การสร้างระบบคัดกรองข้อความการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019

บุษบงก์ คชินทรโรจน์[*1] เดือนเพ็ญ ธีรวรรณวิวัฒน์[2] พาชิตชนัต ศิริพานิช[3]
คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์ 148 หมู่ที่ 3 ถนนเสรีไทย แขวงคลองจั่น เขตบางกะปิ กรุงเทพฯ 10240

Received: 26 January 2021; Revised: 12 April 2021; Accepted: 16 April 2021

บทคัดย่อ

ปัจจุบันการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา หรือ COVID-19 มีการแพร่ระบาดไปยังหลายประเทศทั่วโลก ซึ่งนอกจากจะส่งผลโดยตรงต่อสุขภาพแล้ว ยังมีผลกระทบที่ร้ายแรงที่สุดอย่างหนึ่ง คือ การโจมตีคนต่างชาติต่อคนเชื้อสายเอเชีย (Xenophobia) ในสื่อสังคมออนไลน์รวมทั้งทวิตเตอร์ก็มีรายงานการโจมตีด้วยวาจาหลายครั้ง ซึ่งอาจส่งผลให้เกิดปัญหาสุขภาพจิต สุขภาพกายที่รุนแรงขึ้น รวมถึงอาจก่อให้เกิดเหตุการณ์การใช้ความรุนแรง ผู้วิจัยจึงนำเสนอระบบคัดกรองข้อความการเกลียดกลัวคนต่างชาติ โดยมีการสร้างตัวแบบหัวข้อด้วยวิธีการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation -LDA) และตัวแบบจัดประเภทการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ 2 อัลกอริทึม คือ อัลกอริทึมแรนดอมฟอเรสต์ (Random Forest) และอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และใช้การแปลงเชิงปริมาณทั้งหมด 3 วิธี คือ TF-IDF, Word2vec และ GloVe ในการเพิ่มความแม่นยำของตัวแบบ งานวิจัยนี้ใช้ข้อมูล Xenophobia on Twitter During COVID-19 จาก Kaggle ซึ่งทวีตที่ใช้ศึกษาเป็นทวีตที่ใช้ภาษาอังกฤษของทุกประเทศทั่วโลก จำนวน 1,000,000 ทวีต โดยมีวัตถุประสงค์เพื่อสร้างตัวแบบหัวข้อการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา ซึ่งเป็นแนวทางในการหาประเด็นสำคัญที่พูดถึงในกลุ่มผู้ที่เกลียดกลัวคนต่างชาติ และเพื่อศึกษาอัลกอริทึมจัดประเภทที่เหมาะสมสำหรับการแบ่งแยกทวีตการเกลียดกลัวและไม่เกลียดกลัวคนต่างชาติ ทั้งยังเป็นแนวทางให้ทวิตเตอร์คัดกรองทวีตที่อาจก่อให้เกิดความรุนแรง และสามารถออกมาตรการแก้ไข หรือป้องกันความรุนแรงที่จะเกิดขึ้นได้ ผลการศึกษาการสร้างตัวแบบหัวข้อ ผู้วิจัยได้ทำการแบ่งประเด็นสำคัญที่เหมาะสมออกเป็น 3 หัวข้อซึ่งให้ค่าความสับสน (Perplexity) ต่ำที่สุด เท่ากับ 114186.86 ได้แก่ (1) Arrestment (2) Disgusting (3) Damnation และผลการศึกษาตัวแบบจัดประเภทพบว่า อัลกอริทึมแรนดอมฟอเรสต์เมื่อใช้การแปลงเชิงปริมาณด้วยวิธี TF-IDF ให้ค่าความแม่นยำ F1-Score, Recall และ ROC ไม่แตกต่างกัน เมื่อเปรียบเทียบกับวิธี Word2vec และ GloVe ทั้งนี้เมื่อพิจารณาด้วยค่า Precision ให้ค่าสูงที่สุดเท่ากับ 0.35 ส่วนอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนเมื่อใช้การแปลงเชิงปริมาณด้วยวิธี Word2vec ให้ค่า F1-Score และ Precision สูงที่สุดเท่ากับ 0.43 และ 0.28 ตามลำดับ ดังนั้น อัลกอริทึมแรนดอมฟอเรสต์เมื่อใช้ TF-IDF และอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีนเมื่อใช้ Word2vec เป็นอัลกอริทึมที่เหมาะสมที่สุดกับการจัดประเภทข้อความการเกลียดกลัวและไม่เกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา

คำสำคัญ: การสร้างตัวแบบหัวข้อ แบบจำลอง LDA คุณสมบัติเฉพาะจากข้อความ TF-IDF Word2vec การจำแนกข้อความ

---

* Corresponding author. E-mail: Budsabong.kac@stu.nida.ac.th

[1] นักศึกษามหาบัณฑิต หลักสูตรการวิเคราะห์ธุรกิจและวิทยาการข้อมูล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

[2] รองศาสตราจารย์ หลักสูตรการวิเคราะห์ธุรกิจและวิทยาการข้อมูล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

[3] รองศาสตราจารย์ หลักสูตรการวิเคราะห์ธุรกิจและวิทยาการข้อมูล คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

Topic Modeling and Text Classification for Xenophobia on Twitter during COVID-19

Budsabong Kachintararojn[*1] Duanpen Teerawanviwat[2] Pachitjanut Siripanich[3]

Faculty of Applied Statistics, Graduate School of Applied Statistics,

National Institute of Development Administration,

148 Moo3, Serithai Road, Klong-Chan, Bangkapi, Bangkok THAILAND 10240

Received: 26 January 2021; Revised: 12 April 2021; Accepted: 16 April 2021

Abstract

Currently, the coronavirus epidemic (COVID-19) is spread throughout the country around the world, affecting health and yet one of the serious consequences is the expatriation of Xenophobia. On social media, including Twitter, there have been reports of verbal attacks, which can lead to more serious mental and health problems, including violence. This research presents a topic modeling using the Latent Dirichlet Allocation (LDA) and two classification algorithms, which are Random Forest and Support Vector Machine for Xenophobia on Twitter during COVID-19 with three methods of Word Embedding - TF-IDF, Word2vec and GloVe. The dataset contains Xenophobia on Twitter During COVID-19 around 1,000,000 tweets in English language from Kaggle. This research aimed to build the topic model for Xenophobia on Twitter during COVID-19 and study suitable classification algorithms for Xenophobic or Non-Xenophobic and also provides a way for Twitter to filter out tweets that are potentially violent. The results of topic modeling was showed 3 unique important topics that given the lowest Perplexity of 114186.86 which are (1) Arrestment (2) Disgusting (3) Damnation. As the results of classification algorithms were showed that Random Forest when using TF-IDF given F1-Score, Recall and ROC were not different but Precision was the highest value of 0.35. Therefore, Random Forest when using TF-IDF was the most suitable algorithm for Xenophobic or Non-Xenophobic classify during COVID-19. In addition, Support Vector Machine when using Word2vec given the highest F1-Score and Precision value of 0.43 and 0.28, respectively. Therefore, Support Vector Machine when using Word2vec was the most suitable algorithm for Xenophobic or Non-Xenophobic classify during COVID-19.

Keywords: topic modeling, LDA model, word embedding, TF-IDF, word2vec, text classification

---

* Corresponding author. E-mail: Budsabong.kac@stu.nida.ac.th

[1] Master's student in Department of Business Analytics and Data Science, Faculty of Applied Statistics, Graduate School of Applied Statistics, National Institute of Development Administration

[2] Associate Professor in Department of Business Analytics and Data Science, Faculty of Applied Statistics, Graduate School of Applied Statistics, National Institute of Development Administration

[3] Associate Professor in Department of Business Analytics and Data Science, Faculty of Applied Statistics, Graduate School of Applied Statistics, National Institute of Development Administration

