การจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์

วิทยานิพนธ์

ของ

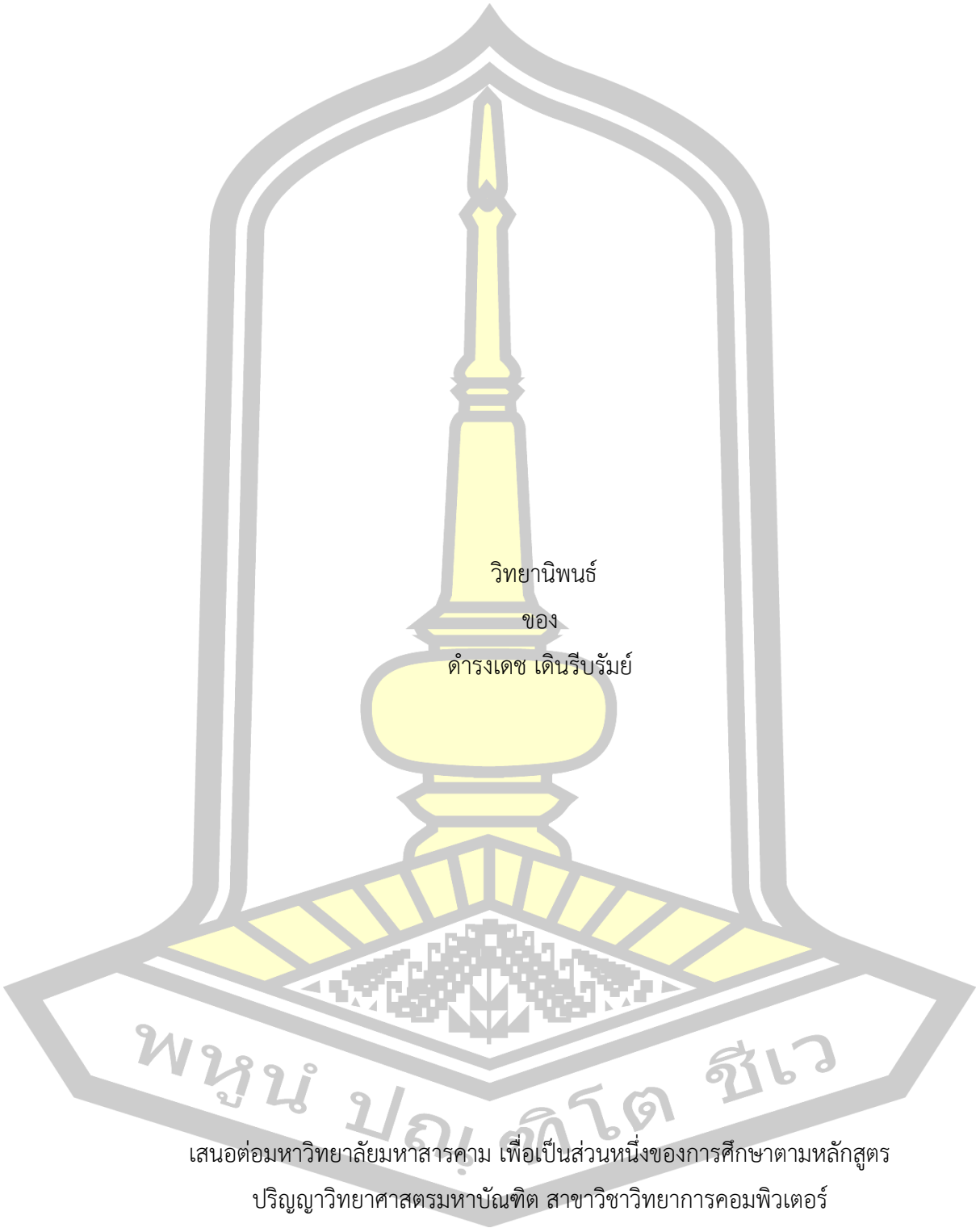
ดำรงเดช เดินรีบรัมย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

มีนาคม 2562

การจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์



วิทยานิพนธ์

ของ

ดำรงเดช เดินรีบรัมย์

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

มีนาคม 2562

Depressive Classification from Posts on Twitter of user Behaviors

Damrongdet Doenribram

A Thesis Submitted in Partial Fulfillment of Requirements

for Master of Science (Computer Science)

March 2019

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนายดำรงเดช เดินรีบรัมย์ แล้วเห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรารัตน์ สงฆ์แป้น)

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

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

.....กรรมการ
(ผศ. ดร. พนิดา ทรงรัมย์)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

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

.....
(ผศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์
ผู้วิจัย ดำรงเดช เดินรีบรัมย์
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. ฉัตรเกล้า เจริญผล
ปริญญา วิทยาศาสตรมหาบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์
มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

ในปี 2017 องค์การอนามัยโลกระบุว่าโรคซึมเศร้าเป็นสาเหตุอันดับ 2 ของการฆ่าตัวตายก่อนวัยอันควรของคนอายุระหว่าง 15-29 ปี และคนที่เป็นโรคซึมเศร้าอยู่ก่อนแล้วเมื่อเข้ามาใช้งานสื่อโซเชียลมีเดียอาจจะมีการแสดงอารมณ์ออกมาทางการโพสต์ ผู้วิจัยจึงนำเสนอการทำเหมืองความคิดเหตุเข้ามาจำแนกพฤติกรรมการโพสต์บน Twitter โดยการใช้งานอัลกอริธึม Bayes สร้างโมเดลเพื่อจำแนก 9 อาการที่บ่งบอกถึงโรซึมเศร้าตามแบบสอบถาม DSM-5 ได้แก่ 1. อารมณ์ซึมเศร้า 2. ความสนใจลดลง 3. น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต 4 . นอนไม่หลับหรือนอนหลับมากกว่าปกติ 5. ร่างกายอ่อนเพลีย 6. รู้สึกตนเองไร้ค่า 7. สมาธิสั้น 8. เคลื่อนไหวช้า และ 9. คิดฆ่าตัวตาย โดยใช้ข้อมูล 2 ชุด ได้แก่ Training set และ Test set ผลการทดลองของ Training set ได้ Accuracy สูงสุด 95.85% และการทดลอง Test set กำหนด Boundary ของความน่าจะเป็นตั้งแต่ 0 ถึง 90 เพื่อกรองข้อความที่ความน่าจะเป็นน้อยกว่าค่า Boundary ผลการทดลองของ Test set ได้ Accuracy สูงสุด 80.00%

คำสำคัญ : โรคซึมเศร้า, การจำแนกข้อมูล, การทำเหมืองโซเชียลมีเดีย, การทำเหมืองข้อความ

TITLE Depressive Classification from Posts on Twitter of user Behaviors

AUTHOR Damrongdet Doenribram

ADVISORS Assistant Professor Chatklaw Jareanpon , Ph.D.

DEGREE Master of Science **MAJOR** Computer Science

UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

In 2017, WHO indicated that the MDD was the second cause of death among the 15-29-year olds. Person has a chance of depression and uses social media, there may be an expression their feeling in the post. Therefore, this research proposes the classification from user behaviors using Bayes algorithm from Twitter that created the 9 various models, based on a symptoms of questionnaire (DSM-5) including as follow: 1) depressive 2) loss of interest 3) appetite 4) abnormal sleep 5) slowed thinking 6) guilt 7) tired 8) unexplained and 9) suicidal ideation. The data set is divided into 2 sets: training set and test set came from real tweets of celebrities. Finally, the results demonstrated of training set showed that the accuracy = 95.85% and the boundary of probability are variously set 0 to 90 for filtering the messages that the probability is less than the boundary, test set showed that the accuracy = 80.00%

Keyword : Major Depressive Disorder, Classification, Social Mining, Text Mining

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สาเร็จสมบูรณ์ได้ด้วยความกรุณาและความช่วยเหลืออย่างสูงยิ่งจาก ผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล ประธานกรรมการควบคุมวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.วรา

รัตน์ สงฆ์แป้น ประธานกรรมการสอบ ผู้ช่วยศาสตราจารย์ ดร.พัฒนพงษ์ ชมภูวิเศษ และผู้ช่วย

ศาสตราจารย์ ดร.พนิดา ทรงรัมย์ กรรมการสอบ

ขอขอบพระคุณบิดา มารดา และผู้เกี่ยวข้องที่เป็นกาลังใจ และให้ทุนสนับสนุนในการเรียน ซึ่งเป็นผลให้วิทยานิพนธ์นี้สาเร็จลุล่วงไปได้ด้วยดี

ดำรงเดช เดินรีบรัมย์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูป

หน้า

บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

โรคซึมเศร้าเป็นสาเหตุหลักของการเสียชีวิตก่อนวัยอันควร ในปี ค.ศ. 2017 องค์การอนามัยโลก (WHO) [1] ระบุว่าโรคซึมเศร้าเป็นสาเหตุอันดับสองของการฆ่าตัวตายในคนอายุระหว่าง 15-29 ปี ปัจจัยสำคัญที่ก่อให้เกิดโรคนี้เกิดจากความเครียดและความผิดหวังที่สะสมมานาน ส่งผลทำให้จิตใจและร่างกายเกิดความบกพร่องในการใช้ชีวิตประจำวัน นำไปสู่สาเหตุการฆ่าตัวตายตามมา ผลการวิจัยจาก University of Pittsburgh [2] ที่ทำการวิจัยเกี่ยวกับผลกระทบจากโซเชียมีเดีย (Social Media) มีผลต่อจิตใจของมนุษย์ผู้ใช้งาน โดยงานวิจัยนี้ระบุว่าคนที่มีอายุระหว่าง 19 - 32 ปี ที่มีการใช้งานโซเชียลมีเดียมาก ๆ นั้นเสี่ยงต่อการเกิดภาวะซึมเศร้าได้สูงมาก ส่วนสาเหตุที่โซเชียลมีเดียมีความเกี่ยวพันกับภาวะซึมเศร้าเป็นเพราะจิตใจของมนุษย์อ่อนไหวได้ง่าย เมื่อได้เห็นโพสต์ของคนอื่น ๆ ใช้ชีวิตออกเดินทางท่องเที่ยวอย่างมีความสุขทำให้เกิดความอิจฉาน้อยเนื้อต่ำใจตัวเอง รวมถึงการตั้งเกณฑ์การมีฝวามสุขตามคนอื่นที่อยู่ในโลกโซเชียลมีเดียไปโดยไม่รู้ตัวทั้ง ๆ ที่ไม่คิดเลยว่าบางเรื่องเล็ก ๆ น้อย ๆ อาจทำให้ตนเองมีความสุขกับเรื่องใกล้ตัว ซึ่งเป็นผลลัพธ์จากงานวิจัยนี้ที่บ่งบอกถึงสาเหตุการเกิดโรคซึมเศร้า

ในปี ค.ศ. 2017 สื่อโซเชียลมีเดียอย่าง Twitter เป็น Micro-blog ที่มีความนิยมมาก โดยปัจจุบันมีบัญชีผู้ใช้งานทั่วโลกกว่า 328 ล้านบัญชี และมีการ Tweets ข้อความกว่า 500 ล้านข้อความต่อวัน จนทำให้เกิดข้อมูลมหาศาลที่สามารถนำมาเข้ากระบวนการวิเคราะห์หาความรู้ (Knowledge) ได้ เหตุผลที่ Twitter สามารถตอบโจทย์พฤติกรรมการแสดงออกของมนุษย์ได้อย่างดี เนื่องจากการโพสต์ข้อความนั้นเป็นข้อความที่มีไม่เกิน 280 ตัวอักษร จึงเป็นข้อความที่สั้นแต่ได้ใจความสำคัญ [3]

ปัจจุบันมีการนำเทคนิคการทำเหมืองความคิดเห็นซึ่งเป็นวิธีการหนึ่งในการทำเหมืองข้อมูลที่มุ่งหมายเพื่อค้นหาความรู้ในฐานข้อมูลนั้น ๆ โดยการอาศัยวิธีการทางคณิตศาสตร์สถิติและการเรียนรู้ของเครื่อง (Machine Learning) เข้ามาสร้างแบบจำลองเพื่อทำการวิเคราะห์และพยากรณ์วินิจฉัยโรคทางด้านการแพทย์ ดังเช่น ได้มีการนำเทคนิคการทำเหมืองความคิดเห็นมาประยุกต์ใช้งานโดยสร้างแบบจำลองตรวจสอบความคิดเห็นบนโซเชียลมีเดียเพื่อหาผู้ที่เข้าข่ายเป็นโรคซึมเศร้า [4-7] แต่ผลลัพธ์ของการทดลองที่ได้ยังมีค่าความถูกต้องและค่าความแม่นยำที่ไม่สูงมากนัก จนกระทั่งมีงานวิจัยของ Punnee และคณะ [8] ได้ทำงานวิจัยเกี่ยวกับการจำแนกประเภทของผู้ป่วยที่เป็น

โรคเบาหวานด้วยการทำเหมืองข้อมูล ซึ่งในงานวิจัยนี้ได้นำเอาเทคนิค Ensemble เข้ามาใช้ในการสร้างแบบจำลองเพื่อเพิ่มความถูกต้องและเพิ่มความแม่นยำได้

ดังนั้นงานวิจัยนี้จะนำวิธีการทำเหมืองความคิดเห็นโดยนำความคิดเห็นจากผู้ใช้งาน Twitter มาทำการจำแนกข้อความที่เข้าข่ายโรคซึมเศร้าด้วยการทำเหมืองความคิดเห็นโดยหวังผลว่าจะมีความถูกต้องและความแม่นยำมากขึ้นในการแยกแยะได้ เพื่อให้หน่วยงานที่เกี่ยวข้องสามารถนำงานวิจัยนี้ไปช่วยวินิจฉัยและช่วยในป้องกันก่อนการเกิดโรคซึมเศร้าได้ในอนาคต

1.2 วัตถุประสงค์ของการวิจัย

เพื่อพัฒนากระบวนการในการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์

1.3 ความสำคัญของการวิจัย

ได้แบบจำลองการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์เพื่อช่วยในการวินิจฉัยของแพทย์หรือให้หน่วยงานที่เกี่ยวข้องนำไปประยุกต์ใช้งานได้ในอนาคต

1.4 ขอบเขตของการวิจัย

1.4.1 ข้อมูลสำหรับการสร้างแบบจำลอง

ข้อมูลสำหรับการสร้างแบบจำลองเป็นข้อความภาษาอังกฤษที่เก็บรวบรวมจากการติด Hashtag บน Twitter ของผู้ใช้งานทั่วไป โดยแบ่งออกตามลักษณะ 9 อาการ ที่บ่งบอกถึงโรคซึมเศร้า ได้แก่

1) ข้อความเกี่ยวกับอารมณ์ซึมเศร้า จำนวน 3,000 ข้อความ และไม่เกี่ยวกับอารมณ์ซึมเศร้า จำนวน 3,000 ข้อความ

2) ข้อความเกี่ยวกับการขาดความสนใจลดลง จำนวน 3000 ข้อความ และไม่เกี่ยวกับการขาดความสนใจลดลง จำนวน 3,000 ข้อความ

3) ข้อความเกี่ยวกับน้ำหนักผิดปกติ จำนวน 3,000 ข้อความ และไม่เกี่ยวกับน้ำหนักผิดปกติ จำนวน 3,000 ข้อความ

4) ข้อความเกี่ยวกับการนอนผิดปกติ จำนวน 3,000 ข้อความ และไม่เกี่ยวกับการนอนผิดปกติ จำนวน 3,000 ข้อความ

5) ข้อความเกี่ยวกับร่างกายอ่อนเพลีย จำนวน 3,000 ข้อความ และไม่เกี่ยวกับร่างกายอ่อนเพลีย จำนวน 3,000 ข้อความ

6) ข้อความเกี่ยวกับการรู้สึกตนเองไร้ค่า จำนวน 3,000 ข้อความ และไม่เกี่ยวกับการรู้สึกตนเองไร้ค่า จำนวน 3,000 ข้อความ

7) ข้อความเกี่ยวกับสมาธิสั้น จำนวน 3,000 ข้อความ และไม่เกี่ยวกับสมาธิสั้น จำนวน 3,000 ข้อความ

8) ข้อความเกี่ยวกับการเคลื่อนไหลช้า จำนวน 3,000 ข้อความ และไม่เกี่ยวกับเคลื่อนไหลช้า จำนวน 3,000 ข้อความ

9) ข้อความเกี่ยวกับการคิดฆ่าตัวตาย จำนวน 3,000 ข้อความ และไม่เกี่ยวกับการคิดฆ่าตัวตาย จำนวน 3,000 ข้อความ

1.4.2 ข้อมูลสำหรับทดสอบแบบจำลอง

ข้อความภาษาอังกฤษจาก Twitter ที่รวบรวมจากผู้ใช้งานที่เป็นดาราป่วยเป็นโรคซึมเศร้าจำนวน 15 คน และผู้ใช้งานที่เป็นดาราแต่ไม่เป็นโรคซึมเศร้าจำนวน 15 คน รวมทั้งหมด 30 คน โดยผู้ใช้งานแต่ละคนมีการโพสต์ข้อความมากกว่า 2 สัปดาห์ขึ้นไป

1.5 นิยามศัพท์เฉพาะ

โรคซึมเศร้า หมายถึง โรคทางจิตเวชชนิดหนึ่งที่มีผลต่อทั้งร่างกายและจิตใจ ผู้ที่ป่วยจะมีอาการทางอารมณ์ซึมเศร้า เช่น รู้สึกท้อแท้ หงอยเหงา เบื่อหน่าย นอนไม่หลับ สมาธิสั้น และรู้สึกว่าตนเองไร้ค่า จนนำไปสู่การฆ่าตัวตายเพื่อหาทางออกในที่สุด

บทที่ 2

ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎี และงานวิจัยที่เกี่ยวข้อง เพื่อสร้างแบบจำลองความคิดเห็นที่เข้าข่ายเป็นโรคซึมเศร้า การทำเหมืองความคิดเห็น ข้อมูลในการวิจัยต่าง ๆ ซึ่งผู้วิจัยได้ศึกษาค้นคว้าโดยมีรายละเอียดดังนี้

2.1 ทฤษฎีที่เกี่ยวข้อง

ในส่วนของทฤษฎีที่เกี่ยวข้อง ผู้วิจัยได้ศึกษาทฤษฎีต่าง ๆ ที่เกี่ยวข้อง คือ โรคซึมเศร้า การทำเหมืองความรู้สึก การทำเหมืองข้อมูล ตัวจำแนกประเภท การแบ่งข้อมูล การสร้างแบบจำลองแบบ Ensemble และการวัดประสิทธิภาพด้วย Confusion Matrix

2.1.1 โรคซึมเศร้า

องค์การอนามัยโลกให้ความหมายว่า เป็นอาการเจ็บป่วยทางจิตที่มีอาการเศร้าหมองอย่างต่อเนื่องและการสูญเสียความสนใจในการทำกิจกรรมที่ทำประจำ เช่น การนอนไม่หลับ มีความวิตกกังวล มีความร้อนรน ไม่กล้าตัดสินใจ มีความรู้สึกว่าตนเองไร้ค่า หรือความคิดที่ทำร้ายตัวเอง โดยอาการเหล่านี้เกิดขึ้นไม่ต่ำกว่าสองสัปดาห์ ถ้าไม่ได้รับการรักษาอาจนำไปสู่การฆ่าตัวตายได้ [1]

โรคซึมเศร้าหรือภาวมีซึมเศร้า หมายถึง โรคทางจิตเวชชนิดหนึ่งที่สามารถเกิดขึ้นได้กับทุกคนทุกเพศทุกวัย โดยจะมีภาวะที่จิตใจแสดงถึงความผิดปกติทางอารมณ์อย่างเด่นชัด ได้แก่ อารมณ์เบื่อหน่าย เศร้าตลอดเวลา ไม่มีความสุข หดหู่ ท้อแท้ง่าย มีอาการนอนไม่หลับ เบื่ออาหาร ความต้องการทางเพศลดลง รู้สึกหดหู่ สมาธิสั้น รู้สึกว่าตนเองไม่มีค่าและมองโลกในแง่ลบ ซึ่งส่งผลทำให้ไม่สะดวกในการใช้ชีวิต และถ้าไม่ได้รับการรักษาผู้ป่วยจะจบชีวิตตัวเองด้วยการฆ่าตัวตาย [9, 10]

สรุปได้ว่า โรคซึมเศร้าเป็นโรคที่เกี่ยวข้องกับอารมณ์ด้านลบเป็นหลัก ซึ่งมีผลในด้านร่างกายและจิตใจทำให้ใช้ชีวิตในประจำวันอย่างยากลำบาก ซึ่งผู้ที่คิดหาทางออกมักจบชีวิตด้วยการฆ่าตัวตาย

2.1.1.1 สาเหตุการเกิดโรคซึมเศร้า

สาเหตุของการเกิดโรคซึมเศร้าเกิดได้ทั้งปัจจัยภายในร่างกายและภายนอกร่างกาย โดยสามารถแบ่งได้เป็น 3 ปัจจัยหลักคือ ปัจจัยด้านชีววิทยา (Biological factors) ปัจจัยทางด้านจิตวิทยา (Psychological factors) และปัจจัยทางสังคม (Social factors) ซึ่งมีรายละเอียดดังนี้ [10]

1. ปัจจัยด้านชีววิทยา (Biological factors) เป็นปัจจัยที่ก่อให้เกิดโรคซึมเศร้าภายในร่างกายที่เกี่ยวข้องกับ พันธุกรรม (Genetic) สารชีวเคมี (Biochemical) ฮอร์โมน (Hormonal) และระบบประสาทสมอง (Nervous system) ซึ่งมีทั้งสามารถสืบทอดทางพันธุกรรมจากพ่อแม่สู่ลูกได้และได้รับสารเคมีจากภาพนอกโดยการฉีด รับประทาน หรือสัมผัส ทำให้สุขภาพจิตของตัวบุคคลนั้นเปลี่ยนไปได้ และเป็นสาเหตุของเกิดโรคซึมเศร้าในที่สุด

