KKU SCIENCE JOURNAL
Journal Home Page : https://ph01.tci-thaijo.org/index.php/KKUSciJ
Published by the Faculty of Science, Khon Kaen University, Thailand

การศึกษาเปรียบเทียบแบบจำลองทางสถิติสำหรับการจำแนกประเภทข่าวในสภาพข้อมูลไม่สมดุล

A Comparative Study of Statistical Models for News Classification under Imbalanced Data

กุลจิรา กิ่งไพร[1*] พิมพิกา วังแก่[1] กมล สนิทธรรม[1] และ กชกร ณ นครพนม[2]

Kunjira Kingphai[1*], Pimpika Wangkae[1], Kamol Sanittham[1] and Kodchakorn Na Nakornphanom[2]

[1]ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏเชียงใหม่ จังหวัดเชียงใหม่ 50300

[2]สาขาวิชาวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยสุโขทัยธรรมาธิราช จังหวัดนนทบุรี 11120

[1]Department of Mathematics and Statistics, Faculty of Science and Technology, Chiang Mai Rajabhat University, Chiang Mai, 50300, Thailand

[2]School of Science and Technology, Sukhothai Thammathirat Open University, Nonthaburi, 11120, Thailand

บทคัดย่อ

การจำแนกประเภทข่าวโดยอัตโนมัติเป็นงานสำคัญในด้านการประมวลผลภาษาธรรมชาติ ซึ่งช่วยอำนวยความสะดวกในการจัดหมวดหมู่และค้นหาข้อมูลจากแหล่งข่าวขนาดใหญ่ งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคการเรียนรู้ของเครื่อง โดยพิจารณาผลกระทบจากวิธีการสกัดฟีเจอร์และการจัดการข้อมูลไม่สมดุล ชุดข้อมูลที่ใช้คือ huffpost news category dataset ขั้นตอนการเตรียมข้อมูลประกอบด้วยการทำความสะอาด การลบ stopwords และการรวมพาดหัวข่าวกับคำอธิบายสั้น ฟีเจอร์ถูกสร้างด้วยเอ็นแกรมและแปลงเป็นเชิงตัวเลขด้วย bag-of-words (BoW) และ term frequency-inverse document frequency (tf-idf) จากนั้นได้ทดสอบอัลกอริทึม 4 แบบ ได้แก่ multinomial naive bayes complement naive bayes logistic regression และ linear support vector classification (linearsvc) โดยใช้ 5-fold cross-validation ผลการทดลองแสดงว่า linearsvc ร่วมกับ tf-idf ให้ประสิทธิภาพสูงสุด (accuracy 82.64% F1-score 81.87%) ขณะที่ multinomial naive bayes แสดงความเหมาะสมมากกว่ากับ BoW นอกจากนี้ การใช้ ไบแกรมช่วยลดความคลุมเครือและเพิ่มบริบทของข้อความได้ดีกว่ายูนิแกรม สำหรับการจัดการข้อมูลไม่สมดุล smote ให้ผลลัพธ์ที่เหนือกว่า adasyn และ undersampling ด้วยค่า accuracy 81.67% และ F1-score 81.68% กล่าวโดยสรุป งานวิจัยนี้นำเสนอหลักฐานเชิงประจักษ์ว่าการใช้ tf-idf ร่วมกับ linearsvc และ smote สำหรับข้อมูลที่ไม่สมดุล เป็นแนวทางที่มีประสิทธิภาพสูงในการจำแนกประเภทข่าว ข้อค้นพบดังกล่าวสามารถนำไปประยุกต์ใช้กับระบบจำแนกข้อความประเภทอื่น ๆ และยังเป็นแนวทางสำหรับการวิจัยในอนาคต

*Corresponding Author, E-mail: kunjira@g.cmru.ac.th

Received date: 2 September 2025 | Revised date: 16 November 2025 | Accepted date: 19 November 2025

doi: 10.14456/kkuscij.2026.16

ABSTRACT

Automatic news classification is an important task in natural language processing, as it facilitates the categorisation and retrieval of information from large news sources. This research aims to compare the performance of machine learning techniques by examining the effects of feature extraction methods and imbalanced data handling. The dataset used in this study is the HuffPost News Category Dataset. Data preparation includes text cleaning, stopword removal, and the combination of news headlines with short descriptions. Features are generated using n-grams and transformed into numerical representations using Bag- of- Words (BoW) and term frequency–inverse document frequency (TF- IDF). Four algorithms are evaluated, namely Multinomial Naive Bayes, Complement Naive Bayes, logistic regression, and linear support vector classification (LinearSVC), using 5-fold cross-validation. The experimental results show that linearsvc combined with tf-idf achieves the highest performance, with an accuracy of 82.64% and an F1-score of 81.87%, while multinomial naive bayes is more suitable for BoW. In addition, the use of bigrams helps reduce ambiguity and provides richer textual context than unigrams. For imbalanced data handling, SMOTE produces better results than adasyn and undersampling, achieving an accuracy of 81.67% and an F1- score of 81.68%. In conclusion, this research provides empirical evidence that using tf-idf together with linearsvc and smote for imbalanced data is a highly effective approach for news classification. These findings can be applied to other types of text classification systems and serve as guidance for future research.

คำสำคัญ: การจำแนกประเภทข้อความ การเรียนรู้ของเครื่อง การประมวลผลภาษาธรรมชาติ การจำแนกประเภทข่าว ข้อมูลไม่สมดุล

Keywords: Text Classification, Machine Learning, Natural Language Processing, News Classification, Imbalanced Data

บทนำ

ในยุคดิจิทัลปัจจุบัน ปริมาณข้อมูลข่าวสารเพิ่มขึ้นอย่างก้าวกระโดดและมีความซับซ้อนมากขึ้นอย่างต่อเนื่อง รายงานของ Marr (2018) ชี้ให้เห็นถึงอุบัติการณ์ดังกล่าว โดยระบุว่าข้อมูลปริมาณมหาศาลกว่า 2.5 ล้านล้านล้านล้านหน่วย (quintillion bytes) ถูกผลิตขึ้นใหม่ในทุก ๆ วันทั่วโลก ซึ่งส่วนสำคัญของข้อมูลเหล่านี้คือข้อมูลข่าวสารที่ต้องได้รับการจัดหมวดหมู่อย่างเป็นระบบ เพื่อสนับสนุนกลไกการสื่อสารมวลชนและระบบแนะนำข้อมูลออนไลน์ อย่างไรก็ตาม กระบวนการคัดกรองและจัดหมวดหมู่โดยอาศัยมนุษย์ (editorial curation) กลับเผชิญกับ “คอขวด” ทางด้านเวลาและทรัพยากรอย่างรุนแรง ข้อมูลจาก Graves (2018) เผยให้เห็นข้อจำกัดนี้อย่างชัดเจน โดยมนุษย์สามารถจัดการข้อมูลข่าวได้เพียง 15 - 20 รายการต่อวัน ในขณะที่ข่าวใหม่เกิดขึ้นมากกว่า 50,000 รายการต่อวัน ช่องว่างระหว่างปริมาณงานและกำลังการประมวลผลนี้สะท้อนถึงความจำเป็นเร่งด่วนของระบบอัตโนมัติที่สามารถจัดหมวดหมู่เนื้อหาอย่างมีประสิทธิภาพ ถูกต้อง และสม่ำเสมอ เพื่อรับมือกับความท้าทายดังกล่าว เทคโนโลยีการประมวลผลภาษาธรรมชาติ (natural language processing; NPL) และการเรียนรู้ของเครื่อง (machine learning; ML) ได้ถูกนำมาประยุกต์ใช้อย่างแพร่หลาย โดย Gasparetto *et al*. (2022) ยืนยันว่าระบบอัตโนมัติสามารถประมวลผลข้อมูลได้เร็วกว่ามนุษย์หลายเท่าตัว ช่วยให้องค์กรสื่อสามารถคัดกรองเนื้อหาปริมาณมากได้อย่างทันท่วงที สอดคล้องกับ McKinsey and Company (2014) ที่ระบุว่าองค์กรที่ขับเคลื่อนด้วยข้อมูลอย่างเข้มข้นสามารถเพิ่มผลกำไรได้มากขึ้นถึง 19 เท่า และขยายฐานผู้ใช้ใหม่ได้มากกว่า 23 เท่า เมื่อเทียบกับองค์กรที่ไม่ใช้เทคโนโลยีดังกล่าว

ในระยะไม่กี่ปีที่ผ่านมา ความก้าวหน้าของเทคนิคการเรียนรู้เชิงลึก (deep learning) โดยเฉพาะสถาปัตยกรรมแบบ Transformer ได้ยกระดับขีดความสามารถของงาน nlp อย่างมีนัยสำคัญ โมเดลภาษาขนาดใหญ่ อาทิ BERT (Devlin *et al*., 2019) RoBERTa (Liu *et al*., 2019) และ XLNet (Yang *et al*., 2019) แสดงให้เห็นถึงศักยภาพในการจับบริบทเชิงความหมาย (semantic context) และการเรียนรู้ความสัมพันธ์ระยะไกล (long-range dependency) ผ่านกลไก self-attention ซึ่งช่วยลดปัญหาความซ้อนทับทางความหมาย (semantic overlap) ในการจำแนกข่าวที่มีความใกล้เคียงกันได้อย่างมีประสิทธิภาพ

อย่างไรก็ดี แม้โมเดลเชิงลึกจะมีประสิทธิภาพสูง แต่ต้องแลกมาด้วยต้นทุนทรัพยากรคำนวณที่มหาศาล ความซับซ้อนในการปรับแต่งพารามิเตอร์ และความต้องการฮาร์ดแวร์ขั้นสูง (GPU/TPU) ซึ่งเป็นอุปสรรคสำคัญสำหรับการใช้งานจริงในองค์กรขนาดกลางหรือหน่วยงานข่าวที่มีทรัพยากรจำกัด ในบริบทดังกล่าว อัลกอริทึมการเรียนรู้ของเครื่องแบบดั้งเดิม (traditional machine learning) กลับเป็นทางเลือกที่น่าสนใจ เนื่องจากมีความเสถียร ตีความผลลัพธ์ได้ง่าย (interpretability) และใช้ทรัพยากรต่ำ งานวิจัยจำนวนมาก (Kim and Gil, 2019; Chowdhury and Schoen, 2020) ยืนยันว่าอัลกอริทึมอย่าง Naïve Bayes Support Vector Machines (SVM) และ Logistic Regression ยังคงให้ประสิทธิภาพสูงในงานจัดหมวดหมู่ข้อความที่มีมิติข้อมูลสูง (high-dimensional) และมีความเบาบาง (sparse representation) โดย Veziroglu *et al*. (2024) และ Alqahtani *et al*. (2022) พบว่า Multinomial Naïve Bayes และ SVM สามารถจัดการกับคำศัพท์ที่กระจายตัวและความซับซ้อนเชิงภาษาได้เป็นอย่างดี

หนึ่งในความท้าทายที่สำคัญยิ่งต่อการจำแนกข่าวคือ ปัญหาความไม่สมดุลของข้อมูล (class imbalance) ซึ่งเป็นปรากฏการณ์ที่พบได้ทั่วไปในหลายภาษาทั่วโลก (Sun *et al*., 2019; Alrashidi, 2023) ปัญหานี้ทวีความรุนแรงขึ้นเมื่อข้อมูลมีความคล้ายคลึงเชิงความหมายระหว่างหมวดหมู่ปัญหานี้ไม่ได้ส่งผลเพียงแค่ตัวเลขทางสถิติ แต่กระทบต่อการตีความเชิงภาษาของโมเดลโดยตรง โดยเฉพาะเมื่อเกิดภาวะ ความซ้อนทับเชิงความหมาย (semantic overlap) ระหว่างหมวดหมู่ เช่น คำว่า "diet" อาจปรากฏทั้งในหมวด healthy living และ wellness ดังนั้น ปัญหาสำคัญที่เทคนิคการจำแนกแบบดั้งเดิมยังแก้ไม่ได้คือ อคติเชิงความหมาย (semantic bias) ซึ่งโมเดลจะเรียนรู้ที่จะเชื่อมโยงคำกำกวมเหล่านี้เข้ากับหมวดหมู่ที่มีจำนวนข้อมูลมากกว่าเสมอ ส่งผลให้บริบททางภาษาที่ละเอียดอ่อนของหมวดหมู่ส่วนน้อยถูกละเลยและนำไปสู่การทำนายผิดพลาด อีกทั้ง งานวิจัยที่เกี่ยวข้องยังชี้ให้เห็นว่า โมเดลการเรียนรู้เชิงลึกมักแสดงความลำเอียง (bias) ไปทางกลุ่มที่มีข้อมูลมากและขาดประสิทธิภาพในการระบุคลาสส่วนน้อย ดังปรากฏในงานศึกษาของ Zhang *et al.* (2015) ซึ่งพบข้อเท็จจริงที่น่าสนใจว่า โมเดลซับซ้อนอย่าง CNN และ LSTM กลับให้ผลลัพธ์ที่ด้อยกว่า SVM เมื่อเผชิญกับชุดข้อมูลที่มีตัวอย่างจำกัดหรือมีความซ้อนทับทางความหมายสูง นอกจากนี้ เทคนิคการฝังคำ (embedding) แบบดั้งเดิม อาทิ Word2Vec และ GloVe ยังมีข้อจำกัดในการแยกแยะความหมายระดับละเอียด (fine-grained meaning) ในบริบทที่หมวดข่าวมีความกำกวม (Dang, 2020)

จากการทบทวนวรรณกรรม พบว่าแนวทางแก้ปัญหาในปัจจุบันยังมีข้อจำกัดในทางปฏิบัติ (practical limitations) กล่าวคือ กลุ่มงานวิจัยที่มุ่งเน้นโมเดลภาษาขนาดใหญ่ (llms) หรือ deep learning มักล้มเหลวในการนำไปใช้งานจริงในองค์กรที่มีทรัพยากรจำกัดเนื่องจากต้นทุนที่สูงเกินความจำเป็น ในขณะที่กลุ่มงานวิจัยที่ใช้อัลกอริทึมดั้งเดิม (traditional ml) มักละเลยการจัดการปัญหาความไม่สมดุลของข้อมูลอย่างเป็นระบบ ทำให้โมเดลขาดความสามารถในการจำแนกหมวดหมู่ที่ยากและมีความคล้ายคลึงกัน แม้จะมีพัฒนาการของเทคนิคขั้นสูงเพื่อบรรเทาปัญหาเหล่านี้ เช่น focal loss (Lin *et al*., 2019) หรือการทำ oversampling บน contextualized embeddings (Stylianou *et al*., 2023) แต่แนวทางดังกล่าวจำเป็นต้องใช้ทรัพยากรคำนวณที่สูงเกินความจำเป็น ในทางกลับกัน การประยุกต์ใช้เทคนิคระดับข้อมูลแบบดั้งเดิมอย่าง smote ร่วมกับ tf–idf กลับยังคงพิสูจน์ประสิทธิภาพได้อย่างโดดเด่น โดยงานของ Mujahidi *et al.* (2024) และ Fadli *et al*.

(2023) แสดงหลักฐานเชิงประจักษ์ว่า การผนวก SMOTE เข้ากับอัลกอริทึมพื้นฐานอย่าง logistic regression หรือ svm สามารถกู้คืนความแม่นยำในคลาสส่วนน้อยได้อย่างมีนัยสำคัญ ภายใต้ต้นทุนทรัพยากรที่ต่ำกว่าอย่างชัดเจน

เมื่อพิจารณาภาพรวม แม้งานวิจัยด้านการจำแนกข้อความจะมีความก้าวหน้าอย่างมาก แต่ยังคงมีช่องว่างงานวิจัยที่สำคัญหลายประการ ได้แก่ การขาดการเปรียบเทียบอัลกอริทึมดั้งเดิมอย่างครอบคลุม การขาดงานวิจัยที่เน้นข้อมูลข่าวโดยเฉพาะ แม้ข่าวจะมีลักษณะเฉพาะทั้งด้านภาษาและความหลากหลายของหมวดหมู่ และการมุ่งเน้นปัญหาข้อมูลไม่สมดุลไม่เพียงพอ ทั้งที่พบได้บ่อยในสถานการณ์จริง ดังนั้น งานวิจัยนี้จึงมีวัตถุประสงค์เพื่อ เปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่องในการจำแนกข่าวสารอย่างครอบคลุม เพื่อสร้างฐานข้อมูลเชิงประจักษ์ที่ช่วยให้องค์กรเลือกใช้อัลกอริทึมที่เหมาะสมต่อการประยุกต์ใช้จริง อีกทั้งยังมุ่งพัฒนาแนวทางแก้ไขปัญหาความไม่สมดุลของข้อมูลข่าวสาร และสนับสนุนการตัดสินใจเชิงนโยบายสำหรับองค์กรสื่อมวลชนในบริบทที่มีข้อจำกัดด้านทรัพยากร ผลลัพธ์ที่ได้จะเป็นแนวทางสำคัญสำหรับทั้งนักวิจัยและผู้ปฏิบัติ ในการสร้างระบบจัดหมวดหมู่ข่าวอัตโนมัติที่ทั้งแม่นยำ มีประสิทธิภาพ และสามารถนำไปใช้ได้จริงในโลกแห่งความเป็นจริง

วิธีการดำเนินการวิจัย

การวิจัยนี้ดำเนินการตามกระบวนการจำแนกประเภทข้อความ (text classification) ด้วยวิธีการเรียนรู้แบบมีผู้สอน (supervised machine learning) เพื่อพัฒนากระบวนการจำแนกข่าวโดยอัตโนมัติ เนื่องจากชุดข้อมูลที่ใช้ในการศึกษาเป็นข้อมูลที่ผ่านการกำกับฉลาก (labeled data) ซึ่งเอื้ออำนวยให้โมเดลสามารถเรียนรู้ความสัมพันธ์ระหว่างรูปแบบทางภาษาและหมวดหมู่เป้าหมายได้อย่างมีประสิทธิภาพ สอดคล้องกับวัตถุประสงค์ที่ต้องการความแม่นยำตามโครงสร้างหมวดหมู่ที่กำหนดไว้ล่วงหน้า ทั้งนี้ ผู้วิจัยพิจารณาเลือกใช้วิธีการดังกล่าวแทนการเรียนรู้แบบไม่มีผู้สอน (unsupervised learning) ซึ่งมีข้อจำกัดในการระบุหมวดหมู่ให้สอดคล้องกับบริบทการใช้งานจริง และหลีกเลี่ยงการใช้โมเดลการเรียนรู้เชิงลึก อาทิ BERT ที่มีความต้องการทรัพยากรการประมวลผลสูง รายละเอียดการดำเนินงานในแต่ละขั้นตอนได้ดังนี้

1. ชุดข้อมูลและการคัดเลือก (Data collection and selection)

ชุดข้อมูลที่ใช้ในการศึกษานี้คือ HuffPost News Category Dataset ซึ่งรวบรวมและเผยแพร่โดย Misra (2022) จากบทความข่าวที่เผยแพร่บนเว็บไซต์ Huffington Post และเผยแพร่สู่สาธารณะผ่านแพลตฟอร์ม Kaggle (เข้าถึงได้ที่ https://www.kaggle.com/datasets/rmisra/news-category-dataset) การรวบรวมข้อมูลดังกล่าวทำขึ้นเพื่อการวิจัยในด้านการจำแนกข้อความ และงานประมวลผลภาษาธรรมชาติ โดยเฉพาะ ชุดข้อมูลประกอบด้วยข่าวจำนวน 210,294 รายการที่ถูกจัดหมวดหมู่ โดยในแต่ละรายการข่าวประกอบไปด้วย หมวดหมู่ของข่าว (category) ซึ่งเป็นตัวแปรเป้าหมาย หัวข้อข่าว (headline) คำอธิบายสั้น (short description) ผู้เขียนบทความ (authors) ลิงก์ไปยังข่าวต้นฉบับ (link) และ วันที่เผยแพร่ข่าว (date)

โดยหมวดหมู่ข่าวที่มีจำนวนข่าวมากที่สุด 15 หมวดแสดงดังตารางที่ 1 สำหรับการวิจัยนี้ผู้วิจัยได้ดำเนินการคัดกรองและเลือกใช้ข้อมูลเพียง 7 หมวดหมู่ ที่ให้ผลการจำแนกประเภทที่เหมาะสมที่สุด การคัดเลือกหมวดหมู่พิจารณาจาก 3 ปัจจัยหลัก คือ ความสมดุลข้อมูลเพื่อป้องกันการลำเอียง ลักษณะเฉพาะที่แตกต่างเพื่อการจำแนกที่ชัดเจน และข้อจำกัดทรัพยากรการประมวลผล

ตารางที่ 1 แสดงจำนวนตัวอย่างในหมวดที่พบบ่อย 15 หมวด ซึ่งใช้ในการทดลองหลักของงานวิจัยนี้

หมวดหมู่ข่าว	ความถี่
politics	35,602
wellness	17,945
entertainment	17,362
travel	9,900
style & beauty	9,814
parenting	8,791
healthy living	6,694
queer voices	6,347
food & drink	6,340
business	5,992
comedy	5,400
sports	5,077
black voices	4,583
home & living	4,320
parents	3,955

จากตารางที่ 1 ข้อมูลแสดงลักษณะความไม่สมดุลอย่างเด่นชัด โดยหมวด Politics มีจำนวนสูงสุดถึง 35,602 รายการ ในขณะที่หมวดอื่น อาทิ healthy living queer voices และ food & drink มีจำนวนต่ำกว่า 7,000 รายการ รวมถึงหมวดกลุ่มน้อยอย่าง black voices และ parents ความแตกต่างของสัดส่วนตัวอย่างนี้สะท้อนถึงปัญหาความไม่สมดุลของข้อมูล ซึ่งเป็นปัจจัยวิกฤตที่ส่งผลต่อประสิทธิภาพการเรียนรู้ของโมเดล และจำเป็นต้องได้รับการจัดการอย่างเหมาะสมในขั้นตอนต่อไป

2. การเตรียมและประมวลผลข้อมูล (Data preprocessing)

เพื่อให้ชุดข้อมูลอยู่ในรูปแบบที่พร้อมสำหรับการเรียนรู้ของเครื่อง ผู้วิจัยได้กำหนดกระบวนการเตรียมข้อมูลตามลำดับ ดังนี้

2.1 การทำความสะอาดข้อมูล (data cleaning) ดำเนินการขจัดสัญญาณรบกวน (noise) โดยเริ่มจากการคัดกรองคุณลักษณะที่ไม่เกี่ยวข้องและไม่มีนัยสำคัญต่อการจำแนกความหมายออก ได้แก่ ชื่อผู้เขียน ลิงก์ และวันที่เผยแพร่ จากนั้นดำเนินการลบอักขระพิเศษ เครื่องหมายวรรคตอน และสัญลักษณ์ยูนิโค้ดที่อาจก่อให้เกิดความคลาดเคลื่อนต่อแบบจำลอง อาทิ เครื่องหมาย Em-dash (—) และ Ellipsis (...) รวมถึงเครื่องหมายอัญประกาศรูปแบบต่าง ๆ (“ ” ‘ ’) ข้อความที่เหลือจะถูกแปลงเป็นตัวพิมพ์เล็กทั้งหมด (lowercasing) เพื่อลดความซ้ำซ้อนของคำศัพท์ (canonicalization) จากนั้นเข้าสู่กระบวนการตัดคำ (tokenization) เพื่อแยกข้อความออกเป็นหน่วยคำ และประยุกต์ใช้การลดรูปคำ (stemming) เพื่อแปลงคำที่มีรากเดียวกันให้อยู่ในรูปแบบมาตรฐาน

2.2 การลบคำหยุด (stopwords removal) ดำเนินการโดยใช้โมดูล nltk.corpus จากไลบรารี nltk (natural language toolkit) สำหรับภาษา python ในการลบคำทั่วไปที่มักปรากฏในภาษาอังกฤษ เช่น “the” “and” “or” ซึ่งเป็นคำที่ไม่มีความหมายเฉพาะเจาะจงและไม่ส่งผลต่อการจำแนกหมวดหมู่ การลบคำหยุดนี้มีวัตถุประสงค์เพื่อลดจำนวนคุณลักษณะที่ไม่จำเป็นและเพิ่มคุณภาพของชุดข้อมูลในการฝึกแบบจำลอง

2.3 การรวมข้อความ (text aggregation) เพื่อเพิ่มบริบทเชิงความหมายให้แก่ข้อมูล ผู้วิจัยได้ดำเนินการรวมฟิลด์พาดหัวข่าวและคำอธิบายสั้นเข้าด้วยกันเป็นข้อความเดียว เทคนิคนี้ช่วยให้โมเดลได้รับข้อมูลที่ครบถ้วนยิ่งขึ้น ส่งผลให้ประสิทธิภาพการจำแนกมีความแม่นยำสูงกว่าการใช้เพียงส่วนใดส่วนหนึ่ง

3. การแบ่งข้อมูล (Train-Test Splitting)

เพื่อให้สอดคล้องกับลำดับการทำงานที่ถูกต้องและป้องกัน การรั่วไหลของข้อมูล ดังนั้น ข้อมูลที่ผ่านการทำความสะอาดจะถูกแบ่งออกเป็นชุดฝึกสอน (training set) และชุดทดสอบ (testing set) ในอัตราส่วน 80:20 โดยกำหนดให้ข้อมูลชุดทดสอบถูกแยกออกไปอย่างเด็ดขาดและจะไม่ถูกนำมาเกี่ยวข้องกับกระบวนการเรียนรู้หรือการสร้างพจนานุกรมคำศัพท์ใด ๆ เพื่อจำลองสถานการณ์การใช้งานจริง (unseen data)

4. การแปลงข้อมูลเป็นคุณลักษณะเชิงตัวเลข

เนื่องจากอัลกอริทึมการเรียนรู้ของเครื่องไม่สามารถประมวลผลข้อมูลในรูปแบบข้อความได้โดยตรง จึงจำเป็นต้องแปลงข้อมูลให้อยู่ในรูปเชิงตัวเลขเพื่อให้โมเดลสามารถทำความเข้าใจโครงสร้างของข้อมูลและเรียนรู้รูปแบบได้อย่างมีประสิทธิภาพ การวิจัยนี้ใช้เทคนิคการแปลงข้อความเป็นเวกเตอร์คุณลักษณะผ่านการสร้างหน่วยพื้นฐานด้วยเอ็นแกรมและวิธีการสร้างคุณลักษณะต่าง ๆ

4.1 การสร้างหน่วยพื้นฐานของข้อความด้วยเอ็นแกรม โดยเอ็นแกรมคือลำดับสัญลักษณ์จำนวน n ตัวที่ปรากฏต่อเนื่องกันในข้อความ เช่น ยูนิแกรม (n = 1) และไบแกรม (n = 2) เทคนิคนี้มีความสำคัญต่อการจับความสัมพันธ์ระหว่างคำที่อยู่ติดกัน ช่วยให้โมเดลเข้าใจบริบทของข้อความได้ลึกซึ้งขึ้น เช่น คำว่า “not good” หากแปลงเป็นยูนิแกรม จะถูกแยกเป็น “not” และ “good” ซึ่งอาจทำให้ความหมายโดยรวมสูญหายไป แต่หากใช้ไบแกรม ความหมายของ “ not good” จะยังคงถูกเก็บรักษาไว้

4.2 Bag-of-Words (BoW) เทคนิค BoW เป็นวิธีการพื้นฐานในการสร้างเวกเตอร์คุณลักษณะ โดยอาศัยการนับความถี่ของคำศัพท์ที่ปรากฏในเอกสาร โดยไม่คำนึงถึงลำดับของคำหรือโครงสร้างทางไวยากรณ์ วิธีการนี้จะสร้างตารางแจกแจงความถี่ของคำ (term frequency matrix) โดยมองว่าเอกสารแต่ละฉบับคือ “ถุงของคำ” ที่บรรจุคำต่าง ๆ อยู่โดยไม่จัดลำดับ แม้ว่ากระบวนการนี้จะทำให้ข้อมูลเชิงลำดับและความสัมพันธ์ทางไวยากรณ์สูญหายไป แต่ข้อได้เปรียบที่สำคัญคือ ความเรียบง่ายในการประมวลผล และความสามารถในการให้ผลลัพธ์ที่น่าพอใจในงานจำแนกข้อความหลายประเภท โดยเฉพาะในกรณีที่ชุดข้อมูลมีขนาดใหญ่และซับซ้อนต่ำ

4.3 Term frequency-inverse document frequency (tf-idf) tf-idf เป็นการพัฒนาต่อยอดจาก BoW โดยให้น้ำหนักกับคำที่มีความสำคัญเชิงบริบท ประกอบด้วย term frequency (ความถี่คำในเอกสาร) และ inverse document frequency (ค่าที่สะท้อนความหายากของคำในชุดข้อมูล) การผสมผสานนี้ช่วยลดความสำคัญของคำทั่วไปและเพิ่มน้ำหนักคำเฉพาะ ส่งผลให้โมเดลระบุคำที่มีบทบาทเชิงจำแนกได้อย่างมีประสิทธิภาพมากขึ้น หลังการประมวลผลด้วยเอ็นแกรมและ BoW หรือ tf-idf คุณลักษณะจะอยู่ในรูปเวกเตอร์เชิงมิติสูง โดยแต่ละมิติแทนคำหรือวลีที่สกัดออกมา งานวิจัยนี้ใช้ทั้งยูนิแกรมและไบแกรมเพื่อรักษาบริบทของข้อความ

5. การจัดการปัญหาข้อมูลไม่สมดุล (Handling class imbalance)

ปัญหาข้อมูลไม่สมดุล เป็นประเด็นที่พบได้บ่อยในชุดข้อมูลจริง โดยเฉพาะในกรณีที่จำนวนตัวอย่างระหว่างคลาสต่าง ๆ แตกต่างกันอย่างมาก ปัญหานี้ส่งผลให้แบบจำลองมีแนวโน้มลำเอียงไปยังคลาสที่มีจำนวนข้อมูลมากและทำงานได้ไม่ดีในการจำแนกคลาสส่วนน้อยเพื่อจัดการกับปัญหานี้ งานวิจัยได้ศึกษาและเปรียบเทียบเทคนิคการปรับสมดุลข้อมูล 3 วิธี ได้แก่ synthetic minority oversampling technique (smote) adaptive synthetic sampling (adasyn) และ undersampling

6. อัลกอริทึมการเรียนรู้ของเครื่อง (Machine learning algorithms)

งานวิจัยนี้ดำเนินการศึกษาและเปรียบเทียบอัลกอริทึมการเรียนรู้ของเครื่องจำนวน 4 แบบ โดยคัดเลือกอัลกอริทึมที่เหมาะสมกับลักษณะของข้อมูลข้อความซึ่งมีมิติสูงและมีความเบาบาง อัลกอริทึมกลุ่มนี้มีข้อได้เปรียบด้านต้นทุนการคำนวณต่ำและสามารถประมวลผลได้อย่างมีประสิทธิภาพ สอดคล้องกับเป้าหมายของงานวิจัย รายละเอียดของแต่ละอัลกอริทึมมีดังต่อไปนี้

6.1 การจำแนกโดยใช้ทฤษฎีเบส์อย่างง่ายแบบพหุนาม (multinomial naive bayes) การจำแนกโดยใช้ทฤษฎีเบส์อย่างง่ายแบบพหุนาม เป็นขั้นตอนวิธีภายใต้กลุ่มตัวจำแนกเบส์อย่างง่าย ซึ่งอิงตามทฤษฎีของเบส์ (bayes' theorem) ที่ระบุความสัมพันธ์ระหว่างความน่าจะเป็นแบบมีเงื่อนไข สำหรับการจำแนกข้อความ สูตรที่ใช้คือ

$$P(c|d) = \frac{P(c)x \prod_{i=1}^{n} P(t_i|c)^{tf(t_i,d)}}{P(d)} \quad (1)$$

โดยที่ $P(c|d)$ คือ ความน่าจะเป็นที่เอกสาร d จะเป็นคลาส c t_i แทนคำที่ i และ $tf(t_i, d)$ คือ จำนวนครั้งที่คำ t_i ปรากฏในเอกสาร d อัลกอริทึมนี้มี สมมติฐานของ feature independence หรือคุณลักษณะต่าง ๆ ไม่ขึ้นต่อกัน

6.2 การจำแนกโดยใช้ทฤษฎีเบส์อย่างง่ายแบบเสริม (complement naive bayes) การจำแนกโดยใช้ทฤษฎีเบส์อย่างง่ายแบบเสริม เป็นการพัฒนาต่อยอดจากทฤษฎีเบส์อย่างง่ายแบบพหุนาม โดยออกแบบมาเพื่อจัดการกับปัญหาข้อมูลไม่สมดุลซึ่งมักทำให้การจำแนกแบบพหุนามมีความลำเอียงต่อคลาสที่มีข้อมูลมาก เทคนิคนี้ใช้ความน่าจะเป็นของแต่ละคลาสจาก complement class แทนการใช้โดยตรงจากคลาสเป้าหมาย กล่าวคือ จะคำนวณจากเอกสารที่ไม่อยู่ในคลาสนั้นเพื่อประเมินโอกาสของการเป็นสมาชิกของคลาสเป้าหมาย วิธีนี้ช่วยลด bias จากคลาสที่มีข้อมูลมาก และมีแนวโน้มให้ผลลัพธ์ที่เสถียรกว่าในชุดข้อมูลที่ไม่สมดุล

6.3 การถดถอยโลจิสติก (logistic regression) การถดถอยโลจิสติก เป็นขั้นตอนวิธีทางสถิติที่ใช้สำหรับการจำแนกประเภท โดยพิจารณาความสัมพันธ์เชิงลอการิทึมระหว่างตัวแปรอิสระและตัวแปรตาม ซึ่งแตกต่างจากการถดถอยเชิงเส้นทั่วไปที่ทำนายค่าต่อเนื่อง โดยสมการพื้นฐาน คือ

$$P(Y = 1|X) = \frac{1}{1 + e^{-(w^T X + b)}} \quad (2)$$

ซึ่งเป็นฟังก์ชัน sigmoid ที่แปลงผลรวมถ่วงน้ำหนักของคุณลักษณะให้เป็นค่าความน่าจะเป็นระหว่าง 0 ถึง 1 ทั้งนี้เนื่องจากข้อมูลข่าวในงานวิจัยนี้มีหลายหมวดหมู่ จึงใช้วิธี multiclass logistic regression แบบ one-vs-rest (OvR) ซึ่งเป็นการขยายโมเดลการถดถอยโลจิสติกจากกรณีทวิภาค (binary logistic regression)ไปสู่การจำแนกหลายคลาส โดยสร้างชุดของ binary logistic regression หนึ่งโมเดลต่อหนึ่งหมวดหมู่ และเลือกค่าความน่าจะเป็นสูงสุดเป็นผลการจำแนกขั้นสุดท้าย

6.4 การจำแนกด้วยเครื่องเวกเตอร์สนับสนุนเชิงเส้น (linear support vector classification) การจำแนกด้วยเครื่องเวกเตอร์สนับสนุนเชิงเส้น เป็นขั้นตอนวิธีที่ปรับใช้จากเครื่องเวกเตอร์สนับสนุน (svm) ซึ่งเป็นขั้นตอนวิธีเชิงเรขาคณิตที่มีพื้นฐานจากแนวคิดในการหาระนาบที่ดีที่สุด (optimal hyperplane) ในการแยกข้อมูล ฟังก์ชันการตัดสินใจ (decision function) มีรูปแบบ

$$f(x) = sign(w^T x + b) \quad (3)$$

โดยที่ w คือเวกเตอร์น้ำหนัก b คือ bias และ $sign$ () เป็นค่าฉลากของข้อมูลแต่ละจุด วัตถุประสงค์คือการหาค่า w และ b ที่ทำให้ระยะห่างระหว่างระนาบแยกกับจุดข้อมูลที่ใกล้ที่สุดมีค่าสูงสุด ซึ่งสามารถแสดงเป็นปัญหาการหาค่าเหมาะที่สุด

$$\min_{w,b} \frac{1}{2}\|w\|^2 \text{ subject to } y_i(w^T x_i + b) \geq 1 \quad (4)$$

การจำแนกด้วยเครื่องเวกเตอร์สนับสนุนเชิงเส้นเหมาะสำหรับชุดข้อมูลที่สามารถแยกออกด้วยระนาบเชิงเส้นและมีประสิทธิภาพสูงเมื่อจำนวนฟีเจอร์มากกว่าจำนวนตัวอย่าง

7. การปรับแต่งพารามิเตอร์ (hyperparameter tuning)

การปรับแต่งพารามิเตอร์ดำเนินการผ่าน Grid Search ร่วมกับ 5-fold Cross-Validation (k=5) ซึ่งเป็นจุดสมดุลระหว่างความเสถียรและทรัพยากรตามแนวทางของ Kuhn and Johnson (2013) เพื่อคัดเลือกชุดพารามิเตอร์ที่ดีที่สุดจากชุดฝึกสอนไปประเมินผลบนชุดทดสอบ โดยมีรายละเอียดช่วงค่าที่ศึกษาดังตารางที่ 2

ตารางที่ 2 ค่าพารามิเตอร์ที่ใช้ในการปรับแต่งของแต่ละแบบจำลอง

แบบจำลอง	พารามิเตอร์	ช่วงค่าที่ทดสอบ
Multinomial NB	α	[0.01, 0.1, 0.5, 1, 2, 5]
Complement NB	α	[0.01, 0.1, 0.5, 1, 2, 5]
Logistic Regression	C	[0.01, 0.1, 1, 10, 100]
Linear SVM	C	[0.01, 0.1, 1, 10, 100]
	loss function	hinge และ squared hinge

8. ตัวชี้วัดการประเมินผลการดำเนินงาน

การประเมินประสิทธิภาพของโมเดลในงานวิจัยนี้ใช้ตัวชี้วัดที่หลากหลายเพื่อสะท้อนภาพรวมและวิเคราะห์จุดแข็งจุดอ่อนของแต่ละอัลกอริทึมได้อย่างรอบด้าน รายละเอียดตัวชี้วัดที่ใช้ในการประเมินคุณภาพของตัวแบบแสดงในตารางที่ 3

ตารางที่ 3 ตัวชี้วัด พร้อมสูตรคำวณและคำอธิบาย

ตัวชี้วัด	สูตรคำนวณ	คำอธิบาย
accuracy	$\frac{\sum_{i=1}^{k} tp_i}{n}$	สัดส่วนของข่าวที่โมเดลทำนายหมวดหมู่ถูกต้องทั้งหมด เมื่อเทียบกับจำนวนข่าวทั้งหมดในชุดทดสอบ
precision	$\frac{tp}{(tp + fp)}$	ค่าความแม่นยำของการทำนายในหมวดหมู่นั้น ๆ เช่น หากโมเดลทายว่าเป็น politics มันเป็น politics จริงหรือไม่
recall	$\frac{tp}{(tp + fn)}$	ความสามารถของโมเดลในการดึงข่าวที่เป็นหมวดนี้จริง ๆ ออกมาได้ครบแค่ไหน เช่น ข่าว healthy living จริง โมเดลหาเจอมากน้อยเพียงใด
F1-score	$2 \times \frac{(precision \times recall)}{(precision + recall)}$	ค่าที่รวม precision และ recall ใช้วัดความสมดุลระหว่างการทำนายถูกและการหาครบในแต่ละหมวดข่าว
macro precision	$\frac{\sum_{i=1}^{k} precision_i}{k}$	ค่า precision เฉลี่ยของทุกหมวดข่าว โดยให้ทุกหมวดมีน้ำหนักเท่ากัน ช่วยบอกว่าโมเดลทำงานสมดุลระหว่างหมวดใหญ่หมวดเล็กเพียงใด

ตารางที่ 3 ตัวชี้วัด พร้อมสูตรคำนวณและคำอธิบาย (ต่อ)

ตัวชี้วัด	สูตรคำนวณ	คำอธิบาย
macro recall	$\frac{\sum_{i=1}^{k} \text{recall}_i}{k}$	ค่าเฉลี่ย recall ของทุกหมวดข่าว ใช้ตรวจสอบว่าโมเดลสามารถจดจำลักษณะของหมวดหมู่ที่มีข้อมูลน้อยได้ดีเพียงใด
macro F1-score	$\frac{\sum_{i=1}^{k} \text{F1}_i}{k}$	ค่าเฉลี่ย F1-score แบบไม่ถ่วงน้ำหนัก เหมาะสำหรับประเมินประสิทธิภาพในภาพรวมโดยไม่ให้หมวดที่มีข้อมูลมากมากลบเกลื่อนปัญหาของหมวดเล็ก
micro precision	$\frac{\sum_{i=1}^{k} \text{tp}}{\sum_{i=1}^{k} \text{tp} + \sum_{i=1}^{k} \text{fp}}$	precision รวมของทุกหมวดข่าว โดยมองจำนวนผลทำนายถูกและผิดจากทั้งระบบ เหมาะสำหรับประเมินภาพรวมของโมเดล
micro recall	$\frac{\sum_{i=1}^{k} \text{tp}}{\sum_{i=1}^{k} \text{tp} + \sum_{i=1}^{k} \text{fn}}$	recall รวมของข่าวทุกหมวด ใช้ดูประสิทธิภาพรวมว่าระบบสามารถดึงข่าวที่ถูกต้องออกมาได้ดีแค่ไหน
micro F1-score	$2 \times \frac{(\text{micro precision} \times \text{micro recall})}{(\text{micro precision} + \text{micro recall})}$	ค่าประสิทธิภาพรวมระดับตัวอย่างข่าว ซึ่งในงานจำแนกประเภทแบบ single-label ค่านี้จะ เท่ากับ accuracy เสมอ

หมายเหตุ: TP (True Positive) คือ จำนวนที่ทำนายถูกว่าเป็นหมวดนั้น

FP (False Positive) คือจำนวนที่ทำนายผิดว่าเป็นหมวดนั้น (จริง ๆ เป็นหมวดอื่น)

FN (False Negative) คือ จำนวนที่เป็นหมวดนั้นจริง ๆ แต่ทำนายผิดไปเป็นหมวดอื่น

K คือ จำนวนหมวดหมู่ทั้งหมด และ n คือจำนวนตัวอย่างข้อมูลทั้งหมด

ผลการวิจัยและวิจารณ์ผล

การดำเนินการทดลองในงานวิจัยนี้แบ่งออกเป็นสามส่วนหลัก ได้แก่ (1) การสำรวจข้อมูลเบื้องต้น (2) การเปรียบเทียบเทคนิคการแปลงข้อมูลเป็นคุณลักษณะ และ (3) การประเมินเทคนิคการจัดการข้อมูลไม่สมดุล ผลการทดลองในแต่ละส่วนมีดังนี้

1. การสำรวจข้อมูลเบื้องต้น

1.1 ความถี่ของคำศัพท์ การวิเคราะห์ความถี่ของคำศัพท์ก่อนการลบคำหยุดที่ไม่สำคัญต่อเนื้อหาของข่าว พบว่าคำที่มีความถี่สูงสุด ได้แก่ "the" "to" "a" และ "of" ซึ่งเป็นคำที่ไม่มีความหมายเฉพาะเจาะจงต่อการจำแนกประเภทข่าว หลังจากดำเนินการลบคำที่ไม่สำคัญแล้ว พบว่าคำที่มีความถี่สูงสุดเปลี่ยนเป็น "new" จำนวน 13,555 ครั้งจากบทความข่าวทั้งหมด 148,122 บทความ รองลงมาได้แก่ "trump" "one" และ "like" ตามลำดับ

1.2 การวิเคราะห์เอ็นแกรม การวิเคราะห์เปรียบเทียบยูนิแกรมและไบแกรมในตารางที่ 4 เผยให้เห็นข้อจำกัดสำคัญของยูนิแกรม คือปัญหาความคลุมเครือของคำที่มีความถี่สูง เช่น คำว่า "new" ที่ปรากฏในหลายหมวดหมู่ wellness (1,717 ครั้ง) และ entertainment (1,976 ครั้ง) ทำให้ไม่สามารถใช้เป็นตัวบ่งชี้ลักษณะเฉพาะของแต่ละหมวดข่าวได้ ในขณะที่ไบแกรมแสดงประสิทธิภาพเหนือกว่าโดยการรวมคำสองคำที่อยู่ติดกัน ช่วยสร้างบริบทที่ชัดเจนและลดความคลุมเครือ ตัวอย่างที่เด่นชัด ได้แก่ "health care" (216 ครั้ง) ในหมวด wellness ที่ให้ความหมายเฉพาะเจาะจง "new york" (189 ครั้ง) ในหมวด travel ที่ระบุจุดหมายปลายทาง และการเปลี่ยนจาก "gay" เป็น "gay men" ในหมวด queer voices ที่ระบุกลุ่มเป้าหมายชัดเจนขึ้น อย่างไรก็ตาม ไบแกรมยังมีข้อจำกัดบางประการ เช่น "donald trump" ที่ปรากฏทั้งใน politics (2,568 ครั้ง) และ entertainment (304 ครั้ง) แสดงว่าบุคคลสาธารณะบางคนครอบคลุมหลายหมวดหมู่ นอกจากนี้ "sure check" ในหมวด style & beauty (770 ครั้ง) อาจเป็นเพียงรูปแบบการเขียนมากกว่าเนื้อหาหลัก โดยสรุป ไบแกรมแสดงประสิทธิภาพเหนือกว่ายูนิแกรมอย่างชัดเจน โดยให้ข้อมูลที่จำเพาะและบริบทที่ชัดเจนกว่า ทำให้เป็นเครื่องมือที่มีประสิทธิภาพสำหรับการจำแนกประเภทข่าวและสื่อดิจิทัล

ตารางที่ 4 คำในรูปแบบยูนิแกรมและไบแกรม ที่พบบ่อยที่สุดในแต่ละหมวดหมู่ข่าว

หมวดหมู่ข่าว	ยูนิแกรม (ความถี่)	ไบแกรม (ความถี่)
politics	trump (8,610)	donald trump (2,568)
wellness	new (1,717)	health care (216)
entertainment	new (1,976)	donald trump (304)
travel	travel (1,414)	new york (189)
style & beauty	fashion (1,791)	sure check (770)
parenting	kids (1,420)	cute kids (99)
healthy living	health (579)	gps guide (107)
queer voices	gay (1,705)	gay men (98)
food & drink	food (788)	ice cream (83)
business	new (627)	women business (120)
comedy	trump (702)	donald trump (322)
sports	nfl (436)	super bowl (127)
black voices	black (1,390)	black lives (87)
home & living	home (1,353)	craft day (142)
parents	kids (537)	say darndest (44)

2. การเปรียบเทียบเทคนิคการแปลงข้อมูลเป็นคุณลักษณะ

การแปลงข้อความให้เป็นการแทนค่าเชิงตัวเลขเป็นขั้นตอนสำคัญในการประมวลผลภาษาธรรมชาติ เทคนิคการสกัดคุณลักษณะที่เลือกใช้ส่งผลโดยตรงต่อประสิทธิภาพของอัลกอริทึมการเรียนรู้ของเครื่อง ผลการทดลองเปรียบเทียบระหว่าง BoW และ tf-idf ในตารางที่ 5 เผยให้เห็นรูปแบบการตอบสนองที่หลากหลายของแต่ละอัลกอริทึม

ตารางที่ 5 การเปรียบเทียบประสิทธิภาพ BoW และ tfidf ในโมเดลต่างๆ

Model	Accuracy			F1-score		
	BoW	tf-idf	(tf-idf vs BoW)	BoW	tf-idf	(tf-idf vs BoW)
Multinomial naive bayes	80.84	69.90	-10.94%	78.75	65.24	-13.51%
Complement naive bayes	80.70	81.49	+0.79%	78.46	79.35	+0.89%
Logistic regression	81.70	82.26	+0.56%	81.55	81.19	-0.36%
Linear support vector classification	80.88	82.64	+1.76%	80.87	81.87	+1.00%

จากผลการวิเคราะห์ที่แสดงในตารางที่ 5 พบว่า โมเดล multinomial naive bayes แสดงความเหมาะสมกับ BoW อย่างชัดเจน ด้วยค่า accuracy ที่ 80.84% และ F1-score ที่ 78.75% ซึ่งสูงกว่าการใช้ tf-idf ถึง 10.94% และ 13.51% ตามลำดับ ผลลัพธ์เชิงประจักษ์นี้สอดคล้องกับรากฐานทางทฤษฎีที่ระบุไว้ในสมการที่ (1) ซึ่งอัลกอริทึมคำนวณความน่าจะเป็นโดยอาศัยตัวแปร ที่ถูกนิยามเป็นจำนวนครั้งของคำที่ปรากฏ รูปแบบข้อมูลจาก BoW จึงมีคุณลักษณะเป็นจำนวนนับที่ตรงตามข้อสมมติฐานของสมการอย่างสมบูรณ์ ในขณะที่ tf-idf แปลงค่าดังกล่าวเป็นทศนิยมถ่วงน้ำหนัก ทำให้ข้อมูลที่ป้อนเข้าสู่โมเดลเบี่ยงเบนไปจากนิยามทางคณิตศาสตร์ดั้งเดิม นอกจากนี้ การลดน้ำหนักคำที่พบบ่อยของ tf-idf ยังส่งผลให้คำสำคัญที่บ่งบอก

เอกลักษณ์ของหมวดหมู่ข่าวถูกลดทอนความสำคัญลง ส่งผลให้ประสิทธิภาพของโมเดลลดต่ำลงเมื่อเทียบกับการใช้ความถี่ดิบจาก BoW