1. บทนำ

การแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา หรือ COVID-19 เกิดจากไวรัสที่มีการระบาดในเมืองอู่ฮั่นประเทศจีนในเดือนธันวาคมปี ค.ศ. 2019 ปัจจุบันมีการแพร่ระบาดไปยังหลายประเทศทั่วโลก ผู้คนต่างถูกครอบงำด้วยอุปทานหมู่ และแสดงความตื่นตระหนกจากการแพร่ระบาดในครั้งนี้ โดยมีการกักตุนน้ำยาฆ่าเชื้อ และหน้ากากอนามัยเพื่อป้องกันตัวเองจาก การแพร่ระบาด ในขณะที่องค์การอนามัยโลกและหน่วยงานต่าง ๆ พยายามเผยแพร่ข้อมูลที่ถูกต้องเกี่ยวกับโรคนี้อย่างโปร่งใส [1] แต่กระแสความเชื่อเกี่ยวกับการเกลียดกลัวคนต่างชาติกลับแพร่ระบาดบนอินเตอร์เน็ตมากขึ้น ซึ่งการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนานอกจากจะส่งผลโดยตรงต่อสุขภาพแล้ว ยังมีผลกระทบค่อนข้างมาก เช่น ตลาดการเงินตกต่ำ ยอดขายเบียร์โคโรนาลดลง และอุตสาหกรรมการท่องเที่ยวก็ได้รับความเสียหาย แต่ผลข้างเคียงที่ร้ายแรงที่สุดอย่างหนึ่ง คือ การโจมตีคนต่างชาติต่อคนเชื้อสายเอเชีย หรือการเกลียดกลัวคนต่างชาติ (Xenophobia) ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา ผู้คนจำนวนมากต้องเผชิญกับโรคเกลียดกลัวคนต่างชาติเนื่องจาก รูปร่างหน้าตา สัญชาติ และอื่น ๆ ในสื่อสังคมออนไลน์รวมทั้งทวิตเตอร์ก็มีรายงานการโจมตีด้วยวาจาหลายครั้ง [2] ซึ่งอาจส่งผลให้เกิดปัญหาสุขภาพจิต สุขภาพกายที่รุนแรงขึ้น และความยากลำบากในการควบคุมโรคติดเชื้อระหว่างการแพร่ระบาด รวมถึงอาจก่อให้เกิดความไม่พอใจกันของคนในพื้นที่กับคนต่างชาติ ทำให้มีเหตุการณ์การใช้ความรุนแรงเกิดขึ้น โดยรัฐสภาคองเกรสแห่งสหรัฐอเมริกาก็ได้ตั้งข้อสังเกตว่าการก่ออาชญากรรมจากความเกลียดชังคนอเมริกันเชื้อสายเอเชียเพิ่มขึ้นเป็นประมาณ 100 เหตุการณ์ต่อวันนับตั้งแต่มีการแพร่ระบาดเกิดขึ้น [3]

ในงานวิจัยนี้ผู้วิจัยจะใช้ชุดเครื่องมือสำหรับการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ในการประมวลผลทวีตที่เกิดขึ้นในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา โดยมีวัตถุประสงค์เพื่อสร้างตัวแบบหัวข้อ หรือ Topic Modeling ในการวิเคราะห์หาสาเหตุ ประเด็นสำคัญที่พูดถึงในกลุ่มผู้ที่เกลียดกลัวคนต่างชาติ และเพื่อให้ได้อัลกอริทึมที่เหมาะสมที่สุดในการจัดประเภทข้อความทวีตการเกลียดกลัวและไม่เกลียดกลัวคนต่างชาติในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา เป็นแนวทางในการคัดกรอง และป้องกันไม่ให้ข้อความเหล่านั้นแพร่กระจายออกไป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 การเกลียดกลัวคนต่างชาติ

การเกลียดกลัวคนต่างชาติ หรือ Xenophobia เกิดจากความกังวลด้านสุขภาพที่มีรากฐานมาจากในช่วงปี ค.ศ. 1800 เมื่อชาวจีนจำนวนมากถูกจ้างเป็นแรงงานข้ามชาติราคาถูก คนงานถูกน้ำท่วมในสหรัฐอเมริกาทำให้เกิดความตื่นตัว และกระตุ้นให้เกิดความรู้สึกต่อต้านชาวจีนด้วยการตีตราว่าเป็นพาหะของ “ความสกปรกและโรค” [4] ปัจจุบันการเกลียดกลัวคนต่างชาติได้แพร่กระจายเหมือนกับไวรัส ซึ่งไม่ได้ส่งผลกระทบต่อคนเชื้อสายจีนเพียงอย่างเดียว แต่ยังส่งผลกระทบต่อคนเชื้อสายเอเชียตะวันออกด้วย [5] เมื่อโรคติดเชื้อไวรัสโคโรนาได้แพร่ระบาดอย่างรวดเร็ว ก็เกิดเหตุการณ์การโจมตีชาวเอเชียในสหรัฐอเมริกาตั้งแต่เดือนมกราคม ปี ค.ศ. 2020 รวมไปถึงการทำร้ายร่างกาย และการพูดจาดูถูกเหยียดหยาม [6] ซึ่งไม่มีสถิติอย่างเป็นทางการที่จะเปิดเผยข้อมูลเกี่ยวกับอาชญากรรมความเกลียดชังที่มีแรงจูงใจทางเชื้อชาติ ทำให้คนเชื้อสายเอเชียก็จะยังคงตกอยู่ในความเสี่ยงอันเป็นผลมาจากความอคติทางเชื้อชาติในการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา และทางสมาชิกรัฐสภาคองเกรสแห่งสหรัฐอเมริกาก็ได้ตั้งข้อสังเกตว่าการก่ออาชญากรรมจากความเกลียดชังกับคนอเมริกันเชื้อสายเอเชีย ได้เพิ่มขึ้นเป็นประมาณ 100 เหตุการณ์ต่อวัน [3] ดังนั้นจึงมีความสำคัญอย่างยิ่งต่อการดูแลสุขภาพ และควรกำจัดความคิดที่ว่าเป็นคนจีนเทียบเท่ากับการเป็นพาหะของโรค และใช้ความรู้ที่มีในการเลือกปฏิบัติ เพื่อเตรียมความพร้อมสำหรับการเพิ่มขึ้นอย่างรวดเร็วของอาชญากรรมที่เกิดการเกลียดกลัวคนต่างชาติ และเพื่อลดการตีตราให้กับเอเชียกลุ่มน้อยที่ได้รับผลกระทบจากการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา [7]

ในงานวิจัย เรื่อง การสร้างระบบคัดกรองข้อความการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่

ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 ได้ให้คำนิยามของคำว่า การเกลียดกลัวคนต่างชาติ (Xenophobia) คือ การพูดจาดูถูกเหยียดหยาม รวมไปถึงมีพฤติกรรมต่อต้าน รังเกียจ และขับไล่คนต่างชาติ โดยแสดงออกผ่านข้อความบนทวิตเตอร์

2.2 ข้อมูลที่ไม่สมดุล

ข้อมูลที่ไม่สมดุล หรือ Imbalanced Datasets หมายถึง ข้อมูลที่มีการกระจายตัวไม่เท่ากัน หรือ ข้อมูลที่มีจำนวนสมาชิกในกลุ่มหนึ่งมากกว่าจำนวนสมาชิกในอีกกลุ่มหนึ่งเป็นจำนวนมาก [8] ซึ่งข้อมูลไม่สมดุลนั้นมีสาเหตุมาจากหลายปัจจัย เช่น ข้อมูลผู้ป่วยโรคมะเร็งชนิดหนึ่ง โดยผู้ที่เป็นมะเร็ง (Positive) มีข้อมูลเพียงประมาณหลักร้อย แต่ผู้ที่ไม่เป็นมะเร็ง มีข้อมูลประมาณหลายแสนคน จึงทำให้ส่งผลกระทบต่อการจัดประเภทข้อมูลส่วนน้อยนั้นให้มีความแม่นยำในการจำแนกลดน้อยลงไป โดยวิธีการแก้ปัญหาข้อมูลไม่สมดุล คือ การสุ่มข้อมูลซ้ำ (Resampling) ซึ่งมีทั้งการสุ่มข้อมูลเพิ่ม และการสุ่มข้อมูลลดลง [9] ซึ่งวิธีการแก้ปัญหาดังกล่าวสามารถจำแนกประเภทข้อมูลข้อมูลส่วนน้อยในข้อมูลที่ไม่สมดุลได้อย่างมีประสิทธิภาพ [10]

2.3 การแปลงเชิงปริมาณ

การแปลงเชิงปริมาณ หรือ Word Embedding คือ การแปลงคำ (Word) เป็นตัวเลข (Numerical) ให้อยู่ในรูปแบบของเวกเตอร์ (Vector) ที่เหมาะสมจะนำเข้าไปประมวลผลกับตัวแบบ

