

การวิเคราะห์ทัศนคติเกี่ยวกับการจัดการภัยพิบัติในประเทศไทยบนทวิตเตอร์ Sentiment Analysis in Disaster Management using Tweets in Thailand

ณัฐพร ศรีสุวรรณ¹, ชนัญชิตา ภาควุฒิ¹, ปกป้อง ส่องเมือง^{2*}

Nuttaporn Seesuan¹, Chananchida Pakkawut¹, Pokpong Songmuang^{2*}

¹สาขาวิชาวิทยาศาสตร์ และนวัตกรรมข้อมูล วิทยาลัยสหวิทยาการ มหาวิทยาลัยธรรมศาสตร์

¹Data Science and Innovation of College of Interdisciplinary Studies of Thammasat University

²สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

²Computer Science of Faculty of Science and Technology Thammasat University

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาความต้องการและทัศนคติของคนไทยต่อการจัดการภัยพิบัติในแต่ละเหตุการณ์ของรัฐบาลไทย จากทวิตเตอร์เป็นจำนวน 3 แชนแท็ก ได้แก่ #สารเคมีรั่ว #โรงงานกิ่งแก้วไฟไหม้ และ #น้ำท่วมอุบล จึงได้นำความรู้เรื่อง Sentiment Analysis มาประยุกต์ใช้ ซึ่งข้อมูลที่ได้จะนำไปใช้มี 2 ชุดข้อมูล คือ ชุดข้อมูลทั้งหมด และชุดข้อมูลที่จะนำไปทำแบ่งกลุ่มข้อมูล (Clustering) เพื่อให้ได้ชุดข้อมูลที่มีความเฉพาะเจาะจงมากขึ้น นั่นคือ เป็นชุดข้อมูลที่สื่อถึงเฉพาะทัศนคติของผู้ใช้งาน และผู้วิจัยได้ทำการสร้างโมเดล 4 ตัว ได้แก่ Logistic Regression, Decision Tree, Random Forest และ Support Vector Machine (SVM) เพื่อเปรียบเทียบการทำนายการวิเคราะห์ความรู้สึกทั้ง 2 ชุดข้อมูลซึ่งพบว่าการทำ Clustering ข้อความก่อนจะช่วยให้ได้ผลลัพธ์ที่ดีมากยิ่งขึ้น แต่ทั้งนี้ทั้งนั้นการแบ่งกลุ่มข้อมูลทำให้ข้อมูลลดขนาดลง จึงใช้วิธี Synthetic Minority Oversampling Technique (SMOTE) เพิ่มเติมเพื่อช่วยแก้ปัญหา โดยที่โมเดลที่ดีที่สุดในทุกแชนแท็ก คือ Decision Tree และให้ค่าความแม่นยำ (Accuracy Score) ใน #สารเคมีรั่ว ถึง 93% และใน #โรงงานกิ่งแก้วไฟไหม้ และ #น้ำท่วมอุบล 91% เช่นเดียวกัน และนอกจากนี้ได้ทำการนำ Word Cloud มาใช้เพื่อศึกษาความต้องการ หรือสิ่งที่ผู้คนส่วนใหญ่กล่าวถึงทั้งในเชิงบวก และเชิงลบ ซึ่งพบว่าคำที่มีมากเจอในทุกแชนแท็ก คือ รัฐบาล และการเมือง ที่มีทั้งในเชิงบวก และเชิงลบ แต่ทว่าส่วนใหญ่มักกล่าวถึงในเชิงลบมากกว่า ซึ่งงานวิจัยนี้แสดงให้เห็นว่าการจัดการภัยพิบัติของประเทศไทยยังไม่ดีเพียงพอต่อความต้องการของคนไทยจึงมีผลให้มีการพูดถึง และให้ความสนใจในเชิงลบเป็นส่วนใหญ่

คำสำคัญ: การวิเคราะห์ความรู้สึก, ทัศนคติ, การแบ่งกลุ่มข้อมูล, ความถี่ของกลุ่มคำ

* Corresponding author : pokpongs@tu.ac.th

Abstract

The purpose of this research was to study the needs and attitudes of Thai people towards disaster management in each situation of the Thai government. The top three hashtags on Twitter are a chemical leak, The King Kaew factory fire, and Flooding in Ubon. As a result, Sentiment Analysis knowledge was applied. The data obtained will be used with two data sets: all data sets and the data set that will be used to group the data (Clustering) to obtain a more specific data set that only conveys the user's attitude. The researchers created four models: Logistic Regression, Decision Tree, Random Forest and Support Vector Machine (SVM). To compare the two dataset's sentiment analysis predictions. Clustering the messages first can help yield better results. However, the data is reduced in size as a result of segmentation. As a result, employing The Synthetic Minority Oversampling Technique (SMOTE) aids in the resolution of the problem. The best model across all hashtags is Decision Tree, which has an Accuracy Score of 93% in #Chemical Leaks and 91% in #KingkaewFire and #Flooding in Ubon. Word Cloud was used to study requirements or what the most people say, both positively and negatively. The words 'government' and 'politics' were discovered to be frequently used in every hashtag's positive and negative connotations. This study concludes that the Thai government's disaster management is insufficient to meet the needs of the Thai people because shows primarily on the negative.

Keywords: Sentiment Analysis, Attitude, Clustering, Word Cloud

1. บทนำ

ในปัจจุบันประเทศไทยต้องเผชิญกับภัยพิบัติที่เกิดขึ้นบ่อยครั้ง และรุนแรงมาก ก่อให้เกิดอันตราย และเกิดความสูญเสียทั้งชีวิต และทรัพย์สินต่อประชาชนชาวไทยเป็นอย่างมาก ซึ่งภาครัฐได้มีการจัดการในภาวะภัยพิบัติอย่างไร จะมีความรวดเร็ว และมีประสิทธิภาพเพียงพอหรือไม่ โดยสามารถเห็นผลลัพธ์ได้จากการแสดงความคิดเห็นต่อเหตุการณ์ต่างๆ ผ่านสื่อโซเชียล อย่างเช่น Twitter ที่เป็นแหล่งรวบรวมการแสดงความคิดเห็นที่เปิดกว้าง และทันต่อเหตุการณ์ ซึ่งเหมาะสมที่จะนำข้อมูลจากแหล่งนี้มาใช้ทำการวิเคราะห์การจัดการภัยพิบัติของภาครัฐในประเทศไทย และในงานวิจัยนี้ได้้นำการทำเหมืองข้อมูลมาใช้ ซึ่งเทคนิคที่ใช้ในการวิเคราะห์ คือ Sentiment Analysis และ Word Cloud เพื่อวิเคราะห์ความต้องการ และความรู้สึกของคนไทยต่อการจัดการภัยพิบัติของภาครัฐผ่านข้อความทาง Twitter ว่าเป็นในเชิงบวก เชิงลบ หรือเป็นกลาง เพื่อนำไปสู่ข้อสรุปว่า ภาครัฐของไทยมีการจัดการต่อเหตุการณ์ในภาวะภัยพิบัติที่ดีหรือไม่ และเนื่องจากข้อความใน Twitter มีการ

