บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

การวิเคราะห์ความรู้สึก (Sentiment Analysis) [1-3] เป็นสาขาวิจัยทางด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ที่ใช้ในการประเมินความรู้สึกของลูกค้าหรือผู้บริโภคจากบทวิจารณ์ของลูกค้า (Customer Review) ที่มีต่อสินค้าหรือบริการว่ามีความรู้สึกเป็นบวก (Positive) ลบ (Negative) หรือเป็นกลาง (Neutral) โดยทั่วไปเทคนิคที่นิยมใช้ในการวิเคราะห์ความรู้สึกคือ เทคนิคด้านการจำแนกเอกสาร (Text Classification) ดังนั้นการวิเคราะห์ความรู้สึกด้วยเทคนิคด้านการจำแนกเอกสาร จึงเรียกว่าการจำแนกความรู้สึก (Sentiment Classification) [2]

การวิเคราะห์ความรู้สึกเป็นประโยชน์สำหรับการดำเนินการทางธุรกิจเป็นอย่างมาก เพราะการประเมินความรู้สึกของลูกค้าที่มีต่อสินค้าและบริการนั้น สามารถนำมาใช้ประโยชน์อย่างมากเช่น เจ้าของสินค้าและบริการสามารถใช้ข้อมูลด้านการวิเคราะห์ความรู้สึกในการปรับปรุงสินค้าและบริการของตนเอง เป็นข้อมูลที่ใช้ในการรักษาแบรนด์หรือฐานลูกค้า รับรู้แนวโน้มทางการตลาดของสินค้าและบริการ หรือการตัดสินใจเกี่ยวกับสินค้าและบริการของตน ในขณะที่ในส่วนของลูกค้าหรือผู้บริโภคข้อมูลดังกล่าวจะช่วยในการตัดสินใจว่าควรจะซื้อสินค้าหรือบริการเหล่านั้นหรือไม่

จากการศึกษาที่ผ่านมาพบว่า กระบวนการในการจำแนกความรู้สึกสามารถแบ่งออกเป็น 3 ประเภท [1] ได้แก่ กระบวนการที่ใช้คำศัพท์ (Lexicon-based Approaches) กระบวนการที่ใช้อัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning-based Approaches) และกระบวนการที่ใช้อัลกอริทึมการเรียนรู้เชิงลึก (Deep Learning-based Approaches) กระบวนการเหล่านี้มีการประยุกต์ใช้งานอย่างกว้างขวาง อย่างไรก็ตามพบว่า เทคนิคเหล่านี้มองข้ามข้อมูลของอารมณ์จากการพิจารณาบริบทในบทวิจารณ์ [4] ดังนั้นเวกเตอร์ที่สร้างขึ้นมาอาจจะเป็นเพียงเวกเตอร์ของคำ และบ่อยครั้งยังพบปัญหาที่เรียกว่า Out-of-Vocabulary (OOV) [5] นั่นคือคำศัพท์ที่ไม่ได้เป็นส่วนหนึ่งของพจนานุกรมทั่วไปที่พบในสภาพแวดล้อมการประมวลผลภาษาธรรมชาติ สาเหตุนี้นำไปสู่การสูญหายของข้อมูล [4] นอกจากนั้นความท้าทายอีกประการหนึ่งสำหรับการจำแนกความรู้สึกก็คือการขาดข้อมูลที่มีคำอธิบายประกอบ (Annotated Data) ซึ่งจะส่งผลให้เกิดการเรียนรู้โมเดลที่ให้ค่า false negative สูง [6] และบางครั้งยังพบว่า มีความขัดแย้งระหว่างบริบทของการรีวิวและคลาสลาเบลซึ่งเป็นสาเหตุของจำแนกข้อมูลที่ผิดพลาด (Misclassification) [4]

ไม่กี่ปีที่ผ่านมามีงานวิจัยด้านการจำแนกความรู้สึก ได้ประยุกต์นะเอากระบวนการแบบทรานสฟอร์มเมอร์ (Transformer-based Approach) มาใช้ ซึ่งเป็นกระบวนการที่มีการพิจารณาโครงสร้างและความหมายของคำควบคู่กับการวิเคราะห์อารมณ์ในบริบทของข้อมูล [4] เนื่องจากโมเดลแบบทรานสฟอร์มเมอร์ (Transformer Model) พัฒนาด้วยโครงข่ายประสาทเทียมด้วยโครงสร้างแบบ Encoder-Decoder (Neural Network with an Encoder-Decoder Structure) ที่เรียนรู้บริบทและความหมายแบบ Self-Attention [7] ที่จะวิเคราะห์และเปรียบเทียบข้อมูลอินพุตทั้งหมดในลักษณะการวิเคราะห์แบบ Sequence-to-Sequence เข้าด้วยกันเพื่อคำนวณหาความสัมพันธ์และลำดับของคำในภาษาหรือประโยค ดังนั้นจึงมักจะมีการกล่าวว่าโมเดลแบบทรานสฟอร์มเมอร์ก็คือโมเดลของภาษา (Language Model)