2.3.1 TF-IDF หรือ Term Frequency Inverse Document Frequency เป็นเทคนิคการคัดแยกคำตามความสำคัญ ที่ถูกใช้ในการสร้างเวกเตอร์ โดยเทคนิคนี้ใช้ในการประเมินความสำคัญของคำในข้อความทั้งหมด ความสำคัญจะมีสัดส่วนเพิ่มตามจำนวนครั้งของคำที่เกิดขึ้นในข้อความทั้งหมด เพื่อเปรียบเทียบกับสัดส่วนผกผันของคำนั้น ๆ ในข้อความทั้งหมด หรือพูดอีกนัยหนึ่งว่า TF-IDF จะช่วยกรองคำเด่น ๆ ออกมาจากข้อความ [11] ซึ่งจะเพิ่มความสำคัญของคำที่ถูกพบโดยการให้น้ำหนักเพิ่มขึ้น ส่งผลทำให้ตัวแบบมีความแม่นยำมากขึ้น สามารถทำนายได้ดีกว่าวิธีเวกเตอร์ของการนับคำ (Count Vectorization) [12] โดย TF-IDF สามารถคำนวณได้ดังสมการที่ (3)

$$TF_{ij} = \frac{f_{ij}}{n_j} \quad (1)$$

เมื่อ f_{ij} แทนจำนวนความถี่ของคำ i ในข้อความ j

n_j แทนจำนวนคำทั้งหมดในข้อความ j

$$IDF_i = 1 + log\frac{N}{c_i} \quad (2)$$

เมื่อ N แทนจำนวนข้อความทั้งหมด

c_i แทนจำนวนข้อความที่มีคำ i ปรากฏอยู่ในข้อความ

$$w_{ij} = TF_{ij} \times IDF_i \quad (3)$$

เมื่อ w_{ij} แทนจำนวนคะแนนของการคัดแยกคำตามความสำคัญ

2.3.2 Word2vec เป็นแบบจำลองที่ใช้สร้างการฝังคำ หรือแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งเวกเตอร์ของคำต่าง ๆ ถูกคำนวณจากบริบทรอบข้าง โดยใช้เทคนิคโครงข่ายประสาทเทียม (Neural Network) แบบ Encoder-Decoder มีเลเยอร์ (Layer) จำนวน 2 เลเยอร์ ซึ่งมีหลักการในการเปรียบเทียบเวกเตอร์ทางความหมายของคำทั้ง 2 คำ แล้วคืนค่าออกมาเป็นตัวเลขตั้งแต่ -1 ถึง 1 บ่งชี้ถึงความใกล้เคียงทางความหมายให้ค่าน้อยไปมาก หรือพูดอีกนัยหนึ่งว่า คำที่มีบริบทการปรากฏคล้าย ๆ กัน ควรเป็นคำที่มีความหมายคล้ายกันด้วย [13] ซึ่งวิธี Word2vec สามารถวัดค่าความละม้ายทางความหมายของเวกเตอร์ของคำ และใช้ร่วมกับการจำแนกข้อมูลประเภทข้อความได้ดี [14] โดยเมื่อมีมิติของข้อมูล (Dimension) มากขึ้นจะทำให้ตัวแบบมีความแม่นยำมากขึ้น

2.3.3 GloVe หรือ Global Vectors for Word Representation เป็นแบบจำลองถดถอยล็อก-ไบลิเนียร์ (Log-bilinear Regression) ที่รวม 2 ตัวแบบเข้าด้วยกัน คือ โกลบอลเมตริกซ์แฟคเตอร์ไรเซชัน (Global Matrix Factorization) และโลคอลคอนเท็กซ์วินโดว์ (Local Context Window) และเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) อัลกอริทึมสำหรับการแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งแบบจำลอง GloVe ได้รับการสอนให้เรียนรู้สถิติการเกิดร่วมกันของคำจากคลังข้อมูล และแสดงผลลัพธ์เป็นเวกเตอร์ของคำในปริภูมิเวกเตอร์ [15] โดยเมื่อมีมิติของข้อมูล (Dimension) มากขึ้นจะทำให้ตัวแบบมีความแม่นยำมากขึ้นเช่นเดียวกันกับวิธี Word2vec [16]

2.4 วิธีที่ใช้สร้างตัวแบบหัวข้อ

ตัวแบบหัวข้อ หรือ Topic modeling เป็นการสร้างแบบจำลองการกระจายตัวของข้อมูล เพื่อนำมาใช้ในการจัดกลุ่มข้อมูล ตัวแบบหัวข้อนี้มีพื้นฐานมาจากแนวคิดที่ว่า ในเอกสารเกิดจากการรวมตัวของหลาย ๆ หัวข้อซึ่งแต่ละหัวข้อมีการแจกแจงความน่าจะเป็นของคำที่เกิดขึ้นหลายๆ คำ ในแต่ละหัวข้อนั้น

การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation - LDA) เป็นตัวแบบความน่าจะเป็นสำหรับข้อมูลไม่ต่อเนื่อง ในกระบวนการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ซึ่งถูกสร้างขึ้นโดยมีแนวคิดว่าในเอกสาร (Document) จะประกอบด้วยหัวข้อ (Topic) รวมกันอยู่แบบสุ่ม ซึ่งหัวข้อเหล่านี้ได้กระจัดกระจายอยู่ในกลุ่มของคำ (Word) ของเอกสารนั้น ๆ การจัดสรรดีรีเคลแฝงถูกใช้ในการหาหัวข้อ หรือ กลุ่มคำที่ต้องสกัดออกมาจากเอกสาร โดยคำนวณหาค่าความน่าจะเป็น (Probabilistic) จากคำที่ปรากฏในเอกสาร ซึ่งจะเป็นหัวข้อแฝง (Latent Topic) ที่ไม่สามารถสังเกตได้อย่างชัดเจน ผลลัพธ์ของการจัดสรรดีรีเคลแฝงจะบอกได้ว่าคำแต่ละคำอยู่ในหัวข้อแฝงที่ 1 2 หรือ 3 ด้วยความน่าจะเป็นเท่าไร และเอกสารหนึ่งมีสัดส่วนของแต่ละหัวข้อแฝงเท่าไร [17] ซึ่งวิธี LDA เป็นวิธีที่ใช้บ่อยที่สุดที่มีความยืดหยุ่น และปรับเปลี่ยนได้สำหรับการสร้างแบบจำลองหัวข้อ [18] สามารถทำการแยกประเด็นสำคัญของข้อความได้ดี [19]

2.5 วิธีที่ใช้สร้างตัวแบบจัดประเภททวีต

2.5.1 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นอัลกอริทึมที่ใช้พื้นฐานการเรียนรู้ที่จำแนกข้อมูลด้วยระนาบแบ่งข้อมูล (Hyperplane) ออกเป็น 2 ส่วน โดยใช้หลักการสร้างเส้นแบ่งกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด (Maximum Margin Hyperplane: MMH) โดยที่อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน จะใช้แม็ปฟังก์ชั่น (Mapping Function) สำหรับแปลงข้อมูลจากอินพุตสเปส (Input Space) ไปยังฟีเจอร์สเปส (Feature Space) และสร้างเคอร์เนลฟังก์ชั่น (Kernel Function) บนฟีเจอร์สเปสเพื่อใช้ในการวัดความคล้ายกันของข้อมูล สำหรับเคอร์เนลฟังก์ชั่น ได้แก่ เคอร์เนลเส้นตรง (Linear Kernel), เคอร์เนลโพลิโนเมียล (Polynomial Kernel), เรเดียลเบสิสฟังก์ชั่น (Radial Basis Function) และเคอร์เนลซิกมอยด์ (Sigmoid) [20] ซึ่งสามารถจำแนกข้อมูลประเภทข้อความได้ดีที่สุด และให้ค่าความถูกต้องแม่นยำที่สุด [21]

2.5.2 แรนดอมฟอเรสต์ (Random Forest) เป็นอัลกอริทึมการสุ่มเลือกใช้ข้อมูล และคุณลักษณะประเภทหนึ่งของอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ที่มีลักษณะแบบไม่ตัดแต่งกิ่ง (Unpruned) ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูล และคุณลักษณะแบบใส่คืน (Sampling with Replacement) แล้วนำมาสร้างเป็นต้นไม้ (Tree) ซึ่งจะมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก เรียกว่า Out-of-Bag (OOB) จะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ วิธีการนี้เรียกว่าแบ็กกิ้ง (Bagging) ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้ตัดสินใจแต่ละต้นจะถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด อัลกอริทึมแรนดอมฟอเรสต์ไม่จำเป็นต้องมีข้อมูลทดสอบ เพื่อประมาณความผิดพลาดเพราะข้อมูลส่วนหนึ่งที่ไม่ถูกเลือก (OOB) นั้นถูกนำมาใช้ทดสอบต้นไม้ตัดสินใจแล้ว [22] ซึ่งสามารถจำแนกข้อมูลประเภทข้อความได้ดี และให้ค่าความถูกต้องแม่นยำมากกว่าอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน [23]

2.6 วิธีการประเมินประสิทธิภาพของตัวแบบ

2.6.1 ค่าความสับสน (Perplexity) คือ วิธีประเมินประสิทธิภาพของตัวแบบหัวข้อ ในรูปแบบเชิงปริมาณ (Quantitative) มีผลลัพธ์เป็นตัวเลข โดยดูการกระจายตัวของความน่าจะเป็น (Probability) ว่าดีมากน้อยเพียงใด หากค่าความสับสนต่ำมากเท่าไร แสดงถึงประสิทธิภาพตัวแบบที่ดีมากเท่านั้น โดยค่าความสับสนคำนวณได้ดังสมการที่ (4)

$$Perplexity = e^{(-1*\log(w))} \quad (4)$$

เมื่อ $log(w)$ คือ log-likelihood ต่อคำ

2.6.2 เมตริกซ์วัดประสิทธิภาพ (Confusion Matrix) คือ การประเมินผลลัพธ์การทำนาย หรือความสามารถในการจำแนกข้อมูลของตัวแบบ จากโปรแกรมเปรียบเทียบกับผลลัพธ์จริง ซึ่งการประเมินผลลัพธ์การทำนายของตัวแบบสามารถวัดจากผลลัพธ์การจัดกลุ่มข้อมูล (Classification) แสดงดังตารางที่ 1