แสดงความคิดเห็นหลากหลายประเภท ไม่ว่าจะเป็นการให้กำลังใจผู้คนในเหตุการณ์ การบอกเล่าเรื่องราวในแต่ละสถานการณ์ หรือทัศนคติของผู้คนต่อเหตุการณ์ นั่นคือ ข้อมูลมีการกระจายไปหลากหลายทิศทาง ผู้วิจัยจึงได้นำเสนอเทคนิคการทำ Clustering มาใช้เพื่อจัดกลุ่มประเภทข้อมูล และดึงเอาข้อความเฉพาะทัศนคติของผู้ใช้ไปวิเคราะห์ความรู้สึก โดยผู้วิจัยได้ทำการเปรียบเทียบประสิทธิภาพของโมเดลระหว่างข้อความทั้งหมด และข้อความที่มีเพียงทัศนคติที่ได้จากการทำจัดกลุ่ม ซึ่งจะสร้างโมเดลอันได้แก่ Logistic Regression, Decision Tree, Random Forest และ Support Vector Machine สำหรับวัดความแม่นยำ และวัดประสิทธิภาพการทำงานเพื่อเปรียบเทียบว่าการทำ Clustering ข้อมูลก่อนจะทำให้ผลของการวิเคราะห์ข้อมูลเปลี่ยนแปลงไปในทิศทางที่ดีขึ้นหรือไม่ อย่างไร และหากใช้เทคนิคเพิ่มเติมอย่าง SMOTE เพื่อเพิ่มจำนวนข้อมูล และการทำให้ข้อมูลมีความสมดุลกันในแต่ละคลาสจะทำให้ผลการวิเคราะห์ที่ดีขึ้นหรือไม่ อย่างไร อีกทั้งผู้วิจัยจะนำ Word Cloud มาใช้เพิ่มเติมเพื่อหาข้อความที่ผู้ใช้พูดถึงมากที่สุดทั้งในเชิงบวก และเชิงลบ นอกจากนี้จะนำมาใช้ในการหาจำนวนข้อความที่ผู้ใช้พูดถึงรัฐบาลทั้งในเชิงบวก และเชิงลบอีกด้วย

เอกสารและงานวิจัยที่เกี่ยวข้อง

Sentiment Analysis

Sentiment Analysis หรือการวิเคราะห์ความรู้สึก คือ การวิเคราะห์อารมณ์จากข้อความ โดยพัฒนาพันธ์ กลินหอม (2564) กล่าวว่า Sentiment Analysis เป็นการวิเคราะห์อารมณ์และความรู้สึกจากข้อความ เพื่อบ่งบอกความรู้สึกของผู้คนที่มีต่อบางสิ่งบางอย่าง เช่น ความรู้สึกดี (Positive) หรือความรู้สึกที่ไม่ดี (Negative) และ Business & Technology (2565) ได้กล่าวว่า Sentiment Analysis เป็นสาขาหนึ่งของงานด้าน NLP (Natural Language Processing) เพื่อช่วยให้เข้าใจพฤติกรรมของคน และนอกจากนี้ ยังถือเป็นส่วนหนึ่งของการทำเหมืองข้อมูล (Text mining) หรือที่รู้จักกันในชื่อว่าการวิเคราะห์ข้อความ (Text analysis) การวิเคราะห์ข้อมูลประเภทนี้จะทำการดึงความหมายออกมาจากข้อความในหลากหลายแหล่งข้อมูล ตัวอย่างเช่น แบบสอบถาม รีวิว และข้อความโซเชียลมีเดีย หรือแม้กระทั่งบทความจากในเว็บไซต์ โดยที่คะแนนจะถูกกำหนดให้ในแต่ละข้อความตามความรู้สึกที่ได้แสดงออกผ่านทางข้อความนั้นๆ ตัวอย่างเช่น -1 สำหรับความรู้สึกเชิงลบ และ +1 สำหรับความรู้สึกที่เป็นเชิงบวก เพื่อวิเคราะห์ว่าผู้คนมีความรู้สึกในเชิงบวก หรือเชิงลบเกี่ยวกับเรื่องใดเรื่องหนึ่ง

Word Cloud

การวิเคราะห์ความรู้สึกไม่เพียงแค่วิเคราะห์อารมณ์จากข้อความ แต่ต้องวิเคราะห์ถึงความถี่ของข้อความด้วย ทางผู้วิจัยจึงได้นำ Word Cloud มาใช้ โดยชนมณีภา จันทรเอี่ยม (2563) กล่าวว่า Word Cloud คือ การนำเทคนิคเหมืองข้อมูลในรูปแบบของ Word Cloud มาใช้ในการวิเคราะห์ความถี่ของคำที่สนใจ (Keywords) และแสดงผลในรูปแบบ Visualization โดย Mindphp (2561) ได้กล่าวว่า หลักการของ Word Cloud คือ การจับกลุ่มคำโดยเรียงจากคำที่มีมากที่สุดไปน้อยที่สุดเป็นประโยชน์ในการทำรายงานข้อความ เพื่อให้มองเห็นคำที่ถูกใช้มากที่สุด

Clustering

เนื่องจากการนำข้อมูลจากสื่อโซเชียลซึ่งเป็นข้อมูลที่กระจัดกระจาย และมีหลากหลายประเภท ทางผู้วิจัยจึงได้นำเทคนิค Clustering มาใช้ในการจัดกลุ่มข้อมูลเหล่านี้ โดยเอกปริยา ไบสนิ (2563) กล่าวว่า Clustering คือ Machine Learning Model ประเภท Unsupervised ที่ไม่มี Target หรือไม่มีต้นแบบของผลลัพธ์ ซึ่งเป็นโมเดลที่นำไปใช้ในการจัดกลุ่มของข้อมูลที่ไม่เคยมีการจัดกลุ่มมาก่อน โดยจะแบ่งกลุ่มข้อมูลจากความคล้าย เช่น การจัดกลุ่มลูกค้าจากพฤติกรรมการซื้อสินค้าของลูกค้าที่มีลักษณะคล้ายกันจะเป็นกลุ่มลูกค้าประเภทเดียวกัน และสุชีพ ต้นติวุฒิพงศ์ และคณะ (2564) ได้กล่าวว่า Clustering เป็นการวิเคราะห์ข้อมูลรูปแบบหนึ่งที่ทำกรแบ่งชุดข้อมูลออกเป็นกลุ่ม หรือแหล่งเก็บสะสม (Collection) โดยนำข้อมูลที่มีคุณลักษณะเหมือนกันหรือคล้ายกันจัดไว้ในกลุ่มเดียวกัน อาศัยความเหมือน (Similarity) หรือความใกล้ชิด (Proximity) ที่คำนวณจากการวัดระยะระหว่างตำแหน่งของข้อมูลเข้าหากัน โดยใช้การวัดระยะแบบต่างๆ เช่น การวัดระยะแบบยูคลิด (Euclidean Distance) การวัดระยะแบบแมนฮัตตัน (Manhattan Distance)

