รายงานวิจัย

เรื่อง

การวิเคราะห์ขั้วอารมณ์แบบผสมจากข่าวหลักทรัพย์ในประเทศไทย
(Mixed Sentiment Analysis for Stock News in Thailand)

ผู้วิจัย
นางสาวจุฑารัตน์ ภู่อมร
(Jutharat Pooamorn)

นางสาวนฤมล คงโอ
(Naruemon Khongao)

รายงานวิจัยนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตร์บัณฑิต
สาขาวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์
มหาวิทยาลัยรามคำแหง
ประจำปี 2568

หัวข้อรายงานวิจัย	การวิเคราะห์ขั้วอารมณ์แบบผสมจากข่าวหลักทรัพย์ในประเทศไทย Mixed Sentiment Analysis for Stock News in Thailand
รายชื่อนักศึกษา	นางสาวจุฑารัตน์ ภู่อมร และ นางสาวนฤมล คงโอ
รหัสนักศึกษา	6405003317 และ 6205002246
ปริญญา	วิทยาศาสตร์บัณฑิต
สาขาวิชา	วิทยาการคอมพิวเตอร์
พ.ศ.	2568
อาจารย์ที่ปรึกษารายงานวิจัย	รศ.ดร.อัณณ์นุพันธ์ รอดทุกข์

รายงานวิจัยฉบับนี้ ได้รับการอนุมัติให้เป็นส่วนหนึ่งของการศึกษา

ตามหลักสูตรวิทยาศาสตร์บัณฑิต คณะวิทยาศาสตร์ มหาวิทยาลัยรามคำแหง

.....

(รศ.ดร.อัณณ์นุพันธ์ รอดทุกข์)

อาจารย์ผู้ควบคุมรายงานวิจัย

บทคัดย่อ

ปัจจุบัน มีการพัฒนาการจำแนกความรู้สึก (sentiment analysis) ของการประมวลผลภาษาธรรมชาติ (NLP) เพื่อเป็นเกณฑ์การตัดสินใจทางเลือกรูปแบบหนึ่ง โดยบทความวิจัยชิ้นนี้ได้นำเสนอ กรณีศึกษาข่าวหลักทรัพย์ในประเทศไทย ซึ่งบริบทของประโยคข่าวสารการลงทุนด้านหุ้นไทย มีความผันผวนสูง เนื่องจากข้อความประกอบด้วย เชิงบวก (Positive) กลาง (Neutral) และเชิงลบ (Negative) โดยสามารถแบ่งประโยคเป็น 2 ประเภท คือ ข้อความขั้วอารมณ์เดียว (Single Sentiment) ข้อความหลายขั้วอารมณ์ (Mixed Sentiment) ซึ่งข้อความข่าวสารการลงุทนประเภทความรู้สึกผสมมีแนวโน้มเพิ่ม และการจำแนกข้อความรู้สึกผสมยังไม่มีงานวิจัยมารองรับมากนัก บทความนี้จึงนำเสนอวิธีวิจัยเพื่อเปรียบเทียบ โดยใช้เทคนิค Logistic Regression (LR) , Naïve Bayes (NB) , Support Vector Machine (SVM) , Recurrent Neural Network (RNN) , WangchanBERTa Model (Thai Language Model) โดยประสิทธิภาพของโมเดลแสดงผ่านค่า Precision () Recall () และค่า F1-score () สำหรับข้อความรู้สึกเดียวและ ข้อความรู้สึกผสม และมีค่า Accuracy () แสดงให้เห็นถึงประสิทธิภาพในการจำแนกข่าวสารหุ้นในแต่ละเทคนิค

ABSTRACT

The Fluctuation of the Thai stock market has significant influence from several domestic factors and international reasons. It makes stock investing a delicate matter due to the amount of investment information, many investors are forced to determine whether to buy or sell stocks based on limited information and time which affects market indexes and individual stock prices. Therefore, news classification is another aspect that makes stock news data analysis more advantageous. To consider the advancements in natural language processing technology (NLP), language models are the top choices. In dataset classification, training the learning model with single sentiment text and adjusting the complexity by training the learning model with mixed sentiment text to enhance the ability to analyze and classify more specifically. The performance of the model is shown through the Precision () Recall () and F1-score () values for single sentiment text and mixed sentiment text, and the Accuracy () values illustrate the efficiency in evaluating the types of stock news data by deep learning techniques.

กิตติกรรมประกาศ

งานวิจัยครั้งนี้สำเร็จลุล่วงได้ด้วยความอนุเคราะห์จากทุกฝ่ายที่เกี่ยวข้องโดยเฉพาะ เพื่อนนักศึกษาและคณาจารย์ภาควิชาวิทยาการคอมพิวเตอร์ โดยเฉพาะอย่างยิ่งจาก รศ.ดร.อัณณ์นุพันธ์ รอดทุกข์ ที่กรุณาให้คำปรึกษาและให้ข้อแนะนำสำหรับการแก้ไขจุดบกพร่องต่าง ๆ ในงานวิจัยชิ้นนี้ ผู้วิจัยรู้สึกขอบพระคุณเป็นอย่างยิ่งที่ได้รับความช่วยเหลือและความทุ่มเทจากอาจารย์ ทั้งนี้ผู้วิจัยขอขอบคุณผู้ที่มีส่วนเกี่ยวข้องช่วยเหลือ สนับสนุนและเป็นกำลังใจเสมอมา ผู้วิจัยหวังเป็นอย่างยิ่งว่าการวิจัยนี้จะเป็นประโยชน์ต่อบุคลากรทางการศึกษา นักวิจัยและผู้ที่สนใจงานวิจัยด้านการประยุกต์ใช้ปัญญาประดิษฐ์ด้านการประมวลผลภาษาธรรมชาติ ตลอดจนจะเป็นประโยชน์ต่อการสร้างองค์ความรู้ใหม่ๆ ทางด้านวิทยาศาสตร์เพื่อเป็นส่วนสนับสนุนต่อส่วนรวม

นางสาวจุฑารัตน์ ภู่อมร และ นางสาวนฤมล คงโอ

15 พฤษภาคม 2568

สารบัญ

บทคัดย่อ **ก**

ABSTRACT **ข**

กิตติกรรมประกาศ **ค**

สารบัญ **ง**

สารบัญตาราง **ฉ**

บทที่ 1 **8**

บทนำ 8

1.1 ที่มาและความสำคัญของปัญหา *8*

1.2 วัตถุประสงค์ *1*

1.3 ขอบเขตของโครงการวิจัย *2*

1.4 นิยามศัพท์เฉพาะ *2*

1.5 ประโยชน์ที่คาดว่าจะได้รับจาการวิจัย *3*

บทที่ 2 **4**

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง 4

2.1 เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (TF-IDF) *4*

2.2 เทคนิคการวิเคราะห์ความถดถอยโลจิสติก (Logistic Regression: LR) *4*

2.3 เทคนิคการจำแนกประเภทโดยใช้ทฤษฎีของเบย์ (Naïve Bayes: NB) *5*

2.4 เทคนิคซัพพอร์ตเวกเตอร์แมชีน (Support Vector Machine: SVM) *7*

2.5 โมเดลประมวลผลภาษาไทยวังจันทร์BERTa (WangchanBERTa) ***Error! Bookmark not defined.***

2.6 งานวิจัยที่เกี่ยวข้อง ***Error! Bookmark not defined.***

บทที่ 3 **15**

วิธีดำเนินการวิจัย ERROR! BOOKMARK NOT DEFINED.

3.1 การออกแบบกระบวนการดำเนินการวิจัย ERROR! BOOKMARK NOT DEFINED.

3.2 การเตรียมข้อมูล ERROR! BOOKMARK NOT DEFINED.

3.3 โมเดลการเรียนรู้เชิงลึก ERROR! BOOKMARK NOT DEFINED.
3.4 การออกแบบชุดฝึกการทดลอง ERROR! BOOKMARK NOT DEFINED.
บทที่ 4 ERROR! BOOKMARK NOT DEFINED.
ผลดำเนินการวิจัยและการวิเคราะห์ข้อมูล ERROR! BOOKMARK NOT DEFINED.
4.1 ตัวชี้วัดประสิทธิภาพของการฝึกสอนโมเดล ERROR! BOOKMARK NOT DEFINED.
4.2 ผลการทดลองฝึกสอนโมเดล ด้วยชุดฝึกการเรียนรู้แบบต่างๆ ERROR! BOOKMARK NOT DEFINED.
4.3 อภิปรายผลการทดลอง ERROR! BOOKMARK NOT DEFINED.
บทที่ 5 ERROR! BOOKMARK NOT DEFINED.
สรุปผลการวิจัย และข้อเสนอแนะ ERROR! BOOKMARK NOT DEFINED.
บรรณานุกรม ERROR! BOOKMARK NOT DEFINED.

