การจำแนกประเภทคนรักสุขภาพจากความคิดเห็นในสื่อสังคมออนไลน์ (พันทิป) เกี่ยวกับประกันโควิด-19 โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติและแบบจำลองการเรียนรู้เชิงลึกที่หลากหลาย

HEALTH LOVER CLASSIFICATION BASED ON COMMENTS ABOUT COVID-19 INSURANCE FROM SOCIAL MEDIA (PANTIP) USING NATURAL LANGUAGE PROCESSING AND VARIOUS DEEP LEARNING MODELS

คงภพ ไชยศร

KONGPOP CHAIYASON

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2565

KMITL-2022-SC-M-017-082

HEALTH LOVER CLASSIFICATION BASED ON COMMENTS ABOUT COVID-19 INSURANCE FROM SOCIAL MEDIA (PANTIP) USING NATURAL LANGUAGE PROCESSING AND VARIOUS DEEP LEARNING MODELS

KONGPOP CHAIYASON

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE PROGRAM IN DATA SCIENCE AND ANALYTICS
KMITL-DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2022

KMITL-2022-SC-M-017-082

หัวข้อวิทยานิพนธ์	การจำแนกประเภทคนรักสุขภาพจากความคิดเห็นในสื่อสังคมออนไลน์ (พันทิป) เกี่ยวกับประกันโควิด-19 โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติและแบบจำลองการเรียนรู้เชิงลึกที่หลากหลาย
ชื่อนักศึกษา	นายคงภพ ไชยศร
รหัสประจำตัว	63605058
ปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์) ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ.	2565
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร.บุษยมาส พิมพ์พรรณชาติ

บทคัดย่อ

ปัจจุบันข้อมูลความคิดเห็นในสังคมออนไลน์กำลังได้รับความนิยมและเข้าถึงได้ง่าย แต่เนื่องจากความคิดเห็นที่เป็นภาษาไทยมีความซับซ้อนมากกว่าภาษาอังกฤษ ทำให้ข้อมูลบางส่วนยังไม่ได้ถูกนำมาวิเคราะห์ เพื่อให้เป็นประโยชน์ต่อธุรกิจในการปรับปรุงผลิตภัณฑ์และการบริการ งานวิจัยนี้นำเสนอการจำแนกประเภทคนรักสุขภาพจากความคิดเห็นในสื่อสังคมออนไลน์ โดยใช้ข้อมูลของผู้ที่แสดงความคิดเห็นในกระทู้ออนไลน์พันทิป ซึ่งเป็นกระดานสนทนาออนไลน์ได้ที่รับความนิยมเป็นอย่างมากในประเทศไทย โดยรวบรวมข้อมูลที่เกี่ยวกับโรคติดเชื้อไวรัสโคโรนา 2019 และประกันภัยในกระทู้ประเภทแสดงความคิดเห็นและประเภทรีวิว งานวิจัยนี้แบ่งออกเป็น 2 ส่วน ได้แก่ 1) การจำแนกประเภทคนรักสุขภาพ โดยประยุกต์ใช้เทคนิคการแปลงประโยคเป็นเวกเตอร์ร่วมกับเทคนิคความคล้ายคลึงของโคไซน์ ทำการทดลองทั้งหมด 297 แบบจำลอง เพื่อเปรียบเทียบแบบจำลองที่ดีที่สุด 2) การจำแนกทัศนคติของผู้ที่แสดงความคิดเห็น โดยใช้เทคนิคการวิเคราะห์ความรู้สึกว่าเป็นเชิงบวกหรือเชิงลบ ซึ่งงานวิจัยนี้ใช้แบบจำลองการเรียนรู้เชิงลึก ทั้งหมด 3 แบบจำลองคือ Gated Recurrent Unit, Bi-Long Short Term Memory และ Convolutional Neural Network ซึ่งแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด 1) ใช้ตัวแปรที่ผ่านกระบวนการสกัดคุณลักษณะเพียงอย่างเดียว 2) เพิ่มตัวแปรประเภทของกระทู้รีวิว เพื่อเปรียบเทียบแบบจำลอง โดยทำการทดลองทั้งหมด 6 แบบจำลอง จากผลการทดลองพบว่า 1) วิธีการที่ทางผู้วิจัยนำเสนอสามารถพัฒนาค่าความถูกต้องจาก 0.539% เป็น 0.725% ด้วยวิธีการใช้เทคนิคการแปลงประโยคเป็นตัวเลขด้วยค่าเฉลี่ย และใช้เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ที่ 91% 2) แบบจำลอง GRU ในชุดที่ 2 ที่ทางผู้วิจัยนำเสนอเป็นแบบจำลองที่ดีที่สุดโดยให้ค่าความถูกต้องสูงที่สุดที่ 0.85 และให้ค่าความผิดพลาดที่ 0.60 ซึ่งผลการทดลองนี้จะเป็นประโยชน์ต่อบริษัทประกันในประเทศไทย สำหรับการพัฒนาประกันโควิด-19 ในการตัดสินใจเลือกบุคคลที่มีแนวโน้มจะเป็นลูกค้า

คำสำคัญ : การประมวลผลภาษาธรรมชาติ โรคติดเชื้อไวรัสโคโรนา 2019 เทคนิคความคล้ายคลึงของโคไซน์ แบบจำลองการเรียนรู้เชิงลึก ประกันภัย สังคมออนไลน์ สุขภาพ

Thesis Title	Health Lover Classification based on Comments about COVID-19 Insurance from Social Media (PANTIP) using Natural Language Processing and Various Deep Learning Models
Student Name	Kongpop Chaiyason
Student ID	63605058
Degree	Master of Science (Data Science and Analytics) KMITL-Digital Analytics and Intelligence Center
Year	2022
Thesis Advisor	Asst. Prof. Dr. Busayamas Pimpunchat

Abstract

Today, social media comments are becoming more popular and accessible. Due to the opinions in Thai are more complicated than English, therefore some information has not yet been analyzed. To benefit the business in improving products and services, this paper proposed health lover classification based on comments from social media used the information about the people who commented on the Pantip online forum which is very popular in Thailand. The data were collected related to coronavirus disease 2019 and insurance in comment and review categories. This research consists of 2 parts: 1) Find customers who tend to be health lovers by applying sentence embedding technique along with the cosine similarity technique comparing with 297 models to find the best model. 2) Classify the attitudes of those who express their opinions whether it is positive or negative using deep learning techniques. The three models were Gated Recurrent Unit, Bi-Long Short Term Memory, and Convolutional Neural Network. Each model was developed into two sets: 1) Used only feature extraction on the process and 2) Added the types of review variable to compare the models by 6 simulations. The experimental results indicated that, the first part improved accuracy from 0.539% to 0.725% using the Mean Sentence Embedding Method with 91% of cosine similarity. In the second part, the second set of GRU model that we provided is the best, with the highest accuracy of 0.85 and the error value of 0.60. This information would benefit insurance companies in Thailand to develop the COVID-19 insurance identifying the person most likely to become a customer.

Keywords: coronavirus disease cosine similarity deep learning insurance health natural language processing social media

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จลุล่วงไปได้ด้วยดีด้วยความอนุเคราะห์จาก ผศ.ดร.บุษยมาส พิมพ์พรรณชาติ อาจารย์ที่ปรึกษาวิทยานิพนธ์ รศ.ดร.ไกรศักดิ์ เกษร และ ผศ.ดร.วรางคณา กิ้มปาน กรรมการสอบวิทยานิพนธ์ที่ได้กรุณาให้ความช่วยเหลือแนะนำช่วยตรวจทานแก้ไข้อผิดพลาดต่างๆ อีกทั้งขอขอบคุณ ดร.โกเมษ จันทวิมล ที่ช่วยให้คำแนะนำต่างๆจนวิทยานิพนธ์เล่มนี้สำเร็จลุล่วงสมบูรณ์ สุดท้ายนี้คุณค่าและประโยชน์อันพึงมีจากวิทยานิพนธ์เล่มนี้ ข้าพเจ้าขอมอบให้แก่ผู้มีพระคุณทุกท่าน

นายคงภพ ไชยศร

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญรูป

รูปที่ **หน้า**

สารบัญตาราง

ตารางที่ **หน้า**

บทที่ 1
บทนำ

งานวิจัยนี้มีที่มาที่แสดงถึง ความสำคัญของการกำหนดปัญหาที่ใช้ในการศึกษาและการกำหนดวัตถุประสงค์ในการวิจัยดังนี้

1.1 ที่มาและความสำคัญ

ในปัจจุบันทั่วโลกกำลังเจอวิกฤตโรคระบาดที่เกิดขึ้นอย่างหนัก นั่นคือโรคติดเชื้อไวรัสโคโรนา 2019 โรคระบาดนี้ส่งผลกระทบในวงกว้างกับภาคธุรกิจ และมีธุรกิจที่ได้รับผลกระทบเป็นจำนวนมาก แต่ในทางกลับกันก็ยังมีบางธุรกิจที่สามารถปรับตัวได้และมีธุรกิจใหม่ที่ประสบความสำเร็จอย่างมาก จากการวิเคราะห์ความต้องการของลูกค้าในปัจจุบัน เนื่องจากธุรกิจนั้นสามารถปรับตัวเพื่อสร้างผลิตภัณฑ์ใหม่ ๆ ที่ตอบสนองความต้องการของลูกค้าได้

เนื่องจากสถานการณ์ปัจจุบัน ทำให้ช่องทางออนไลน์โดยเฉพาะสื่อสังคมออนไลน์ (Social Media) ซึ่งได้รับความนิยมเป็นอย่างมาก โดยเฉพาะความนิยมในการแสดงความคิดเห็นในสถานการณ์ของโรคติดเชื้อไวรัสโคโรนา 2019 (COVID-19) ที่เกิดขึ้น จากที่กล่าวไปข้างต้น ธุรกิจต้องสามารถปรับตัว เพื่อให้เข้ากับสถานการณ์ในปัจจุบันและหาประโยชน์จากข้อมูลความคิดเห็นในสื่อสังคมออนไลน์เหล่านี้ เพื่อหา ผลิตภัณฑ์ใหม่ ๆ ที่ตอบสนองความต้องการของลูกค้า

สำหรับธุรกิจที่ต้องปรับตัวในสถานการณ์ปัจจุบันนั่นก็คือ ธุรกิจประกันภัย ดังนั้นวัตถุประสงค์ของงานวิจัยนี้ จะเน้นไปที่ความคิดเห็นที่เกิดจากธุรกิจประกัน โดยทางคณะผู้จัดทำจะนำเสนอผลิตภัณฑ์ใหม่ ๆ หรือปรับปรุงผลิตภัณฑ์เดิม เช่น นำเสนอขายประกันสุขภาพตัวอื่นจากกลุ่มคนที่คาดว่าจะมีความสนใจ เพื่อให้เกิดประโยชน์ทางธุรกิจประกันมากที่สุด

ทางผู้วิจัยจึงได้เห็นว่า ข้อมูลในสื่อสังคมออนไลน์เหล่านี้เป็นข้อมูลที่สามารถนำมาใช้ประโยชน์ได้กับทางธุรกิจ ซึ่งในสื่อสังคมออนไลน์มีช่องทางในการแสดงความคิดเห็นจำนวนมาก ดังนั้นทางผู้วิจัยจึงเลือกช่องทางนี่เป็นที่นิยมมากในประเทศไทยนั่นคือ กระทู้ออนไลน์พันทิป (Pantip.com) เนื่องจากสามารถนำความคิดเห็นมาวิเคราะห์ได้ และยังสามารถระบุไปยังตัวตนของลูกค้า เพื่อเพิ่มช่องทางการขายให้กับบริษัทได้ โดยเป็นข้อมูลที่สามารถเปิดเผยได้ในสาธารณะ

จากข้อมูล ความคิดเห็นในสังคมออนไลน์ที่มีจำนวนมากในปัจจุบัน ได้มีการประยุกต์ใช้การเรียนรู้ของเครื่อง (Machine Learning) มาช่วยในการวิเคราะห์ข้อมูลกันอย่างแพร่หลาย โดยเฉพาะ

การเรียนรู้เชิงลึก (Deep Learning) เป็นต้น ซึ่งแบบจำลองสำหรับการเรียนรู้ของเครื่องข้อมูลสามารถแบ่งออกได้เป็น 2 ส่วนหลัก ๆ คือ การเรียนรู้แบบมีผู้สอน (Supervised Learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ซึ่งข้อมูลที่อยู่ในสังคมออนไลน์นั้น จัดเป็นการเรียนรู้แบบไม่มีผู้สอน เนื่องจากไม่มีคำตอบของสิ่งที่ผู้วิจัยสนใจ ดังนั้น การเรียนรู้ของเครื่องที่ใช้สำหรับข้อมูลที่เป็นตัวอักษร (Textual Data) จะมีการนำองค์ความรู้ด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) มาประยุกต์ใช้ทำให้แบบจำลองการเรียนรู้ของเครื่องสามารถวิเคราะห์ข้อมูลที่เป็นข้อความได้โดยจะถูกเรียกว่าเหมืองข้อความ (Text Mining) ซึ่งตัวอย่างการประยุกต์ใช้งานสำหรับการทำเหมืองข้อความ ได้แก่ การวิเคราะห์ความรู้สึก (Sentiment Analysis)

จากเหตุผลที่กล่าวไว้ข้างต้นในงานวิจัยนี้จึงขอนำเสนอ การวิเคราะห์ข้อมูลจากสื่อสังคมออนไลน์ โดยทางผู้วิจัยจะนำข้อมูลของผู้ที่แสดงความคิดเห็นเกี่ยวกับการประกัน จากกระทู้ออนไลน์พันทิป ซึ่งเป็นกระดานสนทนาออนไลน์ที่ได้รับความนิยมเป็นอย่างมากในประเทศไทย โดยทางผู้วิจัยได้รวบรวมคำสำคัญ (Keyword) เพื่อสร้างแบบจำลองจากหัวข้อในกระดานสนทนาที่เกี่ยวข้องกับประเภทรีวิว (Review) ใน “คลับสุขภาพ” บนกระทู้ออนไลน์พันทิป และใช้มูลที่เกี่ยวกับการประกันและโรคติดเชื้อไวรัสโคโรนา 2019 ในกระทู้ออนไลน์พันทิปเป็นข้อมูลที่ใช้ในการวัดผลประสิทธิภาพของแบบจำลอง หลังจากนั้นจะใช้การประมวลผลภาษาธรรมชาติ โดยในงานวิจัยนี้จะแบ่งวิธีการเป็น 2 ส่วน ในส่วนที่ 1 จะเป็นการจำแนก (Classification) กลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ เนื่องจากลูกค้าที่อยู่ในกลุ่มนี้มีโอกาสที่จะซื้อประกันที่เกี่ยวกับสุขภาพในแบบอื่น แต่เนื่องจากข้อมูลที่อยู่ในกระทู้ออนไลน์พันทิปยังไม่ได้ถูกนำมาวิเคราะห์เพื่อตอบโจทย์ธุรกิจ ดังนั้นผู้วิจัยจึงใช้เทคนิคการแปลงคำเป็นตัวเลข (Word Embedding) โดยแบบจำลองเวิร์ดทูเวค (Word2Vec) ซึ่งจะทำให้ได้เวกเตอร์ของทุกคนที่อยู่ในความคิดเห็น โดยในทุกคำสำคัญจะถูกเลเบล (Label) ที่บ่งบอกว่าเป็นคำสำคัญในประเภทใด ซึ่งในกระบวนการนี้จะถูกตรวจสอบโดยผู้เชี่ยวชาญ (Expert) แต่เนื่องจากจำนวนคำในแต่ละความคิดเห็นมีจำนวนคำที่ไม่เท่ากันทำให้ยากต่อการจำแนกความคิดเห็น ทางผู้วิจัยจึงนำเสนอเทคนิคการแปลงประโยคเป็นตัวเลข (Sentence Embedding) เพื่อให้สามารถนำไปเปรียบเทียบความคล้ายคลึงกับในแต่ละคำที่ถูกจำแนกโดยผู้เชี่ยวชาญ โดยใช้เทคนิคความคล้ายคลึงของโคไซน์ (Cosine Similarity) มาใช้วิเคราะห์ความคิดเห็นของผู้ที่มีแนวโน้มที่รักสุขภาพ เพื่อมาใช้ในการจำแนกความคิดเห็นว่าเป็นกลุ่มคนที่มีแนวโน้มรักสุขภาพมาก (Strong) หรือกลุ่มคนที่มีแนวโน้มรักสุขภาพน้อย (Weak) ส่วนที่ 2 จะเป็นการจำแนกทัศนคติของผู้ที่แสดงความคิดเห็น เนื่องจากข้อมูลที่อยู่ในรูปแบบความคิดเห็นที่เป็นภาษาไทยมีความซับซ้อนมากกว่ามาภาษาอังกฤษ เลยทำให้ข้อมูล

ในส่วนนี้ยังไม่ค่อยถูกนำมาใช้จำแนกความคิดเห็น เพื่อปรับปรุงและพัฒนาในธุรกิจ ดังนั้นผู้วิจัยจึงใช้เทคนิคการวิเคราะห์ความรู้สึกมาใช้ในการจำแนกความคิดเห็นว่าเป็นเชิงบวก (Positive) หรือเชิงลบ (Negative) โดยกลุ่มเป้าหมายหลักของงานวิจัยนี้คือ กลุ่มของคนที่มีแนวโน้มรักสุขภาพมากและมีทัศนคติเป็นเชิงบวก โดยหลังจากได้กลุ่มเป้าหมายของงานวิจัยแล้ว ทางผู้วิจัยจะนำข้อมูลที่เป็นรหัสของลูกค้า (User ID) ที่อยู่ในกระทู้ออนไลน์พันทิปของกลุ่มเป้าหมาย เพื่อที่ธุรกิจจะนำข้อมูลในส่วนนี้มาวิเคราะห์เพื่อหาช่องทางในการขายในอนาคตต่อไป

1.2 วัตถุประสงค์ของงานวิจัย

1) พัฒนาแบบจำลองที่สามารถจำแนกกลุ่มลูกค้าที่มีแนวโน้มที่รักสุขภาพ เพื่อที่บริษัทจะแนะนำผลิตภัณฑ์ที่เกี่ยวกับประกันสุขภาพตัวอื่นๆ
2) พัฒนาแบบจำลองที่สามารถจำแนกหากลุ่มลูกค้าที่มีทัศนคติที่เป็นเชิงบวกต่อบริษัทประกัน เพื่อนำไปพัฒนาผลิตภัณฑ์ให้ดีขึ้น
3) พัฒนาเทคนิคเพื่อเพิ่มประสิทธิภาพ โดยใช้เทคนิคการแปลงประโยคเป็นตัวเลขร่วมกับเทคนิคความคล้ายคลึงของโคไซน์

1.3 ขอบเขตของงานวิจัย

1) ข้อมูลที่ใช้เป็นข้อมูลความคิดเห็นและประเภทรีวิว ในกระทู้ออนไลน์พันทิปโดยเก็บข้อมูลย้อนหลัง จนถึง วันที่ 31 ธันวาคม 2564 โดยใช้ข้อมูลกระทู้สนทนาทั้งหมด 180 กระทู้ จำนวน 1490 ความคิดเห็น และ ความคิดเห็นประเภทรีวิวทั้งหมด 283 กระทู้
2) ข้อมูลที่ใช้ทำการทดสอบจะเป็นข้อมูลที่เกี่ยวข้องกับธุรกิจประกันและโรคติดเชื้อไวรัสโคโรนา 2019
3) เป็นข้อมูลที่สามารถระบุไปถึงรหัสของลูกค้าที่แสดงความคิดเห็นได้
4) ใช้จัดการและประมวลผลกับข้อความที่เป็นภาษาไทย

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ทำให้ทราบกลุ่มลูกค้าที่มีแนวโน้มที่รักสุขภาพและมีโอกาสที่ทางบริษัทจะแนะนำสินค้าที่เกี่ยวกับประกันสุขภาพตัวอื่น ๆ ได้ เพื่อนำมาใช้สร้างนโยบายและแผนการตลาดที่เหมาะสม
2) ทำให้บริษัทสามารถปรับปรุงและพัฒนาสินค้าให้ตอบโจทย์ธุรกิจมากขึ้นจากการวิเคราะห์ ความรู้สึกของผู้ที่แสดงความคิดเห็น

3) เป็นแนวทางในการประยุกต์ใช้การนำกระทู้ออนไลน์พันทิปมาใช้ในการส่งเสริมและพัฒนา ช่องทางใหม่ ๆ สำหรับการขายสินค้า
4) เป็นแนวทางในการแนะนำช่องทางการขายแบบใหม่ให้กับบริษัท

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาครั้งนี้ ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดของเนื้อหาประกอบ ดังต่อไปนี้

2.1 การเตรียมข้อมูล

2.2 ทฤษฎีที่เกี่ยวข้อง

2.3 การเปรียบเทียบประสิทธิภาพการทำนาย

2.4 เครื่องมือและภาษาที่ใช้ในการพัฒนา

2.5 งานวิจัยที่เกี่ยวข้อง

2.1 การเตรียมข้อมูล

การเตรียมข้อมูลในงานวิจัยฉบับนี้จะเป็นการเตรียมข้อมูลสำหรับข้อความโดยอ้างอิงจาก[1] ซึ่งขั้นตอนที่ใช้ในการเตรียมข้อมูลมีดังนี้

2.1.1 การตัดคำ (Tokenization)

การนำประโยคกระทู้ออนไลน์พันทิปมาแบ่งออกเป็นคำต่าง (Token) ตามพจนานุกรม (Lexicon) โดยใช้ไลบรารี่ PythaiNLP และใช้โมดูล word_tokenize ยกตัวอย่างประโยค “ฉันรักสุขภาพ!” จะถูกแปลงเป็นคำศัพท์ “ฉัน”, “รัก”, “สุขภาพ”, “!”

2.1.2 การตัดคำที่ไม่มีนัยสำคัญ (Stop Words Removal)

การตัดคำที่ไม่มีนัยสำคัญ (stop words) ได้แก่ คำบุพบท คำสันธาน คำสรรพนาม ลักษณะนาม ตัวเลข รวมถึงคำลงท้ายประโยคในภาษาไทย โดยใช้ไลบรารี่ PythaiNLP และใช้โมดูล thai_stopwords ซึ่งมีคำที่ไม่มีนัยสำคัญทั้งหมด 1030 คำ เช่น คำว่า “ฉัน” , “เธอ” , “และ” , “โดย” ยกตัวอย่างประโยค “ฉันรักสุขภาพ!” จะถูกแปลงเป็นคำศัพท์ “รัก”, “สุขภาพ”, “!”

2.1.3 การลบเครื่องหมายวรรคตอน (Punctuation Removal)

การลบเครื่องหมายต่าง ๆ ที่ไม่จำเป็นในการวิเคราะห์ เช่น เครื่องหมายลูกน้ำ (,) เครื่องหมายอัศเจรีย์ (!) โดยใช้ไลบรารี่ PythaiNLP และใช้โมดูล countthai ยกตัวอย่างประโยค “ฉันรักสุขภาพ!” จะถูกแปลงเป็นคำศัพท์ “ฉัน”, “รัก”, “สุขภาพ”

2.1.4 การลบรูปสัญลักษณ์ชนิดข้อความ (Emoji Removal)

การลบรูปสัญลักษณ์ชนิดข้อความที่ไม่จำเป็นในการวิเคราะห์ เช่น การยิ้ม (😀) โดยใช้ไลบรารี่ PythaiNLP ที่ใช้สำหรับตัดคำภาษาไทย pythainlp.util.emoji_to_thai ยกตัวอย่างประโยค “ฉันรักสุขภาพ! 😀” จะถูกแปลงเป็นคำศัพท์ “ฉัน”, “รัก”, “สุขภาพ”, “!”

