การคัดเลือกคุณลักษณะความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ด้วยเหมืองข้อความ

วิทยานิพนธ์

ของ

ศราวุฒิ เกิดถาวร

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

มีนาคม 2565

การคัดเลือกคุณลักษณะความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ด้วยเหมืองข้อความ

วิทยานิพนธ์
ของ
ศราวุฒิ เกิดถาวร

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
มีนาคม 2565

Feature Selection Of Thai People's Sentiment Towards Covid-19 On Social Media With Text Mining

Sarawut Kedtarwon

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Information Technology)

March 2022

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายศราวุฒิ เกิดถาวร แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

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

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

.....กรรมการ
(ผศ. ดร. ฉัตรตระกูล สมบัติธีระ)

.....กรรมการ
(ผศ. ดร. สาธิต แสงประดิษฐ์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

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

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

ชื่อเรื่อง การคัดเลือกคุณลักษณะความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ด้วยเหมืองข้อความ
ผู้วิจัย ศราวุฒิ เกิดถาวร
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. จารี ทองคำ
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2565

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความ และเพื่อสร้างและเปรียบเทียบประสิทธิภาพของแบบบจำลองความรู้สึกของคนไทยต่อโรคโควิด 19 ได้รวบรวมความคิดเห็นด้วยข้อมูลจำนวน 2,920 ความคิดเห็น ผ่านกระบวนการเหมืองความคิดเห็น แล้วสร้างคลังคำศัพท์ได้ทั้งหมด 9,037 คำศัพท์ แล้วคัดเลือกเฉพาะคำวิเศษณ์ที่เป็นคำที่บ่งบอกถึงความรู้สึกได้ดี มาเป็นคำคุณลักษณะ ระบุตามความหมายเชิงบวกและเชิงลบ คงเหลือจำนวน 236 คำ แล้วการคัดเลือกคุณลักษณะ 2 รูปแบบ รูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 รูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 แล้วทำการสร้างแบบจำลองจำแนกแล้วนำมาวัดประสิทธิภาพ 6 เทคนิค ได้แก่ เทคนิคต้นไม้ตัดสินใจ เทคนิคซัพพอร์ทเวกเตอร์แมชชีน เทคนิคนาอีฟเบย์ เทคนิคเคเนียเรสเนเบอร์ เทคนิคเพอร์เซปตรอนหลายชั้น เทคนิคแบบระบบเรียนรู้เชิงลึก จากนั้นใช้หลักการ 10-โฟลด์ครอสวาลิเดชั่น ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลอง พบว่าเมื่อลดมิติข้อมูลลงด้วยการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 การคัดเลือกคุณลักษณะเหมาะสมกับเทคนิคการจำแนกมากที่สุด พบว่า ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-measure) ผลการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 ที่เทคนิค KNN และ MLP สูงที่สุด ร้อยละ 99.20

คำสำคัญ : เหมืองความคิดเห็น, จำแนกความคิดเห็น, โควิด19

TITLE Feature Selection Of Thai People's Sentiment Towards Covid-19 On Social Media With Text Mining

AUTHOR Sarawut Kedtarwon

ADVISORS Assistant Professor Jaree Thongkam , Ph.D.

DEGREE Master of Science **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2022

ABSTRACT

The objective of this research was to study a method of word-weighted trait selection to create an effective model for discriminating the opinions of Thai people on COVID-19 by using text mining principles. and to create and compare the performance of The Thai sentiment model for COVID-19 collected 2,920 opinions through an opinion mining process. Then build a vocabulary of all 9,037 words, and select only the adverbs that express feelings well. come as a feature word Indicated according to positive and negative meanings, 236 words were left, then 2 forms of trait selection were made. Scheme 1 Chi-Square TFIDF and BM25 traits. Characteristics were 83 words. Attributes were selected with TFIDF and BM25, discriminant models were created and performance was measured using 6 techniques: decision tree technique. vector machine support techniques Technique Na Eve Bay Caniers Neighbor Technique Multi-layer perceptron technique deep learning techniques Then apply the principle 10-fold cross validation To segment the data into learning datasets and test datasets. and measure the performance of the model It was found that when reducing the dimensions by selecting the Chi-Square attribute, the characteristic word remained. Then the traits were selected with TFIDF and BM25. The trait selection was most suitable for the classification technique. It was found that the precision, recall, F-measure, trait selection results with TFIDF and BM25 at KNN and MLP techniques were the highest at 99.20%.

Keyword : Opinion Mining, classification, COVID19

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจากอาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.จารี ทองคำ ที่ถ่ายทอดวิชาความรู้ตลอดจนคอยสอนศิษย์ด้วยจิตเมตตา ผู้ซึ่งมีจิตวิญญาณของความเป็นครูโดยแท้จริงและแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่ เพื่อให้วิทยานิพนธ์ฉบับเต็มฉบับนี้สมบูรณ์ที่สุด ขอขอบพระคุณคณะกรรมการ นำโดย รองศาสตราจารย์ ดร.สิทธิชัย บุษหมั่น ประธานกรรมการสอบ ผู้ช่วยศาสตราจารย์ ดร.ฉัตรตระกูล สมบัติธีระ และผู้ช่วยศาสตราจารย์ ดร.สาธิต แสงประดิษฐ์ กรรมการสอบ ที่กรุณาให้คำแนะนำตลอดจนปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ด้วยความเอาใจใส่อย่างดียิ่ง ผู้วิจัยจึงขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณอาจารย์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ให้ความรู้และคำแนะนำในการศึกษาจนสำเร็จการศึกษาในครั้งนี้

ขอขอบพระคุณ บิดามารดา เครือญาติ ครอบครัว และกัลยาณมิตร ที่คอยสนับสนุนและให้กำลังใจ จนทำให้ วิทยานิพนธ์ฉบับเต็มฉบับนี้สมบูรณ์ ครั้งนี้ได้สำเร็จไปด้วยดี

ผู้วิจัยมีความหวังว่าวิทยานิพนธ์ฉบับเต็มนี้จะมีประโยชน์อยู่ไม่มากก็น้อย จึงขอมอบส่วนข้อที่ดีทั้งหมดนี้ แด่ครู อาจารย์ ที่ได้ให้คำแนะนำและแนวทางในการแก้ไขปัญหา สำหรับข้อบกพร่องต่าง ๆ ที่อาจจะเกิดขึ้นนั้น ผู้วิจัยน้อมรับ และยินดีที่จะรับฟังคำแนะนำ เพื่อเป็นประโยชน์ในการพัฒนาวิทยานิพนธ์ และงานวิจัยต่อไป

ศราวุฒิ เกิดถาวร

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1 บทนำ

1.1 หลักการและเหตุผล

การระบาดของโรคติดเชื้อไวรัสโคโรนา2019 (COVID-19) เป็นโรคที่ไม่เคยปรากฏมาก่อน [1] ส่งผลกระทบอย่างรุนแรงต่อ สุขภาพ เศรษฐกิจ และสังคม ในช่วงการแพร่ระบาดของเชื้อไวรัส COVID-19 ในประเทศไทย วันที่ 23 มกราคม 2563 ทำให้วิถีการดำเนินชีวิตของประชาชนเปลี่ยนแปลง และมีมาตรการป้องกันเช่น Work From Home การทำ Social Distancing ถือเป็นมาตรการที่สำคัญที่จะช่วยลดการแพร่กระจายของเชื้อไวรัสที่ได้ผลดี [2] ในช่วงวิกฤตการระบาดของโรคทำให้ประชาชนมีแนวโน้มที่จะใช้แพลตฟอร์มสื่อสังคมออนไลน์ เพื่อที่จะได้รับข้อมูลที่จำเป็นและแลกเปลี่ยนความคิดเห็น ซึ่งแน่นอนว่าจะต้องมีเสียงและการวิพากษ์วิจารณ์บางอย่างเกิดขึ้นบนสื่อสังคมออนไลน์ ตลอดเวลาที่วิกฤติครั้งนี้ดำเนินไป

เหมืองความคิดเห็น (Opinion Mining) [3] เป็นกระบวนการวิเคราะห์ความคิดเห็นที่มีความรู้สึกของคน นำเสนอในรูปของข้อความ ซึ่งพบได้บนเครือข่ายสังคมออนไลน์ เพื่อให้ทราบถึงความพึงพอใจในเชิงบวกหรือเชิงลบ การจัดการข้อความที่ขับเคลื่อนด้วยความคิดเห็นนั้นสามารถนำไปประยุกต์ใช้กับงานที่สำคัญหลายประเภท ตัวอย่างเช่น การติดตามการเปลี่ยนแปลงทัศนคติของประชาชนเกี่ยวกับเรื่องการเมือง หรือ การจำแนกความคิดเห็นของผู้วิจารณ์เกี่ยวกับเรื่องอื่น ๆ บนสังคม โดยแบ่งประเภทบทวิจารณ์สินค้าออนไลน์ แต่ปัญหาหนึ่งในการทำเหมืองความคิดเห็นจะเกี่ยวข้องกับการวิเคราะห์ข้อมูลที่มีจำนวนมหาศาลและ มีความซับซ้อนโดยส่วนใหญ่กระบวนการทำเหมืองความคิดเห็นนั้นจะมีข้อมูลที่ประกอบไปด้วย คุณลักษณะที่ไม่ตรงประเด็นและมีมิติของข้อมูลจำนวนมาก ซึ่งส่งผลให้การทำเหมืองความคิดเห็นต้องใช้เวลาในการวิเคราะห์มากขึ้นถ้าข้อมูลมิติ หรือตัวแปรมากจะทำให้ข้อมูลเกิดการกระจาย และอาจไม่มีความสัมพันธ์กับมิติอื่น ดังนั้นการคัดเลือกคุณลักษณะของข้อมูลโดยการทำให้ข้อมูลเดิมมีขนาดลดลงและสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด จึงเป็นแนวหาหนึ่งที่สามารถช่วยแก้ปัญหาดังกล่าว และยังสามารถช่วยให้การจำแนกข้อมูลได้แม่นยำมากขึ้น

นักวิจัยหลายท่านได้นำเอาการคัดเลือกคุณลักษณะซึ่งเป็นกระบวนการในการทำดัชนี มาใช้เพื่อปรับปรุงประสิทธิภาพการจำแนกความคิดเห็น Hasan และคณะ [4] ใช้แบบจำลอง Bag of Words (Bow) และ Term Frequency-Inverse Document Frequency (TF-IDF) เพื่อวิเคราะห์ความรู้สึกที่ใช้ร่วมกันเพื่อจำแนกทวีตเชิงบวกและเชิงลบอย่างแม่นยำ ของข้อมูลที่ประมวลผลล่วงหน้าโดยใช้ภาษาธรรมชาติ เปรียบเทียบวิธีที่เสนอกับ เทคนิคSupport Vector Machine (SVM), เทคนิค Maximum Entropy, เทคนิคNaive Bayes และ เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ทำให้ความแม่นยำของการวิเคราะห์ความรู้สึก การทดลองพบว่า ความแม่นยำ 85.25% Guo และ Yang [5] วิเคราะห์น้ำหนักของคำข้อมูลจาก People news แบ่งเป็น 4 ประเภทใช้เทคนิค TFIDF

แบบดั้งเดิมและเทคนิค TFIDF ที่ปรับปรุงแล้ว ผลลัพธ์แสดงให้เห็นว่าเทคนิค TFIDF ที่ปรับปรุงแล้วมีความแม่นยำสูงกว่าเทคนิค TFIDF แบบเดิม Kadhim และคณะ [6] ได้ใช้เทคนิคที่แตกต่างกันคือ BM25 และ TF-IDF เพื่อแยกคำ ฟังก์ชัน BM25 ใช้เพื่อจัดลำดับชุดเอกสารที่ไม่มีข้อมูลที่เกี่ยวข้อง ในขณะที่ใช้ TF-IDF เพื่อถ่วงน้ำหนักคุณลักษณะตามความถี่ของแต่ละคำกล่าวคือคำศัพท์ในเอกสารเกี่ยวข้องกับคำอื่น เพื่อแยกคำหลักจากการรวบรวมข้อมูลที่ขึ้นอยู่กับ Twitter เป็นที่ชัดเจนว่าเทคนิค TF-IDF มีประสิทธิภาพดีกว่า BM25 ตามมาตรวัด F1 Nurhayati และคณะ [7] เลือกคุณสมบัติ Chi-Square เพื่อกำจัดคุณสมบัติ การศึกษานี้มีวัตถุประสงค์เพื่อตรวจสอบผลของการเลือกคุณสมบัติ Chi-Square ที่มีต่อประสิทธิภาพของอัลกอริทึมNaïve Bayes ในการวิเคราะห์เอกสารความรู้สึก ข้อมูลถูกนำมาจากข้อมูลการฝึกอบรม Corpus v1.0 Indonesia Movie Review ผลจากการวิเคราะห์ความเชื่อมั่นโดยไม่เลือกคุณลักษณะได้รับความแม่นยำ 73.33% ความแม่นยำ 100.00% การเรียกคืน 65.21% ในขณะที่การเลือกคุณสมบัติ Chi-Square (ระดับนัยสำคัญที่ 0.1) ได้ผลลัพธ์ความแม่นยำ 93.33% และการเรียกคืน 93.33% จากผลลัพธ์จะเห็นได้ว่าการเลือกคุณสมบัติ Chi-Square มีผลต่อประสิทธิภาพอัลกอริทึมของ Naïve Bayes ในการวิเคราะห์เอกสารความรู้สึก

งานวิจัยในการจำแนกข้อความก็ได้มีนักวิจัยหลายท่านได้นำเอาเทคนิคการทำเหมืองความคิดเห็นมาใช้ในการจำแนกข้อความ เช่น Supianto และคณะ [8] ใช้เทคนิคการจำแนกต้นไม้แบบสุ่ม REP Tree และ C4.5 ความแม่นยำเฉลี่ยที่สูงขึ้นใช้เทคนิค C4.5 มากที่สุด มีความแม่นยำ 77.01% REP Tree และ Random Tree ตามลำดับ Chen และคณะ[9] ใช้ Deep Neural Networks เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) และเทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) พบว่า Deep Neural Networks ที่มีประสิทธิภาพที่เหนือกว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) และเทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) ในการการเข้าชมแพลตฟอร์มโซเชียวมีเดีย Sina Weibo Suksangaram และคณะ [10] ใช้เทคนิค: C4.5, Naïve Bayes และ SVM ในการทำนายการเรียนรู้ขององค์กรที่มีผลต่อการปฏิบัติงานของพนักงานธนาคารเพื่อการเกษตรและสหกรณ์การเกษตรในภาคตะวันตก ผลการวิจัยพบว่าเทคนิค SVM มีค่าความแม่นยำ 98.33% ความแม่นยำ 0.025 และการเรียกคืน 0.984 ที่มากกว่าเทคนิค C4.5 และ Naïve Bayes

งานวิจัยที่เกี่ยวข้องดังกล่าวมีการนำเอากระบวนการคัดเลือกคุณลักษณะ แล้วนำเอาเทคนิคต่าง ๆ มาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะ ดังนั้นในงานวิจัยนี้จึงได้นำเอาการกระบวนการคัดเลือกคุณลักษณะ แล้วนำเอาเทคนิคต่าง ๆ เพิ่มประสิทธิภาพในการจำแนกความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ ซึ่งเทคนิคของการคัดเลือกคุณลักษณะที่นำมาใช้ได้แก่ Chi-Square, TFIDF และ BM25 แล้วนำเอาเทคนิคต่าง ๆ มาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะ และทำการเปรียบเทียบเทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความความรู้สึกของคนไทย คือ เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

(Support Vector Machine: SVM) เทคนิคนาอีฟเบย์ (Naive Bayes) เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor: KNN) เทคนิคเพอร์เซปตรอนหลายชั้น (Multi-layer Perceptron) เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) จากนั้นใช้หลักการ 10-fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วยค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure)

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาวิธีการการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความ

1.2.2 เพื่อสร้างและเปรียบเทียบประสิทธิภาพของ แบบบจำลองความรู้สึกของคนไทยต่อโรคโควิด 19

1.3 ความสำคัญของงานวิจัย

1.3.1 ได้วิธีการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักที่เหมาะสมต่อการจำแนกความรู้สึกของคนไทยต่อโรคโควิด19 บนสื่อสังคมออนไลน์

1.3.2 ได้แบบจำลองที่มีประสิทธิภาพในการจำแนกความรู้สึกของคนไทยต่อโรคโควิด19 บนสื่อสังคมออนไลน์

1.4 ขอบเขตของการวิจัย

1.4.1 เก็บรวบรวมความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ ประมาณ 2,000 ความรู้สึก ตั้งแต่วันที่ 23 มกราคม 2563 ถึงวันที่ 1 พฤษภาคม 2563

1.4.2 คัดเลือกคุณลักษณะด้วยการให้น้ำหนักที่เหมาะสม ที่สามารถทำให้แบบจำลอง ที่ได้มีประสิทธิภาพในการจำแนก โดยใช้หลักการเหมืองข้อความ (Chi-Square TFIDF และ BM25)

1.4.3 สร้างแบบจำลองในการวิเคราะห์เหมืองความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ ในรูปแบบของภาษาไทยว่ามีความเห็นไปในทิศทางใดโดยจะแบ่งเป็น 2 กลุ่ม คือความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ โดยการเปรียบเทียบเทคนิคที่ใช้ในการสร้างแบบจำลองเพื่อจำแนกความคิดเห็น 6 เทคนิคคือ 1.เทคนิคต้นไม้ตัดสินใจ แบบ C4.5 (Decision tree :C4.5) 2.เทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine :SVM) 3.เทคนิคนาอีฟเบย์ (Naive Bayes :NB) 4.เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor :KNN) 5.เทคนนิคเพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron :MLP) 6.เทคนิคเทคนิคแบบระบบเรียนรู้เชิงลึก

(Deep learning :DL) จากนั้นใช้หลักการ10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูล ทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วย ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

1.5 นิยามศัพท์เฉพาะ

1.5.1 เครือข่ายสังคมออนไลน์ (Social Media) คือ Social "สังคม" หมายถึง สังคมออนไลน์ Media "สื่อ"หมายถึงสื่อในรูปแบบต่าง ๆ ไม่ว่าจะเป็น ภาพ เสียง วีดิโอ ข้อความ ดังนั้น Social Media หมายความว่าสื่อสังคมออนไลน์ที่ผู้ใช้อินเทอร์เน็ตสามารถแลกเปลี่ยนประสบการณ์ซึ่งกันและกัน โดยใช้สื่อต่าง ๆ เป็นตัวแทนในการสนทนา โดยได้มีการจัดแบ่งประเภทของ Social Media ออกเป็นหลายประเภท เช่น ประเภทสื่อสิ่งพิมพ์ (Publish) ที่มี Wikipedia Blogger เป็นต้น ประเภทสื่อแลกเปลี่ยน (Share) ที่มี YouTube Flickr Slide Share เป็นต้น ประเภทสื่อสนทนา (Discuss) ที่มี messenger Line Skype เป็นต้น และยังมีอีกหลายประเภท โดยอีกประเภทของ Social Media ที่สร้างความสับสนให้บ้าง ก็คือประเภทเครือข่ายสังคมออนไลน์ หรือที่เรียกกันว่า Social Network ที่มี Facebook Twitter เป็นต้น

1.5.2 แบบจำลอง คือ ต้นแบบที่ใช้ในวิเคราะห์เหมืองความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์

1.5.3 ความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ คือ ข้อความแสดงความคิดเห็นที่ประกอบด้วยข้อมูลอันเป็นข้อเท็จจริงกับการแสดงความคิดเห็นความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ มีทั้งเชิงบวกและเชิงลบ

1.5.4 การวิเคราะห์ความคิดเห็น หมายถึง การวิเคราะห์ข้อความที่แสดงความรู้สึกหรืออารมณ์ของความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ โดยแบ่งออกเป็นความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ

1.5.5 เทคนิค C4.5 คือ เทคนิคต้นไม้ตัดสินใจที่พัฒนา มาจากเทคนิคไอดี 3 ซึ่งเป็นเทคนิคพื้นฐานที่ใช้ในการเปรียบเทียบประสิทธิภาพของการทำงานเทคนิคอื่น ๆ โดยการสร้างต้นไม้ประกอบการตัดสินใจสำหรับใช้ในการจำแนกประเภทข้อมูลโดยในการสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่าเกนความรู้ (Information Gain) เพื่อหาความสัมพันธ์ของในแต่ละโหนดคุณลักษณะ ใช้ค่าอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นโหนดในแต่ละระดับ

1.5.6 เทคนิค SVM หมายถึง เทคนิคการจำแนกหมวดหมู่ที่นิยมใช้หลักการสร้างเส้นสมการเส้นตรงเพื่อทำการแบ่งกลุ่มของข้อมูล 2 กลุ่มออกจากกัน โดย SVM พยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มแล้วเพิ่มเส้นขอบทั้งสองข้างให้กับเส้นแบ่ง เส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็น

เส้นแบ่งที่ดีที่สุด และเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ตเวกเตอร์”

1.5.7 เทคนิค NB คือ เทคนิควิธีการจำแนกประเภทที่อาศัยหลักการของทฤษฎีความน่าจะเป็นตามกฎของเบย์ เพื่อหาว่าสมมติฐานใดน่าจะมีความถูกต้องมากที่สุด ซึ่งจะสามารถบ่งบอกถึงความน่าจะเป็นของข้อมูลชุดหนึ่งที่จะอยู่ในหมวดหมู่ของข้อมูลนั้น ๆ ซึ่งเป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยใช้ในการจำแนกหมวดหมู่เอกสารข้อความ (text classification) ได้ดี การทำงานไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมาก

1.5.8 เทคนิค KNN คือ Machine Learning ประเภท Supervised แบบ Classification นั้น คือ จำเป็นจะต้องมี Data Set ที่มีเฉลย (Label) ให้ ชื่อภาษาไทยคือ เพื่อนบ้านที่ใกล้เคียงที่สุด K ตัว หมายความว่า เวลาทำนาย จะดึงข้อมูลที่ใกล้เคียงตัวเองที่สุดมา K ตัว แล้วมานับจำนวนกันว่าเพื่อนบ้านที่ใกล้เคียงของมันทั้งหมด K ตัว เป็น Class อะไรมากที่สุด

1.5.9 เทคนิค MLP [11] เทคนิคหนึ่งในโครงข่ายประสาทเทียม ที่มีโครงสร้างเป็นแบบหลาย ๆ ชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี และใช้การคำนวณด้วยสูตรทางคณิตศาสตร์สำหรับประมวลผลสารสนเทศโดยการจัด เป็นลำดับขั้น (layer) โดยมีข้อมูลนำเข้า และสุดท้ายออกเป็นผลลัพธ์เพื่อจำแนกข้อมูล

1.5.10 เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) คือการเรียนรู้เชิงลึกเป็นสาขาของการเรียนรู้ของเครื่องพื้นฐานของการเรียนรู้เชิงลึกคือ เทคนิคที่พยายามจะสร้างแบบจำลองเพื่อแทนความหมายของข้อมูลในระดับสูงโดยการสร้างสถาปัตยกรรมข้อมูลขึ้นมาที่ประกอบไปด้วยโครงสร้างย่อย ๆหลายอัน และแต่ละอันนั้นได้มาจากการแปลงที่ไม่เป็นเชิงเส้น การเรียนรู้เชิงลึกอาจมองได้ว่าเป็นวิธีการหนึ่งของการเรียนรู้ของเครื่องที่พยายามเรียนรู้วิธีการแทนข้อมูลอย่างมีประสิทธิภาพ ตัวอย่างเช่น รูปภาพภาพหนึ่ง สามารถแทนได้เป็นเวกเตอร์ของความสว่างต่อจุดพิกเซล หรือมองในระดับสูงขึ้นเป็นเซ็ตของขอบวัตถุต่าง ๆ หรือมองว่าเป็นพื้นที่ของรูปร่างใด ๆ ก็ได้การแทนความหมายดังกล่าวจะทำให้การเรียนรู้ที่จะทำงานต่าง ๆ ทำได้ง่ายขึ้น ไม่ว่าจะเป็นการรู้จำใบหน้าหรือการรู้จำการแสดงออกทางสีหน้า การเรียนรู้เชิงลึกถือว่าเป็นวิธีการที่มีศักยภาพสูงในการจัดการกับฟีเจอร์สำหรับการเรียนรู้แบบไม่มีผู้สอนหรือการเรียนรู้แบบกึ่งมีผู้สอน

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้ประกอบด้วย ทฤษฎีที่เกี่ยวข้องกับโรคโควิด 19 การทำเหมืองความคิดเห็น การวัดประสิทธิภาพแบบจำลอง และงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 โรคโควิด 19

โรค COVID-19 คือ โรคติดเชื้อจากไวรัสชนิดหนึ่ง ซึ่งพบการระบาดในช่วงปี 2019 ที่เมืองอู่ฮั่น ประเทศจีน โดยในตอนนั้นเราจะรู้จักกันโรคนี้ในชื่อว่า ไวรัสอู่ฮั่น ก่อนที่ภายหลังจะระบุเชื้อก่อโรคได้ว่าเป็นเชื้อในตระกูลโคโรนาไวรัส แต่เป็นสายพันธุ์ใหม่ที่ไม่เคยเกิดขึ้นมาก่อน ดังนั้น ทางองค์การอนามัยโลก จึงได้ตั้งชื่อโรคติดต่อชนิดนี้ใหม่อย่างเป็นทางการ โดยมีชื่อว่า COVID-19 เพื่อไม่ให้เกิดรอยมลทินกับพื้นที่ที่เกิดการระบาดของโรคด้วย โคโรนาไวรัส เชื้อนี้มีมานานและหลายสายพันธุ์ โคโรนาเป็นเชื้อไวรัสที่ก่อให้เกิดโรคทางเดินระบบหายใจ มีมานานกว่า 60 ปี แล้ว และเป็นเชื้อไวรัสตระกูลใหญ่ที่มีอยู่หลายสายพันธุ์ โดยชื่อโคโรนาก็มีที่มาจากลักษณะของเชื้อไวรัสที่รูปร่างคล้ายมงกุฎ (Corona เป็นภาษาละตินที่แปลว่ามงกุฎ) เนื่องจากเชื้อไวรัสชนิดนี้มีสารพันธุกรรมเป็น RNA มีเปลือกหุ้มด้านนอกที่ประกอบไปด้วยโปรตีนคลุมด้วยกลุ่มคาร์โบไฮเดรต ไขมันเป็นปุ่ม ๆ ยื่นออกไปจากอนุภาคไวรัส อธิบายง่าย ๆ คือเป็นเชื้อไวรัสที่มีหนามอยู่รอบตัว จึงสามารถเกาะตัวอยู่ในอวัยวะที่เป็นเป้าหมายของเชื้อไวรัสได้ โคโรนาไวรัสเป็นเชื้อที่ก่อโรคได้ทั้งในคนและสัตว์ เนื่องจากตัวไวรัสมีสารพันธุกรรม RNA ซึ่งมีโอกาสกลายพันธุ์สูง สามารถติดเชื้อข้ามสปีชีส์กันได้ โดยเฉพาะในสถานที่ที่มีการรวมตัวของสัตว์อย่างหนาแน่น เช่น ตลาดค้าสัตว์ เป็นต้น ดังนั้นต้นตอการแพร่ระบาดของโรคก็อาจจะมาจากสัตว์ปีก เช่น นก ค้างคาว ไก่ หรือสัตว์เลี้ยงลูกด้วยนม เช่น ม้า วัว แมว สุนัข กระต่าย หนู อูฐ รวมไปถึงสัตว์เลื้อยคลานอย่างงู เป็นต้น เราเจอกับโคโรนาไวรัสกันอยู่เนือง ๆ เพราะอย่างที่บอกว่าโคโรนาไวรัสมีอยู่หลายสายพันธุ์ แต่ส่วนใหญ่จะไม่ก่อให้เกิดโรครุนแรง เป็นเพียงไข้หวัดธรรมดา แต่ก็มีโคโรนาไวรัสบางสายพันธุ์ที่ก่ออาการรุนแรงจนถึงขั้นปอดอักเสบได้ เช่น โรคซาร์ส ที่มีสาเหตุมาจากโคโรนาไวรัสสายพันธุ์ SARS-CoV ข้าม สปีชีส์จากค้างคาวมาสู่ตัวชะมด แล้วมาติดเชื้อในคน และโรคเมอร์ส ที่มีสาเหตุมาจากโคโรนาไวรัสสายพันธุ์ MERS-CoV ข้ามสปีชีส์จากค้างคาวสู่อูฐ และมาติดเชื้อในคน และล่าสุดกับเชื้อโคโรนาไวรัสสายพันธุ์ที่ก่อโรค COVID-19 ซึ่งเป็นโคโรนาไวรัสสายพันธุ์ใหม่ โดยโคโรนาไวรัสสายพันธุ์ใหม่ 2019 มีชื่ออย่างเป็นทางการว่า SARS-CoV-2 เป็นเชื้อไวรัสลำดับที่ 7 ในตระกูล coronaviruses lineage B จีนัส beta coronavirus ที่ก่อให้เกิดโรคในคน จากการศึกษาทางพันธุกรรมของไวรัส และการเรียงลำดับของรหัสแต่ละตัวทำให้พบต้นตอของเชื้อ SARS-CoV-2 ว่า ไวรัสสายพันธุ์ใหม่ชนิดนี้มีจำนวน นิวคลีโอไทด์ที่เหมือนกันถึงร้อยละ 89.1 ของเชื้อ

