจะมีเอกลักษณ์ลักษณะการเขียนเป็นคำและเป็นพยางค์ที่ติดกันทั้งประโยค คำในภาษาไทยจะประกอบไปด้วย พยัญชนะ สระ วรรณยุกต์ มาประกอบกันทำให้เกิดขึ้นเป็นคำที่มีความหมาย โดยที่แต่ละคำประกอบเข้าด้วยกันหลายๆ คำเรียกว่า “ประโยค” ซึ่งพยัญชนะไทยจะทำหน้าที่ขึ้นต้นเป็นตัวแรกของคำหรือพยางค์เสมอ และทำหน้าที่เป็นพยัญชนะตัวท้ายหรือตัวสะกด พยัญชนะไทยมีทั้งหมด 44 รูป ดังตารางที่ 2.2

ตารางที่ 2.2 พยัญชนะไทย 44 รูป

1	ก	12	ฌ	23	ท	34	ย
2	ข	13	ญ	24	ธ	35	ร
3	ฃ	14	ฎ	25	น	36	ล
4	ค	15	ฏ	26	บ	37	ว
5	ฅ	16	ฐ	27	ป	38	ศ
6	ฆ	17	ฑ	28	ผ	39	ษ
7	ง	18	ฒ	29	ฝ	40	ส
8	จ	19	ณ	30	พ	41	ห
9	ฉ	20	ด	31	ฟ	42	ฬ
10	ช	21	ต	32	ภ	43	อ
11	ซ	22	ถ	33	ม	44	ฮ

สระ เป็นส่วนประกอบอย่างหนึ่งที่ผสมอยู่ในคำหรือพยางค์ในภาษาไทย มี 21 รูป 32 เสียง มีความแตกต่างกับสระในภาษาอื่นๆ สระในภาษาไทยจะมีรูปสระมีชื่อเรียก ดังตารางที่ 2.3 และตารางที่ 2.4

ตารงที่ 2.3 รูปสระ 21 รูป

ที่	ชื่อสระ	รูปสระ	ที่	ชื่อสระ	รูปสระ	ที่	ชื่อสระ	รูปสระ
1	วิสรรชนีย์	ะ	8	ฟันหนู	็	15	ตัว ออ	อ
2	ไม้หันอากาศ	ั	9	ตีนเหยียด	ุ	16	ตัว ยอ	ย
3	ไม้ไต่คู้	็	10	ตีนคู้	ู	17	ตัว วอ	ว
4	ลากข้าง	า	11	ไม้หน้า	เ	18	ตัว รึ	ฤ
5	พินทุ์อิ	ิ	12	ไม้ม้วน	ใ	19	ตัว รือ	ฤๅ
6	ฝนทอง	่	12	ไม้มลาย	ไ	20	ตัว ลึ	ฦ
7	นฤคหิต	ำ	14	ไม้โอ	โ	21	ตัว ลือ	ฦๅ

ตารางที่ 2.4 รูปเสียง 32 เสียง

อะ	อา	อิ	อี	อึ	อื	อุ	อู
เอะ	เอ	แอะ	แอ	เอียะ	เอีย	เอือะ	เอือ
อัวะ	อัว	โอะ	โอ	เอาะ	ออ	เออะ	เออ
อำ	ใอ	ไอ	เอา	ฤ	ฤๅ	ฦ	ฦๅ

วรรณยุกต์ เป็นเครื่องหมายแทนระดับเสียงที่กำกับพยางค์ของคำในภาษาเสียงวรรณยุกต์หนึ่งอาจมีระดับเสียงต่ำ เสียงสูง เพียงอย่างเดียว หรือเป็นการทอดเสียงจากระดับเสียงหนึ่ง ไปยังอีกระดับเสียงหนึ่งและประโยชน์ของคำที่มีวรรณยุกต์จะช่วยให้คำมีความหมายเพิ่มมากขึ้น แตกต่างจากภาษาอื่น ๆ มีรูปวรรณยุกต์ดังตารางที่ 2.5

ตารางที่ 2.5 รูปวรรณยุกต์

ที่	**รูปวรรณยุกต์**	**เสียงวรรณยุกต์**
1	่	เอก
2	้	โท
3	๊	ตรี
4	๋	จัตวา

ตัวอย่างเช่น

“ปา” หมายถึง ขว้างปา โยน

“ป่า” หมายถึง ที่มีต้นไม้ภูเขา และสัตว์

“ป้า” หมายถึง พี่ของพ่อหรือแม่

“ปา” หมายถึง พ่อในภาษาบางภาษา

“ป๋า” หมายถึง พ่อในภาษาบางภาษา

ซึ่งวรรณยุกต์ทั้งหมดมี 4 รูป 5 เสียง ดังนี้

เสียงสามัญ คือ ระดับเสียงปานกลาง มีระดับเสียงคงที่ เช่น กิน ทอง ลืม ยำ

เสียงเอก คือ ระดับเสียงต่ำสุด มีเสียงคงที่สม่ำเสมอ เช่น เก่ง ผ่า หนัก สุข

เสียงโท คือ ระดับเสียงตอนต้นเป็นเสียงสูงปลายเสียงเปลี่ยนเป็นระดับเสียงต่ำ เช่น ค่า ช่าง มาก ใกล้ ข้า นาบ ลาก พราก

เสียงตรี คือ ระดับเสียงสูง เช่น โต๊ะ จ๊ะ ก๊ก รัก คิด น้อง ไว้ ใช้

เสียงจัตวา คือ ระดับเสียงตอนต้นเสียงต่ำตอนปลายเสียงเน้นระดับเสียงที่สูงขึ้น เช่น แจ๋ว ก๋วยเตี๋ยว เหลือ เขียว ก๋า ขา หมา หลิว สวย หาม ปิ๋ว จิ๋ว

2.3.1 ชนิดของคำในภาษาไทย

คำในภาษาไทยแยกเป็นชนิดต่างๆ ได้ดังนี้

2.3.1.1 คำนาม คือ คำที่ใช้เรียกชื่อคน สัตว์ สิ่งของ สถานที่ รวมทั้งสิ่งมีชีวิตและสิ่งไม่มีชีวิต ทั้งที่เป็นรูปธรรมและนามธรรม เช่น เด็ก พ่อ แม่ นก ช้าง วัว ควาย โรงแรม บ้าน โรงเรียน ความดี ความรัก ฯลฯ

2.3.1.2 คำสรรพนาม คือ คำที่ใช้แทนคำนามที่ผู้พูดหรือผู้เขียนได้กล่าวแล้วหรือเป็นที่เข้าใจกันระหว่างผู้ฟังและผู้พูดเพื่อไม่ต้องกล่าวคำนามซ้ำ เช่น ผม ฉัน ดิฉัน ข้าพเจ้า หนู

2.3.1.3 คำกิริยา คือ คำที่แสดงอาการ สภาพ หรือการกระทำของคำนาม คำกิริยาบางคำอาจจะมีความหมายที่สมบูรณ์ในตัวเอง บางคำต้องมีคำอื่นมาประกอบและบางคำต้องไปประกอยคำอื่นเพื่อขยายความหมาย เช่น น้องยืนบนโต๊ะทำงาน ร้อนวันนี้ร้อนมาก เป็นต้น

2.3.1.4 คำวิเศษณ์ คือ คำที่ใช้ขยายคำนาม คำสรรพนาม คำกริยา และคำวิเศษณ์เพื่อให้ได้ความชัดเจนยิ่งขึ้น เช่น ชนิด ขนาด เสียง สี กลิ่น รส อาการ เป็นต้น

2.3.1.5 คำบุพบท คือ คำที่ทำหน้าที่เชื่อมโยงคำหนึ่งหรือกลุ่มคำหนึ่งให้สัมพันธ์กับคำอื่น หรือกลุ่มคำอื่น เพื่อแสดงความหมายต่างๆ เช่น ความเป็นเจ้าของ ลักษณะ เหตุผล เวลา สถานที่ ประมาณ ความต้องการ เปรียบเทียบ ฯลฯ ได้แก่ คำ ใน แก่ แต่ ต่อ สำหรับ โดย ด้วย ของ แห่ง ใกล้ ไกล ฯลฯ

2.3.1.6 คำสันธาน คือ คำที่ทำหน้าที่เชื่อมคำกับคำ ประโยคกับประโยคข้อความกับข้อความ หรือ ความให้สละสลวย ได้แก่ คำ และ หรือ แต่ เพราะ ฯลฯ

2.3.1.7 คำอุทาน คือ คำที่เปล่งออกมาเพื่อแสดงอารมณ์ ความรู้สึกของผู้พูด ซึ่งอาจเปล่งออกมาในขณะที่ตกใจ ดีใจ เสียใจ ประหลาดใจ คำอุทานส่วนมากไม่มีความหมายตรงตามถ้อยคำแต่จะมีความหมายเน้นความรู้สึกและอารมณ์ของผู้พูดเป็นสำคัญ คำอุทานแบ่งเป็น 3 ลักษณะ

2.3.1.7.1 เป็นคำ เช่น โอ๊ย โถ ว้าย เป็นต้น

2.3.1.7.2 เป็นวลี เช่น คุณพระช่วย ตายล่ะ พุทโธ่เอ๋ย เป็นต้น

2.3.1.7.3 เป็นประโยค เช่น เฮ้ยป้าถูกรถชน ตายล่ะลืมจ่ายค่าไฟ เป็นต้น

2.3.2 ประโยคในภาษาไทย

ประโยค คือ ถ้อยคำที่มีความเกี่ยวข้องกันถูกต้องตามระเบียบของภาษาและมีเนื้อความบริบูรณ์ ประกอบด้วยภาคประธานและภาคแสดง ประโยคในภาษาไทยแบ่งตามลักษณะเป็น 3 ชนิดดังนี้

2.3.2.1 ประโยคความเดียว (เอกรรถประโยค) คือ ประโยคที่มีใจความสำคัญเพียงหนึ่ง คือ มีประธานตัวเดียว กริยาสำคัญเพียงตัวเดียวและอาจมีกรรมหรือไม่มีก็ได้ โดยมีโครงสร้างประธาน+กริยา+(บทกรรม) เช่น เราเดินอย่างรวดเร็ว เจ้าหน้าที่ซื้อปลาจากแหล่งเพาะพันธุ์ การเดินเป็นการออกกำลังกาย โปรดนั่งเงียบ เมื่อคืนนี้พายุพัดบ้านพัง

2.3.2.2 ประโยคความรวม (อเนกรรถประโยค) คือ ประโยคที่เกิดจากประโยคความเดียว ตั้งแต่ 2 ประโยคขึ้นไปมารวมกันโดยมีสันธานเป็นตัวเชื่อม โดยมีโครงสร้าง ประโยคความเดียว+สันธาน+ประโยคความเดียว

2.3.2.3 ประโยคความซ้อน (สังกรประโยค) คือประโยคที่ประกอบด้วยประโยคหลักและประโยคย่อยซึ่งเป็นประโยคซ้อนอยู่ในประโยคหลัก โดยมีสันธานเป็นตัวเชื่อม องค์ประกอบของประโยคความซ้อนมี 2 ชนิด คือ ประโยคหลัก (มุขยประโยค) คือ ประโยคที่เป็นใจความสำคัญของประโยคใหญ่ และประโยคย่อย (อนุประโยค) คือ ประโยคที่ทำหน้าที่ช่วยประโยคหลักให้มีใจความชัดเจนสมบูรณ์มากขึ้น

2.4 แนวคิดและทฤษฎีเกี่ยวกับการตัดคำภาษาไทย

การตัดคำได้รับการพัฒนามาเป็นเวลาพอสมควร ทำให้เกิดแนวคิดเกี่ยวกับการตัดคำขึ้นหลากหลายเทคนิค การตัดคำแบ่งออกเป็นลักษณะใหญ่ๆ ได้ ดังนี้ พยางค์ ได้แก่ ซึ่งใช้กฎในการผสมกันของพยางค์ แบ่งตัวอักษรออกเป็น 5 กลุ่มคือ กลุ่มพยัญชนะ สระ วรรณยุกต์ ตัวเลข และอักขระพิเศษ โดยใช้อักษรดังนี้ แทนแต่ละกลุ่ม

C แทน พยัญชนะต้น

V แทน สระ

S แทน ตัวสะกด

T แทน วรรณยุกต์

G แทน การันต์

ยกตัวอย่าง เช่น สิ้น CTVS, ศิลป CVSSG, กวาง CCVS กฎนี้ไม่ยุ่งยากซับซ้อนมากนักจึงไม่มีความยืดหยุ่นเท่าที่ควร อีกทั้ง อักษรไทยบางตัวสามารถเป็นได้ทั้งตัวสะกด และพยัญชนะต้น นั่นคือเป็นได้ทั้ง C และ S ทำให้การตัดคำบางคำเป็นไปอย่างไม่ถูกต้องเท่าที่ควร

การตัดพยางค์ การตัดพยางค์เป็นการใช้หลักการของภาษาไทยที่มีกฎเกณฑ์ค่อนข้างตายตัว ยกเว้นบางพยางค์งานวิจัยเกี่ยวข้องกับการตัดคำนิยมใช้พจนานุกรมเข้ามาช่วย ได้แก่ งานวิจัยของ ยืน ภู่วรวรรณ และวิวรรธ์ อิ่มอารมณ์ [17] เนื่องจากพจนานุกรมสามารถตัดคำได้ชัดเจนกว่าการตัดพยางค์ เพราะคำอาจเกิดจากการมารวมกันของหลายพยางค์แต่ขอบเขตของคำยังคงซับซ้อนและกำกวม ในงานวิจัยได้ใช้เทคนิคที่เรียกว่าวิธีการย้อนกลับ (Back Tracking) และการเลือกคำที่ยาวที่สุด (Longest Matching) วิธีการนี้มีข้อดีคือย้อนกลับไปเพื่อเลือกคำอีกครั้งได้ แต่ข้อเสียคือเมื่อหากคำนั้นได้ แต่ข้อเสียคือเมื่อหากคำนั้นเมื่อย้อนกลับไปแล้วยังไม่พบตามพจนานุกรมทำให้เสียเวลา

2.4.1 การตัดคำภาษาไทย

การตัดคำภาษาไทย (Thai Word Segmentation) เป็นกระบวนการพื้นฐานของการประมวลผลภาษาธรรมชาติการตัดคำภาษาไทยนั้นมีวิธีที่จะแยกคำแต่ละคำออกจากกัน เนื่องจากทำให้สะดวกต่อการตรวจสอบขอบเขตของคำแต่อยู่ในประโยคได้ง่ายขึ้น เหตุผลที่ต้องแยกคำออกจากประโยคก็เพราะว่าในประโยคของภาษาไทยจะไม่เหมือนกับภาษาอังกฤษที่มีสัญลักษณ์เครื่องหมายเข้ามาเป็นตัวบ่งบอกว่าจบประโยค เช่น การเว้นวรรค และจุดทศนิยม ดังนั้นจึงทำได้มีเทคนิคและวิธีการที่นำไปใช้ในการตัดคำภาษาไทย โดยมีการพัฒนาจากหน่วยงานวิจัยต่างๆ ทั้งของภาครัฐและของภาคเอกชน ได้มีแนวคิดและวิธีการต่างๆ เพื่อทำมาใช้ให้เกิดประโยชน์และถูกต้องมากที่สุด โดยแต่ละวิธีการต่างก็ให้ผลในด้านของความถูกต้องของหลักการทำงาน ความรวดเร็วในการทำงาน รวมถึงทรัพยากรของระบบ วิธีการตัดคำภาษาไทยสามารถแบ่งได้ดังนี้ [18]

2.4.1.1 หลักการตัดคำโดยใช้กฎ (Rule-based) การตักคำโดยวิธีการใช้กฎ เป็นวิธีการตัดคำในแบ่งแรก ๆ ที่มีการถูกคิดค้นขึ้นมาเพื่อใช้ตัดคำในภาษาไทย โดยการตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่กำหนดลักษณะที่มีการผสมอักษรลักษณะการเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้ในการกำหนดกฎเกณฑ์และกำหนดขอบเขตของคำให้มีความทุกต้อง ตัวอย่างเช่น

1) การขึ้นย่อหน้าใหม่เป็นตัวบอกถึงการสิ้นสุดข้อความ หรือประโยค

2) การเว้นวรรคเป็นตัวบอกถึง ความเป็นไปได้ที่จะเป็นการสิ้นสุดคำ หรือสิ้นสุดประโยค

3) กฎทางอักขระวิธีนี้เป็นตัวชี้ให้เห็นถึงความเป็นไปได้ของการตัดคำในตำแหน่งนั้นๆ

2.4.1.2 หลักการตัดคำโดยใช้พจนานุกรม (Dictionary) การตัดคำโดยใช้วิธีพจนานุกรมเป็นการพัฒนามาจากวิธีการตัดคำโดยใช้กฎในการตัดคำ โดยวิธีนี้อักขระหรือข้อมูลที่เป็นคำภาษาไทยจะถูกเก็บไว้ในพจนานุกรมก่อนเพื่อไม่ให้เกิดข้อผิดพลาดในการทำงานของระบบ เมื่อต้องการใช้งานก็

รับคำเข้ามาแล้วนำไปค้นหาเปรียบเทียบในสายอักขระกับคำที่เก็บไว้ในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวที่รับเข้ามาควรตักคำในช่วงบริเวณใดและประกอบไปด้วยคำใดบ้าง

2.4.1.3 หลักการตัดคำโดยใช้คลังข้อความ (Text) การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำโดยนำวิธีการทางสถิติเข้ามาใช้ในการประมวลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลแบ่งออกเป็น 2 วิธี คือการตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic Word Segmentation) และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (Feature-Based Word Segmentation) วิธีการตัดคำโดยอาศัยค่าความน่าจะเป็นจะเป็นการตัดคำโดยใช้แบบจำลองเอนแกรม (Word n-Gram Model) ในการหารูปแบบของการตัดคำและลำดับคำที่เป็นไปได้มากที่สุด โดยวิธีการนี้จะต้องมีการใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำที่เตรียมเอาไว้แล้ว ซึ่งวิธีการนี้ผลลัพธ์ที่ได้จะเป็นการเลือกรูปแบบการตัดคำที่มีความน่าจะเป็นมากที่สุด

<u>ตัวอย่างของแบบจำลองไตรแกรม</u>

“การพัฒนาระบบถาม-ตอบ” จะได้ว่า

การ/ ารพ/รพั /พัฒ/ ัฒน/ ฒนา/ นาร/ าระ/ระบ/บบถ/บถา/ถาม/ ามต/มตอ/ตอบ

หลังจากนั้นจะทำการเลือกคำที่เป็นไปได้เพื่อทำการประมวลผลต่อไปอย่างไร

วิธีการตัดคำโดยอาศัยคุณลักษณะของคำ จะเป็นการแก้ข้อผิดพลาดของการตัดคำโดยอาศัยค่าความน่าจะเป็นของการจำกัดหมวดคำที่จะเป็นแบบจำลองในการตัดคำ ซึ่งวิธีการตัดคำโดยอาศัยคุณลักษณะของคำจะเป็นวิธีการแบบผสม (Hybrid Approach)

<u>2.4.2 ปัญหาการตัดคำในภาษาไทย</u>

การตัดคำภาษาไทยที่มีประสิทธิภาพ หมายถึง การตัดให้ได้หน่วยคำและความหมายที่ถูกต้องอย่างไรก็ตาม ปัญหาที่ทำให้ลดประสิทธิภาพการตัดคำภาษาไทยมีสาเหตุ 2 ประการคือ

2.4.2.1 คำไม่รู้จัก (ชื่อเฉพาะ, คำยืมภาษาต่างประเทศ) คำไม่รู้จักสามารถแบ่งย่อยตามรูปแบบการเกิดได้ 3 แบบ ดังนี้

2.4.2.1.1 แบบชัดเจน เช่น คำว่า “โลตัส” หรือ “สุวิทย์” เป็นต้น

2.4.2.1.2 แบบซ้อนเร้นบางส่วน หมายถึง มีบางส่วนของคำเป็นคำที่มีในพจนานุกรม เช่น คำว่า “แอนติเจน” หรือ “โคโดโมะ” ซึ่งมีคำว่า “ติ” และ “โค”

2.4.2.1.3 แบบซ้อนเร้นทั้งหมด หมายถึง ทุกส่วนของคำเป็นคำที่มีในพจนานุกรม เช่น คำว่า “น้ำดอกไม้” ซึ่งทั้งคำว่า “น้ำ” “ดอก” และ “ไม้” เป็นคำที่มีในพจนานุกรม

ความคลุมเครือในการแบ่งขอบเขตคำ หมายถึง สายอักษรหนึ่งสามารถแบ่งเป็นคำที่มีในพจนานุกรมได้มากกว่า 1 รูปแบบ สามารถแบ่งได้เป็น 2 แบบ ตามลักษณะการเกิด

1) ระดับตัวอักษร ตัวอย่างคือ “มากว่า” ซึ่งตัวอักษร “ก” สามารถรวมได้ 2 แบบ คือ “มาก-ว่า” หรือ “มา-กว่า” 2) ระดับคำ ตัวอย่างคือ “ทางการตลาด” ซึ่งคำว่า “การ” สามารถรวมได้ 2 แบบ คือ “ทางการ-ตลาด” หรือ “ทาง-การตลาด

2.4.2.2 การตัดคำโดยใช้เทคนิคการเรียนรู้จากคลังประโยคโดยไม่ใช้คนเตรียมภาพรวมของระบบ การทำงานของระบบ แบ่งได้เป็น 2 ส่วนคือ การหาขอบเขตคำมูลและการรวมคำมูลให้เป็นหน่วยคำที่ถูกต้อง การตัดคำภาษาไทยเป็นงานพื้นฐานของ Thai NLP ที่ทำกันมาตั้งแต่แรกเริ่มของการใช้ภาษาไทยบนคอมพิวเตอร์ แรกเริ่มเป็นการทำเพื่อช่วยงานพิมพ์ในการปัดข้อความขึ้นบรรทัดใหม่ให้ถูกต้อง ไม่ตัดแยกส่วนคำระหว่างบรรทัดงานแรกๆ ออกมาในรูปของการใช้กฎ เพื่อระบุขอบเขตว่าควรจะตัดที่ไหนได้ กฎต่างๆ ที่เสนอเป็นเรื่องของพยางค์มากกว่าตัดคำ เช่น ถ้าพบ “เ” จะตัดหลังสระนี้ไม่ได้ต้องมีพยัญชนะตามมาเสมอ จากนั้นจึงมีแนวคิดที่ใช้พจนานุกรมช่วยตัด โดยคิดว่าจะต้องมีรายการคำเพื่อให้คอมพิวเตอร์รู้ว่ารูปใดเป็นคำและขอบเขตคำอยู่ที่ไหน แนวคิดพื้นฐานคือเอาคำในพจนานุกรมไปเทียบกับข้อมูลหากพบรูปตรงก็จะตัดได้ ซึ่งก็ทำแบบ Longest Matching ได้การตัดคำให้ยาวที่สุดก่อน แต่ก็จะเห็นว่ามีปัญหา เช่น “ไปหามเหสี” ตัดออกมาเป็น ไป|หาม|เห|สี เลยมีการเสนออีกแนวคิดให้ตัดแบบ Maximum Matching [19]

ตัดคำให้ได้จำนวนน้อยที่สุด ตัวอย่างนี้จึงได้ “ไป|หา|มเหสี” เพราะมีจำนวนคำที่ได้น้อยกว่า การตัดคำที่ได้จำนวนคำมากที่สุด คาดว่าเป็นเพราะภาษาไทยมีการสร้างคำจากการประสมคำจำนวนมาก เมื่อเจอโอกาสที่จะรวมเป็นคำเดียวจะมากกว่าที่จะเป็นสองคำ การเลือกผลที่มีจำนวนคำน้อยสุดจึงมีโอกาสถูกต้องมากกว่า ข้อเสียของการตัดคำด้วยพจนานุกรม คือ ถ้าพบรูปคำที่ไม่รู้จักก็จะตัดคำไม่ได้เป็นปัญหาของ Unknown Word ซึ่งปัญหานี้อาจมาจากพจนานุกรมมีรายการคำไม่ครอบคลุมพอ หรือเป็นคำที่เกิดใหม่ได้เป็นชื่อต่างๆ เป็นคำทับศัพท์ หรือเป็นคำที่สะกดผิดเป็นต้น นอกจากปัญหา Unknown Word ปัญหาพื้นฐานเรื่องความกำกวมก็ยังคงอยู่ โดยเฉพาะเมื่อตัดแล้วได้จำนวนคำเท่ากัน ตัวอย่างเช่น ตาก|ลม กับ ตา|กลม นอกจากการตัดคำแบบเลือกให้มีคำน้อยสุดแล้วก็มีการใช้วิธีการทางสถิติคือใช้ N-Gram ของคำมาช่วยในการแก้ปัญหาความกำกวมในการตัด [20] เช่น มีข้อมูลจริงว่าพบ นั่ง|ตาก|ลม|อยู่ หรือ ทำ|ตา|กลม|แป้ว ก็จะมีบริบทที่ช่วยตัดสินว่าจะเลือกตัดแบบไหนดีกว่า ซึ่งจะทำแบบนี้ได้ก็จะต้องมีการตัดคำด้วยมือก่อนให้มีจำนวนตัวอย่างของ N-Gram ที่มากพอ และเมื่อมีคลังข้อมูลที่ตัดคำเป็นตัวอย่างแล้ว ก็เริ่มมีการใช้ข้อมูลอื่นนอกเหนือจากสถิติจาก N-Gram เช่น ใช้ Feature-Based ตัดคำโดยเรียนรู้จากลักษณ์อื่นๆ ในบริบท เพื่อเลือกการตัดคำที่ดีสุดจากทุกแบบที่ได้จากการใช้ Maximum Matching มาก่อน นอกจากการตัดคำแบบอาศัยพจนานุกรม บางคนก็ใช้วิธีการอื่นเพื่อเลี่ยงปัญหา Unknown Word โดยไม่อิงพจนานุกรมเลย เช่น ใช้ TCC (Thai Character Cluster) เป็นตัวแยกหน่วยย่อยก่อนจะตัดสินใจว่าขอบเขตคำอยู่ที่ไหน TCC อาศัยแนวคิดว่ามีตัวอักษรบางตัวที่รู้ว่าไม่ควรไปตัดคำตรงนั้น เช่น C1 จะไม่ตัดหน้าสระอาแน่ๆ

