วารสารวิทยาศาสตร์ลาดกระบัง ปีที่ 34 ฉบับที่ 1 เดือนมกราคม - มิถุนายน 2568
Journal of Science Ladkrabang Vol. 34 No. 1 January - June 2025

บทความวิจัย (Research article)

การเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกคำบาลีและสันสกฤต ในภาษาไทยด้วยเทคนิคการเรียนรู้ของเครื่อง

Efficiency Comparison of Classification Models for Pali and Sanskrit in Thai Languages Using Machine Learning Techniques

ณัฐรฐนนท์ กานต์รวีกุลธนา รุ่งทิพย์ โคบาล วาสนา ด้วงเหมือน และ สุภษี ดวงใส*
Natratanon Kanraweekultana Rungthip Cobal Wassana Duangmeun and Supasee Duangsai*
สาขาวิชาเทคโนโลยีสารสนเทศและธุรกิจดิจิทัล คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลกรุงเทพ กรุงเทพฯ ประเทศไทย
Department of Information Technology and Digital Business Program, Faculty of Business Administration, Rajamangala University of Technology Krungthep, Bangkok, Thailand
วันที่ส่งบทความ : 12 พฤศจิกายน 2567 วันที่แก้ไขบทความ : 12 มิถุนายน 2568 วันที่ตอบรับบทความ : 14 มิถุนายน 2568
Received: 12 November 2024, Revised: 12 June 2025, Accepted: 14 June 2025

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อทดสอบและเปรียบเทียบประสิทธิภาพของแบบจำลองการจำแนกคำบาลีและสันสกฤตในภาษาไทยโดยใช้เทคนิคการเรียนรู้ของเครื่อง เน้นการพัฒนาความแม่นยำในการแยกแยะคำศัพท์ที่มาจากสองภาษานี้ ซึ่งมีความใกล้เคียงกันในด้านการออกเสียงและการเขียน ในการวิจัยใช้แบบจำลอง 5 ชนิด ได้แก่ แบบจำลองป่าไม้สุ่ม แบบจำลองต้นไม้ตัดสินใจ แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด แบบจำลองนาอีฟเบย์ และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ผ่านกระบวนการตรวจสอบ 10-fold cross-validation เพื่อวัดประสิทธิภาพของแต่ละแบบจำลอง ผลการวิจัยพบว่า แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน มีความสามารถสูงสุดในการจำแนกคำบาลีและสันสกฤตในภาษาไทย โดยมีค่าความแม่นยำสูงถึงร้อยละ 95.75 และค่าความถูกต้องร้อยละ 90.90 รองลงมา คือ แบบจำลองการคำนวณเพื่อนบ้านใกล้สุดมีค่าความแม่นยำร้อยละ 92.86 และแบบจำลองนาอีฟเบย์ที่มีค่าความแม่นยำร้อยละ 81.29 ขณะที่แบบจำลองแบบจำลองป่าไม้สุ่มมีประสิทธิภาพต่ำที่สุด โดยมีค่าความแม่นยำเพียงร้อยละ 55.27 ข้อค้นพบเหล่านี้แสดงให้เห็นถึงความเหมาะสมของการใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ในการแยกแยะภาษาบาลีและสันสกฤตได้อย่างมีประสิทธิภาพ และสามารถนำผลลัพธ์ไปประยุกต์ใช้ในงานด้านการจำแนกภาษา การแปลภาษา รวมถึงการพัฒนาเครื่องมือด้านการเรียนการสอนภาษาในอนาคต

คำสำคัญ : ประสิทธิภาพแบบจำลอง การจำแนกภาษา บาลี สันสกฤต การเรียนรู้ของเครื่อง

*ที่อยู่ติดต่อ E-mail address: supasee.d@mail.rmutk.ac.th https://doi.org/10.55003/scikmitl.2025.265286

Abstract

This research aims to test and compare the performance of classification models for Pali and Sanskrit in Thai language using Machine learning techniques. The study focuses on improving the accuracy in distinguishing between words from these two languages, which often exhibit similarities in pronunciation and spelling. Five models were tested: Random Forest, Decision Tree, K-Nearest Neighbors (K-NN), Naive Bayes, and Support Vector Machine (SVM). The evaluation process employed 10-fold cross-validation to assess model performance. The results indicate that the SVM model is the most efficient, with an accuracy of 95.75% and a precision of 90.90%. K-NN follows closely with an accuracy of 92.86%, while Naive Bayes achieves 81.29%. The Random Forest model, however, shows the lowest performance with an accuracy of only 55.27%. These findings highlight the SVM model's effectiveness in accurately classifying Pali and Sanskrit words in Thai language. The results can be applied to further developments in language classification, translation, and educational technology tools for language learning.

Keywords: Model efficiency, Languages classification, Pali, Sanskrit, Machine learning

1. บทนำ

ภาษาบาลีและสันสกฤตเป็นรากฐานสำคัญของคำศัพท์ในภาษาไทย โดยเฉพาะคำที่เกี่ยวข้องกับศาสนา วัฒนธรรม และการศึกษา การที่มีคำยืมจากทั้งสองภาษาในจำนวนมาก ทำให้เกิดความซับซ้อนในการแยกแยะว่าแต่ละคำมาจากภาษาใด แม้จะมีการสอนภาษาบาลีและสันสกฤตในระบบการศึกษาไทย แต่เนื้อหาที่สอนอาจไม่ครอบคลุมหรือไม่ละเอียดเพียงพอ ทำให้นักเรียนหรือบุคคลทั่วไปขาดความสามารถในการจำแนกคำที่มาจากสองภาษานี้อย่างถูกต้อง อีกทั้งภาษาไทยมีการพัฒนาและปรับเปลี่ยนอย่างต่อเนื่อง ทำให้บางคำจากภาษาบาลีและสันสกฤตอาจถูกนำมาใช้ในบริบทใหม่ ๆ ซึ่งอาจไม่ตรงกับความหมายดั้งเดิม นำไปสู่ความสับสนและการใช้คำที่ไม่ถูกต้อง โดยหลายคำจากภาษาบาลีและสันสกฤตมีเสียงและการเขียนที่คล้ายกัน ผู้ใช้ภาษาอาจสับสนในการแยกแยะ ซึ่งเกิดจากบางคำอาจถูกแปลงเสียงหรือเขียนผิดเพี้ยนไปจากรากศัพท์เดิม ค้นหาในพจนานุกรมไม่เจอซึ่งพจนานุกรมส่วนใหญ่มักแยกบาลีและสันสกฤตออกจากกันไม่ชัดเจน ทำให้เกิดความสับสนในการอ้างอิง ถึงแม้จะมีการนำกฎไวยากรณ์แบบ Rule-based อาจไม่แม่นยำ เนื่องจากคำบาลีและสันสกฤตมีรูปแบบการเปลี่ยนแปลงที่ซับซ้อน (Tammanam et al., 2021) เช่น การเปลี่ยนรูปเมื่ออยู่ในบริบทต่าง ๆ ตามบริบทของภาษาไทยที่มักนำคำบาลีและสันสกฤตมาใช้โดยมีการดัดแปลง เช่น การเติมสระหรือพยัญชนะบางตัวเพื่อให้สอดคล้องกับระบบเสียงของไทย ทำให้ไม่มีชุดกฎที่แน่นอนตายตัว เพราะมีข้อยกเว้นและการใช้ที่หลากหลาย ทำให้เกิดข้อจำกัดในการใช้ Rule-based approach (Zhong et al., 2024) นอกจากนี้ ปัญหาการจำแนกคำบาลีและสันสกฤตในภาษาไทยเป็นเรื่องที่ซับซ้อน เนื่องจากสองภาษานี้มีความใกล้เคียงกันมาก และมักจะใช้ปะปนกันในภาษาไทย โดยเฉพาะใน

ศัพท์ที่เกี่ยวข้องกับศาสนา วรรณกรรม และวิชาการ ปัญหาหลักคือภาษาบาลีและสันสกฤตเป็นภาษากลุ่มอินโด-อารยัน ซึ่งมีคำที่มีรากศัพท์เดียวกันหรือคล้ายกันมาก ทำให้แยกแยะได้ยาก เช่น คำว่า ธรรม (ธรรมะ) มีทั้งในบาลี (Dhamma) และสันสกฤต (Dharma) อีกทั้งคำหลายคำในภาษาไทยเป็นคำผสมระหว่างบาลีและสันสกฤต เช่น วิทยาศาสตร์ (วิทยา = สันสกฤต และ ศาสตร์ = บาลี) ทำให้ไม่สามารถระบุได้ว่าเป็นของภาษาใดเพียงภาษาเดียว และยังมีการเปลี่ยนแปลงในระดับของเสียงและการเขียนเมื่อคำถูกนำมาใช้ในภาษาไทย แต่อย่างไรภาษาบาลีและสันสกฤตยังมีความสำคัญต่อทั้งในด้านศาสนา วัฒนธรรม และการศึกษา หรือแม้แต่ในชีวิตประจำวันยังต้องมีความเข้าใจ แยกแยะ และการจำแนกคำที่มาจากภาษาบาลีและสันสกฤตอย่างถูกต้องตามรากศัพท์และการใช้ภาษาได้อย่างถูกต้อง (Li, 2024) ต้องอาศัยความรู้ความเข้าใจในเรื่องดังกล่าวเพื่อช่วยในการสื่อสารและการเขียนที่ถูกต้องและแม่นยำ ซึ่งการจำแนกแยกแยะต้องอาศัยทักษะและประสบการณ์ของนักวิชาการ ครู อาจารย์ หรือนักภาษาศาสตร์ผ่านเครื่องมือทั้งรูปแบบดั้งเดิม คือ การใช้พจนานุกรม และรูปแบบสมัยใหม่ คือ การใช้เว็บไซต์สืบค้นหรือการใช้แหล่งข้อมูลออนไลน์จากหน่วยงานที่รับผิดชอบด้านภาษา เช่น สำนักงานราชบัณฑิตยสภา ซึ่งความแตกต่างระหว่าง 2 ภาษานี้ คือ ที่มาของคำศัพท์ การใช้ในภาษาไทย และโครงสร้างทางไวยากรณ์ ทั้งสองภาษานี้มีบทบาทสำคัญในการพัฒนาภาษาไทยและวัฒนธรรมไทยอย่างมาก สามารถแบ่งออกเป็น 5 ด้าน ได้แก่

