การจัดประเภทคำถามบทสนทนาภาษาไทยสำหรับแชทบอท โดยใช้โครงข่ายประสาทเทียมและโมเดลเบิร์ตแบบหลายภาษา

QUESTION CLASSIFICATION FOR THAI CONVERSATIONAL CHATBOTS USING ARTIFICIAL NEURAL NETWORKS AND MULTILINGUAL BERT MODELS

กิต ธนานุคุณ

KIT THANANUKHUN

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2566

KMITL-2023-SC-M-002-077

QUESTION CLASSIFICATION FOR THAI CONVERSATIONAL CHATBOTS USING ARTIFICIAL NEURAL NETWORKS AND MULTILINGUAL BERT MODELS

KIT THANANUKHUN

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR
THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE
DEPARTMENT OF COMPUTER SCIENCE SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2023

KMITL-2023-SC-M-002-077

หัวข้อวิทยานิพนธ์	การจัดประเภทคำถามบทสนทนาภาษาไทยสำหรับแชทบอทโดยใช้โครงข่ายประสาทเทียมและเบิร์ตหลายภาษา
ชื่อนักศึกษา	กิต ธนานุคุณ
รหัสประจำตัว	59605068
ปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2566
อาจารย์ที่ปรึกษาวิทยานิพนธ์	ผู้ช่วยศาสตราจารย์ ดร. อนันตพร หรรษคุณาฒัย

บทคัดย่อ

ในงานวิจัยนี้จะนำเสนอการประเมินประสิทธิภาพ การทำงานระหว่างอัลกอริทึมสำหรับการจัดประเภทคำถามบทสนทนาภาษาไทย โดยใช้วิธีการโมเดล Thai2Vec หรือโมเดลเบิร์ตแบบหลายภาษาร่วมกับการเรียนรู้ของเครื่อง ทั้งวิธีต้นไม้ตัดสินใจ, วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว, วิธีเบย์อย่างง่าย, วิธีซัพพอร์ตเวคเตอร์แมชชีนและวิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น โดยจัดกลุ่มข้อมูลส่วนของการเรียนรู้ตามจำนวนคำถามที่อยู่ในหมวดหมู่คำถาม ตั้งแต่จำนวน 3-5 คำถามขึ้นไปที่อยู่ในหมวดหมู่คำถามเดียวกัน โดยเมื่อเปรียบเทียบวิธีการโมเดล Thai2Vec กับวิธีการเรียนรู้ของเครื่องแบบต่าง ๆ พบว่าในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถาม ตั้งแต่ 3 คำถามขึ้นไปวิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัวจะได้ความแม่นยำสูงที่สุดซึ่งมีความแม่นยำอยู่ที่ 48.73% แต่ในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถามตั้งแต่ 4 และ 5 คำถามขึ้นไป วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นจะได้ความแม่นยำสูงที่สุด โดยมีความแม่นยำอยู่ที่ 52.64% และ 84.65% ตามลำดับ ทั้งนี้เมื่อใช้วิธีการโมเดลเบิร์ตแบบหลายภาษากับวิธีการเรียนรู้ของเครื่องแบบต่าง ๆ พบว่าในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถามตั้งแต่ 3 คำถามขึ้นไป วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัวจะได้ความแม่นยำซึ่งมีความแม่นยำอยู่ที่ 43.40% แต่ในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถามตั้งแต่ 4 และ 5 คำถามขึ้นไป วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้นจะได้ความแม่นยำสูงโดยมีความแม่นยำอยู่ที่ 53.57% และ 88.21% ตามลำดับ

คำสำคัญ : การจัดประเภทคำถาม การประมวลผลภาษาธรรมชาติ แชทบอท โมเดลเบิร์ต

Thesis Title	Question Classification for Thai Conversational Chatbots using Artificial Neural Networks and Multilingual BERT Models
Student Name	Kit Thananukhun
Student ID	59605068
Degree	Master of Science (Computer Science)
Department	Computer Science
Year	2023
Thesis Advisor	Asst.Prof.Dr. Anantaporn Hanskunatai

Abstract

In this research, we present a performance evaluation between algorithms for classifying Thai conversation questions using the Thai2Vec model or mBERT model combined with machine learning technique including Decision Trees, K-Nearest Neighbor, Naive Bayes, Support Vector Machine and Multi-Layer Perceptron, then learning by grouping the dataset according to the number of questions ranging from 3-5 questions or more that are in the same question category. When comparing the Thai2Vec with machine learning methods. It was found that in the case of question categories with 3 or more questions, KNN achieves the highest accuracy, with an accuracy of 48.73%, but in the case of with several questions ranging from 4 and 5 questions or more, the MLP will get the highest accuracy. The accuracy was 52.64% and 84.65% respectively. However, when using the mBERT with machine learning methods. It was found that in the case of question categories with 3 or more questions, KNN was more accurate, which has an accuracy of 43.40%, but in the case of question categories with ranging from 4 and 5 questions or more, MLP method will get the highest accuracy. The accuracy was 53.57% and 88.21%, respectively.

Keywords : Question Classification, NLP, Chatbot, BERT Model

กิตติกรรมประกาศ

งานวิจัยเรื่อง การจัดประเภทคำถามบทสนทนาภาษาไทยสำหรับแชทบอทโดยใช้โครงข่ายประสาทเทียมและโมเดลเบิร์ตแบบหลายภาษา สามารถประสบความสำเร็จและลุล่วงไปได้ด้วยดี ผู้จัดทำขอขอบคุณอาจารย์ที่ปรึกษา ผศ.ดร. อนันตพร หรรษคุณาฒัย ที่ได้ให้คำชี้แนะเกี่ยวกับการทำงาน แนะแนวทางการแก้ปัญหา รวมถึง ผศ.ดร. สายชล ใจเย็น และ ผศ.ดร. กุลสวัสดิ์ จิตขจรวานิช อาจารย์ที่ได้ให้คำชี้แนะเกี่ยวกับการเริ่มต้นงานวิจัย ช่วยให้งานวิจัยมีหัวข้อและรายละเอียดที่ชัดเจนมากยิ่งขึ้น

ขอขอบคุณ รศ.ดร. อนุชิต จิตพัฒนกุล ผู้ทรงคุณวุฒิจากภายนอกสถาบันฯ ประธานกรรมการสอบ และ ผศ.ดร. ศรัณย์ อินทโกสุม อาจารย์บัณฑิตประจำภาควิชาฯ กรรมการสอบวิทยานิพนธ์ที่ได้ให้คำชี้แนะจนวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วง

ขอขอบคุณ ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังที่ได้มอบโอกาสและความรู้ให้แก่ผู้จัดทำ

สุดท้ายนี้ ขอขอบคุณครอบครัวที่คอยสนับสนุนและเป็นกำลังใจให้แก่ผู้จัดทำ ในการทำงานวิจัย

กิต ธนานุคุณ

สารบัญ

หน้า

สารบัญ

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญรูป

รูปที่ **หน้า**

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

แชทบอท (Chatbot) เป็นโปรแกรมคอมพิวเตอร์ โดยมีจุดมุ่งหมายให้สามารถโต้ตอบหรือพูดคุยกับมนุษย์ได้อย่างเป็นธรรมชาติ ผ่านภาษาที่มนุษย์ใช้งานในชีวิตประจำวัน สามารถแบ่งออกได้เป็น 2 ประเภทหลัก ๆ ตามขั้นตอนของการพัฒนาโปรแกรม คือ แชทบอทที่ถูกสร้างขึ้นจากกฎการสนทนาที่ตั้งไว้ (Rule-based) คือ ในระหว่างการพัฒนาโปรแกรม ผู้พัฒนาจะทำการกำหนดรูปแบบการโต้ตอบให้กับตัวโปรแกรมเอง ตามกฎการสนทนา จึงทำให้โปรแกรมนั้นสามารถโต้ตอบกับมนุษย์ได้ในเฉพาะรูปแบบที่กำหนดไว้เท่านั้น ซึ่งถ้าเกินรูปแบบที่กำหนดไว้ โปรแกรมจะไม่สามารถดำเนินการโต้ตอบตามที่คาดหวังต่อไปได้ มีข้อดี คือ สามารถพัฒนาได้ง่ายและใช้ทรัพยากรน้อยกว่า แชทบอทที่ถูกสร้างขึ้นจากการเรียนรู้ของเครื่อง (Machine Learning) คือ ระหว่างการพัฒนาโปรแกรม ผู้พัฒนาจะใช้เทคนิคทางด้านคอมพิวเตอร์รูปแบบต่าง ๆ ในการเรียนรู้หรือประมวลผลข้อมูลที่นำเข้าสู่โปรแกรม แล้วโต้ตอบกับมนุษย์โดยอัตโนมัติ ซึ่งโปรแกรมประเภทนี้ทำการพัฒนาได้ยากและใช้ทรัพยากรมากกว่าเมื่อเทียบกับประเภทแรก แต่ผลลัพธ์ที่ได้ คือ โปรแกรมสามารถโต้ตอบกับมนุษย์อย่างเป็นอิสระ แต่ก็มีความเป็นไปได้ที่โต้ตอบนอกเหนือจากรูปแบบที่มีการการเรียนรู้ไว้

ทำให้การจัดเตรียมข้อมูลเพื่อที่จะนำเข้าสู่โปรแกรม ก่อนการพัฒนาแชทบอทที่ถูกสร้างขึ้นจากกฎการสนทนาที่ตั้งไว้ เป็นสิ่งจำเป็นอย่างยิ่งต่อการดำเนินงาน กรณีที่ผู้พัฒนาสามารถแจกแจงหรือแยกแยะข้อมูล ให้เป็นไปในรูปแบบที่โปรแกรมสามารถเรียนรู้หรือประมวลผลได้ง่าย จะทำให้แชทบอทมีประสิทธิภาพดีขึ้นตามไปด้วย โดยส่วนหนึ่งที่ต้องให้ความสำคัญเป็นอย่างยิ่ง คือ ส่วนการจัดประเภทคำถาม (Question Classification) เพราะถ้าแชทบอทสามารถจัดประเภทคำถามว่าเป็นคำถามที่อยู่ในประเภทใดระหว่างการโต้ตอบ แชทบอทก็จะสามารถโต้ตอบกับมนุษย์ได้ โดยอ้างอิงจากผลลัพท์ที่ได้ ทำให้การโต้ตอบระหว่างกันนั้น มีความเป็นอย่างธรรมชาติและอยู่ในรูปแบบเดียวกัน

จากผลลัพธ์ข้างต้น ทำให้ผู้วิจัยมุ่งเน้นดำเนินการวิจัยไปที่ ส่วนการจำแนกประเภทคำถามของแชทบอท เพื่อพัฒนาให้แชทบอทนั้น สามารถจัดการกับประเภทคำถามต่าง ๆ แล้วโต้ตอบกับมนุษย์ได้อย่างเป็นธรรมชาติและอยู่ในรูปแบบเดียวกันได้ดียิ่งขึ้น

1.2 วัตถุประสงค์ของงานวิจัย

1) เพื่อศึกษาและเปรียบเทียบวิธีการแปลงคำให้เป็นเวกเตอร์ สำหรับภาษาไทย
2) เพื่อศึกษาและเปรียบเทียบแบบจำลองการจำแนกประเภทคำถาม สำหรับภาษาไทย
3) พัฒนาโมเดลสำหรับ การจัดประเภทคำถามบทสนทนาภาษาไทย สำหรับแชทบอทโดยใช้โครงข่ายประสาทเทียมและโมเดลเบิร์ตแบบหลายภาษา

1.3 ขอบเขตของการวิจัย

1) ภาษาที่ใช้ในการโต้ตอบเป็นภาษาไทย
2) รูปแบบการโต้ตอบเป็นแบบถาม-ตอบ

1.4 ประโยชน์ที่คาดว่าได้รับ

1) สามารถนำวิธีการแปลงคำให้เป็นเวกเตอร์ สำหรับภาษาไทย ไปใช้กับแชทบอทได้
2) สามารถนำไปใช้ในการจำแนกประเภทคำถาม เพื่อลดเวลาการค้นหาคำตอบของแชทบอทได้
3) สามารถนำผลการวัดประสิทธิภาพ ในแง่ความแม่นยำและเวลาการเรียนรู้ที่ได้ ไปช่วยในการตัดสินใจพัฒนาส่วนจำแนกประเภทคำถามของแชทบอท เพื่อเพิ่มความแม่นยำในการโต้ตอบของแชทบอทได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 แชทบอท

แชทบอท (Chatbot) เป็นโปรแกรมคอมพิวเตอร์ที่เลียนแบบวิธีการสนทนาของมนุษย์ ซึ่งผู้ที่สนทนาจะสื่อสารกับแชทบอทผ่านทางอินเตอร์เฟซการแชท โดยใช้ข้อความหรือเสียงพูดก็ได้ จากนั้นแชทบอทจะแปลความหมายและประมวลผลข้อมูล แล้วโต้ตอบกลับไปยังผู้สนทนาอีกที ซึ่งประกอบไปด้วย ตัวแอปพลิเคชัน ส่วนติดต่อระบบและฐานข้อมูล โดยแชทบอทสามารถแบ่งออกได้เป็น 2 ประเภทหลัก ๆ ตามขั้นตอนของการพัฒนาโปรแกรม ได้แก่

2.1.1 แชทบอทที่สร้างจากกฎ (Rule-based Chatbot)

แชทบอทที่สร้างจากกฎ เป็นแชทบอทประเภทหนึ่งที่มีฐานข้อมูลของกฎการสนทนาเพื่อโต้ตอบกับผู้สนทนา โดยผู้สนทนาจะสื่อสารกับแชทบอทผ่านทางช่องทางของตัวแอปพลิเคชัน จากนั้นนำข้อความไปค้นหาในฐานข้อมูลดังกล่าว ว่ามีการสอดคล้องกันในกฎการสนทนาในข้อใด แล้วนำข้อความที่สอดคล้องกันในกฎการสนทนานั้นตอบกลับไปยังผู้สนทนา โดยแชทบอทประเภทนี้มีข้อดี คือ สามารถพัฒนาได้ง่ายกว่าและใช้ทรัพยากรน้อยกว่า รวมถึงการดูแลรักษาได้ง่าย แต่มีข้อเสีย คือ ผู้สนทนาจะต้องสื่อสารกับแชทบอทภายในขอบเขตที่ตั้งเอาไว้ หากเกินขอบเขตที่ตั้งเอาไว้ แชทบอทจะไม่สามารถตอบกลับหรือตอบแบบไม่สอดคล้องก็ได้ โดยมีหลักการดังนี้ คือ

2.1.1.1 การประมวลอินพุตจากผู้สนทนา (User Input Processing) คือ แชทบอทจะทำการรับอินพุตจากผู้ใช้งาน ซึ่งอาจจะเป็นประเภทข้อความหรือประเภทเสียงก็ได้

2.1.1.2 การจับคู่รูปแบบ (Pattern Matching) คือ แชทบอทจะทำการจับคู่อินพุตที่ได้รับกับชุดของกฎหรือรูปแบบที่ได้มีการกำหนดเอาไว้ล่วงหน้า โดยที่กฎหรือรูปแบบเหล่านี้จะถูกออกแบบมาเพื่อโต้ตอบกับผู้สนทนาในขอบเขตของเรื่องราวใดนั้น ๆ เช่น การกล่าวคำทักทาย การสอบถามข้อมูลหรือการขอความช่วยเหลือในเรื่องใดเรื่องหนึ่ง

2.1.1.3 การประเมินรูปแบบ (Rule Evaluation) คือ แชทบอทจะเริ่มดำเนินการโต้ตอบกับอินพุตนั้น โดยอาจเป็นข้อความที่มีการเตรียมไว้ล่วงหน้าหรือดึงข้อมูลจากแหล่งต่าง ๆ เช่น ฐานข้อมูลโดยตรงหรือ API (Application Program Interface)

2.1.1.4 การสร้างการตอบสนอง (Response Generation) คือ แชทบอทจะสร้างบทสนทนาตอบกลับไปยังผู้สนทนาตามข้อมูลที่ได้รับ โดยข้อความที่จะได้รับอาจเป็นข้อความตอบกลับธรรมดาหรือข้อความที่มีความหมายเฉพาะเจาะจงมากขึ้นก็ได้ เช่น การให้คำแนะนำแก่

ผู้สนทนาโดยรวม การนำผู้สนทนาไปยังแหล่งข้อมูลอื่น ๆ โดยเฉพาะเจาะจงหรือการรับอินพุตเพื่อดำเนินการโต้ตอบ จากนั้นระบบทำการประมวลผลการโต้ตอบเป็นลำดับถัดไป

ดังนั้นแชทบอทที่สร้างจากกฎนี้ จึงเหมาะสำหรับการจัดการสนทนาประเภทใดประเภทหนึ่ง เช่น เหตุการณ์เฉพาะที่สามารถคาดเดาและกำหนดทิศทางการสนทนาล่วงหน้าได้หรือไม่มีความซับซ้อน เนื่องจากแชทบอทประเภทนี้จะมีปัญหาในด้านการทำความเข้าใจในการโต้ตอบที่ซับซ้อนและกำกวม ซึ่งอยู่นอกเหนือจากชุดของกฎหรือรูปแบบที่ได้กำหนดไว้

2.1.2 แชทบอทที่สร้างจากปัญญาประดิษฐ์ (AI-based Chatbot)

