WALAILAK PROCEDIA
NCIs2019; 2019(7): NCIs34
http://wjst.wu.ac.th/index.php/wuresearch

แบบจำลองการวิเคราะห์ความรู้สึกแบบผสมสำหรับความคิดเห็นต่อโรงแรมในประเทศไทยโดยใช้ K-means และ K-NN

A Hybrid Sentiment Analysis Model for Thailand Hotel Review Using K-means and K-NN

วาทิตย์ คำพรมมา*, จักรชัย โสอินทร์ และ เพชร อิ่มทองคำ

ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น ขอนแก่น 40002

*watit_kh@kkumail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อวิเคราะห์ความรู้สึกเกี่ยวกับความคิดเห็นต่อโรงแรมในประเทศไทย งานวิจัยได้ประยุกต์ใช้เทคนิคการวิเคราะห์ความรู้สึกในรูปแบบเหมืองความคิดเห็น (opinion mining) ต่อการเข้ารับบริการ งานวิจัยนี้มีการรวบรวมความคิดเห็นจากเว็บไซต์ APT TUBE จำนวน 10,000 ประโยคแล้วจึงทำการวิเคราะห์ความคิดเห็น (sentiment analysis) ด้วยเทคนิค K-means โดยแบ่งออกเป็น 5 กลุ่มได้แก่ การเข้าถึง (accessibility), กิจกรรมและความบันเทิง (activities & entertainment), อาหารและเครื่องดื่ม (food & beverage), พนักงานผู้ให้บริการ (staff), สถานที่ (place) จากนั้นจึงนำข้อความแต่ล่ะกลุ่มเข้าสู่การแยกประเภทด้วย K-Nearest Neighbors (K-NN) และวัดประสิทธิภาพด้วยค่าความเที่ยง (Precision), ค่าความระลึก (Recall), ค่าความถูกต้อง (Accuracy), ค่า F-Measure ร่วมกับ 10 Fold Cross Validation โดยเปรียบเทียบกับเทคนิค Decision tree, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN) โดยเทคนิคที่ให้ค่าความถูกต้องมากที่สุดคือ เทคนิค K-means ร่วมกับ K-NN ให้ค่าความถูกต้องสูงที่สุดที่ 94.8%

คำสำคัญ: การทำเหมืองข้อมูล, การวิเคราะห์ความรู้สึก, ความคิดเห็นต่อโรงแรม, เหมืองความคิดเห็น

ABSTRACT

This research aims to analyze the feelings about the opinions of hotels in Thailand. The research applying opinion mining technique segment calls and analyze the ideas of service users towards hotel in Thailand. This research has collection of comments from the APT TUBE website in the amount of 10,000 sentences, the sentiment analysis by K-means dividing into 5 groups such as accessibility, activities & entertainment, food & beverage, staff, and place. From there, the message is brought into each group by K-Nearest Neighbors (K-NN) and performance measurement with values Precision, Recall, Accuracy, F-Measure together with 10 Fold Cross Validation by comparison 3 techniques are decision tree , Support Vector Machine (SVM), K-Nearest Neighbors (K-NN) the technique that the K-means technique together with K-NN providing the highest accuracy at 94.8%.

Keyword: Data Mining, Sentiment Analysis, Hotel Review, Opinion Mining

บทนำ

การวิเคราะห์ความรู้สึก (Sentiment Analysis) [1] เป็นศาสตร์แขนงหนึ่งด้านการวิเคราะห์และตรวจสอบความรู้สึกที่มีขั้นตอนวิธีและกระบวนการที่มุ่งเน้นการวิเคราะห์และตรวจสอบความรู้สึกของผู้ใช้บริการของสินค้าหรือบริการ โดยเฉพาะจากข้อความที่ผู้ใช้บริการเหล่านั้นเขียนหรือแสดงความคิดเห็น เพื่อเป็นการบ่งบอกความรู้สึกของตนเองที่มีต่อสินค้าและบริการ เช่น ความรู้สึกดีหรือชื่นชอบ หรือทางบวก (Positive or Good) และในทางตรงกันข้าม ความรู้สึกที่ไม่ดีหรือไม่ชอบ หรือทางลบ (Negative or Bad)

การวิเคราะห์ความรู้สึกของผู้เข้ารับบริการสามารถนำมาประยุกต์ใช้ในการวิเคราะห์ข้อมูลเพื่อใช้ในการปรับปรุงคุณภาพการให้บริการของธุรกิจต่างๆ เนื่องจากความเห็นที่ผู้ใช้บริการเขียนไว้นั้นจะบอกได้ถึงความรู้สึกที่แท้จริงของผู้ใช้บริการในการให้บริการของโรงแรม ซึ่งผู้ประกอบการหรือเจ้าของธุรกิจสามารถนำเอาความคิดเห็นที่มีต่อบริการนำไปเป็นองค์ประกอบสำคัญในการปรับปรุง

การให้บริการหรือแม้แต่ใช้ในการสร้างฐานข้อมูลผู้ใช้บริการเพื่อประโยชน์ในด้านการประชาสัมพันธ์และการส่งเสริมการขายในอนาคต เมื่อข้อความเหล่านั้นมีจำนวนมากและไม่สามารถบอกได้ข้อความนั้นในด้านใดแล้วเป็นความคิดเห็นที่บอกว่าชื่นชอบ หรือไม่ชอบ จะทำให้การวางแผนในการปรับปรุงการให้บริการผิดผลาดได้

ดังนั้นงานวิจัยนี้จึงนำเสนอเทคนิคการวิเคราะห์ข้อความที่แสดงความคิดเห็นของผู้ใช้บริการในด้าน การเข้าถึง(accessibility), กิจกรรมและความบันเทิง (activities & entertainment), อาหารและเครื่องดื่ม (food & beverage), พนักงานผู้ให้บริการ (staff), สถานที่ (place)[2] ต่อผู้เข้าใช้บริการ เช่น มีความรู้สึกดีหรือไม่ดี ต่อบริการ โดยงานวิจัยนี้ประยุกต์ใช้เทคนิคการให้น้ำหนักค่าความถี่ TF-IDF (Term Frequency – Inverted Document Frequency) [3] เพื่อกำหนดค่าน้ำหนักของคำแต่ละคำแล้วนำผลรวมของคำแต่ละประโยคเข้าสู่กระบวนวิเคราะห์ความคิดเห็น (Sentiment Analysis) ที่จะระบุถึงข้อความที่แสดงความรู้สึกว่า “ชอบ” หรือ “ไม่ชอบ” โดยเปรียบเทียบ 4 ขั้นตอนวิธีได้แก่ Decision tree, Support Vector Machine (SVM), K-Nearest Neighbors (K-NN) และใช้วิธีการจัดกลุ่มด้วยเทคนิค K-means (K-means algorithm) ร่วมกับ K-Nearest Neighbors (K-NN) หลังจากนั้นจึงสรุปและเปรียบเทียบความถูกต้องของการวิเคราะห์ของแต่ละเทคนิค

