การวิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ

Sentiment analysis opinion mining using Text Extraction

สมศักดิ์ ศรีสวการย์ [1*] และ สมัย ศรีสวย [2]

Somsak Srisawakarn[1*] และ Samai Srisuay[2]

คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏลำปาง[1,2]

Faculty of Science, Lampang Rajabhat University[1,2]

E-mail: somsak@lpru.ac.th [1*] samai@lpru.ac.th [2]

บทคัดย่อ

ในงานวิจัยนี้ได้นำข้อมูลจากบทวิจารณ์ออนไลน์ผ่านเครือข่ายเฟซบุ๊กสาธารณะของนักท่องเที่ยวมาสกัดคำแยกความคิดเห็นเชิงบวก เชิงลบ และได้ทำเปรียบเทียบประสิทธิภาพจากค่าความถูกต้องด้วยอัลกอริทึมนาอีฟเบย์ อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจ ผลการศึกษา พบว่า อัลกอริทึมนาอีฟเบย์ ให้ค่าความถูกต้อง 87.97% อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด ให้ค่าความถูกต้อง 83.80% และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจ ให้ค่าความถูกต้อง 79.89% ผู้วิจัยได้เลือกวิธีนาอีฟเบย์ มาใช้ในการพัฒนาระบบวิเคราะห์ความคิดเห็นจากเครือข่ายเฟซบุ๊กโรงแรมและที่พักในจังหวัดลำปางมาทดสอบระบบ ผลพบว่า ระบบสามารถสกัดคำแยกความคิดเห็นเชิงลบและเชิงบวกตามเวลาจริงทำงานได้อย่างมีประสิทธิภาพ

คำสำคัญ : การวิเคราะห์เหมืองความคิดเห็น, การจำแนกข้อความ, เครือข่ายสังคมออนไลน์, การสกัดคำ

Abstract

This research retrieved the dataset from the Facebook online articles of the tourism by extract positive and negative review words and compares the efficiency of accuracy rate by Naive Bayes K-Nearest Neighbors and decision tree. The result showed the Naive Bayes algorithm accuracy was 87.97% K-Nearest Neighbors accuracy was 83.80 and decision tree algorithm accuracy was 79.89 %. The researchers selected the Naive Bayes algorithm to be the prototype for develop the online analysis system and select the Facebook online reviews in the Lampang Province legion's accommodation. The result showed the positive and negative review word extraction system can use functionally.

Keywords: Opinion Mining, Text Classification, Social Networks, Extraction

บทนำ

พฤติกรรมของผู้ใช้อินเทอร์เน็ตในประเทศไทยส่วนใหญ่เริ่มเข้าสู่เครือข่ายทางสังคม (Social Networking) มีการติดต่อและสร้างความสัมพันธ์กับบุคคลในโลกออนไลน์ผ่านทางชุมชนออนไลน์ (Online Community) ที่มีลักษณะเป็นสังคมเสมือน (Virtual Community) ทำให้ผู้คนสามารถรู้จัก พูดคุย แลกเปลี่ยนข้อมูลข่าวสาร และสามารถเชื่อมโยงกันได้อย่างรวดเร็วและตลอดเวลาผ่านสื่อสังคมออนไลน์ (Online Social Media) อีกทั้งคุณสมบัติของเครือข่ายออนไลน์ที่มีลักษณะเป็นชุมชน สามารถกระจายข้อมูลข่าวสารได้อย่างรวดเร็วกว่าสื่อที่มีอยู่ทั่วไป ข้อมูลข่าวสารเหล่านี้สามารถสื่อสะท้อนได้ถึงความสนใจ และความคิดเห็นของผู้คนในสังคมออนไลน์ที่มีต่อเหตุการณ์ในสังคมแต่ละช่วงเวลา ด้วยเหตุนี้ทำให้ผู้วิจัยเห็นว่า การศึกษาถึงการนำเสนอเนื้อหาทางเฟซบุ๊กในประเทศไทย จะนำมาซึ่งความเข้าใจเกี่ยวกับข้อมูลทั่วไปและเนื้อหาของข้อมูลข่าวสารที่สอดแทรกในเฟซบุ๊ก เป็นองค์ประกอบที่สำคัญยิ่ง ของโลกสังคมออนไลน์ในโลกธุรกิจอิเล็กทรอนิกส์ ข้อมูลเหล่านั้นมาใช้ประโยชน์กลับเป็น

เรื่องที่ยุ่งยาก เนื่องจากข้อมูลที่จัดเก็บไว้นั้นเป็นข้อมูลดิบที่ไม่ได้ผ่านการคัดกรอง ประมวลผล หรือสกัด ทำให้ข้อมูลที่ได้ยังไม่ละเอียดเพียงพอ ตัวอย่างเช่น การสร้างขั้นตอนวิธีอย่างง่ายในการแยกประเภท คำวิจารณ์ว่า มีความหมายเชิงแนะนำหรือไม่ โดยการพยากรณ์จากการคำนวณค่าเฉลี่ยของกลุ่มคำคุณศัพท์หรือ คำกริยาวิเศษณ์ การจำแนกข้อความแบบอัตโนมัติโดยนำวิธีมาตรฐานในการคำนวณความหมาย จากตำแหน่งของ คำคุณศัพท์ในข้อความมาประยุกต์ใช้ [4]