C1 ก็เลยเป็น Cluster หนึ่ง ส่วนการตัดสินว่า TCC ไหนรวมเป็นคำได้หรือจะตัดคำที่ TCC ไหน ก็อาจใช้วิธีเรียนรู้จากคลังข้อมูลที่ตัดคำไว้ให้ เช่น ใช้ Decision Tree [21] การแยกหน่วยย่อยก่อนการตัดคำ

นอกจากการใช้ TCC ก็มีการใช้รูปพยางค์เป็นหน่วยย่อยด้วย เช่น เลือกตัดพยางค์ก่อนด้วยกฎโครงสร้างพยางค์ ซึ่งก็แม้แต่การตัดพยางค์ยังมีความเป็นไปได้หลายแบบ จึงต้องใช้ N-Gram ของคลังข้อมูลที่ตัดพยางค์แล้วช่วยเลือกการตัดพยางค์ที่ดีสุด ก่อนจะไปตัดสินใจรวบพยางค์เป็นคำภายหลัง การใช้คลังข้อมูลที่ตัดพยางค์ด้วยมือมีข้อดีกว่าการใช้คลังข้อมูลที่ตัดคำด้วยมือ เพราะการตัดพยางค์ไม่มีปัญหาเรื่องการตัดแล้วไม่คงที่สม่ำเสมอ และจำนวนข้อมูลที่เท่ากันก็จะได้จำนวน Token ของพยางค์ที่มากกว่าคำ งานตัดคำหลังๆ ก็อาศัยคลังข้อมูลภาษาที่มีการตัดคำแล้วและใช้การเรียนรู้ด้วยเครื่องแบบต่างๆ เช่น ใช้ CRF, SVM ล่าสุดก็นิยมใช้ Deep Learning เช่น KutCum, Deep Cut แต่ทั้งหมดก็ต้องอาศัยคลังข้อมูลที่มีการตัดคำเป็นตัวอย่างให้เรียนก่อน วิธีการพวกนี้จึงไม่จำเป็นต้องใช้พจนานุกรมได้ เพราะสามารถเริ่มจากหน่วยเล็กสุดคือ ตัวอักษร หรือกลุ่มตัวอักษรก็ได้ หรือจะใช้รูปพยางค์ก็เป็นไปได้ปัญหาของการใช้คลังข้อมูลที่มีการตัดคำ นอกจากต้องลงแรงในการตัดคำเองจำนวนมากแล้ว ยังมีปัญหาเรื่องแต่ละคนตัดคำไม่เหมือนกัน หรือแม้คนๆ เดียวกัน ก็อาจตัดคำไม่เสมอกัน ทำให้ข้อมูลไม่คงที่แบบที่คาดหวัง ปัญหาตัดคำไม่เหมือนกันไม่เสมอกันนี้ ทำให้ต้องมาตอบคำถามว่าคำคืออะไร ทำไมแต่ละคนจึงกำหนดขอบเขตคำไม่เหมือนกัน และการตัดคำทำไปเพื่ออะไร ทำอย่างไรจึงจะสร้างคลังข้อมูลขนาดใหญ่ที่มีการตัดคำได้ [22]

2.5 วิธีที่ใช้ในการตัดคำ

วิธีที่ใช้ตัดคำแบ่งออกเป็น 3 ประเภทใหญ่ๆ ได้แก่ การใช้กฎ การใช้พจนานุกรม และการใช้คลังข้อความ

1) การใช้กฎ การตัดคำโดยการตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่กำหนดลักษณะการประสมอักษร ลักษณะการเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ วิธีการนี้จะมีข้อจำกัดในการทำงานคือ ความถูกต้องของการตัดคำในระดับพยางค์สูงแต่ความถูกต้องของการตัดคำค่อนข้างต่ำแต่ข้อดีของวิธีนี้คือ มีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย

2) การใช้พจนานุกรม การตัดคำโดยพจนานุกรมเป็นการตัดคำ โดยใช้สายอักขระ (String) มาเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม ซึ่งวิธีนี้จะต้องทำการจัดเก็บคำไว้ในพจนานุกรม วิธีนี้ทำให้ได้ความถูกต้องในการตัดคำสูงกว่าการใช้กฎแต่จะใช้เวลามากกว่า

3) การใช้คลังข้อความ การตัดคำโดยใช้คลังข้อมูลเป็นการตัดคำ โดยนำวิธีการทางสถิติเข้ามาใช้ในการประมวลภาษา โดยใช้คลังข้อมูลทางภาษาเป็นฐานความรู้เก็บค่าความถี่ที่ใช้ในการตัดคำ ซึ่งการตัดคำโดยใช้คลังข้อมูลแบ่งออกเป็น 2 วิธีคือ การตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic

Word Segmentation) และวิธีการตัดคำโดยอาศัยคุณลักษณะของคำ (Feature-Based Word Segmentation) วิธีการตัดคำในการหารูปแบบของการตัดคำและลำดับคำที่เป็นไปได้มากที่สุด โดยวิธีการนี้จะต้องมีการใช้คลังข้อมูลที่มีการตัดคำและกำกับหมวดคำที่เตรียมเอาไว้แล้ว ซึ่งวิธีการนี้ผลลัพธ์ที่ได้จะเป็นการเลือกรูปแบบการตัดคำที่มีความน่าจะเป็นมากที่สุด

ตารางที่ 2.6 การเปรียบเทียบการตัดคำของแต่ละวิธี

วิธีการใช้	ข้อดี	ข้อเสีย	ประสิทธิภาพ
กฎ	1. มีความรวดเร็วในการตัดคำ 2. ไม่ต้องการข้อมูลในหน่วยความจำของคอมพิวเตอร์	1. ทำได้ในระดับพยางค์ 2. ยากที่จะสร้างกฎไวยากรณ์ได้สมบูรณ์ 3. ไม่สามารถจัดการคำ ที่เป็นคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words)	70.04%
พจนานุกรม	1. มีความแม่นยำสูงในการตัดคำ 2. สามารถแก้ไขปัญหาคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words) โดยการเพิ่มเข้าไปในพจนานุกรม	1. ต้องการหน่วยความจำมากเพื่อจัดการกับการเก็บพจนานุกรม 2. บริหารจัดการยากที่จะนำคำที่เป็นคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words) เข้าไปในพจนานุกรม	77.25%
คลังข้อความ	1. ไม่จำเป็นต้องเก็บพจนานุกรมในหน่วยความจำ 2. แก้ไขปัญหาคำที่คอมพิวเตอร์ไม่รู้จัก (Unknown Words)ได้บางกรณี โดยการเรียนรู้จากคลังข้อความ 3. สร้างกฎไวยากรณ์ได้จากการเรียนรู้จากคลังข้อความ	1. ต้องการคลังข้อความที่มีขนาดใหญ่และเวลาในการเรียนรู้และต้องการอัลกอริทึม (Algorithm) ที่ดีในการเรียนรู้	68.77%

2.5.1 เทคนิคการตัดคำ

เทคนิคที่ใช้ในการตัดคำที่นิยมใช้กันทั่วไปมี 5 วิธี ซึ่งมีวิธีการใช้นำไปใช้แตกต่างกันขึ้นอยู่กับลักษณะของคำ เพื่อนำไปใช้การตัดคำให้มีประสิทธิภาพที่ดีที่สุด

1) วิธีการเทียบคำที่ยาวที่สุด (Longest Word Pattern Matching)

วิธีนี้จะทำการตรวจสอบสายอักขระ (String) ที่นำเข้ามาจากซ้ายไปขวา จากนั้นนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรม หากตรวจสอบพบว่า พยางค์มากกว่า 1 พยางค์ ในพจนานุกรมจะทำการเลือกพยางค์ที่ยาวที่สุดแล้วทำต่อไปเรื่อยๆ จนจบสายอักขระตัวอย่างคำว่า “กองกลาง” การตัดคำโดยวิธีนี้จะนำสายอักขระไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า ก, กอ และคำว่า กอง ส่วนคำว่า กองกไม่พบอยู่ในพจนานุกรม ดังนั้น จึงได้คำว่า กอง ซึ่งเป็นคำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือคือ กลาง เมื่อนำไปค้นในพจนานุกรมจะได้ว่า ก, กล, กลาง ดังนั้น จึงเลือกคำว่า กลาง คำที่ได้จากการตัดคำโดยวิธีการนี้จึงเป็น กอง|กลาง

2) วิธีการเทียบคำที่สั้นที่สุด (Shortest Word Pattern Matching)

วิธีการนี้คล้ายกับวิธีการเทียบคำที่ยาวที่สุดเพียงแต่จะเลือกคำที่สั้นที่สุดที่พบก่อน แต่วิธีนี้พบว่าได้จำนวนคำมากที่สุดแต่ความถูกต้องของคำหลังทำการตัดคำ แล้วน้อยกว่าการใช้วิธีเทียบคำที่ยาวที่สุดตัวอย่างคำว่า “โคลงเรือ” การตัดคำโดยวิธีนี้จะเลือกเอาคำแรกที่ค้นหาเจอพจนานุกรม ดังนั้น จะได้ว่าโคลงเรือ (โดยไม่เลือกคำว่า “โคลง” ที่จะพบต่อไปภายหลังหากทำการค้นหาต่อ) วิธีนี้ใช้เวลาน้อยกว่าการเทียบคำยาวที่สุดแต่ความถูกต้องที่ได้การตัดคำแบบเทียบคำยาวที่สุดจะมากกว่า

3) วิธีการตัดคำที่ใช้ความถี่ของคำหรือสถิติ (Probabilistic Word Segmentation)

วิธีการนี้เป็นแนวทางหนึ่งในการแก้ปัญหาคำกำกวมของประโยคภาษาไทย โดยการวิเคราะห์ความถี่ของการใช้คำในชีวิตประจำวัน โดยจัดเรียงคำในพจนานุกรมตามความถี่ที่พบและใช้วิธีการตัดคำ ตัวอย่างคำว่า “ก็อด” ในกรณีนี้หากใช้ความถี่ของคำจะได้ว่า ก็อด มีความถี่สูงกว่าก็อด

4) วิธีการย้อนรอยกลับ (Back Tracking)

เมื่อทำการเปรียบเทียบคำที่นำมาตัดคำกับคำที่มีอยู่ในพจนานุกรมอาจพบกรณีที่คำที่พบมีมากกว่า 1 คำ แล้วทำการเลือกคำที่ยาวที่สุดทำให้สายอักขระที่ตามมาจากคำนั้น ไม่สามารถตัดคำได้เนื่องจากไม่พบตามพจนานุกรมกรณีนี้จะทำการย้อนไปอีกคำที่ไม่ถูกเลือก แล้วทำการตัดคำต่อไปตัวอย่างเช่น คำว่า “เมื่อยามนี้” การเปรียบเทียบกับพจนานุกรมจะได้ว่า เมื่อ, เมื่อย ดังนั้นจึงเลือกคำที่ยาวที่สุดจะได้คำว่า เมื่อย ส่วนที่เหลือคือ ตามนี้ ซึ่งไม่พบอยู่ในพจนานุกรม ดังนั้นจะทำการย้อนกลับไปเพื่อเลือกอีกคำหนึ่งคือ เมื่อ จะได้เป็น เมื่อยามนี้ (โดยคำว่า ยาม เกิดจากการเลือกคำที่ยาวที่สุดระหว่าง ยาและ ยาม)

5) วิธีการตัดคำแบบใช้คุณลักษณะ (Feature-Based Approach)

วิธีการนี้จะพิจารณาจากบริบท (Context Words) และการเกิดรวมกันของคำหรือหน้าที่ของคำ (Collocation) เข้ามาช่วยในการตัดคำ ตัวอย่างเช่น "ตากลม" ถ้าพบคำว่า "แป๋ว" ในบริบทก็จะสามารถตัดคำได้ว่า "ตา" "กลม" วิธีการนี้จำเป็นที่จะต้องมีฐานข้อมูลเป็นจำนวนมากและจะต้องมีการเรียนรู้การสร้างคำในบริบท หรือการเกิดรวมกันของคำแต่ละคำเพื่อให้มีข้อมูลที่จะนำมาใช้ในการตัดคำ ซึ่งจากวิธีการข้างต้นเรายังพบว่าแต่ละวิธีการมีข้อจำกัดและไม่สามารถแก้ปัญหาได้ทั้งหมด ในปัจจุบันจึงยังมีการพัฒนาและคิดค้นวิธีการใหม่ๆ ที่จะช่วยในการแบ่งคำให้มีประสิทธิภาพมากยิ่งขึ้น

2.6 อัลกอริทึม การแบ่งประเภท

ในส่วนของงานวิจัยที่ศึกษา ผู้วิจัยจึงขอกล่าวถึง 2 ประเภท ดังนี้

2.6.1 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้จะประกอบด้วย โหนดแทนคุณลักษณะและโหนดล่างสุดแทนหมวดหมู่การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าที่ใช้จะมาจากการคำนวณจากค่า Information Gain การสร้างต้นไม้ตัดสินใจ C 4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วย ค่าที่เป็นไปได้คือ {m_1, m_2, ..., mn} และให้ความน่าจะเป็นที่จะเกิดค่า m_1 มีค่าเท่ากับ P (m_i) จะได้ว่าคาเกนสารสนเทศ (Information Gain) ของ M เขียนแทนด้วย I (M) คำนวณได้ดังสมการ [23]

$$I(M) = \sum_{t=1}^{n} - P(m_i) \log_2 P(m_i) \quad (1)$$

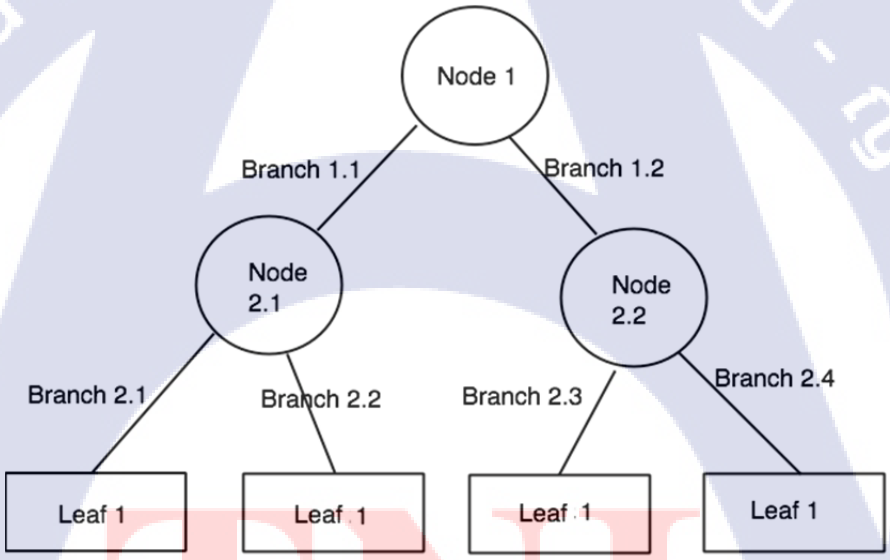
ถ้าให้ข้อมูลสอนคือ T และคูณลักษณะทั้งที่เป็นโหนดคือ x และมีค่าทั้งหมดทั้งที่เป็นไปได้ n ค่าโหนดปัจจุบันจะแบ่งตัวอย่าง T ออกตามกิ่งเป็น {t_1, t_2, ..., t_n} ตามค่าที่เป็นไปได้ของ x ดังนั้นจึงสามารถคำนวณค่าเกนสารสนเทศ หลังจากแบ่งตามคุณลักษณะและค่ามาตรฐานเกน (GAIN) ของคุณลักษณะ x ได้ ดังสมการ

$$Gant(x) = I(T) - I_x(T) \quad (2)$$

จากนั้นคำนวณค่าสารสนเทศของการแบ่งแยก (Split Information) ของคุณลักษณะแต่ละตัว ถ้าให้ T คือ ชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณลักษณะ x จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่ง คือ $\{t_1, t_2, ..., t_n\}$ จำนวน n ชุด ตามค่าที่เป็นไปได้ในคุณสมบัติ x เมื่อคำนวณค่าสารสนเทศของการแบ่งแยกได้ดังสมการ

$$\text{Eplit Information} = \sum_{t=1}^{T} \frac{|t_1|}{|T|} \log_2 \frac{|t_1|}{|T|} \quad (3)$$

คำนวณค่ามาตรฐาน อัตราส่วนเกน (Gain Ratio) ได้จาก Gain Ratio = Gain – Split Information ท้ายสุดจึงเลือกค่า Gain Ratio สูงสดเป็นคุณลักษณะเริ่มต้นและเลือกคุณสมบัติ ถัดไปตามค่า Gain Ratio น้อยลงตามลำดับ ดังรูปที่ 2.2



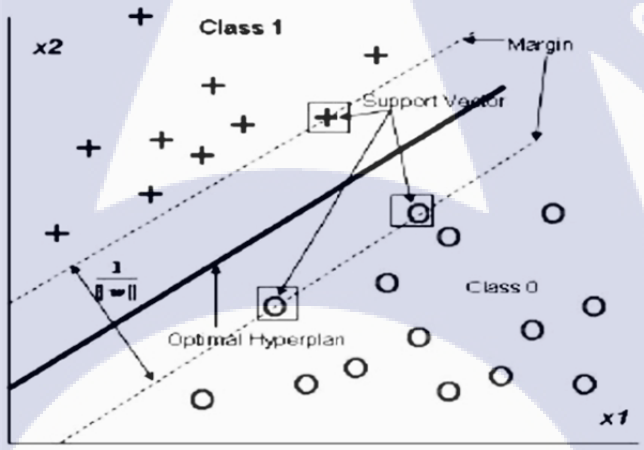
รูปที่ 2.2 ต้นไม้ตัดสินใจ

ซัพพอร์ทเว็กเตอร์แมชชีน (Support Vector Machine) หลักการของวิธีการนี้ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็น 2 ส่วน โดยใช้สมการเส้นตรง เพื่อแบ่งเขตข้อมูล 2 กลุ่ม ออกจากกัน โดยมีวัตถุประสงค์ที่จะพยายามที่จะทำการลดความผิดพลาดจากการทำนาย (Minimize Error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (Maximized Margin) ซึ่งต่างจากเทคนิคโดยทั่วไป เช่น โครงข่ายประสาทเทียม (Artificial Neural Network : ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียว โดยจะใช้ฟังก์ชั่นแม็ปข้อมูลจาก Input Space ไปยัง Feature

Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่า เคอร์เนล ฟังก์ชั่น (Kernel Function) บน Feature Space เหมาะใช้สำหรับข้อมูลที่มีลักษณะมิติของข้อมูลที่มีปริมาณมาก โดยกำหนดให้ $(x_i, y_i),...,(x_n, y_n)$ เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูลตัวอย่าง m คือ จำนวนมิติข้อมูลเข้า และ y คือ ผลลัพธ์มีค่า +1 หรือ -1 ดังสมการ $(x_i, y_i), ..., (x_n, y_n)$ เมื่อ $x \varepsilon R^m$, $y\varepsilon$ {+1, -1}

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจซึ่งคำนวณได้ดังสมการ และเมื่อ w คือ ค่าน้ำหนักและ b คือค่า bias สมการใช้สำหรับจำแนกประเภทของข้อมูล $(w*x) + b = 0$ โดย $(w*x) + b > 0$ ถ้า $y_i = +1$ และ $(w*x) + b < 0$ ถ้า $y_i = -1$

อย่างไรก็ตามซัพพอร์ทเว็กเตอร์แมชชีน มีเคอร์เนล ฟังก์ชั่น (Kernel Function) ผู้ที่สามารถประยุกต์ใช้ในการแก้ปัญหาได้หลายวิธี โดยผู้วิจัยต้องเลือกเคอร์เนลให้เหมาะสมกับลักษณะข้อมูล ดังรูปที่ 2.3



รูปที่ 2.3 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine)

2.6.2 เนอีฟเบย์ (Naïve-Bayes)

หลักการของวิธีการนี้ใช้การคำนวณความน่าจะเป็น ซึ่งถูกใช้ในการทำนายผลจัดเป็นเทคนิคในการแก้ปัญหาแบบ Classification ที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ การเรียนรู้เบย์อย่างง่ายเป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่งที่มีการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่าง มีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะ ดังสมการ

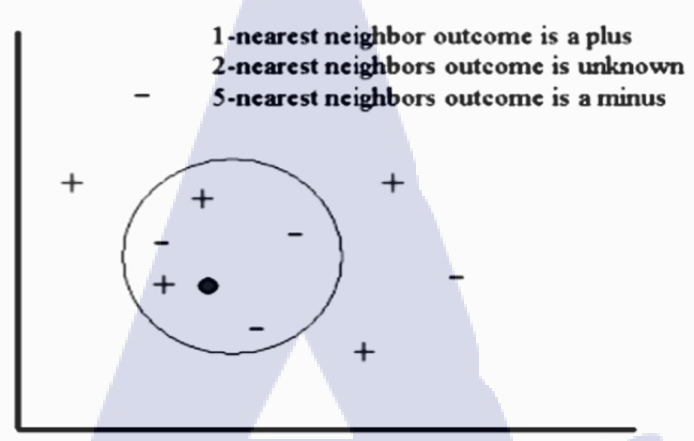
$$P(a_1, a_2,..., a_n \mid \nu_j) = \pi P(a_i \mid \nu_j) \quad (4)$$

กลุ่ม v_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{a_1, a_2, ..., a_n\}$ หรือ ใช้สัญลักษณ์ว่า $P(a_1, a_2, ..., a_n \mid v_j)$ โดยที่ $\bigcap$ หมายถึง ผลคูณของค่า $P(a_i \mid v_j)$ ทั้งหมด i = 1,2,3,..., n และ j = 1, 2, 3, ..., n ดังนั้นเราจะได้ค่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการ

$$\nu_{NB} = \arg.\max P(\nu_1) \times \pi P(a_i \mid \nu_j) \quad (5)$$

เคเนียเรสเนเบอร์ (K-Nearest Neighbor) หลักการของวิธีการนี้จะจำแนกประเภทข้อมูล โดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัวจากข้อมูลบนชุดข้อมูลตัวอย่างทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (Input Query Instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัว แล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่อยู่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วย แอททริบิวต์หลายตัวแปร X_i ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (Case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ อัลกอริทึมแบบ K-NN ได้แก่ 1-NN, 2-NN, 3-NN, ..., K-NN

ตัวอย่าง 2-KNN หมายถึง อัลกอริทึมันจะค้นหา 2 กรณี ที่มีลักษณะใกล้เคียงกับกรณีใหม่ (2 Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกในข้อมูลตัวอย่างฝึกฝนมาเรียงลำดับจากน้อยไปหามาก แล้วเลือกสมาชิกที่มีระยะทาง (Distance) ใกล้เคียงที่สุดออกมา K ตัว โดยใช้การวัดระยะทางแบบ Euclidean Distance มีหลักการคือ การวัดระยะทางระหว่างสองวัตถุ ถ้าวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย แต่ถ้ามีค่าน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก



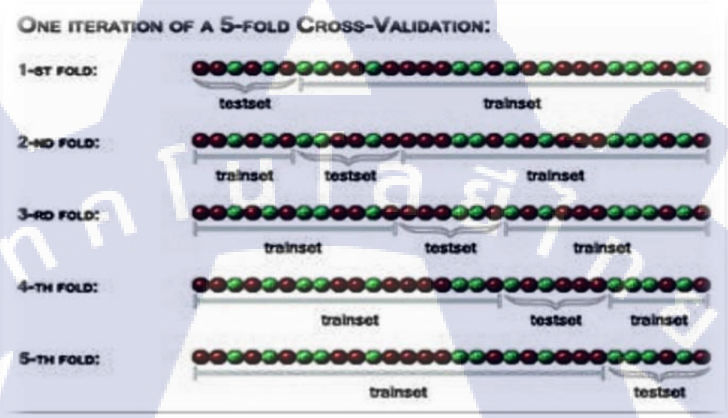
รูปที่ 2.4 เคเนียเรสเนเบอร์

จากรูปที่ 2.4 กฎของริปเปอร์หลักการของวิธีการ ประกอบด้วย 2 เฟสคือ เฟสแรกทำการระบุกฎเริ่มต้น และเฟสที่สองจะระบุค่า Post-Process Rule Optimization โดยข้อมูลที่ถูกการเรียนรู้ (Training) จะถูกแบ่งไปเป็น Growing Set และ Pruning Set โดยอัลกอริทึมนี้จะสร้างกฎความสัมพันธ์ใน Greedy ในขณะที่สร้างกฎริปเปอร์นั้นจะหาค่าที่ดีที่สุดสำหรับ Growing Set ใน Rule Space ซึ่งจะอธิบายได้จาก BNF หลังจากได้ Growing Set ก็จะทำการ Pruning ข้อมูล เมื่อเสร็จจะได้กลุ่มตัวอย่างที่เหมือนกันออกมาที่ครอบคลุมกฎของ Training Set จากนั้นก็จะลบทิ้ง ซึ่ง Training Data ที่เหลือจะถูกแบ่งใหม่อีกครั้ง หลังจากเรียนรู้ตามกฎแล้ว เพื่อช่วยแก้ปัญหาที่เกิดมาจากการแบ่งแยกกลุ่มที่ผิดพลาด ซึ่งกระบวนการนี้จะกระทำซ้ำจนกระทั้งผลเป็นที่น่าพอใจ

