การปรับปรุงเทคนิคการพยากรณ์โรคซึมเศร้าในวัยรุ่น

วิทยานิพนธ์

ของ

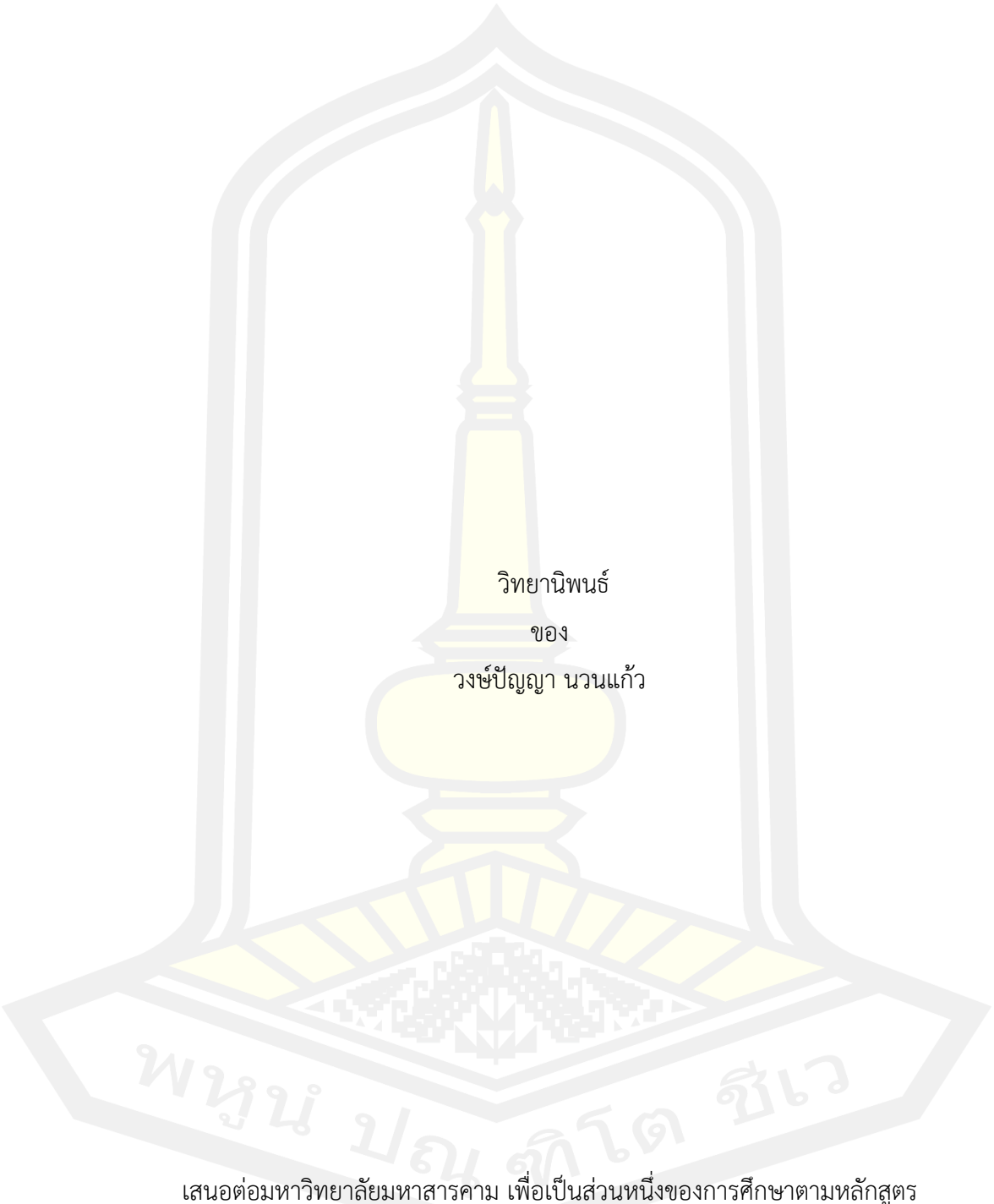
วงษ์ปัญญา นวนแก้ว

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ตุลาคม 2565

การปรับปรุงเทคนิคการพยากรณ์โรคซึมเศร้าในวัยรุ่น



วิทยานิพนธ์

ของ

วงษ์ปัญญา นวนแก้ว

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ตุลาคม 2565

Improvement of prediction technique for adolescent depression disorder

Wongpanya Nuankaew

A Thesis Submitted in Partial Fulfillment of Requirements

for Doctor of Philosophy (Information Technology)

October 2022

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนางวงษ์ปัญญา นวนแก้ว แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(รศ. ดร. วรารัตน์ สงฆ์แป้น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

.....กรรมการ
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

.....กรรมการ
(ผศ. ดร. รพีพร ช่ำชอง)

.....กรรมการ
(ผศ. ดร. โอฬาริก สุรินต๊ะ)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา ปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(รศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การปรับปรุงเทคนิคการพยากรณ์โรคซึมเศร้าในวัยรุ่น
ผู้วิจัย วงษ์ปัญญา นวนแก้ว
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. ฉัตรเกล้า เจริญผล
ปริญญา ปรัชญาดุษฎีบัณฑิต **สาขาวิชา** เทคโนโลยีสารสนเทศ
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2565

บทคัดย่อ

โรคซึมเศร้าเป็นโรคทางจิตเวชเกิดจากความไม่สมดุลของสารเคมีในสมองที่มีผลต่ออารมณ์ ความคิด ความรู้สึก การแสดงอาการของผู้ป่วยมีความแตกต่างกันอยู่ที่ระดับความรุนแรงจนถึงการทำร้ายตนเอง จนนำไปสู่การฆ่าตัวตาย โดยเฉพาะกลุ่มผู้ป่วยวัยรุ่นและวัยทำงานมีความเสี่ยงสูงต่อการเสียชีวิตจากการฆ่าตัวตาย จึงจำเป็นต้องมีวิธีการพยากรณ์โรคซึมเศร้าเพื่อหาแนวทางการป้องกันและการรักษาที่เหมาะสม มีงานวิจัยจำนวนมากได้นำข้อมูลการแสดงความคิดเห็นผ่านโซเชียลมีเดียเป็นการสะท้อนถึงอารมณ์ ความนึกคิดของผู้ใช้งานแต่ละคน ซึ่งได้นำเสนอวิธีการการปรับปรุงการพยากรณ์โรคซึมเศร้า โดยใช้เทคนิคเหมืองข้อมูลและส่วนใหญ่เป็นการพยากรณ์โดยใช้ตัวจำแนกประเภทแบบเดี่ยว แต่ในงานวิจัยนี้ได้นำเสนอวิธีการปรับปรุงค่าน้ำหนักที่เหมาะสมเพื่อการปรับปรุงเทคนิคเหมืองข้อมูลโดยใช้วิธีการเอ็นเซมเบิล ทั้งแบบกำหนดค่าน้ำหนักน้ำหนัก 2 วิธีการ ได้แก่ อัตราการทำนายถูกของคลาสคำตอบ และค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ โดยมีรายละเอียด ดังนี้ 1) คุณลักษณะข้อมูลจากความคิดเห็น ร่วมกับคุณลักษณะภาพ ดำเนินการสกัดคุณลักษณะด้วยวิธีการ Binary term occurrence 2) คัดเลือกคุณลักษณะที่เหมาะสมด้วยวิธีการ Information gain ข้อมูลที่ใช้ในการทดสอบนำมาจาก Twitter และ Instagram ซึ่งเป็นชุดข้อมูลแบบหลายคลาส และแบบไบนารีคลาส 3) เปรียบเทียบและวัดประสิทธิภาพการจำแนกโรคซึมเศร้าจากแบบจำลองทั้งหมด 3 ประเภท ได้แก่ ตัวจำแนกประเภทแบบเดี่ยว วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก และวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก โดยวิธีการเอ็นเซมเบิลแบ่งออกเป็น 4 กลุ่ม ตามจำนวนตัวจำแนกประเภท ประกอบด้วย 3-เอ็นเซมเบิล 4-เอ็นเซมเบิล 5-เอ็นเซมเบิล และ 6-เอ็นเซมเบิล และ 4) ทดสอบนัยสำคัญด้วยวิธีการ Paired samples t-test เพื่อเปรียบเทียบความแตกต่างระหว่างประสิทธิภาพของวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักและแบบไม่กำหนดค่าน้ำหนัก

ผลการทดลองแสดงให้เห็นว่า การปรับปรุงค่าน้ำหนักที่เหมาะโดยวิธีเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักมีประสิทธิภาพดีกว่าวิธีการเอ็นเซมเบิลแบบไม่กำหนดน้ำหนักและตัวจำแนก

ประเภทแบบเดี่ยว มีความถูกต้อง 66.67% และ 87.23% ค่าความแม่นยำ 72.73% และ 88.89% มีค่าระลึก 80.00% และ 92.57% และค่าเฉลี่ยประสิทธิภาพโดยรวม 70.59% และ 88.89% ตามลำดับ โดยวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก มีประสิทธิภาพมากกว่าวิธีการแบบไม่กำหนดค่าน้ำหนักอย่างมีนัยสำคัญที่ 0.05 ทั้งคลาสของบุคคลที่เป็นโรคซึมเศร้าและบุคคลปกติแบบจำลองมีความเหมาะสมสำหรับนำไปประยุกต์หรือพัฒนาเพื่อใช้งานในอนาคต แต่ควรพิจารณาถึงอัตราส่วนระหว่างชุดข้อมูลคลาสอาการของบุคคลปกติและอาการบุคคลคนที่เป็นโรคซึมเศร้าของชุดข้อมูลสำหรับการเรียนรู้แบบจำลองในคลาสอาการของบุคคลปกติมีจำนวนน้อยกว่ามาก จึงทำให้เรียนรู้ชุดคำได้น้อย

คำสำคัญ : การปรับปรุงเทคนิคการพยากรณ์, การจำแนกข้อมูล, ค่าน้ำหนักความน่าจะเป็น, การจำแนกโรคซึมเศร้า, เอ็นเซมเบิล

TITLE Improvement of prediction technique for adolescent depression disorder

AUTHOR Wongpanya Nuankaew

ADVISORS Assistant Professor Chatklaw Jareanpon , Ph.D.

DEGREE Doctor of Philosophy **MAJOR** Information Technology

UNIVERSITY Mahasarakham University **YEAR** 2022

ABSTRACT

Depression disorder is a mental illness caused by unbalanced levels of brain chemicals, which affect emotions, thinking, and feelings. Depressive patients exhibit different symptoms depending on the severity. It can range from self-harm to suicide. Especially, younger patients and working-age patients are more at risk of committing suicide. Early identification of depression disorder is needed so that appropriate preventative and curative care can be implemented and provided in time. Many studies have used social media post data to find indications and interpretations of the emotions, feelings, and thoughts of each social media user. To improve the classification of depression, data mining techniques were widely presented and most of them employed a single classifier. This research proposes two weight adjustment methods for better fitting the weights to improve the weighted voting ensemble of the data mining technique: 1) true positive weighted rate, and 2) average probability weighted as detailed below: 1) Extract the features into opinion features and image features by using binary term occurrence, 2) Select the optimal number of features by using the information gain method. The test data is publicly available on Twitter and Instagram. The datasets are multi-class and binary-class data, 3) Compare and measure the effectiveness of the three classification models: single classifier, unweighted voting ensemble, and weighted voting ensemble, and the ensemble method is grouped into four groups: 3-Ensemble, 4-Ensemble, 5-Ensemble, and 6-Ensemble classifier ensembles, and 4) Test the statistical significance using a paired samples t-test to compare the differences of the

effectiveness of weighted voting ensemble and unweighted voting ensemble models.

The results indicate that the weighted voting ensemble with a better fitting weight adjustment is more effective than the unweighted voting ensemble and single classifier. The result shows a maximum accuracy of 66.67% and 87.23% , a precision of 72.73% and 88.89%, recall of 80% and 92.57%, and F1 of 70.59% and 88.89% , consecutively. The weighted voting ensemble statistically performs better than the unweighted voting ensemble at the significance level of 0.05 for both normalcy and depression classes of data. The method can be further applied or developed in the future with careful consideration of the ratio of datasets of both classes in the training set. The number of datasets of the normalcy class is much less than that of the depression class.

Keyword : Improving data mining, Classification, Probability-weighted, Depressive disorder classification, Ensemble learning

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จสมบูรณ์ลงได้ด้วยดีเพราะได้รับความกรุณาและความช่วยเหลือจาก ผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล อาจารย์ที่ปรึกษาวิทยานิพนธ์ ที่ได้กรุณาให้คำปรึกษา ให้ความรู้ ให้แนวคิด และให้กำลังใจตลอดจนคำแนะนำในการปรับปรุงแก้ไขข้อบกพร่องต่างๆ ผู้วิจัยทราบซึ้งในความกรุณาและขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.วรารัตน์ สงฆ์แป้น ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.พัฒนพงษ์ ชมภูวิเศษ ผู้ช่วยศาสตราจารย์ ดร.รพีพร ช่ำชอง และผู้ช่วยศาสตราจารย์ ดร.โอฬาริก สุรินต๊ะ กรรมการสอบวิทยานิพนธ์ ที่กรุณาให้คำปรึกษา คำแนะนำตลอดจนแนวทางการแก้ไขปัญหาข้อบกพร่องในการทำวิทยานิพนธ์

ขอขอบพระคุณมหาวิทยาลัยมหาสารคามที่สนับสนุนทุนวิจัยสำหรับนิสิตระดับบัณฑิตศึกษา และมหาวิทยาลัยราชภัฏมหาสารคามที่ให้การสนับสนุนการศึกษาต่อระดับปริญญาเอกในครั้งนี้

ขอขอบคุณคณาจารย์ทุกท่าน คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม ที่ได้อบรมสั่งสอน ให้ความรู้ทางด้านวิชาการ เพื่อนนิสิตหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ และเพื่อนร่วมงานที่คอยช่วยเหลือ แนะนำความรู้ ข้อคิดเห็นและเป็นกำลังใจให้กับผู้วิจัยมาโดยตลอด

สุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา คู่สมรส บุตร และญาติพี่น้องและเพื่อนๆ ทุกคน ที่เป็นแรงผลักดันและเป็นกำลังใจสำคัญให้ตลอดเวลาในการศึกษาเล่าเรียน หากวิทยานิพนธ์ฉบับนี้ มีประโยชน์ และคุณค่าทางการศึกษา ผู้เขียนขอยกความดีทั้งหมดให้กับบุคคลที่ผู้วิจัยได้กล่าวถึง

วงษ์ปัญญา นวนแก้ว

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพประกอบ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

โรคซึมเศร้า (Major depressive disorder: MDD) เป็นอาการเจ็บป่วยทางจิตที่ผู้ป่วยจะเกิดอารมณ์ซึมเศร้า หดหู่ ร้องไห้ง่าย สิ้นหวัง รู้สึกด้อยค่า จนไม่มีความสนใจต่อกิจกรรมหรือสถานการณ์ที่ต้องเผชิญ อย่างน้อย 2 สัปดาห์ (Negrão & Gold, 2007) ความรู้สึกเหล่านี้เกิดจากความไม่สมดุลของสารสื่อประสาทสมองที่มีผลต่อความคิด ความรู้สึก อารมณ์และพฤติกรรม ส่งผลทำให้มีสภาวะของโรคซึมเศร้า โดยสามารถเกิดขึ้นได้ในทุกเพศทุกวัยเหมือนกับโรคทางกายอื่นทั่วไป ปัจจัยที่มีผลต่อการเสี่ยงเป็นโรคภาวะซึมเศร้า ได้แก่ ความเครียด การถูกสังคมรังเกียจ สภาพทางจิตใจที่เกิดจากการเลี้ยงดู การเผชิญสถานการณ์เลวร้าย พฤติกรรมและลักษณะนิสัยเฉพาะ เป็นต้น อาการและความรุนแรงของผู้ป่วยโรคซึมเศร้าในแต่ละประเภทมีลักษณะเฉพาะตัว เช่น การขาดความภูมิใจตนเอง การขาดความสนใจในกิจกรรมที่ทำให้สุขใจ อาการอ่อนเพลียไร้เรี่ยวแรง อาการปวดเมื่อยร่างกายโดยไม่มีสาเหตุ นอนหลับยาก นอนหลับมากเกินไป อาการหลงผิดหรือเกิดอาการประสาทหลอน แยกตัวออกจากสังคม ไม่มีสมาธิ อาจร้ายแรงถึงทำร้ายตัวเอง การคิดฆ่าตัวตาย (Mousavian et al., 2018; Ramanuj et al., 2019) องค์การอนามัยโลก (World health organization: WHO) พบว่าในปี 2564 มีผู้ป่วยโรคซึมเศร้าราว 280 ล้านคน (WHO, 2021) ประเทศที่มีจำนวนผู้ป่วยโรคซึมเศร้ามากที่สุด ได้แก่ ยูเครน (6.3%) สหรัฐอเมริกา (5.9%) เอสโตเนีย (5.9%) ออสเตรเลีย (5.9%) และบราซิล (5.8%) ตามลำดับ (Depression Rates by Country 2021, 2021) โรคซึมเศร้าเป็นสาเหตุอันดับสองของการฆ่าตัวตาย โดยเฉพาะผู้ป่วยโรคซึมเศร้าวัยรุ่นและผู้ใหญ่จะมีความเสี่ยงเสียชีวิตจากการฆ่าตัวตายหรือมีอาการร่วมกับความผิดปกติทางอารมณ์ชนิดอื่นเกิดขึ้น จากสถิติมีผู้ฆ่าตัวตายราว 700,000 คน ต่อปี กลุ่มที่มีอัตราการฆ่าตัวตายสูงอายุระหว่าง 15-29 ปี และมีผู้ป่วยโรคซึมเศร้าอีกจำนวนมากที่ล้มเหลวในการฆ่าตัวตาย (WHO, 2021)

การวิจัยเกี่ยวกับผลกระทบจากโซเชียลมีเดีย (Lin et al., 2016) พบว่าคนอายุระหว่าง 19-32 ปี ที่มีการใช้งานโซเชียลมีเดียบ่อยๆ หรือปริมาณมากๆ มีโอกาสเสี่ยงต่อการเกิดภาวะซึมเศร้าสูงมาก สาเหตุที่โซเชียลมีเดียมีความเกี่ยวพันกับภาวะซึมเศร้าเพราะจิตใจของมนุษย์เกิดความอ่อนไหวได้ง่ายเมื่อได้เห็นหรืออ่านโพสต์ของคนอื่น เนื่องจากการเห็นคนอื่นมีความสุขอาจทำให้เกิดความอิจฉาหรือน้อยเนื้อต่ำใจสะสมเป็นระยะๆ อาจรวมถึงการกำหนดเกณฑ์การมีความสุขของตนเองตามคนอื่นที่ได้เห็นจากโซเชียลมีเดียโดยไม่รู้ตัว และในปี 2565 พบว่ามีบัญชีผู้ใช้งานสื่อโซเชียลมีเดียทั่วโลก 465 ล้านบัญชี (Simon, 2022) รูปแบบข้อมูลที่ใช้ในโซเชียลมีเดีย ได้แก่ ข้อความ ภาพนิ่ง ภาพเคลื่อนไหว วิดีโอ อิโมติคอน กิจกรรมการใช้งาน และอื่นๆ จากจำนวนผู้ใช้งานและความถี่ในการใช้งาน จึงทำให้

เกิดข้อมูลมหาศาลที่สามารถนำมาวิเคราะห์หาความรู้ สามารถตอบโจทย์พฤติกรรมการแสดงออกของมนุษย์เพื่อหาแนวทางในการแก้ไขปัญหาเกี่ยวกับโรคซึมเศร้า การสำรวจงานวิจัยด้านสุขภาพจิตที่ได้นำข้อมูลพฤติกรรมการแสดงออกผ่านแพลตฟอร์มโซเชียลมีเดียมาศึกษาและวิเคราะห์ข้อมูล แสดงถึงการให้ความสำคัญในการวิเคราะห์ด้านสุขภาพจิตของผู้ใช้โซเชียลมีเดีย (Harrigian et al., 2021)

งานวิจัยที่วิเคราะห์ข้อมูลจากการใช้งานโซเชียลมีเดียมักจะมีความเกี่ยวข้องกระบวนการเหมืองข้อความได้มีวิธีการทดลองที่ความแตกต่างกันไปตามวัตถุประสงค์ของงาน เช่น การคัดกรองข้อความ การคัดเลือกคุณลักษณะ ตัวจำแนกประเภท และการวัดประสิทธิภาพ การศึกษาเกี่ยวกับการจำแนกภาวะซึมเศร้า ได้แก่ การจำแนกประเภทอาการของโรค การจำแนกการมีภาวะซึมเศร้า การจำแนกระดับความรุนแรง และอื่นๆ ในการจำแนกภาวะซึมเศร้าเหล่านี้ จึงจำเป็นต้องมีตัวจำแนกประเภทและวิธีการที่เหมาะสมกับชุดข้อมูลที่นำมาทดสอบเพื่อนำไปแก้ไขปัญหาที่ถูกต้อง รวมทั้งขั้นตอนการเตรียมข้อมูลถือว่าเป็นขั้นตอนที่สำคัญที่ทำให้ประสิทธิภาพแบบจำลองเพิ่มขึ้นได้ มีการประยุกต์ใช้การทำเมืองข้อมูลในขั้นตอนการคัดเลือกคุณลักษณะของข้อมูลสำหรับพัฒนาแบบจำลองการจำแนกโรคซึมเศร้า (Mousavian et al., 2018; Doenribram et al., 2019; N. Zhang et al., 2019; W. Zhang et al., 2020) วิธีการคัดเลือกคุณลักษณะเหล่านี้จะช่วยในการลดมิติข้อมูล ลดเวลาการประมวลผล ช่วยเพิ่มประสิทธิภาพ และงานวิจัยอื่นที่มุ่งเน้นการปรับปรุงกระบวนการการเตรียมข้อมูลเพื่อช่วยเพิ่มประสิทธิภาพการทำงานของโมเดลให้ดียิ่งขึ้น เช่น การหาค่าผิดปกติ (Kuo et al., 2018) การระบุค่าของคลาสที่ผิด (Nuankaew & Thongkam, 2020) การปรับความไม่สมดุลให้ข้อมูล (Sawangarreerak & Thanathamathee, 2020)

วิธีการคะแนนเสียงข้างมาก (Voting ensemble method) เป็นวิธีการหนึ่งที่มีประสิทธิภาพของวิธีการเรียนรู้เอ็นเซมเบิล (Ensemble learning) พัฒนาแบบจำลองเพื่อจำแนกข้อมูลในรูปแบบวิธีการที่ไม่กำหนดค่าน้ำหนัก (Unweighted) และวิธีการกำหนดค่าน้ำหนัก (Weighted voting) ทั้งสองวิธีการมีจุดเด่นและข้อบกพร่องที่แตกต่างกันตามชุดข้อมูลที่นำไปใช้งาน และข้อจำกัดของแต่ละวิธีการจะมีผลกระทบต่อประสิทธิภาพการจำแนกประเภทของแต่ละวิธีการ วิธีการกำหนดค่าน้ำหนักจะสามารถเลือกผลลัพธ์ที่มีค่าน้ำหนักมากที่สุดเพื่อเป็นคำตอบในการทำนาย ช่วยลดปัญหาของวิธีการที่ไม่กำหนดค่าน้ำหนักได้ โดยเฉพาะกรณีที่การมีการใช้การรวมตัวจำแนกประเภทจำนวนคู่ จะไม่สามารถเลือกคลาสคำตอบได้ถูกต้องได้ รวมทั้งในกรณีที่จำนวนตัวจำแนกประเภทอ่อนแอ (Weak classifier) หรือตัวจำแนกประเภทที่ทายผลลัพธ์ผิดมีมากกว่าจำนวนตัวจำแนกประเภทแข็งแรง (Strong classifier) หรือตัวจำแนกประเภทที่ทายผลลัพธ์ถูกต้อง จะส่งผลกระทบต่อประสิทธิภาพของแบบจำลอง จึงมีนักวิจัยที่ให้ความสนใจศึกษาเกี่ยวกับการปรับปรุงค่าน้ำหนักเพื่อเพิ่มประสิทธิภาพการทำงานของแบบจำลองด้วยวิธีการและชุดข้อมูลที่แตกต่างกัน เช่น การปรับปรุงค่าน้ำหนักที่เหมาะสมให้กับตัวจำแนกประเภทประเภทและคลาสคำตอบตามประสิทธิภาพการทำนาย

ด้วยขั้นตอนวิธีเชิงวิวัฒนาการ (Evolutionary algorithm) สำหรับชุดข้อมูลเอกสารออนไลน์ (Text documents online) (Onan et al., 2016) การเพิ่มประสิทธิภาพแบบจำลองด้วยวิธีการกำหนดค่าน้ำหนักจากค่าความน่าจะเป็นการเกิดคลาสคำตอบสำหรับวิธีการจำแนกประเภท (Rojarath & Songpan, 2021) แบบจำลองการทำนายด้วยวิธีการกำหนดค่าน้ำหนักสำหรับวินิจฉัยวัณโรค (Osamor & Okezie, 2021) แบบจำลองการทำนายด้วยวิธีการกำหนดค่าน้ำหนักสำหรับวินิจฉัยโรคเบาหวานระยะเริ่มต้น (Sannasi Chakravarthy & Rajaguru, 2022)

งานวิจัยนี้ผู้วิจัยได้ปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น โดยวิธีการเอ็นเซมเบิลเพื่อปรับค่าน้ำหนักที่เหมาะสมเพื่อเพิ่มประสิทธิภาพการทำนาย ผู้วิจัยได้นำตัวจำแนกประเภทที่ได้รับความนิยมสำหรับนำมาสร้างแบบจำลองในงานวิจัย จำนวน 7 ตัวจำแนกประเภท ได้แก่ 1) นาอีฟเบย์ (Naive bayes: NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support vector machines: SVM) 3) ต้นไม้ตัดสินใจ (Decision tree: DT) 4) ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด (K-nearest neighbors: KNN) 5) แรนดอมฟอเรสต์ (Random forest: RF) 6) กราเดียนบูทติ้งทรี (Gradient boosting tree: GBT) และ 7) ตัวแบบเชิงเส้นนัยทั่วไป (Generalized linear model: GLMs) ตัวจำแนกประเภทเหล่านี้ถูกนำมาสร้างแบบจำลองเพื่อให้กำหนดค่าน้ำหนักการทำนายคลาสคำตอบ และวิธีการคัดเลือกจำนวนคุณลักษณะที่เหมาะสมด้วยวิธีการ Information gain (IG) เพื่อลดจำนวนคุณลักษณะ ลดเวลาการประมวลผลแต่ได้ผลลัพธ์ที่มีประสิทธิภาพ ซึ่งใช้ข้อมูลการแสดงความคิดเห็นจากการใช้งานโซเชียลมีเดีย (Social media) โดยเปรียบเทียบ ประสิทธิภาพระหว่างแบบจำลองที่สร้างด้วย 1) ตัวจำแนกประเภทแบบเดี่ยว (Single classifier) 2) วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก และ 3) วิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก เพื่อเป็นแนวทางการในการวินิจฉัยโรคซึมเศร้าในเบื้องต้นเพื่อป้องกันการฆ่าตัวตาย ลดจำนวนผู้ป่วย และการวางแผนการรักษาต่อไป

1.2 คำถามการวิจัย

การแสดงความคิดเห็นจากโพสต์ผ่านโซเชียลมีเดียสะท้อนให้เห็นถึงอารมณ์และความนึกคิดของผู้ใช้งานออกมาผ่าน ข้อความ ภาพนิ่ง ภาพเคลื่อนไหว และอื่นๆ เนื่องด้วยลักษณะการแสดงความคิดเห็นของแต่ละผู้ใช้งานมีความแตกต่างกัน บางผู้ใช้งานชอบแสดงความคิดเห็นหรือระบายอารมณ์ด้วยข้อความ แต่ในบางผู้ใช้งานชอบใช้ภาพสื่อสารหรือใช้โทนสีแทนความรู้สึก การนำเฉพาะข้อความแสดงความคิดเห็นมาวิเคราะห์เพียงอย่างเดียวอาจไม่ครอบคลุมในการนำข้อมูลไปวิเคราะห์เพื่อจำแนกโรคซึมเศร้า ดังนั้นมีความเป็นไปได้หรือไม่ที่จะนำข้อความ ข้อความจากภาพ และภาพ มาวิเคราะห์เพื่อจำแนกโรคซึมเศร้าด้วยวิธีการการทำเหมืองข้อมูล แต่เนื่องด้วยข้อความการแสดงความคิดเห็นจากโพสต์ผ่านโซเชียลมีเดียเป็นข้อมูลที่ไม่มีโครงสร้างประโยค (Unstructured data) หรือเป็นภาษาธรรมชาติ ซึ่งไม่สามารถระบุโครงสร้างที่แน่นอนได้ และอาจไม่ถูกต้องตามหลักไวยากรณ์

ของภาษา ในงานวิจัยนี้ได้ใช้ชุดข้อมูลมาจาก Doenribram et al. (2019) ซึ่งเป็นชุดข้อมูลที่ประกอบไปด้วยข้อความจากแฮชแท็กที่เกี่ยวข้องกับอาการของโรคซึมเศร้าและเก็บข้อมูลข้อมูลจากภาพ คือ ข้อความจากภาพ และสีของภาพ เพิ่มเติม เพื่อเพิ่มคุณลักษณะที่อาจมีผลต่อการจำแนกโรคซึมเศร้า โดยเกิดคำถามเพื่อทำให้เกิดการวิจัย 2 ข้อ ดังนี้

คำถามการวิจัยที่ 1: การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล จากคำถามการวิจัยที่ 1: เนื่องจากการทดลองของ Doenribram et al. (2019) เพื่อจำแนกอาการโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบน Twitter โดยใช้วิธีการแบบไม่กำหนดค่าน้ำหนักเพื่อทำนายคำตอบสุดท้าย พบว่า การทำนายผิดเกิดการจากการให้คะแนนโหวตที่ผิดพลาดเนื่องจากตัวจำแนกประเภทอ่อนแอ แต่มีจำนวนที่มีมากกว่าจึงให้ผลทำนายที่ผิดพลาด ดังนั้นหากมีการกำหนดค่าน้ำหนักให้คำตอบที่ตอบถูกต้องมากกว่าคำตอบที่ผิดด้วยการปรับปรุงค่าน้ำหนักให้คลาสคำตอบจากความน่าจะเป็นของการเกิดคลาสคำตอบ จะช่วยเพิ่มประสิทธิภาพการจำแนกโรคซึมเศร้าได้หรือไม่

คำถามการวิจัยที่ 2: การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ จากคำถามการวิจัยที่ 2: จากข้อมูลการโพสต์เพื่อแสดงความคิดเห็นของผู้ใช้งานบนโซเชียลมีเดียมีหลายลักษณะ เช่น ข้อความ ข้อความจากภาพ ภาพนิ่ง ภาพเคลื่อนไหว เป็นต้น ลักษณะข้อมูลเหล่านี้บ่งบอกถึงความรู้สึก อารมณ์ การมีส่วนร่วม ณ เวลานั้น ๆ (Choudhury et al., 2016; Wendlandt et al., 2017) ที่ผู้ใช้งานได้แสดงออกมาในลักษณะแตกต่างกันไป ซึ่งบางผู้ใช้งานมักจะแสดงความความคิดเห็นเป็นลักษณะข้อความ บางผู้ใช้งานอาจแสดงความคิดความรู้สึกผ่านข้อความจากภาพที่มาจากคนอื่น บางผู้ใช้งานอาจแสดงความคิดความรู้สึกผ่านภาพต่าง ๆ ที่อาจจะสามารถบรรยายหรือบ่งบอกถึงความรู้สึกได้มากกว่าข้อความ ดังนั้นการเพิ่มคุณลักษณะข้อมูลจากภาพ ประกอบไปด้วย ข้อความในภาพ และภาพ ที่ผู้ใช้งานโพสต์ผ่านโซเชียลมีเดียจะช่วยให้การจำแนกโรคซึมเศร้าในวัยรุ่นครอบคลุมและมีประสิทธิภาพมากกว่าการทดสอบเฉพาะชุดข้อความจากการโพสต์เท่านั้น โดยทดสอบกับแบบจำลองที่ได้จากคำถามการวิจัยที่ 1 จะได้ประสิทธิภาพที่ดีหรือไม่เมื่อเพิ่มข้อมูลคุณลักษณะแบบอื่นเข้าไปในแบบจำลอง

