การสร้างหัวข้อข่าวภาษาไทยด้วย AI :
ดึงดูดผู้อ่านด้วยหัวข้อข่าวที่สร้างจาก AI พร้อมสื่อถึงความรู้สึก

HEADLINES GENERATION FOR THAI NEWS : ENGAGING READERS WITH AI-GENERATED TITLES WITH EMOTIONS

ธนวัตร แก้วมณี

ธรรธร จงสกุล

โครงงานพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ปีการศึกษา ๒๕๖๖

HEADLINES GENERATION FOR THAI NEWS : ENGAGING READERS WITH AI-GENERATED TITLES WITH EMOTIONS

TANAWAT KAEWMANEE

TUNTORN CHONGSAKUL

A SPECIAL PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF SCIENCE (COMPUTER SCIENCE) DEPARTMENT OF COMPUTER SCIENCE, SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

ACADEMIC YEAR 2022

หัวข้อปัญหาพิเศษ การสร้างหัวข้อข่าวภาษาไทยด้วย AI : ดึงดูดผู้อ่านด้วยหัวข้อข่าวที่สร้างจาก AI พร้อมสื่อถึงความรู้สึก

ชื่อนักศึกษา นาย ธนวัตร แก้วมณี รหัสนักศึกษา 63050138

นาย ธรรธร จงสกุล รหัสนักศึกษา 63050143

ปริญญา วิทยาศาสตรบัณฑิต (วิทยาการคอมพิวเตอร์)

ภาควิชา วิทยาการคอมพิวเตอร์

คณะ วิทยาศาสตร์

มหาวิทยาลัย สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2566

อาจารย์ที่ปรึกษา ดร.จักรพันธ์ เตไชยา

บทคัดย่อ

ปัญหาพิเศษนี้มีวัตถุประสงค์ในการพัฒนาโมเดลการประมวลผลภาษาธรรมชาติเพื่อสังเคราะห์หัวข้อข่าวภาษาไทยที่เน้นย้ำถึงด้านอารมณ์ความรู้สึกจากชุดข้อมูลข่าวภาษาไทย ThaiSum ซึ่งได้แยกประเภทชุดข้อมูลสำหรับการฝึกสอนโมเดลตามแนวทาง 3 รูปแบบ เพื่อหารูปแบบชุดข้อมูลที่มีประสิทธิภาพในการสังเคราะห์หัวข้อข่าวมากที่สุด ได้แก่ การแยกประเภทด้วยการวิเคราะห์ความรู้สึก การแยกประเภทตามความถี่ของคีย์เวิร์ดที่เน้นย้ำความรู้สึก และการแยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐานและใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด การทดสอบได้แบ่งออกเป็น 2 ขั้นตอน โดยขั้นแรกเป็นการทดสอบประสิทธิภาพของชุดข้อมูลทั้ง 3 รูปแบบด้วย BERTScore และ ROUGE เพื่อหารูปแบบที่เหมาะสมและให้ผลลัพธ์ที่ดีที่สุดสำหรับการสังเคราะห์หัวข้อข่าวภาษาไทยที่เน้นย้ำด้านอารมณ์ความรู้สึก จากนั้นขั้นตอนที่ 2 จึงทดสอบประสิทธิภาพของโมเดลที่มีโครงสร้างแตกต่างกัน 3 รูปแบบ ได้แก่ MT5 OpenThaiGPT และ Gemini โดยใช้ชุดข้อมูลข่าวความรู้สึกเศร้าซึ่งเป็นชุดข้อมูลที่ให้ผลลัพธ์ที่ดีที่สุดจากการทดสอบชุดข้อมูล ผลการวัดประสิทธิภาพด้วย BERTScore และ ROUGE แสดงให้เห็นว่าโมเดล MT5 ให้ผลลัพธ์ที่ดีที่สุดในด้านค่า ROUGE ที่คะแนน ROUGE-1 เท่ากับ 39.54 และ ROUGE-L เท่ากับ 36.39 ในขณะที่โมเดล Gemini ให้ค่า BERTScore ที่สูงที่สุดที่คะแนน 77.09 และสำหรับ MT5 ให้ค่า BERTScore ที่ 75.99

เลือกมา model เดียว เช่น (MT5) แล้วสรุปค่าทั้งหมดแล้วบอกสรุปว่าโมเดลนี้เหมาะสมในงานที่เราต้องการที่จะทำมากที่สุดครับ ไม่ต้อง report score ของโมเดลรองใน abstact

คำสำคัญ : การวิเคราะห์ความรู้สึก, การวิเคราะห์ความถี่, คีย์เวิร์ด, ทฤษฎีความรู้สึกพื้นฐาน, การปรับแต่งโมเดล, การปรับแต่งคำสั่ง, การออกแบบคำสั่ง

Title	HEADLINES GENERATION FOR THAI NEWS: ENGAGING READERS WITH AI-GENERATED TITLES WITH EMOTIONS
Students	Mr.Tanawat Kaewmanee Student ID 63050138
	Mr.Tuntorn Chongsakul Student ID 63050143
Degree	Bachelor of Science (Computer Science)
Department	Computer Science
School	Science
University	King Mongkut's Institute of Technology Ladkrabang (KMITL)
Academic Year	2023
Advisor	Dr. Jakapun Tachaiya

Abstract

This special problem aims to develop a natural language processing model to synthesize Thai news headlines that emphasize emotions from the ThaiSum Thai news dataset. The dataset is categorized into three approaches for training the model to find the most effective dataset format for synthesizing emotional news headlines. These are categorization by sentiment analysis, categorization by frequency of emotion-emphasizing keywords, and categorization by basic emotion theory using the most frequent emotion-emphasizing keywords. The evaluation was divided into two stages. The first stage tested the performance of the three dataset formats using BERTScore and ROUGE to find the most suitable format that yields the best results for synthesizing Thai emotional news headlines. The second stage then tested the performance of three different model architectures: MT5, OpenThaiGPT, and Gemini, using the sad news dataset, which was the best-performing dataset from the first stage. The performance evaluation using BERTScore and ROUGE showed that the MT5 model performed best in terms of ROUGE scores, with ROUGE-1 of 39.54 and ROUGE-L of 36.39, while the Gemini model had the highest BERTScore of 77.09, and MT5 had a BERTScore of 75.99. **เด๊ยวแก้ตามภาษาไทยด้วยครับ**

Keywords : Sentiment analysis, Frequency analysis, Keyword, Basic emotions, Fine tuning, Prompt tuning, Prompt engineering

คณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ใบรับรองการค้นคว้าอิสระ

หัวข้อการค้นคว้าอิสระ การสร้างหัวข้อข่าวภาษาไทยด้วย AI : ดึงดูดผู้อ่านด้วยหัวข้อข่าวที่สร้างจาก AI พร้อมสื่อถึงความรู้สึก

ชื่อนักศึกษา นาย ธนวัตร แก้วมณี รหัสนักศึกษา 63050138

นาย ธรรธร จงสกุล รหัสนักศึกษา 63050143

ปริญญา วิทยาศาสตรบัณฑิตสาขาวิชาวิทยาการคอมพิวเตอร์

ภาควิชา วิทยาการคอมพิวเตอร์

อาจารย์ที่ปรึกษา ดร.จักรพันธ์ เตไชยา

คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง (สจล.) อนุมัติให้ปัญหาพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรบัณฑิต (วิทยาการ-คอมพิวเตอร์) ประจำปีการศึกษา 2566

คณะกรรมการสอบปัญหาพิเศษ	**ลายมือ**
ผศ.ดร.อัคคัญญ์ นรบิน (ประธานกรรมการ)	
ดร.บุญหทัย เครือแก้ว (กรรมการ)	
ดร.จักรพันธ์ เตไชยา (อาจารย์ที่ปรึกษา)	

ลิขสิทธิ์ของคณะวิทยาศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

กิตติกรรมประกาศ

ปัญหาพิเศษฉบับนี้สำเร็จลุล่วงด้วยดีเนื่องจากความกรุณา และความช่วยเหลืออย่างยิ่งจาก ท่านดร.จักรพันธ์ เตไชยา อาจารย์ที่ปรึกษาปัญหาพิเศษ ที่ได้ให้ความกรุณาสละเวลาอันมีค่ายิ่ง ในการให้คำปรึกษา การดำเนินงานปัญหาพิเศษ ตลอดจนได้ตรวจสอบ และแก้ไขข้อบกพร่องต่าง ๆ อันเป็นประโยชน์ในการจัดทำปัญหาพิเศษนี้ ตั้งแต่เริ่มดำเนินการจนกระทั่งเสร็จสมบูรณ์ คณะผู้จัดทำขอขอบพระคุณเป็นอย่างสูง ณ โอกาสนี้

ขอขอบพระคุณ ประธานกรรมการ และกรรมการ ที่ได้ชี้แนะให้เล็งเห็นถึงปัญหา แนะนำแนวทาง หรือจุดที่ควรปรับปรุงแก้ไข เพื่อให้ปัญหาพิเศษนี้ได้พัฒนาขึ้น

ขอขอบพระคุณคณาจารย์ทุกท่านในภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ที่ได้ถ่ายทอดองค์ความรู้มาตลอด 4 ปี ที่ผ่านมา ทั้งความรู้ด้านวิชาการ ด้านประสบการณ์ในรายงานต่าง ๆ ที่ทำให้คณะผู้จัดทำสามารถพัฒนาตนเอง และแก้ไขปัญหาต่าง ๆ จนสามารถจัดทำปัญหาพิเศษฉบับนี้ให้เสร็จสิ้นได้

และสุดท้ายนี้ขอขอบคุณเพื่อน ๆ ในกลุ่มที่ได้ค่อยให้ความช่วยเหลือในการให้คำปรึกษาตลอดมา รวมถึงกำลังใจที่ได้รับอยู่เสมอไม่ว่าทางคณะผู้จัดทำจะพบเจอปัญหาใด ๆ ก็ตาม

ธนวัตร แก้วมณี

ธรรธร จงสกุล

สารบัญ

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญตาราง

หน้า

สารบัญตาราง (ต่อ)

หน้า

สารบัญรูป

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันเทคโนโลยีปัญญาประดิษฐ์ (Artificial intelligence: AI) ได้เข้ามามีบทบาทในการช่วยงานมนุษย์มากขึ้นเรื่อย ๆ การประมวลผลภาษาธรรมชาติ (Natural language processing: NLP) ก็เป็นสาขาวิชาในด้านปัญญาประดิษฐ์และคอมพิวเตอร์ที่เกี่ยวข้องกับการประมวลผลและเข้าใจภาษามนุษย์ที่เป็นภาษาธรรมชาติ การสรุปเนื้อหา (Text summarization) และการสร้างข้อความ (Text generation) เป็นส่วนหนึ่งของการประมวลผลภาษาธรรมชาติ โดยการสร้างข้อความเป็นกระบวนการสร้างข้อความหรือประโยคที่มีความหมายโดยใช้เทคนิค และโมเดลปัญญาประดิษฐ์ ในขณะที่การสรุปเนื้อหา คือกระบวนการสรุปใจความสำคัญของเนื้อหา

คณะผู้จัดทำเล็งเห็นว่าสามารถนำความรู้ด้านการสรุปเนื้อหา และการสร้างข้อความมาประยุกต์ใช้กับการสร้างหัวข้อข่าวซึ่งการสร้างหัวข้อข่าว หมายถึง การสรุปใจความสำคัญของเนื้อหาข่าวให้อยู่ในรูปประโยคที่สั้น กระชับ และสื่อถึงภาพรวมของข่าวนั้น ๆ ซึ่งลักษณะเด่นของหัวข้อข่าวภาษาไทย นอกจากจะต้องมีความสั้น และกระชับแล้วยังมักจะมีการใช้คำที่เน้นย้ำความรู้สึก เพื่อทำให้หัวข้อข่าวน่าสนใจ และดึงดูดผู้อ่าน เช่น สลด เศร้า สยอง ผวา ชี้ เผย เป็นต้น อย่างไรก็ตามถึงแม้ว่าปัจจุบันจะสามารถนำปัญญาประดิษฐ์มาสังเคราะห์หัวข้อข่าวได้แต่ก็ยังคงมีข้อจำกัด คือ หัวข้อข่าวที่ถูกสังเคราะห์ยังขาดคำที่เน้นย้ำถึงความรู้สึกอยู่

ทางคณะผู้จัดทำได้เห็นถึงความสำคัญของปัจจัยดังกล่าว และต้องการที่จะสร้างโมเดลสังเคราะห์หัวข้อข่าว ที่สามารถสังเคราะห์หัวข้อข่าวได้ตอบโจทย์ ลักษณะเด่นของหัวข้อข่าวภาษาไทยที่นอกจากจะต้องมีความสั้น และกระชับแล้วยังสามารถสื่อถึงความรู้สึกได้อย่างชัดเจน ทำให้เป็นการนำมาซึ่งการทำปัญหาพิเศษในหัวข้อ "การสร้างหัวข้อข่าวภาษาไทยด้วย AI : ดึงดูดผู้อ่านด้วยหัวข้อข่าวที่สร้างจาก AI พร้อมสื่อถึงความรู้สึก" ซึ่งเป็นการนำความรู้จากการสรุปเนื้อหาและการสร้างข้อความมาใช้ฝึกสอนโมเดลสังเคราะห์หัวข้อข่าวจากข่าวภาษาไทย ที่สามารถเน้นย้ำถึงความรู้สึกได้ชัดเจน โดยชุดข้อมูลที่จะนำมาฝึกสอน คือ ชุดข้อมูลข่าวภาษาไทย ThaiSum เป็นชุดข้อมูลที่รวบรวมข้อมูลข่าวจากแหล่งข่าวไทยต่าง ๆ เช่น ไทยรัฐ ไทยพีบีเอส ประชาไท และเดอะแสตนดาร์ด คุณลักษณะของชุดข้อมูลจะประกอบด้วย หัวข้อข่าว เนื้อหาข่าว เนื้อหาข่าวแบบย่อ ประเภทของข่าว และแท็กของข่าว

1.2 วัตถุประสงค์ของงานวิจัย

1) พัฒนาโมเดลการประมวลผลภาษาธรรมชาติ (NLP Model) สำหรับสังเคราะห์หัวข้อข่าวภาษาไทย
2) สร้างโมเดลที่สามารถสังเคราะห์หัวข้อข่าวภาษาไทยที่สื่อถึงความรู้สึกของข่าวนั้นออกมาได้

1.3 ขอบเขตของงานวิจัย

ในการศึกษาครั้งนี้มีการกำหนดขอบเขตของการพัฒนา คือ ทำการพัฒนาในรูปแบบแบบจำลองโมเดลสำหรับใช้งานร่วมกับโมเดลที่ฝึกสอนมาแล้ว (Pre-trained model)

1) โมเดลสามารถสังเคราะห์หัวข้อข่าวได้เฉพาะภาษาไทยเท่านั้น
2) หัวข้อข่าวที่โมเดลสังเคราะห์ขึ้นจะมีความยาวที่จำกัดให้เหมาะสมในการสร้างหัวข้อข่าว
3) การพัฒนาโมเดลจะใช้เฉพาะข้อมูลจากชุดข้อมูลข่าวภาษาไทย ThaiSum เท่านั้น ซึ่งส่วนของคุณลักษณะที่จะนำมาวิเคราะห์ก็ได้แก่ หัวข้อข่าว และเนื้อหาข่าว โดยนำเนื้อหาข่าวมาใช้ในการสังเคราะห์หัวข้อข่าว

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) สามารถสังเคราะห์หัวข้อข่าวให้มีความน่าสนใจในด้านการสื่อถึงความรู้สึกของข่าวได้
2) ช่วยทำให้การสังเคราะห์ชื่อหัวข้อข่าวมีความกระชับ เข้าใจได้ง่าย และไม่บิดเบือน
3) สามารถนำโมเดลการสังเคราะห์หัวข้อข่าวนี้มาใช้ทดแทนในงานสังเคราะห์หัวข้อข่าวที่มนุษย์เป็นผู้เขียน

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ปัญญาประดิษฐ์ (Artificial intelligent : AI)

ปัญญาประดิษฐ์ เป็นสาขาด้านวิทยาการคอมพิวเตอร์ที่มุ่งเน้นแก้ไขปัญหาความรู้ความเข้าใจที่ปกติเชื่อมโยงกับความฉลาดของมนุษย์ เช่น การเรียนรู้ การสร้าง และการจดจำภาพ องค์กรสมัยใหม่รวบรวมข้อมูลจำนวนมากจากแหล่งที่มาหลากหลาย เช่น เซ็นเซอร์อัจฉริยะ เนื้อหาที่มนุษย์สร้างขึ้น เครื่องมือตรวจติดตาม และข้อมูลบันทึกระบบ เป้าหมายของปัญญาประดิษฐ์คือการสร้างระบบการเรียนรู้ด้วยตนเองซึ่งต่อยอดความหมายที่ได้จากข้อมูล จากนั้นปัญญาประดิษฐ์จะสามารถใช้ความรู้นั้นเพื่อแก้ปัญหาใหม่ ๆ ในรูปแบบที่คล้ายคลึงมนุษย์ได้ ตัวอย่างเช่น เทคโนโลยีปัญญาประดิษฐ์สามารถตอบสนองต่อการสนทนาของมนุษย์โดยมีความหมายเข้าใจได้ สร้างภาพและข้อความต้นฉบับ และตัดสินใจตามข้อมูลรับเข้าได้แบบทันที [1]

2.2 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่องเป็นการให้คอมพิวเตอร์เรียนรู้โดยไม่ต้องโปรแกรมโดยตรงในแบบของขั้นตอน และเงื่อนไขที่ชัดเจน ในขณะที่ให้ความสำคัญกับการใช้ข้อมูล เพื่อเรียนรู้ และปรับปรุงการปฏิบัติของระบบ ความสำคัญของการเรียนรู้ของเครื่องอยู่ที่ความสามารถในการพัฒนาและปรับปรุงโดยอัตโนมัติจากข้อมูลที่มีอยู่ โดยมีวัตถุประสงค์หลักคือการสร้างระบบหรือแอปพลิเคชันที่มีประสิทธิภาพ และฉลาดขึ้นเรื่อย ๆ

ด้วยการใช้การเรียนรู้ของเครื่องผู้พัฒนาสามารถสร้างโมเดลที่สามารถเรียนรู้จากข้อมูลและปรับปรุงการทำงานของตนเอง ซึ่งส่งผลให้สามารถทำงานได้ถูกต้องมากขึ้น และเหมาะสมกับสถานการณ์ที่เปลี่ยนแปลงได้ ยิ่งไปกว่านั้นการเรียนรู้ของเครื่องยังสามารถช่วยให้ระบบมีความสามารถในการทำนาย และการวิเคราะห์ข้อมูลที่ซับซ้อนได้ดีขึ้น ซึ่งเป็นประโยชน์อย่างมากในหลากหลายภาคสาขา เช่น การทำนายการเปลี่ยนแปลงในตลาดการเงิน การวิเคราะห์ข้อมูลทางการแพทย์ และการประมวลผลภาพถ่าย และวิดีโออัตโนมัติ เพื่อช่วยให้ผู้ใช้สามารถตัดสินใจหรือดำเนินการต่าง ๆ ได้อย่างมีประสิทธิภาพมากขึ้น โดยรวมแล้วการเรียนรู้ของเครื่อง เป็นเครื่องมือที่สำคัญในการสร้าง พัฒนาระบบ และแอปพลิเคชันในยุคปัจจุบันจนถึงอนาคต [2]

2.3 การทำความสะอาดข้อมูล (Data cleaning)

การทำความสะอาดข้อมูล เป็นกระบวนการในการตรวจจับ แก้ไข ลบ หรือแทนที่ข้อมูลที่เสียหายหรือไม่ถูกต้องจากตารางหรือฐานข้อมูล โดยให้ชุดข้อมูลมีความสอดคล้องกับความเป็นจริง

2.3 การทำความสะอาดข้อมูล (Data cleaning) (ต่อ)

และชุดข้อมูลอื่น ๆ เพื่อที่ข้อมูลเหล่านั้นจะได้พร้อมสำหรับการนำไปใช้งานมากที่สุด ซึ่ง 4 ขั้นตอนหลักในการทำความสะอาดข้อมูลมีดังนี้

1) **ลบข้อมูลที่ไม่เกี่ยวข้อง** : ตรวจสอบข้อมูล และลบตัวแปรหรือข้อมูลที่ไม่เกี่ยวข้องกับคำถามหรือจุดประสงค์การวิเคราะห์ เพื่อให้งานวิเคราะห์เป็นไปอย่างมีประสิทธิภาพ
2) **ลบข้อมูลที่ซ้ำซ้อน** : ตรวจสอบ และลบข้อมูลที่ซ้ำซ้อน เนื่องจากมันอาจทำให้ข้อมูลหนักขึ้นและส่งผลให้โมเดลทำงานได้ไม่แม่นยำ
3) **แก้ไขข้อมูลที่ผิดปกติ** : ตรวจสอบ และแก้ไขข้อมูลที่ผิดพลาด เช่น ข้อมูลที่ไม่เป็นไปตามรูปแบบ, ข้อมูลที่หายไป, หรือข้อมูลที่มีค่าผิดปกติ
4) **จัดการกับข้อมูลที่หายไป** : ทำการจัดการกับข้อมูลที่หายไปให้เหมาะสม เช่น การแทนค่าด้วยค่าเฉลี่ย, การลบข้อมูลหายไป, หรือการใช้ข้อมูลอื่น ๆ เพื่อแทนที่ข้อมูลที่หายไปได้ [3]

2.4 การตัดคำ (Tokenization)

การตัดคำ เป็นกระบวนการแบ่งประโยค คำ ตัวอักษร หรือเอกสารข้อความหนึ่งหรือหลายเอกสารที่มีข้อความเป็นหน่วยย่อยเรียกว่าโทเคน (Token) ซึ่งโทเคนนี้สามารถแยกได้หลายระดับ เช่น ระดับคำ (Word) ระดับคำย่อย (Sub word) หรือระดับตัวอักษร (Character) แต่ละอัลกอริทึมมีกระบวนการทำ Tokenization ที่แตกต่างกัน ซึ่งการแยกข้อความเป็นประโยคเรียกว่า Sentence - Tokenization ส่วนการแยกคำเรียกว่า Word tokenization โดยการทำ Word tokenization จะแบ่งข้อความ ออกเป็นโทเคนที่เป็นคำ โดยใช้ตัวแบ่งที่เฉพาะเจาะจง และสามารถแบ่งออกเป็น 3 ส่วนหลัก คือ [4]

1) การเน้นที่ระดับคำ (Word-based tokenization) แบ่งข้อความออกเป็นโทเคนโดยใช้ตัวแบ่งที่เฉพาะเจาะจง เช่น เว้นวรรค

   ตัวอย่าง ["สวัสดี", "ครับ", "ยินดี", "ที่", "ได้", "รู้จัก", "ครับ"]

2) การเน้นที่ระดับตัวอักษร (Character-based tokenization) แบ่งข้อความออกเป็นโทเคนโดยแยกออกเป็นตัวอักษรแต่ละตัว

   ตัวอย่าง ["ส", "ว", "ั", "ส ", "ด", "ี", "ค ", "ร ", "ั", "บ", ...]

