การจำแนกความคิดเห็นโซเชียลมีเดียไทยเกี่ยวกับวัคซีนป้องกันโควิดโดยใช้เหมืองข้อความ

Classifying Thai Social Media Opinions in the Covid Vaccine Using Text Mining

พีคอน หมายสุข* และ จารี ทองคำ

Peakon Maisook* and Jaree Thongkam

ภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

Department of Information Technology, Faculty of Informatics, Mahasarakham University

Email: peakon.ms@bru.ac.th

Received : October 3, 2024

Revised : April 17, 2025

Accepted : May 8, 2025

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์ เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองจำนวน 5 เทคนิค ได้แก่ เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคริปเปอร์ เทคนิคฟิวเรีย และเทคนิคป่าสุ่ม มาสร้างแบบจำลอง เพื่อจำแนกความคิดเห็นของคนไทยในสังคม ว่ามีความคิดเห็นต่อการฉีดวัคซีนป้องกันโรคโควิด-19 ให้กับเด็ก โดยข้อมูลนั้นถูกรวบรวมมาจากเครือข่ายสังคมออนไลน์ จำนวนทั้งหมด 2,466 ความคิดเห็น คำคุณลักษณะคำกริยา คำกริยาวิเศษณ์ คำคุณศัพท์ ได้ถูกเลือกมาใช้ในการสร้างแบบจำลอง โดยคำคุณลักษณะประเภทนี้สามารถระบุความรู้สึกในเชิงบวก และเชิงลบได้อย่างชัดเจน ในงานวิจัยนี้ 10 โฟลด์ครอสวาลิเดชั่นได้ถูกนำมาใช้ในการแบ่งกลุ่มข้อมูล เป็นชุดเรียนรู้และชุดทดสอบ นอกจากนี้ค่าความแม่นยำ ค่าความระลึก และค่าความถูกต้องได้ถูกนำมาคำนวณเพื่อใช้ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง ผลการทดลองแสดงให้เห็นว่า นาอีฟเบย์เป็นเทคนิคสร้างแบบจำลองที่มีประสิทธิภาพสูงสุดที่ค่าความแม่นยำร้อยละ 95.40 ค่าความระลึกที่ร้อยละ 95.40 และค่าความถูกต้องที่ร้อยละ 95.40

คำสำคัญ: การจำแนกความคิดเห็น เหมืองข้อความคิดเห็น วัคซีนโควิด-19 ในเด็กไทย

ABSTRACT

This research aims to compare the efficiency of five modeling techniques: Naive Bayes, Support Vector Machine, Ripper, Fourier, and Random Forest. These techniques are used to build models for classifying the opinions of Thai people on social media regarding COVID-19 vaccination for children. The data, consisting of 2,466 opinions, was collected from online social networks. Verbs, adverbs, and adjectives were selected for model building, as these types of words can clearly indicate positive and negative sentiment. In this research, 10-fold cross-validation was used to divide the data into training and testing sets. Furthermore, accuracy, recall, and precision were calculated to compare the performance of the models. The experimental results show that Naive Bayes is the most efficient modeling technique, achieving an accuracy of 95.40%, a recall of 95.40%, and a precision of 95.40%.

Keywords: Opinion classification, Opinion mining, COVID-19 vaccine in Thai children

บทนำ

โรคโควิด-19 ได้มีการเริ่มระบาดครั้งแรก ในเดือนธันวาคม ปี พ.ศ. 2562 โดยเป็นโรคที่เกิดจากเชื้อไวรัสโคโรนาสายพันธ์ใหม่ ที่สามารถแพร่กระจายเชื้อไวรัสจากคนสู่คนโดยมีผู้ติดเชื่อลุกลามไป6ทั่วโลก จนทำให้องค์การอนามัยโลก (WHO) ได้กำหนดให้เป็นภาวะการระบาดใหญ่ (Pandemic) และในประเทศไทยเริ่มมีผู้ป่วย ในเดือนมกราคม พ.ศ. 2563 (สุรัยยา หมานมานะ และคณะ, 2563) และต่อมามีการระบาดอย่างหนัก (อุษา คำประสิทธิ์, 2565) แนวทางในการช่วยลดปัญหาการระบาด คือ วัคซีนโควิด-19 โดยภาครัฐได้มีบริการฉีดวัคซีนกับประชาชน และได้มีบริการฉีดวัคซีนโควิด-19 ให้กับเด็กอายุ 5 - 11 ปี (ประกายแก้ว ศิริพูล และคณะ, 2565) แต่เนื่องจากวัคซีนโควิด-19 มีการผลิตและทดสอบได้ไม่นาน จึงเกิดความรู้สึกสำหรับผู้ปกครองรวมทั้งคนทั่วไปต่อวัคซีน และได้แสดงความคิดเห็นมากมายผ่านโซเชียลมีเดีย (Social Media) ต่าง ๆ ซึ่งมีทั้งความคิดเห็นเชิงลบ และเชิงบวก