ตารางที่ 1 เมตริกซ์วัดประสิทธิภาพ (Confusion Matrix)

ผลลัพธ์จริง (Actual Class)	ค่าที่ทำนายได้ (Predicted Class)	
	Class YES	Class NO
Class YES	True Positive	False Negative
Class NO	False Positive	True Negative

ค่าของผลลัพธ์ที่ได้จากการจัดกลุ่ม ได้แก่ ค่า True Positive (TP) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าผลลัพธ์จริงอยู่ใน Class YES และมีการทำนายว่าอยู่ใน Class YES (ทำนายถูกต้อง), False Negative (FN) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าผลลัพธ์จริงอยู่ใน Class YES และมีการทำนายว่าอยู่ใน Class NO (ทำนายผิด), ค่า False Positive (FP) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าผลลัพธ์จริงอยู่ใน Class NO และมีการทำนายว่าอยู่ใน Class YES (ทำนายผิด) และ True Negative (TN) คือ ค่าที่บอกความถูกต้องในการจำแนกข้อมูลซึ่งมีค่าผลลัพธ์จริงอยู่ใน Class NO และมีการทำนายว่าอยู่ใน Class NO (ทำนายถูกต้อง) ตามลำดับ

ค่า TP, FP, TN และ FN สามารถนำมาใช้คำนวณมาตรวัดต่าง ๆ เพื่อประเมินประสิทธิภาพการจำแนกข้อมูลของตัวแบบ โดยมาตรวัดที่นิยมใช้โดยทั่วไปประกอบด้วย ค่า Accuracy, Precision และ Recall โดยแต่ละมาตรวัดคำนวณได้ดังนี้

ค่า Accuracy คือ ค่าที่บอกว่ามีการทำนายข้อมูลถูกต้อง และมีค่าความแม่นยำเท่าไรในการทำนาย คำนวณได้ดังสมการที่ (5)

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

ค่า Precision คือ ค่าที่บอกว่าโปรแกรมทำนายว่าจริง ถูกต้องเท่าไร คำนวณได้ดังสมการที่ (6)

$$Precision = \frac{TP}{TP+FP} \quad (6)$$

ค่า Recall หรือ Sensitivity คือ ค่าที่บอกว่าโปรแกรมทำนายได้ว่าจริงอยากถูกต้อง เป็นอัตราส่วนเท่าไรของค่าผลลัพธ์จริงที่อยู่ใน Class YES ทั้งหมด คำนวณได้ดังสมการที่ (7)

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

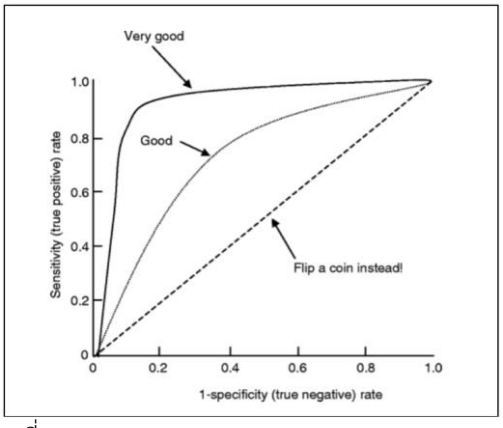
ค่า Specificity คือ ค่าที่บอกว่าโปรแกรมทำนายได้ว่าไม่จริงอย่างถูกต้อง เป็นอัตราส่วนเท่าไรของค่าผลลัพธ์จริงที่อยู่ใน Class NO ทั้งหมด คำนวณได้ดังสมการที่ (8)

$$Specificity = \frac{TN}{TN+FP} \quad (8)$$

ค่า F1-Score คือ ค่าเฉลี่ยฮาร์โมนิกของ Precision และ Recall โดยค่า F1-Score ที่สูงที่สุดคือ 1.0 ในขณะที่ F1-Score ที่ต่ำที่สุดคือ 0.0 คำนวณได้ดังสมการที่ (9)

$$F1 - Score = 2 * (\frac{Precision*Recall}{Precision+Recall}) \quad (9)$$

2.6.3 Receiver Operating Characteristic (ROC) Curve คือ กราฟความสัมพันธ์ระหว่าง True Positive Rate กับ False Positive rate ที่จุดตัด (Cut-off Point) ต่าง ๆ กัน ดังรูปที่ 1



รูปที่ 1 Receiver Operating Characteristic Curve

ตัวแบบที่ดีควรมี Recall และ Specificity สูง ซึ่งจะทำให้มี False Positive Rate ต่ำ ส่งผลให้พื้นที่ใต้โค้ง ROC เข้าชิดมุมซ้ายบนมากที่สุด นอกจากนี้พื้นที่ใต้โค้ง ROC ยังเป็นดัชนีที่ใช้ในการบ่งชี้ความถูกต้อง หรือความน่าเชื่อถือของตัวแบบ ซึ่งตัวแบบใดที่มีพื้นที่ใต้โค้ง ROC มากกว่าแสดงถึงประสิทธิภาพที่สูงกว่า

3. วิธีดำเนินการวิจัย

ในการศึกษางานวิจัยเรื่อง การสร้างระบบคัดกรองข้อความการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 ใช้ข้อมูล Xenophobia on Twitter During COVID-19 จาก Kaggle

ตั้งแต่เดือนกุมภาพันธ์ ถึง เดือนพฤษภาคม ปี ค.ศ. 2020 ที่อยู่ในรูปแบบไฟล์ประเภท Comma Separated Value (CSV) ซึ่งประกอบไปด้วย ข้อความทวีต และประเภทของข้อความที่บ่งบอกว่าทวีตใดเป็นทวีตที่เกลียดกลัวคนต่างชาติ (Xenophobic) หรือ ไม่เกลียดกลัวคนต่างชาติ (Non-Xenophobic) จำนวน 1,000,000 ทวีต โดยผู้วิจัยได้แบ่งวิธีดำเนินการวิจัยออกเป็น 4 ขั้นตอนหลัก โดยมีสาระสำคัญตามลำดับต่อไปนี้

3.1 กระบวนการเตรียมข้อมูล

กระบวนการเตรียมข้อมูล หรือ Text Data Preprocessing ก่อนที่จะนำข้อมูลประเภทข้อความไปให้คอมพิวเตอร์ทำการประมวลผล เป็นกระบวนการนำข้อความทวีตจากทวิตเตอร์ที่อยู่ในรูปแบบไฟล์ประเภท CSV ซึ่งประกอบไปด้วย ข้อความทวีต และประเภทของข้อความที่บ่งบอกว่าทวีตใดเป็นทวีตที่เกลียดกลัวคนต่างชาติ (Xenophobic) หรือ ไม่เกลียดกลัวคนต่างชาติ (Non-Xenophobic) จำนวน 1,000,000 ทวีต โดยใช้ Natural Language Toolkit (Nltk) ที่เป็น Library ของ Python ซึ่งเป็นชุดโปรแกรมสำหรับการประมวลผลภาษาธรรมชาติเชิงสถิติสำหรับภาษาอังกฤษที่ประกอบด้วยอัลกอริทึมที่หลากหลาย เพื่อช่วยในทำการเตรียมข้อมูลตามขั้นตอนดังต่อไปนี้

3.1.1 Missing Value เป็นการตรวจสอบข้อมูลสูญหาย โดยข้อมูลในงานวิจัยนี้ไม่มีข้อมูลสูญหาย

3.1.2 Punctuation Removal เป็นการลบอักขระหรือ สัญลักษณ์ต่าง ๆ ที่ไม่เกี่ยวข้องออก

3.1.3 Tokenization เป็นกระบวนการตัดคำ หรือแบ่งข้อความออกเป็นส่วนที่เล็กลง เช่น การแบ่งข้อความออกเป็นประโยค และแบ่งจากประโยคเป็นคำ แล้วทำการลบเครื่องหมายวรรคตอนออกทั้งหมด

3.1.4 การลบ Stop Words ซึ่งเป็นคำทั่วไปที่พบบ่อย และไม่ได้สื่อความหมายพิเศษ พร้อมทั้งลบคำที่มีอักขระน้อยกว่า 3 ตัวอักษรออก เช่น a, an, the, is, am, are, in, on, at

3.1.5 Lemmatization เป็นกระบวนการแปลงคำด้วยรายการคำศัพท์ในพจนานุกรม (Dictionary) หรือ แปลงสรรพนามบุรุษที่ 3 ให้เป็น สรรพนามบุรุษที่ 1 และทำการเปลี่ยน Past Tenses และ Future Tenses ให้เป็น Present Tenses ให้อยู่ในรูปแบบพื้นฐาน

3.1.6 Stemming เป็นกระบวนการเปลี่ยนคำให้อยู่ในรูปของรากศัพท์ (Root Form) หรือ การตัด ing, ed, s, es, er ออก เช่น winning เมื่อผ่านการทำ Stemming จะได้คำว่า win

