การจำแนกอนุประโยคอัตวิสัยจากบทความภาษาไทยด้วยหน่วยความจำระยะสั้นยาวแบบสองทาง

Clause-Level Subjective Classification for Thai Article Using Bidirectional Long Short-Term Memory

ณัฐดนัย ศรีทิพากร[1*] และ ทรงศักดิ์ รองวิริยะพานิช[1]
Nutdanai Sritiparkorn[1] and Songsakdi Rongviriyapanish[1]
[1]สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
[1]Department of Computer Science, Faculty of Science and Technology, Thammasat University
วันที่ส่งบทความ : 29 พฤษภาคม 2566 วันที่แก้ไขบทความ : 18 มิถุนายน 2566 วันที่ตอบรับบทความ : 24 กรกฎาคม 2566
Received: 29 May 2023, Revised: 18 June 2023, Accepted: 24 July 2023

บทคัดย่อ

การจำแนกประโยคอัตวิสัยเป็นหนึ่งในขั้นตอนที่สำคัญในการวิเคราะห์ความคิดเห็นจากข้อมูลที่มาจากบทความหรือสื่อออนไลน์ ซึ่งมีปริมาณเพิ่มขึ้นเป็นอย่างมาก ความคิดเห็นที่สกัดได้จากประโยคอัตวิสัยสามารถนำมาใช้เป็นข้อมูลในการผลิตหรือปรับปรุงสินค้าให้ดีขึ้น งานวิจัยนี้ได้นำเสนอแนวทางการสร้างโมเดลจำแนกอัตวิสัยระดับอนุประโยคในบทความภาษาไทย โดยใช้โมเดลการเรียนรู้เชิงลึกที่ใช้หน่วยความจำระยะสั้นยาวแบบสองทาง (Bidirectional Long Short-Term Memory) ในการจำแนก ซึ่งเป็นวิธีที่นิยมใช้กันในการจัดการกับข้อมูลที่เป็นลำดับ และได้นำโมเดล FastText มาใช้ในการแปลงคำเป็นเวกเตอร์ตัวเลข งานวิจัยนี้ได้ทดลองสร้างโมเดลจากชุดข้อความในหลายโดเมน และวัดความถูกต้องในการจำแนก โดยใช้ชุดข้อมูล LST20 ซึ่งประกอบด้วยอนุประโยคจำนวน 44,423 อนุประโยคที่ได้ตัดคำไว้แล้ว รวมถึงมีข้อมูลชนิดคำ (Part of speech) และชื่อเฉพาะ (Named entity) ที่ถูกใช้เป็นลักษณะสำหรับการเรียนรู้ของโมเดล ในการวัดประสิทธิภาพของโมเดลได้ใช้การสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง 5 กลุ่ม (5-fold cross-validation) และพบว่าโมเดลที่ใช้จำนวนเซลล์ประสาทของหน่วยความจำระยะสั้นยาวแบบสองทาง 200 เซลล์ และใช้ลักษณะ (Feature) คำและชนิดคำ เป็นโมเดลที่ดีที่สุดให้ค่าความแม่นยำ (Precision) 62.562% ค่าความระลึก (Recall) 51.151% ค่าความถูกต้อง (Accuracy score) 79.407% และค่า F1-score 56.284%

คำสำคัญ : การจำแนกประโยคอัตวิสัย หน่วยความจำระยะสั้นยาวแบบสองทาง หน่วยความจำระยะสั้นยาว การเรียนรู้เชิงลึก การประมวลภาษาธรรมชาติ

[*]ที่อยู่ติดต่อ E-mail address: nutdanai_srit@dome.tu.ac.th

Abstract

Sentence subjective classification is one of the crucial steps in analyzing opinions from such data as articles or online media which the volume has increased greatly. Extracted opinions from sentences can be used as information to produce or improve products. This research presented a method to create a model for classifying opinion at the clause level in Thai language articles using a Bidirectional Long Short-Term Memory (BiLSTM) deep learning model. This model is widely used to deal with sequential data. Moreover, the FastText model was used to convert words into numerical vectors. Our research experimented by creating models from texts in multi-domain and measuring the accuracy of the classification using the LST20 dataset. This dataset contains 44,423 pre-segmented clauses, including Part of Speech and Named Entity annotations, which are used as features for model learning. The evaluation of model performance used 5-fold cross-validation. We found that the BiLSTM model using 200 neurons in the Long Short-Term Memory unit with word and Part of Speech as features is the best model. It achieved precision of 62.562%, recall of 51.151%, accuracy score of 79.407%, and F1-score of 56.284%.

Keywords: Subjective classification, Bidirectional Long Short-Term Memory, Long Short-Term Memory, Deep learning, Natural Language Processing

1. บทนำ

ธุรกิจในปัจจุบันผู้ผลิตมักจะสนใจความคิดเห็นของผู้บริโภคที่มีต่อสินค้า เช่น ข้อดี ข้อเสีย คำแนะนำ และข้อเสนอแนะ ปัจจุบันการแสดงความคิดเห็นของผู้บริโภคไม่จำเป็นต้องอยู่ในช่องความคิดเห็นอย่างเฟซบุ๊ก ยูทูบ อินสตาแกรม หรือในเว็บรีวิวที่สะท้อนความคิดเห็นจากผู้อ่านเท่านั้น อาจจะอยู่ในรูปแบบข่าวหรือบทความที่สามารถแสดงความคิดเห็นของผู้เขียนที่มีต่อสินค้าได้เช่นกันโดยมีทั้งข้อเท็จจริง และข้อคิดเห็นถึงสินค้าปะปนกันอยู่ มนุษย์ไม่สามารถนำข้อมูลเหล่านั้นมาวิเคราะห์ด้วยตัวเองได้ทั้งหมด จำเป็นต้องใช้เครื่องมือสำหรับช่วยวิเคราะห์ข้อมูล การจำแนกอัตวิสัยหรือการจำแนกความคิดเห็น (Subjective classification) ของบทความแบบอัตโนมัติให้มีความถูกต้องแม่นยำจึงเป็นที่มาของการทำวิจัยนี้ โดยต้องการสร้างโมเดลที่ช่วยจำแนกประโยคที่แสดงความคิดเห็นจากบทความภาษาไทย

แม้ว่าที่ผ่านมามีงานวิจัยที่เกี่ยวกับการวิเคราะห์ความรู้สึก (Sentiment analysis) ในข้อความภาษาไทยอยู่หลายงานแต่ไม่ได้เกี่ยวข้องกับการวิเคราะห์อัตวิสัยโดยตรง เพราะอัตวิสัยคือการแสดงถึงข้อคิดเห็น ความรู้สึก การคาดคะเน การเปรียบเทียบหรืออุปมาอุปไมย ข้อเสนอแนะหรือความคิดของผู้สื่อสาร [1]-[2] ดังนั้นข้อความปรวิสัยหรือข้อความข้อเท็จจริงจึงสามารถนำมาวิเคราะห์ความรู้สึกได้เช่นกัน อย่างไรก็ตามสิ่งที่ผู้ใช้สนใจคือในข้อความอัตวิสัยมีผู้กล่าวถึงความรู้สึกแง่บวก หรือแง่ลบของสินค้ามากน้อย

เพียงใด และมีข้อเสนอแนะอย่างไรบ้าง [3] แม้แต่ข่าวหรือบทความก็สามารถแสดงความคิดเห็นได้เช่นกัน ฉะนั้นการจำแนกข้อความอัตวิสัยถือว่าเป็นขั้นตอนหนึ่งที่สำคัญของการวิเคราะห์ความรู้สึก

การวิเคราะห์ความคิดเห็นในข่าวหรือบทความมักจะเขียนในรูปแบบประโยคความซ้อนซึ่งประกอบด้วยหลายอนุประโยคเชื่อมกัน การวิเคราะห์อัตวิสัยของอนุประโยคส่งผลต่ออัตวิสัยของประโยค ดังนั้นการจำแนกอนุประโยคอัตวิสัยในบทความภาษาไทยจึงเป็นเรื่องสำคัญ

