การวิเคราะห์ความรู้สึกของข้อความ กรณีศึกษาชุดข้อมูลของทวิตเตอร์ เรื่อง การศึกษาไทย

Thai sentiment analysis about education in Thailand on Twitter.

วิภาดา ศิลาราช (Wipada Silarach) พิทยารัตน์ พ่อชมภู (Phithayarat Phochompu)
และธนพล ตั้งชูพงศ์ (Thanaphon Tangchoopong)
สาขาวิชาวิทยาการคอมพิวเตอร์ วิทยาลัยการคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น
wipadasi@kkumail.com, phithayarat.mwph@kkumail.com, thanaphon@kku.ac.th

บทคัดย่อ

เหมืองความคิดเห็นบนโลกออนไลน์ มีความจำเป็นมากในปัจจุบัน สามารถนำไปพัฒนาธุรกิจ หรือพิจารณาอารมณ์ของข้อความได้ งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างโมเดลในจำแนกข้อความภาษาไทยที่แสดงความคิดเห็นในแง่มุมของการศึกษาไทยบนทวิตเตอร์ออกมาเป็น 3 ขั้วความคิดเห็นคือ ทางบวก ทางลบ และกลาง โดยนำชุดข้อมูลที่ได้จากการเลเบลไปคัดเลือกคำที่เป็นฟีเจอร์ และพบปัญหาในเรื่องข้อมูลที่ไม่สมดุล จึงนำเสนอการจัดการข้อมูลที่ไม่สมดุลด้วยการปรับข้อมูลที่ไม่สมดุล 3 วิธีคือ การสุ่มตัวอย่างลด การสุ่มตัวอย่างเพิ่ม และ การสังเคราะห์ข้อมูลเพิ่ม (SMOTE) แล้วฝึกโมเดลสำหรับจำแนกอารมณ์ของข้อความด้วยโมเดล Logistic regression, XGBoost, Decision tree, SVM, Random forest, K-NN โดยเลือกไฮเปอร์พารามิเตอร์ (Hyperparameter) ที่ดีที่สุดของแต่ละโมเดลจากการทำการค้นหาแบบกริด (Grid search) และเปรียบเทียบประสิทธิภาพของโมเดลก่อนและหลังการปรับข้อมูลที่ไม่สมดุลด้วยเมทริกซ์ความสับสน (Confusion matrix) พบว่าก่อนปรับข้อมูลค่าที่ได้ส่วนใหญ่ลำเอียงไปที่คลาสลบ และค่าความถูกต้องในแต่ละโมเดลมีค่าต่ำ เฉลี่ยที่ 41% และการทำนายในคลาสบวกมีผลที่น้อยจากค่าเฉลี่ย F1 score 17% หลังจากปรับข้อมูลทำให้ค่าเฉลี่ยความถูกต้องเพิ่มขึ้นทั้ง 3 วิธี โดยวิธีการสุ่มตัวอย่างลดมากที่สุดเฉลี่ยที่ 51% ซึ่งมีโมเดลที่ดีสุดคือ Random forest ที่ 55% อีกทั้งค่าเฉลี่ย F1 score ในคลาสบวกสูงสุดที่ 58% และจาก 3 วิธีโมเดลที่ดีที่สุดคือ Logistic regression มีค่าเฉลี่ยความถูกต้องที่ 53%

คำสำคัญ: การวิเคราะห์ความรู้สึก, เหมืองข้อความ, การประมวลผลภาษาธรรมชาติ

Abstract

Nowadays, text mining from social networks has become very important in order to thrive in business and sentiment monitoring. Therefore, this research aims to create a model that classifies the sentiments of Thai text comments on Twitter into three categories: negative, positive, and neutral. The collected data is labeled and selected as a vector feature. The major results are analyzed and indicate the negativity of the data. Thus, the research presents imbalanced data management and adjustment using three methods: random under sampling, random oversampling, and SMOTE. Afterward, the transformed training data is used to train models for analyzing the sentiment of data with the classification

algorithms: Logistic regression, XGBoost, Decision tree, SVM, Random Forest, K-NN. The hyperparameters of each algorithm are well selected by the grid search method and evaluate the model's performance by a confusion matrix of result pre and post imbalanced data adaptation. Before the adaptation, the data have a tendency to negative classification. The model analysis has low accuracy, 41 percent on average, and the prediction outcome of positive classification has 17 percent from the calculation of the F1 score. After the adaptation, the average accuracy has been increased in all three methods. The most effective method is random under sampling which has 51 percent accuracy, with the Random Forest model at 55 percent accuracy. Moreover, the average F1 score in positive classification has been raised to 58 percent. From all three methods combined with model utilization, the Logistic Regression is the best model.

