โครงการ
การเรียนการสอนเพื่อเสริมประสบการณ์

ชื่อโครงการ การวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์
Sentiment Analysis to Study Depression on Twitter

ชื่อนิสิต นางสาวพรพัทธา อมรรังสรรค์ 5933641023

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์
สาขาวิชา วิทยาการคอมพิวเตอร์

ปีการศึกษา 2562

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์

นางสาวพรพัทธา อมรรังสรรค์

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

SENTIMENT ANALYSIS TO STUDY DEPRESSION ON TWITTER

Phornpatta Amornrungsan

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of Bachelor of Science Program in Computer Science

Department of Mathematics and Computer Science

Faculty of Science

Chulalongkorn University

Academic Year 2019

หัวข้อโครงงาน การวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์

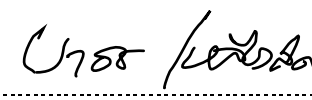
โดย นางสาวพรพัทธา อมรรังสรรค์

สาขาวิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาโครงงานหลัก ผศ.ดร. อาธร เหลืองสดใส

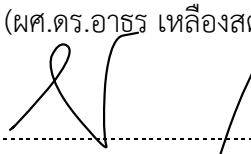
ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับโครงงานฉบับนี้เป็นส่วนหนึ่ง ของการศึกษาตามหลักสูตรปริญญาบัณฑิต ในรายวิชา 2301499 โครงงานวิทยาศาสตร์ (Senior Project)

.....หัวหน้าภาควิชาคณิตศาสตร์
(ศาสตราจารย์ ดร.กฤษณะ เนียมมณี) และวิทยาการคอมพิวเตอร์

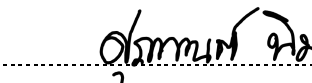
คณะกรรมการสอบโครงงาน



.....อาจารย์ที่ปรึกษาโครงงานหลัก
(ผศ.ดร.อาธร เหลืองสดใส)



.....กรรมการ
(อ.โชติรส สุรพลชัย)



.....กรรมการ
(ผศ.ดร.ศุภกานต์ พิมลธเรศ)

พรพัทธา อมรรังสรรค์ : การวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์.
(SENTIMENT ANALYSIS TO STUDY DEPRESSION ON TWITTER)
อ.ที่ปรึกษาโครงงานหลัก : ผศ.ดร.อาธร เหลืองสดใส, 42 หน้า.

โครงการวิจัยเรื่อง “การวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์” มีวัตถุประสงค์เพื่อการศึกษาวิจัยและวิเคราะห์แยกแยะเกี่ยวกับการวิเคราะห์ความรู้สึกสำหรับภาวะโรคซึมเศร้าบนทวิตเตอร์ จุดมุ่งหมายขอบเขตการวิจัยจะอยู่ภายใต้หลักการวิเคราะห์อารมณ์ความรู้สึกเท่านั้น วิธีการวิจัยจะใช้วิธีรวบรวมชุดข้อมูลจากทวิตเตอร์ และจัดการกับข้อมูล รวมทั้งสกัดคุณลักษณะต่าง ๆ อ้างอิงจากการศึกษาพฤติกรรมภาวะอาการของผู้ที่มีภาวะโรคซึมเศร้าจากงานวิจัยต่าง ๆ กล่าวคือ อาการ เช่น อาการนอนไม่หลับ การมีความคิดเชิงลบที่เกี่ยวกับความตาย เป็นต้น โดยสามารถวิเคราะห์ผ่านคำศัพท์จากข้อความที่ทำการโพสต์ได้ รวมถึงคุณลักษณะการมีปฏิสัมพันธ์กับคนรอบข้างโดยศึกษาผ่านจำนวนการใช้งานบนเครือข่าย จำนวนการติดตามผู้อื่นและจำนวนการถูกติดตาม ทั้งนี้การใช้คำสรรพนามแทนตนเองเป็นส่วนหนึ่งในการนำมาวิเคราะห์แยกแยะสำหรับการเข้าร่วมสังคมได้เช่นกัน ท้ายที่สุดเมื่อนำเข้ากระบวนการฝึกฝนรู้จำ ด้วยแบบจำลองการจำแนกนาอีฟเบย์ แบบจำลองการค้นหาเพื่อนบ้านใกล้สุด k ตัว แบบจำลองต้นไม้ตัดสินใจ และแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน ผลการวิจัยที่ได้แสดงให้เห็นว่า การใช้คุณลักษณะทั้งหมดกับแบบจำลองต้นไม้ตัดสินใจให้ผลประสิทธิภาพดีที่สุด

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต พรพัทธา อมรรังสรรค์

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาโครงงานหลัก อาธร

ปีการศึกษา 2562

5633641023 : MAJOR COMPUTER SCIENCE

KEYWORDS : SENTIMENT ANALYSIS / NATURAL LANGUAGE PROCESSING / DEPRESSION

PHORNPATTA AMORNRUNGSAN: SENTIMENT ANALYSIS TO STUDY DEPRESSION ON TWITTER. ADVISOR : ASSOC. PROF. ARTHON LUEANGSODSAI ,42 pp.

The topic of the classroom action research is "Sentiment analysis to study depression on twitter". The objectives of this research are to research and analyze the sentiment analysis for depression on Twitter. The scope of the research covers the sentiment analysis only. The research methodology uses data collection from Twitter, and preprocess data including feature extraction. Data about insomnia are collected from a study on the time of posting on Twitter. A pattern of having negative thoughts and thoughts about death can be analyzed through vocabularies from the post. The features to study the interaction with those around the user will study through the amount of usage on the network, number of followers and numbers of followings. In this regard, the use of personal pronouns as part of the analysis can be used to distinguish social participation. Finally, we apply processes with Naive Bay classification, K-nearest neighbor, Decision tree and Support vector machine. The research results show that using all features with the Decision tree model gives the best performance.

Department :Mathematics and Computer Science Student's Signature Phornpatta Amornrungsan

Field of Study : Computer Science Advisor's Signature

Academic Year : 2019

กิตติกรรมประกาศ

งานวิจัยในหัวข้อเรื่อง “การวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์” ได้รับการสนับสนุนจากและความช่วยเหลือเต็มที่จาก ผู้ช่วยศาสตราจารย์ ดร.อาธร เหลืองสดใส อาจารย์ที่ปรึกษาโครงงาน ที่ช่วยตรวจทานแก้ไขข้อผิดพลาดและให้คำแนะนำที่มีประโยชน์ในการทำโครงงาน

ขอขอบพระคุณ อาจารย์โชติรส สุรพลชัย และผู้ช่วยศาสตราจารย์ ดร.ศุภกานต์ พิมลธเรศ ผู้เป็นกรรมการคุมสอบ ที่ได้ชี้แนะแนวทางในการทำโครงงานและได้ช่วยสนับสนุนโครงการวิจัยนี้อย่างดี และขอขอบคุณ ผศ.ดร.ภควรรณ ปักษี อาจารย์ที่ปรึกษาที่คอยสนับสนุนแนะแนวและได้ถ่ายทอดความรู้ความเข้าใจในการทำโครงงานทำให้โครงงานนี้สำเร็จได้

ขอขอบพระคุณภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ที่ได้จัดสรรงบประมาณค่าใช้จ่ายในการดำเนินงานวิจัยนี้และหวังว่าผลการวิจัยนี้จะเป็นประโยชน์ในการพัฒนาระบบการเรียนการสอนได้ต่อไป

ขอขอบพระคุณบิดา มารดา และเพื่อน ๆ ทุกท่านที่ได้ชี้แนะแนวทาง สนับสนุน และให้กำลังใจเสมอตลอดการดำเนินงานวิจัยนี้

สารบัญ

สารบัญภาพ

สารบัญตาราง

บทที่ 1

บทนำ

1.1 หลักการและเหตุผล

ภาวะโรคซึมเศร้าเป็นภาวะที่เกิดขึ้นอย่างแพร่หลายมากในปัจจุบัน อาการแห่งความเศร้า เบื่อหน่าย ไม่มี ความสุข เครียดกระวนกระวาย ขาดความมั่นใจ รู้สึกไร้ค่าหรือโทษตัวเอง [7] ขึ้นอยู่กับแต่ละสถานการณ์ และ สภาพแวดล้อมของแต่ละบุคคลต่างกันไป จากสถิติพบว่า มีผู้ป่วยโรคซึมเศร้ามากกว่า 300 ล้านคนทั่วโลก และภาวะโรคซึมเศร้านี้ก็เป็นอีกหนึ่งในสาเหตุหลักซึ่งนำไปสู่การทำร้ายตัวเองและฆ่าตัวตายในที่สุด เราจะเห็นว่า ตัวเลขของอัตราการฆ่าตัวตายในแต่ละประเทศทั่วโลกนั้นมีอัตราที่สูง ซึ่งพบว่าจะมีผู้ที่ลงมือฆ่าตัวตายถึง 80,000 คนในแต่ละปี ซึ่งเฉลี่ยแล้วคือมีการฆ่าตัวตายเกิดขึ้นทุก ๆ 40 วินาทีทั่วโลก ในประเทศไทยนั้นพบว่ามีอัตราการ ฆ่าตัวตายสำเร็จใน ปี พ.ศ. 2562 อยู่ที่ร้อยละ 3.08 ของประชากร 1 แสนคน คือเฉลี่ยประมาณ 11-12 รายต่อวัน และปัจจัยที่มีความสัมพันธ์กับการฆ่าตัวตายนั้น ประกอบด้วยสาเหตุโรคซึมเศร้าถึงร้อยละ 7.8

ภาวะโรคซึมเศร้าเกิดขึ้นได้กับผู้คนทุกเพศทุกวัย ในอดีตความเครียดและความซึมเศร้ามักพบในหมู่คนทำงาน วัยกลางคนจนถึงผู้สูงอายุ แต่ในปัจจุบันภาวะโรคซึมเศร้านั้นพบมากในหมู่วัยรุ่นช่วงอายุ 15-25 ปี ทั้งเด็กนักเรียนมัธยม นิสิตนักศึกษามหาวิทยาลัย รวมทั้งกลุ่มคนทำงานสามารถเกิดภาวะโรคซึมเศร้าขึ้นในอัตราที่สูงกว่าวัยอื่น ๆ กลุ่มคนรุ่นใหม่ที่อาศัยอยู่ ในยุคแห่งความก้าวหน้าทางเทคโนโลยี ความก้าวกระโดดของนวัตกรรม ยุคแห่งข้อมูลข่าวสารที่ส่งถึงกันทั่วโลก การสื่อสารที่เข้าถึงและสามารถในการเชื่อมต่อของผู้คนจากแต่ละมุมโลกอย่างใกล้ชิด แต่กลับทำให้คนรู้สึกเดียวดาย เหงา เครียดและมีภาวะโรคซึมเศร้าได้อย่างไม่น่าเชื่อ

ทวิตเตอร์เป็นเครือข่ายออนไลน์อันเป็นที่นิยมของคนทั่วโลก โดยมีลักษณะให้บริการเป็น micro blogging มีการจำกัดตัวอักษรเพียง 140 ตัวต่อหนึ่งข้อความ ข่าวสารต่าง ๆ ถูกส่งออกอย่างรวดเร็วตลอดเวลา จากการสำรวจ พบว่ามีการส่งข้อความทางทวิตเตอร์ถึง 6,000 ข้อความต่อวินาทีทั่วโลก คือ 500 ล้านทวิตเตอร์ต่อหนึ่งวัน [4] โดยทวิตเตอร์ได้รับความนิยมอย่างมากในหมู่วัยรุ่น ซึ่งเป็นที่ที่ได้เปิดเผยตัวตนความเป็นตัวเองเป็นส่วนใหญ่ ได้ถ่ายทอดความคิด ความรู้สึกออกมาอย่างอิสระเสรี เนื่องจากมักมีการตั้งชื่อบัญชีเป็นนามสมมุติ ไม่ระบุตัวตนที่แท้จริง (Anonymous) ทำให้ข้อความที่ส่งออกมานั้นเป็นดั่งการบ่งบอกสถานะความรู้สึกในใจ ออกมาอย่างเปิดเผย ซึ่งทำให้มีนักวิเคราะห์ข้อมูลมากมายสนใจที่จะศึกษารูปแบบพฤติกรรมความคิดผ่านทางสื่อโซเชียลอย่างทวิตเตอร์

จากการทบทวนบทความพบว่ามีงานวิจัยหลายชิ้นที่ทำการศึกษาเกี่ยวกับการตรวจจับหรือบ่งชี้ภาวะโรคซึมเศร้าบนทวิตเตอร์ผ่านคุณลักษณะต่าง ๆ เช่น คุณลักษณะทางภาษาศาสตร์ การเข้าร่วมต่อเครือข่าย

การตอบสนอง ของเครือข่ายรอบข้าง รูปแบบของกิจกรรมบนเครือข่าย พฤติกรรมการใช้งานยามดึก รวมถึงการมีอิทธิพลต่อ เครือข่าย งานวิจัยของ Amir และคณะ [6] ได้นำเสนอวิธีการในการทำนาย ความรู้สึกบนทวิตเตอร์ด้วยวิธีการ วิเคราะห์อารมณ์ (Sentiment Analysis) [2] โดยมีวิธีการดังนี้ 1) สกัดข้อความทวิตเตอร์ที่เกี่ยวข้องกับการ วิเคราะห์อารมณ์ 2) สกัดความรู้สึกบนทวิตเตอร์ และ 3) จำแนกประเภทของอารมณ์ทั้งบวกและลบ โดยมีการอ้างอิงจาก บทความของ Agarwalและคณะ [1] งานวิจัยของ Bravo-Marquez และคณะ [3] มีการใช้คลังคำศัพท์ (Lexical) เพื่อจับคู่ความสัมพันธ์ระหว่างคำในข้อความและคำในคลังศัพท์เพื่อดูคุณลักษณะ โดยในบทความนี้ได้ศึกษา 6 คลังศัพท์ ด้วยกันคือ Opinion Finder Lexicon, AFINN Lexicon, SentiWordNet Lexicon, SentiStrength Lexicon, Sentiment140 Method และ NRC Lexiconซึ่งแต่ละอันจะได้คุณลักษณะ ทางอารมณ์ที่แตกต่างกันไป อีกทั้งในงานวิจัยของ Suppala และคณะ [5] มีการทำการแยก (classification) โดยวิธีการของนาอีฟเบย์ (Naive Bayes) คือการดูโอกาสความน่าจะเป็นของอารมณ์ความรู้สึกต่าง ๆ บนชุดข้อมูลทวิตเตอร์

จากงานวิจัยดังกล่าวมีการใช้การวิเคราะห์อารมณ์ความรู้สึกความรู้สึกบนข้อมูลทวิตเตอร์ที่น่าสนใจ ผู้พัฒนาจึงเลือกที่จะพัฒนาโครงงานเรื่องการวิเคราะห์อารมณ์เพื่อศึกษาภาวะซึมเศร้าบนทวิตเตอร์ โดยใช้การทำ Sentiment Analysis หรือ การวิเคราะห์อารมณ์บนข้อความทวิตเตอร์

1.2 วัตถุประสงค์

เพื่อศึกษาวิจัยและวิเคราะห์แยกแยะเกี่ยวกับการวิเคราะห์ความรู้สึก (Sentiment Analysis) สำหรับภาวะซึมเศร้าบนทวิตเตอร์

1.3 ขอบเขตของโครงงาน

1.ชุดข้อมูลข้อความทวิตเตอร์ภาษาไทยจาก API ในช่วงเวลา 4 เดือน ไม่เกิน 20,000 ข้อความ

2.ผลลัพธ์จากการวิเคราะห์อารมณ์สามารถแบ่งกลุ่มผู้ที่มีภาวะซึมเศร้ากับกลุ่มปกติ

3.ใช้การแยกแยะกลุ่มของโดย

1.) Naïve Bayes Classification

2.) Support Vector Machine

3.) Decision Tree

4.) K-Nearest Neighbor

4.ใช้ซอฟต์แวร์ Rapidminer และมีการใช้ Jupyter notebook ช่วยในการเขียนโปรแกรมด้วยภาษาไพธอน

