

ชนิดของคำภาษาไทยที่ส่งผลต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็น: กรณีศึกษาของความคิดเห็นของคนไทยต่อโรคโควิด 19

Types of Thai words Affecting the Effectiveness of Opinion Classification Models: Case Study of Thai People's Opinions on COVID-19

ศรวุฒิ เกิดถาวร* และ จารี ทองคำ

Sarawut Kedtrawon* and Jaree Thongkam

กลุ่มสารสนเทศเชิงประยุกต์ ภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม

Applied Information Group, Department of Information Technology, Faculty of Informatics,

Maharakham University,

*E-mail: topzaneverdie@gmail.com

Received: Oct 08, 2021

Revised: Feb 07, 2022

Accepted: Feb 17, 2022

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาผลกระทบของชนิดของคำภาษาไทยที่มีต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็นที่ใช้กับความคิดเห็นของคนไทยต่อโรคโควิด 19 จากสื่อสังคมออนไลน์ ข้อมูลความคิดเห็นเกี่ยวกับโรคโควิด 19 ที่เป็นข้อความภาษาไทยถูกเก็บจากเว็บไซต์ทวิตเตอร์ และพันทิป จำนวนทั้งหมด 2,920 ความคิดเห็น แล้วนำกระบวนการเหมืองข้อความมาแบ่งข้อมูลออกเป็น 2 ชุดข้อมูล คือ คำวิเศษณ์ และคำกริยา ซึ่งเป็นคำที่บ่งบอกว่าข้อมูลเป็นความคิดเห็นเชิงบวก หรือเชิงลบ นอกจากนี้แล้วในงานวิจัยนี้ยังได้เพิ่มข้อมูลอีกหนึ่งชุด ซึ่งเป็นกลุ่มที่มีทั้งคำวิเศษณ์ และคำกริยา ดังนั้นจึงมีข้อมูลทั้งสิ้น 3 ชุด คือ ชุดข้อมูลที่มีแต่คำวิเศษณ์ ชุดข้อมูลที่มีแต่คำกริยา และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา เพื่อใช้ศึกษาผลกระทบของชนิดของคำไทยต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็น ในงานวิจัยนี้ได้นำ 5 เทคนิคมาสร้างแบบจำลอง ได้แก่ เทคนิคต้นไม้ตัดสินใจ 4.5 (C4.5) เทคนิคนาอ็ฟเบย์ (NB) เทคนิคเคเนียร์เรสเนเบอร์ (KNN) เทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) และเทคนิคระบบเรียนรู้เชิงลึก (DL) ในการทดสอบประสิทธิภาพของแบบจำลอง ได้ใช้ 10-โฟลด์ครอสวาไลเดชัน เป็นวิธีการทดสอบ และใช้ความถ่วงดุล ความแม่นยำ ความครบถ้วน และค่าสัมประสิทธิ์สหสัมพันธ์แมทริวส์ เป็นเกณฑ์ในการทดสอบ ผลการทดสอบพบว่าเทคนิค MLP มีค่าวัดประสิทธิภาพสูงสุดถึงร้อยละ 100 สำหรับทุกเกณฑ์การทดสอบเมื่อใช้ข้อมูลที่มีเฉพาะคำวิเศษณ์ และเทคนิค KNN มีประสิทธิภาพสูงสุดถึงร้อยละ 100 สำหรับทุกเกณฑ์การทดสอบเมื่อใช้ข้อมูลที่มีคำวิเศษณ์ และคำกริยารวมกัน งานวิจัยนี้ชี้ให้เห็นว่าชนิดของคำไทยส่งผลต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็น

คำสำคัญ: แบบจำลองจำแนกความคิดเห็น เหมืองข้อความ โควิด 19 ชนิดของคำภาษาไทย

Abstract

The purpose of this research was to study the effect of types of Thai words on the effectiveness of opinion classification models used for Thai people's opinions on COVID-19 from social media. The sentiments about COVID-19 in Thai language were collected from Twitter and Pantip websites up to 2,920 sentiments. The text mining processes were utilized to extract the data into 2 data sets including adverbs and verbs that indicated whether the data were positive or negative opinions. Furthermore, this research also added one more set of data containing both adverbs and verbs. Therefore, there were a total of 3 data sets containing only adverbs, only verbs, and both adverbs and verbs for examining the effect of types of Thai words on the effectiveness of the opinion classification models. In this research, 5 techniques were used to build the models including Decision Tree C4.5 (C4.5), Naive Bayes (NB), K-Nearest Neighbor (KNN), Multi-layer Perceptron (MLP) and Deep learning (DL). To evaluate effectiveness of the models, the 10-fold cross validation was used as the evaluation method and F-Measure, precision, recall and Matthews correlation coefficient were used as evaluation criteria. The results demonstrated that technique MLP had the highest effectiveness scores of 100% for all of the criteria when using the data set containing only adverbs and technique KNN had the highest effectiveness scores of 100% for all of the criteria when using the data set containing only adverbs and verbs. This research indicated that types of Thai words affected the effectiveness of opinion classification models.

Keywords: Opinion classification models, Text mining, COVID-19, Types of Thai words

1. บทนำ

การระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 (COVID-19) ไม่เคยปรากฏมาก่อน [1] ส่งผลกระทบอย่างรุนแรงต่อสุขภาพ เศรษฐกิจ และสังคม เมื่อวันที่ 23 มกราคม 2563 ในช่วงการแพร่ระบาดของเชื้อไวรัส COVID-19 ในประเทศไทย เปลี่ยนแปลงวิถีการดำเนินชีวิตของประชาชน และมีมาตรการป้องกัน เช่น ทำงานที่บ้าน (Work from home) การเว้นระยะห่างระหว่างบุคคล (Social distancing) ถือเป็นมาตรการที่สำคัญที่จะช่วยลดการแพร่กระจายของเชื้อไวรัสที่ได้ผลดี [2] ในช่วงวิกฤตการระบาดของโรคทำให้ประชาชนมีแนวโน้มที่จะใช้แพลตฟอร์มสื่อสังคมออนไลน์ เพื่อที่จะได้รับข้อมูลที่จำเป็นและแลกเปลี่ยนความคิดเห็น ซึ่งแน่นอนว่าจะต้องมีเสียงและการวิพากษ์วิจารณ์บางอย่างเกิดขึ้นบนสื่อสังคมออนไลน์ ตลอดเวลาที่วิกฤตครั้งนี้ดำเนินไป