งานวิจัยที่มีอยู่ในปัจจุบันได้นำเสนอวิธีการจำแนกเอกสารที่ให้ประสิทธิภาพและความถูกต้องสูง โดยสร้างโมเดลการจำแนกเอกสารด้วยวิธีการเรียนรู้เชิงลึก (Deep learning) ซึ่งเป็นส่วนหนึ่งของการเรียนรู้ของเครื่อง (Machine Learning) โมเดลการเรียนรู้เชิงลึกที่ใช้ในการจำแนกในงานวิจัยนี้ คือ หน่วยความจำระยะสั้นยาว (Long Short-Term Memory: LSTM) เนื่องจากวิธีการทำงานของหน่วยความจำระยะสั้นยาวมีลักษณะเหมือนการทำงานของมนุษย์ที่ไม่ได้จำเนื้อหาทั้งหมดแต่จะจำส่วนที่สำคัญของเนื้อหาเท่านั้น ทำให้ไม่เปลืองพื้นที่ความจำมากเกินไป เหมาะกับการวิเคราะห์ข้อมูลที่มีลำดับ (Sequential data) จึงเหมาะกับการวิเคราะห์ประเภทข้อความ (Text analysis) และจากผลการวิจัยพบว่าหน่วยความจำระยะสั้นยาวยังให้ผลลัพธ์เปอร์เซ็นต์ความถูกต้องสูงด้วย [4]-[6] นอกจากนี้มีงานวิจัยที่ใช้หน่วยความจำระยะสั้นยาวแบบสองทางในการจำแนกเอกสาร ซึ่งให้ผลลัพธ์ความถูกต้องสูงเช่นกัน [7]

การเรียนรู้ด้วยวิธีการเรียนรู้เชิงลึกต้องเป็นข้อมูลตัวเลขที่สามารถคำนวณได้ (Numeric data) แต่ข้อมูลที่ต้องการวิเคราะห์เป็นข้อความที่ไม่สามารถคำนวณได้ หนึ่งในวิธีการเปลี่ยนข้อความเป็นตัวเลขคือ Word embedding ซึ่งเป็นโมเดลที่เปลี่ยนคำ วลี หรือประโยคให้แสดงออกมาในรูปของชุดตัวเลขหรือเวกเตอร์ตามความสัมพันธ์ระหว่างคำ วลี หรือประโยค ซึ่งมีหลายงานวิจัยใช้ในการจำแนกข้อความ (Text classification) นอกจากนี้ยังมีโมเดลที่สามารถเรียนรู้จากชุดข้อมูลของผู้ใช้เองได้ เช่น FastText ซึ่งให้ประสิทธิภาพการจำแนกข้อความที่ดีสำหรับการเรียนรู้ของเครื่อง [8] และจากงานวิจัยของ Krungklang และ Sinthupinyo [5] ได้เปรียบเทียบประสิทธิภาพการจำแนกข้อความด้วย Word embedding หลายแบบ ซึ่งจากผลการทดสอบพบว่า FastText ให้ผลการจำแนกข้อความได้ดีที่สุด

ถึงแม้ว่าการจำแนกอัตวิสัยกับการจำแนกความรู้สึกไม่ได้เกี่ยวข้องกันโดยตรง แต่สองอย่างนี้เป็นปัญหาการจำแนกข้อความเหมือนกัน จึงสามารถใช้วิธีการวิจัยที่คล้ายคลึงกันได้ งานวิจัยต่อไปนี้เกี่ยวข้องกับการจำแนกเอกสารและใช้การเรียนรู้เชิงลึกในการแก้ปัญหา

Hajj และคณะ [9] ได้จำแนกอัตวิสัยในบทความกีฬาภาษาอังกฤษโดยใช้คอร์ติคอลอัลกอริทึม (Cortical algorithm) ลักษณะที่นำมาจำแนกคือลักษณะของคำที่ปรากฏในบทความ เช่น จำนวนคำที่เป็นตัวเลข จำนวนคำต่างประเทศ จำนวนคำกริยาวิเศษณ์ จำนวนสัญลักษณ์ และจำนวนคำสันธาน ซึ่งรวมแล้วมีทั้งหมด 52 ลักษณะ หลังจากนั้นนำลักษณะเหล่านี้มาลดจำนวนลงด้วยคอร์ติคอลอัลกอริทึมซึ่งเป็นอัลกอริทึมที่วิวัฒนาการจากโครงข่ายประสาทเทียม (Artificial Neural Networks) เมื่อได้ลดจำนวนลักษณะลงแล้วใช้ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และซัพพอร์ตเวกเตอร์แมชชีนหลายชั้น (Hierarchical Support Vector Machine) ในการจำแนกอัตวิสัย ผลลัพธ์ที่ได้จากการจำแนกให้ความถูกต้องถึง 85.60% เมื่อใช้การสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง 4 กลุ่มในการทดสอบ

Ayutthaya และ Pasupa [4] ได้วิจัยการจำแนกความรู้สึกผ่านหน่วยความจำระยะสั้นยาวแบบสองทางผสมกับโครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neuron Network: CNN) ซึ่งใช้ Word embedding และค่าอารมณ์ (Sentic) เป็นลักษณะในการฝึกข้อมูล Word embedding ที่ใช้คือ Thai2Vec ค่าอารมณ์ต่อหนึ่งคำจะมีข้อมูลเชิงอารมณ์อยู่ 4 ส่วน ได้แก่ ความพึงพอใจ (Pleasantness) ความสนใจ (Attention) ความอ่อนไหว (Sensitivity) และความฉลาด (Aptitude) งานวิจัยดังกล่าวจะวิเคราะห์ถึงลักษณะของข้อมูลนำเข้าว่าลักษณะใดบ้างที่ให้ผลลัพธ์ถูกต้องที่สุด ผลลัพธ์คือใช้ Word embedding ชนิดคำและค่าอารมณ์ ให้ผลลัพธ์ความถูกต้องถึง 78.89%

Krungklang และ Sinthupinyo [5] ได้วิจัยเรื่องข้อสรุปกฎหมายอาญาโดยใช้หน่วยความจำระยะสั้นยาว และการเลือกใช้ Word embedding แบบฝึกข้อมูลก่อนทำงานจริง (Pre-trained word embedding) ในการจำแนก เป้าหมายของงานวิจัยคือการจำแนกกลุ่มในหัวข้อต่าง ๆ ของกฎหมายอาญา เช่น 1) การแสดงเจตนา แบ่งกลุ่มได้เป็นกรณีทั่วไป กรณีพิเศษ หรือไม่เจตนา 2) ผลการกระทำ แบ่งเป็นกลุ่มได้แก่ ถึงแก่ความตาย ได้รับบาดเจ็บ และไม่มีผล 3) การละเว้นโทษ แบ่งเป็นกลุ่มได้แก่ ปกติ ละเว้นเพิ่มขึ้น และลดลง และ 4) การตัดสินโทษ แบ่งเป็นกลุ่มคือมีความผิด และไม่มีความผิด ผลลัพธ์คือหน่วยความจำระยะสั้นยาวแบบสองทางและ Word embedding แบบ FastText ให้ผลลัพธ์การจำแนกถูกต้องมากที่สุด ได้ค่าความถูกต้อง 82.08%, 76.91%, 78.46% และ 98.24% ตามลำดับ

Xu และคณะ [7] ได้วิจัยการจำแนกความคิดเห็นภาษาจีนโดยใช้ทีเอฟไอดีเอฟ (Term Frequency-Inverse Document Frequency: TF-IDF) เป็นลักษณะ และใช้หน่วยความจำระยะสั้นยาวแบบสองทางมาปรับลักษณะให้ดีขึ้น ขั้นตอนคือนำชุดคำที่แปลงเป็น Word embedding เข้ากระบวนการหน่วยความจำระยะสั้นยาวแบบสองทางและนำผลลัพธ์ไปใช้ในโครงข่ายประสาทเทียมแบบป้อนข้อมูลไปข้างหน้า (Feedforward Neural Network) เพื่อจำแนกความคิดเห็น ผลลัพธ์การจำแนกคือได้ F1-score สูงถึง 92.20%

