การวิเคราะห์ความคิดเห็นของประชาชนไทยเกี่ยวกับวัคซีนโควิด-19 โดยใช้เทคนิคเหมืองข้อมูลข้อความ

พรพิมล ชัยวุฒิศักดิ์[*1]
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
1 ถนนฉลองกรุง แขวงลาดกระบัง เขตลาดกระบัง กรุงเทพฯ 10520

Received: 01 April 2025; Revised: 26 May 2025; Accepted: 28 May 2025

บทคัดย่อ

การศึกษานี้ใช้เทคนิคเหมืองข้อมูลข้อความเพื่อวิเคราะห์ความคิดเห็นของประชาชนไทยเกี่ยวกับวัคซีนโควิด-19 ที่แตกต่างกัน ข้อมูลถูกเก็บจาก ทวิตเตอร์ ตั้งแต่เดือนมีนาคมถึงตุลาคม 2564 รวมทั้งสิ้น 451,209 ข้อความที่เกี่ยวข้องกับวัคซีน AstraZeneca, Sinovac, Sinopharm, Pfizer และ Moderna การวิเคราะห์ใช้วิธีการจัดกลุ่มหัวข้อโดย Latent Dirichlet Allocation (LDA) และการวิเคราะห์ความรู้สึกโดยใช้ Multinomial Logistic Regression เพื่อจำแนกความคิดเห็นเกี่ยวกับวัคซีนแต่ละประเภท ผลการวิจัยพบว่าการสนทนาเกี่ยวกับวัคซีน AstraZeneca มุ่งเน้นไปที่การฉีดแบบผสม วัคซีน Sinovac มีการกล่าวถึงการฉีดให้บุคคลมีชื่อเสียง วัคซีน Sinopharm พบปัญหาด้านการจอง วัคซีน Pfizer ถูกกล่าวถึงเรื่องการขาดแคลน และ Moderna มุ่งเน้นไปที่ปัญหาการจัดซื้อ ผลการวิเคราะห์ความรู้สึกพบว่า ความคิดเห็นส่วนใหญ่เป็นกลาง รองลงมาคือเชิงลบเกี่ยวกับปัญหาการขาดแคลนวัคซีน การบริหารจัดการของภาครัฐ และผลข้างเคียง ผลลัพธ์ของงานวิจัยนี้สามารถนำไปใช้เพื่อปรับปรุงนโยบายด้านสาธารณสุขและการสื่อสารเพื่อเสริมสร้างความมั่นใจในวัคซีนของประชาชน

คำสำคัญ: เหมืองข้อมูลข้อความ, การประมวลภาษาธรรมชาติ, การจัดกลุ่มหัวข้อ, การวิเคราะห์ความรู้สึก, วัคซีนโควิด-19

* Corresponding author. E-mail: pornpimol.ch@kmitl.ac.th
[1] ผู้ช่วยศาสตราจารย์ ภาควิชาสถิติ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Understanding what do Thai people think about different kinds of COVID-19 vaccines using text mining

Pornpimol Chaiwuttisak[*1]

King Mongkut's Institute of Technology Ladkrabang (KMITL),

1 Chalong Krung 1 Alley, Lat Krabang, Bangkok 10520, Thailand

Received: 01 April 2025; Revised: 26 May 2025; Accepted: 28 May 2025

Abstract

The research aimed to apply text mining techniques to analyze the opinion of Thai people about the Covid-19 vaccine. Thai Twitter data used in this study was searched through keywords related to Covid-19 vaccine brand names: AstraZeneca, Sinovac, Sinopharm, Pfizer, and Moderna from March to October 2021. There were 451,209 tweets. All the tweets collected were separated on Covid-19 vaccine brand names and then classified topics using with Latent Dirichlet Allocation (LDA) technique. Next, sentiment analysis of the tweets on different topics was done using the Multinomial Logistic Regression Model. As a result, tweets related to the AstraZeneca vaccine can be classified into 8 topics. Most of them illustrated the Mix and match of different Covid-19 vaccines on Thailand's vaccination policy, including the side effects. Sinovac vaccine can be classified into 9 topics. Most of the tweets stated that the celebrity getting vaccinated. Sinopharm vaccine can be classified into 10 topics. Most of the tweets talked about the problems of vaccine booking systems. Pfizer vaccine can be classified into 10 topics. Most of the tweets discussed the delay or shortage of place orders for vaccines, whereas Moderna vaccine can be classified into 4 topics. Most of them focused on the problems with the vaccine booking system. In addition, it revealed that most Thai people's opinions on social media are neutral polarity. Negative sentiment tweets constituted a larger share as compared to positive sentiment tweets about vaccines. Most of the topics in the tweets with negative sentiment involve insufficient vaccines, Covid-19 vaccine procurement management by the government Issues, the online systems of the alternative vaccines defined in Thailand, including side effects after vaccination, and even reviews of vaccinations of Thai celebrities or influencer people.

Keywords: text mining, natural language processing, topic modeling, sentiment analysis, COVID-19 vaccines

* Corresponding author. E-mail: pornpimol.ch@kmitl.ac.th

[1] Assistant Professor in School of Science, King Mongkut's Institute of Technology Ladkrabang

1. บทนำ

การระบาดเชื้อไวรัสโคโรนา สายพันธุ์ใหม่ หรือชื่ออย่างเป็นทางการว่า “โควิด-19” ส่งผลกระทบเศรษฐกิจในวงกว้าง โดยเฉพาะอุตสาหกรรมการท่องเที่ยว จำนวนนักท่องเที่ยวต่างชาติในปี 2563 รวมทั้งสิ้น 6,702,396 คน เทียบกับปี 2562 ที่มียอด 39,916,251 คน ลดลงไปถึง 83.21% และจำนวนนักท่องเที่ยวในปี 2564 รวมทั้งสิ้น 73,608 คน เทียบกับปี 2563 ลดลงถึง 98.90% [3] สาเหตุหลักมาจากการมาตรการงดเดินทางระหว่างประเทศ การปิดสถานที่ท่องเที่ยวต่าง ๆ ในช่วงเวลาหนึ่ง ทำให้เศรษฐกิจไม่สามารถเดินหน้าต่อได้ อุตสาหกรรมอิเล็กทรอนิกส์ ยานยนต์ธุรกิจบางแห่งที่มีผลขาดทุนและปิดตัวลง และธุรกิจโรงแรมต้องหยุดชะงักเนื่องจากจำนวนนักท่องเที่ยวที่ลดน้อยลง ทำให้เกิดการลดลงของรายได้แรงงานที่สำคัญต่อภาคการส่งออกของไทย ซึ่งผลกระทบดังกล่าวเป็นลูกโซ่ไปทั้งระบบ ส่งผลให้เกิดการเลิกจ้างแรงงานเป็นจำนวนมาก ก่อให้เกิดภาวะแรงงานตกงานอัตราการว่างงานสูง

กิริฎา และกิตติพัฒน์ [6] กล่าวว่าปัจจัยหลักสำหรับการฟื้นตัวของเศรษฐกิจคือการฉีดวัคซีนป้องกันโควิด-19 วัคซีนจะช่วยกระตุ้นให้ร่างกายสร้างภูมิคุ้มกันต่อเชื่อไวรัส และลดการแพร่ระบาดการติดต่อโรค ลดความรุนแรงของอาการและลดการเสียชีวิต วัคซีนจึงเป็นเครื่องมือสำคัญที่จะช่วยควบคุมโรค กระทรวงสาธารณสุขเปิดเผยจำนวนผู้ที่ได้รับการฉีดวัคซีนสะสมในประเทศไทย ผู้ที่ได้รับวัคซีน เข็มที่ 1 คิดเป็นร้อยละ 63.8 ของประชากรทั้งหมด ผู้ที่ได้รับวัคซีน เข็มที่ 2 คิดเป็นร้อยละ 46.4 ของประชากรทั้งหมด และผู้ที่ได้รับวัคซีน เข็มที่ 3 คิดเป็นร้อยละ 3.6 ของประชากรทั้งหมด นับตั้งแต่ 28 กุมภาพันธ์ 2564 ถึง 31 ตุลาคม 2564 [1]

ในช่วงสถานการณ์การแพร่ระบาดของโควิด-19 ทำให้คนไทยมีความเครียดสูง เกิดอาการวิตกกังวล การหาข้อมูลหรือข้อเท็จจริงเกี่ยวกับการแพร่ระบาด การติดเชื้อ การดูแลรักษา การป้องกันจากโควิด-19 และการฉีดวัคซีนป้องกันโควิด-19 โดยปัจจุบันวัคซีนป้องกันโควิด-19 ในประเทศไทยที่ผ่านการอนุมัติแล้วมี 6 ชนิดด้วยกัน คือ AstraZeneca, Sinovac, Sinopharm, Johnson & Johnson, Moderna และ Pfizer ซึ่งวัคซีนของแต่ละชนิดมีการใช้เทคโนโลยีที่แตกต่างกัน มีทั้งเทคโนโลยีการผลิตวัคซีนแบบเชื้อตาย (Inactivated Vaccine) เทคโนโลยีการผลิตแบบไวรัลเวกเตอร์ (Viral Vector Vaccines) และเทคโนโลยีการผลิตสมัยใหม่ mRNA Vaccines โดยข่าวเกี่ยวกับวัคซีนที่นำเสนอในแง่มุมต่าง ๆ ส่งผลให้ประชาชนเกิดความรู้สึกกลัว ไม่เชื่อใจและกังวล จึงกลายเป็นเรื่องยากสำหรับการตัดสินใจของประชาชนที่จะเข้ารับการฉีดวัคซีนป้องกันโควิด-19 โดยทวิตเตอร์เผยสถิติ พบว่า คนไทยมีการพูดคุยเกี่ยวกับโควิด-19 ในประเทศไทยบนทวิตเตอร์ ระหว่างวันที่ 1 มีนาคม 2020 ถึง 1 สิงหาคม 2021 มีจำนวนทวีตที่เกี่ยวกับโควิด-19 มากกว่า 73 ล้านทวีต [19] จะเห็นได้ว่าทวิตเตอร์เป็นแหล่งที่ประชาชนไทยจำนวนมากเลือกที่จะเข้ามาหาข้อมูลและร่วมแชร์ประสบการณ์ต่าง ๆ หลายคนจึงหันมาใช้ทวิตเตอร์เป็นสังคม (Community) ในการสื่อสาร การประชาสัมพันธ์ให้ความรู้ ข้อเสนอแนะเกี่ยวกับการแพร่ระบาดของโควิด-19 และวัคซีนป้องกันโควิด-19

งานวิจัยของ Lyu et al. [18] ได้มีการสำรวจเกี่ยวกับความคิดเห็นของประชาชนสหรัฐอเมริกาเกี่ยวกับโรคระบาดโควิด-19 ศึกษาเพื่อสร้างความรู้ความเข้าใจถึงขั้วความรู้สึกของประชาชน ผลการศึกษาพบว่า จำนวนทวีตที่เกี่ยวกับการฉีดวัคซีนโควิด-19 ขับเคลื่อนด้วยเหตุการณ์สำคัญ ซึ่งส่วนใหญ่เป็นเหตุการณ์สำคัญในการพัฒนาวัคซีนและสายพันธุ์ใหม่ของไวรัส การวิเคราะห์ขั้วความรู้สึกแสดงให้เห็นว่าความรู้สึกทั่วไปของประชาชนเปลี่ยนไปตามสถานการณ์ของข่าวที่เปลี่ยนแปลง และ Na et al. [20] พบว่าประชาชนของสหราชอาณาจักรและสหรัฐอเมริกา ส่วนใหญ่มีความต้องการได้รับวัคซีนและรู้สึกสบายใจหลังจากได้รับวัคซีน ความกังวลของประชาชนเกี่ยวกับวัคซีนป้องกันโควิด-19 ส่วนใหญ่มาจากกรณีการเสียชีวิตของผู้ที่ได้รับวัคซีน SV et al. [22] ได้ทำการศึกษาหัวข้อเรื่องมุมมองของชาวอินเดียเกี่ยวกับผลข้างเคียงของวัคซีนโควิด-19 บนทวิต จากผลงานวิจัยสามารถสรุปได้ว่า เกือบ 78.5% ของทวีตที่โพสต์โดยชาวอินเดียเกี่ยวกับผลข้างเคียงของวัคซีนโควิด-19 สรุปโดยรวมได้ว่า ประชาชนมีความกลัวในประสิทธิภาพของวัคซีนโควิด-19 และกลัวการเสียชีวิตจากผลข้างเคียงของวัคซีน ในขณะที่การศึกษาความคิดเห็นของประชาชนไทยที่มีต่อวัคซีน

ป้องกันโควิด-19 รวมถึงการวิเคราะห์ขั้วความรู้สึกที่มีต่อวัคซีนป้องกันโควิด-19 แต่ละชนิดยังไม่ค่อยพบเห็นมากนัก

ดังนั้นผู้วิจัยจึงมีความสนใจที่จะศึกษาถึงความเข้าใจความคิดเห็นของคนไทยที่มีต่อวัคซีนป้องกันโควิด-19 ชนิดต่าง ๆ โดยอาศัยการทำเหมืองข้อความบนทวิตเตอร์ เนื่องจากจำนวนผู้ใช้ทวิตเตอร์ในประเทศไทยสูงถึง 7.35 ล้านบัญชี [14] ด้วยข้อมูลมีอยู่มากและมีขนาดใหญ่ คณะผู้วิจัยซึ่งใช้เทคนิคการวิเคราะห์เหมืองข้อความ (Text Mining) การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation : LDA) เพื่อจัดกลุ่มคำตามหัวข้อและประเด็นต่าง ๆ และใช้การวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis) เพื่อจำแนกตามลักษณะอารมณ์ความคิดเห็นว่าไปในทิศทางบวก ลบ หรือเป็นกลาง

2. การทำเหมืองข้อความ (Text Mining)

Feldman and Sanger [15] กล่าวว่าการทำเหมืองข้อความเป็นกระบวนการที่ใช้แก้ปัญหาภาวะข้อมูลท่วมท้น (Information Overload) โดยใช้เทคนิคจากการเรียนรู้ด้วยเครื่อง (Machine Learning) การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การสืบค้นสารสนเทศ (Information Retrieval) และการจัดการความรู้ (Knowledge Management) การทำเหมืองข้อความได้แรงบันดาลใจและทิศทางงานวิจัยเกี่ยวกับการทำเหมืองข้อมูล (Data Mining) จึงมีกระบวนการคล้ายคลึงกับการทำเหมืองข้อมูลอย่างมาก

Petrović [21] กล่าวว่าเหมืองข้อความเป็นฟิลด์ย่อยของการทำเหมืองข้อมูล และเป็นส่วนหนึ่งของการประมวลผลภาษาธรรมชาติ การทำเหมืองข้อความเป็นสาขาสหวิทยาการที่รวมสาขาต่าง ๆ เช่น การเรียนรู้ด้วยเครื่อง สถิติ และการประมวลผลทางภาษา โดยเน้นการทำงานด้านการประเภทเอกสาร (Text Categorization) การจัดกลุ่มเอกสาร (Text Clustering) การวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis) การสรุปเอกสาร (Document Summarization) และการหาความสัมพันธ์ (Entity Relation Modeling)

กานดา [5] กล่าวว่าสำหรับการวิเคราะห์ข้อความเป็นกระบวนการค้นหาและสกัดความรู้จากฐานข้อมูลขนาดใหญ่ (Big Data) เพื่อให้ได้สารสนเทศที่มีประโยชน์ โดยข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลที่มีลักษณะเป็นภาษาธรรมชาติ จึงเรียกว่าการวิเคราะห์เหมืองข้อความ

กระบวนการทำเหมืองข้อความมีความคล้ายคลึงกับการค้นหาความรู้จากฐานข้อมูล และมีการเอาเทคนิคจากการประมวลผลภาษาธรรมชาติเข้ามาช่วย โดยสามารถแบ่งกระบวนการทำงานออกเป็น 5 ขั้นตอนสำคัญ [13] ดังนี้

1. การเลือกข้อมูล (Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาใช้ในการทำเหมืองข้อความ รวมถึงการนำข้อมูลที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาข้อมูลให้ตรงตามที่ขอบเขตต้องการศึกษากำหนดไว้

2. การเตรียมข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่ใช้ในการทำเหมืองข้อความ ในการรวบรวมข้อมูลควรตรวจสอบข้อมูล เพื่อดูความถูกต้องของข้อมูล และเหมาะสม โดยนำข้อมูลที่ไม่ถูกต้องออกหรือเป็นขั้นตอนที่อาจต้องแก้ไขข้อมูลก่อนนำไปใช้งาน ประกอบด้วย 3 กระบวนการ ได้แก่ การตัดคำ (Word Segmentation) การกำจัดคำหยุด (Stop Words Removal) การแทนข้อความด้วยถุงคำ (Bag of Words)

3. การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลเหมาะสมและตรงกับรูปแบบที่จะประมวลผลต่อไป โดยงานวิจัยนี้ใช้วิธี Term Frequency/Inverse Document Frequency (TF/IDF)

4. การทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนการประมวลผล โดยใช้อัลกอริทึมต่าง ๆ สกัดข้อมูลที่มีอยู่แล้วในฐานข้อมูลที่มีจำนวนมากเพื่อค้นหารูปแบบและความสัมพันธ์ที่ซ้อนอยู่ในชุดข้อมูลนั้นซึ่งการทำเหมืองข้อมูลมีเทคนิคอยู่หลายเทคนิค ซึ่งใช้หลายเทคนิคประกอบกันเพื่อค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery in Database : KDD) ให้ออกมารูปแบบ เพื่อให้ได้สารสนเทศที่ยังไม่รู้ เป็นสารสนเทศที่มีเหตุผล และสามารถนำไปใช้ได้ซึ่งเป็นสิ่งสำคัญที่จะช่วยในการทำธุรกิจ

5. การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความและ

การประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรมีการนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจง่าย

2.1 การประมวลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติ (Natural Language Processing : NLP) เป็นกระบวนการที่จะทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์โดยรับอินพุตเป็นข้อความในภาษาหนึ่ง ๆ ให้คอมพิวเตอร์สามารถเข้าใจข้อความได้เช่นเดียวกับมนุษย์และสามารถนำไปประมวลผลได้ ซึ่งประกอบด้วยหลายขั้นตอนย่อย ตั้งแต่ตอนการจัดเตรียมข้อมูล รวมไปถึงการแปลงข้อมูลให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถนำไปประมวลผลต่อได้ [8]

2.1.1 การทำความสะอาดข้อมูล (Data Cleaning)

ในส่วนของการทำความสะอาดข้อมูล จะทำการลบ อิโมจิทั้งหมด ลบลิงก์ URLที่เป็นสแปม หรือลิงก์ที่ไม่เกี่ยวข้องออก ลบสัญลักษณ์ อักษรพิเศษ และตัวอักษรที่นอกเหนือจากภาษาไทยทิ้ง เช่น “@, $, %, ฿, ?” เป็นต้น คัดแยกข้อความทวีตที่เป็นของสำนักงานข่าวกับบัญชีผู้ใช้ธรรมดาแยกออกจากกันคนละชุดเพื่อนำไปวิเคราะห์แยก ลบข้อความทวิตที่ซ้ำกันตรวจสอบทวีตที่ทวีตโดยบัญชีปลอมหรือบอท และตรวจสอบการเขียนคำย่อให้ถูกต้อง เช่น คำย่อชื่อเดือน อักษรย่อคำนำหน้านาม อักษรย่อหน่วยงานต่าง ๆ

2.1.2 การตัดคำ (Word Segmentation)

เนื่องจากประโยคในภาษาไทยมีวิธีการเขียนตัวหนังสือติดกันหมด ไม่มีการเว้นระหว่างคำด้วยช่องว่างเหมือนภาษาอังกฤษ เพื่อให้การจำแนกมีประสิทธิภาพจึงมีวิธีการตัดคำภาษาไทย โดยมีวิธีการดังนี้

1. การตัดคำโดยใช้กฎ (Rule-Base Approach) เป็นการตัดคำโดยตรวจสอบกฎเกณฑ์ทางอักขระวิธีที่อาศัยหลักไวยากรณ์ภาษาไทย เริ่มจากการตัดพยางค์เนื่องจากพยางค์มีรูปแบบที่แน่นอนมากกว่าคำ จากนั้นนำพยางค์มาเป็นเกณฑ์ในการกำหนดขอบเขตคำ ข้อดีคือ มีการทำงานรวดเร็ว ใช้ทรัพยากรในการประมวลผลน้อย ข้อจำกัดคือ ความถูกต้องและประสิทธิภาพขึ้นอยู่กับเกณฑ์กำหนดขอบเขตคำ

2. การตัดคำโดยใช้พจนานุกรม (Dictionary-Base Approach) ใช้การเปรียบเทียบคำกับคำที่จัดเก็บในพจนานุกรม แล้วนำข้อความที่ป้อนเข้ามาไปเปรียบเทียบอักขระกับคำในพจนุกรม ข้อดีคือ มีความรวดเร็วในการทำงานและใช้ทรัพยากรน้อย การถูกต้องการตัดคำในระดับพยางค์ค่อนข้างสูง แต่ข้อจำกัดคือ ความถูกต้องของการตัดคำค่อนข้างต่ำ และต้องการแหล่งข้อมูลขนาดใหญ่

3. การตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) เป็นการนำหลักทางสถิติและกลไกการเรียนรู้มาใช้การประมวลผลทางภาษามี 2 วิธี คือ วิธีความน่าจะเป็นโดยใช้ตัวแบบ ไตรแกรมกำกับหน้าที่ของคำในการหารูปแบบของการตัดคำและลำดับหมวดหมู่ ซึ่งต้องเตรียมคลังคำที่มีการตัดคำและกำกับหน้าที่ของคำไว้ล่วงหน้าและวิธีคุณลักษณะของคำที่ช่วยแก้ไขความผิดพลาดของการตัดคำที่อาศัยความน่าจะเป็นโดยนำคุณลักษณะคำที่มีความกำกวมมาช่วยเลือกการตัดคำที่ถูกต้องให้ได้จำนวนคำที่ไม่พบในพจนานุกรมน้อยที่สุด โดยการสร้างตัวแบบกลไกการเรียนรู้

2.1.3 การกำจัดคำหยุด (Stop Words Removal)

การกำจัดคำหยุดเป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ความหมายของคำหรือข้อความไม่เปลี่ยนแปลง คำหยุดจะปรากฏอยู่บ่อยครั้งในเอกสาร ถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการจำแนกหมวดหมู่ การกำจัดคำหยุดเป็นกระบวนการที่จะช่วยให้ขนาดของดัชนีลดลงและยังลดขนาดพื้นที่และเวลาในการประมวลผล ประเภทของคำที่เป็นคำหยุดในภาษาไทย ได้แก่ คำบุพบท คำสันธาน คำสรรพนาม คำวิเศษณ์ และคำอุทาน

ตารางที่ 1 คำหยุดในภาษาไทย

คำบุพบท	คำสันธาน	คำสรรพนาม	คำวิเศษณ์	คำอุทาน
แก่	และ	กระผม	ใกล้	555
ต่อ	คือ	ข้าพเจ้า	ครับ	กำ
ซึ่ง	จึง	ฉัน	ง่าย	เอ๊ะ

คำบุพบท	คำสันธาน	คำสรรพนาม	คำวิเศษณ์	คำอุทาน
ด้วย	แต่	เธอ	ด้วย	โธ่
เพื่อ	แล้ว	นาย	ที่สุด	โอ๊ย

จากตารางที่ 1 แสดงตัวอย่างคำหยุดในภาษาไทย อธิบายได้ดังนี้

1. คำบุพบท เป็นคำที่แสดงความสัมพันธ์ระหว่างคำหรือประโยค เพื่อให้ทราบว่าคำหรือกลุ่มคำที่ตามหลังคำบุพบทนั้นเกี่ยวข้องกับกลุ่มคำข้างหน้า
2. คำสันธาน เป็นคำที่ใช้เชื่อมประโยค หรือข้อความต่อข้อความ
3. คำสรรพนาม เป็นคำที่ใช้เรียกแทนชื่อผู้พูดหรือผู้ที่เรากล่าวถึง
4. คำวิเศษณ์ เป็นคำที่ใช้ขยายคำนาม สรรพนาม คำกริยา เพื่อให้ได้ใจความที่ชัดเจนมากขึ้น
5. คำอุทาน เป็นคำที่แสดงอารมณ์ของผู้พูดหรืออาจจะเป็นคำที่ใช้เสริมคำพูด ประโยค ต้องระวังการตัดคำอุทาน เนื่องจากคำอุทานบางคำอาจจะแสดงอารมณ์ในขณะนั้นออกมาด้วย

2.1.4 การแทนข้อความด้วยถุงคำ (Bag of Words)

การแทนข้อความด้วยถุงคำ (Bag of Words) เป็นวิธีการประมวลผลที่ใช้ในการแปลงข้อความให้อยู่ในรูปที่คอมพิวเตอร์นำไปใช้ประมวลผลได้อย่างง่ายที่สุด โดยแทนค่า 1 เมื่อคำเหล่านั้นปรากฏอยู่ในชุดข้อมูล และแทนค่า 0 เมื่อไม่มีคำเหล่านั้นปรากฏอยู่ในชุดข้อมูล และยังเป็นพื้นฐาน ของอัลกอริทึมอื่น ๆ โดยการแทนข้อความด้วยถุงคำเป็นการนำ Token ที่ได้จากการทำ Word Tokenization มาประยุกต์ใช้ร่วมกับความรู้ทางด้านสถิติในการรวบรวมข้อมูลของคำในข้อความโดยทำการสร้างคลังคำศัพท์จากชุดข้อมูลที่จะใช้ในการวิเคราะห์ โดยส่วนใหญ่คลังคำศัพท์นั้นจะถูกจัดเรียงตามลำดับของ Token ที่มีจำนวนมากที่สุดไปน้อยที่สุด จากนั้นทำการกำหนดคีย์หลักสำหรับแทนค่าของ Token นั้น ๆ เพื่อนำไปใช้ในการวิเคราะห์ต่อไป [8] ในส่วนของการใช้งานการแทนข้อความด้วยถุงคำนั้นยังมีปัญหาเรื่องการนับจำนวนคำศัพท์อย่างเดียวแต่ไม่ได้นับในส่วนของเอกสารที่คำนั้นปรากฏ ซึ่งส่งผลให้เกิดความสูญเสียความรู้บางส่วนในชุดข้อมูล

2.1.5 Term Frequency/Inverse Document Frequency (TF/IDF)

Term Frequency/Inverse Document Frequency (TF/IDF) ได้มีการพัฒนามาจาก Bag of Words ใช้หลักการคล้ายคลึงกันในการสร้างคลังคำศัพท์ Term Frequency/Inverse Document Frequency (TF/IDF) เป็นวิธีการให้น้ำหนักของคำ ซึ่งเป็นวิธีการทางด้านสถิติที่ใช้ประเมินความสำคัญของคำหรือคุณลักษณะนั้น ๆ โดยการพิจารณาค่าความสำคัญด้วยด้วยการนับความถี่ของคำ Term Frequency (TF) ที่ปรากฏในกลุ่มเอกสารทั้งหมด และ Inverted Document Frequency (IDF) คือ การวัดจำนวนคำที่หาได้โดยการนำจำนวนเอกสารทั้งหมดในกลุ่มเอกสารมาหารด้วยจำนวนเอกสารที่พบคำ ทั้งนี้คำหรือคุณลักษณะที่มีค่าน้ำหนัก TF/IDF มาก แสดงว่าคำหรือคุณลักษณะนั้นมีประสิทธิภาพในการจำแนก [10] โดยมีสูตรคำนวณดังสมการที่ 1 และ 2

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{1}$$

โดยที่ $tf_{i,j}$ คือ จำนวนครั้งที่คำนั้น ๆ ปรากฏในเอกสาร
$n_{i,j}$ คือ จำนวนคำนั้น ๆ ในเอกสารที่ d_j
$\sum_k n_{i,j}$ คือ จำนวนของทุกคำในเอกสารที่ d_j

$$idf_j = \log \frac{|D|}{\left|\{d_j : t \in d_j\}\right|} \tag{2}$$

โดยที่ $|D|$ คือ จำนวนเอกสารทั้งหมดในกลุ่มเอกสาร
$\left|\{d_j : t \in d_j\}\right|$ คือ จำนวนเอกสารที่คำนั้น ๆ ปรากฏจะได้การคำนวณ **TF/IDF** ดังสมการที่ 3

$$tfidf_{i,j} = tf_{i,j} \times idf_j \tag{3}$$

2.1.6 Word Embedding

Word Embedding คือการสร้างเวกเตอร์คุณลักษณะ (Feature Vector) ขึ้นมาจาก Token ที่ได้ทำการสร้างไว้โดยทำการสร้างเวกเตอร์คุณลักษณะขึ้นมาจากประโยคหรือเอกสารเพื่อทำการสร้างคุณลักษณะที่อยู่ในรูปของตัวเลขที่สามารถ

นำไปใช้ในการคำนวณต่อได้ ซึ่งจุดเด่นของเวกเตอร์คุณลักษณะที่ได้จาก Word Embedding นั้นจะสามารถนำไปใช้คำนวณความคล้ายคลึงกับคำอื่น ๆ ในบริบทของคำที่แตกต่างกันได้

ในปัจจุบันได้มีการพัฒนาตัวแบบการทำ Word Embedding ต่าง ๆ ขึ้นมาให้เลือกใช้งานมากมาย ซึ่งแต่ละตัวแบบนั้นจะมีจุดเด่นที่ไม่เหมือนกัน โดยผู้วิจัยได้ทำการยกตัวอย่างมา 2 แบบ ได้แก่ Word2Vec และ GloVe

1. Word2Vec เป็นตัวแบบที่ใช้สร้างการฝังคำหรือแปลงคำให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งเวกเตอร์ความหมายของคำ ซึ่งมีแนวคิดว่า ความหมายของคำคำหนึ่งในประโยคนั้นมีความสัมพันธ์กับความหมายของคำที่อยู่รอบข้าง ซึ่งแนวคิดนี้ได้ถูกพัฒนาเป็นตัวแบบชื่อ Continuous Bag of Words (CBOW) และ Skip-gram ตามลำดับ โดยสามารถอธิบายได้ดังนี้

- ตัวแบบ Continuous Bag of Words (CBOW) ในการใช้งานผู้ใช้จะทำการสร้างเครือข่ายสมองแบบตื้น (Shallow Neural Network) โดยใช้งาน Word Vector เป็น Input Layer ซึ่งเชื่อมต่อเข้ากับ Hidden Layer จำนวน 1 ชั้น และทำการต่อเข้าสู่ Output Layer ที่มีจำนวนมิติเท่ากับ Input layer จากนั้นจะทำการฝึกฝนตัวแบบ (Training) ในรูปแบบ Classification

- ตัวแบบ Skip-gram ในการใช้งานจะทำการสร้าง เครือข่ายสมองแบบตื้นในลักษณะเดียวกับที่ใช้งานในตัวแบบCBOW โดยมีขนาด Input Layer และ Output Layer เท่ากัน และมี Hidden Layer เป็นจำนวน 1 ชั้น แต่แตกต่างจากตัวแบบ CBOW เพราะแทนที่จะใช้ Word Vectors ของคำต่าง ๆ ในบริบทของคำแต่ละคำมาเพื่อทำนายคำดังกล่าว ตัวแบบ Skip-gram เลือกที่จะใช้คำหนึ่ง ๆ ในการทำนายคำทุกคำที่อยู่ในบริบทของคำนั้นแทน

2. Global Vectors for Word Representation (GloVe) เป็นตัวแบบถดถอยล็อก-ไบลิเนียร์ (Log-Bilinear Regression) ที่รวม 2 ตัวแบบเข้าด้วยกัน คือ โกลบอลเมตริกซ์แฟคเตอร์ไรเซชัน (Global Matrix Factorization) และโลคอลคอนเท็กซ์วินโดว์ (Local Context Window) และเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) อัลกอริทึมสำหรับการแปลงคำ ให้อยู่ในรูปแบบของเวกเตอร์ ซึ่งตัวแบบ GloVe ได้รับการสอนให้เรียนรู้สถิติการเกิดร่วมกันของคำจากคลังข้อมูล และแสดงผลลัพธ์เป็นเวกเตอร์ของคำในปริภูมิเวกเตอร์

2.2 การเลือกคุณลักษณะ (Feature Selection)

วิธีเบื้องต้นในการลดขนาดของเอกสารคือการใช้การนำคำที่ไม่มีนัยสำคัญออก กับการทำรากศัพท์แล้วยังไม่เพียงพอ ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่ โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี การลดขนาดเอกสารจึงเป็นขั้นตอนหนึ่งที่สำคัญ การคัดเลือกคุณลักษณะแบ่งออกเป็น 3 วิธี ได้แก่ Filter Approach, Wrapper Approach และ Embedded Methods

1. วิธี Filter Approach เป็นการคัดเลือกคุณลักษณะที่สำคัญ และจะเลือกคุณลักษณะโดยเรียงลำดับตามค่าน้ำหนักที่คำนวณได้ ซึ่งมีเทคนิคในการคำนวณดังนี้

1.1 เทคนิค Information Gain (IG) เป็นเทคนิคที่ใช้ในการชี้วัดเพื่อเลือกคุณลักษณะข้อมูลที่จำนวนน้อยที่สุดที่เหมาะสมในการนำไปใช้การแบ่งข้อมูลเป็นข้อมูลชุดย่อย โดยการคำนวณหาค่า Gain สำหรับแต่ละมิติข้อมูลซึ่งมิติข้อมูลใดมีค่า Gain สูงสุด จะถูกคัดเลือกให้เป็นกลุ่มย่อย สมการที่ 4 แสดงการคำนวณค่า Information Gain หรือค่า Entropy หมายถึงค่าเฉลี่ยของปริมาณข้อมูลที่ต้องการในการระบุถึงหมวดหมู่ของข้อมูลแถวของข้อมูล หนึ่ง ๆ ในชุดข้อมูลที่จะขึ้นกับอัตราส่วนของจำนวนแถวของข้อมูลที่สอดคล้องกับแต่ละหมวดหมู่ สมการที่ 5 แสดงการคำนวณค่า Entropy ของชุดมิติข้อมูลในแต่ละคุณลักษณะ A สมการที่ 6 เป็นการคำนวณหาค่า Gain สำหรับการพิจารณามิติข้อมูลคุณลักษณะ A

$$E(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \quad (4)$$

$$E_A(D) = \sum_{j=1}^{m} \frac{|D_j|}{D} \times E(D_j) \quad (5)$$

$$Gain(A) = E(D) - E_A(D) \quad (6)$$

โดยที่ p_i คือ ค่าความน่าจะเป็นที่แถวของข้อมูลหนึ่ง ๆ จะมีหมวดหมู่ c_i ซึ่งสามารถคำนวณได้จาก $\frac{|c_{i,D}|}{|D|}$

D คือ จำนวนข้อมูลในฐานข้อมูล

D_j คือ จำนวนข้อมูลในฐานข้อมูล D ที่ j ของคุณลักษณะ A

1.2 เทคนิค Chi-Square การประเมินค่าของคุณลักษณะโดยการใช้การคำนวณค่า Chi-Square ทางสถิติเพื่อศึกษาว่าการแจกแจงความถี่ของตัวแปรคุณลักษณะเป็นไปตามรูปแบบที่กำหนดไว้หรือไม่ ดังสมการที่ 7

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad (7)$$

โดยที่ $O_1, O_2, ..., O_n$ เป็นความถี่ของตัวแปรที่ได้จากการศึกษา

$E_1, E_2, ..., E_n$ เป็นความถี่ที่คาดหวัง (หรือความถี่ที่ควรจะเป็น)

2. วิธี Wrapper Approach เป็นการคัดเลือกคุณลักษณะด้วยการคำนวณค่าน้ำหนักการวัดค่าความถูกต้องในการแบ่งกลุ่มข้อมูลมาสร้างเซตของคุณลักษณะและเลือกเซตคุณลักษณะที่มีประสิทธิภาพมากที่สุด มีเทคนิคดังนี้

2.1 เทคนิค Forward Selection เป็นเทคนิคที่ใช้วิธีการคัดเลือกคุณลักษณะที่เป็นตัวแปรอิสระ เข้ามาในสมการทีละตัว ขั้นแรกพิจารณาจากค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ระหว่างตัวแปรตามกับตัวแปรอิสระที่ให้ค่าสูงสุด หากตัวแปรอิสระตัวนั้นมีคุณสมบัติตามเกณฑ์การนำเข้าก็จะถูกนำเข้าสมการ หลังจากนั้นเป็นการพิจารณาสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรตามกับตัวแปรอิสระที่ไม่อยู่ในสมการถดถอยทีละตัว ตัวแปรอิสระที่มีค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงสุดจะถูกนำเข้าสมการ หากตัวแปรนั้นมีคุณสมบัติตามเกณฑ์การนำเข้าขั้นตอนจะทำซ้ำจนกระทั่ง พบว่าไม่สามารถนำตัวแปรอิสระเข้าสมการได้จึงหยุด

2.2 เทคนิค Backward Elimination เป็นวิธีที่พยายามคัดเลือกตัวแปรที่ดีที่สุดและได้ตัวแบบประหยัด ในการพยากรณ์เช่นเดียวกัน แต่เป็นวิธีที่ตรงข้ามกับวิธี Forward นั้น คือตอนแรกจะนำตัวแปรพยากรณ์ทุกตัวเข้ามาในสมการและดำเนิน การพิจารณาตัวแปรพยากรณ์ที่มีค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation) กับตัวแปรเกณฑ์ โดยควบคุมอิทธิพลของตัวแปรพยากรณ์อื่น ๆ ซึ่งมีค่าต่ำ ที่สุดออกจากสมการ แล้วจึงดำเนินการทดสอบว่า ค่า R^2 ลดลงอย่างมีนัยสำคัญทางสถิติหรือไม่ ถ้าพบว่าลดลงอย่างไม่มีนัยสำคัญทางสถิติแสดงว่าตัวแปร ดังกล่าวไม่ได้มีส่วนทำให้การพยากรณ์ตัวแปรเกณฑ์เพิ่มขึ้นเลย แสดงว่าสามารถขจัดออกจากสมการได้ จากนั้นจึงดำเนินการขจัดตัวแปรพยากรณ์ที่มีความสำคัญน้อยรองลงมาออกไปอีก โดยใช้วิธีพิจารณา เช่นเดียวกัน ซึ่งการขจัดตัวแปรพยากรณ์จะสิ้นสุด เมื่อพบว่ามีผลทำ ให้ค่า R^2 ลดลงอย่างมีนัยสำคัญทางสถิติ หมายความว่า ตัวแปรดังกล่าวมีความสำคัญต่อการพยากรณ์ตัวแปรตาม หากขจัดตัวแปรดังกล่าวออกจากสมการจะทำให้อำนาจการพยากรณ์ตัวแปรเกณฑ์ลดลง จึงต้องคงตัวแปรพยากรณ์ดังกล่าวไว้ในสมการพยากรณ์ต่อไป

3. วิธี Embedded Methods ถูกออกแบบเพื่อแก้ไขข้อเสียของวิธี Filter และ Wrapper Approach ซึ่งจะใช้เวลาการประมวลผลน้อยกว่าวิธี Wrapper โดยวิธีนี้เกี่ยวข้องกับการขึ้นอยู่ต่อกันของฟีเจอร์ด้วยและคำนึงถึงความสัมพันธ์ระหว่างคุณลักษณะ Input/Output โดยเป็นวิธีสำหรับการเรียนรู้เอง มีการเลือกคุณลักษณะที่เหมาะสมสำหรับการสร้างตัวแบบในการแก้ปัญหาต่าง ๆ โดยไม่ต้องเพิ่มกระบวนการวิธีในการคัดเลือกคุณลักษณะที่เหมาะสมอื่น ๆ เข้ามาช่วย

2.3 การสร้างตัวแบบหัวข้อโดยวิธีการจัดสรรแบบดีรีเคลแฝง (Latent Dirichlet Allocation: LDA)

ตัวแบบความน่าจะเป็นที่ถูกสร้างขึ้นโดยมีแนวคิดว่า ในเอกสารจะประกอบไปด้วยหัวข้อรวมกันอยู่แบบสุ่ม ซึ่งหัวข้อเหล่านี้ได้กระจัดกระจายอยู่ในกลุ่มคำของเอกสารนั้น ๆ ซึ่งการทำการจัดสรรแบบดีรีเคลแฝง (Latent Dirichlet Allocation : LDA) ถูกใช้ในการหาหัวข้อหรือสกัดกลุ่มคำออกมาจากเอกสาร โดยการหาค่าความน่าจะเป็นของความสัมพันธ์ของหัวข้อแต่ละประโยคและค่าความน่าจะเป็นของคำที่ปรากฏในแต่ละหัวข้อ ซึ่งใช้หลักความน่าจะเป็นในการหาพารามิเตอร์ โดยหลักการทำงานมีกระบวนการดังนี้

1. เลือกค่า $\boldsymbol{N \sim Poisson(\xi)}$

2. เลือกค่า $\boldsymbol{\theta \sim Dir(\alpha)}$

3. ในแต่ละ W_n จากคำศัพท์ N

   3.1 เลือกหัวข้อ $\boldsymbol{Z_n \sim Multinomial(\theta)}$

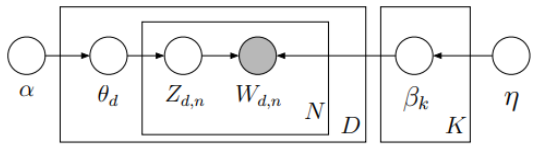
   3.2 เลือกคำศัพท์ W_n จาก $p(W_n \mid Z_{n,\beta})$ ความน่าจะเป็นของคำศัพท์บนหัวข้อ Z_n

โดย W_n คือ คำที่ปรากฏในเอกสาร มีทั้งหมด N

Z_n คือ หัวข้อของคำในแต่ละเอกสาร มีทั้งหมด k หัวข้อ

$\boldsymbol{\theta}$ คือ ความน่าจะเป็นของหัวข้อที่แจกแจงตามเอกสารต่าง ๆ โดยความน่าจะเป็นของแต่ละหัวข้อต้องรวมกันเท่ากับ 1

การแจกแจงความน่าจะเป็นของ LDA สามารถอธิบายได้ในอีกรูปแบบคือ Probabilistic Graphical Model หรือเรียกอีกหนึ่งอย่าง คือ Graphical Model ดังรูปที่ 1



รูปที่ 1 การแสดงตัวแบบกราฟิกของ LDA
ที่มา : [11]