การวิเคราะห์ความคิดเห็นของประชาชนในสังคมไทย ต่อการฉีดวัคซีนป้องกันโรคโควิด-19 ให้กับเด็ก จะนำไปสู่ความเข้าใจความคิดเห็น ซึ่งสามารถทำให้หน่วยงานที่เกี่ยวข้องสามารถเลือกวิธีการที่จะประชาสัมพันธ์หรือรณรงค์ให้ความรู้ความเข้าใจ และประโยชน์ของการฉีดวัคซีนอย่างมีประสิทธิภาพ

การทำเหมืองข้อความ (Khan et al., 2014; Phung et al., 2021) คือการวิเคราะห์ความคิดเห็นโดยใช้เทคนิคการวิเคราะห์ข้อความ (Text Mining) เพื่อระบุความรู้สึกหรือความคิดเห็น เช่น การนำความคิดเห็นเชิงบวก เชิงลบและเป็นกลาง ที่อยู่ภายใต้ข้อความออกมาโดยอัตโนมัติ โดยการทำเหมืองความคิดเห็น จะใช้เทคนิคการค้นหาความรู้ใหม่จากข้อมูลประเภทข้อความที่มีปริมาณมากเป็นพิเศษ รวมเพื่อหาความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความเพื่อให้ทราบความหมาย และนำไปสร้างแบบจำลองเพื่อช่วยค้นหากลุ่มข้อมูลเป้าหมาย เช่น งานวิจัยของ พิศิษฐ์ บวรเลิศสุ และวรภัทร ไพรีเกรง (2565) ได้ทำเหมืองความคิดเห็นวิเคราะห์ความรู้สึกต่อการแนะนำสินค้าออนไลน์ เพื่อจำแนกความคิดเห็นเชิงบวกเป็นกลาง และเชิงลบ มี 5 กระบวนการสร้างตัวแบบในการทดลอง ได้แก่ การเตรียมข้อมูล การตัดคำ การฝึกอบรมข้อมูล การสร้างโมเดลเพื่อแยกประเภทข้อมูล และการประเมินตัวแบบ ทำการทดลองกับข้อมูลตัวอย่างการแสดงความคิดเห็นต่อสินค้า และบริการออนไลน์ภาษาไทย จำนวน 12,900 ข้อมูล และสร้างโมเดลเพื่อแยกประเภทข้อมูลด้วยเทคนิค LSTM, SGD, Logistic Regression และ Support Vector Machines โดยผลการทดลองเทคนิค LSTM ให้ค่าความถูกต้องร้อยละ 81.27 ซึ่งให้ค่าความถูกต้องมากที่สุด Vyas et al. (2022) ได้ทำการพัฒนาระบบเพื่อจำแนกความคิดเห็นในเชิงบวกและเชิงลบต่อโรคโควิด-19 ได้ จากผู้ใช้งาน Twitter โดยในขั้นตอนการสร้างโมเดลใช้เทคนิค Decision Tree, Gaussian Naïve Bayes, Multinominal Naïve Bayes, Logistic Regression, Random Forest และ LSTM โดยค่าที่ใช้ในการวัดประสิทธิภาพของแบบจำลองที่สร้างขึ้นมาในครั้งนี้คือ ค่าความถ่วง (F-Measure) ค่าความแม่นยำ (Precision) และค่าความลึก (Recall) จากผลการทดลองพบเทคนิค LSTM ให้ประสิทธิภาพในการจำแนกความคิดเห็นได้ดีที่สุดที่ร้อยละ 83 งานวิจัยของ Zope and Rajeswari (2022) ได้ทำการวิเคราะห์ความคิดเห็นของผู้ใช้งานทวิตเตอร์เกี่ยวกับโรควิด-19 จำนวน 179,108 ความคิดเห็นว่า มีความรู้สึก มีความสุขมาก กลัว เศร้า หรือกลัว โดยขั้นตอนการดำเนินการวิจัย ได้แก่ ขั้นตอนในการคัดเลือกข้อมูล ขั้นตอนในการเตรียมข้อมูล ขั้นตอนกำหนดป้ายให้กับความคิดเห็น ขั้นตอนในการใช้คัดแยกความคิดเห็น ใช้เทคนิค Random Forest, Decision Tree, Support Vector Machine และขั้นตอนการแสดงผล โดยพบว่า จากข้อมูลตัวอย่างความรู้สึกเศร้า และกลัวมีมากที่สุด ปพิชญา กลางนอก และจารี ทองคำ (2562) ได้ทำการการสร้างแบบจำลองเพื่อพยากรณ์โรคมะเร็งเต้านม และโรคเบาหวาน โดยนำข้อมูลจาก ข้อมูลผู้โรคมะเร็งเต้านม