1.2 วัตถุประสงค์ของโครงงาน

นำเสนอโมเดลแบบทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึก

1.3 ขอบเขตของโครงงาน

1.3.1 ขอบเขตของโมเดล

1) BERT Base จะเป็นโมเดล Pre-train แบบทรานสฟอร์มเมอร์ที่ถูกประยุกต์ใช้สำหรับการ Fine-tune เพื่อสร้างโมเดลเพื่อการจำแนกความรู้สึกจากบทวิจารณ์โรงแรมแบบหลายกลุ่ม
2) โมเดลเพื่อการจำแนกความรู้สึกจากบทวิจารณ์โรงแรมจะเป็นการจำแนกความรู้สึกออกเป็น 3 กลุ่ม ได้แก่ ความรู้สึกเป็นบวก (Positive) ลบ (Negative) หรือเป็นกลาง (Neutral)
3) ชุดข้อมูลที่ใช้ในการศึกษาเป็นบทวิจารณ์โรงแรมที่เป็นภาษาอังกฤษ โดยดาวน์โหลดมาจากเว็บ TripAdvisor จำนวนชุดข้อมูลมากกว่า 30,000 ชุดข้อมูล
   - ข้อมูลบทวิจารณ์ที่ดาวน์โหลดมาและมีค่าเรตติ้งเป็น 5 จะจัดเป็นกลุ่มความรู้สึกแบบบวก จำนวน 10,000 ชุดข้อมูลขึ้นไป
   - ข้อมูลบทวิจารณ์ที่ดาวน์โหลดมาและมีค่าเรตติ้งเป็น 3 จะจัดเป็นกลุ่มความรู้สึกแบบเป็นกลาง จำนวน 10,000 ชุดข้อมูลขึ้นไป
   - ข้อมูลบทวิจารณ์ที่ดาวน์โหลดมาและมีค่าเรตติ้งเป็น 1 จะจัดเป็นกลุ่มความรู้สึกแบบเป็นลบ จำนวน 10,000 ชุดข้อมูลขึ้นไป
4) พัฒนาโมเดลผ่าน Google Colaboratory (Google Colab)

5) ในการ Fine-tune โมเดลแบบ BERT Base เพื่อสร้างโมเดลเพื่อการจำแนกความรู้สึกจากบทวิจารณ์โรงแรมนั้น จะใช้ข้อมูล 70% ที่มีของแต่ละคลาส และอีก 30% ที่เหลือขะใช้เป็นข้อมูลทดสอบ
6) ประเมินประสิทธิภาพของโมเดลเพื่อการจำแนกความรู้สึกจากบทวิจารณ์โรงแรมด้วยค่า Accuracy, Recall, Precision, และ F1

1.3.2 ขอบของเว็บ

1) ขอบเขตของผู้ใช้เว็บ
   - แนบไฟล์ .csv ได้
   - เขียนบทวิจารณ์เป็นข้อความ โดยข้อความเป็นภาษาอังกฤษได้
   - ล้างบทวิจารณ์ที่เขียน และ ยกเลิกการแนบไฟล์ได้
2) ขอบเขตการแสดงผล
   - สามารถแสดงผลลัพธ์จากการทำนายว่าเป็นคลาสได้
   - มีเมนูสำหรับเลือกการส่งข้อมูลคือ แบบแนบไฟล์ และแบบเขียนบทวิจารณ์

1.4 ประโยชน์ที่คาดว่าจะได้รับ

ได้กระบวนการในการจำแนกข้อความแสดงความรู้สึกด้วยกระบวนการแบบทรานสฟอร์มเมอร์

1.5 อุปกรณ์และเครื่องมือที่ใช้ในการดำเนินงาน

1.5.1 เครื่องคอมพิวเตอร์ที่ใช้ในการพัฒนาโมเดล

Material:	Google colab เป็นโฮสต์โปรแกรม Jupyter notebook บน Cloud ของ Google ชื่อเต็ม คือ Google Colaboratory

1.5.2 เครื่องคอมพิวเตอร์ที่ใช้งาน

Hardware:	คอมพิวเตอร์รุ่น Intel® Core™ I5-9400F CPU @ 2.90 GHz , RAM 16 GB BUS 2666 MHz, SSD SATA 240 GB
Operating System:	Windows 10 Pro
Programming Language:	Python, HTML5, CSS3
Application Tools:	visual Studio Code