คำถามการวิจัยที่ 3: การคัดเลือกเฉพาะข้อความมาวิเคราะห์โรคซึมเศร้า โดยตัดการโพสต์ประเภทอื่นๆ เช่น ภาพ อาจทำให้อาการของโรคซึมเศร้าขาดหายไปจากระยะเวลาที่ได้กำหนดไว้ตามมาตรฐานของคู่มือการวินิจฉัยและสถิติสำหรับความผิดปกติทางจิตฉบับที่ 5 ของสมาคมจิตเวชศาสตร์สหรัฐอเมริกา (American Psychiatric Association, 2013) ซึ่งจะทำให้การประเมินผลการเป็นโรคซึมเศร้ามีความผิดพลาดหรือไม่

1.3 ความมุ่งหมายของการวิจัย

1.3.1 เพื่อพัฒนากระบวนการปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่นด้วยวิธีการเอ็นเซมเบิล โดยการปรับปรุงค่าน้ำหนักที่เหมาะสม

1.3.2 เพื่อวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ

1.4 ขอบเขตการวิจัย

การวิจัยเรื่อง “การปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น” ได้กำหนดของเขตของการวิจัยตามโจทย์วิจัย ดังนี้

1.4.1 ชุดข้อมูลสำหรับทดสอบการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล ในคำถามการวิจัยที่ 1: เลือกใช้ชุดข้อมูลที่เป็นข้อความ จาก Twitter แบ่งออกเป็น 2 ชุดข้อมูล จากงานวิจัยของ Doenribram et al. (2019)

1) ชุดข้อมูลสำหรับเรียนรู้แบบจำลอง (Training set) เพื่อจำแนกอาการโรคซึมเศร้าเป็นข้อมูลที่ได้เก็บรวบรวมข้อความ จากแฮชแท็กจำนวน 10 ประเภทอาการ ประกอบด้วย อาการที่เกี่ยวข้องกับโรคซึมเศร้า 9 อาการของโรคซึมเศร้า และ 1 อาการที่เป็นอาการบุคคลปกติ จำนวน 30,000 ข้อความ

2) ชุดข้อมูลสำหรับการทดสอบแบบจำลอง (Test set) เก็บรวบรวมข้อความข้อความจากภาพ ที่เป็นดารา จำนวน 30 คน แบ่งออกเป็น 2 กลุ่ม คือ กลุ่มที่หนึ่ง ผู้ป่วยโรคซึมเศร้า จำนวน 15 คน และกลุ่มที่สอง บุคคลที่ไม่เป็นโรคซึมเศร้า จำนวน 15 คน ซึ่งเป็นข้อมูลที่มีสถานะแบบสาธารณะ จำนวน 56,933 ข้อความ

1.4.2 ชุดข้อมูลสำหรับทดสอบการวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ ในคำถามการวิจัยที่ 2 และ 3 เลือกใช้ชุดข้อมูลที่เป็นข้อความ ข้อความจากภาพ และค่าเฉลี่ยสีของภาพ จาก Twitter และ Instagram แบ่งออก 2 ชุดข้อมูล

1) ชุดข้อมูลสำหรับเรียนรู้แบบจำลอง (Training set) เพื่อจำแนกโรคซึมเศร้า เป็นข้อมูลที่ได้เก็บรวบรวมข้อความจากภาพ และค่าเฉลี่ยสีของภาพ จากแฮชแท็กจำนวน 10 ประเภทอาการ ประกอบด้วย อาการที่เกี่ยวข้องกับโรคซึมเศร้า 9 อาการของโรคซึมเศร้า และ 1 อาการที่เป็นอาการของบุคคลปกติจาก Twitter และ Instagram ภาพ จำนวน 5,300 ภาพ

2) ชุดข้อมูลสำหรับการทดสอบแบบจำลอง (Test set) เก็บรวบรวมข้อความข้อความจากภาพ และโทนสีภาพ จากผู้ใช้งาน Twitter ที่เป็นดาราและผู้ที่มีชื่อเสียง จำนวน 47 คน แบ่งออกเป็น 2 กลุ่ม คือ กลุ่มที่หนึ่ง ผู้ป่วยโรคซึมเศร้า จำนวน 27 คน และกลุ่มที่สอง บุคคลที่ไม่เป็นโรค

ซึมเศร้า จำนวน 20 คน ประกอบไปด้วยภาพ จำนวน 30,341 ภาพ ข้อความ จำนวน 31,916 ข้อความ

1.4.3 กระบวนการปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น

1) การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูล โดยวิธีการเอ็นเซมเบิล ตัวจำแนกประเภทที่ใช้สำหรับสร้างการเรียนรู้แบบจำลอง (Training model) ประกอบด้วย นาอีฟเบย์ ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด และแรนดอมฟอเรสต์ เพื่อให้ค่าน้ำหนักแก่ตัวจำแนกประเภทของวิธีการเอ็นเซมเบิล ด้วยวิธีการอัตราการทำนายถูกของคลาสคำตอบ (True positive weighted ensemble: TPW) และค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ (Average probability weighted ensemble: APW) เพื่อปรับปรุงประสิทธิภาพการพยากรณ์โรคซึมเศร้าในวัยรุ่น

2) การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าในวัยรุ่นด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ จากคำถามการวิจัยที่ 1: ตัวจำแนกประเภทที่ใช้สำหรับสร้างการเรียนรู้แบบจำลอง ประกอบด้วย ได้แก่ นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด กราเดียนบูทติ้งทรี และตัวแบบเชิงเส้นนัยทั่วไป

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้กระบวนการที่มีประสิทธิภาพสำหรับการจำแนกโรคซึมเศร้าจากความคิดเห็นและภาพ

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การวิจัยเรื่อง “การปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น” ผู้วิจัยได้ดำเนินการศึกษาเอกสารแนวคิดและทฤษฎีที่เกี่ยวข้องกับโรคซึมเศร้า กระบวนการทำเหมืองข้อมูล ตัวจำแนกประเภท ทฤษฎีภาพ วิธีการประเมินประสิทธิภาพของแบบจำลอง และ Paired–sample t-test ดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 โรคซึมเศร้า

โรคซึมเศร้า เป็นโรคทางจิตเวชประเภทหนึ่งที่สามารถเกิดขึ้นได้กับทุกคนทุกเพศทุกวัย ผู้ป่วยจะมีภาวะที่จิตใจแสดงถึงความผิดปกติทางอารมณ์ด้านลบเป็นหลัก ได้แก่ เบื่อหน่าย เศร้าหม่นหมอง หดหู่ ท้อแท้ง่าย สมาธิสั้น หมดหวัง วิตกกังวล รู้สึกว่าตนเองไม่มีค่าและมองแต่ด้านลบ ส่งผลในการใช้ชีวิตประจำวันอย่างยากลำบาก หากไม่ได้รับการรักษา อาจทำร้ายตนเอง หรืออาจร้ายแรงจนถึงการจบชีวิตด้วยการฆ่าตัวตาย สาเหตุของการเกิดโรคซึมเศร้าเกิดได้ทั้งปัจจัยภายในร่างกายและภายนอกร่างกาย แบ่งออกเป็น 3 ปัจจัย ได้แก่ ปัจจัยด้านชีววิทยา (Biological factors) ปัจจัยทางด้านจิตวิทยา (Psychological factors) และปัจจัยทางสังคม (Social factors) อาการระหว่างโรคซึมเศร้ากับความเครียดต่างกันตรงความรู้สึก โดยความเครียดจะมีอาการฟ้องทางด้านร่างกายออกมาให้คนรอบข้างเห็นถึงความผิดปกติ จนสามารถสังเกตได้ และสักระยะหนึ่งร่างกายและจิตใจจะปรับสมดุลให้กลับมาร่าเริงตามปกติได้ แต่คนที่มีความเสี่ยงที่จะเป็นโรคซึมเศร้าจะรู้สึกดูถูกตนเองอย่างรุนแรง จนเกิดเป็นความเครียดที่เรื้อรังจนส่งผลกระทบต่อระบบสมองและทำให้ร่างกายเกิดอาการรวน ทำให้การกิน การนอนผิดปกติจนลุกลามไปสู่ความเสื่อมทั้งระบบร่างกาย (ThaiHealth watch 2020, 2020) หากอาการซึมเศร้าเกิดขึ้นเป็นเวลานาน ๆ โดยไม่มีอาการที่จะดีขึ้นอาจจะทำให้มีอาการอื่นๆ ตามมา เช่น นอนไม่หลับ เบื่ออาหาร น้ำหนักลด รู้สึกไม่อยากมีชีวิตอยู่ เป็นต้น โดยลักษณะอาการโรคซึมเศร้าแบ่งออกเป็น 4 กลุ่ม ได้แก่ 1) อาการทางอารมณ์ (Affective symptoms) 2) อาการทางกระบวนการคิด (Cognitive symptoms) 3) อาการทางกาย (Somatic symptoms) และ 4) อาการด้านพฤติกรรม (Behavior symptoms) ซึ่งผู้ป่วยอาจมีทั้งอาการที่แสดงออกอย่างชัดเจนและไม่ชัดเจน ทำให้การวินิจฉัยโรคซึมเศร้าจำเป็นที่ต้องใช้เกณฑ์เพื่อช่วยในการคัดกรองผู้ที่เข้าข่ายเป็นโรคซึมเศร้าและการแยกระดับผู้เป็นโรคซึมเศร้า ซึ่งในการวินิจฉัยโรคซึมเศร้าสมาคมจิตแพทย์อเมริกันได้กำหนดหลักเกณฑ์ในการวินิจฉัยโรคผู้ที่เป็นโรคซึมเศร้าจะต้องมีอาการอย่างน้อย 5 อาการหรือมากกว่า ซึ่งจะมีอาการติดต่อกันนานไม่ต่ำกว่า 2 สัปดาห์ คือ 1) มี

อารมณ์ซึมเศร้า 2) ความสนใจหรือความสนุกสนานในการทำกิจกรรมที่เคยทำลดลงอย่างมาก 3) น้ำหนักลดลงอย่างชัดเจน 4) มีอาการนอนไม่หลับหรือหลับนานหลับบ่อยกว่าปกติ 5) การเคลื่อนไหวช้าลง 6) อ่อนเพลียไม่มีเรี่ยวแรง 7) รู้สึกว่าตนเองไร้ค่าหรือรู้สึกว่าตนเองผิดโดยไม่มีสาเหตุ 8) สมาธิสั้นหรือความสามารถในตัดสินใจลดลง 9) มีความคิดอยากฆ่าตัวตาย โดยกำหนดไว้ในคู่มือการวินิจฉัยและสถิติสำหรับความผิดปกติทางจิตฉบับที่ 5 (The 5th Diagnostic and statistical manual of mental disorders : DSM-5) ของสมาคมจิตเวชศาสตร์สหรัฐอเมริกา (American Psychiatric Association, 2013) ซึ่งทำเป็นแบบสอบถามไว้ ดังตารางที่ 1

ตารางที่ 1 คู่มือการวินิจฉัยและสถิติสำหรับความผิดปกติทางจิตฉบับที่ 5 ของสมาคมจิตเวชศาสตร์สหรัฐอเมริกา

No	Symptoms	None	Someday	Frequently	Everyday
1	อารมณ์ซึมเศร้า	0	1	2	3
2	ความสนใจลดลง	0	1	2	3
3	น้ำหนักผิดปกติ	0	1	2	3
4	การนอนผิดปกติ	0	1	2	3
5	สมาธิสั้น	0	1	2	3
6	รู้สึกไร้ค่า	0	1	2	3
7	ร่างกายอ่อนเพลีย	0	1	2	3
8	กระวนกระวาย หรือ เชื่องช้า	0	1	2	3
9	การอยากฆ่าตัวตาย	0	1	2	3

จากตารางที่ 1 ในการวิเคราะห์ผู้ที่ทำแบบสอบถามที่มีอาการในระยะเวลา 2 สัปดาห์ มีการกำหนดคะแนนตามความถี่ของอาการ โดยที่ 1) Someday หมายถึง มีอาการในระหว่าง 2-4 วัน คะแนนเท่ากับ 1 คะแนน 2) Frequently หมายถึง มีอาการในระหว่าง 6-8 วัน คะแนนเท่ากับ 2 คะแนน และ 3) Everyday หมายถึง มีอาการในระหว่าง 10-14 วัน คะแนนเท่ากับ 3 คะแนน จากนั้นนำคะแนนทั้ง 9 อาการ มาหาผลรวมของคะแนนเพื่อในแบบประเมินของการวินิจฉัยโรคซึมเศร้า หากผลคะแนนประเมิน แบ่งออกเป็น 4 ระดับ

ระดับที่ 1 คะแนนรวม น้อยกว่า 7 คะแนน หมายถึง ปกติ

ระดับที่ 2 คะแนนรวม อยู่ระหว่าง 8-12 คะแนน หมายถึง มีอาการโรคซึมเศร้าน้อย

ระดับที่ 3 คะแนนรวม อยู่ระหว่าง 13-18 คะแนน หมายถึง มีอาการโรคซึมเศร้าปานกลาง

ระดับที่ 4 คะแนนรวม มากกว่า 19 คะแนน หมายถึง มีอาการโรคซึมเศร้ามีอาการรุนแรง

2.1.2 กระบวนการทำเหมืองข้อมูล

การทำเหมืองข้อมูล (Data mining) คือ การวิเคราะห์ข้อมูล เพื่อหารูปแบบ (Patterns) หรือความสัมพันธ์ (Relation) ระหว่างข้อมูลในฐานข้อมูลขนาดใหญ่ (Linoff & Berry, 2011) ซึ่งสอดคล้องกับ Hand (2001) ที่ได้ให้ความเห็นว่า เหมืองข้อมูล คือ การวิเคราะห์ข้อมูลจำนวนมาก เพื่อหาความสัมพันธ์และการสรุปผลข้อมูล ซึ่งสามารถเข้าใจและเป็นประโยชน์ต่อผู้กระทำการเก็บรวบรวมข้อมูล ในปัจจุบันเหมืองข้อความ (Text mining) ได้รับความนิยมและนำมาประยุกต์ใช้ในด้านการค้นหาข้อมูลที่แฝงในชุดข้อความจำนวนมหาศาล (Jimenez-Marquez et al., 2019) ในนำกระบวนการจำแนกข้อความ (Text classification) ได้แก่ การเลือกข้อมูล การเตรียมข้อมูล การสร้างแบบจำลอง และการประเมินผล อธิบายรายละเอียด ดังนี้

1) การเลือกข้อมูล (Data selection) เป็นขั้นตอนการระบุเกี่ยวกับแหล่งข้อมูลที่จะนำมาใช้งานในการทำเหมืองข้อความ จากนั้นนำมาพิจารณาว่าข้อมูลว่าสอดคล้องกับขอบเขตที่ต้องการหรือไม่ รวมทั้งประโยชน์ที่จะได้รับจากการนำมาใช้งาน เพื่อทำงานวางแผนการดำเนินการในขั้นตอนต่อไป

2) การเตรียมข้อมูล (Data preprocessing) เป็นขั้นตอนการทำข้อมูลให้อยู่ในรูปแบบที่ใช้สามารถนำไปใช้งานได้ เนื่องจากข้อมูลส่วนใหญ่ไม่ได้ถูกจัดเก็บในลักษณะที่แตกต่างกัน ดังนั้นจึงจำเป็นต้อง การกรองข้อมูล การแปลงข้อมูลหรือเปลี่ยนชนิดข้อมูล เพื่อให้ข้อมูลอยู่ในลักษณะหรือรูปแบบที่ง่ายต่อการนำไปประมวลผล และการนำข้อมูลที่ไม่ถูกต้องออกไป ในการนำข้อมูลที่อยู่ในรูปแบบข้อความ (Text) ที่นำมาจากโซเชียลมีเดียเพื่อนำมาประมวลผลข้อความ (Text preprocessing) มีขั้นตอนในการทำงาน (Jo, 2019) ดังนี้

ขั้นตอนที่ 1 การตัดคำ (Tokenization) เป็นขั้นตอนการตัดคำแต่ละคำออกจากรูปแบบประโยค การลบคำหยุด (Remove stop words) จากนั้นนำคำเหล่านั้นไปสร้างถุงคำ (Bag of word: BOW) ที่ใช้สำหรับการนับความถี่ที่อยู่ในเอกสาร

ขั้นตอนที่ 2 การกำจัดคำหยุด (Stop word removal) เป็นขั้นตอนลบคำที่สิ้นเปลืองออก ได้แก่ คำสรรพนาม คำนาหน้า คำเชื่อม คำบุพบท เพื่อเพิ่มประสิทธิภาพให้แบบจำลองและลดเวลาการประมวลผล

ขั้นตอนที่ 3 การหารากคำศัพท์ (Stemming) เป็นการจัดกลุ่มคำศัพท์ที่มีความหมายเหมือนกันไว้ด้วยกัน เพื่อลดจำนวนคำ ทำให้จำนวนคุณลักษณะลดลง ลดเวลาในการประมวลผลได้

ขั้นตอนที่ 4 การให้น้ำหนักคำ (Weighting) เป็นขั้นตอนการให้น้ำหนักคำของแต่ละคำในถุงคำ ซึ่งการให้น้ำหนักจะต่างไปตามวิธีการที่ใช้ในการให้น้ำหนัก โดยจะขึ้นอยู่กับประเภทของ

เอกสารนั้น การสกัดคุณลักษณะ (Feature extraction) เป็นขั้นตอนการแปลงข้อมูลหรือการดึงคุณลักษณะเพื่อให้ได้คุณลักษณะที่สามารถเป็นตัวแทนเพื่อจะนำไปใช้งานได้ โดยจะมีการแทนค่าในรูปแบบเวกเตอร์ มีหลายวิธีการ เช่น Binary term occurrences, Term frequencies, Term occurrences และ Term frequency-inverse document frequency (TF-IDF) (Kompan & Bieliková, 2011) ในงานวิจัยนี้ใช้วิธีการ Binary term occurrences หรือ Boolean weighting เป็นวิธีการแทนคำที่เกิดขึ้นหรือไม่เกิดขึ้นในถุงคำ หากพบคำในถุงคำจะแทนค่าเท่ากับ 1 แต่หากไม่พบหรือไม่มีคำนั้นในถุงคำจะแทนค่าเท่ากับ 0 (Atorn, 2006) ดังสมการ (1)

$$\text{Binary term} = \begin{cases} 1, & \text{for term present in document} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

ขั้นตอนที่ 5 การคัดเลือกคุณลักษณะ (Features selection) เป็นขั้นตอนการเลือกคำที่มีแก่การนำไปสร้างแบบจำลอง ซึ่งการคัดเลือกคุณลักษณะช่วยให้สามารถลดเวลาในการสร้างแบบจำลองและสามารถเพิ่มค่าความถูกต้องได้

3) การสร้างแบบจำลอง (Modeling) การนำข้อมูลที่ผ่านกระบวนการการเตรียมข้อมูลมาสร้างแบบจำลองเพื่อดึงเอาความรู้สำคัญออกมาจากข้อมูล โดยจะการเลือกวิธีการที่เหมาะสมกับข้อมูล

4) การประเมินผล (Evaluation) หลังจากได้แบบจำลองแล้ว ต้องทำการประเมินผลลัพธ์ที่ได้ว่าแบบจำลองนั้นมีความถูกต้องแม่นยำมากน้อยเพียงใด โดยการทดลองและประเมินผลสามารถนำไปประมวลผลกับข้อมูลจริงที่มีอยู่เพื่อเปรียบเทียบผลของการวิเคราะห์ว่าถูกต้องเพียงใดยอมรับกับผลที่ได้หรือไม่

2.1.3 ทฤษฎีที่เกี่ยวข้องกับการคัดเลือกคุณลักษณะ

คุณลักษณะข้อมูล คือ ลักษณะหรือคุณสมบัติที่ใช้เพื่อระบุองค์ประกอบ รายละเอียดที่มีในชุดข้อมูล โดยคุณลักษณะข้อมูลที่ดีจะต้องมีความถูกต้องและเชื่อถือได้ เก็บเฉพาะข้อมูลที่จำเป็นต้องใช้ และมีความครบถ้วน สมบูรณ์ หากมีการเก็บคุณลักษณะที่มากเกินความจำเป็นจะทำให้สิ้นเปลืองทรัพยากรในการจัดเก็บ เวลาประมวลผล รวมทั้งการนำข้อมูลไปใช้งานอาจทำให้ประสิทธิภาพลดลงได้

การคัดเลือกคุณลักษณะ (Feature selection) เป็นกระบวนการลดขนาดข้อมูลด้วยการทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุด เนื่องจากข้อมูลแต่ละตัวอาจมีความสำคัญไม่เท่ากัน ดังนั้นวิธีการการเลือกคุณลักษณะข้อมูลที่ดีจะทำให้สามารถเลือกคุณลักษณะข้อมูลที่มีความสำคัญ และสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่และทำให้ได้ประสิทธิภาพที่ดี (Dash & Liu, 1997) วัตถุประสงค์การคัดเลือกคุณลักษณะข้อมูลเพื่อปรับปรุง

ประสิทธิภาพ และเพื่อเตรียมข้อมูลที่ประมวลผลได้รวดเร็ว มีประสิทธิภาพที่ดียิ่งขึ้น ได้แบ่งออกเป็น 3 วิธี คือ การเลือกคุณลักษณะแบบควบรวม การเลือกคุณลักษณะแบบกรอง และการเลือกคุณลักษณะแบบวิธีฝังตัว (Jović et al., 2015; Zhao et al., 2010; Zheng & Casari, 2018) รายละเอียด ดังนี้

1) การเลือกคุณลักษณะแบบควบรวม (Wrapper method) เป็นวิธีการอย่างง่ายอาศัยวิธีการจำแนกประเภทในการวัดความสำคัญของเซตย่อยของตัวแปร โดยเลือกเซตย่อยที่มีความแม่นยำในการจำแนกประเภทข้อมูลสูงหรือใช้ความแม่นยำในการจำแนกประเภทข้อมูลเมื่อใช้เซตย่อยนั้น เป็นดัชนีวัดความสำคัญของเซตย่อย ซึ่งเซตใหม่ที่ให้ผลลัพธ์ที่แม่นยำมากที่สุดจะถูกเลือกใช้ วิธีการคัดเลือกคุณลักษณะนี้ได้ความแม่นยำสูงแต่อาจจะใช้เวลานาน ดังนั้นจึงเหมาะสมที่จะนำไปใช้งานกับข้อมูลขนาดเล็กและขนาดกลาง (Dash & Liu, 1997) เช่น Forward selection, Backward elimination และ Evolutionary selection เป็นต้น

2) การเลือกคุณลักษณะแบบกรอง (Filter method) เป็นวิธีการคัดเลือกคุณลักษณะที่ทำงานได้รวดเร็ว ง่ายต่อการตีความ โดยจะจำกัดคุณลักษณะที่ไม่เกี่ยวข้องในการจำแนกประเภทด้วยคุณสมบัติในเนื้อแท้ของข้อมูล ใช้วิธีการวัดคะแนนและการจัดเรียงลำดับตามคะแนน การเลือกคุณลักษณะแบบกรองสามารถใช้ฟังก์ชันการประเมินค่าด้วยมาตรวัดระยะทาง มาตรวัดสารสนเทศ มาตรวัดความไม่เป็นอิสระ เช่น Correlation based feature selection, Information gain, Gain ratio และ Chi-square (Remeseiro & Bolon-Canedo, 2019) เป็นต้น

3) การเลือกคุณลักษณะแบบวิธีฝังตัว (Embedded method) เป็นวิธีการคัดเลือกคุณลักษณะที่เป็นส่วนหนึ่งในกระบวนการจำแนกประเภท ถูกออกแบบมาเพื่อแก้ไขปัญาข้อเสียของวิธีแบบควบรวม และแบบกรอง ซึ่งใช้เวลาในการประมวลน้อยกว่าวิธีแบบควบรวม โดยจะรวมการเลือกคุณลักษณะให้เป็นส่วนหนึ่งของกระบวนการการเรียนรู้ มีการค้นหาเซตของคุณลักษณะทั้งแบบ Global space และ Local space ทำให้การค้นหามีประสิทธิภาพดีขึ้น แต่ในการเลือกเซตคุณลักษณะไม่มีความยืดหยุ่นเพราะจะขึ้นอยู่กับวิธีการการจำแนกข้อมูล (Zheng & Casari, 2018) เช่น Least absolute shrinkage and selection operator (LASSO) และ Decision trees (Jović et al., 2015) เป็นต้น

การคัดเลือกคุณลักษณะแบบกรองที่ได้ความนิยมในการนำไปใช้ในงานวิจัย เนื่องจากง่ายต่อการตีความ รวดเร็ว ใช้เวลาไม่นาน มีความถูกต้อง ซึ่งเหมาะสมกับชุดข้อมูลที่มีคุณลักษณะจำนวนมากในการจำแนกข้อความ (Text classification) และในงานวิจัยนี้ได้ใช้การคัดเลือกคุณลักษณะแบบกรองด้วยวิธีการ Information gain (Doenribram, 2019; Ong et al., 2015; Wu & Xu, 2016) เป็นวิธีการที่พิจารณาจากค่าความน่าจะเป็นของแต่ละคุณลักษณะที่เป็นไปได้แล้ววัดค่าของเอนโทรปี (Entropy) เพื่อเลือกคุณลักษณะที่สำคัญในการระบุหรือแบ่งข้อมูลเป็นชุดข้อมูลย่อย

ด้วยการคำนวณหาค่า Gain ในแต่ละมิติของข้อมูล หากมิติข้อมูลใดมีค่า Gain สูงสุด มิติข้อมูลนั้นจะถูกเลือกไปเพื่อใช้ในการระบุชุดข้อมูลย่อย เริ่มจากการคำนวณหาค่า *Info(D)* ที่เป็นค่าเอนโทรปีของ *D* ดังสมการ (2) คำนวณค่าเอนโทรปีของชุดข้อมูลทั้งหมด จากนั้นคำนวณหา$Info_A(D)$ *ที่*จะใช้สำหรับแบ่งชุดข้อมูล *D* เป็นชุดข้อมูลย่อย ดังสมการ (3) คำนวณค่าเอนโทรปีชุดมิติข้อมูลแต่ละคุณลักษณะและคำนวนหาค่า *Gain(A)* ดังสมการ (4) คำนวณหาค่า IG สำหรับการพิจารณามิติข้อมูลของคุณลักษณะ *A* (Han et al., 2012) หลังจากคำนวณหาค่า Info(D*)* และ $Info_A(D)$ นำค่าที่คำนวณได้สำหรับการพิจารณาคุณลักษณะของ A

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2 P_i \tag{2}$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info_{D_j} \tag{3}$$

$$Gain(A) = Info(D) - Info_A(D) \tag{4}$$

โดยที่

p_i คือ ค่าความน่าจะเป็นที่เรคคอร์ดหนึ่ง ๆ จะมีหมวดหมู่ของข้อมูล

$\frac{|D_j|}{|D|}$ คือ จำนวนสมาชิกในชุดข้อมูล *D* ที่มีค่าในคุณลักษณะ A เป็น D_j หารด้วยจำนวนสมาชิกทั้งหมดในชุดข้อมูล *D*

2.1.4 ตัวจำแนกประเภท

ตัวจำแนกประเภท (Classifier) เป็นวิธีการที่ใช้ในการจำแนกหมวดหมู่ข้อมูลโดยการเรียนรู้จากคุณลักษณะ จากนั้นจะทำการสร้างโมเดลที่สามารถจำแนกข้อมูลชุดใหม่ที่นำมาทำนายได้ ตัวจำแนกประเภทที่ได้รับความนิยมในการจำแนกประเภท ในงานวิจัยนี้ใช้ตัวจำแนกประเภทจำนวน 8 วิธีการ คือ นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ตัวแบบเชิงเส้นนัยทั่วไป เอ็นเซมเบิล กราเดียนบูทติ้งทรี และแรนดอมฟอเรสต์ อธิบายหลักการ ดังนี้

1) วิธีการนาอีฟเบย์เป็นการเรียนรู้แบบมีผู้สอน (Supervised learning) ใช้สำหรับการจำแนกประเภทด้วยหลักการคำนวณความน่าจะเป็น (Probability) เพื่ออนุมานคำตอบที่ต้องการเรียกทฤษฎีนี้ว่า ทฤษฎีเบย์ (Bayes' Theorem) วิธีการนี้ไม่ได้มีความซับซ้อนมากนัก ทำการเรียนรู้ปัญหาที่เกิดขึ้นเพื่อนำไปสร้างเงื่อนไขในการจำแนกข้อมูลใหม่ โดยตั้งสมมติฐานว่าปริมาณความสนใจจะขึ้นอยู่กับการกระจายความน่าจะเป็น (Conditional probability) (Hand et al., 2001) ดังสมการ (5) และ สมการ (6) วิธีการ Naive bayes เหมาะสมในการจำแนกประเภทในกลุ่มตัวอย่างที่

มีจำนวนมากและคุณลักษณะที่ไม่ขึ้นต่อกัน ซึ่งได้รับความนิยมนำไปประยุกต์ใช้งานด้านการจำแนกประเภทข้อความ (Text classification) (Ignatow & Mihalcea, 2018) ความน่าจะเป็นแบบมีเงื่อนไข เป็นความน่าจะเป็นของการเกิดเหตุการณ์ A เมื่อกำหนดว่าเหตุการณ์ B เกิดขึ้นแล้ว แทนด้วยสัญลักษณ์ P (A | B) สามารถคำนวณจากความน่าจะเป็นร่วมกัน แทนด้วย $P(A \wedge B)$ ความน่าจะเป็นที่เหตุการณ์ A และเหตุการณ์ B เกิดขึ้นร่วมกัน

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)} \tag{5}$$

$$P(A \mid B) = \frac{P(B|A)P(A)}{P(B)} \tag{6}$$

โดยที่

P (A|B) คือ ความน่าจะเป็นของ A เมื่อ B เกิดขึ้นแล้ว

P (B|A) คือ ความน่าจะเป็นของ B เมื่อ A เกิดขึ้นแล้ว

P (A) คือ ความน่าจะเป็นที่จะเกิดหน้าเหตุการณ์ A

P (B) คือ ความน่าจะเป็นที่จะเกิดหน้าเหตุการณ์ B

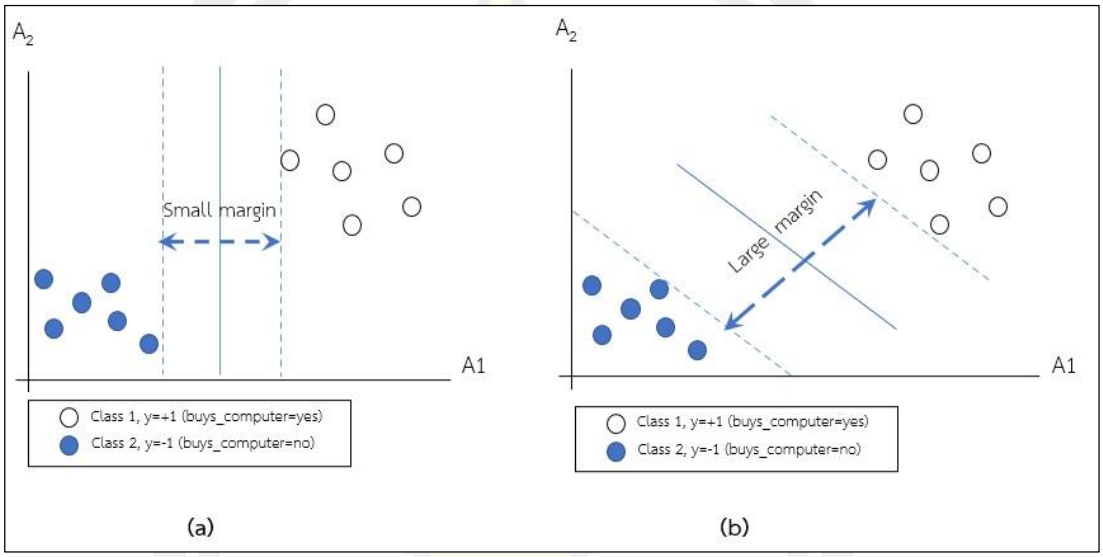
การอธิบายสมการ (5) เมื่อมีเหตุการณ์ฝนตก จะสามารถไปเล่นเทสนิสได้ ดังนี้

$$P(\text{ไปเล่นเทสนิส} \mid \text{ฝนตก}) = \frac{P(\text{ฝนตก} \mid \text{ไปเล่นเทนนิส}) \times P(\text{ไปเล่นเทนนิส})}{P(\text{ฝนตก})}$$

จากสมการตัวอย่างข้างต้นสามารถทำนายได้ว่า การไปเล่นเทนนิสโดยให้สังเกตฝนตก อย่างไรก็ตามเหตุการณ์ที่ทำนายนั้นต้องสอดคล้องกัน เช่น หากต้องการทำนายการไปเล่นเทนนิส จะต้องไม่ใช่เหตุการณ์น้ำท่วมสนามเข้ามาเกี่ยวข้อง เพราะเหตุการณ์ทั้งสองไม่มีความสอดคล้องกัน