สารบัญตาราง

ตารางที่ 1 ชุดข้อมูลที่ถูกออกแบบตามหลักการของชุดฝึกการเรียนรู้สำหรับโมเดลการเรียนรู้เชิงลึก ERROR! BOOKMARK NOT DEFINED.

ตารางที่ 2 แสดงประสิทธิภาพของการฝึกสอนโมเดลด้วยชุดฝึกการเรียนรู้ในรูปแบบต่าง ๆ ERROR! BOOKMARK NOT DEFINED.

สารบัญรูปภาพประกอบ

รูปภาพที่ 1 โครงสร้างจำแนกข้อความ WANGCHANBERTA ERROR! BOOKMARK NOT DEFINED.

รูปภาพที่ 2 สถาปัตยกรรมของ TRANSFORMERS 7

รูปภาพที่ 3 การแปลงข้อความให้เป็นตัวเลขที่โมเดลสามารถเข้าใจ 8

รูปภาพที่ 4 การระบุตำแหน่ง TOKEN 9

รูปภาพที่ 5 ตำแหน่งของเวกเตอร์คำ 9

รูปภาพที่ 6 การใช้ความถี่เพื่อทำ POSITIONAL ENCODING 10

รูปภาพที่ 7 SELF MULTI-HEAD ATTENTION 11

รูปภาพที่ 8 ENCODER SELF-ATTENTION 12

รูปภาพที่ 9 ขั้นตอนการวิเคราะห์ข้อมูล ERROR! BOOKMARK NOT DEFINED.

รูปภาพที่ 10 การจำแนกประเภทของคำ (TOKEN CLASSIFICATION) 15

รูปภาพที่ 11 ขั้นตอนการวิเคราะห์ความรู้สึกเดียว (SINGLE SENTIMENT) ERROR! BOOKMARK NOT DEFINED.

รูปภาพที่ 12 ขั้นตอนการวิเคราะห์ความรู้สึกผสม (MIXED SENTIMENT) 16

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันประเทศไทยมีระบบเศรษฐกิจแบบผสม โดยมีปัจจัยการลงทุนต่างประเทศมีสัดส่วนการลงทุนต่อภาพรวมเศรษฐกิจโดยรวมของประเทศ ทำให้ระบบเศรษฐกิจที่ผสมผสาน ระหว่างภาครัฐเป็นและเอกชนสามารถดำเนินการทางเศรษฐกิจร่วมกัน โดยเปิดโอกาสให้ภาคเอกชนสามารถเป็นผู้นำ การค้า การลงทุนทั้งในและต่างประเทศ และให้ภาครัฐสามารถแทรกแซงกิจกรรมทางเศรษฐกิจตามสมควร ส่งผลให้เมื่อภาคเอกชนต้องการขับเคลื่อนองค์กรให้สามารถเพิ่มโอกาสและความสามารถการแข่งขันจึงเกิดการระดมทุนในตลาดหุ้นเกิดขึ้น ไม่ว่าจะเป็นการเสนอขายหุ้นครั้งแรกแก่ประชาชนทั่วไป (Initial Public Offering - IPO) และการเสนอขายหุ้นเพิ่มทุน ส่งผลให้มีผลกระทบต่อสาธารณะชนโดยเฉพาะนักลงทุนที่เป็นประชาชน จึงต้องมีกฎระเบียบ และข้อกำหนดจาก สำนักงานคณะกรรมการกำกับหลักทรัพย์และตลาดหลักทรัพย์ (ก.ล.ต.) เพื่อให้การระดมทุนเป็นไปอย่างโปร่งใสและเป็นธรรม จึงต้องมีการเปิดเผยตามเหตุการณ์เป็นการรายงาน เช่น งบการเงิน การจ่ายปันผล การผิดชำระหนี้ เมื่อเกิดเหตุการณ์ที่ส่งผลกระทบต่อการตัดสินใจลงุทนในหลักทรัพย์บริษัท เพื่อให้การซื้อขายหลักทรัพย์ในตลาดหลักทรัพย์ เป็นไปอย่างโปร่งใส

“ตลท. มีการกำหนดหลักเกณฑ์และมีกระบวนการในการพิจารณารับหลักทรัพย์จดทะเบียนที่มีความชัดเจน โปร่งใส คำนึงถึงคุณภาพของหลักทรัพย์ที่รับจดทะเบียน มีการกำกับดูแลการเปิดเผยข้อมูลของบริษัทจดทะเบียนและบริษัทผู้ออกหลักทรัพย์ เพื่อให้ผู้ลงทุนมีข้อมูลอย่างเพียงพอ เท่าเทียม ทั่วถึง และภายในระยะเวลาที่เหมาะสม ตลอดจนการพิจารณาเพิกถอนหลักทรัพย์จดทะเบียนที่มีคุณสมบัติไม่เหมาะสม” ทำให้นักลงทุนคัดเลือกหุ้นจากดัชนีที่สะท้อนมูลค่าหุ้น ณ ขนาดที่ทำการประเมินราคา พิจารณาจากปัจจัยต่าง ๆ เช่น ปัจจัยพื้นฐานจากอัตราการเติบโตของรายได้ เงินปันผล ปัจจัยทางเทคนิคปริมาณการซื้อขายหุ้น ภาวะตลาดโดยรวม ดังนั้นการพิจารณาดัชนีราคาหลักทรัพย์เป็นส่วนหนึ่งของตัวชี้วัด เพื่อประเมินความสำเร็จจากการลงทุน

ในขณะที่ข่าวสารภาพรวมดัชนีตลาดหุ้นทั้งแบบ รายตัว กองทุนรวม หรือแม้แต่ ข่าวหุ้นของตลาดประเทศมักส่งผลกระทบต่อราคาหุ้นไทย ประกอบกับข้อมูลที่ใช้พิจารณาหุ้น ในแต่ละช่วงเวลามีปริมาณมหาศาล และต้องใช้ระยะเวลาในการตัดสินใจอย่างจำกัด ทำให้นักลงทุนเกิดภาวะ การลงทุนที่มีอคติและมีอารมณ์เข้ามาเกี่ยวข้อง กล่าวคือ เมื่อผู้ลงทุนได้รับข่าวสารที่มีลักษณะเป็นประโยคที่มีหลายขั้วอารมณ์ เช่น หุ้น NEO คาดไตรมาส 2/68 กำไรลดลงเหลือเพียง 91 ล้านบาท หดตัว -61% และข่าวสารที่มีลักษณะขั้วอารมณ์เดียว เช่น เช้าวันนี้ SET ปิดที่ระดับ 632.32 จุด เพิ่มขึ้น 5.88 จุด หรือ 0.94 % (ที่มา https://www.kaohoon.com/ และ https://www.hooninside.com/news-feed/) จะเห็นได้ว่าหากนักลงทุนได้รับข่าวสารหลายขั้วอารมณ์ในปริมาณที่เยอะอาจทำให้เกิดการตัดสินใจที่ผิดพลาด เพื่อทำความเข้าใจและจัดการกับ ข้อมูลข่าวสารที่มีหลายขั้ว

อารมร์ จึงนำเอาการวิเคราะห์ความรู้สึก ซึ่งเปนการประมวลผลภาษาธรรมชาติ เพื่อตีความข้อมูล ออกเป็น กลาง บวก ลบ โดยประโยคที่มีหลายขั้วอารมณ์มีความท้าทายสูง เนื่องจากมีโครงสร้างซับซ้อน ขาดบริบทที่ชัดเจน มีการเปรียบเทียบ ทำให้เมื่อจัดประเภทประโยคให้เป็นขั้วเดียว

https://www.scb.co.th/th/personal-banking/stories/grow-your-wealth/prejudice-investors-terrible-trap

https://www.krungsriasset.com/TH/Plan-your-investment/Learn-about-Investment/bias-in-investment.aspx

1.2 วัตถุประสงค์

1. เพื่อพัฒนาเทคนิคการวิเคราะห์ความรู้สึกของข่าวสารที่มีหลายขั้วอารมณ์ ซึ่งส่งผลต่อการตัดสินใจลงทุนในตลาดหลักทรัพย์ไทย
2. เพื่อเปรียบเทียบประสิทธิภาพของโมเดลในการจำแนกข่าวสารที่มีความรู้สึกผสม (Mixed Sentiment)
3. เพื่อพัฒนาเครื่องมือช่วยสนับสนุนการตัดสินใจลงทุนโดยใช้โมเดลภาษาไทยที่มีความแม่นยำสูง

