การคัดแยกกลุ่มหัวข้อบทความวิชาการ AUCC ด้วย ซัพพอร์ตเวกเตอร์แมชชีน

AUCC Research Topic Categorization Based on Support Vector Machine

วชิรวิทย์ จันทร์เฟย* และ กฤตกรณ์ ศรีวันนา

สาขาวิศวกรรมคอมพิวเตอร์ คณะเทคโนโลยีดิจิทัล มหาวิทยาลัยราชภัฏเชียงราย

Emails: 651998026@crru.ac.th*, kittakorn.sri@gmail.com

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาการ คัดแยกกลุ่มหัวข้อบทความวิชาการของการประชุมวิชาการระดับปริญญาตรีด้านคอมพิวเตอร์ภูมิภาคเอเชีย The Asia Undergraduate Conference on Computing (AUCC) โดยใช้เทคนิคการเรียนรู้ของเครื่องเพื่อช่วยจำแนกหัวข้อบทความโดยอัตโนมัติ ลดความผิดพลาดที่เกิดขึ้นระหว่างการเลือกหมวดหมู่ของบทความ ข้อมูลบทความภาษาไทยจาก AUCC ถูกนำมาจัดเตรียม และประมวลผลด้วยขั้นตอน NLP ก่อนนำไปสร้างโมเดลจำแนกประเภท โดยใช้โมเดลสามรูปแบบ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (SVM), โครงข่ายประสาทเทียม (ANN) และ WangchanBERTa. ผลการประเมินด้วยค่า Accuracy และ F1-score พบว่า SVM ให้ผลลัพธ์สูงที่สุดเมื่อเทียบกับ ANN และ WangchanBERTa ซึ่งให้ค่าประสิทธิภาพรองลงมาตามลำดับ ผลลัพธ์นี้ชี้ให้เห็นว่า SVM มีความเหมาะสมที่สุดสำหรับงานคัดแยกกลุ่มหัวข้อบทความของ AUCC และเป็นโมเดลที่เสนอให้ใช้ในระบบจริง

คำสำคัญ: การจำแนกประเภทข้อความ, ซัพพอร์ตเวกเตอร์แมชชีน,การเรียนรู้ของเครื่อง, การประมวลผลภาษาธรรมชาติ

ABSTRACT

This study aims to develop a system for classifying academic paper topics submitted to the Asia Undergraduate Conference on Computing (AUCC). Machine learning techniques are applied to automatically categorize article topics and reduce errors that commonly occur during manual topic selection. Thai-language articles from AUCC were collected, preprocessed using NLP techniques, and used to build three classification models: SVM, ANN, and WangchanBERTa. Based on the evaluation using Accuracy and F1-score, the SVM model achieved the highest performance, followed by ANN and WangchanBERTa. These results indicate that SVM is the most suitable and effective model for AUCC's topic classification task and is therefore recommended for use in the actual system.

Keywords: text classification, support vector machine, machine learning, natural language processing

1. บทนำ

ในปัจจุบัน การเผยแพร่บทความวิชาการถือเป็นขั้นตอนสำคัญในการแลกเปลี่ยนองค์ความรู้ระหว่างนักวิจัยและสังคมวิชาการ งานประชุมวิชาการ AUCC มีผู้ส่งบทความเฉลี่ย 500-700 บทความต่อปี ซึ่งจากการวิเคราะห์ข้อมูลย้อนหลัง 3 ปี พบว่า

ร้อยละ 15-20 ของบทความถูกเลือก Track (กลุ่มหัวข้องานวิจัย) ไม่ตรงกับเนื้อหาที่แท้จริง ส่งผลให้เกิดปัญหาสำคัญคือบทความอาจถูกปฏิเสธ (Reject) หรือได้รับการประเมินจากผู้ทรงคุณวุฒิที่ไม่เหมาะสมกับสาขาวิชานั้นๆ และต้องใช้เวลาในการพิจารณาเพื่อจัดหมวดหมู่ใหม่ซ้ำเฉลี่ย 3-5 วันต่อบทความ กระบวนการส่งบทความในระบบปัจจุบันมีหลายขั้นตอน ได้แก่ 1) การลงทะเบียน 2) การจัดเตรียมต้นฉบับ 3) การส่งชื่อบทความ 4) การเลือกรูปแบบนำเสนอ และ 5) การเลือก Track ซึ่งเป็นขั้นตอนที่มีความเสี่ยงต่อความผิดพลาดสูงสุด ตัวอย่างปัญหาที่พบบ่อยคือความกำกวมของเนื้อหา เช่น บทความเรื่อง "Machine Learning for Education" ที่ผู้ส่งมักสับสนระหว่าง Track *Computer Education* และ *Data Science* หรือบทความ "IoT Security System" ที่คาบเกี่ยวระหว่าง *Information Technology* และ *Computer Business*

เพื่อแก้ไขปัญหาการเลือก Track ไม่ตรงกับบทความ ผู้วิจัยได้สำรวจงานวิจัยก่อนหน้าและแบ่งออกเป็น 3 กลุ่มหลักคือ 1) Support Vector Machine(SVM) ดังบทความของ ปกรณ์ สันตกิจ, พงษ์พร พันธ์เพ็ง, ปรีชา โพธิ์แพง, และ เยาวลักษณ์ งามแสนโรจน์ ได้เสนอระบบการจัดหมวดหมู่โดยใช้ SVM ร่วมกับเทคนิคการประมวลผลภาษา (NLP) ที่ครอบคลุมตั้งแต่การตัดคำ การลบคำหยุด และการสกัดคุณลักษณะสำคัญ ซึ่งผลการทดลองพบว่า SVM ให้ความแม่นยำสูงและเหมาะสมกับชุดข้อมูลเอกสารราชการที่มีความซับซ้อน [1] สอดคล้องกับบทความของ V. Lertnattee และ T. Theeramunkong ที่เลือกใช้ SVM สำหรับการจำแนกประเภทเว็บเพจยาแผนไทยและพบว่าเป็นวิธีการที่มีประสิทธิภาพในการจัดการกับข้อมูลข้อความ [2] เช่นเดียวกับบทความของ Y. Wahba, N. Madhavji, และ J. Steinbacher ที่ทำการศึกษาเปรียบเทียบและพบว่า SVM ยังคงเป็นอัลกอริทึมที่เหมาะสมและแม่นยำสำหรับชุดข้อมูลหลายมิติเมื่อเทียบกับโมเดลภาษาที่ซับซ้อนกว่าในบางบริบท [3] 2) Artificial Neural Network (ANN) ดังบทความของ C. Thanajiranthorn, N. Saenkham, T. Saenkham, และ W. Chosungnoen ได้เสนอวิธีการแก้ปัญหาชุดข้อมูลขนาดเล็กและไม่สมดุลโดยใช้เทคนิคการเพิ่มขยายข้อมูล ด้วยวิธีการแทนที่คำพ้องความหมาย เพื่อเพิ่มประสิทธิภาพของ ANN ให้สามารถจำแนกหมวดหมู่ได้ดียิ่งขึ้น[4] 3) WangchanBERTa:บทความของ L. Lowphansirikul, C. Polpanumas, N. Jantrakulchai และ S.Nutanong ได้นำเสนอ WangchanBERTa ซึ่งเป็นโมเดลสถาปัตยกรรม Transformer ที่ผ่านการเรียนรู้ล่วงหน้า (Pre-train) บนคลังข้อมูลภาษาไทยขนาดใหญ่ ทำให้โมเดลมีความเข้าใจบริบทภาษาไทยได้ลึกซึ้งกว่าโมเดลพื้นฐาน [5]

จากการสำรวจงานวิจัยที่ผ่านมา ผู้วิจัยได้วิเคราะห์เปรียบเทียบจุดเด่นและข้อจำกัดของแต่ละเทคนิค พบว่าสำหรับบริบทของบทความวิจัยในงานประชุมวิชาการ AUCC ซึ่งมีลักษณะเฉพาะคือเป็นข้อความสั้นและมีความหลากหลายของหัวข้อ (Tracks) แต่มีข้อจำกัดด้านปริมาณข้อมูลวิธีการใช้ SVM จึงมีความเหมาะสมที่สุด

2. ทฤษฎีที่เกี่ยวข้อง

การจัดหมวดหมู่บทความวิจัยอัตโนมัติจำเป็นต้องอาศัยหลักการทางคอมพิวเตอร์เพื่อแปลงข้อมูลภาษาธรรมชาติให้เครื่องสามารถประมวลผลและเรียนรู้ได้ งานวิจัยนี้ได้ศึกษาทฤษฎีและเทคนิคสำคัญที่เกี่ยวข้อง โดยแบ่งออกเป็น 5 ส่วนหลัก ได้แก่ การจำแนกด้วย SVM , ANNโมเดลภาษา WangchanBERTa, กระบวนการแปลงข้อความเป็นเวกเตอร์, และเกณฑ์การวัดประสิทธิภาพ ดังรายละเอียดต่อไปนี้

2.1 การจำแนกข้อความโดย SVM

เป็นเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอนที่มีประสิทธิภาพสูงสำหรับข้อมูลที่มีมิติมาก เช่น ข้อมูลข้อความ หลักการทำงานของ SVM ไม่ได้เพียงแค่ขีดเส้นแบ่งข้อมูล แต่เป็นการ "ค้นหา" ระนาบแบ่งกั้น ที่ดีที่สุด ซึ่งสามารถแยกกลุ่มข้อมูลออกจากกันโดยมีระยะห่างมากที่สุด กระบวนการค้นหานี้ช่วยให้โมเดลมีความทนทานต่อข้อมูลรบกวนและลดความเสี่ยงในการจำแนกผิดพลาดเมื่อนำไปใช้งานจริง โดยในงานวิจัยนี้เลือกใช้ Linear SVM ซึ่งเหมาะสมกับข้อมูลที่ต้องการความรวดเร็วในการประมวลผล **[1]**

ปัจจัยสำคัญที่มีผลต่อประสิทธิภาพของ SVM คือ ค่าพารามิเตอร์ C (Regularization Parameter) ซึ่งทำหน้าที่ควบคุมความสมดุลระหว่างความกว้างของระยะห่างและความถูกต้องในการจำแนกข้อมูล กล่าวคือ หากกำหนดค่า C สูง โมเดลจะให้ความสำคัญกับการจำแนกข้อมูลฝึกสอนให้ถูกต้องทั้งหมด ซึ่งจะทำให้ระยะห่างแคบลง (Hard Margin) แต่อาจเสี่ยงต่อภาวะการเรียนรู้เกิน (Overfitting) ในขณะที่การกำหนดค่า C ต่ำ

โมเดลจะยอมให้เกิดความคลาดเคลื่อนได้บ้างเพื่อขยายระยะห่างให้กว้างขึ้น (Soft Margin) ซึ่งช่วยให้โมเดลมีความยืดหยุ่นและสามารถจำแนกข้อมูลใหม่ที่ไม่เคยเห็นมาก่อนได้ดียิ่งขึ้น

2.2 การจำแนกข้อความโดย ANN

เป็นโมเดลการเรียนรู้เชิงลึกที่จำลองการทำงานของเซลล์ประสาทมนุษย์ในการออกแบบโมเดลสำหรับงานจำแนกข้อความแต่ละโหนดในชั้นซ่อนจะใช้ ฟังก์ชันกระตุ้น เช่น ReLU เพื่อคัดกรองสัญญาณข้อมูลที่สำคัญ และใช้ฟังก์ชัน Softmax ในชั้นสุดท้ายเพื่อคำนวณความน่าจะเป็นของแต่ละหมวดหมู่ ซึ่งช่วยให้โมเดลสามารถตัดสินใจเลือกคำตอบที่มีความเป็นไปได้สูงสุดได้

2.3 การจำแนกข้อความโดย WangchanBERTa

เป็นโมเดลภาษาไทยขั้นสูงที่พัฒนาบนสถาปัตยกรรม Transformer จุดเด่นคือการผ่านการฝึกฝนบนคลังข้อมูลภาษาไทยขนาดมหาศาล ทำให้โมเดลมีความเข้าใจในบริบท ความหมายแฝง และโครงสร้างของประโยคภาษาไทยที่มีความกำกวมได้ดีกว่าโมเดลสถิติแบบดั้งเดิม โดยใช้กลไก Self-Attention ในการวิเคราะห์ความสัมพันธ์ของคำทุกคำในประโยคพร้อมกัน ทำให้สามารถจำแนกเจตนาหรือหัวข้อของบทความได้อย่างแม่นยำ

2.4 การแปลงข้อความเป็นเวกเตอร์ (Text Vectorization)

เนื่องจากคอมพิวเตอร์ไม่สามารถประมวลผลตัวอักษรได้โดยตรง กระบวนการแปลงข้อความให้เป็นตัวเลขหรือ "เวกเตอร์" จึงมีความสำคัญอย่างยิ่ง งานวิจัยนี้เลือกใช้วิธี TF-IDF ซึ่งเป็นเทคนิคทางสถิติที่เปลี่ยนข้อความให้เป็นเวกเตอร์ตามน้ำหนักความสำคัญของคำ หลักการของ TF-IDF คือการให้คะแนนคำโดยพิจารณาจาก 2 ส่วน คือ 1) คำที่ปรากฏบ่อยในเอกสารนั้น (TF) จะได้คะแนนสูง และ 2) คำที่ปรากฏทั่วไปในทุกเอกสาร (IDF) จะถูกลดคะแนนลง วิธีนี้ช่วยลดความสำคัญของคำเชื่อมทั่วไป (เช่น "การ", "ที่") และเน้นคำเฉพาะเจาะจงที่บ่งบอกหัวข้อของบทความได้ชัดเจน

2.5 เทคนิค TF-IDF (Term Frequency-Inverse Document Frequency)

งานวิจัยนี้เลือกใช้ TF-IDF ซึ่งเป็นเทคนิคทางสถิติที่นิยมใช้ในการแปลงข้อความเป็นเวกเตอร์ตามน้ำหนักความสำคัญของคำ หลักการของ TF-IDF คือการให้คะแนนคำโดยพิจารณาจาก 2 ส่วนประกอบหลัก คือ

2.5.1 Term Frequency (TF): คือความถี่ของคำที่ปรากฏในเอกสารนั้น ๆ หากคำใดปรากฏบ่อยในเอกสาร จะได้คะแนนส่วนนี้สูง

2.5.2 Inverse Document Frequency (IDF) คือค่าที่ลดน้ำหนักของคำที่ปรากฏทั่วไปในทุกเอกสาร (เช่น "การ", "ที่", "และ") และเพิ่มน้ำหนักให้กับคำที่ปรากฏเฉพาะในเอกสารบางกลุ่ม ผลลัพธ์ที่ได้คือ การเน้นคำเฉพาะเจาะจง (Keywords) ที่สามารถบ่งบอกหัวข้อหรือใจความสำคัญของบทความนั้นได้อย่างชัดเจน

2.6 การประเมินผลด้วยค่าความถูกต้องและ F1-Score

เพื่อประเมินประสิทธิภาพของโมเดลในการจำแนกหมวดหมู่บทความวิจัย AUCC ได้อย่างรอบด้านและน่าเชื่อถือ งานวิจัยนี้ได้เลือกใช้เกณฑ์การวัดผลทางสถิติที่สำคัญ 2 ด้าน ดังนี้

2.6.1 ค่าความถูกต้อง (Accuracy) เป็นตัวชี้วัดพื้นฐานที่แสดงภาพรวมความสามารถของโมเดล โดยวัดจากสัดส่วนของจำนวนบทความที่โมเดลทำนายหมวดหมู่ได้ถูกต้องเมื่อเทียบกับจำนวนบทความทั้งหมด ค่าความถูกต้องที่สูงบ่งบอกถึงประสิทธิภาพโดยรวมที่ดีของระบบในการใช้งานทั่วไป

2.6.2 ค่า F1-Score เป็นตัวชี้วัดที่สำคัญที่สุดสำหรับงานวิจัยนี้ เนื่องจากชุดข้อมูลบทความในแต่ละกลุ่ม (Track) มีจำนวนไม่เท่ากัน ซึ่งอาจทำให้ค่า Accuracy ให้ผลลัพธ์ที่ลำเอียงไปทางกลุ่มที่มีข้อมูลจำนวนมาก ค่า F1-Score ช่วยแก้ปัญหานี้โดยการนำค่าความวัดผลสองส่วนมาคำนวณร่วมกัน คือ ค่าความแม่นยำ (Precision) ซึ่งวัดความถูกต้องเมื่อโมเดลทำนายว่าเป็นหมวดหมู่นั้น และ ค่าการดึงกลับ (Recall) ซึ่งวัดความสามารถในการค้นหาข้อมูลในหมวดหมู่นั้นได้ครบถ้วน

ดังนั้น ค่า F1-Score จึงเป็นค่าเฉลี่ยแบบถ่วงน้ำหนักระหว่าง Precision และ Recall ซึ่งเป็นดัชนีชี้วัดที่สะท้อนความสามารถ