2.1.5 การตัดส่วนท้ายของคำ (Word Stemming)

กระบวนการการตัดส่วนท้ายของคำ เพื่อให้สามารถลดความซับซ้อนที่เกิดจากคำพูดโดยที่ไม่รู้บริบทและไม่ตรงกับไวยกรณ์ หรือคำที่เขียนผิด โดยใช้ไลบรารี่ nltk โดยใช้โมดูล stem ยกตัวอย่างคำว่า “กระเจี้ยบ” จะถูกแปลงเป็น “กระเจี๊ยบ”

2.1.6 การบ่งบอกประเภทของความรู้สึก (Class Labeling)

การบ่งบอกความรู้สึกในข้อมูลความคิดเห็นจะถูกกำหนดโดยระดับคะแนนที่ผู้เขียนกระทู้ออนไลน์พันทิปให้ไว้ในข้อมูลที่เก็บมาจากกระทู้ออนไลน์พันทิป การให้คะแนนจะเป็นระบบคะแนนแบบห้าดาว (5 Star Scoring System) โดยระบบการให้คะแนนแบบนี้จะเรียงลำดับความพึงพอใจของลูกค้าได้ โดยเมื่อคะแนนเท่ากับ 0 คือมีความพึงพอใจต่ำสุด และคะแนนเท่ากับ 5 คือมีความพึงพอใจสูงที่สุด ซึ่งในงานวิจัยของ [2] ได้อธิบายไว้ว่าการจำแนกความรู้สึกสามารถแบ่งได้เป็นความรู้สึกเชิงบวก (4-5 คะแนน) และความรู้สึกเชิงลบลบ (1-3 คะแนน) ซึ่งผู้วิจัยสามารถนำประเภทความรู้สึกนี้เป็นชุดข้อมูลฝึก

2.2 ทฤษฎีที่เกี่ยวข้อง

2.2.1 การประมวลผลภาษาทางธรรมชาติ (Natural Language Processing : NLP)

การประมวลผมภาษาทางธรรมชาติเป็นหนึ่งใน การศึกษางานทางด้านปัญญาประดิษฐ์ (Artificial Intelligence : AI) โดยศาสตร์ทางนี้ มีจุดมุ่งหมายเพื่อให้คอมพิวเตอร์สามรถที่จะเข้าใจและประมวลผลภาษาทางธรรมชาติขอมนุษย์ได้ โดยประโยชน์ที่ได้จากการทำให้คอมพิวเตอร์สามารถเข้าใจภาษามนุษย์ได้ จะช่วยอำนวยความสะดวกเป็นอย่างมากในการทำงานมากมายที่มีความเกี่ยวข้องกับทางด้านภาษาของมนุษย์ [3] โดยการประมวลผลภาษาทางธรรมชาติ จากการนำความรู้ในด้านการประมวลผลภาษาทางธรรมชาติ ได้ถูกนำมาใช้ในศาสตร์หลากหลายแขนง เช่น ศาสตร์ทางด้านจิตวิทยา โดยในการศึกษานี้ ผู้วิจัยได้เลือกใช้กระบวนการประมวลผลภาษาธรรมชาติ แบบการจัดจำแนกข้อความเป็นหมวดหมู่ สำหรับในการประมวลผลตามหลักการวิเคราะห์ความรู้สึก

2.2.2 การทำเหมืองข้อความ (Text Mining)

การทำเหมืองข้อความคือ กระบวนการที่นำข้อมูลในลักษณะของข้อความมาใช้ในการวิเคราะห์ โดยจะเป็นการประยุกต์จากศาสตร์ความรู้ทางด้านสถิติและศาสตร์ความรู้ทางด้านคอมพิวเตอร์มาใช้ เพื่อค้นหาประเด็นหรือสาระสำคัญที่อยู่ในข้อมูล ซึ่งการศึกษาการทำเหมืองข้อความนี้จะมุ่งเน้นเพื่อให้คอมพิวเตอร์สามารถทำความเข้าใจข้อมูลได้เช่นเดียวกับมนุษย์ และเพื่อให้คอมพิวเตอร์สามารถที่จะวิเคราะห์ข้อมูลได้อย่างรวดเร็วในสิ่งที่มนุษย์ไม่สามารถที่จะวิเคราะห์ได้ในระยะเวลาอันสั้น เพราะในปัจจุบันปริมาณของข้อมูลนั้นเกิดขึ้นอย่างไม่จำกัด และเพิ่มขึ้นอย่างรวดเร็วด้วยเหตุนี้ ทำให้การวิเคราะห์ด้วยคอมพิวเตอร์จึงเป็นสิ่งจำเป็นต่อการวิเคราะห์ ซึ่งให้ประสิทธิภาพที่รวดเร็วและสามารถทำงานได้อย่างไร้ขีดจำกัด แต่เนื่องจากในปัจจุบัน ลักษณะของข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) เป็นส่วนใหญ่ โดยก่อนที่ จะนำข้อมูลมาให้คอมพิวเตอร์วิเคราะห์นั้น จำเป็นที่จะต้องจัดการข้อมูลให้อยู่ในรูปที่คอมพิวเตอร์สามารถเข้าใจได้ก่อน จึงจะสามารถนำไปใช้ประมวลผลต่อได้ [4]

2.2.3 เทคนิคการแปลงคำเป็นตัวเลข (Word Embedding)

เป็นกระบวนการในการแปลงคำเป็นตัวเลข ถือเป็นหนึ่งในวิธีในการสร้างตัวแปรที่จะนำไปใช้วิเคราะห์จากข้อมูลคำโดยจะทำการลดขนาดปริภูมิเวกเตอร์ (Vector space) โดยแบบจำลองที่นิยมใช้คือ เวิร์ดทูเวค (Word2Vec) ที่เป็นการเปลี่ยนคำให้เป็นเวกเตอร์ ซึ่งแบบจำลองที่ได้รับความนิยมคือแบบจำลองซีโบล (Continuous Bag-of-Words : CBOW) ที่เป็นการรวมกันของการเป็นตัวแทนของคำบริบทหรือคำโดยรอบเพื่อทำนายคำเป้าหมายในระดับกลาง และ แบบจำลองสคริปแกรม (Skip-gram) ที่จะตรงข้ามกับแบบจำลองซีโบล เนื่องจากรูปแบบการป้อนข้อมูลการกระจายของคำเป้าหมายที่ใช้ในการทำนายบริบท

จากงานวิจัยใน [5] การเปลี่ยนคำเป็นปริภูมิเวกเตอร์ โดยแบบจำลองสคริปแกรม ได้ผลดีกว่าแบบจำลองซีโบล แต่จะใช้เวลาฝึกแบบจำลองมากกว่า ทางผู้วิจัยจึงเลือกใช้แบบจำลองสคริปแกรม โดยมีรายละเอียดดังนี้ แบบจำลองสคริปแกรม คือการทำนายคำที่อยู่รอบข้าง (Context Words) โดยมีสูตรดังสมการที่ (2.1)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\log p(w_{t+1}|w_t) \tag{2.1}$$

โดยสูตร แบบจำลองสคริปแกรมพื้นฐาน จะเป็นการกำหนด $\mathrm{p}(\mathrm{w_{t+j}}|\mathrm{w_t})$ โดยใช้ฟังก์ชันที่ใช้ในการแปลงค่าน้ำหนักจากค่าหนึ่งไปเป็นอีกค่าหนึ่งด้วยฟังก์ชันที่ผู้วิจัยกำหนด (SoftMax) ดังสมการที่ (2.2)[6] ,[7]

$$\mathrm{p}(\mathrm{w_0}|\mathrm{w_i}) = \frac{\exp(\mathrm{v'w_0^T vw_i})}{\sum_{\mathrm{w=1}}^{\mathrm{W}} \exp(\mathrm{v'w^T vw_i})} \tag{2.2}$$

เมื่อ $\mathrm{w_1, w_2, w_3, \dots, w_t}$ คือ ลำดับของคำที่ใช้พัฒนาแบบจำลอง

c คือ ขนาดของบริบทการฝึกอบรม

vw คือ Input

$v'w$ คือ output

w คือ เวกเตอร์

W คือ จำนวนคำในคำศัพท์

2.2.4 เทคนิคการแปลงประโยคเป็นตัวเลข (Sentence Embedding)

เป็นการแปลงเวกเตอร์ของคำให้เป็นเวกเตอร์ของประโยค เนื่องจากมีเวกเตอร์ของแต่ละคำจึงแปลงเป็นเวกเตอร์ประโยคด้วยการแบ่งแต่ละประโยคโดยใช้ช่องว่าง จากนั้นจะเป็นการขยายประโยคโดยหาค่าเฉลี่ยของคำทั้งหมดเป็นส่วนหนึ่งของประโยคโดยผลลัพธ์ของการแสดงเอกสารจึงเป็นเวกเตอร์ประโยคที่มีเมทริกซ์ (Matrix) เดียวกันคือ $\mathrm{N} * \mathrm{K}$ ดังสมการที่ (2.3) [8]

$$\mathrm{Sen2Vec(S_j)} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{v_{wi}}}{\mathrm{n}} \tag{2.3}$$

เมื่อ $\mathrm{v_{wi}}$ คือ เวกเตอร์ของคำ

n คือ จำนวนของคำในประโยค

N คือ จำนวนประโยค

K คือ มิติ

2.2.5 เทคนิคความคล้ายคลึงของโคไซน์ (Cosine Similarity)

เป็นกระบวนการสำหรับวัดความเหมือนของเวกเตอร์ A กับเวกเตอร์ B ว่าไปในทิศทางเดียวกันหรือไม่ โดยหากค่าที่ได้ยิ่งน้อยจะทำให้เวกเตอร์ทั้งสองยิ่งมีความเหมือนกันมาก โดยจะอยู่ในช่วง [0,1] ดังสมการที่ (2.4) [9]

$$\cos(\mathrm{A},\mathrm{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\mathbf{AB}} = \frac{\sum_{i=1}^{n} \mathrm{A_i B_i}}{\sqrt{\sum_{i=1}^{n} (\mathrm{A_i})^2} \cdot \sqrt{\sum_{i=1}^{n} (\mathrm{B_i})^2}} \qquad (2.4)$$

เมื่อ i คือ {1, … , n} โดยที่ n คือ จำนวนทั้งหมด

$\mathrm{A_i}$ คือ เวกเตอร์ A

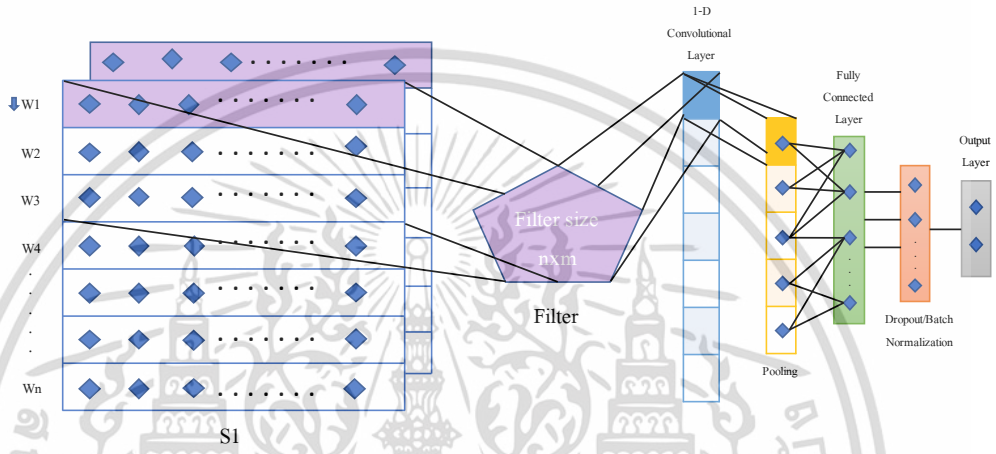
$\mathrm{B_i}$ คือ เวกเตอร์ B

2.2.2 การวิเคราะห์ความรู้สึก (Sentiment Analysis)

หนึ่งในการประมวลผลภาษาทางธรรมชาติที่ได้ถูกนำมาใช้ในการช่วยเหลือและอำนวยความสะดวกในการทำงานของมนุษย์ คือ การวิเคราะห์ความรู้สึก โดยหลักการคือ การวิเคราะห์ข้อความเพื่อค้นหาประเด็นหรือข้อมูลเชิงลึกที่ซ่อนอยู่ เช่น ความรู้สึกนึกคิดของผู้เขียนข้อความ หรือวัตถุประสงค์ที่ผู้เขียนข้อความต้องการจะสื่อสาร เป็นต้น ซึ่งการวิเคราะห์อารมณ์ความรู้สึกนั้น ส่วนใหญ่จะมีการวิเคราะห์ในลักษณะของ ความรู้สึกเชิงบวก ความรู้สึกเชิงลบ และความรู้สึกปกติ (Neutral) เป็นต้น โดยหลักการวิเคราะห์นั้นจะใช้วิธีการทางด้าน การทำเหมืองข้อความ ซึ่งเป็นส่วนหนึ่งของการประมวลผลภาษาทางธรรมชาติ ซึ่งการศึกษากระบวนการวิเคราะห์นี้จะมุ่งเน้นไปทางด้านการจัดจำแนกประเภทของข้อความ (Text Classification) สำหรับในการทำนายความรู้สึกในแต่ละข้อความ โดยการวิเคราะห์ความรู้สึก ซึ่งวิเคราะห์โดยการทำการจัดจำแนกข้อความนั้นมีการถูกนำไปใช้งานอย่างแพร่หลาย โดยเฉพาะในการดำเนินงานทางธุรกิจ อาทิเช่น การวิเคราะห์ความรู้สึกจากการแนะนำร้านอาหารจากช่องทางออนไลน์ สำหรับวิเคราะห์ความคิดเห็นที่มีต่อร้านอาหารว่ามีลักษณะข้อความในทิศทางใด, การวิเคราะห์ข้อความจากข่าวการแพร่ระบาดของโรคเพื่อตรวจสอบข้อเท็จจริงจากข้อความในช่องทางออนไลน์ เป็นต้น [10]

2.2.6 โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network: CNN)

เป็นเทคนิคการเรียนรู้เชิงลึก ซึ่งมีการประยุกต์ใช้หลักการวิเคราะห์ข้อมูลในรูปแบบลำดับเหตุการณ์ที่ชัดเจน โดยข้อดีของข้อมูลในรูปแบบนี้สามารถเปลี่ยนบริบทของเหตุการณ์ตามลำดับได้ในทางกลับกันงานด้านอนุกรมเวลาสามารถนำโครงข่ายประสาทแบบคอนโวลูชันมาใช้ในการสร้างแบบจำลองโดยใช้คอนโวลูชัน 1 มิติ [11]



รูปที่ 2.1 อัลกอริทึมของโครงข่ายประสาทแบบคอนโวลูชัน

จากรูปที่ 2.1 อัลกอริทึมของโครงข่ายประสาทแบบคอนโวลูชันจะมีการแบ่งการประมวลผลออกเป็น 3 ส่วน ดังนี้

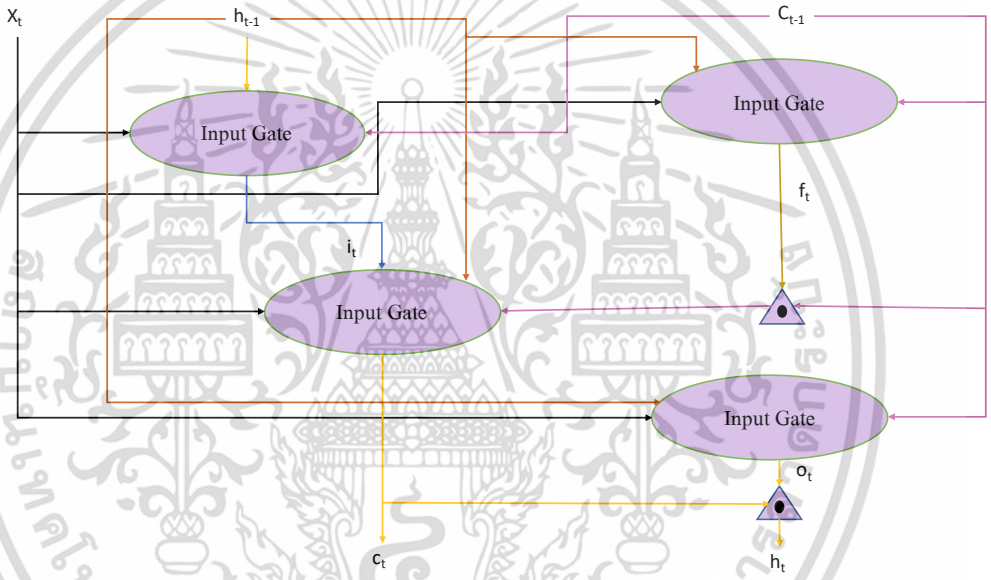
1. การสกัดคุณลักษณะ (Feature Extractions) เป็นการทำงานเพื่อคัดเลือกคุณลักษณะสำหรับการนำไปใช้ในการจัดจำแนก โดยเป็นจุดเด่นของโครงข่ายประสาทแบบคอนโวลูชัน ซึ่งจะมีการใช้เครื่องมือที่เรียกว่าตัวกรอง (Fiter) สำหรับในการคัดเลือกคุณลักษณะดังกล่าวโดยตัวกรองจะเป็นเมทริกซ์ ซึ่งจะต้องกำหนดขนาดของ ตัวกรองที่จะใช้สำหรับในการคัดเลือกข้อมูลก่อน และจากนั้นจะใช้ตัวกรองในการวางบนชุดข้อมูลเพื่อกำหนดขอบเขตของคุณลักษณะที่จะนำไปใช้ในการทำนายในแบบจำลองต่อไป

2. การทำพูลลิ่ง (Pooling) เป็นส่วนหนึ่งของขั้นตอนการจัดเตรียมข้อมูลเพื่อนำเข้าสู่เลเยอร์ชั้นเชื่อมโยงแบบสมบูรณ์ (Fully Connected Layer) โดยจะเป็นการลดขนาดของข้อมูลโดยพิจารณาเงื่อนไขจากค่าที่ได้ในการสกัดคุณลักษณะ ซึ่งจะประกอบไปด้วยการเลือกค่าสูงสุดของพูลลิ่ง (Max Pooling) และค่าเฉลี่ยของพูลลิ่ง (Average Pooling) สำหรับเป็นค่าที่ใช้ในการเลือกค่าการทำ

3. ชั้นเชื่อมโยงแบบสมบูรณ์ จะเป็นการคำนวณค่าจากการทำการสกัดคุณลักษณะ และการทำพูลลิ่งเพื่อหาค่าความน่าจะเป็นในการจัดจำแนกข้อความโดยโครงสร้างของโครงข่ายประสาทแบบคอนโวลูชันได้ก่อนนำเข้าสู่เลเยอร์ชั้นถัดไป

2.2.7 โครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว (Long Short Term Memory : LSTM)

แบบจำลองนี้ถูกออกแบบมาให้มีความสามารถในการอ่านข้อมูลที่เป็นอนุกรมซึ่งมีความเหมาะสมกับการนำมาใช้ในงานด้านการประมวลผลภาษาธรรมชาติและอนุกรมเวลา [12]



รูปที่ 2.2 อัลกอริทึมของโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว

โดยโครงสร้างของโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว จะประกอบไปด้วย 4 ส่วน ได้แก่

1. ฟอร์เก็ทเกท (Forget Gate) เป็นประตูสำหรับการกำหนดข้อมูลเพื่อพิจารณาว่าข้อมูลที่ถูกส่งต่อมาจากโหนด (Node) ก่อนหน้า ต้องจัดเก็บข้อมูลหรือไม่ โดยผลลัพธ์ของการพิจารณาจะมี 2 ค่า [0, 1] โดย 0 เป็นการไม่จัดเก็บข้อมูลจากโหนดก่อนหน้า และ 1 เป็นการจัดเก็บข้อมูลจากโหนดก่อนหน้า ดังสมการที่ (2.5)

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{2.5}$$

เมื่อ f_t คือ ผลลัพธ์ที่ได้จากฟอร์เก็ทเกทที่เวลา t ใดๆ

σ คือ ซิกมอยด์ฟังก์ชั่น (Sigmoid Function)

W_f คือ ค่าน้ำหนัก

h_{t-1} คือ ข้อมูลขาออกในช่วงเวลา t-1 หรือช่วงเวลาก่อนหน้า

x_t คือ ข้อมูลขาเข้าในช่วงเวลา t หรือเวลาใดๆ

b_f คือ ค่าความเอนเอียง (Bias) จากฟอร์เก็ทเกท

2. อินพุทเกท (Input Gate) เป็นประตูสำหรับการรับข้อมูลใหม่ โดยที่จะมีการคำนวณ 2 ส่วน คือ การหาค่าความสำคัญของข้อมูล ดังสมการที่ (2.6) และการหาค่าสถานะของโหนดปัจจุบัน ดังสมการที่ (2.7)

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2.6)$$

$$c_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (2.7)$$

เมื่อ i_t คือ ผลลัพธ์ที่ได้จากอินพุทเกทที่เวลา t ใดๆ

σ คือ ซิกมอยด์ฟังก์ชั่น

W_i คือ ค่าน้ำหนัก

b_i คือ ค่า ค่าความเอนเอียง จาก ประตูเข้า

c_t คือ ผลลัพธ์ที่ได้จากอัพเดทเซลล์สเตรช (Cell State) ที่เวลา t ใดๆ

W_c คือ ค่าน้ำหนัก

b_c คือ ค่า ค่าความเอนเอียงจากอัพเดทเซลล์สเตรช

3. อัพเดทเกท (Update Gate) หลังจากที่คำนวณผลลัพธ์ของฟอร์เก็ทเกทและอินพุทเกทจะนำผลลัพธ์ของสถานะก่อนหน้า รวมกับสถานะปัจจุบัน เพื่อหาค่าสถานะใหม่ของ โหนดปัจจุบัน ดังสมการที่ (2.8)

$$u_t = f_t \cdot c_{t-1} + i_t \cdot c_t \tag{2.8}$$

เมื่อ u_t คือ ผลลัพธ์ที่ได้จากอัพเดทเกท ที่เวลา t ใดๆ

c_{t-1} คือ ผลลัพธ์ที่ได้จากอัพเดทเซลล์สเตรชในช่วงเวลา t-1

4. เอาท์พุตเกท (Output Gate) เป็นประตูสำหรับการเตรียมส่งออกผลลัพธ์จากโหนดปัจจุบันไปยังโหนดถัดไป ดังสมการที่ (2.9) และสุดท้ายจะนำมาคำนวณร่วมกับสถานะของโหนดปัจจุบันที่ได้ทำการอัพเดทแล้ว ดังสมการที่ (2.10) เพื่อส่งต่อไปยังโหนดถัดไป

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{2.9}$$

$$h_t = o_t \tanh u_t \tag{2.10}$$

เมื่อ o_t คือ ผลลัพธ์ที่ได้จากชั้นฟอร์เก็ทเกทที่เวลา t ใดๆ

σ คือ ซิกมอยด์ ฟังก์ชั่น

W_o คือ ค่าน้ำหนัก

b_o คือ ค่า ค่าความเอนเอียงจากชั้นฟอร์เก็ทเกท

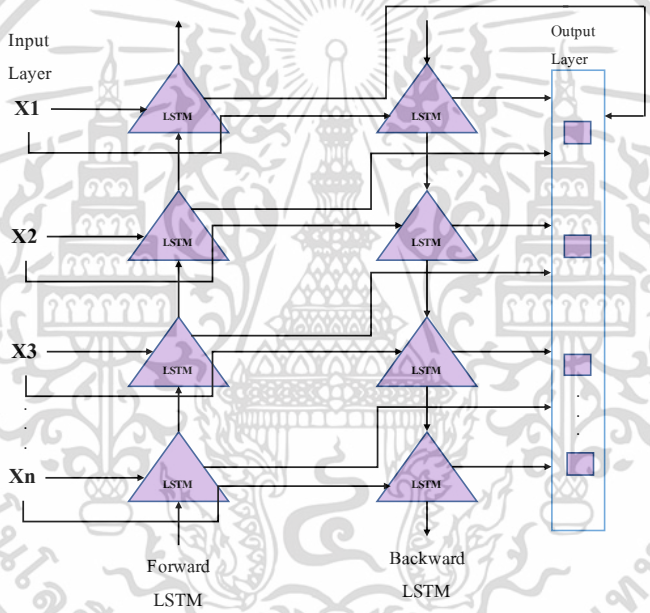
h_t คือ ผลลัพธ์ที่ได้จาก โครงข่ายประสาทเทียมความจำระยะสั้นแบบยาวของเซลล์ ที่เวลา t ใดๆ

ด้วยคุณสมบัติของแบบจำลองโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว ทำให้สามารถแก้ปัญหาในส่วนของแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ ที่ไม่สามารถประมวลผลข้อมูลที่มีความยาวได้ แต่เนื่องจากกระบวนการทำงานของแบบจำลองแบบโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว มีการประมวลผลข้อมูลในทิศทางเดียว เพื่อที่จะทำให้สามารประมวลผล

ข้อมูลได้ครอบคลุมทุกลำดับของข้อมูล จึงได้มีการนำแบบจำลองระยะสั้นแบบยาวสองทิศทาง (Bi-Long Short Term Memory)

2.2.8 โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง (Bi-Long Short Term Memory : Bi-LSTM)

เป็นแบบจำลองที่พัฒนามาจาก โครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว ซึ่งมีการประมวลผลข้อมูลแบบสองทิศทางเพื่อให้ครอบคลุมทั้งข้อมูลในอดีตและอนาคต โดยคุณสมบัติของแบบจำลอง โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง จะมีการประมวลผลข้อมูลจากการอ่านข้อมูลไปข้างหน้าไปข้างหลัง (Forward pass) และอ่านข้อมูลจากข้างหลังมาข้างหน้า (Reverse pass) [13]

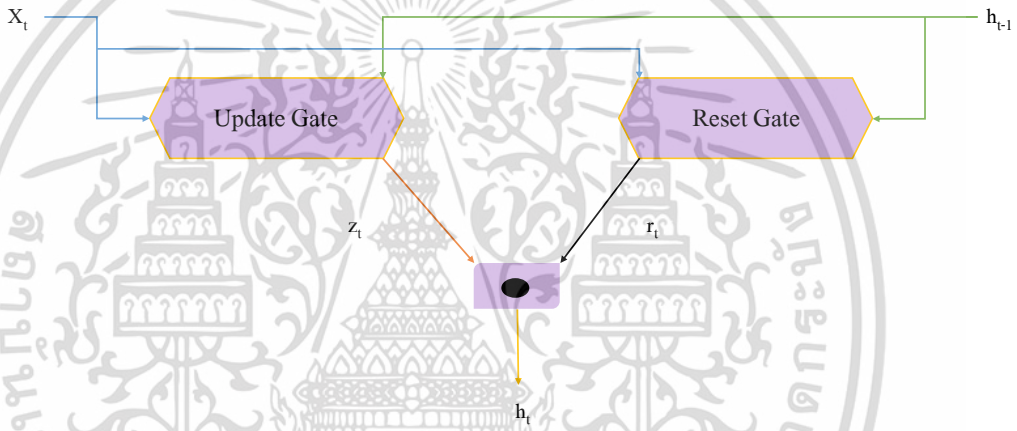


รูปที่ 2.3 อัลกอริทึมของโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาวสองทิศทาง

ซึ่งการอ่านข้อมูลในทั้ง 2 ทิศทางจะมีการทำงานที่พร้อมกันได้ โดยโครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง เป็นอีกรูปแบบหนึ่งของโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว ซึ่งมีหลักการทำงานเหมือนกับโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว เพียงแต่มีการเพิ่มหนึ่งฟังก์ชันการทำงาน คือการป้อนข้อมูลแบบย้อนกลับเข้าไปด้วย นั่นหมายความว่าขั้นตอนในการทดสอบแบบโครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง จะช้ากว่า โครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว

2.2.9 โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู (Gated Recurrent Unit: GRU)

เป็นกลไกปิดเปิดการอัพเดตสถานะภายในโครงข่ายประสาทเทียมแบบวนกลับ ที่คล้ายกับโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว แต่มีพารามิเตอร์ (Parameter) น้อยกว่า โครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว เนื่องจากไม่มีเอาท์พุตเกท โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู มีประสิทธิภาพใกล้เคียงกับโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาวในหลาย ๆ งาน เนื่องจากมีพารามิเตอร์น้อยกว่าทำให้สามารถเทรนได้ง่ายกว่าและเร็วกว่า แต่ในบางงานที่จำนวนของข้อมูลมีขนาดไม่ใหญ่มาก พบว่าโครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตูประสิทธิภาพดีกว่า [14]



รูปที่ 2.4 อัลกอริทึมของโครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู

จากรูปที่ 2.4 โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ได้ทำการลดความซับซ้อนในการทำงานของโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว โดยการลดหน่วยย่อยในเซลล์เหลือเพียง 2 ส่วน ได้แก่อัพเดทเกทและรีเซตเกท (Reset Gate)

1. อัพเดทเกท เป็นหน่วยย่อยที่ทำการนำข้อมูลไปคำนวณ เพื่อกำหนดสถานะขเซลล์ สำหรับใช้ในการคำนวณในขั้นถัดไป โดยทำการคำนวณในทุก ๆ รอบที่มีข้อมูลเข้ามา ดังสมการที่ (2.11)

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \tag{2.11}$$

เมื่อ z_t คือ ผลลัพธ์ที่ได้จากอัพเดทเกทที่เวลา t ใดๆ

$\sum$ คือ ซิกมอยด์ฟังก์ชั่น

W_z คือ ค่าน้ำหนัก

h_{t-1} คือ ข้อมูลขาออกในช่วงเวลา t-1

x_t คือ ข้อมูลขาเข้าในช่วงเวลา t

b_z คือ ค่าความเอนเอียงจากอัพเดทเกท

2. รีเซตเกทเป็นหน่วยย่อยที่ใช้ในการกำหนดข้อมูลว่าควรที่จะเก็บค่าสถานะที่ได้ จากการคำนวณในครั้งที่แล้วมากน้อยเพียงใด ดังสมการที่ (2.12)

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (2.12)$$

เมื่อ r_t คือ ผลลัพธ์ที่ได้จากฟอร์เก็ทเกทที่เวลา t ใดๆ

σ คือ ซิกมอยด์ ฟังก์ชั่น (Sigmoid Function)

W_r คือ ค่าน้ำหนัก

h_{t-1} คือ ข้อมูลขาออกในช่วงเวลา t-1

x_t คือ ข้อมูลขาเข้าในช่วงเวลา t

b_r คือ ค่าความเอนเอียง (Bias) จากฟอร์เก็ทเกท

สำหรับการคำนวณหาค่า ผลลัพธ์ (Output) และ ฮิดเด้นสเตรช(Hidden State) ของโครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตูนั้น หน่วยย่อยสำหรับใช้คำนวณที่ตายตัว โดยทำการนำค่าผลลัพธ์ที่ได้จากทั้ง 2 ประตู มาทำการคำนวณด้วยฟังก์ชัน Tanh จากนั้นค่าที่ได้จากรีเซตเกท จะกำหนดว่าจะทำการจำค่าเดิมหรือทำการล้างค่าเดิมออกไปและทำการควบคุมปริมาณของข้อมูลด้วยค่าจากอัพเดทเกท ดังสมการที่ (2.13)

$$h_h = (1 - z_t) \cdot \tanh(r_t \cdot W_h h_{t-1} + W_x x_t) + z_t h_{t-1} \quad (2.13)$$

เมื่อ h_n คือ ผลลัพธ์ที่ได้จากการคำนวน Hidden State

W_h คือ ค่าน้ำหนักสำหรับคำนวณ Hidden State จากหน่วยเวลาก่อนหน้า

h_{t-1} คือ ข้อมูลขาออกในช่วงเวลา t-1

x_t คือ ข้อมูลขาเข้าในช่วงเวลา t

W_x คือ ค่าน้ำหนักสำหรับคำนวณค่าขาเข้า

โครงข่ายประสาทหน่วยเวียนกลับแบบมีประตู แสดงให้เห็นว่าในส่วนของประสิทธิภาพการวิเคราะห์ข้อมูลเมื่อทำการเปรียบเทียบกับโครงข่ายประสาทเทียมความจำระยะสั้นแบบยาว แต่โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู นั้นสามารถแสดงจุดเด่นในด้านของการลดจำนวนพารามิเตอร์ในการฝึกสอนลง ทำให้แบบจำลองนั้น สามารถทำงานได้อย่างรวดเร็วมากยิ่งขึ้น

2.3 การเปรียบเทียบประสิทธิภาพการทำนาย

2.3.1 เมทริกซ์ความสับสน (Confusion Matrix)

เมทริกซ์ความสับสนเป็นตารางที่ใช้สำหรับวัดผลแบบจำลอง ซึ่งการเลือกผลทำนายจะทำได้โดยการวัดประสิทธิภาพของการทำนายด้วยตัวชี้วัดหลากหลายชนิดซึ่งสามารถคำนวณได้จากเมทริกซ์ประกอบด้วย 4 ชนิดได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F1-Score) และ ค่าความถูกต้อง (Accuracy) ซึ่งสามารถคำนวณได้ตามสมการที่ (2.14) ถึง (2.17) ดังตารางที่ 1 [15]

ตารางที่ 2.1 เมทริกซ์ความสับสน

ผลทำนาย / ค่าจริง	ผลบวก	ผลลบ
ผลบวก	TP	FN
ผลลบ	FP	TN

เมื่อ TP คือ ผลทำนายเป็นจริงและข้อมูลเป็นจริง หรือเรียกว่า True Positive

FP คือ ผลทำนายเป็นจริงแต่ข้อมูลเป็นเท็จ หรือเรียกว่า False Positive

FN คือ ผลทำนายเป็นเท็จแต่ข้อมูลเป็นจริง หรือเรียกว่า False Negative

TN คือ ผลทำนายเป็นเท็จและข้อมูลเป็นเท็จ หรือเรียกว่า True Negative

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{2.14}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{2.15}$$

$$\text{F1 Score} = \frac{2 * (\text{Precision} * \text{Recall})}{(\text{Precisio} + \text{Recall})} \tag{2.16}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{FP} + \text{FN} + \text{TN})} \tag{2.17}$$

2.3.2 ฟังก์ชันการสูญเสีย (Loss function)

ในงานวิจัยนี้จะให้ความสำคัญกับค่าฟังก์ชันการสูญเสียเนื่องจากเป็นฟังก์ชันที่ใช้ในการคำนวณค่าความผิดพลาด (Error) ของโครงข่ายประสาทเทียม ซึ่งผู้วิจัยจะนำความชันที่เกิดจากการหาอนุพันธ์ของฟังก์ชันการสูญเสียไปปรับค่าน้ำหนัก โดยในงานวิจัยได้เลือก ค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error : MSE) ที่ใช้ในการปรับค่าน้ำหนัก ตามสมการที่ (2.18) และ (2.19) [16]

$$\text{Loss}(y, \hat{y}) = \sum_{i=1}^{n} (y - \hat{y})^2 \tag{2.18}$$

$$\text{MSE} = \frac{1}{n} \times \sum (y - \hat{y})^2 \tag{2.19}$$

เมื่อ y คือ ค่าจริง

$\hat{y}$ คือ ค่าทำนาย

n คือ จำนวนข้อมูลทั้งหมด

2.4 เครื่องมือและภาษาที่ใช้ในการพัฒนา

2.4.1 กูเกิลโคแลป (Google Colab)

กูเกิลโคแลปถูกพัฒนาโดยทีมวิจัยของกูเกิล (Google) ที่เปิดโอกาสให้บุคคลทั่วไปสามารถเข้าไปเขียนโค้ด และรันแบบจำลองของการเรียนรู้ของเครื่อง ผ่านเบราว์เซอร์ (Browser) ได้ถ้ามี แอคเคาท์ (Account) ของกูเกิล และ กูเกิลโครม (Google Chrome) ที่สำคัญคือ ไม่เสียค่าใช้จ่ายและมีไลบรารี่พร้อมใช้ที่จำเป็นหลากหลายและหากต้องการใช้งานไลบรารี่อื่น ๆ ก็สามารถนำเข้าไฟล์เพิ่มเติมได้ [17]

2.4.2 ภาษาไพทอน (Python)

ภาษาไพทอนคือ ชื่อภาษาที่ใช้ในการเขียนโปรแกรมภาษาหนึ่ง ซึ่งถูกพัฒนาขึ้นมาโดยไม่ยึดติดกับแพลตฟอร์ม กล่าวคือ สามารถรันภาษา ไพทอน ยังเป็นแหล่งข้อมูลแบบเปิด (open source) เหมือนภาษา พีเอชพี (PHP) ทำให้ทุกคนสามารถที่จะนำ ไพทอน มาพัฒนาโปรแกรมของผู้วิจัยได้ฟรีโดยไม่ต้องเสียค่าใช้จ่าย และความเป็นแหล่งข้อมูลแบบเปิด ทำให้มีคนเข้ามาช่วยกันพัฒนาให้ ไพทอน มีความสามารถสูงมากยิ่งขึ้น และใช้งานได้ครอบคุมกับทุกลักษณะงาน [18]

2.5 งานวิจัยที่เกี่ยวข้อง

จากการทบทวนวรรณกรรม [19] ใช้ข้อมูลจาก กระทู้ออนไลน์พันทิป 67,449 ข้อมูลความคิดเห็นมาวิเคราะห์ความรู้สึกต่อภาพลักษณ์ของแบรนด์ในปี 2016 โดยมีประเภทคำตอบ 4 ประเภท ได้แก่ความรู้สึกเชิงบวก ความรู้สึกเชิงลบ ความรู้สึกเป็นกลาง และความรู้สึกต้องการ ซึ่งแบ่งโดยใช้คนด้วยแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 3 ชนิด ได้แก่ นาอีฟเบย์ (Naïve Bayes) การถดถอยเชิงโลจิสติค (Logistic Regression) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) มาวิเคราะห์ความรู้สึกต่อภาพลักษณ์ของแบรนด์ ซึ่งได้ค่าความถูกต้องของแบบจำลองเป็น 85.12% 67.30% และ 84.80% ตามลำดับ

จากการทบทวนวรรณกรรม [20] ใช้ข้อมูลจาก Twitter 10,000 ข้อมูลความคิดเห็นมาวิเคราะห์ความรู้สึกของนักท่องเที่ยวต่างประเทศที่วิจารณ์เกี่ยวกับกรุงเทพในปี 2017 โดยมีประเภทคำตอบ 3 ประเภทได้แก่ ความรู้สึกเชิงบวก ความรู้สึกเชิงลบ และความรู้สึกเป็นกลาง ซึ่งแบ่ง ด้วยแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 4 ชนิด ได้แก่ ต้นไม้ตัดสินใจ(Decision Tree) นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน และ โครงข่ายประสาทเทียม มาวิเคราะห์ความรู้สึกของชาวต่างชาติที่มีต่อกรุงเทพมหานคร ประเทศไทย ซึ่งได้ค่าความถูกต้องของแบบจำลองเป็น 79.83% 55.66% 80.11% และ 80.33% ตามลำดับ

จากการทบทวนวรรณกรรม [21] ได้นำข้อมูลความคิดเห็นของนักท่องเที่ยวที่มีต่อจังหวัดภูเก็ต ประเทศไทย บนเว็บไซต์ Tripadvisor จำนวน 65,079 ข้อมูลความคิดเห็นซึ่งประกอบด้วย 25,458 ข้อมูลความคิดเห็นชายหาด 12,584 ข้อมูลความคิดเห็นเกาะต่าง ๆ 3,514 ข้อมูลความคิดเห็นตลาด 1,300 ถนนคนเดิน และ 10,519 ข้อมูลความคิดเห็น โดยให้ 1-3 คะแนนเป็นความรู้สึกเชิงลบ และ 4-5 คะแนนเป็นความรู้สึกเชิงบวกเพื่อจำแนกความรู้สึก

จากการทบทวนวรรณกรรม [22] ใช้วิธีการเรียนรู้ของเครื่องในการวิเคราะห์ปัจจัยที่ส่งผลคะแนนความนิยมของโรงแรมสองโรงแรมโดยข้อมูลที่ใช้มีจำนวน 4,276 ข้อมูลความคิดเห็นจาก Tripadvisor มาวิเคราะห์ด้วยแบบจำลองการจัดสรรข้อมูลแฝงเพื่อสกัดข้อมูลออกมาเป็นคำเฉพาะและใช้ซัพพอร์ตเวกเตอร์แมชชีนในการวิเคราะห์ความรู้สึก โดยให้ 1-3 คะแนนเป็นความรู้สึกเชิงลบ และ 4-5 คะแนนเป็นความรู้สึกเชิงบวก เพื่อจำแนกความรู้สึก จากนั้นนำมาสร้างกราฟความสัมพันธ์ระหว่างความพึงพอใจของลูกค้า (Performance) และ ความสำคัญ (Importance) ของแต่ละปัจจัยออกมาจากแบบจำลองการจัดสรรหัวข้อแฝงซึ่งจากผลลัพธ์สามารถอธิบายได้ว่าปัจจัยที่เกี่ยวกับการตรงต่อเวลา ความพร้อม ความเป็นกันเอง ความเป็นระเบียบของการแต่งกายของพนักงานส่งผลให้ลูกค้ามีความพึงพอใจ และโรงแรมได้รับความนิยมเพิ่มขึ้น

จากการทบทวนวรรณกรรม [23] นำเสนอวิธีการตรวจสอบ การตรวจจับความคล้ายคลึงกันของความหมาย (Semantic Textual Similarity Detection) ในเรื่องความซ้ำซ้อนของเอกสารในภาษาอาหรับ โดยใช้ข้อมูล 2 ส่วน คือ ส่วนที่ 1 เอกสารต้นฉบับ จาก KSUCCA AraCorpus และส่วนที่ 2 จาก วิกิพีเดีย (Wikipedia) เป็นเอกสารต้องสงสัยที่มีความหมายเดียวกันจากแหล่งข้อมูลโดยใช้การประมวลผลภาษาธรรมชาติ โดยในส่วนสำคัญของงานวิจัยนี้จะเสนอการพัฒนาคลังข้อมูลแบบถอดความโดยอัตโนมัติเพื่อรักษาคุณสมบัติของภาษาอาหรับ โดยการรวมสคริปแกรม และนำเสนอวิธีการใช้ สถาปัตยกรรมโครงข่ายประสาท (neural network architecture) สำหรับการสร้างแบบจำลองประโยค (sentence modeling) และความคล้ายคลึงกันเชิงความหมาย(semantic textual similarity) โดยวิธีการที่นำเสนอจะแบ่งออกเป็นกระบวนการในการสร้างเวิร์ดทูเวค เปรียบเทียบข้อดีข้อเสียระหว่างแบบจำลองสคริปแกรมและซีโบล โดยเลือกใช้แบบจำลองสคริปแกรมในการพัฒนา และนำเสนอเทคนิคการแปลงประโยคเป็นตัวเลข โดยใช้ค่าเฉลี่ยแบบจำลองโครงข่ายประสาทแบบคอนโวลูชัน ร่วมกับเทคนิคความคล้ายคลึงของโคไซน์ โดยวิธีการวัดผลจะใช้ค่าความแม่นยำและค่าความระลึก ซึ่งผลของงานวิจัยนี้สรุปว่า ค่าความแม่นยำที่ 85% ค่าความระลึกที่

86.8% และเมื่อเปรียบเทียบกับงานวิจัยอื่นโดยใช้ค่าความถ่วงดุล งานวิจัยนี้ให้ค่าสูงสุดที่ 85.8 ซึ่งสูงกว่าในงานวิจัยประเภทเดียวกันที่ใช้ข้อมูลในลักษณะเดียวกัน

จากการทบทวนวรรณกรรม [24] ข้อมูลจาก SemEval 2019 โดยเป็นข้อมูลความคิดเห็นมาวิเคราะห์ความรู้สึกของในการแนะนำการขุดเหมืองแร่ โดยมีประเภทคำตอบ 2 ประเภทได้แก่ ความรู้สึกเชิงบวก ความรู้สึกเชิงลบ ซึ่งแบ่ง ด้วยแบบจำลองการเรียนรู้เชิงลึกทั้งหมด 3 ชนิด ได้แก่ โครงข่ายประสาทแบบคอนโวลูชัน โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง และ โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู และใช้การโหวดแบบจำลองที่ดีที่สุดเพื่อทำนาย โดยสามารถเพิ่มความถูกต้องของแบบจำลองจาก 0.5035 เป็น 0.7329

บทที่ 3
วิธีการดำเนินงานวิจัย

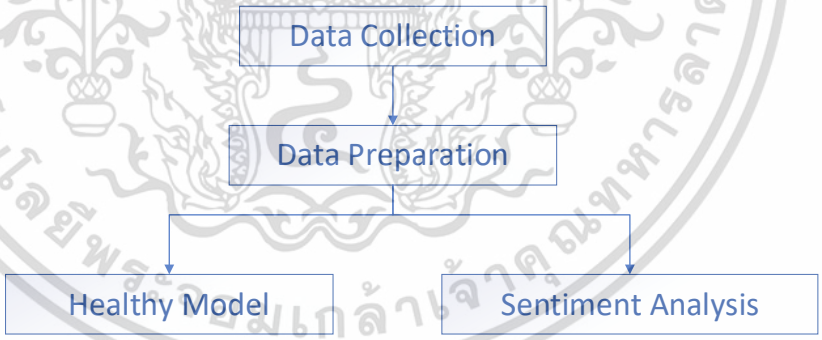
ในงานวิจัยฉบับนี้ได้กำหนดระเบียบวิธีวิจัยได้พัฒนาแนวทางการกำหนดระเบียบวิธีวิจัยการสร้างแบบจำลองในการจำแนกประเภทคนรักสุขภาพจากความคิดเห็นในสื่อสังคมออนไลน์ โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติและแบบจำลองการเรียนรู้เชิงลึกที่หลากหลาย ซึ่งวิธีการวิเคราะห์ข้อมูล เพื่อทำการวิเคราะห์ข้อมูลและตรวจสอบเพื่อสังเคราะห์รูปแบบรายงานขึ้นมาในรูปแบบใหม่ ซึ่งสามารถเข้าถึงข้อมูลได้ตามความต้องการมากขึ้น ทำให้เพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูล และสร้างรายงานสนับสนุนการพยากรณ์และการตัดสินใจของผู้ที่แสดงความคิดเห็นได้เป็นอย่างดี ซึ่งแบ่งเป็นขั้นตอนตามเวิร์กโฟลว์ (Workflow) ดังรูปที่ 3.1

3.1 การรวบรวมข้อมูล (Data Collection)

3.2 การเตรียมข้อมูล (Data Preparation)

3.3 การจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ (Healthy Model)

3.4 การวิเคราะห์ทัศนคติของผู้ที่แสดงความคิดเห็นต่อประกัน (Sentiment Analysis)

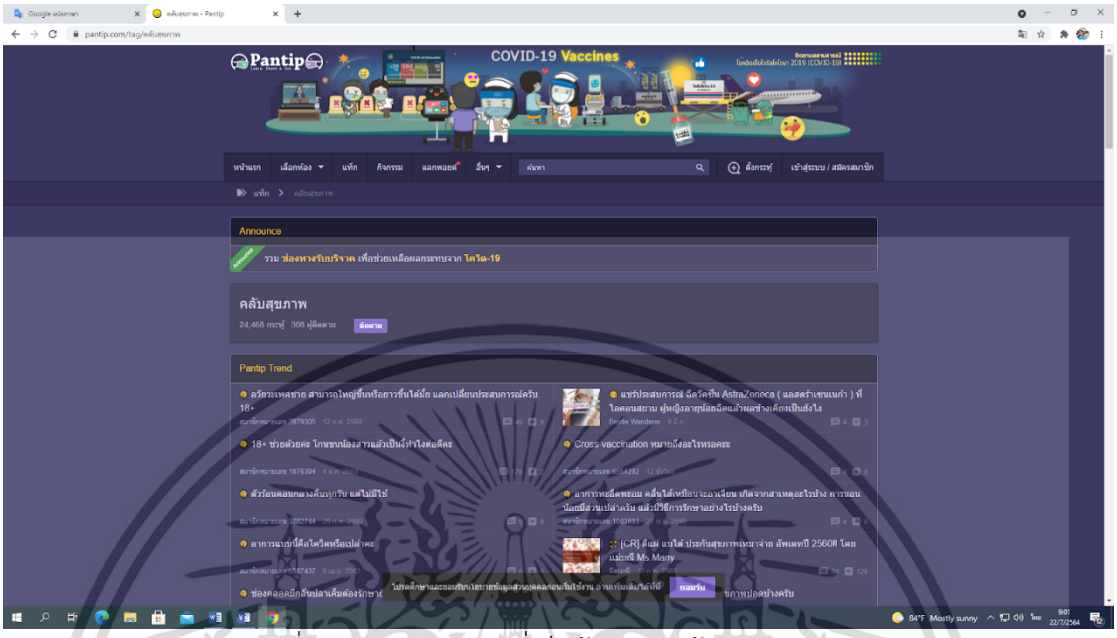


รูปที่ 3.1 เวิร์กโฟลว์ของภาพรวมการทำงานทั้งหมด

3.1 การรวบรวมข้อมูล (Data Collection)

ในการเก็บข้อมูลจากกระทู้ออนไลน์พันทิป ในงานวิจัยฉบับนี้ได้ใช้ภาษาไพธอนเป็นเครื่องมือในการเก็บข้อมูลโดยอัตโนมัติ โดยไลบรารี่ที่สำคัญสำหรับการเก็บข้อมูลผ่านเว็บไซต์ได้ แก่ Beautifulsoup4 ซึ่งเป็นไลบรารี่ที่ช่วยในการค้นหาข้อมูลส่วนที่ผู้วิจัยต้องการในหน้าเว็บไซต์ต่าง ๆ โดยไลบรารี่ตัวนี้จะทำการค้นหาข้อความส่วนที่ผู้วิจัยต้องการจากโครงสร้างของหน้าเว็บไซต์ที่ผู้วิจัย

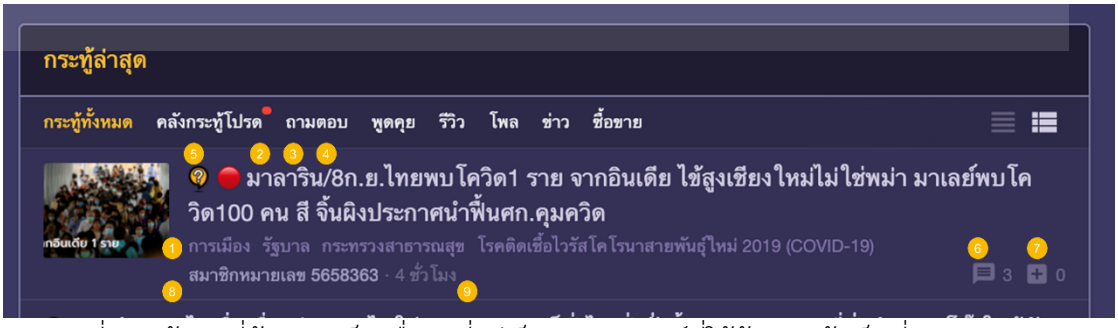
ต้องการ ซึ่งในงานวิจัยฉบับนี้ ที่เกี่ยวข้องกับห้อง "คลับสุขภาพ" ใน กระทู้ออนไลน์พันทิป ดังรูปที่ 3.2 โดยมีรายละเอียดและขั้นตอนดังนี้



รูปที่ 3.2 ตัวอย่างแหล่งที่เก็บข้อมูลจากห้องคลับสุขภาพ

3.1.1 เก็บข้อมูลที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจง (Universal Resource Locator : URL)

เนื่องจากในกระทู้ออนไลน์พันทิป มีกระทู้จำนวนมากดังนั้นจึงไม่สามารถเก็บข้อมูลทั้งหมดได้ ผู้วิจัยจึงจำเป็นต้องเลือกเก็บเฉพาะข้อมูลที่มีความเกี่ยวข้องกับงานวิจัยเท่านั้น โดยเก็บข้อมูล ดังรูปที่ 3.3 และตามรูปแบบใน ตารางที่ 3.1 โดยหลักในการพิจารณา จะพิจารณาจากกระทู้ที่มี แฮชแท็ก “โรคติดเชื้อไวรัสโคโรนา 2019 (COVID-19)” และ แฮชแท็กที่เกี่ยวข้องกับประกัน เช่น “ประกันสุขภาพ” “ประกันออนไลน์” “ประกันชีวิต” หลังจากนั้นผู้วิจัยจะได้ข้อมูลที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจงของกระทู้ที่จะทำการเก็บจำนวน 456 กระทู้ โดยแบ่งเป็นกระทู้ประเภทสนทนาแสดงความคิดเห็น 180 กระทู้ และกระทู้ประเภทรีวิว 276 กระทู้



รูปที่ 3.3 ข้อมูลที่ต้องการเก็บเพื่อหาที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจง

ตารางที่ 3.1 รูปแบบการเก็บข้อมูลที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจง

ชื่อตัวแปร	หมายเลข	รูปแบบข้อมูล	ความหมาย
Tag	1	String	เก็บแฮชแท็กทั้งหมดของกระทู้
Record	2	Numeric	เก็บเลขกระทู้
Topic	3	String	เก็บชื่อกระทู้
URL	4	String	เก็บที่อยู่เว็บ
Type	5	String	เก็บประเภทของกระทู้
Comment	6	String	เก็บจำนวนความคิดเห็น
plustor	7	Numeric	เก็บจำนวนความคิดเห็นเพิ่มเติม
User	8	Numeric	เก็บชื่อของผู้แสดงความคิดเห็น
Time	9	Date	เก็บเวลาในวันที่แสดงความคิดเห็น
ID	10	Numeric	เก็บไอดีของผู้แสดงความคิดเห็น

3.1.2 เก็บข้อมูลกระทู้ประเภทแสดงความคิดเห็น

โดยข้อมูลในส่วนนี้จะถูกรวบรวมไว้ทั้งหมด 1,490 ความคิดเห็น จาก 180 กระทู้ โดยเก็บข้อมูล ดังรูปที่ 3.4 และตามรูปแบบใน ตารางที่ 3.2 โดยทางผู้วิจัยจะนำข้อมูลในส่วนนี้ไปประยุกต์ใช้ในการจำแนกกลุ่มของคนที่มีแนวโน้มรักสุขภาพ

รูปที่ 3.4 ข้อมูลที่ต้องการเก็บในกระทู้ประเภทแสดงความคิดเห็น

ตารางที่ 3.2 รูปแบบการเก็บข้อมูลในกระทู้ประเภทแสดงความคิดเห็น

ชื่อตัวแปร	หมายเลข	รูปแบบข้อมูล	ความหมาย
Topic ID	1	Numeric	เก็บไอดีของกระทู้
Topic	2	String	เก็บชื่อของกระทู้
Detail Topic	3	String	เก็บรายละเอียดของกระทู้
No Comment	4	Numeric	เก็บลำดับความคิดเห็น
Detail Comment	5	String	เก็บรายละเอียดความคิดเห็น
ID Comment	6	String	เก็บไอดีของผู้ที่แสดงความคิดเห็น
Time Comment	7	Date	เก็บเวลาที่แสดงความคิดเห็น