3) การเน้นที่ระดับคำย่อย (Sub word tokenization) แบ่งข้อความออกเป็นโทเคนให้เป็นคำย่อยหรือส่วนย่อยของคำที่เรียกว่า Word piece ซึ่งหลักการทำงานของตัว Word piece จะมีดังนี้

   1) กำหนดคำศัพท์เริ่มต้น (Initial vocabulary) เริ่มต้นด้วยคำศัพท์เริ่มต้นที่เป็นคำหลัก ซึ่งสามารถเป็นคำทั่วไปที่พบบ่อยในข้อมูล
   2) การตัดคำเป็นส่วนย่อย ทำการตัดคำทั้งหมดในข้อมูลเป็นส่วนย่อย โดยใช้เกณฑ์ที่กำหนดไว้ เพื่อให้ได้คำที่เหมาะสมที่สุด

2.4 การตัดคำ (Tokenization) (ต่อ)

3) คำนวณความถี่ (Frequency calculation) นับความถี่ของคำทั้งหมดที่ได้จากขั้นตอนที่ 2

4) การผสม (Merging) ทำการผสมคู่ของตัวอักษรหรือสัญลักษณ์ที่มีความถี่สูงที่สุด โดยเป็นการเพิ่มส่วนย่อยใหม่ลงในคำศัพท์

5) การทำซ้ำ (Iteration) ทำซ้ำขั้นตอน 3 - 4 จนกระทั่งได้ขนาดของคำศัพท์ตามที่กำหนดไว้ล่วงหน้าหรือถึงขนาดที่ต้องการ

6) การตัดคำเป็นโทเคนที่สามารถนำเข้าไปในโมเดลการเรียนรู้เชิงลึก (Deep learning) ได้

ตัวอย่าง "วันนี้เป็นวันที่ดี"

- ทำการกำหนดคำเริ่มต้น

  ["วัน", "นี้", "เป็น", "วัน", "ที่", "ดี"]

- ทำการตัดคำเป็นส่วนย่อย

  ["วัน", "นี้", "เป็น", "วัน", "ที่", "ดี", "วันนี้", "วันที่", "เป็นวัน", "ที่ดี"]

- ทำการนับความถี่ จากนั้นผสมคำที่มีความถี่สูงที่สุด

  ["วัน", "นี้", "เป็น", "วัน", "ที่", "ดี", "วันนี้", "วันที่", "เป็นวัน", "ที่ดี"]

ซึ่งหลังจากนี้คำที่ได้ออกมาจะเป็นโทเคนที่สามารถนำเข้าไปในโมเดลได้ [4]

2.5 SentencePiece

SentencePiece เป็นเครื่องมือสำหรับการตัดคำข้อความที่มีประโยคอยู่ในรูปแบบของตัวอักษรหรือตัวอักษรยูนิโค้ด (Unicode) โดยทำให้เป็นการแบ่งส่วนที่มีประสิทธิภาพและเหมาะสมกับการประมวลผลภายในระบบการเรียนรู้ของเครื่องหรือการเรียนรู้เชิงลึก (Deep learning) โดยเฉพาะ SentencePiece มักถูกนำมาใช้ในงานการแบ่งส่วนภาษาธรรมชาติและภาษาที่มีรูปแบบโครงสร้างที่ซับซ้อน เช่น ภาษาญี่ปุ่น และภาษาไทย เพื่อให้การแบ่งส่วนมีประสิทธิภาพ และเหมาะสมกับการนำเข้าข้อมูลในระบบการเรียนรู้ของเครื่อง อย่างเช่นการสร้างโมเดลภาษาธรรมชาติหรือแปลภาษาอัตโนมัติ โดย SentencePiece มีความยืดหยุ่นในการปรับแต่งการแบ่งส่วนตามความต้องการของผู้ใช้ และสามารถปรับแต่งได้ตามความซับซ้อนของภาษาและข้อมูลที่มีอยู่ [5]

ตัวอย่าง "สวัสดีปีใหม่"

- ["สวัสดี", "ปี", "ใหม่"]

2.6 PyThaiNLP

PyThaiNLP เป็นไลบรารีของภาษาไพทอน (Python) ที่ถูกพัฒนาขึ้น เพื่อการประมวลผลข้อความในภาษาไทย โดยมีวัตถุประสงค์เพื่อช่วยในการวิเคราะห์ และประมวลผลข้อความภาษาไทยให้ง่ายขึ้น ซึ่ง Pythainlp มีฟังก์ชันหลากหลายที่สามารถช่วยในการประมวลผลข้อความได้ เช่น

1) การตัดคำ (Tokenization) ไลบรารีนี้มีเครื่องมือที่ช่วยในการตัดคำภาษาไทยออกเป็นหน่วยต่าง ๆ เช่น คำ, ประโยค, คำย่อย เป็นต้น
2) การแบ่งประโยค (Sentence segmentation) ใช้เพื่อแบ่งประโยคออกจากกันในข้อความภาษาไทย
3) การสร้างรายการคำศัพท์ (Vocabulary creation) ใช้เพื่อสร้างรายการคำศัพท์จากข้อความภาษาไทยที่กำหนด
4) การสร้างโมเดลภาษา (Language modeling) ใช้สร้าง และใช้งานโมเดลภาษาสำหรับการประมวลผลภาษาไทย [6]

2.7 Wordcount

Wordcount เป็นไลบรารีสำหรับนับจำนวนคำทั้งหมดในข้อความหรือเอกสารที่กำหนด ซึ่งเป็นกระบวนการที่ใช้ในการวัดปริมาณของข้อความ หรือเพื่อการวิเคราะห์ข้อมูลต่างๆ เช่น ในงานวิจัยทางด้านภาษาศาสตร์ การเขียนบทความวิชาการ หรือการทำงานที่เกี่ยวข้องกับข้อความ เพื่อให้สามารถปรับแต่งหรือปรับปรุงได้ตามความเหมาะสม [7]

2.8 การวิเคราะห์ความรู้สึก (Sentiment analysis)

ความรู้สึก (Sentiment) หมายถึง การแสดงถึงอารมณ์ความรู้สึกในเชิงบวก เชิงลบ หรือเป็นกลาง ซึ่งสะท้อนออกมาผ่านภาษาเขียน และภาษาพูด

การวิเคราะห์ความรู้สึก (Sentiment analysis) คือวิธีการที่มีประสิทธิภาพในการประเมินและวิเคราะห์ความรู้สึกที่แฝงอยู่ในภาษาเขียน และภาษาพูด เพื่อจะสามารถระบุได้ว่าการแสดงออกทางความคิดเป็นไปในเชิงบวก เชิงลบ หรือเป็นกลาง [8]

รูปที่ 2.1 ตัวอย่างการทำ Sentiment analysis

2.9 ทฤษฎีความรู้สึกพื้นฐาน (Basic emotions)

ทฤษฎีความรู้สึกพื้นฐาน เป็นทฤษฎีที่เกี่ยวข้องกับจิตวิทยาที่กล่าวถึงการแยกแยะประเภทความรู้สึกพื้นฐานที่มนุษย์มีทั้งหมด ทฤษฎีนี้ถูกนำเสนอโดยนักจิตวิทยา Paul Ekman ซึ่ง Ekman ได้ระบุว่าความรู้สึกพื้นฐาน 6 ความรู้สึกประกอบไปด้วย ความสุข ความเศร้า ความโกรธ ความกลัว ความขยะแขยง และความประหลาดใจ [9]

2.10 การวิเคราะห์ความถี่ (Frequency analysis)

การวิเคราะห์ความถี่เป็นเทคนิคที่ใช้ในการวิเคราะห์ข้อมูล เพื่อทำความเข้าใจความถี่ของข้อมูลที่เกิดขึ้นบ่อย ๆ หรือความถี่ของกลุ่มข้อมูลที่แตกต่างกัน การวิเคราะห์ความถี่มักถูกนำมาใช้ในหลายสาขา เช่น สถิติ ธุรกิจ วิทยาศาสตร์ และการเรียนรู้ของเครื่อง เพื่อศึกษาแนวโน้มและลักษณะการกระจายของข้อมูล

วิธีการวิเคราะห์ความถี่อาจรวมถึงการใช้เครื่องมือทางสถิติ เช่น การพล็อตกราฟความถี่ การสร้างฮิสโตแกรม หรือการคำนวณค่าสถิติพื้นฐาน เพื่อตรวจสอบความสัมพันธ์ระหว่างตัวแปรหรือค้นหาโครงสร้างของข้อมูล ซึ่งมีประโยชน์ในหลากหลายบริบท เช่น การวิเคราะห์พฤติกรรมลูกค้า การวางแผนทางธุรกิจ และการพัฒนาแบบจำลองทาง Machine learning วิธีการวิเคราะห์ความถี่อาจรวมถึงการใช้เครื่องมือทางสถิติ เช่น การพล็อตกราฟความถี่ การสร้างฮิสโตแกรม หรือการคำนวณค่าสถิติพื้นฐาน เพื่อตรวจสอบความสัมพันธ์ระหว่างตัวแปรหรือค้นหาโครงสร้างของข้อมูล ซึ่งมีประโยชน์ในหลากหลายบริบท เช่น การวิเคราะห์พฤติกรรมลูกค้า การวางแผนทางธุรกิจ และการพัฒนาโมเดลการเรียนรู้ของเครื่อง

ในภาพรวม การวิเคราะห์ความถี่เป็นเครื่องมือสำคัญในการทำความเข้าใจ และวิเคราะห์ข้อมูลในหลายสาขา เช่น การนับจำนวนครั้งที่แต่ละคำปรากฏในชื่อหัวข้อข่าวทั้งหมด [10]

2.11 การสรุปเนื้อหา (Text summarization)

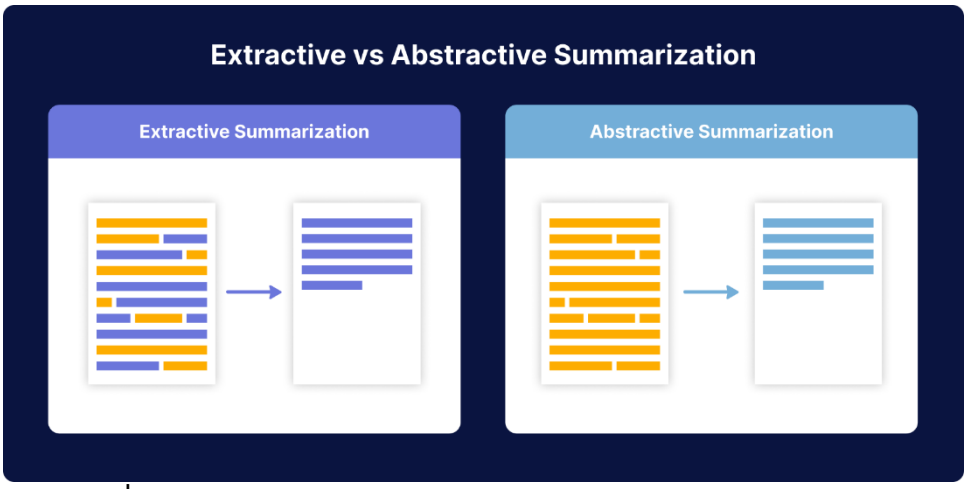
การสรุปเนื้อหา คือกระบวนการย่อยเนื้อหาในเอกสารให้สั้นลงโดยยังคงความหมายเดิมเหมือนกับเอกสารต้นฉบับ การสรุปเนื้อหาสามารถแบ่งออกเป็น 2 แบบหลัก ๆ ได้แก่

1) Extractive summarization เป็นการสรุปความโดยเลือกข้อความหรือประโยคที่สำคัญจากเอกสารต้นฉบับมารวมกัน โดยไม่เปลี่ยนแปลงข้อความหรือประโยคนั้น ๆ
2) Abstractive summarization เป็นการสรุปความโดยสร้างข้อความใหม่จากเอกสารต้นฉบับข้อความใหม่นี้อาจแตกต่างจาก เอกสารต้นฉบับทั้งในด้านไวยากรณ์และเนื้อหาวิธีการนี้มักใช้เทคนิคการเรียนรู้เชิงลึก ถอดความ และสร้างข้อความใหม่ [11]

2.11 การสรุปเนื้อหา (Text summarization) (ต่อ)

ตารางที่ 2.1 เปรียบเทียบความแตกต่างระหว่าง Extractive summarization และ Abstractive summarization

ลักษณะ	Extractive summarization	Abstractive summarization
วิธีการ	เลือกข้อความหรือประโยคจากเอกสารต้นฉบับ	สร้างข้อความใหม่จากเอกสารต้นฉบับ
ความเปลี่ยนแปลง	ไม่เปลี่ยนแปลงข้อความหรือประโยคนั้น ๆ	ข้อความใหม่อาจแตกต่างจากเอกสารต้นฉบับทั้งในด้านไวยากรณ์และเนื้อหา
ความยาก	ง่ายกว่า	ยากกว่า



รูปที่ 2.2 เปรียบเทียบความแตกต่างระหว่าง Extractive summarization และ Abstractive summarization

2.12 การสร้างหัวข้อข่าว (Headline generation)

เป็นกระบวนการสร้างหัวข้อหรือชื่อเรื่องสำหรับบทความต้นฉบับ โดยทั่วไปทำโดยเลือกคำหรือวลีที่สื่อถึงเนื้อหาสำคัญที่สุดของข้อความซึ่งทั้งการสรุปเนื้อหา และการสร้างหัวข้อข่าวต่างก็เป็นการย่อข้อความให้สั้นลงโดยมุ่งเน้นไปที่การนำเสนอข้อมูลสำคัญที่สุดของข้อความ แต่การสร้างหัวข้อข่าวจะเน้นความกระชับของหัวข้อทำให้อาจมีการตัดใจความสำคัญบางอย่างจากบทความต้นฉบับออกไป [12]

2.13 ทรานฟอร์เมอร์โมเดล (Transformer model)

ทรานฟอร์เมอร์โมเดลเป็นสถาปัตยกรรมการเรียนรู้เชิงลึกที่ถูกเสนอครั้งแรกในปี 2017 โดย Vaswani et al. ในงานวิจัยเรื่อง "Attention is All You Need" โมเดลนี้ใช้กลไกความสนใจแบบหลายหัว (Multi-head attention) ในการประมวลผลข้อมูล ซึ่งช่วยให้สามารถเรียนรู้ความสัมพันธ์ที่ซับซ้อนระหว่างคำต่าง ๆ ในข้อความได้ดีกว่าสถาปัตยกรรมการเรียนรู้เชิงลึกแบบเดิมๆ ทรานฟอร์เมอร์โมเดลประกอบด้วย 2 ส่วนหลักได้แก่ [13]

1) **ตัวเข้ารหัส (Encoder)** ทำหน้าที่ประมวลผลข้อมูลรับเข้าเพื่อสร้างเป็นชุดข้อมูลตัวแทน (Representation) ที่สามารถนำมาประมวลผลต่อได้
2) **ตัวถอดรหัส (Decoder)** ทำหน้าที่สร้างข้อมูลส่งออกจากชุดข้อมูลตัวแทน

โมเดลนี้สามารถนำมาใช้งานด้านการประมวลผลภาษาธรรมชาติได้หลากหลาย เช่น การแปลภาษา (Machine translation) การทำความเข้าใจภาษาธรรมชาติ (Natural-language understanding) และการสร้างข้อความ [13]

2.14 Bidirectional Encoder Representations from Transformers : BERT

BERT เป็นโมเดลภาษาขนาดใหญ่ที่ได้รับการพัฒนาโดย Google AI ในปี ค.ศ. 2018 BERT ซึ่งอาศัยสถาปัตยกรรม Transformer model แต่มีหลักการประมวลผลที่แตกต่างกัน คือ BERT เป็นโมเดลแบบสองทิศทาง (Bi-directional) หมายความว่าสามารถประมวลผลข้อมูลได้ทั้งจากซ้ายไปขวา และจากขวาไปซ้าย ในขณะที่ทรานฟอร์เมอร์โมเดลเป็นโมเดลที่มีความสามารถในการทำงานทั้งแบบทิศทางเดียว (Uni-directional) และแบบสองทิศทางซึ่งแตกต่างจากโมเดลอื่น ๆ ที่มีเฉพาะแบบทิศทางเดียวหรือแบบสองทิศทางเท่านั้น เช่น Recurrent Neural Network : RNN [14]

2.15 WangchanBERTa

WangchanBERTa เป็นโมเดลภาษาขนาดใหญ่ (Large Language Model : LLM) ภาษาไทยที่ได้รับการพัฒนาโดยสถาบันวิจัยปัญญาประดิษฐ์ประเทศไทย (AI Research Institute Thailand) ในปี ค.ศ. 2021 WangchanBERTa อาศัยสถาปัตยกรรม RoBERTa ซึ่งเป็นโมเดลภาษาขนาดใหญ่ที่ได้รับการปรับปรุงจาก BERT WangchanBERTa ถูกเทรนบนชุดข้อมูลขนาดใหญ่ของข้อความภาษาไทย ซึ่งรวมถึงข้อความจากแหล่งต่าง ๆ เช่น วิกิพีเดียภาษาไทย ข่าว บทความและโซเชียลมีเดีย WangchanBERTa ประสบความสำเร็จอย่างมากในงานการประมวลผลภาษาธรรมชาติภาษาไทยหลายงาน เช่น การแปลภาษา การทำความเข้าใจภาษาธรรมชาติ และการสร้างข้อความ [15]

2.16 Multilingual Translation Model 5 : MT5

MT5 คือโมเดลภาษาขนาดใหญ่ที่พัฒนาโดย Google AI โมเดลนี้ได้รับการฝึกฝนบนชุดข้อมูลที่มีชื่อว่า MC4 (Multilingual common crawl) ซึ่งเป็นชุดข้อมูลภาษาขนาดใหญ่ที่ประกอบ ไปด้วยภาษากว่า 101 ภาษา MT5 มีสถาปัตยกรรมแบบทรานฟอร์เมอร์โมเดลที่ประกอบไปด้วยตัวเข้ารหัส และตัวถอดรหัสมีโดยมีตัวถอดรหัสเป็นตัวกำหนดโครงสร้างข้อมูลส่งออก (Output) ทำให้ MT5 เหมาะสำหรับงานการประมวลผลภาษาธรรมชาติประเภทการแปลภาษาและการสรุปข้อความ [16]

2.17 Generative Pre-trained Transformer : GPT

GPT คือโมเดลภาษาขนาดใหญ่ที่พัฒนาโดย OpenAI โมเดลนี้ได้รับการฝึกฝนบน ชุดข้อมูลขนาดใหญ่ที่ประกอบไปด้วยหลายภาษา เช่น BookCorpus Common crawl เป็นต้น GPT มีสถาปัตยกรรมแบบทรานฟอร์เมอร์โมเดลที่ประกอบด้วยตัวเข้ารหัสเพียงอย่างเดียวทำให้มีโครงสร้างของข้อมูลส่งออกที่ยืดหยุ่น GPT จึงเหมาะสำหรับงานประเภท การสร้างข้อความ [17]

2.18 Sequence to Sequence Model : Seq2Seq Model

Seq2Seq จะประกอบด้วย 2 ฝั่ง เรียกว่าฝั่งตัวเข้ารหัสภายในเป็นโมเดลแบบ RNN ทำหน้าที่รับข้อความภาษาต้นทางมา แล้วแปลงให้อยู่ในรูปของเว็กเตอร์ (Vector) และฝั่งตัวถอดรหัสภายในเป็นโมเดลแบบ RNN เช่นกัน ทำหน้าที่รับเวกเตอร์ไปสร้างเป็นข้อความภาษาปลายทางที่ต้องการ เปรียบได้ง่าย ๆ ว่า เป็นโมเดลแบบ RNN (LSTM, GRU) 2 ตัวต่อกัน รวมกันเป็นตัวเดียว [18]

2.19 การปรับแต่งโมเดล (Fine tuning)

การปรับแต่งโมเดลเป็นกระบวนการปรับแต่งโมเดลการเรียนรู้ของเครื่องหรือโมเดลการเรียนรู้เชิงลึกที่ฝึกสอนมาแล้วให้ทำงานได้ดีขึ้นกับงานหรือข้อมูลเฉพาะเจาะจง กระบวนการนี้มักใช้เมื่อมีโมเดลที่สร้างเสร็จแล้วต้องการพัฒนาประสิทธิภาพให้ทำงานได้ดีขึ้น โดยใช้ข้อมูลที่มีอยู่เพื่อฝึกสอนและปรับค่าพารามิเตอร์ให้เหมาะสมกับงานหรือข้อมูลใหม่ ขั้นตอนหลักของการการปรับแต่งโมเดลประกอบด้วย

1) **การเลือกโมเดลเริ่มต้น :** เลือกโมเดลที่มีโครงสร้างหรือความสามารถใกล้เคียงกับงานหรือข้อมูลที่ต้องการ
2) **การปรับแต่งโมเดล :** ปรับค่าพารามิเตอร์ของโมเดลให้เหมาะกับงานหรือข้อมูลใหม่ อาจเป็นการปรับค่าน้ำหนักในเลเยอร์ต่าง ๆ หรือปรับการเรียนรู้ในแบบถอยหลัง (Backpropagation)

2.19 การปรับแต่งโมเดล (Fine tuning) (ต่อ)

3) **การฝึกสอน :** ฝึกโมเดลใหม่ด้วยข้อมูลที่มีอยู่ โดยใช้กระบวนการปรับแต่งที่กำหนดไว้
4) **การประเมินและปรับปรุง :** ประเมินประสิทธิภาพของโมเดลที่ปรับแต่งแล้วด้วยข้อมูลทดสอบ และปรับปรุงโมเดลตามผลการประเมิน

การ Fine-tuning มักใช้ในหลาย ๆ งานและสถานการณ์ เช่น :

1) **การถ่ายทอดเรียนรู้ (Transfer learning) :** ใช้โมเดลที่ฝึกสอนไว้กับข้อมูลหนึ่งมาใช้กับงานหรือข้อมูลอื่น ๆ โดยการปรับแต่งโมเดลเพื่อปรับให้ทำงานได้ดีขึ้น
2) **การปรับแต่งโมเดลการทำนาย (Model adaptation) :** ปรับแต่งโมเดลที่มีอยู่ให้ทำงานได้ดีกับข้อมูลหรือเงื่อนไขใหม่
3) **การปรับแต่งสมรรถนะ (Performance tuning) :** ปรับแต่งโมเดลเพื่อปรับปรุงประสิทธิภาพ ความแม่นยำ หรือเวลาการทำงาน [19]