Pugsee และ Ongsirimongkol [10] ได้วิจัยการจำแนกความคิดเห็นโดยใช้โมเดลการเรียนรู้เชิงลึกด้วยโครงข่ายประสาทเทียมแบบคอนโวลูชัน และหน่วยความจำระยะสั้นยาวผสมกัน งานวิจัยนี้เน้นไปที่การประเมินประสิทธิภาพของโมเดลการเรียนรู้เชิงลึกโดยใช้โมเดลโครงข่ายประสาทเทียมแบบคอนโวลูชัน และหน่วยความจำระยะสั้นยาวก่อนใช้โครงข่ายประสาทเทียมแบบเชื่อมโยงสมบูรณ์ (Fully-connected) การจำแนกความคิดเห็นแบ่งออกเป็น 3 กลุ่มคือแง่บวก แง่ลบ และแง่กลาง ใช้ข้อมูลนำเข้าเป็นข้อความและใช้ DeepCut ในการตัดคำ จากนั้นเปลี่ยนคำเป็นตัวเลขลำดับที่อยู่ในชุดฝึกข้อมูล ใช้การสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง 4 กลุ่มในการประเมินความถูกต้อง ลำดับการใช้โมเดลหลายชั้นที่ดีที่สุดคือแบบ CNN(32,2) CNN(64,2) CNN(128,2) และ LSTM(12) ตามลำดับ โดย CNN(32,2) หมายถึง โครงข่ายประสาทเทียมแบบคอนโวลูชันที่ใช้ฟิลเตอร์ 32 ตัว และใช้ขนาดเคอร์เนล (Kernel) เป็น 2 และ LSTM(12) หมายถึงใช้หน่วยความจำระยะสั้นยาวที่มีเซลล์หน่วยความจำ (Memory cell) 12 เซลล์ ใช้ข้อมูลการฝึกทั้งหมด 8,546 ข้อมูล ได้ค่าความถูกต้องจากข้อมูลการประเมิน (Validation data) และข้อมูลทดสอบ (Testing data) เท่ากับ 84.50% และ 84.55% ตามลำดับ เมื่อเทียบตามกลุ่มแล้วพบว่าการจำแนกความคิดเห็นในกลุ่มแง่บวกให้ผลลัพธ์ดีที่สุด โดย F1-score เป็น 89% สำหรับข้อมูลทั้งการประเมินและการทดสอบ

เป้าหมายของงานวิจัยในบทความนี้คือการทดสอบประสิทธิภาพโมเดลการจำแนกอนุประโยคอัตวิสัยภาษาไทยที่สร้างจากหน่วยความจำระยะสั้นยาวแบบสองทางและใช้ Word embedding ด้วย FastText และต้องการทดสอบว่าโมเดลที่สร้างจากชุดข้อมูลหลายโดเมน (Multi-domain dataset) สามารถจำแนกอนุประโยคอัตวิสัยได้มากน้อยเพียงใด เนื่องจากปริมาณชุดข้อความของแต่ละโดเมนมีจำนวนไม่มาก หากสามารถนำชุดข้อมูลหลายโดเมนมาใช้สร้างโมเดลการจำแนกอนุประโยคอัตวิสัยที่มีความถูกต้องสูงได้ถือเป็นเรื่องที่ดี ในงานวิจัยนี้ผู้วิจัยได้ใช้ข่าวบทความภาษาไทยจากชุดข้อมูล LST20 ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ หรือเนคเทค (National Electronics and Computer Technology Center: NECTEC) ในการสร้างโมเดลการจำแนกอนุประโยคอัตวิสัย บทความประเภทข่าวสามารถแสดงความคิดเห็นถึงเนื้อหาข่าวจากผู้เขียน ผู้สัมภาษณ์ข่าวหรือผู้ที่ถูกกล่าวถึงในข่าวได้เช่นกัน นอกจากนี้ชุดข้อมูล LST20 มีข้อมูลการตัดคำและอนุประโยคเรียบร้อยแล้ว

โดยบทความนี้ประกอบด้วยหัวข้อต่อไปนี้ หัวข้อที่ 2 วิธีการทดลอง หัวข้อที่ 3 ผลการทดลองและวิจารณ์ และหัวข้อที่ 4 สรุปผลการทดลอง

2. วิธีการทดลอง

2.1 ชุดข้อมูล

ทดสอบประสิทธิภาพการจำแนกอนุประโยคอัตวิสัยด้วยหน่วยความจำระยะสั้นยาวกับชุดข้อมูล LST20 ของ NECTEC ซึ่งเป็นชุดข้อมูลที่รวบรวมข่าว 15 หมวด มีทั้งหมด 4,751 ข่าว ที่ได้ตัดคำ (Word segmentation) ติดชนิดคำ (Part-of-Speech tagging) แบ่งช่วงคำชื่อเฉพาะ (Named entity recognition) แบ่งคำวลี (Clause segmentation) และแบ่งประโยค (Sentence segmentation) ประเภทข่าวที่มีมากที่สุดตามลำดับ ได้แก่ ข่าวการเมือง ข่าวอาชญากรรมและอุบัติเหตุ และข่าวเศรษฐกิจ

เนื่องจากชุดข้อมูลของ LST20 ไม่ได้ระบุข้อมูลอนุประโยคอัตวิสัยจึงต้องติดคำตอบว่าเป็นอนุประโยคอัตวิสัยหรือไม่ โดยใช้ผู้ตรวจสอบ 3 คนเป็นคนตรวจสอบคลาส (Class labeling) และใช้เสียงส่วนใหญ่ในการตัดสินอนุประโยคอัตวิสัย จากขั้นตอนนี้จึงได้จำนวนอนุประโยคที่ติดคำตอบจำนวน 44,423 อนุประโยค (ข้อมูลที่ใช้ในการทดสอบไม่ใช่อนุประโยคทั้งหมดของ LST20)

2.2 ขั้นตอนการวิจัย

ขั้นตอนการทำวิจัยมีทั้งหมด 4 ขั้นตอน ได้แก่ 1) การทำ Word embedding 2) การเพิ่มลักษณะ (Add features) 3) การสร้างโมเดลหน่วยความจำระยะสั้นยาวแบบสองทาง และ 4) การจำแนกอนุประโยคอัตวิสัย ขั้นตอนการทำวิจัยแสดงเป็นแผนผังดังรูปที่ 1

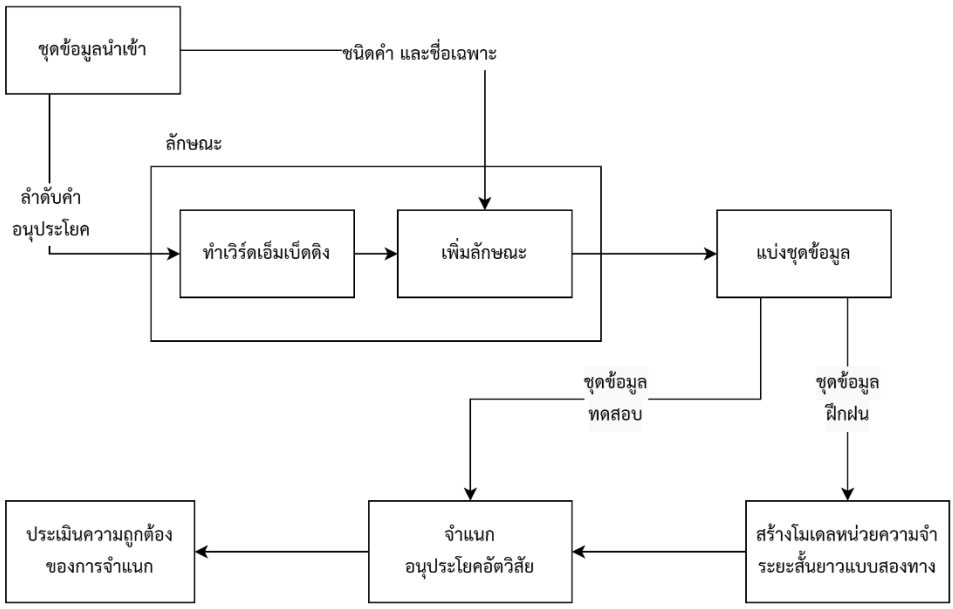
ในการวิจัยนี้ผู้วิจัยคาดว่าลักษณะที่นำมาทดลองนั้นมีความชัดเจนในการจำแนกแม้ว่าจะอยู่คนละโดเมนก็ตาม จึงต้องการเปรียบเทียบประสิทธิภาพโมเดลระหว่างโดเมนรวมและโดเมนเฉพาะ โดยโดเมนรวมหมายถึงโดเมนข่าวทั้ง 15 หมวด ข้อมูลโดยรวมของอนุประโยคที่นำไปสร้างโมเดลการเรียนรู้แสดงได้ดังตารางที่ 1

