การทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร

วิทยานิพนธ์

ของ

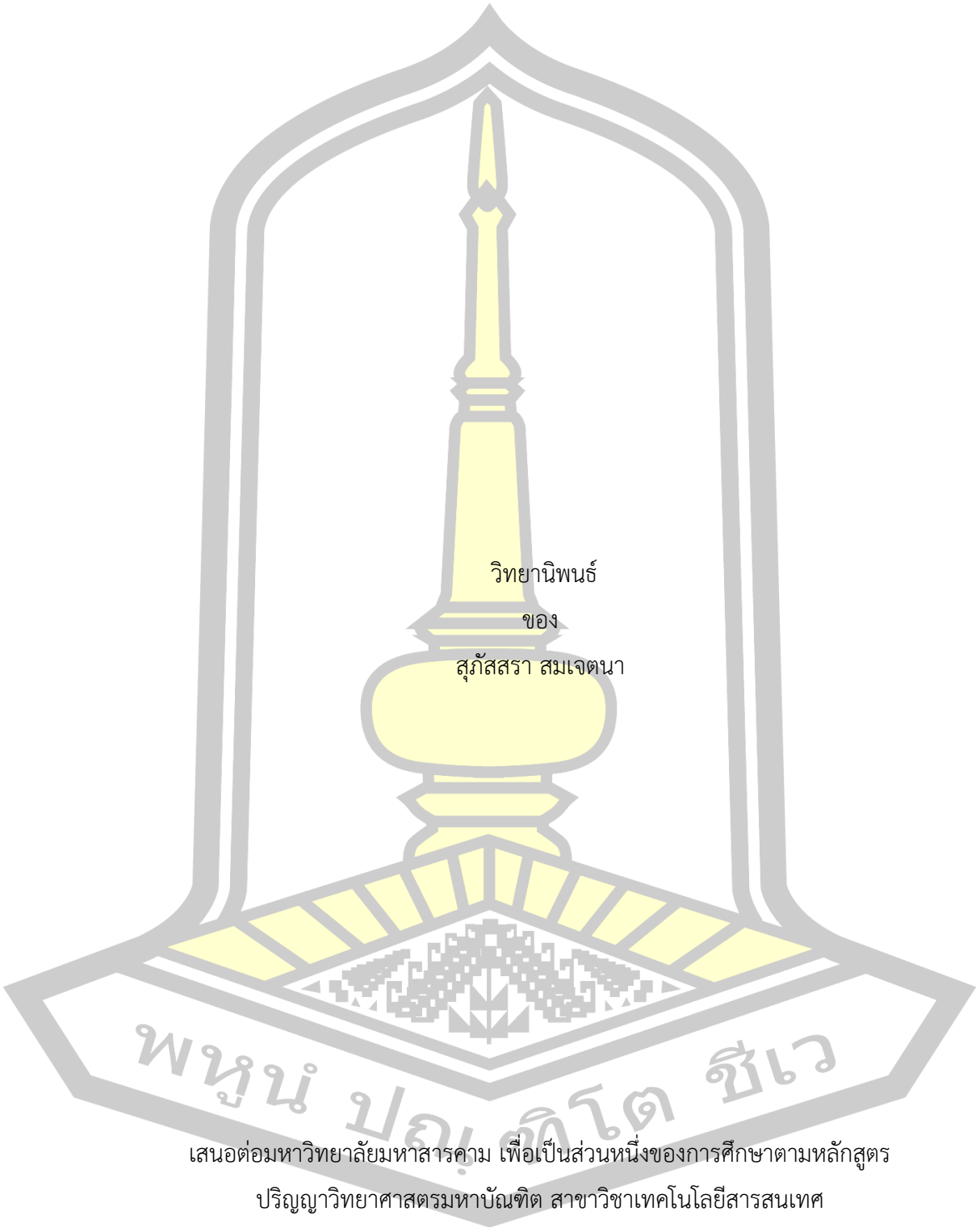
สุภัสสรา สมเจตนา

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มิถุนายน 2564

การทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร



วิทยานิพนธ์

ของ

สุภัสสรา สมเจตนา

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มิถุนายน 2564

Opinion Mining of Parent toward Children who use Smart Phone

Supatsara Somjettana

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

June 2021

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนางสาวสุภัสสรา สมเจตนา แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

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

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. จารี ทองคำ)

.....กรรมการ
(ผศ. ดร. แกมกาญจน์ สมประเสริฐศรี)

.....กรรมการ
(ดร. สาธิต แสงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร
ผู้วิจัย สุภัสสรา สมเจตนา
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จารี ทองคำ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2564

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคในเหมืองข้อมูลในการสร้างแบบจำลองจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร งานวิจัยนี้ใช้ 6 เทคนิค ได้แก่ เทคนิคริปเปอร์ เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวคเตอร์แมชชีน เทคนิคเคเนียร์เรสเนเบอร์ และเทคนิคต้นไม้ป่าสุ่ม มาสร้างแบบจำลองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยข้อมูลนั้นถูกรวบรวมมาเฉพาะความคิดเห็นของผู้ปกครองที่มีลักษณะเป็นข้อความภาษาไทยบนเครือข่ายสังคมออนไลน์ผ่านเว็ปไซต์พันทิปและเฟสบุ๊ค จำนวนทั้งหมด 1,925 ข้อความ คุณลักษณะที่ใช้ในงานวิจัยนี้ผู้วิจัยได้เลือกเฉพาะคำวิเศษณ์ที่สามารถระบุความรู้สึกได้เป็นอย่างดีมาใช้ในการสร้างแบบจำลอง ในส่วนของการทำดัชนีคำ ผู้วิจัยได้ใช้การให้ค่าน้ำหนักตามจำนวนคำที่พบและการใช้ถุงคำในการให้น้ำหนักคำ 10-โฟลด์ครอสวาลิเดชั่นได้ถูกนำมาใช้ในการแบ่งกลุ่มข้อมูลชุดเรียนรู้และชุดทดสอบ รวมถึงวัดประสิทธิภาพของแบบจำลองด้วยโดยใช้ค่าความถ่วงดุล ค่าความแม่นยำ และค่าความระลึก หลังจากทำการทดลองและวัดประสิทธิภาพของแบบจำลองด้วยวิธีการให้ค่าน้ำหนักตามจำนวนคำที่พบ พบว่า เทคนิคต้นไม้ป่าสุ่มเป็นเทคนิคที่ดีที่สุดในการวิเคราะห์ความคิดเห็นสำหรับข้อมูลชุดนี้ โดยให้ค่าความถ่วงดุล 83.55% ค่าความแม่นยำ 89.62% และค่าความระลึก 78.38% ส่วนการวิเคราะห์ข้อมูลด้วยวิธีการใช้ถุงคำในการให้น้ำหนักคำ พบว่า เทคนิคต้นไม้ป่าสุ่มเป็นเทคนิคที่ดีที่สุดเช่นกัน โดยให้ค่าความถ่วงดุล 49.29% ค่าความแม่นยำ 57.27% และค่าความระลึก 43.58%

คำสำคัญ : เหมืองความคิดเห็น, สมาร์ทโฟน, บุตร

TITLE Opinion Mining of Parent toward Children who use Smart Phone
AUTHOR Supatsara Somjettana
ADVISORS Assistant Professor Jaree Thongkam , Ph.D.
DEGREE Master of Science **MAJOR** Information Technology
UNIVERSITY Mahasarakham University **YEAR** 2021

ABSTRACT

The objective of this research was to compare the efficacy of a data mining technique in modeling parental opinion on the use of their children's smartphones. This research proposes the process of opinion mining to compare the model performance of 6 techniques which are the RIPPER, C4.5 decision tree, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, and Random Forest technique to analyze the opinions of parents on the use of smartphones of their children. Therefore, the data was collected from only the Thai text of parent's opinions on the social network which is Pantip and Facebook, a total of 1,925 messages. The characteristic in this research was used only adverbs which can indicate the feeling to create the models. Regrading to the index of words, the researcher has weighted the number of words which found and used the bag of words for weighting. Then, the 10 – fold cross validation was used to separate the data into training and testing sets and measured the performance the models by F-measure, Precision and Recall. After experiment and measurement of weighting the words directly, it was found that the Random Forest technique is the best technique for analyzing the data that F-measure was 83.55%, Precision was 89.62% and Recall was 78.38% while experiment and measurement by using the bag of words, the Random Forest technique is still the best technique for analyzing the data that F-measure was 49.29%, Precision was 57.27% and Recall was 43.58%

Keyword : Opinion Mining, Smart Phone, Children

พหูน ปณฺฑิโต ชีเว

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก รองศาสตราจารย์ ดร.สิทธิชัย บุษหมั่น ประธานกรรมการสอบ ผู้ช่วยศาสตราจารย์ ดร.แกมกาญจน์ สมประเสริฐศรี และดร.สาธิต แสงประดิษฐ์ กรรมการสอบ ที่กรุณาให้คำแนะนำปรึกษาตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่างๆ ด้วยความเอาใจใส่อย่างดียิ่ง ผู้วิจัยจึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ ที่ถ่ายทอดวิชาความรู้ตลอดจนคอยพร่ำสอนศิษย์ด้วยจิตเมตตา ผู้ซึ่งมีจิตวิญญาณของความเป็นครูโดยแท้จริงและแก้ไขข้อบกพร่องต่างๆด้วยความเอาใจใส่ เพื่อให้การศึกษาค้นคว้าอิสระฉบับนี้สมบูรณ์ที่สุด

ขอขอบพระคุณอาจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคามที่ให้ความรู้และคำแนะนำในการศึกษาจนสำเร็จการศึกษาในครั้งนี้

ขอขอบพระคุณ บิดามารดา ครอบครัวและเพื่อนๆที่คอยสนับสนุนและเป็นกำลังใจ จนทำให้งานการศึกษาค้นคว้าวิทยานิพนธ์ครั้งนี้ได้สำเร็จไปด้วยดี

สุภัสสรา สมเจตนา

สารบัญ

หน้า

สารบัญภาพ

สารบัญตาราง

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ในปัจจุบันเป็นที่ทราบกันดีอยู่แล้วว่าสมาร์ทโฟน ได้เข้ามามีบทบาทมาก ทั้งด้านการทำงาน การติดต่อสื่อสาร จนกลายเป็นส่วนหนึ่งในการใช้ชีวิตประจำวันไปแล้ว ไม่ว่าจะเป็นผู้สูงอายุ วัยทำงาน วัยรุ่น รวมไปถึงเด็กเล็กๆ ที่ปัจจุบันนี้ก็มีการใช้สมาร์ทโฟนมากขึ้น และการใช้ก็มีแนวโน้มเพิ่มสูงขึ้นเรื่อยๆ [1] โดยเฉพาะในกลุ่มเด็กๆ ได้มีการใช้สมาร์ทโฟนกันสูงมาก [2] โดยขาดการความเอาใจใส่ของพ่อแม่ ดังมีผลการศึกษาพบว่าเด็กส่วนใหญ่ที่อาศัยอยู่กับพ่อแม่กว่าร้อยละ 80 ที่มีการใช้โทรศัพท์แบบสมาร์ทโฟนที่ตัวเองเป็นเจ้าของและมีการใช้ได้อย่างอิสระ จำนวนชั่วโมงเฉลี่ยที่เล่นต่อวันอยู่ประมาณ 3 ชั่วโมง และมีพฤติกรรมติดเกมมากถึงร้อยละ 45 [3] ซึ่งพ่อแม่ในยุคปัจจุบันต้องทำงานไม่มีเวลาดูแล หรือฝากให้ปู่ย่าตายายหรือญาติคนอื่นๆเลี้ยงให้ การให้เด็กใช้สมาร์ทโฟนทำให้เด็กสงบนิ่ง ส่งผลกระทบพัฒนาการด้านทักษะการสื่อสาร ทักษะการแก้ปัญหา ทักษะการปฏิสัมพันธ์กับสังคมโดยตรงทั้งทางด้านสมองและสายตา [4] [5] แย่ลง แต่บางครั้งการใช้ใช้สมาร์ทโฟนของเด็กสามารถช่วยเพิ่มพัฒนาการภาษา [6] ทำให้ผู้ปกครองเกิดการแลกเปลี่ยนความคิดเห็นในเฟสบุ๊คและเว็ปไซต์พันธุ์ทิพย์ เพื่อแนะนำการใช้สมาร์ทโฟนที่เป็นประโยชน์ให้แก่บุตร ซึ่งความคิดเห็นของผู้ปกครองจะเป็นเชิงบวกและเชิงลบ ในการวิเคราะห์ความคิดเห็นของผู้ปกครองที่มีต่อบุตรที่ใช้สมาร์ทโฟนในเชิงบวกและเชิงลบ เพื่อสนับสนุนการตัดสินใจให้แก่ผู้ปกครองในการเลี้ยงบุตรที่เหมาะสมกับยุคปัจจุบัน

เหมืองความคิดเห็น (Opinion Mining) [7] เป็นกระบวนการวิเคราะห์และจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์ จากความคิดเห็นส่วนใหญ่ที่เกิดขึ้นจะเป็นไปในเชิงบวกและเชิงลบ [8] [9] [10] นักวิจัยได้ใช้เทคนิคริปเปอร์ (RIPPER) เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 (Decision tree C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคซัพพอร์ตเวคเตอร์แมชชีน (SVM Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-NN) และเทคนิคการสุ่มป่าไม้ (Random forest) ในการวิเคราะห์ความคิดเห็น เช่น ผดุง นันอำไพและ จารี ทองคำ [11] ได้ทำการวิจัยเกี่ยวกับการตรวจจับการบุกรุกในระบบเครือข่ายด้วยเทคนิคการจำแนกในการทำเหมืองข้อมูลด้วย 4 เทคนิค คือ เทคนิค Decision Table, เทคนิค Naïve Bayes, เทคนิค RIPPER และเทคนิค PART decision list พบว่าแบบจำลองที่ใช้เทคนิค RIPPER มีค่า Precision มากที่สุดคือ 99.00% ส่วน Gupta และคณะ [9] ได้ทำเหมืองความคิดเห็นของลูกค้าที่มาใช้บริการโรงแรมและให้คะแนนการบริการ กลับพบว่า เทคนิค Decision Tree มีความ แม่นยำถึง 76.22 เปอร์เซ็นต์ และง่ายต่อการเข้าใจของพนักงานโรงแรม ส่วน

นุชนาฎ ปิ่นเมือง และ จารี ทองคำ [8] ได้ทำการวิจัยเกี่ยวกับความคิดเห็นของคนไทยต่อสื่อออนไลน์ด้วยเทคนิค 5 เทคนิค อันได้แก่ เทคนิค Naïve Bayes, SVM, KNN, decision tree และ C4.5 พบว่า เทคนิค Naïve Bayes มีประสิทธิภาพในการจำแนกความคิดเห็น โดยมีค่าความถูกต้องมากถึงร้อยละ 93.88 และค่าความแม่นยำร้อยละ 94.02 ส่วนประพัฒน์ พรมน้ำอ่าง และคณะ [12] ได้ทำการวิจัยเกี่ยวกับการจำแนกกลุ่มข้อความรีวิว โดยใช้เทคนิคเหมืองข้อมูล ซึ่งประกอบด้วย เทคนิค SVM เทคนิค Decition Tree เทคนิค k-NN และ เทคนิค Naïve Bayes จากการทดลองพบว่าโดยเทคนิค SVM ได้ค่าความถูกต้องสูงที่สุดอยู่ที่ 86.26 % ส่วนราชวิทย์ ทิพย์เสนา และคณะ [13] ได้ทำการวิจัยเกี่ยวกับการจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา โดยใช้เทคนิคเหมืองข้อความด้วย 3 เทคนิควิธี คือ เทคนิคการหาเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจ และเทคนิคการเรียนรู้เบย์อย่างง่าย ผลการเปรียบเทียบประสิทธิภาพแสดงให้เห็นว่า เทคนิคการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด โดยค่าความถูกต้องเท่ากับ 0.89 ค่าความเที่ยง เท่ากับ 0.9 ค่าความระลึก เท่ากับ 0.89 และค่า F-Measure เท่ากับ 0.892 วสวัตติ์ อินทร์แปลง และจารี ทองคำ [14] ได้ทำการวิเคราะห์ความคิดเห็นต่อเกมมือถือพับจีด้วยเหมืองข้อความ จาก 5 เทคนิค คือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine (SVM) เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ C4.5 และเทคนิค Random Forest จากการทดสอบและวัดประสิทธิภาพของโมเดล พบว่า เทคนิค K-Nearest Neighbor ให้ผลดีที่สุด โดยให้ค่าความแม่นยำ 99.75% ค่าความระลึก 100% และค่าความถูกต้อง 99.87% และวัชรีวรรณ จิตต์สกุล ได้วิเคราะห์ความเสถียรของเทคนิค Conjunctive Rule, Random Forest, Bayesian Logistic Regression และ Support Vector Machine ในการจำแนกข้อความแสดงความคิดเห็น เชิงบวก และเชิงลบในการให้บริการของเว็ปไซต์ พบว่าเทคนิคที่ให้ผลลัพธ์ดีที่สุด คือ เทคนิค Random Forest

ดังนั้นผู้วิจัยจึงมีแนวคิดที่จะพัฒนาแบบจำลองในการทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยการนำความคิดเห็นของผู้ปกครองบนเครือข่ายสังคมออนไลน์เว็บไซต์ Pantip และ Facebook ว่ามีความคิดเห็นไปในเชิงบวกหรือเชิงลบ โดยการเปรียบเทียบเทคนิคที่ใช้ในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็น ได้แก่ เทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest แล้ววัดประสิทธิภาพของแบบจำลองด้วยค่าความถ่วงดุล (F-Measure) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษากระบวนการในการทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร

1.2.2 เพื่อเปรียบเทียบประสิทธิภาพของจำลองที่สร้างโดยเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร

1.3 ความสำคัญของงานวิจัย

1.3.1 ได้ทราบกระบวนการทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร

1.3.2 ได้แบบจำลองที่มีประสิทธิภาพในจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร

1.4 ขอบเขตของการวิจัย

1.4.1 เก็บรวบรวมข้อความความคิดเห็นบนเครือข่ายสังคมออนไลน์ จากเว็ปไซต์ Pantip และ Facebook ระหว่างวันที่ 1 มกราคม 2561 – 31 ธันวาคม 2562 โดยข้อความความคิดเห็นที่นำมาวิเคราะห์เป็นภาษาไทยเท่านั้น

1.4.2 สร้างแบบจำลองในการวิเคราะห์เหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ในรูปแบบของภาษาไทย ด้วยเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest วัดประสิทธิภาพด้วยค่าความถ่วงดุล ค่าความแม่นยำ และค่าความระลึก

1.5 นิยามศัพท์เฉพาะ

1.5.1 สมาร์ทโฟน คือ โทรศัพท์เคลื่อนที่ที่มีความสามารถเป็นคอมพิวเตอร์พกพาที่ทำงานในลักษณะของโทรศัพท์เคลื่อนที่

1.5.2 แบบจำลอง คือ ต้นแบบที่ใช้ในวิเคราะห์เหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร

1.5.3 ความคิดเห็นของผู้ปกครอง คือ ข้อความแสดงความคิดเห็นที่ประกอบด้วยข้อมูลอันเป็นข้อเท็จจริงกับการแสดงความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร มีทั้งเชิงบวกและเชิงลบ

1.5.4 การวิเคราะห์ความคิดเห็น หมายถึง การวิเคราะห์ข้อความที่แสดงความรู้สึกหรืออารมณ์ของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยแบ่งออกเป็นความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้เป็นการศึกษาความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร การทำเหมืองความคิดเห็น เทคนิคที่ใช้ในการวิจัยและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ความคิดเห็นของผู้ปกครองต่อบุตร