ในการจำแนกที่แท้จริงได้ดีกว่าในกรณีที่ชุดข้อมูลมีความซับซ้อนและไม่สมดุล

2.7 ตัวแปรและพารามิเตอร์สำหรับการฝึกสอนโมเดล

ในการฝึกสอนโมเดลโครงข่ายประสาทเทียม (ANN) และโมเดลภาษา (WangchanBERTa) จำเป็นต้องมีการกำหนดค่าไฮเปอร์พารามิเตอร์ (Hyperparameters) ที่เหมาะสมเพื่อให้โมเดลเกิดการเรียนรู้ที่มีประสิทธิภาพ โดยมีตัวแปรสำคัญดังนี้

2.7.1 จำนวนรอบการฝึกสอน (Epochs)หมายถึงจำนวนครั้งที่โมเดลได้เรียนรู้จากชุดข้อมูลฝึกสอนครบทั้งหมดหนึ่งรอบ การกำหนดจำนวน Epoch ที่เหมาะสมจะช่วยให้โมเดลเรียนรู้รูปแบบข้อมูลได้ครบถ้วนโดยไม่เกิดภาวะการเรียนรู้เกิน (Overfitting)

2.7.2 ขนาดกลุ่มข้อมูล (Batch Size) หมายถึงจำนวนข้อมูลที่ถูกป้อนเข้าสู่โมเดลเพื่อประมวลผลในแต่ละครั้งก่อนที่จะมีการปรับปรุงค่าน้ำหนัก (Weight Update) การเลือก Batch Size ที่เหมาะสมจะช่วยให้การคำนวณ Gradient มีความเสถียรและใช้หน่วยความจำอย่างมีประสิทธิภาพ

2.7.3 อัตราการเรียนรู้ (Learning Rate) เป็นตัวแปรที่กำหนดขนาดของการปรับค่าน้ำหนักในแต่ละขั้นตอนของการฝึกสอน หากกำหนดค่าสูงเกินไปอาจทำให้โมเดลไม่สามารถหาค่าที่เหมาะสมที่สุดได้ แต่หากกำหนดค่าต่ำเกินไปจะทำให้การเรียนรู้ใช้เวลานาน

2.7.4 อัลกอริทึมหาค่าเหมาะสมที่สุด (Optimizer) เป็นอัลกอริทึมที่ใช้ในการปรับค่าน้ำหนักของโมเดลเพื่อลดค่าความสูญเสีย (Loss) ให้น้อยที่สุด โดยงานวิจัยนี้เลือกใช้ Adam (Adaptive Moment Estimation) ซึ่งเป็นอัลกอริทึมที่มีประสิทธิภาพสูงในการปรับอัตราการเรียนรู้ให้เหมาะสมกับพารามิเตอร์แต่ละตัวโดยอัตโนมัติ

3. วิธีดำเนินการวิจัย

3. วิธีดำเนินการวิจัย

งานวิจัยเรื่อง “การคัดแยกกลุ่มหัวข้องานวิจัย AUCC ด้วยซัพพอร์ตเวกเตอร์แมชชีน (SVM)” มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของโมเดลในการจำแนกหมวดหมู่บทความอัตโนมัติ โดยมีขั้นตอนการดำเนินงานดังนี้

3.1 การกำหนดความต้องการ

ขั้นตอนแรกของการพัฒนาระบบ คือ การวิเคราะห์ปัญหาและความต้องการของผู้ใช้งานการคัดแยกกลุ่มหัวข้องานวิจัย AUCC ด้วย ซัพพอร์ตเวกเตอร์แมชชีน (SVM) ซึ่งในปัจจุบันการจัดหมวดหมู่บทความยังอาศัยการคัดเลือกด้วยมือ ทำให้เกิดความคลาดเคลื่อนและใช้เวลานาน ผู้วิจัยจึงกำหนดความต้องการให้จำแนกประเภทบทความภาษาไทยได้อย่างอัตโนมัติ โดยเน้นความถูกต้อง ความรวดเร็ว

3.2 การรวบรวมข้อมูล

ผู้วิจัยได้รวบรวมชุดข้อมูลจากฐานข้อมูลการประชุมวิชาการระดับชาติ AUCC โดยคัดเลือกบทความในส่วนชื่อเรื่องและบทคัดย่อ จำนวนรวมทั้งสิ้น 173 บทความ ข้อมูลทั้งหมดถูกจำแนกออกเป็น 5 กลุ่มหัวข้อ (Tracks) ดังแสดงรายละเอียดใน **ตาราง 1** ข้อมูลและประเภท

ประเภท	จำนวนข้อมูล
Information Technology	46
Computer Education	42
Multimedia Computer Graphics and Games	42
Computer Business	22
Data Science and Analytics	21

3.3 การเตรียมข้อมูล (Data Preprocessing)

เพื่อให้คอมพิวเตอร์สามารถประมวลผลภาษาธรรมชาติ (NLP) ได้ ผู้วิจัยดำเนินการเตรียมข้อมูลตามขั้นตอนดังนี้

3.3.1 การตัดคำ ใช้ไลบรารี PyThaiNLP ในการแยกประโยคภาษาไทยออกเป็นคำย่อย

3.3.2 การกำจัดคำหยุด คัดกรองคำเชื่อมและคำที่ไม่มีนัยสำคัญต่อการจำแนกประเภทออก เพื่อลดสัญญาณรบกวน

3.3.4 การสกัดคุณลักษณะ แปลงข้อมูลข้อความให้เป็นเวกเตอร์ตัวเลข ด้วยเทคนิค TF-IDF เพื่อคำนวณค่าน้ำหนักความสำคัญของคำ โดยมีหลักการคำนวณดังที่ได้กล่าวไว้ในหัวข้อ

2.5 ซึ่งวิธีนี้ช่วยลดมิติของข้อมูลและคัดกรองเฉพาะคำสำคัญที่ส่งผลต่อการจำแนกประเภทบทความ

3.3.5 การแบ่งชุดข้อมูล แบ่งข้อมูลออกเป็นชุดฝึกสอน ร้อยละ 80 และชุด ร้อยละ 20

3.4 การสร้างและฝึกโมเดล (Model Development)