Keywords: Sentiment Analysis, NLP, Text Mining

1. บทนำ

เครือข่ายสังคมออนไลน์ในปัจจุบัน เป็นพื้นที่ที่ใช้ในการแสดงออกต่าง ๆ ของผู้คน ทั้งโพสต์กิจกรรมในชีวิตประจำวัน ติดต่อสื่อสาร ซึ่งทั้งหมดนี้สามารถโพสต์เป็นข้อความ เสียง หรือวิดีโอ แต่ส่วนใหญ่จะเกิดเป็นข้อความมากกว่า ทั้งการแสดงความเห็นใต้โพสต์ การรับ-ส่งข้อความ ซึ่งมีทั้งคำพูดในทางที่ดีและไม่ดีปะปนกันไป แต่ข้อความที่แสดงความคิดเห็นบนสังคมออนไลน์มีอยู่จำนวนมาก การที่จะให้ผู้คนแต่ละคนมาทำความเข้าใจในข้อความแต่ละอันว่ามีความรู้สึกอย่างไรจึงเป็นไปได้ยาก การวิเคราะห์ความคิดเห็นบนสังคมออนไลน์โดยใช้ความสามารถของคอมพิวเตอร์สามารถช่วยลดปัญหาตามที่กล่าวมาได้ แต่ความท้าทายหนึ่ง คือ ข้อความแสดงความคิดเห็นส่วนใหญ่ที่อยู่บนอินเทอร์เน็ตนิยมใช้ภาษาที่มีโครงสร้างไม่แน่นอน (Unstructured data) หรือเป็นภาษาธรรมชาติ(Natural language) ไม่ถูกต้องตามหลักไวยากรณ์ทางภาษาทำให้ยากต่อการวิเคราะห์ ในงานวิจัยนี้จึงได้นำเทคนิคการวิเคราะห์เหมืองข้อความ (Text mining) การประมวลภาษาธรรมชาติ (Natural Language Processing: NLP) เทคนิคการจัดการกับชุดข้อมูลไม่สมดุล (Imbalance data) และเทคนิคอื่น ๆ มาประยุกต์ใช้เพื่อวิเคราะห์ความคิดเห็นที่เป็นภาษาไทยของผู้คนบนเครือข่ายสังคมออนไลน์ ซึ่งเรียกว่าการวิเคราะห์ความรู้สึก (Sentiment analysis) ซึ่งงานวิจัยนี้จะอธิบายแนวคิด ทฤษฎี เทคนิคต่าง ๆ ที่เกี่ยวข้อง รวมถึงกระบวนการในการวิเคราะห์ การสร้างแบบจำลอง และแสดงตัวอย่างจากงานศึกษาวิจัยที่เกี่ยวข้องกับการวิเคราะห์ความรู้สึกของข้อความ โดยจะทำการศึกษาผ่านเว็บไซต์บนสื่อสังคมออนไลน์ ผ่านทางทวิตเตอร์เอพีไอ (Twitter API) โดยกรณีศึกษาที่ผู้วิจัยเลือกมาคือเรื่องการศึกษาไทย ซึ่งหัวข้อนี้เป็นที่ถกเถียงกันในปัจจุบัน

2. ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

2.1 การประมวลผลภาษาธรรมชาติ [1] เป็นการแปลภาษาที่มนุษย์ใช้สื่อสารกัน หรือภาษาธรรมชาติ ให้คอมพิวเตอร์สามารถเข้าใจได้ หรือทำให้เป็นโครงสร้าง

2.2 การตัดคำ (Word segmentation) [2] คือกระบวนการนำข้อความที่เป็นประโยค หรือข้อความที่มีความยาวมาตัดเป็นคำ ๆ เพื่อนำไปวิเคราะห์