ความคิดเห็นของผู้ปกครองต่อบุตร คือ ข้อความแสดงความคิดเห็นที่ประกอบด้วยข้อมูลอันเป็นข้อเท็จจริงกับการแสดงความคิดเห็นของผู้ปกครองต่อบุตรที่ใช้สมาร์ทโฟน มีทั้งเชิงบวกและเชิงลบ ดังที่ได้แสดงความคิดเห็นกันในสื่อต่างๆกันอย่างมากมาย ถึงประโยชน์และโทษของเด็กที่ใช้สมาร์โฟน ดังเช่นเมื่อมีข่าวที่เกี่ยวกับเด็กที่ใช้สมาร์ทโฟนเป็นเวลานานๆ ส่งผลทำให้ดวงตาอักเสบจนถึงขั้นต้องผ่าตัดดวงตา แต่ในขณะเดียวกันก็มีผู้ปกครองบางส่วนที่สามารถใช้สมาร์โฟนให้เกิดประโยชน์กับเด็กได้เป็นอย่างดีเช่นกัน

ข้อมูลความคิดเห็นที่จะนำมาวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตรนั้น จะเป็นข้อความที่มีลักษณะเป็นภาษาไทย ที่ได้มีการแสดงความคิดเห็นบนเครือข่ายสังคมออนไลน์ จากเว็ปไซต์ Pantip และ Facebook ซึ่งจะใช้ข้อความความคิดเห็นจากเพจหลักที่ได้รับความนิยมและน่าเชื่อถือที่ได้รับการจัดอับดับแนะนำ 10 เพจหมอเด็ก ให้ความรู้เรื่องการเลี้ยงลูก ที่พ่อกับแม่ต้องติดตาม เท่านั้น เช่น เพจเลี้ยงลูกนอกบ้าน, นายแพทย์ประเสริฐ ผลิตผลการพิมพ์, Dr.Pam book club, เลี้ยงลูกตามใจหมอ, เลี้ยงลูกให้เป็นคนปกติ, ชมรมจิตแพทย์เด็กและวัยรุ่นแห่งประเทศไทย, เข็นเด็กขึ้นภูเขา, สุธีรา เอื้อไพโรจน์กิจ, เลี้ยงลูกโตไปด้วยกันกับหมออร Hormone for Kids, หมอเสาวภา เลี้ยงลูกเชิงบวก และเพจสารพันปัญหาการเลี้ยงลูก โดยใช้คำสำคัญในการค้นหา เช่น เด็กกับสมาร์โฟน, เด็กเล่นมือถือ, การเล่นมือถือของเด็ก, เด็กติดโทรศัพท์, ประโยชน์ของมือถือ เป็นต้น

2.1.1.1 ความคิดเห็นของผู้ปกครองเชิงบวกต่อบุตร เช่น วสุพล จิตรานนท์และดร.ปัทมาวดี เล่ห์มงคล [15] ได้ศึกษาความคิดเห็นของผู้ปกครองเด็กปฐมวัยที่มีต่อสื่อการ์ตูน ผู้ปกครองเด็กปฐมวัยส่วนใหญ่มีความคิดเห็นเกี่ยวกับสื่อการ์ตูนในประเด็นที่ลูกของตนชอบดูเป็นการ์ตูนฝรั่ง มีความเห็นด้วยมากที่สุดในเรื่อง การแนะนำสื่อการ์ตูนที่ดีให้ลูก ควรดูแลสอดส่องการเข้าถึงสื่อการ์ตูน

ของลูก เพื่อไม่ให้เข้าถึงสื่อที่ไม่เหมาะสมการ์ตูนที่ดีก็มีมากให้สาระความรู้ต่างๆ มากมาย เช่น การ์ตูนสอนทางด้านภาษา เป็นต้น

2.1.1.2 ความคิดเห็นของผู้ปกครองเชิงลบต่อบุตร เช่น สุวภา บุญอุไร และคณะ [16] ได้ศึกษาพฤติกรรมการใช้สมาร์ทโฟนของเด็กและผู้ปกครองที่มีผลต่อความฉลาดทางอารมณ์ของเด็กปฐมวัย พบว่าการใช้สมาร์ทโฟนของเด็กมีอิทธิพลทางตรงเชิงลบต่อความฉลาดทางอารมณ์ของเด็กปฐมวัย เนื่องจากเด็กได้รับสิ่งเร้าที่เป็นแสง สี เสียง ที่เกิดขึ้นอย่างรวดเร็ว ส่งผลทำให้เด็กไม่รู้จักการรอคอย ทำให้เด็กขาดสมาธิและขาดการปฏิสัมพันธ์กับบุคคลรอบข้าง จดจ่ออยู่แต่กับสมาร์ทโฟน ไม่อยู่ในโลกของความจริง ส่งผลถึงพัฒนาการด้านอารมณ์และจิตใจโดยตรง

2.1.2 การทำเหมืองความคิดเห็น

การทำเหมืองความคิดเห็น (Opinion Mining) เป็นการทำเหมืองข้อความอีกประเภทหนึ่ง เพื่อพัฒนาเป็นระบบในส่วนเฉพาะที่เป็นความคิดเห็น ซึ่งลักษณะของข้อความนั้นๆที่แสดงความคิดเห็นออกมาเป็นถ้อยคำต่างๆ จะสะท้อนถึงทัศนคติ ความรู้สึก เหตุผล แสดงให้เห็นถึงความคิดเห็นนั้นว่าเป็นความคิดเห็นเชิงบวกหรือเชิงลบ การทำเหมืองความคิดเห็นจะทำให้การวิเคราะห์ข้อความคิดเห็นจำนวนมากได้อย่างรวดเร็ว เช่น การวิเคราะห์ความคิดเห็นของผู้บริโภคที่มีต่อสินค้าและบริการ เพื่อเพิ่มความสามารถในการแข่งขัน ลดการสำรวจตลาดและสามารถนำข้อมูลที่ผ่านการวิเคราะห์ด้วยเหมืองความคิดเห็นแล้วมาปรับกลยุทธ์ เพื่อให้ตรงตามความต้องการของลูกค้าได้อย่างรวดเร็ว หรือการติดตามการเปลี่ยนทัศนคติของประชาชนทั่วไปเกี่ยวกับการเมือง ที่มีความรวดเร็วสามารถปรับเปลี่ยนได้อยู่ตลอดเวลา [17] กระบวนการทำเหมืองความคิดเห็น มีกระบวนการคล้ายคลึงกับกระบวนการค้นหาความรู้จากฐานข้อมูล โดยสามารถแบ่งกระบวนการทำงานออกเป็น 5 ขั้นตอนสำคัญ มีกระบวนการดังนี้

1. การรวบรวมข้อมูล (Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำเหมืองความคิดเห็น รวมถึงการนำความคิดเห็นที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาในเบื้องต้นตามขอบเขตที่ต้องการทำการศึกษา

2. การเตรียมข้อมูล (Data Preparation) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลความคิดเห็นที่จะนำมาใช้วิเคราะห์ว่ามีความถูกต้อง โดยการนำความคิดเห็นที่ไม่ถูกต้องออกหรือเป็นขั้นตอนที่อาจต้องแก้ไขก่อนนำไปใช้งาน

3. การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลความคิดเห็นให้เหมาะสมและตรงกับรูปแบบที่จะประมวลผลต่อไป เช่น การตัดบางคอลัมน์ที่ไม่จำเป็นออก

ประเภทของคำ ในภาษาไทยโดยใช้แนวคิดของ ORCHID โดยใช้คลังข้อมูล Orchid Corpus ซึ่งเป็นคลังบทความ การกำกับประโยค และชนิดของคำในภาษาไทยที่นิยมนำมาใช้กันอย่างแพร่หลาย ซึ่งได้แบ่งกลุ่มคำภาษาไทยออกเป็น 47 ประเภท ดังตารางที่ 2.1

ตารางที่ 2.1 ประเภทคำโดยใช้คลังข้อมูล Orchid Corpus

No.	POS	Description	Example
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator "ไม่"	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator "ไม่"	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator "ไม่"	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น

ตารางที่ 2.1 ประเภทคำโดยใช้คลังข้อมูล Orchid Corpus

No.	POS	Description	Example
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ 2 ตัว
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็วๆ, เสมอๆ, ช้าๆ
30	ADVP	Adverb with prefixed form	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
35	CFQC	Frequency classifier	ครั้ง, เที่ยว
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก, ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย,โอ้, เออ, เอ๋, อ๋อ

ตารางที่ 2.1 ประเภทคำโดยใช้คลังข้อมูล Orchid Corpus

No.	POS	Description	Example
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentencev	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
47	PUNC	Punctuation	(,), ", ,, ;

4. การสร้างแบบจำลอง เป็นขั้นตอนประมวลผลโดยใช้เทคนิคต่างๆ ในการค้นหารูปแบบของข้อมูลด้วยการจำแนก (Classification) เพื่อการพยากรณ์

5. การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความและการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรมีการนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย

จากกระบวนการวิเคราะห์เหมืองข้อความข้างต้นได้นำมาประยุกต์ใช้ในงานวิจัยนี้เพื่อการวิเคราะห์ความคิดเห็นในรูปแบบของภาษาไทย ซึ่งจำเป็นต้องใช้วิธีการประมวลผลภาษาธรรมชาติโดยอาศัยความรู้ในการเข้าใจภาษาธรรมชาติ

2.1.3 เทคนิคที่ใช้ในงานวิจัย

เทคนิคที่ใช้ในงานวิจัยนี้ เป็นเทคนิคที่อยู่ในกลุ่มต่างๆ ของการทำเหมืองข้อความประกอบด้วย 6 เทคนิค คือ เทคนิคริปเปอร์ (RIPPER) เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 (Decision tree C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคซัพพอร์ตเวคเตอร์แมชชีน (SVM Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-NN) และเทคนิคการสุ่มป่าไม้ (Random forest)

2.1.3.1 แรปเปอร์

เทคนิคแรปเปอร์ (Repeated Incremental Pruning to Produce Error Reduction: RIPPER) [11] [18] ถูกสร้างโดย Cohen ในปี 1995 เป็นเทคนิคที่พัฒนาจากเทคนิค IRIP สามารถสร้างกฎเองได้ โดยการเรียนรู้จากข้อมูลที่เตรียมไว้ให้กฎที่สร้างขึ้นจะอยู่ในรูปของ if then else ประกอบด้วย 2 เฟส คือ เฟสที่ 1 ระบุกฎเริ่มต้น (Building) เฟสที่ 2 ระบุค่า porst –

process rule optimization โดยการเรียนรู้จากข้อมูลที่กำหนดคลาสไว้เรียบร้อยแล้ว (Traning data) ซึ่งแบ่งเป็น Growing set และ pruning set และจะหาค่าที่ดีที่สุด growing set ใน rule space อธิบายได้จาก BNF จากที่ได้ Growing set แล้วจะทำการ pruning ข้อมูลทันที เพื่อให้ได้ผลออกมาแล้วเป็นที่พอใจมากที่สุด ดังแสดงในภาพที่ 2.1

```
Ruleset Ripper(D){ //D เป็นชุดข้อมูลทั้งหมด
  RS = {} //RS เป็นชุดกฎว่าง
  Ci ordered in increasing prior probability
  //ชุดกฎสำหรับแต่ละคลาส Ci เรียงลำดับคลาสตามค่าความน่าจะเป็น
  for p = 1 to K – 1 {
        Pos = Cp, Neg = Cp+1, …, Ck
        //Pos เป็นชุดข้อมูลประเภทบวก, Neg เป็นชุดข้อมูลประเภทลบ
  RSp = {} // RSp เป็นชุดกฎว่างที่จัดเก็บกฎจากกระบวนการเติบโตและตัดแต่งกฎ
  while D contains positive samples {
        Divide D into Grow set G and Prune set P
        //G เป็นกฎที่เติบโต, P เป็นกฎที่ถูกตัดออก
        r = GrowRule(G)
        PruneRule(r,P)
        if CalculateError(r)> 0.5 {
              break
        } else {
              RSp = RSp + r
              Remove examples covered by r from D
         }
        }
        for i = 1 to 2 {
              OptimizeRuleset(RSp, D)
              SimplifyRuleset(RSp, D)
        }
              RS = RS + RSp
  }//RS เป็นชุดกฎที่ได้จากการเรียนรู้กฎทั้งหมด
  return RS
}
```

ภาพที่ 2.1 กระบวนการเรียนรู้กฎด้วยเทคนิค RIPPER

2.1.3.2 ต้นไม้ตัดสินใจ C4.5

ต้นไม้ตัดสินใจ (Decision tree C4.5) [8] เป็นเทคนิควิธีเหมืองข้อมูลที่เป็นที่นิยมกันมาก เนื่องจากเป็นวิธีการวิเคราะห์ที่ง่ายต่อการตีความหมาย ต้นไม้ตัดสินใจนั้นจะประกอบไปด้วยโหนด (Node) ซึ่งจะทำหน้าที่ในการแสดงคุณลักษณะที่ใช้สำหรับการทดสอบข้อมูล กิ่ง (Branch) เป็นส่วนที่จะแสดงคุณสมบัติในโหนดที่ได้มีการแตกออกมา และใบ (Leaf) จะแสดงกลุ่มหรือคลาสที่ได้มีการกำหนดเอาไว้ และในหาความสัมพันธ์ของแต่ละโหนดที่แสดงคุณลักษณะ (Attribute) นั้นจะใช้ค่า Information Gain เพื่อหาความสัมพันธ์ของในแต่ละโหนดคุณลักษณะ โดยการใช้ค่า Information Gain นี้จะช่วยลดจำนวนครั้งที่ใช้ในการทดสอบและทำให้ต้นไม้ตัดสินใจที่ได้ไม่มีความซับซ้อนมากจนเกินไป

ต้นไม้ตัดสินใจ C4.5 เป็นอัลกอริทึมที่พัฒนามาจากอัลกอริทึม ID3 โดย Quinlan ที่พัฒนาขึ้นขึ้นเพื่อแก้ปัญหาในกรณีที่มีข้อมูลในชุดที่นำมาทดสอบมีการกระจายตัวมากๆ หรือไม่เป็นกลุ่มเดียวกัน ทำให้เมื่อจำแนกแล้วเกิดความลำเอียงไปในกลุ่มที่มีข้อมูลอยู่เป็นจำนวนมาก จึงได้ทำการเพิ่มค่าการแบ่งแยกเพื่อใช้ในการคำนวณและลดปัญหาการเกิดการเอนเอียงในการจำแนก ดังสมการที่ 1

$$\text{SplitInfo}_A(D) = -\sum_{i=1}^{n} \frac{|t_i|}{|T|} \log_2 \frac{|t_i|}{|T|} \quad (1)$$

จากการที่ค่าแบ่งแยกมีการกระจายตัว บ่งบอกถึงการกระจายของข้อมูล วิธีแก้ปัญหาความเอนเอียงจึงนำค่าสารสนเทศในการแบ่งแยกหารด้วยค่ามาตรฐานเกน จะได้ค่า Gain Ratio เพื่อเป็นตัวเลือกใช้คุณลักษณะที่จะนำมาเป็นโหนดในลำดับต่อไป ดังสมการที่ 2

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitInfo}_A(D)} \quad (2)$$

เมื่อนำข้อมูลมาคำนวณหาค่า Gain Ratio แล้วเลือกค่าที่ที่มีค่าสูงสุดมาเป็นโหนดเริ่มต้นและทำการร้างโหนดในระดับต่อๆไป โดยใช้คุณลักษณะที่มีค่ารองลงมาต่อๆไป จนข้อมูลมีการกระจายตัวน้อยที่สุดหรือข้อมูลมีค่าเท่ากัน

2.1.3.3 การเรียนรู้แบบเบย์

การเรียนรู้แบบเบย์ (Bayesian Learning) เป็นเทคนิควิธีการจำแนกประเภทมีพื้นฐานมาจากกฎของเบย์ เป็นทฤษฎีทางด้านสถิติโดยนำความน่าจะเป็นมาใช้ ประเมินความไม่

แน่นอนให้เป็นตัวเลขได้ กล่าวถึง ความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น (A) ถ้ามี เหตุการณ์อีกเหตุการณ์หนึ่งเกิดมาแล้ว (B) สามารถ เขียนให้อยู่ในรูปอย่างง่าย ดังสมการที่ 3

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

P(A|B) คือ ความน่าจะเป็นที่เหตุการณ์ A จะเกิดขึ้น ถ้าเหตุการณ์ B เกิดขึ้นแล้ว

P(B|A) คือ ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น ถ้าเหตุการณ์ A เกิดขึ้นแล้ว

P(A) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ A

P(B) คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ B

วิธีการจำแนกหมวดหมู่โดยใช้หลักการความน่าจะเป็น เป็นการแก้ปัญหาแบบ Classification สามารถคาดการณ์ผลลัพธ์และสามารถอธิบายได้ ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เป็นวิธีการจำแนกประเภทข้อมูลที่มี ประสิทธิภาพวิธีหนึ่ง โดยใช้ในการจำแนกหมวดหมู่เอกสารข้อความ (Text Classification) ได้ดี การทำงานไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นดัง สมการที่ 4

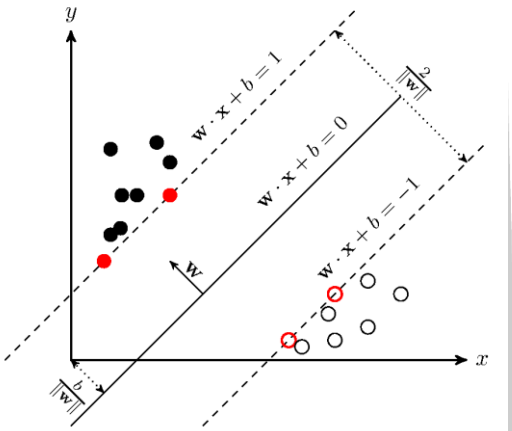
$$P(A_1, A_2, \dots, A_n|C_i) = \prod_{i=1}^{n} P(A_i|C_i) \quad (4)$$

กลุ่ม C_i สำหรับข้อมูลที่มี คุณสมบัติ n ตัว $X = (A_1, A_2, \dots, A_n)$ หรือใช้สัญลักษณ์ว่า $P(A_1, A_2, \dots, A_n|C_i)$ โดยที่ $\prod$ หมายถึงผลคูณของค่า $P(A_i|C_i)$ ทั้งหมด i = 1, 2, 3, ..., n และ j = 1, 2, 3,..., n ดังนั้นจะได้วิธีการจำแนกประเภทแบบเบย์ อย่างง่ายดังสมการที่ 5

$$V_{NB} = argmax P(C_i) \prod_{i=1}^{n} P(A_i|C_i) \quad (5)$$

2.1.3.4 ซัพพอร์ทเวกเตอร์แมชชีน

ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) [19] เป็นทฤษฎีที่ใช้เพื่อลดความผิดพลาดจากการทำนาย (Minimize error) จัดเป็นเทคนิคที่ใช้ในการแก้ปัญหาทางด้านการรู้จำรูปแบบข้อมูลอาศัยหลักการจำแนกหมวดหมู่ข้อมูลด้วยการหาระนาบตัดสินใจและแบ่งข้อมูลออกเป็น 2 ส่วน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะระหว่างขอบเขตทั้ง 2 กลุ่มมากที่สุด (Optimal separating hyperplane) [20] เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูล โดยใช้ฟังก์ชั่นแม็ปสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space ในลักษณะเชิงเส้น ดังแสดงในภาพที่ 2.2