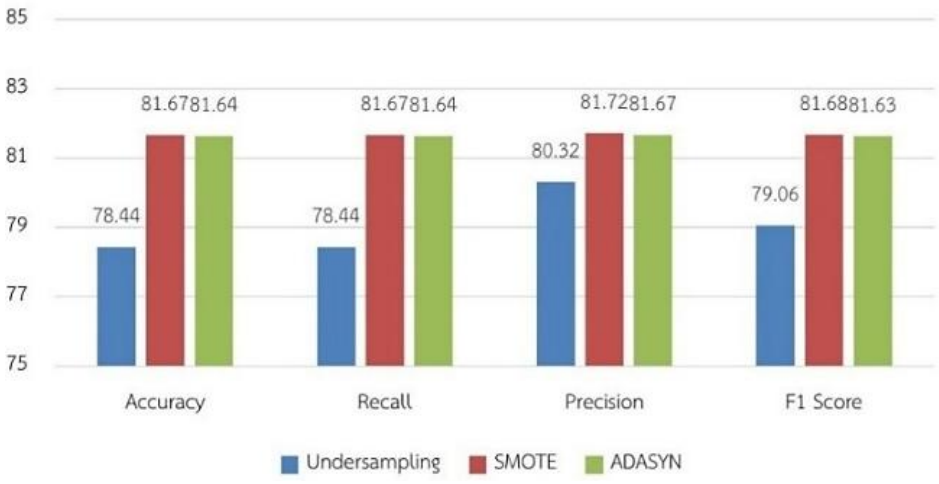
ในทางตรงกันข้าม complement naive bayes แสดงผลที่แตกต่าง โดย tf-idf ให้ประสิทธิภาพที่ดีกว่า BoW เล็กน้อย ด้วยค่า accuracy ที่ 81.49% และ F1-score ที่ 79.35% การปรับปรุงนี้เกิดขึ้น 0.79% และ 0.89% ตามลำดับ ความแตกต่างนี้อาจเกิดจากการที่ complement naive bayes ถูกออกแบบมาเพื่อแก้ไขปัญหาของข้อมูลที่ไม่สมดุล ทำให้สามารถใช้ประโยชน์จากการถ่วงน้ำหนักของ tf-idf ได้ดีกว่า

โมเดล linear support vector classification บรรลุประสิทธิภาพสูงสุดเมื่อใช้กับ tf-idf ด้วยค่า accuracy ที่ 82.64% และ F1-score ที่ 81.87% ซึ่งดีกว่า BoW ถึง 1.76% และ 1.00% ตามลำดับ ความสำเร็จนี้สอดคล้องกับลักษณะของ support vector machine ที่มีความสามารถในการจัดการกับข้อมูลที่มีมิติสูงและความเบาบาง logistic regression นำเสนอรูปแบบที่น่าสนใจ tf-idf ช่วยปรับปรุง accuracy เพิ่มขึ้น 0.56% แต่กลับทำให้ F1-score ลดลง 0.36% ปรากฏการณ์นี้อาจสะท้อนถึงผลกระทบของการถ่วงน้ำหนักต่อการจำแนกประเภทที่มีข้อมูลไม่สมดุล การปรับปรุง accuracy โดยรวมอาจมาพร้อมกับการลดลงของประสิทธิภาพในการจำแนกกลุ่มที่มีขนาดเล็ก

การศึกษานี้เสนอหลักฐานเชิงประจักษ์ที่ยืนยันว่าการเลือกใช้เทคนิคการสกัดคุณลักษณะจำเป็นต้องพิจารณาร่วมกับคุณลักษณะเชิงทฤษฎีของอัลกอริทึมการเรียนรู้ ไม่มีเทคนิคใดที่แสดงความเหนือกว่าในทุกบริบทการประยุกต์ใช้ ดังนั้น ความเข้าใจในหลักการทำงานเชิงอัลกอริทึมจึงมีความสำคัญในการเลือกใช้เทคนิคที่ให้ประสิทธิภาพสูงสุด

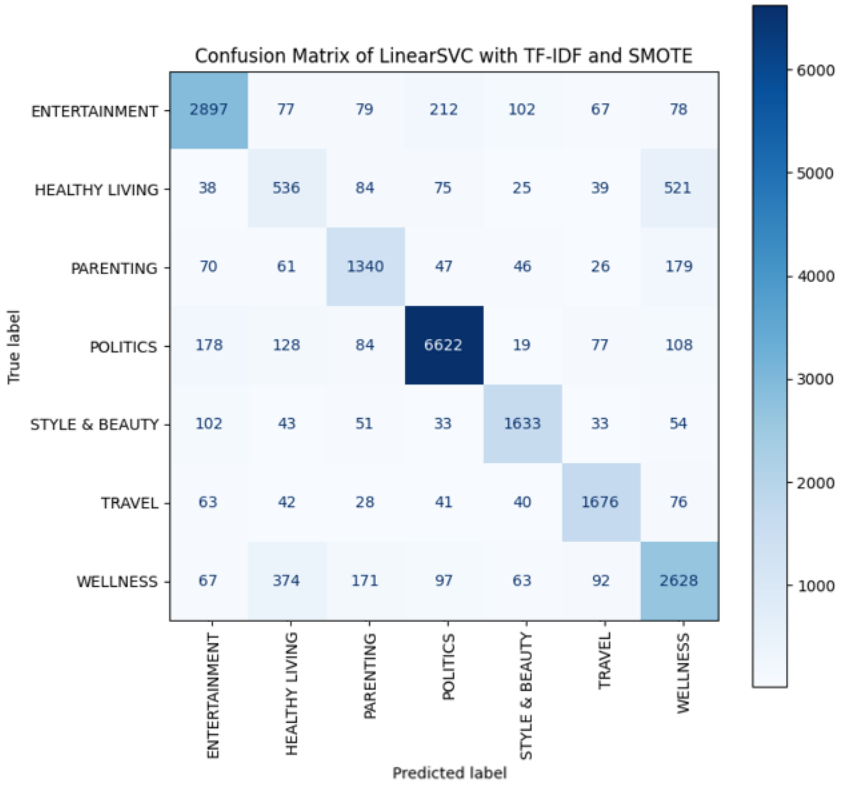
3. การประเมินเทคนิคการจัดการข้อมูลไม่สมดุล

ปัญหาการกระจายข้อมูลไม่สมดุลเป็นความท้าทายที่พบได้บ่อยในชุดข้อมูลจริง โดยส่งผลให้อัลกอริทึมการเรียนรู้มีแนวโน้มลำเอียงไปยังประเภทที่มีข้อมูลมากกว่า ในงานจำแนกประเภทข่าวนี้ ประเภทข่าวบางประเภท เช่น parents และ black voices มีจำนวนตัวอย่างน้อยกว่าประเภทอื่นอย่างมาก จากผลการทดลองในขั้นตอนก่อนหน้าที่แสดงให้เห็นว่า linear support vector classification ร่วมกับ tf-idf ให้ประสิทธิภาพสูงสุด ด้วย accuracy เท่ากับ 82.64% การศึกษาจึงเลือกใช้การตั้งค่านี้เป็นพื้นฐานในการประเมินเทคนิคการจัดการข้อมูลไม่สมดุลสามแบบ ได้แก่ smote adasyn และ undersampling



รูปที่ 1 การเปรียบเทียบประสิทธิภาพของเทคนิคการจัดการข้อมูลไม่สมดุล

จากรูปที่ 1 แสดงให้เห็นว่า smote ให้ประสิทธิภาพสูงสุดด้วยค่า accuracy และ F1-score ที่ 81.67% และ 81.68% ตามลำดับ ความสำเร็จของ smote เกิดจากการสร้างข้อมูลสังเคราะห์ที่มีคุณภาพสำหรับประเภทข่าวที่มีข้อมูลน้อย โดยใช้หลักการ k-nearest neighbors ในการสร้างตัวอย่างใหม่ที่อยู่ในบริเวณใกล้เคียงกับข้อมูลดั้งเดิม วิธีการนี้ช่วยให้โมเดลสามารถเรียนรู้ลักษณะเฉพาะของประเภทข่าวที่มีข้อมูลน้อยได้อย่างมีประสิทธิภาพ จากผลการวิเคราะห์ พบว่า adasyn แสดงประสิทธิภาพที่ต่ำกว่า smote แม้ว่าจะใช้หลักการคล้ายคลึงกัน ความแตกต่างหลักคือ adasyn มุ่งเน้นการสร้างข้อมูลใหม่ในบริเวณที่ยากต่อการเรียนรู้มากกว่า การปรับแต่งที่ซับซ้อนนี้อาจทำให้เกิดการสร้างข้อมูลที่ไม่เหมาะสมกับลักษณะของข้อมูลข้อความ ซึ่งมีความเบาบางและมิติสูง ในขณะที่ undersampling ให้ประสิทธิภาพต่ำสุดในการทดลองนี้ แม้ว่าเทคนิคนี้จะช่วยแก้ปัญหาความไม่สมดุลได้ แต่การลดจำนวนข้อมูลในประเภทที่มีข้อมูลมากอาจทำให้สูญเสียข้อมูลที่มีประโยชน์ โดยเฉพาะในบริบทของข้อมูลข้อความที่ต้องการตัวอย่างจำนวนมากเพื่อเรียนรู้รูปแบบภาษาที่หลากหลาย