2.3 การปรับข้อมูลไม่สมดุล (Imbalance data) การสุ่มตัวอย่างลด (Random Under Sampling: RUS) เป็นการสุ่มลดจำนวนข้อมูลตัวอย่างจากกลุ่มข้อมูลที่มีในกลุ่มมาก ให้มีขนาดน้อยเท่ากับกลุ่มข้อมูลที่น้อย การสุ่มตัวอย่างเกิน (Random Over Sampling: ROS) เป็นการสุ่มเกินจำนวนข้อมูลตัวอย่างจากกลุ่มข้อมูลที่มีขนาดน้อยให้มีขนาดมากเท่ากับกลุ่มข้อมูลที่มากสุด การสังเคราะห์ข้อมูลเพิ่ม (Synthetic Minority Oversampling Technique: SMOTE) [3-4] เป็นการสังเคราะห์สุ่มเกินข้อมูลขึ้นมาใหม่จากข้อมูลที่มีอยู่

2.4 การประเมินหรือวัดประสิทธิภาพของแบบจำลองการวิเคราะห์ข้อมูล (Validation) งานวิจัยนี้ [5] จะใช้ค่าความถูกต้อง (Accuracy), ค่าความแม่นยำ (Precision), ค่าความครบถ้วน (Recall), ค่าความถ่วงดุล (F1-Score) โดยค่าความถูกต้อง คือการวัดค่าความถูกต้องโดยรวมระหว่างค่าจริงและค่าการทำนายของทุก ๆ คลาส

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

ค่าความแม่นยำคือ อัตราส่วนที่โมเดลทำนายถูกหารด้วยจำนวนการทำนายที่ทำนายว่าเป็นคลาสที่กำลังพิจารณาทั้งถูกและผิด

$$\text{Precision} = \frac{TP}{TP+FP}$$

ค่าความครบถ้วนคือ อัตราส่วนการวัดค่าการทำนายที่ทำนายได้ถูกต้องตรงกับค่าจริงจากจำนวนของค่าจริงทั้งหมดของคลาสที่กำลังพิจารณา

$$\text{Recall} = \frac{TP}{TP+FN}$$

ค่าความถ่วงดุลคือ ค่าเฉลี่ยของผลบวกของส่วนกลับแต่ละค่าของข้อมูลสถิติระหว่างค่าความแม่นยำและค่าความครบถ้วน (Harmonic mean)

$$\text{F1} = 2 * \left(\frac{precision*recall}{precision+recall}\right)$$

2.5 งานวิจัยที่เกี่ยวข้อง

กานดา แผ่วัฒนากุล และปราโมทย์ ลือนาม [6] ได้ทำการวิเคราะห์ความเห็นจากเครือข่ายสังคมออนไลน์ ซึ่งได้ใช้ข้อความเป็นภาษาไทย และได้ตัดประโยคก่อนนำไปวิเคราะห์ โดยใช้โมเดล ต้นไม้ตัดสินใจ นาอีฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน มาสร้างโมเดลในการจำแนก

Younis, E. M. [7] วิเคราะห์ความคิดเห็นจากข้อความบนสื่อสังคมออนไลน์ เกี่ยวกับผลิตภัณฑ์และงานบริการ ใช้อัลกอริทึมในการจัดหมวดหมู่ การทำคลัสเตอร์ และอื่น ๆ ซึ่งงานนี้ได้ลบคำแสลงที่มักใช้ในออนไลน์ มีการตัดคำแบบ Term - document matrix

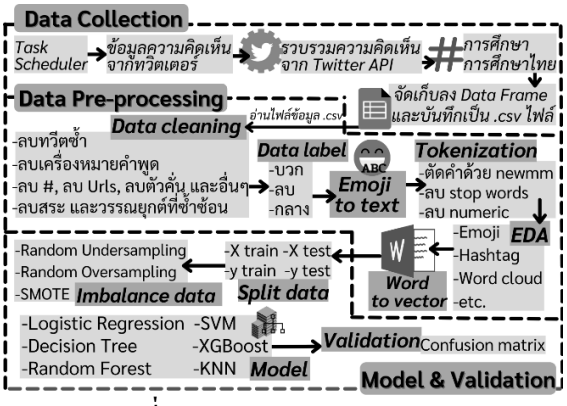
อนัตต์ชัย ชุติภาสเจริญ [8] ได้เปรียบเทียบประสิทธิภาพของโมเดลที่ใช้ทำนายและคุณลักษณะ ที่มีต่อโอกาสความสำเร็จในการโอนเงินข้ามประเทศ โดยใช้ต้นไม้ตัดสินใจ นาอีฟเบย์ และเพื่อนบ้านใกล้สุด มาทำการเปรียบเทียบ และทดสอบประสิทธิภาพได้ค่าความถูกต้องเป็น 99.90% 96.71% และ99.55% ตามลำดับ