ในงานวิจัยนี้ผู้วิจัยจึงสนใจที่จะศึกษาความคิดเห็นของลูกค้าที่เข้าใช้งานกลุ่มเฟซบุ๊กชื่อที่พักในจังหวัดลำปาง ในกลุ่มมียอดสมาชิกประมาณ 5 พันคน โดยเนื้อหาที่สนใจและต้องการตรวจสอบคือ การตั้งโพสต์และการแสดงความคิดเห็นที่เป็นเชิงบวก และเชิงลบ โดยได้ทดสอบความแม่นยำในการประเมินความน่าเชื่อถือ โดยการตรวจจับคำที่มีความหมายในเชิงบวกและเชิงลบ โดยเลือกคุณลักษณะที่นำมาทดลองสร้างโมเดลจำแนกข้อมูลมี การใช้ 3 อัลกอริทึมมาเปรียบเทียบหาค่าความแม่นยำ ได้แก่ อัลกอริทึมนาอีฟเบย์ อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจ แล้วเลือกโมเดลที่มีค่าความถูกต้องมากที่สุด มาเป็นโมเดลเพื่อพัฒนาระบบตรวจจับข้อคิดเห็นจากเฟซบุ๊กแบบเวลาจริง ความคิดเห็นของลูกค้าทั้งเชิงบวกและเชิงลบดังกล่าวจะสะท้อนกลับให้เจ้าของที่พักให้รักษามาตรฐานในการบริการและให้มีการปรับปรุงห้องพักให้ดีขึ้น เพื่อเพิ่มยอดจำนวนคนเข้าพักและปรับห้องพักและการบริการให้ตรงกับความต้องการของลูกค้ามากขึ้น

1. วัตถุประสงค์ของการวิจัย

1.1 เพื่อศึกษาและเปรียบเทียบการจำแนกตามความแม่นยำการสกัดคำบทวิจารณ์ออนไลน์

1.2 เพื่อพัฒนาระบบสกัดคำจากบทวิจารณ์ออนไลน์ของลูกค้า จากความคิดเห็นบนเฟซบุ๊กแบบเวลาจริง

2. เอกสารและงานวิจัยที่เกี่ยวข้อง

การประมวลผลภาษาธรรมชาติ คือ การแปลความจากภาษาธรรมชาติที่มนุษย์ใช้สื่อสารกันให้อยู่ในรูปแบบที่เป็นโครงสร้าง (Structured Data) ที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้ 2 แนวทาง [1] คือแนวทางการศึกษาและทำความเข้าใจกับโครงสร้างทางภาษาศาสตร์ และอีกแนวทางคือ อาศัยความรู้ด้านปัญญาประดิษฐ์ โดยการแทนความรู้ด้วยคลังคำ หน่วยคำหลายหน่วยคำประกอบกันเป็นคำ (Word) ที่มีความหมาย คำหลายคำประกอบกันเป็นวลี (Phrase) แบ่งเป็นนามวลี (Noun Phrase: NP) และกริยาวลี (Verb Phrase: VP) วลีหลายวลีประกอบกันเป็นประโยค (Sentence: S) ในการแบ่งประโยคออกเป็นส่วนดังนี้ [5]

$$S = NP + VP \quad (1)$$

$$NP = N \mid N + (ADJ) + (ADV) + (PP) \mid PRON \quad (2)$$

$$VP = V \mid V + (ADV) \mid AUX + V \mid VP + NP \quad (3)$$

$$PP = PREP + NP \mid PREP + VP \mid PREP + NP + VP \quad (4)$$

การจำแนกประเภท (Classification) คือ การทำเหมืองข้อมูล (Data Mining) คือ กระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น เป็นส่วนช่วยในการจัดเก็บและตีความหมาย ข้อมูลจากเดิมที่มีการจัดเก็บข้อมูลมาสู่การจัดเก็บในรูปแบบฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้มีเทคนิคที่นิยมใช้ คือ การจำแนกประเภทข้อมูล ซึ่งเป็นกระบวนการสร้างโมเดลเพื่อตรวจจับกลุ่มข้อมูลใหม่ โดยข้อมูลทั้งหมดจะมีการแบ่งเป็น 2 กลุ่มคือ กลุ่มข้อมูลสอน (Training Data) เป็นชุดข้อมูลที่มีบทบาทในการสร้างโมเดลจำแนกประเภทข้อมูลขึ้นมา และกลุ่มข้อมูลทดสอบ (Test Data) เป็นชุดข้อมูลประเมินความถูกต้องของโมเดลโดยกระบวนการสร้างตัวโมเดลจำแนกประเภทข้อมูล การสร้างโมเดลการจำแนกประเภท การประเมินผลโมเดลเป็นขั้นตอนตรวจสอบความถูกต้องโดยใช้ข้อมูลทดสอบ และปรับปรุงโมเดลจนกว่าจะได้ค่าความถูกต้อง [6]