เหมืองความคิดเห็น (Opinion mining) [3] เป็นกระบวนการวิเคราะห์ความคิดเห็นที่มีความคิดเห็นของคน นำเสนอในรูปแบบของข้อความ ซึ่งพบได้บนเครือข่ายสังคมออนไลน์ เพื่อให้ทราบถึงความพึงพอใจในเชิงบวกหรือเชิงลบ การจัดการข้อความที่ซับซ้อนด้วยความคิดเห็นนั้นสามารถนำไปประยุกต์ใช้กับงานที่สำคัญหลายประเภท มีนักวิจัยหลายท่านได้นำเทคนิคการทำเหมืองความคิดเห็นมาใช้ในการจำแนกข้อความ เช่น สุภัสสรา สมเจตนา และจारी ทองคำ [4] เปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร โดยใช้เฉพาะคำวิเศษณ์ใช้ เทคนิคริปปเปอร์ (Ripper) เทคนิคต้นไม้ตัดสินใจซี 4.5 (Decision tree C4.5) เทคนิคนาอิวเบย์ (Naive Bayes) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เทคนิคเคเนียร์เรสเนเบอร์ (K-

nearest neighbor) และ เทคนิค ต้นไม้ป่าสุ่ม (Random forest) ให้ค่าประสิทธิภาพในการจำแนกคุณลักษณะที่ดีที่สุดมีค่าความแม่นยำ 89.62% สถิติช็อกโพรสเซอร์ และสถิติภูมิ โพรสวัง [5] ได้เสนอการจำแนกพฤติกรรมการขับชิ่งโดยสารสนเทศโดยใช้วิธีการสกัดข้อความและเทคนิคการเรียนรู้ของเครื่องเพื่อจำแนกประเภทข้อร้องเรียนแบบอัตโนมัติ ที่เกี่ยวกับพฤติกรรมการขับชิ่งและคุณภาพการบริการ จำนวน 4 คลาส ผลการทดลองพบว่าตัวแบบโครงข่ายประสาทเทียมสามารถจำแนกข้อความพฤติกรรมการขับชิ่งโดยสารสนเทศได้ความครบถ้วนสูงสุดซึ่งมีค่าเท่ากับ 90.23% ความแม่นยำ เท่ากับ 91.9% ความครบถ้วน เท่ากับ 90.2% และ ความถ่วงดุลเท่ากับ 90.5% อีรียุทธ คุณสุข และจารี ทองคำ [6] ได้เสนอกระบวนการคัดเลือกคุณลักษณะสำหรับเพิ่มประสิทธิภาพในการจำแนกความคิดเห็นของลูกค้าร้านอาหารข้อมูลจากเว็บไซต์ wongnai.com จำนวน 4,487 ข้อความ ใช้เทคนิค 3 ประการในการเลือกคุณสมบัติข้อความ ได้แก่ ไค สแควร์ (Chi-Square), อิน ฟอ เม ชัน เคน (Information gain) และ อินฟอ เม ชัน เคน เรดิโอ (Information gain ratio) เพื่อวัดประสิทธิภาพของเทคนิคการเลือกคุณสมบัติและใช้เทคนิคนาอ็ฟเบย์, เทคนิคซัพพอร์ตเวกเตอร์แมชชีน, เทคนิคเคเนียร์เรสเนเบอร์และ เทคนิคต้นไม้ตัดสินใจ 4.5 เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) สำหรับการจัดประเภทไอออนบวก นอกจากนี้ยังมีการใช้ 10-โพลด์ครอสวาเลชัน (10-fold cross validation) เพื่อแบ่งข้อมูลออกเป็นชุดการเรียนรู้และวัดความถูกต้อง ความแม่นยำ และการเรียกคืน จากการทดลองพบว่าเทคนิคการเลือกพีเจอร์ อินฟอ เม ชัน เคน ร่วมมือกับเทคนิคนาอ็ฟเบย์ ให้ผลลัพธ์ที่ดีที่สุดในการจำแนกประเภทของความคิดเห็นโดยความถูกต้องคือ 89.08% ความแม่นยำเท่ากับ 89.12% และการเรียกคืนเท่ากับ 89.10% Supianto และคณะ [7] ใช้เทคนิคการจำแนกต้นไม้แบบสุ่ม และ ต้นไม้ตัดสินใจแบบซี 4.5 ความแม่นยำเฉลี่ยที่สูงขึ้นใช้เทคนิคต้นไม้ตัดสินใจแบบซี 4.5

มากที่สุด มีความแม่นยำ 77.01% เทคนิคต้นไม้ป่าสุ่มตามลำดับ Chen และคณะ [8] ใช้เทคนิคแบบระบบเรียนรู้เชิงลึก เทคนิคซัพพอร์ตเวกเตอร์แมชชีน และเทคนิคเพอร์เซปตรอนแบบหลายชั้น (Multi-layer perceptron) พบว่า เทคนิคแบบระบบเรียนรู้เชิงลึก ที่มีประสิทธิภาพที่เหนือกว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีน และเทคนิคเพอร์เซปตรอนแบบหลายชั้น ในการเข้าชมแพลตฟอร์มโซเชียลมีเดีย Suksangaram และคณะ [9] ใช้เทคนิคต้นไม้ตัดสินใจแบบซี 4.5, เทคนิคนาอ็ฟเบย์ และ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน ในการทำนายการเรียนรู้ขององค์กรที่มีผลต่อการปฏิบัติงานของพนักงานธนาคารเพื่อการเกษตรและสหกรณ์การเกษตรในภาคตะวันตก ผลการวิจัยพบว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีน มีค่าความแม่นยำ 98.33% ความแม่นยำ 0.025 และการเรียกคืน 0.984 ที่มากกว่าเทคนิคต้นไม้ตัดสินใจแบบซี 4.5 และ เทคนิคนาอ็ฟเบย์ นุชนาฏ ปิ่นเมือง และ จารี ทองคำ [10] ได้ทำการวิจัยเกี่ยวกับความคิดเห็นของคนไทยต่อสื่อออนไลน์ด้วยเทคนิค 5 เทคนิค อันได้แก่ เทคนิคนาอ็ฟเบย์ เทคนิคจำแนกเชิงเส้น เทคนิคเคเนียร์เรสเนเบอร์ เทคนิคต้นไม้ตัดสินใจและเทคนิคต้นไม้ตัดสินใจซี 4.5 พบว่า เทคนิคนาอ็ฟเบย์ มีประสิทธิภาพในการจำแนกความคิดเห็น โดยมีค่าความถูกต้องมากถึงร้อยละ 93.88 และค่าความแม่นยำร้อยละ 94.02 ส่วน Gupta และคณะ [11] ได้ทำเหมืองความคิดเห็นของลูกค้าที่มาใช้บริการโรงแรมและให้คะแนนการบริการ กลับพบว่าเทคนิคต้นไม้ตัดสินใจมีความ แม่นยำถึง 76.22 เปอร์เซนต์ และง่ายต่อการเข้าใจของพนักงานโรงแรมสรุปเทคนิค 5 เทคนิคที่เป็นที่นิยมจากงานวิจัยที่เกี่ยวข้อง เทคนิคต้นไม้ตัดสินใจซี 4.5 เทคนิคนาอ็ฟเบย์ เทคนิคเคเนียร์เรสเนเบอร์ เทคนิคเพอร์เซปตรอนแบบหลายชั้น เทคนิคแบบระบบเรียนรู้เชิงลึก