1.3 ขอบเขตของโครงการวิจัย

การวิจัยนี้ ศึกษาประสิทธิภาพของโมเดลจากการเปรียเทียบแบบจำลองการเรียนรู้ของเครื่อง การเรียนรู้เชิงลึก และ โมเดลภาษาไทยที่ถูกฝึกมาแล้ว (pre-trained model) ในการจำแนกวิเคราะห์ความรู้สึก (Sentiment Analysis) ขั้วอารมณ์ที่ซ่อนอยู่ในข่าวสารหุ้น ตามจำนวนชุดข้อมูลข่าวสารหุ้นที่ประกอบด้วยข้อความ เชิงบวก (Positive) เชิงลบ (Negative) เป็นกลาง (Neutral) และความรู้สึกผสม (Mixed) เพื่อใช้ในการฝึกสอนและทดสอบแบบจำลองทั้งหมด

1.4 นิยามศัพท์เฉพาะ

โมเดลที่ผ่านการเทรน (Pre-Trained Model) คือ โมเดลที่ผ่านการเทรน บนชุดข้อมูลที่ได้จากแหล่งต่างๆ เช่น ข่าว, วิกิพีเดีย, ข้อความในโซเชียลมีเดีย และข้อมูลที่ได้จากการ crawl เว็บไซต์ในอินเทอร์เน็ต ซึ่งมีขนาดข้อมูลขนาดใหญ่ และได้วัดประสิทธิภาพของโมเดลที่ผ่านกระบวนการปรับแต่งโมเดลทาง (Fine-tuning)

ปัญญาประดิษฐ์ (Artificial Intelligence) คือ โปรแกรมคอมพิวเตอร์ที่มีความสามารถในการเลียนแบบความฉลาดของมนุษย์ เรียนรู้ในด้านของกระบวนการคิด การกระทำ การให้เหตุผล การปรับตัว หรือการอนุมานและการทำงานของสมอง

การเรียนรู้ของเครื่อง (Machine learning) คือ อัลกอริทึมที่สร้างขึ้นมาเพื่อให้สอนคอมพิวเตอร์ให้เรียนรู้ด้วยตัวเอง และสามารถพัฒนาอัลกอริทึมได้ด้วยการป้อนชุดข้อมูลมหัต

การเรียนรู้เชิง (Deep learning) คือ การออกแบบอัลกอริทึมด้วยแนวคิดที่เลียนแบบการทำงานของระบบโครงข่ายประสาท (Neural Networks) ของสมองมนุษย์ ใช้ในการเรียนรู้และแก้ไขปัญหา สามารถตัดสินใจได้ใกล้เคียงกับมนุษย์

โครงข่ายประสาทเทียม (Artificial neural network) คือ การสร้างคอมพิวเตอร์ที่จำลองเอาวิธีการทำงานของสมองมนุษย์ หรือทำให้คอมพิวเตอร์รู้จักคิดและจดจำในแนวเดียวกับโครงข่ายประสาทของมนุษย์

1.5 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. ผลผลิต (Output) เปรียบเทียบประสิทธิภาพของโมเดลที่สามารถประมวลผลและจำแนกความรู้สึกจากข่าวสารการลงทุนในตลาดหุ้น โดยใช้เทคนิคการเรียนรู้ตามชุดฝึกที่เพื่อให้โมเดลสามารถปรับตัวและเรียนรู้จากข้อมูลที่มีความหลากหลายและลักษณะความรู้สึกผสม (Mixed Sentiment) ได้ดียิ่งขึ้น
2. ผลลัพธ์ (Outcome) สามารถจำแนกข้อความที่แสดงความรู้สึกผสม (Mixed Sentiment) จากชุดข้อมูลในลักษณะ unknown

ผลกระทบ (Impact) โมเดลการเรียนรู้เชิงลึกที่มีความสามารถในการจำแนกข้อความที่แสดงความรู้สึกผสม (Mixed Sentiment) ช่วยทำหน้าที่เป็นเครื่องมือในการช่วยสนับสนุนการตัดสินของนักลงทุน หรือบุคคลทั่วไปที่มีความสนใจการลงทุนประเภท หลักทรัพย์

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency : TF-IDF)

เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency – Inverse Document Frequency: TF-IDF) เป็นวิธีทางสถิติ ในการคำนวณค่าน้ำหนักของคำที่มีความสำคัญในแต่ละคำในเอกสารเมื่อเทียบกับเอกสารทั้งหมด ประกอบไปด้วยสองส่วน ความถี่ของคำ (Term Frequency) และ ส่วนกลับความถี่ของเอกสาร (Inverse Document Frequency)

ความถี่ของคำ คำนวณจากสูตร จำนวนครั้งที่คำปรากฏในเอกสาร หารด้วย จำนวนคำทั้งหมดในเอกสาร

$$TF(term, document) = \frac{f(term, document)}{\sum term' \in document\ f(term', document)}$$

ส่วนกลับความถี่ของเอกสาร คำนวณจากสูตร log จำนวนเอกสารทั้งหมด หารด้วย จำนวนเอกสารที่มีคำนั้นปรากฏ

$$IDF(term, allDocument) = log\frac{N}{df(t)}$$

การรวมค่า ความถี่ของคำ และ ส่วนกลับความถี่ของเอกสาร คำนวณจากสูตร ความถี่ของคำ คูณด้วยส่วนกลับความถี่ของเอกสาร

$$TF\,IDF = TF(term,document) \times IDF(term,allDocument)$$

2.2 เทคนิคการวิเคราะห์ความถดถอยโลจิสติก (Logistic Regression: LR)

Logistic Regression (LR) ถือเป็นเครื่องมือทางสถิติที่มีประสิทธิภาพในการวิเคราะห์ความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ (Independent Variable) กับตัวแปรตาม (Dependent Variable) ซึ่งอยู่ในรูปแบบข้อมูลทวินาม (Binary Outcome) ได้แก่ 0 และ 1 การประเมินค่าความน่าจะเป็นของตัวแปรตามดำเนินการผ่านแบบจำลองทางคณิตศาสตร์ที่พัฒนาในรูปของฟังก์ชันซิกมอยด์ (Sigmoid Function)

โดยทั่วไป หากค่าของตัวแปรอิสระมีระดับต่ำ ความน่าจะเป็นที่ตัวแปรตามจะมีค่าเท่ากับ 0 จะมีแนวโน้มสูงขึ้น ขณะที่เมื่อค่าตัวแปรอิสระมีค่ามากขึ้น ความน่าจะเป็นที่ตัวแปรตามจะมีค่าเท่ากับ 1 ก็จะเพิ่มขึ้นตามลำดับ กระบวนการนี้สามารถแสดงออกผ่านสมการฟังก์ชันซิกมอยด์ ดังนี้

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

โดยสัญลักษณ์ต่าง ๆ ของสมการอธิบายได้ดังนี้

z คือ ค่าที่ป้อนเข้า (input) ของฟังก์ชัน

e คือ ค่าคงที่ของลอการิทึมธรรมชาติ ≈ 2.71828

-z คือ ค่าตรงข้ามของ z

2.3 เทคนิคการจำแนกประเภทโดยใช้ทฤษฎีของเบย์ (Naïve Bayes: NB)

Naïve Bayes (NB) เป็นอัลกอริทึมการจำแนกประเภทที่อยู่ในกลุ่มของอัลกอริทึมเชิงกำหนด (Generative Algorithms) โดยมีรากฐานมาจากทฤษฎีของเบย์ (Bayes' Theorem) ซึ่งมุ่งเน้นการประเมินความน่าจะเป็นแบบมีเงื่อนไขเพื่อใช้ในการจำแนกข้อมูลออกเป็นกลุ่มต่าง ๆ อย่างมีประสิทธิภาพ อัลกอริทึมดังกล่าวจะสร้างแบบจำลองของการแจกแจงความน่าจะเป็นของข้อมูลอินพุตสำหรับแต่ละคลาสหรือหมวดหมู่ที่ระบุไว้

Naïve Bayes ดำเนินการภายใต้สมมติฐานสำคัญที่ว่า คุณลักษณะของข้อมูลอินพุตแต่ละตัวมีความเป็นอิสระต่อกันภายใต้เงื่อนไขของคลาส หรือที่เรียกว่า Conditional Independence Assumption สมการพื้นฐานของทฤษฎีเบย์ที่ใช้อธิบายหลักการจำแนกมีรูปแบบ ดังนี้