2.6.3 การประเมินโมเดล

ครอสวาลิเดชั่น (Cross Validation) คือ วิธีการในการคาดการค่าความผิดพลาดของโมเดล หรือวิธีการที่เรานำเสนอ โดยพื้นฐานของวิธีการครอสวาลิเดชั่นคือ การสุ่มตัวอย่าง (Resampling) โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชั่น มันถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network Architecture) โมเดลในการคิดแยกประเภท (Classification Model) เช่น ในการทำ Classify ข้อมูล โดยใช้เทคนิคของ Data Mining เช่น Neural Network หรือ Decision Tree นั้น จะต้องมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบทำให้ผลการ Classify นั้นดีเกินจริง ดังนั้นจะมีการคิดวิธี K-fold Cross Validation ขึ้นมาแก้ปัญหา

การแบ่งข้อมูลแบบ K-fold Cross-validation คือ การแบ่งข้อมูลออกเป็น K ชุดเท่าๆ กันและทำการคำนวณค่าความผิดพลาด K รอบ โดยแต่ละรอบการคำนวณข้อมูลชุดหนึ่งจากข้อมูล K ชุด จะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบและข้อมูลอีก K-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ดังตัวอย่างต่อไปนี้ K-fold Cross Validation (K=5) ชุดข้อมลหลังจากทำการแบ่งออกเป็น 5 ชุดข้อมูลย่อยเท่าๆ กัน โดยแต่ละกล่องคือ ชุดข้อมูลย่อย 1 ชุด ดังรูปที่ 2.5



รูปที่ 2.5 K-fold Cross Validation (K=5)

จากวิธีการข้างต้นจะได้ค่าความผิดพลาดของแต่ละรอบการคำนวณ ซึ่งประกอบด้วย e1, e2, e3, e4 และ e5 โดยปกติแล้วนั้น การหาค่าเฉลี่ยความผิดพลาด และใช้ค่านั้นเป็นตัวแทนของความผิดพลาดของโมเดลหรือวิธีการที่นำเสนอ ซึ่งสามารถแสดงได้ดังสมการต่อไปนี้ Average Error = (e1 + e2 + ... + eK)/K จากตัวอย่างในข้างต้นนั้น ข้อดีของวิธีการนี้คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกเรียนรู้ทั้งหมด K -1 ครั้ง โดยในขั้นตอนเหล่านี้เราสามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการคำนวณเป็นจำนวนรอบเทาใด แต่อย่างไรก็ตามเมื่อมองในมุมกลับกันวิธีการนี้ใช้เวลาในการคำนวณเป็น K เท่า ซึ่งในความเห็นส่วนตัวเวลานั้นไม่เป็นปัจจัยสำคัญต่อการวัดผลเมื่อเทียบกับกับความถูกต้องของการวัดผล

2.7 การให้น้ำหนักคำ

การกำหนดคำน้ำหนักของคำหรือวลี เพื่อการเรียนรู้และทดสอบโครงข่ายประสาทเทียม โดยเป็นหลักการให้ความสำคัญกับคำ หรือวลีในด้านของค่าสถิติศาสตร์ และไวยากรณ์ภาษาศาสตร์ คือ 1) ค่าสถิติความถี่การเกิดขึ้นของคำ จากหลักไวยากรณ์ภาษาในการใช้คำ หรือวลีเดิมซ้ำๆ เพื่อเป็นการเน้นย้ำถึงความสำคัญของคำ หรือวลีนั้นในเอกสาร 2) ค่าความสำคัญทางไวยากรณ์ภาษา ในด้าน

ตำแหน่งการเกิดขึ้นของแต่ละคำ หรือวลีในส่วนต่างๆ ของประโยค 3) ค่าความสำคัญทางไวยากรณ์ภาษาในด้านตำแหน่งการเกิดขึ้นของแต่ละคำ หรือวลีในส่วนต่างๆ ของย่อหน้า 4) ค่าความสำคัญทางไวยากรณ์ภาษา ในตำแหน่งการเกิดขึ้นของแต่ละคำ หรือวลีในส่วนต่างๆ ของเอกสาร 5) ค่าความสำคัญทางไวยากรณ์ภาษา ในการให้ความสำคัญกับหน้าที่ของแต่ละคำ หรือวลีในการเขียน

การให้น้ำหนักคำ (Term Weighting) หรือการกำหนดน้ำหนักคำเป็นวิธีการให้น้ำหนักสำหรับคำที่มีความสำคัญ หรือใช้เป็นตัวแทนของเอกสารที่ควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยในชุดของเอกสารที่เหลือทั้งหมด แต่ถ้าคำนั้นปรากฏเป็นจำนวนมากในทุกๆ เอกสาร แสดงว่าคำดังกล่าวไม่สามารถเป็นตัวแทนของเอกสารใดๆ ได้ ซึ่งคำเหล่านั้นเรียกว่าคำหยุด (Stop Word) เช่น a, and, the เป็นต้น ดังนั้น การให้น้ำหนักคำๆ หนึ่งในเอกสารฉบับหนึ่งจะพิจารณาจากความถี่ของคำ (Term Frequency) ที่ปรากฏในเอกสารนั้น และจำนวนของเอกสารทั้งหมดที่มีคำๆ นั้นปรากฏอยู่ วิธีการให้น้ำหนักของคำวิธีหนึ่งคือ TF IDF (Term Frequency. Inverted Document Frequency) โดย TF คือ ความถี่ของคำที่ปรากฏในเอกสาร [24] สามารถแสดงได้ดังสมการที่ 1

$$tf_{t,d} = \frac{0.5\ xf_{t,d}}{\max\{f_{t,d} : t \in d\}} \quad (6)$$

โดยที่ tf คือ จำนวนครั้งที่คำ t ปรากฏในเอกสารหารด้วยคำที่มีความถี่มากที่สุดในเอกสาร ส่วน IDF คือ ส่วนกลับความถี่ของเอกสารสามารถแสดงได้ดังสมการ

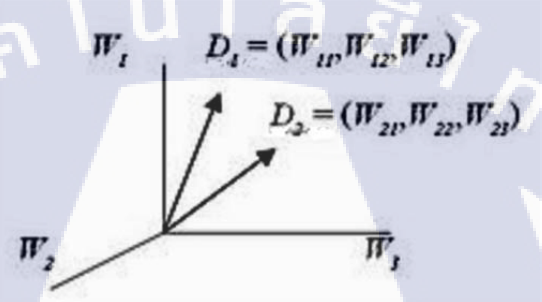
$$idf_{t,d} = \log \frac{N}{|\{d \in D : t \in d\}|} \quad (7)$$

โดยที่ N คือ จำนวนเอกสารทั้งหมดที่อยู่ในคลังเอกสารหารด้วยจำนวนเอกสารที่มีคำ t ปรากฏอยู่ และ TF-IDF สามารถแสดงได้ดังสมการ

$$w_{t,d,D} = tf_{t,d}\ X\ idf_{t,d} \quad (8)$$

2.7.1 แบบจำลองปริภูมิเวกเตอร์ (Vector Space Model : VSM)

การให้น้ำหนักคำ (Term Weighting) เป็นหนึ่งในวิธีการแทนเอกสารที่ไม่มีโครงสร้าง (Unstructured Text Document) ด้วยแบบจำลองเวกเตอร์สเปซ โดยกำหนดให้เอกสารแต่ละฉบับเปรียบเสมือนเวกเตอร์ของคำ ขนาดของเวกเตอร์ขึ้นอยู่กับจำนวนของคำที่ปรากฏอยู่ในเอกสารฉบับนั้น กำหนดให้ W_{ik} คือน้ำหนักของคำ K ที่ปรากฏในเอกสารฉบับที่ i เวกเตอร์สำหรับเอกสาร D_i เขียนแทนด้วย $D_i = (W_{i1}, W_{i2},..., W_{it})$ ซึ่ง t คือจำนวนของคำที่ไม่ซ้ำกัน ในชุดของเอกสารทั้งหมด ดังนั้นในช่องว่าง (Space) ของเอกสารชุดหนึ่งจะมีมิติเท่ากับ t มิติ เช่น เวกเตอร์ของเอกสารใน 3 มิติ [25] แสดงได้ดังรูปที่ 2.6



รูปที่ 2.6 เวกเตอร์ของเอกสารใน 3 มิติ

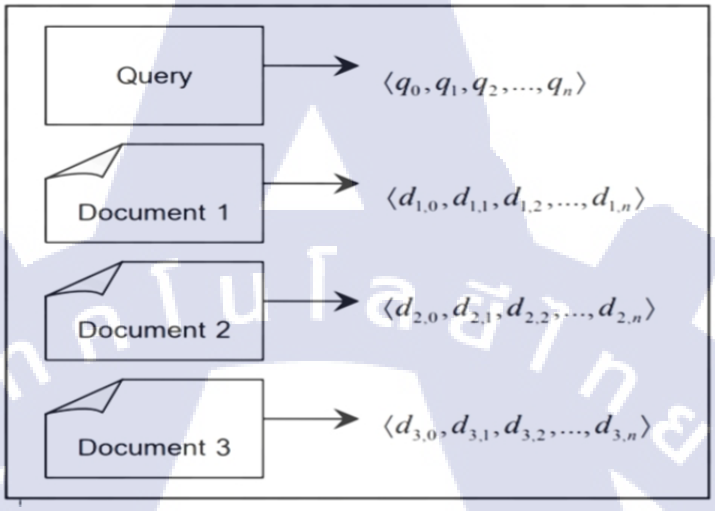
2.7.2 คำนิยามของแบบจำลองเวกเตอร์

สำหรับแบบจำลองเวกเตอร์นั้น ประกอบไปด้วยสมาชิกที่เป็นน้ำหนักของคำ (The Weight, $w_{i,j}$) ซึ่งมีความสัมพันธ์กันกับคู่อันดับของดัชนีของคำ โดยทั่วไป (Generic Index Term, k_i) กับเอกสาร (Document Term, d_j) โดยสามารถเขียนเป็นคู่อันดับได้ดังนี้ (k_i,d_j)

คำนิยามของแบบจำลองเวกเตอร์มีการกำหนดให้ $w_{i,j}$ เป็นน้ำหนักของคำที่มีความสัมพันธ์กันกับคู่อันดับ [k_i, d_j] โดยค่าของน้ำหนักของคำต้องมากกว่าหรือเท่ากับศูนย์ ($w_{i,j} \geq 0$) จากนั้นจึงนำน้ำหนักของคำมาสร้างเวกเตอร์ที่เป็นตัวแทนของคำที่ต้องการสืบค้น (Query Vector, q) ได้ดังนี้ $q = (w_{1,q}, w_{2,q},....w_{t,q})$ โดย t คือผลรวมของดัชนีของคำทั้งหมด (Index Term) ที่ปรากฏในเอกสาร และสามารถสร้างเวกเตอร์ที่เป็นตัวแทนของเอกสาร (Document Vector, d_j) ได้ดังนี้ $d_j = (w_{1,j}, w_{2,j},..., w_{t,j})$ โดย t คือผลรวมของดัชนีของคำทั้งหมด (Index Term) ที่ปรากฏในเอกสารเช่นเดียวกัน

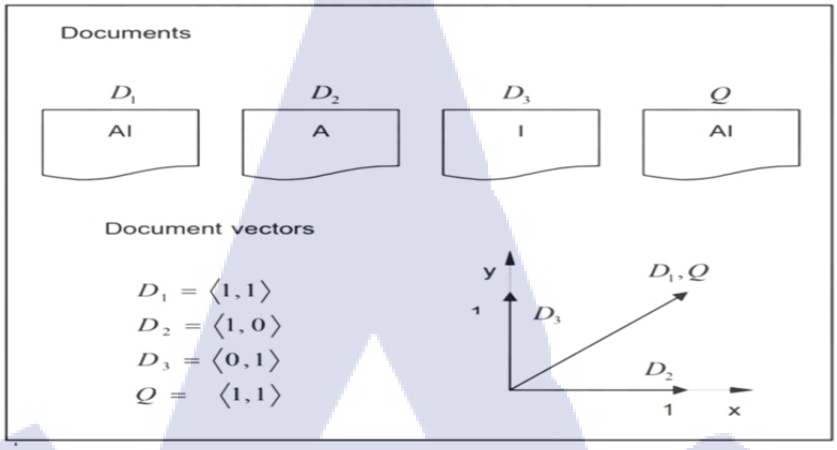
การนำแบบจำลองเวกเตอร์ไปประยุกต์ใช้ โดยรูปแบบหนึ่งในการนำแบบจำลองเวกเตอร์ไปประยุกต์ใช้นั้นก็คือ การวัดค่าความคล้ายคลึงกันระหว่างเอกสารกับคำที่ต้องการสืบค้น โดยกำหนดให้นำเวกเตอร์ตัวแทนแต่ละเอกสาร และเวกเตอร์ตัวแทนของคำที่ต้องการสืบค้นมาทำการคำนวณเพื่อ

หาตัวเลขสัมประสิทธิ์ความคล้ายคลึงกัน (Similarity Coefficient) หากเอกสารดังกล่าวประกอบไปด้วยคำที่ใกล้เคียงกัน หรือตรงกันกับคำที่ต้องการสืบค้น จึงจะถือได้ว่าเป็นเอกสารที่ตรงประเด็นที่ (Relevant Document) ดังแสดงหลักการพื้นฐานของแบบจำลองเวกเตอร์ดังรูปที่ 2.7



รูปที่ 2.7 เวกเตอร์ตัวแทนของเอกสารและเวกเตอร์ตัวแทนของคำที่ต้องการสืบค้นบนหลักการพื้นฐานของแบบจำลองเวกเตอร์

จากรูปที่ 2.8 แสดงเวกเตอร์ที่เป็นตัวแทนคำในแต่ละเอกสาร และเวกเตอร์ที่เป็นตัวแทนของคำที่ต้องการสืบค้น ซึ่งกระบวนการต่อไปจะนำตัวแทนเวกเตอร์ดังกล่าวไปทำการวัดมุมระหว่างเวกเตอร์แต่ละเวกเตอร์ โดยใช้วิธีการคำนวณแบบ Dot Product และการคำนวณ โดยใช้สูตรโคไซน์วัดค่าความคล้ายคลึงกัน (Cosine Similarity) เพื่อหาสัมประสิทธิ์ความคล้ายคลึงกัน (Similarity Coefficient) พิจารณายกตัวอย่างโดยให้เอกสารที่รวบรวมไว้ประกอบไปด้วยองค์ประกอบ 2 ส่วน องค์ประกอบแรกเป็นตัวแทนการเกิดขึ้นของคำ α และองค์ประกอบที่สองเป็นตัวแทนการเกิดขึ้นของคำ β ซึ่งจะให้ค่าเป็น 1 เมื่อมีคำนั้นปรากฏในเวกเตอร์เอกสาร และให้ค่าเป็น 0 เมื่อคำนั้นไม่ปรากฏในเวกเตอร์เอกสาร ยกตัวอย่างเช่น เอกสารที่ 1 ประกอบไปด้วยการเกิดขึ้น 2 ครั้งสำหรับคำ α และไม่มีการเกิดขึ้นของคำ β ดังนั้นเวกเตอร์ตัวแทนเอกสารในตัวอย่างนี้ก็คือ {2, 0} เป็นต้น แต่สำหรับเวกเตอร์ที่เป็นตัวแทนเอกสารใน 2 มิติที่มีสมาชิกของเวกเตอร์ ดังนั้น จะสามารถทำการคำนวณหาค่าความคล้ายคลึงกันได้แต่ไม่สามารถที่จะนับจำนวนความถี่ของคำที่เกิดขึ้นภายในเอกสารได้



รูปที่ 2.8 แบบจำลองเวกเตอร์ที่ประกอบไปด้วยคำศัพท์สองคำ

จากรูปที่ 2.8 แสดงตัวอย่างการสืบค้นตัวอักษรภาษาอังกฤษ 2 ตัวนั้นก็คือ A และ I โดยที่ตัวอักษรทั้งสองตัวนั้นถือได้ว่าเป็นองค์ประกอบหลักของเวกเตอร์ที่ต้องการค้นหาการแสดงเวกเตอร์ในมุมมอง 2 มิติที่เวกเตอร์ตัวแทนคำที่ต้องการสืบค้น และเวกเตอร์ตัวแทนเอกสารทั้ง 2 เอกสาร ทำให้สามารถหาค่าสัมประสิทธิ์ความคล้ายคลึงระหว่างคำที่ต้องการสืบค้น กับเอกสารทั้งหมดได้ด้วยการคำนวณระยะทางจากเวกเตอร์ตัวแทนคำที่ต้องการสืบค้นไปจนถึงเวกเตอร์ตัวแทนเอกสารอีก 2 เวกเตอร์ และจากตัวอย่างนี้เวกเตอร์ตัวแทนเอกสารที่ 1 มีความคล้ายคลึงกันกับเวกเตอร์ตัวแทนคำที่ต้องการสืบค้นมากที่สุด ดังนี้ การค้นคืนคำที่ต้องการค้นหาจากเอกสารที่ 1 จะตรงประเด็นมากที่สุด

2.8 งานวิจัยที่เกี่ยวข้อง

สมศักดิ์ วิชัยกิจ และมาลีรัตน์ โสดานิล [26] ศึกษาเกี่ยวกับการจัดกลุ่มลูกค้าสินเชื่อจากการปฏิเสธ ด้วยวิธีการทำเหมืองข้อความ จากการศึกษางานวิจัยที่เกี่ยวกับการรับรู้การสื่อสารการตลาดของผู้บริโภคที่มีต่อภาพลักษณ์ผลิตภัณฑ์ขององค์การส่งเสริมกิจการโคนมแห่งประเทศไทย ในเบื้องต้นมี 12 รายการ ดังนี้ วัตถุประสงค์เพื่อนำข้อมูลที่มีอยู่มาใช้ให้เกิดประโยชน์เป็นแนวทางเพิ่มความได้เปรียบทางการแข่งขันทางธุรกิจได้ โดยใช้วิธีการ Clustering วิธีการ K-means และวิธี SOM ซึ่งวิธีการ K-means ได้ผลดีที่สุด ผลการดำเนินงานข้อมูล จำนวน 5,898 โดยสามารถแบ่งกลุ่มออกเป็น 3 กลุ่ม ดังนี้ คือ 1. กลุ่มลูกค้าที่ต้องการสอบถามข้อมูลและยังตัดสินใจไม่ได้ 2. กลุ่มลูกค้าที่ต้องการรีไฟแนนซ์ 3 กลุ่ม ลูกค้าที่มีสินทรัพย์ไม่เข้าเกณฑ์ ซึ่งเป็นกลุ่มที่มีจำนวนมากที่สุด

เมธาพร คงทอง [27] ศึกษาเกี่ยวกับการจัดกลุ่มลูกค้าค้างชำระของสินเชื่อที่อยู่อาศัยของธนาคารพาณิชย์แห่งหนึ่ง มีวัตถุประสงค์เพื่อจัดกลุ่มลูกค้าค้างชำระของสินเชื่อที่อยู่อาศัยและหาลักษณะปัจจัยสำคัญประจำกลุ่ม วิธีการศึกษาเป็นการวิจัยเชิงทดลอง โดยใช้ข้อมูลลูกค้าที่ค้างชำระ

จำนวน 744 ราย ทำการวิเคราะห์การจัดกลุ่ม (Cluster Analysis) แบบวิธี K Means โดยใช้โปรแกรมสำเร็จรูปทางสถิติ โดยผลการศึกษาพบว่า สามารถจัดกลุ่มค้างชำระได้เป็น 3 กลุ่ม แต่ละกลุ่มมีลักษณะปัจจัยสำคัญประจำกลุ่ม ดังนี้ ประกันสินเชื่อ โดยนำผลที่ได้ไปจัดแนวทางการประยุกต์ เพื่อลดการเกิดหนี้ที่ไม่ก่อให้เกิดรายได้ จัดการความเสี่ยงด้านเครดิต และลดค่าใช้จ่าย ในขั้นตอนการควบคุมสินเชื่อ

ธิดาพร วงศ์พิพันธ์ [28] ศึกษาเกี่ยวกับการใช้เหมืองข้อมูลช่วยตัดสินใจในการให้สินเชื่อ กรณีศึกษาของกรุงไทยคาร์เร้นท์ มีวัตถุประสงค์เพื่อนำความรู้ที่ได้รับจากการทำเหมืองข้อมูลช่วยในการตัดสินใจการให้สินเชื่อมาเป็นแนวทางสนับสนุนการตัดสินใจ การอนุมัติสินเชื่อของบริษัทได้อย่างมีประสิทธิภาพมากขึ้น ซึ่งอาจมีส่วนช่วยลดปริมาณหนี้สูญได้ เริ่มต้นเตรียมข้อมูล โอนย้ายข้อมูล การลดขนาดและทำความสะอาดข้อมูล จากนั้นจึงให้ทำการทดสอบข้อมูลโดยใช้เทคนิคการจำแนกกลุ่มเพื่อหากฎที่ใช้ในการจำแนกลูกค้ากลุ่มที่ดี และลูกค้ากลุ่มที่ไม่ได้ ซึ่งได้ทดลองโดยใช้การแบ่งกลุ่ม (Cluster) โดยใช้เทคนิคโมเดล Classifies Tree J 48 และโมเดล Classifies PART และโมเดล Classifies Naive Bayes ผลลัพธ์ที่ได้คือกำที่ใช้ในการจำแนกลูกค้ากลุ่มที่ดีและไม่ดี ซึ่งพบว่าการจำแนกกลุ่มข้อมูลด้วยเทคนิค Classifies Tree J 48 ให้ผลลัพธ์ดีที่สุด เนื่องจากมีความถูกต้องมากที่สุด และสามารถแบ่งกลุ่มได้ ตามเงื่อนไขมากที่สุด เมื่อเทียบกับอีก 2 เทคนิค ซึ่งท้ายสุดจึงนำกฎที่ได้จากผลลัพธ์มาเป็นแนวทางและเงื่อนไขสำหรับการเขียนโปรแกรมเป็นระบบการตัดสินใจอนุมัติสินเชื่อออนไลน์ได้อย่างมีประโยชน์

Jiao et al. [29] ในการสกัดคำสำคัญภาษาจีนนั้นมีจุดประสงค์หลักคือ การแบ่งคำแบบอัตโนมัติก่อนการนำไปแก้ปัญหา วิธีการนี้เป็นการถอดรหัสคำภาษาจีนบน Word Platform และสร้างรูปแบบของเอกสารขึ้นใหม่ในคอมพิวเตอร์ วิธีการนี้สร้างจากหน่วยคำศัพท์เล็ก ๆ ในสารสนเทศการสกัดสำคำสำคัญภาษาจีน ทำโดยไม่อาศัยการตัดแบ่งคำ โดยวิธีการใหม่นี้ส่งเสริมให้การทำงานมีประสิทธิภาพมากขึ้น ซึ่งการนำการวิเคราะห์ทางด้านสถิติมาใช้ทำให้ผลที่ได้รับมีประสิทธิภาพมากยิ่งขึ้น

Li et al. [30] นำเสนออัลกอริทึมใหม่ในการสกัดคำสำคัญสำหรับเว็บข่าวในภาษาจีน โดยใช้lexical Chains คือ การพิจารณาความสัมพันธ์ของคำ โดยใช้ความหมายของคำใน HowNet มาใช้ในการสร้าง Lexical Chains และใช้ Word Co-occurrence ร่วมกับ Feature หรือคุณลักษณะ ได้แก่ Frequency Features, Cohesion Feature และ Correlation Features ในงานวิจัยนี้ได้นำเสนอ KELCC Algorithm ซึ่งมีการหลักการทำงานคือ นำข้อมูลข่าวจากเว็บมาผ่านการทำ Preprocessing แล้วหาค่า TFIDF ของแต่ละคำ และเลือก Candidate Words คือ คำที่มีคุณสมบัติพอที่จะเป็นคำสำคัญได้ ซึ่งสามารถมีได้มากกว่า 1 คำ จากนั้นเลือกคำจาก HowNet มาสร้าง Lexical Chains และคำนวณหาค่าความถี่ของ Word Occurrence จากนั้นคำนวณหาค่าน้ำหนักของคำตาม Feature ที่กำหนดไว้และเลือกคำที่มีค่าน้ำหนักที่มากที่สุดของเอกสารออกมาเป็นคำสำคัญ ซึ่งอัลกอริทึมนี้ไม่มี