จากรูปที่ 1 แสดงตัวแบบกราฟิกของ LDA สามารถอธิบายตัวแปรในรูปได้ ดังนี้

$W_{d,n}$ เป็นคำที่ปรากฏอยู่ในข้อมูลที่มีทั้งหมด N โดยจะเรียกตัวแปร $W_{d,n}$ ว่า Observed Word เป็นตัวแปรที่สามารถสังเกตได้โดยตรง จึงเรียกว่า ตัวแปรค่าสังเกต

$Z_{d,n}$ เป็นหัวข้อของคำในแต่ละข้อมูลที่มีทั้งหมด k หัวข้อ โดยจะเรียกตัวแปร $Z_{d,n}$ ว่า Latent Variable เป็นตัวแปรที่ไม่สามารถสังเกตได้โดยตรง จึงเรียกว่า ตัวแปรแฝง

θ_d เป็นความน่าจะเป็นของแต่ละหัวข้อกับข้อมูล โดยความน่าเป็นของแต่ละหัวข้อเมื่อรวมกันต้องมีค่าเท่ากับ 1

α เป็นตัวควบคุมสัดส่วนของหัวข้อในแต่ละข้อมูล หาก α มีค่าสูง หมายความว่า การกระจายของหัวข้อในแต่ละข้อมูลค่อนข้างสม่ำเสมอ แต่หาก α มีค่าต่ำ หมายความว่า การกระจายของหัวข้อในแต่ละข้อมูลจะเบี่ยงไปทางหัวข้อใดหัวข้อหนึ่ง

β_k เป็นความน่าจะเป็นของแต่ละคำที่แจกแจงอยู่ในหัวข้อ โดยความน่าจะเป็นของคำแต่ละคำในหัวข้อนั้นเมื่อรวมกันต้องมีค่าเท่ากับ 1

η เป็นตัวควบคุมการกระจายของคำในแต่ละหัวข้อ หากมีค่าสูง การกระจายของคำในหัวข้อค่อนข้างสม่ำเสมอ แต่หากมีค่าต่ำ การกระจายของคำในแต่ละหัวข้อจะเบี่ยงไปทางหัวข้อใดหัวข้อหนึ่ง

หลังจากได้ผลลัพธ์ในรูปของกลุ่มคำของหัวข้อนั้นจะได้ผลลัพธ์ตามจำนวนพารามิเตอร์ (*k*) ที่เราตั้งไว้จากนั้นจะทำการหาจำนวนหัวข้อ มีวิธีดังนี้

1. ค่า Perplexity นั้นเป็นการประเมินประสิทธิภาพของตัวแบบในการอธิบายข้อมูล โดยค่า Perplexity ต่ำมากเท่าไรนั้นหมายความว่า ตัวแบบนั้นมีประสิทธิภาพดีมาก วิธีการหาค่า Perplexity นั้นต้องกำหนดค่า *k* โดยที่ค่า *k* นั้นถูกกำหนดมาให้แล้วจากการทำ LDA และนำคำเหล่านั้นที่ได้กำหนดความเชื่อมโยงกับแต่ละหัวข้อเปรียบเทียบกับการกระจายของคำในเอกสารเพื่อหาว่าเอกสารนั้นสมควรอยู่ในกลุ่มคำของหัวข้อ (Topic) ใด หลังจากที่ได้ค่า *k* ที่ทำให้ค่า Perplexity ต่ำที่สุดแล้วสามารถใช้ค่า *k* เพื่อมาหาจำนวนหัวข้อ (Topic) สามารถคำนวณได้จากสมการที่ 8

$$perplexity = \exp\left(-1 * \log(w)\right) \tag{8}$$

เมื่อ $\log(w)$ คือ log-likelihood ของคำที่ปรากฏในเอกสาร และ w คือ คำที่ปรากฏในเอกสาร

2. ค่า Log – Likelihood Function เป็นการหาค่าความใกล้เคียง โดยผลลัพธ์ที่นั้นหากมีค่าสูง หมายความว่าค่ามีความใกล้เคียงกันมาก โดย Log – Likelihood Function สามารถเขียนได้ดังสมการที่ 9

$$L(\theta, m \mid D) = p(D \mid \theta, m) \tag{9}$$

โดยที่ θ คือ พารามิเตอร์ทั้งหมดของตัวแบบ

m คือ จำนวนตัวแบบ

D คือ ข้อมูลเอกสาร

3. ค่า Coherence เป็นการวัดค่าความคล้ายคลึงกันของความหมายระหว่างคำที่ให้คะแนนสูงในหัวข้อ เพื่อช่วยแยกแยะความหมายของแต่ละหัวข้อ และหัวข้อที่สร้างขึ้นด้วยการอนุมานทางสถิติ คำนวณค่า Coherence ของหัวข้อเป็นผลรวมของคะแนนความคล้ายคลึงกันของการแจกแจงแบบคู่โดยรวมชุดของคำหัวข้อ $T = \langle w_1, \ldots w_n \rangle$ โดยที่ UMass-coherence ถูกกำหนดดังสมการที่ 10

$$C_{UMass}(T) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{p(w_m, w_l) + 1/D}{p(w_l)} \tag{10}$$

โดยที่ p คือ ค่าความน่าจะเป็นของคำ

$p(w_m, w_l)$ คือ อัตราส่วนของจำนวนเอกสารที่มีทั้งคำ w_m, w_l

D คือ จำนวนเอกสารทั้งหมดในคลัง (เพิ่มจำนวนการปรับเรียบ 1/D เพื่อหลีกเลี่ยงการคำนวณลอการิทึมของศูนย์)

2.4 การวิเคราะห์จำนวนหัวข้อ (K)

Hyperparameters คือ พารามิเตอร์ต่าง ๆ ที่ผู้ใช้สามารถกำหนดเองได้ก่อนที่จะนำไปใช้กับตัวแบบการเรียนรู้ ดังนั้นการทำ Hyperparameters Tuning การเพิ่มประสิทธิภาพของตัวแบบให้สูงสุด สามารถแบ่งออกเป็น 2 ประเภทใหญ่ ๆ ได้แก่

1. Traditional Hyperparameter Tuning คือ วิธีการปรับค่าด้วยการเทียบผลของตัวแบบทุกตัว หรือปรับค่าไปเรื่อย ๆ จนกว่าจะพบค่าที่ส่งผลให้ตัวแบบบรรลุผลตามที่คาดหมาย

2. Automated Hyperparameter Tuning คือ การปรับค่าที่เหมาะสมโดยอัตโนมัติด้วยอัลกอริทึมประเภทต่าง ๆ

เมื่อได้จำนวนหัวข้อที่เหมาะสมจึงนำมาทำการวัดประสิทธิภาพ ยกตัวอย่างมา 2 อัลกอริทึม ได้แก่

1. Elbow Method เป็นหนึ่งในอัลกอริทึมที่เอาไว้วัดประสิทธิภาพในการจัดกลุ่มของ K-Mean คือวิธีการวัดหาค่า Error ผลรวมของระยะห่างระหว่าง วัตถุ กับ จุดศูนย์กลาง เพื่อการเลือกค่า K ที่เหมาะสมหรือการหาจำนวนกลุ่มที่เหมาะสมที่สุด จุดที่เหมาะสมของจำนวนกลุ่ม (Clusters) คือจุดที่กราฟมีลักษณะหักศอก จะเป็นจุดที่ให้ค่าจำนวนกลุ่มที่ดีที่สุดหาจำนวนกลุ่มที่เหมาะสมที่สุดจะถูกนำมาจากค่าผลรวมความคลาดเคลื่อนกำลังสอง Sum of Square Error ซึ่งลดลงอย่างมีนัยสำคัญในการคำนวณค่า Sum of Square Error เมื่อค่าที่ได้จากการคำนวณมีความผิดพลาดน้อยลงความชันของเส้นโค้งจะเริ่มเรียบและจะเกิดเป็นมุมที่มีลักษณะคล้ายข้อศอก โดยใช้สมการในการคำนวนหาค่า SSE ดังสมการที่ 11

$$SSE = \sum_{i=1}^{n} (y_i - y')^2 \tag{11}$$

โดยที่ n คือจำนวนตัวอย่าง , y_i คือ คือค่าของตัวอย่างที่ i และ y' คือ ค่าเฉลี่ยของตัวอย่าง

2. Silhouette Score เป็นอัลกอริทึมหนึ่งที่เอาไว้วัดประสิทธิภาพในการจัดกลุ่มของ K-Mean เป็นเทคนิคที่ใช้วัดว่าตัวอย่างนั้นว่ามีความเหมือนกับกลุ่มที่อยู่มากเพียงใด เมื่อเทียบกับกลุ่มอื่น ๆ ค่าของ Silhouette อยู่ในช่วง -1 ถึง +1 ยิ่งมีค่ามากแสดงว่า ตัวอย่างมีความคล้ายกับกลุ่มที่อยู่นั้นมากและมีความคล้ายกับกลุ่มอื่นน้อย ถ้าค่า Silhouette Score มีค่ามากหรือเข้าใกล้ +1 หมายความว่าเป็นการหาจำนวนกลุ่มที่ดี Silhouette เป็นมาตรวัดที่อาศัยการยึดเหนี่ยวภายในกลุ่มและมีความสามารถในการแยกกันระหว่างกลุ่ม Silhouette โดยมีสมการในการคำนวณหา Silhouette Score ดังสมการที่ 12

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{12}$$

โดยที่ $b(i)$ คือ ค่าเฉลี่ยของระยะห่างระหว่างกลุ่ม และ $a(i)$ คือ ค่าเฉลี่ยระยะห่างภายในกลุ่ม

2.5 การวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis)

การวิเคราะห์ขั้วความรู้สึกเป็นกระบวนการที่มุ่งเน้นไปยังการตรวจสอบวัดความรู้สึก ความคิดเห็น ทัศนคติ และอารมณ์ที่มาอยู่ในรูปแบบของข้อความทั่วไป ที่เป็นข้อเท็จจริง (Facts) หรือ ข้อความแสดงความคิดเห็น (Opinion) ที่เป็นข้อความแสดงออกทางความคิดเห็น เป็นสาขาหนึ่งในการประมวลผลภาษาธรรมชาติ โดยการวิเคราะห์ขั้วความรู้สึกนี้มีจุดประสงค์เพื่อเป็นการตรวจวัดหรือบ่งบอกความรู้สึกของข้อความนั้น ๆ เช่น ความรู้สึกดีหรือชอบ ความรู้สึกที่ไม่ดีหรือไม่ชอบ และความรู้สึกปกติ โดยพื้นฐานผลลัพธ์ของการวิเคราะห์ขั้วความรู้สึกนั้นถูกแบ่งออกเป็น 3 แบบ คือ ขั้วบวก (Positive) ขั้วลบ (Negative) และเป็นกลาง (Neutral) และสามารถแบ่งการทำงานออกเป็น 3 ระดับ [17] ได้แก่

- ระดับเอกสาร (Document Level) เป็นการวิเคราะห์อารมณ์ความรู้สึกและความคิดเห็นโดยรวมทั้งเอกสาร
- ระดับประโยค (Sentence Level) ก่อนการวิเคราะห์ขั้วความรู้สึก เอกสารจะถูกแบ่งออกเป็นประโยคและจำทำการวิเคราะห์ทีละประโยค
- ระดับคุณลักษณะ (Feature Level) เป็นการวิเคราะห์ขั้วความรู้สึกตามลักษณะหรือหัวข้อที่สนใจหรือที่ถูกกล่าวถึงในข้อความ

2.6 การถดถอยลอจิสติกพหุ (Multinomial Logistic Regression: MLR)

การถดถอยลอจิสติกพหุ ใช้เพื่อทำนายการจัดหมวดหมู่ ในตัวแปรอิสระหลายตัว ตัวแปรอิสระสามารถเป็นได้ทั้งแบบไม่ต่อเนื่องหรือแบบต่อเนื่องก็ได้ เช่นเดียวกับการถดถอยลอจิสติกแบบทวิภาค (Binary Logistic Regression) การถดถอยลอจิสติกพหุโดยพิจารณาค่าภาวะน่าจะเป็นสูงสุดสำหรับการจำแนกกลุ่ม ในงานวิจัยนี้ใช้กรณีการถดถอยลอจิสติกพหุมาใช้ โดยเลือกใช้ไลบรารี sklearn โดยมีพารามิเตอร์ดังต่อไปนี้ ‘liblinear’, ‘newton-cg’, ‘sag’, ‘saga’, และ ‘lbfgs’

การเพิ่มประสิทธิภาพของตัวแบบการถดถอยโลจิสติกพหุโดยใช้เทคนิค Regularization ได้แก่ Lasso Regression (L1) เป็นการทำการยกกำลังสองของสัมประสิทธิ์เป็นการลงโทษ (Penalty) สำหรับฟังก์ชันการสูญเสีย และ Ridge Regression (L2) เป็นการเพิ่มค่าสัมบูรณ์ของขนาดของสัมประสิทธิ์ เป็นการลงโทษสำหรับฟังก์ชันการสูญเสีย (Cost Function) ซึ่งสามารถอธิบายได้ดังสมการที่ 13 และ 14

การทำให้ L1 เป็นมาตรฐาน

$$\mathrm{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda\sum_{j=0}^{M}\left|W_j\right| \qquad (13)$$

การทำให้ L2 เป็นมาตรฐาน

$$\mathrm{Cost} = \sum_{i=0}^{N}(y_i - \sum_{j=0}^{M} x_{ij}W_j)^2 + \lambda\sum_{j=0}^{M}W_j^2 \qquad (14)$$

โดยที่ λ คือ พารามิเตอร์ที่ทำให้เป็นมาตรฐาน และ W_j คือ น้ำหนักพารามิเตอร์ j

จากสมการเห็นได้ว่าสูตรการทำให้ L1 เป็นมาตรฐานจะเพิ่มเงื่อนไขการลงโทษ (Penalty) ในฟังก์ชันการสูญเสีย (Cost Function) โดยการเพิ่มค่าสัมบูรณ์ของน้ำหนักพารามิเตอร์ (W_j) ในขณะที่การทำให้ L2 เป็นมาตรฐาน จะเพิ่มค่ากำลังสองของน้ำหนักในฟังก์ชันต้นทุน ถ้า แลมป์ดา (λ) เป็น 0 จะได้ค่า Ordinary Least Square (OLS) กลับมา ในขณะที่แลมป์ดา (λ) มีค่ามาก จะทำให้สัมประสิทธิ์เป็นศูนย์

ข้อมูลที่ไม่สมดุล (Imbalance Data) คือ ข้อมูลของคลาสใดคลาสหนึ่งสูงมากเมื่อเทียบกับคลาสอื่นที่มีอยู่ อัลกอริทึมการเรียนรู้ด้วยเครื่องส่วนใหญ่ถือว่าข้อมูลมีการกระจายอย่างสม่ำเสมอภายในคลาส ในกรณีข้อมูลที่ไม่สมดุลจะทำให้อัลกอริทึมมีอคติในการทำนาย การปรับสมดุลข้อมูลจึงเป็นขั้นตอนสำคัญ โดยค่าเริ่มต้นของ พารามิเตอร์ Class_Weight = ‘None’ คือ ทั้งสองคลาสมีน้ำหนักที่เท่ากัน สำหรับกรณีข้อมูลที่ไม่สมดุล สามารถกำหนด Class_Weight = ‘Balance’ ซึ่งตัวแบบจะกำหนดน้ำหนักของคลาสโดยแปรผกผันกับความถี่ ดังสมการที่ 15

$$w_j = \frac{n\_samples}{(n\_classes \times n\_samples_j)} \qquad (15)$$

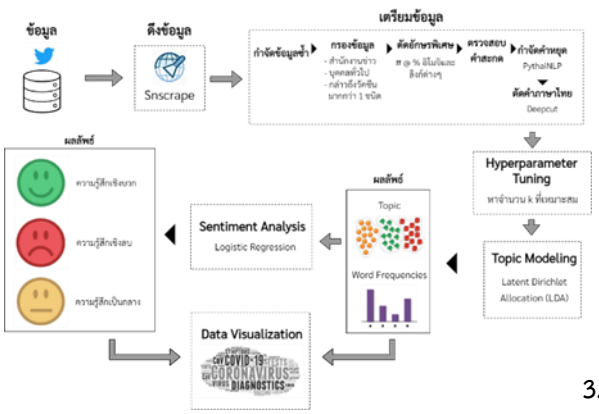
โดยที่ w_j คือ น้ำหนักของคลาสที่ j

$n\_samples$ คือ จำนวนตัวอย่างข้อมูลทั้งหมดในชุดข้อมูล

$n\_classes$ คือ จำนวนคลาสที่ไม่ซ้ำทั้งหมด

$n\_samples_j$ คือ จำนวนตัวอย่างข้อมูลของคลาสที่ j

3 วิธีการดำเนินงานวิจัย



รูปที่ 2 ขั้นตอนการดำเนินงาน

จากรูปที่ 2 แสดงให้เห็นถึงขั้นตอนการดำเนินงาน โดยเริ่มต้นจากการดึงข้อมูลความคิดเห็น ภาษาไทยที่มีการทวีตต่อวัคซีนป้องกันโควิด-19 ด้วยการใช้ชุดคำสั่ง Snscrape ขั้นตอนถัดไปจะเป็น การเตรียมข้อมูลโดยตรวจสอบและกำจัดข้อมูลซ้ำซ้อน จากนั้นกรองเอาเฉพาะข้อมูลทวีตของบุคคลทั่วไปที่กล่าวถึงวัคซีนเพียง 1 ชนิด อีกทั้งนำไปทำการตัดอักษรพิเศษ ได้แก่ # @ % อิโมจิ และลิ้งก์ต่าง ๆ ออกจากข้อมูลทวีตรวมทั้งตรวจสอบคำสะกดผิด จากนั้นทำการกำจัดคำหยุดโดยใช้คลังคำศัพท์ภาษาไทยของ PythaiNLP และทำการตัดคำภาษาไทยโดยใช้ฟังก์ชัน Deepcut ขั้นตอนต่อไปหลังจากทำการเตรียมข้อมูลจะเป็นการหาจำนวนหัวข้อที่เหมาะสม (Hyperparameter Tuning) สำหรับการสร้างตัวแบบหัวข้อ (Topic Modeling) โดยใช้วิธีการ Latent Dirichlet Allocation (LDA) เพื่อวิเคราะห์หา หัวข้อจากข้อมูลทวีตความคิดเห็น จากนั้นทำการจัดกลุ่มข้อความทวีตความคิดเห็นตามหัวข้อโดย พิจารณาจากค่า Dominant ขั้นตอนต่อไปคือการวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis) โดย วิเคราะห์ขั้วความรู้สึกของแต่ละหัวข้อ ซึ่งขั้วความรู้สึกแบ่งออกเป็น 3 รูปแบบ ได้แก่ ขั้วความรู้สึกเชิงบวก ขั้วความรู้สึกเชิงลบ และขั้วความรู้สึกเป็นกลาง หลังจากนั้นจะทำการนำผลลัพธ์จากการสร้างตัวแบบหัวข้อและการวิเคราะห์ขั้วความรู้สึกมานำเสนอในรูปแบบของภาพข้อมูล

3.1 ประชากรและกลุ่มตัวอย่าง

ประชากรที่ศึกษา คือ ข้อความคิดเห็นที่มีการทวีตต่อวัคซีนป้องกันโควิด-19 บนทวิตเตอร์

กลุ่มตัวอย่างที่ศึกษา คือ ข้อความคิดเห็นภาษาไทยที่มีการทวีตต่อวัคซีนป้องกันโควิด-19 บนทวิตเตอร์ จำนวน 493,368 ทวีต ซึ่งข้อความคิดเห็นที่ใช้มีกลุ่มคำสำคัญ ได้แก่ โมเดอร์นา (Moderna) ซิโนแวค (Sinovac) ไฟเซอร์ (Pfizer) ซิโนฟาร์ม (Sinopharm) และแอสตร้าเซนเนกา (AstraZeneca) ช่วง 1 มีนาคม 2564 ถึง 31 ตุลาคม 2564

3.2 การเก็บรวบรวมข้อมูล

สำหรับการเก็บรวบรวมข้อมูลทวีตความคิดเห็นที่มีต่อวัคซีนป้องกันโควิด-19 นั้น ผู้วิจัยได้ทำการดึงข้อมูลโดยการเขียนคำสั่งในโปรแกรม Jupyter Notebook เรียกว่า Web Scrapping โดยอาศัยคำสั่ง Snscrape ในการดึงข้อมูลทวีตที่ต้องการได้ตามที่ระบุ แทนการใช้ Application Programming Interface (API) ของทวิตเตอร์ที่สามารถดึงข้อมูลย้อนหลังได้เพียง 7 วัน โดยผู้วิจัยได้ทำการกำหนดกลุ่มคำที่ใช้สำหรับการดึงข้อมูลทวีตความคิดเห็นที่มีต่อวัคซีนป้องกันโควิด-19 ได้แก่ ซิโนแวค ซิโนฟาร์ม แอสตร้า ไฟเซอร์ โมเดอร์นา

ข้อมูลที่ใช้ในการวิเคราะห์นั้นเป็นข้อมูลที่อยู่ในช่วงระยะเวลาตั้งแต่ 1 มีนาคม 2564 ถึง 31 ตุลาคม 2564 ได้จำนวนทวีตของแต่ละกลุ่มคำ ดังนี้ โมเดอร์นา 84,088 ทวีต ซิโนแวค 83,169 ทวีต ไฟเซอร์ 100,716 ทวีต ซิโนฟาร์ม 109,470 ทวีต แอสตร้า 115,925 ทวีต รวมทั้งสิ้น 493,638 ทวีต