SARS-like coronaviruses ในค้างคาวที่เคยพบในประเทศจีน และในภายหลังก็มีข้อมูลที่ยืนยันว่าต้นตอของโคโรนาไวรัสสายพันธุ์ใหม่ 2019 เกิดจากการผสมสารพันธุกรรมระหว่างโคโรนาไวรัสของค้างคาวกับโคโรนาไวรัสในงูเห่า กลายพันธุ์เป็นโคโรนาไวรัส สายพันธุ์ SARS-CoV-2 ที่แพร่เชื้อจากงูเห่ามายังคนได้

โคโรนาไวรัสเป็นเชื้อที่ไม่สามารถอยู่เดี่ยว ๆ ได้ แต่จะแฝงตัวอยู่ในละอองฝอยจากการไอ จาม และสารคัดหลั่งอย่างน้ำมูก น้ำลาย หรืออุจจาระ ดังนั้นการแพร่เชื้อโคโรนาไวรัสสายพันธุ์ใหม่ ผู้ที่อยู่ใกล้ชิดก็ต้องได้รับเชื้อผ่านการสูดดมละอองฝอยขนาดใหญ่และละอองฝอยขนาดเล็กในอากาศ รับเชื้อเข้าไปในทางเดินหายใจ หรือใครที่อยู่ใกล้ผู้ป่วยในระยะ 1-2 เมตร ก็อาจจะติดเชื้อจากการสูดฝอยละอองขนาดใหญ่ และฝอยละอองขนาดเล็กจากการไอ จาม รดกันโดยตรง หรือหากอยู่ห่างจากผู้ติดเชื้อในระยะ 2 เมตรขึ้นไป ก็อาจติดเชื้อจากการสูดฝอยละอองขนาดเล็กได้เหมือนกัน

นอกจากนี้โคโรนาไวรัส สายพันธุ์ใหม่ ยังอาจแพร่เชื้อโดยการสัมผัสได้ เช่น การจับของใช้สาธารณะร่วมกัน แล้วมาสัมผัสเยื่อบุต่าง ๆ ในร่างกาย เช่น ขยี้ตา สัมผัสปาก หรือหยิบของกินเข้าปาก เป็นต้น การที่เชื้อไวรัสจะก่อโรคในร่างกายเราได้ เราต้องได้รับเชื้อไวรัสดังกล่าวผ่านเยื่อบุต่าง ๆ จนนำไปสู่การติดเชื้อที่ระบบทางเดินหายใจส่วนบน เช่น เซลล์เยื่อบุหลอดลม ซึ่งไวรัสจะใช้ผิวเซลล์ของไวรัสจับกับเอนไซม์ที่ผิวเซลล์มนุษย์ จากนั้นไวรัสจะค่อย ๆ เพิ่มจำนวนเชื้อในตัวเรา ซึ่งหากภูมิต้านทานของเราไม่สามารถจัดการกับเชื้อไวรัสนี้ได้ จำนวนเชื้อไวรัสก็จะเพิ่มมากขึ้น และกระจายไปยังเซลล์ข้างเคียง ทำลายเซลล์ในหลอดลมและปอด ทำให้ปอดอักเสบและเกิดภาวะทางเดินหายใจล้มเหลวได้ ติดเชื้อโคโรนาไวรัสแล้วอันตรายต่อปอดแค่ไหน โดยกรมควบคุมโรคเคยให้ข้อมูลไว้ว่า มีเพียง 15-20% ที่เชื้อลงปอดแล้วทำให้เป็นปอดอักเสบ แต่เมื่อลงปอดไปแล้วจะก่อความรุนแรงแค่ไหน ขึ้นอยู่กับภูมิต้านทานร่างกายของแต่ละคน ขณะที่ข้อมูลผู้ติดเชื้อในประเทศจีนพบว่า การลงปอดมักเกิดขึ้นในสัปดาห์ที่สองหลังจากได้รับเชื้อแล้ว แต่มีผู้ติดเชื้อประมาณ 80% ที่เชื้อไม่ลงปอด เป็นเพียงไข้หวัดธรรมดา กรณีเชื้อไวรัสลงปอดจะเกิดขึ้นเมื่อเชื้อไวรัสเข้าสู่ร่างกายแล้วจะแบ่งตัวและเจริญเติบโตในเซลล์มนุษย์ เช่น เซลล์ของเยื่อบุหลอดลม จึงจะก่อโรคได้ และเซลล์มนุษย์ที่ติดเชื้อจะเพิ่มจำนวนและปล่อยเชื้อไวรัสออกมานอกเซลล์ เพื่อไปก่อโรคในเซลล์ข้างเคียง เมื่อเชื้อไวรัสเพิ่มมากขึ้นเรื่อย ๆ จะทำลายเซลล์มนุษย์ในหลอดลม ถุงลม และเนื้อปอด รวมทั้งเซลล์ข้างเคียงด้วย หากภูมิคุ้มกันของร่างกายไม่แข็งแรงพอ หรือสร้างภูมิต้านทานขึ้นมาช้า เพราะเม็ดเลือดขาวเพิ่งพบกับเชื้อไวรัสเป็นครั้งแรก ทำให้ภูมิต้านทานทำลายเชื้อไม่ทัน ผู้ป่วยจะมีอาการปอดอักเสบ และเมื่อเซลล์ที่ติดเชื้อจำนวนมากตาย จะถูกทดแทนด้วยพังผืดในเวลา 2-3 สัปดาห์หลังการเจ็บป่วย มีข้อมูลว่า ผู้ป่วยที่มีอาการปอดอักเสบส่วนใหญ่ เนื้อปอดจะถูกทำลายไปราว 20% ซึ่งหากเนื้อปอดถูกทำลายไม่ถึง 50% ร่างกายฟื้นฟูเองได้ตามสภาพแต่ละคน ทว่าจะมีผู้ป่วยราว 5% ที่เนื้อปอดถูกทำลาย 70-80% กรณีนี้ถือว่า วิกฤต ร่างกายอาจฟื้นตัวไม่ไหว หรือแพทย์อาจต้องใช้เครื่อง ECMO หรือเครื่องหัวใจ-

ปอดเทียมแบบเคลื่อนย้าย มาทำงานแทนหัวใจและปอดของผู้ป่วย ซึ่งหากช่วยไม่ไหว สุดท้ายแล้วระบบหายใจจะล้มเหลวและเป็นเหตุให้ผู้ติดเชื้อโคโรนาไวรัสเสียชีวิต โดยทั่วไปแล้ว หากเป็นคนที่มีภูมิต้านทานแข็งแรง ไม่มีโรคประจำตัว ไม่มีปัญหาที่ปอด ส่วนใหญ่จะสามารถทนต่อการก่อโรคของเชื้อโคโรนาไวรัส สายพันธุ์ใหม่ ที่ค่อย ๆ เพิ่มจำนวนขึ้น พร้อมกันนั้นภูมิคุ้มกันของร่างกายก็จะพยายามต่อสู้กับเชื้อไวรัสได้ทันกาล ก่อนที่ปอดจะเสียหายหนัก แต่สำหรับคนที่มีภูมิต้านทานไม่แข็งแรง เช่น ผู้สูงอายุ ผู้ที่มีโรคประจำตัว ผู้ที่ได้รับยากดภูมิคุ้มกัน ทำให้ร่างกายผลิตเซลล์เม็ดเลือดขาวมาสู้โรคได้ไม่ทัน หรือผู้ที่มีโรคปอดเรื้อรังอยู่แล้ว รวมทั้งคนที่สูบบุหรี่บ่อย ๆ ก็อาจทำให้ปอดติดเชื้ออย่างรุนแรงและรวดเร็วขึ้น ข้อมูลจาก ศ. นพ.ธีระรัฒน์ เหมะจุฑา เผยว่า เชื้อโคโรนาไวรัสจะมีชีวิตอยู่ได้ที่อุณหภูมิประมาณ 20-40 องศาเซลเซียส โดยสามารถอยู่บนพื้นผิวได้นานถึง 20 วัน ในสภาพอากาศเย็น และในสภาพอากาศร้อน เชื้อไวรัสจะอยู่ได้ 3-9 วัน จากการศึกษาไวรัสที่มีลักษณะคล้ายกัน พบว่าสามารถอยู่บนพื้นผิวโลหะ แก้ว ไม้ หรือพลาสติก ประมาณ 4-5 วัน ณ อุณหภูมิห้อง แต่ในสภาพภูมิอากาศประมาณ 4 องศาเซลเซียส เชื้อจะอยู่ได้ราว ๆ 28 วัน ในกรณีอุณหภูมิมากกว่า 30 องศาเซลเซียส อายุเชื้อไวรัสจะสั้นลง และในสภาพความชื้นที่มากกว่า 50% เชื้อไวรัสจะ อยู่ได้นานกว่าสภาพความชื้นที่ 30%เชื้อไวรัสตัวนี้ไม่ทนความร้อน ดังนั้นแค่เจออุณหภูมิ 70 องศาเซลเซียสก็ทำให้เชื้อตายได้ นอกจากนี้เชื้อไวรัสตัวนี้ยังจะตายได้ง่าย ๆ ด้วยแอลกอฮอล์ที่ความเข้มข้น 70% และการทำความสะอาดด้วยสบู่อย่างเหมาะสม กล่าวคือ ล้างมือด้วยสบู่เป็นระยะเวลา 15-30 วินาที รวมไปถึงสารลดแรงตึงผิวต่าง ๆ เช่น ผงซักฟอก สารฟอกขาว (Sodium hypochlorite) ที่ความเข้มข้น 0.1-0.5% โพวิโด ไอโอดีน 1% หรือไฮโดรเจนเปอร์ออกไซด์ 0.5-7.0% เป็นต้น สิ่งเหล่านี้แหละที่โคโรนาไวรัสจะไม่ทน เพราะไวรัสชนิดนี้มีไขมันหุ้มอยู่ด้านนอก ดังนั้น หากใช้สารลดแรงตึงผิวทำลายไขมันที่หุ้มอยู่ได้ ก็จะฆ่าไวรัสได้ วิธีป้องกันโคโรนาไวรัส สามารถป้องกันการติดเชื้อโคโรนาไวรัสได้ ดังนี้

1. หลีกเลี่ยงการอยู่ในพื้นที่ที่มีการระบาดของโรค
2. สวมหน้ากากอนามัย ซึ่งจะช่วยลดความเสี่ยงในการสูดดมละอองฝอยขนาดใหญ่ได้ถึง 80%
3. อยู่ห่างจากผู้ป่วย หรือผู้ที่มีอาการไอ จาม อย่างน้อย 2 เมตร
4. ล้างมือบ่อย ๆ โดยเฉพาะหลังจับหรือใช้ของสาธารณะหลังเข้าห้องน้ำ เป็นต้น
5. หลีกเลี่ยงการเอามือสัมผัสใบหน้า และดวงตา
6. กินอาหารปรุงร้อน สดใหม่ และใช้ช้อนกลางทุกครั้ง

2.1.2 การทำเหมืองความคิดเห็น

การทำเหมืองความคิดเห็น (Opinion Mining) เป็นกระบวนการวิเคราะห์ความคิดเห็นจากข้อความต่าง ๆ ตัวอย่างเช่น สมศักดิ์ วิชัยกิจ [12] การทำเหมืองความคิดเห็นนั้นเป็นกระบวนการ

ที่จัดการกับข้อความจำนวนมาก เพื่อค้นหารูปแบบ แนวทางทัศนคติเชิงบวกและเชิงลบ ความสัมพันธ์ที่ซ่อนอยู่ในความคิดเห็นนั้น ๆ โดยเป็นการใช้หลักการทางสถิติ การรู้จำ หลักการทางคณิตศาสตร์ การประมวลผลเอกสาร การประมวลผลข้อความและการประมวลผลภาษาธรรมชาติ มีการนำเทคนิคการทำเหมืองความคิดเห็นมาประยุกต์ใช้ในการวิเคราะห์เหมืองความคิดเห็นซึ่ง กานดา แผ่วัฒนากุล [13] กล่าวถึงการวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายออนไลน์ว่า เป็นกระบวนการวิเคราะห์ข้อคิดเห็นที่ผู้บริโภคที่พบได้อย่างมากมายในเครือข่ายสังคมออนไลน์ ทำให้ทราบความรู้สึกของผู้บริโภคทั้งเชิงบวกและเชิงลบ ทำให้ได้ข้อมูลนำไปสู่การปรับปรุงคุณภาพของสินค้าหรือบริการที่จะช่วยเพิ่มความสามารถในการแข่งขัน ลดระยะเวลาและค่าใช้จ่ายในการสำรวจตลาด รวมถึงปรับกลยุทธ์สนองตอบต่อความต้องการของผู้บริโภคที่เปลี่ยนแปลงอย่างรวดเร็ว พัชรนิกานต์ พงษ์ธนู [14] โมเดลการวิเคราะห์เหมืองข้อความจากการเก็บข้อมูลการแสดงความพึงพอใจของลูกค้าจากข้อความคิดเห็น คำแนะนำบนเว็บไซต์ ในความคิดเห็นด้านดีและด้านไม่ดี เพื่อสรุปการให้บริการของเว็บไซต์บริการ ซึ่งได้เปรียบเทียบผลจากการสร้างโมเดลด้วยเทคนิควิธีต้นไม้ตัดสนใจและวิธีการแบบเบย์อย่างง่าย โดยมีวัตถุประสงค์เพื่อหาแนวทางในการปรับปรุงการบริการของเว็บไซต์ผู้ให้บริการโรงแรมให้มีประสิทธิภาพมากขึ้น

กระบวนการทำเหมืองความคิดเห็น มีกระบวนการคล้ายคลึงกับกระบวนการค้นหาความรู้จากฐานข้อมูล ทั้งนี้หากผลจากการวิเคราะห์ในแต่ละขั้นตอนมีความถูกต้องหรือความน่าเชื่อถือต่ำเกินไป จะต้องกลับไปขั้นตอนที่ต่ำกว่า หรือทำการเลือกข้อมูลมาใหม่เพื่อทำให้กระบวนการทำเหมืองความคิดเห็นมีคุณภาพ โดยสามารถแบ่งกระบวนการทำงานออกเป็น 5 ขั้นตอนสำคัญ ดังภาพที่ 2.1

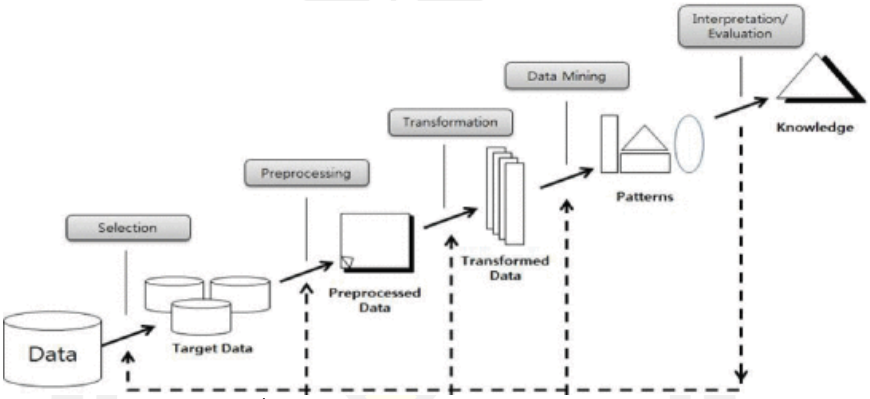
1. การเลือกข้อมูล (Selection) เป็นการรวบรวมข้อมูล ที่จะนำมาใช้ในการทำเหมืองความคิดเห็น รวมถึงการนำความคิดเห็นที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาในเบื้องต้นตามขอบเขตที่ต้องการทำการศึกษา

2. การเตรียมข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลความคิดเห็นที่จะนำมาใช้วิเคราะห์ว่ามีความถูกต้อง โดยการนำความคิดเห็นที่ไม่ถูกต้องออกหรือเป็นขั้นตอนที่อาจต้องแก้ไขก่อนนำไปประมวลผลต่อไป เช่น การตัดอักขระพิเศษ สัญลักษณ์บ่งบอกอารมณ์

3. การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลความคิดเห็นให้เหมาะสมและตรงกับรูปแบบที่จะประมวลผลต่อไป เช่น การตัดบางคอลัมน์ที่ไม่จำเป็นออก

4. การทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนประมวลผล โดยการสร้างแบบจำลองเพื่อการจำแนก จากเทคนิคเหมืองข้อมูล หรือเทคนิคแมชชีนเลิร์นนิ่ง

5. การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความและการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรมีการนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย



ภาพที่ 2. 1 กระบวนการเหมืองความคิดเห็น

2.1.3 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะมี 3 วิธี การคัดเลือกคุณลักษณะด้วย Chi-Square การคัดเลือกคุณลักษณะด้วย TFIDF และการคัดเลือกคุณลักษณะด้วย BM25

2.1.3.1 การคัดเลือกคุณลักษณะด้วย Chi-Square

สถิติที่ใช้ทดสอบความแตกต่างค่าเฉลี่ย ของกลุ่มตัวอย่างที่มีเพียงกลุ่มหรือสองกลุ่ม จะใช้ ทดสอบด้วยค่า Z-test หรือ T-test ข้อมูลที่นำมาทดสอบนั้นจะต้องเป็นข้อมูลที่อยู่ในระดับการวัด (Measurement Scale) ระดับอันตรภาคชั้น (Interver Scale) หรือระดับอัตราส่วน (Ratio Scale) เท่านั้นในงานวิจัยบางเรื่องข้อมูลอาจอยู่ในรูปของความถี่ที่เป็นอิสระต่อกัน (Discrete Data) เป็น ข้อมูลที่อยู่ในระดับนามบัญญัติ (Nominal Scale)หรือ ข้อมูลเรียงลำดับ(Ordinal Scale) การทดสอบ ข้อมูลในลักษณะนี้ จะเป็นการทดสอบว่า ข้อมูลที่ได้เป็นไปตามคาดหวัง (Expected) ไว้หรือไม่ หรืออาจจะทดสอบว่าตัวแปร (Variable) มีความสัมพันธ์กันหรือไม่ ข้อมูลดังกล่าวไม่สามารถทดสอบได้ด้วย Z-test หรือ T-test ซึ่งเป็นสถิติแบบพารามิตริก (Parametric Statistics) แต่จะสามารถทดสอบได้ด้วยไคสแควร์ (x^2) ซึ่งเป็นสถิติแบบนอนพารามิตริก (Nonparametric Statistics) ซึ่งเป็นสถิติที่ไม่คำนึงถึงลักษณะการแจกแจงของประชากร

การทดสอบไคสแควร์

การทดสอบโดยใช้ x^2 จะมีอยู่ 2 กรณี คือ

1. การทดสอบกรณีตัวแปรเดียว (The x^2 one - variable case หรือ บางครั้งอาจเรียกว่า การทดสอบภาวสารูปสนิทดี (Goodness of fit test)

2. การทดสอบกรณีสองตัวแปร (The x^2 two - variable case) เป็นการทดสอบเพื่อดูว่าตัวแปรสองตัวมีความสัมพันธ์หรือเกี่ยวข้องกันหรือไม่ ดังนั้นบางทีจึงเรียกว่า การทดสอบความเป็นอิสระ (The x^2 test for independence)

การทดสอบภาวสารูปสนิทดี มีวัตถุประสงค์เพื่อ ใช้เพื่อทดสอบว่าการแจกแจงความถี่สอดคล้องกับการแจกแจงคาดหวังหรือไม่ การทดสอบนี้สามารถใช้กับข้อมูลเชิงคุณภาพ

การทดสอบภาวสารูปสนิทดี เป็นการทดสอบสำหรับตัวอย่าง 1 กลุ่มทดสอบเกี่ยวกับการแจกแจงของตัวแปร 1 ตัว ว่าเป็นไปตามสัดส่วนที่กำหนดไว้หรือไม่ ตัวแปรมีสเกลการวัดแบบแบ่งประเภท (Nominal scale) ข้อมูลมาจากประชากรกลุ่มเดียว

การทดสอบ x^2 กรณีกลุ่มตัวอย่างกลุ่มเดียว เป็นการทดสอบตัวแปรเพียงด้านเดียวเพื่อต้องการ ทราบว่าความถี่ที่ได้จากการสังเกต (Observed Frequency)จากกลุ่มตัวอย่างเป็นไปตามความถี่ที่คาดหวัง (Expected Frequency) ไว้หรือไม่ตามนัยสำคัญที่กำหนด การทดสอบโดยการใช้สูตร คำนวณ x^2 test ดังสมการที่ 2.1

$$x^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \qquad (2.1)$$

x^2 = ค่าสถิติไคสแควร์

Oi = ความถี่ที่ได้จากการสังเกต (Observed Frequency)

Ei = ความถี่ที่คาดหวัง (Expected Frequency)

ซึ่งมีค่าเท่ากับ จำนวนข้อมูลคูณด้วย สัดส่วนที่คาดหวัง

k = จำนวนกลุ่มหรือจำนวนระดับ

n = จำนวนตัวอย่าง

pi = สัดส่วนของประชากรที่มีลักษณะหรือระดับที่ i

** ซึ่ง Ei = npi**

การทดสอบความเป็นอิสระ การทดสอบในกรณีตัวแปรสองตัวนี้เป็นการทดสอบเพื่อดูว่า ตัวแปรสองตัวนี้มีความเกี่ยวข้องหรือสัมพันธ์กันหรือไม่ ถ้าไม่สัมพันธ์กันหมายความว่าเป็นอิสระจากกัน ดังนั้นบางครั้งเรา จึงเรียกว่า การทดสอบความเป็นอิสระ (x' test for dependence) ข้อมูลที่ได้จะอยู่ในระดับนามบัญญัติ (Norminal scale) ซึ่งอาจเป็นจำนวนความถี่ สัดส่วน ร้อยละ ก็ได้ โดยแต่ละตัวแปรจะแบ่งเป็น 2 กลุ่ม หรือ 2 ประเภทขึ้นไป เช่น เพศ (ชาย - หญิง) กับวุฒิการศึกษา (ป.ตรี ป .โท ป.เอก) จะได้รูปแบบเป็น 2 x 3 ดังนั้นรูปแบบการวิเคราะห์อาจเป็นได้หลายรูปแบบขึ้นอยู่กับจำนวน กลุ่มของแต่ละตัวแปร (2 x 2, 2 x 4, 3 x 2 เป็นต้น)

2.1.3.2 การคัดเลือกคุณลักษณะด้วยTFIDF

term frequency-inverse document frequency เป็นเทคนิคที่พิจารณาองค์ประกอบของคำภายในประโยค (และเอกสาร) เป็นหลักโดยจะไม่นำลำดับของคำภายในเอกสารมาใช้วิเคราะห์ประกอบด้วย เทคนิคนี้มีอยู่ 2 องค์ประกอบด้วยกันคือ Term Frequency (TF) และ Inverse Document Frequency (IDF) ซึ่งมีที่มาจากสองไอเดียหลักๆ ต่อไปนี้ การคำนวณค่าของทั้ง TF และ IDF นั้น มีหลายรูปแบบ สิ่งที่เลือกมานำเสนอเป็นรูปแบบพื้นฐานของทั้งสองค่า

Term-Frequency (TF) ค่าของ Term Frequency เป็นค่าที่บอกความถี่ของคำแต่ละคำที่ปรากฏในเอกสารเอกสารหนึ่ง โดยคิดคำนวณจากการนำจำนวนครั้งที่คำนั้น ๆ ปรากฏในเอกสารมาหารด้วยจำนวนคำทั้งหมดในเอกสาร การหาคำที่เกิดขึ้นในเอกสารที่ได้จากความถี่ของคำ ๆ นั้น ในจำนวนเอกสารทั้งหมดซึ่งเป็นวิธีการที่เป็นที่นิยมและมีประสิทธิภาพมากพอสมควร แต่วิธีการนี้ยังถือว่ามีข้อบกพร่องอยู่ เพราะเมื่อเกิดค่าความถี่ของคำสูงนั่นก็คือ คำนั้น ๆ มีโอกาสเกิดขึ้นในหลายข้อความพร้อมกัน ซึ่งมีโอกาสเกิดข้อผิดพลาดตในการวิเคราะห์เพื่อการจำแนกข้อความ เช่น ถ้าหากคำไหนถูกพูดถึงอยู่บ่อยในเอกสารนั้น จะมีความเป็นไปได้สูงว่าคำนั้นมีความเกี่ยวข้องกับใจความสำคัญของเอกสารนั้นมาก ถ้าหากเลือกที่จะพิจารณาบทความเกี่ยวกับโรคโควิด อาจจะเห็นคำว่า “โควิด” และ “ติด” หลายครั้งภายในบทความนั้น ดังสมการที่ 2.2

$$TF(\text{ของคำคำหนึ่ง}) = \frac{\text{จำนวนของคำนั้นๆ ในเอกสาร}}{\text{จำนวนของคำทั้งหมดในเอกสาร}} \quad (2.2)$$