การกำหนด Domain-Specific และยังสามารถนำไปประยุกต์ใช้กับ Single Web Page แทน Corpus ได้อีกด้วย ในการทดลองได้ทำการสุ่มเลือก Web Pages มาใช้เพื่อช่วยในการเพิ่มประสิทธิภาพในงานวิจัย

Wartena et al. [31] นำเสนอการกำหนดคำสำคัญให้เอกสาร โดยการเลือกคำที่เหมาะสมที่สุดจากข้อความในเอกสาร โดยมีเกณฑ์ที่สำคัญคือ คำสำคัญต้องเป็นคำที่มีความสัมพันธ์กันกับข้อความในเอกสารนั้น ๆ โดยมีการใช้ข้อมูล 2 ชุด คือ ACM Abstract และ BBC Synopses โดยมีการพิจารณาการหาค่าความสำคัญของแต่ละคำด้วยวิธีการทางสถิติ 4 คือ TFIDF, JSD, InfGain, Correlation และนำค่าที่ได้ไปหาค่า Co-Occurrence Distribution ผลการทดลองปรากฏว่า เมื่อนำ TFIDF มาใช้ร่วมกับการ Co-Occurrence Distribution ทำให้มีผลดีกว่าการทำงานร่วมกับวิธีอื่นๆ

Zheng and Ye [32] ได้ทำการทดลองกับบทวิจารณ์ของนักท่องเที่ยวในภาษาจีนบนเว็บไซต์ ctrip.com โดยนำขั้นตอนวิธีการให้คำแนะนำเพื่อการเรียนรู้ของเครื่องคอมพิวเตอร์ (SVM) มาจำแนกทัศนคติของบทวิจารณ์ภาษาจีน จำนวน 40 ที่พักในเมืองเทียนจิน (Tianjin) และชองชิง (Chongqing) ผลการศึกษาพบว่า (1) ขั้นตอนวิธีในการจำแนกทัศนคติของบทวิจารณ์นี้สามารถใช้ได้ดีกับบทวิจารณ์ในภาษาจีน (2) บทวิจารณ์ของนักท่องเที่ยวบนอินเทอร์เน็ตส่งผลกระทบที่สำคัญต่อการจองที่พักในประเทศจีน และ (3) ผลที่ได้จากงานวิจัยนี้สามารถนำไปใช้เพิ่มศักยภาพทางธุรกิจที่พักในประเทศจีนได้อย่างไร้ข้อสงสัย

J. M. Kassim and M. Rahmany [33] ได้นำเสนอวิจัยเรื่อง “Introduction to Semantic Search Engine” โดยได้นำเสนอแนวคิดเกี่ยวกับหลักการของการพัฒนาระบบช่วยบริการสืบค้น โดยใช้หลักการเว็บเชิงความหมายมาช่วยพัฒนา ซึ่งในงานวิจัยนี้ได้อธิบายภาพรวมของหลักการและนำเสนอเทคโนโลยีที่จำเป็นต่อการพัฒนาระบบ และยังได้นำเสนอเว็บไซต์ที่ให้บริการสืบค้นสารสนเทศซึ่งใช้หลักการเชิงความหมาย สำหรับผู้ที่สนใจสามารถเข้าไปทดลองใช้บริการสืบค้น และสุดท้ายได้สรุปข้อแตกต่างระหว่างเทคโนโลยีเว็บเชิงความหมาย ซึ่งเป็นเทคโนโลยีเกี่ยวกับ 3.0 กับเทคโนโลยีปัจจุบัน ซึ่งเป็นเทคโนโลยี2.0

B. Fang and S. Ma [34] วิจัยเรื่อง “Data Mining and Its Applications in CRM of Commercial Banks” ได้วิจัยงานเกี่ยวกับการทำเหมืองข้อมูล เรื่องการทำการบริหารลูกค้าสัมพันธ์ของธนาคารพาณิชย์ว่าการบริหารลูกค้าสัมพันธ์ เริ่มต้นในปี 1980 และได้มีการพัฒนาเรื่อยๆ โดยการที่ใช้ลูกค้าเป็นศูนย์กลางในปี 1990 ในทุกวันนั้นเรื่องนี้ได้กลายเป็นเรื่องที่สำคัญและมีการบริหารลูกค้าสัมพันธ์ ในหลายธุรกิจ และทำให้เห็นว่าการบริหารลูกค้าสัมพันธ์เป็นแนวคิดของการตลาดแบบใหม่ และการวางกยุทธ์ที่ใช้ลูกค้าเป็นศูนย์กลาง โดยที่การบริหารลูกค้าสัมพันธ์จะช่วยให้ได้รับผลกำไรจากลูกค้าเพิ่มมากขึ้น รักษาฐานลูกค้าไว้ได้และช่วยลดค่าใช้จ่ายในการดำเนินงานได้ การที่จะเก็บข้อมูลของลูกค้าของธนาคารพาณิชย์ สามารถเก็บได้จากการใช้บริการต่างๆ โดยที่ข้อมูลเหล่านี้จะถูกเชื่อม ต่อ

กับฐานข้อมูลของธนาคาร โดยที่ข้อมูลเหล่านั้นได้มาจากการฝากเงิน ใช้บริการทางโทรศัพท์ แฟกซ์หรือทางสาขาของธนาคาร และการที่เรานำข้อมูลเหล่านี้มาใช้งาน เนื่องจากข้อมูลเหล่านี้มีจำนวนมากและถูกจัดเก็บไว้ ไม่ได้ใช้งานทั้งที่ ข้อมูลเหล่านี้ สามารถช่วยให้เราสามารถตัดสินใจในการดำเนินการได้ ซึ่งการทำเหมืองข้อมูลของธนาคารพาณิชย์ในเรื่องของการบริหารลูกค้าสัมพันธ์ สามารถแบ่งได้หลายทาง 1. แบ่งข้อมูลตามแต่ละประเภท 2.แบ่งตามกลุ่มข้อมูลที่นำมาวิเคราะห์ 3. แบ่งตามความสัมพันธ์ที่นำมาวิเคราะห์ 4. แบ่งตามข้อมูลที่เกี่ยวข้องกัน ซึ่งการทำเหมืองข้อมูลของธนาคารพาณิชย์ ช่วยในการพยากรณ์พฤติกรรมลูกค้า เพื่อใช้วางแผนในการทำการตลาด และประโยชน์ของการทำเหมืองข้อมูลของธนาคารพาณิชย์ คือทำให้เรา สามารถแบ่งกลุ่มลูกค้า รักษาฐานลูกค้า ต่อยอดการขายอย่างต่อเนื่อง วิเคราะห์ความเสี่ยงของลูกค้าและกล่าวถึง ปัจจัยความสำเร็จของการทำเหมืองข้อมูลของธนาคารพาณิชย์ ต้องมีการวางแผนเก็บข้อมูลของลูกค้า 2 บริษัท มียอดขาย บริการหลังการขาย 2 และข้อมูลลูกค้าอื่นๆ 2 ลงในฐานข้อมูล 2 ของลูกค้าใน 2 คลังข้อมูล (Data Warehouse) ในการจัดเก็บข้อมูลเหล่านี้ ซึ่งข้อมูลเหล่านี้จะเป็นพื้นฐานของการประสบความสำเร็จในการวางแผน เพื่อนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์

Noroozi and Fotouhi [35] ศึกษาอิทธิพลของเว็บเชิงความหมายที่มีต่อการตัดสินใจของผู้บริโภคในอุตสาหกรรมการท่องเที่ยว โดยมุ่งเน้นการเก็บข้อมูลจากระบบรับบทวิจารณ์สินค้าหรือบริการ ผลการศึกษาพบว่า ผู้บริโภคที่มีความเกี่ยวข้องกับสินค้าหรือบริการต่ำจะวิเคราะห์ข้อมูลและตัดสินใจอย่างระมัดระวัง หรือไม่เชื่อถือบทวิจารณ์เลย ในขณะที่ผู้บริโภคที่มีความเกี่ยวข้องกับสินค้าหรือบริการสูงจะให้ความสำคัญกับบทวิจารณ์ และสิ่งที่เกี่ยวข้องด้วย อาทิ จำนวนผู้วิจารณ์ งานวิจัยนี้แสดงให้เห็นว่า นอกจากการอ่านเนื้อหาบทวิจารณ์หลักแล้วบทบาทของปัจจัยทางสังคมยังมีอิทธิพลอย่างมีนัยสำคัญต่อทัศนคติของผู้บริโภคที่มีต่อบทวิจารณ์บนเว็บไซต์

J. L. Seng [36] งานวิจัย “An Analytic Approach to Select Data Mining for Business Decision” ศึกษาโดยการทบทวนวรรณกรรมต่างๆ ซึ่งข้อมูลของวิธีการทำเหมืองข้อมูลๆ หมายถึง การเตรียมเครื่องมือ ฟังก์ชั่นที่ใช้ทำเหมืองข้อมูล โดยความหมาย แนวคิดของการทำเหมืองข้อมูลของแต่ละวิธีการ และเกณฑ์การแบ่งประเภทแตกต่างกัน โดยสามารถจำแนกได้เป็น แบบการพยากรณ์ การจัดกลุ่ม โดยทั่วไปในการทำงานที่แตกต่างกัน จะมีการจำแนกข้อมูลและวิธีใช้งานที่แตกต่างกัน การสรุปกฎลำดับอาจไม่เหมือนกัน ข้อมูลวิธีการทำเหมืองข้อมูลรวมถึงเทคนิคที่คิดค้นจาก สถิติ การเรียนรู้เครื่อง OLAP และอื่นๆ แล้วได้สรุปถึงวิธีการทำเหมืองข้อมูล โดยสรุปว่าควรใช้ตัวแปรไหนสำหรับจะทำเหมืองข้อมูล

Q. Luo [37] งานวิจัยเรื่อง “Advancing Knowledge Discovery and Data Mining” กล่าวถึงการค้นหาความรู้ และการทำเหมืองข้อมูลได้กลายมาเป็นสิ่งสำคัญ เนื่องจากความต้องการที่เพิ่มมากขึ้น ซึ่งรวมถึงการใช้ในการเรียนรู้ของเครื่อง (Machine Learning) ฐานข้อมูล (Databases) สถิติ (Statistics)

การจัดหาความรู้ (Knowledge Acquisition) ข้อมูลจินตทัศน์ (Data Visualization) และการประมวลผลคุณภาพสูง (High Performance Computing) การค้นหาความรู้ และการทำเหมืองข้อมูลมีประโยชน์อย่างมากในหลายๆ ด้านของปัญญาประดิษฐ์ (Artificial Intelligence) เช่น อุตสาหกรรม พาณิชย์ รัฐบาล และการศึกษา เป็นต้น ในงานวิจัยนี้จะแสดงถึงความสัมพันธ์ระหว่างความรู้กับการทำเหมืองข้อมูลและกระบวนการค้นคว้าความรู้ฐานข้อมูล (KDD) ทฤษฎีการทำเหมืองข้อมูล ขั้นตอนการทำเหมืองข้อมูล เทคโนโลยีเหมืองข้อมูล

Y. Jin et al. [38] ศึกษาเรื่อง “The Research of Search Engine Based on Semantic Web” โดยได้นำเสนอการนำอัลกอริทึม TFIDF เข้ามาช่วยในการประมวลผลในการแยกคำหรือการตัดคำ ซึ่งเป็นการนำหลักการทางสถิติเข้ามาช่วย โดยการทำงานจะใช้การจัดทำดัชนีเป็นตัวชี้ความสำคัญของข้อมูล ที่ช่วยประเมินความสำคัญของข้อมูลในการนำมาแสดงผล เพื่อเพิ่มความถูกต้องในการสืบค้นทำให้การสืบค้นมีการคัดกรองข้อมูลผลลัพธ์ ก่อนนำมาแสดงผลหลากหลายขึ้น จากการทดสอบระบบ โดยทำการทดสอบเปรียบเทียบกับระบบสืบค้นที่ไม่ใช้อัลกอริทึม TFIDF ผลลัพธ์ได้จากระบบที่พัฒนาขึ้น มีความถูกต้องแม่นยำมากกว่าระบบที่ไม่ได้ใช้อัลกอริทึม TFIDF

นลินี พฤกษาชาติ และธนพล เจนสุทธิเวชกุล [39] การจำแนกความคิดเห็นออนไลน์ของนักท่องเที่ยวโดยใช้ Word Vector และอัลกอริทึมเนอีฟ เบย์ ได้เสนอกระบวนการจำแนกความคิดเห็นออนไลน์โดยเทคนิค Word2vec ร่วมกับอัลกอริทึมเนอีฟ เบย์ โดยเก็บข้อมูลจากข้อคิดเห็นจากลูกค้าใน 20 โรงแรม โดยดึงข้อมูลข้อคิดเห็นจาก Agoda.com ด้วยซอฟต์แวร์ WebHarvy และ Tripadvisor.com ด้วยซอฟต์แวร์ Octoparse จากนั้นใช้ RapidMiner 7.6 ในการทำเหมืองข้อมูลผลการวิจัยนี้พบว่าสถานที่ตั้งของโรงแรมมีผลต่อการตัดสินใจเลือกโรงแรมของนักท่องเที่ยวมากที่สุด

ทิพย์อรุณ เขี้ยวแก้ว และณัฐพงศ์ ทองเทพ [40] การวิเคราะห์การสกัดรูปแบบการตอบกลับที่สุภาพต่อข้อความแสดงความเห็นที่มีต่อสินค้าและบริการ ได้เสนอรูปแบบการตอบที่สุภาพต่อความคิดเห็นของลูกค้าหลังซื้อหรือใช้สินค้าและบริการ โดยผู้วิจัยได้รวบรวม 231 ข้อความจากแหล่ง ข้อมูลสาธารณะ คือ เฟสบุ๊คและพันทิป โดยเลือกจาก 2 ธุรกิจ คือ KFC และธนาคารกสิกรไทย จากนั้นนำมาวิเคราะห์เพื่อจัดกลุ่มและสร้างรูปแบบการตอบที่มีความสุภาพ จากผลการวิจัยพบว่าข้อความตอบกลับที่สุภาพจากธุรกิจต่อความคิดเห็นของลูกค้าสามารถแบ่งออกได้ 35 รูปแบบจากสองกลุ่ม คือ รูปแบบการตอบกลับที่สุภาพสำหรับความคิดเห็นเชิงลบมีจำนวน 21 รูปแบบ และรูปแบบการตอบกลับที่สุภาพสำหรับความคิดเห็นเชิงบวกมีจำนวน 14 รูปแบบ

รณกร ขำเพชร และวรารัตน์ สงฆ์แป้น [41] นำเสนอโมเดลการวิเคราะห์ข้อความข่าวเพื่อพยากรณ์แนวโน้มราคายางพาราในวันถัดไปด้วยเทคนิคการทำเหมืองข้อความ โดยมีการจำแนกข้อความข่าวที่ส่งผลราคายางเปลี่ยนแปลงในวันถัดไป จากการวิเคราะห์คำในข้อความข่าวที่ส่งผลต่อราคายางมีทั้งหมด 73 ปัจจัย จำนวน 326 ระเบียน แบ่งเป็นชุดข้อมูลแบบสอนและทดสอบโดยใช้เทคนิคการ

ทำเหมืองข้อมูล 3 วิธีพบว่า วิธีของ K-Nearest Neighbors (K-NN) ให้ความถูกต้องสูงที่สุดคิดเป็น 96% และวิธีต้นไม้ตัดสิน (Decision Tree) ให้ความถูกต้องคิดเป็น 92% และวิธี Naive Bayes ให้ความถูกต้องคิดเป็น 89.33% ดังนั้นการทำเหมืองข้อมูลโดยใช้โมเดล K-Nearest Neighbors (K-NN) จะสามารถช่วยให้ผู้อ่านข่าวสามารถคาดการณ์การเปลี่ยนแปลงของราคายางพาราได้

Krishnaveni and Hemalatha [42] ศึกษาการวิเคราะห์การเกิดอุบัติเหตุทางจราจรด้วยเทคนิคเหมืองข้อมูล พบว่า รูปแบบการจัดหมวดหมู่เพื่อทำนายความรุนแรงของการบาดเจ็บที่เกิดขึ้นระหว่างการเกิดอุบัติเหตุจราจรด้วย Naive Bayesian AdaBoost M1 Meta classifier, PART Rule classifier, J48 Decision Tree Classifier และ Random Forest Tree Classifier จำแนกประเภทความรุนแรงการบาดเจ็บของอุบัติเหตุจราจรต่างๆ สุดท้ายแสดงให้เห็นว่า Random Forest ทำได้ดีกว่าอัลกอริทึมอื่นๆ

T. Niu et al. [43] ศึกษาเรื่อง Sentiment Analysis on Multi-View Social Data เป็นการแนะนำชุดขอมูลใหม่ที่เรียกว่า MVSA ซึ่งประกอบด้วยทวีตหลายมุมมองสำหรับการวิเคราะห์ความเชื่อมั่น ผลของเราสามารถใช้ประโยชน์ได้เป็นพื้นฐานนอกเหนือจากการวิเคราะห์ความเชื่อมั่นแล้วชุดข้อมูลของเรายังสามารถใช้ในการตรวจสอบปัญหาที่น่าสนใจอื่นๆ ที่เปิดเช่นการตรวจสอบ Tweets นอกจากนี้ยังได้แสดงให้เห็นว่ามีป้ายข้อความที่ไม่สอดคล้องกันหลายอย่างระหว่างข้อความมุมมองและภาพในทวีตที่เก็บรวบรวม

J. G. Kang et al.. [44] ศึกษาเรื่อง Semantic Network Analysis of Vaccine Sentiment in Online Social Media เพื่อตรวจสอบความเชื่อมั่นของวัคซีนในปัจจุบันเกี่ยวกับสื่อทางสังคมโดยการสร้างและวิเคราะห์เครือข่ายความหมายของข้อมูลวัคซีนจากเว็บไซต์ที่ใช้งานร่วมกันของผู้ใช้ Twitter ในสหรัฐอเมริกา และเพื่อช่วยในการสื่อสารเรื่องสาธารณสุขกับวัคซีน มีการรวบรวมทวีตทั้งหมด 26,389 รายการระหว่างวันที่ 16 เมษายน 2015 และวันที่ 29 พฤษภาคม 2015 ซึ่งเราได้รับลิงก์เว็บที่ไม่ซ้ำกัน 8,416 รายการ

N. Kumar et al. [45] ศึกษาเรื่อง Sentiment Dynamics in Social Media News Channels วิเคราะห์เปรียบเทียบความเชื่อมั่นของโพสต์ข่าวสื่อสังคมออนไลน์ทางโทรทัศน์วิทยุและสื่อสิ่งพิมพ์วิเคราะห์ปฏิกิริยา ของผู้ใช้และความคิดเห็นเกี่ยวกับโพสต์ข่าวสารที่มีความรู้สึกต่างกัน ซึ่งชุดข้อมูลที่สกัดจากหน้า Facebook ของ 5 ช่องข่าวยอดนิยม ได้แก่ CNN, Fox News, The Economist, NYT และ NPR ชุดข้อมูลมีข่าว 0.15 ล้านฉบับ และปฏิกิริยาผู้ใช้ 1.13 พันล้านคน ผลการทดลองแสดงให้เห็นวาความรูสึกของความคิดเห็นของผู้ใช้มีความสัมพันธ์กันอย่างมากกับความเชื่อมั่นของข่าวสารและประเภทของแหล่งข้อมูล การศึกษาของเรายังแสดงให้เห็นถึงความแตกต่างระหว่างช่องทางสื่อสังคมออนไลน์ของแหล่งข่าวต่างๆ

ดังนั้นจึงสรุปได้ว่าในงานวิจัยเล่มนี้ ขอเสนอวิธีการตัดคำโดยใช้เทคนิคแบบผสมผสาน ได้แก่ แบบกฎ แบบพจนานุกรม และแบบคลังข้อมูล ที่เข้ามาช่วยในส่วนของการแยกคำออกจากประโยคให้มีความถูกต้องและแม่นยำมากยิ่งขึ้น เพื่อนำความคิดเห็นที่ได้ไปวิเคราะห์หารูปแบบความคิดเห็นโดยแยกออกเป็นความคิดเห็นเชิงบวก ความคิดเห็นเชิงลบ ความคิดเห็นตรงกลางหรือที่เรียกว่า “เป็นกลาง” เพื่อนำไปประเมินประสิทธิภาพการวิเคราะห์รูปแบบความคิดเห็นต่อไป

บทที่ 3
วิธีการดำเนินงาน

งานวิจัยเรื่อง การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมโดยใช้เทคนิคการตัดคำแบบผสมผสาน มีขั้นตอนการทำงานดังนี้

ขั้นตอนที่ 1 การวิเคราะห์ข้อมูลปัญหาของงานวิจัย

ขั้นตอนที่ 2 การศึกษาข้อมูลและกลุ่มตัวอย่างที่ใช้ในงานวิจัย

ขั้นตอนที่ 3 การวิเคราะห์และออกแบบกระบวนการทางเทคนิคการตัดคำแบบผสมผสาน

ขั้นตอนที่ 4 การประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็น

ขั้นตอนที่ 5 การสรุปผลการวิจัย

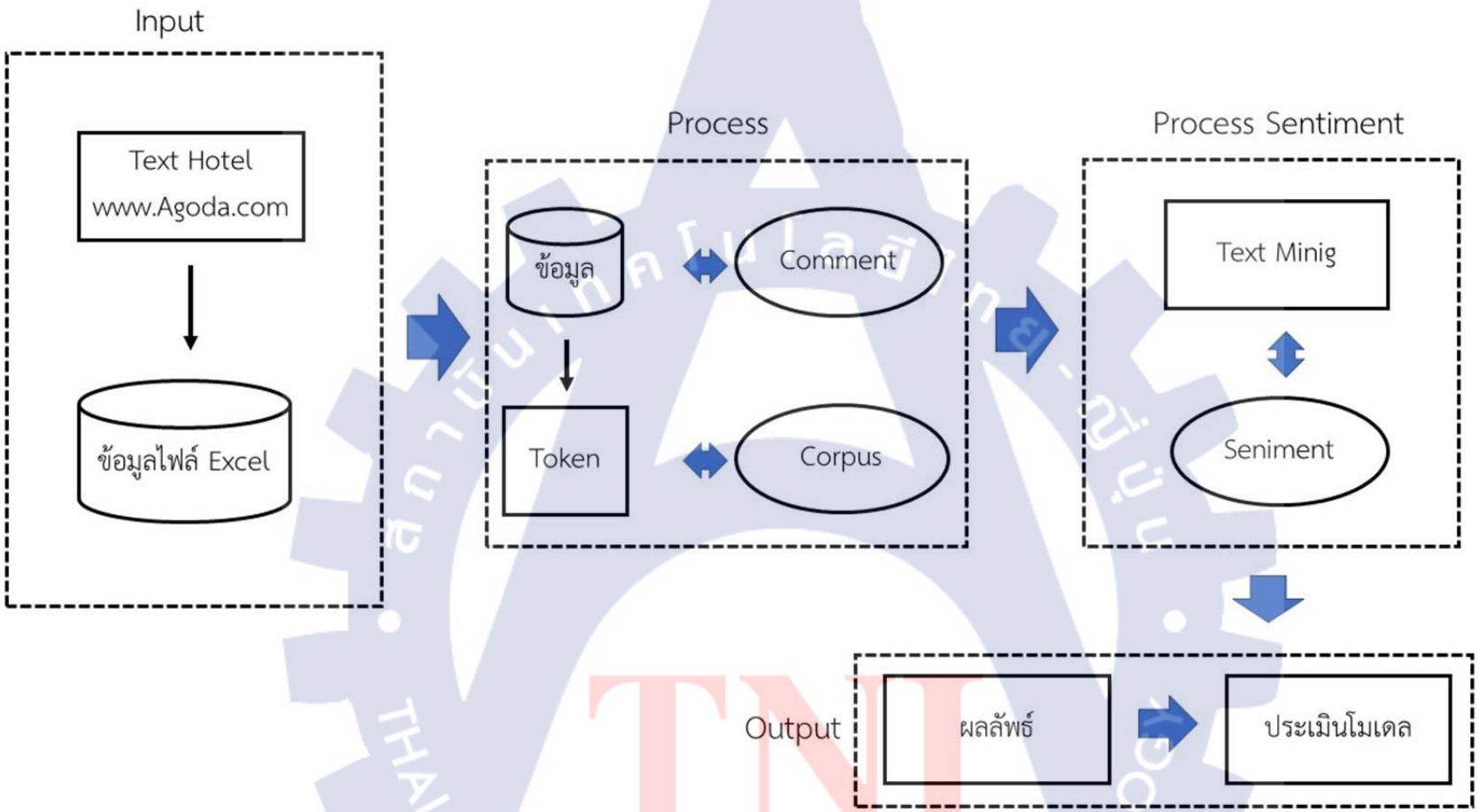
รูปที่ 3.1 แผนผังกระบวนการทำงาน

โดยแผนการดำเนินงานวิจัยมีทั้งหมด 3 ขั้นตอน ดังตารางที่ 3.1 แผนการดำเนินงานวิจัย

1. ขั้นตอนเบื้องต้น เป็นขั้นตอนการวิเคราะห์ปัญหาเพื่อนำไปสู่หัวข้อวิจัยที่จะใช้ในการพัฒนาและการทบทวนวรรณกรรมเพื่อเลือกวิธีการดำเนินงานการวิจัย

2. ขั้นตอนการดำเนินการ เป็นขั้นตอนการเก็บรวบรวมข้อมูล การเตรียมข้อมูล การทดลองประมวลผลข้อมูล และทำไปประเมินประสิทธิภาพ