1.6 แผนการดำเนินงาน

โครงงานปริญญานิพนธ์ฉบับนี้ ดำเนินงาน ณ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ระหว่างเดือนตุลาคม 2565 ถึง กันยายน 2566 ดังที่แสดงในตารางที่ 1.1

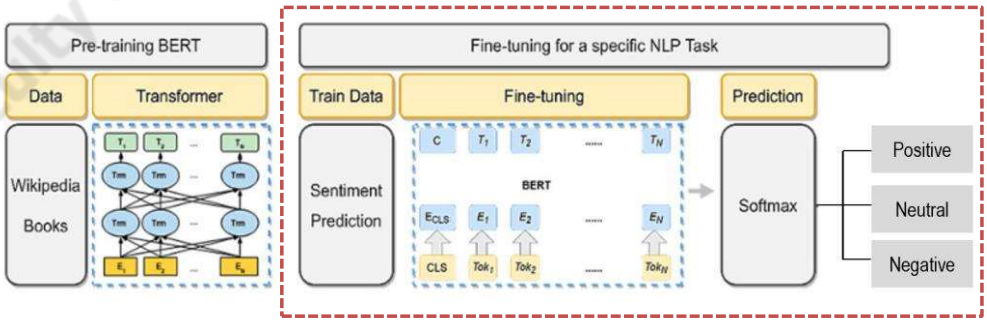
ตารางที่ 1.1 แผนการดำเนินงาน

กิจกรรม	เดือน											
	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.	พ.ค.	มิ.ย.	ก.ค.	ส.ค.	ก.ย.
1. ศึกษาเทคนิคและรวบรวมข้อมูล												
2. ออกแบบกระบวนการวิจัย												
3. เตรียมข้อมูล												
4. พัฒนาโปรแกรม วัดประสิทธิภาพ และปรับปรุง												
5. เขียนเอกสารฉบับสมบูรณ์												

1.7 กรอบการดำเนินงาน

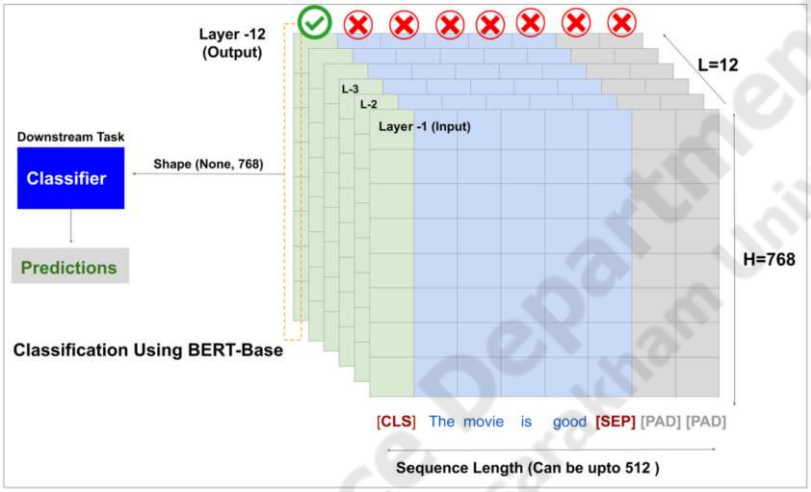
1.7.1 การสร้างโมเดลทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึก

กรอบการดำเนินงานในการประยุกต์โมเดลแบบทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึกสามารถแสดงได้ดังภาพประกอบที่ 1.1 และสามารถอธิบายการทำงานในระบบฯ ได้ดังนี้



ภาพประกอบที่ 1.1 กรอบการดำเนินงานของโมเดลทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึก

ขั้นตอนที่ 1 : คือการเลือก Pre-training BERT ที่เหมาะสมต่อการประยุกต์ใช้งาน โดยในที่นี้จะทำการทดสอบกับ BERT Base ซึ่งจะใช้พารามิเตอร์ทั้งหมด 110 ล้านพารามิเตอร์ ซึ่งใช้จำนวนเลเยอร์ทั้งสิ้น 12 เลเยอร์ ซึ่งจำนวนเลเยอร์ก็คือจำนวน Transformer Blocks ในขณะที่ขนาดของ Hidden คือ 768 และจำนวนของ Self-attention Heads เท่ากับ 12 ส่วน ในการประมวลผลต้องการ 1 GPU



ภาพประกอบที่ 1.2 Fine-tuning โมเดล BERT สำหรับ Sentiment Analysis ด้วย Google Colab

ที่มา: https://www.analyticsvidhya.com/blog/2021/12/fine-tune-bert-model-for-sentiment-analysis-in-google-colab/