699, ข้อมูลโรคเบาหวาน จำนวน 768 โดยสร้างแบบจำลองใช้เทคนิค FURIA, MODLEM และ RIPPER ใช้ร่วมกับเทคนิค Bagging และ Weighted Instances Handler Wrapper วัดประสิทธิภาพจะทำการคำนวณหาค่าความไว (Sensitivity) ค่าความจำเพาะ (Specificity) ความถูกต้อง (Accuracy) ซึ่งได้ผลการทดลองคือ Bagging และ FURIA ให้ค่าความไวและความถูกต้อง มากที่สุดที่ร้อยละ 97.86 และร้อยละ 92.12 และ Bagging และ MODLEM ให้ค่าความจำเพาะมากที่สุดที่ร้อยละ 90.37

ดังนั้น ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาแบบจำลอง เพื่อการจำแนกความคิดเห็นของคนไทยต่อการฉีดวัคซีนโควิด-19 ให้กับเด็กบนโซเชียลมีเดียด้วยการทำเหมืองความคิดเห็น โดยการนำเอาความคิดเห็นของคนไทยต่อการฉีดวัคซีนโควิด-19 ให้กับเด็กอายุ 5 - 11 ปี บนโซเชียลมีเดียว่ามีความคิดเห็นไปในทิศทางใดซึ่งจะแบ่งออกเป็น 2 กลุ่ม ได้แก่ ความคิดเห็นเชิงบวก และความคิดเห็นเชิงลบโดยการเปรียบเทียบเทคนิคที่ใช้ในการสร้างแบบจำลองเพื่อ จำแนกความคิดเห็น คือ นาอีฟเบย์ (Naïve Bayes: NB) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Classifier :SVM) ริปเปอร์ (Ripper: RP) ฟิวเรีย (Furia: FR) และเทคนิคป่าสุ่ม (Random Forest: RF) ซึ่งจะวัดประสิทธิภาพโดยการใช้ 10-fold cross validation เพื่อแบ่งกลุ่มข้อมูลออกเป็นชุดข้อมูล 2 ชุด คือ ชุดข้อมูล สำหรับเรียนรู้และชุดข้อมูลสำหรับทดสอบ โดยจะวัดประสิทธิภาพแบบจำลองด้วยค่าวัดประสิทธิภาพค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความลึก (Recall)

วัตถุประสงค์การวิจัย

1. สร้างแบบจำลองที่มีประสิทธิภาพในการจำแนกความคิดเห็นบนโซเชียลมีเดียต่อการฉีดวัคซีนโควิด-19 ให้กับเด็กอายุ 5 - 11 ปี

2. เปรียบเทียบประสิทธิภาพของเทคนิคที่นำมาใช้สร้างแบบจำลองเพื่อการจำแนกความคิดเห็นบนโซเชียลมีเดียต่อการฉีดวัคซีนโควิด-19 ให้กับเด็กอายุ 5 - 11 ปี

วิธีการดำเนินการวิจัย

1. การเตรียมข้อมูล

การเตรียมข้อมูลคือขั้นตอนในการค้นหารวบรวมข้อมูลความคิดเห็น และนำมาปรับให้อยู่ในรูปแบบที่เหมาะสมที่สามารถวิเคราะห์ และใช้กับเครื่องมือที่นำมาช่วยในการวิเคราะห์ มีขั้นตอนดังนี้

1.1 การรวบรวมข้อมูล คือขั้นตอนการเลือกข้อมูล เพื่อนำมาทำการวิเคราะห์ความคิดเห็นของคนไทย การฉีดวัคซีนสำหรับเด็กอายุ 5 - 11 ปี จะนำข้อมูลมาจากโซเชียลมีเดีย ได้แก่ ติ๊กตอก

ยูทูป และพันทิป โดยเป็นข้อมูลจากหลายกระทู้ที่มีหัวข้อที่เกี่ยวข้องกับการฉีดวัคซีนสำหรับเด็กที่มีอายุ 5 - 11 ปี และเลือกเอาข้อมูลที่เป็นความคิดเห็น (Comment) ที่แสดงไว้ในกระทู้นั้น ๆ จำนวนทั้งหมด 2,466 ความคิดเห็น ตั้งแต่วันที่ 1 กันยายน 2561 ถึงวันที่ 31 มกราคม 2563 แล้วนำมาเก็บไว้ในไมโครซอฟท์เอ็กเซล

1.2 การกลั่นกรองข้อมูลเป็นการจัดเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการวิเคราะห์และจัดการด้วยโปรแกรมที่จะใช้ในการวิเคราะห์ โดยประกอบด้วยขั้นตอน การแก้ไขคำที่ผิด กำจัดตัวเลข กำจัดสัญลักษณ์ และกำจัดช่องว่างออกไป โดยในขั้นตอนนี้จะนำข้อมูลเข้าสู่โปรแกรมไมโครซอฟเอ็กเซล และดำเนินการเตรียมข้อมูล ดังแสดงในตารางที่ 1