การจัดกลุ่มข้อมูลแบบ K-means

ในส่วนของ การจัดกลุ่มทางผู้วิจัยได้เลือกใช้อัลกอริทึมการจัดกลุ่มข้อมูลแบบ K-means ด้วยวิธี Elbow Method โดยภัทรพล อาภาอาษา (2564) กล่าวว่า K-means คือ การวิเคราะห์กลุ่มแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) หรือการแบ่งส่วน (Partitioning) โดยใช้ทฤษฎี Euclidean Distance เป็นมาตรวัดระยะพื้นฐานที่ใช้สำหรับหาระยะทางระหว่างจุดสองจุด เป็นที่นิยมใช้ในงานประเภทต่างๆ เป็นอย่างมาก เพราะง่ายต่อความเข้าใจ และลักษณะการคำนวณที่คล้ายกับทฤษฎีบทพีทาโกรัส โดยอาศัยการคำนวณระยะทางระหว่าง Centroid กับข้อมูลที่จะนำมาจัดกลุ่ม และได้กล่าวว่า Elbow Method เป็นหนึ่งในอัลกอริทึมที่เอาไว้วัด Performance ในการทำ Clustering ของ K-Mean นั่นคือวิธีการวัดค่า Error ผลรวมของระยะห่างระหว่าง Object กับ Centroid เพื่อการเลือกค่า k ที่เหมาะสมหรือการหาค่า Optimal Cluster Number จุดที่เหมาะสมของจำนวนกลุ่ม คือ จุดที่กราฟมีลักษณะหักศอกซึ่งจุดนี้จะเป็นจุดที่ให้ค่าจำนวนกลุ่มที่ดีที่สุด

SMOTE

เนื่องจากหลังการทำ Clustering ทำให้ข้อมูลน้อยลง อีกทั้งข้อมูลมีความไม่สมดุล (Imbalanced data) อาจทำให้ผลลัพธ์มีความโน้มเอียงไปทางข้อมูลกลุ่มมาก ทางผู้วิจัยจึงได้นำเทคนิค SMOTE มาใช้เพื่อแก้ปัญหา โดยพุทธิพร ธนธรรมเมธี และ เขียวเรศ ศิริสถิตย์กุล (2562) กล่าวว่า Oversampling คือ การจัดการความไม่สมดุลของข้อมูล (Imbalanced Data) เกิดขึ้นเมื่อข้อมูลที่ใช้งานมีจำนวนข้อมูลในแต่ละคลาสแตกต่างกันมาก ทำให้ผลลัพธ์จากการจำแนกข้อมูลมีความโน้มเอียงไปทางข้อมูลกลุ่มมาก ซึ่งเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่มเป็นการสุ่มเพิ่มจำนวนข้อมูลของข้อมูลกลุ่มน้อย ให้มีปริมาณข้อมูลใกล้เคียงกับข้อมูลกลุ่มมาก นอกจากนี้ อัจฉรา แผ้วบาง และสายชล สินสมบูรณ์ทอง (2563) ได้กล่าวว่าเทคนิค SMOTE คือ การสุ่มสร้างข้อมูลจากกลุ่มส่วนน้อยตามจำนวนที่กำหนด โดยการวัดระยะห่างจากจุดข้อมูลตัวอย่างไปยังจุดข้อมูลใกล้เคียงแล้วสุ่มสร้างข้อมูลสังเคราะห์ขึ้นให้ใกล้เคียงกับกลุ่มส่วนมาก และจากการศึกษางานวิจัยของ อัจฉรา แผ้วบาง

และ สายชล สิ้นสมบุรณ์ทอง (2563) ซึ่งได้ศึกษาถึงวิธีการปรับข้อมูลที่ไม่สมดุล 4 วิธี คือ วิธีการสุ่มเกินวิธีการสุ่มเกินโดยเทคนิค SMOTE วิธีการสุ่มลด และวิธีการสุ่มผสมผสาน โดยวิธีการจำแนก 5 วิธี คือ วิธีเพื่อนบ้านใกล้สุด k ตัว วิธีโครงข่ายประสาทเทียม วิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ และวิธีลาดลงสโตแคสติก ว่าวิธีใดมีประสิทธิภาพในการจำแนกดีที่สุดใน โดยพิจารณาจากค่าความถูกต้อง ค่าความไว ค่าความจำเพาะ ค่าคลาดเคลื่อนกำลังสองเฉลี่ย และค่าคลาดเคลื่อนสัมบูรณ์เฉลี่ย พบว่าวิธีที่มีประสิทธิภาพสูงสุด คือ วิธีฐานกฎ โดยการปรับความไม่สมดุลด้วยวิธีการสุ่มเกินเทคนิค SMOTE

Machine Learning

ในการสร้างโมเดลเพื่อวิเคราะห์ความรู้สึกหรือความคิดเห็น จากการศึกษางานวิจัยที่เกี่ยวข้อง มีการสร้างแบบจำลองด้วยอัลกอริทึมหลายวิธี โดยในงานวิจัยของวิสุดา เทศเมือง และ นิเวศ จิระวิชิตชัย (2560) ได้ทำการสร้างแบบจำลองด้วยอัลกอริทึม 4 วิธี ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ นาอ็ฟเบย์ และเคเนียร์เรนเนอร์ เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ซึ่งผลการทดลองพบว่าคุณลักษณะที่ดีที่สุด คือ ซัพพอร์ตเวกเตอร์แมชชีน โดยมีค่าความถูกต้อง (Accuracy) เท่ากับ 83.38% ระดับรองลงมาเป็นนาอ็ฟเบย์, ต้นไม้ตัดสินใจ และเคเนียร์เรนเนอร์ โดยมีค่าความถูกต้อง เท่ากับ 83.31%, 79.92% และ 71.26% ตามลำดับ ส่วนในงานวิจัยของสันติ สุขเกษม (2561) ได้ทำการวิเคราะห์ความคิดเห็นโดยใช้เทคนิคการจำแนกนาอ็ฟเบย์ (Naive Bayes) แบบต้นไม้ตัดสินใจ (Decision Tree) และแบบเคเนียร์เรนเนอร์ (K-Nearest Neighbor) ซึ่งผลการวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน การวิเคราะห์จำแนกแบบนาอ็ฟเบย์ ได้ค่าความถูกต้อง 99.66% การจำแนกแบบต้นไม้ตัดสินใจ ได้ค่าความถูกต้อง 65.49% และการจำแนกแบบเคเนียร์เรนเนอร์ ได้ค่าความถูกต้อง 70.21% ส่วนงานวิจัยของ กิตติศักดิ์ กิตติฐานุสรณ์ (2563) ได้ใช้อัลกอริทึมการจัดแบ่งประเภทของเอกสาร และการเรียนรู้ของเครื่อง เพื่อจำแนกประเภทข่าว กระบวนการการจำแนกประเภทจะสำรวจเทคนิค Bag of Word และ Term Frequency Inverse Document Frequency (TFIDF) ด้วย 5 การเรียนรู้ คือ Multinomial Naive Bayes, Complement Naive Bayes, Logistic Regression, LinearSVC และ Random Forest และแก้ไขปัญหาที่ไม่สมดุลโดยใช้อัลกอริทึมการสุ่มตัวอย่าง 3 วิธี คือ Under Sampling, Synthetic Minority Oversampling Technique และ Adaptive Synthetic Sampling ผลลัพธ์จากการทดลองพบว่า Logistic Regression ที่ใช้เทคนิค Bag of Word และ SMOTE มีประสิทธิภาพสูงที่สุดในการจำแนกประเภทข่าว นอกจากนี้ในงานวิจัยของ วันสวรรณ มีประเสริฐ และ เอกรัฐ รัชฎากัญจน (2564) ได้สร้างโมเดลสำหรับการแยกหัวข้อประเด็นสำคัญของข้อความด้วยวิธี Support Vector Machine (SVM), Random Forest และ Logistic Regression และใช้ Model ในการแยกความรู้สึกแต่ละประเด็นสำคัญด้วยวิธี WangchanBERTa, SVM และ Logistics Regression ซึ่งพบว่ากรวิเคราะห์ข้อมูลความคิดเห็นของลูกค้าด้วยวิธี Random Forest สามารถแยกประเด็นสำคัญได้แม่นยำที่สุด และวิธี WangchanBERTa สามารถแยกความรู้สึกแต่ประเด็นสำคัญในชุดข้อมูลภาษาไทยได้แม่นยำที่สุด และในงานวิจัยของพิศิษฐ์ บวรเลิศสุธี และ วรภัทร ไพรีเกรง (2565) ได้ทำการสร้างตัวแบบสำหรับการวิเคราะห์ระดับความรู้สึกของการแสดงความคิดเห็น โดยใช้ 4 เทคนิค ได้แก่

