การเปรียบเทียบประสิทธิภาพโครงสร้างเหมืองข้อมูลเพื่อจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์

Comparison of data mining structure performance for depressive classification if twitter users from their posts on twitter of user behaviors

ดำรงเดช เดินรีบรัมย์[1], ฉัตรเกล้า เจริญผล[2], จริยา จิรานุกูล[3]
Damrongdet Doenribram[1], Chatklaw Jareanpon[2], Jariya Jiranukool[3]

Received: 18 September 2019 ; Revised: 6 January 2020 ; Accepted: 28 January 2020

บทคัดย่อ

ในปี ค.ศ. 2018 องค์กรอนามัยโลกและกรมสุขภาพจิตระบุว่า โรคซึมเศร้าทำให้เกิดการสูญเสียของสุขภาพเป็นอันดับ 2 โดยเกิดจากการใช้งานโซเชียลมีเดียอย่างไม่เหมาะสม เสพสื่ออย่างไม่ระมัดระวัง ทำให้เกิดความเครียด ความรุนแรง และส่งผลทำให้เกิดโรคซึมเศร้าได้ งานวิจัยนี้มีจุดประสงค์ของการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์ โดยมีการเปรียบเทียบการจำแนกระหว่างแบบหนึ่งระดับและแบบสองระดับ การจำแนกหนึ่งระดับจะใช้งานอัลกอริทึม Bayes และอัลกอริทึม SVM จำแนกข้อความทั่วไปและข้อความที่บ่งบอกถึงลักษณะอาการซึมเศร้าตามแบบสอบถาม DSM-5 ได้แก่ อารมณ์ซึมเศร้า ความสนใจลดลง น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต นอนไม่หลับหรือนอนหลับมากกว่าปกติ ร่างกายอ่อนเพลีย รู้สึกตนเองไร้ค่า สมาธิสั้น เคลื่อนไหวช้าและคิดฆ่าตัวตาย การจำแนกสองระดับจะใช้งานอัลกอริทึม SVM เพื่อจำแนกข้อความทั่วไปกับข้อความที่บ่งบอกถึงโรคซึมเศร้า ต่อมาใช้งานอัลกอริทึม Bayes เปรียบเทียบกับอัลกอริทึม Random Forest เพื่อจำแนก 9 อาการที่บ่งบอกถึงโรคซึมเศร้าตามแบบสอบถาม DSM-5 โดยใช้ข้อมูล 2 ชุด ได้แก่ ชุดเรียนรู้ (Training set) และชุดทดสอบ (Test set) ที่มาจากการโพสต์ของดาราต่างประเทศ ผลการทดลองของชุด Training set ระหว่างการจำแนกหนึ่งระดับและการจำแนกสองระดับ คือ หนึ่งระดับอัลกอริทึม Bayes ได้ Accuracy=82.55% และอัลกอริทึม SVM ได้ Accuracy=96.18% การจำแนกสองระดับอัลกอริทึม SVM ได้ Accuracy=98.20% ส่วนผลการทดลองอัลกอริทึม SVM จับคู่กับอัลกอริทึม Bayes ได้ Accuracy=83.23% และอัลกอริทึม SVM จับคู่กับอัลกอริทึม Random Forest ได้ Accuracy=94.45% ผลการทดลองของชุด Test set โดยมีการกำหนดค่าความน่าจะเป็นตั้งแต่ 0.1-0.9 การจำแนกหนึ่งระดับอัลกอริทึม Bayes ได้ Accuracy=76.67% และอัลกอริทึม SVM ได้ Accuracy=70.00% การจำแนกสองระดับอัลกอริทึม SVM จับคู่กับอัลกอริทึม Bayes ได้ Accuracy=73.33% ส่วนอัลกอริทึม SVM จับคู่กับอัลกอริทึม Random Forest ได้ Accuracy=70.00%

คำสำคัญ: โรคซึมเศร้า การทำเหมืองความคิดเห็น เทคนิคการจำแนกข้อมูล การทำเหมืองข้อความคำ

Abstract

In 2018, The World Health Organization (WHO) and Department of Mental Health (DMH), specified that major depressive disorder (MDD) was the second most important disease that it is probably caused by social media usage affecting stress and leading to violence, and depression . This research proposes the depressive classification from posts on twitter of user behaviors and compared the accuracy of two classifiers between one level and two levels:- (1) one level: using the Bayes algorithm created a model for classification between general and symptoms based on a symptoms detailed in a questionnaire (DSM-5) including as follows: depression, loss of interest, loss of

---

[1] นิสิตปริญญาโท, สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150
[2] ผู้ช่วยศาสตราจารย์ สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาการสารสนเทศ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150
[3] ผู้ช่วยศาสตราจารย์ สาขาจิตเวชศาสตร์ คณะแพทยศาสตร์ มหาวิทยาลัยมหาสารคาม อำเภอกันทรวิชัย จังหวัดมหาสารคาม 44150
[1] Master's degree, Computer Science, Faculty of Informatics, Mahasarakham University, Kantharawichai, District, Maha Sarakham 44150, Thailand.
[2] Assistant professor, Computer Science, Faculty of Informatics, Mahasarakham University, Kantharawichai, District, Maha Sarakham 44150, Thailand.
[3] Assistant professor, Department of Psychiatry, Faculty of Medicine, Mahasarakham University, Kantharawichai, District, Maha Sarakham 44150, Thailand.

appetite, abnormal sleep, slowed thinking, guilt, tiredness, unexplained and suicidal ideation. (2) Two levels: Using the SVM algorithm created a model for classification between general and depression. Using the Bayes algorithm compared with the Random Forest algorithm for classification of symptoms in a questionnaire (DSM-5). The data came from real postings of international celebrities. The dataset is divided into 2 sets: a training set and a test set. Finally, the results are demonstrated in a training set prediction between one level and two levels: One level: the Bayes algorithm showed that the accuracy=82.55%, and the SVM algorithm showed that the accuracy=96.18%. Two level: the SVM algorithm showed that the accuracy=98.20%. SVM algorithm pair with Bayes algorithm showed that the accuracy=82.23%, and SVM algorithm pair with the Random Forest algorithm showed that the accuracy=91.45%. The results of test set, by the boundary of probability are variously set 0.1 to 0.9 that prediction between one level and two levels : One level: the Bayes algorithm showed that the accuracy=76.67%, and the SVM algorithm showed that the accuracy=70.00%. Two level: SVM algorithm pair with Bayes algorithm showed that the accuracy=73.33%. SVM algorithm pair with Random Forest algorithm showed that the accuracy=70.00%.

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

บทนำ

ในปี ค.ศ. 2018 องค์กรอนามัยโลกระบุว่า โรคซึมเศร้า (Major Depressive Disorder ตัวย่อ MDD)[1,2] ทำให้เกิดการสูญเสียของสุขภาพเป็นอันดับ 2 โดยส่งผลทำให้ผู้ป่วยเกิดความเสี่ยงต่อการฆ่าตัวตายมากกว่าคนทั่วไปถึง 20 เท่า สาเหตุของการเกิดโรคซึมเศร้าเกิดจาก 3 ปัจจัยหลัก ได้แก่ ปัจจัยทางกรรมพันธุ์ ปัจจัยทางจิตใจและปัจจัยด้านสิ่งแวดล้อม อาการของโรคซึมเศร้าจะมีลักษณะ 9 อาการ ได้แก่ 1. อารมณ์ซึมเศร้า 2. ความสนใจลดลง 3. น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต 4. นอนไม่หลับหรือนอนหลับมากกว่าปกติ 5. ร่างกายอ่อนเพลีย 6. รู้สึกตนเองไร้ค่า 7. สมาธิสั้น 8. เคลื่อนไหวช้า และ 9. คิดฆ่าตัวตาย

ปัจจุบันโซเชียลมีเดียมีการใช้งานอย่างแพร่หลายในชีวิตประจำวันไม่ว่าจะเป็น Facebook Instagram หรือ Twitter โดยกรมสุขภาพจิตระบุว่า การใช้งานโซเชียลมีเดียส่งผลทำให้เกิดโรคซึมเศร้า เกิดจากการใช้งานอย่างไม่เหมาะสม หรือเสพสื่ออย่างไม่ระมัดระวัง ส่งผลทำให้เกิดความเครียด ความรุนแรง รวมไปถึงภาวะซึมเศร้าได้ นอกจากนี้ยังแสดงภาวะซึมเศร้าผ่านการโพสต์และการแชทได้

การทำเหมืองข้อมูล (Data mining) เป็นเทคนิคการวิเคราะห์ข้อมูลจำนวนมหาศาลด้วยอัลกอริทึมมา จำแนก ทำนาย หาความสำคัญ หรือหาความสัมพันธ์ของข้อมูล โดยปัจจุบันข้อมูลมหาศาลที่เกิดขึ้นตลอดเวลาล้วนมาจากโซเชียลมีเดียและได้มีงานวิจัยที่นำข้อมูลจากโซเชียลมีเดียเข้ามาทำเหมืองข้อมูลเป็นจำนวนมากดังงานวิจัยต่อไปนี้