1) ด้านแหล่งที่มา ภาษาบาลีมาจากแถบอินเดียตอนเหนือและเป็นภาษาของคัมภีร์พุทธศาสนาในฝ่ายเถรวาท ซึ่งมีอิทธิพลต่อไทยมาก เนื่องจากไทยเป็นประเทศที่นับถือศาสนาพุทธแบบเถรวาท ในขณะที่ภาษาสันสกฤตเป็นภาษาที่เก่าแก่และมีความซับซ้อนมากกว่าภาษาบาลี มาจากแถบอินเดียเช่นกัน แต่มักเกี่ยวข้องกับศาสนาฮินดูและวรรณกรรมคลาสสิก เช่น มหาภารตะและรามายณะ เป็นต้น

2) การใช้ในภาษาไทย ภาษาบาลีในภาษาไทยมักเป็นคำที่เกี่ยวข้องกับศาสนาและวัฒนธรรมทางพระพุทธศาสนา เช่น พระ ศีล บริจาค และบุญ เป็นต้น ในขณะที่ภาษาสันสกฤตในภาษาไทยมักใช้ในคำที่เป็นทางการหรือลึกซึ้ง เช่น กษัตริย์ ประชาธิปไตย องค์ และปรัชญา เป็นต้น

3) ลักษณะของคำศัพท์ ในภาษาบาลี คำบาลีมักสั้นและง่ายต่อการออกเสียงในภาษาไทยเป็นภาษาที่ไม่มีวรรณยุกต์ ส่วนใหญ่เป็นคำพยางค์เปิด (Open syllable) คือ ลงท้ายด้วยสระ เช่น พุท-ธะ สง-ฆะ และ วิ-หา-ระ เป็นต้น ไม่เน้นเสียงควบกล้ำและพยัญชนะสะกด เช่น ศีล (Pali: s**ī**la) และธรรม (Pali: dhamma) เป็นต้น ในขณะที่ภาษาสันสกฤต คำสันสกฤตมักยาวและซับซ้อนกว่า เป็นภาษาที่มีการใช้วรรณยุกต์และพยัญชนะควบกล้ำ ภาษาคำผัน (Inflectional language) มีการเปลี่ยนรูปของรากศัพท์ตามหน้าที่ในประโยค และมีคำที่เป็นพยางค์ปิดมากกว่าบาลี คือ ลงท้ายด้วยพยัญชนะสะกด เช่น โลก พรหม วรรค และกรรม เป็นต้น (Duraiswamy, 2024) เช่น ประชาธิปไตย (Sanskrit: praj**ā**dhipatya) และปฐมฤกษ์ (Sanskrit: prathama-rik**Ṣ**a) เป็นต้น

4) โครงสร้างทางไวยากรณ์ ในภาษาบาลีมีโครงสร้างทางไวยากรณ์ที่ชัดเจนและเป็นระบบ โดยมีการผันคำนาม คำสรรพนามหรือคำคุณศัพท์ (Declension) และการผันคำกริยา (Conjugation) แต่เข้าใจและจดจำได้ง่าย โดยมีรูปแบบการประสมคำแบบสมาส (การรวมคำโดยไม่มีการเปลี่ยนแปลงคำเดิม) เช่น พุทธศาสนา = พุทธ (พระพุทธเจ้า) + ศาสนา (คำสอน) และ อริยสัจ = อริยะ (ประเสริฐ) + สัจ (ความจริง) เป็นต้น ในขณะที่ภาษาสันสกฤตมีโครงสร้างทางไวยากรณ์ที่ซับซ้อนและเป็นระบบมาก ซึ่งเป็นหนึ่งในเอกลักษณ์สำคัญของภาษาโบราณ โดยคำที่ถูกยืมมาใช้มักถูกดัดแปลงให้เข้ากับเสียงของภาษาไทย ซึ่งเป็นรูปแบบการ

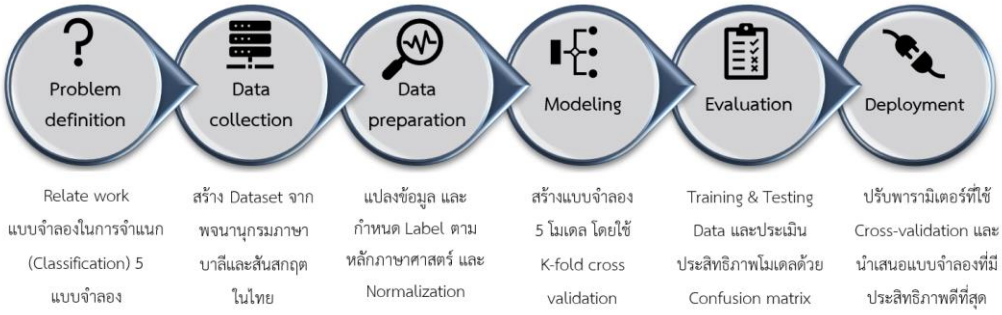
ประสมคำแบบตัทธิต (การเติมปัจจัยขยายความ) และสมาส เช่น วรรณคดี = วรรณะ (สี) + คดี (เรื่องราว) หมายถึงวรรณกรรม และนฤมิต = นฤ (ไม่มี) + มิต (ขอบเขต) หมายถึงการสร้างสรรค์ เป็นต้น (Saechan & Ruangjaroon, 2024)

5) การออกเสียง ในภาษาบาลี การออกเสียงของคำบาลีในภาษาไทยมักจะคล้ายกับเสียงในภาษาไทย ทำให้ง่ายต่อการออกเสียง ในขณะที่ ภาษาสันสกฤตการออกเสียงของคำสันสกฤตในภาษาไทยอาจจะมีความซับซ้อนมากขึ้น เนื่องจากมีเสียงที่ไม่คุ้นเคยในภาษาไทย (Bradley, 2024)

ดังนั้น ด้วยความซับซ้อนของภาษาจึงมีการประยุกต์เทคนิค และเครื่องมือทางด้านเทคโนโลยีมาประยุกต์ใช้ในการจำแนกภาษาบาลีและสันสกฤตด้วยการเรียนรู้ของเครื่อง (Machine learning) และการเรียนเชิงลึก (Deep learning) เพื่อช่วยให้การจำแนกภาษาเกิดความแม่นยำ และตรงตามหลักภาษาผ่านการผสานการเรียนรู้ของเครื่อง และการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ในภาษาบาลีสันสกฤต (Gudadhe et al., 2024) เพื่อค้นหาและดึงข้อมูลจากวรรณกรรม แยกความแตกต่างระหว่างความหมายของคำบาลี โดยใช้เครื่องมือและเทคโนโลยีการประมวลผลภาษาธรรมชาติที่ทำงานร่วมกันระหว่างศาสตร์ด้านภาษาและศาสตร์ด้านวิทยาการคอมพิวเตอร์ เพื่อถอดรหัสโครงสร้างภาษาบาลีและวลีทางไวยากรณ์ เพื่อสร้างแบบจำลองที่สามารถทำความเข้าใจ แยกย่อย และแยกรายละเอียดสำคัญได้อย่างเป็นระบบ จึงนำมาสู่วัตถุประสงค์การวิจัยเพื่อทดสอบ และเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกภาษาบาลีและสันสกฤตในภาษาไทยด้วยเทคนิคการเรียนรู้ของเครื่อง

2. วิธีดำเนินการวิจัย

การศึกษาครั้งนี้ใช้แนวทางการวิจัยเชิงทดลองผ่านการวิเคราะห์ข้อมูล เพื่อทดสอบพร้อมทั้งเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกภาษาบาลีและสันสกฤตในภาษาไทยด้วยเทคนิคการเรียนรู้ของเครื่อง โดยมีขั้นตอนในการวิจัยตามหลักวงจรชีวิตของการเรียนรู้ของเครื่อง (Machine learning life cycle) ดังรูปที่ 1



รูปที่ 1. กรอบแนวคิดในการดำเนินการวิจัย

2.1 การกำหนดปัญหา (Problem definition)

เนื่องด้วยภาษาบาลีและสันสกฤตเป็นคำที่เกี่ยวข้องกับศาสนา วัฒนธรรม และการศึกษา อีกทั้งภาษาไทยยังยืมคำมาจากทั้งสองภาษาโดยใช้ในจำนวนมาก ทำให้เกิดความซับซ้อนในการแยกแยะว่าคำใดมาจากภาษาใด ส่งผลให้ในการจำแนกคำบาลีและสันสกฤตต้องอาศัยทักษะและความเชี่ยวชาญ จึงนำมาสู่แนวคิดในการนำเครื่องมือทางด้านการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการจำแนก (Classification) โดยมีคำถามการวิจัยว่า อะไรคือแบบจำลองในการจำแนกภาษาบาลีและสันสกฤตในภาษาไทยโดยใช้เทคนิคการเรียนรู้ของเครื่องที่มีประสิทธิภาพดีที่สุด การวิจัยเริ่มต้นจากการศึกษาจากงานวิจัยที่เกี่ยวข้องในด้านการจำแนกภาษาต่าง ๆ โดยการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งสอดคล้องกับงานวิจัยของ Zhang et al. (2024) ที่ได้ประยุกต์นำเอาเทคนิคทางเทคโนโลยีมาใช้กับงานด้านภาษาศาสตร์และการเขียนเชิงวิชาการ โดยเปรียบเทียบระหว่างการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกด้วยแบบจำลองทางภาษาที่ผ่านการฝึกอบรมล่วงหน้า ได้แก่ Bidirectional Encoder Representations from Transformers (BERT) Robustly optimized BERT approach (RoBERTa) BigBird และ Long Document Transformer (Longformer) ผลการศึกษาบ่งชี้ว่า แบบจำลอง BigBird และ Longformer มีประสิทธิภาพเหนือกว่า BERT และ RoBERTa อย่างมีนัยสำคัญ

เช่นเดียวกับ Wang et al. (2024) ที่ได้ศึกษาและวิจัยเกี่ยวกับระบบอัจฉริยะโดยใช้แบบจำลองทางภาษาขนาดใหญ่สำหรับจำแนกข้อความ ประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง โดยใช้แบบจำลองป่าไม้สุ่ม (Random forest) แบบจำลองต้นไม้ตัดสินใจ (Decision tree) แบบจำลองเพื่อนบ้านใกล้สุด (K-Nearest Neighbors: KNN) แบบจำลองนาอีฟเบย์ (Naive Bayes) และแบบจำลองการถดถอยโลจิสติกส์ (Logistic Regression) เพื่อนำมาเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกที่ดีที่สุด ซึ่งเป็นไปในทิศทางเดียวกับงานวิจัยของ Brena et al. (2021) ที่ได้นำเอากระบวนการการประเมินประสิทธิภาพการพูดภาษาต่างประเทศโดยใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งเปรียบเทียบคลาสของการจำแนกตามการฝึก (Training) เพื่อให้เกิดความแม่นยำของแบบจำลอง ประกอบด้วยค่าความแม่นยำ (Accuracy) และความเที่ยงตรง (Precision) เป็นหลัก โดยทดสอบและเปรียบเทียบประสิทธิภาพแบบจำลองป่าไม้สุ่ม แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) และแบบจำลองโครงข่ายประสาทเทียม (Neural Network: NN) เช่นเดียวกับ Kashif et al. (2024) ที่ได้ศึกษาแบบจำลองการจำแนกหลายประเภทเพื่อใช้ในการระบุสำเนียงต่างประเทศโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน และแบบจำลองโครงข่ายประสาทเทียมในภาษาญี่ปุ่นที่มีความซับซ้อนทั้งนี้ได้ดำเนินการลดมิติของฟีเจอร์เพื่อคัดเลือกคุณลักษณะสำคัญ เพื่อประเมินประสิทธิภาพในการจำแนกข้อมูล

อีกทั้ง Joshi et al. (2024) ได้นำไปประยุกต์ใช้ในบริบทของภาษาถิ่นด้วยเทคนิคการประมวลผลภาษาธรรมชาติ ซึ่งแสดงให้เห็นว่าเครื่องมือทางเทคโนโลยีสามารถนำไปใช้กับงานด้านภาษาที่มีความซับซ้อน ซึ่งสอดคล้องกับงานวิจัย Saputra et al. (2024) ได้ศึกษาในครั้งนี้มุ่งเน้นการจำแนกปัจจัยที่ส่งผลต่อการพัฒนาความสามารถทางภาษาต่างประเทศในบริบทสังคม โดยใช้แบบจำลองต้นไม้การตัดสินใจในการวิเคราะห์ข้อมูล โดยมีตัวแปรอิสระ ได้แก่ ปัจจัยด้านสิ่งแวดล้อม แรงจูงใจ การใช้เทคโนโลยี และการปฏิบัติ ข้อมูลถูกรวบรวมผ่านแบบสอบถามที่ใช้มาตราส่วนลิเคิร์ทสเกล และวิเคราะห์ด้วยซอฟต์แวร์ RapidMiner เพื่อระบุปัจจัยที่มีอิทธิพลสูงสุดต่อการพัฒนาทักษะภาษาต่างประเทศ

ยังมีงานวิจัยของ Mousa et al. (2024) ที่ได้นำเอาแบบจำลองการเรียนรู้ของเครื่องสำหรับการตรวจจับภาษาอาหรับที่ไม่เหมาะสมในโซเชียลมีเดีย โดยผลการทดสอบประสิทธิภาพแบบจำลองสูงที่สุดคือ KNN หรือแม้แต่งานวิจัยในภาคธุรกิจ Makayasa et al. (2024) ได้นำเทคนิคการประมวลผลภาษาธรรมชาติมาจำแนกรูปแบบของภาษาในการบันทึกธุรกรรมทางการเงินทางบัญชี โดยเปรียบเทียบประสิทธิภาพของอัลกอริทึมการจำแนกประเภท 3 แบบ ได้แก่ แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองเพื่อนบ้านใกล้สุด และแบบจำลองป่าไม้สุ่ม โดยผลการวิจัยพบว่า แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน มีค่า Accuracy เท่ากับ 92.5% ค่า Precision เท่ากับ 92.5% ค่า Recall เท่ากับ 93.33% และค่า F1-score เท่ากับ 92% ซึ่งเป็นแบบจำลองที่มีประสิทธิภาพสูงสุดในการจำแนกภาษาในงานการเงินทางบัญชี

ในภาคการศึกษาได้มีการวิจัยของ Wei (2024) ได้นำเอาเทคนิคการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการจำแนกข้อความภาษาจีนและภาษาอังกฤษได้อย่างน่าเชื่อถือ โดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองเพื่อนบ้านใกล้สุด และแบบจำลองนาอีฟเบย์ ซึ่งแสดงให้เห็นว่า เทคนิคดังกล่าวมีความเหมาะสมกับภาษาต่างชาติ หรือภาษาที่มีลักษณะโครงสร้างที่มีความแตกต่างกัน เช่นเดียวกับงานวิจัย Alzoubi et al. (2023) ที่ได้ศึกษาเปรียบเทียบการจำแนกข้อความในภาษาตุรกี โดยอิงตามการเรียนรู้ของเครื่องโดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองนาอีฟเบย์ โครงข่ายประสาทเทียมแบบ Long Term-Short Memory (LTSM) แบบจำลองป่าไม้สุ่ม และแบบจำลองการถดถอยโลจิสติกส์ ในการจำแนกประเภทข้อมูลประเภทภาษาได้ และมีงานวิจัย Khairy et al. (2024) ได้ศึกษาการตรวจจับภาษาที่ไม่เหมาะสมทางอินเทอร์เน็ตในภาษาอาหรับโดยนำเอาการเรียนรู้ของเครื่องมาทดสอบ เพื่อเปรียบเทียบประสิทธิภาพการจำแนกโดยใช้แบบจำลองป่าไม้สุ่ม แบบจำลองต้นไม้ตัดสินใจ แบบจำลองเพื่อนบ้านใกล้สุด และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

อีกทั้ง Kaewpanitch & Tongngam (2020) ได้ศึกษาคุณลักษณะสำคัญที่ส่งผลต่อประสิทธิผลของการเรียนรู้ภาษาอังกฤษในหมู่นักศึกษาปริญญาตรี โดยใช้เทคนิคเหมืองข้อมูลและอัลกอริทึมการเรียนรู้ของเครื่อง เพื่อเปรียบเทียบประสิทธิภาพ ความแม่นยำ และความถูกต้องของอัลกอริทึมการเรียนรู้ของเครื่องผ่านแบบจำลองต้นไม้ตัดสินใจและแบบจำลองป่าไม้สุ่ม โดยแบบจำลองที่มีประสิทธิภาพสูงสุดคือแบบจำลองป่าไม้สุ่ม มีค่าความแม่นยำเท่ากับ 85.14% ซึ่งเป็นไปในทางเดียวกับ Maqsood et al. (2022) ที่ใช้การเรียนรู้ของเครื่องในการจำแนกความสามารถในการอ่านประโยคภาษาอังกฤษ โดยทดสอบประสิทธิภาพแบบจำลองการเรียนรู้ของเครื่อง 5 วิธี คือ แบบจำลองเพื่อนบ้านใกล้สุด แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองต้นไม้ตัดสินใจ แบบจำลองการถดถอยโลจิสติกส์ และแบบจำลองนาอีฟเบย์ ซึ่งเมื่อศึกษาในบริบทที่แตกต่างกันพบว่า งานวิจัย Kanraweekultana et al. (2024) ได้ศึกษาและวิจัยในด้านจิตวิทยา โดยนำเอาเทคนิคการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการจำแนกข้อมูลที่เป็นตัวอักษรที่เกิดขึ้นจากพฤติกรรมมนุษย์ที่มีความซับซ้อน และมีรูปแบบข้อมูลที่มีความคล้ายคลึงกัน โดยใช้แบบจำลองต้นไม้ตัดสินใจ แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองนาอีฟเบย์ และแบบจำลองเพื่อนบ้านใกล้สุด โดยวิธีที่ดีที่สุด คือ แบบจำลองต้นไม้ตัดสินใจให้ค่าความแม่นยำเท่ากับ 84.62%

รวมไปถึงภาษาต่างชาติที่ได้รับอิทธิพลเช่นเดียวกับภาษาไทย คือ ภาษาเขมร (Khmer language) งานวิจัย Phann et al. (2023) ได้ศึกษาเกี่ยวกับการจำแนกข้อความหลายคลาสในข่าวเขมรผ่านวิธีการ

เรียนรู้ของเครื่องเป็นตัวจำแนกกับงานด้านภาษา โดยใช้แบบจำลองการถดถอยโลจิสติกส์ แบบจำลองนาอีฟเบย์ แบบจำลองเพื่อนบ้านใกล้สุด แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองต้นไม้ตัดสินใจ และแบบจำลองป่าไม้สุ่ม ดังนั้น จากการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องกับงานวิจัยด้านการประสิทธิภาพแบบจำลองการจำแนกภาษาบาลีและสันสกฤตในภาษาไทย มักนำเอาเทคนิคการเรียนรู้ของเครื่องมาประยุกต์ใช้ในการจำแนกงานด้านภาษาที่มีความคล้ายคลึงและซับซ้อน สามารถสรุปแบบจำลองที่มีความเหมาะสมในการจำแนก ได้แก่ แบบจำลองป่าไม้สุ่ม แบบจำลองต้นไม้ตัดสินใจ แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด แบบจำลองนาอีฟเบย์ และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ดังรายละเอียดตารางที่ 1

ตารางที่ 1. สรุปผลการศึกษางานวิจัยที่เกี่ยวข้องกับ Machine Learning Classification Models

Research	Year	Algorithms						
		Random forest	Decision tree	K-Nearest Neighbors	Naive Bayes	SVM	Logistic Regression	Neural Network
Wang et al. (2024)	2024	✓	✓	✓	✓		✓	
Brena et al. (2021)	2021	✓				✓		✓
Kashif et al. (2024)	2024			✓		✓		✓
Joshi et al. (2024)	2024	✓				✓		✓
Saputra et al. (2024)	2024		✓					
Mousa et al. (2024)	2024			✓	✓	✓		
Makayasa et al. (2024)	2024	✓		✓		✓		
Wei (2024)	2024			✓	✓	✓		
Alzoubi et al. (2023)	2023	✓			✓	✓	✓	
Khairy et al. (2024)	2024	✓	✓	✓		✓		
Kaewpanitch & Tongngam (2020)	2020	✓	✓					
Maqsood et al. (2022)	2022		✓	✓	✓	✓	✓	
Kanraweekultana et al. (2024)	2024		✓		✓	✓		✓
Phann et al. (2023)	2023	✓	✓	✓	✓	✓	✓	
Result		**8**	**7**	**8**	**7**	**11**	4	4

2.2 การรวบรวมข้อมูล (Data collection)

ขั้นตอนการรวบรวมข้อมูลโดยสร้าง Dataset จากพจนานุกรมภาษาบาลีและสันสกฤตในภาษาไทยและเว็บไซต์พจนานุกรมภาษาไทยฉบับราชบัณฑิตสถาน ทั้งรูปแบบออนไลน์และออฟไลน์ โดยการรวบรวมชุดข้อมูลซึ่งเป็นการเก็บรวบรวมโดยใช้วิธีการคีย์ข้อมูลด้วยทีมนักวิจัยทั้งหมด 3 เล่ม ผ่านการเก็บรวบรวมด้วยทีมนักวิจัยได้ข้อมูลจำนวนทั้งสิ้น 21,664 records ด้วยการแบ่งข้อมูลแบบ Cross Validation (CV) ใช้

ชุดทดสอบ (Test Set) จำนวน 1 ชุดในการทดสอบประสิทธิภาพโดยแบ่งข้อมูลออกเป็นหลายชุดแล้วใช้ทุกชุดทั้งในการฝึก (Training) และทดสอบ (Testing) แบบจำลองในรอบที่แตกต่างกัน โดยมีข้อมูลทั้งหมด 3 Attribute ประกอบด้วย 1) คำศัพท์ (Vocabulary) 2) ประเภทของคำ (Part of speech) และ 3) ชนิดของคำภาษา (Language) โดยการทำความเข้าใจข้อมูล (Data understanding) ข้อมูลที่จะนำไปสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง เนื่องจากปัจจุบันยังไม่มีการพัฒนาคลังข้อมูล (Corpus) ขนาดใหญ่ที่รวบรวมคำบาลีและสันสกฤตที่ปรากฏในภาษาไทยอย่างเป็นระบบ ดังนั้นเทคนิคดังกล่าวจึงมีความเหมาะสม สามารถฝึกจากตัวอย่างคำที่มีการระบุชนิดของคำและที่มาไว้แล้ว และนำไปคาดการณ์กับคำใหม่ได้ คำบาลีและสันสกฤตที่ใช้ในภาษาไทยอาจมีการแปลงเสียงหรือเปลี่ยนรูป เช่น “ปัญญา” (บาลี) แปลงเป็น “ปั๊ญญา” (สำเนียงไทย) โดยการเรียนรู้ของเครื่องจะสามารถเรียนรู้ลักษณะการเปลี่ยนแปลงเสียงหรือรูปได้ ดังนั้น คำศัพท์ภาษาบาลีและสันสกฤตในภาษาไทยจะมีหลักการจำแนกโดยอาศัยประเภทของคำที่แบ่งตามชนิดของคำที่เป็นตัวจำแนก มีรายละเอียดข้อมูลสำหรับการวิจัยดังตารางที่ 2

ตารางที่ 2. รายละเอียดข้อมูล

ลำดับ	Attribute	ประเภทของข้อมูล	รายละเอียด
1.	Vocabulary	TEXT	คำศัพท์ภาษาบาลีและสันสกฤตในภาษาไทย
2.	Part of speech	TEXT	ประเภทของคำต่าง ๆ ภาษาบาลีและสันสกฤตในภาษาไทย
3.	Language (Label)	TEXT	ชนิดของคำ (4 Class) 1) คำบาลี เช่น สิขร วิริย ยุทธภูมิ มธุ พาหุ ปิติ ทุรชาติ เป็นต้น 2) คำสันสกฤต เช่น ศาป วิถี ปิฑ กษัย โศก ไทย ตุริ ทศ เป็นต้น 3) คำใช้ร่วมกัน เช่น เกศปาศ จมร นิวิต ปเร มานส เป็นต้น 4) คำพ้องรูป เช่น เกเศย ชิมุ ทฤฒ นลิน ศุนก อาภา เป็นต้น

2.3 การเตรียมข้อมูล (Data preparation)

จัดการข้อมูลและการเลือกใช้ข้อมูลเพื่อนำมาสู่วิธีการสร้างแบบจำลอง (Train model) จึงต้องจัดเตรียมชุดข้อมูลในรูปแบบของตาราง (Table) ไฟล์ .CSV โดยโครงสร้างข้อมูลแบบพื้นฐาน (Primitive data structure) อยู่ในรูปแบบตัวอักษร (Character) ที่เป็นไปตามลักษณะของข้อมูลเชิงโครงสร้าง (Structured data) และการแปลงข้อมูลเชิงตัวเลข (Numerical transformation) เมื่อเข้าสู่ขั้นตอน Modeling เพื่อนำข้อมูลไปแปลงเป็น Matrix แล้ว Train Model จำเป็นต้องมี 3 คุณสมบัติ 1) คำศัพท์ 2) ประเภทของคำ 3) ชนิดของคำ เป็น Feature Vector ในการวิเคราะห์ภาษาธรรมชาติ (NLP) เพื่อช่วยให้โมเดลเข้าใจโครงสร้างภาษาและความหมายของคำในบริบทที่ดีขึ้น โดยใช้วิธีการแทนค่าด้วย One-hot Encoding แต่ละประเภทของคำ ตัวอย่างเช่น Noun Verb Adjective เมื่อถูกเข้ารหัสเป็นเวกเตอร์ที่มีค่า 1 ในตำแหน่งที่แทนชนิดคำนั้น และค่า 0 ที่เหลือ ถ้ามี Part of Speech เมื่อเป็นเวกเตอร์จะแปลงค่าเป็น [0,1,0,0] โดยที่ข้อมูลทั้งหมดต้องไม่มีค่า Null ส่วนการคัดกรองข้อมูล (Filter data) ใช้วิธีการกรองร่วมกัน (Collaborative filtering) ที่เป็นเทคนิคแนะนำข้อมูลตามความคล้ายคลึงของภาษาสามารถสรุปได้ดังนี้ 1) ข้อมูลของแต่ละคำจะต้องประกอบด้วยตัวอักษรภาษาไทยทั้งหมด 2) พื้นที่ข้อมูลคุณสมบัติของแต่ละคำ

ต้องไม่มีค่าว่าง หรือค่าที่ไม่ตรงกับที่กำหนดไว้ หลังจากนั้นจะเข้าสู่ขั้นตอนการแปลงข้อมูล (Data transformation) โดยแบ่งเป็น 2 ขั้นตอน ดังนี้

1) การรวมข้อมูล (Data aggregation) จากพจนานุกรมภาษาบาลีและสันสกฤตในภาษาไทย และเว็บไซต์พจนานุกรมภาษาไทยฉบับราชบัณฑิตสถาน ทั้งในรูปแบบออนไลน์และออฟไลน์ โดยประกอบด้วยคำบาลี เช่น จิต - จิตใจ (Citta - Mind) ธรรม – หลักความจริง คำสอน (Dhamma - Truth, Teaching) กรุณา - ความสงสาร (Karu**ṇā** - Compassion) ปัญญา - ความรู้ (Pañ**ñā** - Wisdom) เป็นต้น ในขณะที่ภาษาสันสกฤต เช่น กรรม - การกระทำ (Karma - Action) ราชา - กษัตริย์ (R**ā**ja - King) ศิลป์ - งานศิลปะ (**ś**ilpa - Art) ภาษา - การพูด (Bh**āṣā** - Language) เป็นต้น แต่เนื่องจากคำในภาษาบาลีและภาษาสันสกฤต จะมีบางคำที่เป็นคำใช้ร่วมกัน ซึ่งหากมีข้อมูลคำที่ซ้ำกันแต่มี Class ที่ต่างกัน จะทำให้เกิดการชนกันของข้อมูล (Data collision) โดยการจะทำขั้นตอนการรวมข้อมูลนั้น ต้องรวมคำภาษาบาลีและภาษาสันกฤตที่ใช้ร่วมกัน และมีส่วนของคำพูด เหมือนกันให้กลายเป็น Class “ใช้ร่วมกัน” และคำภาษาบาลีและภาษาสันสกฤตที่ใช้ร่วมกัน เช่น กฐิน ขมา ครรภ์ ฆราวาส จันทน์ เป็นต้น และมีส่วนของคำพูด ต่างกันให้กลายเป็น Class “คำพ้องรูป” เมื่อไม่มีการชนกันของข้อมูล เช่น เกต เศาฑฤ ครุฑ ปเภท ปกาสโก เป็นต้น โดยการขอความเห็นจากนักภาษาศาสตร์ นักวิชาการ ครู และอาจารย์ภาษาไทยที่เกี่ยวข้อง เพื่อช่วยให้ความผิดพลาดในการจัดการ Label โดยข้อมูลมีความถูกต้องตามหลักภาษาศาสตร์ อีกทั้งช่วยให้การจำแนกข้อมูลจะมีความแม่นยำมากขึ้น

2) การสร้างแผนผังข้อมูล (Data mapping) ในการจำแนกภาษาด้วยการเรียนรู้ของเครื่อง ด้วยการแปลงค่าหมวดหมู่ให้เป็นตัวเลข (Integer) การปรับโครงสร้างข้อมูลโดยการแปลงข้อมูลเชิงตัวเลข โดยการปรับเปลี่ยนค่าของข้อมูลเชิงตัวเลขให้อยู่ในรูปแบบที่เหมาะสมกับการวิเคราะห์หรือการสร้างโมเดลทางสถิติและการเรียนรู้ของเครื่องผ่านการปรับแต่ง Features (Feature engineering) ในชุดข้อมูล เพื่อให้แบบจำลองการเรียนรู้ของเครื่องสามารถการจัดการค่าที่หายไป (Missing Values) และลบหรือแทนค่าผิดปกติ (Outliers) จากนั้นการแปลงข้อมูล (Feature transformation) ประเภทข้อความให้เป็นตัวเลข จากนั้นสร้าง Feature ใหม่ (Feature Creation) ด้วยวิธีการแปลงค่าตัวเลขให้ค่าของข้อมูลอยู่ในช่วงที่กำหนด เช่น [0,1] หรือ [-1,1] ที่เป็นไปตามรูปแบบการปรับขนาดข้อมูล (Normalization) โดยวิธีการแปลงข้อมูลให้เป็นระดับคำ (word level) เพื่อให้โมเดลสามารถเรียนรู้จากหน่วยข้อมูลที่เป็นคำสำหรับแปลงข้อมูลเป็นระดับ subword เพื่อรับมือกับคำที่ไม่รู้จักในภาษา สำหรับการเชื่อมโยงกับ Part of Speech ที่เป็นตัวช่วยจำแนกข้อมูลภายใต้อัตลักษณ์ทางภาษาบางประการที่จะช่วยให้แบบจำลองสามารถอ่าน และประมวลผลแบบจำลองทั้งหมดผ่านการทำ Normalization เพื่อปรับค่าข้อมูลให้อยู่ในช่วงที่เหมาะสม เพื่อให้การวิเคราะห์หรือการนำไปใช้แบบจำลองการเรียนรู้ของเครื่องมีประสิทธิภาพมากขึ้น ลดผลกระทบจากขนาดของข้อมูลที่แตกต่างกันของคำบาลี คำสันสกฤต คำใช้ร่วมกัน และคำพ้องรูป ที่สามารถช่วยให้โมเดลเรียนรู้ได้ดีขึ้น

2.4 การสร้างแบบจำลอง (Modeling)

การออกแบบจำลองตามการจำแนกประเภทข้อมูล มาวิเคราะห์โดยใช้เทคนิคการเรียนรู้ของเครื่องผ่านเครื่องมือ RapidMiner Studio ประกอบด้วย 5 แบบจำลอง คือ แบบจำลองป่าไม้สุ่ม แบบจำลอง

ต้นไม้ตัดสินใจ แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด แบบจำลองนาอีฟเบย์ และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ซึ่งก่อนการวิเคราะห์ได้ปรับโครงสร้างข้อมูลให้มีความเหมาะสม และสามารถใช้งานร่วมกันกับทั้ง 5 แบบจำลองตามขั้นตอนวิธีการในการปรับสร้างแผนผังข้อมูล และใช้การวิเคราะห์ความเที่ยงตรงของแบบจำลอง (K-fold cross validation) ในการตรวจสอบไขว้ (Cross validation) เป็นวิธีการตรวจสอบความผิดพลาด ในการคาดการณ์ของแบบจำลอง โดยลักษณะวิธี การตรวจสอบไขว้กันด้วยวิธีการ 10-fold Cross-validation ซึ่งจะแบ่งข้อมูลเป็น 10 ชุดเท่ากัน แล้วให้ 1 ชุดเป็นกลุ่มทดสอบ (Test set) และอีก 9 ชุดเป็นกลุ่มฝึกฝน (Train set) สำหรับการวัดประสิทธิภาพแบบจำลองของการจำแนกข้อมูล สำหรับเปรียบเทียบประสิทธิภาพของแบบจำลอง โดยมีรายละเอียดดังนี้

1) แบบจำลองป่าไม้สุ่ม เป็นแบบจำลองที่สร้างขึ้นจากการรวมกันของต้นไม้ตัดสินใจ หลายๆ ต้น โดยใช้หลักการ Ensemble Learning แบบ Bagging (Bootstrap Aggregating) โดยการทำงานจะ Bootstrap Sampling สำหรับต้นไม้แต่ละต้น t ใน T ต้น ผ่านการวิเคราะห์นี้มีการกำหนดพารามิเตอร์ของแบบจำลอง Number of tree = 100 เพื่อให้แบบจำลองสมดุลระหว่างความแม่นยำและเวลาในการคำนวณโดยไม่ Overfitting และ Criterion depth = gini_index เพื่อวัดความไม่บริสุทธิ์ของข้อมูลในแต่ละโหนดของต้นไม้การตัดสินใจ อีกทั้งเพื่อป้องกันการเกิด Overfitting โดยจำกัดความซับซ้อนของแบบจำลองให้ไม่เกิน 10 ระดับกำหนด Maximum depth = 10 และ Voting strategy = Majority vote ช่วยให้แบบจำลองป่าไม้สุ่ม มีความแม่นยำสูงขึ้น (Kanraweekultana et al., 2024) โดยมีกระบวนการดังสมการที่ (1) ถึงสมการที่ (6)

Gini Impurity
$$G(S) = 1 - \sum_{i=1}^{c} p_i^2 \quad (1)$$

Reduction in Gini Impurity
$$\Delta G = G(S) - \sum_{j=1}^{k} \frac{|S_j|}{|S|} G(S_j) \quad (2)$$

Leaf Node Classification
$$\hat{y} = arg\max_i |S_i| \quad (3)$$

Stopping Criteria
$$G(S) = 0 \quad (4)$$

Complexity Pruning
$$C(T) = \sum_{t \in T_L} [|S_t| \cdot Impurity(S_t)] + \alpha |T_L| \quad (5)$$

Majority voting
$$\hat{y} = argmax_{C_j \in \{C_1, C_2, \ldots, C_n\}} \sum_{i=1}^{k} I(y_i = C_j) \quad (6)$$

2) แบบจำลองต้นไม้ตัดสินใจ เป็นหนึ่งในอัลกอริทึมที่ใช้งานง่ายและแปลผลได้ชัดเจน ในแต่ละโหนดต้นไม้จะเลือกคุณลักษณะ (Feature) และค่าที่เหมาะสมที่สุดเพื่อแบ่งข้อมูล โดยใช้ตัววัดความบริสุทธิ์ (Purity measures) การวิเคราะห์นี้ได้กำหนดค่าพารามิเตอร์ของโมเดลเป็น Criterion = gini_index เพื่อวัดความไม่แน่นอนหรือความหลากหลายในชุดข้อมูล (Impurity) และใช้ max-depth = 10 เป็นการเริ่มต้น

ที่ปลอดภัย ซึ่งช่วยให้โมเดลมีประสิทธิภาพในการเรียนรู้ข้อมูลที่ไม่ซับซ้อนเกินไป อีกทั้งเป็นการป้องกันปัญหา Overfitting หรือ Underfitting กำหนดให้ Confidence = 0.1 เป็นค่าสมดุลที่ช่วยให้ต้นไม้ไม่ซับซ้อนเกินไป แต่ยังคงรักษาความแม่นยำ (Modhugu & Ponnusamy, 2024) โดยมีกระบวนการดังสมการที่ (7) – (9)

Gini Impurity $$Gini(t) = 1 - \sum_{i=1}^{C} p_i^2 \quad (7)$$

Entropy $$Entropy(S) = -\sum_{i=1}^{C} p_i log_2(p_i) \quad (8)$$

Information Gain $$IG(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \quad (9)$$

3) แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด เป็นอัลกอริทึมแบบ Instance-based learning หรือ Lazy learning โดยกำหนดค่าพารามิเตอร์ของแบบจำลองเป็น k = 5 โดยเลือกจุดข้อมูล 5 จุดที่อยู่ใกล้กับ $\boldsymbol{x}$ มากที่สุด (จากการคำนวณระยะทาง) และใช้เป็นเกณฑ์ในการตัดสินใจว่าคลาสใดปรากฏมากที่สุด (Majority voting) กำหนด Measure types = MixedMeasures เนื่องจากจำเป็นต้องรวมคุณลักษณะทุกประเภทเข้าด้วยกันตามลักษณะของข้อมูลที่มีตัวแปร (Perroni et al., 2024) โดยมีกระบวนการดังสมการที่ (10) และสมการที่ (11)

Distance Calculation $$d(x, x_i) = \sqrt{\sum_{j=1}^{m} (x_j - x_{ij})^2} \quad (10)$$

Majority Voting $$Class(x) = \underset{C}{argmax} \sum_{i=1}^{k} I(C_i = C) \quad (11)$$

4) แบบจำลองนาอีฟเบย์ การวิเคราะห์นี้กำหนดพารามิเตอร์ของแบบจำลองเป็น Laplace correlation = Smoothing เพื่อปรับค่าความน่าจะเป็นในการคำนวณความน่าจะเป็นของแต่ละฟีเจอร์ภายใต้คลาส โดยใช้ทฤษฎีเบย์ในการคำนวณความน่าจะเป็นของคลาสเป้าหมาย (y) เมื่อกำหนดชุดคุณลักษณะ $x = (x_1, x_2, \dots, x_n)$ (Hemal et al., 2024) โดยมีกระบวนการดังสมการที่ (12)

Naïve Bayes classification $$P(C_k|X) = \frac{P(C_k) \prod_{i=1}^{n} P(x_i|C_k)}{P(X)} \quad (12)$$

5) แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน พารามิเตอร์ของแบบจำลองถูกตั้งค่า Kernel type = dot สำหรับการวิเคราะห์เคอร์เนลเชิงเส้น Kernel cache = 200 เพื่อกำหนดขนาดหน่วยความจำที่ใช้เก็บค่า kernel สำหรับการวิเคราะห์ SVM convergence epsilon = 0.001 การฝึกหยุดเมื่อความแตกต่างระหว่างค่าในการวนซ้ำปัจจุบัน และการวนซ้ำก่อนหน้าน้อยกว่าหรือเท่ากับ 0.001 และ Max iterations = 10,000 แบบจำลองจะฝึกซ้ำสูงสุด 10,000 ครั้ง ก่อนหยุดการฝึก (Dinesh et al., 2024) โดยมีกระบวนการดังสมการที่ (13) – (15)

Decision Function
$$f(x) = w^T x + b \tag{13}$$

Optimization
$$\min_{w,b} \frac{1}{n} \sum_{i=1}^{n} (y_i - (w^T x_i + b))^2 \tag{14}$$

Polynomial Kernel
$$K(x, z) = (x^t z + c)^d \tag{15}$$

2.5 การวัดประสิทธิภาพของแบบจำลอง (Evaluation)

การวัดประสิทธิภาพแบบจำลอง เพื่อวัดว่าแบบจำลองมีประสิทธิภาพเพียงพอต่อการนำไปใช้งานหรือไม่ โดยนำชุดข้อมูลที่ใช้ในการฝึกฝนข้อมูล (Training data) กับแบบจำลองของการเรียนรู้ของเครื่องเพื่อให้แบบจำลองสามารถเรียนรู้รูปแบบและความสัมพันธ์ในข้อมูล หลังจากการฝึกฝนข้อมูล แล้วแบบจำลองจะต้องถูกทดสอบข้อมูล (Testing data) เพื่อนำมาประเมินประสิทธิภาพของแบบจำลองในการจำแนกภาษาบาลีและสันกฤตในภาษาไทย ประกอบด้วย

1) ค่าความถูกต้อง (Accuracy) การหาค่าความถูกต้องของแบบจำลอง ซึ่งจะพิจารณารวมทุก Class ที่เกี่ยวข้องมีสมการที่ (16)

$$\text{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{True Positive+False Positive + True Negative+ False Negative}} \tag{16}$$

2) ค่าความแม่นยำ (Precision) การหาค่าความแม่นยำของแบบจำลอง โดยการพิจารณาแยกเป็นราย Class มีสมการที่ (17)

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{17}$$

3) ค่าความระลึก (Recall) การวัดค่าความถูกต้องของแบบจำลอง โดยพิจารณาแยกเป็นราย Class มีสมการที่ (18)

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative"}} \tag{18}$$

4) ค่าความถ่วงดุล (F1-Score) การวัดค่า Precision และ Recall พร้อมกันของแบบจำลอง โดยพิจารณาแยกเป็นราย Class มีสมการที่ (19)

$$\mathrm{F1-Score} = \frac{2\times \mathrm{Precision} \times \mathrm{Recal}}{\mathrm{Precision+Recall}} \quad (19)$$

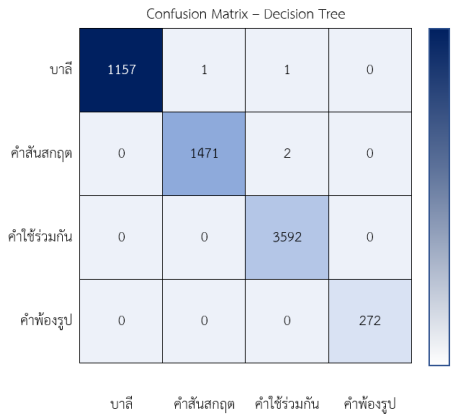
2.6 การนำแบบจำลองไปใช้งานจริง (Deployment)

การนำแบบจำลองไปใช้กับข้อมูลชุดทดสอบเพื่อประเมินประสิทธิภาพของแบบจำลอง และปรับแต่งพารามิเตอร์ (Parameter optimization) โดยการปรับปรุงแบบจำลองผ่านการ Parameter tuning ด้วยวิธีการปรับพารามิเตอร์ที่ใช้ Cross-validation ช่วยในการทดสอบความแม่นยำของแบบจำลองในจำนวน K-fold ที่แตกต่างกันในแต่ละรอบ เพื่อป้องกันการ Overfitting โดยกำหนดให้ k=10 เพื่อสร้างความสมดุลระหว่าง Bias และ Variance เนื่องจาก Dataset มีขนาดของข้อมูลไม่สมดุล จึงปรับให้เหมาะสมด้วยการทดสอบหลายค่า และประสิทธิภาพการทำนายหรือผลลัพธ์ที่ดีที่สุด ก่อนนำแบบจำลองที่ฝึกเสร็จไปใช้งานจริงด้วยการนำเสนอแบบจำลองที่มีความแม่นยำที่ดีที่สุด และเหมาะสมสำหรับการจำแนกภาษาบาลีและสันสกฤตในภาษาไทยต่อไป

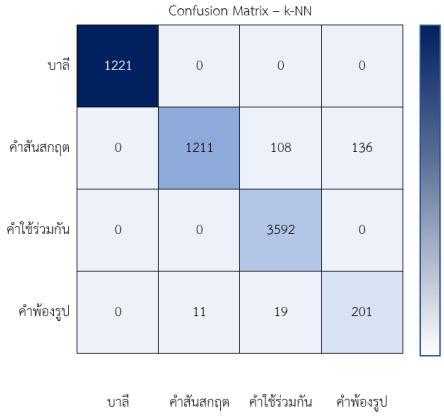
3. ผลการวิจัยและอภิปรายผล

ผลการเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนกภาษาบาลีและสันสกฤตในภาษาไทยด้วยเทคนิคการเรียนรู้ของเครื่อง ประกอบด้วยแบบจำลองป่าไม้สุ่ม แบบจำลองต้นไม้ตัดสินใจ แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด แบบจำลองนาอีฟเบย์ และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ด้วยวิธีการ 10-fold cross validation และการวิเคราะห์หาประสิทธิภาพด้วยการหาค่า Accuracy ค่า Precision ค่า Recall และค่า F1-Score เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลอง ผลการวิเคราะห์พบว่าแบบจำลองที่มีประสิทธิภาพสูงที่สุดคือ 1) แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน มีค่า Accuracy เท่ากับร้อยละ 95.75 ค่า Precision เท่ากับร้อยละ 90.90 ค่า Recall เท่ากับร้อยละ 91.81 และค่า F1-Score เท่ากับร้อยละ 90.90 ตามผล Confusion matrix ดังรูปที่ 1(จ)

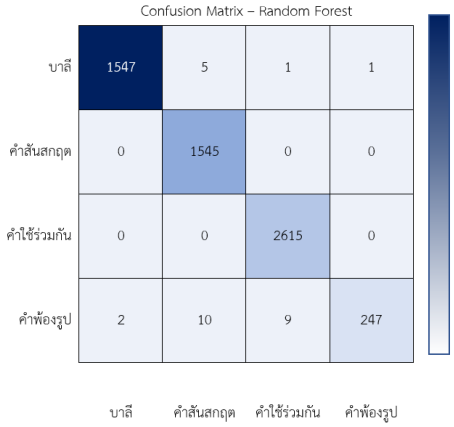
รองลงมาคือ 2) แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด มีค่า Accuracy เท่ากับร้อยละ 92.86 ค่า Precision เท่ากับร้อยละ 90.88 ค่า Recall เท่ากับร้อยละ 92.37 และค่า F1-Score เท่ากับร้อยละ 90.88 ตามผล Confusion matrix ดังรูปที่ 1(ข) ถัดมาคือ 3) แบบจำลองนาอีฟเบย์ มีค่า Accuracy เท่ากับร้อยละ 81.29 ค่า Precision เท่ากับร้อยละ 59.47 ค่า Recall เท่ากับร้อยละ 51.14 และค่า F1-Score เท่ากับร้อยละ 59.47 ตามรูปที่ 1(ง) ถัดมาคือ 4) แบบจำลองต้นไม้ตัดสินใจ มีค่า Accuracy เท่ากับร้อยละ 77.97 ค่า Precision เท่ากับร้อยละ 73.59 ค่า Recall เท่ากับร้อยละ 65.02 และค่า F1-Score เท่ากับร้อยละ 73.59 ตามผล Confusion matrix ดังรูปที่ 1(ก) และสุดท้ายคือ 5) แบบจำลองป่าไม้สุ่ม มีค่า Accuracy เท่ากับร้อยละ 55.27 ค่า Precision เท่ากับร้อยละ 35.25 ค่า Recall เท่ากับร้อยละ 25.00 และค่า F1-Score เท่ากับร้อยละ 35.25 ตามผล Confusion matrix ดังรูปที่ 1(ค) ตามผลการเปรียบเทียบดังในตารางที่ 3



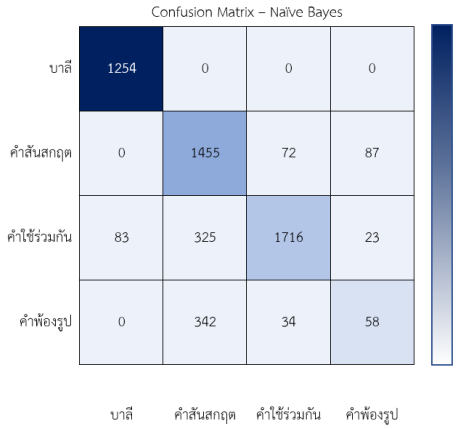
(ก) แบบจำลองต้นไม้ตัดสินใจ



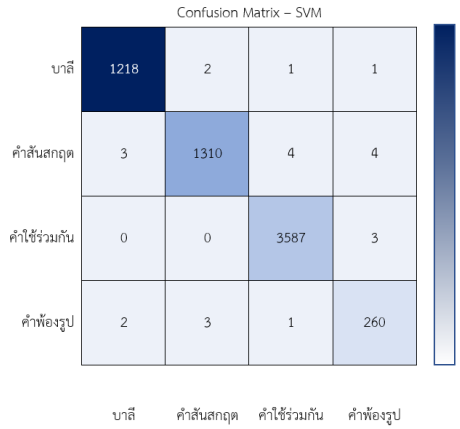
(ข) แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด



(ค) แบบจำลองป่าไม้สุ่ม



(ง) แบบจำลองนาอีฟเบย์



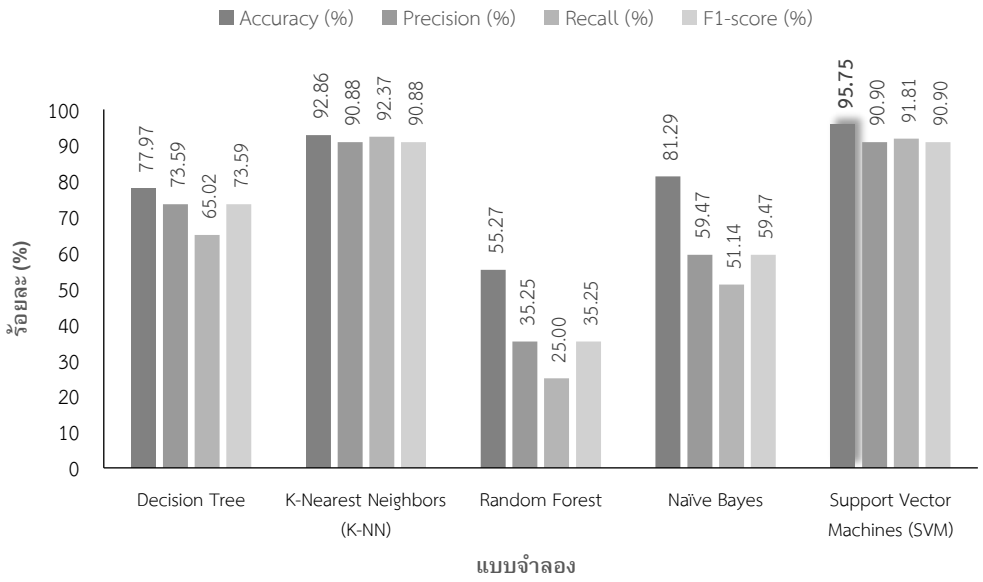
(จ) แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

รูปที่ 1. Confusion matrix ของแบบจำลองต่าง ๆ

ตารางที่ 3. ผลการทดสอบประสิทธิภาพแบบจำลอง

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
แบบจำลองต้นไม้ตัดสินใจ	77.97	73.59	65.02	73.59
แบบจำลองการคำนวณเพื่อนบ้านใกล้สุด	92.86	90.88	92.37	90.88
แบบจำลองป่าไม้สุ่ม	55.27	35.25	25.00	35.25
แบบจำลองนาอีฟเบย์	81.29	59.47	51.14	59.47
แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน	**95.75**	**90.90**	**91.81**	**90.90**

จากผลการทดลองข้อมูลที่ได้จากการวิเคราะห์สามารถสะท้อนถึงความเฉพาะตัว และความซับซ้อนของภาษาบาลีและสันสกฤตในภาษาไทย ซึ่งความแตกต่างระหว่าง 2 ภาษา คือ ที่มาของคำศัพท์ การใช้ในภาษาไทย และโครงสร้างทางไวยากรณ์ ทั้ง 2 ภาษานี้มีบทบาทสำคัญในการพัฒนาภาษาไทยและวัฒนธรรมไทยอย่างมาก ผลลัพธ์ดังกล่าวแสดงให้เห็นว่า แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนมีความสามารถในการจำแนกภาษาบาลีและสันสกฤตในภาษาไทยได้อย่างมีประสิทธิภาพมากที่สุด มีค่า Accuracy เท่ากับร้อยละ 95.75 ค่า Precision เท่ากับร้อยละ 90.90 ค่า Recall เท่ากับร้อยละ 91.81 และค่า F1-Score เท่ากับร้อยละ 90.90 เมื่อเทียบกับแบบจำลองอื่น ๆ ที่ใช้ในการทดลอง โดยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนมีความแม่นยำสูงและยังคงรักษาสมดุลระหว่างค่าความแม่นยำ และความระลึก ซึ่งสะท้อนให้เห็นถึงความสามารถของแบบจำลองในการระบุข้อความที่เป็นภาษาที่ถูกต้องได้ โดยไม่เกิดการทำนายที่ผิดพลาด และได้การปรับปรุงแบบจำลองผ่านการ Parameter tuning โดยใช้ Cross-validation โดยใช้เทคนิค 10-fold cross-validation ในการประเมินผลช่วยลดปัญหาการ Overfitting และเพิ่มความน่าเชื่อถือของผลลัพธ์สำหรับการจัดการกับปัญหาที่มีลักษณะของข้อมูลเชิงภาษา (linguistic data) และการจำแนกประเภทของภาษาที่ใกล้เคียงกัน โดยผลลัพธ์นี้สามารถนำไปพัฒนาและประยุกต์ใช้ในการจำแนกภาษาอื่น ๆ หรืองานที่มีลักษณะคล้ายคลึงกันต่อไป ดังผลการเปรียบเทียบตามรูปที่ 2



รูปที่ 2. แผนภาพเปรียบเทียบประสิทธิภาพแบบจำลองการจำแนก

4. สรุปผลการวิจัย

การศึกษาวิจัยครั้งนี้มีจุดมุ่งหมายในการประเมินประสิทธิภาพของแบบจำลองที่ใช้ในการจำแนกคำบาลีและสันสกฤตในภาษาไทย ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยเปรียบเทียบประสิทธิภาพของแบบจำลองทั้ง 5 แบบ ได้แก่ แบบจำลองต้นไม้ตัดสินใจ แบบจำลองป่าไม้สุ่ม แบบจำลองเพื่อนบ้านใกล้ที่สุด แบบจำลองนาอีฟเบย์ และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ผ่านกระบวนการตรวจสอบด้วยวิธี 10-fold cross-validation เพื่อประเมินความสามารถในการจำแนกคำของแต่ละแบบจำลอง

ผลการทดลองชี้ว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนมีความสามารถสูงที่สุด โดยมีค่าความถูกต้องอยู่ที่ 95.75% และค่าความแม่นยำสูงสุดที่ 90.9% ขณะที่แบบจำลองป่าไม้สุ่มเป็นแบบจำลองที่มีประสิทธิภาพต่ำที่สุด แสดงให้เห็นว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน มีศักยภาพในการจัดการกับความใกล้เคียงกันของภาษาบาลีและสันสกฤตในภาษาไทยได้อย่างมีประสิทธิภาพ ซึ่งภาษาทั้งสองมีคำที่ใช้ร่วมกันจำนวนมากและอาจทำให้เกิดความสับสน การใช้เทคโนโลยีเข้ามาช่วยจำแนกจึงสามารถลดความคลาดเคลื่อนและเพิ่มความแม่นยำได้อย่างชัดเจน เมื่อเปรียบเทียบกับงานวิจัยก่อนหน้า ผลลัพธ์ของงานนี้สอดคล้องกับงานของ Brena et al. (2021) ที่ประยุกต์ใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ในการจำแนกภาษาต่างประเทศ รวมทั้งงานของ Saputra et al. (2024) ซึ่งศึกษาการพัฒนาทักษะภาษาต่างประเทศผ่านการวิเคราะห์ด้วยโปรแกรม RapidMiner เช่นเดียวกับในงานนี้ นอกจากนี้ยังมีงานที่ใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ในการวิเคราะห์ข้อมูลภาษาธรรมชาติที่มีความเฉพาะ เช่น การจำแนกข้อความในบันทึกบัญชีการเงิน อีกทั้งผลลัพธ์จากงานวิจัยนี้ยังตรงกับการศึกษาของ Wei (2024) ที่ใช้ SVM จำแนกข้อความระหว่างภาษาจีนและอังกฤษ ซึ่งยืนยันว่าแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