2) วิธีการซัพพอร์ตเวกเตอร์แมชชีนเป็นการเรียนรู้แบบมีการสอนใช้สำหรับการจำแนกประเภทข้อมูล ซึ่งมีพื้นฐานมาจากการจำแนกเชิงเส้นตรงที่ได้รับความนิยมอย่างแพร่หลายในการนำไปแก้ปัญหาการจำแนกประเภท (Cortes & Vapnik, 1995) โดยอาศัยหลักการทางสถิติกับการหาค่าที่เหมาะสม (Optimization theory) เพื่อหาสัมประสิทธิ์ของสมการเส้นตรงเพื่อสร้างเส้นแบ่งกลุ่มให้ข้อมูล การหาเส้นตรงที่มีขนาดมาร์จินที่โตสุด (Maximal margin) ในการจำแนกคุณลักษณะข้อมูล ในขณะเดียวกันจะใช้ฟังก์ชันเคอร์เนลแปลงข้อมูลไปยังมิติที่สูงขึ้น ในปริภูมิคุณลักษณะ (Feature space) ก่อนทำการจำแนกประเภท สามารถทำงานได้ในรูปแบบ Linear, Non-linear, Regression และ Outlier detection เหมาะในการจำแนกข้อมูลที่มีความซับซ้อน ข้อมูลขนาดเล็ก และข้อมูลขนาดกลาง กระบวนการทำงานของวิธีการซัพพอร์ตเวกเตอร์แมชชีนจะใช้มาร์จินโตสุดเพื่อเพิ่ม

ประสิทธิภาพในการจำแนกด้วยเส้นตรงและการแปลงข้อมูลด้วยฟังก์ชันเคอร์เนลจากปริภูมิ ที่ยังไม่เหมาะสมในการจำแนกเชิงเส้นตรงไปยังมิติที่สูงขึ้นไปปริภูมิคุณลักษณะ แล้วค่อยทำการจำแนกข้อมูลในปริภูมิคุณลักษณะ จึงทำให้วิธีการซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนกประเภทที่มีประสิทธิภาพสูงในภาพรวม เนื่องจากเป็นตัวจำแนกเชิงเส้นตรง ทำให้สามารถจำแนกข้อมูลแบบสองกลุ่ม และข้อมูลแบบหลายกลุ่ม (Multiclass classification) ได้ดี (Wongthanavasu, 2017)



ภาพที่ 1 แสดง Hyperplanes ระหว่าง (a) ขอบขนาดเล็ก และ (b) ขอบขนาดใหญ่
(Han et al., 2012)

จากภาพประกอบที่ 1 ได้แสดงถึงระยะของ (a) ขอบขนาดเล็ก และ (b) ขอบขนาดใหญ่ และลักษณะของเส้นตรงที่แบ่งกลุ่มข้อมูลเพื่อจำแนกข้อมูลคลาส 1 และคลาส 2 ซึ่งวิธีการหาระยะขอบ ดังสมการ (7) (8) และ (9)

$$wx + b = 0 \quad (7)$$

$$w_0 + w_1x_1 + w_2x_2 > 0 \quad (8)$$

$$w_0 + w_1x_1 + w_2x_2 < 0 \quad (9)$$

โดยที่

w คือ ค่าน้ำหนักของเวกเตอร์ $w = \{w_1, w_2, w_3, ..., w_n\}$

x คือ Training tuples $x = (x_1, x_2)$

3) วิธีการต้นไม้ตัดสินใจเป็นการเรียนรู้แบบมีการสอนใช้ในการจำแนกประเภทข้อมูล มีพื้นฐานจาก Hunt's algorithm โดยสร้างแบบจำลองที่ใช้ทำนายเหตุการณ์ล่วงหน้าทำให้ได้ผลลัพธ์ที่มีความแม่นยำ ไม่ซับซ้อนและเป็นอีกวิธีการหนึ่งได้รับความนิยมในนำไปใช้งาน เนื่องจากใช้เข้าใจง่าย

เพราะมีโครงสร้างที่คล้ายต้นไม้ประกอบด้วยโหนดราก (Root node) ที่มีการแตกกิ่งก้าน (Branch) แทนคุณลักษณะของข้อมูลเชื่อมโยงต่อกันไปโหนดลูกเป็นตัวทดสอบกิ่งของต้นไม้ที่แสดงค่าที่เป็นไปได้ของคุณลักษณะ และโหนดใบ (Leaf node) แทนคลาสหรือผลลัพธ์การทำนาย ดังนั้นประสิทธิภาพของแบบจำลองจึงขึ้นอยู่กับคุณลักษณะข้อมูลที่นำมาใช้จำแนกประเภท ซึ่งค่าของคุณลักษณะเป็นค่าที่ไม่ต่อเนื่อง (Discrete value) ในการสร้างต้นไม้โดยคัดเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นโหนดราก โดยใช้ค่า Gain ratio ที่สูงที่สุดเป็นโหนดราก แต่ก่อนที่จะได้ค่า Gain ratio ต้องหาค่า Split information ค่า Information gain และค่า Entropy (Han et al., 2011) สมการ Entropy เป็นสมการที่ใช้ในการหาค่าสารสนเทศของข้อมูล (Entropy measure) ดังสมการ (10)

$$\text{Entropy(s)} = -\sum_{i=1}^{c} P_i \log_2 P_i \tag{10}$$

โดยที่

s คือ คุณลักษณะ ที่นำมาวัดค่า Entropy

P_i คือ สัดส่วนของจำนวนสมาชิกในกลุ่ม i เท่ากับจำนวนสมาชิกทั้งหมดของกลุ่มตัวอย่าง

สมการ Information gain เป็นสมการที่ใช้ในการหาค่าสารสนเทศเกน ก่อนนำไปใช้ในการหาค่ามาตรฐานอัตราส่วนเกน ดังสมการ (11)

$$\text{Gain}(S,A) = \text{Entropy}(s) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{11}$$

โดยที่

A คือ คุณลักษณะ A

$|S_v|$ คือ สมาชิกของ คุณลักษณะ A ที่มีค่า v

$|S|$ คือ จำนวนสมาชิกของกลุ่มตัวอย่าง

สมการ Split information ใช้ในการหาค่าสารสนเทศของการแบ่ง ดังสมการ (12)

$$\text{Split information}(S,A) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|} \tag{12}$$

โดยที่

S_i คือ สัดส่วนของจำนวนสมาชิกในกลุ่ม i

สมการ Gain ratio เป็นสมการที่เพิ่มขึ้นจากอัลกอริทึม ID3 เพื่อลดความลำเอียงของข้อมูล ดังสมการ (13)

$$\text{Gain ratio}(S,A) = \frac{Gain(S,A)}{Split\ Information(S,A)} \tag{13}$$

4) ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดเป็นการเรียนรู้แบบมีการสอน ซึ่งเป็นการจำแนกประเภทที่มีหลักการไม่ยุ่งยาก โดยมีจะพิจารณาจากระยะห่างที่น้อยที่สุดระหว่างข้อมูลที่ต้องการจำแนกกับชุดการเรียนรู้แบบจำลอง โดยที่ K คือ ค่าที่เป็นจำนวนที่ต้องการเลือกเป็นเพื่อนบ้านที่ใกล้ที่สุด เพื่อไม่ให้เกิดการจำนวนเพื่อนบ้านเท่ากันควรกำหนดค่า K เป็นจำนวนคี่ เช่น กลุ่มข้อมูลตามจำนวน K=3, K=5 และ K=9 การหาระยะทางเพื่อนบ้านที่ใกล้ที่สุด ซึ่งหาระยะทางสามารถทำได้หลายวิธีการ (Kuhn & Johnson, 2013) เช่น ระยะห่างแบบยูคลิด (Euclidean distance) ระยะห่างแบบมินโควสกี (Minkowski distance) ความเหมือนแบบโควไซน์ (Cosine similarity) ระยะห่างแบบแมนฮัตตัน (Manhattan distance) ความเหมือนแบบแจคการ์ด (Jaccard similarity) เป็นต้น การหาระยะห่างระหว่างจุดแบบยูคลิดเป็นวิธีการที่ง่ายและได้รับความนิยม ในการนับไปใช้หาระยะทางเพื่อนบ้านที่ใกล้ที่สุด ดังสมการ (14)

$$D = \sqrt{\sum_{i=1}^{n} \left(p_i - q_i\right)^2} \tag{14}$$

โดยที่

p คือ ค่าของชุดข้อมูลที่ต้องการจำแนก

q คือ ค่าของชุดข้อมูลเพื่อนบ้านที่นำมาพิจารณา ชุดข้อมูลสำหรับการเรียนรู้แบบจำลอง

5) ตัวแบบเชิงเส้นนัยทั่วไปเป็นวิธีการที่มีการปรับปรุงหรือเป็นส่วนขยายของตัวแบบเชิงเส้นทั่วไป (General linear model) ที่ใช้วิธีหลักการทางสถิติเพื่อหาความสัมพันธ์ระหว่างตัวแปรต้นและตัวแปรตาม โดยสามารถวิเคราะห์ได้ทั้งข้อมูลตัวเลขและข้อมูลไม่ใช่ตัวเลข มีประโยชน์สำหรับการแจกแจงแบบไม่ปกติและไม่ต่อเนื่อง การคำนวณเป็นแบบขนานกัน ประมวลผลเร็ว และปรับขยายได้ดี สามารถใช้กับชุดข้อมูลใหม่ (Unseen data) เพื่อระบุคลาสคำตอบ (Ebrahimi et al., 2019)

6) วิธีการเอ็นเซมเบิลเป็นวิธีการหนึ่งที่ใช้ในการจำแนกประเภทข้อมูลโดยนำตัวจำแนก

หลายๆ ตัวมาผสมผสานเข้าด้วยกันเพื่อการทำนายคำตอบสุดท้าย ถือว่าเป็นวิธีการที่มีประสิทธิภาพ มักนำมาใช้เพื่อเพิ่มประสิทธิภาพการทำงานให้แบบจำลอง เนื่องจากเป็นการรวมผลการทำนายคำตอบของแต่ละตัวจำแนกประเภทแบบเดี่ยวที่เกิดจากการเรียนรู้ที่หลากหลายเข้าด้วยกัน (Swamynathan, 2017) วิธีการเอ็นเซมเบิลแบ่งออกเป็น 2 แบบ คือ แบบที่ 1 ตัวจำแนกที่ทำงานร่วมกันและใช้ตัวจำแนกเหมือนกัน (Homogeneous ensemble) เช่น แรนดอมฟอเรสต์ แบบที่ 2 ตัวจำแนกที่ทำงานร่วมกันแต่ใช้ตัวจำแนกที่แตกต่างกัน (Heterogeneous ensemble) (Rojarath, 2020) ในงานวิจัยนี้ใช้ตัวจำแนกเอ็นเซมเบิล รายละเอียด ดังนี้

วิธีการที่ 1 วิธีการคะแนนเสียงข้างมากเป็นวิธีการจำแนกประเภทข้อมูลหนึ่งของวิธีการเอ็นเซมเบิล โดยการนำข้อมูลสำหรับการเรียนรู้แบบจำลองชุดเดียวกันไปใช้ในแบบจำลองที่สร้างจากวิธีการที่แตกต่างกัน จากนั้นนำข้อมูลสำหรับการทดสอบชุดเดียวกันมาเข้าแบบจำลอง เพื่อทำนายผลลัพธ์ออกมาและใช้การคะแนนเสียงข้างมากเป็นคำตอบสุดท้าย โดยทั่วไปแล้วการวิธีการคะแนนเสียงข้างมาก แบ่งออกเป็น 2 วิธีการ ได้แก่ วิธีการแบบไม่กำหนดค่าน้ำหนักและวิธีการแบบกำหนดค่า (Onan et al., 2016)

Unweighted voting schemes เป็นการโหวตให้คะแนนแบบง่ายและนำไปใช้งานอย่างแพร่หลาย ใน LC*i (x)* คือ ค่าความน่าจะเป็นมากที่สุด สามารถคำนวณจาก สมการ (15) เพื่อหาคำตอบสุดท้าย (Dehzangi & Karamizadeh, 2011) โดยที่ n คือ จำนวนคลาส

$$H(X) = arg_{i=1...n} max\ (LC_i(x)) \tag{15}$$

วิธีการหาค่าความน่าจะเป็นแบบต่างๆ สามารถคำนวณ จากสมการ (16-20)

Average of probabilities: $LC_i(X) = \frac{1}{m} \sum_{j=1}^{m} P_i(w_i\ |x)$ (16)

Product of probabilities: $LC_i(X) = \frac{1}{m} \prod_{j=1}^{m} P_i(w\ |x)$ (17)

Minimum of probabilities: $LC_i(X) = min_{j=1...m}\{P_j(w_i|x)\}$ (18)

Maximum of probabilities: $LC_i(X) = max_{j=1,...,m}\{P_j(w_i|x\}$ (19)

Majority voting: $H(X) = arg_{i=1...m} max\ \{\ S_i = \sum_{j=1}^{m} I\ (h_j(x))=Y\ \}$ (20)

โดยที่

i คือ จำนวนของคลาสตั้งแต่ 1 ถึง n

j คือ จำนวนของของตัวจำแนกประเภทตั้งแต่ 1 ถึง m ที่ใช้ในวิธีการเอ็นเซมเบิล

Weighted voting schemes เป็นวิธีการโหวตแบบให้ค่าถ่วงน้ำหนักที่มีความครอบคลุมภายใต้กรอบการหาค่าของ Simple weighted voting, Rescaled weighted voting, Best–worst weighted voting และ Quadratic best–worst weighted voting โดยค่าถ่วงน้ำหนัก (Weight values) หรือ w_k สำหรับแต่ละคลาส (Onan et al., 2016) สามารถคำนวณหาค่าตามสมการ (21)

$$w_k = \frac{a_k}{\sum_l a_l} \tag{21}$$

โดยที่

a_l คือ ค่าความถูกต้อง ของตัวจำแนกประเภท l จากตัวจำแนกทั้งหมด

a_k คือ ค่าความถูกต้อง ของตัวจำแนกประเภท k ของ ชุดข้อมูลสำหรับการเรียนรู้แบบจำลอง

การปรับค่าน้ำหนักตามสัดส่วน ดังสมการ (22)

$$a_k = max\{\, 0,1 - \frac{M.e_k}{N(M-1)} \,\} \tag{22}$$

โดยที่

e_k คือ จำนวนของค่าความผิดพลาดที่ได้รับ

N คือ จำนวนของตัวอย่างข้อมูล และ M คือ จำนวนคลาส

การหาค่า Best–worst weighted voting ดังสมการ (23)

$$a_k = 1 - \frac{e_k - e_B}{e_w - e_B} \tag{23}$$

โดยที่

e_B คือ ค่าความผิดพลาดต่ำสุด และ e_w คือ ค่าความพลาดสูงสุด

การหาค่า Quadratic best– worst weighted voting ดังสมการ (24)

$$a_k = \left(\frac{e_w - e_k}{e_w - e_B} \right)^2 \tag{24}$$

โดยที่

e_k คือ จำนวนของค่าความผิดพลาดที่ได้รับ

วิธีการที่ 2 วิธีการกราเดียนบูทติ้งทรี เป็นวิธีการที่ได้ปรับปรุงจากวิธีการต้นไม้ตัดสินใจและเป็นตัวจำแนกประเภทหนึ่งของวิธีการเอ็นเซมเบิล โดยสร้างการสุ่มสร้างต้นไม้ในรูปแบบลำดับชั้นโดยจะสร้างต้นไม้เพื่อลดค่าความผิดพลาดที่เกิดขึ้นจากต้นไม้ก่อนหน้า จากนั้นนำผลลัพธ์ที่ได้มา

รวมกัน วิธีการนี้จะทำให้ค่าความเอนเอียงและค่าความแปรปรวนลดลง เพราะค่าความผิดพลาดที่เกิดขึ้นก่อนหน้าได้ถูกแก้ไข แต่อาจจะมีตัวแปรหลายตัวที่ต้องปรับเพื่อให้ได้ประสิทธิภาพที่ดีขึ้น และหลีกเลี่ยงปัญหาการเกิดข้อมูลเกินจำนวนมาก (Ebrahimi et al., 2019)

วิธีการที่ 3 แรนดอมฟอเรสต์เป็นวิธีการหนึ่งของวิธีการเอ็นเซมเบิลทำงานโดยจะสุ่มข้อมูลสำหรับการเรียนรู้แบบจำลองและสุ่มคุณลักษณะ ออกเป็นหลายชุด จากนั้นจะสร้างแบบจำลองด้วยวิธีการต้นไม้ตัดสินใจขึ้นมาหลายๆ แบบจำลอง จากนั้นนำข้อมูลสำหรับการทดสอบชุดเดียวกันเข้าไปทำนาย แล้วได้ผลลัพธ์ของแต่ละแบบจำลอง จากนั้นนำผลลัพธ์มาทำการโหวตแล้วเลือกผลลัพธ์ได้คะแนนโหวตมากที่สุดเป็นคำตอบ (Han et al., 2012)

2.2 งานวิจัยที่เกี่ยวข้อง

ผู้วิจัยได้ทำการศึกษาปัญหาและแนวทางงานวิจัยที่มีความเกี่ยวข้องกับการจำแนกโรคซึมเศร้าที่วิเคราะห์ข้อมูลจากโซเชียลมีเดีย และงานวิจัยที่ศึกษาเกี่ยวกับวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก รวมทั้งงานวิจัยที่เกี่ยวข้องกับการแสดงความรู้สึกนึกคิดด้วยข้อมูลแบบหลายลักษณะ รายละเอียดดังนี้

2.2.1 การจำแนกประเภทโรคซึมเศร้าจากโซเชียลมีเดีย

กระบวนการจำแนกประเภทโรคซึมเศร้าโดยนำข้อมูลโซเชียลมีเดียมาวิเคราะห์ มีการพัฒนางานวิจัยอย่างต่อเนื่องเพื่อหาแนวทางในการช่วยแก้ปัญหาการจำแนกประเภทของโรคซึมเศร้าจากพฤติกรรมการใช้โซเชียลมีเดีย ดังนี้

Tsugawa และคณะ (2015) ได้นำเสนองานวิจัยเพื่อหาระดับของภาวะซึมเศร้าของผู้ใช้งานบน Twitter โดยใช้ข้อมูลจากผู้ใช้งาน Twitter จำนวน 3,200 tweets ซึ่งข้อมูลเหล่านี้ มีคำศัพท์ที่เกี่ยวข้องกับโรคซึมเศร้า จำนวน 1,622 ข้อความ แบ่งออกเป็น 2 กลุ่ม ได้แก่ คำศัพท์ที่บ่งบอกเป็นโรคซึมเศร้าจำนวน 862 คำ และคำศัพท์ที่บ่งบอกว่าไม่เป็นโรคซึมเศร้าจำนวน 760 คำ ในการหาระดับของภาวะซึมเศร้านั้น ความถี่คำของหัวข้อที่มีความยาว และอาจจะทำให้ได้ผลลัพธ์ที่ไม่ชัดเจน ดังนั้นจึงนำ คุณลักษณะกิจกรรมของผู้ใช้งานและหัวข้อของ tweets มาพิจารณาร่วมด้วย เพื่อทำให้เกิดความสมบูรณ์ต่อการนำข้อมูลไปวิเคราะห์ ในการพิจารณาคุณลักษณะของภาวะซึมเศร้า (Munmun, De et al., 2013) ประกอบด้วย 1) พิจารณาจากในถุงคำ 2) หัวข้อ แบ่งเป็น 5, 10 และ 20 หัวข้อ 3) อัตราส่วนของคำที่มีผลกระทบเชิงบวก 4) อัตราส่วนของคำที่มีผลกระทบเชิงลบ 5) ความถี่ในการโพสต์ต่อ 1 ชั่วโมง 6) ความถี่ในการโพสต์ต่อ 1 วัน 7) ค่าเฉลี่ยคำต่อการโพสต์ 8) อัตราการรีทวีต 9) อัตราโดยรวม 10) อัตราส่วนที่มีที่อยู่เว็บไซต์ 11) จำนวนผู้ใช้งานที่ติดตาม และ 12) จำนวนผู้ที่ถูกติดตาม การทดลองนี้ได้สังเกตุการณ์เป็นระยะเวลา 2 เดือน จากนั้นใช้วิธีการซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกประเภท โดยนำคุณลักษณะที่ 2) หัวข้อจำนวน 10 หัวข้อ 3) อัตราส่วน

ของคำที่มีผลกระทบเชิงบวก 4) อัตราส่วนของคำที่มีผลกระทบเชิงลบ 5) ความถี่ในการโพสต์ต่อ 1 ชั่วโมง 8) อัตราการรีทวีต 10) อัตราส่วนที่มีที่อยู่เว็บไซต์ 11) จำนวนผู้ใช้งานที่ติดตาม และ 12) จำนวนผู้ที่ถูกติดตาม ไปพิจารณาได้ค่าความถูกต้องมากที่สุด จากผลการทดลองค่าความถูกต้องเท่ากับ 69% ของงานวิจัยคาดการณ์ว่าปัญหาที่ทำให้ความถูกต้องของการทดลองได้ไม่สูงนั้น เพราะจำนวนข้อความที่นำเข้ามาสร้างแบบจำลองมีน้อย และเข้ารับการทำนายนั้นอาจสั้นเกินไป จนทำให้ไม่สามารถทำนายได้ และระยะเวลาในการสังเกตการณ์ที่มากขึ้นไม่ได้ช่วยเพิ่มความแม่นยำให้แบบจำลอง และในบางครั้งอาจจะได้ผลลัพธ์ที่แย่ลง

Kang และคณะ (2016) ได้ทำเสนองานวิจัยการจำแนกหาผู้คนที่เข้าข่ายเป็นโรคซึมเศร้าจากการข้อความความคิดเห็น อิโมติคอน และรูปของการโพสต์บน Twitter เพื่อการสร้างแบบจำลองที่สามารถนำเอาข้อความ อิโมติคอน และภาพ มาจำแนกว่าเข้าข่ายเป็นโรคซึมเศร้า มีชุดข้อมูลในการทดลอง 3 ชุดข้อมูล ได้แก่ 1) ชุดข้อมูลที่เป็นข้อความ 2) ชุดข้อมูลที่เป็นอิโมติคอน และ 3) ชุดข้อมูลที่เป็นภาพ จากนั้นนำมาเข้ากระบวนการทำเหมืองข้อมูล โดยใช้วิธีการซัพพอร์ตเวกเตอร์แมชชีน พบว่าการใช้เฉพาะข้อความความคิดเห็นและอิโมติคอน มีความถูกต้อง 84.45 % ความแม่นยำ 81.01 % ความระลึก 83.50 % และค่าเฉลี่ยประสิทธิภาพโดยรวม 82.24 % การใช้ข้อความความคิดเห็น อิโมติคอน และรูป มีความถูกต้อง 90.04 % ความแม่นยำ 86.01 % ความระลึก 87.51 % และค่าเฉลี่ยประสิทธิภาพโดยรวม 86.72 % จากการวิเคราะห์ผลการทดลองทำให้ทราบว่าการนำเอาข้อความความคิดเห็น อิโมติคอน และภาพ เข้ามาทำนายทำให้มีค่าความถูกต้องที่สูงตามไปด้วย เพราะมีการให้น้ำหนักอิโมติคอนด้วย จึงทำให้ข้อความจากประโยคสั้น ๆ ได้รับการทำนายถูกต้องมากขึ้น จึงทำให้ได้ผลลัพธ์ดีขึ้นไปด้วย

Aldarwish และคณะ (2017) ได้นำเสนอวิธีการจัดระดับความรุนแรงของโรคซึมเศร้าจากการโพสต์บนโซเชียลมีเดีย ซึ่งการแยกระดับผู้ป่วยที่เป็นโรคซึมเศร้าเพื่อช่วยวินิจฉัยและรักษาให้ถูกต้องตามระดับความรุนแรงของโรค งานวิจัยนี้ได้มีการพิจารณาอาการโรคซึมเศร้า 9 อาการตาม DSM-5 criteria ได้แก่ Sadness, Loss of Interest, Appetite, Sleep, Thinking Guilt, Tired, Movement และ Suicidal ideation ซึ่งได้รวบรวมข้อความจากการโพสต์บนโซเชียลมีเดีย ได้แก่ Facebook, Twitter และ LiveJournal จำนวน 6,773 ข้อความ เป็นข้อมูลที่แสดงความคิดเห็นถึงเกี่ยวกับอาการโรคซึมเศร้า จำนวน 2,073 ข้อความ และข้อความที่ไม่ได้แสดงออกถึงอาการเป็นโรคซึมเศร้า จำนวน 4,700 ข้อความ จากอัตราส่วนระหว่างข้อความที่แสดงถึงอาการเป็นโรคซึมเศร้าและไม่แสดงอาการถึงการเป็นโรคซึมเศร้ามีอัตราส่วนที่แตกต่างกันมากกว่าสองเท่า จึงดำเนินการแก้ไขปัญาด้วยการเลือกข้อความที่ไม่แสดงถึงอาการของโรคซึมเศร้าไปใช้ในการสร้างแบบจำลอง จำนวน 2,073 ข้อความ ด้วยวิธีการสุ่มเลือก จากนั้นแบ่งข้อความที่แสดงออกว่าเป็นโรคซึมเศร้าและข้อความที่ไม่แสดงออกถึงโรคซึมเศร้า แก้ไขการเกิดปัญหาความไม่สมดุลของข้อมูลด้วยการสุ่ม

ข้อความที่ไม่แสดงออกถึงโรคซึมเศร้าให้เท่ากับจำนวนข้อความที่แสดงออกถึงโรคซึมเศร้า ใช้วิธีการ Porter stemming algorithm ในการลดความยาวคำจนกว่าจะถึงความยาวคำขั้นต่ำ จากนั้นนำมาจำแนกระดับความรุนแรงด้วย วิธีการซัพพอร์ตเวกเตอร์แมชชีนและวิธีการนาอีฟเบย์ พบว่าการใช้วิธีการนาอีฟเบย์ได้ผลลัพธ์ ความถูกต้อง 63% ความแม่นยำ 100% และค่าความระลึก 57% และวิธีการซัพพอร์ตเวกเตอร์แมชชีน ได้ ความถูกต้อง 57 % ความแม่นยำ 67 % และความระลึก 56 % ซึ่งเป็นค่าที่ไม่สูงทั้งสองวิธีการ เมื่อเปรียบเทียบกับวิธีการ Social network sites (SNS) based predictive model ที่ได้ผลลัพธ์ ความถูกต้อง 77 % ความแม่นยำ 78 % และความระลึก 85 % จากการวิเคราะห์คาดว่าปัญหาที่ทำให้ผลการทดลองได้ค่าความถูกต้องที่ไม่สูง เพราะอาจนำข้อความที่เป็นคำสแลงเข้ามาทำนาย ทำให้ค่าความถูกต้องลดน้อยลงได้ จากงานวิจัยนี้ เมื่อพิจารณาอาจไม่ใช่เฉพาะคำสแลงเพียงอย่างเดียวที่ทำให้ประสิทธิภาพของแบบจำลองไม่ดี อาจเป็นเพราะวิธีการเลือกข้อมูลแต่ละอาการมีจำนวนข้อความที่แตกต่างกันมาก 1) Sadness มีจำนวน 1,195 ข้อความ 2) Loss of Interest มีจำนวน 15 ข้อความ 3) Appetite มีจำนวน 14 ข้อความ 4) Sleep มีจำนวน 67 ข้อความ 5) Thinking มีจำนวน 292 ข้อความ 6) Guilt มีจำนวน 234 ข้อความ 7) Tired มีจำนวน 77 ข้อความ 8) Movement มีจำนวน 6 ข้อความ และ 9) Suicidal ideation มีจำนวน 173 ข้อความ ซึ่งจำนวนข้อความที่ได้ทำการสุ่มเลือกเพื่อมาทดสอบอาจไม่เพียงพอต่อการนำมาวิเคราะห์อาการโรคซึมเศร้า หรือข้อความในบางอาการมีค่อนข้างน้อยทำให้ชุดคำเกิดการเรียนรู้น้อย ทำให้การทำนายไม่ถูกต้อง ซึ่งอาจใช้วิธีการอื่นที่ทำช่วยแก้ไขปัญหาความไม่สมดุลของข้อมูล และอาจจำเป็นต้องเก็บรวบรวมข้อมูลเพิ่ม

Doenribram และคณะ (2019) ได้นำเสนอการจำแนกอาการของโรคซึมเศร้า 9 อาการ โดยแบ่งชุดข้อมูล 1) ชุดข้อมูลสำหรับการเรียนรู้แบบจำลอง ใช้ข้อความการโพสต์แสดงความคิดเห็นภาษาอังกฤษจาก 9 แฮชแท็กที่เป็นอาการของโรคซึมเศร้าบน Twitter จำนวนข้อความ 27,000 ข้อความ ในงานวิจัยนี้ประกอบด้วยแบบจำลอง 9 แบบจำลองตามอาการโรคซึมเศร้า ในแต่ละแบบจำลองมีคลาสคำตอบจำนวน 2 คลาส คือ คลาสที่เป็นอาการโรคซึมเศร้าและคลาสอาการคนปกทั่วไป และ 2) สำหรับเป็นชุดการทดสอบแบบจำลอง ได้รวบรวมข้อความการแสดงคงความคิดเห็นจาก 30 คน แบ่งเป็นที่ป่วยเป็นโรคซึมเศร้า 15 คน และคนที่ปกติ 15 คน ที่มีการโพสต์ข้อความต่อเนื่องอย่างน้อย 2 สัปดาห์ และมีการโพสต์แสดงความคิดเห็นทุก ๆ 1 วัน เพื่อหาประสิทธิภาพแบบจำลองและเวลาที่ใช้ในการประมวล ในขั้นตอนการเตรียมข้อมูลใช้วิธีการ Information gian ในการคัดเลือกคุณลักษณะให้ข้อชุดมูล โดยแบ่งจำนวน k = 2000, 4000, 6000 และคุณลักษณะทั้งหมด จากนั้นได้นำเสนอคุณลักษณะ Top 10 คุณลักษณะที่มีความถี่มากที่สุดในทั้ง 9 อาการ ขั้นตอนการการจำแนกประเภทอาการโรคซึมเศร้าใช้วิธีการ นาอีฟเบย์ โดยกำหนด boundary ค่าความน่าจะเป็น 0, 10, 20, 30, 40, 50, 60, 70, 80 และ 90 ในการกรองชุดข้อความ จากนั้นทำการ

โหวตเพื่อหาคำตอบการทำนายด้วยวิธีการคะแนนเสียงข้างมากแบบไม่กำหนดค่าน้ำหนัก ในการวิเคราะห์อาการซึมเศร้าจากแบบฟอร์ม DSM-5 criteria การคำนวณตามทฤษฎี person's ผลการทดลองพบว่า 1) ในชุดข้อมูลสำหรับการเรียนรู้แบบจำลองของแบบจำลอง Loss of interest ได้ค่าเฉลี่ยความถูกต้องมากที่สุด คือ 95.85 % คุณลักษณะ 2,000 ใช้เวลาเรียนรู้แบบจำลองน้อยที่สุด คุณลักษณะทั้งหมดใช้เวลาเรียนรู้แบบจำลองมากที่สุด และการใช้คุณลักษณะทั้งหมดใช้เวลาในการขั้นตอนการเตรียมข้อมูลน้อยที่สุด โดยที่ 2,000 4,000 และ 6,000 ใช้เวลาในขั้นตอนการเตรียมข้อมูลใกล้เคียงกัน และเวลาที่ใช้ไปในทุกขั้นตอน 6,000 คุณลักษณะ ใช้เวลาประมวลผลรวมมากที่สุด และ 2) ชุดข้อมูลสำหรับการทดสอบแบบจำลอง พบว่า boundary ความน่าจะเป็น คือ 80 และ 90 ค่าเฉลี่ยความถูกต้อง 80.00 % และ Doenribram (2020) ได้นำเสนอวิธีการเปรียบเทียบแบบจำลองด้วยวิธีการเหมืองข้อมูล โดยแบ่งออกเป็น 2 ระดับ โดยระดับที่หนึ่งใช้วิธีการนาอีฟเบย์ วิธีการซัพพอร์ตเวกเตอร์แมชชีน จากนั้นนำผลคำตอบไปทำการโหวตด้วยวิธีการคะแนนเสียงข้างมากแบบไม่กำหนดค่าน้ำหนัก ผลการทดสอบพบว่า ในชุดสำหรับการเรียนรู้แบบจำลอง ระดับที่หนึ่ง วิธีการนาอีฟเบย์ ได้ค่าความถูกต้อง 82.55 % และวิธีการซัพพอร์ตเวกเตอร์แมชชีน ได้ค่าความถูกต้อง 96.18% ในระดับที่สอง จึงนำวิธีการซัพพอร์ตเวกเตอร์แมชชีนที่ได้ค่าความถูกต้องไปใช้ทดสอบในระดับที่สอง ใช้วิธีการซัพพอร์ตเวกเตอร์แมชชีน+แรนดอมฟอเรสต์ ได้ค่าความถูกต้อง 94.45 % และวิธีการซัพพอร์ตเวกเตอร์แมชชีน+วิธีการนาอีฟเบย์ ความถูกต้อง 83.23 % ชุดข้อมูลสำหรับทดสอบระดับที่หนึ่ง วิธีการนาอีฟเบย์ ความถูกต้อง 76.67 % และวิธีการซัพพอร์ตเวกเตอร์แมชชีน ความถูกต้อง 70.00 % ระดับที่สองวิธีการซัพพอร์ตเวกเตอร์แมชชีน+แรนดอมฟอเรสต์ ความถูกต้อง 70.00 % วิธีการซัพพอร์ตเวกเตอร์แมชชีน+วิธีการนาอีฟเบย์ ความถูกต้อง 73.33 % จากงานวิจัยทั้งสอง พบว่าเมื่อจำนวนของคำตอบที่ทายผิดมีมากกว่าจำนวนคำตอบที่ทายถูกทำให้การทำนายคำตอบสุดท้ายผิด และทำให้ประสิทธิภาพของแบบจำลองลดลง และใช้เวลาในการประมวลค่อนข้างนาน ดังนั้นจึงเกิดคำถามในการวิจัยว่า หากนำชุดข้อมูลดังกล่าว มากำหนดค่าน้ำหนักให้ชุดคำตอบที่ดีมากกว่าชุดคำตอบที่ไม่ดี จะทำให้ประสิทธิภาพแบบจำลองและเวลาในการประมวลผลดีขึ้น และหากนำชุดข้อมูลสำหรับการเรียนรู้ 9 อาการของโรคซึมเศร้า และ 1 อาการของบุคคลปกติ มาเป็นคลาสคำตอบ พบว่าอาจมีบางคุณลักษณะในคลาสหนึ่งที่อาจจะส่งผลต่อคำตอบในคลาสอื่นได้ ดังนั้นการใช้วิธีการหาค่าถ่วงน้ำหนักที่เหมาะสมให้คำตอบตามประสิทธิภาพการทำนายอาจเป็นช่วยแก้ไขปัญหาดังกล่าวได้