แชทบอทที่สร้างจากปัญญาประดิษฐ์ เป็นแชทบอทประเภทหนึ่งที่มีความสามารถในการทำเข้าใจและทำการสื่อสารรูปแบบที่เป็นภาษาของมนุษย์ได้ อาจถูกสร้างขึ้นโดยใช้เทคนิคทางการเรียนรู้ของเครื่อง (Machine Learning) หรือการเรียนรู้เชิงลึก (Deep Learning) ในรูปแบบต่าง ๆ โดยทำการเรียนรู้หรือประมวลผลข้อมูลบทการสนทนาเป็นจำนวนมาก จนสามารถโต้ตอบกับมนุษย์ได้ ซึ่งการทำความเข้าใจในภาษาของมนุษย์นั้น ถูกจัดอยู่ในสาขาหนึ่งของการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ดังนั้นการจะทำให้แชทบอทสามารถเข้าใจในภาษาของมนุษย์ได้นั้น ต้องอาศัยความรู้และความเข้าใจทางด้านการประมวลผลภาษาธรรมชาติร่วมด้วย เพื่อสามารถทำงานได้อย่างมีประสิทธิภาพ โดยมีส่วนประกอบดังนี้ คือ

2.1.2.1 การเข้าใจในภาษาธรรมชาติ (Natural Language Understanding) คือ การทำความเข้าใจและแปลความหมายของอิตพุตที่ได้รับมา อาจใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติต่าง ๆ เช่น การจดจำเจตนา (Intent Recognition), การสกัดเอนทิตี (Entity Extraction) หรือการวิเคราะห์ความรู้สึกเพื่อที่จะทำความเข้าใจในความหมาย บริบท หรืออารมณ์ที่จะสื่อถึงในข้อความ (Sentiment Analysis)

2.1.2.2 การเรียนรู้ของเครื่อง (Machine Learning Model) คือ การใช้เทคนิคทางด้านการเรียนรู้ของเครื่องต่าง ๆ ในการเรียนรู้ข้อมูลเพื่อปรับปรุงประสิทธิภาพ อาจใช้เทคนิคภายใต้วิธีการที่เรียกว่าเรียนรู้แบบมีผู้สอน (Supervised Learning) ในการเรียนรู้ข้อมูล โดยได้มีการกำหนดคำตอบ (Label) ไว้ก่อนแล้วหรือเทคนิคภายใต้วิธีการเรียนรู้แบบเสริมแรง (Reinforcement Learning) ในการเรียนรู้ข้อมูลแบบมีการปรับการทำงานของแชทบอทในรูปแบบต่าง ๆ เพื่อให้เกิดผลลัพธ์ที่แตกต่างกันออกไปตามแต่สภาพแวดล้อมระหว่างการเรียนรู้

2.1.2.3 การจัดการโต้ตอบ (Dialog Management) คือ การจดจำประวัติของการสนทนา โดยจดจำบริบทของการโต้ตอบและทำความเข้าใจว่ามีการโต้ตอบไปอย่างไรบ้าง เพื่อที่จะสร้างการโต้ตอบตามประวัติการสนทนาที่ทำการจดจำไว้

2.1.2.4 การสร้างภาษาธรรมชาติ (Natural Language Generation) คือ การสร้างการตอบสนองเพื่อให้สามารถโต้ตอบได้เหมือนกับมนุษย์ โดยสร้างข้อความตามหลักไวยากรณ์และที่มีการใช้อยู่ในชีวิตประจำวันของมนุษย์ผ่านการเรียนรู้

2.1.2.5 การรวมฐานความรู้ (Knowledge Base Integration) คือ การรวมข้อมูลเข้ากับฐานความรู้หรือแหล่งข้อมูลที่ถูกต้องและเป็นปัจจุบันให้แก่ผู้สนทนา ซึ่งจะทำให้แชทบอทนั้นสามารถตอบคำถามหรือให้คำแนะนำตามความรู้ที่มีได้

ดังนั้นแชทบอทที่สร้างจากปัญญาประดิษฐ์นี้ เหมาะสำหรับการสนทนาประเภทที่มีความซับซ้อน มีความหลากหลายมากยิ่งขึ้น โดยสามารถปรับทิศทางการโต้ตอบที่มีความแตกต่างกันและเรียนรู้การโต้ตอบของผู้สนทนา เพื่อนำไปปรับปรุงการโต้ตอบให้มีประสิทธิภาพให้ดียิ่งขึ้นเมื่อมีการใช้งาน ซึ่งต้องมีการใช้ข้อมูลในการเรียนรู้ขนาดใหญ่และจำนวนมหาศาล เพื่อให้สามารถทำการโต้ตอบให้เหมือนกับมนุษย์ได้ แต่มีข้อเสีย คือ ในบางกรณีแชทบอทประเภทนี้ไม่อาจโต้ตอบได้ดีเท่าในรูปแบบที่มีการกำหนดขอบเขตเอาไว้แล้ว เนื่องจากระหว่างการโต้ตอบอาจมีการโต้ตอบนอกเหนือจากขอบเขตที่คาดหวังไว้ในการสนทนา

จากผลลัพธ์ข้างต้นแสดงให้เห็นว่า แชทบอทที่สร้างจากกฎกับแชทบอทที่สร้างจากปัญญาประดิษฐ์ มีความแตกต่างดังนี้ คือ

ตารางที่ 2.1 ความแตกต่างของแชทบอทที่สร้างจากกฎกับแชทบอทที่สร้างจากปัญญาประดิษฐ์

หัวข้อ	ที่สร้างจากกฎ	ที่สร้างจากปัญญาประดิษฐ์
เทคโนโลยีและวิธีการใช้งาน	มีการกำหนดกฎเกณฑ์หรือรูปแบบเอาไว้ล่วงหน้า แบบเฉพาะเจาะจง รวมถึงมีความจำเป็นที่จะต้องมีการกำหนดแผนของการพัฒนาโปรแกรม เพื่อให้เกิดความชัดเจนในการทำงาน	มีการใช้เทคนิคหรือวิธีต่าง ๆ ทางด้านปัญญาประดิษฐ์ ทางด้านการประมวลผลภาษาธรรมชาติ ทางด้านการเรียนรู้ของเครื่องหรือทางด้านการเรียนรู้เชิงลึกเพื่อพัฒนาการโปรแกรมให้สามารถตอบสนองได้
การปรับตัวและการเรียนรู้	ไม่มีความสามารถในการปรับตัว อยู่ภายใต้ขอบเขตที่จำกัดและไม่สามารถทำการเรียนรู้เพิ่มเติมหรือทำการ	มีความสามารถในการปรับตัว ตามข้อมูลอินพุตที่ได้รับจากผู้ใช้งาน โดยที่จะมีความแตกต่างกันออกไป

	ปรับปรุงใดเอง ๆ หากไม่มีการอัปเดต ต้องมาจากผู้พัฒนาเท่านั้น	เนื่องจากมีการเรียนรู้ จากชุดข้อมูลขนาดใหญ่และเรียนรู้ผ่านการโต้ตอบของผู้ใช้งานเอง โดยวิธีการเรียนรู้ของเครื่องแบบต่าง ๆ
บริบทและทิศทางการโต้ตอบ	มีปัญหากับการรักษาบริบทของการสนทนแบบต่อเนื่อง เพราะขาดความสามารถในการจดจำและโต้ตอบกับอดีตหรือการรักษาลำดับของการโต้ตอบที่สอดคล้องกัน โดยนอกเหนือไปจากกฎที่ถูกกำหนดล่วงหน้าเอาไว้	มีความสามารถเข้าใจและจดจำในบริบทของการสนทนาที่เกิดก่อนหน้านี้ได้ รวมถึงทำการสร้างการโต้ตอบกลับ เนื่องจากมีการเรียนรู้จากประวัติของการสนทนากัน ที่ได้มีการรวบรวมเอาไว้
ความยืดหยุ่นและการพัฒนา	มีความง่ายต่อการออกแบบและใช้งาน ทำให้เหมาะสำหรับกรณีที่ต้องการใช้งานในรูปแบบที่เรียบง่ายและมีความเฉพาะเจาะจง แต่อย่างไรก็ตาม อาจต้องใช้เวลาและทรัพยากรมากในการปรับขนาดหรือทำให้มีการรองรับสถานการณ์การโต้ตอบในรูปแบบใหม่	มีความสามารถจัดการกับการโต้ตอบที่มีขอบเขตที่กว้างขึ้นและเรียนรู้ในชุดข้อมูลที่มีความหลากหลายได้ ซึ่งทำให้สามารถปรับตัวเองให้เข้ากับสถานการณ์ต่าง ๆ และจัดการกับการโต้ตอบที่มีความซับซ้อนแตกต่างกัน

2.2 การเรียนรู้ของเครื่อง

การเรียนรู้ของเครื่อง (Machine Learning) เป็นสาขาหนึ่งของวิชาปัญญาประดิษฐ์ (Artificial Intelligence) ที่มีจุดประสงค์ในการมุ่งเน้นไปในส่วนของการออกแบบและพัฒนาอัลกอริทึมด้วยเทคนิคต่าง ๆ ที่ทำให้คอมพิวเตอร์สามารถเรียนรู้จากข้อมูลที่มีได้ โดยหลักการทำงานนั้นจะใช้โมเดลทางคณิตศาสตร์ในการอธิบายถึงวิธีการเรียนรู้ ซึ่งขั้นตอนการเรียนรู้ประกอบไปด้วยการใช้ชุดข้อมูลฝึกสอน (Training Dataset) ในการปรับค่าพารามิเตอร์หรือตัวแปรของโมเดล โดยเมื่อกระทำเช่นนี้ไปเรื่อย ๆ ค่าพารามิเตอร์หรือตัวแปรของโมเดลจะมีการปรับเปลี่ยนตามข้อมูลที่มี

การนำเข้าสู่โมเดล จากนั้นจึงใช้ชุดข้อมูลทดสอบ (Test Dataset) ลองนำเข้าสู่โมเดล เพื่อใช้ในการตรวจสอบความถูกต้อง

โดยการเรียนรู้ของเครื่องนั้นสามารถแบ่งออกได้ 3 ประเภท ตามวิธีการเรียนรู้ คือ การเรียนรู้แบบมีผู้สอน (Supervised Learning), การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) และการเรียนรู้แบบเสริมแรง (Reinforcement Learning)

2.2.1 การเรียนรู้แบบมีผู้สอน (Supervised Learning)

การเรียนรู้แบบมีผู้สอน เป็นการเรียนรู้ชุดจากข้อมูลฝึกสอน (Training Dataset) ที่มีการกำหนดคำตอบ (Label) ประกอบอยู่ด้วย ซึ่งภายในชุดข้อมูลจะประกอบด้วยข้อมูลต้น (Input) และคำตอบ (Output) ที่เราต้องการ โดยในส่วนของการเรียนรู้ ระบบจะพยายามทำนายผลลัพธ์ให้ตรงกับคำตอบที่มีการเตรียมเอาไว้ ด้วยการปรับพารามิเตอร์หรือตัวแปรที่อ้างอิง โดยชุดข้อมูลฝึกสอนที่นำเข้า มีองค์ประกอบหลักดังนี้ คือ

2.2.1.1 ชุดข้อมูลฝึกสอน (Training Data) คือ ชุดข้อมูลที่มีการกำหนดคำตอบประกอบอยู่ด้วย ซึ่งภายในประกอบด้วย ข้อมูลต้น (ตัวแปรต้น) และคำตอบ (ตัวแปรตาม) ที่มีความสอดคล้องกัน

2.2.1.2 การฝึกสอนโมเดล (Model Training) คือ การเรียนรู้ข้อมูลจากชุดข้อมูลฝึกสอนโดยโมเดล (Model) ซึ่งเรียนรู้จากข้อมูลต้น (Input) และคำตอบ (Output) ที่นำเข้าสู่ส่วนการเรียนรู้และปรับค่าพารามิเตอร์ (Parameter) ภายในโมเดล เพื่อลดความแตกต่างระหว่างคำตอบที่ทำนายกับคำตอบที่สอดคล้องกันในชุดข้อมูลฝึกสอน

2.2.1.3 การทำนาย (Prediction) คือ เมื่อโมเดลได้รับการเรียนรู้แล้วจะสามารถใช้โมเดลนั้นเพื่อทำนายผลลัพธ์เกี่ยวกับข้อมูลต้นใหม่ ที่นำเข้าสู่โมเดลได้เพื่อต้องการทำนายผลลัพธ์ได้ โดยคำตอบที่ได้นั้น อ้างอิงมาจากการเรียนรู้จากข้อมูลต้น (Input) และคำตอบ (Output) ที่นำเข้าสู่ส่วนการเรียนรู้และปรับพารามิเตอร์ (Parameter) ภายในโมเดล

การเรียนรู้ประเภทนี้มักใช้ในการแก้ไขปัญหาประเภท การจำแนกประเภทข้อมูล (Classification) คือ การให้โมเดลเรียนรู้ในการจำแนกข้อมูลออกมาเป็นหมวดหมู่ที่ถูกกำหนดไว้ล่วงหน้า เช่น การตรวจจับสแปมของอีเมล (Spam Detection), การระบุวัตถุในรูปภาพ (Object Recognition), การวิเคราะห์ความรู้สึก (Sentiment Analysis) แล้วมักใช้ในการแก้ไขปัญหาประเภทการหาความถดถอย (Regression) เช่นกัน คือ การให้โมเดลเรียนรู้ในการทำนายค่าตัวเลขแบบต่อเนื่อง เช่น การทำนายราคาบ้าน (Predicting House Prices), การประมาณยอดขาย (Estimating Sales Figures) หรือการคาดการณ์ราคาหุ้น (Forecasting Stock Prices)

ตัวอย่างการเรียนรู้แบบมีผู้สอน เช่น การถดถอยเชิงเส้น (Linear Regression), การถดถอยโลจิสติก (Logistic Regression), ต้นไม้ตัดสินใจ (Decision Trees), ซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine), ป่าสุ่ม (Random Forest) หรือโครงข่ายประสาทเทียม (Artificial Neural Networks) ซึ่งการเลือกวิธีการเรียนรู้แบบมีผู้สอนนี้จะขึ้นอยู่กับลักษณะของปัญหา ข้อมูลที่มี รวมถึงความแม่นยำที่ต้องการ

2.2.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning)

การเรียนรู้แบบไม่มีผู้สอน เป็นการเรียนรู้จากชุดข้อมูลฝึกสอน (Training Dataset) ที่ไม่มีการกำหนดคำตอบ (Unlabeled) ประกอบอยู่ด้วย ซึ่งภายในชุดข้อมูลจะประกอบด้วยข้อมูลต้น (Input) เพียงอย่างเดียว ในส่วนของการเรียนรู้ข้อมูล ระบบจะพยายามทำนายผลลัพธ์ โดยค้นหารูปแบบ โครงสร้าง ความสัมพันธ์ แล้วจัดกลุ่มของข้อมูลจากความเชื่อมโยงกันของข้อมูลด้วยการปรับพารามิเตอร์หรือตัวแปรที่อ้างอิงโดยชุดข้อมูลฝึกสอนที่นำเข้า มีลักษณะสำคัญดังนี้ คือ

2.2.2.1 ข้อมูลที่ไม่มีการกำหนดคำตอบ (Unlabeled Data) คือ การเรียนรู้แบบไม่มีผู้สอนจะเรียนรู้ข้อมูลฝึกสอบแบบไม่มีการกำหนดคำตอบ เพื่อค้นหาความเชื่อมโยง

2.2.2.2 การค้นหารูปแบบ (Pattern Discovery) คือ การค้นหารูปแบบ โครงสร้าง และความสัมพันธ์ โดยเรียนจากข้อมูลฝึกสอนที่ได้รับเท่านั้น ไม่จำเป็นต้องวางแนวทาง

2.2.2.3 การทำคลัสเตอร์ (Clustering) คือ การจัดกลุ่มข้อมูลที่คล้ายกันออกเป็นกลุ่ม ๆ ตามคุณสมบัติหรือลักษณะเฉพาะที่ได้ค้นพบจากที่มีการเรียนรู้

การเรียนรู้ประเภทนี้มักใช้ในการแก้ไขปัญหาประเภท การจัดข้อมูลกันเป็นกลุ่ม (Clustering) หรือการหาความเชื่อมโยง (Association) เช่น การแบ่งกลุ่มลูกค้า (Customer Segmentation), การตรวจจับความผิดปกติ (Anomaly Detection), ระบบคำแนะนำ (Recommendation System) หรือการวิเคราะห์รูปภาพและข้อความ (Image and Text Analysis)

ตัวอย่างการเรียนรู้แบบไม่มีผู้สอน เช่น การจัดกลุ่มด้วยค่า K-means (K-means Clustering), การจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering), การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) หรือการเข้ารหัสอัตโนมัติ (Autoencoder) ซึ่งการเลือกวิธีการเรียนรู้แบบมีผู้สอนนี้จะขึ้นอยู่กับลักษณะของปัญหา ข้อมูลที่มี รวมถึงความแม่นยำที่ต้องการ

2.2.3 การเรียนรู้แบบเสริมแรง (Reinforcement Learning)

การเรียนรู้แบบเสริมแรง เป็นการเรียนรู้ที่มุ่งเน้นไปที่การปฏิสัมพันธ์กับสภาพแวดล้อมที่พบ โดยกระทำการบางอย่างที่เป็นผลให้เกิดการแก้ไขปัญหาและได้ผลลัพธ์ที่ดีที่สุด การเรียนรู้ประเภทนี้ระบบจะรับรู้สภาพแวดล้อมรอบข้างผ่านทางสถานะ (State) จากนั้นตัวแทนระบบ (Agent) และการดำเนินการ (Action) ต่อสภาพแวดล้อม (Environment) เพื่อให้เกิดผลลัพธ์

