การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง

NEWS CATEGORY CLASSIFICATION WITH MACHINE LEARNING METHOD

กิตติศักดิ์ กิตติธานุสรณ์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง

กิตติศักดิ์ กิตติธานุสรณ์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563

NEWS CATEGORY CLASSIFICATION WITH MACHINE LEARNING METHOD

KITTISAK KITTITHANUSORN

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2020

สารนิพนธ์

เรื่อง

การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง

ของ

กิตติศักดิ์ กิตติธานุสรณ์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

.....

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

.....

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.วีระ สอิ้ง)

..... ประธาน

(อาจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วีรยุทธ เจริญเรืองกิจ)

ชื่อเรื่อง	การจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่อง
ผู้วิจัย	กิตติศักดิ์ กิตติธานุสรณ์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	อาจารย์ ดร. วีระ สอิ้ง

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 : การจำแนกประเภทของข่าว, การจัดแบ่งประเภทของเอกสาร, การเรียนรู้ของเครื่อง, การสุ่มตัวอย่าง

Title	NEWS CATEGORY CLASSIFICATION WITH MACHINE LEARNING METHOD
Author	KITTISAK KITTITHANUSORN
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Vera Sa-ing , Ph.D.

The purpose of this research is to study the methods of categorizing news using machine learning techniques with a news dataset. This dataset consisted of 41 news categories and 202,372 headlines from 2012 to 2018, provided by news website HuffPost. In this research, techniques such as bag-of-word and Term Frequency Inverse Document Frequency (TFIDF) were explored, along 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. The results showed that logistic regression using bag-of-word techniques and SMOTE had the highest accuracy in terms of 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 classification category was healthy living news which yielded 89% but the performance of classifying sports news was quite low.

Keyword : News category classification, Text classification, Machine learning, Sampling

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

กิตติศักดิ์ กิตติธานุสรณ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

ในปัจจุบันมีการนำสื่อใหม่มาใช้กับวงการสื่อมวลชนผ่าน สมาร์ทโฟน แท็บเล็ต หรือเครื่องคอมพิวเตอร์รับส่งข่าวสารผ่านทางอีเมลเป็นต้น ปัจจุบัน ผู้รับข่าวสารสามารถอ่านหนังสือพิมพ์และนิตยสาร ได้จากหน้าจอโทรศัพท์เคลื่อนที่ในรูปแบบหนังสืออิเล็กทรอนิกส์ ตลอดจนสามารถรับข้อมูลข่าวสารได้ด้วยตนเองผ่านอินเตอร์เน็ตและสื่อสังคมออนไลน์ โดยปรากฏการณ์ดังกล่าวสร้างความเปลี่ยนแปลงให้แก่กระบวนทัศน์เดิมของการสื่อสารมวลชนในเกือบทุกแง่มุม สิ่งสะท้อนลักษณะ สัมพันธ์ระหว่างเทคโนโลยีการสื่อสารและสังคมของมนุษย์ได้อย่างชัดเจน การเปลี่ยนแปลงกระบวนการการสื่อสารมวลชนที่เห็นได้ชัดเจนอย่างหนึ่งคือ ในแง่เนื้อหาของสื่อมวลชนเนื้อหาสื่อที่เดิมมีลักษณะตายตัวได้เปลี่ยนแปลงไปตามเทคโนโลยีดิจิทัล ทำให้เนื้อหามีความยืดหยุ่น ปรับเปลี่ยนได้ตลอดเวลา โดยการมีส่วนร่วมในการสร้างและผลิตเนื้อหาของผู้รับสาร แม้ว่าเทคโนโลยีใดช่วยในการผลิตเนื้อหาของสื่อมวลชนมีความสะดวก แปลกใหม่และความน่าสนใจขึ้นแต่ขณะเดียวกันผู้ผลิตเนื้อหาบนสื่อดิจิตอลก็ต้องเผชิญกับภาวะแข่งขันและความกดดันจากความก้าวหน้าของเทคโนโลยีด้วย เพราะผู้รับสาร สามารถผลิตเนื้อหาและเผยแพร่ทางสื่อสังคมของตนเองได้ ก็ทำให้มีนักข่าวพลเมืองเกิดขึ้นจำนวนมาก และเนื้อหาของสื่อที่ได้จากผลเมืองก็เป็นปัญหาที่น่าสนใจและสามารถรายงานข่าวได้รวดเร็วกว่าผู้สื่อข่าวมืออาชีพ อีกทั้งปริมาณของข้อมูลข่าวสารจำนวนมากที่ไหลเวียนอยู่ในสื่อดิจิทัล ก็เป็นข้อมูลที่หลากหลายและปะปนกันทั้งเรื่องจริง เรื่องเท็จ และความคิดเห็นส่วนตัวของผู้ใช้อินเตอร์เน็ต โดยธรรมชาติของสื่อดิจิทัล ค่อนข้างมีความยืดหยุ่นและไม่มีแบบแผนที่ตายตัว ทำให้การนำเสนอข่าวในปัจจุบันมีความหลากหลาย

ข่าวดิจิตอลและสำนักข่าวนิยมนำเสนอในปัจจุบัน มี 5 รูปแบบ คือหนังสือพิมพ์ออนไลน์ การนำเสนอข่าวแบบข่าวด่วน การนำเสนอข่าวแบบยั่งยืน การนำเสนอข่าวแบบอธิบายความ และการนำเสนอข่าวแบบให้ติดตามต่อจากลิงค์ข่าวต้นฉบับ โดยรูปแบบแรกคือหนังสือพิมพ์ออนไลน์เป็นการนำข้าวสื่อหลักมานำเสนอบนสื่อออนไลน์โดยใช้รูปแบบการเขียนข่าวเหมือนกับข่าวหนังสือพิมพ์ ซึ่งนิยมใช้การเขียนข่าวแบบปิระมิดหัวกลับคือเรียงเนื้อหาส่วนที่สำคัญสุดมาก่อนแล้วตามด้วยเนื้อหาที่มีความสำคัญรองลงมา การนำเสนอข่าวในลักษณะนี้ สามารถทำได้บนสื่อดิจิทัล ซึ่งอยู่ในรูปแบบของหนังสือพิมพ์ออนไลน์เนื่องจากสื่อดิจิทัลมีพื้นที่มากพอที่จะให้รายละเอียดของข่าวได้เช่นเดียวกับหนังสือพิมพ์ โดยการนำเสนอข่าวลักษณะนี้เกิดขึ้นในช่วงแรก

ที่องค์กรหนังสือพิมพ์เริ่มปรับตัวมาใช้สื่อดิจิทัลเป็นสื่อเสริมหนังสือพิมพ์ฉบับกระดาษ แต่ในปัจจุบันหน้าเว็บไซต์ของหนังสือพิมพ์ก็จะมีการออกแบบที่แตกต่างกันไป รูปแบบถัดไปคือการนำเสนอข่าวแบบข่าวด่วน การนำเสนอข่าวในรูปแบบนี่คือการใช้คุณลักษณะ ด้านความรวดเร็วของสื่อดิจิทัลได้เป็นประโยชน์มาก ซึ่งนิยมใช้การนำเสนอข่าวในลักษณะเป็นการสรุปข่าวสั้นๆให้ผู้รับสารได้ทราบเป็นเบื้องต้นก่อนว่ามีเหตุการณ์อะไรเกิดขึ้นแล้วหลังจากนั้นเมื่อสามารถรวบรวมรายละเอียดได้ก็จะนำเสนอเป็นรายงานข่าวในฉบับเต็มอีกครั้ง โดยการนำเสนอข่าวด่วนนี้มักจะใช้ช่องทางสื่อสังคม ขององค์กรในการนำเสนอ เพราะมีคุณลักษณะ ที่เอื้อให้สามารถนำเสนอเนื้อข่าวและภาพข่าวได้แบบทันทีทันใด บางครั้งอาจจะนำเสนอแค่เพียงพาดหัวข่าวและโปรยข่าวไปก่อน แล้วจึงนำเสนอรายละเอียดอีกครั้ง โดยการนำเสนอในลักษณะนี้มักมีความผิดพลาดเกิดขึ้นได้ง่าย เพราะเน้นในเรื่องความรวดเร็วแต่ ขาดการตรวจสอบความถูกต้อง ทำให้ข่าวขาดความน่าเชื่อถือได้ รูปแบบถัดไปคือการนำเสนอข่าวแบบยั่งยืน ข่าวประเภทนี้ค่อนข้างทำยากแต่มีคุณค่า มีความยั่งยืนและมีประเด็นที่มีนัยสำคัญทางสังคม และสามารถดึงดูดความสนใจและความเห็นหลากหลายจากผู้คนที่ติดตาม เมื่ออยู่ในรูปแบบดิจิทัล แล้วรายงานข่าวชิ้นนั้น ๆ ก็ยังปรากฏอยู่ในรูปแบบรายงานสดต่อไปอย่างยาวนานและเพิ่มพูนเนื้อหาสาระไปได้ตลอดเวลาตามสถานการณ์ และรายละเอียดเพิ่มเติมหรือข้อมูลเสริมจากประชาชนทั่วไปที่เกี่ยวข้องกับเรื่อง นั้น ๆ โดย จะนำเสนอข่าวรูปแบบนี้ อาจต้อง อาศัยผู้สื่อข่าวและองค์กรข่าวมืออาชีพพอใช้ความสามารถในการวิเคราะห์และนำเสนอเนื้อข่าวที่ถูกต้องครบถ้วน มีข้อมูลข่าวสาร ที่มีความน่าเชื่อถือตลอดจนศักยภาพในการเข้าถึงแหล่งข่าวได้ การใช้สื่อดิจิทัลในการนำเสนอข่าวสามารถช่วยสนับสนุนข้อมูล ให้สื่อสารเนื้อหาข่าวออกไปได้อย่างรวดเร็ว และส่งเสริมเกิดการขยายข้อมูลและให้รายละเอียดของข่าวได้จากผู้อ่านที่เข้ามาอ่านและแสดงความคิดเห็น รูปแบบถัดไปคือ การนำเสนอข่าวแบบอธิบายความ เป็นการรายงานข่าวที่เน้นการอธิบาย เรื่องที่สลับซับซ้อนหรือมีหลายมิติ ซึ่งอาจจะมีส่วนที่คล้ายกันกับการนำเสนอข่าวแบบยั่งยืนอยู่มาก แต่ละประเด็นข่าวที่นำเสนอในลักษณะอธิบายความ อาจจะไม่ใช่เรื่องฮือฮาที่สร้างความเกรี้ยวกราดในพาดหัวข่าว แต่เป็นประเด็นที่มีผลกระทบต่อผู้คนทั้งในระยะสั้นและระยะยาว โดยการนำเสนอข่าว แบบอธิบายความมีจุดเด่นที่การให้รายละเอียดของข่าวอย่างรอบคอบ พร้อมทั้งต้องใช้ความสามารถในการสกัดและวิเคราะห์ข้อมูลมาประกอบในรายการข่าวให้เกิดคุณค่าและมีประโยชน์กับผู้อ่านให้มากที่สุด โดย รายงานข่าวลักษณะนี้สามารถใช้เป็นข้อมูลอ้างอิงทางวิชาการได้ รูปแบบสุดท้ายคือการนำเสนอข่าวแบบให้ติดตามต่อจากลิงค์ข่าวต้นฉบับ รูปแบบการนำเสนอข่าวลักษณะนี้นิยมใช้กันอย่างแพร่หลาย เช่นเดียวกันโดยเฉพาะเพจข่าว นอกจากเนื้อข่าวโดยสรุปแล้วยังประกอบไป

ด้วยแง่มุมความคิดเห็นของผู้ดูแลเพจนั้น ๆ ด้วย โดยเพจดังจะมีคุณลักษณะ ที่เฉพาะแตกต่างกันไปตามความคาดหวังและรสนิยมของผู้ติดตาม การให้ข้อมูลข่าวต้นฉบับในลักษณะของลิงค์ที่อยู่ตอนท้ายๆของการนำเสนอข่าว ของการนำเสนอข่าวรูปแบบนี้ อาจไม่จำเป็นเสียด้วยซ้ำสำหรับผู้ติดตามเพราะรสนิยมที่ชื่นชอบการโพสการแชร์ข่าวแบบนี้อยู่แล้ว การนำเสนอข่าวในรูปแบบนี้เกิดจากลักษณะผู้รับสารของสื่อดิจิทัลที่มีเฉพาะกลุ่มมากกว่าผู้รับสารจากสื่อมวลชนกระแสหลัก ดังนั้น การนำเสนอข่าวจากเพจข่าวจึงมีวัตถุประสงค์หลักเพื่อตอบสนองกลุ่มเป้าหมายของตนเอง โดยเฉพาะการใช้ภาษาลีลาในการนำเสนอ และการแสดงความคิดเห็นของผู้เขียนหรือผู้ดูแลเพจรูปแบบการนำเสนอข่าว ดังกล่าวนับว่าเป็นข้อมูลทุติยภูมิ ซึ่งผู้เขียนอาจจะสนใจประเด็นข่าวและหยิบมาพูดเพียงแค่ในบางมุม เท่านั้นในส่วนรายละเอียดของหมดผู้อ่านต้องไปติดตามอ่านจากลิงค์ข่าวต้นฉบับ ดังนั้นการนำเสนอข่าวทั้ง5 รูปแบบ (สุขประเสริฐ, 2560)ที่ถูกกล่าวไว้ข้างต้น เป็นรูปแบบที่เกิดจากการปรับตัวยอมรับนวัตกรรมขององค์กรสื่อ

ในระยะเริ่มต้นที่อินเตอร์เน็ตเข้ามามีบทบาทต่อกระบวนการสื่อสารมวลชนจนมาถึงในยุคปัจจุบันที่เทคโนโลยีดิจิทัลมีความเจริญก้าวหน้าไปอย่างมาก ดั้งนั้นจากการที่ข่าวมีหลายรูปแบบและระบบเครือข่ายอินเทอร์เน็ตที่มีผู้ใช้งานจำนวนมากประกอบด้วยส่วนของเครือข่ายสังคมออนไลน์ที่ปัจจุบันกำลังได้รับความนิยมใช้งานกันมากและยังเป็นแหล่งกำเนิดข้อมูลข่าวสารต่าง ๆ ที่ผู้คนส่วนใหญ่ใช้รับข่าวสารเป็นเรื่องปกติในชีวิตประจำวันข้อมูลข่าวที่อยู่ในรูปแบบอิเล็กทรอนิกส์สามารถส่งต่อได้ง่ายและรวดเร็วเนื้อหาสามารถแพร่กระจายไปได้อย่างกว้างขวางรวดเร็วเครือข่ายสังคมออนไลน์กลายเป็นแหล่งข่าวสำคัญมีข้อมูลปริมาณมากมายมหาศาลสามารถส่งต่อเผยแพร่ไปบนสื่อสังคมออนไลน์ผู้อ่านจึงได้รับรู้เนื้อหาข่าวจากหลายแหล่งในคราวเดียวกัน ปัจจุบันข่าวสารที่เกิดขึ้นจากเครือข่ายสังคมออนไลน์ ทำให้มีปริมาณข่าวเป็นจำนวนมาก ดังนั้น จึงควรมีการจัดหมวดหมู่หรือประเภทของข่าว เพื่อให้ง่ายต่อการเข้าถึงข่าว โดยการจัดหมวดหมู่ข่าว ข่าวที่มีเนื้อหาอย่างเดียวกันหรือคล้ายคลึงกันจะรวมอยู่ในหมวดหมู่เดียวกัน ช่วยให้ผู้รับข่าวมีโอกาสเลือก ข่าวหรือเนื้อหาข่าวตามที่ต้องการจากแหล่งข่าวหลายๆที่ และข่าวที่มีเนื้อข่าวเกี่ยวเนื่องกันหรือสัมพันธ์กันจะอยู่ใกล้ๆกันซึ่งจะ ช่วยให้ผู้รับข่าวสามารถหาข่าวที่มีเรื่องราวเหมือนกันมาใช้ประกอบเนื้อหาข่าวได้อย่างสมบูรณ์ยิ่งขึ้น และทำให้ผู้รับข่าวสามารถค้นหาข่าวสารได้อย่างสะดวกรวดเร็ว และช่วยให้ทราบจำนวนข่าวสารของแต่ละหมวดหมู่ว่ามีจำนวนมากน้อยแค่ไหน เพราะฉะนั้นการจำแนกหมวดหมู่หรือประเภทของข่าวจึงมีความสำคัญในการเพิ่มประสิทธิภาพของข่าวและผู้ที่รับข่าวสารมากยิ่งขึ้น

1.1 ความเป็นมาและความสำคัญของปัญหาการวิจัย

ข่าว หมายถึง ความเป็นจริงสมบูรณ์ (Completely true) เป็นเรื่องราวหรือเหตุการณ์ที่เกิดขึ้นจากอดีตสู่ปัจจุบันอย่างมีความสัมพันธ์ต่อเนื่อง รวมถึงข้อคิดเห็นด้วย ดังนั้น สิ่งที่ถูกบันทึกในเนื้อข่าวต้องเป็นข้อเท็จจริงที่ยืนยันได้ไม่ว่าอีกกี่ปีข้างหน้า ข้อสำคัญ ข้อเท็จจริงจะเป็นเท็จหรือสมมติขึ้นเองหาได้ไม่ เรื่องราวเหล่านั้นบางครั้งอาจส่งผลกระทบต่อคนหมู่มากทั้งระดับท้องถิ่นหรือระดับประเทศ หรือมวลมนุษย์ในโลก และเมื่อปรากฏเป็นข่าวสู่สาธารณชนสามารถก่อให้เกิดความเข้าใจในตัวมันเองได้ ข้อเท็จจริงที่เป็นข่าวนั้นต้องสามารถจะพิสูจน์ได้ไม่ว่าจะอีกกี่ปีข้างหน้า โดยข่าวมีองค์ประกอบ 3 ส่วน ส่วนแรกคือ พาดหัวข่าว (Headline) เป็นส่วนนำที่สร้างความสนใจโดยใช้คำที่สะดุดตา และตัวอักษรขนาดใหญ่กว่าเนื้อข่าว ส่วนต่อไปคือ ความนำ (Lead) คือ เนื้อเรื่องย่อของข่าว เป็นการเขียนอธิบายให้ผู้อ่านทราบโดยสรุปว่าเหตุการณ์ที่นำมาเขียนข่าวมีเนื้อความอย่างไร ความนำที่ดีต้องชัดเจน และทำให้ผู้อ่านเข้าใจเรื่องราว และส่วนสุดท้ายคือ เนื้อข่าว (Body) คือ รายละเอียดทั้งหมดของข่าว ส่วนใหญ่นิยมเขียนเป็นย่อหน้าสั้นๆ หากมีรายละเอียดมาก ก็จะเขียนแยกออกเป็นหลายย่อหน้า โดยเรียงลำดับเหตุการณ์ที่เกิดขึ้น หรือเรียงลำดับความสำคัญจากมากไปหาน้อย โดยแสดงองค์ประกอบทั้ง 3 ส่วน ในภาพประกอบ 1