1.4 วิธีการดำเนินงาน

ก. แผนการศึกษา

1.ศึกษาค้นคว้าทฤษฎีที่เกี่ยวข้อง

2.ศึกษางานวิจัยที่เกี่ยวข้อง

3.ศึกษาเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีต่าง ๆ

4.เก็บตัวอย่างข้อมูลทวิตเตอร์

5. วิจัยและศึกษาจำแนกภาวะซึมเศร้า

6. เปรียบเทียบประสิทธิภาพการจำแนกภาวะซึมเศร้า

7. ประเมินผลและอธิบายผล

8. จัดทำเอกสาร

ข.ระยะเวลาที่ศึกษา

ขั้นตอน	ปี พ.ศ. 2562					ปี พ.ศ. 2563			
ขั้นตอนการดำเนินงาน	ส.ค.	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1.ศึกษาค้นคว้าทฤษฎีที่เกี่ยวข้อง									
2. ศึกษางานวิจัยที่เกี่ยวข้อง									
3. ศึกษาเปรียบเทียบประสิทธิภาพขั้นตอนวิธีต่าง ๆ									
4. เก็บตัวอย่างข้อมูลทวิตเตอร์									
5. วิจัยและศึกษาจำแนกภาวะซึมเศร้า									

6. เปรียบเทียบประสิทธิภาพการจำแนกภาวะซึมเศร้า									
7. ประเมินผลและอธิบายผล									
8. จัดทำเอกสาร									

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ก. ประโยชน์ด้านความรู้และประสบการณ์ต่อนิสิต

1. ได้เรียนรู้พัฒนาทักษะการศึกษาค้นคว้างานวิจัย หาข้อมูลจากแหล่งต่าง ๆ
2. ได้รับความรู้เกี่ยวกับการรวบรวมข้อความบน API ทวิตเตอร์
3. ได้ฝึกฝนทักษะการเขียนโปรแกรม ในการทำงานกับข้อมูลจริง
4. ได้ฝึกทักษะการวางแผนและแก้ไขปัญหา
5. ได้เรียนรู้ที่จะรับผิดชอบ

ข. ประโยชน์ที่ได้จากโครงงานที่พัฒนาขึ้น

1. สามารถเห็นสัญญาณของผู้ที่จะมีแนวโน้มมีภาวะโรคซึมเศร้า
2. ให้ตระหนักถึงความสำคัญและความเกี่ยวข้องของภาวะซึมเศร้าผ่านทางสื่อออนไลน์ในชีวิตประจำวัน

1.6 โครงสร้างของรายงาน

บทที่ 2 กล่าวถึงหลักการและทฤษฎีที่เกี่ยวข้องกับภาวะโรคซึมเศร้าและการวิเคราะห์อารมรณ์
บทที่ 3 กล่าวถึงวิธีการศึกษาวิจัยที่ทำการฝึกฝนบน 4 แบบจำลองได้แก่แบบจำลองนาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และแบบจำลองเพื่อนบ้านที่ใกล้เคียงที่สุดจำนวน K ตัว
บทที่ 4 กล่าวถึงผลการทดลอง และการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลอง
บทที่ 5 กล่าวถึงการอภิปรายผลการวิจัยและข้อเสนอแนะ

บทที่ 2

งานวิจัยที่เกี่ยวข้อง

ในบทนี้ผู้ทำวิจัยได้ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องดังนี้

2.1 หลักการทฤษฎีที่เกี่ยวข้อง

2.1.1 ความรู้ทั่วไปเกี่ยวกับภาวะโรคซึมเศร้า

2.1.2 หลักการการวิเคราะห์อารมณ์ความรู้สึก (Sentiment Analysis)

2.2 งานวิจัยที่เกี่ยวข้อง

2.1 หลักการและทฤษฎีที่เกี่ยวข้อง

2.1.1 ความรู้ทั่วไปเกี่ยวกับภาวะโรคซึมเศร้า

งานวิจัยนี้ไม่ได้มีเป้าหมายเพื่อการวินิจฉัยโรคซึมเศร้า แต่เพื่อทำการศึกษาวิจัยอารมณ์ความรู้สึกผ่านคำและข้อความเท่านั้น

ข้อมูลทฤษฎีเกี่ยวกับภาวะโรคซึมเศร้าทำให้เห็นภาพกว้างและรูปแบบของพฤติกรรม คำพูด รวมถึงความรู้สึกของผู้ที่มีภาวะของโรคนี้ กล่าวคือ พฤติกรรมการมีปฏิสัมพันธ์กับคนรอบข้าง พฤติกรรมการนอน การแสดงออกความคิดเห็น ลักษณะการพูด และลักษณะการใช้คำ ซึ่งส่งผลต่อการบ่งชี้ภาวะโรคซึมเศร้าจะออกมาผ่านทางข้อความทวิตเตอร์และข้อมูลต่าง ๆ ที่เราทำการศึกษาวิจัยอย่างสังเกตุได้

2.1.1.1 อาการของโรคซึมเศร้า

อาการหลักของโรคซึมเศร้ารุนแรง คือ มีอารมณ์เศร้าหมอง เบื่อหน่าย ไม่มีความสุข ติดต่อกันอย่างต่อเนื่องเป็นเวลานาน โดยอาการเหล่านี้แสดงออกได้ชัดในรูปแบบของพฤติกรรมเปลี่ยนแปลง ได้แก่ [8]

- อารมณ์แปรปรวนตามเวลา คือเศร้ามากที่สุดในช่วงเช้าตรู่ และดีขึ้นในช่วงบ่ายถึงค่ำ ซึ่งมีผลมาจากฮอร์โมน cortisol
- อาการคิดหมกมุ่นเกี่ยวกับความตาย
- อาการย้ำคิดย้ำทำ เป็นอาการที่ความคิดความรู้สึกเกิดขึ้นอย่างซ้ำ ๆ โดยไม่ทราบสาเหตุ จึงวนเวียนกับความคิดวิตกกังวลและการกระทำบางอย่าง

ผู้ป่วยโรคซึมเศร้านอกจากอาการเศร้าแล้วยังพบกลุ่มอาการร่วมดังนี้คือ เบื่ออาหาร อ่อนเพลีย นอนไม่หลับ ไม่มีสมาธิ ไม่สนในสิ่งแวดล้อม มองโลกในแง่ร้าย รู้สึกตนเองไร้ค่า และมีความรู้สึกผิดในระดับที่มากและรุนแรง [9]

2.1.1.2 สาเหตุของโรคซึมเศร้า [9]

1. การขาดความรู้ความเข้าใจเกี่ยวกับโรค ยารักษาโรค และการจัดการกับอาการข้างเคียงของโรคซึมเศร้า

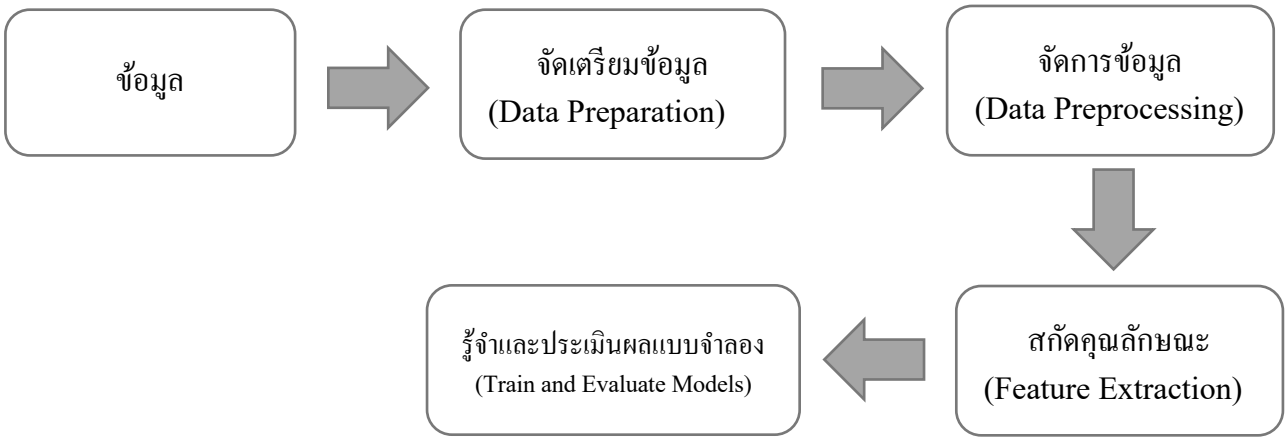
2.การขาดทักษะการดูแลตนเองด้านจิตใจ เช่น การคิดเชิงบวก การยับยั้งความคิดเชิงลบ การผ่อนคลายความเครียด ทักษะการแก้ปัญหา และการเห็นคุณค่าในตัวเอง เป็นต้น มีผลทำให้ไม่สามารถจัดการปัญหาด้านต่าง ๆ ที่เชื่อมโยงกับภาวะซึมเศร้าได้อย่างมีประสิทธิภาพ

3.ขาดทักษะทางสังคม เช่นการสร้างสัมพันธภาพ การสื่อสารและการหาแหล่งสนับสนุนทางสังคม ซึ่งเป็นสาเหตุหลักของการกลับไปเป็นซ้ำของโรค

2.1.2 หลักการการวิเคราะห์อารมณ์ความรู้สึก (Sentiment Analysis)

การวิเคราะห์อารมณ์ความรู้สึก เป็นความรู้แขนงหนึ่งในการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ซึ่งเป็นสาขาความรู้เกี่ยวกับวิทยาการคอมพิวเตอร์และภาษาศาสตร์ อัลกอริทึมของการประมวลผลภาษาธรรมชาติในปัจจุบันนั้นมักมีพื้นฐานอยู่บนการเรียนรู้ของเครื่อง (Machine Leaning) โดยเฉพาะการเรียนรู้ของเครื่องเชิงสถิติ (Statistic Machine Leaning) ซึ่งมีประโยชน์ในงานหลากหลายด้าน เช่น การวิเคราะห์ระดับพยางค์และคำ การระบุหน้าที่ของคำในประโยค และการวิเคราะห์อารมณ์ ที่งานวิจัยนี้ได้ทำการศึกษา เป็นต้น

การวิเคราะห์อารมณ์ความรู้สึก คือ กระบวนการเชิงคำนวณเพื่อกำกับข้อความหรือส่วนของข้อความที่แสดงความคิดเห็นด้วยทัศนคติของผู้เขียนข้อความนั้น ๆ ซึ่งมักจะเป็น ทัศนคติเชิงบวก เชิงลบ หรือเป็นกลาง มักใช้เพื่อการตีความและบ่งบอกประเภทของอารมณ์ความรู้สึกของข้อมูล



รูปภาพที่ 2.1 แผนภาพขั้นตอนการวิเคราะห์อารมณ์

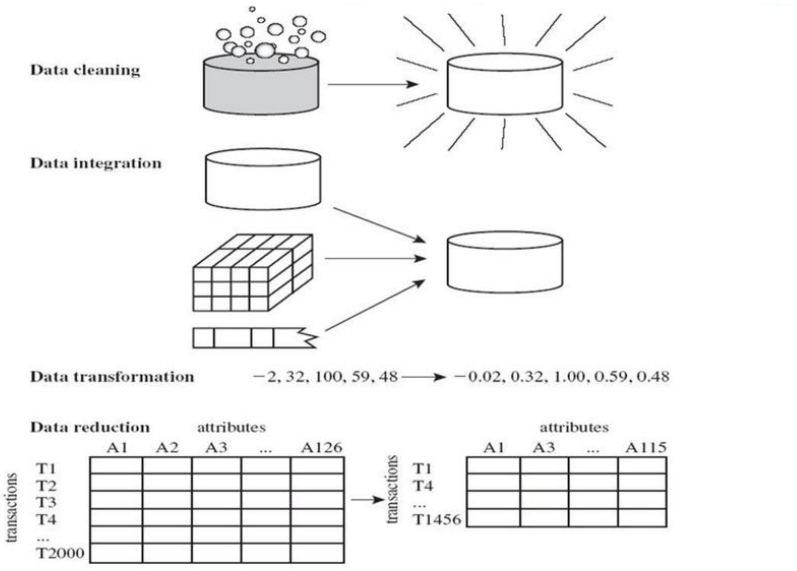
2.1.2.1 การเตรียมข้อมูล (Data Preparation)

ขั้นตอนการรวบรวมข้อมูล ซึ่งประกอบด้วย attribute ต่าง ๆ และทำให้ชุดข้อมูลอยู่ในโครงสร้าง (Structure) ที่เหมาะสม

2.1.2.2 การจัดการข้อมูล (Data Pre-processing)

เป็นขั้นตอนการจัดการกับข้อมูล ประมวลผล คัดกรองและปรับปรุงคุณภาพของข้อมูลก่อนนำไปใช้ซึ่งนับเป็นขั้นตอนที่สำคัญอย่างมากในการทำการศึกษาวิจัยเกี่ยวกับกระบวนการประมวลผลทางภาษา (Natural Language Processing) หากไม่มีการจัดการข้อมูลที่ดี ก็จะมีโอกาสสูงในการเกิดความเสียหายขึ้นในกระบวนการรู้จำ ชุดข้อมูลที่ถูกรบกวน ซ้ำกัน หรือมีช่วงระยะห่างและความแตกต่างที่มากเกินไป อาจส่งผลให้ผลการวิเคราะห์ตีความผิดเพี้ยนได้ โดยเทคนิควิธีการของการจัดการข้อมูล (Data Pre-processing Technique)หลากหลายแบบ ดังนี้ [10]

- การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนสำหรับการคัดข้อมูลที่เป็นส่วนรบกวน หรือข้อมูลที่ไม่เกี่ยวข้องออก
- การรวบรวมข้อมูล (Data Integration) เป็นขั้นตอนการรวมแหล่งข้อมูลซึ่งมีข้อมูลหลายแหล่งมารวมไว้ที่เดียวกัน
- การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนการแปลงข้อมูลในขั้นตอนการคัดเลือก ให้เหมาะสำหรับขั้นตอนการทำเหมืองข้อมูล เช่น แปลงข้อมูลตัวเลข (Numeric) ให้เป็นข้อมูลกลุ่ม (Norminal) หรือการ Normalization ปรับช่วงของข้อมูลให้อยู่ในมาตราส่วน (Scale) เดียวกัน
- การลดข้อมูล (Data Reduction) เป็นขั้นตอนการลดมิติข้อมูล เพื่อเป็นตัวแทนจำนวนข้อมูลทั้งหมด



รูปภาพที่ 2.2 แสดงแบบแผนของการจัดการข้อมูล (Data pre-processing)

(รูปภาพจาก Han & Kamber , 2006)

โดยลักษณะข้อมูลของที่ควรจัดการและควรพึงตรวจสอบมีลักษณะดังนี้

-ข้อมูลไม่สมบูรณ์ (Incomplete data) เช่น ค่าข้อมูลขาดหาย (Missing value) ขาดคุณลักษณะที่น่าสนใจ รายละเอียดของข้อมูลขาดหาย

-ข้อมูลรบกวน (Noisy data) เช่น ข้อมูลที่มีค่าผิดพลาด (Error) หรือมีค่าผิดปกติ (Outliers)

-ข้อมูลไม่สอดคล้อง (Inconsistent data) เช่น ข้อมูลเดียวกันที่มีชื่อต่างกัน หรือมีการใช้ค่าข้อมูลที่ต่างกัน

2.1.2.3 การกลั่นคุณลักษณะ (Feature Extraction)

การกลั่นหรือการดึงเอาคุณลักษณะจากข้อมูลเป็นขั้นตอนที่สำคัญมากและมีผลต่อค่าประสิทธิภาพของการรู้จำ เพราะผลลัพธ์ของการจำแนกกลุ่มข้อมูลนั้นขึ้นอยู่กับความสามารถในการสกัดคุณลักษณะตามรูปแบบความจำเพาะของแต่ละกลุ่มข้อมูล ส่งผลต่ออัตราความถูกต้องของการรู้จำนั้นโดยตรง ซึ่งการสกัดคุณลักษณะนั้นคือ วิธีการแปลงจากชุดรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่งที่มีจำนวนของคุณลักษณะลดลงและช่วยลดเวลาในการรู้จำ ดังนั้นหากเลือกลักษณะที่เหมาะสมแล้วอัตราความถูกต้องของการรู้จำก็จะมากขึ้นเช่นกัน [11]

เนื่องจากการวิเคราะห์ความรู้สึกนั้นมักต้องใช้ข้อมูลและคุณลักษณะจำนวนมากเพื่อการรู้จำของแบบจำลองที่น่าเชื่อถือ ซึ่งคุณลักษณะบางอย่างอาจจะไม่มีผลต่อการรู้จำหรือไม่ได้มีความสำคัญในการแบ่งแยกคลาส (Class) จึงควรมีการตรวจวัดความเหมาะสมในการเลือกใช้คุณลักษณะ กล่าวคือการหาค่าความเกี่ยวข้องของข้อมูลในแอททริบิวต์ (Attribute) อ้างอิงกับชุดข้อมูลที่มีการระบุชนิดแล้ว (Labeled data) เพื่อช่วยในการตัดสินใจเลือกใช้คุณลักษณะนั้น ๆ จากน้ำหนักความเกี่ยวข้อง โดยการวัดค่าน้ำหนักความสัมพันธ์เกี่ยวข้องหลายวิธี โดยจะยกมา 3 วิธีดังนี้

1.Weight by Information Gain

2.Weight by Gain Ratio

3.Weight by Relief

วิธี Information Gain จะใช้วัดค่าความเกี่ยวข้องของข้อมูล โดยทำการวัดค่าเอนโทรปี (Entropy) หรือค่าความแปรปรวน ซึ่งเป็นการวัดความแตกต่างหรือการกระจัดกระจายของข้อมูล เมื่อข้อมูลมีความแตกต่างกันมาก ค่าเอนโทรปีจะสูง ในทางกลับกันถ้าข้อมูลมีความคล้ายคลึงกันมาก ก็จะส่งผลให้ค่าเอนโทรปีต่ำ กล่าวคือประสิทธิภาพของคุณลักษณะที่สำคัญเหมาะสม ค่า Information Gain ก็จะสูง

$$Information\ Gain = Entropy(initial) - [P(c_1) \times Entropy(c_1) + P(c_2) \times Entropy(c_2) + \cdots]$$

โดยที่ $Entropy(c_1) = -P(c_1)log_2P(c_1)$ และ $P(c_1)$ คือ ความน่าจะเป็นของ c_1

วิธี Information Gain Ratio คือ สัดส่วน (Ratio) ของ Information gain ต่อข้อมูลเนื้อแท้ (Intrinsic Information) ที่มีค่าหรือความหมายโดยตัวเอง ไม่ได้พึ่งความสัมพันธ์จากภายนอก วิธีการนี้เป็นการช่วยเสริมวิธี Information Gain ได้อย่างดีในกรณีที่ข้อมูลแอททริบิวต์นั้นมีความเฉพาะเจาะจง เช่น ข้อมูลรหัสประจำตัว ที่มีความจำเพาะ (Unique) สูง กล่าวคือข้อมูลตัวเลขรหัสนั้น ๆ ไม่ได้มีความหมายต่อการวิเคราะห์

วิธี Relief เป็นอัลกอริทึมที่สามารถรับมือกับความสัมพันธ์ของคุณลักษณะที่อ่อนไหวได้ โดยจะทำการคำนวณค่าคะแนนของแต่ละคุณลักษณะ และจัดลำดับคะแนนเพื่อเลือกลำดับที่มีคะแนนสูง

2.1.2.4 แบบจำลองเพื่อการรู้จำ

ในงานวิจัยนี้สนใจวิธีการจำแนกของแบบจำลองทั้งหมด 4 วิธี ได้แก่

1.) แบบจำลองการจำแนกแบบนาอีฟเบย์ (Naïve-Bayes Classification)

เป็นการทำเหมืองข้อมูลในแบบแบ่งกลุ่มจำแนกประเภท (Classification) ที่ถูกสร้างขึ้นโดยหลักความน่าจะเป็นจากทฤษฎีของเบย์ โดยใช้วิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น [12] โดยการคาดเดาจากสิ่งที่เคยเกิดขึ้นมาก่อน ตัวจำแนกนาอีฟเบย์เป็นวิธีการจำแนกข้อมูลออกเป็นประเภทต่าง ๆ ซึ่งเป็นสมการที่อยู่บนพื้นฐานของความน่าจะเป็นสำหรับวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ โดยสามารถคำนวณความน่าจะเป็นของสมมติฐานต่าง ๆ จากทฤษฎีของเบย์ (Bayes' Theorem) ในสมการ ให้ D แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็นของสมมติฐาน h

$$P(\,h|D)\;=\;\frac{P(D|h)\times P(h)}{P(D)}$$

$\mathrm{P(h)}$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ h

$\mathrm{P(D)}$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ D

$\mathrm{P(h|D)}$ คือ ความน่าจะเป็นของ h เมื่อเกิดเหตุการณ์ D ก่อน

$\mathrm{P(D|h)}$ คือ ความน่าจะเป็นของ D เมื่อเกิดเหตุการณ์ h ก่อน

$$\text{Naïve Bayesian Classification} = \frac{Max(P(C_i)\prod_{j=1}^{n}P(A_j|C_i)}{P(A_{1,}\ldots,A_n)}$$

$\mathrm{P(C_i)}$ คือ ความน่าจะเป็นของคลาส i

$\mathrm{P(A_j|C_i)}$ คือ ค่าความน่าจะเป็นของแอททริบิวต์ j ที่อยู่ในคลาส i

วิธีของนาอีฟเบย์มีข้อดีคือ สามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior knowledge) เข้ามาช่วยในการเรียนรู้ได้ ซึ่งพบว่าวิธีนี้ให้ประสิทธิภาพในการเรียนรู้ได้ดี เหมาะสมกับข้อมูลที่มีจำนวนไม่มาก และข้อมูลที่ไม่ขึ้นต่อกัน ในทางทฤษฎีแล้วการทำนายผลของแบบจำลองนาอีฟเบย์จะถูกต้อง ถ้าตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน โดยไม่ขึ้นกับตัวแปรอิสระตัวใดตัวหนึ่ง แบบจำลองไม่สามารถรองรับข้อมูลที่เป็นข้อมูลต่อเนื่อง (Continuous Data) ซึ่งข้อมูลในปัจจุบันนั้นมีไม่มากนักที่ตัวแปรอิสระทั้งหมดจะเป็นอิสระต่อกัน ด้วยดังนั้นตัวแปรอิสระหรือตัวแปรตามที่มีค่าเป็นค่าต่อเนื่อง จะต้องถูกแบ่งออกเป็นช่วง ซึ่งการแบ่งช่วงนั้นถ้ามีการแบ่งที่ไม่ดีก็จะทำให้ผลลัพธ์ของแบบจำลองที่ได้มีคุณภาพไม่ดีตามไปด้วย [15]

2.) การจำแนกวิธีเพื่อนบ้านที่ใกล้เคียงที่สุดจำนวน K ตัว (K-Nearest Neighbors Classification)

การจำแนกข้อมูลด้วยวิธีการค้นหาเพื่อนบ้านใกล้ที่สุดจำนวน K ตัว เป็นวิธีที่มีการเรียนรู้แบบขี้เกียจ (Lazy Learning) เนื่องจากไม่มีการสร้างแบบจำลองสำหรับจำแนกประเภทข้อมูลเตรียมไว้ล่วงหน้า คือ เมื่อมีข้อมูลใหม่ที่ต้องการจำแนกประเภท วิธีนี้จะนำข้อมูลนั้นมาเปรียบเทียบกับข้อมูลเดิมและดูความคล้ายคลึงกันของข้อมูลใหม่กับข้อมูลเดิม สามารถจำแนกประเภทของข้อมูลใหม่ได้โดยให้เป็นประเภทเดียวกับข้อมูลเดิมที่อยู่ใกล้เคียง

ก่อนการประมวลผลข้อมูลเพื่อการจำแนกประเภทด้วยวิธีนี้ จะต้องมีการกำหนดค่า K ซึ่งหมายถึง จำนวนของข้อมูลเดิมที่ใกล้เคียงกับข้อมูลที่ต้องการจำแนกประเภท การเลือกค่า K ที่เหมาะสม จะพิจารณาจากการวิเคราะห์ลักษณะของคลาสและข้อมูลเดิมที่มี

ข้อมูลเดิมมีลักษณะเฉพาะเจาะจงและมีจำนวนคลาสน้อย การจำแนกเพื่อให้ได้ค่าความแม่นสูง อาจไม่จำเป็นต้องเลือกค่า K ที่สูง ในทางกลับกัน หากเลือกค่า K เท่ากับ 1 ค่าความแม่นก็ควรจะสูง เนื่องจากข้อมูลมีลักษณะเฉพาะตัวและมีจำนวนคลาสน้อยทำให้แยกประเภทได้ง่าย กล่าวคือถ้าค่า K น้อยเกินไป จะส่งผลถึงความอ่อนไหวต่อจุดที่รบกวน และถ้าค่า K ใหญ่เกินไป จุดเพื่อนบ้านใกล้เคียง (Neighborhood) มีโอกาสจะถูกเลือกเป็นคลาสที่ไม่ถูกต้อง

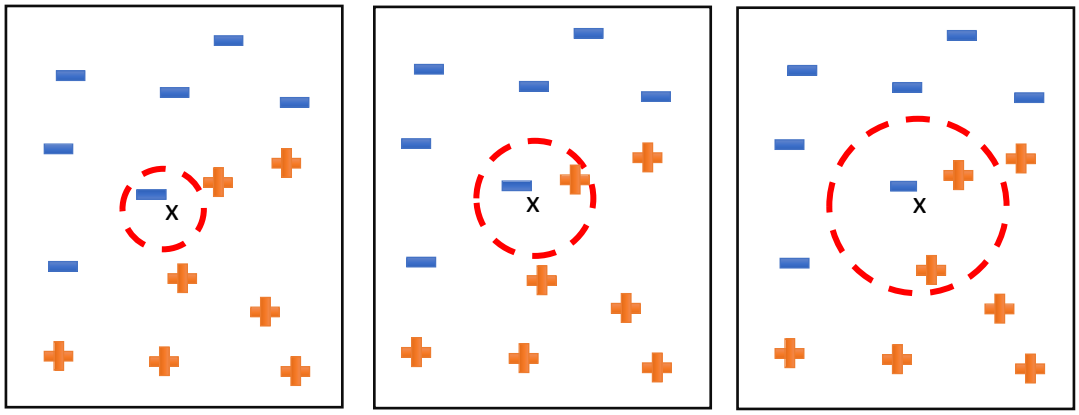
โดยสามารถอธิบายเป็น 5 ขั้นตอนได้ดังนี้ [13]

1. กำหนดค่า K และมาตรวัดระยะทางเพื่อนบ้านที่ใกล้เคียงที่สุด (Nearest Neighbor)
2. คำนวณระยะทางระหว่างข้อมูลใหม่ที่ต้องการจำแนกกับข้อมูลเดิมที่มีทั้งหมด
3. เรียงลำดับระยะทางและกำหนดเพื่อนบ้านที่ใกล้ที่สุดตามค่า K
4. รวบรวมคลาสเป้าหมายของเพื่อนบ้าน
5. กำหนดคลาสให้กับข้อมูลใหม่โดยพิจารณาจากประเภทคลาสเป้าหมายของเพื่อนบ้านว่าเป็นประเภทใดมากที่สุด การกำหนดคลาสให้ข้อมูลใหม่ก็จะเป็นคลาสนั้น

ถ้า x ประกอบไปด้วยแอททิบิวต์ $< a_1(x), a_2(x), \dots a_n(x) >$ สามารถค่าระยะทางที่เรียกว่า Euclidean Distance ได้โดยสูตร

$$d_{Euclidean}(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2}$$

ข้อดีของวิธีการเพื่อนบ้านที่ใกล้เคียงที่สุดจำนวน K ตัวทคือมีกระบวนการรู้จำที่ง่ายและประสิทธิภาพดี มีความสามารถในการเรียนรู้ฟังก์ชั่นที่ซับซ้อน และไม่ทำให้ข้อมูลสูญหาย

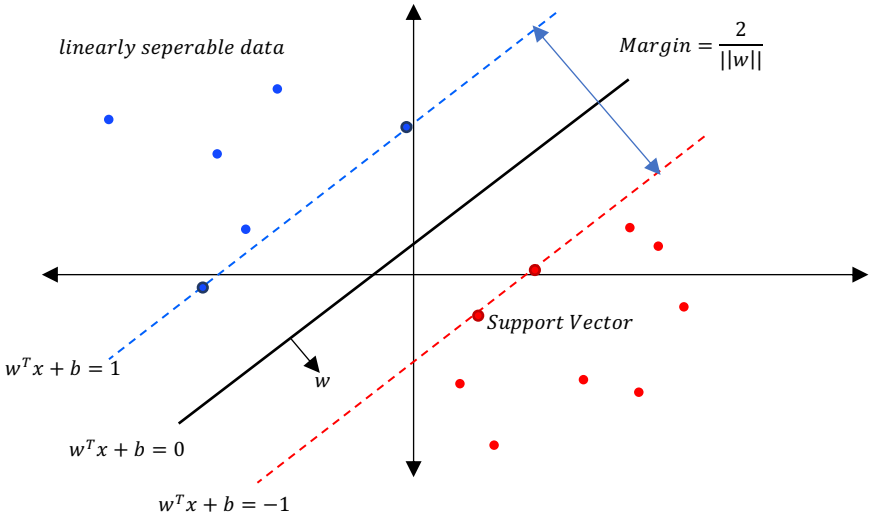


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

รูปภาพที่ 2.3 แสดงการแบ่งแบบเพื่อนบ้านใกล้ที่สุด K ตัว

3.) ซัพพอร์ตเวกเตอร์แมชชีน (Support vector machine Classification)

เป็นเทคนิคที่ใช้ความรู้เรื่องของสมการเส้นตรงเพื่อการแบ่งกลุ่มของข้อมูล ช่วยในการเรียนรู้จดจำรูปแบบและช่วยแก้ปัญหาการจัดกลุ่ม ใช้การหาค่าสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูล ซึ่งมุ่งเน้นที่จะหาเส้นแบ่งแยกกลุ่มที่ดีที่สุด (optimal separating hyperplane) และนิยามค่าขอบเขตข้อมูล (Margin) เป็นผลรวมของระยะห่างสูงสุดที่แบ่งแยกข้อมูลสองชนิดออกจากกัน [14]



รูปภาพที่ 2.4 แสดงการแบ่งเชิงเส้นของซัพพอร์ตเวกเตอร์แมชชีนการเลือกตัวแบ่งแยกทางสถิติที่ทำให้ระยะห่างจาก positive training sample กับ negative training sample ระยะห่างเท่ากัน

ซัพพอร์ตเวกเตอร์แมชชีนสามารถนำมาใช้แก้ไขปัญหาที่มีความซับซ้อนไม่เป็นเชิงเส้น ได้โดยอาศัยการแปลงข้อมูลไปยังอีกปริภูมิหนึ่งที่มีจำนวนมิติมากกว่าเดิมโดยอาศัยเคอร์เนลฟังก์ชัน (Kernel Function) ทำให้แก้ไขปัญหาได้ง่ายยิ่งขึ้น โดยการใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนนี้อย่างประสิทธิภาพได้นั้นต้องมีการปรับและเตรียมข้อมูลอย่างเหมาะสม กล่าวคือการจำแนกข้อมูลบนระนาบหลายมิติ จะใช้ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า โครงสร้างในการคัดเลือกคุณลักษณะ (Feature selection) ซึ่งโครงสร้างในการคัดเลือกคุณลักษณะมาจากข้อมูลที่สอนให้ระบบเรียนรู้ โดยที่จำนวนเซตของโครงสร้างที่ใช้อธิบายในกรณีหนึ่ง เรียกว่า เวกเตอร์ (Vector) ดังนั้นจุดมุ่งหมายของตัวแบบซัพพอร์ตเวกเตอร์แมชชีน คือ แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรเป้าหมายที่อยู่ข้างหนึ่งของระนาบ และกรณีของกลุ่มอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า ซัพพอร์ตเวกเตอร์ (Support Vectors) [15]

4.ต้นไม้ตัดสินใจ (Decision Tree Classification)