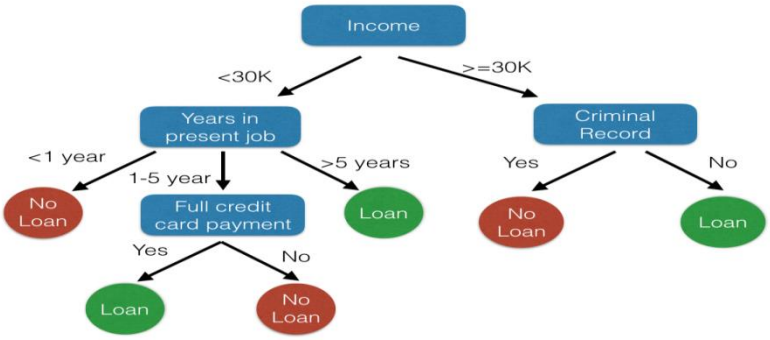
อัลกอริทึมนาอีฟเบย์ (Naive Bayes algorithm) คือ ตัวจำแนกประเภทแบบเบย์อย่างง่าย คือ โมเดล การจำแนกประเภทข้อมูลที่ใช้หลักความน่าจะเป็น อยู่บนพื้นฐานของ Bayes' Theorem และสมมติฐานที่ให้การเกิดของเหตุการณ์ต่าง ๆ เป็นอิสระต่อกัน ถ้ากำหนดให้ P(h) ความน่าจะเป็นที่จะเกิดเหตุการณ์ h และ P(h|D) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D จากตัวแปลที่กำหนด เราสามารถตรวจจับเหตุการณ์ที่พิจารณาจากเหตุการณ์ต่าง ๆ ได้ดังสมการที่ 5 [7]

$$P(h \mid D) = P(D \mid h) * P(h) / P(D) \quad (5)$$

จากสมการข้างต้น สามารถสร้างคำนวณความน่าจะเป็นของการจำแนกประเภทแบบเบย์ ดังสมการที่ (6)

$$P(d \mid h) = P(a, \ldots, a \mid h) = P(a \mid h) \quad (6)$$

อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree Classification) คือ กระบวนการทางด้านเหมืองข้อมูลนั้น (Data mining) ได้นำการตัดสินใจแบบโครงสร้างต้นไม้ช่วยในการทำงานด้านการตัดสินใจต่าง ๆ ไม่ว่าจะเป็นทางด้านระบบธุรกิจหรือด้านอื่น ๆ โดยปกติมักประกอบด้วยกฎในรูปแบบ "ถ้า เงื่อนไข แล้ว ผลลัพธ์" เช่น "If Income = High and Married = No THEN Risk = Poor" โดยลักษณะของการตัดสินใจแบบโครงสร้างต้นไม้นั้นมีลักษณะคล้ายกับต้นไม้กลับหัว โดยโหนดแรกสุดจะเป็นรากของต้นไม้ (Root node) แต่ละโหนดแสดงคุณลักษณะ (attribute) แต่ละกิ่งจะแสดงค่าผลในการทดสอบ และโหนดใบ (Leaf node) แสดงคลาสที่กำหนด [8] แสดงดังภาพที่ 1



ภาพที่ 1 ตัวอย่างต้นไม้ตัดสินใจ (Decision Tree)

อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbors Classification) หมายถึง วิธีที่ใช้ใน การจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ โดยการตรวจสอบจำนวนบางจำนวน" K" ในขั้นตอนวิธีการหาเพื่อนบ้านใกล้ที่สุดของกรณี หรือเงื่อนไขที่เหมือนกัน หรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวม (Count Up) ของจำนวนเงื่อนไข หรือกรณีต่าง ๆ สำหรับแต่ละคลาส และกำหนดเงื่อนไขใหม่ ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด การนำเทคนิคของขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดไปใช้นั้นเป็นการหาระยะห่างระหว่างแต่ละตัวแปร (Attribute) ในข้อมูลจากนั้นก็คำนวณค่าออกมา วิธีนี้จะเหมาะสำหรับข้อมูลแบบตัวเลขแต่ตัวแปรที่เป็นค่าแบบไม่ต่อเนื่องนั้น ก็สามารถทำได้แต่ต้องการการจัดการแบบพิเศษเพิ่มขึ้น [9]

การสกัดคำ (Text Extraction) เป็นกระบวนการของการทำเหมืองข้อความ เพื่อใช้การวิเคราะห์คำออกจากเอกสาร ข่าวสาร ข้อความ และสารสนเทศต่าง ๆ ที่เป็นตัวอักษร โดยสามารถนำไปทำการแบ่งกลุ่ม (Clustering) จำแนกกลุ่ม และหาความสัมพันธ์ขั้นตอนการแบ่งกลุ่มโดยใช้เทคนิคต่าง ๆ เช่น 1) การตัดคำ (Word Segmentation) เป็นการแยกแต่ละคำจากเอกสารออกจากกัน โดยยังคงมีความหมายที่ถูกต้องสมบูรณ์อยู่โดยการตัดคำนั้นใช้ฐานข้อมูลพจนานุกรมคำศัพท์ ในการแบ่งคำออกมา 2) การกำจัดคำหยุดเป็นการตัดคำที่ไม่มีความหมายออกจากเอกสาร โดยการกำจัดคำหยุดนั้นใช้ฐานข้อมูลคำศัพท์ที่เป็นคำที่ไม่มีความหมาย เช่น "กับ" "แก่" "แด่"

“ต่อ” “ที่” “ซึ่ง” เป็นต้น ในการกำจัดคำออก เมื่อทำการตัดคำเรียบร้อยแล้วจึงทำการเลือกคำที่ต้องการใช้ใน การวิเคราะห์ (Feature Selection) [2]