ที่แตกต่างกันออกไป ซึ่งในกรณีที่กระทำแล้วเป็นผลลัพธ์ที่ดี ระบบจะได้รับสิ่งแทนรางวัล (Reward) ที่เรียกว่า การเสริมแรงบวก (Positive Reinforcement) และในกรณีที่กระทำแล้วเป็นผลลัพธ์ที่ไม่ดี ระบบจะได้รับสิ่งแทนการลงโทษที่เรียกว่า การเสริมแรงลบ (Negative Reinforcement) และท้ายที่สุดระบบจะใช้การอ้างอิงในส่วนของการเสริมแรงบวกและการเสริมแรงลบ ในการปรับวิธีการดำเนินงานเพื่อให้ได้ผลลัพธ์ที่ดีที่สุด มีลักษณะสำคัญดังนี้ คือ

2.2.3.1 ตัวแทนระบบ (Agent) คือ ผู้เรียนหรือผู้มีอำนาจในการตัดสินใจที่จะมีการปฏิสัมพันธ์กับสภาพแวดล้อมโดยรับรู้สถานะปัจจุบันของสภาพแวดล้อม จากนั้นดำเนินการกระทำตามที่ตัดสินใจต่อสภาพแวดล้อมนั้น

2.2.3.2 สภาพแวดล้อม (Environment) คือ ระบบที่ตัวแทนระบบได้ดำเนินการบางอย่าง เพื่อให้เกิดผลลัพธ์ที่แตกต่างกันโดยได้รับสิ่งที่เรียกว่า การเสริมแรงบวกหรือการเสริมแรงลบเป็นผลตอบแทนการดำเนินการนั้น

2.2.3.3 สถานะ (State) คือ สถานการณ์ปัจจุบันหรือค่าของสภาพแวดล้อม ณ เวลาที่กำหนด โดยจะส่งข้อมูลให้กับตัวแทนระบบเพื่อตัดสินใจและดำเนินการ

2.2.3.4 การดำเนินการ (Action) คือ ตัวเลือกที่ตัวแทนระบบได้กระทำต่อสภาพแวดล้อมเพื่อตอบสนองสถานะที่ได้รับ

2.2.3.5 รางวัล (Reward) คือ สัญญาณตัวเลขจากสภาพแวดล้อมที่ให้กับตัวแทนระบบตอบแทนที่มีการดำเนินงานต่อสภาพแวดล้อม โดยเป้าหมายของตัวแทนระบบนั้น คือ การเพิ่มค่าการเสริมแรงบวกเมื่อเวลาผ่านไป

การเรียนรู้ประเภทนี้มักใช้ในการแก้ไขปัญหาประเภทที่ต้องมีการลองผิดลองถูกการใช้ประโยชน์หรือการสำรวจ เนื่องจากตัวแทนระบบต้องมีการโต้ตอบซ้ำ ๆ กับสภาพแวดล้อมเพื่อให้ได้รับรางวัล เช่น การเล่นเกม การทำงานของหุ่นยนต์ ยานพาหนะแบบอัตโนมัติระบบให้คำแนะนำหรือการจัดการทรัพยากร

ตัวอย่างการเรียนรู้แบบเสริมแรง เช่น Q-Learning, Monte Carlo หรือ SARSA ซึ่งการเลือกวิธีการเรียนรู้แบบมีผู้สอนนี้จะขึ้นอยู่กับลักษณะของปัญหา ข้อมูลที่มี รวมถึงความแม่นยำตามที่ต้องการ

ตารางที่ 2.2 ความแตกต่างของการเรียนรู้แบบมีผู้สอนกับไม่มีผู้สอนและแบบเสริมแรง

หัวข้อ	แบบมีผู้สอน	แบบไม่มีผู้สอน	แบบเสริมแรง
คำจำกัดความ	เรียนรู้โดยใช้ข้อมูลที่มีการกำหนดคำตอบ	เรียนรู้โดยใช้ข้อมูลที่ไม่มีการกำหนดคำตอบ	เรียนรู้โดยใช้วิธีการที่มีการดำเนินการต่อสภาพแวดล้อม
ข้อมูลฝึกสอน	ข้อมูลที่มีการกำหนดคำตอบ	ข้อมูลที่ไม่มีการกำหนดคำตอบ	ไม่มีการกำหนดข้อมูลไว้ล่วงหน้า
การแก้ปัญหา	การจัดจำแนกประเภทข้อมูลหรือการหาความถดถอย	การจัดจำแนกข้อมูลเป็นกลุ่มหรือการหาความเชื่อมโยง	การลองผิดลองถูกการใช้ประโยชน์หรือการสำรวจ
การกำกับดูแล	ต้องมีการกำกับ	ไม่ต้องมีการกำกับ	ไม่ต้องมีการกำกับ
จุดมุ่งหมาย	ทำนายผลลัพธ์	ค้นหารูปแบบ	เรียนรู้การดำเนินการ

2.3 การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ (Natural Language Processing) เป็นสาขาหนึ่งของวิชาปัญญาประดิษฐ์ (Artificial Intelligence) ที่มีจุดประสงค์ในการมุ่งเน้นไปที่การปฏิสัมพันธ์ระหว่างคอมพิวเตอร์และภาษาของมนุษย์ พัฒนาอัลกอริทึมและแบบจำลองการคำนวณ เพื่อช่วยให้คอมพิวเตอร์ได้เข้าใจ ตีความและสร้างภาษามนุษย์อย่างมีความหมายโดยใช้แบบจำลองทางสถิติ เช่น การเรียนรู้ของเครื่องหรือการเรียนรู้เชิงลึกในการเรียนรู้ข้อมูล ซึ่งการทำงานเบื้องต้นมีดังนี้ คือ

2.3.1 การประมวลผลข้อความล่วงหน้า (Text Preprocessing) คือ การจัดการข้อมูลที่ได้รับให้อยู่ในรูปแบบที่เหมาะสม ก่อนทำการวิเคราะห์ เช่น การแยกข้อความออกเป็นคำ การลบเครื่องหมายที่ไม่จำเป็น การแปลงข้อความเป็นตัวพิมพ์เล็กหรือการจัดการกับอักขระพิเศษ

2.3.2 การแท็กส่วนหนึ่งของคำพูด (Part-of-Speech Tagging) คือ การกำหนดแท็กทางไวยากรณ์ให้กับคำในประโยค เช่น คำนาม คำกริยาหรือคำคุณศัพท์ ที่สัมพันธ์กันในรูปประโยค

2.3.3 การการแยกวิเคราะห์ประโยค (Syntactic Parsing) คือ การวิเคราะห์โครงสร้างทางไวยากรณ์ของประโยคเพื่อทำความเข้าใจความสัมพันธ์ระหว่างคำ เช่น หัวเรื่อง กรรมหรือกริยา

2.3.4 การรู้จำเอนทิตีที่มีชื่อ (Named Entity Recognition) คือ การระบุและแยกเอนทิตีที่มีชื่อออกจากข้อความ เช่น ชื่อของผู้คน องค์กร สถานที่ วันที่หรือเอนทิตีที่มีความเฉพาะอื่น ๆ

2.3.5 การเรียนรู้แบบจำลอง (Model Training) คือ การนำข้อมูลที่ดำเนินการมาเข้าสู่การเรียนรู้ของเครื่องแบบมีผู้สอน เพื่อเรียนรู้ข้อมูลจากข้อมูลที่มีการกำหนดคำตอบไว้ก่อนแล้ว

2.3.6 การประเมินผลแบบจำลอง (Model Evaluation) คือ การประเมินประสิทธิภาพของแบบจำลองที่ผ่านการเรียนรู้มาแล้วว่ามีความถูกต้องหรือแม่นยำเท่าใด

จากหลักการข้างต้น ทำให้การประมวลผลภาษาธรรมชาติได้ถูกประยุกต์ใช้ในงานประเภทต่าง ๆ ที่เกี่ยวข้องกับการปฏิสัมพันธ์ในชีวิตประจำวัน เช่น

1) การวิเคราะห์ความรู้สึก (Sentiment Analysis) คือ การพิจารณาความรู้สึกหรือน้ำเสียงทางอารมณ์ที่มีการแสดงในข้อความ มักถูกจัดประเภทเป็นเชิงบวก เชิงลบหรือเป็นกลาง เช่น ความคิดเห็นของลูกค้าหรือแนวโน้มของโซเชียลมีเดีย

2) การแปลด้วยเครื่อง (Machine Translation) คือ การแปลข้อความจากภาษาหนึ่งเป็นอีกภาษาหนึ่งโดยอัตโนมัติ ทำให้สามารถสื่อสารและทำความเข้าใจในภาษาต่าง ๆ

3) การตอบคำถาม (Question Answering) คือ การสร้างระบบที่สามารถเข้าใจและให้คำตอบสำหรับคำถามที่ถามเป็นภาษาของมนุษย์ โดยประมวลผลคำถามของมนุษย์และดึงข้อมูลที่เกี่ยวข้องเพื่อให้ได้คำตอบที่กระชับและถูกต้อง มักอ้างอิงตามฐานความรู้ที่กำหนดไว้

4) การสร้างข้อความ (Text Generation) คือ การสร้างข้อความที่สอดคล้องกันและมีความหมาย เช่น การสร้างบทสรุป การถอดความหรือการเขียนในเชิงสร้างสรรค์

2.4 การจัดประเภทคำถาม

การจัดประเภทคำถาม (Question Classification) เป็นการจัดประเภท กำหนดประเภทหรือหมวดหมู่ที่มีการกำหนดไว้ล่วงหน้าให้กับคำถามที่กำหนดไว้ ตามความหมายหรือวัตถุประสงค์ของคำถามนั้นโดยมีจุดมุ่งหมาย คือ การอำนวยความสะดวกในการดึงข้อมูลหรือการตอบคำถามให้มีประสิทธิภาพ เนื่องจากการจัดประเภทคำถามจะช่วยให้มีจับคู่คำถามกับคำตอบหรือแหล่งข้อมูลที่เกี่ยวข้องได้ง่ายยิ่งขึ้น ซึ่งมีวิธีการในการจัดประเภทคำถาม เช่น

2.4.1 วิธีการตามคุณลักษณะ (Feature-based Approaches) คือ การดึงคุณลักษณะทางด้านภาษาศาสตร์ออกจากคำถาม เช่น วิธีการแท็กส่วนหนึ่งของคำพูด (Part-of-Speech Tagging) หรือการการแยกวิเคราะห์ประโยค (Syntactic Parsing) จากนั้นจึงใช้การเรียนรู้ของเครื่อง เช่น ต้นไม้ตัดสินใจ (Decision Trees) หรือซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) ในการจัดประเภทคำถาม

2.4.2 วิธีการเรียนรู้ของเครื่อง (Machine Learning Approaches) คือ การใช้การเรียนรู้แบบมีผู้สอน (Supervised Learning) เช่น การถดถอยโลจิสติก (Logistic Regression) หรือโครงข่ายประสาทเทียม (Artificial Neural Networks) ในการเรียนรู้ข้อมูลที่มีการกำหนดคำตอบ

ไว้แล้ว ซึ่งเรียนรู้รูปแบบพื้นฐานหรือความสัมพันธ์ระหว่างคุณสมบัติของข้อมูลต้นและคำตอบ เพื่อใช้ในการจัดประเภทคำถาม

2.4.3 วิธีการเรียนรู้เชิงลึก (Deep Learning Approaches) คือ การใช้การเรียนรู้เชิงลึก (Deep Learning) เช่น โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Networks) หรือโครงข่ายประสาทเทียมแบบวนรอบ (Convolutional Neural Networks) ในการเรียนรู้ข้อมูล ซึ่งสามารถเรียนรู้ข้อมูลคำต้นและคำถามที่มีความซับซ้อน รวมทั้งจับความสัมพันธ์ระหว่างคำต่างได้ดี

2.5 วิธีต้นไม้ตัดสินใจ

วิธีต้นไม้ตัดสินใจ (Decision Trees) เป็นการเรียนรู้ของเครื่องแบบมีผู้สอน ซึ่งจำลองการตัดสินใจหรือการกระทำ โดยการสร้างโครงสร้างคล้ายแผนผังต้นไม้หรือแผนผังลำดับงาน ซึ่งแต่ละโหนดภายในแผนผังต้นไม้หรือแผนผังลำดับงานนั้น แสดงถึงคุณลักษณะหรือแอตทริบิวต์ ซึ่งแต่ละสาขาแสดงถึงกฎการตัดสินใจและโหนดปลายของแต่ละโหนด แสดงถึงการทำนายผลลัพธ์ของกฎการตัดสินใจนั้น มีลักษณะสำคัญดังนี้ คือ

1) โครงสร้างต้นไม้ (Trees Construction) คือ ปกติอัลกอริทึมต้นไม้ตัดสินใจเริ่มจากการนำชุดข้อมูลตั้งต้นเริ่มที่รูทโหนด (Root Node) จากนั้นมีการประเมินคุณลักษณะต่าง ๆ แล้วแยกข้อมูลตามแอตทริบิวต์ (Attribute) ที่ใกล้เคียงที่สุดในแต่โหนดแล้วดำเนินการซ้ำ ๆ จนกว่าจะหมดแผนผังลำดับงาน

2) การเลือกแอตทริบิวต์ (Attribute Selection) คือ ในแต่ละโหนดภายในอัลกอริทึมต้นไม้ตัดสินใจจะทำการกำหนดแอตทริบิวต์หรือคุณสมบัติที่ดีที่สุด เพื่อแยกข้อมูลตามเกณฑ์ที่กำหนด

3) การแยกโหนด (Node Splitting) คือ เมื่อเลือกแอตทริบิวต์ที่ดีที่สุดแล้ว ข้อมูลจะถูกแบ่งออกเป็นชุดย่อยตามค่าแอตทริบิวต์ที่เลือกไว้ โดยสอดคล้องกับที่อยู่โหนดปัจจุบัน จากนั้นไปยังโหนดย่อยชุดถัดไปที่อยู่ภายใต้โหนดปัจจุบัน

4) การสร้างโหนดลีฟ (Leaf Node Creation) คือ กระบวนการของอัลกอริทึมต้นไม้ตัดสินใจจะทำการดำเนินต่อไปจนกว่าตรงตามเงื่อนไขการหยุด เช่น ถึงความลึกสูงสุดหรือจำนวนตัวอย่างขั้นต่ำต่อโหนดลีฟ

5) การตัดแต่งกิ่งต้นไม้ (Trees Pruning) คือ หลังจากสร้างแผนผังต้นไม้หรือแผนผังลำดับงานครั้งแรก อาจใช้เทคนิคการตัดแต่งโหนดเพื่อลดจำนวนโหนดที่มากเกินไปหรือจะเป็นการลบหรือยุบโหนดที่ไม่ได้ใช้หรือไม่ได้มีส่วนสำคัญต่อการดำเนินงานก็ได้

6) การคาดคะเน (Prediction) คือ การเรียนรู้ข้อมูลจากแผนผังต้นไม้หรือแผนผังลำดับงาน โดยที่ดูโครงสร้างต้นไม้ตั้งแต่รูทโหนดจนถึงโหนดสุดท้าย ซึ่งโหนดสุดท้ายนั้นจะถูกนับเป็นคำตอบของการทำนายผลลัพทธ์นั้น

จากลักษณะข้างต้น วิธีต้นไม้ตัดสินใจมีความสามารถในการจัดการข้อมูลที่มีลักษณะที่เป็นเชิงหมวดหมู่และเชิงตัวเลขหรือความสัมพันธ์ที่ไม่ใช่เชิงเส้น จึงมักใช้ในงานประเภท การเงิน การดูแลสุขภาพ การตลาดหรือจัดการลูกค้าสัมพันธ์

โดยงานวิจัยนี้จะใช้วิธีต้นไม้ตัดสินใจ ดังสมการที่ 2.1

$$p_{mk} = \frac{1}{n_m} \sum_{y \in Q_m} I(y = k) \tag{2.1}$$

เมื่อ $k = 0, 1, 2, \ldots, K-1$ และ K แทนจำนวนคลาสทั้งหมด

p_{mk} คือ ความน่าจะเป็นของคลาส k ที่โหนด m

n_m คือ จำนวนข้อมูลที่โหนด m

Q_m คือ ข้อมูลในโหนด m

$$I(x) = \begin{cases} 1, & x \; is \; true \\ 0, & otherwise \end{cases}$$

2.6 วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว

วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (K-nearest Neighbors) เป็นการเรียนรู้ของเครื่องแบบมีผู้สอนซึ่งใช้สำหรับการจัดหมวดหมู่และการหาค่าถดถอย โดยเป็นอัลกอริทึมแบบไม่มีพารามิเตอร์ หมายความว่าไม่ได้มีการตั้งสมมติฐานใด ๆ เกี่ยวกับการกระจายข้อมูลพื้นฐาน ซึ่งมีหลักการเบื้องต้นคือ เมื่อมีการระบุจุดข้อมูลใหม่ อัลกอริทึมจะคำนวณระยะทาง (เช่น ระยะทางแบบยุคลิด) ระหว่างจุดใหม่และจุดข้อมูลการฝึกอบรมทั้งหมด จากนั้นเลือกเพื่อนบ้านที่ใกล้ที่สุด “k” ตามระยะทางที่สั้นที่สุด แล้วดูจากคะแนนเสียงข้างมากในหมู่เพื่อนบ้าน “k” นั้นมีใกล้คำตอบหรือคลาสอะไรมากที่สุด มีขั้นตอนการทำงาน คือ