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

โดยสัญลักษณ์ต่าง ๆ ของสมการอธิบายได้ดังนี้

P (c | x) คือ ความน่าจะเป็นภายหลังของ*คลาส* (c *เป้าหมาย*) ที่กำหนดให้*ตัวทำนาย* (x *คุณลักษณะ*)

P (c) คือ ความน่าจะเป็นก่อนหน้าของ*คลาส*

P (x / *c*) คือ ความน่าจะเป็น ซึ่งเป็นความน่าจะเป็นของ*ตัวทำนาย*ที่กำหนด*คลาส*

P (*x*) คือ ความน่าจะเป็นก่อนหน้าของตัว*ทำนาย*

2.4 เทคนิคซัพพอร์ตเวกเตอร์แมชีน (Support Vector Machine: SVM)

Support Vector Machine (SVM) เป็นหนึ่งในอัลกอริทึมประเภทการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่ได้รับความนิยมอย่างแพร่หลายในงานด้านการจำแนกประเภท (Classification) และการถดถอย (Regression) โดยมีหลักการทำงานอิงจากการค้นหาขอบเขตการแบ่งข้อมูลที่เหมาะสมที่สุด ซึ่งเรียกว่า “ไฮเปอร์เพลน” (Hyperplane) เพื่อใช้ในการแยกกลุ่มข้อมูลต่าง ๆ ออกจากกันอย่างมีประสิทธิภาพ

แบบจำลองของ SVM มุ่งเน้นไปที่การหาขอบเขตที่มีระยะห่าง (Margin) ระหว่างคลาสมากที่สุด โดยแสดงออกในรูปสมการของเส้นแบ่งดังนี้

$$w^T x + b = 0$$

โดยสัญลักษณ์ต่าง ๆ ของสมการอธิบายได้ดังนี้

w คือ ค่าน้ำหนักของเวกเตอร์

x คือ ค่าเบี่ยงเบนจากการเรียนรู้ของ SVM

b คือ ค่าของข้อมูลที่เรียนรู้

อัลกอริทึม SVM ทำงานบนแนวคิดของการเพิ่มระยะห่างระหว่างไฮเปอร์เพลนกับข้อมูลที่อยู่ใกล้ขอบเขตการจำแนกมากที่สุด หรือที่เรียกว่า “ซัพพอร์ตเวกเตอร์” (Support Vectors) ซึ่งมีบทบาทสำคัญในการกำหนดขอบเขตการแบ่งข้อมูลที่เหมาะสมและมีความแม่นยำในการทำนาย

2.5 โมเดลเวกเตอร์คำ (Word Embeddings)

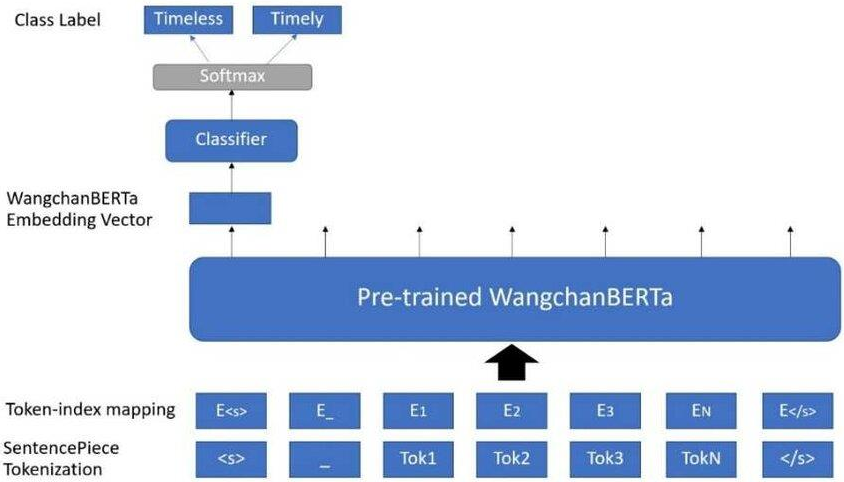
Thai2Vec เป็นโมเดลเวกเตอร์คำ (word embeddings) ที่ได้รับการออกแบบมาเพื่อรองรับการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ในบริบทของภาษาไทยโดยเฉพาะ โดยอิงหลักการจากเทคนิค Word2Vec ซึ่งเป็นวิธีการเรียนรู้แบบไร้การกำกับ (unsupervised learning) สำหรับการสร้างเวกเตอร์แทนคำในมิติสูง

ด้วยความซับซ้อนของภาษาไทย เช่น การไม่มีการเว้นวรรคระหว่างคำ การใช้คำซ้อน และลักษณะทางไวยากรณ์เฉพาะ Thai2Vec จึงถูกพัฒนาให้สามารถจัดการกับความท้าทายเหล่านี้ได้อย่างมีประสิทธิภาพ โดยอาศัยข้อมูลขนาดใหญ่จาก Thai Wikipedia และแหล่งข้อมูลออนไลน์อื่น ๆ ในการฝึกฝนโมเดล เพื่อสร้างเวกเตอร์คำในขนาด 300 มิติ ซึ่งสามารถสะท้อนบริบทและความหมายของคำในลักษณะเชิงกลไกได้อย่างแม่นยำ

2.6 โมเดลประมวลผลภาษาไทยวังจันทร์BERTa (WangchanBERTa)

https://arxiv.org/pdf/2101.09635

โมเดลประมวลผลภาษาไทยวังจันทร์ เป็นโมเดลทางภาษาไทยที่ได้ทำการเทรน (Pre-trained Thai Language Model) โดยโมเดลวังจันทร์เป็นแบบจำลองภาษาที่ใช้โครงสร้าง Transformer โดยเฉพาะสถาปัตยกรรมที่อิงตาม BERT ซึ่งการเทรนโมเดลใช้เทคนิค Masked Language Model (MLM) หรือการทำนายคำในประโยคที่ถูกแทนที่ด้วย masked token โดยก่อนการเทรนโมเดล ชุดข้อมูลที่นำมาเทรนต้องผ่านกระบวนการตัดแบ่งคำ (word tokenization) โดยการตัดแบ่งคำเป็น 4 รูปแบบการตัดแบ่งคำคือ การตัดแบ่งหน่วยคำย่อย (subword-level tokenization) การตัดแบ่งคำจาก dictionary ของคำในภาษาไทย ด้วยไลบรารี่ PyThaiNLP การตัดแบ่งพยางค์ในภาษาไทย จาก dictionary ของพยางค์ในภาษาไทย ด้วยไลบรารี่ PyThaiNLP และ การตัดแบ่งคำจากโมเดลการเรียนรู้ จากบทความทางวิชาการชื่อ “Stacked Ensemble Filter and Refine for Word Segmentation” ดังรูป



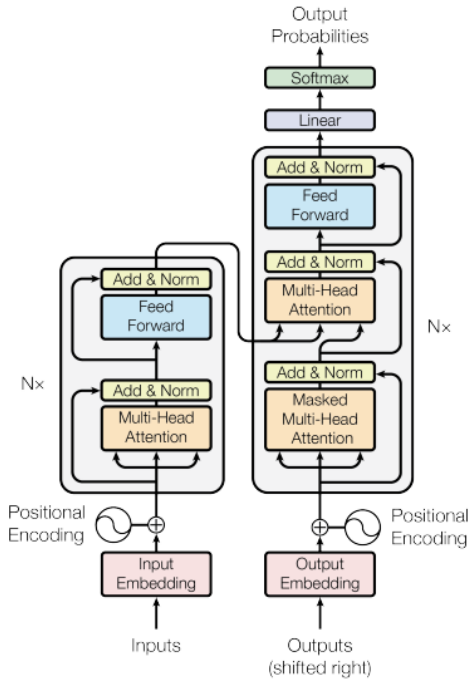
โครงสร้างจำแนกข้อความ WangchanBERTa

(Gatchalee et al.,(2023))

https://www.researchgate.net/publication/366902061_THAI_TEXT_CLASSIFICATION_EXPERIMENT_USING_CNN_AND_TRANSFORMER_MODELS_FOR_TIMELY-TIMELESS_CONTENT_MARKETING#fullTextFileContent