เทคนิคที่ใช้ในการสร้างแผนภูมิต้นไม้จากข้อมูล ด้วยกฎในรูปแบบ “ถ้า เงื่อนไข แล้ว ผลลัพธ์” ต้นไม้ตัดสินใจเป็นเทคนิคที่ใช้กันอย่างแพร่หลาย เนื่องจากผลลัพธ์ที่ออกมาง่ายต่อการเข้าใจ เทคนิคต้นไม้ตัดสินใจ จะทำการจำกัดข้อมูลที่เป็นตัวแปรตาม (Dependent Variable) 1 ตัวต่อ 1 แบบ จำลอง ซึ่งหากต้องการทำนายหลายตัว จะต้องทำการสร้างแบบจำลองสำหรับตัวแปรทุกตัว เช่นเดียวกันกับเทคนิคนาอีฟเบย์ เทคนิคต้นไม้ตัดสินใจส่วนใหญ่จะไม่รองรับข้อมูลแบบต่อเนื่อง ดังนั้นจึง ต้องมีการแบ่งข้อมูลให้เป็นแบบไม่ต่อเนื่องก่อน

ต้นไม้ตัดสินใจ หมายถึง ต้นไม้ที่ใช้ในการสนับสนุนการตัดสินใจ ซึ่งมีลักษณะเป็นโครงสร้างต้นไม้หัวกลับที่มีรากอยู่ด้านบนและใบอยู่ด่านล่างสุด โดยที่ภายในต้นไม้ประกอบไปด้วยโหนด (Node) ซึ่งแต่ละโหนดนั้นจะแสดงถึงการตัดสินใจบนข้อมูลของคุณสมบัติต่าง ๆ กิ่งของต้นไม้แสดงถึงค่าหรือผลลัพธ์ที่ได้จากการทดสอบ และใบซึ่งเป็นสิ่งที่อยู่ล่างสุดของต้นไมตัดสินใจจะแสดงถึงกลุ่มของข้อมูล (Class) หรือผลลัพธ์ โหนดที่อยู่บนสุดจะเรียกว่า โหนด ราก (Root Node) [16]

มีขั้นตอนการสร้างต้นไม้ตัดสินใจดังนี้ [17]

1. เลือกแอททริบิวต์ (Attribute) ที่ทำหน้าที่เป็นโหนดราก (Root Node)
2. จาก Root Node สร้างเส้นเชื่อมโยงไปยังโหนดลูก จำนวนเส้นเชื่อมโยงจะเท่ากับจำนวนค่าที่เป็นไปได้ทั้งหมดของ Attribute ที่เป็น root node

3. ถ้าโหนดลูก เป็นกลุ่มของข้อมูลที่อยู่ในคลาสเดียวกันทั้งหมดให้หยุดสร้างต้นไม้ แต่ถ้าโหนดลูกมีข้อมูลของหลายคลาสปะปนกันอยู่ ต้อง สร้างต้นไม้ย่อยเพื่อจำแนกข้อมูลต่อไป โดยเลือก ต้นไม้ย่อยมาทำหน้าที่เป็นโหนดรากของต้นไม้ย่อย มาทำซ้ำในขั้นตอนที่ 2 และ 3

แม้การแบ่งกลุ่มแบบต้นไม้ตัดสินใจนั้นจะไม่มีสมการกำกับความสัมพันธ์ระหว่างคุณลักษณะกับเป้าหมาย แต่ทว่าทุกครั้งที่เกิดการแบ่งแยก (Split) ค่าคุณลักษณะต่าง ๆ จะมีการทำให้ค่า Cost Function น้อยที่สุด (Minimize) โดยค่า Cost Function ที่ใช้ในการแบ่งกลุ่มของต้นไม้ตัดสินใจ ในที่นี้คือ ค่าความไม่บริสทธ์ของจีนี (Gini Impurity) กับค่าเอนโทรปี (Entropy) [18]

ค่าความไม่บริสุทธ์ของจีนี (Gini impurity) เป็นการวัดความไม่บริสุทธิ์ของคลาส ในแต่ละกลุ่มข้อมูลที่แบ่งตามแต่ละจุดแบ่งแยก (Split point) สำหรับปัญหาการแบ่งกลุ่มแบบไบนารี (binary) นั้นการแบ่งแยกที่ดี ควรจะได้กลุ่มข้อมูลออกมา 2 กลุ่มที่สามารถแยก 2 คลาสออกมาได้ชัดเจนในแต่ละกลุ่ม ยิ่งสามารถแบ่งแยกคลาสของเป้าหมายออกมาได้ดี ค่า Gini impurity ก็จะยิ่งต่ำ

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

โดยที่ p_i คือสัดส่วนของตัวอย่างที่อยู่ในคลาส C และ C คือ จำนวนคลาสทั้งหมด

ค่าเอ็นโทปี้เป็นการวัดความไม่แน่นอนของข้อมูล ใช้ในการดูความไม่บริสุทธิ์ของข้อมูล โดยมีสมการการคำนวณดังนี้

$$Entropy = -\sum_{i=1}^{C} p_i log_2 (p_i)$$

โดยที่ p_i คือสัดส่วนของตัวอย่างที่อยู่ในคลาส C และ C คือ จำนวนคลาสทั้งหมด

2.1.2.5 การรู้จำและการประเมินแบบจำลอง

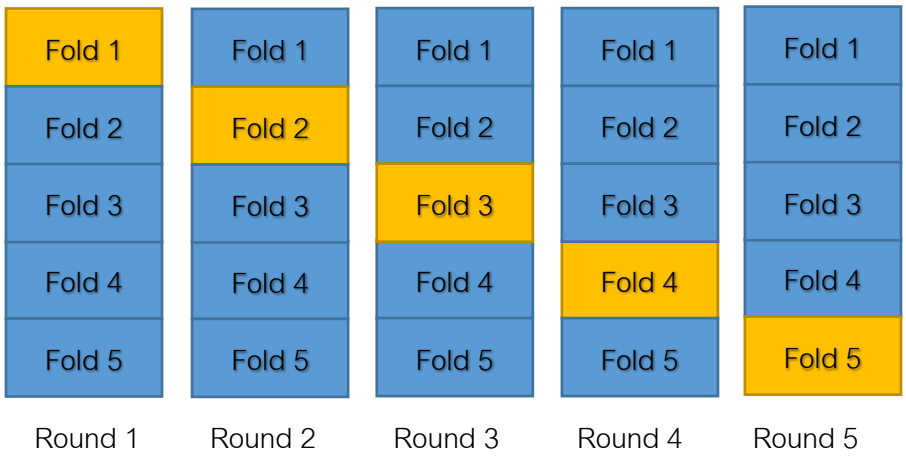
ก่อนที่จะทำการนำข้อมูลไปฝึกฝนในแบบจำลอง จำเป็นจะต้องมีการแบ่งข้อมูลสำหรับการรู้จำเพื่อใช้ในการประเมินประสิทธิภาพของแบบจำลอง โดยมี 3 วิธีหลักดังนี้ [20]

1. Self Consistency Test หรือ Use Training Set นี้เป็นวิธีการที่ง่ายที่สุด กล่าวคือข้อมูลที่ใช้ในการสร้างแบบจำลอง (model) และข้อมูลที่ใช้ในการทดสอบเป็นข้อมูลชุดเดียวกัน กระบวนการนี้เริ่มด้วยการสร้างแบบจำลองด้วยข้อมูลสำหรับรู้จำ (training data) หลังจากนั้นนำแบบจำลองที่สร้างไปทำนายข้อมูลสำหรับรู้จำชุดเดิม การวัดประสิทธิภาพด้วยวิธีนี้จะให้ผลการวัดประสิทธิภาพที่มีค่าสูง เนื่องจากเป็นการการนำข้อมูลชุดเดิมที่ระบบได้ทำการเรียนรู้มาแล้วมาวัดผล ผลการวัดด้วยวิธีนี้จึงไม่เหมาะสำหรับรายงานผลในงานวิจัย แต่เหมาะสำหรับใช้ในการทดสอบประสิทธิภาพเพื่อดูแนวโน้มของแบบจำลองที่สร้างขึ้น ถ้าได้ผล

การวัดที่น้อย แสดงว่าแบบจำลองไม่เหมาะสมกับข้อมูล จึงไม่ควรจะนำไปทดสอบด้วยวิธีการแบ่งข้อมูลแบบต่าง ๆ

2. Split Test เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น อัตราส่วน 70 ต่อ 30 หรือ 80 ต่อ 20 โดยข้อมูลส่วนที่หนึ่งจะใช้ในการสร้างแบบจำลองและข้อมูลส่วนที่สอง (ส่วนที่น้อยกว่า) จะใช้ในการทดสอบประสิทธิภาพของแบบจำลอง การทดสอบแบบ Split Test นี้ทำการสุ่มข้อมูลเพียงครั้งเดียวซึ่งในบาง กรณีถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะคล้ายกับข้อมูลที่ใช้สร้างแบบจำลอง จะส่งผลทำให้ผลการวัดประสิทธิภาพออกมาดี ในทางตรงกันข้ามถ้าการสุ่มข้อมูลที่ใช้ในการทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้สร้างแบบจำลองมาก จะทำให้ผลการวัดประสิทธิภาพออกมาแย่ ดังนั้นจึงควรใช้วิธีแบ่งข้อมูลนี้ ในการการสุ่มหลาย ๆ ครั้ง วิธีนี้มีข้อดีคือใช้เวลาในการสร้างแบบจำลองน้อยซึ่งเหมาะกับชุดข้อมูลที่มีขนาดใหญ่

3. Cross-Validation Test เป็นวิธีนี้เป็นวิธีที่นิยมในการทำงานวิจัย เพื่อใช้ในการทดสอบประสิทธิภาพของแบบจำลองเนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธี Cross-validation นี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน เป็นจำนวน k ตัว เช่น 5-fold cross-validation คือ ทำการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หรือ 10-fold cross-validation คือ การแบ่งข้อมูลออกเป็น 10 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลัง จากนั้นข้อมูลหนึ่งส่วนจะใช้เป็นตัวทดสอบประสิทธิภาพของแบบจำลอง ทำซ้ำเช่นนี้จนครบจำนวนที่แบ่งไว้ k ส่วน



รูปภาพที่ 2.5 แสดงการประเมินผลแบบ 5-fold Cross validation โดยที่กล่องสีฟ้าคือ ข้อมูลสำหรับฝึกฝนและกล่องสีเหลืองคือข้อมูลสำหรับทดสอบ การรู้จำจะทำการวนรอบการเรียนรู้และทดสอบไปจนครบจำนวนของ k ที่กำหนดไว้

2.2 งานวิจัยที่เกี่ยวข้อง

จากงานวิจัยของ Minsu Park และคณะ [20] ซึ่งได้ทำการศึกษาภาวะอารมณ์ต่าง ๆ ของผู้ที่เป็นโรคซึมเศร้าผ่านทวิตเตอร์ ด้วยการสำรวจดูการใช้ภาษาที่อธิบายภาวะโรคซึมเศร้าโดยเริ่มจากการสุ่มตัวอย่างจากผู้ใช้ทวิตเตอร์เป็นเวลา 2 เดือน เพื่อลักษณะรูปแบบการพูดถึงโรคซึมเศร้า ผลปรากฏว่ามีข้อมูลรายละเอียดเกี่ยวกับอารมณ์ซึมเศร้า สถานะความซึมเศร้า รวมถึงประวัติการรักษาภาวะโรคซึมเศร้าแสดงออกมาผ่านข้อความนั้น ๆ อย่างชัดเจน และงานวิจัยนี้ทำการศึกษาจากกลุ่มผู้เข้าร่วมจำนวน 69 คน เพื่อศึกษาความแตกต่างของการใช้กับผู้ใช้ทั่วไป ผลลัพธ์ที่ได้คือพบความสัมพันธ์ระหว่างการใช้คำกับอารมร์ในเชิงลบและอารมณ์โกรธท่ามกลางผู้ใช้ที่มีภาวะโรคซึมเศร้าเพิ่มขึ้นอย่างมีนัยสำคัญ อย่างไรก็ตามไม่พบความแตกต่างในการใช้คำที่มีความสัมพันธ์กับอารมณ์ในเชิงบวกของทั้งสองกลุ่ม

งานวิจัยนี้ได้ทำการแยกหมวดหมู่ของข้อความที่พูดถึงภาวะโรคซึมเศร้า โดยแยกเป็น 5 หมวด ดังนี้

1.หมวดความรู้สึกจากภาวะซึมเศร้าของผู้ใช้
2.หมวดการส่งต่อข้อมูลข่าวสารเกี่ยวกับภาวะโรคซึมเศร้า
3.หมวดการแสดงความคิดเห็นที่เกี่ยวข้องกับภาวะโรคซึมเศร้า
4.หมวดอื่น ๆ
5.หมวดเกี่ยวกับภาวะโรคซึมเศร้าของผู้อื่น

ผลสรุปคือ ภาวะโรคซึมเศร้าถูกพบในกลุ่มของข้อความบ่งบอกสถานะจริง ๆ ของผู้ใช้งานถึง 42.40 เปอร์เซ็นต์จากข้อความของผู้เข้าร่วม กล่าวคือ มี 113 ข้อความที่แสดงถึงการรายงานการเข้ารับการรักษา 3 ข้อความที่ใช้สื่อว่าผู้ใช้คนนั้นไม่ใช่โรคซึมเศร้า มี 216 ข้อความที่บอกถึงความรู้สึกซึมเศร้าของตน โดยพบว่ามีการบอกรายละเอียดเชิงลึกของสาเหตุ รูปแบบ และเหตุการณ์ที่ทำให้เกิดภาวะโรคซึมเศร้าของผู้เข้าร่วม ทั้งนี้ข้อความของผู้เข้าร่วมเหล่านั้นมีการเปิดเผยข้อมูลส่วนตัวของตน เช่น ประวัติการใช้ยา ประวัติการเข้ารับการรักษา และแนะนำวิธีการรักษาของตนแก่ผู้อื่นอีกด้วย

งานวิจัยของ Nikhita และ Srinivasan[6] ศึกษาเกี่ยวกับความหมายทางอารมณ์และภาษาศาสตร์ของภาวะโรคซึมเศร้าจากสื่อสังคมออนไลน์ โดยได้ทำการดึงลักษณะเฉพาะออกมาจากข้อความของผู้เข้าร่วม ซึ่งสรุปได้ 7 ข้อดังนี้

1.รูปแบบกิจกรรมที่ลดลง
2.พฤติกรรมการเข้าถึงสื่อสังคมออนไลน์ในยามดึก
3.การมีส่วนร่วมกับเครือข่ายน้อยลง
4.การเพิ่มขึ้นของความรู้สึกแง่ลบ
5.การใช้คำสรรพนามที่แทนตนเอง
6.มีการกระจุกตัวของกลุ่มที่มีภาวะโรคซึมเศร้า
7.การแลกเปลี่ยนอิทธิพลต่อเครือข่ายรอบข้างน้อย

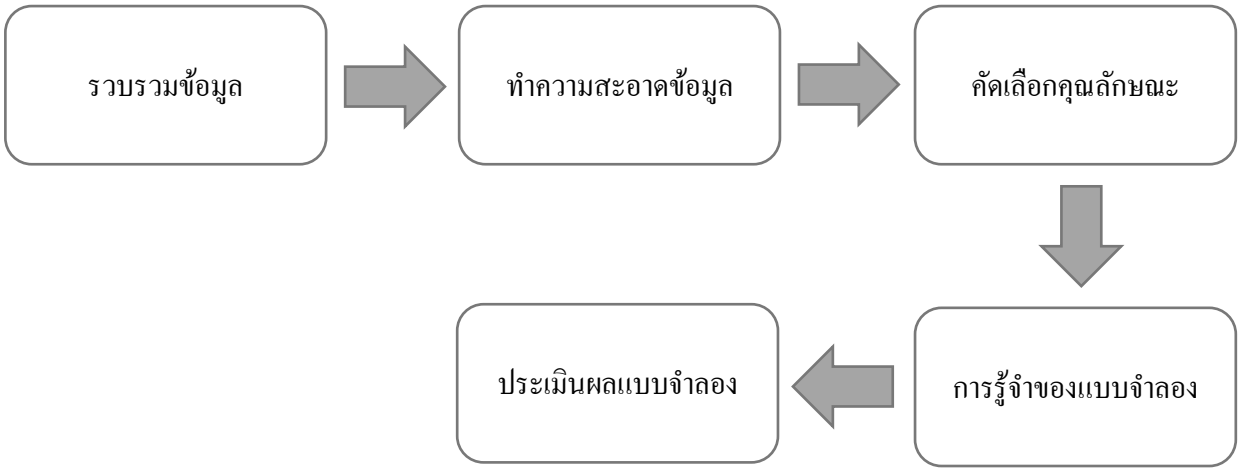
ในงานวิจัยนี้มุ่งที่จะทำการสร้างแบบจำลองที่จะทำนายถึงผู้ใช้งานที่มีแนวโน้มจะเป็นโรคซึมเศร้าทางการแพทย์ในอนาคต นอกจากนั้นยังมีการนำเสนอทฤษฎีจาก [22] ที่กล่าวว่ากลุ่มผู้มีภาวะซึมเศร้ามักจะรวมกลุ่มอยู่ใกล้กัน โดยทำการทดสอบได้ผลเปอร์เซ็นต์ของการมีปฏิกริยาตอบกลับของเครือข่ายรอบข้างพบว่ามากกว่าครึ่งหนึ่งของสัดส่วนผู้เข้าร่วมที่มีภาวะซึมเศร้า จะมีปฏิกิริยาตอบกลับจากเครือข่ายน้อยกว่า 1 เปอร์เซ็นต์ และอีกหลายการทดลองที่ให้ผลสอดคล้องกับทฤษฎีที่ส่งผลต่อการเลือกทั้ง 7 คุณลักษณะที่กล่าวไปข้างต้น