2. ปัจจัยทางด้านจิตวิทยา (Psychological factors) เป็นปัจจัยจากภายนอกร่างกายซึ่งอาจจะเป็นผลสืบเนื่องมาจากการใช้ชีวิตในวัยเด็กและสิ่งแวดล้อมการเลี้ยงดูในวัยเด็ก หากทั้ง 2 สิ่งนี้ได้รับประสบการณ์ด้านลบมาก ๆ เป็นเวลานาน จนทำให้ความคิดและบุคลิกบิดเบือนไปจากความเป็นจริงจนนำไปสู่สาเหตุของการเกิดโรคซึมเศร้าได้ในที่สุด

3. ปัจจัยทางสังคม (Social factors) เป็นปัจจัยจากภายนอกร่างกายที่มีผลกระทบต่อความสมดุลของอารมณ์ มักเกิดจากการปฏิสัมพันธ์ระหว่างบุคคลกับสิ่งแวดล้อม หรือบุคคลกับบุคคล เช่น ปัญหาด้านครอบครัว ปัญหาด้านความรัก การปรับตัวเข้ากับเพื่อน ความเชื่อ ศาสนาและวัฒนธรรมการเลี้ยงดูของแต่ละครอบครัวมีผลในการปรับตัวต่อปัญหาต่าง ๆ จนนำไปสู่สาเหตุของการเกิดโรคซึมเศร้าได้ในที่สุด

2.1.1.2 อาการการเกิดโรคซึมเศร้า

โรคซึมเศร้าจัดเป็นโรคที่อยู่ในกลุ่มของโรคที่เกี่ยวกับการที่มีอารมณ์แปรปรวน (Mood disorder) ซึ่งสามารถเกิดขึ้นได้กับทุกเพศทุกวัย อาการสำคัญที่เห็นได้ชัดส่วนใหญ่ ได้แก่ ความรู้สึกเบื่อหน่ายในสิ่งรอบตัว ความสนใจสิ่งแวดล้อมรอบข้างน้อยลง หมดหวัง วิตกกังวล รู้สึกว่าตนเองเป็นคนไร้ค่า ถ้าหากอารมณ์ซึมเศร้าเกิดขึ้นเป็นเวลานาน ๆ โดยไม่มีท่าว่าจะดีขึ้น อาจมีอาการต่าง ๆ ตามมา เช่น นอนไม่หลับ เบื่ออาหาร น้ำหนักลด รู้สึกไม่อยากมีชีวิตอยู่ ซึ่งลักษณะอาการที่สำคัญของโรคซึมเศร้า สามารถแบ่งเป็น 4 กลุ่มอาการใหญ่ ๆ ดังนี้ [10]

1. กลุ่มอาการทางอารมณ์ (Affective symptoms) เป็นการแสดงออกทางอารมณ์ ได้แก่ โกรธง่าย เศร้าง่าย ไม่สดชื่นแจ่มใส มีอารมณ์ท้อแท้ ไม่มีความสุข สะเทือนใจง่าย ร้องไห้บ่อย รู้สึกเบื่อหน่ายกับทุกสิ่งทุกอย่างรอบตัว และเก็บตัวไม่พูดจากับใคร เป็นต้น อาการเหล่านี้อาจส่งผลให้การมีความสัมพันธ์กับคนรอบข้างเปลี่ยนไปทางที่ไม่ดีได้

2. กลุ่มอาการทางกระบวนการคิด (Cognitive symptoms) ได้แก่ ลังเลในการตัดสินใจในเรื่องต่าง ๆ เกิดความไม่มั่นใจในตนเอง เกิดความคิดด้านลบ เกิดความคิดว่าตนเองว่าไร้คุณค่า ตำหนิตนเองว่าเป็นภาระให้แก่ผู้อื่น มีทัศนคติมองโลกในแง่ร้าย มีอาการเหม่อลอยบ่อย และสมาธิสั้น ส่งผลให้การใช้ชีวิตประจำวันติดขัด หรือเกิดความน้อยเนื้อต่ำใจจนพยายามฆ่าตัวตายได้

3. กลุ่มอาการทางกาย (Somatic symptoms) ได้แก่ มีอาการเหนื่อยง่ายอ่อนเพลียง่ายโดยไม่ทราบสาเหตุ นอนไม่หลับ หรือเบื่ออาหาร ซึ่งเกิดได้กับทุกเพศทุกวัย และมักจะเป็นอาการที่แสดงออกในช่วงแรก ๆ เมื่อเริ่มเป็นอาการซึมเศร้า

4. กลุ่มอาการด้านพฤติกรรม (Behavior symptoms) ได้แก่ มีพฤติกรรมเหม่อลอย มีพฤติกรรมเชื่องซึมเชื่องช้า มีพฤติกรรมพูดช้า หรือพูดเบากว่าปกติ มีพฤติกรรมชอบอยู่ตัวคนเดียวโดยไม่สนใจสิ่งแวดล้อม หรืออาจมีพฤติกรรมในการใช้สารเสพติดเพิ่มขึ้นจากเดิม เช่น ดื่มสุราบ่อยขึ้น สูบบุหรี่บ่อยขึ้น ใช้ยานอนหลับเมื่อมีอาการนอนไม่หลับ หรือการเรียกร้องความสนใจ เป็นต้น

สรุปได้ว่าอาการทั้ง 4 กลุ่ม อาจมีทั้งอาการที่แสดงออกอย่างชัดเจน และมีทั้งอาการแสดงไม่ชัดเจน ดังนั้นการวินิจฉัยโรคซึมเศร้าจำเป็นที่ต้องใช้เกณฑ์เพื่อช่วยในการคัดกรองผู้ที่เข้าข่ายเป็นโรคซึมเศร้าและการแยกระดับผู้เป็นโรคซึมเศร้า

2.1.1.3 การวินิจฉัยโรคซึมเศร้า

สมาคมจิตแพทย์อเมริกันได้มีการวินิจฉัยโรคซึมเศร้าโดยกำหนดไว้ในหนังสือ The 5th Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [11] มีหลักเกณฑ์ว่าผู้ที่เป็นโรคซึมเศร้าจะต้องมีอาการอย่างน้อย 5 อาการหรือมากกว่า โดยเกิดขึ้นติดต่อกันนานไม่ต่ำกว่า 2 สัปดาห์ ได้แก่ 1) มีอารมณ์ซึมเศร้า 2) ความสนใจหรือความสนุกสนานในการทำกิจกรรมต่าง ๆ ที่เคยทำทั้งหมดลดลงอย่างมาก 3) น้ำหนักลดลงอย่างชัดเจน 4) มีอาการนอนไม่หลับหรือหลับนานหลับบ่อยกว่าปกติ 5) การเคลื่อนไหวช้าลง 6) อ่อนเพลียไม่มีเรี่ยวแรง 7) รู้สึกว่าตนเองไร้ค่าหรือรู้สึกว่าตนเองผิดโดยไม่มีสาเหตุ 8) สมาธิสั้นหรือความสามารถในตัดสินใจลดลง 9) มีความคิดอยากฆ่าตัวตาย

2.1.1.4 การรักษาโรคซึมเศร้า

การบำบัดรักษาผู้ที่ป่วยที่เป็นโรคซึมเศร้ามีหลากหลายวิธีมีรายละเอียดดังนี้ [10]

1. การรักษาด้วยการใช้ยา จะใช้ยาต้านเศร้า (Anti-Depressant) ในบุคคลตั้งแต่อายุ 16 ปีขึ้นไป และยาต้านเศร้าในกลุ่มต่าง ๆ (Monoamine oxidase inhibitors, selective serotonin reuptake inhibitors, or tricyclic antidepressant) ยาพวกนี้ช่วยลดอาการซึมเศร้าอย่างฉับพลันในโรคซึมเศร้าทุกชนิดแต่อาจจะมีผลข้างเคียงบางประการ

2. การบำบัดทางจิต เป็นกระบวนการรักษาที่นำมาใช้กับผู้ที่มีอาการซึมเศร้าเล็กน้อยถึงปานกลางที่ไม่มีอาการทางจิต ควบคู่ไปกับการรักษาด้วยยา เนื่องจากการรักษาด้วยยาเป็น

เพียงการบำบัดอาการของผู้ป่วยให้ดีขึ้น แต่ไม่สามารถปรับกระบวนการคิดของผู้ป่วยให้ดีขึ้น ดังนั้น การนำกระบวนการจิตบำบัดมาใช้ในการรักษาจะช่วยให้บุคคลเรียนรู้วิธีการจัดการกับปัญหาชีวิต เช่น เรื่องสัมพันธภาพ กระบวนการหรือรูปแบบการคิด และพฤติกรรมที่อาจนำไปสู่การเป็นโรคซึมเศร้า ผลที่ได้จากการบำบัดจะทำให้บุคคลมีคุณภาพชีวิตที่ดีขึ้น

3. พฤติกรรมบำบัด (Behavior therapy) เป็นกระบวนการรักษาที่มุ่งเน้นเพิ่มพฤติกรรมที่ต้องการ หรือลดพฤติกรรมที่ไม่ต้องการโดยการสร้างทักษะการเรียนรู้ใหม่ขึ้น เช่น การฝึกสติสมาธิ การฝึกจังหวะหายใจ ฝึกลดความไวต่อสิ่งเร้า ฝึกการเผชิญหน้ากับสิ่งที่กลัว ฝึกทักษะการมีปฏิสัมพันธ์กับผู้อื่น และเบี่ยงเบนความคิด

4. การบำบัดทางจิตร่วมกับยาต้านเศร้า เป็นวิธีที่ต้องใช้ยาต้านโรคซึมเศร้าและการบำบัดรักษาทางจิตใจร่วมกัน โดยการรวมเข้าด้วยกันนั้นจะมีประสิทธิภาพดีกว่าการบำบัดรักษาวิธีใดวิธีหนึ่ง

5. การรักษาด้วยกระแสไฟฟ้า โดยจะทำในผู้ป่วยที่มีอายุมากกว่า 16 ปี ด้วยการรักษาด้วยกระแสไฟฟ้า วิธีนี้ทำให้อาการซึมเศร้าดีขึ้นเนื่องจากไฟฟ้าจะช่วยปรับสมดุลของสารสื่อประสาททำให้อาการซึมเศร้าของผู้ป่วยดีขึ้น

2.1.2 การวิเคราะห์ความรู้สึก (Sentiment Analysis)

การวิเคราะห์ความรู้สึก (Sentiment Analysis) คือ การนำเอาความคิดเห็นบนเครือข่ายสังคมออนไลน์ที่ความคิดเห็นส่วนใหญ่นิยมใช้ภาษาที่มีโครงสร้างประโยคที่ไม่แน่นอน (Unstructured Data) หรือเป็นภาษาธรรมชาติ (Natural Language) ที่ไม่ถูกต้องตามหลักไวยากรณ์ทางภาษา มาเข้ากระบวนการ Text Mining และ การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) เพื่อวิเคราะห์ความคิดเห็น ซึ่งเรียกการวิเคราะห์นี้ว่าการวิเคราะห์ความรู้สึก (Sentiment Analysis) หรือการทำเหมืองความคิดเห็น (Opinion Mining) [12, 13]

ดังนั้นสรุปได้ว่า Sentiment Analysis คือ การนำข้อความที่เป็นภาษาพูดและไม่เป็นทางการเข้ากระบวนการเพื่อวิเคราะห์บ่งบอกอารมณ์ และความรู้สึกจากข้อความนั้น เช่น ความรู้สึกดี (Positive) หรือความรู้สึกที่ไม่ดี (Negative)

2.1.3 การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล คือ การทำเหมืองข้อมูลเป็นการสกัดข้อมูลจำนวนมาก ๆ เพื่อหาความรู้ที่บ่งบอกถึงใจความสำคัญที่ซ่อนอยู่ในข้อมูลมหาศาล [14]

การทำเหมืองข้อมูล คือ การนำเอาข้อมูลสารสนเทศจำนวนมาก ๆ บนฐานข้อมูลต่าง ๆ มาเข้ากระบวนการทำเหมืองเพื่อหารูปแบบความสัมพันธ์ที่ซ่อนอยู่ในข้อมูลนั้น เพื่อนำไปประยุกต์ใช้งานเกี่ยวกับงานด้านต่าง ๆ เช่น ด้านธุรกิจ ด้านวิทยาศาสตร์ หรือด้านการแพทย์ เป็นต้น [15]

ดังนั้นสรุปได้ว่าการทำเหมืองข้อมูล คือ การนำเอาข้อมูลสารสนเทศบนฐานข้อมูลต่าง ๆ นำมามาเข้ากระบวนการทำเหมืองเพื่อวิเคราะห์และทำนายสิ่งต่าง ๆ ที่จะเกิดขึ้น

การทำเหมืองข้อมูลมีขั้นตอนดังนี้ [16]

2.1.3.1 การจัดเตรียมข้อมูล (Pre-Processing) เป็นกระบวนการแรกที่ต้องเตรียมข้อมูลให้อยู่ในรูปแบบที่นำไปเข้ากระบวนการทำเหมืองข้อมูลได้ ประกอบด้วยขั้นตอนต่อไปนี้

1. การคัดเลือกข้อมูล (Data Selection) คือ ขั้นตอนในการจัดเตรียมข้อมูลจากฐานข้อมูลต่าง ๆ มาไว้ในที่เดียวและต้องเป็นข้อมูลที่ต้องมีแหล่งที่มาชัดเจน

2. การกลั่นกรองข้อมูล (Data Cleaning) คือ ขั้นตอนในการตรวจสอบและแก้ไขข้อมูลที่ไม่ถูกต้อง เช่น 1) ข้อมูลขาดหาย (Missing Value) คือ ข้อมูลที่ขาดหายไปบางส่วน อาจจะเกิดจากข้อผิดพลาดของคน หรือเกิดจากความเสียหายของข้อมูลเอง โดยต้องใช้กระบวนการแก้ไขข้อมูลเหล่านี้ก่อนจะเข้ากระบวนการทำเหมือง มิฉะนั้นจะทำให้ผลลัพธ์คาดเคลื่อนได้ 2) ข้อมูลที่คาดเคลื่อนจากความเป็นจริง (Inaccurate Data) คือ ข้อมูลที่เกิดจากความตั้งใจของผู้ที่จงใจใส่ให้ข้อมูลเกินความเป็นจริง ทำให้ผลลัพธ์คาดเคลื่อนได้ 3) ข้อมูลที่มีความซ้ำซ้อน (Duplicate Data) คือ ข้อมูลที่ซ้ำกันทุกประการอย่างน้อย 2 ข้อมูลขึ้นไป โดยอาจเกิดจากการรวมข้อมูล 2 ฐานข้อมูลเข้าด้วยกัน

3. การรวบรวมข้อมูล (Data Integration) เป็นการรวบรวมข้อมูลจากหลาย ๆ แหล่งมาเก็บไว้ที่เดียวกัน เพื่อให้ผู้ใช้งานมองเห็นเป็นก้อนเดียวกัน

4. การแปลงข้อมูล (Data Transformation) เป็นการแปลงข้อมูลที่รวบรวมมาให้อยู่ในรูปแบบที่สามารถนำไปใช้กับอัลกอริธึมที่อยู่ในขั้นตอนการทำเหมืองข้อมูลได้

2.1.3.2 การตัดคำ (Word Segmentation) เป็นการตัดคำแต่ละคำออกจากรูปแบบประโยคเพื่อนำคำเหล่านั้นสร้างถุงคำ (Bag of Word) ที่ใช้สำหรับการนับความถี่ที่อยู่ในเอกสาร

2.1.3.3 การให้น้ำหนักคำ (Weighting) เป็นการให้น้ำหนักคำของแต่ละคำในถุงคำ โดยการให้น้ำหนักจะต่างไปตามอัลกอริธึมที่ใช้ในการให้น้ำหนัก ขึ้นอยู่กับประเภทของเอกสารเหล่านั้น

2.1.3.4 การคัดเลือกคุณลักษณะ (Features Selection) เป็นการเลือกคำที่มีแก่การนำไปสร้างแบบจำลอง ซึ่งการคัดเลือกคุณลักษณะช่วยให้สามารถลดเวลาในการสร้างแบบจำลองและสามารถเพิ่มค่าความถูกต้องได้

2.1.3.5 การสร้างแบบจำลองหรือการสร้างตัวแบบ (Modeling) เป็นกระบวนการในการเลือกเทคนิคที่เหมาะสมกับข้อมูลและงานที่ต้องการ ซึ่งสามารถทำการเลือกเทคนิคมากกว่าหนึ่งเทคนิคมาใช้งานได้ ซึ่งเทคนิคนี้เป็นการประมวลผลข้อมูลที่ผ่านกระบวนการ Pre-Processing มาแล้วเพื่อดึงเอาความรู้สำคัญออกมาจากข้อมูล โดยมีกระบวนการดังนี้

1. การเลือกอัลกอริธึม (Select Modeling Technique) เป็นขั้นตอนในการเลือกอัลกอริธึมที่จะนำมาใช้ในการเรียนรู้ข้อมูล เช่น Bayes, SVM และ Decision Table

2. การกำหนดรูปแบบของผลลัพธ์ (Build a Model) เป็นขั้นตอนในการเลือกรูปแบบผลลัพธ์ที่จะนำไปสรุปผล

3. การสร้างแบบจำลอง (Build a Model) เป็นขั้นตอนในการใช้อัลกอริธึมที่เลือกมาให้เรียนรู้ข้อมูลที่เราได้เตรียมไว้ เมื่อได้แบบจำลองออกมาแล้วสามารถนำไปทำนายผลข้อมูลชุดใหม่ได้

2.1.3.3 การประเมินผล (Evaluation) เป็นการนำตัวแบบมาเข้ากระบวนการประเมินโดยจะใช้ข้อมูลที่ใช้ในการสร้างมาทดสอบความถูกต้องก่อนนำไปใช้งานจริง

2.1.3.4 การนำไปใช้งานจริง (Deployment) เมื่อได้ทำการทดสอบตัวแบบเสร็จสิ้น โดยจะต้องมีความถูกต้อง (Accuracy) ในการทำนายข้อมูลและพยากรณ์ข้อมูลที่สูง เพื่อให้สามารถใช้ได้กับข้อมูลที่เป็นเป้าหมายได้อย่างถูกต้องที่สุด

2.1.4 ตัวจำแนกประเภท (Classifier)

ตัวจำแนกประเภท คือ อัลกอริธึมที่ใช้ในการจำแนกหมวดหมู่ของข้อมูลโดยการเรียนรู้จากแอตทริบิวต์ (Attribute) และทำการสร้างแบบจำลองที่สามารถพยากรณ์ข้อมูลชุดใหม่ที่นำมาทำนายได้โดยประกอบไปด้วย

2.1.4.1 กฎของเบย์ (Bayes)

เทคนิค Bayes เป็นเทคนิคอัลกอริธึมประเภทการเรียนรู้แบบมีผู้สอน (Supervised Learning Algorithm) ที่คิดค้นโดย Thorem Bayes โดยใช้หลักการความน่าแบบมีเงื่อนไขเข้ามา

พัฒนาทฤษฎีดังกล่าว จากแนวคิดของ Bayes สามารถทำนายเหตุการณ์ต่าง ๆ [17] ได้ด้วยสมการ 2.1

$$P(A|B) = \frac{P(A|B)\,P(A)}{P(B)} \tag{2.1}$$

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

กล่าวอธิบายสมการคือ เมื่อมีเหตุการณ์ฝนตกหนักจะสามารถไปเล่นฟุตบอลได้ดังสมการต่อไปนี้

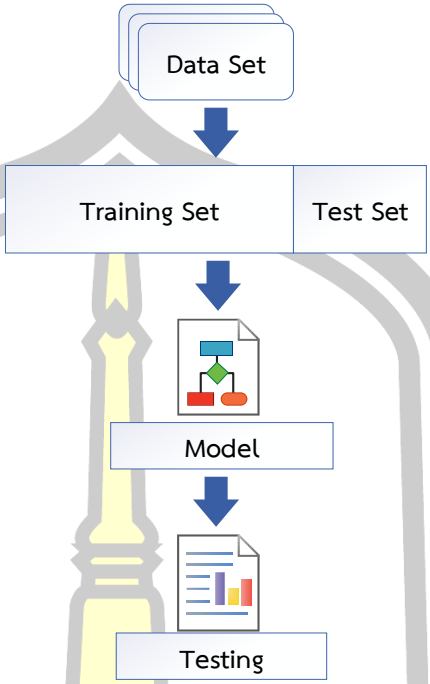
P (การไปเล่นฟุตบอล | ฝนตก) = P (ฝนตก | การไปเล่นฟุตบอล) x P (การไปเล่นฟุตบอล) / P (ฝนตก)

จากสมการตัวอย่างข้างต้นสามารถทำนายได้ว่า การไปเล่นฟุตบอลโดยให้สังเกตฝนตก อย่างไรก็ตามเหตุการณ์ที่ทำนายนั้นต้องสอดคล้องกัน เช่น ถ้าหากต้องการทำนายการไปเล่นฟุตบอลจะต้องไม่ใช้เหตุการณ์น้ำท่วมเข้ามาเกี่ยวข้อง เพราะเหตุการณ์ทั้งสองไม่มีความสอดคล้องกัน