1) กำหนดขนาดของ “k” (ควรกำหนดเป็นเลขคี่)

2) คำนวณระยะห่างของข้อมูลที่ต้องการพิจารณากับกลุ่มข้อมูลในชุดข้อมูลฝึกสอน

3) เรียงลำดับตามระยะห่าง จากน้อยไปมากและเลือกชุดข้อมูลจำนวน “k” ตัวแรก

4) พิจารณาข้อมูล “k” ตัวที่เลือกว่าส่วนใหญ่อยู่ในคำตอบหรือคลาสอะไร

5) กำหนดคำตอบหรือคลาสให้กับข้อมูลที่พิจารณาตามคะแนนเสียงข้างมากนั้น

จากลักษณะข้างต้น วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว มีความสามารถจัดการข้อมูลประเภทหมวดหมู่และการหาค่าถดถอยได้ดี แต่มีข้อเสียในกรณีที่ชุดข้อมูลมีขนาดใหญ่จะทำงานล่าช้า เนื่องจากต้องมีการคำนวณระยะทางสำหรับการทำนายผลลัพธ์ในแต่ละครั้ง

โดยงานวิจัยนี้จะใช้วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว ดังสมการที่ 2.2

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i| \qquad (2.2)$$

เมื่อ

$d(x,y)$ คือ ระยะทางจุด $x = (x_1, x_2, \dots, x_n)$ และ $y = (y_1, y_2, \dots, y_n)$

2.7 วิธีเบย์อย่างง่าย

วิธีเบย์อย่างง่าย (Naive Bayes) เป็นการเรียนรู้ของเครื่องที่ใช้ทฤษฎีความน่าจะเป็นของ Bayes หรือ Bayesian ในการอธิบายถึงความน่าจะเป็นของเหตุการณ์ตามข้อมูลที่มีอยู่ก่อน มีลักษณะสำคัญดังนี้ คือ

1) ทฤษฎีของเบย์ (Bayes' Theorem) คือ ทฤษฎีที่พูดถึงความน่าจะเป็นในการเกิดสิ่งใดสิ่งหนึ่ง ก็ต่อเมื่ออีกสิ่งที่ได้เกิดขึ้นแล้ว

2) การเรียนรู้ (Training) คือ อัลกอริทึมจะประเมินความน่าจะเป็นของข้อมูลคำตอบว่ามีจำนวนความถี่ที่อยู่ในชุดข้อมูลฝึกสอนว่ามีความถี่เท่าไหร่ รวมถึงคุณสมบัติของแต่ละค่าที่กำหนดให้เป็นข้อมูลคำตอบนั้น

3) การทำนาย (Prediction) คือ อัลกอริทึมจะคำนวณความน่าจะเป็นของข้อมูล โดยใช้ทฤษฎีความน่าจะเป็นของ Bayes หรือ Bayesian ซึ่งค่าที่ได้สูงสุดของข้อมูลที่นำเข้าจะถูกจัดอยู่ในคำตอบหรือคลาสนั้น

จากลักษณะข้างต้น วิธีเบย์อย่างง่าย มีความสามารถในการจัดการข้อมูลที่เป็นประเภทข้อความ เช่น การคัดกรองสแปม การวิเคราะห์ความคิดเห็นหรือการจัดประเภทเอกสาร

โดยงานวิจัยนี้จะใช้วิธีเบย์อย่างง่าย ดังสมการที่ 2.3 และ 2.4

$$P(y|x_1, \dots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, \dots, x_n)} \quad (2.3)$$

$$P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y))(1 - x_i) \quad (2.4)$$

เมื่อ

$P(y|x_1, \dots, x_n)$ คือ ความน่าจะเป็นของ y เมื่อกำหนด $x_1, \dots, x_n$

$P(y)$ คือ ความน่าจะเป็นของ y

$P(x_1, \dots, x_n)$ คือ ความน่าจะเป็นร่วมของ $x_1, \dots, x_n$

2.8 วิธีซัพพอร์ตเวคเตอร์แมชชีน

วิธีซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) เป็นการเรียนรู้ของเครื่องแบบมีผู้สอน ซึ่งใช้สำหรับการจัดหมวดหมู่และการหาค่าถดถอย อาศัยหลักการของการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ ซึ่งเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลที่ใกล้ที่สุด โดยเรียนรู้จากการนำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature Space) จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกันโดยจะสร้างเส้นแบ่ง (Hyperplane) ที่เป็นเส้นตรงขึ้นมา แล้วอัลกอริทึมจำทำการปรับเส้นตรงปรับปรุงเส้นแบ่งไปเรื่อย ๆ จนได้ เส้นแบ่งที่ดีที่สุด (Optimal Hyperplane) ขึ้นมา เพื่อแยกกลุ่มข้อมูลคำตอบให้ออกเป็นกลุ่มต่าง ๆ แยกออกจากกัน พอเวลาที่มีข้อมูลใหม่เข้าไปในอัลกอริทึม อัลกอริทึมจะดูว่าข้อมูลที่เข้าไปใหม่นั้น อยู่ภายใต้เส้นแบ่งที่ดีที่สุดเส้นใด จากนั้นจึงจะจัดข้อมูลต้นนั้นเข้าไปอยู่ในคำตอบหรือคลาสนั้น ๆ

จากลักษณะข้างต้น วิธีซัพพอร์ตเวคเตอร์แมชชีน มักใช้กับกับงานที่มีคุณสมบัติของข้อมูลเป็นจำนวนมากหรือข้อมูลที่ไม่ใช่ในเชิงเส้น (ผ่านฟังก์ชันเคอร์เนล) เช่น การจำแนกข้อความ การจดจำรูปภาพ ชีวสารสนเทศ หรือการวิเคราะห์ทางการเงิน

โดยงานวิจัยนี้จะใช้วิธีซัพพอร์ตเวคเตอร์แมชชีน ดังสมการที่ 2.5 และ 2.6

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \tag{2.5}$$

$$subject\ to\ y^T \alpha = 0 \text{ และ } 0 \leq a_i \leq C, i = 1, \ldots, n \tag{2.6}$$

เมื่อ

α คือ Dual Coefficients

$Q = [Q_{ij}]$ คือ Positive Semidefinite Matrix โดยที่ $Q_{ij} = y_i y_j K(x_i, x_j)$

$K(x_i, x_j)$ คือ Kernel Function

e คือ เวกเตอร์ที่มีสมาชิกเป็น 1 ทุกตัว

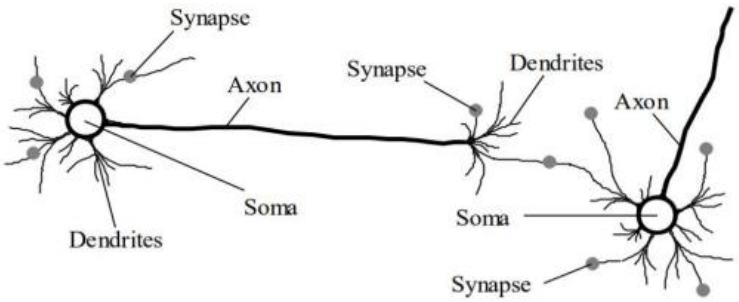
เมื่อหาค่า α_i จากสมการที่ 2.5 และ 2.6 จะได้ Optimal Hyperplane ดังสมการที่ 2.7

$$\sum_{i \in SV} y_i \, \alpha_i K(x_i, x) + b \tag{2.7}$$

2.9 วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น

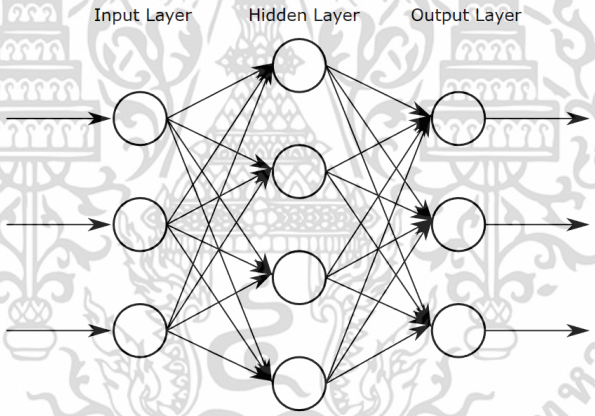
2.9.1 โครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม (Artificial Neural Networks) เป็นแบบจำลองทางคณิตศาสตร์ที่พยายามเลียนการทำงานของโครงข่ายระบบประสาทของมนุษย์ ประกอบไปด้วยเซลล์ประสาท (Soma หรือ Neuron), แขนงประสาท (Dendrite), ใยประสาท (Axon) และจุดประสานประสาท (Synapses) จะเทียบเท่าโหนด (Node), อินพุต (Input), เอาต์พุต (Output) และค่าถ่วงน้ำหนัก (Weight) ตามลำดับ โดยในโครงสร้างนั้นโหนดต่าง ๆ จะเชื่อมโยงกันผ่านแขนงประสาทที่เป็นส่วนที่นำข้อมูลจากโหนดหนึ่งไปสู่อีกโหนดหนึ่ง โดยเมื่อโหนดใด ๆ นั้นถูกกระตุ้นจนถึงระดับหนึ่งแล้วมันจะส่งสัญญาณไปยังโหนดถัดไป ซึ่งเลียนแบบการทำงานของโครงข่ายระบบประสาทของมนุษย์ แสดงได้ดัง รูปที่ 2.1



รูปที่ 2.1 โครงข่ายประสาทของมนุษย์

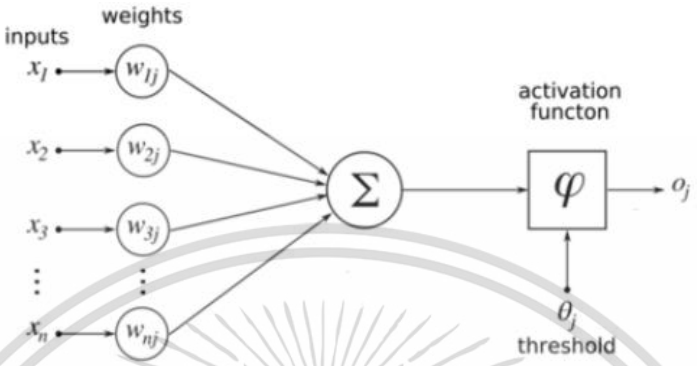
โครงข่ายประสาทเทียมนั้น สามารถแบ่งออกได้เป็น 3 ส่วน คือ ส่วนอินพุต (Input), ส่วนฮิดเดน (Hidden) และส่วนเอาต์พุต (Output) โดยจะมีส่วนฮิดเดนเป็นตัวกลางเชื่อมอยู่ระหว่างส่วนอินพุตและส่วนเอาต์พุต แสดงได้ดัง รูปที่ 2.2



รูปที่ 2.2 โครงข่ายประสาทเทียม

หลักการทำงานโครงข่ายประสาทเทียม คือ เมื่อมีอินพุต (Input) เข้ามายังโครงข่ายก็จะนำค่าอินพุตมาคูณกับค่าถ่วงน้ำหนัก (Weight) ของแต่ละโหนด ผลที่ได้จากอินพุตทุก ๆ โหนดจะนำมารวมกันแล้วเอามาเทียบกับค่าเกณฑ์ (Threshold) ที่กำหนดไว้ ในกรณีที่ผลรวมมีค่ามากกว่าค่าเกณฑ์แล้ว โหนดก็จะส่งค่าออกไป แต่ถ้าน้อยกว่าจะไม่มีการส่งค่า ทั้งนี้ส่วนสำคัญที่เราต้องทราบคือ ค่าถ่วงน้ำหนักและค่าเกณฑ์ เนื่องจากเป็นค่าที่ต้องให้ระบบทำการจดจำเพื่อใช้ในการเรียนรู้ซึ่งเป็นค่าที่ไม่แน่นอน แต่ได้จากการที่ระบบทำการปรับผ่านการประมวลผลกับชุดข้อมูลฝึกสอน โดยจะมีกระบวนการที่เรียกว่า “Back Propagation” เพื่อใช้ในการปรับปรุงค่าถ่วงน้ำหนัก

คือ หลังจากที่มีการประมวลผลกับชุดข้อมูลฝึกสอนแล้ว ค่าที่ส่งออกไปจะถูกเปรียบเทียบกับผลที่คาดหวัง แล้วคำนวณหาค่าความผิดพลาด ซึ่งค่าความผิดพลาดนี้ก็จะถูกส่งกลับไปยังโครงข่ายเพื่อใช้แก้ไขค่าถ่วงน้ำหนักในรอบต่อไป แสดงได้ดัง รูปที่ 2.3



รูปที่ 2.3 หลักการทำงานโครงข่ายประสาทเทียม

2.9.2 โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น

โครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (Multilayer Perceptron: MLP) จัดเป็นประเภทหนึ่งของโครงข่ายประสาทเทียม มีความแตกต่างกันที่ส่วนฮิดเดน (Hidden) จำนวนมากกว่า 1 ชั้นขึ้นไป แสดงได้ดัง รูปที่ 2.4 โดยมีส่วนประกอบที่สำคัญ ดังนี้

2.9.2.1 ส่วนอินพุต (Input Layer) คือ ข้อมูลอินพุตซึ่งอาจเป็นเวกเตอร์ของคุณสมบัติหรือค่าต่าง ๆ ที่จะนำเข้าสู่โครงข่าย

2.9.2.2 ส่วนฮิดเดน (Hidden Layer) คือ ส่วนที่อยู่ตรงกลางระหว่างส่วนอินพุตและส่วนเอาท์พุต ประกอบด้วยโหนดหลายตัวที่ใช้การแปลงข้อมูลแบบไม่เชิงเส้นกับข้อมูลอินพุต แล้วมีการคำนวณอินพุตเหล่านี้แบบถ่วงน้ำหนัก จากนั้นจะมีการรวมกับฟังก์ชันการใช้งาน เพื่อส่งไปยังส่วนฮิดเดนถัดไป

2.9.2.3 ฟังก์ชันการเปิดงาน (Activation Function) คือ ฟังก์ชันที่รวมเข้าไปกับอินพุตในช่วงที่อยู่ในส่วนฮิดเดน โดยสามารถเรียนรู้และประมวลผลความสัมพันธ์ระหว่างข้อมูลได้ เช่น ฟังก์ชันซิกมอยด์ (Sigmoid: Logistic), ไฮเปอร์โบลิกแทนเจนต์ (Hyperbolic Tangent: Tanh) หรือเรกไฟต์ลิเนียร์ยูนิต (Rectified Linear Unit: ReLU)

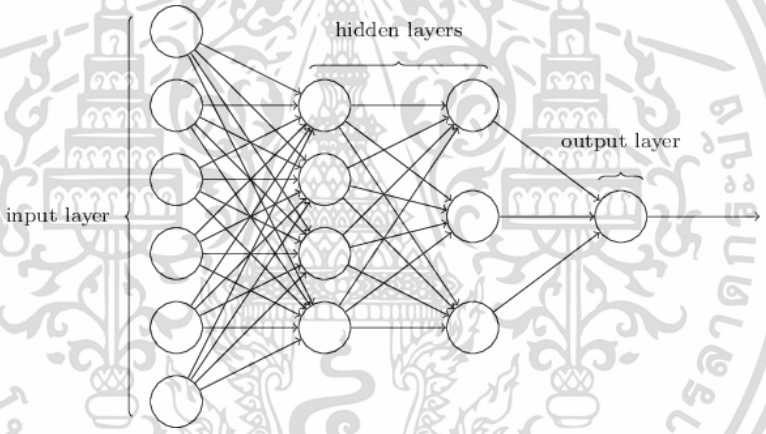
2.9.2.4 ค่าถ่วงน้ำหนัก (Weight) และค่าอคติ (Bias) คือ ค่าที่ปลี่ยนขอบเขตการตัดสินใจภายในโครงข่าย โดยหลังจากที่มีการคำนวณจะนำค่าเอามาเทียบกับค่าเกณฑ์

(Threshold) ที่กำหนดไว้ ในกรณีที่ผลรวมมีค่ามากกว่าค่าเกณฑ์แล้ว โหนดก็จะส่งค่าออกไป แต่ถ้าน้อยกว่าจะไม่มีการส่งค่า

2.9.2.5 ส่วนเอาต์พุต (Output Layer) คือ ส่วนสุดท้ายของโครงข่าย ใช้สร้างส่วนที่เป็นคำตอบของการเรียนรู้นั้น โดยจำนวนของเอาต์พุตจะอยู่กับประเภทของปัญหาที่กำลังแก้ไขและแทนกลุ่มคำตอบหรือคลาสในโครงข่ายนั้น

2.9.2.6 กระบวนการกระจายไปข้างหน้า (Forward Propagation) คือ การที่ข้อมูลไหลไปข้างหน้าผ่านโครงข่าย โดยจะรับอินพุตก่อนแล้วนำไปคำนวณร่วมกับ ค่าถ่วงน้ำหนัก ค่าอคติ และค่าฟังก์ชันการเปิดงาน จากนั้นก็ส่งผลลัพธ์ผ่านไปยังส่วนฮิดเดนชั้นต่อไป

2.9.2.7 กระบวนการกระจายย้อนกลับ (Back Propagation) คือ การที่ข้อมูลไหลย้อนกลับผ่านโครงข่าย โดยมีการปรับปรุงค่าถ่วงน้ำหนักและค่าอคติ เพื่อลดความแตกต่างระหว่างเอาต์พุตของเครือข่ายและอินพุตที่เข้าโครงข่าย