2.20 การออกแบบคำสั่งพร้อมพ์ (Prompt engineering) และการปรับแต่งคำสั่งพร้อมพ์ (Prompt tuning)

การออกแบบคำสั่งพร้อมพ์เป็นการออกแบบคำสั่งพร้อมพ์อย่างเป็นระบบเพื่อให้ได้ผลลัพธ์ที่ต้องการจากโมเดลประมวลผลภาษาธรรมชาติโดยจะครอบคลุมทั้งการกำหนดโครงสร้าง เนื้อหา และรูปแบบของคำสั่งพร้อมพ์เพื่อให้เหมาะสมกับงานที่ใช้ การออกแบบคำสั่งพร้อมพ์จะเน้นการออกแบบคำสั่งพร้อมพ์ที่มีประสิทธิภาพและสามารถใช้ได้ดีกับโมเดลประมวลผลภาษาธรรมชาติ

การปรับแต่งคำสั่งพร้อมพ์เป็นการปรับแต่งทั้งพารามิเตอร์และคำสั่งพร้อมพ์ (Prompt) เพื่อให้ได้ผลลัพธ์ที่ดีขึ้นจากการใช้โมเดลประมวลผลภาษาธรรมชาติ โดยจะมีการปรับค่าของพารามิเตอร์ต่างๆ ในโครงสร้างของคำสั่งพร้อมพ์ เช่น การจัดรูปแบบ การใช้ข้อความเพื่อให้ได้ผลลัพธ์ที่ต้องการ การปรับแต่งคำสั่งพร้อมพ์จะเน้นการปรับแต่งส่วนของคำสั่งพร้อมพ์เป็นหลัก เพื่อให้ได้ผลลัพธ์ที่ดีขึ้นจากการใช้โมเดลประมวลผลภาษาธรรมชาติ [20]

2.21 ค่าเฉลี่ยที่ถ่วงน้ำหนัก (Weighted average)

ค่าเฉลี่ยที่ถ่วงน้ำหนัก เป็นเครื่องมือที่ช่วยในการทำคำนวณค่าเฉลี่ยโดยให้ความสำคัญแตกต่างกันให้แก่ข้อมูลแต่ละส่วนตามน้ำหนักที่กำหนดไว้ล่วงหน้า การคำนวณค่าเฉลี่ยด้วยวิธีนี้มีข้อได้เปรียบที่สำคัญ เนื่องจากมันช่วยให้เสามารถให้ความสำคัญกับข้อมูลที่สำคัญมากขึ้นในกระบวนการวิเคราะห์ เมื่อต้องการคำนวณค่าเฉลี่ยที่ถ่วงน้ำหนักมีขั้นตอนต่อไปนี้

1) กำหนดน้ำหนัก (Weight) กำหนดน้ำหนักที่ต้องการให้แก่แต่ละข้อมูล โดยให้ความสำคัญสูงสุดเป็น 1 หรือ 100% และความสำคัญต่ำสุดเป็น 0 หรือ 0% ซึ่งความสำคัญนี้สามารถเป็นค่าที่กำหนดล่วงหน้า หรืออาจไปอิงจากเกณฑ์ที่เกี่ยวข้องกับการวิเคราะห์

2.21 ค่าเฉลี่ยที่ถ่วงน้ำหนัก (Weighted average) (ต่อ)

2) คูณค่าข้อมูลด้วยน้ำหนัก หลังจากกำหนดน้ำหนักแล้ว ต้องคูณค่าข้อมูลแต่ละชุดด้วยน้ำหนักที่กำหนด
3) รวมผลลัพธ์ทั้งหมด นำผลคูณทุกตัวมาบวกกันเพื่อให้ได้ผลรวมของผลลัพธ์ทั้งหมด
4) หารผลรวมด้วยผลรวมของน้ำหนัก หลังจากที่คูณและรวมผลลัพธ์ทุกตัวแล้ว ให้นำผลรวมมาหารด้วยผลรวมของน้ำหนักทั้งหมด เพื่อให้ได้ค่าเฉลี่ยที่สมบูรณ์ [21]

2.22 การวัดประสิทธิภาพ (Evaluation standard)

2.22.1 Recall-Oriented Understudy for Gisting Evaluation : ROUGE

ROUGE เป็นการวัดประสิทธิภาพมีหน้าที่ช่วยให้ทราบถึงประสิทธิภาพของโมเดลที่ถูกสร้างขึ้นมา ในส่วนของการสรุปเนื้อหาจะนิยมใช้ ROUGE ซึ่งเป็นวิธีการวัดความคล้ายคลึงกันโดยคำนวณจากค่าเมตริก เช่น Recall ซึ่ง ROUGE ก็จะมีเกณฑ์การให้คะแนนหลายแบบ [19] เช่น

1) ROUGE-N คำนวณจากคำที่ตรงกันเป็นชุด ๆ ชุดละ n คำระหว่างมาตรฐานอ้างอิงและบทสรุปที่ต้องการทดสอบ ส่วนใหญ่จะใช้ 1 และ 2 คำ
2) ROUGE-L โดย L มาจาก (Longest common subsequence : LCS) จะเป็นการเลือกค่า n จาก n-gram แบบอัตโนมัติ โดยนำมาจากจำนวนคำที่ตรงกันต่อเนื่องมากที่สุด
3) ROUGE-W มาจาก (Weighted longest common Subsequence : WLCS) คำนวณโดย ให้คะแนนคำที่ตรงตามมาตรฐานอ้างอิง และจะได้คะแนนเพิ่มขึ้นถ้ามีจำนวนคำที่ตรงกัน ต่อเนื่องมากๆ
4) ROUGE-S คำนวณจากคำที่ตรงกันแบบสคิปไบแกรม (Skip-bigram) ระหว่างมาตรฐานอ้างอิงและบทสรุปที่ต้องการทดสอบ ลักษณะเหมือนการจับคู่คำให้ครบโดยห้ามสลับตำแหน่ง เช่นประโยค ABCD จะได้ ไบแกรมในรูปแบ AB, AC, AD, BC, BD และ CD
5) ROUGE-SU เป็นวิธีการที่เสริมจาก ROUGE-S คำนวณจากคำที่ตรงกันแบบสคิปไบแกรมร่วมกับยูนิแกรม (Unigram) คือมีการนับคำเดี่ยวที่ตรงกันระหว่างมาตรฐานอ้างอิงและ บทสรุปที่ต้องการทดสอบ หมายความว่าถ้าในไบแกรมนั้นมีเพียง 1 คำที่ตรงกันก็จะได้ คะแนนต่างจากวิธีก่อนหน้าจะถือว่าไม่ตรงกันเลย

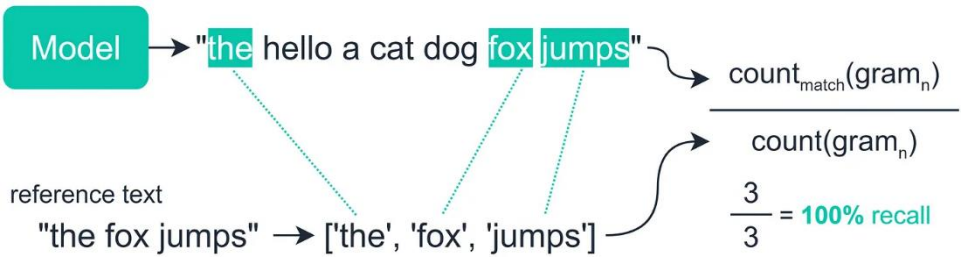
จากที่อธิบายมาข้างต้น ผู้จัดทำได้ใช้การวัดประสิทธิภาพในรูปแบบ ROUGE-N (N-gram) โดยจะคำนวณได้ดังนี้ [22]

$$\mathbf{ROUGE-N} = \frac{\mathbf{Number\ of\ overlapping\ N-grams\ in\ reference\ and\ generated\ summaries}}{\mathbf{Number\ of\ N-grams\ in\ reference\ summaries}}$$

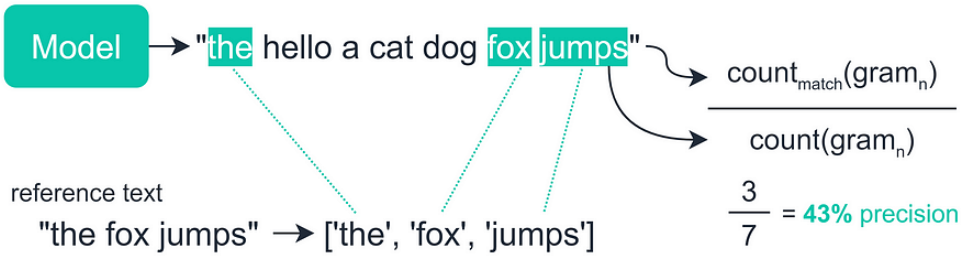
2.22 การวัดประสิทธิภาพ (Evaluation standard) (ต่อ)

โดย Number of overlapping N-grams in reference and generated summaries หมายถึงจำนวนของ N-grams ที่ซึ่งเป็นลำดับของ N คำหรือวลีที่ตรงกันระหว่างข้อความที่สร้างขึ้น (Generated summaries) และข้อความต้นฉบับ (Reference summaries) ที่ถูกให้เปรียบเทียบ

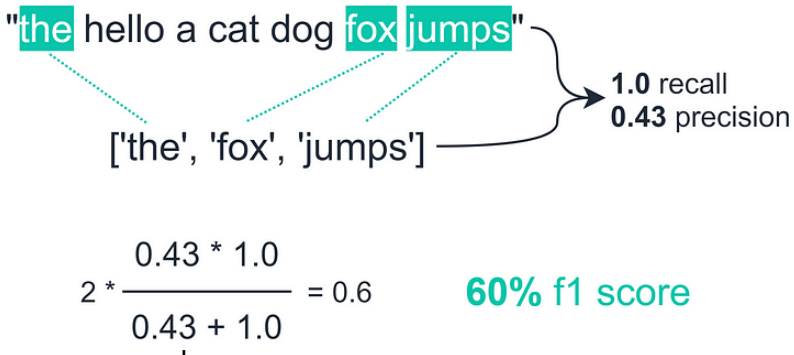
Number of N-grams in reference summaries หมายถึงจำนวนของ N-grams ทั้งหมดในข้อความต้นฉบับที่ถูกใช้เพื่อเปรียบเทียบความคล้ายคลึงในการคำนวณ ROUGE-N



รูปที่ 2.3 ตัวอย่างการคำนวณค่า Recall ของ ROUGE-1



รูปที่ 2.4 ตัวอย่างการคำนวณค่า Precision ของ ROUGE-1



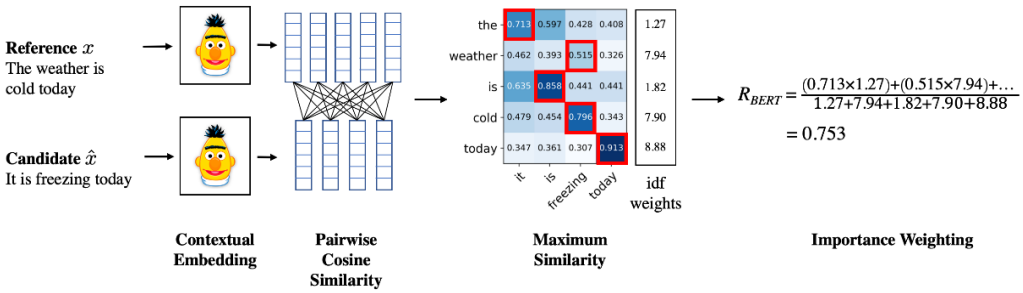
รูปที่ 2.5 ตัวอย่างการคำนวณค่า F1 ของ ROUGE-1

2.22.2 BERTScore

BERTScore เป็นเมตริกที่ใช้ในการประเมินคุณภาพของข้อความที่ถูกสร้างขึ้นโดยโมเดลการสร้างข้อความโดย BERTScore มีข้อได้เปรียบเหนือเมตริกอื่น ๆ เช่น ROUGE ในบางงาน เนื่องจาก BERTScore สามารถพิจารณาความหมายของคำแทนที่จะเน้นเพียงการ

2.22.2 BERTScore (ต่อ)

ตรงกันของคำตัวอักษร ตัวอย่างเช่น คำว่า "Cold" และ "Freezing" มีความหมายที่ใกล้เคียงกันแต่เขียนต่างกัน ซึ่งเป็นข้อจำกัดของเมตริกอย่าง ROUGE ในการประเมินคุณภาพของข้อความที่ถูกสร้างขึ้นมาโดยคำนวณจากคำที่เขียนเหมือนกัน ในขณะที่ BERTScore จะใช้โมเดล BERT ในการเปรียบเทียบความหมายของคำที่ใช้ในข้อความที่ถูกประเมิน ทำให้สามารถประเมินคุณภาพได้ดีกว่าเมตริกที่เน้นเฉพาะความเหมือนกันของคำ [23]



รูปที่ 2.6 ตัวอย่างการคำนวณค่า BERTScore

2.23 งานวิจัยที่เกี่ยวข้อง

PING LI et al. (2021) ได้ศึกษาการสร้างโมเดลสรุปหัวข่าวโดยใช้เทคนิคของระบบประสาทเทียม ซึ่งแสดงผลลัพธ์ที่ดีเป็นอย่างมาก เมื่อนำวิธีการของระบบประสาทเทียมมาใช้ในการสรุปข้อความ บทความนี้มุ่งเน้นการสร้างโมเดลสังเคราะห์หัวข้อข่าว โดยได้นำเสนอโมเดลสังเคราะห์หัวข้อข่าวที่มีพื้นฐานอยู่บนโมเดลการใช้งานร่วมกับโมเดลที่ฝึกสอนมาแล้ว และได้นำเสนอโมดูลการนำเข้าคุณลักษณะที่หลากหลาย โมเดลสังเคราะห์หัวข้อข่าวที่นำเสนอในบทความนี้ มีเพียงตัวถอดรหัสที่รวมกลไกตัวชี้และคุณลักษณะภาษา N-gram ในขณะที่โมเดลสร้างอื่น ๆ ใช้โครงสร้างการเข้ารหัส-การถอดรหัส การทดลองบนชุดข้อมูลข่าวแสดงให้เห็นว่า โมเดลที่นำเสนอในบทความนี้มีผลลัพธ์ในการสังเคราะห์หัวข้อข่าวที่เทียบเท่ากับโมเดลอื่น ๆ [24]

FUCHENG YOU et al. (2020) ได้ศึกษาการผสมข้อมูลที่เกี่ยวข้องกับหัวข้อและความสัมพันธ์ทางด้านไวยากรณ์สำหรับการสรุปข้อความ ด้วยการพัฒนาล้ำหน้าของการเรียนรู้ลึก โมเดลที่ถูกฝึกก่อนหน้านี้ได้ทำให้เกิดผลลัพธ์ที่ดีในด้านการประมวลผลภาษาธรรมชาติ อย่างไรก็ตาม การวิจัยในการสรุปข้อความยังไม่ได้ถึงตรงนั้นที่คนต้องการ ระบบสรุปที่มีคุณภาพสูงต้องให้ความสำคัญกับเนื้อหาของเอกสารและความคล้ายคลึงระหว่างสรุปและเอกสารต้นฉบับ ในบทความนี้ได้นำเสนอการผสมข้อมูลที่เกี่ยวข้องกับหัวข้อและความสัมพันธ์นั้นในการสรุปข้อความ ซึ่งขึ้นอยู่บน

2.23 งานวิจัยที่เกี่ยวข้อง (ต่อ)

FUCHENG YOU et al. (2020) (ต่อ) Fine-tuning BERT (TIF-SR) โดยหลักการ โดยพิจารณาบทบาทสำคัญของข้อมูลเกี่ยวกับหัวข้อในการสร้างสรุป บทความได้สกัดคำหลักของหัวข้อและผสมเข้ากับเอกสารต้นฉบับเป็นส่วนหนึ่งของข้อมูลนำเข้า นอกจากนี้ เพิ่มความคล้ายคลึงระหว่างสรุปที่สร้างขึ้นและเอกสารต้นฉบับ โดยการคำนวณความคล้ายคลึงปรากฏทางด้านไวยากรณ์ ทำให้คุณภาพของบทคัดย่อเพิ่มขึ้น ข้อมูลการทดลองแสดงว่าดัชนี ROUGE และความอ่านง่ายได้ปรับปรุงในโมเดลนี้ [25]

Pongsatorn Harnmetta และ Taweesak Samanchuen (2022) ผู้เขียนนำเสนอแนวคิดของการวิเคราะห์ความรู้สึกในตลาดหุ้นและอธิบายวิธีการนำมันมาใช้ในการพยากรณ์ราคาหุ้นและสัญญาซื้อขาย ทางบทความยังหาถึงข้อจำกัดของวิธีการวิเคราะห์ความรู้สึกที่ทั่วไปและความจำเป็นต้องใช้เทคนิคที่ซับซ้อนมากขึ้น ทางบทความได้เสนอการวิเคราะห์ความรู้สึกของรีวิวหุ้นไทยโดยใช้ทรานฟอร์เมอร์โมเดล บทความอธิบายว่าระบบถูกออกแบบมาเพื่อที่จะแก้ไขข้อจำกัดของวิธีการทั่วไปและทำนายความรู้สึกของหุ้นในบริบทไทยได้อย่างแม่นยำ [26]

Sheher Bano และ Shah Khalid (2022) ได้ศึกษาเกี่ยวกับการสรุปที่มีหลากหลายรูปแบบซึ่งแต่ละวิธีมีข้อได้เปรียบและข้อเสีย อย่างไรก็ตามไม่มีวิธีใดเป็นวิธีที่สมบูรณ์ ซึ่งหมายความว่ายังมีที่ว่างในการพัฒนาในสาขานี้ของ BERT เป็นเครือข่ายทรานส์ฟอร์มหลายชั้นที่ได้รับการฝึกก่อนสำหรับหลายแอปพลิเคชันที่ใช้การสอนเอง อย่างไรก็ตาม เนื่องจากจำกัดความยาวของข้อมูลนำเข้า BERT เหมาะกับข้อความที่สั้นเท่านั้น ดังนั้น เชื่อว่าการใช้ BERT ในการสรุปเอกสารที่ยาวจะเป็นงานที่ท้าทาย บทความได้นำเสนอวิธีการใหม่ที่ BERT สามารถใช้สรุปเอกสารที่ยาว บทความใช้วิธีการแบ่งเอกสารทั้งหมดเป็นชิ้นย่อย ๆ และแต่ละชิ้นมีประโยคหนึ่ง แนวคิดพื้นฐานคือการรับภาพสำรวจจาก BERT แล้วนำมาใช้กับโมเดล การเข้ารหัส-การถอดรหัส ทดลองด้วยชุดข้อมูลวิชาการสองรายการ (arXiv และ PubMed) ผลทดลองแสดงให้เห็นว่าเทคนิคของบทความมีประสิทธิภาพเหนือหลายโมเดลที่ทันสมัยวิจัยนี้ได้กล่าวถึงข้อจำกัดของวิธีการสรุปการสกัดที่มีอยู่และเสนอสถาปัตยกรรมใหม่ที่ใช้ BERT ในการสรุปเอกสารยาววิธีการนี้เกี่ยวข้องกับการแบ่งเอกสารออกเป็นชิ้น ๆ รับการฝังประโยคจาก BERT และการใช้ตัวเข้ารหัส-ถอดรหัสที่ด้านบน BERT เพื่อสรุปผลการทดลองแสดงให้เห็นว่าเทคนิคนี้มีประสิทธิภาพเหนือกว่าโมเดลที่ล้ำสมัยหลายแบบอย่างสม่ำเสมอ ดังที่แสดงให้เห็นในชุดข้อมูลทางวิชาการสองชุด [27]

2.23 งานวิจัยที่เกี่ยวข้อง (ต่อ)

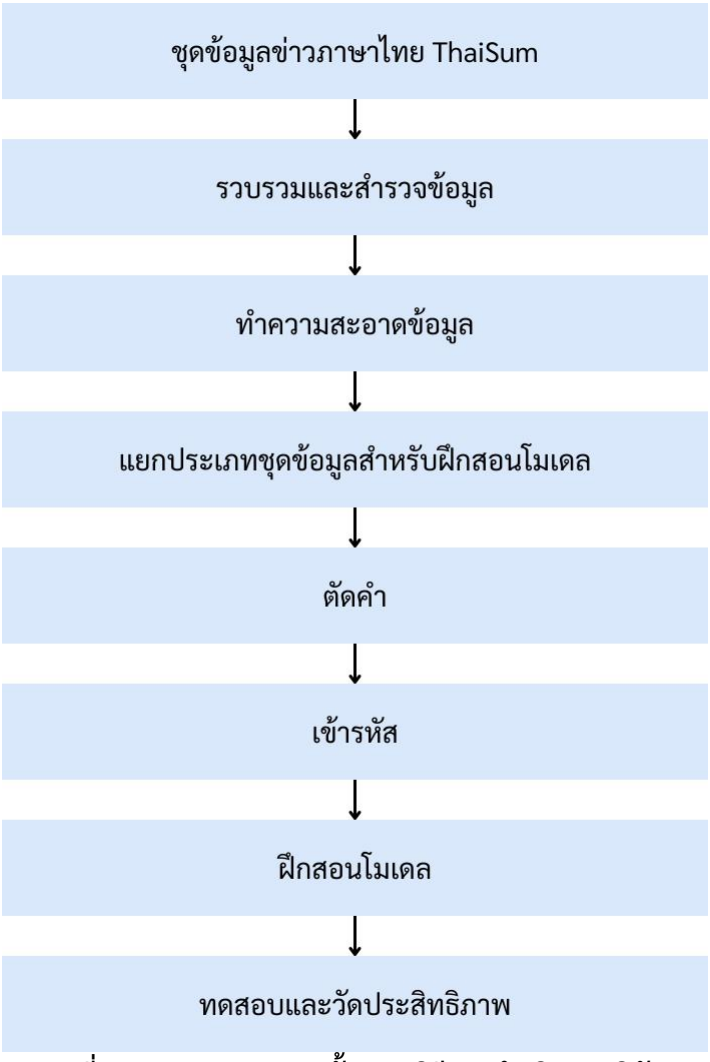
Anandan Chinnalagu และ Ashok Kumar Durairaj (2022) ในงานวิจัยนี้ได้ศึกษาเกี่ยวกับการใช้โมเดล BERT และโมเดลการเรียนรู้เชิงลึกในการประมวลผลภาษาธรรมชาติได้รับการนำมาใช้ในแอปพลิเคชันต่าง ๆ การตลาดทางสังคมและความคิดเชิงบวกของลูกค้าเป็นปัจจัยสำคัญสำหรับธุรกิจออนไลน์หลาย ๆ ราย การทำนายความรู้สึกของลูกค้าอย่างแม่นยำเป็นงานที่สำคัญสำหรับบริษัทที่ต้องการพยายามคาดการณ์ความรู้สึกของลูกค้าจากบทวิจารณ์ออนไลน์ การทำนายความรู้สึกที่แม่นยำเป็นงานที่ใช้เวลามากและท้าทายเนื่องจากปริมาณข้อมูลบทวิจารณ์ลูกค้าที่ไม่มีโครงสร้างมีมากมายมีผลการทดลองก่อนหน้านี้ที่เปิดเผยปัญหาทางประสิทธิภาพ และความไม่แม่นยำในข้อมูลบทวิจารณ์ลูกค้าขนาดใหญ่ [28]