2.1.5 การแบ่งข้อมูล Cross Validation

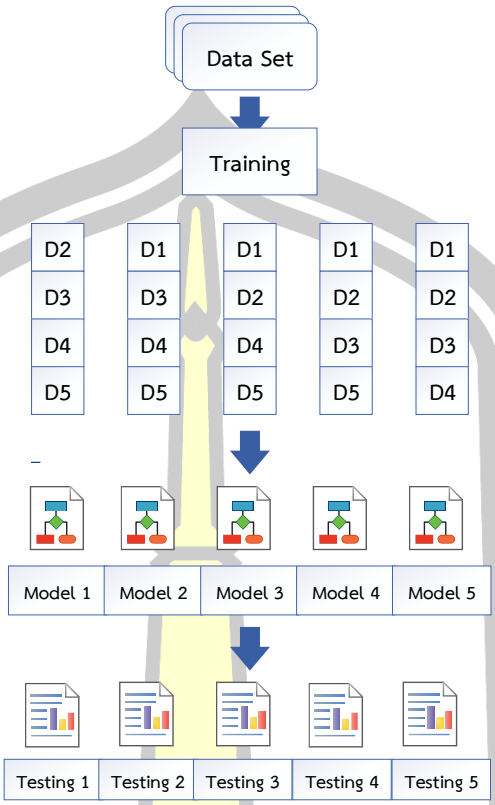
การแบ่งข้อมูล Cross Validation เป็นวิธีการแบ่งข้อมูลเพื่อนำมาทดสอบวัดประสิทธิภาพการทำนายของแบบจำลองเพื่อให้เกิดความน่าเชื่อถือของแบบจำลอง โดยประกอบไปด้วย [19, 20]

2.1.5.1 Hold Out Method เริ่มต้นจะแบ่งข้อมูลออกเป็น 2 ชุด คือ ชุดการสอน (Training Set) และชุดทดสอบ (Test Set) โดยส่วนมากจะนิยมแบ่งชุดการสอนออกเป็น 80% จากข้อมูลทั้งหมด และชุดทดสอบ 20% จากข้อมูลทั้งหมดดังรูปที่ 2.7 จากนั้นก็จะนำข้อมูลการสอนไปสร้างแบบจำลอง และนำชุดทดสอบเข้ามาทำการทดสอบรอบเดียว



รูปที่ 2.1 การแบ่งข้อมูลแบบ Hold Out Method

2.1.5.2 K-Fold Cross Validation เริ่มต้นจะทำการแบ่งข้อมูลออกเป็นหลาย ๆ ส่วน โดยการกำหนดค่าด้วย K โดยมีความหมายดังนี้ 5-Fold Cross Validation คือ การแบ่งข้อมูลออกเป็น 5 ส่วน โดยแต่ละส่วนมีข้อมูลเท่ากัน หลังจากนั้นก็นำส่วนแรงเข้าไปทดสอบประสิทธิภาพของแบบจำลองจนเสร็จ แล้ววนซ้ำจนครบจำนวน K ที่กำหนดไว้ดังรูปที่ 2.8

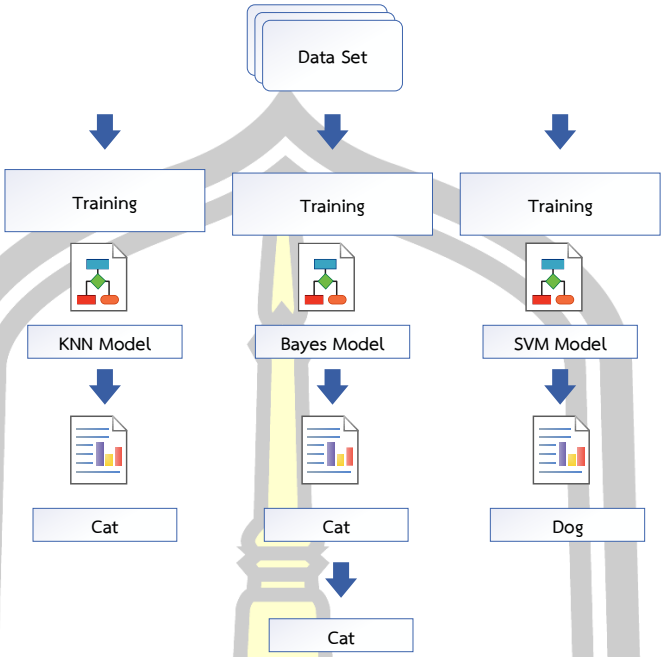


รูปที่ 2.2 การแบ่งข้อมูลแบบ K-Fold Cross Validation

2.1.6 การสร้างโมเดลแบบ Ensemble

เทคนิค Ensemble เป็นเทคนิคที่มีประสิทธิภาพสูงในการสร้างโมเดลเพื่อให้สามารถทำนายได้อย่างมีประสิทธิภาพสูงสุด โดยหลักการจะใช้ Classification หลาย ๆ โมเดลเข้ามาช่วยในการหาผลลัพธ์ หลักการของเทคนิค Ensemble มีอยู่ 3 หลักการ [24] ได้แก่

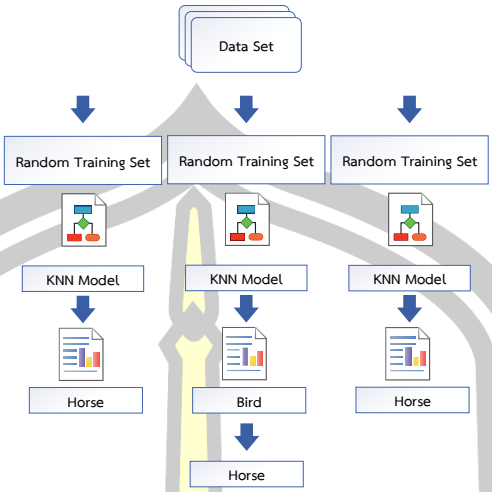
2.1.6.1 Vote Ensemble เป็นเทคนิคที่นำ Data Train ชุดเดียวมาให้ Classification ที่แตกต่างกันสร้างโมเดล ดังรูป มีการใช้งาน Training Data ชุดเดียวแต่มีการใช้งาน Classification ทั้งหมด 3 ตัวในการสร้างโมเดลขึ้นมา จากนั้นนำ Data Test ชุดเดียวกันมาเข้าโมเดลทั้ง 3 ตัว เพื่อให้ทำนายผลลัพธ์ออกมา ดังรูปที่ 2.3



รูปที่ 2.3 แนวคิดของเทคนิค Vote Ensemble

จากรูปที่ 2.3 ผลลัพธ์ที่ได้จากโมเดลที่ 1 ทำนายได้เป็น Cat โมเดลที่ 2 ทำนายได้เป็น Cat แต่โมเดลที่ 3 ทำนายเป็น Dog จากนั้นจะมีการ Vote เอาผลลัพธ์ที่มากสุดคือ Dog เนื่องจากโมเดลที่ 1 และ 2 ทำนายว่าเป็น Dog โดยมีอัตราส่วนมากกว่าโมเดลที่ 3 ที่ทำนายว่าเป็น Cat

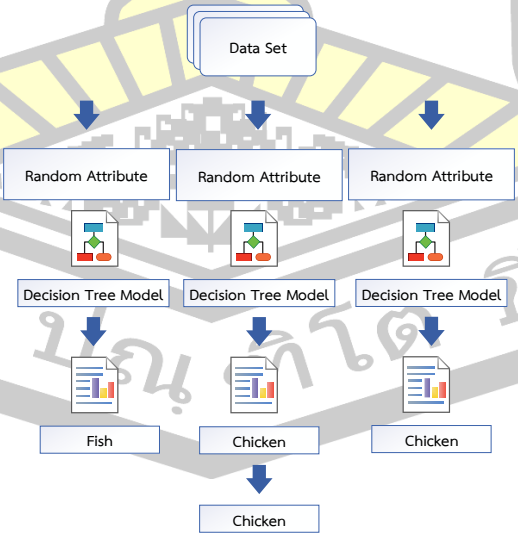
2.1.6.2 Bootstrap Aggregating หรือ Bagging เป็นเทคนิคการสุ่ม Data Train ให้ Classification ตัวเดียวสร้างหลาย ๆ โมเดลแล้วทำการ Vote หาผลลัพธ์ที่มีผล Vote มากที่สุด โดยจะทำการสุ่ม Data Train มาให้ต่างกัน จากนั้นทำการสร้างโมเดลด้วยอัลกอริธึมต่าง ๆ จากนั้นนำ Data Test ชุดเดียวกันมาเข้าไปทำนาย ต่อมาจะได้ผลลัพธ์ของแต่ละโมแดลจากนั้นนำมาทำการ Vote เลือกเอาผลลัพธ์ที่ดีที่สุด ดังรูปที่ 2.4 ที่ใช้อัลกอริธึมเทคนิค KNN ในการทำนาย



รูปที่ 2.4 แนวคิดของเทคนิค Bootstrap Aggregating

จากรูปที่ 2.4 ผลลัพธ์ที่ได้จากโมเดลที่ 1 ทำนายเป็น Horse โมเดลที่ 2 ทำนายเป็น Bird และโมเดลที่ 3 ทำนายเป็น Horse จากนั้นจะมีการ Vote เอาผลลัพธ์ที่มากที่สุดคือ Horse เนื่องจากโมเดลที่ 1 และ 3 ทำนายว่าเป็น Horse โดยมีอัตราส่วนมากกว่าโมเดลที่ 2 ที่ทำนายว่าเป็น Bird

2.1.6.3 Random Forest เป็นเทคนิคการสุ่ม Data Train และสุ่มแอตทริบิวต์ (ฟิวเจอร์) ต่าง ๆ ออกเป็นหลายชุด แล้วสร้างโมเดลด้วย Classification Decision Tree อย่างเดียว จากนั้นนำ Data Test ชุดเดียวกันเข้าไปทำนาย ต่อมาจะได้ผลลัพธ์ของแต่ละโมแดลและนำมาทำการ Vote เลือกเอาผลลัพธ์ที่ดีที่สุด ดังรูปที่ 2.5



รูปที่ 2.5 แนวคิดของเทคนิค Random Forest

จากรูปที่ 2.5 ผลลัพธ์ที่ด้จากโมเดลที่ 1 ทำนายเป็น Fish โมเดลที่ 2 ทำนายเป็น Chicken และโมเดลที่ 3 ทำนายเป็น Chicken จากนั้นจะมีการ Vote เอาผลลัพธ์ที่มากที่สุดคือ Chicken เนื่องจากโมเดลที่ 2และ 3 ทำนายว่าเป็น Chicken โดยมีอัตราส่วนมากกว่าโมเดลที่ 1 ที่ทำนายว่าเป็น Fish

2.1.7 การวัดประสิทธิภาพ

การวัดประสิทธิภาพด้วย Confusion Matrix เป็นการประเมินผลประสิทธิภาพการทำนายของแบบจำลองโดยการนำผลการทดลองที่เกิดขึ้นจากการทำนายมาเปรียบเทียบกับผลลัพธ์จริง โดยอาศัยการคำนวณค่าต่าง ๆ จากตารางที่ Confusion ดังตารางที่ 2.1 [21]

ตารางที่ 2.1 Confusion Matrix

		ค่าทำนาย	
ค่าความจริง	**Classes**	**Yes**	**No**
	Yes	TP	FN
	No	FP	TN

โดยที่

True Positive (TP) คือ สิ่งที่แบบจำลองทำนายว่าจริงและผลลัพธ์บอกว่าจริง

True Negative (TN) คือ สิ่งที่แบบจำลองทำนายว่าไม่จริงและผลลัพธ์บอกว่าไม่จริง

False Positive (FP) คือ สิ่งที่แบบจำลองทำนายว่าจริงและผลลัพธ์บอกว่าไม่จริง

False Negative (FN) คือ สิ่งที่แบบจำลองทำนายว่าไม่จริงและผลลัพธ์บอกว่าจริง

สิ่งที่ได้จากการทำนายของตารางที่ Confusion ได้แก่

2.1.7.1 ค่าความถูกต้อง Accuracy คือ ค่าที่บ่งบอกว่าแบบจำลองสามารถทำนายได้ถูกต้องเท่าไรจากผลลัพธ์ทั้งหมด ดังสมการ 2.2

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2.2}$$

2.1.7.2 ค่าความระลึก Recall (True Positive Rate) คือ ค่าที่บ่งบอกว่าแบบจำลองทำนายว่าจริงเป็นอัตราส่วนเท่าไรของผลลัพธ์จริงทั้งหมด ดังสมการ 2.3

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2.3}$$

2.1.7.3 True Negative Rate (TNR) คือ ค่าที่บ่งบอกว่าแบบจำลองทำนายว่าไม่จริงเป็นอัตราส่วนเท่าไรของผลลัพธ์จริงทั้งหมด ดังสมการ 2.4

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{2.4}$$

2.1.7.4 False Positive Rate (TPR) คือ ค่าที่บ่งบอกว่าแบบจำลองทำนายว่าจริงเป็นอัตราส่วนเท่าไรของผลลัพธ์ไม่จริงทั้งหมด ดังสมการ 2.5

$$\text{TPR} = \frac{\text{FP}}{\text{TN} + \text{FP}} \tag{2.5}$$

2.1.7.5 False Negative Rate (FNR) คือ ค่าที่บ่งบอกว่าแบบจำลองทำนายว่าไม่จริงเป็นอัตราส่วนเท่าไรของผลลัพธ์จริงทั้งหมด ดังสมการ 2.6

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \tag{2.6}$$

2.1.7.6 Precision คือ ค่าความแม่นยำที่บ่งบอกว่าแบบจำลองทำนายถูกต้องทั้งหมดเป็นค่าเท่าไร ดังสมการ 2.7

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.7}$$

2.2 งานวิจัยที่เกี่ยวข้อง

งานวิจัยในส่วนนี้ผู้วิจัยได้ศึกษาวิจัยต่าง ๆ ที่เกี่ยวกับการวิเคราะห์หาผู้ที่เข้าข่ายเป็นโรคซึมเศร้าด้วยวิธีการทำเหมืองต่าง ๆ โดยมีรายละเอียดดังนี้

2.2.1 งานวิจัยที่เกี่ยวข้องกับการจำแนกข้อมูลโรคซึมเศร้า

ในปี ค.ศ. 2013 งานวิจัยของ Wang และคณะ [7] ได้ทำการวิจัยเกี่ยวกับการสร้างแบบจำลองตรวจหาผู้ป่วยโรคซึมเศร้าในโซเชียลมีเดียจำพวก Micro-blog วัตถุประสงค์ที่ศึกษานั้นต้องการนำแบบจำลองไปสร้างโปรแกรมช่วยทำนายหาผู้ที่เข้าข่ายผู้ป่วยโรคซึมเศร้าให้กับแพทย์ไว้เป็นตัวช่วยประกอบการตัดสินใจสำหรับการวินิจฉัยโรค โดยข้อมูลที่ใช้จะเป็นภาษาจีนที่ได้มาจาก Twitter และ Sina Micro-blog เช่น Weibo, Renren และ QQ เป็นต้น จากนั้นนำมาเข้ากระบวนการทำเหมืองและเปรียบเทียบ Classifier ทั้ง 3 ตัวได้แก่ Bayes, Trees และ Decision Table และทำการแบ่งข้อมูลโดยใช้เทคนิค K-Fold Cross Validation ให้ K = 10 ผลการวิจัยพบว่า Classifier ที่ดีที่สุดคือ Bayes โดยได้ค่า Mean absolute error = 1.86%, ROC Area = 90.8% และ ค่า F-measure = 85% จากการวิเคราะห์ผลการทดลองทำให้ทราบว่าการนำวิธีการคำนวณขั้วของประโยค (Polarity calculation of sentence) เข้ามาทำในขั้นตอนเตรียมข้อมูลทำให้ความถูกต้องของการทำนายเพิ่มขึ้น

ในปี ค.ศ. 2015 งานวิจัยของ Tsugawa และคณะ [5] ได้ทำวิจัยเกี่ยวกับการใช้เทคนิคการทำเหมืองความคิดเห็นเพื่อหาผู้ป่วยที่เข้าข่ายเป็นโรคซึมเศร้า โดยวัตถุประสงค์ที่ศึกษาผู้วิจัยต้องการนำแบบจำลองที่ผ่านการทดสอบแล้วมาเป็นตัวช่วยให้แพทย์วินิจฉัยได้อย่างถูกต้องที่สุด งานวิจัยนี้ใช้ข้อมูลจากผู้ใช้งาน Twitter จำนวน 3,200 tweets และมีคำศัพท์ที่เกี่ยวข้องกับโรคซึมเศร้าจำนวน 1,622 ข้อความ แบ่งออกเป็น 2 Class ได้แก่ คำศัพท์ที่บ่งบอกเป็นโรคซึมเศร้าจำนวน 862 คำ และคำศัพท์ที่บ่งบอกว่าไม่เป็นโรคซึมเศร้าจำนวน 760 คำ จากนั้นมาเข้ากระบวนการทำเหมืองด้วย Classifier SVM และทำการแบ่งข้อมูลโดยใช้เทคนิค K-Fold Cross Validation ให้ K = 10 เพื่อให้ได้ค่าความถูกต้องที่ดีที่สุด จากผลวิจัยพบว่าค่า Accuracy = 69% , ค่า precision = 64%, ค่า recall = 43% และค่า F-measure = 52% จากผลการทดลองของงานวิจัยคาดการณ์ว่าปัญหาที่ทำให้ความถูกต้องของการทดลองไม่สูงเนื่องมาจากข้อความที่นำเข้ามาสร้างแบบจำลองและเข้ารับการทำนายนั้นอาจสั้นเกินไปจนไม่สามารถทำนายได้เพราะอุปนิสัยการพูดของคนญี่ปุ่นนั้นมีนิสัยการพูดที่น้อย

ในปี ค.ศ. 2016 งานวิจัยของ Kang และคณะ [4] ได้ทำการวิจัยเกี่ยวกับการจำแนกหาผู้ที่เข้าข่ายเป็นโรคซึมเศร้าจากการข้อความความคิดเห็น, Emoticon และรูปของการโพสต์บน Twitter วัตถุประสงค์ที่ศึกษานั้นเพื่อต้องการสร้างแบบจำลองที่สามารถนำเอาข้อความและรูปเข้ามาจำแนกข้อความที่เข้าข่ายเป็นโรคซึมเศร้าได้ โดยข้อมูลที่ใช้นั้นได้จาก Twitter จำนวน 10,000 ข้อความ โดยแบ่งเป็นข้อความที่แสดงออกถึงการเป็นโรคซึมเศร้าจำนวน 2,500 ข้อความ และข้อความที่ไม่แสดงออกถึงโรคซึมเศร้าจำนวน 2,500 ข้อความ จากนั้นนำมาเข้ากระบวนการทำเหมืองข้อมูลด้วย Classifier SVM พบว่าการใช้เฉพาะข้อความความคิดเห็นและ Emoticon มีค่า Accuracy =

84.45%, ค่า Precision = 81.01%, ค่า Recall = 83.50% และค่า F1 = 82.24% ส่วนการใช้ข้อความความคิดเห็น, Emoticon และรูป มีค่าความถูกต้อง 90.04%, ค่า Precision = 86.01%, ค่า Recall = 87.51% และค่า F1 = 86.72% จากการวิเคราะห์ผลการทดลองทำให้ทราบว่าการนำเอาข้อความความคิดเห็น, Emoticon และรูป เข้ามาทำนายทำให้มีค่าความถูกต้องที่สูงตามไปด้วย เนื่องจากมีการให้น้ำหนักจาก Emoticon เข้าช่วยทำให้ประโยคสั้นได้รับการทำนายถูกมากขึ้น

ในปี ค.ศ. 2016 งานวิจัยของ Nadeem และคณะ [22] ได้ทำการวิจัยเกี่ยวกับการตรวจจับโรคซึมเศร้าจาก Twitter วัตถุประสงค์ที่ศึกษานั้นเพื่อต้องการสร้างแบบจำลองที่สามารถจำแนกผู้ที่เข้าข่ายเป็นโรคซึมเศร้าเพื่อให้หน่วยงานด้านสุขภาพใช้งาน โดยข้อมูลที่ใช้นั้นได้จาก Twitter จำนวน 3,000 ข้อความ จากนั้นนำมาเข้ากระบวนการทำเหมืองข้อมูลด้วย Classifier ทั้ง 4 ตัว ได้แก่ Decision Tree, SVM, Ridge และ Bayes อีกทั้งยังใช้วิธีการทางสถิติ Logistic Regression พบว่าการใช้งาน Logistic Regression ให้ Precision สูงสุดอยู่ที่ 0.86 และ F1-Score สูงสุดอยู่ที่ 0.84 ส่วน SVM ให้ Recall สูงสุดอยู่ที่ 0.83 ส่วน Bayes ให้ Accuracy สูงสุดอยู่ที่ จากผลการทดลองพบว่าการใช้ Bayes สร้างแบบจำลองทำให้มีค่า Accuracy มากกว่า Classifier อื่น ๆ ถึง 4% – 19% และการใช้งานอัลกอริธึม Bayes ใช้เวลาในการ Training Data น้อยกว่า Classifier อื่น ๆ

ในปี ค.ศ. 2017 งานวิจัยของ Aldarwish และคณะ [6] ได้ทำการวิจัยเกี่ยวกับการจัดระดับความรุนแรงของโรคซึมเศร้าจากการโพสต์บนโซเชียลมีเดีย วัตถุประสงค์ที่ศึกษานั้นเพื่อต้องการสร้างแบบจำลองที่สามารถแยกระดับผู้ป่วยที่เป็นโรคซึมเศร้าเพื่อช่วยให้แพทย์วินิจฉัยและรักษาให้ถูกต้องตามระดับของโรคซึมเศร้า โดยข้อมูลที่ใช้นั้นได้จาก Facebook, LiveJournal และ Twitter จำนวน 6,773 ข้อความ โดยแบ่งเป็นข้อความที่แสดงออกถึงการเป็นโรคซึมเศร้าจำนวน 2,073 ข้อความ และข้อความที่ไม่แสดงออกถึงโรคซึมเศร้าจำนวน 4,700 ข้อความ แต่แก้ไขการเกิดปัญหา Imbalanced data ด้วยการสุ่มข้อความที่ไม่แสดงออกถึงโรคซึมเศร้าให้เหลือ 2,073 ข้อความ ตัวอย่างข้อความ ดังตารางที่ 2.2 จากนั้นนำมาเข้ากระบวนการทำเหมืองข้อมูลด้วย Classifier ทั้ง 2 ตัว ได้แก่ SVM และ Bayes พบว่าการใช้งาน Bayes มีค่า Accuracy 63%, ค่า Precision = 100% และค่า Recall = 57% จากการวิเคราะห์ผลการทดลองคาดการณ์ว่าปัญหาที่ทำให้ผู้ทดลองได้ค่าความถูกต้องที่ไม่สูงเนื่องมาจากปัญหาที่ผู้ทำการวิจัยอาจนำข้อความที่เป็นคำสแลงเข้ามาทำนาย ทำให้ค่าความถูกต้องลดน้อยลงได้

ตารางที่ 2.2 ตัวอย่างข้อมูลในงานวิจัยของ M. M. Aldarwish [6]

ลักษณะอาการ	ตัวอย่างข้อความ
Sadness	I just found out my mom never wanted me in the first place; That just ruined my day.
Loss of Interest	What, if anything, is there to live for?
Appetite	I was a little depressed that I ate so much last night there were no leftovers today.
Sleep	It was a sleepless night.
Thinking	I can't concentrate.
Guilt	I feel bad for doing it.
Tired	too worried and tired to post tonight
Movement	I think I just gave myself permission to be lazy.
Suicidal ideation	I did it again. I don't know what I was thinking. I cut a star-like design into my upper left arm, and then took a whole bunch of pills and strong scotch. life is not going well.

