# การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง

กิตติศักดิ์ กิตติธานุสรณ์[1*], วีระ สอิ้ง[2], ศุภร คนธภักดี[2]

## บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการจำแนกประเภทของข่าว โดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยใช้ชุดข้อมูลประเภทข่าว ชุดข้อมูลนี้ประเภทข่าวอยู่ 41 ประเภท และหัวข้อข่าว 202,372 หัวข้อตั้งแต่ปี 2555 ถึงปี 2561 ที่ได้รับจากเว็บไซต์ข่าว HuffPost งานวิจัยนี้ใช้อัลกอริทึม การจัดแบ่งประเภทของเอกสาร และการเรียนรู้ของเครื่อง เพื่อจำแนกประเภทข่าว กระบวนการการจำแนกประเภทจะสำรวจเทคนิค bag-of-word และ Term Frequency Inverse Document Frequency (TFIDF) ด้วย 5 การเรียนรู้ คือ Multinomial Naive Bayes, Complement Naive Bayes, Logistic regression, LinearSVC และ Random Forest บนคลาสที่ไม่สมดุล ปัญหาที่ท้าทายนี้จัดการโดยใช้อัลกอริทึมการสุ่มตัวอย่าง 3 วิธี คือ undersampling, synthetic minority oversampling technique (SMOTE) และ adaptive synthetic sampling ผลลัพธ์จากการทดลองพบว่า Logistic regression ที่ใช้เทคนิค bag-of-word และ SMOTE มีประสิทธิภาพสูงที่สุดในการจำแนกประเภทข่าว แสดงค่า Accuracy, Recall, Precision และ F1 score เป็น 80.69, 77.63, 77.04 และ 77.31 ตามลำดับ และจาก confusion matrix แสดงให้เห็นว่ามีความแม่นยำในการตรวจจับข่าวประเภท Healthy Living มากที่สุดคือ 89% แต่มีประสิทธิภาพการตรวจจับข่าวประเภท Sports ค่อนข้างต่ำ

**คำสำคัญ** : กาจำแนกประเภทของข่าว, การจัดแบ่งประเภทของเอกสาร, การเรียนรู้ของเครื่อง, การสุ่มตัวอย่าง

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 062-5565809 E-mail address: kittisak.film@g.swu.ac.th

# NEWS CATEGORY CLASSIFICATION WITH MACHINE LEARNING METHOD

Kittisak Kittitanusorn[1*], Vera Sa-ing[2], Subhorn Khonthapagdee[2]

## Abstract

The purpose of this research is to study the methods of categorizing news using machine learning techniques with a news dataset. This dataset contains 41 news categories and 202,372 headlines from 2012 to 2018, provided by news website HuffPost. In this research, we explore techniques such as bag-of-word and Term Frequency Inverse Document Frequency (TFIDF) techniques with five machine learning methods: Multinomial Naive Bayes, Complement Naive Bayes, Logistic regression, LinearSVC, and Random Forest on asymmetric classes. This challenging problem was addressed by using three sampling algorithms: undersampling, synthetic minority oversampling technique (SMOTE), and adaptive synthetic sampling. Results showed that logistic regression using bag-of-word techniques and SMOTE had the highest accuracy in news classification, with Accuracy, Recall, Precision, and F1 scores of 80.69, 77.63, 77.04, and 77.31, respectively. Using the confusion matrix, it showed that the most accurate in classifying category was Healthy Living news which yielded 89% but the performance of classifying Sport news was quite low.

**Keywords:** News Category Classification, Text Classification, Machine Learning, Sampling

---

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 062-5565809 E-mail address: kittisak.film@g.swu.ac.th

## บทนำ

ข่าว หมายถึง ความเป็นจริงสมบูรณ์ )Completely true) เป็นเรื่องราวหรือเหตุการณ์ที่เกิดขึ้นจากอดีตสู่ปัจจุบันอย่างมีความสัมพันธ์ต่อเนื่อง รวมถึงข้อคิดเห็นด้วย ดังนั้น สิ่งที่ถูกบันทึกในเนื้อข่าวต้องเป็นข้อเท็จจริงที่ยืนยันได้ไม่ว่าอีกกี่ปีข้างหน้า ข้อสำคัญ ข้อเท็จจริงจะเป็นเท็จหรือสมมติขึ้นเองหาได้ไม่ เรื่องราวเหล่านั้นบางครั้งอาจส่งผลกระทบต่อคนหมู่มากทั้งระดับท้องถิ่น หรือระดับประเทศ หรือมวลมนุษย์ในโลก และเมื่อปรากฏเป็นข่าวสู่สาธารณชนสามารถก่อให้เกิดความเข้าใจในตัวมันเองได้ ข้อเท็จจริงที่เป็นข่าวนั้นต้องสามารถจะพิสูจน์ได้ไม่ว่าจะอีกกี่ปีข้างหน้า โดยข่าวมีองค์ประกอบ 3 ส่วน ส่วนแรก คือ พาดหัวข่าว) Headline) เป็นส่วนนำที่สร้างความสนใจ โดยใช้คำที่สะดุดตา และตัวอักษรขนาดใหญ่กว่าเนื้อข่าว ส่วนต่อไป คือ ความนำ) Lead) คือ เนื้อเรื่องย่อของข่าวเป็นการเขียนอธิบายให้ผู้อ่านทราบโดยสรุปว่าเหตุการณ์ที่นำมาเขียนข่าวมีเนื้อความอย่างไร สุดท้ายคือ เนื้อข่าว) Body) คือ รายละเอียดทั้งหมดของข่าว [1]

โดยการจัดประเภทของข่าวสามารถแบ่งออกได้เป็นหลายวิธีตามการแบ่งพิจารณาของในแต่ละแง่มุม โดยอาจแบ่งประเภทข่าวได้จากความรู้สึกของผู้อ่านเนื้อข่าวโดยสามารถแบ่งได้เป็น 2 ประเภท คือ ข่าวหนัก )Head News) หมายถึง ข่าวที่มีเนื้อเรื่องในเชิงสาระ และมีอิทธิพลต่อคนส่วนใหญ่ในสังคม เช่น ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวธุรกิจ ข่าวการศึกษา เป็นต้น และ ข่าวเบา )Soft News) หมายถึง ข่าวที่เกิดขึ้นในกลุ่มคนกลุ่มย่อย ๆ ไม่มีอิทธิพลต่อส่วนใหญ่ในสังคมมากนัก เช่น ข่าวชาวบ้าน ข่าวสังคม บันเทิง ข่าวกีฬา ข่าวอาชญากรรม เป็นต้น