รวิสุดา เทศเมือง และนิเวศ จิระวิชิตชัย [3] ได้วิจัยเรื่อง การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน นำเสนอวิธีการจำแนกความคิดเห็นโดยการวิเคราะห์ความคิดเห็นภาษาไทย เกี่ยวกับการรีวิวสินค้าออนไลน์ ด้านการบริการห้องพัก โรงแรม รีสอร์ทจากอโกด้า ประเทศไทยและทวิตเตอร์ ประเทศที่จดทะเบียนหลักทรัพย์ โดยใช้เทคนิคเหมืองข้อความวิเคราะห์ ความคิดเห็นภาษาไทยโดยสร้างแบบจำลองด้วยอัลกอริทึม 4 วิธี ได้แก่ ซัพพอร์ตเวกเตอร์แมทชีน, ต้นไม้ตัดสินใจ, นาอีฟเบย์และวิธีการเพื่อนบ้านใกล้ที่สุด เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ จากการทดลองพบว่า คุณลักษณะที่ดีสุดคือ นาอีฟเบย์ ระดับรองลงมาเป็นต้นไม้ตัดสินใจ และวิธีการเพื่อนบ้านใกล้ที่สุด ซัพพอร์ตเวกเตอร์แมทชีนตามลำดับ

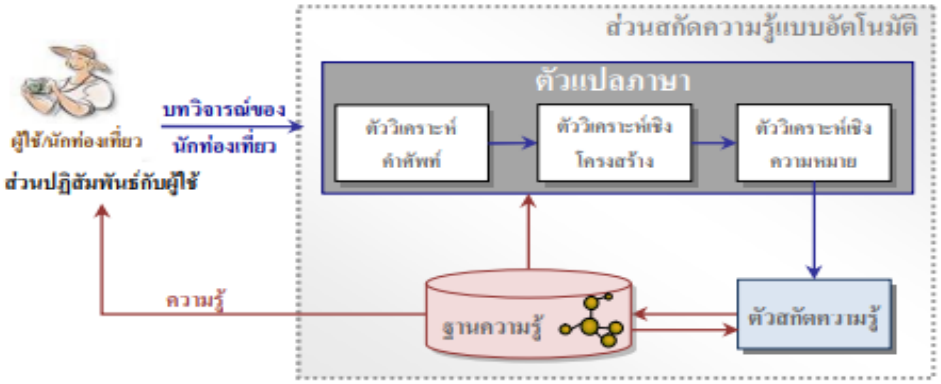
Pedro, André, Pedro and Mario. [7] ได้ทำการศึกษาการคัดเลือกคุณลักษณะสำหรับการวิเคราะห์ความรู้สึก โดยการวิเคราะห์ข้อความโดยการจัดหมวดหมู่ข้อความเป้าหมาย คือเพื่อกำหนดความเป็นกลาง (เชิงบวกหรือลบ) ของข้อความที่แสดงไว้ในเอกสาร โดยบทความนี้จะศึกษาเชิงเปรียบเทียบ เกี่ยวกับวิธีการคัดเลือกคุณลักษณะ โดยใช้อัลลักอริทึมในการจำแนกที่เป็นมาตรฐานสากลคือ นาอีฟเบย์และเทคนิคต้นไม้ตัดสินใจ นอกจากนั้นผู้วิจัยได้นำเสนอวิธีถ่วงน้ำหนัก สำหรับการคัดเลือกคุณลักษณะนาอีฟเบย์ ประสิทธิภาพการวิเคราะห์ซึ่งมีความเชื่อมั่นมากกว่าเทคนิคต้นไม้ตัดสินใจทั่วไป โดยเมื่อใช้ร่วมกับขบวนการคัดเลือกคุณลักษณะร่วมกัน

วิธีดำเนินการวิจัย

วิธีการดำเนินการวิจัยในงานวิจัยสำหรับงานวิจัยนี้ได้นำเอาขั้นตอนการทำเหมืองข้อมูลมาใช้เพื่อ การประยุกต์ใช้เทคนิคการจำแนกเพื่อพัฒนาประสิทธิภาพของแบบจำลองซึ่งมีขั้นตอนการดำเนินการดังต่อไปนี้

1. การสกัดคำ

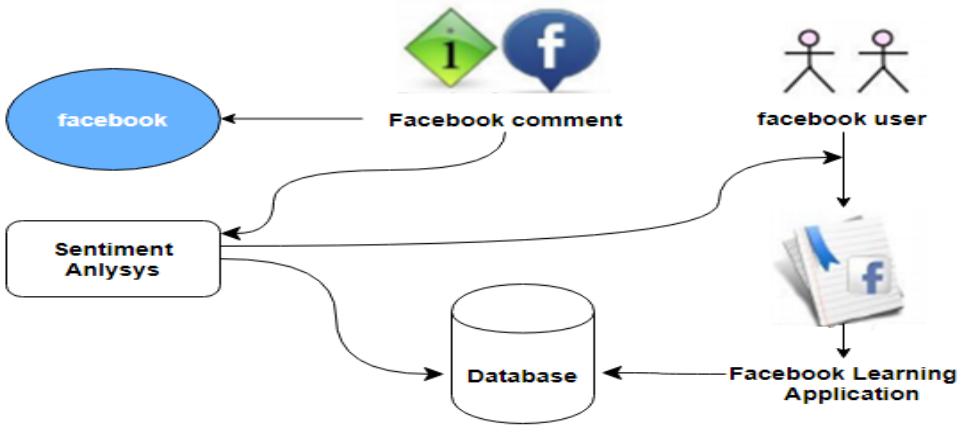
ได้เตรียมชุดข้อมูล การตัดคำหยุด การสกัดคำ การเตรียมชุดข้อมูล การวิเคราะห์คำศัพท์ วิเคราะห์โครงสร้างประโยค โดยมีการออกแบบขั้นตอนการดำเนินการ [6] แสดงดังภาพที่ 2