ดังนั้นผู้วิจัยมีแนวคิดศึกษาชนิดของคำภาษาไทยที่ส่งผลต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 จากสื่อสังคมออนไลน์ โดยการแบ่งชุดข้อมูลด้วยการกำหนด

ชนิดของคำภาษาไทยแบ่งชุดข้อมูลออกเป็น 3 ชุดข้อมูล ได้แก่ ชุดข้อมูลคำวิเศษณ์ ชุดข้อมูลคำกริยา ชุดข้อมูล คำวิเศษณ์และคำกริยา เพื่อหาประสิทธิภาพของชุด ข้อมูลที่ดีที่สุด เพื่อนำไปใช้กับเทคนิคเหมืองข้อมูลใน การสร้างแบบจำลองจำแนกความคิดเห็นของ ความ คิดเห็นของคนไทยต่อโรคโควิด 19 บนสื่อสังคม ออนไลน์ ซึ่งประกอบด้วย 5 เทคนิค คือ เทคนิคต้นไม้ ตัดสินใจ 4.5 เทคนิคนาอิวเบย์ เทคนิคเคเนียร์เรสเน เบอร์ เทคนิคเพอร์เซปตรอนแบบหลายชั้น เทคนิคแบบ ระบบเรียนรู้เชิงลึก จากนั้นใช้หลักการ 10-โฟลด์ครอส วาลิเดชัน ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้ และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของ แบบจำลองด้วย ความถ่วงดุล (F-Measure) ความ แม่นยำ (Precision) ความครบถ้วน (Recall) และค่า สัมประสิทธิ์สหสัมพันธ์แมทริคซ์ (MCC)

2. วิธีดำเนินการวิจัย

งานวิจัยฉบับนี้ มีอุปกรณ์ที่ใช้คือ เครื่อง คอมพิวเตอร์ รองรับ ระบบปฏิบัติการ Windows 10, โปรแกรม Microsoft Excel, โปรแกรม google Colaboratory, โปรแกรม weka และกระบวนการ เหมืองความคิดเห็นซึ่งมี 5 ขั้นตอน รวมถึง การรวบรวม ข้อมูล การเตรียมข้อมูล การจัดทำดัชนีข้อมูล การสร้าง แบบจำลอง และการวัดประสิทธิภาพแบบจำลอง

2.1. การรวบรวมข้อมูล

งานวิจัยฉบับนี้ได้รวบรวมความคิดเห็นของคน ไทย ต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ เว็บไซต์ทวิต เทอร์ และเว็บไซต์พันทิป ตั้งแต่วันที่ 23 มกราคม 2563 เริ่มต้นโรคโควิด 19 ในประเทศไทย ถึงวันที่ 1 พฤษภาคม 2563 จำนวน 2,920 ความคิดเห็น การดึง ข้อมูลจากทวิตเตอร์ต้องมีการดึงผ่าน authentication code ที่สร้างขึ้นมายืนยันตัวตนแล้วใช้งานคำสั่ง API และการดึงข้อมูลจากเว็บไซต์พันทิป ใช้เทคนิค Web scraping โดยใช้ไลบรารีของ selenium, beautiful soup4 ในภาษา Python มาเก็บไว้ใน Microsoft Excel ดังแสดงใน Table 1

Table 1 Examples of format data in Microsoft Excel

Format Data in Microsoft Excel	
1	กัปตันอเมริกันนำทัพดำหลังทรมับบอกประชาชนว่าไม่ต้องกลัวเพราะตัวเองติด
2	RSVและโรคทางเดินหายใจในช่วงที่มีการระบาดของใหม่ ๆ มีการป้องกันกัน
3	วัคซีนโควิด19การศึกษาระยะสุดท้ายแบบสมบูรณ์ คงจะมีผลประกาศออกมาในเดือนหน้าเชื่อว่า
4	Facebookเตรียมบริจาคเงิน40ล้านดอลลาร์ช่วยเหลือธุรกิจคนผิวดำจากพิษFacebook
5	พม่ามีผู้ป่วยเพิ่ม1010รายสรุปผู้ป่วยทั้งหมด14383 ราย

2.2. การเตรียมข้อมูล

งานวิจัยฉบับนี้ได้เตรียมข้อมูลจากสื่อสังคม ออนไลน์ และเป็นขั้นตอนที่สำคัญขั้นตอนหนึ่งในการทำ เหมืองความคิดเห็น ประกอบด้วย การทำความสะอาด ข้อมูล และการตัดคำ

2.2.1. การทำความสะอาดข้อความ

ข้อความที่นำมาจากสื่อสังคมออนไลน์ ต้อง ตัดออก เช่น เครื่องหมายคำพูด การใช้ข้อความไม่ถูก หลักภาษาไทย มีเครื่องหมายอักขระพิเศษ จึง จำเป็นต้องทำความสะอาดก่อน และคงไว้เฉพาะตัว อักขระ พยัญชนะ ภาษาไทย โดยใช้ regular expression ด้วยโปรแกรม colab google ภาษา Python ใช้ไลบรารีของ re, emoji ผลจากการทำความสะอาด ดังแสดงใน Table 2

Table 2 Examples of text cleaning

Text cleaning
1 กับตันอเมริกานำทัพดำหลังทรมัปปอกประชาชนว่า ไม่ต้องกลัวเพราะตัวเองติด
2 โรคทางเดินหายใจในช่วงที่มีการระบาดของใหม่ ๆ มีการป้องกัน
3 วัคซีนโควิดการศึกษาจะสุดท้ายแบบสมบูรณ์คง จะมีผลประกาศออกมาในเดือนหน้า
4 เตรียมบริจาคเงินล้านดอลลาร์ช่วยเหลือธุรกิจคนผิวดำ
5 พม่ามีผู้ป่วยเพิ่มสรูปผู้ป่วยทั้งหมด

2.2.2. การตัดคำ

การตัดคำ (Word segmentation) งานวิจัยฉบับนี้มีขั้นตอนการแบ่งตัวอักษรจากข้อความ (string) เพื่อหาขอบเขตของแต่ละหน่วยคำ (morpheme) เพื่อให้การจำแนกข้อความมีประสิทธิภาพ ด้วยโปรแกรม colab google ภาษา Python ใช้ไลบรารีของ PyThaiNLP ผลที่ได้ดังแสดงใน Table 3