2.2.2 งานวิจัยที่เกี่ยวข้องกับเทคนิคที่จะใช้ในงานวิจัย

ในปี ค.ศ. 2015 งานวิจัยของ Punnee และคณะ [8] ได้ทำงานวิจัยเกี่ยวกับการจำแนกประเภทของผู้ป่วยที่เป็นโรคเบาหวานด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งงานวิจัยนี้ได้นำเอาหลักการ Ensemble เข้ามาช่วยให้แบบจำลองมีความน่าเชื่อถือในการจำแนกข้อมูลผู้ป่วยโรคเบาหวานได้อย่างถูกต้องที่สุด โดยนำ Classifier ทั้ง 5 ตัว ได้แก่ Decision Tree, Back Propagation Neural Network, SVM, KNN และ Bayes จากนั้นสร้างแบบจำลองการทำนายด้วย Train Set ตัวเดียวกันแล้วนำเอาข้อมูลชุด Test Set เข้ามารับการทำนาย โดยผลลัพธ์ที่ได้จากการทำนายแต่ละ Classifier จะนำมาเข้ารับการ Vote เลือกเอาผลลัพธ์ที่ดีที่สุด เพื่อให้ได้ผลลัพธ์ที่แม่นยำที่สุดในการทำนายแต่ละข้อมูล จากขั้นตอนการทดลองทำให้ทราบว่าการนำวิธีการ Ensemble เข้ามาใช้งานในการหาผลลัพธ์ที่ดีที่สุดนั้นเป็นวิธีการที่เหมาะกับการทำข้อมูลเกี่ยวกับด้านการแพทย์ เพราะข้อมูลทางการแพทย์ต้องการผลลัพธ์ที่ถูกต้องและแม่นยำที่สุด เนื่องจากข้อมูลด้านนี้มีผลต่อชีวิตมนุษย์โดยตรง

จากงานวิจัยที่เกี่ยวข้องทั้งหมดทำให้เรียนรู้ว่างานวิจัยที่เกี่ยวกับด้านการแพทย์ต้องพัฒนาเอาหลักการต่าง ๆ เข้ามาผสมผสานเพื่อให้ได้ความถูกต้องที่สูงที่สุด เนื่องจากการวินิจฉัยโรคเป็นเรื่องที่ละเอียดอ่อนมีผลต่อชีวิตของมนุษย์โดยตรง ฉะนั้นการนำเอาวิธีการต่าง ๆ ที่ได้จากงานวิจัยที่เกี่ยวข้องมาผสมผสานเพื่อให้ได้ค่าความถูกต้องที่ดีที่สุด เพื่อใช้ประกอบการตัดสินใจของแพทย์ใช้วินิจฉัยโรคซึมเศร้าได้อย่างแม่นยำ โดยในงานวิจัยนี้เลือกนำเอาเทคนิคหลักการดังนี้เข้ามาช่วยในงานวิจัย ได้แก่ การนำเอาข้อความความคิดเห็นเข้ามาทำนาย, การนำเอาข้อความแสดงความรู้สึกเข้ามาทำนาย และการนำอัลกอริธึม Bayes เข้ามาหาผลลัพธ์ที่ดีที่สุดในการทำนายเพื่อเพิ่มความถูกต้องในการทำนายและลดระยะเวลาในการสร้างแบบจำลอง

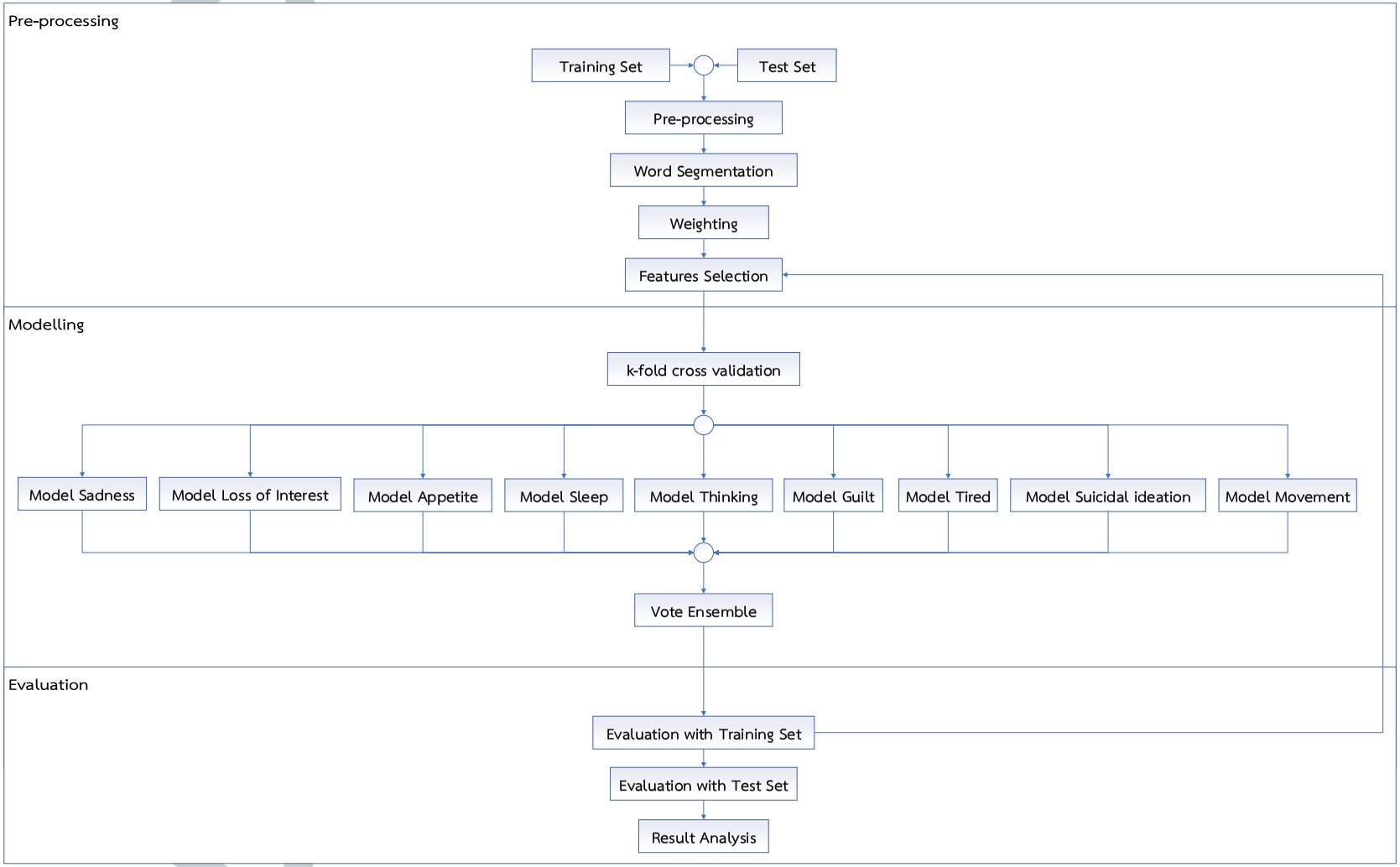
บทที่ 3

วิธีดำเนินการวิจัย

วิธีการดำเนินงานวิจัยนี้ ใช้วิธีการทำเหมืองความคิดเห็น ซึ่งมีขั้นตอนในงานวิจัยทั้งหมด 6 ขั้นตอน ประกอบไปด้วย การเตรียมข้อมูล การสร้างแบบจำลอง การวัดประสิทธิภาพของแบบจำลอง การนำเสนอการวิเคราะห์โรคซึมเศร้า และการพัฒนาระบบ โดยมีขั้นตอนดังรูปที่ 3.1

3.1 ทำความเข้าใจปัญหา

ในสังคมปัจจุบันการสื่อสารในโลกโซเชียลมีเดียมีความสะดวกสบาย ซึ่งทำให้การปฏิสัมพันธ์ในโลกความจริงนั้นเปราะบางลงและด้วยสิ่งเร้าต่าง ๆ ทำให้เกิดผลกระทบทางจิตใจนำไปสู่การเกิดโรคซึมเศร้าตามมา ซึ่งการระบายอารมณ์ด้านลบและการแสดงความรู้สึกที่อัดอยู่ในใจ บางครั้งก็ไม่สามารถพูดกับบุคคลในโลกความจริงได้อย่างเปิดใจ ทำให้ผู้ที่เป็นโรคซึมเศร้าหันมาแสดงความรู้สึกด้านลบในโซเชียลมีเดีย ดังนั้นการที่จะสามารถแยกแยะผู้ที่เข้าข่ายเป็นโรคซึมเศร้า ได้อย่างคร่าว ๆ สามารถนำข้อความที่แสดงออกถึงความรู้สึกด้านลบมาทำนายข้อความที่เข้าข่ายเป็นโรคซึมเศร้าได้

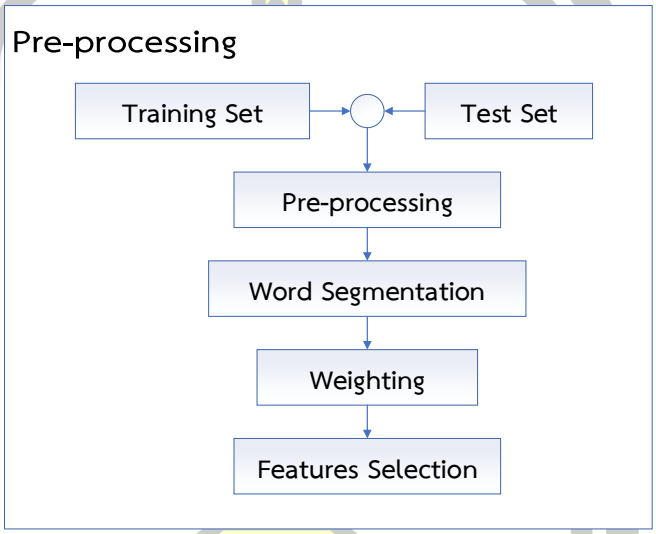


รูปที่ 3.1 ขั้นตอนวิธีการดำเนินงานวิจัย

3.2 การเตรียมข้อมูล

การเตรียมข้อมูล (Data Pre-processing)

ขั้นตอนในการเตรียมข้อมูลเป็นขั้นตอนที่สำคัญอย่างมากในการทำเหมืองความคิดเห็น ในงานวิจัยนี้มีขั้นตอนในการเตรียมข้อมูลดังรูปที่ 3.2



รูปที่ 3.2 แผนผังขั้นตอนการทำเหมืองความคิดเห็น

จากรูปที่ 3.2 สามารถแบ่งขั้นตอนในการเตรียมข้อมูลออกเป็นขั้นตอนย่อยดังต่อไปนี้

3.2.1 ข้อมูลสำหรับการเรียนรู้แบบจำลอง (Training Set)

ข้อมูลสำหรับการเรียนรู้แบบจำลอง คือ ข้อความภาษาอังกฤษที่เก็บรวบรวมจากการติด Hashtag บน Twitter ของผู้ใช้งานทั่วไป โดยแบ่งออกตามลักษณะ 9 อาการ ที่บ่งบอกถึงโรคซึมเศร้า มีจำนวนทั้งหมดดังตารางที่ 3.1

ตารางที่ 3.1 การเก็บข้อมูลโดยเก็บจาก Hashtag

กลุ่มของอาการ	**จำนวนข้อความตามอาการ**	**จำนวนข้อความที่ไม่เป็นไปตามอาการ**
1) ข้อความเกี่ยวกับอารมณ์ซึมเศร้า	3,000 ข้อความ	3,000 ข้อความ
2) ข้อความเกี่ยวกับการขาดความสนใจลดลง	3,000 ข้อความ	3,000 ข้อความ
3) ข้อความเกี่ยวกับน้ำหนักผิดปกติ	3,000 ข้อความ	3,000 ข้อความ
4) ข้อความเกี่ยวกับการนอนผิดปกติ	3,000 ข้อความ	3,000 ข้อความ

ตารางที่ 3.2 การเก็บข้อมูลโดยเก็บจาก Hashtag (ต่อ)

กลุ่มของอาการ	**จำนวนข้อความตามอาการ**	**จำนวนข้อความที่ไม่เป็นไปตามอาการ**
1) ข้อความเกี่ยวกับอารมณ์ซึมเศร้า	3,000 ข้อความ	3,000 ข้อความ
2) ข้อความเกี่ยวกับการขาดความสนใจลดลง	3,000 ข้อความ	3,000 ข้อความ
3) ข้อความเกี่ยวกับน้ำหนักผิดปกติ	3,000 ข้อความ	3,000 ข้อความ
4) ข้อความเกี่ยวกับการนอนผิดปกติ	3,000 ข้อความ	3,000 ข้อความ
5) ข้อความเกี่ยวกับร่างกายอ่อนเพลีย	3,000 ข้อความ	3,000 ข้อความ
6) ข้อความเกี่ยวกับการรู้สึกตนเองไร้ค่า	3,000 ข้อความ	3,000 ข้อความ
7) ข้อความเกี่ยวกับสมาธิสั้น	3,000 ข้อความ	3,000 ข้อความ
8) ข้อความเกี่ยวกับการเคลื่อนไหลช้า	3,000 ข้อความ	3,000 ข้อความ
9) ข้อความเกี่ยวกับการคิดฆ่าตัวตาย	3,000 ข้อความ	3,000 ข้อความ
ทั้งหมด	54,000 ข้อความ	

จากตารางที่ 3.1 มีการเก็บรวบรวมข้อมูลจาก Twitter ในส่วนข้อมูลสำหรับสร้างแบบจำลอง ผู้วิจัยได้ใช้งาน RapidMiner studio ในการดึงข้อความที่มีการติด Hashtag ดังตารางที่ 3.2

ตารางที่ 3.3 การเก็บข้อมูลโดยเก็บจาก Hashtag

ลำดับ	อาการ	Hashtag
1	อารมณ์ซึมเศร้า	#Sadness #Depressive
2	ขาดความสนใจลดลง	#Loss of Interest #Lose interest
3	น้ำหนักผิดปกติ	#Appetite #Hunger
4	การนอนผิดปกติ	#Sleepless #Hethargy
5	ร่างกายอ่อนเพลีย	#Un thinking #Out thinking
6	การรู้สึกตนเองไร้ค่า	#Guilt #Disgrace #Dishonor
7	สมาธิสั้น	#Tired #Bored #Fatigued
8	การเคลื่อนไหลช้า	#Lackadaisical #Lazy #Loafing #Phlegmatic
9	การอยากฆ่าตัวตาย	#Suicidal #Dangerous #Destructive

จากตารางที่ 3.1 สามารถแสดงตัวอย่างข้อมูลใน 1 Instant ได้ดังตารางที่ 3.3

ตารางที่ 3.4 ตัวอย่างข้อมูล Data Train

ตัวแปร	ตัวอย่างข้อมูล	คำอธิบายตัวแปร
Id	672849834257158144	รหัสของโพสต์นั้น รูปแบบ Integer
Text	I am so sad, angry, heart broken... I hate today.	ข้อความสถานะของโพสต์รูปแบบ UTF-8

3.2.2 ข้อมูลสำหรับทดสอบแบบจำลอง (Test Set)

ข้อมูลสำหรับทดสอบแบบจำลอง คือ ข้อความภาษาอังกฤษจาก Twitter ที่รวบรวมจากผู้ใช้งานที่เป็นดาราป่วยเป็นโรคซึมจำนวน 15 คน และผู้ใช้งานที่เป็นดาราแต่ไม่เป็นโรคซึมเศร้า 15 คน รวมทั้งหมด 30 คน โดยผู้ใช้งานแต่ละคนมีการโพสต์ข้อความมากกว่า 2 สัปดาห์ขึ้นไป ตัวอย่างดาราที่ป่วยเป็นโรคซึมเศร้า [25, 26, 27] เช่น Chester Bennington, Kristen Bell และ Lady Gaga เป็นต้น ในการใช้งานข้อมูลสำหรับทดสอบแบบจำลอง ผู้วิจัยเลือกใช้ข้อมูลดังตารางที่ 3.4

ตารางที่ 3.5 ตัวอย่างข้อมูล Data Test

ตัวแปร	ตัวอย่างข้อมูล	คำอธิบายตัวแปร
Id	832141009068765184	รหัสของโพสต์นั้น รูปแบบ Integer
Created-At	2017-02-16 15:14:11	เวลา UTC เมื่อสร้างโพสต์
Text	I'm going back into my ultra all for sleepy time.	ข้อความสถานะของโพสต์รูปแบบ UTF-8

จากตารางที่ 3.4 ได้มีการเพิ่มตัวแปร Created-At เนื่องจากการวิเคราะห์โรคซึมเศร้าต้องอาศัยความถี่ต่อเวลาช่วงระยะเวลาหนึ่งในการวิเคราะห์

3.2.3 การกรองข้อมูล

สำหรับการกรองข้อมูล ผู้วิจัยเลือกกรองข้อความบางส่วนออก โดยใช้ฟังก์ชัน Regular expression คือ รูปแบบการหากลุ่มคำที่กำหนดขึ้นมา เพื่อค้นหาตัวอักษรหรือคำที่ต้องการ โดยกำหนดค่าฟังก์ชันดังต่อไปนี้

1) การกรองข้อความที่เป็นการ Retweet เนื่องจากข้อความเหล่านี้มีเนื้อหาที่ไม่เกี่ยวกับการจำแนกลักษณะอาการ โดยการกำหนดค่าฟังก์ชันเท่ากับ RT(.*) ซึ่งมีตัวอย่างดังตารางที่ 3.5

ตารางที่ 3.6 ตัวอย่างการกรองข้อความ Retweet

ลำดับ	ข้อความก่อนกรอง Retweet	ข้อความหลังกรอง Retweet
1	@Loudwire @CoreyTaylorRock Thanks Corey! You're the best	@Loudwire @CoreyTaylorRock Thanks Corey! You're the best
2	RT @gwatsky: His fleshlight wouldn't let him smash	Null

จากตารางที่ 3.5 หลังจากการกรองข้อความ Retweet แล้ว ข้อความในแถวข้อมูลนั้นจะว่าง ซึ่งเป็นการตัดออกไม่นำมาพยากรณ์

2) การกรองข้อความที่เกี่ยวกับลิ้งเข้าใช้งานเว็บไซต์ซึ่งไม่เกี่ยวข้องกับการใช้พยากรณ์ โดยกำหนดค่าฟังก์ชันเท่ากับ (https?|http)://[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|] ซึ่งมีตัวอย่างดังตารางที่ 3.6

ตารางที่ 3.7 ตัวอย่างการกรองลิ้งเข้าใช้งานเว็บไซต์

ลำดับ	ข้อความก่อนกรองลิ้งเข้าใช้งานเว็บไซต์	ข้อความหลังกรองลิ้งเข้าใช้งานเว็บไซต์
1	@Loudwire @CoreyTaylorRock Thanks Corey! You're the best	@Loudwire @CoreyTaylorRock Thanks Corey! You're the best
2	#NewProfilePic https://t.co/eDzbXf5qAd	#NewProfilePic

จากตารางที่ 3.6 หลังจากการกรองข้อความที่เป็นลิ้งเข้าใช้งานเว็บไซต์ออกแล้ว ข้อความในแถวข้อมูลนั้นจะยังคงอยู่นอกจากลิ้งเข้าใช้งานเว็บไซต์ที่ถูกตัดออก

3) การกรองชื่อบุคคลในโพสต์ ซึ่งเป็นการตัดชื่อของผู้ใช้งานที่ถูกแทกในโพสต์ เนื่องจากชื่อบุคคลที่ถูกแทกในโพสต์ไม่มีผลในการจำแนก โดยการกำหนดค่าฟังก์ชันเท่ากับ (@)[-a-zA-Z0-9+&@#/%?=~_|!:,.;]*[-a-zA-Z0-9+&@#/%=~_|:] ซึ่งมีตัวอย่างดังตารางที่ 3.7

ตารางที่ 3.8 ตัวอย่างการกรองชื่อบุคคลในโพสต์

ลำดับ	ข้อความก่อนกรองชื่อบุคคลในโพสต์	ข้อความหลังกรองชื่อบุคคลในโพสต์
1	@Loudwire @CoreyTaylorRock Thanks Corey! You're the best	Thanks Corey! You're the best
2	Stormzy is da best !!!	Stormzy is da best !!!