3.1.3 เก็บข้อมูลกระทู้ประเภทรีวิว

เป็นการเก็บข้อมูลข้อความแสดงความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยรายละเอียดของข้อมูลผู้วิจัยจะใช้ข้อมูลของผู้ที่แสดงความคิดเห็นประเภทรีวิวที่เกี่ยวข้องกับประกันภัย โดยมีทั้งหมด 276 กระทู้ ข้อมูลประเภทรีวิวในกระทู้ออนไลน์พันทิป จะมีการให้คะแนนในการแสดงความคิดเห็นใน

กระทู้นั้น มีคะแนนตั้งแต่ 0 ถึง 5 โดยเก็บข้อมูล ดังรูปที่ 3.5 และตามรูปแบบใน ตารางที่ 3.3 โดยทางผู้วิจัยได้ประยุกต์ใช้ข้อมูลในส่วนนี้ในการจำแนกทัศนคติของผู้ที่แสดงความคิดเห็นต่อประกัน

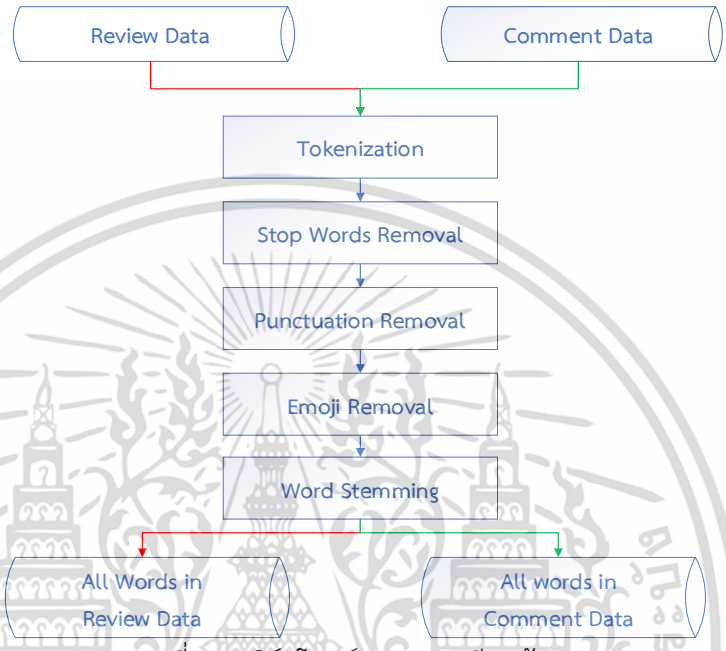
รูปที่ 3.5 ข้อมูลที่ต้องการเก็บในกระทู้ประเภทรีวิว

ตารางที่ 3.3 รูปแบบการเก็บข้อมูลในกระทู้ประเภทรีวิว

ชื่อตัวแปร	หมายเลข	รูปแบบข้อมูล	ความหมาย
Topic ID	1	Numeric	เก็บไอดีของกระทู้
Topic	2	String	เก็บชื่อของกระทู้
Detail Topic	3	String	เก็บรายละเอียดของกระทู้
Review	4	String	เก็บข้อมูลรีวิว
Score	5	Numeric	เก็บคะแนนที่ได้
Review Type	6	String	เก็บประเภทของกระทู้รีวิว

3.2 การเตรียมข้อมูล (Data Preparation)

หลังจากทำการเก็บข้อมูลจากกระทู้ออนไลน์พันทิปแล้ว พบว่าข้อมูลยังไม่สามารถนำไปใช้ในการวิเคราะห์ได้ทันที จึงต้องมีการนำข้อมูลความคิดเห็น มาเข้าสู่ขั้นตอนการเตรียมข้อมูล โดยผู้วิจัยจะพิจารณาเฉพาะข้อมูลความคิดเห็นที่เป็นภาษาไทยเท่านั้น โดยข้อมูลทั้ง 2 ส่วนที่ถูกเก็บมาจะใช้กระบวนการเตรียมข้อมูลเหมือนกัน โดยมีขั้นตอนอ้างอิงจากในหัวข้อที่ 2.1 และแสดงดังรูป 3.6



รูปที่ 3.6 เวิร์กโฟลว์ของการเตรียมข้อมูล

ตัวอย่างการเตรียมข้อมูลนี้ ประโยคที่ใช้คือประโยค “ฉันดื่มน้ำกระเจี๊ยบ! 😀” ซึ่งจะต้องผ่านขั้นตอนการเตรียมข้อมูลทั้ง 5 ขั้นตอน ได้แก่ การตัดคำ การลบคำที่ไม่มีนัยสำคัญ การลบเครื่องหมายวรรคตอน การลบรูปสัญลักษณ์ชนิดข้อความ การตัดส่วนท้ายของคำ ซึ่งจากประโยคตัวอย่างดังกล่าวเมื่อนำมาผ่านขั้นตอนการเตรียมข้อมูลทั้ง 5 ขั้นตอน ผลลัพธ์แต่ละขั้นตอนสามารถแสดงได้ดังตารางที่ 3.4

ตารางที่ 3.4 ตัวอย่างขั้นตอนการเตรียมข้อมูล

ขั้นตอน	คำอธิบาย	ผลลัพธ์
1. การตัดคำ	นำประโยคมาแบ่งออกเป็นคำ	“ฉัน”, “ดื่ม”, “น้ำ”, “กระเจี๊ยบ”, “!”, “😀”
2. การลบคำที่ไม่มีนัยสำคัญ	ลบ “ฉัน” ออก	“ดื่ม”, “น้ำ”, “กระเจี๊ยบ”, “!”, “😀”

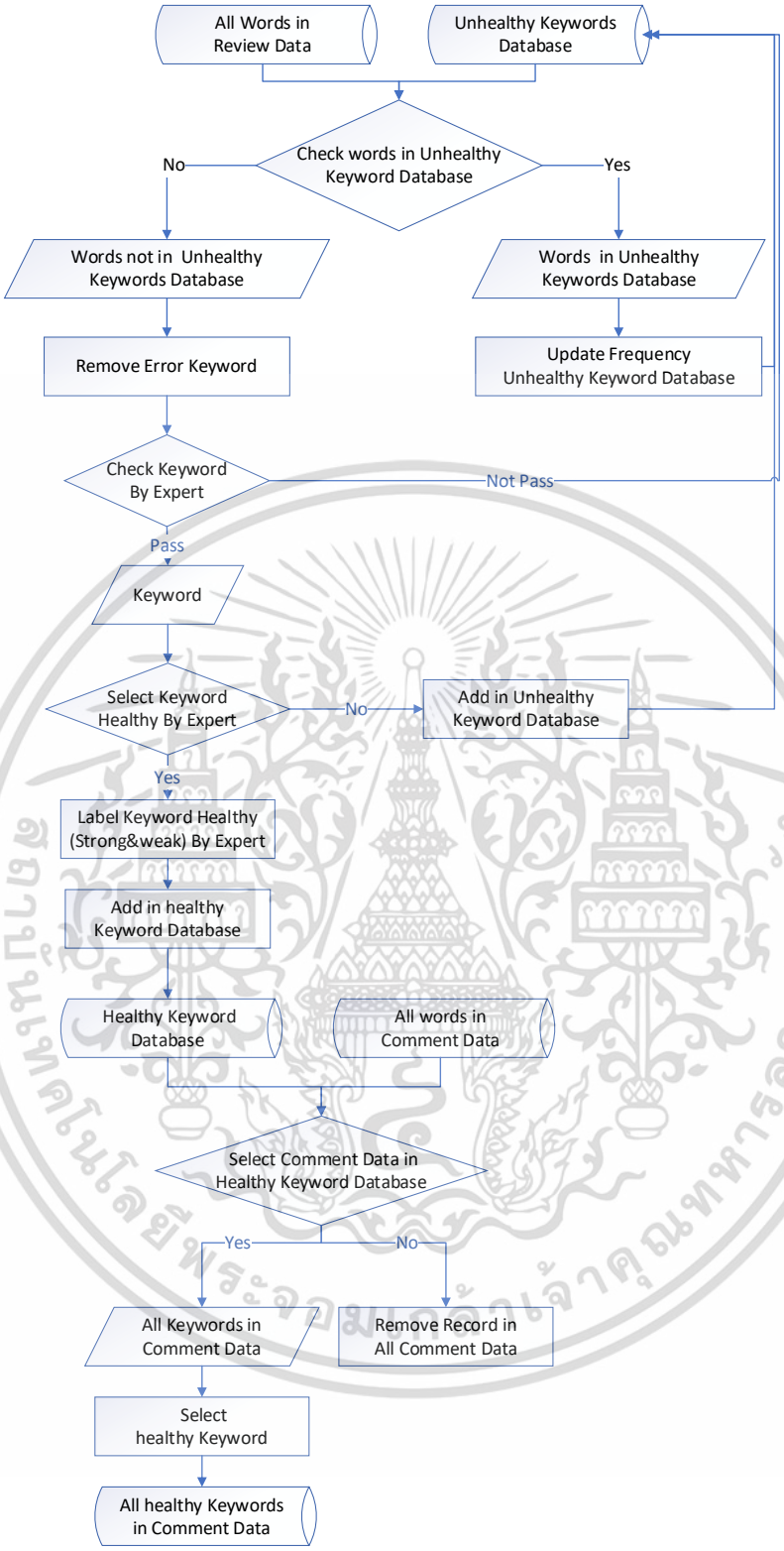
3. การลบเครื่องหมายวรรคตอน	ลบเครื่องหมาย “!” ออก	“ดื่ม”, “น้ำ”, “กระเจี๊ยบ”, “😀”
4. การลบรูปสัญลักษณ์ชนิดข้อความ	ลบสัญลักษณ์ “😀” ออก	“ดื่ม”, “น้ำ”, “กระเจี๊ยบ”
5. การตัดส่วนท้ายของคำ	การทำให้คำศัพท์อยู่ในรูปปกติโดยในตัวอย่างนี้คือการทำให้ “กระเจี๊ยบ” เป็น “กระเจี๊ยบ”	“ดื่ม”, “น้ำ”, “กระเจี๊ยบ”

3.3 การจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ (Healthy Model)

มีวัตถุประสงค์เพื่อจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพในกระทู้ออนไลน์พันทิป โดยทำการทดลอง 2 ส่วนคือ ส่วนที่ 1 การทดลองปรับพารามิเตอร์ที่เหมาะสมสำหรับเทคนิคการแปลงคำเป็นตัวเลข และส่วนที่ 2 นำเสนอวิธีการจำแนกโดยใช้เทคนิคการแปลงประโยคเป็นตัวเลขทั้ง 3 วิธีร่วมกับเปอร์เซ็นต์ความคล้ายคลึงของโคไซด์ (Percentage of cosine similarity) เปรียบเทียบกับการจำแนกแบบปกติ (Original Classify) โดยมีขั้นตอนดังนี้

3.3.1 การเลือกข้อมูล

หลังจากกระบวนการเตรียมข้อมูลจะได้คำสำคัญโดยแบ่งออกเป็น 2 ส่วน ส่วนที่ 1 มาจากกระทู้ประเภทรีวิว และส่วนที่ 2 มาจากกระทู้ประเภทแสดงความคิดเห็น ในขั้นตอนนี้ผู้วิจัยจะให้ความสำคัญกับคำสำคัญที่มาจากกระทู้ประเภทรีวิวเป็นหลัก เนื่องจากทางผู้วิจัยได้มีเป้าหมายที่จะหาคำสำคัญที่เกี่ยวกับสุขภาพจากกระทู้ประเภทรีวิวในคลับสุขภาพ เพื่อเป็นคำสำคัญที่จะใช้ในการพัฒนาแบบจำลองต่อไป โดยมีขั้นตอนดังรูป 3.7



รูปที่ 3.7 เวิร์กโฟลว์ของการเลือกข้อมูล

จากรูปที่ 3.7 จะได้คำสำคัญในกระทู้ประเภทรีวิว ที่ผ่านกระบวนการเตรียมข้อมูลแล้ว ทางผู้วิจัยจะนำคำสำคัญทั้งหมด ไปตรวจสอบด้วยคำในฐานข้อมูลคำสำคัญที่ไม่เกี่ยวกับสุขภาพ (Unhealthy Keywords Database) ถ้าหากพบคำสำคัญที่ตรงกับ ฐานข้อมูลคำสำคัญที่ไม่เกี่ยวกับ

สุขภาพ คำสำคัญนั้นจะถูกคัดกรองออก และไปรวมในฐานข้อมูลคำสำคัญที่ไม่เกี่ยวกับสุขภาพ เพื่อหาความถี่ของคำดังกล่าวที่เกิดขึ้นเพื่อนำไปต่อยอดทางธุรกิจต่อไป แต่หากคำสำคัญไม่ตรงกันจะถูกพิจารณาในกระบวนการต่อไป

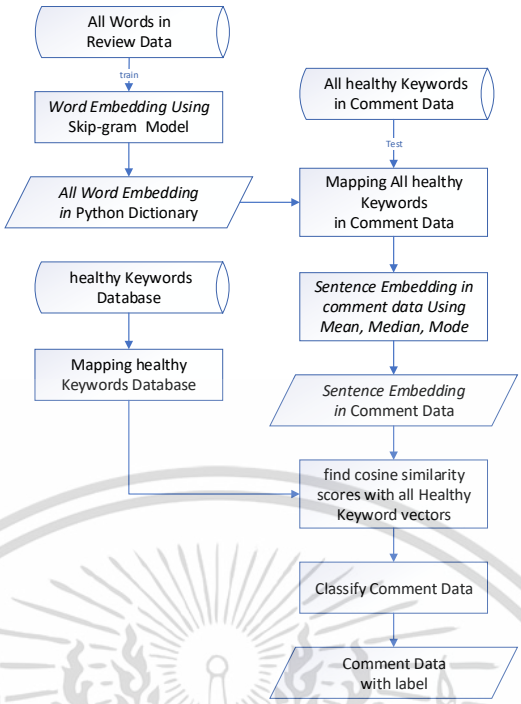
หลังจากผู้วิจัยตรวจสอบคำสำคัญที่ไม่เกี่ยวกับสุขภาพ (Unhealthy Keywords) ผู้วิจัยจะทำการลบคำที่อาจจะเกิดจากความผิดพลาดการตัดคำ (Error Keywords) เช่น “รักสุขภาพมากกกกกกกกก” ในบางครั้งคำที่ถูกตัดออกมา อาจจะเป็น “รัก” , “สุขภาพ” , “มาก” , “กกกกกกกก” ผู้วิจัยจึงจำเป็นต้องตัดคำสำคัญที่เป็น “กกกกกกกก” ออกไป ซึ่งคำสำคัญที่ถูกตัดออกจะไปถูกส่งไปที่ฐานข้อมูลคำสำคัญที่ไม่เกี่ยวกับสุขภาพ ส่วนคำที่ไม่ถูกตัดออกจะถูกส่งไปยังกระบวนการถัดไป

ขั้นตอนถัดไปจะเป็นการตรวจสอบโดยผู้เชี่ยวชาญ โดยจะพิจารณาว่าเป็นคำสำคัญที่เกี่ยวกับสุขภาพ (Healthy Keywords) หรือไม่ หากตรวจสอบแล้วพบว่าเป็นคำสำคัญที่ไม่เกี่ยวกับสุขภาพทางผู้วิจัยจะนำไปรวมกับฐานข้อมูลคำสำคัญที่ไม่เกี่ยวกับสุขภาพ แต่หากเป็นคำสำคัญที่เกี่ยวกับสุขภาพ จะถูกเลเบลคำสำคัญนั้นว่าเป็นประเภทรักสุขภาพมากหรือรักสุขภาพน้อย และจะถูกนำไปเพิ่มที่ ฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพ (Healthy Keywords Database)

จากนั้นผู้วิจัยจะนำข้อมูลกระทู้ประเภทแสดงความคิดเห็นในแต่ละข้อมูลไปตรวจสอบกับฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพ โดยหากไม่พบว่าความคิดเห็นนั้นมีคำสำคัญที่ตรงกันในฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพ ความคิดเห็นนั้นจะถูกตัดออกจากการวิเคราะห์ แต่ถ้าหากความคิดเห็นนั้นมีคำสำคัญที่อยู่ในฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพความคิดเห็นนั้นจะถูกไปพัฒนาแบบจำลอง ซึ่งจากข้อมูลความคิดเห็นทั้งหมด 1,490 ความคิดเห็น จะเหลือเพียง 555 ความคิดเห็น

3.3.2 การพัฒนาแบบจำลอง (Data Modelling)

ในการพัฒนาแบบจำลองจะใช้ข้อมูล 2 ส่วนคือ ข้อมูลที่ใช้ในการสอน (Training) และ ข้อมูลที่ใช้ในการทดสอบ (Testing) ในส่วนที่ 1 สำหรับข้อมูลที่ใช้ในการสอนจะใช้ข้อมูลกระทู้รีวิวทั้งหมดที่อยู่ในคลับสุขภาพ และในส่วนที่ 2 สำหรับข้อมูลที่ใช้ในการทดสอบจากในกระบวนการเลือกข้อมูลจะมีข้อมูลความคิดเห็นจำนวน 555 ข้อความ แต่ด้วยข้อความภาษาไทยมีความกำกวมในบางส่วนผู้วิจัยจึงให้ผู้เชี่ยวชาญทางธุรกิจเลือกข้อมูลที่ถูกเลเบลเพียงแค่ 30% ของข้อมูลทั้งหมดที่ทางผู้เชี่ยวชาญมั่นใจเพื่อให้ความคิดเห็นที่ถูกเลเบลมีคุณภาพและแม่นยำที่สุด เพื่อใช้วัดประสิทธิภาพของแบบจำลองต่อไป



รูปที่ 3.8 เวิร์กโฟลว์ของพัฒนาแบบจำลอง

จากรูปที่ 3.8 ขั้นตอนแรกคือการใช้ข้อมูลที่เก็บจากกระทู้ประเภทรีวิวที่ผ่านกระบวนการเตรียมข้อมูลแล้วในการแปลงโดยใช้เทคนิคการแปลงคำเป็นตัวเลขในหัวข้อ 3.3.2.1 ในการพัฒนาแบบจำลอง โดยผลลัพธ์ของกระบวนการนี้คือในทุกคำสำคัญจะถูกแปลงเป็นเวกเตอร์ โดยในแต่ละคำสำคัญจะมีเลเบลระบุประเภทของคำสำคัญนั้น

3.3.2.1 เทคนิคการแปลงคำเป็นตัวเลข (Word Embedding)

จากงานวิจัยเนื่องจากแบบจำลองสคริปแกรมให้ผลลัพธ์ที่ดีกว่าแบบจำลองซีโบลแต่จะใช้เวลาในการสอนแบบจำลองมากกว่า ซึ่งผู้วิจัยเลือกที่จะใช้เทคนิคการแปลงคำเป็นตัวเลข โดยวิธีแบบจำลองสคริปแกรมโดยทำการทดลองในการสร้างแบบจำลองดังนี้

การทดลองที่ 1

ผู้วิจัยได้ทำการปรับค่าพารามิเตอร์ที่ใช้ในการสร้างแบบจำลอง เพื่อหาพารามิเตอร์ที่ดีที่สุดที่ทำให้ฟังก์ชันการสูญเสียต่ำที่สุดและใช้เวลาในการพัฒนาแบบจำลองที่เหมาะสม โดยทำการทดลองดังนี้

1. กำหนดค่าระยะทางสูงสุดจากคำที่กำลังใช้ฝึกฝน (Window Size) ที่นับว่าคำที่เจอเป็นบริบทของคำที่กำลังใช้ฝึกฝนอยู่หรือ เป็น 2 ในสำหรับทุก แบบจำลอง [25]

2. กำหนดจำนวนรอบที่ใช้สอนแบบจำลอง (Epoch) ทั้งหมด 10 ค่าได้แก่ 10, 20, 30, 40, 50, 60, 70, 80, 90, 100

3. กำหนดจำนวนมิติ (Dimensions) ทั้งหมด 3 ค่าได้แก่ 300, 400, 500 , 600, 700

โดยผลลัพธ์ของการทดลองทุกคำที่ถูกฝึกฝนจะมีเวกเตอร์ของคำนั้น หลังจากนั้นทางผู้วิจัยจะทำการเชื่อม (Mapping) เพื่อดึงเอาเวกเตอร์ของคำสำคัญที่เกี่ยวกับสุขภาพที่อยู่ในข้อมูลความคิดเห็น ดังรูปที่ 3.8 หลังจากนั้นในทุกข้อมูลความคิดเห็นจะมีเวกเตอร์ของคำสำคัญที่เกี่ยวกับสุขภาพ แต่เนื่องจากในแต่ละแต่ละข้อมูลความคิดเห็นมีจำนวนคำที่ไม่เท่ากัน ทางผู้วิจัยจึงประยุกต์ใช้เทคนิคการแปลงประโยคเป็นตัวเลข เพื่อแปลงข้อมูลความคิดเห็นแต่ละรายการเป็นเวกเตอร์เดียว

3.3.2.2 เทคนิคการแปลงประโยคเป็นตัวเลข

หลังจากที่ผู้วิจัยได้ เทคนิคการแปลงคำเป็นตัวเลข โดยการเชื่อมแต่ละคำที่ผู้วิจัยใช้ฝึกฝนกับเวกเตอร์ที่ได้จากแบบจำลอง โดยผู้วิจัยจะแปลงจาก รูปที่ 3.8 ผู้วิจัยจะทำการทดลองโดยการแปลงเป็นเทคนิคการแปลงประโยคเป็นตัวเลข ทั้งหมด 3 ค่า ได้แก่ ค่าเฉลี่ย (Mean) ค่ากลาง (Median) และ ค่าฐานนิยม (Mode) เพื่อหาวิธีการที่ดีที่สุด โดยมีรายละเอียด ดังนี้

1.วิธีที่ใช้ค่าเฉลี่ย คือผู้วิจัยจะแปลงเวกเตอร์โดยที่แต่ละมิติของเวกเตอร์ เป็นค่าเฉลี่ยของเวกเตอร์ทั้งหมด

2.วิธีที่ใช้ค่าฐานนิยม คือผู้วิจัยจะแปลงเวกเตอร์โดยที่แต่ละมิติของเวกเตอร์ เป็นค่าความถี่สูงสุดของเวกเตอร์ทั้งหมด

3.วิธีที่ใช้ค่ากลาง คือผู้วิจัยจะแปลงเวกเตอร์โดยที่แต่ละมิติของเวกเตอร์เป็น ค่ากลางของเวกเตอร์ทั้งหมด

ดังนั้นจากทั้ง 3 วิธีที่กล่าวมาผลลัพธ์ของการแสดงความคิดเห็นเป็นเวกเตอร์เดียว

3.3.2.3 หาค่าความคล้ายคลึงของโคไซน์

จากที่ผู้วิจัยได้ใช้เทคนิคการแปลงประโยคเป็นตัวเลขแล้ว ในขั้นตอนนี้ผู้วิจัยจะนำเวกเตอร์ของความคิดเห็นที่ได้จากกระบวนการก่อนมาหาค่าความคล้ายคลึงของโคไซน์กับทุกเวกเตอร์ของคำสำคัญในฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพ แล้วเลือกคำสำคัญที่มีค่าค่าความคล้ายคลึงของโคไซน์มากที่สุด เนื่องจากทุกคำสำคัญในฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพ จะถูกเลเบลจากผู้เชี่ยวชาญทางธุรกิจ

การทดลองที่2

ผู้วิจัยได้ทำการทดลองเปรียบเทียบวิธีการทำเทคนิคการแปลงประโยคเป็นตัวเลขร่วมกับเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ โดยมีรายละเอียด ดังนี้

1. ทดลองทำเทคนิคการแปลงประโยคเป็นตัวเลข ทั้งหมด 3 วิธี ได้แก่ ค่าเฉลี่ย ค่าฐานนิยม และค่ากลาง

2. ใช้เทคนิคการแปลงประโยคเป็นตัวเลข ที่ได้มาหาค่าความคล้ายคลึงของโคไซน์ กับทุกคำสำคัญในฐานข้อมูลคำสำคัญที่เกี่ยวกับสุขภาพ หลังจากนั้นทางผู้วิจัยจะทำการทดลองการจำแนกแบบใช้เปอร์เซ็นต์ของค่าความคล้ายคลึงของโคไซน์ (Percentage of Cosine Similarity Classify) โดยทางผู้วิจัยจำแนกและจะทำการทดลองตั้งแต่ เปอร์เซ็นต์ที่ 0 ถึง 99 โดยมีเกณฑ์คือ จะใช้เปอร์เซ็นต์ของค่าความคล้ายคลึงของโคไซน์ที่เลือกในการจำแนก โดยถ้าเปอร์เซ็นต์ของค่าความคล้ายคลึงของโคไซน์ มากกว่าเท่ากับค่าความคล้ายคลึงของโคไซน์ที่เลือกจะจำแนกตามจะถูกเลเบลของคำสำคัญนั้น แต่ในทางกลับกัน ถ้าค่าความคล้ายคลึงของโคไซน์น้อยกว่าค่าความคล้ายคลึงของโคไซน์ที่เลือกจะจำแนกตรงข้ามเลเบลของคำสำคัญนั้น

โดยทางผู้วิจัยจะใช้เทียบกับการจำแนกแบบปกติซึ่งเป็นการไม่ใช้ เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ โดยจะจำแนกตามเลเบลของคำสำคัญที่มีค่าความคล้ายคลึงของโคไซน์มากที่สุด

3.4 การจำแนกทัศนคติของผู้ที่แสดงความคิดเห็นต่อประกัน

มีวัตถุประสงค์ เพื่อจำแนกความรู้สึกของความคิดเห็นที่มีต่อประกัน ว่าเป็นความรู้สึกเชิงบวกหรือเชิงลบจากข้อมูลกระทู้ออนไลน์พันทิป ในงานวิจัยนี้ มุ่งเน้นในการศึกษาการนำการวิเคราะห์ความรู้สึกแบบโครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง และ โครงข่ายประสาทแบบคอนโวลูชัน ไปประยุกต์ใช้ในการสร้างแบบจำลองการทำนายความรู้สึกจากข้อความในข้อมูลความคิดเห็นและการเสนอความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยมีวิธีการตาม

3.4.1 การบ่งบอกประเภทของความรู้สึกในกระทู้ออนไลน์พันทิป

กำหนดโดยระดับคะแนนที่ผู้เขียนกระทู้ออนไลน์พันทิปให้ไว้ในข้อมูลที่เก็บมาจากกระทู้รีวิวในพันทิป การจำแนกความรู้สึกสามารถแบ่งได้เป็นความรู้สึกเชิงบวก (4-5 คะแนน) และความรู้สึกเชิงลบ (0-3 คะแนน) ซึ่งผู้วิจัยสามารถนำประเภทความรู้สึกนี้เป็นชุดข้อมูลฝึก

3.4.2 การพัฒนาแบบจำลอง (Data Modelling)

ก่อนการพัฒนาแบบจำลองจำเป็นต้องมีการแบ่งข้อมูล (Split Data) ในการศึกษาครั้งนี้จะทำการแบ่งทั้งหมด 3 ส่วน คือ ข้อมูลที่ใช้สอน ข้อมูลตรวจสอบ (Validation) และข้อมูลทดสอบ โดยมีสัดส่วน เป็น 80% 10% 10% ตามลำดับ โดยข้อมูลฝึกเป็นข้อมูลเพื่อสอนแบบจำลอง ข้อมูลตรวจสอบนั้นเพื่อปรับจูนพารามิเตอร์แบบจำลอง และข้อมูลทดสอบใช้สำหรับทดสอบความถูกต้อง

ของแบบจำลองที่ผู้วิจัยสร้างขึ้น โดยในงานวิจัยนี้ผู้วิจัยจะใช้เทคนิคการเรียนรู้เชิงลึกในการวิเคราะห์ความพึงพอใจของผู้ใช้งานต่อประกันภัยจากกระทู้ออนไลน์พันทิป

ในส่วนของการวิเคราะห์ข้อมูล มีเป้าหมายในการวิเคราะห์ความรู้สึก เพื่อดึงความรู้สึกจากข้อความเป็นไปในเชิงใด โดยใช้แบบจำลองแบบมีผู้สอน ซึ่งจะใช้เทคนิคการจำแนก ทั้ง 3 แบบจำลองต่อไปนี้ โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู และ โครงข่ายประสาทแบบคอนโวลูชัน

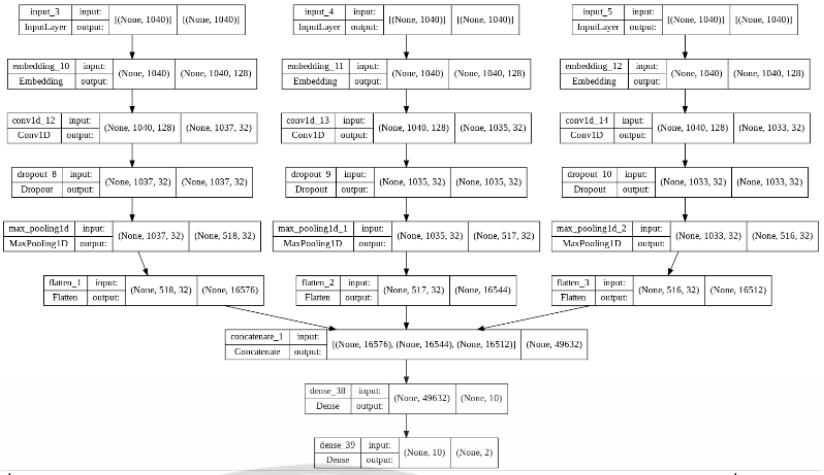
ในงานวิจัยนี้ในแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด (version) โดยแบบจำลองชุดที่ 1 จะเป็นการสร้างแบบจำลองแบบโดยใช้ตัวแปรที่ผ่านกระบวนการ การสกัดคุณลักษณะ ส่วนในแบบจำลองชุดที่ 2 จะใช้ตัวแปรที่ผ่านกระบวนการการสกัดคุณลักษณะและเพิ่มตัวแปรที่ถูกเพิ่มเติมที่ทางผู้วิจัยคาดว่าจะสามารถเพิ่มประสิทธิภาพของแบบจำลองและสามารถปรับใช้กับทางธุรกิจได้ โดยตัวแปรดังกล่าวคือ ประเภทของกระทู้รีวิวซึ่งตัวแปรนี้มีอยู่ในทุกกระทู้รีวิวในกระทู้ออนไลน์พันทิปโดยมีรายละเอียด ดังนี้

1. กระทู้รีวิวจากธุรกิจ (Business Review : BR)
2. กระทู้รีวิวจากเจ้าของกระทู้ (Consumer Review : CR)
3. กระทู้รีวิวจากผู้สนับสนุน (Sponsored Review : SR)

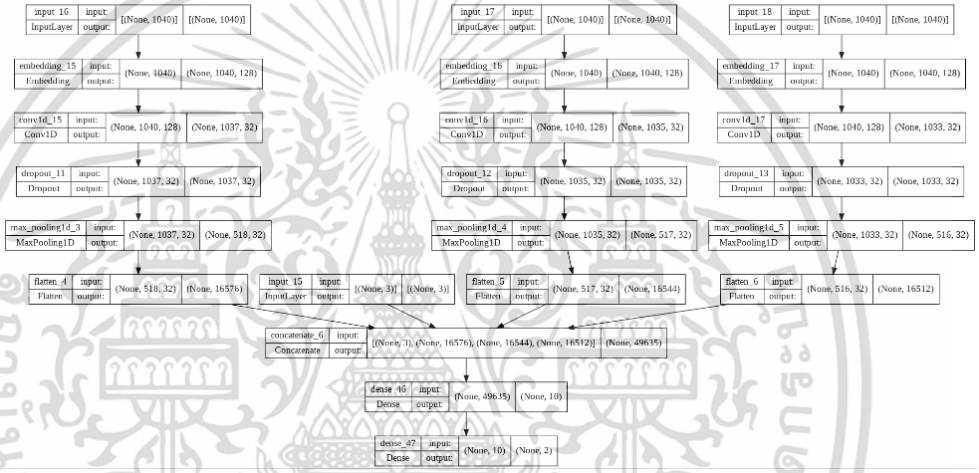
ซึ่งทางผู้วิจัยทำการทดลองทั้งหมด 3 แบบจำลอง โดยมีรายละเอียด ดังนี้

3.4.2.1 โครงข่ายประสาทแบบคอนโวลูชัน

โครงข่ายประสาทแบบคอนโวลูชัน ที่ใช้ในงานวิจัยนี้มีโครงสร้าง ดังรูปที่ 3.9 และ รูปที่ 3.10 ในการศึกษาครั้งนี้ได้กำหนดพารามิเตอร์สำหรับ โครงข่ายประสาทแบบคอนโวลูชัน ดังนี้ Epochs = 300, Batch size = 64, Adam Optimization, Learning rate = 0.0001, Kernel size = 4, 6, 8, Dropout = 0.5, Max Pooling, Activation function (ReLU - Hidden Layer, Softmax - Output Layer) อ้างอิงจากหัวข้อ 2.2.4



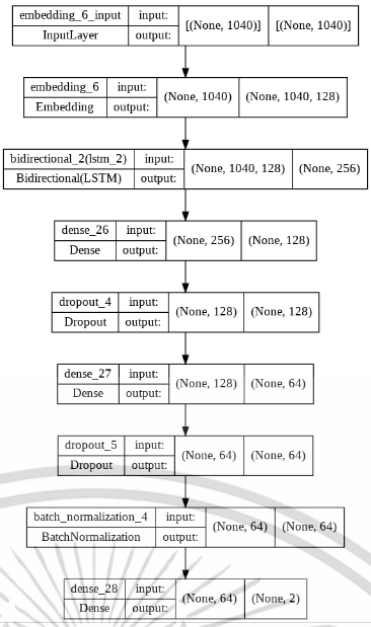
รูปที่ 3.9 โครงสร้างแบบจำลองโครงข่ายประสาทแบบคอนโวลูชัน ชุดที่ 1



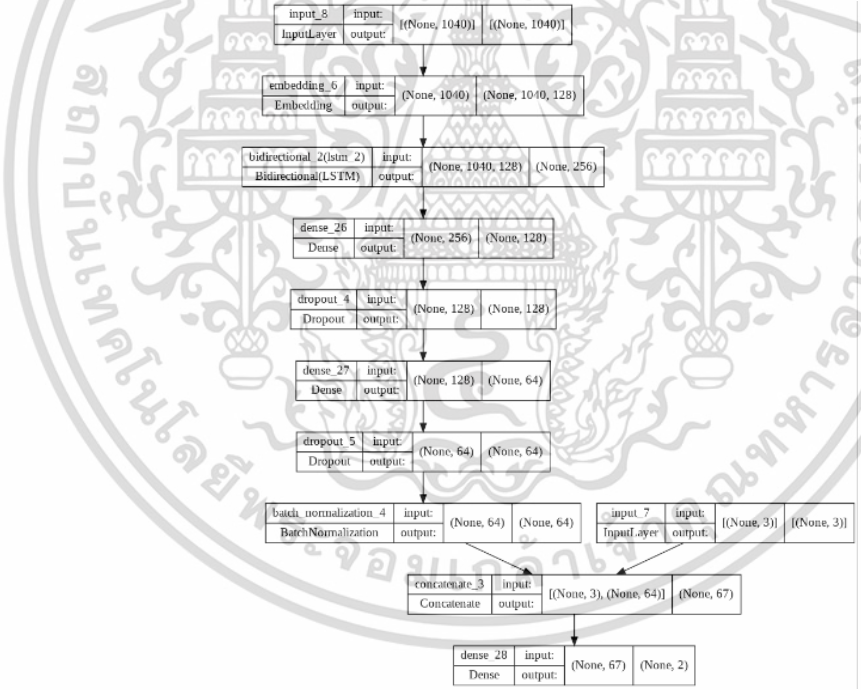
รูปที่ 3.10 โครงสร้างแบบจำลองโครงข่ายประสาทแบบคอนโวลูชัน ชุดที่ 2

3.4.2.2 หน่วยความจำระยะยาว-ระยะสั้น (Bi-Long Short-Term Memory: Bi-LSTM)

โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง ที่ใช้ในงานวิจัยนี้มีโครงสร้าง ดังรูปที่ 3.11 และ รูปที่ 3.12 ในการศึกษาครั้งนี้ได้กำหนดพารามิเตอร์สำหรับ โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง ดังนี้ Epochs = 300, Batch size = 64, Optimization Adam, Learning rate = 0.0001, Dropout = 0.25, Batch Normalization, Activation function (ReLu - Hidden Layer, Softmax - Output layer) อ้างอิงจากหัวข้อ 2.2.5 และ 2.2.6



รูปที่ 3.11 โครงสร้างแบบจำลอง โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง ชุดที่ 1

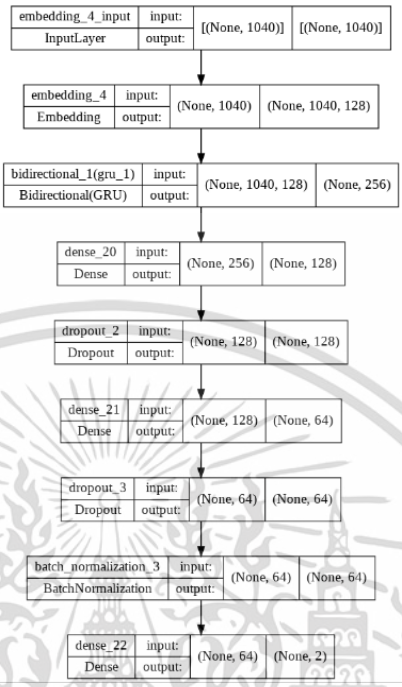


รูปที่ 3.12 โครงสร้างแบบจำลอง โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง ชุดที่ 2

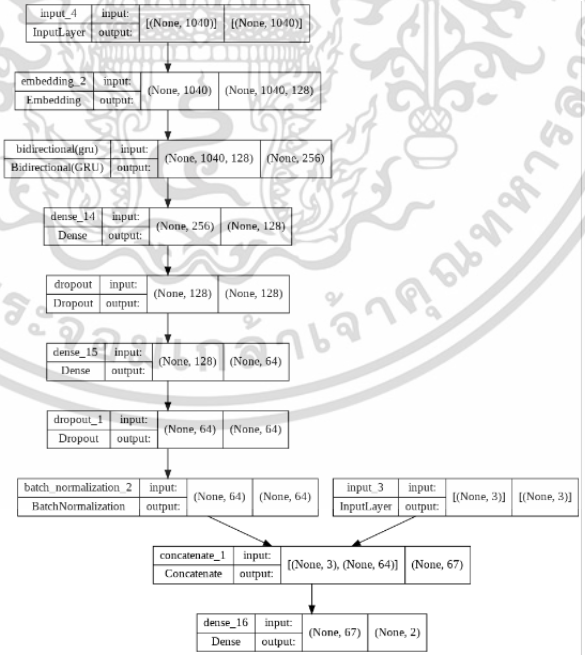
3.4.2.3 หน่วยเวียนกลับแบบมีประตู (Gated Recurrent Unit: GRU)

โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู มีโครงสร้างดังที่ใช้ในงานวิจัยนี้มีโครงสร้าง ดังรูปที่ 3.13 และ รูปที่ 3.14 ในการศึกษาครั้งนี้ได้กำหนดพารามิเตอร์สำหรับ โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ดังนี้ Epochs = 300, Batch size = 64, Optimization

Adam, Learning rate = 0.0001, Dropout = 0.25, Batch Normalization, Activation function (ReLu - Hidden Layer, Softmax - Output layer) อ้างอิงจากหัวข้อ 2.2.7



รูปที่ 3.13 โครงสร้างแบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 1



รูปที่ 3.14 โครงสร้างแบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2

บทที่ 4
ผลการวิจัยและการอภิปรายผล

จากบทที่ 3 ที่ได้มีการนำเสนอขั้นตอนการดำเนินงานวิจัยและได้นำเสนอดังรูปที่ 3.1 สำหรับบทที่ 4 นี้จะเป็นผลการศึกษาในส่วนต่างๆ ซึ่งประกอบไปด้วย

4.1 ผลลัพธ์ของการรวบรวมข้อมูล

4.2 ผลลัพธ์ของการจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ

4.3 ผลลัพธ์ของการจำแนกทัศนคติของผู้ที่แสดงความคิดเห็นต่อประกัน

4.1 ผลลัพธ์การรวบรวมข้อมูล (Data Collection)

ในการเก็บข้อมูลจากกระทู้ออนไลน์พันทิปในแต่ละขั้นตอนจะมีผลลัพธ์ดังนี้

4.1.1 ผลลัพธ์เก็บข้อมูลที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจง

จากรูปที่ 4.1 จะเห็นได้ว่าผู้วิจัยสามารถเก็บข้อมูลที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจงในกระทู้ออนไลน์ทั้งหมดได้ โดยข้อดีของการที่ออกแบบการเก็บข้อมูลดังรูป 4.1 จะทำให้ผู้วิจัยสามารถที่จะปรับใช้ข้อมูลได้ในอนาคต เช่น การเลือกประเภทของกระทู้ การเลือกช่วงเวลาที่จะทำการเก็บ การเลือกแฮชแท็กที่ผู้วิจัยต้องการจะใช้ การเก็บข้อมูลจำนวนคนที่แสดงความคิดเห็นในแต่ละกระทู้รีวิวที่ทำให้ผู้วิจัยสามารถตรวจสอบจำนวนของความคิดเห็นที่มีได้เบื้องต้น รวมไปถึงสามารถเก็บข้อมูลของผู้ที่ตั้งกระทู้ได้

topic	tag	url	id	type	n_comment	n_vote	user_id	user_name	day	month	year
สงสัยว่า ทำไม เวลาฟันกำไรค	ประกันชีวิตสำหรับผู้	https://pantip.com	40142136	question	2	0	https://pantip.	สมาชิกหมายเลข	20	8	2563
ตอนนี้ยังมีประกัน COVID-19 แบบ เจอแ	ประกันสุขภาพ	https://pantip.com	39985765	question	1	0	https://pantip.	akkhastock	15	6	2563
ซิม AIS ฟรีประกันโควิด อ่ "โคม่ารักษาไม่หายห	ประกันสุขภาพ	https://pantip.com	39974772	question	1	0	https://pantip.	สมาชิกหมายเลข	11	6	2563
ประกันสังคมวินิจฉัยจ่ายเงินเยียวยามาตรา33ยังไง	ประกันสังคม	https://pantip.com	39966941	question	1	0	https://pantip.	Elise	8	6	2563
ประกันไวรัสโควิดของวิริยะ	ประกันสุขภาพ	https://pantip.com	39916555	question	1	0	https://pantip.	สมาชิกหมายเลข	21	5	2563
จนถึง ณ ปัจจุบันนี้มีใคร เคยเคลมประกันโควิด-19 แล้วบ้างมั้ย	ประกันสุขภาพ	https://pantip.co m/topic/3991512 7	39915127	talking	2	0	https://panti p.com/profile /737381	The Mash VR	20	5	2563
ปสก ตรวจโควิด-19 ที่ ออสเตรเลีย สวัสดิการแห่งรัฐวิคตอเรีย ฟรีเซอร์วิสแบบ Drive Thru ยก รพ มาอยู่ที่ลานจอดรถ ห้างใกล้บ้าน	ประกันสุขภาพ	https://pantip.co m/topic/3990272 7	39902727	talking	0	0	https://panti p.com/profile /5441842	สมาชิกหมายเลข 5441842	16	5	2563
มีใครได้ใช้ประกัน Covid-19 บ้างรึยัง จ่า	ประกันสุขภาพ	https://pantip.com	39890122	question	6	0	https://pantip.	สมาชิกหมายเลข	12	5	2563
"ประกันสังคม" กองทุนที่รอวันล่ม สลาย...	ประกันสังคม	https://pantip.co m/topic/3986886 6	39868866	talking	30	26	https://panti p.com/profile /680484	พายุน้ำแข็ง	5	5	2563
ประกันสังคม ต้องจ่ายหรือไม่จ่ายก	ประกันสังคม	https://pantip.com	39859620	question	1	0	https://pantip.	สมาชิกหมายเลข	2	5	2563
ถูกเลิกจ้างช่วงโควิด ท้อแท้ แอบเครียด	ประกันสังคม	https://pantip.com	39854732	question	10	0	https://pantip.	สมาชิกหมายเลข	30	4	2563
การเลือกซื้อ(ต่อ)ประกันภัยรถในช่วงสถานการณ์ COVID-19	ประกันภัยรถ	https://pantip.com	39848510	question	2	0	https://pantip.	สมาชิกหมายเลข	28	4	2563
มีใครได้เงินจากประกันสังคมกรณีว่างงานช่วงโควิดแล	ประกันสังคม	https://pantip.com	39836152	question	1	0	https://pantip.	สมาชิกหมายเลข	24	4	2563
Www sso go th ประกัน สังคมมาตรา 33 เริ่มจ่ายเงินทดแทน กรณีว่างงานเหตุสุดวิสัยแล้ว	ประกันสังคม	https://pantip.co m/topic/3982673 4	39826734	talking	1	0	https://panti p.com/profile /2560032	สมาชิกหมายเลข 2560032	21	4	2563

รูปที่ 4.1 ผลลัพธ์เก็บข้อมูลที่อยู่เว็บแบบสมบูรณ์ที่ใช้ค้นหาหน้าเว็บที่เฉพาะเจาะจง

4.1.2 ผลลัพธ์ของการเก็บข้อมูลกระทู้ประเภทแสดงความคิดเห็น

จากรูปที่ 4.2 จะเห็นได้ว่าผู้วิจัยสามารถเก็บข้อมูลข้อมูลกระทู้ประเภทแสดงความคิดเห็นในกระทู้ออนไลน์ทั้งหมดได้ โดยข้อดีของการที่ออกแบบการเก็บข้อมูลดังรูป 4.2 จะทำให้ผู้วิจัยสามารถตรวจสอบความน่าเชื่อถือของ หัวข้อกระทู้และข้อความที่แสดงความคิดเห็นได้ รวมไปถึงสามารถดูลำดับในการแสดงความคิดเห็นได้

topic_id	topic_title	topic_story	comment_number	comment_story	comment_user_name	comment_day	comment_month	comment_year
39985765	ตอนนี้ยังมีประกัน COVID-19 แบบ เจอแล้วจ่ายเป็นเงินก้อนทันทีไหมครับ	ขอบคุณมากครับ	1	ตอนผมสมัคร ช่วงระบาดแรกๆ หาง่ายจะตาย มีเพียบ จิ้มไปเจ้าไหนก็มี	Disaster Magnet	15	6	2563
39974772	ซิม AIS ฟรีประกันโควิด อ่านเงื่อนไขดีๆแล้ว "โคม่ารักษาไม่หาย หรือตายสถานเดียว" ถึงจะจ่าย	https://www.ais.co.th/soodkoom-insurance / ไม่คิดมากครับ สมกับราคาซิม 1 บาท เป็นข้อมูลเผื่อใครนึกว่าช่วยรักษาโควิดนะครับ ตายก่อนค่อยให้ญาติมารับดังครับ ไม่ได้เอาไว้รักษาตัว	1	สวัสดีค่ะ สมายล์ - คารัตน์มีความยินดีเป็นอย่างยิ่งที่ได้ดูแลข้อมูลคุณสมาชิกหมายเลข 5661398 ในวันนี้นะคะ ขอขอบคุณที่สนใจแคมเปญซิมสุดคุ้ม ราคาพิเศษ 1 บาทค่ะ ช่องทางการซื้อซิมได้ทั้ง Website Become : https://become-ais-family.ais.co . th/find-by-sim-type?categoryId =1047455 AIS Online Store และ Shopee ถึงวันที่ 30 มิถุนายน 2563 ค่ะ พร้อม รับฟรี ประกันโควิด+ประกันอุบัติเหตุ และสามารถติดต่อสอบถามข้อมูลกรมธรรม์เพิ่มเติมได้ที่ บริษัท ทิพยประกันภัย จำกัด (มหาชน) โทร. 1736 ทุกวันตลอด 24 ชั่วโมง รายละเอียด www.ais.co.th/soodkoom-	callcenter@ais.co.th	11	6	2563
39974772	ซิม AIS ฟรีประกันโควิด อ่านเงื่อนไขดีๆแล้ว "โคม่ารักษาไม่หาย หรือตายสถานเดียว" ถึงจะจ่าย	https://www.ais.co.th/soodkoom-insurance / ไม่คิดมากครับ สมกับราคาซิม 1 บาท เป็นข้อมูลเผื่อใครนึกว่าช่วยรักษาโควิดนะครับ ตายก่อนค่อยให้ญาติมารับดังครับ ไม่ได้เอาไว้รักษาตัว	1-1			11	6	2563

รูปที่ 4.2 ผลลัพธ์ของการเก็บข้อมูลกระทู้ประเภทแสดงความคิดเห็น

4.1.3 ผลลัพธ์ของการเก็บข้อมูลกระทู้ประเภทรีวิว

จากรูปที่ 4.3 จะเห็นได้ว่าผู้วิจัยสามารถเก็บข้อมูลกระทู้ประเภทรีวิวในกระทู้ออนไลน์ทั้งหมดได้ โดยข้อดีของการที่ออกแบบการเก็บข้อมูลดังรูป 4.3 ตรวจสอบได้ว่าข้อมูลที่รีวิว เป็นกระทู้รีวิวในแบบใด ได้คะแนนจากผู้ที่แสดงความคิดเห็นเท่าไหร่ โดยข้อมูลในส่วนนี้จะเป็นข้อมูลสำคัญที่ผู้วิจัยจะใช้การวิเคราะห์ข้อมูลต่อไป

BR	CR	SR	Titel	comment	rating	labels
No	No	Yes	[SR] เตือนภัย!! ประกันรถยนต์ อาคาเนย์ ใบเคลมเอาไปใช้ซ่อมรถไม่ได้	เตือนภัย!! ใบเคลมอาคาเนย์ เอาไปใช้ซ่อมรถไม่ได้ ต้องมารอทางอาคาเนย์อนุมัติอีกที ซึ่งไม่มีท่าทีว่าจะอนุมัติได้ง่ายๆ ซึ่งคุยคร่าวๆกับทางอู่ คือทางอาคาเนย์ไม่ค่อยอนุมัติงานซ่อม ทางอู่ก็จะไม่ได้รับเงิน เลย ไม่มีอู่ไหนอยากซ่อมให้กับประกันเจ้านี้ แต่ถ้าเป็นใบเคลมของบริษัทอื่น ทางอู่ก็จะรับซ่อมได้เลย ขอให้เป็นบทเรียนกับผู้ที่กำลังจะตัดสินใจทำประกันรถยนต์ครับ อย่าเห็นแค่ว่าค่าเบี้ยถูก ควรเลือกบริษัทที่ดีด้วย ไม่งั้นคุณก็แค่จ่ายราคาถูกแต่พอถึงเวลาเกิดเหตุเค้ากลับเบี้ยว ไม่อนุมัติซ่อมให้คุณเนะครับ	5	Neg
No	Yes	No	[CR] ล่าสุดพึ่งเจอมาครับประสบการณ์กับเครมประกันมิตรแท้	ผมว่า ปัญหาอยู่ที่ อู่ มากกว่านะ 1. ปกติเอารถเข้าเคลมที่อู่ แล้วเค้าทำเรื่องขออนุมัติ .. คนที่จะติดต่อคุณ จะเป็นอู่ครับ ประกันหมดหน้าที่ในการติดต่อคุณแล้ว 2. คนสั่งอะไหล่ นี่ก็โยนไป โยนมา .. อู่บอกประกัน / ประกันบอกอู่ .. ถ้าเช็คประกันว่าให้อู่สั่ง ก็อู่นั่นแหละ พวกอาจจะหาไม่ได้ แล้วโยนกลับประกันหรือป่าว ? 3. รถอายุขนาดนี้ เค้าจะหามือสอง ของเทียมให้ นะครับ ที่บอกว่า ไม่มียอง เพราะเค้ารอหาพวกนี้อยู่ครับ ส่วนเรื่องเบิกอะไหล่จากศูนย์ปกติอู่ หรือ ประกัน มีร้านอะไหล่ใน คอนแทค เบิกได้อยู่แล้วครับ ผมยังเบิกจากศูนย์ไปให้ร้านข้างนอกทำบ่อยๆ ไม่เคยมีปัญหานะ สั่งเข้าได้ง่าย 4. รถคุณมีประกัน ต่อให้เป็นฝ่ายถูก ก็ควรใช้ประกันคุณ แล้วเดี๋ยวประกันคุณไปเก็บเงินกับฝ่ายผิดเอง แต่อย่าคิดว่า เป็นฝ่ายถูกแล้ว จะเปลี่ยนอะไหล่ยังไงก็ได้นะ ประกันเค้ามีเงื่อนไข ข้อตกลงมาตรฐานกันอยู่ครับ	5	Neg