Jiaohong Yao (2022) ผู้คิดค้นได้นำเสนอการสังเคราะห์หัวข้อข่าวที่ปรับตามบุคคลเป้าหมายมุ่งเน้นที่จะสรุปบทความข่าวเป็นหัวข้อข่าวตามความชอบของผู้ใช้ที่ระบุไว้ เช่นนี้สามารถช่วยผู้ใช้กรองข่าวที่น่าสนใจได้เร็วขึ้น และเพิ่มอัตราการคลิกข่าวสำหรับผู้ให้บริการ อย่างไรก็ตาม ในสายงานนี้ เมื่อเรียนรู้ความสนใจของผู้ใช้จากข่าวที่ผู้ใช้คลิกเป็นประวัติ การวิจัยที่มีอยู่เดิมมักเรียนรู้ความสนใจของผู้ใช้ที่ระดับคำ และระดับข่าวเท่านั้น ทำไม่คำนึงถึงระดับประโยคที่สรุปข้อมูลนี้ บทความนี้ ได้นำเสนอ โมเดลผู้ใช้โดยเพิ่มความสนใจในระดับประโยคเพื่อเรียนรู้ความสนใจของผู้ใช้และนำทางการสังเคราะห์หัวข้อข่าว เพื่อให้เป็นไปอย่างละเอียดมากขึ้น จากชั้นการให้ความสนใจประโยคและข่าวถูกแสดงในรูปแบบผลรวมที่น้ำหนักของคำและประโยคตามลำดับ เพื่อสำรวจความสัมพันธ์ระหว่างเนื้อหาข่าวที่แตกต่างกัน (หัวข้อข่าว เนื้อหา และข้อมูลเนื้อหา) เวกเตอร์คำค้นในชั้นการให้ความสนใจถูกแทนที่ด้วยเนื้อหาข่าว การทดลองบนชุดข้อมูล PENS แสดงให้เห็นว่าประสิทธิภาพของโมเดลเหล่านี้ดีกว่าโมเดลเบสไลน์ตามเกณฑ์การประเมิน ROUGE ในที่สุดทิศทางของงานที่จะมีในอนาคตบางประการ รวมถึงการกระทำที่เกี่ยวข้องกับระดับความสนใจ และเนื้อหาที่แตกต่างกัน [29]

Nutthanit Wiwatbutsiri et al. (2022) ศึกษาเกี่ยวกับการสร้างคำถาม (Question generation : QG) ในภาษาอังกฤษและ ภาษาไทย แต่ภาษาไทยมีชุดข้อมูลที่น้อยกว่าซึ่งภาษาอังกฤษมีมากกว่าหนึ่งล้าน คู่คำถาม-คำตอบ เทียบกับในภาษาไทยมีเพียงประมาณ 12,000 คู่คำถาม-คำตอบ บทความนี้ได้นำเสนอวิธีการปรับปรุงการสร้างคำถามที่ไม่ขึ้นกับคำตอบในภาษาไทยจากชุดข้อมูลขนาดไม่เพียงพอ การประเมินของบทความแสดงให้เห็นว่าโมเดล QG ที่ได้รับการฝึกฝนจากโมเดลที่ได้รับการฝึกฝนก่อนหน้า MT5 จากชุดข้อมูลไทยได้คะแนน BLEU-1 ที่ 56.19 ทางบทความได้นำเสนอวิธีการสร้างข้อมูลสังเคราะห์และกลไกเพิ่มเติมโดยใช้โมเดลที่ได้รับการฝึกฝนก่อนหน้าเพียงตัวเดียว [30]

บทที่ 3
วิธีการดำเนินงานวิจัย

ในบทนี้จะกล่าวถึงขั้นตอนวิธีดำเนินงานวิจัยในการพัฒนาโมเดลสังเคราะห์หัวข้อข่าวที่สื่อถึงความรู้สึกโดยนำแนวคิดของการเรียนรู้เชิงลึกมาใช้ ซึ่งขั้นตอนการดำเนินงานจะประกอบไปด้วยขั้นตอนหลัก ๆ 4 ขั้นตอน คือ 1) การรวบรวมและสำรวจข้อมูล 2) การทำความสะอาดข้อมูล 3) แยกประเภทชุดข้อมูล และ 4) การฝึกสอนโมเดล แต่ละขั้นตอนประกอบด้วยรายละเอียดย่อย ๆ ดังที่แสดงในรูป



รูปที่ 3.1 แผนภาพแสดงขั้นตอนวิธีการดำเนินงานวิจัย

3.1 การรวบรวมและสำรวจข้อมูล

คณะผู้จัดทำรวบรวมข้อมูลข่าวมาจากชุดข้อมูล ThaiSum บนเว็บไซต์ Huggingface ซึ่งเป็นชุดข้อมูลที่ประกอบด้วยข่าวจากหลาย ๆ แหล่ง ได้แก่ ไทยรัฐ ไทยพีบีเอส ประชาไท และเดอะแสตนดาร์ด โดยชุดข้อมูลนี้มีจำนวน 358,868 ตัวอย่าง คอลัมน์ของชุดข้อมูลนี้จะประกอบไปด้วย 5 คอลัมน์ ได้แก่

1) คอลัมน์ Title เป็นคอลัมน์ที่เก็บข้อมูลชื่อหัวข้อข่าว
2) คอลัมน์ Body เป็นคอลัมน์ที่เก็บข้อมูลเนื้อหาข่าว
3) คอลัมน์ Summary เป็นคอลัมน์ที่เก็บข้อมูลเนื้อหาข่าวแบบย่อ
4) คอลัมน์ Type เป็นคอลัมน์ที่เก็บข้อมูลประเภทของข่าว
5) คอลัมน์ Tag เป็นคอลัมน์ที่เก็บข้อมูลแท็กของข่าว

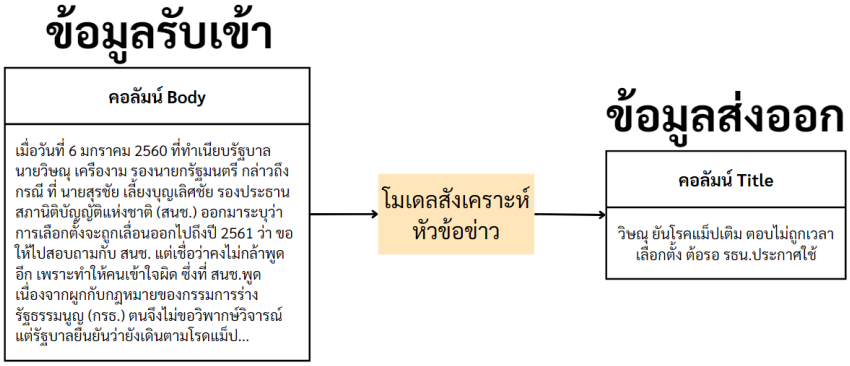
ตารางที่ 3.1 ตัวอย่างของข้อมูลในแต่ละคอลัมน์

คอลัมน์	ตัวอย่างของข้อมูล
คอลัมน์ Title	วิษณุ ยันโรคแม็ปเดิม ตอบไม่ถูกเวลาเลือกตั้ง ต้อรอ รธน.ประกาศใช้
	อังคณา เผยเตรียมร้อง ปอท. หลังถูกโพสต์สร้างความเกลียดชังในโซเชียลฯ
	รองผู้ว่าฯภูเก็ตตรวจสอบที่ดินบนเกาะนาคาน้อยกรมที่ดินตั้งกรรมการสอบสวนแล้ว
คอลัมน์ Body	เมื่อวันที่ 6 ม.ค.60 ที่ทำเนียบรัฐบาล นายวิษณุ เครืองาม รองนายกรัฐมนตรี กล่าวถึงกรณี ที่ นายสุรชัย เลี้ยงบุญเลิศชัย รองประธานสภานิติบัญญัติแห่งชาติ (สนช.) ออกมาระบุว่า การเลือกตั้งจะถูกเลื่อนออกไปถึงปี 2561 ว่าขอให้ไปสอบถามกับ สนช. แต่เชื่อว่าคงไม่กล้าพูดอีก เพราะทำให้คนเข้าใจผิด ซึ่งที่ สนช.พูดเนื่องจากผูกกับกฎหมายของกรรมการร่างรัฐธรรมนูญ(กรธ.)...
	โดยล่าสุดถูกปั่นว่าเรียกร้องให้สังคมให้อภัยเปรี้ยว 6 มิ.ย. 2560 ผู้สื่อข่าวรายงานว่า เมื่อเวลา 21.59 น. ที่ผ่านมาอังคณา นีละไพจิตรกรรมการสิทธิมนุษยชนแห่งชาติ โพสต์ภาพ พร้อมข้อความผ่านเฟซบุ๊ก ซึ่งเป็นภาพบิดเบือนข้อมูลเกี่ยวกับ อังคณา พร้อระบุว่า ทำงานตรวจสอบการละเมิดสิทธิฯมานานพรุ่งนี้(7 มิ.ย.60) 10.00 น. จะไปแจ้ง กองบังคับการปราบปราม...
	วันนี้ (7 เม.ย.2559) นายโชคดี อมรวัฒน์ รองผู้ว่าราชการจังหวัดภูเก็ต เข้าตรวจสอบที่ดินของบริษัท ภูเขาหกลูก ผู้ครอบครองเอกสารสิทธิ์ น.ส.3 ก เลขที่ 3977 เนื้อที่ 24 ไร่เศษ บนเกาะนาคาน้อย ต.ป่าคลอก อ.ถลาง จ.ภูเก็ต ซึ่งตั้งติดกับที่ดินของครอบครัว หิรัญพฤกษ์ หลังจากครอบครัวดังกล่าวโพสต์ข้อความผ่านเฟซบุ๊กว่า มีคนของบริษัทดังกล่าวซึ่งมีอาวุธครบมือ...

ตารางที่ 3.1 ตัวอย่างของข้อมูลในแต่ละคอลัมน์ (ต่อ)

คอลัมน์	ตัวอย่างของข้อมูล
คอลัมน์ Summary	วิษณุ ยันโรดแม็ปตามขั้นตอนเดิม เชื่อ สนช.หยุดพูดขยับเลือกตั้ง ปัดวิจารณ์ยึดตามกรอบเวลา ย้ำเริ่มนับโรดแม็ปเมื่อ รธน.ประกาศใช้
	อังคณา กสม. โพสต์ระบุเตรียมเข้าร้องต่อ ปอท.เรื่องการถูกละเมิดสิทธิ และศักดิ์ศรีความเป็นมนุษย์ ทำให้ถูกเกลียดชังโดยการใช้ภาพ และข้อความเผยแพร่ทางเฟซบุ๊กต่างๆ
	รองผู้ว่าราชการจังหวัดภูเก็ตลงพื้นที่ตรวจสอบเกาะนาคาน้อย ภายหลังที่ครอบครัว หิรัญพฤกษ์ ออกมาระบุว่า มีกลุ่มชายฉกรรจ์เตรียมเข้ามาแผ้วถางพื้นที่บนเกาะ และข่มขู่
คอลัมน์ Type	NaN
	สิทธิมนุษยชน,ไอซีที
	ภูมิภาค
คอลัมน์ Tag	เลือกตั้ง,โรดแม็ป,วิษณุ เครืองาม,ร่างรัฐธรรมนูญ,สนช.
	ปอท.,สร้างความเกลียดชัง,อังคณา นีละไพจิตร,เปรี้ยว
	บุกรุก,ภูเก็ต,เกาะนาคาน้อย,หิรัญพฤกษ์,ThaiPBSnews,ข่าวไทยพีบีเอส

ซึ่งคณะผู้จัดทำได้เลือกคอลัมน์ที่จะใช้ฝึกสอนโมเดลมา 2 คอลัมน์ ได้แก่ คอลัมน์หัวข้อข่าว (Title) และคอลัมน์เนื้อหาข่าว (Body) โดยโครงสร้างของโมเดลจะใช้คอลัมน์เนื้อหาข่าวเป็นส่วนของข้อมูลรับเข้าและใช้คอลัมน์หัวข้อข่าวเป็นข้อมูลส่งออก



รูปที่ 3.2 โครงสร้างข้อมูลรับเข้า/ข้อมูลส่งออกของโมเดล

3.2 การทำความสะอาดข้อมูล

คณะผู้จัดทำได้ทำความสะอาดชุดข้อมูลให้อยู่ในรูปแบบที่เหมาะสมกับการฝึกสอนโมเดลซึ่งประกอบไปด้วยขั้นตอนดังนี้

1) ลบคำหรือสัญลักษณ์ที่ไม่จำเป็นออก

คณะผู้จัดทำได้ลบคำหรือสัญลักษณ์ที่ไม่มีความสำคัญต่อความหมายของหัวข้อข่าวและเนื้อหาข่าว

ตารางที่ 3.2 ตัวอย่างในชุดข้อมูลก่อนลบคำหรือสัญลักษณ์ที่ไม่จำเป็น (คำหรือสัญลักษณ์ที่ไม่จำเป็นจะถูกระบุด้วยเครื่องหมายสัญประกาศ)

คอลัมน์ Title	เจอแล้ว กระบะทับคนตาย รับเหมือนเหยียบอะไร แต่ไม่เห็นคนเจ็บใต้รถ (คลิป)
	สะเทือนแรงมากจนภูเขาถล่ม ธรณีพิโรธที่อัฟกานิสถาน ตายแล้วกว่า 300 (ชมคลิป)
คอลัมน์ Body	จากกรณีคลิปวงจรปิดจับภาพเหตุการณ์เหตุรถจักรยานยนต์ชนกันกลางถนนศรีสะเกษ-ขุขันธ์ ใกล้สี่แยกไทยประสาน ต.ห้วยเหนือ อ.ขุขันธ์ จ.ศรีสะเกษ และต่อมามี.รถกระบะขับมาทับซ้ำผู้บาดเจ็บ.ที่นอนอยู่บนถนน ก่อนจะเดินหน้าถอยหลังและขับออกไป กระทั่งภายหลังทราบว่าผู้บาดเจ็บเสียชีวิตนั้น.ร.ต.อ.บุญลักษณ์ เหลี่ยมแก้ว รอง สว.สส.สภ.ขุขันธ์ กล่าวว่า...
	เมื่อวันที่ 27 ต.ค.58 สำนักข่าวต่างประเทศรายงานความคืบหน้าเหตุการณ์เกิดแผ่นดินไหว ขนาด 7.5 มีจุดศูนย์กลางอยู่ทางภาคตะวันออกเฉียงเหนือของอัฟกานิสถาน เมื่อวันที่ 26 ต.ค.ที่ผ่านมา และแรงสั่นสะเทือนสามารถรับรู้ได้ทั่วภูมิภาคเอเชียใต้ ทั้งปากีสถาน อินเดีย และคาซัคสถาน จนสร้างความเสียหายแก่ชีวิตและอาคารบ้านเรือนพังเสียหายจำนวนมากว่า ยอดผู้เสียชีวิตจาก...

ตารางที่ 3.3 ตัวอย่างในชุดข้อมูลหลังลบคำหรือสัญลักษณ์ที่ไม่จำเป็น

คอลัมน์ Title	เจอแล้ว กระบะทับคนตาย รับเหมือนเหยียบอะไร แต่ไม่เห็นคนเจ็บใต้รถ
	สะเทือนแรงมากจนภูเขาถล่ม ธรณีพิโรธที่อัฟกานิสถาน ตายแล้วกว่า 300
คอลัมน์ Body	จากกรณีคลิปวงจรปิดจับภาพเหตุการณ์เหตุรถจักรยานยนต์ชนกันกลางถนนศรีสะเกษ-ขุขันธ์ ใกล้สี่แยกไทยประสาน ต.ห้วยเหนือ อ.ขุขันธ์ จ.ศรีสะเกษ และต่อมามีรถกระบะขับมาทับซ้ำผู้บาดเจ็บที่นอนอยู่บนถนน ก่อนจะเดินหน้าถอยหลังและขับออกไป กระทั่งภายหลังทราบว่าผู้บาดเจ็บเสียชีวิตนั้นร.ต.อ.บุญลักษณ์ เหลี่ยมแก้ว รอง สว.สส.สภ.ขุขันธ์ กล่าวว่า...

ตารางที่ 3.3 ตัวอย่างในชุดข้อมูลหลังลบคำหรือสัญลักษณ์ที่ไม่จำเป็น (ต่อ)

คอลัมน์ Body	เมื่อวันที่ 27 ต.ค.58 สำนักข่าวต่างประเทศรายงานความคืบหน้าเหตุการณ์เกิดแผ่นดินไหว ขนาด 7.5 มีจุดศูนย์กลางอยู่ทางภาคตะวันออกเฉียงเหนือของอัฟกานิสถาน เมื่อวันที่ 26 ต.ค.ที่ผ่านมา และแรงสั่นสะเทือนสามารถรับรู้ได้ทั่วภูมิภาคเอเชียใต้ ทั้งปากีสถาน อินเดีย และคาซัคสถาน จนสร้างความเสียหายแก่ชีวิตและอาคารบ้านเรือนพังเสียหายจำนวนมากว่า ยอดผู้เสียชีวิตจากโศกนาฏกรรมธรณีพิโรธครั้งนี้ ยังคงเพิ่มขึ้นอย่างต่อเนื่องตามรายงานของเจ้าหน้าที่...

2) ลบตัวอย่างข้อมูลที่ยากต่อการทำความสะอาด

คณะผู้จัดทำได้ลบตัวอย่างหัวข้อข่าวบางหัวข้อที่ไม่สามารถทำความสะอาดได้โดยการลบคำหรือสัญลักษณ์ที่ไม่จำเป็นออกไป ได้แก่ หัวข้อข่าวที่ประกอบไปด้วยสัญลักษณ์ "@" ซึ่งสัญลักษณ์นี้ เมื่อลบออกจากหัวข้อข่าวแล้ว จะทำให้หัวข้อข่าวสูญเสียความหมายเดิมไป

ตารางที่ 3.4 ตัวอย่างในชุดข้อมูลที่ยากต่อการทำความสะอาด

คอลัมน์ Title	รัฐเดินหน้าจัดถนนคนเดิน กระตุ้นเศรษฐกิจทุกวันอาทิตย์ @ถนนสีลม (คลิป)
	สคบ.ชวน นิสิต-นศ. ร่วมประกวดหนังสั้น @ สคบ.ครั้งที่ 2
	@PravitR: ทวีตนี้แด่อากง SMS

3) ลบตัวอย่างหัวข้อข่าวที่เป็นประโยคคำถาม

คณะผู้จัดทำได้ลบตัวอย่างหัวข้อข่าวที่เป็นประโยคคำถามเพื่อป้องกันการสับสนของโมเดลในระหว่างขั้นตอนการฝึกสอนโมเดล

ตารางที่ 3.5 ตัวอย่างในชุดข้อมูลที่เป็นประโยคคำถาม

คอลัมน์ Title	เลเซอร์กำจัดขน เกลี้ยง เนียน ถาวรจริงมั้ย?
	แล้งนี้จะอยู่กันอย่างไร?

4) เลือกเฉพาะตัวอย่างในชุดข้อมูลที่มีความยาวของหัวข้อข่าวอยู่ระหว่าง 8 ถึง 30 คำ

คณะผู้จัดทำเลือกเฉพาะตัวอย่างในชุดข้อมูลที่มีความยาวของหัวข้อข่าวอยู่ระหว่าง 8 ถึง 30 คำ เนื่องจากหัวข้อข่าวที่มีความยาวของหัวข้อข่าวน้อยกว่า 8 คำ มีจำนวนคำที่น้อย

4) เลือกเฉพาะตัวอย่างในชุดข้อมูลที่มีความยาวของหัวข้อข่าวอยู่ระหว่าง 8 ถึง 30 คำ (ต่อ)

เกินไปทำให้ไม่สามารถ สื่อความหมายที่อยู่ในเนื้อหาข่าวได้อย่างชัดเจน และหัวข้อข่าวที่มีความยาวของหัวข้อข่าวมากกว่า 30 คำ มีจำนวนคำที่มากเกินไปซึ่งไม่ตรงกับลักษณะของหัวข้อข่าวที่จำเป็นต้องมีความสั้นและกระชับ

ตารางที่ 3.6 ตัวอย่างในชุดข้อมูลที่มีความยาวของหัวข้อข่าวน้อยกว่า 8 คำ

คอลัมน์ Title	เอาจริงซักที
	แรงงานคนละกิ่ง
	อารมณ์ขันเลือน สมองเสื่อมพาล

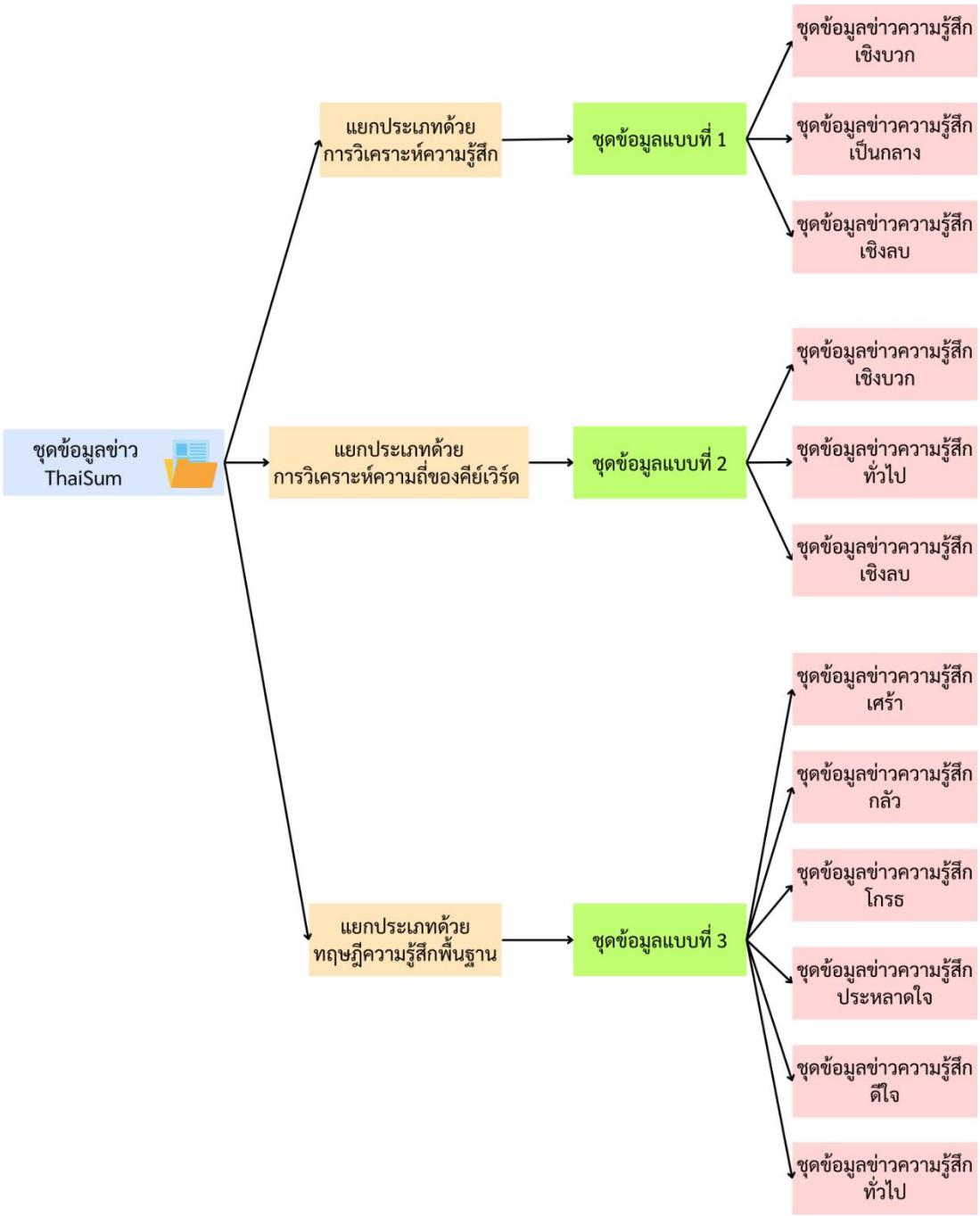
ตารางที่ 3.7 ตัวอย่างในชุดข้อมูลที่มีความยาวของหัวข้อข่าวมากกว่า 30 คำ