ในการวิจัยครั้งนี้ผู้วิจัยได้นำข้อมูลพาดหัวข่าว) Headline) จากแหล่งข้อมูล HuffPost มาทำการจัดการจำแนกประเภทของข่าวโดยใช้เครื่องมือการเรียนรู้) Machine Learning) เป็นตัวจำแนกประเภทของข่าวโดยวิเคราะห์จากบทความที่พาดหัวข่าวว่าข่าวนี้ถูกจัดอยู่ในประเภทใด โดยมีบทความวิจัยที่ทำการวิจัยโดยการศึกษาการจำแนกประเภทข่าว [2, 3, 4, 5] ดังนั้นจึงมีการปรับปรุงหรือเพิ่มเติมการทดลองจากบทความวิจัย โดยงานวิจัยนี้ได้ใช้หลักการ Text Classification โดยเป็นการจำแนกประเภท การจำแนกหรือแยกแยะประเภทข้อความ เพื่อบอกว่าข้อความนั้นจัดอยู่ในประเภทใด ซึ่งเหมาะสำหรับการสร้างระบบอัตโนมัติในการจัดประเภทหรือประเภทของข่าว โดยผู้วิจัยจะนำกระบวนการและหลักการที่กล่าวขั้นต้นมาใช้กับการวิจัย เพื่อให้มีประสิทธิภาพในการจัดประเภทของข่าว และเนื่องจากตัวข้อมูล มีความไม่สมดุลกันผู้วิจัยจึงทำการปรับตัวข้อมูลให้มีความสมดุลกัน เพื่อเวลานำเข้าไปสู่โมเดลจะทำให้ประสิทธิภาพในการจำแนกประเภทของโมเดลมีประสิทธิภาพมากยิ่งขึ้น โดยผู้วิจัยได้ใช้โมเดล ที่เหมาะแก่การจำแนกประเภทหลากหลายแบบในการเปรียบเทียบการทำงานของโมเดล ว่าโมเดลไหนมีประสิทธิภาพในการจำแนกประเภทที่สุดและเลือกโมเดลที่มีคะแนนมากที่สุดในการจำแนกประเภทของพาดหัวข่าวว่าข่าวนั้นอยู่ในประเภทไหน ดังนั้นงานวิจัยนี้จะเน้นทางด้านการทำให้ข้อมูลสมดุล การใช้กระบวนการ Text Classification และ การเปรียบเทียบประสิทธิภาพโมเดลที่ใช้ในการจำแนกประเภทหรือประเภทของพาดหัวข่าว

## วิธีดำเนินการ

**ขั้นตอนที่ 1 :** แนะนำชุดข้อมูลที่ใช้ในการศึกษานี้

โดยงานวิจัยนี้ได้ใช้ชุดข้อมูล News Category Dataset (Misra, 2018) โดยชุดข้อมูลนี้มีบทความข่าวที่ 202,372 ข่าว ตั้งแต่ปี 2012 ถึง 2018 จากเว็บไซต์ HuffPost ซึ่งเป็นผู้รวบรวมข่าวสารของอเมริกา โดยบทความแต่ละข่าวมีจำนวนประเภทข่าวที่เกี่ยวข้องกัน โดยยกตัวอย่างเช่น ข่าวประเภท POLITICS มีจำนวนบทความ 32739 ข่าว เป็นต้น

ทำการเลือก Feature ที่นำไปใช้ในการเข้าสู่โมเดล โดยจะเปลี่ยน category เป็น label และจะทำการรวม headline, shot description เป็น text และทำการลบ authors กับ data ออก ดังที่แสดงในตารางที่ 1

**ตารางที่ 1** News Category Dataset detail หลังเลือก Feature

| Field name | ความหมาย |
|---|---|
| text | เนื้อหาของข่าว |
| label | ประเภทของข่าว |

**ขั้นตอนที่ 2 :** การเตรียมข้อมูล

การเลือกประเภทข่าวในขั้นตอนนี้จะทำการเลือกประเภทข่าวที่จะไปทำการแก้ความไม่สมดุลของข้อมูล (Imbalanced Datasets) โดยจะทำการเลือกข่าวประเภทที่มีจำนวนบทความเยอะที่สุด 15 อันดับแรก โดยมีข่าวทั้งหมด 142,745 บทความ แสดงในตารางที่ 2

**ตารางที่ 2** แสดงจำนวนบทความของประเภทข่าว 15 อันดับแรก

| Category | Number of text |
|---|---|
| POLITICS | 32739 |
| WELLNESS | 17827 |
| ENTERTAINMENT | 16058 |
| TRAVEL | 9887 |
| STYLE & BEAUTY | 9649 |
| PARENTING | 8677 |
| HEALTHY LIVING | 6694 |
| QUEER VOICES | 6314 |
| FOOD & DRINK | 6226 |
| BUSINESS | 5937 |
| COMEDY | 5175 |
| SPORTS | 4884 |
| BLACK VOICES | 4528 |
| HOME & LIVING | 4195 |
| PARENTS | 3955 |

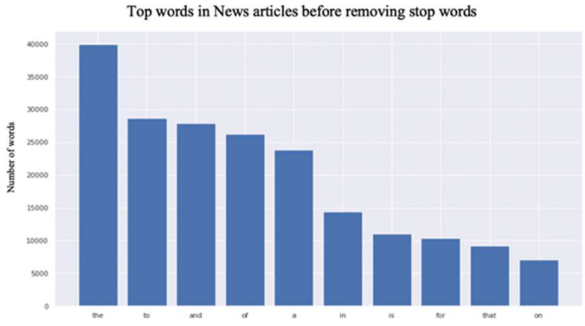
ผู้วิจัยได้ใช้ Label Encoding ในการแทนประเภทข่าวทั้ง 15 ประเภท ด้วยเลข 0 ถึง 14 ตามลำดับ เนื่องจากข้อมูลที่เป็นข้อความเป็นส่วนใหญ่ ทำให้ไม่สามารถนำข้อความใช้เป็น Feature ตรง ๆ จะต้องทำการเปลี่ยนให้เป็นค่าตัวเลขหรือดัชนี เพื่อให้คอมพิวเตอร์สามารถประมวลผลแลแยกแยะได้