ภาพที่ 2 ตัวอย่างลำดับขั้นตอนการดำเนินงานการสกัดคำ

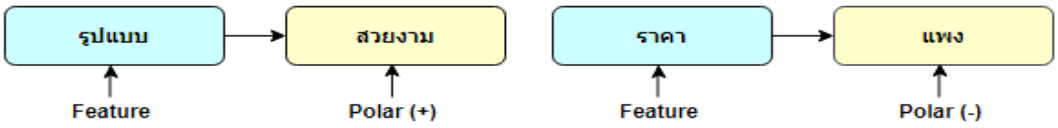
2. การเตรียมข้อมูล

ข้อมูลที่ได้ผู้วิจัยได้รวบรวม โพสต์ และความความคิดเห็นในเครือข่ายกลุ่มเฟซบุ๊กโรงแรมและที่พักจังหวัดลำปาง ในช่วง 5 ปีที่ผ่านมา ปัจจุบันมียอดสมาชิกในกลุ่มประมาณ 2 หมื่นคน ข้อมูลที่ใช้ในการทดสอบ ช่วงวันที่ 1 กุมภาพันธ์ 2558 ถึงวันที่ 28 กุมภาพันธ์ 2563 จำนวน 958 โพสต์ และแสดงความคิดเห็นรวม 13,564 ความคิดเห็น สำหรับวัดประสิทธิภาพโมเดล และนำอัลกอลิทึมที่มีความค่าความถูกต้องที่สุดมาพัฒนาระบบตรวจจับโดยเริ่มตรวจจับตั้งแต่วันที่ 1 มีนาคม ถึงวันที่ 30 มิถุนาคม 2563 จำนวน 1,056 ความคิดเห็น สำหรับใช้ในการทดสอบระบบ โดยมีการออกแบบขั้นตอนการดำเนินงาน แสดงดังภาพที่ 3



ภาพที่ 3 ตัวอย่างลำดับการรวบรวมข้อมูล

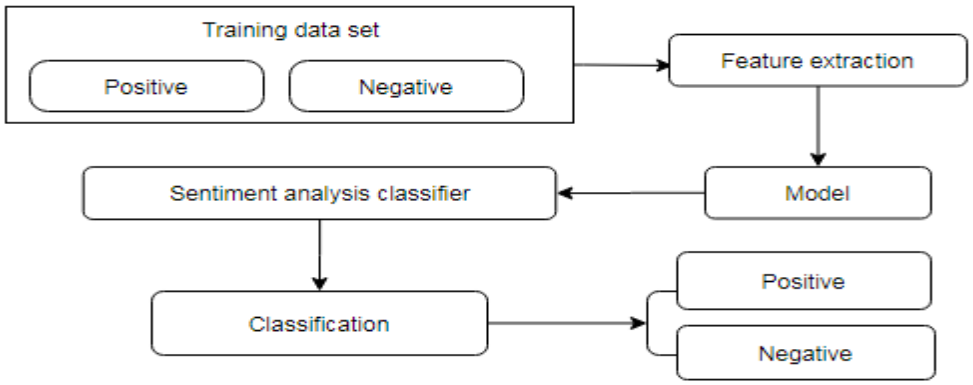
จากภาพที่ 2 มีการรวบรวมความคิดเห็นจากกลุ่มเฟซบุ๊ก แล้วทำการสกัดคำรูปประโยคที่มีเนื้อหา เชิงบวกและเชิงลบการคำนวณหาค่าคำสำคัญ ทำการวัดประสิทธิภาพและทำการพัฒนาระบบเพื่อสรุปผลข้อมูลโดยการสกัดความคิดเห็นแบ่งเป็นเชิงบวกและเชิงลบ จากคำศัพท์ที่มีการบันทึกไว้ ตัวอย่างดังภาพที่ 4



ภาพที่ 4 ตัวอย่างการกำกับคำคุณลักลักษณะและแสดงขั้วความคิดเห็น

3. ขั้นตอนการสร้างแบบจำลอง

ขั้นตอนก่อนการสร้างแบบจำลองเป็นขั้นตอนที่ใช้ในการวิเคราะห์ข้อมูล เพื่อเตรียมข้อมูลเข้าสู่กระบวนการจำแนกข้อความแสดงความคิดเห็น โดยจะนำข้อความคิดเห็นที่ผ่านกระบวนการคัดแยกข้อความที่เกี่ยวข้องกับข้อคิดเห็นและผ่านกระบวนการตัดคำ การกำจัดคำหยุด แล้วนำมาทดสอบการจำแนกกับโมเดลที่มีการเลือกไว้สำหรับการทดสอบเปรียบเทียบหาค่าความถูกขั้นตอนการดำเนินงาน แสดงดังภาพที่ 5



