

กระบวนการคัดเลือกคุณลักษณะสำหรับเพิ่มประสิทธิภาพในการจำแนก
ความคิดเห็นของลูกค้า

Feature Selection Method for Improving Customer Reviews Classification

ธีรยุทธ คุณสุข^{1*} และจारी ทองคำ¹

Teerayut Khunsuk^{1*} and Jaree Thongkam¹

Received: May 8, 2019; Revised: October 7, 2019; Accepted: October 7, 2019

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อค้นหาเทคนิคคัดเลือกคุณลักษณะเพื่อเพิ่มประสิทธิภาพของการจำแนกความคิดเห็นของลูกค้าเกี่ยวกับร้านอาหาร ข้อมูลถูกเก็บรวบรวมจากเว็บไซต์ wongnai.com จำนวนทั้งหมด 4,487 ข้อความ คณะผู้วิจัยได้นำเอาเทคนิค 3 เทคนิคในการคัดเลือกคุณลักษณะของข้อความ คือ Chi-Square, Information Gain และ Information Gain Ratio ในการวัดประสิทธิภาพของเทคนิคคัดเลือกคุณลักษณะ เทคนิค Naive Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor และ C4.5 มาใช้ในการจำแนก ยิ่งไปกว่านั้น 10-fold Cross Validation ได้ถูกนำมาใช้ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้ และวัดค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึก จากการทดลองพบว่า เทคนิคการคัดเลือกคุณลักษณะแบบ Information Gain ร่วมกับเทคนิค Naive Bayes ให้ผลที่ดีที่สุดในการจำแนกความคิดเห็น โดยให้ค่าความถูกต้อง 89.08 % ค่าความแม่นยำ 89.12 % และค่าความระลึก 89.10 %

คำสำคัญ : การคัดเลือกคุณลักษณะ; การจำแนกความคิดเห็น; เหมืองความคิดเห็น

¹ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

¹ Faculty of Informatics, Mahasarakham University

* Corresponding Author E - mail Address: teerayut041995@gmail.com

Abstract

This research proposes improvement of search for feature selection techniques to increase the efficiency of customer feedback classification about restaurants, Information is collected from the website wongnai.com, a total of 4,487 messages. The research team adopted 3 techniques for selecting text features: Chi-Square, Information Gain and Information Gain Ratio to measure the effectiveness of feature selection techniques and applied Naive Bayes, Support Vector Machine, K-Nearest Neighbor and C4.5 for classification. Moreover, the 10-fold Cross Validation has been used to divide the data into a learning set and measure accuracy (Accuracy), precision (Precision) and recall (Recall). From the experiment found that Information Gain feature selection technique cooperate with the Naive Bayes technique provides the best results in the classification of comments by the accuracy is 89.08 %, the precision is 89.12 % and the recall is 89.10 %.

Keywords: Feature Selection; Text Classification; Opinion Mining

บทนำ

ปัจจุบันมีข้อความแสดงความคิดเห็นของผู้คนอย่างมากมายบนอินเทอร์เน็ตซึ่งเป็นที่นิยมสำหรับผู้ที่ชื่นชอบใช้งาน ทั้งจากเว็บไซต์ บล็อก และเครือข่ายสังคมออนไลน์ที่อนุญาตให้คนสามารถเข้ามาแสดงความคิดเห็น และยังมีการให้คะแนนของสินค้าและการบริการด้วย ความคิดเห็นเหล่านี้ช่วยเปิดพื้นที่ใหม่ให้กับธุรกิจต่าง ๆ และตลาดออนไลน์ ความคิดเห็นจากเว็บไซต์ที่ถูกค่าได้มีการพูดถึงสินค้าหรือการบริการที่ได้รับในแง่บวกต่าง ๆ ย่อมมีความเป็นจริงของความจริงหรือไม่พอใจในสินค้าหรือการบริการ [1] ดังนั้นการวิเคราะห์ข้อความเหล่านี้จะช่วยให้เห็นทัศนคติ และความคิดเห็นที่แท้จริงของลูกค้า มีผลทำให้ลูกค้าเลือกใช้หรือไม่ใช้สินค้าและการบริการได้ แต่ข้อมูลข้อความแสดงความคิดเห็นบนอินเทอร์เน็ตมีปริมาณมากขึ้น และมีความหลากหลายทำให้ยากต่อการค้นหาและการจัดหมวดหมู่เอกสาร จึงได้มีการศึกษาและพัฒนากระบวนการในการจำแนกความคิดเห็น [2] ที่มีจำนวนมากให้เป็นไปอย่างมีประสิทธิภาพและรวดเร็ว

เหมืองความคิดเห็น (Opinion Mining) [3] เป็นกระบวนการวิเคราะห์ความคิดเห็นที่ผู้บริโภคนำเสนอในรูปแบบของข้อความ ซึ่งพบได้บนเครือข่ายสังคมออนไลน์ เพื่อให้ทราบถึงความพึงพอใจในเชิงบวกหรือเชิงลบ การจัดการข้อความที่ขับเคลื่อนด้วยความคิดเห็นนั้นสามารถนำไปประยุกต์ใช้กับงานที่สำคัญหลายประเภท ตัวอย่างเช่น การจำแนกความคิดเห็นของผู้วิจารณ์เกี่ยวกับสินค้าโดยแบ่งประเภทบทวิจารณ์สินค้าออนไลน์ หรือการติดตามการเปลี่ยนแปลงทัศนคติของประชาชนเกี่ยวกับเรื่องการเมือง แต่ปัญหาหนึ่งในการทำเหมืองความคิดเห็นจะเกี่ยวข้องกับการวิเคราะห์ข้อมูลที่มีจำนวนมหาศาลและมีความซับซ้อน โดยส่วนใหญ่กระบวนการทำเหมืองความคิดเห็นนั้นจะมีข้อมูลที่ประกอบไปด้วย คุณลักษณะที่ไม่ตรงประเด็น และมีมิติของข้อมูลจำนวนมาก ซึ่งส่งผลให้การทำเหมืองความคิดเห็นต้องใช้เวลาในการวิเคราะห์มากขึ้น ถ้าข้อมูลมีมิติ หรือตัวแปรมากจะทำให้ข้อมูลเกิดการกระจาย และอาจไม่มีความสัมพันธ์กับมิติอื่น [4]

ดังนั้นการคัดเลือกคุณลักษณะของข้อมูลโดยการทำให้ข้อมูลเดิมมีขนาดลดลงและสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด จึงเป็นแนวทางหนึ่งที่สามารถช่วยแก้ปัญหาดังกล่าว และยังสามารถช่วยให้การจำแนกข้อมูลได้แม่นยำมากขึ้น