Pong-Inwong, C. และ Songpan, W. [9] ได้รวบรวมความเห็นจากคำถามปลายเปิดจากนักศึกษาต่อการสอน เสนอวิธีการปรับปรุงด้วยการทำเหมืองข้อมูล กฎการเชื่อมโยงเพื่อวิเคราะห์วลีความรู้สึก ซึ่งใช้เทคนิค SPPM ในการจำแนกทัศนคติการสอน มีความแม่นยำสูงสุด 87.94% เมื่อเทียบกับโมเดลลักษณนามอื่น ๆ

พัชรียา ทองพูล และคณะ [10] ได้เปรียบเทียบวิธีการปรับข้อมูลไม่สมดุล 4 วิธีคือ การสุ่มเกิน SMOTE การสุ่มลด และการสุ่มผสมผสาน กับการจำแนก 4 วิธี เพื่อทดสอบว่าวิธีใดที่จะทำให้โมเดลจำแนกได้ดีที่สุด

3. วิธีการดำเนินงาน

ดังภาพที่ 1 งานวิจัยนี้ประกอบด้วย 3 ขั้นตอนใหญ่คือ การเก็บข้อมูล การเตรียมข้อมูล ท้ายสุดคือการสร้างโมเดลและการประเมินประสิทธิภาพของโมเดลรายละเอียด ดังนี้



ภาพที่ 1: ลำดับการทำงาน (Process)

3.1 การเก็บข้อมูล (Data collection) การเก็บข้อมูลแบบอัตโนมัติโดยใช้โปรแกรมตัวกำหนดตารางเวลางาน (Task Scheduler) โดยสร้างไฟล์ชุด (Batch Files) ด้านในมีไพธอนพาร์ทและไพธอนสคริปต์ไฟล์ จากนั้นตั้งค่าให้โปรแกรมรันสัปดาห์ละครั้ง ทุก ๆ วันจันทร์เวลา 01.00 น. ซึ่งภายในไฟล์มีการขอ Twitter API เพื่อเก็บ

ข้อความบนทวิตเตอร์ โดยใช้แฮชแท็กในการค้นหาข้อมูล คือ #การศึกษาไทย และ#การศึกษา

3.2 การเตรียมข้อมูล (Data pre-processing) การทำความสะอาดข้อมูล (Data cleaning) เป็นการตรวจสอบและแก้ไขข้อมูล เพื่อให้ข้อมูลอยู่ในรูปแบบที่ถูกต้อง การเลเบล (Label) ผู้วิจัยได้เลเบลข้อความด้วยตนเอง ซึ่งแบ่งออกเป็น 3 คลาสคือ คลาสบวก คลาสลบ และคลาสกลาง โดยทำการเลเบลในระดับเอกสาร (Document level) การแปลงอิโมจิเป็นคำไทย เป็นการเปลี่ยนรูปแบบของอิโมจิให้เป็นคำภาษาไทย โดยใช้ไลบรารี่ของ Pythainlp ที่นำเข้า emoji_to_thai มาทำการแปลง เช่น 🙄 เป็นคำว่า กลอกตา เป็นต้น การตัดคำ เป็นการนำข้อความมาทำการตัดแบ่งเป็นคำ ๆ โดยใช้ newmm มาช่วยในการตัดคำ ซึ่งเป็นหนึ่งในไลบรารี่ของ Pythainlp การสำรวจข้อมูล (Exploratory Data Analysis: EDA) เป็นการตรวจสอบข้อมูลที่ได้มาก่อนนำไปใช้ สำรวจข้อมูลในมุมต่าง ๆ ในทุก ๆ ตัวแปร หรือเปรียบเทียบกันระหว่างตัวแปร ดังตัวอย่างในภาพที่ 2 ที่แสดงให้เห็นถึงข้อความที่มักพบมากที่สุดในคลาสบวก การแทนข้อความ คือก่อนการนำไปสร้างโมเดลต้องแปลงข้อมูลที่อยู่ในรูปแบบที่สามารถนำไปประมวลผลได้

3.3 การแบ่งข้อมูล และการจัดการข้อมูลไม่สมดุล

ผู้วิจัยได้แบ่งข้อมูลออกเป็นข้อมูลฝึกสอน และข้อมูลทดสอบจำนวน 4,096 และ 1,756 ข้อความ ตามลำดับ จาก 5,852 ข้อความ ซึ่งในแต่ละคลาสของข้อมูลฝึกสอนมีจำนวนไม่เท่ากัน จึงแก้ปัญหาของข้อมูลที่ไม่สมดุลนี้จาก 3 วิธีคือ การสุ่มตัวอย่างลด การสุ่มตัวอย่างเกิน และ SMOTE