คอลัมน์ Title	ศูนย์ทนายเพื่อสิทธิมนุษยชนเผย นศ.มธ.เคยชู 3 นิ้วหน้า วิษณุ ถูกตำรวจตามประกบติด 48 ชั่วโมง ก่อน ประวิตร ลงพื้นที่ลำปาง
	ทรูประกาศใช้ ‘Growable Bag’ ถุงรีไซเคิลออกดอกได้แทนถุงพลาสติกตั้งเป้าลด CO2 600,000 กิโลกรัมคาร์บอนไดออกไซด์เทียบเท่าต่อปี

3.3 การแยกประเภทชุดข้อมูลสำหรับฝึกสอนโมเดล

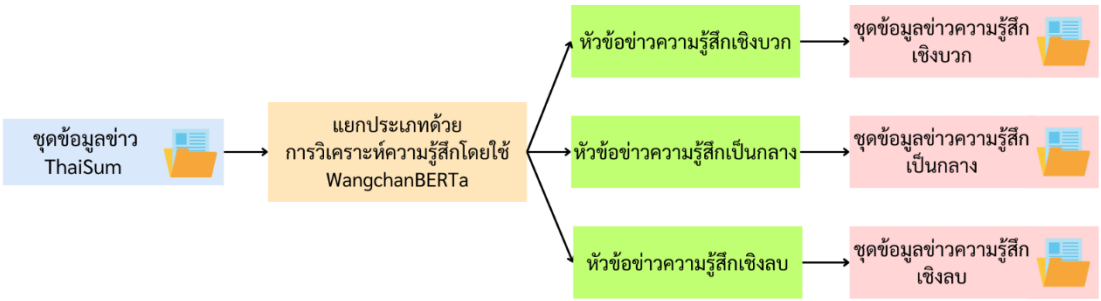
คณะผู้จัดทำได้แยกประเภทชุดข้อมูลตามความรู้สึกที่หัวข้อข่าวได้สื่อออกมาโดยมีเกณฑ์ในการแยกที่แตกต่างกัน 3 รูปแบบ ดังนี้

1) ชุดข้อมูลแบบที่ 1 แยกประเภทด้วยการวิเคราะห์ความรู้สึก
2) ชุดข้อมูลแบบที่ 2 แยกประเภทด้วยการวิเคราะห์ความถี่ของคีย์เวิร์ดที่เน้นย้ำความรู้สึก
3) ชุดข้อมูลแบบที่ 3 แยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐานและใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด



รูปที่ 3.3 แผนภาพแสดงเกณฑ์การแยกประเภทชุดข้อมูล

1) ชุดข้อมูลแบบที่ 1 แยกประเภทด้วยการวิเคราะห์ความรู้สึก



รูปที่ 3.4 แผนภาพแสดงการแยกประเภทชุดข้อมูลแบบที่ 1

คณะผู้จัดทำได้แยกประเภทหัวข้อข่าวด้วยการวิเคราะห์ความรู้สึก (Sentiment analysis) โดยแบ่งประเภทของหัวข้อข่าวออกเป็น 3 ประเภท คือ

1) หัวข้อข่าวความรู้สึกเชิงบวก
2) หัวข้อข่าวความรู้สึกเชิงลบ
3) หัวข้อข่าวความรู้สึกเป็นกลาง

ในการวิเคราะห์ความรู้สึกนี้ คณะผู้จัดทำใช้โมเดล WangchanBERTa เป็นเครื่องมือช่วยในการวิเคราะห์ความรู้สึก

ตารางที่ 3.8 ตัวอย่างหัวข้อข่าวที่ได้ผ่านการวิเคราะห์ความรู้สึก

ความรู้สึก	ตัวอย่างหัวข้อข่าว
ความรู้สึกเชิงลบ	อ้างเป็น รองโจ๊ก รีดเงินโยกย้าย
ความรู้สึกเป็นกลาง	พลอย เฌอมาลย์ ชวนดู เพลิงนาง แซ่บ สด ใหม่ ไม่มีรีรัน บทแรงสุด โสเภณีร้อยล้าน
ความรู้สึกเชิงบวก	เล้งแซ่บอร่อยสุด ห้วยขวาง อร่อยเหาะรอบดึกไม่ไปไม่ได้

คณะผู้จัดทำแยกชุดข้อมูลออกเป็น 3 ชุด ตามประเภทที่ได้ทำการจำแนกไว้ ได้แก่

1) ชุดข้อมูลข่าวความรู้สึกเชิงบวก 2,068 ตัวอย่าง
2) ชุดข้อมูลข่าวความรู้สึกเชิงลบ 2,443 ตัวอย่าง
3) ชุดข้อมูลข่าวความรู้สึกเป็นกลาง 2,274 ตัวอย่าง

ตารางที่ 3.9 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเชิงบวก

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเชิงบวก
คอลัมน์ Title	ใครๆก็ชอบของฟรี
	เชียงใหม่เฮ ฝนหยุดตก งานยี่เป็งเชียงใหม่เริ่มมีชีวิตชีวา
คอลัมน์ Body	บรรดานโยบาย แจกแหลก แจกกระจาย แจกกระหน่ำซัมเมอร์เซลส์ ที่พรรคพลังประชารัฐตีปี๊บหาเสียงไว้มากมายมโหฬารก็กลายเป็น ห่วงผูกคอ รัฐบาลต้องเร่งทำตามสัญญาที่หาเสียงไว้กับประชาชน ลืมไม่ได้ เลิกไม่ได้ ลดไม่ได้ เบี้ยวไม่ได้ทุกกรณี แม่ลูกจันทร์ เชื่อว่า นายกฯลุงตู่ จะทำตามนโยบายที่หาเสียงไว้อย่างแน่นอนแต่เมื่อเม็ดเงินในกระเป๋ามีจำกัด รัฐบาลจึงไม่สามารถทำตามนโยบายที่หาเสียงไว้ครบถ้วนพร้อมกันทันที บางนโยบายจำเป็นต้องเลื่อนไปก่อนชั่วคราว นโยบายแรกที่เจอโรคเลื่อนแน่นอน...
	เมื่อวันที่ 6 พ.ย. ผู้สื่อข่าวรายงานบรรยากาศงานประเพณีเดือนยี่เป็งเจียงใหม่ประจำปี 2557 หรืองานลอยกระทง โดยในช่วงเช้าวันที่ 5 พ.ย.จนถึงเวลา 19.00 น.วันที่ 6 พ.ย. ฝนได้ตกลงมาตลอดเวลาทำให้โคมยี่เป็งที่นำมาประดับนับหมื่นใบได้รับความเสีย หายบางส่วน และไม่สามารถเปิดไฟโชว์ได้ รวมทั้งขบวนแห่ที่งดงาม ซึ่งมีนักท่องเที่ยวชมอย่างบางตาตามท้องถนน..

ตารางที่ 3.10 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเชิงลบ

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเชิงลบ
คอลัมน์ Title	ชาวนาทุกข์ซ้ำ โจรขโมยเกี่ยวข้าวในนา หอมมะลิกว่า 8 ไร่
	เซ็งกันไป ผู้โดยสารหลายพัน ติดแหง็กสนามบินฝรั่งเศส เจอหยุดงานประท้วง
คอลัมน์ Body	เมื่อวันที่ 28 พ.ค. 58 ผู้สื่อข่าวได้รับแจ้งว่า มีชาวบ้านหมู่ 6 บ้านแหลมยาง ต.วังบัว อ.คลองขลุง จ.กำแพงเพชร ว่า ได้รับความเดือดร้อนจากคนร้ายที่เข้ามาขโมยเกี่ยวข้าวในนาไปเกือบ 10 ไร่ จึงเดินทางไปตรวจสอบข้อเท็จจริง พบนายจำรัส พิลึก อายุ 52 ปี ผู้ใหญ่บ้านแหลมยาง หมู่ 6 ต.วังบัว อ.คลองขลุง และนายสมบัติ ตาคะ อายุ 47 ปี อยู่บ้านเลขที่ 840 หมู่ 2 ต.ท่ามะเขือ อ.คลองขลุง ชาวนา พร้อมด้วยชาวบ้าน กำลังตรวจดูแปลงนาข้าวหอมมะลิพันธุ์เตี้ย ซึ่งถูกคนร้ายใช้เคียวเกี่ยวเอาแต่เฉพาะรวงข้าวไป เหลือทิ้งไว้แต่ต้นข้าว รวมเนื้อ 8 ไร่ เป็นข้าวเปลือกร่วม 7 เกวียน ค่าเสียหายประมาณ 56,000 บาท

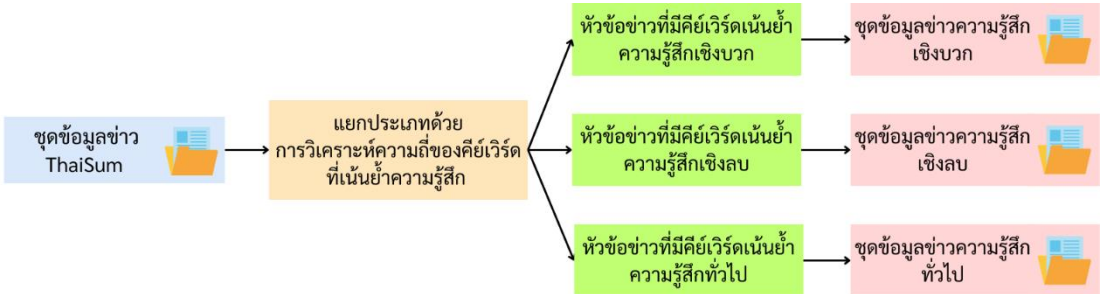
ตารางที่ 3.10 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเชิงลบ (ต่อ)

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเชิงลบ
คอลัมน์ Body	สำนักข่าวต่างประเทศรายงานว่า ผู้โดยสารหลายพันคน ส่วนใหญ่เป็นชาวอังกฤษต้องติดค้างอยู่ที่ท่าอากาศยานและสนามบินในฝรั่งเศสเป็นวันที่ 2 เมื่อ 8 เม.ย. เนื่องจากมีเที่ยวบินหลายร้อยเที่ยวจากฝรั่งเศส ต้องถูกยกเลิก อันเป็นผลมาจากเจ้าหน้าที่หอควบคุมการบินในฝรั่งเศส สังกัดสหภาพแรงงาน SNCTA ได้รวมตัวกัน สไตรก์ หยุดงานเป็นเวลา 2 วัน

ตารางที่ 3.11 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเป็นกลาง

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเป็นกลาง
คอลัมน์ Title	ออกหมายจับอีก 1 ราย คดียูฟัน จ่อสรุปสำนวนต้นเดือนหน้า
	รู้ก่อนซื้อ 2 สิ่งสำคัญ เช็ก รถหรู ถูกกฎหมาย ดารานิยมจัดไฟแนนซ์
คอลัมน์ Body	เมื่อช่วงบ่ายที่ผ่านมา พลตำรวจเอกสมยศ พุ่มพันธุ์ม่วง ผู้บัญชาการตำรวจแห่งชาติ ได้เดินทางมาที่ กองบังคับการปราบปรามการกระทำความผิดเกี่ยวกับการคุ้มครองผู้บริโภค หรือ ปคบ. เพื่อติดตามความคืบหน้า คดีฉ้อโกงประชาชนจากธุรกิจยูฟัน ซึ่งในวันนี้ยังคงมีผู้เสียหาย...
	รถหรู นำเข้ากลายเป็นประเด็นร้อนทันที หลัง หนิง ปณิตา ธรรมวัฒนะ ดารานักแสดงหญิงชื่อดัง ตกเป็นข่าวรถหรูราคาแพง มาเซราติ ที่กำลังอยู่ระหว่างการตรวจสอบ กรณีรถไม่ตรงกับข้อมูลป้ายทะเบียนรถ จากกรณีดังกล่าว วันนี้,ทีมข่าวเฉพาะกิจไทยรัฐออนไลน์...

2) ชุดข้อมูลแบบที่ 2 แยกประเภทด้วยการวิเคราะห์ความถี่ของคีย์เวิร์ดที่เน้นย้ำความรู้สึก



รูปที่ 3.5 แผนภาพแสดงการแยกประเภทชุดข้อมูลแบบที่ 2

2) ชุดข้อมูลแบบที่ 2 แยกประเภทด้วยการวิเคราะห์ความถี่ของคีย์เวิร์ดที่เน้นย้ำความรู้สึก (ต่อ)

คณะผู้จัดทำได้แยกประเภทหัวข้อข่าวจากชุดข้อมูล ThaiSum โดยใช้การวิเคราะห์ความถี่ของคีย์เวิร์ดเน้นย้ำความรู้สึกที่ปรากฏในหัวข้อข่าว ซึ่งจะใช้ไลบรารี PyThaiNLP ในการตัดคำแต่ละคำและนับความถี่โดยใช้ไลบรารี Wordcount โดยแบ่งประเภทหัวข้อข่าวตามคีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากออกเป็น 3 ประเภท ได้แก่

1) หัวข้อข่าวที่มีคีย์เวิร์ดเน้นย้ำความรู้สึกเชิงบวก
2) หัวข้อข่าวที่มีคีย์เวิร์ดเน้นย้ำความรู้สึกเชิงลบ
3) หัวข้อข่าวที่มีคีย์เวิร์ดเน้นย้ำความรู้สึกทั่วไป

ตารางที่ 3.12 คีย์เวิร์ดที่เน้นย้ำความรู้สึกเชิงบวกที่มีความถี่มาก

"ดีใจ","ยินดี","ข่าวดี","ตื่นเต้น","ประทับใจ","ความสุข","โล่งอก","สบายใจ","อบอุ่น","ฉลอง","ภูมิใจ","ยกย่อง","ชื่นชม","สนับสนุน","ส่งเสริม","ปลื้ม","สุขใจ","ชื่นใจ","ซาบซึ้ง","น่ารัก","ทึ่ง","ฮือฮา"

ตารางที่ 3.13 คีย์เวิร์ดที่เน้นย้ำความรู้สึกเชิงลบที่มีความถี่มาก

"สลด","โศกนาฏกรรม","เศร้า","เสียใจ","ประณาม","ช็อก","สะเทือนใจ","ใจสลาย","เสียว","กลัว","โกรธ","โมโห","แค้น","อาฆาต","สิ้นหวัง","เครียด","กังวล","ทุกข์ใจ","สงสาร","วุ่นวาย","โกลาหล","รุนแรง","น่ากลัว","เกลียด","อาลัย","เวทนา","รังเกียจ","ผวา","สยอง","ระทึก"

ตารางที่ 3.14 คีย์เวิร์ดที่เน้นย้ำความรู้สึกทั่วไปที่มีความถี่มาก

"ชี้","ชัด","ฟาด","ลั่น","แฉ","เตือน","ด่วน","เผย","คาด","แจง","อ้าง","ย้ำ","เปิดใจ","ยืนยัน","แถลง"

คณะผู้จัดทำดึงหัวข้อข่าวที่ประกอบด้วยคีย์เวิร์ดที่เน้นย้ำความรู้สึกเชิงบวก คีย์เวิร์ดที่เน้นย้ำความรู้สึกเชิงลบและคีย์เวิร์ดที่เน้นย้ำความรู้สึกทั่วไปที่มีความถี่มากออกมาคีย์เวิร์ดละ 100 ตัวอย่าง แล้วแยกชุดข้อมูลได้เป็น 3 ชุดข้อมูล คือ

1) ชุดข้อมูลข่าวความรู้สึกเชิงบวก จำนวน 1,994 ตัวอย่าง
2) ชุดข้อมูลข่าวความรู้สึกเชิงลบ จำนวน 2,546 ตัวอย่าง
3) ชุดข้อมูลข่าวความรู้สึกทั่วไป จำนวน 1,438 ตัวอย่าง

ตารางที่ 3.15 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเชิงบวก

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเชิงบวก
คอลัมน์ Title	สุดดีใจ เจ้าของปล่อยโฮ หลังพบบิ๊กบู สุนัขตกรถ หายไป 7 เดือน
	ข่าวดี ผู้ติดเชื้อโคโรนาอาการดีขึ้นมากกว่าผู้เสียชีวิต
คอลัมน์ Body	กรณี นายธงชัย ใจสบาย ผู้ใหญ่บ้านหมู่ 10 ต.ควนทอง อ.ขนอม จ. นครศรีธรรมราช ได้แจ้งสื่อมวลชนว่ามีสุนัขตัวหนึ่งตกจากรถระหว่างทางริมถนน 401 ไม่ได้รับบาดเจ็บใดๆ ที่บริเวณเขาหัวช้าง ซึ่งอาศัยอยู่ริมทางกว่า 7 เดือน แล้ว...
	วันนี้ (3 ก.พ.2563) ไทยพีบีเอสตรวจสอบข้อมูลจาก ซึ่งเป็นเว็บไซต์รวบรวมข้อมูลสถิติผู้ติดเชื้อไวรัสโคโรนาสายพันธุ์ใหม่ 2019 จาก WHO ECDC NHC และ DXY พบสถิติผู้ติดเชื้อเพิ่มขึ้นต่อเนื่อง โดยล่าสุด วันนี้มีรายงานผู้ติดเชื้อไวรัสโคโรนาแล้ว 17318 คน เสียชีวิตสะสม 362 คน และอาการดีขึ้นสะสม 487 คน...

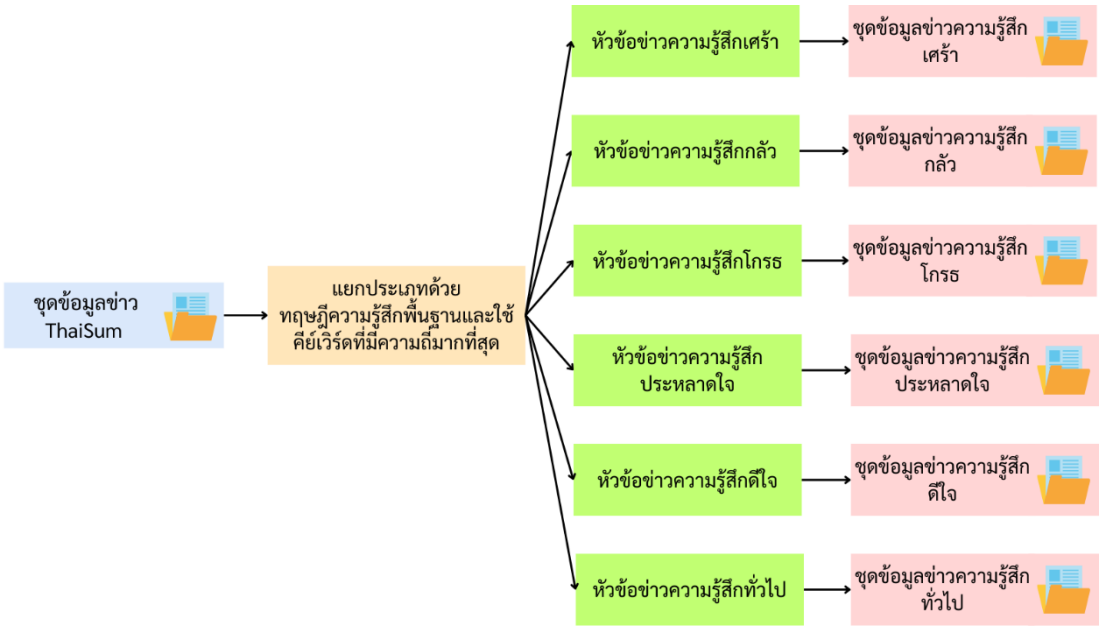
ตารางที่ 3.16 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเชิงลบ

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเชิงลบ
คอลัมน์ Title	โศกนาฏกรรม ซ้ำรอย ช้างป่าเขาใหญ่ ตกเหวนรก ยืนยันตาย 11 ตัว
	โกรธมาก น้าเทิด โวยแหลกหลัง คุนญา โดนต่อย เกมบุกพ่ายราชบุรี
คอลัมน์ Body	วันที่ 8 ต.ค. หลังเมื่อวานนี้ ชุดบินโดรนจากสมาคมตอบโต้ภัยพิบัติแห่งประเทศไทยบินสำรวจพื้นที่อีกครั้ง พบซากช้างป่าเพิ่มจากเดิมอีก 5 ตัว รวมเป็น 11 ตัว และคืนที่ผ่านมาเจ้าหน้าที่อุทยานแห่งชาติเขาใหญ่เดินสำรวจตามแนวลำธารน้ำตกเหวนรกบริเวณชั้น 2 พบซากช้างป่า 5 ตัว...
	วันที่ 8 ก.ค. เทิดศักดิ์ ใจมั่น หัวหน้าผู้ฝึกสอนของทีม ฉลามชล ชลบุรี เอฟซี ออกมาเปิดใจเกี่ยวกับความวุ่นวายช่วงจบครึ่งแรก ของเกมโตโยต้า ไทยลีก นัดที่ราชบุรี มิตรผล เอฟซี เปิดบ้านเอาชนะ ชลบุรี เอฟซี ไป 5-1 เมื่อวันเสาร์ที่ผ่านมา...