ภาพประกอบ 1 แสดงองค์ประกอบของข่าวทั้ง 3 ส่วน

ที่มา: พีพีทีวีออนไลน์. (2020). ตัวอย่างข่าว(ออนไลน์). สืบค้นจาก : https://www.pptvhd36.com/sport/news/147125 [15 พฤษาภาคม 2564]

ในปัจจุบันมีการนำเสนอข้อความข่าวผ่านรูปแบบอิเล็กทรอนิกส์หลากหลายรูปแบบ และทำให้เนื้อข่าวที่ถูกนำมาเขียนในรูปแบบอิเล็กทรอนิกส์อาจมีหลากหลายประเภทข่าวอยู่ในนั้น โดยการจัดหมวดหมู่หรือประเภทของข่าวสามารถแบ่งออกได้เป็นหลายวิธีตามการแบ่ง พิจารณาของในแต่ละแง่มุม โดยอาจแบ่งประเภทข่าวได้จากความรู้สึกของผู้อ่านเนื้อข่าวโดยสามารถแบ่งได้เป็น 2 ประเภท คือ ข่าวหนัก (Head News) หมายถึง ข่าวที่มีเนื้อเรื่องในเชิงสาระ และมีอิทธิพลต่อคนส่วนใหญ่ในสังคม เช่น ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวธุรกิจ ข่าวการศึกษา เป็นต้น และข่าวเบา (Soft News) หมายถึง ข่าวที่เกิดขึ้นในกลุ่มคนกลุ่มย่อย ๆ ไม่มีอิทธิพลต่อส่วนใหญ่ในสังคมมากนัก เช่น ข่าวชาวบ้าน ข่าวสังคม บันเทิง ข่าวกีฬา ข่าวอาชญากรรม เป็นต้น หรือประเภทของข่าวซึ่งพิจารณาในแง่ระดับความรู้สึกตอบสนองของผู้อ่านได้แก่ ข่าวที่ ผู้อ่านรู้สึกตอบสนองได้ทันที แต่เป็นการตอบสนองในระยะสั้นๆ ได้แก่ ข่าวบันเทิง อาชญากรรม อุบัติเหตุ และข่าวที่ผู้อ่านรู้สึกตอบสนองช้า เพราะต้องใช้ความคิดพิจารณา ได้แก่ ข่าวการเมือง เศรษฐกิจ การศึกษา และ ประเภทของข่าวซึ่งพิจารณาจากวิธีการนำเสนอข่าว ซึ่งแบ่งเป็น ข่าวที่เสนอโดยเน้นเหตุการณ์ คือ เสนอเฉพาะข้อเท็จจริง และข่าวที่เสนอโดยเน้นที่กระบวนเกี่ยวเนื่องของข่าว คือ เน้นการอธิบาย ตีความ ใช้ลีลาการเขียนแบบสารคดี เพื่อให้รายละเอียดที่เร้าใจ ดึงดูดความสนใจผู้อ่าน โดยถ้าอ้างอิงตามหนังสือพิมพ์ทั่วไปจะแบ่งประเภทข่าวได้ดังนี้ ข่าวในประเทศ (Home News หรือ Local News) คือข่าวประจำแต่ละท้องถิ่น, ข่าวภูมิภาค (Regional News) คือข่าวที่อยู่ในประเทศทวีปเดียวกัน ข่าวประเทศเพื่อนบ้านใกล้เคียง, ข่าวต่างประเทศ (World News) คือหนังสือพิมพ์ในประเทศไทยก็จะหมายถึง ต่างประเทศ ข่าวของอเมริกา ข่าวประเทศแถบยุโรป, บทความ (Comment) คือ การวิจารณ์ในคอลัมน์หนังสือพิมพ์หรือ บทความข่าวที่เป็นประโยชน์, ข่าวธุรกิจ ข่าวนี้แยกย่อยออกมาเป็น เช่น ข่าวธุรกิจในประเทศ (Business) , ข่าวธุรกิจในประเทศเพื่อนบ้าน (Regional Business), ข่าวธุรกิจทั่วโลก (World Business) , ตลาดหุ้น (Stocks Exchange Markets),การเงิน (Finance), ข่าวสังคมธุรกิจ (Social), และ บทความพิเศษ (Features) ,ข่าวเทคโนโลยี (Technology), ข่าวกีฬา (Sports) ซึ่งแยกออกเป็น World Sportsและ Local Sports และข่าวทั่วไป สาระ-บันเทิง แยกออกเป็น ข่าวทั่วไป (General), สารคดี (Documentary), บันเทิง (Entertainment) เป็นต้น และสุดท้ายอาจแบบตามหัวข้อข่าวเช่น ภัยพิบัติ (Disasters), อาชญากรรมและการดำเนินคดี (Crime and Courts),สงครามและการก่อการร้าย (Wars and Terrorisms), การเมือง (Politics), การนัดหยุดงานและกรณีพิพากแรงงาน (Strikes and Disputes), ธุรกิจ การเงิน และเศรษฐกิจ (Business Finance and Economic), วิทยาศาสตร์และสิ่งแวดล้อม (Science and the Environment), กีฬา (Sports), ข่าวทั่วไป

(General) เป็นต้น เพราะฉะนั้นประเภทของข่าวอาจมีผลต่อการเลือกที่จะอ่านของบุคคลนั้น โดยในปัจจุบันข่าวแบบไม่ได้จำแนกอาจส่งผลทำให้คนที่อายุน้อยอาจทำให้ได้รับเนื้อหาข่าวที่รุนแรงเกินไป โดยงานวิจัยพบว่าเรื่องราวข่าวที่ได้รับความนิยมในการเผยแพร่ มากที่สุดเป็นเรื่องที่มีความเกี่ยวข้องกับการเมือง ตามด้วยเรื่องราวความลี้ลับที่เกี่ยวข้องกับตำนาน ข่าวสารทางด้านเศรษฐกิจ ข่าวอาชญากรรมและการก่อการร้าย ข่าววิทยาศาสตร์และเทคโนโลยี ข่าว บันเทิง และข่าวภัยธรรมชาติ การประเมินประเภทข่าวไม่เพียงแต่เกี่ยวข้องกับความน่าเชื่อถือของเนื้อหาข่าวเท่านั้น แต่ยังรวมถึงความน่าเชื่อถือของแหล่งสื่อสังคมออนไลน์อีกด้วย

ดังนั้นการทำ News classification เพื่อประโยชน์ในการแก้ปัญหาการจำแนกประเภทของข่าวสารท่มีปริมาณมากและช่วยประหยัดแรงงานมนุษย์เพราะไม่ต้องใช้มนุนย์ในการจำแนกประเภทข่าว

1.2 วัตถุประสงค์ของการวิจัย

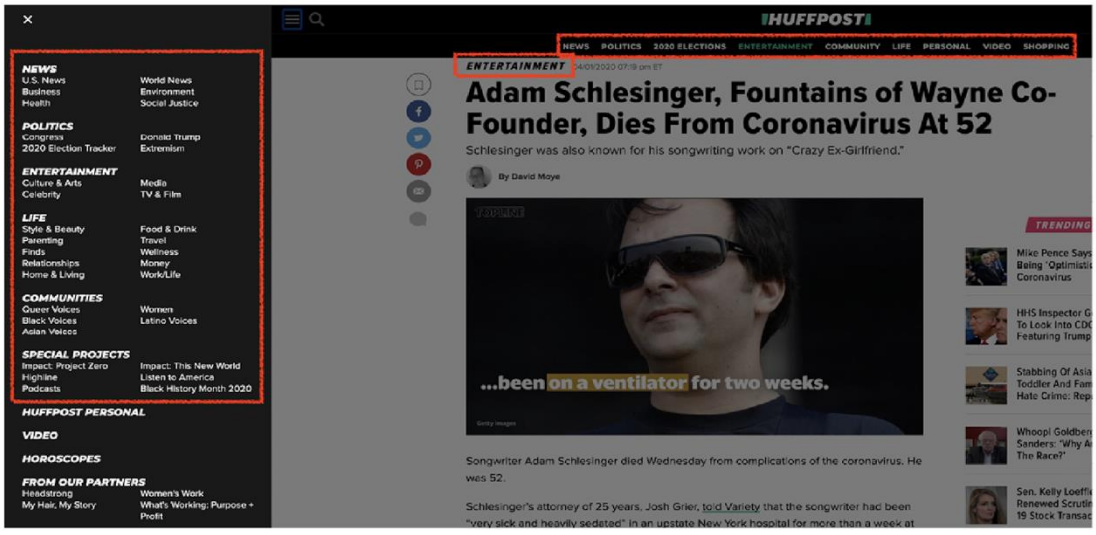
ในการวิจัยครั้งนี้ผู้วิจัยได้นำข้อมูลพาดหัวข่าว (Headline) จากแหล่งข้อมูล HuffPost มาทำการจัดการจำแนกประเภทของข่าวโดยใช้เครื่องมือการเรียนรู้ (Machine Learning) เป็นตัวคัดแยกประเภทของข่าวโดยวิเคราะห์จากบทความที่พาดหัวข่าวว่าข่าวนี้ถูกจัดอยู่ในประเภทใด โดยการใช้หลักการ Text Classification โดยเป็นการจำแนกประเภท การคัดแยกหรือแยกแยะหมวดหมู่ข้อความ เพื่อบอกว่าข้อความนั้นจัดอยู่ในหมวดหมู่ใด ซึ่งเหมาะสำหรับการสร้างระบบอัตโนมัติในการจัดหมวดหมู่หรือประเภทของข่าว โดยผู้วิจัยจะนำกระบวนการและหลักการที่กล่าวขั้นต้นมาใช้กับการวิจัย เพื่อให้มีประสิทธิภาพในการจัดประเภทของข่าว และเนื่องจากตัวข้อมูล มีความไม่สมดุลกันผู้วิจัยจึงทำการปรับตัวข้อมูลให้มีความสมดุลกันด้วยเทคนิคต่าง ๆ เพื่อเวลานำเข้าไปสู่โมเดลจะทำให้ประสิทธิภาพในการจำแนกหมวดหมู่ของโมเดลมีประสิทธิภาพมากยิ่งขึ้น โดยผู้วิจัยได้ใช้โมเดล ที่เหมาะแก่การจำแนกหมวดหมู่หลากหลายแบบในการเปรียบเทียบการทำงานของโมเดล ว่าโมเดลไหนมีประสิทธิภาพในการจำแนกหมวดหมู่ที่สุดและเลือกโมเดลที่มีคะแนนมากที่สุดในการจำแนกหมวดหมู่ของพาดหัวข่าว ว่าข่าวนั้นอยู่ในประเภทไหน ดังนั้นงานวิจัยนี้จะเน้นทางด้านการทำให้ข้อมูลสมดุล การใช้กระบวนการ Text Classification และ การเปรียบเทียบประสิทธิภาพโมเดลที่ใช้ในการจำแนกหมวดหมู่หรือประเภทของพาดหัวข่าว

1.3 ประโยชน์ของการวิจัย

งานวิจัยชิ้นนี้จัดทำขึ้นมาเพื่อศึกษาและทดสอบประสิทธิภาพของแบบจำลองประเภทต่าง ๆ เพื่อค้นหาแบบจำลองที่ดีที่สุดในการจัดประเภทข่าวจากข้อมูลพาดหัวข่าวและทำให้รู้วิธีการปรับข้อมูลที่ไม่สมดุลให้สมดุลขึ้นโดยการใช้เทคนิคต่าง ๆ อีกทั้งยังทำให้ผู้อ่านเนื้อหาสามารถหาอ่านประเภทข่าวที่เหมาะกับตนเองได้ และยังสามารถนำไปต่อยอดใส่ใน Application อื่น ๆ ในการจำแนกประเภทหรือหมวดหมู่เมื่อมีข้อมูลปริมาณเยอะเพื่อลดเวลาในการต้องมาจัดการเรียงหมวดหมู่หรือประเภทด้วยตนเอง และทำให้เข้าใจหลักการของ Text Classification เพื่อใช้ในการจำแนกหมวดหมู่ข้อความอย่างอื่นนอกจากพาดหัวข่าว และยังสามารถเอาไปใช้ในการจำแนกข่าวได้อย่างถูกต้องโดยเฉพาะอย่างยิ่งข่าวที่ยังไม่ได้รับการจัดหมวดหมู่

1.4 ขอบเขตของการวิจัย

การศึกษาการวิจัยนี้เน้นศึกษาชุดข้อมูลเกี่ยวกับการพาดหัวข่าวปี 2012 ถึง 2018 จากแหล่งข้อมูล HuffPost ซึ่งเป็นเว็บไซต์ข่าวออนไลน์รายใหญ่ของสหรัฐอเมริกา โดยจะเป็นข่าวที่เป็นภาษาอังกฤษ แสดงในภาพประกอบ 2



ภาพประกอบ 2 แสดงการจัดประเภทข่าวของเว็บไซต์ HuffPost

ที่มา: HuffPost. (2021). Website home page (ออนไลน์). สืบค้นจาก : https://www.huffpost.com/ [15 พฤษาภาคม 2564]

จากภาพประกอบ 2 แสดงให้เห็นว่าในเว็บไซต์ HuffPost แสดงให้เห็นว่ามีข่าวหลายประเภท โดยชุดข้อมูลที่ได้มาใช้ในงานวิจัยถูกจัดเก็บอยู่ในรูปแบบ json และเมื่อถูกแปลงเป็นตารางจะประกอบไปด้วย ประเภทข่าว, พาดหัวข่าว, ผู้เขียน, ลิงค์ข่าว, บทความนำข่าว และ วันที่ โดยในการวิจัยนี้จะเลือกใช้ ประเภทข่าว และ พาดหัวข่าว ในการจำแนกประเภทของข่าว ดังตารางที่ 1

ตาราง 1 Feature Classification

ตัวแปร	ความหมาย
Category	ประเภทข่าว
Headline	พาดหัวข่าว

โดยตัวแปร category ชุดข้อมูลได้มามีการจัดประเภทอยู่ 41 ประเภทข่าวแต่เนื่องจากความไม่สมดุลของข่าวในแต่ละประเภทผู้วิจัยจึงทำการจัดการให้เหลือเพียง 15 ประเภทข่าว โดยได้แก่ Black Voice, Business, Comedy, Entertainment, Food Drink, Healthy Living, Home Living, Parenting, Parents, Politics, Queer Voice, Sprots, Style Beauty, Travel และ Wellness

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

การศึกษางานวิจัย (วาทินี นุ้ยเพียรม & พยุง มีสัจ, 2556)ได้เปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความคัดเลือกคุณลักษณะการ กรองแบบไคสแควร์ ใช้วิธีการจำแนกประเภทแบบวิธีซัพพอร์ตเวกเตอร์ แมทชีน (Support VectorMachine: SVM)โดยใช้เคอร์เนลแบบโพลิโนเมียลและเรเดียลเบสิสฟังก์ชัน, เนอีฟเบย์ (Naïve Bayes), เบย์เซี่ยนเน็ตเวิร์ก (Bayesian Networks) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ผลการวิจัยพบว่า วิธีซัพพอร์ตเวกเตอร์แมทชีนโดยใช้เคอร์เนลเรเดียลเบสิสฟังก์ชัน (SVMR) ให้ผลการวัด ประสิทธิภาพโดยรวมสูงที่สุดคือ 92.2% และเคอร์เนลแบบโพลิโนเมียล 86.5 % เนอีฟเบย์ 91.7% เบย์ เซี่ยนเน็ตเวิร์ก 91.4% และเคเนียเรสเนเบอร์ 88.7 % ตามลำดับ

การศึกษางานวิจัย (Ghaidaa A. Al-Sultany & Hatim, 2018) ได้ศึกษาการจัดหมวดหมู่ของข่าวและบทความจำนวนมากของชุดข่าวออนไลน์โดยใช้โมเดล Machine learning พร้อมกับใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural language processing) ที่ใช้กันแพร่หลายอย่างมากสำหรับการจัดหมวดหมู่ แต่ในบางครั้งก็เป็นเรื่องยากในการจัดหมวดหมู่ข่าวหรือบทความที่เป็นประเภทข้อความสั้นๆ งานวิจัยนี้จึงเน้นไปที่การสำรวจการจำแนกจากข้อความพาดหัวข่าวโดยใช้อรรถศาสตร์ (Semantics) คือ การศึกษาเกี่ยวกับความหมาย ให้ความสนใจเกี่ยวกับความสัมพันธ์ระหว่างคำ วลี สัญลักษณ์ และความหมาย และเรียนรู้การปรับปรุงการจัดหมวดหมู่ข้อความที่เป็นประเภทสั้นๆ โดยการใช้ Semantics เป็นวิธีเริ่มต้นก่อนการนำไปประมวลผล จากนั้นนำไปคำนวณโดยหาค่า Features โดยใช้ word2vec กับ TFIDF เวกเตอร์ หลังจากนั้นจะนำเข้าสู่โมเดลแบบจำลองในการทำนายการจัดหมู่หมวด โดยเป็นโมเดลที่มีการพัฒนารูปแบบการจำแนกหมวดหมู่ที่แตกต่างกัน โดยโมเดลที่ใช้ได้แก่ K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Naïve Bayes และ Gradient boosting จากผลการทดลองพบว่า Multinomial Naïve Bayes มีประสิทธิภาพในการจำแนกหมวดหมู่ที่ดีที่สุด ด้วยค่าความแม่นยำ (Accuracy) ที่ 90.12% และ ค่าความถูกต้อง (Recall) 90%