ภาพที่ 2.2 การวางตัวของข้อมูลในลักษณะเชิงเส้น

สร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) Kernal Function ในสื่อตีพิมพ์เกี่ยวกับ SVM [21] จะเรียกตัวแปรในการตัดสินใจว่าคุณสมบัติ และตัวแปรที่เปลี่ยนแปลงใช้ในการกำหนดระนาบหลายมิติเรียกว่า คุณลักษณะ (feature) ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า การคัดเลือกคุณลักษณะ (feature selection) จำนวนเซตของคุณลักษณะที่ใช้อธิบายในกรณีหนึ่ง (เช่น แถวของการค่าคาดการณ์) เรียกว่า เวกเตอร์ (vector) ดังนั้น จุดมุ่งหมายของตัวแบบ SVM คือการได้ประโยชน์สูงสุดจากระนาบหลายมิติที่แบ่งแยกกลุ่มของเวกเตอร์ ในกรณีนี้ด้วย Feature Space เหมาะใช้สำหรับข้อมูลที่มีมิติของข้อมูลสูง กำหนดให้ $(x_i, y_i), \ldots, (x_n, y_n)$ เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูลตัวอย่าง m คือจำนวนมิติข้อมูลเข้า และ y คือ ผลลัพธ์ มีค่า +1 หรือ -1 ดังสมการที่ 6

$$(xR\{+1,-1(5)_i,y_i),\ldots,(x_n,y_n)$$ เมื่อ x m , y ∈∈ (6)

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูก แบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจ ซึ่งคำนวณได้ดังสมการที่ 7

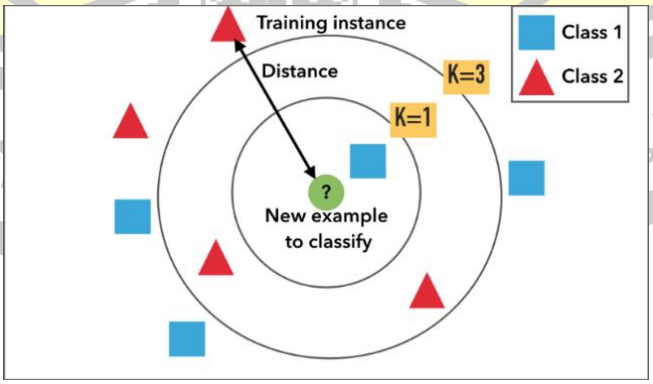
$$(w \times x) + b = 0 \quad (7)$$

เมื่อ w คือ ค่าน้ำหนักและ b คือค่า bias สมการ ใช้สำหรับจำแนกประเภทของข้อมูล ดังสมการที่ 8

$$(w \times x) + b > 0 \text{ ถ้า } y_i + 1 \text{ และ } (w \times x) + b < 0 \text{ ถ้า } y_i - 1 \quad (8)$$

2.1.3.5 เคเนียเรสเนเบอร์

เคเนียเรสเนเบอร์ (K-Nearest Neighbor: K-NN) เป็นวิธีการในการจัดแบ่งคลาส โดยเทคนิคนี้จะตัดสินใจว่า คลาสใดที่จะแทนเงื่อนไขหรือกรณีใหม่ ๆ ได้บ้าง หลักการของวิธีการนี้ [22] จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงกันมากที่สุด K ตัวจากข้อมูลบนชุดข้อมูลตัวอย่าง ทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่หรือข้อมูลที่ป้อนถาม (Input Query Instance) กับข้อมูลตัวอย่างฝึกฝนจะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ดังแสดงในภาพที่ 2.3



ภาพที่ 2.3 การจัดกลุ่มของเทคนิค K-NN

ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง k ตัว ประกอบด้วยแอตทริบิวท์หลายตัวแปร $\mathbf{X_i}$ ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายหากรณีใหม่ อัลกอริทึมแบบ K-NN ได้แก่ 1-NN , 2-NN , 3-NN , ... k-NN โดยค่า k ต้องระบุในการสร้างโมเดล [22] มาตรวัดความถูกต้อง (Distance Measure) การหาความยาวระหว่างจุดที่ต้องการโดยใช้เครื่องมือ และวิธีต่าง ๆ งานวิจัยได้เลือกวิธีการหามาตรวัดความแม่นยำ โดย Euclidean Distance ระยะทาง ระหว่าง 2 จุด จุดที่จะวัดนั้นมีเงื่อนไขมีหลายค่าจากหลายมิติหรือขนาดขึ้นกับรูปแบบ ซึ่งสามารถพิสูจน์หาค่าได้ด้วยทฤษฎีของ Pythagorean เมื่อมีการใช้สูตรเพื่อหาระยะทางขนาดของ Euclidean ระยะทางระหว่างจุด $P = (p_1, p_2, \ldots, p_3)$ และ $Q = (q_q, q_2, \ldots, q_n)$ ใน Euclidean หลายขนาดระบุได้เป็น ดังสมการที่ 9

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (9)$$

2.1.3.6 เทคนิคป่าไม้สุ่ม

เทคนิคป่าไม้สุ่ม (Random Forest: RF) เทคนิคที่ทำการสุ่มเลือกคุณสมบัติออกมาจากชุดข้อมูลหลายๆชุด [23] จากนั้นนำเอาชุดของคุณสมบัติเหล่านั้นมาสร้างแบบจำลองด้วยเทคนิคต้นไม้ตัดสินใจหลายๆต้น โดยเทคนิคการสุ่มป่าไม้ถูกนำเสนอครั้งแรกในปี ค.ศ. 1995 โดย Tin Kam ซึ่งต่อมาเทคนิคนี้ถูกต่อยอดโดย Leo Breiman ลักษณะของต้นไม้ที่อยู่ภายในป่าของเทคนิคการสุ่มป่าไม้จะถูกควบคุมด้วย 3 ปัจจัยคือ

1. ต้นไม้แต่ละต้นจะถูกสอน (Train) โดยการใช้เซตย่อยจากข้อมูลตัวอย่าง

2. เมื่อต้นไม้โตขึ้น จะสามารถค้นหาโนด (Node) แต่ละโนดที่อยู่ในกิ่งที่ดีที่สุดของต้นไม้โดยใช้การสุ่มเลือกคุณสมบัติจาก N คุณสมบัติ

3. ต้นไม้แต่ละต้นจะไม่มีการตัดออก แต่จะปล่อยให้ต้นไม้โตขึ้นไปเรื่อยๆ จนได้ผลลัพธ์ที่ดีที่สุดหลังจากการสร้างป่า แล้วทำการให้คะแนน (Vote) โดยต้นไม้ภายในป่า หากต้นไม้ต้นใดได้คะแนนสูงสุดก็จะนำเอาต้นไม้นั้นออกมาสร้างเป็นโมเดล

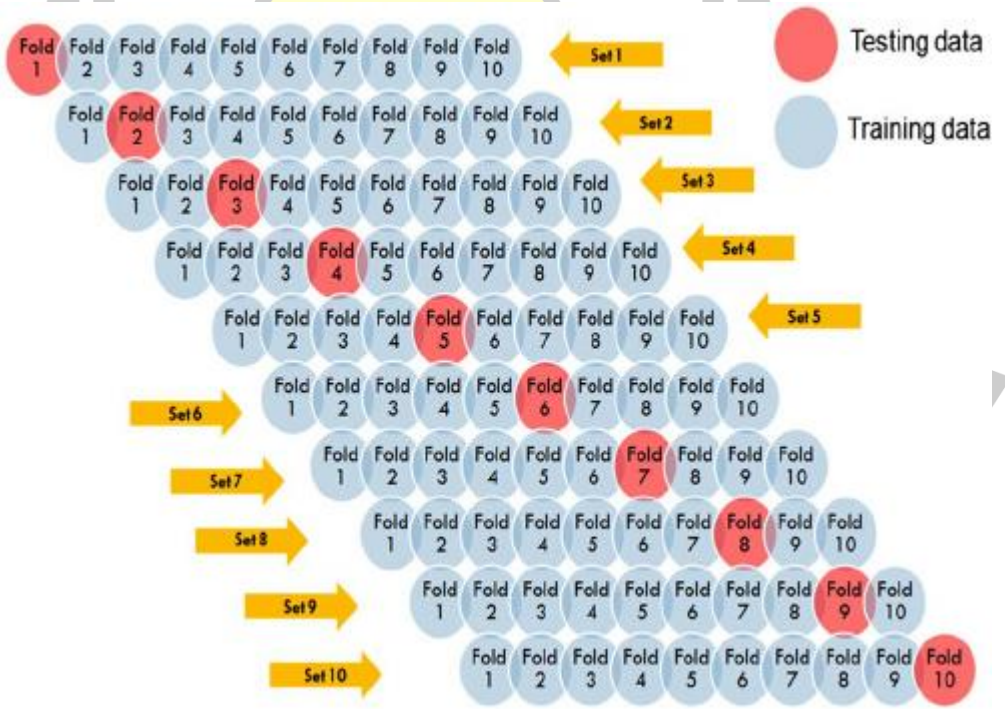
มีนักวิจัยได้นำเสนอเทคนิค Random Forest ในการจำแนก เช่น วัชรีวรรณ จิตต์สกุล [24] ได้วิเคราะห์ความเสถียรของอัลกอริทึมเพื่อการจำแนก 4 รูปแบบพื้นฐาน ได้แก่ Conjunctive Rule, Random Forest , Bayesian Logistic Regression, Support Vector

Machine กับข้อความคิดเห็น เชิงบวก และเชิงลบในการให้บริการของเว็ปไซต์ พบว่าอัลกอริทึมที่ให้ผลลัพธ์ดีที่สุด คือ อัลกอริทึม Random Forest

2.1.4 การวัดประสิทธิภาพของแบบจำลอง

2.1.4.1 10-fold cross validation

10-fold cross validation คือ การเลือกสุ่มข้อมูลแบบความเที่ยงตรง ซึ่งวิธีนี้เป็นที่นิยมในการทำงานวิจัยเพื่อใช้ในการทดสอบประสิทธิภาพของโมเดล เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี 10-fold cross-validation จะทำการเลือกสุ่มข้อมูลออกเป็น K ชุดเท่าๆ กัน จากนั้นจะทำการทดลองครั้งแรกด้วยข้อมูลชุดที่ 1 ซึ่งเป็นข้อมูลทดสอบและกำหนดให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั่งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ โดยวิธีการทดสอบประสิทธิภาพแบบ 10-fold cross validation มีข้อเสียคือ จะต้องทำการเริ่มทดสอบใหม่โดยจะต้องทำทั้งหมด K รอบ



ภาพที่ 2.4 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation

จากภาพที่ 2.4 จะเป็นการทดสอบประสิทธิภาพแบบ 10-fold cross validation ซึ่งจะทำการแบ่งชุดข้อมูลออกเป็น 10 ชุด โดยในแต่ละรอบจะใช้ชุดข้อมูลเพื่อเป็นชุดข้อมูลทดสอบ 1 ชุด และให้ชุดข้อมูลอื่นๆ เป็นข้อมูลชุดสอน โดยจะทำการทดสอบทั้งหมด 10 รอบ จากนั้นนำค่าความถ่วงดุล ค่าความแม่นยำ และค่าความระลึก

2.1.4.2 การวิเคราะห์ประสิทธิภาพ

การวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาค่าค่าความถ่วงดุล ค่าความแม่นยำ และค่าความระลึก โดย

ค่าความถ่วงดุล (F-Measure) คือ การวัดประสิทธิภาพโดยรวมของทั้งสองค่าระหว่างค่าความแม่นยำและค่าความครบถ้วน ซึ่งนำค่าทั้งสองมาคำนวณร่วมกัน

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (10)$$

ค่าความแม่นยำ (Precision) คือ การวัดความสามารถในการที่จะขจัดเอกสารที่ไม่เกี่ยวข้องออกไปโดยที่ค่าความแม่นยำนั้นจะเป็นอัตราส่วนของจำนวนเอกสารที่เกี่ยวข้องและได้มีการถูกดึงออกมา เพื่อเทียบกับจำนวนเอกสารที่ถูกดึงออกมาทั้งหมด

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

ค่าความระลึก (Recall) [20] คือ การวัดความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องออกมา โดยค่าความระลึกนั้นจะใช้อัตราส่วนของจำนวนเอกสารที่เกี่ยวข้องและได้มีการถูกดึงออกมา เทียบกับจำนวนเอกสารที่เกี่ยวข้องทั้งหมด

ซึ่งสามารถหาได้จากสมการดังต่อไปนี้

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

โดยที่ TP คือ จำนวนข้อมูลที่ถูกนำมาใช้อย่างถูกต้อง

TN คือ จำนวนข้อมูลที่ผิดที่ถูกนำมาใช้

FP คือ จำนวนข้อมูลที่ถูกต้องแต่ไม่นำมาใช้

FN คือ จำนวนข้อมูลที่ผิดแต่ไม่นำมาใช้

2.2 งานวิจัยที่เกี่ยวข้อง

ผดุง นันอำไพและ จารี ทองคำ [11] ได้ทำการวิจัยเกี่ยวกับการตรวจจับการบุกรุกด้วยเทคนิคการจำแนกในการทำเหมืองข้อมูลในระบบเครือข่าย ด้วยเทคนิคการจำแนกในการทำเหมืองข้อมูลด้วย 4 เทคนิค คือ เทคนิค Decision Table, เทคนิค Naïve Bayes, เทคนิค RIPPER และเทคนิค PART decision list เพื่อพัฒนาแบบจำลองและเปรียบเทียบประสิทธิภาพการจำแนกรูปแบบการบุกรุกในระบบเครือข่าย ในโดยใช้ชุดข้อมูลการบุกรุกระบบเครือข่ายจากฐานข้อมูลความรู้ KDD Cup'99 หลักการ 10-Fold Cross Validationสถิติที่ใช้ในการวิจัย 1) ค่าความแม่นยา Precision 2) ค่าระลึก Recall และ 3) ค่า F-Measure ผลการทดลองพบว่าแบบจำลองที่ใช้เทคนิค RIPPER มีค่า Precision มากที่สุดคือ 99.00% และผลเปรียบเทียบการวิเคราะห์แบบจำลองการจำแนกการตรวจจับการบุกรุกของแต่ละเทคนิค พบว่า แบบจำลองที่ใช้เทคนิค RIPPER ให้ค่าเฉลี่ยทางสถิติเป็นเปอร์เซ็นต์มากที่สุดอย่างมีนัยสำคัญ จากเทคนิคทั้งหมด 4 เทคนิค

Gupta และคณะ [9] ได้ทำเหมืองความคิดเห็นของลูกค้าที่มาใช้บริการโรงแรมและให้คะแนนการบริการ กลับพบว่า เทคนิค Decision Tree มีความ แม่นยำถึง 76.22 เปอร์เซ็นต์ และง่ายต่อการเข้าใจของพนักงานโรงแรม

จุฑาทิพย์ ทิพย์พูล [25] ได้นำเสนอเทคนิคการทำเหมืองข้อความ 3 เทคนิค ซึ่งประกอบด้วยเทคนิคต้นไม้ตัดสินใจ เทคนิค Naïve Bayes และเทคนิค k-NN โดยทำการเก็บรวบรวมข้อมูลจดหมายทั้งหมด 5,172 ข้อความ ซึ่งผลการทดลองพบว่าเทคนิควิธีของ Naïve Bayes ให้ค่าความถูกต้องมากที่สุด

นุชนาฎ ปิ่นเมือง และ จารี ทองคำ [8] ได้ทำการวิจัยเกี่ยวกับความคิดเห็นของคนไทยต่อสื่อออนไลน์ด้วยเทคนิค 5 เทคนิค อันได้แก่ เทคนิค Naïve Bayes, SVM, KNN, decision tree และ C4.5 พบว่า เทคนิค Naïve Bayes มีประสิทธิภาพในการจำแนกความคิดเห็น โดยมีค่าความถูกต้องมากถึงร้อยละ 93.88 และค่าความแม่นยำร้อยละ 94.02

สมศักดิ์ ศรีสวการย์และสมัย ศรีสวย [26] ได้ทำการวิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ โดยนำข้อมูลจากบทวิจารณ์ออนไลน์ผ่านเครือข่ายเฟซบุ๊กสาธารณะของ

นักท่องเที่ยวมาสกัดคำแยกความคิดเห็นเชิงบวก เชิงลบ และได้ทำเปรียบเทียบประสิทธิภาพจากค่าความถูกต้องด้วยอัลกอริทึมนาอีฟเบย์ อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุด และอัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจผลการศึกษา พบว่า อัลกอริทึมนาอีฟเบย์ให้ค่าความถูกต้อง 87.97% อัลกอริทึมการหาเพื่อนบ้านใกล้ที่สุดให้ค่าความถูกต้อง 83.80% และ อัลกอริทึมการเรียนรู้ของต้นไม้ตัดสินใจให้ค่าความถูกต้อง 79.89%

ประพัฒน์ พรมน้ำอ่าง [12] ได้นำเสนอการจำแนกกลุ่มข้อความ โดยการใช้เทคนิคเหมืองข้อมูล ซึ่งประกอบด้วยเทคนิค SVM เทคนิคต้นไม้ตัดสินใจ เทคนิค K-NN และ Naïve Bayes โดยทำการเก็บข้อมูลจากข้อความรีวิวทั้งหมด 12,500 ข้อความ และแบ่งออกมาเป็นคำต่าง ๆ ตามคุณลักษณะเชิงบวกและเชิงลบจำนวนทั้งหมด 1,433 คำ ผลการทดลองพบว่าเทคนิค SVM ได้ให้ค่าความถูกต้องมากที่สุดที่ 86.26%

สุวิมล วงศ์สิงทอง [27] ได้ใช้วิธีจัดกลุ่มของการแชร์ข้อมูลบนเฟซบุ๊คของนักศึกษาทั้งหมด 700 คน ซึ่งได้ใช้วิธีเปรียบเทียบการจัดกลุ่มโดยการใช้ K-NN การจำแนกแบบต้นไม้ การเรียนรู้แบบเบย์ และ SVM ซึ่งผลการวิจัยพบว่าวิธีการของ SVM ให้ค่าความถูกต้องมากที่สุดที่ 87.95%

อดิเทพ ไชยสาร [28] ได้นำเสนอการเปรียบเทียบการประมาณอารมณ์จากความคิดเห็นภาษาไทย โดยใช้ 3 เทคนิค คือ เทคนิค Naïve Bayes เทคนิค SVM และเทคนิคต้นไม้ตัดสินใจ ซึ่งได้รวบรวมความคิดเห็นจากเว็บไซต์บริการข่าววาไรตี้บันเทิงและบริการวิจารณ์สินค้า จำนวน 6,000 ความคิดเห็น โดยได้แบ่งกลุ่มความคิดเห็นออกเป็น 6 กลุ่มอารมณ์ โดยวิธี SVM นั้นสามารถประมาณอารมณ์ได้ถูกต้องที่สุด 69.15% เมื่อเทียบกับเทคนิค Naïve Bayes และเทคนิคต้นไม้ตัดสินใจ