ชุดคำสั่งของ Snscrape มีรูปแบบการทำงานเริ่มจากการสร้าง List เพื่อเตรียมสำหรับเก็บข้อมูลหลังจากทำการดึงและใช้ TwitterSearchScraper ในการดึงข้อมูลโดยการระบุคำสำคัญที่ต้องการดึงข้อมูลและระบุจำนวนทวีตที่ต้องการดึงข้อมูล หลังจากนั้นจะนำข้อความทวีตที่ทำการดึงและถูกเก็บอยู่

ใน List แปลงให้อยู่ในรูปแบบของ Dataframe เพื่อให้ข้อมูลอยู่ในรูปของตาราง ชุดคำสั่งของ Snscrape ที่ใช้สำหรับการดึงข้อมูลจากทวีตเตอร์ด้วยโปรแกรม Jupyter Notebook ซึ่งชุดคำสั่งนี้ไม่สามารถระบุช่วงเวลาที่ต้องการดึงข้อมูลได้

3.3 การเตรียมข้อมูล (Data Cleaning)

การเตรียมข้อมูลสำหรับการทำเหมืองข้อความภาษาไทยแตกต่างจากภาษาอังกฤษ เนื่องด้วยลักษณะเฉพาะของภาษาไทย เช่น ไม่มีการเว้นวรรค ไม่มีตัวพิมพ์ใหญ่ และมีรูปแบบการใช้คำที่หลากหลาย ทำให้ขั้นตอนต่าง ๆ ต้องปรับเปลี่ยนไปตามความซับซ้อนของภาษา

3.3.1 การกำจัดข้อมูลซ้ำ

ข้อมูลทวีตความคิดเห็นของคนไทยที่มีผลต่อวัคซีนป้องกัน โควิด-19 ที่ใช้ในการวิเคราะห์นั้นมีจำนวนมาก ดังนั้นจึงทำการตรวจสอบความซ้ำซ้อนของข้อมูล เพราะหากนำข้อมูลที่มีความซ้ำซ้อนไปวิเคราะห์สามารถทำให้ผลการวิเคราะห์นั้นคลาดเคลื่อนได้

3.3.2 การกรองข้อมูลทวีต

หลังจากที่ได้ทำการดึงข้อมูลทวีตความคิดเห็นที่มีผลต่อวัคซีนป้องกันโควิด-19 เรียบร้อยแล้ว ข้อมูลที่ดึงมานั้นจะประกอบด้วย ข้อมูลทวีตที่เป็นของสำนักงานข่าว ข้อมูลทวีตที่เป็นของบุคคลทั่วไป ซึ่งในการวิจัยครั้งนี้จะทำการวิเคราะห์ข้อมูลจากทวีตวามคิดเห็นของบุคคลทั่วไป ดังนั้นจึงต้องทำการแยกข้อมูลทวีตที่เป็นของสำนักงานข่าวออกจากข้อมูลทวีตความคิดเห็นของบุคคลทั่วไป

โดยที่ผู้วิจัยได้ทำการค้นหาบัญชีทวิตเตอร์ที่เป็นของสำนักงานข่าวได้จำนวน 24 บัญชี ดังนี้ Ch3, ThailandNews, Thairath_TV, Thairath_Pol, Thairath_News, amarintvhd, tnnthailand, VoiceTVOffical, onenews31, thaich8news, workpointTODAY, TNAMCOT, PPTVHD36, ThaiPBS, ThaiPBSNews, NationTv22, Ch7HD, Ch7HDNews, Mono29News, thestandardth, Sanook, Kom_chad_luek, Thansettakij และ KhaosodOnline

3.3.3 การตัดอักษรพิเศษ

ข้อมูลทวีตความคิดเห็นของบุคคลทั่วไปที่มีต่อวัคซีนป้องกันโควิด-19 ที่ดึงมาจากทวิตเตอร์นั้นไม่มีเพียงแค่ข้อมูลที่เป็นข้อความเพียงอย่างเดียว ทวีตแต่ละทวีตนั้นจะมีการใส่ตัวอักษรพิเศษ ยกตัวอย่างเช่น # @ % อิโมจิและลิงก์ต่าง ๆ ซึ่งเป็นส่วนที่ต้องทำการตัดออกจากข้อมูลความคิดเห็น หากไม่ทำการตัดอักษรพิเศษอาจทำให้ข้อมูลหลังจากการวิเคราะห์เกิดความผิดพลาด

ชุดคำสั่งที่ใช้สำหรับการตัดอักษรพิเศษมีรูปแบบการทำงานเริ่มต้นจากการระบุไฟล์ที่ต้องการตัดอักษรพิเศษ จากนั้นทำการระบุภาษาและเงื่อนไขสำหรับการตัดอักษรพิเศษ เช่น ระบุภาษาไทยโดยใช้ ก-๙

3.3.4 การตรวจสอบคำสะกด

ขั้นตอนต่อมาหลังจากทำการตัดอักษรพิเศษแล้วต้องทำการตรวจสอบคำภาษาไทยเนื่องจากข้อมูลความคิดเห็นนั้นจะมีการพิมพ์ภาษาไทยที่ผิดและถูกต้องรวมกันอยู่ในข้อมูล ดังนั้นจึงต้องทำการตรวจสอบคำภาษาไทยที่ผิดแล้วทำการแก้ไขให้ถูกต้อง ตัวอย่างเช่น เป็น ที่เป็นคำที่สะกดถูกต้องแต่ในทวีตความคิดเห็นจะพบคำที่สะกดผิด ได้แก่ เปน เป้น ซึ่งต้องทำการแก้ไขให้เป็นคำที่สะกดถูกต้อง เนื่องจากคำที่สะกดผิดนั้นสามารถทำให้ผลการวิเคราะห์ข้อมูลผิดพลาดได้ รวมถึง การใช้ภาษาพูด เช่น “วัคซีนนนนน” แทนที่“วัคซีน” แก้ไขด้วยการใช้ Regular Expressions ร่วมกับฟังก์ชันใน PyThaiNLP

3.3.5 การกำจัดคำหยุด

ขั้นตอนก่อนการตัดคำภาษาไทยนั้นจะต้องทำการกำจัดคำหยุดของภาษาไทย ลบคำที่ไม่สำคัญ ต้องใช้ Thai_Stopwords เช่น “ที่” “คือ” “แต่” และ “จาก” เป็นต้น จาก PyThaiNLP เพราะหากไม่ทำการกำจัดคำหยุดภาษาไทย อาจทำให้ขั้นตอนในการตัดคำภาษาไทยนั้นเกิดความผิดพลาดในการตัดคำและส่งผลให้ข้อมูลผลการวิเคราะห์ผิดพลาดได้

3.3.6 การตัดคำภาษาไทย

ข้อมูลทวีตความคิดเห็นของคนไทยที่มีผลต่อวัคซีนป้องกันโควิด-19 นั้นหลังจากที่ทำการตัดอักษรพิเศษและกำจัดข้อมูลซ้ำนั้นข้อมูลความคิดเห็นนั้นจะอยู่ในรูปของประโยค ซึ่งก่อนที่จะนำข้อมูลไปใช้ในการสร้างตัวแบบหัวข้อ (Topic

Modeling) ต้องทำการตัดประโยคให้แยกออกเป็นแต่ละคำ จึงสามารถนำไปใช้ในการสร้างตัวแบบหัวข้อได้ โดยใช้วิธีการตัดคำโดยใช้พจนานุกรม (Dictionary-Base Approach) และ การตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) เนื่องจากภาษาไทยไม่มีการเว้นวรรคระหว่างคำ จึงต้องใช้เครื่องมือตัดคำ เช่น PyThaiNLP ซึ่งต่างจากภาษาอังกฤษที่ใช้ whitespace ในการแบ่งคำโดยอัตโนมัติ ทั้งนี้การตัดคำภาษาไทยโดยใช้พจนานุกรมนั้นจะใช้ PythaiNLP มาใช้เป็นพจนานุกรมของการตัดคำซึ่งสามารถทำให้การตัดคำภาษาไทยนั้นทำได้อย่างรวดเร็ว แต่ความถูกต้องของการตัดคำค่อนข้างต่ำ ดังนั้นผู้วิจัยจึงใช้วิธีการตัดคำโดยอาศัยคลังข้อมูลร่วมกับพจนานุกรม อย่างไรก็ตามการเตรียมคลังคำที่ต้องการตัดจึงช่วยแก้ไขความผิดพลาดของการตัดคำภาษาไทย ซึ่งการตัดคำภาษาไทยมีฟังก์ชันสำหรับการตัดคำหลายฟังก์ชัน ได้แก่ ฟังก์ชัน Newmm Longest Deepcut Pyicu Ulmfit Tcc Etcc Multicut โดยแต่ละฟังก์ชันจะให้ผลลัพธ์การตัดคำที่แตกต่างกัน จากการศึกษาและทดลองใช้พบว่าฟังก์ชัน Deepcut สามารถให้ผลลัพธ์การตัดคำภาษาไทยที่ดีกว่าฟังก์ชันอื่นเพราะฟังก์ชัน Deepcut มีการใช้ข้อมูลบทความ ข่าว นวนิยาย ในการฝึกฝน ฟังก์ชัน Deepcut ให้สามารถตัดคำภาษาไทยเพื่อผลลัพธ์ที่ถูกต้อง ดังนั้นจึงเลือกใช้ฟังก์ชัน Deepcut ในการตัดคำภาษาไทย

3.4 การสร้างตัวแบบหัวข้อ (Topic Modeling)

หลังจากที่ทำการตัดคำภาษาไทยโดยนำความคิดเห็นในรูปของข้อความประโยคมาตัดแยกเป็นแต่ละคำเพื่อใช้ในการสร้างตัวแบบหัวข้อ โดยการจัดกลุ่มคำสำหรับจำแนกตามหัวข้อที่มีกล่าวถึงในทวีตความคิดเห็นของแต่ละวัคซีนนั้น วิธีการสร้างตัวแบบหัวข้อนั้นมีหลายวิธี ทางผู้วิจัยเลือกใช้วิธี Latent Dirichlet Allocation (LDA) เนื่องจากเป็นวิธีที่นิยมกันแพร่หลาย ในหลาย ๆ งานวิจัยเลือกใช้วิธี LDA เพราะแตกต่างจากวิธีอื่น ๆ ที่มีความสามารถในการจัดกลุ่มเอกสารคือ มีการแบ่งการทำงานออกเป็น 3 วิธี และมีการกระบวนการทำซ้ำ (Repeat) ในการเลือกหัวข้อ เป็นผลให้เอกสารหนึ่ง ๆ อาจจะมีความเชื่อมโยงกับหลาย ๆ หัวข้อได้ [7] ทั้งนี้ วิธี LDA มักจะให้ผลลัพธ์ที่ดีกว่า LSA โดย LSA ใช้ Singular Value Decomposition เพื่อลดมิติของข้อมูล และแปลงเอกสารให้เป็นเวกเตอร์ใน Latent Space เพื่อค้นหาความสัมพันธ์เชิงความหมาย โดยไม่แนวคิดของความน่าจะเป็น ในขณะที่ LDA ใช้ Bayesian Inference สามารถตีความได้ชัดเจนกว่าเพราะใช้ Probabilistic Distribution สามารถบอกได้ว่าเอกสารนี้มีโอกาสเป็นหัวข้อใดบ้าง แต่ประมวลผลช้ากว่า LSA

การใช้โมเดล LDA จำเป็นต้องเตรียมข้อมูลภาษาไทยอย่างรอบคอบ อาทิเช่น

- ต้องตัดคำอย่างถูกต้องก่อนป้อนเข้าสู่โมเดล
- ต้องเลือก stopwords ภาษาไทยให้เหมาะสม เพื่อป้องกันการแทรกแซงของคำที่ไม่มีความหมาย
- การตั้งค่าพารามิเตอร์ เช่น จำนวนหัวข้อ (topics) ซึ่งต้องทดสอบซ้ำหลายรอบ (cross-validation) เพราะเนื้อหาภาษาไทยมักมีลักษณะไม่เป็นทางการ

3.5 การวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis)

ขั้นตอนต่อมาคือการวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis) เป็นวิธีการวิเคราะห์ข้อความคิดเห็นทวีตว่าทวีตนั้น ๆ แสดงถึงความรู้สึกอย่างไร ซึ่งผลลัพธ์ของการวิเคราะห์ขั้วความรู้สึกนั้นประกอบด้วย 3 ขั้วความรู้สึก คือ ขั้วความรู้สึกเชิงบวก (Positive) ขั้วความรู้สึกเชิงลบ (Negative) และขั้วความรู้สึกเป็นกลาง (Neutral) ซึ่งวิธีการวิเคราะห์ขั้วความรู้สึกแบ่งออกเป็น 2 วิธีหลัก ๆ คือ การเรียนรู้ด้วยเครื่อง (Machine Learning) และการใช้พจนานุกรม (Lexicon-based) ทางผู้วิจัยเลือกใช้วิธีการเรียนรู้ด้วยเครื่อง โดยใช้ตัวแบบ Logistic Regression ใน Library sklearn

ข้อมูลสำหรับนำไปใช้ในการสร้างตัวแบบวิเคราะห์ขั้วความรู้สึก ทางคณะผู้วิจัยเลือกใช้ข้อมูลทวีตจำนวน 1,100 ทวีตจากทวีตความคิดเห็นของวัคซีนทุกชนิดโดยสุ่มมาจำนวนอย่างละเท่า ๆ กัน และกำหนดผลเฉลยขั้วความรู้สึก เพื่อสร้างตัวแบบวิเคราะห์ขั้วความรู้สึก ซึ่งกำหนดให้อัตราส่วนจำนวนข้อมูลชุดฝึกสอน (Training Dataset) ต่อจำนวนข้อมูลชุดทดสอบ (Test Dataset) เท่ากับ 70:30 กล่าวคือ จำนวนข้อมูลที่ใช้ในการสร้างตัวแบบเท่ากับ 1,100 ตัว และจำนวน

ข้อมูลที่ใช้ทดสอบเท่ากับ 330 แสดงตัวอย่างข้อความทวีตและผลเฉลยขั้วความรู้สึก

การประเมินตัวแบบการจำแนกขั้วความรู้สึกจะพิจารณาจากค่าความถูกต้อง (Accuracy) ค่า ความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่า F1-Score และ ค่าพื้นที่ใต้เส้นโค้ง ROC (AUC-Score) ที่มีค่ามากที่สุดและ ค่า False Positive Rate (FPR) ของขั้วความรู้สึกทั้ง 3 กลุ่มที่มีค่าต่ำที่สุด

สำหรับภาษาไทย ยังไม่มีโมเดลที่มีความแม่นสูงได้เช่นเดียวกับ VADER ที่ใช้กับภาษาอังกฤษ แนวทางการดำเนินการคือ Finetune โมเดลด้วยชุดข้อมูลความคิดเห็นที่ติดป้ายกำกับ

3.6 การจัดทำ Dashboard สำหรับนำเสนอขั้วความรู้สึกในแต่ละหัวข้อ (Aspect-Level Sentiment Classification)

ขั้นตอนต่อมาคือ การจัดทำ Dashboard สำหรับนำเสนอขั้วความรู้สึกในแต่ละหัวข้อ (Aspect Level Sentiment Classification) เป็นวิธีการนำเสนอผลลัพธ์ขั้วความรู้สึกในรูปแบบกราฟโดยใช้โปรแกรม Power BI และสร้าง Wordcloud โดยใช้ชุดคำสั่ง Python วัคซีนทั้ง 5 ชนิดได้แก่ แอสตร้าเซนเนกา (AstraZeneca) ซิโนแวค (Sinovac) ซิโนฟาร์ม (Sinopharm) ไฟเซอร์ (Pfizer) และโมเดอร์นา (Moderna) เพื่อให้สามารถนำมาใช้ในการนำเสนอและวิเคราะห์ผลลัพธ์ได้สะดวกยิ่งขึ้น

หลังจากสร้าง Wordcloud ขั้วความรู้สึกทั้ง 3 กลุ่มของแต่ละวัคซีนขั้นตอนต่อไปของการจัดทำ Dashboard นำข้อมูลจำนวนทวีตของขั้วความรู้สึกของแต่ละวัคซีนนำไปพล็อตในโปรแกรม Power BI สำหรับการจัดทำ Dashboard

4. ผลการวิจัยและการอภิปรายผล

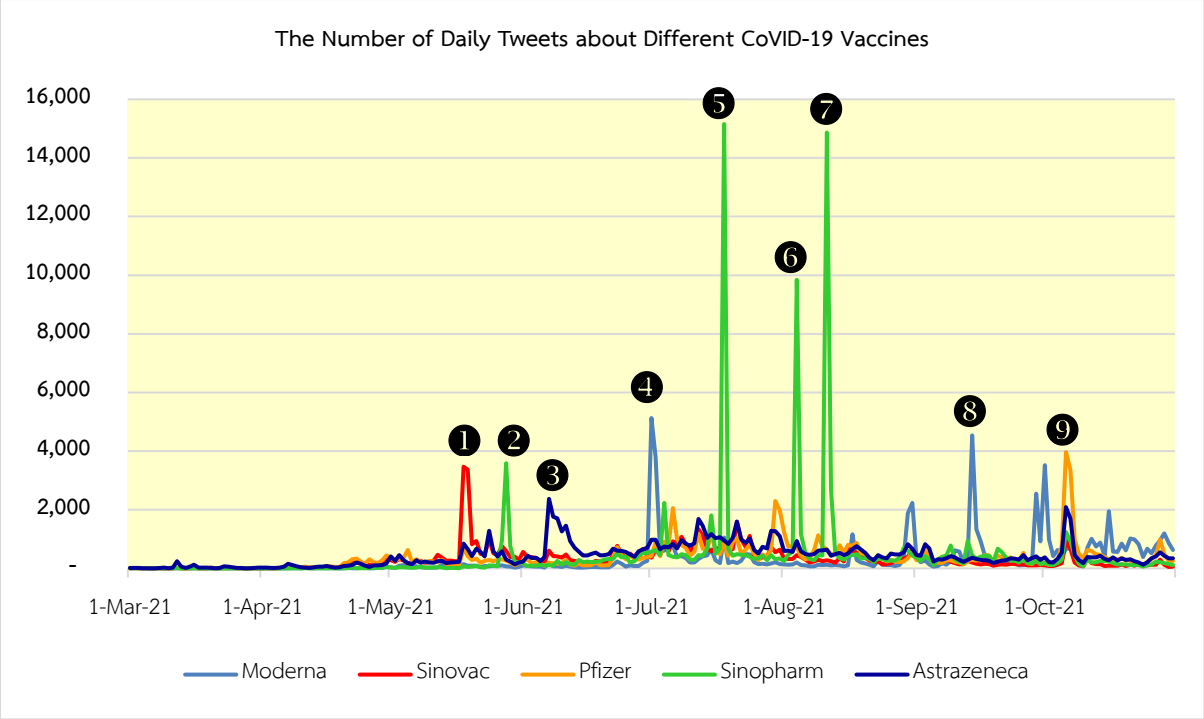
4.1 ผลการประมวลภาษาธรรมชาติของทวีตเกี่ยวกับวัคซีนป้องกันโควิด -19

จากการเก็บรวบรวมทวีตเกี่ยวกับวัคซีนป้องกันโควิด - 19 ชนิดต่าง ๆ ได้แก่ แอสตร้าเซเนกา (AstraZeneca) ซิโนแวค (Sinovac) ซิโนฟาร์ม (Sinopharm) ไฟเซอร์ (Pfizer) และโมเดอร์นา (Moderna) ตั้งแต่เดือน มีนาคม ถึง ตุลาคม 2564 พบว่ามีจำนวนทวีตทั้งสิ้น 451,209 ทวีต โดยจำนวนทวีตรวมจำแนกตามชื่อวัคซีนป้องกันโควิด - 19 และจัดเรียงลำดับตามจำนวนทวีตจากค่าน้อยไปค่ามาก สามารถแสดงได้ดังตารางที่ 2

ตารางที่ 2 จำนวนทวีตที่เกี่ยวกับวัคซีนป้องกันโควิด - 19 ชนิดต่าง ๆ

วัคซีนป้องกันโควิด-19	จำนวนข้อความทวีต	คิดเป็นร้อยละ (%)
โมเดอร์นา	76,262	16.90
ซิโนแวค	76,362	16.92
ไฟเซอร์	88,792	19.68
ซิโนฟาร์ม	103,125	22.86
แอสตร้าเซเนกา	106,668	23.64
รวม	451,209	100.00

จากตารางที่ 2 พบว่า วัคซีนแอสตร้าเซเนกามีจำนวนทวีตสูงกว่าวัคซีนป้องกันโควิด-19 ชนิดอื่น ๆ เนื่องจากเป็นวัคซีนที่มีการนำเข้าและเริ่มฉีดเป็นชนิดแรก ๆ ในประเทศไทย พร้อมกับวัคซีนซิโนแวค เช่นเดียวกับวัคซีนซิโนฟาร์มที่มีจำนวนทวีตสูงที่สุดในวัคซีนทางเลือก เนื่องจากการนำเข้ามาฉีดก่อนวัคซีนทางเลื่อนชนิดอื่น ในส่วนของวัคซีนทางเลือกโมเดอร์นามีจำนวนน้อยสุด อาจเป็นเพราะการอนุมัตินำเข้ามาฉีดล่าช้าและการนำเข้ามาฉีดอยู่หลังช่วงเวลาที่ผู้วิจัยเก็บรวบรวมข้อมูล โดยจำนวนทวีตรายวันของวัคซีนป้องกันโควิด–19 ชนิดต่าง ๆ ตั้งแต่เดือน มีนาคม ถึง ตุลาคม 2564 สามารถแสดงได้ดังรูปที่ 2 ซึ่งแกนแนวนอนแสดงเวลารายวัน ส่วนแกนแนวตั้งแสดงจำนวนทวีต (หน่วยเป็นข้อความ)