Alabdulkreem (2020) ได้นำเสนองานวิจัยนี้ที่นำข้อมูลจากการโพสต์แสดงความคิดเห็นและความรู้สึกของผู้ใช้งานบน Twitter ด้วยภาษาอาหรับของหญิงอาหรับในสถานการณ์การระบายของไวรัสไวรัสโคโรนาสายพันธุ์ 2019 ระหว่างเดือน มีนาคม-สิงหาคม ปี 2563 จากผู้ใช้งานจำนวน 200 คน มากกว่า 10,000 ข้อความ แบ่งเป็น 2 ชุด คือ 7,000 ข้อความสำหรับการเรียนรู้

แบบจำลอง และ 3,000 ข้อความสำหรับการทดสอบแบบจำลอง โดยคัดเลือกเฉพาะผู้ใช้งานที่มีการโพสต์มากกว่า 50 โพสต์เป็นต้นไปเพื่อให้แน่ใจว่ามีข้อมูลเพียงพอต่อการนำมาวิเคราะห์ข้อมูล โดยมีจำนวน 2 คลาส คือ มีภาวะซึมเศร้าและไม่มีภาวะซึมเศร้า เนื่องด้วยโครงสร้างของภาษาอาหรับมีความซับซ้อนมากกว่าภาษาอังกฤษจึงต้องมีวิธีการจัดการเกี่ยวโครงสร้างทางภาษาให้ดี จากนั้นได้เปรียบเทียบประสิทธิภาพการจำแนกภาวะซึมเศร้าโดยแบ่งเป็น 2 แบบ ได้แก่ 1) วิธีการสอนแบบดั้งเดิม (Traditional learning) ประกอบด้วย วิธีการซัพพอร์ตเวกเตอร์แมชชีน การถดถอยโลจีสติก แรนดอมฟอเรสต์ และต้นไม้ตัดสินใจ และ 2) การเรียนรู้เชิงลึก (Deep learning) ประกอบด้วย CNN, DNN และ Recurrent neural network - Long short-term memory (RNN-LSTM) พบว่าการจำแนกภาวะซึมเศร้าแบบ วิธีการสอนแบบดั้งเดิม วิธีการต้นไม้ตัดสินใจมีความถูกต้องมากที่สุด 60 % และ ค่าเฉลี่ยประสิทธิภาพโดยรวม 57 % ซึ่งได้ผลลัพธ์ที่น้อยกว่าวิธีการการเรียนรู้เชิงลึก โดย RNN-LSTM เป็นวิธีการที่ได้ประสิทธิภาพมากกว่าวิธีการอื่น ความถูกต้อง 72 % ความแม่นยำ 71 % ความระลึก 68 % และค่าเฉลี่ยประสิทธิภาพโดยรวม 69 % และใช้เวลาการประมวลน้อยที่สุด

Yazdavar และคณะ (2020) ได้นำเสนอ Multimodal framework ประกอบ Content-based models, Image-based models และ Network-based models เพื่อวิเคราะห์สุขภาพทางจิตโดยใช้ข้อมูลจากคุณลักษณะภาพ Profile คุณลักษณะข้อมูลภาพจากการใช้โซเชียลมีเดีย (จำนวนเพื่อน จำนวนการติดตาม สถานะ การรีทวีต รายการโปรด) คุณลักษณะข้อความ เพื่อจำแนกกลุ่มที่มีมีภาวะซึมเศร้า ใช้วิธีการเอ็มเซมเบิลในการคัดเลือกคุณลักษณะร่วมกับหลักการทางสถิติ ได้แก่ Chi-square, Pearson correlation และ ANOVA เพื่อคัดลือกคุณลักษณะไปในสร้างแบบจำลองทั้ง 3 วิธีการที่นำเสนอได้ค่าเฉลี่ยประสิทธิภาพโดยรวมเพิ่มขึ้น 5%

2.2.2 วิธีการเอ็นเซมเบิล

Onan และคณะ (2016) ได้นำเสนอวิธีการกำหนดค่าน้ำหนักใน Voting ensemble method โดยพัฒนารูปแบบการให้คะแนนแบบถ่วงน้ำหนักด้วยการปรับค่าน้ำหนักที่เหมาะสมให้กับตัวจำแนกประเภทและคลาสคำตอบแต่ละตัวตามค่าความแม่นยำและความระลึก จากการทำนายของตัวจำแนกประเภททั้งหมด เพื่อแก้ไขปัญหาจำนวนคำตอบที่ไม่ดี แต่มีจำนวนมากกว่าคำตอบที่ดี การแก้ไขปัญหาดังกล่าว จะสามารถเพิ่มประสิทธิภาพในการทำนายในการจำแนกประเภทข้อความ ในพัฒนารูปแบบการให้คะแนนแบบถ่วงน้ำหนักแบบ Multi-objective optimization โดยจะกำหนดค่าน้ำหนักที่เหมาะสมให้กับตัวจำแนกประเภทและคลาสคำตอบแต่ละตัวตามประสิทธิภาพการทำนายของตัวจำแนกประเภททั้งหมด เพื่อเพิ่มประสิทธิภาพในการทำนาย ซึ่งประสิทธิภาพของค่าความแม่นยำและความระลึก จะเป็นตัวกำหนดการปรับค่าน้ำหนัก ดังนั้นจึงจำเป็นต้องหาวิธีการในการกำหนดค่าน้ำหนัก ในงานวิจัยนี้ได้ใช้ชุดข้อมูลในการทดลอง 9 ชุด ได้แก่ Camera, Camp, Doctor, Drug, Laptop, Lawyer, Radio, TV และ Music ซึ่งเป็นชุดข้อมูลที่มีคลาสจำนวน 2 คลาส

คือ คลาสเชิงบวกและคลาสเชิงลบ ทดสอบเพื่อหาประสิทธิภาพแบบจำลอง โดยการนำวิธีการจำแนกประเภท Stacking, Voting schemes และ Metaheuristic-based weighted voting ที่ประกอบไปด้วย Genetic algorithm (GA), Multi-objective simulated annealing (SA), Differential evolution (DE), Multi-objective particle swarm optimization (CMDPSO) และ Multi-objective differential evolution (MODE) ซึ่งวิธีการ Metaheuristic-based weighted voting จะทำการคำนวณหาค่า Multi-objective weight adjustment โดยพิจารณาจากค่าความแม่นยำและความระลึก จากนั้นนำค่าน้ำหนักที่ได้มาทำการเลือกคำตอบสุดท้ายในการทำนาย จากผลการทดลองพบว่า MODE-based weighted voting ได้ผลลัพธ์ดีที่สุด SA-based weighted voting ได้ผลลัพธ์รองลงมา ซึ่งผลลัพธ์ที่ได้มีค่าที่ใกล้เคียงกันมาก จากงานวิจัยนี้พบว่า การปรับปรุงค่าถ่วงน้ำหนักให้คำตอบแต่ละตัวในแต่ละคลาส จากนั้นแล้วนำค่าถ่วงน้ำหนักที่มากที่สุดมาเป็นคำตอบสุดท้ายในแต่ละคลาสทำให้ได้ประสิทธิภาพเพิ่มขึ้น ซึ่งช่วยแก้ปัญหาการเลือกคำตอบที่ดีแต่มีจำนวนน้อยกว่าคำตอบที่ไม่ดีได้

Zul และคณะ (2018) ได้นำเสนอการผสมผสานวิธีการเหมืองข้อมูลการวิเคราะห์ความเชื่อมั่นของ Facebook และ Twitter โดยรวบรวมข้อมูล Feeds จาก Facebook ในขั้นตอนการเตรียมข้อมูลได้ ลบค่าที่ผิดปกติออก กำหนดลาเบลด้วยวิธีการ Sentiwordnet คัดเลือกคุณลักษณะด้วยวิธีการ Frequent itemset mining (FIM) และนำชุดข้อมูลไปทำการจำแนกประเภทด้วยวิธีการวิธีการนาอีฟเบย์ และ วิธีการนาอีฟเบย์+การแบ่งกลุ่มแบบเคมีน โดยกำหนดให้ k=6, 7, 8, 9 และ 10 ผลการทดลองพบว่า วิธีการนาอีฟเบย์ มีค่าเฉลี่ยความถูกต้องอยู่ระหว่าง 80.52% - 82.50% และ ธีการนาอีฟเบย์+การแบ่งกลุ่มแบบเคมีน มีค่าเฉลี่ยความถูกต้องอยู่ระหว่าง 80.32% - 81.52% จากกงานวิจัยนี้พบว่า ให้กำหนดลาเบลด้วยวิธีการ Sentiwordnet ทำให้ได้ผลลัพธ์ที่ดีกว่าการนำ K-Means มาเป็นคำตอบในการทดลองด้วยวิธีการนาอีฟเบย์+การแบ่งกลุ่มแบบเคมีน

AlHamed และ AlGwaiz (2020) ได้นำเสนองานวิจัยเกี่ยวกับการนำวิธีการเหมืองข้อมูลแบบผสมผสานในการวิเคราะห์ข้อมูลของบริษัทบน Twitter จำนวน 14,200 จากบริษัท Starbucks, Aramex, Uber และ Pizza เป็นข้อความเชิงบวก 4,590 และเชิงลบ 4,070 ในการวิจัยนี้ได้แบ่งกระบวนการทำงานเป็น 2 ระดับ คือ ระดับที่ 1 Lexicon classifier จะกำหนดลาเบลให้ชุดข้อมูลเป็นคลาสเชิงบวก เป็นกลาง และเชิงลบ จากนั้นทำการจำแนกประเภทด้วย ANN ระดับที่ 2 ผู้บริโภคสามารถระบุผลิตภัณฑ์ บริการ และประกาศ การรับรู้ในเชิงบวกและเชิงลบ จากนั้นจัดอันดับรายการคำที่ถูกพูดถึงมากที่สุดในเชิงบวกและเชิงลบด้วยวิธีการ Term frequency ผลการทดลองได้ค่าความถูกต้อง 89% บริษัทสามารถนำไปใช้ประโยชน์ในการระบุจุดอ่อนและจุดแข็ง ในการโฆษณาและผลกระทบต่อการตัดสินใจเชิงกลยุทย์ได้ดี

Rojarath และ Songpan (2020) นำเสนอการเลือกจำนวนตัวจำแนกประเภทจากตัวจำแนกประเภทพื้นฐานและตัวจำแนกแบบเอ็นเซมเบิลเพื่อให้ค่าน้ำหนักแต่ละคลาสสำหรับการโหวตแบบให้ค่าถ่วงน้ำหนักด้วยอัตราการทำนายถูกของคลาส และได้เปรียบเทียบผลลัพธ์ระหว่างวิธีการที่ได้นำเสนอกับตัวจำแนกแบบเอ็นเซมเบิลโดยทดสอบกับข้อมูลจำนวน 5 ชุด ที่ประกอบด้วยคลาสคำตอบแบบไบนารีและแบบหลายคลาส และ Rojarath (2021) ได้นำเสนอกระบวนการเพื่อปรับปรุงประสิทธิภาพการทำนายของแต่ละแบบจำลองดีขึ้นสำหรับการให้ค่าน้ำหนักแก่คลาสด้วยการกำหนดค่าแก่คลาสด้วยค่าความน่าจะเป็นการเกิดคลาสผลลัพธ์ โดยทดสอบกับข้อมูลจำนวน 10 ชุด ที่ประกอบด้วยคลาสคำตอบแบบไบนารีและแบบหลายคลาส พบว่าวิธีการทั้งสองวิธีการเมื่อให้ค่าน้ำหนักแล้วสามารถเพิ่มความเสถียรภาพแก่ตัวแบบการเรียนรู้แบบผสมผสานทำให้ได้ผลลัพธ์ที่ดี มีความถูกต้องเพิ่มมากขึ้นในชุดข้อมูลแบบหลายคลาส และช่วยแก้ไขปัญหาความไม่สมดุลของชุดข้อมูลได้ ในงานวิจัยนี้ไม่ได้กล่าวถึงกระบวนการคัดเลือกคุณลักษณะ

Khan และคณะ (2021) ให้นำเสนอวิธีการผสมผสานการจำแนกประเภทภาวะซึมเศร้าทดสอบกับชุดข้อมูล Genres-tags movielens โดยคัดเลือกคุณลักษณะด้วยวิธีการ Recursive feature elimination (RFE) จากนั้นนำคุณลักษณะที่ได้มาจำแนกประเภทด้วยวิธีการ วิธีการซัพพอร์ตเวกเตอร์แมชชีน วิธีการเพื่อนบ้านที่ใกล้ที่สุด วิธีการนาอีฟเบย์ การวิเคราะห์การถดถอยโลจีสติก การวิเคราะห์การจำแนกประเภทเชิงเส้น และ Classification and regression trees (CART) ผลการทดลองพบว่า CART ผลลัพธ์ที่ดีที่สุด จากนั้นทำการทดลองด้วยวิธีการเอ็นเซมเบิลแบบต่าง ๆ ผลการทดลองพบว่า วิธีการแรนดอมฟอเรสต์ให้ผลลัพธ์ที่ดีที่สุด

นอกจากงานวิจัยที่เกี่ยวข้องทั้ง 2 ประเด็นข้างต้นนั้นเป็นงานวิจัยที่นำข้อความไปวิเคราะห์เพื่อสร้างแบบจำลอง ยังมีงานวิจัยที่เกี่ยวข้องที่ใช้ข้อมูลแบบ Multimodal analysis ซึ่งใช้ข้อมูล เช่น ข้อความ สีภาพ การกระจายสี ความสว่างของภาพ ฉากที่อยู่ในภาพ และอื่น ๆ ในการวิจัย (Choudhury et al., 2016; Wendlandt et al., 2017) และพบว่าคุณลักษณะข้อมูลเหล่านี้มีความสัมพันธ์กับลักษณะทางจิตใจและความรู้สึกนึกคิดของผู้ใช้งาน เช่น การวิจัยเกี่ยวกับผู้ป่วยที่เป็นโรคอารมณ์สองขั้วโดยเฉพาะช่วงของอารมณ์ซึมเศร้าจะมีแนวโน้มที่จะเปิดเผยอารมณ์ของตนจากการเลือกสีหรือชอบเฉดสีเข้มในสถานการณ์ในชีวิตประจำวัน (Fernandes et al., 2017) การเปรียบเทียบความเกี่ยวข้องกันระหว่างคุณลักษณะที่ใช้ในการทำนายโดยใช้ข้อมูลข้อความ ภาพ และใช้ข้อความร่วมกับภาพ (Wendlandt et al., 2017) การวิจัยเพื่อพัฒนาแบบจำลองแบบ multimodal ด้วยการใช้วิธีการทางสถิติเพื่อรวมชุดคุณลักษณะที่แตกต่างกัน ได้แก่ ข้อมูลภาพ ข้อความ และการใช้งานระบบของผู้ใช้งานบน Twitter เพื่อแยกแยะพฤติกรรมซึมเศร้าโดยใช้คุณลักษณะข้อมูลที่หลากหลาย (Yazdavar et al., 2020) เป็นต้น

2.3 ทฤษฎีของภาพ

ภาพสามารถใช้สื่อสารเพื่อบ่งบอกความหมายและการแสดงออกทางด้านอารมณ์ได้ องค์ประกอบของภาพ ประกอบด้วย สี พื้นผิว รูปร่าง ขนาด ที่ว่าง จุด และลักษณะเฉพาะอื่นๆ คุณสมบัติของภาพมีความคล้ายคลึงและแตกต่างกัน การนำคุณสมบัติของภาพนำไปใช้ประโยชน์เพื่อวิเคราะห์การแสดงออกด้านอารมณ์ เพื่อการดึงคุณลักษณะของภาพเพื่อทำให้ได้ข้อมูลที่ต้องการ จึงจำเป็นต้องศึกษาวิธีการที่เหมาะสมและเข้าใจคุณสมบัติของภาพก่อน ในงานวิจัยได้มุ่งเน้นศึกษาเกี่ยวกับภาพดิจิทัลขนาด 2 มิติ และการประมวลผลภาพดิจิทัล (Digital image processing) เพื่อประมวลผลภาพสีและนำคุณลักษณะสีภาพเพื่อนำข้อมูลไปวิเคราะห์ภาวะซึมเศร้า การเก็บภาพในระบบดิจิทัลจะมีการกำหนดตำแหน่งเป็นเมตริก์ (Matrix) ตามจำนวนคอลัมน์และแถว แทนจำนวนจุดพิกเซล (Pixel) ของภาพ ขนาด m*n โดยมีขนาดคอลัมน์ m และ n แถว เมตริกซ์ (Carruthers et al., 2010; Reece & Danforth, 2017)

ฮิสโตแกรม (Histogram) ของภาพระบุถึงการกระจายความเข้มแสงด้วยค่าความถี่ โดยแสดงระดับความเข้มแสงจากนับจำนวนพิกเซลที่มีค่าความเข้มเหมือนกันเรียงค่าตั้งแต่ 0-255 (มืดสุด - สว่างสุด) หากการกระจายส่วนใหญ่อยู่ทางซ้ายภาพนั้นจะมีความสว่างน้อย หากการกระจายส่วนใหญ่อยู่ทางขวาภาพนั้นมีความสว่างมาก (Singh & Hemachandran, 2012) ในงานวิจัยนี้ใช้ การสร้างฮิสโตแกรมภาพสี (Color histogram) แบบแยกองค์ประกอบจากแบบจำลองสี RGB (Red, Green, Blue) โดยแยกองค์ประกอบของแต่ละสี R [0, 255] - G [0, 255] - B [0, 255] การคำนวณการกระจายสีในงานวิจัยนี้ ใช้การหาค่าเฉลี่ย (El-Bendary, Hariri, Hassanien, & Badr, 2015) เพื่อนำคุณลักษณะสีไปใช้งาน

$$\bar{\square} = \frac{\sum_{\square=1}^{\square\cdot\square} \square_{\square\square}}{\square \cdot \square} \tag{25}$$

เมื่อ $M \cdot N$ คือ จำนวนพิกเซล (Pixels) ของภาพ ขนาด 2 มิติ

x_{ij} คือ ค่าพิกเซลภาพ j ของสี i

2.4 การประเมินประสิทธิภาพ

สมการที่ใช้ในการประเมินประสิทธิภาพแบบจำลองด้วยวิธีการแบ่งข้อมูลเป็น 10 ส่วน (10-fold cross validation)ในงานวิจัยนี้ ได้แก่ Accuracy, Precision, Recall และ F-measure หรือ F1-score (Han et al., 2011) รายละเอียดสมการ ดังนี้

Actual Values

Predicted Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

ภาพที่ 2 Confusion Matrix ขนาด 2x2

โดยที่

True Positive (TP) คือ ทำนายว่า จริง และสิ่งที่เกิดขึ้น เป็นจริง

True Negative (TN) คือ ทำนายว่า ไม่จริง และสิ่งที่เกิดขึ้น ไม่จริง

False Positive (FP) คือ ทำนายว่า จริง แต่สิ่งที่เกิดขึ้น ไม่จริง

False Negative (FN) คือ ทำนายว่าไม่จริง แต่สิ่งที่เกิดขึ้น เป็นจริง

ค่าความถูกต้อง (Accuracy) เป็นการหาความถูกต้องการจำแนกประเภทข้อมูลโดยรวม หรือการพิจาณาทุกคลาส คำนวณจากผลการทำนายถูกหารด้วยจำนวนของข้อมูลทั้งหมดที่นำมาทำนาย ดังสมการ (26)

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN} \tag{26}$$

การวัดความแม่นยำ (Precision) โดยพิจารณาแยกแต่ละคลาส โดยคำนวณหาจากจำนวนครั้งที่ทายว่า Positive แล้วถูก หารด้วยจำนวนครั้งที่ทายว่า Positive ทั้งหมด ดังสมการ (27)

$$\text{Precision}= \frac{TP}{TP+FP} \tag{27}$$

การวัดค่าความระลึก (Recall) หรือ Senstivity เป็นการวัดความถูกต้องของวิธีการ ซึ่งพิจารณาทีละคลาส โดยคำนวณหาจาก จำนวนครั้งที่ทายว่า Positive แล้วถูก หารด้วยจำนวน Positive ทั้งหมดในข้อมูล ดังสมการ (28)

$$\text{Recall}= \frac{TP}{TP+FN} \tag{28}$$

ค่าเฉลี่ยประสิทธิภาพโดยรวม (F1-score) ซึ่งพิจารณาทีละคลาส โดยคำนวณหาค่า Harmonic Mean ของ Precision และ Recall ดังสมการ (29)

$$\text{F1}=2 \times \left(\frac{precision \times recall}{precision+recall}\right) \tag{29}$$

2.5 Paired – sample t-test

ทฤษฎี Paired – sample t-test เป็นการเปรียบเทียบข้อมูลผลก่อนการทดสอบกับข้อมูลผลหลังการทดสอบกับกลุ่มตัวอย่างเดียวกัน จากสมมติฐาน *(Hypothesis)* ทางสถิติการพิสูจน์ว่าสมมติฐานเป็นจริงหรือไม่ จะนำข้อมูลไปไปทดสอบสมมติฐานทางสถิติ *(Statistic Hypothesis Testing)* เพื่อสรุปสมมติฐานนั้นว่าเป็นจริงหรือไม่จริง โดยบอกค่าระดับนัยสำคัญหรือโอกาสที่จะผิดพลาดไว้ด้วย ขั้นตอนการทดสอบสมมติฐาน ได้แก่ การตั้งสมมติฐาน กำหนดระดับนัยสำคัญ เลือกวิธีการทางสถิติและการคำนวณทางสถิติ หาค่า p-value ตัดสินใจและสรุปผล สัญลักษณ์ที่ใช้ในการทดสอบสมมติฐาน ได้แก่ H_0 คือ สมมติฐานที่กำหนดขึ้นเพื่อทดสอบ โดยค่าที่ระบุใน H_0 ต้องมีเครื่องหมายเท่ากับ (=) ด้วยเสมอ μ_0 คือ ค่าเฉลี่ยการคาดการณ์หรือค่าเฉลี่ยของสมมติฐานที่ตั้งขึ้น H_1 คือ สมมติฐานแย้ง จะมีความหมายในทางตรงกันข้ามกับ H_0 ซึ่ง μ_1 คือ ค่าเฉลี่ยของประชากรที่ไม่ทราบค่า และ $\square^2$ คือ ค่าแปรปรวนประชากร หากผลการทดสอบ ยอมรับ H_0 จะปฏิเสธ H_1 แต่หากผลการทดสอบเป็นไปในทางตรงกันข้ามด้วยการปฏิบเสธ H_0 จะยอมรับ H_1 การทดสอบสมมติฐานทางสถิติแบบข้างเดียว (One-sided test) มักจะมีเครื่องหมายมากกว่า (>) หรือน้อยกว่า (<) อยู่ใน H_1 และ H_0 โดยจะมีเครื่องหมายเท่ากับอยู่ในเครื่องหมายมากกว่าหรือน้อยกว่าด้วยเสมอ และการทดสอบสมมติฐานทางสถิติแบบสองทาง (Two-sided test) มักจะมีเครื่องหมายเท่ากับที่ H_0 เช่น

$H_0 : \mu_0 \leq 75.25$

$H_1 : \mu_1 > 75.25$ และ

$H_0 : \mu_1 = \mu_0$ ผลการทดสอบกอนและหลังการทดสอบไม่แตกต่างกัน

$H_1 : \mu_1 \neq \mu_0$ ผลการทดสอบกอนและหลังการทดสอบแตกต่างกัน

วิธีการทางสถิติที่มีส่วนเกี่ยวข้องกับการทดสอบสมมติฐาน ได้แก่ ค่าเฉลี่ย (Mean) ส่วนเบี่ยงเบนมาตราฐาน (Standard deviation) และความคลาดเคลื่อนมาตรฐาน (Standard error of measurement: S.E.) (Jennings et al., 2002) สถิติสำหรับการทดสอบ t-test

$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \qquad (30)$$

เมื่อ t คือ ค่าทางสถิติที่ได้จากการคำนวณ จากสมการ (30)

$\bar{\square}$ คือ ค่าเฉลี่ยกลุ่มตัวอย่าง

$\square$ คือ ค่าเฉลี่ยการคาดการณ์

n คือ จำนวนกลุ่มตัวอย่าง

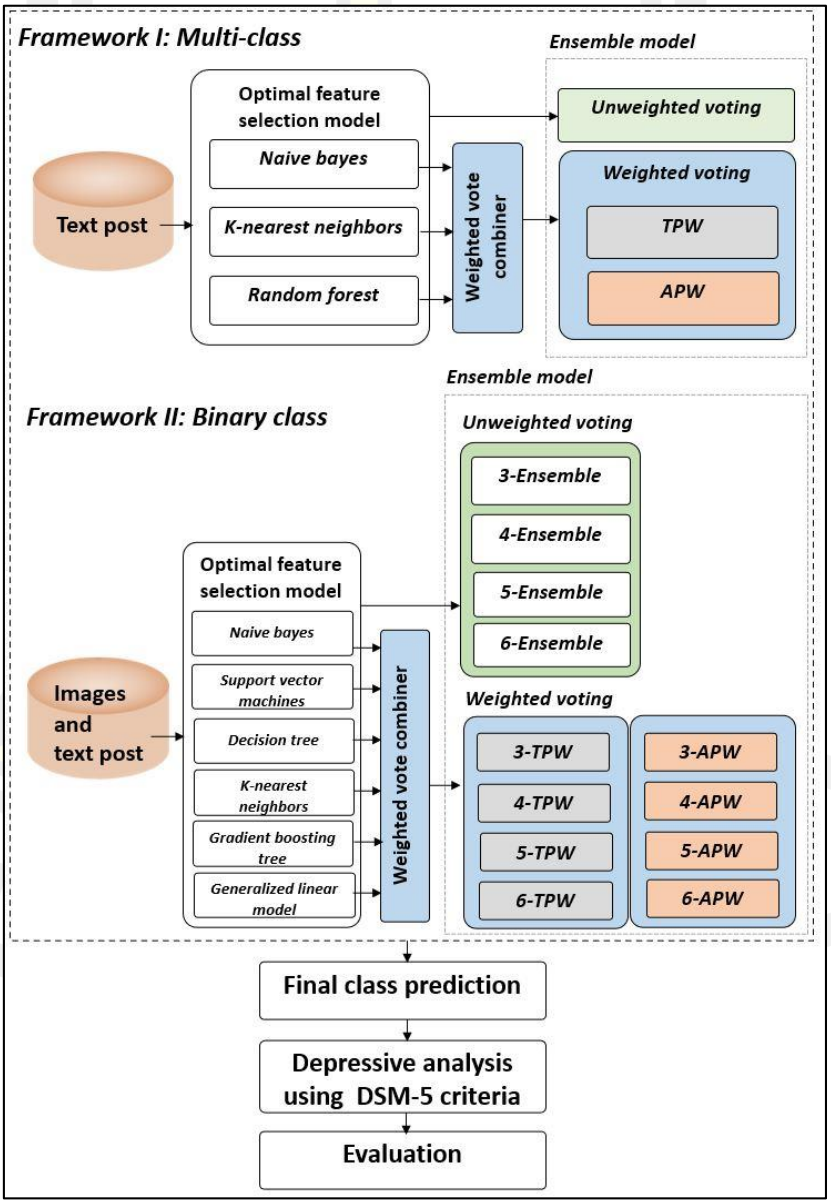
$\square$ คือ ความคลาดเคลื่อนมาตรฐาน

ในการแปลความหมายระดับความเชื่อมั่น (Confidence level) ในการทดสอบสมมติฐาน จะบอกโอกาสความถูกต้องของผลในการทดสอบสมมติฐาน ระดับความเชื่อมั่นที่ใช้ เช่น 99% 95% 90% เป็นต้น โดยระดับความเชื่อมั่นจะตรงกันข้ามกับระดับนัยสำคัญ (Significance level) หรือความผิดพลาดของการทดสอบมมติฐาน แทนด้วยสัญลักษณ์ **α**

จากการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง จึงได้นำการสกัดคุณลักษณะด้วยวิธีการ Binary term occurrences การคัดเลือกคุณลักษณะด้วยวิธีการ Information gain มาทดสอบเพื่อคัดเลือกคุณลักษณะที่คำสัญของชุดข้อมูลแบบหลายคลาสและไบนารีคลาส และเลือกตัวจำแนกประเภทแบบเดี่ยวด้วยวิธีการ นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ เพื่อนบ้านที่ใกล้ที่สุด ตัวแบบเชิงเส้นนัยทั่วไป กราเดียนบูทติ้งทรี แรนดอมฟอเรสต์ ซึ่งเป็นวิธีการที่ให้ประสิทธิภาพที่ดีกับชุดข้อมูลในงานวิจัย เพื่อนำไปใช้วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก และได้นำเสนอวิธีการเอ็นเซมเบิลแบบการกำหนดค่าน้ำหนัก ด้วยวิธีการอัตราการทำนายถูกของคลาสคำตอบ *(*True Positive weighted ensemble*:* TPW*)* และค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ *(*Average probability weighted ensemble*:* APW*)* เพื่อปรับปรุงประสิทธิภาพการพยากรณ์โรคซึมเศร้าในวัยรุ่น จากนั้นทำการวัดประเมินประสิทธิภาพวิธีการที่ได้นำเสนอ และทำการทดสอบ Paired – sample t-test เพื่อเปรียบเทียบความแตกต่างกันของผลลัพธ์ระหว่างวิธีการที่นำเสนอ

บทที่ 3
วิธีดำเนินการวิจัย

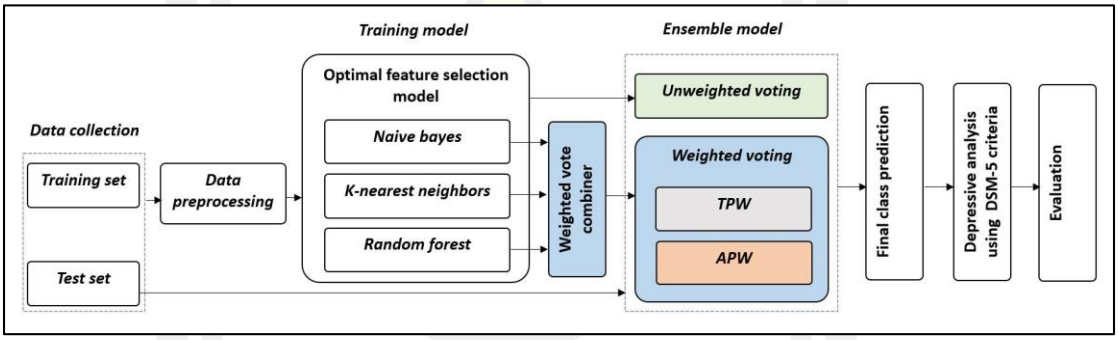
การดำเนินการวิจัยนี้เพื่อบรรลุตามวัตถุประสงค์ ผู้วิจัยได้ดำเนินการวิจัยตามทฤษฎีและงานวิจัยที่เกี่ยวข้องกับการวิธีการการพยากรณ์โรคซึมเศร้า โดยการดำเนินการงานวิจัยแบ่งวิธีการวิจัยออกเป็น 2 ส่วน ได้แก่ 1) การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล และ 2) การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ ดังนี้