รูปที่ 2.4 หลักการทำงานโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น

โดยงานวิจัยนี้จะใช้วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น ดังสมการที่ 2.8

$$y_j = \varphi\left(\sum_{i=1}^{n} w_{ij}x_i + b\right) \tag{2.8}$$

เมื่อ

y_j คือ Output ของนิวรอนที่ j

$\varphi(\cdot)$ คือ Activation Function

w_{ij} คือ ค่าถ่วงน้ำหนักจากนิวรอน i ไปนิวรอน j

b คือ ค่าอคติ หรือ Bias

2.10 โมเดล Word2Vec

โมเดล Word2Vec เป็นโมเดลที่ใช้สำหรับการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) โดยภายในนั้นจะใช้โมเดลโครงข่ายประสาทเทียม (Artificial Neural Networks: ANN) ในการเรียนรู้ความสัมพันธ์ของคำต่าง ๆ จากแหล่งข้อมูล ซึ่งเมื่อผ่านการเรียนรู้แล้ว โมเดล Word2Vec จะสามารถทำการตรวจจับคำที่มีความหมายเหมือนกันหรือแนะนำคำที่ขาดหายไปของประโยคได้ มีหลักการทำงานเบื้องต้น คือ

1) การประมวลผลข้อมูลล่วงหน้า (Data Preprocessing) คือ การทำให้ข้อความเป็นคำ จากนั้นลบเครื่องหมายหรืออักขระที่ไม่จำเป็นต่อการประมวลผลออก

2) การสร้างคำศัพท์ (Vocabulary Creation) คือ การกำหนดค่าเวกเตอร์ของคำที่เป็นเอกลักษณ์ไม่ซ้ำกับคำอื่น โดยนำไปเข้ากับโมเดล Continuous Bag of Words (CBOW) และ Skip-gram ตามลำดับเพื่อทำการสร้างเวกเตอร์ที่ไม่ซ้ำกับคำอื่น ๆ ขึ้น รวมทั้งในกรณีที่คำที่มีความหมายใกล้เคียงกัน เวกเตอร์ที่ได้ก็จะใกล้เคียงกันตามไปด้วย

2.11 โมเดลเบิร์ตแบบหลายภาษา

โมเดลเบิร์ต (Bidirectional Encoder Representations from Transformers: BERT) เป็นโมเดลที่พัฒนาโดย Google ออกแบบมาเพื่อเรียนรู้ข้อความแบบที่ไม่มีการกำหนดคำตอบ (Unlabeled) ซึ่งใช้งานอย่างแพร่หลายสำหรับงานประมวลผลภาษาธรรมชาติ เช่น การจัดหมวดหมู่ประโยค (Sentence Classification), การจดจำชื่อเอนทิตี (Named Entity Recognition) หรือการตอบคำถาม (Question Answering)

โมเดลเบิร์ตแบบหลายภาษา (Multilingual BERT) คือ โมเดลเบิร์ตที่ได้รับการเรียนรู้ในหลากหลายภาษา เพื่อที่จะสามารถใช้งานในภาษาอื่น ๆ นอกจากภาษาอังกฤษได้ โดยเรียนรู้

ผ่านคลังข้อมูลที่มีขนาดใหญ่และหลากหลายภาษาซึ่งมีข้อดี คือ ไม่จำเป็นต้องพัฒนาโมเดลไว้สำหรับแต่ละภาษาโดยเฉพาะ เนื่องจากมีการเรียนรู้ในภาษาพร้อมกันหลากหลายภาษา ทำให้ใช้งานได้ง่าย ซึ่งภายในโมเดลจะรองรับทั้งหมด 104 ภาษา 12 เลเยอร์ (ชั้น) และส่วนฮิดเดนจำนวน 768 โหนด แสดงได้ดัง รูปที่ 2.5 โดยมีภาพรวมของการดำเนินงานดังนี้ คือ

1) การเรียนรู้ล่วงหน้า (Pretraining) คือ โมเดลจะได้รับการเรียนรู้จากข้อมูลในคลังข้อมูลเป็นจำนวนมากและหลากหลายภาษา ซึ่งกรณีที่เป็นข้อความจะถูกแบ่งออกเป็นคำ ๆ โดยใช้วิธีการที่เรียกว่า “WordPiece Tokenization” ซึ่งช่วยให้โมเดลสามารถจัดการคำที่ไม่อยู่ในคำศัพท์และรูปแบบทางภาษาศาสตร์ได้

2) การสร้างแบบจำลองภาษามาสก์ (Masked Language Modeling) คือ ระหว่างการเรียนรู้ล่วงหน้า โมเดลจะทำการสุ่มอินพุตบางส่วนขึ้นมาแล้วทำการเรียนรู้การทำนายผลลัพธ์โดยซ่อนบริบทโดยรอบ เป้าหมายเพื่อที่จะทำนายคำโดยพิจารณาจากบริบทของคำที่อยู่รอบ ๆ ที่เหลืออยู่และเป็นการเรียนรู้แบบ 2 ทิศทาง ซึ่งหมายถึงทั้งทางซ้ายและทางขวาของคำ

3) การทำนายประโยคถัดไป (Next Sentence Prediction) คือ โมเดลจะได้รับการเรียนรู้ให้ทำนายว่าประโยต 2 ประโยคถัดจากตัวอย่างจะเป็นอะไร เพื่อให้เข้าใจในความสัมพันธ์ระหว่างประโยคและความสอดคล้องกัน

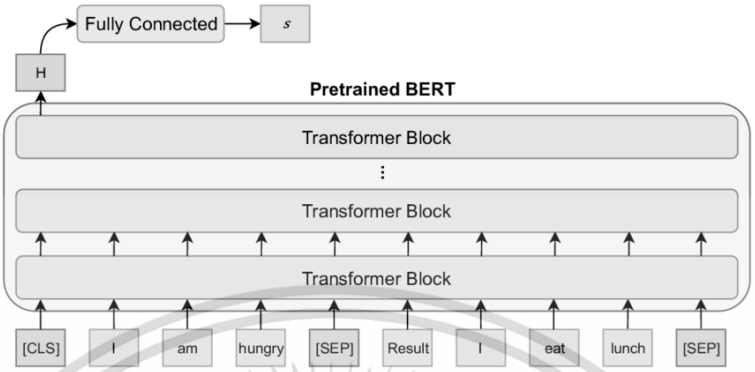
4) การเรียนรู้โดยใช้สถาปัตยกรรม Transformer (Transformer Architecture Training) คือ การเรียนรู้โดยใช้สถาปัตยกรรม Transformer ประกอบด้วยโครงข่ายประสาทเทียมแบบฟีดฟอร์เวิร์ดหลายชั้น (Feed-Forward Neural Networks) ซ้อนกัน ซึ่งแต่ละเลเยอร์ (ชั้น) จะประมวลผลอินพุตอย่างเป็นอิสระและสามารถหาความสัมพันธ์ระหว่างคำทั้งหมดในลำดับทั้งก่อนหน้าและตามหลังคำที่กำหนดได้

5) การปรับแต่งอย่างละเอียด (Fine-Tuning) คือ หลักจากที่มีการเรียนรู้ โมเดลจะทำการปรับแต่งพารามิเตอร์ให้รองรับงานแบบเฉพาะเจาะจงได้ โดยเรียนรู้ข้อมูลที่มีรายละเอียดแบบที่มีความเฉพาะเจาะจงมากยิ่งขึ้น

จากลักษณะข้างต้น โมเดลเบิร์ตแบบหลายภาษา จึงมักใช้ในงานด้านภาษาศาสตร์กันอย่างแพร่หลาย เนื่องจากเหตุผลดังกล่าว คือ

- ความครอบคลุมหลายภาษา เนื่องจากโมเดลได้รับการเรียนรู้ในหลากหลายภาษา ซึ่งช่วยในกรณีที่เกิดข้อมูลในแต่ละภาษามีจำกัด
- ลดความซ้ำซ้อนของโมเดล เนื่องจาก ไม่ต้องทำการเรียนรู้โมเดลในแต่ละภาษาโดยเฉพาะเจาะจง จึงทำให้สามารถใช้โมเดลเดียวในการทำงานกับหลากหลายภาษาได้

- ประสิทธิภาพของทรัพยากร เนื่องจากโมเดลจะช่วยลดความต้องการทรัพยากรของการเรียนรู้หลากหลายภาษาพร้อมกัน ทำให้เป็นประโยชน์กรณีที่มีทรัพยากรจำกัด



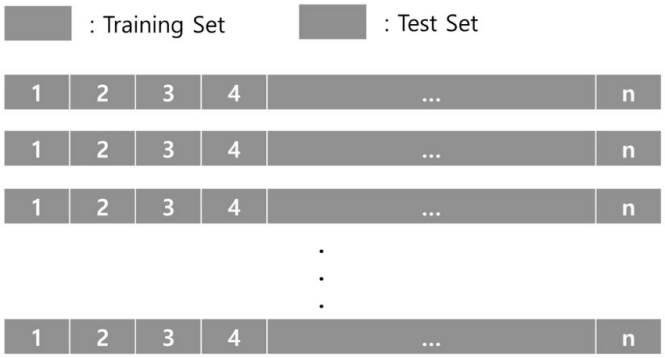
รูปที่ 2.5 หลักการทำงานโมเดลเบิร์ตแบบหลายภาษา

2.12 การวัดประสิทธิภาพแบบดึงทีละตัว

การวัดประสิทธิภาพแบบดึงทีละตัว (Leave-One-Out: LOO) เป็นเทคนิคที่ใช้ในการเรียนรู้ของเครื่องและสถิติ เพื่อประเมินประสิทธิภาพของโมเดลในชุด ซึ่งมีขั้นตอนการดำเนินงานดังนี้ คือ

1) ละจุดข้อมูลหนึ่งจุดออกจากชุดข้อมูล
2) ใช้จุดข้อมูลที่เหลือนำเข้าโมเดลเพื่อทำการเรียนรู้
3) นำจุดข้อมูลหนึ่งจุดที่ถูกละไว้เป็นตัวทดสอบความแม่นยำ
4) ประเมินผลการทดสอบความแม่นยำ
5) ละจุดข้อมูลเป็นจุดข้อมูลถัดไปในชุดข้อมูล แล้วดำเนินการใหม่ตั้งแต่ข้อ 1)

ในส่วนของการวัดประสิทธิภาพโดยรวมนั้นสามารถดำเนินการ โดยใช้การคำนวณหาค่าเฉลี่ยหรือผลรวมของการประเมินในแต่ละรายการ ซึ่งมีข้อดีในชุดข้อมูลที่มีขนาดเล็กแต่จะใช้ทรัพยากรมากขึ้นในกรณีที่ชุดข้อมูลมีขนาดใหญ่ เนื่องจากต้องทำการฝึกโมเดลและทดสอบตามจำนวนข้อมูลที่มีแสดงได้ดัง รูปที่ 2.6



รูปที่ 2.6 หลักการทำงานการวัดประสิทธิภาพแบบดึงทีละตัว

2.13 งานวิจัยที่เกี่ยวข้อง

ระบบถามตอบ (Question Answering) นั้น ในปัจจุบันเริ่มเป็นที่พูดถึงและเริ่มมีการใช้งานกันอย่างแพร่หลาย ทั้งในมุมมองในด้านภาษาอื่น ๆ ที่ไม่ใช่ภาษาอังกฤษ โดยในปี ค.ศ. 2020 Nguyen Thi Mai Trang และ Maxim Shcherbakov ได้ดำเนินการวิจัยระบบถามตอบร่วมกับภาษาเวียดนาม โดยใช้โมเดลเบิร์ตแบบหลายภาษาในการเรียนรู้ภาษาเวียดนามแล้วแปลงโมเดลเบิร์ตนั้นให้เป็นโมเดลเบิร์ตที่รองรับภาษาเวียดนามอย่างเดียว ซึ่งพบว่าค่าความแม่นยำของโมเดลเบิร์ตที่รองรับภาษาเวียดนามอย่างเดียวนั้นมีมากกว่าโมเดลเบิร์ตแบบหลายภาษาหรือในปี ค.ศ. 2021 Bineet Kumar Jha และพวก ได้ทำการประยุกต์ให้เข้ากับภาษาอินเดียจำนวน 12 รูปแบบทำให้สามารถใช้งานรูปแบบภาษาอินเดียที่มากขึ้นและโมเดลเบิร์ตนั้น สามารถนำมาประยุกต์ใช้ในภาษาที่หลากหลายเช่นกัน ซึ่งในบางกรณีกลับพบว่าการใช้งานโมเดลเบิร์ตนั้น อาจได้ผลลัพธ์ที่ไม่ได้คาดหวังเช่นเดียวกัน โดย Narayana Darapaneni และพวก ได้ทำการวิจัยในการประยุกต์โมเดลเบิร์ตร่วมกับระบบถามตอบในเรื่องข่าว พบว่าความแม่นยำได้น้อยกว่าความแม่นยำที่ได้จากมนุษย์หรือในปี ค.ศ. 2022 Jie Yin ได้นำโมเดลเบิร์ตร่วมกับวิธีการที่เรียกว่า “co-attention” และ “self-attention” ไปพัฒนาร่วมกัน ผลที่ได้พบว่าได้ค่า F_1 Score อยู่ที่ 58.90% แต่ถึงอย่างไรนั้นก็ยังมีการนำโมเดลเบิร์ตไปใช้ร่วมกับระบบถามต่อประเภทอื่นอีก เช่น ใช้ในงานด้านสาธารณสุขโดย Eslam Amer และพวกได้นำไปใช้ร่วมกับระบบถามตอบของโรค COVID-19 พบว่าได้ค่าความแม่นยำถึง 96% ซึ่งถือว่ามีความแม่นยำ ทั้งนี้ในระบบถามตอบ ควรที่จะมีประยุกต์ร่วมกับการเรียนรู้ของเครื่องหรือการเรียนรู้เชิงลึก เช่น Yihan Bian และ Kaiwen Peng ได้อธิบายถึงสถาปัตยกรรมการทำงานของระบบการตอบคำถามว่าแต่ละแบบมีรายละเอียดเป็นอย่างไร เพื่อให้ได้ผลลัพธ์ที่ดีขึ้น ในการใช้งานร่วมกับระบบงานภายนอก โดย Nikita และพวก ได้ทำการนำระบบถามตอบ ซึ่งใช้โมเดล

เบิร์ตร่วมกับระบบ Google Dialogflow เพื่อพัฒนาเป็นแชทบอท ทำให้ผู้คนสามารถใช้งานได้ง่ายยิ่งขึ้น

ทั้งนี้ภายในระบบถามตอบ ส่วนที่สำคัญอย่างยิ่งส่วนหนึ่ง คือ ส่วนที่เรียกว่าส่วนการจำแนกคำถาม (Question Classification) เนื่องจากถ้าหากว่าระบบถามตอบทราบว่าได้มีการโต้ตอบอยู่ในกลุ่มคำถามใด ซึ่งจะทำให้การโต้ตอบนั้นอยู่ในหมวดหมู่เดียวกัน โดยในปี ค.ศ. 2022 Guanhyi และพวกได้นำโมเดลเบิร์ตไปใช้ในส่วนการจำแนกคำถาม เพื่อโต้ตอบกันในเรื่องของกฎหมาย โดยใช้ภาษามองโกเลีย พบว่าค่า F_1 Score มีค่าถึง 86.98% โดยในส่วนของการจำแนกคำถาม ในรูปแบบภาษาไทย Saranlita และพวก ได้ทำการเพิ่มขั้นตอนที่เรียกว่า Part-of-Speech ก่อนที่จะมีการใช้งานโมเดลเบิร์ตแบบพื้นฐานซึ่งพบว่ามีค่า F_1 Score อยู่ที่ 91.40%

ตารางที่ 2.3 สรุปข้อแตกต่างระหว่างงานวิจัยที่เกี่ยวข้อง

งานวิจัย	จุดเด่น	จุดด้อย
Mongolian Questions Classification in the Law Domain	ค่า F_1 Score มีค่าถึง 86.98%	ใช้โมเดลเบิร์ตแบบพื้นฐานในการประยุกต์ใช้กับภาษามองโกเลีย ไม่ใช่โมเดลเบิร์ตแบบหลายภาษา
Vietnamese Question Answering System from Multilingual BERT Models to Monolingual BERT Model	มีการประยุกต์ใช้โมเดลเบิร์ตให้เข้ากับภาษาเวียดนามอย่างเดียว จึงทำให้ค่า F_1 Score นั้นสูงกว่าในกรณีที่เป็นโมเดลเบิร์ตแบบหลายภาษา	ในกรณีที่มีต้องมีการใช้งานหลายภาษาพร้อมกัน จะเป็นผลให้ความแม่นยำน้อยกว่าโมเดลเบิร์ตแบบหลายภาษา
Question Answering System with Indic multilingual-BERT	มีการประยุกต์ใช้โมเดลเบิร์ตแบบหลายภาษาให้เข้ากับรูปแบบของภาษาอินเดียจำนวน 12 รูปแบบ ทำให้สามารถรองรับภาษาอินเดียได้หลายรูปแบบ	ในกรณีที่มีรูปแบบอื่นที่เกินจากการเรียนรู้แล้ว จะส่งผลให้ความแม่นยำในการทำนายผลลัพธ์นั้นลดลง
A Proposed Chatbot Framework for COVID-19	ค่าความแม่นยำ มีค่าถึง 96%	มีการใช้ในรูปแบบของภาษาอังกฤษอย่างเดียว

Building a Question and Answer System for News Domain	เมื่อเทียบกับโมเดลในงานวิจัย พบว่าโมเดลเบิร์ตให้ผลค่า F_1 Score ดีกว่าโมเดลอื่น ๆ ในงานวิจัย	ค่า F_1 Score ที่ได้กลับน้อยว่าค่า F_1 Score ที่ได้จากมนุษย์
Question Answering System Analysis Based on Machine Learning	พูดถึงสถาปัตยกรรมการทำงานของระบบการตอบคำถามว่าแต่ละแบบมีรายละเอียดเป็นอย่างไร	ไม่มีการวัดประสิทธิภาพระหว่างโมเดล
Question Answering Model Based Conversational Chatbot using BERT Model and Google Dialogflow	พูดถึงขั้นตอนในส่วนของการประยุกต์ใช้โมเดลเบิร์ต ร่วมกับ Google Dialogflow เพื่อพัฒนาเป็นแชทบอท	ไม่มีการวัดประสิทธิภาพระหว่างโมเดล
Research on Question Answering System Based on BERT Model	มีการประยุกต์ใช้โมเดลเบิร์ตให้เข้ากับวิธีการทำงานที่เรียกว่า “co-attention” และ “self-attention” เพื่อค้นหาส่วนย่อหน้าและสร้างคำตอบเพื่อโต้ตอบกลับไปยังผู้ใช้	ค่า F_1 Score มีค่าแค่ 58.90%
Question Classification from Thai Sentences by Considering Word Context to Question Generation	มีการใช้ Part-of-Speech (POS) ในการระบุชนิดของคำตามหน้าที่ก่อนการใช้งานร่วมกับโมเดลเบิร์ต	ใช้โมเดลเบิร์ตแบบพื้นฐานในการประยุกต์ใช้กับภาษาไทย ไม่ใช่โมเดลเบิร์ตแบบหลายภาษา

บทที่ 3

วิธีดำเนินการวิจัย

3.1 การเตรียมข้อมูล

ในขั้นตอนการเตรียมข้อมูลนี้จะใช้ข้อมูลจากคลังข้อมูลถามตอบภาษาไทยของศูนย์เทคโนโลยีอีเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (The National Electronics and Computer Technology Center: NECTEC) โดยเป็นข้อมูลประเภทถามตอบ เก็บข้อมูลมาจากเว็บไซต์ Wikipedia ภาษาไทย ประกอบด้วยข้อความคำถามและคำตอบ รวมถึงรายละเอียดที่เกี่ยวข้อง จำนวน 4,000 ตัวอย่าง ใน 2,289 หมวดหมู่ประเภทของคำถาม แสดงได้ดัง รูปที่ 3.1 โดยทำเป็นไฟล์ประเภท JSON แล้วใช้ดำเนินการในงานวิจัย ซึ่งภายในประกอบด้วย

3.1.1 question_id แทน รหัสคำถาม (ไม่ซ้ำกัน)

3.1.2 question แทน ข้อความคำถาม

3.1.3 answer แทน ข้อความคำตอบ

3.1.4 answer_begin_position แทน ตำแหน่งเริ่มต้นของข้อความคำตอบ

3.1.5 answer_end_position แทน ตำแหน่งสิ้นสุดของข้อความคำตอบ

3.1.6 article_id แทน หมวดหมู่คำถาม (กรณีที่เลขเดียวกัน หมายถึงอยู่ในหมวดเดียวกัน)


```
{
  "question_id": 1,
  "question": "สุนัขตัวแรกรับบทเป็นเบนจี้ในภาพยนตร์เรื่อง Benji ที่ออกฉายในปี พ.ศ. 2517 มีชื่อว่าอะไร",
  "answer": "ฮิกกินส์",
  "answer_begin_position ": 529,
  "answer_end_position": 538,
  "article_id": 115035
},
```


รูปที่ 3.1 ตัวอย่างข้อมูลจากคลังข้อมูลถามตอบภาษาไทย

งานวิจัยนี้จะใช้หัวข้อ “article_id” แทนคำตอบหรือคลาสของชุดข้อมูลฝึกสอน จากนั้นทำการจัดกลุ่มข้อมูลโดยแบ่งออกเป็น 3 กลุ่ม โดยใช้เกณฑ์จากจำนวนคำถามและคำตอบที่อยู่ในหมวดหมู่คำถามเดียวกัน ตั้งแต่ 3-5 คำถามและคำตอบขึ้นไป (ในกรณีอื่นไม่ได้ใช้งาน เนื่องจากจำนวนคำตอบหรือคลาสของชุดข้อมูลฝึกสอนที่ได้จะได้น้อยเกินไปต่อการดำเนินการวิจัย) จากนั้นจึงนำข้อมูลที่ได้เข้าสู่กระบวนการเรียนรู้ ซึ่งกลุ่มทั้ง 3 กลุ่มมีรายละเอียดดังนี้ คือ

- กลุ่ม article_id ที่มี หมวดหมู่คำถามเดียวกันตั้งแต่ 3 คำถามขึ้นไป หมายถึง การแบ่งกลุ่มข้อมูลซึ่งใช้เกณฑ์จากจำนวนคำถามและคำตอบที่อยู่ในหมวดหมู่ article_id เดียวกันตั้งแต่ 3 คำถามและคำตอบขึ้นไป โดยตัวอย่างแสดงได้ดัง รูปที่ 3.2


```
{
  "question_id": 16,
  "question": "ท่าอากาศยานสุโขทัย เปิดทำการครั้งแรกเมื่อไร",
  "answer": "วันที่ 12 เมษายน พ.ศ. 2539",
  "answer_begin_position ": 172,
  "answer_end_position": 198,
  "article_id": 152042
},
{
  "question_id": 2791,
  "question": "ท่าอากาศยานสุโขทัยเปิดทำการเมื่อปี พ.ศ. ใด",
  "answer": "2539",
  "answer_begin_position ": 194,
  "answer_end_position": 198,
  "article_id": 152042
},
{
  "question_id": 3905,
  "question": "การออกแบบอาคารผู้โดยสารของท่าอากาศยานสุโขทัยเป็นลักษณะใด",
  "answer": "ทรงไทย",
  "answer_begin_position ": 427,
  "answer_end_position": 433,
  "article_id": 152042
},
```


รูปที่ 3.2 ตัวอย่างข้อมูลจากคลังฯ ในหมวดหมู่เดียวกันตั้งแต่ 3 คำถามขึ้นไป

จากรูปที่ 3.2 แสดงถึงกลุ่มคำถามและคำตอบที่อยู่ในหมวดหมู่คำถาม หรือ article_id เดียวกันโดย article_id ที่ 152042 แทนกลุ่มคำถามที่เกี่ยวข้องกับ “ท่าอากาศยานสุโขทัย” ซึ่งมีอยู่ตั้งแต่ 3 คำถามขึ้นไป และเมื่อใช้เกณฑ์นี้เราจะแบ่งหมวดหมู่คำถามออกมาได้ทั้งหมด 102 กลุ่ม ซึ่งค่าที่ได้แทนหมายเลข article_id แสดงได้ดัง รูปที่ 3.3


```
dict_keys([
  152042, 74213, 299739, 153987, 873656, 485468, 888869, 28335, 48391, 139197,
  571724, 41334, 48127, 86461, 134557, 348996, 46482, 327130, 58404, 587901,
  2051, 41074, 53185, 141417, 5449, 224450, 479762, 285056, 574000, 605272,
  198596, 945001, 299618, 224258, 565618, 142307, 305470, 285522, 17368, 619806,
  423130, 616556, 228438, 602733, 211448, 55059, 43559, 144434, 604036, 47992,
  630062, 179218, 179217, 105416, 59741, 172366, 50876, 233396, 171262, 514247,
  218354, 850651, 287943, 116956, 194880, 574014, 616563, 40396, 594113, 34446,
  52044, 407314, 550489, 20246, 306886, 76722, 465021, 133595, 291591, 742390,
  149647, 18078, 142353, 225020, 846532, 130001, 25143, 73830, 78353, 3898,
  6130, 437712, 10334, 209298, 121402, 125814, 27907, 276207, 182497, 6697,
  246918, 87205
])
```


รูปที่ 3.3 ตัวอย่างหมวดหมู่คำถามจากคลังฯ ในหมวดหมู่เดียวกันตั้งแต่ 3 คำถามขึ้นไป

- กลุ่ม article_id ที่มี หมวดหมู่คำถามเดียวกันตั้งแต่ 4 คำถามขึ้นไป หมายถึง การแบ่งกลุ่มข้อมูลซึ่งใช้เกณฑ์จากจำนวนคำถามและคำตอบที่อยู่ในหมวดหมู่ article_id เดียวกันตั้งแต่ 4 คำถามและคำตอบขึ้นไป โดยตัวอย่างแสดงได้ดัง รูปที่ 3.4


```
{
  "question_id": 178,
  "question": "ละครเวทีของถกลเกียรติ วีรวรรณ ที่จุฑาภัทร เหล่าธรรมทัศน์ได้เล่นมีชื่อเรื่องว่าอะไร",
  "answer": "บัลลังก์เมฆ",
  "answer_begin_position ": 284,
  "answer_end_position": 295,
  "article_id": 299739
},
{
  "question_id": 1497,
  "question": "จุฑาภัทร เหล่าธรรมทัศน์ นักร้องชาวไทย เกิดเมื่อวันที่เท่าไร",
  "answer": "9",
  "answer_begin_position ": 163,
  "answer_end_position": 164,
  "article_id": 299739
},
{
  "question_id": 3248,
  "question": "ผลงานเพลงชุดแรกของ จุฑาภัทร เหล่าธรรมทัศน์ มีชื่อว่าอะไร",
  "answer": "พลับ",
  "answer_begin_position ": 427,
  "answer_end_position": 431,
  "article_id": 299739
},
{
  "question_id": 3249,
  "question": "จุฑาภัทร เหล่าธรรมทัศน์ หรือน้องพลับ เคยเล่นละครเวทีเรื่องใด",
  "answer": "บัลลังก์เมฆ",
  "answer_begin_position ": 284,
  "answer_end_position": 295,
  "article_id": 299739
},
```


รูปที่ 3.4 ตัวอย่างข้อมูลจากคลังฯ ในหมวดหมู่เดียวกันตั้งแต่ 4 คำถามขึ้นไป

จากรูปที่ 3.4 แสดงถึงกลุ่มคำถามและคำตอบที่อยู่ในหมวดหมู่คำถาม หรือ article_id เดียวกันโดย article_id ที่ 299739 แทนกลุ่มคำถามที่เกี่ยวข้องกับ “จุฑาภัทร เหล่าธรรมทัศน์” ซึ่งมีอยู่ตั้งแต่ 4 คำถามขึ้นไป และเมื่อใช้เกณฑ์นี้เราจะแบ่งหมวดหมู่คำถามออกมาได้ทั้งหมด 81 กลุ่ม ซึ่งค่าที่ได้แทนหมายเลข article_id แสดงได้ดัง รูปที่ 3.5

```
dict_keys([
  299739, 153987, 873656, 28335, 134557, 348996, 46482, 327130, 58404, 587901,
  41074, 53185, 141417, 224450, 479762, 285056, 574000, 605272, 198596, 945001,
  299618, 224258, 565618, 142307, 305470, 285522, 17368, 619806, 423130, 616556,
  228438, 602733, 211448, 55059, 43559, 144434, 604036, 47992, 630062, 179218,
  179217, 105416, 59741, 172366, 50876, 233396, 171262, 514247, 218354, 850651,
  287943, 116956, 194880, 574014, 616563, 40396, 594113, 34446, 52044, 407314,
  550489, 20246, 306886, 76722, 465021, 133595, 291591, 742390, 149647, 142353,
  225020, 846532, 130001, 73830, 6130, 10334, 121402, 125814, 27907, 182497,
  6697
])
```

รูปที่ 3.5 ตัวอย่างหมวดหมู่คำถามจากคลังฯ ในหมวดหมู่เดียวกันตั้งแต่ 4 คำถามขึ้นไป

- กลุ่ม article_id ที่มี หมวดหมู่คำถามเดียวกันตั้งแต่ 5 คำถามขึ้นไป หมายถึง การแบ่งกลุ่มข้อมูลซึ่งใช้เกณฑ์จากจำนวนคำถามและคำตอบที่อยู่ในหมวดหมู่ article_id เดียวกันตั้งแต่ 5 คำถามและคำตอบขึ้นไป โดยตัวอย่างแสดงได้ดัง รูปที่ 3.6


```
{
  "question_id": 657,
  "question": "หม่อมราชวงศ์จัตุมงคล โสณกุล เป็นโอรสของใคร",
  "answer": "พลตรี หม่อมเจ้าฉัตรมงคล โสณกุล",
  "answer_begin_position ": 476,
  "answer_end_position": 506,
  "article_id": 46482
},
{
  "question_id": 786,
  "question": "หม่อมราชวงศ์จัตุมงคล โสณกุล สมรสครั้งที่สองกับใคร",
  "answer": "คุณหญิงบุลย์วิภา โสณกุล ณ อยุธยา",
  "answer_begin_position ": 1399,
  "answer_end_position": 1431,
  "article_id": 46482
},
{
  "question_id": 787,
  "question": "พระบิดาของหม่อมราชวงศ์จัตุมงคล โสณกุล มีพระนามว่าอะไร",
  "answer": "พลตรี หม่อมเจ้าฉัตรมงคล โสณกุล",
  "answer_begin_position ": 476,
  "answer_end_position": 506,
  "article_id": 46482
},
{
  "question_id": 891,
  "question": "หม่อมราชวงศ์จัตุมงคล โสณกุล เป็นพระโอรสของผู้ใด",
  "answer": "พลตรี หม่อมเจ้าฉัตรมงคล โสณกุล",
  "answer_begin_position ": 476,
  "answer_end_position": 506,
  "article_id": 46482
},
{
  "question_id": 892,
  "question": "หม่อมราชวงศ์จัตุมงคล สมรสครั้งที่สองกับผู้ใด",
  "answer": "คุณหญิงบุลย์วิภา โสณกุล ณ อยุธยา",
  "answer_begin_position ": 1399,
  "answer_end_position": 1431,
  "article_id": 46482
},
```


รูปที่ 3.6 ตัวอย่างข้อมูลจากคลังฯ ในหมวดหมู่เดียวกันตั้งแต่ 5 คำถามขึ้นไป

จากรูปที่ 3.6 แสดงถึงกลุ่มคำถามและคำตอบที่อยู่ในหมวดหมู่คำถาม หรือ article_id เดียวกันโดย article_id ที่ 46482 แทนกลุ่มคำถามที่เกี่ยวข้องกับ “หม่อมราชวงศ์จัตุมงคล โสณกุล” ซึ่งมีอยู่ตั้งแต่ 5 คำถามขึ้นไป และเมื่อใช้เกณฑ์นี้เราจะแบ่งหมวดหมู่คำถามออกมาได้ทั้งหมด 11 กลุ่ม ซึ่งค่าที่ได้แทนหมายเลข article_id แสดงได้ดังรูปที่ 3.7

```
dict_keys([
  46482, 224450, 285056, 17368, 233396, 171262, 20246, 465021, 291591, 742390,
  130001
])
```

รูปที่ 3.7 ตัวอย่างหมวดหมู่คำถามจากคลังฯ ในหมวดหมู่เดียวกันตั้งแต่ 5 คำถามขึ้นไป

3.2 การตัดคำภาษาไทย

ในภาษาไทยนั้น คำแต่ละคำในประโยคมักจะเขียนติดต่อกัน ดังนั้นก่อนที่จะนำประโยคเข้าไปประมวลผลจะต้องมีการตัดประโยคนั้นออกเป็นคำแต่ละคำเสียก่อน ซึ่งมีหลากหลายวิธีด้วยกัน เช่น การใช้พจนานุกรมช่วยในการตัดคำโดยมีแนวคิด คือ การใช้คำในพจนานุกรมในการเปรียบเทียบกับประโยค หากพบก็จะสามารถตัดคำนั้นออกมาได้เลย มีข้อเสีย คือ หากไม่มีคำนั้นในพจนานุกรมจะมีสามารถตัดคำนี้ได้หรือใช้วิธีการทางด้านปัญญาประดิษฐ์ เช่น Library ที่สร้างจากการเรียนรู้เชิงลึก เช่น Deepcut ในการตัดคำ โดยมีแนวคิด คือ ใช้เทคนิคการเรียนรู้ของเครื่องในการประมวลผล วิธีการนี้จะไม่ต้องใช้พจนานุกรมในการเปรียบเทียบ แต่จะต้องมีชุดการฝึกสอนให้ระบบทำการเรียนรู้ก่อน จึงจะสามารถใช้งานได้หรืออัลกอริทึมที่ใช้การนับคำในประโยคต่าง ๆ เป็นต้น