ราชวิทย์ ทิพย์เสนา และคณะ [13] ได้ทำการวิจัยเกี่ยวกับการจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความ โดยใช้เทคนิคในการจำแนกข้อความ (Text Classification) เป็นกระบวนการในการทำเหมืองข้อความ (Text Mining) มาช่วยในการจำแนกหมวดหมู่ของคำถามแบบอัตโนมัติบนกระดานสนทนา โดยจะพิจารณาจากคุณลักษณะของคำในแต่ละคำถามในการจำแนกหมวดหมู่ เพื่อแยกประเภทและความแตกต่างของคำถาม ทำให้ผู้ตอบสามารถตอบคำถามได้ตรงกับการให้บริการของหน่วยงาน และผู้ตั้งคำถามได้รับคำตอบที่มีความถูกต้องชัดเจน เทคนิคที่ใช้มี 3 เทคนิควิธี คือ เทคนิคการหาเพื่อนบ้านใกล้ที่สุด เทคนิคต้นไม้ตัดสินใจและเทคนิคการเรียนรู้เบย์อย่างง่าย ผลการเปรียบเทียบประสิทธิภาพแสดงให้เห็นว่า เทคนิคการหาเพื่อนบ้านใกล้ที่สุดให้ประสิทธิภาพในการจำแนกดีที่สุด โดยค่าความถูกต้องเท่ากับ 0.89 ค่าความเที่ยง เท่ากับ 0.9 ค่าความระลึก เท่ากับ 0.89 และค่า F-Measure เท่ากับ 0.892

วสวัตติ์ อินทร์แปลง และจารี ทองคำ [14] ได้ทำการวิเคราะห์ความคิดเห็นต่อเกมมือถือพับจีด้วยเหมืองข้อความ โดยเป็นกระบวนการวิเคราะห์ข้อมูลตัวอักษรเพื่อสกัดข้อมูลที่เป็นประโยชน์จากแหล่งข้อมูล เก็บรวบรวมข้อมูลความคิดเห็นต่อเกมมือถือผับจีจำ นวน 3,798 ข้อความ บ่งชี้เพื่อใช้ในการแยกคุณลักษณะได้เลือกใช้คำ วิเศษณ์ และคำ สแลงบางคำ ที่ความหมายของคำ เป็นคำ วิเศษณ์เพื่อทำ การแยกคุณลักษณะเชิงบวกและเชิงลบ ใช้การแก้ปัญหาโดยการปรับความสมดุลของข้อมูลด้วยวิธี SMOTE (Synthetic Minority Over-sampling Technique) และใช้หลักการ 10-fold cross validation จาก 5 เทคนิค คือ เทคนิค Naïve Bayes เทคนิค Support Vector Machine (SVM) เทคนิค K-Nearest Neighbor เทคนิคต้นไม้ตัดสินใจ C4.5 และเทคนิค Random Forest จากการทดสอบและวัดประสิทธิภาพของโมเดลพบว่า เทคนิค K-Nearest Neighbor ให้ผลดีที่สุดโดยให้ค่าความแม่นยำ 99.75% ค่าความระลึก 100% และค่าความถูกต้อง 99.87%

บรรหาร จันทะวงค์ และคณะ [29] ได้นำเสนอการจัดประเภทเอกสารงานวิจัยทางโลจิสติกส์เพื่อสร้างระบบช่วยในการทำนายประเภทของเอกสารงานวิจัย โดยใช้เทคนิคการทำเหมืองข้อมูลด้านการจัดประเภทข้อมูล ซึ่งได้เลือกวิธี k-Nearest Neighbors โดยจากการทดลองได้ค่าความแม่นยำสูงสุดในการทำนายที่ 99.00% โดยใช้วิธีการมาตรวัดระยะห่างแบบ Overlap Similarity

ภรัณยา ปาลวิสุทธิ์ [30] ได้ศึกษาการเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุล โดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต โดยใช้ เทคนิคต้นไม้ตัดสินใจ J48, ID3, LMT, CART และ RandomForest โดยใช้ 10-fold cross validation ในการแบ่งข้อมูล และกลุ่มตัวอย่างที่ใช้ในงานวิจัยเป็นกลุ่มเยาวชน ซึ่งมีอายุระหว่าง 15 - 24 ปี ในเขตอำเภอเมือง จังหวัดนครปฐม จากการเก็บรวบรวมข้อมูล ระหว่างวันที่ 3 - 25 สิงหาคม พ.ศ. 2558 และการเก็บรวบรวมข้อมูลได้จากการสำรวจโดยการใช้แบบประเมินอาการติดอินเทอร์เน็ต (Internet Addiction Test) ของศูนย์โรคติดอินเทอร์เน็ต พบว่าตัวแบบการพยากรณ์จากเทคนิค Random Forest เป็นตัวแบบที่มีประสิทธิภาพในการพยากรณ์สูงที่สุด โดยมีค่าความแม่นยำเท่ากับร้อยละ 87.15 ค่าความไวเท่ากับร้อยละ 85.89 และค่าความจำเพาะเท่ากับร้อยละ 87.53

วัชรีวรรณ จิตต์สกุล [24] ได้ศึกษาการวิเคราะห์การจำแนกข้อความ เพื่อศึกษาความเสถียรของอัลกอริทึมเพื่อการจำแนก 4 รูปแบบพื้นฐาน ได้แก่ ฐานกฎ เลือกอัลกอริทึม Conjunctive Rule ฐานต้นไม้ตัดสินใจ เลือกอัลกอริทึม Random Forest ฐานความน่าจะเป็น เลือกอัลกอริทึม Bayesian Logistic Regression และฐานการเรียนรู้ เลือก Support Vector Machine กับข้อความทดสอบนำมาจากฐานข้อมูลUCI Machine Learning Repository (http://archive.ics.uci.edu/ml) จำนวน 3 ชุด ได้แก่ ข้อมูลแสดงความคิดเห็นเกี่ยวกับภาพยนตร์จากเว็บไซต์ www.imdb.com ข้อมูลแสดงความคิดเห็นเกี่ยวกับร้านอาหารจากเว็บไซต์ www.yelp.com และ ข้อมูลแสดงความ

คิดเห็นเกี่ยวกับสินค้าจากเว็บไซต์ www.amazon.com ซึ่งข้อมูลทั้ง 3 ชุดเป็นลักษณะข้อความแสดงความคิดเห็น แบ่งออกเป็น 2 กลุ่ม คือ “1” แทนการแสดงความคิดเห็นเชิงบวก และ “0” แทนการแสดงความคิดเห็นในเชิงลบ ทำการปรับค่าพารามิเตอร์ตามที่กำหนด ใช้การทดสอบ 2 รูปแบบ คือ วิธี 10-fold cross validation และวิธี Test Dataset ผลลัพธ์ พบว่า อัลกอริทึมที่ให้ผลลัพธ์ดีที่สุดสองอันดับแรกของข้อมูลทั้ง 3 ชุด คือ อัลกอริทึม Random Forest และอัลกอริทึม Bayesian Logistic Regression ต่อมาทำการวิเคราะห์ ROC พบว่า อัลกอริทึมฐาน Random Forest แสดงความเสถียรการจำแนกข้อความได้สูงกว่าอัลกอริทึมที่ศึกษาอื่นๆ ด้วยค่าเฉลี่ย ROC > 0.80 และผลต่างของค่ากลางความถูกต้องเท่ากับ 0 เมื่อทดสอบระหว่างวิธี 10-fold cross validation และวิธี Test Dataset อย่างมีนัยสำคัญ ซึ่งงานวิจัยนี้สามารถนำไปประยุกต์ใช้ในเรื่องของการสนับสนุนการตัดสินใจ การให้คำแนะนำได้

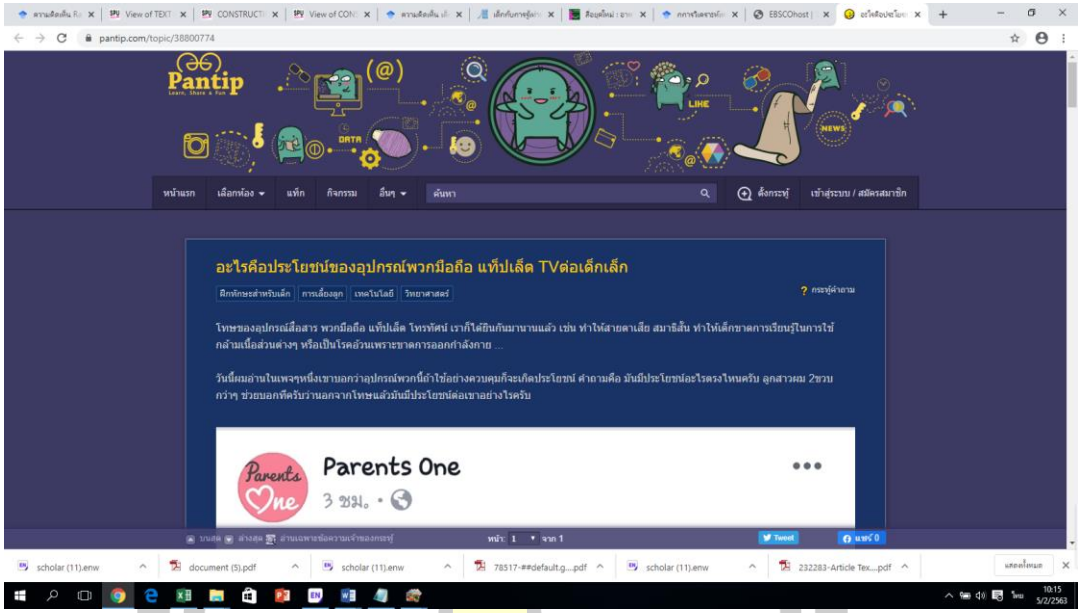
บทที่ 3
วิธีดำเนินการวิจัย

ในการดำเนินการวิจัยเพื่อให้ได้ตามวัตถุประสงค์ที่ตั้งไว้ กระบวนการในการทำเหมืองความคิดเห็นซึ่งมี 5 ขั้นตอนได้ถูกนำมาใช้โดยเริ่มจาก การรวบรวมข้อมูล การเตรียมข้อมูล การจัดทำดัชนีข้อมูล การสร้างแบบจำลองและการวัดประสิทธิภาพของแบบจำลอง โดยได้ทำการวิเคราะห์ด้วยกัน 2 แบบ คือ การให้ค่าน้ำหนักตามจำนวนคำก่อนการใช้ถุงคำ (bag of word) และหลังการใช้ถุงคำ (bag of word)

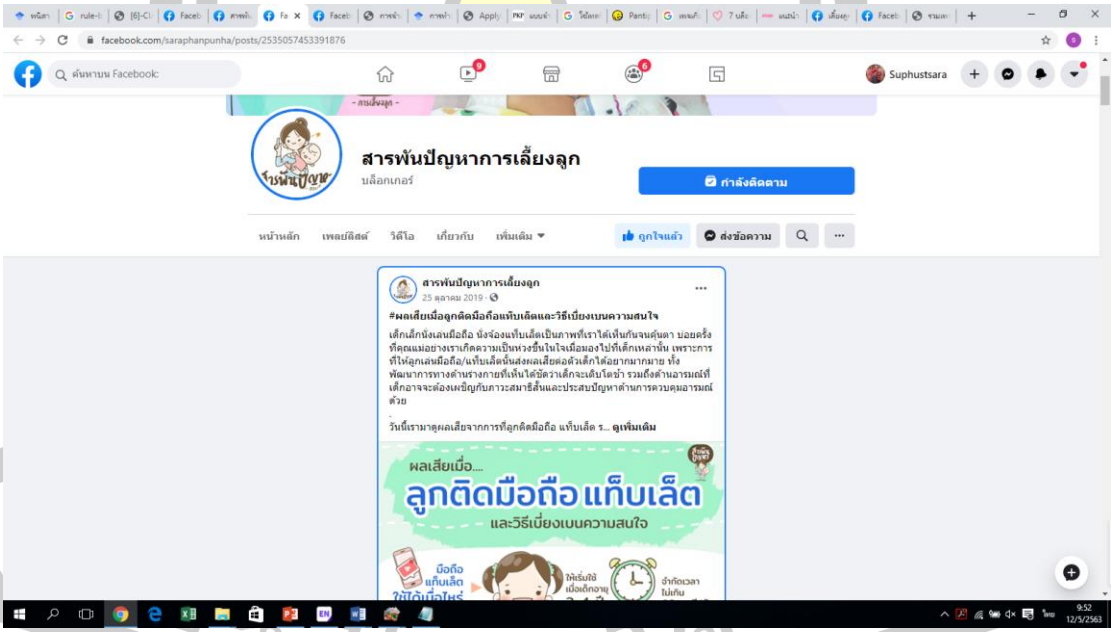
3.1 การรวบรวมข้อมูล

ในการเก็บรวบรวมข้อมูล ผู้วิจัยจะทำการเก็บรวบรวมข้อมูลความคิดเห็นเพื่อวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ตั้งแต่วันที่ 1 มกราคม 2561 ถึงวันที่ 31 ธันวาคม 2562 ทั้งนี้ความคิดเห็นที่ได้ทำการเก็บรวบรวมนั้นจะเป็นข้อความที่มีลักษณะเป็นภาษาไทยที่ได้มีการแสดงความคิดเห็นบนเครือข่ายสังคมออนไลน์ จากเว็ปไซต์ Pantip และ Facebook ซึ่งจะใช้ข้อความความคิดเห็นจากเพจหลักที่ได้รับความนิยมและน่าเชื่อถือที่ได้รับการจัดอับดับแนะนำ 10 เพจหมอเด็ก ให้ความรู้เรื่องการเลี้ยงลูก ที่พ่อกับแม่ต้องติดตาม เท่านั้น เช่น เพจเลี้ยงลูกนอกบ้าน, นายแพทย์ประเสริฐ ผลิตผลการพิมพ์, Dr.Pam book club, เลี้ยงลูกตามใจหมอ, เลี้ยงลูกให้เป็นคนปกติ, ชมรมจิตแพทย์เด็กและวัยรุ่นแห่งประเทศไทย, เข็นเด็กขึ้นภูเขา, สุธีรา เอื้อไพโรจน์กิจ, เลี้ยงลูกโตไปด้วยกันกับหมออร Hormone for Kids, หมอเสาวภา เลี้ยงลูกเชิงบวก และเพจสารพันปัญหาการเลี้ยงลูก โดยใช้คำสำคัญในการค้นหาเช่น เด็ก สมาร์ทโฟน, เด็กเล่นมือถือ, การเล่นมือถือของเด็ก, เด็กติดโทรศัพท์, ประโยชน์ มือถือ เป็นต้น

ซึ่งจำนวนของความคิดเห็นที่ได้เก็บรวบรวมนั้นมีจำนวนทั้งสิ้น 1,925 ความคิดเห็น ดังแสดงตัวอย่างในภาพที่ 3.1 และ ภาพที่ 3.2

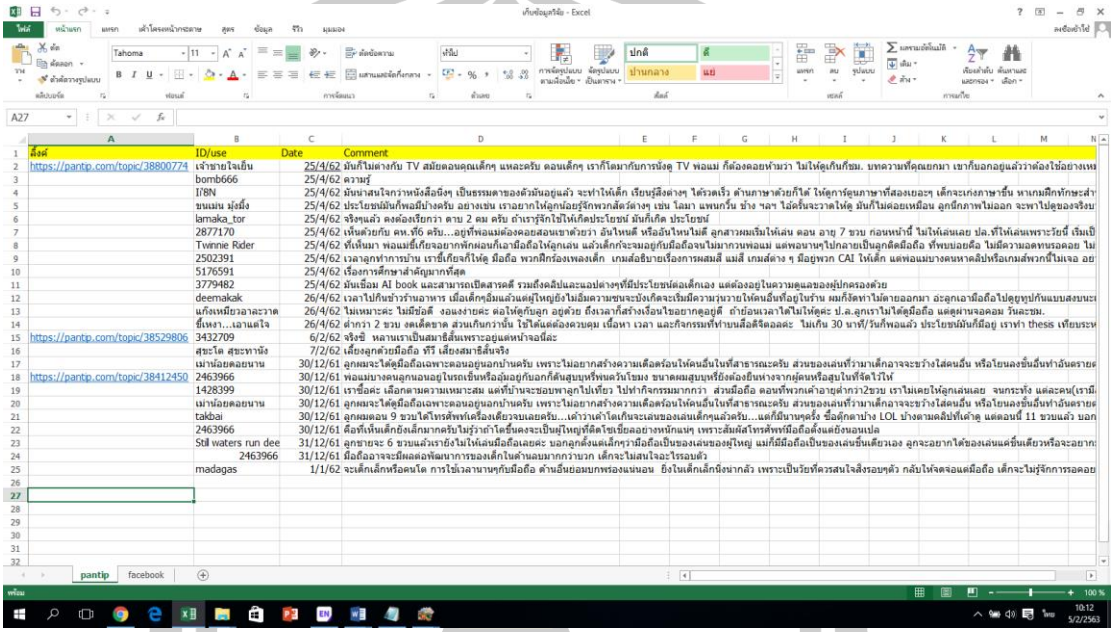


ภาพที่ 3.1 การเก็บข้อมูลจากเว็บไซต์ Pantip



ภาพที่ 3.2 การเก็บข้อมูลจากเว็บไซต์ Facebook

จากนั้นทำการดึงข้อมูลความคิดเห็น โดยในแต่ละเรคคอร์ดจะประกอบไปด้วยชื่อผู้ใช้ วันที่ และข้อความความคิดเห็นของผู้ใช้งาน ดังแสดงในภาพที่ 3.3