การศึกษางานวิจัย (Akanksha Patro, Mahima Patel, Richa Shukla, & Jagurti Save, 2020) ได้กล่าวถึงการมีแหล่งข้อมูลจำนวนมากบนอินเทอร์เน็ตที่สร้างข่าวรายวันจำนวนมหาศาล จึงมีความจำเป็นที่จะต้องจัดประเภทบทความข่าวเพื่อให้ข้อมูลพร้อมใช้งานแก่ผู้ใช้ได้อย่างรวดเร็วและมีประสิทธิภาพ ดังนั้นงานวิจัยนี้จึงทำการจัดประเภทข่าว โดยเริ่มต้นด้วยการรวบรวมบทความข่าวแบบเรียลไทม์จากเว็บไซต์ข่าวด้วยเทคนิคการทำ Web Scraping แล้วทำการจัดประเภทข่าวโดยอัตโนมัติ โดยใช้โมเดลในการจำแนกประเภทต่าง ๆ บทความนี้กล่าวถึงโมเดลในการจัดประเภทข่าวได้แก่ Naïve Bayes, Multinomial Logistic Regression, Support Vector Machine (SVM), Decision tree และ Random Forest สำหรับการจัดหมวดหมู่บทความข่าวโดยอัตโนมัติให้เป็นประเภทต่าง ๆ โดยใช้ชุดข้อมูล จากเว็บไซต์รายงานข่าวนานาชาติของสำนักข่าวบีบีซี (BBC News) ที่มีบทความประเภทข่าวที่แตกต่างกันได้แก่ ธุรกิจ, บันเทิง, การเมือง, กีฬา และเทคโนโลยี โดยบทความนี้จะตรวจสอบผลลัพธ์ของโมเดลการจำแนกประเภทและทำการเปรียบเทียบประสิทธิภาพของโมเดลต่าง ๆ โดยจากผลการทดลองโดยพิจารณาค่าความแม่นยำ (Accuracy), ค่าความเที่ยงตรง (Precision), ค่าความถูกต้อง (Recall) และ ค่าเฉลี่ยระหว่าง Precision และ Recall ที่เรียกว่า F1-Score พบว่าโมเดล Multinomial Logistic Regression ให้ค่าที่ดีที่สุดที่ถึง 95.5% โดยเหมือนทำการตรวจสอบผลลัพธ์การจัดประเภทของข่าวพบ ข่าวประเภทธุรกิจมีการจัดประเภทที่ถูกต้องที่สุดตามด้วยข่าวประเภทกีฬา

การศึกษางานวิจัย (Tsai & Chen-Wei, 2017) ได้เสนอตัวแยกประเภทข่าวแบบเรียลไทม์ที่รวบรวมข่าวจากเว็บไซต์ข่าวของสหรัฐฯ และจำแนกออกข่าวเป็น 7 ประเภท ได้แก่ สหรัฐอเมริกา, โลก, ความคิดเห็น, ธุรกิจ, เทคโนโลยี, ความบันเทิงและกีฬา โดยเลือกใช้โมเดล Naïve Bayes Multinomial(NBM) และ linear kernel Support Vector Machine (SVM) ในการแยกประเภทข่าว และมีการใช้ไลบรารี Python Scikit-learn ที่มีจุดเด่นคือฟังก์ชันในการแบ่งประเภทข้อมูลและการแบ่งกลุ่มข้อมูล โดยมีสูงสุด 30 ความถี่ของคำ(Term) จากการใช้ Term frequency-inverse document frequency (TF-IDF) โดยได้มีแสดงการคำนวณคะแนนสำหรับแต่ละประเภทข่าว โดยผลลัพธ์ที่ได้จากการทดลองในแง่ของข้อผิดพลาดในการทดสอบ (Testing error) จะเห็นได้ว่าประเภทข่าวเทคโนโลยีและความบันเทิง มีค่าความผิดพลาดในการทดสอบค่อนข้างต่ำโดยมีค่าอยู่ที่ 0.1776 และ 0.1107 ตามลำดับ

การศึกษางานวิจัย (J. E. Sembodo, E. B. Setiawan, & M. A. Bijaksana, 2018) มีการเสนอการจัดหมวดหมู่อัตโนมัติจากข้อมูลทวีตตามหมวดหมู่ข่าว โดยจะมีกระบวนการในการจัดหมวดหมู่อยู่ด้วยกันสามส่วน ได้แก่ การรวบรวมข้อมูล (Data collection), การแยกแยะข้อมูล (Data labelling) และการเรียนรู้และการทดลองของเครื่อง (Machine learning and Experiment) โดยผู้วิจัยได้ใช้โมเดลในการจัดหมวดหมู่ข่าวคือ ZeroR, Naïve Bayes Multinomial (NBM), Support Vector Machine (SVM), Random Forest Sequential Minimal Optimization(SMO) และได้มีการนำเอา เวกา (WEKA) ซึ่งเป็นซอฟต์แวร์สำหรับการพัฒนาโปรแกรมโดยใช้การเรียนรู้ของเครื่องและการทำเหมืองข้อมูล นำมาใช้เป็นเครื่องมือจำแนกประเภทสำหรับการเรียนรู้ของเครื่อง ในขณะที่ Microsoft excel (Ms. Excel) จะถูกใช้สำหรับการประมวลผลล่วงหน้า (Preprocessing) และแยกแยะ (Labelling) โดยมีการใช้เทคนิค K-Fold Cross Validation (k-fold cv) โดยบทความได้ตั้งค่า k = 10 ซึ่งเทคนิคนี้จะทำให้ข้อมูลมีการกระจายเท่าๆ กัน และข้อมูลจะถูกดึงโดยใช้ Term frequency-inverse document frequency (TF-IDF) ซึ่งจะสะท้อนให้เห็นถึงความสำคัญของคำในบทความ จากนั้นจะนำข้อมูลที่ผ่านกระบวนการและเทคนิคข้างต้นมาเข้าสู่โมเดล โดยจากผลการทดลองจะพบว่าโมเดล Multinomial Naïve Bayes ให้ค่าความแม่นยำ (Accuracy) ในการทำงานของโมเดลสูงสุดที่ 77.47% ในขณะที่โมเดล ZeroR ให้ผลลัพธ์ค่าความแม่นยำต่ำสุดคือ 19.17% ซึ่งแสดงให้เห็นว่าโมเดล ZeroR ไม่เหมาะกับการใช้จัดหมวดหมู่

การศึกษางานวิจัย (Melek & Mokhtar, 2020) ได้นำชุดข้อมูล “News Category Dataset ซึ่งเป็น data เดียวกันที่ผู้วิจัยสนใจ บทความงานวิจัยนี้ได้ทำการเปรียบเทียบการทำงานของโมเดลโดยการใช้โมเดลที่เป็น Machine learning (ML) 3 โมเดล ได้แก่ Logistic regression, Linear Support Vector Machine และ Naive Bayes Classifier และโมเดลที่เป็น Deep learning (DL) 1 โมเดล คือ LSTM ในการแยกประเภท Topic จากหัวข้อข่าว ซึ่งผลลัพธ์จากบทความวิจัยแสดงให้เห็นว่าโมเดล LSTM มีค่าการทำงานที่ดีที่สุด หรือก็คือมีค่าความแม่นยำ (Accuracy) อยู่ที่ 72% ซึ่งจากข้างต้นจะเห็นว่า บทความงานวิจัยนั้นได้เปรียบเทียบการทำงานของ ML และ DL พบว่า DL ทำงานได้ดีกว่า แต่ผู้วิจัยพบว่ายังมี ML ตัวอื่น ๆ ที่น่าสนใจควรเอามาศึกษาเปรียบเทียบเพิ่มเติม และบทความงานวิจัยไม่ได้แก้ปัญหา imbalance ก่อนการใช้ ML ดังนั้นผู้วิจัยจึงใช้บทความวิจัยนี้ซึ่งใช้ชุดข้อมูลเดียวกันในการสร้างภาพประกอบตาราง 3

	งานวิจัย	บทความวิจัยของ Mina A. Melek
ML Algorithm	Multinomial NB	Naïve bayes Classifier
	Complement NB	
	Logistic Regression	
	LinearSVC	Logistic Regression
	Randomforest	LinearSVC
Imbalance dataset	Undersampling	ไม่ได้ระบุอย่างชัดเจน
	SMOTE	
	ADASYN	

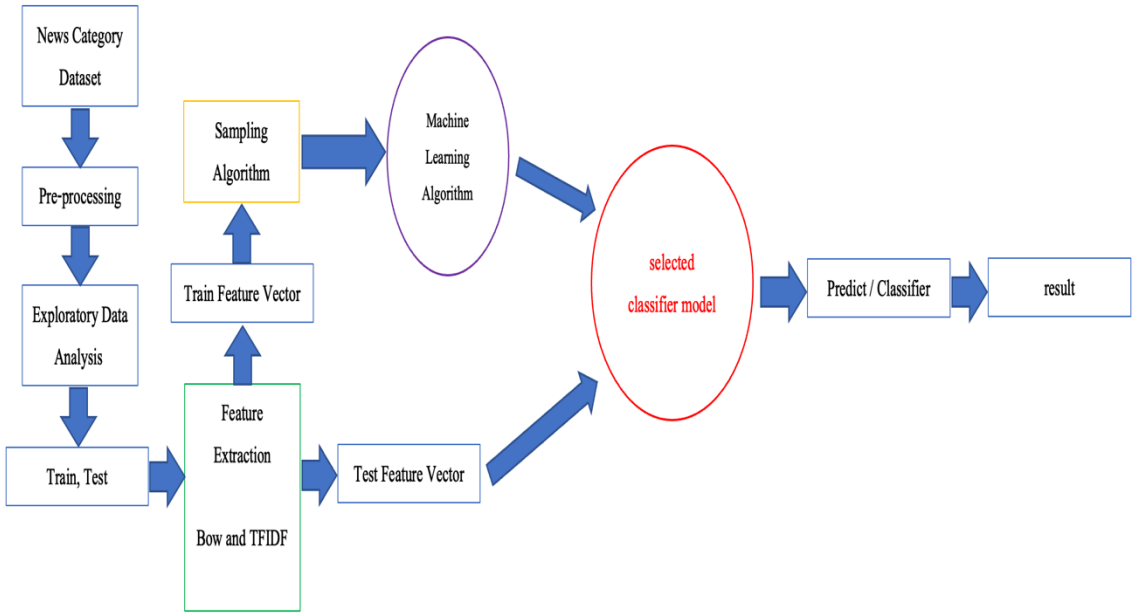
ภาพประกอบ 3 เปรียบเทียบเทคนิคในการแยกประเภทข่าวของผู้วิจัยและบทความวิจัย

ดังนั้นผลการศึกษาจากบทความวิจัยข้างต้น แสดงให้เห็นว่า โมเดลประเภท Naive Bayes, SVM, Logistic Regression และ Random Forest สามารถนำมาใช้ได้ดีมีผลเป็นที่น่าพอใจสำหรับการจัดประเภทข้อมูลข่าว นอกจากนี้ผู้วิจัยยังได้ทำการศึกษาเทคนิคการแก้ข้อมูลไม่สมดุล (Imbalance) ด้วยเทคนิค Undersampling, SMOTE และ ADASYN ในงานวิจัยนี้ด้วย

บทที่ 3
วิธีการดำเนินงานวิจัย

3.1 กระบวนการทำงานวิจัย

ในการวิจัยครั้งนี้ ผู้วิจัยได้มีการใช้กระบวนการ Text Classification ในการจำแนกประเภทของข่าว



ภาพประกอบ 4 กระบวนการ Text Classification

จากภาพประกอบ 4 แสดงกระบวนการ Text Classification โดยกระบวนการนี้เป็นกระบวนการที่ผู้วิจัยได้ใช้ในการจำแนกประเภทหรือหมวดหมู่ของข้อความข่าว โดยการทำงานของ Text Classification แบบ Supervised Machine Learning ประกอบด้วยการนำข้อความมาหาค่า Feature จากนั้นก็ทำการ Train เพื่อสร้างโมเดล เมื่อต้องการจำแนกประเภทจากเนื้อหาข่าว ก็ต้องนำข้อความเนื้อหาข่าวที่ได้มาใหม่ไปผ่าน Feature Extraction เหมือนกับขั้นตอนก่อน Train แล้วจึงทำนายว่าเนื้อหาข่าวใหม่ที่เข้านั้นจัดอยู่ในประเภทใด

3.2 ข้อมูลที่เอามาใช้ในงานวิจัย

โดยงานวิจัยนี้ได้ใช้ชุดข้อมูลจาก Kaggle โดยมีชื่อชุดข้อมูลว่า News Category Dataset (Misra, 2018) โดยชุดข้อมูลนี้มีบทความข่าวที่ 202,372 ข่าว ตั้งแต่ปี 2012 ถึง 2018 จากเว็บไซต์ HuffPost ซึ่งเป็นผู้รวบรวมข่าวสารของอเมริกา โดยบทความแต่ละข่าวมีจำนวนประเภทข่าวที่เกี่ยวข้องกัน โดยยกตัวอย่างเช่น ข่าวประเภท POLITICS มีจำนวนบทความ 32739 ข่าว เป็นต้น โดยจะแสดงตัวอย่างของข่าวในภาพประกอบ 5 และตัว Fied name ของ Dataset ในตาราง 2

	category	headline	authors	link	short_description	date
0	CRIME	There Were 2 Mass Shootings In Texas Last Week...	Melissa Jeltsen	https://www.huffingtonpost.com/entry/texas-ama...	She left her husband. He killed their children...	2018-05-26
1	ENTERTAINMENT	Will Smith Joins Diplo And Nicky Jam For The 2...	Andy McDonald	https://www.huffingtonpost.com/entry/will-smit...	Of course it has a song.	2018-05-26
2	ENTERTAINMENT	Hugh Grant Marries For The First Time At Age 57	Ron Dicker	https://www.huffingtonpost.com/entry/hugh-gran...	The actor and his longtime girlfriend Anna Ebe...	2018-05-26
3	ENTERTAINMENT	Jim Carrey Blasts 'Castrato' Adam Schiff And D...	Ron Dicker	https://www.huffingtonpost.com/entry/jim-carre...	The actor gives Dems an ass-kicking for not fi...	2018-05-26
4	ENTERTAINMENT	Julianna Margulies Uses Donald Trump Poop Bags...	Ron Dicker	https://www.huffingtonpost.com/entry/julianna-...	The "Dietland" actress said using the bags is ...	2018-05-26

ภาพประกอบ 5 ตัวอย่างของข่าว

ตาราง 2 News Category Dataset detail

Field name		
category	headline	authors
link	short description	date

ทำการเลือก Feature ที่นำไปใช้ในการเข้าสู่โมเดล โดยจะเปลี่ยน category เป็น label และจะทำการรวม headline, shot description เป็น text และทำการลบ authors กับ date ออก ดังที่แสดงในตาราง 3

ตาราง 3 News Category Dataset detail หลังเลือก Feature

Field name	ความหมาย
text	เนื้อหาของข่าว
label	ประเภทของข่าว

3.3 การเตรียมข้อมูล

3.3.1 การเลือกประเภทข่าว

ในขั้นตอนนี้จะทำการเลือกประเภทข่าวที่จะไปทำการแก้ความไม่สมดุลของข้อมูล (Imbalanced Datasets) โดยจะทำการเลือกข่าวประเภทที่มีจำนวนบทความเยอะที่สุด 15 อันดับแรก โดยมีข่าวทั้งหมด 142,745 บทความ แสดงในภาพประกอบ 6

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

ภาพประกอบ 6 ตาราง แสดงจำนวนบทความของประเภทข่าว 15 อันดับแรก

3.3.2 การแทนที่ด้วยตัวเลข (Label Encoding)

โดยปกติข้อมูลประเภทหมวดหมู่ (Categorical Data) ที่ไม่ใช่ค่าตัวเลขตรง ๆ ก็ต้องทำการแทนที่ด้วยตัวเลข จะเห็นว่าค่า Feature เป็นตัวเลขทั้งสิ้น แต่สำหรับ Text Classification มีข้อมูลที่เป็นข้อความเป็นส่วนใหญ่ ทำให้ไม่สามารถนำข้อความใช้เป็น Feature ตรง ๆ จะต้องทำการเปลี่ยนให้เป็นค่าตัวเลขหรือดัชนีเพื่อให้คอมพิวเตอร์สามารถประมวลผลแลแยกแยะได้ โดยผู้วิจัยได้ใช้ Label Encoding ในการแทนประเภทข่าวทั้ง 15 ประเภท ด้วยเลข 0 ถึง 14 ตามลำดับ

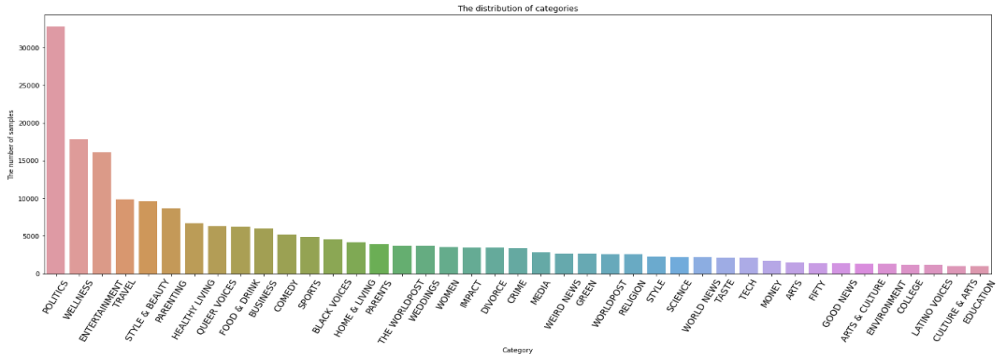
3.4 การสำรวจข้อมูล

3.4.1 สำรวจและสรุปหาสิ่งที่อยู่ในข้อมูล (Exploratory Data Analysis)

หลังจากที่ได้ทำการเตรียมข้อมูลแล้ว ก็เป็นขั้นตอนการวิเคราะห์สรุปค้นหาข้อมูลที่สำคัญ ที่แฝงอยู่ในกลุ่มข้อมูลนั้น โดย EDA จะทำการหารูปแบบความเชื่อมโยงระหว่างกัน โดยปกติชุดข้อมูล จะแสดงอยู่ในตัวเลข อักขระและข้อความ ซึ่งอาจยากต่อการสรุปทำความเข้าใจ หรือการค้นหาความหมายจากชุดข้อมูล ดังนั้นจึงจำเป็นต้องมีกระบวนการใช้วิธีการนำข้อมูลมาแสดงในรูปแบบของภาพกราฟิก แผนภูมิกราฟ ซึ่งช่วยให้เปรียบเทียบ เห็นความสัมพันธ์ของข้อมูล ว่าข้อมูลมีความสัมพันธ์อย่างไร ทำให้การศึกษาข้อมูลและวิเคราะห์ง่ายขึ้น โดยการนำเสนอข้อมูลแบบนี้เรียกว่า การสร้างมโนทัศน์ข้อมูล (Data Visualization) ซึ่งมีการใช้เครื่องมือและไลบรารี Matplotlib และ Seaborn