3.4 การสร้างแบบจำลองข้อมูล เป็นขั้นตอนในการสร้างโมเดลด้วยเทคนิคต่าง ๆ ได้แก่ Logistic Regression, XGBoost, Decision Tree, SVM, Random Forest, K-Nearest Neighbor ในการสร้างกฎเพื่อทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น

3.5 การประเมินประสิทธิภาพแบบจำลองข้อมูล ในงานวิจัยนี้จะใช้ Confusion matrix ในการวัดประสิทธิภาพของโมเดล ค่าที่นำมาวัดได้แก่ ค่าความถ่วงดุล, ค่าความครบถ้วน, ค่าความแม่นยำ และ ค่าความถูกต้อง

ภาพที่ 2: ความถี่ของคำที่ปรากฏในคลาสบวก

Tweets : บางทีครูก็สั่งงานเหมือนไม่อยากให้เด็กจบ😒😒#การศึกษาไทย #การบ้าน #เด็กมีคนเดียว #วิชาสามัญ #โรงเรียน

Cleaning : บางทีครูก็สั่งงานเหมือนไม่อยากให้เด็กจบ อดทน โล่งอก

ภาพที่ 3: ตัวอย่างการทำความสะอาดข้อมูล

4. ผลการศึกษา

4.1 การเก็บข้อมูล เก็บข้อมูลแบบอัตโนมัติโดยค้นหาทวีตจาก #การศึกษาไทย และ #การศึกษา ทั้งหมด 5,852 ข้อความ มีคลาสลบ 3,346 ข้อความ คลาสบวก 257 ข้อความ และคลาสกลาง 2,249 ข้อความ

4.2 การเตรียมข้อมูล ขั้นแรกทำความสะอาดข้อมูล โดยลบข้อมูลที่ซ้ำกันออก ลบสัญลักษณ์ที่ไม่ได้ใช้ ลบ URLs ลบแฮชแท็ก ลบข้อความที่อยู่ใน <> ลบ @Users และอื่น ๆ หลังทำความสะอาดข้อมูลจึงจะทำการเลเบลข้อมูลในระดับประโยค แบ่งเป็น 3 คลาสคือ คลาสบวก คลาสลบ และคลาสกลาง จากนั้นแปลงอิโมจิให้เป็นคำภาษาไทยโดยใช้ emoji_to_thai ตัวอย่างในภาพที่ 3 และตัดคำโดยใช้ newmm โดยลบตัวเลข คำฟุ่มเฟือย (Stop word) ทำการสำรวจข้อมูลเพื่อตรวจสอบข้อมูลก่อนนำไปใช้ และสุดท้ายเป็นการทำให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเข้าใจได้ โดยแปลงข้อความเป็นเวกเตอร์ ซึ่งใช้วิธีแบบ TF-IDF เพื่อให้โมเดลวิเคราะห์ได้

4.3 การแบ่งข้อมูล การจัดการกับข้อมูลที่ไม่สมดุล การสร้างโมเดล และการวัดประสิทธิภาพ ทำการแบ่งชุดข้อมูลออกเป็นชุดฝึกสอน 4,096 ข้อความ และชุดทดสอบ 1,756 ข้อความ จากนั้นเปรียบเทียบค่าที่ได้ก่อน-หลังการ