โดยงานวิจัยนี้ผู้วิจัยจะใช้ Library ของภาษา Python ที่ชื่อ “PyThaiNLP” ซึ่งเป็น Library ไว้สำหรับการตัดคำภาษาไทย โดยใช้เทคนิคที่เรียกว่า “Maximum Matching Algorithm” ในการแบ่งคำภาษาไทยออกจากประโยคด้วยวิธีการ คือ การตัดคำแบบสอดคล้องมากที่สุด ด้วยการหาวิธีในการตัดคำที่สามารถเป็นไปได้ทั้งหมดก่อน จากนั้นให้เลือกข้อความที่แบ่งแล้วมีจำนวนคำน้อยที่สุด ซึ่งในกรณีที่มีจำนวนคำเท่ากัน ให้เลือกวิธีการตัดคำแบบยาวที่สุดเข้ามาช่วย แสดงได้ดัง รูปที่ 3.8

```
from pythainlp import word_tokenize

text = "ก็จะรู้ความชั่วร้ายที่ทำไว้      และคงจะไม่ยอมให้ทำนาบนหลังคน "

print("default (newmm):")
print(word_tokenize(text))
print("\nnewmm and keep_whitespace=False:")
print(word_tokenize(text, keep_whitespace=False))
```

```
default (newmm):
['ก็', 'จะ', 'รู้ความ', 'ชั่วร้าย', 'ที่', 'ทำ', 'ไว้', '      ', 'และ', 'คงจะ', 'ไม่', 'ยอมให้', 'ทำนาบนหลังคน', ' ']

newmm and keep_whitespace=False:
['ก็', 'จะ', 'รู้ความ', 'ชั่วร้าย', 'ที่', 'ทำ', 'ไว้', 'และ', 'คงจะ', 'ไม่', 'ยอมให้', 'ทำนาบนหลังคน']
```

รูปที่ 3.8 ตัวอย่างการตัดคำด้วยวิธี Maximum Matching Algorithm

3.3 การแทนคำภาษาไทยด้วยเวกเตอร์

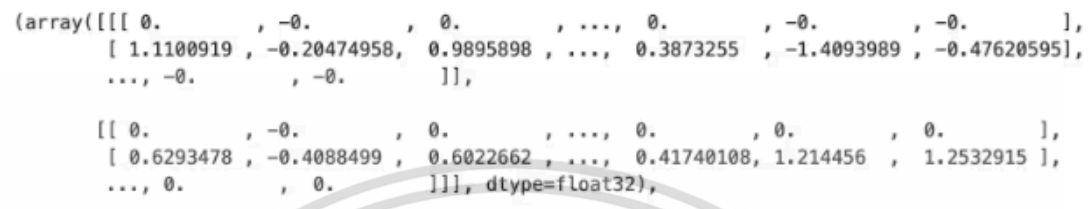
หลักจากที่ได้ตัดคำภาษาไทยเป็นคำแต่ละคำออกเป็นที่เรียบร้อยแล้ว ขั้นตอนต่อไป คือ การแปลงคำที่ได้ให้อยู่ในรูปของเวกเตอร์ โดยงานวิจัยนี้ผู้วิจัยจะใช้วิธีการตัดคำภาษาไทย 2 แบบ แล้วนำมาเปรียบเทียบกัน โดยมี

3.3.1 วิธีการโมเดล Thai2Vec คือ โมเดล Word2Vec ที่ทำการแปลง “คำ” ให้อยู่ในรูปของ “ตัวเลขหรือเวกเตอร์” โดยทำให้รองรับอยู่ในรูปแบบภาษาไทย แสดงได้ดัง รูปที่ 3.9

```
array([-5.3622725e-04,  2.3643016e-04,  5.1033497e-03,  9.0092728e-03,
       -9.3029495e-03, -7.1168090e-03,  6.4588715e-03,  8.9729885e-03,
       -5.0154282e-03, -3.7633730e-03,  7.3805046e-03, -1.5334726e-03,
       -4.5366143e-03,  6.5540504e-03, -4.8601604e-03, -1.8160177e-03,
        2.8765798e-03,  9.9187379e-04, -8.2852151e-03, -9.4488189e-03,
        7.3117660e-03,  5.0702621e-03,  6.7576934e-03,  7.6286553e-04,
        6.3508893e-03, -3.4053659e-03, -9.4640255e-04,  5.7685734e-03,
       -7.5216386e-03, -3.9361049e-03, -7.5115822e-03, -9.3004224e-04,
        9.5381187e-03, -7.3191668e-03, -2.3337698e-03, -1.9377422e-03,
        8.0774352e-03, -5.9308959e-03,  4.5161247e-05, -4.7537349e-03,
       -9.6035507e-03,  5.0072931e-03, -8.7595871e-03, -4.3918253e-03,
       -3.5099984e-05, -2.9618264e-04, -7.6612402e-03,  9.6147414e-03,
        4.9820566e-03,  9.2331432e-03, -8.1579182e-03,  4.4957972e-03,
       -4.1370774e-03,  8.2453492e-04,  8.4986184e-03, -4.4621779e-03,
        4.5175003e-03, -6.7869616e-03, -3.5484887e-03,  9.3985079e-03,
       -1.5776539e-03,  3.2137157e-04, -4.1406299e-03, -7.6826881e-03,
       -1.5080094e-03,  2.4697948e-03, -8.8802812e-04,  5.5336617e-03,
       -2.7429771e-03,  2.2600652e-03,  5.4557943e-03,  8.3459523e-03,
       -1.4537406e-03, -9.2081428e-03,  4.3705511e-03,  5.7178497e-04,
        7.4419067e-03, -8.1328390e-04, -2.6384138e-03, -8.7530091e-03,
       -8.5655687e-04,  2.8265619e-03,  5.4014279e-03,  7.0526553e-03,
       -5.7031228e-03,  1.8588186e-03,  6.0888622e-03, -4.7980524e-03,
       -3.1072616e-03,  6.7976285e-03,  1.6314745e-03,  1.8991709e-04,
        3.4736372e-03,  2.1777629e-04,  9.6188262e-03,  5.0606038e-03,
       -8.9173913e-03, -7.0415614e-03,  9.0145587e-04,  6.3925339e-03],
      dtype=float32)
```

รูปที่ 3.9 ตัวอย่างข้อมูลที่ถูกสร้างขึ้นโดยโมเดล Word2Vec

3.3.2 วิธีการโมเดลเบิร์ตแบบหลายภาษา คือ โมเดลที่ได้รับการเรียนรู้ในหลากหลายภาษา โดยใช้โมเดลเบิร์ตแบบหลายภาษาในการเรียนรู้ข้อมูลกับคลังข้อมูลที่มีขนาดใหญ่และมีหลากหลายภาษา เพื่อที่จะที่ทำการแปลง “คำ” ให้อยู่ในรูปของ “ตัวเลขหรือเวกเตอร์” ในภาษาไทยรวมถึงภาษาอื่น ๆ ได้ แสดงได้ดัง รูปที่ 3.10



รูปที่ 3.10 ตัวอย่างข้อมูลที่ถูกสร้างขึ้นโดยโมเดลเบิร์ตแบบหลายภาษา

3.4 การเรียนรู้ข้อมูล

หลังจากที่ได้มีการแทนคำภาษาไทยด้วยเวกเตอร์โดยใช้วิธีการต่าง ๆ แล้ว ขั้นตอนต่อไปคือการนำข้อมูลเวกเตอร์ที่ได้ เข้าสู่การเรียนรู้ของเครื่องหรือโมเดลในรูปแบบต่าง ๆ โดยวิธีการ คือ นำข้อมูลเวกเตอร์ที่ได้ ทั้งในส่วนที่ได้จาก Word2Vec และโมเดลเบิร์ตแบบหลายภาษา (mBERT) นำเข้าไปในการเรียนรู้ของเครื่องหรือโมเดลต่าง ๆ พร้อมตั้งค่าพารามิเตอร์ที่เกี่ยวข้อง ที่ใช้ในการวิจัยทั้งสิ้นจำนวน 5โมเดล โดยมีรายชื่อดังนี้ คือ

- Word2Vec ร่วมกับ วิธีต้นไม้ตัดสินใจ (Decision Trees)
- Word2Vec ร่วมกับ วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (K-nearest Neighbors)
- Word2Vec ร่วมกับ วิธีเบย์อย่างง่าย (Naive Bayes)
- Word2Vec ร่วมกับ วิธีซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine)
- Word2Vec ร่วมกับ วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (MLP)
- mBERT ร่วมกับ วิธีต้นไม้ตัดสินใจ (Decision Trees)
- mBERT ร่วมกับ วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (K-nearest Neighbors)
- mBERT ร่วมกับ วิธีเบย์อย่างง่าย (Naive Bayes)
- mBERT ร่วมกับ วิธีซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine)
- mBERT ร่วมกับ วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (MLP)

3.5 การวัดประสิทธิภาพ

หลังจากที่ทำการเรียนรู้ข้อมูลแล้ว ขั้นตอนต่อไป คือ การวัดประสิทธิภาพ ทั้งนี้จะใช้วิธีที่เรียกว่าการวัดประสิทธิภาพแบบดึงทีละตัว (Leave-One-Out: LOO) มีหลักการ คือ แบ่งข้อมูลออกเป็นชุดฝึกสอนและชุดทดสอบ จากนั้นทำการวนซ้ำหลาย ๆ รอบโดยเท่ากับจำนวนข้อมูลที่มี โดยละจุดข้อมูลหนึ่งจุดออกจากชุดข้อมูลเพื่อเป็นชุดทดสอบ ส่วนจุดข้อมูลที่เหลือเป็นแบบชุดฝึกสอน ทยอยเปลี่ยนไปทีละตัวไปเรื่อย ๆ เพื่อดำเนินการทดสอบจนครบ แล้วนำค่าความถูกต้องไปหาค่าเฉลี่ยเพื่อวัดประสิทธิภาพโดยรวมของโมเดล

โดยงานวิจัยนี้จะใช้วิธีการวัดประสิทธิภาพแบบดึงทีละตัว ดังสมการที่ 3.1

$$MSE = \left(\frac{1}{n}\right) * \sum \left(y_i - f(x_i)\right)^2 \quad (3.1)$$

เมื่อ

n คือ จำนวนตัวอย่างทั้งหมด

y_i คือ ค่าคำตอบของจำนวนตัวอย่างที่ i

$f(x_i)$ คือ ค่าทำนายของจำนวนตัวอย่างที่ i

บทที่ 4

ผลการวิจัย

4.1 การทดลองจัดประเภทคำถามบทสนทนาภาษาไทย

ในการทดลองนี้จะใช้ข้อมูลจากคลังข้อมูลถามตอบภาษาไทยของศูนย์เทคโนโลยีอีเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) แบ่งเป็นข้อมูลชุดฝึกสอนและข้อมูลชุดทดสอบ โดยเป็นข้อมูลประเภทถามตอบ ประกอบด้วยคำถามและคำตอบจำนวน 4,000 ตัวอย่าง ในหมวดหมู่ที่แตกต่างกัน เช่น วิทยาศาสตร์, สังคมและบันเทิง ซึ่งภายในจะประกอบไปด้วย “question_id” หมายถึง รหัสคำถาม, “question” หมายถึง ข้อความคำถาม, “answer” หมายถึง ข้อความคำตอบ, “answer_begin_position” หมายถึง ตำแหน่งเริ่มต้นของข้อความคำตอบ, “answer_end_position” หมายถึง ตำแหน่งสิ้นสุดของข้อความคำตอบ และ “article_id” หมายถึง หมวดหมู่คำถาม โดยงานวิจัยนี้จะใช้หัวข้อ “article_id” ในการเป็นคำตอบหรือคลาสของข้อมูลชุดฝึกสอน แล้วแบ่งออกเป็น 3 กลุ่มตามจำนวนขั้นต่ำของคำถามที่อยู่ในหมวดหมู่คำถามนั้น ๆ จากนั้นใช้ Library ของภาษา Python ชื่อ “PyThaiNLP” ในการตัดคำภาษาไทย โดยเทคนิคที่เรียกว่า “Maximum Matching Algorithm” ในการแบ่งคำภาษาไทยออกจากประโยค แล้วใช้วิธีการโมเดล Thai2Vec หรือโมเดลเบิร์ตแบบหลายภาษา ในการแปลงคำให้อยู่ในรูปเวกเตอร์เพื่อนำเข้าสู่โมเดลการเรียนรู้ของเครื่อง ทั้งวิธี วิธีต้นไม้ตัดสินใจ (Decision Trees), วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (K-nearest Neighbors), วิธีเบย์อย่างง่าย (Naïve Bayes), วิธีซัพพอร์ตเวคเตอร์แมชชีน (Support Vector Machine) และ วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (MLP) ท้ายที่สุดจะใช้วิธีที่เรียกว่า Leave One Out: LOO ในการวัดประสิทธิภาพโดยรวมของโมเดลต่าง ๆ โดยจะมีการวัดประสิทธิภาพมุมของความแม่นยำ

4.2 ผลการใช้ Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 3 คำถามขึ้นไป

ตารางที่ 4.1 ผลทดลอง Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 3 คำถามขึ้นไป

อังกอริทึม	พารามิเตอร์	ความแม่นยำ (เปอร์เซ็นต์)	เวลาที่ใช้ (วินาที)
DTs	gini	29.54±0.456	357.00
DTs	entropy	31.72±0.465	310.00
KNN	minkowski	48.43±0.499	0.71
KNN	cosine	48.91±0.499	0.90
KNN	euclidean	48.43±0.499	0.70
KNN	manhattan	49.15±0.499	0.67
NB	GaussianNB	30.27±0.459	4.95
NB	BernoulliNB	51.33±0.499	5.00
SVM	rbf	16.46±0.370	36.20
SVM	linear	48.91±0.499	33.70
SVM	poly	17.43±0.379	31.30
SVM	sigmoid	05.08±0.219	34.40
MLP	relu	50.85±0.499	1,088.00
MLP	identity	54.96±0.497	1,035.00
MLP	logistic	06.30±0.242	557.00
MLP	tanh	54.72±0.497	1,233.00

จากตารางที่ 4.1 แสดงถึงผลทดลอง Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 3 คำถามขึ้นไป โดยพบว่า วิธีการเรียนรู้แบบ KNN จะได้ค่าเฉลี่ยความแม่นยำสูงที่สุด เมื่อเทียบกับวิธีการอื่น ๆ แต่มีจุดสังเกตที่ วิธีการเรียนรู้แบบ MLP จะใช้เวลาในการเรียนรู้มากที่สุด แต่ค่าเฉลี่ยความแม่นยำสูงได้น้อยกว่าแบบ KNN

4.3 ผลการใช้ Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 4 คำถามขึ้นไป

ตารางที่ 4.2 ผลทดลอง Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 4 คำถามขึ้นไป

อังกอริทึม	พารามิเตอร์	ความแม่นยำ (เปอร์เซ็นต์)	เวลาที่ใช้ (วินาที)
DTs	gini	31.71±0.465	267.00
DTs	entropy	35.14±0.477	181.00
KNN	minkowski	50.86±0.499	0.58
KNN	cosine	50.57±0.499	0.72
KNN	euclidean	50.86±0.499	0.57
KNN	manhattan	50.86±0.499	0.56
NB	GaussianNB	35.14±0.477	3.49
NB	BernoulliNB	54.00±0.498	3.57
SVM	rbf	22.29±0.416	21.00
SVM	linear	51.43±0.499	19.90
SVM	poly	23.71±0.425	18.60
SVM	sigmoid	06.86±0.252	20.50
MLP	relu	64.00±0.480	675.00
MLP	identity	63.14±0.482	679.80
MLP	logistic	18.00±0.384	693.60
MLP	tanh	65.43±0.475	791.40

จากตารางที่ 4.2 แสดงถึงผลทดลอง Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 4 คำถามขึ้นไป โดยพบว่า วิธีการเรียนรู้แบบ MLP จะได้ค่าเฉลี่ยความแม่นยำสูงที่สุด เมื่อเทียบกับวิธีการอื่น ๆ แต่มีจุดสังเกตที่ วิธีการเรียนรู้แบบ MLP จะใช้เวลาในการเรียนรู้มากที่สุด มากกว่าวิธีการอื่น ๆ เช่นเดียวกัน

4.4 ผลการใช้ Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 5 คำถามขึ้นไป

ตารางที่ 4.3 ผลทดลอง Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 5 คำถามขึ้นไป

อังกอริทึม	พารามิเตอร์	ความแม่นยำ (เปอร์เซ็นต์)	เวลาที่ใช้ (วินาที)
DTs	gini	64.29±0.479	0.90
DTs	entropy	60.00±0.489	1.36
KNN	minkowski	74.29±0.437	0.10
KNN	cosine	72.86±0.444	0.13
KNN	euclidean	74.29±0.437	0.11
KNN	manhattan	75.71±0.428	0.10
NB	GaussianNB	72.86±0.444	0.16
NB	BernoulliNB	75.71±0.428	0.16
SVM	rbf	68.57±0.464	0.23
SVM	linear	85.71±0.349	0.21
SVM	poly	61.43±0.486	0.21
SVM	sigmoid	34.29±0.474	0.23
MLP	relu	87.14±0.334	17.80
MLP	identity	84.29±0.363	16.70
MLP	logistic	82.86±0.376	18.20
MLP	tanh	84.29±0.363	19.70

จากตารางที่ 4.3 แสดงถึงผลทดลอง Word2Vec กับหมวดหมู่คำถาม ตั้งแต่ 5 คำถามขึ้นไป โดยพบว่า วิธีการเรียนรู้แบบ MLP จะได้ค่าเฉลี่ยความแม่นยำสูงที่สุด เมื่อเทียบกับวิธีการอื่น ๆ แต่มีจุดสังเกตที่ วิธีการเรียนรู้แบบ MLP จะใช้เวลาในการเรียนรู้มากที่สุด มากกว่าวิธีการอื่น ๆ เช่นเดียวกัน