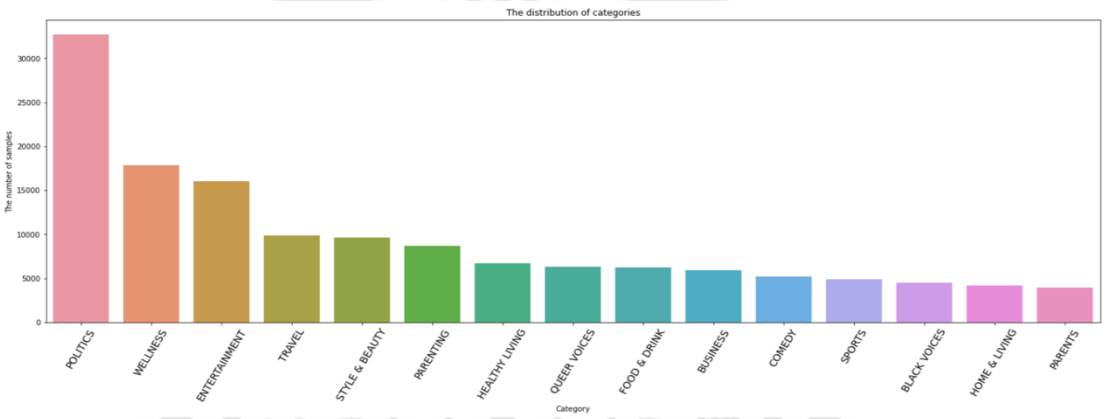
3.4.2 ฮิสโตแกรม (Histogram)

การสร้างฮิสโตแกรมหรือกราฟแท่งด้วย matplotlib โดยมีขั้นตอนประกอบด้วยการ import ไลบรารี โมดูล, การเตรียมข้อมูล x และ y สำหรับพล็อตกราฟ และการสั่งพล็อต โดยข้อมูล x,y อาจได้มาจากการรันสมการในโปรแกรม การอ่านจากไฟล์ Excle หรือ CSV เป็นต้น โดยจากงานวิจัยจะทำการกราฟแท่งโดยจะแสดงความสัมพันธ์ระหว่างจำนวนบทความกับประเภทของข่าว โดยพล๊อตจากข้อมูลตั้งต้นก่อนการทำ Sampling ดังแสดงภาพประกอบที่ 7



ภาพประกอบ 7 แสดงจำนวนบทความของประเภทข่าว 41 ประเภท

จากภาพประกอบ 7 แสดงถึงจำนวนบทความที่แตกต่างกันจากมากไปหาน้อย จากทั้งหมด 41ประเภทข่าว การที่มีจำนวนบทความที่แตกต่างกันมากเกินจะส่งผลโมเดลที่ใช้ในการคัดแยกประเภทข่าวมีการเอนเอียงในการคัดแยกประเภทข่าวได้ โดยถ้าดูจากกราฟที่พล๊อตจะเห็นได้ว่าสามารถทำการตัดบางประเภทข่าวออกได้เช่น EDUCATION, CULTRE & ART, LATION VOICE, COLLEGE เป็นต้น เพราะมีจำนวนบทความที่น้อยมากอยู่ในระดับไม่เกิน 2000 บทความ เมื่อเทียบกับประเภทข่าวที่มีจำนวนเยอะ ๆ เช่น POLITTICS ที่มีถึง 32739 บทความ จึงไม่ส่งผลต่อการคัดแยกโมเดลเท่าไร และเมื่อทำการเลือกประเภทข่าวที่มีจำนวนวนสูงสุด โดยจะทำการสร้างกราฟแท่งเพื่อดูจำนวนบทความของ 15 ประเภทข่าวแรก แสดงดังภาพประกอบ 8

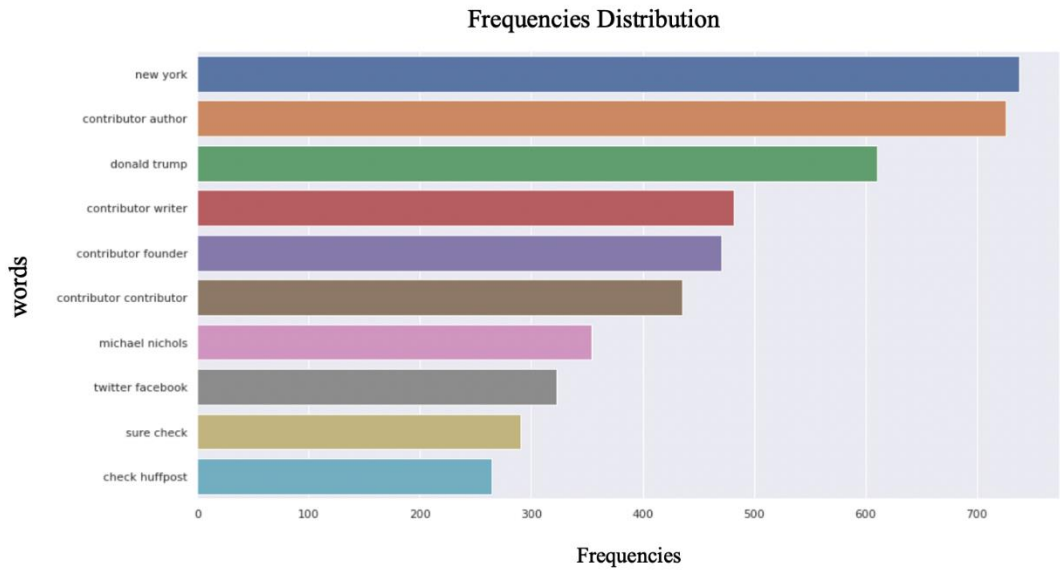


ภาพประกอบ 8 แสดงจำนวนบทความของประเภทข่าว 15 ประเภท หลังการเตรียมข้อมูล

จากภาพประกอบ 8 แสดงถึงจำนวนบทความข่าวทั้งหมดคือ 142,745 บทความ และจำนวนประเภทข่าวที่ทำการคัดเลือกจากประเภทที่มีบทความมากที่สุด 15 ประเภทแรก

3.4.3 ชุดลำดับคำ (N-Gram)

เป็นวิธีของการประมวลผลภาษาที่หาความเป็นไปได้ของชุดลำดับ อักขระหรือคำศัพท์ โดยมีการระบุชื่อตามหน่วยของขนาดชุดลำดับย่อย ได้แก่ 1 หน่วยย่อย เรียกว่า Uni-gram, 2 หน่วยย่อย เรียวกว่า Bi-gram โดยอันดับแรกจะทำการสร้างแผนภูมิด้วยเทคนิค Bi-gram เพื่อแสดงความถี่1รูปประโยคที่มี 2 คำ ในบทความข่าวทั้งหมดรวมกัน ดังภาพประกอบที่ 9

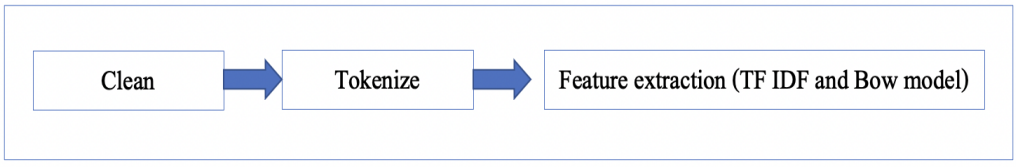


ภาพประกอบ 9 แสดง รูปประโยคคำที่ใช้ถี่ในบทความข่าวทั้งหมดด้วยวิธี N-gram

จากภาพประกอบ 9 พบว่า การใช้เทคนิค Bi-gram ได้ประโยคคำที่เข้าใจความหมายได้มากกว่า Uni-gram ซึ่งเป็นคำเดียวโดด ๆ เพราะบางคำก็ไม่ควรแยกเป็นคำเดียวเช่น Trump และ Donald เนื่องจากเป็นชื่อของบุคคลจึงควรอยู่รวมกันมากกว่า ดังนั้นความถี่รูปประโยคคำที่พบมากที่สุดในบทความข่าวทั้งหมดคือคำว่า New York ซึ่งมีความถี่ในการใช้ประโยคคำที่ 738 ครั้ง โดยประโยคนี้เป็นชื่อเมืองสำคัญของเว็บไซต์ข่าวที่เป็นฐานข้อมูลจึงทำให้มีการใช้เยอะในบทความข่าว

3.5 Feature Extraction

คือ กระบวนการหาค่า Feature ของข้อความ โดยการสกัดหรือคำนวณค่า Feature ของข้อความ ซึ่งเป็นส่วนหนึ่งของขั้นตอนการทำ Text Classification จะมีขั้นตอนย่อยดังแสดงใน (กอบเกียรติ สระอุบล, 2020) ภาพประกอบ 10



ภาพประกอบ 10 แสดงขั้นตอนการหาค่า Feature จากข้อความ

3.5.1 ทำความสะอาด (Clean)

หมายถึงการล้างเอาเครื่องหมายหรืออักขระที่ไม่มีประโยชน์ตอนการประมวลผลออกไป เช่น ข้อความที่มาจากระบบเว็บและมีแท็กปนมา โดยใช้ฟังก์ชัน BeautifulSoup ในการทำความสะอาด ยกตัวอย่างเช่น ข้อความก่อนทำความสะอาด "
<ht>Cybersecurity></ht></td>" หลังทำความสะอากดจะได้ข้อความว่า "Cybersecurity" เป็นต้น และยังใช้เทคนิคอื่นในการทำความสะอาดข้อมูลดังนี้

การตัดคำ (Stop words) โดยเป็นการตัดคำที่มักพบทั่วไปแต่ไม่มีผลกับความหมายหลัก เช่น a, an, the, of และ which เป็นต้น โดยจะเห็นได้ว่ามีจำวนของ Stop words มากในบทความข่าวทั้งหมด ดังที่เคยแสดงในการสร้างกราฟแท่งภาพประกอบ 10 แต่บางครั้งการตัด Stop words ออกไปก็อาจทำให้ความแม่นยำของโมเดลลดลง

การเปลี่ยนรูปคำให้อยู่ในรูปแบบดั้งเดิม (Lemmatization) คือ กระบวนการในการแปลง Word ด้วยรายการคำศัพท์ใน Dictionary, การวิเคราะห์หลักไวยกรณ์ของภาษาอย่างเหมาะสม ในการแปรคำ ผันคำ เพื่อกำจัด Inflection ของคำ เช่น เพศ, Tense, เสียง, อารมณ์, จำนวน และอื่น ๆ ส่วนใหญ่จะตัดส่วนท้าย ให้เหลือแต่รูปฟอร์มพื้นฐาน เป็นคำใน Dictionary เรียกว่า Lemma

3.5.2 ตัดคำ (Tokenize)

เนื่องจากภาษาอังกฤษไม่มีปัญหาในการตัดคำอยู่แล้วเพราะมีการแยกคำมา อยู่แล้วแต่ถ้าเป็นภาษาไทยต้องทำการตัดคำก่อน เช่น "การเรียนรู้ของเครื่องมือ" จะถูกตัดเป็น "การ", "เรียน", "รู้", "ของ", "เครื่อง", "มือ" เป็นต้น

3.5.3 สร้าง Feature

เป็นการคำนวณการสร้างค่าตัวเลขขึ้นมาค่าหนึ่งเพื่อเป็นตัวแทนคุณลักษณะหรือ Feature ของ ข้อความ (Text) นั้น ๆ

3.5.3.1 TFIDF

เป็นวิธีในการคำนวณสร้าง Feature โดยย่อมากจาก Term Frequency และ Inverse Document Frequency โดยจะใช้สูตรสมการโดยทั่วไป ดังสมการที่ 1

$$\text{ค่าดัชนี TFIDF} = TF * IDF \quad (1)$$

โดยความหมายของตัวแปร

TF คือ ค่าความถี่คำนั้นในบทความข่าว

TF(z) = จำนวนคำ z ที่มีในบทความข่าว / คำทั้งหมดในบทความข่าว

IDF คือ ค่าส่วนกลับความถี่

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

โดยมีหลักการและเหตุผลที่ได้จากสมการดังนี้ คือ ในข้อความ หรือ ประโยคในบทความข่าว ถ้ามีคำใดจำนวนมาก ตัวเลขค่าความถี่ของคำนั้นจะสูง (ค่า TF สูง) เท่ากับว่าคำนั้นมีค่า Feature คุณลักษณะเด่นหรือมีความสำคัญสูง แต่ถ้าคำใด มีอยู่ในข้อความ หรือ ประโยค ในบทความข่าวอื่น ๆ ด้วยจะมีค่าดัชนีต่ำ (ค่า IDF ต่ำ) เพราะถือว่าบทความข่าวไหน ๆ ก็มีคำนั้น แสดงให้เห็นว่าไม่ควรหยิบคำนี้มาเป็นคุณลักษณะเด่นแล้ว โดย Feature (Bijoyan Das & Sarit Chakraborty, 2018) ของงานวิจัยนี้ คือ list ของ unigram และ bigram จะเห็นได้ว่าการหาค่า Feature ของข้อความ จะมีการใช้ list ของทั้ง unigram และ bigram

3.5.3.2 Bag-of-word (BOW)

เป็นการสร้างตารางความถี่คำ มีความคล้ายกับการทำตารางแจกแจงว่ามีคำใดจำนวนเท่าไรโดยการสร้าง ตารางแจกแจงโดยไม่ได้คำนึงถึงหลักไวยากรณ์ ความถี่ที่พบ และลำดับของคำ โดยนำมาใช้เป็น Feature ในการเทรนตัวจัดแบ่งข้อความ Classifier (George K & Joseph, 2014)

3.6 Sampling Algorithm

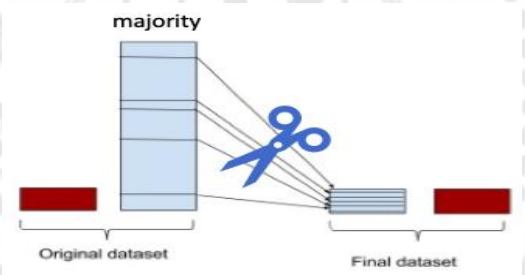
3.6.1 **ชุดข้อมูลที่ไม่สมดุล** (Imbalanced Datasets)

Imbalance data set คือ การที่ข้อมูลของเรามีจำนวนข้อมูลในคลาสที่แตกต่างกันมากโดยปกติแล้วชุดข้อมูลที่ดีนั้นควรมีความสมดุล (Balance) เพราะข้อมูลที่ไม่สมดุลของกลุ่มของตัวแปรเป้าหมายที่นำมาศึกษามีผลต่อความถูกต้องของสมการทำนายหรือจำแนกหมวดหมู่ (กาญจน์ ณ ศรีธะ, กิตติศักดิ์ เกิดประสพ, & นิตยา เกิดประสพ, 2561)โดยชุดข้อมูล News Category Dataset ที่ได้นำมาใช้นั้นเป็นชุดข้อมูลที่ไม่สมดุลกัน เนื่องจากประเภทของข่าวมีจำนวนบทความที่แตกต่างกันมาก

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

3.6.2 การสุ่มลด (Undersampling)

วิธีสุ่มลด (Undersampling) เป็นวิธีการลดจำนวนข้อมูลประเภทข่าวที่อยู่ในกลุ่มที่มีจำนวนข่าวมากเป็นส่วนมาก ให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลที่อยู่ในกลุ่มประเภทข่าวที่มีจำนวนข่าวน้อย ๆ (Ristu Saptono & Winarno Win, 2018) โดยแสดงวิธี Undersampling ในภาพประกอบ 11

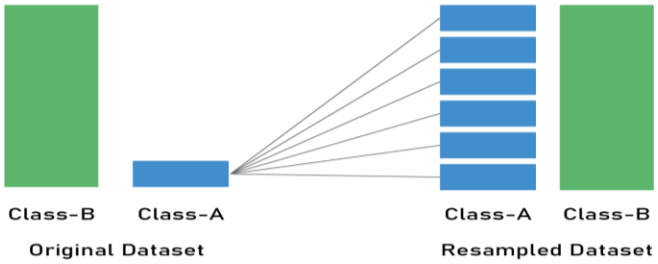


ภาพประกอบ 11 วิธีการสุ่ม ลด (Undersampling)

ที่มา: Dataman. (2018). Using Under-Sampling Techniques for Extremely Imbalanced Data (ออนไลน์). สืบค้นจาก : https://medium.com/dataman-in-ai/sampling-techniques-for-extremely-imbalanced-data-part-i-under-sampling-a8dbc3d8d6d8 [20 พฤษาภาคม 2564]

3.6.3 การสังเคราะห์ข้อมูลเพิ่ม (SMOTE)

หรือ Synthetic Minority Oversampling Technique เป็นเทคนิคการสุ่มตัวอย่างแบบพิเศษของการสุ่มเพิ่ม แทนที่จะสุ่มเพิ่มโดยใช้ข้อมูลเดิมแต่จะทำการสังเคราะห์ข้อมูลขึ้นมากใหม่จากข้อมูลเดิมที่มีอยู่ โดยจะสุ่มให้มีจำนวนใกล้เคียง หรือเท่ากับจำนวนข้อมูลในคลาสส่วนมาก (พุทธิพร ธนธรรมเมธี & เยาวเรศ ศิริสถิตย์กุล, 2562) โดยแสดงวิธี SMOTE ในภาพประกอบ 12

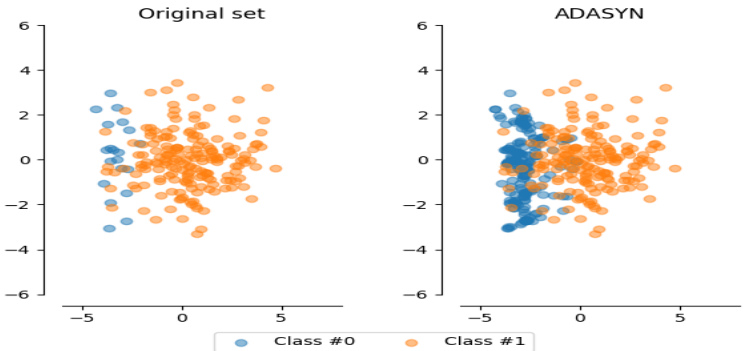


ภาพประกอบ 12 วิธีการสังเคราะห์ข้อมูลเพิ่ม (SMOTE)

ที่มา: Younes Charfaoui. (2019). Resampling to Properly Handle Imbalanced Datasets in Machine Learning (ออนไลน์). สืบค้นจาก : https://heartbeat.fritz.ai/resampling-to-properly-handle-imbalanced-datasets-in-machine-learning-64d82c16ceaa [20 พฤษาภาคม 2564]

3.6.4 การสุ่มเพิ่มข้อมูลในกลุ่ม (ADASYN)