เพื่อแสดงการตัวอย่างการคำนวณค่า TF (รวมถึง ค่า IDF และ TF-IDF ในตัวอย่างต่อไป) ดังนั้นเราสามารถคำนวณค่า Term Frequency ของคำแต่ละคำ อย่างไรก็ตามการที่ใช้คำที่ปรากฏบ่อย ๆ เพียงอย่างเดียวเพื่อหาใจความสำคัญของข้อความนั้นอาจไม่ดีพอ ยกตัวอย่างเช่นในกรณีที่เรามีหลายเอกสารในหมวดหมู่เดียวกัน เราต้องการคำที่สามารถแยกแยะเอกสารแต่ละชิ้นออกจากกันได้ด้วย (เช่นในกรณีที่เราต้องการหาใจความสำคัญของบทความเกี่ยวกับโรคโควิดก็คงไม่ต้องการให้คำสำคัญที่สุดของทุกบทความที่หามาได้เป็นคำว่า “โควิด”,”ระบาด” หรือ ”แพร่เชื้อ” ไปทั้งหมด) ดังนั้นคำที่ถูกเลือกมาว่าเป็นคำที่มีความสำคัญที่สุดต่อความหมายของบทความนั้นควรจะเป็นคำที่ปรากฏอยู่ในเอกสารจำนวนไม่มากในบรรดาเอกสารทั้งหมดที่เรานำมาวิเคราะห์ ความต้องการในจุดนี้ทำให้ไอเดียของ Invert Document Frequency (IDF) ถูกนำมาใช้

Inverse Document Frequency (IDF) เป็นวิธีการสร้างตัวแทนเอกสารในรูปแบบของเวกเตอร์ที่มีความน่าเชื่อถือและได้รับความนิยมเป็นอย่างยิ่ง โดยจะมีการหาค่าความถี่ของคำที่ปรากฏในเอกสารและจากนั้นจะทำการหาค่าส่วนกลับของเอกสารหรือค่าน้ำหนักของความถี่เอกสารผกผันเป็นการคำนวณค่าน้ำหนัก (weight) ความสำคัญของแต่ละคำโดยจะคำที่พบเจอได้บ่อย

(ในหลายๆเอกสาร) จะมีค่า IDF ต่ำ ซึ่งบ่งบอกว่าคำเหล่านั้นจะไม่สามารถดึงเอาจุดเด่นของเอกสารที่คำเหล่านั้นปรากฏอยู่ออกมาได้ดี ค่า IDF สามารถคำนวณได้ ดังสมการที่ 2.3

$$IDF\left(\text{ของคำคำหนึ่ง}\right) = \log\left(\frac{\text{จำนวนเอกสารทั้งหมดที่ใช้พิจารณา}}{\text{จำนวนเอกสารที่มีคำคำนั้นปรากฏอยู่}}\right) \quad (2.3)$$

สำหรับตัวอย่างการคำนวณค่า IDF นั้น เราจะใช้ตัวอย่างเอกสารชุดเดียวกับกรณียกตัวอย่างการคำนวณ ค่า TF ด้านบน เมื่อพิจารณาเอกสารแล้วมีทั้งหมด 2,920 เอกสารทั้งหมดที่ใช้ในการพิจารณา ดังนั้นจึงสามารถคำนวณค่า IDF

คำนวณค่า TF-IDF เมื่อนำการคำนวณทั้งสองส่วนมารวมกัน เราจะได้การคำนวณ TF-IDF ดังสมการที่ 2.4

$$TFIDF = TF \times IDF \quad (2.4)$$

2.1.3.3 การคัดเลือกคุณลักษณะ ด้วย BM25

เป็นฟังก์ชันอันดับที่จัดลำดับกลุ่มของเอกสารขึ้นอยู่กับคำค้นหาที่ปรากฏในเอกสารแต่ละฉบับ เกี่ยวกับความสัมพันธ์ระหว่างคำค้นหาที่อยู่ในเอกสาร ฟังก์ชันนี้ไม่ใช่ฟังก์ชันเดียวแต่โดยพื้นฐานแล้วเป็นฟังก์ชันการให้คะแนนทั้งหมดที่มีกลไกและข้อ จำกัด ที่แตกต่างกันเล็กน้อย [15] นอกจากนี้เครื่องมือค้นหายังใช้เพื่อจัดอันดับเอกสารที่เกี่ยวข้องตามความเกี่ยวข้องกับการดึงข้อมูลที่จัดอันดับชุดของเอกสารตามคำค้นหาที่ปรากฏในเอกสารแต่ละฉบับโดยไม่คำนึงถึงความใกล้ชิดกันภายในเอกสาร เป็นกลุ่มฟังก์ชันการให้คะแนนที่มีส่วนประกอบและพารามิเตอร์ต่างกันเล็กน้อย หนึ่งในอินสแตนซ์ที่โดดเด่นที่สุดของฟังก์ชันมีดังสมการที่ 2.5

$$BM25 = \frac{f(q,D)*(k+1)}{f(t,D)+k*(1-b+b*\frac{D}{d_{avg}})} * \log\left(\frac{N-N(q)+0.5}{N(q)+0.5}+1\right) \quad (2.5)$$

f(q,D) คือจำนวนครั้งที่คำว่า q เกิดขึ้นในเอกสาร D

q คือจำนวนครั้งที่คำเกิดขึ้นในเอกสารนั้น

D คือจำนวนคำในเอกสารทั้งหมด

d_{avg} คือจำนวนคำเฉลี่ยต่อเอกสารทั้งหมด

k ค่าเริ่มต้นประมาณ 1.2

b ค่าเริ่มต้นประมาณ 0.75

b และ k เป็นไฮเปอร์พารามิเตอร์ที่ปรับได้สำหรับ BM25

N จำนวนเอกสาร

N(q) เอกสารที่มีคำที่ค้นหานั้น

0.5 ค่าคงที่

1 ค่าคงที่

อย่างไรก็ตาม BM25 ไม่ใช่ฟังก์ชันเดียวที่ใช้ในการแก้ไขเอกสารที่ยาวมาก ปัญหานี้ไม่สามารถระบุถึงความเป็นประโยชน์หรือความสำคัญของคุณลักษณะที่มั่นใจได้ และลดประสิทธิภาพของการจัดประเภท

2.1.4 เทคนิคที่ใช้ในงานวิจัย

การสร้างแบบจำลองเพื่อการจำแนก (Text Mining)ที่เหมาะสมที่สุดในการจัดกลุ่มความรู้สึกของคนไทยต่อโรค COVID 19 บน Social Media โดยใช้เทคนิคการของการทำเหมืองข้อความมาใช้ในการวิเคราะห์ทั้งหมด 6 เทคนิค คือ เทคนิคต้นไม้ตัดสินใจ แบบ C4.5 (Decision tree C4.5) เทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine :SVM) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (multi-layer perceptron) เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) จากนั้นใช้หลักการ 10-fold cross validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลเรียนรู้และชุดข้อมูล ทดสอบและวัดประสิทธิภาพของแบบจำลองด้วย ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

2.1.4.1 เทคนิคต้นไม้ตัดสินใจ แบบ C4.5

เทคนิคต้นไม้ตัดสินใจ แบบ C4.5 (Decision tree C4.5) คือ เป็นการเรียนรู้โดยการจำแนก(Classification) ชุดข้อมูล ออกเป็นกลุ่ม (Class) ต่าง ๆ โดยใช้คุณสมบัติ (Attribute) ของข้อมลูในการแยกแยะต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ท าให้ทราบคุณสมบัติใดของ ข้อมูลที่เป็นตัวก าหนดการจำแนก และคุณสมบัติแต่ล่ะตัวของข้อมูลนั้นมีความสำคัญมากน้อยอย่างไร ซึ่งเป็นประโยชน์ที่จะช่วยให้ ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้น ผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจประกอบด้วย 1) โหนดภายใน (Internal Node) 2) กิ่ง (Branch, Link) คือ ค่าคุณสมบัติของคุณสมบัติในโหนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่ง เป็นจำนวนเท่ากับจำนวนค่าคุณสมบัติของโหนดภายในนั้น 3) โหนดใบ (Leaf Node) คือ กลุ่ม (Class) ต่าง ๆ ซึ่งเป็นผลลัพธ์ในการ จำแนกข้อมูล ซึ่งเป็นขั้นตอนวิธีหนึ่งที่ใช้ในการสร้างต้นไม้ตัดสินใจจะใช้ค่าที่วัดที่เรียกว่าค่าแกน (Gain) เป็นตัวตัดสินใจว่าจะใช้แอตทริบิวท์ (Attribute) ใดในการแบ่งข้อมูลเพื่อใช้ในการตัดสินใจ โดยวิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นตามลำดับของค่าตัวชี้วัดหรือค่าเกนของ Attribute ที่มีค่าสูงที่สุด ซึ่งเริ่มหาค่าสารสนเทศ (Information) แต่ละ Attribute ซึ่งบ่งบอกความสามารถของแอตทริบิวท์ ในการแยกแต่ละคลาส (Class) ดังสมการที่ 2.6

$$I(S_1, S_2, \ldots, S_n) = -\sum_{i=1}^{n} \frac{si}{s} log_2 \frac{si}{s} \qquad (2.6)$$

โดย S คือ จำนวนข้อมูลทั้งหมด ซึ่งมีจำนวนข้อมูลของ n คือ จำนวน Class ทั้งหมด

จากนั้นหาค่าเอนโทปี (Entropy) เป็นค่าของผลรวมแต่ละ Attribute A ที่สามารถแยกประเภทข้อมูลสารสนเทศ (Information) แต่ละคลาสของแต่ละคฎ ของAttribute m จำนวน ดังสมการที่ 2.7

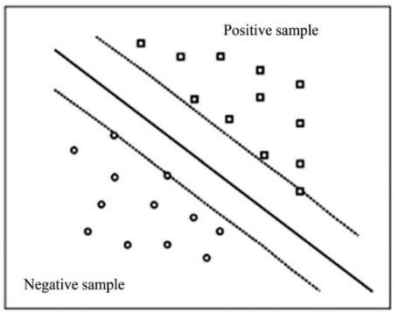
$$E(A) = \sum_{j=1}^{m} \frac{s_{1j}+\ldots+s_{nj}}{s} I(S_{1f},\ldots,S_{2j}) \qquad (2.7)$$

ดังนั้นค่าเกน (Gain) ของแอตทริบิวท์ที่ถูกเลือกหาได้จากการแยกประเภทข้อมลู สารสนเทศทั้งหมดของ แอตทริบิวท์นั้น ดังสมการที่ 2.8

$$\mathrm{Gain(A)} = \mathrm{I(S1, S2, \ldots, Sn)} - \mathrm{E(A)} \qquad (2.8)$$

2.1.4.2 เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

เทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine :SVM) คือขั้นตอนวิธีการที่มีความรวดเร็วและเป็นเทคนิคที่มีความสามารถนำมาช่วยแก้ปัญหาการจำแนกข้อมูล โดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มขอข้อมูลที่ถูกป้อนเข้าสู้กระบวนการสอบแบบให้ระบบเรียนรู้โดยเน้นไปที่เส้นแบ่งแยกกลุ่มของข้อมูลได้ดีที่สุด แนวความคิดของเทคนิควิธี SVM นั้นเกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางในฟีเจอร์สเปซ (Feature Space)ในลักษณะเชิงเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา เพื่อให้ทราบว่าเส้นตรงที่แบ่งกลุ่มสองกลุ่มออกจากกันนั้นเส้นใดเป็นเส้นที่ดีที่สุดดังตัวอย่างดังภาพที่ 2.2



ภาพที่ 2. 2 Support Vector Machine :SVM

2.1.4.3 เทคนิคนาอีฟเบย์

เทคนิคนาอีฟเบย์ (Naïve Bayes) มีพื้นฐานมาจากกฎของเบย์ เป็นทฤษฎีทางด้านสถิติโดยน่าความน่าจะเป็นมาใช้ ประเมินความไม่แน่นอนให้เป็นตัวเลขได้ กล่าวถึงความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น (A) ถ้ามีเหตุการณ์อีกเหตุการณ์หนึ่งเกิดมาแล้ว (B) สามารถเขียนให้อยู่ในรูปอย่างง่าย ดังสมการที่ 2.9

$$P(A|B) = \frac{\mathrm{P(B|A)P(A)}}{\mathrm{P(B)}} \qquad (2.9)$$

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 สามารถคาดการณ์ผลลัพธ์ และสามารถอธิบายได้ ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์เป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยใช้ในการจำแนกหมวดหมู่เอกสารข้อความ (Text Classification) ได้ดี การทำงานไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ไม่ขึ้นต่อกัน โดย กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็น ดังสมการที่ 2.10

$$P(A_1, A_2, \ldots, A_n|C_j) = \prod_{i=1}^{n} P(A_i|C_i) \qquad (2.10)$$

กลุ่ม **Ci** สำหรับข้อมูลที่มีคุณสมบัติ n ตัว **X** = (**A1, A2,..., An**) หรือใช้สัญญาลักษณ์ว่า P (**A1, A2,..., An|Cj**) โดย ที่ **∏** หมายถึงผลคูณของค่า **P(Ai|Ci)** ทั้งหมด i = 1, 2, 3, ..., n และ j = 1, 2, 3,..., n ดังนั้นจะได้วิธีการจำแนกประเภทแบบเบย์ อย่างง่าย ดังสมการที่ 2.11

$$V_{NB} = argmax\, P\,(C_j) \prod_{i=1}^{n} P(A_i|C_i) \qquad (2.11)$$

2.1.4.4 เทคนิคเคเนียเรสเนเบอร์

เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor: KNN) คือ เป็นวิธีการในการจัดแบ่งคลาส หลักการของวิธีการนี้ จะจำแนกประเภทข้อมลูโดยขึ้นกับข้อมูล มีคุณสมบัติใกล้เคียงที่สุด k ตัวจากข้อมลูบนชุดข้อมูล ตัวอย่างทางานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิก ใหม่หรือข้อมูลที่ป้อนถาม (Input Query Instance) กับข้อมูลตัวอย่างฝึกฝนจะคำนวณหาเพื่อนบ้านที่ใกล้

ที่สุด k ตัว หลังจากนั้นเรา จะรวบรวมสมาชิกที่ใกล้เคียงที่สุด k ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม k ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมลูการจำแนกโดยใช้ข้อมลูข้างเคียง k ตัว ประกอบด้วยแอตทริบิวท์ หลายตัวแปร ซึ่งจะ นำมาใช้ในการแบ่งกลุ่ม โดยระบุ ค่าตัวเลขจำนวนเต็มบวกให้กับ k ซึ่งค่านี้จะเป็นตัวบอกจำนวนของ กรณี (case) ที่จะต้องค้นหาในการทำนายหากรณีใหม่ เทคนิค แบบ KNN ได้แก่ 1-NN , 2-NN , 3-NN , ... k-NN โดยค่า k ต้องระบุในการสร้างโมเดล [13, 14] มาตรวัดความถูกต้อง (Distance Measure) การหาความยาวระหว่างจุดที่ต้องการโดยใช้เครื่องมือ และวิธีต่าง ๆ งานวิจัยได้เลือกวิธีการหามาตรวัดความแม่นยำ โดย Euclidean Distance ระยะทาง ระหว่าง 2 จุด จุดที่จะวัดนั้นมีเงื่อนไขมีหลายค่าจากหลายมิติหรือขนาดขึ้นกับรูปแบบ ซึ่งสามารถ พิสูจน์หาค่าได้ด้วยทฤษฎีของ Pythagorean เมื่อมีการใช้สูตรเพื่อหาระยะทางขนาดของ Euclidean ระยะทางระหว่างจุด P = (p1,p2...,p3) และ (Q = q1,q2,...,q3) ใน Euclidean หลายขนาดระบุได้เป็น ดังสมการที่ 2.12

$$\sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \qquad (2.12)$$

2.1.4.5 เทคนิคเพอร์เซปตรอนหลายชั้น

เพอร์เซ็ปตรอนหลายชั้น (multi-layer perceptron) โครงข่ายประสาทเทียมแบบ MLP เป็นรูปแบบหนึ่งของโครงข่ายประสาทเทียมที่มีโครงสร้างเป็นแบบหลายๆชั้น ใช้สำหรับงานที่มีความซับซ้อนได้ผลเป็นอย่างดี โดยมีกระบวนการฝึกฝนเป็นแบบมีผู้สอน (Supervise) และใช้ขั้นตอนการส่งค่าย้อนกลับ (Backpropagation) สำหรับการฝึกฝนกระบวนการส่งค่าย้อนกลับประกอบด้วย 2 ส่วนย่อยคือการส่งผ่านไปข้างหน้า (Forward Pass) การส่งผ่านย้อนกลับ (Backward Pass) สำหรับการส่งผ่านไปข้างหน้าข้อมูลจะผ่านเข้าโครงข่ายประสาทเทียมที่ชั้นข้อ มูลเข้า และจะส่งผ่าน จากอีกชั้นหนึ่งไปสู่อีกชั้นหนึ่งจนกระทั่งถึงชั้นข้อมูลออก ส่วนการส่งผ่านย้อนกลับค่าน้ำหนักการเชื่อมต่อจะถูกปรับเปลี่ยนให้สอดคล้องกับกฎการแก้ข้อผิดพลาด (Error-Correction) คือผลต่างของผลตอบที่แท้จริง (Actual Response) กับผลตอบเป้าหมาย (Target Response) เกิดเป็นสัญญาณผิดพลาด (Error Signal) ซึ่งสัญญาณผิดพลาดนี้จะถูกส่งย้อนกลับเข้าสู่โครงข่ายประสาทเทียมในทิศทางตรงกันข้ามกับการเชื่อมต่อ และค่าน้ำหนักของการเชื่อมต่อจะถูกปรับจนกระทั่งผลตอบที่แท้จริงเข้าใกล้ผลตอบเป้าหมาย สัญญาณที่มีโครงข่ายประสาทเทียมแบบ MLP มี 2 ประเภท คือ Function Signal และ Error Signal 1.Function Signal เป็นสัญญาณเข้าที่มาจากโหนดในชั้นก่อนหน้า และจะส่งผ่านไปข้างหน้าจากโหนดหนึ่งไปสู่อีกโหนดหนึ่ง 2.Error Signal เป็นสัญญาณย้อนกลับที่เกิดขึ้นที่โหนดในชั้นข้อมูลออกของโครงข่ายประสาทเทียม และถูกส่งผ่านย้อนกลับจากชั้นหนึ่งไปสู่อีกชั้นหนึ่ง

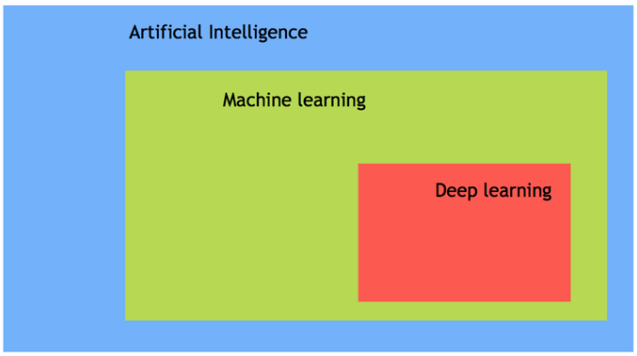
หลักการทำงานของ MLP คือในแต่ละชั้นของชั้นซ่อนตัว (Hidden Layer) จะมีฟังก์ชันสำหรับคำนวณเมื่อได้รับสัญญาณ (Output) จากโหนดในชั้นก่อนหน้านี้เรียกว่า Activation Function โดยในแต่ละชั้นไม่จำเป็นต้องเป็นฟังก์ชันเดียวกันก็ได้ ชั้นซ่อนตัวนั้นมีหน้าที่สำคัญคือ จะพยายามแปลงข้อมูลที่เข้ามาในชั้น (Layer) นั้นๆให้สามารถแยกแยะความแตกต่างโดยใช้เส้นตรงเส้นเดียว (Linearly Separable) และก่อนที่ข้อมูลจะถูกส่งไปถึงชั้นข้อมูลออก (Output Layer) ในบางครั้งอาจจำเป็นต้องใช้ชั้นซ่อนตัวมากกว่า 1 ชั้นในการแปลงข้อมูลให้อยู่ในรูป Linearly Separable

ในการคำนวณหา Output ในปัญหาการจำแนกทำได้โดยการใส่ข้อมูล Input เข้าไปในโครงข่ายประสาทเทียมที่เราได้ทำการหาไว้แล้ว จากนั้นให้ทำการเปรียบเทียบค่าของ Output ใน Output Layer และให้ทำการเลือกค่าของ Output ที่มีค่าสูงกว่า (Neuron ที่มีค่าสูงกว่า) และทำการรับค่าของพยากรณ์ที่ตรงกับ Neuron ที่เลือก และให้นำค่าของ มาเปรียบเทียบกับค่าที่ยอมรับได้ หากค่าของ อยู่ในช่วงที่รับได้ (Error น้อยกว่า Error ที่เรากำหนด) ก็ให้ทำการรับข้อมูลชุดถัดไป แต่หากค่าของ มากกว่าค่าที่ยอมรับได้ ให้ทำการปรับค่าน้ำหนักและ Biased ตามขั้นตอนที่ได้กล่าวไว้ข้างต้น เมื่อทำการปรับน้ำหนักเรียบร้อยแล้ว ให้ทำการรับข้อมูลชุดถัดไปและทำตามขั้นตอนซ้ำอีกรอบจนกระทั่งถึงข้อมูลชุดสุดท้าย และเมื่อทำข้อมูลชุดสุดท้ายเสร็จจะนับเป็น 1 รอบของการคำนวณ (1 Epoch) จากนั้นจะทำการหาค่าผิดพลาดรวมเฉลี่ย จากค่าเฉลี่ยของ ที่ได้เก็บค่าเอาไว้ เพื่อใช้ในการตรวจสอบว่าค่า โดยเฉลี่ยในการจำแนกนั้น มีค่าน้อยกว่าค่าผิดพลาดที่ยอมรับได้หรือไม่ ถ้าใช่แสดงว่าโครงข่ายประสาทเทียมที่สร้างขึ้นนั้นสามารถให้ผลลัพธ์ที่ถูกต้องของทุก ๆ ข้อมูลแล้วจึงทำการจบการเรียนรู้ได้ แต่ถ้าไม่ใช่ ให้กลับไปทำตามขั้นตอนแรกโดยเริ่มรับข้อมูลชุดที่ 1

2.1.4.6 เทคนิคแบบระบบเรียนรู้เชิงลึก

เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) หมายถึงเทคนิคในการสร้างปัญญาประดิษฐ์โดยใช้โครงข่ายประสาทเทียมหรือข่ายงานประสาทเทียมหลายๆ ชั้นเหมือนแบบจำลองอันเรียบง่ายของสมองมนุษย์ การเรียนรู้เชิงลึก เป็นสาขาของการเรียนรู้ของเครื่อง พื้นฐานของการเรียนรู้เชิงลึกคือ เทคนิคที่พยายามจะสร้างแบบจำลองเพื่อแทนความหมายของข้อมูลในระดับสูงโดยการสร้างสถาปัตยกรรมข้อมูลขึ้นมาที่ประกอบไปด้วยโครงสร้างย่อย ๆหลายอัน และแต่ละอันนั้นได้มาจากการแปลงที่ไม่เป็นเชิงเส้นการเรียนรู้เชิงลึก อาจมองได้ว่าเป็นวิธีการหนึ่งของการเรียนรู้ของเครื่องที่พยายามเรียนรู้วิธีการแทนข้อมูลอย่างมีประสิทธิภาพ ตัวอย่างเช่น รูปภาพภาพหนึ่ง สามารถแทนได้เป็นเวกเตอร์ของความสว่างต่อจุดพิกเซล หรือมองในระดับสูงขึ้นเป็นเซ็ตของขอบของวัตถุต่าง ๆ หรือมองว่าเป็นพื้นที่ของรูปร่างใด ๆ ก็ได้ การแทนความหมายดังกล่าวจะทำให้การเรียนรู้ที่จะทำงานต่าง ๆ ทำได้ง่ายขึ้น ไม่ว่าจะเป็นการรู้จำใบหน้าหรือการรู้จำการแสดงออกทางสีหน้า การเรียนรู้เชิงลึกถือว่าเป็นวิธีการที่มีศักยภาพสูงในการจัดการกับฟีเจอร์สำหรับการเรียนรู้

แบบไม่มีผู้สอนหรือการเรียนรู้แบบกึ่งมีผู้สอนเหมาะกับเทคนิคMachine learning ที่มีจุดมุ่งหมายในการสอนเครื่องจักรกลให้วิเคราะห์ข้อมูลตามการตัดสินใจของตัวมันเองแทนการใช้เทคนิคเทคนิคที่ถูกกำหนดโดยมนุษย์ ซึ่งจะกำหนดล่วงหน้าสำหรับงานเฉพาะด้านไว้กระบวนการ Deep learning มีรากฐานมาจาก Neocortex หรือส่วนหนึ่งของเปลือกสมองในสัตว์เลี้ยงลูกด้วยนมโดย Deep learning จะจัดเรียงโนดการวิเคราะห์ในชุดเส้นทางสำหรับข้อมูลที่ไหลระหว่างการเชื่อมต่อในโครงข่ายดังเช่นเครือข่ายโนดที่ซ้อนทับกันหลายเลเยอร์ อย่างไรก็ตาม ปัจจุบันนี้มนุษย์ยังไม่สามารถจำลองข้อมูลการเชื่อมต่อหลายชั้นที่ซับซ้อนจนทำให้สมองเป็นดั่งคอมพิวเตอร์ที่ทรงพลังได้เลย ความสามารถในการวิเคราะห์นี้กำลังช่วยขับเคลื่อนเทคโนโลยีแห่งอนาคต เช่น รถยนต์ที่ไร้คนขับซึ่งจะสามารถช่วยให้ระบบรถยนต์รับรู้และจดจำสัญญาณจราจรหรือแยกความแตกต่างของวัตถุต่าง ๆ ในเส้นทางได้เอง



ภาพที่ 2. 3 Artificial Intelligence, Machine Learning, Deep learning

ดังภาพที่ 2.3 แนวคิดของ AI ไม่ใช่เรื่องใหม่อันที่จริงมีบันทึกของปัญญาประดิษฐ์ตั้งแต่ต้นศตวรรษที่ 18 แล้ว ซึ่งมันมาพร้อมกับการคุกคามของเครื่องจักรที่อัจฉริยะพอๆกับมนุษย์ โดยเป็นที่นิยมอย่างกว้างขวางจากวงการภาพยนตร์อย่าง 2001: A Space Odyssey และ The Terminatorและแนวคิดเหล่านี้ก็ไม่ได้เป็นแนวความคิดที่นำเสนอผ่านเพียงภาพยนตร์อีกต่อไป เพราะแนวคิดเหล่านี้ได้กลายเป็นส่วนหนึ่งของชีวิตประจำวันของพวกเรามากขึ้นอย่างการเปิดตัวเทคโนโลยีอย่าง Chat Bot และแอพพลิเคชั่นสุดอัจฉริยะ หรือในอนาคตอาจจะมีเครื่องจักรที่มีความอัจฉริยะเหนือกว่าความฉลาดของมนุษย์ก็ได้ซึ่งไม่ใช่เรื่องเกินจริงเลย คำว่า "ปัญญาประดิษฐ์" เป็นการอธิบายแนวคิดกว้างๆ ของเครื่องจักรที่มีความคิด การตัดสินใจเพื่อตัวมันเอง แต่ในความเป็นจริง ปัญญาประดิษฐ์เพียงหนึ่งคำอาจไม่สามารถอธิบายครอบคลุมความหมายของมันได้ หากเราเริ่มมองดูเทคโนโลยีนี้ ปัญญาประดิษฐ์นั้นสามารถแบ่งออกเป็นสองมิติที่แตกต่างกัน ได้แก่: ปัญญาประดิษฐ์ทั่วไป (General AI) และปัญญาประดิษฐ์เชิงแคบ (Narrow AI) (หรือ Applied AI) ปัญญาประดิษฐ์ทั่วไป (General AI) หมายถึงการศึกษาและออกแบบระบบที่สามารถปฏิบัติงานได้เท่าที่มนุษย์สามารถทำได้ มันอาจเป็นความหมายที่พบบ่อยที่สุดของ AI และสิ่งที่ทำให้เกิดโรค