ในส่วนของการสร้างแบบจำลองเพื่อทำนายในงานวิจัยนี้ ทำการสร้างโดยใช้แบบจำลองแบ่งกลุ่มต้นไม้ตัดสินใจแบบ Gradient Boosted พร้อมกับ ตัวแบ่งประเมินผลแบบ 5-fold Cross Validation บนทั้งชุดข้อมูลของผู้ใช้งาน โดยการประเมินประสิทธิภาพจะใช้ค่าความผู้ต้องแม่นยำของเมทริกซ์พื้นฐาน ร่วมกับค่าคะแนน F1 ประสบความสำเร็จได้ ค่าความถูกต้อง 0.9 สำหรับกลุ่มข้อมูลผู้มีภาวะโรคซึมเศร้า และ 0.87 สำหรับกลุ่มผู้ใช้ปกติ เมื่อเจาะลึกลงไปพบว่ามากการทำทายผิดพลาดไปเพียง 5 ผู้ใช้งานที่มีภาวะโรคซึมเศร้าที่ถูกทำนายว่าเป็นผู้ใช้งานปกติ ยิ่งไปกว่านั้นมีการพบว่าในกลุ่มผู้ใช้ปกติ แม้ว่าจะไม่มีการสื่อชัดเจนถึงภาวะโรคซึมเศร้า แต่มีการตรวจพบการใช้คำศัพท์ที่มีขั้วอารมณ์ในเชิงลบ อย่างความโกรธและความรุนแรงอยู่เช่นกัน

บทที่ 3
วิธีการดำเนินการศึกษา

วิธีการศึกษาจะทำการวิจัยบนชุดข้อมูลทวิตเตอร์ภาษาไทย ที่ทำการฝึกฝนรู้จำ ด้วย 4 แบบจำลอง ได้แก่

1. แบบจำลองนาอีฟเบย์ (Naïve Bayes)
2. แบบจำลองต้นไม้ตัดสินใจ (Decision Tree)
3. แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
4. แบบจำลองการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว (K-Nearest neighbor)



รูปภาพที่ 3.1 แสดงกระบวนการการดำเนินงานศึกษาวิจัย

3.1 การรวบรวมข้อมูลและการระบุชนิดของข้อความ

การรวบรวมข้อมูลคำที่บ่งบอกความรู้สึกในงานวิจัยนี้ จะทำการรวบรวมในช่วงเวลา 4 เดือน คือตั้งแต่เดือนพฤศจิกายนถึงเดือนกุมภาพันธ์ ข้อมูลจากทวิตเตอร์ภาษาไทยจะถูกรวบรวมผ่านการค้นหาด้วยแฮชแท็ก เพื่อเป็นการแบ่งกลุ่มความคิดเห็นออกมาอย่างคร่าวๆ เช่น #โรคซึมเศร้า หรือ #ความสุข เป็นต้น ชุดข้อมูลทั้งสองที่จะนำมาศึกษาวิจัยนั้นจะแบ่งออกมาในรูปของ ชุดข้อมูลที่บ่งบอกภาวะซึมเศร้า (Depressed Set) และ ชุดข้อมูลที่ปกติ (Normal Set)

ชุดข้อมูล	**จำนวนข้อความ**
Normal Set	6,407
Depressed Set	8,165
รวม	14,572

ตารางที่ 3.1 แสดงจำนวนและกลุ่มของข้อมูลที่ทำการรวบรวม

ทำการรวบรวมข้อมูลจาก API Twitter ด้วย Tweepy เมื่อรวบรวมออกมาจะพบว่าได้กรอบข้อมูลที่มี 6 คอลัมน์ และมีการระบุชนิด (Label) เป็น 2 ชุดคือ ชุดปกติ (Normal) กับชุดซึมเศร้า (Depressed) เพิ่มเติมหลังจากนี้ด้วย

ข้อมูลประกอบไปด้วย 8 คอลัมน์ ได้แก่

-Text เป็นส่วนที่เก็บข้อความทวิตเตอร์

-Token เก็บข้อมูลการตัดคำ (Token) จากข้อความทวิตเตอร์

-Follower เก็บข้อมูลตัวเลขของจำนวนผู้ที่มาติดตาม

-Following เก็บข้อมูลตัวเลขของจำนวนคนที่ผู้ใช้กำลังติดตาม

-Total Tweet เก็บข้อมูลตัวเลขของจำนวนข้อความที่ผู้ใช้ได้โพสต์

-Retweet เก็บข้อมูลตัวเลขของจำนวนที่ข้อความถูกรีทวีต (Retweet)

-Tweet Created เก็บข้อมูลระบุช่วงเวลาที่ข้อความถูกโพสต์

-Label เก็บตัวระบุชนิดของชุดข้อความ ได้แก่ dep คือ ซึมเศร้า และ nor คือ ปกติ

3.2 การทำความสะอาดข้อมูล

ภาษาแต่ละภาษามีรายละเอียดและความซับซ้อนแตกต่างกันไปในแต่ละแบบ ภาษาไทยมีลักษณะแตกต่างจากภาษาอังกฤษ หรือภาษาจีน เนื่องจากในภาษาไทยมีการเขียน ติดกันไปทั้งประโยค อีกทั้งคำไทยหนึ่งคำอาจประกอบไปด้วยสระที่เป็นสระประกอบ คือมาจาก สระอื่นอีกหลายตัวประกอบกัน และพยัญชนะบางตัวยังสามารถทำหน้าที่ตัวสะกดได้ ทั้งนี้ข้อความภาษาไทยในสื่อออนไลน์ก็มักมีความผิดเพี้ยนด้านการสะกดคำ คำศัพท์ที่หลากหลายหรือมีความหมายแฝง การใช้คำสรรพนาม หรือการใช้ตัวอักษรซ้ำ ๆ เพื่อสื่ออารมณ์ ในงานวิจัยนี้จะใช้โมดูล (Module) ของ PyThaiNLP และ NLTK มาช่วยในการตัดคำและทำความสะอาดชุดข้อมูลจากทวิตเตอร์ อีกทั้งยังมีการใช้เทคนิคด้าน (Regular expression) ในการตัดคำที่มีรายละเอียดเป็นพิเศษเช่นกัน [21]

ส่วนแรกจะทำการลบส่วนที่ซ้ำกันของข้อความทวิตเตอร์ที่รวบรวมมา เช่น ข้อความที่ถูกรีทวิตซ้ำ ๆ หรือข้อความที่ถูกคัดลอกมาโพสต์ และส่วนที่ไม่เกี่ยวข้องออกจากชุดข้อมูล เช่น ลิ้งค์ URL แฮชเท็ก (Hashtag) เครื่องหมายคำพูด อิโมติคอน (Emoticon) และเครื่องหมายต่าง ๆ ก็จะทำการลบออกจากชุดข้อมูล ตัวเลข รวมถึงตัวอักษรภาษาอังกฤษออกไปจะไม่เป็นประโยชน์ต่อการวิเคราะห์ จึงทำการลบออกเพื่อลดความยุ่งยากในการวิเคราะห์ข้อมูลเช่นกัน ต่อมาเมื่อทำความสะอาดส่วนแรกไปแล้ว จึงทำการตัดคำ (Tokenize) เพื่อได้คำศัพท์แต่ละคำจากข้อความนั้น ๆ เหมาะสำหรับการรู้จำ และการวิเคราะห์รูปแบบการใช้คำของแต่ละกลุ่มข้อมูลได้อย่างดี

```
original text:
 ผมหยุดเสียเวลาอธิบาย ความคิดตัวเอง
เพราะ ธรรมชาติของคน เขาจะเข้าใจเรื่องที่เขา อยากจะเข้าใจมันเท่านั้น..

#ข้อคิดดีๆ #ข้อคิดชีวิต #ข้อคิดเตือนใจ #ข้อคิด #คำคมความรู้สึก #คำคมโดนๆ #คำคม #แรงบันดาลใจ #โรคซึมเศร้า #ความรู้สึกล้วนๆ #ความรัก #ArtAekkaphob..
 this is 🅰rt🤘
clean text:
 ผมหยุดเสียเวลาอธิบาย ความคิดตัวเอง เพราะ ธรรมชาติของคน เขาจะเข้าใจเรื่องที่เขา อยากจะเข้าใจมันเท่านั้น ข้อคิดดีๆ ข้อคิดชีวิต ข้อคิดเตือนใจ ข้อคิด คำคมความรู้สึก คำคมโดนๆ คำคม แรงบันดาลใจ โรคซึมเศร้า ความรู้สึกล้วนๆ ความรัก ArtAekkaphob this is rt
```

รูปภาพที่ 3.2 แสดงตัวอย่างของผลลัพธ์หลังการทำความสะอาดข้อความส่วนแรก

3.3 การคัดเลือกคุณลักษณะ

จากการทบทวนวรรณกรรมต่าง ๆ และการสำรวจข้อมูลที่รวบรวมมานั้น พบว่าจะทำการเลือกศึกษาไปที่คุณลักษณะ 7 ประการคือ 1.คลังคำศัพท์ 2.จำนวนการถูกติดตาม 3.จำนวนการไปติดตามผู้อื่น 4.จำนวนการโพสต์ข้อความโดยรวม 5.ช่วงเวลาในการโพสต์ข้อความ 6.จำนวนการปรากฏของคำสรรพนามแทนคนเอง 7.จำนวนการถูกรีทวีต

คุณลักษณะที่สำคัญสำหรับการวิเคราะห์อารมณ์ความรู้สึก คือผ่านทางข้อความและคำที่สื่อความหมาย การรวบรวมการปรากฏของคำศัพท์นั้น ๆ ในชุดข้อมูลนั้น เพื่อศึกษารูปแบบแนวทางการใช้คำศัพท์ของแต่ละกลุ่ม ในที่นี้ได้ทำการสร้างคลังที่นับจำนวนความถี่ของคำศัพท์ที่พบในแต่ละกลุ่มมาเปรียบเทียบกัน จากการศึกษาถึงคำศัพท์ที่มีผลต่อการสื่อถึงภาวะโรคซึมเศร้านั้นชี้ว่า คำส่วนใหญ่ที่พบในชุดข้อมูลซึมเศร้านั้นมักเป็นคำในแง่ลบ ขั้วของอารมณ์นั้นเป็นในเชิงลบและมีความคิดเห็นหมกมุ่นเกี่ยวกับความตายอย่างที่งานวิจัย [9] ได้กล่าวไว้ โดยพบคำว่า ตาย ถึง 908 ครั้งในกลุ่มซึมเศร้า อย่างไรก็ตามแม้กลุ่มปกติจะมีการใช้คำศัพท์บางคำที่เป็นในเชิงลบเหมือนกับกลุ่มซึมเศร้า แต่จำนวนการพบนั้นแตกต่างและน้อยกว่ากลุ่มซึมเศร้าอย่างชัดเจน

เนื่องจากจากงานวิจัย [6] ได้กล่าวไว้ว่า ผู้ที่มีภาวะโรคซึมเศร้ามักจะใช้คำสรรพนามแทนตนเองมากกว่าคำสรรพนามที่กล่าวถึงผู้อื่น โดยมักจะสนใจเกี่ยวกับตนเองและสิ่งที่เกิดขึ้นกับตนเองมากกว่าสนใจสังคมรอบข้าง ในที่นี้จะทำการตรวจสอบจากแอททริบิวต์ของคำศัพท์ (Token) ต่อชุดของคำสรรพนามแทนตนเองที่เตรียมไว้ โดยคำในชุดข้อมูลได้แก่ ฉัน ผม ตนเอง ตัวเอง เรา ชั้น เค้า เป็นต้น เมื่อทำการตรวจสอบและนับจำนวนครั้งที่คำสรรพนามเหล่านั้นปรากฏ ก็ทำการเพิ่มแอททริบิวต์ที่ชื่อว่า Self Pronoun ขึ้นเพื่อนำสู่การรู้จำต่อไป

Token	Number of Document in Nor	Number of Document in Dep
ยิ้ม	812	204
ดีใจ	381	33
ความสุข	346	508
รัก	330	235
ชอบ	295	310
ตนเอง	291	1025
ชีวิต	257	571
น่ารัก	235	16
รอยยิ้ม	197	5
ขอบใจ	184	97
เค้า	168	214
บ้าน	160	147
กำลังใจ	154	129
เพื่อน	152	315
แม่	149	239
เย้	147	3
ฟิลลิ่ง	141	0
ความรัก	140	77
สู้	133	208
เหนื่อย	132	685

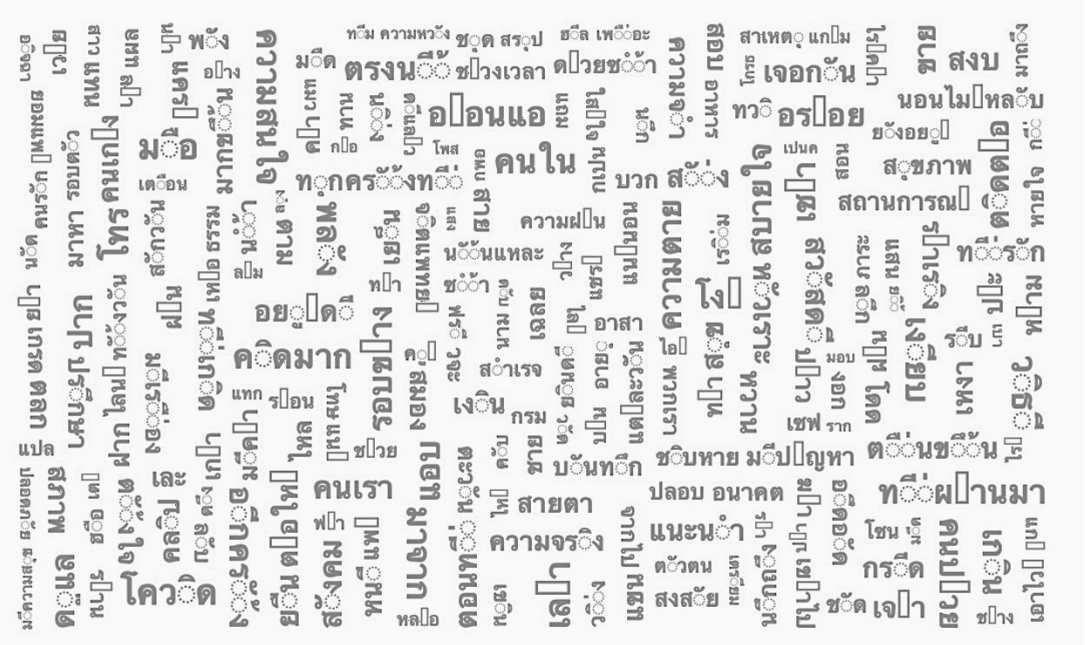
ตารางที่ 3.2 ตารางแสดงการปรากฏของคำที่มากที่สุด 20 อันดับของชุดข้อมูลปกติ (Normal Set)

Token	Number of Document in Dep	Number of Document in Nor
ตนเอง	1025	291
ตาย	908	82
โรคซึมเศร้า	799	0
เหนื่อย	685	132
ชีวิต	571	257
ความสุข	508	346
รู้สึก	497	184
หาย	453	83
ร้องไห้	395	49
คนอื่น	349	97
แย่	339	59
โลก	333	92
เพื่อน	315	152
ไหว	309	74
คนเดียว	295	64
นอน	288	86
โรค	270	39
ครอบครัว	263	53
หมอ	254	45
เจ็บ	248	75