ภาพที่ 3 กรอบการวิจัยการปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น

จากภาพที่ 3 แสดงถึงกรอบการวิจัยและขั้นตอนการทำงาน 2 ส่วน โดยการดำเนินการกรอบการวิจัยที่ 1 ได้ทำการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล (Framework I) สำหรับวิเคราะห์โรคซึมเศร้าด้วยชุดข้อมูลแบบหลายคลาส (Multi-class) และการดำเนินการกรอบการวิจัยที่ 2 ได้ทำการวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยการเพิ่มคุณลักษณะข้อมูลจากภาพ (Framework II) สำหรับวิเคราะห์โรคซึมเศร้าด้วยชุดข้อมูลแบบไบนารีคลาส (Binary class) โดยเปรียบเทียบการหาค่าน้ำหนักที่เหมาะสมจากตัวจำแนกประเภทที่แตกต่างกันเพื่อปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล ด้วยวิธีการจากกรอบการวิจัยที่ 1

3.1 การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล



ภาพที่ 4 ขั้นตอนการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล

การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล มีวัตถุประสงค์และขั้นตอนการทำงาน ดังนี้

1. การหาจำนวนคุณลักษณะที่เหมาะสมสำหรับการเรียนรู้แบบจำลอง ตัวจำแนกประเภทที่ใช้ได้แก่ นาอีฟเบย์ ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด แรนดอมฟอเรสต์ สำหรับชุดข้อมูลแบบหลายคลาส

2. การปรับปรุงค่าน้ำหนักจำนวน 2 วิธีการ คือ 1) อัตราการทำนายถูกของคลาสคำตอบ (True positive weighted ensemble: TPW) และ 2) ค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ (Average probability weighted ensemble: APW) จากผลลัพธ์ที่ได้จาก ชุดข้อมูลสำหรับเรียนรู้แบบจำลอง

3. การเปรียบเทียบผลลัพธ์ระหว่าง ตัวจำแนกประเภทแบบเดี่ยว วิธีการเอ็นเซมเบิลแบบไม่กำหนดน้ำหนัก และวิธีการเอ็นเซมเบิลแบบกำหนดน้ำหนักแบบ TPW และ APW

ขั้นตอนการดำเนินการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล ดังนี้

3.1.1 การรวบรวมข้อมูล

งานวิจัยนี้ ได้เก็บรวบรวมข้อมูลในการทดสอบจาก จากงานวิจัย การจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบน Twitter (Doenribram et al., 2019) ซึ่งเป็นข้อมูลการโพสต์ Twitter ประกอบด้วยชุดข้อมูล 2 ชุด ดังนี้

ชุดข้อมูลที่ 1 สำหรับการเรียนรู้แบบจำลอง รวบรวมจากการติดแฮชแท็ก บน Twitter ของผู้ใช้งานทั่วไป ประกอบด้วย 9 อาการ ที่ได้บ่งบอกถึงอาการของโรคซึมเศร้า และ 1 อาการที่บ่งบอกถึงบุคคลปกติ ข้อมูลนี้เป็นข้อมูลที่มีจำนวนเท่ากันทุกอาการเป็นข้อความภาษาอังกฤษ แต่ละอาการคือคลาสคำตอบ ดังตารางที่ 2

ตารางที่ 2 จำนวนข้อมูลจากแฮชแท็กตามลักษณะอาการชุดข้อมูลสำหรับเรียนรู้แบบจำลอง

ลำดับ	อาการ	แฮชแท็ก	ข้อความ
1	อารมณ์ซึมเศร้า	#sadness #depressive	3,000
2	ความสนใจลดลง	#loss of interest #lose interest	3,000
3	น้ำหนักผิดปกติ	#appetite #hunger	3,000
4	การนอนผิดปกติ	#sleepless #hethargy	3,000
5	สมาธิสั้น	#un thinking #out thinking	3,000
6	รู้สึกไร้ค่า	#guilt #disgrace #dishonor	3,000
7	ร่างกายอ่อนเพลีย	#tired #bored #fatigued	3,000
8	กระวนกระวาย หรือเชื่องช้า	#lackadaisical #lazy #loafing #phlegmatic	3,000
9	การอยากฆ่าตัวตาย	#suicidal #dangerous #destructive	3,000
10	ปกติ	#Happy	3,000
รวม			**30,000**

ตารางที่ 3 ตัวอย่างข้อความจากแฮชแท็กตามลักษณะอาการ

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ตัวอย่างข้อมูล
1	Id	รหัสการโพสต์	104642xxxxxx0000
2	Text	ข้อความที่โพสต์	I've realized i've been drenched in a whole lot more depressive stories than usual

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ตัวอย่างข้อมูล
3	Label	ประเภทตามลักษณะอาการ	Depressed mood (1)

ตารางที่ 4 ตัวอย่างข้อมูลสำหรับการเรียนรู้แบบจำลอง

Id	Text	Label
104620677718xxxxx29	Supermodel Gisele Bündchen speaks out about her struggle with severe panic attacks and suicidal thoughts. https://t.co/brfugVAyZI	9
104644822507xxxxx97	RT @PeterStefanovi2: Yet another gut-wrenching indictment of Tory Britain. A disabled man left ‘suicidal’ and ‘living off water’ after his…	9
104644818430xxxxx32	RT @emileypaigeargo: no suicidal shit but do y’all ever wonder what the world would be like if you weren’t here anymore?	9
104644716952xxxxx46	SUICIDAL DOORS CALL IT KURT COBAIIIN https://t.co/N077Z3eMQc	9

ชุดข้อมูลที่ 2 สำหรับทดสอบแบบจำลอง รวบรวมจากการโพสต์ข้อความภาษาอังกฤษบน Twitter ของนักแสดงและเป็นผู้ที่มีชื่อเสียง จำนวน 30 คน ประกอบด้วย 15 คน ที่เปิดเผยว่าตนเองเป็นโรคซึมเศร้า (34,435 ข้อความ) และ 15 คนไม่เป็นโรคซึมเศร้า (22,498 ข้อความ) โดยผู้ใช้งานแต่ละคนต้องโพสต์ข้อความต่อเนื่องกันในระยะเวลามากกว่า 2 สัปดาห์ขึ้นไป

ตารางที่ 5 จำนวนข้อความจากการโพสต์ของชุดทดสอบแบบจำลอง

บุคคลที่ไม่เป็นโรคซึมเศร้า	จำนวนข้อความ	บุคคลที่เป็นโรคซึมเศร้า	จำนวนข้อความ
1	459	1	1,021
2	3,190	2	3,199
3	3,228	3	1285
4	1,697	4	3,125
5	848	5	310

บุคคลที่ไม่เป็นโรคซึมเศร้า	จำนวนข้อความ	บุคคลที่เป็นโรคซึมเศร้า	จำนวนข้อความ
6	1,657	6	3,223
7	1,265	7	1,895
8	454	8	3,225
9	2,090	9	3,212
10	419	10	3,088
11	547	11	3,207
12	3,180	12	155
13	1,287	13	1,046
14	872	14	3,235
15	1,305	15	3,209
รวม	22,498		**34,435**
ร้อยละ	**39.52%**		**60.48%**

ตารางที่ 6 ตัวอย่างข้อความจากการโพสต์ของบุคคล

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ตัวอย่างข้อมูล
1	Id	รหัสการโพสต์	10327xxxxx460609
2	Created_At	วันและเวลาโพสต์ YYYY-MM-DD HH: MM	2018-08-24 03:18
3	Text	ข้อความที่โพสต์	we love u more https://t.co/y5hRxC1rSQ

ตารางที่ 7 ตัวอย่างข้อมูลสำหรับทดสอบแบบจำลอง

Id	Created_At	Text
986776763698xxxx00	2018/04/19	But then I resurrected it with a blow dryer it was a christening.
985360308720xxxx00	2018/04/15	Love you #BeyHive Have fun celebrating all that love in the desert. When it’s dry and we are thirsty, music will bring the water that quenches our

Id	Created_At	Text
		soul.
984528458234xxxx00	2018/04/13	So crazy!!! Happened in 2010! ????
983913140009xxxx00	2018/04/11	Don’t forget to watch!! I had so much fun performing for my pops! #EltonSalute #EltonREVAMP

3.1.2 การเตรียมข้อมูล

ชุดข้อมูลที่ได้จาก Twitter ประกอบด้วยแฮชแท็ก ผู้ใช้งาน การรีทวีต ข้อความแสดงอารมณ์ ที่อยู่เว็บไซต์ สัญลักษณ์พิเศษ อักษรที่ใช้ซ้ำทำให้ความหมายเปลี่ยน จึงจำเป็นต้องคัดเฉพาะข้อมูลที่เหมาะสำหรับนำใช้ในการทดลอง ขั้นตอนการเตรียมข้อมูล มีดังนี้

ขั้นตอนการทำความสะอาดชุดข้อมูลด้วยการคัดกรองสิ่งที่ไม่จำเป็นสำหรับการประมวลผลออกไป ในงานวิจัยนี้ ได้ดำเนินการลบข้อมูลต่อไปนี้ออก ได้แก่ 1) ข้อความที่ขึ้นต้นด้วย “@” เนื่องจากเป็นชื่อผู้ใช้งาน 2) ข้อความที่ขึ้นต้นด้วย RT เนื่องจากเป็นข้อความที่เป็นการรีทวีต อักษรที่พิมพ์ซ้ำๆ เช่น “woowww” หากมีอักษรที่พิมพ์ติดกันซ้ำมากกว่า 2 ตัวอักษร จะถูกแทนที่ด้วยตัวอักษรเพียง 1 ตัวอักษร 3) ลบข้อความที่อยู่เว็บไซต์ 4) ข้อความแสดงอารมณ์ และ 5) สัญลักษณ์พิเศษ ข้อความเหล่านี้ล้วนไม่มีความเกี่ยวข้องกับการจำแนกโรคซึมเศร้า

ตารางที่ 8 ตัวอย่างการทำ Regular expression

ลำดับ	ข้อความก่อนการกรอง	ข้อความหลังการกรอง
1	@ArianaGrande @NBCTheVoice SO CUTEEE □□ https://t.co/pyMR778XQZ	SO CUTE
2	RT @NBCTheVoice: performance of just look up will leave you in awe. ??	performance of just look Up will leave you in awe
3	@NBCTheVoice Everything was just perfect wooowww	Everything was just perfect wow

1) การตัดคำเป็นตอนสำหรับแยกคำศัพธ์ต่างๆ ให้เป็นคำเดียวจากประโยค ก่อนนำคำเหล่านั้นไปสู่ขั้นตอนต่อไป ในภาษาอังกฤษจะดำเนินการแยกคำด้วยช่องว่าง

ตารางที่ 9 ตัวอย่างการตัดคำ

ลำดับ	ข้อความก่อนการตัดคำ	ข้อความหลังการตัดคำ
1	SO CUTE	SO | CUTE
2	performance of just look up will leave you in awe.	performance| of | just | look up | will | leave | you | in | awe
3	Everything was just perfect wow	Everything | was | just | perfect | wow

2) การแปลงข้อความเป็นการเปลี่ยนตัวอักษรให้เป็นตัวอักษรพิมพ์เล็กทั้งหมด เช่น “SO CUTE” จะถูกแก้ไขเป็น “so cute” และ “Everything” จะถูกแก้ไขเป็น “everything”

3) การกำจัดคำหยุดเป็นขั้นตอนการคัดกรองและลบคำหยุดที่เป็นคำที่ไม่มีความหมายออกไป เช่น a, an, the, of, is, are, so, you เป็นต้น ในงานวิจัยนี้ใช้ชุดคำหยุด Default English stop words list (Stopwords, n.d.) เช่น “everything was just perfect wow” คำว่า “was” จะถูกกรองออก

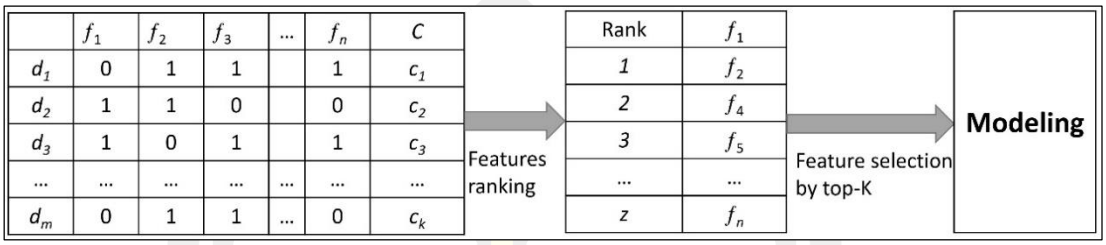
ตารางที่ 10 ตัวอย่างการคัดกรองและลบคำหยุด

ลำดับ	ข้อความก่อนการกรองคำหยุด	ข้อความหลังการกรองคำหยุด
1	so cute	cute |
2	performance of just look up will leave you in awe.	performance| just | look | up | leave | awe
3	everything was just perfect wow	everything | just | perfect | wow

4) การหารากคำศัพท์เป็นการตรวจสอบเพื่อหารากศัพท์ เพื่อการลดจำนวนกลุ่มคำที่มีความหมายเหมือนกัน จากนั้นนำชุดคำเข้าสู่กระบวนการทำถุงคำที่ใช้สำหรับการนับความถี่ที่อยู่ในเอกสาร

5) การสกัดคุณลักษณะ เป็นขั้นตอนการให้ค่าน้ำหนักแก่คำหรือคุณลักษณะ ในงานวิจัยนี้ ใช้วิธีการ Binary term occurrence หลักการทำงาน “หากพบคำที่มีในเอกสาร” จะถูกกำหนดค่าเป็น 1 “หากไม่พบคำในเอกสาร”จะกำหนดค่าเป็น 0 จากนั้นนับความถี่คำที่ถูกพบ และนำมาคำนวนค่าน้ำหนักให้แต่ละคำ ด้วยวิธีการ Information gain และจัดลำดับค่าน้ำหนักตามความสำคัญของคุณลักษณะ จากนั้นนำไปคัดเลือกคุณลักษณะตามลำดับของ Top-k เนื่องด้วยเป็น

วิธีการที่ใช้ช่วยลดเวลาการประมวลผล ใช้จำนวนคุณลักษณะน้อยแต่ไม่ลดประสิทธิภาพลง IG สามารถคัดกรองคำแม้จะมีค่าน้ำหนักต่ำมากได้ (Pintas et al., 2021)



ภาพที่ 5 ขั้นตอนการคัดกรองคุณลักษณะ

ตารางที่ 11 ตัวอย่างคุณลักษณะที่ได้รับการคัดเลือก

Feature id	Rank	Features
f_1	1	appetite
f_2	2	hunger
f_3	3	help
f_4	4	depressive
f_5	5	lose
f_6	6	tired

ขั้นตอนการแทนที่ค่าการเกิดคุณลักษณะในถุงคำ หากพบคุณลักษณะในถุงคำ จะให้ค่า 1 หากไม่พบจะให้ 0 โดยขนาดของเวกเตอร์เท่ากับ ขนาดของเอกสาร x ขนาดของคุณลักษณะ จากตัวอย่างตารางที่ 11 มีขนาดของเวกเตอร์เท่ากับ 4x6 เช่น $\mathbf{d_1}$= “Sunday is so sad and **depressive**” จะพบคำว่า “depressive” ดังนั้นจะให้ค่า t_3=1 ส่วนคำอื่นที่ไม่พบ จะให้ค่า 0

ตารางที่ 12 ตัวอย่างการแทนที่ค่าในเอกสาร

	t_1	t_2	$*t_3$	t_4	t_5	t_6
d_1	**0**	**0**	**1**	**0**	**0**	**0**
d_2	0	0	1	1	0	0
d_3	1	0	0	0	1	0
d_4	0	1	0	0	0	1

จากนั้นนำมาคำนวณเค่า IG เพื่อจัดลำดับความสำคัญให้คุณลักษณะ โดยคัดเลือกคุณลักษณะจากค่าน้ำหนักด้วยวิธีการ Top-k จากค่าน้ำหนักที่สูงสุด จนถึงจำนวน k

3.1.3 การเรียนรู้แบบจำลอง

ขั้นตอนนี้จะสร้างแบบจำลองตามจำนวนของคุณลักษณะ ตั้งแต่ 200 – 6,000 คุณลักษณะ โดยเลือกคุณลักษณะเรียงลำดับจากค่าน้ำหนักตาม Top-k แล้วนำไปสร้างแบบจำลองด้วยตัวจำแนกประเภทนาอีฟเบย์ ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด และแรนดอมฟอเรสต์ จำนวน 93 แบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพและเลือกแบบจำลองที่เหมาะสมที่สุด โดยพิจารณาจากค่าความถูกต้องกับเวลาที่ใช้ประมวลทั้ง 3 ตัวจำแนก เพื่อให้ค่าน้ำหนักสำหรับวิธีการเอ็นเซมเบิลแบบกำหนดน้ำหนักแบบ TPW และ APW

TPW ให้ค่าน้ำหนักด้วยอัตราการทำนายถูกต้อง (True positive-rate: TP-rate) ของคลาสคำตอบ คำนวณได้จากสมการ (31) และ (32) (Rojarath & Songpan, 2021) และ APW ให้ค่าน้ำหนักด้วยค่าความน่าจะเป็นการเกิดคลาสคำตอบ คำนวณจาก (33) (Rojarath & Songpan, 2020)

$$\text{TP-rate} = \frac{\textit{True positive class } n}{\textit{Total member of class } n} \tag{31}$$

เมื่อ n คือ คลาสคำตอบ n= {1, 2, 3, ..., 10}

$$\text{TPW class} = \textit{Probability}_i \; x \; \textit{TP-rate}_n \tag{32}$$

$$\text{APW class} = \text{weight max} \frac{\sum_{n=c}^{class} \left(P \; x \; (T_r)\right)_w}{C} \tag{33}$$

โดยที่

P คือ ค่าความน่าจะเป็นในแต่ละคลาส

T_r คือ ค่าน้ำหนักของคลาส n ที่การทำนายถูกต้อง (TP-rate)

w คือ สมาชิกแต่ละ Instance

C คือ จำนวนตัวจำแนก

ขั้นตอนการคำนวณค่าน้ำหนัก TPW และ APW ดังแสดงใน Algorithm 1 ขั้นตอนการนำค่าน้ำหนักไปใช้เพื่อทำนายคลาส ดังแสดงใน Algorithm 2 และขั้นตอนการคำนวณคะแนนเพื่อประเมิณการเป็นโรคซึมเศร้า ดังแสดงใน Algorithm 3

Algorithm 2: Calculation the TPW and APW vote ensemble

Input: Training set

1: Given a set of the classifiers $c=(c_1, c_2, \ldots c_k)$, a set of the member $m=(m_1, m_2, \ldots)$,
a set of the class $n=(n_1, n_2, \ldots, n_l)$

2: **for** $c \leftarrow 1$ **to** l **do**

3: **for** $n \leftarrow 1$ **to** l **do**

4: $TPrate_n \leftarrow (TP_n) / (\text{total\_member}_n)$

5: **for** $l \leftarrow 1$ **to** m **do**

6: $TPW_n \leftarrow Probability_m \times TPrate_n$

7: $newTPW_n \leftarrow TPW_n$

8: $APW_n \leftarrow sum\ (Probability_m \times TPrate_n) / c$

9: $newAPW_m \leftarrow \max (APW_n)$

10: **end for**

11: **end for**

12: **end for**

Output: the newTPW and newAPW

3.1.4 การทดสอบแบบจำลอง

หลังจากได้ค่าน้ำหนักแต่ละแบบจำลอง จากนั้นนำค่าน้ำหนักที่ได้มาทำขั้นตอนทดสอบแบบจำลองเพื่อทำนายคลาสของแต่ละแถวข้อมูลของแต่ละผู้ใช้งาน (z_l) โดยใช้ชุดข้อมูลสำหรับทดสอบแบบจำลอง ดัง Algorithm 3

Algorithm 3: Calculating the final prediction for the weighted voting ensemble

Input: Test set

1: Given a set of the classifiers $c=(c_1, c_2, \ldots, c_k)$,
a set of the member of the test set $z=(z_1, z_2, \ldots, z_l)$,
a set of the class $n=(n_1, n_2, \ldots, n_l)$

2: **for** $z \leftarrow 1$ to l **do**

3: **for** $i \leftarrow 1$ to c **do**

4: **for** $j \leftarrow 1$ to m **do**

5: $newTPW\_probability_m \leftarrow Probability_m \times newTPW_n$

6: $newAPW\_probability_m \leftarrow Probability_m \times newAPW_n$

7: $pred\_TPW_j \leftarrow max\ (newTPW\_probability_m)_n$

8: $pred\_APW_j \leftarrow max\ (sum\ (newAPW\_probability_m) / ((the\ number\ of\ c))$

9: **end for**

10: **end for**

11: **end for**

Output: the final prediction class of each instance of the member (z_l)

ตารางที่ 13 ตัวอย่างผลการทำนายคลาสของชุดข้อมูลสำหรับทดสอบแบบจำลองแบบหลายคลาส

C(1)	C(2)	C(3)	C(4)	C(5)	C(6)	C(7)	C(8)	C(9)	C(10)	Created-At
1	0	0	0	0	0	0	0	0	0	2018-04-23
1	0	0	0	0	0	0	0	0	0	2018-04-20
1	0	0	0	0	0	0	0	0	0	2018-04-04
0	1	0	0	0	0	0	0	0	0	2018-04-03
0	1	0	0	0	0	0	0	0	0	2018-03-11

ผลการทำนายคลาสของชุดข้อมูลสำหรับทดสอบ หากการทำนายตรงกับอาการของคลาสนั้น จะกำหนดค่าเท่ากับ 1 หากไม่ใช่จะกำหนดค่าเท่ากับ 0 จากนั้นนำข้อมูลที่ได้ไปวิเคราะห์ภาวะโรคซึมเศร้า

3.1.5 การวิเคราะห์ภาวะโรคซึมเศร้า

เมื่อได้คลาสคำตอบของชุดข้อมูลสำหรับทดสอบแบบจำลองครบทั้ง 30 คน จากนั้นนำผลลัพธ์ที่ได้ทำประเมินการมีสภาวะเป็นโรคซึมเศร้า โดยนำความถี่ที่เกิดขึ้นในแต่ละวันของคลาสอาการโรคซึมเศร้า (คลาส 1-9) มาคำนวณและประเมิน แต่ไม่นำคลาสที่ 10 มาคำนวณเพราะเป็นคลาสอาการบุคคลปกติ ดัง Algorithm 4

Algorithm 4: Calculating the score of the depressive symptom

Input: The final prediction class of each instance of the member

1: Given a set of the member of the test set $z = (z_1, z_2, ..., z_l)$

a set of the final prediction class of each instance m

a set of dates of the month $date\_time = (date\_time_1, date\_time_2, ...)$

a set of score of the depressive symptom of each instance s = (s1, s2, ...)

2: **for** $z \leftarrow 1$ **to** l **do**

3: **for** $j \leftarrow 1$ **to** m **do**

4: **for** $date\_time \leftarrow 1$ **to** 14 **do**

5: **if** $prediction\_class_j$ == *depressive_ symptoms* **then**

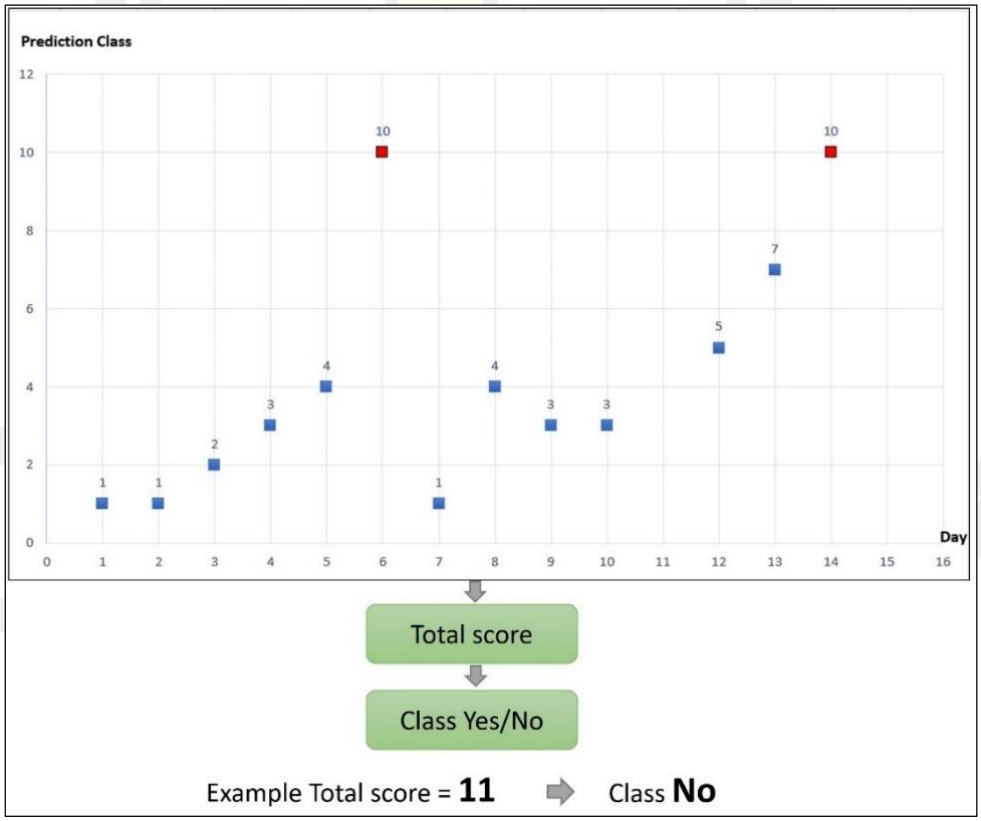
6: $total\_score(z_l) \leftarrow$ sum($score(s_j)$)

7: **end if**

8: **end for**

9: **if** *total_score*(z_l) <= 27 **and** *total_score*(z_l) >= 13 **then**
10: z_l ← *true*
11: **else**
12: z_l ← *false*
13: **end for**
14: **end for**
Output: the final prediction of the member (z_l)

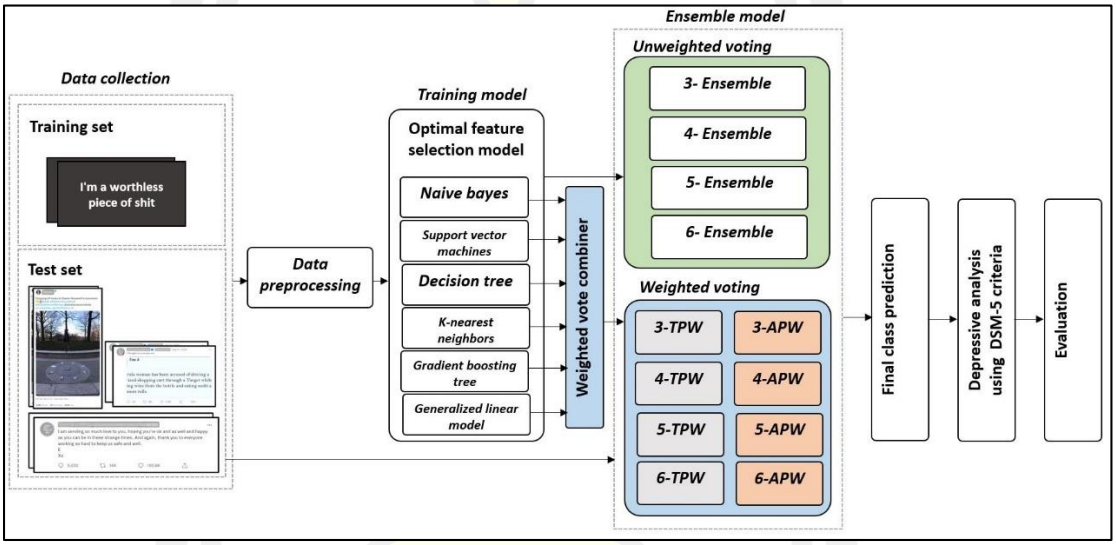
จาก Algorithm 4 ได้อธิบายถึงการนำ Prediction class มาคำนวณคะแนนตามมาตราฐานของ DSM-5 criteria (บรรทัดที่ 4-8) ในระยะเวลา 2 สัปดาห์ที่ต่อเนื่อง อาการของโรคซึมเศร้าทั้ง 9 อาการหรือ 9 คลาส จะถูกนำมาคำนวณคะแนน แต่คลาส 10 หรืออาการของคนทั่วไปจะไม่ถูกนำมาคำนวณคะแนน จากนั้นนำคะแนนรวมมาประเมินว่ามีภาวะซึมเศร้า (บรรทัดที่ 9-13) หรือไม่ โดยผลการประเมินหรือผลลัพธ์ที่ได้ คือ หากบุคคลที่ได้คะแนนรวมตั้งแต่ 13 คะแนน (z_i = true) บุคคลนั้นมีภาวะซึมเศร้าโดยกำหนดให้เป็นคลาส Yes หากคะแนนรวมน้อยกว่า 13 คะแนน (z_i = false) บุคคลนั้นปกติหรือไม่มีภาวะซึมเศร้าโดยกำหนดให้เป็นคลาส No



ภาพที่ 6 แสดงภาพรวมการได้มาของคลาสการประเมินภาวะซึมเศร้า

ดังภาพที่ 6 ในวันที่ 6 และ วันที่ 14 คลาสทำนายเป็นคลาสที่ 10 เป็นคลาสอาการของคน ทั่วไปไม่นำมาคำนวณคะแนน คะแนนรวมเท่ากับ 11 คะแนน ดังนั้นจึงเป็นคลาส No ซึ่งไม่มีภาวะซึมเศร้า

3.2 การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ



ภาพที่ 7 ขั้นตอนการวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ

การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ มีวัตถุประสงค์และขั้นตอนการทำงาน ดังนี้

1. การหาจำนวนคุณลักษณะที่เหมาะสมสำหรับการเรียนรู้แบบจำลอง ตัวจำแนกประเภทที่ใช้ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ นาอีฟเบย์ กราเดียนบูทติ้งทรี และตัวแบบเชิงเส้นนัยทั่วไป สำหรับชุดข้อมูลไบนารีคลาส

2. การปรับปรุงค่าน้ำหนัก 2 วิธีการ คือ อัตราการทำนายถูกของคลาสคำตอบ และค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ จากผลลัพธ์ที่ได้จาก ชุดข้อมูลสำหรับเรียนรู้แบบจำลองการหาค่าน้ำหนักให้แต่ละวิธีการ จากตัวจำแนกประเภทสำหรับสร้างแบบจำลองการเรียนรู้ จำนวน 3, 4, 5, และ 6 ตัวจำแนก โดยกำหนดแต่ละกลุ่มของตัวจำแนกประเภทเพื่อให้ค่าน้ำหนักตามประสิทธิภาพลำดับจากมากไปน้อย ดังนี้

1) จำนวน 3 ตัวจำแนก ประกอบด้วย ซัพพอร์ตเวกเตอร์แมชชีน ตัวแบบเชิงเส้นนัยทั่วไป และนาอีฟเบย์

2) จำนวน 4 ตัวจำแนก ประกอบด้วย ซัพพอร์ตเวกเตอร์แมชชีน ตัวแบบเชิงเส้นนัยทั่วไป นาอีฟเบย์ และกราเดียนบูทติ้งทรี

3) จำนวน 5 ตัวจำแนก ประกอบด้วย ซัพพอร์ตเวกเตอร์แมชชีน ตัวแบบเชิงเส้นนัยทั่วไป นาอีฟเบย์ กราเดียนบูทติ้งทรี และต้นไม้ตัดสินใจ

4) จำนวน 6 ตัวจำแนก ประกอบด้วย ซัพพอร์ตเวกเตอร์แมชชีน ตัวแบบเชิงเส้นนัยทั่วไป นาอีฟเบย์ กราเดียนบูทติ้งทรี ต้นไม้ตัดสินใจ และขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด

3. การเปรียบเทียบผลลัพธ์ระหว่างตัวจำแนกประเภทแบบเดี่ยว วิธีการเอ็นเซมเบิลแบบไม่กำหนดน้ำหนัก และวิธีการเอ็นเซมเบิลแบบกำหนดน้ำหนัก