ตารางที่ 1 ความคิดเห็นก่อนและหลังการเตรียม

ความคิดเห็น ก่อนเตรียม	ความคิดเห็นหลังเตรียม
ไม่ฉีดแน่นอน สงสารเด็ก ใครจะว่ายังไง ผมก็ขอไม่ให้ฉีด ผมโดนมา 2 เข็มแล้ว เกือบเดี้ยง!!!!	ไม่ฉีดแน่นอนสงสารเด็กใครจะว่ายังไงผมก็ขอไม่ให้ฉีดผมโดนมาเข็มแล้วเกือบเดี้ยง
ไม่สมควรบังคับเด็กฉีดวัคซีนค่ะ	ไม่สมควรบังคับเด็กฉีดวัคซีน
สงสารเด็กค่ะ ไม่อยากให้ฉีดเลย	สงสารเด็กไม่อยากให้ฉีดเลย

2. การสร้างคุณลักษณะ

2.1 การตัดคำ เป็นการนำข้อมูลที่ได้จากขั้นตอนการเตรียมข้อมูล โดยขั้นตอนนี้จะพิจารณาเพิ่มเติมว่าการตัดคำมีความเหมาะสมหรือไม่ โดยใช้เครื่องมือเป็น Python และ Library PyThaiNLP

2.2 การทำถุงคำ (Bag of Word) คือ การนำคำที่แยกจากกระบวนการตัดคำมาใช้ โดยถ้าเป็นคำที่ซ้ำกันนำมาเพียงหนึ่งคำและคำที่ไม่ซ้ำมาหนึ่งคำ ซึ่งจะทำให้ได้คำจำนวนหนึ่งที่ไม่ซ้ำกัน จากนั้นนำคำแต่ละคำมากำหนดประเภทของคำตามพจนานุกรม แล้วเลือกเฉพาะคำที่เป็นประเภท คำวิเศษณ์ คำกริยา คำกริยาวิเศษณ์ เพื่อให้ผู้เชี่ยวชาญทางด้านภาษาไทยระบุคำว่ามีความหมายเชิงบวก หรือเชิงลบ โดยใช้เครื่องมือเป็น Python และ Library PyThaiNLP

3. การเลือกคุณลักษณะ

3.1 การคัดเลือกคุณลักษณะ คือ การเอาคำคุณศัพท์ คำกริยา คำกริยาวิเศษณ์ ที่ได้จากขั้นตอนการทำถุงคำมาระบุว่ามีความหมายเป็นเชิงบวกหรือเชิงลบ แล้วกำหนดให้เป็นตัวเลข โดยคำที่มี

ความหมายเชิงบวกจะมีค่าเป็น 1 คำที่มีความหมายเชิงลบจะกำหนดค่าให้เป็น -1 ส่วนคำที่ไม่ได้กำหนดว่ามีความหมายสื่อถึงเชิงบวกหรือลบจะมีการกำหนดค่าให้เป็น 0 เพื่อจะนำตัวเลขเหล่านี้ไปคำนวณในขั้นตอนการกำหนดคลาส ดังตัวอย่างในตารางที่ 2

ตารางที่ 2 ตัวอย่างคำที่ผ่านขั้นตอนการเตรียมข้อมูล การสร้างคุณลักษณะ การคัดเลือกคุณลักษณะ

คำ	ประเภทคำ	คำเชิงลบ/เชิงบวก
ตกค้าง	คำกริยา (VERB)	-1
ชี้แนะ	คำกริยา (VERB)	1
ทรุด	คำกริยาวิเศษณ์ (ADV)	-1

3.2 การกำหนดคลาส เป็นการนำคำ ที่ได้จากขั้นตอนการทำถุงคำ แล้วแปลงข้อมูลคำที่มีความหมายเชิงบวกให้มีค่าเป็น 1 และเชิงลบมีค่าเป็น -1 แล้วนำไปหาค่าความถี่ของคำเชิงบวกเชิงลบที่อยู่ในความคิดเห็น โดยถ้าพบคำที่มีความหมายเชิงบวกจะให้ค่าความถี่เป็น 1 หากพบคำเดียวกัน 2 ครั้ง หรือมากกว่าก็จะนับรวมค่าความถี่เพิ่มเข้าด้วยกัน และถ้าพบคำที่มีความหมายเชิงลบจะให้ค่าความถี่เป็นลบ -1 หากพบคำเดียวกัน 2 ครั้งหรือมากกว่าก็จะนำค่า -1 เข้าไปรวมเข้าด้วยกัน และเมื่อได้ค่าความถี่แล้วจะนำมารวมกัน ผลที่ได้จะสามารถทำให้แบ่งคลาสออกได้ 3 คือ กรณีที่ 1 หากผลรวมมีค่าเป็นบวก ความคิดเห็นนั้นจะอยู่ในคลาส P กรณีที่ 2 หากผลรวมมีค่าเป็นลบ ความคิดเห็นนั้นจะอยู่ในคลาส N และกรณีที่ 3 หากผลรวมมีค่าเป็น 0 ความคิดเห็นนั้นจะอยู่ในคลาส NE และความคิดเห็นที่มีคลาสเป็น NE นั้นจะถูกตัดออกจากชุดข้อมูล โดยการใช้ เครื่องมือเป็น Python และ Library PyThaiNLP กำหนดคลาส จะได้ผลลัพธ์ ดังตัวอย่างการกำหนดคลาสในตารางที่ 3