ตารางที่ 3.3 ตารางแสดงการปรากฏของคำที่มากที่สุด 20 อันดับของชุดข้อมูลที่มีภาวะซึมเศร้า (Depressed Set)

การวิเคราะห์คุณลักษณะเพื่อศึกษาการมีปฏิสัมพันธ์กับผู้อื่น สามารถใช้ข้อมูลจากแอททริบิวต์ต่าง ๆ เช่น จำนวนการถูกรีทวีต (Retweet) จำนวนการติดตามผู้ใช้รายอื่น (Following) หรือการถูกติดตามโดยผู้อื่น (Follower) ร่วมกันเพื่อผลของการจำแนกแบ่งกลุ่มที่แม่นยำและน่าเชื่อถือ โดยจากงานวิจัย [6] กล่าวถึงการศึกษาการมีส่วนร่วมต่อสื่อสังคมออนไลน์และการมีปฏิกิริยาโต้ตอบระหว่างผู้ใช้งานกับเครือข่ายรอบข้าง วิเคราะห์การมีอิทธิพลของผู้ใช้งานในเครือข่ายนั้น ๆ เช่น ภาวะซึมเศร้าของผู้ใช้งานสามารถมีอิทธิพลกลุ่มคนรอบข้างได้หรือไม่ การโพสต์ข้อความนั้น ๆ ส่งผลให้เกิดพฤติกรรมการสนับสนุนมากน้อยเพียงใด ผ่านการแชร์ข้อความ โดยฟังก์ชั่นรีทวิต ทั้งนี้ยังสามารถพบรูปแบบของขนาดสังคมเครือข่ายของผู้ใช้ ผ่านการถูกติดตามและการติดตามผู้อื่นอีกด้วย

จำนวนการโพสต์ข้อความทั้งหมด (Total Tweet) เป็นอีกหนึ่งแอททริบิวต์ที่บอกลักษณะเฉพาะของแต่ละกลุ่มข้อมูลออกมาได้ ผลรวมของจำนวนการโพสต์ข้อความ มักจะสัมพันธ์กับจำนวนกิจกรรมในสังคมออนไลน์ที่มีมากน้อยตามแต่ละผู้ใช้งาน รวมถึงตามแต่ละกลุ่ม

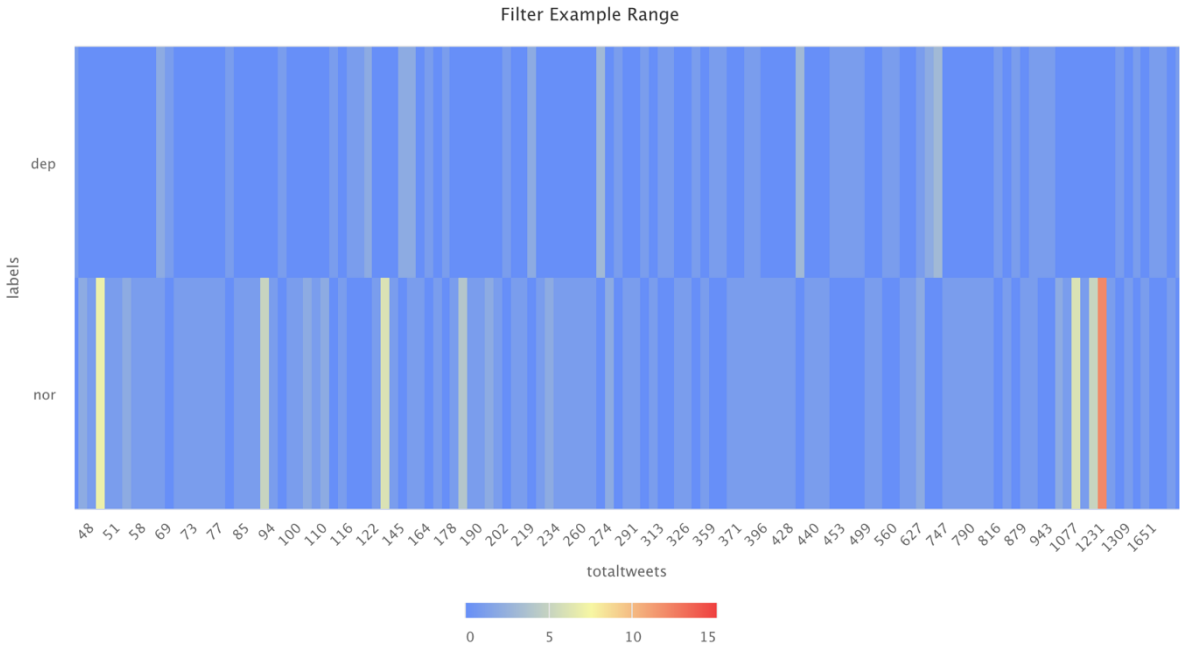


รูปภาพที่ 3.3 WordCloud ที่แสดงคำตามขนาดเมื่อเทียบจำนวนการปรากฏ

ความคิดที่วนเวียนกับความตาย ความสิ้นหวัง ความเศร้า หรือความโดดเดี่ยวเป็นอาการทั่วไปพบได้บ่อยในผู้ที่มีภาวะโรคซึมเศร้า นอกจากนี้ยังพบว่ามีการใช้คำศัพท์ในแง่ลบมากกว่ากลุ่มปกติอีกด้วย จากภาพที่ 3.3 แสดงคำศัพท์ที่การใช้บ่อยกว่าคำอื่น ๆ ได้แก่ อ่อนแอ ความตาย คิดมาก คนป่วย ความสนใจ โง่ หรือ เงียบ เป็นต้น

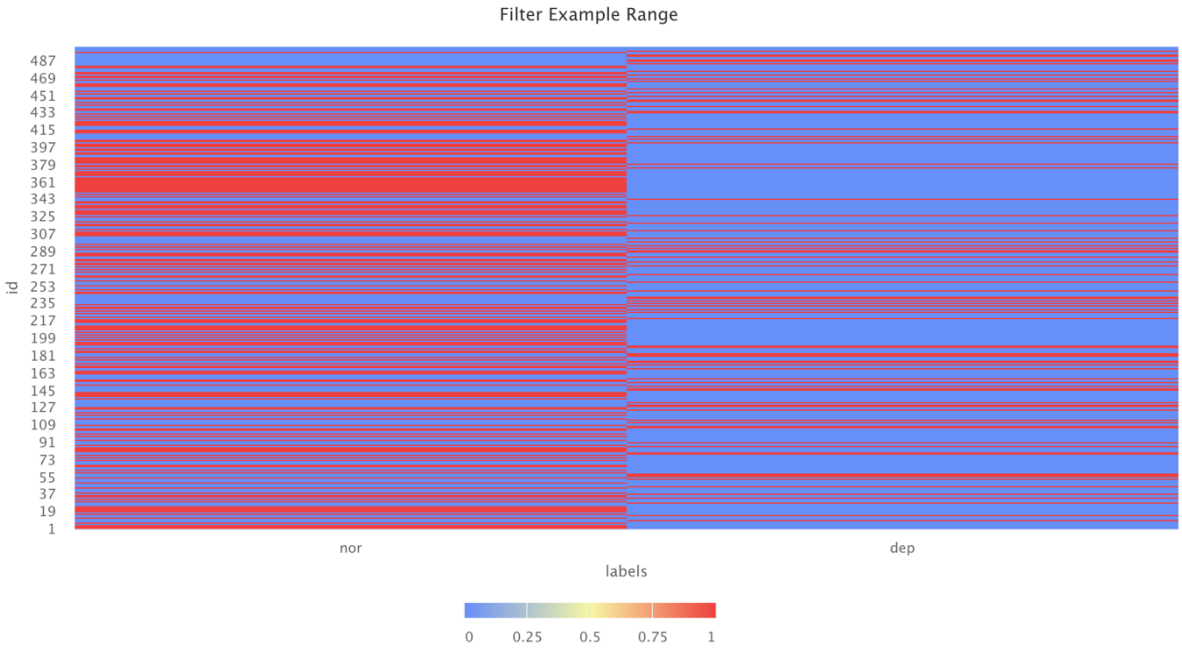
รูปภาพที่ 3.4 แสดงการกระจายตัวของจำนวนการติดตามผู้อื่น (Follower) ของผู้ใช้นั้น ๆ

การติดตามผู้อื่น (Following) และการถูกติดตามโดยผู้อื่น (Follower) ทำให้เห็นถึงการมีส่วนร่วมต่อชุมชนของผู้ใช้นั้น ๆ หลักการที่กล่าวว่าผู้มีภาวะซึมเศร้าส่วนมากมักจะมีปฏิสัมพันธ์กับคนรอบข้างน้อย หรือแยกตัวออกจากสังคมนั้นสามารถแสดงออกมาผ่านรูปแบบการมีปฏิสัมพันธ์กับชุมชนออนไลน์เช่นกัน แต่การที่มีการติดตามผู้อื่นหรือการถูกติดตามโดยผู้อื่นจำนวนมากน้อยก็ไม่ได้เป็นตัวชี้วัดถึงภาวะโรคซึมเศร้าที่น่าเชื่อถือเสมอไป เนื่องจากข้อมูลที่งานวิจัยนี้ทำการรวบรวมเพื่อศึกษามีการพบว่าจำนวนการถูกติดตามของกลุ่มซึมเศร้าบางรายกลับมีจำนวนสูงกว่ากลุ่มปกติ โดยสามารถเกิดขึ้นได้แล้วกรณีของแต่ละบุคคล อย่างไรก็ตามการสำรวจจำนวนยอดผู้ติดตามกลับสามารถทำให้พบความแตกต่างบางอย่างจากชุดข้อมูลทั้ง 2 อย่างดี จากภาพที่ 3.4 การกระจายตัวของจำนวนผู้ที่กำลังติดตาม (Following) ในกลุ่มซึมเศร้าหรือสีแดง มีจำนวนการติดตามผู้อื่นเฉลี่ยอยู่ที่ 96.61 คน โดยมีการกระจายตัวต่ำเมื่อเทียบกับค่าเฉลี่ยของกลุ่มปกติ คือ 334.02 คน



รูปภาพที่ 3.5 แสดงแผนภาพความร้อน (Heatmap) ของจำนวนข้อความรวมที่มีการโพสต์ (Total tweet) ส่วนด้านบนเป็นกลุ่มซึมเศร้า ด้านล่างเป็นส่วนของกลุ่มปกติ

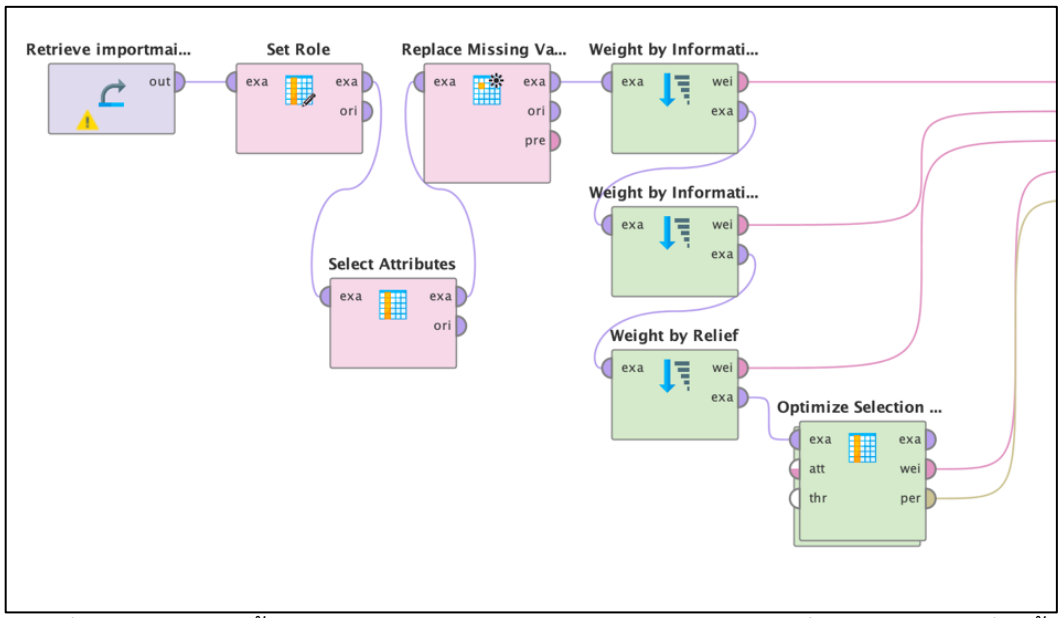
จากภาพที่ 3.5 แสดงแผนภาพความร้อนพบว่าจำนวนข้อความโดยรวมของกลุ่มซึมเศร้ามีค่าที่น้อยมาก คือเป็นสีฟ้าเข้มถึงเหลืองอ่อน เมื่อเทียบกับกลุ่มปกติจะมีจำนวนการโพสต์ข้อความโดยรวมมากกว่า คือเป็นสีฟ้าอ่อนจนถึงส้มเข้มอย่างเห็นได้ชัด จำนวนข้อความโดยรวมที่มีการโพสต์มีผลในการดูพฤติกรรมทางออนไลน์ของผู้ใช้งาน



รูปภาพที่ 3.6 แสดง แผนภาพความร้อนของจำนวนการถูกรีทวีตของทั้งสองกลุ่ม

จำนวนการรีทวีต แสดงให้เห็นถึงการมีส่วนร่วมและการมีอิทธิพลกับคนรอบข้างของผู้ใช้งานที่มีต่อเครือข่าย จากภาพเห็นได้ชัดว่ากลุ่มปกติมีจำนวนครั้งการถูกรีทวีตสูงกว่ากลุ่มซึมเศร้า คือได้ถูกมีการนำข้อความของผู้ใช้นั้น ๆ ไปกระจายต่อ การริทวิตเป็นเหมือนการแสดงความเห็นด้วย โดยการแชร์ข้อความนั้นออกไป

เนื่องจากมีหลายคุณลักษณะที่ทำการศึกษา จึงมีการตรวจสอบดูความเหมาะสมของคุณลักษณะเพื่อใช้ในการตัดสินในเลือกคุณลักษณะ โดยใช้โปรแกรม Rapidminer ด้วยตัวดำเนินการ (Operator) คัดเลือกคุณลักษณะ (Feature Selection) เพื่อการวัดค่าน้ำหนักความสัมพันธ์เกี่ยวข้อง คือ ค่า Weight Information Gain ค่า Weight Information Gain Ratio และค่า Weight Relief



รูปภาพที่ 3.7 แสดงการใช้ Rapidminer และตัวดำเนินการต่าง ๆในการหาค่าคะแนนความเกี่ยวข้อง

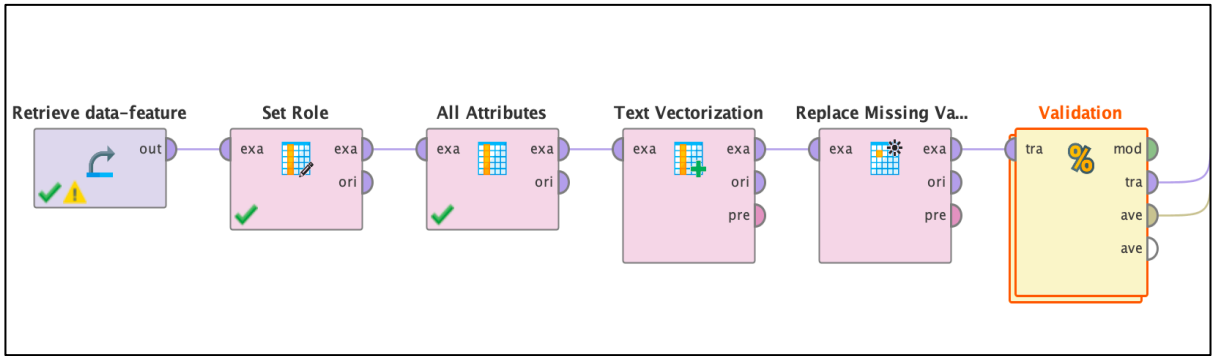
Information Gain		Gain Ratio		Relief	
Attribute	Weight	Attribute	Weight	Attribute	Weight
1.Text	0.977	1.Retweet	0.183	1.Totaltweet	3.543
2.Token	0.968	2.Follower	0.122	2.Retweet	1.226
3.TweetCreated	0.896	3.Totaltweet	0.082	3.Following	0.450