ปรับข้อมูลที่ไม่สมดุลของชุดข้อมูลฝึกสอน ด้วย 3 วิธีการ คือ การสุ่มตัวอย่างลด การสุ่มตัวอย่างเพิ่ม และ SMOTE ซึ่งแบ่งข้อมูลได้ตามตารางที่ 1 จากตารางที่ 2-5 แสดงผลค่าความถ่วงดุล, ค่าความครบถ้วน, ค่าความแม่นยำ และ ค่าความถูกต้อง จากเมทริกซ์ความสับสน โดยกำหนดไฮเปอร์พารามิเตอร์ ให้แต่ละโมเดล ดังนี้ **Logistic Regression**: C=2, max_iter=100, penalty='l2', solver='liblinear', **XGBoost**: max_depth=5, gamma=2, subsample=0.6, colsample_bytree=1.0, **Decision tree**: max_depth=20, min_samples_leaf=20, criterion='entropy', **Random forest**: min_samples_leaf=3, min_samples_split=2, max_depth=6, min_sample_leaf=3, max_leaf_node=5 , n_estimators=200, max_samples=0.8, **KNN**: p=2, n_jobs=-1, metric='minkowki', weights='distance', algorithm='auto', n_neighbors=5, leaf_size=30,SVM: C=100, gamma=0.001, kernel='rbf' ซึ่งกำหนดโดยใช้การค้นหาแบบกริดในการค้นหาตัวไฮเปอร์พารามิเตอร์ ที่ดีที่สุดที่เหมาะสมกับโมเดลแต่ละตัว จากตาราง LGR คือ Logistic Regression, XGB คือ XGBoost, DT คือ Decision Tree, RFT คือ Random Forest, KNN คือ K-Nearest Neighbor, SVM คือ Support Vector Machine และ Avg คือค่าเฉลี่ยของทุกโมเดล ซึ่ง (-1) แทนคลาสลบ (0) แทนคลาสกลาง และ (1) แทนคลาสบวก ซึ่งใช้ข้อมูลฝึกสอนในการปรับข้อมูลที่ไม่สมดุล และใช้ข้อมูลทดสอบเป็นชุดข้อมูลเดียวกันทั้งหมด จากตารางที่ 2 ค่าความถ่วงดุลของคลาสบวกมีค่าที่ต่ำเฉลี่ย 17% เนื่องจากจำนวนข้อมูลน้อยที่สุดจากทั้ง 3 คลาส ค่าความถูกต้องของแต่ละโมเดลมีค่าต่ำเช่นกันเฉลี่ยที่ 41% และค่าที่ได้ส่วนใหญ่ทำนายได้ดีในคลาสลบ ทางผู้วิจัยจึงทำการปรับข้อมูลที่ไม่สมดุลของชุดข้อมูลฝึกสอน เพื่อเพิ่มประสิทธิภาพของค่าความถ่วงดุลในคลาสบวก

ตารางที่ 1 ข้อมูลในแต่ละคลาสก่อน-หลัง ข้อมูลที่ไม่สมดุล

	Pos	Neg	Neu
Train data (pre)	186	2,339	1,571
Test data	71	1007	678
Random under sampling	186	186	186
Random over sampling	2,339	2,339	2,339
SMOTE	2,339	2,339	2,339

ตารางที่ 2 แสดงค่าที่ได้ก่อนการปรับข้อมูลที่ไม่สมดุล

B4 Imbalance		LGR	XGB	DT	RFT	KNN	SVM	Avg
F1 Score	-1	0.59	0.56	0.48	0.62	0.29	0.53	0.51
	0	0.45	0.41	0.38	0.36	0.46	0.43	0.42
	1	0.05	0.14	0.19	0.1	0.33	0.23	0.17
Recall	-1	0.7	0.62	0.55	0.77	0.24	0.68	0.59
	0	0.58	0.54	0.48	0.44	0.63	0.48	0.53
	1	0.03	0.08	0.11	0.06	0.27	0.14	0.12
Precision	-1	0.51	0.52	0.43	0.51	0.38	0.43	0.46
	0	0.36	0.33	0.31	0.31	0.37	0.39	0.35
	1	1.00	0.5	0.57	0.57	0.42	0.67	0.62
Accuracy		0.44	0.41	0.38	0.42	0.38	0.43	0.41

ตารางที่ 3 การปรับข้อมูลที่ไม่สมดุลด้วยวิธี RUS

RUS		LGR	XGB	DT	RFT	KNN	SVM	Avg
F1 Score	-1	0.48	0.55	0.5	0.57	0.49	0.47	0.51
	0	0.42	0.44	0.42	0.43	0.41	0.39	0.42
	1	0.58	0.62	0.52	0.63	0.49	0.65	0.58
Recall	-1	0.46	0.54	0.55	0.56	0.49	0.41	0.5
	0	0.42	0.44	0.38	0.37	0.42	0.39	0.4
	1	0.59	0.63	0.52	0.72	0.46	0.73	0.61
Precision	-1	0.49	0.56	0.45	0.57	0.49	0.55	0.52
	0	0.42	0.44	0.47	0.51	0.39	0.39	0.44
	1	0.57	0.6	0.53	0.55	0.51	0.58	0.56
Accuracy		0.49	0.54	0.48	0.55	0.46	0.51	0.51