ชื่อผู้	ID/use	Date	Comment
https://pantip.com/topic/38800774	เจ้าชายใจเย็น	25/4/62	มันก็ไม่ต่างกับ TV สมัยตอนคุณเด็กๆ แหละครับ ตอนเด็กๆ เราก็โดนมากับการนั่งดู TV พ่อแม่ ก็ด่วงตอยห้ามว่า ไม่ให้ดูเกินกำหนด บทความที่คุณยกมา เขาก็บอกอยู่แล้วว่าต้องใช้อย่างเห
	bomb666	25/4/62	ความรู้
	IIBN	25/4/62	มันหาสนใจกว่าหนังสือซึ่งๆ เป็นธรรมดาของตัวมันอยู่แล้ว จะทำให้เด็ก เรียนรู้สิ่งต่างๆ ได้รวดเร็ว ตำนภาษาด้วยก็ได้ ให้ดูการ์ตูนภาษาที่สองเยอะๆ เด็กจะเก่งภาษาขึ้น หากเมมีหักษะสำ
	ขนมเม่น มังมึ่ง	25/4/62	ประโยชน์มันก็พอมีบ้างครับ อย่างเช่น เราอยากให้ลูกน้อยรู้จักพวกสัตว์ต่างๆ เช่น ไดโนา แพนกวิ้น ช้าง ฯลฯ ไอ้ครั้นจะวาดให้ดู มันก็ไม่ต่อยเหมือน ลูกนึกภาพไม่ออก จะพาไปดูของจริงบ
	lamaka_tor	25/4/62	จริงๆแล้ว ลงต้องเรียกว่า ตาม 2 ดม ครับ ถ้าเรารู้จักใช้ให้ก็ดีประโยชน์ มันก็เกิด ประโยชน์
	2877170	25/4/62	เห็นด้วยกับ คห.ที่6 ครับ...อยู่ที่พอแม่ต้องคอยสอนเขาด้วยว่า อันไหนดี หรืออันไหนไม่ดี ลูกสาวผมเริ่มให้เล่น ตอน อายุ 7 ขวบ ก่อนหน้านี้ ไม่ให้เล่นเลย ปล.ที่ให้เล่นเพราะวัยนี้ เริ่มเป็
	Twinnie Rider	25/4/62	ที่เห็นมา พอแม่ซื้อเกี่ยวอยากหักผ่อนก็เอามือถือให้ลูกเล่น แล้วเด็กก็จะจมอยู่กับมือถือจนไม่มากวนพ่อแม แต่พอนานๆไปกลายเป็นลูกติดมือถือ ที่พบบ่อยคือ ไม่มีความอดทนรอคอย ไม
	2502391	25/4/62	เวลาลูกทำการบ้าน เราซื้อเกี่ยวก็ให้ดู มือถือ พวกฝึกร้องเพลงเด็ก เกมส์อธิบายเรื่องการผสมสี แม่สี เกมส์ต่าง ๆ มีอยู่พวก CAI ให้เด็ก แต่พอแม่บางคนหาคลิปหรือเกมส์พวกนี้ไม่เจอ อย
	5176591	25/4/62	เรื่องการศึกษาสำคัญมากที่สุด
	3779482	25/4/62	มันเยื่อม AI book และสามารถเปิดสารคดี รวมถึงคลิปและแอปต่างๆที่มีประโยชน์ต่อเด็กเอง แต่ต้องอยู่ในความดูแลของผู้ปกครองด้วย
	deemakak	26/4/62	เวลาไปกินข้าวร้านอาหาร เมื่อเด็กๆอิ่มแล้วแต่ผู้ใหญ่ยังไม่อิ่มความชนจะบังเกิดจะเริ่มมีความมุ่งมายให้คนอื่นที่อยู่ในร้าน ผมก็รัดทำไม่ดายออกมา อะลูกเอามือถือไปดูทุปกันแบบสงบนะ
	แก้งเหมียวลาละวาด	26/4/62	ไม่เหมาะค่ะ ไม่มีข้อดี งอแง่งายค่ะ ต่อให้ดูกับลูก อยู่ด้วย ถึงเวลาก็สร้างเงื่อนไขอยากดูอยู่ดี ถ้าย่อนเวลาให้ไม่ให้ดูค่ะ ป.ล.ลูกเราไม่ได้ดูมือถือ แต่ดูผ่านจอคอม วันละชม.
	ขี้เหงา...เอาแต่ใจ	26/4/62	ต่ากว่า 2 ขวบ อดเด็ดขาด ส่วนเกินกว่านั้น ใช้ได้แต่ต้องควบคุม เนื้อหา เวลา และกิจกรรมที่ทำบนมือถือดีจิตอลค่ะ ไม่เกิน 30 นาที/วันก็พอแล้ว ประโยชน์มันก็มีอยู่ เราทำ thess เทียบระห
https://pantip.com/topic/38529806	3432709	6/2/62	จริงชี้ หลานเราเป็นสมาธิสั้นเพราะอยู่แต่หน้าจอนี่ละ
	สุขะโต สุขะทาทัง	7/2/62	เลี้ยงลูกด้วยมือถือ ที่วี เลี้ยงสมาธิสั้นจริง
	เม่าน้อยคอยนาน	30/12/61	ลูกผมจะได้ดูมือถือเฉพาะตอนอยู่นอกบ้านครับ เพราะไม่อยากสร้างความเคือดร้อนให้คนอื่นในที่สาธารณะครับ ส่วนของเล่นที่ว่ามาเด็กอาจจะขว้างใส่คนอื่น หรือโยนลงชั้นอื่นทำอันตรายร
https://pantip.com/topic/38412450	2463966	30/12/61	ผอมแบบนางคนลูกนอนอยู่ในกอเย็นหรือผู้หมอยู่กับอกที่ดันสุขมุหรือพนวันโยษง ขนาดผมสุขมุหรือยังต่องยืนห่างจากผู้คนหรือสุขในที่จัดไว้ให้
	1428399	30/12/61	เรายื่อค่ะ เลือกตามความเหมาะสม แต่ที่บ้านจะขอบพาลูกไปเที่ยว ไปทำกิจกรรมมากกว่า ส่วนมือถือ ดอนที่พวกค้าอายุต่ำกว่า2ขวบ เราไม่เคยให้ลูกเล่นเลย จนกระทั่ง แต่ละคน(เรามี
	เม่าน้อยคอยนาน	30/12/61	ลูกผมจะได้ดูมือถือเฉพาะตอนอยู่นอกบ้านครับ เพราะไม่อยากสร้างความเคือดร้อนให้คนอื่นในที่สาธารณะครับ ส่วนของเล่นที่ว่ามาเด็กอาจจะขว้างใส่คนอื่น หรือโยนลงชั้นอื่นทำอันตรายร
	takbai	30/12/61	ลูกผมตอน 9 ขวบได้โทรศัพท์เครื่องเดียวขนาดครับ...เด้ว่าเด้ว่าโดนเรื่องของเล่นของเล่นเด็กๆแล้วครับ...แต่ที่มีบางครั้ง ชื่อเด็กตาบ้าง LOL บ้างตามคลิปที่เด้ดู แต่ตอนนี้ 11 ขวบแล้ว บอก
	2463966	30/12/61	คือที่เห็นเด็กยิ่งเล็กมากครับไม่รู้ว่าด้วยโตขึ้นคงจะเป็นผู้ใหญ่ที่ติดโรเบียดอย่างหนักแน่ๆ เพราะสัมผัสโทรศัพท์มือถือตั้งแต่ยังนอนแปล
	Stil waters run dee	31/12/61	ลูกชายจะ 6 ขวบแล้วเรายังไม่ให้เล่นมือถือเลยค่ะ บอกลูกตั้งแต่เล็กๆว่ามือถือเป็นของเล่นของผู้ใหญ่ แม้ที่มีถือเป็นของเล่นชิ้นเดียวเอง ลูกจะอยากได้ของเล่นแค่ชิ้นเดียวหรือจะอยาก
	2463966	31/12/61	มือถือเราจะมีผลต่อพัฒนาการของเด็กในด้านอมมากกว่าบวก เด็กจะไม่สนใจอะไรรอบตัว
	madagas	1/1/62	จะเด็กเล็กหรือคนโต การใช้เวลานานๆกับมือถือ ด้านอื่นย่อมบกพร่องแน่นอน ยิ่งในเด็กเล็กนี่ยากลัว เพราะเป็นวัยที่ควรสนใจสิ่งรอบๆตัว กลับให้จดจ่อแต่มือถือ เด็กจะไม่รู้จักการรอคอย

ภาพที่ 3.3 ตัวอย่างข้อความคิดเห็นที่เก็บรวบรวมมาจัดเก็บในโปรแกรม Microsoft Excel

เมื่อทำการเก็บข้อมูลความคิดเห็นออกมาแล้ว นำมาจัดเก็บใช้โปรแกรม Microsoft Excel ในการรวบรวมข้อมูล หรือแก้ไขข้อมูล ต่าง ๆ ได้

3.2 การเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูลนี้เป็นกระบวนการที่ใช้ในการวิเคราะห์ข้อมูลเพื่อเตรียมข้อมูลก่อนเข้าสู่กระบวนการวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยจะนำความคิดเห็นที่ผ่านการตัดคำและจำกัดคำหยุด การคัดแยกข้อความที่เกี่ยวข้อง เพื่อแยกประเภทของความคิดเห็นที่เป็นคุณลักษณะเชิงบวกและเชิงลบ กระบวนการเตรียมข้อมูลนี้จะทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่จะนำมาใช้วิเคราะห์ว่ามีความถูกต้องและอยู่ในรูปแบบที่สามารถนำไปวิเคราะห์ได้ เนื่องจากข้อมูลที่ได้มาอาจจะมีการพิมพ์ผิด พิมพ์ตกหรือพิมพ์เป็นภาษาพูด ซึ่งในงานวิจัยนี้ได้มีขั้นตอนการเตรียมข้อมูลก่อนการจัดทำดัชนีคำ ดังนี้

3.2.1 การตัดคำและกำจัดคำหยุด

หลังจากผ่านกระบวนการเตรียมข้อมูลแล้ว จะเป็นขั้นตอนตัดคำและจำกัดคำหยุด การแบ่งตัวอักษรจากข้อความ เพื่อหาขอบเขตของแต่ละหน่วยคำ การตัดคำนี้จะส่งผลทำให้การจำแนกข้อความมีประสิทธิภาพมากยิ่งขึ้น จากนั้นทำการกำจัดคำหยุดต่าง ๆ ออกไป เช่นคำว่า ครับ ค่ะ เป็นต้น โดยจะไม่ทำให้ความหมายของข้อความนั้นเปลี่ยนไป ซึ่งในการตัดคำและกำจัดคำหยุดในงานวิจัยนี้ได้ใช้ภาษา Lexto ในการตัดคำภาษาไทยและระบุชนิดของคำ ซึ่งจะใช้วิธีการตัดคำโดยใช้โปรแกรมภาษา Python ไลบลารีของ PyThaiNLP และเพื่อความถูกต้องมากยิ่งขึ้นผู้วิจัยได้ให้ผู้เชี่ยวชาญเข้ามาช่วยในการส่วนของการระบุชนิดของคำและแยกประเภทของความคิดเห็นที่เป็นคุณลักษณะเชิงบวกและเชิงลบ หลังจากผ่านกระบวนการแล้วจะได้ดังตารางที่ 3.1

ตารางที่ 3.1 ตัวอย่างข้อความที่ถูกตัดคำ

	Comment Cut
1	คุณพ่อ,คุณแม่,ติดมือ,มั้ย,คะ,แนะนำ,กรณี,ลูกสาว,เรา,ที่,ทำ,แล้ว,ได้ผล,นะคะ,เรา,สามี,จะ,ไม่,เล่น,โทรศัพท์,เวลา,ที่อยู่,ลูก,ค่ะ,เรา,จะ,ตกลง,กัน,ว่า,จะ,ไม่,ติด,โทรศัพท์,ทุกคน,แล้ว,ให้,ลูก,ใช้,ตามกำหนด,เวลา,ลูก,เรา,เรา,ปล่อย,ให้,เล่น,ได้,ชม.,แล้ว,...
2	หลาน,เรา,ก็,ติด,เกมส์,แต่,ไม่,ก้าวร้าว,บอก,อะไร,ก็,เชื่อ,รับผิดชอบ,เรื่อง,เรียน,เพราะ,แม่,เค้า,ไม่,ตามใจ,ทั้งหมด,ให้,รับผิดชอบ,ส่วน,ของ,ตัวเอง,ดูแลตัวเอง,ให้,ทำ,อะไร,เอง,และ,พ่อ,ดุ,มาก,แต่,มี,ตผล,ไม่,ดุ,พร่ำเพรื่อ,ดุ,เมื่อ,ทำผิด,หนัก,เกรงใจ,...
3	เอา,ง่ายๆ,ผู้ใหญ่,ยัง,ติดงอมแงม,เด็ก,นี่,ยัง,ไม่,มี,วิจารณญาณ,เลย,ต้อง,คอย,คุม,ดี
4	ที่,เห็น,มา,พ่อแม่,ขี้เกียจ,อยาก,พักผ่อน,ก็,เอา,มือถือ,ให้,ลูกเล่น,แล้ว,เด็ก,จะ,จมอยู่กับ,มือถือ,จน,ไม่,มา,กวน,พ่อแม่,แต่,พอ,นานๆ,ไป,กลายเป็น,ลูกติด,มือถือ,ที่,พบ,บ่อย,คือ,ไม่,มี,ความอดทน,รอคอย,ไม่,มี,ระเบียบวินัย,ก้าวร้าว,สมาธิ,สั้น,ขาด,จินตนาการ,...
5	ขอ,เข้ามา,ตอบ,บ้าง,นะคะ,มี,ลูกสาว,ค่ะ,แต่,เรา,ไม่,ได้,ห้าม,เรื่อง,จอ,นะคะ,ส่วนมาก,จะ,เป็น,ทีวี,ที่,ดู,เกือบ,ทุกวัน,ประมาณ,วัน,ละ,นาที,แต่,ไม่,ตายตัว,ค่ะ,มือถือ,นี่,ให้,นั่ง,ดู,บ้าง,แต่,ค่อน,ข้างน้อย,ถึง,แทบ,ไม่,มี,ไม่,ใช่,อะไร,หรอก,ลูก,ใช้,ไม่,เป็น,ค่ะ,ลูก,....

3.3 การจัดทำดัชนีข้อมูล

การจัดทำดัชนีเป็นการจัดข้อมูลความคิดเห็นให้เหมาะสมและตรงกับรูปแบบที่จะประมวลผล เนื่องจากคอมพิวเตอร์ไม่สามารถที่จะเข้าใจภาษาธรรมชาติของมนุษย์ได้ ดังนั้นจึงต้องมีการแปลงเอกสารให้อยู่ในรูปแบบที่สามารถเข้าใจได้ เช่น การตัดบางคอลัมน์ที่ไม่จำเป็นออก การแทนคำหรือข้อความ เป็นต้น ในส่วนของการระบุประเภทของคำและการแทนคำคุณลักษณะที่สามารถแสดงถึงความคิดเห็นเชิงบวก เชิงลบ หรือเป็นกลางนั้น ผู้วิจัยได้ขอให้ผู้เชี่ยวชาญทางด้านภาษาไทยเข้ามาช่วยการตรวจสอบ เพื่อความถูกต้องของข้อมูลมากยิ่งขึ้น

3.3.1 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะ (Feature selection)หลังจากได้ทำการตัดคำและกำจัดคำหยุดแล้ว จะได้ถุงคำทั้งหมด 5,409 คำ ในกระบวนการนี้เป็นกระบวนการที่สำคัญเบื้องต้น คือการนำคำมาจัดประเภท ในงานวิจัยนี้ได้ใช้ภาษา Lexto ในการตัดคำภาษาไทยและระบุชนิดของคำ ซึ่งจะใช้วิธีการตัดคำโดยใช้โปรแกรมภาษา Python ไลบลารีของ PyThaiNLP ได้คัดเลือกคุณลักษณะคำที่เป็นคำวิเศษณ์ จำนวน 671 คำ ซึ่งเป็นประเภทของคำที่เหมาะที่จะนำมาใช้ในการจำแนกความคิดเห็นหลังจากผ่านกระบวนการตัดคำและการกำกับชนิดของคำ สามารถที่จะสกัดเอาคำจากความคิดเห็นทั้งหมดและทำการตัดคำที่ซ้ำกันออก โดยจะนำเอาคำวิเศษณ์มาใช้เพื่อการจำแนก เนื่องจากคำวิเศษนี้สามารถสื่อถึงการแสดงอารมณ์เชิงบวกและเชิงลบได้ดี แล้วทำการแบ่งคำบ่งชี้คุณลักษณะออกมา ดังแสดงในตารางที่ 3.2

ตารางที่ 3.2 ตัวอย่างคำคุณลักษณะที่มีการกำหนดประเภทคำตามความหมายในพจนานุกรม

No.	คำคุณลักษณะ	ประเภทคำ
1	กวน	คำกริยา
2	การ์ตูน	คำนาม
3	การบ้าน	คำนาม
4	การศึกษา	คำนาม
5	ก้าวร้าว	คำวิเศษณ์
6	บ่อย	คำวิเศษณ์
7	ดี	คำวิเศษณ์
8	ความรู้	คำนาม
9	ง่าย	คำวิเศษณ์

ตารางที่ 3.2 ตัวอย่างคำคุณลักษณะที่มีการกำหนดประเภทคำตามความหมายในพจนานุกรม

No.	คำคุณลักษณะ	ประเภทคำ
10	ฝึก	คำกริยา

คำวิเศษณ์ เป็นคำที่อธิบายหรือแทนความหมายของคำกริยาคำวิเศษณ์อื่นหรือประโยค สามารถแบ่งออกเป็น 4 หมวดหมู่ย่อย ดังนี้

คำวิเศษณ์ที่มีรูปแบบปกติ (ADVN) เป็นคำวิเศษณ์ที่ใช้ในรูปแบบฐานและไม่ได้อยู่ในรูปแบบของการทำซ้ำหรือไม่ได้มาจากคำกริยาโดยการเพิ่มคำนำหน้าคำวิเศษณ์ "โดย" หรือ "อย่าง" ตัวอย่างเช่นเก่ง (อย่างชาญฉลาด), เร็ว (เร็ว), ช้า (ช้า), หนา (เสมอ)

คำวิเศษณ์ที่มีรูปแบบซ้ำ (ADVI) เป็นคำวิเศษณ์ที่ใช้ในรูปแบบซ้ำโดยรวมกับ "ๆ" ตัวอย่างเช่น เร็ว ๆ (เร็ว ๆ), เสมอ ๆ (เสมอ), ช้าๆ (ช้า)

คำวิเศษณ์ที่มีแบบฟอร์มนำหน้า (ADVP) เป็นคำวิเศษณ์ซึ่งได้มาจากคำกริยาโดยการเพิ่มคำนำหน้าเช่น "โดย" หรือ "อย่าง" ตัวอย่างเช่นโดยเร็ว (เร็ว)

ประโยคคำวิเศษณ์ (ADVS) เป็นคำวิเศษณ์ที่แสดงทัศนคติของผู้พูดหรือประเมินสิ่งที่พูดในประโยคที่เหลือ มันมักจะอยู่ที่จุดเริ่มต้นของประโยค แต่บางส่วนก็ใช้ในตำแหน่งอื่น ตัวอย่างเช่น โดยปกติ, ธรรมดา

3.3.2 การแทนคำคุณลักษณะ

หลังจากจัดประเภทของคำเรียบร้อยแล้ว จากคำวิเศษณ์ทั้งหมด 671 คำ จะคัดเลือกคำวิเศษณ์ที่สามารถแสดงถึงความคิดเห็นเชิงบวกและเชิงลบเท่านั้น ได้จำนวนคำวิเศษณ์ 394 คำมาวิเคราะห์ โดยการแทนค่าของจำนวนคำคุณลักษณะที่ไม่ปรากฏเป็น 0 แทนค่าจำนวนคำคุณลักษณะเชิงบวกที่ปรากฏเป็น 1 ดังตารางที่ 3.3 และแทนค่าจำนวนคำคุณลักษณะเชิงลบที่ปรากฏเป็น -1 ดังตารางที่ 3.4

ตารางที่ 3.3 ตัวอย่างข้อมูลของคำคุณลักษณะเชิงบวก

ลำดับ	ดี	ง่าย	เก่ง	ค่าความถี่
1	0	1	0	1
2	0	0	0	0
3	0	0	1	1
4	0	0	0	0
5	0	0	0	0
6	1	0	0	2
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

ตารางที่ 3.4 ตัวอย่างข้อมูลของคำคุณลักษณะเชิงลบ

ลำดับ	ก้าวร้าว	เกินตัว	ไม่	ค่าความถี่
1	0	0	0	0
2	0	0	0	0
3	0	0	1	1
4	0	0	1	2
5	0	0	0	0
6	0	0	0	0
7	1	1	1	4
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0

หลังจากแทนค่าจำนวนคำคุณลักษณะเรียบร้อยแล้วจะเห็นความถี่ของคำคุณลักษณะที่ถูกจำแนกออกมาเป็นเชิงบวกและเชิงลบในแต่ละความคิดเห็น และทำการแบ่งความคิดเห็นเชิงบวกและเชิงลบ โดยใช้หลักเกณฑ์

ถ้าจำนวนความถี่คุณลักษณะเชิงบวกมากกว่า จะแทน ความคิดเห็นเชิงบวก ด้วย P

ถ้าจำนวนความถี่คุณลักษณะเชิงลบมากกว่า จะแทน ความคิดเห็นเชิงลบ ด้วย N

แต่ถ้าหากความถี่ของคุณลักษณะเชิงบวกและเชิงลบเท่ากัน จะแทน ด้วย B และข้อความนั้นจะถูกตัดออก ดังตารางที่ 3.5