ค.ศ. 2014 งานวิจัยของ Yoon[3] และคณะได้มีการใช้งานการทำเหมืองข้อมูลเพื่อค้นหาความสัมพันธ์ของการเกิดภาวะซึมเศร้าเรื้อรังด้วยข้อมูลจากหน่วยงาน Behavioral Risk Factor Surveillance System (BRFSS) ที่เป็นหน่วยงานเกี่ยวกับการสำรวจสุขภาพทางโทรศัพท์ของของสหรัฐอเมริกา โดยในการสร้างแบบจำลองครั้งนี้ใช้อัลกอริทึม Tree J48, Random Forest, Multilayer Perception, Adaboost และ Support Vector Machine ถึงแม้ว่า Random Forest จะให้ผลลัพธ์ดีกว่าแต่ผู้วิจัยจึงเลือก Tree J48 เพราะง่ายต่อการสร้างต้นไม้ตัดสินใจเพื่อวิเคราะห์ผลในครั้งนี้ โดยแบบจำลองมีความถูกต้อง 80-82%

ค.ศ. 2014 งานวิจัยของ Maryam[4] และคณะได้มีการใช้งาน Hashtags จำแนกอารมณ์ของบุคคลที่ใช้งานโซเชียลมีเดียอย่าง Twitter โดยจะมุ่งการศึกษาไปการจำแนกอารมณ์ซึมเศร้า ในการทดลองได้เปรียบเทียบอัลกอริทึม SVM และ KNN โดยใช้กับข้อมูลที่เก็บจาก Hashtags ที่บ่งบอกถึงอารมณ์ซึมเศร้าและอารมณ์ที่บ่งบอกว่ามีความสุข ผลปรากฏว่า SVM มีประสิทธิภาพที่ดีที่สุด โดยมีความถูกต้องถึง 90%

ค.ศ. 2015 งานวิจัยของ McManus[5] และคณะได้มีการใช้เทคนิคการทำเหมืองข้อมูลในการวิเคราะห์หาผู้ป่วยจิตเวทด้วยข้อมูลจากโซเชียลมีเดียอย่าง Twitter ด้วยอัลกอริทึม SVM, Neural Network และ Naïve Bayes โดยกระทำกับข้อมูลโปรไฟล์ของผู้เข้ารับการทดลองอย่างจำนวนเพื่อน ช่วงเวลาในการทวีต คำที่เกี่ยวข้องกับจิตเวทและอีโมติคอน ผลปรากฏว่า SVM มีประสิทธิภาพที่ดีที่สุด โดยมีความถูกต้องถึง 92.3%

ค.ศ. 2017 Anees[6] และคณะ ได้มีการทำเหมืองข้อมูลเพื่อหาภาวะซึมเศร้าจากข้อมูลโซเชียลมีเดีย โดยจำแนกเฉพาะข้อความที่เป็นภาวะซึมเศร้าและไม่เป็นภาวะซึมเศร้าด้วยใช้อัลกอริทึม SVM Bayes และ Maximum Entropy ผลปรากฏว่า SVM มีประสิทธิภาพที่ดีที่สุด โดยมีความถูกต้องถึง 91%

ดังนั้นงานวิจัยนี้จะนำเสนอวิธีการทำเหมืองข้อมูลโดยนำความคิดเห็นจากผู้ใช้งาน Twitter มาทำการจำแนกโรคซึมเศร้าจากเหมืองข้อมูลโดยมีการเปรียบเทียบการจำแนกระหว่างแบบหนึ่งระดับและแบบสองระดับ ได้แก่ การจำแนกหนึ่งระดับจะใช้งานอัลกอริทึม Bayes และอัลกอริทึม SVM จำแนกข้อความทั่วไปและข้อความที่บ่งบอกถึงลักษณะอาการซึมเศร้าตามแบบสอบถาม DSM-5 ได้แก่ 1. อารมณ์ซึมเศร้า 2. ความสนใจลดลง 3. น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต 4. นอนไม่หลับหรือนอนหลับมากกว่าปกติ 5. ร่างกายอ่อนเพลีย 6. รู้สึกตนเองไร้ค่า 7. สมาธิสั้น 8. เคลื่อนไหวช้า และ 9. คิดฆ่าตัวตาย ส่วนการจำแนกสองระดับจะใช้งานอัลกอริทึม SVM เพื่อจำแนกข้อความทั่วไปกับข้อความที่บ่งบอกถึงโรคซึมเศร้า ต่อมาใช้งานอัลกอริทึม Bayes เปรียบเทียบกับอัลกอริทึม Random Forest เพื่อจำแนก 9 อาการที่บ่งบอกถึงโรคซึมเศร้าตามแบบสอบถาม DSM-5 ทั้งหมดเพื่อให้สามารถเพิ่มความถูกต้องในการจำแนกผู้ที่เข้าข่ายโรคซึมเศร้าได้อย่างถูกต้องและแม่นยำมากขึ้น

งานที่เกี่ยวข้อง

1. โรคซึมเศร้า

โรคซึมเศร้า[7] คือ โรคทางจิตเวชชนิดหนึ่งที่เกิดจากสารเคมีเซโรโทนินในสมองมีปริมาณลดลงจนเสียสมดุลส่งผลทำให้ร่างกายและจิตใจเกิดภาวะซึมเศร้า เบื่อหน่าย ท้อแท้ นอนไม่หลับ หรือ อยากฆ่าตัวตาย เป็นต้น ซึ่งอาการป่วยจะเป็นอย่างต่อเนื่องมากกว่า 2 สัปดาห์ขึ้นไป

Table 1 DSM-5 criteria's

No.	Symptoms	Non	Someday	Frequently	Everyday
1	Depressed mood	0	1	2	3
2	Diminished interest	0	1	2	3
3	Change in appetite	0	1	2	3
4	Change in sleep	0	1	2	3
5	Slowed thinking	0	1	2	3
6	Worthlessness or guilt	0	1	2	3
7	Fatigue	0	1	2	3
8	Agitation or retardation	0	1	2	3
9	Suicidal ideation	0	1	2	3

Table 2 Training set

No.	Symptoms	Hashtag	Examples of Training set	Number of messages
1	Depressed mood	#Sadness #Depressive	I'm in such a depressive spiral today.	3,000
2	Diminished interest	#Loss of Interest #Lose interest	Loss of interest in friends, family & favorite activities.	3,000
3	Change in appetite	#Appetite #Hunger	I'm hungry but I have no appetite can anyone relate.	3,000
4	Change in sleep	#Sleep #Lethargy	I just want to sleep for real!	3,000
5	Slowed thinking	#Un thinking #Out thinking	These neighbors really out here thinking that I'm selling dope.	3,000
6	Worthlessness or guilt	#Guilt #Disgrace #Dishonor	This guy is a perpetual disgrace.	3,000
7	Fatigue	#Tired #Bored #Fatigued	I'm just tired that's all.	3,000
8	Agitation or retardation	#Lackadaisical #Lazy #Loafing #Phlegmatic	Feeling lazy of today.	3,000
9	Suicidal ideation	#Suicidal #Dangerous #Destructive	Suicidal thoughts will never get out of my head.	3,000
10	Normal	#Happy	Happy birthday, family.	3,000

ในการวินิจฉัยสามารถวินิจฉัยโดยใช้เกณฑ์การวินิจฉัยของสมาคมจิตแพทย์อเมริกัน โดยกำหนดไว้ในหนังสือ The 5th Diagnostic and Statistical Manual of Mental Disorders (DSM-5)[8] มีหลักเกณฑ์ว่าผู้ที่เป็นโรคซึมเศร้าจะต้องมีลักษณะอาการอย่างน้อย 5 อาการหรือมากกว่า ได้แก่ 1. มีอารมณ์ซึมเศร้า 2. ความสนใจหรือความสนุกสนานในการทำกิจกรรมต่างๆ ที่เคยทำทั้งหมดลดลงอย่างมาก 3. น้ำหนักลดลงอย่างชัดเจน 4. มีอาการนอนไม่หลับหรือหลับนานหลับบ่อยกว่าปกติ 5. การเคลื่อนไหวช้าลง 6. อ่อนเพลียไม่มีเรี่ยวแรง 7. รู้สึกว่าตนเองไร้ค่าหรือรู้สึกว่าตนเองผิดโดยไม่มีสาเหตุ 8. สมาธิสั้นหรือความสามารถในตัดสินใจลดลง 9. มีความคิดอยากฆ่าตัวตาย โดยเกิดขึ้นติดต่อกันนานไม่ต่ำกว่า 2 สัปดาห์ ซึ่งจัดทำเป็นแบบสอบถามไว้ดัง Table 1