หรือ ADAptive SYNthetic เป็นวิธีปรับปรุงการทำงานของ SMOTH ให้ดีขึ้น ซึ่งในขั้นตอนการสร้างข้อมูลเทียม (Synthetic data) ไม่จำเป็นต้องพิจารณาข้อมูลทุกตัวที่อยู่ในกลุ่มน้อย โดย ADASYN จะใช้ค่าการแจกแจงแบบถ่วงน้ำหนัก (Weight distribution) ของข้อมูลตัวอย่างในกลุ่มน้อย โดยการสร้างข้อมูลเทียมซึ่งขึ้นอยู่กับความสำคัญของข้อมูลนั้น ๆ ถ้าข้อมูลใดยากต่อการแบ่งกลุ่มก็จะให้ค่าของน้ำหนักข้อมูลนั้นมากและสร้างชุดข้อมูลเทียมขึ้นมาในบริเวณนั้น ๆ ซึ่งจะทำให้มีการปรับขอบเขตของการตัดสินใจในการแบ่งกลุ่มดีขึ้น (Sabina Pun, Sharan Thapa, & Suresh Timilsina, 2019)



ภาพประกอบ 13 วิธีการสุ่มเพิ่มข้อมูลในกลุ่ม (ADASYN)

ที่มา: Rui Nian. (2019). Fixing Imbalanced Datasets: An Introduction to ADASYN (ออนไลน์). สืบค้นจาก : https://medium.com/@ruinian/an-introduction-to-adasyn-with-code-1383a5ece7aa [20 พฤษาภาคม 2564]

3.7 การสร้าง Training dataset และ Test dataset

หลังจากที่ทำการเตรียมข้อมูล, การสำรวจข้อมูล และ Feature Extraction ก็จะทำการแบ่งข้อมูลเป็น Train และ Test โดย Train คือ ชุดข้อมูลที่นำไปทำการสอนให้กับคอมพิวเตอร์ ส่วน Test คือ ชุดข้อมูลที่แบ่งมาจาก Data set เพื่อนำมาทดสอบความแม่นยำ ความถูกต้องของ Model ที่ Train เรียบร้อยแล้ว โดยงานวิจัยนี้จะทำการแบ่งข้อมูลในสัดส่วน Test 20% ของข้อมูลทั้งหมด

3.8 Machine Learning Algorithm /Classifiers

โดยขั้นตอนนี้จะเป็นเลือก หลักการคิดการคำนวณประมวลผล (Algorithm) ของการเรียนรู้ด้วยเครื่อง (Machine Learning) หรือเรียกกว่า โมเดล (Model) ในการคัดแยกประเภทของข่าว โดยเป็นวิธีหรือขั้นตอนกระบวนคิดคำนวณทางคณิตศาสตร์เพื่อให้ได้ผลลัพธ์ออกมา โดยการเรียนรู้ด้วยเครื่อง (Machine Learning) จะมี Algorithm อยู่เป็นจำนวนมาก โดยงานวิจัยนี้จะเลือกใช้โมเดลในการเปรียบเทียบค่าประสิทธิภาพของโมเดลในกระบวนการ Feature Extraction ระหว่าง TFID และ BOW ในการจำแนกประเภทข่าว โดยโมเดลที่นำมาใช้ได้แก่

Naïve bayes Classifier เป็นอัลกอริทึมพื้นฐานที่สำคัญตัวหนึ่ง ที่สามารถใช้ได้ทั้งการจำแนก (Classification) และ Regression โดยใช้ทฤษฎีของเบย์ (Bayes' theorem) โดยอาศัยความน่าจะเป็น (Probability) เพื่ออนุมานคำตอบที่ต้องการ เช่น คลาสหรือประเภทของตัวอย่างบนสมมติฐานว่า ปริมาณที่สนใจจะอยู่ภายใต้การแจกแจงความน่าจะเป็น (Probability Distribution) ดังนั้น ความน่าจะเป็นของตัวอย่างจึงสามารถใช้ประกอบการตัดสินใจอย่างมีเหตุผลได้ในงานจำแนกประเภท เพื่อใช้ในการจำแนกประเภท (Gurmeet Kaur & Karan Bajaj, 2016) จากทฤษฎีของเบย์ สามารถกำหนดสมการที่ 2

$$\boldsymbol{P(h|D)} = \frac{\boldsymbol{P(D|h)P(h)}}{\mathbf{P(D)}} \qquad (2)$$

โดยสัญลักษณ์ ต่าง ๆ ของสมการอธิบายได้ดังนี้

เมื่อ P(h | D) เป็นความน่าจะเป็น ของการเกิด h บนเงื่อนไข การเกิดของ D มาก่อน

P(h) เป็นความน่าจะเป็นของการเกิด h

P(B | A) เป็นความน่าเป็น ของการเกิด D บนเงื่อนไขของการเกิด h มาก่อน

Multinomial Naïve Bayes (Multinomial NB) เป็นอัลกอริทึมใช้ใน Scikit-learn โดยส่วนหนึ่งของ Naïve bayes Classifier เหมาะสำหรับงานสำหรับแยกประเภทความรู้สึก โดยคิดว่า ประเภทหนึ่ง ๆ(ที่แยกประเภท) มีคุณสมบัติได้มากมาย ในตัวอย่างการคำนวณประเภทเอกสารเพื่อจำแนกความรู้สึก สำหรับการคำนวณค่าแต่ละค่าจะเริ่มจากการประมาณค่า Prior ซึ่งเป็นความน่าจะเป็นของประเภทในแต่ละด้าน โดยสมการจะให้ N_c แทน จำนวนเอกสาร ประเภท c และ N_{doc} แทนจำวนเอกสารทั้งหมด (Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon, & Anees Ahmed, 2019) ดังนั้นแล้วจะแสดงเป็นดังสมการที่ 3

$$\boldsymbol{P(c)} = \frac{N_c}{N_{doc}} \tag{3}$$

ในการคำนวณค่า Likelihood (P(f |c)) คำนวณจาก จำนวนคำที่เลือกในประเภท c ที่ปรากฏในเอกสารต่อจำนวนเอกสารทั้งหมด โดยจะใช้สัญลักษณ์ w แทน คำ และ v แทนคำทุกประเภท จึงแสดงดังสมการที่ 4

$$P(w_i|c) = \frac{count(w_i, c)+1}{(\sum_{w\in} v^{count(w_i, c))+v}} \tag{4}$$

จากสมการที่ 4 จำเป็นต้องบวกหนึ่ง ทั้งเศษและส่วนเพราะอาจเป็นไปได้ว่า จะไม่พบคำประเภทหนึ่งได้ ซึ่งจะให้ค่าเป็นสูตรได้ เมื่อได้ Likelihood เป็นศูนย์จะทำให้นำไปคูณกับ Prior เป็นศูนย์ได้ จึงต้องบวกหนึ่งเพื่อป้องกันค่าเป็นศูนย์

Complement Naïve Bayes (Complement NB) เป็นอัลกอริทึมที่ดัดแปลงมากจาก Multinomial Naïve Bayes โดยจะมีประสิทธิภาพมากขึ้นในกรณีที่ชุดข้อมูลที่ไม่สมดุล โดยได้ใช้สถิตจากองค์ประกอบของแต่ละคลาสเพื่อคำนวณน้ำหนักของโมเดล จะแสดงให้เห็นค่าพารามิเตอร์การประมาณความน่าจะเป็น ของ Complement Naïve Bayes มีความเสถียรมากกว่าและมีประสิทธิภาพสูงกว่าในการจัดประเภทข้อความ ดังนั้นสมการการประมาณค่าความน่าจะเป็นแบบมีเงื่อนไข (Siva RamaKrishna Reddy V, D V L N. Somayajulu, & Ajay R. Dani, 2010) จะแสดงดังในสมการที่ 5

$$\theta_{ci} = \frac{N_{ci}+\alpha_i}{N_c+\alpha} \tag{5}$$

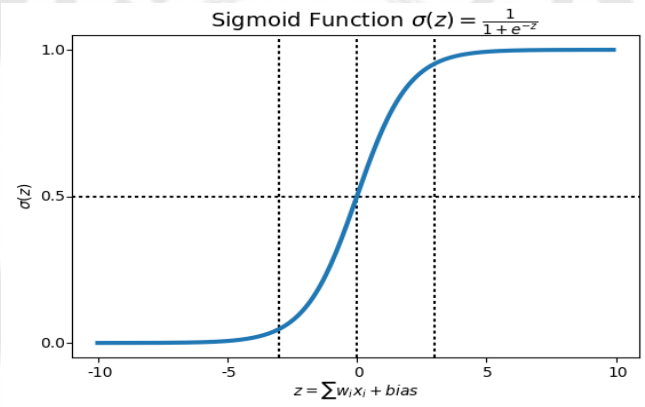
โดยสัญลักษณ์ ต่าง ๆ ของสมการอธิบายได้ดังนี้

เมื่อ N_{ci} คือจำนวนครั้งของคำ i ที่ปรากฏขึ้นในเอกสารในคลาสอื่น ๆ ที่ไม่ใช่ c

N_c คือผลรวมจำนวนคำที่เกิดขึ้นในคลาสอื่นที่ไม่ใช่ c

α และ α_i คือ พารามิเตอร์ที่ทำให้ smoot

Logistic Regression การวิเคราะห์ความถดถอยโลจิสติก เป็นเทคนิคในการทำวิเคราะห์สมการถดถอยโลจิสติก เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ โดยตัวแปรมีเพียงสองค่าคือ 0 และ 1 หากตัวแปรอิสระมีค่าน้อย ค่าของตัวแปรตามจะมีค่าเท่ากับ 0 และหากตัวแปรอิสระมีค่ามากค่าของตัวแปรตามจะมีค่าเท่ากับ 1 และนำสมการความถดถอยที่ได้ไปประมาณหรือพยากรณ์ค่าตัวแปรตาม เมื่อกำหนดตัวแปรอิสระ(สุรวัชร ศรีเปารยะ & สายชล สินสมบูรณ์ทอง, 2560) โดยแสดงตัวอย่างกราฟฟังก์ชันในภาพประกอบ 14



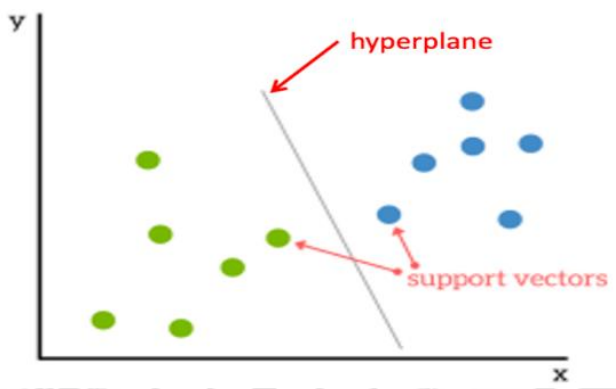
ภาพประกอบ 14 ฟังก์ชันโลจิสติก

ที่มา: Puja Sengupta. (2019). ORIGINAL RESEARCH PAPER CONTINUOUS MONITORING SYSTEM FOR ANALYSING VEGETATIVE STATE AND COMATOSE PATIENTS WITH THE HELP OF SOMATOSENSORY EVOKED POTENTIAL AND DEEP LEARNING (ออนไลน์). สืบค้นจาก : https://www.researchgate.net/figure/Fig-3-Sigmoid-function-graph-CourtseyInternet-sources_fig3_330761350 [25 พฤษาภาคม 2564]

จากภาพประกอบ 14 แสดงการสร้างกราฟขึ้นมา โลจิสติกจะสร้างกราฟโดยพยายาม fit ชุดข้อมูลให้เข้ากับ sigmoid function ซึ่งการที่จะประมาณค่าได้ซักค่านึง สามารถทำได้ด้วยการเอาค่า x ป้อนไปในฟังก์ชัน หาค่าของโลจิสติกถ้ามากกว่า 0.5 ก็จะจัดเป็นหมวดเดียวกัน ถ้าต่ำกว่าก็จะจัดเป็นอีกหมวด ซึ่ง Logistic Regression สามารถทำได้มากกว่าแค่แบ่งประเภท

ของสิ่งของแค่ 2 สิ่งเหมือนกัน โดยจำนวนกลุ่มของตัวแปรตามมี 2 กลุ่มจะเรียกว่า การวิเคราะห์ความถดถอยโลจิสติก 2 กลุ่ม (Binary Logistic Regression) จำนวนกลุ่มของตัวแปรตามมี 2 กลุ่มจะเรียกวาการวิเคราะห์ความถดถอยโลจิสติกพหุกลุ่ม (Multinomial Logistic Regression)

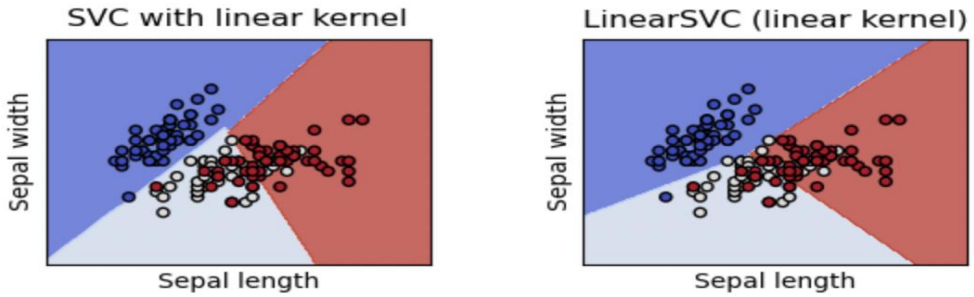
Support Vector Machine (SVM) จัดเป็นอัลกอริทึมประเภท Supervised Learning โดยใน Scikit-learn สามารถใช้ทั้งในการจำแนก (Classification) และ Regression ซึ่งถ้าเป็น Regression จะเรียกว่า Support Vector Regression (SVR) โดยงานวิจัยนี้เป็นแบบการจำแนก จึงเลือกใช้อัลกอริทึม SVM โดยแสดงกราฟของอัลกอริทึม SVM ดังภาพประกอบ 15



ภาพประกอบ 15 แสดง Support Vector Machine

ที่มา: Noel Bambrick. (2016). Support Vector Machines: A Simple Explanation (ออนไลน์). สืบค้นจาก : https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html [25 พฤษาภาคม 2564]

จากภาพประกอบ 15 จะแสดงหลักการของ SVM คือการนำค่าของกลุ่มข้อมูลมาสร้างลงในพื้นที่เรียกว่า ฟีเจอร์สเปซ (Feature Space) จากนั้นหาระนาบเกิน (Hyperplane) ที่ใช้แบ่งข้อมูลกลุ่ม (Class) ออกจากกัน โดยจะสร้างแนวที่ดีที่สุดในการแบ่งกลุ่ม ซึ่งจะมีกลไกเส้นแบ่ง (Kernel) ในการเป็นเป็นกลไกหรือตัวสร้างเส้นที่ใช้แบ่งกลุ่มข้อมูล โดยงานวิจัยนี้จะสนใจ Kernel ที่เป็นแบบ Linear (kernel = ‘linear’) ซึ่งเป็นการสร้างแบบเส้นตรง โดยจะแสดง Kernel ในแบบ Linear (J SreeDevi, M Rama Bai, & M Chandrashekar Reddy, 2020) โดยแสดงตัวอย่างในภาพประกอบ 16



ภาพประกอบ 16 แสดง Kernel แบบ Linear

ที่มา: Scikit-learn . (2016). Plot different SVM classifiers in the iris dataset (ออนไลน์). สืบค้นจาก : https://scikit-learn.org/0.18/auto_examples/svm/plot_iris.html [25 พฤษาภาคม 2564]

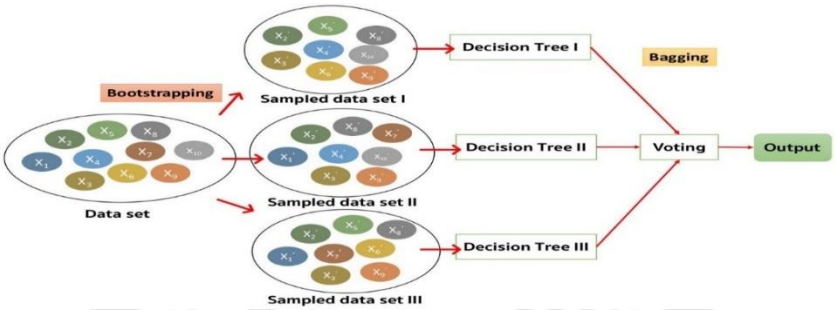
Linear Support Vector Classification (LinearSVC) เป็นวิธีที่งานวิจัยนี้ได้ใช้ในการจำแนกประเภทข่าว โดย LinearSVC มีพารามิเตอร์เป็นแบบ Linear (kernel = 'linear') และอยู่ในรูปแบบซัพพอร์ตเวกเตอร์แมชชีนแบบหลายคลาส (Multi-Class SVM) โดยทั่วไปแล้ว SVM เป็นตัวจำแนกประเภทแบบไบนารี (2 คลาส) แต่ในการประยุกต์ใช้งานจริงนั้น หลายครั้งจำนวนคลาสที่มากกว่า 2 คลาส ซึ่งการทำให้ SVM สามารถนำไปใช้กับปัญหาหลายคลาส (Multi – Class) คือการรวม SVM แบบไบนารีเข้าด้วยกันวิธีการ one-versus-rest โดยวิธีการนี้เป็นการพิจารณาคลาสใดคลาสหนึ่งเทียบกับคลาสอื่น ๆ ที่เหลือทั้งหมด(Suliana Sulaiman, Rohaizah Abdul Wahid, Asma Hanee Ariffin, & Che Zalina Zulkifli, 2020) โดยกำหนดค่าคำตอบ y จากสมการที่ 6

$$y_i = +1, \quad x, \in C_k \text{ และ } -1, \quad x, \notin C_k \tag{6}$$

เมื่อ C_k คือคลาสใดคลาสหนึ่งที่ถูกพิจารณาในแต่ละครั้ง โดย $1 \leq k \leq K$ และถูกพิจารณาทั้งหมดทุกคลาสจำนวน K คลาส

Random Forest เป็นหนึ่งในกลุ่มของโมเดลที่เรียกว่า Ensemble learning โดยการเรียนรู้แบบ Ensemble จะทำงานได้ดีบนเงื่อนไขที่ว่า โมเดลผู้ทำนายแต่ละตัวจะต้องเรียนรู้อย่างเป็นอิสระต่อกันให้มากที่สุด ซึ่ง Random Forest เป็นโมเดลที่ใช้พื้นฐานจากต้นไม้ตัดสินใจ (Decision Tree) โดยเป็นการเทรนโมเดลที่เหมือนกันหลาย ๆ ครั้ง (หลาย Instance) โดยมีหลักการคือ สร้างโมเดลจากต้นไม้ตัดสินใจหลาย ๆ โมเดลย่อย (ตั้งแต่ 10 หรือมากกว่า 1000