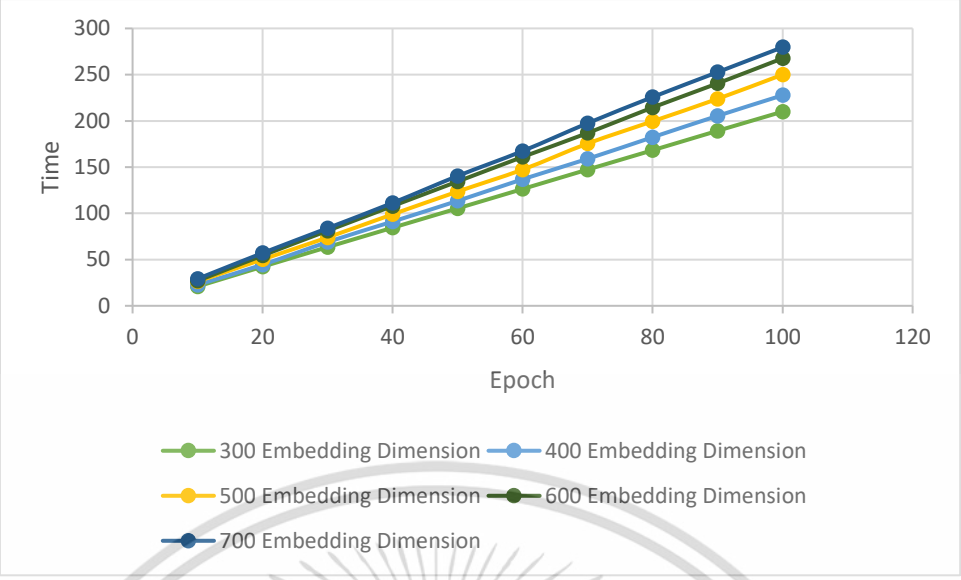
รูปที่ 4.3 ผลลัพธ์ของการเก็บข้อมูลกระทู้ประเภทรีวิว

4.2 ผลลัพธ์ของการจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ

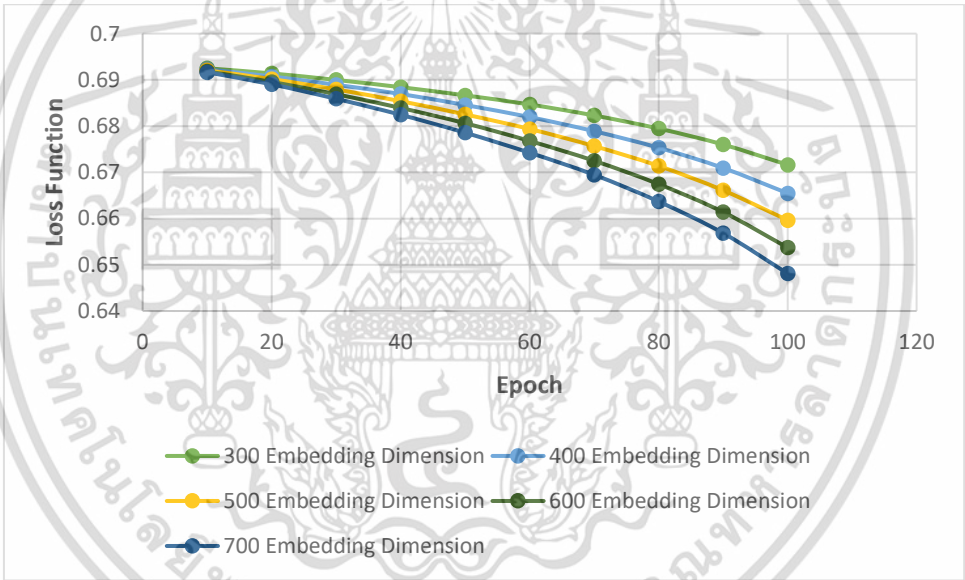
ในส่วนนี้ผู้วิจัยจะแสดงผลลัพธ์ของทั้ง 2 การทดลอง โดยรายละเอียดในแต่ละการทดลอง ผู้วิจัยจะนำเสนอสิ่งที่น่าสนใจสำหรับการทดลอง ผลการทดลอง และคำแนะนำเพิ่มเติมสำหรับงานการทดลองในอนาคต และสรุปภาพรวมของทั้ง 2 การทดลอง โดยมีรายละเอียดดังนี้

4.2.1 ผลลัพธ์ของการทดลองที่ 1

โดยทางผู้วิจัยได้พัฒนาทั้งหมด 50 แบบจำลองสำหรับใช้ในการเปรียบเทียบ เพื่อหาแบบจำลองที่ดีที่สุดที่ทำให้ค่าฟังก์ชันการสูญเสียต่ำที่สุด และใช้เวลาในการพัฒนาแบบจำลองที่เหมาะสม โดยมีผลของการทดลองดังนี้



รูปที่ 4.4 กราฟความสัมพันธ์ของจำนวนรอบที่ใช้กับเวลา



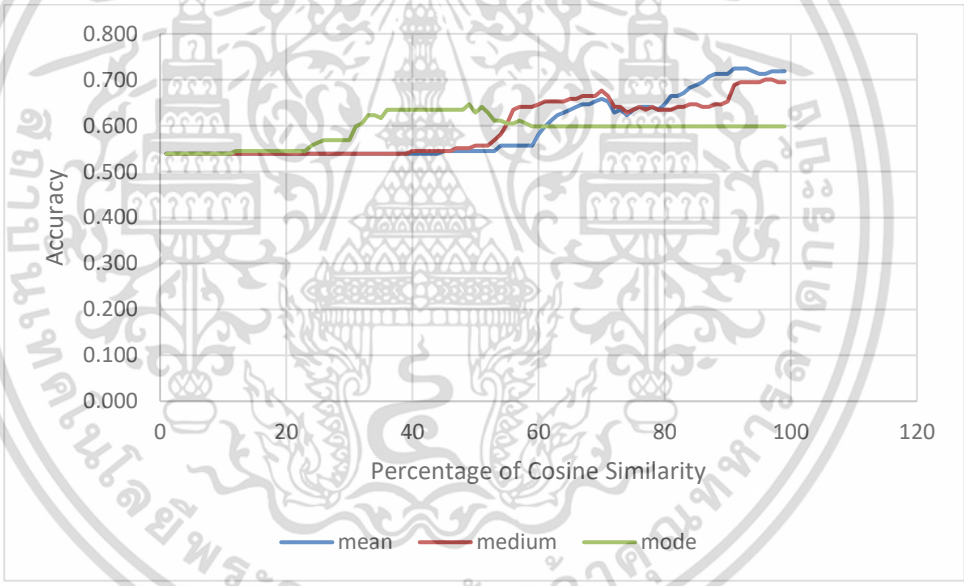
รูปที่ 4.5 กราฟความสัมพันธ์ของจำนวนรอบที่ใช้กับ ฟังก์ชันการสูญเสีย

จากการทดลองจะเห็นได้ว่าจำนวนรอบที่ใช้สอนแบบจำลองที่เพิ่มขึ้นมีผลต่อการลดฟังก์ชันการสูญเสียในทุกการทดลองแต่ถ้านำผลของการทดลองทั้ง 5 มาเปรียบเทียบกันจะเห็นได้ว่าการกำหนด จำนวนรอบที่ใช้สอนแบบจำลอง ที่ไม่เกิน 20 จะให้ค่า ฟังก์ชันการสูญเสีย ไม่แตกต่างกัน ดังรูปที่ 4.5 แต่ถ้าเริ่มเพิ่มจำนวนรอบที่ใช้สอนแบบจำลองขึ้น โดยสังเกตที่จำนวนรอบที่ใช้สอนแบบจำลองเป็น 60 จะเริ่มเห็นความแตกต่างของค่าฟังก์ชันการสูญเสียและจากการทดลองจะเห็นความแตกต่างจากการลดลงของค่าฟังก์ชันการสูญเสียมากสุดที่จำนวนรอบที่ใช้สอนแบบจำลองเป็น 100 และในส่วนของจำนวนมิติจะมีผลต่อการลดลงของฟังก์ชันการสูญเสียเช่นกัน โดยจากการทดลองจะเห็นที่ว่าเมื่อกำหนดมิติ เป็น 700 จะทำให้ฟังก์ชันการสูญเสีย เป็น 0.649 ซึ่งเป็นค่าที่ต่ำที่สุดใน

การทดลองครั้งนี้ โดยที่กราฟสามารถสรุปได้ว่ายิ่งกำหนดมิติมากขึ้นจะทำให้ค่าฟังก์ชันการสูญเสียลดลง แต่เนื่องจากว่าผู้วิจัยไม่สามารถใช้ฟังก์ชันการสูญเสียในการพิจารณาเพียงอย่างเดียวได้ ดังนั้นอีกปัจจัยหนึ่งที่ทำให้มีผลต่อการเลือก แบบจำลอง คือ เวลาที่ใช้ในการพัฒนาแบบจำลอง ถ้าดูจากรูป 4.4 จะเห็นได้ว่ายิ่งจำนวนรอบที่ใช้สอนแบบจำลองสูงขึ้นก็จะยิ่งใช้เวลามากขึ้น แต่สิ่งที่น่าสนใจคือจำนวนของมิติที่มีผลต่อเวลาผู้วิจัยสังเกตว่าที่ 600 และ 700 มิติ ค่อนข้างใช้เวลาที่ใกล้เคียงกันเมื่อมีจำนวนรอบที่ใช้สอนโมเดลที่เท่ากัน ดังนั้นทางผู้วิจัยจึงจะเลือกใช้ค่ามิติเป็น 500 เนื่องจากในงานวิจัย [25] ใช้ในการทดสอบ และได้ข้อสรุปว่าค่ามิติเป็น 500 เป็นค่าที่เหมาะสมสำหรับการทดลอง

4.2.2 ผลลัพธ์ของการทดลองที่ 2

โดยทางผู้วิจัยได้ทำการทดลองเปรียบเทียบทั้งหมด 297 วิธี ที่ใช้ในสำหรับใช้ในการจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพเพื่อหาวิธีการทำเทคนิคการแปลงประโยคเป็นตัวเลขและการใช้เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ที่ดีที่สุดที่ให้ค่าความถูกต้องสูงที่สุดโดยมีผลของการทดลองดังนี้



รูปที่ 4.6 กราฟเปรียบเทียบค่าความถูกต้องของแต่ละวิธีในการทำเทคนิคการแปลงประโยคเป็นตัวเลขในทุกเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์

จากรูปที่ 4.6 จะเห็นได้ว่าวิธีการทำเทคนิคการแปลงประโยคเป็นตัวเลข ทั้ง 3 วิธี ให้ค่าความถูกต้องต่างกันในแต่ละเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ ถ้าดูจาก เส้นสีฟ้าที่แทนค่าเฉลี่ย สีแดงที่แทนค่ากลาง และสีเขียวที่แทนค่าฐานนิยม จะเห็นได้ว่านอกจากที่ให้ความแม่นยำที่ไม่เท่ากันแล้ว จากกราฟนี้ยังสามารถสรุปได้ว่า ในกรณีที่ผู้วิจัยกำหนดค่าเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ที่ไม่สูงมาก ค่าฐานนิยมจะเหมาะสมที่สุด เพราะเนื่องจากว่าถ้าหากมีจำนวนคำในประโยคที่มาก การใช้ค่าฐานนิยมในการพิจารณาจะดีที่สุด แต่ในทางกลับกันจะเห็นได้ว่าค่าตั้งแต่ความแม่นยำที่ 50 ขึ้นไป การใช้ฐาน

นิยมจะให้ค่าที่คงที่ และจากกราฟจะเห็นได้ว่าถ้าค่าเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ของผู้วิจัยอยู่ที่ประมาณ 50 เปอร์เซ็นต์ จะเห็นได้ว่าค่าที่ให้ผลดีที่สุดคือค่ามัทยะฐานจะเหมาะสมที่สุด เนื่องจากการใช้ค่ากลางนั้น การคิดขนาดของเวกเตอร์ก็จะยิ่งมีความใกล้เคียงมากกับค่ากลางมากที่สุด แต่ในกรณีถ้าผู้วิจัยต้องการเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ที่สูง การใช้ค่าเฉลี่ยจะเป็นวิธีที่ดีที่สุด

ตารางที่ 4.1 เปรียบเทียบวิธีการทำใช้ เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ กับวิธีการปกติ

Sentence Embedding Methods	Original Classify	Percentage of Cosine Similarity Classify	Percentage of Cosine Similarity
Mean	0.539	0.725	91
Medium	0.539	0.701	96
Mode	0.539	0.647	49

จากตาราง 4.1 จะเห็นว่าวิธีการจำแนกแบบปรกติให้ค่าความถูกต้องที่เท่ากันในทุกเทคนิคการแปลงประโยคเป็นตัวเลข เนื่องจากข้อมูลของผู้วิจัยมีจำนวนน้อย ผู้วิจัยจึงนำเสนอวิธีการจำแนกด้วยเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ จะเห็นได้ว่าในทุกเทคนิคการแปลงประโยคเป็นตัวเลข วิธีการให้ค่าความถูกต้องที่แตกต่างกันละสูงกว่าวิธีการจำแนกแบบปรกติ ในทุกเทคนิคการแปลงประโยคเป็นตัวเลข วิธีการ ซึ่งแสดงให้เห็นว่า วิธีการที่ทางผู้วิจัยนำเสนอสามารถพัฒนา ประสิทธิภาพของงานให้ดีขึ้น โดยพัฒนาจาก ค่าความถูกต้อง เป็น 0.539% เพิ่มขึ้นเป็น ค่าความถูกต้อง เป็น 0.725% ด้วยวิธีการทำ เทคนิคการแปลงประโยคเป็นตัวเลข ด้วย ค่าเฉลี่ย และใช้เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ ที่ 91% ในการ การจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ

4.3 ผลลัพธ์ของการจำแนกทัศนคติของผู้ที่แสดงความคิดเห็นต่อประกัน

จากการสร้างเทคนิคการเรียนรู้เชิงลึกเพื่อวิเคราะห์ความพึงพอใจของผู้ใช้งานต่อประกันภัยจากกระทู้ออนไลน์พันทิป ในทั้ง 3 แบบนั้นได้ค่าความถูกต้องดังตารางที่ 4.2 จะเห็นได้ว่าจากแบบจำลองทั้งหมด มี 4 แบบจำลอง ที่ให้ค่าด้วยความถูกต้องสูงสุดที่ 0.85 คือ แบบจำลอง โครงข่ายประสาทแบบคอนโวลูชัน ชุดที่ 1 แบบจำลองโครงข่ายประสาทแบบคอนโวลูชัน ชุดที่ 2 แบบจำลองโครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2 แบบจำลอง โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง ชุดที่ 1 ซึ่งจากค่าความถูกต้องยังไม่สามารถสรุปได้ว่า

วิธีการที่ทางผู้วิจัยนำเสนอสามารถพัฒนาแบบจำลองได้หรือไม่ ทางผู้วิจัยจึงพิจารณาจากค่าฟังก์ชันการสูญเสียซึ่งเป็นค่าความผิดพลาดดังตารางที่ 4.3

ตารางที่ 4.2 เปรียบเทียบค่าความแม่นยำในการวิเคราะห์ค่าความรู้สึก

Model	Version	Precision POS	Precision NEG	Recall POS	Recall NEG	F1 Score POS	F1 Score NEG	Accuracy
CNN	1	1.00	0.81	0.60	1.00	0.75	0.89	0.85
	2	1.00	0.81	0.60	1.00	0.75	0.89	0.85
GRU	1	0.60	0.76	0.60	0.76	0.60	0.76	0.70
	2	0.75	0.93	0.90	0.82	0.82	0.88	0.85
Bi-LSTM	1	0.88	0.84	0.70	0.94	0.78	0.89	0.85
	2	0.62	0.86	0.80	0.71	0.70	0.77	0.74

ตารางที่ 4.3 เปรียบเทียบค่าฟังก์ชันการสูญเสียในการวิเคราะห์ค่าความรู้สึก

Model	Loss function	
	Version 1	Version 2
CNN	0.61	0.60
GRU	0.66	0.60
Bi-LSTM	0.62	0.51

จากตารางที่ 4.3 จะเห็นได้ว่า การเปรียบเทียบประสิทธิภาพแบบของแบบจำลองจากค่าฟังก์ชันการสูญเสีย แบบจำลองที่ทางผู้วิจัยนำเสนอ ชุดที่ 2 ให้ค่าผิดพลาดน้อยกว่า แบบจำลองปกติ ชุดที่ 1 ในทุกแบบจำลองจากการเปรียบเทียบค่าความถูกต้องแบบจำลองทั้ง 4 แบบจะให้ค่าความแม่นยำเท่ากัน แต่แบบจำลองที่ทางผู้วิจัยนำเสนอ สามารถลดค่าความผิดพลาดต่ำสุดที่ 0.51 คือ แบบจำลองโครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง ชุดที่ 2 แต่แบบจำลองนี้ทางผู้วิจัยจะไม่เลือกใช้ เพราะให้ค่าความถูกต้องเพียงแค่ 0.74 ซึ่งน้อยกว่า แบบจำลองทั้ง 4 อยู่ถึง 11% ทางผู้วิจัยจึงเลือกพิจารณาที่ค่าความผิดพลาดที่ 0.60 ซึ่งมีทั้งหมด 2 แบบจำลองคือ แบบจำลอง โครงข่ายประสาทแบบคอนโวลูชัน ชุดที่ 2 และแบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2

เนื่องจากงานวิจัยนี้เป็นงานวิจัยที่จะนำไปปรับใช้กับธุรกิจ โดยส่วนมากจะใช้ค่าความระลึก ซึ่งทางผู้วิจัยจะใช้ค่าค่าความระลึกเชิงบวก เพราะในเชิงธุรกิจถ้าค่าค่าความระลึกเชิงบวกจะเป็นค่าความแม่นยำที่จะทำให้บริษัทมั่นใจได้ว่า หากมีค่าสูงหมายความว่า บริษัทจะไม่พลาดโอกาสที่จะเชิญชวนคนที่อยากได้สินค้าในกลุ่มเป้าหมายเข้ามาเป็นลูกค้า ในทางกลับกันหากมีค่าน้อย อาจจะทำให้บริษัทไม่ได้ชวนคนที่อยากได้สินค้าในกลุ่มเป้าหมายเข้ามาเป็นลูกค้า หรืออาจทำให้คู่แข่งได้ลูกค้ากลุ่มนี้ไปแทน เพราะเป็นกลุ่มลูกค้ากลุ่มนี้บริษัทสามารถนำไปต่อยอด โดยการเสนอขายประกันภัยเพิ่มเติมได้ จากตารางที่ 1 จะเห็นได้ว่า แบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2 ใช้ค่า ค่าความระลึกเชิงบวก ที่ 0.90 ซึ่งสูงกว่าแบบจำลอง โครงข่ายประสาทแบบคอนโวลูชัน ชุดที่ 2 อยู่ถึง 30% และเป็นค่า ค่าความระลึกเชิงบวก ที่สูงที่สุดในทุกแบบจำลอง แต่ในเชิงธุรกิจจะมีอีกค่าหนึ่งที่นิยมใช้คือ ความแม่นยำเชิงบวก นั้นหมายความว่ายิ่งมีค่ามากงบประมาณที่ผู้วิจัยใช้ในการเชิญชวนลูกค้าจะยิ่งคุ้มค่า ไม่สิ้นเปลือง ในทางกลับกันหากความแม่นยำต่ำบริษัทอาจจะสิ้นเปลืองเงินไปกับการเชิญชวนคนที่จะไม่เข้ามาเป็นลูกค้า แต่เนื่องจากในการทดลองนี้อาจจะยังไม่ได้นำค่าความแม่นยำเชิงบวก เนื่องจากทางผู้เชี่ยวชาญมองว่าการใช้ค่า ความระลึกเชิงบวก เหมาะสมกับงานวิจัยนี้มากกว่า

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในงานวิจัยนี้สามารถสรุป ผลการวิจัย และข้อเสนอแนะต่างๆ ได้ดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้แบ่งออกเป็น 2 ส่วน ส่วนที่ 1 คือการจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ และส่วนที่ 2 คือการจำแนกทัศนคติของผู้ที่แสดงความคิดเห็นต่อประกัน โดยทั้ง 2 ส่วนนี้ใช้ข้อมูลจากกระทู้ออนไลน์พันทิป และใช้วิธีการเตรียมข้อมูลเดียวกัน โดยมีรายละเอียดดังนี้

ส่วนที่ 1 ในการหาแบบจำลองที่ทำให้ค่าฟังก์ชันการสูญเสียต่ำสุดและใช้ที่เวลาเหมาะสม จากการทดลองทั้งหมด 50 แบบจำลอง โดยชุดพารามิเตอร์ของแบบจำลองที่ดีที่สุด จะเป็นกำหนดค่าระยะทางสูงสุดจากคำที่กำลังใช้ฝึกฝนเป็น 2 จำนวนรอบที่ใช้สอนโมเดลเป็น 100 และ จำนวนมิติเป็น 500 โดยให้ค่า ฟังก์ชันการสูญเสีย เป็น 0.659 และจากการทดลองเปรียบเทียบวิธีการทำเทคนิคการแปลงประโยคเป็นตัวเลข ทั้ง 3 วิธีได้แก่ ค่าเฉลี่ย ค่ากลาง และ ค่าฐานนิยม ร่วมกับการใช้เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ ตั้งแต่ เปอร์เซ็นต์ที่ 1 จนถึง 99 จากการทดลองทั้งหมด 297 การทดลอง โดยงานวิจัยนี้มีข้อมูลในการพัฒนาแบบจำลองที่จำกัด ซึ่งจะเห็นว่าวิธีการการใช้ เทคนิคการแปลงประโยคเป็นตัวเลข อย่างเดียวทั้ง 3 วิธีให้ค่าความถูกต้องที่เท่ากัน แต่จากวิธีการที่ทางผู้วิจัยนำเสนอในการใช้ เทคนิคการแปลงประโยคเป็นตัวเลขร่วมเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ สามารถพัฒนาประสิทธิภาพของงานให้ดีขึ้น โดยพัฒนาจาก ค่าความถูกต้อง เป็น 0.539% เพิ่มขึ้นเป็น ค่าความถูกต้อง เป็น 0.725% ด้วยวิธีการทำเทคนิคการแปลงประโยคเป็นตัวเลขด้วย ค่าเฉลี่ยและใช้เปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ ที่ 91% ในการการจำแนกกลุ่มของคนที่มีแนวโน้มที่รักสุขภาพ โดยในส่วนนี้จะสอดคล้องกับวัตถุประสงค์ในข้อที่ 1 และข้อที่ 3

ส่วนที่ 2 งานวิจัยนี้จึงขอนำเสนอวิธีการรับรู้เสียงของลูกค้าที่อยู่ในรูปของข้อมูลที่เป็นข้อความ ซึ่งในงานวิจัยนี้คือความพึงพอใจหรือดาวที่ผู้ตั้งกระทู้ให้คะแนนในแต่ละกระทู้รีวิว โดยในงานวิจัยนี้จะใช้ข้อมูลความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยในงานวิจัยนี้ผู้วิจัยจะใช้การวิเคราะห์ความรู้สึกเพื่อวิเคราะห์ความรู้สึกจากข้อความว่าเป็นเชิงบวก หรือเชิงลบ โดยใช้แบบจำลองแบบมีผู้สอนจะใช้เทคนิคการจำแนก ทั้ง 3 เทคนิคต่อไปนี้ โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู โครงข่ายประสาทเทียมแบบความจำระยะสั้นแบบยาวสองทิศทาง และ โครงข่ายประสาทแบบคอนโวลูชัน ซึ่งงานวิจัยนี้ในแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด โดยแบบจำลองชุดที่ 1 จะ

เป็นการสร้างแบบจำลองแบบโดยใช้ตัวแปรที่ผ่านกระบวนการ การสกัดคุณลักษณะ ส่วนในแบบจำลองชุดที่ 2 จะใช้ตัวแปรที่ผ่านกระบวนการการสกัดคุณลักษณะ และเพิ่มตัวแปรคือประเภทของกระทู้รีวิวในกระทู้ออนไลน์พันทิป ซึ่งตัวแปรนี้มีอยู่ในทุกกระทู้รีวิวในรวมแล้วทั้งสิ้น 6 แบบจำลอง จะเห็นได้ว่าแบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2 เป็นแบบจำลองที่ดีที่สุด ซึ่งให้ค่าความถูกต้องสูงที่สุดที่ 0.85 และให้ค่าความผิดพลาดที่ 0.60 ถึงแม้ว่าจะไม่ได้เป็นค่าที่น้อยที่สุดในทุกแบบจำลอง แต่เมื่อเทียบกับ แบบจำลองที่ให้ค่าความถูกต้องที่เท่ากันถือว่าแบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2 ให้ค่าต่ำที่สุด และให้ค่าค่าความระลึกเชิงบวกซึ่งจะเป็นค่าความแม่นยำที่จะทำให้บริษัทมั่นใจได้ว่า หากมีค่าสูงหมายความว่าบริษัทจะไม่พลาดโอกาสที่จะเชิญชวนคนที่อยากได้สินค้าในกลุ่มเป้าหมายเข้ามาเป็นลูกค้า สูงสุดที่ 0.90 โดยบริษัทประกันสามารถนำแบบจำลอง โครงข่ายประสาทเทียมหน่วยเวียนกลับแบบมีประตู ชุดที่ 2 ที่มีการใส่ตัวแปรประเภทของกระทู้รีวิวในกระทู้ออนไลน์พันทิปเข้าไปเพิ่มในแบบจำลอง โดยสามารถแนะนำกลุ่มลูกค้าที่มีความคิดเห็นเชิงบวกต่อการทำประกันภัย เพื่อที่บริษัทจะสามารถเสนอขายผลิตภัณฑ์อื่นๆของทางบริษัท ผ่านทางช่องทางกระทู้ออนไลน์พันทิป เพราะสามารถระบุไปยังตัวตนของผู้ที่แสดงความคิดเห็น ซึ่งเพิ่มช่องทางการขายให้กับบริษัทประกันได้ โดยในส่วนนี้จะสอดคล้องกับวัตถุประสงค์ในข้อที่ 2