3. ขั้นตอนการสรุป เป็นขั้นตอนสุดท้ายในการทำงานวิจัยการสรุปผลงานวิจัยทั้งหมดที่ได้ทำมาในทั้งขั้นตอนเบื้องและขั้นตอนการดำเนินการ ทั้งสองขั้นตอนข้างต้นเพื่อนำไปเขียนรูปเล่มวิทยานิพนธ์และบทความวิชาการ



รูปที่ 3.2 กรอบแนวความคิด

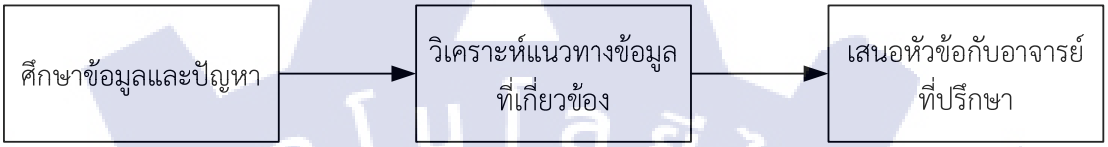
ตารางที่ 3.1 แผนการดำเนินงานวิจัย

การดำเนินงาน	ระยะเวลา								
	ปี 2561			ปี 2562					
	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.
ขั้นตอนเบื้องต้น									
- การหาหัวข้อ									
- การทบทวนวรรณกรรม									
- ร่างเค้าโครงวิทยานิพนธ์									
- นำเสนอเค้าโครงวิทยานิพนธ์									
ขั้นตอนการดำเนินการ									
- การเก็บข้อมูลและเตรียมข้อมูล									
- การประมวลผล									
ขั้นตอนการสรุป									
- สรุปและตีพิมพ์งานวิจัย									
- เขียนวิทยานิพนธ์เล่มสมบูรณ์									
- สอบป้องกันวิทยานิพนธ์									

โดยขั้นตอนการดำเนินงานวิจัยในครั้งนี้มีทั้งหมด 5 ขั้นตอน และมีวิธีการแบ่งเป็นรายละเอียดในแต่ละขั้นตอนตามรูปภาพที่ 3.1 ดังนี้

3.1 ขั้นตอนที่ 1 การวิเคราะห์ข้อมูลปัญหาของงานวิจัย

ผู้วิจัยได้ทำการวิเคราะห์ข้อมูลปัญหาของงานวิจัย เพื่อศึกษาหาผลกระทบหรือปัญหาการทำเหมืองข้อมูลด้านเทคนิคการตัดคำ ดังต่อไปนี้



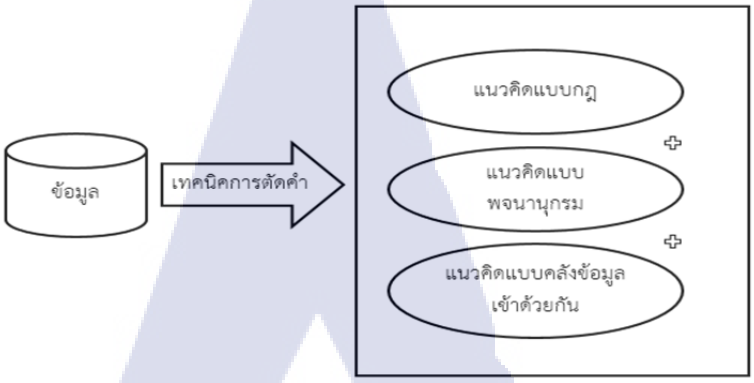
รูปที่ 3.3 ขั้นตอนการวิเคราะห์ข้อมุลปัญหาของงานวิจัย

3.1.1 ศึกษาข้อมูลและปัญหา

การศึกษาข้อมูลและปัญหาวิจัยในครั้งนี้ เป็นส่วนหนึ่งของการทำเหมืองข้อความให้ได้ผลลัพธ์ที่ดีต้องมากจากการศึกษาวิธีการวิเคราะห์รูปแบบการตัดคำ เช่น ลักษณะของคำลักษณะการเขียนของคำลักษณะความหมายของคำ เป็นต้น สิ่งเหล่านี้จะเป็นตัวบ่งบอกถึงกระบวนการการตัดคำที่ทำให้มีประสิทธิภาพของการวิเคราะห์ข้อมูล

3.1.2 วิเคราะห์แนวทางข้อมูลที่เกี่ยวข้อง

การเลือกหัวข้อวิจัยของปัญหาที่กล่าวมาจากข้างต้น ทำให้ผู้วิจัยเห็นถึงปัจจัยความสำคัญที่มีผลต่อการตัดคำเพื่อนำไปประยุกต์ใช้ในงานด้านการทำเหมืองข้อความ สำหรับการวิเคราะห์ข้อมูลที่เป็นในด้านภาษาไทยนั้น ในความเป็นจริงการระบุหน้าที่ของคำเพียงบอกว่าเป็นคำนาม คำกริยา คำสรรพนาม คำคุณศัพท์ คำวิเศษณ์ ฯลฯ นั้นไม่เพียงพอที่จะนำมาใช้งานการวิเคราะห์ ดังนั้นผู้วิจัยจึงศึกษาคิดค้นเทคนิคการตัดคำแบบผสมผสานเข้ามาปรับปรุงให้การทำงานในการคำตัดภาษาไทยมีความถูกต้องแม่นยำมากขึ้นกว่าเดิม โดยการผสมผสานระหว่าง แนวคิดแบบกฎ แนวคิดแบบพจนานุกรม และแนวคิดแบบคลังข้อมูลเข้าด้วยกัน



รูปที่ 3.4 เทคนิคการตัดคำแบบผสมผสาน

3.1.3 เสนอหัวข้อกับอาจารย์

จากการวิเคราะห์ข้อมูลและปัญหา ทฤษฎีการเรียนรู้กาศึกษางานวิจัยที่เกี่ยวข้องต่างๆ ผู้วิจัยได้นำทฤษฎีการตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่ ซึ่งทฤษฎีการตัดคำภาษาไทยนี้ยังมีข้อกำหนดบ้างอย่างที่ทำให้ผลลัพธ์การทำเหมืองข้อความมีประสิทธิภาพไม่มากเพียงพอ เช่น คำกำกวม คำเฉพาะ คำทับศัพท์จากภาษาอื่น หรือคำที่เขียนเหมือนกันแต่แปลได้หลากหลายความหมายหลังจากนั้นผู้วิจัยจึงได้รวบรวมข้อมูลที่ศึกษามาเสนอหัวข้อต่ออาจารย์ที่ปรึกษา โดยใช้ชื่อหัวข้อว่า “การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน”

3.2 ขั้นตอนที่ 2 การศึกษาข้อมูลและกลุ่มตัวอย่างที่ใช้ในงานวิจัย

การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน ใช้ข้อมูลจากเว็บไซต์อโกด้า (https://www.agoda.com/th-th/?cid=-218) ซึ่งเป็นข้อมูลการแสดงความคิดเห็นใช้บริการผ่านความรู้สึกที่มีการล็อกอินจำนวน 2,040 ความคิดเห็น จาก 10 จังหวัดที่นักท่องเที่ยวเที่ยวมากที่สุดในประเทศไทยจำนวน 40 โรงแรม โดยมีการเก็บข้อมูลการแสดงความคิดเห็นรายละเอียดดังตารางที่ 3.2

ตารางที่ 3.2 รายละเอียดการจัดเก็บข้อมูล

ที่	ชื่อคอลัมน์	หมายถึง
1	URL	ที่อยู่ของเว็บไซต์ที่ใช้ในการเก็บข้อมูล
2	ระดับคะแนน	ระดับความพึงพอใจของการใช้บริการ
3	ผู้ใช้งาน	ผู้ใช้งานที่ผ่านการล็อกอินของเว็บไซต์อโกด้า
4	วันที่	วันเดือนปี การแสดงความคิดเห็นในแต่ละครั้ง
5	เน้นหัวข้อ	การพูดถึงการในเรื่องนั้นๆ
6	คอมเม้น	การแสดงความคิดเห็นของผู้ใช้บริการ
7	โรงแรม	ชื่อโรงแรม
8	จังหวัด	สถานที่ตั้งของโรงแรม
9	เกรด	ระดับดาวของโรงแรม

ขั้นตอนการเก็บข้อมูลนั้นแบ่งแยกย่อยออกเป็นขั้น ๆ เพื่อให้เห็นถึงระดับของการเก็บข้อมูลเพื่อนำมาวิเคราะห์รูปแบบ ดังรูปที่ 3.5

รูปที่ 3.5 ขั้นตอนการเก็บข้อมูลที่ใช้ในงานวิจัย

ตารางที่ 3.3 ตัวอย่างข้อมูลที่จัดเก็บ

URL	ระดับคะแนน	ผู้ใช้	วันที่	เน้นหัวข้อ	คอมเมนต์	โรงแรม	จังหวัด	เกรด
https://www.agoda.com/th-th/siam-siam-design-hotel-pattaya	10	Nidhiprabha	รีวิวเมื่อ 27 กันยายน 2018	แนะนำให้มาพัก	พอดีวันนั้นไปพักแล้วที่พักที่แรกที่จองแบบไม่ไหวจริงๆ ไม่ขอเอ่ยนามนะคะ ก็เลยย้ายโรงแรมมาเป็นที่นี่แทน คือแบบ ประทับใจเวอร์มาก ทั้งห้องพักเองง ห้องน้ำเอง การบริการ อาหาร และวิวจากในห้อง โดยรวมคือดีแบบคุ้มมากกับราคาที่จ่ายไป คือชอบมาก ถ้ากลับไปพัทยาอีก ก็คงพักที่นี่แหละค่ะ ชอบมากจริงๆ	สยาม แอท สยาม ดีไซน์ โฮเต็ล พัทยา	ชลบุรี	4 ดาว

3.3 ขั้นตอนที่ 3 การวิเคราะห์และออกแบบกระบวนการทางเทคนิคการตัดคำแบบผสมผสาน

วิธีการในการตัดคำภาษาไทยมีขั้นตอนดังนี้ คือ การตัดอนุประโยคโดยอาศัยอักขระพิเศษและช่องว่าง การตรวจความหมายของคำบริเวณโดยรอบประโยคหรือข้อความ การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย การวิเคราะห์คำที่มีอยู่ในพจนานุกรม การเพิ่มคำลงให้คลังข้อมูล

3.3.1 อนุประโยคโดยอาศัยอักขระพิเศษและช่องว่าง

วิธีการนี้จะพิจารณาส่วนที่ติดกันหรือส่วนที่ใกล้กันของอักษรในภาษาไทย หากมีอักขระพิเศษตัวอักษรอื่นๆ หรือช่องว่างระหว่างคำที่ไม่ใช้อักษรภาษาไทยให้ถือว่าเป็นคนละประโยคหรือคนละข้อความกัน

3.3.2 การตรวจความหมายของคำบริเวณโดยรอบประโยคหรือข้อความ

เป็นการหาความหมายของคำโดยที่อยู่ในประโยคนั้นๆ หรืออยู่ในพยางค์นั้นๆ เพื่อบ่งบอกถึงคำที่ต้องการหาว่ามีความหมายไปในทิศทางเดียวกันกับประโยคหรือไม่ จะเป็นตัวช่วยในการตัดสินใจของการตัดคำให้มีประสิทธิภาพและได้ค่าความถูกต้องที่ชัดเจนแม่นยำ

3.3.3 การตัดคำโดยอาศัยกฎการผสมอักษรในภาษาไทย

สามารถแบ่งประเภทของพยัญชนะไทย 44 ตัว ออกเป็น 3 หมู่หรือเรียกว่า “ไตรยางค์” ได้แก่ อักษรสูง อักษรกลาง อักษรต่ำ การที่จะทำไปผสมเป็นระดับพยางค์หรือคำจะแบ่งได้ 3 แบบ ดังนี้

1) พยัญชนะ+สระ+วรรณยุกต์ เช่น ดำ ไก่

2) พยัญชนะ+สระ+วรรณยุกต์+ตัวสะกด เช่น คน ตน

3) พยัญชนะ+สระ+วรรณยุกต์+ตัวสะกด+ตัวการันต์ เช่น ฤทธิ์ แพทย์ ยักษ์

โดยกฎการแบ่งตามพยางค์จะได้ว่าสามารถตัดประโยคหรืออนุประโยคที่ประกอบด้วยอักษรน้อยกว่า 4 ตัวอักษรให้เป็น 1 คำหรือพยางค์ได้ทันทีโดยไม่ต้องเปรียบเทียบกับกฎหรือพจนานุกรม เนื่องจากพยางค์ที่สั้นที่สุดคือ แบบที่ 1) พยัชนะ+สระ+วรรณยุกต์ หากในกรณีที่วรรณยุกต์อยู่ในรูปสามัญจะประกอบด้วยอักษร 2 ตัว (กรณีนี้ไม่รวมถึง ธ ณ ฤ ที่จะมีช่องว่างอยู่หน้าและหลังเสมอ) นั่นหมายถึงหากจะเป็น 2 คำหรือ 2 พยางค์ขึ้นไปต้องประกอบไปด้วย 4 ตัวอักษรขึ้นไป นอกจากคำไทยแท้นี้แล้วภาษาไทยได้มีการรับภาษาต่างประเทศเข้ามาใช้ เช่น บาลี สันสกฤต อังกฤษ เป็นต้น ทำให้เกิดคำที่ไม่ตรงกับหลักการผสมคำอยู่มาก เช่น พรหม แลนด์ มาร์ค เลาน์จ ซึ่งในปัจจุบันนี้จะพบคำที่มาจากภาษาต่างประเทศมากขึ้นและมีการถอดความเป็นภาษาไทยไม่ตรงกับหลักไวยากรณ์ไทย

3.3.4 การวิเคราะห์คำที่มีอยู่ในพจนานุกรม

การตัดคำโดยอ้างอิงกับพจนานุกรม จะแบ่งเป็น 2 ส่วน ดังนี้

1) พบคำตามพจนานุกรม หากพบคำในพจนานุกรมมากกว่าสองครั้งจะถือเอาคำที่ยาวที่สุดเป็นหลักคำที่เก็บอยู่ในพจนานุกรมนี้เป็นคำที่พบได้ทั่วไปหรือคำที่มีความยาวหลายพยางค์หรือคำที่มีการสะกดตรงตามอักขรวิธี

2) ไม่พบตามพจนานุกรม โดยปรกติส่วนนี้หากเป็นคำเดี่ยว ๆ จะนำไปพิจารณากับคำรอบข้างซึ่งมีความเป็นไปได้ที่จะเป็นคำเดียวกันโดยอาศัยกฎการวิเคราะห์ความเป็นไปได้ที่จะเป็นการทับศัพท์ภาษาต่างประเทศหรือคำเฉพาะ คำเฉพาะบางส่วนจะถูกเก็บอยู่ในพจนานุกรมคำเฉพาะและสามารถปรับปรุงได้เพื่อความเหมาะสมกับข้อมูลแต่ละประเภท คำเฉพาะที่เหมาะสมได้แก่คำที่มาจากบาลี สันสกฤต ที่นำตัวอักษรเหล่านี้มาใช้ ฆ , ญ , ฒ , ฎ , ฏ , ฒ , ผ , ภ , ฌ , ศ , ษ ซึ่งมักไม่ค่อยพบอยู่กับคำที่มาจากการทับศัพท์ภาษาต่างประเทศ เป็นต้น เช่นชื่อคน ชื่อสถานที่ ชื่อเรียกเฉพาะ

3.3.5 การเพิ่มคำลงให้คลังข้อมูล

เมื่อตรวจสอบคำจากการเทียบแบบกฎและแบบพจนานุกรมแล้ว คำๆ นั้นไม่มีความหมายหรือมีความหมายแต่ไม่ได้ถูกเก็บในพจนานุกรม จึงต้องทำการนำคำนั้นๆ มาเพิ่มลงในฐานของคลังข้อมูลเพื่อให้มีคำและความหมายที่จะใช้ในการตัดคำได้อย่างสมบูรณ์แบบ แต่จะต้องพิจารณาจากพจนานุกรมก่อนเสมอ

3.4 ขั้นตอนที่ 4 การประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็น

การแบ่งข้อมูลทดสอบแบ่งโดยใช้วิธีการประเมินผล K-fold Cross Validation กำหนดให้ K-fold เป็น 3 กลุ่ม ได้แก่ k=10, k=15, k=20 สำหรับการทดสอบข้อมูลจะใช้โปรแกรม Rapid Miner แล้วเลือกค่า k ที่ให้ผลลัพธ์ดีที่สุด จากนั้นหาค่าเฉลี่ยการวัดประสิทธิภาพของแต่ละเทคนิควิธีนำแบบจำลองของแต่ละเทคนิควิธีที่ให้ประสิทธิภาพดีที่สุดมาพัฒนา โดยใช้การวัดประสิทธิภาพโดยการหาค่าความเที่ยง (Precision) ค่าความลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่า F-Measure ดังสมการที่ 9, 10, 11 และ 12 ตามลำดับ

$$\text{Precision} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (11)$$

$$F = \frac{2 \times Precision \times Recall}{Precision \times Recall} \quad (12)$$

โดยที่ TP คือ จำนวนข้อมูลที่ถูกดึงออกมาอย่างถูกต้อง
FP คือ จำนวนข้อมูลที่ผิดพลาดที่ถูกดึงออกมา
TN คือ จำนวนข้อมูลที่ถูกต้องแต่ไม่ถูกดึงออกมา
FN คือ จำนวนข้อมูลที่ผิดพลาดแต่ไม่ถูกดึงออกมา

3.5 ขั้นตอนที่ 5 การสรุปผลการวิจัย

ในส่วนของการสรุปผลการวิจัย จะกล่าวถึง ผลที่ได้จากการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน รวมทั้งประสิทธิภาพของโมเดลว่าตรงตามสมมติฐานที่ได้ให้ไว้ในบทที่ 1 หรือไม่ โดยการหาค่าความถูกต้อง จากสมการนี้

$$\%Accuracy = 100 - \%Error \quad (13)$$

โดยที่ %Error = Relative Error x 100
และ Relative Error = + |xmea-xtxt|
เมื่อ xmea คือ ค่าที่ได้จากการวัด (Measure Value)
xt คือ ค่าจริง (True Value)

บทที่ 4
ผลการวิจัย

การศึกษาวิจัยเรื่องการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน มีวัตถุประสงค์เพื่อพัฒนาเทคนิคการตัดคำแบบผสมผสานเพื่อวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม ผู้วิจัยจึงได้พัฒนาเทคนิคการตัดคำแบบผสมผสานขึ้นมาให้ง่ายต่อการนำไปวิเคราะห์หารูปแบบความคิดเห็นของข้อมูลโรงแรม พร้อมทั้งประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็นโดยการจำแนกแบบต้นไม้ตัดสินใจแบบนาอีฟ เบย์และแบบเคเนียเรสเนเบอร์ โดยสามารถแบ่งผลการดำเนินการและการวิเคราะห์ข้อมูล ดังต่อไปนี้

4.1 เทคนิคการตัดคำ

4.2 ผลการตัดคำด้วยเทคนิคผสมผสาน

4.3 ปัจจัยที่มีผลต่อการตัดสินใจเลือกโรงแรม

4.4 ผลการประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็น

4.1 เทคนิคการตัดคำ

เทคนิคการตัดคำนี้ได้นำเอกสารข่าวจากหนังสือพิมพ์จำนวน 600 คำ หรือ 1 คอลัมน์ข่าว เป็นเกณฑ์ในการวัดประสิทธิภาพของแต่ละเทคนิคดังตารางที่ 4.1 ต่อไปนี้

วิธีที่ 1 เทคนิคตัดคำด้วยกฎ เป็นการแบ่งในส่วนของอนุประโยคอาศัยช่องว่างระหว่างตัวอักษร อักขระพิเศษ และใช้การตัดคำที่อาศัยหลักการทางภาษาไทยเข้ามาเป็นตัวประกอบ ได้แก่ พยัญชนะ วรรณยุกต์ สระ ตัวสะกด ซึ่งวิธีนี้ตัดคำได้ถูกต้องและมีความหมายถึง 422 คำ ส่วนข้อผิดพลาดที่ตัดคำแล้วไม่ถูกต้องเกิดจากมีตัวอักษรพยัญชนะเกิดและคิดเป็นเปอร์เซ็นดังวิธีต่อไปนี้

$$(422 \times 100) / 600 = 42{,}200 / 600$$
$$= 70.33\%$$

วิธีที่ 2 เทคนิคตัดคำด้วยกฎ + พจนานุกรม เป็นการตัดคำที่อาศัยการตัดอักขระพิเศษ เช่น (,) , “ , ‘ , - , / ออกจากประโยคทั้งหมดก่อนเพื่อทำให้ประโยคมีข้อความที่สมบูรณ์ขึ้น หลังจากนั้นนำประโยคทั้งหมดที่ได้ไปทำการตัดคำโดยอาศัยการเทียบคำในพจนานุกรมหากมีคำที่ตรงตามพจนานุกรมก็จะได้คำออกมาได้อย่างถูกต้อง ซึ่งวิธีนี้ตัดคำได้ถูกต้องและมีความหมายถึง 458 คำ และคิดเป็นเปอร์เซ็นต์ได้ดังวิธีต่อไปนี้

(458 x 100) / 600 = 45,800 / 600

= 76.33%

วิธีที่ 3 เทคนิคตัดคำด้วยกฎ +คลังข้อมูล เป็นการตัดคำที่อาศัยการตัดอักขระพิเศษ เช่น (,) , “ , ‘ , - , / ออกจากประโยคทั้งหมดก่อนเพื่อทำให้ประโยคมีข้อความที่สมบูรณ์ขึ้นเหมือนกับวิธีที่ 2 แต่วิธีนี้แตกต่างตรงใช้คลังข้อมูลในหมวดหมู่โดเมนของข้อมูลนั้นๆ และการนับความถี่ของคำเข้ามาช่วยในการตัดคำ ข้อเสียของการตัดคำประเภทนี้ คือ หากมีคำศัพท์ในคลังข้อมูลน้อยเกินไปจะทำให้การตัดคำมีความผิดพลาดสูง ซึ่งวิธีนี้ตัดคำได้ถูกต้องและมีความหมายถึง 419 คำ และคิดเป็นเปอร์เซ็นต์ได้ดังวิธีต่อไปนี้

(419 x 100) / 600 = 41,900 / 600

= 69.83%

วิธีที่ 4 เทคนิคตัดคำด้วยพจนานุกรม + คลังข้อมูล เป็นการตัดคำที่อาศัยพจนานุกรมร่วมกับคลังข้อมูลโดยเน้นการตัดคำจากพจนานุกรมก่อนเพื่อที่จะให้ได้ความหมายที่ถูกต้องตรงตามหลักภาษาไทย หลังจากนั้นหากพบคำที่ไม่สามารถตัดคำได้จึงนำไปเปรียบเทียบกับคลังข้อมูลเพื่อนับความถี่ของคำต่อไป ซึ่งวิธีนี้ตัดคำได้ถูกต้องและมีความหมายถึง 415 คำ และคิดเป็นเปอร์เซ็นต์ได้ดังวิธีต่อไปนี้

(415 x 100) / 600 = 41,500 / 600

= 69.16%

วิธีที่ 5 เทคนิคการตัดคำด้วยพจนานุกรม + กฎ + คลังข้อมูล การตัดคำวิธีนี้เป็นการตัดคำที่เทียบโดยพจนานุกรมจากนั้นนำมาตัดอักขระพิเศษ เช่น (,) , “ , ‘ , - , / ออกทีหลัง หากคำไหนที่ไม่พบในพจนานุกรมก็จะทำการเปรียบเทียบคำศัพท์ในคลังข้อมูลต่อไป วิธีนี้มีข้อเสีย คือ หากประโยคที่จะตัดนั้นมีอักขระแฝงอยู่ด้วย เช่น ข้อดี-ข้อเสียของโรงแรม เป็นต้น จะทำให้การตัดคำนั้นไม่ถูกต้องทันที ซึ่งวิธีนี้ตัดคำได้ถูกต้องและมีความหมายถึง 436 คำ และคิดเป็นเปอร์เซ็นต์ได้ดังวิธีต่อไปนี้