ตารางที่ 3 ผลของการกำหนดคลาสด้วย Python

ID	เดี้ยง	รับผิดชอบ	กระทบ	...	SUM	CLASS
1	-1	0	0	...	-1	N
2	0	2	0	...	2	P
3	0	0	0	...	0	NE

จากความคิดเห็นที่ได้จากขั้นตอนการรวบรวมข้อมูล ได้นำมาดำเนินการเตรียมข้อมูลโดยการกลั่นกรองข้อมูลและตัดคำ ทำให้ได้ความคิดเห็นที่มีรูปแบบที่พร้อมทำการทดลอง จำนวน 2,466 ความคิดเห็น และนำความคิดเห็นที่ได้มาทำดัชนีโดยใช้หลักการทำถุงคำซึ่งทำให้ได้คำทั้งหมด 3,279 คำ และเลือกคำที่มีคุณลักษณะเป็น คำกริยา คำกริยาวิเศษณ์ คำคุณศัพท์ ที่มีหมายเป็นเชิงบวกหรือลบอย่างชัดเจน เพื่อไปใช้ในการกำหนดคลาสให้กับความคิดเห็นทั้งหมด โดยได้ทั้งหมด 254 คำ เป็นคำที่มีความหมายเชิงบวก 120 คำ และเชิงลบ 134 คำ จากขั้นตอนการกำหนดคลาส ทำให้ได้คลาส NE จำนวน 1,271 คลาส P จำนวน 638 และคลาส N จำนวน 557 โดยจะเลือกเฉพาะคลาส P และ N รวมกันจำนวน 1,195 ความคิดเห็นไปสร้างแบบจำลอง

3.3 การแปลงข้อมูลตัวเลขเป็นข้อมูลนามบัญญัติ คือ การเปลี่ยนแปลงค่าความถี่ของคำคุณลักษณะพึ่พบในแต่ละความคิดเห็นให้เป็นข้อมูลนาม ได้แก่การเปลี่ยนค่าความถี่ที่เป็นลบที่มีค่าน้อยกว่า 0 เปลี่ยนค่าเป็น 1 เปลี่ยนค่าความถี่ที่เป็นบวกที่มีค่ามากกว่า 0 ให้เป็น 1 และค่าความถี่ที่เป็น 0 ให้มีค่าเป็น 0 ดังเดิม ดังตารางที่ 4

ตารางที่ 4 ตัวอย่างการแปลงข้อมูลตัวเลขเป็นข้อมูลนามบัญญัติ

ID	เดี้ยง	รับผิดชอบ	กระทบ	...	CLASS
1	1	0	0	...	N
2	0	1	0	...	P

4. การสร้างแบบจำลองเพื่อจำแนก

4.1 นาอีฟเบย์ คือ โมเดลง่ายและไม่ซ้ำซ้อน โดยอาศัยทฤษฎีความน่าจะเป็น (Probability) เป็นหลักซึ่งถูกใช้ในการทำนายผลจัดเป็นเทคนิคในการแก้ปัญหาแบบ Classification ที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ โดยจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เหมาะกับการจัดข้อความเป็นสองประเภท (Banluesapy & Jirapanthong, 2022; Suharsono et al., 2022)

4.2 ซัพพอร์ตเวกเตอร์แมชชีน คือ อัลกอริทึมใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูลโดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวน การฝึกฝนให้ระบบเรียนรู้โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลซึ่งเหมาะกับข้อมูลที่มีการแบ่งกลุ่มอย่างชัดเจน (Kovacs et al., 2023; พิศิษฐ์ บวรเลิศสุ และวรภัทร ไพรีเกรง, 2565)

4.3 เทคนิคริปเปอร์ คือ เทคนิคที่ได้พัฒนาจากอัลกอริทึม IREP โดยใช้หลักการ Growing and Pruning เป็นการแยกกฎการจำแนกประเภทข้อมูลให้เป็นไปตามกลุ่มหมวดหมู่ ซึ่งมีเป้าหมายในการใช้หลักการตัดกิ่ง ในการเลือกกฎที่สามารถทำให้เกิดผลในการทำงานมากกว่าอัตราความผิดพลาด (Error Rate) ประกอบด้วยขั้นตอน คือการระบุกฎเริ่มต้น และขั้นตอน ระบุค่า Porst – Process Rule Optimization โดยการเรียนรู้จากข้อมูลที่ได้กำหนดคลาสไว้เรียบร้อยแล้ว ซึ่งแบ่งเป็น Growing Set และ Pruning Set สำหรับเป็นการช่วยแก้ปัญหาที่เกิดมาจากการแบ่งแยกกลุ่มที่ผิดพลาด จะทำจนกระทั้งจะได้ผลที่พอใจ ซึ่งเหมาะกับข้อมูลต้องการความแม่นยำในการสร้างกฎ (ปพิชญา กลางนอก และจารี ทองคำ, 2019)