ตารางที่ 3.17 ตัวอย่างชุดข้อมูลข่าวความรู้สึกทั่วไป

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกทั่วไป
คอลัมน์ Title	เพื่อไทยซัดพลังประชารัฐเบี้ยวแก้รัฐธรรมนูญจ่อตรวจสอบนโยบายแถลงแล้วไม่ทำ
	นิคม ลั่นพร้อมแจงสู้ คดีถอดถอน เตือน สนช.เลือกแรงกดดันหรือ ก.ม.
คอลัมน์ Body	วันที่ 26 ส.ค. 2562 นายอนุสรณ์ เอี่ยมสะอาด โฆษกพรรคเพื่อไทย กล่าวถึงกรณีที่ นายสนธิรัตน์ สนธิจิรวงศ์ รัฐมนตรีว่าการกระทรวงพลังงาน และเลขาธิการพรรคพลังประชารัฐ ระบุยังไม่มีแนวคิดในการแก้ไขรัฐธรรมนูญ ว่าการที่ฝ่ายค้านเสนอญัตติตั้งคณะกรรมาธิการวิสามัญศึกษาแก้รัฐธรรมนูญ...
	วันที่ 2 พ.ย. นายนิคม ไวยรัชพานิช อดีตประธานวุฒิสภา กล่าวถึงกรณี สนช.เตรียมพิจารณาการถอดถอนนายสมศักดิ์ เกียรติสุรนนท์ อดีตประธานรัฐสภา และนายนิคม ไวยรัชพานิช กรณีการแก้ไขรัฐธรรมนูญเรื่องที่มา ส.ว.ไม่ชอบ ในวันที่ 6 พ.ย. ว่า ขึ้นอยู่กับ สนช.จะมีมติอย่างไร แต่ตนเตรียมความพร้อมที่จะชี้แจง...

3) ชุดข้อมูลแบบที่ 3 แยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐานและใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด



รูปที่ 3.6 แผนภาพแสดงการแยกประเภทชุดข้อมูลแบบที่ 3

3) ชุดข้อมูลแบบที่ 3 แยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐานและใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด (ต่อ)

คณะผู้จัดทำแยกประเภทของหัวข้อข่าวเป็น 6 ประเภท ดังนี้

1) ข่าวความรู้สึกเศร้า
2) ข่าวความรู้สึกกลัว
3) ข่าวความรู้สึกโกรธ
4) ข่าวความรู้สึกประหลาดใจ
5) ข่าวความรู้สึกดีใจ
6) ข่าวความรู้สึกทั่วไป

โดย 5 ประเภทแรกอิงมาจากทฤษฎีความรู้สึกพื้นฐาน (Basic emotions) ของ Paul Ekman ได้แก่ ความรู้สึกเศร้า ความรู้สึกกลัว ความรู้สึกโกรธ ความรู้สึกประหลาดใจ ความรู้สึกรังเกียจ และความรู้สึกดีใจ

อย่างไรก็ตาม คณะผู้จัดทำได้ตัดความรู้สึกรังเกียจออกจากประเภทข่าว เนื่องจากไม่พบคีย์เวิร์ดในหัวข้อข่าวที่เน้นย้ำถึงความรู้สึกรังเกียจ

นอกจากนี้ ยังได้เพิ่มประเภทข่าวความรู้สึกทั่วไปเพื่อจัดเก็บหัวข้อข่าวที่มีคีย์เวิร์ดเน้นย้ำความรู้สึกแต่ไม่สามารถจัดอยู่ในอีก 5 ประเภทได้

ตารางที่ 3.18 ตัวอย่างหัวข้อข่าวที่มีคีย์เวิร์ดเน้นย้ำความรู้สึกแต่ไม่สามารถจัดอยู่ในอีก 5 ประเภทได้

ตัวอย่างหัวข้อข่าว	คีย์เวิร์ดที่เน้นย้ำความรู้สึก
ซูเปอร์โพลชี้คน 70.5% อยากให้ ยุบสภา 17.8% อยากให้ปรับ ครม.	ชี้
อนาคตใหม่ เผยตัวแทนทูต-นักวิชาการเตรียมสังเกตการณ์ ปิยบุตร พบตำรวจ 17 เม.ย. นี้	เผย

หลังจากแยกประเภทของหัวข้อข่าวออกเป็น 6 ประเภท คณะผู้จัดทำได้นับความถี่ของคีย์เวิร์ดที่ปรากฏในชุดข้อมูลแบบที่ 2 โดยใช้ไลบรารี Wordcount จากนั้นเลือกเฉพาะ 2-3 คีย์เวิร์ดที่มีความถี่มากที่สุดในแต่ละประเภทมาใช้เป็นตัวแทนในการดึงตัวอย่างหัวข้อข่าวที่มีการใช้คีย์เวิร์ด

3) ชุดข้อมูลแบบที่ 3 แยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐานและใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด (ต่อ)

เหล่านั้น เหตุผลที่เลือกเฉพาะ 2-3 คีย์เวิร์ดคือเพื่อให้โมเดลเรียนรู้ไปในทิศทางเดียวกัน และเพื่อหลีกเลี่ยงตัวแปรที่ไม่จำเป็น (Noise) ซึ่งอาจเกิดจากการใช้คีย์เวิร์ดที่มีความถี่น้อย

ตารางที่ 3.19 คีย์เวิร์ดที่มีความถี่มากที่สุดในแต่ละความรู้สึก

ประเภทความรู้สึกของข่าว	คีย์เวิร์ดที่มีความถี่มากที่สุด
ข่าวความรู้สึกเศร้า	สลด: 1,939 คำ เศร้า: 1,122 คำ อาลัย: 491 คำ
ข่าวความรู้สึกกลัว	สยอง: 1,296 คำ ผวา: 964 คำ
ข่าวความรู้สึกโกรธ	แค้น: 818 คำ โมโห: 286 คำ
ข่าวความรู้สึกประหลาดใจ	ช็อก: 1,598 คำ ฮือฮา: 570 คำ ทึ่ง: 177 คำ
ข่าวความรู้สึกดีใจ	ดีใจ: 455 คำ ยินดี: 415 คำ ข่าวดี: 316 คำ
ข่าวความรู้สึกทั่วไป	ชี้: 11,353 คำ เผย: 10,593 คำ คาด: 6,132 คำ

คณะผู้จัดทำได้คัดเลือกตัวอย่างที่เหมาะสม โดยเลือกเฉพาะหัวข้อข่าวที่มีคีย์เวิร์ดเน้นย้ำความรู้สึกที่อยู่บริเวณส่วนด้านหน้าของหัวข้อ (คำที่ 1-3) เนื่องจากสื่อถึงความรู้สึกได้มากกว่าคีย์เวิร์ดที่อยู่ส่วนท้ายของประโยค ทำให้ชุดข้อมูลแบบที่ 3 มีจำนวนตัวอย่างดังนี้

1) ข่าวความรู้สึกเศร้า 1,699 ตัวอย่าง
2) ข่าวความรู้สึกกลัว 519 ตัวอย่าง
3) ข่าวความรู้สึกโกรธ 1,079 ตัวอย่าง

3) ชุดข้อมูลแบบที่ 3 แยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐานและใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด (ต่อ)

4) ข่าวความรู้สึกประหลาดใจ 1,504 ตัวอย่าง

5) ข่าวความรู้สึกดีใจ 958 ตัวอย่าง

6) ข่าวความรู้สึกทั่วไป 2,057 ตัวอย่าง

โดยคณะผู้จัดทำได้เลือกจำนวนตัวอย่างที่ใกล้เคียงกันในแต่ละประเภทความรู้สึกเพื่อป้องกันปัญหาความไม่สมดุลของชุดข้อมูล (Imbalance) ซึ่งอาจส่งผลกระทบต่อประสิทธิภาพและความน่าเชื่อถือของโมเดลที่พัฒนาขึ้น

ตารางที่ 3.20 ตัวอย่างชุดข้อมูลข่าวความรู้สึกเศร้า

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกเศร้า
คอลัมน์ Title	สลด คนเก็บขวดพบศพทารกอายุครรภ์ 8 เดือน ถูกทิ้งถังขยะที่สมุทรปราการ
	เศร้า ตำนานนักกอล์ฟผิวสีคนแรก ลาโลกด้วยวัย 92 ปี
คอลัมน์ Body	คนเก็บขวดตะลึง พบศพทารกเพศหญิงอายุครรภ์ 8 เดือน ยังมีสายรกติดอยู่ ห่อด้วยผ้าขนหนูและถุงพลาสติกมิดชิด ถูกทิ้งไว้ในถังขยะที่สมุทรปราการ ตร.เร่งสอบสวนหาพ่อแม่ใจร้ายมาดำเนินคดี...
	สำนักข่าวต่างประเทศ รายงานวันที่ 6 ก.พ. ว่า ชาร์ลี ซิฟฟอร์ด นักกอล์ฟผิวสีชาวอเมริกันคนแรก ที่ได้เล่นในพีจีเอ ทัวร์ เมื่อปี 1961 พร้อมกับคว้าแชมป์ได้ 2 รายการ สิ้นใจในวัย 92 ปีด้วยโรคชรา...

ตารางที่ 3.21 ตัวอย่างชุดข้อมูลข่าวความรู้สึกกลัว

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกกลัว
คอลัมน์ Title	ผวาเสียงเดิน หมุนลูกบิดบ้านพักครู นักศึกษาโทรเรียกคนช่วย เผยเรื่องเล่าสุดขนลุก
	สยองแก่งคอย ฆ่าหั่นสาว ยัดชิ้นส่วน5ถุง
คอลัมน์ Body	เมื่อเวลา 05.00 น. วันที่ 12 ม.ค.63 ตำรวจ สภ.เกาะคา อ.เกาะคา จ.ลำปาง รับแจ้งจากเจ้าหน้าที่ของโรงเรียนแห่งหนึ่ง ใน อ.เกาะคา จ.ลำปาง ว่า มีนักศึกษาสถาบันราชภัฏ ชั้นปี 4 ที่มาฝึกงานเป็นครูภายในโรงเรียน ...
	ทิ้งใต้สะพานลอยริมถนนมิตรภาพสภาพศพแห้งกรัง สยองสาวนิรนามถูกฆ่าหั่นศพแยกชิ้นส่วน ศีรษะ แขน ขา ลำตัว และข้อเท้า ยัดถุงปุ๋ยและถุงดำ 5 ถุง ทิ้งในป่ารกใต้สะพานต่างระดับแก่งคอย-บ้านนา คาดตายมาราว 1 เดือน ...

ตารางที่ 3.22 ตัวอย่างชุดข้อมูลข่าวความรู้สึกโกรธ

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกโกรธ
คอลัมน์ Title	แค้นขโมยเสื้อชุดโปรดโยนทิ้ง หนุ่มใหญ่ฆ่าทุบหัวเพื่อนร่วมงาน
	หัวโจกรุมกระทืบ รปภ.มอบตัว อ้างโมโหถูกต่อว่าหยาบคาย ขอโทษทำรุนแรง
คอลัมน์ Body	เมื่อเวลา 16.00 น. วันที่ 21 เมษายน 2559 ที่ห้องสืบสวน สภ.คลองหลวง ต.คลองสอง อ.คลองหลวง จ.ปทุมธานี พ.ต.อ.สมหมาย ประสิทธิ์ รอง ผบก.ภ.จว.ปทุมธานี พ.ต.ท.กานตพล วรรณา รอง ผกก.สส.สภ.คลองหลวง พ.ต.ท.จิรวัฒน์ เปี่ยมปิ่นเศรษฐ สว.สส.สภ.คลองหลวง กำลังชุดสืบสวนร่วมกันจับกุมตัว...
	เมื่อเวลา 14.30 น. วันที่ 14 มี.ค.60 ที่ สภ.รัตนาธิเบศร์ จ.นนทบุรี นายวิษณุ หรือต่อ ทับผึ้ง อายุ 28 ปี ชาว ต.ท่าทราย อ.เมืองนนทบุรี ผู้ต้องหาตามหมายจับศาลจังหวัดนนทบุรี เลขที่ 60/2560 ลงวันที่ 14 มี.ค.60 คดี 3 วัยรุ่นรุมทำร้าย รปภ.โครงการบ้านเอื้ออาทรประชานิเวศน์ 3 จนได้รับบาดเจ็บ...

ตารางที่ 3.23 ตัวอย่างชุดข้อมูลข่าวความรู้สึกประหลาดใจ

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกประหลาดใจ
คอลัมน์ Title	ช็อก ระเบิดโบสถ์คริสต์บน เกาะโจโล ฟิลิปปินส์
	ระทึก เมียนมาเตือน 31นักโทษแหกคุกที่รัฐกะเหรี่ยง อาจหลบหนีเข้าไทย
คอลัมน์ Body	วันนี้ (27 ม.ค.2562) กองทัพฟิลิปปินส์ เผยแพร่ภาพความเสียหายภายในโบสถ์คริสต์นิกายโรมันคาทอลิกแห่งหนึ่ง บนเกาะโจโล ทางภาคใต้ของฟิลิปปินส์หลังจากเกิดเหตุระเบิด 2 ครั้งซ้อนภายในโบสถ์และลานจอดรถ เมื่อเวลาประมาณ 08.45 น.ตามเวลาท้องถิ่น...
	เมื่อ 17 ก.ย.61 เว็บไซต์ Channelnewsasia รายงาน เจ้าหน้าที่เมียนมาออกคำเตือน นักโทษแหกคุกที่เรือนจำในเมืองพะอัน รัฐกะเหรี่ยงที่ยังหลบหนีไปได้อีกกว่า 30 คน อาจหนีเข้ามาในประเทศไทย ขณะเจ้าหน้าที่พยายามติดตามจับกุมนักโทษเหล่านี้ หลังนักโทษ 41 คน ร่วมมือกันก่อเหตุ...

ตารางที่ 3.24 ตัวอย่างชุดข้อมูลข่าวความรู้สึกดีใจ

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวทั่วไป
คอลัมน์ Title	ข่าวดี มข.จับมือ รพ.ลาว ผ่าตัดรักษา มะเร็งท่อน้ำดี ครั้งแรก
	ชื่นชมพ่อค้าหวยเก็บ 90 ล้านให้ลูกค้า ชาวเน็ตบอกงวดหน้าจะไปอุดหนุน
คอลัมน์ Body	วันนี้ (31 พ.ค.2562) ผู้สื่อข่าวรายงานว่า เมื่อวันที่ 28-30 พ.ค.ที่ผ่านมา มหาวิทยาลัยขอนแก่น ร่วมกับศูนย์ความเป็นเลิศมะเร็งท่อน้ำดี โรงพยาบาลศรีนครินทร์ พร้อมด้วยทีมงานกว่า 50 คน ทั้งทีมศัลยแพทย์ ทีมวิสัญญีแพทย์ ทีมพยาบาล ICU และทีมงานสนับสนุน ดำเนินกิจกรรมผ่าตัดรักษาและดูแลผู้ป่วยมะเร็งท่อน้ำดีหลังการผ่าตัด...
	จากกรณีพ่อค้าขายลอตเตอรี่แผงมหาเฮง เก็บสลากไว้ให้ลูกค้าที่ถูกรางวัล 90 ล้านบาท ตามที่ได้เสนอข่าวไปก่อหน้านี้ (ถูกจริง 90 ล้าน พ่อค้าสลากสมุทรสาคร ถือสัจจะไม่เผยข้อมูลเศรษฐีคนใหม่)ทั้งนี้แผงขายลอตเตอรี่ของพ่อค้าคนนี้ตั้งอยู่ในปั๊มน้ำมัน ปตท.สาขาเอกชัย ต.โคกขาม อ.เมือง จ.สมุทรสาคร...

ตารางที่ 3.25 ตัวอย่างชุดข้อมูลข่าวความรู้สึกทั่วไป

คอลัมน์	ตัวอย่างข้อมูลของชุดข้อมูลข่าวความรู้สึกทั่วไป
คอลัมน์ Title	ซูเปอร์โพลชี้คน 70.5% อยากให้ ยุบสภา 17.8% อยากให้ปรับ ครม.
	อนาคตใหม่ เผยตัวแทนทูต-นักวิชาการเตรียมสังเกตการณ์ ปิยบุตร พบตำรวจ 17 เม.ย. นี้
คอลัมน์ Body	ทำงานต่อไปคือคำตอบ13 มิ.ย. 2563 ดร.นพดล กรรณิกา ผอ.สำนักวิจัยซูเปอร์โพล (SUPER POLL) เปิดเผยผลสำรวจภาคสนาม เรื่อง คนดีการเมือง กรณีศึกษาตัวอย่างประชาชนทุกสาขาอาชีพทั่วประเทศ โดยดำเนินการเก็บข้อมูลแบบผสมผสาน ทั้งการสัมภาษณ์ทางโทรศัพท์ การลงพื้นที่และการเก็บข้อมูลในโลกโซเชียล...
	รวมทั้งจะมีกลุ่มนักวิชาการและลูกศิษย์เดินทางไปให้กำลังใจด้วย ด้าน รังสิมันต์ โรม โพสต์ให้กำลังใจอยากเห็นสังคมไทยเปลี่ยนแปลง14 เม.ย. 2562 รายงานว่า น.ส.พรรณิการ์ วานิช โฆษกพรรคอนาคตใหม่ กล่าวถึงกระแสข่าวว่านายปิยบุตร แสงกนกกุล เลขาธิการพรรคอนาคตใหม่...

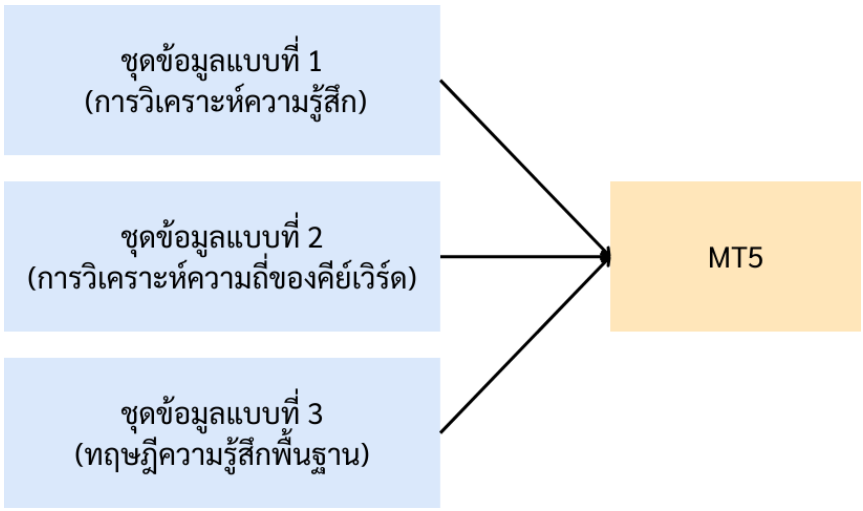
3.3 การฝึกสอนโมเดล

คณะผู้จัดทำได้แบ่งขั้นตอนการฝึกสอนโมเดลออกเป็น 2 ขั้นตอน โดยได้ผ่านการเข้ารหัสด้วย SentencePiece แล้ว ดังนี้

3.3.1 การทดสอบขั้นตอนที่ 1 : การทดสอบประสิทธิภาพของชุดข้อมูล 3 รูปแบบ

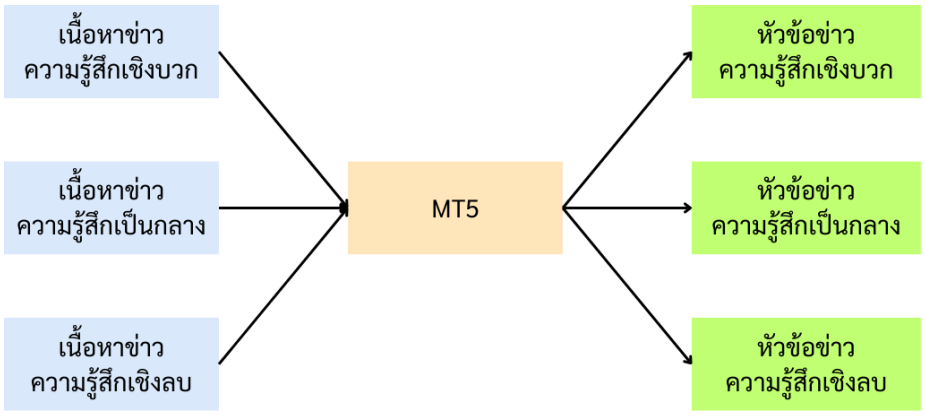
ในขั้นตอนนี้ คณะผู้จัดทำได้ดำเนินการทดสอบเพื่อหารูปแบบของชุดข้อมูลที่เมื่อนำมาฝึกสอนกับโมเดลแล้วจะให้ประสิทธิภาพที่ดีที่สุดโดยแบ่งการทดลองออกเป็น 3 รูปแบบ ดังนี้

1) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 1 โดยใช้วิธีการปรับแต่งโมเดล
2) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 2 โดยใช้วิธีการปรับแต่งโมเดล
3) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 3 โดยใช้วิธีการปรับแต่งโมเดล



รูปที่ 3.7 แผนภาพการทดสอบขั้นตอนที่ 1

1) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 1 โดยใช้วิธีการปรับแต่งโมเดล



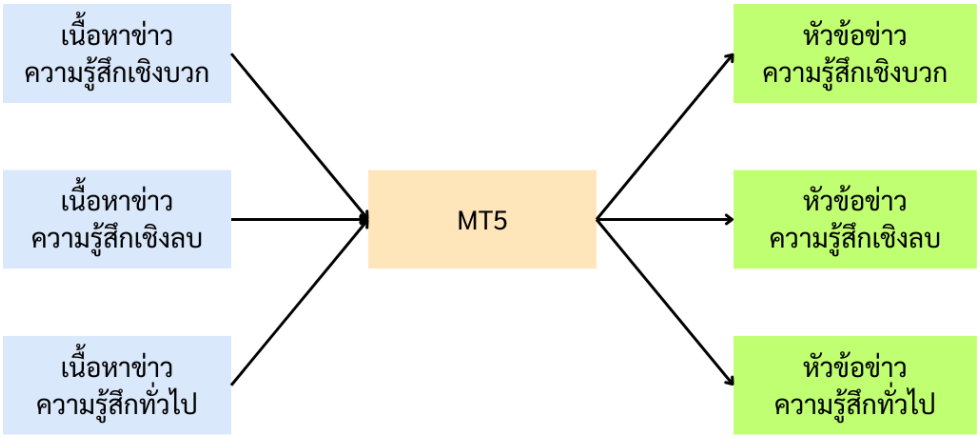
รูปที่ 3.8 แผนภาพการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลรูปแบบที่ 1

1) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 1 โดยใช้วิธีการปรับแต่งโมเดล (ต่อ)

การทดลองแผนภาพการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลรูปแบบที่ 1 ได้มีการแบ่งชุดข้อมูลออกเป็น