โมเดล) โดยแต่ละโมเดลได้รับ Data set ไม่เหมือนกัน ซึ่งเป็น Subset ของData set ทั้งหมด โดยในขั้นตอนการทำนาย (Prediction) จะให้แต่ละต้นไม้ตัดสินใจ ทำการทำนายในส่วนของตัวเอง และคำนวณผลของการทำนายด้วยการโหวด (Vote Output) ที่มาที่สุดมาเป็นผลลัพธ์ โดยจะเรียกว่า Bagging หรือ Bootstrapping (Baoxun Xu, Xiufeng Guo, Yunming Ye, & Jiefeng Cheng, 2012) ดังตัวอย่างภาพประกอบ 17



ภาพประกอบ 17 หลักการทำ Random Forest

ที่มา: Witchapong Daroontham. (2018). เจาะลึก Random Forest Part 2 รู้จัก Decision Tree, Random Forest, และ XGBoost (ออนไลน์). สืบค้นจาก : https://medium.com/@witchapongdaroontham/random-forest-part-2-of -xgboost-79b9f41a1c1c [25 พฤษาภาคม 2564]

3.9 Confusion Matrix

ตาราง Confusion Matrix เป็นตารางที่ใช้ประเมินการวัดประสิทธิภาพของการจำแนกข้อมูล (Classifier) โดยใช้ผลทำนาย (Prediction) เปรียบเทียบกับค่าจริง ๆ (Actual) ซึ่งจะบอกความน่าเชื่อถือของโมเดล รูปแบบของ Confusion Matrix แสดงในภาพประกอบที่ 18

		Predicted Class	
		Positive	**Negative**
Actual Class	**Positive**	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

ภาพประกอบ 18 แสดงเมตริกซ์วัดประสิทธิภาพ (Confusion Matrix)

ที่มา: Vipul Jain. (2018). Idiot's Guide to Precision, Recall, and Confusion Matrix (ออนไลน์). สืบค้นจาก : https://www.kdnuggets.com/2020/01/guide-precision-recall-confusion-matrix.html [27 พฤษาภาคม 2564]

จากภาพประกอบ 18 Confusion Matrix มีหลักการอ่านคือ ลากคอลัมนแนวนอนและแนวดิ่งเข้ามาชนกัน โดยคอลัมน์แนวนอน เป็นผลลัพธ์ (Class) ที่โมเดลทำนายได้ว่าเป็นอะไร ส่วนคอลัมน์แนวดิ่ง เป็น Class, ชื่อ (Label) หรือหมวดหมู่จริง ๆ ของตัวทดสอบ และความหมายในแต่ละช่องสำหรับ Confusion Matrix มีดังนี้

True Positive (TP) คือ จำนวนของข้อมูลทดสอบที่เป็นคลาส Positive และโมเดลจำแนกได้ถูกต้องว่าเป็น Positive

True Negative (TN) คือ จำนวนของข้อมูลทดสอบที่เป็นคลาส Negative และโมเดลจำแนกได้ถูกต้องว่าเป็น Negative

False Positive (FP) คือ จำนวนของข้อมูลทดสอบที่ไม่เป็นคลาส Positive และโมเดลจำแนกผิดว่าเป็น Positive

False Negative (FN) คือ จำนวนของข้อมูลทดสอบที่ไม่เป็นคลาส Negative และโมเดลจำแนกผิดว่าเป็น Negative

โดยถ้าค่า TP และ TN ยิ่งสูงยิ่งดี แสดงว่าจำแนกได้ถูกต้อง ส่วน FP และ FN ค่ายิ่งต่ำยิ่งดีเพราะเป็นความผิดพลาดของการจำแนก และค่า TP, TN, FP และ FN สามารถนำมาคำนวณเพื่อประเมินประสิทธิภาพการจำแนกข้อมูลของโมเดล (Rivera, Florencia, García, Ruiz, & Sánchez-Solís, 2020) โดยทั่วไปจะประกอบไปด้วย Accuracy, Recall, Precision และ F1 score โดยสามารถคำนวณได้ดังต่อไป

Accuracy คือ สัดส่วนเปอร์เซ็นต์ความถูกต้อง หมายถึงจำนวนที่จำแนกถูกต่อจำนวนทั้งหมดหรือประเมินประสิทธิภาพการจำแนกโดยรวมของทุกคลาสในโมเดล จะแสดงดังสมการที่ 7

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

Precision คือ ความแม่นยำผลการจำแนก เป็นการวัดความแม่นยำการทำนายคลาส Positive โดยคำนวณได้จากการหาอัตราส่วนของการทำนายว่าถูกต้องเมื่อเทียบกับข้อมูลที่ถูกต้องจริง (TP) หารด้วยส่วนที่เป็นจริงทั้งหมดของข้อมูลจริง (TP + FP) จะแสดงดังสมการที่ 8

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

Recall คือ การวัดความแม่นยำอีกมิติ ที่สนใจผลลัพธ์เทียบกับที่เป็นของจริง โดยเป็นการวัดความสามารถในการค้นหาข้อมูลที่เป็นคลาส Positive สามารถคำนวณได้จากการหาอัตราส่วนของการทำนายว่าถูกต้องเมื่อเทียบกับข้อมูลที่ถูกต้องจริง (TP) หารด้วยค่าที่ทำนายว่าถูกต้องทั้งหมด จะแสดงในสมการที่ 9

$$Recall = \frac{TN}{TN+FP} \quad (9)$$

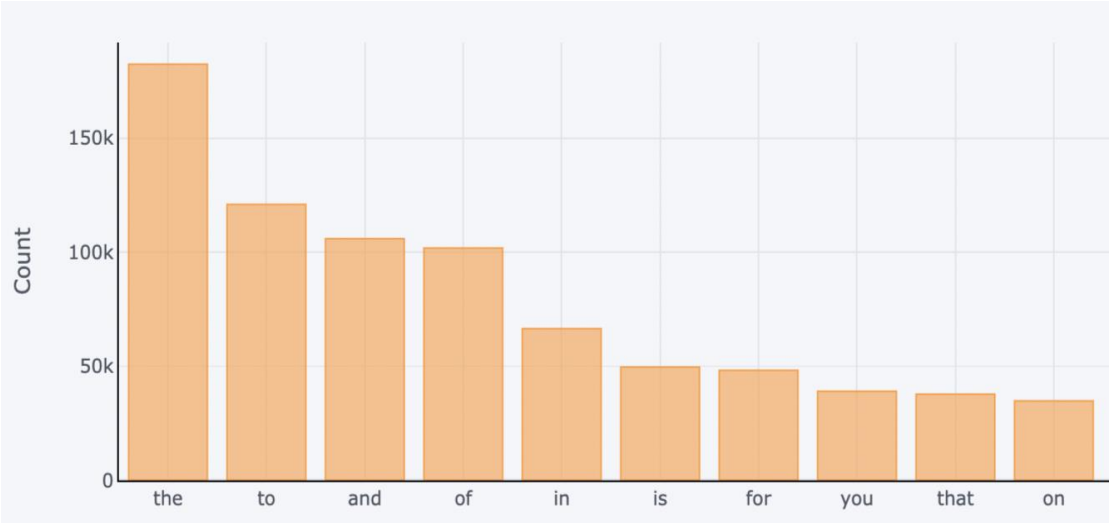
F1 score คือ ค่าที่แสดงประสิทธิภาพ โดยการนำ Precision กับ Recall มา3ค่าเฉลี่ย ถ้ามีค่าสูง ๆ แสดงว่าโมเดลมีประสิทธิภาพดี จะแสดงในสมการที่ 10

$$F1\ score = \frac{2(Precision)(Recall)}{Precision+Recall} \quad (10)$$

บทที่ 4
การทดลอง

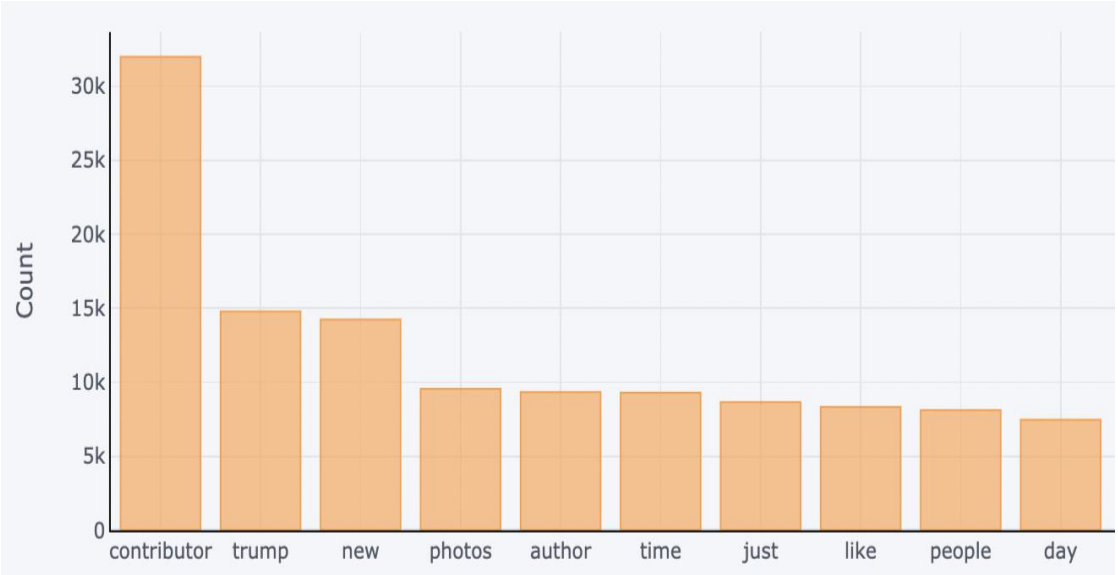
4.1 ผลการสำรวข้อมูล

ทำการสร้างกราฟแท่งเพื่อดูว่าบทความข่าวทั้ง 15 ประเภทข่าวมีการใช้คำไหนมากที่สุดในบทความข่าว แสดงในภาพประกอบ 19



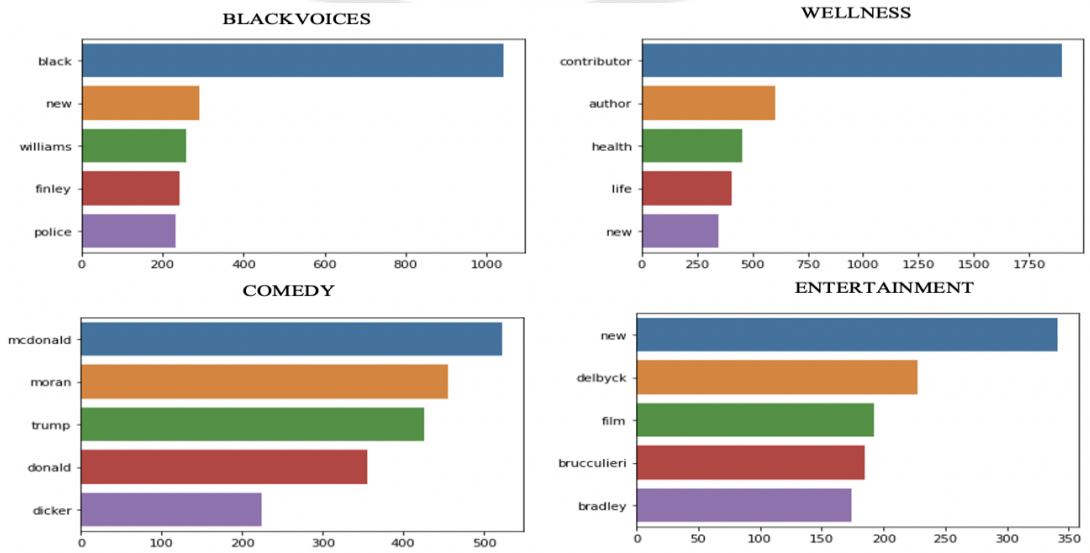
ภาพประกอบ 19 แสดง คำที่ถูกใช้มากที่สุดในบทความข่าวทั้งหมดก่อนลบ stop word

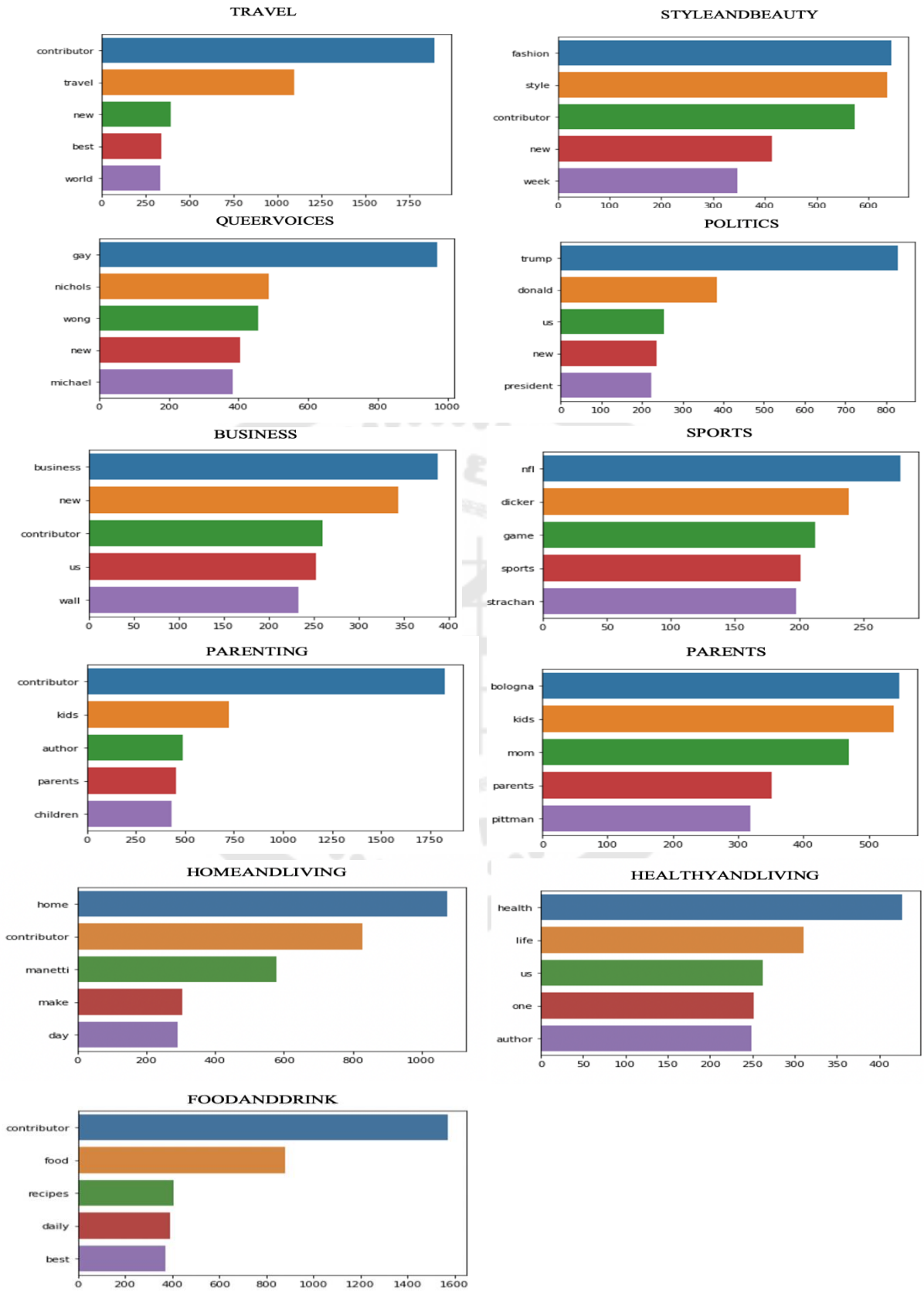
จากภาพประกอบ 19 จะเห็นได้ว่าในคำในบทความข่าวที่ถูกใช้มากที่สุดจะเป็นคำว่า the, to, and, of เป็นต้น โดย the มีมากถึง 56317 คำ จากบทความทั้งหมด ซึ่งคำเหล่านี้เรียกว่า stop word ดังนั้นจากกราฟเห็นได้ว่าควรลบคำเหล่านั้นออกไปจากรายการคำศัพท์ได้เลย ถึงแม้คำเหล่าถูกนำมาใช้มากจริงในบทความแต่ไม่ค่อยช่วยในการสื่อความหมายสักเท่าไรและอาจทำให้การแยกประเภทของโมเดลผิดพลาดได้ ดังนั้นเมื่อทำการลบ stop word ออกแล้วมาสร้างกราฟใหม่อีกครั้ง ดังภาพประกอบ 20



ภาพประกอบ 20 แสดง คำที่ถูกใช้มากที่สุดในบทความข่าวทั้งหมดหลังลบ stop word

จากภาพประกอบ 20 แสดงให้เห็นจำนวนคำที่ถูกใช้มากที่สุดหลังการลบ stop word โดยคำที่มีมากที่สุดในบทความข่าวทั้งหมดคือคำว่า contributor มีอยู่ 9769 คำ จากบทความข่าวทั้งหมด 45000 บทความ และคำอื่น ๆ รองลงมาได้ new, one, us เป็นต้น และจะใช้วิธี N-gram ด้วยเทคนิค Uni-gram และ Bi-gram ในการสร้างแผนภูมิแสดงความถี่ของคำที่ใช้มากที่สุดในแต่ละประเภทข่าวทั้ง 15 ประเภท โดยเริ่มจากจะแสดงแผนภูมิ Uni-gram ที่แสดงของคำเดียว ดังภาพประกอบ 21





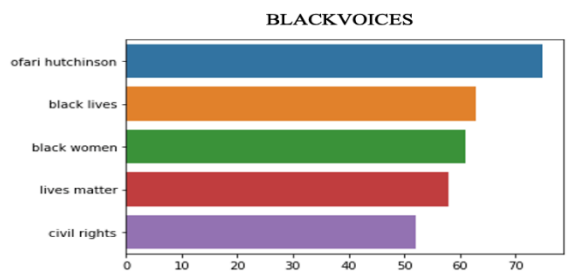
ภาพประกอบ 21 แสดง ความคำที่มีการใช้ถี่มากที่สุดในแต่ละประเภทข่าว