ในการวิจัยโดเมนเฉพาะในตารางที่ 1 จะใช้ข้อมูลเฉพาะหมวดข่าวการเมือง และข่าวอาชญากรรมและอุบัติเหตุเนื่องจากมีจำนวนอนุประโยค 18,058 และ 10,833 ประโยคซึ่งมีปริมาณที่มากพอที่จะใช้ใน

การทดสอบประสิทธิภาพของหน่วยความจำระยะสั้นยาวได้ ส่วนโดเมนรวมคือการนำอนุประโยคจากโดเมนทั้งหมดที่ได้ติดคำตอบซึ่งมีทั้งหมด 44,423 ประโยค

ข้อมูลที่ต้องเตรียมก่อนการทดสอบได้แก่ ชุดลำดับคำ (Word sequence) ชนิดคำ และชื่อเฉพาะของอนุประโยค ขั้นตอนการทำงานมีรายละเอียดต่อไปนี้

ขั้นตอนที่ 1 การทำ Word embedding เริ่มแรกนำลำดับคำในชุดข้อมูล LST20 ที่มีการตัดคำไว้แล้วเข้าสู่กระบวนการแปลงคำเป็นเวกเตอร์ ซึ่งการแปลงคำเป็นเวกเตอร์คือการแปลงประเภทคำให้เป็นชุดตัวเลขในรูปของเวกเตอร์ เพราะข้อมูลนำเข้าของหน่วยความจำระยะสั้นยาวต้องเป็นข้อมูลตัวเลขที่คำนวณได้ การทำ Word embedding เป็นการแปลงคำเป็นเวกเตอร์ตัวเลขที่แสดงถึงบริบทของประโยคจากคำนี้



รูปที่ 1. ขั้นตอนการวิจัยการจำแนกอนุประโยคอัตวิสัย

ตารางที่ 1. ข้อมูลโดยรวมของอนุประโยคอัตวิสัยสำหรับสร้างโมเดลการเรียนรู้

โดเมนข่าว	คำ	อนุประโยค	คำชื่อเฉพาะ	คำไม่ซ้ำ	อนุประโยคที่เป็นอัตวิสัย	อนุประโยคที่ไม่เป็นอัตวิสัย
การเมือง	226,580	18,058	18,887	7,603	5,818	12,240
อาชญากรรมและอุบัติเหตุ	154,527	10,833	14,227	7,330	1,994	8,839
เศรษฐกิจ	37,102	2,705	2,761	3,042	659	2,046
ต่างประเทศ	27,732	2,089	2,708	2,904	486	1,603
สุขภาพ	19,824	1,441	1,355	2,163	281	1,160
กีฬา	22,061	1,426	2,314	2,476	450	976
ทั่วไป	17,374	1,261	1,389	2,197	326	935
วัฒนธรรม	16,513	1,202	1,438	2,206	274	928

ตารางที่ 1. (ต่อ) ข้อมูลโดยรวมของอนุประโยคอัตวิสัยสำหรับสร้างโมเดลการเรียนรู้

โดเมนข่าว	คำ	อนุประโยค	คำชื่อเฉพาะ	คำไม่ซ้ำ	อนุประโยคที่เป็นอัตวิสัย	อนุประโยคที่ไม่เป็นอัตวิสัย
ภัยพิบัติ	17,194	1,168	1,525	1,894	236	932
การศึกษา	14,362	1,052	1,103	1,979	281	771
พระราชสำนัก	13,728	904	1,477	2,202	202	702
บันเทิง	11,244	821	849	1,760	191	630
สิ่งแวดล้อม	7,516	568	523	1,199	159	409
พยากรณ์อากาศ	9,130	563	882	1,278	97	466
การพัฒนา	5,375	332	565	1,009	59	273
ทั้งหมด	600,262	44,423	52,003	16,869**	11,513	32,910

หมายเหตุ : รวมคำทุกโดเมนและนับคำไม่ซ้ำ

โมเดลที่จะนำมาใช้คือ FastText ของเฟซบุ๊กเนื่องจากสามารถฝึกฝนกับข้อมูลใหม่ได้ งานวิจัยนี้ให้ FastText สร้างโมเดล Word embedding โดยให้เรียนรู้แบบมีผู้สอน (Supervised learning) กำหนดอัตราการเรียนรู้ Learning rate เท่ากับ 0.5 จำนวนรอบในการทำงาน (Epoch) 100 รอบ และขนาดเวกเตอร์ผลลัพธ์เป็น 300 เมื่อได้โมเดล Word embedding แล้วจะแปลงลำดับคำที่เป็นข้อมูลนำเข้าให้เป็นเวกเตอร์ตัวเลข 300 ตัวแทนลักษณะข้อมูลต่อ 1 คำ ในการทดลองถ้าคำที่นำมาฝึกไม่อยู่ในชุดแปลงคำ ระบบจะกำหนดเวกเตอร์ของคำนั้นเป็นค่า 0 ทั้ง 300 ตัว

ขั้นตอนที่ 2 การเพิ่มลักษณะ การวิจัยครั้งนี้ต้องการเปรียบเทียบข้อมูลนำเข้าว่าลักษณะใดบ้างจะให้ผลลัพธ์ออกมาดีที่สุด ข้อมูลนำเข้าที่นำมาเปรียบเทียบมี 4 แบบ ได้แก่ 1) คำ 2) คำและชนิดคำ 3) คำและชื่อเฉพาะ 4) คำ ชนิดคำ และชื่อเฉพาะ เนื่องจากคำได้เปลี่ยนเป็นค่าตัวเลขจากการทำ Word embedding แล้ว ข้อมูลที่เหลือต้องแปลงเป็นข้อมูลตัวเลขด้วยเช่นกัน ข้อมูลชนิดคำเป็นข้อมูลประเภทหมวดหมู่ที่ไม่สามารถนำมาคำนวณได้ จึงเปลี่ยนเป็นเวกเตอร์ตัวเลขโดยให้ขนาดของเวกเตอร์คือจำนวนชนิดคำที่เป็นไปได้ทั้งหมดซึ่งมีทั้งหมด 15 ชนิด ถ้าคำนี้อยู่ในชนิดคำที่ p ให้เวกเตอร์มิติที่ p เป็นค่า 1 ที่เหลือเป็นค่า 0 ในชุดข้อมูล LST20 มีชนิดคำอยู่ประเภทหนึ่งเป็นชนิดคำที่ไม่สามารถแยกหมวดหมู่จึงกำหนดค่าเป็น 0 ทั้งหมด ส่วนกรณีข้อมูลชื่อเฉพาะสามารถทำแบบเดียวกับข้อมูลชนิดคำ ในข้อมูลชื่อเฉพาะในชุดข้อมูล LST20 แบ่งข้อมูลเป็น 2 ส่วน ได้แก่ ประเภทของชื่อเฉพาะซึ่งมีทั้งหมด 10 ประเภท และช่วงของชื่อเฉพาะ
ที่ประกอบไปด้วยช่วงเริ่มต้นคำ ช่วงระหว่าง และช่วงท้ายคำ ระบบจะใช้ข้อมูลชื่อเฉพาะที่เป็นประเภทของชื่อเฉพาะเท่านั้นโดยตัดส่วนที่เป็นช่วงชื่อเฉพาะออกไป ลักษณะที่เพิ่มขึ้นมาจะถูกนำมาเป็นส่วนหนึ่งของเวกเตอร์ของคำสรุปคือ ขนาดเวกเตอร์ต่อ 1 คำของข้อมูลนำเข้าแต่ละแบบเป็น 300, 315, 310 และ 325 ตามลำดับ เมื่ออิงตัวอย่างการเปลี่ยนลักษณะของข้อมูลให้เป็นตัวเลขดังตารางที่ 2