สถาปัตยกรรม Transformers เป็นแบบจำลองการแปลงลำดับข้อมูล (sequence transduction models) ที่ใช้กันแพร่หลายนั้นมีพื้นฐานมาจากโครงข่ายประสาทเทียมแบบวนซ้ำ (RNN) หรือโครงข่ายประสาทเทียมแบบสังวัตนาการ (CNN) ที่ซับซ้อน ซึ่งประกอบด้วยตัวเข้ารหัส (encoder) และตัวถอดรหัส (decoder)" ผ่านกระบวนกลไกความสนใจ (attention mechanism) โดยที่โมเดลจะสนใจ คำที่เกี่ยวข้องกับบริบท และคำสำคัญของประโยคแล้วทำการประมวลผล

โมเดล Transformer โดยภาพรวมประกอบด้วยการซ้อนกันของกลไก self-attention และเลเยอร์แ fully connected ที่ทำงานแบบจุดต่อจุด (point-wise) ทั้งในส่วนของ encoder และ decoder ดังรูป



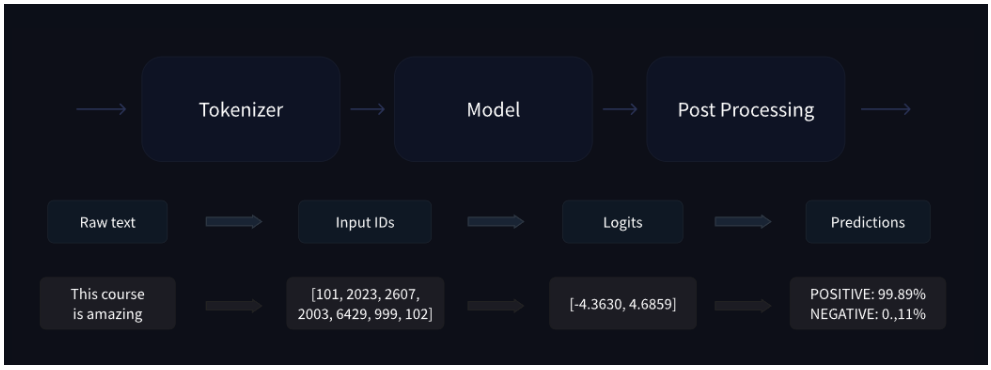
สถาปัตยกรรมของ Transformers

(Vaswani et al.,(2017))

https://arxiv.org/pdf/1706.03762

โครงสร้างสถาปัตยกรรม Transformer โดยรวม

1. Input Embedding เริ่มต้นด้วยกระบวนการจัดการข้อมูลและเตรียมข้อมูลดิบให้เหมาะสมกับการวิเคราะห์และประมวลผล โครงข่ายประสาทเทียมอื่น ๆและ โมเดล Transformer ไม่สามารถประมวลผลข้อมูลดิบได้โดยตรง ดังนั้นขั้นตอนแรกของกระบวนการคือการแปลงข้อความให้เป็นตัวเลขที่โมเดลสามารถเข้าใจได้ ดังรูปซึ่งมีกระบวนการ ดังรูป



การแปลงข้อความให้เป็นตัวเลขที่โมเดลสามารถเข้าใจ

(Hugging Face (2025))

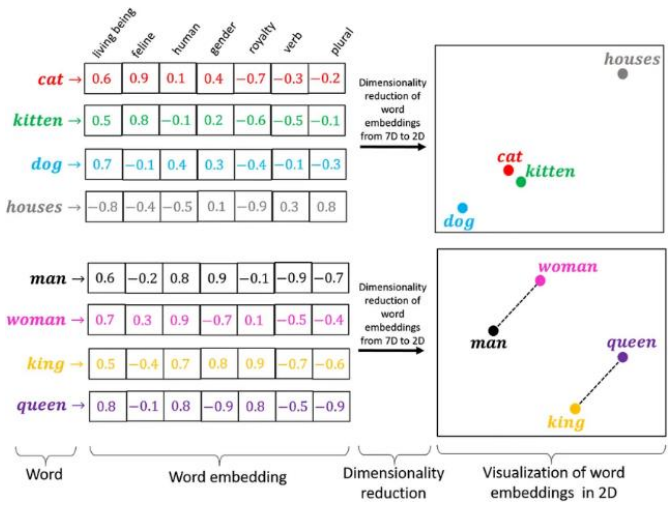
1.1. tokenization แยกข้อความออกเป็นคำ, คำย่อย (subwords) หรือสัญลักษณ์ (เช่น เครื่องหมายวรรคตอน) ซึ่งเรียกว่า token เช่น

"Don't you love 🤗 Transformers? We sure do."

["Don't", "you", "love", "🤗", "Transformers?", "We", "sure", "do."]

1.2 word embedding ในความเป็นจริง จะมีชุดข้อมูลพื้นฐาน จะประกอบด้วยอักขระ ASCII ทั้งหมดอย่างน้อยที่สุด และอาจรวมถึงอักขระ Unicode บางส่วนด้วย โดยแต่ละคำจะถูกระบุตำแหน่งแต่ละ token ให้เป็นตัวเลขจำนวนเต็ม (integer) ด้วยเลข byte ซึ่งเป็น การบีบอัดข้อมูล Byte Pair Encoding โดยจะทำการแทนค่าคู่ของ byte ที่พบมากที่สุดในลำดับ ด้วย byte ใหม่เพียงตัวเดียวที่ยังไม่เคยถูกใช้งานมาก่อน โดยทำซ้ำไปเรื่อย ๆ และผลลัพธ์ที่ได้ คือเวกเตอร์คำ (Word Vectors) หรือ word embeddings ซึ่งสามารถจับความสัมพันธ์ทางความหมายระหว่างคำต่าง ๆ ได้โดยอิงจากบริบท นั่นหมายความว่า หากมีคำที่มีความหมายใกล้เคียงกัน ตำแหน่งที่อยู่ของเวกเตอร์คำก็จะใกล้กันด้วยเช่นกัน ดังรูป

http://www.pennelynn.com/Documents/CUJ/HTML/94HTML/19940045.HTM



การระบุตำแหน่ง token

(Hugging Face (2025))

ตำแหน่งของเวกเตอร์คำ

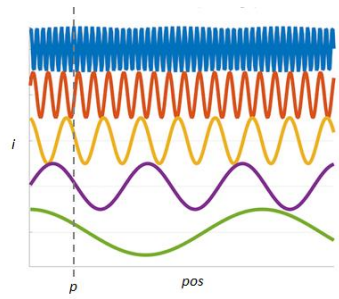
(Dipanjan (2018))

https://medium.com/data-science/understanding-feature-engineering-part-4-deep-learning-methods-for-text-data-96c44370bbfa

2. Positional Encoding เนื่องจากโมเดล transformer ไม่มีการใช้โครงสร้างแบบวนซ้ำ (recurrence) หรือคอนโวลูชัน (convolution) ดังนั้น เพื่อให้โมเดลสามารถเข้าใจลำดับของข้อมูล จึงจำเป็นต้องใส่ข้อมูลเกี่ยวกับ ตำแหน่งเข้ารหัส (positional encodings) เข้ากับเวกเตอร์คำ (embeddings) ที่ส่วนล่างของ encoder และ decoder โดยใช้ cosine functions ซึ่งก็คือ sin และ cos ที่ความถี่ต่างๆ กัน มาหาค่าในแต่ละตำแหน่ง

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

นอกจากนี้ การใช้ความถี่เพื่อทำ positional encoding เป็นการใช้ position vector ของตำแหน่ง p ตัวที่ j จะได้จากค่าของ periodic function ตัวที่ j ในตำแหน่ง p โดยช่วงความถี่สูงจะแยกแยะตำแหน่งที่อยู่ใกล้กัน และช่วงความถี่ต่ำจะแยกแยะตำแหน่งที่อยู่ไกล ซึ่งการทำ Positional Encoding แบบนี้ เป็นการใช้ข้อมูลเชิงความถี่มาประกอบกัน ดังรูป