Table 3 Examples of word segmentation

Word segmentation
1 กับตันอเมริกา, นำทัพ, ดำ, หลัง, ทรมัปป, บอก, ประชาชน, ว่า, ไม่ต้องกลัว, เพราะตัวเองติด
2 โรคทางเดินหายใจ, ในช่วง, ที่มี, การระบาด, ของใหม่ๆ, มีการป้องกัน
3 วัคซีนโควิด, การศึกษา, ระยะสุดท้าย, แบบ สมบูรณ์, คงจะมีผล, ประกาศ, ออกมา, ในเดือน หน้า
4 เตรียม, บริจาค, เงิน, ล้าน, ดอลลาร์, ช่วย, ธุรกิจ, คนผิวดำ
5 พม่า, มี, ผู้ป่วย, เพิ่ม, สรูป, ผู้ป่วย, ทั้งหมด

2.3. การจัดทำดัชนีข้อมูล

การจัดทำดัชนีข้อมูลทำโดยการนำข้อมูลมาแยกประเภทของคำหรือชนิดของคำที่อยู่ในประโยค เช่น คำนาม (Noun) คำสรรพนาม (Pronoun) คำกริยา

(Verb) คำวิเศษณ์ (Adverb) คำบุพบท (Preposition) คำสันธาน (Conjunction) ดังแสดงใน Table 4

Table 4 Examples of Word type classification

Word	Word type
เครียดมาก	Adverb
จุดด้อย	Adverb
เข้าใจกัน	Adverb
เลิกหวัง	Verb
ขัดขวาง	Verb
ภูมิภาค	Noun
สัญชาติ	Noun
ใกล้เคียง	Preposition
ไปไม่ได้	Verb

เนื่องจาก ประโยค วลี คำพูด ต่าง ๆ ที่ใช้สื่อสารกันล้วนเกิดขึ้นจากการนำคำต่าง ๆ มาประกอบกันเป็น ส่วนต่าง ๆ ที่ทำหน้าที่ต่างกันประโยคแล้วนำไป สร้างคลังคำศัพท์ และการกำหนดคลาส

2.3.1. การสร้างคลังคำศัพท์

การสร้างคลังคำศัพท์ เป็นการนำคำที่แยกประเภท ในกรณีนี้เข้ากันนำมาเพียงหนึ่งคำ ส่วนกรณีคำที่ไม่เข้ากันนำมาทั้งหมด คงเหลือทั้งหมด 9,037 คำ ด้วยโปรแกรม colab google ภาษา Python ไลบรารีของ Scikit-Learn CountVectorizer และ PyThaiNLP Part-of-Speech โดยการใช้พจนานุกรมไทยในการแยกชนิดของคำ จากนั้นทำการคัดเลือกเอาเฉพาะคำกริยา คำวิเศษณ์ ที่เป็นคำที่บ่งบอกความรู้สึกได้ดี [12] แล้วทำการคัดเลือกคำคุณลักษณะที่มีความหมายเชิงบวก เชิงลบแล้วแทนค่าของจำนวนคำคุณลักษณะที่ไม่ปรากฏเป็น 0 แทนค่าจำนวนคำคุณลักษณะเชิงบวกที่ปรากฏเป็น 1 และแทนค่าจำนวนคำคุณลักษณะเชิงลบที่ปรากฏเป็น -1 จะเห็นความถี่ของคำคุณลักษณะที่ถูกจำแนกออกมาเป็นเชิงบวก หรือเชิงลบ ในแต่ละความ

คิดเห็น และทำการแบ่งความคิดเห็นเชิงบวกและเชิงลบ โดยใช้หลักเกณฑ์ของภาษาไทยดังแสดงใน Table 5

Table 5 Examples of feature word selection

Word	Word type	Feature word
เครียดมาก	Adverb	-1
จุดด้อย	Adverb	-1
เข้าใจกัน	Adverb	1
เหมือนกัน	Adverb	1
เลิกหวัง	Verb	-1
ขัดขวาง	Verb	-1
ไปไม่ได้	Verb	-1
มีความหวัง	Verb	1
ยิ้ม	Verb	1
มาตรการป้องกัน	Adverb	1
ไม่ยอมให้	Verb	-1
อาจตกงาน	Verb	-1
อ่านไม่ออก	Verb	-1
วูบ	Verb	-1
ยอมแพ้	Verb	-1

จากวัตถุประสงค์งานวิจัยนี้ต้องการศึกษาชนิดของคำที่ส่งผลกระทบต่อประสิทธิภาพการจำแนก คณะผู้วิจัยจึงได้ทำการแบ่ง 3 ชุดข้อมูลดังนี้ ชุดข้อมูลคำวิเศษณ์ จำนวน 314 คำ ชุดข้อมูลคำกริยา จำนวน 1,604 คำ ชุดข้อมูลคำวิเศษณ์และคำกริยา จำนวน 1,918 คำ มาเป็นคุณลักษณะ ระบุตามความหมาย เชิงบวก หรือเชิง

ลบ ที่เหลือไม่เอามาใช้ คงเหลือคำที่เป็นคุณลักษณะเชิงบวกและเชิงลบ ที่สามารถนำไปใช้ในงานวิจัยนี้ จำนวน 236 คำ 254 คำ 490 คำ ตามลำดับ โดยแบ่งคำออกเป็น 2 กลุ่ม คือคำแทนคุณลักษณะเชิงบวกและคำแทนคุณลักษณะเชิงลบ

2.3.2. การกำหนดคลาส

การกำหนดคลาสทำโดยการนำค่าบ่งชี้คุณลักษณะในชุดข้อมูลมาเรียงกันเพื่อทำการนับความถี่ของการเกิดขึ้นของคำนั้น ๆ จากนั้นจึงนำค่าจำนวนความถี่ของคำมาสร้างเวกเตอร์ตัวแทนเอกสาร และค่าบ่งชี้ที่ไม่ปรากฏในเอกสารจะมีค่าเป็น 0 จากนั้นทำการนับจำนวนคำในแต่ละคุณลักษณะ โดยใช้การนับจำนวนความถี่ของค่าคุณลักษณะว่ามีจำนวนเท่าใด เพื่อนำมาเปรียบเทียบกัน เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงบวกมากกว่าความถี่ของคุณลักษณะเชิงลบให้ตัวแปรตามเป็นความคิดเห็นเชิงบวก แทนด้วย P เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงลบมากกว่าความถี่ของคุณลักษณะเชิงบวกให้ตัวแปรตามเป็นความคิดเห็นเชิงลบ แทนด้วย N แต่ถ้าหากความถี่ของคุณลักษณะเชิงบวกและเชิงลบเท่ากัน จะแทน ด้วย NE และข้อความจะถูกตัดออก จึงทำการลบแถวตั้งที่ข้อความ P และ N ออก จึงนำข้อมูลดังกล่าวไปสร้างแบบจำลองดังแสดงใน Table 6

Table 6 Examples of opinion classification

Message	เครียดมาก	อันตรายมาก	เลิกหวัง	ขัดขวาง	ไปไม่ได้	มีความหวัง	ยิ้ม	P	N	Class
1	-1	0	0	0	0	0	0	0	-1	N
2	0	0	-1	0	0	0	0	0	-1	N
3	0	0	0	-1	0	0	0	0	-1	N
4	0	0	0	0	0	1	0	1	0	P
5	0	0	0	0	0	0	0	0	0	NE