ขั้นตอนการดำเนินการ ประกอบด้วย ดังนี้

3.2.1 การรวบรวมข้อมูล

งานวิจัยนี้ผู้วิจัยได้เก็บรวบรวมข้อมูล ซึ่งนำมาจาก Twitter และ Instagram ประกอบด้วยชุดข้อมูล 2 ชุด ดังนี้

ชุดข้อมูลที่ 1 สำหรับการเรียนรู้แบบจำลอง รวบรวมภาพที่มีข้อความภาษาอังกฤษในภาพจากแฮชแท็กบน Twitter และ Instagram ของผู้ใช้งานทั่วไป ประกอบด้วย 9 อาการ ที่ได้บ่งบอกถึงอาการของโรคซึมเศร้า และ 1 อาการที่บ่งบอกถึงบุคคลปกติ จำนวน 5,300 ภาพ โดยกำหนดให้ 9 อาการเป็นคลาสคำตอบคนที่เป็นโรคซึมเศร้า (Depression class) และ 1 อาการบุคคลปกติเป็นคลาสคนที่ไม่เป็นโรคซึมเศร้า (Non-depression class)

ชุดข้อมูลที่ 2 สำหรับทดสอบแบบจำลอง รวบรวมจากการโพสต์ข้อความภาษาอังกฤษและภาพที่ถูกโพสต์โดยนักแสดง นักร้องและผู้ที่มีชื่อเสียง บน Twitter จำนวน 47 คน ประกอบด้วยบุคคลที่เปิดเผยว่าเป็นโรคซึมเศร้า จำนวน 27 คน และ บุคคลที่ไม่เป็นโรคซึมเศร้า จำนวน 20 คน โดยผู้ใช้งานแต่ละคนต้องโพสต์ข้อความต่อเนื่องกันในระยะเวลา 2 สัปดาห์ขึ้นไป

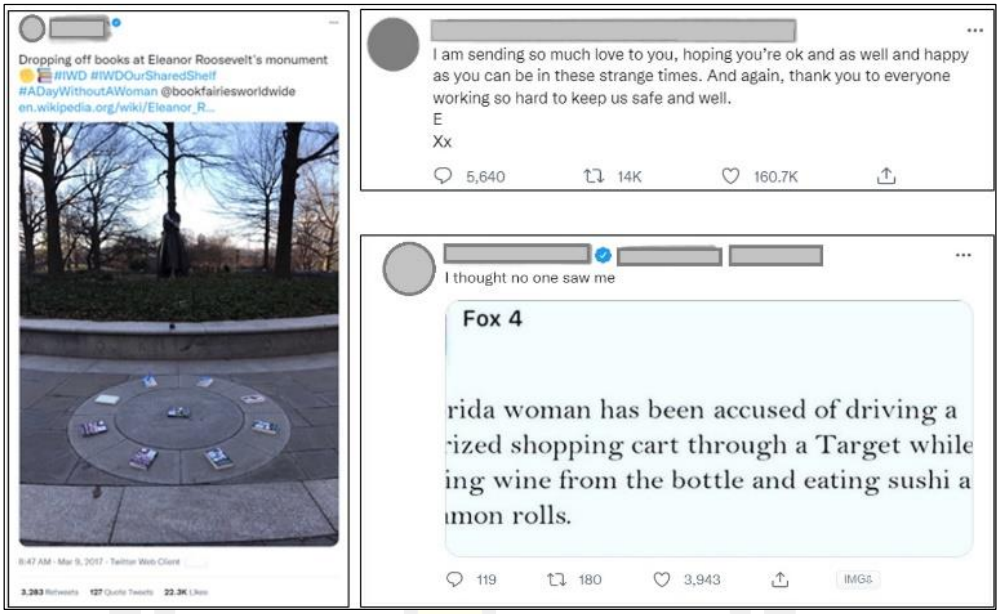
ตารางที่ 14 แสดงจำนวนภาพจากการโพสต์ของชุดทดสอบ

ลำดับ	จำนวน (คน)	ภาพ	ข้อความ
คนที่เป็นโรคซึมเศร้า	27	20,830	17,415
คนที่ไม่เป็นโรคซึมเศร้า	20	9,511	14,501
รวม	47	30,341	31,916

ตารางที่ 15 แสดงจำนวนภาพจากแฮชแท็กตามลักษณะอาการ

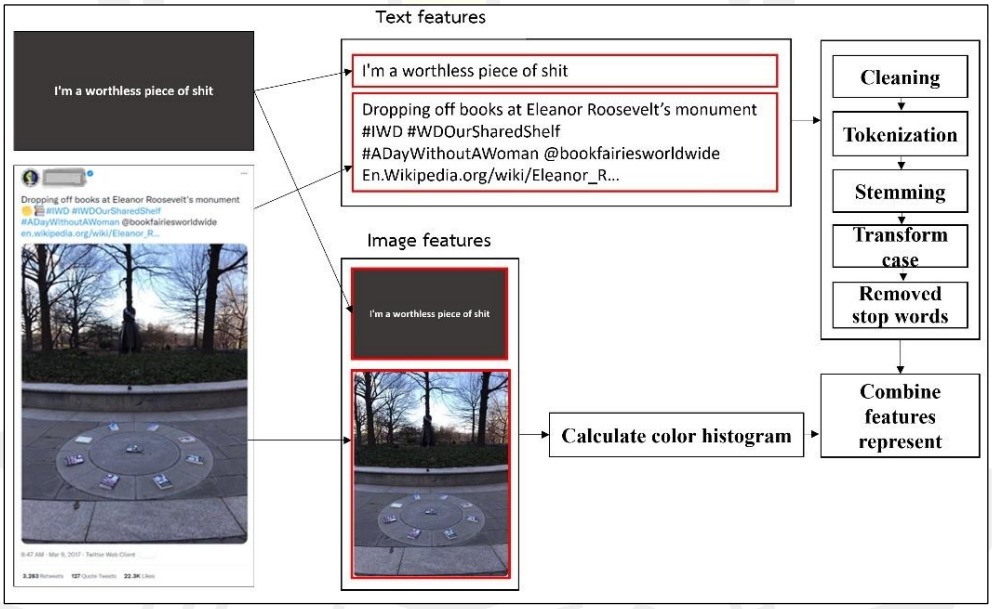
ลำดับ	อาการ	แฮชแท็ก	จำนวนภาพ	คลาส
1	อารมณ์ซึมเศร้า	#sadness #depressive	357	depression
2	ความสนใจลดลง	#loss of interest #lose interest	252	
3	น้ำหนักผิดปกติ	#appetite #hunger	842	
4	การนอนผิดปกติ	#sleepless #hethargy	317	
5	สมาธิสั้น	#un thinking #out thinking	299	
6	รู้สึกไร้ค่า	#guilt #disgrace #dishonor	295	
7	ร่างกายอ่อนเพลีย	#tired #bored #fatigued	263	
8	การเคลื่อนไหวช้า	#lackadaisical #lazy #loafing #phlegmatic	281	
9	การอยากฆ่าตัวตาย	#suicidal #dangerous #destructive	293	
10	ปกติ	#Happy #Enjoy #Smile	2,100	non-depression
รวม			5,300	

ภาพที่ 8 ตัวอย่างข้อมูลสำหรับการเรียนรู้แบบจำลอง



ภาพที่ 9 ตัวอย่างข้อมูลสำหรับทดสอบแบบจำลอง

3.2.2 การเตรียมข้อมูล



ภาพที่ 10 แสดงขั้นตอนการแยกคุณลักษณะ

ในงานวิจัยนี้นำภาพที่นำมาทดสอบที่ประกอบด้วยข้อความและคุณลักษณะของสีภาพ ดังนั้นในขั้นตอนการเตรียมข้อมูลแบ่งออก 2 ส่วน ดังนี้

1) คุณลักษณะข้อความ (Text features) ในขั้นตอนนี้นำข้อความในภาพออกมา

เมื่อได้ข้อความตามในภาพที่ 10 จากนั้นนำเข้าสู่ขั้นตอนการเตรียมข้อมูลในส่วนของข้อความตามขั้นตอนและวิธีการที่ 3.1.1

2) คุณลักษณะสีของภาพ (Image features) ในขั้นตอนนี้ได้นำภาพมาประมวลผลเพื่อคำนวณหาค่าการกระจายสีของภาพด้วยวิธีการ RGB histogram โดยใช้ค่าเฉลี่ยการกระจายสีของสามสีหลัก ได้แก่ สีแดง สีเขียว และสีน้ำเงิน มีช่วงค่าสีตั้งแต่ 0 ถึง 255 จากนั้นจะนำค่าเฉลี่ยการกระจายสีที่ได้เป็นคุณลักษณะของภาพนั้นๆ การคำนวนการกระจายสี ทฤษฎีทางสถิติที่ได้รับความนิยมนำมาใช้เพื่อหาคุณลักษณะสีของภาพ ได้แก่ ค่าเฉลี่ย (Mean) ส่วนเบี่ยงเบนมาตรฐาน (Standard deviation: SD) และค่าความเบ้ (Skewness) เป็นต้น (El-Bendary, Hariri, Hassanien, & Badr, 2015) ในงานวิจัยนี้ใช้ค่าเฉลี่ยของ สีแดง เขียว และน้ำหนัก การคำนวณหาค่าเฉลี่ยสีของสามสีหลัก จากสมการ (25) มีลำดับการประมวลผล ดัง Algorithm 5

Algorithm 5: Calculating the mean RGB histogram of images

Input: Image (a_j)

```
image ← imread(a_j)
meanR ← 0; meanG ← 0; meanB ← 0
[row, col, rgb] ← size(image)
channel_color ← 256
RGB ← ['r','g','b']
if len(shape(image)) == 3
    then img ← rgb2gray(image)
    else img ← img
end if
histogram(img) ← zeros(rgb, channel_color)
for color ← 1: rgb
    for M ← 1 : row
        for N ← 1 : col
            for i ← 0 : (channel_color - 1)
                if img (M, N, color) == i
                    histogram(color, i + 1) ← histogram(color, i + 1) + 1
                end if
            end for
        meanR ← sumRed(r, i+1)/(M*N)
        meanG ← sumGreen(g, i+1) /(M*N)
        meanB ← sumBlue(b, i+1) /(M*N)
```

23: meanRGB ← *meanRGB*(meanR, meanG, meanB)
24: **end for**
25: **end for**
26: **end for**
Output: *meanRGB(Red, Green, Blue) value of the image(a_j)*

รวมชุดข้อความ และ RGB histogram เข้าด้วยกันได้แก่ ค่าเฉลี่ยสีแดง ค่าเฉลี่ยสีเขียว ค่าเฉลี่ยสีน้ำเงิน และข้อความ จากนั้นทำการปรับช่วงขอบเขตแต่ละคุณลักษณะ (Data normalization) ให้มีขนาดสเกลเดียวกันทุกคุณลักษณะและให้ค่าน้ำหนักแก่คุณลักษณะด้วยวิธีการ Binary occurrence และคัดเลือกคุณลักษณะที่จะนำไปใช้งานด้วยวิธี Information gain โดยคัดเลือกตาม Top-k โดยนำ ข้อความ ค่าเฉลี่ยสีแดง ค่าเฉลี่ยเขียว ค่าเฉลี่ยน้ำเงิน และคลาส ไปใช้สำหรับสร้างการเรียนรู้แบบจำลอง โดยในชุดข้อมูลสำหรับการเรียนรู้แบบจำลองจะคัดเลือกมาเฉพาะภาพที่มีข้อความในภาพมาเท่านั้น

ตารางที่ 16 ตัวอย่างคุณลักษณะภาพ

ลำดับ	ชื่อตัวแปร	คำอธิบาย	ตัวอย่างข้อมูล
1	Id	ลำดับภาพ	10xxxxx
2	Text	ข้อความจากภาพ	Smile at strangers and you just might change a life
3	MeanRed	ค่าเฉลี่ยสีแดง	221.42
4	MeanGreen	ค่าเฉลี่ยสีเขียว	112.60
5	MeanBlue	ค่าเฉลี่ยสีน้ำเงิน	112.93
6	Class	คลาส	non-depression (0)

ตารางที่ 17 ตัวอย่างข้อมูลสำหรับการเรียนรู้แบบจำลอง

No	Text	Mean red	Mean green	Mean blue	Label
1	A single wish that has never come true	46.32	56.16	92.19	1
2	You look pretty but you sound like a lie	90.39	106.33	114.35	1
3	because right now death sounds so	87.23	113.36	127.65	1

No	Text	Mean red	Mean green	Mean blue	Label
	Peaceful				
4	It's not you it's my anxiety	30.23	27.57	28.45	1
5	I guess I just want a love story in the end	76.40	97.05	113.74	1

3.2.3 การเรียนรู้แบบจำลอง

สร้างแบบจำลองตามจำนวนของคุณลักษณะ ตั้งแต่ 200 – 1,700 คุณลักษณะ โดยเลือกคุณลักษณะเรียงลำดับจากค่าน้ำหนักตาม Top-k แล้วนำไปสร้างแบบจำลองด้วยตัวจำแนกประเภท ซัพพอร์ตเวกเตอร์แมชชีน ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด ต้นไม้ตัดสินใจ นาอีฟเบย์ กราเดียนบูทติ้งทรี และตัวแบบเชิงเส้นนัยทั่วไป สำหรับไบนารีคลาส จำนวน 96 แบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพแต่ละแบบจำลองและเลือกแบบจำลองที่เหมาะสมที่สุด โดยพิจารณาจากค่าความถูกต้องกับเวลา เลือกแบบจำลองที่ให้ประสิทธิภาพที่ดีและใช้เวลาการประมวลที่เหมาะสมเพื่อนำมาให้ค่าน้ำหนักตาม Algorithm 1

3.2.4 การทดสอบแบบจำลอง

เมื่อได้ค่าน้ำหนักของแต่ละแบบจำลองจาก Algorithm 2 นำข้อมูลชุดที่ 2 จำนวน 47 คน มาทดสอบกับแบบจำลองดัง Algorithm 3 เพื่อให้แบบจำลองทำนายคลาสในแต่ละ Instance ของชุดทดสอบแต่ละบุคคล โดยที่ 1 หมายถึง มีอาการของในคลาสนั้นๆ และ 0 หมายถึง ไม่มีอาการในคลาสนั้น ซึ่ง class(1) เป็นคลาสที่มีอาการของโรคซึมเศร้า และ class(0) เป็นคลาสบุคคลปกติ

ตารางที่ 18 ตัวอย่างการทำนายคลาสของชุดข้อมูลสำหรับทดสอบแบบจำลองแบบไบนารีคลาส

Prediction class(1)	Prediction class(0)	Date Created
1	0	2018-06-07
1	0	2018-06-06
1	0	2018-05-24
1	0	2018-05-20
0	1	2018-05-21
1	0	2018-05-24
0	1	2018-05-25
0	1	2018-05-26

3.2.5 การวิเคราะห์ภาวะโรคซึมเศร้า

นำผลลัพธ์ที่ได้จากขั้นตอน 3.2.4 ซึ่งเป็นคลาสคำตอบของชุดข้อมูลสำหรับทดสอบแบบจำลองครบทั้ง 47 คน จากนั้นนำผลลัพธ์ที่ได้ทำการประเมินสภาวะการเป็นโรคซึมเศร้าตามขั้นตอนของ Algorithm 4

3.3 การวัดประสิทธิแบบจำลอง

งานวิจัยนี้วัดประสิทธิภาพ แบบจำลองด้วยค่า ความถูกต้อง ความแม่นยำ ความระลึก และค่าเฉลี่ยประสิทธิภาพโดยรวม และการทดสอบประสิทธิภาพของแบบจำลองด้วยวิธีการแบ่งข้อมูลเป็น 10 ส่วน (10-fold cross validation)

บทที่ 4
ผลการวิจัยและอภิปรายผล

การวิจัยเรื่อง “การปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น” มีวัตถุประสงค์เพื่อพัฒนากระบวนการปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่นด้วยวิธีการเอ็นเซมเบิล โดยการปรับปรุงค่าน้ำหนักที่เหมาะสมและวัดประสิทธิภาพการจำแนกด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ ผลการดำเนินการวิจัย และการอภิปรายผลการทดลอง มีรายละเอียด ดังนี้

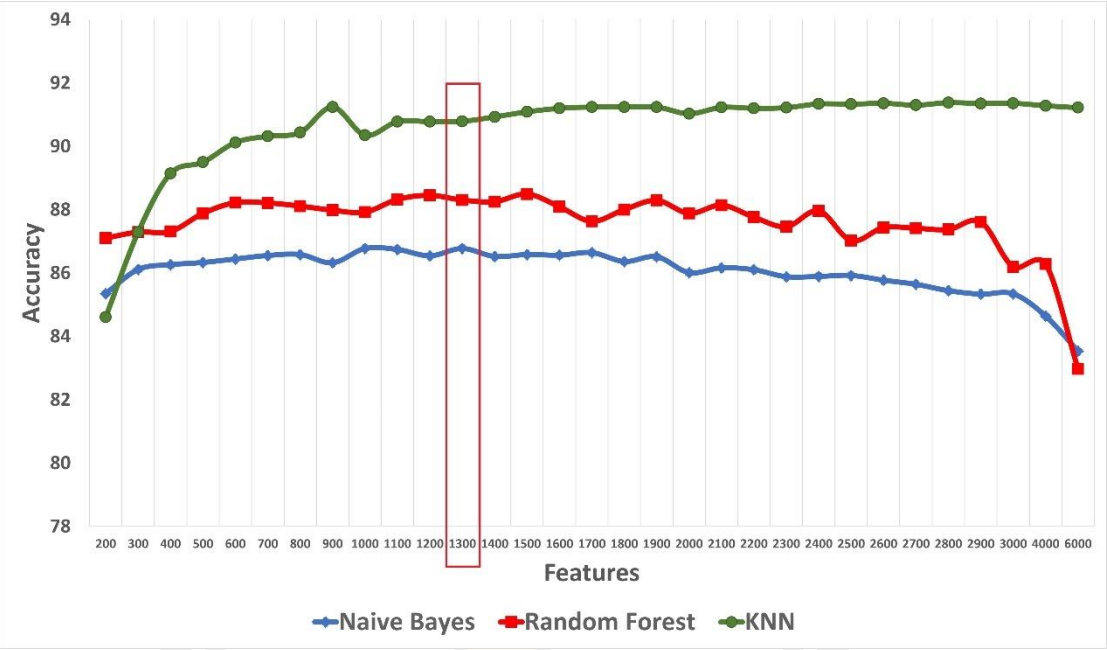
ตารางที่ 19 แสดงจำนวนกลุ่มการทดลองการจำแนกประเภทแบบเดี่ยว

การทดลอง	ตัวจำแนก	จำนวนกลุ่มคุณลักษณะ	จำนวนแบบจำลอง
การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล	3 ตัวจำแนก ได้แก่ NB, RF และ KNN	31	93
การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ	6 ตัวจำแนก ได้แก่ SVM, DT, NB, KNN, GBT และ GLMs	16	96

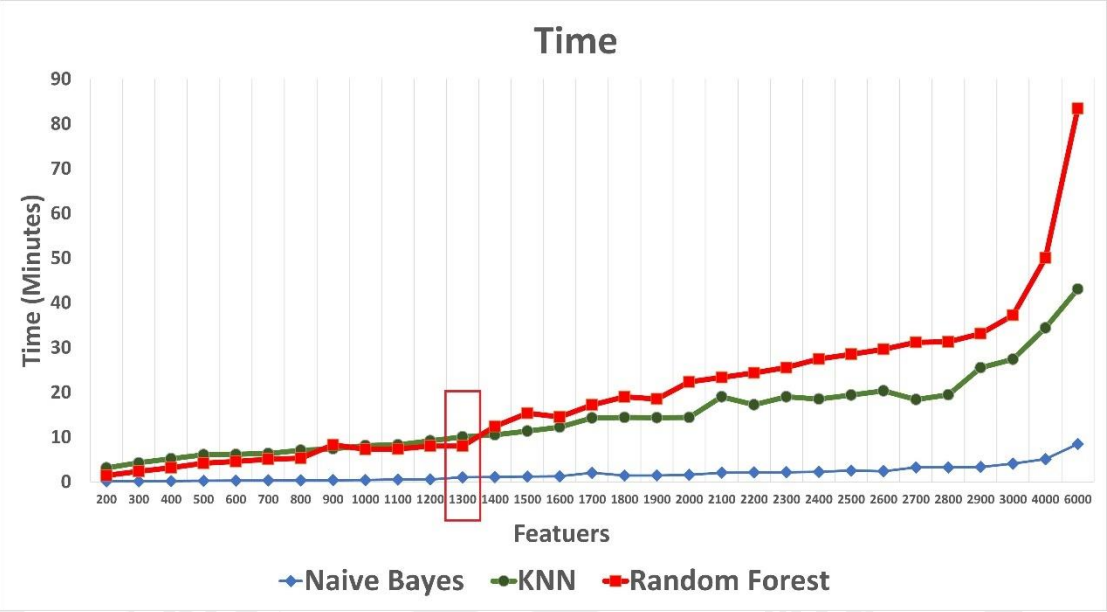
4.1 ผลการดำเนินการวิจัย

4.1.1 ผลการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล

1) การหาจำนวนคุณลักษณะที่เหมาะสมสำหรับการเรียนรู้แบบจำลอง จำนวน 93 แบบจำลอง ประกอบด้วย แบบจำลองใช้จำนวนคุณลักษณะ เริ่มจาก 200-6,000 คุณลักษณะ ด้วยตัวจำแนกประเภท นาอีฟเบย์ แรนดอมฟอเรสต์ และขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด



ภาพที่ 11 กราฟแสดงค่าความถูกต้องตัวจำแนกประเภทแบบเดี่ยวของกรอบการวิจัยที่ 1



ภาพที่ 12 กราฟแสดงเวลาประมวลผลตัวจำแนกประเภทแบบเดี่ยวของกรอบการวิจัยที่ 1

จากผลลัพธ์ภาพที่ 11 และ 12 พบว่า แบบจำลองที่สร้างด้วยจำนวนคุณลักษณะ 1,300 คุณลักษณะมีความถูกต้องที่ดีและใช้เวลาประมวลผลที่เหมาะสมทั้ง 3 ตัวจำแนก ดังนั้นจึงนำแบบจำลองนี้เป็นแบบจำลองสำหรับการเรียนรู้แบบจำลอง โดยนำค่าความน่าจะเป็นในแต่ละคลาสมา

คำนวนเพื่อหาอัตราการทำนายถูกเพื่อให้ค่าน้ำหนักแต่ละคลาสตาม Algorithm 2 แก่วิธีการ TPW และ APW นำมาทดสอบประสิทธิภาพของแบบจำลอง ตาม Algorithm 3 ด้วยชุดข้อมูลสำหรับทดสอบแบบจำลอง จากนั้นทำการวิเคราะภาวะการเป็นโรคซึมเศร้าตาม Algorithm 4 จะได้ผลลัพธ์ว่า z_i มีภาวะซึมเศร้า และการนับจำนวนการทำนายถูกและผิดของแต่บุคคลทั้ง 30 คน เพื่อนำมาคำนวณหาค่า ความถูกต้อง ความแม่นยำ ความระลึก และค่าเฉลี่ยประสิทธิภาพโดยรวม

ตารางที่ 20 ความถูกต้องและเวลาประมวลผลตัวจำแนกประเภทแบบเดี่ยวของกรอบการวิจัยที่ 1

Features	Accuracy			Time (Minute)		
	NB	RF	KNN	NB	RF	KNN
200	85.35	87.10	84.61	0.1	1.44	3.12
300	86.11	87.29	87.29	0.13	2.34	4.26
400	86.26	87.31	89.14	0.18	3.18	5.19
500	86.33	87.88	89.5	0.24	4.2	6.09
600	86.44	88.22	90.12	0.29	4.58	6.15
700	86.55	88.21	90.32	0.34	5.07	6.32
800	86.58	88.11	90.44	0.37	5.28	7.06
900	86.32	87.98	91.24	0.38	8.27	7.38
1000	86.77	87.92	90.35	0.4	7.27	8.1
1100	86.74	88.32	90.78	0.51	7.35	8.28
1200	86.54	88.45	90.78	0.55	8.02	9.18
1300	86.77	88.3	90.79	1.01	8.03	10.05
1400	86.52	88.25	90.93	1.09	12.38	10.52
1500	86.58	88.49	91.09	1.18	15.32	11.32
1600	86.56	88.09	91.2	1.25	14.51	12.25
1700	86.64	87.63	91.24	2.03	17.19	14.28
1800	86.36	88	91.24	1.4	19	14.38
1900	86.51	88.29	91.24	1.45	18.51	14.33
2000	86.01	87.88	91.03	1.58	22.37	14.42
2100	86.16	88.14	91.23	2.06	23.37	19.02
2200	86.1	87.76	91.2	2.11	24.34	17.22
2300	85.88	87.46	91.22	2.17	25.54	19.01

Features	Accuracy			Time (Minute)		
	NB	RF	KNN	NB	RF	KNN
2400	85.89	87.96	91.34	2.24	27.44	18.52
2500	85.92	87.02	91.33	2.52	28.53	19.42
2600	85.77	87.43	91.36	2.38	29.6	20.36
2700	85.64	87.41	91.3	3.25	31.19	18.41
2800	85.44	87.37	91.38	3.24	31.3	19.45
2900	85.33	87.61	91.35	3.34	33.12	25.51
3000	85.34	86.19	91.36	4.07	37.22	27.42
4000	84.64	86.28	91.28	5.09	50.06	34.43
6000	83.53	82.97	91.22	8.48	83.4	43.08

2) การเรียนรู้แบบจำลอง

ประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลอง ด้วยจำนวนคุณลักษณะ 1,300 คุณลักษณะ

ตารางที่ 21 ประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลองของกรอบการวิจัยที่ 1

Class	NB		RF		KNN	
	Precision	Recall	Precision	Recall	Precision	Recall
Depressive	87.90	91.07	**96.56**	89.83	84.08	96.30
Loss of interest	**98.95**	87.70	96.33	97.17	97.17	**98.60**
Appetite	94.30	87.13	**97.09**	90.17	92.82	95.20
Sleep	82.02	78.30	69.30	83.07	73.34	**91.07**
Thinking	91.97	86.23	**93.46**	89.53	91.49	78.17
Guilt	87.28	89.40	**96.54**	88.40	95.44	93.53
Tired	66.21	91.70	75.47	88.60	**94.28**	85.77
Movement	87.58	83.00	78.43	86.43	**95.23**	88.43
Suicidal	91.21	83.67	**96.46**	83.67	95.94	89.77
Normal	88.79	87.63	**95.56**	87.60	95.02	91.63
Accuracy	86.77		88.30		**90.79**	

จากตารางที่ 21 แสดงประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลองด้วยจำนวนคุณลักษณะ 1,300 คุณลักษณะ พบว่าแบบจำลองที่สร้างจากตัวนำแจก KNN มีความถูกต้องสูงสุดและให้ค่าความระลึกสูงที่สุดที่คลาส Loss of interest แบบจำลองที่สร้างจากแรนดอมฟอเรสต์มีความแม่นยำในการทำนายแต่ละคลาสได้ดี โดยมีค่าความแม่นยำสูงสุดในแต่ละคลาส ได้แก่ Depressive, Appetite, Thinking, Guilt, Suicidal และ Normal และแบบจำลองที่สร้างจากนาอีฟเบย์ มีความแม่นยำในการทำนายคลาส Loss of interest สูงสุด เท่ากับ 98.95% วิธีการเพื่อนบ้านที่ใกล้ที่สุดมีความถูกต้องมากที่สุดเท่ากับ 90.79% แรนดอมฟอเรสต์มีความถูกต้องเท่ากับ 88.30% และนาอีฟเบย์มีความถูกต้องน้อยที่สุดเท่ากับ 86.77%

3) การทดสอบแบบจำลอง

ประสิทธิภาพการทดสอบแบบจำลองตัวจำแนกประเภทแบบเดี่ยว วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก และแบบกำหนดค่าน้ำหนัก

ตารางที่ 22 ประสิทธิภาพการทดสอบแบบจำลองของกรอบการวิจัยที่ 1

Models	Precision		Recall		F1		Accuracy	Time
	Yes	No	Yes	No	Yes	No		(Minute)
Single model								
NB	56.25	57.14	60.00	53.33	58.06	55.17	56.67	10:02
RF	60.00	52.00	20.00	**86.67**	30.00	65.00	53.33	15:58
KNN	47.37	45.45	60.00	33.33	52.94	38.46	46.67	60:46
Ensemble model								
Unweighted	45.00	45.00	60.00	26.67	51.43	32.00	43.33	72:13
TPW	61.11	66.67	73.33	53.33	66.67	59.26	63.33	-
APW	63.16	72.73	**80.00**	53.33	**70.59**	61.54	**66.67**	-

จากตารางที่ 22 แสดงประสิทธิภาพการทดสอบแบบจำลอง พบว่าแบบจำลองที่สร้างวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักแบบ TPW และ APW มีความถูกต้องมากกว่า วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก โดยวิธีการกำหนดค่าน้ำหนักแบบ APW มีค่าความถูกต้องสูงสุด ที่ 66.67% มีค่าความระลึกสูงสุดที่ 80.00% และค่าเฉลี่ยประสิทธิภาพโดยรวม 70.59% ในคลาสของบุคคลคนที่เป็นโรคซึมเศร้า แบบจำลองที่สร้างด้วยแรนดอมฟอเรสต์ให้ค่าความระลึกสูงสุดที่ 86.67% ในคลาสของบุคคลคนที่ไม่เป็นโรคซึมเศร้า โดยวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักแบบ TPW และ APW คำนวณจาก Algorithm 1- Algorithm 4 ที่ได้ค่าน้ำหนักมาจากตัวจำแนกประเภทแบบเดี่ยว จึงไม่แสดงเวลาในการประมวลผล

จากนั้นนำผลลัพธ์ของการทดสอบแบบจำลองมาวิเคราะห์ความแตกต่างระหว่างวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักและแบบไม่กำหนดค่าน้ำหนักด้วยวิธีการทางสถิติ Paired samples t-test เพื่อทดสอบสมมติฐานทางสถิติ วัดระดับความเชื่อมั่น และระดับนัยสำคัญของประสิทธิภาพในแต่ละวิธีการ กำหนดให้ H0 คือ ประสิทธิภาพของวิธีการเอ็นเซมเบิลแบบไม่กำหนดค่า H1 คือ ประสิทธิภาพของวิธีการเอ็นเซมเบิลแบบกำหนดค่า โดยวัดระดับความเชื่อมั่นที่ 95% และระดับนัยสำคัญที่ 0.05

ตารางที่ 23 ผลการวิเคราะห์ Paired samples t-test

Class	Methods	Mean	S.D.	SEM	95% confidence interval of this difference		t	Sig (2-tailed)
					Lower	Upper		
Yes	Unweighted	52.143	7.525	4.344	-18.426	-11.360	18.140	0.0030
	TPW	67.037	6.118	3.532				
	Unweighted	52.143	7.525	4.344	-22.878	-13.336	35.926	0.0008
	APW	71.250	8.439	4.873				
No	Unweighted	34.557	9.429	5.444	-32.820	-17.573	14.221	0.0049
	TPW	59.756	6.684	3.859				
	Unweighted	34.557	9.429	5.444	-31.593	24.3604	33.286	0.0009
	APW	62.533	9.738	5.622				

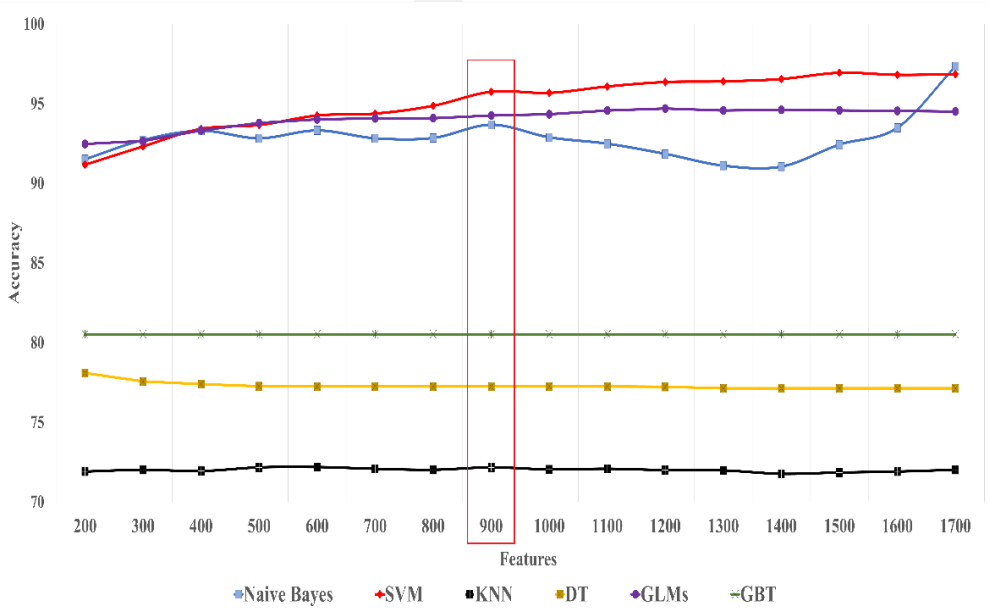
จากตารางที่ 23 แสดงเห็นว่าวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก มีประสิทธิภาพมากกว่าวิธีการแบบไม่กำหนดค่าน้ำหนักอย่างมีนัยสำคัญที่ 0.05 ทั้งคลาสของบุคคลที่เป็นและไม่เป็นโรคซึมเศร้า สะท้อนให้เห็นว่าวิธีการปรับปรุงค่าหนักที่ได้นำเสนอมีผลลัพธ์ที่แตกต่างจากวิธีการแบบไม่กำหนดค่าน้ำหนัก เหมาะสมสำหรับนำไปประยุกต์ใช้ต่อหรือเพื่อพัฒนาให้ดียิ่ง