ภาพที่ 5 ขั้นตอนดำเนินงานการจำแนกความคิดเห็น

4. การวัดประสิทธิภาพแบบจำลอง

การประเมินประสิทธิภาพจะดูจากผลการทำนายของโมเดล ซึ่งหาได้จาก confusion matrix แสดงโดยเมตริกซ์นี้ จะเป็นการประเมินผลลัพธ์การทำนายกับผลลัพธ์จริง ๆ ที่หาได้ ค่าความระลึกและค่าความความถูกต้องมาคำนวณด้วยสมการที่ 7 ถึงสมการที่ 9 ตามลำดับ [10] ดังตารางที่ 1

ตารางที่ 1 confusion matrix

	predict class	
actual class	class=Yes	class=No
class=Yes	TP	FN
class=No	FP	TN

$$\text{accuracy}=\frac{TP+TN}{(TP+TN+FP+FN)}\text{x}100\% \quad (7)$$

$$\text{precision}=\frac{TP+TN}{(T+FP)}\text{x}100\% \quad (8)$$

$$\text{recall}=\frac{TP}{(TP+FN)}\text{x}100\% \quad (9)$$

TP (true positive) หมายถึง ข้อมูลที่สามารถสกัดได้และมีความเกี่ยวข้อง; FP (false positive) หมายถึง ข้อมูลที่สามารถสกัดได้ แต่ไม่มีความเกี่ยวข้อง TN (true negative) หมายถึง ข้อมูลที่ไม่สามารถสกัดได้และไม่มีความเกี่ยวข้อง และ FN (false negative) หมายถึง ข้อมูลที่ไม่สามารถสกัดได้แต่มีความเกี่ยวข้องการศึกษา

ผลการวิจัย

1. ผลการเปรียบเทียบการจำแนกตามความแม่นยำการสกัดคำบทวิจารณ์ออนไลน์

จากการทดลองโมเดลการตรวจจับความคิดเห็นจำแนกเป็นขั้วความคิดเห็นประกอบด้วยความคิดเห็นเชิงบวก และเชิงลบ วัดค่าประสิทธิภาพ ความแม่นยำ ค่าความระลึก ค่าความถูกต้อง ดังแสดงในตารางที่ 2

ตารางที่ 2 ผลการตรวจจับค่าความแม่นยำ ค่าความระลึก และค่าความถูกต้อง

model	confusion matrix			evaluation		accuracy
naïve Bayes		correct label		precision	recall	87.97
		positive	negative			
	positive	9578	2152	83.54	85.27	
	negative	2042	9688	84.20	84.23	
k-Nearest neighbors		correct label		precision	recall	83.80
		positive	negative			
	positive	9435	2295	78.32	81.65	
	negative	2310	9420	80.73	81.57	
decision tree		correct label		precision	recall	79.89
		positive	negative			
	positive	8985	2745	76.81	79.45	
	negative	2750	8980	77.64	78.95	

จากตารางที่ 2 พบว่า เทคนิควิธีนาอีฟเบย์ มีประสิทธิภาพสูงสุด มีค่าเฉลี่ยความถูกต้อง ร้อยละ 87.97 เทคนิควิธีการหาเพื่อนบ้านใกล้ที่สุด มีค่าเฉลี่ยความถูกต้อง ร้อยละ 83.80 และเทคนิควิธีการเรียนรู้ของต้นไม้ตัดสินใจ มีค่าเฉลี่ยความถูกต้อง ร้อยละ 79.89 ผลการทดลองพบว่า วิธีนาอีฟเบย์ มีค่าเฉลี่ยความถูกต้องมากที่สุด

นำไปพัฒนาระบบเนื่องจาก การทดลองโมเดลการตรวจจับความคิดเห็นจำแนกเป็นขั้วความคิดเห็น (Polarity) ประกอบด้วยความคิดเห็นเชิงบวก และ ความความคิดเห็นเชิงลบ ทำให้ได้ค่าประสิทธิภาพซึ่งประกอบไปด้วยความแม่นยำ และค่าความระลึก ค่าความถูกต้อง ดังแสดงในตารางที่ 3

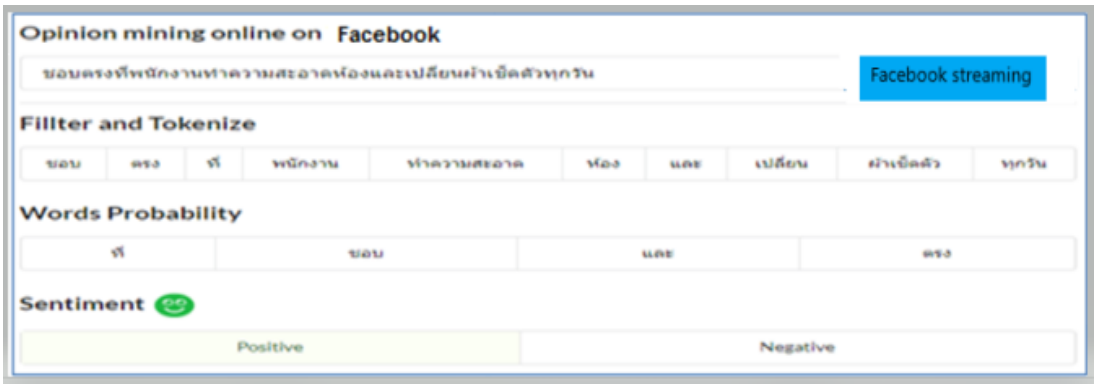
ตารางที่ 3 ผลการตรวจจับค่าความคิดเห็น ค่าความแม่นยำ และค่าความระลึก ค่าความถูกต้อง