ซึ่งมีนักวิจัยหลายท่านได้นำเอาเทคนิคและกระบวนการคัดเลือกคุณลักษณะมาใช้ในการปรับปรุงประสิทธิภาพการจำแนกความคิดเห็น เช่น [5] ได้ใช้วิธีการคัดเลือกคุณลักษณะเป็นวิธีที่สำคัญสำหรับการเพิ่มประสิทธิภาพและการหาค่าความถูกต้องของการจัดกลุ่ม โดยเทคนิคในการคัดเลือกคุณลักษณะที่ใช้ได้แก่ Information Gain, Information Gain Ratio และ Chi-Square และหลังจากนั้นจึงวัดประสิทธิภาพของการจำแนกข้อความด้วยเทคนิค K-Nearest Neighbor, Naive Bayes และ Support Vector Machine ซึ่งพบว่า Information Gain ใช้ในการคัดเลือกคุณลักษณะได้ดีที่สุดในการจำแนกข้อความประมวลผลได้เร็วและใช้ทรัพยากรน้อย ส่วนงานวิจัยของ [6] ได้ทำการเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะของการทำเหมืองข้อความเพื่อจำแนกข้อความ โดยใช้เทคนิคการคัดเลือกคุณลักษณะ คือ Information Gain, Gain Ratio และ Chi-Square ผลการเปรียบเทียบพบว่าวิธีการคัดเลือกคุณลักษณะด้วยเทคนิค Chi-Square โดยใช้การจำแนกประเภทแบบ Support Vector Machine ให้ผลการวัดประสิทธิภาพดีที่สุด เช่นเดียวกันกับงานของ [7] ที่ได้ทำการเปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่เหมาะสมด้วยเทคนิค Information Gain, Gain Ratio และ Chi-Square โดยได้ใช้เทคนิค Support Vector Machine ในการจำแนกประเภทข้อความ ผลการทดลองพบว่าเทคนิค Chi-Square ให้ผลลัพธ์ที่มีประสิทธิภาพดีที่สุด

ในส่วนของงานวิจัยในการจำแนกข้อความก็ได้มีนักวิจัยหลายท่านได้นำเอาเทคนิคการทำเหมืองความคิดเห็นมาใช้ในการจำแนกข้อความ เช่น [8] ได้นำเสนอวิธีการจำแนกความคิดเห็นโดยการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ด้านการบริการห้องพักโรงแรม โดยใช้เทคนิคเหมืองข้อความวิเคราะห์ความคิดเห็นภาษาไทย และสร้างแบบจำลองด้วยอัลกอริทึม 4 วิธี ได้แก่ เทคนิค Support Vector Machine เทคนิค Decision Tree เทคนิค Naive Bayes และเทคนิค K-Nearest Neighbor เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์ความคิดเห็นภาษาไทย ผลการทดลองพบว่าเทคนิค Support Vector Machine ให้ประสิทธิภาพในการจำแนกดีที่สุด มีความถูกต้อง 83.38 % แต่ [9] ได้นำเสนอการจำแนกความคิดเห็นของคนไทยเกี่ยวกับสินค้าออนไลน์โดยใช้การทำเหมืองข้อความ โดยมีวัตถุประสงค์เพื่อค้นหาเทคนิคการจำแนกจาก 5 เทคนิคที่มีประสิทธิภาพ คือ เทคนิค Naive Bayes เทคนิค Support Vector Machine เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ (Decision Tree) และเทคนิค C4.5 ผลการทดลองพบว่าเทคนิค Naive Bayes ให้ประสิทธิภาพในการจำแนกที่ดีที่สุด และงานวิจัยของ [10] ได้นำเสนอการจัดกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ ซึ่งได้ทำการศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกข้อความ 3 เทคนิควิธี คือ เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ และเทคนิค Naive Bayes ผลการเปรียบเทียบประสิทธิภาพแสดงให้เห็นว่า เทคนิคการทำเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด โดยค่าความถูกต้องเท่ากับ 0.89 แสดงให้เห็นว่าประสิทธิภาพการจำแนกข้อความขึ้นอยู่กับเทคนิควิธีที่นำมาใช้จำแนก

จากงานวิจัยที่เกี่ยวข้องดังกล่าว การนำเอากระบวนการคัดเลือกคุณลักษณะสามารถที่จะช่วยเพิ่มประสิทธิภาพในการจำแนกข้อความ ดังนั้นในงานวิจัยนี้จึงได้นำเอากระบวนการคัดเลือกคุณลักษณะเพื่อเพิ่มประสิทธิภาพในการจำแนกความคิดเห็นของลูกค้า ซึ่งเทคนิคของการคัดเลือกคุณลักษณะที่นำมาใช้

ได้แก่ Chi-Square, Information Gain และ Gain Ratio และทำการเปรียบเทียบเทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความคิดเห็นของลูกค้า คือ เทคนิค Naive Bayes เทคนิค Support Vector Machine เทคนิคต้นไม้ตัดสินใจ และ K-Nearest Neighbor จากนั้นใช้หลักการ 10-fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) โดยแบบจำลองนี้สามารถนำไปจำแนกความคิดเห็นบนเว็บไซต์ของลูกค้าที่มีต่อร้านอาหารได้อย่างรวดเร็วและแม่นยำขึ้น

วิธีดำเนินการวิจัย

ในงานวิจัยนี้ผู้วิจัยได้ทำการศึกษาค้นคว้าทฤษฎีและทบทวนงานวิจัยที่เกี่ยวข้อง เพื่อนำมาใช้ในการวิเคราะห์ออกแบบและพัฒนาระบบการวิเคราะห์ความคิดเห็น ซึ่งสามารถแบ่งวิธีการดำเนินงานวิจัยได้ 5 ขั้นตอน ประกอบด้วย การเก็บรวบรวมข้อมูล ขั้นตอนก่อนการประมวลผล ขั้นตอนการคัดเลือกคุณลักษณะ การสร้างแบบจำลองเพื่อจำแนกความคิดเห็น และการประเมินประสิทธิภาพของแบบจำลอง แสดงขั้นตอนการดำเนินงานได้ดังรูปที่ 1



รูปที่ 1 ขั้นตอนการคัดเลือกคุณลักษณะสำหรับเพิ่มประสิทธิภาพในการจำแนกความคิดเห็น

1. การเก็บรวบรวมข้อมูล

งานวิจัยนี้ได้เก็บรวบรวมข้อมูลการแสดงความคิดเห็นของลูกค้าที่มีต่อร้านอาหาร จากเว็บไซต์ wongnai.com ซึ่งมีผู้เก็บรวบรวมข้อมูลไว้แล้วโดยสามารถดาวน์โหลดชุดข้อมูลจาก <https://github.com/wongnai/wongnai-corpus> เป็นความคิดเห็นที่มีคะแนนจากผู้แสดงความคิดเห็นในระดับที่ 1 - 5 ในงานวิจัยนี้ได้ใช้เฉพาะความคิดเห็นระดับ 1 2 4 และ 5 ซึ่งเป็นข้อความที่แสดงความคิดเห็นที่ทั้งด้านบวก ด้านลบ โดยเก็บข้อมูลตั้งแต่เดือนพฤษภาคม - กันยายน 2561 จำนวน 4,487 ความคิดเห็น จากนั้นกำหนดให้ข้อมูลความคิดเห็นที่มีคะแนนจากผู้แสดงความคิดเห็นในระดับที่ 1 กับ 2 เป็นความคิดเห็นเชิงลบ มีทั้งหมด 2,239 ความคิดเห็น และกำหนดให้ข้อมูลความคิดเห็นที่มีคะแนนจากผู้แสดงความคิดเห็นในระดับที่ 4 กับ 5 เป็นความคิดเห็นเชิงบวก มีทั้งหมด 2,248 ความคิดเห็น แสดงได้ดังตัวอย่างในรูปที่ 2