ขั้นตอนที่ 3 การสร้างโมเดลหน่วยความจำระยะสั้นยาวแบบสองทาง ขั้นตอนนี้สร้างโมเดลเพื่อนำข้อมูลที่ได้มาตอนเรียนรู้และส่งผลลัพธ์ให้กับขั้นตอนถัดไป ข้อมูลนำเข้าสำหรับสร้างโมเดลเป็นลำดับเวกเตอร์คำที่ได้จากขั้นตอนที่ 1 และขั้นตอนที่ 2 โดยกำหนดจำนวนคำ 100 คำ ถ้าอนุประโยคใดมีลำดับคำไม่ถึง 100 คำ จะกำหนดเวกเตอร์คำที่เหลือเป็นเวกเตอร์ศูนย์ และถ้าอนุประโยคใดมีลำดับคำเกิน 100 คำ

ตารางที่ 2. ตัวอย่างการเปลี่ยนข้อมูลนำเข้าตามลักษณะข้อมูล

ลักษณะของข้อมูล	ตัวอย่างข้อมูล	คำอธิบาย	ขนาดเวกเตอร์
คำ (Word embedding)	[-0.09883000701665878, 0.16018599271774292, ..., -0.03424237295985222, 0.007027868181467056]	ข้อมูล Word embedding แทนคำว่า “นาย”	300
ชนิดคำ	[0,0,0,0,0,0,0,0,1,0,0,0,0,0,0]	ข้อมูลแทนชนิดคำ ประเภทคำนาม	15
ชื่อเฉพาะ	[0,0,1,0,0,0,0,0,0,0]	ข้อมูลแทนชื่อเฉพาะ ประเภทชื่อบุคคล	10

จะใช้เฉพาะเวกเตอร์คำ 100 คำแรกเท่านั้น จากข้อมูลนำเข้าในการสร้างโมเดลที่แสดงดังรูปที่ 2 ข้อมูล 300 แถวแรกคือลักษณะของคำ อีก 15 แถวถัดมาคือลักษณะของชนิดคำ และ 10 แถวสุดท้ายคือลักษณะของชื่อเฉพาะ การสร้างโมเดลต้องกำหนดพารามิเตอร์ต่อไปนี้ คือ จำนวนเซลล์ประสาท (Neurons) อัตราการเรียนรู้ขนาดแบทช์ (Batch size) และจำนวนรอบการทำงานงานวิจัยนี้ตั้งพารามิเตอร์จำนวนเซลล์ประสาทที่ 50, 100 และ 200 เซลล์ อัตราการเรียนรู้เท่ากับ 0.01 ขนาดแบทช์เท่ากับ 256 และจำนวนรอบการทำงาน 30 รอบ เมื่อนำข้อมูลเข้าโมเดลนี้แล้วจะได้ผลลัพธ์เป็นเวกเตอร์ตัวเลขตามจำนวนเซลล์ประสาทที่กำหนดไว้

ขั้นตอนที่ 4 การจำแนกอนุประโยคอัตวิสัย เมื่อได้ผลลัพธ์จากขั้นตอนที่ 3 จะนำมาจำแนกด้วยโครงข่ายประสาทเทียมแบบเชื่อมโยงสมบูรณ์โดยไม่มีฟังก์ชันกระตุ้น ผลลัพธ์สุดท้ายจะได้ค่าตัวเลขแทนคลาสคำตอบแต่ละคลาส วิธีจำแนกอนุประโยคอัตวิสัยคือค่าในคลาสใดมากที่สุดคือคำตอบของอนุประโยคนั้น

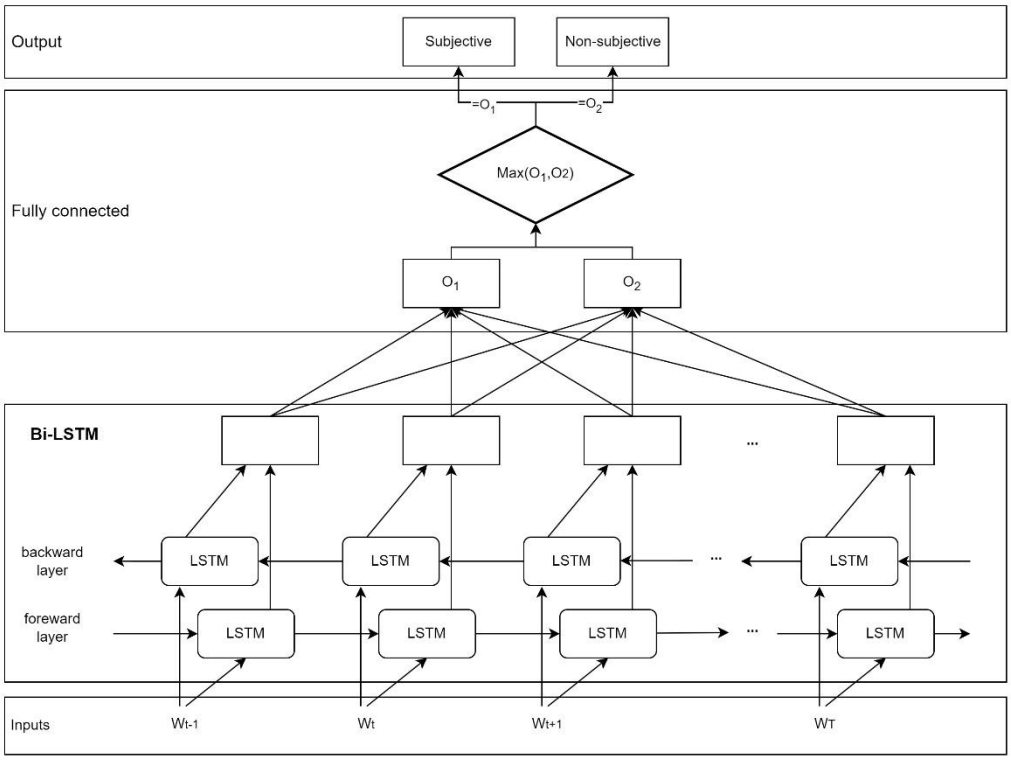
กระบวนการตั้งแต่การสร้างโมเดลหน่วยความจำระยะสั้นยาวแบบสองทางจนถึงผลการจำแนกได้แสดงดังรูปที่ 3 กำหนดให้ W คือข้อมูลนำเข้า และ W_t คือข้อมูลนำเข้าตัวที่ t

	w_1	w_2		w_i	w_{i+1}		w_{100}
	เชื่อ	ต้องการ	...	ภาคใต้	-	...	-
เวิร์ดเอ็มเบ็ดดิง (ขนาด 300)	0.144856	0.016414	...	0.022936	0	...	0
	-0.15811	-0.01621	...	-0.01979	0	...	0
	...	...	...	...	...	...	...
	-0.01551	6.92E-06	...	0.000642	0	...	0
ชนิดคำ	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	1	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	1	1	...	0	0	...	0
คำเฉพาะ	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	1	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0
	0	0	...	0	0	...	0