จากตารางที่ 3.7 หลังจากการกรองข้อความที่มีชื่อบุคคลในโพสต์ออกแล้ว ข้อความในแถวข้อมูลนั้นจะยังคงอยู่นอกจากชื่อบุคคลในโพสต์ที่ถูกตัดออก

3.2.4 การตัดคำ (Word Segmentation)

การตัดคำเป็นกระบวนการกรองคำศัพท์ออกจากรูปประโยคเพื่อให้สามารถนำไปเข้าสู่กระบวนการทำ Bag of word [13] [28] โดยการใช้งานโปรแกรม Rapidminer มีตัวอย่างดังตารางที่ 3.8 โดยคำที่ตัดออกมาจะนำไปเปรียบเทียบและนับความถี่ในขั้นตอนถัดไป

ตารางที่ 3.9 ตัวอย่างการตัดคำ

Text	Word Segmentation
my time life depression back I’m	my | time | lift | depression | back

จากตารางที่ 3.8 พบว่าคำศัพท์บางคำขาดคุณลักษณะที่น่าสนใจและเป็นคำที่ไม่มีความหมาย เช่นคำ my และ I'm แล้วยังพบอีกว่าในชุดข้อมูลอื่น ๆ จึงได้ทำการลบคำ Stopword จำพวกนี้ออก โดยใช้ Stopword [29] จาก ranks.nl แบบ Default English stopwords list.

3.2.5 การกำหนดน้ำหนักให้คำในเอกสารด้วยวิธีการ Binary Occurrence

หลังจากที่ได้คำในเอกสารที่เหมาะสมในการวิเคราะห์โรคซึมเศร้าแล้ว ต่อมาจะเป็นการนำคำในเอกสารเข้ามาทำการให้น้ำหนักโดยใช้วิธีการ Binary Occurrence มีหลักการว่า ถ้าคำที่เกิดขึ้นในเอกสารจะกำหนดค่าให้เป็น 1 ถ้าไม่เกิดขึ้นในเอกสารจะกำหนดค่าให้เป็น 0 ดังตารางที่ที่ 3.9

ตารางที่ 3.10 ตัวอย่างการให้น้ำหนักด้วยวิธี Binary Occurrence

Message	**Word**						
	depression	***back***	***best***	***honestly***	***funny***	***anxiety***	***...***
my **depression**? i'm **back** on it	1	1	0	0	0	0	...
how to stop **depression**?	1	0	0	0	0	0	...
It's **honestly** **best** for your mental health.	0	0	1	1	0	0	...
That's the **funny** thing about **anxiety**	0	0	0	0	1	1	...

จากตารางที่ 3.9 จะเห็นได้ว่าในข้อความ my depression? i'm back on it เมื่อเทียบกับคำใน Bog of Word โดยใช้หลักการ Binary Occurrence ในการให้น้ำหนัก เมื่อพบเจอคำอย่าง depression และ back ที่ตรงกัน จะให้น้ำหนักคำที่เหมือนกันเป็น 1 แต่ถ้าไม่มีคำที่ตรงกันจะเป็น 0

3.2.5 การคัดเลือกแอตทริบิวต์ด้วย Information Gain

การคัดเลือกแอตทริบิวต์ด้วย Information Gain เป็นการคัดเลือกแอตทริบิวต์ หรือ Feature ที่จะนำมาใช้ในการสร้างแบบจำลอง เนื่องจากแอตทริบิวต์มีจำนวนมากเกินไป ทำให้เวลาในการประมวลผลนั้นล่าช้าและค่าความถูกต้องในการทำนายนั้นอาจจะลดลง ผู้วิจัยจึงนำเอาหลักการ Information Gain มาคำนวณหาน้ำหนักความสำคัญของแอตทริบิวต์แต่ละแอตทริบิวต์ แล้วคัดเอา

เฉพาะแอตทริบิวต์ที่ผ่านค่ามาตรฐาน จึงสามารถลดจำนวนแอตทริบิวต์ลงได้ โดยหลักการของ Information Gain มีวิธีการดังต่อไปนี้

1) คำนวณค่า Entropy ของแต่ละแอตทริบิวต์โดยใช้สมการ 3.4

$$Entopy(c1) = -P(c1)log_2P(c1) \quad (3.1)$$

โดยที่ P(c1) คือ ความน่าจะเป็น (probability) ของ c1

2) คำนวณค่า Information Gain (IG) ของแต่ละแอตทริบิวต์โดยใช้สมการ 3.5

$$IG = Entropy(initial) - [P(c1) * Entropy(c1) + P(c2) * Entropy(c2)] \quad (3.2)$$

จากสมการ 3.4 และ 3.5 สามารถนำมาคำนวณนำหนักของแอตทริบิวต์ ดังตารางที่ 3.10

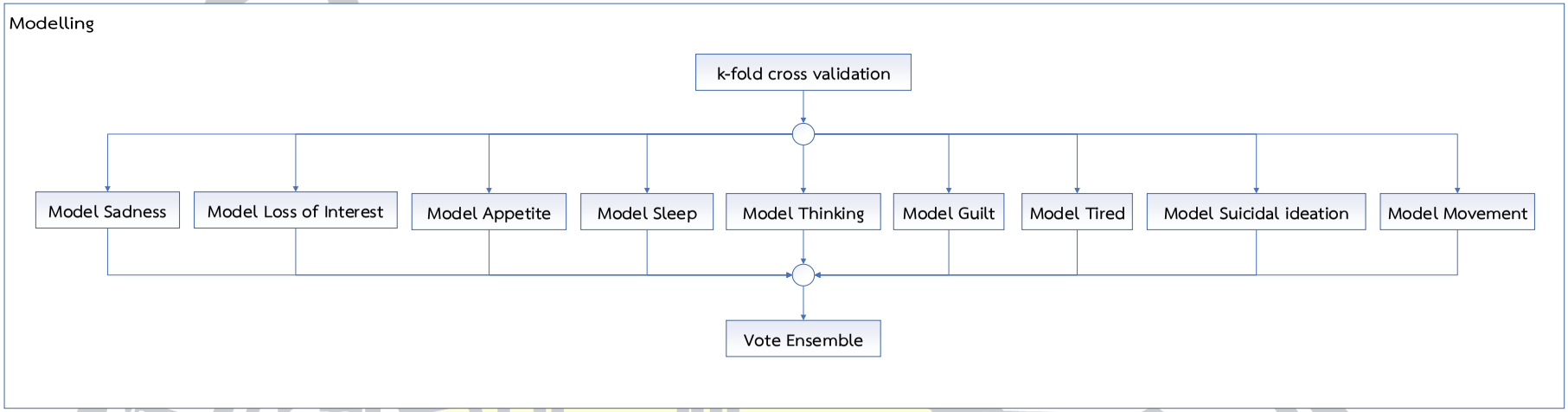
ตารางที่ 3.11 ตัวอย่างการให้น้ำหนักแอตทริบิวต์ด้วยวิธี Information Gain

Attribute	Weight
depression	1
back	0.384
best	0.382
honestly	0.517
funny	0.450
anxiety	1
...	...

จากตารางที่ 3.10 จะเห็นได้ว่าน้ำหนักของแต่ละแอตทริบิวต์มีค่าแตกต่างกัน โดยการเลือกแอตทริบิวต์จากการคำนวณด้วย Information Gain มีหลักการว่าให้เลือกค่าที่เข้าใกล้ 1 มากที่สุด เนื่องจากเหมาะแก่การนำแอตทริบิวต์นั้นไปเป็นแอตทริบิวต์ในการสร้างแบบจำลองให้มีประสิทธิภาพสูงสุดทั้งด้านเวลาและความถูกต้อง

3.3 การสร้างแบบจำลอง

ในงานวิจัยนี้ได้นำกระบวนการที่ใช้ในการทำเหมืองความคิดเห็นที่มีประสิทธิภาพโดยผ่านการยอมรับในงานวิจัยด้านการจำแนกข้อความโรคซึมเศร้า เพื่อสร้างแบบจำลองที่มีประสิทธิภาพในการจำแนกข้อความซึมเศร้าสูงสุดดังรูปที่ 3.3



รูปที่ 3.3 การสร้างแบบจำลอง

โดยในการสร้างแบบจำลองการจำแนกข้อความที่เข้าข่ายเป็นโรคซึมเศร้าและไม่เข้าข่ายโรคซึมเศร้ามีขั้นตอนดังต่อไปนี้

3.3.1 การแบ่งข้อมูลโดยใช้วิธีการ K – Fold Cross Validation

การแบ่งข้อมูลที่ใช้ในการจำแนกข้อมูลที่เข้าข่ายเป็นโรคซึมเศร้าได้ใช้งาน K – Fold Cross Validation โดยการกำหนดให้ K = 10 โดยจะแบ่งออกเป็น Training Set 90% และ Testing Set 10% ในการสร้างโมเดล 10 รอบ จนกระทั่งใช้ข้อมูลทุกตัวที่มีความหลากหลายในการเรียนรู้ ซึ่งวิธีการนี้สามารถช่วยให้การีนตีค่าความถูกต้องที่สูงได้

3.3.2 การสร้างแบบจำลองการทำเหมือง

ในการสร้าง Model การทำเหมืองใช้สำหรับในกรณีข้อความที่เข้ามาไม่ใช่สัญลักษณะ Emoticon เพียงอย่างเดียว ผู้วิจัยจึงเลือกใช้อัลกอริธึม Bayes ในการสร้าง Model เนื่องจากอัลกอริธึม Bayes ได้ผลลัพธ์เป็นค่าความน่าจะเป็น (Probability) โดยจะใช้ค่าความน่าจะเป็นที่ได้เป็นตัวตัดสินข้อความนั้น ๆ ผู้วิจัยจึงเลือกใช้การสร้างแบบจำลองตามลักษณะอารมณ์ที่เกี่ยวข้องกับโรคซึมเศร้า ดังต่อไปนี้

1) แบบจำลองจำแนกอารมณ์ซึมเศร้า

2) แบบจำลองจำแนกการขาดความสนใจลดลง

3) แบบจำลองจำแนกน้ำหนักผิดปกติ

4) แบบจำลองจำแนกการนอนผิดปกติ

5) แบบจำลองจำแนกร่างกายอ่อนเพลีย

6) แบบจำลองจำแนกการรู้สึกตนเองไร้ค่า

7) แบบจำลองจำแนกสมาธิสั้น

8) แบบจำลองจำแนกการเคลื่อนไหลช้า

9) แบบจำลองจำแนกกับการคิดฆ่าตัวตาย

โดยแบบจำลอง 1 แบบจำลองมีคำตอบเป็น เข้าข่ายเป็นอาการ (Yes) และ ไม่เข้าข่ายเป็นอาการ (No) มีตัวอย่างการสร้างแบบจำลองดังตารางที่ 3.11

ตารางที่ 3.12 ตัวอย่างการสร้างแบบจำลองด้วยอัลกอริธึม Bayes

Class Sadness	Word						
	cry	*want*	*fat*	*sleep*	*read*	*feel*	*lazy*
Yes	1	0	0	0	0	1	0
Yes	0	1	0	0	0	1	1
No	0	0	1	0	0	0	1
No	0	0	0	1	0	1	0
Yes	1	1	0	0	1	0	0
No	0	0	0	0	0	1	0
Unknow	1	0	0	1	0	1	0

จากตารางที่ 3.11 มีวิธีการหาความน่าจะเป็นตามอัลกอริธึม Bayes โดยใช้ข้อมูล Unknow Class สำหรับทำใน Sadness Class มีวิธีต่อไปนี้

1) การทำนายว่าเข้าข่ายเป็นข้อความซึมเศร้า

P(Yes | Sadness) = P(Yes)*P(cry | Yes)* P(want | No)* P(fat | No)* P(sleep | Yes)* P(read | No)* P(feel | Yes)* P(lazy | No)

โดยที่

P(sadness) = 3/6

P(cry | Yes) = 2/3

P(want | No) = 1/3

P(fat | No) = 3/3

P(sleep | Yes) = 0/3

P(read | No) = 2/3

P(feel | Yes) = 2/3

P(lazy | No) = 2/3

P(Yes | Sadness) = 3/6 * 2/3 * 1/3 * 3/3 * 0/3 * 2/3 * 2/3 * 2/3

P(Yes | Sadness) = 0/13,122

กล่าวได้ว่าข้อมูล Unknow มีความน่าจะเป็นที่จะเข้าข่ายเป็นข้อความซึมเศร้า คือ 0

2) การทำนายว่าไม่เข้าข่ายเป็นข้อความซึมเศร้า

P(No | Sadness) = P(Yes)*P(cry | Yes)* P(want | No)* P(fat | No)* P(sleep | Yes)* P(read | No)* P(feel | Yes)* P(lazy | No)

โดยที่

P(sadness) = 3/6

P(cry | Yes) = 2/3

P(want | No) = 1/3

P(fat | No) = 3/3

P(sleep | Yes) = 0/3

P(read | No) = 2/3

P(feel | Yes) = 2/3

P(lazy | No) = 2/3

P(Yes | Sadness) = 3/6 * 3/3 * 2/3 * 1/3 * 2/3 * 3/3 * 2/3 * 1/3

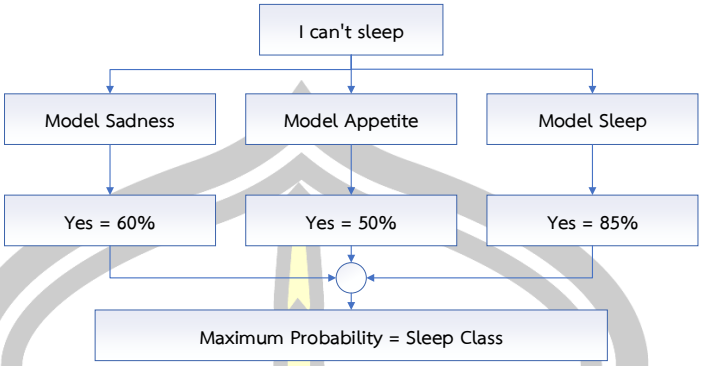
P(Yes | Sadness) = 216/13,122 = 0.016

กล่าวได้ว่าข้อมูล Unknow มีความน่าจะเป็นที่ไม่เข้าข่ายเป็นข้อความซึมเศร้า คือ 0.016

3) สรุปได้ว่าข้อมูล Unknow ไม่เข้าข่ายเป็นข้อความซึมเศร้าเพราะความน่าจะเป็นที่ไม่เข้าข่ายซึมเศร้ามีมากกว่าเข้าข่ายซึมเศร้า

3.3.3 การเลือกผลลัพธ์โดยการ Vote ensemble

การเลือกผลลัพธ์โดยการใช้งาน Vote ensemble คือ การเลือกค่าความน่าจะเป็นที่มีค่าสูงที่สุดโดยที่ข้อมูล 1 Instant ผ่านแบบจำลองทั้ง 9 แบบจำลอง โดย มีแบบจำลองจะได้ค่าความน่าจะเป็นของ Class คำตอบคือ 1 ค่า แล้วนำค่าความน่าจะเป็นของ 9 แบบจำลองมาทำการโหวตเลือกค่าความน่าจะเป็นสูงสุด เพื่อเป็นคำตอบของการทำนายครั้งนั้น ดังตัวอย่างเบื้องต้นกระบวนการเลือกค่า Maximum Probability ดังรูปที่ 3.4

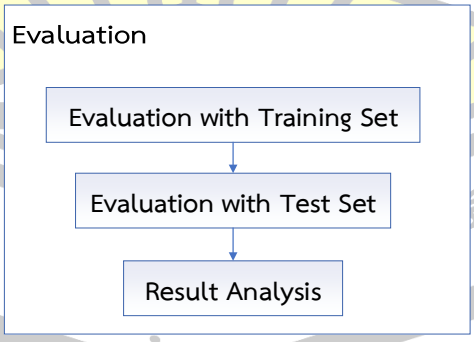


รูปที่ 3.4 กระบวนการเลือกค่า Maximum Probability

จากรูปที่ 3.4 มีการนำข้อความ I can’s sleep เข้าไปทำนาย 3 Model ผลปรากฏว่า Model Sadness ทำนายความน่าจะเป็นข้อความซึมเศร้า 60% และ Model Appetite ทำนายความน่าจะเป็นข้อความน้ำหนักผิดปกติ 50% แต่ Model Sleep ทำนายความน่าจะเป็นการนอนหลับไม่ปกติ 60% ซึ่งสรุปได้ว่าข้อความ I can’s sleep มีความเป็นไปได้ที่จะเป็นข้อความที่เข้าข่ายอาการนอนไม่หลับในการทำนายครั้งนี้ ในการทดลองครั้งนี้ผู้วิจัยได้กำหนดค่า Boundary ออกเป็นช่วง 0-90 เพื่อหาเพดานที่เหมาะสมในการคัดเลือกข้อความ

3.4 การวัดประสิทธิภาพของแบบจำลอง

การวัดประสิทธิภาพของแบบจำลองในงานวิจัยนี้ใช้ค่าสถิติที่ใช้ในการวัดประสิทธิภาพของการการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์โดยใช้ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และ F-1 จากข้อมูลผู้ที่เป็นโรคซึมเศร้าและไม่เป็นโรคซึมเศร้า โดยการวัดประสิทธิภาพมี 2 ขั้นตอน ดังรูปที่ 3.5



รูปที่ 3.5 การวัดประสิทธิภาพของแบบจำลอง

3.4.1 การวัดประสิทธิภาพข้อมูลสำหรับการเรียนรู้แบบจำลอง

การวัดประสิทธิภาพข้อมูลสำหรับทดสอบแบบจำลอง เป็นการวัดประสิทธิภาพของชุดข้อมูลที่นำเข้าไปสร้างแบบจำลอง โดยเป็นข้อมูล 10% จากการแบ่งข้อมูลในขั้นตอน 10 – Fold Cross Validation ซึ่งข้อมูล 10% ที่แบ่งไว้จะได้เข้าทดสอบแบบจำลองแต่ละแบบจำลอง เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองนั้น ๆ จะใช้การวัดประสิทธิภาพด้วย Confusion Matrix โดยมีตัวอย่างการวัดประสิทธิภาพแบบจำลอง 1 แบบจำลอง ดังตารางที่ 3.12

ตารางที่ 3.13 ตัวอย่างการวัดประสิทธิภาพ Model Sadness ด้วย Confusion Matrix

		ค่าทำนาย	
ค่าความจริง	**Classes**	**Yes**	**No**
	Yes	1500	500
	No	500	1500

จากตารางที่ 3.12 พบว่าได้ค่า Accuracy = 75% Precision = 75% Recall = 75% และ F-1 = 75% ซึ่งการวัดประสิทธิภาพเช่นนี้ จะวัดทั้ง 9 แบบจำลองเพื่อเป็นแนวทางในการปรับพารามิเตอร์ของอัลกอริธึมต่อไป

3.4.2 การวัดประสิทธิภาพข้อมูลสำหรับทดสอบแบบจำลอง

การวัดประสิทธิภาพข้อมูลสำหรับทดสอบแบบจำลอง เป็นการวัดประสิทธิภาพของการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์ โดยใช้ชุดข้อมูลสำหรับทดสอบที่ประกอบด้วยชุดข้อมูลการโพสต์ของผู้ที่เป็นโรคซึมเศร้าจำนวน 15 คน และไม่เป็นโรคซึมเศร้าจำนวน 15 คน ที่มีระยะการโพสต์มากกว่า 2 สัปดาห์ขึ้นไป

3.4.2.1 การวิเคราะห์โรคซึมเศร้า

หลังจากแบบจำลองสามารถจำแนกข้อมูลที่อยู่ในลักษณะอาการ 9 อาการได้อย่างแม่นยำแล้ว ในการวิเคราะห์ผู้ที่เข้าข่ายเป็นโรคซึมเศร้านั้น สามารถทำได้โดยการนับความถี่ของอาการแต่ละอาการในระยะเวลา 2 สัปดาห์ หรือ 14 วัน โดยจะต้องขยับไปทุก ๆ 1 วัน แล้วคำนวณความถี่ใหม่จนกว่าจะครบข้อมูลที่ทำนายของบุคคลนั้น [23] มีตัวอย่างดังตารางที่ 3.13

ตารางที่ 3.14 ตัวอย่างการนับความถี่จากการทำนายของแบบจำลอง

อาการ	วันที่ 1 – 14 หรือ 2 สัปดาห์													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
อารมณ์ซึมเศร้า	มี	-	มี	-	-	มี	-	-	มี	-	-	มี	-	-
ความสนใจลดลง	-	มี	-	-	-	มี	-	-	มี	-	-	-	มี	มี
น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต	-	-	-	มี	-	-	มี	-	-	-	-	-	มี	-
นอนไม่หลับหรือนอนหลับมากกว่าปกติ	-	มี	-	-	-	มี	-	-	มี	-	-	มี	-	-
ร่างกายอ่อนเพลีย	-	มี	-	มี	-	-	-	มี	-	มี	-	-	-	มี
รู้สึกตนเองไร้ค่า	มี	-	-	มี	-	-	มี	-	-	-	มี	-	-	-
สมาธิสั้น	-	มี	-	-	-	-	มี	-	-	-	มี	-	-	มี
เคลื่อนไหวช้า	-	-	-	-	-	-	-	มี	-	มี	-	มี	-	-
อยากฆ่าตัวตาย	-	-	-	-	-	-	-	-	-	มี	-	-	มี	-

จากตารางที่ 3.13 สามารถใช้ร่วมกับการวินิจฉัยโรคซึมเศร้าของจิตแพทย์ในการวิเคราะผู้ที่เข้าข่ายเป็นโรคซึมเศร้านั้น มีหลักเกณฑ์ว่า บุคคลนั้นต้องมีการเกิดของลักษณะอาการอย่างน้อย 5 อาการ โดยเกิดขึ้นติดต่อกันเป็นระยะเวลา 2 สัปดาห์ขึ้นไป หลังจากทราบความถี่ของอาการเหล่านั้น จิตแพทย์จะนำความถี่ของอาการในช่วงเวลา 2 สัปดาห์ ไปให้คะแนนความถี่โดยมีหลักเกณฑ์ว่า ไม่มี คือ ในระยะเวลา 2 สัปดาห์ ไม่มีอาการนั้นเลย, บางวัน คือ มีอาการนั้นใน 2-4 วันในระยะเวลา 1 สัปดาห์, บ่อยๆ คือ มีอาการนั้นใน 6-8 วันในระยะเวลา 2 สัปดาห์ และ ทุกวัน คือ มีอาการนั้นใน 10-14 วัน ในระยะเวลา 2 สัปดาห์แล้วนำไปหาผลรวมของคะแนนในแบบประเมินของการวินิจฉัยโรคซึมเศร้า ดังตารางที่ 3.14