เหมาะสมกับภาษาโครงสร้างซับซ้อน เช่นเดียวกับงานของ Alzoubi et al. (2023) ที่นำมาใช้กับภาษาในตุรกี และ Maqsood et al. (2022) ที่ใช้ในการวิเคราะห์ความสามารถในการอ่านภาษาอังกฤษ นอกจากนี้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนยังถูกนำไปประยุกต์ใช้ในการวิเคราะห์พฤติกรรมมนุษย์ในมิติทางจิตวิทยา ผ่านการจำแนกข้อมูลอักษรที่มีรูปแบบคล้ายคลึงกันอีกด้วย ท้ายที่สุด งานของ Phann et al. (2023) ที่นำแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนมาใช้กับภาษาเขมร ซึ่งได้รับอิทธิพลจากภาษาไทย ก็สะท้อนให้เห็นถึงศักยภาพของแบบจำลองนี้ในการจำแนกภาษาต่างชาติ หรือภาษาที่มีความซับซ้อนสูงได้อย่างมีประสิทธิภาพ ดังนั้น ความสามารถของแบบจำลองซัพพอร์ตเวกเตอร์แมชชีนในการรักษาสมดุลระหว่างความแม่นยำและความครอบคลุมในการจำแนกจึงส่งผลให้เกิดผลลัพธ์ที่น่าเชื่อถือ ลดข้อผิดพลาดจากการทำนาย และเหมาะสมกับงานด้านภาษาศาสตร์

อย่างไรก็ตาม งานวิจัยนี้ยังพบข้อจำกัดบางประการ เช่น ประสิทธิภาพที่ต่ำกว่าของแบบจำลองป่าไม้สุ่ม ซึ่งอาจมาจากการจัดการกับข้อมูลภาษาศาสตร์ที่ซับซ้อนได้ไม่ดีพอ อีกทั้งการเลือกใช้พารามิเตอร์และข้อมูลยังมีความสำคัญต่อผลลัพธ์ แม้การทำ Cross-validation จะช่วยลดความเสี่ยงของ Overfitting ได้ในระดับหนึ่งก็ตาม ดังนั้นการใช้เทคนิคการเรียนรู้ของเครื่องในงานวิจัยภาษาศาสตร์ลักษณะนี้จึงมีความเหมาะสมกับข้อมูลขนาดไม่ใหญ่และไม่ซับซ้อนมาก ทำให้สามารถอธิบายผลลัพธ์ได้อย่างชัดเจน ทั้งนี้ในอนาคตสามารถขยายการวิจัยไปยังข้อมูลที่มีขนาดใหญ่ขึ้น หรือประยุกต์ใช้การเรียนรู้เชิงลึก เพื่อศึกษาความสามารถเปรียบเทียบของเทคนิคประมวลผลภาษาธรรมชาติให้แม่นยำยิ่งขึ้น รวมถึงนำข้อมูลที่ได้ไปพัฒนาเครื่องมือด้านการสอนภาษา การแปลภาษา และขยายผลไปสู่การศึกษาภาษาต่างประเทศอื่น ๆ ซึ่งจะช่วยส่งเสริมการวิจัยทั้งในด้านภาษาศาสตร์ และเทคโนโลยีการเรียนรู้ของเครื่องอย่างมีประสิทธิภาพตลอดจนสร้างคลังข้อมูลคำบาลีและสันสกฤตในภาษาไทยสำหรับใช้งานในอนาคตต่อไป

เอกสารอ้างอิง (References)

Alzoubi, Y. I., Topcu, A. E., & Erkaya, A. E. (2023). Machine learning-based text classification comparison: Turkish language context. *Applied Sciences, 13*(16), Article 9428. https://doi.org/10.3390/app13169428

Bradley, D. (2024). Sociolinguistics in Mainland Southeast Asia. In M. J. Ball, R. Mesthrie & C. Meluzzi (Eds.), The Routledge handbook of sociolinguistics around the world (2nd ed., pp. 227-237). Routledge. https://doi.org/10.4324/9781003198345-21

Brena, R. F., Zuvirie, E., Preciado, A., Valdiviezo, A., Gonzalez-Mendoza, M., & Zozaya-Gorostiza, C. (2021). Automated evaluation of foreign language speaking performance with machine learning. *International Journal on Interactive Design and Manufacturing (IJIDeM), 15*(2), 317-331. https://doi.org/10.1007/s12008-021-00759-z

Dinesh, P., Vickram, A. S., & Kalyanasundaram, P. (2024). Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM,

KNN, logistic regression, random forest and decision tree to measure accuracy. *AIP Conference Proceedings, 2853*(1). Article 020140. https://doi.org/10.1063/5.0203746

Duraiswamy, D. (2024). Cultural and trade links between India and Siam: Their impact on the Maritime Silk Road. *Acta Via Serica, 9*(1), 67-90. https://doi.org/10.22679/avs.2024.9.1.003

Gudadhe, S. R., Bardekar, A. A., & Ranit, A. B. (2024). Integrating machine learning and NLP: Efficient retrieval of characters in Pali script preservation. *Juni Khyat Journal (जून ख्यात), 14*(4), 94-103.

Hemal, S. H., Khan, M. A. R., Ahammad, I., Rahman, M., Khan, M. A. S. D., & Ejaz, S. (2024). Predicting the impact of internet usage on students' academic performance using machine learning techniques in Bangladesh perspective. *Social Network Analysis & Mining, 14*(1), Article 66. https://doi.org/10.1007/s13278-024-01234-9

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., & Dippold, D. (2024). *Natural language processing for dialects of a language: A survey*. arXiv. https://doi.org/10.48550/arXiv.2401.05632

Kaewpanitch, C., & Tongngam, S. (2020). Analysis of key features affecting the effectiveness of English language learning among undergraduate students utilizing the data mining techniques. *NIDA Development Journal, 60*(1-2), 1-29. (in Thai)

Kanraweekultana, N., Waijanya, S., Promrit, N., Nopnapaporn, U., Korsanan, A., & Poolphol, S. (2024). Comparison of capability of data classification models to predict consistent results for depression analysis based on user-behaviour tracking and facial expression recognition during PHQ-9 assessment. *Engineering and Applied Science Research, 51*(1), 11-21. https://doi.org/10.14456/easr.2024.2

Kashif, K., Alwan, A., Wu, Y., De Nardis, L., & Di Benedetto, M. G. (2024). MKELM based multi-classification model for foreign accent identification. *Heliyon, 10*(16), Article e36460. https://doi.org/10.1016/j.heliyon.2024.e36460

Khairy, M., Mahmoud, T. M., Omar, A., & Abd El-Hafeez, T. (2024). Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection. *Language Resources and Evaluation, 58*(2), 695-712. https://doi.org/10.1007/s10579-023-09683-y

Li, A. (2024). Nominalizations and its grammaticalization in Standard Thai. *Folia Linguistica, 58*(2), 449-481. https://doi.org/10.1515/flin-2024-2006

Makayasa, B. A., Siregar, M. U., Sugiantoro, B., & Fatwanto, A. (2024). Comparison of classification algorithm and language model in accounting financial transaction record: A natural language processing approach. *International Journal on Advanced*

Science, Engineering & Information Technology, 14(3), 1044-1052. https://doi.org/10.18517/ijaseit.14.3.19179

Maqsood, S., Shahid, A., Afzal, M. T., Roman, M., Khan, Z., Nawaz, Z., & Aziz, M. H. (2022). Assessing English language sentences readability using machine learning models. *PeerJ Computer Science, 8*, Article e818. https://doi.org/10.7717/peerj-cs.818

Modhugu, V. R., & Ponnusamy, S. (2024). Comparative analysis of machine learning algorithms for liver disease prediction: SVM, logistic regression, and decision tree. *Asian Journal of Research in Computer Science, 17*(6), 188-201. https://doi.org/10.9734/ajrcos/2024/v17i6467

Mousa, A., Shahin, I., Nassif, A. B., & Elnagar, A. (2024). Detection of Arabic offensive language in social media using machine learning models. *Intelligent Systems with Applications, 22*, Article 200376. https://doi.org/10.1016/j.iswa.2024.200376

Perroni, M. G., Veiga, C. P. D., Forteski, E., Marconatto, D. A. B., da Silva, W. V., Senff, C. O., & Su, Z. (2024). Integrating relative efficiency models with machine learning algorithms for performance prediction. *SAGE Open, 14*(2), Article 21582440241257800. https://doi.org/10.1177/21582440241257800

Phann, R., Soomlek, C., & Seresangtakul, P. (2023). Multi-class text classification on Khmer news using ensemble method in machine learning algorithms. *Acta Informatica Pragensia, 12*(2), 243-259. https://doi.org/10.18267/j.aip.210

Saechan, A., & Ruangjaroon, S. (2024). *The influences of orthographic forms, stress placement and consonantal manners on syllabification and acoustic durations of intervocalic consonants with geminate graphemes by Thai L2 speakers of English* [Doctoral dissertation, Srinakharinwirot University]. Srinakharinwirot University. http://ir-ithesis.swu.ac.th/dspace/handle/123456789/2717

Saputra, N., Antoni, R., Widodo, A., Solissa, E. M., & Arief, I. (2024). Improving foreign language proficiency in society by decision tree classification. *AIP Conference Proceedings, 3001*(1). Article 070006. https://doi.org/10.1063/5.0183888

Tammanam, K., Promrit, N., & Waijanya, S. (2021). A hybrid approach to Pali Sandhi segmentation using BiLSTM and rule-based analysis. *Engineering and Applied Science Research, 48*(5), 614-626. https://doi.org/10.14456/easr.2021.63

Wang, Z., Pang, Y., & Lin, Y. (2024). *Smart expert system: Large language models as text classifiers*. arXiv. https://doi.org/10.48550/arXiv.2405.10523

Wei, Y. (2024). Chinese and English text classification techniques incorporating CHI feature selection for ELT cloud classroom. *Open Computer Science, 14*(1), Article 20240007. https://doi.org/10.1515/comp-2024-0007

Zhang, C., Hofmann, F., Plößl, L., & Gläser-Zikuda, M. (2024). Classification of reflective writing: A comparative analysis with shallow machine learning and pre-trained language models. *Education and Information Technologies, 29*, 21593-21619. https://doi.org/10.1007/s10639-024-12720-0

Zhong, X., Jin, C., An, M., & Cambria, E. (2024). XTime: A general rule-based method for time expression recognition and normalization. *Knowledge-Based Systems, 297*, Article 111921. https://doi.org/10.1016/j.knosys.2024.111921