**ขั้นตอนที่ 3 :** สำรวจและสรุปหาสิ่งที่อยู่ในข้อมูล

หลังจากที่ได้ทำการเตรียมข้อมูลแล้ว ก็เป็นขั้นตอนการวิเคราะห์สรุปค้นหาข้อมูลที่สำคัญ ที่แฝงอยู่ในกลุ่มข้อมูลนั้น โดย EDA จะทำการหารูปแบบความเชื่อมโยงระหว่างกัน โดยปกติชุดข้อมูล จะแสดงอยู่ในตัวเลข อักขระและข้อความ ซึ่งอาจ

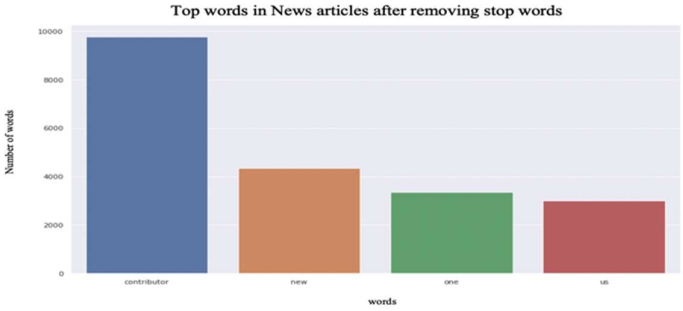
ยากต่อการสรุปทำความเข้าใจหรือการค้นหาความหมายจากชุดข้อมูล ดังนั้นจึงจำเป็นต้องมีกระบวนการใช้วิธีการนำข้อมูลมาแสดงในรูปแบบของภาพกราฟิก แผนภูมิกราฟ ซึ่งช่วยให้เปรียบเทียบ เห็นความสัมพันธ์ของข้อมูลว่าข้อมูลมีความสัมพันธ์อย่างไร ทำให้การศึกษาข้อมูลและวิเคราะห์ง่ายขึ้น โดยการนำเสนอข้อมูลแบบนี้เรียกว่า การสร้างมโนทัศน์ข้อมูล (Data Visualization) ซึ่งมีการใช้เครื่องมือและไลบรารี Matplotlib และ Seaborn

ทำการสร้างกราฟแท่งเพื่อดูว่าบทความข่าวทั้ง 15 ประเภทข่าวมีการใช้คำไหนมากที่สุดในบทความข่าว แสดงในรูปที่ 1




รูปที่ 1 แสดง คำที่ถูกใช้มากที่สุดในบทความข่าวทั้งหมดก่อนลบ stop word

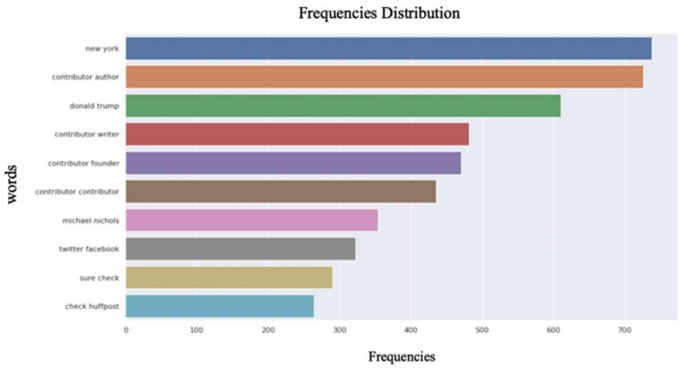
จากรูปที่ 1 จะเห็นได้ว่าในคำในบทความข่าวที่ถูกใช้มากที่สุดจะเป็นคำว่า the, to, and, of เป็นต้น โดย the มีมากถึง 56317 คำ จากบทความทั้งหมด ซึ่งคำเหล่านี้เรียกว่า stop word ดังนั้นจากกราฟเห็นได้ว่าควรลบคำเหล่านั้นออกไปจากรายการคำศัพท์ได้เลย ถึงแม้คำเหล่าถูกนำมาใช้มากจริงในบทความแต่ไม่ค่อยช่วยในการสื่อความหมายสักเท่าไรและอาจทำให้การคัดแยกประเภทของโมเดลผิดพลาดได้ ดังนั้นเมื่อทำการลบ stop word ออกแล้วมาสร้างกราฟใหม่อีกครั้ง ดังรูปที่ 2




รูปที่ 2 แสดง คำที่ถูกใช้มากที่สุดในบทความข่าวทั้งหมดหลังลบ stop word

จากรูปที่ 2 แสดงให้เห็นจำนวนคำที่ถูกใช้มากที่สุดหลังการลบ stop wordโดยคำที่มีมากที่สุดในบทความข่าวทั้งหมดคือคำว่า contributor มีอยู่ 9769 คำ จากบทความข่าวทั้งหมด 45000 บทความ และคำอื่น ๆ รองลงมาได้ new, one, us เป็นต้น

สร้างชุดลำดับคำ (N-Gram) เป็นวิธีของการประมวลผลภาษาที่หาความเป็นไปได้ของชุดลำดับ อักขระหรือคำศัพท์ โดยอาศัยข้อมูลทางสถิติในการคำนวณความเป็นไปได้ของประโยค โดยการคำนวณหาค่าความน่าจะเป็นของความหมายต่าง ๆ โดยมีการระบุชื่อตามหน่วยของขนาดชุดลำดับย่อย ได้แก่ 1 หน่วยย่อย เรียกว่า Uni-gram, 2 หน่วยย่อย เรียวกว่า Bi-gram โดยอันดับแรกจะทำการสร้างแผนภูมิด้วยเทคนิค Bi-gram เพื่อแสดงความถี่1รูปประโยคที่มี 2 คำ ในบทความข่าวทั้งหมดรวมกัน ดังรูปที่ 3




รูปที่ 3 แสดง รูปประโยคคำที่ใช้ถี่ในบทความข่าวทั้งหมดด้วยวิธี N-gram

จากรูปที่ 3 พบว่า การใช้เทคนิค Bi-gram ได้ประโยคคำที่เข้าใจความหมายได้มากกว่า Uni-gram ซึ่งเป็นคำเดียว โดด ๆ เพราะบางคำก็ไม่ควรแยกเป็นคำเดียวเช่น Trump และ Donald เนื่องจากเป็นชื่อของบุคคลจึงควรอยู่รวมกันมากกว่า ดังนั้นความถี่รูปประโยคคำที่พบมากที่สุดในบทความข่าวทั้งหมดคือคำว่า New York ซึ่งมีความถี่ในการใช้ประโยคคำที่ 738 ครั้ง โดยประโยคคำนี้เป็นชื่อเมืองสำคัญของเว็บไซต์ข่าวที่เป็นฐานข้อมูลจึงทำให้มีการใช้เยอะในบทความข่าว และจะใช้วิธี N-gram ด้วยเทคนิค Uni-gram และ Bi-gram ในการสร้างแผนภูมิแสดงความถี่ของคำที่ใช้มากที่สุดในแต่ละประเภทข่าวทั้ง 15 ประเภท โดยจะสรุปเป็นตารางแสดงคำและจำนวนที่ใช้ถี่มากที่สุดอันดับแรกของแต่ละประเภทข่าว โดยเริ่มจาก Uni-gram ที่แสดงของคำเดียว แสดงในตารางที่ 3

**ตารางที่ 3** ตารางแสดงคำและจำนวนที่ใช้ถี่มากที่สุดอันดับแรกของแต่ละประเภทข่าว ด้วย Uni-gram

| Category | Uni-gram Top word | Frequencies |
|---|---|---|
| WELLNESS | contributor | 1899 |
| TRAVEL | contributor | 1895 |
| PARENTING | contributor | 1828 |
| FOOD & DRINK | contributor | 1573 |
| HOME & LIVING | home | 1076 |
| BLACK VOICES | black | 1040 |
| QUEER VOICES | gay | 969 |
| POLITICS | trump | 830 |
| STYLE & BEAUTY | fashion | 645 |
| PARENTS | bologna | 547 |
| COMEDY | mcdonald | 523 |
| HEALTHY LIVING | health | 427 |
| BUSINESS | business | 388 |
| ENTERTAINMENT | new | 341 |
| SPORTS | nfl | 279 |

จากตารางที่ 2 แสดงให้เห็นการนำกลุ่มคำที่ถูกใช้มากที่สุดในแต่ละประเภทข่าวมาเรียงจำนวนการใช้คำจากมากสุดไปน้อยสุด จะเห็นได้ชัดเจนว่ามีประเภทข่าวที่มีการใช้คำเหมือนกันและจำนวนการความถี่ในการใช้คำใกล้กัน ได้แก่ ข่าวประเภท WELLNESS, TRAVEL, PARENTING และ FOOD&DRINK โดยคำที่ใช้คือคำว่า contributor โดยมีความถี่ในการใช้คำอยู่ที่ 1899, 1895, 1828 และ 1573 ตามลำดับ โดยจะเห็นคำ ๆ หนึ่งอาจสามารถถูกตีความให้อยู่ได้หลายประเภทข่าวอาจ

ทำให้เห็นความไม่ชัดเจนของคำในการแยกประเภทข่าว ดังนั้นจึงทำการสร้างแผ่นภูมิ Bi-gram ที่เกิดจากนำคำ2 คำมารวมกันเป็น 1 รูป ประโยค แล้วสรุปแสดงเป็นตาราง ดังแสดงในตารางที่

**ตารางที่ 4** ตารางแสดงคำและจำนวนที่ใช้ถี่มากที่สุดอันดับแรกของแต่ละประเภทข่าว ด้วย Bi-gram

| Category | Bi-gram Top word | Frequencies |
|---|---|---|
| QUEER VOICES | michael nichols | 354 |
| POLITICS | donald trump | 275 |
| STYLE & BEAUTY | sure check | 240 |
| PARENTING | contributor author | 234 |
| WELLNESS | contributor author | 203 |
| COMEDY | donald trump | 194 |
| FOOD & DRINK | food drink | 186 |
| TRAVEL | contributor contributor | 162 |
| HOME & LIVING | craft day | 142 |
| BUSINESS | 247 wall | 122 |
| BLACK VOICES | ofari hutchinson | 75 |
| ENTERTAINMENT | oliver whitney | 71 |
| HEALTHY LIVING | huffington post | 68 |
| SPORTS | super bowl | 63 |
| PARENTS | contributorwriter blogger | 46 |

จากตารางที่ 4 แสดงให้เห็นว่า การซ้ำกันของคำที่อยู่ในแต่ละประเภทข่าวลดลงเหมือนรูปประโยคคำที่ชัดเจนขึ้น โดยจำนวนความถี่ของประโยคคำที่พบมากที่สุดอยู่ที่ประเภทข่าว QUEER VOICE โดยเป็นประโยคคำว่า Michael Nichols ซึ่งเป็นชื่อของบุคคลที่มีความเกี่ยวข้องกับข่าวประเภท QUEER VOICE โดยถูกพบความถี่ของชื่อบุคคลในข่าวประเภทนี้ที่ 354 ครั้ง แต่ก็มีความบางประเภทที่มีการใช้ประโยคคำซ้ำกันอยู่ได้แก่ข่าวประเภท PARENTING และ WELLNESS โดยความถี่ประโยคคำที่พบมากที่สุดของทั้งสองประเภทข่าวคือประโยคคำว่า contributor author ซึ่งมีจำนวนความถี่เท่ากับ 234 และ 203 ตามลำดับ โดยอาจบอกได้ว่าข่าวทั้งสองประเภทนี้อาจมีความใกล้เคียงในลักษณะของประเภทข่าว ดังนั้นการใช้เทคนิคพล๊อต Bi-gram จะทำให้เห็นความชัดเจนของคำหรือประโยคที่มีต่อประเภทข่าว

**ขั้นตอนที่ 4 :** เปรียบเทียบวิธีการสร้างคุณลักษณะข้อมูล

ผู้วิจัยทำการเปรียบเทียบวิธี Term Frequecy Invese Document Frequecy (TFIDF) โดยจะใช้สูตรสมการโดยทั่วไป ดังสมการที่ 1

$$\text{ค่าดัชนี TFIDF} = TF * IF \qquad (1)$$

โดยความหมายของตัวแปร

TF คือ ค่าความถี่คำนั้นในบทความข่าว

TF(z) = จำนวนคำ z ที่มีในบทความข่าว / คำทั้งหมดในบทความข่าว

IDF คือ ค่าส่วนกลับความถี่

IDF(z) = log (จำนวนคำทั้งหมดในบทความข่าว / จำนวนบทความข่าวที่มีคำ z)

โดยมีหลักการและเหตุผลที่ได้จากสมการดังนี้ คือ ในข้อความ หรือ ประโยคใน บทความข่าว ถ้ามีคำใดจำนวนมากตัวเลขค่าความถี่ของคำนั้นจะสูง (ค่า TF สูง) เท่ากับว่าคำนั้นมีค่า Feature คุณลักษณะเด่นหรือมีความสำคัญสูง แต่ถ้าคำใด มีอยู่ในข้อความ หรือ ประโยค ในบทความข่าวอื่น ๆ ด้วยจะมีค่าดัชนีต่ำ (ค่า IDF ต่ำ) เพราะถือว่าบทความข่าวไหน ๆ ก็มีคำนั้น

แสดงให้เห็นว่าไม่ควรหยิบคำนี้มาเป็นคุณลักษณะเด่นแล้ว โดย Feature [6] ของงานวิจัยนี้ คือ list ของ unigram และ bigram จะเห็นได้ว่าการหาค่า Feature ของข้อความ จะมีการใช้ list ของทั้ง unigram และ bigram

และอีกวิธีหนึ่งคือ Bag-of-word (BOW) เป็นการสร้างตารางความถี่คำ มีความคล้ายกับการทำตารางแจกแจงว่ามีคำใด จำนวนเท่าไรโดยการสร้าง ตารางแจกแจงโดยไม่ได้คำนึงถึงหลักไวยากรณ์ ความถี่ที่พบ และลำดับของคำ โดยนำมาใช้เป็น Feature ในการเทรนตัวจัดแบ่งข้อความ Classifier [7]

**ขั้นตอนที่ 5 :** เปรียบเทียบวิธีการสุ่มตัวอย่าง

เนื่องจากการที่ข้อมูลของมีจำนวนข้อมูลในคลาสที่แตกต่างกันมากโดยปกติแล้วชุดข้อมูลที่ดีนั้นควรมีความสมดุล (Balance) เพราะข้อมูลที่ไม่สมดุลของกลุ่มของตัวแปรเป้าหมายที่นำมาศึกษามีผลต่อความถูกต้องของสมการทำนายหรือจำแนกประเภท [8] โดยชุดข้อมูล News Category Dataset ที่ได้นำมาใช้นั้นเป็นชุดข้อมูลที่ไม่สมดุลกัน เนื่องจากประเภทของข่าวมีจำนวนบทความที่แตกต่างกันมาก

ประเภทข่าวที่ได้มาจะมีจำนวนทั้งหมด 15 ประเภทข่าวด้วยกัน โดยจะมีประเภทข่าวที่มีจำนวนบทความเยอะสุดคือ POLITICS ซึ่งมีจำนวนบทความมากถึง 32739 บทความ ในขณะที่ประเภทข่าว PARENTS มีจำนวนบทความเพียงแค่ 3955 บทความเท่านั้น และประเภทข่าวอื่น ๆ ก็มีจำนวนบทความมากน้อยแตกต่างกันไป โดยความแตกต่างที่มากกันเกินไปส่งผลทำให้ตัวข้อมูลไม่มีความสมดุล (Imbalance) ซึ่งหากไม่ทำการจัดการให้ข้อมูลมีความสมดุลจะส่งผลทำให้โมเดลที่ใช้ในการจัดประเภทข่าวมีข้อผิดพลาดได้ ดังนั้นจึงทำการเปรียบเทียบ Sampling Algorithm ด้วยเทคนิคต่าง ๆ ต่อไปนี้

วิธีสุ่มลด (Undersampling) เป็นวิธีการลดจำนวนข้อมูลประเภทข่าวที่อยู่ในกลุ่มที่มีจำนวนข่าวมากเป็นส่วนมากให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มประเภทข่าวที่มีจำนวนข่าวน้อย ๆ [8]

การสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique, SMOTH) เป็นเทคนิคการสุ่มตัวอย่างแบบพิเศษของการสุ่มเพิ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมากใหม่จากข้อมูลเดิมที่มีอยู่ โดยจะสุ่มให้มีจำนวนใกล้เคียง หรือเท่ากับจำนวนข้อมูลในคลาสส่วนมาก [9]

การสุ่มเพิ่มข้อมูลในกลุ่ม (ADASYN) หรือ ADAptive SYNthetic เป็นวิธีปรับปรุงการทำงานของ SMOTH ให้ดีขึ้น ซึ่งในขั้นตอนการสร้างข้อมูลเทียม (Synthetic data) ไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย โดย ADASYN จะใช้ค่าการแจกแจงแบบถ่วงน้ำหนัก (Weight distribution) ของข้อมูลตัวอย่างในกลุ่มน้อย โดยการสร้างข้อมูลเทียมซึ่งขึ้นอยู่กับความสำคัญของข้อมูลนั้น ๆ ถ้าข้อมูลใดยากต่อการแบ่งกลุ่มก็จะให้ค่าของน้ำหนักข้อมูลนั้นมากและสร้างชุดข้อมูลเทียมขึ้นมาในบริเวณนั้น ๆ ซึ่งจะทำให้มีการปรับขอบเขตของการตัดสินใจในการแบ่งกลุ่มดีขึ้น [10]

**ขั้นตอนที่ 6 :** การสร้างแบบจำลองสำหรับจำแนกประเภทข่าว

ในการศึกษานี้ ผู้วิจัยเปรียบเทียบเครื่องมือการเรียนรู้ 5 โมเดล โดยเปรียบเทียบแบบจำลองที่ใช้เทคนิค Bag-of-word และ TFIDF และเลือกโมเดลที่มีประสิทธิภาพสูงสุด 3 อันดับแรก มาทำการเปรียบเทียบด้วยเทคนิคการสุ่มตัวอย่างได้แก่ Undersampling, SMOTH และ ADASYN เพื่อหาโมเดลที่ดีสุดในการจำแนกประเภทข่าว โดยแบบจำลองแต่ละแบบมีรายละเอียดดังนี้