5.2 ข้อเสนอแนะ

1) ในด้านของข้อมูลเนื่องจากข้อมูลที่ผู้วิจัยต้องการใช้มีคำความเฉพาะเจาะจงไปที่กระทู้ที่เกี่ยวกับประกันภัยและโรคติดเชื้อไวรัสโคโรนา 2019 ทำให้ข้อมูลที่ทางผู้วิจัยมีในปริมาณที่ไม่มาก แต่สิ่งที่น่าสนใจสำหรับการเก็บข้อมูลในกระทู้ออนไลน์พันทิปคือ ข้อมูลสามารถเก็บได้ง่าย แต่ไม่ติดในเรื่องของสิทธิในการเข้าถึง ที่สำคัญแหล่งข้อมูลที่คนไทยให้ความสำคัญ ดังนั้นจึงเหมาะสมกับการที่จะนำไปต่อยอดทางธุรกิจพัฒนาแบบจำลองที่สามารถจำแนกหากลุ่มลูกค้าที่มีทัศนคติที่เป็นเชิงบวกต่อบริษัทประกัน เพื่อนำไปพัฒนาผลิตภัณฑ์ให้ดีขึ้น
2) ในด้านของวิธีการเตรียมข้อมูลในกระบวนการทำความสะอาดข้อมูล ในงานวิจัยนี้ได้มีการลบอิโมจิออกไปก่อนการวิเคราะห์ แต่เนื่องจากมีบางงานวิจัยพบว่า การใช้ อิโมจิเป็นปัจจัยที่สำคัญในการวิเคราะห์ความรู้สึก เช่นการรีวิวสินค้าต่างๆ ดังนั้นงานวิจัยในอนาคต สามารถใช้ อิโมจิ ในพัฒนาเทคนิคการวิเคราะห์ความรู้สึกได้

3) ในด้านของแบบจำลองการวิเคราะห์บทวิจารณ์ปัจจุบันได้มีเทคโนโลยีการวิเคราะห์เชิงลึกอีกหลายแบบ ที่ช่วยทำให้คอมพิวเตอร์สามารถเรียนรู้คำพ้องความหมาย (Synonyme) ได้ซึ่งจะช่วยทำให้โมเดลมีความฉลาดในการแยกแยะ และจับกลุ่มคำศัพท์ที่มีความหมายคล้ายคลึงกันให้เป็นคำเดียวกันได้ดียิ่งขึ้น

4) ในด้านของการวัดผลการทดลองยังมีตัวชี้วัดแบบอื่นที่เหมาะสมกับภาคธุรกิจเช่นค่าความแม่นยำเชิงบวก นั้นหมายความว่ายิ่งมีค่ามากงบประมาณที่ผู้วิจัยใช้ในการเชิญชวนลูกค้าจะยิ่งคุ้มค่า ไม่สิ้นเปลือง ในทางกลับกันหากความแม่นยำต่ำบริษัทอาจจะสิ้นเปลืองเงินไปกับการเชิญชวนคนที่จะไม่เข้ามาเป็นลูกค้า แต่เนื่องจากในการทดลองนี้อาจจะยังไม่ได้นำค่าความแม่นยำเชิงบวก เนื่องจากทางผู้เขี่ยวชาญมองว่าการใช้ค่าความระลึกเชิงบวก เหมาะสมกับงานวิจัยนี้มากกว่า

ข้อแนะนำสำหรับงานวิจัยในอนาคต เนื่องจากข้อมูลจาก กระทู้ออนไลน์พันทิป เป็นข้อมูลที่มีความหลากหลายของข้อมูล ซึ่งสิ่งที่จะเป็นตัวชี้วัดของงานวิจัยอีกด้านหนึ่งก็คือธุรกิจ เพราะในบางครั้งวิธีการที่ทำให้ ค่าความถูกต้องสูงสุด อาจไม่ได้ถูกนำมาใช้เสมอไป ถ้าดูจากการทดลองการเลือกเปอร์เซ็นต์ความคล้ายคลึงของโคไซน์ อาจจะไม่จำเป็นต้องเลือกที่ 91% อาจจะเลือกน้อยกว่านี้ได้ถ้าธุรกิจสามารถยอมรับได้ อีกด้านคือเรื่องของข้อมูลธุรกิจจะเชื่อถือข้อมูลที่มีคุณภาพจากแหล่งข้อมูลที่น่าเชื่อถือ ดังนั้นการใช้ข้อมูลจริงที่เป็นข้อมูลของบริษัทจะทำให้งานดูมีความน่าเชื่อถือมากกว่าการรวบรวมข้อมูลจากสังคมออนไลน์

5.3 ข้อจำกัดของงานวิจัย

ข้อจำกัดที่เกิดขึ้นกับงานวิจัยนี้มีดังนี้

1) จำนวนของข้อมูลความคิดเห็นที่มีปริมาณน้อยและมีความซ้ำซ้อนกัน
2) การทำความสะอาดข้อมูลทำให้ความสมบูรณ์ของข้อมูลบางส่วนหายไปทำให้ยากต่อการวิเคราะห์
3) แบบจำลองที่ใช้เปรียบเทียบมีจำนวนไม่มาก
4) ตัวชี้วัดทางสถิติบางค่ายังไม่สามารถสรุปผลได้
5) คะแนนเกิดที่เกิดจากการแสดงความคิดเห็นเกิดจากความชอบส่วนบุคคลซึ่งผู้ใช้งานมีความยากง่ายในการให้แตกต่างกัน

เอกสารอ้างอิง

[1] Goularas, D., & Kamis, S. (2019). Evaluation of deep learning techniques in sentiment analysis from Twitter data. 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML). https://doi.org/10.1109/deep-ml.2019.00011

[2] Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. Mining Text Data, 415-463. https://doi.org/10.1007/978-1-4614-3223-4_13

[3] Yang, S., Yoo, S., & Jeong, O. (2020). DeNERT-KG: Named entity and relation extraction model using DQN, knowledge graph, and BERT. Applied Sciences, 10(18), 6429. https://doi.org/10.3390/app10186429

[4] Jang, B., Kim, M., Harerimana, G., Kang, S., & Kim, J. W. (2020). Bi-LSTM model to increase accuracy in text classification: Combining Word2vec CNN and attention mechanism. Applied Sciences, 10(17), 5841. https://doi.org/10.3390/app10175841

[5] Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.

[6] Mikolov, Tomas, et al. “Efficient estimation of word representations in vector space.” arXiv preprint arXiv:1301.3781 (2013).

[7] Mikolov, Tomas, et al. “Distributed representations of words and phrases and their compositionality.” arXiv preprint arXiv:1310.4546 (2013).

[8] A. Poornima และ K. Sathiya Priya “A comparative Sentiment Analysis Of Sentence Embedding Using Machine Learning Techniques” in 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020

[9] Gunawan, D., Sembiring, C. A., & Budiman, M. A. (2018). The implementation of cosine similarity to calculate text relevance between two documents. Journal of Physics: Conference Series, 978, 012120. https://doi.org/10.1088/1742-6596/978/1/012120

เอกสารอ้างอิง (ต่อ)

[10] Otter, D. W., Medina, J. R., & Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 32(2), 604-624. https://doi.org/10.1109/tnnls.2020.2979670

[11] Liang Yao, Chengsheng Mao และ et al., "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks" Yao et al. BMC Medical Informatics and Decision Making 2019, 19(Suppl 3):71 https://doi.org/10.1186/s12911-019-0781-4

[12] Sirinart Tangruamsub, "Long Short-Term Memory (LSTM)," Available https://medium.com

[13] [Sentiment Analysis of Comment Texts based on BiLSTM., 2017, 19] (Guixizn Xu et al., 2017)

[14] Li Qing, Weng Linghong และ et al,. "A Novel Neural Network-Based Method for Medical Text Classification" www.mdpi.com/journal/futureinternet Future Internet 2019, 11, 255; doi:10.3390/fi11120255

[15] Tatsuya Iwasa and Kenji Kita. 1995. Error Correction of Speech Recognition Outputs Using Generalized LR Parsing and Confusion Matrix. In ROCLING 1995 Poster Papers, pages 101–110, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

[16] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, 240-248. https://doi.org/10.1007/978-3-319-67558-9_28

[17] Francisco Regis Vieira Alves และ Reneta Passos Machado Vieira. "The Newton Fractal's Lonado Sequence Study with the Google Colab" International Electronic Journal of Mathematics Education, 2020 - Volume 15 Issue 2, Article No: em0575 https://doi.org/10.29333/iejme/6440

[18] G. Van Rossum (Guido). 1995. "Python Reference Manual" CWI

เอกสารอ้างอิง (ต่อ)

[19] Parinya Sangauansat. “Paragraph2Vex-based Sentiment Analysis on Social Media for Business in thailand” in 2016 8th International Conference on Knowledge and Smart Technology (KST). 2016. DOI: 10.1109/KST.2016.7440526

[20] Kuhamanee et al. 2017. “Sentiment Analysis of Foreign Tourists to Bangkok using Data Mining through Online Social Network” in 2017 IEEE 15th International Conference on Industrial Informatics (INDIN). 2016. DOI: 10.1109/INDIN/.2017.8104921

[21] Viriya Taecharungroj and Boonyanit Mathayomchan. 2019. “Analysing Trip Advisor Reviews of Tourist Attractions in Phuket, Thailand” Tourism Management – Volume 75 pages 550-568

[22] Jian-Wu Bi, Yang Liu, Zhi-Ping Fan, and Jin Zhang. “Exploring Asymmetric Effects of Attribute Performance on Customer Satisfaction in the Hotel Industry” Tourism Management – Volume 77, 104006

[23] Mahmoud, A., Zrigui, M. Sentence Embedding and Convolutional Neural Network for Semantic Textual Similarity Detection in Arabic Language. Arab J Sci Eng 44, 9263–9274 (2019). https://doi.org/10.1007/s13369-019-04039-7

[24] Yue, P., Wang, J., & Zhang, X. (2019). YNU-HPCC at semeval-2019 task 9: Using a BERT and CNN-bilstm-GRU model for suggestion mining. Proceedings of the 13th International Workshop on Semantic Evaluation. https://doi.org/10.18653/v1/s19-2224

[25] Poornima, A., & Priya, K. S. (2020). A comparative sentiment analysis of sentence embedding using machine learning techniques. 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). https://doi.org/10.1109/icaccs48705.2020.9074312

ภาคผนวก

...

Proceedings of the
26th Annual Meeting in Mathematics (AMM 2022)
School of Mathematics, Institute of Science
Suranaree University of Technology, Thailand

การวิเคราะห์ความรู้สึกที่มีต่อการทำประกันภัย ด้วยเทคนิคการเรียนรู้เชิงลึก กรณีศึกษา กระทู้ออนไลน์พันทิป

คงภพ ไชยศร[1,†,‡] บุษยมาส พิมพ์พรรณชาติ[2] และ งามเฉิด ด่านพัฒนามงคล[3]

[1] ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง 10520

[2,3] ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง 10520

บทคัดย่อ

ในปัจจุบันข้อมูลความคิดเห็นในสังคมออนไลน์ (Social Media) กำลังได้รับความนิยมและสามารถเข้าถึงได้ง่าย โดยข้อมูลดังกล่าวเป็นประโยชน์ต่อธุรกิจ ในงานวิจัยนี้จึงขอนำเสนอวิธีการรับรู้เสียงของลูกค้า (Voice of Customer: VOC) ที่อยู่ในรูปของข้อมูลที่เป็นข้อความจากกระทู้ออนไลน์พันทิป ซึ่งได้รับความนิยมอย่างมาก โดยงานวิจัยนี้ VOC คือความพึงพอใจหรือดาวที่ผู้ตั้งกระทู้ให้คะแนนในแต่ละกระทู้รีวิว และใช้ข้อมูลจากบทวิจารณ์และการเสนอความคิดเห็นทั้งหมด 276 กระทู้ โดยในงานวิจัยนี้เราจะใช้การวิเคราะห์ความรู้สึก (Sentiment analysis) ซึ่งนิยมใช้เทคนิคการเรียนรู้เชิงลึก ซึ่งงานวิจัยนี้ใช้แบบจำลองทั้งหมด 3 แบบคือ GRU, Bi- LSTM และ CNN ซึ่งแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด โดยชุดที่ 1 จะใช้ตัวแปรที่ผ่านกระบวนการสกัดคุณลักษณะ (Feature Extraction) เพียงอย่างเดียว และชุดที่ 2 จะเพิ่มตัวแปรที่ประเภทของกระทู้รีวิว เพื่อเปรียบเทียบแบบจำลองชุดที่ 2 ที่ทางผู้วิจัยพัฒนา สามารถเพิ่มประสิทธิภาพให้แบบจำลองได้หรือไม่ โดยทำการทดลองทั้งหมด 6 แบบจำลอง จากการทดลองสามารถสรุปได้ว่าแบบจำลอง GRU ชุดที่ 2 เป็นแบบจำลองที่ดีที่สุด เนื่องจากให้ค่าความถูกต้อง (Accuracy) สูงที่สุดที่ 0.85 และให้ค่าฟังก์ชันการสูญเสีย (Loss function) ที่ 0.60 ถึงแม้ว่าจะไม่ได้เป็นค่าที่น้อยที่สุดในทุกแบบจำลอง แต่เมื่อเทียบกับ แบบจำลองที่ให้ค่าความถูกต้อง ที่เท่ากันถือว่าแบบจำลอง GRU ชุดที่ 2 ให้ค่าต่ำที่สุด และให้ค่าความระลึก (Recall POS) สูงสุดที่ 0.90

คำสำคัญ: ประกันภัย, การวิเคราะห์ความรู้สึก, กระทู้ออนไลน์พันทิป, เทคนิคการเรียนรู้เชิงลึก

2020 MSC: 68M10

*งานวิจัยเรื่องนี้ได้รับทุนสนับสนุนจาก สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

†ผู้นำเสนอ คงภพ ไชยศร ‡ผู้แต่งหลัก คงภพ ไชยศร

อีเมล: 63605058@kmitl.ac.th, busayamas.pi@kmitl.ac.th , ngarmcherd.da@kmitl.ac.th

...

1 บทนำ

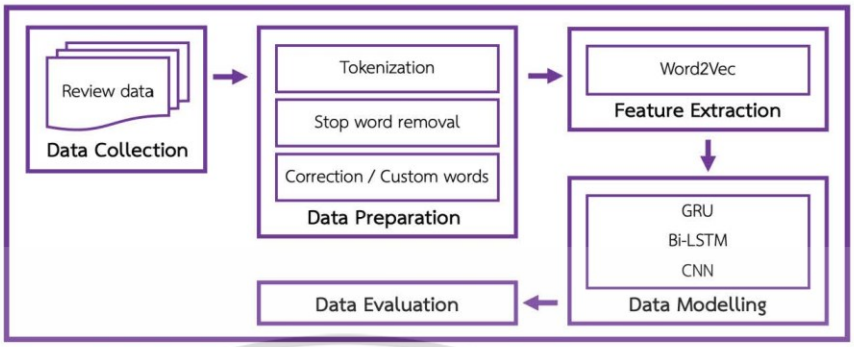
ในปัจจุบันทั่วโลกกำลังเจอวิกฤตจากการระบาดของโรค COVID-19 [1] และจากการวิเคราะห์ความต้องการของลูกค้าในปัจจุบันพบว่า ถึงแม้จะมีธุรกิจที่ได้รับผลกระทบเป็นจำนวนมากก็ตามแต่ในทางกลับกันก็ยังมีบางธุรกิจที่สามารถปรับตัวได้ อีกทั้งยังมีธุรกิจที่เกิดขึ้นใหม่ และยังสามารถประสบความสำเร็จได้เป็นอย่างมากอีกด้วย และเพื่อพัฒนาผลิตภัณฑ์ใหม่ให้สอดคล้องกับสถานการณ์ที่เกิดขึ้นนี้ โดยมุ่งเน้นไปที่ธุรกิจที่ได้รับผลกระทบเยอะและมีความจำเป็นต้องปรับตัวให้เท่าทันกับวิกฤตที่เกิดขึ้นนี้นั่นคือบริษัทประกัน [2] ดังนั้นวัตถุประสงค์ของงานวิจัยนี้ จึงให้ความสำคัญกับความคิดเห็นที่เกิดจากธุรกิจประกันภัยโดยครอบคลุมในส่วนของ การประกันชีวิต (Life Insurance) และการประกันวินาศภัย (Non-Life Insurance) ซึ่งจากสถานการณ์ในปัจจุบันพบว่า Social Media เป็นแพลตฟอร์มที่ได้รับความนิยมเป็นอย่างมาก โดยข้อมูลใน Social Media เป็นข้อมูลที่สามารถนำมาใช้ประโยชน์ได้ เพราะมีช่องทางในการแสดงความคิดเห็นค่อนข้างมาก ดังนั้นทางผู้วิจัยจึงเลือกแพลตฟอร์มที่เป็นที่นิยมในประเทศไทย คือ กระทู้ออนไลน์พันทิป เนื่องจากสามารถระบุไปยังตัวตนของผู้ที่แสดงความคิดเห็น เพื่อเพิ่มช่องทางการขายให้กับธุรกิจได้ ซึ่งข้อมูลที่อยู่ใน Social Media เป็น ข้อมูลตัวอักษร (Textual Data) จะมีการนำองค์ความรู้ด้าน Natural Language Processing มาประยุกต์ใช้ [3] ในปัจจุบันข้อมูลความคิดเห็นได้มีการประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) มาช่วยในการวิเคราะห์ข้อมูลกันอย่างแพร่หลาย [4]

จากเหตุผลที่กล่าวไว้ข้างต้นในงานวิจัยนี้จึงมีความประสงค์ในการนำเสนอวิธีการรับรู้เสียงของลูกค้า (Voice of Customer: VOC) ที่อยู่ในรูปของข้อมูลที่เป็นข้อความซึ่ง VOC [5] ในงานวิจัยนี้คือความพึงพอใจหรือตาว ในกระทู้ประเภทรีวิวที่ผู้ตั้งกระทู้ให้คะแนนในแต่ละกระทู้รีวิว โดยในงานวิจัยนี้จะใช้ข้อมูลจากบทวิจารณ์และการเสนอความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยในส่วนของการวิเคราะห์ข้อมูลมีเป้าหมายในการวิเคราะห์ความรู้สึก (Sentiment analysis) เพื่อดึงความรู้สึกจากข้อความว่าเป็นไปในเชิงใด โดยใช้แบบจำลองแบบมีผู้สอนจะใช้เทคนิคการจำแนก (Classification) ทั้ง 3 เทคนิคต่อไปนี้ Gated Recurrent Unit (GRU), Bi-Long Short-Term Memory (Bi-LSTM) และ Convolutional Neural Networks (CNN) ซึ่งแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด โดยชุดที่ 1 จะใช้ตัวแปรที่ผ่านกระบวนการสกัดคุณลักษณะเพียงอย่างเดียว และชุดที่ 2 จะเพิ่มตัวแปรที่ประเภทของกระทู้รีวิว เพื่อเปรียบเทียบระหว่างตัวแบบมาใช้ในการจำแนกแสดงความคิดเห็นว่าเป็นเชิงบวก (Positive: POS) หรือเชิงลบ (Negative: NEG) เพื่อให้บริษัทประกันสามารถกำหนดเป้าหมายและสร้างความสัมพันธ์กับลูกค้าในกลุ่มนี้ได้และสามารถใช้ประโยชน์จากข้อมูลที่ได้นี้ไปปรับใช้ให้เกิดประโยชน์ในการปรับรูปแบบการประกันภัยให้สอดคล้องกับสถานการณ์ที่เกิดขึ้นในปัจจุบันได้อย่างมีประสิทธิภาพ

2 วิธีการดำเนินงาน

ในงานวิจัยนี้มุ่งเน้นในการศึกษาการนำการวิเคราะห์ความรู้สึกแบบ Gated Recurrent Unit (GRU), Bi-Long Short-Term Memory (Bi-LSTM) และ Convolutional Neural Networks (CNN) ไปประยุกต์ใช้ในการสร้างแบบจำลองการทำนายความรู้สึกจากข้อความในบทวิจารณ์และการเสนอความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยมีวิธีการตาม รูปที่ 1 โดยแบ่งออกเป็น 5 ขั้นตอนได้แก่ การรวบรวมข้อมูล การเตรียมข้อมูล การแปลงคำให้เป็นตัวเลข การพัฒนาแบบจำลอง และการวัดผลการทดลอง

...



รูปที่1 ขั้นตอนการดำเนินงานของระบบ

2.1 การเก็บรวบรวมข้อมูล (Data Collection)

การศึกษาครั้งนี้รวบรวมข้อมูลข้อความแสดงความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยรายละเอียดของข้อมูลเราจะใช้ข้อมูลของผู้ที่แสดงความคิดเห็นประเภทรีวิวที่เกี่ยวข้องกับประกันภัย โดยมีทั้งหมด 276 กระทู้ ซึ่งเป็นการรวบรวมกระทู้ทั้งหมดที่มีการแสดงความคิดเห็นตั้งแต่ปี 2543 จนถึง 2564 ข้อมูลประเภทรีวิวในกระทู้ออนไลน์พันทิปจะมีการให้คะแนนในการแสดงความคิดเห็นในกระทู้นั้น มีคะแนนตั้งแต่ 0 ถึง 5 โดยทางผู้วิจัยได้ประยุกต์ใช้ข้อมูลในส่วนนี้ในการวิเคราะห์ความรู้สึก จะเห็นได้ว่าข้อมูลที่ใช้เป็นข้อมูลก่อนการเกิดเหตุการณ์และหลังเหตุการณ์ COVID-19 ซึ่งจะทำให้มองเห็นภาพรวมของการสนใจการทำประกันตั้งแต่อดีตจนถึงปัจจุบัน ซึ่งสามารถเป็นต้นแบบในการวิเคราะห์ความสนใจในการทำประกันด้านอื่นๆ ที่มีลักษณะเดียวกันกับความสนใจในการทำประกันแบบเดียวกับเหตุการณ์ COVID-19

2.2 การเตรียมข้อมูล (Data Preparation)

เป็นขั้นตอนที่จะนำมาใช้ในการสร้างแบบจำลองการทำนายโดยมีกระบวนการต่อไปนี้ [6]

2.2.1 การบ่งบอกประเภทของความรู้สึกในบทวิจารณ์ (Class Labeling)

กำหนดโดยระดับคะแนนที่ผู้เขียนบทวิจารณ์ให้ไว้ในข้อมูลที่เก็บมาจากกระทู้รีวิวในพันทิป การให้คะแนนจะเป็นระบบคะแนนแบบห้าดาว (5 Star Scoring System) โดยระบบการให้คะแนนแบบนี้จะเรียงลำดับความพึงพอใจของลูกค้าได้ โดยเมื่อคะแนนเท่ากับ 0 คือผู้เขียนบทวิจารณ์มีความพึงพอใจต่ำสุด และคะแนนเท่ากับ 5 คือมีความพึงพอใจสูงที่สุด การจำแนกความรู้สึกสามารถแบ่งได้เป็นความรู้สึกเชิงบวก (4-5 คะแนน) [7] และความรู้สึกเชิงลบ (0-3 คะแนน) ซึ่งเราสามารถนำประเภทความรู้สึกนี้เป็นชุดข้อมูลฝึก

2.2.2 การทำความสะอาดข้อมูล (Data Cleansing)

คือการแปลงข้อมูลให้พร้อมใช้งานโดยเราใช้ Natural Language Processing: NLP ใน Python โดยใช้ Library pythainlp ในการตัดคำ (Tokenization) เพื่อแปลงประโยคเป็นคำ รวมถึงการตัดคำฟุ่มเฟือย (Stop word removal) ในการตัดคำที่ไม่สำคัญหรือคำที่พบบ่อยในเอกสารแต่ไม่ได้เป็นประเด็นสำคัญในเอกสาร และการตัดคำที่ไม่จำเป็นทิ้ง (Punctuation) ในการลบข้อความอักขระพิเศษทิ้งซึ่งเป็นสิ่งที่ไม่ต้องการ

...