รูปที่ 2. ข้อมูลนำเข้าสำหรับสร้างโมเดลหน่วยความจำระยะสั้นยาวแบบสองทาง

2.3 วิธีการประเมิน

ในการประเมินประสิทธิภาพของโมเดลได้ใช้การสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง k กลุ่ม แบ่งชุดข้อมูลออกเป็น k ส่วนเท่ากันและให้ชุดข้อมูล 1 ส่วนเป็นข้อมูลทดสอบและ k-1 ส่วนที่เหลือเป็นข้อมูลฝึกฝน และทำกระบวนการนี้ k รอบ โดยแต่ละรอบใช้ข้อมูลทดสอบคนละส่วน เช่น รอบแรกใช้ข้อมูลก้อนที่



รูปที่ 3. โมเดลการจำแนกอนุประโยคอัตวิสัย

1 เป็นข้อมูลทดสอบ รอบที่ 2 ใช้ข้อมูลก้อนที่ 2 เป็นข้อมูลทดสอบ วนทำกระบวนการนี้จนถึงรอบที่ k งานวิจัยนี้กำหนด k เท่ากับ 5 กรณีโดเมนรวมระบบจะแบ่งชุดข้อมูลให้แต่ละชุดมีอัตราส่วนของข่าวแต่ละประเภทเท่ากันและสุ่มข้อมูลให้ข้อมูลจากข่าวเดียวกันกระจายตามชุดข้อมูลอื่น ในการทดสอบแต่ละรอบข้อมูลสำหรับการสร้าง Word embedding มาจากชุดข้อมูลฝึกฝนเท่านั้น

ในการวัดผลจะนำผลลัพธ์ของแต่ละรอบของการสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง k กลุ่มมาคำนวณค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง ฟังก์ชันค่าเสียหาย (Loss function) F1-score และเวลาในการประมวลผล ฟังก์ชันค่าเสียหายได้ใช้สูตรของ Binary Cross Entropy ที่เหมาะกับงานจำแนกสองคลาส (Binary classification) ฟังก์ชันค่าเสียหายยิ่งน้อยความถูกต้องยิ่งมาก สามารถคำนวณด้วยดังสมการที่ (1) โดยให้ N คือจำนวนอนุประโยคในชุดข้อมูลทดสอบ เวกเตอร์ y คือเวกเตอร์ของคำตอบหรือคลาสที่ถูกต้องขนาด N มิติ และเวกเตอร์ p คือเวกเตอร์ของคำตอบหรือคลาสที่ทำนายขนาด N มิติ y_i และ p_i แทนคลาสคำตอบที่ถูกต้องและคำตอบที่ทำนายของอนุประโยคที่ i ค่าของ y_i และ p_i สามารถเป็นได้ 2 ค่าคือ 0 หรือ 1 เมื่อ 0 แทนคลาสไม่ใช่อนุประโยคอัตวิสัย และ 1 แทนคลาสอนุประโยคอัตวิสัย

$$Loss = \frac{1}{N}\sum_{i=0}^{N} -\big(y_i * log(p_i) + (1 - y_i) * log(1 - p_i)\big) \quad (1)$$

ค่าความแม่นยำ ค่าความระลึก ค่าความถูกต้อง และ F1-score มาจากตาราง Confusion Matrix ของการสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรงทั้ง 5 รอบมารวมกันและคำนวณตามตัวชี้วัดแต่ละตัว ส่วนฟังก์ชันค่าเสียหายและเวลาการประการมวลผลของการทดลองเป็นค่าเฉลี่ยจากการสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรงทั้ง 5 รอบ

3. ผลการทดลองและวิจารณ์

ผลการทดลองแบ่งออกเป็น 6 ส่วน คือ 1) ค่าความแม่นยำ 2) ค่าความระลึก 3) ค่าความถูกต้อง 4) F1-score 5) ค่าฟังก์ชันค่าเสียหาย และ 6) เวลาการประมวลผลของการจำแนกอนุประโยคอัตวิสัยประเภทความคิดเห็น และข้อเสนอแนะโดยใช้การสุ่มเลือกแบ่งข้อมูลแบบความเที่ยงตรง 5 กลุ่ม

ผลการทดลองดังตารางที่ 3 ถึงตารางที่ 5 แสดงหน่วยวัดและเวลาประมวลผลของการจำแนกอนุประโยคอัตวิสัยตามข้อมูลนำเข้าแต่ละแบบโดยใช้จำนวนเซลล์ประสาท 50, 100 และ 200 เซลล์ตามลำดับ

ตารางที่ 3. การวัดประสิทธิภาพด้วยตัวชี้วัดต่าง ๆ และเวลาประมวลผล เมื่อใช้จำนวนเซลล์ประสาท 50 เซลล์

โดเมน	ข้อมูล	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)	Loss	เวลาเฉลี่ย (นาที)
รวม	คำ	62.436	45.722	78.804	52.788	3.270	49
	คำ + ชนิดคำ	62.605	46.721	78.959	53.509	3.246	53
	คำ + ชื่อเฉพาะ	60.577	48.354	78.459	53.780	3.323	53
	คำ + ชนิดคำ + ชื่อเฉพาะ	62.815	47.260	79.081	53.938	3.227	55
การเมือง	คำ	74.543	65.933	81.770	69.974	2.812	24
	คำ + ชนิดคำ	73.178	65.418	81.133	69.081	2.910	26
	คำ + ชื่อเฉพาะ	74.641	64.249	81.449	69.056	2.862	26
	คำ + ชนิดคำ + ชื่อเฉพาะ	73.546	64.988	81.188	69.003	2.902	27
อญอ.	คำ	65.329	58.776	86.670	61.880	2.056	15
	คำ + ชนิดคำ	63.578	59.880	86.301	61.674	2.113	16
	คำ + ชื่อเฉพาะ	64.975	59.077	86.606	61.886	2.066	15
	คำ + ชนิดคำ + ชื่อเฉพาะ	61.581	61.334	85.840	61.457	2.184	16

หมายเหตุ : อญอ. คือ อาชญากรรมและอุบัติเหตุ

ตารางที่ 4. การวัดประสิทธิภาพด้วยตัวชี้วัดต่าง ๆ และเวลาประมวลผล เมื่อใช้จำนวนเซลล์ประสาท 100 เซลล์

โดเมน	ข้อมูล	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)	Loss	เวลาเฉลี่ย (นาที)
รวม	คำ	59.547	47.512	78.032	52.853	3.389	62
	คำ + ชนิดคำ	62.385	50.117	79.240	55.582	3.202	65
	คำ + ชื่อเฉพาะ	59.544	43.759	77.719	50.446	3.437	66
	คำ + ชนิดคำ + ชื่อเฉพาะ	62.037	48.571	78.968	54.484	3.244	67
การเมือง	คำ	74.452	64.163	81.360	68.925	2.875	28
	คำ + ชนิดคำ	73.431	64.936	81.133	68.923	2.910	29
	คำ + ชื่อเฉพาะ	73.723	65.246	81.310	69.226	2.883	30
	คำ + ชนิดคำ + ชื่อเฉพาะ	72.536	65.916	80.978	69.068	2.934	32
อญอ.	คำ	64.616	59.529	86.550	61.968	2.075	17
	คำ + ชนิดคำ	66.606	55.216	86.661	60.378	2.058	18
	คำ + ชื่อเฉพาะ	63.981	59.328	86.366	61.566	2.103	18
	คำ + ชนิดคำ + ชื่อเฉพาะ	62.132	61.384	86.006	61.756	2.159	18