3.2 กระบวนการแปลงเชิงปริมาณ

กระบวนการแปลงเชิงปริมาณเป็นกระบวนการแทนคำ (Words) ด้วยเวกเตอร์ (Vector) หรือ Word Embedding โดยหลังจากได้ทำการเตรียมข้อมูลจากขั้นตอนที่ผ่านมาแล้ว จะเป็นการนำเอาคำที่ถูกตัด หรือ คำที่เหมาะสมจะนำเข้าตัวแบบ ไปแปลงให้อยู่ในรูปตัวเลข (Numerical) โดยใช้เทคนิคการคัดแยกคำตามความสำคัญ (TF-IDF) สำหรับการสร้างตัวแบบหัวข้อ และใช้ TF-IDF, Word2vec และ GloVe ในการเปรียบเทียบประสิทธิภาพของการแปลงเชิงปริมาณสำหรับการสร้างตัวแบบจัดประเภท

3.3 การสร้างตัวแบบหัวข้อ

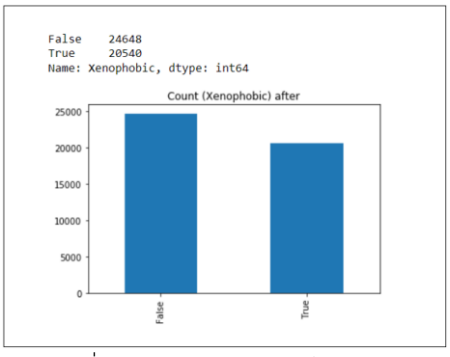
การสร้างตัวแบบหัวข้อ หรือ การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation - LDA) ในกระบวนการนี้ใช้เฉพาะข้อความทวีตที่เป็นประเภทเกลียดกลัวคนต่างชาติ (Xenophobic) เพื่อหาหัวข้อ สาเหตุ หรือ ประเด็นสำคัญที่พูดถึงในกลุ่มผู้ที่เกลียดกลัวคนต่างชาติ โดยผู้วิจัยใช้ Library จาก sklearn.decomposition.LatentDirichletAllocation ในการสร้างตัวแบบ และใช้ Library จาก sklearn.model_selection.GridSearchCV เพื่อปรับพารามิเตอร์ที่สำคัญและเหมาะสมที่สุดสำหรับตัวแบบการจัดสรรดีรีเคลแฝง ในที่นี้ได้กำหนดพารามิเตอร์ที่ต้องการ คือ n_components (จำนวนของหัวข้อ) เท่ากับ 3, 4, 5, 6, 7, 10, 15, 20 และ learning_decay (อัตราการเรียนรู้) เท่ากับ 0.5, 0.7, 0.9 [24] เพื่อให้ได้ค่าความสับสน (Perplexity) ที่ต่ำที่สุดซึ่งแสดงถึงประสิทธิภาพของตัวแบบที่ดี

การตั้งชื่อประเด็นสำคัญ ผู้วิจัยได้พิจารณาถึงคำที่โดดเด่นของแต่ละประเด็น จากนั้นทำการจับกลุ่มว่าประเด็นสำคัญใดสอดคล้องกับเรื่องใดบ้าง

3.4 การสร้างตัวแบบจัดประเภท

3.4.1 การแบ่งข้อมูลทดสอบ (Split Testing) ในงานวิจัยนี้ได้แบ่งข้อมูลออกเป็น 2 ชุด โดยข้อมูลชุดแรก คือ ข้อมูลเรียนรู้ (Training Data) ซึ่งสุ่มมาจากข้อมูลที่ทำการเตรียมข้อมูล (Preprocess) แล้ว 75% เพื่อนำไปใช้ในการสร้างตัวแบบ และข้อมูลชุดที่สอง คือ ข้อมูลทดสอบ (Test Data) อีก 25% เพื่อนำไปใช้ทดสอบประสิทธิภาพของตัวแบบ และใช้วิธี 10-Fold Cross Validation คือแบ่งข้อมูลออกเป็น 10 ส่วนเท่า ๆ กัน หลังจากนั้นใช้ข้อมูลส่วนหนึ่งเป็นตัวทดสอบประสิทธิภาพของตัวแบบ

3.4.2 การแก้ปัญหาข้อมูลไม่สมดุล (Resampling Imbalanced Data) เนื่องจากประเภทของข้อความที่บ่งบอกว่าเป็นทวีตที่เกลียดกลัวคนต่างชาติ (Xenophobic) มีจำนวน 745,892 ทวีต และประเภทของข้อความที่บ่งบอกว่าเป็นทวีตที่ไม่เกลียดกลัวคนต่างชาติ (Non-Xenophobic) มีจำนวน 4,108 ทวีต ผู้วิจัยจึงทำการสุ่มข้อมูลเพิ่มและลด (Resampling) ได้ข้อความที่เป็นทวีตที่เกลียดกลัวคนต่างชาติจำนวน 20,540 ทวีต และข้อความที่เป็นทวีตที่ไม่เกลียดกลัวคนต่างชาติจำนวน 24,648 ทวีต ดังรูปที่ 2



รูปที่ 2 การแก้ปัญหาข้อมูลไม่สมดุล

3.4.3 การแทนคำ (Words) ด้วยเวกเตอร์ (Vector) หรือ Word Embedding ใช้เทคนิคการคัดแยกคำตามความสำคัญ (TF-IDF), Word2vec ในงานวิจัยนี้ใช้แบบจำลองซึ่งสร้างขึ้นจากข้อมูลภายในคลังข้อมูล Google News Corpus ที่มีจำนวน 3 ล้านคำ 300 มิติข้อมูล (Dimensions) และ GloVe ในงานวิจัยนี้ใช้แบบจำลองซึ่งสร้างขึ้นจากข้อมูลภายในคลังข้อมูล Wikipedia 2014 + Gigaword 5 ที่มีจำนวน 6 พันล้านคำ 300 มิติข้อมูล (Dimensions) เพื่อให้อยู่ในรูปแบบตัวเลขที่สามารถนำไปเข้าตัวแบบจัดประเภทได้

3.4.4 การสร้างตัวแบบจัดประเภท ในงานวิจัยครั้งนี้มุ่งเน้นการนำตัวแบบ 2 รูปแบบมาใช้ในการทดสอบเพื่อหาตัวแบบที่มีประสิทธิภาพที่สุด โดยตัวแบบที่ถูกเลือกมาใช้จะมีโครงสร้างพื้นฐานแตกต่างกันออกไป คือ ตัวแบบโครงสร้างต้นไม้ (Tree based Model) และตัวแบบฐานการเรียนรู้ (Learning based Model) มีตัวแบบในการวิเคราะห์ คือ อัลกอริทึมแรนดอมฟอเรสต์ (Random Forest) และอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ตามลำดับ

3.4.5 การเปรียบเทียบประสิทธิภาพของตัวแบบจะใช้ค่า F1-Score เป็นอันดับแรกในการวัดผล ซึ่งจะคำนึงถึงค่า Accuracy รองลงมา ดังนั้นผลลัพธ์ที่ได้จะไม่ได้มีค่าความแม่นยำสูงสุดสำหรับ Positive Class (Xenophobic) และ Negative Class (Non-Xenophobic) แต่จะสามารถรวบรวมจำนวน Positive Class (Xenophobic) ได้มากและแม่นยำที่สุด และค่า ROC Curve ซึ่งตัวแบบใดที่มีค่าดังกล่าวสูงที่สุด จะเป็นตัวแบบจัดประเภทที่ดีที่สุด

4. ผลการศึกษา

4.1 ผลลัพธ์ของจำนวนหัวข้อ หรือ ประเด็นสำคัญ

จากการปรับพารามิเตอร์ด้วย sklearn.model selection.GridSearchCV ได้ผลลัพธ์เป็น learning_decay (อัตราการเรียนรู้) เท่ากับ 0.5, n_components (จำนวนของหัวข้อ) เท่ากับ 3 หัวข้อ และให้ค่าความสับสน (Perplexity) ต่ำที่สุด เท่ากับ 114186.86 ดังรูปที่ 3

```
Best Model's Params:  {'learning_decay': 0.5, 'n_components': 3}
Model Perplexity:  114186.86200357963
```

รูปที่ 3 พารามิเตอร์ของตัวแบบ LDA ที่ให้ค่าความสับสน (Perplexity) ต่ำที่สุด

ผลลัพธ์ของคำสำคัญ 15 อันดับแรก ของแต่ละหัวข้อจากการสร้างตัวแบบการจัดสรรดีรีเคลแฝง (LDA) แสดงดังตารางที่ 2

ตารางที่ 2 คำสำคัญ 15 อันดับแรกของแต่ละหัวข้อ

Topic / Word	1	2	3
1	fear	case	widespread
2	spread virus	resident	send
3	center	disease	death
4	temporarily	confirm	treatment
5	territory	total	country
6	legal	treat	containment
7	rebel	know	seek
8	occupy	country	stop
9	past year	recovery	epidemic
10	automatically	fake	possible
11	prosecute	rally	work
12	arrest	region	uninsured
13	control	whistleblower	infection
14	subside	claim	hospital
15	leader	hate	stay

จากตารางที่ 2 จะทราบถึงจำนวนหัวข้อ หรือประเด็นสำคัญที่เหมาะสม ที่ให้ค่าความสับสน (Perplexity) ต่ำที่สุด เท่ากับ 3 หัวข้อ โดยมีการแปลผลและตีความผลลัพธ์จากคำสำคัญของแต่ละหัวข้อได้ดังนี้

หัวข้อที่ 1 : Arrestment (การต่อต้านผู้ป่วยที่ติดเชื้อโดยให้จับกุม หรือควบคุมผู้ป่วยไม่ให้อยู่ร่วมกันในสังคม)

ตัวอย่างทวีต : He should be arrested, this is the lowest of the fucking low. The dregs of humanity.