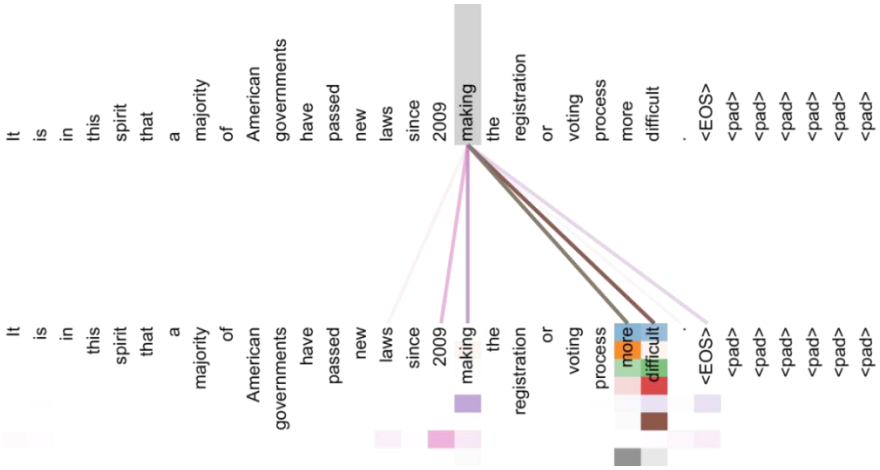


การใช้ความถี่เพื่อทำ positional encoding

(ppp (2018))

https://medium.com/@u41ppp/all-you-need-is-attention-%E0%B9%81%E0%B8%84%E0%B9%88%E0%B9%83%E0%B8%AA%E0%B9%88%E0%B9%83%E0%B8%88%E0%B8%81%E0%B8%B1%E0%B8%99%E0%B9%80%E0%B8%97%E0%B9%88%E0%B8%B2%E0%B8%99%E0%B8%B1%E0%B9%89%E0%B8%99%E0%B8%81%E0%B9%87%E0%B8%9E%E0%B8%AD-f473f112db12

3. Attention คือการเรียนรู้ของ attention layer ซึ่งสามารถอธิบายได้ว่าเป็นการจับคู่ระหว่าง Query กับชุดของ Key-Value pairs เพื่อให้ได้ผลลัพธ์ ซึ่งทั้ง Query, Key, Value และผลลัพธ์นั้นจะอยู่ในรูปเวกเตอร์ทั้งหมด โดยผลลัพธ์จะถูกคำนวณจากผลรวมถ่วงน้ำหนัก (weighted sum) แต่การคำนวณเวกเตอร์เพียงแค่ 1 attention head สามารถเรียนรู้ได้แค่ความสัมพันธ์รูปแบบเดียว ไม่เพียงพอในการเรียนรู้ภาพรวมของทั้งประโยค ซึ่งแก้ด้วยวิธีง่ายๆ คือ ทำ attention หลายๆรอบ โดยแต่ละรอบใช้ Wq, Wk และ Wv คนละตัวกัน แล้วค่อยนำมารวมกัน จึงเกิดเป็น Multi-Head Attention ดังรูป



self multi-head attention

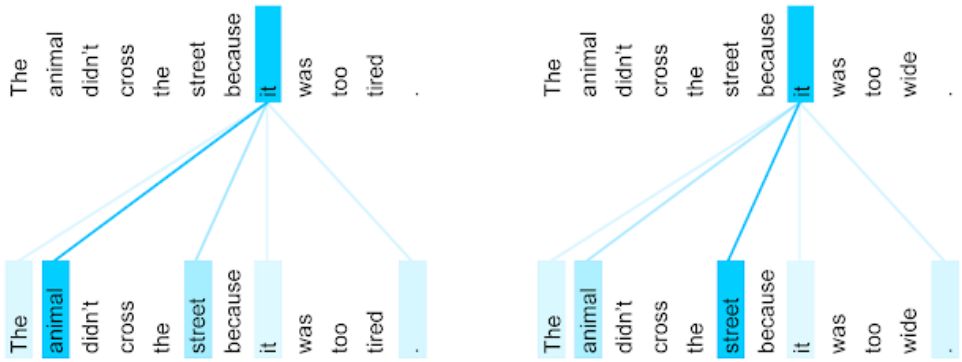
(Vaswani et al.,(2017))

Multi-head Attention ช่วยให้โมเดลสามารถโฟกัสข้อมูลจากหลาย ๆ มุมมองของเวกเตอร์ แทน (representation subspaces) ในตำแหน่งต่าง ๆ พร้อมกันได้ ในขณะที่หากใช้แค่ 1 attention head ไม่เหมาะสมในการประมวล

The animal didn't cross the street because it was too tired.
L'animal n'a pas traversé la rue parce qu'il était trop fatigué.

The animal didn't cross the street because it was too wide.
L'animal n'a pas traversé la rue parce qu'elle était trop large.

จากตัวอย่างประโยค it ในประโยคแรกหมายถึง สัตว์ และ it ในประโยคสองหมายถึง ถนน เมื่อนำประโยคแปลเป็นภาษาฝรั่งเศส คำแปลของ it จะขึ้นกับ เพศของคำนามที่ it อ้างถึง ซึ่งในภาษาฝรั่งเศส คำว่า สัตว์ และ ถนน มีเพศที่แตกต่างกันอย่างชัดเจน ซึ่งสามารถแปลเป็น “il” หรือ “elle” ก็ได้ ขึ้นอยู่กับว่า “it” ตัวนี้ใช้แทนคำนามเพศไหน ซึ่งสรุปได้ว่า จำนวน Attention ที่ให้แต่ละคำก็สะท้อนให้เห็นถึงการเลือกของโมเดลในบริบทที่แตกต่างกัน ดังรูป



encoder self-attention

(Jakob *(2017))*

https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

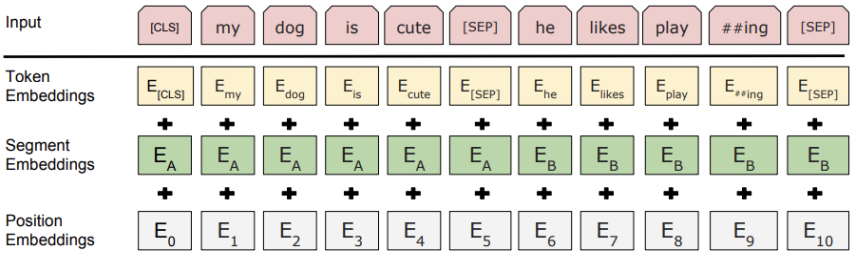
4. Feed Forward สังเกตุได้ว่า ในแต่ละชั้นของ attention จะมี encoder และ decoder ด้วยด้วยโครงข่ายฟีดฟอร์เวิร์ดแบบเชื่อมต่อเต็มรูปแบบ (Fully Connected Feed-Forward Network) ทำหน้าที่รับส่งข้อมูล ไปแต่ละ layer จนไปถึง output ทำให้โมเดลสามารถประมวลผลที่ซับซ้อนได้

5. Softmax ใช้ฟังก์ชัน Softmax เพื่อแปลงผลลัพธ์จาก Decoder ให้เป็นความน่าจะเป็นของ Token

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

https://arxiv.org/pdf/1810.04805

สถาปัตยกรรม BERT (Bidirectional Encoder Representations from Transformers) ถูกพัฒนาจากโมเดลที่ผ่านการเทรนมาแล้ว (Pretrained Model) ผ่านชุดข้อมูลขนาดใหญ่ ที่ไม่มีป้ายกำกับ (unlabeled data) ด้วยเทคนิค Masked Language Model (MLM) และ Next Sentence Prediction (NSP) โดยมีโครงสร้างข้อมูลนำเข้า สำหรับโทเคนหนึ่งตัว การสร้างเวกเตอร์แทนข้อมูลขาเข้าของมัน (input representation) จะทำโดยการบวกเวกเตอร์ต่อไปนี้เข้าด้วยกัน คือ เวกเตอร์แทนโทเคน (token embedding) เวกเตอร์แทนตำแหน่งของโทเคนในประโยค (position embedding) เวกเตอร์แทนว่าโทเคนนั้นอยู่ในประโยคใด (segment embedding) ดังรูป



ทำให้เมื่อเทรนเสร็จสิ้น โมเดล BERT สามารถเรียนรู้ และจัดการปัญหาได้หลากหลาย เช่น การตอบคำถาม การอนุมานทางภาษา โดยไม่จำเป็นต้องเปลี่ยนแปลงสถาปัตยกรรมสำหรับงานใด งานหนึ่ง ผ่านกระบวนการปรับแต่งโมเดลทาง (Fine-tuning) นอกจากนี้ BERT พิจารณา บริบท บริบททั้ง ก่อนหน้า (ซ้าย) และ ถัดไป (ขวา) ของคำที่กำลังดูอยู่ โดยใชช้โครงสร้างของ Transformer แบบ Encoder ที่สามารถอ่านและประมวลผลข้อมูล ทั้งสองทิศทางพร้อมกัน