K-Nearest Neighbors, Random Forest, Logistic Regression และ Support Vector Machines (SVM) ผลการทดลองสร้างตัวแบบการวิเคราะห์ระดับความรู้สึกทางอารมณ์ 3 ระดับ คือ ความรู้สึกเชิงบวก เป็นกลาง (Neutral) และเชิงลบ โดยการสร้างตัวแบบใช้เทคนิคการแยกประเภทแบบ Support Vector Machines (SVM) ร่วมกับเทคนิคการตัดค่าแบบ Longest ให้ค่าความแม่นยำดีที่สุดเมื่อเปรียบเทียบกับเทคนิคอื่นๆ

Scatter Chart

เนื่องจากผู้วิจัยต้องการดูการกระจายตัวของข้อมูลเพื่อใช้ในการวิเคราะห์ข้อมูลต่อไป ทางผู้วิจัยจึงได้นำ Scatter Chart มาใช้ โดย Automated (2564) ได้กล่าวว่า Scatter Chart คือ กราฟที่นำมาเพื่อวิเคราะห์ข้อมูลความสัมพันธ์ของข้อมูล โดยเปรียบเทียบระหว่างแกน X และ แกน Y ซึ่งในทางธุรกิจนิยมนำมาใช้ดูยอดขาย (Sale Amount), %Achieve, %YoY ของสินค้า (Products) หรือดูกลุ่มลูกค้า (Segment) และอื่นๆ ตามที่ผู้วิจัยต้องการวิเคราะห์

Confusion Matrix

ในการประเมินประสิทธิภาพของการทำนายของโมเดล ทางผู้วิจัยได้นำ Confusion Matrix มาใช้ โดย PL (2561) ได้กล่าวว่า Confusion Matrix เป็นอีกหนึ่งการประเมินโมเดล โดยการนำเอาองค์ความรู้ด้านเมทริกซ์มาประกอบกันเป็นการทำ Classification Evaluate โดย Confusion Matrix สามารถอธิบายค่า TP, TN, FN และ FP ของคำตอบได้ เพื่อนำมาตรวจสอบโมเดลในการทำนายมีถูกผิดมากน้อยเพียงใด

2. วัตถุประสงค์การวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาความต้องการและทัศนคติของคนไทยต่อการจัดการภัยพิบัติในแต่ละเหตุการณ์ของรัฐบาลไทย จากทวิตเตอร์เป็นจำนวน 3 แชนแท็ก ได้แก่ สารเคมีรั่ว โรงงานกิ่งแก้วไฟไหม้ และน้ำท่วมอุบล ซึ่งจะนำความรู้เรื่อง Sentiment Analysis มาประยุกต์ใช้ โดยข้อมูลที่ได้นำไปใช้เป็น 3 ชุดข้อมูล คือ ชุดข้อมูลทั้งหมด, ชุดข้อมูลที่จะนำไปทำแบ่งกลุ่มข้อมูล และชุดข้อมูลที่ทำแบ่งกลุ่มข้อมูล พร้อมทั้งใช้วิธี Synthetic Minority Oversampling Technique เพิ่มเติมเพื่อช่วยแก้ปัญหาความไม่สมดุลของข้อมูล ซึ่งการนำเทคนิคการแบ่งกลุ่มข้อมูลมาใช้เพื่อให้ได้ชุดข้อมูลที่มีความเฉพาะเจาะจงมากขึ้น นั่นคือ เป็นชุดข้อมูลที่สื่อถึงเฉพาะทัศนคติของผู้ใช้งาน และนอกจากนี้ได้ทำการนำ Word Cloud มาใช้เพิ่มเติมเพื่อศึกษาความต้องการ หรือสิ่งที่ผู้คนส่วนใหญ่กล่าวถึงคือ “รัฐบาล” และ “การเมือง” ทั้งในเชิงบวก และเชิงลบของแต่ละแชนแท็กเป็นจำนวนเท่าไร และในเชิงใดมากกว่ากัน เพื่อสรุปการวิเคราะห์ทัศนคติของผู้ใช้งานบนทวิตเตอร์ต่อการจัดการภัยพิบัติในประเทศไทย

3. วิธีดำเนินการวิจัย

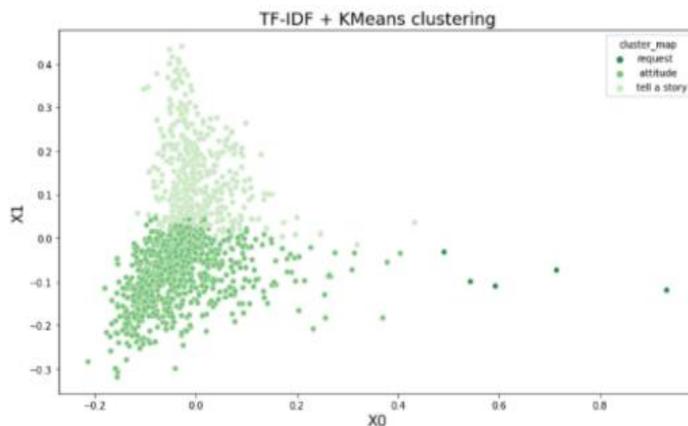
3.1 กระบวนการ

การทำ Sentiment Analysis จะได้ผลลัพธ์เป็นเชิงบวก เชิงลบ และเป็นกลาง ซึ่งการทำ Sentiment Analysis กับข้อมูลจากโซเชียลที่มีลักษณะข้อมูลไม่เป็นหมวดหมู่ และหลากหลาย จะทำให้ผลลัพธ์การทำ Sentiment Analysis มีค่าเป็นกลางมากเกินไป เนื่องจากบางข้อความไม่ได้มีเนื้อหาเกี่ยวกับการแสดงทัศนคติ