รูปที่ 3 จำนวนทวีตรายวันที่เกี่ยวกับวัคซีนป้องกันโควิด-19 ตั้งแต่เดือน มีนาคม ถึง ตุลาคม 2564

รูปที่ 3 แสดงจำนวนทวีตรายวันของวัคซีนป้องกันโควิด – 19 ชนิดต่าง ๆ จะเห็นว่าโดยเฉพาะในช่วงแรกข้อความทวีตเกี่ยวกับวัคซีนชนิดต่าง ๆ ยังมีจำนวนไม่มากนัก และมีจำนวนทวีตของแต่ละวัคซันไม่แตกต่างกัน ในขณะที่ช่วงหลังของระยะเวลาที่ศึกษา มีข้อความทวีตเกี่ยวกับวัคซีนเพิ่มมากขึ้น โดยพบว่า มี 9 ตำแหน่ง ที่มีค่าสูงผิดปกติ ได้แก่

ตำแหน่งที่ ❶ แสดงจำนวนทวีตที่กล่าวถึงวัคซีนซิโนแวคสูงถึง 3,469 ระหว่างวันที่ 18 – 19 พฤษภาคม 2564 ตามลำดับ เนื่องจาก ชมพู่ อารยา เป็นดาราไทยชื่อดังออกมารีวิวอาการ ผลข้างเคียง และการตัดสินใจในการเข้ารับวัคซีนซิโนแวคเข็มแรกในอินสตราแกรมส่วนตัว

ตำแหน่งที่ ❷ แสดงจำนวนทวีตที่กล่าวถึงวัคซีนซิโนฟาร์มสูงถึง 3,585 ทวีต ในวันที่ 28 พฤษภาคม 2564 เป็นผลจากองค์การอนามัยโลก (WHO) ได้ลงรายการวัคซีนซิโนฟาร์มเพื่อใช้ในกรณีฉุกเฉิน [23]

ตำแหน่งที่ ❸ แสดงจำนวนทวีตที่กล่าวถึงวัคซีนแอสตร้าเซเนกาที่เพิ่มสูงขึ้นอย่างชัดเจน เนื่องจากกระทรวงสาธารณสุขประกาศฉีดวัคซีนป้องกันโควิด – 19 ให้กับประชาชนพร้อมกันทั้งประเทศ [4] และมีจำนวนทวีตสูงถึง 2,372 ทวีต

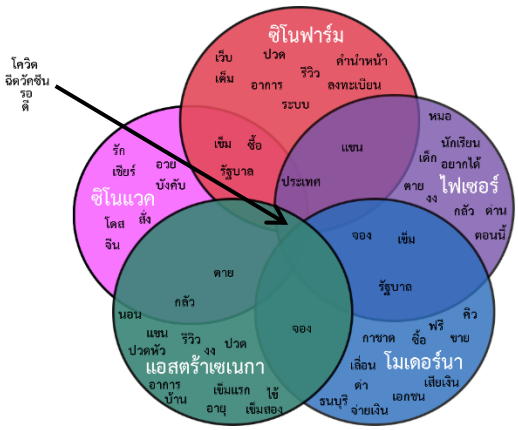
ตำแหน่งที่ ❹ แสดงจำนวนทวีตวัคซีนโมเดอร์นาสูงถึง 5,129 ทวีต ในวันที่ 1 กรกฎาคม 2564 เนื่องจากโรงพยาบาลเอกชนหลายแห่งเปิดระบบจองวัคซีนทางเลือกชนิดโมเดอร์นา

ตำแหน่งที่ ❺, ❻ และ ❼ แสดงจำนวนทวีตวัคซีนซิโนฟาร์มเพิ่มสูงอย่างมากถ้าเทียบกับวัคซีนชนิดอื่น ๆ โดยมีจำนวนทวีต เท่ากับ 15,158, 9,838 และ 14,861 ทวีต ในวันที่ 18 กรกฎาคม 4 สิงหาคม และ 11 สิงหาคม 2564 ตามลำดับ เนื่องจาก วันที่ 18 กรกฎาคม 2564 เปิดให้ลงทะเบียนจองวัคซีนซิโนฟาร์มเป็นวันแรกของราชวิทยาลัยจุฬาภรณ์ ส่วนวันที่ 4 สิงหาคม 2564 เปิดจองลงทะเบียนวัคซีนซิโนฟาร์มรอบที่ 2 ของราชวิทยาลัยจุฬาภรณ์ และวันที่ 11 สิงหาคม 2564 เปิดจองลงทะเบียนวัคซีนซิโนฟาร์ม รอบที่ 3 ของราชวิทยาลัยจุฬาภรณ์ [9]

ตำแหน่งที่ ❽ แสดงจำนวนทวีตวัคซีนโมเดอร์นาเพิ่มสูงขึ้นถึง 4,542 ทวีตในวันที่ 14 กันยายน 2564 เนื่องจากมีรายงานข่าวจาก ประชุมคณะรัฐมนตรีว่า ครม.ได้อนุมัติงบกลาง ปี 2564 เป็นจำนวนเงิน 946.13 ล้านบาท

ให้กับสภากาชาดไทย สำหรับใช้ในโครงการฉีดวัคซีนโมเดอร์นาฟรีให้กับประชาชน

ตำแหน่งที่ ❾ แสดงจำนวนทวีตเกี่ยวกับวัคซีนไฟเซอร์ ในวันที่ 6 -7 ตุลาคม 2564 ที่เพิ่มสูงถึง 3,697 และ 3,344 ทวีต ตามลำดับ เนื่องจากกรมควบคุมโรคได้ชี้แจงวัคซีนไฟเซอร์ 1.5 ล้านโดสถึงประเทศไทยตามกำหนด คาดว่าเพียงพอสำหรับกลุ่มนักเรียนที่อายุ 12 ปีขึ้นไป [2]



รูปที่ 4 แผนภาพเวนน์ (Venn Diagram) แสดงคำที่ปรากฏบ่อย ๆ ในวัคซีนชนิดต่าง ๆ

จากรูปที่ 4 แสดงคำที่ปรากฏในวัคซีนป้องกันโควิด – 19 ชนิดต่าง ๆ ได้แก่ “โควิด” “ฉีดวัคซีน” “รอ” “ดี” โดยในวงกลมของวัคซีนซิโนแวค และ วัคซีนแอสตร้าเซเนกา พบคำว่า “ตาย” “กลัว” ปรากฏบ่อย ๆ คล้ายกัน วงกลมของวัคซีนโมเดอร์นา วัคซีนไฟเซอร์ และวัคซีนซิโนฟาร์ม มีคำที่ปรากฏบ่อย ๆ คล้ายกัน พบคำว่า “จอง” “รัฐบาล” “เข็ม” เนื่องจากวัคซีนทั้ง 3 ชนิดนี้ เป็นวัคซีนทางเลือกจึงกล่าวถึงเรื่องการจองวัคซีน

4.2 ผลการวิเคราะห์หัวข้อ

4.2.1 การวิเคราะห์จำนวนหัวข้อสำหรับวัคซีน

การวิเคราะห์หัวข้อของทวีตในวัคซีนแต่ละชนิดจำเป็นต้องระบุพารามิเตอร์ Hyperparameter ต่าง ๆ ได้แก่ แอลฟ่า (Alpha) เบต้า (Beta) และจำนวนหัวข้อ โดยพิจารณาค่า Coherence ที่สูงสุดในการหาจำนวนหัวข้อที่เหมาะสม โดยค่า Alpha = 0.01 Beta = 0.1 และจำนวนหัวข้อที่มีค่าระหว่าง 2 - 19 จำแนกตามวัคซีนชนิดต่าง ๆ ดังนี้

- วัคซีนโมเดอร์นา (Moderna) จะกำหนดจำนวนหัวข้อเท่ากับ 4
- วัคซีนซิโนแวค (Sinovac) จะกำหนดจำนวนหัวข้อเท่ากับ 9
- วัคซีนไฟเซอร์ (Pfizer) จะกำหนดจำนวนหัวข้อเท่ากับ 10
- วัคซีนซิโนฟาร์ม (Sinopharm) จะกำหนดจำนวนหัวข้อเท่ากับ 10
- วัคซีนแอสตร้าเซนเนกา (AstraZeneca) จะกำหนดจำนวนหัวข้อเท่ากับ 8

4.2.2 ผลลัพธ์ของการวิเคราะห์หัวข้อ (Topic Modeling) โดยใช้ตัวแบบการจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation)

ตารางที่ 3 น้ำหนักของคำในแต่ละหัวข้อของวัคซีนโมเดอร์นา (Moderna)

หัวข้อ	คำที่ปรากฏ
การขอเลื่อนฉีดวัคซีนโมเดอร์นาสำหรับผู้ที่จอง	0.040*"ฉีด" , 0.027*"จอง" , 0.023*"รอ" , 0.011*"จ่ายเงิน" , 0.010*"วัคซีน" , 0.009*"ฟรี" , 0.008*"เสียเงิน" , 0.007*"เลื่อน" , 0.005*"บ้าน" , 0.005*"คิว"
การจัดสรรวัคซีนโมเดอร์นาที่จองไว้ให้แก่โรงพยาบาลเอกชนและสภากาชาดไทยผ่านองค์การเภสัช	0.012*"เอกชน" , 0.011*"วัคซีน" , 0.008*"ล้าน" , 0.008*"รพ" , 0.008*"จอง" , 0.008*"โดส" , 0.006*"กาชาด" , 0.006*"ฉีด" , 0.005*"อภ" , 0.005*"รอ"
นโยบายบริหารจัดการของรัฐบาลและปัญหาการเมือง	0.017*"รัฐบาล" , 0.013*"โควิด" , 0.010*"วัคซีน" , 0.007*"ฆาตกร" , 0.007*"ประชาชน" , 0.007*"ฉีด" , 0.006*"ภูเก็ตแซนด์บ็อกซ์" , 0.006*"ฟรี" , 0.005*"ซื้อ" , 0.005*"มือบ"
การจองวัคซีนโมเดอร์นากับทางโรงพยาบาลเอกชนต่าง ๆ	0.016*"วัคซีน" , 0.015*"จอง" , 0.014*"โควิด" , 0.014*"รพ" , 0.010*"ธนบุรี" , 0.009*"คิว" , 0.009*"เกษมราษฎร์" , 0.007*"ทางเลือก" , 0.007*"โรงพยาบาล" , 0.006*"ฉีด"

ตารางที่ 4 น้ำหนักของคำในแต่ละหัวข้อของวัคซีนซิโนแวค (Sinovac)

หัวข้อ	คำที่ปรากฏ
ดาราไทยเข้ารับการฉีดวัคซีนซิโนแวค	0.016*"ชมพู่", 0.016*"อั๋ม", 0.015*"พัชราภา", 0.015*"อารยา", 0.013*"ดารา", 0.013*"วัคซีน", 0.013*"โดส", 0.011*"ล้าน", 0.011*"โควิด", 0.009*"ไทย"
การบังคับให้ฉีดซิโนแวค	0.021*"กลัว", 0.017*"โดน", 0.013*"บังคับ", 0.012*"วัคซีน", 0.008*"ฉีดวัคซีน", 0.007*"พ่อแม่", 0.006*"โควิด", 0.006*"อยากได้", 0.005*"บ้าน", 0.005*"อวย"
ดาราไทยออกมารีวิวอาการและผลข้างเคียงหลังฉีดซิโนแวค	0.011*"โควิด", 0.011*"วัคซีน", 0.008*"ชมพู่", 0.008*"อารยา", 0.008*"ภูมิ", 0.007*"อาการ", 0.007*"ผลข้างเคียง", 0.007*"ฉีดวัคซีน", 0.006*"รัฐบาล", 0.006*"เจอ"
คุณหมอยงและกลุ่มประชาชนบางกลุ่มสนับสนุนการฉีดวัคซีนซิโนแวค	0.020*"โควิด", 0.018*"สลิ่ม", 0.016*"รัก", 0.012*"วัคซีน", 0.010*"หมอยง", 0.009*"บ้าน", 0.009*"ตู่", 0.008*"รอ", 0.007*"ฉีดวัคซีน", 0.006*"รอด"
รัฐบาลเปิดลงทะเบียนฉีดวัคซีนซิโนแวคฟรี	0.025*"วัคซีน", 0.007*"ลงทะเบียน", 0.007*"โควิด", 0.006*"ฉีดวัคซีน", 0.005*"ฟรี", 0.005*"ครู", 0.005*"ไทย", 0.005*"หมอ", 0.004*"บังคับ", 0.004*"รัฐบาล"
ความวิตกกังวลเกี่ยวกับภูมิคุ้มกันและการเสียชีวิตจากการฉีดวัคซีน	0.012*"วัคซีน", 0.008*"ฉีดวัคซีน", 0.007*"โควิด", 0.007*"ประชาชน", 0.006*"ไอ", 0.006*"รัฐบาล", 0.005*"กลัวตาย", 0.005*"ภูมิคุ้มกัน", 0.005*"ซื้อ", 0.004*"รีบ"
การบังคับให้บุคลากรฉีดวัคซีนซิโนแวคที่ผลิตจากสาธารณรัฐประชาชนจีน	0.012*"โควิด", 0.012*"วัคซีน", 0.010*"บังคับ", 0.008*"จีน", 0.007*"เข็มแรก", 0.007*"ฉีดวัคซีน", 0.005*"บุคลากร", 0.005*"ตัวเอง", 0.005*"ไทย", 0.005*"เข็มสอง"
รัฐบาลไทยออกมาชี้แจงประเด็นการด้อยค่าวัคซีนซิโนแวค	0.022*"ฉีดวัคซีน", 0.013*"วัคซีน", 0.011*"ซื้อ", 0.009*"อย่า", 0.008*"ชนิด", 0.008*"โควิด", 0.007*"ไทย", 0.006*"รัฐบาล", 0.005*"ด้อยค่า", 0.004*"รอ"
แพทย์ไทยแสดงความคิดเห็นต่อรัฐบาลให้สั่งซื้อวัคซีนซิโนแวค	0.013*"หมอ", 0.012*"รัฐบาล", 0.011*"หยุด", 0.010*"วัคซีน", 0.007*"เสี่ยง", 0.007*"ฉีดวัคซีน", 0.006*"โควิด", 0.005*"ซื้อ", 0.005*"ไม่ยอม", 0.004*"เป็นไร"

ตารางที่ 5 น้ำหนักของคำในแต่ละหัวข้อของวัคซีนไฟเซอร์ (Pfizer)

หัวข้อ	คำที่ปรากฏ
การรออนุมัติรับรองวัคซีนไฟเซอร์จากสำนักงานคณะกรรมการอาหารและยา	0.028*"รอ", 0.013*"วัคซีน", 0.008*"โควิด", 0.008*"อนุมัติ", 0.008*"เด็ก", 0.007*"หมอบุญ", 0.007*"อย", 0.007*"ล้าน", 0.007*"โดส", 0.006*"ไทย"
นายกรัฐมนตรีฉีดวัคซีนไฟเซอร์เพื่อเดินทางไปต่างประเทศ	0.009*"วัคซีน", 0.008*"ตู่", 0.007*"บิน", 0.007*"ฉีดวัคซีน", 0.006*"สลิ่ม", 0.006*"โควิด", 0.006*"นักเรียน", 0.005*"รอ", 0.005*"ไม่ยอม", 0.005*"สาม"
ข่าววัคซีนไฟเซอร์ที่ได้รับบริจาค 1.5 ล้านโดสหายไป ทำให้แพทย์ด่านหน้าไม่ได้รับวัคซีน	0.020*"วัคซีน", 0.013*"โควิด", 0.010*"ด่าน", 0.008*"หาย", 0.008*"ทวง", 0.008*"ล้าน", 0.007*"หน่วย", 0.007*"โดส", 0.005*"หมอ", 0.005*"อย่า"
เด็กวัยรุ่นมีความต้องการฉีดวัคซีนไฟเซอร์	0.023*"อยากได้", 0.012*"วัยรุ่น", 0.012*"เด็ก", 0.009*"วัคซีน", 0.008*"ทหาร", 0.007*"อ่าน", 0.007*"ไขว้", 0.006*"กลัว", 0.005*"ฉีดวัคซีน", 0.005*"เกาหลี"
ประเทศสหรัฐอเมริกาอนุญาตให้ฉีดวัคซีนในเด็กได้ แต่ข่าวลือเกี่ยวกับวัคซีนไฟเซอร์ทำให้เกิดอาการบางอย่างในเด็ก	0.012*"หมอ", 0.010*"เด็ก", 0.008*"วัคซีน", 0.008*"ข่าวลือ", 0.008*"เมกา", 0.006*"พยาบาล", 0.006*"รอ", 0.006*"ดีใจ", 0.006*"โควิด", 0.005*"ด่าน"

หัวข้อ	คำที่ปรากฏ
ผลข้างเคียงจากการฉีดวัคซีนไฟเซอร์	0.013*"ฉีดวัคซีน", 0.009*"วัคซีน", 0.008*"ไทย", 0.008*"กลัว", 0.006*"รอ", 0.006*"ผลข้างเคียง", 0.006*"ญี่ปุ่น", 0.006*"อาการ", 0.005*"เข็มแรก", 0.005*"ฟรี"
สาธารณรัฐประชาธิปไตยประชาชนลาวได้รับวัคซีนไฟเซอร์กว่าแสนโดส โดยมีข่าวคนไทยแอบไปเก็บเห็ดและถูกจับที่ลาวได้รับการฉีดวัคซีนไฟเซอร์	0.020*"ลาว", 0.010*"จอง", 0.010*"วัคซีน", 0.008*"รอ", 0.006*"จีน", 0.006*"เห็ด", 0.005*"รัฐบาล", 0.005*"โง่", 0.005*"ปชช", 0.005*"ด่าน"
ปัญหาการสั่งซื้อวัคซีนไฟเซอร์เพื่อป้องกันโควิดของรัฐบาล	0.012*"ซื้อ", 0.012*"สั่ง", 0.011*"รัฐบาล", 0.011*"วัคซีน", 0.008*"โควิด", 0.007*"กราบ", 0.007*"เงิน", 0.006*"ไทย", 0.006*"ภูมิ", 0.006*"ล้าน"
การรอวัคซีนไฟเซอร์และวัคซีนจอห์นสันแอนด์จอห์นสันเข้าประเทศไทย	0.025*"โดส", 0.024*"ล้าน", 0.011*"วัคซีน", 0.009*"สัน", 0.009*"อย่า", 0.008*"ไทย", 0.006*"จอห์น", 0.006*"รอ", 0.006*"รร", 0.004*"โควิด"
บุคลากรทางการแพทย์ด่านหน้าไม่ได้รับวัคซีนไฟเซอร์	0.018*"บุคลากร", 0.011*"ด่าน", 0.011*"ทางการแพทย์", 0.010*"วัคซีน", 0.008*"ขอให้", 0.008*"โควิด", 0.007*"อายุ", 0.007*"แขน", 0.006*"ปวด", 0.006*"ฉีดวัคซีน"

ตารางที่ 6 น้ำหนักของคำในแต่ละหัวข้อของวัคซีนซิโนฟาร์ม (Sinopharm)