2.4. การสร้างแบบจำลอง

งานวิจัยฉบับนี้ได้จำแนกความคิดเห็น ให้การจำแนกความคิดเห็นที่เหมาะสมที่สุดต่อชนิดของคำภาษาไทยที่ส่งผลกระทบต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 จากสื่อสังคมออนไลน์ โดยใช้เทคนิคการของการทำเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 5 เทคนิค ด้วยโปรแกรม Weka ดังนี้

เทคนิคต้นไม้ตัดสินใจ 4.5 (C4.5) เป็นตัวแบบในการแบ่งประเภทโดยนำ คุณลักษณะต่าง ๆ ของข้อมูลนำเข้ามานำเป็นน้ำหนักในแต่ละโหนดของโครงสร้างข้อมูลต้นไม้โหนดล่างสุดหรือใบ จะเป็นผลลัพธ์หรือประเภทที่จัดแบ่ง

เทคนิคนาอิวเบย์ (NB) [13] เป็นตัวแบบที่ทำให้ความน่าจะเป็นกับข้อมูลนำเข้าเพื่อจัดแบ่งประเภทข้อมูลนำเข้าเป็นรูปแบบที่ง่ายที่สุดรูปแบบหนึ่งของอัลกอริทึมการแบ่งประเภท

เทคนิคเคเนียร์เรสนเบอร์ (KNN) [14] เป็นการจำแนกกลุ่มข้อมูลที่มีคุณสมบัติใกล้เคียงกันที่สุด K กลุ่ม โดยการคำนวณค่าระยะห่างระหว่างจุด (Euclidean Distance) ซึ่งเป็นการวัดระยะห่างระหว่างข้อมูลสองข้อมูล ถ้าข้อมูลมีความคล้ายคลึงกันมากจะมีระยะห่างน้อย

เทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) [15] คือ เทคนิคหนึ่งในโครงข่ายประสาทเทียม ที่มีโครงสร้างเป็นแบบหลาย ๆ ชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี และใช้การคำนวณด้วยสูตรทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศโดยการจัดเป็นลำดับชั้น (layer) โดยมีข้อมูลนำเข้า และสุดท้ายออกเป็นผลลัพธ์เพื่อจำแนกข้อมูล

เทคนิคแบบระบบเรียนรู้เชิงลึก (DL) คือ เป็นส่วนหนึ่งของวิธีการการเรียนรู้ของเครื่องบนพื้นฐานของโครงข่ายประสาทเทียมหรือข่ายงานประสาทเทียมหลาย ๆ ชั้นเหมือนแบบจำลองอันเรียบง่ายของสมองมนุษย์ และการเรียนเชิงคุณลักษณะ การเรียนรู้สามารถเป็นได้ทั้งแบบการเรียนรู้แบบมีผู้สอน การเรียนรู้แบบ

กึ่งมีผู้สอน และการเรียนรู้แบบไม่มีผู้สอนคำว่า ลึก ในความหมายมาจากการที่มีชั้นของโครงข่ายหลายชั้น ที่มีประสิทธิภาพมากการเรียนรู้ที่สะดวกขึ้น และการเข้าใจในโครงสร้างที่ชัดเจนขึ้น

2.5. การวัดประสิทธิภาพของแบบจำลอง

งานวิจัยฉบับนี้ใช้การวัดประสิทธิภาพของแบบจำลองด้วยวิธี 10-โฟลด์ครอสวาเลชัน จะทำการเลือกสุ่มข้อมูลออกเป็น K ชุดเท่า ๆ กัน จากนั้นจะทำการทดลองครั้งแรกด้วยข้อมูลชุดที่ 1 ซึ่งเป็นข้อมูลทดสอบและกำหนดให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั่งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ โดยวิธีการทดสอบประสิทธิภาพแบบ 10-โฟลด์ครอสวาเลชัน มีข้อเสียคือ จะต้องทำการเริ่มทดสอบใหม่โดยจะต้องทำทั้งหมด K รอบ

การวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาความถ่วงดุล ความแม่นยำ ความครบถ้วน และค่าสัมประสิทธิ์สหสัมพันธ์แมทริส โดยความถ่วงดุล (F-Measure) คือค่าประสิทธิภาพโดยรวมที่ได้จากค่าความแม่นยำและความครบถ้วน พร้อมกันของแบบจำลอง โดยพิจารณาแยกทีละคลาส ดังสมการที่ 1

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (1)$$

โดยที่

Precision = ความแม่นยำของแบบจำลอง

Recall = ความครบถ้วนของแบบจำลอง

ความแม่นยำ (Precision) คือ ค่าความสามารถในการที่จะจัดเอกสารที่ไม่เกี่ยวข้องออกไปโดยที่ความแม่นยำนั้นจะเป็นอัตราส่วนของจำนวนเอกสารที่เกี่ยวข้องและได้มีการถูกดึงออกมาเพื่อเทียบกับจำนวนเอกสารที่ถูกดึงออกมาทั้งหมด ดังสมการที่ 2

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

โดยที่

TP = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ถูกต้อง

FP = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ไม่ถูกต้อง

ความครบถ้วน (Recall) เป็นการวัดความสามารถของระบบในการดึงเอกสารที่เกี่ยวข้องออกมา โดย recall เป็นอัตราส่วนของจำนวนเอกสารที่เกี่ยวข้องและถูกดึงออกมากับจำนวนเอกสารที่เกี่ยวข้องทั้งหมด ดังสมการที่ 3

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

โดยที่

TP = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ถูกต้อง

FN = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ไม่ถูกต้อง

ค่าสัมประสิทธิ์สหสัมพันธ์แมทธิวส์ (Matthews correlation coefficient: MCC) หรือค่าสัมประสิทธิ์ phi ใช้ในการเรียนรู้ของเครื่องเป็นการวัดคุณภาพของไบนารี (สองชั้น) ค่าสัมประสิทธิ์คำนึงถึงผลบวกและเชิงลบที่เป็นจริงและเท็จและโดยทั่วไปถือว่าเป็นหน่วยวัดที่สมดุลซึ่งสามารถใช้ได้แม้ว่าชั้นเรียนจะมีขนาดแตกต่างกันมากก็ตาม MCC เป็นค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างการจำแนกประเภทไบนารีที่สังเกตและทำนาย จะส่งคืนค่าระหว่าง -1 และ +1 ค่าสัมประสิทธิ์ +1 หมายถึงการทำนายที่สมบูรณ์แบบ 0 ไม่ได้ไปกว่าการทำนายแบบสุ่มและ -1 หมายถึงความไม่เห็นด้วยทั้งหมดระหว่างการทำนายและการสังเกต MCC ดังสมการที่ 4

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP)}} \quad (4)$$

โดยที่

TP = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ถูกต้อง

TN = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ถูกต้อง

FN = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ไม่ถูกต้อง

FP = จำนวนข้อมูลที่แบบจำลองพยากรณ์ได้ไม่ถูกต้อง

3. ผลการวิจัยและอภิปรายผลการวิจัย

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาชนิดของคำภาษาไทยที่ส่งผลกระทบต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยผู้วิจัยได้ทำการเก็บรวบรวมจากสื่อสังคมออนไลน์ เริ่มต้นโรคโควิด 19 ในประเทศไทย วันที่ 23 มกราคม 2563 ถึงวันที่ 1 พฤษภาคม 2563 โดยคัดเลือกจากคำสำคัญ COVID19 โควิด19 และไวรัสโคโรนาสายพันธุ์ใหม่ ข้อความความคิดเห็น จำนวนทั้งหมด 2,920 ความคิดเห็น แล้วนำกระบวนการเหมืองข้อความมาทำการสกัดคุณลักษณะ แบ่งออกเป็น 2 กลุ่มที่บ่งบอกถึงความคิดเห็นเชิงบวก หรือเชิงลบ ได้แก่ คำวิเศษณ์ และคำกริยา ยิ่งไปกว่านั้นนักวิจัยได้นำคำวิเศษณ์ร่วมกับคำกริยา จึงทำให้ได้ชุดข้อมูล 3 ชุดข้อมูล คือ ชุดข้อมูลที่มีแต่คำวิเศษณ์จำนวน 314 คำ ชุดข้อมูลที่มีแต่คำกริยาจำนวน 1,604 คำ และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยาจำนวน 1,918 คำ เพื่อหาชนิดของคำที่ส่งผลกระทบต่อประสิทธิภาพแบบจำลอง พบว่ามีข้อความที่มีความคิดเห็นเชิงบวกหรือเชิงลบ การระบุจะพิจารณาเป็นคำ ๆ โดยไม่สนใจประโยค แล้วจะคงเหลือคำที่บ่งชี้ลักษณะคำ จำนวน 236 คำ 254 คำ 490 คำ ตามลำดับ โดยแบ่งคำออกเป็น 2 กลุ่ม คือ คำแทนคุณลักษณะเชิงบวกและคำแทนคุณลักษณะเชิงลบ มีการกำหนดชนิดของคำภาษาไทย เพื่อหาประสิทธิภาพของชุดข้อมูลที่ดีที่สุด เพื่อนำไปใช้เทคนิคในเหมืองข้อมูลในการสร้างแบบจำลองจำแนกความคิดเห็น งานวิจัยนี้ใช้ 5 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ 4.5 (C4.5) เทคนิคนาอิวเบย์ (NB) เทคนิคเคเนียร์เรส

เนเบอร์ (KNN) เทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) และเทคนิคระบบเรียนรู้เชิงลึก (DL) การวัดประสิทธิภาพของชนิดของคุณลักษณะผู้วิจัยใช้หลักการ 10-โฟลด์ครอสวาเลชัน ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ และเปรียบเทียบประสิทธิภาพด้วยค่า ความถ่วงดุล ความแม่นยำ ความครบถ้วน และค่าสัมประสิทธิ์สหสัมพันธ์แมทธิวส์ จากการทดลองพบว่าคุณลักษณะที่เป็นคำวิเศษณ์ ส่งผลให้แบบจำลองมีค่าสูงกว่าคุณลักษณะที่เป็นคำกริยา และคุณลักษณะที่ใช้คำวิเศษณ์รวมกับคำกริยา

ในการเปรียบเทียบประสิทธิภาพของแบบจำลอง งานวิจัยฉบับนี้ใช้ความถ่วงดุล (F-Measure) ความแม่นยำ (Precision) ความครบถ้วน (Recall) และค่าสัมประสิทธิ์สหสัมพันธ์แมทธิวส์ (MCC) ดังแสดงใน Table 7, Table 8, Table 9 และ Table 10 ตามลำดับ

Table 7 แสดงความถ่วงดุล (F-Measure) ของแบบจำลอง 5 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ ซี 4.5 (C4.5) เทคนิคนอว์เพีย (NB) เทคนิคเคเนียร์เรสเนเบอร์ (KNN) เทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) และเทคนิคระบบเรียนรู้เชิงลึก (DL) ผลปรากฏว่าเปรียบเทียบความถ่วงดุลเฉพาะชุดข้อมูลพบว่าการใช้คำวิเศษ โดยใช้เทคนิค MLP และการใช้คำวิเศษณ์รวมกับคำกริยา โดยใช้เทคนิค KNN ให้ความถ่วงดุลสูงสุด ร้อยละ 100.00 แต่ถ้าเปรียบเทียบความถ่วงดุลค่าเฉลี่ยเฉพาะเทคนิค เทคนิค DL ให้ความถ่วงดุลเฉลี่ยทุกชุดข้อมูลสูงสุดร้อยละ 99.47 รองลงมาคือเทคนิค KNN มีความถ่วงดุลเฉลี่ยร้อยละ 99.37 และเทคนิคที่มีค่าเฉลี่ยต่ำที่สุดคือเทคนิค C4.5 ร้อยละ 97.87 และเปรียบเทียบความถ่วงดุลค่าเฉลี่ยเฉพาะชุดข้อมูล ชุดข้อมูลที่มีความถ่วงดุลเฉลี่ยสูงสุดคือชุดข้อมูลคำกริยา ร้อยละ 99.14 รองลงมาคือชุดข้อมูลคำวิเศษณ์ความถ่วงดุลเฉลี่ยร้อยละ 99.08 และชุดข้อมูลที่ทำให้ความถ่วงดุลเฉลี่ยต่ำที่สุดคือชุดข้อมูลที่มีคำวิเศษณ์และคำกริยาร้อยละ 98.34

Table 7 F-Measure of classification models

	Adverb	Verb	Adverb & Verb	Average
C4.5	98.6	98.9	96.1	97.87
NB	98.8	99	97	98.27
KNN	98.8	99.3	100	99.37
MLP	100	99.2	98.7	99.3
DL	99.2	99.3	99.9	99.47
Average	99.08	99.14	98.34	

Table 8 ความแม่นยำ (Precision) ของ 5 เทคนิค ได้แก่ C4.5, NB, KNN, MLP และ DL ผลปรากฏว่าเปรียบเทียบความแม่นยำเฉพาะชุดข้อมูล พบว่าการใช้คำวิเศษณ์โดยใช้เทคนิค MLP และการใช้คำวิเศษณ์รวมกับคำกริยาโดยใช้เทคนิค KNN ให้ความถ่วงดุลสูงสุดที่ร้อยละ 100.00 แต่ถ้าเปรียบเทียบความแม่นยำ ค่าเฉลี่ยเฉพาะเทคนิค DL ให้ความแม่นยำเฉลี่ยทุกชุดข้อมูลสูงสุดร้อยละ 99.50 รองลงมาคือเทคนิค KNN มีความแม่นยำเฉลี่ยร้อยละ 99.37 และเทคนิคที่มีความแม่นยำเฉลี่ยต่ำที่สุดคือเทคนิค C4.5 ร้อยละ 97.87 และเปรียบเทียบความแม่นยำค่าเฉลี่ยเฉพาะชุดข้อมูล ชุดข้อมูลที่มีความแม่นยำเฉลี่ยสูงสุดคือชุดข้อมูลคำกริยาร้อยละ 99.16 รองลงมาคือชุดข้อมูลคำวิเศษณ์ความแม่นยำเฉลี่ยร้อยละ 99.08 และชุดข้อมูลที่ทำให้ความแม่นยำเฉลี่ยต่ำที่สุดคือชุดข้อมูลที่มีคำวิเศษณ์และคำกริยาร้อยละ 98.3

Table 8 Precision of classification models

	Adverb	Verb	Adverb & Verb	Average
C4.5	98.6	98.9	96.1	97.87
NB	98.8	99	97	98.27
KNN	98.8	99.3	100	99.37
MLP	100	99.2	98.7	99.3
DL	99.2	99.4	99.9	99.5
Average	99.08	99.16	98.34	

Table 9 แสดงความครบถ้วน (Recall) ของ 5 เทคนิค ได้แก่ C4.5, NB, KNN, MLP และ DL ผลปรากฏว่าเปรียบเทียบความครบถ้วนเฉพาะชุดข้อมูล พบว่าการใช้คำวิเศษณ์โดยใช้เทคนิค MLP และการใช้คำวิเศษณ์รวมกับคำกริยาโดยใช้เทคนิค KNN ให้ความกว้างสูงสุดร้อยละ 100.00 แต่ถ้าเปรียบเทียบความกว้างค่าเฉลี่ยเทคนิค DL ให้ความครบถ้วนเฉลี่ยทุกชุดข้อมูลสูงสุดร้อยละ 99.47 รองลงมาคือเทคนิค KNN มีความครบถ้วนเฉลี่ยร้อยละ 99.37 และเทคนิคที่มีความครบถ้วนเฉลี่ยต่ำที่สุดคือเทคนิค C4.5 ร้อยละ 97.87 และเปรียบเทียบความครบถ้วนค่าเฉลี่ยเฉพาะชุดข้อมูล ชุดข้อมูลที่มีความครบถ้วนเฉลี่ยสูงสุดคือชุดข้อมูลคำกริยาร้อยละ 99.14 รองลงมาคือชุดข้อมูลคำวิเศษณ์ ความครบถ้วนเฉลี่ยร้อยละ 99.08 และชุดข้อมูลที่ทำให้ความครบถ้วนเฉลี่ยต่ำที่สุด คือชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา ร้อยละ 97.8

Table 9 Recall of classification models

	Adverb	Verb	Adverb & Verb	Average
C4.5	98.6	98.9	96.1	97.87
NB	98.8	99	97	98.27
KNN	98.8	99.3	100	99.37
MLP	100	99.2	98.7	99.3
DL	99.2	99.3	99.9	99.47
Average	99.08	99.14	98.34	

Table 10 แสดงค่าสัมประสิทธิ์สหสัมพันธ์ แมทริคซ์ (MCC) ของ 5 เทคนิค ได้แก่ C4.5, NB, KNN, MLP และ DL ผลปรากฏว่าเปรียบเทียบค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉพาะชุดข้อมูล พบว่าการใช้คำวิเศษณ์โดยใช้เทคนิค MLP และการใช้คำวิเศษณ์รวมกับคำกริยา โดยใช้เทคนิค KNN ให้ความกว้างสูงสุดร้อยละ 100.00 แต่ถ้าเปรียบเทียบค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์ค่าเฉลี่ยเทคนิค DL ให้ค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉลี่ยทุกชุดข้อมูลสูงสุด

ร้อยละ 99.53 รองลงมาคือ MLP มีค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉลี่ยร้อยละ 98.40 และเทคนิคที่มีค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉลี่ยต่ำที่สุดคือเทคนิคต้นไม้ตัดสินใจ C4.5 ร้อยละ 94.43 และเปรียบเทียบความครบถ้วน ค่าเฉลี่ยเฉพาะชุดข้อมูล ชุดข้อมูลที่มีค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉลี่ยสูงสุดคือชุดข้อมูลคำกริยาร้อยละ 99.26 รองลงมาคือชุดข้อมูลคำวิเศษณ์ค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉลี่ยร้อยละ 96.78 และชุดข้อมูลที่ทำให้ค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉลี่ยต่ำที่สุด คือชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา ร้อยละ 95.92

Table 10 MCC of classification models

	Adverb	Verb	Adverb & Verb	Average
C4.5	95.2	97.7	90.4	94.43
NB	95.6	97.9	92.7	95.4
KNN	95.9	98.6	100	98.17
MLP	100	98.4	96.8	98.4
DL	97.2	98.7	99.7	98.53
Average	96.78	98.26	95.92	

จากการศึกษานี้สรุปได้ว่าเทคนิค MLP มีค่าวัดประสิทธิภาพสูงสุดถึงร้อยละ 100 สำหรับทุกเกณฑ์การทดสอบเมื่อใช้ข้อมูลที่มีเฉพาะคำวิเศษณ์ และเทคนิค KNN มีประสิทธิภาพสูงสุดถึงร้อยละ 100 สำหรับทุกเกณฑ์การทดสอบเมื่อใช้ข้อมูลที่มีคำวิเศษณ์ และคำกริยารวมกัน

ที่ผ่านมามีงานวิจัยที่คล้ายคลึงกัน ซึ่งเป็นงานวิจัยของ สุภัทสร สมนเจตนา และจารี ทองคำ [4] ที่ทำการเปรียบเทียบประสิทธิภาพเทคนิคเหมืองข้อมูลในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็นของผู้ปกครองต่อการใช้สมาร์ทโฟนของบุตร ซึ่งพบว่าเทคนิค KNN เหมาะกับชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา ให้ค่าวัดประสิทธิภาพของแบบจำลองสูงสุดร้อยละ 100.00