จาก Table 1 ในการวิเคราะห์จะให้ผู้ป่วยทำแบบสอบถามดังกล่าว Someday คือ มีอาการนั้นใน 2-4 วันในระยะเวลา 2 สัปดาห์ Frequently คือ มีอาการนั้นใน 6-8 วันในระยะเวลา 2 สัปดาห์และ Everyday คือ มีอาการนั้นใน 10-14 วัน ในระยะเวลา 2 สัปดาห์ แล้วจึงนำไปหาผลรวมของคะแนนในแบบประเมินของการวินิจฉัยโรคซึมเศร้า ในการสรุปผลสามารถนำคะแนนจากแบบประเมินไปประเมินได้ว่าถ้าหากน้อยกว่า 7 คะแนน แสดงว่าปกติ ถ้าอยู่ระหว่าง 8-12 คะแนน แสดงว่ามีอาการน้อย ถ้าอยู่ระหว่าง 13-18 แสดงว่ามีอาการปานกลาง ถ้ามากกว่า 19 แสดงว่ามีอาการรุนแรง

2.การคัดเลือกคุณลักษณะด้วย Information Gain

Information Gain[9] เป็นการคัดเลือกคุณลักษณะ (Feature) สำหรับใช้ในการสร้างแบบจำลอง เนื่องมาจากคุณลักษณะในข้อมูลนั้นมีจำนวนมากเกินไป ทำให้เวลาในการสร้างแบบจำลองและการทดสอบแบบจำลองนั้นล่าช้าและอาจจะส่งผลทำให้ค่าความถูกต้องในการทำนายนั้นลดลง ส่วนสมการ Information Gain มีดังสมการที่ (1) เมื่อ S คือ ตัวอย่างที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลายๆ กรณี *E* คือ เอนโทรปีของตัวอย่าง *A* คือ ตัวแปรต้นที่พิจารณา *V = value(A)* คือ เซตของค่าของ A ที่เป็นไปได้และ S_v คือ ตัวอย่างที่ A มีค่า *V* ทั้งหมด โดยที่ Entropy หาได้จากสมการที่ (2) เมื่อ *S* คือ ตัวอย่างที่ประกอบด้วยชุดของตัวแปรต้นและตัวแปรตามหลายๆ กรณีและ *Ps(J)* คือ อัตราส่วนของกรณีใน *s* ที่ตัวแปรตามหรือผลลัพธ์มีค่า *j*

$$Gain(S,A) = E(S) - \sum_{v=value(A)} \frac{|S_v|}{|S|} E(S_v) \quad (1)$$

$$E(S) = -\sum_{j=i}^{n} p_s(j) log_2 p_s(j) \quad (2)$$

3. อัลกอริทึม Support Vector Machine (SVM)

เทคนิค SVM[10] เป็นเทคนิคอัลกอริทึมประเภท Supervised Learning Algorithm ที่คิดค้นโดย Vladimir N. Vapnik และ Alexey Ya. Chervonenkis ในปี 1963 โดยใช้หลักการสร้าง Hyperplane ที่เป็นเส้นตรงขึ้นมาดังสมการที่ (3) เมื่อ W^T คือ ความชันของเส้นตรง b คือ จุดตัดแกน y *g(x)* คือ พิกัดแกน y และ x คือ พิกัดแกน x เพื่อแบ่งกลุ่มของข้อมูลออกจากกันและคำนวณหาเส้นตรงเส้นใดที่ดีที่สุด โดย SVM มีข้อดีที่ไม่ค่อยเกิดปัญหา Overfitting มากเหมือนกับ Neural Network

$$g(x) = w^T x + b \quad (3)$$

4. อัลกอริทึม Naïve Bayes

เทคนิค Naïve Bayes[11] เป็นเทคนิคอัลกอริทึมประเภท Supervised Learning Algorithm โดยใช้หลักการความน่าจะเป็นแบบมีเงื่อนไขที่คิดค้นโดย Theorem Bayes เข้ามาพัฒนาทฤษฎีดังกล่าว สมมติฐานของสมการที่กำหนดให้การเกิดของเหตุการณ์ต่างๆ มีอิสระต่อกัน (Independence) ซึ่งการใช้งาน Bayes มีการใช้งานอย่างแพร่หลายในงานด้าน Machine Learning เช่น sentiment analysis มีเหตุผลเนื่องจาก Bayes มีการทำงานที่ไม่ซับซ้อนแต่ให้ประสิทธิภาพที่สูง ใช้เวลาในการสร้างโมเดลไวกว่าอัลกอริทึมอื่นๆ ดังสมการที่ (4) เมื่อ *p(c|x)* คือ Posterior probability เป็นค่าความน่าจะเป็นของข้อมูลที่มีแอตทริบิวต์เป็น x จะเป็นคลาส *c P(c)* คือ Prior probability เป็นค่าความน่าจะเป็นของคลาส คือ ค่าความน่าจะเป็นที่ข้อมูลเป็นคลาส c มีแอตทริบิวต์ *x* และ *P(x)* คือ Predictor Prior probability เป็นจำนวนที่มีมีแอตทริบิวต์ *x* ทั้งหมดในข้อมูล

$$p(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (4)$$

5. อัลกอริทึม Random Forest

เทคนิค Random Forest[12] เป็นเทคนิคอัลกอริทึมประเภท Supervised Learning Algorithm ที่คิดค้นโดย Ho ในปี 1995 โดยใช้หลักการสร้างโมเดลด้วย Decision Tree หลายๆ ต้น โดยในแต่ละ Decision Tree จะได้ Data และ Feature แบบ Random ไปเพื่อสร้างโมเดล หลังจากนั้นเมื่อจำแนกผลก็จะนำผลลัพธ์ที่ได้มาโหวตหาค่าที่มากที่สุด ข้อดีของ Random Forest คือ ระยะเวลาในการจำแนกผลลัพธ์ที่สั้นเนื่องจากเป็น Decision Tree ที่ข้างในประกอบไปด้วยเงื่อนไข If-Else แต่มีระยะเวลาในการสร้างแบบจำลองที่นานถ้าหาก

เลือกจำนวน Decision Tree เยอะขึ้นและถ้าปรับพารามิเตอร์ Maximal depth ที่ซึ่งหมายถึงระดับความลึกของ Decision Tree แต่ละต้น ระยะเวลาสร้างแบบจำลองก็จะมากขึ้นตามลำดับ

วิธีการดำเนินงานวิจัย

ในการดำเนินงานวิจัยประกอบไปด้วย 4 ขั้นตอน ดัง Figure 1

1. Data collection

ในการรวบรวมข้อมูลผู้วิจัยจะใช้ข้อความจากการโพสต์บน Twitter ของผู้ใช้งานทั่วไป โดยการเก็บข้อมูลผ่าน Twitter API บน RapidMiner Studio โดยในงานวิจัยจะแบ่งข้อมูล (Data set) ออกเป็น 2 ส่วน ได้แก่ Training set และ Test set

Training set จะเป็นข้อมูลสำหรับสร้างแบบจำลอง ซึ่งเป็นข้อความที่รวบรวมจาก Twitter

โดยกำหนดจาก Hashtag ที่เกี่ยวกับ 9 ลักษณะอาการที่เกี่ยวข้องกับโรคซึมเศร้าดัง Table 2 จาก Table 2 ในการใช้งาน Training set จะติด Label โดยการกำหนดอาการทั้ง 9 คือ ลำดับที่ 1-9 เป็น Negative คือ เป็นข้อความที่เข้าข่ายเป็นโรคซึมเศร้าและลำดับที่ 10 เป็น Positive คือ ข้อความที่ไม่เข้าข่ายเป็นโลกซึมเศร้า มีตัวอย่างข้อความดัง Table 2

Test set จะเป็นข้อมูลสำหรับทดสอบแบบจำลอง ซึ่งเป็นข้อความภาษาอังกฤษจาก Twitter ที่รวบรวมจากผู้ใช้งานที่เป็นดาราต่างประเทศที่ป่วยเป็นโรคซึมเศร้าจำนวน 15 คน และผู้ใช้งานที่เป็นดาราต่างประเทศแต่ไม่เป็นโรคซึมเศร้า 15 คน รวมทั้งหมด 30 คน โดยผู้ใช้งานแต่ละคนมีการโพสต์ข้อความมากกว่า 2 สัปดาห์ขึ้นไป

2. Data Preprocessing

Data Preprocessing เป็นขั้นตอนในการเตรียมข้อมูลก่อนนำเข้าอัลกอริทึม โดยใช้กับข้อมูล Train set และ Test set โดยมีวิธีการดังต่อไปนี้ Regular Expression เป็นขั้นตอนในการกรองข้อความที่ไม่จำเป็นบางส่วนออกโดยใช้งาน Regular expression ผู้วิจัยเลือกกรองข้อความที่เป็นข้อความ Retweet โดยกำหนดค่าฟังก์ชันคือ “RT(.*)” กรองข้อความที่เป็นลิงค์เข้าใช้งานเว็บไซต์โดยกำหนดค่าฟังก์ชันคือ “(https?|http)://[-a-zA-Z0-9+&@#/%?=~_|!:,. ;]*[-a-zA-Z0-9+&@#/%=~_|]” และกรองข้อความที่เป็นชื่อคนภายในโพสต์ โดยกำหนดค่าฟังก์ชันคือ “(@)[-a-zA-Z0-9+&@#/%?=~_|!:,. ;]*[-a-zA-Z0-9+&@#/%=~_|:]”

Word Segmentation เป็นการทำข้อมูล Training set และ Test set เข้าไปขั้นตอนการตัดคำออกมาเก็บในถุงคำ (Bag of word)

Transform Cases เป็นการนำคำศัพท์ใน Bag of word มาแปลงเป็น Lower case เพื่อลดความหลากหลายของคำศัพท์

Filter Stop words เป็นการนำคำศัพท์มาตัดคำหยุดของภาษาอังกฤษ (Stop words) [13] เพื่อลดคำที่ไม่เกี่ยวข้องออกไปจาก Bog of word

Weighting เป็นขั้นตอนในนำคำใน Bag of word มานับคำในประโยคของ Train set และ Test set โดยให้อยู่ในรูปแบบ Numeric data เพื่อให้สามารถเข้าอัลกอริทึมคำนวณได้ โดยใช้วิธีการ Binary term occurrence ซึ่งเป็นการนับความถี่ในประโยค โดยคำที่ตรวจพบจะให้ค่าน้ำหนักเป็น 1 และคำที่ไม่พบจะให้ค่าน้ำหนักเป็น 0

Features Selection คือการลดจำนวนของคุณลักษณะ โดยใช้วิธีการ Information Gain ผู้วิจัยทำการเลือกคุณลักษณะโดยใช้เกณฑ์ Top-K=2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะและคุณลักษณะทั้งหมด

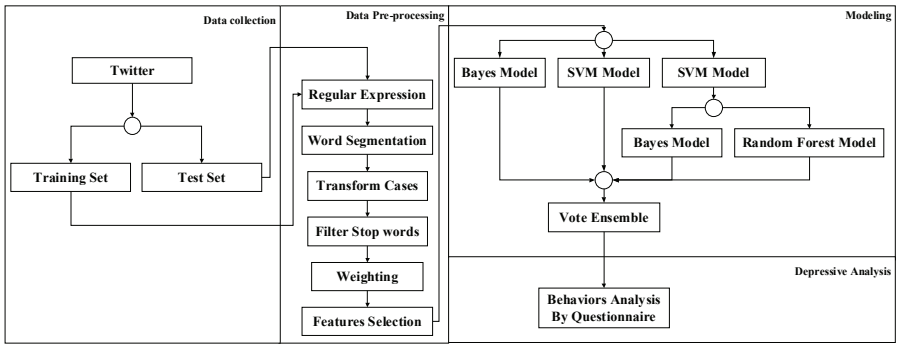

Figure 1 The proposed diagram

Table 3 Top 20 feature frequency

No.	Feature	No.	Feature	No.	Feature	Feature	No.	Feature
1	things	6	tiredness	11.	tired	tired	16.	people
2	loss	7	esteem	12.	sleep	sleep	17.	bored
3	feeling	8	hopeless	13.	appetite	appetite	18.	sadness
4	self	9	happy	14.	get	get	19.	bra

3. Modelling

ในการสร้างแบบจำลองใช้งาน K-Fold Cross Validation โดยการกำหนดให้ K=10 และจะแบ่งเป็นการจำแนกเป็น 2 ระดับ ดังต่อไปนี้

จำแนกหนึ่งระดับ คือ ใช้อัลกอริทึม Bayes และอัลกอริทึม SVM ในการจำแนก Training set และ Test set โดยจำแนกข้อความทั่วไปและข้อความที่บ่งบอกถึงลักษณะอาการซึมเศร้าตามแบบสอบถาม DSM-5 ได้แก่ 1. อารมณ์ซึมเศร้า 2. ความสนใจลดลง 3. น้ำหนักลดลงหรือเพิ่มขึ้นอย่างผิดสังเกต 4. นอนไม่หลับหรือนอนหลับมากกว่าปกติ 5. ร่างกายอ่อนเพลีย 6. รู้สึกตนเองไร้ค่า 7. สมาธิสั้น 8. เคลื่อนไหวช้า และ 9. คิดฆ่าตัวตาย

จำแนกสองระดับ คือ การใช้งานอัลกอริทึมซ้อนกัน 2 อัลกอริทึมโดยจะแบ่งเป็น 2 ขั้น ดังต่อไปนี้ ขั้นที่ 1 ใช้งานอัลกอริทึม SVM ในการจำแนก Training set และ Test set โดยที่จะจำแนกข้อความที่เข้าข่ายเป็นโรคซึมเศร้าและไม่เข้าข่ายเป็นโรคซึมเศร้าเท่านั้น

ขั้นที่ 2 ใช้งานอัลกอริทึม Bayes เปรียบเทียบกับอัลกอริทึม Random Forest โดยจำแนกข้อความที่เกี่ยวข้องกับโรคซึมเศร้าออกเป็น 9 ลักษณะอาการ (อัลกอริทึม Random Forest กำหนดค่าตัวแปรดังนี้ Criterion ใช้ Gain ratio, Number of trees=100 และ Maximal=100)

สุดท้ายผลลัพธ์แต่ละ Class จะได้ค่าความน่าจะเป็น (Probability) ของแต่ละอาการที่เข้าข่ายซึมเศร้า ผู้วิจัยจึงได้ทำการเลือกค่า Maximum Probability คือ การเลือกค่าความน่าจะเป็นที่มีค่าสูงที่สุด โดยที่ข้อมูล 1 Instant ผ่านแบบจำลองทั้ง 2 แบบจำลอง จะได้ค่าความน่าจะเป็นของ Class คำตอบคือ 9 Class แล้วนำค่าความน่าจะเป็นของ 9 Class มาทำการโหวตเลือกค่าความน่าจะเป็นสูงสุด (Maximum Probability) เพื่อเป็นคำตอบของการทำนายครั้งนั้น การวัดประสิทธิภาพของแบบจำลองในงานวิจัยนี้ใช้ค่าสถิติที่ใช้ในการวัดประสิทธิภาพของการการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์โดยใช้ค่า Accuracy Precision Recall และ F-1

4. Depressive Analysis

หลังจากแบบจำลองสามารถจำแนกข้อมูลที่อยู่ในลักษณะอาการ 9 อาการได้อย่างแม่นยำแล้ว ในการวิเคราะห์ผู้ที่เข้าข่ายเป็นโรคซึมเศร้านั้น สามารถทำได้โดยการนับความถี่ของอาการแต่ละอาการในระยะเวลา 2 สัปดาห์ โดยจะต้องขยับไปทุกๆ 1 วัน แล้วคำนวณความถี่ใหม่จนกว่าจะครบข้อมูลที่ทำนายของบุคคลนั้น

ผลการทดลอง

1. Features Selection

ในการสร้างแบบจำลองผู้วิจัยเลือกใช้งาน Information Gain ในการคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะ และคุณลักษณะทั้งหมด มีตัวอย่าง 20 คุณลักษณะแรกดังแสดงใน Table 3

Table 4 Performance modeling SVM algorithm (2 Levels)

Feature	Accuracy	Precision Yes	Precision No	Recall Yes	Recall No	F-1
2,000	98.03%	99.27%	86.00%	98.60%	92.18%	98.93%
4,000	98.11%	99.27%	96.52%	98.65%	92.29%	98.96%
6,000	98.20%	99.27%	87.50%	98.75%	92.37%	99.01%
All	94.29%	99.82%	39.32%	94.24%	95.57%	96.95%

Table 5 Algorithm performance modeling Bayes, SVM + Bayes and SVM + Random Forest

Feature	1 Level				2 Levels			
	Bayes		SVM		SVM + Bayes		SVM + Random Forest	
	Accuracy	F-1	Accuracy	F-1	Accuracy	F-1	Accuracy	F-1
2,000	82.55%	80%	96.18%	96%	83.23%	86%	91.45%	94%
4,000	78.58%	75%	94.64%	94%	78.81%	80%	88.31%	93%
6,000	76.66%	74%	92.28%	92%	79.00%	79%	82.77%	88%
All	70.52%	67%	91.76%	92%	72.34%	72%	63.82%	73%

Table 6 Algorithm performance modeling Bayes, SVM + Bayes and SVM + Random Forest 2,000 features