หัวข้อที่ 2 : Disgusting (การรังเกียจผู้ป่วยที่ติดเชื้อซึ่งถูกกักตัวอยู่ใกล้บริเวณถิ่นฐานที่อยู่)

ตัวอย่างทวีต : Shame on you! Your country is infected and you sit on the sidelines, wet your pants (literally) and whine.

หัวข้อที่ 3 : Damnation (การประณาม หรือสาปแช่งผู้ป่วยที่ติดเชื้อให้เสียชีวิต เพื่อไม่ให้ไวรัสแพร่กระจาย)

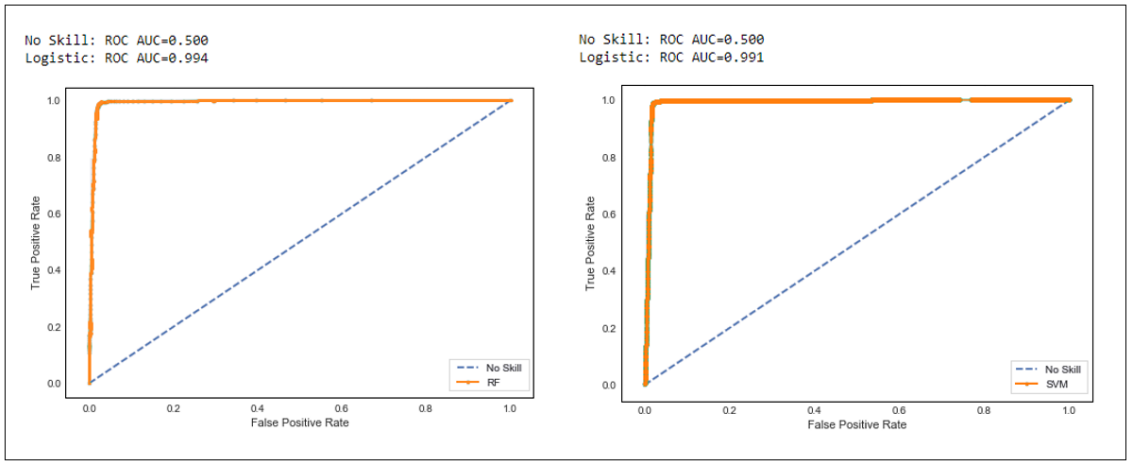
ตัวอย่างทวีต : Had you been smart enough to do the job we wouldn't have covid19 running wild. Every single death is on your head. You were too stupid to listen to China when they were telling you this is dangerous.

4.2 ผลลัพธ์การจัดแบ่งแยกทวีต

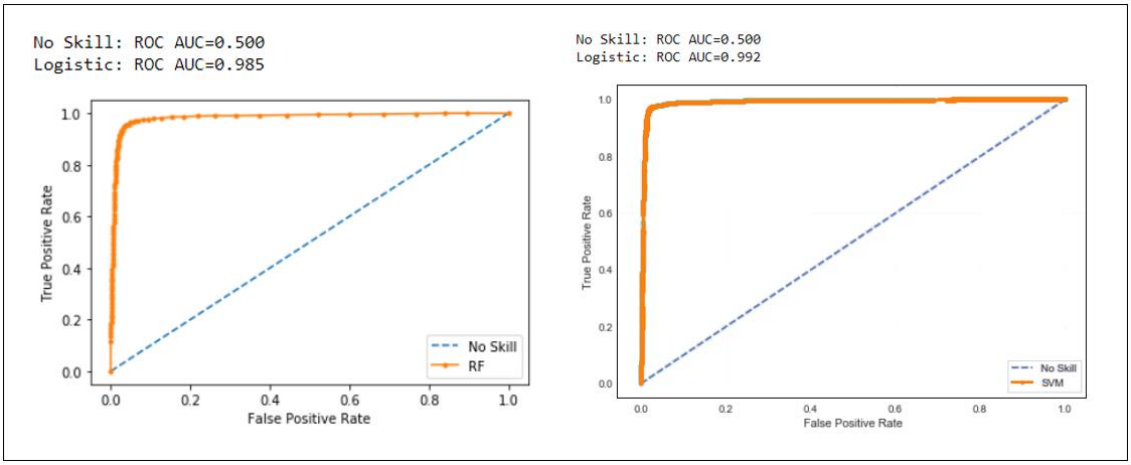
ในการแก้ปัญหาข้อมูลไม่สมดุล (Resampling Imbalanced Data) ผู้วิจัยได้ทดสอบการทำงานของอัลกอริทึมกับข้อมูลขนาดเล็ก (Small Data) โดยทำการสุ่มลดข้อมูล (Undersampling) ได้ข้อความทวีตที่เกลียดกลัวคนต่างชาติและไม่เกลียดกลัวคนต่างชาติจำนวนประเภทละ 4,108 ทวีต ซึ่งให้ค่าความแม่นยำ F1-Score น้อยกว่าการสุ่มข้อมูลเพิ่มและลด (Resampling) ประมาณ 15-20% ผู้วิจัยจึงเลือกทำการทดสอบกับข้อความที่เป็นทวีตที่เกลียดกลัวคนต่างชาติจำนวน 20,540 ทวีต และข้อความที่เป็นทวีตที่ไม่เกลียดกลัวคนต่างชาติจำนวน 24,648 ทวีต

ตารางที่ 3 ค่าความแม่นยำของตัวแบบจัดประเภทสำหรับการแปลงเชิงปริมาณ (Word Embedding) ทั้ง 3 วิธี

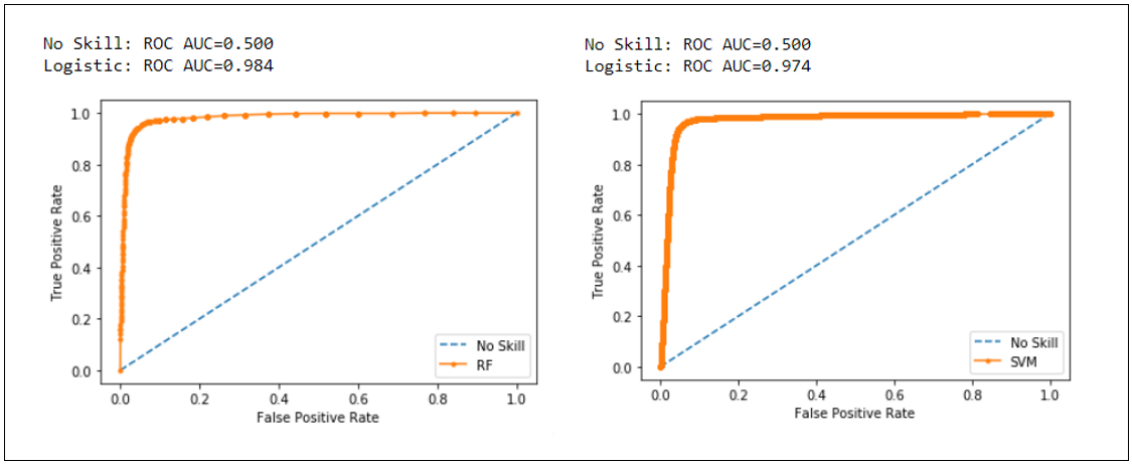
Word Embedding	Evaluation Metrics	Classifiers	
		Random Forest	Support Vector Machine
TF-IDF	Accuracy	0.99	0.98
	Precision	0.35	0.23
	Recall	0.65	0.99
	F1-Score	0.45	0.37
	ROC	0.99	0.99
Word2vec	Accuracy	0.99	0.99
	Precision	0.29	0.28
	Recall	0.68	0.93
	F1-Score	0.41	0.43
	ROC	0.99	0.99
GloVe	Accuracy	0.99	0.95
	Precision	0.28	0.09
	Recall	0.65	0.95
	F1-Score	0.39	0.17
	ROC	0.98	0.97



รูปที่ 4 กราฟ ROC Curve ของ 2 อัลกอริทึม เมื่อใช้การแปลงเชิงปริมาณด้วยวิธี TF-IDF