หมายเหตุ : อญอ. คือ อาชญากรรมและอุบัติเหตุ

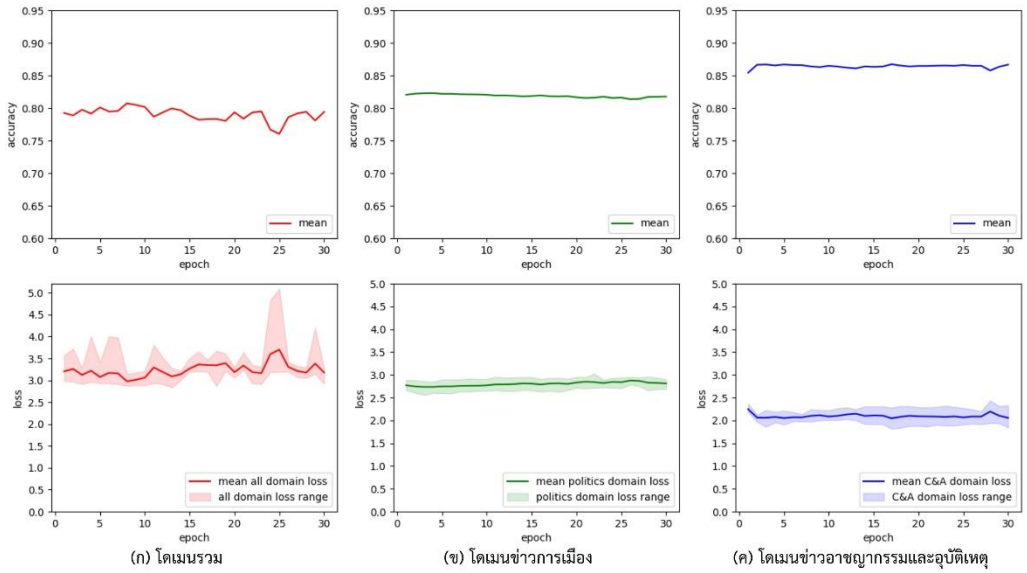
ตารางที่ 5. การวัดประสิทธิภาพด้วยตัวชี้วัดต่าง ๆ และเวลาประมวลผล เมื่อใช้จำนวนเซลล์ประสาท 200 เซลล์

โดเมน	ข้อมูล	Precision (%)	Recall (%)	Accuracy (%)	F1-score (%)	Loss	เวลาเฉลี่ย (นาที)
รวม	คำ	57.742	45.800	77.266	51.083	3.507	96
	คำ + ชนิดคำ	62.562	51.151	79.407	56.284	3.176	101
	คำ + ชื่อเฉพาะ	55.409	46.400	76.431	50.506	3.636	101
	คำ + ชนิดคำ + ชื่อเฉพาะ	63.046	47.242	79.150	54.012	3.216	105
การเมือง	คำ	69.950	67.257	80.142	68.577	3.063	40
	คำ + ชนิดคำ	72.705	67.257	81.316	69.875	2.882	43
	คำ + ชื่อเฉพาะ	71.654	68.735	81.166	70.164	2.905	41
	คำ + ชนิดคำ + ชื่อเฉพาะ	72.397	67.033	81.144	69.612	2.909	46
อญอ.	คำ	57.024	62.086	84.409	59.448	2.405	24
	คำ + ชนิดคำ	59.148	64.042	85.240	61.498	2.277	26
	คำ + ชื่อเฉพาะ	61.997	61.033	85.941	61.511	2.169	25
	คำ + ชนิดคำ + ชื่อเฉพาะ	58.826	65.848	85.230	62.139	2.278	26

หมายเหตุ : อญอ. คือ อาชญากรรมและอุบัติเหตุ

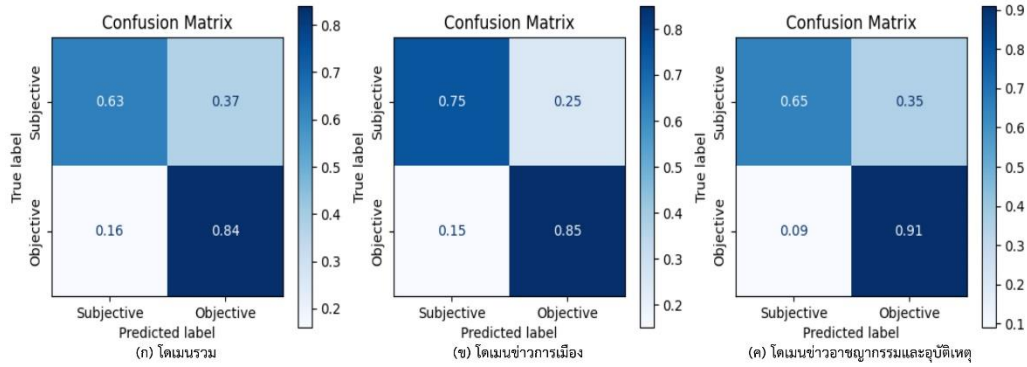
รูปที่ 4 แสดงกราฟค่าความถูกต้องและฟังก์ชันค่าเสียหายของโมเดลที่ให้ค่าความถูกต้องมากที่สุดในโดเมนรวม โดเมนข่าวการเมือง และโดเมนข่าวอาชญากรรมและอุบัติเหตุตามลำดับ และรูปที่ 5 แสดง

ตาราง Confusion Matrix ของโมเดลที่ให้ค่าความถูกต้องมากที่สุดในโดเมนรวม โดเมนข่าวการเมือง และโดเมนข่าวอาชญากรรมและอุบัติเหตุตามลำดับ



(ก) โดเมนรวม (ข) โดเมนข่าวการเมือง (ค) โดเมนข่าวอาชญากรรมและอุบัติเหตุ

รูปที่ 4. ค่าความถูกต้องและฟังก์ชันค่าเสียหายของโมเดลที่ได้ค่าความถูกต้องมากที่สุดในแต่ละโดเมน



(ก) โดเมนรวม (ข) โดเมนข่าวการเมือง (ค) โดเมนข่าวอาชญากรรมและอุบัติเหตุ

รูปที่ 5. Confusion Matrix ของโมเดลที่ได้ค่าความถูกต้องมากที่สุดในแต่ละโดเมน

จากผลการทดลองในโดเมนรวม โมเดลที่จำแนกความคิดเห็นและข้อเสนอแนะได้ถูกต้องมากที่สุด เมื่อใช้ลักษณะเป็นคำและชนิดคำ และใช้เซลล์ประสาท 200 เซลล์ ให้เปอร์เซ็นต์ความถูกต้องที่ 79.407% ส่วนโดเมนข่าวการเมือง และโดเมนข่าวอาชญากรรมและอุบัติเหตุ โมเดลที่จำแนกความคิดเห็นและข้อเสนอแนะได้ถูกต้องมากที่สุด เมื่อใช้ลักษณะเป็นคำเท่านั้น และใช้เซลล์ประสาท 50 เซลล์ ให้เปอร์เซ็นต์ความถูกต้องที่ 81.770% และ 86.670% ตามลำดับ

เมื่อเทียบกับลักษณะของข้อมูลนำเข้าโดยให้คำเป็นลักษณะหลัก การเพิ่มลักษณะไม่ว่าจะเป็นชนิดคำหรือชื่อเฉพาะไม่ได้ช่วยทำให้ประสิทธิภาพดีขึ้นอย่างมีนัยยะ จากรูปที่ 4 กราฟค่าความถูกต้องของโมเดลที่มีค่าความถูกต้องสูงที่สุดทั้งโดเมนรวม โดเมนข่าวการเมือง และข่าวอาชญากรรมและอุบัติเหตุค่อนข้างเป็นแนวเส้นตรงตามแกน x ซึ่งหมายถึงเมื่อจำนวนรอบการทำงานมากขึ้นค่าความถูกต้องไม่ได้เปลี่ยนแปลงไป สาเหตุอาจมาจากลักษณะที่เพิ่มเข้ามาในงานวิจัยไม่เกี่ยวข้องกับการจำแนกอนุประโยคอัตวิสัย