4.4 เทคนิคฟิวเรีย (Furia: FR) คือ เทคนิคในการเหนี่ยวนำกฎที่ไม่เรียงลำดับตามฟัซซี่ เป็นวิธีการสร้างกฎเพื่อนำมาใช้ในการพยากรณ์ข้อมูล ซึ่งกฎที่นำมาสร้างเป็นกฎความสัมพันธ์ได้มาจากการประมวลผลของการทำเหมืองข้อมูลเพื่อค้นหากฎความสัมพันธ์ (Association Rule Mining) หรืออธิบายการจำแนกข้อมูลด้วยกฎความสัมพันธ์ โดยเป็นเทคนิคที่นำข้อดีของการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์และเทคนิคการจำแนกประเภทข้อมูล เพื่อค้นหากฎความสัมพันธ์ ซึ่งเหมาะกับข้อมูลที่มีความไม่แน่นอนและมีการผันแปร (ปพิชญา กลางนอก และ จารี ทองคำ, 2562)

4.5 เทคนิคป่าสุ่ม (Random Forest: RF) คือ การสร้าง model จาก Decision Tree หลาย ๆ model ย่อย ๆ โดยแต่ละโมเดลจะได้รับ data set ไม่เหมือนกัน ซึ่ง subset ของ data set ทั้งหมด ตอนทำ Prediction จะให้ Decision Tree ทำการ Prediction ในแต่ละโหนด และคำนวณผล Prediction ด้วยการ Vote Output ที่ถูกเลือกโดย Decision Tree มากที่สุด ซึ่งเป็นวิธีการวิธีการที่มีประสิทธิภาพและยืดหยุ่นมากกว่า Decision Tree (Banluesapy & Jirapanthong, 2022; Karthika et al., 2019)

2.6 การประเมินประสิทธิภาพแบบจำลอง

การวัดประสิทธิภาพของแบบจำลองจะใช้วิธี 10-fold cross validation คือ การแบ่งข้อมูลออกเป็น 10 กลุ่ม แต่ละกลุ่มมีจำนวนข้อมูลเท่ากัน ฝึกฝนแบบจำลองจากข้อมูล 9 กลุ่ม แล้วใช้ 1 กลุ่มที่เหลือสำหรับทดสอบวนซ้ำการทำงานนี้จนกระทั้งข้อมูลทุกกลุ่มถูกใช้เป็นชุดทดสอบหรือครบ 10 ครั้ง โดยหลังจากการทดลองแล้วจะทำการวิเคราะห์ประสิทธิภาพจากการประเมินประสิทธิภาพของตัวแบบ โดยพิจารณาจาก ความถูกต้อง (Accuracy) ดังสมการที่ 1 ความแม่นยำ (Precision) ดังสมการที่ 2 ค่าความระลึก (Recall) และดังสมการที่ 3

$$Accuracy = \frac{TN + TP}{TP + FP + TP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

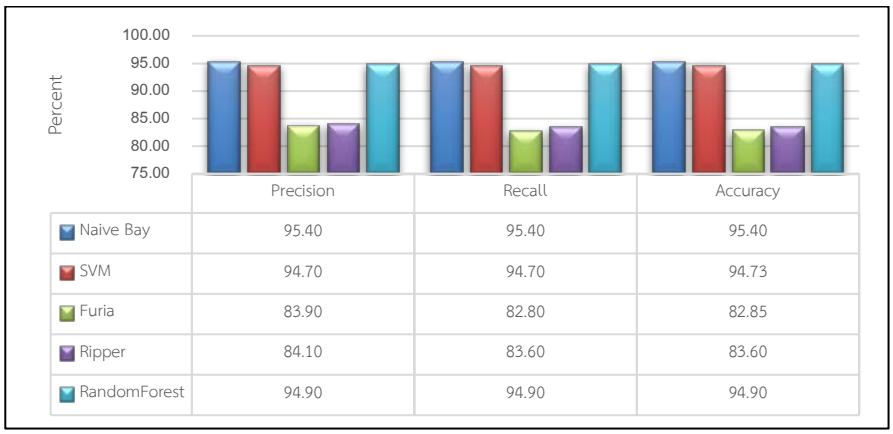
$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

โดยที่ TP คือ ข้อมูลที่ทำนายถูกต้องเมื่อเทียบกับเฉลย FP คือ ข้อมูลที่ทำนายแล้วไม่ถูกต้องเมื่อเทียบกับเฉลย FN คือ ข้อมูลที่อยู่ในเฉลยแต่ไม่มีการทำนาย (ตรงข้าม กับ FN) Precision คือ ค่าความแม่นยำ เกิดจากการนำค่า TP มาเทียบกับ FP Recall คือ ค่าความระลึก เกิดจากการนำค่า TP มาเทียบกับ FN (Angelopoulou et al., 2024; Devi & Sharmila, 2021)