หัวข้อ	คำที่ปรากฏ
ระบบลงทะเบียนจองวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือกล่ม	0.053*"เต็ม", 0.027*"วัคซีน", 0.026*"ล่ม", 0.017*"จอง", 0.012*"เว็บ", 0.010*"ทางเลือก", 0.010*"เสียเวลา", 0.010*"ระบบ", 0.009*"ลงทะเบียน", 0.009*"ห่วยแตก"
การเลื่อนการจองฉีดวัคซีนซิโนฟาร์ม	0.073*"จอง", 0.015*"จีน", 0.013*"วัคซีน", 0.011*"สายพันธุ์", 0.009*"ไทย", 0.009*"โดน", 0.008*"โควิด", 0.006*"เลื่อน", 0.006*"อยู่เลย", 0.006*"รอ"
ราชวิทยาลัยจุฬาภรณ์นำเข้าวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือก	0.022*"วัคซีน", 0.019*"โควิด", 0.017*"ราชวิทยาลัยจุฬาภรณ์", 0.010*"จอง", 0.010*"ประชาชน", 0.009*"ฉีดวัคซีน", 0.008*"นำเข้า", 0.007*"ลงทะเบียน", 0.007*"องค์กร", 0.007*"ฟรี"
เว็บการจองของวัคซีนซิโนฟาร์ม	0.017*"เว็บ", 0.015*"เสียเงิน", 0.012*"จอง", 0.011*"วัคซีน", 0.010*"จ่าย", 0.010*"เหนื่อย", 0.010*"ตื่น", 0.009*"ระบบ", 0.009*"ขนาด", 0.007*"สำเร็จ"
การลงทะเบียนจองรับวัคซีนซิโนฟาร์มของจังหวัดต่าง ๆ	0.028*"รอ", 0.015*"จังหวัด", 0.012*"ขาย", 0.010*"วัคซีน", 0.010*"จอง", 0.009*"หวย", 0.008*"กลัว", 0.006*"กค", 0.006*"เว็ป", 0.005*"อก"
ภูมิป้องกันโควิดหลังจากการฉีดวัคซีนซิโนฟาร์ม	0.015*"โควิด", 0.013*"กรอก", 0.009*"วัคซีน", 0.008*"ตรวจ", 0.008*"ภูมิ", 0.007*"จอง", 0.007*"เข้าไป", 0.006*"รีบ", 0.005*"เก่ง", 0.005*"ชื่อ"
ขั้นตอนและข้อมูลที่ต้องใช้ในการจองวัคซีนซิโนฟาร์ม	0.015*"คำนำหน้าชื่อ", 0.011*"อำเภอ", 0.010*"จอง", 0.010*"ข้อมูล", 0.009*"ตอน", 0.008*"เรียบร้อย", 0.007*"ที่อยู่", 0.007*"วัคซีน", 0.007*"โทร", 0.007*"ตำบล"
รีวิวอาการของการฉีดวัคซีนซิโนฟาร์มเป็นเข็มแรก	0.025*"อาการ", 0.021*"ปวด", 0.019*"รีวิว", 0.019*"แขน", 0.016*"ระบบ", 0.016*"ง่วง", 0.013*"เข็มแรก", 0.009*"ปกติ", 0.008*"ผลข้างเคียง", 0.008*"รู้สึก"
ความต้องการวัคซีนซิโนฟาร์มอย่างสูงของบริษัทและองค์กรต่าง ๆ เพื่อฉีดให้แก่พนักงานของตนเอง	0.025*"ลงทะเบียน", 0.015*"ซื้อ", 0.014*"วัคซีน", 0.012*"จ่ายเงิน", 0.012*"จง", 0.011*"แย่ง", 0.010*"คิว", 0.009*"บริษัท", 0.008*"ฟรี", 0.008*"ภาษี"

หัวข้อ	คำที่ปรากฏ
กลุ่มม็อบเรียกร้องให้รัฐบาลนำวัคซีนซิโนฟาร์มมาฉีดฟรีแก่ประชาชน	0.020*"ทิพย์", 0.017*"ฟรี", 0.013*"รัฐบาล", 0.013*"วัคซีน", 0.011*"จอง", 0.009*"ไอ", 0.008*"ม็อบ", 0.007*"เข้าสู่ระบบ", 0.007*"เชียร์", 0.007*"เด็ก"

ตารางที่ 7 น้ำหนักของคำในแต่ละหัวข้อของวัคซีนแอสตร้าเซเนกา (AstraZeneca)

หัวข้อ	คำที่ปรากฏ
การผลิตวัคซีนแอสตร้าโดยบริษัทสยามไบโอไซเอนซ์	0.022*"วัคซีน", 0.017*"ผลิต", 0.016*"อายุ", 0.015*"ไทย", 0.013*"โดส", 0.013*"ล้าน", 0.012*"ม อ", 0.010*"โควิด", 0.008*"สยามไบโอ", 0.008*"ล็อต"
อาการและผลข้างเคียงของวัคซีนแอสตร้าในของผู้สูงอายุ	0.011*"หมอ", 0.010*"วัคซีน", 0.010*"อายุ", 0.008*"ผลข้างเคียง", 0.006*"ขอให้", 0.006*"คนแก่", 0.006*"โควิด", 0.006*"ปลอดภัย", 0.005*"มหาลัย", 0.005*"ฉีดวัคซีน"
การฉีดวัคซีนแอสตร้าเป็นเข็มแรกให้แก่ผู้สูงอายุ	0.015*"เข็มแรก", 0.014*"จอง", 0.009*"อายุ", 0.009*"วัคซีน", 0.007*"ฉีดวัคซีน", 0.007*"รอบ", 0.006*"เหมาะกับ", 0.006*"โควิด", 0.006*"กลัว", 0.005*"รอ"
การฉีดวัคซีนสูตรผสม	0.013*"เข็มสอง", 0.013*"วัคซีน", 0.009*"โดน", 0.007*"เข็มแรก", 0.006*"ฉีดวัคซีน", 0.005*"เรียบร้อย", 0.005*"บุคลากร", 0.005*"ผสม", 0.004*"โควิด", 0.004*"รอ"
ความเครียดที่มีต่อการฉีดวัคซีนแอสตร้า	0.015*"ฉีดวัคซีน", 0.009*"แรง", 0.008*"เหมือนกัน", 0.007*"วัคซีน", 0.007*"เครียด", 0.006*"แขน", 0.006*"ไทย", 0.006*"กลัว", 0.006*"ปวด", 0.006*"อายุ"
ความกลัวเรื่องอายุที่เหมาะสมที่จะฉีดวัคซีนแอสตร้าเซเนกา	0.018*"กลัว", 0.010*"ดีกว่า", 0.010*"รอด", 0.009*"วัคซีน", 0.008*"เกาหลี", 0.007*"น่ากลัว", 0.007*"อายุ", 0.007*"ไทย", 0.006*"รอ", 0.006*"วัยรุ่น"
รีวิวอาการหลังฉีดวัคซีนแอสตร้า	0.024*"ปวด", 0.021*"ไข้", 0.019*"อาการ", 0.018*"แขน", 0.015*"ปวดหัว", 0.014*"รีวิว", 0.010*"ปกติ", 0.010*"มีไข้", 0.008*"หนัก", 0.008*"ตอน"
การรอลงทะเบียนฉีดวัคซีนแอสตร้าเซเนกาของผู้สูงอายุ	0.017*"รอ", 0.012*"ลงทะเบียน", 0.010*"เสี่ยง", 0.010*"คิว", 0.009*"ผู้สูงอายุ", 0.009*"อายุ", 0.009*"วัคซีน", 0.008*"ฉีดวัคซีน", 0.008*"พ่อแม่", 0.006*"ลิ่มเลือด"

4.3 ผลการวิเคราะห์ขั้วความรู้สึก

ค่าความถูกต้อง (Accuracy) ค่าพื้นที่ใต้โค้ง ROC (AUC Score) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และ F1 Score ของตัวแบบการจำแนกขั้วความรู้สึกที่ใช้กับข้อมูลชุดทดสอบโดยกำหนดให้ Class Weight = 'Balanced' และ Class Weight = 'None' ดังตารางที่ 8 และ 9 ตามลำดับ

ผลการทดลองพบว่า วิธี L2-sag-None ให้ค่าความถูกต้องสูงสุดที่ 0.70 หรือคิดเป็นร้อยละ 70 เมื่อกำหนดให้ Class Weight = 'None' และ วิธี L2-saga-balance และวิธี L2-sag-balance ให้ค่าความถูกต้องสูงสุดที่ 0.71 หรือคิดเป็นร้อยละ 71 เมื่อกำหนดให้ Class Weight = 'Balanced ' ซึ่งมีค่าสูงกว่าเมื่อเปรียบเทียบกับวิธีอื่น ๆ นอกจากนี้ ทุก ๆ วิธี เมื่อกำหนดให้ Class Weight = 'Balanced' จะพบว่า พื้นที่ใต้โค้ง ROC มีค่าเพิ่มสูงขึ้น โดยถ้าค่า AUC-Score มากกว่า 0.80 แสดงว่าตัวแบบนั้น ๆ มีประสิทธิภาพในการทำงานได้เป็นอย่างดี (Hosmer and Lemeshow, 2013) ทั้งนี้ วิธี L2-saga-balanced, L1-sag-balanced, elasticnet-saga-balanced ให้ค่า AUC Score เท่ากับ 0.82 ซึ่งสูงกว่าวิธีอื่น ๆ

ตารางที่ 8 ประสิทธิภาพของตัวแบบจำแนกตามกลุ่ม โดยกำหนดให้ Class Weight = 'Balanced'

Parameters (Penalty – Solver)	Polarity	Precision	Recall	F1 score
L2 – balanced - newton cg	Negative	0.59	0.71	0.64
	Neutral	0.80	0.67	0.73
	Positive	0.72	0.85	0.78
L2 - balanced - lbfgs	Negative	0.59	0.71	0.64
	Neutral	0.80	0.67	0.73
	Positive	0.72	0.85	0.78
L2 - balanced - sag	Negative	0.60	0.71	0.65
	Neutral	0.81	0.67	0.73
	Positive	0.72	0.90	0.80
L2 - balanced - saga	Negative	**0.61**	**0.72**	**0.66**
	Neutral	**0.81**	**0.68**	**0.74**
	Positive	**0.72**	**0.90**	**0.80**
L1 - balanced - saga	Negative	0.61	0.69	0.65
	Neutral	0.77	0.67	0.72
	Positive	0.72	0.80	0.76
elasticnet - balanced - saga	Negative	0.61	0.69	0.65
	Neutral	0.77	0.67	0.72
	Positive	0.72	0.80	0.76

ตารางที่ 9 ประสิทธิภาพของตัวแบบจำแนกตามกลุ่ม โดยกำหนดให้ Class Weight = 'None'

Parameters (Penalty – Solver)	Polarity	Precision	Recall	F1 score
L2 – newton cg	Negative	0.59	0.67	0.63
	Neutral	0.81	0.65	0.72
	Positive	0.46	0.59	0.62
L2 - lbfgs	Negative	0.59	0.67	0.63
	Neutral	0.81	0.65	0.72
	Positive	0.46	0.95	0.62
L2 - sag	Negative	0.61	0.67	0.63
	Neutral	0.81	0.67	0.74
	Positive	0.51	0.95	0.67
L2 - saga	Negative	0.60	0.67	0.63
	Neutral	0.82	0.67	0.74
	Positive	0.51	0.95	0.67
L1 - saga	Negative	**0.61**	**0.68**	**0.64**
	Neutral	**0.82**	**0.67**	**0.74**
	Positive	**0.56**	**0.96**	**0.71**
elasticnet - saga	Negative	**0.61**	**0.68**	**0.64**
	Neutral	**0.82**	**0.67**	**0.74**
	Positive	**0.56**	**0.96**	**0.71**

จะเห็นได้ว่า วิธี L2-saga เมื่อกำหนด Class Weight = 'Balanced' นั้น จะทำให้ความแม่นยำ (Precision) ความระลึก (Recall) และ F1 Score ของทุกขั้วความรู้สึกมีค่าสูงขึ้น และ มากกว่า ทั้ง 2 วิธีการ

ตารางที่ 10 ผลลัพธ์ไคสแควร์ขั้วความรู้สึกของวัคซีนแต่ละชนิด

	Chi-Square	DF	p - value
Pearson	19,008.890	8	0.000
Likelihood Ratio	20,315.121	8	0.000

จากตารางที่ 10 ได้นำขั้วความรู้สึกของวัคซีนแต่ละชนิดมาวิเคราะห์ความสัมพันธ์ของวัคซีนและขั้วความรู้สึก โดยกำหนด α = 0.05 สามารถตั้งสมมติฐานได้ว่า

H_0 คือ วัคซีนทั้ง 5 ชนิดไม่มีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่ม

H_1 คือ วัคซีนทั้ง 5 ชนิดมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่ม

สามารถสรุปจากตารางที่ 10 ได้ว่าค่า p – value ที่คำนวณได้มีค่าน้อยกว่า 0.05 จึงปฏิเสธ H_0 ดังนั้น วัคซีนทั้ง 5 ชนิดมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่มที่ระดับนัยสำคัญ 0.05

4.3.1 วัคซีนโมเดอร์นา (Moderna)

จากตารางที่ 11 สามารถอธิบายได้ว่า วัคซีนโมเดอร์นา (Moderna) สามารถจำแนกหัวข้อ (Topic) ได้ 4 หัวข้อ ได้แก่

Topic 1: การขอเลื่อนฉีดวัคซีนโมเดอร์นาสำหรับผู้ที่จอง

Topic 2: การจัดสรรวัคซีนโมเดอร์นาที่จองไว้ให้แก่โรงพยาบาลเอกชนและสภากาชาดไทยผ่านองค์การเภสัช

Topic 3: นโยบายบริหารจัดการของรัฐบาลและปัญหาการเมือง

Topic 4: การจองวัคซีนโมเดอร์นากับทางโรงพยาบาลเอกชนต่าง ๆ

ตารางที่ 11 ขั้วความรู้สึกของวัคซีนโมเดอร์นา (Moderna) ของแต่ละหัวข้อ (Topic)

หัวข้อ (Topic)	ขั้วความรู้สึกเชิงบวก (Positive)	ขั้วความรู้สึกเป็นกลาง (Neutral)	ขั้วความรู้สึกเชิงลบ (Negative)	รวม
Topic 1	49	8,786	8,130	16,965
Topic 2	26	8,656	7,519	16,201
Topic 3	75	5,132	9,655	14,862
Topic 4	81	7,398	2,720	10,199
รวม	231	29,972	28,024	58,227

ในแต่ละหัวข้อสามารถจำแนกขั้วความรู้สึกออกเป็น 3 กลุ่มได้แก่ ขั้วความรู้สึกเชิงลบ (Negative) ขั้วความรู้สึกเป็นกลาง (Neutral) และขั้วความรู้สึกเชิงบวก (Positive) ส่วนใหญ่ข้อมูลมีขั้วความรู้สึกเป็นกลางมากที่สุดและเชิงลบรองลงมา ซึ่งจากตารางแสดงจำนวนขั้วความรู้สึกของวัคซีนโมเดอร์นาของแต่ละหัวข้อ พบว่า หัวข้อที่ 1 มีจำนวนทวีตมากที่สุด โดยส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นการขอเลื่อนฉีดวัคซีนโมเดอร์นาสำหรับผู้ที่จองและรองลงมาจากความคิดเห็นส่วนใหญ่มีความรู้สึกไม่พอใจกับการขอเลื่อนฉีดวัคซีนโมเดอร์นา หัวข้อที่ 2 มีจำนวนทวีตมากที่สุดอันดับที่ 2 ซึ่งส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นการจัดสรรวัคซีนโมเดอร์นาที่จองไว้ให้แก่โรงพยาบาลเอกชนและสภากาชาดไทยผ่านองค์การเภสัช รองลงมาจากความคิดเห็นส่วนใหญ่นั้นมีความรู้สึกไม่พอใจเกี่ยวกับประเด็นนี้

ผลลัพธ์การวิเคราะห์ไคสแควร์ทดสอบความสัมพันธ์ระหว่างหัวข้อกับขั้วความรู้สึกของวัคซีนโมเดอร์นา พบว่าค่า p – value ที่คำนวณมีค่าน้อยกว่า 0.05 จึงปฏิเสธดังนั้น หัวข้อทั้ง 4 หัวข้อมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่ม ที่ระดับนัยสำคัญ 0.05

4.3.2 วัคซีนซิโนแวค (Sinovac)

จากตารางที่ 12 สามารถอธิบายได้ว่า วัคซีนซิโนแวค (Sinovac) สามารถจำแนกหัวข้อ (Topic) ได้ 9 หัวข้อได้แก่

Topic 1: ดาราไทยเข้ารับการฉีดวัคซีนซิโนแวค

Topic 2: การบังคับให้ฉีดซิโนแวค

Topic 3: ดาราไทยออกมารีวิวอาการและผลข้างเคียงหลังฉีดซิโนแวค

Topic 4: คุณหมอยงและกลุ่มประชาชนบางกลุ่มสนับสนุนการฉีดวัคซีนซิโนแวค

Topic 5: รัฐบาลเปิดลงทะเบียนฉีดวัคซีนซิโนแวคฟรี

Topic 6: ความวิตกกังวลเกี่ยวกับภูมิคุ้มกันและการเสียชีวิตจากการฉีดวัคซีน

Topic 7: การบังคับให้บุคลากรฉีดวัคซีนซิโนแวคที่ผลิตจากสาธารณรัฐประชาชนจีน

Topic 8: รัฐบาลไทยออกมาชี้แจงประเด็นการด้อยค่าวัคซีนซิโนแวค

Topic 9: แพทย์ไทยแสดงความคิดเห็นต่อรัฐบาลให้หยุดการสั่งซื้อวัคซีนซิโนแวค

ตารางที่ 12 ขั้วความรู้สึกของวัคซีนซิโนแวค (Sinovac) ของแต่ละหัวข้อ (Topic)

หัวข้อ (Topic)	ขั้วความรู้สึกเชิงบวก (Positive)	ขั้วความรู้สึกเป็นกลาง (Neutral)	ขั้วความรู้สึกเชิงลบ (Negative)	รวม
Topic 1	816	3,640	1,991	6,447
Topic 2	674	2,619	2,054	5,347
Topic 3	768	2,591	1,772	5,131
Topic 4	533	2,933	1,369	4,835
Topic 5	592	2,479	1,480	4,551
Topic 6	563	2,536	1,734	4,833
Topic 7	752	2,779	1,826	5,357
Topic 8	678	2,880	2,005	5,563
Topic 9	568	2,580	1,914	5,062
รวม	5,944	25,037	16,145	47,126

ในแต่ละหัวข้อสามารถจำแนกขั้วความรู้สึกออกเป็น 3 กลุ่มได้แก่ ขั้วความรู้สึกเชิงลบ (Negative) ขั้วความรู้สึกเป็นกลาง (Neutral) และขั้วความรู้สึกเชิงบวก (Positive) ส่วนใหญ่ข้อมูลมีขั้วความรู้สึกเป็นกลางมากที่สุดและเชิงลบรองลงมา ซึ่งจากตารางแสดงจำนวนขั้วความรู้สึกของวัคซีนซิโนแวคของแต่ละหัวข้อ พบว่า หัวข้อที่ 1 มีจำนวนทวีตมากโดยส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นดาราไทยเข้ารับการฉีดวัคซีนซิโนแวคและหัวข้อที่ 8 มีจำนวนทวีตมากที่สุดอันดับที่ 2 ซึ่งส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นรัฐบาลไทยออกมาชี้แจงประเด็นการด้อยค่าวัคซีนซิโนแวค รองลงมาจากความคิดเห็นส่วนใหญ่นั้นมีความรู้สึกไม่พอใจกับการชี้แจงของรัฐบาล

ผลลัพธ์การวิเคราะห์ไคสแควร์ทดสอบความสัมพันธ์ระหว่างหัวข้อกับขั้วความรู้สึกของวัคซีนซิโนแวคพบว่าค่า p – value ที่คำนวณมีค่าน้อยกว่า 0.05 จึงปฏิเสธ ดังนั้นหัวข้อทั้ง 9 หัวข้อมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่มที่ระดับนัยสำคัญ 0.05

4.3.3 วัคซีนไฟเซอร์ (Pfizer)

จากตารางที่ 13 สามารถอธิบายได้ว่า วัคซีนไฟเซอร์ (Pfizer) สามารถจำแนกหัวข้อ (Topic) ได้ 10 หัวข้อได้แก่

Topic 1: การรออนุมัติรับรองวัคซีนไฟเซอร์จากสำนักงานคณะกรรมการอาหารและยา

Topic 2: นายกรัฐมนตรีฉีดวัคซีนไฟเซอร์เพื่อเดินทางไปต่างประเทศ

Topic 3: ข่าววัคซีนไฟเซอร์ที่ได้รับบริจาค 1.5 ล้านโดสหายไป ทำให้แพทย์ด่านหน้าไม่ได้รับวัคซีน

Topic 4: เด็กวัยรุ่นมีความต้องการฉีดวัคซีนไฟเซอร์

Topic 5: ประเทศสหรัฐอเมริกาอนุญาตให้ฉีดวัคซีนในเด็กได้ แต่ข่าวลือเกี่ยวกับวัคซีนไฟเซอร์ทำให้เกิดอาการบางอย่างในเด็ก

Topic 6: ผลข้างเคียงจากการฉีดวัคซีนไฟเซอร์

Topic 7: สาธารณรัฐประชาธิปไตยประชาชนลาวได้รับวัคซีนไฟเซอร์กว่าแสนโดส โดยมีข่าวคนไทยแอบไปเก็บเห็ดและถูกจับที่ลาวได้รับการฉีดวัคซีนไฟเซอร์

Topic 8: ปัญหาการสั่งซื้อวัคซีนไฟเซอร์เพื่อป้องกันโควิดของรัฐบาล

Topic 9: การรอวัคซีนไฟเซอร์และวัคซีนจอห์นสันแอนด์จอห์นสันเข้าประเทศไทย

Topic 10: บุคลากรทางการแพทย์ด่านหน้าไม่ได้รับวัคซีนไฟเซอร์

ตารางที่ 13 ขั้วความรู้สึกของวัคซีนไฟเซอร์ (Pfizer) ของแต่ละหัวข้อ (Topic)

หัวข้อ (Topic)	ขั้วความรู้สึกเชิงบวก (Positive)	ขั้วความรู้สึกเป็นกลาง (Neutral)	ขั้วความรู้สึกเชิงลบ (Negative)	รวม
Topic 1	324	4,261	1,936	6,521
Topic 2	222	2,924	1,405	4,551
Topic 3	416	3,424	1,960	5,800
Topic 4	246	3,663	1,636	5,545
Topic 5	335	3,392	1,851	5,578
Topic 6	397	3,328	1,960	5,685
Topic 7	206	3,322	1,564	5,092
Topic 8	306	3,487	2,098	5,891
Topic 9	257	3,648	1,437	5,342
Topic 10	465	3,446	2,194	6,105
รวม	3,174	34,895	18,041	56,110