ตารางที่ 4 การปรับข้อมูลที่ไม่สมดุลด้วยวิธี ROS

ROS		LGR	XGB	DT	RFT	KNN	SVM	Avg
F1 Score	-1	0.63	0.58	0.44	0.62	0.51	0.56	0.56
	0	0.49	0.41	0.42	0.3	0.43	0.52	0.43
	1	0.53	0.35	0.56	0.61	0.27	0.54	0.48
Recall	-1	0.63	0.62	0.42	0.65	0.54	0.52	0.56
	0	0.59	0.52	0.41	0.24	0.56	0.68	0.5
	1	0.42	0.23	0.59	0.7	0.17	0.42	0.42
Precision	-1	0.62	0.54	0.46	0.59	0.49	0.62	0.55
	0	0.42	0.34	0.43	0.4	0.34	0.43	0.39
	1	0.71	0.76	0.53	0.54	0.63	0.73	0.65
Accuracy		0.55	0.46	0.47	0.53	0.42	0.54	0.5

จากตารางที่ 3-5 พบว่าโดยผลจากวิธี RUS วิธี ROS และวิธี SMOTE ค่าความถ่วงดุลคลาสบวกโดยเฉลี่ยเพิ่มขึ้นที่ 58% 48% และ 36% ตามลำดับ และค่าความถูกต้องเพิ่มขึ้นเฉลี่ยในแต่ละวิธีที่ 51% 50% และ 47% ตามลำดับ เมื่อเปรียบเทียบการปรับข้อมูลที่ไม่สมดุล จากทั้ง 3 แบบพบว่าค่าความถูกต้องส่วนใหญ่มีค่าเพิ่มขึ้นหลังจากปรับแก้ข้อมูลแล้วโดยโมเดลที่ทำนายแม่นยำน้อยส่วนใหญ่คือ K-NN เฉลี่ยจากทั้ง 3 วิธีที่ 43.33% ส่วนโมเดลที่ทำนายได้แม่นยำ

ส่วนใหญ่จากทั้ง 3 วิธี เฉลี่ย 52.67% คือ Logistic Regression โดยวิธี RUS ให้ค่าความถูกต้องเฉลี่ยของการทำนายที่ดีที่สุด 51% โดยมีโมเดล Random forest ที่ให้ค่าความถูกต้องที่ 55%

ตารางที่ 5 การปรับข้อมูลที่ไม่สมดุลด้วยวิธี SMOTE

SMOTE		LGR	XGB	DT	RFT	KNN	SVM	Avg
F1 Score	-1	0.63	0.62	0.48	0.55	0.53	0.59	0.57
	0	0.48	0.46	0.43	0.46	0.42	0.44	0.45
	1	0.49	0.24	0.31	0.41	0.26	0.46	0.36
Recall	-1	0.65	0.7	0.49	0.56	0.58	0.61	0.6
	0	0.56	0.59	0.54	0.58	0.52	0.52	0.55
	1	0.39	0.14	0.23	0.3	0.17	0.37	0.27
Precision	-1	0.61	0.56	0.46	0.54	0.48	0.58	0.54
	0	0.43	0.38	0.36	0.38	0.35	0.38	0.38
	1	0.64	0.77	0.5	0.66	0.52	0.62	0.62
Accuracy		0.54	0.48	0.42	0.48	0.42	0.5	0.47