4. บทสรุป

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาชนิดของคำภาษาไทยที่ส่งผลกระทบต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 จากสื่อสังคมออนไลน์ ผู้วิจัยได้ทำการเก็บรวบรวมจากเว็บไซต์ทวิตเตอร์ และพันทิป มีลักษณะเป็นข้อความภาษาไทย จำนวนทั้งหมด 2,920 ความคิดเห็น แล้วนำกระบวนการเหมืองข้อความมาทำการสกัดคุณลักษณะ แบ่งออกเป็น 2 กลุ่มที่บ่งบอกถึงความคิดเห็นเชิงบวก หรือเชิงลบ ได้แก่ คำวิเศษณ์ และคำกริยา นอกจากนั้นนักวิจัยได้นำคำวิเศษณ์ร่วมกับคำกริยา จึงทำให้ได้ชุดข้อมูล 3 ชุดข้อมูล คือ ชุดข้อมูลที่มีแต่คำวิเศษณ์ ชุดข้อมูลที่มีแต่คำกริยา และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา เพื่อหาชนิดของคำที่ส่งผลกระทบต่อประสิทธิภาพแบบจำลอง ในงานวิจัยนี้ได้นำ 5 เทคนิคมาทำแบบจำลองได้แก่ C4.5, NB, KNN, MLP และ DL การวัดประสิทธิภาพของชนิดของคุณลักษณะผู้วิจัยใช้หลักการ 10-โฟลด์ครอสวาไลเดชัน ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ และเปรียบเทียบประสิทธิภาพด้วย ความถ่วงดุล ความแม่นยำ ความครบถ้วนและค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์ เมื่อใช้แบบจำลองการจำแนกความคิดเห็น แล้ว ชนิดของคำที่มีประสิทธิภาพสูงสุด เมื่อพิจารณาจาก ความถ่วงดุล (F-Measure) ความแม่นยำ (Precision) ความครบถ้วน (Recall) และ ค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์ (MCC) ได้ผลดังนี้

ผลการเปรียบเทียบความถ่วงดุลเฉพาะชุดข้อมูลพบว่าการใช้ชุดข้อมูลที่มีคำวิเศษณ์ โดยใช้เทคนิค MLP ให้ความถ่วงดุลสูงสุดร้อยละ 100.00 และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา โดยใช้เทคนิค KNN ให้ความถ่วงดุลสูงสุดร้อยละ 100.00

ผลการเปรียบเทียบความแม่นยำเฉพาะชุดข้อมูลพบว่าการใช้ชุดข้อมูลที่มีคำวิเศษณ์ โดยใช้เทคนิค MLP ให้ความถ่วงดุลสูงสุดที่ร้อยละ 100.00 และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา โดยใช้เทคนิค KNN ให้ความถ่วงดุลสูงสุดที่ร้อยละ 100.00

ผลการเปรียบเทียบความครบถ้วนเฉพาะชุดข้อมูล พบว่าชุดข้อมูลที่มีคำวิเศษณ์ โดยใช้เทคนิค MLP ให้ความถ่วงดุลสูงสุดร้อยละ 100.00 และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา โดยใช้เทคนิค KNN ให้ความถ่วงดุลสูงสุดร้อยละ 100.00

ผลการเปรียบเทียบค่าสัมประสิทธิ์สหสัมพันธ์แมทริคซ์เฉพาะชุดข้อมูล พบว่าชุดข้อมูลที่มีคำวิเศษณ์ โดยใช้เทคนิค MLP ให้ความถ่วงดุลสูงสุดร้อยละ 100.00 และชุดข้อมูลที่มีคำวิเศษณ์และคำกริยา โดยใช้เทคนิค KNN ให้ความถ่วงดุลสูงสุดร้อยละ 100.00 งานวิจัยนี้ชี้ให้เห็นว่าชนิดของคำไทยส่งผลกระทบต่อประสิทธิภาพของแบบจำลองการจำแนกความคิดเห็น

5. กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณข้อมูล ความคิดเห็นของคนไทยต่อโรคโควิด 19 และกลุ่มสารสนเทศเชิงประยุกต์ ภาควิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ให้คำปรึกษา แสดงความคิดเห็นที่เป็นประโยชน์ ขอขอบคุณอาจารย์คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ให้คำปรึกษาและคำแนะนำต่าง ๆ เพื่อให้งานออกมาได้ดีและมีประสิทธิภาพมากที่สุด

6. References

- [1] Agarwal, V. and Sunitha, B.K. 2020. COVID-19 current pandemic and its societal impact. *International Journal of Advanced Science and Technology*. 29: 432-439.
- [2] Phumkhaichom, P. and Rattanachaiyaporn, P. 2020. Coronaviruses: Biological disasters from SARS and MERS to COVID-19. *Journal of Science and Technology, Mahasarakham University*. 39(6): 598-604. (in Thai)

- [3] Malik, M. and et al. 2018. A novel approach to web-based review analysis using opinion mining. **Procedia Computer Science**. 132: 1202-1209.
- [4] Somjettana, S. and Thongkam, J. 2021. Performance comparison of data mining techniques for building classification models of the Parent toward children who use smart phone. **Journal of Science and Technology, Ubon Ratchathani University**. 23(1): 21-30. (*in Thai*)
- [5] Phosaard, S. and Posawang, P. 2019. Classification for bus driver's behaviors using text extraction and machine learning technique. **Information Technology Journal**. 15(1): 71-80. (*in Thai*)
- [6] Khunsuk, T. and Thongkam, J. 2020. Feature selection method for improving customer reviews classification, **RMUTI Journal Science and Technology**. 13(1): 129-143. (*in Thai*)
- [7] Supianto, A.A. and et al. 2018. Decision tree usage for student graduation classification: A comparative case study in Faculty of Computer Science Brawijaya University. In: **Proceedings of the 2018 International Conference on Sustainable Information Engineering and Technology (SIET)**, 10-12 November 2018. Malang, Indonesia.
- [8] Chen, Y. and et al. 2017. A convolutional neural network for traffic information sensing from social media text. In: **Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)**, 16-19 October 2017. Yokohama, Japan.
- [9] Suksangaram, W. and et al. 2018. Predicting learning organization factors that affect performance by data mining techniques. In: **Proceedings of the 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (ISAI-NLP)**, 15-17 November 2018. Pattaya, Thailand.
- [10] Pinmuang, N. and Thongkam, J. 2018. Classifying Thai opinions on online media using text mining. **Journal of Science and Technology, Mahasarakham University**. 37(3): 372-379. (*in Thai*)
- [11] Gupta, S. and et al. 2018. Opinion mining for hotel rating through reviews using decision tree classification method. **International Journal of Advanced Research in Computer Science**. 9(2) : 180-184.
- [12] Thongkam, J. and Khunsuk, T. 2021. Translate Thai subtitles at anonymous. **Journal of Industrial Technology, Ubon Ratchathani Rajabhat University**. 11(1): 83-94. (*in Thai*)

- [13] Watchareewan, J. and Sunantha, S. 2017. Text classification analysis by stability comparison of algorithms. **Sripatum Review of Science and Technology**. 9(1): 19-31. (*in Thai*)
- [14] Hossain, T. and et al. 2017. Machine learning based class level prediction of restaurant reviews. In: **Proceedings of the 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)**, 21-23 Dec 2017. Dhaka, Bangladesh.
- [15] Xu, X. and et al. 2020. Application of neural network algorithm in fault diagnosis of mechanical intelligence. **Mechanical Systems and Signal Processing**. 141: 1-13.