Model	Confusion Matrix	Polarity	
		Positive	Negative
Naïve Bayes	Precision	83.54	84.20
	Recall	85.27	84.23
	Accuracy	83.08	83.86
K-Nearest Neighbors	Precision	78.32	80.73
	Recall	81.65	81.57
	Accuracy	78.32	79.28
Decision Tree	Precision	76.81	77.64
	Recall	79.45	78.95
	Accuracy	75.10	74.68

จากตารางที่ 3 ในการทำนายบทวิจารณ์ออนไลน์ด้วยเทคนิคเหมืองข้อมูล 3 โมเดลด้วยเทคนิควิธีนาอีฟเบย์, วิธีการหาเพื่อนบ้านใกล้ที่สุด และวิธีการเรียนรู้ของต้นไม้ตัดสินใจ ผลที่ได้โมเดลที่สร้างด้วยเทคนิควิธีนาอีฟเบย์ มีประสิทธิภาพสูงสุดมีค่าเฉลี่ยความถูกต้องร้อยละ 82.97 เทคนิควิธีการหาเพื่อนบ้านใกล้ที่สุด มีค่าเฉลี่ยความถูกต้องร้อยละ 78.80 และ เทคนิควิธีการเรียนรู้ของต้นไม้ตัดสินใจ มีค่าเฉลี่ยความถูกต้องร้อยละ 74.89 จากการเปรียบเทียบประสิทธิภาพจำลองพบว่า วิธีนาอีฟเบย์ มีค่าความถูกต้องมากที่สุด ผู้วิจัยได้เลือกวิธีนี้เพื่อนำไปพัฒนาระบบต่อไป

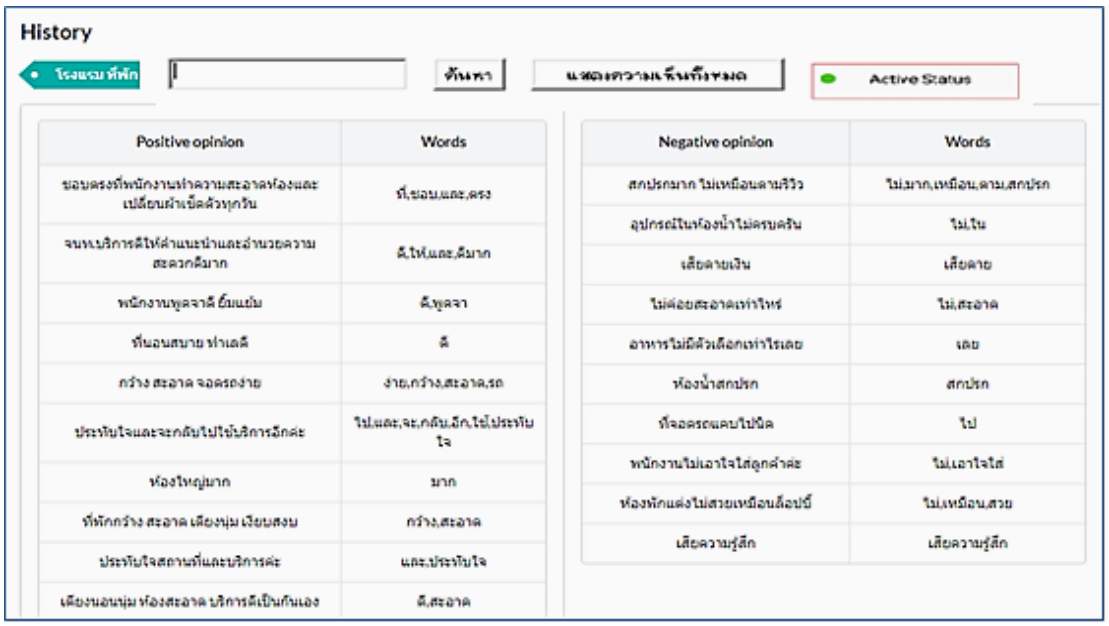
2. ผลการพัฒนาระบบสกัดคำจากบทวิจารณ์ออนไลน์ของลูกค้า

การทดสอบระบบมีการจำแนกข้อมูลและการสกัดคำ เป็นการนำบทวิจารณ์ออนไลน์มาสกัดความคิดเห็นแบ่งเป็นเชิงบวกและเชิงลบจากคำศัพท์ที่มีการบันทึก ผลการทดสอบดังภาพที่ 6



ภาพที่ 6 ผลการทดสอบการสกัดคำที่ใช้อัลกอริทึมนาอีฟเบย์ (Naive Bayes)

จากภาพที่ 6 ผู้วิจัยได้เลือกอัลกอริทึมนาอีฟเบย์ที่มีค่าถูกต้องดีที่สุด มาทดสอบข้อมูลจากบทวิจารณ์ออนไลน์แบบเวลาจริง ผลพบว่า สามารถสกัดคำจากบทวิจารณ์จากรูปประโยคได้เป็นอย่างดี จากนั้นได้ทำการนำระบบที่ได้ไปทดสอบกับเครือข่ายเฟซบุ๊กโรงแรมและที่พักจังหวัดลำปาง ผลการทดสอบพบว่า ระบบสามารถแยก บทวิจารณ์ที่เป็นเชิงบวกและเชิงลบตามเวลาจริงได้อย่างถูกต้อง ดังภาพที่ 7