ทฤษฎีที่เกี่ยวข้อง

เหมืองข้อมูล (Data Mining)

เหมืองข้อมูล[4] เป็นวิธีการในการค้นหาความรู้ (Knowledge) ในข้อมูลขนาดใหญ่ โดยที่เป็นกระบวนการที่กระทำกับข้อมูลจำนวนมากเพื่อค้นหารูปแบบ (Pattern) และความสัมพันธ์ (Association) ที่ซ่อนอยู่ในชุดข้อมูล โดยกระบวนการทำงานของเหมืองข้อมูลถูกพัฒนาขึ้นในปี ค.ศ. 1996 โดยแบ่งกระบวนการทำงานออกเป็น 6 ขั้นตอน หรือที่เรียกว่า Cross-Industry Standard Process for Data Mining (CRISP-DM) [5] ดังนี้

- **การทำความเข้าใจกับปัญหาและวิเคราะห์ปัญหา (Business understanding):** เป็นการเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางเหมืองข้อมูลพร้อมทั้งวางแผนในการดำเนินการ
- **การทำความเข้าใจกับข้อมูล (Data understanding):** เป็นกระบวนการทำงานเริ่มจากการเก็บรวบรวมข้อมูล แล้วจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล
- **การเตรียมข้อมูล (Data preparation):** เป็นการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา หรือข้อมูลดิบ (Raw data) ให้เป็นข้อมูลที่สามารถนำไปวิเคราะห์ได้ โดยการแปลงข้อมูลให้ถูกต้อง หรือที่เรียกว่า การทำความสะอาดข้อมูล (Data cleaning)
- **การสร้างโมเดล (Modeling):** เป็นการวิเคราะห์ข้อมูลด้วยเทคนิคทางเหมืองข้อมูล เช่น การจำแนกประเภทข้อมูล หรือการแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้อาจจะต้องมีการย้อนกลับไปที่ขั้นตอนการเตรียมข้อมูลเพื่อทำการแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิค
- **การประเมินรูปแบบ (Model evaluation):** เป็นการได้มาซึ่งผลการวิเคราะห์ข้อมูลด้วยเทคนิคทางเหมืองข้อมูล ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อไป ซึ่งจะต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ให้ตรงกับวัตถุประสงค์ที่ตั้งไว้ในขั้นตอนแรก หรืออีกนัยหนึ่งก็คือ มีความน่าเชื่อถือมากน้อยเพียงใด
- **การนำไปใช้ (Deployment):** เป็นการนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริง

การวิเคราะห์เหมืองความคิดเห็น (Opinion Mining)

การวิเคราะห์เหมืองความคิดเห็น เป็นกระบวนการวิเคราะห์ข้อคิดเห็นที่ผู้บริโภคนำเสนอในรูปของข้อความ ซึ่งปรากฏส่วนใหญ่บนเครือข่ายสังคมออนไลน์ (social network) โดยมีวัตถุประสงค์หลักเพื่อให้ทราบถึงความพึงพอใจในแง่บวกหรือลบ โดยกระบวนการดังกล่าวได้นำความรู้ทางด้านการค้นหาลักษณะแฝงของข้อมูล (knowledge discovery) มาประยุกต์ใช้ หรือการทำเหมือนข้อมูลนั้นเอง [6] แต่เนื่องจากข้อมูลที่นำมาวิเคราะห์มีลักษณะเป็นข้อความ กระบวนการค้นหาลักษณะแฝงจึงถูกเรียกว่าการวิเคราะห์เหมืองข้อความ

(text mining) นำมาวิเคราะห์ที่มีลักษณะของ ภาษาธรรมชาติ ดังนั้นจึงต้องนำความรู้ในเรื่องการประมวลผลภาษาธรรมชาติมาใช้ร่วมด้วย Natural Language Processing (NLP) [7]

การวิเคราะห์ความรู้สึก (Sentiment analysis)

การวิเคราะห์ความรู้สึก [1] เป็นวิธีการในการประมวลผลภาษาธรรมชาติ ที่มีขั้นตอนและกระบวนการที่มุ่งเน้นการวิเคราะห์และตรวจสอบความรู้สึกของผู้ใช้บริการของสินค้าหรือบริการนั้น ๆ จากข้อความที่ผู้ใช้บริการเหล่านั้นเขียนหรือแสดงไว้ เพื่อเป็นการบ่งบอกความรู้สึกของตนเองที่มีต่อสินค้าและบริการ เช่น ความรู้สึกดีหรือชื่นชอบ และ ความรู้สึกที่ไม่ดีหรือไม่ชอบโดยสามารถจำแนกประเภทด้วยเทคนิคการวิเคราะห์เหมืองความคิดเห็น (opinion mining) เป็นองค์ประกอบสำคัญในการปรับปรุงการให้บริการหรือแม้แต่ใช้ในการสร้างฐานข้อมูลผู้ใช้บริการเพื่อประโยชน์ในด้านการประชาสัมพันธ์และการส่งเสริมการขายในอนาคต

Term Frequency – Inverted Document Frequency (TF-IDF)