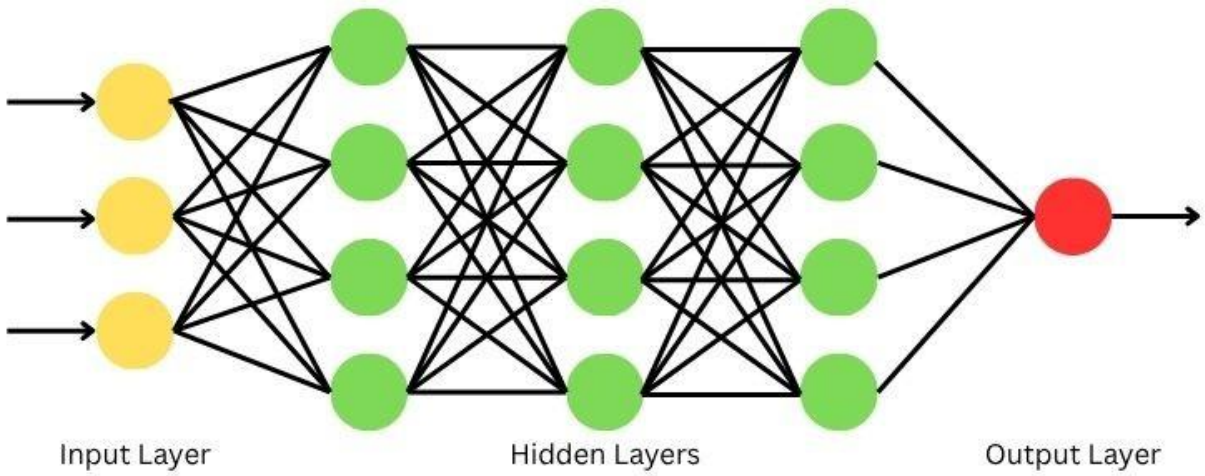
https://www.researchgate.net/publication/228340819_Multilayer_perceptron_and_neural_networks#fullTextFileContent

2.7 เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)

โครงข่ายประสาทเทียมแบบฟีดฟอร์เวิร์ด (Feedforward Neural Network) ที่ประกอบด้วยหลายชั้นของนิวรอน (neurons) ซึ่งอย่างน้อยต้องมี หนึ่งชั้นซ่อน (Hidden Layer) เพื่อให้สามารถเรียนรู้ฟังก์ชันที่ไม่เป็นเชิงเส้น (Nonlinear Functions) ได้ โดยโครงสร้างของ MLP ประกอบด้วย Input Layer – รับข้อมูลอินพุต

Hidden Layers – ประมวลผลข้อมูลด้วยการเชื่อมโยงแบบ Fully Connected และ Output Layer – ส่ง

การทำงานของ MLP ใช้ อัลกอริทึม Backpropagation ร่วมกับเทคนิค Gradient Descent เพื่อปรับปรุงน้ำหนักของการเชื่อมต่อภายในเครือข่าย โดยการ: ส่งข้อมูลผ่านชั้นต่าง ๆ เพื่อทำนายผลลัพธ์ (Forward Propagation) คำนวณค่าความผิดพลาดจากผลลัพธ์ที่คาดหวัง ส่งค่าความผิดพลาดย้อนกลับและปรับน้ำหนัก (Backward Propagation) โดยมีฟังก์ชันกระตุ้น (Activation Function) จำกัดช่วงของค่าให้อยู่ในกรอบที่เหมาะสม


Input Layer
Hidden Layers
Output Layer

2.6 งานวิจัยที่เกี่ยวข้อง

2.6.1 ระบบแนะนำวารสารวิชาการให้กับผู้เขียนบทความ โดยใช้ข้อมูลภาษาไทยและภาษาอังกฤษจาก บทความ

https://digital.car.chula.ac.th/cgi/viewcontent.cgi?article=6394&context=chulaetd

งานวิจัยของ นิธิรันดร์ นุ่มนนท์ ทำแบบทำแบบจำลองสำหรับระบบแนะนำ โดยจะทำแบบจำลองจากการคำนวณหาความสำคัญจากข้อความด้วยเทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (Term Frequency - Inverse Document Frequency:TF-IDF) และการวิเคราะห์ความคล้ายคลึงระหว่างบทความและวารสารโดยใช้ Cosine Similarity แล้วจึงจัดอันดับค่าคะแนนเพื่อแนะนำบทความที่เหมาะสม จากผลการทดลองในงานวิจัยนี้ การ ตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 10 พับ (10-fold cross-validation) พบว่าเมื่อเรานำข้อมูลคำสำคัญและบทคัดย่อจากทั้งภาษาไทยและภาษาอังกฤษมารวมกัน ระบบสามารถแนะนำ ออกมาได้ค่าความแม่นยำที่วัดด้วย Hit Rate ได้ค่าสูงสุดที่ 0.87965 ซึ่งมากกว่าแบบจำลองที่ใช้ ข้อมูลภาษาอังกฤษอย่างเดียว (0.84948) หรือ แบบจำลองที่ใช้ข้อมูลภาษาไทยอย่างเดียว (0.80383) และได้ค่าความแม่นยำที่สูงกว่าการตรวจสอบความสมเหตุสมผลแบบไขว้จำนวน 5 พับ และการทดลองแบบจำลองในลักษณะความคล้ายระหว่างบทความ

2.6.2 การเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มการเป็นโรคไตเรื้อรัง: กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย

https://li01.tci-thaijo.org/index.php/tstj/article/view/85101/67778

งานวิจัยของ สุรวัชร ศรีเปารยะ และสายชลสิน สมบูรณ์ทอง ทำการเปรียบเทียบประสิทธิภาพของวิธีการจำแนกกลุ่ม โดยเลือกใช้วิธีความใกล้เคียงกันมากที่สุดวิธีต้นไม้ตัดสินใจวิธีโครงข่ายประสาทเทียมวิธีซัพพอร์ตเวกเตอร์แมชชีน วิธีฐานกฎ วิธีการถดถอยลอจิสติก และวิธีนาอีฟเบย์ เพื่อวัดประสิทธิภาพการจำแนกกลุ่มโดยใช้ข้อมูลผู้ป่วยโรคไตเรื้อรังของโรงพยาบาลอพอลโลประเทศอินเดีย โดยแบ่งข้อมูลเป็นชุดสร้างตัวแบบ และชุดทดสอบตัวแบบ ในอัตราส่วน 70 และ 30 ตามลำดับ จากการเปรียบเทียบประสิทธิภาพวิธีการจำแนกกลุ่มผู้ป่วยโรคไตเรื้อรังโดยเปรียบเทียบจากค่าความถูกต้องและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยวิธีการจำแนกกลุ่มที่มีประสิทธิภาพการจำแนกดีที่สุด คือ วิธีต้นไม้ตัดสินใจ ซึ่งให้ค่าความถูกต้องคือ 100% และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยคือ 0.0059

2.6.3 News Classification and Its Techniques: A Review

https://www.iosrjournals.org/iosr-jce/papers/Vol18-issue1/Version-3/D018132226.pdf

งานวิจัยของ Gurmeet Kaur , Karan Bajaj ทำการจำแนกข่าวและเทคนิคที่ใช้ การขุดข้อมูล (Data Mining) ได้กลายเป็นเทคนิคที่มีความสำคัญอย่างยิ่ง เนื่องจากปริมาณข้อมูลที่ผู้ใช้สามารถเข้าถึงได้จากหลากหลายแหล่ง เช่น สื่ออิเล็กทรอนิกส์ สื่อดิจิทัล และช่องทางออนไลน์อื่นๆ ซึ่งข้อมูลเหล่านี้มักอยู่ในรูปแบบที่ไม่มีโครงสร้าง (Unstructured Data) จึงมีความจำเป็นต้องแปลงให้อยู่ในรูปแบบที่มีโครงสร้าง เพื่อให้สามารถนำไปประมวลผลและวิเคราะห์ได้อย่างมีประสิทธิภาพ

2.6.4 Newspaper Article Classification using Machine Learning Techniques

https://www.researchgate.net/publication/363921132_Newspaper_Article_Classification_using_Machine_Learning_Techniques

งานวิจัยของ J Sree Devi, M. Rama Bai, Chandrashekar Reddy ทำการการจำแนกบทความจากหนังสือพิมพ์ซึ่งมีเนื้อหาครอบคลุมหลากหลายหมวดหมู่ เช่น กีฬา การเมือง วิทยาศาสตร์ และเทคโนโลยี โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อจัดประเภทข้อมูลอย่างมีประสิทธิภาพ ในยุคที่ข้อมูลมีปริมาณมหาศาลและอยู่ในรูปแบบที่ไม่เป็นโครงสร้าง การจำแนกข้อความจึงเป็นสิ่งจำเป็นในชีวิตประจำวัน เพื่อให้สามารถจัดการและเข้าถึงข้อมูลได้อย่างรวดเร็วและแม่นยำ งานวิจัยนี้เปรียบเทียบอัลกอริทึมหลายประเภท เช่น Support Vector Machine (SVM), Naive Bayes, K-Nearest Neighbors (KNN) และ Convolutional Neural Networks (CNN) ซึ่งเป็นเทคนิคการเรียนรู้เชิงลึกที่ได้รับความนิยมและมีประสิทธิภาพสูงในการจำแนกข้อมูล