Class	1 Level				2 Levels			
	Bayes		SVM		SVM + Bayes		SVM + Random Forest	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Depressive	83.86%	80.67%	98.48%	95.20%	83.32%	87.92%	90.39%	97.11%
Loss of interest	98.57%	86.21%	99.96%	92.23%	86.28%	98.67%	98.79%	97.82%
Appetite	72.85%	89.39%	99.14%	96.37%	89.18%	74.68%	95.31%	84.99%
Sleep	79.36%	76.20%	97.94%	93.63%	75.59%	79.31%	80.73%	95.46%
Thinking	82.40%	69.29%	98.20%	96.60%	70.00%	84.77%	87.58%	96.28%
Guilt	83.36%	88.95%	95.55%	88.00%	89.12%	82.45%	93.84%	96.79%
Tired	75.72%	86.54%	98.63%	96.30%	86.55%	76.10%	91.72%	83.72%
Movement	86.29%	85.54%	98.87%	90.47%	85.02%	88.40%	92.92%	82.64%
Suicidal	76.59%	78.24%	99.21%	91.97%	77.81%	78.54%	87.35%	97.15%
Normal	88.66%	77.34%	-	-	-	-	-	-

Table 7 Algorithm performance modeling Bayes, SVM + Bayes and SVM + Random Forest 4,000 features

Class	1 Level				2 Levels			
	Bayes		SVM		SVM + Bayes		SVM + Random Forest	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Depressive	77.23%	73.81%	98.53%	91.47%	76.55%	81.14%	90.96%	95.13%
Loss of interest	98.57%	86.70%	99.96%	90.37%	86.82%	98.12%	97.94%	96.81%
Appetite	72.79%	86.04%	99.18%	93.27%	86.87%	74.77%	90.01%	85.04%
Sleep	72.76%	70.61%	98.54%	92.10%	70.53%	72.57%	79.11%	85.17%
Thinking	70.78%	57.04%	98.47%	94.37%	56.76%	71.65%	86.78%	94.56%
Guilt	77.11%	85.54%	99.52%	83.23%	85.52%	76.49%	84.83%	94.59%
Tired	70.89%	80.09%	98.45%	95.57%	78.92%	72.08%	89.75%	77.65%
Movement	79.35%	81.18%	98.99%	85.10%	80.09%	81.48%	88.13%	79.88%
Suicidal	68.82%	73.07%	99.09%	86.70%	71.76%	70.83%	83.01%	95.45%
Normal	84.17%	70.64%			-	-	-	-

Table 8 Algorithm performance modeling Bayes, SVM + Bayes and SVM + Random Forest 6,000 features

Class	1 Level				2 Levels			
	Bayes		SVM		SVM + Bayes		SVM + Random Forest	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Depressive	76.60%	72.35%	98.12%	85.33%	77.26%	83.45%	83.05%	94.23%
Loss of interest	98.33%	87.19%	99.96%	89.07%	86.82%	98.37%	97.62%	96.46%
Appetite	72.94%	85.05%	99.33%	79.30%	86.04%	73.64%	78.48%	69.40%
Sleep	71.51%	69.71%	97.80%	91.80%	70.53%	72.53%	71.37%	91.93%
Thinking	66.41%	55.85%	97.47%	91.17%	58.10%	74.72%	82.53%	92.01%
Guilt	76.00%	84.92%	99.79%	79.27%	85.70%	77.77%	85.39%	90.69%
Tired	69.91%	78.39%	96.63%	94.60%	80.79%	72.10%	86.23%	64.00%
Movement	78.70%	80.36%	98.54%	78.90%	80.54%	83.07%	77.62%	75.70%
Suicidal	68.39%	72.61%	98.94%	83.80%	72.54%	69.52%	78.12%	90.64%
Normal	82.72%	68.56%			-	-	-	-

Table 9 Algorithm performance modeling Bayes, SVM + Bayes and SVM + Random Forest all features

Class	1 Level				2 Levels			
	Bayes		SVM		SVM + Bayes		SVM + Random Forest	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Depressive	69.48%	66.25%	98.00%	85.10%	77.26%	83.45%	83.05%	94.23%
Loss of interest	98.13%	87.01%	99.96%	89.07%	86.82%	98.37%	97.62%	96.46%
Appetite	68.99%	79.88%	99.08%	86.10%	86.04%	73.64%	78.48%	69.40%
Sleep	63.20%	62.60%	97.43%	89.77%	70.53%	72.53%	71.37%	91.93%
Thinking	54.82%	42.96%	96.79%	87.57%	58.10%	74.72%	82.53%	92.01%
Guilt	70.07%	78.64%	99.58%	78.43%	85.70%	77.77%	85.39%	90.69%
Tired	63.28%	72.30%	95.73%	94.07%	80.79%	72.10%	86.23%	64.00%
Movement	70.19%	73.28%	97.59%	74.33%	80.54%	83.07%	77.62%	75.70%
Suicidal	60.68%	63.58%	98.53%	82.80%	72.54%	69.52%	78.12%	90.64%
Normal	77.72%	61.78%			-	-	-	-

2. Performance of Training Set

จากการสร้างแบบจำลองที่ใช้งาน Information Gain ในการคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะและคุณลักษณะทั้งหมด การจำแนกหนึ่งระดับมีประสิทธิภาพดัง Table 5–9

จากผลการทดลองพบว่าแบบจำลองที่ถูกสร้างด้วยการคัดเลือกคุณลักษณะ 2,000 คุณลักษณะมีค่าความถูกต้องดีที่สุด โดยอัลกอริทึม Bayes ได้ Accuracy=82.55% และอัลกอริทึม SVM ได้ Accuracy=96.18% การจำแนกแบบสองระดับ ขั้นที่ 1 อัลกอริทึม SVM ได้ประสิทธิภาพดัง Table 4 จากผลการทดลองพบว่าผลลัพธ์ทั้งหมดของแบบจำลองที่ถูกสร้างด้วยการคัดเลือกคุณลักษณะ 6,000 คุณลักษณะได้ค่าความถูกต้องมากที่สุด โดยได้ค่า Accuracy=98.20% และขั้นที่ 2 อัลกอริทึม Bayes และอัลกอริทึม Random Forest มีประสิทธิภาพดัง Table 5–9 จากผลการทดลองพบว่าแบบจำลองที่ถูกสร้างด้วยการคัดเลือกคุณลักษณะ 2,000 ได้ค่าความถูกต้องมากที่สุด โดยอัลกอริทึม Random Forest ได้ค่า Accuracy คือ 83.23% และอัลกอริทึม Random Forest ได้ค่า Accuracy คือ 91.45%

3. Performance of Test Set

สำหรับการวัดความถูกต้องของข้อมูลทดสอบแบบจำลอง ซึ่งเป็นข้อมูลของผู้ที่เป็นโรคซึมเศร้าจำนวน 15 คน และคนที่ไม่เป็นโรคซึมเศร้าจำนวน 15 คน โดยทดสอบกับแบบจำลองที่เลือกใช้งาน Information gain ในการคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะและคุณลักษณะทั้งหมด และมีการกำหนดค่า Boundary ออกเป็น 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 และ 0.9 ซึ่งเป็นการกำหนดเพดานของค่าความน่าจะเป็น ถ้าข้อความไหนมีค่าไม่ถึงค่าเพดานก็จะไม่นำมาทำการ Vote ensemble โดยมีผลการทดลองการจำแนกที่แบ่งออกเป็นสองระดับ คือ

การจำแนกหนึ่งระดับมีผลการทดลองดัง Figure 2–5 จากผลการทดลองทั้งหมดพบว่าค่าความน่าจะเป็นที่เหมาะสมแก่การเป็นค่า Boundary ของ อัลกอริทึม Bayes อยู่ที่ 0.4 โดยได้ค่า Accuracy สูงสุด=76.67% และค่า Boundary ของอัลกอริทึม SVM อยู่ที่ 0.5 โดยได้ค่า Accuracy สูงสุด คือ 70.00%

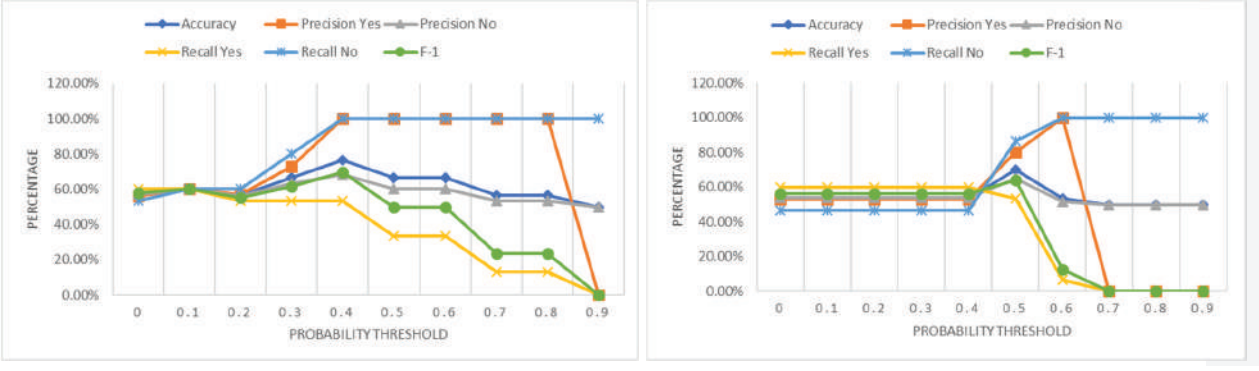

Figure 2 Performance Bayes model (Left) and SVM model (Right) by 2,000 features (1 Level)

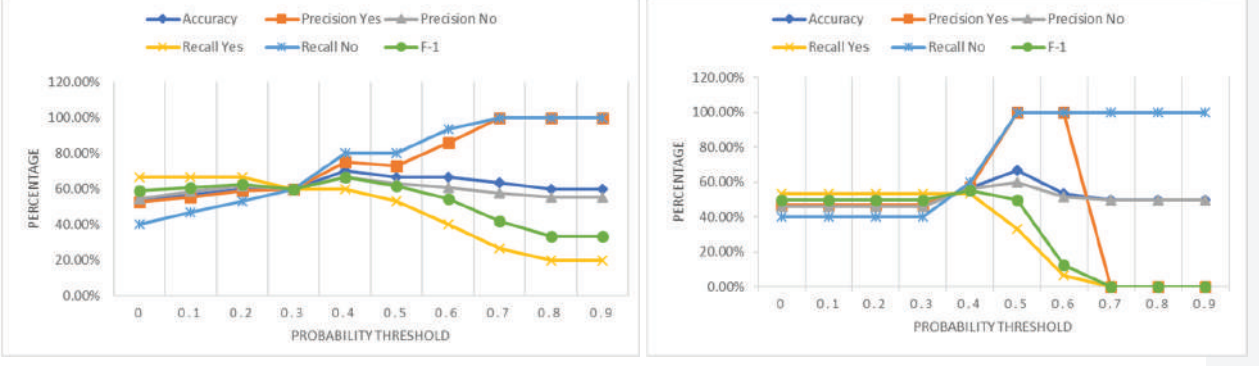

Figure 3 Performance Bayes model (Left) and SVM model (Right) by 4,000 features (1 Level)

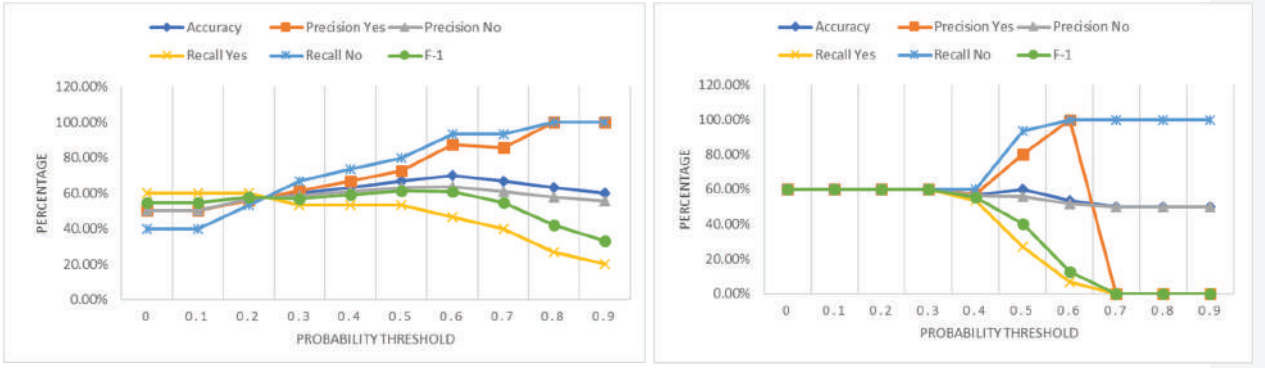

Figure 4 Performance Bayes model (Left) and SVM model (Right) by 6,000 features (1 Level)

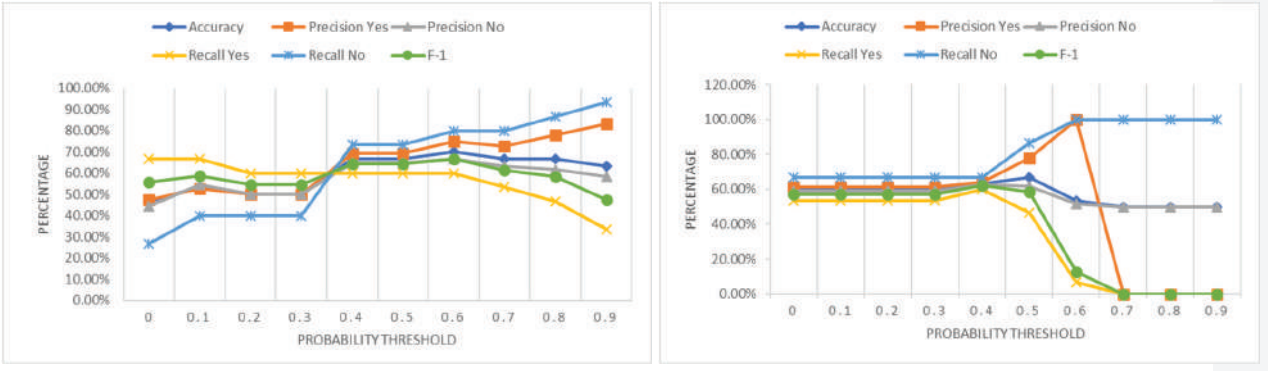

Figure 5 Performance Bayes model (Left) and SVM model (Right) by all features (1 Level)

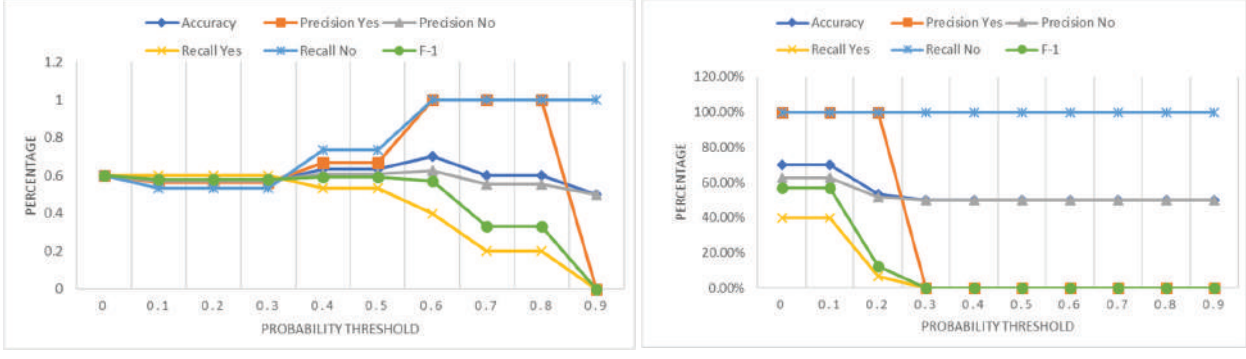

Figure 6 Performance SVM + Bayes model (Left) and SVM + Random Forest model (Right) by 2,000 features (2 Levels)

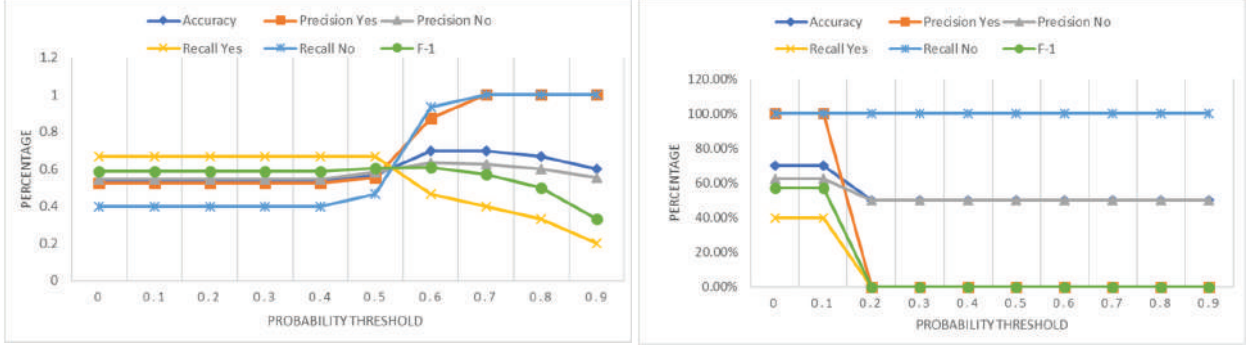

Figure 7 Performance SVM + Bayes model (Left) and SVM + Random Forest model (Right) by 4,000 features (2 Levels)