การคำนวณค่า TF-IDF [3] เป็นวิธีการให้น้ำหนักของคำที่นิยมใช้กันมากในทางการสืบค้นสารสนเทศ และ Text mining ซึ่งเป็นวิธีการทางด้านสถิติที่ใช้ในการประเมินความสำคัญของคำต่อเอกสาร ซึ่งจะคำนวณโดยเป็นสัดส่วนโดยตรงกับจำนวนที่คำนั้นปรากฏในเอกสารนั้น ๆ โดยการพิจารณาค่าความสำคัญด้วยด้วยการนับความถี่ของคำ TF (Term Frequency) ที่ปรากฏในกลุ่มเอกสารทั้งหมด และ IDF (Inverted Document Frequency) คือ การวัดจำนวนคำที่หาได้โดยการนำจำนวนเอกสารทั้งหมดในกลุ่มเอกสาร มาหารด้วยจำนวนเอกสารที่พบคำ โดยมีสูตรในการคำนวณดังนี้

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \quad ((1)$$

$tf_{i,j}$ คือ จำนวนครั้งที่คำนั้น ๆ ปรากฏในเอกสาร
$n_{i,j}$ คือ จำนวนคำนั้น ๆ ในเอกสารที่ d_j
$\sum_k n_{k,j}$ คือ จำนวนของทุกคำในเอกสารที่ d_j

$$idf_j = log \frac{|D|}{|\{d_j : t_j \in d_j\}|} \quad ((2)$$

$|D|$ คือ จำนวนเอกสารทั้งหมดในกลุ่มเอกสาร
$|\{d_j : t_j \in d_j\}|$ คือ จำนวนเอกสารที่คำนั้น ๆ ปรากฏ

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad ((3)$$

การแบ่งกลุ่มข้อมูลแบบค่ามาตรฐานเคมีน (K-means algorithm)

การแบ่งกลุ่มข้อมูลแบบค่ามาตรฐานเคมีน [8] เป็นวิธีหนึ่งในวิธีการแบ่งเวกเตอร์ (vector quantization) ที่มีรากฐานมาจากการประมวลผลสัญญาณ โดยวิธีนี้เป็นที่นิยมสำหรับการแบ่งกลุ่มข้อมูล (cluster analysis) โดยที่การแบ่งกลุ่มข้อมูลแบบค่ามาตรฐานเคมีนใช้สำหรับการแบ่งการสังเกตจำนวน N สิ่งเป็น K กลุ่ม โดยแต่ละการสังเกตจะอยู่ในกลุ่มที่มีค่าเฉลี่ยที่ใช้เป็นต้นแบบใกล้เคียงกันที่สุดอยู่ในกลุ่มเดียวกันตามจำนวน K กลุ่ม

K-Nearest Neighbors (K-NN)

K-Nearest Neighbors [1] คือ วิธีการในการจัดแบ่งกลุ่ม โดยเทคนิคนี้จะทำการตัดสินใจว่ากลุ่มใดที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง โดยการตรวจสอบจำนวนบางจำนวน เช่น “K” ใน K-nearest neighbor ของกรณีหรือเงื่อนไขที่เหมือนกันหรือใกล้เคียงกันมากที่สุด โดยจะหาผลรวมของจำนวนเงื่อนไข หรือกรณีต่าง ๆ ในแต่ละคลาส และกำหนดเงื่อนไขใหม่ ๆ ให้คลาสที่เหมือนกันกับคลาสที่ใกล้เคียงกันมากที่สุด

ต้นไม้ตัดสินใจ (Decision Tree)

เป็นขั้นตอนวิธีการสร้างกฎจากต้นไม้ตัดสินใจ โดยมีจุดเด่น [10] คือใช้กับข้อมูลที่มีลักษณะต่อเนื่องและข้อมูลไม่ต่อเนื่องได้, กับข้อมูลชุดฝึกสอนที่ไม่มีค่าของคุณลักษณะได้ โดยจะทำเครื่องหมายที่ไม่มีคุณลักษณะเป็นเครื่องหมาย "?", ใช้งานได้กับค่าที่มีความผิดปกติหรือมีความเสียหาย และ สามารถปรับแต่งต้นไม้ตัดสินใจในขณะที่กำลังสร้าง นอกจากนี้ วิธีการกำหนดคุณลักษณะสำคัญ จะเป็นการเลือกข้อมูลตามลำดับของตัวชี้วัดค่ามาตรฐานเกน Gain สูงที่สุด เป็นข้อมูลเริ่มต้นและข้อมูลถัดไปที่มีค่าน้อยลงตามลำดับ ซึ่งค่านี้สามารถคำนวณได้โดยใช้ความรู้จากทฤษฎีสารสนเทศ Information Gain (IG) อ้างอิงจาก $Info_A(D)$ ดังสมการที่ 4

$$Info(D) = -\sum_{i=1}^{m} p_i \, log_2(P_i) \tag{4}$$

P_i คือ ความน่าจะเป็นในข้อมูล D อยู่ในกลุ่ม Ci ซึ่งมีค่า $|Ci, D|/|D|$
m คือ จำนวนกลุ่มทั้งหมดที่ต่างกันของข้อมูล
Ci คือ กลุ่มลำดับที่ i โดยที่ i มีค่าระหว่าง 1 ถึง m
$|Ci, D|$ คือ จำนวนข้อมูลข้อมูล D ที่อยู่ในกลุ่ม Ci
$|D|$ คือ จำนวนข้อมูลในฐานข้อมูล D

เมื่อทำการพิจารณาเลือกคุณลักษณะเป็นตัวเลือกจะใช้ค่าความรู้จาก IG ซึ่งสามารถคำนวณ ค่า $Info_A(D)$ ดังสมการที่ 5

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times info(D_j) \tag{5}$$

v คือ จำนวนค่าที่เป็นไปได้ของคุณลักษณะ
$|D|$ คือ จำนวนข้อมูลในฐานข้อมูล D
$|D_j|$ คือ จำนวนข้อมูลในฐานข้อมูล D ที่มีค่าที่ j ของคุณลักษณะ A

นอกจากนี้จะใช้พิจารณาเลือกคุณลักษณะของ $Gain(A)$ มาเป็นโหนดของต้นไม้จะมีค่าเท่ากับผลต่างของความรู้จากทฤษฎีสารสนเทศกับ ความรู้จากทฤษฎีสารสนเทศของคุณลักษณะ ซึ่งสามารถคำนวณ ค่า $Gain(A)$ ได้ดังสมการที่ 6

$$Gain(A) = Info(D) - Info_A(D) \tag{6}$$

ในส่วนท้ายสุดการสร้างต้นไม้ตัดสินใจจะต้องพิจารณาคุณลักษณะที่มีของข้อมูลเรียนรู้เพื่อเลือกเป็นโหนดราก แล้วจึงตัดสินใจเลือกคุณลักษณะที่มีค่ามาตรฐานเกนสูงที่สุดนั้นเอง