1	text	Polarity	topics
2	ใหม่อีกแล้วเธอ ครันเยอะมาก เอาจริง	negative	['สุขภาพกาย']
3	จีโนต้องมีประโยชน์แล้วนะ กิ่งแก้วไฟ	neutral	['รถยนต์']
4	รบกวนผู้รู้ทั้งหลายนะคะ ช่วยมาอภิปรา	negative	['บ้าน']
5	ตอนไฟใหม่ที่กิ่งแก้วรถฉีดน้ำของรัฐบ	negative	['รถยนต์']
6	ตอน กิ่งแก้วไฟใหม่ รถพวกนี้ไปไหนกัน	negative	['ถนนกิ่งแก้ว']
7	ดูอุปกรณ์ของนักผจญเพลิงที่ กิ่งแก้วไ	negative	['การเมือง']
8	รถน้ำดับเพลิงมีไว้เอามาฉีดมีอบ ตอน ก	negative	['อุโมงรถ']
9	วัน กิ่งแก้วไฟใหม่ บอก จีโนเสียหายต วั	negative	['รถยนต์']

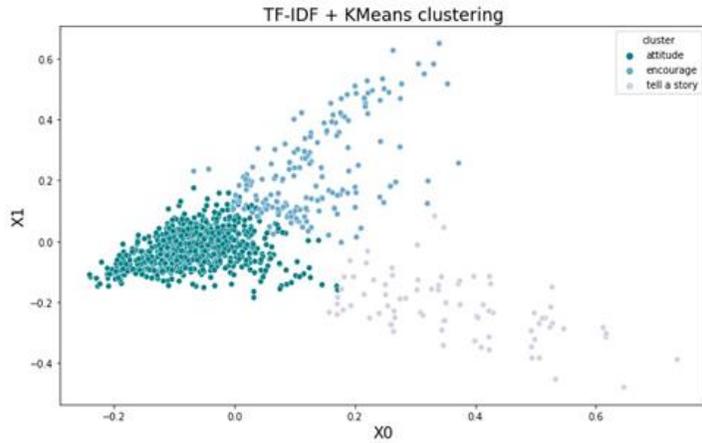
ภาพที่ 2 แสดงตัวอย่างข้อมูลหลังการทำ Label ของ #โรงงานกิ่งแก้วไฟไหม้

3.2.4 การแบ่งกลุ่มข้อมูล จะทำการแบ่งกลุ่มข้อมูลด้วยอัลกอริทึม K-means ซึ่งแต่ละแฮชแท็ก ได้ผลการทำ Elbow Method เท่ากับ 3 จึงทำการแบ่งกลุ่มออกเป็น 3 กลุ่ม ดังแสดงในภาพที่ 1, 2 และ 3 โดยจะใช้กราฟ Scatter เพื่อแสดงให้เห็นถึงการกระจายตัวของข้อมูล อีกทั้งแสดงถึงการเป็นกลุ่มข้อมูลหรือประเภทข้อมูลเดียวกันผ่านสี



ภาพที่ 3 แสดงกราฟ Scatter ของ #สารเคมีรั่ว

- จากภาพที่ 3: สีเขียวเข้ม ● แทนกลุ่มข้อมูลการเรียกร้อง หรือความต้องการให้มีการดำเนินการเร่งด่วน (Request)
- สีเขียวกลาง ● แทนกลุ่มข้อมูลที่สนคดี
- สีเขียวอ่อน ● แทนกลุ่มข้อมูลการบอกเล่าเรื่องราวหรือสถานการณ์ที่เกิดขึ้น (Tell a story)

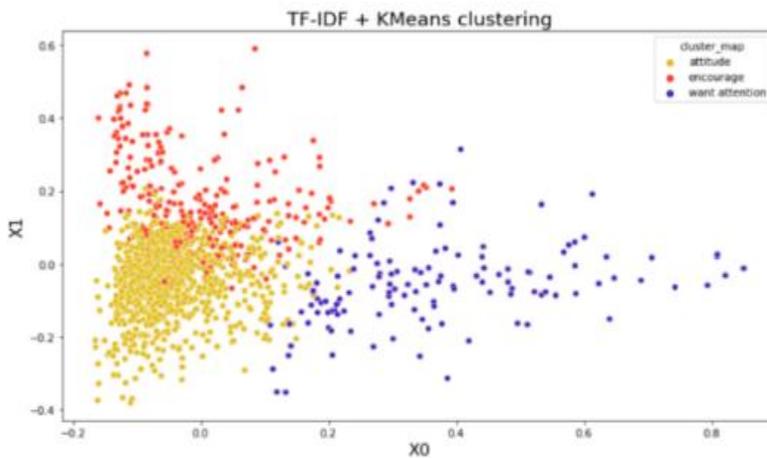


ภาพที่ 4 แสดงกราฟ Scatter ของ #โรงงานกึ่งแก้วไฟใหม่

จากภาพที่ 4: สีน้ำเงิน ● แทนกลุ่มข้อมูลทัศนคติ

สีฟ้า ● แทนกลุ่มข้อมูลส่งกำลังใจ (Encourage)

สีขาว ● แทนกลุ่มข้อมูลการบอกเล่าเรื่องราวหรือสถานการณ์ที่เกิดขึ้น



ภาพที่ 5 แสดงกราฟ Scatter ของ #น้ำท่วมอุบล

จากภาพที่ 5: สีเหลือง ● แทนกลุ่มข้อมูลทัศนคติ

สีแดง ● แทนกลุ่มข้อมูลส่งกำลังใจ

สีน้ำเงิน ● แทนกลุ่มข้อมูลต้องการการสนใจหรือต้องการให้เรื่องราวของ
แฮชแท็กเป็นที่น่าสนใจในสังคม (Want attention)

การวิเคราะห์ความรู้สึกจากข้อมูลทั้งหมดทำให้การทำนายมีความแม่นยำไม่มากพอจึงนำเทคนิค Clustering มาใช้เพื่อจัดกลุ่มข้อมูลของแต่ละแฮชแท็ก ซึ่งจากการจัดกลุ่มจะเห็นว่าในแต่ละแฮชแท็กจะมีกลุ่มข้อมูลที่แสดงถึงทัศนคติผู้ใช้งานทวิตเตอร์ ปะปนอยู่ จึงนำเฉพาะกลุ่มข้อมูลนี้ไปใช้งานต่อไป

3.2.5 การปรับเพิ่มข้อมูลด้วยวิธีสุ่ม หรือเรียกว่า Oversampling เนื่องจากหลังการทำ Clustering ทำให้ข้อมูลน้อยลง อีกทั้งข้อมูลมีความไม่สมดุล อาจทำให้ผลลัพธ์มีความโน้มเอียงไปทางข้อมูลกลุ่มมาก จึงนำเทคนิค SMOTE มาใช้เพื่อแก้ปัญหา

3.2.6 การสร้างโมเดล จะทำการสร้างโมเดล 3 ชุด นั่นคือ สร้างจากข้อมูลทั้งหมด 1 ชุด, สร้างจากข้อมูลที่ได้จากการทำ Clustering แล้วเหลือข้อมูลเฉพาะทัศนคติ (Attitude Dataset) 1 ชุด และสร้างจากข้อมูลชุด Attitude Dataset พร้อมใช้เทคนิค SMOTE ในการแก้ปัญหา 1 ชุด