(436 x 100) / 600 = 43,600 / 600

= 72.66%

วิธีที่ 6 เทคนิคการตัดคำด้วยกฎ + พจนานุกรม + คลังข้อมูล (ผสมผสาน) วิธีนี้จะเป็นการตัดคำที่อาศัยการตัดคำเป็นลำดับขั้นตอนจากกฎไปพจนานุกรมไปคลังข้อมูล ในส่วนของกฎจะทำการตัดอักขระพิเศษ เช่น (,) , “ , ‘ , - , / ออกจากประโยคทั้งหมดก่อนเพื่อทำให้ประโยคมีข้อความที่สมบูรณ์ขึ้น จากนั้นจึงนำประโยคที่ผ่านกระบวนการขั้นแรกไปทำการตัดคำกับพจนานุกรมจะทำการตัดคำตามหลักการทางภาษาไทยเข้ามาเป็นตัวประกอบ ได้แก่ พยัญชนะ วรรณยุกต์ สระ ตัวสะกด จะตัดจากซ้ายไปขวาเสมอโดยเลือกการตัดคำแบบการเทียบคำที่ยาวที่สุด หากคำไหนที่ไม่พบในพจนานุกรมก็จะส่งไปเข้าสู่กระบวนการเทียบคำของคลังข้อมูลโดยจะมีหมวดหมู่โดเมนคำศัพท์ของดาต้านั้นพร้อมทั้งนับความถี่คำที่เทียบเอาไว้ด้วย ซึ่งวิธีนี้ตัดคำได้ถูกต้องและมีความหมายถึง 507 คำ และคิดเป็นเปอร์เซ็นต์ได้ดังวิธีต่อไปนี้

$$(507 \times 100) / 600 = 50{,}700 / 600$$
$$= 84.50\%$$

ตารางที่ 4.1 เปรียบเทียบประสิทธิภาพเทคนิคการตัดคำแต่ละวิธี

วิธีที่	เทคนิค	ประสิทธิภาพ
1.	กฎ	70.33%
2.	กฎ + พจนานุกรม	76.33%
3.	กฎ + คลังข้อมูล	69.83%
4.	พจนานุกรม + คลังข้อมูล	69.16%
5.	พจนานุกรม + กฎ + คลังข้อมูล	72.66%
6.	กฎ + พจนานุกรม + คลังข้อมูล (ผสมผสาน)	84.50%

จะเห็นได้ว่าแต่ละวิธีจะมีค่าเปอร์เซ็นต์ที่แตกต่างกันออกไป ซึ่งในการทดลองครั้งนี้สามารถสรุปสรุปได้ว่า การทดลองของวิธีที่ 1 ยังมีข้อจำกัดในเรื่องของการตัดที่มีตัวอักษร สระ วรรณยุกต์ที่เกินมาจึงทำให้ตัดคำได้แต่ไม่สามารถหาความหมายของคำนั้นๆ ได้วิธีนี้จึงไม่นิยมนำมาตัดคำเพราะยิ่งประโยคข้อความยาวมากๆ การตัดคำยิ่งลดประสิทธิภาพลงรวมถึงไม่สามารถแก้ไขคำกำกวน คำทับศัพท์คำแสลงได้ การทดลองของวิธีที่ 2 สามารตัดคำได้ถูกต้องมากยิ่งขึ้นกว่าวิธีที่ 1 แต่ยังมีข้อจำกัดในส่วนของพจนานุกรมหากในพจนานุกรมไม่พบคำศัพท์ก็ไม่สามารถตัดข้อความได้เช่นกันและพจนานุกรมต้องอัพเดทตลอดเวลา เนื่องด้วยปัจจุบันมีคำศัพท์ใหม่ๆ ที่มีการเกิดขึ้นตลอดเวลาจึงเป็นสิ่งจำเป็นที่จะต้องปรับปรุงในส่วนนี้ให้ดีขึ้น ส่วนการการทดลองวิธีที่ 3, 4 พบปัญหาลักษณะเดียวกันหากมีคำศัพท์

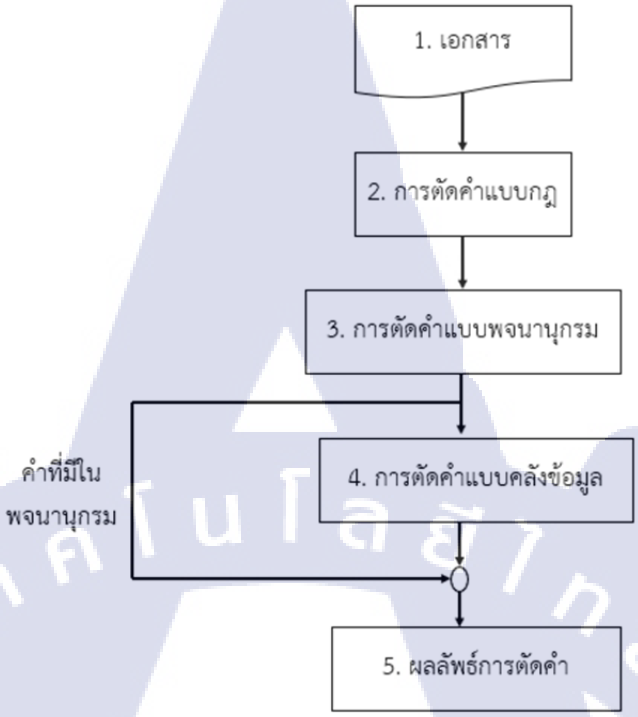
ในคลังข้อมูลน้อยเกินไปจะทำให้การตัดคำมีความผิดพลาดสูง การทดลองวิธีที่ 5 วิธีนี้เราทำการตัดคำโดยแบบพจนานุกรมก่อนแล้วตามด้วยแบบกฎแบบคลังข้อมูลตามลำดับ พบว่าการตัดคำวิธีนี้มีข้อเสียหากประโยคที่จะตัดนั้นมีอักขระแฝงอยู่ด้วย ตัวอย่างเช่น “เวลาเช้า-บ่าย(พนักงานจะนำอาหารว่างมาให้ที่ห้องพัก)” จะทำให้การตัดคำไม่ถูกต้องทันที ซึ่งการทดลองวิธีที่ 6 หรือเรียกว่า “เทคนิคตัดคำผสมผสาน” สามารถแก้ไขปัญหาในส่วนของคำที่เกิดจากคำกำกวน คำทับศัพท์ คำแสลงได้ดี เพราะในการตัดคำวิธีที่ 1-5 ไม่สามารถแก้ไขในส่วนนี้ได้จึงทำให้วิธีที่ 6 นี้สามารถตัดคำได้สมบูรณ์ยิ่งขึ้น

จากตารางที่ 4.1 จะเห็นได้ว่าเทคนิคที่มีประสิทธิภาพมากที่สุดในแต่ละวิธี ได้แก่ วิธีที่ 6 เทคนิคแบบ “กฎ + พจนานุกรม + คลังข้อมูล” หรือเรียกว่า “เทคนิคผสมผสาน” มีประสิทธิภาพ 84.50% และเทคนิคที่ให้ประสิทธิภาพน้อยที่สุดในแต่ละวิธี ได้แก่ วิธีที่ 4 เทคนิคแบบ “พจนานุกรม + คลังข้อมูล” มีประสิทธิภาพ 69.16%

4.2 ผลการตัดคำด้วยเทคนิคผสมผสาน

4.2.1 การตัดคำ

ในงานวิจัยครั้งนี้ได้เลือกใช้กระบวนการตัดคำภาษาไทยด้วยวิธีการเทคนิคผสมผสาน ได้แก่ แบบกฎ แบบพจนานุกรม และแบบคลังข้อมูล พิจารณาการวนลูปจากซ้ายไปขวาตามหลักกฎภาษาไทยเพื่อที่จะนำไปเปรียบเทียบกับพจนานุกรมและเก็บเป็นคลังข้อมูลไว้ตามลำดับ โดยจะแสดงผลลัพธ์การตัดคำดังตารางที่ 4.2 ต่อไปนี้ และขั้นตอนวิธีการตัดคำเทคนิคผสมผสาน ดังรูปที่ 4.1 โดยงานวิจัยชิ้นนี้ถูกพัฒนาขึ้นด้วยภาษา PHP ซึ่งตัวอย่างหน้าจอแสดงการทำงาน ดังรูปที่ 4.2



รูปที่ 4.1 ขั้นตอนวิธีการตัดคำเทคนิคผสมผสาน

จากรูปภาพที่ 4.1 สามารถออกแบบขั้นตอนวิธีการตัดคำเทคนิคผสมผสานเพื่อนำไปเขียนโค้ดโปรแกรมได้ดังนี้

ขั้นตอนที่ 1 เริ่มต้นนำเอกสารเข้าสู่การตัดคำ

ขั้นตอนที่ 2 ตัดคำแบบกฎโดยแยกอักขระพิเศษออกจากประโยค เช่น (,) , “ , ‘ , - , / และตัดคำอาศัยหลักการในภาษาไทย

ขั้นตอนที่ 3 ตัดคำกับพจนานุกรมจะทำการตัดคำตามหลักการทางภาษาไทยเข้ามาเป็นตัวประกอบ ได้แก่ พยัญชนะ วรรณยุกต์ สระ ตัวสะกด จะตัดจากซ้ายไปขวาเสมอโดยเลือกการตัดคำแบบการเทียบคำที่ยาวที่สุด หลังจากตัดคำจากผ่านพจนานุกรมแล้วคำไหนที่มีตรงตามพจนานุกรมก็จะแสดงไปในขั้นของผลลัพธ์การตัดคำ ส่วนคำไหนที่ไม่มีในพจนานุกรมก็จะทำการส่งไปยังส่วนของของการตัดคำแบบคลังข้อมูล

ขั้นตอนที่ 4 การตัดคำแบบคลังข้อมูลจะรับคำที่ไม่ได้ตัดจากพจนานุกรมมาทำการเปรียบเทียบคำศัพท์ในคลังข้อมูลโดยมีหมวดหมู่โดเมนตามเรื่องนั้นๆ ที่ทำการตัดคำ หากพบว่าคำที่จะตัดมีตามคำศัพท์ในคลังข้อมูลก็จะส่งข้อมูลไปยังผลลัพธ์การตัดคำพร้อมทั้งนับความถี่ของคำ

ขั้นตอนที่ 5 ผลลัพธ์การตัดคำ โดยเทคนิคการตัดคำแบบผสมผสาน

รูปที่ 4.2 ตัวอย่างหน้าจอแสดงการทำงาน

ตารางที่ 4.2 ผลลัพธ์การตัดคำเทคนิคผสมผสาน

ตัวอย่างข้อความ	การตัดคำ
พนักงานบริการดี ยิ้มแย้มแจ่มใส ห้องน้ำสะอาด มีความส่วนตัวสูง ช่องทางเดินกว้าง มีมุมถ่ายรูปเยอะ	พนักงาน | บริการดี | ยิ้มแย้มแจ่มใส | ห้อง | น้ำ สะอาด | มี | ความ | ส่วนตัว | สูง | ช่องทางเดิน | กว้าง | มี | มุม | ถ่ายรูป | เยอะ
อาหารไม่คุ้มค่ากับราคาที่พักเลย	อาหาร | ไม่ | คุ้มค่า | กับ | ราคา | ที่พัก | เลย
ห้องสะอาด เดินทางง่าย เหมาะกับคนที่บินไฟลท์ดึกมากๆ หรือเช้ามากๆ พักผ่อนก่อนออกเดิน	ห้อง | สะอาด | เดินทาง | ง่าย | เหมาะกับ | คน | ที่ | บิน | ไฟลท์ | ดึก | มาก | ๆ | หรือ | เช้า | มาก | ๆ | พักผ่อน | ก่อน | ออกเดินทาง
ห้องที่ทำการจองแบบ 3 เตียงแคบมาก ต้องชำระเงินเพิ่มกับทางโรงแรมและขอย้ายห้องค่อนข้างเสียงดังทั้งห้องข้างเคียงและชั้นบน	ห้อง | ที่ทำการ | จอง | แบบ | 3 | เตียง | แคบ | มาก | ต้อง | ชำระเงิน | เพิ่ม | กับ | ทาง | โรงแรม | และ | ขอ | ย้าย | ห้อง | ค่อนข้าง | เสียง | ดัง | ทั้ง | ห้อง | ข้างเคียง | และ | ชั้นบน
อยู่ใกล้รถไฟฟ้า	อยู่ | ใกล้ | รถไฟฟ้า

4.2.2 การวิเคราะห์ความรู้สึก

งานวิจัยชิ้นนี้ได้แบ่งการวิเคราะห์ความรู้สึกออกเป็น 3 ด้าน ได้แก่ เชิงบวก (Positive) เชิงลบ (Negative) และเป็นกลาง (Neutral) เพื่อนำไปประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็น ซึ่งให้ผลลัพธ์การวิเคราะห์ความรู้สึก ดังตารางที่ 4.3 ต่อไปนี้

การจะแยกความรู้สึกออกเป็นเชิงบวก (Positive) เชิงลบ (Negative) และเป็นกลาง (Neutral) ต้องอาศัยการตัดประโยคออกจากกันให้เป็นคำ จากนั้นจึงแยกคำออกตามลักษณะคำในภาษาไทย เช่น คำนาม คำสรรพนาม คำกิริยา คำวิเศษณ์ คำบุพบท คำสันธาน เพื่อที่จะทำการกำกับคำด้วยหน้าที่ของคำตามคลังคำที่เก็บไว้เป็นการกำหนดคุณลักษณะและสร้างคำแสดงความรู้สึกที่เป็นเชิงบวก (Positive) เชิงลบ (Negative) ออกมาดังรูปที่ 4.3 ต่อไปนี้

รูปที่ 4.3 ตัวอย่างการกำกับคุณลักษณะการแยกความรู้สึก

ตารางที่ 4.3 ผลลัพธ์การวิเคราะห์ความรู้สึก

ตัวอย่างข้อความ	การตัดคำ	ผลวิเคราะห์ความรู้สึก
พนักงานบริการดี ยิ้มแย้ม แจ่มใส ห้องน้ำสะอาด มีความส่วนตัวสูง ช่องทางเดินกว้าง มีมุมถ่ายรูปเยอะ	พนักงาน | บริการดี | ยิ้มแย้มแจ่มใส | ห้อง | น้ำสะอาด | มี | ความ | ส่วนตัว | สูง | ช่องทางเดิน | กว้าง | มี | มุม | ถ่ายรูป | เยอะ	Positive
อาหารไม่คุ้มค่ากับราคาที่พักเลย	อาหาร | ไม่ | คุ้มค่า | กับ | ราคา | ที่พัก | เลย	Negative
ห้องสะอาด เดินทางง่าย เหมาะกับคนที่บินไฟลท์ดึกมากๆ หรือเช้ามากๆ พักผ่อนก่อนออกเดิน	ห้อง | สะอาด | เดินทาง | ง่าย | เหมาะกับ | คน |ที่ | บิน | ไฟลท์ | ดึก | มาก | ๆ | หรือ | เช้า | มาก | ๆ | พักผ่อน | ก่อน | ออกเดินทาง	Positive

ตารางที่ 4.3 ผลลัพธ์การวิเคราะห์ความรู้สึก (ต่อ)

ตัวอย่างข้อความ	การตัดคำ	ผลวิเคราะห์ความรู้สึก
ห้องที่ทำการจองแบบ 3 เตียง แคบมาก ต้องชำระเงินเพิ่มกับทางโรงแรมและขอย้ายห้อง ค่อนข้างเสียงดังทั้งห้อง ข้างเคียงและชั้นบน	ห้อง | ที่ทำการ | จอง | แบบ | 3 | เตียง | แคบ | มาก | ต้อง | ชำระเงิน | เพิ่ม | กับ | ทาง | โรงแรม | และ | ขอ | ย้าย | ห้อง | ค่อนข้าง | เสียง | ดัง | ทั้ง | ห้อง | ข้างเคียง | และ | ชั้นบน	Negative
อยู่ใกล้รถไฟฟ้า	อยู่ | ใกล้ | รถไฟฟ้า	Neutral

จากผลการศึกษาการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสานได้รวบรวมข้อมูลจากเว็บไซต์อโกด้า Agoda (https://www.agoda.com/th-th/) ผู้วิจัยจึงได้เก็บรวบรวมข้อมูลความคิดเห็นของโรงแรมที่จะนำมาใช้ในการวิเคราะห์ ซึ่งสรุปรายละเอียดข้อมูลทั้งหมดได้จากตารางที่ 4.4 ดังนี้

ตารางที่ 4.4 สรุปข้อมูลที่ใช้วิเคราะห์

ข้อมูล	จำนวน	หน่วย
เลือกจังหวัด	10	จังหวัด
เลือกโรงแรม	40	โรงแรม
พจนานุกรม (Dictionary)	76314	คำ
ความคิดเห็น	2040	ข้อความ
ความคิดเห็นเชิงบวก (Positive)	1336	ข้อความ
ความคิดเห็นเชิงลบ (Negative)	344	ข้อความ
ความคิดเห็นที่เป็นค่ากลาง (Neutral)	360	ข้อความ

ผลการศึกษาการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสานได้รวบรวมข้อมูลจากเว็บไซต์อโกด้า ผู้วิจัยได้สรุปรายละเอียดคำถามที่พบบ่อย 5 อันดับแรก ที่มีสถิติการพบคำศัพท์ในข้อความสูงสุดของผลการวิเคราะห์ความรู้สึกโดยแบ่งออกเป็นความรู้สึกเชิงบวก (Positive) เชิงลบ (Negative) และเป็นกลาง (Neutral) ดังตารางที่ 4.5 ถึง 4.7 ตามลำดับต่อไปนี้

ตารางที่ 4.5 ผลการวิเคราะห์ความรู้สึกเชิงวก (Positive) 5 อันดับแรก

ลำดับ	คำศัพท์	จำนวน	ตัวอย่างข้อความ
1.	บริการดี	614	พนักงานบริการดี ยิ้มแย้มแจ่มใส ห้องน้ำสะอาด มีความส่วนตัวสูง ช่องทางเดินกว้าง มีมุมถ่ายรูปเยอะ
2.	ข้อดี	267	ข้อดี พนักงานบริการดี การทำความสะอาดห้องสะอาด อาหารอร่อยมีหลากหลาย
3.	ประทับใจ	166	ประทับใจในบริการของพนักงานและสถานที่สวยมากๆ ห้องสวยสะอาด วิวสระน้ำสวยมองได้รอบพัทยา
4.	เยี่ยม	45	บริการยอดเยี่ยม
5.	คุ้มค่า	32	คุ้มค่าเงินที่จ่าย

ตารางที่ 4.6 ผลการวิเคราะห์ความรู้สึกเชิงลบ (Negative) 5 อันดับแรก

ลำดับ	คำศัพท์	จำนวน	ตัวอย่างข้อความ
1.	ข้อเสีย	156	ข้อเสีย บรรยากาศ Lobby ในโรงแรมค่อนข้างมืดไปนิดนึง
2.	แย่	51	บริการแย่มากๆ พนักงานไม่มีจิตบริการ
3.	ไม่คุ้มค่า	49	อาหารไม่คุ้มค่ากับราคาที่พักเลย
4.	เหม็น	21	ห้องเหม็นอับ ผ้าเช็ดตัวก็เหม็น ดูห้องเก่ากว่าในภาพเยอะเลย
5.	ไม่สุภาพ	4	พนักงานไม่สุภาพ

ตารางที่ 4.7 ผลการวิเคราะห์ความรู้สึกเป็นกลาง (Neutral) 5 อันดับแรก

ลำดับ	คำศัพท์	จำนวน	ตัวอย่างข้อความ
1.	ทำเล	112	ทำเลที่ตั้ง
2.	บีทีเอส	36	ใกล้บีทีเอส
3.	ห้องน้ำ	30	มีห้องน้ำ 1 ห้องไม่มีห้องน้ำในห้องนอน
4.	ขนาด	29	ขนาดห้องพัก
5.	ซื้อของ	2	ทำเลเหมาะกับมาเที่ยวกับซื้อของ แต่ห้องพักเก่าหน่อย

4.3 ปัจจัยที่มีผลต่อการตัดสินใจเลือกโรงแรม

การนับจำนวนการถูกกล่าวถึงและการให้ค่าน้ำหนักของปัจจัยย่อยในแต่ละด้านใช้หลักการ (จำนวการถูกกล่าวถึง * 100) \ จำนวนความคิดเห็นทั้งหมด) การกำหนดค่าน้ำหนักนี้เพื่อเป็นตัวบ่งชี้ว่าแต่ละปัจจัยมีค่าน้ำหนักไปในด้านบวกหรือด้านลบ เนื่องจากปัจจัยย่อยที่ประกอบอยู่ในปัจจัยหลักแต่ละด้าน มีจำนวนตั้งแต่ 1-4 ชนิด หากไม่ได้กำหนดค่าน้ำหนักของปัจจัยย่อยและให้ผลรวมค่าน้ำหนักในการสนับสนุนการตัดสินใจอาจทำให้การวิเคราะห์ไม่มีความแน่นอน ดังนั้นจึงจำเป็นต้องกำหนดค่าน้ำหนักให้แต่ละปัจจัย ได้แก่ ด้านสภาพห้องพัก ด้านที่ตั้งและสภาพแวดล้อมโรงแรม ด้านการให้บริการ ด้านสิ่งอำนวยความสะดวก และด้านราคา ดังตารางที่ 4.8 ถึง 4.12

ตารางที่ 4.8 รายละเอียดของปัจจัยด้านสภาพห้องพักและค่าน้ำหนัก

ด้านสภาพห้องพัก		
ปัจจัยย่อย	**รายละเอียด**	**ค่าน้ำหนัก**
ความพึงพอใจโดยรวม	ความสะดวกสบาย ความเก่าใหม่ การตกแต่ง	0.48
ความสะอาด	ความสะอาดของห้องพัก ห้องน้ำ และอุปกรณ์ต่างๆ ภายในห้องพัก	0.44
ขนาดห้องพัก	ขนาดห้องพักและพื้นที่ใช้งาน	0.08

ตารางที่ 4.9 รายละเอียดของปัจจัยด้านที่ตั้งและสภาพแวดล้อมโรงแรมและค่าน้ำหนัก

ด้านที่ตั้งและสภาพแวดล้อมโรงแรม		
ปัจจัยย่อย	**รายละเอียด**	**ค่าน้ำหนัก**
ตำแหน่งที่ตั้งของโรงแรม	แหล่งท่องเที่ยว แหล่งธุรกิจ ห้างสรรพสินค้า ร้านค้า	0.32
บริการสาธารณะ	การเดินทางโดยใช้การบริการของโรงแรม หรือใช้บริการสาธารณะ	0.23
สภาพแวดล้อมในโรงแรม	บรรยากาศภายในโรงแรม ทางเดิน การตกแต่งสวน บริเวณรอบๆ (พื้นที่นอกห้องพัก)	0.45

ตารางที่ 4.10 รายละเอียดของปัจจัยด้านการใช้บริการและค่าน้ำหนัก

ด้านการให้บริการ		
ปัจจัยย่อย	**รายละเอียด**	**ค่าน้ำหนัก**
การบริกาของพนักงาน	อัธยาศัยดียิ้มแย้มแจ่มใส่ ความพึงพอใจต่อการปฏิบัติหน้าที่ของพนักงาน	0.77
เช็คอิน-เช็คเอาท์	ความรวดเร็วของการเช็คอิน-เช็คเอาท์	0.23

ตารางที่ 4.11 รายละเอียดของปัจจัยด้านสิ่งอำนวยความสะดวกและค่าน้ำหนัก

ด้านสิ่งอำนวยความสะดวก		
ปัจจัยย่อย	**รายละเอียด**	**ค่าน้ำหนัก**
อาหาร	ความอร่อยของรสชาติอาหาร ความร้อน-เย็นของอาหาร ความหลากหลายของเมนูอาหาร	0.67
สิ่งอำนวยความสะดวกอื่นๆ	สระว่ายน้ำ ฟิตเนส ร้านขายของภายในโรงแรม	0.25
อินเตอร์เน็ต	ความเร็วของอินเตอร์เน็ต การบริการฟรี wifi ความเสถียรของอินเตอร์เน็ตภายในโรงแรม	0.08

ตารางที่ 4.12 รายละเอียดของปัจจัยด้านราคาและค่าน้ำหนัก

ด้านราคา		
ปัจจัยย่อย	**รายละเอียด**	**ค่าน้ำหนัก**
ราคาห้องพัก	เฉพาะราคาห้องพัก	0.64
ค่าสินค้าและบริการอื่น ๆ	ค่าบริการต่าง ๆ นอกเหนือจากข้อกำหนด	0.36

4.4 ผลการประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็น

ผู้วิจัยได้นำข้อความที่แสดงความคิดเห็นบนเว็บไซต์อโกด้า มาวิเคราะห์ความรู้สึก (Sentiment Analysis) โดยแบ่งออกเป็น 3 กลุ่ม คือ ความรู้สึกเชิงบวก (Positive) ความรู้สึกเชิงลบ (Negative) และความรู้สึกเป็นกลาง (Neutral) ในงานวิจัยนี้มีการทดสอบประสิทธิภาพด้วยการวัดค่าความแม่นยำเพื่อทดสอบระดับความถูกต้องของข้อมูล ผู้วิจัยจึงเลือกกระบวนการทดสอบโดยใช้โปรแกรม Rapid Miner Studio ในการช่วยวิเคราะห์ความถูกต้องแม่นยำของการวิเคราะห์รูปแบบความคิดเห็น