การทดลองนี้ดำเนินการบนสภาพแวดล้อม Google Colab โดยใช้ภาษา Python และไลบรารีมาตรฐานทางด้านปัญญาประดิษฐ์ ได้แก่ Scikit-learn, TensorFlow/Keras และ Hugging Face Transformers ผู้วิจัยได้พัฒนาและทดสอบโมเดลการเรียนรู้ 3 รูปแบบ เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกหมวดหมู่บทความ ดังรายละเอียดต่อไปนี้

3.4.1 ซัพพอร์ตเวกเตอร์แมชชีน (SVM)

เนื่องจากเป็นโมเดลหลักในการศึกษานี้ ผู้วิจัยเลือกใช้ไลบรารี LinearSVC ร่วมกับเวกเตอร์คุณลักษณะจาก TF-IDF โดยมีการกำหนดค่าและการปรับปรุงโมเดลดังนี้

การกำหนดเคอร์เนล (Kernel Selection) เลือกใช้ Linear Kernel เนื่องจากข้อมูลข้อความที่ผ่านการแปลงด้วย TF-IDF มีลักษณะเป็นข้อมูลหลายมิติ ซึ่งเคอร์เนลแบบเชิงเส้นสามารถประมวลผลได้รวดเร็วและให้ผลลัพธ์ที่มีประสิทธิภาพสูง

การปรับจูนพารามิเตอร์ ดำเนินการค้นหาค่าพารามิเตอร์ที่เหมาะสมที่สุดด้วยวิธี GridSearchCV โดยเน้นการปรับค่า C (Regularization Parameter) เพื่อหาจุดสมดุลที่ดีที่สุดระหว่างความกว้างของระยะห่าง และความถูกต้องในการจำแนกกลุ่มข้อมูล

3.4.2 โครงข่ายประสาทเทียม(ANN)

ผู้วิจัยออกแบบสถาปัตยกรรมแบบ Multilayer Perceptron (MLP) สำหรับการเรียนรู้เชิงลึกโดยมีโครงสร้างและกระบวนการดังนี้

โครงสร้างโมเดล ประกอบด้วยเลเยอร์ซ่อน (Hidden Layers) จำนวน 2 ชั้น ใช้ฟังก์ชันกระตุ้นแบบ ReLU ในชั้นซ่อน และฟังก์ชัน Softmax ในชั้นผลลัพธ์เพื่อคำนวณความน่าจะเป็นของทั้ง 5 หมวดหมู่ การปรับพารามิเตอร์ ใช้อัลกอริทึม Adam Optimizer ในการปรับปรุงค่าน้ำหนักของโมเดลระหว่างการฝึกสอน

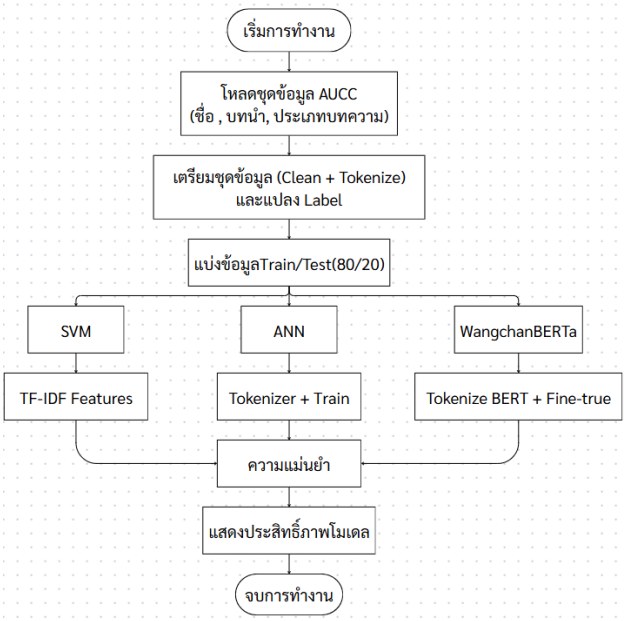
การจัดการข้อมูล เนื่องจากชุดข้อมูล AUCC มีขนาดเล็กและไม่สมดุล ผู้วิจัยจึงประยุกต์ใช้เทคนิค การเพิ่มขยายข้อมูลด้วยวิธีการแทนที่คำด้วยคำพ้องความหมาย เพื่อเพิ่มจำนวนตัวอย่างในกลุ่มที่มีข้อมูลน้อย ช่วยให้โมเดลเรียนรู้ได้ดีขึ้นและลดโอกาสเกิด Overfitting

3.4.3 WangchanBERTa

ผู้วิจัยนำโมเดลภาษาไทยสถาปัตยกรรม Transformer ที่ผ่านการเรียนรู้ล่วงหน้า ชื่อรุ่น wangchanberta-base-att-spm-uncased มาประยุกต์ใช้เพื่อเปรียบเทียบประสิทธิภาพกับโมเดลพื้นฐาน

กระบวนการ นำโมเดลมาเข้าสู่กระบวนการปรับจูน ด้วยชุดข้อมูลบทความวิจัย AUCC โดยใช้ Trainer API ของ Hugging Face เพื่อปรับค่าน้ำหนักของโมเดลให้เข้ากับบริบทของภาษาทางวิชาการและการจัดหมวดหมู่เฉพาะทาง

เพื่อให้เห็นภาพรวมของกระบวนการวิจัยทั้งหมด ตั้งแต่การนำเข้าข้อมูล การประมวลผล จนถึงการฝึกสอนและวัดผลโมเดล ผู้วิจัยได้สรุปขั้นตอนการทำงานไว้ในแผนผังงาน(Flowchart) ดังแสดงในภาพ 1



ภาพ 1 ผังงานของการคัดแยกกลุ่มหัวข้องานวิจัย AUCC ด้วยซัพพอร์ตเวกเตอร์แมชชีน (SVM)