ในแต่ละหัวข้อสามารถจำแนกขั้วความรู้สึกออกเป็น 3 กลุ่มได้แก่ ขั้วความรู้สึกเชิงลบ (Negative) ขั้วความรู้สึกเป็นกลาง (Neutral) และขั้วความรู้สึกเชิงบวก (Positive) ส่วนใหญ่ข้อมูลมีขั้วความรู้สึกเป็นกลางมากที่สุดและเชิงลบรองลงมา ซึ่งจากตารางแสดงจำนวนขั้วความรู้สึกของวัคซีนไฟเซอร์ของแต่ละหัวข้อ พบว่า หัวข้อที่ 1 มีจำนวนทวีตมากที่สุด โดยส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นการรออนุมัติรับรองวัคซีนไฟเซอร์จากสำนักงานคณะกรรมการอาหารและยา อีกทั้งหัวข้อที่ 10 มีจำนวนทวีตมากที่สุดอันดับที่ 2 ซึ่งส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นรัฐบาลไทยออกมาชี้แจงประเด็นบุคลากรทางการแพทย์ด่านหน้าไม่ได้รับวัคซีนไฟเซอร์

รองลงมาจากความคิดเห็นส่วนใหญ่นั้นมีความรู้สึกไม่พอใจกับบุคลากรทางการแพทย์ด่านหน้าไม่ได้รับวัคซีนไฟเซอร์

ผลลัพธ์การวิเคราะห์ไคสแควร์ทดสอบความสัมพันธ์ระหว่างหัวข้อกับขั้วความรู้สึกของวัคซีนไฟเซอร์ (พบว่าค่า p – value ที่คำนวณมีค่าน้อยกว่า 0.05 จึงปฏิเสธ ดังนั้น หัวข้อทั้ง 10 หัวข้อมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่ม ที่ระดับนัยสำคัญ 0.05

4.3.4 วัคซีนซิโนฟาร์ม (Sinopharm)

จากตารางที่ 14 สามารถอธิบายได้ว่า วัคซีนซิโนฟาร์ม (Sinopharm) สามารถจำแนกหัวข้อ (Topic) ได้ 10 หัวข้อได้แก่

Topic 1: ระบบลงทะเบียนจองวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือกล่ม

Topic 2: การเลื่อนการจองฉีดวัคซีนซิโนฟาร์ม

Topic 3: ราชวิทยาลัยจุฬาภรณ์นำเข้าวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือก

Topic 4: เว็บการจองของวัคซีนซิโนฟาร์ม

Topic 5: การลงทะเบียนจองรับวัคซีนซิโนฟาร์มของจังหวัดต่างๆ

Topic 6: ภูมิป้องกันโควิดหลังจากการฉีดวัคซีนซิโนฟาร์ม

Topic 7: ขั้นตอนและข้อมูลที่ต้องใช้ในการจองวัคซีนซิโนฟาร์ม

Topic 8: รีวิวอาการของการฉีดวัคซีนซิโนฟาร์มเป็นเข็มแรก

Topic 9: ความต้องการวัคซีนซิโนฟาร์มอย่างสูงของบริษัทและองค์กรต่าง ๆ เพื่อฉีดให้แก่พนักงานของตนเอง

Topic 10: กลุ่มม็อบเรียกร้องให้รัฐบาลนำวัคซีนซิโนฟาร์มมาฉีดฟรีแก่ประชาชน

ตารางที่ 14 ขั้วความรู้สึกของวัคซีนซิโนฟาร์ม (Sinopharm) ของแต่ละหัวข้อ (Topic)

หัวข้อ (Topic)	**ขั้วความรู้สึกเชิงบวก (Positive)**	**ขั้วความรู้สึกเป็นกลาง (Neutral)**	**ขั้วความรู้สึกเชิงลบ (Negative)**	**รวม**
Topic 1	337	5,839	1,923	8,099
Topic 2	425	5,162	1,475	7,062
Topic 3	702	7,179	1,678	9,559
Topic 4	391	4,634	2,029	7,054
Topic 5	398	4,172	1,434	6,004
Topic 6	502	3,847	1,351	5,700
Topic 7	443	4,531	1,396	6,370
Topic 8	1,407	4,783	2,706	8,896
Topic 9	564	5,305	2,035	7,904
Topic 10	340	4,209	1,659	6,208
รวม	5,509	49,661	17,686	72,856

ในแต่ละหัวข้อสามารถจำแนกขั้วความรู้สึกออกเป็น 3 กลุ่มได้แก่ ขั้วความรู้สึกเชิงลบ (Negative) ขั้วความรู้สึกเป็นกลาง (Neutral) และขั้วความรู้สึกเชิงบวก (Positive) ส่วนใหญ่ข้อมูลมีขั้วความรู้สึกเป็นกลางมากที่สุดและเชิงลบรองลงมา ซึ่งจากตารางแสดงจำนวนขั้วความรู้สึกของวัคซีนซิโนฟาร์มของแต่ละหัวข้อ พบว่า หัวข้อที่ 3 จำนวนทวีตมากที่สุด โดยส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นราชวิทยาลัยจุฬาภรณ์นำเข้าวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือกและหัวข้อที่ 8 มีจำนวนทวีตมากที่สุดอันดับที่ 2 ซึ่งส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นรีวิวอาการของการฉีดวัคซีนซิโนฟาร์มเป็นเข็มแรก รองลงมาจากความคิดเห็นส่วนใหญ่นั้นมีความรู้สึกไม่พอใจ เนื่องจากมีอาการปวดแขนและอาการเวียนหัวหลังฉีดวัคซีนซิโนฟาร์มเป็นเข็มแรก

ผลลัพธ์การวิเคราะห์ไคสแควร์ทดสอบความสัมพันธ์ระหว่างหัวข้อกับขั้วความรู้สึกของวัคซีนซิโนฟาร์ม พบว่าค่า p – value ที่คำนวณมีค่าน้อยกว่า 0.05 จึงปฏิเสธ ดังนั้น หัวข้อทั้ง 10 หัวข้อมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่ม ที่ระดับนัยสำคัญ 0.05

4.3.5 วัคซีนแอสตร้าเซนเนกา (AstraZeneca)

จากตารางที่ 15 สามารถอธิบายได้ว่า วัคซีนแอสตร้าเซนเนกา (AstraZeneca) สามารถจำแนกหัวข้อ (Topic) ได้ 8 หัวข้อได้แก่

Topic 1: การผลิตวัคซีนแอสตร้าโดยบริษัทสยามไบโอไซเอนซ์

Topic 2: อาการและผลข้างเคียงของวัคซีนแอสตร้าในของผู้สูงอายุ

Topic 3: การฉีดวัคซีนแอสตร้าเป็นเข็มแรกให้แก่ผู้สูงอายุ

Topic 4: การฉีดวัคซีนสูตรผสม

Topic 5: ความเครียดที่มีต่อการฉีดวัคซีนแอสตร้า

Topic 6: ความกลัวเรื่องอายุที่เหมาะสมสมที่จะฉีดวัคซีนแอสตร้าเซเนกา

Topic 7: รีวิวอาการหลังฉีดวัคซีนแอสตร้าเซเนกา

Topic 8: การรอลงทะเบียนฉีดวัคซีนแอสตร้าเซเนกาของผู้สูงอายุ

ตารางที่ 15 ขั้วความรู้สึกของวัคซีนแอสตร้าเซนเนกา (AstraZeneca) ของแต่ละหัวข้อ (Topic)

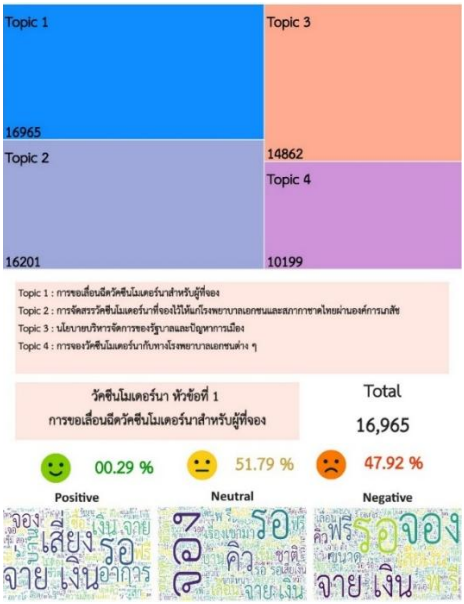
หัวข้อ (Topic)	ขั้วความรู้สึกเชิงบวก (Positive)	ขั้วความรู้สึกเป็นกลาง (Neutral)	ขั้วความรู้สึกเชิงลบ (Negative)	รวม
Topic 1	76	6,553	3,119	9,748
Topic 2	103	4,574	2,717	7,394
Topic 3	65	4,505	1,901	6,471
Topic 4	56	3,955	2,154	6,165
Topic 5	87	3,687	2,473	6,247
Topic 6	58	4,304	2,558	6,920
Topic 7	219	4,640	7,649	12,508
Topic 8	121	5,237	2,864	8,222
รวม	785	37,455	25,435	63,675

ในแต่ละหัวข้อสามารถจำแนกขั้วความรู้สึกออกเป็น 3 กลุ่มได้แก่ ขั้วความรู้สึกเชิงลบ (Negative) ขั้วความรู้สึกเป็นกลาง (Neutral) และขั้วความรู้สึกเชิงบวก (Positive) ส่วนใหญ่ข้อมูลมีขั้วความรู้สึกเป็นกลางมากที่สุดและเชิงลบรองลงมา ซึ่งจากตารางแสดงจำนวนขั้วความรู้สึกของวัคซีนแอสตร้าเซเนกาของแต่ละหัวข้อ พบว่า หัวข้อที่ 7 มีจำนวนทวีตมากที่สุด โดยส่วนใหญ่ไม่พอใจแก่ประเด็นรีวิวอาการหลังฉีดวัคซีนแอสตร้าเซเนกา เนื่องจากมีอาการปวดแขนข้างที่ฉีดและมีอาการไข้ขึ้นสูงในช่วงกลางคืน หัวข้อที่ 1 มีจำนวนทวีตมากที่สุดอันดับที่ 2 ซึ่งส่วนใหญ่มีการแสดงความรู้สึกเป็นกลางแก่ประเด็นการผลิตวัคซีนแอสตร้าโดยบริษัทสยามไบโอไซเอนซ์

ผลลัพธ์การวิเคราะห์ไคสแควร์ทดสอบความสัมพันธ์ระหว่างหัวข้อกับขั้วความรู้สึกของวัคซีนแอสตร้าเซนเนกา พบว่าค่า p – value ที่คำนวณมีค่าน้อยกว่า 0.05 จึงปฏิเสธ ดังนั้น หัวข้อทั้ง 8 หัวข้อมีความสัมพันธ์กับขั้วความรู้สึกทั้ง 3 กลุ่ม ที่ระดับนัยสำคัญ 0.05

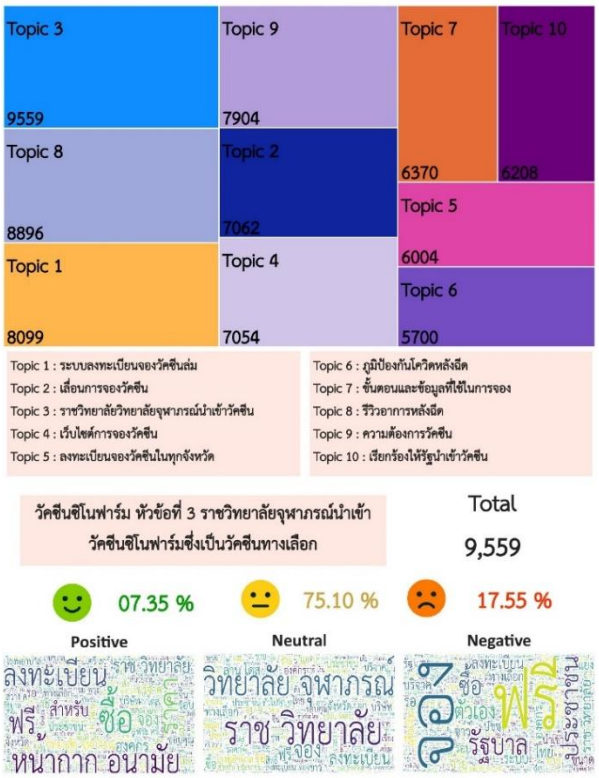
4.4 Dashboard จำแนกตามชื่อทางธุรกิจของวัคซีน

จากรูปที่ 5 - 9 แสดง Dashboard จำแนกตามชื่อทางธุรกิจของวัคซีน โดยภาพแสดงสัดส่วนของจำนวนทวีตในแต่ละหัวข้อ พร้อมทั้งนำเสนอร้อยละของทวีตในหัวข้อที่ประชาชนแสดงความคิดเห็นบนสื่อออนไลน์มากที่สุด จำแนกตามขั้วความรู้สึก และ คำที่ใช้ในการแสดงความคิดเห็นตามขั้วความรู้สึก

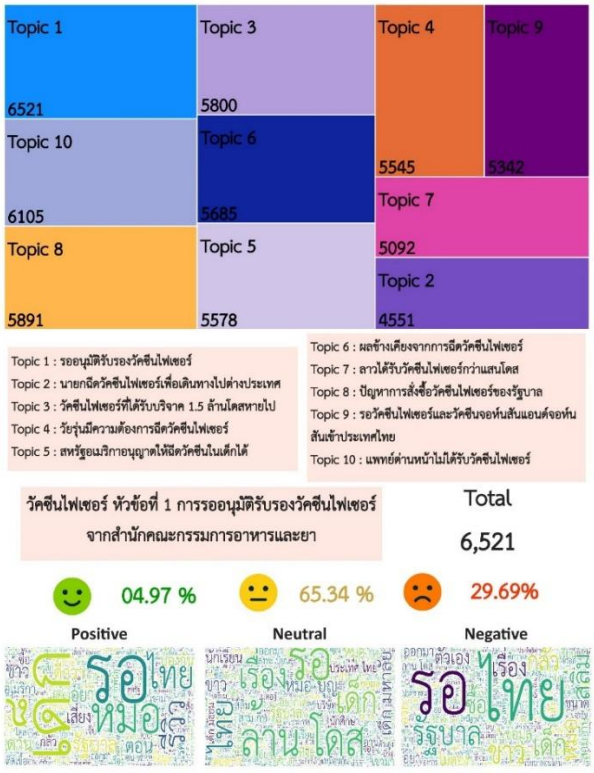


รูปที่ 5 ผลลัพธ์ Dashboard ของวัคซีนโมเดอร์นา

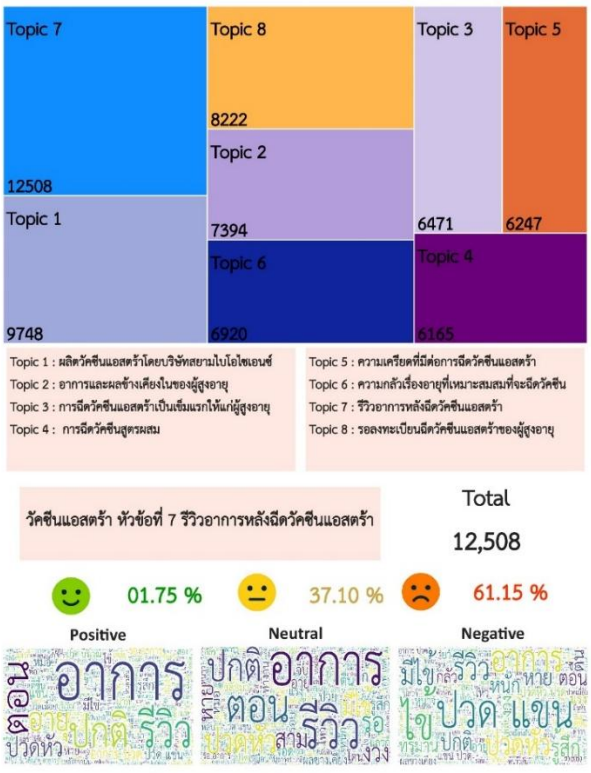
รูปที่ 6 ผลลัพธ์ Dashboard ของวัคซีนซิโนแวค



รูปที่ 8 ผลลัพธ์ Dashboard ของวัคซีนซิโนฟาร์ม



รูปที่ 7 ผลลัพธ์ Dashboard ของวัคซีนไฟเซอร์



รูปที่ 9 ผลลัพธ์ Dashboard ของวัคซีนแอสตร้าเซเนกา

4.5 อภิปรายผลการวิจัย

การวิเคราะห์หัวข้อด้วยตัวแบบ LDA จำเป็นต้องระบุไฮเปอร์พารามิเตอร์ (Hyperparameters) ได้แก่ จำนวนหัวข้อ (k), แอลฟ่า (Alpha), เบต้า (Beta) โดยค่าพารามิเตอร์ที่เหมาะสมสามารถพิจารณาได้จากค่า Coherence Score ซึ่งงานวิจัยของ [12] ได้ทำการศึกษาเกี่ยวกับตัวแบบ LDA และการปรับแต่งค่าพารามิเตอร์ของตัวแบบ LDA โดยมีการทดลองกำหนดค่าแอลฟ่าเท่ากับ 0.01, 0.31, 0.61, 0.91 ผลการศึกษาพบว่าควรกำหนดค่าแอลฟ่าให้มีค่าน้อย เนื่องจากแอลฟ่ายิ่งมีค่าต่ำจะหมายถึงการกระจายตัวของคำในแต่ละหัวข้อจะค่อนข้างต่ำ

ข้อความทวีตในช่วงระยะเวลาที่ศึกษาที่ถูกกล่าวถึงมากสุด 3 อันดับแรก ได้แก่วัคซีนซิโนฟาร์ม วัคซีนแอสตร้าเซนเนกา และ วัคซีนโมเดอร์นา ตามลำดับ โดยทั้งวัคซีนซิโนฟาร์มและวัคซีนแอสตร้าเซนเนกา เป็นวัคซีนทางเลือกซึ่งหมายถึงวัคซีนโควิด-19 ที่รัฐบาลอนุมัติให้โรงพยาบาลเอกชนเป็นผู้จัดซื้อผ่านองค์การเภสัช นอกเหนือจากวัคซีนที่อยู่ในแผนการจัดซื้อของกระทรวงสาธารณสุข ดังนั้นหัวข้อที่กล่าวถึงวัคซีนทั้งสอง เกี่ยวข้องกับ ความยุ่งยากของระบบการจองวัคซีน รวมถึงความล่าช้าในการนำเข้าวัคซีน ส่วนวัคซีนแอสตร้าเซนเนกาเป็นวัคซีนหลักของประเทศไทยซึ่งจะกล่าวถึงอาการหลังการฉีดวัคซีน เป็นส่วนใหญ่ เช่นเดียวกับวัคซีนซิโนแวค สำหรับวัคซีนไฟเซอร์เป็นวัคซีนหลักลำดับที่สามของประเทศที่ประชาชนรอคอยให้ได้รับอนุมัติขึ้นทะเบียนจากสำนักคณะกรรมการอาหารและยา

ส่วนใหญ่ขั้วความรู้สึกของประชาชนไทยที่แสดงความคิดเห็นบนสื่อสังคมออนไลน์ มักเป็นกลาง ตามด้วยขั้วความรู้สึกเชิงลบ และมีขั้วความรู้สึกเชิงบวกเพียงเล็กน้อย นอกจากนี้ยังดูเหมือนว่าข้อความขั้วความรู้สึกเชิงลบจะมากกว่าขั้วความรู้สึกเชิงบวก ซึ่งแตกต่างจากงานวิจัยของ [16] กล่าวว่า ส่วนใหญ่ของผู้ใช้ทวิตเตอร์แสดงขั้วความรู้สึกเชิงบวกสูงถึง 46.51% ขั้วความรู้สึกเป็นกลาง 28.70 % และขั้วความรู้สึกเชิงลบเพียง 23.81 % โดยมักพบข้อความเชิงลบในช่วงแรก ต่อจากนั้นจะเริ่มคงที่ และข้อความเชิงบวกจะค่อย ๆ เพิ่มสูงขึ้น เช่นเดียวกับงานวิจัยของ [24] รายงานว่า ขั้วความรู้สึกความคิดเห็นเกี่ยวกับวัคซีนในประเทศอินเดีย สหรัฐอเมริกา แคนาดา และอังกฤษ โดยส่วนใหญ่ของผู้ใช้ทวิตเตอร์มักแสดงความรู้สึกเชิงบวก

นอกจากนี้ผลการศึกษาหัวข้อความรู้สึกที่มีต่อวัคซีนชนิดต่าง ๆ ส่วนใหญ่หัวข้อที่แสดงถึงขั้วความรู้สึกเชิงลบมีการกล่าวถึงประเด็นต่าง ๆ ได้แก่ การลงทะเบียนจองวัคซีนโดยเฉพาะปัญหาการลงทะเบียนผ่านเว็บไซต์ ความไม่เพียงพอของวัคซีนเมื่อเทียบกับความต้องการภายในประเทศไทย การบริหารการจัดการหรือจัดซื้อวัคซีนของรัฐบาล ปัญหาการจัดสรรวัคซีนไปยังโรงพยาบาลต่าง ๆ ไม่ทั่วถึง ปัญหาความล่าช้าในการสั่งซื้อที่ล่าช้า รวมทั้งอาการและผลข้างเคียงของวัคซีนเช่น ปวดแขน ไข้ขึ้นสูง ซึ่งสอดคล้องกับงานวิจัยของ [24] ที่ศึกษาขั้วความรู้สึกและหัวข้อของความคิดเห็นเกี่ยวกับวัคซีนในประเทศอินเดีย สหรัฐอเมริกา แคนาดา และอังกฤษ ซึ่งหัวข้อที่แสดงถึงขั้วความรู้สึกเชิงลบส่วนใหญ่มีการกล่าวถึงอาการ ผลข้างเคียงของวัคซีนที่ไม่พึงประสงค์ เช่น ปวดแขน ไข้ขึ้นสูง อีกทั้งยังสอดคล้องกับงานวิจัยของ [16] ที่ศึกษาหัวข้อขั้วความรู้สึกของวัคซีนป้องกันโควิด – 19 หลังจากการเปิดตัวของวัคซีน พบว่า ประเด็นหัวข้อที่แสดงถึงขั้วความรู้สึกเชิงลบ โดยส่วนใหญ่มีการกล่าวถึงอาการและผลข้างเคียง เช่น มีอาการปวดในบริเวณที่มีการวัคซีน เป็นต้น