4.4.1 Rapid Miner Studio Performance Test

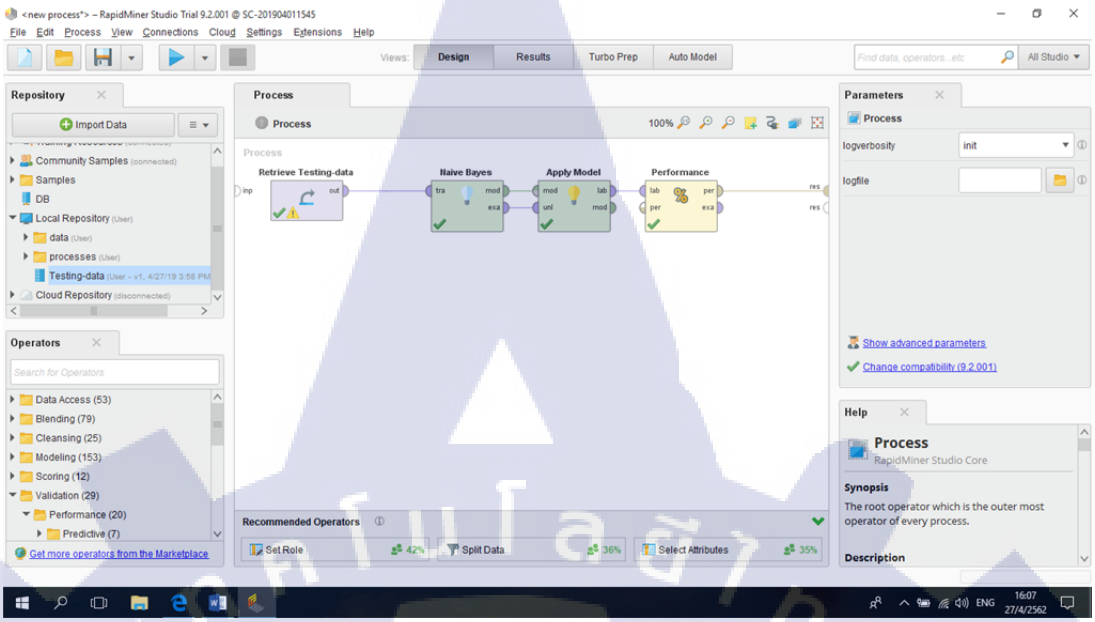
4.4.1.1 Import Data นำข้อมูลที่ได้เข้าโปรแกรมเพื่อใช้ในการวิเคราะห์ด้วยโปรแกรม Rapid Miner Studio ตรวจสอบข้อมูลเบื้องต้นหลังจากนั้นเลือกที่ปุ่ม “Next” เพื่อเปลี่ยนเข้าสู่หน้าจอการเลือกโฟลเดอร์ในการจัดเก็บไฟล์ข้อมูลหลังจากเลือกโฟลเดอร์จัดเก็บข้อมูลเรียบร้อยแล้วให้กดปุ่ม “Finish” จะทำการนำเข้าข้อมูลสำเร็จ

4.4.1.2 เลือกข้อมูลผ่านปุ่ม “Add Data” เพื่อเลือกการนำเข้าของชุดข้อมูล โปรแกรมจะแสดงหน้าจอให้กดเลือก “My Computer” เพื่อเลือกชุดข้อมูลที่อยู่ภายในเครื่องเลือก “Last Directory” ในหัวข้อ “Bookmarks” จากนั้นหน้าจอจะแสดงให้เลือกไฟลด์ที่ทำการนำเข้ามาเพื่อทำการตรวจสอบ กดเลือกไฟล์ที่จัดเก็บไว้จากนั้นกดปุ่ม “Next” ต่อไป

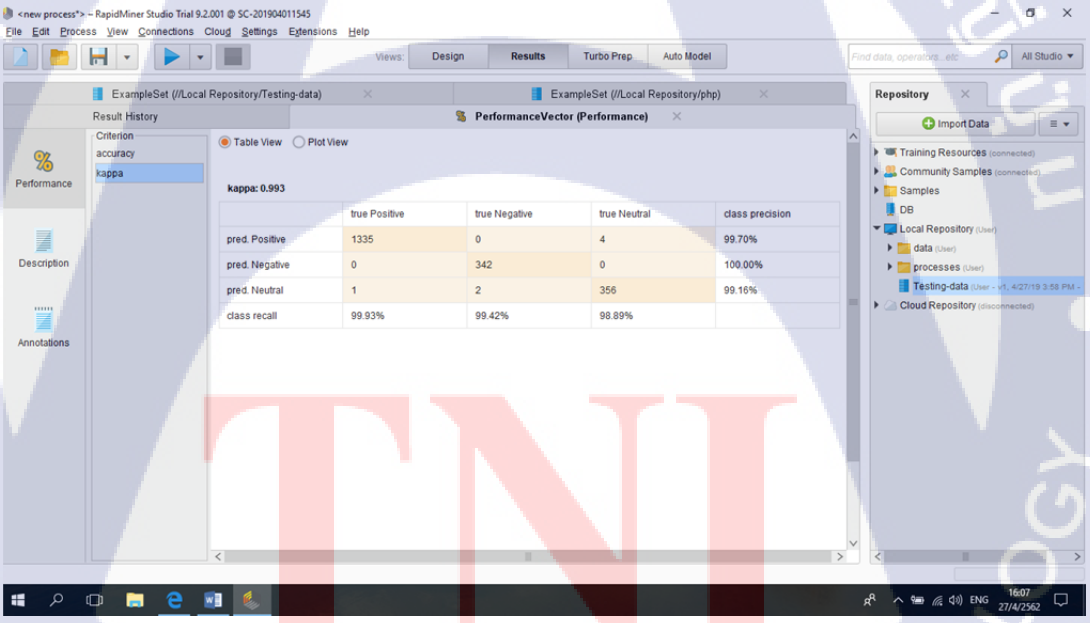
4.4.1.3 จัดรูปแบบไฟล์ หลังจากที่นำไฟล์เข้ามาเก็บไว้ในตัวโปรแกรมจะมีการแสดงตัวอย่างของชุดข้อมูลที่เลือก พร้อมตรวจสอบเบื้องต้นให้ทำการกดปุ่ม “Next” ทำการตั้งค่าข้อมูลที่นำเข้าในคอลัมน์ที่ตั้งชื่อไว้แล้วกดเลือก “Exclude Column” ในไข้กำหนดกดเลือกเป็น “Change Role” พร้อมทั้งตั้งค่าเป็น “Label” แล้วทำการกดปุ่ม “Next” จากนั้นกด “Ok” ตามด้วยกดปุ่ม “Finish”

4.4.1.4 Naïve Bayes Process สร้างกระบวนการทดสอบประสิทธิภาพการวิเคราะห์รูปแบบความคิดเห็นด้วยนาอีฟ เบย์ (Naïve Bayes) ด้วยชุด Operator Naïve Bayes จะได้หน้าจอการทำงานต่อจากนั้นทำการเลือก Operator Apply Model แล้วทำการเลือก Operator Performance

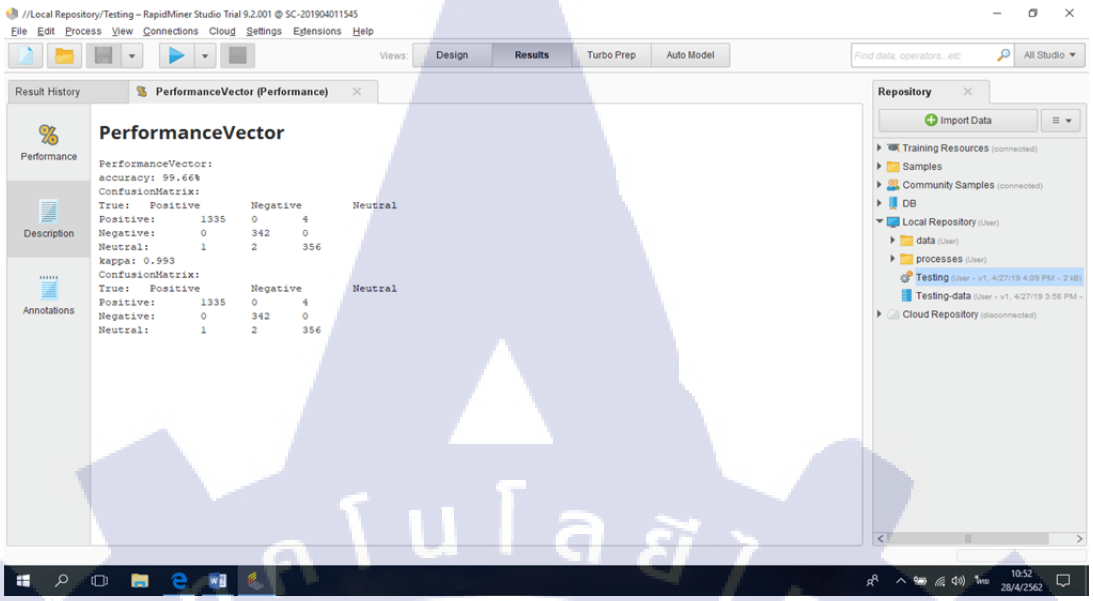
4.4.1.5 หลังจากทำการเลือก Operator ทั้งหมดเสร็จแล้วจึงนำชุดข้อมูลที่ได้ทำการ “Add Data” ไว้ในข้างต้นมาวิเคราะห์ประสิทธิภาพของข้อมูลการเชื่อต่อชุดข้อมูลเข้ากับ Operator จะทำการโดยการลากเส้นให้เชื่อมต่อกัน เมื่อทำการลากเส้นเชื่อมต่อกันครบแล้วทำการเลือกปุ่ม “Execute” โปรแกรมจะทำการวิเคราะห์และแสดงผลการวิเคราะห์ออกมาให้ ดังรูปที่ 4.4 และรูปที่ 4.5 ตามลำดับ



รูปที่ 4.4 หน้าจอแสดงการเชื่อมต่อชุดข้อมูลกับ Operator Naïve Bayes

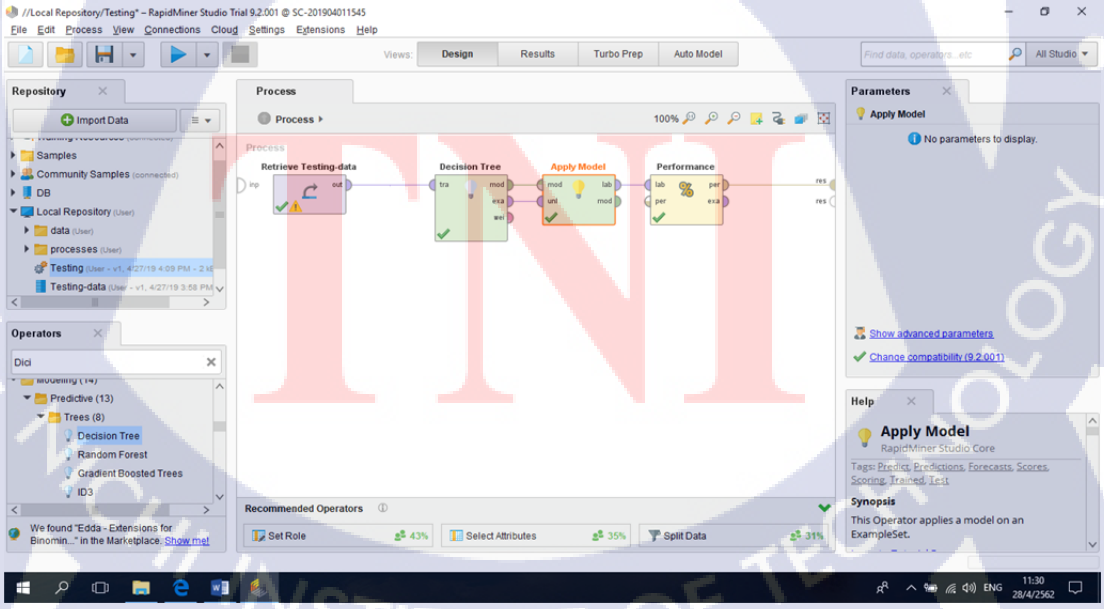


รูปที่ 4.5 หน้าจอแสดงผลการทดสอบชุดข้อมูล Naïve Bayes

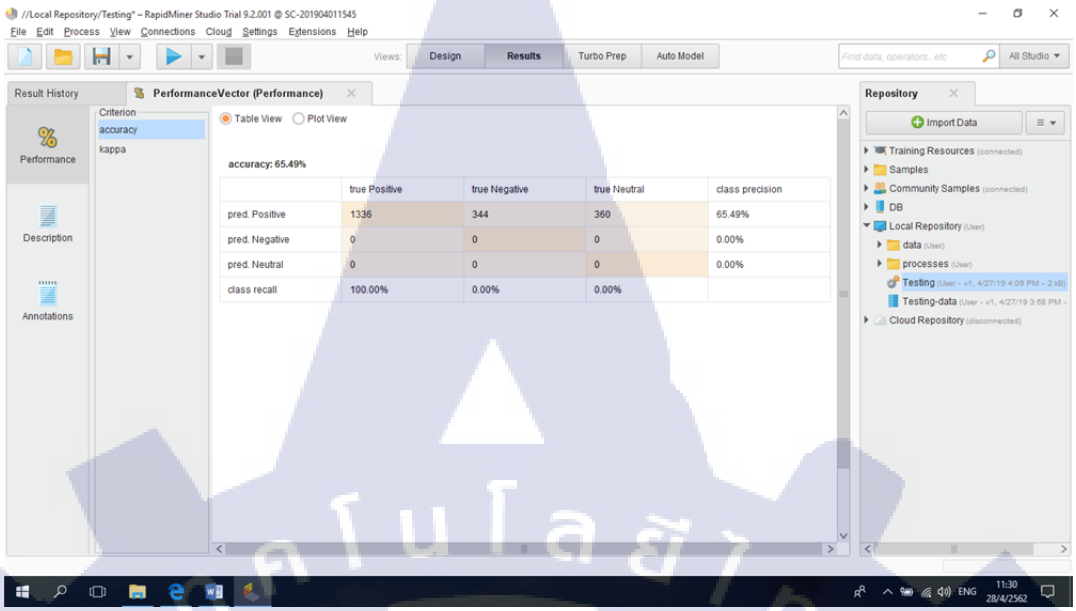


รูปที่ 4.6 หน้าจอแสดรายละเอียดผลการวิเคราะห์ที่ได้จากการทำสอบ Naïve Bayes

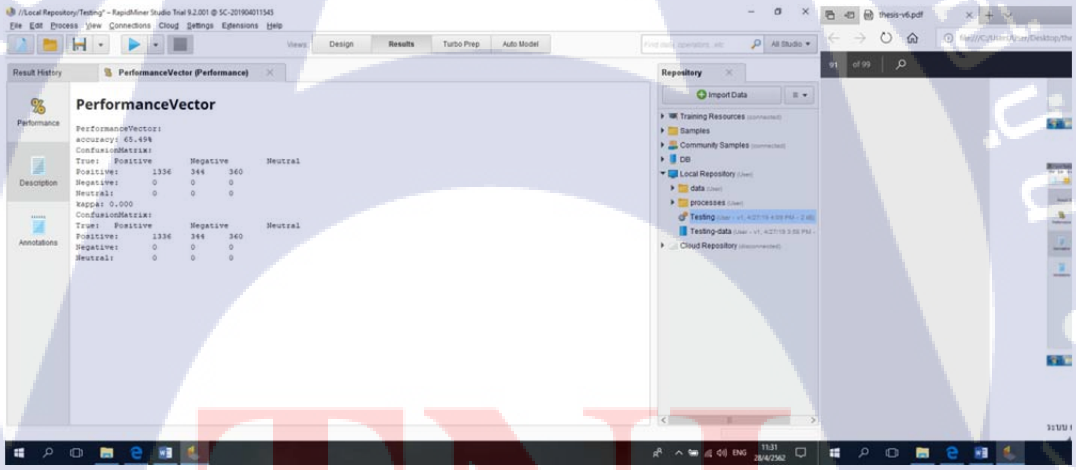
4.4.1.6 Decision Tree สร้างกระบวนการทดสอบประสิทธิภาพการวิเคราะห์รูปแบบความคิดเห็นด้วยต้นไม้ตัดสินใจ (Decision Tree) ด้วยชุด Operator Decision Tree จะได้หน้าจอการทำงานต่อจากนั้นทำการเลือก Operator Apply Model แล้วทำการเลือก Operator Performance ดังรูปที่ 4.67



รูปที่ 4.7 หน้าจอแสดงการเชื่อมต่อชุดข้อมูลกับ Operator Decision Tree



รูปที่ 4.8 หน้าจอแสดงผลการทดสอบชุดข้อมูล Decision Tree



รูปที่ 4.9 หน้าจอแสดรายละเอียดผลการวิเคราะห์ที่ได้จากการทำสอบ Decision Tree

ผลการเปรียบเทียบการวิเคราะห์รูปแบบความคิดเห็นระหว่างอัลกอริทึม Naïve Bayes Decision Tree และ K-Nearest Neighbor จากโปรแกรม Rapid Miner Studio ผู้วิจัยได้ทำการวิเคราะห์แยกข้อมูลเป็น 3 ด้าน ได้แก่ เชิงบวก (Positive) เชิงลบ (Negative) และเป็นกลาง (Neutral) โดยใช้ข้อความคิดเห็นในการวิเคราะห์ทั้งหมด 2,040 ความคิดเห็น จาก 10 จังหวัด จำนวน 40 โรงแรม ซึ่งผู้วิจัยได้ใช้วิธีทดสอบทีละประโยคข้อความด้วยตัวเองผ่านทางเครื่องมือที่ได้พัฒนาขึ้นมาในการตัดคำเพื่อหาคำที่บ่งบอกถึงเชิงบวก เชิงลบ และค่าเป็นกลาง ทั้งนี้ตัวแปรที่สำคัญที่ทำให้เกิดข้อผิดพลาด

มากที่สุดคือ คำศัพท์ที่ไม่มีอยู่ในพจนานุกรม คำที่เกิดจากการเขียนผิด คำที่เกิดจากมีวรรณยุกต์ซ้ำกันหลายตัว โดยผลการทดสอบได้ผลลัพธ์จากอัลกอริทึม Naïve Bayes มีดังนี้ ผลการวิเคราะห์เชิงบวก (Positive) 99.39% ผลการวิเคราะห์เชิงลบ (Negative) 99.42% ผลวิเคราะห์เป็นกลาง (Neutral) 98.89% และทดสอบค่าความถูกต้อง 99.66% ดังตารางที่ 4.13

ตารางที่ 4.13 ผลการทดสอบ Naïve Bayes ด้วยโปรแกรม Rapid Miner Studio

Accuracy (99.66%)	**True Positive**	**True Neutral**	**True Negative**	**Class Precision**
Pred. Positive	1335	0	4	99.70%
Pred. Neutral	0	342	0	100%
Pred. Negative	1	2	356	99.16%
Class Recall	99.36%	99.42%	98.89%	

ผลการทดสอบได้ผลลัพธ์จากอัลกอริทึม Decision Tree มีดังนี้ ผลการวิเคราะห์เชิงบวก (Positive) 100% ผลการวิเคราะห์เชิงลบ (Negative) 0% ผลวิเคราะห์เป็นกลาง (Neutral) 0% และทดสอบค่าความถูกต้อง 65.49% ดังตารางที่ 4.14

ตารางที่ 4.14 ผลการทดสอบ Decision Tree ด้วยโปรแกรม Rapid Miner Studio

Accuracy (65.49%)	**True Positive**	**True Neutral**	**True Negative**	**Class Precision**
Pred. Positive	1336	344	360	65.49%
Pred. Neutral	0	0	0	0.00%
Pred. Negative	0	0	0	0.00%
Class Recall	100%	0%	0%	

ผลการทดสอบได้ผลลัพธ์จากอัลกอริทึม K-Nearest Neighbor มีดังนี้ ผลการวิเคราะห์เชิงบวก (Positive) 83.68% ผลการวิเคราะห์เชิงลบ (Negative) 26.16% ผลวิเคราะห์เป็นกลาง (Neutral) 45.00% และทดสอบค่าความถูกต้อง 70.37% ดังตารางที่ 4.15

ตารางที่ 4.15 ผลการทดสอบ K-Nearest Neighbor ด้วยโปรแกรม Rapid Miner Studio

Accuracy (70.21%)	True Positive	True Neutral	True Negative	Class Precision
Pred. Positive	1118	196	180	74.83%
Pred. Neutral	37	90	18	62.07%
Pred. Negative	181	58	162	40.40%
Class Recall	83.68%	26.16%	45.00%	

บทที่ 5
บทสรุป อภิปราย และข้อเสนอแนะ

การศึกษาวิจัยเรื่องการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมโดยใช้เทคนิคการตัดคำแบบผสมผสาน ซึ่งการวิจัยนี้มีวัตถุประสงค์ตามบทที่ 1 ได้นำเสนอข้อสรุปตามลำดับดังนี้

5.1 สรุปผลการวิจัย

5.2 อภิปรายผลวิจัย

5.3 ปัญหาและอุปสรรคในการทำวิจัย

5.4 ข้อเสนอแนะงานวิจัย

5.1 สรุปผลการวิจัย

ในการสรุปผลของงานวิจัย ผู้วิจัยได้ทำการสรุปผลตามวัตถุประสงค์ในแต่ละข้อที่ได้กล่าวไว้ข้างต้นในบทที่ 1 ดังต่อไปนี้

การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมสมาน โดยการรวบรวมข้อมูลจากเว็บไซต์โกด้า แล้วนำข้อมูลมาตัดคำโดยใช้เทคนิคผสมผสาน คือ การนำการตัดคำแบบกฎ การตัดคำแบบพจนานุกรม และการตัดคำแบบคลังข้อมูล พิจารณาการวนลูปจากซ้ายไปขวาตามกฎหลังภาษาไทย แล้วนำมาวิเคราะห์ความคิดเห็นออกมาเป็นเชิงบวก เชิงลิบ และเป็นกลาง เพื่อสรุปภาพรวมของการวิเคราะห์ความรู้สึก (Sentimenti Analysis) จำแนกระหว่างแบบนาอีฟ เบย์ Naïve Bayes แบบต้นไม้ตัดสินใจ Decision Tree และแบบเคเนียเรสเนเบอร์ K-Nearest Neighbor

สรุปงานวิจัยนี้ ผู้วิจัยมุ่งเน้นเสนอการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมโดยใช้เทคนิคการตัดคำแบบผสมผสานและประเมินค่าความแม่นยำและค่าความถูกต้องของการวิเคราะห์รูปแบบความรู้สึกจากการจำแนก 3 อัลกอริทึม ดังนี้

1) แบบนาอีฟ เบย์ Naïve Bayes

โดยแบ่งออกเป็นผลวิเคราะห์ความคิดเห็นเชิงบวก (Positive) 99.36%, ผลวิเคราะห์ความคิดเห็นเชิงลบ (Negative) 98.89%, ผลวิเคราะห์ความคิดเห็นเป็นกลาง (Neutral) 99.43%, ความถูกต้อง (Accuracy) 99.66%

2) แบบต้นไม้ตัดสินใจ Decision Tree

โดยแบ่งออกเป็นผลวิเคราะห์ความคิดเห็นเชิงบวก (Positive) 100%, ผลวิเคราะห์ความคิดเห็นเชิงลบ (Negative) 0.00%, ผลวิเคราะห์ความคิดเห็นเป็นกลาง (Neutral) 0.00%, ความถูกต้อง (Accuracy) 65.49%

3) แบบเคเนียเรสเนเบอร์ K-Nearest Neighbor

โดยแบ่งออกเป็นผลวิเคราะห์ความคิดเห็นเชิงบวก (Positive) 83.68%, ผลวิเคราะห์ความคิดเห็นเชิงลบ (Negative) 26.16%, ผลวิเคราะห์ความคิดเห็นเป็นกลาง (Neutral) 45.00%, ค่าความถูกต้อง (Accuracy) 70.37%

5.2 อภิปรายผลวิจัย

จากการทำวิทยานิพนธ์เพื่อการศึกษา เรื่อง การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน

5.2.1 การตัดคำภาษาไทยโดยใช้เทคนิคผสมผสานมุ่งเน้นเพื่อแก้ปัญหาการตัดคำด้วยวิธีเดิมที่มีอยู่แล้ว ซึ่งไม่ครอบคลุมคำในภาษาไทยที่มีลักษณะที่แตกต่างกันออกไปจากเดิมโดยเฉพาะจำพวกของคำเฉพาะ คำทับศัพท์ที่มีจากภาษาต่างประเทศ คำแสลง และคำที่อ่านออกเสียงได้หลากหลายแบบแต่ยังคงรูปเดิม การนำเทคนิคแบบผสมผสานมาใช้นี้สามารถเพิ่มความยืดหยุ่นได้มากขึ้น อีกทั้งมีความถูกต้องของการตัดคำโดยเฉพาะคำที่ไม่มีอยู่ในพจนานุกรมเพราะว่าได้นำคลังข้อมูลเข้ามาช่วยในการตัดคำเพิ่มขึ้นอีกขั้นตอนหนึ่งจึงทำให้การตัดคำได้ถูกต้องแม่นยำขึ้นถึง 84.50% เมื่อเทียบกับการตัดคำแบบเดิมแล้วได้ค่าความแม่นยำเพิ่มขึ้นอยู่ที่ 8.17%

5.2.2 การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมมีปัจจัยที่ทำให้เกิดการวิเคราะห์ความคิดเห็นเชิงบวก การวิเคราะห์ความคิดเห็นเชิงลบ และการวิเคราะห์ความคิดเห็นเป็นกลางอยู่ 5 ปัจจัยหลัก ได้แก่ ด้านสภาพห้องพัก ด้านที่ตั้งและสภาพแวดล้อมโรงแรม ด้านการให้บริการ ด้านสิ่งอำนวยความสะดวก และด้านราคา งานวิจัยที่เกี่ยวข้องนี้อาจจะเป็นการพัฒนาโมเดลในครั้งต่อไปที่จะวิเคราะห์ความคิดเห็นของข้อมูลโรงแรมได้ในทุกๆ ระดับดาวของโรงแรมไม่ว่าจะเป็นเชิงบวก เชิงลบ และเป็นกลาง