3.4.4 การกำหนดค่าพารามิเตอร์

เพื่อให้การทดลองมีความน่าเชื่อถือและสามารถทำซ้ำได้ (Reproducibility) ผู้วิจัยได้กำหนดค่าพารามิเตอร์ที่สำคัญสำหรับการฝึกสอนโมเดลทั้ง 3 รูปแบบ โดยเลือกใช้ค่าที่ให้ประสิทธิภาพสูงสุดจากการทดสอบเบื้องต้น ซึ่งมีรายละเอียดการกำหนดค่าตามหลักการทฤษฎีดังนี้

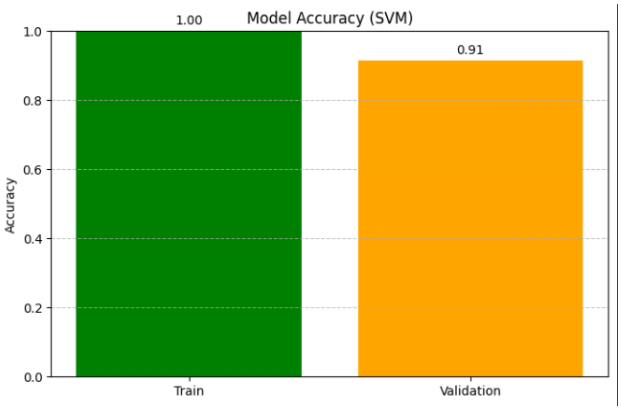
3.4.4.1 โมเดล SVM ผู้วิจัยได้กำหนดค่าพารามิเตอร์ C (Regularization Parameter) เพื่อควบคุมความสมดุลระหว่างระยะห่าง (Margin) และความถูกต้องในการจำแนกข้อมูล โดยอ้างอิงหลักการทำงานตามที่ได้อธิบายไว้ใน หัวข้อ 2.1

3.4.4.1โมเดล ANN และ WangchanBERTa ผู้วิจัยได้กำหนดค่าไฮเปอร์พารามิเตอร์สำหรับการฝึกสอน ได้แก่ จำนวนรอบการฝึกสอน (Epochs), ขนาดกลุ่มข้อมูล (Batch Size), อัตราการเรียนรู้ (Learning Rate) และ อัลกอริทึมหาค่าเหมาะสมที่สุด (Optimizer) โดยอ้างอิงนิยามและความสำคัญของตัวแปรเหล่านี้จาก หัวข้อ 2.7 เพื่อให้โมเดลเกิดการเรียนรู้ที่มีประสิทธิภาพสูงสุด

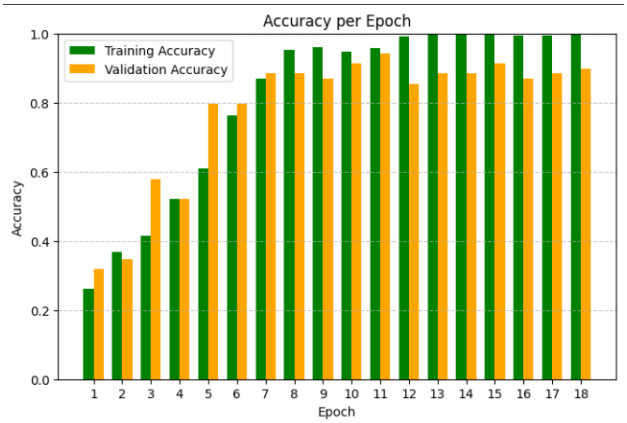
รายละเอียดค่าพารามิเตอร์ที่ดีที่สุด (Best Parameters) ที่ผู้วิจัยเลือกใช้ในการทดลอง แสดงดังตารางที่ 2

ตาราง 2 การกำหนดพารามิเตอร์

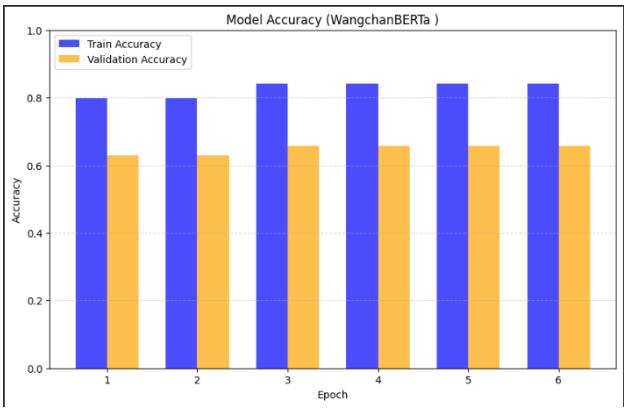
พารามิเตอร์	SVM	ANN	WangchanBERTa
Optimizer	-	Adam	AdamW
Learning Rate	-	0.001	2e-5
Batch Size	-	32	16
Epochs/Max Iter	1,000	50	5
Specific Config	Linear Kernel, C	ReLU, Softmax	Pre-trained



ภาพ 2 กราฟค่าความแม่นยำในการเทรนของ SVM



ภาพ 3 กราฟค่าความแม่นยำในการเทรนของ ANN



ภาพ 4 กราฟค่าความแม่นยำในการเทรนของ WangchanBERTa

4. ผลของการวิจัย

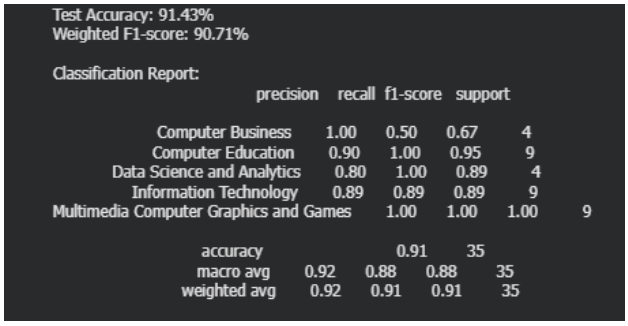
ผลจากการทดลอง สามารถแสดงการทำงานได้แก่

4.1 กราฟการทำงานของการเทรนโมเดล

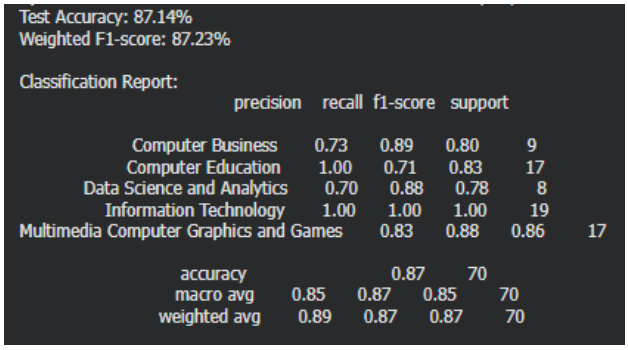
โมเดลทั้ง 3 โมเดลได้แก่ SVM ในภาพ 2 ANN ในภาพ 3 และ WangchanBERTa ในภาพ 4

4.2 ผลการเทรนโมเดล

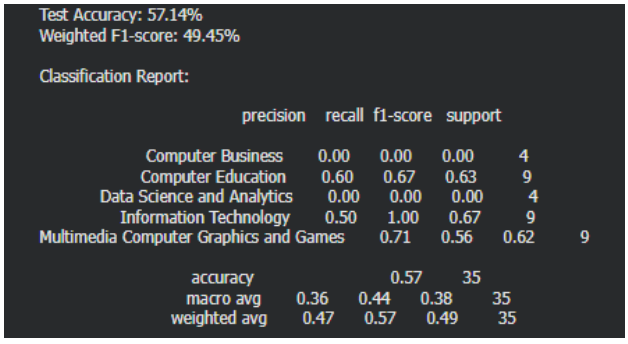
จะแสดงผลการเทรนค่าความแม่นยำ(Accuracy) และF1-Score แสดงผลจาก SVM ในภาพ 5 ANN ในภาพ 6 และ WangchanBERTa ในภาพ 7



ภาพ 5 ผลการทำงานของ SVM



ภาพ 6 ผลการทำงานของ ANN



ภาพ 7 ผลการทำงานของ WangchanBERTa

4.3 ผลการประเมินความแม่นยำของโมเดล

ผลการประเมินความแม่นยำจากการทดลอง จะพบว่า ข้อมูลจาก Data Set มีความแม่นยำแตกต่างกันต่อโมเดลที่พัฒนาขึ้น ดังแสดงในตาราง 3

ตาราง 3 ตารางเปรียบเทียบความแม่นยำของโมเดลต่างๆ

Model	Accuracy (%)	f1-score
SVM	91.71	90.71
ANN	87.14	87.23
WangchanBERTa	57.14	49.45

จากตาราง 3 โมเดลที่มีประสิทธิภาพสูงสุด คือ SVM โดยสามารถทำค่าความแม่นยำ (Accuracy) ได้สูงถึง 91.71% และมีค่า f1-score อยู่ที่ 90.71% ซึ่งสะท้อนถึงความสามารถในการจำแนกข้อมูลได้อย่างถูกต้องและมีความเสถียรสูงสุดในการทดลองนี้

โมเดลที่มีประสิทธิภาพรองลงมา คือ ANN ซึ่งให้ค่าความแม่นยำที่ 87.14% แม้จะมีประสิทธิภาพด้อยกว่า SVM เล็กน้อย แต่ยังถือว่าอยู่ในเกณฑ์ที่สูงและยอมรับได้สำหรับการนำไปใช้งาน

โมเดลที่มีประสิทธิภาพต่ำสุด คือ WangchanBERTa โดยทำความแม่นยำได้เพียง 57.14% ซึ่งแสดงให้เห็นว่าโมเดลประเภท Transformer ขนาดใหญ่อาจไม่เหมาะสมกับลักษณะหรือปริมาณของชุดข้อมูล (Data Set) ที่ใช้ในการทดลองนี้เท่ากับโมเดลพื้นฐานอย่าง SVM และ ANN

4.4 การอภิปรายผลการวิจัย

จากผลการทดลองเปรียบเทียบประสิทธิภาพของทั้ง 3 โมเดล ผู้วิจัยสามารถอภิปรายประเด็นสำคัญที่ส่งผลต่อความแตกต่างของคะแนนประสิทธิภาพได้ดังนี้ โมเดล SVM แสดงประสิทธิภาพได้โดดเด่นที่สุด เนื่องจากลักษณะโครงสร้างทางคณิตศาสตร์ของ SVM มีความสามารถในการจัดการกับข้อมูลที่มีมิติสูงแต่มีจำนวนตัวอย่างจำกัด ได้อย่างมีประสิทธิภาพ เมื่อนำมาใช้งานร่วมกับการสกัดคุณลักษณะด้วย TF-IDF จึงทำให้โมเดลสามารถให้น้ำหนักกับคำสำคัญ ที่ระบุเอกลักษณ์ของแต่ละกลุ่มหัวข้อ ได้อย่างแม่นยำ แม้จะมีชุดข้อมูลเพียง 173 บทความก็ตาม ในส่วนของ โมเดล ANN ที่ให้ผลลัพธ์ในระดับปานกลางนั้น วิเคราะห์ได้ว่าแม้ผู้วิจัยจะนำเทคนิคการเพิ่มขยายข้อมูล มาช่วยเสริมแล้ว แต่โดยธรรมชาติของโครงข่ายประสาทเทียมยังคงจำเป็นต้องใช้ฐานข้อมูลการเรียนรู้ขนาดใหญ่ เพื่อให้ค่าน้ำหนักในชั้นซ่อน สามารถจดจำรูปแบบความซับซ้อนของภาษาไทยได้ลึกซึ้งพอ ปริมาณข้อมูลในงานวิจัยนี้จึงอาจยังไม่เพียงพอที่จะดึงศักยภาพสูงสุดของ ANN ออกมาได้เมื่อเทียบกับ SVM สำหรับกรณีของ WangchanBERTa ที่ให้ผลลัพธ์น้อยที่สุดนั้น มีสาเหตุ