1) ชุดฝึกสอน (Training set) 80% สำหรับฝึกสอนโมเดลซึ่งเท่ากับ 5,429 ตัวอย่าง
2) ชุดทดสอบ (Testing set) 10% สำหรับทดสอบโมเดลซึ่งเท่ากับ 678 ตัวอย่าง
3) ชุดประเมิน (Evaluation set) 10% สำหรับวัดประสิทธิภาพโมเดลระหว่างการฝึกสอนซึ่งเท่ากับ 678 ตัวอย่าง

2) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 2 โดยใช้วิธีการปรับแต่งโมเดล

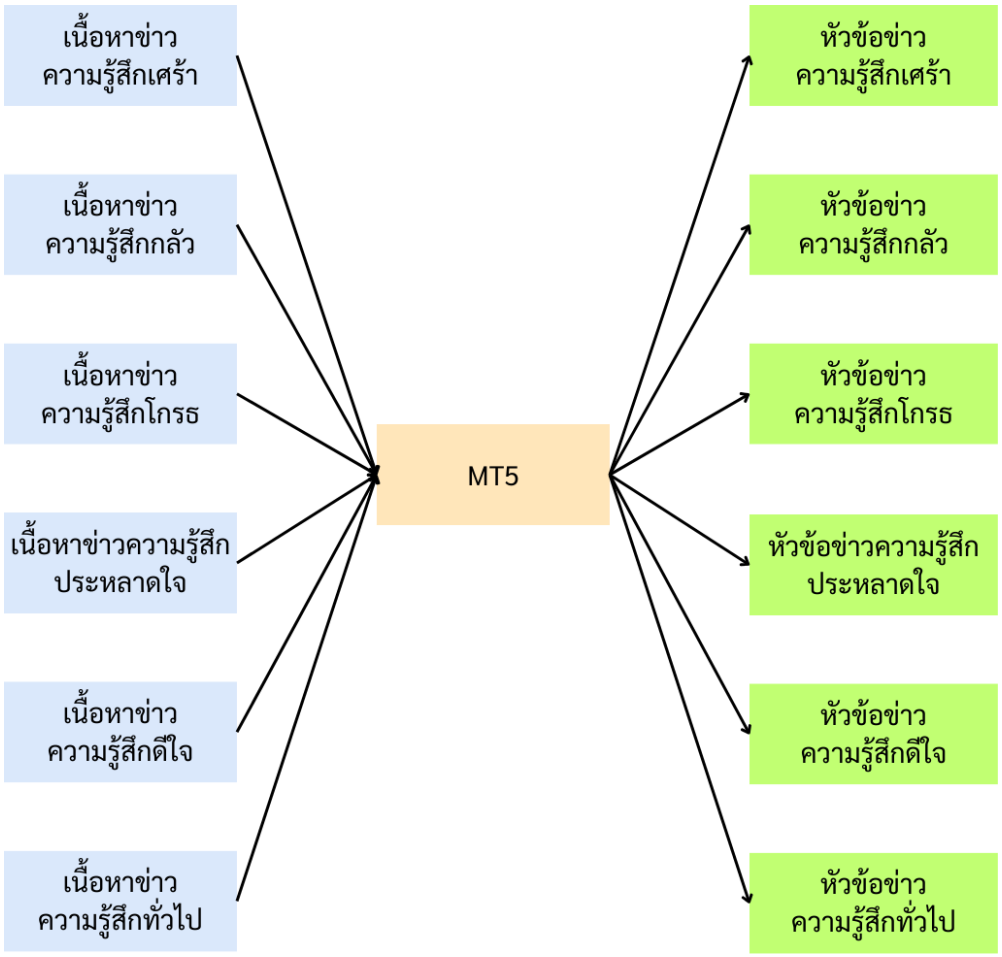


รูปที่ 3.9 แผนภาพการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลรูปแบบที่ 2

การทดลองแผนภาพการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลรูปแบบที่ 2 ได้มีการแบ่งชุดข้อมูลออกเป็น

1) ชุดฝึกสอน (Training set) 80% สำหรับฝึกสอนโมเดลซึ่งเท่ากับ 4,784 ตัวอย่าง
2) ชุดทดสอบ (Testing set) 10% สำหรับทดสอบโมเดลซึ่งเท่ากับ 597 ตัวอย่าง
3) ชุดประเมิน (Evaluation set) 10% สำหรับวัดประสิทธิภาพโมเดลระหว่างการฝึกสอนซึ่งเท่ากับ 597 ตัวอย่าง

3) ฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลแบบที่ 2 โดยใช้วิธีการปรับแต่งโมเดล



รูปที่ 3.10 แผนภาพการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลรูปแบบที่ 3

การทดลองแผนภาพการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลรูปแบบที่ 3 ได้มีการแบ่งชุดข้อมูลออกเป็น

1) ชุดฝึกสอน (Training set) 80% สำหรับฝึกสอนโมเดลซึ่งเท่ากับ 6,254 ตัวอย่าง
2) ชุดทดสอบ (Testing set) 10% สำหรับทดสอบโมเดลซึ่งเท่ากับ 781 ตัวอย่าง
3) ชุดประเมิน (Evaluation set) 10% สำหรับวัดประสิทธิภาพโมเดลระหว่างการฝึกสอนซึ่งเท่ากับ 781 ตัวอย่าง

3.3.1 การทดสอบขั้นตอนที่ 1 : การทดสอบประสิทธิภาพของชุดข้อมูล 3 รูปแบบ (ต่อ)

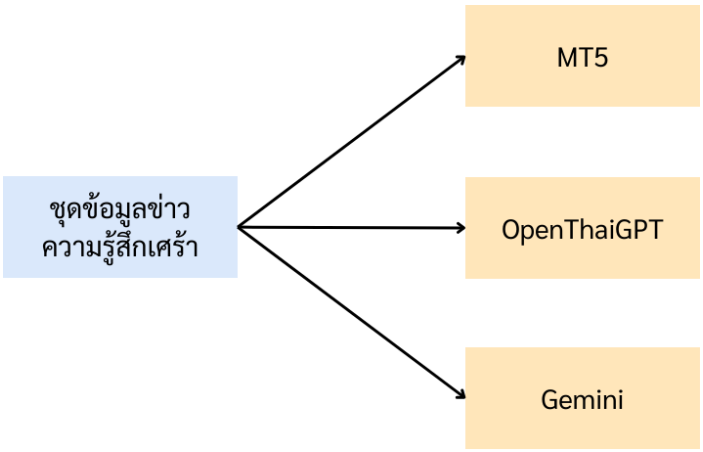
นอกจากนี้การแบ่งการทดลองออกเป็น 3 รูปแบบ มีการตั้งค่าพารามิเตอร์ที่เท่ากันสำหรับการทดลองทุกรูปแบบ ได้แก่

1) จำนวนรอบ (Epoch) เท่ากับ 10 เพราะเป็นจำนวนรอบที่ให้ประสิทธิภาพดีที่สุด
2) อัตราการเรียนรู้ (Learning rate) เท่ากับ 0.00005 เพื่อป้องกันไม่ให้เกิดเหตุการณ์โมเดลเรียนรู้ไม่เพียงพอ (Underfit) หรือโมเดลเรียนรู้มากเกินไป (Overfit)
3) ขนาด batch (Batch size) เท่ากับ 2 เพื่อไม่ให้ใช้ GPU RAM มากเกินไประหว่างการฝึกสอน

3.3.2 การทดสอบขั้นตอนที่ 2 : การทดสอบประสิทธิภาพของโมเดลที่มีโครงสร้างต่างกัน

ในขั้นตอนนี้ คณะผู้จัดทำได้นำชุดข้อมูลรูปแบบที่ 3 เฉพาะชุดข้อมูลข่าวความรู้สึกเศร้าซึ่งเป็นชุดข้อมูลที่ให้ผลการวัดประสิทธิภาพดีที่สุดจากการทดสอบขั้นตอนที่ 1 มาทดสอบกับโมเดลที่มีโครงสร้างแตกต่างกัน 3 รูปแบบ ดังนี้

1) การปรับแต่งโมเดล (Fine tuning) โมเดล MT5 ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า
2) การปรับแต่งคำสั่งพร้อมพ์ (Prompt tuning) โมเดล OpenThaiGPT ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า
3) การออกแบบคำสั่งพร้อมพ์ (Prompt engineering) โมเดล Gemini ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า



รูปที่ 3.11 แผนภาพการทดสอบขั้นตอนที่ 2

3.3.2 การทดสอบขั้นตอนที่ 2 : การทดสอบประสิทธิภาพของโมเดลที่มีโครงสร้างต่างกัน (ต่อ)

1) การปรับแต่งโมเดล MT5 ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า

รูปที่ 3.12 แผนภาพการปรับแต่งโมเดล MT5 ด้วยชุดข้อมูลข่าวความรู้สึกเศร้าพร้อมทั้งฝึกสอนโมเดล

ในการทดลองนี้ คณะผู้จัดทำได้แบ่งชุดข้อมูลสำหรับฝึกสอนโมเดลออกเป็น

1) ชุดฝึกสอน 80% สำหรับฝึกสอนโมเดล ซึ่งเท่ากับ 1359 ตัวอย่าง
2) ชุดทดสอบ 10% สำหรับทดสอบโมเดล ซึ่งเท่ากับ 170 ตัวอย่าง
3) ชุดประเมิน 10% สำหรับวัดประสิทธิภาพโมเดลระหว่างการฝึกสอน ซึ่งเท่ากับ 170 ตัวอย่าง

นอกจากนี้ คณะผู้จัดทำได้ทำการตั้งค่าพารามิเตอร์ที่ใช้ในการฝึกสอนโมเดล ดังนี้

1) จำนวนรอบเท่ากับ 10 เพราะเป็นจำนวนรอบที่ให้ประสิทธิภาพดีที่สุด
2) อัตราการเรียนรู้เท่ากับ 0.00005 เพื่อป้องกันไม่ให้เกิดเหตุการณ์โมเดลเรียนรู้ไม่เพียงพอหรือโมเดลเรียนรู้มากเกินไป
3) ขนาด batch เท่ากับ 2 เพื่อไม่ให้ใช้ GPU RAM มากเกินไประหว่างการฝึกสอน

2) การปรับแต่งคำสั่งพร้อมพ์ (Prompt tuning) OpenThaiGPT ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า

รูปที่ 3.13 แผนภาพการปรับแต่งคำสั่งพร้อมพ์ OpenThaiGPT ด้วยชุดข้อมูลข่าวความรู้สึกเศร้าพร้อมทั้งฝึกสอนโมเดล

2) การปรับแต่งคำสั่งพร้อมพ์ (Prompt tuning) OpenThaiGPT ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า (ต่อ)

ในการทดลองนี้ คณะผู้จัดทำได้มีการเตรียมข้อมูลรับเข้าที่ประกอบด้วย คำสั่งพร้อมพ์ที่เน้นย้ำถึงความรู้สึกเศร้า "จงสร้างหัวข้อข่าวที่เน้นย้ำถึงความรู้สึกเศร้าจากเนื้อหาข่าวต่อไปนี้" และเนื้อหาข่าวที่เน้นย้ำความรู้สึกเศร้า โดยได้แบ่งชุดข้อมูลสำหรับฝึกสอนโมเดลออกเป็น

1) ชุดฝึกสอน 80% สำหรับฝึกสอนโมเดล ซึ่งเท่ากับ 1359 ตัวอย่าง
2) ชุดทดสอบ 10% สำหรับทดสอบโมเดล ซึ่งเท่ากับ 170 ตัวอย่าง
3) ชุดประเมิน 10% สำหรับวัดประสิทธิภาพโมเดลระหว่างการฝึกสอน ซึ่งเท่ากับ 170 ตัวอย่าง

นอกจากนี้ คณะผู้จัดทำได้ทำการตั้งค่าพารามิเตอร์ที่ใช้ในการฝึกสอนโมเดล ดังนี้

1) จำนวนรอบเท่ากับ 10 เพราะเป็นจำนวนรอบที่ให้ประสิทธิภาพดีที่สุด
2) อัตราการเรียนรู้เท่ากับ 0.0001 เพื่อป้องกันไม่ให้เกิดเหตุการณ์โมเดลเรียนรู้ไม่เพียงพอหรือโมเดลเรียนรู้มากเกินไป
3) ขนาด batch เท่ากับ 2 เพื่อไม่ให้ใช้ GPU RAM มากเกินไประหว่างการฝึกสอน

3) การออกแบบคำสั่งพร้อมพ์ (Prompt engineering) โมเดล Gemini ด้วยชุดข้อมูลข่าวความรู้สึกเศร้า

รูปที่ 3.14 แผนภาพการออกแบบคำสั่งพร้อมพ์ Gemini ด้วยชุดข้อมูลข่าวความรู้สึกเศร้าโดยไม่ฝึกสอนโมเดล

ในการทดลองนี้คณะผู้จัดทำได้มีการเตรียมข้อมูลรับเข้าที่ประกอบด้วยคำสั่งพร้อมพ์ที่เน้นย้ำถึงความรู้สึกเศร้า "จงสร้างหัวข้อข่าวที่เน้นย้ำถึงความรู้สึกเศร้าจากเนื้อหาข่าวต่อไปนี้" และเนื้อหาข่าวที่เน้นย้ำความรู้สึกเศร้า โดยไม่ได้มีการฝึกสอนโมเดล Gemini เพิ่มเติม

บทที่ 4
ผลการดำเนินงาน

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาโมเดลประมวลผลภาษาธรรมชาติสำหรับการสังเคราะห์หัวข้อข่าว โดยคณะผู้จัดทำได้ใช้ค่า ROUGE (Recall-oriented understudy for gisting evaluation) และค่า BERTScore (Bidirectional encoder representations from transformers score) เป็นค่าในการวัดประสิทธิภาพของโมเดล โดยแบ่งขั้นตอนการวัดประสิทธิภาพออกเป็น 2 ขั้นตอน ดังนี้

4.1 การวัดประสิทธิภาพขั้นตอนที่ 1 : การวัดประสิทธิภาพของชุดข้อมูล 3 รูปแบบ

1) ผลการวัดประสิทธิภาพชุดข้อมูลแบบที่ 1 ที่ฝึกสอนด้วยโมเดล MT5
2) ผลการวัดประสิทธิภาพชุดข้อมูลแบบที่ 2 ที่ฝึกสอนด้วยโมเดล MT5
3) ผลการวัดประสิทธิภาพชุดข้อมูลแบบที่ 3 ที่ฝึกสอนด้วยโมเดล MT5

ตารางที่ 4.1 ตารางคะแนนเฉลี่ยผลการวัดประสิทธิภาพของชุดข้อมูล 3 รูปแบบ

ชุดข้อมูล	ROUGE		BERTScore
	ROUGE-1 F1	ROUGE-L F1	F1
ชุดข้อมูลแบบที่ 1	25.46	22.68	70.97
ชุดข้อมูลแบบที่ 2	28.96	26.09	73.60
ชุดข้อมูลแบบที่ 3	**34.70**	**31.56**	**73.94**

จากการวิเคราะห์ผลการทดสอบในตารางที่ 4.1 พบว่า ชุดข้อมูลแบบที่ 3 มีคะแนนเฉลี่ยของ BERTScore และ ROUGE สูงที่สุด ซึ่งหมายความว่าชุดข้อมูลแบบที่ 3 สามารถให้ผลลัพธ์การสังเคราะห์หัวข้อข่าวที่มีความสอดคล้องกับข้อความต้นฉบับมากที่สุด เมื่อเทียบกับชุดข้อมูลแบบอื่น ๆ

ตารางที่ 4.2 ตารางผลการวัดประสิทธิภาพชุดข้อมูลแบบที่ 1 ที่ฝึกสอนด้วยโมเดล MT5

ชุดข้อมูลแบบที่ 1	ROUGE			BERTScore		
	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	Precision	Recall	F1
ชุดข้อมูลข่าวความรู้สึกเชิงบวก	**27.33**	10.81	23.25	72.71	68.87	70.58
ชุดข้อมูลข่าวความรู้สึกเชิงลบ	21.72	7.73	19.65	71.67	68.44	69.85
ชุดข้อมูลข่าวความรู้สึกเป็นกลาง	**27.33**	**11.65**	**25.14**	**74.13**	**71.20**	**72.48**

จากการวิเคราะห์ผลการทดสอบในตารางที่ 4.2 พบว่า ชุดข้อมูลความรู้สึกเป็นกลางมีคะแนน BERTScore และ ROUGE สูงที่สุด ชุดข้อมูลความรู้สึกเชิงบวกมีคะแนน BERTScore และ ROUGE อยู่ในระดับรองลงมา และชุดข้อมูลความรู้สึกเชิงลบมีคะแนน BERTScore และ ROUGE ต่ำที่สุด

ตารางที่ 4.3 ตัวอย่างหัวข้อข่าวที่ได้จากการสังเคราะห์หัวข้อข่าวด้วยโมเดล MT5 ที่ฝึกสอนบนชุดข้อมูลแบบที่ 1

ตัวอย่างหัวข้อข่าวความรู้สึกเชิงบวก	ผัดไทย วัดท้องคุ้ง อร่อย ก๋วยเตี๋ยวต้มยำราคา 15 บาท
	ยูริ ปลื้มไทย พาสมาชิกวงเกิร์ลส์เจเนอเรชั่นมาเจอพร้อมหน้า
ตัวอย่างหัวข้อข่าวความรู้สึกเชิงลบ	ร้องเรียน รถตู้ วิ่งออกจาก ม.ศิลปากร น้ำประปากร่อยหมู่บ้าน มีรสเค็ม ใช้อุปโภคบริโภคไม่ได้ บางวันก็ไหลอ่อน เดือดร้อนไม่รู้จะพึ่งใคร
	เกิดเหตุทะเลาะวิวาทกับพนักงานใน จ.นครนายก บาดเจ็บนับสิบ
ตัวอย่างหัวข้อข่าวความรู้สึกเป็นกลาง	มือปืนลั่นไกสังหาร ภรรยากินกันกว่า 10 ปี ขู่ปันใจตีตัวออกห่าง
	ชาวอินเดียอพยพหนีไซโคลน ฆ่าตัวตาย 20 ศพ

ตารางที่ 4.4 ตารางผลการวัดประสิทธิภาพชุดข้อมูลแบบที่ 2 ที่ฝึกสอนด้วยโมเดล MT5

ชุดข้อมูลแบบที่ 2	ROUGE			BERTScore		
	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	Precision	Recall	F1
ชุดข้อมูลข่าวความรู้สึกเชิงบวก	**29.61**	**12.29**	**26.81**	75.10	74.13	74.56
ชุดข้อมูลข่าวความรู้สึกเชิงลบ	28.49	10.43	25.06	**75.46**	**74.16**	**74.76**
ชุดข้อมูลข่าวความรู้สึกทั่วไป	28.78	11.80	26.41	73.44	69.94	71.47

จากการวิเคราะห์ผลการทดสอบในตารางที่ 4.4 พบว่า ชุดข้อมูลข่าวความรู้สึกเชิงลบมีคะแนน BERTScore สูงที่สุด ชุดข้อมูลข่าวความรู้สึกเชิงบวกมีคะแนน ROUGE สูงที่สุด และชุดข้อมูลความรู้สึกทั่วไปมีคะแนนโดยรวมต่ำที่สุด

ตารางที่ 4.5 ตัวอย่างหัวข้อข่าวที่ได้จากการสังเคราะห์หัวข้อข่าวด้วยโมเดล MT5 ที่ฝึกสอนบนชุดข้อมูลแบบที่ 2

ตัวอย่างหัวข้อข่าวความรู้สึกเชิงบวก	น่ารักน่าเอ็นดู พี่น้องฝาแฝด สายฟ้า-พายุ ชมพ่อน็อต
	เบิร์ด ซาบซึ้งมากราบลาในหลวง รัชกาลที่ 9 อย่างใกล้ชิดที่สุดในชีวิต
ตัวอย่างหัวข้อข่าวความรู้สึกเชิงลบ	ลุงเมืองอุดรฯ เครียดเรื่องขับรถชน ผูกคอดับคาโรงเก็บรถ
	สุดสยอง สาวนิรนามยัดปุ๋ยทิ้งใต้สะพาน ตายมา 1 เดือน เหลือหนังหุ้มกระดูก
ตัวอย่างหัวข้อข่าวความรู้สึกทั่วไป	กรมอุตุฯ เตือน 24 ชั่วโมง อากาศร้อนจัด ภาคอีสานฝนฟ้าคะนอง ลมกระโชกแรง ฟ้าผ่าได้
	จรัญ ชี้ หลักนิติธรรมต้องรวม 10 ข้อ ย้ำไม่ลำเอียงปราศจากอคติ

ตารางที่ 4.6 ตารางผลการวัดประสิทธิภาพชุดข้อมูลแบบที่ 3 ที่ฝึกสอนด้วย MT5

ชุดข้อมูลแบบที่ 3	ROUGE			BERTScore		
	ROUGE-1 F1	ROUGE-2 F1	ROUGE-L F1	Precision	Recall	F1
ชุดข้อมูลข่าวความรู้สึกเศร้า	**39.54**	**15.41**	**36.39**	**77.61**	**74.74**	**75.99**
ชุดข้อมูลข่าวความรู้สึกกลัว	33.96	12.44	31.66	73.58	70.22	71.68
ชุดข้อมูลข่าวความรู้สึกโกรธ	31.73	8.86	27.04	75.10	71.09	72.88
ชุดข้อมูลข่าวความรู้สึกประหลาดใจ	38.09	16.62	35.37	76.42	74.39	75.23
ชุดข้อมูลข่าวความรู้สึกดีใจ	34.74	14.20	31.14	75.18	71.76	73.26
ชุดข้อมูลข่าวความรู้สึกทั่วไป	30.11	13.00	27.76	74.89	74.41	74.60

จากการวิเคราะห์ผลการทดสอบในตารางที่ 4.6 พบว่า ชุดข้อมูลข่าวความรู้สึกเศร้า และชุดข้อมูลข่าวความรู้สึกประหลาดใจมีคะแนน BERTScore และ ROUGE สูงที่สุด อันดับรองลงมาคือชุดข้อมูลข่าวความรู้สึกโกรธ ชุดข้อมูลข่าวความรู้สึกกลัว ชุดข้อมูลข่าวความรู้สึกดีใจ และชุดข้อมูลข่าวความรู้สึกทั่วไปซึ่งมีคะแนนใกล้เคียงกัน

ตารางที่ 4.7 ตัวอย่างหัวข้อข่าวที่ได้จากการสังเคราะห์หัวข้อข่าวด้วยโมเดล MT5 ที่ฝึกสอนบนชุดข้อมูลแบบที่ 3