5.2.3 ผลการประเมินประสิทธิภาพการวิเคราะห์ความคิดเห็นที่นำมาใช้ในงานวิจัยครั้งนี้ ได้แก่ การทดสอบแบบ Naïve Bayes การทดสอบแบบ Decision Tree และการทดสอบแบบ K-Nearest Neighbor พบว่าการประเมินประสิทธิภาพแบบ Naïve Bayes ให้ค่าความถูกต้องมากที่สุดถึง 99.66% และพบกว่าการประเมินประสิทธิภาพแบบ Decision Tree ให้ค่าความถูกต้องน้อยที่สุดในการประเมินประสิทธิภาพทั้ง 3 แบบ อาจจะเนื่องด้วยการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรมในบางส่วนยังไม่สามารถแยกขั้วความคิดเห็นในเชิงบวกและเชิงลบที่อยู่รวมกันในประโยคได้ดังนั้นจึงควรแก้ด้วย

วิธีการแยกขั้วความคิดเห็นที่อยู่รวมของของประโยคที่ผ่านการวิเคราะห์รูปแบบความคิดเห็นให้เป็น 2 ส่วนชัดเจนและทำการจัดการแบ่งข้อมูลทั้งหมดให้เท่าๆ กันในการวิเคราะห์

5.2.4 กระบวนการวิเคราะห์ความคิดเห็นที่ผ่านผู้บริโภคในเชิงบวก เชิงลบ และเป็นกลาง ยังสามารถนำไปปรับปรุงใช้เกี่ยวกับการวิเคราะห์ข้อความออนไลน์อื่นๆ ได้อีก เช่น การวิเคราะห์ข้อเสนอแนะของผู้บริโภคที่นำไปใช้ในด้านธุรกิจ การวิเคราะห์วิกกฤตการณ์ต่างๆ ที่เกิดขึ้นในปัจจุบัน เป็นต้น แต่ปัญหาที่เกิดขึ้น คือ ข้อความคิดเห็นจำนวนมหาศาลล้วนมีความหลากหลาย จึงจำเป็นต้องแบ่งย่อยเป็นหมวดหมู่ปัจจัยเสียก่อน เพื่อให้ง่ายต่อการวิเคราะห์ข้อมูลขนาดใหญ่ที่จะนำไปใช้งานต่อไป

5.3 ปัญหาและอุปสรรคในการทำงานวิจัย

5.3.1 ปัญหาการเก็บข้อมูลที่มีความคลาดเคลื่อนและผิดพลาด เนื่องจากความคิดเห็นที่เราเก็บรวบรวมมานั้น บางความคิดเห็นจะมีการพิมพ์อักษรที่ตกหล่นไปบางความคิดเห็นจะมีการพิมพ์สระและวรรณยุกต์เกินมา จึงทำให้การวิเคราะห์มีประสิทธิภาพลดลง

5.3.2 ปัญหาในการใช้เครื่องมือโปรแกรม Rapid Miner Studio ซึ่งต้องศึกษาให้เชี่ยวชาญในด้านการใช้งานมากขึ้น จึงทำให้เกิดปัญหาในขั้นตอนการวิเคราะห์ Sentiment

5.4 ข้อเสนอแนะงานวิจัย

5.4.1 พจนานุกรมการแบ่งของคำยังมีน้อย สำหรับการนำมาใช้ในการวิเคราะห์ข้อมูลจำนวนมาก ควรมีการเพิ่มคำที่เป็นคำกำกวม คำสแลง คำทับศัพท์จากภาษาต่างประเทศให้มากขึ้นกว่าเดิม เพื่อผลลัพธ์ของงานวิจัยที่ดีขึ้น

5.4.2 ควรมีการจัดการแก้ไขคำที่ผิดเกิดจากการพิมพ์สระ วรรณยุกต์ อักษรที่ขาดและเกินมา เพื่อการนำไปวิเคราะห์จะได้มีประสิทธิภาพมากยิ่งขึ้น

บรรณานุกรม

บรรณานุกรม

[1] อานนท์ ศักดิ์วรวิชญ์, *"Rwedictive Analytics จะทำนายอะไร ทำไมต้องทำนาย?,"* [Online]. Available : https://mgronline.com/daily/detail/9590000054507. [Accessed : 30 มกราคม 2561].

[2] สมัคร ชัยสงวน, *"การพัฒนาระบบวิเคราะห์ความรู้สึกแบบเรียลไทม์ของนักศึกษาบนเฟซบุ๊กโดยใช้ตัวจำแนกข้อมูล นาอีฟ เบย์ สำหรับภาษาไทย,"* วิทยานิพนธ์ วท.ม. (วิทยาศาสตร์และเทคโนโลยี), มหาวิทยาลัยสุโขทัยธรรมาธิราช, นนทบุรี, 2560.

[3] นพมาศ ธีรเวคิน, *จิตวิทยาสังคมกับชีวิต 2,* กรุงเทพฯ : ภาควิชาจิตวิทยา คณะจิตวิทยา มหาวิทยาลัยธรรมศาสตร์, 2539.

[4] อัมพาพร นพรัตน์, *"ความคิดเห็นของนักศึกษาที่มีต่อการจัดการเรียนการสอนตามหลักสูตรของสถาบันการพลศึกษา วิทยาเขตยะลาปีการศึกษา 2556,"* วิทยานิพนธ์ บธ.ม. (การจัดการ), มหาวิทยาลัยหาดใหญ่, สงขลา, 2558.

[5] สมรรถชัย คันธมาทน์, *"ความคิดเห็นของนักศึกษาที่มีต่อการเรียนการสอนตามหลักสูตรของสถาบันการพลศึกษาในเขตภาคใต้ ปีการศึกษา 2555,"* วิทยานิพนธ์ บธ.ม. (การจัดการ), มหาวิทยาลัยศรีนครินทรวิโรฒ, กรุงเทพฯ, 2556.

[6] มัณฑนา โพธิ์แก้ว, *"ความคิดเห็นของบุคลากรที่มีต่อการบริหารงานขององค์การบริหารส่วนตำบลในเขตอำเภอแหลมสิงห์ จังหวัดจันทบุรี,"* วิทยานิพนธ์ รป.ม. (รัฐประศาสนศาสตร์), มหาวิทยาลัยบูรพา, ชลบุรี, 2550.

[7] จิตตินันท์ เดชะคุปต์, *เจตคติและความพึงพอใจในการบริการ,* นนทบุรี : มหาวิทยาลัยสุโขทัยธรรมาธิราช, 2543.

[8] นพพร ไพบูลย์, *"ความคิดเห็นของพนักงาน บริษัท เอสวีไอ จำกัด (มหาชน) ที่มีต่อการยอมรับมาตรฐานการจัดสิ่งแวดล้อม ISO 14000,"* วิทยานิพนธ์ วท.ม. (การวัดและประเมินผลการศึกษา), มหาวิทยาลัยรามคำแหง, กรุงเทพฯ, 2546.

[9] Foster, "Psychology for life adjustment chicago," *American Technical So-Ciety,* vol. 30, no. 2, pp. 502-505, August 1952.

[10] M. P. Fled Man, *Psychology in the Industrial Environment,* London : Butterworth and Co., Ltd., 1971.

[11] W. John and Best, *Research in Education,* New Jersey : Prentice Hall, 1977.

[12] สัตตยา กระแสชล, *"ความคิดเห็นของประชาชนต่อการจัดตั้งอุทยานสายใจธรรม จังหวัดฉะเชิงเทรา,"* วิทยานิพนธ์ วท.ม. (วิทยาศาสตร์สิ่งแวดล้อม), มหาวิทยาลัยเกษตรศาสตร์, กรุงเทพฯ, 2538.

[13] E. Hurlock, *Adolescent Development,* New York : McGraw-Hill, 1995.

[14] ประดินันท์ อุปรมัย, *จิตวิทยากับการศึกษา,* กรุงเทพฯ : ภาควิชาจิตวิทยา คณะจิตวิทยา มหาวิทยาลัยสุโขทัยธรรมาธิราช, 2533.

[15] H. Jiao et al., "Chinese keyword extraction based on word platform," *Fourth International Conference on Fuzzy Systems and Knowledge Discovery, NEDSI 2007 Proceedings,* Wuhan, China, June 2007, pp. 360-363.

[16] K. Sukhum et al., "Opinion detection in thai political news columns based on subjectivity analysis," *Annual International Conference Proceedings : 9th Computer Games & Allied Technologies, CGAT 2016,* Tampines, Singapore, March 6-9, 2016, pp. 28-34.

[17] ยืน ภู่วรวรรณ และวิวรรธ์ อิ่มอารมณ์, "การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี," *การประชุมวิชาการวิศวกรรมไฟฟ้าครั้งที่ 9, EECON-29,* กรุงเทพฯ, 9-10 พฤศจิกายน 2529, หน้า 152-158.

[18] กานดา รุณนะพงศา และปโยธร อุราธรรมกุล, "การตัดภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่," *งานประชุมวิชาการด้านวิศวกรรมคอมพิวเตอร์ 2549, NGRC,* ขอนแก่น, 22 พฤษภาคม 2549, หน้า 25-30.

[19] สมปรารถนา รัทยานนท์, *"โครงสร้างข้อมูลสำหรับพจนานุกรมอิเล็กทรอนิกส์ภาษาไทย,"* วิทยานิพนธ์ บธ.ม. (การจัดการ), จุฬาลงกรณ์มหาวิทยาลัย, กรุงเทพฯ, 2535.

[20] วิรัช ศรเลิศล้ำวาณิช, "การตัดคำในระบบแปลภาษา (Word Segmentation for Thai in Machine Translation System) การแปลภาษาด้วยคอมพิวเตอร์," *งานประชุมวิชาการด้านวิศวกรรมคอมพิวเตอร์, NECTEC,* กรุงเทพฯ, 13 สิงหาคม 2536, หน้า 109.

[21] Thanaruk Theeramunkong and Sasiporn Usanavasin, "Non-dictionary-based thai word segmentation using decision trees," *The Eighth National Conference on Computing and Information Technology, NCCIT 2012,* Bangkok, May 9-19, 2012, pp. 729-736.

[22] W. Aroonmanakun, "Collocation and thai word segmentation," *The Tenth National Conference on Computing and Information Technology, NCCIT 2014,* Bangkok, October 7, 2014, pp. 245-251.

[23] พรพล ธรรมรงค์รัตน์ ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์ประสิทธิ์, "การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเว็กเตอร์แมชชีน," *The 12th National Computer Science and Engineering Conference 2008, NCSEC,* Chonburi, 19-21 พฤศจิกายน 2551, หน้า 72-77.

[24] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing and Management,* vol. 24, no. 5, pp. 513-523, May 2012.

[25] V. Raghavan and S.K.M. Wong, "A critical analysis of vector space model for information retrieval," *Journal of the American Society for Information Science,* vol. 37, no. 5, pp. 279-287, April 1986.

[26] สมศักดิ์ วิชัยกิจ และมาลีรัตน์ โสดานิล, *การจัดกลุ่มลูกค้าสินเชื่อจากการปฏิเสธด้วยวิธีการทำเหมืองข้อความ,* นนทบุรี : มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2556.

[27] เมธาพร คงทอง, *"การจัดกลุ่มลูกค้าค้างชำระของสินเชื่อที่อยู่อาศัยของธนาคารพาณิชย์แห่งหนึ่ง,"* สารนิพนธ์ บธ.ม. (การจัดการ), มหาวิทยาลัยเทคโนโลยีราชมงคลธัญบุรี, ปทุมธานี, 2556.

[28] ธิดาพร วงศ์พิพันธ์, *"การใช้เหมืองข้อมูลช่วยตัดสินใจในการให้สินเชื่อ กรณีศึกษาของกรุงไทยคาร์เร้นท์,"* สารนิพนธ์ บธ.ม. (การจัดการ), มหาวิทยาลัยธุรกิจบัณฑิต, กรุงเทพฯ, 2556.

[29] Jiao et al., "Chinese keyword extraction on word platform," *Fourth International Conference on Fuzzy Systems and Knowledge Discovery, NEDSI 2007 Proceedings,* Wuhan, China, June 2007, pp. 365-367.

[30] Li et al., "Keyword extraction based on lexical chains and word co-occurrence for chinese news web pages," *IEEE International Conference on Data Mining Workshops, AACC 2008,* Eagan, United States America, May 2008, pp. 744-751.

[31] Wartena et al., "Keyword extraction using word co-occurrence," *Worlshops on Database and Expert Systems Applications,* vol. 1, no. 8, pp. 54-58, September 2010.

[32] Zheng and Ye, "บทวิจารณ์ของนักท่องเที่ยวในภาษาจีนบนเว็บไซต์ ctrip.com," *Engineering Journal,* vol. 35, no. 2, pp. 256-261, May 2017.

[33] J. M. Kassim and M. Rahmany, "Introduction to semantic search engine," *Electrical Engineering and Informatics 2009, ICEEI 09 International Conference,* Bangkok, August 5, 2009, pp. 923-928.

[34] B. Fang and S. Ma, “Theoretical foundations of instructional applications of computer-generated animated visuals,” *Journal of Computer-Based Instruction,* vol. 18, no. 3, pp. 83-88, November 2009.

[35] Noroozi and Fotouhi, “The influence of semantic web on decision making of customers in tourism industry,” *International Journal of Information Science and Management,* vol. 34, no. 3, pp. 77-98, June 2010.

[36] J. L. Seng, “An analytic approach to select data mining for business decision,” *In Expert Systems with Applications 37,* vol. 3, no. 2, pp. 8,042-8,057, September 2010.

[37] Q. Luo, “Advancing knowledge discovery and data mining,” *First International Workshop, The Eighth National Conference on Computing and Information Technology, NCCIT 2012,* Bangkok, May 9-19, 2012, pp. 598–604.

[38] Y. Jin et al., “The research of search engine based on semantic web,” *Proceedings of the 2004 Northeast Decision Science Institute Conference, NEDSI 2008 Proceedings,* New Jersey, USA, September 12-20, 2008, pp. 32-39.

[39] นลินี พฤกษาชาติ และธนพล เจนสุทธิเวชกุล, “การจำแนกความคิดเห็นออนไลน์ของนักท่องเที่ยวโดยใช้ Word Vector และอัลกอริทึมเนอีฟ เบย์,” *การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 9, IE Network 2010,* อุบลราชธานี, 13-15 ตุลาคม 2553, หน้า 57-63.

[40] ทิพย์อรุณ เขี้ยวแก้ว และณัฐพงศ์ ทองเทพ, “การวิเคราะห์การสกัดรูปแบบการตอบกลับที่สุภาพต่อข้อความแสดงความเห็นที่มีต่อสินค้าและบริการ,” *การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 7, CITT,* กรุงเทพฯ, 29-30 ตุลาคม 2558, หน้า 72-77.

[41] รณกร ขำเพชร และวรารัตน์ สงฆ์แป้น, “การวิเคราะห์ข้อความข่าวราคายางพาราโดยใช้เทคนิคเหมืองข้อความ,” *การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 9, CITT,* กรุงเทพฯ, 22-24 มีนาคม 2560, หน้า 102-107.

[42] Krishnaveni and Hemalatha, “ศึกษาการวิเคราะห์การเกิดอุบัติเหตุทางจราจรด้วยเทคนิคเหมืองข้อมูล,” *การประชุมวิชาการระดับประเทศด้านเทคโนโลยีสารสนเทศ ครั้งที่ 9, CITT,* กรุงเทพฯ, 22-24 มีนาคม 2560, หน้า 24-29.

[43] T. Niu et al., *Sentiment Analysis on Multi View Social Data,* New Jersey : Prentice Hall,

[44] J. G. Kang et al., “Semantic network analysis of vaccine sentiment in online social media,” *Engineering Journal,* vol. 35. no. 2. pp. 256-259, June 2017.

[45] N. Kumar et al., “Sentiment dynamics in social media news channels,” *Engineering Journal,* vol. 12, no. 2, pp. 789-793, May 2018.

ภาคผนวก

ภาคผนวก ก.
โค้ดโปรแกรมพีเฮชพี (PHP)

โค้ดการออกแบบ UI หน้าเว็บไซต์

```
<style type="text/css">
    body{
        font-size:16px;
    }
</style>

<div style="float:left; width:400px;">
    <form method="post">
        <b>ข้อความ : </b><br/>
        <textarea name="text_to_segment" cols="40" rows="50"
style="width:300px;height:300px;"><?php echo isset($_POST['text_to_segment'])?
trim($_POST['text_to_segment']):'' ?></textarea>
        <br/>
        <input type="submit" value="ตัดคำ" style="width:301px;height:40px;font-
size:18px;background: #eee"/>

    </form>
</div>

<div style="width:1000px;">
  <?php
   if ($_POST) {

        $time_start = microtime(true);

        $text_to_segment = trim($_POST['text_to_segment']);
        echo '<hr/>';
        //echo '<b>ประโยคที่ต้องการตัดคือ: </b>' . $text_to_segment . '<br/>';
        include(dirname(__FILE__) . DIRECTORY_SEPARATOR .
```

```
'THSplit/segment.php');
        $segment = new Segment();
        //echo '<hr/>';
        $result = $segment->get_segment_array($text_to_segment);
        echo implode(' | ', $result);

        function convert($size) {
            $unit = array('b', 'kb', 'mb', 'gb', 'tb', 'pb');
            return @round($size / pow(1024, ($i = floor(log($size, 1024)))), 2) . ' ' .
$unit[$i];
        }
        $time_end = microtime(true);
        $time = $time_end - $time_start;
        echo '<br/><b>ประมวลผลใน: </b> '.round($time,4).' วินาที';
        echo '<br/><b>คำที่อาจจะตัดผิด:</b> ';
        foreach($result as $row)
        {
            if (mb_strlen($row) > 12)
            {
                echo $row.'<br/>';
            }
        }
    }
    ?>

  </div>
</body>
```

โค้ดการทำงานภายในโปรแกรมที่ใช้ในการตัดคำ

```
<?php

class Segment {

    private $_input_string;
    private $_dictionary_array = array();
    private $_thcharacter_obj;
    private $_unicode_obj;
    private $_segmented_result = array();

    function __construct() {
        if (!class_exists('Thchracter')) {
        include(dirname(__FILE__) . DIRECTORY_SEPARATOR . 'thcharacter.php');
                                }

                                if (!class_exists('Unicode')) {
        include(dirname(__FILE__) . DIRECTORY_SEPARATOR . 'unicode.php');
                                }

        $file_handle = fopen(dirname(__FILE__) . DIRECTORY_SEPARATOR .
'dictionary' . DIRECTORY_SEPARATOR . 'dictionary.txt', "rb");
        while (!feof($file_handle)) {
            $line_of_text = fgets($file_handle);
            $this->_dictionary_array[crc32(trim($line_of_text))] = trim($line_of_text);
        }
        fclose($file_handle);

        // Load Helper Class/
        $this->_unicode_obj = new Unicode();
```

```
        $this->_thcharacter_obj = new Thchracter();
    }

    private function clear_duplicated($string) {
        //หดรูปตัวอักษรซ้ำๆ//
        $input_string_split = $this->_unicode_obj->uni_strsplit($string);
        $previous_char = '';
        $previous_string = '';
        $dup_list_array = array();
        $dup_list_array_replace = array();
        foreach ($input_string_split as $current_char) {

            if ($previous_char == $current_char) {
                $previous_char = $current_char;
                $previous_string .= $current_char;
            } else {
                if (mb_strlen($previous_string) > 3) {
                    $dup_list_array[] = $previous_string;
                    $dup_list_array_replace[] = $current_char;
                    $string = str_replace($previous_string, $previous_char, $string);
                }
                $previous_char = $current_char;
                $previous_string = $current_char;
            }
        }
        if (mb_strlen($previous_string) > 3) {
            $dup_list_array[] = $previous_string;
            $dup_list_array_replace = $current_char;
        }
        return str_replace($dup_list_array, $dup_list_array_replace, $string);
    }
```

```
    public function get_segment_array($input_string) {
        $this->_input_string = $input_string;

        // ลบเครื่องหมายคำพูด, ตัวแบ่งประโยค //
        $this->_input_string = str_replace(array('\'', '‘', '’', '“', '”', '"', '-', '/', '(', ')', '{', '}', '...',
'..', '…', '', ',', ':', '|', '\\', '^',';','.'), '', $this->_input_string);
        // เปลี่ยน newline ให้กลายเป็น Space เพื่อที่ใช้สำหรับ Trim
        $this->_input_string = str_replace(array("\r", "\r\n", "\n"), ' ', $this-
>_input_string);

        // กำจัดซ้ำ //
        $this->_input_string = $this->clear_duplicated($this->_input_string);

        // แยกประโยคจากช่องว่าง (~เผื่อไว้สำหรับภาษาอังกฤษ) //
        $this->_input_string_exploded = explode(' ', $this->_input_string);

        // Reverse Array สำหรับการใช้ Dictionary แบบ Reverse //
        foreach ($this->_input_string_exploded as $input_string_exploded_row) {
            $current_string_reverse_array = array_reverse($this->_unicode_obj-
>uni_strsplit(trim($input_string_exploded_row)));

            $current_array_result = $this-
>_segment_by_dictionary_reverse($current_string_reverse_array);
            foreach ($current_array_result as $each_result) {
                if (trim($each_result) != '')
                    $this->_segmented_result[] = trim($each_result);
            }
        }

        // จัดการคำที่ตัดที่ยาวผิดปกติ (~อาจจะเป็นเพราะว่าพิมผิด) โดยการตัดตาม Dict แบบ
ธรรมดา//
```

```
        $tmp_result = array();
        foreach ($this->_segmented_result as $result_row) {
            if (mb_strlen($result_row) > 10) {

                $current_string_array = $this->_unicode_obj-
>uni_strsplit(trim($result_row));
                $current_array_result = $this-
>_segment_by_dictionary($current_string_array);

                foreach ($current_array_result as $current_result_row) {
                    $tmp_result[] = trim($current_result_row);
                }
            } else {
                $tmp_result[] = $result_row;
            }
        }
        $this->_segmented_result = $tmp_result;
        return $this->_segmented_result;
    }

    private function _segment_by_dictionary($input_array) {

        $result_array = array();
        $tmp_string = '';

        $pointer = 0;
        $length_of_string = count($input_array)-1;

        while ($pointer <= $length_of_string) {

            $tmp_string .= $input_array[$pointer];
```

```
if (isset($this->_dictionary_array[crc32($tmp_string)])) { // ถ้าเจอใน Dict //
    $dup_array = array();
    $dup_array[] = array(
        'title' => $tmp_string,
        'to_mark' => $pointer + 1,
    );
    $count_more = 0;
    $more_tmp = $tmp_string;
    //echo $more_tmp.'<br/>';

    for ($i = $pointer + 1; $i <= $length_of_string; $i++) {
        $more_tmp .= $input_array[$i];
        //echo $more_tmp.'<br/>';
        //echo $more_tmp.'<br/>';
        if (isset($this->_dictionary_array[crc32($more_tmp)])) {
            $dup_array[] = array(
                'title' => $more_tmp,
                'to_mark' => $i + 1,
            );
            //print_r($dup_array);
        }

        $count_more++;
    }

    if (count($dup_array) > 0) {
        $result_array[] = $dup_array[count($dup_array) - 1]['title'];

        $pointer = $dup_array[count($dup_array) - 1]['to_mark'];

        //echo $to_mark;
```

```
            } else {


            }
            //echo '----------<br/>';
            $dup_array = array();
            $tmp_string = '';
            continue;
        }

        $pointer++;
    }

    if ($tmp_string != '') { //  ส่วนที่เหลือ ถ้าไม่เจอใน Dict
        $result_array[] = $tmp_string;
    }

    if (count($result_array) == 0) {
        return array(implode($input_array));
    }
    return $result_array;
}

private function _segment_by_dictionary_reverse($input_array) {

    $result_array = array();
    $tmp_string = '';

    $pointer = 0;
    $length_of_string = count($input_array)-1;

    while ($pointer <= $length_of_string) {
```

```
$tmp_string = $input_array[$pointer] . $tmp_string;

if (isset($this->_dictionary_array[crc32($tmp_string)])) { // ถ้าเจอใน Dict //
    $dup_array = array();
    $dup_array[] = array(
        'title' => $tmp_string,
        'to_mark' => $pointer + 1,
    );
    $count_more = 0;
    $more_tmp = $tmp_string;
    //echo $more_tmp.'<br/>';

    for ($i = $pointer + 1; $i <= $length_of_string; $i++) {
        $more_tmp = $input_array[$i] . $more_tmp;
        //echo $more_tmp.'<br/>';
        //echo $more_tmp.'<br/>';
        if (isset($this->_dictionary_array[crc32($more_tmp)])) {
            $dup_array[] = array(
                'title' => $more_tmp,
                'to_mark' => $i + 1,
            );
            //print_r($dup_array);
        }

        $count_more++;
    }

    if (count($dup_array) > 0) {
        $result_array[] = $dup_array[count($dup_array) - 1]['title'];

        $pointer = $dup_array[count($dup_array) - 1]['to_mark'];
```

```
                //echo $to_mark;
            } else {

            }
            //echo '--------------------<br/>';
            $dup_array = array();
            $tmp_string = '';
            continue;
        }

        $pointer++;
    }

    if ($tmp_string != '') { //  ส่วนที่เหลือ ถ้าไม่เจอใน Dict
        $result_array[] = $tmp_string;
    }

    if (count($result_array) == 0) {
        return array(implode(array_reverse($input_array)));
    }
    return array_reverse($result_array);
}
```