ตารางที่ 3.15 ตัวอย่างแบบประเมินโรคซึมเศร้า 9Q

ข้อ	ลักษณะอาการใน 2 สัปดาห์	ไม่มี	บางวัน	บ่อยๆ	ทุกวัน
1	อารมณ์ซึมเศร้า	0	1	2	**3**
2	ความสนใจลดลง	0	1	2	**3**
3	น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต	0	1	**2**	3
4	นอนไม่หลับหรือนอนหลับมากกว่าปกติ	0	1	**2**	3
5	ร่างกายอ่อนเพลีย	0	1	2	**3**
6	รู้สึกตนเองไร้ค่า	0	1	**2**	3
7	สมาธิสั้น	0	1	**2**	3
8	เคลื่อนไหวช้า	0	1	**2**	3
9	อยากฆ่าตัวตาย	0	1	**2**	3
	รวมคะแนนทั้งหมด	**รวม**		**21**	

จากตารางที่ 3.14 ในการสรุปผลสามารถนำคะแนนจากแบบประเมินไปประเมินได้ว่า ถ้าน้อยกว่า 7 แสดงว่าปกติ, ถ้า 8 – 12 อาการน้อย, 13-18 อาการปานกลาง, มากกว่า 19 อาการรุนแรงควรปรึกษาแพทย์โดยด่วน ดังตัวอย่าง ตารางที่ 3.10 ผู้ป่วยมีอาการรุนแรงเนื่องจากคะแนนรวม 21 คะแนน

3.4.2.2 การวัดประสิทธิภาพข้อมูลสำหรับทดสอบแบบจำลอง

การวัดประสิทธิภาพด้วย Confusion Matrix เป็นการประเมินผลประสิทธิภาพการทำนายของแบบจำลองโดยในการวิจัยครั้งนี้ผลการทดลองที่เกิดขึ้นจากการทำนาย Data Test จะนำมาวัดประสิทธิภาพด้วย Confusion Matrix โดยมี 30 คน แบ่งเป็นคนที่เป็นโรคซึมเศร้าจำนวน 15 คน กำหนดเป็น Class Yes และไม่เป็นโรคซึมเศร้าจำนวน 15 คน กำหนดเป็น Class No มีตังอย่างดังตารางที่ 3.15

ตารางที่ 3.16 ตัวอย่างการวัดประสิทธิภาพของข้อมูลทดสอบด้วย Confusion Matrix

		ค่าทำนาย	
ค่าความจริง	**Classes**	**Yes**	**No**
	Yes	7	8
	No	5	10

จากตารางที่ 3.15 พบว่าได้ค่า Accuracy = 56%, Precision = 46% และ Recall = 58%

3.5 การใช้งานทรัพยากรในการพัฒนาระบบ

การใช้งานทรัพยากรในการพัฒนาระบบในการสร้างแบบจำลองเพื่อจำแนกข้อความที่บ่งบอกถึงโรคซึมเศร้า ใช้ทรัพยากรในการพัฒนาระบบดังนี้

3.5.1 ด้าน Hardware ของคอมพิวเตอร์ มีรายละเอียดดังนี้

1) Processor Intel Core i7 3770 @3.90GHz

2) RAM DDR3 Bus 1600 24GB

3) SSD 120GB

3.5.2 ด้าน Software ของคอมพิวเตอร์ มีรายละเอียดดังนี้

Windows 10 Pro 64Bit

3.5.3 ด้าน Software ในการสร้างแบบจำลอง มีรายละเอียดดังนี้

RapidMiner Version 9.0

บทที่ 4

ผลการวิจัยและการอภิปราย

ในงานวิจัยนี้มีจุดมุ่งหมายเพื่อพัฒนาการทำเหมืองความคิดเห็นโดยการนำความคิดเห็นจากการโพสต์ของผู้ใช้งาน Twitter มาทำการจำแนกหาผู้ที่มีพฤติกรรมเข้าข่ายเป็นโรคซึมเศร้าด้วยการทำเหมืองความคิดเห็นโดยใช้อัลกอริธึม Bayes ในการสร้างแบบจำลอง ซึ่งผลการทดลองมีดังนี้ ผลของการทำความเข้าใจปัญหา ผลการเตรียมข้อมูล และผลการวัดประสิทธิภาพการทดลอง

4.1 ผลของการทำความเข้าใจปัญหา

เนื่องจากปัจจุบันมีผู้ป่วยที่เป็นโรคซึมเศร้าใช้งานโซเชียลมีเดียอย่างแพร่หลาย ทั้งคนที่เป็นอยู่และคนที่เป็นอยู่แล้วแต่ไม่รู้ตัว ผู้วิจัยจึงได้นำปัญหาเหล่านี้เข้าปรึกษาจิตแพทย์ทำให้ทราบว่าการวิเคราะห์ผู้ป่วยที่จะเข้าข่ายเป็นโรคซึมเศร้าจะถูกวินิจฉัยเบื้อต้นจากการทำแบบสอบถาม Q9 ผู้วิจัยจึงนำข้อมูลพฤติกรรมการโพสต์บน Twitter เข้ามาพยากรณ์เพื่อหาผู้ที่มีพฤติกรรมเข้าข่ายเป็นโรคซึมเศร้าโดยหวังว่าจะช่วยให้จิตแพทย์ตัดสินใจได้สะดวกยิ่งขึ้น

4.2 ผลการเตรียมข้อมูล

4.2.1 ผลของการเตรียมข้อมูลสำหรับการเรียนรู้แบบจำลอง

สำหรับการเก็บข้อมูลสำหรับการเรียนรู้แบบจำลองจาก Twitter ผู้วิจัยเลือกใช้งาน RapidMiner ดึงข้อมูลจาก Twitter โดยมีจำนวนข้อมูลสำหรับการเรียนรู้แบบจำลองดังตารางที่ 4.1

ตารางที่ 4.1 จำนวนข้อมูลสำหรับการเรียนรู้แบบจำลอง

กลุ่มของอาการ	**จำนวนข้อความตามอาการ**	**จำนวนข้อความที่ไม่เป็นไปตามอาการ**
1) ข้อความเกี่ยวกับอารมณ์ซึมเศร้า	3,000 ข้อความ	3,000 ข้อความ
2) ข้อความเกี่ยวกับการขาดความสนใจลดลง	3,000 ข้อความ	3,000 ข้อความ
3) ข้อความเกี่ยวกับน้ำหนักผิดปกติ	3,000 ข้อความ	3,000 ข้อความ
4) ข้อความเกี่ยวกับการนอนผิดปกติ	3,000 ข้อความ	3,000 ข้อความ
5) ข้อความเกี่ยวกับร่างกายอ่อนเพลีย	3,000 ข้อความ	3,000 ข้อความ
6) ข้อความเกี่ยวกับการรู้สึกตนเองไร้ค่า	3,000 ข้อความ	3,000 ข้อความ
7) ข้อความเกี่ยวกับสมาธิสั้น	3,000 ข้อความ	3,000 ข้อความ
8) ข้อความเกี่ยวกับการเคลื่อนไหลช้า	3,000 ข้อความ	3,000 ข้อความ
9) ข้อความเกี่ยวกับการคิดฆ่าตัวตาย	3,000 ข้อความ	3,000 ข้อความ
ทั้งหมด	54,000 ข้อความ	

จากตารางที่ 4.1 ประเภทข้อความเป็นข้อความที่เกี่ยวกับลักษณะอาการที่ก่อให้เกิดโรคซึมเศร้าจำนวน 3,000 ข้อความ โดยรวมกับข้อความทั่วที่ไม่เกี่ยวข้องกับลักษณะอาการที่ก่อให้เกิดโรคซึมเศร้าจำนวน 3,000 ข้อความ ทั้งหมดรวมกันเป็นจำนวน 6,000 ข้อความ

4.2.2 ผลของการเตรียมข้อมูลสำหรับการทดสอบแบบจำลอง

สำหรับการเก็บข้อมูลสำหรับการทดสอบแบบจำลองจาก Twitter ผู้วิจัยเลือกใช้งาน Rapidminer ดึงข้อมูลโดยข้อมูลจาก Twitter สำหรับการทดสอบแบบจำลองที่สร้างเสร็จแล้วผู้วิจัย

เลือกใช้งานข้อมูลของบุคคลที่เป็นโรคซึมเศร้าดังตารางที่ 4.2 และบุคคลที่ไม่เป็นโรคซึมเศร้าดังตารางที่ 4.3

ตารางที่ 4.2 จำนวนข้อมูลสำหรับการทดสอบแบบจำลองของบุคคลที่เป็นโรคซึมเศร้า

บุคคลที่เป็นโรคซึมเศร้า	จำนวนข้อความ
1	1,021
2	3,199
3	1285
4	3,125
5	310
6	3,223
7	1,895
8	3,225
9	3,212
10	3,088
11	3,207
12	155
13	1,046
14	3,235
15	3,209

จากตารางที่ 4.2 เป็นจำนวนข้อความของคนดังต่างประเทศทั้ง 15 คน ที่เคยเป็นโรคซึมเศร้า โดยมีระยะเวลาการโพสต์มากกว่า 2 สัปดาห์ขึ้นไป

ตารางที่ 4.3 จำนวนข้อมูลสำหรับการทดสอบแบบจำลองของบุคคลที่ไม่เป็นโรคซึมเศร้า

บุคคลที่ไม่เป็นโรคซึมเศร้า	จำนวนข้อความ
1	459
2	3,190
3	3,228
4	1,697
5	848
6	1,657
7	1,265
8	454
9	2,090
10	419
11	547
12	3,180
13	1,287
14	872
15	1,305

จากตารางที่ 4.3 เป็นจำนวนข้อความของคนดังต่างประเทศทั้ง 15 คน ที่ไม่เคยเป็นโรคซึมเศร้า โดยมีระยะเวลาการโพสต์มากกว่า 2 สัปดาห์ขึ้นไป

4.2.3 ผลของการเตรียมข้อมูล

ในการเตรียมข้อมูล ผู้วิจัยเลือกกรองข้อความบางส่วนออกโดยใช้ฟังก์ชัน Regular expression โดยกรองข้อความที่เป็นการ Retweet การกรองข้อความที่เกี่ยวกับลิ้งเข้าใช้งานเว็บไซต์ และการกรองชื่อบุคคลในโพสต์ จำนวนของการกรองข้อมูลสำหรับการเรียนรู้แบบจำลองดังตารางที่ 4.4 จำนวนการกรองข้อมูลสำหรับการทดสอบแบบจำลองของบุคคลที่เป็นโรคซึมเศร้าดังตารางที่ 4.5 และ จำนวนการกรองข้อมูลสำหรับการทดสอบแบบจำลองของบุคคลที่ไม่เป็นโรคซึมเศร้าดังตารางที่ 4.6

ตารางที่ 4.4 จำนวนข้อมูลสำหรับการเรียนรู้แบบจำลองที่ถูกกรองข้อความ Retweet

ประเภทข้อความ	**Retweet**	**HTTP**	**Tag**
อารมณ์ซึมเศร้า	3,023	476	754
ความสนใจลดลง	2,242	329	629
น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต	2,632	534	824
นอนไม่หลับหรือนอนหลับมากกว่าปกติ	2,663	492	928
ร่างกายอ่อนเพลีย	3,217	482	1,034
รู้สึกตนเองไร้ค่า	2,828	489	1,131
สมาธิสั้น	2,544	557	880
เคลื่อนไหวช้า	2,494	637	1,038
คิดฆ่าตัวตาย	3,169	502	850

จากตารางที่ 4.4 เป็นจำนวนข้อความสำหรับการใช้เรียนรู้แบบจำลอง โดยข้อความเหล่านี้ถูกกรองข้อความที่เป็นการ Retweet ออก ถูกกรอก HTTP ออก และถูกกรอก Tag ที่เป็นชื่อบุคคลออก

ตารางที่ 4.5 จำนวนการกรองข้อมูลสำหรับการทดสอบแบบจำลองของบุคคลที่เป็นโรคซึมเศร้า

บุคคลที่เป็นโรคซึมเศร้า	Retweet	HTTP	Tag
1	404	453	279
2	427	2,428	1,078
3	13	861	78
4	562	208	1,804
5	39	183	206
6	1,642	644	670
7	1,032	491	356
8	758	1,695	573
9	496	1,099	1,712
10	721	1481	941
11	644	1,922	1,079
12	41	49	17
13	644	175	104
14	731	1,777	1,315
15	216	541	1,682

จากตารางที่ 4.5 เป็นจำนวนข้อความสำหรับทดสอบแบบจำลองที่เป็นข้อความของผู้ที่เป็นโรคซึมเศร้า โดยข้อความเหล่านี้ถูกกรองข้อความที่เป็นการ Retweet ออก ถูกกรอก HTTP ออก และถูกกรอก Tag ที่เป็นชื่อบุคคลออก

ตารางที่ 4.6 จำนวนการกรองข้อมูลสำหรับการทดสอบแบบจำลองของบุคคลที่ไม่เป็นโรคซึมเศร้า

บุคคลที่เป็นไม่เป็นโรคซึมเศร้า	Retweet	HTTP	Tag
1	79	268	192
2	426	1,186	693
3	187	2,255	1,395
4	372	109	385
5	164	456	276
6	486	406	547
7	117	865	479
8	88	276	198
9	434	676	704
10	103	221	186
11	70	316	168
12	305	2,006	872
13	114	1,128	355
14	24	605	93
15	255	474	421

จากตารางที่ 4.6 เป็นจำนวนข้อความสำหรับทดสอบแบบจำลองที่เป็นข้อความของผู้ที่ไม่เคยเป็นโรคซึมเศร้า โดยข้อความเหล่านี้ถูกกรองข้อความที่เป็นการ Retweet ออก ถูกกรอก HTTP ออก และถูกกรอก Tag ที่เป็นชื่อบุคคลออก

4.2.4 ผลของการตัดคำ

สำหรับการตัดคำผู้วิจัยใช้งานฟังก์ชันการตัดคำใน RapidMiner ในการตัดคำเพื่อสร้างถุงคำ โดยเก็บไว้ใน Bag of word เพื่อเข้าขั้นตอนการให้น้ำหนัก ซึ่งได้คำในถุงคำทั้งหมดเมื่อนำมานับความถี่ของคำจะมีผลลัพธ์ ดังตารางที่ 4.7

ตารางที่ 4.7 จำนวนคำใน Bag of word

แบบจำลอง	**Bag of word**
อารมณ์ซึมเศร้า	8,124
ความสนใจลดลง	6,596
น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต	9,343
นอนไม่หลับหรือนอนหลับมากกว่าปกติ	8,940
ร่างกายอ่อนเพลีย	8,994
รู้สึกตนเองไร้ค่า	9,091
สมาธิสั้น	8,671
เคลื่อนไหวช้า	9,844
คิดฆ่าตัวตาย	8,936

จากตารางที่ 4.7 เป็นจำนวนคำใน Bag of word ของทั้ง 9 ลักษณะอาการที่บ่งบอกถึงการเกิดโรคซึมเศร้า สำหรับความถี่ของคำใน Bag of word ของแต่ละชุดข้อมูลก่อนนำเข้าคัดเลือกคุณลักษณะมีความถี่สูงสุด 10 คุณลักษณะดังตารางที่ 4.8

ตารางที่ 4.8 Top 10 อันดับคำใน Bag of word

Top 10-word frequency									
No.	**Depressive**	**Loss of interest**	**Appetite**	**Sleep**	**Slowed thinking**	**Guilt**	**Tired**	**Unexplained**	**Suicidal**
1	Sadness	Interest	Appetite	Sleep	Thinking	Guilt	Tired	Loafing	Suicidal
2	Depressive	Loss	Bra	Lethargy	People	Dishonor	Bored	Lazy	Destructive
3	Episode	Things	Hunger	Tired	Time	Step	Fatigued	Movement	Thoughts
4	Hope	Feeling	Eat	Secs	Stop	Dems	Secs	Peace	Dangerous
5	Heart	Low	Food	Time	Day	Fbi	Im	Nature	People
6	Murder	Guilt	Things	Day	Good	Stop	Time	Green	Feel
7	Abortion	Esteem	Cana	Night	Love	Evidence	Hearing	Symbolizes	Depression
8	Forever	Hopeless	Lost	Good	Thing	Flake	People	Phlegmatic	Depressed
9	Tear	Tiredness	People	Sleeping	Feel	Republicans	Pm	Agree	Life
10	Wipe	Lose	Day	People	Bra	Investigation	Feel	Investigation	Make

จากตารางที่ 4.8 เป็น อันดับ 10 คำแรกที่มีความถี่สูงสุดใน Bag of word ที่เข้าข่ายเป็นอาการแต่ละอาการของทั้ง 9 ชุดข้อมูลสำหรับการสร้างแบบจำลอง

4.2.5 ผลของการคัดเลือกคุณลักษณะ

สำหรับการคัดเลือกคุณลักษณะโดยการใช้งาน Information Gain ผลของการคัดเลือกคุณลักษณะ 10 อันดับแรกมีผลลัพธ์ดังตารางที่ 4.9

ตารางที่ 4.9 ผลการดัดเลือกแอตทริบิวต์

	Model								
No.	**Depressive**	**Loss of interest**	**Appetite**	**Sleep**	**Slowed thinking**	**Guilt**	**Tired**	**Unexplained**	**Suicidal**
1	happy	interest	happy	happy	happy	happy	happy	happy	happy
2	depressive	loss	appetite	sleep	thinking	guilt	tired	loafing	suicidal
3	sadness	guilt	bra	lethargy	birthday	birthday	bored	lazy	birthday
4	birthday	esteem	birthday	birthday	sunday	dishonor	fatigued	birthday	destructive
5	episode	hopeless	hunger	tired	bi	dems	birthday	movement	thoughts
6	abortion	tiredness	cana	secs	love	flake	secs	symbolizes	dangerous
7	murder	feeling	eat	sunday	euphoria	republicans	sunday	green	sunday
8	wipe	low	sunday	sleeping	inlove	despicable	hearing	phlegmatic	depression
9	sunday	things	love	love	singularity	caving	love	nature	depressed
10	tear	happy	food	bi	ake	flips	upset	peace	love

จากตารางที่ 4.9 เป็นการคัดเลือกคุณลักษณะโดยใช้งาน Information Gain โดยแสดง 10 อันดับแรกที่มี Gain สูงสุดของคุณลักษณะทั้งหมด

4.3 ผลการวัดประสิทธิภาพแบบจำลอง

ผลของการสร้างแบบจำลองโดยใช้งานอัลกอริธึม Bayes ใช้การวัดประสิทธิภาพด้วย Confusion Matrix เพื่อหาค่า Accuracy, Precision, Recall และ F-1 เข้ามาประเมินผลของการวัดประสิทธิภาพของแบบจำลอง โดยมีผลการทดสอบดังนี้

4.3.1 ผลการวัดประสิทธิภาพแบบจำลองโดยข้อมูลสำหรับการเรียนรู้แบบจำลอง

ในการสร้างแบบจำลองผู้วิจัยเลือกใช้งาน Information Gain ในการคัดเลือกคุณลักษณะโดยกำหนด Top-K คือ 2,000 4,000 6,000 และเลือกคุณลักษณะทั้งหมด มีผลการทดลองดังต่อไปนี้

1) ผลการวัดประสิทธิภาพของแบบจำลองที่ใช้งาน 2,000 คุณลักษณะ

จากการสร้างแบบจำลองโดยการกำหนด Top k = 2,000 ซึ่งเป็นการเลือกคุณลักษณะที่มีค่า Gain สูงสุด 2,000 อันดับแรกเข้ามาสร้างแบบจำลอง ผลลัพธ์ดังตารางที่ 4.10

ตารางที่ 4.10 ผลการวัดประสิทธิภาพแบบจำลองที่เรียนรู้ด้วยด้วย 2,000 คุณลักษณะ

	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
1. Depressive	96.03%	95.31%	96.78%	96.83%	95.25%	96.00%
2. Loss of interest	98.92%	99.39%	98.45%	98.43%	99.40%	98.92%
3. Appetite	96.65%	96.70%	96.60%	96.60%	96.70%	96.65%
4. Sleep	94.45%	92.32%	96.81%	96.97%	91.93%	94.31%
5. Slowed thinking	95.22%	94.99%	95.44%	95.47%	94.97%	95.20%
6. Guilt	96.42%	95.88%	96.96%	97.00%	95.83%	96.39%
7. Tired	94.88%	92.52%	97.53%	97.67%	92.10%	94.73%
8. Unexplained	93.87%	92.07%	95.82%	96.00%	91.73%	93.72%
9. Suicidal	96.23%	95.57%	96.92%	96.97%	95.50%	96.20%
Average	95.85%	94.97%	96.81%	96.88%	94.82%	95.79%

จากตารางที่ 4.10 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการทดลองทั้ง 9 แบบจำลอง พบว่าแบบจำลองการจำแนกอาการความสนใจลดลงมีความถูกต้องในการจำแนกมากที่สุด

2) ผลการวัดประสิทธิภาพของแบบจำลองที่ใช้งาน 4,000 คุณลักษณะ

จากการสร้างแบบจำลองโดยการกำหนด Top k = 4,000 ซึ่งเป็นการเลือกคุณลักษณะที่มีค่า Gain สูงสุด 4,000 อันดับแรกเข้ามาสร้างแบบจำลอง ผลลัพธ์ดังตารางที่ 4.11