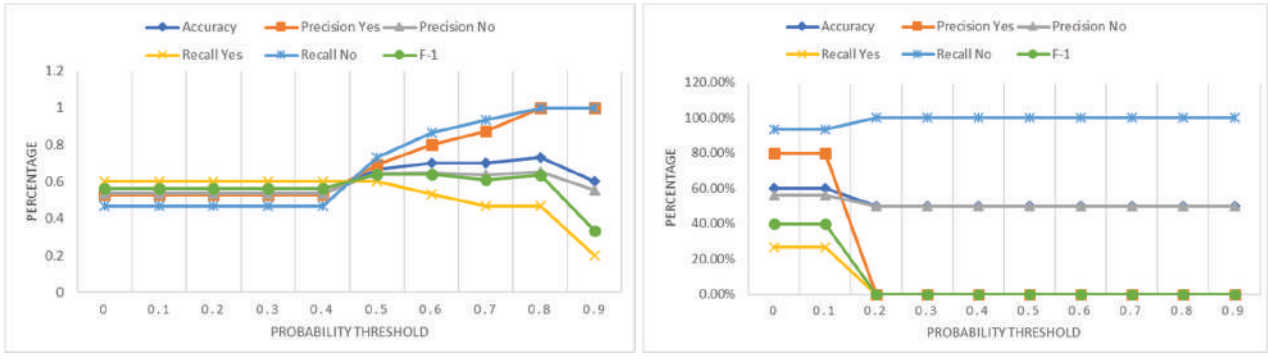

Figure 8 Performance SVM + Bayes model (Left) and SVM + Random Forest model (Right) by 6,000 features (2 Levels)

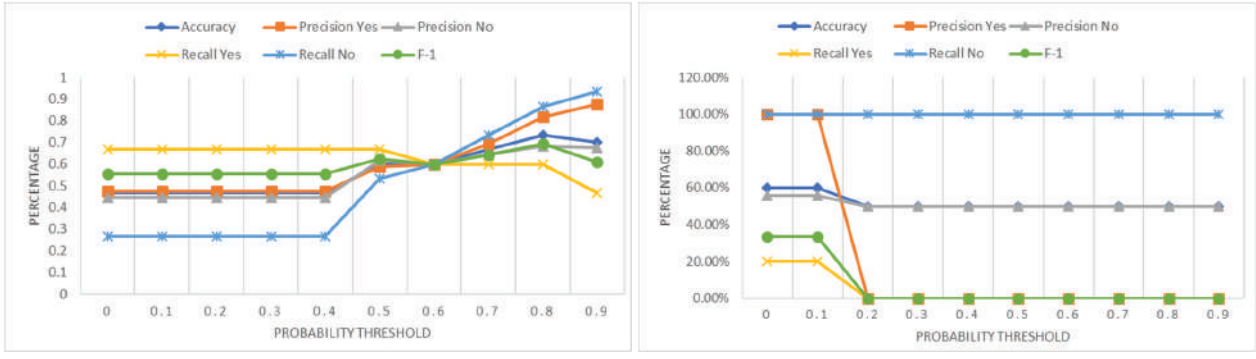

Figure 9 Performance SVM + Bayes model (Left) and SVM + Random Forest model (Right) by all features (2 Levels)

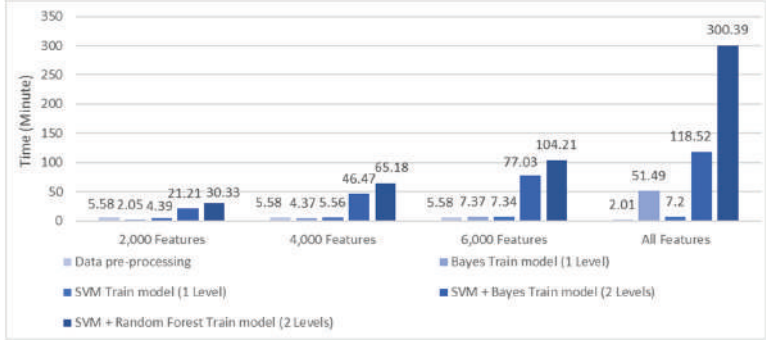

Figure 10 Performance time of train model

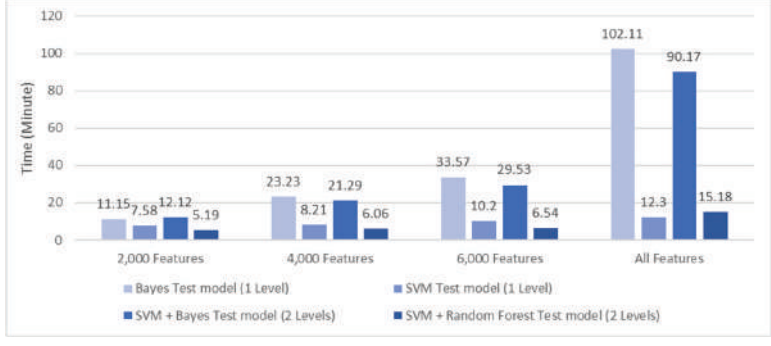

Figure 11 Performance time of test model

การจำแนกสองระดับมีผลการทดลองดัง Figure 6-9 จากการทดลองทั้งหมดพบว่าค่าความน่าจะเป็นที่เหมาะสมแก่การเป็นค่า Boundary ของอัลกอริทึม SVM จับคู่กับอัลกอริทึม Bayes อยู่ที่ 0.6 โดยได้ค่า Accuracy สูงสุด คือ 73.33% และอัลกอริทึม SVM จับคู่กับอัลกอริทึม Random Forest อยู่ที่ 0.1 โดยได้ค่า Accuracy สูงสุด คือ 70.00%

จากการทดลองทั้งหมดพบว่าค่าความน่าจะเป็นที่สูงสามารถกรองข้อความที่ไม่เข้าข่ายในอาการที่เกี่ยวข้องกับโรคซึมเศร้าออกไป ทำให้จำการจำแนกข้อความตามอาการได้ดีและการคัดเลือกคุณลักษณะเพื่อสร้างแบบจำลองให้ผลลัพธ์ที่ดีกว่าการใช้งานการเลือกคุณลักษณะทั้งหมด เนื่องจากข้อมูลหลังจากการทำกระบวนการเตรียมข้อมูล คุณลักษณะมีมากเกินไป ทำให้การเลือกใช้งานการคัดเลือกคุณลักษณะ ให้ผลที่ได้กว่าการเลือกคุณลักษณะทั้งหมด

4. Time (Performance)

ประสิทธิภาพของการใช้เวลาในการทดลองจะแบ่งออกเป็น 2 การทดลอง ได้แก่

1. เวลาในการสร้างแบบจำลอง โดยจะแบ่งเป็น 2 ช่วง ได้แก่ 1. ช่วงเวลาในการเตรียมข้อมูลชุด Training set โดยการทดลองจะเริ่มตั้งแต่การคัดเลือกคุณลักษณะ โดยกำหนด Top-K คือ 2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะและคุณลักษณะทั้งหมดจนถึงการสร้างแบบจำลอง

2. ช่วงเวลาในการสร้างแบบจำลอง โดยการสร้างแบบจำลองจะมีการจำแนกหนึ่งระดับ คือ การใช้งานอัลกอริทึม Bayes และอัลกอริทึม SVM การจำแนกสองระดับ คือ การใช้งานอัลกอริทึม SVM คู่กับอัลกอริทึม Bayes เปรียบเทียบกับการใช้งานอัลกอริทึม SVM คู่กับกับอัลกอริทึม Random Forest ผลการทดลองดัง Figure 10 จากการทดลองพบว่าการสร้างแบบจำลองโดยการเลือก 2,000 คุณลักษณะได้เวลาน้อยที่สุด คือ 2.05 นาที เนื่องจากในขั้นตอนการเลือกคุณลักษณะโดยการใช้งาน Information gain มีการใช้เวลาในการคำนวณที่มาพอสมควรและการจำแนกหนึ่งระดับ มีการใช้เวลาน้อยกว่าสองระดับมากสุดถึง 28.28 นาทีที่ 2,000 คุณลักษณะ ส่วนการจำแนกสองระดับพบว่าการใช้งานอัลกอริทึม SVM คู่กับอัลกอริทึม Bayes มีการสร้างแบบจำลองที่เร็วกว่า SVM กับอัลกอริทึม Random Forest มากถึง 9.12 นาทีที่ 2,000 คุณลักษณะ