- โมเดลที่ 1 Multinomial Naïve Bayes (Multinomial NB) เป็นอัลกอริทึมใช้ใน Scikit-learn โดยส่วนหนึ่งของ Naïve bayes Classifier เหมาะสำหรับงานสำหรับแยกประเภทความรู้สึก

- โมเดลที่ 2 Complement Naïve Bayes (Complement NB) เป็นอัลกอริทึมที่ดัดแปลงมากจาก Multinomial Naïve Bayes โดยจะมีประสิทธิภาพมากขึ้นในกรณีที่ชุดข้อมูลที่ไม่สมดุล

- โมเดลที่ 3 Logistic Regression การวิเคราะห์ความถดถอยโลจิสติก เป็นเทคนิคในการทำวิเคราะห์สมการถดถอยโลจิสติก เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ

- โมเดลที่ 4 Linear Support Vector Classification (LinearSVC) เป็นวิธีที่งานวิจัยนี้ได้ใช้ในการจำแนกประเภทข่าวสามารถนำไปใช้กับปัญหาหลายคลาส (Multi – Class) โดยวิธีการนี้เป็นการพิจารณาคลาสใดคลาสหนึ่งเทียบกับคลาสอื่น ๆ ที่เหลือทั้งหมด
- โมเดลที่ 5 Random Forest เป็นหนึ่งในกลุ่มของโมเดลที่เรียกว่า Ensemble learning โดยการเรียนรู้แบบ Ensemble จะทำงานได้ดีบนเงื่อนไขที่ว่า โมเดลผู้ทำนายแต่ละตัวจะต้องเรียนรู้อย่างเป็นอิสระต่อกันให้มากที่สุด

## ผลการวิจัยและอภิปรายผลการวิจัย

จากศึกษาวิจัยการจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่องโดยการเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้ในการจำแนกประเภทข่าว ด้วยวิธีการสร้างคุณลักษณะข้อมูลระหว่าง TFIDF และ BOW โดยโมเดลที่ทำการเปรียบเทียบได้แก่ Multinomial NB, Complement NB, Logistic Regression, LinearSVC และ Random Forest โดยแบ่งข้อมูลในสัดส่วน Test 20% ของข้อมูลทั้งหมด โดยได้ตั้งค่าพารามิเตอร์เป็น default ซึ่งผลจากการวิจัยแสดงอยู่ในรูปแบบ Classification Report ที่จะแสดงค่าต่าง ๆ ได้แก่ Accuracy, Precision, Recall และ F1 score โดยผลของการทดลองจะแสดงดังตารางที่ 5

**ตารางที่ 5** ผลลัพธ์จากการทดสอบของโมเดลระหว่าง TFIDF และ BOW

| BOW | | | | |
|---|---|---|---|---|
| Model Classification | %Accuracy | %Recall | %Precision | %F1 score |
| Multinomial Naïve Bayes | 80.83 | 77.85 | 76.98 | 77.32 |
| Complement Naïve Bayes | 77.20 | 67.58 | 80.38 | 70.23 |
| Logistic Regression | 82.48 | 78.22 | 80.46 | 79.26 |
| LinearSVC | 80.28 | 76.15 | 76.84 | 76.47 |
| Random Forest | 78.45 | 71.95 | 78.02 | 74.34 |

| TFIDF | | | | |
|---|---|---|---|---|
| Model Classification | %Accuracy | %Recall | %Precision | %F1 score |
| Multinomial Naïve Bayes | 77.20 | 67.50 | 80.80 | 71.40 |
| Complement Naïve Bayes | 75.40 | 65.40 | 77.10 | 68.00 |
| Logistic Regression | 81.80 | 75.50 | 81.10 | 77.80 |
| LinearSVC | 82.20 | 76.70 | 80.30 | 78.20 |
| Random Forest | 78.10 | 71.10 | 77.60 | 73.60 |

จากตารางที่ 5 ผลการเปรียบเทียบระหว่าง TFIDF และ BOW เห็นได้ว่าโมเดล Logistic Regression ที่ใช้เทคนิค BOW มีค่าประสิทธิภาพในการจำแนกประเภทข่าวสูงที่สุดเมื่อดูจากค่าเปอร์เซ็นต์ของ Accuracy, Recall, Precision และ F1 score ซึ่งจะมีค่าอยู่ที่ 82.48, 78.22, 80.46 และ 79.26 ตามลำดับ ในขณะที่เทคนิค TFIDF โมเดลที่มีค่าสูงสุดคือ LinearSVC

ซึ่งมีค่าเปอร์เซ็นต์อยู่ที่ 82.20 ดังนั้นจึงสรุปผลงานวิจัยได้ว่าในภาพรวมค่าประสิทธิภาพของแต่ละเทคนิค จะพบว่า เทคนิค BOW มีประสิทธิภาพมากกว่า TFIDF ทำให้มีความแม่นยำในการจำแนกประเภทดีที่สุดเมื่อเปรียบเทียบกัน

จากผลลัพธ์การเปรียบเทียบของโมเดลที่ใช้ TFIDF และ BOW พบว่า โมเดลที่ใช้ BOW มีค่าประสิทธิภาพสูงกว่า ดังนั้นจึงเลือก 3 โมเดลแรกของ BOW ที่มีค่าประสิทธิสูง ได้แก่ Logistic Regression, LinearSVC และ Multinomial NB มาทำการเปรียบเทียบการใช้ Sampling Algorithm เพื่อแก้ไขข้อมูลที่ไม่สมดุล (Imbalance) ในแต่ละประเภทข่าว โดยเทคนิค Sampling Algorithm ที่ใช้ได้แก่ Undersampling, SMOTE และ ADASYN โดยพารามิเตอร์ที่ใช้ของ SMOTE และ ADASYN คือ default ส่วน Undersampling ตั้งพารามิเตอร์ n_sample =3000 ดังนั้นจึงทำการเปรียบเทียบค่าประสิทธิภาพของทั้ง 3 เทคนิค Sampling Algorithm และนำมาแยกประเภทข่าวด้วย 3 โมเดลที่เลือกไว้ที่ใช้เทคนิค BOW โดยผลของการทดลองจะมาจาก Test set 20 % ซึ่งจะแสดงตารางที่ 6