ตารางที่ 3.4 แสดงค่าน้ำหนักของความเกี่ยวข้องของแต่ละคุณลักษณะ โดยที่ค่าที่มากก็จะบ่งบอกถึงประสิทธิภาพที่ดีของคุณลักษณะที่มากเช่นกัน

จากตารางข้างต้นพบว่าคุณลักษณะที่มีค่าน้ำหนักของ Information Gain ดีที่สุด 3 อันดับ ได้แก่ คุณลักษณะของข้อความ (Text) คำศัพท์ (Token) และช่วงเวลาที่มีการโพสต์ข้อความ (Tweet created) ซึ่งเหมาะสมดีเนื่องจาก Information Gain จะตรวจดูคุณลักษณะที่มีความหมาย ในส่วนของค่าน้ำหนัก Gain ratio พบว่าคุณลักษณะที่มีประสิทธิภาพสูงกว่าคุณลักษณะอื่น คือ จำนวนการรีทวีต (Retweet) จำนวนผู้ติดตาม (Follower) กับจำนวนข้อความทั้งหมดที่ถูกโพสต์ (Total Tweet) ผลจากการหาค่าน้ำหนักแบบ Relief ให้ผล 3 คุณลักษณะที่ดีเช่นกันได้แก่ จำนวนข้อความทั้งหมดที่ถูกโพสต์ (Total Tweet) จำนวนการรีทวีต (Retweet) และจำนวนผู้ที่ผู้โพสต์กำลังติดตาม (Following)

3.4การรู้จำและการประเมินผล

ในส่วนของการรู้จำ ในงานวิจัยนี้ได้มีการทดลองปรับค่าต่าง ๆ ในแบบจำลองและปรับค่าตัวประเมินผลแบบจำลองในรูปแบบต่าง ๆ เพื่อให้ได้ผลค่าความแม่นยำที่ดีที่สุด

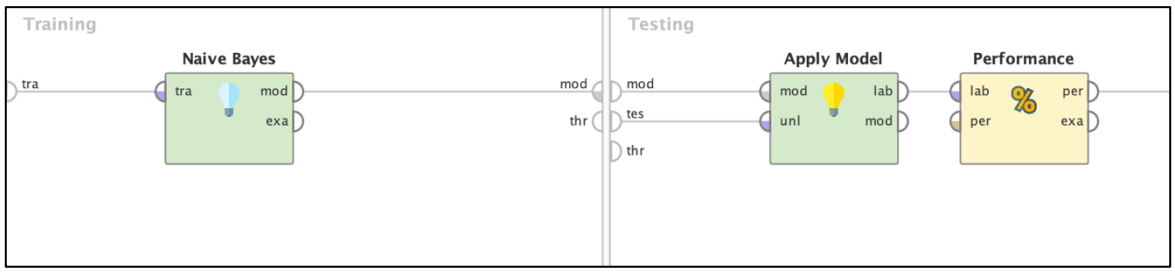


รูปภาพที่ 3.8 แสดงการดำเนินการรู้จำโดยดำเนินการ (Operator) ต่าง ๆใน Rapidminer

ขั้นตอนในการรู้จำและประเมินผลแบบจำลอง โดยใช้เครื่องมืออย่าง Rapidminer เริ่มต้นด้วยการนำเข้าข้อมูลเป็นไฟล์ซีเอสวี (CSV) โดยมีทั้งหมด 9 แอททริบิวต์เมื่อรวมกับคุณลักษณะของจำนวนคำสรรพนามแทนตนเอง ต่อมาทำการกำหนดบทบาท (Set Role) ให้กับตัวระบุกลุ่ม (Label) และกำหนดแอททริบิวต์ (Set Attribute) หรือคอลัมน์ที่จะนำเข้าเพื่อฝึกฝน (Train) หลังจากนั้นในงานวิจัยนี้มีการใช้ ตัวดำเนินการที่ชื่อว่า Text Vectorization ในการหาค่าของการปรากฏของคำในแต่ละข้อความ แล้วจึงตรวจสอบแทนที่ข้อความที่สูญหายก่อนการนำเข้าแบบจำลอง เป็นอันเสร็จสิ้นการกำหนดขั้นพื้นฐาน

ส่วนถัดมาเป็นขั้นตอนในการเตรียมแบบจำลอง โดยทำการเลือกใช้ตัวดำเนินการประเมินผล (Validation) จากรูปที่ ด้านล่างที่ถูกแบ่งออกเป็น 2 ฝั่ง ซ้ายและขวา แสดงการแบ่งจำนวนข้อมูลในการฝึกฝนหรือทดสอบ ในที่นี้จะแบ่งข้อมูลด้วยอัตราส่วน 70 ต่อ 30 กล่าวคือส่วนด้านซ้ายหรือส่วนการฝึกฝนจะมีการนำเข้าข้อมูลร้อยละ 70 จากทั้งหมดเพื่อเรียนรู้ ส่วนฝั่งด้านขวาจะนำเข้าข้อมูลร้อยละ 30 เพื่อการทดสอบความแม่นยำ โดยในที่นี้จะเริ่มต้นเลือกใช้แบบจำลองของนาอีฟเบย์ (Naïve Bayes) เป็นตัวอย่าง

หลังจากนั้นในการจัดวางตัวดำเนินการต่อไป จะใช้คำสั่งที่ชื่อว่า Apply Model และ Performance เพื่อเดินหน้าใช้แบบจำลองและให้ผลของการรู้จำออกมาเป็นค่าประสิทธิภาพ



รูปภาพที่ 3.9 แสดงการใช้ตัวดำเนินการ (Operator) ภายใต้ตัวประเมินผล (Validation)

ก่อนทำการทดลองได้มีการตั้งค่าพารามิเตอร์ในบางแบบจำลองดังนี้

-แบบจำลองการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว มีการเลือกค่า k เท่ากับ 5

-แบบจำลองต้นไม้ตัดสินใจ มีการใช้ดัชนีจีนี่ (Gini index) ช่วยในเรื่องการตัดสินใจนำคุณลักษณะมาใช้สำหรับแยกโหนด

ส่วนสุดท้ายจะเป็นส่วนของการทดลองโดยในที่นี้จะทำการทดลองปรับปรุงค่าพารามิเตอร์ต่าง ๆ เพื่อให้ได้ค่าความถูกต้องดีที่สุด เริ่มต้นการทดลองด้วยการเลือกแอททริบิวต์ของคุณลักษณะในการนำเข้า 2 แบบคือ 1.เลือกนำเข้าคุณลักษณะทั้งหมด และ 2.การเลือกนำเข้าบางคุณลักษณะ ส่วนของการประเมินผลก็มีการทดลองใช้ 2 แบบเช่นกัน คือ 1.Split Validation ด้วยอัตราส่วน 70:30 และ 2.Cross Validation แบบ 10-fold ทำการทดสอบประสิทธิภาพการรู้จำและเปรียบเทียบประสิทธิภาพเพื่อหาแบบจำลองและแอททริบิวต์ที่ดีที่สุด โดยผลจะทำการกล่าวต่อไปในบทที่ 4

บทที่ 4
ผลการดำเนินการศึกษา

ในบทนี้จะนำเสนอการสรุปผลการทดลองของแต่ละแบบจำลองและนำมาเปรียบเทียบกัน

ผู้วิจัยได้นำข้อมูลจากทวิตเตอร์จำนวนรวม 14,572 เรคคอร์ด (Record) จาก 2 กลุ่ม แบ่งเป็นกลุ่มซึมเศร้าจำนวน 8,165 เรคคอร์ด และกลุ่มปกติจำนวน 6,407 เรคคอร์ด โดยมีรายละเอียดขั้นตอนการทำดังที่กล่าวไว้ในบทที่ 3.1 แล้วทำการสกัดคุณลักษณะและนำเข้าข้อมูลเพื่อการฝึกฝนรู้จำของแบบจำลอง

การประเมินประสิทธิภาพของแบบการวิเคราะห์อารมณ์เพื่อวิจัยภาวะซึมเศร้าในงานวิจัยนี้จะใช้การประเมินประสิทธิภาพการทำงานด้วยค่าความเที่ยง (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าคะแนน F1 จากแบบจำลองทั้ง 4 แบบ

การตั้งค่าการทดลองแบ่งเป็น 2 ส่วนคือ 1.เปรียบเทียบผลการทดลองระหว่าง การใช้ตัวแบ่งประเมินแบบ Split Validation ด้วยอัตราส่วน 70 : 30 กับตัวแบ่งประเมินแบบ 10-fold Cross Validation และ 2.เปรียบเทียบการใช้แอททริบิวต์แค่บางตัว กับการใช้แอททริบิวต์ทั้งหมดในการฝึกฝนรู้จำ

Model	Attribute	Validation	Accuracy	Precision		Recall		F1-score
				nor	dep	nor	dep	
KNN	All	Split Validation 0.7	71.71	76.61	64.65	75.81	65.66	0.71
	Selected	Split validation 0.7	65.60	68.85	58.94	77.47	48.00	0.63
	All	Cross validation-10	74.16	67.61	81.10	79.14	70.25	0.75
	Selected	Cross validation-10	67.08	70.49	60.62	77.24	52.01	0.65

ตารางที่ 4.1 แสดงผลประสิทธิภาพความถูกต้อง ค่าความเที่ยงตรง ค่าความระลึก และค่าคะแนน F1 จากการทดลองด้วยแบบจำลองการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว

จากตารางที่ 4.1 การทดลองด้วยแบบจำลองการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว พบว่าได้ผลต่ำที่สุดเมื่อใช้แอททริบิวต์แค่บางตัวกับตัวการประเมินแบบแบ่งแยกกลุ่ม (Split Validation) เท่ากับ 65.60 เปอร์เซ็นต์ และให้ผลประสิทธิภาพความถูกต้องสูงที่สุดเมื่อใช้แอททริบิวต์ทั้งหมดในการฝึกฝนพร้อมกับการใช้ตัวแบ่งการประเมินแบบ 10-fold Cross Validation โดยให้ค่าประสิทธิภาพความถูกต้องเท่ากับ 74.16 เปอร์เซ็นต์ ให้ค่าความเที่ยง (Precision) ของชุดซึมเศร้าสูงกว่าชุดปกติ ในทางกลับกันให้ค่าการระลึก (Recall) ของชุดปกติสูงกว่าชุดซึมเศร้า ซึ่งหมายความว่าแบบจำลองทำการเลือกแบ่งกลุ่มปกติได้ดีกว่ากลุ่มซึมเศร้าหรือเอียงไปในทางกลุ่มปกติมากกว่า แต่การแบ่งกลุ่มซึมเศร้านั้นง่ายกว่ากลุ่มปกติ

Model	Attribute	Validation	Accuracy	Precision		Recall		F1-score
				nor	dep	nor	dep	
Naive	All	Split Validation 0.7	72.58	72.58	69.16	51.93	84.61	0.70
	Selected	Split Validation 0.7	61.78	63.30	55.34	85.73	26.25	0.58
	All	Cross validation-10	70.76	73.36	69.5	52.61	85.91	0.70
	Selected	Cross validation-10	62.61	63.59	58.06	87.54	25.61	0.59

ตารางที่ 4.2 แสดงผลประสิทธิภาพความถูกต้อง ค่าความเที่ยงตรง ค่าความระลึกและค่าคะแนน F1 จากการทดลองด้วยแบบจำลองนาอีฟเบย์

จากตารางที่ 4.2 พบว่าค่าประสิทธิภาพความถูกต้องของการทดลองที่ดีที่สุดมาจากการใช้ชุดแอททริบิวต์ทั้งหมดกับตัวแบ่งประเมิน Split Validation คือเท่ากับ 72.58 เปอร์เซ็นต์ ในทางกลับกันการใช้แอททริบิวต์บางตัวกับตัวแบ่งประเมินแบบ Split Validation นั้นให้ผลประสิทธิภาพความถูกต้องที่ต่ำที่สุดคือ 61.78 เปอร์เซ็นต์

Model	Attribute	Validation	Accuracy	Precision		Recall		F1-score
				nor	dep	nor	dep	
SVM	All	Split Validation 0.7	62.18	62.42	60.39	92.23	17.58	0.58
	Selected	Split Validation 0.7	63.03	63.08	62.76	91.96	20.11	0.59
	All	Cross validation-10	62.09	63.40	56.31	86.45	25.93	0.58
	Selected	Cross validation-10	63.37	63.07	65.72	93.39	18.82	0.60

ตารางที่ 4.3 แสดงผลเปรียบเทียบค่าประสิทธิภาพความถูกต้อง ค่าความเที่ยงตรง ค่าความระลึกและค่าคะแนน F1 จากการทดลองด้วยแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน

จากตารางที่ 4.3 พบว่าค่าประสิทธิภาพความถูกต้องของการทดลองทั้ง 4 กรณีมีค่าไม่ต่างกันมากนัก กรณีที่ให้ผลดีที่สุดมาจากการใช้ชุดแอททริบิวต์แค่บางตัวกับตัวแบ่งประเมิน Cross Validation โดยจะให้ค่าประสิทธิภาพความถูกต้องเท่ากับ 63.37 เปอร์เซ็นต์ จากค่าประสิทธิภาพความถูกต้องสูงสุดนั้นเมื่อดูค่าความเที่ยงเปรียบเทียบระหว่างสองกลุ่มผลคือกลุ่มซึมเศร้าให้ค่าดีกว่าคือ 65.72 เปอร์เซ็นต์ และค่าความระลึก 93.39 จากกลุ่มปกติ ทั้งนี้หมายถึงว่าค่าผลการแบ่งกลุ่มให้คำตอบไปทางกลุ่มปกติมากกว่ากลุ่มซึมเศร้า

Model	Attribute	Validation	Accuracy	Precision		Recall		F1-score
				nor	dep	nor	dep	
Decision Tree	All	Split Validation 0.7	78.06	74.43	81.04	76.33	79.42	0.78
	Selected	Split Validation 0.7	72.20	63.26	86.17	87.72	60.02	0.76
	All	Cross validation-10	79.14	76.50	81.19	75.87	81.71	0.79
	Selected	Cross validation-10	72.18	62.91	87.69	87.53	58.57	0.74

ตารางที่ 4.4 แสดงผลเปรียบเทียบค่าประสิทธิภาพความถูกต้อง ค่าความเที่ยงตรง ค่าความระลึก และค่าคะแนน F1 จากการทดลองด้วยแบบจำลองต้นไม้ตัดสินใจ

จากตารางที่ 4.4 พบว่าผลลัพธ์ที่ดีที่สุดจากแบบจำลองนี้นั้นสูงกว่ากรณีอื่นอย่างชัดเจน คือ เท่ากับ 79.14 เปอร์เซ็นต์ สูงกว่าทุกแบบจำลองที่ทำการทดลองมา โดยที่เมื่อเปรียบเทียบค่าความเที่ยงของสองกลุ่มเห็นว่า กลุ่มซึมเศร้าให้คะแนนสูงกว่าคือ 81.19 หมายความว่า กลุ่มซึมเศร้าถูกแบ่งได้ง่ายกว่ากลุ่มปกติ และผลของค่าความระลึกเท่ากับ 81.71 ในกลุ่มซึมเศร้า ซึ่งหมายถึงว่า ผลการแบ่งกลุ่มมีแนวโน้มไปทางกลุ่มซึมเศร้ามากกว่า

Model	Attribute	Validation	Accuracy	Precision		Precision		F1-score
				nor	dep	nor	dep	
SVM	Selected	Cross validation-10	63.37	63.07	65.72	93.39	18.82	0.60
Naive	All	Split Validation 0.7	72.58	72.58	69.16	51.93	84.61	0.70
KNN	All	Cross validation-10	74.16	67.61	81.10	79.14	70.25	0.75
Decision Tree	All	Cross validation-10	79.14	76.50	81.19	75.87	81.71	0.79

ตารางที่ 4.5 แสดงผลเปรียบเทียบประสิทธิภาพความถูกต้อง ค่าความเที่ยงตรง ค่าความระลึกและค่าคะแนน F1 จากการทดลองที่ดีที่สุดของทุกแบบจำลอง

จากตารางที่ 4.5 แสดงผลการทดลองจากทุกแบบจำลองพบว่าในแต่ละแบบจำลองให้ค่าที่ดีที่สุดจากการตั้งค่าการทดลองด้วยรูปแบบแอททริบิวต์ต่างกันและวิธีการแบ่งประเมินข้อมูลต่างกัน ให้ค่าคะแนน F1 ตามลำดับจากน้อยไปมากดังนี้คือ แบบจำลองซัพพอร์ตเวกเตอร์แมชชีนให้ผลค่าประสิทธิภาพความถูกต้องสูงสุดเท่ากับ 0.60 แบบจำลองการจำแนกนาอีฟเบย์ได้ผลค่าประสิทธิภาพความถูกต้องสูงสุดเท่ากับ 0.70 แบบจำลองการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัวให้ค่าประสิทธิภาพความถูกต้องสูงสุดเท่ากับ 0.75 และสุดท้ายคือแบบจำลองต้นไม้ตัดสินใจให้ค่าประสิทธิภาพความถูกต้องสูงสุดเท่ากับ 0.79 ซึ่งมากเป็นอันดับที่ 1 เมื่อเปรียบเทียบกับการฝึกฝนด้วยแบบจำลองอื่น ๆ

บทที่ 5
อภิปรายผล

5.1 การอภิปรายผลการวิจัย

ผลของการทดลองเพื่อแบ่งแยกกลุ่มข้อมูลภาวะซึมเศร้า (Depressed class) และกลุ่มปกติ (Normal class) โดยมีการใช้คุณลักษณะทั้ง 7 อย่าง ร่วมกันในการรู้จำของแบบจำลองทั้ง 4 ประเภท ได้ผลว่าแบบจำลองต้นไม้ตัดสินใจให้ค่าความถูกต้องและค่าคะแนน F1 มากที่สุด คือค่าความถูกต้อง 79.14 และค่าคะแนน F1 0.79 เปรียบเทียบกับแบบจำลองซัพพอร์ตเวกเตอร์แมชชีน แบบจำลองนาอีฟเบย์ และแบบจำลองการค้นหาเพื่อนบ้านใกล้ที่สุด k ตัว เมื่อมีการใช้ตัวแบ่งประเมินแบบ 10-fold Cross Validation พร้อมกับการนำเข้าแอททริบิวต์คุณลักษณะทั้งหมดเข้าไปเพื่อฝึกฝน กล่าวคือต้นไม้ตัดสินใจสามารถรับมือกับแอททริบิวต์ประเภทแบ่งกลุ่ม (Norminal attribute) ประเภทแบ่งเป็นกลุ่มโดยมีการเรียงลำดับของกลุ่ม (Ordinal attribute) และข้อมูลต่อเนื่อง (Continuous attribute) ได้ดีกว่าแบบจำลองอื่น เนื่องจากต้นไม้ตัดสินใจจะให้ค่าประสิทธิภาพออกมาดีเมื่อใช้งานกับข้อมูลที่มีความขึ้นต่อกันสูง แอททริบิวต์ของคุณลักษณะต่าง ๆ อย่างเช่น จำนวนผู้ติดตาม จำนวนการติดตามผู้อื่น จำนวนการรีทวีต และจำนวนข้อความโดยรวม นั้นมีความสัมพันธ์กันในบางแง่มุม คือเมื่อมีจำนวนผู้ติดตามมาก ก็อาจส่งผลต่อจำนวนการรีทวีตที่มากแตกต่างกันไปในแต่ละกรณี แสดงว่าชุดข้อมูลที่รวบรวมแต่ละแอททริบิวต์นั้นมีความสัมพันธ์ต่อกันไม่ทางใดก็ทางหนึ่ง

จากการทดลองใช้อัตราการแบ่งของ Split Validation หลาย ๆ แบบ เช่น 80 ต่อ 20 ผลปรากฏว่าอัตราส่วน 70 ต่อ 30 ให้ค่าประสิทธิภาพที่ดีกว่า โดยพบว่ายิ่งค่าอัตราส่วนการทดสอบมาก จะสามารถคาดการณ์ความผิดพลาดได้ดีมากยิ่งขึ้น ทั้งนี้การปรับอัตราการแบ่งแยกก่อนการรู้จำที่เหมาะสมขึ้นอยู่กับลักษณะของข้อมูลที่ใช้งานด้วย

อีกทั้งพบว่าการแบ่งข้อมูลในการประมวลผลแบบ 10-fold Cross Validation ในงานวิจัยนี้เป็นการแบ่งประเมินผลข้อมูลที่เหมาะสมและให้ผลที่ดีกว่า Split Validation แบบ 70 ต่อ 30 เนื่องจาก 10-fold Cross Validation มีการเรียนรู้มากกว่า คือ ภายใน 10 ส่วน จะมีการเรียนรู้ถึง 9 ส่วนและการทดสอบ 1 ส่วนในแต่ละรอบการรู้จำ ซึ่งจะวนการรู้จำไปตามจำนวนที่กำหนด เมื่อเทียบกับ Split Validation ซึ่งมีการเรียนรู้เพียง 7 ส่วนเท่านั้น อัตราการเรียนรู้ (Learning rate) ที่มากกว่าบ่งบอกถึงความรู้ที่มากขึ้น ทำให้แบบจำลองสามารถตัดสินใจแบ่งแยกกลุ่มชุดข้อมูลได้แม่นยำมากขึ้นเช่นกัน

5.2 ข้อเสนอแนะ

1.งานวิจัยนี้ได้ทำการปรับใช้ขั้นตอนวิธีบางอย่างจากงานวิจัยที่ทำการค้นคว้า โดยนำมาปรับใช้กับการวิเคราะห์อารมณ์บนชุดข้อมูลภาษาไทย ทั้งนี้มีการใช้คุณลักษณะตามทฤษฎีต่าง ๆ มาประยุกต์กับชุดข้อมูลภาษาไทยที่รวบรวม และทำการตั้งค่าพารามิเตอร์เพิ่มเติมเพื่อทดสอบการทำงานของแบบจำลองทั้ง 4

แบบดังที่กล่าวไปข้างต้น เพื่อศึกษาวิธีการที่ดีที่สุดในการรู้จำแบบจำลอง โดยมุ่งเน้นการแบ่งกลุ่มที่ถูกต้องเหมาะสม

2.การวิเคราะห์อารมณ์กับข้อมูลภาษาไทยในปัจจุบันยังไม่ได้มีแหล่งศึกษาค้นคว้ามากนัก คลังคำศัพท์บอกขั้วอารมณ์ทั้งเชิงบวกและเชิงลบยังไม่เปิดกว้าง ทั้งนี้ในอนาคตผู้วิจัยหวังเป็นอย่างยิ่งว่าการวิเคราะห์อารมณ์ในภาษาไทยจะสามารถต่อยอดการวิเคราะห์ภาวะโรคซึมเศร้าได้แม่นยำมากขึ้นกว่าในงานวิจัยนี้ โดยการเพิ่มคุณลักษณะการคำนวนคะแนนอารมณ์ความรู้สึกผ่านคลังศัพท์ที่เก็บค่าอารมณ์ เช่น คลังศัพท์ SentiStrength ในภาษาอังกฤษ ที่ทำให้สามารถทราบถึงค่าคะแนนโดยรวมของข้อความนั้น ๆ โดยพบว่าในภาษาไทยนั้นยังไม่มีการทำคลังศัพท์ที่ระบุค่าอารมณ์มากนัก ทั้งนี้เพื่อเพิ่มประสิทธิภาพในการจำแนกกลุ่มข้อมูล

3.การตัดคำภาษาไทยในปัจจุบันมีการพัฒนาไปอย่างมาก จากการศึกษาค้นคว้าพบว่ามีหลักการและทฤษฎีมากมายในการตัดคำ และทำให้ได้ผลลัพธ์ที่มีความแม่นยำสูง เช่นการใช้ Deepcut หรือการใช้หลักการตัดคำโดยคลังคำศัพท์ (Dictionary-based) หรืออีกหลายวิธีในการช่วยตัดคำ ที่งานวิจัยนี้ไม่ได้เลือกใช้ เพื่อการตัดคำที่แม่นยำยิ่งขึ้น ส่งผลให้การเรียนรู้ของแบบจำลองให้ผลที่ดีด้วยเช่นกัน

4.การศึกษาวิจัยภาวะโรคซึมเศร้าสามารถใช้อีกหลายคุณลักษณะใช้วิเคราะห์ เช่นเดียวกับในงานวิจัยอื่น ๆ ที่ใช้คลังคำศัพท์เกี่ยวกับภาวะโรคซึมเศร้าโดยตรงในการตรวจจับขั้วอารมณ์ หรือการใช้คุณลักษณะค่าทางอารมณ์ (Sentiment Score) การตรวจจับขั้วอารมณ์ต่อช่วงเวลาแต่ละวัน หรือการดูการจับกลุ่ม (Cluster) ของเครือข่ายรอบข้างของผู้ใช้ เพื่อความมีประสิทธิภาพมากขึ้นของการวิเคราะห์ภาวะโรคซึมเศร้า

5.การรวบรวมข้อมูลจากสื่อออนไลน์จำเป็นต้องมีความระมัดระวัง เนื่องจากมีโอกาสที่จะได้ข้อมูลที่ไม่เกี่ยวข้องกับงานวิจัยมาจำนวนมาก อาจเกิดผลให้เกิดความผิดพลาดและเสียเวลาได้ จึงต้องใช้เทคนิคการค้นหาที่เหมาะสม บวกกับการทำความสะอาดข้อมูลที่ดีจะช่วยให้ประสิทธิภาพของการวิจัยดียิ่งขึ้นแน่นอน นอกจากนี้การเลือกกลุ่มผู้คนที่จะทำการศึกษานั้นก็สำคัญอย่างมาก ในงานวิจัยนี้ใช้การค้นหาข้อความทวิตเตอร์ด้วยแฮชแท็ก (Hashtag) ทำให้สามารถรวบรวมข้อมูลได้ในวงกว้าง ซึ่งมีโอกาสที่จะพบข้อความที่ไม่เกี่ยวข้องหรือข้อความที่ผิดกลุ่มได้ แม้กลุ่มข้อมูลชุดซึมเศร้า (Depressed set) จะสามารถรวบรวมข้อมูลได้อย่างค่อนข้างแม่นยำ แต่กลับกันกับชุดข้อมูลปกติ (Normal set) ที่ยากต่อการรวบรวมและค้นหา เนื่องจากการชี้วัดความปกติจากคำค้นหาแบบแฮชแท็กนั้นเป็นสิ่งที่ต้องใช้เวลาสำรวจในกลุ่มผู้ใช้งานทวิตเตอร์อย่างระเอียด โดยถ้าสามารถทำการรวบรวมข้อมูลของผู้เข้าร่วมแต่ละคนได้ติดต่อกันเป็นเวลานาน จะทำให้ได้รูปแบบพฤติกรรมที่มีชัดเจนและน่าสนใจต่อการทำวิจัย

เอกสารอ้างอิง

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. (2011, June) Proceedings of the Workshop on Languages in Social Media Page 30-38

[2] A. Asiaee , M. Tepper,Arindam Banerjee ,Guillermo Sapiro. If you are happy and you know it...tweet. (2012, October) Proceedings of the 21st ACM international conference on Information and knowledge management Page 1602-1606

[3] F. Bravo-Marquez , M. Mendoza and B. Poblete. Combining Strengths, Emotions and Polarities for Boosting Twitter Sentiment Analysis. (2013, August) Proceedings of the Second International Workshop on Issue of Sentiment Discovery and Opinion Mining Article No.2

[4] Internet Live Stat. Twitter Usage Statistics. (2016) Retrieved from https://www.internetlivestats.com/twitter-statistics/ [Available on October 20, 2019]

[5] K. Suppala, N. Rao. Sentiment Analysis Using Naïve Bayes Classifier. (2019,June) International Jounal of Innovative Technology and Exploring Engineering (IJITEE) ISSN:2278-3075, Volumn-8

[6] N. Vedula, S. Parthsarathy. Emotional and linguistic cues of depression from social media. Proceedings of the 2017 International Conference on Digital Health Page 127-136

[7] กองบรรณาธิการ HonestDocs (2019, November) Retrieved from https://www.honestdocs.co/most-common-psychiatric-disorders. [Available on November 9 ,2019]

[8] อรพรรณ ลือบุญธวัชชัย,พีรพนธ์ ลือบุญธวัชชัย . การบำบัดรักษาทางจิตสังคมสำหรับโรคซึมเศร้า Psychosocial treatment for depressive disorder.(2553)

[9] พรพรรณ สุดใจ. ผลของโปรแกรมกลุ่มบำบัดทางจิตสังคมแบบบูรณาการต่อภาวะซึมเศร้าและการทำหน้าที่ทางสังคม.(2009)

[10] อนุพงศ์ สุขประเสริฐ. Data Pre-Processing.[ออนไลน์] ได้จาก: https://slideplayer.in.th/slide/15935722/ [สืบค้นเมื่อ 8 เมษายน 2562]

[11] อัจฉรินทร์ กล่อมแสร์.การสกัดคุณลักษณะภาพเพื่อการรรู้จำตัวเขียนอักษรธรรมล้านนา.(กันยายน 2555)

[12] วิภาวรรณ บัวทอง. Naive Bayes Classification.(มิถุนายน 2557)

[13] พงศกร ธีรรัศมี. วิธีการหาค่า เค ที่เหมาะสมในการจำแนกแบบเคเนียร์เรสเนเบอร์กับข้อมูลทางการแพทย์.(2558)

[14] สุภาพร คลังเพ็ชร.ประสิทธิภาพการจัดกลุ่มของวิธีซัพพอร์ตเวกเตอร์แมทชีนและวิธี เคเนียเรสเนเบอร์ กรณีข้อมูลที่มีการแจกแจงแบบเสถียรที่มี ลักษณะหางหนา (2553)

[15]ภู ริพัทร์ ทองคำ.อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล.(2559)

[16] นฤพนต์ ว่องประชุกูล.วิธีที่เหมาะสมสำหรับการตัดกิ่งต้นไม้ตัดสินใจของการทำเหมืองข้อมูลทางด้านวิทยาศาสตร์.(2547)

[17]การทำเหมืองข้อมูลแบบจำแนก Classification Data mining.(2013) ได้จาก : https://wipawanblog.files.wordpress.com/2013/08/chapter-5-classification-decision-tree.pdf [สืบค้นเมื่อ 8 เมษายน 2562]

[18] วิชพงศ์ ดรูญธรรม(2018) รู้จัก Decision Tree, Random Forest, และ XGBoost [ออนไลน์] ได้จาก: https://medium.com/@witchapongdaroontham/%E0%B8%A3%E0%B8%B9%E0%B9%89%E0%B8%88%E0%B8%B1%E0%B8%81-decision-tree-random-forrest-%E0%B9%81%E0%B8%A5%E0%B8%B0-xgboost-part-1-cb49c4ac1315 [สืบค้นเมื่อ 5 เมษายน 2562]

[19] เอกสิทธิ์ พัชรวงศ์ศักดา.(2557).การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น An Introduction to Data Mining Techniques Thai version.พิมพ์ครั้งที่ 1.[ออนไลน์].บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด 21/19-20 ถ.งามวงศ์วาน แขวงลาดยาว เขตจตุจักร กรุงเทพมหานคร 10900 ได้จาก : http://dataminingtrend.com/2014/wpcontent/uploads/2014/08/intro_data_mining_preview.pdf [สืบค้นเมื่อ 8 เมษายน 2562]

[20] M. Park, C. Cha and M. Cha . Depressive Moods of Users Portrayed in Twitter.(2012)

[21]กานดา รุณนะพงศา และ ปโยธร อุราธรรมกุล.การตัดคาภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่.(เมษายน 2549)

[22] I. Kawachi และ Lisa F Berkman.Social ties and mental health.(2001). Journal of urban health (2001)

ประวัติผู้เขียน

นางสาวพรพัทธา อมรรังสรรค์

รหัสนิสิต 5933641023

วันเดือนปีเกิด 9 ตุลาคม 2540

ภูมิลาเนา จังหวัดกรุงเทพมหานคร กำลังศึกษาใน
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์
และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
จุฬาลงกรณ์มหาวิทยาลัย