	review	star	class
0	กลับมาอีกแล้วกับไออิชิบุฟเฟ้ วันนี้อะมาที่สาขา...	5	positive
1	เป็นร้านหนึ่งที่ตั้งใจว่าจะไปกินหลายต่อหลายครั...	4	positive
2	มากินช่วงเทศกาล มาวันแรกตอนสามโมงของหมด วันที...	2	negative
3	ร้านนี้ตั้งอยู่ที่ ถนนนิมมานเหมินท์ ซอย 5 ร้าน...	4	positive
4	เคยได้ยินว่าร้านบุฟเฟ้ดีปิ้งย่าง แถวสามย่าน ...	2	negative
5	วันนี้ นื่องไปทำธุระแถวกระทรวงมหาดไทยเลยซื้อข้า...	1	negative
6	ไปสยามหิวๆเลยแวะไปทานร้านโปรดร้านนี้ครับ ปกติท...	1	negative
7	ร้านนี้มีชื่อร้านว่า Riverside park tak ตั้งอย...	5	positive
8	ร้านเค็กสัฟ ที่โดนใจที่สุดคือการสังฆารอนและ...	5	positive
9	เข้าแอฟงในเห็นว่าเค้ามีโปรโมชัน พิซซ่า 1 แด...	2	negative

รูปที่ 2 ความคิดเห็นของลูกค้าที่มีต่อร้านอาหาร

2. ขั้นตอนก่อนการประมวลผล

ขั้นตอนก่อนการประมวลผล (Preprocessing) เป็นการเตรียมข้อมูล ซึ่งในงานวิจัยนี้ได้มีขั้นตอนการเตรียมข้อมูลก่อนการประมวลผล ดังนี้

2.1 การตัดคำ (Word Segmentation) เป็นขั้นตอนการแบ่งตัวอักษรจากข้อความ (String) เพื่อหาขอบเขตของแต่ละหน่วยคำ (Morpheme) เพื่อให้การจำแนกข้อความมีประสิทธิภาพ ในงานวิจัยนี้ได้ใช้ภาษา Python โดยใช้ในการตัดคำภาษาไทย ซึ่งจะใช้วิธีการตัดคำแบบ Maximal Matching ซึ่งเป็นเทคนิควิธีการตัดคำแบบสอดคล้องมากที่สุด แล้วเลือกแบบที่ได้จำนวนน้อยที่สุด ตัวอย่างเช่น เมื่อมีข้อความว่า “ไปหามเหสี” ก็จะตัดคำได้ 2 เส้นทางคือ เส้นทางที่ 1 ไป หาม เท สี มีจำนวนคำเท่ากับ 4 คำ เส้นทางที่ 2 ไป หา มเหสี มีจำนวนคำเท่ากับ 3 คำ ซึ่งจะเลือกเส้นทางที่ 2 เนื่องจากมีจำนวนคำน้อยที่สุด การตัดข้อความความคิดเห็นภาษาไทยด้วยโปรแกรมภาษา Python โดยใช้วิธีการตัดคำแบบ Maximal Matching ได้ผลลัพธ์ดังตัวอย่างในรูปที่ 3

```
In [3]: from pythainlp.tokenize import word_tokenize

text = "กลับมาอีกแล้ววันนี้เรามาที่สาขาเซ็นทรัลลองสั่งเมนูประจำร้านปรากฏว่าอาหารไม่อร่อยเลย"
proc = word_tokenize(text, engine='newmm')
print(proc)

['กลับมา', 'อีกแล้ว', 'วันนี้', 'เรา', 'มา', 'ที่', 'สาขา', 'เซ็นทรัล', 'ลอง', 'สั่ง', 'เมนู', 'ประจำ', 'ร้าน', 'ปรากฏ', 'ว่า', 'อาหาร', 'ไม่', 'อร่อย', 'เลย']
```

รูปที่ 3 การตัดข้อความความคิดเห็นภาษาไทยด้วยโปรแกรมภาษา Python

2.2 การเลือกประเภทคำเพื่อใช้เป็นตัวแทนเอกสาร ในงานวิจัยนี้จะทำการเลือกตัวแทนเอกสาร โดยใช้เฉพาะคำนามกับคำกริยาและคำวิเศษณ์ ซึ่งเป็นประเภทคำที่เหมาะสมสำหรับการจำแนกความคิดเห็น หลังจากผ่านกระบวนการตัดคำ ผู้วิจัยได้ทำการกำกับประเภทของคำโดยใช้แนวคิดของ ORCHID โดยใช้คลังข้อมูล Orchid corpus ซึ่งเป็นคลังข้อความภาษาไทยที่มีการกำกับขอบเขตของประโยค ขอบเขตของคำ และชนิดของคำ หลังจากทำการกำกับชนิดของคำให้แต่ละคำแล้ว จึงได้ทำการคัดเลือกเอาเฉพาะ

ประเภทของคำที่ต้องการที่จะนำมาเป็นตัวแทนเอกสาร หลังจากผ่านกระบวนการตัดคำและการกำกับชนิดของคำ สามารถที่จะสกัดเอาคำจากความคิดเห็นทั้งหมด และทำการตัดคำที่ซ้ำกันออก จะได้คำนามกับคำกริยา และคำวิเศษณ์ 15,462 คำ ในงานวิจัยนี้ได้ใช้ขั้นตอนการกำกับชนิดของคำ โดยใช้คลังข้อมูล Orchid Corpus ซึ่งเป็นไฟล์ที่เก็บข้อความภาษาไทยที่มีการกำกับชนิดของคำ หลังจากนั้นใช้ภาษา Python ในการดึงข้อความทั้งหมดเพื่อทำเป็นพจนานุกรมเพื่อนำไปใช้งาน ผลลัพธ์จากการกำกับชนิดของคำโดยใช้ภาษา Python แสดงได้ดังรูปที่ 4

```
In [1]: import re
import os
import codecs,pickle
a={}
path = 'orchid97.crp.utf'
lines=codecs.open(path, 'r',encoding='utf8')
for line in lines:
    if len(re.findall('/', line)) == 1:
        lines1=line.split('/')
        a[lines1[0]]=re.sub("\n", "",lines1[1]) # เอา \n เมื่ออ่านไฟล์ออกมา
    else:
        lines1=0
with open('thaipos.pickle', 'wb+') as handle:
    pickle.dump(a, handle)
```

```
In [6]: import nltk.tag, nltk.data
import pickle
with open('thaipos.pickle', 'rb') as handle:
    model = pickle.load(handle)
text= ['กลับมา', 'อีกแล้ว', 'วันนี้', 'อาหาร', 'ไม่', 'อร่อย', 'เลย']
tagger = nltk.tag.UnigramTagger(model=model)
pos = tagger.tag(text)
print(pos)
```

[(‘กลับมา’, ‘VACT’), (‘อีกแล้ว’, ‘ADVN’), (‘วันนี้’, ‘NCMN’), (‘อาหาร’, ‘NCMN’), (‘ไม่’, ‘NEG’), (‘อร่อย’, None), (‘เลย’, ‘XVAE’)]