**ตารางที่ 6** ผลลัพธ์จากการทดสอบของโมเดลระหว่างด้วยเทคนิค Sampling Algorithm

| Classifier (BOW) | Sampling Algorithm | %Accuracy | %Recall | %Precision | %F1 score |
|---|---|---|---|---|---|
| Multinomial Naïve Bayes | Undersampling | 78.20 | 78.30 | 78.40 | 78.20 |
| | SMOTE | 80.32 | 77.20 | 76.55 | 71.89 |
| | ADASYN | 80.41 | 78.34 | 76.45 | 77.26 |
| Logistic Regression | Undersampling | 78.40 | 78.50 | 78.60 | 78.50 |
| | SMOTE | 80.69 | 77.63 | 77.04 | 77.31 |
| | ADASYN | 80.48 | 77.28 | 76.71 | 76.97 |
| LinearSVC | Undersampling | 74.30 | 74.50 | 74.50 | 74.40 |
| | SMOTE | 79.24 | 75.76 | 75.15 | 75.43 |
| | ADASYN | 79.10 | 75.66 | 74.96 | 75.28 |

จากผลลัพธ์ตารางที่ 6 จะเห็นได้ว่า Sampling Algorithm ที่ใช้เทคนิค SMOTE และใช้โมเดล Logistic Regression มีค่าประสิทธิภาพในการจำแนกประเภทข่าวสูงที่สุดเมื่อดูจากค่าเปอร์เซ็นต์ของ Accuracy, Recall, Precision และ F1 score ซึ่งจะมีค่าอยู่ที่ 80.69, 77.63, 77.04 และ 77.31 ตามลำดับ

ผลลัพธ์จากงานวิจัยได้ว่า Logistic Regression ที่ใช้ BOW และ SMOTH มีค่าประสิทธิภาพสูงที่สุดของของโมเดลทั้งหมด ซึ่งจะสามารถอภิปรายได้ว่า ทำไมถึงเป็นวิธีที่มีประสิทธิภาพดีที่สุดในงานวิจัย โดยจะทำการพิจารณาได้จากดังต่อไปนี้ ผลลัพธ์ค่า Accuracy, Recall, Precision และ F1 score ของโมเดล Logistic Regression ของประเภทข่าวทั้ง 15 ประเภท ดังแสดงในรูปที่ 4

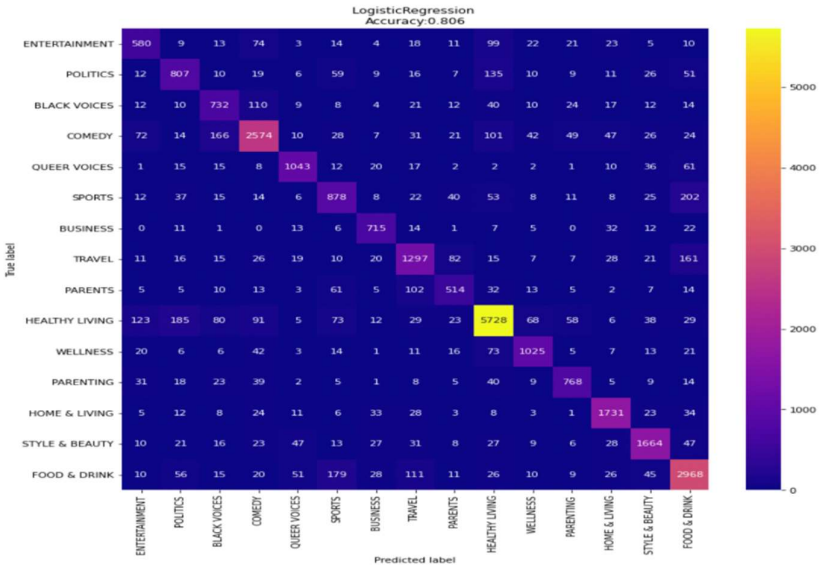
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.67 | 0.64 | 0.66 | 906 |
| 1 | 0.69 | 0.67 | 0.68 | 1187 |
| 2 | 0.70 | 0.71 | 0.70 | 1035 |
| 3 | 0.84 | 0.81 | 0.82 | 3212 |
| 4 | 0.83 | 0.85 | 0.84 | 1245 |
| 5 | 0.66 | 0.64 | 0.65 | 1339 |
| 6 | 0.80 | 0.87 | 0.83 | 839 |
| 7 | 0.72 | 0.76 | 0.74 | 1735 |
| 8 | 0.68 | 0.63 | 0.66 | 791 |
| 9 | 0.90 | 0.88 | 0.89 | 6548 |
| 10 | 0.85 | 0.84 | 0.84 | 1263 |
| 11 | 0.79 | 0.80 | 0.80 | 977 |
| 12 | 0.87 | 0.90 | 0.88 | 1930 |
| 13 | 0.84 | 0.86 | 0.85 | 1977 |
| 14 | 0.81 | 0.84 | 0.82 | 3565 |
| accuracy | | | 0.81 | 28549 |
| macro avg | 0.78 | 0.78 | 0.78 | 28549 |
| weighted avg | 0.81 | 0.81 | 0.81 | 28549 |

รูปที่ 4 ผลลัพธ์ของโมเดล Logistic Regression ของแต่ละประเภทข่าว

จากรูปที่ 4 จะเห็นค่าผลลัพธ์ต่าง ๆของข่าวแต่ละประเภทในโมเดล Logistic Regression โดยเริ่มจากค่า Support ซึ่งคือจำวนตัวทดสอบหรือจำนวนข้อมูล (Sample) ของข่าวที่นำมาใช้ประเมินทดสอบ ซึ่งจากค่าจะเห็นได้ว่า จำวนตัวทดสอบของข่าวแต่ละประเภทอยู่ที่ 791 ถึง 6548 ตัวทดสอบ โดยประเภทข่าวที่มีจำนวนตัวทดสอบเยอะสุดคือ ประเภทข่าวที่ 9 หรือ HEALTHY LIVING และต่ำสุดคือประเภทข่าวที่ 8 หรือ PARENTS ดังนั้นสัดส่วนของ ค่า Support ในแต่ละประเภทยังมีความไกลกันจึงส่งผลอาจทำให้เกิดการเอนเอียง (Bias) และดูค่า Precision และ Recall ควบคู่กันในแต่ละประเภทข่าว