ตัวอย่างหัวข้อข่าว ความรู้สึกเศร้า	สลด วิศวกรหนุ่มอกหักส่งข้อความสุดท้ายถึงหญิงคนรักก่อนไปแขวนคอตาย
	สลด สองเฒ่าอุทัยธานีตกบ่อน้ำดับคาที่2ศพ คาดขาดอากาศหายใจ
ตัวอย่างหัวข้อข่าว ความรู้สึกกลัว	สยอง ศพเฒ่าถูกฟันคอขาดดับปริศนากลางเมืองหนองบัวลำภู
	ผวา อาบู ไซยาฟขู่ฆ่าตัวประกันฟิลิปปินส์3ศพ
ตัวอย่างหัวข้อข่าว ความรู้สึกโกรธ	ตร.เลยจับมือยิงสาวทอมอ้างแค้นไม่คบหาคนใหม่
	ผัวเก่าแค้นอดีตเมียใช้มีดตัดอ้อยฟันคอดับ สารภาพโมโหด่าบุพการี
ตัวอย่างหัวข้อข่าว ความรู้สึกประหลาดใจ	ฮือฮา ไนกี้รองเท้าหายากสุดในโลกMoonShoeเปิดประมูลในแคนาดาแพงสุดเป็นประวัติการณ์13ล้าน
	ช็อก ตร.จีนขุดพบโครงกระดูกอดีตพนักงานบูรณะสนามกีฬา
ตัวอย่างหัวข้อข่าว ความรู้สึกดีใจ	ข่าวดี รพ.ศิริราชทดสอบCOVID-19ได้ผลชัดเจน
	สุดดีใจ คู่รักเทิดทูนณบางรักรับทะเบียนสมรสทองคำ
ตัวอย่างหัวข้อข่าว ความรู้สึกทั่วไป	วันชัยชี้สัญญาณดีปฏิรูปตร.มั่นใจปชช.มีที่พึ่งขอความเป็นธรรมได้
	วราวุธเผยเตรียมเร่งรัดกฎหมายป่าชุมชนปี63-คนช.เห็นชอบ

จากการวัดประสิทธิภาพของชุดข้อมูลทั้ง 3 รูปแบบ พบว่าชุดข้อมูลแบบที่ 3 ให้ผลลัพธ์ที่ดีที่สุด เนื่องจากตัวอย่างข้อมูลในชุดนี้ได้รับการแยกประเภทตามทฤษฎีความรู้สึกอย่างชัดเจน ซึ่งช่วยแก้ปัญหาเรื่องคำศัพท์ที่มีความหมายโดยนัยต่างกัน ซึ่งเป็นจุดอ่อนของชุดข้อมูลแบบที่ 2 ที่มีคำเน้นย้ำความรู้สึกที่หลากหลายแต่มีความหมายโดยนัยแตกต่างกัน เช่น คำว่า "สลด" และ "สยอง" ถึงแม้จะมีความหมายในเชิงลบคล้ายคลึงกัน แต่มีความหมายโดยนัยที่แตกต่างกัน ในขณะที่ชุดข้อมูลแบบที่ 1 นั้นมีข้อจำกัดที่ตัวอย่างข้อมูลที่แยกประเภทด้วยโมเดล WangchanBERTa ผสมกันระหว่างตัวอย่างข่าวที่เน้นความรู้สึก และตัวอย่างข่าวที่เกี่ยวกับรีวิวอาหารหรือสินค้าซึ่งไม่ได้เน้นความรู้สึกมากนัก ดังนั้นชุดข้อมูลแบบที่ 3 จึงเป็นชุดข้อมูลที่มีประสิทธิภาพ และสอดคล้องกับวัตถุประสงค์มากที่สุด

4.2 การวัดประสิทธิภาพขั้นตอนที่ 2 : การวัดประสิทธิภาพของโมเดลที่มีโครงสร้างต่างกัน

ในขั้นตอนนี้ คณะผู้จัดทำได้นำชุดข้อมูลรูปแบบที่ 3 เฉพาะชุดข้อมูลข่าวความรู้สึกเศร้าซึ่งเป็นชุดข้อมูลที่ให้ประสิทธิภาพดีที่สุดจากการทดสอบขั้นตอนที่ 1 มาทดสอบกับโมเดลที่มีโครงสร้างแตกต่างกัน ได้แก่ โมเดล MT5 โมเดล OpenThaiGPT และโมเดล Gemini ซึ่งผลการวัดประสิทธิภาพจะมีดังนี้

ตารางที่ 4.8 ตารางเปรียบเทียบคะแนนการวัดประสิทธิภาพของโมเดลโดยใช้ชุดข้อมูลแบบที่ 3 (ชุดข้อมูลข่าวความรู้สึกเศร้า)

โมเดล	ROUGE		BERTScore
	ROUGE-1 F1	ROUGE-L F1	F1
โมเดล MT5	**39.54**	**36.39**	75.99
โมเดล OpenThaiGPT	22.60	20.27	68.80
โมเดล Gemini	33.67	28.67	**77.09**
โมเดลของคณะผู้จัดทำท่านอื่น	30.35	26.91	73.81

จากการวิเคราะห์ผลการทดสอบในตารางที่ 4.8 พบว่า โมเดล MT5 ให้ผลลัพธ์ที่ดีที่สุดในค่า ROUGE-1 และ ROUGE-L โดยได้คะแนน 39.54 และ 36.39 ตามลำดับ ซึ่งเป็นคะแนนที่สูงที่สุดเมื่อเทียบกับโมเดลอื่นๆ อันดับรองลงมาคือโมเดล Gemini ที่ได้คะแนน ROUGE-1 และ ROUGE-L เท่ากับ 33.67 และ 28.67 ตามลำดับ และอันดับสุดท้ายคือ OpenThaiGPT ซึ่งให้ผลลัพธ์ที่ไม่ดีเท่าโมเดลอื่นๆ อย่างไรก็ตามเมื่อพิจารณาที่ค่า BERTScore โมเดล Gemini ให้ค่า F1 สูงที่สุดที่ 77.09 อันดับรองลงมาคือโมเดล MT5 ได้คะแนน 75.99

จากผลการวัดประสิทธิภาพของโมเดลสามารถสรุปได้ว่า โมเดล MT5 เป็นโมเดลที่ประกอบด้วยทั้งตัวเข้ารหัสและตัวถอดรหัส โดยตัวถอดรหัสจะช่วยกำหนดรูปแบบและโครงสร้างของข้อมูลที่สังเคราะห์ออกมาให้มีความชัดเจน ซึ่งเหมาะสมกับงานสังเคราะห์หัวข้อข่าวที่ต้องการความชัดเจนของรูปแบบ เช่น "ข่าวดี รพ.ศิริราชทดสอบ COVID-19 ได้ผลชัดเจน" ในทางตรงกันข้าม โมเดล OpenThaiGPT เป็นโมเดลที่ประกอบด้วยเพียงตัวเข้ารหัสเท่านั้น โดยไม่มีตัวถอดรหัส ส่งผลให้ข้อมูลที่สังเคราะห์ออกมามีความไม่ตายตัว ซึ่งอาจไม่เหมาะสมสำหรับงานสังเคราะห์หัวข้อข่าวที่ต้อง การความชัดเจน สำหรับโมเดล Gemini แม้จะได้รับการฝึกสอนจากชุดข้อมูลที่มีขนาดใหญ่กว่าโมเดล MT5 ประมาณ 2.6 เท่า และครอบคลุมโดเมนที่หลากหลายมากกว่า เช่น หนังสือ บทความ โค้ด และข้อมูลสนทนา ในขณะที่ MT5 ได้รับการฝึกสอนจากชุดข้อมูลที่มุ่งเน้นเฉพาะบทความและโค้ดเท่านั้น แต่โมเดล Gemini อาจขาดความชัดเจนในรูปแบบเมื่อเทียบกับ MT5

ตารางที่ 4.9 ตัวอย่างหัวข้อข่าวความรู้สึกเศร้าที่ได้จากการสังเคราะห์หัวข้อข่าวด้วยโมเดลที่มีโครงสร้างต่างกัน

โมเดล	ตัวอย่างหัวข้อข่าวความรู้สึกเศร้า
โมเดล MT5	สลด วิศวกรหนุ่มอกหักส่งข้อความสุดท้ายถึงหญิงคนรักก่อนไปแขวนคอตาย
	สลด สองเฒ่าอุทัยธานีตกบ่อน้ำดับคาที่2ศพ คาดขาดอากาศหายใจ
โมเดล OpenThaiGPT	สลด!หนุ่มรับเหมาก่อสร้างนั่งกินร้านดังเมืองปากช่องเมาแล้วขับเก๋งไปจอดหน้าบ้านแล้วดันประตูค้างตายคาที่ลำตัว##
	พบศพเด็กชายวัย3ขวบจมน้ำเสียชีวิตในป่าหญ้าริมทางรถไฟอยุธยา### Instruction:จงสร้างหัวข้อข่าวที่เน้นย้ำถึงความ
โมเดล Gemini	โศก! ชายเมาเดินข้ามถนนพลาด ลื่นล้มถูกรถเมล์ทับดับ
	เศร้า! พบศพ "น้องไทเกอร์" ลอยอืดในท้องนา หลังหายตัวปริศนา 4 วัน
โมเดลของคณะผู้จัดทำท่านอื่น	สลดทารกแรกเกิดแอฟริกาเสียชีวิตจากโควิด-19แล้ว หลังติดเชื้อไวรัสโคโรนา 2019 แล้ว
	สลดอดีตทหารผ่านศึกเชียงรายเครียดดื่มยาฆ่าหญ้าดับคาบ้านอดีตนายทหารจากศึก เชียงราย

บทที่ 5
สรุปผลและข้อเสนอแนะ

5.1สรุปผล
เขียนเป็น paragraph เดียวครับ
คือ เอา abstact มาเขียนใหม่อีกรอบ
โดย parapahse ให้มีเนื้อหาเดียวกัน

5.1 สรุปผลการดำเนินงาน

การพัฒนาโมเดลการประมวลผลภาษาธรรมชาติสำหรับสังเคราะห์หัวข้อข่าวภาษาไทย เริ่มจากการรวบรวมข้อมูลจากชุดข้อมูลข่าวภาษาไทย ThaiSum และทำความสะอาดข้อมูล จากนั้นจึงแยกประเภทชุดข้อมูลสำหรับการฝึกสอนโมเดล โดยแบ่งออกเป็น 3 รูปแบบ ได้แก่

1) ชุดข้อมูลแบบที่ 1 แยกประเภทด้วยการวิเคราะห์ความรู้สึก
2) ชุดข้อมูลแบบที่ 2 แยกประเภทด้วยการวิเคราะห์ความถี่ของคีย์เวิร์ดที่เน้นย้ำความรู้สึก
3) ชุดข้อมูลแบบที่ 3 แยกประเภทด้วยทฤษฎีความรู้สึกพื้นฐาน และใช้คีย์เวิร์ดเน้นย้ำความรู้สึกที่มีความถี่มากที่สุด

ซึ่งจะแบ่งการฝึกสอนโมเดลออกเป็น 2 ขั้นตอน คือ

5.1.1 การทดสอบชุดข้อมูล

คณะผู้จัดทำได้ดำเนินการทดสอบเพื่อหารูปแบบของชุดข้อมูลที่เมื่อนำมาฝึกสอนกับโมเดลแล้วจะให้ประสิทธิภาพที่ดีที่สุดโดยการฝึกสอนโมเดล MT5 ด้วยชุดข้อมูลที่แตกต่างกัน 3 รูปแบบ ผลการวัดประสิทธิภาพพบว่าชุดข้อมูลแบบที่ 3 ให้ประสิทธิภาพดีที่สุด โดยมีค่าเฉลี่ย F1 ของ BERTScore เท่ากับ 73.94 และค่าเฉลี่ย F1 ของ ROUGE-1 เท่ากับ 34.56

5.1.2 การทดสอบโมเดล

คณะผู้จัดทำได้ดำเนินการทดสอบกับโมเดลที่มีโครงสร้างแตกต่างกัน ได้แก่ MT5 OpenThaiGPT และ Gemini เพื่อหาโมเดลที่เหมาะกับงานสังเคราะห์หัวข้อข่าวภาษาไทยมากที่สุด โดยใช้ชุดข้อมูลแบบที่ 3 เฉพาะในส่วนของชุดข้อมูลข่าวความรู้สึกเศร้า ซึ่งเป็นชุดข้อมูลที่ให้ประสิทธิภาพดีที่สุดจากการทดสอบชุดข้อมูล ผลการวัดประสิทธิภาพพบว่าโมเดล MT5 ให้ผลลัพธ์ที่ดีที่สุดในค่า ROUGE-1 โดยได้คะแนน 39.54 ซึ่งเป็นคะแนนที่สูงที่สุดเมื่อเทียบกับโมเดลอื่นๆ อันดับรองลงมาคือโมเดล Gemini ที่ได้ค่า ROUGE-1 เท่ากับ 33.67 อย่างไรก็ตามเมื่อพิจารณาที่ค่า BERTScore โมเดล Gemini ให้ค่า F1 สูงที่สุดที่ 77.09 อันดับรองลงมาคือโมเดล MT5 ให้ค่า F1 เท่ากับ 75.99

5.1.2 การทดสอบโมเดล (ต่อ)

โดยสรุปแล้วในการพัฒนาโมเดลสังเคราะห์หัวข้อข่าวที่สามารถสื่อถึงความรู้สึกได้อย่างชัดเจน การใช้ชุดข้อมูลแบบที่ 3 จะให้ผลลัพธ์ที่ดีที่สุด แต่เนื่องด้วยข้อจำกัดของจำนวนตัวอย่างที่มีอยู่อย่างจำกัด จึงทำให้ประสิทธิภาพของโมเดลอาจจะไม่ได้ดีเท่าที่ควร ซึ่งน่าจะดีกว่านี้ถ้ามีจำนวนตัวอย่างที่มากขึ้น อย่างไรก็ตาม ในกรณีที่มีจำนวนตัวอย่างในการฝึกสอนไม่มากนัก โมเดล MT5 จะมีประสิทธิภาพที่ดีกว่าโมเดล OpenThaiGPT และโมเดล Gemini สำหรับงานนี้ เนื่องจากโมเดล MT5 ที่มีขนาดเล็ก จึงสามารถเรียนรู้พารามิเตอร์ได้ดีกว่าแม้จะมีตัวอย่างไม่มากนัก

5.2 ปัญหาที่พบ

1) สำหรับชุดข้อมูลแบบที่ 1 ซึ่งแยกประเภทโดยใช้โมเดล WangchanBERTa นั้น มีข้อจำกัด เนื่องจากโมเดลดังกล่าวถูกฝึกสอนจากชุดข้อมูลที่ส่วนใหญ่เป็นรีวิวร้านอาหารและสินค้า เช่น Wongnai Reviews และ English-Thai Generated Reviews ทำให้หัวข้อข่าวที่ได้จากการจำแนกด้วย WangchanBERTa ค่อนข้างมีความเอนเอียงไปในเรื่องของอาหารและสินค้า ซึ่งอาจไม่เหมาะสำหรับการพัฒนาโมเดลการประมวลผลภาษาธรรมชาติเพื่อสังเคราะห์หัวข้อข่าวภาษาไทยที่มุ่งเน้นไปที่ความรู้สึก

2) สำหรับชุดข้อมูลแบบที่ 3 ซึ่งแยกประเภทโดยใช้ทฤษฎีความรู้สึกพื้นฐานและคีย์เวิร์ดที่มีความถี่สูงสุดในการเน้นย้ำความรู้สึก พบว่าข้อมูลหัวข้อข่าวที่มีคุณภาพหลังการแยกประเภทตามทฤษฎีความรู้สึกพื้นฐานนั้นมีจำนวนค่อนข้างน้อย ซึ่งอาจส่งผลให้ประสิทธิภาพในการสังเคราะห์หัวข้อข่าวยังไม่เป็นที่น่าพอใจเท่าที่ควร

3) ในการใช้โมเดล OpenThaiGPT เพื่อสังเคราะห์หัวข้อข่าวภาษาไทยที่มุ่งเน้นไปที่ความรู้สึก พบว่ามีข้อจำกัดในการปรับแต่งคำสั่งพร้อมพ์เนื่องจากโมเดลดังกล่าวมีขนาดใหญ่ ในขณะที่ชุดข้อมูลที่นำมาใช้ในการฝึกสอนมีจำนวนจำกัด ซึ่งอาจไม่เพียงพอต่อการเรียนรู้และปรับแต่งโมเดลให้มีประสิทธิภาพในการสังเคราะห์หัวข้อข่าวที่สะท้อนความรู้สึกได้อย่างมีประสิทธิภาพ

5.3 ข้อเสนอแนะ

5.3.1 การจัดสรรเวลาที่เหมาะสม

เนื่องจากโมเดลการสร้างข้อความต้องใช้เวลาในการฝึกสอนค่อนข้างนาน ดังนั้นจึงควรมีการจัดสรรเวลาและทรัพยากรให้เพียงพอต่อการฝึกสอนและพัฒนาโมเดลอย่างรอบคอบ เพื่อให้ได้โมเดลที่มีประสิทธิภาพสูงสุด

5.3.2 การปรับปรุงฐานข้อมูล

หัวข้อข่าวภาษาไทยมีคำเน้นย้ำความรู้สึกที่หลากหลายและอาจเปลี่ยนแปลงไปตามยุคสมัย ดังนั้น ชุดข้อมูล ThaiSum ที่ใช้ในการพัฒนาโมเดลอาจไม่สามารถตอบสนองต่อความต้องการในอนาคตได้อย่างเพียงพอ ดังนั้น อาจต้องมีการปรับปรุงและขยายฐานข้อมูลให้มีความทันสมัยและครอบคลุมมากขึ้น เพื่อให้โมเดลที่พัฒนาขึ้นสามารถใช้งานได้อย่างมีประสิทธิภาพในระยะยาว

อ้างอิงและแหล่งที่มา

[1] Amazon, (n.d.), "ปัญญาประดิษฐ์ หรือ AI", Retrieved 30 April 2023, from https://aws. amazon.com/th/what-is/artificial-intelligence/

[2] machine-learning https://cloud-ace.co.th/blogs/o0v9a6-ai-machine-learning-ml-ai-ml-goog

[3] Data Innovation and Governance Institute, (2022), Retrieved 30 April 2023, from https://digi.data.go.th/blog/how-to-cleansing-data/

[4] Chetna Khanna, (2021), "tokenization", Retrieved 25 January 2023, from https://towardsdatascience com/word-subword-and-character-based-tokenization-know-the-difference-ea0976b64e17

[5] “VISTEC-depa AI Research Institute of Thailand” , Retrieved 25 January 2023, from https://airesearch.in.th/

[6] Phatthiyaphaibun et al., (2023), "PyThaiNLP: Thai Natural Language Processing in Python", in aclanthology Access, pp. 25–36.

[7] "Wordcount", (2022), Retrieved 30 April 2023, from https://support.microsoft.com/th-th/office/

[8] Business & Technology, (2022), "Sentiment analysis"), Retrieved 22 January 2023, from https://aigencorp com/what-is-sentiment-analysis/#:~:text=Sentiment

[9] Urbinner, (2021), "รู้จักอารมณ์พื้นฐานของมนุษย์จาก Emotion Wheel", Retrieved 22 January 2023, from https://www.urbinner.com/post/know-basic-human-emotion-from-emotion-wheel

[10] "การวิเคราะห์ความถี่ (Frequency analysis) " , (2022), Retrieved 30 April 2023, from https://digi.data.go.th/blog/what-is-a-data-frequency/

[11] "text-summarization", https://blog pjjop org/thai-text-summarization-with-bert-and-pagerank

[12] "การสร้างหัวข้อข่าว หรือ Headline generation" , (2022), Retrieved 30 April 2023, from https://medium.com/@yeepoon24/headline-generator-using-bart-model-bart

[13] Amazon, (n.d.), "Transformer model", Retrieved 20 December 2023, from https://aws.amazon.com/th/what-is/transformers-in-artificial-intelligence/

[14] BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding https://arxiv.org/abs/1810.04805

อ้างอิงและแหล่งที่มา (ต่อ)

[15] WangchanBERTa https://airesearch.in.th/releases/wangchanberta-pre-trained-thai-language-model/

[16] Tom B Brown, (n.d.), "Language Models are Few-Shot LearnersMT5", "MT5" , Retrieved 30 April 2023, from https://arxiv.org/abs/2005.14165

[17] "Generative Pre-trained Transformer : GPT" , (2022), Retrieved 30 April 2023, from https://www.bualabs.com/archives/4419/what-is-gpt-4-generative-pre-trained-transformer-gpt

[18] "Sequence to Sequence Model : Seq2Seq Model" , (2022), Retrieved 30 April 2023, from https://attapol.github.io/compling/seq2seq.html

[19] "การปรับแต่งโมเดล (Fine tuning model) " , (2022), Retrieved 30 April 2023, from https://kittimasak.com/transfer-learning-deep-learning-python/

[20] "การออกแบบคำสั่งพร้อมพ์ (Prompt engineering) และการปรับแต่งคำสั่งพร้อมพ์ (Prompt tuning)" , (2024), Retrieved 21 Febuary 2024, from https://medium.com/@aabhi02/prompt-engineering-vs-prompt-tuning-a-detailed-explanation-19ea8ce62ac4

[21] "Weighted Average" , (2022), Retrieved 30 April 2023, from https://www.investopedia.com/terms/w/weightedaverage.asp

[22] "Recall-Oriented Understudy for Gisting Evaluation : ROUGE" , (2022), Retrieved 30 April 2023, from https://huggingface.co/spaces/evaluate-metric/rouge

[23] "BERTSCORE: EVALUATING TEXT GENERATION WITH BERT" , (2022), Retrieved 30 April 2023, in IEEE Access

[24] PING LI, JIONG Y, JIAYING CHEN, BINGLEI GUO, (2021), "HG-News: News Headline Generation Based on a Generative Pre-Training Model", in IEEE Access, pp. 110039-110040.

[25] FUCHENG YOU, SHUAI ZHAO, JINGJING CHENA, (2020), "Topic Information Fusion and Semantic Relevance for Text Summarization", in IEEE Access, pp. 178946-178947.

[26] Pongstorn Harnmetta and Taweesak Samanchuen. (2022), "Sentiment Analysis of Thai Stock Reviews Using Transformer Models", in IEEE Access.

อ้างอิงและแหล่งที่มา (ต่อ)

[27] Sheher Bano and Shah Khalid, (2022), "BERT-based Extractive Text Summarization of Scholarly Articles: A Novel Architecture", in IEEE Access.

[28] Anandan Chinnalagu and Ashok Kumar Durairaj, (2022), "Comparative Analysis of BERT-base Transformers and Deep Learning Sentiment Prediction Models", in IEEE Access, pp. 874-877.

[29] Jiaohong Yao (2022), "Personalized News Headline Generation System with Fine-grained User Modeling", in IEEE Access.

[30] Nutthanit Wiwatbutsiri, Atiwong Suchato, Proadpran Punyabukkana, Nuengwong, "Question Generation in the Thai Language Using MT5", in IEEE Access.