3. เวลาในการทดสอบแบบจำลอง โดยทดสอบข้อมูลชุด Test set กับแบบจำลองที่สร้างโดยคุณลักษณะ 2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะและคุณลักษณะทั้งหมด มีผลการทดลองดัง Figure 11 จากผลการทดลองทั้งหมดพบว่า การใช้งานการจำแนกสองระดับด้วยอัลกอริทึม SVM กับอัลกอริทึม Random Forest ที่ 2,000 คุณลักษณะให้ผลที่ไวกว่าการใช้จำแนกหนึ่งระดับและสองระดับมากสุดถึง 7.33 นาที เนื่องจากอัลกอริทึม Random Forest เป็นอัลกอริทึมที่มีต้นไม้ตัดสินใจอยู่ภายในและไม่ได้ใช้งานคุณลักษณะทุกคุณลักษณะมาคำนวณเหมือนอัลกอริทึม Bayes และ SVM จึงทำให้ผลลัพธ์ในการทดสอบข้อมูลได้ผลดีที่สุด

สรุปผลและวิจารณ์ผลการทดลอง

ในงานวิจัยในครั้งนี้ผู้วิจัยได้ทำการกรองข้อมูลจาก Twitter โดยการตัดข้อความที่เป็นการ Retweet การตัดข้อความที่เป็นลิงค์เข้าใช้งานเว็บไซต์ การตัดข้อความที่เกี่ยวข้องกับชื่อบุคคลที่ถูกแทกในโพสต์และการตัด Stop word สำหรับการให้น้ำหนักเลือกใช้งานวิธีการ Binary term occurrence แล้วคัดเลือกแอตทริบิวต์ด้วย Information Gain ที่กำหนด Top-K=2,000 คุณลักษณะ 4,000 คุณลักษณะ 6,000 คุณลักษณะและคุณลักษณะทั้งหมด ในการสร้างแบบจำลองผู้วิจัยเลือกใช้การแบ่งข้อมูลโดยใช้วิธีการ 10–Fold Cross Validation โดยมีผลการทดลองดังต่อไปนี้

สำหรับการสร้างแบบจำลองด้วย Training set การจำแนกหนึ่งระดับที่ใช้งานอัลกอริทึม Bayes ในการสร้างแบบจำลองได้ Accuracy สูงสุด 82.55% และอัลกอริทึม SVM ได้ Accuracy สูงสุด 96.18% การจำแนกสองระดับที่ใช้งานอัลกอริทึม SVM ในการสร้างแบบจำลองการจำแนกข้อความทั่วไปกับข้อความที่เกี่ยวข้องกับโรคซึมเศร้า มีค่า Accuracy สูงสุด 98.20% ต่อมาการสร้างแบบจำลองสำหรับจำแนกอาการที่เกี่ยวข้องกับโรคซึมเศร้าอัลกอริทึม SVM จับคู่กันกับอัลกอริทึม Bayes มีค่า Accuracy สูงสุด 83.23% และอัลกอริทึม SVM จับคู่กับอัลกอริทึม Random Forest มีค่า Accuracy สูงสุด 91.45%

สำหรับการทดสอบแบบจำลองด้วย Test set ที่เป็นข้อมูลผู้ใช้งานที่เป็นโรคซึมเศร้าและไม่เป็นโรคซึมเศร้าจำนวนทั้งหมด 30 บุคคล โดยมีการกำหนด Boundary เพื่อหาช่วงการคัดเลือกค่าความน่าจะเป็นที่เหมาะสมสำหรับการโหวตโดยกำหนด Boundary ที่ 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 และ 0.9 การทดลองแบ่งออกการจำแนกทั้ง 2 ระดับ การจำแนกหนึ่งระดับด้วยอัลกอริทึม Bayes ได้ค่าความน่าจะเป็นที่เหมาะสม คือ 0.4 มี Accuracy สูงสุด 76.67% และ อัลกอริทึม SVM ได้ค่าความน่าจะเป็นที่เหมาะสม คือ 0.5 มี Accuracy สูงสุด 70.00% การจำแนกสองระดับ อัลกอริทึม SVM จับคู่กับอัลกอริทึม Bayes ได้ค่าความน่าจะเป็นเหมาะสม คือ 0.6 มี Accuracy สูงสุด 73.00% และอัลกอริทึม SVM จับคู่กับอัลกอริทึม Random Forest ได้ค่าความน่าจะเป็นเหมาะสม คือ 0.1 มี Accuracy สูงสุด 70.00%

จากผลการทดลองทั้งหมดสรุปได้ว่าเทคนิคการคัดเลือกคุณลักษณะด้วย Information Gain มีผลทำให้เวลาในการสร้างแบบจำลองลดน้อยลงอย่างมาก โดยเฉพาะอัลกอริทึมที่ใช้เวลานานอย่าง Random Forest ลดเวลาการทำงานได้ถึง 270.06 นาที อีกทั้งการลดคุณลักษณะให้ได้จุดเหมาะสมทำให้ประสิทธิภาพของการทดสอบแบบจำลองเพิ่มขึ้นและในการเลือกใช้อัลกอริทึมเพื่อจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์ การใช้งานที่ดีที่สุดคือการจำแนกหนึ่งระดับด้วยอัลกอริทึม Bayes เนื่องจากให้ผลลัพธ์ในการทดสอบกับชุดข้อมูลในโลกความจริง (Real world data) ได้ความถูกต้องที่ดีที่สุดคือ Accuracy=76.67% และได้ค่า Boundary ความน่าจะเป็นที่เหมาะสมแก่การทำ Vote ensemble ที่ 0.4

จากจุดประสงค์ในการทำวิจัยครั้งนี้ ผู้วิจัยได้ทำการเปรียบเทียบประสิทธิภาพการสร้างแบบจำลองจำแนกหนึ่งระดับและจำแนกสองระดับเพื่อให้มีประสิทธิภาพสูงสุดในการการจำแนกโรคซึมเศร้าจากพฤติกรรมการโพสต์ข้อความบนทวิตเตอร์ จึงมีข้อเสนอแนะว่าข้อมูลที่นำมาสร้างแบบจำลองเพียงแค่ข้อความตัวอักษรในการทำนายโรคซึมเศร้าอาจจะไม่เพียงพอต่อการจำแนกพฤติกรรม ซึ่งการใช้งานข้อมูลอื่นๆ ของตัวบุคคล เช่น รูปภาพ อายุ ความสัมพันธ์ และเพศ อาจจะมีผลต่อการจำแนกพฤติกรรมที่เข้าข่ายเป็นโรคซึมเศร้าได้ และยังพบว่าคุณลักษณะที่ได้จากอาการแต่ละอาการที่มี 3,000 ข้อความ อาจจะยังไม่เพียงพอต่อการจำแนกลักษณะอาการดีพอ เนื่องจากในโลกความจริงนั้น มีข้อมูลจำนวนมาก อาจทำให้ความถูกต้องลดลงได้

กิตติกรรมประกาศ

โครงการวิจัยนี้ได้รับการสนับสนุนเงินทุนอุดหนุนการวิจัยจากงบประมาณเงินรายได้ ประจำปี 2562 มหาวิทยาลัยมหาสารคาม

เอกสารอ้างอิง

1. World Health Organization (2018), *Depression.* (2019). https://www.who.int/news-room/fact-sheets/detail/depression.
2. สถาบันวิจัยระบบสาธารณสุข (2015), *โรคซึมเศร้า ภัยร้ายใกล้ตัว*. (2019). https://hsri.or.th/people/media/care/detail/6268.
3. Yoon S, Taha B, Bakken S, Using a data mining approach to discover behavior correlates of chronic disease: a case study of depression. *Stud Health Technol Inform ;* 2014. P. 71–78.
4. Maryam H, Emmanuel A, Elke A. R, Using Hashtags as Labels for Supervised Learning of Emotions in Twitter Messages. In: *ACM Workshop on Workshop on Health Informatics*. New York USA *;* 2014.
5. McManus K, Mallory E, Goldfeder R, Haynes W, Tatum J, Mining Twitter Data to Improve Detection of Schizophrenia. *AMIA Joint Summits on Translational Science ;* 2015. P. 122–126.
6. Anees U. H, Jamil H, Musarrat H, Muhammad S, Sungyoung L, Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. 2017 *International Conference on Information and Communication Technology Convergence (ICTC) ;* 2017. P. 138-140.
7. กรมสุขภาพจิต (2019). *โรคซึมเศร้ากับการเอาชีวิตรอด*. (2019). https://www.dmh.go.th/news-dmh/view.asp?id=29531.
8. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Washington DC: American Psychiatric Association *;* 2013.
9. Quinlan JR. *Mach Learn 1: Induction of Decision Trees ;* 1986. P. 81-106.
10. Vapnik, Vladimir N. *Machine Learning: Support-vector networks ;* 1995. P. 273-297.
11. Russell S, Norvig P. *Artificial Intelligence: A Modern Approach ;* 1995. P. 495–499.
12. Ho TK, Random Decision Forests. *Proceedings of the 3rd International Conference on Document Analysis and Recognition ;* 1995. P. 278–282.
13. RANKS NL (2019), *Default English stop words list.* (2019). https://www.ranks.nl/stopwords.