สำหรับในเรื่องของจำนวนเซลล์ประสาท ผู้วิจัยคาดว่าจำนวนที่มากขึ้นจะให้ค่าความถูกต้องมากขึ้น แต่จากผลการทดลองพบว่าจำนวนเซลล์ประสาทไม่มีผลแต่อย่างใด ในชุดข้อมูลบางชุดและลักษณะของข้อมูลนำเข้าบางลักษณะเมื่อเพิ่มเซลล์ประสาทมากขึ้นค่าความถูกต้องมีทั้งที่มากขึ้นและลดลง และจากผลการทดลองของแต่ละชุดข้อมูลและลักษณะข้อมูลนำเข้าเดียวกันให้ค่าความถูกต้องที่แทบไม่ต่างกัน เมื่อเปรียบเทียบจำนวนเซลล์ประสาทที่ต่างกันซึ่งให้ค่าความถูกต้องต่างกันประมาณ 1%

จากการพิจารณาเวลาประมวลผล เมื่อใช้ข้อมูลที่น้อยที่สุดในการทดลองโดยใช้ลักษณะข้อมูลนำเข้าเป็นคำเท่านั้นและใช้จำนวนเซลล์ประสาท 50 เซลล์ใช้เวลา 49 นาที สำหรับโดเมนรวมที่มีข้อมูล 44,423 ข้อมูล และเมื่อใช้ข้อมูลที่มากที่สุดในการทดลองโดยใช้ลักษณะข้อมูลนำเข้าเป็นคำ ชนิดคำ และชื่อเฉพาะและใช้จำนวนเซลล์ประสาท 200 เซลล์ ใช้เวลา 105 นาทีสำหรับโดเมนรวม

ค่าความแม่นยำในการจำแนกอนุประโยคอัตวิสัยที่แสดงในตารางที่ 5 โมเดลของโดเมนรวมทำนายถูก 63% โดเมนข่าวการเมือง 75% และโดเมนข่าวอาชญากรรมและอุบัติเหตุ 65%

ปัจจัยที่คาดว่าส่งผลต่อประสิทธิภาพมีดังนี้ ในข้อมูลทดสอบคำที่ได้จากการตัดคำของ LST20 มีบางคำที่ถูกตัดเป็นคำย่อยจนการวิเคราะห์ระดับคำให้ความหมายผิดไปจากเดิม เช่น “ความเป็นไปได้” (คำวิเศษณ์) ตัดคำออกมาเป็น “ความ” (คำนาม) “เป็น” (คำกริยา) “ไป” (คำกริยา) และ “ได้” (คำกริยา) ด้วยเหตุนี้อาจเป็นสาเหตุที่ทำให้โมเดลจำแนกผิด และมีบางคำที่ใช้รูปคำเหมือนกันแต่ให้คนละความหมายกันแม้ว่าจะเป็นคำชนิดเดียวกัน เช่น คำว่า “เห็น” สามารถมีความหมายได้ทั้งการมองเห็นด้วยตาและแสดงความเห็นได้ โดยทั้งสองความหมายนี้ไม่มีความคล้ายคลึงกัน ส่วน F1-score ที่ได้ประมาณ 60% อาจเนื่องจากข้อมูลที่นำมาสร้างโมเดลเป็นชุดข้อมูลที่ไม่สมดุล (Imbalanced data) โดยมีจำนวนอนุประโยคอัตวิสัยที่น้อยกว่าเมื่อเทียบกับจำนวนที่ไม่ใช่อนุประโยคอัตวิสัย

4. สรุปผลการทดลอง

งานวิจัยนี้ได้สร้างระบบการจำแนกอนุประโยคอัตวิสัยภาษาไทยโดยใช้ Word embedding ของ FastText และหน่วยความจำระยะสั้นยาวแบบสองทางและใช้ลักษณะต่าง ๆ จากผลทดสอบพบว่าโมเดลการจำแนกอนุประโยคอัตวิสัยสำหรับชุดข้อมูลหลายโดเมนที่ให้ผลลัพธ์ดีที่สุดคือ โมเดลที่ใช้จำนวนเซลล์ประสาท 200 เซลล์ และใช้ลักษณะเป็นลักษณะคำและชนิดคำ ซึ่งให้ค่าความถูกต้องในการจำแนก 79.407% สิ่งที่ควรปรับปรุงในงานวิจัยนี้คือควรจัดทำชุดข้อมูลให้มากขึ้นให้มีถึงระดับแสนหรือล้านประโยค เนื่องจากการเรียนรู้เชิงลึกต้องอาศัยข้อมูลจำนวนมาก ในขั้นตอนการตัดคำควรรวมคำให้มีความหมายตรงกับบริบทของประโยคก่อนนำมาเข้าโมเดล สำหรับการพัฒนาในงานวิจัยในอนาคตจะใช้เทคนิคของ Bidirectional Encoder Representations from Transformers (BERT) เข้าไปใส่ในโมเดลด้วย เนื่องจากเป็นเทคนิคที่ทันสมัยและมีความสามารถมากกว่าหน่วยความจำระยะสั้นยาว และอาจมีการปรับชุดข้อมูลให้

เป็นข้อมูลที่สมดุล (Balanced data) เพื่อให้การเรียนรู้ข้อมูลของแต่ละคลาสมีความเท่าเทียมกัน ถ้าได้ผลลัพธ์ที่ดีแล้วสามารถนำไปต่อยอดในการจำแนกอัตวิสัยระดับประโยคได้

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับทุนสนับสนุนจาก คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ และขอขอบคุณ ดร.เทพชัย ทรัพย์นิธิ ที่ปรึกษางานวิจัยร่วมช่วยแนะนำออกแบบวิธีการวิจัย และให้คำปรึกษาการเลือกลักษณะเพื่อสร้างโมเดลการเรียนรู้เชิงลึก และ ดร.ปรัชญา บุญขวัญ ที่ช่วยให้คำปรึกษาวางแนวคิด ขอบเขตงานวิจัย แนะนำการพัฒนาโมเดลการเรียนรู้ และการวิเคราะห์ข้อมูล

เอกสารอ้างอิง (References)

[1] Regmi, S., Bal, B.K. and Kultsova, M. 2017. Analyzing facts and opinions in Nepali subjective texts. 2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA), Larnaca, Cyprus, 1-4.

[2] นงคราญ เจริญพงษ์. 2555. การแยกข้อเท็จจริง ข้อคิดเห็น. แหล่งข้อมูล : https://kunkrunongkran.wordpress.com/ภาษาไทย-ม-2/ภาษาไทย-ม-2-เทอม-2/การแยกข้อเท็จจริง-ข้อคิ/. ค้นเมื่อวันที่ 23 พฤษภาคม 2563.

[3] Liu, B. 2010. Sentiment analysis and subjectivity. Handbook of natural language processing. 2nd Edition, Chapman and Hall/CRC, New York.

[4] Ayutthaya, T.S.N. and Pasupa, K. 2018. Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features. 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 1-6.

[5] Krungklang, W. and Sinthupinyo, S. 2020. An Analysis of Natural Language Text Relating to Thai Criminal Law. 2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), Bucharest, Romania, 1-6.

[6] Zhang, Y. and Rao, Z. 2020. n-BiLSTM: BiLSTM with n-gram Features for Text Classification. 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 1056-1059.

[7] Xu, G., Meng, Y., Qiu, X., Yu, Z. and Wu, X. 2019. Sentiment Analysis of Comment Texts Based on BiLSTM. *IEEE Access,* 7, 51522-51532.

[8] Yao, T., Zhai, Z. and Gao, B. 2020. Text Classification Model Based on fastText. 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Dalian, China, 154-157.

[9] Hajj, N., Rizk, Y. and Awad, M. 2019. A subjectivity classification framework for sports articles using improved cortical algorithms. *Neural Computing and Applications,* 11(31), 8069-8085.

[10] Pugsee, P. and Ongsirimongkol, N. 2020. A Classification Model for Thai Statement Sentiments by Deep Learning Techniques. Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems, 22-27.