2.3 การสกัดคุณลักษณะ (Feature Extraction)

การแปลงคำเป็นตัวเลขโดยที่ผลลัพธ์นั้นอยู่ในรูปของเวกเตอร์ โดยวิธีการทำ Word Embedding ซึ่งเป็นการทำคำหนึ่งคำสามารถแทนค่าด้วยตัวเลขชุดหนึ่งใช้วิธี Word2Vec จากไลบรารี่จาก Gensim สำหรับในการกำหนด Pretrained ของแบบจำลอง โดยทำการทดลองสร้าง Skip-gram ซึ่งเป็นที่นิยมใช้กันอย่างแพร่หลาย [8] โดยกำหนดโดยที่กำหนดขนาดของเวกเตอร์ที่ 128 มิติ และ จำนวนรอบเป็น 1000 รอบ [9]

2.4 การพัฒนาแบบจำลอง (Data Modelling)

ก่อนการพัฒนาแบบจำลองจำเป็นต้องมีการแบ่งข้อมูล (Split Data) ในการศึกษาครั้งนี้จะทำการแบ่งทั้งหมด 3 ส่วน คือ ข้อมูลฝึก (Training set), ข้อมูลตรวจสอบ (Validation set) และข้อมูลทดสอบ (Test set) โดยมีสัดส่วนที่อ้างอิงจาก Xu และคณะ[10] เป็น 80% 10% 10% ตามลำดับ โดยข้อมูลฝึกเป็นข้อมูลเพื่อสอนแบบจำลอง ข้อมูลตรวจสอบ (Validation set) นั้นใช้เพื่อปรับจูนพารามิเตอร์แบบจำลอง และข้อมูลทดสอบ (Test set) ใช้สำหรับทดสอบความแม่นยำของแบบจำลองที่เราสร้างขึ้น โดยในงานวิจัยนี้เราจะใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) ในการวิเคราะห์ความพึงพอใจของผู้ใช้งานต่อประกันภัยจากกระทู้ออนไลน์พันทิป

ในส่วนของการวิเคราะห์ข้อมูล มีเป้าหมายในการวิเคราะห์ความรู้สึก (Sentiment Analysis) เพื่อดึงความรู้สึกจากข้อความเป็นไปในเชิงใด โดยใช้แบบจำลองแบบมีผู้สอนจะใช้เทคนิคการจำแนก (Classification) ทั้ง 3 แบบจำลองต่อไปนี้ Gated Recurrent Unit (GRU), Bi-Long Short-Term Memory (Bi-LSTM) และ Convolutional Neural Networks (CNN)

ในงานวิจัยนี้ในแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด (version) โดยแบบจำลองชุดที่ 1 จะเป็นการสร้างแบบจำลองแบบโดยใช้ตัวแปรที่ผ่านกระบวนการ Feature Extraction ส่วนในแบบจำลองชุดที่ 2 จะใช้ตัวแปรที่ผ่านกระบวนการ Feature Extraction และเพิ่มตัวแปรที่เพิ่มเติมที่ทางผู้วิจัยคาดว่าจะสามารถเพิ่มประสิทธิภาพของแบบจำลอง และสามารถปรับใช้กับทางธุรกิจได้ โดยตัวแปรดังกล่าวคือประเภทของกระทู้รีวิว ซึ่งตัวแปรนี้มีอยู่ในทุกกระทู้รีวิวใน Pantip.com โดยมีรายละเอียดดังนี้

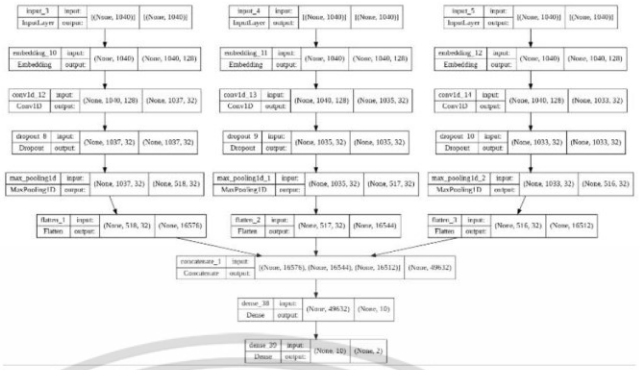
1. BR - Business Review: กระทู้นี้เป็นกระทู้รีวิวจากผู้สนับสนุน
2. CR - Consumer Review: กระทู้รีวิวนี้เป็นกระทู้ CR โดยที่เจ้าของกระทู้
3. SR - Sponsored Review: กระทู้รีวิวนี้เป็นกระทู้ SR โดยที่เจ้าของกระทู้

ซึ่งทางผู้วิจัยทำการทดลองทั้งหมด 3 แบบจำลอง โดยมีรายละเอียดดังนี้

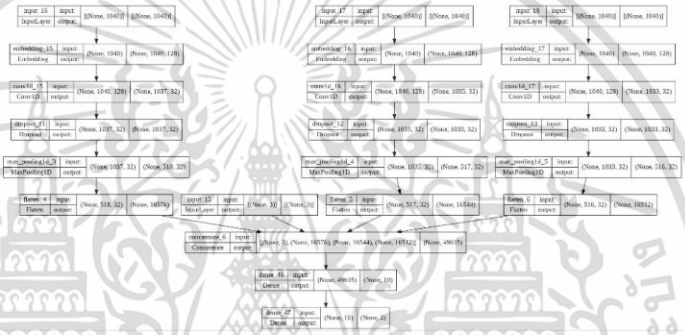
2.4.1 โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional Neural Network: CNN)

เป็นเทคนิคการเรียนรู้เชิงลึกซึ่งมีการประยุกต์หลักการวิเคราะห์ข้อมูลในรูปแบบลำดับเหตุการณ์ที่ชัดเจน ซึ่งข้อดีของข้อมูลในรูปแบบนี้สามารถเปลี่ยนบริบทของเหตุการณ์ตามลำดับได้ ประกอบไปด้วยส่วนประกอบที่สำคัญ 2 ส่วน คือ Convolutional Layer และ Pooling Layer ซึ่ง Convolutional Layer ใช้ในการอ่านข้อมูลที่เข้ามาตัวอย่างเช่น รูปภาพ และใช้เคอร์เนล (Kernel) ในการอ่านข้อมูลเล็ก ๆ ในรูปภาพตามขนาดของเคอร์เนลที่ถูกกำหนดไว้จนทั่วทั้งรูปภาพ โดยแต่ละครั้งที่เคอร์เนลเคลื่อนไปจะมีการใช้การกรอง (Filter) เพื่อรวม (Map) ข้อมูลออกมาใหม่โดยเป็นตัวแทนของข้อมูลที่รับเข้ามา (Input) ส่วน Max Pooling อีกด้วยเพื่อที่จะเอาไปประมวลผลต่อในขั้นตอนถัดๆ ไป ในทางกลับกันงานด้านอนุกรมเวลาสามารถนำ CNN มาใช้ในการสร้างแบบจำลองได้เช่นกันโดยการใช้ Convolutional 1D [11] โดย CNN ที่ใช้ในงานวิจัยนี้มีโครงสร้างดัง รูปที่ 2 และ รูปที่ 3

...



รูปที่ 2 โครงสร้างแบบจำลอง CNN ชุดที่ 1



รูปที่ 3 โครงสร้างแบบจำลอง CNN ชุดที่ 2

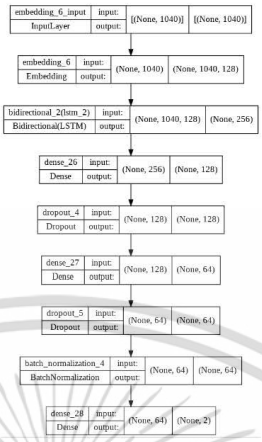
ในศึกษาครั้งนี้ได้กำหนดพารามิเตอร์สำหรับ CNN ดังนี้ Epochs = 300, Batch size = 64, Adam Optimization, Learning rate = 0.0001, Kernel size = 4, 6, 8, Dropout = 0.5, Max Pooling, Activation function (ReLU - Hidden Layer, Softmax - Output Layer)

2.4.2 หน่วยความจำระยะยาว-ระยะสั้น (Bi-Long Short-Term Memory: Bi-LSTM)

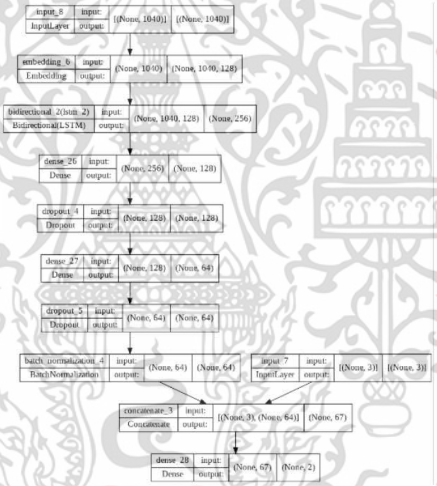
แบบจำลองนี้ถูกออกแบบมาให้มีความสามารถในการอ่านข้อมูลที่เป็นอนุกรมซึ่งมีความเหมาะสมกับการนำมาใช้ในงานด้าน Natural Language Processing และอนุกรมเวลา โดยโครงสร้างของ LSTM จะประกอบไปด้วยเซลล์และประตู (Gate) ทำหน้าที่ในการรับข้อมูลจากอนุกรมก่อนหน้าเข้ามาประมวลผลแล้วเลือกที่จะละเลย หรือทิ้งข้อมูลที่ไม่มีความสำคัญรวมทั้งทำหน้าที่ในการจดจำเฉพาะแต่สิ่งที่สำคัญไว้

ส่วน Bi-LSTM เป็นอีกรูปแบบหนึ่งของ LSTM มีหลักการทำงานเหมือนกับ LSTM เพียงแต่มีการเพิ่มหนึ่งฟังก์ชันการทำงาน คือการป้อนข้อมูลแบบย้อนกลับเข้าไปด้วย นั่นหมายความว่าขั้นตอนในการทดสอบแบบ Bi-LSTM จะช้ากว่า LSTM [12] โดย Bi-LSTM ที่ใช้ในงานวิจัยนี้มีโครงสร้างดัง รูปที่ 4 และ รูปที่ 5

...



รูปที่ 4 โครงสร้างแบบจำลอง Bi-LSTM ชุดที่ 1



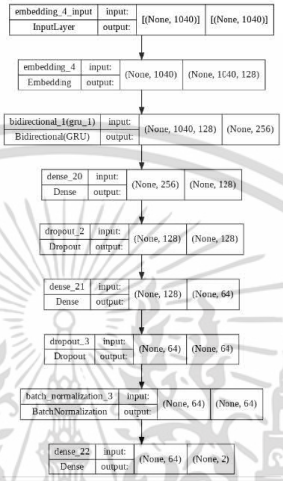
รูปที่ 5 โครงสร้างแบบจำลอง Bi-LSTM ชุดที่ 2

ในศึกษาครั้งนี้ได้กำหนดพารามิเตอร์สำหรับ Bi-LSTM ดังนี้ Epochs = 300, Batch size = 64, Optimization Adam, Learning rate = 0.0001, Dropout = 0.25, Batch Normalization, Activation function (ReLu - Hidden Layer, Softmax - Output layer)

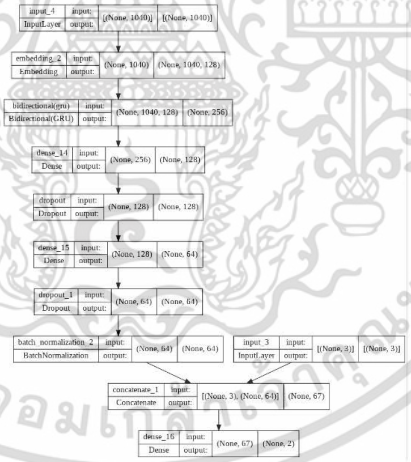
...

2.4.3 หน่วยเวียนกลับแบบมีประตู (Gated Recurrent Unit: GRU)

เป็นกลไกปิดเปิดการอัพเดตสถานะภายใน Recurrent Neural Network ที่คล้ายกับ LSTM แต่มี Parameter น้อยกว่า LSTM เนื่องจากไม่มี Output Gate GRU มีประสิทธิภาพใกล้เคียงกับ LSTM ในหลาย ๆ งาน แต่เนื่องจาก Parameter น้อยกว่าทำให้เทรนได้ง่ายกว่าและเร็วกว่า ในบางงานที่จำนวนของข้อมูลมีขนาดไม่ใหญ่มาก พบว่า GRU ประสิทธิภาพดีกว่า [13] โดย GRU มีโครงสร้างดังที่ใช้ในงานวิจัยนี้มีโครงสร้างดังรูปที่ 6 และ รูปที่ 7



รูปที่ 6 โครงสร้างแบบจำลอง GRU ชุดที่ 1



รูปที่ 7 โครงสร้างแบบจำลอง GRU ชุดที่ 2

...

ในศึกษาครั้งนี้ได้กำหนดพารามิเตอร์สำหรับ GRU ดังนี้ Epochs = 300, Batch size = 64, Optimization Adam, Learning rate = 0.0001, Dropout = 0.25, Batch Normalization, Activation function (ReLu - Hidden Layer, Softmax - Output layer)

2.5 การวัดประสิทธิภาพแบบจำลอง (Data Evaluation)

มาตรวัดประสิทธิภาพในการทดลองด้านของการจำแนกประเภทของข้อมูลมีการวัดของแบบจำลองได้หลายค่า โดยทางผู้วิจัยเลือกค่าที่จะใช้พิจารณาเลือกแบบจำลองดังนี้

2.5.1 เมทริกซ์ความสับสน (Confusion Matrix)

เป็นตารางที่ใช้สำหรับวัดผลแบบจำลองการจำแนกความรู้สึกที่เป็นเชิงบวกและเชิงลบ การเลือกผลทำนายจะทำได้โดยการวัดประสิทธิภาพของการทำนายด้วยตัวชี้วัดหลากหลายชนิดซึ่งสามารถคำนวณได้จากเมทริกซ์ความสับสน [14] ในงานวิจัยนี้ใช้เทคนิคการทำเหมืองข้อความจำแนกบทวิจารณ์โดยให้ผลบวกเป็นความรู้สึกเชิงบวก และผลลบเป็นความรู้สึกเชิงลบ และในส่วนของตัวชี้วัด งานวิจัยนี้ได้ใช้ตัวชี้วัด 4 ชนิดได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถ่วงดุล (F1-Score) และ ค่าความถูกต้อง (Accuracy) ซึ่งสามารถคำนวณได้ตามสมการที่ (2.1) ถึง (2.4)

$$Precision = TP/(TP + FP) \quad (2.1)$$

$$Recall = TP/(TP + FN) \quad (2.2)$$

$$F1\,Score = 2 * (Precision * Recall)/(Precisio + Recall) \quad (2.3)$$

$$Accuracy = (TP + TN)/(TP + FP + FN + TN) \quad (2.4)$$

โดยที่ TP แทน ผลบวกจริง , TNแทน ผลลบจริง , FP แทนผลบวกลวง , FN แทนผลลบลวง

2.5.2 ฟังก์ชันการสูญเสีย (Loss function)

ในงานวิจัยนี้จะให้ความสำคัญกับค่า Loss function เนื่องจากเป็นฟังก์ชันที่ใช้ในการคำนวณค่าความผิดพลาด (Error) ของ Neural Network ซึ่งเราจะนำความชันที่เกิดจากการหาอนุพันธ์ของ Loss Function ไปปรับค่า Weight โดยในงานวิจัยได้เลือกค่า Mean Square Error (MSE) ที่ใช้ในการปรับ Weight [15] ตามสมการที่ (2.5) และ (2.6)

$$Loss(y, \hat{y}) = \sum_{i=1}^{n}(y - \hat{y})^2 \quad (2.5)$$

$$MSE = \frac{1}{n} \times \sum(y - \hat{y})^2 \quad (2.6)$$

โดยที่ y แทน ค่าจริง , $\hat{y}$ แทน ค่าทำนาย , n แทน จำนวนข้อมูลทั้งหมด

3 ผลทดลอง

จากการสร้างเทคนิคการเรียนรู้เชิงลึก (Deep learning) เพื่อวิเคราะห์ความพึงพอใจของผู้ใช้งานต่อประกันภัยจากกระทู้ออนไลน์พันทิป ในทั้ง 3 แบบนั้นได้ค่าความถูกต้องดังตารางที่ 1 จะเห็นได้ว่าจากแบบจำลองทั้งหมด มี 4 แบบจำลอง ที่ให้ค่าด้วยความถูกต้องสูงสุดที่ 0.85 คือ แบบจำลอง CNN ชุดที่ 1 ,แบบจำลอง CNN ชุดที่ 2 , แบบจำลอง GRU ชุดที่ 2 ,แบบจำลอง BI-LSTM ชุดที่ 1 ซึ่งจากค่าความถูกต้องยังไม่สามารถสรุปได้ว่า วิธีการที่ทาง

...

ผู้วิจัยนำเสนอ สามารถพัฒนาแบบจำลองได้หรือไม่ ทางผู้วิจัยจึงพิจารณาจากค่าฟังก์ชันการสูญเสียซึ่งเป็นค่าความผิดพลาดดังตารางที่ 2

ตารางที่ 1 เปรียบเทียบค่าความแม่นยำในการวิเคราะห์ค่าความรู้สึก

Model	Version	Precision POS	Precision NEG	Recall POS	Recall NEG	F1 Score POS	F1 Score NEG	Accuracy
CNN	1	1.00	0.81	0.60	1.00	0.75	0.89	0.85
	2	1.00	0.81	0.60	1.00	0.75	0.89	0.85
GRU	1	0.60	0.76	0.60	0.76	0.60	0.76	0.70
	2	0.75	0.93	0.90	0.82	0.82	0.88	0.85
Bi-LSTM	1	0.88	0.84	0.70	0.94	0.78	0.89	0.85
	2	0.62	0.86	0.80	0.71	0.70	0.77	0.74

ตารางที่ 2 เปรียบเทียบค่าฟังก์ชันการสูญเสียในการวิเคราะห์ค่าความรู้สึก

Model	Loss function	
	Version 1	Version 2
CNN	0.61	0.60
GRU	0.66	0.60
Bi-LSTM	0.62	0.51

จากตารางที่ 2 จะเห็นได้ว่า การเปรียบเทียบประสิทธิภาพแบบของแบบจำลองจากค่าฟังก์ชันการสูญเสียแบบจำลองที่ทางผู้วิจัยนำเสนอ (ชุดที่ 2) ให้ค่าผิดพลาดน้อยกว่า แบบจำลองปกติ (ชุดที่ 1) ในทุกแบบจำลอง

จากการเปรียบเทียบค่าความถูกต้องแบบจำลองทั้ง 4 แบบจะให้ค่าความแม่นยำเท่ากัน แต่แบบจำลองที่ทางผู้วิจัยนำเสนอ สามารถลดค่าความผิดพลาดต่ำสุดที่ 0.51 คือ แบบจำลอง Bi-LSTM ชุดที่ 2 แต่แบบจำลองนี้ทางผู้วิจัยจะไม่เลือกใช้ เพราะให้ค่าความถูกต้องเพียงแค่ 0.74 ซึ่งน้อยกว่า แบบจำลองทั้ง 4 อยู่ถึง 11% ทางผู้วิจัยจึงเลือกพิจารณาที่ค่าความผิดพลาดที่ 0.60 ซึ่งมีทั้งหมด 2 แบบจำลองคือ แบบจำลอง CNN ชุดที่ 2 และแบบจำลอง GRU ชุดที่ 2

เนื่องจากงานวิจัยนี้เป็นงานวิจัยที่จะนำไปปรับใช้กับธุรกิจ ซึ่งโดยส่วนมากจะใช้ค่าความระลึก [16] ซึ่งทางผู้วิจัยจะใช้ค่า Recall POS เพราะในเชิงธุรกิจถ้าค่า Recall POS จะค่าความแม่นยำที่จะทำให้บริษัทมั่นใจได้ว่า บริษัทจะไม่เสียลูกค้าที่มีความคิดเชิงบวก เพราะเป็นกลุ่มลูกค้ากลุ่มนี้บริษัทสามารถนำไปต่อยอด โดยการเสนอขายประกันภัยเพิ่มเติมได้ จากตารางที่ 1 จะเห็นได้ว่า แบบจำลอง GRU ชุดที่ 2 ใช้ค่า Recall POS ที่ 0.90 ซึ่งสูงกว่าแบบจำลอง CNN ชุดที่ 2 อยู่ถึง 30% และเป็นค่า Recall POS ที่สูงที่สุดในทุกแบบจำลอง

...

4 สรุปผลการวิจัย

งานวิจัยนี้จึงขอนำเสนอวิธีการรับรู้เสียงของลูกค้าที่อยู่ในรูปของข้อมูลที่เป็นข้อความ ซึ่งในงานวิจัยนี้คือความพึงพอใจหรือดาวที่ผู้ตั้งกระทู้ให้คะแนนในแต่ละกระทู้รีวิว โดยในงานวิจัยนี้จะใช้ข้อมูลจากบทวิจารณ์และการเสนอความคิดเห็นจากกระทู้ออนไลน์พันทิป โดยในงานวิจัยนี้เราจะใช้การวิเคราะห์ความรู้สึกเพื่อวิเคราะห์ความรู้สึกจากข้อความว่าเป็นเชิงบวก หรือเชิงลบ โดยใช้แบบจำลองแบบมีผู้สอนจะใช้เทคนิคการจำแนก ทั้ง 3 เทคนิคต่อไปนี้ Gated Recurrent Unit (GRU) ,Bi-Long Short-Term Memory (Bi-LSTM) และ Convolutional Neural Networks (CNN) ซึ่งงานวิจัยนี้ในแต่ละแบบจำลองจะถูกพัฒนาทั้งหมด 2 ชุด โดยแบบจำลองชุดที่ 1 จะเป็นการสร้างแบบจำลองแบบโดยใช้ตัวแปรที่ผ่านกระบวนการ feature extraction ส่วนในแบบจำลองชุดที่ 2 จะใช้ตัวแปรที่ผ่านกระบวนการ feature extraction และเพิ่มตัวแปรคือประเภทของกระทู้รีวิวในกระทู้ออนไลน์พันทิป ซึ่งตัวแปรนี้มีอยู่ในทุกกระทู้รีวิวในรวมแล้วทั้งสิ้น 6 แบบจำลองจะเห็นได้ว่าแบบจำลอง GRU ชุดที่ 2 เป็นแบบจำลองที่ดีที่สุด ซึ่งให้ค่าความถูกต้องสูงที่สุดที่ 0.85 และให้ค่าความผิดพลาดที่ 0.60 ถึงแม้ว่าจะไม่ได้เป็นค่าที่น้อยที่สุดในทุกแบบจำลอง แต่เมื่อเทียบกับ แบบจำลองที่ให้ค่าความถูกต้องที่เท่ากันถือว่าแบบจำลอง GRU ชุดที่ 2 ให้ค่าต่ำที่สุด และให้ค่า Recall POS สูงสุดที่ 0.90 โดยบริษัทประกันสามารถนำแบบจำลอง GRU ชุดที่ 2 ที่มีการใส่ตัวแปรประเภทของกระทู้รีวิวในกระทู้ออนไลน์พันทิปเข้าไปเพิ่มในแบบจำลอง โดยสามารถแนะนำกลุ่มลูกค้าที่มีความคิดเห็นเชิงบวกต่อการทำประกันภัย เพื่อที่บริษัทจะสามารถเสนอขายผลิตภัณฑ์อื่นๆของทางบริษัท ผ่านทางช่องทางกระทู้ออนไลน์พันทิป เพราะสามารถระบุไปยังตัวตนของผู้ที่แสดงความคิดเห็น ซึ่งเพิ่มช่องทางการขายให้กับบริษัทประกันได้

เอกสารอ้างอิง

[1] World Health Organization. (2020). WHO | World Health Organization. https://www.who.int/thailand/emergencies/novel-coronavirus-2019/q-a-on-p-19

[2] Choi, S., Simon, L., Riedy, C., and Barrow, J. (2020). Modeling the impact of COVID-19 on dental insurance coverage and utilization. Journal of Dental Research, 100(1), 50-57. https://doi.org/10.1177/0022034520954126

[3] Coppersmith, G., Leary, R., Crutchley, P., and Fine, A. (2018). Natural language processing of social media as screening for suicide risk. Biomedical Informatics Insights, 10, 117822261879286. https://doi.org/10.1177/1178222618792860

[4] Parraga-Alava, J., Calcedo, R. A., Gomez, J. M., and Inostroza-Ponta, M. (2019). An unsupervised learning approach for automatically to categorize potential suicide messages in social media. 2019 38th International Conference of the Chilean Computer Science Society (SCCC). https://doi.org/10.1109/sccc49216.2019.8966443

[5] Aguwa, C., Olya, M. H., and Monplaisir, L. (2017). Modeling of fuzzy-based voice of customer for business decision analytics. Knowledge-Based Systems, 125, 136-145. https://doi.org/10.1016/j.knosys.2017.03.019

...

[6] Rajput, D., Thakur, R., and Basha, S. (2018). Sentiment analysis and knowledge discovery in contemporary business. IGI Global.

[7] Liu, Bing and Lei Zhang. 2012. 'A survey of opinion mining and sentiment analysis.' in, Mining text data (Springer).

[8] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[9] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." arXiv preprint arXiv:1310.4546 (2013).

[10] Xu, Y. and Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of Analysis and Testing, 2(3), 249-262. https://doi.org/10.1007/s41664-018-0068-2

[11] Liang Yao, Chengsheng Mao et al., "Clinical text classification with rule-based features and knowledge-guided convolutional neural networks" Yao et al. BMC Medical Informatics and Decision Making 2019, 19(Suppl 3):71 https://doi.org/10.1186/s12911-019-0781-4

[12] Sherstinsky, A. (2020). Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306. https://doi.org/10.1016/j.physd.2019.132306

[13] Dey, R. and Salem, F. M. (2017). Gate-variants of gated recurrent unit (GRU) neural networks. 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS). https://doi.org/10.1109/mwscas.2017.8053243

[14] Marom, N. D., Rokach, L., and Shmilovici, A. (2010). Using the confusion matrix for improving ensemble classifiers. 2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel. https://doi.org/10.1109/eeei.2010.5662159

[15] Xu, Y., Ran, X., Sun, W., Luo, X., and Wang, C. (2019). Gated neural network with regularized loss for multi-label text classification. 2019 International Joint Conference on Neural Networks (IJCNN). https://doi.org/10.1109/ijcnn.2019.8851686

[16] Kim, H., Kim, H. H., Han, B., Kim, K. H., Han, K., Nam, H., Lee, E. H., and Kim, E. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: A retrospective, multireader study. The Lancet Digital Health, 2(3), e138-e148. https://doi.org/10.1016/s2589-7500(20)30003-0

ประวัติผู้เขียน

ชื่อ	นายคงภพ ไชยศร
วัน เดือน ปีเกิด	7 มิถุนายน 2541
ที่อยู่ปัจจุบัน	39/279 หมู่บ้านอรุณนิเวศน์ ซอยนวมินทร์ 163 แยก11 ถนนนวมินทร์ เขตบึงกุ่ม แขวงนวลจันทร์ กรุงเทพ 10230
ประวัติการศึกษา	(2562) วิทยาศาสตรบัณฑิต สาขาคณิตศาสตร์ประยุกต์ (สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง)
ผลงานวิชาการ	การวิเคราะห์ความรู้สึกที่มีต่อการทำประกันภัย ด้วยเทคนิคการเรียนรู้เชิงลึก กรณีศึกษา กระทู้ออนไลน์พันทิป กรณีศึกษา ถนนเยาวราช ประเทศไทย การประชุมวิชาการทางคณิตศาสตร์ระดับชาติ ครั้งที่ 26 วันที่ 18-20 พฤษภาคม 2565