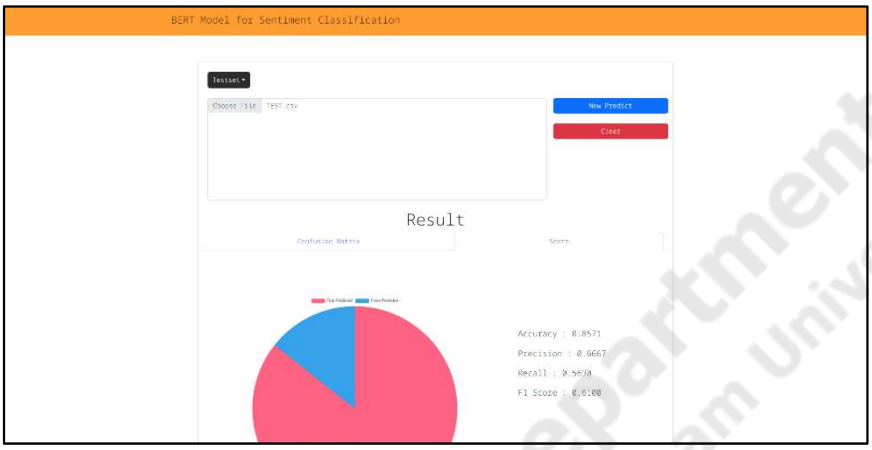
ขั้นตอนที่ 2 : คือการ Fine-tune BERT ผ่าน Google Colab ให้เหมาะสมกับการประยุกต์ใช้งานในด้านการจำแนกความรู้สึก โดยจะใช้ข้อมูลบทวิจารณ์โรงแรมในภาษาอังกฤษที่ดาวน์โหลดมาจากเว็บ TripAdvisor ซึ่งข้อมูลที่ใช้ในการ Fine-tune จะมี 3 คลาสคือ ความรู้สึกเป็นบวก (Positive) ลบ (Negative) หรือเป็นกลาง (Neutral) ซึ่งการ Fine-tune BERT สำหรับงานในด้านการจำแนกความรู้สึกสามารถแสดงได้ดังภาพประกอบที่ 1.2

ขั้นตอนที่ 3 : ภายหลังจากการ Fine-tune ก็จะได้โมเดลทรานสฟอร์มเมอร์สำหรับการทำนายความรู้สึกจากบทวิจารณ์โรงแรมร่วมกับฟังก์ชัน SoftMax

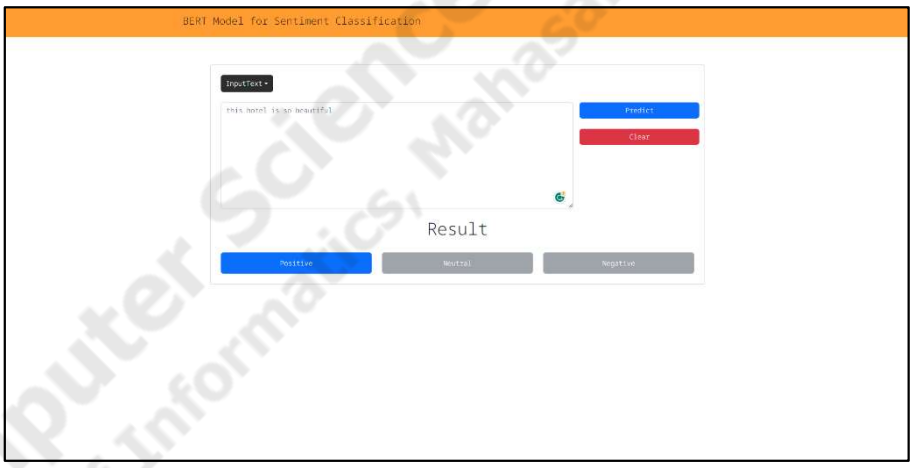
เมื่อได้โมเดลที่เหมาะสมกับการใช้งาน จะนำโมเดลที่ได้ไปใช้ในหน้าจอ GUI สำหรับแสดงการทดลองการใช้งานโมเดล

1.7.2 หน้าจอสำหรับแสดงการทดลองการใช้งานโมเดล

หน้าจอสำหรับแสดงการทดลองการใช้งานโมเดลทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึก สามารถแสดงได้ดังภาพประกอบที่ 1.3 และภาพประกอบที่ 1.4



ภาพประกอบที่ 1.3 การใช้โมเดลทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึกด้วย Test Set



ภาพประกอบที่ 1.4 การใช้โมเดลทรานสฟอร์มเมอร์สำหรับการจำแนกความรู้สึกด้วยการเขียนรีวิว