ผลการวิจัย

งานวิจัยนี้ ได้นำความคิดเห็นของคนไทยในสังคมว่า มีความคิดเห็นต่อการฉีดวัคซีนป้องกันโรคโควิด-19 ให้กับเด็กจากโซเชียลมีเดีย ทั้งหมดจำนวน 2,466 ความคิดเห็น ผ่านขั้นตอนการเตรียมข้อมูลทำดัชนี และกำหนดคลาสให้กับความคิดเห็น จนอยู่รูปแบบที่เหมาะสมในการทดลอง จำนวน 1,195 ความคิดเห็น เป็นความคิดเห็นเชิงบวกและลบ จำนวน 638 และ 557 ความคิดเห็น ตามลำดับ และนำข้อมูลที่ได้มาใช้ในการสร้างแบบจำลองเพื่อวิเคราะห์ทั้งสิ้นจาก 5 เทคนิค ได้แก่ นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน เทคนิคริปเปอร์ เทคนิคฟิวเรีย และเทคนิคป่าสุ่ม โดยทำการวิเคราะห์ผ่านการใช้โปรแกรม WEKA เวอร์ชั่น 3.8.6

ในการเปรียบเทียบประสิทธิภาพแบบจำลองใช้การทดสอบแบบจำลองด้วยวิธี 10-fold cross validation ในการแบ่งกลุ่มชุดทดสอบ และชุดเรียนรู้ ซึ่งได้ผลการวัดประสิทธิภาพแบบจำลองด้วยค่าความแม่นยำ ค่าความระลึก และค่าความถูกต้อง ดังแสดงในภาพประกอบ 1



	Precision	Recall	Accuracy
Naive Bay	95.40	95.40	95.40
SVM	94.70	94.70	94.73
Furia	83.90	82.80	82.85
Ripper	84.10	83.60	83.60
RandomForest	94.90	94.90	94.90

ภาพประกอบ 1 แสดงการเปรียบเทียบประสิทธิภาพทั้งหมดของโมเดล

จากภาพประกอบ 1 แสดงค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ของการทดสอบแบบจำลองโดยใช้ของเทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคฟิวเรีย เทคนิคริปเปอร์ และเทคนิคป่าสุ่ม ในการจำแนกความคิดเห็นต่อการฉีดวัคซีนป้องกันโรคโควิด-19 ให้กับเด็ก ผลปรากฏว่า เทคนิคนาอีฟเบย์ ให้ค่าความถูกต้อง ความแม่นยำ และความระลึกมากที่สุด

อภิปรายผลการวิจัย

งานวิจัยนี้ได้มีวัตถุประสงค์ในการสร้างแบบจำลอง และวัดประสิทธิภาพระหว่างแบบจำลองที่ใช้เทคนิคเหมืองข้อความ โดยเป็นการใช้การทำเหมืองข้อความจากความคิดเห็นต่อการฉีดวัคซีนป้องกันโรคโควิด-19 ให้กับเด็ก จากสื่อสังคมออนไลน์ ได้แก่ ยูทูป, ติ๊กตอก และพันทิป ตั้งแต่วันที่ 1 มกราคม 2562 จนถึง 31 ธันวาคม 2563 เป็นความคิดเห็นจำนวน 2,466 ความคิดเห็น ที่มีความยาวมากกว่า 20 คำ ซึ่งการสร้างแบบจำลองของงานวิจัยนี้ได้ใช้คำบงชี้ โดยเป็นคำประเภทคำกริยา คำกริยาวิเศษณ์ คำคุณศัพท์ ที่สามารถแสดงอารมณ์เชิงลบหรือเชิงบวกของความคิดเห็นได้อย่างชัดเจน ส่งผลดีต่อประสิทธิภาพในการเรียนรู้ของแบบจำลองจากข้อมูล สำหรับขั้นตอนในการเปรียบเทียบประสิทธิภาพของแบบจำลอง ใช้วิธี 10-fold cross validation เพื่อวัดประสิทธิภาพแบบจำลองที่พัฒนาจากเทคนิค

ในการจำแนกของ 5 เทคนิควิธี ได้แก่ เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคฟิวเรีย เทคนิคริปเปอร์ และเทคนิคป่าสุ่ม มาสร้างแบบจำลอง

สรุปผล

โดยขั้นตอนการวัดประสิทธิภาพของทั้ง 5 แบบจำลอง ที่ได้จาก 5 เทคนิควิธี จะใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดสำหรับเรียนรู้และทดลอง และวัดประสิทธิภาพของแบบจำลองจาก ค่าความแม่นยำ ค่าความระลึก และค่าความถูกต้อง ผลการวิจัยพบว่า เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน และเทคนิคป่าสุ่ม เป็นกลุ่มเทคนิคที่ให้ประสิทธิภาพที่ดีมากโดยเฉพาะกับข้อมูลการแบ่งคลาสอย่างชัดเจน โดยที่เทคนิคนาอีฟเบย์ โดยที่ให้ค่าความแม่นยำร้อยละ 95.40 ให้ค่าความระลึกร้อยละ 95.40 และให้ค่าความถูกต้องร้อยละ 95.40 ดังเช่นงานวิจัยของ Devi and Sharmila (2021) ได้ใช้เทคนิคนาอีฟเบย์ทดสอบร่วมกับเทคนิคเหมืองข้อมูลอื่นซึ่งผลปรากฏว่า เทคนิคนาอีฟเบย์ให้ประสิทธิภาพในการจำแนกความคิดเห็นมากที่สุดเช่นเดียวกัน