ภาพที่ 7 ผลการทดสอบการสกัดคำแยกข้อมูลบทวิจารณ์ออนไลน์

จากภาพที่ 7 ผู้วิจัยได้ทำการพัฒนาระบบจัดเก็บข้อมูลบทวิจารณ์ออนไลน์ โดยทำการทดสอบข้อมูลแบบเวลาจริงพบว่า ระบบสามารถค้นหาบทวิจารณ์ออนไลน์แยกตามโรงแรมและที่พัก แยกตามช่วงวันที่ และแยกตามประเด็นที่มีการวิจารณ์ได้ รวมทั้งสามารถสกัดคำและทำการแยกความคิดเห็นที่เป็นเชิงลบและเชิงบวกได้อย่างมีประสิทธิภาพ

การอภิปรายผลการวิจัย

งานวิจัยนี้ได้พัฒนาระบบสกัดคำจากบทวิจารณ์ออนไลน์ของลูกค้า โดยประมวลผลที่ได้จากการสกัดความรู้พบว่า ผลการวัดค่าความถูกต้อง อัลกอริทึมนาอีฟเบย์ อยู่ที่ร้อยละ 82.97 อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด อยู่ที่ร้อยละ78.80 และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจนั้นจะอยู่ที่ร้อยละ 74.89 ผู้วิจัยได้นำอัลกอริทึมนาอีฟเบย์มาพัฒนาระบบ โดยทดสอบกับบทวิจารณ์จากกลุ่มเฟซบุ๊ก ผลสรุปพบว่า ระบบสามารถแยกบทวิจารณ์ที่เป็นเชิงบวกและเชิงลบได้เป็นอย่างดีและผลงานวิจัยสามารถแยกประเด็นบทวิจารณ์เชิงบวกและเชิงลบสามารถแยกบทวิจารณ์โรงแรมแต่ละแห่ง และภาพรวมของธุรกิจโรงแรมในจังหวัดลำปางทั้งหมดได้

ข้อเสนอแนะ

สำหรับแนวทางในการพัฒนางานวิจัยต่อไปในอนาคต ให้ผลลัพธ์มีความถูกต้องมากที่สุด เช่น การเพิ่มเติมคำศัพท์ให้มากขึ้น ทั้งในส่วนของคำที่มีความหมายเหมือนกัน แต่เขียนต่างกัน คำคุณศัพท์ คำกริยาวิเศษณ์ ความรู้ที่ส่งผลต่อการคำนวณระดับคะแนนและการสกัด ระบบสกัดความรู้แบบอัตโนมัติที่นำเสนอนี้จะมีประโยชน์อย่างยิ่งหากผู้ประกอบการธุรกิจอิเล็กทรอนิกส์นำไปประยุกต์ใช้เก็บและวิเคราะห์ความคิดเห็นของลูกค้าเกี่ยวกับสินค้าและบริการ แล้วนำข้อมูลเหล่านั้นมาพัฒนาสินค้าหรือบริการของตนเองให้ตอบสนอง ความต้องการของลูกค้าให้มากที่สุด นอกจากนี้ระบบจะช่วยให้ลูกค้าสามารถนำความรู้ในเรื่องที่สกัดไปใช้ประกอบการตัดสินใจได้อย่างสะดวกและรวดเร็วอีกด้วย

เอกสารอ้างอิง

[1] กนกวรรณ เขียนวรรณ. (2555). *การประมวลผลภาษาธรรมชาติ*. ค้นวันที่ 3 ธันวาคม 2561 เข้าถึงได้จาก : www.mbs.mut.ac.th/paper/pdf/29.pdf

[2] สมนึก สิทธิปวน. (2546). *การวิเคราะห์กระจายคำในประโยคภาษาไทย โดยการโปรแกรมเชิงเจนเนติก*. (วิทยานิพนธ์ปริญญามหาบัณฑิต) สถาบันบัณฑิตพัฒนบริหารศาสตร์

[3] รวิสุดา เทศเมือง นิเวศ จิระวิชิตชัย. (2017). *การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้า ออนไลน์ โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน*. Engineering Journal of Siam University. ปีที่18 ฉบับที่.34 (2017)

[4] M. G. Miniwatts. (2010, Dec. 7). *World Internet Usage and Population Statistics*. Online].Available: http://www.internetworldstats.com/stats3.htm World Tourism Organization, Yearbook of Tourism Statistics: Data 2001-2005 the 59th Edition, Madrid: WTO, 2007, pp. 8-9.

[5] P.D. Turney, *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*, in Proc. the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 417-424.

[6] M. Taboada and J. Grieve .*Analyzing Appraisal Automatically*. in Proc. the AAAI Spring Symposium, 2004, pp. 158-161.

[7] K. Zhang, R. Narayanan, and A. Choudha. *Voice of the Customers: Mining Online Customer Reviews for Product Feature-Based Ranking*. in Proc. the 3rd conference on Online social networks (WOSN'10), 2010, pp. 11.

[8] M. Hu and B. Liu. *Mining and Summarizing Customer Reviews*. in Proc. the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 168-177.

[9] T.L. Saaty. *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*. (2nd edition), Pittsburgh, PA: RWS, 2006, pp. 4-5.

[10] Thongjor N. *Machine Learning#2 [Internet].Babel Coder*. 2017. Available from: https://www.babelcoder.com/blog/posts/k-nearest-neighbors