ตารางที่ 3.5 ตัวอย่างการกำหนด Class ของข้อมูล

ลำดับ	ดี	ง่าย	เก่ง	ก้าวร้าว	เกินตัว	ไม่	...	P	N	Class
1	0	1	0	0	0	0		1	0	P
2	0	0	0	0	0	0		0	0	B
3	0	0	1	0	0	1		1	1	B
4	0	0	0	0	0	1		0	2	N
5	0	0	0	0	0	0		0	0	B
6	1	0	0	0	0	0		2	0	P
7	0	0	0	1	1	1		0	4	N
8	0	0	0	0	0	0		0	0	B
9	0	0	0	0	0	0		0	0	B
10	0	0	0	0	0	0		0	0	B

ในงานวิจัยนี้การใช้งานค่าน้ำหนักแบ่งเป็น 2 แบบ คือ การให้ค่าน้ำหนักตามจำนวนคำที่พบและการใช้ถุงคำ (bag of word)

โดยในส่วนของการวิเคราะห์หลังการใช้ bag of word ก่อนนำข้อมูลไปวิเคราะห์จะต้องทำการตัดข้อมูลในส่วนของคอลัมน์ Class P และ N ออกก่อนที่จะนำข้อมูลไปวิเคราะห์ จากนั้นทำการให้น้ำหนักของคำด้วยหลักการ bag of word โดยแทนค่า 1 คือ แทนคำที่กำหนดโดยเกิดขึ้นในเอกสาร และ 0 คือ แทนคำที่กำหนดโดยไม่เกิดในเอกสาร ดังตารางที่ 3.6

ตารางที่ 3.6 ตัวอย่างข้อมูลการใช้ถุงคำในการให้น้ำหนักคำ

ลำดับ	ดี	ง่าย	เก่ง	ก้าวร้าว	เกินตัว	ไม่	...	Class
1	0	1	0	0	0	0		P
2	0	0	1	0	0	0		P
3	0	0	0	0	0	1		P
4	0	0	0	0	0	0		P
5	1	0	0	0	0	0		P
6	0	0	0	1	0	1		N
7	0	0	0	0	0	1		N
8	0	0	0	0	0	1		N
9	0	0	0	0	0	1		N
10	0	0	0	0	0	0		N

3.4 การสร้างแบบจำลอง

ในกระบวนการสร้างแบบจำลองเพื่อหาแบบจำลองที่เหมาะสมที่สุดในการวิเคราะห์ความคิดเห็นเพื่อวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ใช้ข้อมูลความคิดเห็นที่เป็นภาษาไทย โดยได้นำเทคนิคการจัดกลุ่มมาทำการจัดกลุ่มข้อมูลและใช้เทคนิคการทำเหมืองความคิดเห็นมาใช้ในการวิเคราะห์ทั้งหมด 6 เทคนิค ยิ่งไปกว่านั้นผู้วิจัยได้ใช้ค่าพารามิเตอร์ของแต่ละแบบจำลองที่ทำให้แต่ละเทคนิคมีประสิทธิภาพสูงสุด ดังต่อไปนี้

1. เทคนิค RIPPER ในงานวิจัยได้มีการกำหนดค่าพารามิเตอร์ ดังนี้
   - ค่า minNo = 2
   - ค่า numDecimalPlace = 2
   - ค่า optimizations = 2
2. เทคนิค Decision tree C4.5 ในงานวิจัยได้มีการกำหนดค่าพารามิเตอร์ ดังนี้
   - ค่า confidenceFactor = 0.25
   - ค่า minNumObj = 2
   - ค่า numDecimalPlace = 2
3. เทคนิค Naïve Bayes ในงานวิจัยได้มีการกำหนดค่าพารามิเตอร์ ดังนี้
   - ค่า confidenceFactor = *

- ค่า minNumObj = *,
- ค่า numDecimalPlace = 2

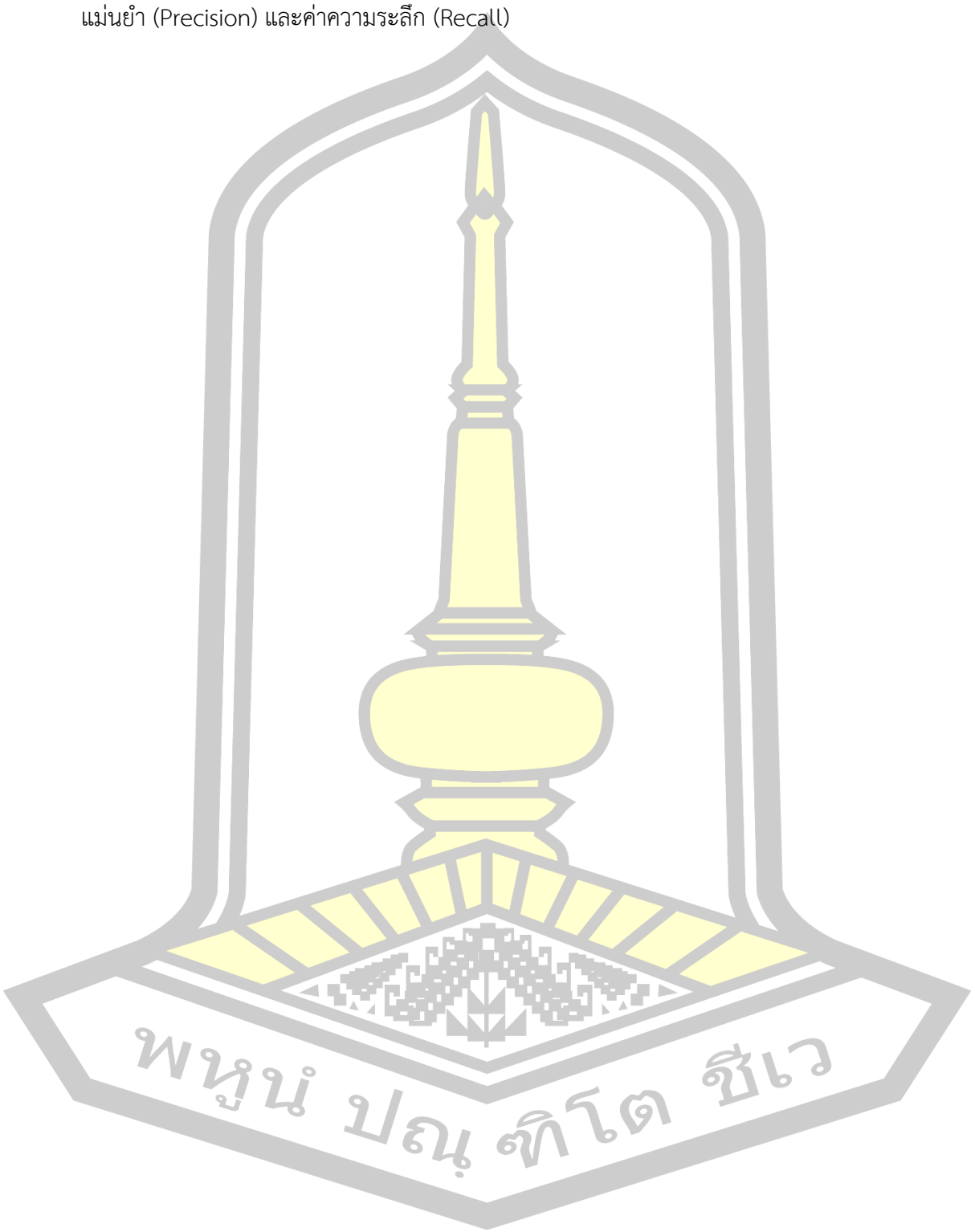
4. เทคนิค SVM ในงานวิจัยได้มีการกำหนดค่าพารามิเตอร์ ดังนี้
   - ค่า SVMType: nu-SVC (classification)
   - ค่า kernel type: Radial basis function
   - ค่า normalize: True
5. เทคนิค K-NN ในงานวิจัยได้มีการกำหนดค่าพารามิเตอร์ ดังนี้
   - ค่า k: 3
   - ค่า numDecimalPlace = 2
6. เทคนิค Random forest ในงานวิจัยได้มีการกำหนดค่าพารามิเตอร์ ดังนี้
   - ค่า confidenceFactor = *
   - ค่า minNumObj = *
   - ค่า numDecimalPlace = 2

ในงานวิจัยนี้ผู้วิจัยได้ใช้แบบจำลองในการสร้างแบบจำลองเพื่อวัดประสิทธิภาพของการทำงานแบบจำลองเพื่อทำการเปรียบเทียบเทคนิคที่ใช้ในการวิเคราะห์ความคิดเห็นเพื่อวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ในงานวิจัยนี้ผู้วิจัยได้ใช้โปรแกรม WEKA ในการสร้างแบบจำลองเพื่อวัดประสิทธิภาพของเทคนิคทั้ง 6

3.5 การวัดประสิทธิภาพของแบบจำลอง

เป็นขั้นตอนการแปลความหมาย การตีความและการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งมีการนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย และงานวิจัยนี้ได้สร้างแบบจำลองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง จึงได้ทำการวิเคราะห์ประสิทธิภาพการทำงานของแบบจำลอง โดยใช้เทคนิคการวัดประสิทธิภาพแบบ 10-fold cross validation โดยทำการเลือกสุ่มข้อมูลออกเป็น 10 ชุดเท่าๆกัน จากนั้นในการทดลองครั้งแรกจะใช้ข้อมูลชุดที่ 1 เป็นชุดทดสอบ (Test data) ส่วนชุดข้อมูลที่เหลืออีก 9 ชุดนั้นจะเป็นข้อมูลชุดสอน (Training data) ในการทดลองครั้งที่ 2 จะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบ (Test data) และให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน (Training data) ทำจนกระทั้งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งคิดเป็นอัตราข้อมูลทดสอบต่อปริมาณข้อมูลฝึก คิดเป็นอัตราร้อยละ 10:90 โดยผลลัพธ์ที่ได้นั้นจะ

นำมาคำนวณหาค่าประสิทธิภาพโดยรวมของแบบจำลอง คือ ค่าความถ่วงดุล (F-Measure) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

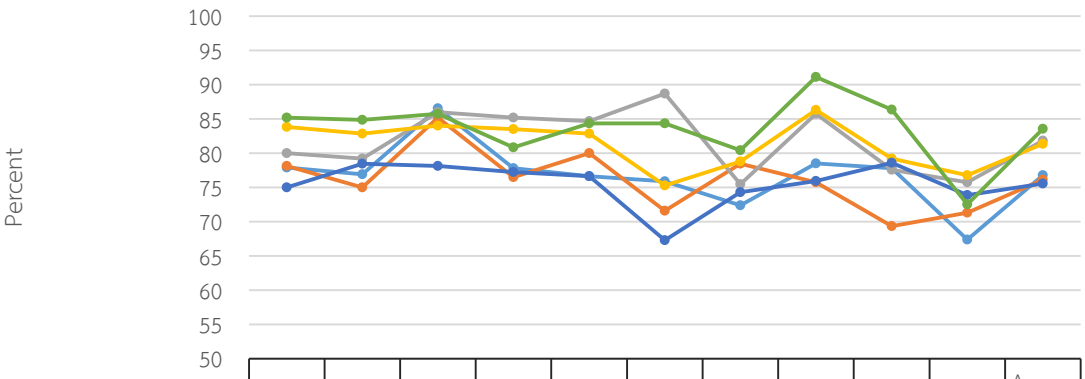


บทที่ 4
ผลการวิจัย

งานวิจัยนี้ได้ทำการวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยในบทนี้เป็นการอภิปรายผลการวิจัยตามวัตถุประสงค์ที่ตั้งไว้ มีจุดมุ่งหมายที่สำคัญ คือ ศึกษากระบวนการในการทำเหมืองความคิดเห็นในรูปแบบภาษาไทย ด้วยเทคนิคการทำเหมืองข้อความเพื่อทำการเปรียบเทียบในการทำเหมืองข้อความที่ดีที่สุด โดยได้เลือกใช้เทคนิค 6 เทคนิค ได้แก่ เทคนิคริปเปอร์ (RIPPER) เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 (Decision tree C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคซัพพอร์ตเวคเตอร์แมชชีน (SVM Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-NN) และเทคนิคการสุ่มป่าไม้ (Random forest) ข้อมูลที่ใช้ในงานวิจัยนี้เก็บรวบรวมจากการแสดงความคิดเห็นบนเครือข่ายสังคมออนไลน์ จากเว็ปไซต์ Pantip และ Facebook ตั้งแต่วันที่ 1 มกราคม 2561 ถึงวันที่ 31 ธันวาคม 2562 จำนวน 1,925 ข้อความคิดเห็น โดยในงานวิจัยนี้จะเลือกใช้เฉพาะคำวิเศษณ์ที่มีคุณลักษณะเชิงบวกและเชิงลบเท่านั้นมาวิเคราะห์ ซึ่งมีจำนวน 394 คำ ซึ่งหลังจากผ่านกระบวนการต่างๆ แล้วจะเหลือชุดข้อมูลที่ใช้ในกระบวนการสร้างแบบจำลอง จำนวน 1,925 ข้อความ โดยใช้ 10-Fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูลทดสอบ โดยได้ทำการวิเคราะห์ด้วยกัน 2 แบบ คือ การให้ค่าน้ำหนักตามจำนวนคำที่พบและการใช้ถุงคำ (bag of word) ในการให้น้ำหนักคำ

4.1 ผลการวิจัยจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ

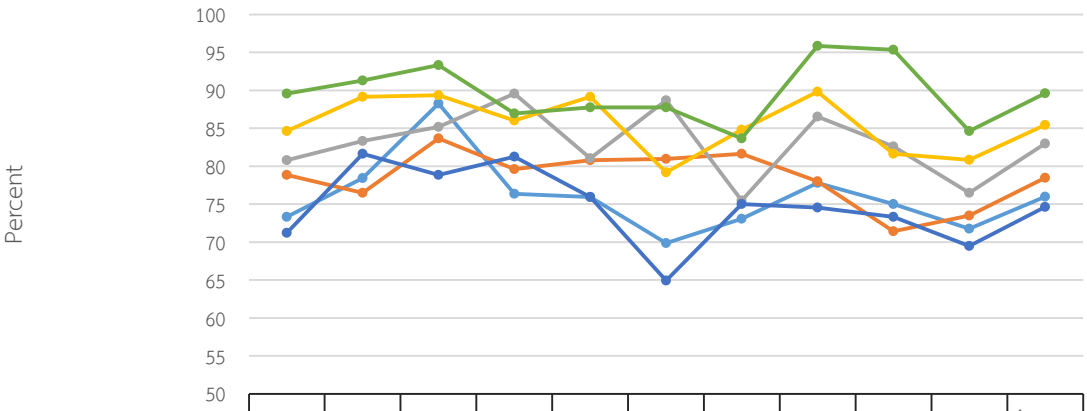
ผลการวิจัยจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบมาใช้ในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยใช้เทคนิค 6 เทคนิค ได้แก่ เทคนิคริปเปอร์ (RIPPER) เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 (Decision tree C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคซัพพอร์ตเวคเตอร์แมชชีน (SVM Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-NN) และเทคนิคการสุ่มป่าไม้ (Random forest) ในการประเมินความสามารถของแบบจำลองใช้ค่าความถ่วงดุล ค่าความแม่นยำและค่าความระลึกมาเป็นเกณฑ์ในการเปรียบเทียบสามารถแสดงได้ดัง ภาพที่ 4.1 ภาพที่ 4.2 และภาพที่ 4.3 ตามลำดับ



	1	2	3	4	5	6	7	8	9	10	Average
RIPPER	77.88	76.92	86.54	77.78	76.64	75.86	72.38	78.50	77.78	67.35	76.76
Decision tree C4.5	78.10	75.00	85.19	76.47	80.00	71.58	78.43	75.73	69.31	71.29	76.11
Naïve Bayes	80.00	79.21	85.98	85.15	84.68	88.68	75.47	85.71	77.55	75.73	81.82
SVM	83.81	82.83	84.00	83.50	82.83	75.25	78.79	86.27	79.21	76.77	81.33
KNN	75.00	78.43	78.10	77.23	76.64	67.27	74.29	75.93	78.57	73.87	75.53
Random Forest	85.15	84.85	85.71	80.81	84.31	84.31	80.39	91.09	86.32	72.53	83.55

ภาพที่ 4.1 ค่าความถ่วงดุลของแบบจำลองจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ

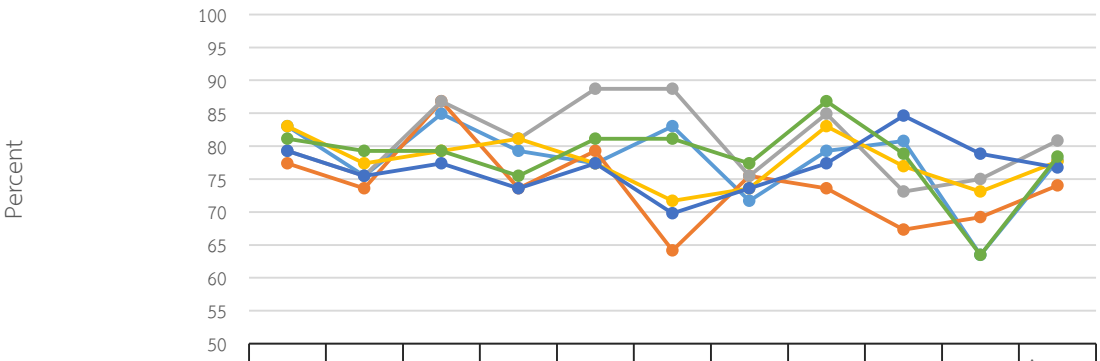
จากภาพที่ 4.1 แสดงการเปรียบเทียบค่าความถ่วงดุลของแบบจำลองจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ ของเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ผลปรากฏว่า เทคนิค Random Forest ที่ให้ค่าความถ่วงดุลสูงสุด เมื่อพิจารณาตามรอบแล้วจะเห็นได้ว่าในรอบที่ 1-7 มีค่าเฉลี่ยใกล้เคียงกัน อยู่ที่ระหว่าง 80-85% และมาพุ่งสูงขึ้นที่สุดในรอบที่ 8 ที่ 83.55% แล้วค่อยๆลดลงมาในรอบที่ 9 ที่ 86.32% จนมาในรอบสุดท้ายรอบที่ 10 ค่าได้ลดลงมาต่ำสุดกว่าทุกรอบ จะเห็นได้ว่าเทคนิค Random Forest ให้ค่าความถ่วงดุลเฉลี่ยสูงสุดที่ 83.55% รองลงมาคือเทคนิค Naïve Bayes ที่ 81.82% ส่วนเทคนิคที่ให้ค่าค่าความถ่วงดุลน้อยที่สุด คือ เทคนิค K-NN ที่ 75.53%



	1	2	3	4	5	6	7	8	9	10	Average
RIPPER	73.33	78.43	88.24	76.36	75.93	69.84	73.08	77.78	75.00	71.74	75.97
Decision tree C4.5	78.85	76.47	83.64	79.59	80.77	80.95	81.63	78.00	71.43	73.47	78.48
Naïve Bayes	80.77	83.33	85.19	89.58	81.03	88.68	75.47	86.54	82.61	76.47	82.97
SVM	84.62	89.13	89.36	86.00	89.13	79.17	84.78	89.80	81.63	80.85	85.45
KNN	71.19	81.63	78.85	81.25	75.93	64.91	75.00	74.55	73.33	69.49	74.61
Random Forest	89.58	91.30	93.33	86.96	87.76	87.76	83.67	95.83	95.35	84.62	89.62

ภาพที่ 4.2 ค่าความแม่นยำของแบบจำลองจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ

จากภาพที่ 4.2 แสดงการเปรียบเทียบค่าความแม่นยำของแบบจำลองจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ ของเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ผลปรากฏว่า เทคนิค Random Forest ให้ค่าความแม่นยำสูงสุดที่ 89.62% เมื่อพิจารณาตามรอบแล้วจะเห็นได้ว่าในรอบที่ 1-6 มีค่าเฉลี่ยใกล้เคียงกัน อยู่ที่ระหว่าง 87–93 % และในรอบที่ 7 ได้ลดลงต่ำสุดกว่าทุกรอบที่ 83.67% หลังจากลดลงต่ำสุดแล้วได้พุ่งสูงขึ้นที่สุดในรอบที่ 8 ที่ 95.83% แล้วค่อยๆลดลงมาในรอบที่ 9 ที่ 95.35% จนมาในรอบสุดท้ายรอบที่ 10 ค่าได้ลดลงมาใกล้เคียงกับรอบที่ 1-6 จะเห็นได้ว่าเทคนิค Random Forest ที่ให้ค่าความแม่นยำเฉลี่ยสูงสุดที่ 89.62% รองลงมาคือเทคนิค SVM ที่ 85.45% ส่วนเทคนิคที่ให้ค่าความแม่นยำเฉลี่ยน้อยที่สุดคือ เทคนิค K-NN ที่ 74.61%



	1	2	3	4	5	6	7	8	9	10	Average
RIPPER	83.02	75.47	84.91	79.25	77.36	83.02	71.70	79.25	80.77	63.46	77.82
Decision tree C4.5	77.36	73.58	86.79	73.58	79.25	64.15	75.47	73.58	67.31	69.23	74.03
Naïve Bayes	79.25	75.47	86.79	81.13	88.68	88.68	75.47	84.91	73.08	75.00	80.85
SVM	83.02	77.36	79.25	81.13	77.36	71.70	73.58	83.02	76.92	73.08	77.64
KNN	79.25	75.47	77.36	73.58	77.36	69.81	73.58	77.36	84.62	78.85	76.72
Random Forest	81.13	79.25	79.25	75.47	81.13	81.13	77.36	86.79	78.85	63.46	78.38

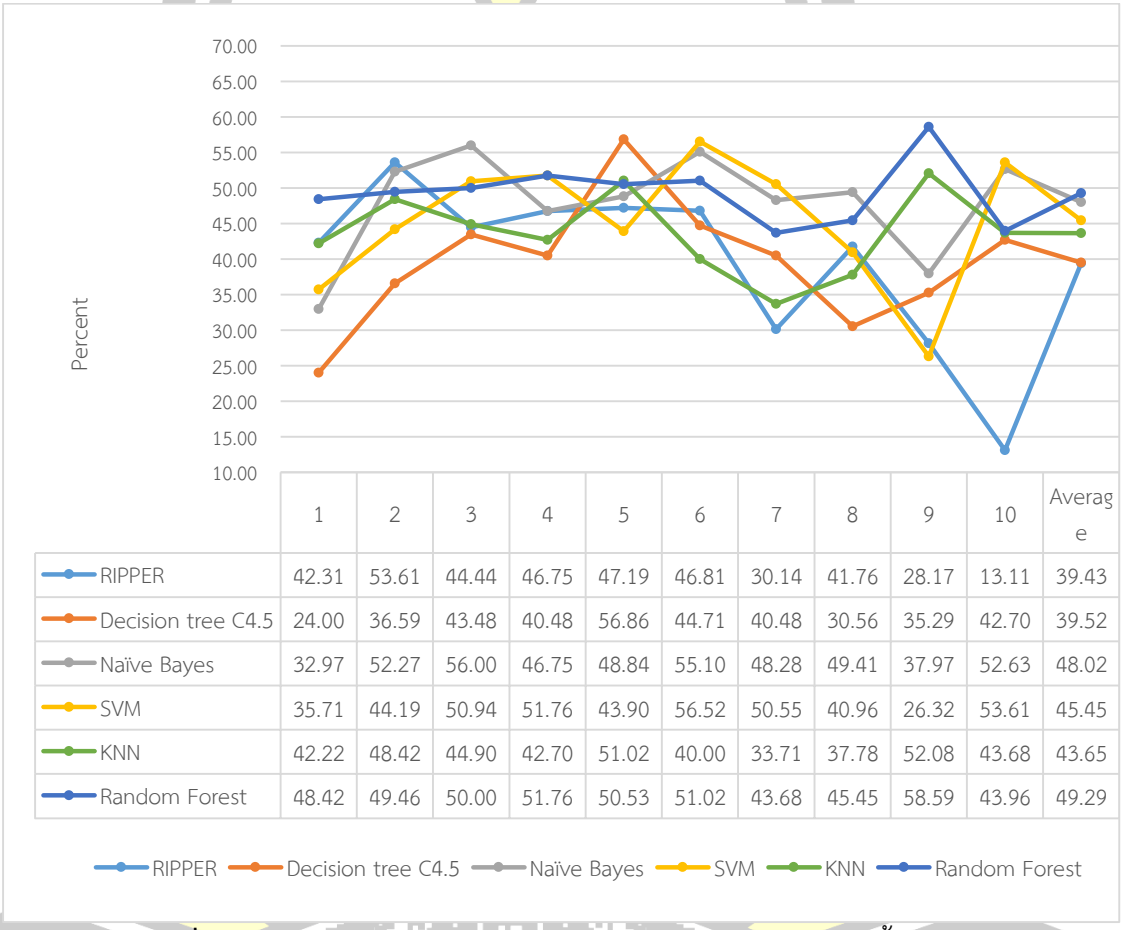
ภาพที่ 4.3 ค่าความระลึกของแบบจำลองจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ

จากภาพที่ 4.3 แสดงการเปรียบเทียบค่าความระลึกของแบบจำลองจากการให้ค่าน้ำหนักคำตามจำนวนคำที่พบ ของเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ผลปรากฏว่า เทคนิค Naïve Bayes ให้ค่าความระลึกเฉลี่ยสูงสุดที่ 80.85% เมื่อพิจารณาตามรอบแล้วจะเห็นได้ว่าในรอบที่ 1 และรอบที่ 2 มีค่าเฉลี่ยอยู่ที่ 75-79% และในรอบที่ 3 ได้สูงขึ้นมาจนเกือบสูงสุดที่ 86.79% แล้วลดลงในรอบที่ 4 และมาในรอบที่ 6-7 ได้พุ่งสูงขึ้นที่สุดที่ 88.68% ทั้ง 2 รอบ หลังจากพุ่งขึ้นสูงสุดแล้วก็ค่อยๆลดลงมาในรอบที่ 8-10 จะเห็นได้ว่าเทคนิค Naïve Bayes ให้ค่าความระลึกเฉลี่ยสูงสุดที่ 80.85% รองลงมาคือเทคนิค Random Forest ที่ 78.38% ส่วนเทคนิคที่ให้ค่าความระลึกน้อยที่สุด คือ เทคนิค Decision tree C4.5 ที่ 74.03%

4.2 ผลการวิจัยจากการใช้ถุงคำในการให้น้ำหนักคำ

ผลการวิจัยจากการใช้ถุงคำในการให้น้ำหนักคำมาใช้ในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยใช้เทคนิค 6 เทคนิค ได้แก่ เทคนิคริปเปอร์

(RIPPER) เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 (Decision tree C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคซัพพอร์ตเวคเตอร์แมชชีน (SVM Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-NN) และเทคนิคการสุ่มป่าไม้ (Random forest) ในการประเมินความสามารถของแบบจำลองใช้ค่าความถ่วงดุล ค่าความแม่นยำและค่าความระลึกมาเป็นเกณฑ์ในการเปรียบเทียบสามารถแสดงได้ดังภาพที่ 4.4 ภาพที่ 4.5 และภาพที่ 4.6 ตามลำดับ

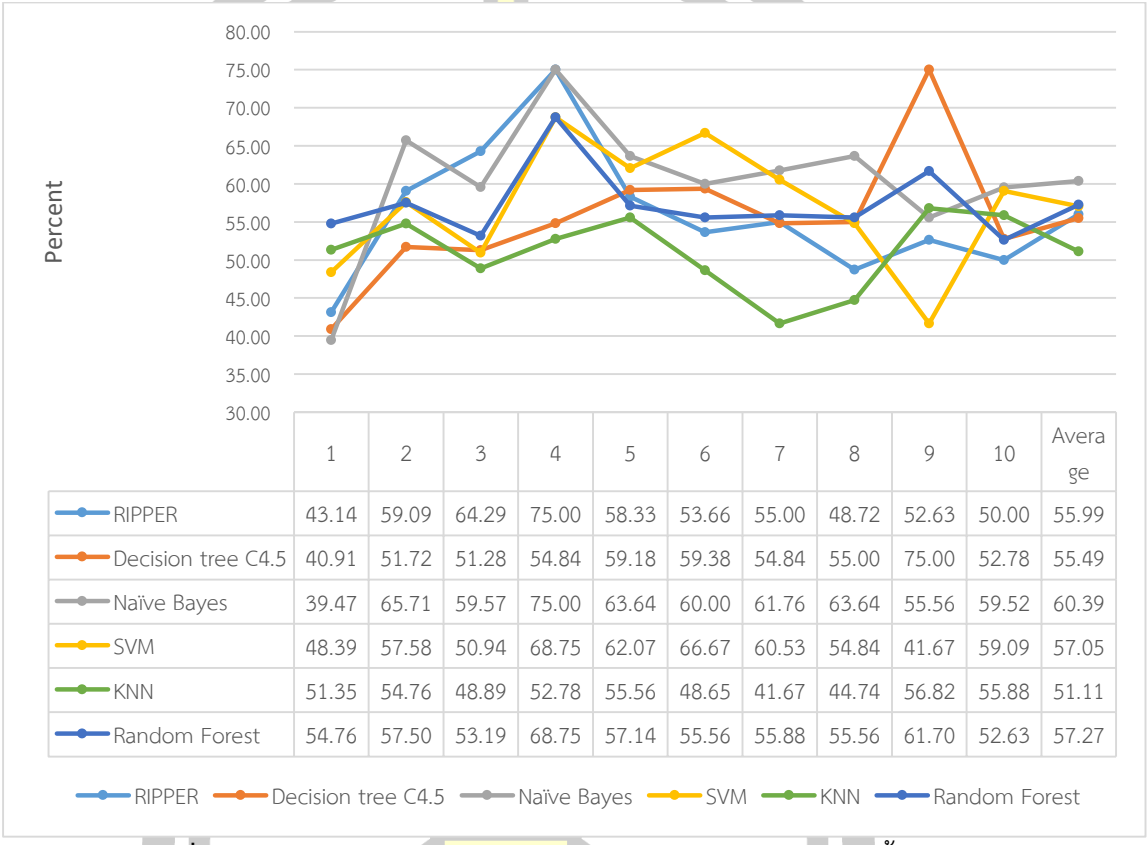


	1	2	3	4	5	6	7	8	9	10	Average
RIPPER	42.31	53.61	44.44	46.75	47.19	46.81	30.14	41.76	28.17	13.11	39.43
Decision tree C4.5	24.00	36.59	43.48	40.48	56.86	44.71	40.48	30.56	35.29	42.70	39.52
Naïve Bayes	32.97	52.27	56.00	46.75	48.84	55.10	48.28	49.41	37.97	52.63	48.02
SVM	35.71	44.19	50.94	51.76	43.90	56.52	50.55	40.96	26.32	53.61	45.45
KNN	42.22	48.42	44.90	42.70	51.02	40.00	33.71	37.78	52.08	43.68	43.65
Random Forest	48.42	49.46	50.00	51.76	50.53	51.02	43.68	45.45	58.59	43.96	49.29

ภาพที่ 4.4 ค่าความถ่วงดุลของแบบจำลองจากการใช้ถุงคำในการให้น้ำหนักคำ

จากภาพที่ 4.4 แสดงการเปรียบเทียบค่าความถ่วงดุลของแบบจำลองจากการใช้ถุงคำในการให้น้ำหนักคำ ของเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ผลปรากฏว่า เทคนิค Random Forest ที่ให้ค่าความถ่วงดุลสูงสุด เมื่อพิจารณาตามรอบแล้วจะเห็นได้ว่าในรอบที่ 1-6 มีค่าเฉลี่ยใกล้เคียงกัน อยู่ที่ระหว่าง 48-51% และมาลดลงต่ำสุดในรอบที่ 7 ที่ 43.68% แล้วค่อยๆเพิ่มขึ้นจนสูงที่สุดในรอบที่ 9 ที่ 58.59% จนมาในรอบ

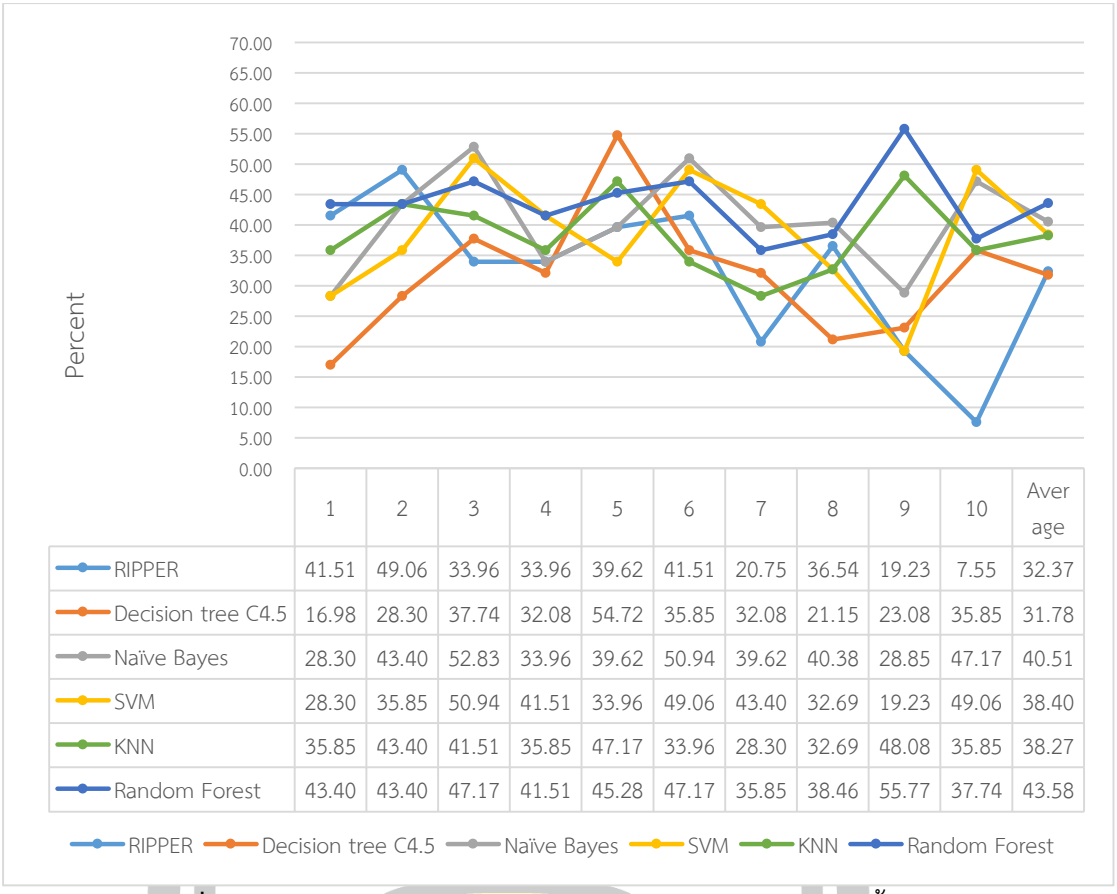
สุดท้ายรอบที่ 10 ค่าได้ลดลงมาจนเกือบต่ำสุดกว่าทุกรอบ จะเห็นได้ว่าเทคนิค Random Forest ให้ค่าความถ่วงดุลเฉลี่ยสูงสุดที่ 49.29% รองลงมาคือเทคนิค Naïve Bayes ที่ 48.02% ส่วนเทคนิคที่ให้ค่าความถ่วงดุลน้อยที่สุด คือ เทคนิค RIPPER ที่ 39.43%



	1	2	3	4	5	6	7	8	9	10	Average
RIPPER	43.14	59.09	64.29	75.00	58.33	53.66	55.00	48.72	52.63	50.00	55.99
Decision tree C4.5	40.91	51.72	51.28	54.84	59.18	59.38	54.84	55.00	75.00	52.78	55.49
Naïve Bayes	39.47	65.71	59.57	75.00	63.64	60.00	61.76	63.64	55.56	59.52	60.39
SVM	48.39	57.58	50.94	68.75	62.07	66.67	60.53	54.84	41.67	59.09	57.05
KNN	51.35	54.76	48.89	52.78	55.56	48.65	41.67	44.74	56.82	55.88	51.11
Random Forest	54.76	57.50	53.19	68.75	57.14	55.56	55.88	55.56	61.70	52.63	57.27

ภาพที่ 4.5 ค่าความแม่นยำของแบบจำลองจากการใช้ถุงคำในการให้น้ำหนักคำ

จากภาพที่ 4.5 แสดงการเปรียบเทียบค่าความแม่นยำของแบบจำลองจากการใช้ถุงคำในการให้น้ำหนักคำ ของเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ผลปรากฏว่า เทคนิค Naïve Bayes ที่ให้ค่าความแม่นยำสูงสุด เมื่อพิจารณาตามรอบแล้วจะเห็นได้ว่าในรอบที่ 1 มีค่าต่ำที่สุดกว่าทุกรอบ ที่ 39.47% จากนั้นจึงเพิ่มขึ้นและลดลงตามลำดับ จนมาในรอบที่ 4 จึงพุ่งสูงขึ้นจบสูงที่สุด ที่ 75.00% จากนั้นจึงค่อยๆลดลงในรอบที่ 5-10 มีค่าเฉลี่ยใกล้เคียงกัน อยู่ที่ระหว่าง 55-63% จะเห็นได้ว่าเทคนิค Naïve Bayes ให้ค่าความแม่นยำเฉลี่ยสูงสุดที่ 60.39% รองลงมาคือเทคนิค Random Forest ที่ 57.27% ส่วนเทคนิคที่ให้ค่าความแม่นยำน้อยที่สุด คือ เทคนิค K-NN ที่ 51.11%



	1	2	3	4	5	6	7	8	9	10	Average
RIPPER	41.51	49.06	33.96	33.96	39.62	41.51	20.75	36.54	19.23	7.55	32.37
Decision tree C4.5	16.98	28.30	37.74	32.08	54.72	35.85	32.08	21.15	23.08	35.85	31.78
Naïve Bayes	28.30	43.40	52.83	33.96	39.62	50.94	39.62	40.38	28.85	47.17	40.51
SVM	28.30	35.85	50.94	41.51	33.96	49.06	43.40	32.69	19.23	49.06	38.40
KNN	35.85	43.40	41.51	35.85	47.17	33.96	28.30	32.69	48.08	35.85	38.27
Random Forest	43.40	43.40	47.17	41.51	45.28	47.17	35.85	38.46	55.77	37.74	43.58

ภาพที่ 4.6 ค่าความระลึกของแบบจำลองจากการใช้ถุงคำในการให้น้ำหนักคำ

จากภาพที่ 4.6 แสดงการเปรียบเทียบค่าความระลึกของแบบจำลองจากการใช้ถุงคำในการให้น้ำหนักคำ ของเทคนิค RIPPER เทคนิค Decision tree C4.5 เทคนิค Naïve Bayes เทคนิค SVM เทคนิค K-NN และเทคนิค Random forest ในการจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ผลปรากฏว่า เทคนิค Random Forest ที่ให้ค่าความระลึกสูงสุด เมื่อพิจารณาตามรอบแล้วจะเห็นได้ว่าในรอบที่ 1-6 มีค่าเฉลี่ยใกล้เคียงกัน อยู่ที่ระหว่าง 41-47% และมาลดลงต่ำสุดในรอบที่ 7 ที่ 35.85% แล้วค่อยๆเพิ่มขึ้นในรอบที่ 8-9 และ จนมาในรอบสุดท้ายรอบที่ 10 ค่าได้ลดลงมาจนเกือบต่ำสุดกว่าทุกรอบที่ 37.74% จะเห็นได้ว่าเทคนิค Random Forest ให้ค่าความระลึกเฉลี่ยสูงสุดที่ 43.58% รองลงมาคือเทคนิค Naïve Bayes ที่ 40.51% ส่วนเทคนิคที่ให้ค่าความระลึกน้อยที่สุด คือ เทคนิค Decision tree C4.5 ที่ 31.78%