3.2.7 การเปรียบเทียบโมเดล ทำการเปรียบเทียบโมเดลทั้ง 3 ชุดโดยใช้ค่าความแม่นยำ (Accuracy Score) และผลการประเมินผลลัพธ์การทำนาย (Performance Matrix)

3.2.8 การสร้าง Word Cloud จากข้อมูลทั้งหมด 1 ชุด และจากการทำ Clustering 1 ชุดโดยข้อมูลชุดนี้จะนำมาแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มที่เป็นเชิงบวก 1 กลุ่ม และเชิงลบ 1 กลุ่ม (ได้จาก Label ของ Sentiment)

3.2.9 การเปรียบเทียบผลลัพธ์จาก Word Cloud จะทำการเปรียบเทียบ Word Cloud ที่ได้จากข้อมูลชุด Clustering ทั้งในเชิงบวก และเชิงลบ เพื่อดูสิ่งที่คุณพุดถึงมากที่สุด และมีการพุดถึงรัฐบาลมากเพียงใด และมีการพุดถึงในเชิงใดมากกว่า

3.3 ประชากรและกลุ่มตัวอย่าง

ตารางที่ 1 แสดงจำนวนข้อมูลทั้งหมด จำนวนข้อมูลหลังการทำ Clustering และจำนวนข้อมูลหลังการทำ SMOTE กับข้อมูล Clustering ในแต่ละชั่ววอรณณ์ของแต่ละแฮชแท็ก

	Non-clustering				Clustering				Clustering + SMOTE			
	Negative	Neutral	Positive	Total	Negative	Neutral	Positive	Total	Negative	Neutral	Positive	Total
#สารเคมีรั่ว	836	373	113	1328	498	249	82	829	394	394	394	1182
#โรงงานกึ่ง แก้วไฟไหม้	993	79	90	1162	768	49	71	888	540	540	540	1620
#น้ำท่วมอุบล	1214	97	85	1396	898	66	58	1022	629	629	629	1887

4. ผลการวิจัย

4.1 สารเคมีรั่ว

4.1.1 ผลลัพธ์การทำ Word Cloud จากชุดข้อมูลทั้งหมด



ภาพที่ 6 แสดงผลการทำ Word Cloud จากชุดข้อมูลทั้งหมดของ #สารเคมีรั่ว

พบว่า สิ่งที่ผู้คนพูดถึงมากที่สุดอันดับต้นๆ คือ สถานที่เกิดเหตุการณ์ นั่นคือ จังหวัดนครปฐม รองลงมา คือ การเรียกร้องให้หน่วยงานของรัฐจัดทำระบบแจ้งเตือนภัยให้กับประชาชน

4.1.2 ผลค่าความแม่นยำของการทดลองการเปรียบเทียบโมเดลระหว่างชุดข้อมูลทั้งหมด, ชุดข้อมูล Clustering และชุดข้อมูล Clustering + SMOTE

ตารางที่ 2 แสดงผลค่าความแม่นยำระหว่างชุดข้อมูลทั้งหมด, Clustering และ Clustering + SMOTE ของ #สารเคมีรั่ว

Model	Non-clustering	Clustering	Clustering + SMOTE
Logistic Regression	72%	72%	86%
Decision Tree	60%	64%	93%
Random Forest	71%	68%	93%
SVM	69%	71%	84%

4.1.3 ผลการประเมินผลลัพธ์การทำนายของการทดลองการเปรียบเทียบโมเดลระหว่างชุดข้อมูลทั้งหมด, ชุดข้อมูล Clustering และชุดข้อมูล Clustering + SMOTE

ตารางที่ 3 แสดงการประเมินผลลัพธ์การทำนาย (Confusion Matrix) ของชุดข้อมูลทั้งหมด ของ #สารเคมีรั่ว

- Non-clustering

Logistic Regression			Decision Tree		
159	25	10	129	25	8
25	27	7	39	24	7
3	3	5	19	6	7
Random Forest			SVM		
162	29	14	154	26	13
25	26	6	33	29	9
0	0	2	0	0	0

ตารางที่ 4 แสดงการประเมินผลลัพธ์การทำนายของชุดข้อมูลแบ่งกลุ่ม ของ #สารเคมีรั่ว

- Clustering

Logistic Regression			Decision Tree		
97	24	10	77	15	8
7	22	5	16	26	4
0	0	1	11	5	4
Random Forest			SVM		
91	22	11	98	25	12
13	23	5	6	21	4
0	1	0	0	0	0

ตารางที่ 5 แสดงการประเมินผลลัพธ์การทำนาย ของชุดข้อมูลแบ่งกลุ่มและทำ SMOTE ของ #สารเคมีรั่ว

- Clustering + SMOTE

Logistic Regression			Decision Tree		
375	6	5	393	0	0
11	278	15	1	329	13
8	110	374	0	65	381
Random Forest			SVM		
394	0	0	368	11	10
13	23	5	6	21	4
0	1	0	0	0	0

4.1.4 ผลลัพธ์การทำ Word Cloud จากชุดข้อมูล Clustering โดยแยกแต่ละหัวข้ออารมณ์ (Polarity)



ภาพที่ 7 แสดงผลการทำ Word Cloud จากชุดข้อมูล Clustering โดยแสดงผลเป็นหัวข้ออารมณ์เชิงบวก

พบว่า ผู้คนพูดถึง “เคมี”, “บ้าน” และ “การเมือง” มากที่สุด 3 อันดับแรก ตามลำดับ และ “รัฐบาล” อยู่ในอันดับที่ 33 เป็นจำนวน 1 ข้อความ และ “การเมือง” อันดับที่ 3 เป็นจำนวน 4 ข้อความ



ภาพที่ 8 แสดงผลการทำ Word Cloud จากชุดข้อมูล Clustering โดยแสดงผลเป็นหัวข้ออารมณ์เชิงลบ

พบว่า ผู้คนพูดถึง “เคมี”, “สุขภาพกาย” และ “บ้าน” มากที่สุด 3 อันดับแรก ตามลำดับ และ “รัฐบาล” อยู่ในอันดับที่ 21 เป็นจำนวน 7 ข้อความ และ “การเมือง” อันดับที่ 4 เป็นจำนวน 16 ข้อความ

จะเห็นได้ว่า #สารเคมีรั่ว มีจำนวน Keyword “รัฐบาล” และ “การเมือง” ใน Word Cloud เชิงลบมากกว่าเชิงบวก ดังนั้นจึงสรุปได้ว่าการพูดถึงรัฐบาล และการเมืองไปในเชิงลบมากกว่าเชิงบวกในสถานการณ์ภัยพิบัติสารเคมีรั่ว

ตารางที่ 7 แสดงการประเมินผลสัมฤทธิ์การทำนายของชุดข้อมูลทั้งหมด
ของ #โรงงานกึ่งแก้วไฟไหม้

- Non-clustering

Logistic Regression			Decision Tree		
297	16	23	267	11	18
2	4	3	20	6	5
0	0	4	12	3	7
Random Forest			SVM		
297	15	26	283	13	20
2	5	3	9	6	6
0	0	1	7	1	4