รูปที่ 5 กราฟ ROC Curve ของ 2 อัลกอริทึม เมื่อใช้การแปลงเชิงปริมาณด้วยวิธี Word2Vec



รูปที่ 6 กราฟ ROC Curve ของ 2 อัลกอริทึม เมื่อใช้การแปลงเชิงปริมาณด้วยวิธี GloVe

จากตารางที่ 3 การจัดแบ่งแยกทวีตด้วย 2 อัลกอริทึม พบว่า อัลกอริทึมแรนดอมฟอเรสต์ (Random Forest) เมื่อใช้การแปลงเชิงปริมาณ 3 วิธี คือ TF-IDF ให้ค่าความแม่นยำ F1-Score เท่ากับ 0.45 Precision เท่ากับ 0.35 Recall เท่ากับ 0.65, Word2vec ให้ค่าความแม่นยำ F1-Score เท่ากับ 0.41 Precision เท่ากับ 0.29 Recall เท่ากับ 0.68 และ GloVe ให้ค่าความแม่นยำ F1-Score เท่ากับ 0.39 Precision เท่ากับ 0.28 Recall เท่ากับ 0.65 ซึ่งจะเห็นได้ว่าค่า Precision ของวิธี TF-IDF ให้ค่ามากที่สุด รองลงมาเป็นวิธี Word2Vec และ GloVe ให้ค่าไม่แตกต่างกัน ในส่วนของค่า Recall, F1-Score และ ROC ของทั้ง 3 วิธีให้ค่าไม่แตกต่างกัน ส่วนอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เมื่อใช้การแปลงเชิงปริมาณ 3 วิธี คือ TF-IDF ให้ค่าความแม่นยำ F1-Score เท่ากับ 0.37 Precision เท่ากับ 0.23 Recall เท่ากับ 0.99, Word2vec ให้ค่าความแม่นยำ F1-Score เท่ากับ 0.43 Precision เท่ากับ 0.28 Recall เท่ากับ 0.93 และ GloVe ให้ค่าความแม่นยำ F1-Score เท่ากับ 0.17 Precision เท่ากับ 0.09 Recall เท่ากับ 0.95 ซึ่งจะเห็นได้ว่าค่า Precision ของวิธี Word2Vec ให้ค่ามากที่สุด รองลงมาเป็นวิธี TF-IDF และ GloVe ตามลำดับ ในส่วนของค่า Recall ของวิธี TF-IDF ให้ค่ามากที่สุด รองลงมาเป็นวิธี Word2Vec และ GloVe ให้ค่าไม่แตกต่างกัน และค่า F1-Score ของวิธี Word2Vec ให้ค่ามากที่สุด รองลงมาเป็นวิธี TF-IDF และ GloVe ตามลำดับ และค่า ROC ของทั้ง 3 วิธี ให้ค่าไม่แตกต่างกัน

5. สรุปและอภิปรายผลการศึกษา

จากผลการศึกษาการสร้างตัวแบบหัวข้อ และการหาจำนวนประเด็นสำคัญจากการปรับพารามิเตอร์ สรุปได้ว่าจำนวนประเด็นสำคัญที่เหมาะสมสำหรับการเกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 เท่ากับ 3 หัวข้อ ที่ให้ค่าความสับสน (Perplexity) ต่ำที่สุด เท่ากับ 114186.86 ได้แก่ (1) Arrestment (การต่อต้านผู้ป่วยที่ติดเชื้อโดยให้จับกุม หรือควบคุมผู้ป่วยไม่ให้อยู่ร่วมกันในสังคม) (2) Disgusting (การรังเกียจผู้ป่วยที่ติดเชื้อซึ่งถูกกักตัวอยู่ใกล้บริเวณถิ่นฐานที่อยู่) (3) Damnation (การประณาม หรือสาปแช่งผู้ป่วยที่ติดเชื้อให้เสียชีวิต เพื่อไม่ให้ไวรัสแพร่กระจาย) ซึ่งได้มาจากการแปลผลและตีความผลลัพธ์จากคำสำคัญของแต่ละหัวข้อ ผู้วิจัยพบความเกี่ยวข้องของแต่ละหัวข้อ คือ ประโยคที่พูดถึงจะทวีตไปในเชิงเหยียดหยาม ดูหมิ่น และมีการใช้คำหยาบคาย ซึ่งจากการศึกษามีทวีตที่เป็นประเภทการเกลียดกลัวคนต่างชาติจำนวนน้อยมาก อาจเป็นเพราะสังคมปัจจุบันเริ่มให้ความสำคัญกับเรื่องนี้ และมีการรณรงค์ให้หยุดกลั่นแกล้ง ยุติความความรุนแรง (Stop Bullying) ในหลาย ๆ ประเทศ จากงานวิจัยอื่น ๆ ที่เกี่ยวข้องก็ได้มีการใช้วิธีการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation - LDA) เป็นการสร้างตัวแบบหัวข้อเช่นกัน ซึ่งได้หาประเด็นสำคัญเกี่ยวกับความวิตกกังวลระหว่างการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา หัวข้อที่ได้คือ (1) Origin of COVID-19, (2) Source of the Novel Coronavirus, (3) Impact of COVID-19 on People and Countries, (4) Methods for Decreasing the Spread of COVID-19 ซึ่งแต่ละหัวข้อก็จะมีคำที่คล้ายกับคำในงานวิจัยนี้ เช่น Death และ Fear [25] โดยส่วนมากงานวิจัยอื่น ๆ จะวิเคราะห์ความรู้สึกจากข้อความบนทวิตเตอร์ระหว่างการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา ซึ่งพูดเกี่ยวกับผู้ป่วย, การติดเชื้อที่เพิ่มขึ้น, วัคซีน [26], การป้องกัน และผลกระทบทางเศรษฐกิจ [27] แต่งานวิจัยนี้มีการเจาะลึกถึงกลุ่มคนที่เกลียดกลัวคนต่างชาติ ดังนั้นหัวข้อที่ได้จึงมีความแตกต่าง มีคำที่พูดไปในเชิงเหยียดหยาม

จากผลการศึกษาตัวแบบจัดประเภท สรุปได้ว่าอัลกอริทึมแรนดอมฟอเรสต์ (Random Forest) เมื่อใช้การแปลงเชิงปริมาณด้วยวิธี TF-IDF ให้ค่าความแม่นยำ F1-Score, Recall และ ROC ไม่แตกต่างกัน เมื่อเปรียบเทียบกับวิธี Word2vec และ GloVe ทั้งนี้เมื่อพิจารณาด้วยค่า Precision วิธี TF-IDF ให้ค่าความแม่นยำสูงที่สุดเท่ากับ 0.35 ดังนั้น อัลกอริทึมแรนดอมฟอเรสต์ (Random Forest) เมื่อใช้ TF-IDF เป็นอัลกอริทึมที่เหมาะสมที่สุดกับการจัดประเภทข้อความการเกลียดกลัวและไม่เกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 ส่วนอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เมื่อใช้การแปลงเชิงปริมาณ

ด้วยวิธี Word2vec ให้ค่าความแม่นยำ F1-Score และ Precision สูงที่สุดเท่ากับ 0.43 และ 0.28 ตามลำดับ เมื่อเปรียบเทียบกับวิธี TF-IDF และ GloVe แม้ว่า Recall ของวิธี Word2vec จะมีค่าต่ำกว่าวิธี TF-IDF ส่วนค่า ROC มีค่าไม่แตกต่างกัน ทั้งนี้เมื่อพิจารณาด้วยค่า F1-Score ซึ่งเป็นค่าเฉลี่ยฮาร์โมนิกของ Precision และ Recall แล้ว วิธี Word2vec จึงมีประสิทธิภาพมากที่สุด ดังนั้น อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เมื่อใช้ Word2vec เป็นอัลกอริทึมที่เหมาะสมที่สุดกับการจัดประเภทข้อความการเกลียดกลัวและไม่เกลียดกลัวคนต่างชาติบนทวิตเตอร์ในช่วงการแพร่ระบาดของโรคติดเชื้อไวรัสโคโรนา 2019 จากการศึกษา ผู้วิจัยได้ทดสอบการทำงานของอัลกอริทึมกับข้อมูลขนาดเล็ก ให้ค่าความแม่นยำ F1-Score น้อยกว่าประมาณ 15-20% และผู้วิจัยไม่สามารถทดสอบกับข้อมูลขนาดใหญ่ได้ เนื่องจากประสิทธิภาพในการประมวลผลของคอมพิวเตอร์ (Hardware) ไม่เพียงพอ ซึ่งผู้วิจัยคาดว่าหากข้อมูลมีขนาดใหญ่ขึ้น และคอมพิวเตอร์ (Hardware) มีประสิทธิภาพในการประมวลผลมากเพียงพอ จะสามารถให้ค่าความแม่นยำเพิ่มขึ้นได้