จากผลที่ได้ จึงนำวิธีการเอ็นเซมเบิ้ลแบบกำหนดค่าน้ำหนักไปทดสอบประสิทธิภาพกับชุดข้อมูลใหม่ในตามกรอบงานวิจัยที่ 2

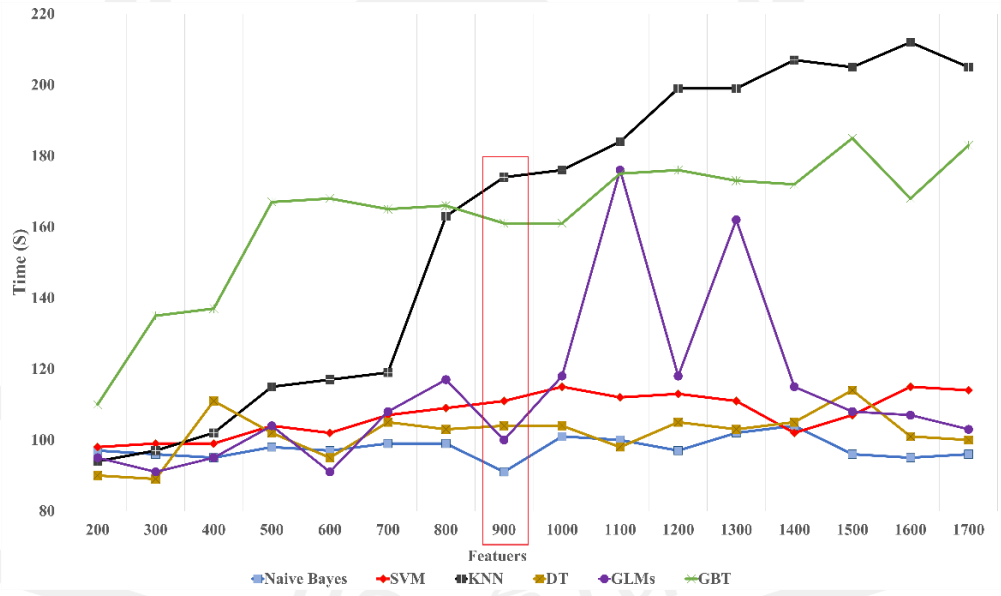
4.1.2 ผลการวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ

1) การหาจำนวนคุณลักษณะที่เหมาะสมสำหรับการเรียนรู้แบบจำลอง จำนวน 96 แบบจำลอง ประกอบด้วย แบบจำลองใช้จำนวนคุณลักษณะ เริ่มจาก 200-1,700 คุณลักษณะ ด้วยตัว

จำแนกประเภท ซัพพอร์ตเวกเตอร์แมชชีน ตัวแบบเชิงเส้นนัยทั่วไป นาอีฟเบย์ กราเดียนบูทติ้งทรี ต้นไม้ตัดสินใจ และขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุด



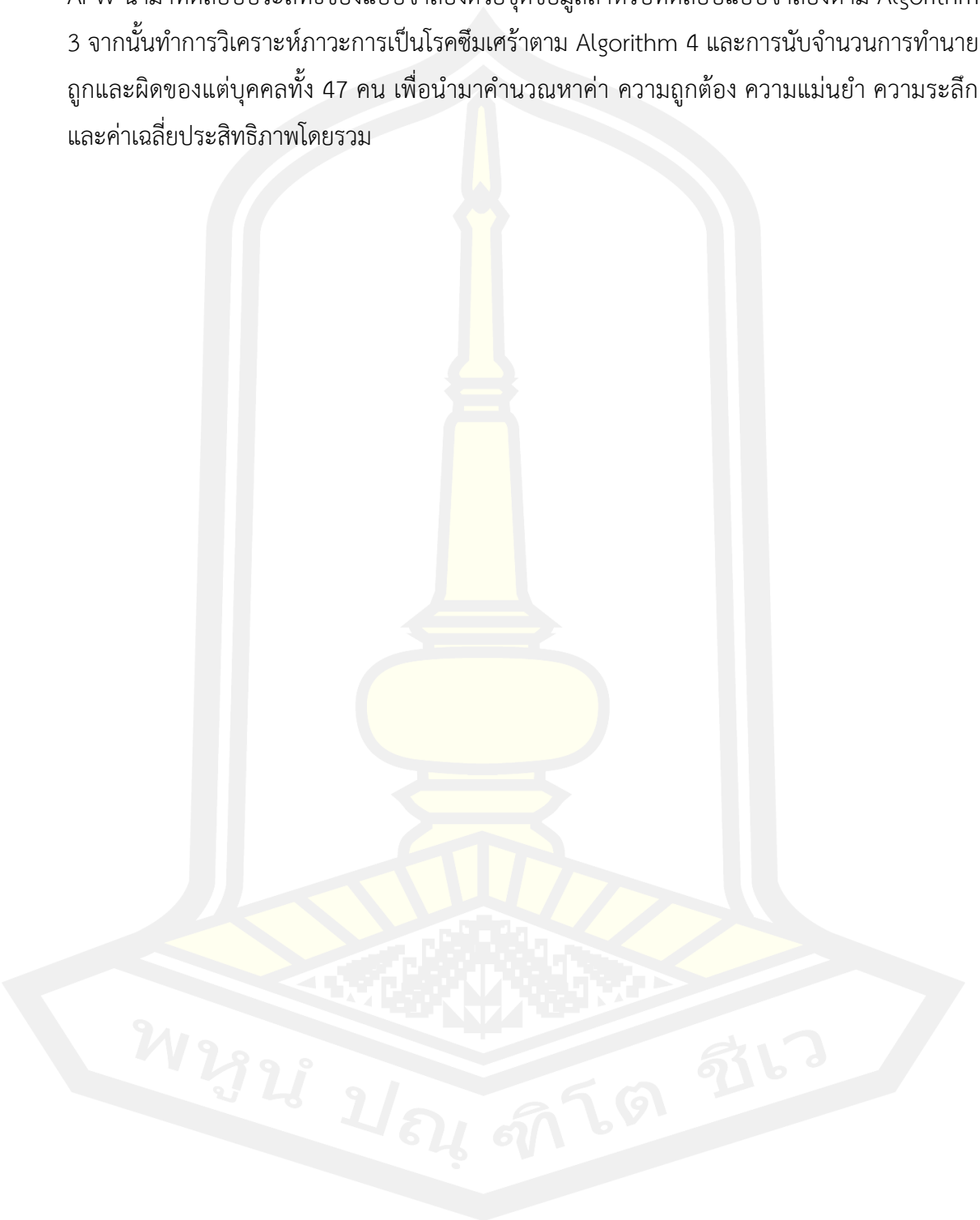
ภาพที่ 13 กราฟแสดงค่าความถูกต้องตัวจำแนกประเภทแบบเดี่ยวของกรอบการวิจัยที่ 2



ภาพที่ 14 กราฟแสดงเวลาประมวลผลตัวจำแนกประเภทแบบเดี่ยวของกรอบการวิจัยที่ 2

จากผลลัพธ์ภาพที่ 13 และ 14 พบว่า แบบจำลองที่สร้างด้วยจำนวนคุณลักษณะ 900 คุณลักษณะมีความถูกต้องที่ดีและใช้เวลาประมวลผลที่เหมาะสมจาก 6 ตัวจำแนก ดังนั้นนำแบบจำลองนี้มาสร้างแบบจำลองการเรียนรู้ โดยแต่นำค่าความน่าจะเป็นในแต่ละคลาสไปคำนวนเพื่อ

หาอัตราการทำนายถูกต้องเพื่อให้ค่าน้ำหนักแต่ละคลาสตาม Algorithm 2 แก่วิธีการ TPW และ APW นำมาทดสอบประสิทธิของแบบจำลองด้วยชุดข้อมูลสำหรับทดสอบแบบจำลองตาม Algorithm 3 จากนั้นทำการวิเคราะห์ภาวะการเป็นโรคซึมเศร้าตาม Algorithm 4 และการนับจำนวนการทำนายถูกและผิดของแต่บุคคลทั้ง 47 คน เพื่อนำมาคำนวณหาค่า ความถูกต้อง ความแม่นยำ ความระลึก และค่าเฉลี่ยประสิทธิภาพโดยรวม



ตารางที่ 24 ความถูกต้องและเวลาประมวลผลตัวจำแนกประเภทแบบเดี่ยวของกรอบการวิจัยที่ 2

Features	NB		SVM		KNN		DT		GLMs		GBT	
	Acc	Time(s)	Acc	Time(s)	Acc	Time(s)	Acc	Time(s)	Acc	Time(s)	Acc	Time(s)
200	91.51	97	91.17	98	71.91	94	78.09	90	92.47	95	80.53	110
300	92.72	96	92.32	99	72.02	97	77.57	89	92.66	91	80.53	135
400	93.30	95	93.42	99	71.94	102	77.40	111	93.32	95	80.53	137
500	92.83	98	93.66	104	72.17	115	77.26	102	93.79	104	80.53	167
600	93.32	97	94.26	102	72.19	117	77.25	95	94.02	91	80.53	168
700	92.83	99	94.38	107	72.08	119	77.25	105	94.08	108	80.53	165
800	92.85	99	94.87	109	72.02	163	77.25	103	94.09	117	80.53	166
900	93.68	91	95.75	111	72.17	174	77.25	104	94.26	100	80.53	161
1000	92.89	101	95.68	115	72.04	176	77.25	104	94.34	118	80.53	161
1100	92.49	100	96.08	112	72.08	184	77.25	98	94.57	176	80.53	175
1200	91.85	97	96.36	113	72.00	199	77.23	105	94.70	118	80.53	176
1300	91.11	102	96.40	111	71.98	199	77.13	103	94.57	162	80.53	173
1400	91.05	104	96.55	102	71.77	207	77.13	105	94.62	115	80.53	172
1500	92.43	96	96.94	107	71.85	205	77.13	114	94.58	108	80.53	185
1600	93.49	95	96.81	115	71.91	212	77.13	101	94.55	107	80.53	168
1700	97.34	96	96.85	114	72.02	205	77.13	100	94.51	103	80.53	183

2) การเรียนรู้แบบจำลอง

ประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลอง ด้วยจำนวนคุณลักษณะ 900 คุณลักษณะ

ตารางที่ 25 ประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลองของกรอบการวิจัยที่ 2

Classifiers	Precision		Recall		F1		Accuracy
	Yes	No	Yes	No	Yes	No	
SVM	**97.75**	92.91	95.16	92.91	**96.44**	94.75	**95.75**
GLMs	95.71	92.12	94.75	93.52	95.71	92.12	94.26
NB	97.89	88.22	91.50	**97.00**	94.59	**92.40**	93.68
GBT	86.51	72.84	80.19	80.95	83.23	76.68	80.49
DT	74.00	**93.97**	**98.00**	47.52	94.33	63.12	78.00
KNN	75.60	66.18	79.59	60.86	77.54	63.41	72.04

จากตารางที่ 25 แสดงประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลองด้วยจำนวนคุณลักษณะ 900 คุณลักษณะ พบว่าแบบจำลองที่สร้างจาก ซัพพอร์ตเวกเตอร์แมชชีน มีความถูกต้องสูงสุด มีค่าความแม่นยำในการทำนายแต่ละคลาสที่ดี รองลงมา คือ ตัวแบบเชิงเส้นนัยทั่วไปและนาอีฟเบย์ ตามลำดับ โดยแบบจำลองที่สร้างจากต้นไม้ตัดสินใจให้ค่าความระลึก สูงที่สุดที่คลาสที่เป็นโรคซึมเศร้า จากประสิทธิภาพของแบบจำลอง จึงนำมาประมวลผลตาม Algorithm 2 เพื่อให้ค่าน้ำหนักในแต่ละวิธีการและนำไปทดสอบประสิทธิภาพแบบจำลองด้วยชุดข้อมูลสำหรับทดสอบแบบจำลองตาม Algorithm 3 จากนั้นนำคลาสที่ได้แต่บุคคลของชุดข้อมูลสำหรับทดสอบแบบจำลองมาทำการวิเคราะภาวะการเป็นโรคซึมเศร้าตาม Algorithm 4 จะได้ผลลัพธ์ว่า z_i มีภาวะซึมเศร้า และการนับจำนวนการทำนายถูกและการทำนายผิดของแต่บุคคลทั้ง 47 คน เพื่อนำมาคำนวณหาประสิทธิภาพแบบจำลอง

3) การทดสอบแบบจำลอง

ประสิทธิภาพการทดสอบแบบจำลองแบบเดี่ยว วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก และวิธีการเอ็นเซมเบิลกำหนดค่าน้ำหนัก

ตารางที่ 26 ประสิทธิภาพการทดสอบแบบจำลองของกรอบการวิจัยที่ 2

Model	Precision		Recall		F1		Accuracy	Time (Minute)
	Yes	No	Yes	No	Yes	No		
Single model								
SVM	88.46	80.95	85.19	85.00	86.79	82.93	**85.11**	5.19

GLMs	85.19	80.00	85.19	80.00	85.19	80.00	82.98	5.40
NB	88.24	60.00	55.56	**90.00**	68.18	72.00	70.21	5.45
GBT	88.46	80.95	85.19	85.00	86.79	82.93	**85.11**	5.32
DT	88.46	80.95	85.19	85.00	86.79	82.93	**85.11**	**4.18**
KNN	81.25	54.84	48.15	85.00	60.47	66.67	63.83	37.45
Unweighted ensemble								
3-ensemble	85.71	84.21	88.89	80.00	87.27	82.05	85.11	32.39
4-ensemble	83.33	69.57	74.07	80.00	78.43	74.42	76.60	31.54
5-ensemble	85.71	84.21	88.89	80.00	87.27	82.05	85.11	31.36
6-ensemble	85.19	80.00	85.19	80.00	85.19	80.00	82.98	33.03
Weighted ensemble								
3-TPW	85.71	84.21	88.89	80.00	87.27	82.05	85.11	32.43
4-TPW	88.89	85.00	88.89	85.00	**88.89**	**85.00**	**87.23**	32.51
5-TPW	85.71	84.21	88.89	80.00	87.27	82.05	85.11	33.04
6-TPW	85.71	84.21	88.89	80.00	87.27	82.05	85.11	33.16
3-APW	85.71	84.21	88.89	80.00	87.27	82.05	85.11	31.29
4-APW	88.89	85.00	88.89	85.00	**88.89**	**85.00**	**87.23**	31.54
5-APW	83.33	88.24	**92.57**	75.00	87.72	81.08	85.11	32.39
6-APW	85.71	84.21	88.89	80.00	87.27	82.05	85.11	31.33

จากตารางที่ 26 แสดงประสิทธิภาพการทดสอบแบบจำลอง พบว่าแบบจำลองที่สร้างด้วยต้นไม้ตัดสินใจและกราเดียนบูทติ้งทรี มีประสิทธิภาพโดยรวมที่ดีขึ้นเมื่อเปรียบเทียบกับประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลองตามลำดับ ซึ่งตรงกันข้ามกับประสิทธิภาพของวิธีการตัวแบบเชิงเส้นนัยทั่วไป ที่ได้ประสิทธิภาพการทดสอบแบบจำลองลดลงมากหากเปรียบเทียบกับประสิทธิภาพแบบจำลองสำหรับการเรียนรู้แบบจำลอง จากวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก มีความถูกต้องมากกว่า วิธีการเอ็นเซมเบิลแบบ ไม่กำหนดค่าน้ำหนัก และแบบจำลองที่สร้างด้วยตัวจำแนกประเภทแบบเดี่ยว โดยวิธีการกำหนดค่าน้ำหนักด้วย 4-TPW และ 4-APW มีค่าความถูกต้องสูงสุด ที่ 87.23% วิธีการ 4-APW ใช้เวลาประมวลน้อยกว่าวิธีการ 4-TPW และวิธีการ 5-APW มีค่าความระลึกสูงสุดที่ 92.57% คลาสของบุคคลคนที่เป็นโรคซึมเศร้า แบบจำลองที่สร้างด้วยนาอีฟเบย์ ให้ค่าความระลึกสูงสุดที่ 90.00% ในคลาสของบุคคลคนที่ไม่เป็นโรคซึมเศร้า แบบจำลองที่

สร้างด้วยต้นไม้ตัดสินใจเป็นวิธีการที่ใช้เวลาประมวลน้อยที่สุด วิธีการเอ็นเซมเบิลแบบ ไม่กำหนดค่าน้ำหนัก 3-ensemble ใช้เวลาประมวลผลมากที่สุด

จากผลการทดลองวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก 4-TPW และ 4-APW มีความถูกต้องมากที่สุด เป็นผลจากการสร้างการเรียนรู้แบบจำลองด้วยการเพิ่มคุณลักษณะจากการคัดเลือกจำนวนคุณลักษณะที่เหมาะสม และการเลือกจำนวนตัวจำแนกประเภทตามประสิทธิภาพการทำนายคลาสคำตอบที่ถูกต้อง ทำให้การกำหนดค่าน้ำหนักเหมาะสำหรับตัวจำแนกประเภทจำนวน 4 ตัวจำแนก จึงทำให้ได้ผลลัพธ์ที่ดีกว่าการใช้ตัวจำแนกประเภทอื่น โดยเฉพาะคลาสของบุคคลที่ไม่เป็นโรคซึมเศร้า วิธีการแบบกำหนดค่าน้ำหนักด้วยจำนวน 4 ตัวจำแนก มีผลลัพธ์ที่ดีกว่าวิธีการกำหนดค่าน้ำหนักด้วยจำนวนตัวจำแนกประเภทอื่น

4.2 การอภิปรายผลการทดลอง

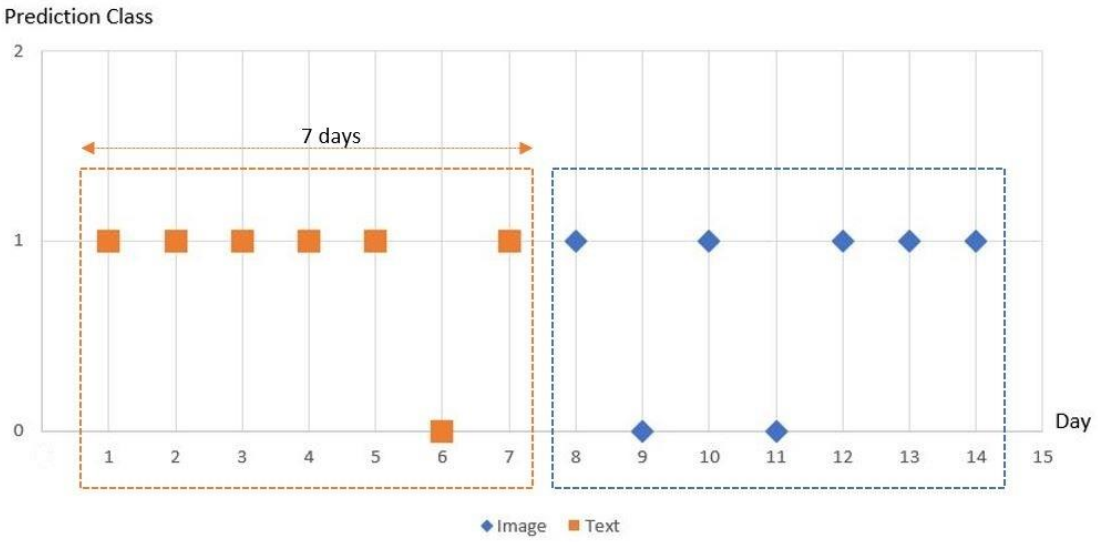
4.2.1 การปรับปรุงค่าน้ำหนักที่เหมาะสมและการปรับปรุงวิธีการเหมืองข้อมูลโดยใช้วิธีการเอ็นเซมเบิล

จากผลการทดสอบประสิทธิภาพแบบจำลองโดยวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนัก โดยวิธีการให้ค่าน้ำหนักจากอัตราการทำนายถูกของคลาสคำตอบ และค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ ทำให้ผลลัพธ์ที่ดีกว่า วิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก โดยเฉพาะการทดสอบกับชุดข้อมูลที่มีจำนวนคลาสคำตอบหลายคลาส เนื่องจากวิธีการให้ค่าน้ำหนักจากอัตราการทำนายถูกของคลาสคำตอบจะทำนายคำตอบจากค่าความน่าจะเป็นที่สูงที่สุด ดังนั้นโอกาสที่จะทำนายถูกต้องจึงมีมากกว่าวิธีการแบบไม่กำหนดค่าน้ำหนัก ซึ่งใช้หลักการทำนายด้วยวิธีการคะแนนเสียงข้างมาก และวิธีการกำหนดค่าน้ำหนักจากค่าเฉลี่ยสูงสุดของแต่ละคลาสคำตอบทำให้การทำนายมีความเป็นเสียรภาพมากกว่าวิธีการแบบไม่กำหนดค่าน้ำหนัก เห็นได้ชัดว่าวิธีการกำหนดค่าน้ำหนักทั้งสองวิธีการ ช่วยเพิ่มโอกาสการทำนายที่มีการทำนายผิดพลาดจากแบบจำลองแบบเดี่ยวและแบบจำลองแบบไม่กำหนดค่า เพราะทำให้การทำนายคลาสคำตอบใหม่มีโอกาสทำนายถูกต้องเพิ่มขึ้น และทำให้แบบจำลองมีประสิทธิภาพเพิ่มมากขึ้นอย่างมีนัยสำคัญ จึงเหมาะสำหรับนำไปปรับปรุงการสร้างแบบจำลองเพื่อช่วยเพิ่มประสิทธิภาพแบบจำลอง โดยวิธีการที่ได้นำเสนอนี้พบว่าอัตราการเกิดข้อมูลเกินจำนวนมาก (Overfitting) ค่อนข้างสูง แม้ว่าการให้ค่าน้ำหนักจากแบบจำลองการเรียนมีประสิทธิภาพที่ดี แต่ผลการวัดประสิทธิภาพการทดสอบแบบจำลองจากตารางที่ 22 พบว่าวิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ มีค่าความระลึกของคลาสบุคคลปกติและคนที่เป็นโรคซึมเศร้ามีความแตกต่างกันเนื่องด้วยจำนวนการโพสต์ต่อวันมีผลต่อการทำนายอาการของโรคซึมเศร้า จึงทำให้การทำนายอาจผิดพลาดหากมีการโพสต์จำนวนหลายครั้งต่อวัน โดยเฉพาะบุคคลที่ไม่เป็นโรคซึมเศร้าที่มีการแสดง

ความเสียใจเหตุการ์ที่เกิดขึ้นจาก ข่าวสาร กิจกรรม เหตุการณ์สำคัญ ที่ทำให้เกิดความเศร้า เสียใจ หดหู่ วิตกกังวล เหนื่อย เพลีย จากสิ่งรอบข้าง จะส่งผลกระทบต่อการแบบจำลอง เช่น ข้อความที่แสดงความเสียใจจากเหตุการณ์ทำร้ายและยิงนักเรียน ถูกโพสต์มาเป็นจำนวนหนึ่งและต่อเนื่อง ได้แก่ “We are lost. Our children are seeing yet another school shooting”, “ How to reduce mass shooting deaths? Experts rank gun laws”, “Another school shooting last Thursday w/3 dead and I’m only now hearing about it”, “ Watching yet another school shooting, big loss, condolences to their families” ในกรณีดังกล่าวให้คลาสเป็นโรคซึมเศร้า เนื่องจากแฮชแท็กส่วนใหญ่เป็นความรู้สึกที่รู้สึกเศร้าหมอง หดหู่ใจ รู้สึกไม่ดี ซึ่งอาจจะแก้ไขได้ด้วยการตรวจสอบสถานะการณ์ของการโพสต์ว่าเป็นการแชร์ การรีทวีต จากข่าว และอาจกรองหรือลบการโพสต์นั้นออก แต่อาจจะทำให้การเกิดการขาดหายของระยะเวลาตามมาตรฐานของ DSM-5 criteria

4.2.2 การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ

ในการทดลองนี้ได้เก็บรวบรวมข้อมูลที่ประกอบด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ เนื่องด้วยแต่ละบุคคลมีความชอบการโพสต์เพื่อแสดงความรู้สึกที่แตกต่างกัน ผู้ใช้งานโซเชียลมีเดียบางคนชอบโพสต์รูปภาพ บางผู้ใช้งานชอบโพสต์ข้อความ หรือในบางระยะเวลาผู้ใช้งานอาจจะชอบโพสต์เฉพาะข้อความและอาจจะชอบโพสต์รูปภาพในบางวัน จะทำให้ได้ข้อมูลการโพสต์เป็นไปตามระยะเวลาที่ได้กำหนดไว้ตามมาตรฐานของ DSM-5 criteria การทดลองนี้ทำให้ทราบว่าการคัดเลือกเฉพาะข้อความมาวิเคราะห์โรคซึมเศร้า โดยตัดการโพสต์ประเภทอื่นๆ เช่น ภาพ อาจทำให้อาการของโรคซึมเศร้าขาดหายไปจากระยะเวลาที่ได้กำหนดไว้ตามมาตรฐาน ซึ่งจะทำให้การประเมินผลการเป็นโรคซึมเศร้ามีความผิดพลาดหรือไม่ พบว่า การนำข้อมูลที่ประกอบประกอบด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ ซึ่งจะมีผลต่อการวิเคราะห์ภาวะการเป็นโรคซึมเศร้า หากนำเฉพาะข้อความจากการโพสต์เหมือนกับการทดลองของกรอบการวิจัยที่ 1 ไปพิจารณาเพื่อจำแนกอาการของโรคซึมเศร้าเพียงอย่างเดียว อาจทำให้ไม่สามารถวิเคราะห์ได้ว่าบุคคลนั้น มีภาวะของโรคซึมเศร้าหรือไม่ เนื่องจากข้อความที่นำมาพิจารณามีระยะเวลาไม่เป็นไปตามมาตรฐานของ DSM-5 criteria ซึ่งไม่ได้หมายความว่าบุคคลนั้นไม่มีภาวะของโรคซึมเศร้า เพราะคะแนนการประเมินไม่เป็นไปตามกำหนด แต่เพราะระยะเวลาของบุคคลนั้นไม่เป็นไปตามมาตรฐานของ DSM-5 แต่หากนำคุณลักษณะภาพมาพิจารณาร่วมกับคุณลักษณะจากความคิดเห็นที่เป็นข้อความ จะสามารถเพิ่มระยะเวลาตามที่ได้กำหนดไว้ จะทำให้สามารถนำข้อมูลมาจำแนกและวิเคราะห์โรคซึมเศร้าได้ดียิ่งขึ้น



ภาพที่ 15 ตัวอย่างระยะเวลาการวิเคราะห์ภาวะโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ

จากภาพที่ 15 แสดงตัวอย่างระยะเวลาการวิเคราะห์ภาวะโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ หากนำเฉพาะคุณลักษณะจากข้อความที่มีการโพสต์ตั้งแต่วันที่ 1 ถึง 7 จะทำให้ระยะเวลาไม่เป็นไปตามมาตรฐานของ DSM-5 criteria หากนำคุณลักษณะภาพที่มีการโพสต์ตั้งแต่วันที่ 8 ถึง 14 มาทำการวิเคราะห์ร่วมกับคุณลักษณะจากข้อความจะทำให้ระยะเวลาของบุคคลตัวอย่างเป็นไปตามมาตรฐานของ DSM-5 criteria และจะสามารถนำมาประเมินภาวะโรคซึมเศร้าได้

การทดลองทำให้ทราบว่าวิธีการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิลด้วยวิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ ที่ได้นำเสนอจะสามารถทำงานได้ดีกับข้อมูลจากภาพและข้อความหรือไม่ จากผลการทดลอง พบว่า วิธีการกำหนดค่าน้ำหนักทั้งสองวิธีการด้วยวิธีการให้ค่าน้ำหนักจากอัตราการทำนายถูกของคลาสคำตอบ ค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ ให้ประสิทธิภาพที่ดีกว่าวิธีการแบบไม่กำหนดค่าน้ำหนัก และแบบจำลองด้วยตัวจำแนกแบบเดี่ยว มีการคัดเลือกชุดข้อมูลสำหรับการเรียนรู้แบบจำลองที่ดี เห็นได้อย่างชัดเจนว่าวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักสามารถแก้ไขปัญหาการวิธีการแบบไม่กำหนดค่าน้ำหนักของแบบจำลองที่ใช้ตัวจำแนกประเภทจำนวนคู่ มีความแม่นยำทั้งคลาสคนที่เป็นโรคซึมเศร้าและไม่เป็นโรคซึมเศร้าและมีเสถียรภาพทั้งสองวิธีการ แต่ใช้เวลาในการประมวลไม่ต่างจากวิธีการแบบไม่กำหนดค่าน้ำหนัก

ข้อดีของการทดลองนี้ ในกรณีที่ข้อความถูกขั้นตอนการเตรียมข้อมูลลบหรือตัดออกจนเป็นข้อมูลว่าง ยังมีข้อมูลคุณลักษณะภาพเหลือและสามารถนำไปทำนายได้ จึงไม่ทำให้ข้อมูลบางช่วงเวลาขาดหายไป และสามารถนำมาประเมินภาวะโรคซึมเศร้าได้

บุคคลที่ไม่เป็นโรคซึมเศร้าที่มีผลการทำนายผิดนั้น พบว่า มีการโพสต์ภาพจำนวนหลายภาพต่อวันที่เกี่ยวกับกิจกรรมการแสดง เช่น การถ่ายภาพ หนังสือ การเดินแบบ การแสดงในสถานที่มืดหรือมีแสงค่อนข้างน้อย ภาพถ่ายที่มาจากการร่วมประท้วง ซึ่งภาพส่วนใหญ่เป็นภาพโทนสีเทา ดำ และสีน้ำเงิน ดังภาพที่ 16 ซึ่งเป็นสีที่ใช้สำหรับการเรียนรู้แบบจำลองของบุคคลที่เป็นโรคซึมเศร้า จึงทำให้ผลการทำนายผิด ส่วนบุคคลที่เป็นโรคซึมเศร้าที่มีผลการทำนายผิด เนื่องจากข้อมูลการโพสต์มีเพียงข้อความหรือรูปภาพเพียงอย่างใดอย่างหนึ่งหรือมีข้อมูลไม่เพียงพอต่อการนำมาวิเคราะห์ โรคซึมเศร้า ดังภาพที่ 17

ภาพที่ 16 ตัวอย่างการโพสต์ที่มีเพียงข้อความหรือรูปภาพเพียงอย่างใดอย่างหนึ่ง

ภาพที่ 17 ตัวอย่างการโพสต์ที่มีเพียงข้อความหรือรูปภาพเพียงอย่างใดอย่างหนึ่ง

งานวิจัยนี้ได้ทำการทดสอบวิธีการต่างๆ ในกระบวนการสกัดคุณลักษณะเพื่อให้ค่าน้ำหนักแก่คุณลักษณะด้วยวิธีการ Binary term occurrence, Term frequency inverse document frequency (TF-IDF), Term frequency และ Term occurrence จากนั้นทำการคัดเลือกคุณลักษณะด้วยวิธีการ Information gain ลำดับตาม Top-k พบว่า วิธีการ Binary term

occurrence มีประสิทธิภาพที่ดีที่สุด แม้ลดจำนวนคุณลักษณะลงแต่ยังให้ผลลัพธ์ที่ดีกว่าวิธีการอื่น จึงได้นำวิธีการดังกล่าวมาใช้ในการสร้างแบบจำลองการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล จำนวน 93 แบบจำลองจากตารางที่ 20 พบว่าแบบจำลองที่สร้างจากคุณลักษณะ จำนวน 1,300 คุณลักษณะ ให้ประสิทธิภาพที่ดีและใช้เวลาประมวลผลไม่มาก และการวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ จำนวน 96 แบบจำลองจากตารางที่ 24 ได้จำนวน 900 คุณลักษณะ ให้ประสิทธิภาพที่ดีและใช้เวลาประมวลผลไม่มาก จึงนำผลของแบบจำลองมาให้ค่าน้ำหนักแก่วิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ

จากผลการทดลอง การปรับปรุงค่าน้ำหนักที่เหมาะสมเพื่อการปรับปรุงวิธีการเหมืองข้อมูล โดยใช้วิธีการเอ็นเซมเบิล ใช้คุณลักษณะจำนวน 1,300 คุณลักษณะ ซึ่งใช้คุณลักษณะน้อยกว่าการทดลองของ Doenribram et al. (2019) ชุดข้อมูลแบบไบนารีคลาส ที่ใช้คุณลักษณะ 2,000 4,000 6,000 และใช้ทั้งหมด (24,000) เมื่อเปรียบเทียบผลลัพธ์ของชุดข้อมูลสำหรับทดสอบประสิทธิภาพแบบจำลอง พบว่าวิธีการที่นำเสนอได้ค่าความระลึกที่ดีกว่า โดยวิธีการกำหนดค่าน้ำหนักด้วยวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบได้ผลลัพธ์มากกว่า 20.00% และวิธีการอัตราการทำนายถูกของคลาสคำตอบได้ผลลัพธ์มากกว่า การทดลองของ 13.33% ในคลาสคนที่เป็นโรคซึมเศร้า ซึ่งการทดลองสำหรับการเรียนรู้งานวิจัยนี้เป็นแบบหลายคลาส แต่ได้ผลลัพธ์ของชุดข้อมูลสำหรับทดสอบประสิทธิภาพสูงสุดไม่ต่างแตกกัน และการคัดเลือกจำนวนคุณลักษณะที่เหมาะสมทำให้ใช้เวลาประมวลที่น้อย จุดด้อยของการทดลองนี้ ชุดข้อมูลสำหรับเรียนรู้แบบจำลองในคลาสอาการบุคคลปกติมีจำนวนอัตราส่วนที่น้อยกว่าอาการของโรคซึมเศร้า แม้ว่าแต่ละคลาสจะมีจำนวนเท่ากัน การนำชุดข้อมูลที่เป็นอาการของโรคซึมเศร้า 9 อาการ ทำให้คุณลักษณะที่สำคัญของคนอาการปกติน้อยเกินไป การเรียนรู้ชุดคำสำคัญได้น้อยกว่าอาการคนที่เป็นโรคซึมเศร้า เห็นได้จากค่าความระลึกของวิธีการเอ็นเซมเบิลแบบกำหนดน้ำหนักวิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบได้มากกว่าการทดลองของ Doenribram et al. (2019) ค่อนข้างมากในคลาสบุคคลที่เป็นโรคซึมเศร้า แต่คลาสบุคคลที่ไม่เป็นโรคซึมเศร้าได้ผลลัพธ์น้อยกว่า เพราะจำนวนข้อมูลของคลาสสำหรับการเรียนรู้ของบุคคลที่ไม่เป็นโรคซึมเศร้ามีจำนวนน้อยกว่าบุคคลที่เป็นโรคซึมเศร้าจำนวนมาก ทำให้เกิดความไม่สมดุลของข้อมูลขึ้นและในบางแถวข้อความได้ถูกลบทิ้งไปทำให้เกิดข้อผิดพลาดในการทำนาย

จากผลการทดลอง การวัดประสิทธิภาพการจำแนกโรคซึมเศร้าด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพด้วยวิธีการวิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ โดยใช้จำนวนตัวจำแนกประเภทที่แตกต่างกัน และการคัดเลือกจำนวนคุณลักษณะ จำนวน 900 คุณลักษณะ ทำให้ได้แบบจำลองการเรียนรู้ที่ดี

การให้ค่าน้ำหนักที่ทำให้เพิ่มประสิทธิภาพการทำนาย ใช้เวลาการประมวลน้อยและได้ผลลัพธ์ที่ดี การใช้คุณลักษณะภาพร่วมกับข้อความ แม้อาจจะมีการทำความสะอาดข้อความออกไปทั้งหมด แต่คุณลักษณะภาพยังคงอยู่ ทำให้ยังสามารถทำนายคำตอบสุดท้ายได้ จึงไม่ทำให้อาการของโรคซึมเศร้าขาดหายไปจากระยะเวลาตามข้อกำหนดของมาตราฐานของ DSM-5 criteria

บทที่ 5
สรุปผลการวิจัย

การวิจัยเรื่อง “การปรับปรุงวิธีการการพยากรณ์โรคซึมเศร้าในวัยรุ่น” ประกอบด้วย 2 ส่วน คือ 1) การปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล และ 2) การวัดประสิทธิภาพการจำแนกด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพ ผลการวิจัยสามารถสรุปผลและอภิปรายผล และข้อเสนอแนะ ดังนี้

5.1 สรุปผลและอภิปราย

การวิจัยนี้ได้เสนอวิธีการปรับปรุงค่าน้ำหนักที่เหมาะสมและปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล และการเพิ่มคุณลักษณะภาพสามารถการจำแนกโรคซึมเศร้า ด้วยวิธีการให้ค่าน้ำหนักจากอัตราการทำนายถูกของคลาสคำตอบ (True positive weighted ensemble) ค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ (Average probability weighted ensemble) โดยใช้ข้อมูลจาก Twitter และ Instagram การให้ค่าน้ำหนักแก่คุณลักษณะด้วยวิธีการ Binary term occurrence และคัดเลือกคุณลักษณะด้วยวิธีการ Information gain ตาม Top-k ทำให้ได้คุณลักษณะที่เหมาะสมในการสร้างแบบจำลองการเรียนรู้ที่มีประสิทธิภาพสามารถ ช่วยลดเวลาการประมวลผลได้ กรอบการวิจัยที่ 1ใช้ คุณลักษณะที่เหมาะสมจากกรอบการวิจัยที่ 1 จำนวน 1,300 คุณลักษณะสร้างแบบจำลองสำหรับการเรียนรู้ พบว่าประสิทธิภาพเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักมีความถูกต้องมากกว่าแบบไม่กำหนดค่าน้ำหนักทั้งสองวิธีการและมีเสถียรภาพมากกว่า โดยวิธีการวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบ มีค่าความถูกต้อง 66.67% มีค่าความระลึกสูงสุดที่ 80.00% และค่าเฉลี่ยประสิทธิภาพโดยรวม 70.59% แม้ว่า แรนดอมฟอเรสต์ ให้ค่าความระลึกสูงสุดที่ 86.67% ในคลาสของบุคคลคนที่ไม่เป็นโรคซึมเศร้า แต่ในคลาสของบุคคลที่เป็นโรคซึมเศร้าได้ผลลัพธ์ที่ต่ำมาก พบว่า การให้ค่าน้ำหนักคุณลักษณะด้วยวิธีการ Binary term occurrence สามารถทำงานได้ดี วิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักมีประสิทธิภาพมากกว่าวิธีการแบบไม่กำหนดค่าน้ำหนักอย่างมีนัยสำคัญที่ 0.05 ทั้งคลาสของบุคคลที่เป็นและไม่เป็นโรคซึมเศร้า โดยวิธีการวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบมีค่าความระลึกมากกว่าการทดลองของ Doenribram et al. (2019) 20.00% และวิธีการวิธีการอัตราการทำนายถูกของคลาสคำตอบมีค่าความระลึกมากกว่า 13.33% ในคลาสของคนที่เป็นโรคซึมเศร้า แต่ใช้คุณลักษณะจำนวนน้อยกว่า

การวัดประสิทธิภาพการจำแนกด้วยคุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพให้ประสิทธิภาพแบบกำหนดค่าน้ำหนัก ด้วยวิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบได้ประสิทธิภาพที่ดีกว่าแบบไม่กำหนดค่าน้ำหนักและแบบจำลองด้วยตัวจำแนกแบบเดี่ยว ในการทดลองนี้ ใช้คุณลักษณะที่เหมาะสมจากกรอบการวิจัยที่ 2 จำนวน 900 คุณลักษณะสร้างแบบจำลองสำหรับการเรียนรู้ โดยวิธีการอัตราการทำนายถูกของคลาสคำตอบและวิธีการค่าเฉลี่ยความน่าจะเป็นของการเกิดคลาสคำตอบมีค่าความถูกต้องมากที่สุด 87.23% ซึ่งมากกว่าวิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนัก แต่ใช้ตัวจำแนกประเภทจำนวนเท่ากัน พบว่าบุคคลที่ไม่เป็นโรคซึมเศร้าแต่ผลการทำนายผิด เขาได้โพสต์ภาพที่เกี่ยวกับกิจกรรมการทำงานในการถ่ายภาพที่มีโทนสีเทา น้ำเงิน และดำ ทำให้ผลการทำนายผิด

การปรับปรุงค่าน้ำหนักเพื่อปรับปรุงวิธีการเหมืองข้อมูลโดยวิธีการเอ็นเซมเบิล สามารถจำแนกโรคซึมเศร้าได้อย่างมีประสิทธิภาพ ช่วยแก้ไขปัญหาของวิธีการเอ็นเซมเบิลแบบไม่กำหนดค่าน้ำหนักได้ทั้งข้อมูลที่มีคลาสคำตอบแบบหลายคลาสและไบนารีคลาส เนื่องจากวิธีการกำหนดค่าน้ำหนักช่วยเพิ่มโอกาสการทำนายที่มีการทำนายผิดพลาดเพราะทำให้การทำนายคลาสคำตอบใหม่ที่ได้จากการคำนวณด้วยวิธีการเอ็นเซมเบิลแบบกำหนดค่าน้ำหนักมีโอกาสที่จะทำนายถูกต้องเพิ่มขึ้น ลดปัญหาที่เกิดขึ้นหากจำนวนตัวจำแนกประเภทเป็นจำนวนคู่ รวมทั้งช่วยแก้ไขปัญหาการเลือกตัวจำแนกประเภทที่เหมาะสมกับชุดข้อมูลทดลองได้ การเพิ่มคุณลักษณะภาพทำให้เพิ่มประสิทธิภาพการทำนาย เนื่องการคำนวณความน่าจะเป็นให้แต่ละคลาสคำตอบจะพิจารณาจากค่าน้ำหนักของคุณลักษณะ อาจมีคำบางคำที่มีค่าน้ำหนักใกล้เคียงกันมากแล้วทำให้ผลการทำนายคลาสอยู่คนละคลาส อาจจะทำให้ทำนายผิดได้ เช่น destructive กับ beauty มีค่าน้ำหนักต่างกัน 0.0001 เป็นคำที่มีค่าน้ำหนักใกล้เคียงกันมากแต่เป็นคำที่อยู่ในคลาสที่ตรงข้ามกัน เมื่อมีคุณลักษณะภาพทำให้การพิจารณาค่าน้ำหนักของแถวข้อมูลนั้นเปลี่ยนแล้วทำให้การคำนวณค่าความน่าจะเป็นของแต่ละคลาสแตกต่างกันได้ และในกรณีที่ข้อความในแถวถูกลบออกไปหมด แต่คุณลักษณะภาพยังคงอยู่ จึงสามารถนำไปพิจารณาเพื่อคำนวณค่าความน่าจะเป็นของคลาสนั้นได้ ช่วยลดการทำนายผิด จากการไม่มีเหลือข้อมูลในแถวนั้น จึงทำให้ประสิทธิภาพการทดลองที่ใช้คุณลักษณะข้อมูลจากความคิดเห็นร่วมกับคุณลักษณะภาพได้ผลลัพธ์ที่ดีกว่าการใช้คุณลักษณะจากข้อความเพียงอย่างเดียว และช่วยให้การจำแนกโรคซึมเศร้าครอบคลุมมากยิ่งขึ้น

การคัดเลือกเฉพาะการโพสต์ข้อความมาวิเคราะห์อาการโรคซึมเศร้า โดยตัดการโพสต์ประเภทอื่นออกไป อาจทำให้อาการของโรคซึมเศร้าขาดหายไปจากระยะเวลาที่ได้กำหนดไว้ตามมาตราฐานของ DSM-5 criteria ซึ่งอาจจะทำให้การประเมินผลโรคซึมเศร้ามีความผิดพลาดหรือไม่ พบว่า ผลการทดลองกรอบการวิจัยที่ 1 ที่ใช้เฉพาะข้อความการโพสต์แสดงความคิดเห็นทั้ง สำหรับการเรียนรู้แบบจำลอง และสำหรับทดสอบแบบจำลอง มีหลายแถวที่มีการทำความสะอาดข้อความออกไปหมด

ทำให้แถวนั้นเป็นข้อมูลว่าง ซึ่งทำให้ประสิทธิภาพลดลง และทำให้การประเมินผลการเป็นโรคซึมเศร้าผิดพลาด ซึ่งแตกต่างจากผลการทดลองกรอบการวิจัยที่ 2 ที่ใช้ชุดข้อมูลที่ประกอบด้วยคุณลักษณะข้อความจากความคิดเห็นร่วมกับคุณลักษณะภาพ ได้ผลการประเมินการเป็นโรคซึมเศร้าที่มีประสิทธิภาพที่ดี

5.2 ข้อเสนอแนะ

1. การปรับค่าพารามิเตอร์ (Parameter) ของแต่ละแบบจำลองอาจจะมีผลกระทบต่อประสิทธิภาพของแบบจำลอง และหากมีการปรับปรุงการกำหนดค่าพารามิเตอร์ที่เหมาะสมอาจจะช่วยเพิ่มประสิทธิภาพให้แบบจำลองได้

2. ควรมีการวิเคราะห์เพิ่มเติมในกรณีการโพสต์เชิงบวกและการโพสต์เชิงลบในแถวข้อมูลเดียวกันหรือในโพสต์เดียวกัน เพราะเมื่อนำมาประมวลผลแล้วอาจจะมีการหักล้างกัน อาจจะทำให้ได้ผลการทำนายที่ผิดหรืออาจทำนายผลไม่ได้

3. เนื่องจากข้อมูลสำหรับนำมาทดสอบประสิทธิภาพแบบจำลองมีข้อจำกัดในการคัดเลือกจำนวนข้อมูลและช่วงเวลาในการโพสต์แสดงความคิดเห็น ส่งผลกระทบโดยตรงต่อการวิเคราะห์และประเมินภาวะการเป็นโรคซึมเศร้า อาจจะทำให้คะแนนประเมินมีคะแนนหลายคะแนน ซึ่งโปรแกรมที่ใช้ในการคำนวณและประเมินโรคซึมเศร้าในงานวิจัยกำหนดให้เลือกคะแนนที่สูงที่สุดมาประเมินว่ามีภาวะของโรคซึมเศร้าหรือไม่ อาจจะทำให้ผลการประเมินให้บุคคลนั้นเป็นโรคซึมเศร้า เนื่องจากเป็นโปรแกรมจะเลือกคะแนนที่มากที่สุด ถือเป็นข้อกำจัดของงานวิจัยนี้ อาจทำให้คะแนนการประเมินสูง เช่น ข้อมูลของ user01 มีระยะเวลาที่เป็นไปตามระยะเวลาตามข้อกำหนดของมาตราฐานของ DSM-5 criteria ในปี ค.ศ. 2021 จำนวน 5 ครั้ง เมื่อนำข้อมูลมาทำตามกระบวนการของงานวิจัยนี้ พบว่า user01 มีคะแนนประเมิน 11, 9, 10, 13 และ 14 คะแนน โดยคะแนน 13 และ 14 จะอยู่ในเกณฑ์เป็นโรคซึมเศร้า แต่คะแนน 11, 9 และ 10 ไม่เป็นโรคซึมเศร้า แต่โปรแกรมจะประเมินว่า user01 เป็นโรคซึมเศร้าเพราะถูกกำหนดให้เลือกคะแนนที่สูงที่สุดมาทำการประเมิน จากปัญหาและข้อบกพร่องที่พบ หากเพิ่มวิธีการการคัดเลือกจำนวนข้อมูลสำหรับมาทดสอบที่เหมาะสมอาจช่วยลดความผิดพลาดในการวิเคราะห์ผลได้หรือควรมีวิธีการเลือกช่วงระยะเวลาที่การประเมินโรคซึมเศร้าและเป็นไปตามมาตราฐานของ DSM-5 criteria

บรรณานุกรม

Alabdulkreem, E. (2020). Prediction of depressed Arab women using their tweets. *Journal of Decision Systems*, 1–16. https://doi.org/10.1080/12460125.2020.1859745

Aldarwish, M. M., & Ahmad, H. F. (2017). Predicting depression levels using social media posts. In *Proceedings - IEEE 13th International Symposium on Autonomous Decentralized Systems, (ISADS)*, 277–280. https://doi.org/10.1109/ISADS.2017.41

AlHamed, F., & AlGwaiz, A. (2020). A hybrid social mining approach for companies current reputation analysis. In *Advances in Intelligent Systems and Computing: Vol. 978 AISC*. Springer International Publishing. https://doi.org/10.1007/978-3-030-36056-6_40

American Psychiatric Association. (2013). *Diagnostic and statistical manual psychiatric, mental disorder edition "DSM-5".* (5th ed.). American Psychiatric Association.

Atorn, N. (2006). *Text categorization & retrieval for Thai item bank using patterned keyword in phrase (PKIP).* Doctoral dissertation, Mahidol University.

Carruthers, H.R., Morris, J., Tarrier, N. et al. The Manchester color wheel: development of a novel way of identifying color choice and its validation in healthy, anxious and depressed individuals. *BMC Med Res Methodol* 10, 12 (2010). https://doi.org/10.1186/1471-2288-10-12

Choudhury, M. De, Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016). Discovering shifts to suicide ideation from mental health. *Proc SIGCHI Conf Hum Factor Comput Syst.*, 2098–2110. https://doi.org/10.1145/2858036.2858207

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data*

Analysis, *1*(3), 131–156. https://doi.org/10.3233/IDA-1997-1302

Dehzangi, A., & Karamizadeh, S. (2011). Solving protein fold prediction problem using fusion of heterogeneous classifiers. *Information*, *14*(11), 3611–3621.

Depression rates by country 2021. (2021). World Population Review. https://worldpopulationreview.com/country-rankings/diabetes-rates-by-country

Derrac, J., García, S., & Herrera, F. (2010). A Survey on Evolutionary instance selection and generation. *International Journal of Applied Metaheuristic Computing*, *1*(1), 60–92. https://doi.org/10.4018/jamc.2010102604

Doenribram, D. (2019). *Depressive Classification from posts on twitter of user behaviors,* Master's dissertation, Mahasarakham University. http://202.28.34.124/dspace/bitstream/123456789/50/1/60011252001.pdf

Doenribram, D., Jareanpon, C., & Jiranukool, J. (2020). Comparison of data mining structure performance for depressive classification if twitter users from their posts on twitter of user behaviors. *Journal of Science and Technology Mahasarakham University*, *39*(3), 331–343.

Doenribram, D., Jareanpon, C., & Jiranukool M.D., Jariya Sonbua, S. (2019). Major depressive disorder classification from user behaviors from twitter. In *The Twenty-Fourth International Symposium on Artificial Life and Robotics 2019 (AROB), Beppu, Japan, January 23-25, 2019*, 241–246.

Ebrahimi, M., Mohammadi-Dehcheshmeh, M., Ebrahimie, E., & Petrovski, K. R. (2019). Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: deep learning and gradient-boosted trees outperform other models. *Computers in Biology and Medicine*, *114*(May), 103456. https://doi.org/10.1016/j.compbiomed.2019.103456

El-Bendary, N., Hariri, E. E., Hassanien, A. E., & Badr, A. (2015). Using machine learning techniques for evaluating tomato ripeness. *Expert Systems with Applications, 42*(4), 1892-1905. https://doi.org/10.1016/j.eswa.2014.09.057

Fernandes, T. M. P., Andrade, S. M., De Andrade, M. J. O., Nogueira, R. M. T. B. L., & Santos, N. A. (2017). Colour discrimination thresholds in type 1 Bipolar Disorder: a pilot study. *Scientific Reports*, *7*(1), 1–11. https://doi.org/10.1038/s41598-017-16752-0

Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. In *Morgan Kaufmann Publishers is an imprint of Elsevier* (2nd ed.). Morgan Kaufmann Publishers is an Imprint of Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: data mining concepts and techniques. In *Proceedings - 2013 International Conference on Machine Intelligence Research and Advancement (ICMIRA)*. Morgan Kaufmann Publishers is an Imprint of Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Hand, D., Mannila, H., & Smyth, P. (1999). *Principles of Data Mining Chapter 9 Descriptive Modeling*. The MIT Press.

Hand, D., Mannila, H., & Smyth, P. (2001). Principles of data mining. In *Bradford Book*. http://link.springer.com/10.1007/978-1-4471-4884-5

Harrigian, K., Aguirre, C., & Dredze, M. (2021). On the state of social media data for mental health research. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, 15–24. https://doi.org/10.18653/v1/2021.clpsych-1.2

Ignatow, G., & Mihalcea, R. (2018). *An introduction to text mining*. SAGE Publications, Inc.

Jennings, M. J., Zumbo, B. D., & Joula, J. F. (2002). The robustness of validity and efficiency of the related samples t-test in the presence of outliers. *Psicologica*, 23(2), 415-450.

Jimenez-Marquez, J. L., Gonzalez-Carrasco, I., Lopez-Cuadrado, J. L., & Ruiz-Mezcua, B. (2019). Towards a big data framework for analyzing social media content. *International Journal of Information Management*, *44*(May 2018), 1–12. https://doi.org/10.1016/j.ijinfomgt.2018.09.003

Jo, V. (2019). Text mining concepts, Implementation, and big data challenge. In *Seminars in Diagnostic Pathology*, 36(2), 83-84. https://doi.org/10.1053/j.semdp.2019.02.002

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2015 - Proceedings*. Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/MIPRO.2015.7160458

Kang, K., Yoon, C., & Kim, E. Y. (2016). Identifying depressive users in twitter using multimodal analysis. In *2016 International Conference on Big Data and Smart Computing, BigComp 2016*, 231–238. https://doi.org/10.1109/BIGCOMP.2016.7425918

Khan, A., Li, J. P., Haq, A. U., Memon, I., Patel, S. H., & ud Din, S. (2021). Emotional-physic analysis using multi-feature hybrid classification. *Journal of Intelligent and Fuzzy Systems*, *40*(1), 1681–1694. https://doi.org/10.3233/JIFS-201069

Kompan, M., & Bieliková, M. (2011). News article classification based on a vector representation including words' collocations. In: Dicheva, D., Markov, Z., Stefanova, E. (eds) Third International Conference on Software, Services and Semantic Technologies S3T 2011. *Advances in Intelligent and Soft Computing*,

101. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23163-6_1

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. In *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Kuo, Y.-H., Li, Z., & Kifer, D. (2018). *Detecting Outliers in Data with Correlated Measures*. 10. https://doi.org/10.1145/3269206.3271798

Lin, L. Y., Sidani, J. E., Shensa, A., Radovic, A., Miller, E., Colditz, J. B., Hoffman, B. L., Giles, L. M., & Primack, B. A. (2016). Association between social media use and depression among U.S. young adults. *Depression and Anxiety*, *33*(4), 323–331. https://doi.org/10.1002/da.22466

Linoff, G., & Berry, M. J. A. (2011). *Data mining techniques : for marketing, sales, and customer relationship management*. Wiley.

Mondal, M., Semwal, R., Raj, U., Aier, I., & Varadwaj, P. K. (2020). An entropy-based classification of breast cancerous genes using microarray data. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-018-3864-8

Mousavian, M., Chen, J., & Greening, S. (2018). Feature selection and imbalanced data handling for depression detection. *Brain Informatics. BI 2018. Lecture Notes in Computer Science*, *11309 LNAI*, 349–358. https://doi.org/10.1007/978-3-030-05587-5_33

Munmun, De, C., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, *36*(1–2), 128–137. https://doi.org/10.3109/01460862.2013.798190

Negrão, A. B., & Gold, P. W. (2007). Major depressive disorder. *Encyclopedia of Stress*, *28*, 640–645. https://doi.org/10.1016/B978-012373947-6.00245-2

Nuankaew, W., & Thongkam, J. (2020). Improving student academic performance prediction models using feature selection. In *17th International Conference on*

Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 392–395. https://doi.org/10.1109/ECTI-CON49241.2020.9158286

Onan, A., Korukoğlu, S., & Bulut, H. (2016). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, *62*, 1–16. https://doi.org/10.1016/j.eswa.2016.06.005

Ong, B. Y., Goh, S. W., & Xu, C. (2015). Sparsity adjusted information gain for feature selection in sentiment analysis. In *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, 2122–2128. https://doi.org/10.1109/BigData.2015.7363995

Osamor, V. C., & Okezie, A. F. (2021). Enhancing the weighted voting ensemble algorithm for tuberculosis predictive diagnosis. *Scientific Reports*, *11*(1), 1–11. https://doi.org/10.1038/s41598-021-94347-6

Psychas, I. D., Delimpasi, E., & Marinakis, Y. (2015). Hybrid evolutionary algorithms for the Multiobjective Traveling Salesman Problem. *Expert Systems with Applications*, *42*(22), 8956–8970. https://doi.org/10.1016/j.eswa.2015.07.051

Ramanuj, P., Ferenchick, E. K., & Pincus, H. A. (2019). Depression in primary care: part 2-management. *BMJ (Clinical Research Ed.)*, *365*, l835. https://doi.org/10.1136/bmj.l835

Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1), 15.

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications. In *Computers in Biology and Medicine*, 112, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Rojarath, A. (2020). *Propability-weighted voting ensemble learning for classification*

model. Doctoral dissertation, Khon Kaen University.

Rojarath, A. A., & Songpan, W. B. (2020). Probability-weighted voting ensemble learning for classification model. *Journal of Advances in Information Technology*, 11(4), 217–227. https://doi.org/10.12720/jait.11.4.217-227

Rojarath, A., & Songpan, W. (2021). Cost-sensitive probability for weighted voting in an ensemble model for multi-class classification problems. *Applied Intelligence*, *51*(7), 4908–4932. https://doi.org/10.1007/s10489-020-02106-3

Sannasi Chakravarthy, S. R., & Rajaguru, H. (2022). Ensemble-based weighted voting approach for the early diagnosis of diabetes mellitus. In *Sustainable Communication Networks and Application. Lecture Notes on Data Engineering and Communications Technologies*, 93, 451–460. https://doi.org/10.1007/978-981-16-6605-6_33

Sawangarreerak, S., & Thanathamathee, P. (2020). Random forest with sampling techniques for handling imbalanced prediction of university student depression. *Information (Switzerland)*, *11*(11), 1–13. https://doi.org/10.3390/info11110519

Schapire, R. E. (2013). Explaining adaboost. *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, 37–52. https://doi.org/10.1007/978-3-642-41136-6_5

Sierra, M. R., & Coello Coello, C. A. (2005). Improving PSO-based multi-objective optimization using crowding, mutation and **ε**-dominance. *Lecture Notes in Computer Science*, *3410*, 505–519. https://doi.org/10.1007/978-3-540-31880-4_35

Simon, K. (2022). *Digital 2022: April Global Statshot*. 2022-04-21. https://datareportal.com/reports/digital-2022-april-global-statshot

Singh, S. M., & Hemachandran, K. (2012). Content based image retrieval based on the integration of color histogram, color moment and gabor texture. *International Journal of Computer Applications*, 59(17), 13-22. https://doi.org/10.5120/9639-4325

Stopwords, E. (n.d.). *Default english stopwords list*. Retrieved January 20, 2022, from https://www.ranks.nl/stopwords

Swamynathan, M. (2017). Mastering machine learning with python in six steps. In *Mastering Machine Learning with Python in Six Steps*. Apress. https://doi.org/10.1007/978-1-4842-2866-1

Tama, B. A., Im, S., & Lee, S. (2020). Improving an intelligent detection system for coronary heart disease using a two-tier classifier ensemble. *BioMed Research International*, *2020*. https://doi.org/10.1155/2020/9816142

ThaiHealth watch 2020. (n.d.). https://www.thaihealth.or.th/data/ecatalog/614/pdf/614.pdf

Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing depression from twitter activity. In *Conference on Human Factors in Computing Systems*, *April 18–2*, 3187–3196. https://doi.org/10.1145/2702123.2702280

Wendlandt, L., Mihalcea, R., Boyd, R. L., & Pennebaker, J. W. (2017). Multimodal analysis and prediction of latent user dimensions. In *International Conference on So- Cial Informatics*, *10539 LNCS*, 323–340. https://doi.org/10.1007/978-3-319-67217-5_20

WHO. (2021). *Depression*. WHO. https://www.who.int/news-room/fact-sheets/detail/depression

Wongthanavasu, S. (2017). *Artificial intelligence : theory, programs and applications*. Khon Kaen University.

Wu, G., & Xu, J. (2016). Optimized Approach of Feature Selection Based on Information Gain. In *Proceedings - 2015 International Conference on Computer Science and Mechanical Automation, CSMA 2015*, *Mi*, 157–161. https://doi.org/10.1109/CSMA.2015.38

Yazdavar, A. H., Mahdavinejad, M. S., Bajaj, G., Romine, W., Sheth, A., Monadjemi, A. H., Thirunarayan, K., Meddar, J. M., Myers, A., Pathak, J., & Hitzler, P. (2020). Multimodal mental health analysis in social media. *PLoS ONE*, *15*(4), 1–27. https://doi.org/10.1371/journal.pone.0226248

Zhang, N., Liu, C., Chen, Z., An, L., Ren, D., Yuan, F., Yuan, R., Ji, L., Bi, Y., Guo, Z., Ma, G., Xu, F., Yang, F., Zhu, L., Robert, G., Xu, Y., He, L., Bai, B., Yu, T., & He, G. (2019). Prediction of adolescent subjective well-being: A machine learning approach. *General Psychiatry*, *32*(5), 100096. https://doi.org/10.1136/gpsych-2019-100096

Zhang, W., Liu, H., Silenzio, V. M. B., Qiu, P., & Gong, W. (2020). Machine learning models for the prediction of postpartum depression: Application and comparison based on a cohort study. *JMIR Medical Informatics*, *8*(4), e15516. https://doi.org/10.2196/15516

Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., & Liu, H. (2010). Advancing feature selection research. *ASU Feature Selection Repository Arizona State University*, 1–28. http://featureselection.asu.edu/featureselection_techreport.pdf

Zheng, A., & Casari, A. (2018). Feature engineering for machine learning: principles and techniques for data scientists. *O'Reilly Media*, Inc.

Zul, M. I., Yulia, F., & Nurmalasari, D. (2018). Social media sentiment analysis using K-means and naïve bayes algorithm. In *Proceedings - 2018 2nd International Conference on Electrical Engineering and Informatics: Toward the Most Efficient Way of Making and Dealing with Future Electrical Power System and Big Data Analysis (ICon EEI), October*, 24–29. https://doi.org/10.1109/ICon-EEI.2018.8784326

ประวัติผู้เขียน

ชื่อ	นางวงษ์ปัญญา นวนแก้ว
วันเกิด	12 กรกฎาคม 2524
สถานที่อยู่ปัจจุบัน	194/181 หมู่ 6 ถนนศรีจันทร์ อำเภอเมือง จังหวัดขอนแก่น
ตำแหน่งหน้าที่การงาน	อาจารย์
สถานที่ทำงานปัจจุบัน	คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยราชภัฏมหาสารคาม
ประวัติการศึกษา	พ.ศ. 2543 สำเร็จการศึกษา ระดับชั้นมัธยมศึกษา โรงเรียนแก่นนครวิทยาลัย อำเภอเมือง จังหวัดขอนแก่น พ.ศ. 2547 สำเร็จการศึกษา ระดับปริญญาตรี (วท.บ.) วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร พ.ศ. 2550 สำเร็จการศึกษา ระดับปริญญาโท (วท.ม.) วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร พ.ศ. 2565 สำเร็จการศึกษา ระดับปริญญาเอก (ปร.ด.) ปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม
ผลงานวิจัย	- Nuankaew, W., & Thongkam, J. (2020). Improving student academic performance prediction models using feature selection. In 17th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2020, 392–395. - Nuankaew, W., Nuankaew, P., Doenribram, D.,& Jareanpon, C. (2023). Weighted voting ensemble for depressive disorder analysis with multi-objective optimization. Current Applied Science and Technology. Vol. 23 No.1. (January-February 2023). - Nuankaew, W., Nuankaew, P., Doenribram, D., Jareanpon, C.,& Thanarat, P. (2022). A new probabilistic weighted voting model for depressive disorder classification from captions and colors of images. ICIC Express Letters. (Accept).

พหุน ปณ ฑิโต ชีเว