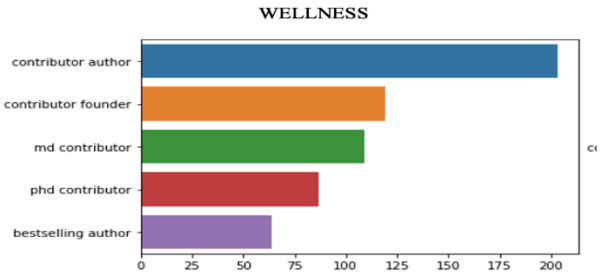
จากภาพประกอบ 21 จะพบมีคำที่ใช้มากสุดในแต่ละประเภทข่าวซึ่งเป็นคำเดียวๆ โดยคำที่ได้ในแต่ละประเภทข่าวจะมีคำซ้ำกันโดยอาจจะไม่ได้อยู่ในประเภทข่าวเดียวกัน โดยจะแสดงตารางภาพประกอบ 22 เพื่อง่ายต่อความเข้าใจ

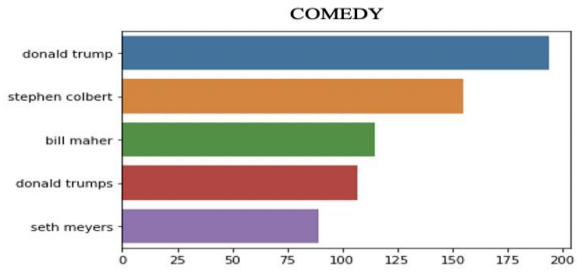
Category	Uni-gram Top word	Frequencies
WELLNESS	contributor	1899
TRAVEL	contributor	1895
PARENTING	contributor	1828
FOOD & DRINK	contributor	1573
HOME & LIVING	home	1076
BLACK VOICES	black	1040
QUEER VOICES	gay	969
POLITICS	trump	830
STYLE & BEAUTY	fashion	645
PARENTS	bologna	547
COMEDY	mcdonald	523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

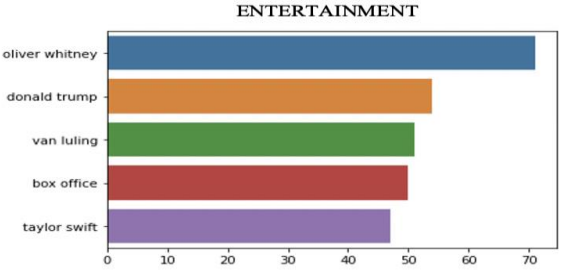
ภาพประกอบ 22 แสดงคำและจำนวนที่ใช้ถี่มากที่สุดอันดับแรกของแต่ละประเภทข่าว

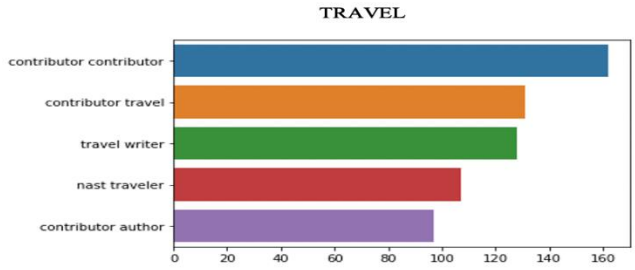
จากภาพประกอบ 22 ตาราง แสดงให้เห็นการนำกลุ่มคำที่ถูกใช้มากที่สุดในแต่ละประเภทข่าวมาเรียงจำนวนการใช้คำจากมากสุดไปน้อยสุด จะเห็นได้ชัดเจนว่ามีประเภทข่าวที่มีการใช้คำเหมือนกันและจำนวนการความถี่ในการใช้คำใกล้กัน ได้แก่ ข่าวประเภท WELLNESS, TRAVEL, PARENTING และ FOOD&DRINK โดยคำที่ใช้คือคำว่า contributor โดยมีความถี่ในการใช้คำอยู่ที่ 1899, 1895, 1828 และ 1573 ตามลำดับ โดยจะเห็นคำ ๆ หนึ่งอาจสามารถถูกตีความให้อยู่ได้หลายประเภทข่าวอาจทำให้เห็นความไม่ชัดเจนของคำในการแยกประเภทข่าว ดังนั้นจึงทำการสร้างแผ่นภูมิ Bi-gram ที่เกิดจากนำคำ 2 คำมารวมกันเป็น 1 รูป ประโยค แสดงดังภาพประกอบ 23 เพื่อให้เห็นความชัดเจนของคำมากขึ้น

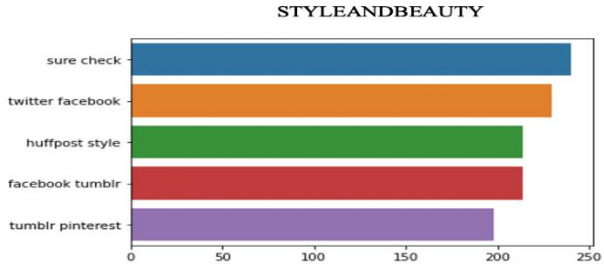

BLACKVOICES
ofari hutchinson
black lives
black women
lives matter
civil rights
0 10 20 30 40 50 60 70

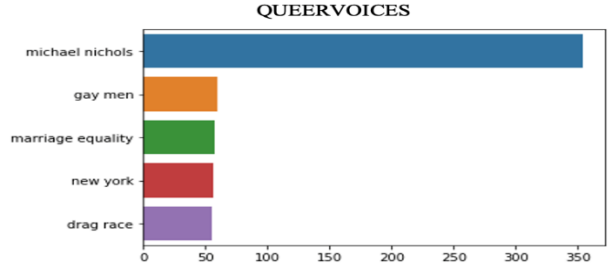

WELLNESS
contributor author
contributor founder
md contributor
phd contributor
bestselling author
0 25 50 75 100 125 150 175 200

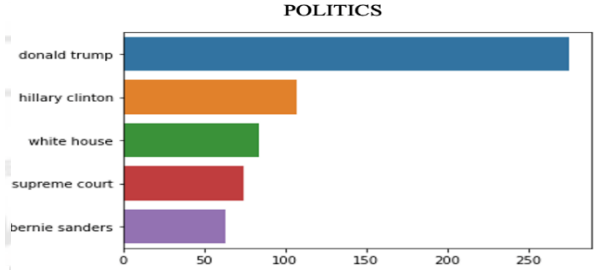

COMEDY
donald trump
stephen colbert
bill maher
donald trumps
seth meyers
0 25 50 75 100 125 150 175 200

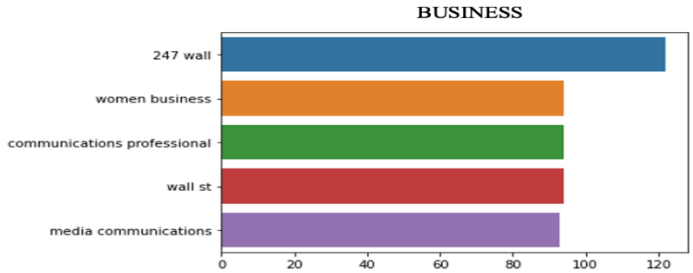

ENTERTAINMENT
oliver whitney
donald trump
van luling
box office
taylor swift
0 10 20 30 40 50 60 70

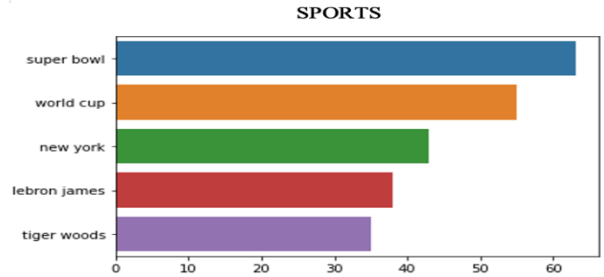

TRAVEL
contributor contributor
contributor travel
travel writer
nast traveler
contributor author
0 20 40 60 80 100 120 140 160

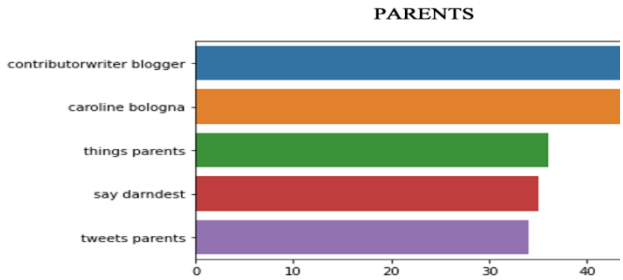

STYLEANDBEAUTY
sure check
twitter facebook
huffpost style
facebook tumblr
tumblr pinterest
0 50 100 150 200 250

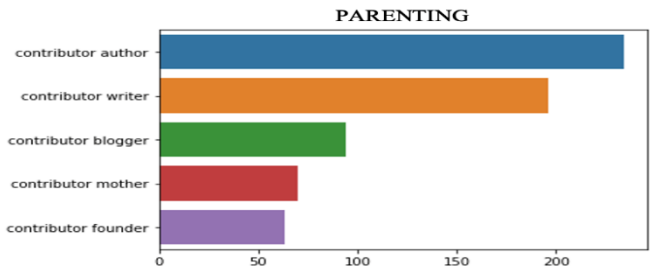

QUEERVOICES
michael nichols
gay men
marriage equality
new york
drag race
0 50 100 150 200 250 300 350

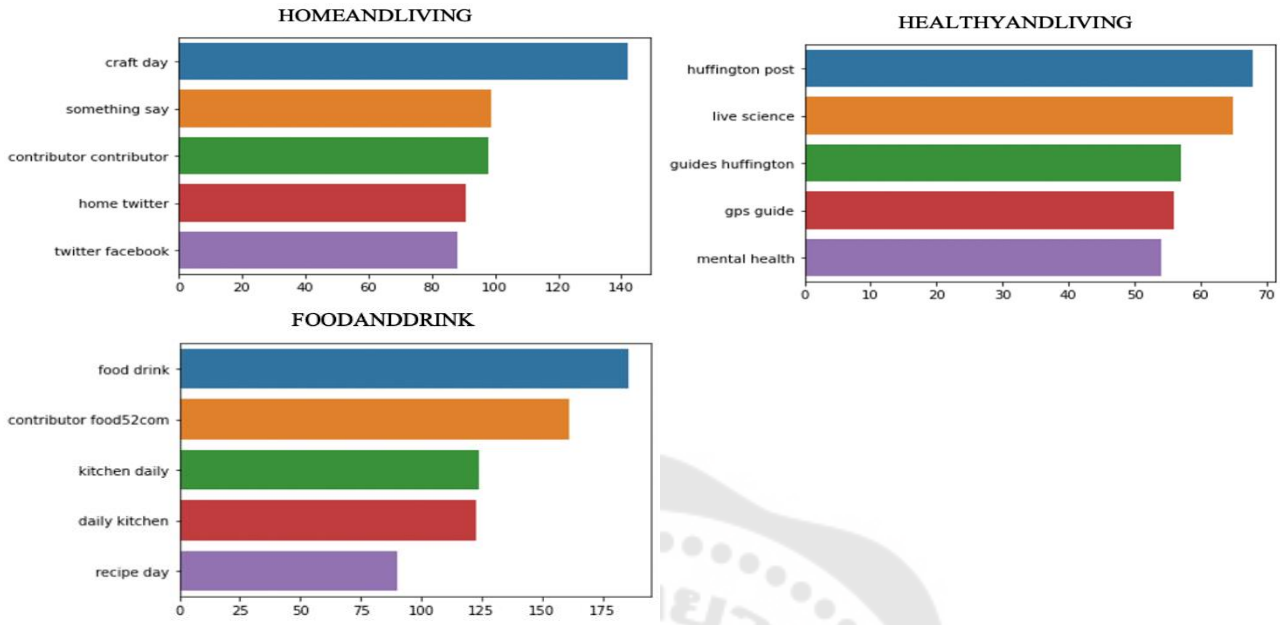

POLITICS
donald trump
hillary clinton
white house
supreme court
bernie sanders
0 50 100 150 200 250


BUSINESS
247 wall
women business
communications professional
wall st
media communications
0 20 40 60 80 100 120


SPORTS
super bowl
world cup
new york
lebron james
tiger woods
0 10 20 30 40 50 60


PARENTS
contributorwriter blogger
caroline bologna
things parents
say darndest
tweets parents
0 10 20 30 40


PARENTING
contributor author
contributor writer
contributor blogger
contributor mother
contributor founder
0 50 100 150 200



ภาพประกอบ 23 แสดง ความคำที่มีการใช้ถี่มากที่สุดในแต่ละประเภทข่าวด้วยเทคนิค Bi-gram

จากภาพประกอบ 23 เห็นได้ว่าเมื่อมีการรวมคำกันเกิดขึ้น จะทำให้ได้ความหมายของประโยคหนึ่งซึ่งทำให้เห็นถึงความเกี่ยวเนื่องของประโยคที่มีต่อประเภทข่าวมากขึ้น แต่จะสังเกตได้ว่าจำนวนความถี่ในการพบประโยคคำจะลดลงอย่างมาก โดยจะแสดงในภาพประกอบตาราง 24

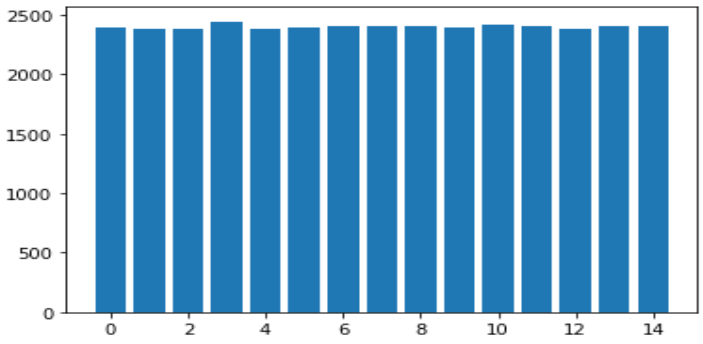
Category	Bi-gram Top word	Frequencies
QUEER VOICES	michael nichols	354
POLITICS	donald trump	275
STYLE & BEAUTY	sure check	240
PARENTING	contributor author	234
WELLNESS	contributor author	203
COMEDY	donald trump	194
FOOD & DRINK	food drink	186
TRAVEL	contributor contributor	162
HOME & LIVING	craft day	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

ภาพประกอบ 24 แสดงประโยคคำและจำนวนที่ใช้ถี่มากที่สุดอันดับแรกของแต่ละประเภทข่าว

จากภาพประกอบ 24 ตารางแสดงให้เห็นว่าการซ้ำกันของคำที่อยู่ในแต่ละประเภทข่าวลดลงเหมือนรูปประโยคคำที่ชัดเจนขึ้น 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

4.2 ผลการทำ Sampling algorithm

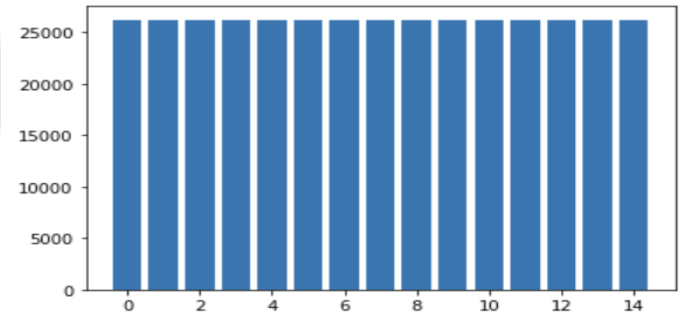
ทำการเลือกโมเดลสามโมเดลที่ดีที่สุดจากการเปรียบเทียบค่าประสิทธิภาพของโมเดลในกระบวนการ Feature Extraction ระหว่าง TFID และ BOW แล้วนำมาเปรียบเทียบค่าประสิทธิภาพจากกระบวน Sampling Algorithm Method ซึ่งทำกับ Trian set ที่ 80% โดยในเทคนิคแรก คือ Undersampling ผู้วิจัยได้ตั้งค่าคำสั่งฟังก์ชันไว้ คือ n_sample =3000 เป็นการกำหนดบทความข่าวที่ต้องการจำนวน 3000 ข่าวต่อ 1 ประเภทข่าว และ random_state = 123 เป็นการกำหนดการสุ่มแต่ละครั้งให้ได้ผลออกมาเหมือนกัน ซึ่งเป็นประโยชน์ในการทดสอบโมเดล โดย 123 เป็นการตั้งเลขจำนวนเต็มซึ่งใส่ตัวเลขจำนวนเต็มอะไรก็ได้ลงไป โดยเมื่อทำการใช้คำสั่ง resample() ในข่าวแต่ละประเภทที่ได้ผู้วิจัยได้ทำการเลือกแล้ว ดังแสดงในภาพประกอบ 25



ภาพประกอบ 25 กราฟแสดงจำนวนบทความข่าวของ Train หลัง Undersampling

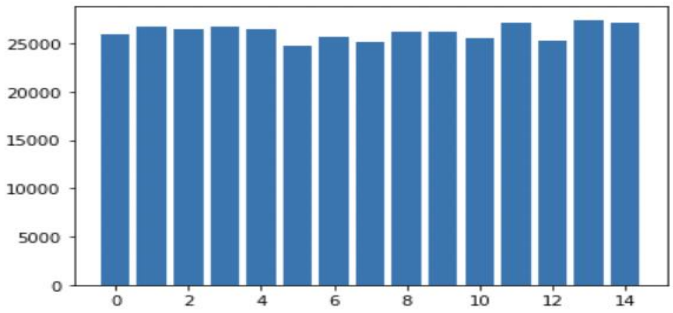
จากภาพประกอบ 25 จะทำให้เห็นประเภทข่าวทั้ง 15 ประเภท ที่มีจำนวนบทความข่าว 3000 ข่าวเท่ากัน ทำให้ชุดข้อมูลนี้มีความสมดุลมากขึ้น เพราะเมื่อชุดข้อมูลมีความสมดุลกันจะทำให้ประสิทธิภาพของโมเดลที่ใช้ในการแยกประเภทข่าวมีประสิทธิภาพมากขึ้น โดยมีบทความข่าวรวมกัน 450000 บทความ การที่มีบทความเท่ากันในแต่ละประเภทจะทำให้ง่ายต่อการเข้าโมเดลเพื่อทำการจำแนกประเภท และทำการใช้เทคนิค SMOTE โดยทำหลังจากมีการแบ่ง Test และ Train แล้ว

ต่อไปทำเทคนิคต่อไปเรียกว่าการ SMOTE ที่ตัว Train ซึ่งในงานวิจัยนี้ได้แบ่งตัว Train ไว้ที่ 80% โดยมีจำนวนบทความแต่ละคลาสหลังการแบ่งที่มากสุดจำนวน 26191 บทความและน้อยสุดที่ 3164 บทความ เมื่อทำการ SMOTE ก็จะทำการสังเคราะห์ข้อมูลเพิ่ม โดยทุกคลาสจะถูกเพิ่มจำนวนบทความให้เท่ากับคลาสที่มีจำนวนบทความมากสุด แสดงดังภาพประกอบที่ 26



ภาพประกอบ 26 กราฟแสดงจำนวนบทความข่าวของ Train หลัง SMOTH

และเทคนิคสุดท้ายคือ เทคนิค ADASYN ได้แบ่งตัว Train ไว้ที่ 80% โดยมีจำนวนบทความแต่ละคลาสหลังการแบ่งที่มากสุดจำนวน 27090 บทความและน้อยสุดที่ 24781 บทความ แสดงกราฟดังภาพประกอบที่ 27



ภาพประกอบ 27 กราฟแสดงจำนวนบทความข่าวของ Train หลัง ADASYN

ดังนั้นจึงทำภาพประกอบแสดงตารางจำนวนบทความข่าวของประเภทข่าวและบทความรวมทั้งหมดหลังจากการใช้ Sampling Algorithm ในการแก้ปัญหาข้อมูลที่ไม่สมดุล (imbalance) ด้วยเทคนิค Undersampling, SMOTE และ ADASYN โดยแสดงในภาพประกอบ 28

Sampling Algorithm	None	Undersampling	SMOTE	ADASYN
จำนวนบทความแต่ละ class	26259 - 3,230	2,444- 2,376	26,191	27,090-24,781
จำนวนบทความทั้งหมด (Train set)	114,196	36,000	392,865	392,772

ภาพประกอบ 28 กราฟแสดงจำนวนบทความข่าวหลังใช้ Sampling Algorithm

4.3 ผลการทดลองจากการทดสอบประสิทธิภาพ

4.3.1 ผลการทดลอง

ในบทความวิจัยเรื่องนี้ผู้วิจัยได้วัดประสิทธิภาพของโมเดลในการจำแนกประเภทของข่าวด้วยวิธีการแต่ละเทคนิคเพื่อ นำมาเปรียบเทียบและสรุปผลงานวิจัย โดยสามารถแบ่งหัวข้อในการสรุปผลดังต่อไปนี้

จากศึกษาวิจัยการจำแนกประเภทข่าวด้วยวิธีการเรียนรู้ด้วยเครื่องโดยการเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้ในการจำแนกประเภทข่าว ด้วยการทำ Feature Extraction ระหว่าง TFIDF และ BOW โดยโมเดลที่ทำการเปรียบเทียบได้แก่ Multinomial Naïve Bayes, Complement Naïve Bayes, Logistic Regression, LinearSVC และ Random Forest โดยแบ่งข้อมูลในสัดส่วน Test 20% ของข้อมูลทั้งหมด โดยได้ตั้งค่าพารามิเตอร์เป็น default ยกเว้น โมเดล Complement Naïve Bayes ที่ตั้ง alpha เท่ากับ 1 ซึ่งผลจากการวิจัยแสดงอยู่ในรูปแบบ Classification Report ที่จะแสดงค่าต่าง ๆ ได้แก่ Accuracy, Precision, Recall และ ค่าแสดงประสิทธิภาพ (F1 score) โดยผลของการทดลองจะแสดงดังภาพประกอบ 29

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

ภาพประกอบ 29 ผลลัพธ์จากการทดสอบของโมเดลระหว่าง TFIDF และ BOW

จากภาพประกอบ 29 ผลการเปรียบเทียบระหว่าง TFIDF และ BOW เห็นได้ว่า โมเดล 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 Naïve Baye 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 3เทคนิค Sampling Algorithm และ

นำมาแยกประเภทข่าวด้วย 3 โมเดลที่เลือกไว้ที่ใช้เทคนิค BOW โดยผลของการทดลองจะมาจาก Test set 20 % ซึ่งจะแสดงดังภาพประกอบ 30

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

ภาพประกอบ 30 ผลลัพธ์จากการทดสอบของโมเดลระหว่างด้วยเทคนิค Sampling Algorithm

จากผลลัพธ์ภาพประกอบ 30 จะเห็นได้ว่า Sampling Algorithm ที่ใช้เทคนิค SMOTE และใช้โมเดล Logistic Regression มีค่าประสิทธิภาพในการจำแนกประเภทข่าวสูงที่สุด เมื่อดูจากค่าเปอร์เซ็นต์ของ Accuracy, Recall, Precision และ F1 score ซึ่งจะมีค่าอยู่ที่ 80.69, 77.63, 77.04 และ 77.31 ตามลำดับ

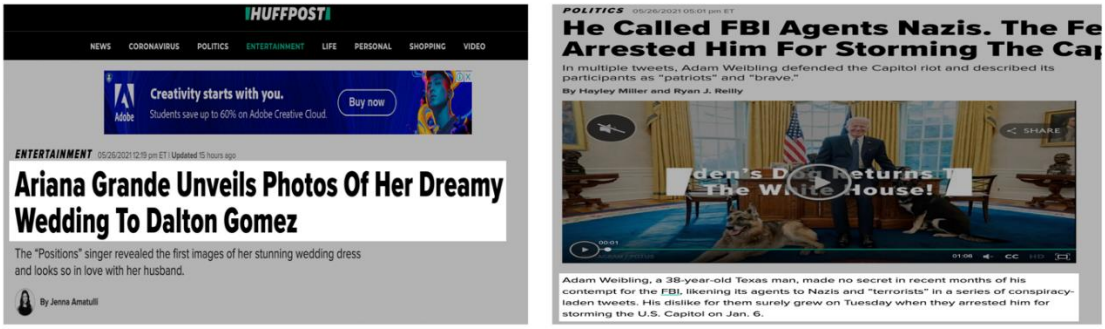
จึงสรุปผลลัพธ์งานวิจัยได้ว่าโมเดล Logistic Regression ที่ใช้มีการใช้ Feature Extraction ด้วย Bag-of-Word (BOW) และทำการแก้ข้อมูลที่ไม่สมดุลในแต่ประเภทข่าวด้วยเทคนิค SMOTE สามารถเพิ่มประสิทธิภาพในการจำแนกประเภทข่าวของโมเดลได้ ดังนั้นจึงทำการเลือกใช้เทคนิค BOW และเลือกโมเดล Logistic Regression ที่ทำการ SMOTH ของข้อมูลแล้ว ซึ่งพบว่ามีความแม่นยำในการจำแนกประเภทข่าวดีที่สุด ในการวิจัยมาทำการทดสอบในการจำแนกประเภทข่าวโดยการนำประโยคข้อความข่าวและพาดหัวข่าวจากเว็บไซต์ HuffPost มาเป็นตัวอย่างในการให้โมเดลทดสอบการจำแนกประเภทข่าว โดยนำประโยคข้อความข่าวมาเพื่อให้โมเดลทำการจำแนกประเภทข่าว ว่าได้เป็นประเภทข่าวอะไร โดยผลลัพธ์จะแสดงในภาพประกอบ 31

```
headline = "Ariana Grande Unveils Photos Of Her Dreamy Wedding To Dalton Gomez"
news = "Adam Weibling, a 38-year-old Texas man, made no secret in recent months of his contempt
print(logistic_model.predict([headline, news]))
```

```
['ENTERTAINMENT' 'POLITICS']
```

ภาพประกอบ 31 จำแนกประเภทของข่าวจากโมเดลที่ดีที่สุดในการวิจัย

โดยผลลัพธ์ได้นำพาดหัวข่าวและเนื้อข่าวจากเว็บไซต์ข่าว HuffPost ที่มีการจัดประเภทข่าวไว้แล้ว โดยเลือกพาดหัวข่าวที่อยู่ในประเภท Entertainment และเนื้อข่าวที่อยู่ในประเภท Politics ของเว็บไซต์ มาให้โมเดลได้จำแนกแยกประเภทข่าว ซึ่งผลลัพธ์โมเดลสามารถจำแนกได้เป็น Entertainment และ Politics เช่นเดียวกับ ที่เว็บไซต์ได้จัดประเภทไว้ ดังตัวอย่างภาพประกอบ 32



ภาพประกอบ 32 พาดหัวข่าวและเนื้อข่าวจากเว็บไซต์ HuffPost ที่ใช้ในการทดสอบโมเดล

ดังนั้นเมื่อทำการเปรียบเทียบกับผลการวิจัยที่ได้ทำกับของบทความวิจัย(Melek & Mokthtar, 2020) จะแสดงได้ว่าจากบทความวิจัยมีการจำแนกประเภทข่าวด้วยวิธีการใช้ Deep Learning ด้วยโมเดล LSTM ในการจำแนกประเภทของข่าวโดยได้เปอร์เซ็นต์อยู่ที่ 72 แต่ในขณะที่งานวิจัยของผู้วิจัยได้เลือกใช้ Machine Learning ด้วยโมเดล Logistic Regression ที่ใช้เทคนิค BOW และ การปรับสมดุลด้วย SMOTH ซึ่งได้เปอรเซ็นต์อยู่ที่ 81 แสดงให้เห็นว่าการใช้วิธีนี้สามารถทดแทนการทำ Deep Learning ได้ในการจำแนกประเภทข่าว

บทที่ 5
สรุปผลการทดลอง

5.1 อภิปรายผลการทดลอง

งานวิจัยการจำแนกประเภทข่าวด้วยวีการเรียนรู้ด้วยเครื่องที่ได้นำชุดข้อมูลจากเว็บไซต์ข่าว HuffPost ที่มีข้อมูลหลังการเตรียมข้อมูลอยู่ 28549 ข่าว 15 ประเภท โดยข้อมูลข่าวทั้งหมดเป็นภาษาอังกฤษจะเห็นผลลัพธ์จากงานวิจัยได้ว่า Logistic Regression ที่ใช้ BOW และ SMOTH มีค่าประสิทธิภาพสูงที่สุดของของโมเดลทั้งหมด ซึ่งจะสามารถอภิปรายได้ว่า ทำไมถึงเป็นวิธีที่มีประสิทธิภาพดีที่สุดในงานวิจัย โดยจะทำการพิจารณาได้จากดังต่อไปนี้

ผลลัพธ์ค่า Accuracy, Recall, Precision และ F1 score ของโมเดล Logistic Regression ของประเภทข่าวทั้ง 15 ประเภท ดังแสดงในภาพประกอบ 33

	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

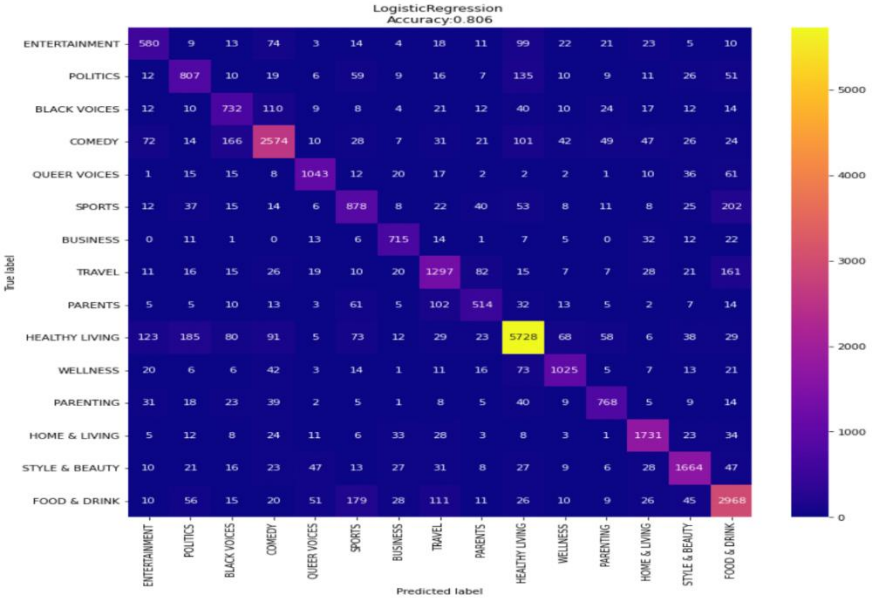
ภาพประกอบ 33 ผลลัพธ์ของโมเดล Logistic Regression ของแต่ละประเภทข่าว

จากภาพประกอบ 33 จะเห็นค่าผลลัพธ์ต่าง ๆของข่าวแต่ละประเภทในโมเดล Logistic Regression โดยเริ่มจากค่า Support ซึ่งคือจำนวนตัวทดสอบหรือจำนวนข้อมูล (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 ได้ในภาพประกอบ 34



ภาพประกอบ 34 Confusion Matrix ของ Logistic Regression, BOW และ SMOTH ของแต่ละประเภทข่าว

โดยจากภาพประกอบ 35 จะเห็นได้ว่า Logistic Regression 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 หนึ่งสามารถอยู่ได้ในหลายประเภทข่าว

5.2 ข้อเสนอแนะ

ค่า Accuracy, Recall, Precision และ F1 score ของโมเดล Logistic Regression ที่เลือกใช้เทคนิค BOW และทำการ SMOTH ถึงเป็นโมเดลที่ดีที่สุดแต่ยังมีค่าไม่สูงมากเท่าที่ควร อาจมาจากการเลือกประเภทข่าวที่น้อยไปหรือจำนวนข่าวที่น้อยทำให้มีข้อมูลในการใช้เป็นตัวเปรียบเทียบในการจำแนกประเภทข่าวไม่เพียงพอ หรืออาจมาจากปัจจัยอื่น ๆ อีก แต่การได้ทดลองนำพาดข่าวมาให้โมเดล Logistic Regression ได้ลองจำแนกประเภทข่าวแล้วพบว่ามีความถูกต้องในการจำแนกประเภท แสดงให้เห็นว่าถึงแม้มีค่าประสิทธิภาพที่ไม่สูงแต่ก็สามารถจำแนกประเภทได้ถูกต้อง และการใช้เทคนิค Bag-of-Word ทำให้เกิดการสร้างตารางความถี่คำมากซึ่งหากมีข้อมูลเป็นจำนวนมากจะทำให้เวลาในการรันโมเดลใช้เวลามาก

ในอนาคตการจำแนกประเภทข่าวให้ได้แม่นยำขึ้น จำเป็นต้องมีการรวมโมเดลเพื่อเพิ่มประสิทธิภาพในการจำแนกข่าวหรือการนำ Deep learning มาช่วย ส่วนในขั้นตอนการสกัดคำ Text Mining ในการช่วยสกัดคำ หรือลองใช้เทคนิค TFIDF ในการ Feature Vector เพื่อทำให้โมเดลมีค่าประสิทธิภาพที่สูงขึ้น

บรรณานุกรม

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.

Baoxun Xu, Xiufeng Guo, Yunming Ye, & Jiefeng Cheng. (2012). An Improved Random Forest Classifier for Text Categorization. *Journal of Computers, 7*(12), 2913-2920.

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

George K, S., & Joseph, S. (2014). Text Classification by Augmenting Bag of Words (BOW) Representation with Co-occurrence Feature. *IOSR Journal of Computer Engineering, 16*(1), 34-38.

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.

Gurmeet Kaur, & Karan Bajaj. (2016, Jan – Feb). News Classification and Its Techniques: A Review. *IOSR Journal of Computer Engineering (IOSR-JCE), 18*(1), 22-26.

J SreeDevi, M Rama Bai, & M Chandrashekar Reddy. (2020). Newspaper Article Classification using Machine Learning Techniques. *International Journal of Innovative Technology and Exploring Engineering, 9*(5), 872-877.

J. E. Sembodo, E. B. Setiawan, & M. A. Bijaksana. (2018). Automatic Tweet Classification Based on News Category in Indonesian Language. *International Conference on Information and Communication Technology (ICoICT), 6*, 389-393.

Melek, M. A., & Mokhtar, B. (2020). *Towards Intelligent Web Context-Based Content On-Demand Extraction Using Deep Learning*. Paper presented at the 2020 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT).

Muhammad Abbas, Kamran Ali Memon, Abdul Aleem Jamali, Saleemullah Memon, & Anees Ahmed. (2019). Multinomial Naive Bayes Classification Model for Sentiment Analysis. *IJCSNS International Journal of Computer Science and Network*

Security, 19(3), 62-67.

Ristu Saptono, & Winarno Win. (2018). Comparison of Under-Sampling and Over-Sampling Techniques in Diabetic Mellitus (DM) Patient Data Classification by Using Naive Bayes Classifier (NBC). *Journal of Telecommunication, 10*, 2-4.

Rivera, G., Florencia, R., García, V., Ruiz, A., & Sánchez-Solís, J. P. (2020). News Classification for Identifying Traffic Incident Points in a Spanish-Speaking Country: A Real-World Case Study of Class Imbalance Learning. *Applied Sciences, 10*(18).

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.

Siva RamaKrishna Reddy V, D V L N. Somayajulu, & Ajay R. Dani. (2010). Classification of Movie Reviews Using Complemented Naive Bayesian Classifier. *nternational Journal of Intelligent Computing Research (IJICR), 1*(4), 162-167.

Suliana Sulaiman, Rohaizah Abdul Wahid, Asma Hanee Ariffin, & Che Zalina Zulkifli. (2020, March – April). Question Classification Based on Cognitive Levels using Linear SVC. *Test Engineering and Management*, 6463 - 6470.

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

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

กาญจน์ณ ศรีธะ, กิตติศักดิ์ เกิดประสพ, & นิตยา เกิดประสพ. (2561). การเปรียบเทียบเทคนิคการสุ่มตัวอย่างเพื่อการจำแนกข้อมูลที่ไม่สมดุล. วารสารวิทยาการสารสนเทศและเทคโนโลยีประยุกต์, 21-35.

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

วาทินี นุ้ยเพียรม, & พยุง มีสัจ. (2556, กันยายน – ธันวาคม). การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวม ของการทำเหมืองข้อความเพื่อการจำแนกข้อความ. วารสารวิชาการเทคโนโลยีอุตสาหกรรม, *9*(3), 118-129.

สุขประเสริฐ, น. (2560). องค์ประกอบของข่าว. สืบค้นจาก https://sites.google.com/site/pesavena/xngkh-prakxb-khxng-khaw

สุรวัชร ศรีเปารยะ, & สายชล สินสมบูรณ์ทอง. (2560, กันยายน - ตุลาคม). การเปรียบเทียบประสิทธิภาพวิธีการจาแนกกลุ่มการเป็นโรคไตเรื้อรัง : กรณีศึกษาโรงพยาบาลแห่งหนึ่งในประเทศอินเดีย. วารสารวิทยาศาสตร์และเทคโนโลยี, *25*(5), 840-853.

ประวัติผู้เขียน

ชื่อ-สกุล	นายกิตติศักดิ์ กิตติธานุสรณ์
วัน เดือน ปี เกิด	20 กรกฎาคม 2536
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	พ.ศ. 2554 วิทยาศาสตรบัณฑิต สาขาเคมีประยุกต์ จาก มหาวิทยาลัยรังสิต พ.ศ.2562 วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล จาก มหาวิทยาลัยศรีนครินทรวิโรฒ