ผลการศึกษาข้างต้นอาจเป็นแนวทางให้ทวิตเตอร์สร้างระบบการคัดกรองข้อความทวีตการเกลียดกลัวคนต่างชาติโดยดูจากหัวข้อ หรือประเด็นสำคัญที่พูดถึงในกลุ่มผู้ที่เกลียดกลัวคนต่างชาติซึ่งอาจก่อให้เกิดความรุนแรง และยังเป็นแนวทางในการปรับทัศนคติผู้ที่ทวีตข้อความในเชิงเกลียดกลัวคนต่างชาติของแต่ละประเทศ และออกมาตรการแก้ไข หรือป้องกันความรุนแรงที่จะเกิดขึ้นได้

6. ข้อเสนอแนะ

1. เป็นแนวทางในการคัดกรองข้อความบนทวิตเตอร์ที่มีคำตรงกับคำสำคัญของแต่ละหัวข้อ หรือประเด็นสำคัญการเกลียดกลัวคนต่างชาติทั้ง 3 ประเด็น คือ (1) Arrestment (2) Disgusting (3) Damnation และยังสามารถเป็น Sensor ให้ผู้ที่มีส่วนเกี่ยวข้องเข้าไปพิจารณาหากพบข้อความทวีตที่เกี่ยวข้องกับประเด็นสำคัญดังกล่าว

2. งานวิจัยนี้ศึกษาเฉพาะข้อความการเกลียดกลัวและไม่เกลียดกลัวคนต่างชาติบนทวิตเตอร์เท่านั้น สำหรับงานวิจัยในอนาคตสามารถใช้ข้อมูลบนสื่อสังคมออนไลน์อื่น ๆ เช่น Facebook, Instagram, Youtube ฯลฯ มาพิจารณาร่วมด้วย เนื่องจากลักษณะข้อความของแต่ละสื่อสังคมออนไลน์มีความแตกต่างกัน

3. งานวิจัยในอนาคตอาจมีการนำข้อมูลอื่น ๆ ใน Twitter มาใช้ร่วมด้วย เช่น Location ซึ่งสามารถนำมาใช้ประโยชน์ในด้านอื่น ๆ ได้อีก

4. หากมีคอมพิวเตอร์ (Hardware) ที่มีประสิทธิภาพในการประมวลผลมากเพียงพอ ก็อาจจะช่วยเพิ่มความแม่นยำของตัวแบบได้ และยังสามารถศึกษาอัลกอริทึม Deep Learning และวิธีการแปลงเชิงปริมาณ (Word Embedding) อื่น ๆ เพิ่มเติมที่อาจจะมีประสิทธิภาพและเหมาะสมกับข้อมูลมากกว่า

7. เอกสารอ้างอิง

[1] World Health Organization, “Coronavirus disease (COVID-19) questions and answers,” [Online]. Available: https://www.who.int/thailand/emergencies/novel-coronavirus-2019/q-a-on-p-19. [Accessed 9 October 2020].

[2] J. Bauomy, “COVID-19 and xenophobia: Why outbreaks are often accompanied by racism,” [Online]. Available: https://www.euronews.com/2020/03/05/covid-19-and-xenophobia-why-outbreaks-are-often-accompanied-by-racism. [Accessed 9 October 2020].

[3] A. Kelly, “Attacks on Asian Americans skyrocket to 100 per day during coronavirus pandemic,” [Online]. Available: https:// thehill.com/changing-america/respect/equality/490373-attacks-onasian-americans-at-about-100-per-day-due-to. [Accessed 9 October 2020].

[4] J. B. Trauner, "Chinese as Medical Scapegoats, 1870-1905," [Online]. Available: https://www.foundsf.org/index.php?title=Chinese_as_Medical_Scapegoats%2C_1870-1905. [Accessed 9 October 2020].

[5] S. O. Cheng, "Xenophobia due to the coronavirus outbreak – a letter to the editor in response to The socio-economic implications of the coronavirus pandemic (COVID-19): A review," *International Journal of Surgery*., vol. 79, pp. 13-14, 2020.

[6] Anti-Defamation League, "Reports of Anti-Asian Assaults, Harassment and Hate Crimes Rise as Coronavirus Spreads," [Online]. Available: https://www.adl.org/blog/reports-of-anti-asian-assaults-harassment-and-hate-crimes-rise-as-coronavirus-spreads. [Accessed 9 October 2020].

[7] J. Huang and R. Liu, "Xenophobia in America in the Age of Coronavirus and Beyond," *Journal of Vascular and Interventional Radiology*., vol. 31, no. 7, pp. 1187-1188, 2020.

[8] N. V. Chawla, N. Japkowicz and A. Kolcz, "Editorial: Special Issue on Learning from Imbalanced Data Sets," *ACM SIGKDD Explorations Newsletter.*, vol. 6, no. 1, pp. 1-6, 2004.

[9] A. Estabrooks and N. Japkowicz, "A Mixture-of-Experts Framework for Learning from Imbalanced Data Sets," in *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, Cascais, Portugal, 2001.

[10] S. Cateni, V. Colla and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing*., vol. 135, pp. 32–41, 2014.

[11] U. Erra, S. Senatore, F. Minnella and G. Caggianese, "Approximate TF–IDF based on topic extraction from massive message stream using the GPU," *Information Sciences*., vol. 292, pp. 143–161, 2015.

[12] ปฏิญญา หัสกุล, "การวิเคราะห์ SMS ด้วยคุณสมบัติเฉพาะจาก Bag-of-Words และทำนายอัตราการคลิก," ปริญญามหาบัณฑิต, สาขาการวิเคราะห์ธุรกิจและวิทยาการข้อมูล, คณะสถิติประยุกต์, สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2563.

[13] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the* 26th International Conference *on Neural Information Processing Systems*, Lake Tahoe, Nevada, 2013.

[14] ศุภวัจน์ แต่รุ่งเรือง, "การตรวจเทียบภายนอกหาการลักลอกในงานวิชาการโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนและการวัดค่าความละม้ายของข้อความ," วิทยานิพนธ์ ปริญญาอักษรศาสตร์ดุษฎีบัณฑิต, สาขาวิชาภาษาศาสตร์, คณะอักษรศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย, 2560.

[15] J. Pennington, R. Socher and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, United States, 2014.

[16] A. Khatua, A. Khatua and E. Cambria, "A tale of two epidemics: Contextual Word2Vec for classifying twitter streams during outbreaks," *Information Processing & Management*., vol. 56, no. 1, pp. 247–257, 2019.

[17] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*., vol. 3, pp. 993-1022, 2003.

[18] I. Vayansky and S. A. P. Kumar, "A review of topic modeling methods," *Information Systems*., vol. 94, 2020.

[19] ผไทรัช สีดา, "การแยกประเด็นสำคัญโดยอัตโนมัติจากข้อความรีวิวอาหารภาษาไทย," ปริญญามหาบัณฑิต, สาขาปัญญาและการวิเคราะห์ธุรกิจ, คณะสถิติประยุกต์, สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2563.

[20] W. Ali, S. M. Shamsuddin and A. S. Ismail, "Web Proxy Cache Content Classification based on Support Vector Machine," *Journal of Artificial Intelligence*., vol. 4, no. 1, pp. 100-109, 2011.

[21] วลัยลักษณ์ สุขสมบูรณ์ และ สมชาย ปราการเจริญ, "การเปรียบเทียบประสิทธิภาพการจำแนกประเภทปัญหาสำหรับระบบถามตอบโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และเคเนียรเรสต์เนเบอร์," ใน *รายงานสืบเนื่องการประชุมวิชาการมหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน ครั้งที่ 7*, มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน, 2553.

[22] L. Breiman, "Random Forests," *Machine Learning*., vol. 45, no. 1, pp. 5-32, 2001.

[23] วัชรีวรรณ จิตต์สกุล, "การวิเคราะห์การจำแนกข้อความด้วยการเปรียบเทียบความเสถียรของอัลกอริทึม," ปริญญาดุษฎีบัณฑิต, สาขาวิชาเทคโนโลยีสารสนเทศ, คณะเทคโนโลยีสารสนเทศ, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2560.

[24] S. Prabhakaran, "LDA in Python - How to grid search best topic models," [Online]. Available: https://www.machinelearningplus.com/nlp/topic-modeling-python-sklearn-examples/. [Accessed 9 October 2020].

[25] A. Abd-Alrazaq, D. Alhuwail, M. Househ, M. Hamdi and Z. Shah, "Top Concerns of Tweeters During the COVID-19 Pandemic: Infoveillance Study," *Journal of Medical Internet Research*., vol. 22, no. 4, pp. 1-9, 2020.

[26] S. V. Praveen, R.Ittamalla and G. Deepak, "Analyzing Indian general public's perspective on anxiety, stress and trauma during Covid-19 - A machine learning study of 840,000 tweets," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*., vol. 15, no. 3, pp. 667-671, 2021.

[27] K. Garcia and L. Berton, "Topic detection and sentiment analysis in Twitter content related to COVID-19 from Brazil and the USA," *Applied Soft Computing Journal*., vol. 101, no. 1, 2020.