ตารางที่ 4.11 ผลการวัดประสิทธิภาพแบบจำลองที่เรียนรู้ด้วยด้วย 4,000 คุณลักษณะ

	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
1. Depressive	95.28%	94.53%	96.16%	94.43%	94.43%	95.24%
2. Loss of interest	98.88%	99.19%	98.58%	98.57%	99.20%	98.89%
3. Appetite	96.22%	95.65%	96.79%	96.83%	95.60%	96.19%
4. Sleep	94.88%	93.28%	96.61%	96.73%	93.03%	94.78%
5. Slowed thinking	94.45%	94.00%	94.91%	94.97%	93.93%	94.42%
6. Guilt	96.60%	95.99%	97.23%	97.27%	95.93%	96.58%
7. Tired	94.55%	92.63%	96.95%	96.80%	92.30%	94.42%
8. Unexplained	93.43%	92.89%	93.99%	94.07%	92.80%	93.39%
9. Suicidal	96.10%	95.28%	96.95%	97.00%	95.20%	96.06%
Average	95.60%	94.83%	96.46%	96.30%	94.71%	95.55%

จากตารางที่ 4.11 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการทดลองทั้ง 9 แบบจำลอง พบว่าแบบจำลองการจำแนกอาการความสนใจลดลงมีความถูกต้องในการจำแนกมากที่สุด

3) ผลการวัดประสิทธิภาพของแบบจำลองที่ใช้งาน 6,000 คุณลักษณะ

จากการสร้างแบบจำลองโดยการกำหนด Top k = 6,000 ซึ่งเป็นการเลือกคุณลักษณะที่มีค่า Gain สูงสุด 6,000 อันดับแรกเข้ามาสร้างแบบจำลอง ผลลัพธ์ดังตารางที่ 4.12

ตารางที่ 4.12 ผลการวัดประสิทธิภาพแบบจำลองที่เรียนรู้ด้วยด้วย 6,000 คุณลักษณะ

	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
1. Depressive	92.08%	91.30%	92.90%	93.03%	91.13%	91.31%
2. Loss of interest	97.77%	98.28%	97.26%	97.23%	98.30%	97.78%
3. Appetite	95.87%	94.88%	96.90%	96.97%	94.77%	95.82%
4. Sleep	94.30%	92.60%	96.15%	96.30%	92.30%	94.18%
5. Slowed thinking	92.30%	91.15%	93.52%	93.70%	90.90%	92.19%
6. Guilt	96.38%	95.43%	97.38%	97.43%	95.33%	96.34%
7. Tired	93.05%	91.20%	95.08%	95.30%	90.80%	92.89%
8. Unexplained	92.70%	91.78%	93.66%	93.80%	91.60%	92.61%
9. Suicidal	95.35%	93.87%	96.93%	97.03%	93.67%	95.27%
Average	94.42%	93.39%	95.53%	95.64%	93.20%	94.27%

จากตารางที่ 4.12 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการทดลองทั้ง 9 แบบจำลอง พบว่าแบบจำลองการจำแนกอาการความสนใจลดลงมีความถูกต้องในการจำแนกมากที่สุดแต่ก็ยังมีค่า Precision No และ Recall Yes ของแบบจำลองความรู้สึกผิดจำแนกได้ดีกว่า

4) ผลการวัดประสิทธิภาพของแบบจำลองที่ใช้งานคุณลักษณะทั้งหมด

จากการสร้างแบบจำลองโดยการเลือกคุณลักษณะที่ทั้งหมดเข้ามาสร้างแบบจำลอง ผลลัพธ์ดังตารางที่ 4.13

ตารางที่ 4.13 ผลการวัดประสิทธิภาพแบบจำลองที่เรียนรู้ด้วยด้วยคุณลักษณะทั้งหมด

	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
1. Depressive	91.73%	90.81%	92.70%	92.87%	90.60%	91.64%
2. Loss of interest	97.77%	98.28%	97.26%	97.23%	98.30%	97.78%
3. Appetite	93.03%	90.98%	95.30%	95.53%	90.53%	92.85%
4. Sleep	90.63%	88.92%	92.50%	92.83%	88.43%	90.63%
5. Slowed thinking	89.42%	88.21%	90.71%	91.00%	87.83%	89.25%
6. Guilt	95.25%	92.73%	98.09%	98.20%	92.30%	95.10%
7. Tired	90.60%	88.81%	92.56%	92.90%	88.30%	90.37%
8. Unexplained	88.97%	87.09%	91.05%	91.50%	86.43%	88.68%
9. Suicidal	93.30%	91.11%	95.74%	95.97%	90.63%	93.11%
Average	92.30%	90.77%	93.99%	94.23%	90.37%	92.16%

จากตารางที่ 4.13 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการทดลองทั้ง 9 แบบจำลอง พบว่าแบบจำลองการจำแนกอาการความสนใจลดลงมีความถูกต้องในการจำแนกมากที่สุดแต่ก็ยังมีค่า Precision No ของแบบจำลองความรู้สึกผิดจำแนกได้ดีกว่า

5) ผลการวัดประสิทธิภาพค่าเฉลี่ยของแบบจำลองทั้งหมด

จากการสร้างแบบจำลองที่ใช้งาน Information Gain ในการคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 4,000 6,000 และเลือกคุณลักษณะทั้งหมด ได้ค่าเฉลี่ยของทั้ง 9 แบบจำลองดังตารางที่ 4.14

ตารางที่ 4.14 ผลการวัดประสิทธิภาพค่าเฉลี่ยของแบบจำลองทั้งหมด

	Feature			
	2,000	**4,000**	**6,000**	**All**
Accuracy	95.85%	95.60%	94.42%	92.30%
Precision Yes	94.97%	94.83%	93.39%	90.77%
Precision No	96.81%	96.46%	95.53%	93.99%
Recall Yes	96.88%	96.30%	95.64%	94.23%
Recall No	94.82%	94.71%	93.20%	90.37%
F-1	95.79%	95.55%	94.27%	92.16%

จากตารางที่ 4.18 ค่าเฉลี่ยผลลัพธ์ทั้งหมดของแบบจำลองที่ถูกสร้างด้วยการคัดเลือกคุณลักษณะ 2,000 คุณลักษณะได้ผลลัพธ์ที่ดีที่สุด เนื่องจากคุณลักษณะที่คัดเลือกโดย Information gain สามารถแบ่งกลุ่มได้ดี

4.3.2 ผลการวัดประสิทธิภาพแบบจำลองโดยข้อมูลสำหรับทดสอบแบบจำลอง

สำหรับการวัดความถูกต้องของข้อมูลทดสอบแบบจำลอง ซึ่งเป็นข้อมูลของผู้ที่เป็นโรคซึมเศร้าจำนวน 15 คน และ คนที่ไม่เป็นโรคซึมเศร้าจำนวน 15 คน โดยทดสอบกับแบบจำลองที่เลือกใช้งาน Information Gain ในการคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 4,000 6,000 และเลือกคุณลักษณะทั้งหมด และมีการกำหนดค่า Boundary ออกเป็น 0, 10, 20, 30, 40, 50, 60, 70, 80 และ 90 ซึ่งเป็นการกำหนดเพดานของค่าความน่าจะเป็น ถ้าข้อความไหนมีค่าไม่ถึงค่าเพดานก็จะไม่นำมาทำการ Vote ensemble ได้ผลการทดลองดังต่อไปนี้

1) ผลการวัดประสิทธิภาพของการทำนายแบบจำลองที่ใช้งาน 2,000 คุณลักษณะ มีผลลัพธ์ดังตารางที่ 4.15

ตารางที่ 4.15 ผลการทำนายของแบบจำลองที่สร้างด้วย 2,000 คุณลักษณะ

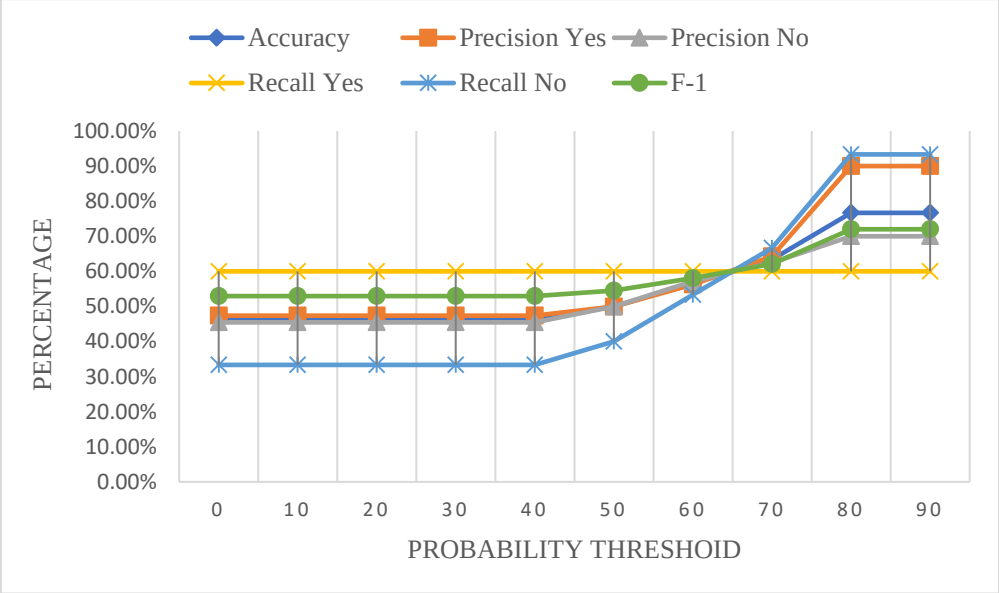
	ค่าทำนาย																			
ความน่าจะเป็น	**0**		**0.10**		**0.20**		**0.30**		**0.40**		**0.50**		**0.60**		**0.70**		**0.80**		**0.90**	
ค่าความจริง	**Yes**	**No**	**Yes**	**No**	**Yes**	**No**	**yes**	**no**	**yes**	**no**	**yes**	**no**	**yes**	**no**	**yes**	**no**	**yes**	**no**	**yes**	**no**
	9	6	9	6	9	6	9	6	9	6	9	6	9	6	9	6	9	6	9	6
	10	5	10	5	10	5	10	5	10	5	9	6	7	8	5	10	1	14	1	14

จากตารางที่ 4.15 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการเลือกคุณลักษณะที่ 2,000 สำหรับสร้างแบบจำลองได้ผลลัพธ์ดังตารางที่ 4.20

ตารางที่ 4.16 ผลการวัดประสิทธิภาพการทำนายของแบบจำลองที่สร้างด้วย 2,000 คุณลักษณะ

Probability	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
0.0	46.67%	47.37%	45.45%	60.00%	33.33%	52.94%
0.10	46.67%	47.37%	45.45%	60.00%	33.33%	52.94%
0.20	46.67%	47.37%	45.45%	60.00%	33.33%	52.94%
0.30	46.67%	47.37%	45.45%	60.00%	33.33%	52.94%
0.40	46.67%	47.37%	45.45%	60.00%	33.33%	52.94%
0.50	50.00%	50.00%	50.00%	60.00%	40.00%	54.55%
0.60	56.67%	56.25%	57.14%	60.00%	53.33%	58.06%
0.70	63.33%	64.29%	62.50%	60.00%	66.67%	62.07%
0.80	76.67%	90.00%	70.00%	60.00%	93.33%	72.00%
0.90	76.67%	90.00%	70.00%	60.00%	93.33%	72.00%

จากตารางที่ 4.15 สามารถนำ Accuracy, Precision Yes, Precision No, Recall Yes, Recall No และ F-1 มาแสดงเป็นกราฟดังรูปที่ 4.1



รูปที่ 4.1 กราฟแสดงประสิทธิภาพของแบบจำลองที่สร้างด้วย 2,000 คุณลักษณะ

จากรูปที่ 4.1 จะเห็นได้ว่าค่าเกือบทุกค่ายกเว้นค่า Recall yes จะมีการปรับตัวขึ้นไปหลังจากการกำหนดค่าความน่าจะเป็นที่ 80 ถึง 90

2) ผลการวัดประสิทธิภาพของการทำนายแบบจำลองที่ใช้งาน 4,000 คุณลักษณะ มีผลลัพธ์ดังตารางที่ 4.17

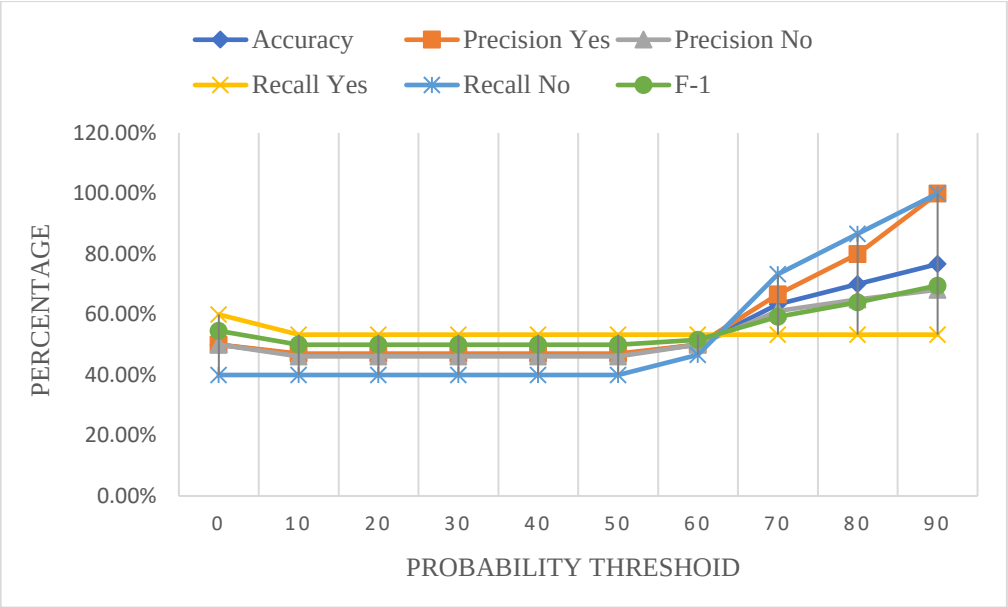
ตารางที่ 4.17 ผลการทำนายของแบบจำลองที่สร้างด้วย 4,000 คุณลักษณะ

	ค่าทำนาย																			
ความน่าจะเป็น	0		0.10		0.20		0.30		0.40		0.50		0.60		0.70		0.80		0.90	
ค่าความจริง	Yes	No	Yes	No	Yes	No	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no
	9	6	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7
	9	6	9	6	9	6	9	6	9	6	9	6	8	7	4	11	0	15	0	15

จากตารางที่ 4.17 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการเลือกคุณลักษณะที่ 4,000 สำหรับสร้างแบบจำลองได้ผลลัพธ์ดังตารางที่ 4.18

ตารางที่ 4.18 ผลการวัดประสิทธิภาพการทำนายของแบบจำลองที่สร้างด้วย 4,000 คุณลักษณะ

Probability	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
0	50.00%	50.00%	50.00%	60.00%	40.00%	54.55%
0.10	46.67%	47.06%	46.15%	53.33%	40.00%	50.00%
0.20	46.67%	47.06%	46.15%	53.33%	40.00%	50.00%
0.30	46.67%	47.06%	46.15%	53.33%	40.00%	50.00%
0.40	46.67%	47.06%	46.15%	53.33%	40.00%	50.00%
0.50	46.67%	47.06%	46.15%	53.33%	40.00%	50.00%
0.60	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.70	63.33%	66.67%	61.11%	53.33%	73.33%	59.26%
0.80	70.00%	80.00%	65.00%	53.33%	86.67%	64.00%
0.90	76.67%	100.00%	68.18%	53.33%	100.00%	69.57%



รูปที่ 4.2 กราฟแสดงประสิทธิภาพของแบบจำลองที่สร้างด้วย 4,000 คุณลักษณะ

จากรูปที่ 4.2 จะเห็นได้ว่าค่าเกือบทุกค่ายกเว้นค่า Recall yes จะมีการปรับตัวขึ้นไปหลังจากการกำหนดค่าความน่าจะเป็นที่ 70 ถึง 90

3) ผลการวัดประสิทธิภาพของการทำนายแบบจำลองที่ใช้งาน 6,000 คุณลักษณะ มีผลลัพธ์ดังตารางที่ 4.19

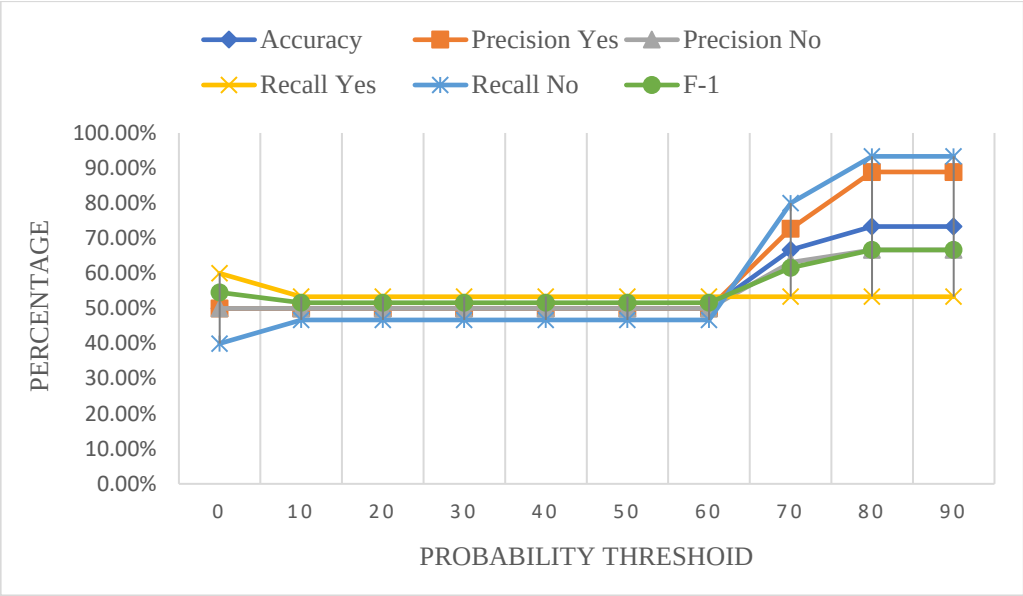
ตารางที่ 4.19 ผลการทำนายของแบบจำลองที่สร้างด้วย 6,000 คุณลักษณะ

	ค่าทำนาย																			
ความน่าจะเป็น	0		0.10		0.20		0.30		0.40		0.50		0.60		0.70		0.80		0.90	
ค่าความจริง	Yes	No	Yes	No	Yes	No	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no
	9	6	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7	8	7
	9	6	8	7	8	7	8	7	8	7	8	7	8	7	3	12	1	14	1	14

จากตารางที่ 4.19 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการเลือกคุณลักษณะที่ 6,000 สำหรับสร้างแบบจำลองได้ผลลัพธ์ดังตารางที่ 4.20

ตารางที่ 4.20 ผลการวัดประสิทธิภาพการทำนายของแบบจำลองที่สร้างด้วย 6,000 คุณลักษณะ

Probability	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
0	50.00%	50.00%	50.00%	60.00%	40.00%	54.55%
0.10	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.20	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.30	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.40	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.50	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.60	50.00%	50.00%	50.00%	53.33%	46.67%	51.61%
0.70	66.67%	72.73%	63.16%	53.33%	80.00%	61.54%
0.80	73.33%	88.89%	66.67%	53.33%	93.33%	66.67%
0.90	73.33%	88.89%	66.67%	53.33%	93.33%	66.67%



รูปที่ 4.3 กราฟแสดงประสิทธิภาพของแบบจำลองที่สร้างด้วย 6,000 คุณลักษณะ

จากรูปที่ 4.3 จะเห็นได้ว่าค่าเกือบทุกค่ายกเว้นค่า Recall yes จะมีการปรับตัวขึ้นไปหลังจากการกำหนดค่าความน่าจะเป็นที่ 70 ถึง 90

4) ผลการวัดประสิทธิภาพของการทำนายแบบจำลองที่ใช้งานคุณลักษณะทั้งหมด มีผลลัพธ์ดังตารางที่ 4.21

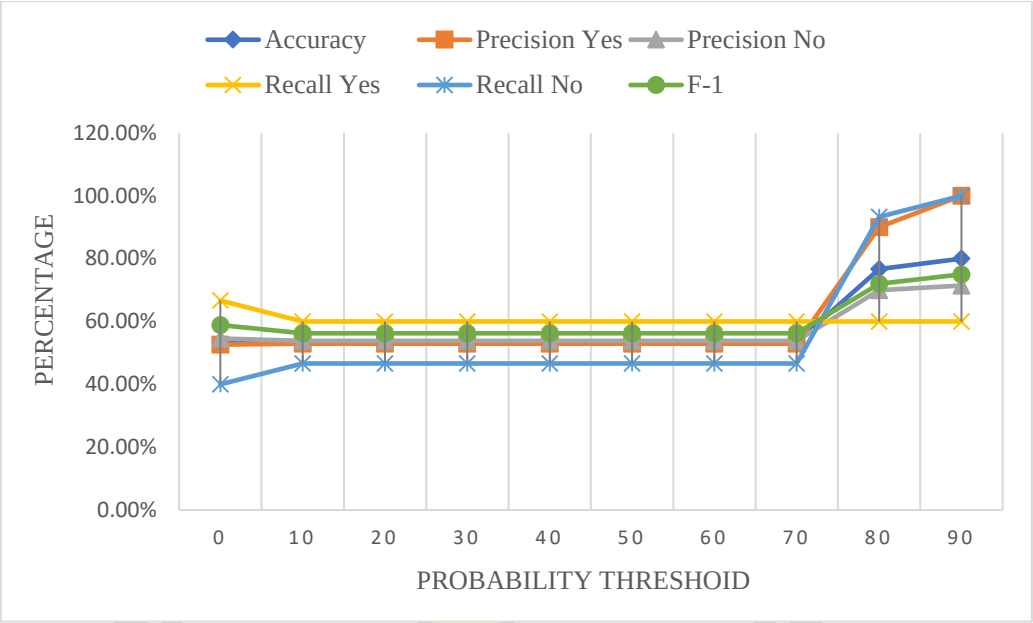
ตารางที่ 4.21 ผลการทำนายของแบบจำลองที่สร้างด้วยคุณลักษณะทั้งหมด