ตารางที่ 8 แสดงการประเมินผลสัมฤทธิ์การทำนายของชุดข้อมูลแบ่งกลุ่ม
ของ #โรงงานกึ่งแก้วไฟไหม้

- Clustering

Logistic Regression			Decision Tree		
227	19	17	213	14	11
0	0	0	5	4	4
1	0	3	10	1	5
Random Forest			SVM		
228	19	20	223	17	15
0	0	0	3	2	2
0	0	0	2	0	3

ตารางที่ 9 แสดงการประเมินผลสัมฤทธิ์การทำนายของชุดข้อมูลแบ่งกลุ่มและทำ SMOTE
ของ #โรงงานกึ่งแก้วไฟไหม้

- Clustering + SMOTE

Logistic Regression			Decision Tree		
563	3	1	571	0	0
8	544	152	0	557	140
0	24	148	0	14	431
Random Forest			SVM		
571	0	0	570	0	0
0	566	163	0	538	0
0	5	408	1	33	147

4.3.3 ผลการประเมินผลลัพท์การทำนายการทดลองการเปรียบเทียบโมเดลระหว่างชุดข้อมูลทั้งหมด, ชุดข้อมูล Clustering และชุดข้อมูล Clustering + SMOTE

ตารางที่ 11 แสดงการประเมินผลลัพท์การทำนายของชุดข้อมูลทั้งหมด ของ #น้ำท่วมอุบล

- Non-clustering

Logistic Regression			Decision Tree		
363	24	17	342	20	10
6	2	2	22	6	4
5	0	0	10	2	5
Random Forest			SVM		
373	24	19	341	22	17
1	2	0	16	4	1
0	0	0	17	0	1

ตารางที่ 12 แสดงการประเมินผลลัพท์การทำนายของชุดข้อมูลแบ่งกลุ่ม ของ #น้ำท่วมอุบล

- Clustering

Logistic Regression			Decision Tree		
263	20	14	248	13	11
3	0	2	12	5	4
3	0	2	9	2	3
Random Forest			SVM		
269	20	18	242	17	13
0	0	0	16	3	2
0	0	0	11	0	3

ตารางที่ 13 แสดงการประเมินผลลัพท์การทำนาย ของชุดข้อมูลแบ่งกลุ่มและทำ SMOTE ของ #น้ำท่วมอุบล

- Clustering + SMOTE

Logistic Regression			Decision Tree		
616	4	0	629	0	0
8	601	187	0	628	163
5	24	442	0	1	466
Random Forest			SVM		
629	0	0	626	2	0
0	611	164	2	621	194
0	18	465	1	6	435

4.3.4 ผลลัพธ์การทำ Word Cloud จากชุดข้อมูล Clustering โดยแยกแต่ละชั่ววாரมณั์



ภาพที่ 13 แสดงผลการทำ Word Cloud จากชุดข้อมูล Clustering โดยแสดงผลเป็นชั่ววாரมณั์เขียว

พบว่า ผู้คนพูดถึง “การเมือง”, “จังหวัดอุบลราชธานี” และ “รัฐบาล” มากที่สุด 3 อันดับแรก ตามลำดับ และ “รัฐบาล” อยู่ในอันดับที่ 3 เป็นจำนวน 6 ข้อความ และ “การเมือง” อันดับที่ 1 เป็นจำนวน 8 ข้อความ



ภาพที่ 14 แสดงผลการทำ Word Cloud จากชุดข้อมูล Clustering โดยแสดงผลเป็นชั่ววாரมณั์เขียว

พบว่า ผู้คนพูดถึง “จังหวัดอุบลราชธานี”, “รัฐบาล” และ “การเมือง” มากที่สุด 3 อันดับแรก ตามลำดับ และ “รัฐบาล” อยู่ในอันดับที่ 2 เป็นจำนวน 117 ข้อความ และ “การเมือง” อันดับที่ 3 เป็นจำนวน 98 ข้อความ

จะเห็นได้ว่า #น้ำท่วมอุบล มีจำนวน Keyword “รัฐบาล” และ “การเมือง” ใน Word Cloud เขียว มากกว่าเชิงบวก ดังนั้นจึงสรุปได้ว่าการพูดถึงรัฐบาล และการเมืองไปในเชิงลบมากกว่าเชิงบวกในสถานการณ์ภัยพิบัติน้ำท่วมอุบล

5. อภิปรายผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย และอภิปรายผล

จากผลการทดลองเปรียบเทียบโมเดลระหว่างชุดข้อมูลทั้งหมดกับชุดข้อมูล Clustering และชุดข้อมูล Clustering + SMOTE สรุปได้ว่า การทำ Clustering ข้อมูลเพื่อเลือกกลุ่มข้อมูลที่เฉพาะเจาะจงมากขึ้น แล้วนำไปทำ SMOTE เพิ่มเติม จะทำให้ได้ค่าความแม่นยำของโมเดลสูงที่สุด และมีผลการประเมินผลลัพธ์การทำนายที่ดีที่สุด นอกจากนี้ การทำ Word Cloud จากชุดข้อมูลทั้งหมดจะเห็นเพียงภาพรวม และ Keyword หลักจะเป็นเพียงสถานที่ที่เกิดภัยพิบัติ การนำเทคนิค Clustering มาใช้ให้เหลือเพียงข้อความทัศนคติ และทำ Tag Suggestion ให้กับข้อความจะทำให้ Word Cloud เห็น Keyword ได้เฉพาะเจาะจงมากขึ้น ทั้งในเชิงบวก และเชิงลบ อีกทั้งสามารถบอกได้ว่าการจัดการภัยพิบัติของรัฐในแต่ละเหตุการณ์เป็นไปในเชิงบวกหรือลบ มากกว่ากันอีกด้วย

5.2 ข้อเสนอแนะ

การแบ่งกลุ่มข้อมูลที่นำเสนอในงานวิจัยนี้ มีจุดเด่น คือ ทำให้ได้กลุ่มข้อมูลที่เฉพาะเจาะจงมากขึ้น แต่เนื่องจากชุดข้อมูลที่มีจำนวนน้อย การแบ่งกลุ่มจึงยังลดขนาดของข้อมูล หากจะนำเทคนิคนี้ไปประยุกต์ใช้งานต่อ ควรจะใช้กับชุดข้อมูลที่มีปริมาณมาก เพื่อที่จะได้รับผลลัพธ์ และประสิทธิภาพที่ดีมากยิ่งขึ้น

5.3 ปัญหา และอุปสรรค

ขั้นตอนการทำ Preprocessing เป็นขั้นตอนที่สำคัญมากที่สุด เนื่องจากคุณภาพของข้อมูลจะส่งผลต่อประสิทธิภาพของโมเดล แต่ในข้อความภาษาไทยที่ได้จากสื่อสังคมออนไลน์มักมีการเขียนผิด หรือเป็นคำใหม่ที่กำลังสร้าง ซึ่งไม่ได้อยู่ในพจนานุกรม ทำให้การสื่อความหมายที่ระบุอย่างแน่ชัดทำได้ค่อนข้างยาก ซึ่งปัจจัยเหล่านี้จะส่งผลต่อการวิเคราะห์ความรู้สึกที่ไม่ดีนัก