รูปที่ 2 confusion matrix ของการจำแนกหมวดหมู่ข่าว 7 หมวดหมู่

จากรูปที่ 2 ซึ่งแสดง confusion matrix การวิเคราะห์ผลการจำแนกพบว่า โมเดลมีประสิทธิภาพที่แตกต่างกันอย่างชัดเจนในแต่ละหมวดหมู่ข่าว เมื่อเปรียบเทียบกับประสิทธิภาพโดยรวม 81.67% พบว่า หมวดหมู่ politics แสดงประสิทธิภาพสูงสุดด้วยการจำแนกที่ถูกต้อง 6,622 รายการจากทั้งหมด 7,216 รายการ คิดเป็นอัตราความแม่นยำ 91.8% ซึ่งสูงกว่าค่าเฉลี่ยโดยรวมถึง 10.1% บ่งชี้ว่าข่าวการเมืองมีลักษณะเฉพาะทางภาษาที่ชัดเจนและแยกแยะได้ง่าย เช่น คำศัพท์เฉพาะด้านการเมือง ชื่อนักการเมืองและประเด็นทางการเมือง หมวดหมู่ travel และ wellness แสดงประสิทธิภาพที่ดีในระดับรองและใกล้เคียงกับค่าเฉลี่ยโดยรวม โดย travel มีการจำแนกที่ถูกต้อง 1,676 รายการจาก 1,966 รายการ (85.2%) และ wellness มีการจำแนกที่ถูกต้อง 2,628 รายการจาก 3,492 รายการ (75.3%) ความสำเร็จของหมวดหมู่เหล่านี้อาจมาจากการมีคำศัพท์เฉพาะโดเมนที่ชัดเจน เช่น สถานที่ท่องเที่ยว กิจกรรมการเดินทาง สำหรับ travel และคำศัพท์เกี่ยวกับสุขภาพ การออกกำลังกาย การดูแลตัวเอง สำหรับ wellness ในขณะที่หมวดหมู่ healthy living แสดงประสิทธิภาพต่ำสุด โดยมี

ความแม่นยำเพียง 40.7% (536 จาก 1,318 รายการ) ซึ่งต่ำกว่าค่าเฉลี่ยโดยรวมถึง 40.9% บ่งชี้ถึงปัญหาสำคัญในการแยกแยะหมวดหมู่นี้จากหมวดหมู่อื่น รูปแบบการจำแนกผิดที่สำคัญ การวิเคราะห์รูปแบบการจำแนกผิดเผยให้เห็นประเด็นสำคัญหลายประการ การจำแนกผิดที่เห็นได้ชัดที่สุดคือการที่ข่าว healthy living ถูกจำแนกผิดเป็น wellness ถึง 521 รายการ ซึ่งเป็นจำนวนสูงสุดในเมทริกซ์ทั้งหมด ปรากฏการณ์นี้สามารถอธิบายได้จากความคาบเกี่ยวทางเนื้อหาระหว่างสองหมวดหมู่ โดยทั้งสองหมวดหมู่มีการใช้คำศัพท์ที่เกี่ยวข้องกับสุขภาพ เช่น "health" "nutrition" "fitness" "diet" "exercise" "wellbeing" ทำให้โมเดลไม่สามารถแยกแยะได้อย่างชัดเจน

การจำแนกผิดรูปแบบอื่นที่น่าสนใจ ได้แก่ parenting ที่ถูกจำแนกผิดเป็น healthy living จำนวน 84 รายการ ซึ่งสะท้อนให้เห็นว่าเนื้อหาเกี่ยวกับการเลี้ยงดูเด็กมักเกี่ยวข้องกับสุขภาพของเด็ก การดูแลสุขภาพครอบครัว และคำแนะนำด้านสุขภาพสำหรับผู้ปกครอง นอกจากนี้ หมวดหมู่ entertainment แสดงรูปแบบการกระจายตัวของการจำแนกผิดไปยังหลายหมวดหมู่ ซึ่งอาจบ่งชี้ว่าเนื้อหาบันเทิงมีความหลากหลายและอาจครอบคลุมหัวข้อต่างๆ ที่ทับซ้อนกับหมวดหมู่อื่น ปัจจัยที่ส่งผลต่อประสิทธิภาพการจำแนก ความแตกต่างในประสิทธิภาพการจำแนกระหว่างหมวดหมู่สามารถอธิบายได้จากหลายปัจจัย ปัจจัยแรกคือ ความเฉพาะเจาะจงของคำศัพท์ หมวดหมู่ที่มีประสิทธิภาพสูง เช่น politics มักมีคำศัพท์เฉพาะโดเมนที่ไม่ปรากฏในหมวดหมู่อื่น ในขณะที่หมวดหมู่ที่มีการจำแนกผิดสูง เช่น healthy living และ wellness มีการใช้คำศัพท์ร่วมกันมาก ปัจจัยที่สองคือ ความเป็นนามธรรมของหัวข้อ หมวดหมู่ที่เป็นรูปธรรม เช่น travel ที่มีการกล่าวถึงสถานที่เฉพาะ กิจกรรมเฉพาะ มีแนวโน้มที่จะถูกจำแนกได้แม่นยำกว่าหมวดหมู่ที่เป็นนามธรรมหรือครอบคลุมหัวข้อกว้าง ปัจจัยที่สามคือ ความสมดุลของข้อมูล แม้ว่าจะมีการใช้เทคนิค smote แล้ว แต่ลักษณะเฉพาะของเนื้อหาในแต่ละหมวดหมู่ยังคงส่งผลต่อประสิทธิภาพการเรียนรู้ของโมเดล

เพื่อให้การประเมินประสิทธิภาพของโมเดลครอบคลุมมากกว่าการรายงานค่า accuracy เพียงตัวเดียว งานวิจัยนี้จึงคำนวณ precision recall และ F1-score แยกตามหมวดหมู่ข่าว รวมถึงรายงานค่าแบบ macro-average และ micro-average ซึ่งสะท้อนผลลัพธ์ทั้งในระดับหมวดหมู่และภาพรวมของชุดข้อมูล

ตารางที่ 6 ค่าประสิทธิภาพรายหมวดหมู่

หมวดหมู่	Precision	Recall	F1-score
entertainment	0.8483	0.8249	0.8364
healthy living	0.4251	0.4067	0.4157
parenting	0.7295	0.7575	0.7432
politics	0.9291	0.9177	0.9234
style & beauty	0.8470	0.8379	0.8424
travel	0.8338	0.8525	0.8431
wellness	0.7212	0.7526	0.7365

ตารางที่ 6 แสดงผลการจำแนกในแต่ละหมวดหมู่ โดยหมวด politics travel style & beauty และ entertainment ให้ค่า F1-score สูงกว่า 0.80 สะท้อนถึงลักษณะคำศัพท์เฉพาะที่ช่วยให้โมเดลจำแนกได้อย่างแม่นยำ ส่วน parenting และ wellness ให้ค่า F1-score อยู่ที่ประมาณ 0.73 - 0.74 ขณะที่ healthy living มีค่า F1-score ต่ำที่สุด (0.416) และมีความคลาดเคลื่อนสูง โดยข้อมูลจำนวนมากถูกจำแนกผิดไปเป็น wellness อ้างอิงจากผลที่แสดงใน รูปที่ 2 confusion matrix

ตารางที่ 7 ผลการประเมินประสิทธิภาพภาพรวมของแบบจำลองด้วยตัวชี้วัดแบบ Macro และ Micro

ประเภทตัวชี้วัด	Precision	Recall	F1-score
Macro (เฉลี่ยเท่ากันทุกหมวด)	0.7620	0.7642	0.7630
Micro (เฉลี่ยตามจำนวนข้อมูล)	0.8167	0.8167	0.8167

ผลการประเมินภาพรวมด้วย Macro และ Micro สรุปในตารางที่ 7 พบว่า ค่า Macro F1-score มีค่า 0.7630 ซึ่งเป็นการเฉลี่ยน้ำหนักเท่ากันทุกหมวด สะท้อนความสามารถของโมเดลเมื่อให้ความสำคัญกับทุกหมวดหมู่ข่าวเท่าเทียมกันในขณะที่ค่า Micro F1-score เท่ากับ 0.8167 ซึ่งสัมพันธ์กับ accuracy โดยตรง เนื่องจากงานนี้เป็น single-label classification ทำให้ micro-precision micro-recall และ micro-F1 มีค่าเท่ากันโดยนิยาม ส่วนต่างระหว่างค่า Macro-F1 (0.763) และ Micro-F1 (0.817) สะท้อนให้เห็นว่าโมเดลมีแนวโน้มลำเอียง ไปยังหมวดหมู่ที่มีจำนวนข้อมูลมาก โดยยังคงประสบปัญหาในการจำแนกหมวดหมู่ที่มีความกำกวมทางความหมาย โดยเฉพาะอย่างยิ่งในคู่ของ healthy living และ wellness ซึ่งชุดคำศัพท์มีความคล้ายคลึงกันสูง จนเทคนิคพื้นฐานอย่าง tf-idf ไม่สามารถแยกแยะบริบทได้ละเอียดเพียงพอ

โดยภาพรวม โมเดลที่ใช้สามารถจำแนกหมวดหมู่ข่าวได้อย่างมีประสิทธิภาพ โดยเฉพาะหมวดที่มีคำศัพท์เฉพาะหรือรูปแบบภาษาชัดเจน ซึ่งให้ค่า F1-score อยู่ในระดับสูงอย่างต่อเนื่อง ข้อจำกัดที่พบในบางหมวดหมู่ เช่น healthy living ซึ่งมีความหมายใกล้เคียงกับ wellness สะท้อนลักษณะของข้อมูลที่มีความทับซ้อนเชิงความหมายมากกว่าข้อจำกัดของโมเดลเอง ทั้งนี้ผลลัพธ์ดังกล่าวชี้ให้เห็นถึงโอกาสในการพัฒนาเพิ่มเติม โดยการประยุกต์คุณลักษณะเชิงบริบทที่ลึกกว่าเพื่อเพิ่มความสามารถของโมเดลในการแยกแยะหมวดหมู่ที่มีความซับซ้อนสูงในงานวิจัยต่อไป

สรุปผลการวิจัย

ผลการวิจัยครั้งนี้ยืนยันถึงบทบาทสำคัญของกระบวนการเตรียมข้อมูลและการเลือกใช้เทคนิคการสกัดคุณลักษณะที่มีผลโดยตรงต่อประสิทธิภาพของการจำแนกประเภทข่าวสารในบริบทของข้อมูลขนาดใหญ่ การศึกษาพบว่าการลบคำหยุดที่ไม่ช่วยในการจำแนกร่วมกับการใช้เทคนิคเอ็นแกรม โดยเฉพาะไบแกรม แสดงให้เห็นถึงศักยภาพในการลดความกำกวมและเพิ่มความจำเพาะของคำ ซึ่งสอดคล้องกับแนวคิดด้านการเรียนรู้ความสัมพันธ์ของคำตามการปรากฏร่วมกัน (co-occurrence statistics) ในงานสมัยใหม่ เช่น Word2Vec (Mikolov *et al*., 2013) และ GloVe (Pennington *et al*., 2014) ที่ชี้ว่าการใช้ข้อมูลลำดับคำสามารถช่วยเพิ่มความแม่นยำในการจำแนกข้อความได้อย่างมีนัยสำคัญ