2.6.5 การศึกษาขั้วอารมณ์ (SENTIMENT ANALYSIS) ของการเปิดเผยด้าน ESG ของบริษัทที่ได้รับการคัดเลือกเป็นหลักทรัพย์ที่ถูกคำนวณ ในดัชนี SETTHSI เฉพาะกลุ่มทรัพยากร (RESOURCE)

https://ethesisarchive.library.tu.ac.th/thesis/2023/TU_2023_6502110502_18757_28390.pdf

งานวิจัยของ ธารทรายทอง หิรัญศรี ทำการวิเคราะห์ข้อความ (Textual Analysis) ของการเปิดเผยข้อมูลด้าน ESG ซึ่งเป็นวิธีการวิเคราะห์ที่นอกเหนือจากการวิเคราะห์ ตัวเลขทางการเงิน ผ่านการพัฒนาโมเดลภาษาที่มุ่งเน้นเฉพาะด้าน ESG ที่อยู่ภายในรายงานประจำปี จากนั้นจึงทำประเมินผลของประสิทธิภาพโมเดลที่มุ่งเน้นไปที่ด้าน ESG โดยเฉพาะเทียบกับโมเดลที่มุ่งเน้นโดยภาพรวม พบว่าโมเดลที่มุ่งเน้นเฉพาะด้าน ESG มีความแม่นยำในการทำนายที่สูงกว่า ขั้นถัดมาจึงได้นำโมเดลเฉพาะด้าน ESG นี้มาใช้ในการทำ Sentiment Analysis ของการเปิดเผยข้อความระหว่างปี 2563-2565 ของบริษัทในกลุ่มทรัพยากรจำนวน 27 บริษัท ผลการศึกษาพบว่า บริษัท

ส่วนใหญ่ให้น้ำหนักไปกับการเปิดเผย Aspect ด้าน ESG ไปที่ด้าน Social มากที่สุด และมีการ เปิดเผย Sentiment ของข้อความด้าน ESG ไปในขั้วที่เป็นกลาง (Neutral)

2.2.6 Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features

https://www.researchgate.net/publication/332526260_Thai_Sentiment_Analysis_via_Bidirectional_LSTM-CNN_Model_with_Embedding_Vectors_and_Sentic_Features

งานวิจัยของ Thititorn Seneewong Na Ayutthaya และ Kitsuchart Pasupa ทำการวิเคราะห์ความรู้สึกของคนไทยผ่านแบบจำลอง LSTM-CNN แบบสองทิศทางพร้อมเวกเตอร์ฝังตัวและคุณลักษณะเซนติก เมื่อเร็ว ๆ นี้ มีการพัฒนาแบบจำลอง Word2Vec ซึ่งเป็นเทคนิคการฝังคำ (การแปลงข้อความเป็นตัวเลข) และประสบความสำเร็จในระดับหนึ่งในการดึงความรู้สึกของลูกค้าจากข้อความตอบกลับที่พวกเขาให้มา งานวิจัยนี้พยายามรวมคุณสมบัติเพิ่มเติมอีกสองอย่างเข้าด้วยกัน ได้แก่ คุณสมบัติส่วนของคำพูดและคุณสมบัติประโยค เพื่อให้การวิเคราะห์มีความแม่นยำมากขึ้น คุณลักษณะส่วนของคำพูดจะระบุประเภทของคำที่สื่อถึงความรู้สึกต่างๆ ได้ดีกว่า ในขณะที่คุณลักษณะของความรู้สึกจะระบุอารมณ์ที่อยู่เบื้องหลังคำบางคำ เมื่อรวมหน่วยความจำระยะสั้นแบบทวิภาคีและโมเดลเครือข่ายประสาทเทียมแบบ Convolutional เข้ากับการผสมผสานคุณลักษณะต่างๆ ที่กล่าวถึงหลายๆ อย่างแล้ว เราได้ดำเนินการวิเคราะห์ความรู้สึกของเรื่องราวสำหรับเด็กไทย และพบว่าการผสมผสานคุณลักษณะทั้งสามอย่างให้ผลลัพธ์ดีที่สุดที่คะแนน F1 78.89%

บทที่ 4

ผลดำเนินการวิจัยและการวิเคราะห์ข้อมูล

Dataset Type : ชุดที่ 1 **Mixed Types :** No (ง่าย)

Number of Classes : 2 **รายละเอียด :** ประโยคตรงไปตรงมา ไม่มีคลาส neutral

Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.97	0.99	0.94	0.96
			pos		0.96	0.99	0.98
		SVM	neg	0.98	0.97	0.96	0.97
			pos		0.98	0.98	0.98
		NB	neg	0.89	0.87	0.87	0.87
			pos		0.91	0.91	0.91
Group 2	Thai2Vec	MLP	neg				
			pos				
Group 3	-	WangchanBERTa	neg	0.98	0.97	0.98	0.97
			pos		0.98	0.98	0.98

Dataset Type : ชุดที่ 2 **Mixed Types :** Yes (4 รูปแบบ)

Number of Classes : 2 **รายละเอียด :** ข้อมูลปนกันในพารากราฟ

Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.91	0.89	0.92	0.91
			pos		0.94	0.90	0.92
		SVM	neg	0.92	0.89	0.93	0.91
			pos		0.94	0.90	0.92
		NB	neg	0.84	0.79	0.89	0.84

			pos		0.90	0.80	0.85
Group 2	Thai2Vec	MLP	neg				
			pos				
Group 3	-	WangchanBERTa	neg	0.98	0.99	0.98	0.98
			pos		0.98	0.99	0.99

Dataset Type : ชุดที่ 3 **Mixed Types :** Yes (6 รูปแบบ)

Number of Classes : 2 **รายละเอียด :** ข้อมูลปนกันในพารากราฟ

Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.92	0.93	0.90	0.91
			pos		0.91	0.94	0.93
		SVM	neg	0.92	0.92	0.91	0.91
			pos		0.92	0.93	0.93
		NB	neg	0.87	0.83	0.91	0.87
			pos		0.92	0.84	0.88
Group 2	Thai2Vec	MLP	neg				
			pos				
Group 3	-	WangchanBERTa	neg	0.98	0.98	0.97	0.98
			pos		0.97	0.98	0.98

Dataset Type : ชุดที่ 4 **Mixed Types :** No

Number of Classes : 3 **รายละเอียด :** มี positive, negative, neutral

Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.90	0.99	0.89	0.94
			pos		0.83	0.93	0.88
			neu		0.92	0.87	0.90
		SVM	neg	0.91	0.98	0.91	0.95
			pos		0.85	0.94	0.89
			neu		0.93	0.88	0.90
		NB	neg	0.85	0.87	0.85	0.86
			pos		0.84	0.88	0.86
			neu		0.84	0.82	0.83
Group 2	Thai2Vec	MLP	neg				
			pos				
			neu				
Group 3	-	WangchanBERTa	neg	0.94	0.98	0.96	0.97
			pos		0.93	0.93	0.93
			neu		0.93	0.94	0.94

Dataset Type : ชุดที่ 5 **Mixed Types :** Yes (6 รูปแบบ)

Number of Classes : 3 **รายละเอียด :** ข้อมูลปนกันในพารากราฟ

Model Group	Feature Extract	Classifier	Class	Accuracy	Precision	Recall	F1
Group 1	TF-IDF	Logistic	neg	0.90	0.92	0.91	0.91
			pos		0.80	0.97	0.87
			neu		0.93	0.98	0.89

		SVM	neg	0.91	0.93	0.91	0.92
			pos		0.83	0.97	0.89
			neu		0.93	0.89	0.91
		NB	neg	0.82	0.79	0.90	0.84
			pos		0.78	0.92	0.84
			neu		0.89	0.72	0.80
Group 2	Thai2Vec	MLP	neg				
			pos				
			neu				
Group 3	-	WangchanBERTa	neg	0.96	0.98	0.98	0.98
			pos		0.94	0.90	0.92
			neu		0.96	0.97	0.97