พิจารณาจากค่า Precision หรือ ผลการจำแนกโดยจะดูว่า Model จำแนกข่าวแต่ละประเภทถูกต้องกี่เปอร์เซ็นต์ โดยโมเดลมีการจำแนกประเภทข่าวถูกต้องที่สุดคือข่าวประเภทที่ 9 หรือ HEALTHY LIVING ซึ่งจำแนกถูกต้องถึง 90% จากจำนวนตัวทดสอบทั้งหมด 6548 ตัวทดสอบ ทั้งนี้ประเภทข่าวที่จำแนกได้ถูกต้องน้อยสุดคือ ประเภทข่าวที่ 5 หรือ SPORTS จำแนกถูกต้องแค่ 66% สรุปได้ว่าโมเดล Logistic Regression ที่ทำการ BOW และ SMOTH สามารถจำแนกประเภทข่าว HEALTHY LIVING ได้ถูกต้องที่สุด และพิจารณาจากค่า Recall หรือการดูผลลัพธ์เทียบกับที่เป็นของจริง (Actual) จะบอกได้ที่โมเดลจำแนกมานั้นถูกต้องกี่เปอร์เซ็นต์เมื่อเทียบกับของจริง โดยโมเดลมีความแม่นยำในการจำแนกข่าวประเภทที่ 12 หรือ HOME & LIVING อยู่ถึง 90% เมื่อเทียบกับของจริง

เมื่อดูภาพรวมจากค่า F1 score ซึ่งเป็นค่าที่แสดงประสิทธิภาพ โดยการนำ Precision และ Recall มาคำนวณหาค่าเฉลี่ยจะบอกได้ว่าโมเดล Logistic Regression สามารถจำแนกประเภทข่าวแต่ละประเภทอยู่ในเกณฑ์ดี ยกเว้นข่าวประเภทที่ 5 หรือ SPORTS ที่มีค่าต่ำอยู่ที่ 65% โดยที่ F1 score ยิ่งมีค่าสูงแสดงว่าโมเดลมีประสิทธิภาพดี จากผลลัพธ์ทั้งหมดจะแสดงเป็นกราฟิก Confusion Matrix ได้ในรูปที่ 5




รูปที่ 5 Confusion Matrix ของ Logistic Regression, BOW และ SMOTH ของแต่ละประเภทข่าว

โดยจากรูปที่ 5 จะเห็นได้ว่า Logistic Regression, BOW และ SMOTH มีความแม่นยำในการตรวจจับข่าวประเภท HEALTHY LIVING มากที่สุด ที่จำนวน 5728 จาก ตัวทดสอบ 6548 และประเภทข่าว SPORTS มีการจำแนกผิดมากสุด โดยถูกจำแนกผิดเป็นประเภทข่าว FOOD & DRINK ถึง 202 ตัวทดสอบ โดยอาจมาจากการที่มีคำ ๆ หนึ่งสามารถอยู่ได้ในหลายประเภท

## สรุปผลการวิจัย

งานวิจัยนี้เป็นการศึกษาการจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง โดยจากผลการวิจัยโมเดล Logistic Regression ที่ใช้วิธีการสร้างคุณลักษณะข้อมูลใช้ด้วย Bag-of-word และทำการแก้ข้อมูลที่ไม่สมดุลในแต่ประเภทข่าวด้วยเทคนิค SMOTE เป็นโมเดลที่มีประสิทธิภาพสูงสุดในการจำแนกประเภทข่าวและสามารถจำแนกได้ประเภทข่าวได้ถูกต้องจากหัวข้อข่าวที่ไม่เคยมีการเรียนรู้มาก่อน และจาก confusion matrix แสดงให้เห็นว่ามีความแม่นยำในการตรวจจับข่าวประเภท Healthy Living มากที่สุดคือ 89% แต่มีประสิทธิภาพการตรวจจับข่าวประเภท Sports ค่อนข้างต่ำ โดยการทำ Bag-of-word และ SMOTE สามารถเพิ่มประสิทธิภาพในการจำแนกประเภทข่าวของโมเดลได้

## กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

## เอกสารอ้างอิง

[1] U Suleymanov and S Rustamov, “Automated News Categorization using Machine Learning methods,” 2018
*IOP Conf. Ser.: Mater. Sci. Eng.* 459 012006.

.[2] วาทินี นุ้ยเพียรม, & พยุง มีสัจ. (2556, กันยายน – ธันวาคม). การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวม ของการทำเหมืองข้อความเพื่อการจำแนกข้อความ. วารสารวิชาการ

[3] Ghaidaa A. Al-Sultany, & Hatim, R. M. ( 2018). Semantic Based Short Messages Classification with Topic Modeling Support. *Journal of Engineering and Applied Sciences, 13*, 2407-2412.

[4] Akanksha Patro, Mahima Patel, Richa Shukla, & Jagurti Save. (2020). Real Time News Classification Using Machine Learning. *International Journal of Advanced Science and Technology, 29*, 620-630.

[5] Tsai, & Chen-Wei. (2017). Real-time News Classifier Search Engine Architecture Final Report.

[6] Bijoyan Das, & Sarit Chakraborty. (2018). An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation. *ArXiv, abs/1806.06407*.

[7] กอบเกียรติ สระอุบล. (2020). เรียนรู้ *Data Science* และ *AI:Machine Learning* ด้วย *Python*. กรุงเทพฯ: หสม มีเดีย เนทเวิร์ค.

[8] กาญจน์ ณ ศรีธะ, กิตติศักดิ์ เกิดประสพ, & นิตยา เกิดประสพ. (2561). การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล. วารสารวิทยาการสารสนเทศและเทคโนโลยีประยุกต์, 21-35.

[9] พุทธิพร ธนธรรมเมธี, & เยาวเรศ ศิริสถิตย์กุล. (2562, พฤศจิกายน - ธันวาคม). เทคนิคการจำแนกข้อมูลที่พัฒนาสำหรับชุดข้อมูลที่ไม่สมดุลของภาวะข้อเข่าเสื่อมในผู้สูงอายุ. วารสารวิทยาศาสตร์และเทคโนโลยี, *27*(6), 1165-1178.

[10] Sabina Pun, Sharan Thapa, & Suresh Timilsina. (2019). Customer Churn Prediction Using ADASYN Sampling Technique and Ensemble Model. *Proceedings of IOE Graduate Conference, 6*, 513-518.