Support Vector Machine (SVM)

Support Vector Machine [9] เป็นตัวแบบที่ใช้ในการระบุตัวบุคคล หรือ Object โดย SVM จะทำการแบ่งชั้นของข้อมูลด้วยระนาบหลายมิติจากข้อมูล 2 กลุ่มชุดข้อมูล โดยตัวแบบของ SVM เกี่ยวข้องกับเครือข่ายประสาทเทียม โดยใช้ Sigmoid Kernel Function ซึ่งจะมีค่าเท่ากันทั้ง 2 ระนาบ โดย SVM จะทำการแบ่งชั้นของข้อมูลด้วยระนาบหลายมิติจากข้อมูลเป็น 2 กลุ่ม

งานวิจัยที่เกี่ยวข้อง

ในส่วนนี้จะมุ่งเน้นการอธิบายเฉพาะการประยุกต์ใช้ Sentimental Analysis กับโรงแรมเท่านั้น เช่น P. Prameswari, Z. I. Surjandari, and E. Laoh. [2] ได้นำเสนอการวิเคราะห์ข้อมูลความคิดเห็นภายในประเทศอินโดนีเซีย โดยงานวิจัยนี้ใช้ขั้นตอนการ

กำหนดค่าน้ำหนักของคำด้วยวิธีการ TF-IDF เพื่อนำค่าที่ได้เข้าสู่กระบวกการแบ่งประเภทด้วย SVM, KNN และ Naïve Bayes โดยแบ่งประเภทของความคิดเห็นเป็น 10 กลุ่ม โดยสรุปผลการวิเคราะห์ด้วยวิธีการหาค่า F-Measure อยู่ที่ 78%, 62% และ 53%

I. P. Windasari and D. Eridani [4] นำเสนองานวิจัยชื่อ Sentiment Analysis on Travel Destination in Indonesia นำเสนอวิธีการวิเคราะห์ความรู้สึกของนักท่องเที่ยวในประเทศอินโดนีเซีย โดยวิธีการให้ค่าน้ำหนัก TF-IDF มากำหนดค่าให้กับข้อความที่มีการรีวิวของนักท่องเที่ยวแล้วใช้กระบวนการ SVM เข้ามาแบ่งกลุ่มข้อความว่าข้อความนั้นให้ความคิดเห็นกับสถานที่เป็น ความรู้สึกดี หรือความรู้สึกไม่ดี โดยความถูกต้องอยู่ที่ 85 %

V. B. Raut and D. D. Londhe [12] นำเสนองานวิจัยการวิเคราะห์ความคิดเห็นของผู้ใช้งานห้องพักโดยใช้กระบวนการวิเคราะห์ข้อมูลโดยแบ่งข้อมูลออกเป็น 2 ชุด คือ ข้อมูลชุดสอน และข้อมูลชุดเรียนรู้ โดยใช้ 4 อัลกอริทึมในการเรียนรู้คือ Naïve Bayes, SVM, Decision Tree และ SentiWordNet ในการวิเคราะห์โดยใช้ข้อมูลจำนวน 1000 ประโยค โดยแบ่งเป็นข้อมูลเชิงบวก 500 และข้อมูล เชิงลบ 500 ประโยคโดยวัดประสิทธิภาพของการวิเคราะห์ด้วยวิธีการหาค่า F-Measure โดยสรุปผลค่าความถูกต้องได้ 87%, 83%, 78% และ 88% ตามลำดับ

H. X. Shi and X. J. Li [13] นำเสนองานวิจัยการวิเคราะห์ความรู้สึกของผู้ใช้งานโรงแรมโดยใช้วิธีกระบวนการเรียนรู้แบบ Supervised learning หรือการเรียนรู้แบบมีผู้สอนโดยสร้างเงื่อนไขจากข้อมูลสอน (training data) ด้วยการนับจำนวนความถี่ของคำและวิธีการกำหนดความถี่ของคำด้วย TF-IDF ในขั้นตอนของการ Pre-processing เพื่อนำข้อมูลที่ผ่านการนับความถี่ เข้าสู่การทำนายผลด้วย SVM ซึ่งแบ่งกลุ่มเป็นสองกลุ่มข้อมูลคือ กลุ่มข้อความเชิงบวก และ กลุ่มข้อความเชิงลบ

Wenying ZHENG and Qiang YE [14] นำเสนองานวิจัยชื่อ Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm โดยบทความนี้นำเสนอการวิเคราะห์ความรู้สึกของของนักท่องเที่ยวของเว็บไซต์ www.ctrip.com จำนวน 40 โรงแรม 755 ข้อความ โดยใช้ค่าคะแนนของแต่ล่ะข้อความเข้าสู่กระบวนการแยกประเภทด้วยเทคนิค KNN, Naïve Bayes และ SVM โดย SVM ให้ค่าความถูกต้องสูงสุดที่ 91.15%

Puteri Prameswari, Isti Surjandari and Enrico Laoh[15] นำเสนอการทำเหมืองความคิดเห็นจากบทวิจารณ์ออนไลน์ในแหล่งท่องเที่ยวบาหลี โดยรวบรวมความคิดเห็นออนไลน์จาก Tripadvisor.com แล้วจึงนำเข้าสู่กระบวนการกำหนดค่าความเชื่อมัน จากนั้นจึงนำข้อมูลเข้าการแยกประเภท Recursive Neural Tensor Network (RNTN) โดยให้ค้าความถูกต้องอยู่ที่ 85%

Xiaobo Zhang and Qingsong Yu[16] นำเสนองานวิจัยชื่อ Hotel Reviews Sentiment Analysis Based on Word Vector Clustering โดยรวบรวมความคิดเห็นจากเว็บไซต์ http://www.dianping.com จำนวน 4000 ความคิดเห็นโดยแบ่งข้อมูลออกเป็นข้อมูลเทรน 3000 ข้อมูลทดสอบ 1000 โดยใช้กระบวนการ Word2Vec ในการกำหนดค่าของข้อความรวมกับการจักกลุ่มแบบ K-means ค่าเฉลี่ยความถูกต้องอยู่ที่ 96 %

Burçin Buket Ogul and Gönenç Ercan[17] นำเสนอการจำแนกประเภทของการวิเคราะห์ความคิดเห็นสำหรับโรงแรมในประเทศตรุกี โดยรวบรวมข้อความจากเว็บไซต์ booking.com และ tripadvisor.com จำนวน 331 ข้อความ ด้วยวิธีการกำหนดค่าน้ำหนักข้อมูลด้วย TF-IDF และการจำแนกประเภทด้วย Decision Tree และ Naïve Bayes แล้วจึงประเมินประสิทธิภาพด้วย AUC (Area Under Curve) โดยให้ค่าอยู่ที่ 90 %

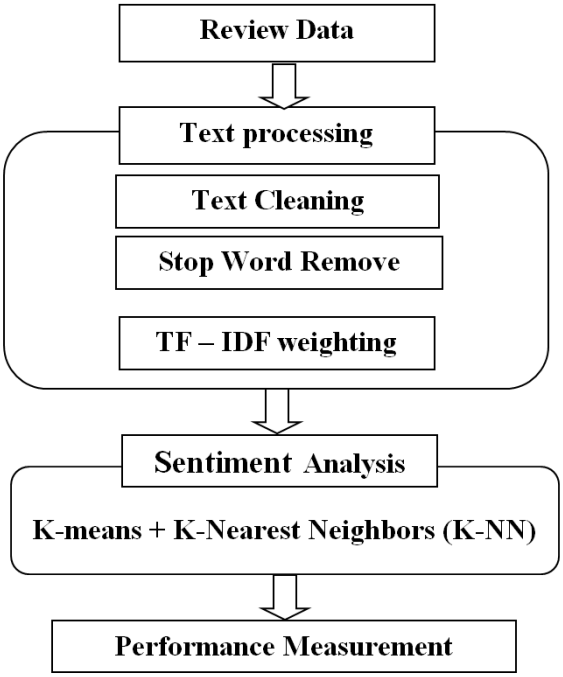
Sachio Hirokawa and Kiyota Hashimoto[18] นำเสนอวิธีการจำแนกประเภทการวิเคราะห์ความคิดเห็นด้วยวิธีการ SVM จำนวน73,589 ความคิดเห็น โดยแบ่งประเภทออกเป็น 2 ประเภทคือความคิดเห็นในเชิงบวก และความคิดเห็นในเชิงลบ แล้วจึงแบ่งกลุ่มขอแต่ล่ะประเภทออกเป็น 8 กลุ่มคือ มูลค่า, ห้องพัก, สถานที่, ความสะอาด, การเช็คอิน, การบริการ, ธุรกิจ และภาพรวม โดยแบ่งออกความคิดเห็นในเชิงบวก 74 % และความคิดเห็นในเชิงลบ 24% ของข้อความทั้งหมด

Qianjun Shuai,Yamei Huang, Libiao Jin and Long Pang[19] นำเสนอการวิเคราะห์ความคิดเห็นโรงแรมในประเทศจีน ด้วยการกำหนดค่าข้อมูล Doc2Vec เป็นการแปลงค่าข้อมูลในเอกสารออกมาในรูปแบบของค่าตัวเลข แล้วจัดประเภทข้อมูลด้วย SVM และ Naïve Bayes แล้ววัดประสิทธิภาพด้วยค่า Precision, Recall และ F-Measure โดย SVM ให้ค่า F-Measure สูงสุดที่ 81.1 %

จากงานวิจัยที่กล่าวมานั้นจะเห็นได้ว่าจำนวนของความคิดเห็นที่ใช้ในการเทรนแบบจำลองค่อนข้างน้อย งานวิจัยส่วนใหญ่จะใช้เพียงเทคนิคเดียวในการจัดกลุ่มข้อมูล หรือการจำแนกประเภทของข้อมูล โดยไม่มีการเปรียบเทียบกับเทคนิคอื่น ๆ ที่มีลักษณะที่คล้ายคลึงกัน และการนำเอาตัวแบบนั้น ๆ ไปสร้างเป็นแอปพลิเคชันในการวิเคราะห์

วิธีดำเนินการวิจัย

จากการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง ในบทนี้ผู้วิจัยได้นำเสนอขั้นตอนการดำเนินงานวิจัยซึ่งประกอบไปด้วยกระบวนการประมวลผลข้อความ (Text Processing), การวิเคราะห์ความคิดเห็น (Sentiment Analysis) และทดสอบประสิทธิภาพความถูกต้อง (Performance Measurement) ดังรูปที่ 1



รูปที่ 1 กรอบแนวคิดงานวิจัย

กระบวนการประมวลผลข้อความ (Text Processing)

Text cleaning ในขั้นตอนนี้จะเป็นการนำประโยคที่ได้เข้ามาตรวจสอบอักษรพิเศษหากประโยคนั้น ๆ มีอักขระพิเศษจะทำการลบอักขระพิเศษออกจากประโยค ดังตัวอย่างแสดงในตารางที่ 1

ตารางที่ 1 ข้อมูลตัวอย่างการลบอักขระพิเศษ

ประโยคเดิม	ประโยคหลังตัดอักขระพิเศษ
Comfortable clean room ...friendly staffs...	Comfortable clean room friendly staffs
Each room/floor is different	Each room floor different
Hotel that's near train station.	Hotel that is near train station

Stop Word Removal เนื่องจากข้อมูลที่ผู้วิจัยได้นำมาใช้งานนั้นเป็นภาษาอังกฤษจึงใช้คุณลักษณะการเว้นวรรคในการตัดคำออกจากประโยค โดยการนำเอาคำเชื่อมประโยคออก เช่น I, is, and, or และ for ดังตัวอย่างด้านล่าง

“Good for last minute option” → good, last minute, option

“Decor friendly staff pool” → Decor, friendly, staff, pool

TF – IDF weighting [3] เป็นขั้นตอนการกำหนดค่าความน้ำหนักความสำคัญของคำด้วยวิธีการ TF-IDF และการกำหนดค่า Category และ Class ให้กับข้อมูล ดังตัวอย่างแสดงในตารางที่ 2 และ 3

ตารางที่ 2 ตัวอย่างขั้นตอน TF

ประโยค	Decor	friendly	staff	pool	attention	…	evening
Decor and friendly staff, pool	1	1	1	1			
We loved the attention to little details: the evening dress code, tea time, etc.					1		1
Awesome staff			1				
Staff not helpful and not professional!			1				

ตารางที่ 3 ตัวอย่างขั้นตอน TF-IDF

ประโยค	Decor	friendly	staff	pool	attention	…	evening
Decor and friendly staff, pool	4.6052	3.5066	2.5257	3.912			
We loved the attention to little details: the evening dress code, tea time, etc.					4.6052		4.6052
Awesome staff			2.5257				
Staff not helpful and not professional!			2.5257				

จากตารางที่ 3 จะเห็นได้ว่าค่า TF-IDF ค่อนข้างที่จะใกล้เคียงกันถึงแม้ข้อความนั้นมีความหมายที่ไม่คล้ายคลึงกันเลย ผู้วิจัยจึงมีแนวคิดที่จะให้ข้อความเหล่านั้นถูกมีค่า TF-IDF แตกต่างกันออกไป โดยกำหนดให้ Category = “a” คือ การเข้าถึง (accessibility), “b” คือ กิจกรรมและความบันเทิง (activities & entertainment), “c” คือ อาหารและเครื่องดื่ม (food & beverage), “d” คือ พนักงานผู้ให้บริการ (staff), “e” คือ สถานที่ (place)

กำหนดให้ Class T = Positive และ F = Negative โดยตารางที่ 4 และ 5 เป็นตัวอย่างการกำหนด TF-IDF, Category และ Class ตามลำดับ แสดงได้ดังตารางที่ 4 และ 5

ตารางที่ 4 ตัวอย่างการกำหนด TF-IDF

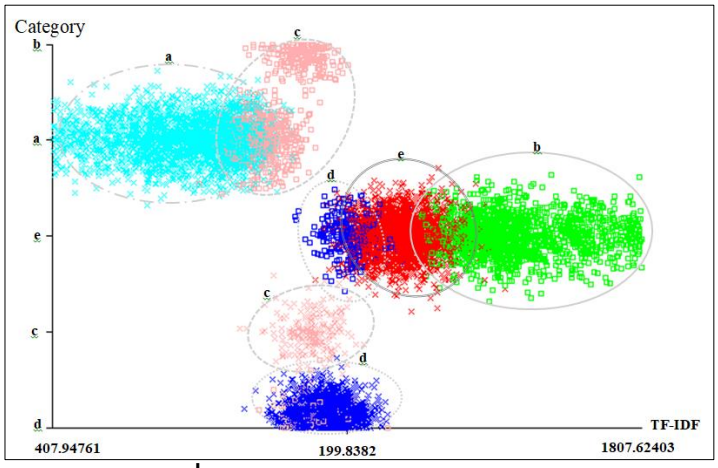
IF	THEN
Category = “a”	TF-IDF = TF-IDF * -100
Category = “b”	TF-IDF = TF-IDF * -10
Category = “c”	TF-IDF = TF-IDF * 1
Category = “d”	TF-IDF = TF-IDF * 10
Category = “e”	TF-IDF = TF-IDF * 100

ตารางที่ 5 ตัวอย่างการกำหนด Category และ Class

ประโยค	TF-IDF	Category	Class
Decor and friendly staff, pool	145.495	d	T
We loved the attention to little details: the evening dress code, tea time, etc.	-184.207	b	T
Awesome staff	64.378	d	T
Staff not helpful and not professional!	110.429	d	F

การวิเคราะห์ความคิดเห็น (Sentiment Analysis)

การแบ่งประเภทของข้อมูล K-means จำนวน 10,000 ประโยค โดยการแบ่งประเภทของข้อมูลผู้วิจัยเลือกใช้ Simple K Means ด้วยโปรแกรม WEKA[20] โดยนำเข้าข้อมูล TF-IDF และ Category กำหนดให้ Category เป็น Class สำหรับคำตอบ โดยค่าความถูกต้องของการแบ่งประเภทของข้อมูลอยู่ที่ 77 % ดังตัวอย่างในรูปที่ 2



รูปที่ 2 ผลการจัดประเภทด้วย K-means

จำแนกประเภทด้วย K-Nearest Neighbors (K-NN) จากขั้นตอนการแบ่งประเภทของข้อมูลด้วยเทคนิค K-means จะได้ผลลัพธ์ออกเป็น 5 กลุ่ม แล้วผู้วิจัยจึงนำเอาความคิดเห็นแต่ล่ะกลุ่มเข้าสู่กระบวนการหาค่า TF-IDF ก่อนนำความคิดเห็นนั้นเข้าสู่กระบวนจำแนกประเภทด้วย K-NN โดยใช้การวัดประสิทธิภาพแบบ 10 Fold Cross Validation สามารถแสดงผลการทดลองดังตารางที่ 7

ตารางที่ 7 ผลการเปรียบเทียบประสิทธิภาพความถูกต้องของแต่ล่ะกลุ่ม

	การเข้าถึง	กิจกรรมและความบันเทิง	อาหารและเครื่องดื่ม	พนักงานให้บริการ	สถานที่	ค่าเฉลี่ย
Accuracy	89.1 %	98.4 %	92.8 %	94.7 %	95.8 %	94.2 %
Recall	89.2 %	98.4 %	92.9 %	94.8 %	95.8 %	94.2 %
Precision	89.1 %	98.4 %	92.7 %	94.9 %	95.8 %	94.2 %
F-Measure	89.2 %	99.2 %	92.8 %	94.8 %	97.9 %	94.8 %

การทดสอบประสิทธิภาพความถูกต้อง

ในงานวิจัยนี้มีการวัดประสิทธิภาพการทำงานโดยการหา ค่าความเที่ยง (Precision), ค่าความระลึก (Recall), ค่าความถูกต้อง (Accuracy) และค่า F-Measure ดังสมการที่ 7 ถึง 10

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

$$Recall = \frac{TP}{TP + FN} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{9}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

TP คือ จำนวนข้อมูลที่ถูกนำมาใช้อย่างถูกต้อง

FP คือ จำนวนข้อมูลที่ผิดผลาดที่ถูกนำมาใช้อย่างถูกต้อง

TN คือ จำนวนข้อมูลที่ถูกต้องที่ไม่ถูกนำมาใช้

FN คือ จำนวนข้อมูลที่ผิดผลาดที่ไม่ถูกนำมาใช้

ผลการดำเนินงาน

งานวิจัยนี้ทำการเปรียบเทียบประสิทธิภาพความถูกต้องในการวิเคราะห์ความรู้สึกโดยแบ่งเป็นเทคนิคดั้งเดิม คือ 4 เทคนิคคือ Decision-Tree, K-Nearest Neighbors, Support Vector Machine และการปรับปรุงโดยเพิ่มการจัดกลุ่มโดยใช้ K-Means ร่วมกับเทคนิค K-Nearest Neighbors ซึ่งสามารถสรุปได้ดังต่อไปนี้ เทคนิค K-Means ร่วมกับ K-Nearest Neighbors ให้ความถูกต้องสูงสุดที่ 94.8 % รองลงมาคือ Decision-Tree 89.0%, K-Nearest Neighbors (IBk) 87.7 % และเทคนิค Support Vector Machine 86.2% แสดงได้ดังตารางที่ 8

ตารางที่ 8 ผลการเปรียบเทียบประสิทธิภาพความถูกต้อง

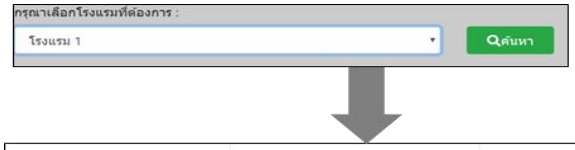
	Decision-Tree (J48)	**Support Vector Machine (SMO)**	**K-Nearest Neighbors (IBk)**	**K-Means + K-Nearest Neighbors**
Accuracy	89.4 %	86.2 %	88.2 %	94.2 %
Recall	89.4 %	86.2 %	87.5 %	95.0 %
Precision	88.7 %	86.2 %	87.7 %	94.7 %
F-Measure	89.0 %	86.2 %	87.7 %	94.8 %

แอปพลิเคชัน

การทำงานของการวิเคราะห์ความคิดเห็นโดยใช้ PHP 7.3, JAVA และ Weka[20] 3.8.3 โดยเริ่มจากการฟังชั่น CallAPT() ที่จะเรียกใช้ WebAPI ไปยังเว็บไซต์ API TUBEใน แล้วทำการนำข้อความที่ได้วนลูปเข้าสู่การกระบวนการประมวลผลข้อความ (Text Processing) ในฟังชั่น calTFIDF() แล้วจึงนำเข้าสู่การการวิเคราะห์ความคิดเห็น (Sentiment Analysis) โดยแบ่งกลุ่มข้อความด้วย k-means ในฟังชั่น callKmeansModel() จะทำการเรียกโมเดลที่ได้จาก Weka ผ่านคำสั่ง java –cd แล้วเก็บผลลัพธ์ไว้ที่ $groupList แล้วจำทำการวนลูปแต่ล่ะกลุ่มเพื่อจำแนกประเภทของข้อมูลโดยการทำงานในแต่ล่ะลูปจะทำการหาค่า TFIDF ในฟังชั่น calTFIDF() ของประโยคในแต่ล่ะกลุ่มใหม่ แล้วจึงเรียกโมเดลการจำแนกประเภทด้วย K-NN ด้วยฟังชั่น callKNNModel() แล้วเก็บผลลัพธลงตัวแปร $result เมื่อการทำงานครบทุกลูปจะนำผลลัพธ์ที่ได้ไปแสดงในฟังชั่น callDisPlayResult() ดังรูปที่ 3

```
// call raw data in website APT TUBE
$reviewList = CallAPI($propertyId);
//K-means clustering
foreach ($reviewList as $row) {
        $raw['review'] = $row['comment'];
        $raw['sunTFIDF'] = calTFIDF($reviewList,$row);
        // weka.clusterers.SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 5 -A
"weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10
        $raw['category'] = callKmeansModel($raw['sunTFIDF'],$groupList);
}
// K-nn classifiers
foreach ($groupList as $group) {
        foreach ($group['reviewList'] as $row) {
                $raw['sunTFIDF'] = calTFIDF($group['reviewList'],$row);
        // weka.classifiers.lazy.IBk -K 2 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
                $result[] = callKNNModel($raw['sunTFIDF'],$groupList);
        }
}
callDisPlayResult($result);
```

รูปที่ 3 ตัวอย่างโค้ดในการทำงาน



การเข้าถึง (Accessibility)	กิจกรรมและความบันเทิง (Activities & Entertainment)	อาหารและเครื่องดื่ม (Food & beverage)	พนักงานผู้ให้บริการ (Human resources)	สถานที่ (Physical)
Exceptional	I liked it well enough.	We loved the attention to little details: the evening dress code, tea time, etc.	Decor and friendly staff, pool	Comfortable clean room ...friendly staffs ....
Loved the location, hopefully you are a smoker	Nice up to date room..	Happy shopping like cosmetics, good Korean food	Awesome staff	lousy stay pool service
Good centrally located hotel			Staff not helpful and not professional!	Nice common area
comfortable			Value for money!	Good hotel good price comfy bed
Service not good at all			Great location, attentive staff	Convenient Hotel that's near train station.
hotel with excellent service			clean efficient hotel with courteous staff	Comfortable stay
Brilliant Hotel			Great value, big rooms	nice hotel

รูปที่ 4 แอปพลิเคชัน

จากรูปที่ 4 แอปพลิเคชันที่นำเอาโมเดลไปประยุกต์ใช้งานโดยขั้นตอนการทำงานเริ่มจากดึงรายชื่อโรงแรมในประเทศไทยจากเว็บไซต์ APITUBE โดยงานวิจัยได้เปลี่ยนชื่อของโรงแรมเพื่อไม่ให้มีผมเสียต่อโรงแรมนั้นๆ เมื่อทำการเลือกโรงแรมที่ต้องการแล้วกดค้นหาระบบจะส่งรหัสของโรงแรมนั้นไปยังการวิเคราะห์ข้อมูล แล้วจึงแสดงผลโดยจะแยกข้อความเป็นหวดหมู่และระบุความรู้สึกผ่านสีพื้นหลังโดย สีเขียว คือ ชื่นชอบ และ สีแดง คือ ไม่ชอบ

สรุปและข้อเสนอแนะ

งานวิจัยนี้เป็นการเปรียบเทียบประสิทธิภาพความถูกต้องในการวิเคราะห์ความรู้สึกด้วยเทคนิค 4 เทคนิคคือ Decision Tree (J48), K-Nearest Neighbors (IBk), Support Vector Machine (SVM) และ การแบ่งกลุ่มด้วย K-means ร่วมกับ K-Nearest Neighbors (K-NN) โดยวัดประสิทธิภาพแบบ 10 Fold Cross Validation จากข้อมูลจำนวน 10,000 ประโยค โดยปรากฏว่าเทคนิค K-means ร่วมกับ K-Nearest Neighbors (K-NN) ให้ความถูกต้องสูงสุดที่ 94.8% รองลงมาคือเทคนิค Decision tree 89.7%, K-Nearest Neighbors (IBk) 88.2 % และเทคนิค Support Vector Machine 87.5% ซึ่งรวบรวมข้อความจากเว็บไซต์ APITUDE โดยข้อความแสดงความคิดเห็นเป็นภาษาอังกฤษเท่านั้น โดยงานวิจัยนี้ได้มีการแบ่งกลุ่มออกเป็น 5 กลุ่ม คือ การเข้าถึง (accessibility), กิจกรรมและความบันเทิง (activities & entertainment), อาหารและเครื่องดื่ม (food & beverage), พนักงานผู้ให้บริการ (staff) และ สถานที่ (place) เท่านั้นซึ่งข้อความยังถูกพูดถึงในหลายด้านออกไป และรูปแบบวิธีการในการกำหนดค่าของคำแบบอื่น ๆ

บรรณานุกรม

Y.-H. Hu and K. Chen. (2016). "Predicting Hotel Review Helpfulness: TheImpact of Review Visibility, and Interaction Between Hotel Stars and Review Ratings". International Journal of Information Management. vol.36. pp. 929–944.

P. Prameswari, Z. I. Surjandari, and E. Laoh. (2017). "Mining Online Reviews in Indonesia's Priority Tourist Destinations Using Sentiment Analysis and Text Summarization Approach". IEEE 8th International Conference on Awareness Science and Technology.pp.121-126.

รวงทอง ถูระวรณ์. (2554). "ปัจจัยที่มีผลต่อการตัดสินใจใช้บริการโรงแรมรีสอร์ทในเขตกรุงเทพมหานคร". วิทยานิพนธ์ปริญญามหาบัณฑิต มหาวิทยาลัยราชภัฏวไลยอลงกรณ์ในพระบรมราชูปถัมภ์.

B. Pang and L. Lee. (2008). "Opinion Mining and Sentiment Analysis". Foundations and Trends in Information Retrieval Vol.2.pp.1-90.

C. Shearer. (2000). "The CRISP-DM model: the new blueprint for data mining". Journal Of Data Warehousing. pp.13-24.

I. P. Windasari and D. Eridani. (2017). "Sentiment Analysis on Travel Destination in Indonesia". Int. Conf. on Information Tech, Computer and Electrical Engineering (ICITACEE). pp. 276-279.

กนกวรรณ เขียนวรรณ."การประมวลผลภาษาธรรมชาติ".ค้นวันที่ 10 พฤษจิกายน 2561 จาก www.mbs.mut.ac.th/paper/pdf/29.pdf

Joachims. (1998). "Text Categorization with Support Vector Machines: Learning with Many Features". Proceeding 10 European Conference on Machine Learning (ECML), Chemnitz.

ชัดชัย แก้วตา อัจฉรา และ อัจฉรา มหาวีรวัฒน์. (2010). "การวินิจฉัยคดีด้วยเทคนิคต้นไม้ตัดสินใจ". The 6th National Conference on Computing and Information Tec hnology (NCCIT2010),pp.308-313,

Derek Abbott and Brain Ng. (2011). "Final Report 2011: Who wrote the Letter to the Hebrews". ค้นเมื่อ 01 มกราคม 2562, จาก "https://www.eleceng.adelaide.au /personal/dabbott".

Vijay B Raut and D.D. Londhe. (2014). "Opinion Mining and Summarization of Hotel Reviews". 6th International Conference on Computational Intelligence and Communication Networks. pp.556-559.

H. X. Shi and X. J. Li. (2011). "A Sentiment analysis model for hotel reviews based on supervised learning". International Conference on Machine Learning and Cybernetics .pp.950-954.

Wenying ZHENG and Qiang YE. (2009). "Sentiment Classification of Chinese Traveler Reviews by Support Vector Machine Algorithm". Third International Symposium on Intelligent Information Technology Application. pp335-338

Puteri Prameswari, Isti Surjandari and Enrico Laoh. (2017). "Opinion mining from online reviews in Bali tourist area". 3rd International Conference on Science in Information Technology (ICSITech).pp.226-230

Xiaobo Zhang and Qingsong Yu. (2017). "Hotel Reviews Sentiment Analysis Based on Word Vector Clustering". 2nd IEEE International Conference on Computational Intelligence and Applications. pp260-264.

Burçin Buket Ogul and Gönenç Ercan. (2016). "Türkçe Otel Yorumlarından Duygu AnaliziSentiment Classification on Turkish Hotel Reviews". 24th Signal Processing and Communication Application Conference (SIU).

Sachio Hirokawa, Kiyota Hashimoto. (2018). "Simplicity of Positive Reviews and Diversity of Negative Reviews in Hotel Reputation". 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing.

Qianjun Shuai,Yamei Huang, Libiao Jin and Long Pang. (2018). "Sentiment Analysis on Chinese Hotel Reviews with Doc2Vec and Classifiers". 3rd Advanced Information Technology, Electronic and Automation Control Conference.

Ian H. Witten, Eibe Frank and Mark A. Hall. (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition Morgan Kaufmann, San Francisco.