5. สรุปผลการวิจัยและข้อเสนอแนะ

5.1 การศึกษาการทำเหมืองข้อความ (Text Mining) เกี่ยวกับความคิดเห็นที่มีต่อวัคซีนจากทวิตเตอร์ พบว่า จำนวนทวีตที่มีการทวีตเกี่ยวกับวัคซีนในเดือนมีนาคม 2654 ถึง ตุลาคม 2564 มากที่สุดเป็นลำดับ 1 คือ วัคซีนแอสตร้าเซเนกา (AstraZeneca) มีด้วยกันทั้งสิ้น 106,668 ทวีต คิดเป็นร้อยละ 23.64 ของทวีตทั้งหมด ลำดับที่ 2 คือ วัคซีนซิโนฟาร์ม (Sinopharm) มีด้วยกันทั้งสิ้น 103,125 ทวีต คิดเป็นร้อยละ 22.86 ของทวีตทั้งหมด ลำดับที่ 3 คือ วัคซีนไฟเซอร์ (Pfizer) มีด้วยกันทั้งสิ้น 88,792 ทวีต คิดเป็นร้อยละ 19.68 ของทวีตทั้งหมด ลำดับที่ 4 คือ วัคซีนซิโนแวค (Sinovac) มีด้วยกันทั้งสิ้น 76,362 ทวีต คิดเป็นร้อยละ 16.92 ของทวีตทั้งหมด และลำดับที่ 5 คือ วัคซีนโมเดอร์นา (Moderna) มีด้วยกันทั้งสิ้น 76,262 ทวีต คิดเป็นร้อยละ 16.92 ของทวีตทั้งหมด และในการตัดคำภาษาไทยได้ใช้

วิธีการตัดคำโดยใช้พจนานุกรม (Dictionary-Base Approach) และ การตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) โดยใช้พจนานุกรมนั้นจะใช้ของ PythaiNLP ทั้งนี้การตัดคำจะเริ่มพิจารณาจากซ้ายไปขวาตามหลักการเขียนภาษาไทย จากนั้นการนำคำหยุด (Stop Word) ออกเพื่อลดขนาดของข้อมูลที่ใช้ในการประมวลผล และการให้น้ำหนักของคำแบบ TF-IDF กล่าวคือ น้ำหนักของคำจะสูง ถ้าคำนั้น ๆ ปรากฏบ่อยครั้งในทวีตจำนวนไม่มากนัก และน้ำหนักของคำจะต่ำลง ถ้าคำนั้น ๆ ปรากฏบ่อยครั้งในเอกสารหนึ่ง หรือปรากฏบ่อยครั้งในเอกสารจำนวนมาก แสดงว่าเป็นคำที่ไม่สามารถเป็นตัวแทนของเอกสารได้

5.2 การสำรวจหัวข้อความคิดเห็นโดยสร้างตัวแบบการจัดสรรดีรีเคลแฝง (LDA) ซึ่งต้องกำหนดจำนวนหัวข้อที่เหมาะสมโดยพิจารณาจากค่า Coherence สามารถสรุปผลได้ดังนี้

- จำนวนหัวข้อที่เหมาะสมของทวีตที่มีวัคซีนโมเดอร์นาเป็นคำสำคัญ เท่ากับ 4 หัวข้อ ประกอบด้วยหัวข้อ 1) การขอเลื่อนฉีดวัคซีนโมเดอร์นาสำหรับผู้ที่จอง 2) การจัดสรรวัคซีนโมเดอร์นาที่จองไว้ให้แก่โรงพยาบาลเอกชนและสภากาชาดไทยผ่านองค์การเภสัช 3) นโยบายบริหารจัดการของรัฐบาลและปัญหาการเมือง 4) การจองวัคซีนโมเดอร์นากับทางโรงพยาบาลเอกชนต่าง ๆ
- จำนวนหัวข้อที่เหมาะสมของทวีตที่มีวัคซีนซิโนฟาร์มเป็นคำสำคัญ เท่ากับ 10 หัวข้อ ประกอบด้วยหัวข้อ 1) ระบบลงทะเบียนจองวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือกล่ม 2) การเลื่อนการจองฉีดวัคซีนซิโนฟาร์ม 3) ราชวิทยาลัยจุฬาภรณ์นำเข้าวัคซีนซิโนฟาร์มซึ่งเป็นวัคซีนทางเลือก 4) เว็บการจองของวัคซีนซิโนฟาร์ม 5) การลงทะเบียนจองรับวัคซีนซิโนฟาร์มของจังหวัดต่าง ๆ 6) ภูมิป้องกันโควิดหลังจากการฉีดวัคซีนซิโนฟาร์ม 7) ขั้นตอนและข้อมูลที่ต้องใช้ในการจองวัคซีนซิโนฟาร์ม 8) รีวิวอาการของการฉีดวัคซีนซิโนฟาร์มเป็นเข็มแรก 9) ความต้องการวัคซีนซิโนฟาร์มอย่างสูงของบริษัทและองค์กรต่าง ๆ เพื่อฉีดให้แก่พนักงานของตนเอง 10) กลุ่มม็อบเรียกร้องให้รัฐบาลนำวัคซีนซิโนฟาร์มมาฉีดฟรีแก่ประชาชน
- จำนวนหัวข้อที่เหมาะสมของทวีตที่มีวัคซีนไฟเซอร์เป็นคำสำคัญ เท่ากับ 10 หัวข้อ ประกอบด้วย 1) การรออนุมัติรับรองวัคซีนไฟเซอร์จากสำนักงานคณะกรรมการอาหารและยา 2) นายกรัฐมนตรีฉีดวัคซีนไฟเซอร์เพื่อเดินทางไปต่างประเทศ 3) ข่าววัคซีนไฟเซอร์ที่ได้รับบริจาค 1.5 ล้านโดสหายไป ทำให้แพทย์ด่านหน้าไม่ได้รับวัคซีน 4) เด็กวัยรุ่นมีความต้องการฉีดวัคซีนไฟเซอร์ 5) ประเทศสหรัฐอเมริกาอนุญาตให้ฉีดวัคซีนในเด็กได้ แต่ข่าวลือเกี่ยวกับวัคซีนไฟเซอร์ทำให้เกิดอาการบางอย่างในเด็ก 6) ผลข้างเคียงจากการฉีดวัคซีนไฟเซอร์ 7) สาธารณรัฐประชาธิปไตยประชาชนลาวได้รับวัคซีนไฟเซอร์กว่าแสนโดส โดยมีข่าวคนไทยแอบไปเก็บเห็ดและถูกจับที่ลาวได้รับการฉีดวัคซีนไฟเซอร์ 8) ปัญหาการสั่งซื้อวัคซีนไฟเซอร์เพื่อป้องกันโควิดของรัฐบาล 9) การรอวัคซีนไฟเซอร์และวัคซีนจอห์นสันแอนด์จอห์นสันเข้าประเทศไทย 10) บุคลากรทางการแพทย์ด่านหน้าไม่ได้รับวัคซีนไฟเซอร์
- จำนวนหัวข้อที่เหมาะสมของทวีตที่มีวัคซีนซิโนแวคเป็นคำสำคัญเท่ากับ 9 หัวข้อ ได้แก่ 1) ดาราไทยเข้ารับการฉีดวัคซีนซิโนแวค 2) การบังคับให้ฉีดซิโนแวค 3) ดาราไทยออกมารีวิวอาการและผลข้างเคียงหลังฉีดซิโนแวค 4) คุณหมอยงและกลุ่มประชาชนบางกลุ่มสนับสนุนการฉีดวัคซีนซิโนแวค 5) รัฐบาลเปิดลงทะเบียนฉีดวัคซีนซิโนแวคฟรี 6) ความวิตกกังวลเกี่ยวกับภูมิคุ้มกันและการเสียชีวิตจากการฉีดวัคซีน 7) การบังคับให้บุคลากรฉีดวัคซีนซิโนแวคที่ผลิตจากสาธารณรัฐประชาชนจีน 8) รัฐบาลไทยออกมาชี้แจงประเด็นการด้อยค่าวัคซีนซิโนแวค 9) แพทย์ไทยแสดงความคิดเห็นต่อรัฐบาลให้หยุดการสั่งซื้อวัคซีนซิโนแวค
- จำนวนหัวข้อที่เหมาะสมของทวีตที่มีวัคซีนแอสตร้าเซเนกาเป็นคำสำคัญ เท่ากับ 8 หัวข้อ ได้แก่ 1) การผลิตวัคซีนแอสตร้าโดยบริษัทสยามไบโอไซเอนซ์ 2) อาการและผลข้างเคียงของวัคซีนแอสตร้าในของผู้สูงอายุ 3) การฉีดวัคซีนแอสตร้าเป็นเข็มแรกให้แก่

ผู้สูงอายุ 4) การฉีดวัคซีนสูตรผสม 5) ความเครียด ที่มีต่อการฉีดวัคซีนแอสตร้า 6) ความกลัวเรื่องอายุที่เหมาะสมที่จะฉีดวัคซีนแอสตร้าเซเนกา 7)รีวิวอาการหลังฉีดวัคซีนแอสตร้า 8)การรอลงทะเบียนฉีดวัคซีนแอสตร้าเซเนกาของผู้สูงอายุ

5.3 การวิเคราะห์ขั้วความรู้สึก (Sentiment Analysis)

จะเห็นได้ว่า ในภาพรวมของผู้ใช้ทวิตเตอร์ที่เป็นประชาชนไทยที่มีความรู้สึกเป็นกลางต่อวัคซีนมากที่สุด นอกจากนี้จำนวนทวีตของขั้วความรู้สึกทางเชิงลบมากกว่าเชิงบวก โดยหัวข้อของทวีตที่มีขั้วความรู้สึกเชิงลบ ได้แก่ ประเด็นความไม่เพียงพอของวัคซีน ประเด็นการบริหารจัดการสั่งซื้อของรัฐบาล ประเด็นระบบการจองหรือสั่งซื้อวัคซีนทางเลือก ประเด็นอาการและผลข้างเคียงหลังการฉีดวัคซีน หรือแม้แต่ประเด็นการรีวิวการฉีดวัคซีนของดาราไทยหรือผู้มีชื่อเสียงก็ตาม

5.4 แนวทางการประยุกต์ใช้ผลการวิเคราะห์ในเชิงนโยบายหรือการกำหนดมาตรฐาน

จากผลการวิเคราะห์ความคิดเห็นของประชาชนที่มีต่อวัคซีน COVID-19 ชนิดต่าง ๆ ผ่านเทคนิคเหมืองข้อความ หน่วยงานด้านสาธารณสุขสามารถนำไปใช้เป็นแนวทางในการกำหนดมาตรฐานการสื่อสารสาธารณะในช่วงวิกฤต โดยเฉพาะในการตอบสนองต่อความรู้สึกของประชาชน อันจะช่วยเพิ่มความเชื่อมั่นในการบริหารจัดการวัคซีนในอนาคต เช่น การออกแบบเนื้อหาสื่อประชาสัมพันธ์ให้ตรงกับความคาดหวังและข้อกังวลของแต่ละกลุ่มประชากร

5.5 แนวทางการใช้งานสำหรับภาคอุตสาหกรรม

ผลการศึกษาในครั้งนี้ยังสามารถนำไปต่อยอดในภาคอุตสาหกรรม โดยเฉพาะบริษัทผู้ผลิตหรือจัดจำหน่ายวัคซีน ที่สามารถนำผลวิเคราะห์ไปใช้ในการวางแผนกลยุทธ์การสื่อสารทางการตลาด การปรับปรุงผลิตภัณฑ์ และการสร้างความเชื่อมั่นต่อแบรนด์ นอกจากนี้ แนวทางการวิเคราะห์ข้อความจากสื่อสังคมออนไลน์ยังสามารถประยุกต์ใช้กับผลิตภัณฑ์สุขภาพชนิดอื่น ๆ เพื่อใช้เป็นเครื่องมือวิเคราะห์แนวโน้มความพึงพอใจของลูกค้าในอนาคต

5.6 แนวทางการศึกษาต่อยอดงานวิจัยในอนาคต ได้แก่

- ผลการศึกษาเป็นการศึกษาความคิดเห็นของประชาชนที่แสดงความคิดเห็นเกี่ยวกับวัคซีนผ่านสื่อสังคมออนไลน์ ดังนั้นข้อมูลที่ใช้วิเคราะห์นั้นเป็นข้อมูลของผู้ใช้ทวิตเตอร์ภาษาไทย อาจไม่ครอบคลุมความคิดเห็นของประชาชนที่ใช้ช่องทางหรือ แพลตฟอร์มอื่น ๆ ในการแสดงความคิดเห็น โดยอาจสำรวจความคิดเห็นของคนไทยในแพลตฟอร์มอื่นๆ เช่นเฟซบุ๊ก (Facebook) ยูทูป (YouTube) และเว็บไซต์ต่าง ๆ เป็นต้น
- อาจทำการวิเคราะห์ขั้วความรู้สึกจำแนกตามช่วงเวลา นอกเหนือจากการวิเคราะห์ความรู้สึกจำแนกตามหัวข้อ
- สามารถศึกษาเพิ่มเติม และเปรียบประสิทธิภาพของอัลกอริทึมต่าง ๆ ที่ใช้ในการสร้างตัวแบบหัวข้อ (Topic Modeling) เช่น LSA, NFM เป็นต้น
- อาจศึกษาเพิ่มเติมเกี่ยวกับอัลกอริทึมที่ใช้ในการทำการวิเคราะห์ขั้วความรู้สึก นอกเหนือ Logistic Regression เช่น Long Short-Term Memory (LTSM), Support Vector Machine (SVM), Gated Recurrent Unit (GRU), Recurrent Neural Network (RNN), Convolutional Neural Network (CNN) เป็นต้น

6. เอกสารอ้างอิง

[1] กรมควบคุมโรค, “รายงานความก้าวหน้าการให้บริการฉีดวัคซีนโควิด19,” [ออนไลน์]. แหล่งที่มา: https://ddc.moph.go.th/. [วันที่เข้าถึง 31 มีนาคม 2565].

[2] กรมควบคุมโรค, “เผยวัคซีนไฟเซอร์ 1.5 ล้านโดส ถึงไทยตามกำหนด,” [ออนไลน์]. แหล่งที่มา: https://ddc.moph.go.th/brc/news.php?news=21341&deptcode=brc. [วันที่เข้าถึง 31 มีนาคม 2565].

[3] กระทรวงการท่องเที่ยวและกีฬา, “สถิตินักท่องเที่ยว,” [ออนไลน์]. แหล่งที่มา:

https://mots.go.th/more_news_new.php?cid=411.[31 มีนาคม 2565].

[4] กระทรวงสาธารณสุข, “สธ.เซ็นสัญญาซื้อวัคซีนแอสตร้าเซนเนก้า 60 ล้านโดสสำหรับปี 2565,” [ออนไลน์]. แหล่งที่มา: https://pr.moph.go.th/?url=pr/detail/2/04/164808/.[วันที่เข้าถึง 31 มีนาคม 2565].

[5] กานดา แผ่วัฒนากุล, “การวิเคราะห์เหมืองข้อเสนอแนะจากบทรายการโทรทัศน์,” วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, สาขาบริหารเทคโนโลยีสารสนเทศ คณะสถิติประยุกต์, สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2555.

[6] กิริฎา เภาพิจิตร และกิตติพัฒน์ บัวอุบ, “เศรษฐกิจไทยปี 64 ในวิกฤติโควิดระลอกใหม่,” [ออนไลน์]. แหล่งที่มา: https://tdri.or.th/2021/01/economic-outlook-2021/. [วันที่เข้าถึง 3 เมษายน 2565].

[7] ธนกร ญานกาย และวนิดา แก่นอากาศ, “การเพิ่มประสิทธิภาพแบบจำลองหัวข้อด้วยสภาพแวดล้อมแบบข้อมูลขนาดใหญ่,” *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยอุบลราชธานี*., ปีที่ 21, ฉบับที่ 3, น. 166-174, 2562.

[8] ภูมิรพี ภูมิค้า, “เทคนิคการเรียนรู้เชิงลึกเพื่อวิเคราะห์ความรู้สึกจากผู้ใช้ผลิตภัณฑ์,” วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต, สาขาวิศวกรรมโทรคมนาคมและคอมพิวเตอร์, มหาวิทยาลัยเทคโนโลยีสุรนารี, 2562.

[9] ราชวิทยาลัยจุฬาภรณ์, “ราชวิทยาลัยจุฬาภรณ์ กระทรวงสาธารณสุข และ อย. ผสานความร่วมมือนำเข้าวัคซีนโควิดทางเลือก ซิโนฟาร์ม,” [ออนไลน์]. แหล่งที่มา: https://www.chulabhornchannel.com/news-activities/vaccine/2021/05/. [วันที่เข้าถึง 16 มีนาคม 2565].

[10] วาทิตย์ คำพรมมา, จักรชัย โสอินทร์ และเพรช อิ่มทองคำ, “แบบจำลองกำรวิเคราะห์ความรู้สึกแบบผสมสำหรับความคิดเห็นต่อโรงแรมในประเทศไทยโดยใช้ K-means และ K-NN,” ในการประชุมวิชาการระดับชาติ สารสนเทศศาสตร์วิชาการ 2019, 2562.

[11] D. M. Blei, A. Y. Ng, M. I. Jordan, 2003. “Latent dirichlet allocation,” *Journal of Machine Learning Research*., vol. 3, pp. 993–1022, 2003.

[12] D. Buenaño-Fernandez, M. González, D. Gil and S. Luján-Mora, "Text mining of open-ended questions in self-assessment of university teachers: an LDA topic modeling approach," in *IEEE Access*, vol. 8, pp. 35318-35330, 2020.

[13] H. Cherfi, A. Napoli and Y. Toussaint, “Towards a text mining methodology using association rule extraction,” *Soft Computing*., vol. 10, no. 5, pp. 431-441, 2005.

[14] Data Reportal, “Digital 2021: Thailand.” [ออนไลน์]. แหล่งที่มา: https://datareportal.com/reports/digital-2021-thailand. [วันที่เข้าถึง 3 เมษายน 2565].

[15] R. Feldman and J. Sanger, *The Text Mining Handbook*, Boston: Cambridge University Press, 2006.

[16] L. Huangfu, Y. Mo, P. Zhang, D. D. Zeng, S. He, “COVID-19 vaccine tweets after vaccine rollout: sentiment–based topic modeling,” *Journal of Medical Internet Research*., vol. 24, no. 2, pp. e31726, 2022.

[17] B. Liu, *Sentiment Analysis and Opinion Mining*, California: Morgan & Claypool Publishers, 2012.

[18] J. C. Lyu, E. L. Han and G. K. Luli, “COVID-19 Vaccine–Related Discussion on Twitter: Topic modeling and sentiment analysis,” *Journal of Medical Internet Research*., vol. 23, no. 6, pp. 24303, 2021.

[19] MGR Online, “Twitter ไทย อัดแน่นข้อความเกี่ยวกับโควิด-19 ทะลุ 73 ล้านทวีต,” [ออนไลน์]. แหล่งที่มา:

https://mgronline.com/cyberbiz/detail/9640000086249. [วันที่เข้าถึง 8 เมษายน 2565].

[20] T. C W. Na, D. Li, W. Lu and H. Li, “Insight from NLP Analysis: Covid-19 Vaccines Sentiments on Social Media,” [ออนไลน์]. แหล่งที่มา: https://arxiv.org/abs/2106.04081. [วันที่เข้าถึง 8 เมษายน 2565].

[21] S. Petrović, “Collocation Extraction Measures for Text Mining Applications,” Doctoral dissertation, Fakultetelektrotehnike i računarstva, Sveučilište u Zagrebu, 2007.

[22] SV. Praveen, J. M. Lorenz, R. Ittamalla, K. Dhama, C. Chakraborty, C. V. S. Kumar and T. Mohan, “twitter-based sentiment analysis and topic modeling of social media posts using natural language processing, to understand people’s perspectives regarding COVID-19 booster vaccine shots in india: crucial to expanding vaccination coverage,” *Vaccines*, vol. 10, no. 11, pp. 1929, 2022.

[23] World Health Organization, “Coronavirus (COVID-19) question and answers,” [Online]. Available: https://www.who.int/thailand/emergencies/novel-coronavirus-2019/q-a-on-covid-19. [Accessed 4 March 2022].

[24] H. Yin, X. Song, S. Yang, J. Li, “sentiment analysis and topic modeling for COVID-19 vaccine discussions,” *World Wide Web*., vol. 25, no. 1, pp. 1067-1083, 2022.