รูปที่ 4 การกำกับชนิดของคำโดยใช้ภาษา Python

3. ขั้นตอนการคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะของเอกสาร [4], [6] คือการนำเอกสารจากระบบต่าง ๆ เช่น เอกสารข่าวจากเว็บไซต์ บทคัดย่อ ความคิดเห็น มาแปลงเพื่อให้เอกสารอยู่ในรูปแบบเดียวกัน มีองค์ประกอบของเอกสารที่เหมือนกัน และแทนคุณลักษณะในเอกสารโดยใช้คำเดี่ยว พยางค์ วลี กลุ่มของคำ ประโยค เพื่อเป็นตัวแทนคุณลักษณะของเอกสาร และทำการคำนวณหาค่าน้ำหนักของคุณลักษณะ อาจใช้เทคนิคการวิเคราะห์ทางภาษาเข้ามาช่วย (Language Analysis) เพื่อตัดคำที่ไม่จำเป็นออกหรือตัดคำตามสถิติของคลังประโยค เพื่อลดมิติของขนาดเอกสาร เนื่องจากชุดของเอกสารเป็นชุดข้อความที่มีเทอมไม่ซ้ำหรือคำที่เกิดขึ้นในเอกสารทั้งหมด โอกาสเกิดของคำสามารถเกิดขึ้นเป็นหลักล้านคำทำให้ข้อมูลมีมิติมาก จึงต้องประยุกต์ใช้การคัดเลือกคุณลักษณะที่ดีเพื่อนำมาหาประสิทธิภาพ ลดเวลาในการประมวลผลและทรัพยากร [11] อัลกอริทึมในการคัดเลือกคุณลักษณะที่ใช้ในการวิจัยนี้ 3 วิธี ได้แก่

3.1 Chi-Square คือ การประเมินค่าของแอททริบิวต์ โดยคำนวณค่า Chi-Square ทางสถิติ [12] ดังแสดงในสมการที่ (1)

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

3.2 Information Gain [5] คือการประเมินค่าเพื่อใช้ในการแบ่งข้อมูลด้วยการคำนวณค่า Gain สำหรับแต่ละมิติข้อมูลถ้ามิติข้อมูลใดมีค่า Gain สูงสุด จะถูกเลือกให้เป็นกลุ่มย่อยที่มีอำนาจจำแนกดังสมการที่ (2) [6] แสดงการคำนวณค่า Information Gain

$$\text{InfoGain}(S, A) = \text{Entropy}(S) - \left| \frac{S_v}{S} \right| E(S_v) \quad (2)$$

โดยที่

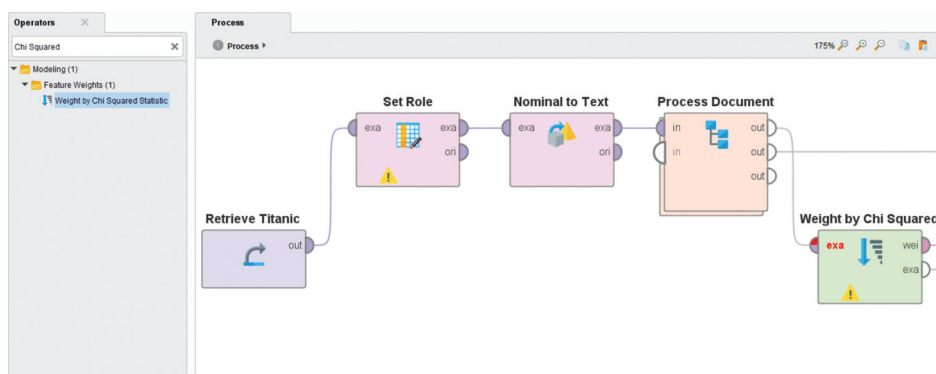
A	คือ	Attribute A
$ S_v $	คือ	สมาชิกของ Attribute A ที่มีค่า V
$ S $	คือ	จำนวนสมาชิกของกลุ่มตัวอย่าง

3.3 Gain Ratio [13] มีหลักการทำงานในลักษณะเดียวกับการทำงานของต้นไม้ตัดสินใจ โดยทำการหาตัวแปรที่เป็นตัวแบ่งข้อมูลออกเป็นกลุ่มย่อยที่มีสมาชิกภายในกลุ่มเป็นชนิดเดียวกันมากที่สุด ด้วยสิ่งที่เรียกว่า อัตราส่วนเกน (Gain Ratio) ซึ่งอัตราส่วนเกนคือ มาตรการได้ประโยชน์จากการแบ่งกลุ่มย่อย โดยค่าของอัตราส่วนเกนมาจากอัตราส่วนระหว่างค่าเกน (Gain) กับค่าสารสนเทศการแบ่งกลุ่ม (Split Info) [6] ซึ่งเป็นการลดปัจจัยของตัวแปรที่มีค่าหลากหลาย โดยสมการหาอัตราส่วนเกนสามารถเขียนได้ดังสมการที่ (3) - (4)

$$\text{SplitINFO} = - \sum_{i=1}^k \frac{n_i}{n} \log \frac{n_i}{n} \quad (3)$$

$$\text{GainRatio}(A) = \frac{\text{InfoGain}(A)}{\text{SplitINFO}(A)} \quad (4)$$

ในการคัดเลือกคุณลักษณะด้วยเทคนิค Chi-Square, Information Gain, Gain Ratio โดยใช้โปรแกรม RapidMiner Studio 9 โดยทำการเลือกเทคนิคการคัดเลือกคุณลักษณะ ได้จากส่วนของ Operators ซึ่งจะอยู่ในหมวด Modeling/Feature Weights ซึ่งการคัดเลือกคุณลักษณะด้วยโปรแกรม RapidMiner Studio 9 ได้ผลลัพธ์ดังรูปที่ 5



รูปที่ 5 การคัดเลือกคุณลักษณะด้วยโปรแกรม RapidMiner Studio 9

4. การสร้างแบบจำลองเพื่อจำแนกความคิดเห็น

ในกระบวนการสร้างแบบจำลองจะนำชุดคุณลักษณะของคำแต่ละประเภทมาทำการแบ่งข้อมูลออกเป็น 2 ชุด คือ (1) ชุดข้อมูลเรียนรู้ (2) ชุดข้อมูลทดสอบ และนำข้อมูลทั้งสองชุดไปสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการจัดกลุ่มความคิดเห็นที่ได้ใช้ข้อมูลความคิดเห็นที่เป็นภาษาไทย โดยใช้เทคนิคการจำแนกของการทำเหมืองข้อความมาใช้ในการจำแนกข้อความ ในงานวิจัยนี้ได้ศึกษาค้นคว้างานวิจัยที่เกี่ยวข้องกับการจำแนกข้อความ จึงได้เทคนิคที่มีประสิทธิภาพในการสร้างแบบจำลอง ผู้วิจัยได้เลือกเทคนิคในการจำแนกข้อความ เพื่อนำมาสร้างและเปรียบเทียบประสิทธิภาพของแบบจำลองเพื่อจำแนกความคิดเห็นของลูกค้าต่อร้านอาหาร ซึ่งมีทั้งหมด 4 เทคนิค ได้แก่

4.1 เทคนิค Naive Bayes [9], [14] คือ เทคนิควิธีการจำแนกประเภทโดยอาศัยหลักการของทฤษฎีความน่าจะเป็นตามกฎของเบย์ เพื่อหาว่าสมมติฐานใดน่าจะเป็นความถูกต้องมากที่สุด ซึ่งจะสามารถบ่งบอกถึงความน่าจะเป็นของข้อมูลชุดหนึ่งที่จะอยู่ในหมวดหมู่ของข้อมูลนั้น ๆ ซึ่งเป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยใช้ในการจำแนกหมวดหมู่เอกสารข้อความ (Text Classification) ได้ดี การทำงานไม่ซับซ้อนเหมาะกับการรันของเซตตัวอย่างที่มีจำนวนมาก เป็นวิธีที่ใช้งานได้ดีและง่ายต่อการประมวลผล มีการทำงานที่ไม่ซับซ้อนเหมือนวิธีการอื่น ๆ ซึ่งทฤษฎีของเบย์สามารถคำนวณได้ดังสมการที่ (5)

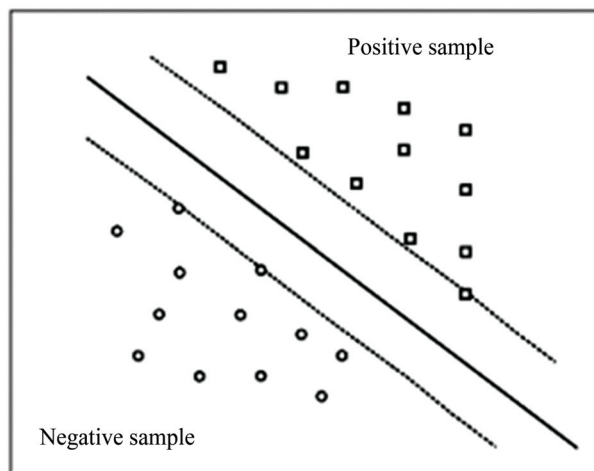
$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (5)$$

โดย

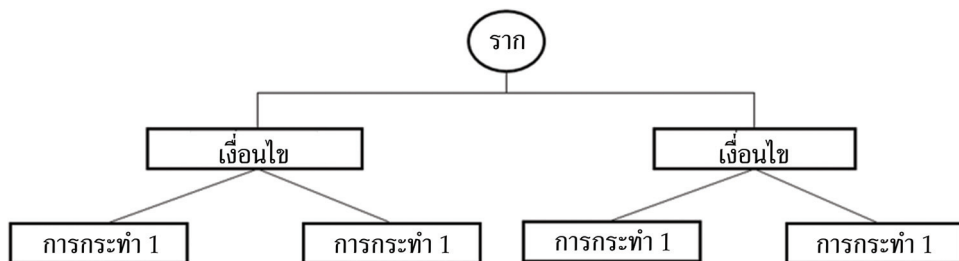
X		แทนข้อมูลการแจกแจงความน่าจะเป็น
H	คือ	คลาสของข้อมูล
$P(H)$	คือ	ความน่าจะเป็นก่อนหน้าของ H
$P(X)$	คือ	ความน่าจะเป็นก่อนหน้าของ X
$P(H X)$	คือ	ความน่าจะเป็นของ H เมื่อรู้ X
$P(X H)$	คือ	ความน่าจะเป็นของ X เมื่อรู้ H

4.2 เทคนิค Support Vector Machine [8], [15] คือ ขั้นตอนวิธีการที่มีความรวดเร็วและเป็นอัลกอริทึมที่มีความสามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มของข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปที่เส้นแบ่งแยกแยะกลุ่มข้อมูลได้ดีที่สุด แนวความคิดของเทคนิควิธี SVM นั้นเกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางในฟีเจอร์สเปซ (Feature Space) ในลักษณะเชิงเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่าเส้นตรงที่แบ่งกลุ่มสองกลุ่มออกจากกันนั้นเส้นใดเป็นเส้นที่ดีที่สุดดังตัวอย่างในรูปที่ 6

4.3 เทคนิคต้นไม้ตัดสินใจแบบ C4.5 [16] - [17] เป็นเทคนิควิธีหนึ่งที่ได้ทำการพัฒนามาจากเทคนิค ID3 ซึ่งได้กลายมาเป็นอัลกอริทึมพื้นฐานที่ใช้ในการเปรียบเทียบประสิทธิภาพของการทำงานอัลกอริทึมต่าง ๆ ใช้ในการสร้างต้นไม้ประกอบการตัดสินใจสำหรับใช้ในการจำแนกประเภทข้อมูล โดยในการสร้างต้นไม้ตัดสินใจ C4.5 จะใช้ค่ามาตรฐานส่วนเกิน (Gain Ratio) เพื่อทำการคัดเลือกคุณลักษณะเพื่อทำการคัดเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ดังตัวอย่างในรูปที่ 7



รูปที่ 6 ตัวอย่างระนาบการตัดสินใจของ Support Vector Machine



รูปที่ 7 ตัวอย่างต้นไม้ตัดสินใจสำหรับการจำแนกคุณลักษณะ

4.4 เทคนิค K-Nearest Neighbor [10], [18] คือ เป็นเทคนิคที่ใช้ในการจัดหมวดหมู่ข้อมูลที่เรียบง่ายและมีประสิทธิภาพ แต่เทคนิคนี้มีวิธีการจัดหมวดหมู่ที่แตกต่างจากเทคนิคอื่นตรงที่ไม่ได้ใช้ข้อมูลชุดฝึกในการสร้างแบบจำลอง แต่จะใช้ข้อมูลทั้งชุดมาเป็นตัวแบบจำลองเลย หลักการจัดหมวดหมู่ข้อมูลจะพิจารณาจากค่า k เช่น ค่า $k = 5$ หมายถึง ขั้นตอนวิธีจะพิจารณาข้อมูลจำนวน 5 ชุด มีลักษณะใกล้เคียงกับข้อมูลชุดตัวอย่าง การคำนวณความใกล้เคียงจะใช้วิธีการวัดระยะทางยูคลิดีเนียน (Euclidian Distance) จากสมการที่ (6) ภายหลังจากที่รวบรวมจำนวนสมาชิกที่ใกล้เคียงที่สุด k ตัวได้แล้ว โดยการเลือกหมวดหมู่จะเลือกหมวดหมู่ที่สมาชิกส่วนใหญ่ที่อยู่ในกลุ่ม k สังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ที่มาเข้ากลุ่มด้วย ในงานวิจัยนี้กำหนดค่า k เท่ากับ 1

$$dist(X_i, X_j) = \sqrt{\sum_{k=1}^n (X_{i,k} - X_{j,k})^2} \quad (6)$$

โดย

$dist(X_i, X_j)$ คือ ระยะห่างระหว่างข้อมูลตัวอย่าง X_i กับข้อมูลตัวอย่าง X_j

n คือ จำนวนคุณลักษณะทั้งหมดของข้อมูลตัวอย่าง

$X_{i,k}$ คือ คุณลักษณะตัวที่ k ของข้อมูลตัวอย่าง X_i

5. การวัดประสิทธิภาพของแบบจำลอง

ในการประเมินผลของแบบจำลองใช้เทคนิคการวัดประสิทธิภาพแบบ 10-fold Cross Validation โดยทำการแบ่งข้อมูลออกเป็น 10 กลุ่มเท่า ๆ กัน จากนั้นในแต่ละรอบการทดสอบจะใช้ข้อมูล 1 ชุดเป็นชุดข้อมูลทดสอบ และให้อีกชุดที่เหลือเป็นชุดฝึกสอน ซึ่งผู้วิจัยได้ทำการทดสอบแบบจำลองในการจำแนกความคิดเห็นของลูกค้าที่มีต่อร้านอาหาร โดยใช้แบบจำลองที่ละ 1 แบบจำลอง และนำข้อมูลทั้งหมด 4,965 ชุดข้อมูล ทำการแบ่งออกเป็นทั้งหมด 10 กลุ่ม ทั้งนี้จะแบ่งกลุ่มข้อมูลเพื่อใช้เป็นข้อมูลชุดทดสอบ 1 ชุด และที่เหลือจะเป็นข้อมูลเรียนรู้ ซึ่งข้อมูลทดสอบคิดเป็นอัตราร้อยละ 10 และข้อมูลเรียนรู้คิดเป็นอัตราร้อยละ 90 โดยค่าที่ใช้ในการวัดประสิทธิภาพของแบบจำลอง คือได้ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ดังสมการที่ (7) - (9) ตามลำดับ

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (7)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (8)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (9)$$

โดย

<i>TP</i>	คือ	จำนวนข้อมูลความคิดเห็นเชิงบวกที่จำแนกได้ถูกต้อง
<i>TN</i>	คือ	จำนวนข้อมูลความคิดเห็นเชิงลบบวกที่ทำนายได้ถูกต้อง
<i>FP</i>	คือ	จำนวนข้อมูลความคิดเห็นเชิงบวกที่จำแนกผิด
<i>FN</i>	คือ	จำนวนข้อมูลความคิดเห็นเชิงลบที่จำแนกผิด

ผลการวิจัย

งานวิจัยนี้ได้ทำการทดลองการจำแนกความคิดเห็นของลูกค้าต่อร้านอาหารด้วยเทคนิคการจำแนกของเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 4 เทคนิค คือ เทคนิค Naive Bayes เทคนิค Support Vector Machine เทคนิคต้นไม้ตัดสินใจแบบ C4.5 และเทคนิค K-Nearest Neighbor ผ่านการใช้งานภาษาโปรแกรม Python โดยชุดข้อมูลความคิดเห็นที่ใช้ในงานวิจัยนี้รวบรวมจากเว็บไซต์ wongnai.com จำนวน 4,487 ชุดข้อมูล แบ่งเป็นความคิดเห็นเชิงบวกจำนวน 2,248 ความคิดเห็น และความคิดเห็นเชิงลบ 2,239 ความคิดเห็น จากนั้นทำการคัดเลือกเฉพาะประเภทของคำนามกับคำกริยาและคำวิเศษณ์ เพื่อคัดเลือกเป็นคำคุณลักษณะ จากนั้นทำการคัดเลือกคุณลักษณะโดยใช้เทคนิค Chi-Square, Information Gain และ Gain Ratio ซึ่งจะทำให้การเลือกโดยใช้ค่าที่มีค่าคะแนนสูงสุดจำนวน 1,000 คำ เพื่อสร้างเป็นตัวแทนเอกสารที่จะเตรียมเข้าสู่แบบจำลอง และประเมินประสิทธิภาพของแบบจำลองทั้ง 4 เทคนิค โดยใช้วิธีการ 10-fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึก ได้ผลการเปรียบเทียบดังนี้

1. ค่าความถูกต้อง (Accuracy) คือ การวัดประสิทธิภาพความถูกต้องของแบบจำลอง จากการทดลองผลการเปรียบเทียบค่าความถูกต้องแสดงได้ดังตารางที่ 1

ตารางที่ 1 ผลการเปรียบเทียบค่าความถูกต้องของเทคนิคการจำแนกข้อความร่วมกับการคัดเลือกคุณลักษณะ

Algorithms	Chi- Square	Information Gain	Gain Ratio
Naive Bayes	87.10	89.08	87.94
SVM	87.61	88.14	86.03
C4.5	70.85	71.03	71.76
K-Nearest Neighbor	66.30	66.70	65.68

จากตารางที่ 1 การจำแนกข้อความโดยใช้คำนาม คำกริยา และคำวิเศษณ์ร่วมกับการคัดเลือกคุณลักษณะ ผลปรากฏว่า การคัดเลือกคุณลักษณะแบบ Chi- Square โดยใช้เทคนิคการจำแนกข้อความแบบ Support Vector Machine ให้ค่าความถูกต้องมากที่สุดคือ 87.61 % รองลงมาคือเทคนิค Naive Bayes 87.10 % เมื่อใช้เทคนิค Information Gain ในการคัดเลือกคุณลักษณะพบว่าเทคนิค Naive Bayes ให้ค่าความถูกต้องสูงที่สุดที่ 89.08 % รองลงมาคือ Support Vector Machine ที่ 88.14 % และเทคนิคการคัดเลือกคุณลักษณะแบบ Gain Ratio พบว่าเทคนิค Naive Bayes ให้ค่าความถูกต้องสูงที่สุดที่ 87.94 % รองลงมาคือ Support Vector Machine ที่ 86.03 % จากการเปรียบเทียบค่าความถูกต้องของการจำแนกข้อความพบว่า การคัดเลือกคุณลักษณะแบบ Information Gain โดยใช้เทคนิคการจำแนกข้อความแบบ Naive Bayes ให้ค่าความถูกต้องสูงที่สุดที่ 89.08 %

2. ค่าความแม่นยำ (Precision) คือ การวัดความสามารถในการจำแนกเอกสารที่ไม่เกี่ยวข้องออกไปโดยค่าความแม่นยำเป็นอัตราส่วนของจำนวนข้อมูลที่ถูกนำมาใช้อย่างถูกต้องกับจำนวนเอกสารที่ถูกดึงออกมาทั้งหมด ผลการเปรียบเทียบค่าความแม่นยำแสดงได้ดังตารางที่ 2

ตารางที่ 2 ผลการเปรียบเทียบค่าความแม่นยำของเทคนิคการจำแนกข้อความร่วมกับการคัดเลือกคุณลักษณะ

Algorithms	Chi- Square	Information Gain	Gain Ratio
Naive Bayes	87.14	89.12	87.97
SVM	87.64	88.17	86.05
C4.5	71.95	72.14	72.08
K-Nearest Neighbor	66.49	67.36	65.96

จากตารางที่ 2 การจำแนกข้อความโดยใช้คำนาม คำกริยา และคำวิเศษณ์ร่วมกับการคัดเลือกคุณลักษณะ ผลปรากฏว่า การคัดเลือกคุณลักษณะแบบ Chi- Square โดยใช้เทคนิคการจำแนกข้อความแบบ Support Vector Machine ให้ค่าความแม่นยำมากที่สุดคือ 87.64 % รองลงมาคือ เทคนิค

Naive Bayes 87.14 % เมื่อใช้เทคนิค Information Gain ในการคัดเลือกคุณลักษณะพบว่าเทคนิค Naive Bayes ให้ค่าความแม่นยำสูงที่สุดที่ 89.12 % รองลงมาคือ Support Vector Machine ที่ 88.17 % และเทคนิคการคัดเลือกคุณลักษณะแบบ Gain Ratio พบว่าเทคนิค Naive Bayes ให้ค่าความแม่นยำสูงที่สุดที่ 87.97 % รองลงมาคือ Support Vector Machine ที่ 86.05 % จากการเปรียบเทียบค่าความแม่นยำของการจำแนกข้อความพบว่า การคัดเลือกคุณลักษณะแบบ Information Gain โดยใช้เทคนิคการจำแนกข้อความแบบ Naive Bayes ให้ค่าความแม่นยำสูงที่สุดที่ 89.12 %

3. ค่าความระลึก (Recall) คือ การวัดความสามารถของแบบจำลองในการดึงเอกสารที่เกี่ยวข้องออกมา โดยค่าความระลึกเป็นอัตราส่วนของจำนวนข้อมูลที่ถูกนำมาใช้อย่างถูกต้องกับจำนวนเอกสารที่เกี่ยวข้องทั้งหมด ผลการเปรียบเทียบค่าความระลึกแสดงได้ดังตารางที่ 3

ตารางที่ 3 ผลการเปรียบเทียบค่าความระลึกของเทคนิคการจำแนกข้อความร่วมกับการคัดเลือกคุณลักษณะ

Algorithms	Chi- Square	Information Gain	Gain Ratio
Naive Bayes	87.11	89.10	87.96
SVM	87.62	88.15	86.04
C4.5	70.84	71.02	71.77
K-Nearest Neighbor	66.29	66.66	65.74

จากตารางที่ 3 การจำแนกข้อความโดยใช้ค่านาม คำกริยา และคำวิเศษณ์ร่วมกับการคัดเลือกคุณลักษณะ ผลปรากฏว่า การคัดเลือกคุณลักษณะแบบ Chi- Square โดยใช้เทคนิคการจำแนกข้อความแบบ Support Vector Machine ให้ค่าความระลึกมากที่สุดคือ 87.62 % รองลงมาคือเทคนิค Naive Bayes 87.11 % เมื่อใช้เทคนิค Information Gain ในการคัดเลือกคุณลักษณะพบว่าเทคนิค Naive Bayes ให้ค่าความระลึกสูงที่สุดที่ 89.10 % รองลงมาคือ Support Vector Machine ที่ 88.15 % และเทคนิคการคัดเลือกคุณลักษณะแบบ Gain Ratio พบว่าเทคนิค Naive Bayes ให้ค่าความแม่นยำสูงที่สุดที่ 87.96 % รองลงมาคือ Support Vector Machine ที่ 86.04 % จากการเปรียบเทียบค่าความระลึกของการจำแนกข้อความพบว่า การคัดเลือกคุณลักษณะด้วย Information Gain และใช้เทคนิคการจำแนกข้อความแบบ Naive Bayes ให้ค่าความระลึกสูงที่สุดที่ 89.10 %

จากการเปรียบเทียบการคัดเลือกคุณลักษณะที่เหมาะสมสำหรับนำมาจำแนกข้อความ โดยใช้ค่านาม คำกริยา และคำวิเศษณ์มาร่วมกันโดยใช้วิธีการคัดเลือกคุณลักษณะแบบ Information Gain แล้วเลือกค่าที่มีค่าคะแนนสูงสุดจำนวน 1,000 คำ พบว่าเทคนิคการจำแนกข้อความแบบ Naive Bayes ให้ผลการวัดประสิทธิภาพดีที่สุด ให้ค่าความถูกต้องเท่ากับ 89.08 % ค่าความแม่นยำเท่ากับ 89.12 % และค่าความระลึกเท่ากับ 89.10 %

สรุปผลการวิจัย

งานวิจัยนี้ได้ทำการค้นหาเทคนิคการทำเหมืองข้อความที่มีประสิทธิภาพในการจำแนกความคิดเห็นของลูกค้าที่มีต่อร้านอาหารที่ถูกเขียนขึ้นด้วยภาษาไทย โดยชุดข้อมูลความคิดเห็นที่ใช้ในงานวิจัยนี้ รวบรวมจากเว็บไซต์ wongnai.com จำนวน 4,487 ชุดข้อมูล แบ่งเป็นความคิดเห็นเชิงบวกจำนวน 2,248 ความคิดเห็น และความคิดเห็นเชิงลบ 2,239 โดยงานวิจัยนี้ได้ใช้ คำนาม คำกริยา และคำวิเศษณ์ เพื่อนำมาทำการจำแนกความคิดเห็น แล้วทำการคัดเลือกคุณลักษณะโดยใช้เทคนิค Chi-Square, Information Gain และ Gain Ratio ซึ่งจะทำให้การเลือกโดยใช้ค่าที่คะแนนสูงสุดจำนวน 1,000 คำ เพื่อสร้างเป็นตัวแทนเอกสาร จากนั้นได้นำเอาเทคนิควิธีการจำแนกข้อความมาทำการจำแนกความคิดเห็นภาษาไทย โดยทำการเปรียบเทียบทั้งหมด 4 เทคนิค คือ เทคนิค Naive Bayes เทคนิค Support Vector Machine เทคนิค ต้นไม้ตัดสินใจแบบ C4.5 และเทคนิค K-Nearest Neighbor สำหรับขั้นตอนการวัดประสิทธิภาพของทั้ง 5 เทคนิค ได้ใช้หลักการ 10-fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ วัดประสิทธิภาพของแบบจำลองด้วยค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึก ผลจากการวิจัยพบว่าเมื่อใช้การคัดเลือกคุณลักษณะแบบ Information Gain ร่วมกับเทคนิคการจำแนกข้อความแบบ Naive Bayes ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่าความถูกต้อง 89.08 % ค่าความแม่นยำ 89.12 % และค่าความระลึก 89.10 % รองลงมาคือการคัดเลือกคุณลักษณะแบบ Information Gain ร่วมกับเทคนิคการจำแนกข้อความแบบ Support Vector Machine โดยให้ค่าความถูกต้อง 88.14 % ค่าความแม่นยำ 88.17 % และค่าความระลึก 88.15 %

จากผลการวิจัยพบว่า เทคนิค Naive Bayes มีประสิทธิภาพมากที่สุดในการจำแนกข้อความ โดยพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ และค่าความระลึก เหตุผลที่เทคนิค Naive Bayes มีประสิทธิภาพสูงที่สุดเนื่องจากมีการทำงานที่ไม่ซับซ้อนเหมาะกับกรณีที่มีข้อมูลจำนวนมาก มี Attribute ที่เป็นอิสระต่อกันเหมาะที่จะนำมาใช้ในการจำแนกข้อความ เทคนิคที่มีประสิทธิภาพในการจำแนกข้อความรองลงมาคือเทคนิค Support Vector Machine ที่ให้ผลลัพธ์ในการจำแนกใกล้เคียงกับเทคนิค Naive Bayes เนื่องจากเทคนิค Support Vector Machine สามารถใช้กับลักษณะข้อมูลที่เป็นตัวเลขหรือข้อมูลเชิงกลุ่มแบบข้อความ รองรับจำนวนตัวแปรที่หลากหลายได้เป็นจำนวนมาก ซึ่งจะแตกต่างจากเทคนิค K-Nearest Neighbor ที่ได้ผลการวัดประสิทธิภาพที่ต่ำกว่าอย่างมีนัยสำคัญ เนื่องจากถ้ามี Attribute หรือตัวแปรจำนวนมากจะเกิดปัญหาในการคำนวณทำให้ประสิทธิภาพลดลง เช่นเดียวกับเทคนิค C4.5 หากมีกิ่งที่มากเกินไปอาจทำให้การจำแนกมีประสิทธิภาพที่ลดลง

จากผลดังกล่าวผู้วิจัยจึงได้นำเอาคำนาม คำกริยา และคำวิเศษณ์รวมกันแล้วทำการคัดเลือกคุณลักษณะเพื่อลดมิติของข้อมูลด้วยเทคนิค Information Gain มาใช้ในการปรับปรุงการจำแนกความคิดเห็นของลูกค้า โดยใช้เทคนิคการจำแนกข้อความ Naive Bayes มาประยุกต์ใช้ในการสร้างแบบจำลองให้ค่าความถูกต้องสูงที่สุดเท่ากับ 89.08 % ซึ่งจะให้ประสิทธิภาพการจำแนกความคิดเห็นของลูกค้าได้เป็นอย่างดี และกระบวนการคัดเลือกคุณลักษณะด้วย Information Gain สามารถลดมิติของข้อมูลได้เป็นอย่างมาก ไม่ส่งผลกระทบต่อประสิทธิภาพในการจำแนกข้อความแต่อย่างใด ซึ่งกระบวนการที่นำเสนอนี้สามารถนำไปปรับใช้กับความคิดเห็นเกี่ยวกับสินค้าหรือบริการอย่างอื่น ๆ ได้

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณเว็บไซต์ www.wongnai.com ที่ให้ข้อมูลมาใช้ในการวิเคราะห์และดำเนินการวิจัย เพื่อให้งานออกมาได้ดีและมีประสิทธิภาพมากที่สุด

References

- [1] Laowsungsuk, P., Jinda, A., and Sitthisarn, S. (2017). Sentiment Analysis of Restaurant Reviews on Review Web Sites. **Thaksin University Journal**. Vol. 20, No. 1, pp. 39-47
- [2] Srisuan, J. and Hanskunatai, A. (2014). An Application of Hotel Searching Based on Opinion Mining. In **The 10th National Conference on Computing and Information Technology (NCCIT 2014)**. Faculty of Information Technology, King Mongkut's University of Technology North Bangkok. pp. 95-100
- [3] Chotnasiaw, P., Songpan, W., Arch-int, S., and Saiyod, S. (2017). Analysis of Affecting Factors to Customer Reviews using Opinion Mining. In **The 13th National Conference on Computing and Information Technology (NCCIT2017)**. Arnoma Grand Bangkok Hotel, Bangkok. pp. 44-50
- [4] Shah, F. P. and Patel, V. (2016). A Review on Feature Selection and Feature Extraction for Text Classification. In **International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)**. pp. 2264-2268. DOI: 10.1109/WiSPNET.2016.7566545
- [5] Haruechaiyasak, C., Jitkritum, W., Sangkeettrakarn, C., and Damrongrat, C. (2008). Implementing News Article Category Browsing Based on Text Categorization Technique. In **Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology**. Vol. 3, pp. 143-146. DOI: 10.1109/WIIAT.2008.61
- [6] Nuipian, V. and Meesad, P. (2013). A Comparison of Filter and Wrapper Approaches with Text Mining for Text Classification. **The Journal of Industrial Technology**. Vol. 9, No. 3, pp. 118-129
- [7] Marknakorn, N. (2013). **Comparison of Feature Selection Methods for Inappropriate Webpage Classification by Data Mining Technique**. Master of Science, Information Technology, King Mongkut's University of Technology North Bangkok.
- [8] Thetmueang, R. and Jirawichitchai, N. (2017). Thai Sentiment Analysis of Product Review Online Using Support Vector Machine. **Engineering Journal of Siam University**. Vol. 18, Issue 1, No. 34, pp. 1-12
- [9] Pinmuang, N. and Thongkam, J. (2018). Classifying Thai Opinions on Online Media using Text Mining. **Journal of Science and Technology Mahasarakham University**. Vol. 37, No. 3, pp. 372-379
- [10] Tipsena, R., Jareanpon, C., and Somprasertsri, G. (2013). Automatic Question Classification on Webboard Using Text Mining Techniques. **Journal of Science and Technology Mahasarakham University**. Vol. 33, No. 5, pp. 493-502

- [11] Phopli, W., Boonmatham, S., and Meesad, P. (2017). Ensemble Feature Selection for Sparse Data. In **The 13th National Conference on Computing and Information Technology (NCCIT2017)**. Arnoma Grand Bangkok Hotel, Bangkok. pp. 373-378
- [12] Saengsiri, P., Meesad, P., Wichian, S. N., and Herwig, U. (2010). Comparison of Hybrid Feature Selection Models on Gene Expression Data. In **2010 Eighth International Conference on ICT and Knowledge Engineering**. pp. 13-18. DOI: 10.1109/ICTKE.2010.5692905
- [13] Buathong, W. (2014). **Efficiency Improvement of Dimension Reduction by Feature Selection Methods for Data Classification**. Master of Science, Information Technology, King Mongkut's University of Technology North Bangkok
- [14] Pukkhem, N., Junmanee, C., and Ouisui, S. (2017). Automatic Thai Folk Wisdom Classification using Data Mining Approach. **Thaksin University Journal**. Vol. 20, No. 3, Special Issue 2017, pp. 300-307
- [15] Samal, B., Behera, A. K., and Panda, M. (2017). Performance Analysis of Supervised Machine Learning Techniques for Sentiment Analysis. In **IEEE 3rd International Conference on Sensing, Signal Processing and Security**. pp. 128-133
- [16] Kongthong, A., Haruechaiyasak, C., Pailai, J., and Kongyoung, S. (2012). The Role of Twitter During a Natural Disaster: Case Study of 2011 Thai Flood. In **2012 Proceedings of Technology Management for Emerging Technologies (PICMET)**. Vol. 12, pp. 2227-2232
- [17] Kaewta, C. and Mahawirawat, A. (2010). Diagnosis of Cases by Decision Tree Techniques. In **National Conference on Information Technology (NCIT2010)**. The Grand Ayudhaya Hotel, Bangkok. pp. 308-313
- [18] Hossain, F. M. T., Hossain, M. I., and Nawshin, S. (2017). Machine Learning Based Class Level Prediction of Restaurant Reviews. In **2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)**. pp. 420-423. DOI: 10.1109/R10-HTC.2017.8288989