เอกสารอ้างอิง

ปพิชญา กลางนอก, & จารี ทองคำ. (2562). การประยุกต์ใช้เทคนิคแบบรวมเพื่อเพิ่มประสิทธิภาพแบบจำลองตามกฎในเหมืองข้อมูล. วารสารเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏอุบลราชธานี, *9*(1),97-108. https://ph01.tci-thaijo.org/index.php/jitubru/ article/view/183422.

ประกายแก้ว ศิริพูล, ไพรินทร์ ยอดสุบัน, เรืองอุไร อมรไชย, อรรถพงษ์ ฤทธิทิศ, เวหา เกษมสุข, นิตยา บัวสาย, นลินี กินาวงศ์, ชลิตดา ขันแก้ว, & สุภา เพ่งพิศ. (2565). สถานการณ์และปัญหาการรับวัคซีนโควิด 19 ในเด็กอายุ 5-11 ปี ในชุมชน: การวิจัยเชิงคุณภาพ. วารสารการพยาบาลและการดูแลสุขภาพ, *40*(3), 34-43. https://he01.tci-thaijo.org/index.php/jnat-ned/ article/ view/258141.

พิศิษฐ์ บวรเลิศสุ, & วรภัทร ไพรีเกรง. (2565). ตัวแบบการวิเคราะห์ความรู้สึกทางอารมณ์ สำหรับจำแนกประเภทบทความแนะนำสินค้าออนไลน์. *Journal of Engineering and Digital Technology (JEDT)*, *10*(1), 71-79.

สุรัยยา หมานมานะ, โสภณ เอี่ยมศิริถาวร, & สุมนมาลย์ อุทยมกุล. (2563). โรคติดเชื้อไวรัสโคโรนา 2019 (COVID-19). วารสารสถาบันบำราศนราดูร, *14*(2), 124 - 133. https://he01.tci-thaijo.org/index.php/bamrasjournal/article/view/240349.

อุษา คำประสิทธิ์. (2565). การพัฒนารูปแบบการบริหารการพยาบาลในสถานการณ์การระบาดของโควิด-19 แผนกผู้ป่วยใน โรงพยาบาลโนนไทย. วารสารศูนย์อนามัยที่ *9*, *16*.

Angelopoulou, A., Mykoniatis, K., & Smith, A. E. (2024). Analysis of Public Sentiment on COVID-19 Mitigation Measures in Social Media in the United States Using Machine Learning. *IEEE Transactions on Computational Social Systems*, *11*(1), 307-318. https://doi.org/10.1109/TCSS.2022.3214527.

Banluesapy, S., & Jirapanthong, W. (2022). Towards Machine Learning Algorithm for Screening Prediction of COVID-19 Patients. *JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY*, *12*(1), 47-60. https://doi.org/10.14456/jist.2022.5.

Devi, N. S., & Sharmila, K. (2021). Fine Grainded Sentiment Analysis on COVID-19 Vaccine. *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)*.

Karthika, P., Murugeswari, R., & Manoranjithem, R. (2019). Sentiment Analysis of Social Media Network Using Random Forest Algorithm. *2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*.

Khan, F. H., Bashir, S., & Qamar, U. (2014). TOM: Twitter opinion mining framework using hybrid classification scheme. *Decision Support Systems*, *57*, 245-257. https://doi.org/https://doi.org/10.1016/j.dss.2013.09.004.

Kovacs, E. R., Cotfas, L. A., Delcea, C., & Florescu, M. S. (2023). 1000 Days of COVID-19: A Gender-Based Long-Term Investigation into Attitudes With Regards to Vaccination. *IEEE Access*, *11*, 25351-25371. https://doi.org/10.1109/ACCESS.2023.3254503.

Phung, T. K., Te, N. A., & Ha, T. T. T. (2021). A machine learning approach for opinion mining online customer reviews. 2021 21st ACIS International Winter Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD-Winter).

Suharsono, T. N., Fauzan, A., & Mardiati, R. (2022). Sentiment Analysis of Covid-19 on Indonesian Twitter by Implementing the Naïve Bayes Method. *2022 8th International Conference on Wireless and Telematics (ICWT)*.

Vyas, P., Reisslein, M., Rimal, B. P., Vyas, G., Basyal, G. P., & Muzumdar, P. (2022). Automated Classification of Societal Sentiments on Twitter With Machine Learning. *IEEE Transactions on Technology and Society*, *3*(2), 100-110. https://doi.org/10.1109/TTS.2021.3108963.

Zope, T., & Rajeswari, K. (2022). Sentiment Analysis of Covid-19 Tweets using Twitter Database– A Global Scenario. *2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA)*.