5. สรุป

ข้อมูลจากทวิตเตอร์ในหัวข้อการศึกษาไทย เป็นข้อมูลที่รวบรวมมาทั้งสิ้น 5,852 ข้อความ แบ่งเป็นคลาสบวก คลาสลบ และคลาสกลาง เมื่อทดสอบและวัดประสิทธิภาพแล้วพบว่าค่าความถูกต้องในแต่ละโมเดลมีค่าโดยเฉลี่ยที่ 41% ผลการทำนายมีแนวโน้มไปทางคลาสลบ ซึ่งคลาสบวกมีการทำนายที่แม่นยำได้น้อยที่สุดจากทั้ง 3 คลาส เนื่องจากข้อมูลมีเกิดอคติ (Bias) ดังนั้นผู้วิจัยจึงปรับข้อมูลให้สมดุลด้วยวิธี RUS วิธี ROS และ SMOTE เพื่อลดปัญหาดังกล่าว และการค้นหาแบบกริดเพื่อหาไฮเปอร์พารามิเตอร์ที่ดีที่สุดของแต่ละโมเดล เมื่อทดสอบและวัดประสิทธิภาพพบว่า สามารถเพิ่มประสิทธิภาพให้โมเดลได้จากทั้ง 3 วิธี ซึ่งสอดคล้องกับงานวิจัยของพัชรียา ทองพูล และคณะ[10] ส่งผลให้ค่าเฉลี่ยความถูกต้องเพิ่มขึ้น เมื่อพิจารณาที่ค่าเฉลี่ยวิธี RUS ให้ผลดีที่สุด ก่อนและหลังการปรับข้อมูลที่ไม่สมดุลโมเดล Logistic regression ให้ค่าความถูกต้องมากที่สุด เฉพาะวิธี RUS เท่านั้นที่โมเดลทำนายดีที่สุดคือ Random forest และยังส่งผลให้ค่าความถ่วงดุลในคลาสบวกเพิ่มขึ้น พบว่าวิธี RUS ให้ประสิทธิภาพที่ดีที่สุดในค่าความถ่วงดุลในคลาสบวก เมื่อดูจากความถี่คำที่ปรากฏในคลาสบวกพบคำที่ปรากฏถี่ เช่น “ชอบ”, “ดี” และ “น่ารัก” เป็นต้น ทั้งนี้ในชุดข้อมูลยังพบอุปสรรคในเรื่องของภาษา บนแพลตฟอร์มนี้นิยมใช้ภาษาที่กระชับ พบคำย่อ คำที่เขียนผิดอาจส่งผลให้การทำฟีเจอร์เวกเตอร์ให้ค่าที่ต่างกัน ซึ่งทางผู้วิจัยจะนำประเด็นนี้ไปพัฒนาต่อไปเพื่อประสิทธิภาพที่ดีขึ้น

เอกสารอ้างอิง

[1] ยืน ภู่วรวรรณ. (2535). **การประมวลผลภาษาธรรมชาติ.** กรุงเทพฯ: สถาบันเทคโนโลยีพระจอมเกล้าธนบุรี.

[2] สุรศักร์ ตั้งสกุล และฐาปนี เฮงสนั่นกูล. (2554). การตัดคำภาษไทยสำหรับข้อความในพิพิธภัณฑ์ปลาน้ำจืด. **วารสารวิทยาศาสตร์บูรพา, 16**(1), 84-93.

[3] Fawcett, T. (2016). **Learning from Imbalanced Classes**. Retrieved November 1, 2021, from https://www.svds.com/learning-imbalanced-classes/?utm_source=kdnuggets&utm_medium=

[4] Garbled. (2013). **Class Imbalance Problem**. Retrieved November 1, 2021, from http://www.chioka.in/class-imbalance-problem/

[5] Chengz. (2560). **วัดประสิทธิภาพ Model จาก Confusion Matrix.** ค้นเมื่อ 2 ธันวาคม 2563 จาก https://medium.com/@cheng-confusion-matrix

[6] กานดา แผ่วัฒนากุล, ปราโมทย์ ลือนาม. (2556). การวิเคราะห์เหมืองความคิดเห็นบนเครือข่ายสังคม ออนไลน์. **วารสารการจัดการสมัยใหม่, 11**(2), 12-20.

[7] Younis, E. M. (2015). **Sentiment analysis and text mining for social media microblogs using Opensource tools: an empirical study**. International Journal of Computer Applications, 112(5).

[8] Anutchai Chutipascharoen and Charun Sanrach. (2559). **A Comparison of the Efficiency of Algorithms and Feature Selection Methods for Predicting the Success of Personal Overseas Money Transfer**. ค้นเมื่อ 1 สิงหาคม 2564, จาก https://so04.tci-thaijo.org/index.php/gskkuhs/article/view/156370.

[9] Pong-Inwong, C., & Songpan, W. (2019). Sentiment analysis in teaching evaluations using sentiment phrase pattern matching (SPPM) based on association mining. **International Journal of Machine Learning and Cybernetics, 10**(8), 2177-2186.

[10] พัชรียา ทองพูล, พิมพ์ชนก จำเรือง, รมย์นลิน บุญฤทธิ์, & สายชลสิน สมบูรณ์ทอง. (2019). การเปรียบเทียบประสิทธิภาพ ในการทำนายผลการปรับความไม่สมดุลของข้อมูล ในการจำแนกด้วยเทคนิคการทำเหมืองข้อมูล. **Thai Journal of Science and Technology, 8**(6), 565-584.