เมื่อเปรียบเทียบเทคนิคการสร้างตัวแทนข้อมูล พบว่า tf-idf ให้ผลลัพธ์ที่ดีกว่าในโมเดล complement naive bayes, logistic regression และ linear svc ในขณะที่ multinomial naive bayes มีประสิทธิภาพสูงกว่าเมื่อใช้ร่วมกับ BoW ซึ่งสอดคล้องกับข้อสังเกตของ Rennie *et al.* (2003) ที่ระบุว่า multinomial naive bayes ตั้งอยู่บนสมมติฐานการกระจายความถี่ของคำแบบพหุนาม จึงทำงานได้ดีเมื่อข้อมูลไม่ได้ถูกปรับน้ำหนักคำเพิ่มเติม อย่างไรก็ตาม ในภาพรวมของการทดลองพื้นฐาน linear support vector classification (linear svc) ร่วมกับ tf-idf ให้ความถูกต้องสูงสุดที่ 82.64% สะท้อนถึงความเหมาะสมของ svm ในการจัดการกับข้อมูลที่มีมิติสูงและเบาบาง สำหรับการจัดการปัญหาข้อมูลไม่สมดุล ผลการทดลองชี้ว่าเทคนิค smote ให้ผลลัพธ์เหนือกว่า adasyn และ undersampling โดยให้ค่าความถูกต้องที่ 81.67% และค่า F1-score เฉลี่ยระดับ Micro ที่ 0.8167 แม้ค่าความถูกต้องจะลดลงเล็กน้อยเมื่อเทียบกับกรณีไม่ปรับสมดุล แต่ช่วยเพิ่มความครอบคลุมของขอบเขตการตัดสินใจและลดอคติในหมวดหมู่ที่มีข้อมูลน้อยได้อย่างมีประสิทธิภาพ

เมื่อวิเคราะห์ประสิทธิภาพเชิงลึกผ่าน confusion matrix พบว่าโมเดลมีความสามารถในการจำแนกที่แตกต่างกันอย่างมีนัยสำคัญในแต่ละหมวดหมู่ โดยหมวดหมู่ที่มีคำศัพท์เฉพาะเจาะจงชัดเจนอย่างการเมือง (politics) มีความแม่นยำสูงสุดถึง 91.8% ในขณะที่หมวดหมู่การใช้ชีวิตเพื่อสุขภาพ (healthy living) กลับมีประสิทธิภาพต่ำที่สุด (40.7%) และพบความผิดพลาดสูงในการจำแนกสลับกับหมวดสุขภาวะ (wellness) ข้อค้นพบนี้ชี้ให้เห็นถึงปัญหา การซ้อนทับเชิง-

ความหมาย (semantic overlap) เนื่องจากทั้งสองหมวดมีการใช้คำศัพท์ร่วมกันสูง เช่น diet fitness และ health ทำให้เทคนิคพื้นฐานอย่าง tf-idf ไม่สามารถแยกแยะบริบทที่ละเอียดอ่อนนี้ได้เพียงพอ ซึ่งสอดคล้องกับข้อสังเกตในงานของ Mihalcea *et al.* (2006) เกี่ยวกับความท้าทายในการจำแนกข้อความที่มีความใกล้เคียงกันเชิงความหมาย

อย่างไรก็ตาม จากผลการศึกษาพบข้อควรระวังสำคัญสำหรับการนำโมเดลไปใช้งานในสถานการณ์ กล่าวคือ แม้โมเดลจะมีค่าความแม่นยำโดยรวมสูง แต่มีความเสี่ยงที่จะเกิดความผิดพลาดในหมวดหมู่ที่มีความกำกวมทางความหมายสูง เช่น ข่าวสุขภาพและไลฟ์สไตล์ ดังนั้น ในทางปฏิบัติองค์กรสามารถออกแบบโครงสร้างหมวดหมู่ข่าวในองค์กร โดยควรหลีกเลี่ยงการกำหนดหมวดหมู่ที่มีขอบเขตทับซ้อนกันเพื่อลดความสับสนของทั้งโมเดลและผู้ใช้งาน

จากข้อจำกัดดังกล่าว งานวิจัยนี้เสนอแนะแนวทางสำหรับการศึกษาในอนาคต 3 ประเด็นหลัก ได้แก่ (1) การประยุกต์ใช้เทคนิคการฝังคำ (word embeddings) อาทิ word2vec เพื่อจับความสัมพันธ์ของคำที่มีความหมายใกล้เคียงกันได้ลึกซึ้งยิ่งขึ้น (2) การทดลองใช้สถาปัตยกรรมการเรียนรู้เชิงลึกที่มีความซับซ้อนและสามารถจับบริบทได้ดีกว่าโมเดลเชิงเส้น เช่น BERT หรือ Transformer-based models และ (3) การขยายผลการทดสอบไปยัง ชุดข้อมูลภาษาอื่น เช่น ภาษาไทย หรือข้ามโดเมน (cross-lingual or cross-domain) เพื่อประเมินความสามารถในการทั่วไป (generalization) ซึ่งจะนำไปสู่การพัฒนาตัวจำแนกข่าวสารอัตโนมัติที่มีความเสถียรและแม่นยำสำหรับการใช้งานจริงต่อไป

เอกสารอ้างอิง

Alqahtani, A., Ullah Khan, H., Alsubai, S., Sha, M., Almadhor, A., Iqbal, T. and Abbas, S. (2022). An efficient approach for textual data classification using deep learning. Frontiers in computational neuroscience 16: 992296. doi: 10.3389/fncom.2022.992296.

Alrashidi, B., Jamal, A. and Alkhathlan, A. (2023). Abusive Content Detection in Arabic Tweets Using Multi-Task Learning and Transformer-Based Models. Applied Sciences 13(10): 5825. doi: 10.3390/app13105825.

Chowdhury, S. and Schoen, M.P. (2020). Research Paper Classification using Supervised Machine Learning Techniques. In: 2020 Intermountain Engineering, Technology and Computing (IETC). USA: IEEE. 1 - 6. doi: 10.1109/IETC47856.2020.9249211.

Dang, N.C., Moreno-García, M.N. and De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. Electronics 9(3): 483. doi: 10.3390/electronics9030483.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). 4171 - 4186. doi: 10.18653/v1/N19-1423.

Fadli, M., Wijaya, V., Pribadi, M.R. and Widhiarso, W. (2023). Effect of TF-IDF Extraction and Application of SMOTE on Model Performance in Detecting Spam Email. In: 2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Palembang, Indonesia. 637 - 641. doi: 10.1109/EECSI59885.2023.10295851.

Gasparetto, A., Marcuzzo, M., Zangari, A. and Albarelli, A. (2022). A Survey on Text Classification Algorithms: From Text to Predictions. Information 13(2): 83. doi: 10.3390/info13020083.

Graves, L. (2018). Understanding the promise and limits of automated fact-checking (Technical report). Reuters Institute, University of Oxford. doi: 10.60625/risj-nqnx-bg89.

Kim, S.W. and Gil, J.M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. Human-centric Computing and Information Sciences 9: 30. doi: 10.1186/s13673-019-0192-7.

Kuhn, M. and Johnson, K. (2013). Applied predictive modeling. New York: Springer.

Lin, T.Y., Goyal, P., Girshick, R., He, K. and Dollár, P. (2019). Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice. 2980 - 2988. doi: 10.1109/ICCV.2017.324

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692.

Marr, B. (2018). How much data do we create every day? The mind-blowing stats everyone should read. Source: https://www.forbes.com/. Retrieved date 10 July 2024.

McKinsey and Company. (2014). Five facts*: How customer analytics boosts corporate performance.* Source: https://www.mckinsey.com/. Retrieved date 28 August 2024.

Mihalcea, R., Corley, C. and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06). 775 - 780. AAAI Press.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv: 1301.3781. doi: 10.48550/arXiv.1301.3781.

Misra, R. (2022). News category dataset. arXiv: 2209.11429. doi: 10.48550/arXiv.2209.11429.

Mujahidi, M., Kına, E.R.O.L., Rustam, F., Villar, M.G., Alvarado, E.S., De La Torre Diez, I. and Ashraf, I. (2024). Data oversampling and imbalanced datasets: An investigation of performance for machine learning and feature engineering. Journal of Big Data 11(1): 87. doi: 10.1186/s40537-024-00941-x.

Pennington, J., Socher, R. and Manning, C.D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). doi: 10.3115/v1/D14-1162.

Rennie, J.D.M., Shih, L., Teevan, J. and Karger, D.R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In Proceedings of the 20th International Conference on Machine Learning (ICML-03). 616 - 623. AAAI Press.

Stylianou, N., Chatzakou, D., Tsikrika, T., Vrochidis, S. and Kompatsiaris, I. (2023). Domain-aligned data augmentation for low-resource and imbalanced text classification. In: 45 European Conference on Information Retrieval. Springer-Verlag.

Sun, C., Qiu, X., Xu, Y. and Huang, X. (2019). How to fine-tune BERT for text classification? In: Chinese Computational Linguistics. 194 - 206. doi: 10.48550/arXiv.1905.05583.

Veziroglu, M., Veziroglu, E. and Bucak, I.O. (2024). Performance comparison between Naive Bayes and machine learning algorithms for news classification. In: Bayesian Inference-Recent Trends. IntechOpen. doi: 10.5772/intechopen.1002778.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R. and Le, Q.V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. arXiv: 1906.08237. doi: 10.48550/arXiv.1906.08237.

Zhang, X., Zhao, J. and LeCun, Y. (2015). Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems 28 doi: 10.48550/arXiv.1509.01626.

❑❑❑❑❑