4.5 ผลการใช้ mBERT กับหมวดหมู่คำถาม ตั้งแต่ 3 คำถามขึ้นไป

ตารางที่ 4.4 ผลทดลอง mBERT กับหมวดหมู่คำถาม ตั้งแต่ 3 คำถามขึ้นไป

อังกอริทึม	พารามิเตอร์	ความแม่นยำ (เปอร์เซ็นต์)	เวลาที่ใช้ (วินาที)
DTs	gini	25.91±0.438	6,355.00
DTs	entropy	31.96±0.466	3,131.00
KNN	minkowski	43.34±0.495	3.55
KNN	cosine	42.86±0.494	8.51
KNN	euclidean	43.34±0.495	3.56
KNN	manhattan	44.07±0.496	3.10
NB	GaussianNB	26.15±0.439	13.30
NB	BernoulliNB	08.72±0.282	16.90
SVM	rbf	00.00±0.000	174.00
SVM	linear	48.43±0.499	141.00
SVM	poly	00.00±0.000	154.00
SVM	sigmoid	00.00±0.000	162.00
MLP	relu	11.14±0.314	7,923.00
MLP	identity	47.22±0.499	7,465.00
MLP	logistic	24.70±0.431	8,902.00
MLP	tanh	40.92±0.491	4,283.00

จากตารางที่ 4.4 แสดงถึงผลทดลอง mBERT กับหมวดหมู่คำถาม ตั้งแต่ 3 คำถามขึ้นไป โดยพบว่า วิธีการเรียนรู้แบบ KNN จะได้ค่าเฉลี่ยความแม่นยำสูงที่สุด เมื่อเทียบกับวิธีการอื่น ๆ แต่มีจุดสังเกตที่ วิธีการเรียนรู้แบบ MLP จะใช้เวลาในการเรียนรู้มากที่สุด แต่ค่าเฉลี่ยความแม่นยำสูงได้น้อยกว่าแบบ KNN รวมถึง วิธีการเรียนรู้แบบ SVM อาจกล่าวได้ว่าไม่เหมาะสมกับโมเดล mBERT เนื่องจากได้ความแม่นยำที่ 0 ในหลายกรณี

4.6 ผลการใช้ mBERT กับหมวดหมู่คำถาม ตั้งแต่ 4 คำถามขึ้นไป

ตารางที่ 4.5 ผลทดลอง mBERT กับหมวดหมู่คำถาม ตั้งแต่ 4 คำถามขึ้นไป

อังกอริทึม	พารามิเตอร์	ความแม่นยำ (เปอร์เซ็นต์)	เวลาที่ใช้ (วินาที)
DTs	gini	30.29±0.459	3,153.00
DTs	entropy	25.71±0.437	1,824.00
KNN	minkowski	47.14±0.499	1.96
KNN	cosine	48.00±0.499	3.93
KNN	euclidean	47.14±0.499	1.96
KNN	manhattan	47.71±0.499	1.80
NB	GaussianNB	30.29±0.459	9.16
NB	BernoulliNB	10.29±0.303	11.70
SVM	rbf	00.00±0.000	103.00
SVM	linear	52.86±0.499	83.00
SVM	poly	00.00±0.000	92.00
SVM	sigmoid	00.00±0.000	97.00
MLP	relu	38.86±0.487	2,415.00
MLP	identity	58.86±0.492	4,815.00
MLP	logistic	54.57±0.497	4,888.00
MLP	tanh	62.00±0.485	5,088.00

จากตารางที่ 4.5 แสดงถึงผลทดลอง mBERT กับหมวดหมู่คำถาม ตั้งแต่ 4 คำถามขึ้นไป โดยพบว่า วิธีการเรียนรู้แบบ MLP จะได้ค่าเฉลี่ยความแม่นยำสูงที่สุด เมื่อเทียบกับวิธีการอื่น ๆ แต่มีจุดสังเกตที่ วิธีการเรียนรู้แบบ MLP จะใช้เวลาในการเรียนรู้มากที่สุด มากกว่าวิธีการอื่น ๆ เช่นเดียวกัน รวมถึง วิธีการเรียนรู้แบบ SVM อาจกล่าวได้ว่าไม่เหมาะสมกับโมเดล mBERT เนื่องจากได้ความแม่นยำที่ 0 ในหลายกรณี

4.7 ผลการใช้ mBERT กับหมวดหมู่คำถาม ตั้งแต่ 5 คำถามขึ้นไป

ตารางที่ 4.6 ผลทดลอง mBERT กับหมวดหมู่คำถาม ตั้งแต่ 5 คำถามขึ้นไป

อังกอริทึม	พารามิเตอร์	ความแม่นยำ (เปอร์เซ็นต์)	เวลาที่ใช้ (วินาที)
DTs	gini	57.14±0.494	9.31
DTs	entropy	67.14±0.469	12.10
KNN	minkowski	77.14±0.419	0.15
KNN	cosine	81.43±0.388	0.21
KNN	euclidean	77.14±0.419	0.15
KNN	manhattan	81.43±0.388	0.14
NB	GaussianNB	74.29±0.437	0.30
NB	BernoulliNB	51.43±0.499	0.46
SVM	rbf	00.00±0.000	0.84
SVM	linear	85.71±0.349	0.65
SVM	poly	00.00±0.000	0.78
SVM	sigmoid	00.00±0.000	0.81
MLP	relu	87.14±0.334	154.00
MLP	identity	88.57±0.318	149.00
MLP	logistic	88.57±0.318	160.00
MLP	tanh	88.57±0.318	161.00

จากตารางที่ 4.6 แสดงถึงผลทดลอง mBERT กับหมวดหมู่คำถาม ตั้งแต่ 8 คำถามขึ้นไป โดยพบว่า วิธีการเรียนรู้แบบ MLP จะได้ค่าเฉลี่ยความแม่นยำสูงที่สุด เมื่อเทียบกับวิธีการอื่น ๆ แต่มีจุดสังเกตที่ วิธีการเรียนรู้แบบ MLP จะใช้เวลาในการเรียนรู้มากที่สุด มากกว่าวิธีการอื่น ๆ เช่นเดียวกัน รวมถึง วิธีการเรียนรู้แบบ SVM อาจกล่าวได้ว่าไม่เหมาะสมกับโมเดล mBERT เนื่องจากได้ความแม่นยำที่ 0 ในหลายกรณี

4.8 สรุปผลการเปรียบเทียบภาพรวมความแม่นยำระหว่างโมเดล

ตารางที่ 4.7 ผลการเปรียบเทียบภาพรวมความแม่นยำระหว่างโมเดล

	Word2Vec			mBERT		
	หมวดหมู่คำถามตั้งแต่ 3 ขึ้นไป (%)	หมวดหมู่คำถามตั้งแต่ 4 ขึ้นไป (%)	หมวดหมู่คำถามตั้งแต่ 5 ขึ้นไป (%)	หมวดหมู่คำถามตั้งแต่ 3 ขึ้นไป (%)	หมวดหมู่คำถามตั้งแต่ 4 ขึ้นไป (%)	หมวดหมู่คำถามตั้งแต่ 5 ขึ้นไป (%)
DTs	30.63	33.43	62.15	28.94	28.00	62.14
KNN	48.73	50.79	74.29	43.40	47.50	79.29
NB	40.80	44.57	74.29	17.44	20.29	62.86
SVM	21.97	26.07	62.50	12.11	13.22	21.43
MLP	41.71	52.64	84.65	31.00	53.57	88.21

จากตารางที่ 4.7 แสดงถึงผลการเปรียบเทียบภาพรวมความแม่นยำระหว่างโมเดล โดยพบว่าโมเดล Word2Vec นั้น ได้ค่าความแม่นยำที่ดีกว่าโมเดล mBERT ในทุกวิธีของการเรียนรู้ของเครื่องกับหมวดหมู่คำถามที่มีคำถามตั้งแต่ 3 คำถามขึ้นไป โดยโมเดล mBERT จะได้ค่าความแม่นยำที่มากกว่าใน 2 กรณี คือ ต้องใช้วิธีการเรียนรู้ของเครื่องแบบ MLP กับหมวดหมู่คำถามที่มีคำถามตั้งแต่ 4 คำถามขึ้นไป ทั้งนี้มีจุดสังเกตว่า โมเดล mBERT อาจไม่เหมาะกับวิธีการเรียนรู้ของเครื่องแบบ SVM เนื่องจากค่าความแม่นยำที่ได้น้อย เมื่อเทียบกับวิธีการอื่น ๆ

4.9 สรุปผลการเปรียบเทียบภาพรวมระหว่างโมเดลที่ดีที่สุด

ตารางที่ 4.8 ผลการเปรียบเทียบภาพรวมระหว่างโมเดลที่ดีที่สุด

	โมเดล	ความแม่นยำ (เปอร์เซ็นต์)	Precision	Recall	F1 Score	เวลาที่ใช้ (วินาที)
Word2Vec กับหมวดหมู่คำถามตั้งแต่ 3 ขึ้นไป	KNN	48.73	0.53	0.48	0.48	0.75
Word2Vec กับหมวดหมู่คำถามตั้งแต่ 4 ขึ้นไป	MLP	52.64	0.69	0.65	0.66	709.95
Word2Vec กับหมวดหมู่คำถามตั้งแต่ 5 ขึ้นไป	MLP	84.65	0.84	0.82	0.82	18.10
mBERT กับหมวดหมู่คำถามตั้งแต่ 3 ขึ้นไป	KNN	43.40	0.42	0.41	0.40	4.68
mBERT กับหมวดหมู่คำถามตั้งแต่ 4 ขึ้นไป	MLP	53.57	0.47	0.47	0.46	4,301.50
mBERT กับหมวดหมู่คำถามตั้งแต่ 5 ขึ้นไป	MLP	88.21	0.85	0.81	0.82	156.00

จากตารางที่ 4.8 แสดงถึงผลการเปรียบเทียบภาพรวมระหว่างโมเดลที่ดีที่สุด โดยพบว่าในกรณีหมวดหมู่ที่มีคำถามตั้งแต่ 3 คำถามซึ่งอยู่ในหมวดหมู่เดียวกันขึ้นไป วิธีการเรียนรู้ของเครื่องแบบ KNN จะได้ค่าความแม่นยำดีที่สุดเมื่อเทียบกับในกรณีอื่น ๆ แต่ในกรณีหมวดหมู่ที่มีคำถามตั้งแต่ 4 คำถามขึ้นไปกลับพบว่า วิธีการเรียนรู้ของเครื่องแบบ MLP จะได้ค่าความแม่นยำที่ดีกว่าเมื่อเทียบกับในกรณีอื่น ๆ แต่มีข้อสังเกตุ คือ วิธีการเรียนรู้ของเครื่องแบบ MLP นั้นจะใช้ระยะเวลาในการเรียนรู้ที่มากกว่า เมื่อเทียบกับวิธีการอื่น ๆ

บทที่ 5

สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ในงานวิจัยนี้จะนำเสนอการประเมินประสิทธิภาพการทำงานระหว่างอัลกอริทึมสำหรับการจัดประเภทคำถามบทสนทนาภาษาไทย โดยใช้การแปลงคำให้อยู่ในรูปเวกเตอร์ คือ วิธีการโมเดล Thai2Vec หรือโมเดลเบิร์ตแบบหลายภาษา (mBERT) ร่วมกับการเรียนรู้ของเครื่อง ทั้งวิธีต้นไม้ตัดสินใจ (DTs), วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (KNN), วิธีเบย์อย่างง่าย (NB), วิธีซัพพอร์ตเวคเตอร์แมชชีน (SVM) และวิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (MLP) โดยจัดกลุ่มข้อมูลส่วนของการเรียนรู้ ตามจำนวนคำถามที่อยู่ในหมวดหมู่คำถาม ตั้งแต่จำนวน 3-5 คำถาม ขึ้นไปที่อยู่ในหมวดหมู่คำถามเดียวกันโดยเมื่อเปรียบเทียบวิธีการโมเดล Thai2Vec กับวิธีการเรียนรู้ของเครื่องแบบต่าง ๆ พบว่าในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถาม ตั้งแต่ 3 คำถามขึ้นไป วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (KNN) จะได้ความแม่นยำสูงที่สุดเมื่อเทียบกับวิธีการเรียนรู้ของเครื่องแบบอื่น ๆ ซึ่งมีความแม่นยำอยู่ที่ 48.73% โดยในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถาม ตั้งแต่ 4 และ 5 คำถามขึ้นไป วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (MLP) จะได้ความแม่นยำสูงที่สุดเมื่อเทียบกับวิธีการเรียนรู้ของเครื่องแบบอื่น ๆ โดยมีความแม่นยำอยู่ที่ 52.64% และ 84.65% ตามลำดับ ทั้งนี้เมื่อเปรียบเทียบวิธีการโมเดลเบิร์ตแบบหลายภาษา (mBERT) กับวิธีการเรียนรู้ของเครื่องแบบต่าง ๆ พบว่าในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถาม ตั้งแต่ 3 คำถามขึ้นไป วิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว (KNN) จะได้ความแม่นยำสูงที่สุดเมื่อเทียบกับวิธีการเรียนรู้ของเครื่องแบบอื่น ๆ เช่นเดียวกัน ซึ่งมีความแม่นยำอยู่ที่ 43.40% โดยในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถาม ตั้งแต่ 4 และ 5 คำถามขึ้นไป วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น (MLP) จะได้ความแม่นยำสูงที่สุดเมื่อเทียบกับวิธีการเรียนรู้ของเครื่องแบบอื่น ๆ เช่นกัน โดยมีความแม่นยำอยู่ที่ 53.57% และ 88.21% ตามลำดับ

สรุปได้ว่าในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถาม ตั้งแต่ 3 คำถามขึ้นไป วิธีการโมเดล Thai2Vec ที่ใช้ร่วมกับวิธีเพื่อนบ้านที่ใกล้ที่สุด k ตัว จะได้ค่าความแม่นยำสูงที่สุดเมื่อเทียบกับวิธีการอื่น ๆ แต่ในกรณีที่มีหมวดหมู่คำถามที่มีจำนวนคำถามมากกว่านั้น วิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น จะได้ความแม่นยำที่ดีกว่า ทั้งนี้เมื่อใช้งานโมเดลเบิร์ตแบบหลายภาษาร่วมกับวิธีโครงข่ายประสาทเทียมแบบเพอร์เซฟตรอนหลายชั้น จะได้ค่าความแม่นยำสูงที่สุดเมื่อเทียบกับทุก ๆ โมเดลและทุก ๆ วิธีการเรียนรู้ข้อมูลที่ทำการวิจัย

5.2 ข้อเสนอแนะ

งานวิจัยในอนาคตควรมุ่งเน้นไปที่ การศึกษาวิธีการจัดประเภทคำถามบทสนทนาภาษาไทยเพิ่มเติม เช่น การหาคลังข้อมูลถามตอบภาษาไทย เพื่อขยายขอบเขตการโต้ตอบหรือเพิ่มความแม่นยำของการโต้ตอบเพิ่มเติม รวมถึงการใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติแบบอื่น ๆ เพื่อช่วยให้การตัดคำหรือการแปลงคำให้อยู่ในรูปแบบเวกเตอร์ ให้มีประสิทธิภาพในการจัดประเภทคำถามบทสนทนาภาษาไทยดียิ่งขึ้น

เอกสารอ้างอิง

Guangyi Wang, Feilong Bao, Weihua Wang. 2020. Mongolian Questions Classification in the Law Domain. *International Conference on Asian Language Processing* : 56-59

Nguyen Thi Mai Trang and Maxim Shcherbakov. 2020. Vietnamese Question Answering System from Multilingual BERT Models to Monolingual BERT Model. *9th International Conference on System Modeling & Advancement in Research Trends* : 201-205

Bineet Kumar Jha, Chandra Mouli Venkata Srinivas Akana, Anand R. 2021 Question Answering System with Indic multilingual-BERT. *5th International Conference on Computing Methodologies and Communication* : 1631-1638

Eslam Amer, Ahmed Hazem, Omar Farouk, Albert Louca, Youssef Mohamed, Michel Ashraf. 2021. A Proposed Chatbot Framework for COVID-19. *International Mobile, Intelligent, and Ubiquitous Computing Conference* : 263-268

Narayana Darapaneni, Pooja Chetan, Great Learning, Aravind Gaddala, Garima Tiwari, Sandipan Basu, Sadwik Parvathaneni. 2021. Building a Question and Answer System for News Domain. *Second International Conference in Secure Cyber Computing and Communication* : 78-83

Yihan Bian and Kaiwen Peng. 2021. Question Answering System Analysis Based on Machine Learning. *IEEE International Conference on Computer Science, Electronic Information Engineering, and Intelligent Control Technology* : 279-283

Nikita Kanodia, Khandakar Ahmed, Yuan Miao. 2021. Question Answering Model Based Conversational Chatbot using BERT Model and Google Dialogflow. *31st International Telecommunication Networks and Applications Conference* : 19-22

Jie Yin. 2022. Research on Question Answering System Based on BERT Model *3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications* : 68-71

Saranlita Chotirat, Phayung Meesad, Herwig Unger. 2022. Question Classification from Thai Sentences by Considering Word Context to Question Generation. *Research, Invention, and Innovation Congress: Innovative Electricals and Electronics* : 9-14

ประวัติผู้เขียน

ชื่อ	นายกิต ธนานุคุณ
วัน เดือน ปีเกิด	6 มกราคม พ.ศ.2532
ที่อยู่ปัจจุบัน	69/149 ซอยนวมินทร์ 153 ถนนนวมินทร์ นวลจันทร์ บึงกุ่ม กรุงเทพฯ
ประวัติการศึกษา	2556 วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ เกรดเฉลี่ย 3.58 มหาวิทยาลัยราชภัฏจันทรเกษม