ฮิสทีเรียมากที่สุด เนื่องจากระบบนี้สร้างความหวาดกลัวให้คน เกี่ยวกับระบบอัตโนมัติต่างและการเพิ่มขึ้นของหุ่นยนต์สังหารมาทำงานแทนที่มนุษย์ ที่ทราบแล้วว่าความสำเร็จในด้านนี้ค่อนข้างจำกัด ในทางกลับกันปัญญาประดิษฐ์เชิงแคบ (Narrow AI) นั้นกลับประสบความสำเร็จมากกว่า เพราะแทนที่จะมุ่งเน้นในการสร้างระบบที่สามารถเลียนแบบมนุษย์โดยทั่วไปได้ แต่ด้านนี้จะเน้นการสร้างเครื่องจักรที่สามารถทำงานเฉพาะทางหรือชุดงานใดก็ตามให้ดีกว่ามนุษย์คนใด ยกตัวอย่าง Chat Bot ที่ออกแบบโดย บริษัท AI Luka ถูกสร้างขึ้นเพื่อวิเคราะห์ตอบสนองข้อความและตอบข้อความในโซเชียวมีเดียอัตโนมัติ เพื่อส่งไปยังเพื่อนของหนึ่งในนักพัฒนาของ Luka ที่เพิ่งเสียชีวิตไปโดยโปรแกรมนี้ได้รับมอบหมายให้วิเคราะห์ข้อมูลเมื่อสี่ปีก่อนเพื่อสร้างความประทับใจในการสื่อสารกับเพื่อน ดังนั้นเมื่อโปรแกรมดึงข้อมูล เหล่านี้มาใช้ก็จะสามารถตอบกลับข้อความด้วยสไตล์ของเพื่อน โดยสะท้อนโทนเสียงและภาษาของเขาได้หลายคนยอมรับว่าเป็นตัวอย่างที่น่าตกใจพอสมควรเพราะเกี่ยวข้องกับความตายของมนุษย์ แต่ก็เป็นการแสดงให้เห็นว่าปัญญาประดิษฐ์เชิงแคบไม่จำเป็นต้องทะเยอทะยานเหมือนกับปัญญาประดิษฐ์ทั่วไปและถึงแม้ว่ามันจะไม่ได้ก้าวหน้าอย่างหุ่นยนต์สังหารจากจิตนาการด้าน Sci-fi ของเราแต่ก็ไม่มีตัวช่วยหรือระบบใดที่สามารถจำลองข้อมูลอีกชุดขึ้นมาจนเกือบเท่าระดับความฉลาดของมนุษย์ได้อย่างปัญญาประดิษฐ์เช่นกัน สิ่งใดที่เป็นไปได้ ส่วนใหญ่ต้องขอบคุณ Machine learning เพราะแทนที่จะเป็นเครื่องจักที่คัดลอกเฉพาะการกระทำของมนุษย์ด้วยคำแนะนำที่ตั้งไว้ล่วงหน้าแต่เทคนิคที่สร้างขึ้นด้วยหลักการ Machine learningนั้นจะถูกสอนด้วยระบบปัญญาประดิษฐ์เชิงแคบเพื่อเรียนรู้จากข้อมูลที่ดำเนินการตัวอย่างเช่น ระบบที่ระบุรูปภาพของลูกโป่งวันเกิดเครื่องจักรอาจถูกสอนให้ใช้ข้อมูลจากการทำงานประจำที่กำหนดไว้ล่วงหน้าเช่น เครื่องหนึ่งเอาไว้ตรวจจับรูปร่างอีกเครื่องหนึ่งเอาไว้ระบุตัวเลขและอีกเครื่องหนึ่งเอาไว้เพื่อวิเคราะห์สีโดยรูปแบบMachine learning ที่กล่าวมาระบบจะดึงรหัสที่มนุษย์ใช้เข้าเป็นประจำมาใช้และพัฒนาเทคนิคเพื่อช่วยให้มันเรียนรู้การระบุวัตถุได้อย่างถูกต้องในขณะที่สิ่งนี้เป็นการพัฒนาระบบ AI ที่ค่อนข้างแหวกแนว แต่ข้อบกพร่องของแบบจำลองก็ปรากฏขึ้นอย่างรวดเร็วเช่นกันซึ่งปัญหาที่ใหญ่ที่สุดคือการใช้การวิเคราะห์แบบประจำที่ต้องกำหนดไว้ล่วงหน้านั้น ต้องการข้อมูลที่ป้อนจากมนุษย์มากเกินไปตลอดช่วงการใช้งาน นอกจากนี้ยังมีปัญหาหากมีรูปภาพที่ยากต่อการประมวลผล เช่น รูปใบหน้าหรือรูปวัตถุต่าง ๆ นั้นเบลอเนื่องจากแบบจำลองนี้ได้ดึงความเข้าใจของเราเกี่ยวกับสมองมนุษย์ทุกวันนี้เราจึงเรียกว่า Deep learning.คำว่า 'ลึก' หรือ ‘Deep’ หมายถึงการสร้างโครงข่ายประสาทเทียมแบบเลเยอร์คล้ายกับตาข่ายของเซลล์ประสาทที่เชื่อมต่อกันซึ่งอยู่ภายในสมอง แต่ไม่เหมือนกับสมองที่ทำหน้าที่เหมือนตาข่ายสาม มิติที่เซลล์ประสาทหนึ่งสามารถสื่อสารกับส่วนอื่น ๆ ภายในบริเวณใกล้เคียงได้เครือข่ายประดิษฐ์เหล่านี้มีโครงสร้างที่ทำเป็นชั้นโดยมีเลเยอร์บนเลเยอร์ (layer upon layer)ของเส้นทางเชื่อมต่อเพื่อให้ข้อมูลนั้นไหลได้เทคนิคนี้เรียกว่า Backpropagation เทคนิคที่ปรับน้ำหนักระหว่างโนดในเครือข่ายเพื่อให้แน่ใจว่าจุดข้อมูลที่เข้ามาจะนำไปสู่ผลลัพธ์ที่

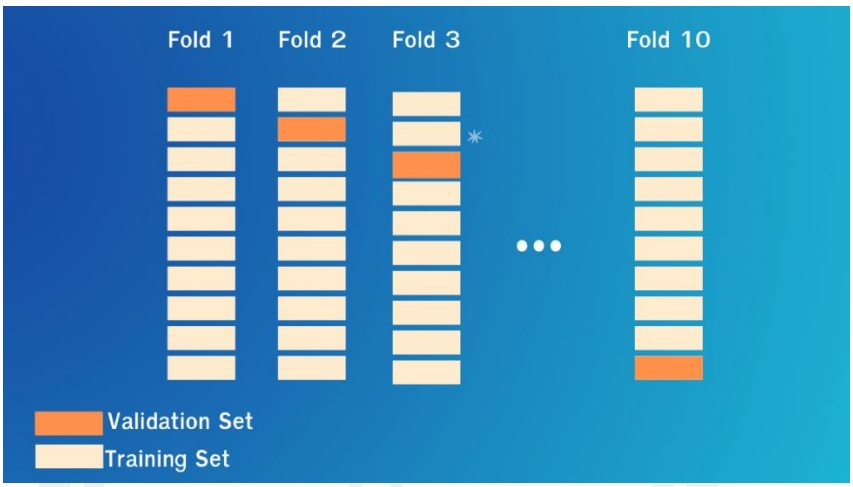
ถูกต้อง นักวิจัยต้องการสร้างกระบวนการวิเคราะห์ที่ซับซ้อนของสมองขึ้นมาใหม่แต่ละเลเยอร์จึงถูกออกแบบให้วิเคราะห์ข้อมูลและเพิ่มเติมข้อมูลประกอบสำหรับข้อมูลนั้น ๆ ทุกครั้งอีกด้วย เมื่อวัตถุผ่านแต่ละเลเยอร์ ความแม่นยำของภาพที่ ถูกต้องและความเข้าใจก็จะมีความเป็นไปได้มากขึ้นอย่าง ในตัวอย่างของลูกโป่งวันเกิดที่ได้กล่าวไปแล้วนั้น ภาพจะถูกแบ่งออกเป็นส่วนประกอบต่าง ๆ ของลูกโป่งวันเกิด ไม่ว่าจะเป็นสี หมายเลขหรือตัวอักษรใด ๆ บนพื้นผิว รูปร่าง หรือแม้กระทั่งแยกว่าว่าลูกโป่งถูกจับไว้หรือลอยอยู่บนอากาศ จากนั้นแต่ละส่วนจะถูกวิเคราะห์โดยเซลล์ประสาทเลเยอร์ที่หนึ่งเพื่อประมาณการและส่งผ่านข้อมูลไปยังเลเยอร์ถัดไประบบนี้ยังสามารถทำงานได้ดี โดยเฉพาะอย่างยิ่งถูกใช้ในการป้องกันการฉ้อโกง ยกตัวอย่างเช่น ระบบสามารถออกแบบมาเพื่อระบุกิจกรรมของบัญชีฉ้อโกงที่โยงกับเครือข่ายประสาทโดยใช้ข้อมูลดิบก่อนหน้า แล้วหลังจากนั้นมันจะเพิ่มข้อมูลเพิ่มเติมที่ไหลผ่านในระบบด้วย เช่น ราคาซื้อขายของที่นำเข้าและข้อมูลด้านตำแหน่งต่าง ๆ ในขณะที่บางเครือข่ายอาจมีเพียงไม่กี่เลเยอร์ แต่บางโปรแกรมอย่าง Alpha Go ของ Google นั้นมีหลายร้อยเลเยอร์ จนสามารถเอาชนะผู้เล่นระดับแชมป์ของเกมกระดานจีนได้ในปี2559 ได้ โดยธรรมชาติแล้ว สิ่งนี้ต้องการอานุภาพในการคำนวณที่ยิ่งใหญ่ อย่างไรก็ตาม โครงข่ายประสาทนั้นเป็นความใฝ่ฝันของผู้บุกเบิก AI ยุคแรกเสมอแม้ในปัจจุบันมันก็ยังไม่สามารถทำได้ ในปัจจุบัน ระบบMachine learning ที่ทันสมัยที่สุดหลายแห่งใช้เครือข่ายประสาทเทียมในการประมวลผลข้อมูล ความสำเร็จล่าสุดในอุตสาหกรรมรถยนต์ไร้คนขับเป็นไปได้เนื่องจากDeep learning ในขณะที่หลักการยังถูกนำไปใช้ในภาคการป้องกันและการบินเพื่อระบุวัตถุจากอวกาศอีกด้วยในขณะที่ศักยภาพของDeep learning นั้นมีมหาศาล แต่ก็มีข้อจำกัดเมื่อต้องทำงานที่คล้ายกับมนุษย์มากขึ้นเช่นกัน เพราะ Deep learning นั้นมีการจดจำรูปแบบอย่างกฎที่ซับซ้อนแต่ตายตัวของ Goแต่นักวิจัยชี้ให้เห็นว่า Training Data จำนวนมากจำเป็นต้องสอนเครื่องจักรแค่กฎที่เฉพาะเจาะจง การจดจำรูปแบบอาจเป็นตัวอย่างที่เด่นชัดที่สุดของ Aiในด้านการติดต่อ โดยมี Deep learningทำหน้าที่เป็นเครือข่ายสนับสนุนอินพุตแบบ Multimodal ที่รวมความสามารถด้านเสียงและการรับรู้จะถูกประมวลผลควบคู่ไปกับเอาต์พุตแบบMultimodal เช่น รูปภาพและเสียงสังเคราะห์บริษัทอย่างStarbucks ไปจนถึง Apple กำลังปรับใช้ระบบข่าวกรองนี้ ทำให้ลูกค้ามีทางเลือกในการสั่งซื้อผ่านแอพพลิเคชั่นของพวกเขาผ่านทางคำสั่งเสียงและความสะดวกในการเข้าสู่อุปกรณ์ด้วยสายตาเพียงอย่างเดียวในการขั้นของการพัฒนาในปัจจุบัน ดูเหมือนจะเป็นไปไม่ได้ที่ Deep learning จะดำเนินการกระบวนการที่ซับซ้อนและมีกระบวนการคิดแบบปรับได้อย่างมนุษย์ อย่างไรก็ตามเทคโนโลยียังคงพัฒนาอย่างต่อเนื่อง Deep learningอาจไม่ส่งผลสำหรับหุ่นยนต์สังหารในช่วงเวลาอันใกล้ นี้แต่นั่นไม่ได้หมายความว่ามันจะไม่เปลี่ยนแปลงแง่มุมพื้นฐานของสังคมในรูปแบบอื่นกลุ่มวิจัย Google Brainแสดงให้เห็นถึง วิธีที่ AI นั้นเรียนรู้อย่างลึกซึ้งได้อย่างไร โดยไม่ต้องระบุวงจำกัดของการทดลองสำหรับการระบุแมวแต่ละตัว หลังจากนั้น ข้อมูลของแมวหลายล้านตัวถูกส่งต่อไปยัง Google Brain และเครือข่ายสามารถระบุ

รูปภาพได้โดยไม่ต้องใช้ข้อมูลกำกับการระบุแมวแต่ละตัวอาจดูเหมือนเป็นแค่การทดสองพื้นฐาน แต่นั่นก็ทำให้เห็นว่าการพัฒนาดังกล่าวสามารถนำไปใช้ในทางปฏิบัติได้มากขึ้นผลการศึกษาเกี่ยวกับ AI ในวงการแพทย์ โดยมหาวิทยาลัยเบอร์มิงแฮม พบว่าDeep learning นั้นมีความเชี่ยวชาญในการตีความภาพทางการแพทย์เทียบเท่าความเชี่ยวชาญของมนุษย์เลย นี่อาจเป็นการปูทางให้ AI ได้มีบทบาทในวงการแพทย์ที่กำลังจะก้าวไปข้างหน้ามากขึ้น เพื่อลดความเครียดของเหล่าบุคลากรทางการแพทย์และช่วยให้แพทย์ใช้เวลากับผู้ป่วยได้เยอะมากขึ้นสิ่งน่าตื่นเต้นที่สุดของ Deep learning คือมันกำลังถูกขนานนามว่าเป็นดั่งกระดานกระโดดน้ำที่ถูกค้นพบในจักรวาล โดยนักวิจัยจากมหาวิทยาลัย ETH ซูริค ได้เปิดเผยงานวิจัยที่พวกเขาได้ลองใช้เครือข่ายประสาทเทียมเพื่อศึกษาสสารมืดโดยเปรียบเทียบกับกล้องโทรทรรศน์ฮับเบิล พวกเขาพบว่า Deep learning บ่งบอกค่าได้แม่นยำกว่า 30% เมื่อแยกองค์ประกอบของจักรวาลสารแบริออนสารมืด และพลังงานมืด นักวิจัยได้สรุปว่า Deep learning สร้างโอกาสที่ดีสำหรับการวิเคราะห์ข้อมูลทางดาราศาสตร์ในอนาคตและสิ่งที่จะเป็นไปอย่างแน่นอนคือการระดมทุนเข้าสู่ AI - Pentagon has เพื่อจัดสรรงบประมาณเกือบ 1 พันล้านดอลลาร์ให้ AI โดยเน้นการศึกษาวิจัยDeep learning ในปี 2563เพื่อที่อิทธิพลของ AI และ Deep จะได้ เติบโตขึ้นไปในอนาคต

2.1.5 การวัดประสิทธิภาพของแบบจำลอง

2.1.5.1 การวัดประสิทธิภาพของแบบจำลองด้วย 10-fold cross validation

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



ภาพที่ 2. 4 ตัวอย่างการทดสอบประสิทธิภาพแบบ 10- fold cross validation

ดังภาพที่ 2.4 จะเป็นการทดสอบประสิทธิภาพแบบ 10-fold ซึ่งจะทำการแบ่งชุดข้อมูลออกเป็น 10 ชุด โดยในแต่ละรอบจะใช้ชุดข้อมูลเพื่อเป็นชุดข้อมูลทดสอบ 1 ชุด และให้ชุดข้อมูลอื่น ๆ เป็นข้อมูลชุดสอน โดยจะทำการทดสอบทั้งหมด 10 รอบ จากนั้นนำ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-Measure)

2.1.5.2 การวิเคราะห์ประสิทธิภาพ

การประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ซึ่งควรนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย วัดประสิทธิภาพของแบบจำลองโดยใช้เทคนิคการวัดประสิทธิภาพแบบ 10-fold cross validation โดยการแบบข้อมูลออกเป็น 10 กลุ่ม เท่า ๆ กัน โดยในแต่ละรอบการทดสอบจะใช้ข้อมูล 1 ชุด เป็นชุดทดสอบ และใช้ชุดที่เหลือเป็นชุดฝึกสอน โดยค่าที่ใช้ในการวัดประสิทธิภาพของแบบจำลองคือค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure) ดังสมการที่ 2.13 2.14 2.15

$$\text{Precision} = \frac{TP}{TP+TP} \quad (2.13)$$

$$\text{Recall} = \frac{TP}{TP+TF} \quad (2.14)$$

$$\text{F} = \frac{2\ X\ Precision\ X\ Recall}{Precision + Recall} \quad (2.15)$$

โดย

TP คือ จำนวนข้อมูลความคิดเห็นเชิงบวกที่จำแนกได้ถูกต้อง

TN คือ จำนวนข้อมูลความคิดเห็นเชิงลบบวกที่ทำนายได้ถูกต้อง

FP คือ จำนวนข้อมูลความคิดเห็นเชิงบวกที่จำแนกผิด

FN คือ จำนวนข้อมูลความคิดเห็นเชิงลบที่จำแนกผิด

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องในงานวิจัยนี้แบ่งเป็น 3 ส่วนคือ เหมืองข้อความที่เกี่ยวกับโควิด 19 การคัดเลือกคุณลักษณะ การจำแนกข้อความ(Classification) มีดังนี้

2.2.1 เหมืองข้อมูลสำหรับโควิด

ก็ได้มีนักวิจัยหลายท่านได้ทำเหมืองข้อความที่เกี่ยวกับโควิด 19 มาใช้ในการวิเคราะห์ข้อความ Sciandra [16] มุ่งเน้นไปที่การสื่อสารผ่านโซเชียลมีเดียของอิตาลีเกี่ยวกับ COVID-19 การวิเคราะห์การสนทนาที่เกิดขึ้นบน Twitter มีการตรวจสอบเพิ่มเติมและวิธีการที่แตกต่างกันในขณะที่การทดลอง พบว่าผลของการฝังคำต่อการจำแนกประเภทนั้นคล้ายคลึงกับความถี่ของคำสมมติฐานแรกคือเรารวบรวมชุดข้อมูลขนาดใหญ่ที่มีข้อความที่สามารถจับความแตกต่างของภาษาระหว่างก่อนและหลังประกาศจากอิตาลี การวิเคราะห์ชี้ให้เห็นว่าข้อมูลที่จัดทำโดยข้อมูลอาจเปิดเผยคำศัพท์ที่ใช้แตกต่างกันในช่วงเวลาต่าง ๆ ซึ่งจะทำให้เกิดการเปลี่ยนแปลงในการสนทนาออนไลน์อันเป็นผลมาจากมาตรการของรัฐบาลที่มีผลกระทบอย่างมากต่อชีวิตของผู้คน สมมติฐานที่สองเกี่ยวข้องกับการเปรียบเทียบผลของการวิเคราะห์ความรู้สึกโดยใช้ศัพท์สองคำสำหรับภาษาอิตาลี ในกรณีนี้การวิเคราะห์แสดงให้เห็นถึงความไม่ลงรอยกันและความจำเป็นในการจัดประเภทที่เชื่อถือได้ เพื่อให้สามารถประมาณการขั้วของข้อความที่ยอมรับได้ ในที่สุดเทคนิคการฝังคำแสดงให้เห็นในระดับเชิงพรรณนาถึงความสามารถในการจับความคล้ายคลึงทางความหมายและบริบทในขณะที่แบบจำลองการคาดคะเนตามการฝังคำไม่สามารถปรับปรุงความแม่นยำของการจำแนกตัวแปรบางตัวได้อย่างมีนัยสำคัญ

Sharma และคณะ [17] การศึกษาให้ข้อมูลเชิงลึกเกี่ยวกับความรู้สึกของผู้ใช้ทวีตในสองประเทศหลัก การระบาดของโรคได้เขย่าโลกและมีผู้ติดเชื้อและเสียชีวิตเพิ่มขึ้นทำให้ผู้คนมีความทุกข์ อย่างไรก็ตามการศึกษานี้รวบรวมข้อมูลทวีตเป็นเวลา 7 เดือนหลังจากที่ WHO ประกาศให้ไวรัสโคโรนาสายพันธุ์ใหม่กำลังระบาด อารมณ์ได้พัฒนาขึ้นเมื่อชาวทวีตในช่วงหลายเดือนนี้ได้รับการเสริมพลังด้วยข้อมูลที่ดีขึ้นเกี่ยวกับมาตรการป้องกัน การศึกษานำเสนอสถานะปัจจุบันของอารมณ์โดยใช้ตัวอย่างข้อมูลขนาดใหญ่ การวิเคราะห์ความรู้สึกช่วยรัฐบาลและองค์กรในการวัดอารมณ์สาธารณะในปัจจุบันและสามารถแนะนำพวกเขาในการตัดสินใจเชิงกลยุทธ์ที่เกี่ยวข้องกับธุรกิจและการกำกับดูแล อย่างไรก็ตามการศึกษานี้ไม่ได้ไร้ข้อ จำกัด การศึกษานี้รวบรวมเฉพาะข้อมูลของสองประเทศหลัก ได้แก่ สหรัฐอเมริกาและอินเดีย การวิเคราะห์เชิงสำรวจที่ละเอียดยิ่งขึ้นของทวีตแต่ละรายการจะให้ข้อมูล

เชิงลึกที่ดีขึ้น นอกจากนี้ยังแนะนำว่าการสร้างแบบจำลองเฉพาะที่พร้อมกับการวิเคราะห์ความรู้สึกสามารถทำได้เพื่อการศึกษาที่ครอบคลุมมากขึ้น

Jelodar และคณะ [18] อินเทอร์เน็ตและโซเชียลมีเดียสาธารณะเช่นฟอรัมการดูแลสุขภาพออนไลน์เป็นช่องทางที่สะดวกสำหรับผู้ใช้ (ผู้คน / ผู้ป่วย) ที่กังวลเกี่ยวกับปัญหาสุขภาพในการพูดคุยและแบ่งปันข้อมูลซึ่งกันและกัน ในช่วงปลายเดือนธันวาคม 2019 มีรายงานการระบาดของไวรัสโคโรนาสายพันธุ์ใหม่ (การติดเชื้อซึ่งส่งผลให้เกิดโรคชื่อ COVID-19) และเนื่องจากการแพร่กระจายของไวรัสอย่างรวดเร็วในส่วนอื่น ๆ ของโลกองค์การอนามัยโลกจึงประกาศ ภาวะฉุกเฉิน ในบทความนี้เราใช้การแยกการอภิปรายที่เกี่ยวข้องกับ COVID-19 โดยอัตโนมัติจากโซเชียลมีเดียและวิธีการใช้ภาษาธรรมชาติ (NLP) ตามการสร้างแบบจำลองหัวข้อเพื่อเปิดเผยประเด็นต่าง ๆ ที่เกี่ยวข้องกับ COVID-19 จากความคิดเห็นสาธารณะ นอกจากนี้เรายังตรวจสอบวิธีการใช้เครือข่ายประสาทเทียม LSTM สำหรับการจำแนกความคิดเห็นของ COVID-19 การค้นพบของเราชี้ให้เห็นถึงความสำคัญของการใช้ความคิดเห็นสาธารณะและเทคนิคการคำนวณที่เหมาะสมเพื่อทำความเข้าใจปัญหารอบ ๆ COVID-19 และเพื่อเป็นแนวทางในการตัดสินใจที่เกี่ยวข้อง นอกจากนี้การทดลองแสดงให้เห็นว่าแบบจำลองการวิจัยมีความแม่นยำ 81.15% ซึ่งเป็นความแม่นยำที่สูงกว่าอัลกอริทึมการเรียนรู้ของเครื่องอื่น ๆ ที่รู้จักกันดีหลายประการสำหรับ COVID-19-Sentiment Classification

Ito และคณะ [19] ในการศึกษานี้ได้รวบรวมทวีตที่เกี่ยวข้องกับ COVID-19 และไม่รวมรีทวีต จากนั้นทวีตที่เหลือจะถูกวิเคราะห์โดยการสร้างแบบจำลองหัวข้อการจัดกลุ่มด้วยการฝังประโยคที่แตกต่างกันประการแรกหัวข้อทวีตที่ LDA ได้รับการยืนยันอย่างชัดเจนเกี่ยวกับสิ่งที่เกิดขึ้นเกี่ยวกับ COVID-19 LDA เป็นเครื่องมือที่ดีในการดึงภาพรวมขององค์กรจากนั้นแต่ละประโยคจะถูกเปรียบเทียบโดยผลการจัดกลุ่ม k-mean เพื่อตรวจสอบว่าพื้นที่การฝังใดเหมาะสมที่สุดในการเปรียบเทียบทวีตในเชิงความหมาย ในการประเมินเชิงปริมาณ Sentence-BERT จะให้คะแนนค่า Pseudo F ที่ดีที่สุด อย่างไรก็ตามไม่ได้หมายความว่า Sentence-BERT จะดีกว่าสำหรับการสร้างกลุ่มหัวข้อ อันที่จริงผลการจัดกลุ่มด้วยการฝังประโยค Word 2 Vec ได้ดึง word clouds ที่คล้ายกันกับหัวข้อ LDA การวิเคราะห์คลัสเตอร์แบบไดนามิกยังไม่เพียงพอเนื่องจากดูเหมือนว่าไม่มีเครื่องมือที่เหมาะสมในการแสดงภาพรวมของคลัสเตอร์ประโยค ตัวอย่างเช่นจำเป็นต้องใช้เครื่องมือที่แยกประโยคที่เหมือนกันโดยทั่วไปน้อยที่สุดออกจากคลัสเตอร์ กระบวนการสร้างคลัสเตอร์แบบไดนามิกควรได้รับการพิจารณาใหม่เนื่องจากคลัสเตอร์ที่ถูกขัดจังหวะจะไม่ปรากฏขึ้นอีกเลยหลังจากนั้นนั่นหมายความว่าคลัสเตอร์สองคลัสเตอร์จะถือว่าเป็นคลัสเตอร์ที่แตกต่างกันหากคลัสเตอร์เดียวกันปรากฏขึ้นอีกครั้งหลังจากช่วงเวลาหนึ่ง การศึกษานี้แสดงให้เห็นว่าสามารถพัฒนาวิธีการที่มีประสิทธิภาพบางอย่างสำหรับการวิเคราะห์ข้อมูลทวีตเพื่อเปิดเผยการเปลี่ยนแปลงข้อมูลที่อยู่ภายใต้

ช่วงเวลาหนึ่ง ขณะนี้เรากำลังดำเนินการเพื่อปรับแต่งเพิ่มเติมเกี่ยวกับเทคนิคในการดึงข้อมูลการเปลี่ยนแปลงแบบไดนามิกจากการวิเคราะห์ข้อมูลทวีต

2.2.2 การคัดเลือกคุณลักษณะ

ซึ่งเป็นกระบวนการในการทำดัชนี มาใช้เพื่อปรับปรุงประสิทธิภาพการจำแนกความคิดเห็น Hasan และคณะ [4] ใช้แบบจำลอง Bag of Words และ Term Frequency-Inverse Document Frequency (TF-IDF) เพื่อวิเคราะห์ความรู้สึก ที่ใช้ร่วมกันเพื่อจำแนกทวีตเชิงบวกและเชิงลบอย่างแม่นยำ ของข้อมูลที่ประมวลผลล่วงหน้าโดยใช้ภาษาธรรมชาติ เปรียบเทียบวิธีที่เสนอกับเทคนิคSupport Vector Machine (SVM), เทคนิค MaxiMum Entropy, เทคนิคNaive Bayes และ เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ทำให้ความแม่นยำของการวิเคราะห์ความรู้สึก การทดลองพบว่า ความแม่นยำ 85.25%

Guo และ Yang [5] วิเคราะห์น้ำหนักของคำข้อมูลจาก People news แบ่งเป็น 4 ประเภทใช้เทคนิค TFIDF แบบดั้งเดิมและเทคนิค TFIDF ที่ปรับปรุงแล้ว ผลลัพธ์แสดงให้เห็นว่าเทคนิค TFIDF ที่ปรับปรุงแล้วมีความแม่นยำสูงกว่าเทคนิค TFIDF แบบเดิม

Kadhim และคณะ [6] ได้ใช้เทคนิคที่แตกต่างกันคือ BM25 และ TF-IDF เพื่อแยกคำ ฟังก์ชัน BM25 ใช้เพื่อจัดลำดับชุดเอกสารที่ไม่มีข้อมูลที่เกี่ยวข้อง ในขณะที่ใช้ TF-IDF เพื่อถ่วงน้ำหนักคุณลักษณะตามความถี่ของแต่ละคำกล่าวคือคำศัพท์ในเอกสารเกี่ยวข้องกับคำอื่น เพื่อแยกคำหลักจากการรวบรวมข้อมูลที่ขึ้นอยู่กับ Twitter เป็นที่ชัดเจนว่าเทคนิค TF-IDF มีประสิทธิภาพดีกว่า BM25 ตามมาตรวัด F1

Nurhayati และคณะ [7] เลือกคุณสมบัติ Chi-Square เพื่อกำจัดคุณสมบัติ การศึกษานี้มีวัตถุประสงค์เพื่อตรวจสอบผลของการเลือกคุณสมบัติ Chi-Square ที่มีต่อประสิทธิภาพของอัลกอริทึมNaïve Bayes ในการวิเคราะห์เอกสารความรู้สึก ข้อมูลถูกนำมาจากข้อมูลการฝึกอบรม Corpus v1.0 Indonesia Movie Review ผลจากการวิเคราะห์ความเชื่อมั่นโดยไม่เลือกคุณลักษณะได้รับความแม่นยำ 73.33% ความแม่นยำ 100.00% การเรียกคืน 65.21% ในขณะที่การเลือกคุณสมบัติ Chi-Square (ระดับนัยสำคัญที่ 0.1) ได้ผลลัพธ์ความแม่นยำ 93.33% และการเรียกคืน 93.33% จากผลลัพธ์จะเห็นได้ว่าการเลือกคุณสมบัติ Chi-Square มีผลต่อประสิทธิภาพอัลกอริทึมของ Naïve Bayes ในการวิเคราะห์เอกสารความรู้สึก

Zhai และคณะ [20] น้ำหนักและการประเมินประสิทธิภาพการจัดหมวดหมู่ ในหมู่พวกเขาการเลือกคุณสมบัติเป็นขั้นตอนสำคัญในการจัดประเภทข้อความซึ่งส่งผลต่อความแม่นยำในการจัดหมวดหมู่ การเลือกคุณสมบัติสามารถช่วยบ่งชี้ความเกี่ยวข้องของเนื้อหาข้อความและสามารถจัดประเภทข้อความได้ดีขึ้น ในขณะเดียวกันการเลือกคุณสมบัติก็มีอิทธิพลอย่างมากต่อผลการจัดหมวดหมู่ การจัดหมวดหมู่ข้อความเป็นโมดูลที่สำคัญมากในการประมวลผลข้อความและมีการ

นำไปใช้กันอย่างแพร่หลายในด้านต่าง ๆ เช่นการกรองสแปมการจัดประเภทข่าวสารการจัดประเภทความรู้สึกและการติดแท็กบางส่วน บทความนี้เสนอวิธีการแยกคำที่มีคุณลักษณะตามสถิติ Chi-Square เนื่องจากคุณลักษณะคำที่ปรากฏร่วมกันหรือแยกกันอาจแตกต่างกันไปในสถานการณ์ต่าง ๆ เราจึงจัดประเภทข้อความโดยใช้คำเดี่ยวและคำคู่เป็นคุณลักษณะในเวลาเดียวกัน จากวิธีการของเรา เราทำการทดลองโดยใช้เทคนิคการจำแนกประเภท Naive Bayes และ Support Vector Machine ประสิทธิภาพของวิธีการของเราแสดงให้เห็นโดยการเปรียบเทียบและวิเคราะห์ผลการทดลอง

Haryanto และคณะ [21] ในการศึกษานี้เราใช้ SVM ในการจัดประเภทข้อความ มีการทำให้เป็นมาตรฐานหรือ lemmatization word ด้วยการเพิ่มการเลือก Chi-Square ในการจำแนกประเภทที่เราทำขึ้น นอกจากนี้ยังมีการดำเนินการข้อมูลก่อนการประมวลผล ได้แก่ การลบคำหยุดและโทเค็น เราใช้ชุดข้อมูล BBC ที่มีเอกสาร 2,225 รายการและ 5 หมวดหมู่ มี 21,813. คุณลักษณะที่เป็นผลมาจากการใช้การกำหนดคำและคุณลักษณะ 31,007 อันเป็นผลมาจากการใช้คำขยาย แต่ละคุณลักษณะแทนจำนวนคำที่ออกมาในเอกสาร เราใช้เมทริกซ์ ความสับสนเพื่อประเมินผลลัพธ์ของการจัดกลุ่มข้อความ ประสิทธิภาพการจัดหมวดหมู่ข้อความ SVM โดยใช้ Stemming ที่ปรับปรุงโดย Chi-Square (วิธีที่ 1) ได้ผลลัพธ์ที่ดีกว่าการใช้ lemmatization ที่ปรับปรุงโดย Chi-Square (วิธีที่ 2) ประสิทธิภาพที่ดีที่สุดได้มาจากการลดคุณสมบัติ 80% โดยที่วิธีที่ 1 ได้รับค่าความแม่นยำ 95% ค่าการเรียกคืน 95% และค่าความแม่นยำ 95.05% วิธีที่ 2 รับเฉพาะค่าความแม่นยำ 93% ค่าการเรียกคืน 93% และค่าความแม่นยำ 93.24% โดยใช้การลดคุณสมบัติจำนวนเท่ากัน

Li [22] การเลือกคุณลักษณะที่มีอยู่ส่วนใหญ่เช่น Chi-Square และการได้รับข้อมูลไม่ได้พิจารณาว่าคุณลักษณะมีความสำคัญเพียงใดในเอกสาร คุณสมบัติไม่ว่าจะมีข้อมูลทางความหมายที่มั่นคงหรือไม่ก็ตามจะได้รับการปฏิบัติอย่างเท่าเทียมกัน โดยสังหรณ์ใจแล้วเมตริกการเลือกคุณสมบัติที่คำนวณจึงมีแนวโน้มที่จะทำให้เกิดเสียงรบกวน ดังนั้นในการศึกษานี้เราจึงขยายผลงานของ Li et al [1] เกี่ยวกับเมตริกความถี่ของเอกสารเสนอกลยุทธ์การเลือกคุณลักษณะแบบถ่วงน้ำหนักที่มีความสำคัญโดยทั่วไปสำหรับการจัดประเภทข้อความซึ่งค่าความสำคัญของคุณลักษณะในเอกสารได้มาจากความถี่สัมพัทธ์ในเอกสารนั้น การทดลองเกี่ยวกับชุดข้อมูลที่เปิดเผยต่อสาธารณะสามชุดพร้อมด้วยเมตริกยอดนิยมสองรายการแสดงให้เห็นว่ากลยุทธ์การเลือกคุณลักษณะแบบถ่วงน้ำหนักความสำคัญที่เสนอช่วยให้เมตริก Chi-square และ Information Gain แบบดั้งเดิมได้ผลลัพธ์ที่ดีขึ้นโดยเฉพาะเมื่อใช้คุณลักษณะน้อยลงในชุดข้อมูลที่ไม่สมดุล

Meng และคณะ [23] ในการจัดประเภทข้อความหลายคลาสประสิทธิภาพของตัวแยกประเภทมักจะต่ำมากในขณะที่การกระจายข้อมูลไม่สม่ำเสมอมาก สาเหตุอาจเป็นเพราะคลาสส่วนใหญ่มีอิทธิพลอย่างมากต่อลักษณนามและชนชั้นกลุ่มน้อยไม่สนใจ การศึกษานี้พยายามแก้ไขปัญหานี้โดยการปรับปรุงอัลกอริทึมการเลือกคุณลักษณะ CHI-square และโดยการปรับปรุงอัลกอริทึม

น้ำหนักความถี่ - ผกผันของเอกสารระยะ (TF-IDF) และรวมกับวิธีการจัดกลุ่ม K-mean สร้างกลุ่มตัวอย่างของชนกลุ่มน้อยเพื่อลดอิทธิพล ของการกระจายคลาสแบบเอียง ผลการทดลองแสดงให้เห็นว่าวิธีการที่ได้รับการปรับปรุงมีการปรับปรุงอย่างมีนัยสำคัญต่อประสิทธิภาพการจำแนกบนชุดข้อมูลอคติ

Setiyaningrum และคณะ [24] ใช้อัลกอริทึม Chi-Square และ KNN เป็นวิธีการ วิธีไคสแควร์เป็นอัลกอริทึมสำหรับการเลือกคุณสมบัติและสำหรับกระบวนการจำแนกผู้เขียนใช้อัลกอริทึม KNN การศึกษานี้สรุปได้ว่าอัลกอริทึม KNN ที่มีการเลือกคุณลักษณะ (ไคสแควร์) ได้ผลลัพธ์ของความแม่นยำที่เห็นได้จากค่าของระยะเพื่อนบ้านที่ใกล้ที่สุด K = 3 ได้ผลลัพธ์เดียวกันโดยมีค่าความแม่นยำ 65.00% ซึ่งระบุว่าทั้งสองวิธี ด้วยการเลือกคุณลักษณะและไม่มีการเลือกคุณลักษณะในชุดข้อมูลปรากฏว่าวิธีการที่มีการเลือกคุณลักษณะมีข้อดีคือมีประสิทธิภาพในกระบวนการคำนวณมากกว่า

ธีรยุทธ และจารี ทองคำ[25] ได้เสนอ กระบวนการคัดเลือกคุณลักษณะสำหรับเพิ่มประสิทธิภาพในการจำแนกความคิดเห็นของลูกค้า เกี่ยวกับร้านอาหารข้อมูลรวบรวมจากเว็บไซต์ wongnai.com จำนวน 4,487 ข้อความ ได้ใช้เทคนิค 3 ประการในการเลือกคุณสมบัติข้อความ ได้แก่ Chi-Square, Information Gain และ Information Gain Ratio เพื่อวัดประสิทธิภาพของเทคนิคการเลือกคุณสมบัติและใช้ Naive Bayes, Support Vector Machine, K-Nearest Neighbor และ C4.5 สำหรับการจัดประเภทไอออนบวก นอกจากนี้ยังมีการใช้ Cross Validation 10 เท่าเพื่อแบ่งข้อมูลออกเป็นชุดการเรียนรู้และวัดความความถูกต้อง (Accuracy) ความแม่นยำ (Precision) และการเรียกคืน (Recall) จากการทดลองพบว่าเทคนิคการเลือกฟีเจอร์ Information Gain ร่วมมือกับเทคนิค Naive Bayes ให้ผลลัพธ์ที่ดีที่สุดในการจำแนกประเภทของความคิดเห็นโดยความถูกต้องคือ 89.08% ความแม่นยำเท่ากับ 89.12% และการเรียกคืนเท่ากับ 89.10%

ดังนั้นสรุปการคัดเลือกคุณลักษณะที่ดีที่สุด คือ Term Frequency-Inverse Document Frequency (TF-IDF), BM25, Chi-Square และนำไปเป็นเทคนิค ในงานวิจัย

2.2.3 การจำแนกข้อความ (Classification)

ก็ได้มีนักวิจัยหลายท่านได้นำเอาเทคนิคการทำเหมืองความคิดเห็น มาใช้ในการจำแนกข้อความ เช่น Supianto และคณะ [8] ใช้เทคนิคการจำแนกต้นไม้แบบสุ่ม REP Tree และ C4.5 ความแม่นยำเฉลี่ยที่สูงขึ้นใช้เทคนิค C4.5 มากที่สุด มีความแม่นยำ 77.01% REP Tree และ Random Tree ตามลำดับ

Chen และคณะ[9] ใช้ Deep Neural Networks เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) และเทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) พบว่า Deep Neural Networks ที่มีประสิทธิภาพที่เหนือกว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) และเทคนิคเพอร์เซปตรอนแบบหลายชั้น (MLP) ในการการเข้าชมแพลตฟอร์มโซเชียวมีเดีย Sina Weibo

Suksangaram และคณะ [10] ใช้เทคนิค: C4.5, Naïve Bayes และ SVM ในการทำนายการเรียนรู้ขององค์กรที่มีผลต่อการปฏิบัติงานของพนักงานธนาคารเพื่อการเกษตรและสหกรณ์การเกษตรในภาคตะวันตก ผลการวิจัยพบว่าเทคนิค SVM มีค่าความแม่นยำ 98.33% ความแม่นยำ 0.025 และการเรียกคืน 0.984 ที่มากกว่าเทคนิค C4.5 และ Naïve Bayes

Saengthongpattana และคณะ [26] คุณภาพของบทความ Wikipedia ยังคงเป็นประเด็นหลักในทุกภาษา Wikipedia ส่วนใหญ่อาศัยบรรณาธิการและผู้ดูแลระบบเพื่อให้คุณภาพของเนื้อหา แต่ขนาดของเนื้อหา Wikipedia ทำให้การค้นหาทุกอินสแตนซ์ ของบทความใช้เวลานานมาก ดังนั้นเราจึงต้องการการตรวจจับคุณภาพอัตโนมัติที่ช่วยให้ผู้ใช้ประเมินคุณภาพของบทความได้ ในบทความนี้เราขอเสนอชุดคุณลักษณะเพื่อใช้สำหรับบทความ Wikipedia ภาษาอาเซียน เราตรวจสอบคุณลักษณะทางสถิติ การทดลองดำเนินการโดยใช้เทคนิค Naïve Bayes และ Decision tree เราพบว่าความแม่นยำของ Decision tree (96.34%) ดีกว่าNaïve Bayes (86.47%) นอกจากนี้เราพบว่าคุณลักษณะทางสถิติมีบทบาทสำคัญในการจำแนกคุณภาพของบทความวิกิพีเดีย ภาษาเวียดนามชาวอินโดนีเซียมาเลเซียไทยและตากาล็อกฟิลิปปินส์

Suksangaram และคณะ [10] การศึกษานี้มีวัตถุประสงค์เพื่อศึกษาและเปรียบเทียบแบบจำลองที่เหมาะสม การคาดการณ์ปัจจัยองค์กรที่มีผลต่อความสามารถในการทำงานของพนักงานธนาคารเพื่อการเกษตรและสหกรณ์การเกษตรในภาคตะวันตก แบบจำลองนี้ใช้เพื่อเปรียบเทียบ 3 เทคนิค: C4.5, Naïve Bayes และ SVM ผลการวิจัยพบว่าเทคนิค SVM เหมาะสมที่สุดตามด้วย C4.5 และ Naïve Bayes ในการทำนายปัจจัยการเรียนรู้ขององค์กรที่มีผลต่อการปฏิบัติงานของพนักงานธนาคารเพื่อการเกษตรและสหกรณ์การเกษตรในภาคตะวันตก โดยการวัดประสิทธิภาพของแบบจำลองด้วยความแม่นยำ 98.33% ความแม่นยำ 0.025 และการเรียกคืน 0.984

Zhang และ Rao [27] ด้วยการพัฒนาและความเป็นที่นิยมของ Internet of Things (IoT) ทำให้ IoT สร้างข้อมูลจำนวนมากและการจัดหมวดหมู่ ในระบบกริดไฟฟ้า อย่างไรก็ตามความแม่นยำของอัลกอริทึมเหล่านี้ไม่เป็นที่น่าพอใจ ในเอกสารของเราจะใช้แบบจำลองการเรียนรู้เชิงลึกที่มีประสิทธิภาพผลการทดลอง ได้เปรียบเทียบวิธีการจำแนกข้อความกริดที่มีอยู่นั่นคือ deep neural networks , Bayes, SVM ตามลำดับ

Zheng [28] บทความนี้เราจึงเสนอให้ใช้โค้ดเพื่อให้บรรลุฟังก์ชันแทนที่จะใช้ไลบรารีของบุคคลที่สาม เราประเมินโค้ดของเราในรูปแบบการจำแนกประเภทต่าง ๆ และผลการทดสอบแสดงให้เห็นว่าโค้ดของเรามีความเป็นไปได้ การจำแนกประเภทข้อความที่เราใช้ ได้แก่ Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR) และ Logistic Regression CV (LRCV) และ Naïve Bayes (NB) ดีที่สุด

Sawanglok และคณะ [29] งานนี้นำเสนอวิธีการพัฒนาการจดจำกิจกรรมโดยใช้ Kinect เป็นอุปกรณ์ตรวจจับการเคลื่อนไหวและการเรียนรู้ภายใต้การดูแลสำหรับการจำแนกประเภทจากนั้นข้อมูลจะได้รับการเพื่อเปรียบเทียบการเรียนรู้ภายใต้การดูแลสำหรับการจำแนกประเภทในงานอัลกอริทึม 4 แบบ ได้แก่ neural networks, naive bayes, decision tree and support vector machine ถูกนำไปใช้เพื่อสร้างแบบจำลองการจำแนกประเภท ผลการทดลองรูปแบบการจำแนกโดยรวมที่ดีที่สุดคือ neural networks ที่มีความแม่นยำประมาณ 75% ในขณะที่ดีที่สุดอันดับสองคือ support vector machine

Nassif และคณะ [30] โรคหลอดเลือดหัวใจตีบ เป็นหนึ่งในสาเหตุการเสียชีวิตอันดับต้น ๆ ของโลก ดังนั้นจึงเป็นเรื่องสำคัญมากที่จะต้องวินิจฉัยผู้ป่วยที่เป็นโรคนี้อย่างถูกต้อง คุณลักษณะที่เลือกใช้เพื่อฝึกตัวแยกประเภทที่แตกต่างกันสามประเภท ได้แก่ SVM, Naïve Bayes และ KNN Naïve Bayes ทำงานได้ดีที่สุดกับชุดข้อมูลและคุณสมบัตินี้มีประสิทธิภาพดีกว่าหรือตรงกับ SVM และ KNN ในพารามิเตอร์การประเมินทั้งสี่ที่ใช้และมีความแม่นยำ 84%

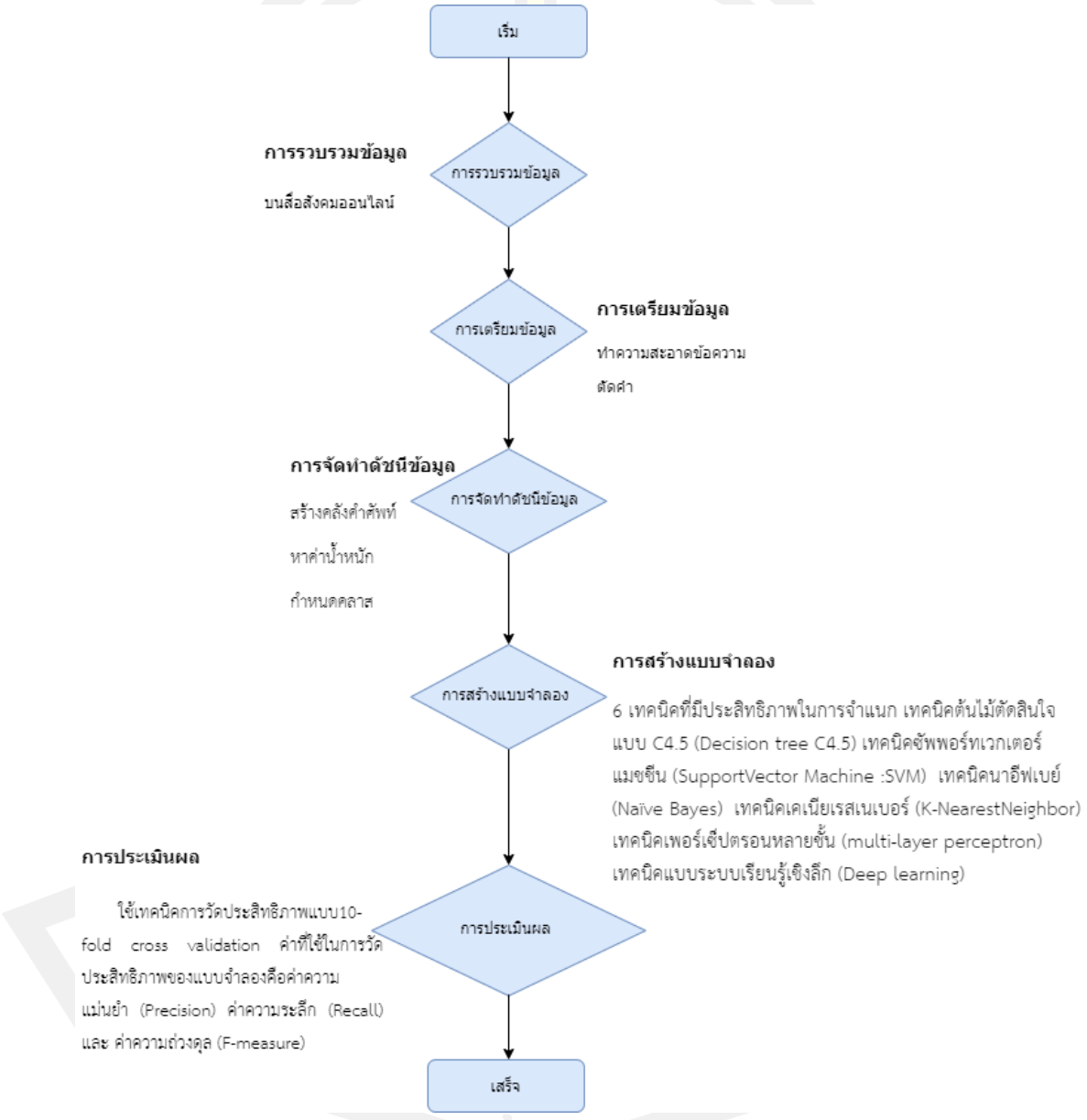
สถิตโชคและคณะ[2]ได้เสนอการจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะโดยใช้วิธีการสกัดข้อความและเทคนิคการเรียนรู้ของเครื่อง เพื่อจำแนกประเภทข้อร้องเรียนแบบอัตโนมัติ ที่เกี่ยวกับพฤติกรรมการขับขี่และคณภาพการการบริการ จำนวน 4 คลาส ผลการทดลองพบว่าตัวแบบโครงข่ายประสาทเทียมสามารถจำแนกข้อความพฤติกรรมการขับรถโดยสารสาธารณะได้ค่าความแม่นตรงสูงสุดซึ่งมีค่าเท่ากับ 90.23% ค่าความแม่นยำ เท่ากับ 91.9% ค่าความระลึกเท่ากับ 90.2% และค่าเอฟเมเชอร์เท่ากับ90.5% งานวิจัยนี้สามารถนำ ไปประยุกต์ใช้สร้างระบบจำแนกหมวดหมู่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะและระบบจำแนกหมวดหมู่เอกสารอัตโนมัติได้

2.2.4 สรุปงานวิจัยที่เกี่ยวข้อง

เทคนิคการจำแนกข้อความ คือ Decision tree C 4.5 Naïve Bayes SVM K-Nearest Neighbor: KNN Multi-layer Perceptron Deep Learning แล้วนำไปเป็นเทคนิค ในงานวิจัย

บทที่ 3 วิธีดำเนินงานวิจัย

ในการดำเนินการวิจัยเพื่อให้ได้ตามวัตถุประสงค์ที่ตั้งไว้ การวิเคราะห์ข้อมูลในงานวิจัยนี้ได้ใช้ขั้นตอน 5 ขั้นตอนรวมถึง การรวบรวมข้อมูล การเตรียมข้อมูล การจัดทำดัชนีข้อมูล การสร้างแบบจำลอง การประเมินผล ดังภาพที่ 3.1



ภาพที่ 3. 1 Flowchart การดำเนินการวิจัย

3.1 การรวบรวมข้อมูล

เก็บรวบรวมความรู้สึกของคนไทย ต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ เว็บไซต์ทวิตเตอร์ และพันทิป ตั้งแต่วันที่ 23 มกราคม 2563 ถึงวันที่ 1 พฤษภาคม 2563 (เริ่มต้นโรคโควิด 19 ในประเทศไทย) โดยคัดเลือกจาก Keyword COVID19, โควิด19, ไวรัสโคโรนาสายพันธุ์ใหม่

3.2 การเตรียมข้อมูล

นำข้อมูลความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ ด้วยเทคนิคเหมืองข้อความ โดยนำข้อมูลความคิดเห็นของผู้คนไทยที่ใช้สื่อสังคมออนไลน์ประมาณ 2,000 ความคิดเห็น ดึงข้อมูลด้วยการเขียนโปรแกรมโดยผู้วิจัย โดยใช้ภาษา Python ใช้เทคนิค Web Scraping โดยใช้ไลบลารีของ selenium, beautifulsoup4 โดยใช้ประโยชน์ ได้ตามความต้องการ ดังภาพที่ 3.2 เก็บข้อมูลอยู่ในรูปของไฟล์ประเภทแบบ Microsoft Excel ดังตารางที่ 3.1

```python
from selenium import webdriver
from selenium.webdriver.common.keys import Keys
from bs4 import BeautifulSoup as soup
import time

driver = webdriver.Chrome()

pixel = 0

def HashTag(keyword):

    global pixel

    url = 'https://pantip.com/search?q=' + keyword

    driver.get(url)

    time.sleep(3)

    # totalpage =  news // 20 # per scrolling is 20 news
    for i in range(20):
        driver.execute_script("window.scrollTo(0, {})".format(pixel))
        time.sleep(3)
        pixel = pixel + 10000
        #pixel += 10000

    page_html = driver.page_source

    data = soup(page_html, 'html.parser')

    pantiptext = data.findAll('a',{'class':'datasearch-in'})

    for i,tw in enumerate(pantiptext):
        print(i+1)
        print(tw.text)
        print('---------------')


HashTag('COVID19')
# HashTag('ไวรัสโคโรนาสายพันธุ์ใหม่')
# HashTag('โควิด19')
```

ภาพที่ 3. 2 ดึงข้อมูลโดยใช้ ภาษา Python

ตารางที่ 3. 1 เก็บข้อมูลอยู่ในรูปของไฟล์ประเภทแบบ Microsoft Excel

ข้อมูลอยู่ในรูปของไฟล์ประเภทแบบ Microsoft Excel	
1	กัปตันอเมริกานำทัพด่าหลังทรัมป์บอกประชาชนว่าไม่ต้องกลัวเพราะตัวเองติด
2	RSVและโรคทางเดินหายใจในช่วงที่มีการระบาดของใหม่ๆมีการป้องกันกัน
3	วัคซีนโควิด19การศึกษาระยะสุดท้ายแบบสมบูรณ์คงจะมีผลประกาศออกมาในเดือนหน้าเชื่อว่า
4	Facebookเตรียมบริจาคเงิน40ล้านดอลลาร์ฯช่วยธุรกิจคนผิวดำจากพิษFacebook
5	พม่ามีผู้ป่วยเพิ่ม1010รายสรุปผู้ป่วยทั้งหมด14383ราย

เป็นการทำให้เกิดความมั่นใจในคุณภาพของข้อมูลความคิดเห็นที่จะนำมาใช้วิเคราะห์ว่ามีความถูกต้องโดยการนำความคิดเห็นที่ไม่ถูกต้องออกหรือเป็นขั้นตอนที่อาจต้องแก้ไขก่อนนำไปใช้งาน ขั้นตอนการเตรียมข้อมูลประเภท text จะมีขั้นตอนดังนี้

3.2.1 ทำความสะอาดข้อความ

ทำความสะอาดข้อความ text cleaning เพราะโดยส่วนใหญ่ข้อมูล text ที่เราดึงมาจาก social จะเต็มไปด้วยสิ่งที่จะต้องตัดออกพอสมควร เช่น html tag ต่าง ๆ, url, หรือเครื่องหมายคำพูด (punctuation) บางครั้ง text ที่เราได้มามันจะมีความสกปรก เช่น มี html tag (กรณี scrape มาจาก web) หรือ มีเครื่องหมายต่าง ๆ ที่เราไม่ต้องการ (!€@%#*&~) ติดมาด้วยจึงจำเป็นต้องทำความสะอาดก่อน ดังตารางที่ 3.2

ตารางที่ 3. 2 ตัวอย่างทำความสะอาดข้อความ

	ก่อน ทำความสะอาดข้อความ	หลัง ทำความสะอาดข้อความ
1	กัปตันอเมริกานำทัพด่าหลังทรัมป์บอกประชาชนว่าไม่ต้องกลัวเพราะตัวเองติด	กัปตันอเมริกานำทัพด่าหลังทรัมป์บอกประชาชนว่าไม่ต้องกลัวเพราะตัวเองติด
2	RSVและโรคทางเดินหายใจในช่วงที่มีการระบาดของใหม่ๆมีการป้องกันกัน	โรคทางเดินหายใจในช่วงที่มีการระบาดของใหม่ๆมีการป้องกัน
3	วัคซีนโควิด19การศึกษาระยะสุดท้ายแบบสมบูรณ์คงจะมีผลประกาศออกมาในเดือนหน้าเชื่อว่า	วัคซีนโควิดการศึกษาระยะสุดท้ายแบบสมบูรณ์คงจะมีผลประกาศออกมาในเดือนหน้า
4	Facebookเตรียมบริจาคเงิน40ล้านดอลลาร์ฯช่วยธุรกิจคนผิวดำจากพิษFacebook	เตรียมบริจาคเงินล้านดอลลาร์ฯช่วยธุรกิจคนผิวดำ
5	พม่ามีผู้ป่วยเพิ่ม1010รายสรุปผู้ป่วยทั้งหมด14383ราย	พม่ามีผู้ป่วยเพิ่มสรุปผู้ป่วยทั้งหมด

3.2.2 ตัดคำ

การตัดคำ (Word Segmentation) งานวิจัยฉบับนี้มีขั้นตอนการการแบ่งตัวอักษรจากข้อความ (string) เพื่อหาขอบเขตของแต่ละหน่วยคำ (morpheme) เพื่อให้การจำแนกข้อความมีประสิทธิภาพ ด้วยโปรแกรม colab google ภาษา Python ใช้ไลบรารี (Library)ของ PyThaiNLP ผลที่ได้ ดังตารางที่ 3.3

ตารางที่ 3. 3 การตัดคำ

	การตัดคำ
1	กัปตันอเมริกา,นำทัพ,ด่า,หลัง,ทรัมป์,บอก,ประชาชน,ว่า,ไม่ต้องกลัว,เพราะตัวเองติด
2	โรคทางเดินหายใจ,ในช่วง,ที่มี,การระบาด,ของใหม่ๆ,มีการป้องกัน
3	วัคซีนโควิด,การศึกษา,ระยะสุดท้าย,แบบสมบูรณ์,คงจะมีผล,ประกาศ,ออกมา,ในเดือนหน้า
4	เตรียม,บริจาค,เงิน,ล้าน,ดอลลาร์,ฯ,ช่วย,ธุรกิจ,คนผิวดำ
5	พม่า,มี,ผู้ป่วย,เพิ่ม,สรุป,ผู้ป่วย,ทั้งหมด

3.2.3 กำจัดคำสะกดผิด

กำจัดคำสะกดผิด ว่าแต่ละ Token สะกดถูกหรือไม่ และแก้ไขให้ถูกต้อง Spelling Correction ก่อนที่จะนำไปใช้งานจริง เช่น การรับ Input จาก User สิ่งที่เราจะได้พบเจออยู่ตลอด คือ User กรอกข้อมูลผิด ยิ่งข้อมูลที่ไม่ใช่ข้อมูลที่มีโครงสร้าง Structure Data, ไม่มี Data Type ตัวเลข หรือวันเวลา ที่มีรูปแบบแน่นอน แต่เป็นข้อความ Free Text ที่มีคำสะกดผิดปนอยู่ ทำให้การ Validate ข้อมูล อาจจะทำได้ยาก

Spell checker คือ โปรแกรมตรวจการสะกด ตรวจคำ ผิด ว่าข้อความคำที่ User กรอกเข้ามามีปรากฏอยู่ใน Dictionary หรือไม่ โดยอาจจะแนะนำคำใกล้เคียง ที่น่าจะเป็นคำที่ถูกต้องให้ User เลือก หรือแม้กระทั่งเลือกให้โดยอัตโนมัติ เรียกว่า Spelling Correction

ในส่วนของภาษาไทย โดย Default แล้ว Spellchecker ของ PyThaiNLP จะใช้อัลกอริทึม ของ Peter Norvig ที่จะหารายการคำใกล้เคียงจาก Dictionary โดยใช้จำนวนอักษรที่ผิด 1, 2, ...,n ตัวอักษร ผสมกับความน่าจะเป็น จากความถี่ของคำนั้นที่ปรากฏใน Corpus ดัง Source Code ตัวอย่าง ภาษา Python ดังภาพที่ 3.3

```python
import re
from collections import Counter

def words(text): return re.findall(r'\w+', text.lower())

WORDS = Counter(words(open('big.txt').read()))

def P(word, N=sum(WORDS.values())):
    "Probability of `word`."
    return WORDS[word] / N

def correction(word):
    "Most probable spelling correction for word."
    return max(candidates(word), key=P)

def candidates(word):
    "Generate possible spelling corrections for word."
    return (known([word]) or known(edits1(word)) or known(edits2(word)) or [word])

def known(words):
    "The subset of `words` that appear in the dictionary of WORDS."
    return set(w for w in words if w in WORDS)

def edits1(word):
    "All edits that are one edit away from `word`."
    letters    = 'abcdefghijklmnopqrstuvwxyz'
    splits     = [(word[:i], word[i:])    for i in range(len(word) + 1)]
    deletes    = [L + R[1:]               for L, R in splits if R]
    transposes = [L + R[1] + R[0] + R[2:] for L, R in splits if len(R)>1]
    replaces   = [L + c + R[1:]           for L, R in splits if R for c in letters]
    inserts    = [L + c + R               for L, R in splits for c in letters]
    return set(deletes + transposes + replaces + inserts)

def edits2(word):
    "All edits that are two edits away from `word`."
    return (e2 for e1 in edits1(word) for e2 in edits1(e1))
```

ภาพที่ 3. 3 ความถี่ของคำนั้นที่ปรากฏใน Corpus

โดยอัลกอริทึม ของ Peter Norvig ไม่ได้ใช้ Context หรือคำแวดล้อมที่มาก่อนหน้าหรือต่อจากนั้น และไม่ได้ใช้ตำแหน่งปุ่มใกล้เคียง บน Keyboard มาคำนวนความน่าจะเป็น

3.2.4 การหาประเภทของคำ

การหาประเภทของคำ เป็นกระบวนการในการกำหนดชนิดของคำ (part of speech) กำกับหน้าที่ของคำ ประเภทของคำหรือชนิดของคำที่อยู่ในประโยค เช่น คำนาม (Noun) คำสรรพนาม (Pronoun) คำกริยา (Verb) คำวิเศษณ์ (Adverb) คำบุพบท (Preposition) คำสันธาน (Conjunctions) เนื่องจาก ประโยค วลี คำพูด ต่าง ๆ ที่เราใช้สื่อสารกันล้วนเกิดขึ้นจากการนำคำต่าง ๆ มาประกอบกันเป็นส่วนต่าง ๆ ที่ทำหน้าที่ต่างกันในประโยค เหมาะสำหรับนำไปแบ่งคำเพื่อหานิพจน์ระบุนาม (Named Entities) หรือชื่อเฉพาะต่าง ๆ ซึ่งพัฒนาโดย NECTEC ดังตารางที่ 3.4

คุณสมบัติ

สามารถทำงานได้กับทุกระบบปฏิบัติการ (Windows, Unix based, OSX)

รองรับการทำงานในรูปแบบเซอร์วิส (REST Full Service)

ประมวลผลได้อย่างรวดเร็ว

รองรับการจัดการคำที่เกิดขึ้นใหม่ คำในภาษาต่างประเทศ หรือคำแสลงใหม่

รองรับการจัดการคำที่ไม่อยู่ในพจนานุกรมอย่างชาญฉลาด
เรียนรู้จากคลังข้อมูลของ BEST2009 ขนาด 9 ล้านคำ
มีโมเดลการเรียนรู้ให้เลือกหลายขนาดตามความเหมาะสมของการใช้งาน

ตารางที่ 3. 4 ชนิดของคำ (part of speech)

#	ชนิดของคำ	คำอธิบาย	#	ชนิดของคำ	คำอธิบาย
1	ADV	คำกริยาคุณศัพท์	18	IJ	คำหรือน้ำเสียงแสดงอารมณ์
2	AUX	คำกริยาช่วย	19	JJA	คำขยายหน้านามและกริยา
3	CD	ตัวเลขและจำนวน	20	JJV	คำขยายหน้านามและกริยา
4	CL	คำลักษณนาม (Classifiers)	21	NEG	กำกับคำว่า "ไม่" และ "มิ"
5	CNJ	คำเชื่อม (Conjunction)	22	NN	คำนามทั่วไปหรือ สามานยนาม
6	COMP	Complementize	23	NR	ชื่อเฉพาะหรือวิสามานยนาม
7	DDEM	คำสรรพนาม นั่น, โน่น, นี่	24	OD	ตัวเลขและจำนวน
8	DINT	คำบ่งชี้ที่ใช้ตั้งคำถาม	25	P	คำบุพบท (Prepositions)
9	DPER	คำบ่งชี้ความเป็นเจ้าของ ของคำนามหน่วยหลัก	26	PAR	คำลงท้าย (Particles)
10	FXAJ	คำเสริม หน้าหรือคำอปุสรรค (น่า)	27	PDEM	คำสรรพนาม นั่น, โน่น, นี่ ที่อยู่ในตำแหน่งคำนามหลัก
11	FXAV	คำเสริม หน้าหรือคำอปุสรรค (อย่าง, แบบ, โดย)	28	PDT	คำบ่งชี้ที่มีลักษณะคล้ายการบอกจำนวน
12	FXG	คำเสริม หน้าหรือคำอปุสรรค (เหล่า, บรรดา, ชาว, พวก, นาน, ผอง)	29	PINT	กำกับคำถาม
13	FXN	คำเสริม หรือคำอปุ สรรค	30	PPER	บุรุษสรรพนาม
14	FWA	คำในภาษาต่างประเทศ	31	PU	เครื่องหมาย (Punctuations)
15	FWN	คำในภาษาต่างประเทศ	32	REFX	คำReflexive, คำReciprocals
16	FWV	คำในภาษาต่างประเทศ	33	VA	กริยาที่คล้ายคุณศัพท์
17	FWX	คำในภาษาต่างประเทศ	34	VV	คำกริยา
			35	X	อื่น ๆ

3.3 การจัดทำดัชนีข้อมูล

การจัดทำดัชนีข้อมูล เป็นวิธีหนึ่งโดยการนำข้อมูลการแยกประเภทของคำหรือชนิดของคำที่อยู่ในประโยค ดังตารางที่ 3.5

ตารางที่ 3. 5 ข้อมูลการแยกประเภทของคำหรือชนิดของคำ

คำ	ประเภทของคำ
เครียดมาก	Adverb
จุดด้อย	Adverb
เข้าใจกัน	Adverb
จุดด้อย	Verb
ด่างพร้อย	Verb
ภูมิภาค	Noun
สัญชาติ	Noun
ใกล้เคียง	Preposition
ไปไม่ได้	Verb

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

3.3.1 การสร้างคลังคำศัพท์ (Bag of Words)

เป็นการนำคำที่แยกประเภท ในกรณีที่ซ้ำกันนำมาเพียงหนึ่งคำ ส่วนกรณีคำที่ไม่ซ้ำกันนำมาทั้งหมด คงเหลือทั้งหมด 9,037 คำ ด้วยโปรแกรม colab google ภาษา Python ไลบรารี (Library)ของ Scikit-Learn CountVectorizer และ PyThaiNLP Part-of-Speech โดยการใช้พจนานุกรมไทยในการแยกชนิดของคำ จากนั้นทำการคัดเลือก เอาเฉพาะคำวิเศษณ์ โดยผู้วิจัยระบุคำความหมายเชิงบวก เชิงลบ แล้วให้ผู้เชี่ยวชาญด้านภาษาไทยให้ระบุคำความหมายเชิงบวก เชิงลบ อาจารย์ปุ่น ชมภูพระ อาจารย์ประจำสาขาวิชาภาษาไทย คณะครุศาสตร์ มหาวิทยาลัยราชภัฏชัยภูมิ [31] แล้วทำการคัดเลือกคำคุณลักษณะที่มีความหมายเชิงบวก เชิงลบแล้วแทนค่าของจำนวนคำคุณลักษณะที่ไม่ปรากฏเป็น 0 แทนค่าจำนวนคำคุณลักษณะเชิงบวกที่ปรากฏเป็น 1 และแทนค่าจำนวนคำคุณลักษณะเชิงลบที่ปรากฏเป็น -1 จะเห็นความถี่ของคำคุณลักษณะที่ถูกจำแนกออกมาเป็น เชิงบวก หรือเชิงลบ ในแต่ละความคิดเห็น และทำการแบ่งความคิดเห็นเชิงบวกและเชิงลบ โดยใช้หลักเกณฑ์ของภาษาไทยดังตารางที่ 3.6

ตารางที่ 3. 6 ตัวอย่างคำคุณลักษณะประเภทของคำ

คำ	ประเภทของคำ	คำคุณลักษณะ
เครียดมาก	Adverb	-1
จุดด้อย	Adverb	-1
เข้าใจกัน	Adverb	1
เหมือนกัน	Adverb	1
พุ่งขึ้น	Adverb	-1
เพราะมีระบาด	Adverb	-1
ล่าช้า	Adverb	-1
ถี่ถ้วน	Adverb	-1
อบอุ่น	Adverb	-1
มาตรการป้องกัน	Adverb	1
ยังไม่คลี่คลาย	Adverb	-1
นานวัน	Adverb	-1
กลับช้าสุด	Adverb	-1
ยากจน	Adverb	-1
ในวิกฤตนี้	Adverb	-1

จากวัตถุประสงค์งานวิจัยนี้ต้องการศึกษาชนิดของคำที่ส่งผลต่อประสิทธิภาพการจำแนก ผู้วิจัยจึงได้ทำการคัดเลือกเอา ชุดข้อมูลคำวิเศษณ์ คงเหลือคำที่เป็นคุณลักษณะเชิงบวกและเชิงลบ ที่สามารถนำไปใช้ในงานวิจัยนี้ จำนวน 236 คำ โดยแบ่งคำออก เป็น 2 กลุ่ม คือคำแทนคุณลักษณะเชิงบวก 76 คำ และคำแทนคุณลักษณะเชิงลบ 160 คำ

3.3.2 กำหนดคลาส

การกำหนดคลาส งานวิจัยฉบับนี้ได้สร้างตัวแทนเอกสารนั้นจะใช้วิธีในการนำคำบ่งชี้คุณลักษณะในชุดข้อมูลมาเรียงกันเพื่อทำการนับความถี่ของการเกิดขึ้นของคำนั้น ๆ จากนั้นจึงนำค่าจำนวนความถี่ของคำมาสร้างเวกเตอร์ตัวแทนเอกสาร และคำบ่งชี้ที่ไม่ปรากฏในเอกสารจะมีค่าเป็น 0 จากนั้นทำการนับจำนวนคำในแต่ละคุณลักษณะ โดยใช้การนับจำนวนความถี่ของค่าคุณลักษณะว่ามีจำนวนเท่าใด เพื่อนำมาเปรียบเทียบกัน เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงบวกมากกว่าความถี่ของคุณลักษณะเชิงลบให้ตัวแปรตามเป็น ความคิดเห็นเชิงบวก แทนด้วย P เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงลบมากกว่าความถี่ของคุณลักษณะเชิงบวกให้ตัวแปร

ตาม เป็นความคิดเห็นเชิงลบ แทนด้วย N แต่ถ้าหากความถี่ของคุณลักษณะเชิงบวกและเชิงลบเท่ากัน จะแทน ด้วย NE และข้อความจะถูกตัดออก จึงทำการลบแถวตั้งที่ข้อความ P และ N ออก จึงนำข้อมูลดังกล่าวไปสร้างแบบจำลองดังตารางที่ 3.7

ตารางที่ 3. 7 ตัวอย่างคุณลักษณะเชิงบวกและเชิงลบ

Message	เครียดมาก	อันตรายมาก	เลิกหวัง	ขัดขวาง	ไปไม่ได้	ถี่ถ้วน	อบอุ่น	**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

3.3.3 การคัดเลือกคุณลักษณะ

การคัดเลือกคุณลักษณะมี 3 วิธี การคัดเลือกคุณลักษณะด้วย Chi-Square การคัดเลือกคุณลักษณะด้วย TFIDF และการคัดเลือกคุณลักษณะ ด้วย BM25

3.3.3.1 การคัดเลือกคุณลักษณะด้วย Chi-Square

จากสร้างคลังคำศัพท์ (Bag of Words) ผู้วิจัยจึงได้ทำการคัดเลือกเอา คงเหลือคำที่เป็นคุณลักษณะเชิงบวกและเชิงลบ ที่สามารถนำไปใช้ในงานวิจัยนี้ จำนวน 236 คำ โดยแบ่งคำออก เป็น 2 กลุ่ม คือคำแทนคุณลักษณะเชิงบวกและคำแทนคุณลักษณะเชิงลบ

แล้วนำมาใช้วิธีการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะเชิงบวกเชิงลบ และลบคำ (attribute) ที่มีค่าเท่ากับศูนย์ออก สามารถนำไปใช้ในงานวิจัยนี้ จำนวน 83 คำ โดยแบ่งคำออก เป็น 2 กลุ่ม คือคำแทนคุณลักษณะเชิงบวกและคำแทนคุณลักษณะเชิงลบดังตารางที่ 3.8 และตารางที่ 3.9 การตั้งค่า (parameter) การคัดเลือกคุณลักษณะด้วย Chi-Square

ตารางที่ 3. 8 ตัวอย่างผลการลบคำ (attribute) ที่มีค่าเท่ากับศูนย์ออก

Message	เครียดมาก	เลิกหวัง	ขัดขวาง	ถี่ถ้วน	**P**	**N**	**Class**
1	-1	0	0	0	0	-1	N
2	0	-1	0	0	0	-1	N
3	0	0	-1	0	0	-1	N
4	0	0	0	1	1	0	P

ตารางที่ 3. 9 การตั้งค่า (parameter) การคัดเลือกคุณลักษณะด้วย Chi-Square

weka.attributeSelection.CfsSubsetEval	
หัวข้อ	การตั้งค่า
debug	False
doNotCheckCapabilities	False
locallyPredictive	True
missingSeparate	False
numThreads	1
poolSize	1
preComputeCorrelationMatrix	False

3.3.3.2 การคัดเลือกคุณลักษณะด้วยTFIDF

term frequency-inverse document frequency เป็นเทคนิคที่พิจารณาองค์ประกอบของคำภายในประโยค (และเอกสาร) เป็นหลักโดยจะไม่นำลำดับของคำภายในเอกสารมาใช้วิเคราะห์ประกอบด้วย เทคนิคนี้มีอยู่ 2 องค์ประกอบด้วยกันคือ Term Frequency (TF) และ Inverse Document Frequency (IDF)

Term-Frequency (TF) ค่าของ Term Frequency เป็นค่าที่บอกความถี่ของคำแต่ละคำที่ปรากฏในเอกสารเอกสารหนึ่ง โดยคิดคำนวณจากการนำจำนวนครั้งที่คำนั้น ๆ ปรากฏในเอกสารมาหารด้วยจำนวนคำทั้งหมดในเอกสาร ดังสมการที่ 3.1

$$TF(\text{ของคำคำหนึ่ง}) = \frac{\text{จำนวนของคำนั้นๆ ในเอกสาร}}{\text{จำนวนของคำทั้งหมดในเอกสาร}} \quad (3.1)$$

เพื่อแสดงการตัวอย่างการคำนวณค่า TF (รวมถึง ค่า IDF และ TF-IDF ในตัวอย่างต่อไป) กรณียกตัวอย่าง ประกอบด้วยคำทั้งหมด 49 คำ ดังนั้นเราสามารถคำนวณค่า Term Frequency ของคำแต่ละคำได้ดังต่อไปนี้ ดังตารางที่ 3.10

ตารางที่ 3. 10 ตัวอย่างการคำนวณค่า TF

คำ	จำนวนคำ	Term-Frequency (TF)
เครียดมาก	2	2/49
เลิกหวัง	1	1/49
ขัดขวาง	1	1/49
ถี่ถ้วน	2	2/49

Inverse Document Frequency (IDF) เป็นการคำนวณค่าน้ำหนัก (weight) ความสำคัญของแต่ละคำโดยจะคำที่พบเจอได้บ่อย (ในหลายๆเอกสาร) จะมีค่า IDF ต่ำ ซึ่งบ่งบอกว่าคำเหล่านั้นจะไม่สามารถดึงเอาจุดเด่นของเอกสารที่คำเหล่านั้นปรากฏอยู่ออกมาได้ดี ค่า IDF สามารถคำนวณได้ ดังสมการที่ 3.2

$$IDF\left(\text{ของคำคำหนึ่ง}\right) = \log\left(\frac{\text{จำนวนเอกสารทั้งหมดที่ใช้พิจารณา}}{\text{จำนวนเอกสารที่มีคำคำนั้นปรากฏอยู่}}\right) \quad (3.2)$$

สำหรับตัวอย่างการคำนวณค่า IDF นั้น เราจะใช้ตัวอย่างเอกสารชุดเดียวกับกรณียกตัวอย่างการคำนวณ ค่า TF ด้านบน เมื่อพิจารณาเอกสารแล้วมีทั้งหมด 2,920 เอกสารทั้งหมดที่ใช้ในการพิจารณา ดังนั้นจึงสามารถคำนวณค่า IDF ได้ดังตารางที่ 3.11

ตารางที่ 3. 11 ตัวอย่างการคำนวณค่า IDF

คำ	จำนวนเอกสารที่ปรากฏ	Inverse Document Frequency (IDF)
เครียดมาก	2	Log(2920/2) = 3.16
เลิกหวัง	1	Log(2920/1) = 3.46
ขัดขวาง	1	Log(2920/1) = 3.46
ถี่ถ้วน	1	Log(2920/1) = 3.46

คำนวณค่า TF-IDF เมื่อนำการคำนวณทั้งสองส่วนมาคูณกัน จะได้การคำนวณ TF-IDF ดังสมการที่ 3.3

$$TFIDF = TF \times IDF \quad (3.3)$$

เพราะฉะนั้นค่า TF-IDF ของแต่ละคำในเอกสารที่ 1 ถูกคำนวณได้ดังต่อไปนี้ได้ดังตารางที่ 3.12

ตารางที่ 3. 12 ตัวอย่างค่าคำนวณ TF-IDF

คำ	**TF**	**IDF**	**TFIDF = TF x IDF**
เครียดมาก	2/49	3.16	0.13
เลิกหวัง	1/49	3.46	0.07
ขัดขวาง	1/49	3.46	0.07
ถี่ถ้วน	2/49	3.46	0.14

3.3.3.3 การคัดเลือกคุณลักษณะด้วย BM25

ฟังก์ชันการดึงข้อมูลที่จัดอันดับชุดของเอกสารตามคำค้นหาที่ปรากฏในเอกสารแต่ละฉบับโดยไม่คำนึงถึงความใกล้ชิดกันภายในเอกสาร เป็นกลุ่มฟังก์ชันการให้คะแนนที่มีส่วนประกอบและพารามิเตอร์ต่างกันเล็กน้อย หนึ่งในอินสแตนซ์ที่โดดเด่นที่สุดของฟังก์ชันมีดังสมการที่ 3.4

$$BM25 = \frac{f(q,D)*(k+1)}{f(t,D)+k*(1-b+b*\frac{D}{d_{avg}})} * \log(\frac{N-N(q)+0.5}{N(q)+0.5}+1) \quad (3.4)$$

f(q,D) คือจำนวนครั้งที่คำว่า q เกิดขึ้นในเอกสาร D

q คือจำนวนครั้งที่คำเกิดขึ้นในเอกสารนั้น

D คือจำนวนคำในเอกสารทั้งหมด

d_{avg} คือจำนวนคำเฉลี่ยต่อเอกสารทั้งหมด

k ค่าเริ่มต้นประมาณ 1.2

b ค่าเริ่มต้นประมาณ 0.75

b และ k เป็นไฮเปอร์พารามิเตอร์ที่ปรับได้สำหรับ BM25

N จำนวนเอกสาร

N(q) เอกสารที่มีคำที่ค้นหานั้น

0.5 ค่าคงที่

1 ค่าคงที่

โดยการแบ่งผลเป็นตารางสมการย่อย BM25 ส่วนที่ 1 ดังสมการที่ 3.5 และตารางที่ 3.13 ตัวอย่างการคำนวณค่าสมการย่อย BM25 ส่วนที่ 1

$$\frac{f(q,D)*(k+1)}{f(t,D)+k*(1-b+b*\frac{D}{d_{avg}})} \quad (3.5)$$

ตารางสมการย่อย BM25 ส่วนที่ 2 ดังสมการที่ 3.6 และตารางที่ 3.14 ตัวอย่างการคำนวณค่าสมการย่อย BM25 ส่วนที่ 2

$$\log(\frac{N-N(q)+0.5}{N(q)+0.5}+1) \quad (3.6)$$

และ ตัวอย่างค่าสมการย่อย ส่วนที่ 1 คูณกับ ส่วนที่ 2 ดังตารางที่ 3.15 ตัวอย่างค่าสมการย่อย ส่วนที่ 1 คูณกับ ส่วนที่ 2

ตารางที่ 3. 13 ตัวอย่างการคำนวณค่าสมการย่อย BM25 ส่วนที่ 1

คำ	$\frac{f(q,D)*(k+1)}{f(t,D)+k*(1-b+b*\frac{D}{d_{avg}})}$
เครียดมาก	0.000823497
เลิกหวัง	0.001646392
ขัดขวาง	0.001646392
ถี่ถ้วน	0.000823497

ตารางที่ 3. 14 ตัวอย่างการคำนวณค่าสมการย่อย BM25 ส่วนที่ 2

คำ	$\log(\frac{N-N(q)+0.5}{N(q)+0.5}+1)$
เครียดมาก	2.453823066
เลิกหวัง	5.451553342
ขัดขวาง	5.451553342
ถี่ถ้วน	2.453823066

ตารางที่ 3. 15 ตัวอย่างค่าสมการย่อย ส่วนที่ 1 คูณกับ ส่วนที่ 2

คำ	**สมการย่อย BM25 ส่วนที่ 1**	**สมการย่อย BM25 ส่วนที่ 2**	**สมการ BM25**
เครียดมาก	0.000823497	2.453823066	0.00202072
เลิกหวัง	0.001646392	5.451553342	0.00897539
ขัดขวาง	0.001646392	5.451553342	0.00897539
ถี่ถ้วน	0.000823497	2.453823066	0.00202072

3.4 การสร้างแบบจำลอง

การสร้างแบบจำลอง งานวิจัยฉบับนี้ได้จำแนกความคิดเห็น ให้การจำแนกความคิดเห็นที่เหมาะสมที่สุดแล้วนำเอาเทคนิคต่าง ๆ เพิ่มประสิทธิภาพในการจำแนกความรู้สึก ซึ่งการคัดเลือกคุณลักษณะที่นำมาใช้ ได้แก่ Bag of Words Chi-Square TFIDF และ BM25 แล้วนำเอาเทคนิคต่าง ๆ มาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะ และทำการเปรียบเทียบ เทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความความรู้สึกของคนไทย ในขั้นตอนนี้ผู้วิจัยนำ 6 เทคนิค ที่มีประสิทธิภาพและการตั้งค่าเทคนิคในโปรแกรม Weka ในการจำแนกดังนี้ เทคนิคต้นไม้ตัดสินใจ แบบ C4.5 (Decision tree C4.5) เทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine :SVM) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (multi-layer perceptron) เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) ดังตารางที่ 3.16 การตั้งค่า (parameter) เทคนิคต้นไม้ตัดสินใจ แบบ C4.5 (Decision tree C4.5) ตารางที่ 3.17 การตั้งค่า(parameter)เทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector ตารางที่ 3.18 การตั้งค่า (parameter) เทคนิคนาอีฟเบย์ (Naïve Bayes) ตารางที่ 3.19 การตั้งค่า (parameter) เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ตารางที่ 3.20 การตั้งค่า (parameter) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (multi-layer perceptron) ตารางที่ 3. 21 การตั้งค่า (parameter) เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning)

ตารางที่ 3. 16 การตั้งค่า (parameter) เทคนิคต้นไม้ตัดสินใจ แบบ C4.5

weka.classifiers.trees.J48	
หัวข้อ	การตั้งค่า
batchSize	100
binarySplits	False
collapseTree	True
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
doNotMakeSplitPointActualValue	False
minNumObj	2
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	True
unpruned	False
useLaplace	False
useMDLcorrection	True

ตารางที่ 3. 17 การตั้งค่า (parameter) เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

weka.classifiers.functions.LibSVM	
หัวข้อ	การตั้งค่า
SVMType	C-SVC (dlassification)
batchSize	100
cacheSize	40.0
coefo	0.0
cost	1.0
debug	False
degree	3
doNotCheckCapabilities	False
doNotReplaceMissingValues	False
eps	0.001
gamma	0.0
doNotCheckCapabilities	False
doNotReplaceMissingValues	False
eps	0.001
gamma	0.0
kernelType	radial basis function: exp(-gamma*lu-v)^2)
loss	0.1
modelFile	ad
normalize	False
nu	0.5
numDecimalPlaces	2
probabilityEstimates	False
seed	1
shrinking	True
weights	

ตารางที่ 3. 18 การตั้งค่า (parameter) เทคนิคนาอีฟเบย์

weka.classifiers.bayes.NaiveBayes	
หัวข้อ	การตั้งค่า
batchSize	100
debug	False
displayModelInOldFormat	False
doNotCheckCapabilities	False
numDecimalPlaces	2
useKernelEstimator	False
useSupervisedDiscretization	False

ตารางที่ 3. 19 การตั้งค่า (parameter) เทคนิคเคเนียเรสเนเบอร์

weka.classifiers.lazy.IBk	
หัวข้อ	การตั้งค่า
KNN	1
batchSize	100
crossValidate	False
debug	False
distanceWeighting	No distance weighting
doNotCheckCapabilities	False
meanSquared	False
nearestNeighbourSearchAlgorithm	LineerNNSearch
numDecimalPlaces	2
windowSize	0

ตารางที่ 3. 20 การตั้งค่า (parameter) เทคนิคเพอร์เซ็ปตรอนหลายชั้น

weka.classifiers.functions.MultilayerPerceptron	
หัวข้อ	การตั้งค่า
GUI	False
autoBuild	True
batchSize	100
debug	False
decay	False
doNotCheckCapabilities	False
hiddenLayers	a
learningRate	0.3
momentum	0.2
nominalToBinaryFilter	True
normalizeAttributes	True
normalizeNumericClass	True
numDecimalPlaces	2
reset	True
resume	False
seed	0
trainingTime	500
validationSetSize	0
validationThreshold	20

ตารางที่ 3. 21 การตั้งค่า (parameter) เทคนิคแบบระบบเรียนรู้เชิงลึก

weka.classifiers.functions.Dl4jMlpClassifier	
หัวข้อ	การตั้งค่า
log config	LogConfiguration
layer specifcation.	1 weka.dl4j.layers.Layer
Preview zoo model layer specification in GUI	False
number of epochs	10
instance iterator	DefaultInstanceIterator -bs 1
early stopping	EarlyStopping
network configuration	NeuralNetConfiguration
set the iteration listener	EpochListener
zooModel	CustomNet
attribute normalization	Standardize training data
set the cache mode	MEMORY
data queue size	0
resume False	False
Preserve flesystem cache	False
Number of GPUS	1
Size of prefetch buffer for multiple GPUs	24
Model parameter averaging frequency	10
batchSize	100
debug	False
doNotCheckCapabilities	False
numDecimalPlaces	2
seed	1

3.5 การวัดประสิทธิภาพของแบบจำลอง

การวัดประสิทธิภาพของแบบจำลอง งานวิจัยฉบับนี้ใช้การวัดประสิทธิภาพของแบบจำลองด้วยวิธี 10-fold Cross Validation ทำการเลือกสุ่มข้อมูลออกเป็น K ชุดเท่าๆ กัน จากนั้นจะทำการทดลองครั้งแรกด้วยข้อมูลชุดที่ 1 ซึ่งเป็นข้อมูลทดสอบและกำหนดให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน และในการทดลองครั้งที่สองจะใช้ข้อมูลชุดที่ 2 เป็นชุดข้อมูลทดสอบและให้ข้อมูลชุดที่เหลือเป็นข้อมูลชุดสอน ทำจนกระทั้งข้อมูลทุกชุดข้อมูลได้ถูกนำมาเป็นชุดข้อมูลทดสอบทั้งหมด ซึ่งจำนวนในการทดสอบมีจำนวนเท่ากับ K ครั้ง โดยผลลัพธ์ที่ได้นั้นจะมาคำนวณหาค่าเฉลี่ยความถูกต้องของการจำแนกข้อมูลในแต่ละรอบ โดยวิธีการทดสอบประสิทธิภาพแบบ 10-fold Cross Validation มีข้อเสียคือ จะต้องทำการเริ่มทดสอบใหม่โดยจะต้องทำทั้งหมด K รอบ การวัดประสิทธิภาพการทำงานในแต่ละขั้นตอนวิธี สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

บทที่ 4 ผลการวิจัยและการการอภิปราย

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความ และเพื่อสร้างและเปรียบเทียบประสิทธิภาพของ แบบบจำลองความรู้สึกของคนไทยต่อโรคโควิด 19

วัตถุประสงค์ในการศึกษาวิธีการการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความ ดังกระบวนการการทำเหมืองข้อมูลมี 5 ขั้นตอนรวมถึง การรวบรวมข้อมูลได้รวบรวมความคิดเห็นของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ เว็บไซต์ทวิตเตอร์ และพันทิป ตั้งแต่วันที่ 23 มกราคม 2563 ถึงวันที่ 1 พฤษภาคม 2563 เริ่มต้นโรคโควิด 19 ในประเทศไทยจำนวน 2,920 ความคิดเห็น การเตรียมข้อมูลประกอบด้วยการทำความสะอาดข้อความที่นำมาจากสื่อสังคมออนไลน์

การจัดทำดัชนีข้อมูลนำข้อมูลมาแยกประเภทของคำหรือชนิดของคำที่อยู่ในประโยค เช่น คำนาม (Noun) คำสรรพนาม (Pronoun) คำกริยา (Verb) คำวิเศษณ์ (Adverb) คำบุพบท (Preposition) คำสันธาน (Conjunctions) เนื่องจาก ประโยค วลี คำพูด ต่าง ๆ ที่ใช้สื่อสารกันล้วนเกิดขึ้นจากการนำ คำต่าง ๆ มาประกอบกันเป็น ส่วนต่าง ๆ ที่ทำหน้าที่ต่างกันในประโยคแล้วนำไปสร้างคลังคำศัพท์ และการกำหนดคลาส

การสร้างคลังคำศัพท์ (Bag of Words) เป็นการนำคำที่แยกประเภท ในกรณีที่ซ้ำกันนำมาเพียงหนึ่งคำ ส่วนกรณีคำที่ไม่ซ้ำกันนำมาทั้งหมด คงเหลือทั้งหมด 9,037 คำ โดยการใช้พจนานุกรมไทยแยกชนิดของคำ จากนั้นทำการคัดเลือก เอาเฉพาะคำวิเศษณ์ ที่เป็นคำที่บ่งบอกถุงความรู้สึกได้ดี แล้วทำการคัดเลือกคำคุณลักษณะที่มีความหมายเชิงบวกเชิงลบแล้วแทนค่าของจำนวนคำคุณลักษณะที่ไม่ปรากฏเป็น 0 แทนค่าจำนวนคำคุณลักษณะเชิงบวกที่ปรากฏเป็น 1 และแทนค่าจำนวนคำคุณลักษณะเชิงลบที่ปรากฏเป็น -1 จะเห็นความถี่ของคำคุณลักษณะที่ถูกจำแนกออกมาเป็น เชิงบวกหรือเชิงลบ ในแต่ละความคิดเห็น และทำการแบ่งความคิดเห็นเชิงบวกและเชิงลบ โดยใช้หลักเกณฑ์ของภาษาไทย แล้วคัดเลือกเอา ชุดข้อมูลคำวิเศษณ์ จำนวน 236 คำ โดยแบ่งคำออก เป็น 2 กลุ่ม การกำหนดคลาส

งานวิจัยฉบับนี้ได้สร้างตัวแทนเอกสารนั้นจะใช้วิธีในการนำคำบ่งชี้คุณลักษณะในชุดข้อมูลมาเรียงกันเพื่อทำการนับความถี่ของการเกิดขึ้นของคำนั้น ๆ จากนั้นจึงนำค่าจำนวนความถี่ของคำมาสร้างเวกเตอร์ตัวแทนเอกสาร และคำบ่งชี้ที่ไม่ปรากฏในเอกสารจะมีค่าเป็น 0 จากนั้นทำการนับจำนวนคำในแต่ละคุณลักษณะ โดยใช้การนับจำนวนความถี่ของค่าคุณลักษณะว่ามีจำนวนเท่าใด เพื่อ

นำมาเปรียบเทียบกัน เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงบวกมากกว่าความถี่ของคุณลักษณะเชิงลบให้ตัวแปรตามเป็น ความคิดเห็นเชิงบวก แทนด้วย P เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงลบมากกว่าความถี่ของคุณลักษณะเชิงบวกให้ตัวแปรตาม เป็นความคิดเห็นเชิงลบ แทนด้วย N แต่ถ้าหากความถี่ของคุณลักษณะเชิงบวกและเชิงลบเท่ากัน จะแทน ด้วย NE และข้อความจะถูกตัดออก คงเหลือแถวที่มีข้อความ P และ N แล้วการคัดเลือกคุณลักษณะ 2 รูปแบบ รูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 รูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25

การสร้างแบบจำลอง จำแนกความคิดเห็นที่เหมาะสมที่สุดแล้วนำเอาเทคนิคต่าง ๆ เพิ่มประสิทธิภาพ แล้วนำเอาเทคนิคมาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะและทำการเปรียบเทียบ 6 เทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความความรู้สึกของคนไทย มาใช้ในการวิเคราะห์ และวัดประสิทธิภาพของแบบจำลองวัดประสิทธิภาพของแบบจำลองด้วยวิธี 10-fold Cross Validation สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

4.1 วิธีการวัดผลการวิจัย

ตามวัตถุประสงค์เพื่อสร้างและเปรียบเทียบประสิทธิภาพของ แบบบจำลองความรู้สึกของคนไทยต่อโรคโควิด 19 งานวิจัยฉบับนี้ได้จำแนกความคิดเห็นที่เหมาะสมที่สุดแล้วนำเอาเทคนิคต่าง ๆ เพิ่มประสิทธิภาพในการจำแนกความรู้สึก ซึ่งการคัดเลือกคุณลักษณะที่นำมาใช้ ได้แก่ Bag of Words Chi-Square TFIDF และ BM25 แล้วนำเอาเทคนิคต่าง ๆ มาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะ และทำการเปรียบเทียบ เทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความความรู้สึกของคนไทย มาใช้ในการวิเคราะห์ทั้งหมด 6 เทคนิค เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เทคนิคซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) เทคนิคนาอีฟเบย์ (Naive Bayes) เทคนิคเคเนียเรสเนเบอร์ (K-Nearest Neighbor: KNN) เทคนิคเพอร์เซปตรอนหลายชั้น (Multi-layer Perceptron) เทคนิคแบบระบบเรียนรู้เชิงลึก (Deep learning) จากนั้นใช้หลักการ 10-fold Cross Validation ในการแบ่งกลุ่มข้อมูลเป็นชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ และวัดประสิทธิภาพของแบบจำลองด้วย ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

4.2 ผลการวิจัย

ผลการวิจัยจากการวิธีการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักที่เหมาะสมต่อการจำแนกความรู้สึกของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ โดยได้ทำการวิเคราะห์การจำแนกความคิดเห็นที่เหมาะสมที่สุดแล้วนำเอาเทคนิคต่าง ๆ เพิ่มประสิทธิภาพในการจำแนกความรู้สึก ซึ่งการคัดเลือกคุณลักษณะที่นำมาใช้ แล้วนำเอาเทคนิคต่าง ๆ มาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะ 2 รูปแบบ รูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 รูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 และทำการเปรียบเทียบ เทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความความรู้สึกของคนไทย มาใช้ในการวิเคราะห์ทั้งหมด 6 เทคนิค ในการเปรียบเทียบรูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 สามารถแสดงได้ดัง ภาพที่ 4.1 ค่าความแม่นยำ (Precision) ภาพที่ 4.2 ค่าความระลึก (Recall) และภาพที่ 4.3 ค่าความถ่วงดุล (F-measure)

และการเปรียบเทียบรูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 สามารถแสดงได้ดัง ตารางที่ 4. 4 ค่าความแม่นยำ (Precision) ผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square ตารางที่ 4. 5 ค่าความระลึก (Recall) ผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square ตารางที่ 4. 6 ค่าความถ่วงดุล (F-measure) ผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square ตามลำดับ

ตารางที่ 4. 1 ค่าความแม่นยำ (Precision)

Precision (%)	C4.5	SVM	NB	KNN	MLP	DL
Chi-Squared	91.50	88.80	94.60	99.20	99.20	**99.30**
TFIDF	91.70	82.80	97.20	99.40	**99.60**	98.60
BM25	91.70	82.80	97.20	99.40	**99.60**	98.60

จากตารางที่ 4.1 แสดงการเปรียบเทียบค่าความแม่นยำ (Precision) ของ 6 เทคนิค เปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความแม่นยำสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะของเทคนิค C4.5 ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 91.70 เทคนิค SVM ได้ค่าความแม่นยำสูงสุดที่ Chi-Squared ร้อยละ 88.80 เทคนิค NB ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 97.20 เทคนิค KNN ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ

99.40 เทคนิค MLP ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.60 เทคนิค DL ได้ค่าความแม่นยำสูงสุดที่ Chi-Squared ร้อยละ 99.30

เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่ให้ค่าความแม่นยำสูงสุด การคัดเลือกคุณลักษณะการคัดเลือกคุณลักษณะ Chi-Squared ที่เทคนิค DL ร้อยละ 99.30 การคัดเลือกคุณลักษณะ TFIDF ที่เทคนิค MLP ร้อยละ 99.60 การคัดเลือกคุณลักษณะ BM25 ที่เทคนิค MLP ร้อยละ 99.60

ตารางที่ 4. 2 ค่าความระลึก (Recall)

Recall (%)	C4.5	SVM	NB	KNN	MLP	DL
chi-squared	91.20	87.20	94.40	99.20	99.20	**99.30**
TFIDF	91.40	82.80	97.20	99.40	**99.60**	98.60
BM25	91.40	82.80	97.20	99.40	**99.60**	98.60

จากตารางที่ 4.2 แสดงการเปรียบเทียบค่าความระลึก (Recall) ของ 6 เทคนิค เปรียบเทียบ กับการคัดเลือกคุณลักษณะที่มีค่าความระลึกสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะของเทคนิค C4.5 ได้ค่าความความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 91.40 เทคนิค SVM ได้ค่าความระลึกสูงสุดที่ Chi-Squared ร้อยละ 87.20 เทคนิค NB ได้ค่าความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 97.20 เทคนิค KNN ได้ค่าความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.40 เทคนิค MLP ได้ค่าความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.60 เทคนิค DL ได้ค่าความระลึกสูงสุดที่ Chi-Squared ร้อยละ 99.30

เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่ให้ค่าความระลึกสูงสุด การคัดเลือกคุณลักษณะ Chi-Squared ที่เทคนิค DL ร้อยละ 99.30 การคัดเลือกคุณลักษณะ TFIDF ที่เทคนิค MLP ร้อยละ 99.60 การคัดเลือกคุณลักษณะ BM25 ที่เทคนิค MLP ร้อยละ 99.60

ตารางที่ 4. 3 ค่าความถ่วงดุล (F-measure)

F-measure (%)	C4.5	SVM	NB	KNN	MLP	DL
chi-squared	90.10	83.90	94	99.20	99.20	**99.30**
TFIDF	90.30	90.60	97.20	99.40	**99.60**	98.60
BM25	90.30	90.60	97.10	99.40	**99.60**	98.60

จากตารางที่ 4.3 แสดงการเปรียบเทียบและค่าความถ่วงดุล (F-measure) ของ 6 เทคนิค เปรียบเทียบกับการคัดเลือกคุณลักษณะที่มีค่าความถ่วงดุลสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะของเทคนิค C4.5 ได้ค่าความถ่วงดุลสูงสุดที่ Chi-Squared ร้อยละ 90.10 เทคนิค SVM ได้ค่าความถ่วงดุลสูงสุดที่ TFIDF และ BM25 ร้อยละ 90.60 เทคนิค NB ได้ค่าความถ่วงดุลสูงสุดที่ TFIDF ร้อยละ 97.20 เทคนิค KNN ได้ค่าความถ่วงดุลสูงสุดที่ Chi-Squared ร้อยละ 99.20 เทคนิค MLP ได้ค่าความถ่วงดุลสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.60 เทคนิค DL ได้ค่าความถ่วงดุลสูงสุดที่ Chi-Squared ร้อยละ 99.30

เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่ให้ค่าความถ่วงดุลสูงสุด การคัดเลือกคุณลักษณะ Chi-Squared ที่เทคนิค DL ร้อยละ 99.30 การคัดเลือกคุณลักษณะ TFIDF ที่เทคนิค MLP ร้อยละ 99.60 การคัดเลือกคุณลักษณะ BM25 ที่เทคนิค MLP ร้อยละ 99.60

ตารางที่ 4. 4 ค่าความแม่นยำ (Precision) ผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square

Precision (%)	C4.5	SVM	NB	KNN	MLP	DL
TFIDF	90.90	82.80	94.60	**99.20**	**99.20**	93.30
BM25	91.50	82.80	94.60	**99.20**	**99.20**	93.30

จากตารางที่ 4.4 แสดงการเปรียบเทียบค่าความแม่นยำ (Precision) เมื่อผ่านวิธีการการคัดเลือกคุณลักษณะด้วย Chi-Square ของ 6 เทคนิค เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความแม่นยำสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF ที่เทคนิค KNN และ MLP ร้อยละ 99.20 การคัดเลือกคุณลักษณะด้วย BM25 ที่เทคนิค KNN และ MLP ร้อยละ 99.20

ตารางที่ 4. 5 ค่าความระลึก (Recall) ผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square

Recall (%)	C4.5	SVM	NB	KNN	MLP	DL
TFIDF	99.30	100	94.40	**99.20**	**99.20**	93.30
BM25	91.20	100	94.40	**99.20**	**99.20**	93.30

จากตารางที่ 4.5 แสดงการเปรียบเทียบค่าความระลึก (Recall) เมื่อผ่านวิธีการการคัดเลือกคุณลักษณะด้วย Chi-Square ของ 6 เทคนิค เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความระลึกสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF ที่เทคนิค SVM ร้อยละ 100 การคัดเลือกคุณลักษณะด้วย BM25 ที่เทคนิค SVM ร้อยละ 100

ตารางที่ 4. 6 ค่าความถ่วงดุล (F-measure) ผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square

F-measure (%)	C4.5	SVM	NB	KNN	MLP	DL
TFIDF	94.90	90.60	94	**99.20**	**99.20**	93.30
BM25	90.10	90.60	94	**99.20**	**99.20**	93.30

จากตารางที่ 4.6 แสดงการเปรียบเทียบและค่าความถ่วงดุล (F-measure) เมื่อผ่านวิธีการการคัดเลือกคุณลักษณะด้วย Chi-Square ของ 6 เทคนิค เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความถ่วงดุลสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF ที่เทคนิค KNN และ MLP ร้อยละ 99.20 การคัดเลือกคุณลักษณะด้วย BM25 ที่เทคนิค KNN และ MLP ร้อยละ 99.20

บทที่ 5 สรุปผล อภิปรายผล และข้อเสนอแนะ

งานวิจัยนี้ได้นำเอาการกระบวนการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความ เพื่อสร้างและเปรียบเทียบประสิทธิภาพของแบบบจำลองความรู้สึกของคนไทยต่อโรคโควิด 19 จึงได้นำเอาการกระบวนการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็น

การทำเหมืองข้อมูลมี 5 ขั้นตอนรวมถึง การรวบรวมข้อมูลได้รวบรวมความคิดเห็นของคนไทยต่อโรคโควิด 19 บนสื่อสังคมออนไลน์ เว็บไซต์ทวิตเตอร์ และพันทิป ตั้งแต่วันที่ 23 มกราคม 2563 ถึงวันที่ 1 พฤษภาคม 2563 เริ่มต้นโรคโควิด 19 ในประเทศไทยจำนวน 2,920 ความคิดเห็น การเตรียมข้อมูลประกอบด้วยการทำความสะอาดข้อความที่นำมาจากสื่อสังคมออนไลน์

การจัดทำดัชนีข้อมูลนำข้อมูลมาแยกประเภทของคำหรือชนิดของคำที่อยู่ในประโยค เช่น คำนาม (Noun) คำสรรพนาม (Pronoun) คำกริยา (Verb) คำวิเศษณ์ (Adverb) คำบุพบท (Preposition) คำสันธาน (Conjunctions) เนื่องจาก ประโยค วลี คำพูด ต่าง ๆ ที่ใช้สื่อสารกันล้วนเกิดขึ้นจากการนำ คำต่าง ๆ มาประกอบกันเป็น ส่วนต่าง ๆ ที่ทำหน้าที่ต่างกันในประโยคแล้วนำไปสร้างคลังคำศัพท์ และการกำหนดคลาส

การสร้างคลังคำศัพท์ (Bag of Words) เป็นการนำคำที่แยกประเภท ในกรณีที่ซ้ำกันนำมาเพียงหนึ่งคำ ส่วนกรณีคำที่ไม่ซ้ำกันนำมาทั้งหมด คงเหลือทั้งหมด 9,037 คำ โดยการใช้พจนานุกรมไทยแยกชนิดของคำ จากนั้นทำการคัดเลือก เอาเฉพาะคำวิเศษณ์ ที่เป็นคำที่บ่งบอกถุงความรู้สึกได้ดีแล้วทำการคัดเลือกคำคุณลักษณะที่มีความหมายเชิงบวกเชิงลบแล้วแทนค่าของจำนวนคำคุณลักษณะที่ไม่ปรากฏเป็น 0 แทนค่าจำนวนคำคุณลักษณะเชิงบวกที่ปรากฏเป็น 1 และแทนค่าจำนวนคำคุณลักษณะเชิงลบที่ปรากฏเป็น -1 จะเห็นความถี่ของคำคุณลักษณะที่ถูกจำแนกออกมาเป็น เชิงบวกหรือเชิงลบ ในแต่ละความคิดเห็น และทำการแบ่งความคิดเห็นเชิงบวกและเชิงลบ โดยใช้หลักเกณฑ์ของภาษาไทย แล้วคัดเลือกเอา ชุดข้อมูลคำวิเศษณ์ จำนวน 236 คำ โดยแบ่งคำออก เป็น 2 กลุ่ม การกำหนดคลาส

งานวิจัยฉบับนี้ได้สร้างตัวแทนเอกสารนั้นจะใช้วิธีในการนำคำบ่งชี้คุณลักษณะในชุดข้อมูลมาเรียงกันเพื่อทำการนับความถี่ของการเกิดขึ้นของคำนั้น ๆ จากนั้นจึงนำค่าจำนวนความถี่ของคำมาสร้างเวกเตอร์ตัวแทนเอกสาร และคำบ่งชี้ที่ไม่ปรากฏในเอกสารจะมีค่าเป็น 0 จากนั้นทำการนับจำนวนคำในแต่ละคุณลักษณะ โดยใช้การนับจำนวนความถี่ของค่าคุณลักษณะว่ามีจำนวนเท่าใด เพื่อนำมาเปรียบเทียบกัน เมื่อจำนวนความถี่ของคุณลักษณะความคิดเห็นเชิงบวกมากกว่าความถี่ของคุณลักษณะเชิงลบให้ตัวแปรตามเป็น ความคิดเห็นเชิงบวก แทนด้วย P เมื่อจำนวนความถี่ของ

คุณลักษณะความคิดเห็นเชิงลบมากกว่าความถี่ของคุณลักษณะเชิงบวกให้ตัวแปรตาม เป็นความคิดเห็นเชิงลบ แทนด้วย N แต่ถ้าหากความถี่ของคุณลักษณะเชิงบวกและเชิงลบเท่ากัน จะแทน ด้วย NE และข้อความจะถูกตัดออก คงเหลือแถวที่มีข้อความ P และ N แล้วนำวิธีการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะเชิงบวกและเชิงลบ แล้วการคัดเลือกคุณลักษณะ 2 รูปแบบ รูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 รูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25

การสร้างแบบจำลอง จำแนกความคิดเห็นที่เหมาะสมที่สุดแล้วนำเอาเทคนิคต่าง ๆ เพิ่มประสิทธิภาพ แล้วนำเอาเทคนิคมาใช้ในการวัดประสิทธิภาพของการคัดเลือกคุณลักษณะและทำการเปรียบเทียบ 6 เทคนิคที่ใช้สร้างแบบจำลองเพื่อจำแนกความความรู้สึกของคนไทย มาใช้ในการวิเคราะห์ และวัดประสิทธิภาพของแบบจำลองวัดประสิทธิภาพของแบบจำลองด้วยวิธี 10-fold Cross Validation สามารถวัดได้จากผลของการจำแนกกลุ่มข้อมูล และสามารถหาค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

5.1 สรุปผล

จากผลการวิจัยการคัดเลือกคุณลักษณะ 2 รูปแบบ รูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 รูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 การคัดเลือกคุณลักษณะที่ดีที่สุดสามารถหาค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าความถ่วงดุล (F-measure)

สรุปได้ว่า รูปแบบที่ 1 เปรียบเทียบการคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 การคัดเลือกคุณลักษณะเหมาะสมกับเทคนิคการจำแนกใดมากที่สุด พบว่า

เปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความแม่นยำสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะของเทคนิค C4.5 ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 91.70 เทคนิค SVM ได้ค่าความแม่นยำสูงสุดที่ Chi-Squared ร้อยละ 88.80 เทคนิค NB ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 97.20 เทคนิค KNN ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.40 เทคนิค MLP ได้ค่าความแม่นยำสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.60 เทคนิค DL ได้ค่าความแม่นยำสูงสุดที่ Chi-Squared ร้อยละ 99.30 และเปรียบเทียบการคัดเลือกคุณลักษณะที่ให้ค่าความแม่นยำสูงสุด การคัดเลือกคุณลักษณะ การคัดเลือกคุณลักษณะ Chi-Squared ที่เทคนิค DL ร้อยละ 99.30 การคัดเลือกคุณลักษณะ TFIDF ที่เทคนิค MLP ร้อยละ 99.60 การคัดเลือกคุณลักษณะ BM25 ที่เทคนิค MLP ร้อยละ 99.60

เปรียบเทียบ กับการคัดเลือกคุณลักษณะที่มีค่าความระลึกสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะของเทคนิค C4.5 ได้ค่าความความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 91.40 เทคนิค SVM ได้ค่าความระลึกสูงสุดที่ Chi-Squared ร้อยละ 87.20 เทคนิค NB ได้ค่าความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 97.20 เทคนิค KNN ได้ค่าความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.40 เทคนิค MLP ได้ค่าความระลึกสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.60 เทคนิค DL ได้ค่าความระลึกสูงสุดที่ Chi-Squared ร้อยละ 99.30 และเปรียบเทียบการคัดเลือกคุณลักษณะที่ให้ค่าความระลึกสูงสุด การคัดเลือกคุณลักษณะ Chi-Squared ที่เทคนิค DL ร้อยละ 99.30 การคัดเลือกคุณลักษณะ TFIDF ที่เทคนิค MLP ร้อยละ 99.60 การคัดเลือกคุณลักษณะ BM25 ที่เทคนิค MLP ร้อยละ 99.60

เปรียบเทียบกับการคัดเลือกคุณลักษณะที่มีค่าความถ่วงดุลสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะของเทคนิค C4.5 ได้ค่าความถ่วงดุลสูงสุดที่ Chi-Squared ร้อยละ 90.10 เทคนิค SVM ได้ค่าความถ่วงดุลสูงสุดที่ TFIDF และ BM25 ร้อยละ 90.60 เทคนิค NB ได้ค่าความถ่วงดุลสูงสุดที่ TFIDF ร้อยละ 97.20 เทคนิค KNN ได้ค่าความถ่วงดุลสูงสุดที่ Chi-Squared ร้อยละ 99.20 เทคนิค MLP ได้ค่าความถ่วงดุลสูงสุดที่ TFIDF และ BM25 ร้อยละ 99.60 เทคนิค DL ได้ค่าความถ่วงดุลสูงสุดที่ Chi-Squared ร้อยละ 99.30 และเปรียบเทียบการคัดเลือกคุณลักษณะที่ให้ค่าความถ่วงดุลสูงสุด การคัดเลือกคุณลักษณะ Chi-Squared ที่เทคนิค DL ร้อยละ 99.30 การคัดเลือกคุณลักษณะ TFIDF ที่เทคนิค MLP ร้อยละ 99.60 การคัดเลือกคุณลักษณะ BM25 ที่เทคนิค MLP ร้อยละ 99.60

สรุปได้ว่ารูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 การคัดเลือกคุณลักษณะเหมาะสมกับเทคนิคการจำแนกใดมากที่สุด พบว่า

เปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความแม่นยำ (Precision) เมื่อผ่านวิธีการการคัดเลือกคุณลักษณะด้วย Chi-Square ของ 6 เทคนิค เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความแม่นยำสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF ที่เทคนิค KNN และ MLP ร้อยละ 99.20 การคัดเลือกคุณลักษณะด้วย BM25 ที่เทคนิค KNN และ MLP ร้อยละ 99.20

เปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความระลึก (Recall) เมื่อผ่านวิธีการการคัดเลือกคุณลักษณะด้วย Chi-Square ของ 6 เทคนิค เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความระลึกสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF ที่เทคนิค SVM ร้อยละ 100 การคัดเลือกคุณลักษณะด้วย BM25 ที่เทคนิค SVM ร้อยละ 100

เปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่าความถ่วงดุล (F-measure) เมื่อผ่านวิธีการการคัดเลือกคุณลักษณะด้วย Chi-Square ของ 6 เทคนิค เมื่อเปรียบเทียบการคัดเลือกคุณลักษณะที่มีค่า

ความถ่วงดุลสูงสุด ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF ที่เทคนิค KNN และ MLP ร้อยละ 99.20 การคัดเลือกคุณลักษณะด้วย BM25 ที่เทคนิค KNN และ MLP ร้อยละ 99.20

เปรียบเทียบการคัดเลือกคุณลักษณะ จากผลการวิจัยการคัดเลือกคุณลักษณะทั้ง 2 รูปแบบ รูปแบบที่ 1 การคัดเลือกคุณลักษณะด้วย Chi-Square TFIDF และ BM25 รูปแบบที่ 2 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25

ผลปรากฏว่า การคัดเลือกคุณลักษณะเมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square แล้วค่อยคัดเลือกคุณลักษณะด้วย TFIDF มีค่าความระลึก (Recall) เพิ่มขึ้นที่เทคนิค C4.5 ที่ร้อยละ 99.30 จากเดิมร้อยละ 91.40 ที่เทคนิค SVM ที่ร้อยละ 100 จากเดิมร้อยละ 82.80 และคัดเลือกคุณลักษณะด้วย BM25 มีค่าความระลึก (Recall) เพิ่มขึ้น ที่เทคนิค SVM ที่ร้อยละ 100 จากเดิมร้อยละ 82.80 มีค่าความถ่วงดุล (F-measure) เพิ่มขึ้นที่เทคนิค C4.5 ที่ร้อยละ 94.90 จากเดิมร้อยละ 90.30 และเมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะ จำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 การคัดเลือกคุณลักษณะเหมาะสมกับเทคนิคการจำแนกมากที่สุด พบว่า ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-measure) ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 ที่เทคนิค KNN และ MLP ร้อยละ 99.20

5.2 อภิปรายผล

ตามวัตถุประสงค์ของการวิจัย เพื่อศึกษาวิธีการการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความเพื่อสร้างและเปรียบเทียบประสิทธิภาพของ แบบบจำลองความรู้สึกของคนไทยต่อโรคโควิด 19

จากการศึกษาวิธีการการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำเพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 โดยใช้หลักการเหมืองข้อความ พบว่า สาเหตุของการคัดเลือกคุณลักษณะที่ส่งผลต่อแบบจำลอง ทำให้การจำแนกมีประสิทธิภาพมากดีขึ้นที่ มีค่าความระลึก (Recall) เพิ่มขึ้น ที่เทคนิค SVM ที่ร้อยละ 100 จากเดิมร้อยละ 82.80 มีค่าความถ่วงดุล (F-measure) เพิ่มขึ้นที่เทคนิค C4.5 ที่ร้อยละ 94.90 จากเดิมร้อยละ 90.30 โดยเปรียบเทียบกับการคัดเลือกคุณลักษณะด้วย TFIDF และ BM 25 เมื่อผ่านการคัดเลือกคุณลักษณะด้วย Chi-Square มีการลดมิติของคำ ส่วนการทำงานของการคัดเลือกคุณลักษณะด้วย TFIDF เป็นเทคนิคที่พิจารณาองค์ประกอบของคำภายในประโยค เป็นหลักโดยจะไม่นำลำดับของคำภายในเอกสารมาใช้ จึงสามารถดึงคำภายในเอกสารมาใช้ วิเคราะห์ได้อย่างมีประสิทธิภาพ และ

BM25 เป็นฟังก์ชันอันดับที่จัดลำดับกลุ่มของเอกสารขึ้นอยู่กับคำค้นหาที่ปรากฏในเอกสาร เกี่ยวกับความสัมพันธ์ระหว่างคำค้นหาที่อยู่ในเอกสาร เพื่อจัดอันดับเอกสารที่เกี่ยวข้องตามคำค้นหาที่ปรากฏแต่ละฉบับโดยไม่คำนึงถึงความใกล้ชิดกัน ซึ่งสอดคล้องกับการวิจัยของ รวิสุดา และนิเวศ [32] ได้ทำการการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน พบว่าเมื่อแทนค่าดัชนีด้วยค่า TFIDF และลดคุณลักษณะด้วย ค่าไครแสควร์ (Chi-Square) พบว่า ซัพพอร์ทเวกเตอร์แมชชีน ให้ประสิทธิภาพในการจำแนกดีที่สุด

เมื่อลดมิติข้อมูลลงด้วยการคัดเลือกคุณลักษณะด้วย Chi-Square คงเหลือคำที่เป็นคุณลักษณะจำนวน 83 คำ แล้วจึงทำการคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 การคัดเลือกคุณลักษณะเหมาะสมกับเทคนิคการจำแนกมากที่สุด พบว่า ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F-measure) ผลปรากฏว่า การคัดเลือกคุณลักษณะด้วย TFIDF และ BM25 ที่เทคนิค KNN และ MLP ร้อยละ 99.20

5.3 ข้อเสนอแนะ

การศึกษาการนำเทคนิคเหมืองข้อมูลเพื่อการกระบวนการคัดเลือกคุณลักษณะด้วยการให้น้ำหนักคำ เพื่อใช้ในการสร้างแบบจำลองที่ได้ประสิทธิภาพในการจำแนกความคิดเห็นของคนไทยต่อโรคโควิด 19 จึงควรให้มีการเพิ่มจำนวนปริมาณข้อมูลให้ครอบคลุมแก่การวิเคราะห์ข้อมูล และศึกษาเทคนิคการทำเหมืองข้อมูลด้วย เทคนิคอื่น ๆ เพื่อหาความสัมพันธ์ภายในชุดข้อมูล นอกจากนี้ควรศึกษาการคัดเลือกคุณลักษณะด้วยรูปแบบวิธีอื่น ๆ รวมทั้งศึกษาการปรับตั้งค่าพารามิเตอร์การคัดเลือกคุณลักษณะ และเทคนิคต่าง ๆ การทำเหมืองข้อมูล โปรแกรม WAKA และภาษา Python หรือภาษาคอมพิวเตอร์อื่น ๆ เพื่อให้เกิดความเหมาะสม มากที่สุด

การนำไปประยุกต์ใช้เป็นแนวทางพัฒนาระบบสารสนเทศยุคใหม่ เพื่อให้ตอบสนองต่อความต้องการ โดยศึกษาจากโครสร้างรูปแบบและวิธีดำเนินการเตรียมข้อมูลให้รองรับกับการทำเหมืองข้อมูลที่นำไปใช้วิเคราะห์การทำเหมืองข้อมูลด้วยรูปแบบอื่น ๆ และเพื่อเป็นข้อมูลเสนอแนวทางเพื่อให้เห็นภาพรวมของการนำเทคนิคเหมืองข้อมูล และสามารถนำข้อมูลที่จัดเก็บในตารางฐานข้อมูลใช้สำหรับประกอบการวางแผน ศึกษาแนวโน้ม และบริการจัดการเป็นรูปแบบเพื่อประชาสัมพันธ์ในเชิงสื่อสารที่ให้ความรู้ความเข้าใจตรงกันกับเป้าหมายเพื่อการศึกษาต่อไป

สำหรับงานวิจัยครั้งต่อไป แนะนำให้ปรับพารามิเตอร์ ของ BM25 เพื่อให้ได้ค่าน้ำหนักที่ดีขึ้นและนอกเหนือจาก Bag of Word Chi-Square TFIDF และBM25 จึงแนะนำเวกเตอร์ (Vector) การวัดค่าน้ำหนักเพิ่มเติม คือ BERT

บรรณานุกรม

บรรณานุกรม

[1] W. Director-General's, "WHO Director-General’s remarks at the media briefing on 2019-nCoV on 11 February 2020," 15 May 2020, 2020 11 February 2020.

[2] S. Phosaard. and P. Posawang., "Classification for Bus Driver’s Behaviors Using Text Extraction and Machine Learning Technique," *ITD KMUTNB,* vol. Vol. 15 No. 1 (2019): January - June, 2019.

[3] U. o. I. a. C. Bing Liu, *Sentiment Analysis and Opinion Mining*. 2012.

[4] M. R. Hasan, M. Maliha, and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," in 2019 *International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 2019, pp. 1-4.

[5] A. Guo and T. Yang, "Research and improvement of feature words weight based on TFIDF algorithm," in 2016 *IEEE Information Technology, Networking, Electronic and Automation Control Conference*, 2016, pp. 415-419.

[6] A. I. Kadhim, "Term Weighting for Feature Extraction on Twitter: A Comparison Between BM25 and TF-IDF," in 2019 *International Conference on Advanced Science and Engineering (ICOASE)*, 2019, pp. 124-128.

[7] Nurhayati, A. E. Putra, L. K. Wardhani, and Busman, "Chi-Square Feature Selection Effect On Naive Bayes Classifier Algorithm Performance For Sentiment Analysis Document," in 2019 *7th International Conference on Cyber and IT Service Management (CITSM)*, 2019, vol. 7, pp. 1-7.

[8] A. A. Supianto, A. J. Dwitama, and M. Hafis, "Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University," in 2018 *International Conference on Sustainable Information Engineering and Technology (SIET)*, 2018, pp. 308-311.

[9] Y. Chen, Y. Lv, X. Wang, and F. Wang, "A convolutional neural network for traffic information sensing from social media text," in 2017 *IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1-6.

[10] W. Suksangaram, W. Hemtong, and S. Klamsakul, "Predicting learning organization factors that affect performance by data mining techniques," in

2018 *International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1-4.

[11] X. Xu, D. Cao, Y. Zhou, and J. Gao, "Application of neural network algorithm in fault diagnosis of mechanical intelligence," *Mechanical Systems and Signal Processing,* vol. 141, p. 106625, 2020/07/01/ 2020.

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

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

[14] พัชรนิกานต์ พงษ์ธนู, "วิเคราะห์ความพึงพอใจของลูกค้าจากข้อความคำแนะนำโดยการทำเหมืองความคิดเห็น," โครงการประชุมวิชาการนานาชาติ *Knowledge and Smart Technologies,* no. 1, pp. 53-60 2012.

[15] T.-Y. Liu, "Learning to Rank for Information Retrieval," *Foundations and Trends® in Information Retrieval,* vol. 3, no. 3, pp. 225-331, 2009.

[16] A. Sciandra, "COVID-19 Outbreak through Tweeters' Words: Monitoring Italian Social Media Communication about COVID-19 with Text Mining and Word Embeddings," in 2020 *IEEE Symposium on Computers and Communications (ISCC)*, 2020, pp. 1-6.

[17] S. Sharma and A. Sharma, "Twitter Sentiment Analysis During Unlock Period of COVID-19," in 2020 *Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, 2020, pp. 221-224.

[18] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach," *IEEE Journal of Biomedical and Health Informatics,* vol. 24, no. 10, pp. 2733-2742, 2020.

[19] H. Ito and B. Chakraborty, "Social Media Mining with Dynamic Clustering: A Case Study by COVID-19 Tweets," in 2020 11*th International Conference on Awareness Science and Technology (iCAST)*, 2020, pp. 1-6.

[20] Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, "A Chi-Square Statistics Based Feature Selection Method in Text Classification," in 2018 *IEEE 9th International*

Conference on Software Engineering and Service Science (ICSESS), 2018, pp. 160-163.

[21] A. W. Haryanto, E. K. Mawardi, and Muljono, "Influence of Word Normalization and Chi-Squared Feature Selection on Support Vector Machine (SVM) Text Classification," in 2018 *International Seminar on Application for Technology of Information and Communication*, 2018, pp. 229-233.

[22] B. Li, "Importance weighted feature selection strategy for text classification," in 2016 *International Conference on Asian Language Processing (IALP)*, 2016, pp. 344-347.

[23] F. Meng and L. Xu, "An Improved Native Bayes Classifier for Imbalanced Text Categorization Based on K-Means and Chi-Square Feature Selection," in 2018 *Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, 2018, pp. 894-898.

[24] Y. D. Setiyaningrum, A. F. Herdajanti, C. Supriyanto, and Muljono, "Classification of Twitter Contents using Chi-Square and K-Nearest Neighbour Algorithm," in 2019 *International Seminar on Application for Technology of Information and Communication (iSemantic)*, 2019, pp. 1-4.

[25] T. Khunsuk. and J. Thongkam., "Feature Selection Method for Improving Customer Reviews Classification," *RMUTI JOURNAL Science and Technology* vol. Vol. 13 No. 1 (2020):January - April 2020, 2019-10-07 2020.

[26] K. Saengthongpattana, T. Supnithi, and N. Soonthornphisaj, "Quality Classification of ASEAN Wikipedia Articles using Statistical Features," in 2018 *International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1-6.

[27] Y. Zhang and Z. Rao, "Hierarchical Attention Networks for Grid Text Classification," in 2020 *IEEE International Conference on Information Technology,Big Data and Artificial Intelligence (ICIBA)*, 2020, vol. 1, pp. 491-494.

[28] Y. Zheng, "An Exploration on Text Classification with Classical Machine Learning Algorithm," in 2019 *International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, 2019, pp. 81-85.

[29] T. Sawanglok, T. Thampairoj, and P. Songmuang, "Activity Recognition using Kinect and Comparison of Supervised Learning Models for Activity Classification," in 2018 *International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1-6.

[30] A. B. Nassif, O. Mahdi, Q. Nasir, M. A. Talib, and M. Azzeh, "Machine Learning Classifications of Coronary Artery Disease," in 2018 *International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, 2018, pp. 1-6.

[31] J. T. a. T. Khunsuk., "Translate Thai subtitles at Anonymous," *J of Ind. Tech. UBRU,* 2021.

[32] R. T. a. N. Chirawichitchai, "Thai Sentiment Analysis of Product Review Online Using Support Vector Machine," *Engineering Journal of Siam University,* vol. 18, no. 1, January-June2017 2017.

ภาคผนวก

ภาคผนวก ก
คำแสดงคุณลักษณะที่ใช้ในงานวิจัย

คำแสดงคุณลักษณะที่ใช้ในงานวิจัย

คำแสดงคุณลักษณะ ทั้งหมด 236 คำ คำแสดงคุณลักษณะเชิงลบ จำนวน 160 คำ คำแสดงคุณลักษณะเชิงบวก จำนวน 76 คำ

คำแสดงคุณลักษณะเชิงลบ จำนวน 160 คำ

ก็เพราะว่า	โดยเฉพาะ	นาน	ยังไม่คลี่คลาย
กระโดดขึ้น	โดยตรง	นานวัน	ยังอีก
กลับช้าสุด	โดยส่วนใหญ่	แน่น	ยากจน
ก่อน	ได้แตกต่างกัน	ในไม่ช้า	ยากเย็น
กักตัวแล้ว	ตรงท้าย	ในไลน์อัพแรก	ยาวนาน
การเที่ยว	ตลอดเวลา	ในวิกฤตนี้	ยิ่ง
การเที่ยวแล้ว	ต่อกัน	บ้าง	ยิ่งขึ้น
เกิน	ต้องจ่าย	เบื้องหลัง	ยิ่งทวีขึ้น
เกินไป	ต้องทำไงดี	เผยว่า	ยื่นเรื่อง
แก้ด้วย	ต่อไป	เผื่อ	แย่มาก ๆ
ใกล้กัน	ตั้งใหม่	พอสมควร	ร่วม
ใกล้ชิด	ต่างหาก	พุ่งขึ้น	รวมตัวกัน
ไกลกันมาก	ตามเดิม	พูดไม่ชัด	รู้กัน
ไกลมาก	ตามลำดับ	เพราะมีระบาด	ลงเอง
ค่อนข้าง	ติดกัน	เพิ่มเติม	ลดวันกักตัว
เครียดมาก	ติดโควิดแล้ว	มองว่า	ล่วงหน้า
จะจบเร็วขึ้น	ติดต่อกัน	มาก	ล่าช้า
จะพ่ายแพ้	เต็มไปด้วย	มากกกกก	แล้วก็เลย
จากรอบ	แตกต่างกัน	มากกว่า	แล้วด้วย
จากรอบรอผลช้า	ถ้างานหนัก	มากเกินไป	ไว้หมดแล้ว
จุดด้อย	ถ้าจะไปเที่ยว	มากขึ้น	สุด
จุดสูงสุด	ถ้าได้	มากยิ่งขึ้น	สูงขึ้น

คำแสดงคุณลักษณะเชิงลบ จำนวน 160 คำ(ต่อ)

เจอดี	ถึงชีวิต	มากหลาย	สูงถึง
แจ้งมา	ถึงยังคง	มาก่อน	สูงสุด
ช้า	ถึงเอาเปรียบ	มาแรง	เสี่ยงมากขึ้น
ช้าเร็ว	ถือโอกาส	มาเล่นกัน	เสียมาก
ชิ้น	ที่กิน	มาแล้ว	หมดแล้ว
ชุมนุมกัน	ที่ไกลมาก	มาว่า	เหลือเกิน
ใช้จ่าย	ที่คล้ายกัน	มาอาทิตย์แล้ว	ใหญ่หลวง
ใช้เวลานาน	ที่ค่อนข้างต่ำนี้	มีความหมาย	อย่างที่ทราบกัน
ใช่สมควร	ที่จัดขึ้น	มีปัญหา	อย่างมาก
ซึ่งกีฬากลางแจ้ง	ที่ดังพอๆกัน	เมื่อปีก่อน	อยู่ในชุมชน
ซึ่งนี่	ที่แตกต่างกัน	ไม่ครบ	อ้วน
ซึ่งสร้าง	ที่เลิกกัน	ไม่ต้องมาพูดกันเลย	ออกวีคละคนโหดมาก
เซ	ที่สุด	ยังไง	อ่อน
ด้วย	ที่ใหญ่ที่สุด	ยังไงดี	อันตรายมาก
ด่างพร้อย	น้อย	ยังไงดีคะ	อายุน้อยสุด
ดำเนินคดี	น้อยกว่า	ยังเจอ	อีกรอบ
ดิ	น้อยมาก	ยังมีอีก	อีกแล้ว
ดูต่าง ๆ	น้อยลง	ยังไม่แน่	เอาหน้า

คำแสดงคุณลักษณะเชิงบวก จำนวน 90 คำ

กระเป๋าหนัก	ด้วยดี	พอ	เลือกตั้ง
การป้องกัน	ดำเนินต่อไป	พอควร	สม่ำเสมอ
เข้าใจกัน	ดีมาก ๆ	พอดี	ส่วนตัวเรา
เข้าใจนะ	โดยยาตัวนี้	พอเพียง	สวยที่สุด
ครบ	โดยรวม	พอแล้ว	สูงกว่ากัน
ค่อย ๆ	ถี่ถ้วน	พูดคุย	เสมอ
ง่ายมาก ๆ	ถูกต้อง	เพียงพอ	หายกัน
จริง ๆ	ถูกสุด	มาแชร์กัน	เห็นชัด ๆ
จะได้มีภูมิคุ้มกัน	ทันที	มาแชร์	เห็นด้วย
จะพร้อมใจกัน	ที่ต้อง	มาโดยตลอด	เหมือนกัน
จะสำเร็จ	ที่ป้องกัน	มาตรการป้องกัน	เหมือนปกติ
จากข้อสังเกตแรก	ที่สูงที่สุด	มาเพื่อกัน	อบอุ่น
ใจตรงกัน	ที่ใหม่	ยืนยันพร้อม	อย่างเท่าเทียม
ใจมาก	น่ารัก	ร่วมกัน	อย่างยิ่ง
ช่วยกัน	บอกกัน	ระดับสูงสุด	อย่างรวดเร็ว
เช่นกัน	เป็นการดีเสียอีก	เรียบร้อย	อย่างหมดใจ
เช้า	พร้อมกัน	ลดลงเลย	ออกกำลังกาย
ซึ่งกันและกัน	พร้อมใจกัน	ลดหย่อน	อันยิ่งใหญ่
ด้วยกัน	พร้อมแล้ว	เล็กน้อย	อ่านแล้วชอบมาก

ประวัติผู้เขียน

ชื่อ	นาย ศราวุฒิ เกิดถาวร
วันเกิด	2 กันยายน 2530
สถานที่อยู่ปัจจุบัน	32/9 หมู่ 4 ตำบลในเมือง อำเภอเมือง จังหวัดชัยภูมิ 36000
ตำแหน่งหน้าที่การงาน	นักวิชาการโสตทัศนศึกษา
สถานที่ทำงานปัจจุบัน	2557 - ปัจจุบัน สำนักวิทยบริการและเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏชัยภูมิ
ประวัติการศึกษา	พ.ศ. 2550 อุตสาหกรรมศาสตร์บัณฑิต (อส.บ.) สาขาเทคโนโลยีอิเล็กทรอนิกส์และโทรคมนาคม คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏนครราชสีมา
ผลงานวิจัย	1.การพัฒนาบทเรียนผ่านเครือข่ายคอมพิวเตอร์วิชาการจัดการฐานข้อมูลทางการศึกษาสำหรับนักศึกษาระดับปริญญาตรีฯ 2.เว็บฝึกอบรมห้องสมุดเสมือนจริงเพื่อพัฒนาการรู้สารสนเทศ โดยใช้กรณีศึกษาเรื่องการใช้สารสนเทศในห้องสมุด มหาวิทยาลัยราชภัฏชัยภูมิ 3.การศึกษาพัฒนาเว็บแอพพลิเคชั่นต้นแบบระบบการประชาสัมพันธ์หน่วยงานมหาวิทยาลัยราชภัฏชัยภูมิ โดยใช้แองกูลาร์เจเอส