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

ในการศึกษาเพื่อค้นหาเทคนิคการทำเหมืองข้อความที่มีประสิทธิภาพในการวิเคราะห์ความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตรที่ได้มีการแสดงความคิดเห็นด้วยภาษาไทย ซึ่งเก็บรวบรวมจากการแสดงความคิดเห็นบนเครือข่ายสังคมออนไลน์ จากเว็ปไซต์ Pantip และ Facebook ซึ่งจะใช้ข้อความความคิดเห็นจากเพจหลักที่ได้รับความนิยมและน่าเชื่อถือที่ได้รับการจัดอับดับแนะนำ 10 เพจหมอเด็ก ให้ความรู้เรื่องการเลี้ยงลูก ที่พ่อกับแม่ต้องติดตาม ตั้งแต่วันที่ 1 มกราคม 2561 ถึงวันที่ 31 ธันวาคม 2562 จำนวน 1,925 ข้อความคิดเห็น โดยงานวิจัยนี้ได้ทำการแบ่งคุณลักษณะออกเป็น 2 กลุ่ม คือ คุณลักษณะเชิงบวกและคุณลักษณะเชิงลบ ซึ่งได้นำเอาเทคนิคของการทำเหมืองข้อความมาใช้ในการวิเคราะห์และเปรียบเทียบทั้งหมด 6 เทคนิค ได้แก่ เทคนิคริปเปอร์ (RIPPER) เทคนิคต้นไม้ตัดสินใจแบบ ซี4.5 (Decision tree C4.5) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคซัพพอร์ตเวคเตอร์แมชชีน (SVM Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-NN) และเทคนิคการสุ่มป่าไม้ (Random forest) และทำการวัดประสิทธิภาพด้วยวิธีการ 10-fold-cross-validation

5.1 สรุปผล

งานวิจัยนี้ได้ทำการวิเคราะห์และเปรียบเทียบประสิทธิภาพของแบบจำลองและศึกษากระบวนการในการทำเหมืองความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตรที่เป็นภาษาไทยจากข้อความคิดเห็นทั้งหมด จำนวน 1,925 ข้อความ โดยงานวิจัยนี้ได้ทำการแบ่งคำคุณลักษณะออกเป็น 2 กลุ่ม คือ คำคุณลักษณะเชิงบวกและคำคุณลักษณะเชิงลบ ซึ่งคำวิเศษณ์นี้จะสามารถแสดงถึงอารมณ์เชิงบวกและเชิงลบได้ดี [31] ด้วยการวิเคราะห์ 2 แบบ คือ การให้ค่าน้ำหนักตามจำนวนคำก่อนการใช้ถุงคำ (bag of word) และหลังการใช้ถุงคำ (bag of word) ด้วยเทคนิคทั้งหมด 6 เทคนิค

ในกระบวนการเตรียมข้อมูลเพื่อหาคำบ่งชี้คำคุณลักษณะได้เลือกวิธีในการตัดคำแบบอิงพจนานุกรม ซึ่งข้อดีของการตัดคำด้วยวิธีนี้คือ สามารถเพิ่มคำศัพท์อื่นๆ ได้ จากนั้นจึงทำการแบ่งประเภทของคำที่ใช้ในการกำหนดคุณลักษณะ โดยจะใช้เพียงคำวิเศษณ์ที่สามารถระบุความหมายที่เป็นบวกและเชิงลบได้เท่านั้น ดังนั้นผลของการตัดคำอาจมีผลทำให้คำบางคำอาจไม่พบในพจนานุกรมและไม่สามารถระบุประเภทของคำคำนั้นได้ ดังนั้นคำที่ใช้ในการวิเคราะห์อาจไม่ได้มีเฉพาะคำที่เป็น

คำวิเศษณ์อย่างเดียวเท่านั้น ผลสรุปที่ได้จากการเปรียบเทียบของทั้ง 6 โมเดล สามารถสรุปได้เป็น 2 ส่วน ดังนี้

5.1.1 ผลการวิเคราะห์ข้อมูลด้วยวิธีการให้ค่าน้ำหนักตามจำนวนคำก่อนการใช้ bag of word สรุปได้ว่า เทคนิค Random forest ให้ผลการทดสอบดีที่สุด โดยให้ค่าประสิทธิภาพของแบบจำลองที่ 83.85% ค่าความถ่วงดุล 83.55% ค่าความแม่นยำ 89.62% และค่าความระลึก 78.38% ส่วนเทคนิคที่ให้ค่าประสิทธิภาพของแบบจำลองน้อยที่สุดคือ เทคนิค K-NN ที่ 75.62% โดยให้ค่าความถ่วงดุล 75.53% ค่าความแม่นยำ 74.61% และค่าความระลึก 76.72%

5.1.2 ผลการวิเคราะห์ข้อมูลด้วยวิธีการให้ค่าน้ำหนักตามจำนวนคำหลังการใช้ bag of word สรุปได้ว่า เทคนิค Random forest ให้ผลการทดสอบดีที่สุด โดยให้ค่าประสิทธิภาพของแบบจำลองที่ 50.04% โดยให้ค่าความถ่วงดุล 49.29% ค่าความแม่นยำ 57.27% และค่าความระลึก 43.58% ส่วนเทคนิคที่ให้ค่าประสิทธิภาพของแบบจำลองน้อยที่สุดคือ เทคนิค Decision tree C4.5 ที่ 42.26% โดยให้ค่าความถ่วงดุล 39.52% ค่าความแม่นยำ 55.49% และค่าความระลึก 31.78%

5.2 อภิปรายผล

จากผลการวิจัยการให้ค่าน้ำหนักตามจำนวนคำก่อนการใช้ bag of word และหลังการใช้ bag of word ของความเห็นของผู้ปกครอง ได้ผลการวิจัยที่ตรงกัน ทำให้เห็นว่าเทคนิค Random forest ที่ให้ประสิทธิภาพของแบบจำลองสูงสุดนั้นเหมาะกับชุดข้อมูลที่มีการจำแนกคำ เนื่องจากเทคนิค Random forest ที่อยู่ภายใต้เทคนิคฐานต้นไม้ จะมีความได้เปรียบกว่าเทคนิคอื่นๆที่ได้นำมาทดสอบ เพราะเทคนิคนี้มีความสามารถในการทำนายผลได้อย่างแม่นยำ เป็นอัลกอลิทึมที่มีลักษณะแบบไม่ตัดกิ่งหรือต้นไม้ถดถอย ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูลและคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือกจะถูกนำมาใช้ในการทดสอบ ทำให้ไม่มีปัญหาในเรื่องของ overfitting ซึ่งสอดคล้องกับการวิจัยของ วัชรีวรรณ จิตต์สกุล [32] ได้วิเคราะห์ความเสถียรของอัลกอริทึม ในการจำแนกข้อความแสดงความคิดเห็นเชิงบวกและเชิงลบในการให้บริการของเว็ปไซต์ พบว่าเทคนิคที่ให้ผลลัพธ์ดีที่สุด คือ เทคนิค Random forest ด้วยเช่นกัน ดังนั้นทำเห็นว่าเทคนิค Random forest เหมาะที่จะใช้ในการวิเคราะห์ชุดข้อมูลความคิดเห็นได้ดีที่สุด ส่วนเทคนิค K-NN และเทคนิค Decision tree C4.5 ให้ผลของค่าประสิทธิภาพของแบบจำลองน้อยที่สุดทั้งนี้อาจเกิดจากลักษณะของข้อมูลไม่เหมาะต่อการทำงานของโมเดลด้วยเช่นกัน

แม้ว่าโมเดลการจำแนกกลุ่มของชุดข้อมูลที่ใช้ในการทดสอบนี้จะได้ผลลัพธ์ที่ดี แต่อย่างไรก็ตามข้อผิดพลาดที่เกิดขึ้นจากการจำแนกกลุ่มของโมเดลทั้ง 6 นี้ โดยส่วนใหญ่จะเกิดข้อผิดพลาดขึ้นในกระบวนการของการตัดคำ เนื่องจากข้อความภาษาไทยนั้นมีโครงสร้างของประโยคที่ไม่แน่นอนอีกทั้ง

ความหมายของคำที่มีต่อประโยคนั้นๆ มีหลายความหมายด้วยกัน รวมไปถึงความยืดหยุ่นของการใช้ภาษาไทย แต่ทั้งนี้คำที่ใช้ในงานวิจัยนี้ยังไม่ครอบคลุมถึงการจำแนกกลุ่มข้อความในปัญหาอื่นๆ เนื่องจากคำบ่งชี้คุณลักษณะที่ได้ใช้ในงานวิจัยนี้นำมาจากข้อความที่ใช้ในการวิจัยนี้เท่านั้น

5.3 ข้อเสนอแนะ

สำหรับงานวิจัยครั้งต่อไป อาจมีการนำคำบางคำที่เป็นศัพท์เฉพาะมาใช้เป็นคำบ่งชี้คุณลักษณะเพื่อเพิ่มความแม่นยำในการวิเคราะห์ประเภทของคำและประโยคได้ดีมากขึ้น ซึ่งจะส่งผลทำให้แบบจำลองที่ได้มีประสิทธิภาพมากยิ่งขึ้นอีกด้วย

บรรณานุกรม

บรรณานุกรม

[1] G. S. O'Keeffe, K. Clarke-Pearson, C. Council on, and Media, "The impact of social media on children, adolescents, and families," *Pediatrics,* vol. 127, no. 4, pp. 800-4, Apr 2011, doi: 10.1542/peds.2011-0054.

[2] อภิรพี เศรษฐรักษ์ ตันเจริญวงศ์, ศรีรัฐ ภักดีรณชิต, and ญาณวุฒิ เศวตธิติกุล, "<พฤติกรรมการใช้หน้าจอของเด็กไทยวัย 0-3 ปี ในเขตกรุงเทพมหานคร.pdf>."

[3] N. Apisit and K. Supaporn, "Factors related to Game Addiction among Prathom Suksa 4-6 Schoolchildren," *J. of Soc Sci & Hum,* vol. 46, no. 2, pp. 111-141 (in Thai), 2020.

[4] ศิริธ ยิ่งเรงเริง, "<การติดเกมในเด็ก การคัดกรองและการแก้ไข.pdf>."

[5] โสภณา จิรวงศ์นุสรณ์, ณัฐวดี จิตรมานะศักดิ์, สาริน ฤทธิสาร, and พวงผกา ภูยาดาว, "อันตรายที่แฝงมากับโทรศัพท์มือถือ."

[6] ฉัตร ชูชื่น, พูนทรัพย์ จันต๊ะ, พัฒนชัย กุลจันทร์, and ณัฐดนัย เขียววาท, "ผลการเรียนรู้วิชาภาษาอังกฤษ โดยใช้แอปพลิเคชันการสอนภาษาอังกฤษเลทเซินวันฟอร์คิดส์โดยใช้เทคโนโลยีเสมือนจริง," *Journal of Liberal Arts, Maejo University,* vol. 7, no. 2, pp. 81-95, 2019.

[7] F. H. Khan, S. Bashir, and U. Qamar, "TOM: Twitter opinion mining framework using hybrid classification scheme," *Decision Support Systems,* vol. 57, pp. 245-257, 2014, doi: 10.1016/j.dss.2013.09.004.

[8] น. ปิ่นเมือง and จ. ทองคำ, "การจำแนกความคิดเห็นของคนไทยเกี่ยวกับสื่อออนไลน์โดยใช้การทำเหมืองข้อความ," *Journal of Science & Technology MSU,* vol. 37, no. 3, 2018.

[9] S. Gupta, S. Jain, S. Gupta, and A. Chauhan, "Opinion Mining for Hotel Rating through Reviews Using Decision Tree Classification Method," *International Journal of Advanced Research in Computer Science,* vol. 9, no. 2, p. 180, 2018.

[10] กานดา แผ่วัฒนากุล and ดร.ปราโมทย์ ลือนาม, "การ วิเคราะห์ เหมือง ความ คิดเห็น บนเครือ ข่าย สังคม ออนไลน์," *Modern Management Journal,* vol. 11, no. 2, pp. 11-20, 2013.

[11] P. Nanaumphai, "Intrusion Detection Using Classification Techniques in Data Mining," *Journal of Information Technology Management and Innovation,* vol. 6, no. 2, pp. 111-118 (in Thai), 2019.

[12] ประพัฒน์ พรมน้ำอ่าง, วสุวรรธน์ พงศ์ขจร, and นิเวศ จิระวิชิตชัย, "การจำแนกกลุ่มข้อความรีวิวโดยใช้เทคนิคเหมืองข้อมูล," วารสารวิทยาศาสตร์และเทคโนโลยี มทร. ธัญบุรี ปีที่ : *6* vol. 1, pp. 94-102, 2016.

[13] ราชวิทย์ ทิพย์เสนา, ฉัตรเกล้า เจริญผล, and แ. สมประเสริฐศรี, "การจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา โดยใช้เทคนิคเหมืองข้อความ," *Journal of Science and Technology Mahasarakham University,* vol. 33, no. 5, pp. 493-493, 2014.

[14] อ. วสวัตติ์ and ท. จารี, "การวิเคราะห์ความคิดเห็นต่อเกมมือถือพับจีด้วยเหมืองข้อความ," *Journal of Science & Technology MSU,* vol. 39, no. 5, 2020.

[15] วสุพล จิตรานนท์ and ป. เล่ห์มงคล, "ความคิดเห็นของผู้ปกครองเด็กปฐมวัยที่มีต่อสื่อการ์ตูน," วารสารศึกษาศาสตร์ปริทัศน์, vol. 30, no. 3, pp. 168-174, 2015.

[16] สุวภา บุญอุไร *et al.*, "โมเดลความสัมพันธ์เชิงสาเหตุของการใช้สมาร์ทโฟนของเด็กและผู้ปกครองที่ส่งผลต่อความ ฉลาดทางอารมณ์ของเด็กปฐมวัย," *Hatyai Academic Journal,* vol. 16, no. 2, pp. 127-137, 2018.

[17] สมศักดิ์ วิชัยกิจ and มาลีรัตน์ โสดานิล, "การจัดกลุ่มลูกค้าสินเชื่อจากการปฏิเสธด้วยวิธีการทำเหมืองข้อความ," การประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ *11,* pp. 359-363, 2015.

[18] P. Klangnok and J. Thongkam, "Applying the Ensemble Technique for Improving Rule-based Models Performance," Mahasarakham University, 2018.

[19] ปริญญา สงวนสัตย์, ปัญญาประดิษฐ์ด้วยการเรียนรู้ของเครื่องฉบับภาษา *python*: ปริญญา สงวนสัตย์, 2019.

[20] รวิสุดา เทศเมือง and นิเวศ จิระวิชิตชัย, "การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน,," วารสารวิศวกรรมศาสตร์ มหาวิทยาลัยสยาม,, vol. 18, no. 1, pp. 1-11, 2017.

[21] อรทิพย์ เลื่อยงาม and ชัยพร เขมะภาตะพันธ์, "การจัดประเภทเอกสารด้วยวิธีเอสวีเอ็มเพื่อการป้องกันเอกสารรั่วไหล," การประชุมวิชาการ "นเรศวรวิจัย" ครั้งที่ *7,* pp. 3-12, 2011.

[22] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *Machine Learning: ECML-98*, Berlin,

Heidelberg, C. Nédellec and C. Rouveirol, Eds., 1998// 1998: Springer Berlin Heidelberg, pp. 137-142.

[23] M.sujatha and Dr. G. Lavanya Devi, "<feature-selection-techniques-using-for-high-dimensional-data-in-machine-learning-IJERTV2IS90912.pdf>," 2013.

[24] วัชรีวรรณ จิตต์สกุล and ส. สดสี, "การเคราะห์การจำแนกข้อความด้วยการเปรียบเทียบความเสถียรของอัลกอริทึม," *Sripatum Review of Science and Technology,* vol. 9, no. 1, pp. 19-31, 2017.

[25] จุฑาทิพย์ ทิพย์พูล and นิเวศ จิระวิชิตชัย, "การจำแนกจดหมายอิเล็กทรอนิกส์ที่เป็นสแปมโดยการใช้เทคนิคเหมืองข้อมูล," *Science and Technology RMUTT Journal,* vol. 6, no. 8, pp. 102-109, 2016.

[26] สมศักดิ์ ศรีสวการย์ and ส. ศรีสวย, "การวิเคราะห์เหมืองความคิดเห็นโดยใช้เทคนิคการสกัดคำ," การ ประยุกต์ ใช้ เทคโนโลยี สารสนเทศ, vol. 6, no. 2, pp. 95-104, 2020.

[27] S. Vongsingthong and N. Wisitpongphan, "Classification of university students' behaviors in sharing information on Facebook," in *2014 11th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 14-16 May 2014 2014, pp. 134-139, doi: 10.1109/JCSSE.2014.6841856.

[28] อดิเทพ ไชยสาร and รัฐสิทธิ์ สุขะหุต, "การประมาณอารมณ์จากความคิดเห็นภาษาไทยโดยใช้เทคนิคการเรียนรู้ของเครื่อง," in การประชุมวิชาการระดับชาติ ด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ ครั้งที่ *9*, 2013, pp. 260-266.

[29] บรรหาร จันทะวงค์ *et al.*, "การจัดประเภทเอกสารงานวิจัยทางโลจิสติกส์ กรณีศึกษามหาวิทยาลัยราชภัฏเชียงราย," in การประชุมวิชาการ งานวิจัยและพัฒนาเชิงประยุกต์ ครั้งที่ *7*, 2015, pp. 491-494.

[30] ภรัณยา ปาลวิสุทธิ์, "<การเพิ่มประสิทธิภาพเทคนิคต้นไม้ตัดสินใจบนชุดข้อมูลที่ไม่สมดุลโดยวิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยสำหรับข้อมูลการเป็นโรคติดอินเทอร์เน็ต.pdf>," 2559.

[31] Y. Sun, C. Quan, X. Kang, Z. Zhang, and F. Ren, "Customer emotion detection by emotion expression analysis on adverbs," *Information Technology and Management,* vol. 16, no. 4, pp. 303-311, 2015.

[32] J. Watchareewan and S. Sunantha, "TEXT CLASSIFICATION ANALYSIS BY STABILITY COMPARISON OF ALGORITHMS," *Sripatum Review of Science and Technology,* vol. 9, no. 1, pp. 19-31 (in Thai), 2017.

ประวัติผู้เขียน

ชื่อ

วันเกิด 4 ธันวาคม พ.ศ.2525

สถานที่เกิด โรงพยาบาลสุรินทร์

สถานที่อยู่ปัจจุบัน 170/8 ถนนถีนานนท์ ตำบลตลาด อำเภอเมือง จังหวัดมหาสารคาม รหัสไปรษณีย์ 44000

ประวัติการศึกษา พ.ศ. 2548 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.)
สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยราชภัฏสุรินทร์
พ.ศ. 2564 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.)
สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ
มหาวิทยาลัยมหาสารคาม