หลักมาจากสถาปัตยกรรมของโมเดลที่เป็น Transformer ขนาดใหญ่ ซึ่งถูกออกแบบมาให้ต้องผ่านการปรับจูน (Fine-tuning) ด้วยชุดข้อมูลเฉพาะทางปริมาณมหาศาล การฝึกสอนด้วยข้อมูลจำนวน 173 บทความ ถือเป็นปริมาณที่น้อยเกินไปสำหรับโมเดลระดับนี้ ส่งผลให้เกิดภาวะการเรียนรู้ที่ต่ำกว่าเกณฑ์ โมเดลจึงไม่สามารถแยกแยะบริบทที่ซับซ้อนของหัวข้อวิจัยที่คาบเกี่ยวกันได้อย่างแม่นยำ

5. สรุปผลการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาการคัดแยกกลุ่มหัวข้อบทความวิชาการ AUCC ด้วย ซัพพอร์ตเวกเตอร์แมชชีน (SVM) แบบอัตโนมัติ โดยดำเนินการทดลองกับโมเดลหลัก 3 รูปแบบ ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (SVM), โครงข่ายประสาทเทียม (ANN) และโมเดลภาษา WangchanBERTa เพื่อค้นหาแนวทางที่มีประสิทธิภาพสูงสุดในการลดภาระงานและความผิดพลาดจากการคัดแยกด้วยมนุษย์

ผลการทดลองพบว่า โมเดล SVM ให้ประสิทธิภาพสูงสุดในทุกมิติการวัดผล โดยมีค่าความถูกต้อง (Accuracy) สูงถึง 91.43% และมีค่าเฉลี่ย F1-Score เท่ากับ 90.71ซึ่งสูงกว่าโมเดล ANN ที่มี Accuracy 87.14%, F1-Score 87.23% และ WangchanBERTa ที่มี Accuracy 57.14%, F1-Score 49.45% ตามลำดับดัง ตาราง 3

การที่ SVM มีค่า F1-Score สูงที่สุด ทำให้เห็นได้ว่าโมเดลมีความสามารถในการจำแนกบทความได้อย่างแม่นยำ และมีความเหมาะสมอย่างยิ่งกับข้อมูลข้อความภาษาไทย

โดยสรุป SVM เป็นโมเดลที่มีประสิทธิภาพและความคุ้มค่าที่สุดสำหรับการนำไป คัดแยกกลุ่มหัวข้อบทความวิชาการ AUCC เนื่องจากมีความแม่นยำสูง ประมวลผลได้รวดเร็ว และมีความเสถียรในการจำแนกข้อความภาษาไทย จะสามารถช่วยลดข้อผิดพลาดจากการจัดหมวดหมู่ด้วยมือ

คำชี้แจงการใช้ AI

ต้นฉบับนี้มีเนื้อหาที่สร้างขึ้นหรือได้รับการสนับสนุนจากเครื่องมือปัญญาประดิษฐ์(AI) ได้แก่ CHATGPT และ GEMINI ซึ่งถูกนำมาใช้เพื่อช่วยในการดำเนินงาน เช่น การแก้ไขภาษา การวิเคราะห์ข้อมูล และช่วยในการเขียนโค้ด ทั้งนี้ ผู้เขียนได้ตรวจสอบเนื้อหาที่สร้างโดย AI อย่างรอบคอบ เพื่อให้มั่นใจในความถูกต้อง ความเป็นต้นฉบับ และการปฏิบัติตามหลักจริยธรรมทางวิชาการ

เอกสารอ้างอิง

[1] ปกรณ์ สันตกิจ, พงษ์พร พันธ์เพ็ง, ปรีชา โพธิ์แพง, และ เยาวลักษณ์ งามแสนโรจน์, “ระบบการจัดหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ SVM ร่วมกับการประมวลผลภาษา,” *The Journal of KMUTNB*, ปีที่ 34, ฉบับที่ 4, ตุลาคม–ธันวาคม 2567. [ออนไลน์]. [สืบค้นวันที่ 17 กันยายน 2568] https://www.researchgate.net/publication/380446370_Text_Classification_Using_Machine_Learning_for_Thai_Official_Letters

[2] V.Lertnattee and T. Theeramunkong, ‘Text classification for thai medicinal web pages’, in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2007, pp. 631–638. [ออนไลน์].[สืบค้นวันที่ 17 กันยายน2568] จาก https://link.springer.com/chapter/10.1007/978-3-540-71701-0_67

[3] Y. Wahba, N. Madhavji, and J. Steinbacher, ‘A comparison of svm against pre-trained language models (plms) for text classification tasks’, in *International Conference on Machine Learning, Optimization, and Data Science*, 2022, pp. 304–313. [ออนไลน์].[สืบค้นวันที่ 17 กันยายน 2568] จาก https://link.springer.com/chapter/10.1007/978-3-031-25891-6_23

[4] C. Thanajiranthorn, N. Saenkham, T. Saenkham, and W. Chosungnoen, ‘Thai Text-Based Classification for Small and Imbalanced Dataset’, *Available at SSRN 4932993*. [ออนไลน์]. [สืบค้นวันที่ 17 กันยายน 2568] จาก https://www.ijcesen.com/index.php/ijcesen/article/view/3015

[5] L. Lowphansirikul, C. Polpanumas, N. Jantrakulchai,

andS.Nutanong,‘Wangchanberta:Pretrainingtra
nsformer-based thai language models’,*arXiv
preprintarXiv:2101. 09635*, 2021. [ออนไลน์]
[สืบค้นวันที่ 17 กันยายน 2568] จาก
https://arxiv.org/abs/2101.09635