5.4 ประโยชน์

จากการรวบรวมข้อมูลการแสดงความคิดเห็นของผู้คนบนทวิตเตอร์ และจัดทำการวิเคราะห์ข้อมูลด้วย Machine Learning ทำให้เห็นถึงข้อมูลเชิงลึกที่สะท้อนถึงความต้องการ และความรู้สึกของประชาชนต่อการรับมือของรัฐบาลในการจัดการปัญหาที่ไม่คาดคิด ซึ่งส่งผลกระทบต่อชีวิตและทรัพย์สิน เพราะการเห็นถึงปัญหาจึงจะสามารถแก้ไขปัญหาได้อย่างถูกต้อง และรวดเร็ว ทำให้ผลกระทบที่จะส่งผลถึงประชาชนรุนแรงน้อยลง อีกทั้งผลกระทบที่เกิดขึ้นไม่ใช่เพียงแค่ด้านทรัพย์สิน แต่จิตใจของประชาชนเป็นสิ่งที่รัฐบาลต้องคำนึงถึงเป็นอันดับแรก และจัดการเยียวยาอย่างรวดเร็วและเหมาะสม เพื่อเป็นกำลังใจในการร่วมด้วยช่วยกันในการต่อสู้กับภัยพิบัติที่เกิดขึ้นต่อไป

6. กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณ อาจารย์ ปกป้อง ส่องเมือง อาจารย์ที่ปรึกษาทางวิจัยอย่างสูง ซึ่งได้ให้คำแนะนำถึงประเด็นต่าง ๆ ในการศึกษาและชี้แนวทางในการแก้ปัญหา การค้นคว้าหาข้อมูลเพิ่มเติม อันเป็นประโยชน์ในการวิเคราะห์และสรุปผลการศึกษา รวมทั้งการแก้ไขงานให้สมบูรณ์เป็นอย่างยิ่ง

ขอกราบขอบพระคุณ คณาจารย์ เจ้าหน้าที่สาขาวิชาวิทยาศาสตร์ และนวัตกรรมการศึกษา วิทยาลัยสหวิทยาการ มหาวิทยาลัยธรรมศาสตร์ทุกท่านที่ให้ความดูแล และช่วยเหลือตลอดระยะเวลาการดำเนินงานวิจัยนี้

7. เอกสารอ้างอิง

กิตติศักดิ์ กิตติธานุสรณ์. (2563). การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง [สารนิพนธ์].

สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ.

ชนม์นิภา จันท์เอี่ยม. (2563). การศึกษาและจัดการข้อมูลทางทวิตเตอร์เพื่อเป็นตัวแทนความเชื่อมั่นของนักลงทุนในตลาดหลักทรัพย์ กรณีศึกษาระยะสั้น: สงครามการค้าสหรัฐอเมริกา-จีน [วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ]. คณะเศรษฐศาสตร์ มหาวิทยาลัยเชียงใหม่.

พัฒนพันธ์ กลิ่นหอม. (2564). การวิเคราะห์มุมมองของผู้บริโภคผ่านการแสดงความคิดเห็น (ข้อความเมนต์) ของผู้บริโภคที่มีต่อสินค้าสมาร์ตโฟนบนโซเชียลมีเดีย [วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ]. สาขาวิชาการบริหารเทคโนโลยีสารสนเทศเชิงกลยุทธ์ มหาวิทยาลัยนเรศวร.

พิศิษฐ์ บวรเลิศสุธี และ วรภัทร ไพรีเกรง. (2565). การวิเคราะห์ความรู้สึกทางอารมณ์ในบทความแนะนำสินค้าออนไลน์. วารสารวิจัย มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย, 14(3), 755-769. <https://li01.tci-thaijo.org/index.php/rmutsvrj/article/view/245470/175660>

พุทธิพร ธนธรรมเมธี และ ยาวเรศ ศิริสถิตย์กุล. (2562). เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่มีสมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ. วารสารวิทยาศาสตร์และเทคโนโลยี, 27(6), 1164-1178. <https://li01.tci-thaijo.org/index.php/tstj/article/download/205346/143001/635829>

ภัครพล อาจอาษา. (2564). การวิเคราะห์คุณภาพน้ำด้วยเทคนิคการจัดกลุ่มข้อมูล [วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ]. สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยมหาสารคาม.

รวีสุดา เทศเมือง และ นิเวศ จิระวิชิตชัย. (2560). การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ทเวกเตอร์แมชชีน. วารสารวิศวกรรมศาสตร์ มหาวิทยาลัยสยาม, 18(1), 1-12. https://e-library.siam.edu/e-journal/wp-content/uploads/2018/11/EJSU_No.34_pp_1-12.pdf

วันสวรรค์ มีประเสริฐ และ เอกรัฐ รัชฎาภยจน. (2564). การวิเคราะห์ความคิดเห็นจากทวิตเตอร์ของลูกค้าบริษัทข้อปี่ประเทศไทย. วารสารระบบสารสนเทศด้านธุรกิจ, 7(3), 6-18. <http://doi:10.14456/jisb.2021.11>

- สันติ สุขเกษม. (2561). *การวิเคราะห์รูปแบบความคิดเห็นของข้อมูลโรงแรม โดยใช้เทคนิคการตัดคำแบบผสมผสาน* [วิทยานิพนธ์ปริญญา มหาบัณฑิต]. สาขาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีไทย-ญี่ปุ่น.
- สุชีพ ตันติวุฒิมงคล และคณะ. (2564). การประยุกต์ระบบสารสนเทศภูมิศาสตร์ เพื่อหาพื้นที่ให้บริการของหมวดบำรุงทางหลวงชนบท:กรณีศึกษาแขวงทางหลวงชนบทราชบุรี. *วารสารอักษรศาสตร์*, 50(1), 136-156. <https://so03.tci-thaijo.org/index.php/jletters/article/view/245805/169637>
- อัจฉรา แผ้วบาง และ สายชล สิ้นสมบูรณ์ทอง. (2563). การปรับความไม่สมดุลของข้อมูลด้วยการจำแนก 5 วิธี. *วิทยาศาสตร์และเทคโนโลยี*, 9(4), 418-435. <https://doi.org/10.14456/tjst.2020.46>
- เอกปรียา ไบสนิ. (2563). *การศึกษาวีธีการแบ่งกลุ่มลูกค้าด้วยเทคนิคอาร์เอฟเอ็มร่วมกับการวิเคราะห์ข้อความ*. สารนิพนธ์ สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ.
- Automated. (2564). *การประยุกต์ใช้งาน Scatter Chart*. <https://www.autosoft.in.th/data-visualization/scatter-chart/>.
- Business & Technology. (2565). *รู้จักกับ Sentiment analysis ตัวช่วยธุรกิจในการวิเคราะห์ลูกค้าได้ดียิ่งขึ้น*. <https://aigencorp.com/what-is-sentiment-analysis>.
- Mindphp. (2561). *มาทำความเข้าใจกับ Word Cloud*. <https://www.mindphp.com/forums/viewtopic.php?f=144&t=49427>.
- PL. (2561). *Evaluate Model นั้นสำคัญอย่างไร*. <https://medium.com/mmp-li/evaluate-model-precision-recall-f1-score-machine-learning-101-89dbbada0c96>