	ค่าทำนาย																			
ความน่าจะเป็น	0		0.10		0.20		0.30		0.40		0.50		0.60		0.70		0.80		0.90	
ค่าความจริง	Yes	No	Yes	No	Yes	No	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no
	10	5	9	6	9	6	9	6	9	6	9	6	9	6	9	6	9	6	9	6
	9	6	8	7	8	7	8	7	8	7	8	7	8	7	8	7	1	14	0	15

จากตารางที่ 4.21 เป็นการวัดประสิทธิภาพด้วย Confusion Matrix ซึ่งผลการเลือกคุณลักษณะทั้งหมด สำหรับสร้างแบบจำลองได้ผลลัพธ์ดังตารางที่ 4.22

ตารางที่ 4.22 ผลการวัดประสิทธิภาพการทำนายของแบบจำลองที่สร้างด้วยคุณลักษณะทั้งหมด

Probability	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
0	53.33%	52.63%	54.55%	66.67%	40.00%	58.82%
0.10	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.20	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.30	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.40	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.50	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.60	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.70	53.33%	52.94%	53.85%	60.00%	46.67%	56.25%
0.80	76.67%	90.00%	70.00%	60.00%	93.33%	72.00%
0.90	80.00%	100.00%	71.43%	60.00%	100.00%	75.00%



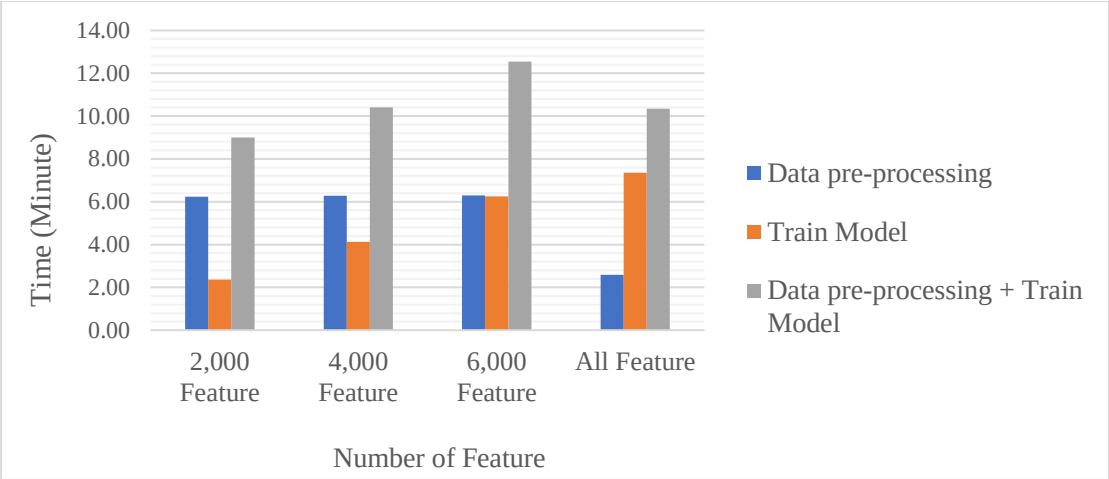
รูปที่ 4.4 กราฟแสดงประสิทธิภาพของแบบจำลองที่สร้างด้วยคุณลักษณะทั้งหมด

จากรูปที่ 4.4 จะเห็นได้ว่าค่าเกือบทุกค่ายกเว้นค่า Recall yes จะมีการปรับตัวขึ้นไปหลังจากการกำหนดค่าความน่าจะเป็นที่ 80 ถึง 90

จากการทดลองทั้งหมดพบว่าค่าความน่าจะเป็นที่เหมาะสมแก่การเป็นค่า Boundary อยู่ที่ 80 ถึง 90 เนื่องจากค่าเป็นค่าความน่าจะเป็นที่สูงสามารถกรองข้อความที่ไม่เข้าข่ายในอาการที่เกี่ยวข้องกับโรคซึมเศร้าออกไป ทำให้จำการแนกข้อความตามอาการได้ดี และการเลือกนำเข้าคุณลักษณะทั้งหมดเพื่อสร้างแบบจำลองให้ผลลัพธ์ที่ดีกว่าการใช้งานการเลือกคุณลักษณะ เนื่องจากข้อมูลหลังจากการทำกระบวนการเตรียมข้อมูล ข้อความที่นำเข้าจะเหลือคำเพียงไม่กี่คำ ทำให้การเลือกใช้งานคุณลักษณะจำนวนมากเข้ามาเพื่อทำนายกับข้อมูลในโลกความจริง ให้ผลที่ได้ดีกว่าการตัดคุณลักษณะออกไป

4.3.3 ประสิทธิภาพด้านเวลาในการสร้างแบบจำลอง

ในการสร้างแบบจำลองทั้ง 9 แบบจำลอง ซึ่งใช้งานการคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 4,000 6,000 และเลือกคุณลักษณะทั้งหมด มีผลการทดลองดังรูปที่ 4.5



รูปที่ 4.5 ผลการวัดประสิทธิภาพเวลาเฉลี่ยของแบบจำลองทั้งหมด

จากรูปที่ 4.5 พบว่าการสร้างแบบจำลองโดยการเลือก 2,000 คุณลักษณะได้เวลาน้อยที่สุด และการเลือก 6,000 คุณลักษณะมีการใช้เวลานานที่สุดแทนที่จะเป็นการเลือกใช้คุณลักษณะทั้งหมด เนื่องจากในขั้นตอนการเลือกคุณลักษณะโดยการใช้งาน Information gain มีการใช้เวลาในการคำนวณที่มาพอสมควร และการใช้งาน Bayes มีการสร้างแบบจำลองที่ไวแม้จะมีคุณลักษณะที่มาก ทำให้เวลาของการใช้งานการเลือกคุณลักษณะทั้งหมดน้อยกว่า แต่ถ้าหากมีคุณลักษณะที่มากกว่านี้ การเลือกใช้งานการคัดเลือกคุณลักอาจจะดีกว่า

บทที่ 5

สรุปผล อภิปรายผล และข้อเสนอแนะ

จากจุดประสงค์ในการทำวิจัยครั้งนี้ผู้วิจัยได้ทำการทดลองทำเหมืองข้อความเพื่อพัฒนากระบวนการในการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์ ซึ่งได้นำเอาอัลกอริธึม Bayes เข้ามาสร้างแบบจำลองการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์เพื่อช่วยในการวินิจฉัยของแพทย์หรือให้หน่วยงานที่เกี่ยวข้องนำไปประยุกต์ใช้งานได้ในอนาคต

5.1 สรุปผลการศึกษาและอภิปรายผล

ในงานวิจัยในครั้งนี้ผู้วิจัยได้ทำการกรองข้อมูลจาก Twitter โดยการตัดข้อความที่เป็นการ Retweet การตัดข้อความที่เป็นลิ้งเข้าใช้งานเว็บไซต์ การตัดข้อความที่เกี่ยวข้องกับชื่อบุคคลที่ถูกแทกในโพสต์ และการตัด Stopword สำหรับการให้น้ำหนักเลือกใช้งานวิธีการ Binary term occurrence แล้วคัดเลือกแอตทริบิวต์ด้วย Information Gain ที่กำหนด Top k = 2,000 4,000 6,000 และเลือกคุณลักษณะทั้งหมด ในการสร้างแบบจำลองผู้วิจัยเลือกใช้การแบ่งข้อมูลโดยใช้วิธีการ 10 – Fold Cross Validation ซึ่งใช้งานอัลกอริธึม Bayes ในการสร้างแบบจำลอง พบว่าทั้ง 9 แบบจำลองมีค่าเฉลี่ย คือ Accuracy = 95.85%, Precision Yes = 94.97%, Precision No = 96.81%, Recall Yes = 96.88%, Recall No = 94.82% และ F-1 = 95.79% สำหรับการทดสอบแบบจำลองด้วยข้อมูลผู้ใช้งานที่เป็นโรคซึมเศร้าและไม่เป็นโรคซึมเศร้าพบว่าการทำนายผลทั้ง 30 บุคคล โดยมีการกำหนด Boundary เพื่อหาช่วงการคัดเลือกค่าความน่าจะเป็นที่เหมาะสมสำหรับการโหวต โดยกำหนด Boundary ที่ 0, 10, 20, 30, 40, 50, 60, 70, 80 และ 90 ผลลัพธ์ที่ได้ปรากฏว่าช่วงค่าความน่าจะเป็นที่เหมาะสมอยู่ที่ 80 – 90 ได้ผลลัพธ์ที่ดีที่สุด คือ Accuracy = 80.00%, Precision Yes = 100.00%, Precision No = 71.43%, Recall Yes = 60.00%, Recall No = 100.00% และ F-1 = 75.00% จากการทดลองทั้งหมดสรุปได้ว่า

1. พบว่าเทคนิคในการตัดข้อความที่เป็นการ Retweet บน Twitter มีผลทำให้ค่าความถูกต้องในการทำนายเพิ่มขึ้นเนื่องจากข้อความที่เป็นการ Retweet บน Twitter ส่วนมากเป็นข้อความที่เกี่ยวกับข่าว ซึ่งถ้าหากเป็นข้อความที่เข้าข่ายใน 9 ลักษณะอาการที่เกี่ยวข้องกับโรคซึมเศร้า จะทำให้การทำนายของแบบจำลองผิดผลาดได้

2. พบว่าเทคนิคการให้น้ำหนักแบบ Binary term occurrence เหมาะสำหรับการทำนายข้อความที่มีลักษณะสั้น เนื่องจากข้อความที่ได้หลังจากการกรองข้อมูล มีคุณลักษณะสำคัญเพียงไม่กี่คำ

3. พบว่าเทคนิคการคัดเลือกแอตทริบิวต์ด้วย Information Gain มีผลทำให้เวลาในการสร้างแบบจำลองลดน้อยลงเนื่องจากคุณลักษณะที่ได้จาก Bag of word มีจำนวนมาก ทำให้การสร้างแบบจำลองและทำนายข้อมูลนั้นล่าช้า และการลดคุณลักษณะให้ได้จุดเหมาะสมทำให้ประสิทธิภาพของการทดสอบแบบจำลองเพิ่มขึ้น

4. พบว่าข้อมูลของบุคคลที่นำมาทดสอบการทำนายจากแบบจำลอง พบว่าจำนวนในการโพสต์ต่อวันมีผลต่อการทำนายเนื่องจากผู้ป่วยที่ไม่ป่วยเป็นโรคซึมเศร้า ถ้าโพสต์ข้อความติดต่อกันแล้วเป็นข้อความที่เกี่ยวกับข่าวหรือการแสดงทัศนคติต่อเหตุการณ์รอบข้างที่มีลักษณะเข้าข่ายกับอาการซึมเศร้า เช่น แสดงความเสียใจในเหตุการณ์ต่าง ๆ เช่น “We have lost a pillar of our industry, and a huge supporter of Aussie film. RIP Greg Cote. You will be sorely missed.” ซึ่งเป็นข้อความที่ไม่เกี่ยวข้องกับอาการแต่มีคำที่เข้าข่าย เช่น lost กับ RIP และข้อความที่มีการใช้ Hastag กำกวมอย่าง “Can't wait for today's matches! #ADDICTED!!! knvb fifaworldcup” ซึ่งเป็นข้อความที่ใช้ Hastag คำว่า #ADDICTED ทำให้การทำนายนั้นคาดเคลื่อนได้

5. งานวิจัยในครั้งนี้ยังไม่สามารถพยากรณ์ครอบคลุมบุคคลบนโลกความจริงมากนัก เนื่องจากข้อมูลที่นำมาสร้างแบบจำลองยังขาดคุณลักษณะที่เหมาะสมที่จะครอบคลุมคนส่วนใหญ่

5.2 ข้อเสนอแนะ

1. จากจุดประสงค์ในการทำวิจัยครั้งนี้ ผู้วิจัยได้ทำการวิธีการในการสร้างแบบจำลองให้มีประสิทธิภาพสูงสุดในการการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์ ผลปรากฏว่าข้อมูลที่นำมาสร้างแบบจำลองมีผลต่อความถูกต้องในการทำนายข้อมูลบนโลกความจริง และการใช้งานเพียงแค่ข้อความในการทำนายโรคซึมเศร้าอาจจะไม่เพียงพอต่อการจำแนกพฤติกรรม ซึ่งการใช้งานข้อมูลอื่น ๆ ของตัวบุคคล เช่น รูปภาพ อายุ ความสัมพันธ์ และ เพศ อาจจะมีผลต่อการจำแนกพฤติกรรมที่เข้าข่ายเป็นโรคซึมเศร้าได้

2. ในขั้นตอนการคัดเลือกคุณลักษณะที่เหมาะสมจำนวน 2,000 4,000 และ 6,000 คุณลักษณะ ผู้วิจัยไม่ได้ทำการปรับสมดุลของจำนวนคำในแต่ละคลาส จึงทำให้คำในคลาสที่คัดเลือกมามีการเกิดปัญหา Imbalanced จึงอาจจะทำให้ผลการทำนายเอียงไปด้านใดด้านหนึ่ง

บรรณานุกรม

บรรณานุกรม

[1] Organization WH. 2017 World Health Day. [ออนไลน์].สืบค้นเมื่อ 29 กันยายน 2561]; http://www.who.int/campaigns/world-health-day/2017/campaign-essentials/en/.

[2] Lin LY, Sidani JE, Shensa A, Radovic A, Miller E, Colditz JB, et al. ASSOCIATION BETWEEN SOCIAL MEDIA USE AND DEPRESSION AMONG U.S. YOUNG ADULTS. Depression and anxiety 2016; 33[4]: 323-331.

[3] Bifet A, Frank E. Sentiment Knowledge Discovery in Twitter Streaming Data. In: Pfahringer B, Holmes G, Hoffmann A, eds. Discovery Science: 13th International Conference, DS 2010, Canberra, Australia, October 6-8, 2010 Proceedings. Berlin, Heidelberg: Springer Berlin Heidelberg 2010; 1-15.

[4] Keumhee K, Chanhee Y, Eun Yi K. Identifying depressive users in Twitter using multimodal analysis. 2016 International Conference on Big Data and Smart Computing (BigComp); 18-20 Jan. 2016: 231-238.

[5] Sho T, Yusuke K, Fumio K, Kosuke N, Yuichi I, Hiroyuki O. Recognizing Depression from Twitter Activity. ACM, 2015: 3187-3196.

[6] Aldarwish MM, Ahmad HF. Predicting Depression Levels Using Social Media Posts. 2017 IEEE 13th International Symposium on Autonomous Decentralized System (ISADS); 22-24 March 2017; 277-280.

[7] Wang X, Zhang C, Ji Y, Sun L, Wu L, Bao Z. A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In: Li J, Cao L, Wang C, Tan KC, Liu B, Pei J, et al., eds. Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2013 International Workshops: DMApps, DANTH, QIMIE, BDM, CDA, CloudSD, Gold Coast, QLD, Australia, April 14-17, 2013, Revised Selected Papers. Berlin, Heidelberg: Springer Berlin Heidelberg 2013: 201-213.

[8] Punnee S, Nongyao N-a, Ian T. N. Bagging Model with Cost Sensitive Analysis on Diabetes Data. Information Technology Journal 2015; 1182-90.

[9] นันทิรา หงษ์ศรีสุวรรณ์. "ภาวะซึมเศร้า" วารสาร มฉกวิชาการ ปีที่ 19, 2559: 105-118.

[10] ดาวรุ่ง งามเลิศ. "ผลของโปรแกรมการฝึกไบโอฟีดแบคชนิดควบคุมอัตราการแปรปรวนการเต้นของหัวใจต่ออาการซึมเศร้าในผู้ใหญ่โรคซึมเศร้า". พยาบาลศาสตรมหาบัณฑิต การพยาบาลจิตเวชและสุขภาพจิต คณะพยาบาลศาสตร์ มหาวิทยาลัยธรรมศาสตร์, 2558.

[11] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders (DSM-5). Washington DC: American Psychiatric Association; 2013.

[12] Pang B, Lee L. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval 2008; 2[1–2]: 1-135.

[13] กานดา แผ่วัฒนากุล. "การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์". วิทยาศาสตรมหาบัณฑิต บริหารเทคโนโลยีสารสนเทศ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์, 2555.

[14] Jiawei H, Micheline K. Data Mining Concepts and Techniques. Kaufmann Publishers: Oxford Morgan; 2006.

[15] สุวนีย์ กุลกรินีธรรม. "การใช้เทคนิคเหมืองข้อมูลเพื่อการจัดกลุ่มหลักสูตรตามกลุ่มสาขาวิชา ISCED กรณีศึกษากลุ่มสาขาวิชาวิทยาศาสตร์". วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศิลปากร, 2549.

[16] ชัชชฎา วันดี. การศึกษาปัจจัยที่ 1 ผลต่อการเลือกอาชีพของนิสิตระดับปริญญาตรีหลังสำเร็จการศึกษา โดยใช้เทคนิคเหมืองข้อมูล. The 9th National Conference on Computing and Information Technology 2013; 80-85.

[17] ฉัตรเกล้า เจริญผล. DATA MINING. พิมพครั้งที่ 4. Mahasarakham University; 2014.

[18] sit. เทคนิค Support Vector Machine (SVM) และซอฟต์แวร์ HR-SVM. [ออนไลน์].สืบค้นเมื่อ 1 ตุลาคม 2560]; http://dataminingtrend.com/2014/support-vector-machine-svm/.

[19] Mae K. การวัดประสิทธิภาพ. [ออนไลน์].สืบค้นเมื่อ 1 ตุลาคม 2560]; http://slideplayer.in.th/slide/2126510/.

[20] Eakasit P. การแบ่งข้อมูลเพื่อนำมาทดสอบประสิทธิภาพของแบบจำลอง. [ออนไลน์].สืบค้นเมื่อ 1 ตุลาคม 2560]; https://th.linkedin.com/pulse/การแบงขอมลเพอนำมาทดสอบประสทธภาพของแบบจำลอง-eakasit-pacharawongsakda.

[21] Napong W. Confusion Matrix. [ออนไลน์].สืบค้นเมื่อ 1 ตุลาคม 2560]; https://plagad.wordpress.com/2010/08/26/confusion-matrix/.

[22] Nadeem M. Identifying Depression on Twitter. CoRR 2016; abs/1607.07384

[23] กรมสุขภาพจิต. แบบประเมินโรคซึมเศร้า 9 คำถาม (9Q). [ออนไลน์].สืบค้นเมื่อ 1 พฤษภาคม 2561]; http://www.prdmh.com/แบบประเมินโรคซึมเศร้า-9-คำถาม-9q.html.

[24] sit. การสร้างโมเดล Ensemble แบบต่างๆ. [ออนไลน์].สืบค้นเมื่อ 20 พฤศจิกายน 2560]; http://dataminingtrend.com/2014/data-mining-techniques/ensemble-model/.

[25] ไทยรัฐออนไลน์. คนเป็นก็ต้องสู้ต่อไป ย้อนรอย 'คนดัง' มีคนตายและไม่ตายด้วย 'โรคซึมเศร้า'. [ออนไลน์].สืบค้นเมื่อ 1 สิงหาคม 2561]; https://www.thairath.co.th/content/1011892.

[26] Okadmin. รู้หรือไม่! 12 คนดังเหล่านี้เคยป่วยเป็นโรคซึมเศร้า. [ออนไลน์].สืบค้นเมื่อ 1 สิงหาคม 2561]; http://www.okmagazine-thai.com/celebrities-who-battled-depression/.

[27] Alpaca. 8 คนดังฮอลลิวูด ที่ออกมาเผยว่า ต้องเผชิญหน้ากับอาการซึมเศร้า. [ออนไลน์]. สืบค้นเมื่อ 1 สิงหาคม 2561]; https://gossipstar.mthai.com/hollywood/inter/66143.

[28] sit. การทำ Text Mining ภาษาไทยด้วย RapidMiner Studio 9 และ Python. [ออนไลน์]. สืบค้นเมื่อ 20 พฤศจิกายน 2560]; http://dataminingtrend.com/2014/thai-text-mining-rapidminer-python/.

[29] Damian D. English Stopwords. [ออนไลน์].สืบค้นเมื่อ 11 สิงหาคม 2561]; http://dataminingtrend.com/2014/thai-text-mining-rapidminer-python/.

ประวัติผู้เขียน

ชื่อ	นายดำรงเดช เดินรีบรัมย์
วันเกิด	วันที่ 25 มิถุนายน พ.ศ. 2537
สถานที่เกิด	อำเภอเมืองยาง จังหวัดนครราชสีมา
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 232 หมู่ 1 ตำบลเมืองยาง อำเภอเมืองยาง จังหวัดนครราชสีมา รหัสไปรษณีย์ 30270
ตำแหน่งหน้าที่การงาน	โปรแกรมเมอร์
สถานที่ทำงานปัจจุบัน	บริษัท เทอร์มินอล เอธ จำกัด 160-161 ถนนเฉลิมพระเกียรติ ร.9 ตำบลตลาด อำเภอเมืองมหาสารคาม จังหวัดมหาสารคาม รหัสไปรษณีย์ 44000
ประวัติการศึกษา	พ.ศ. 2551 มัธยมศึกษาตอนต้น โรงเรียนจุฬาภรณราชวิทยาลัย บุรีรัมย์ อำเภอสตึก จังหวัดบุรีรัมย์ พ.ศ. 2554 มัธยมศึกษาตอนปลาย โรงเรียนจุฬาภรณราชวิทยาลัย บุรีรัมย์ อำเภอสตึก จังหวัดบุรีรัมย์ พ.ศ. 2558 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ มหาวิทยาลัยราชภัฏบุรีรัมย์ พ.ศ. 2562 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม