การพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า โดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

สุดา ทิพย์ประเสริฐ

มหาวิทยาลัยเทคโนโลยีสุรนารี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาปรัชญาดุษฎีบัณฑิต
สาขาวิชาเทคโนโลยีสารสนเทศ
มหาวิทยาลัยเทคโนโลยีสุรนารี
ปีการศึกษา 2564

THE DEVELOPMENT OF A DEPRESSION RISK ANALYSIS MODEL USING ONLINE SOCIAL NETWORK DATA

Suda Tipprasert

มหาวิทยาลัยเทคโนโลยีสุรนารี

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Information Technology)

Suranaree University of Technology

Academic Year 2022

**การพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า
โดยใช้ข้อมูลเครือข่ายสังคมออนไลน์**

มหาวิทยาลัยเทคโนโลยีสุรนารี อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาดุษฎีบัณฑิต

คณะกรรมการสอบวิทยานิพนธ์

(ผู้ช่วยศาสตราจารย์ ดร.ศุภกฤษฏิ์ นิวัฒนากูล)
ประธานกรรมการ

(รองศาสตราจารย์ ดร.จิติมนต์ อั่งสกุล)
กรรมการ (อาจารย์ที่ปรึกษาวิทยานิพนธ์)

(รองศาสตราจารย์ ดร.ชูพันธุ์ รัตนโภคา)
กรรมการ

(รองศาสตราจารย์ ดร.ศิรปัฐช์ บุญครอง)
กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.หนึ่งหทัย ขอผลกลาง)
กรรมการ

รองอธิการบดีฝ่ายวิชาการและประกันคุณภาพ
(รองศาสตราจารย์ ดร.ฉัตรชัย โชติษฐยางกูร)

คณบดีสำนักวิชาศาสตร์และศิลป์ดิจิทัล
(รองศาสตราจารย์ ดร.ธรา อั่งสกุล, รักษาการแทน)

มหาวิทยาลัยเทคโนโลยีสุรนารี

สุดา ทิพย์ประเสริฐ : การพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ (THE DEVELOPMENT OF A DEPRESSION RISK ANALYSIS MODEL USING ONLINE SOCIAL NETWORK DATA)
อาจารย์ที่ปรึกษา : รองศาสตราจารย์ ดร.จิติมนต์ อั่งสกุล, 168 หน้า.

ภาวะซึมเศร้าเป็นหนึ่งในสาเหตุหลักที่ก่อให้เกิดปัญหาการฆ่าตัวตายตามมา ปัจจุบันจำนวนผู้ป่วยโรคซึมเศร้าที่เพิ่มขึ้นอย่างต่อเนื่อง ซึ่งส่วนใหญ่ไม่ได้รับการรักษา และอาศัยการแสดงออกผ่านทางสื่อสังคมออนไลน์ งานวิจัยนี้จึงมุ่งเน้นการวิเคราะห์ข้อมูลจากเครือข่ายสังคมออนไลน์เพื่อนำไปใช้ในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าที่เรียกว่า ดีแรม

ในการพัฒนาแบบจำลองดีแรมได้เก็บรวบรวมข้อมูลคุณลักษณะด้านกายภาพส่วนบุคคลและผลการประเมินภาวะซึมเศร้าจากการทำแบบสอบถาม ร่วมกับการสกัดข้อมูลการแสดงความคิดเห็นในทวิตเตอร์ ได้แก่ ข้อมูลทั่วไปของผู้ใช้ทวิตเตอร์ ข้อความ สัญรูปอารมณ์ และรูปภาพ และประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องในรูปแบบต่าง ๆ เพื่อใช้ในการสร้างแบบจำลอง โดยมีสมมติฐานการวิจัยคือ แบบจำลองที่พัฒนาขึ้นสามารถประมวลผลได้ความถูกต้องร้อยละ 80 ขึ้นไป

ผลการวิจัยพบว่า แบบจำลองดีแรมที่นำเสนอในงานวิจัยนี้ มีค่าประสิทธิภาพโดยรวม ร้อยละ 88.51 และให้ค่าความถูกต้อง ร้อยละ 89.00 ซึ่งสูงกว่าเทคนิคการเรียนรู้ของเครื่องอื่นทั้งหมด โดยแบบจำลองดีแรมสามารถนำไปประยุกต์ใช้ในการพัฒนาระบบวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าเพื่อตรวจสอบภาวะซึมเศร้าในระดับเริ่มต้น นอกจากนี้ระบบสามารถนำไปประยุกต์ใช้ในสถานศึกษา ครอบครัว หรือสถานที่ทำงานได้

สาขาวิชาเทคโนโลยีสารสนเทศ	ลายมือชื่อนักศึกษา ____________
ปีการศึกษา 2564	ลายมือชื่ออาจารย์ที่ปรึกษา ____________

มหาวิทยาลัยเทคโนโลยีสุรนารี

SUDA TIPPRASERT : THE DEVELOPMENT OF A DEPRESSION RIKS ANALYSIS MODEL USING ONLINE SOCIAL NETWORK DATA. THESIS ADVISOR : ASSOC. PROF. JITIMON ANGSKUN, D.ENG., 168 PP.

DEPRESSION/SOCIAL NETWORK/RISK ANALYSIS MODEL/MACHINE LEARNING

Depression is one of the leading causes of suicide problems. Currently, the number of people suffering from depression is constantly increasing. Most depressed people are not treated and express behavior via social network posts. Thus, this research focuses on analyzing data from social networks for use in designing and developing a Depression Risk Analysis Model called DeRAM.

To develop the DeRAM, demographic characteristics and depression assessment outcomes are collected from questionnaires, and opinion data are extracted from Twitter, such as Twitter users' Information, text, emoticons, and image. Machine learning technique is applied as the model construction. The research hypothesis is that the developed model can process with 80% of accuracy.

The experimental results revealed that the proposed DeRAM was achieved with 88.51% of F-measure and 89.00% of accuracy, which is higher than the other machine learning techniques. The DeRAM can be applied to develop a depression risk analysis system for early detection of depression. In addition, this system can be applied in schools, families, or workplaces.

School of Information Technology | Student's Signature____________

Academic Year 2021 | Advisor's Signature____________

กิตติกรรมประกาศ

วิทยานิพนธ์เล่มนี้สำเร็จด้วยดี ข้าพเจ้าขอกราบขอบพระคุณ รองศาสตราจารย์ ดร.จิติมนต์ อั่งสกุล ผู้ซึ่งเป็นอาจารย์ที่ปรึกษาวิทยานิพนธ์ที่ผู้วิจัยรักและเคารพยิ่ง ที่ได้กรุณาให้คำปรึกษา คำแนะนำ ให้กำลังใจ และให้การช่วยเหลือด้วยความเมตตาเสมอมา ทั้งในด้านวิชาการ และด้านการดำเนินการวิจัย

ขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.ศุภกฤษฏิ์ นิวัฒนากูล ประธานกรรมการในการสอบวิทยานิพนธ์ รองศาสตราจารย์ ดร.ชูพันธุ์ รัตนโภคา รองศาสตราจารย์ ดร.ศิรปัฐช์ บุญครอง และผู้ช่วยศาสตราจารย์ ดร.หนึ่งหทัย ขอผลกลาง ที่ได้กรุณาเสียสละเวลาอันมีค่าในการพิจารณาและให้คำแนะนำในการแก้ไขปรับปรุงวิทยานิพนธ์ เสนอแนวทาง ความรู้ และให้ข้อเสนอแนะที่เป็นประโยชน์ต่อการวิจัยในครั้งนี้เป็นอย่างยิ่ง

ขอกราบขอบพระคุณคณาจารย์ทุกท่านในสาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารีทุกท่าน ที่ให้ความรู้ และประสบการณ์ที่เป็นประโยชน์ต่อกระบวนการวิจัยในครั้งนี้

ขอขอบคุณมหาวิทยาลัยเทคโนโลยีราชมงคลอีสานที่ให้การสนับสนุนทุนการศึกษา และขอบคุณคณาจารย์สาขาวิชาระบบสารสนเทศ คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ซึ่งเป็นต้นสังกัดของข้าพเจ้า ที่ให้การสนับสนุน ช่วยเหลือ ให้คำปรึกษา และให้กำลังใจในการทำวิทยานิพนธ์เล่มนี้ให้สำเร็จลุล่วง

ขอขอบคุณเพื่อน ๆ พี่ ๆ และ น้อง ๆ เจ้าหน้าที่ และบัณฑิตศึกษาทุกท่าน ที่คอยช่วยเหลือและเป็นกำลังใจที่ดีเสมอมา

ท้ายนี้ ขอกราบขอบพระคุณทุกคนในครอบครัว ที่คอยให้กำลังใจ สนับสนุน ส่งเสริมผลักดันให้การทำวิทยานิพนธ์สำเร็จลงได้ด้วยดี คุณงามความดีอันใดที่เกิดจากวิทยานิพนธ์เล่มนี้ ข้าพเจ้าขอมอบให้กับ บิดา มารดา และครูอาจารย์ที่เคารพทุกท่าน ที่ได้ให้วิชาความรู้ และเป็นกำลังใจให้ข้าพเจ้าเสมอมาจนสำเร็จการศึกษา

สุดา ทิพย์ประเสริฐ

สารบัญ

หน้า

บทคัดย่อ (ภาษาไทย)......ก
บทคัดย่อ (ภาษาอังกฤษ)......ข
กิตติกรรมประกาศ......ค
สารบัญ......ง
สารบัญตาราง......ซ
สารบัญภาพ......ญ
บทที่
1 บทนำ......1
1.1 ความสำคัญและที่มาของปัญหาการวิจัย......1
1.2 วัตถุประสงค์การวิจัย......4
1.3 สมมติฐานการวิจัย......4
1.4 ข้อตกลงเบื้องต้น......4
1.5 ขอบเขตของการวิจัย......4
1.6 ประโยชน์ที่คาดว่าจะได้รับ......5
1.7 คำอธิบายศัพท์......5
2 ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง......6
2.1 ข้อมูลพื้นฐานที่นำมาใช้ในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์......7
2.1.1 ภาวะซึมเศร้า......7
2.1.2 เครือข่ายสังคมออนไลน์ (Social Network)......14
2.1.3 สื่อสังคม (Social Media)......14
2.1.4 ทวิตเตอร์ (Twitter)......16
2.2 เทคโนโลยีในการรวบรวมข้อมูล......23
2.2.1 โปรแกรมประยุกต์บนเว็บ (Web Application)......23
2.2.2 ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ (Twitter API)......25
2.2.3 ส่วนต่อประสานโปรแกรมประยุกต์ในไพทอน (Python API)......26

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญ (ต่อ)

หน้า

2.3 เทคโนโลยีในการประมวลผลข้อมูล........28

2.3.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)........28

2.3.2 การวิเคราะห์ความรู้สึก (Sentiment Analysis)........35

2.3.3 การประมวลผลรูปภาพ (Image Processing)........38

2.3.4 การตรวจจับใบหน้า (Face Detection)........43

2.3.5 การประมวลผลสัญรูปอารมณ์ (Emoticon Processing)........47

2.4 เทคนิคที่นำมาใช้ในการสร้างแบบจำลอง........49

2.4.1 การเรียนรู้ของเครื่อง (Machine Learning)........49

2.4.2 นาอีฟเบย์ (Naïve Bayes)........50

2.4.3 ต้นไม้การตัดสินใจ (Decision Tree)........52

2.4.4 การสุ่มป่าไม้ (Random Forest)........57

2.4.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)........59

2.4.6 การเรียนรู้เชิงลึก (Deep Learning)........62

2.4.7 เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)........67

2.5 การประเมินผลแบบจำลอง........70

2.5.1 เทคนิคการประเมินแบบจำลอง (Model Evaluation Techniques)........70

2.5.2 การประเมินแบบจำลองด้วยเมตริกซ์ (Model Evaluation Metrics)........71

2.6 งานวิจัยที่เกี่ยวข้อง........73

2.6.1 การวิเคราะห์ความรู้สึกจากข้อมูลในเครือข่ายสังคมออนไลน์........73

2.6.2 การวิเคราะห์ภาวะซึมเศร้าจากข้อมูลในเครือข่ายสังคมออนไลน์........74

2.6.3 การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า........75

2.7 กรอบแนวคิดการวิจัย........83

3 วิธีการดำเนินการวิจัย........85

3.1 วิธีวิจัย........85

3.1.1 ศึกษา รวบรวมข้อมูล และวิเคราะห์คุณลักษณะที่เกี่ยวข้องกับการเกิดภาวะซึมเศร้า โดยใช้ข้อมูลเครือข่ายสังคมออนไลน์........85

สารบัญ (ต่อ)

หน้า

3.1.2 ออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ 86
3.1.3 ทดสอบและปรับปรุงประสิทธิภาพของแบบจำลองที่พัฒนา 102
3.2 ประชากรและกลุ่มตัวอย่าง 102
3.2.1 ประชากร 102
3.2.2 กลุ่มตัวอย่าง 102
3.3 เครื่องมือที่ใช้ในการวิจัย 103
3.4 การเก็บรวบรวมข้อมูล 104
3.4.1 การเก็บข้อมูลเพื่อนำมาใช้ในการสร้างแบบจำลอง 104
3.4.2 การเก็บข้อมูลเพื่อวัดค่าความถูกต้องของแบบจำลอง 104
3.5 การวิเคราะห์ข้อมูล 104
4 ผลการวิจัยและอภิปรายผล 107
4.1 ผลการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ 107
4.1.1 การเก็บรวบรวมข้อมูลจากเครือข่ายสังคมออนไลน์และแบบประเมินภาวะซึมเศร้า 9 คำถาม 108
4.1.2 การสกัดคุณลักษณะของข้อมูล (Feature Extraction 111
4.1.3 การหาเทคนิคในการคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection Techniques) 114
4.1.4 การคัดเลือกคุณลักษณะที่เหมาะสม (FEATURE SELECTION) 137
4.2 ผลการประเมินแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์ 141
4.2.1 การเปรียบเทียบค่าประสิทธิภาพแบบจำลองที่พัฒนา กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) อื่น ๆ 141
4.2.2 การอภิปรายผล 144
5 สรุปและข้อเสนอแนะ 148
5.1 สรุปผลการวิจัย 148

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญ (ต่อ)

หน้า

5.2 ข้อจำกัดของการวิจัย........149
5.3 การประยุกต์ผลการวิจัย........150
5.4 ข้อเสนอแนะในการวิจัยต่อไป........151
รายการอ้างอิง........152
ภาคผนวก ก........162
ภาคผนวก ข........166
ประวัติผู้เขียน........168

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญตาราง

ตารางที่ **หน้า**

2.1 สัญลักษณ์ที่ใช้ในการกำกับหน้าที่ของคำและความหมายของสัญลักษณ์ 33
2.2 โครงสร้างของ SentiWrodNet 36
2.3 งานวิจัยที่เกี่ยวข้อง 79
3.1 คุณลักษณะที่เกี่ยวข้องกับภาวะซึมเศร้า 85
3.2 ตัวอย่างคลังคำเหมือน (Synonym Word Corpus) 91
3.3 คุณลักษณะของข้อมูลในไฟล์ข้อมูลที่ใช้สำหรับการสร้างแบบจำลอง 98
4.1 คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics) 112
4.2 คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information) 112
4.3 คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) 113
4.4 คุณลักษณะจากทวิตเตอร์ที่เป็นเป็นสัญรูปอารมณ์ (Emoticon) 113
4.5 คุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) 114
4.6 ผลการคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (RFE-SVM) 115
4.7 การเปรียบเทียบผลการคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ 117
4.8 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test) 119
4.9 การตัดคุณลักษณะตามลำดับคะแนนจากเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test) 122
4.10 การเปรียบเทียบผลการคัดเลือกคุณลักษณะด้วยเทคนิค ANOVA F-test Feature Selection 124
4.11 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) 126
4.12 การตัดคุณลักษณะตามลำดับคะแนนจากเทคนิคการสุ่มป่าไม้ 129
4.13 เปรียบเทียบผลการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ 131

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญตาราง (ต่อ)

ตารางที่ หน้า

4.14 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยไม่ใช้เทคนิคการคัดเลือกคุณลักษณะ (No FS)........134
4.15 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ ที่ 29 คุณลักษณะ (RFE-SVM(29))........134
4.16 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน ที่ 15 คุณลักษณะ (ANOVA(15))........134
4.17 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้การคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ ที่ 23 คุณลักษณะ (Random(23))........135
4.18 การเปรียบเทียบค่าประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้........135
4.19 การเปรียบเทียบค่าประสิทธิภาพโดยรวม (F-measure) ของแบบจำลองจากการใช้ตัวแปรนำเข้า (Input Variables) ที่แตกต่างกัน........138
4.20 ผลการประเมินประสิทธิภาพแบบจำลองที่พัฒนากับเทคนิคการเรียนรู้ของเครื่องอื่น ๆ........142

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญภาพ

ภาพที่ หน้า

2.1 ภาวะซึมเศร้า 9
2.2 CES-D : Center for Epidemiological Studies-Depression Scale 10
2.3 PHQ-9 : The Patient Health Questionnaire depression scale 11
2.4 แบบคัดกรองภาวะซึมเศร้า 2 คำถาม 12
2.5 แบบประเมินโรคซึมเศร้า 9 คำถาม 13
2.6 คลาสไดอะแกรมของทวิตเตอร์ 17
2.7 ยูสเคสไดอะแกรมของทวิตเตอร์ 18
2.8 การแสดงความคิดเห็นผ่านทวิตเตอร์ 19
2.9 ตัวอย่างข้อความทวีต (Tweet) 20
2.10 ตัวอย่างจำนวนรีทวีต (Retweet) 20
2.11 ตัวอย่างจำนวนการกดถูกใจ (Like) 21
2.12 ตัวอย่างจำนวนการติดตาม (Following) 21
2.13 ตัวอย่างจำนวนผู้ติดตาม (Follower) 22
2.14 สถาปัตยกรรมโปรแกรมประยุกต์บนเว็บ 23
2.15 ขั้นตอนการประมวลผลภาษาธรรมชาติ 29
2.16 โครงสร้างของภาษา 30
2.17 กระบวนการวิเคราะห์ความรู้สึกโดยใช้ SentiWordNet 37
2.18 ขั้นตอนการวิเคราะห์ความรู้สึกโดยอาศัยการเรียนรู้ของเครื่อง 38
2.19 ขั้นตอนพื้นฐานของการประมวลผลภาพดิจิทัล 40
2.20 แสงสีหลักในระบบสีอาร์จีบี 41
2.21 ระบบสี HSV 42
2.22 วิธีการต่าง ๆ ที่ใช้สำหรับการตรวจจับใบหน้า 44
2.23 ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ 46
2.24 คุณลักษณะของฮาร์ไลค์ฟีเชอร์ (Haar-like features) 47
2.25 ฐานข้อมูลการวิเคราะห์ความรู้สึกสัญรูปอารมณ์ 48

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

2.26 องค์ประกอบของต้นไม้การตัดสินใจ.....52
2.27 ขั้นตอนการทำงานของขั้นตอนวิธี CHAID.....56
2.28 หลักการทำงานของขั้นตอนวิธีการสุ่มป่าไม้.....57
2.29 ความแตกต่างของโครงข่ายประสาทเทียมและการเรียนรู้เชิงลึก.....63
2.30 หลักการทำงานของโครงข่ายประสาทเทียมแบบสังวัตนาการ.....65
2.31 โครงสร้างของโครงข่ายประสาทเทียมแบบวนซ้ำ.....66
2.32 โครงสร้างของโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว.....66
2.33 ฟังก์ชันฮาร์ดลิมิต และฟังก์ชันฮาร์ดลิมิตแบบสมมาตร.....67
2.34 ฟังก์ชันเชิงเส้น (Linear).....68
2.35 ชั้นของแบบจำลองเพอร์เซ็ปตรอนหลายชั้น.....69
2.36 กรอบแนวคิดการวิจัย.....84
3.1 กรอบการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์.....86
3.2 โปรแกรมประยุกต์บนเว็บสำหรับเก็บข้อมูลตามแบบประเมินภาวะซึมเศร้า 9 คำถาม.....87
3.3 การสกัดคุณลักษณะจากข้อมูลที่เป็นประเภทข้อความในทวิตเตอร์.....89
3.4 ขั้นตอนการตรวจสอบข้อมูล.....90
3.5 ขั้นตอนวิธีการตัดคำแบบเหมือนมากที่สุด.....90
3.6 กระบวนการการประมวลผลภาษาธรรมชาติ.....91
3.7 คำสั่งในการค้นหาคำศัพท์ในคลังคำเหมือน.....91
3.8 คำสั่งที่ใช้สำหรับแปลภาษาด้วย Google Translate ผ่าน Python API.....92
3.9 ขั้นตอนวิธีการหาค่าคะแนนความรู้สึก.....93
3.10 ตัวอย่างไฟล์ sentiwordnet3.0.txt.....93
3.11 ฐานข้อมูล Emoji Sentiment Ranking.....95
3.12 การสกัดคุณลักษณะจากข้อมูลที่เป็นสัญรูปอารมณ์ในทวิตเตอร์.....95
3.13 ขั้นตอนการสกัดคุณลักษณะจำนวนรูปภาพจากการตรวจจับใบหน้า.....96
3.14 ขั้นตอนวิธีการวิเคราะห์ค่าเฉลี่ยระดับพิกเซล.....97

มหาวิทยาลัยเทคโนโลยีสุรนารี

สารบัญภาพ (ต่อ)

ภาพที่ **หน้า**

3.15 ขั้นตอนการสร้างแบบจำลองที่ใช้ในการพยากรณ์ 100
4.1 โปรแกรมประยุกต์บนเว็บสำหรับจัดเก็บข้อมูล 108
4.2 ส่วนการเชื่อมต่อกับทวิตเตอร์เอพีไอ 109
4.3 ส่วนของแบบสอบถามข้อมูลพื้นฐาน 109
4.4 ส่วนของแบบประเมินภาวะซึมเศร้า 9 คำถาม 110
4.5 ส่วนของผลจากแบบประเมินภาวะซึมเศร้า 9 คำถาม 111
4.6 การประเมินประสิทธิภาพเพื่อหาคุณลักษณะที่ควรตัดออกเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (RFE-SVM) 118
4.7 ประสิทธิภาพโดยรวมจากการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน เมื่อใช้งานร่วมกับเทคนิคการสุ่มป่าไม้ 125
4.8 ประสิทธิภาพโดยรวมจากการคัดเลือกคุณลักษณะด้วยการสุ่มป่าไม้ 132
4.9 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ ร่วมกับการคัดเลือก คุณลักษณะด้วยวิธีการต่าง ๆ 136
4.10 การเปรียบเทียบค่าประสิทธิภาพของแบบจำลองจากการใช้ตัวแปรนำเข้า (Input Variables) ที่แตกต่างกัน 140
4.11 การเปรียบเทียบค่าประสิทธิภาพแบบจำลองที่พัฒนา กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) อื่น ๆ 143
4.12 เวลาในการสร้างและเรียกใช้แบบจำลองรูปแบบต่าง ๆ 143
4.13 คุณลักษณะที่สำคัญสำหรับการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะ โรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์ 146

มหาวิทยาลัยเทคโนโลยีสุรนารี

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของปัญหาการวิจัย

เทคโนโลยีดิจิทัล ได้เข้ามามีบทบาทในการดำเนินชีวิตประจำวันอย่างครอบคลุมในทุก ๆ ด้าน โดยเฉพาะด้านการสื่อสาร ส่งผลให้เกิดการพัฒนากระบวนการหรือรูปแบบการสื่อสารให้สะดวกรวดเร็ว และเชื่อมต่อกันได้อย่างอิสระ ไม่ว่าจะเป็นการสื่อสารผ่านทางเครือข่ายสังคมออนไลน์ (Social Network) ซึ่งได้รับความนิยม และมีแนวโน้มการใช้งานเพิ่มขึ้นอย่างต่อเนื่อง โดยผู้ใช้งานเครือข่ายสังคมออนไลน์ อาจมีวัตถุประสงค์ในการใช้งานแตกต่างกันไป เช่น ใช้งานเครือข่ายสังคมออนไลน์เพื่องานธุรกิจ ใช้เพื่อการเข้าถึงข้อมูลข่าวสารในปัจจุบัน ใช้สำหรับเป็นสมุดบันทึกประจำวันเพื่อแสดงออกด้านความรู้สึก หรือใช้ในการแสดงอารมณ์ ทั้งในด้านที่ดี และด้านที่เป็นความเสี่ยงในเรื่องต่าง ๆ เช่น แสดงอารมณ์สำหรับผู้ที่มีความเสี่ยงที่จะเกิดภาวะซึมเศร้าซึ่งภาวะซึมเศร้า คือ ภาวะที่จิตใจหม่นหมอง หดหู่ เศร้า ท้อแท้ หมดหวังมองโลกแง่ร้าย รู้สึกมีคุณค่าในตนเองต่ำ ตำหนิตนเอง ซึ่งความรู้สึกเหล่านี้จะเกี่ยวข้องกับความสูญเสีย โดยภาวะซึมเศร้าเป็นหนึ่งในสาเหตุที่ทำให้เกิดโรคซึมเศร้า

โรคซึมเศร้า หมายถึง "ภาวะจิตใจที่แสดงออกถึงความผิดปกติของอารมณ์ เช่น อารมณ์เศร้า ไม่มีความสุข เบื่อหน่าย หดหู่ ท้อแท้ สิ้นหวัง เซื่องซึม นอนไม่หลับ เบื่ออาหาร ขาดสมาธิ วิตกกังวล มองโลกในแง่ลบ ซึ่งอาจจะส่งผลเสียต่อคุณภาพชีวิต และอาจนำไปสู่การฆ่าตัวตายในที่สุด" (นันทิรา หงษ์ศรีสุวรรณ์, 2559) การวินิจฉัยโรคซึมเศร้าสามารถวินิจฉัยด้วยตนเองร่วมกับการวินิจฉัยโดยแพทย์ ซึ่งการวินิจฉัยด้วยตนเองจะกระทำโดยการทำแบบทดสอบประเมินภาวะซึมเศร้า 9 คำถาม และการวินิจฉัยโดยแพทย์จะใช้การตรวจทางห้องปฏิบัติการร่วมด้วย โดยสาเหตุของโรคซึมเศร้ามีปัจจัยสำคัญที่เกี่ยวข้องได้แก่ กรรมพันธุ์ สารเคมีในสมอง และลักษณะนิสัยเฉพาะบุคคล (มาโนช หล่อตระกูล, 2560) ซึ่งโรคซึมเศร้านั้นก่อให้เกิดการสูญเสียด้านสุขภาพอย่างมีนัยสำคัญ และส่งผลกระทบต่อประชากรทั่วโลก องค์การอนามัยโลกได้ประมาณการว่ามีประชากรมากกว่า 280 ล้านคนเป็นโรคซึมเศร้า ซึ่งมากกว่าร้อยละ 4 ของประชากรโลก และมีถึงร้อยละ 80 ที่ไม่ได้รับการรักษา (WHO, 2564)

มหาวิทยาลัยเทคโนโลยีสุรนารี

ในประเทศไทยอัตราการป่วยโรคซึมเศร้า ได้เพิ่มขึ้นอย่างต่อเนื่องในช่วงปี พ.ศ. 2552-2562 โดยในปี พ.ศ. 2552 พบผู้ป่วยโรคซึมเศร้าในอัตรา 270 รายต่อแสนประชากร ในขณะที่ปี พ.ศ. 2562 พบผู้ป่วยโรคซึมเศร้าในอัตรา 345.1 รายต่อประชากรแสนคน (สำนักงานพัฒนาระบบข้อมูลข่าวสารสุขภาพ, 2564) โรคซึมเศร้าเป็นหนึ่งในสาเหตุหลักที่ก่อให้เกิดปัญหาการฆ่าตัวตายตามมา โดยมีอัตราการฆ่าตัวตายอยู่ที่ 7.37 รายต่อประชากรแสนคน ในปี พ.ศ. 2563 เพิ่มขึ้นจาก 5.9 รายต่อประชากรแสนคน ในปี พ.ศ. 2558 และในปี พ.ศ. 2560 ผู้ที่ได้รับการค้นหาคัดกรองว่ามีแนวโน้มป่วยเป็นโรคซึมเศร้าด้วยแบบคัดกรองภาวะซึมเศร้า 2 คำถาม (แบบประเมิน 2Q) มีจำนวนถึง 14 ล้านคน และมีถึงร้อยละ 64 ที่ไม่ได้รับการรักษา (ศูนย์โรคซึมเศร้า โรงพยาบาลพระศรีมหาโพธิ์, 2560; กรมสุขภาพจิต, 2560) นอกจากนั้นยังพบว่าจำนวนผู้ป่วยด้วยโรคซึมเศร้าในปี พ.ศ. 2560 เพิ่มขึ้นถึงร้อยละ 12 จากปี พ.ศ. 2559 และในประเทศไทยอัตราการฆ่าตัวตายสำเร็จประมาณ 4,000 คน ในปี พ.ศ. 2563 ซึ่งมีถึงร้อยละ 60 ที่มีสาเหตุมาจากโรคซึมเศร้า (กรมสุขภาพจิต, 2564) จากสถิติดังกล่าวสะท้อนให้เห็นถึงความจำเป็นและความสำคัญในการสนับสนุนงานป้องกัน รักษา และส่งเสริมสุขภาวะทางจิตของคนไทย (สถาบันวิจัยประชากรและสังคม มหาวิทยาลัยมหิดล, 2560)

จากข้อมูลดังกล่าวจะเห็นได้ว่า ผู้ที่มีความเสี่ยงที่จะเกิดโรคซึมเศร้า มีจำนวนเพิ่มขึ้น และส่วนใหญ่จะไม่ได้รับการรักษาหรือไม่ทราบว่าตนเองเกิดภาวะซึมเศร้า ซึ่งนำไปสู่การเกิดโรคซึมเศร้า รวมถึงบุคคลในครอบครัวหรือผู้ที่ใกล้ชิดก็ไม่ทราบเช่นเดียวกัน ทำให้ไม่ได้รับการรักษาที่เหมาะสมและทันท่วงที ดังนั้นผู้ที่มีความเสี่ยงที่จะเกิดโรคซึมเศร้า จึงแสดงออกผ่านทางเครือข่ายสังคมออนไลน์ เนื่องจากเครือข่ายสังคมออนไลน์เป็นรูปแบบการสื่อสารผ่านช่องทางที่ไม่ต้องอาศัยน้ำเสียง การสบตา และการแสดงออกทางสีหน้า (กันยา พาณิชย์ศิริ และเบญจพร ตันตสูติ, 2559) ซึ่งสามารถแสดงออกได้จากการโพสต์ข้อความ หรือรูปภาพ ในเครือข่ายสังคมออนไลน์อย่างต่อเนื่อง นอกจากนั้นแล้วเครือข่ายสังคมออนไลน์ยังมีสัญญาณที่บ่งบอกถึงภาวะซึมเศร้ารายบุคคล ความกังวลเกี่ยวกับความสัมพันธ์ต่าง ๆ โดยวัดการเพิ่มขึ้นของเนื้อหาในด้านลบ (Choudhury, 2013)

ข้อมูลปริมาณมหาศาลจากเครือข่ายสังคมออนไลน์ซึ่งมีผู้ใช้งานมากกว่า 3,200 ล้านคนทั่วโลก (Tjepkema, 2019) ได้ถูกนำมาใช้ประโยชน์ในด้านต่าง ๆ เช่น ด้านธุรกิจ โดยนำมาวิเคราะห์หรือคาดคะเนเหตุการณ์ในอนาคต ด้านสาธารณสุข ซึ่งใช้ในการพัฒนาแนวทางใหม่สำหรับการวิเคราะห์ข้อมูลรายงานการป่วยของกระทรวงสาธารณสุข (ณัฐพงษ์ หนูสิงห์, 2560) ด้านข้อมูลจราจรทางคอมพิวเตอร์ (Log file) โดยใช้ในการเก็บรายละเอียดการติดต่อสื่อสารผ่านทางระบบเครือข่ายในรูปแบบของไฟล์ และใช้ในการสืบค้นข้อมูล และยังสามารถใช้ข้อมูลจากสื่อสังคมออนไลน์อื่น ๆ โดยการใช้ส่วนต่อประสานโปรแกรมประยุกต์ของเครือข่ายสังคมออนไลน์ (Social

Network API) ในการวิเคราะห์ข้อมูลจากสื่อสังคมออนไลน์ เช่น การใช้ข้อมูลจากผู้ใช้งานทวิตเตอร์ (Twitter) ในการวิเคราะห์ข้อมูลด้านสุขภาพ โดยสามารถจำแนกกลุ่มของผู้รักสุขภาพและกลุ่มผู้ที่ไม่ดูแลสุขภาพได้จากข้อความที่ผู้ใช้ทวิตเตอร์โพสต์ รวมถึงสามารถวัดระดับคะแนนการดูแลสุขภาพได้โดยการดึงข้อมูลจากทวิตเตอร์ และใช้การวิเคราะห์ความรู้สึก (Sentiment Analysis) โดยใช้ข้อมูลที่ได้จากทวิตเตอร์ (Kashyap, 2014)

จากการศึกษางานวิจัยที่เกี่ยวกับการวิเคราะห์ภาวะซึมเศร้าจากเครือข่ายสังคมออนไลน์ พบว่า งานวิจัยส่วนใหญ่เน้นไปที่การใช้การวิเคราะห์ความรู้สึก โดยเป็นการจำแนกประเภทและความหมายของข้อความแสดงความคิดเห็นที่ได้จากเครือข่ายสังคมออนไลน์ เพื่อใช้ในการวิเคราะห์ภาวะซึมเศร้า ซึ่งในประเทศไทย ทวิตเตอร์เป็นเครือข่ายสังคมออนไลน์ที่ได้รับความนิยมสูงในปัจจุบัน และสามารถเผยแพร่ข้อมูลได้รวดเร็ว โดยผู้ใช้สามารถแสดงความคิดเห็นผ่านทางทวิตเตอร์ได้อย่างอิสระและรวดเร็ว และมีอัตราการใช้งานในประเทศไทยเพิ่มขึ้นร้อยละ 35 (Hootsuite, 2019) โดยแบ่งเป็นเพศหญิง ร้อยละ 51 และเพศชาย ร้อยละ 49 โดยช่วงวัยที่ใช้งานมากที่สุด ได้แก่ ช่วงอายุ 16-24 ปี (คิดเป็นร้อยละ 40) รองลงมาคือ ช่วงอายุ 25-34 ปี (คิดเป็นร้อยละ 26) ช่วงอายุ 35-44 ปี (คิดเป็นร้อยละ 19) ช่วงอายุ 45-54 ปี (คิดเป็นร้อยละ 11) และสุดท้าย ช่วงอายุ 55-64 ปี (คิดเป็นร้อยละ 4) ซึ่งสอดคล้องกับความชุกของภาวะซึมเศร้าในนักศึกษาซึ่งเป็นวัยรุ่นตอนปลาย (อายุ 18-24 ปี) ที่มีตัวเลขสูงถึงร้อยละ 53.50 (โปรยทิพย์ สันตะพันธุ์, ศิริญพร บุสหงส์ และเชาวลิต ศรีเสริม, 2562) และจากการศึกษาเอกสารที่เกี่ยวข้องกับสุขภาพจิตและภาวะซึมเศร้า พบว่า ภาวะซึมเศร้าเป็นสิ่งที่ต้องรายงานด้วยตนเองบ่อยที่สุด โดยเฉพาะการประเมินตนเองในขั้นต้น ซึ่งสอดคล้องกับบริบทของผู้ใช้งานทวิตเตอร์ ซึ่งส่วนใหญ่จะแสดงความคิดเห็นโดยใช้นามแฝง และแสดงความคิดเห็นสั้น ๆ แต่มีแนวโน้มที่จะแสดงความคิดเห็นตลอดทั้งวัน

ดังนั้น ในงานวิจัยนี้จึงใช้ข้อมูลที่ได้จากการแสดงความคิดเห็นในทวิตเตอร์สำหรับการวิเคราะห์ภาวะซึมเศร้า ซึ่งการวิเคราะห์ภาวะซึมเศร้าจากข้อความแสดงความคิดเห็นทวิตเตอร์เพียงอย่างเดียวอาจไม่เพียงพอ จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่า งานวิจัยส่วนใหญ่เน้นไปที่การใช้ข้อความในเครือข่ายสังคมออนไลน์มาวิเคราะห์เพียงอย่างเดียว โดยไม่มีการใช้ปัจจัยอื่นหรือคุณลักษณะอื่น ๆ ร่วมในการวิเคราะห์ ได้แก่ คุณลักษณะด้านกายภาพ เช่น เพศ อายุ จำนวนเพื่อนในเครือข่ายสังคมออนไลน์ จำนวนผู้ติดตาม จำนวนผู้กดไลค์ ช่วงเวลาที่ใช้งานเครือข่ายสังคมออนไลน์ และคุณลักษณะของข้อความ เช่น การใช้งานอีโมติคอนในข้อความ และความยาวของข้อความแสดงความคิดเห็น รวมถึงการใช้งานรูปภาพในการแสดงความคิดเห็น นอกจากนั้นยังพบว่างานวิจัยส่วนใหญ่มีค่าความถูกต้องอยู่ในระดับต่ำ (Choudhury, 2013; Reece and Danforth, 2017; Aldarwish and Ahmad, 2017; Katchapakirin et al., 2018) อีกทั้งงานวิจัยบางส่วนไม่ได้มีการเปรียบเทียบเทคนิคอื่น ๆ เพื่อหาวิธีที่ดีที่สุด และยังไม่มีงานวิจัยใดที่วิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ที่สามารถบ่งบอกระดับคะแนนภาวะซึมเศร้าได้

จึงนำมาสู่แนวคิดในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยสามารถวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ซึ่งบ่งบอกถึงระดับคะแนนของความเสี่ยงที่จะเกิดภาวะซึมเศร้าที่เหมาะสมกับบริบทในสังคมไทย และมีค่าความถูกต้องในระดับที่ยอมรับได้ ในส่วนของข้อมูลที่นำมาใช้ในการวิเคราะห์นั้นได้ใช้ปัจจัย หรือคุณลักษณะอื่น ๆ นอกเหนือจากข้อความ เช่น คุณลักษณะทางกายภาพส่วนบุคคล คุณลักษณะของรูปภาพ มาร่วมในการวิเคราะห์ โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การประมวลผลรูปภาพ (Image Processing) การวิเคราะห์ความรู้สึก (Sentiment Analysis) และเทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

1.2 วัตถุประสงค์ของการวิจัย

เพื่อออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

1.3 สมมติฐานการวิจัย

แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์สามารถประมวลผลได้ความถูกต้องร้อยละ 80 ขึ้นไป

1.4 ข้อตกลงเบื้องต้น

1.4.1 การเก็บรวบรวมข้อมูลในอุปกรณ์ที่แตกต่างกัน ไม่มีผลกระทบต่อขั้นตอนวิธีในงานวิจัยนี้

1.4.2 ข้อมูลที่ได้จากเครือข่ายสังคมออนไลน์ เป็นข้อความที่มีความหมาย และสามารถเข้าใจได้

1.5 ขอบเขตของการวิจัย

การวิจัยนี้เป็นการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ที่มุ่งเน้นการวิเคราะห์ความเสี่ยงจากข้อมูลที่อยู่ในเครือข่ายสังคมออนไลน์แบบสาธารณะเท่านั้น คือ ทวิตเตอร์ ซึ่งข้อมูลที่ใช้ ได้แก่ ข้อความแสดงความคิดเห็น คุณลักษณะทางกายภาพ และคุณลักษณะของรูปภาพ โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติ การประมวลผลรูปภาพ การวิเคราะห์ความรู้สึก และเทคนิคการเรียนรู้ของเครื่อง ซึ่งแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า สามารถบ่งบอกถึงระดับคะแนนภาวะซึมเศร้า

มหาวิทยาลัยเทคโนโลยีสุรนารี

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1.6.1 ได้แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าที่มีความถูกต้องในระดับที่ยอมรับได้

1.6.2 ได้รับรู้ระดับภาวะสุขภาพจิตเพื่อให้สามารถเข้าถึงการรักษาได้โดยเร็ว

1.7 คำอธิบายศัพท์

1.7.1 ภาวะซึมเศร้า หมายถึง ภาวะที่จิตใจหม่นหมอง หดหู่ เศร้า รวมถึงความรู้สึกท้อแท้ หมดหวังมองโลกแง่ร้าย รู้สึกมีคุณค่าในตนเองต่ำ ตำหนิตนเอง ความรู้สึกเหล่านี้เป็นอยู่เป็นระยะเวลานานและเกี่ยวข้องกับการสูญเสีย ซึ่งภาวะซึมเศร้านั้นเป็นอาการหนึ่งของโรคซึมเศร้า (สถาบันสุขภาพจิต กระทรวงสาธารณสุขและสมาคมจิตเวชแห่งประเทศไทย, 2536)

1.7.2 โรคซึมเศร้า หมายถึง โรคทางจิตเวชประเภทหนึ่งที่ส่งผลให้ผู้ป่วยมีอาการทั้งทางด้านร่างกาย จิตใจ และความคิด โดยอาการต่าง ๆ เหล่านี้จะคอยสร้างผลกระทบต่อการดำเนินชีวิตประจำวันอย่างมาก เช่น รับประทานอาหารได้น้อยลง เบื่ออาหาร นอนไม่หลับ มีความรู้สึกว่าตัวเองไม่มีความสุขกับชีวิต มีความวิตกกังวล ที่สำคัญ ผู้ป่วยไม่มีความสามารถในการประสานความคิดและความรู้สึกของตัวเองเพื่อใช้ในการแก้ไขปัญหาได้อย่างดีพอ (พจนา เปลี่ยนเกิด, 2557)

1.7.3 แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (Depression Risk Analysis Model: DeRAM) หมายถึง การจำลองแบบปัญหา หรือการจำลองรูปแบบทางเลือกในการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ด้วยการวิเคราะห์จากข้อมูลที่อยู่ในเครือข่ายสังคมออนไลน์ และทำการเรียงลำดับความเสี่ยงของการเกิดภาวะซึมเศร้าจากสูงสุดสู่ระดับต่ำกว่าตามลำดับ

1.7.4 เครือข่ายสังคมออนไลน์ (Online Social Network) หมายถึง เครือข่ายที่มนุษย์สามารถเข้าถึงได้ผ่านทางอินเทอร์เน็ต สามารถติดต่อสื่อสาร เล่าเรื่อง แสดงความคิดเห็นแลกเปลี่ยนประสบการณ์ หรือทำกิจกรรมร่วมกันผ่านระบบเครือข่ายอินเทอร์เน็ตที่ให้บริการ ตัวอย่างของเครือข่ายสังคมออนไลน์ เช่น เฟสบุค (Facebook) ทวิตเตอร์ (Twitter) ฯลฯ

มหาวิทยาลัยเทคโนโลยีสุรนารี

บทที่ 2
ปริทัศน์วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

จากการศึกษาทฤษฎีและเอกสารงานวิจัยที่เกี่ยวข้องในการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์นั้น พบว่า มีแนวคิด ทฤษฎี และงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

2.1 ข้อมูลพื้นฐานที่นำมาใช้ในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

2.1.1 ภาวะซึมเศร้า

2.1.2 เครือข่ายสังคมออนไลน์ (Social Network)

2.1.3 สื่อสังคมออนไลน์ (Social Media)

2.1.4 ทวิตเตอร์ (Twitter)

2.2 เทคโนโลยีในการรวบรวมข้อมูล

2.2.1 โปรแกรมประยุกต์บนเว็บ (Web Application)

2.2.2 ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ (Twitter API)

2.2.3 ส่วนต่อประสานโปรแกรมประยุกต์ในไพทอน (Python API)

2.3 เทคโนโลยีในการประมวลผลข้อมูล

2.3.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

2.3.2 การวิเคราะห์ความรู้สึก (Sentiment Analysis)

2.3.3 การประมวลผลรูปภาพ (Image Processing)

2.3.4 การตรวจจับใบหน้า (Face Detection)

2.3.5 การประมวลผลสัญรูปอารมณ์ (Emoticon Processing)

2.4 เทคนิคที่นำมาใช้ในการสร้างแบบจำลอง

2.4.1 การเรียนรู้ของเครื่อง (Machine Learning)

2.4.2 นาอีฟเบย์ (Naïve Bayes)

2.4.3 ต้นไม้การตัดสินใจ (Decision Tree)

2.4.4 การสุ่มป่าไม้ (Random Forest)

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.4.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
2.4.6 การเรียนรู้เชิงลึก (Deep Learning)
2.4.7 เพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron)
2.5 การประเมินผลแบบจำลอง
2.5.1 เทคนิคการประเมินแบบจำลอง (Model Evaluation Techniques)
2.5.2 การประเมินแบบจำลองด้วยเมตริกซ์ (Model Evaluation Metrics)
2.6 งานวิจัยที่เกี่ยวข้อง
2.6.1 การวิเคราะห์ความรู้สึกจากข้อมูลในเครือข่ายสังคมออนไลน์
2.6.2 การวิเคราะห์ภาวะซึมเศร้าจากข้อมูลในเครือข่ายสังคมออนไลน์
2.6.3 การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า
2.7 กรอบแนวคิดการวิจัย

โดยมีรายละเอียดของแต่ละหัวข้อ ดังนี้

2.1 ข้อมูลพื้นฐานที่นำมาใช้ในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

2.1.1 ภาวะซึมเศร้า

ภาวะซึมเศร้า หรือ Depression มาจากคำว่า Depriver ในภาษาละติน มีความหมายว่ากดต่ำหรือจมลงต่ำกว่าตำแหน่งเดิม

สถาบันสุขภาพจิต กระทรวงสาธารณสุขและสมาคมจิตเวชแห่งประเทศไทย (2536) ได้ให้ความหมายของภาวะซึมเศร้าไว้ว่า ภาวะซึมเศร้า หมายถึง ภาวะที่จิตใจหม่นหมอง หดหู่ เศร้า รวมถึงความรู้สึกท้อแท้ หมดหวังมองโลกแง่ร้ายรู้สึกมีคุณค่าในตนเองต่ำ ตำหนิตนเอง ความรู้สึกเหล่านี้เป็นอยู่เป็นระยะเวลานานและเกี่ยวข้องกับการสูญเสีย

WHO (2018) ได้อธิบายถึงภาวะซึมเศร้า คือ อาการของโรคทางจิตเวชที่แสดงถึงอารมณ์เศร้า สูญเสียความสนใจและคุณค่าในตัวเอง รู้สึกผิดและต่ำต้อย นอนไม่หลับ เบื่ออาหาร และส่วนใหญ่เกิดขึ้นด้วยอาการวิตกกังวล

Parekh (2017) กล่าวว่า อาการซึมเศร้า หรือโรคซึมเศร้า เป็นโรคที่พบบ่อยและร้ายแรงซึ่งส่งผลเสียต่อความรู้สึกนึกคิด อาการซึมเศร้าทำให้เกิดความรู้สึกเศร้าและสูญเสียความสนใจในกิจกรรมที่เคยทำ นำไปสู่ปัญหาทางอารมณ์ และร่างกายส่งผลให้ความสามารถในการทำงานลดลง

Marra (2004) ได้ให้ความหมายของ ภาวะซึมเศร้าคือ ภาวะที่สูญเสียความคิดริเริ่ม ความสามารถ และความสนใจในกิจกรรมต่าง ๆ มีความกระวนกระวายในทุกสิ่งที่เกิดขึ้น และสูญเสียบางสิ่งที่สำคัญไป สูญเสียพลังงานในการทำบางสิ่งบางอย่าง มีความตกใจ กลัว และคิดในแง่ลบเกี่ยวกับอนาคตที่จะเกิดขึ้น

นันทิรา หงษ์ศรีสุวรรณ์ (2559) กล่าวว่า ภาวะซึมเศร้า หมายถึง ภาวะจิตใจที่แสดงออกถึงความผิดปกติของอารมณ์ เช่น อารมณ์เศร้า ไม่มีความสุข เบื่อหน่าย หดหู่ ท้อแท้ สิ้นหวัง เซื่องซึม นอนไม่หลับ เบื่ออาหาร ขาดสมาธิ วิตกกังวล มองโลกในแง่ลบ ซึ่งอาจจะส่งผลเสียต่อคุณภาพชีวิต และอาจนำไปสู่การฆ่าตัวตายในที่สุด

บุษราคัม จิตอารีย์ (2555) ได้ให้ความหมายของภาวะซึมเศร้าไว้ว่า ภาวะซึมเศร้าคือแปรปรวนทางด้านความคิด อารมณ์ แรงจูงใจ ร่างกายและพฤติกรรม ซึ่งเป็นผลจากการประมวลเรื่องราวผ่านกระบวนการคิดจนเกิดความคิดอัตโนมัติในทางลบเกี่ยวกับตนเอง สิ่งแวดล้อมและอนาคต เมื่อบุคคลประสบภาวะวิกฤตในชีวิตการประมวลเรื่องราวที่ผิดพลาดในอดีตจะกระตุ้นบุคลนั้นให้เกิดความคิดอัตโนมัติในทางลบ และ ส่งผลให้เกิดภาวะซึมเศร้าตามมาภาวะซึมเศร้ายิ่งเพิ่มมากขึ้นความคิดอัตโนมัติในทางลบจะเพิ่มมากขึ้นทั้งความถี่และความรุนแรงความคิดที่เป็นเหตุผลลดลงและเกิดอาการและอาการแสดงของภาวะซึมเศร้า ทั้งทางด้านอารมณ์ความคิด แรงจูงใจ ร่างกาย และพฤติกรรม เช่น จิตใจที่หม่นหมอง หดหู่ เศร้า รวมถึงความรู้สึกท้อแท้หมดหวังไม่สนใจกิจกรรมนอนไม่หลับ น้ำหนักตัวเพิ่มหรือลด ปวดท้อง ปวดศีรษะ ฯลฯ

จากความหมายข้างต้นดังนั้นจึงสรุปได้ว่า ภาวะซึมเศร้าหมายถึง อาการหนึ่งของโรคซึมเศร้าซึ่งแสดงออกทั้งทางด้านจิตใจและอารมณ์ ด้านพฤติกรรม และด้านร่างกาย เช่น เศร้าใจ ท้อแท้หมดหวัง รู้สึกแย่กับตนเอง เบื่ออาหาร นอนไม่หลับ มีความคิดอยากทำร้ายตนเอง ฯลฯ ส่งผลให้ความสามารถในการทำงานลดลง และนำไปสู่การเกิดโรคซึมเศร้า ซึ่งก่อให้เกิดผลเสียต่อคุณภาพชีวิต และปัญหาการฆ่าตัวตายตามมา

อาการบ่งชี้ของภาวะซึมเศร้าสามารถแสดงออกได้ทั้งทางด้านจิตใจ ด้านพฤติกรรม และด้านร่างกาย และนำไปสู่การเกิดโรคซึมเศร้าได้ ดังแสดงในภาพที่ 2.1 ซึ่งมีรายละเอียดดังต่อไปนี้

ด้านจิตใจและอารมณ์

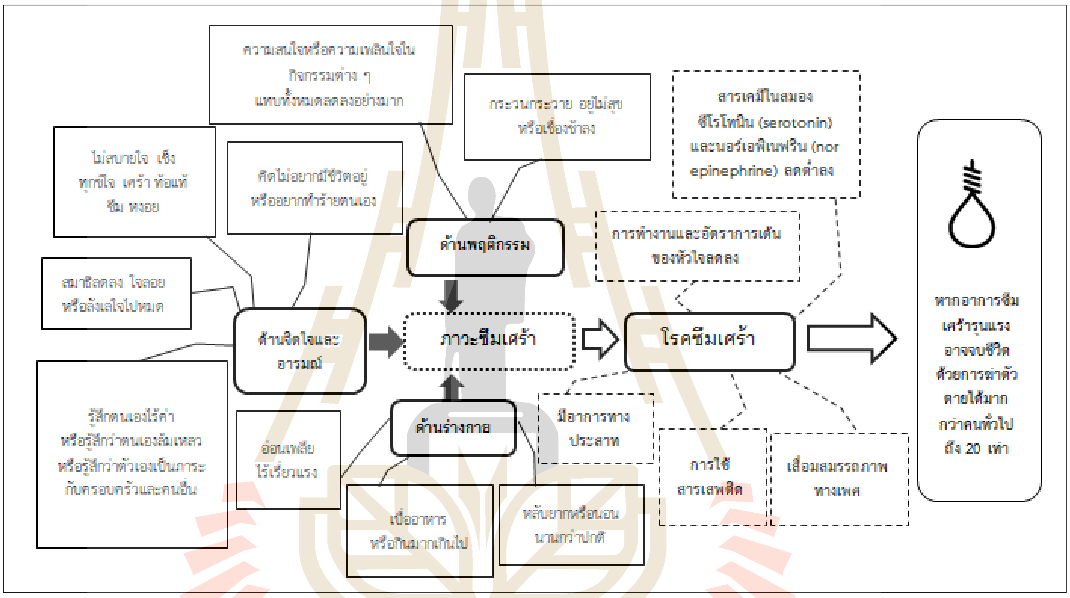
1. มีอารมณ์เศร้า ไม่สบายใจ เซ็ง ทุกข์ใจ เศร้า ท้อแท้ ซึม หงอย
2. รู้สึกตนเองไร้ค่า หรือรู้สึกว่าตนเองล้มเหลว
3. สมาธิลดลง ใจลอย หรือลังเลใจไปหมด
4. คิดไม่อยากมีชีวิตอยู่ หรืออยากทำร้ายตนเอง

ด้านพฤติกรรม

5. ความสนใจหรือความเพลินใจในกิจกรรมต่างๆ แทบทั้งหมดลดลงอย่างมาก
6. กระวนกระวาย อยู่ไม่สุข หรือเชื่องช้าลง

ด้านร่างกาย

7. น้ำหนักลดลงหรือเพิ่มขึ้นมาก (น้ำหนักเปลี่ยนแปลงมากกว่าร้อยละ 5 ต่อเดือน) หรือมีการเบื่ออาหารหรือเจริญอาหารมาก
8. นอนไม่หลับ หรือหลับมากไป
9. อ่อนเพลีย ไร้เรี่ยวแรง



ภาพที่ 2.1 ภาวะซึมเศร้า

การวินิจฉัยโรคซึมเศร้าสามารถวินิจฉัยได้จากการประเมินภาวะซึมเศร้าร่วมกับการวินิจฉัยทางห้องปฏิบัติการ โดยแพทย์ โดยมีกระบวนการดังนี้

1. ซักประวัติ สัมภาษณ์ ทำแบบทดสอบประเมินภาวะซึมเศร้า โดยแบ่งแบบทดสอบเป็น 2 แบบ คือ 1) แบบสัมภาษณ์เชิงโครงสร้าง และแบบประเมินภาวะซึมเศร้า 9 คำถาม

แบบสัมภาษณ์เชิงโครงสร้างเพื่อการวินิจฉัยโรคตามเกณฑ์การวินิจฉัยโรคซึมเศร้าของ DSM-IV และ ICD 10 คือ M.I.N.I.14 ในหมวด A-Major depressive episode, B-Dysthymia, C-Suicidality และ D-Mania/Hypomania (พันธุ์นภา กิตติรัตนไพบูลย์ และคณะ, 2550)

แบบประเมินภาวะซึมเศร้า 9 คำถาม (9Q) และเครื่องมือมาตรฐานวัดระดับความรุนแรงของโรคซึมเศร้า HRSD-17 (อัจฉราพร สี่หิรัญวงศ์, 2544) ประกอบไปด้วยคำถาม 17 ข้อ ซึ่งประเมิน 4 ด้าน ได้แก่ กลุ่มอาการซึมเศร้าและวิตก กังวล การนอนหลับ การแสดงออกทางกายภาพ และความคิด

2. ตรวจร่างกาย ในเบื้องต้น แพทย์จะใช้การตรวจร่างกายและถามคำถามเกี่ยวกับสุขภาพของผู้ป่วย เพราะบางครั้งภาวะซึมเศร้าที่เกิดขึ้นอาจเกี่ยวเนื่องกับปัญหาสุขภาพทางกายชนิดอื่นได้

3. ตรวจทางห้องปฏิบัติการ เช่น ตรวจปัสสาวะ ตรวจเลือด CT scan หรือ MRT ของสมอง

ปัจจุบันได้มีการพัฒนาแบบคัดกรองภาวะซึมเศร้าหลายรูปแบบ เช่น แบบประเมินภาวะซึมเศร้าสำหรับวัยรุ่น (CES-D : Center for Epidemiological Studies-Depression Scale) (Radloff, 1997) ดังแสดงในภาพที่ 2.2 เป็นเครื่องมือประเภทให้ผู้เข้ารับการประเมินตอบคำถามด้วยตนเอง (Self-report) ประกอบด้วยข้อคำถาม 20 ข้อ ซึ่งเกี่ยวกับอาการซึมเศร้าในระยะเวลา 1 สัปดาห์ที่ผ่านมา ลักษณะแบบสอบถามเป็นแบบมาตราส่วนประมาณค่า (Rating Scale) 4 ระดับ

Center for Epidemiologic Studies Depression Scale (CES-D), NIMH

Below is a list of the ways you might have felt or behaved. Please tell me how often you have felt this way during the past wee

	During the Past Week			
	Rarely or none of the time (less than 1 day)	Some or a little of the time (1-2 days)	Occasionally or a moderate amount of time (3-4 days)	Most or all of the time (5-7 days)
1. I was bothered by things that usually don't bother me.	☐	☐	☐	☐
2. I did not feel like eating; my appetite was poor.	☐	☐	☐	☐
3. I felt that I could not shake off the blues even with help from my family or friends.	☐	☐	☐	☐
4. I felt I was just as good as other people.	☐	☐	☐	☐
5. I had trouble keeping my mind on what I was doing.	☐	☐	☐	☐
6. I felt depressed.	☐	☐	☐	☐
7. I felt that everything I did was an effort.	☐	☐	☐	☐
8. I felt hopeful about the future.	☐	☐	☐	☐
9. I thought my life had been a failure.	☐	☐	☐	☐
10. I felt fearful.	☐	☐	☐	☐
11. My sleep was restless.	☐	☐	☐	☐
12. I was happy.	☐	☐	☐	☐
13. I talked less than usual.	☐	☐	☐	☐
14. I felt lonely.	☐	☐	☐	☐
15. People were unfriendly.	☐	☐	☐	☐
16. I enjoyed life.	☐	☐	☐	☐
17. I had crying spells.	☐	☐	☐	☐
18. I felt sad.	☐	☐	☐	☐
19. I felt that people dislike me.	☐	☐	☐	☐
20. I could not get "going."	☐	☐	☐	☐

SCORING: zero for answers in the first column, 1 for answers in the second column, 2 for answers in the third column, 3 for answers in the fourth column. The scoring of positive items is reversed. Possible range of scores is zero to 60, with the higher scores indicating the presence of more symptomatology.

ภาพที่ 2.2 CES-D : Center for Epidemiological Studies-Depression Scale (Radloff, 1997)

มหาวิทยาลัยเทคโนโลยีสุรนารี

ส่วนแบบประเมินภาวะซึมเศร้า 9 คำถาม (PHQ-9 : The Patient Health Questionnaire Depression Scale) เป็นแบบประเมินประกอบด้วยข้อคำถามจำนวน 9 ข้อคำถาม โดยผู้เข้ารับการประเมินสามารถตอบคำถามได้ด้วยตนเอง ซึ่งข้อคำถามจะเกี่ยวข้องกับอาการซึมเศร้าในระยะเวลา 2 สัปดาห์ที่ผ่านมา ดังแสดงในภาพที่ 2.3

PATIENT HEALTH QUESTIONNAIRE (PHQ-9)

NAME:____________________________ DATE:______________

Over the last *2 weeks*, how often have you been bothered by any of the following problems? *(use "✓" to indicate your answer)*	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself—or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite — being so figety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead, or of hurting yourself	0	1	2	3
	add columns		+	+

(Healthcare professional: For interpretation of TOTAL, please refer to accompanying scoring card). TOTAL:

10. If you checked off *any problems*, how *difficult* have these problems made it for you to do your work, take care of things at home, or get along with other people?	Not difficult at all ______ Somewhat difficult ______ Very difficult ______ Extremely difficult ______

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาพที่ 2.3 PHQ-9 : The Patient Health Questionnaire depression scale (Pfizer, 2005)

ซึ่งงานวิจัยพบว่า แบบประเมินทั้ง 2 แบบ CES-D และ PHQ-9 ไม่มีความแตกต่างกันอย่างมีนัยสำคัญ (Milette et al., 2010) สำหรับประเทศไทยแบ่งแบบประเมินภาวะซึมเศร้าออกเป็น 3 รูปแบบคือ แบบคัดกรองภาวะซึมเศร้า 2 คำถาม (2Q) ดังแสดงในภาพที่ 2.4 และแบบประเมินโรคซึมเศร้า 9 คำถาม (9Q) ดังแสดงในภาพที่ 2.5 และแบบประเมินการฆ่าตัวตาย 8 คำถาม (8Q) (ประชาสัมพันธ์กรมสุขภาพจิต, 2561) สำหรับงานวิจัยนี้ได้ใช้แบบประเมินภาวะซึมเศร้า 9 คำถาม ซึ่งสอดคล้องวัตถุประสงค์ของงานวิจัย และสอดคล้องกับแบบประเมิน PHQ-9 ที่ใช้กันแพร่หลาย

แบบคัดกรองโรคซึมเศร้า 2 คำถาม (2Q)

__ข้อแนะนำ__ ถามด้วยภาษาที่สอดคล้องกับท้องถิ่นหรือเหมาะสมกับผู้ถูกสัมภาษณ์ ควรถามให้ได้คำตอบทีละข้อ ถ้าไม่เข้าใจให้ถามซ้ำ ไม่ควรอธิบายขยายความเพิ่มเติม และให้ผู้ประเมินกาเครื่องหมาย "✓" ในช่องที่ตรงกับคำตอบของผู้รับบริการ

คำถาม	ภาษากลาง	ภาษาอีสาน	ภาษาท้องถิ่นใต้	มี	ไม่มี
1	**ใน 2 สัปดาห์ที่ผ่านมา รวมวันนี้ ท่านรู้สึก** *หดหู่ เศร้า หรือท้อแท้ สิ้นหวัง หรือไม่*	**ใน 2 สัปดาห์ที่ผ่านมา รวมมื่อนี่ เจ้ามีอาการมูนี่บ่** *: อุกอั่ง หงอย เซ็ง หมหวย บ่เป็นตาอยู่ มีแต่อยากไฮ่*	ตั้งแต่ 2 อาทิตย์ก่อน จนถึงวันนี้ (สรรพนาม) รู้สึก เศร้า ใจมันหดหู้ม่าย หรือเหม็ดหวัง มั้งม่าย		
2	**ใน 2 สัปดาห์ที่ผ่านมา รวมวันนี้ ท่านรู้สึก** *เบื่อ ทำอะไร ก็ไม่เพลิดเพลิน หรือไม่*	**ใน 2 สัปดาห์ที่ผ่านมา รวมมื่อนี่ เจ้ามีอาการมูนี่บ่** *: บ่สนใจหยัง บ่อยากเฮ็ดหยัง บม่วนบ่ซืน*	ตั้งแต่ 2 อาทิตย์ก่อน จนถึงวันนี้ (สรรพนาม) รู้สึก เบื่อ ไม่อยากทำไหร ทำไหรกะไม่หนุกมั้งม่าย		

ภาพที่ 2.4 แบบคัดกรองภาวะซึมเศร้า 2 คำถาม

มหาวิทยาลัยเทคโนโลยีสุรนารี

แบบประเมินโรคซึมเศร้า 9 คำถาม (9Q)

<u>ข้อแนะนำ</u> ถามด้วยภาษาที่สอดคล้องกับท้องถิ่นหรือเหมาะสมกับผู้ถูกสัมภาษณ์ ควรถามให้ได้คำตอบทีละข้อ ถ้าไม่เข้าใจให้ถามซ้ำ ไม่ควรอธิบายขยายความเพิ่มเติม พยายามให้ได้คำตอบทุกข้อ จากนั้นรวมคะแนนแล้วแปรผลตามตารางแปรผล ควรแจ้งผลตามแนวทางการแจ้งผลอย่างเหมาะสมพร้อมให้คำแนะนำในการปฏิบัติตัวหรือการดูแลรักษาในขั้นตอนต่อไป

หมายเหตุ ให้เน้นการถามเพื่อค้นหาอาการที่มีใน<u>**ระยะ 2 สัปดาห์**</u>ที่ผ่านมา

ลำดับคำถาม	ภาษากลาง ในช่วง 2 สัปดาห์ที่ผ่านมา รวมทั้งวันนี้ ท่านมีอาการเหล่านี้ บ่อยแค่ไหน	ภาษาอีสาน ใน 2 สัปดาห์ที่ผ่านมา รวมมื่อนี่ เจ้ามีอาการมูนี่ดู่ซำใด่	ภาษาท้องถิ่นใต้ ที่ผ่านมา ตั้งแต่ 2 อาทิตย์ก่อนจนถึงวันนี้ (สรรพนาม) มีอาการพันนี้ บ่อยแค่ไหน	ไม่มีเลย (บ่เคยมี บ่เคยเป็น)	เป็นบางวัน 1-7 วัน (เป็นลางเทื่อ)	เป็นบ่อย >7 วัน (เป็นดู่)	เป็นทุกวัน (เป็นซุมื่อ)
1	เบื่อ ไม่สนใจอยากทำอะไร	บ่อยากเฮ็ดหยัง บ่สนใจเฮ็ดหยัง	เบื่อไม่อยากทำไหร	0	1	2	3
2	ไม่สบายใจ ซึมเศร้า ท้อแท้	บ่ม่วนบ่ซื่น เซิง หงอย	ไม่บายใจ เศร้า ท้อ	0	1	2	3
3	หลับยาก หรือหลับๆ ตื่นๆ หรือหลับมากไป	นอนบ่หลับ หรือหลับๆ ตื่นๆ หรือ นอนบ่อยากลุก	นอนไม่ไคร่หลับ หลับไม่ทีน หรือว่าหลับมากเกิน	0	1	2	3
4	เหนื่อยง่าย หรือ ไม่ค่อยมีแรง	เมื่อย บ่มีแฮง	เหนื่อยง่าย หรือเหม็ดแรง	0	1	2	3
5	เบื่ออาหาร หรือกินมากเกินไป	บ่อยากเข่า บ่อยากน่าม หรือกินหลายโพด	เบื่อไม่อยากกินไหร หรือ กินมากเกิน/กินเหมือนจุกบอก	0	1	2	3
6	รู้สึกไม่ดีกับตัวเอง คิดว่าตัวเองล้มเหลวหรือทำให้ตนเองหรือครอบครัวผิดหวัง	คิดว่าเจ้าของบ่ดี	รู้สึกว่าตัวเองไม่ได้เรื่อง ทำไหรก้าไม่เข้าท่า/ทำไหรก้าไม่ดี ทำให้ตัวเองหรือครอบครัวผิดหวัง	0	1	2	3
7	สมาธิไม่ดีเวลาทำอะไร เช่น ดูโทรทัศน์ ฟังวิทยุ หรือทำงานที่ต้องใช้ความตั้งใจ	คึดหยังกะบ่ออก เฮ็ดหยังกะลืม	ใจไม่นิ่ง ใจไม่อยู่กับตัว ใจไม่ทีน เช่น แลโทรทัศน์ ฟังวิทยุ หรือว่าทำไหรที่ต้องตั้งใจ	0	1	2	3
8	พูดช้า ทำอะไรช้าลงจนคนอื่นสังเกตเห็นได้ หรือกระสับกระส่ายไม่สามารถอยู่นิ่งได้เหมือนที่เคยเป็น	เว่ากะซ้า เฮ็ดหยังกะซ้า หรือ หนหวย บ่เป็นตาอยู่	เนื้อย แหลงช้า ทำไหรช้า หรือว่าอยู่ไม่นิ่ง ซุ่นหว่าเดิม	0	1	2	3
9	คิดทำร้ายตนเอง หรือคิดว่าถ้าตายไปคงจะดี	คึดอยากตาย บ่อยากอยู่	คิดทำร้ายตัวเอง หรือตายดีหวา	0	1	2	3

คะแนนรวมทั้งหมด [ ]

คะแนนรวม	การแปรผล
<7	ไม่มีอาการของโรคซึมเศร้าหรือมีอาการของโรคซึมเศร้าระดับน้อยมาก
7 - 12	มีอาการของโรคซึมเศร้า ระดับน้อย
13 - 18	มีอาการของโรคซึมเศร้า ระดับปานกลาง
≥19	มีอาการของโรคซึมเศร้า ระดับรุนแรง

ภาพที่ 2.5 แบบประเมินโรคซึมเศร้า 9 คำถาม

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.1.2 เครือข่ายสังคมออนไลน์ (Social Network)

เครือข่ายสังคม หมายถึง กลุ่มบุคคลที่ติดต่อสื่อสารกันโดยผ่านสื่อสังคม (Social Media) ซึ่งนอกจากจะส่งข่าวสารข้อมูลแลกเปลี่ยนกันแล้ว ยังอาจจะร่วมกันทำกิจกรรมที่สนใจด้วยกัน (Aroonpipoon, 2012) เครือข่ายสังคมออนไลน์เป็นสิ่งที่เกิดจากการพัฒนาเทคโนโลยีเว็บในระยะที่ 2 (WEB 2.0) ก่อให้เกิดปฏิสัมพันธ์ระหว่างกลุ่มคนบนเครือข่ายอินเทอร์เน็ตรวมตัวกันโดยมีโครงสร้างทางสังคมตามทฤษฎีเครือข่าย (Network Theory) คือประกอบด้วยส่วนที่เป็นจุด (Node) และส่วนของความสัมพันธ์ (Ties หรือ Relationship) ทั้งนี้ จุด (Node) เปรียบเสมือนคน ๆ หนึ่งในระบบเครือข่ายที่อาจจะมีลักษณะความสัมพันธ์หลายรูปแบบกับคนอื่น ๆ ทำให้เกิดลักษณะความสัมพันธ์ในเครือข่ายออนไลน์ (Social Contacts) (นนทรัฐ ไผ่เจริญ, 2557) นอกจากนั้นแล้วภายในเครือข่ายสังคมออนไลน์ยังมีสื่อสังคมออนไลน์ (Social Media) ในรูปแบบของเว็บไซต์ซึ่งสามารถสื่อสารได้ทั้ง 2 ทาง ผ่านทางรูปภาพ วิดีโอภาพเคลื่อนไหว และข้อความ เพื่อเป็นการเชื่อมโยงกับความสนใจและกิจกรรมของผู้อื่น แบ่งปันให้ผู้อื่นที่อยู่ในเครือข่ายของตนได้รับรู้ โดยผู้ใช้บริการสามารถสร้างประวัติส่วนตัวเพื่อแนะตัวตนเองและสร้างเครือข่ายส่วนบุคคลของตนเพื่อติดต่อกับผู้อื่นได้ เช่น ทวิตเตอร์ (Twitter) เฟซบุ๊ก (Facebook) ไลน์ (Line) อินสตราแกรม (Instagram) และ ยูทูบ (YouTube) (เอมิกา เหมมินทร, 2556) ในขณะที่อรรถศิษฐ์ พัฒนะศิริ และ วรลักษณ์ วงศ์โดยหวัง ศิริเจริญ (2559) กล่าวว่า เครือข่ายสังคมจัดเป็นส่วนหนึ่งของสื่อสังคมซึ่งถูกนำ มาประยุกต์ใช้กับสินค้าและบริการต่าง ๆ หรือเนื้อหาที่ผู้ใช้สร้างขึ้นมาซึ่งประกอบไปด้วยบทสนทนา บทความ รูปภาพ สูตรอาหาร และอื่น ๆ ที่ผู้ใช้เผยแพร่ไปยังผู้ใช้รายอื่น ๆ บนสื่อสังคมต่าง ๆ

ดังนั้น เครือข่ายสังคมออนไลน์ (Social Network) จึงเป็นเสมือนสังคมที่เชื่อมโยงผู้คนไว้ด้วยกันผ่านทางเทคโนโลยีบนอินเทอร์เน็ตหรือเว็บไซต์ โดยผู้ใช้มีส่วนร่วมในการสร้างสรรค์สิ่งต่าง ๆ ลงบนเว็บไซต์ร่วมกัน สามารถสร้างเนื้อหาเพื่อแบ่งปันกัน และโต้ตอบข้อมูลระหว่างผู้ใช้งานด้วยกัน ผ่านทางบริการต่าง ๆ ที่เกิดขึ้นบนอินเทอร์เน็ต

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.1.3 สื่อสังคม (Social Media)

ราชบัณฑิตยสถานได้บัญญัติคำว่า “Social Media” ไว้ว่า “สื่อสังคม” หมายถึงสื่ออิเล็กทรอนิกส์ ซึ่งเป็นสื่อกลางที่ให้บุคคลทั่วไปมีส่วนร่วมสร้างและแลกเปลี่ยนความคิดเห็นต่าง ๆ ผ่านอินเทอร์เน็ตได้

อรรถศิษฐ์ พัฒนะศิริ และ วรลักษณ์ วงศ์โดยหวัง ศิริเจริญ (2559) กล่าวว่า สื่อสังคม (Social Media) หมายถึง การประยุกต์ใช้เทคโนโลยีซึ่งอยู่บนพื้นฐานของเว็บ (Web-based Technology) และเทคโนโลยีเคลื่อนย้ายได้ (Portable Technology) ซึ่งเปลี่ยนรูปแบบการสื่อสารให้อยู่ในรูปแบบของการสนทนาเชิงโต้ตอบ (Interactive Conversation)

Kaplan และ Haenlein (2010) ได้ให้นิยามของสื่อสังคมไว้ว่า สื่อสังคมคือกลุ่มของโปรแกรมประยุกต์บนอินเทอร์เน็ต (Internet-based Application) ที่ช่วยให้ผู้ใช้สามารถสร้างและแลกเปลี่ยนเนื้อหาซึ่งกันและกันได้

สื่อสังคมประกอบไปด้วย บล็อก (Blog) การค้นหาทางสังคม (Social Search) เครือข่ายสังคม (Social Networking) โดยองค์ประกอบต่าง ๆ เหล่านี้จัดเป็นเครื่องมือที่มีประสิทธิภาพในด้านการตลาดสำหรับการสร้างธุรกิจรูปแบบใหม่ สื่อสังคมถูกนำไปประยุกต์ใช้ในรูปแบบของเครื่องมือและแพลตฟอร์มซึ่งอนุญาตให้ผู้ใช้มีปฏิสัมพันธ์ทางสังคมออนไลน์และสามารถเผยแพร่เนื้อหาที่ตนเองต้องการได้ (Siricharoen, 2012)

ลักษณะสำคัญของสื่อสังคม ที่ถูกนำมาประยุกต์ใช้มากที่สุดคือคำบรรยายลักษณะผู้ใช้ (User Profile) โดยคำบรรยายลักษณะผู้ใช้ นี้ทำให้ผู้ใช้เครือข่ายสังคมต่างๆสามารถรวบรวมบัญชีรายชื่อของเพื่อนและยังสามารถค้นหาผู้ใช้รายอื่น ๆ ที่มีความสนใจในเรื่องใดเรื่องหนึ่งตรงกันอีกด้วย เครือข่ายสังคมออนไลน์ที่มีชื่อเสียงมากที่สุดในปัจจุบันคือทวิตเตอร์ (Twitter) เฟซบุ๊ก (Facebook) และอินสตาแกรม (Instagram) สื่อสังคมมีอิทธิพลอย่างสูงต่อผู้ใช้งานทั้งในด้านของการเลือกซื้อสินค้า ความสัมพันธ์กับผู้อื่น และการศึกษา เป็นต้น

ลักษณะสำคัญของสื่อสังคมที่ทำให้สื่อสังคมแตกต่างจากการสื่อสารในรูปแบบอื่น ๆ มี 6 ประการประกอบด้วย

1) การมีส่วนร่วม (Participation) จัดเป็นประโยชน์ที่สำคัญประการหนึ่งที่ผู้ใช้สื่อสังคม ได้รับจากการรวมรวบผู้ใช้อื่น ๆ ให้มีปฏิสัมพันธ์ร่วมกันในการทำกิจกรรมต่าง ๆ

2) สิ่งที่ร่วมกันทำ (Collective) ผู้ใช้สื่อสังคม ส่วนใหญ่จะแนะนำตัวตนของตนเองโดยผ่านคำบรรยายลักษณะผู้ใช้บนสื่อสังคมตลอดจนรวบรวมเนื้อหาต่าง ๆ ที่ตนเองสนใจบนสื่อสังคมที่ตนเองเป็นสมาชิกอยู่และแบ่งปันเนื้อหาต่าง ๆ เหล่านั้นให้แก่สมาชิกอื่น ๆ โดยเนื้อหาที่มีการแบ่งปันนั้นส่วนใหญ่จะเป็นเนื้อหาที่รวบรวมมาจากผู้ใช้อื่น ๆ ผู้ใช้ที่ทำการแบ่งปันไม่ได้เป็นผู้สร้างและเป็นเจ้าของเนื้อหานั้น ๆ เอง

3) ความโปร่งใส (Transparency) สมาชิกในสื่อสังคมสามารถตรวจดู ช่วยแบ่งปันเพิ่มเติม วิจารณ์ และประเมิน สิ่งต่าง ๆ ที่สมาชิกอื่นแบ่งปันกันบนสื่อสังคมได้

4) ความเป็นอิสระ (Independence) สมาชิกของสื่อสังคมสามารถเผยแพร่ข้อมูลและมีส่วนร่วมในกิจกรรมต่างๆบนสื่อสังคมได้โดยอิสระและไม่ขึ้นต่อสมาชิกอื่น ๆ บนสื่อสังคมนั้น ๆ

มหาวิทยาลัยเทคโนโลยีสุรนารี

5) การคงอยู่ (Persistence) สำหรับสื่อสังคมนั้น สิ่งต่างๆที่สมาชิกแบ่งปันหรือเผยแพร่ผ่านสื่อสังคม ไม่ว่าจะเป็นเนื้อหาสาระหรือกิจกรรมใด ๆ จะไม่สูญหายไป ซึ่งสมาชิกสามารถดู เผยแพร่ หรือเพิ่มเติมได้

6) การปรากฏออกมาหรือการเกิดขึ้น (Emergence) สมาชิกของสื่อสังคมไม่สามารถคาดเดา สร้างรูปแบบ ออกแบบ และควบคุมการมีส่วนร่วมหรือการโต้ตอบของสมาชิกในสื่อสังคมได้ (Siricharoen, 2012)

2.1.4 ทวิตเตอร์ (Twitter)

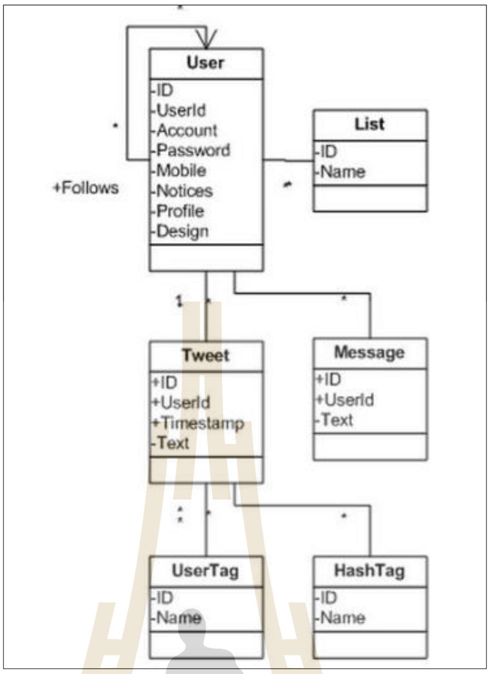
ทวิตเตอร์ (Twitter) เป็นบริการเครือข่ายสังคมออนไลน์จำพวกไมโครบล็อก โดยผู้ใช้สามารถส่งข้อความยาวไม่เกิน 280 ตัวอักษร ว่าตนเองกำลังทำอะไรอยู่ โดยเรียกการส่งข้อความนี้ว่า ทวีต (Tweet)

ทวิตเตอร์ก่อตั้งขึ้นเมื่อเดือนมีนาคม พ.ศ. 2549 โดย แจ็ก ดอร์ซี, บิซ สโตน และอีวาน วิลเลียมส์ เจ้าของบริษัทอ็อบวีไออุสและต่อมาในเดือนกรกฎาคม ก็ได้เปิดตัวเว็บไซต์สังคมออนไลน์ ซึ่งต่อมาได้รับความนิยมจากทั่วโลก โดยมีผู้ใช้ที่ลงทะเบียนเข้าใช้มากกว่า 500 ล้านคนในปี พ.ศ. 2555 รวมไปถึงมีทวีตมากกว่า 340 ล้านทวีตต่อวัน และมีการค้นหาข้อมูลผ่านเว็บไซต์มากกว่า 1,600 ล้านครั้งต่อวันนับตั้งแต่วันเปิดตัว ทวิตเตอร์เป็นเว็บไซต์ 1 ใน 10 อันดับที่มีผู้เข้าใช้งานมากที่สุด โดยในประเทศไทยมีผู้ใช้งานในปี พ.ศ. 2562 จำนวน 4.1 ล้านคน ซึ่งเพิ่มขึ้นจากปี พ.ศ. 2561 ถึง ร้อยละ 35 นอกจากนั้นแล้วผู้ใช้ทวิตเตอร์ ยังมีถึง 40% ที่อยู่ในช่วงอายุ 16-24 ปี (Statista Research Department, 2019) ซึ่งในภายหลัง นอกจากที่จะสามารถทวีตบนเว็บไซต์แล้ว ได้มีการเปิดให้ใช้งานการส่งทวีต ด้วยการส่งบริการข้อความสั้น (SMS) และบนโปรแกรมประยุกต์ในโทรศัพท์มือถือ และสมาร์ตโฟน

สถาปัตยกรรมของทวิตเตอร์สามารถอธิบายได้ด้วย แบบจำลองของครัชเทน (Kruchten’s view model) ประกอบด้วย มุมมองเชิงตรรกะ (Logical View) มุมมองกระบวนการ (Process View) มุมมองทางกายภาพ (Physical View) มุมมองการพัฒนา (Development View) และมุมมองสถานการณ์ (Scenario View) (Janssen and Boer, 2011) โดยมีรายละเอียดดังนี้

1) มุมมองเชิงตรรกะ (Logical View) จะอธิบายถึง คุณสมบัติที่ผู้ใช้ทั่วไปสามารถใช้งานได้ ดังแสดงในภาพที่ 2.6

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 2.6 คลาสไดอะแกรมของทวิตเตอร์

คลาสไดอะแกรมประกอบไปด้วย 6 คลาส ได้แก่ คลาส User คลาส List คลาส Tweet คลาส Message คลาส UserTag และ คลาส HashTag

คลาส User คือคลาสหลักของทวิตเตอร์โดยมีกระบวนการทำงานดังนี้

(1) ผู้ใช้สามารถสร้างทวีตได้

(2) ผู้ใช้สามารถส่งข้อความส่วนตัว (Direct Message) ไปยังผู้ใช้อื่น ๆ ได้

(3) ผู้ใช้สามารถจัดกลุ่มได้

- คลาส List เป็นคลาสที่ใช้สำหรับการระบุกลุ่มของผู้ใช้
- คลาส Tweet ใช้สำหรับการเก็บทวีต ซึ่งสามารถระบุผู้ใช้ เวลา และข้อความในการทวีตได้
- คลาส Message ใช้ในการเก็บข้อความส่วนตัว (Direct Message)
- คลาส UserTag และ HashTag ใช้ในการเก็บแฮชแท็ก (Hashtag) ซึ่งเปรียบเสมือนคำระบุหมวดหมู่ว่าข้อความที่ผู้ใช้เผยแพร่นั้นอยู่ในหมวดหมู่ใด

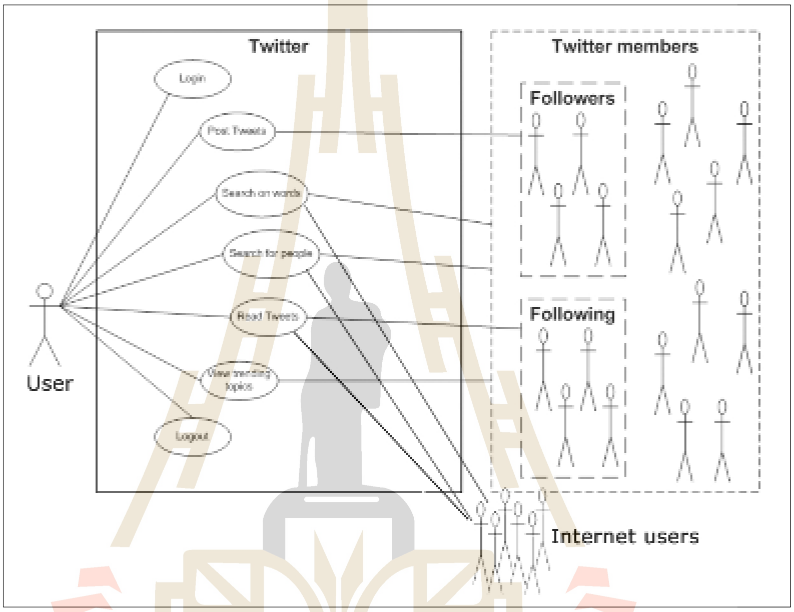
2) มุมมองกระบวนการ (Process View) อธิบายกระบวนการของระบบในแง่ของการสื่อสารและจัดการพฤติกรรมของระบบในขณะทำงาน

3) มุมมองทางกายภาพ (Physical View) มุมมองทางกายภาพอธิบายถึงโทโพโลยีส่วนประกอบซอฟต์แวร์ของทวิตเตอร์และการสื่อสารภายในทวิตเตอร์

มหาวิทยาลัยเทคโนโลยีสุรนารี

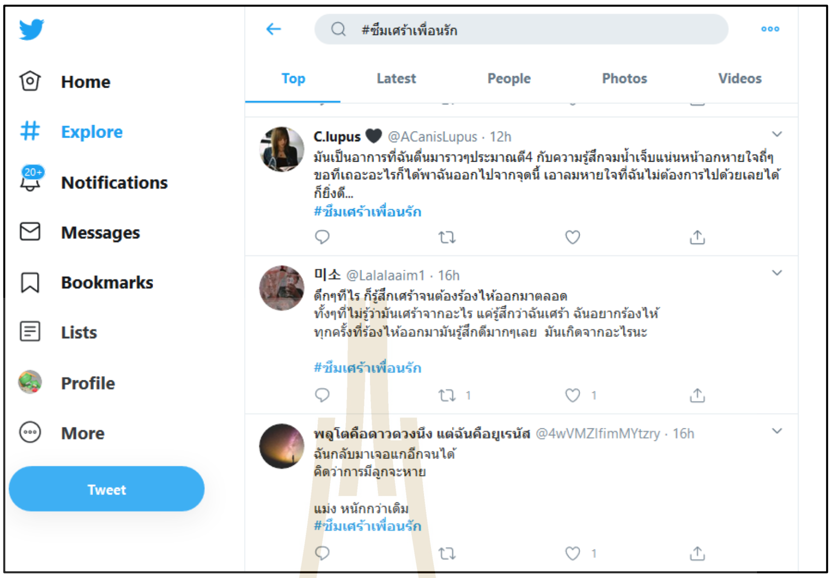
4) มุมมองการพัฒนา (Development View) มุมมองการพัฒนาแสดงให้เห็นทวิตเตอร์จากมุมมองของโปรแกรมเมอร์ผู้พัฒนาและเกี่ยวข้องกับการจัดการซอฟต์แวร์

5) มุมมองสถานการณ์ (Scenario View) มุมมองนี้แสดงให้เห็นถึงสถาปัตยกรรมของทวิตเตอร์โดยแบ่งเป็นชุดขึ้นอยู่กับกรณีและสถานการณ์ที่เกิดขึ้นระหว่างการปฏิสัมพันธ์และการเชื่อมต่อระหว่างผู้ติดตาม / กำลังติดตาม / ทวีตและกระบวนการที่เกี่ยวข้องจะนำเสนอในรูปแบบยูสเคสไดอะแกรม ดังแสดงในภาพที่ 2.7



ภาพที่ 2.7 ยูสเคสไดอะแกรมของทวิตเตอร์

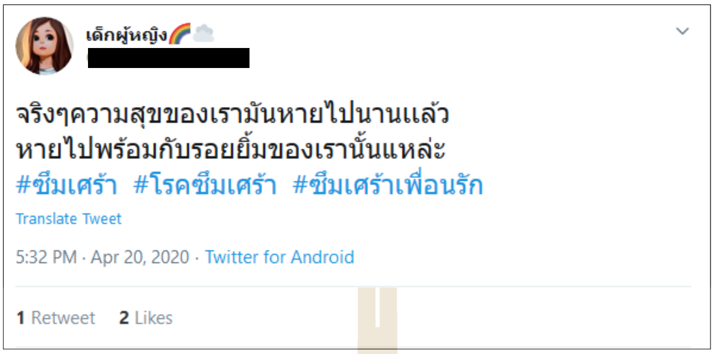
เนื้อหาที่ผู้ใช้งานแสดงความคิดเห็นผ่านทวิตเตอร์ ประกอบด้วย ข้อความ รูปภาพ สัญรูปอารมณ์ และภาพเคลื่อนไหว โดยรูปแบบข้อความความคิดเห็นในทวิตเตอร์นั้นจะเป็นข้อความที่จำกัดตัวอักษร ในหนึ่งข้อความจะมีตัวอักษรได้ไม่เกิน 280 ตัวอักษร และข้อความในทวิตเตอร์จะมีการใช้แฮชแท็กซึ่งสามารถใช้เป็นตัวกรองเบื้องต้น แสดงให้เห็นว่าข้อความนั้นกล่าวถึงเรื่องใด จึงทำให้การวิเคราะห์ข้อความความคิดเห็นของผู้ใช้งานเป็นไปได้อย่างสะดวกมากยิ่งขึ้น (กษิดิ์เดช ทาแป้ง, 2559) ดังแสดงตัวอย่างในภาพที่ 2.8



ภาพที่ 2.8 การแสดงความคิดเห็นผ่านทวิตเตอร์

จากคุณลักษณะดังกล่าวของทวิตเตอร์นั้นทำให้ทวิตเตอร์ถูกนำมาใช้สำหรับงานวิจัยด้านต่าง ๆ ไม่ว่าจะเป็นการวิเคราะห์ข้อมูลเพื่อพยากรณ์ด้านเศรษฐกิจ ด้านการตลาด และด้านการเมือง เช่น การพยากรณ์ทิศทางการเปลี่ยนแปลงของหุ้นโดยดูจากความคิดเห็นของผู้ใช้งานทวิตเตอร์ ผลการวิจัยพบว่ามีความแม่นยำสูงถึงร้อยละ 86.4 (Bollen et al., 2011) การสร้างขั้นตอนวิธีเพื่อทำนายการเกิดการจราจลด้วยการใช้ข้อมูลจากทวิตเตอร์ โดยผลการวิจัยพบว่า ขั้นตอนวิธีดังกล่าวสามารถใช้ทำนายเหตุการณ์ความวุ่นวายต่าง ๆ ได้ก่อนที่ตำรวจจะทราบข้อมูล (Alsaedi, 2017) ดังนั้นจะเห็นได้ว่าทวิตเตอร์มีความสามารถในการรวบรวมข้อมูลจำนวนมากได้เป็นอย่างดี และนอกจากนั้นแล้วด้วยปริมาณตัวอักษรที่จำกัดของทวิตเตอร์ส่งผลให้ขอบเขตเนื้อหาของข้อความนั้นมีความเฉพาะเจาะจงมากกว่าบริการเครือข่ายสังคมออนไลน์รูปแบบอื่น อีกทั้งรูปแบบข้อมูลจากทวิตเตอร์มีความหลากหลาย โดยประกอบไปด้วย 5 ส่วนหลัก คือ ข้อความทวีต (Tweet) จำนวนรีทวีต (Retweet) จำนวนการกดถูกใจ (Like) จำนวนผู้ติดตาม (Follower) และจำนวนการติดตาม (Following) ซึ่งข้อความทวีต (Tweet) สามารถจำแนกออกเป็น ข้อความ (Text) สัญรูปอารมณ์ (Emoticon) แฮชแท็ก (Hashtag) รูปภาพ (Image) และภาพเคลื่อนไหว (Video) ดังรายละเอียดต่อไปนี้

1) ข้อความทวีต (Tweet) คือข้อความที่ผู้ใช้งานทวิตเตอร์เขียนเพื่อแสดงความคิดเห็น ความรู้สึกไปบนทวิตเตอร์เพื่อให้ผู้ใช้อื่น ๆ รับรู้ โดยข้อความทวีตจะมีจำนวนตัวอักษรสูงสุดคือ 280 ตัวอักษร โดยข้อความทวีตจะประกอบด้วย ข้อความ สัญรูปอารมณ์ แฮชแท็ก รูปภาพ และภาพเคลื่อนไหว ดังแสดงในภาพที่ 2.9



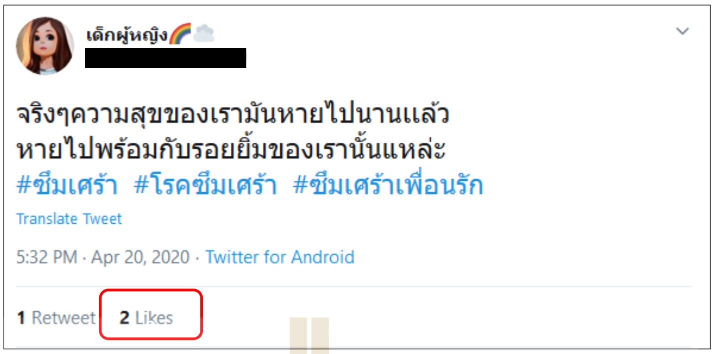
ภาพที่ 2.9 ตัวอย่างข้อความทวีต (Tweet)

2) จำนวนรีทวีต (Retweet) คือจำนวนที่ผู้ใช้รายอื่น ๆ เผยแพร่ข้อความทวีตซ้ำโดยการกดปุ่ม Retweet โดยจำนวนครั้งที่รีทวีตจะปรากฏอยู่ด้านล่างข้อความทวีต ดังแสดงในภาพที่ 2.10

ภาพที่ 2.10 ตัวอย่างจำนวนรีทวีต (Retweet)

3) จำนวนการกดถูกใจ (Like) คือจำนวนครั้งที่ผู้ใช้งานอื่น ๆ กดปุ่มถูกใจข้อความ โดยการกดปุ่มถูกใจข้อความจะทำได้เพียงครั้งเดียวในแต่ละข้อความทวีต ซึ่งจำนวนการกดถูกใจจะแสดงอยู่ด้านล่างของข้อความทวีตเช่นเดียวกับจำนวนรีทวีต ดังแสดงในภาพที่ 2.11

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 2.11 ตัวอย่างจำนวนการกดถูกใจ (Like)

4) จำนวนการติดตาม (Following) คือจำนวนที่ผู้ใช้กดปุ่มติดตาม (Follow) ผู้ใช้รายอื่น โดยการติดตามนั้นจะทำให้มองเห็นข้อความทวีตของผู้ที่ถูกติดตาม แสดงให้เห็นถึงความสนใจของผู้ใช้งานที่สนใจผู้ใช้รายอื่น ซึ่งจำนวนการติดตามจะแสดงที่หน้าโพรไฟล์ของผู้ใช้งาน ดังแสดงในภาพที่ 2.12

ภาพที่ 2.12 ตัวอย่างจำนวนการติดตาม (Following)

5) จำนวนผู้ติดตาม (Follower) คือจำนวนที่ผู้ใช้รายอื่นกดปุ่มติดตาม (Follow) โดยการติดตามนั้นจะทำให้ผู้ที่กดติดตามมองเห็นข้อความทวีตของผู้ใช้ แสดงให้เห็นว่าผู้ใช้รายอื่นมีความสนใจมากน้อยเพียงใด ซึ่งจำนวนผู้ติดตามจะแสดงที่หน้าโพรไฟล์ของผู้ใช้งานเช่นเดียวกับจำนวนการติดตาม ดังแสดงในภาพที่ 2.13

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาพที่ 2.13 ตัวอย่างจำนวนผู้ติดตาม (Follower)

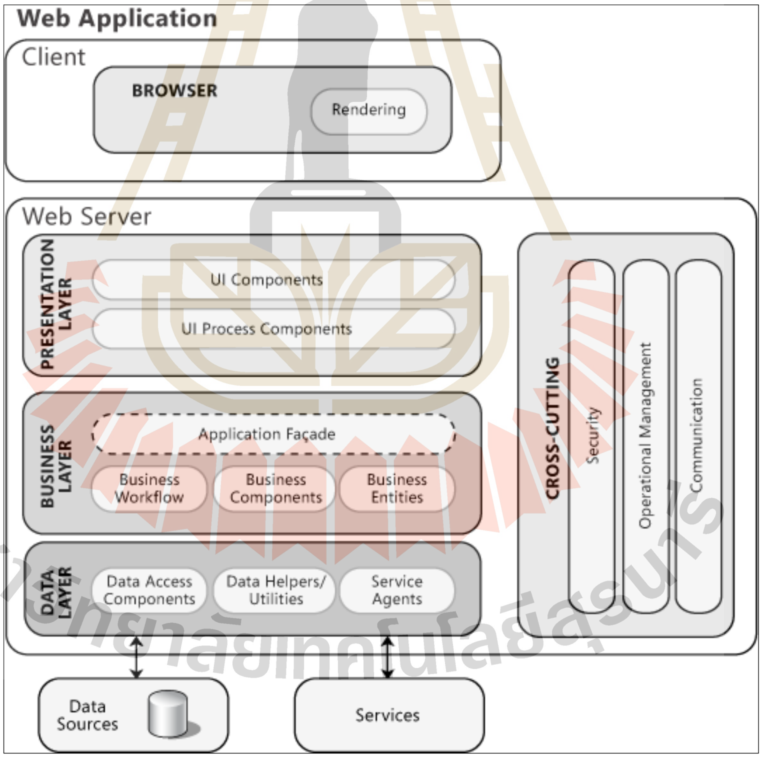
ด้วยคุณลักษณะของทวิตเตอร์ดังที่กล่าวมาแล้วข้างต้น และความสอดคล้องของจำนวนผู้ใช้งานที่มีอัตราการใช้งานที่สูงขึ้น และช่วงอายุของผู้ใช้งานสอดคล้องกับความชุกของภาวะซึมเศร้าในนักศึกษาซึ่งเป็นวัยรุ่นตอนปลายโดยมีตัวเลขสูงถึงร้อยละ 53.50 (โปรยทิพย์ สันตะพันธุ์ และคณะ, 2562) งานวิจัยนี้จึงเลือกใช้การเก็บรวบรวมข้อมูลจากผู้ใช้งานเครือข่ายสังคมออนไลน์คือ ทวิตเตอร์เพื่อใช้ในการวิจัยเรื่องการพัฒนาแบบจำลองการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.2 เทคโนโลยีในการรวบรวมข้อมูล

2.2.1 โปรแกรมประยุกต์บนเว็บ (Web Application)

โปรแกรมประยุกต์บนเว็บ หรือ เว็บแอปพลิเคชัน คือโปรแกรมที่อยู่ในเครื่องบริการเว็บ (Web Server) ที่ทำหน้าที่ให้บริการเมื่อมีการร้องขอใช้บริการจากทางลูกข่าย (Client) ผ่านทางโปรโตคอลสำหรับการสื่อสารผ่านเว็บ เช่น เอชทีทีพี (HTTP) หรือ เอชทีทีพีเอส (HTTPS) ซึ่งการแสดงผลนั้นจะแสดงผลลัพธ์ที่ได้จากการร้องขอในรูปแบบของเอกสารเอชทีเอ็มแอล (HTML) ผ่านทางเว็บเบราว์เซอร์ (Web Browser) (ศุภชัย สมพานิช, 2549) แกนหลักของโปรแกรมประยุกต์บนเว็บคือ ตรรกะการทำงานฝั่งแม่ข่าย (Server –side logic) สถาปัตยกรรมของโปรแกรมประยุกต์บนเว็บประกอบด้วย 3 ส่วนคือ ส่วนผู้ใช้งานหรือลูกข่าย (Client) ส่วนเว็บแม่ข่าย (Web Server) และส่วนแหล่งข้อมูล (Data Source) กับการให้บริการต่าง ๆ (Service) (Meier et al., 2008) ดังแสดงในภาพที่ 2.14



ภาพที่ 2.14 สถาปัตยกรรมโปรแกรมประยุกต์บนเว็บ (Meier et al., 2008)

สถาปัตยกรรมของโปรแกรมประยุกต์บนเว็บมีรายละเอียดในแต่ละส่วน ดังนี้

1) ส่วนผู้ใช้งานหรือลูกข่าย (Client) เป็นส่วนแรกที่ผู้ใช้ใช้ในการเข้าถึงเว็บแม่ข่าย โดยเว็บเบราว์เซอร์ (Web browser) เช่น กูเกิลโครม (Google Chrome) ไฟร์ฟอกซ์ (Firefox) อินเทอร์เน็ตเอ็กโพเลอร์ (Internet Explorer) ฯลฯ ผ่านอินเทอร์เน็ตโดยคำร้องขอจากผู้ใช้จะถูกส่งผ่านโปรโตคอล (Protocol) HTTP ไปยังเว็บแม่ข่าย และเว็บแม่ข่ายจะส่งข้อมูลในรูปแบบ HTML มาแสดงยังเว็บเบราว์เซอร์ของผู้ใช้งาน

2) ส่วนเว็บแม่ข่าย (Web Server) ประกอบด้วยลำดับชั้นที่แตกต่างกัน คือ ชั้นการนำเสนอ (Presentation Layer) ชั้นการบริหารจัดการ (Business Layer) ชั้นข้อมูล (Data Layer) และครอสคัตติง (Cross Cutting)

- ชั้นการนำเสนอ (Presentation Layer) เป็นส่วนที่รับผิดชอบในการนำเสนอผลลัพธ์เพื่อให้ผู้ใช้งานสามารถเข้าใจได้ ซึ่งในส่วนนี้จะประกอบด้วย องค์ประกอบของส่วนต่อประสานกับผู้ใช้ (User Interface Components : UI Components) และกระบวนการของส่วนต่อประสานกับผู้ใช้ (User Interface Process Components : UI Process Components) ในส่วนของชั้นนำเสนอส่วนใหญ่แล้วจะเปิดใช้งานโดยตรงผ่านเว็บเบราว์เซอร์ ดังนั้นเทคโนโลยีที่ใช้สำหรับชั้นนำเสนอส่วนใหญ่จะทำงานผ่านเว็บเบราว์เซอร์ได้โดยตรง เช่น จาวาสคริปต์ (JavaScript) Cascading Style Sheets (CSS) และ HTML

- ชั้นการบริหารจัดการ (Business Layer) ในชั้นนี้จะใช้ตรรกะทางธุรกิจ (Business logic) ในการออกแบบ เพื่อใช้ในการปรับปรุง ดูแล และทดสอบความสามารถของโปรแกรมประยุกต์ โดยเทคโนโลยี และเครื่องมือที่ใช้ในชั้นนี้ต้องสามารถทำงานและประมวลผลที่ฝั่งแม่ข่ายได้เช่น PHP Laravel ซึ่งเป็นเฟรมเวิร์คของ PHP และ Symfony ซึ่งเป็นคอมโพแน้นของ PHP เช่นเดียวกัน

- ชั้นข้อมูล (Data Layer) เป็นส่วนที่ใช้เป็นส่วนที่ใช้สำหรับปรับปรุงแหล่งข้อมูล ถ่ายโอนข้อมูลเมื่อมีปฏิสัมพันธ์กับชั้นอื่น ๆ และเชื่อมโยงส่งข้อมูลระหว่างชั้นการแยกชั้นข้อมูลออกมาจากชั้นอื่น ๆ ช่วยให้สามารถกำหนดค่า และบำรุงรักษาโปรแกรมประยุกต์ได้ง่าย นอกจากนี้ยังช่วยซ่อนรายละเอียดของฐานข้อมูลจากชั้นอื่น ๆ ของโปรแกรมประยุกต์เพื่อความปลอดภัยของข้อมูลได้ ตัวอย่างเทคโนโลยีที่ใช้สำหรับชั้นข้อมูล เช่น Hibernate framework และ RabbitMQ

- ครอสคัตติง (Cross Cutting) เป็นชั้นที่ใช้ในการสื่อสารระหว่างโปรแกรมประยุกต์อื่น ๆ โดยพิจารณาในเรื่องความปลอดภัยของการสื่อสาร การยืนยันตัวตน ซึ่งมีเทคโนโลยีที่สามารถใช้กับครอสคัตติงได้ เช่น OAuth Spring Security และ OpenID

3) ส่วนแหล่งข้อมูล (Data Source) กับการให้บริการต่าง ๆ (Service) เป็นส่วนที่อยู่ของฐานข้อมูล เช่น MySQL Oracle และ Microsoft SQL Server ฯลฯ

2.2.2 ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ (Twitter API)

ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ หรือเรียกอีกอย่างว่า ทวิตเตอร์

เอพีไอ พัฒนาขึ้นเพื่อให้ผู้พัฒนาโปรแกรมประยุกต์อื่น ๆ ใช้ในการติดต่อสื่อสารและเก็บรวบรวมข้อมูลจากทวิตเตอร์ เช่น ข้อมูลการแสดงความคิดเห็น หรือข้อความทวีต สถานะ และข้อมูลส่วนตัวของผู้ใช้งาน และข้อมูลที่ถูกส่งกลับมายังผู้พัฒนาจะอยู่ในขอบเขตจำกัดซึ่งต้องผ่านการยืนยันตัวต้นกับทางทวิตเตอร์ ซึ่งรูปแบบของเอพีไอในทวิตเตอร์จะแบ่งเป็น 4 รูปแบบ (Twitter, 2020) ดังต่อไปนี้

1) เอพีไอมาตรฐาน (Standard API) เป็นเอพีไอที่ผู้เรียกใช้สามารถเข้าถึงข้อมูลต่าง ๆ ในทวิตเตอร์ได้ ประกอบด้วย ข้อความทวีต และไทม์ไลน์ (Timelines) การรับส่งข้อความส่วนตัว (Direct message) การเข้าถึงข้อมูลส่วนบุคคลที่เปิดเป็นสาธารณะ การสร้างและการจัดการลิสต์รายการ การติดตามและการค้นหาผู้ใช้งานทวิตเตอร์

2) เอพีไอพรีเมียม (Premium API) ฟังก์ชันการกรองข้อมูลขั้นสูง และการเข้าถึงข้อมูลผ่านทางการค้นหาข้อมูลด้วยส่วนต่อประสานการค้นหาข้อความทวีต (Search Tweets API) และเข้าถึงผู้ใช้สาธารณะแบบเรียลไทม์ การมีส่วนร่วม และข้อมูลการทวีตผ่านทางส่วนต่อประสานบัญชีผู้ใช้งาน (Account Activity API)

3) เอพีไอวิสาหกิจ (Enterprise API) เป็นการเข้าถึงข้อมูลในทวิตเตอร์ระดับวิสาหกิจ ประกอบด้วย ข้อความทวีตแบบเรียลไทม์ และข้อมูลผู้ใช้งานแบบสาธารณะ ประวัติการแสดงข้อความทวีต และข้อมูลเชิงลึกของข้อความทวีต

4) เอพีไอโฆษณา (Ads API) เป็นการผสานรวมทางโปรแกรมกับแพลตฟอร์มโฆษณาในทวิตเตอร์

ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ เป็นส่วนที่ทำให้ผู้พัฒนาสามารถเข้าถึงข้อมูลต่าง ๆ ในฐานข้อมูลหลักของทวิตเตอร์ได้ ซึ่งการนำข้อมูลในแต่ละส่วนออกมานั้นจะมีวิธีการที่แตกต่างกันออกไป ประกอบด้วย สตรีมมิงเอพีไอ (Streaming API) และ เรสเอพีไอ (REST API)

- สตรีมมิงเอพีไอ ใช้ดึงข้อมูลทวีตของผู้ใช้ทวิตเตอร์โดยการใช้คำสั่งในรูปแบบสคริปต์ (Script) เพื่อดึงข้อมูลออกมาตามรูปแบบที่กำหนดไว้ เช่น เอ็กซ์เอ็มเอล (XML) ซีเอสวี (CSV) และเจสัน (JSON)

มหาวิทยาลัยเทคโนโลยีสุรนารี

- เรสเอพีไอ การดึงข้อมูลจากทวิตเตอร์ด้วยเรสเอพีไอนั้นจำเป็นต้องมีการยืนยันตัวตนก่อนจะเข้าถึงฐานข้อมูล โดยการลงทะเบียนการใช้งานที่เว็บไซต์ของทวิตเตอร์ เพื่อเป็นการกำหนดข้อตกลงร่วมกันระหว่างผู้ใช้และทวิตเตอร์ว่าสามารถเข้าถึงข้อมูลได้ในระดับใด โดยผู้ใช้งานสามารถดึงข้อมูลผ่านเรสเอพีไอ ร่วมกับภาษาทางคอมพิวเตอร์อื่น ๆ เช่น ไพทอน (python) จาวา (Java) พีเอชพี (PHP) ฯลฯ

สำหรับงานวิจัยนี้ได้ใช้ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ ในรูปแบบเอพีไอมาตรฐานด้วยวิธีการเรสเอพีไอในการรวบรวมข้อมูลจากทวิตเตอร์เพื่อใช้ในการวิเคราะห์ข้อมูล เนื่องจากเอพีไอมาตรฐานสามารถเข้าถึงข้อมูลต่าง ๆ ในทวิตเตอร์ได้อย่างครอบคลุม

2.2.3 ส่วนต่อประสานโปรแกรมประยุกต์ในไพทอน (Python API)

ไพทอน (Python) เป็นภาษาคอมพิวเตอร์ในรูปแบบภาษาเชิงวัตถุ (Dynamic Object-Oriented Programming) ที่พัฒนาขึ้นในปี ค.ศ. 1990 โดย Guido von Rossum ภาษาไพทอนมีลักษณะการทำงานในรูปแบบข้ามแพลตฟอร์ม (Cross Platform) คือสามารถนำโปรแกรมที่พัฒนาด้วยภาษาไพทอนไปใช้ในระบบปฏิบัติการอื่น ๆ ได้ เช่น ระบบปฏิบัติการวินโดว์ ระบบปฏิบัติการลีนุกซ์ ระบบปฏิบัติการโอเอสเอ็กซ์ นอกจากนั้นแล้วภาษาไพทอนยังมีอิสระในการแก้ไขไลบราลี (Library) ต่าง ๆ ได้

ส่วนต่อประสานโปรแกรมประยุกต์ในไพทอน หรือไพทอนเอพีไอ (Python API) ช่วยให้นักพัฒนาสามารถรวมโปรแกรมประยุกต์หนึ่งเข้ากับโปรแกรมประยุกต์อื่น โดยการเผยแพร่การทำงานบางส่วน และจำกัดการทำงานในบางส่วน โดยสามารถใช้ส่วนต่อประสานโปรแกรมประยุกต์เพื่อรับข้อมูลจากโปรแกรมอื่น ๆ (King, 2019) เช่น ทวีตพายเอพีไอ (Tweepy API) ซึ่งใช้สำหรับการดึงข้อมูลจากโปรแกรมทวิตเตอร์ โอเพนโนติฟายเอพีไอ (OpenNotify API) ซึ่งใช้สำหรับการสร้างการแจ้งเตือนผ่านระบบ

การทำงานของส่วนต่อประสานโปรแกรมประยุกต์ในไพทอนส่วนใหญ่มีการทำงานในลักษณะเดียวกันคือร้องขอ (Request) ผ่านโปรโตคอลเอชทีทีพี (http) โดยสามารถร้องขอผ่านการเก็ต/โพสต์ (GET/POST) และส่งข้อมูลกลับในรูปแบบของเจสัน (JSON) ซึ่งในงานวิจัยนี้ได้เลือกใช้ทวีตพาย (Tweepy) ในการดึงข้อมูลจากผู้ใช้งานทวิตเตอร์ โดยมีขั้นตอนการทำงานดังต่อไปนี้

1) การนำเข้าไลบลารี (Import library) โดยใช้คำสั่ง

```
Import tweepy
```

2) การตรวจสอบสิทธิ์การใช้งานเพื่อเชื่อมต่อกับทวิตเตอร์เอพีไอ

```
auth = tweepy.OAuthHandler(consumer_key,
consumer_secret)
auth.set_access_token(access_token, access_token_secret)
api = tweepy.API(auth)
```

3) การดึงข้อมูลจากทวิตเตอร์มาเก็บไว้ในตัวแปรของภาษาไพทอน

```
public_tweets = api.home_timeline()
for tweet in public_tweets:
    print(tweet.text)
```

มหาวิทยาลัยเทคโนโลยีสุรนารี

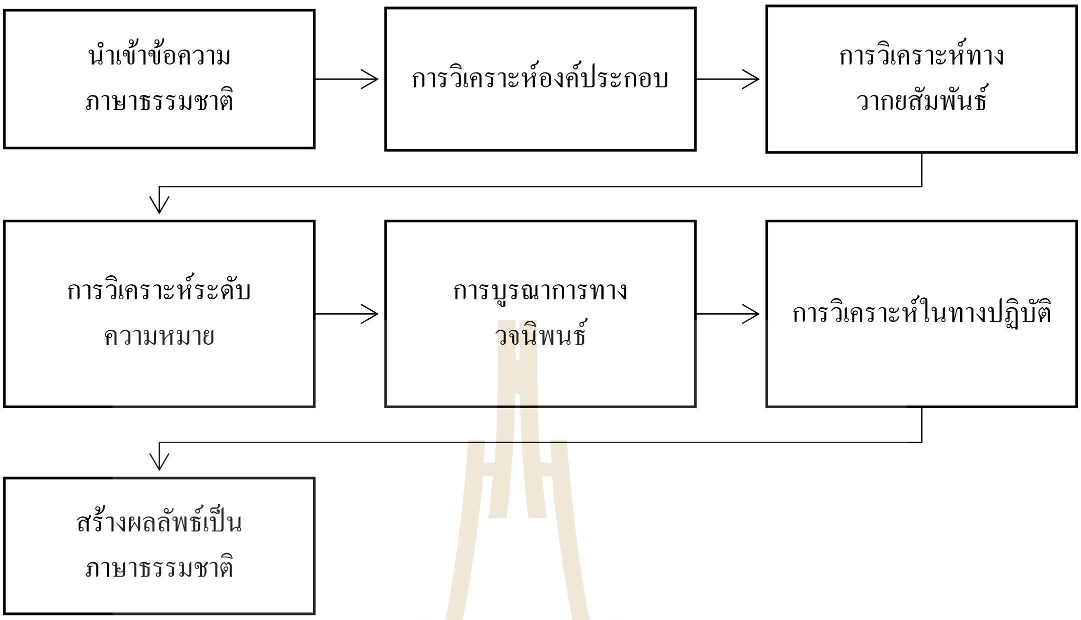
2.3 เทคโนโลยีในการประมวลผลข้อมูล

2.3.1 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติ คือการแปลความจากภาษาธรรมชาติที่มนุษย์ใช้สื่อสารกันให้อยู่ในรูปแบบที่เป็นโครงสร้างที่เครื่องคอมพิวเตอร์สามารถเข้าใจได้ 2 แนวทาง (ยืน ภู่วรรณ, 2535) คือแนวทางการศึกษาและทำความเข้าใจกับโครงสร้างทางภาษาศาสตร์และแนวทางคือการอาศัยความรู้ด้านปัญญาประดิษฐ์ และการใช้คลังคำศัพท์ โดยทั่วไปแล้วการประมวลผลภาษาธรรมชาติจะใช้เพื่ออธิบายฟังก์ชันของซอฟต์แวร์หรือส่วนประกอบฮาร์ดแวร์ในระบบคอมพิวเตอร์ที่วิเคราะห์หรือสังเคราะห์ภาษาพูดหรือภาษาเขียน คำคุณศัพท์ตามธรรมชาติมีไว้เพื่อแยกความแตกต่างของคำพูดและการเขียนของมนุษย์จากภาษาที่เป็นทางการมากขึ้นเช่นสัญลักษณ์ทางคณิตศาสตร์หรือภาษาโปรแกรมที่คำศัพท์และไวยากรณ์ถูกจำกัดคำพูดอย่างเคร่งครัด (Jackson and Moulinier, 2007)

การประมวลผลภาษาธรรมชาติ เป็นการนำความรู้ทางด้านภาษาศาสตร์มาวิเคราะห์รูปแบบโครงสร้างของประโยคตามหลักไวยากรณ์ และแปลความหมายของคำเพื่อทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์ แล้วนำข้อความนั้นเก็บไว้ในฐานความรู้เพื่อให้คอมพิวเตอร์เกิดการเรียนรู้และสามารถนำไปสร้างเป็นแบบจำลอง (Model) เพื่อนำไปใช้ประโยคตามความต้องการ (นงเยาว์ สอนจะโป๊ะ, 2561) นอกจากนั้นแล้ว การประมวลผลภาษาธรรมชาติยัง เป็นกระบวนการที่ทำให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติของมนุษย์ที่ใช้สื่อสารกัน ให้อยู่ในรูปแบบโครงสร้างที่คอมพิวเตอร์สามารถเข้าใจ และนำไปประมวลผลได้ดี โดยการประมวลผลภาษาธรรมชาติได้แบ่งระดับขั้นการวิเคราะห์ ดังแสดงในภาพที่ 2.15 ซึ่งมีรายละเอียดดังนี้ (กรมวุฒิ นงนุช, อนุชา ซาเฮาะ และสุวุฒิ ตุ้มทอง, 2559; มาสวีร์ มาศดิศรโชติ, 2557; Larsson et al., 2017)

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 2.15 ขั้นตอนการประมวลผลภาษาธรรมชาติ

1) การวิเคราะห์ทางองค์ประกอบ หรือการวิเคราะห์ระดับวจีวิภาค (Morphological analysis) เป็นการวิเคราะห์ในระดับของคำ โดยวิเคราะห์ว่าในแต่ละคำสามารถแยกย่อยได้อีกหรือไม่ เช่น “งานประจำ” แยกได้ “งาน” และ “ประจำ”

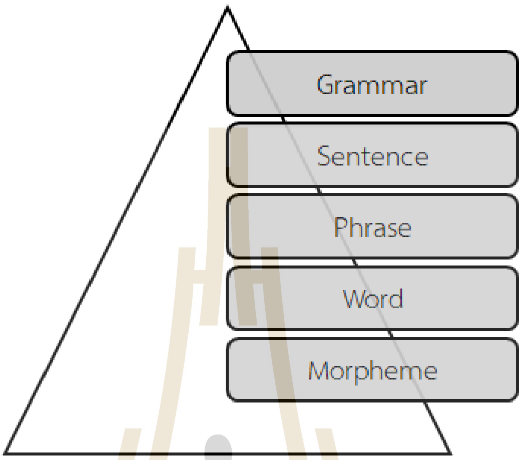
2) การวิเคราะห์ทางวากยสัมพันธ์ (Syntactic analysis) เป็นการวิเคราะห์ทางไวยกรณ์เพื่อดูโครงสร้างของประโยค และความเกี่ยวข้องของส่วนต่าง ๆ ในประโยคที่รับเข้ามาว่า คำไหนเป็นประธาน กริยา และกรรม หรือส่วนใดเป็นวลี เพื่อใช้แสดงควาสัมพันธ์ของคำต่าง ๆ

3) การวิเคราะห์ระดับความหมาย (Semantic analysis) เป็นการวิเคราะห์ความหมายของคำ หลังจากผ่านการวิเคราะห์ทางวากยสัมพันธ์มาแล้ว จึงมากำหนดค่าของแต่ละคำว่าหมายถึงสิ่งใด ซึ่งบางคำเขียนถูกต้องตามหลักไวยกรณ์ แต่บางครั้งความหมายที่ได้เป็นความหมายที่กำกวม หรือไม่มีความหมาย หรือเป็นเรื่องที่เป็นไปไม่ได้ เช่น “ฉันกินบ้าน” ซึ่งถูกต้องตามหลักไวยกรณ์เพราะประกอบด้วยโครงสร้างประโยค ประธาน กริยา กรรม

4) การบูรณาการทางวจนิพนธ์ (Discourse integration) เป็นการวิเคราะห์ความหมายของประโยค โดยพิจารณาจากประโยคข้างเคียง เนื่องจากคำบางคำในประโยคจะเข้าใจความหมายได้ต้องดูประโยคก่อนหรือหรือประโยคหลังประกอบด้วย

5) การวิเคราะห์ในทางปฏิบัติ (Pragmatic analysis) คือการแปลความหมายหรือการตีความของประโยคใหม่อีกครั้งว่าผู้พูดตั้งใจจะสื่อความหมายอะไร หลังจากผ่านกระบวนการทั้ง 4 ขั้นตอนที่กล่าวมา

กระบวนการวิเคราะห์ภาษาจะเริ่มต้นที่ระดับต่ำสุดก่อนคือการวิเคราะห์ทางองค์ประกอบ หรือการวิเคราะห์ระดับคำ ไปจนถึงระดับวากยสัมพันธ์ ที่สามารถอธิบายได้ด้วยโครงสร้างของภาษาที่ประกอบด้วยหน่วยต่าง ๆ ดังแสดงในภาพที่ 2.16



ภาพที่ 2.16 โครงสร้างของภาษา

โดยโครงสร้างของภาษาที่ประกอบด้วยหน่วยต่าง ๆ มีรายละเอียดดังนี้

1. หน่วยคำ (Morpheme) คือ หน่วยที่เล็กที่สุดที่มีความหมาย จะอยู่เป็นอิสระหรือไม่ก็ได้ หน่วยคำเป็นองค์ประกอบของคำ คำบางคำประกอบด้วยหน่วยคำ 1 หน่วย เช่น “เรียน” บางคำประกอบด้วยหน่วยคำมากกว่า 1 หน่วย เช่น “นักเรียน” โดย ประกอบด้วย “นัก” และ “เรียน” ซึ่งถือเป็นหน่วยคำ และ “นักเรียน” จึงถือเป็นคำ

2. คำ (Word) คือ หน่วยอิสระที่เล็กที่สุดที่มีความหมาย โดยมีคุณสมบัติ 3 ประการ คือ อยู่เป็นอิสระได้ เล็กที่สุด และแบ่งแยกให้เล็กลงเป็นหน่วยอิสระไม่ได้ มีความหมาย ดังนั้น ความแตกต่างระหว่าง หน่วยคำ กับ คำคือ หน่วยคำอาจเกิดขึ้นอิสระหรือไม่อิสระก็ได้ แต่คำต้องเป็นหน่วยอิสระเสมอ ในกรณีที่หน่วยใด ๆ เป็นคำอิสระ หน่วยคำนั้นก็สามารถเป็นคำได้ด้วย ในหลาย ๆ ภาษาในการเขียนจะมีการเว้นวรรคระหว่างคำทำให้สังเกตคำได้ง่าย แต่สำหรับภาษาไทยซึ่งไม่มีช่องว่างระหว่างคำ เช่น คำ “ลูกเสือ” เป็นคำ (Word) เพราะถ้าแยกเป็น ลูก และ เสือ อาจไม่ได้ความหมายเดิมว่า “ลูกเสือ” จึงถือว่า “ลูกเสือ” เป็นหน่วยที่เล็กที่สุดที่แบ่งแยกลงไปไม่ได้อีกแล้ว และถือว่าเป็น 1 คำ

3. วลี (Phrase) คือกลุ่มของคำหรือคำ ๆ เดียวก็ได้ ที่เป็นส่วนประกอบของประโยค (วลีเป็นส่วนประกอบของประโยค แต่คำไม่ใช่ส่วนประกอบของประโยค) ดังนั้น ไม่ว่าจะเป็นคำ ๆ เดียวหรือกลุ่มที่ประกอบด้วยหลาย ๆ คำ เมื่อเป็นส่วนประกอบของประโยคเราเรียกเป็น วลี

4. นามวลี (Noun phrase) คือวลีที่ทำหน้าที่เป็นส่วนของประโยคที่เป็นหน่วยประธาน หน่วยกรรม นามวลี

5. กริยาวลี (Verb phrase) คือวลีที่ทำหน้าที่เป็น ภาคแสดงของประโยค

6. ประโยค (Sentence) คือคำหลายคำเรียงกันในการพูดหรือเขียน เพื่อเป็นการแสดงความคิด 1 ความคิดอย่างสมบูรณ์ทางไวยากรณ์ โดยปกติมักประกอบไปด้วยประธานและภาคแสดง

7. ไวยากรณ์ (Grammar) คือ ภาษาที่มีกฎเกณฑ์ และเป็นไปตามระเบียบวิธีการประกอบรูปคำให้เป็นประโยค

การวิเคราะห์ในระดับที่สูงขึ้นกว่าระดับวากยสัมพันธ์หรือโครงสร้างของภาษาจำเป็นต้องอาศัยฐานความรู้ที่สูงขึ้น เช่น ฐานความรู้เครือข่ายคำ (WordNet) สำหรับกระบวนการวิเคราะห์ภาษาธรรมชาติในภาษาไทยส่วนใหญ่ เป็นการวิเคราะห์ภาษาในระดับวจีวภาคและระดับวากยสัมพันธ์เท่านั้น เนื่องจากเครือข่ายคำไทย (Thai WordNet) ค่อนข้างมีจำนวนจำกัดและยังอยู่ในขั้นพัฒนา (Alisa et al., 2010)

การวิเคราะห์ภาษาในระดับวจีวภาคและวากยสัมพันธ์สำหรับภาษาไทย มีความยุ่งยากกว่าภาษาอังกฤษเล็กน้อย เนื่องจากภาษาไทยเป็นภาษาที่ไม่มีการแบ่งวรรคตอนระหว่างประโยคหรือระหว่างคำที่แน่นอน มีการเขียนในรูปแบบเขียนติดกันไปทั้งย่อหน้า ดังนั้นการวิเคราะห์ในระดับต่ำสุดหรือคำ จึงจำเป็นต้องแบ่งเอกสารตั้งแต่หน่วยใหญ่สุดคือประโยค ให้มีหน่วยเล็กสุดที่มีความหมายก่อน ซึ่งหมายถึงคำ แล้วจึงนำคำแต่ละคำมาประกอบเป็นโครงสร้างประโยคดังนั้นการวิเคราะห์ภาษาสำหรับภาษาไทยต้องอาศัย 3 กระบวนการพื้นฐานคือ 1) กระบวนการตัดคำ (Tokenization) เพื่อให้มีความสามารถในการวิเคราะห์ระดับวจีวภาคได้ 2) กระบวนการกำกับคำตามหน้าที่ของคำ (Part-of-Speech tagging) และ 3) กระบวนการวิเคราะห์โครงสร้างภาษา (Syntactic analysis) เพื่อให้สามารถนำคำแต่ละคำมาประกอบเป็นโครงสร้างประโยคได้ โดยมีรายละเอียดในแต่ละกระบวนการดังนี้

1) กระบวนการตัดคำ

ภาษาไทยเป็นภาษาที่เขียนโดยไม่มีการเว้นวรรคเหมือนภาษาอังกฤษ ดังนั้นการตัดคำในภาษาไทยจึงมีความซับซ้อน และการตัดคำที่ถูกต้องมีความสำคัญต่อการนำข้อมูลคำศัพท์ไปประมวลผลต่อ ในงานวิจัยด้านการตัดคำได้แบ่งการตัดคำออกเป็น 3 เทคนิคหลัก ได้แก่

- การตัดคำโดยใช้กฎ (Rule based approach) เป็นวิธีการพิจารณาการตัดคำจากพยัญชนะ สระ วรรณยุกต์ตัดสะกด การันต์ วิธีนี้เป็นวิธีที่ง่ายที่สุด และทำงานได้เร็วที่สุด แต่ไม่สามารถแก้ปัญหาความกำกวมของกฎได้ เช่น พยัญชนะบางตัวสามารถเป็นได้ทั้งพยัญชนะต้นและตัวสะกด

- การตัดคำโดยใช้ฐานความรู้จากพจนานุกรม (Dictionary approach) ดังตัวอย่าง วิธีการเลือกตัดคำจากคำที่พบในพจนานุกรมและมีความยาวมากที่สุด (Longest matching) หรือเลือกตัดคำจากความเหมือนมากที่สุด (Maximal matching) วิธีการตัดคำด้วยพจนานุกรมนี้มีความถูกต้องมากกว่าการตัดคำโดยใช้กฎ แต่ยังไม่สามารถแก้ปัญหาคำกำกวมทั้งหมดได้
- การตัดคำโดยใช้ค่าสถิติจากคลังเอกสาร (Corpus based approach) วิธีการนี้นำเอาค่าสถิติการเกิดคำและหน้าที่ของคำเข้ามาช่วยในการคำนวณหาความน่าจะเป็น เพื่อเลือกแบบที่มีโอกาสการเกิดมากที่สุด วิธีการนี้มีความถูกต้องมากกว่า 2 วิธีการแรก แต่มีข้อจำกัดคือต้องมีฐานข้อมูลที่มีการตัดคำไว้อย่างถูกต้อง ฐานข้อมูลต้องมีขนาดใหญ่มากพอ เพื่อให้ได้สถิติข้อมูลที่มีความน่าเชื่อถือ

สำหรับงานวิจัยนี้ได้เลือกใช้การตัดคำจากส่วนต่อประสานโปรแกรมประยุกต์ในไพทอน (Python API) คือ PyThaiNLP โดยใช้ฟังก์ชัน word_tokenize ซึ่งถูกออกแบบมาให้รองรับการตัดคำหลากหลายขั้นตอนวิธี เช่น newmm longest multi_cut pyicu deepcut attacut tcc และ etcc ซึ่งในงานวิจัยนี้ได้เลือกใช้ขั้นตอนวิธี newmm ซึ่งเป็นขั้นตอนวิธีใหม่ที่สุด และเป็นค่ามาตรฐานของ PyThaiNLP โดยขั้นตอนวิธี newmm เป็นขั้นตอนวิธีซึ่งอาศัยเทคนิคการตัดคำแบบใช้ฐานความรู้จากพจนานุกรม แบบเลือกตัดคำจากความเหมือนมากที่สุด (Maximum matching) ซึ่งมีความเร็วในการประมวลผลสูงและมีความถูกต้องมาก แต่ปัญหาหลักของการประมวลผลภาษาธรรมชาติคือ ความกำกวมของภาษา เนื่องจากในภาษาธรรมชาติมีกฎเกณฑ์และข้อยกเว้นมากมาย คำแต่ละคำอาจตีความได้แตกต่างกัน ถ้าอยู่ในบริบทที่แตกต่างกัน ดังนั้นผลลัพธ์ที่ได้จากกระบวนการตัดคำจึงต้องนำมาตรวจสอบและแก้ไขให้มีความถูกต้องการเข้าสู่กระบวนการวิเคราะห์ข้อมูล (กานดา แผ่วัฒนากุล, 2555)

2) กระบวนการกำกับคำตามหน้าที่ของคำ

การกำกับคำตามหน้าที่ของคำและการวิเคราะห์ในระดับวากยสัมพันธ์จะช่วยให้เข้าใจประโยคได้ถูกต้องยิ่งขึ้น ซึ่งการกำกับคำตามหน้าที่ของคำเป็นวิธีการบ่งบอกว่าคำแต่ละคำในประโยคเป็นคำชนิดใด จึงทำให้การวิเคราะห์ไวยากรณ์โครงสร้างประโยคและการแปลความถูกต้องยิ่งขึ้น

หน้าที่หลักของคำ แบ่งได้ดังนี้ คำนาม คำสรรพนาม คำกริยา คำกริยาช่วย คำวิเศษณ์ คำบุพบท คำสันธาน คำนามชี้เฉพาะ คำนามบอกลักษณะ คำปฏิเสธ และคำหยุด สัญลักษณ์ที่ใช้ในการกำกับคำจะใช้คำย่อเป็นภาษาอังกฤษเพื่อเป็นสัญลักษณ์บอกชนิดของคำ ดังตารางที่ 2.1

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 2.1 สัญลักษณ์ที่ใช้ในการกำกับหน้าที่ของคำและความหมายของสัญลักษณ์

สัญลักษณ์	คำอธิบายหน้าที่ของคำ
N	คำนามใช้เรียก คน สัตว์ สิ่งของ
PRON	คำสรรพนามที่ใช้แทนคำนาม
V	คำกริยา แสดงอาการหรือการกระทำของนามหรือสรรพนาม
AUX	คำกริยาช่วย คือคำที่เติมหน้าคำกริยาหลักในประโยคเพื่อช่วยขยายความหมายของคำกริยาให้ได้ใจความชัดเจนยิ่งขึ้น
ADJ	คำคุณศัพท์ใช้ขยายได้คำนามและคำสรรพนาม
ADV	คำกริยาวิเศษณ์ ใช้ขยายคำกริยาและคำวิเศษณ์เอง
PREP	คำบุพบทที่ทำหน้าที่เชื่อมคำหรือกลุ่มคำ
CONJ	คำที่ใช้เชื่อมประโยคกับประโยค
DET	คำนามชี้เฉพาะ
CLAS	คำนามบอกลักษณะ ขนาด หรือปริมาณ
NEG	คำปฏิเสธ
END	คำหยุด

3) การวิเคราะห์กฎไวยากรณ์

การประมวลผลภาษาธรรมชาติโดยอาศัยโครงสร้างทางภาษาศาสตร์ที่ประกอบด้วยหน่วยที่เล็กที่สุดคือหน่วยคำ (Morpheme) ซึ่งอาจมีหรือไม่มีความหมายก็ได้ หน่วยคำหลายหน่วยคำประกอบกันเป็นคำ (Word) ที่มีความหมาย คำหลายคำประกอบกันเป็นวลี (phrase) ซึ่งแบ่งเป็นนามวลี (Noun Phrase: NP) และกริยาวลี (Verb Phrase: VP) วลีหลายวลีประกอบกันเป็นประโยค (Sentence: S) ซึ่งอธิบายได้ด้วยไวยากรณ์ (Grammar) ตามหลักภาษาศาสตร์หรือเรียกว่าสูตรไวยากรณ์ (Grammar Formalism) สูตรไวยากรณ์ที่นิยมใช้มากที่สุดได้แก่ไวยากรณ์แบบไม่พึ่งบริบท (Context Free Grammar: CFG) กล่าวคือไม่พิจารณาถึงความหมายของคำ แต่จะอาศัยหน้าที่ของคำ (Part-of-Speech: POS) เป็นหลักคือ คำนาม (N) คำสรรพนาม (PRON) คำกริยา (V) คำกริยาช่วย (AUX) คำคุณศัพท์ (ADJ) คำกริยาวิเศษณ์ (ADV) และคำบุพบท (PREP) แล้วจึงอธิบายตามกลุ่มคำนาม (นามวลี) และกลุ่มคำกริยา (กริยาวลี) ในการแบ่งประโยคออกเป็นส่วน ดังนี้ (สมนึก สินธุปวน, 2546)

S = NP + VP

NP = N | N + (ADJ) + (ADV) + (PP) | PRON

VP = V | V + (ADV) | AUX + V | VP + NP

PP = PREP + NP | PREP + VP | PREP + NP + VP

เทคนิคการอธิบายโครงสร้างภาษาหรือการแจงประโยค มี 2 เทคนิค ได้แก่

1. การแจกแจงจากบนลงล่าง (Top-down parsing) เริ่มต้นด้วยสัญลักษณ์ S แล้วเขียนใหม่ด้วยสัญลักษณ์ทางซ้ายมือ แจงประโยคจนกว่าจะพบสัญลักษณ์สิ้นสุด โดยแสดงตัวอย่าง ดังนี้

S

NP + VP

NP + V + NP

NP + V + Art + N

Art + N + V + Art + N

Art + N + V + Art + worms.

Art + N + V + the worms.

Art + N + eat the worms.

Art + birds eat the worms.

The birds eat the worms.

2. การแจกแจงจากล่างขึ้นบน (Bottom-up parsing) เริ่มจากคำศัพท์หรือสัญลักษณ์สิ้นสุด แทนคำด้วยหน้าที่ของคำ จากนั้นใช้สัญลักษณ์ที่อยู่ทางซ้ายของกฎแทนด้วยกลุ่มของสัญลักษณ์ไม่จบท้ายทำไปจนกว่าจะพบสัญลักษณ์ S

The birds eat the worms.

Art + birds eat the worms.

Art + N + eat the worms.

Art + N + V + the worms.

Art + N + V + Art + worms.

Art + N + V + Art + N

NP + V + Art + N

NP + V + NP

NP + VP

S

การประมวลผลภาษาธรรมชาติโดยอาศัยโครงสร้างทางภาษาศาสตร์สำหรับภาษาไทยบางส่วนไม่สามารถวิเคราะห์ประโยคได้ทั้งหมด เนื่องจากภาษาไทยมีรูปแบบไม่ตายตัว และมีคำใหม่เกิดขึ้นมากมาย ซึ่งคำบางคำเป็นลักษณะคำจำเพาะเจาะจงในหัวข้อใดหัวข้อหนึ่ง จึงต้องอาศัยวิธีการแทนความรู้ด้วยคลังคำ คือการรวบรวมคำจากเอกสาร หรือแหล่งข้อมูลที่

มหาวิทยาลัยเทคโนโลยีสุรนารี

เกี่ยวข้องโดยให้ผู้เชี่ยวชาญพิจารณา และนำมาเก็บไว้เป็นคลังคำ (Corpus) ซึ่งปัจจุบันได้มีหลายหน่วยงานรวบรวมคลังคำเพื่อเป็นทรัพยากรที่สำคัญในการศึกษาวิจัยภาษาต่าง ๆ เช่น British National Corpus (100 ล้านคำ แล้วเสร็จในปี ค.ศ. 1994) The Bank of English (ปัจจุบันมีมากกว่า 450 ล้านคำ) American National Corpus (เป้าหมาย 100 ล้านคำ ปัจจุบันเก็บได้ 22 ล้านคำ) Czech National Corpus และ Hellenic National Corpus (คลังข้อมูลภาษากรีก ปัจจุบันมีมากกว่า 34 ล้านคำ) National Corpus of Irish (30 ล้านคำ แล้วเสร็จในปี ค.ศ. 1999) Hungarian National Corpus (มากกว่า 150 ล้านคำ) Slovak National Corpus (30 ล้านคำในปัจจุบัน) Croatian National Corpus ฯลฯ คลังข้อมูลภาษาเหล่านี้ถูกนำมาใช้ในงานต่างๆ อาทิ การจัดทำพจนานุกรม การจัดทำหนังสืออ้างอิงต่าง ๆ และในการศึกษาวิจัยทางภาษาศาสตร์อีกมากมาย

สำหรับภาษาไทยได้มีหน่วยงานและสถาบันการศึกษาหลายแห่งได้จัดทำคลังคำเพื่อนำมาใช้ในด้านการศึกษา และในเชิงพาณิชย์ เช่น คลังข้อมูลออร์คิด (ORCHID) ของศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (NECTEC) แต่ก็เป็นข้อมูลขนาดเล็ก (4 แสนคำ) และเป็นงานเขียนวิชาการประเภทเดียว NAiST Corpusที่รวบรวมคำศัพท์โดยมหาวิทยาลัยเกษตรศาสตร์ และ Thai National Corpus ที่รวบรวมคำศัพท์โดยจุฬาลงกรณ์มหาวิทยาลัย

2.3.2 การวิเคราะห์ความรู้สึก (Sentiment Analysis)

การวิเคราะห์ความรู้สึก (Sentiment Analysis) เรียกอีกอย่างว่า “ Opinion mining” หรือ“ Emotion Artificial Intelligence” และการใช้ประโยชน์จากการประมวลผลภาษาธรรมชาติ (NLP) การทำเหมืองข้อความ ภาษาศาสตร์เชิงคำนวณและการวัดทางชีวภาพในการรับรู้อย่างเป็นระบบ เพื่อประเมินผล ตรวจสอบสถานะทางอารมณ์และข้อมูลส่วนตัว การวิเคราะห์ความรู้สึกโดยทั่วไปเกี่ยวข้องกับการแสดงความคิดเห็นของผู้ใช้งาน ตัวอย่าง เช่น การสำรวจและความคิดเห็นบนเว็บและเครือข่ายสังคมออนไลน์ (Alsaeedi and Khan, 2019)

การวิเคราะห์ความรู้สึก เป็นศาสตร์หนึ่งในการวิเคราะห์ข้อมูลขนาดใหญ่ (Big data) เพื่อนำไปใช้ประโยชน์ในด้านต่าง ๆ การวิเคราะห์ความรู้สึกอาศัยเทคนิคของการประมวลภาษาธรรมชาติ (Natural Language Processing) ในการระบุขั้วความรู้สึก (Sentiment polarity) ของข้อความ โดยจำแนกออกเป็น 3 ขั้วคือ ขั้วบวก (Positive) ขั้วลบ (Negative) และ เป็นกลาง (Neutral) (อิสรภาพ ล้อรัตนไชยยงค์, 2560) โดยการวิเคราะห์ความรู้สึกสามารถทำได้ 3 ระดับ ประกอบด้วย

1) การจำแนกความรู้สึกระดับเอกสาร ในระดับนี้ เอกสารสามารถจำแนกออกเป็น 3 ระดับ คือ ขั้วบวก (Positive) ขั้วลบ (Negative) และ เป็นกลาง (Neutral)

2) การจำแนกความรู้สึกระดับประโยค ในระดับนี้ แต่ละประโยคจะถูกจัดประเภทเป็น ขั้วบวก (Positive) ขั้วลบ (Negative) และ เป็นกลาง (Neutral)

3) การจำแนกความรู้สึกระดับมุมมอง ในระดับนี้ประโยค หรือเอกสารสามารถจัดประเภทเป็น "บวก" (Positive) "ลบ" หรือ "ไม่เข้าข้าง" (Negative or non-partisan) ในแง่มุมของประโยค หรือเอกสารสำคัญ ในระดับนี้เรียกอีกอย่างว่า การจัดกลุ่มการประเมินระดับเปอร์สเปคทีฟ (perspective-level assessment grouping)

การวิเคราะห์ความรู้สึกของข้อความสามารถทำได้ 2 รูปแบบคือ

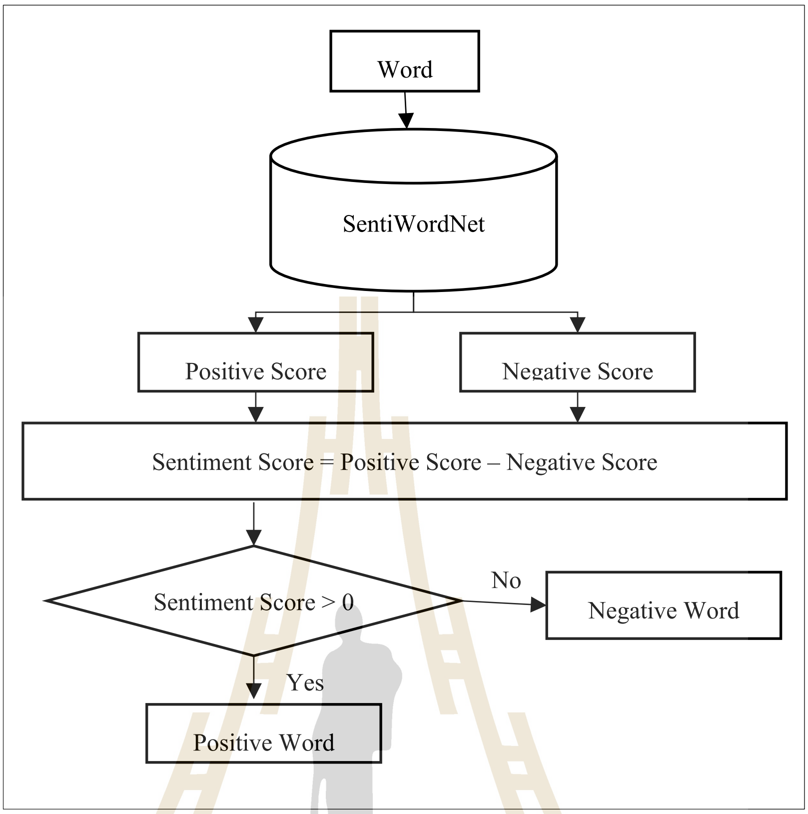
1) การวิเคราะห์แบบอิงคลังศัพท์ (lexicon-based) การวิเคราะห์ความรู้สึกแบบอิงคลังศัพท์อาศัยการระบุคำที่บ่งบอกความรู้สึกในประโยคโดยอ้างอิงจากคลังศัพท์บอกความรู้สึก ซึ่งรวบรวมคำบอกความรู้สึกขั้วบวกและขั้วลบทั้งหมดเอาไว้ เช่น การใช้คลัง SentiWordNet ซึ่งมีคลังคำศัพท์มากที่สุด โดยมีมากถึง 117,695 คำ (Musto, 2014) การใช้งาน SentiWordNet เพื่อวิเคราะห์ความรู้สึก (Sentiment Analysis) ในการหาค่าระดับคะแนนความรู้สึก (Sentiment Score) โดยหาค่าคะแนนของคำศัพท์ในแต่ละคำ ซึ่งคำศัพท์แต่ละคำ จะประกอบด้วย ค่าคะแนนด้านบวก (Positive score: PosScore) และค่าคะแนนด้านลบ (Negative Score: NegScore) ดังแสดงในตารางที่ 2.2

ตารางที่ 2.2 โครงสร้างของ SentiWrodNet (Amarouche et al., 2015)

POS	Offset	PosScore	NegScore	SynsetTerms
a	01150475	0	0.625	sorry#1
a	02273643	0.5	0	secure#5
a	01838253	0.625	0	fine#2
n	03931044	0	0	image#3
n	03931044	0	0	picture#1
v	01824736	0.125	0	like#1

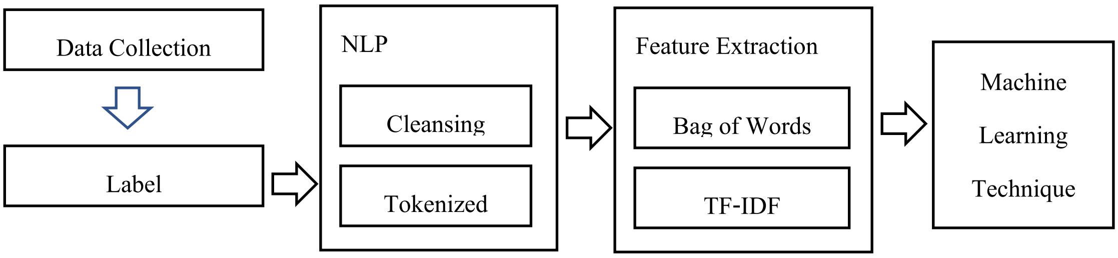
มหาวิทยาลัยเทคโนโลยีสุรนารี

กระบวนการวิเคราะห์ความรู้สึกโดยใช้ SentiWordNet ดังแสดงในภาพที่ 2.17 ใช้การค้นหาคำศัพท์ และค่าคะแนนด้านบวก (Positive Score) ค่าคะแนนด้านลบ (Negative Score) เพื่อคำนวณหาค่าคะแนนความรู้สึก (Sentiment Score) ถ้าค่าคะแนนความรู้สึกเป็นบวก หมายถึงคำศัพท์นั้นเป็นคำศัพท์ด้านบวก ถ้าค่าคะแนนความรู้สึกเป็นลบ หมายถึงคำศัพท์นั้นเป็นคำศัพท์ด้านลบ



ภาพที่ 2.17 กระบวนการวิเคราะห์ความรู้สึกโดยใช้ SentiWordNet

2) การวิเคราะห์โดยอาศัยการเรียนรู้ของเครื่อง (Machine Learning) การวิเคราะห์ความรู้สึกจากการเรียนรู้ด้วยเครื่องอาศัยการใช้ข้อมูลจำนวนมากเพื่อให้เครื่อง (Machine) ใช้ในการฝึก (Train) จำแนก (Classifier) ให้เรียนรู้การจำแนกประเภทของข้อความออกเป็นข้อความที่แสดงความรู้สึกเชิงบวกและลบ โดยมีขั้นตอนดำเนินงานดังแสดงในภาพที่ 2.18 ประกอบด้วยการเก็บข้อมูลสำหรับการสร้างข้อมูลฝึกฝน (Training Set) และการทำป้ายกำกับ (Label) ในแต่ละข้อความ หลังจากนั้นจึงเข้าสู่กระบวนการประมวลผลภาษาธรรมชาติ เพื่อใช้สำหรับการทำความสะอาดข้อมูล การตัดคำ (Cleansing and Tokenized) และการสกัดคุณลักษณะจากข้อความ (Feature Extraction) ซึ่งเป็นกระบวนการแปลงคุณลักษณะต่าง ๆ ให้อยู่ในรูปแบบที่สามารถนำไปใช้ในการสร้างตัวแบบได้ เช่น การสร้างคลังคำศัพท์ (Bag of Words) การค้นหาคำสำคัญจากเอกสาร (Term Frequency-Inverse Document Frequency: TF-IDF) และเข้าสู่กระบวนการเลือกใช้เทคนิคการเรียนรู้ของเครื่อง เช่น ซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ การสุ่มป่าไม้



ภาพที่ 2.18 ขั้นตอนการวิเคราะห์ความรู้สึกโดยอาศัยการเรียนรู้ของเครื่อง

ในงานวิจัยนี้ได้ใช้การวิเคราะห์ความรู้สึกด้วยการวิเคราะห์แบบอิงคลังคำศัพท์ โดยใช้ SentiWordNet ร่วมกับการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยอาศัยการเรียนรู้ของเครื่อง

2.3.3 การประมวลผลรูปภาพ (Image Processing)

2.3.3.1 แนวคิดทฤษฎีเกี่ยวกับการประมวลผลรูปภาพ

Gonzalez and Woods (1977) กล่าวว่า ภาพดิจิทัลคือภาพที่ถูกกำหนดให้เป็นฟังก์ชัน 2 มิติ (Two-dimensional) f(x, y) โดยที่ x และ y เป็นพิกัดเชิงพื้นที่ (ในแนวระนาบ) และแอมพลิจูด (Amplitude) ของ f ที่คู่ใด ๆ ของพิกัด (x, y) เรียกว่าความเข้ม หรือระดับสีเทาของรูปภาพ เมื่อ x, y และค่าความเข้มของ f ทั้งหมดมีปริมาณจำกัด การประมวลผลรูปภาพหมายถึงการประมวลผลภาพดิจิทัลด้วยคอมพิวเตอร์ โดยภาพดิจิทัลประกอบด้วยองค์ประกอบในแต่ละภาพคือตำแหน่ง และค่าเฉพาะ ซึ่งเรียกองค์ประกอบเหล่านี้ว่า องค์ประกอบรูปภาพ (Picture elements หรือ Image elements) ภาพพิกเซล (Pels หรือ Pixel) ซึ่งพิกเซล (Pixel) เป็นคำที่ถูกใช้กันอย่างแพร่หลาย

Acharya and Ray (2005) กล่าวว่า การประมวลผลรูปภาพคือ การบีบอัดและการเข้ารหัสข้อมูลภาพ ด้วยเทคนิคของการจัดเก็บ การประมวลผล การส่งการรับรู้ และการตีความของภาพ โดยมีโปรแกรมประยุกต์สำหรับการประมวลผลภาพจำนวนมากที่สามารถนำมาใช้ประโยชน์ในด้านต่าง ๆ ไม่ว่าจะเป็นการตีความภาพจากระยะไกล (Remote) การใช้งานด้านชีวการแพทย์

ไกรฤกษ์ เชยชื่น และคณะ (2557) กล่าวว่า การประมวลผลภาพคือการกระทำอย่างใดอย่างหนึ่งกับภาพต้นฉบับ (Input Image) เพื่อให้ได้ภาพผลลัพธ์ (Output Image) มีลักษณะของภาพเป็นไปตามที่ต้องการ

สัญญา สมัยมาก และคณะ (2560) กล่าวว่า การประมวลผลภาพ หมายถึงการนำภาพมาประมวลผลหรือคำนวณด้วยคอมพิวเตอร์เพื่อให้ได้ข้อมูลที่ต้องการทั้งในเชิงคุณภาพและปริมาณ โดยมีขั้นตอนต่าง ๆ ที่สำคัญ คือ การทำให้ภาพมีความคมชัดขึ้น การกำจัดสัญญาณรบกวนออกจากภาพ การแบ่งส่วนของวัตถุที่ออกมาจากภาพ เพื่อนำภาพวัตถุที่ได้ไปวิเคราะห์หาข้อมูลเชิงปริมาณ เช่น ขนาด รูปร่าง และทิศทางการเคลื่อนที่ของวัตถุในภาพ โดยสามารถนำข้อมูลเหล่านี้ไปใช้สำหรับการวิเคราะห์ และสร้างเป็นระบบเพื่อใช้ประโยชน์ในด้านต่าง ๆ เช่น ระบบคัดแยกคุณภาพทางการเกษตร ระบบรู้จำลายนิ้วมือ ระบบรู้จำใบหน้าเพื่อประโยชน์ในด้านการรักษาความปลอดภัย

ดังนั้น การประมวลผลรูปภาพในงานวิจัยนี้จึงหมายถึง การนำภาพมาประมวลผลด้วยเทคนิคทางคอมพิวเตอร์เพื่อให้ได้ข้อมูลที่สามารถนำไปวิเคราะห์ สังเคราะห์ และนำไปใช้ประโยชน์ในด้านต่าง ๆ ได้

การประมวลผลภาพสามารถแบ่งการดำเนินการรูปภาพ (Image Operator) ได้เป็น 4 ประเภท คือ การดำเนินการแบบจุด (Point Operators) และการดำเนินการแบบบริเวณ (Local Operators) การดำเนินการทางเลขาคณิต (Geometrical Operators) และการดำเนินการแบบภาพรวม (Global Operators) (Boomgaard, 2017)

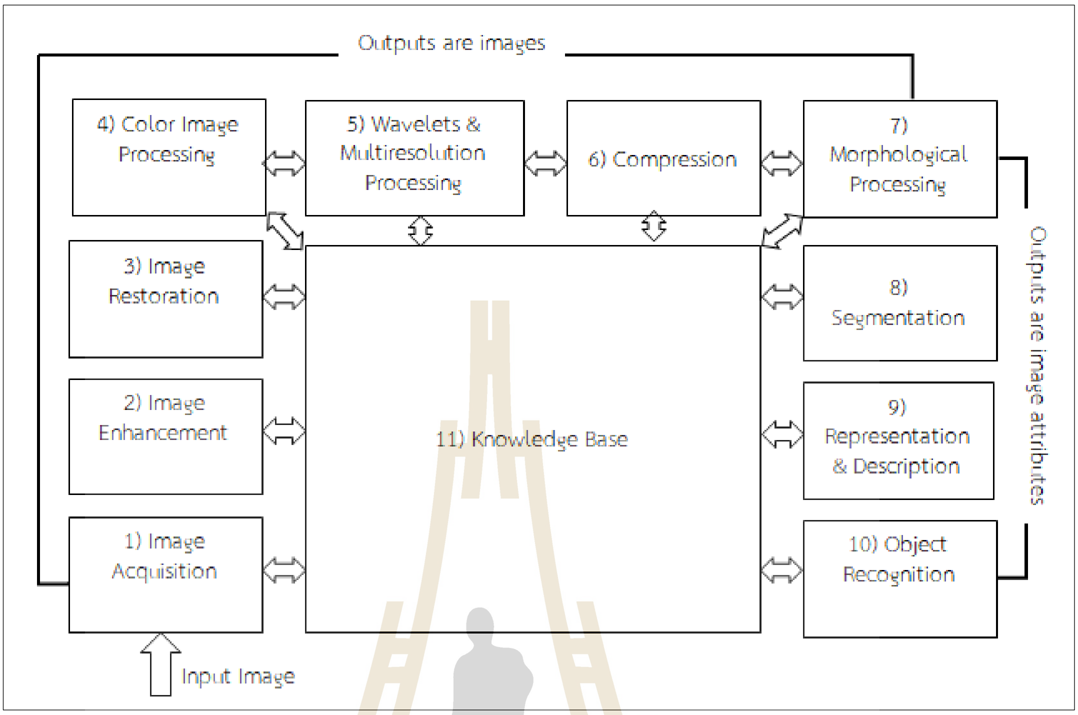
1) การดำเนินการแบบจุด (Point Operators) เป็นวิธีการกระทำการกับภาพต้นฉบับที่ค่าระดับความเข้มเท่าที่แสดงในแต่ละพิกเซล (Pixel) ที่ตำแหน่งสมนัยกันโดยที่ค่าเปลี่ยนแปลงของพิกเซลของภาพผลลัพธ์ไม่ได้ขึ้นอยู่กับค่าพิกเซลที่อยู่บริเวณใกล้เคียงของภาพต้นแบบ

2) ดำเนินการแบบบริเวณ (Local Operators) จะคำนวณค่าของพิกเซลโดยขึ้นกับค่าในตำแหน่งเดียวกันในภาพนำเข้า (Input Image) และใช้ค่าหลายจุดในตำแหน่งใกล้เคียง การดำเนินการแบบบริเวณเป็นหัวใจสำคัญของการประมวลผลภาพส่วนใหญ่

3) การดำเนินการทางเลขาคณิต (Geometrical Operators) เป็นการย้ายพิกเซลไปที่ตำแหน่งใหม่ ซึ่งต่างจากการดำเนินการแบบจุดที่ใช้การเปลี่ยนค่าของพิกเซล

4) การดำเนินการแบบภาพรวม (Global Operators) ต้องการค่าพิกเซลทั้งหมดในรูปภาพนำเข้าเพื่อใช้ในการคำนวณค่าสำหรับพิกเซลเดียวในรูปภาพที่ส่งออก

โดยขั้นตอนพื้นฐานของการประมวลผลภาพดิจิทัล ในระบบคอมพิวเตอร์ประกอบด้วย 11 ขั้นตอน ดังแสดงในภาพที่ 2.19 โดยมีรายละเอียดดังนี้



ภาพที่ 2.19 ขั้นตอนพื้นฐานของการประมวลผลภาพดิจิทัล (Gonzalez and Woods., 2008)

1) การได้มาซึ่งรูปภาพ (Image Acquisition) การได้มาซึ่งรูปภาพคือการนำเข้ารูปภาพที่มีอยู่แล้วในรูปแบบดิจิทัล

2) การปรับปรุงภาพ (Image Enhancement) มีแนวคิดคือการนำรายละเอียดที่ไม่ชัดเจนออก หรือเพื่อเน้นคุณสมบัติที่น่าสนใจในภาพ

3) การคืนค่ารูปภาพ (Image Restoration) คือแนวคิดการปรับปรุงลักษณะที่ปรากฏของภาพ

4) การประมวลผลสีภาพ (Color image processing) คือการสร้างแบบจำลองสีและการประมวลผลในรูปแบบดิจิทัล

5) การประมวลผลเวฟเล็ตและการประมวลผลแบบหลายจุด (Wavelets and Multiresolution Processing) เป็นรากฐานสำหรับการแสดงรูปภาพด้วยความละเอียดระดับต่าง ๆ

6) การบีบอัด (Compression) เพื่อลดพื้นที่การจัดเก็บในการบันทึก หรือรับส่งภาพ

7) การประมวลผลทางสัณฐานวิทยา (Morphological processing) คือการแยกส่วนประกอบเพื่ออธิบายถึงลักษณะของรูปร่าง

8) การแบ่งส่วน (Segmentation) แบ่งรูปภาพออกเป็นส่วนหรือวัตถุที่เป็นส่วนประกอบ

9) การเป็นตัวแทนและคำอธิบาย (Representation and Description) คือการแปลงข้อมูลดิบเป็นรูปแบบที่เหมาะสมสำหรับการประมวลผลคอมพิวเตอร์ในภายหลัง

10) การรับรู้วัตถุ (Object Recognition) เป็นกระบวนการที่กำหนดชื่อ (Label) ให้กับวัตถุ

11) ฐานความรู้ (Knowledge Base) เป็นส่วนที่รวบรวมข้อมูล สารสนเทศ และคุณลักษณะของภาพ

2.3.3.2 แบบจำลองสี (Color Model)

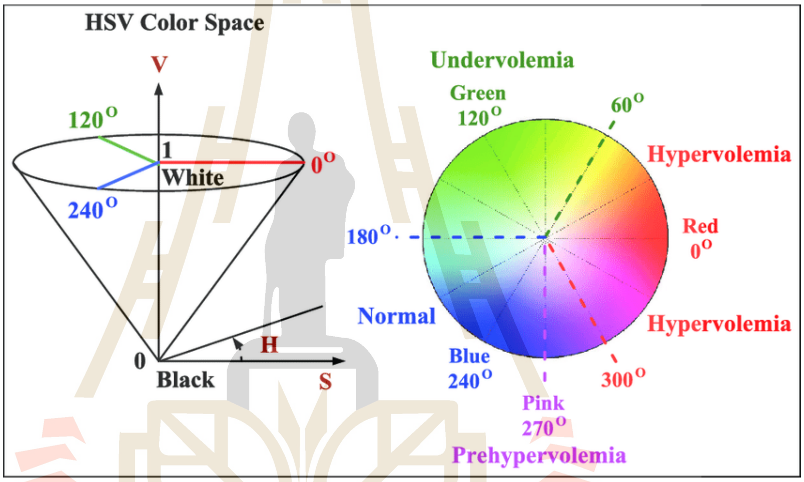
แบบจำลองสีหรือระบบสี แบบจำลองของสีนั้นใช้ในการกำหนดสีต่าง ๆ ให้เป็นไปตามแบบมาตรฐาน ซึ่งแบบจำลองสีในแต่ละแบบจะมีคุณสมบัติที่แตกต่างกัน จึงหมาะกับการใช้งานที่แตกต่างกันออกไป แบบจำลองสีในแต่ละแบบจะใช้แม่สีหลักแตกต่างกันไป มาผสมกันเพื่อให้เกิดค่าสีอื่น ๆ สำหรับแบบจำลองสีที่เกี่ยวข้องในงานวิจัยนี้ ประกอบด้วยแบบจำลองสี 2 แบบ คือ

1) แบบจำลองสีอาร์จีบี (RGB Color Model) คือระบบสีที่มีพื้นฐานมาจากการรับรู้สีของมนุษย์ ถือเป็นโมเดลสีพื้นฐานสำหรับรูปภาพส่วนใหญ่ เกิดจากการผสมกันของแสงสีหลัก (Primary colors) ที่รวมเข้าด้วยกัน คือแสงสีแดง (Red) แสงสีเขียว (Green) และแสงสีน้ำเงิน (Blue) ทำให้เกิดแสงสีใหม่ ดังแสดงในภาพที่ 2.20

ภาพที่ 2.20 แสงสีหลักในระบบสีอาร์จีบี (Ibraheem et al., 2012)

แบบจำลองสีอาร์จีบี แบ่งออกเป็น 2 ประเภท ได้แก่ ปริภูมิสีอาร์จีบีเชิงเส้น(Linear RGB Color: RGB) และปริภูมิสีอาร์จีบีไม่เชิงเส้น (Nonlinear RGB Color Space: R’G’B’) ปริภูมิสีอาร์จีบีเชิงเส้นให้ความสม่ำเสมอของสี โดยใช้ระบบจัดการสี ซึ่งไม่เหมาะสำหรับการวิเคราะห์เชิงตัวเลข ใช้สำหรับโปรแกรมคอมพิวเตอร์กราฟิก มีค่าตั้งแต่ 0 ถึง 1 ในขณะที่ปริภูมิสีอาร์จีบีไม่เชิงเส้นคือข้อมูลนำเข้าจากกล้องถ่ายภาพหรือเครื่องแสกน หรือสีที่อิงอุปกรณ์ (Device Dependent Color) ซึ่งข้อมูลนี้ใช้สำหรับการประมวลผลภาพ โดยมีค่าตั้งแต่ 0 ถึง 255

2) แบบจำลองสี HSV (Hue, Saturation, Value) หรือ HSB (Hue, Saturation, Brightness) ดังแสดงในภาพที่ 2.21 เป็นระบบสีที่ใกล้เคียงกับความคิดของมนุษย์ ประกอบด้วย



ภาพที่ 2.21 ระบบสี HSV (Chen et al., 2017)

- สีของภาพ (Hue) ซึ่งแทนค่าด้วยตัวเลข 0 ถึง 360 องศา (Degree)
- สีแดงอยู่ระหว่าง 0 ถึง 60 องศา
- สีเหลืองอยู่ระหว่าง 61 ถึง 120 องศา
- เขียวอยู่ระหว่าง 121 ถึง 180 องศา
- สีฟ้าอยู่ระหว่าง 181 ถึง 240 องศา
- สีน้ำเงินอยู่ระหว่าง 241 ถึง 300 องศา
- สีม่วงแดงอยู่ระหว่าง 301 ถึง 360 องศา
- ความอิ่มตัว (Saturation) แสดงถึงปริมาณสีเทาในสีใดสีหนึ่ง ตั้งแต่ 0 ถึง 100 เปอร์เซ็นต์ หากมีค่าน้อยจะทำให้ภาพเกิดสีเทามากขึ้น หากมีค่ามากภาพจะมีสีสดขึ้น
- ความสว่าง (Value หรือ Brightness) เป็นค่าที่แสดงถึงปริมาณความสว่างของภาพ โดยมีค่าตั้งแต่ 0 ถึง 100 เปอร์เซ็นต์ หากมีค่ามากภาพจะยิ่งมีความสว่างมาก

HSV สามารถแปลงจาก RGB ได้ดังนี้

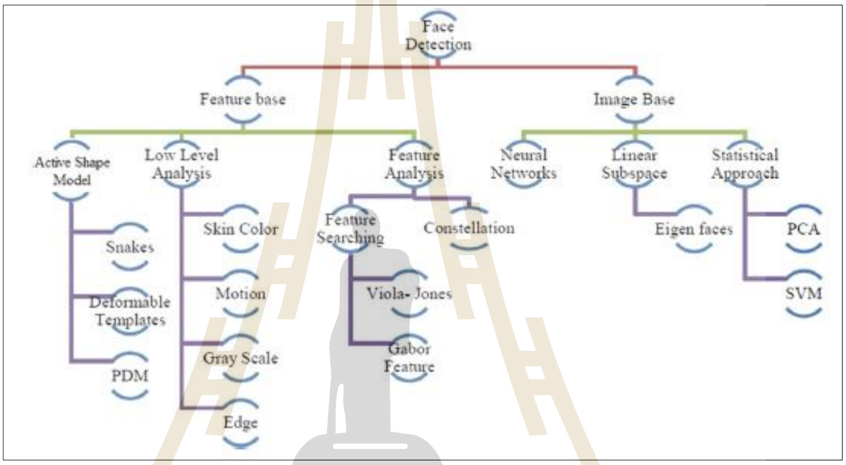
```
Input: RGB
Output: HSV
Method:
Step1: [Find the max and min values]
        M = max(R, G, B), m = min(R, G, B)
Step2: [normalized the RGB values to be in the range [0, 1]]
        r = (M-R)/(M-m)
        g = (M-G)/(M-m)
        b = (M-B)/(M-m)
Step3: [Calculate V value]
        V = max(R, G, B)
Step4: [Calculate S value]
        if M = 0 then S = 0 and H = 180 degrees
        if M <> 0 then S = (M - m) / M
Step5: [Calculate H value]
        if R = M then H = 60(b-g)
        if G = M then H = 60(2+r-b)
        if B = M then H = 60(4+g-r)
        if H >= 360 then H = H - 360
        if H < 0 then H = H + 360
        Where H in the range [0,360], S and H in the
        range[0,100]
Step6: [output HSV]
The calculated H, S, and V are the output of the algorithm.
                        End
```

งานวิจัยนี้ใช้การประมวลผลรูปภาพในการวิเคราะห์ค่าเฉลี่ยระดับพิกเซล (Pixel-level averages) เพื่อหาค่าเฉลี่ยของ เฉดสี (Hue: H) ความอิ่มตัวของสี (Saturation: S) และความสว่างของสี (Value: V) ด้วยไพทอนอิมเมจไลบรารี (Python Image Library: PIL Image) หรือเรียกอีกอย่างว่า Pillow Fork เป็นไลบรารีที่ช่วยเพิ่มความสามารถในการประมวลผลภาพให้กับภาษาไพทอน ออกแบบมาเพื่อให้เข้าถึงข้อมูลที่จัดเก็บในรูปแบบพิกเซล ซึ่งใช้สำหรับการจัดเก็บภาพ (Image Archives) การแสดงผลภาพ (Image Display) และการประมวลผลภาพ (Image Processing) ซึ่งในการประมวลผลภาพด้วย PIL Image ไลบรารีประกอบด้วยฟังก์ชันการประมวลผลภาพพื้นฐาน รวมถึงการดำเนินการตามจุด การกรองด้วยชุดเคอร์เนลการบิดในตัวและการแปลงพื้นที่สี (Lundh, 2011)

2.3.4 การตรวจจับใบหน้า (Face Detection)

การตรวจจับใบหน้า (Face Detection) คือกระบวนการค้นหาใบหน้าของบุคคลจากภาพหรือวิดีโอหลังจากนั้นก็จะทำการประมวลผลภาพใบหน้าที่ได้สำหรับขั้นตอนถัดไป เช่นใช้ในกระบวนการรู้จำใบหน้า (Face Recognition) ซึ่งกระบวนการรู้จำใบหน้าคือการนำภาพที่ตรวจจับ และประมวลผลแล้วจากการตรวจจับใบหน้ามาเปรียบเทียบกับฐานข้อมูลของใบหน้า เพื่อระบุว่าใบหน้านั้นตรงกับบุคคลใด

การตรวจจับใบหน้า (Face Detection) เป็นส่วนหนึ่งของการประมวลผลภาพ โดย Kumar et al. (2018) ได้แบ่งขั้นตอนวิธีที่ใช้ในการตรวจจับใบหน้าออกเป็น 2 หมวดหมู่ ได้แก่ วิธีการตามฐานคุณสมบัติ (Feature Base) และวิธีการตามฐานรูปภาพ (Image Base) โดยที่วิธีการตามฐานคุณสมบัติ ใช้การดึงคุณสมบัติของภาพและจับคู่กับคุณสมบัติของใบหน้า ในขณะที่วิธีการตามฐานรูปภาพใช้การจับคู่ระหว่างรูปภาพสำหรับการฝึก (Training Images) และรูปภาพสำหรับการทดสอบ (Testing Image) ซึ่งวิธีการต่าง ๆ ที่ใช้สำหรับการตรวจจับใบหน้า แสดงได้ดังภาพที่ 2.22



ภาพที่ 2.22 วิธีการต่าง ๆ ที่ใช้สำหรับการตรวจจับใบหน้า

ในขณะที่ Rizvi (2011) ได้แบ่งวิธีการสำหรับการตรวจจับใบหน้าออกเป็น 4 กลุ่ม เช่นเดียวกับ รุสลี่ สุทธวีร์กูล และ วิไลพร แซ่ลี้ (2554) คือ วิธีการ Knowledge-based วิธีการ Feature invariant วิธีการ Template Matching และวิธีการ Appearance – based โดยระบุว่าวิธีการ Appearance – based เป็นวิธีการที่มีการพัฒนามากที่สุด และเป็นที่ยอมรับสามารถนำมาใช้งานได้จริงในหลาย ๆ ด้าน ซึ่งในแต่ละวิธีการมีรายละเอียดดังต่อไปนี้

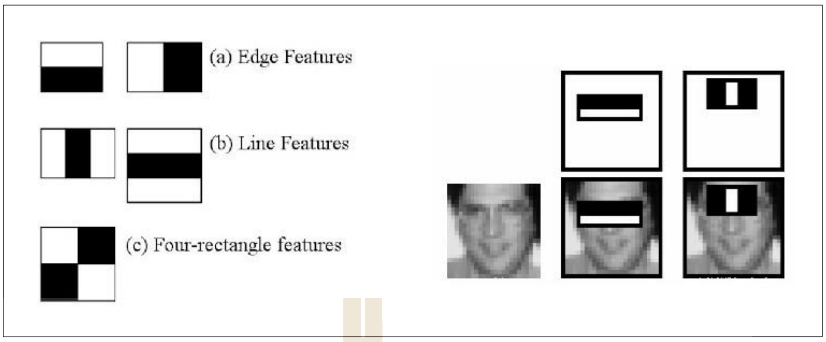
1) วิธีการใช้กฎพื้นฐานความรู้ (Knowledge-based methods) วิธีการนี้ใช้กฎเกณฑ์พื้นฐานความรู้ของมนุษย์ในเรื่องส่วนประกอบสำคัญบนใบหน้า เช่น กฎที่ระบุว่าในรูปที่มีใบหน้าจะต้องประกอบด้วยดวงตา 2 ดวงอยู่ในตำแหน่งที่สมมาตรกัน จะต้องมีจมูก และปากอย่างละ 1 ซึ่งจุดสำคัญดังกล่าวจะมีความสัมพันธ์กันโดยใช้ระยะทางและตำแหน่งตามกฎเกณฑ์ที่ตั้งไว้ อุปสรรคที่สำคัญของวิธีการนี้คือ การหากฎเกณฑ์ที่จำเพาะเจาะจงซึ่งใช้ในการจำแนกหน้าของมนุษย์นั้นทำได้ยาก เพราะหากมีกฎเกณฑ์มากเกินไปจะมีผลให้การตรวจจับใบหน้าทำได้ยาก หรือหากมีกฎเกณฑ์น้อยเกินไปก็อาจเกิดข้อผิดพลาดได้ง่าย

2) วิธีการหาลักษณะเด่นบนใบหน้า (Feature invariant approaches) วิธีการนี้มีเป้าหมายเพื่อค้นหาโครงสร้างลักษณะเด่น (Feature) ที่มีอยู่แม้สภาพของตำแหน่งการจัดวาง หรือสภาพแสงจะแตกต่างกันไป ซึ่งวิธีการนี้ตั้งอยู่บนพื้นฐานของการสังเกตของมนุษย์ที่สามารถจดจำวัตถุใด ๆ ได้ไม่ว่าจะมีการเปลี่ยนแปลงตำแหน่งการจัดวาง หรือสภาพแสดงที่แตกต่างกัน ซึ่งจะต้องมีคุณลักษณะบางอย่างที่ไม่มีผลต่อการเปลี่ยนแปลงดังกล่าว ในการตรวจจับใบหน้าโดยการหาลักษณะเด่นบนใบหน้าสามารถใช้ร่วมกับการตรวจจับขอบ (Edge Detection) เพื่อดึงลักษณะเด่น เช่น คิ้ว ดวงตา จมูก และปาก ออกมา ปัญหาของวิธีการนี้คือ สภาพแสง และสัญญาณรบกวนต่าง ๆ สามารถทำลายคุณสมบัติ หรือคุณลักษณะเด่นบนใบหน้าได้โดยง่าย และจุดเด่นบนใบหน้าอาจตรวจพบได้ยากในบางกลุ่มใบหน้า

3) วิธีการจับคู่โดยแม่แบบมาตรฐาน (Template matching methods) วิธีการนี้ใช้แม่แบบมาตรฐานหลายแบบของใบหน้าถูกจัดเก็บเพื่ออธิบายใบหน้าโดยรวมหรือแยกจากกัน ความสัมพันธ์ระหว่างภาพนำเข้าและแม่แบบจะถูกคำนวณเพื่อการตรวจจับและใช้สำหรับการค้นหาใบหน้า ในงานวิจัยของ Scassellati (1998) ได้กำหนดแม่แบบของใบหน้าประกอบด้วย 16 รูปแบบ และ 23 ความสัมพันธ์ ใบหน้าที่นำเข้ามาจะนำมาหาค่าสหสัมพันธ์ (Correlation Value) กับใบหน้าที่เป็นรูปแบบมาตรฐาน (Standard Pattern) วิธีการนี้สามารถทำได้ง่าย แต่ประสิทธิภาพในการตรวจจับใบหน้ายังได้ผลลัพธ์ที่ไม่ดี ซึ่งเกิดจากผลของขนาด ตำแหน่งการวาง และรูปทรงของใบหน้าที่แตกต่างจากแม่แบบมาตรฐาน

4) วิธีการตามลักษณะที่ปรากฏ (Appearance-based methods) วิธีการนี้อยู่บนพื้นฐานการวิเคราะห์ทางสถิติ และการเรียนรู้ของเครื่อง (Machine Learning) ซึ่งใช้ในการหาคุณลักษณะและความสัมพันธ์ของรูปใบหน้า และรูปที่ไม่ใช่ใบหน้า โดยใช้แบบจำลองการกระจาย (Distribution Models) และใช้การลดมิติ (Dimension Reduction) เพื่อเพิ่มประสิทธิภาพในการตรวจจับใบหน้า

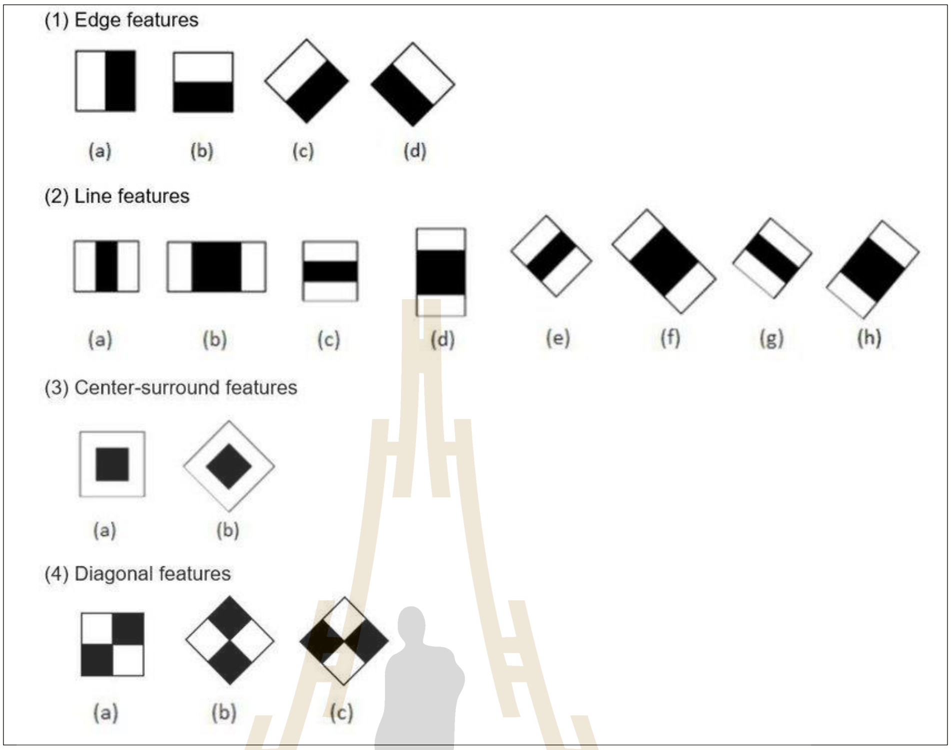
สำหรับงานวิจัยนี้ใช้การตรวจจับใบหน้าโดยใช้ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของ ฮาร์ (Haar feature-based cascade classifiers) ของ Open CV ซึ่งมีวิธีการที่ง่ายและทำงานได้อย่างรวดเร็ว การทำงานของการตรวจจับใบหน้าโดยใช้ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของ ฮาร์ คือการใช้รูปสี่เหลี่ยมที่มีลักษณะตามภาพที่ 2.23 จำนวน 6,061 รูป เปรียบเทียบกับภาพเป็นชั้น ๆ ถ้าภาพในพื้นที่นั้น ๆ ผ่านการเปรียบเทียบคุณลักษณะจนครบทุกชั้น จะตีความเป็นภาพของใบหน้า



ภาพที่ 2.23 ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ (OpenCV, 2022)

โอเพนซีวี (Open Source Computer Vision Library: OpenCV) เป็นไลบรารีโอเพนซอร์สที่ประกอบด้วย ขั้นตอนวิธีคอมพิวเตอร์วิทัศน์ (Computer Vision Algorithms) จำนวนมาก โอเพนซีวี (OpenCV) ประกอบด้วย ไลบรารีจำนวนมาก เช่น โมดูลโครงสร้างข้อมูลพื้นฐาน(Core functionality) และฟังก์ชันพื้นฐานที่ใช้โดยโมดูลอื่น ๆ ทั้งหมด โมดูลการประมวลผลภาพ (Image Processing) ที่รวมถึงการกรองภาพเชิงเส้นและไม่เป็นเชิงเส้น การแปลงภาพเชิงเรขาคณิต (การปรับขนาด ความเกี่ยวข้องและการแปรปรวนของเปอร์สเปคทีฟ การรีแมปตารางทั่วไป) การแปลงพื้นที่สี ฮิสโตแกรม (OpenCV, 2022)

การตรวจจับใบหน้าโดยใช้ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ (Haar feature-based cascade classifiers) เป็นวิธีการตรวจจับวัตถุที่มีประสิทธิภาพซึ่งเป็นส่วนหนึ่งของไลบรารี OpenCV นำเสนอโดย Viola and Jones (2001) ในการวิจัยเรื่อง “Rapid Object Detection using a Boosted Cascade of Simple Features” ซึ่งใช้การเรียนรู้ของเครื่อง (Machine Learning) ในการวิเคราะห์รูปภาพ ซึ่งถูกแบ่งออกเป็น 2 กลุ่ม คือ กลุ่มรูปที่ถูกต้องในรูปแบบต่าง ๆ ที่มีการตัดส่วนที่ไม่เกี่ยวข้องออก ให้เหลือเพียงส่วนที่ต้องการเท่านั้น และกลุ่มรูปทั่วไปที่ไม่เกี่ยวข้องกับสิ่งที่ต้องการ ซึ่งวิธีที่นำเสนอนี้ช่วยลดการใช้ทรัพยากร และเวลาในการประมวลผลอย่างมาก โดยใช้การแบ่งพื้นที่ของภาพด้วยรูปแบบ 4 แบบ คุณลักษณะของเส้น (Line Feature) คุณลักษณะของขอบ (Edge Feature) คุณลักษณะรอบทิศทางตรงกลาง (Center-Surround Features) และลักษณะแนวทแยง (Diagonal Features) ดังแสดงในภาพที่ 2.24



ภาพที่ 2.24 คุณลักษณะของฮาร์ไลค์ฟีเชอร์ (Haar-like features) (Yang, Zhang and Li, 2019)

การตรวจจับใบหน้าโดยใช้ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ ด้วยโอเพนซีวี (OpenCV) มีวิธีการที่ง่ายและทำงานได้อย่างรวดเร็ว ซึ่งได้รับความนิยมมานาน แม้ในปัจจุบันจะมีการตรวจจับใบหน้าหลายวิธีแต่วิธีนี้ก็ยังเป็นที่นิยมใช้กันอย่างแพร่หลาย (ชัชวาลย์ บุตรมาศ, 2560)

2.3.5 การประมวลผลสัญรูปอารมณ์ (Emoticon Processing)

สัญรูปอารมณ์ได้รับการพิสูจน์แล้วว่ามีความสำคัญในการจำแนกข้อความอัตโนมัติอย่างไม่เป็นทางการ ความแตกต่างระหว่างสัญรูปอารมณ์ด้านบวกและลบคือ ใช้เพื่อสร้างตัวอย่างข้อความที่เป็นบวกและลบโดยอัตโนมัติ ตัวอย่างเหล่านี้นั้นใช้ในการฝึกอบรมและทดสอบแบบจำลองการจำแนกความเชื่อมั่นโดยใช้เทคนิคการเรียนรู้ของเครื่อง

มีงานวิจัยจำนวนมากได้วิเคราะห์ความรู้สึกผ่านการแสดงความคิดเห็นบนเครือข่ายสังคมออนไลน์ เช่น เฟสบุค ทวิตเตอร์ แสดงให้เห็นถึงอารมณ์ในข้อความแสดงความคิดเห็น โดยมีหลายงานวิจัยได้ศึกษาถึงการใช้งานสัญรูปอารมณ์ เนื่องจากสัญรูปอารมณ์ได้ถูกนำมาใช้งานในเครือข่ายสังคมออนไลน์มากขึ้น ไม่ว่าจะเป็นการศึกษาถึงความคิดเห็นที่มีการวิเคราะห์เพื่อประเมินทัศนคติทางอารมณ์ของขนาดใหญ่ สัดส่วนของประชากรต่อประเด็นต่าง ๆ

Novak et al. (2015) ได้ศึกษาถึงการวิเคราะห์ความรู้สึกผ่านทางสัญรูปอารมณ์ การจัดอันดับความรู้สึกของสัญรูปอารมณ์ โดยใช้ข้อมูลจากทวิตเตอร์จำนวน 70,000 ทวีต และใช้สัญรูปจำนวน 969 ในการวิเคราะห์ โดยใช้สถิติการวิเคราะห์สัมประสิทธิ์สหสัมพันธ์เสปียร์แมน (Spearman) และสัมประสิทธิ์สหสัมพันธ์เพียร์สัน (Pearson) ผลการศึกษาทำให้ได้ฐานข้อมูลสัญรูปอารมณ์ประกอบด้วยค่าคะแนนความรู้สึก ค่าคะแนนด้านบวก ด้านลบ ของสัญรูปอารมณ์ในแบบต่าง ๆ ดังแสดงในภาพที่ 2.25

Emoji Sentiment Ranking v1.0

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name	Unicode block
		0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY	Emoticons
		0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART	Dingbats
		0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT	Miscellaneous Symbols
		0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES	Emoticons
		0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
		0x1f618	3648	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS	Emoticons
		0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES	Emoticons

ภาพที่ 2.25 ฐานข้อมูลการวิเคราะห์ความรู้สึกสัญรูปอารมณ์

ในงานวิจัยนี้ได้ใช้ฐานข้อมูลการวิเคราะห์ความรู้สึกสัญรูปอารมณ์ มาใช้ในการวิเคราะห์สัญรูปอารมณ์ ซึ่งเป็นส่วนหนึ่งในการแสดงความคิดเห็นผ่านทวิตเตอร์ ร่วมกับการวิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็น และนำผลลัพธ์ที่ได้จากกระบวนการนี้เข้าสู่กระบวนการสร้างแบบจำลองต่อไป

มหาวิทยาลัยเทคโนโลยีสุรนารี

2.4 เทคนิคที่นำมาใช้ในการสร้างแบบจำลอง

2.4.1 การเรียนรู้ของเครื่อง (Machine Learning)

การเรียนรู้ของเครื่อง หมายถึงการเปลี่ยนแปลงในระบบที่ทำงานที่เกี่ยวข้องกับปัญญาประดิษฐ์ (Artificial Intelligence: AI) หรือเป็นศาสตร์แขนงหนึ่งของปัญญาประดิษฐ์ โดยสามารถใช้การเรียนรู้ของเครื่องเพื่อปรับปรุงการออกแบบเครื่องจักรที่มีอยู่ในปัจจุบัน งานดังกล่าวเกี่ยวข้องกับการจดจำ การวินิจฉัย การวางแผนการควบคุมหุ่นยนต์ และการทำนาย (Nilsson, 1998) ซึ่งปัญญาประดิษฐ์ คือศาสตร์ที่ว่าด้วยการสร้างความฉลาดให้กับเครื่องจักร โดยเฉพาะอย่างยิ่งกับระบบคอมพิวเตอร์ เพื่อให้สามารถคำนวณ คิดหาเหตุผล มีการเรียนรู้ได้เสมือนกับสมองของมนุษย์ และตอบสนองต่อสถานการณ์ต่าง ๆ ได้ เพื่อช่วยให้คอมพิวเตอร์สามารถทำงานแทนมนุษย์ได้อย่างมีประสิทธิภาพ โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning)

นอกจากนั้นแล้ว การเรียนรู้ของเครื่องยังเป็นกระบวนการที่ทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้ด้วยตนเอง หลักการทำงานเกี่ยวข้องกับการศึกษาและการพัฒนาขั้นตอนวิธีขั้นตอนวิธีที่สามารถเรียนรู้ และปรับตัวตามข้อมูลที่ได้รับโดยอาศัยแบบจำลอง (Model) ที่สร้างจากชุดข้อมูลที่มีความสัมพันธ์กัน และสามารถนำไปทำนายหรือใช้ในการตัดสินใจภายหลัง จากการศึกษาแนวคิดทฤษฎีที่เกี่ยวข้องกับการเรียนรู้ของเครื่องพบว่า การเรียนรู้ของเครื่องประกอบด้วย 3 รูปแบบ (บุญเสริม กิจศิริกุล, 2548; Mitchell, 1997; Armstrong, 2015) ดังนี้

1) การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนรู้โดยอาศัยข้อมูลที่ป้อนเข้าไปเก็บไว้เป็นตัวอย่าง เพื่อให้คอมพิวเตอร์ใช้ในการเปรียบเทียบกับข้อมูลที่เข้ามาใหม่ แล้วทำนายหรือจัดหมวดหมู่ที่มีความเหมือนกันมากที่สุดให้อยู่ด้วยกัน ผลลัพธ์ที่ได้คือ การจัดหมวดหมู่ (Classification) และการวิเคราะห์การถดถอย (Regression)

2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการเรียนรู้ที่ไม่มีข้อมูลตัวอย่างที่ใช้บอกว่าข้อมูลนั้นคืออะไร แต่จะเรียนรู้จากการหาความสัมพันธ์จากข้อมูลนำเข้า (input) โดยพิจารณาจากรูปแบบ (patterns) หรือโครงสร้างของข้อมูล (data structure) แล้วนำมาจัดเป็นกลุ่ม (cluster) บนพื้นฐานของความเหมือน (similarities) และความแตกต่าง (differences) ระหว่างรูปแบบของข้อมูลนำเข้า (input patterns) ตัวอย่าง การหาโครงสร้างของข้อมูลที่ซ่อนอยู่ประกอบด้วย การลดมิติข้อมูล (dimensionality reduction) การจัดกลุ่ม (clustering) และการเรียนรู้แบบซับซ้อน (manifold learning)

3) การเรียนรู้แบบเสริมกำลัง (reinforcement learning) เป็นการเรียนรู้แบบไม่มีผู้สอนแต่จะมีปฏิสัมพันธ์กับสิ่งแวดล้อม ทำให้เกิดการกระตุ้นในการสร้างแบบจำลองเพื่อตอบสนองว่าจะต้องทำอย่างไรต่อไป การกระทำจะเปลี่ยนไปตามสภาพแวดล้อมในขณะนั้น ตัวอย่าง การนำไปประยุกต์ใช้งาน เช่น ควบคุมหุ่นยนต์ การเล่นเกม และการแนะนำเส้นทางการขับรถยนต์

การเรียนรู้ของเครื่องคือการออกแบบและพัฒนาขั้นตอนวิธีเพื่อทำให้คอมพิวเตอร์มีความสามารถในการเรียนรู้คอมพิวเตอร์เชิงอุปนัย และเชิงอนุมาน โดยเชิงอุปนัยเป็นการค้นหากฎลักษณะแบบแผนหรือข้อสรุปต่างๆจากการสังเกตกลุ่มข้อมูลขนาดใหญ่ และเชิงอนุมาน เป็นการหาข้อสรุปจากหลักฐานหรือข้อเท็จจริงที่มีอยู่ สิ่งสำคัญของเทคนิคการเรียนรู้ของเครื่องคือการสกัดเอาความรู้หรือสารสนเทศจากข้อมูลโดยอัตโนมัติด้วยวิธีการคำนวณ หรือวิธีการทางสถิติ ดังนั้นเทคนิคการเรียนรู้ของเครื่องจึงมีความสัมพันธ์อย่างใกล้ชิดกับการทำเหมืองข้อมูล (Data mining) (อังศุมาลี สุทธภักติ, 2553) ซึ่งใช้ในการจำแนกหมวดหมู่ (Classification) หรือการจัดกลุ่ม (Clustering) โดยมีขั้นตอนวิธีการจำแนกหมวดหมู่ที่ใช้ในงานวิจัยนี้ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ต้นไม้การตัดสินใจ (Decision Tree) นาอีฟเบย์ (Naïve Bayes) การสุ่มป่าไม้ (Random Forest) และเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron : MLP)

2.4.2 นาอีฟเบย์ (Naïve Bayes)

นาอีฟเบย์ (Naïve: NB) เป็นวิธีการจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง ซึ่งนาอีฟเบย์ได้รับการพิสูจน์แล้วว่ามีประสิทธิภาพในการใช้งานที่หลากหลาย เช่น การจำแนกข้อความ การวินิจฉัยทางการแพทย์ และการจัดการประสิทธิภาพของระบบ (Domingosand and Pazzani, 1997) ใช้ในการสร้างโมเดลการจำแนกประเภทข้อมูลที่ใช้หลักความน่าจะเป็นซึ่งอยู่บนพื้นฐานของ Bayes' Theorem และสมมติฐาน ที่ให้การเกิดของเหตุการณ์ต่าง ๆ เป็นอิสระต่อกัน (Independence) กำหนดให้ P(h) ความน่าจะเป็นที่จะเกิดเหตุการณ์ h และ P(h|D) คือความน่าจะเป็นที่จะเกิดเหตุการณ์ h เมื่อเกิดเหตุการณ์ D จากตัวแปรที่กำหนดและแนวคิดของ Bayes' Theorem นั้นสามารถทำนายเหตุการณ์ที่พิจารณาได้จากการเกิดของเหตุการณ์ต่าง ๆ ได้ดังสมการ ที่ 2.1

$$P(h|D) = \frac{[P(D|h) * P(h)]}{P(D)} \tag{2.1}$$

โดย

D แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็นposteriori probability ของสมมติฐาน h คือ P(h|D) ตามทฤษฎี

P(h) คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h

P(D) คือ ความน่าจะเป็นก่อนหน้าของชุดข้อมูลตัวอย่าง D

P(h|D) คือ ความน่าจปีนของ h เมื่อรู้ D

P(D|h) คือ ความน่าจะเป็นของ D เมื่อรู้ h

กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม V_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว X= $\{a_1, a_2, \ldots a_n\}$ หรือใช้สัญลักษณ์ว่า $P(a_1, a_2, \ldots a_n)$ ดังสมการที่ 2.2

$$P(a_1, a_2, \ldots a_n | V_j) = \prod_{i=1}^{n} P(a_i|V_j) \tag{2.2}$$

โดยที่ II หมายถึง ผลคูณของค่า $P(a_i|V_j)$ ทั้งหมด

$i = 1,2,3,\ldots,n$ และ $j = 1,2,3,\ldots,n$

การนำวิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) ไปใช้ มีวิธีการดังต่อไปนี้

1. หาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า P($a_1, a_2, \ldots, a_n | v_j$) จากสมการที่ 2.2 มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้นๆ คือ $P(v_j)$ ได้เท่ากับ v_{NB}

2. นำค่าที่ได้ มาเปรียบเทียบกัน กลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการที่ 2.3

$$V_{NB} = \underset{V_j \in V}{\operatorname{argmax}} P(v_j) \times \prod_{i=1}^{n} P(a_i|V_j) \tag{2.3}$$

จากสมการข้างต้นสามารถสร้างเป็นขั้นตอนวิธีการเรียนรู้เบย์อย่างง่ายได้ดังนี้

$Assume\ target\ function\ : X \rightarrow V\ , where\ each$
$instance\ x\ described\ by\ attributes\ (a_1, a_2 \ldots a_n).$
$Most\ probable\ value\ of\ f(x)\ is:$

$$v_{MAP} = \underset{v_j \in V}{argmax}\ P(v_j|a_i, a_2 \ldots a_n)$$

$$v_{MAP} = \underset{v_j \in V}{argmax} \frac{P(a_i, a_2 \ldots a_n\ |v_j)\ P(v_j)}{P(a_i, a_2 \ldots a_n)}$$

$$v_{MAP} = \underset{v_j \in V}{argmax}\ P(a_i, a_2 \ldots a_n\ |v_j)\ P(v_j)$$

$Navie\ Bayes\ assumption:$

$$P(a_i, a_2 \ldots a_n\ |v_j) = \prod_i P(a_i|v_j)$$

$which\ gives$

$$Naive\ Bayes\ classifier: v_{NB} = \underset{v_j \in V}{argmax}\ P(v_j)\ \textstyle\prod_i P(a_i|v_j)$$

มหาวิทยาลัยเทคโนโลยีสุรนาร

2.4.3 ต้นไม้การตัดสินใจ (Decision Tree)

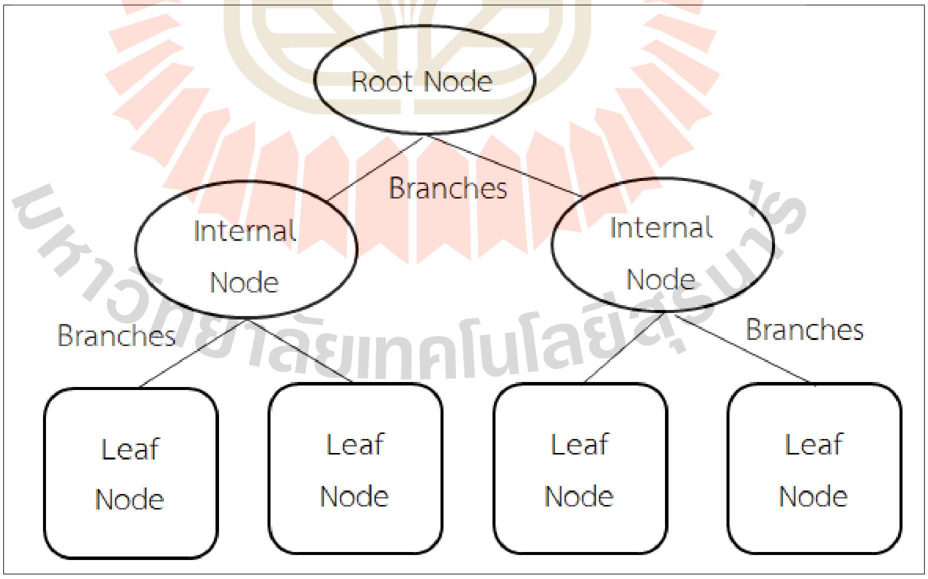
ต้นไม้การตัดสินใจ (Decision Tree: DT) เป็นเครื่องมือที่ช่วยในการตัดสินใจ โดยการใช้โครงสร้างกราฟที่มีลักษณะคล้ายต้นไม้ ทำให้มนุษย์สามารถแปลผลได้ง่ายขึ้น เพื่อใช้ประโยชน์ในการจำแนกประเภท (Classification) ทั้งในปัญหาแบบไบนารี (Binary Classes) และปัญหาแบบหลายคลาส (Multiple Classes) รวมทั้งยังสามารถนำไปใช้แสดงโครงสร้างข้อมูลในฐานข้อมูล (Database) เพื่อความรวดเร็วในการค้นหา หรือการวิเคราะห์การถดถอย (Regression Analysis) โดยต้นไม้การตัดสินใจจัดอยู่ในกลุ่มของการเรียนรู้แบบมีผู้สอน (Supervised Learning)

องค์ประกอบของต้นไม้การตัดสินใจ ดังแสดงในภาพที่ 2.26 ประกอบด้วยองค์ประกอบหลัก 3 องค์ประกอบ ได้แก่

1) โหนด (Node) ใช้สำหรับระบุชื่อของลักษณะประจำ (Attribute) เป็นจุดที่แยกข้อมูลว่าจะให้ไปในทิศทางใด ซึ่งโหนดที่อยู่สูงสุดเรียกว่า โหนดราก (Root Node) และโหนดที่อยู่ระหว่าง ใบ กับ โหนดราก เรียกว่า โหนดภายใน (Internal Node)

2) กิ่ง (Branch) ใช้สำหรับระบุค่าของแต่ละลักษณะประจำของปมที่เหนือกิ่งกิ่งนั้น หรือเรียกอีกอย่างว่า เส้นเชื่อม (Edge) หรือการโยง (Link) โดยจำนวนกิ่งจะเท่ากับคุณสมบัติของโหนด

3) ใบ (Leaf) ใช้สำหรับระบุค่าเป้าหมาย (Target) หรือคลาส (Class) ของปัญหา



ภาพที่ 2.26 องค์ประกอบของต้นไม้การตัดสินใจ

ต้นไม้การตัดสินใจสามารถนำไปใช้ได้กับปัญหาและข้อมูลหลายประเภท เช่น ข้อมูลค่าไม่ต่อเนื่อง (Discrete Values) ข้อมูลค่าต่อเนื่อง (Continuous Values) ข้อมูลค่าสูญหาย (Missing Values) และข้อมูลที่มีค่าถ่วงน้ำหนัก (Weighted Values) ซึ่งขั้นตอนวิธีในการสร้างต้นไม้การตัดสินใจขึ้นอยู่กับ ประเภทของตัวแปรเป้าหมาย โดยขั้นตอนวิธีในการสร้างต้นไม้การตัดสินใจแบ่งออกได้ 4 ขั้นตอนวิธี ได้แก่

1) ID3 (Iterative Dichotomiser 3)
2) C4.5
3) CART (Classification And Regression Tree)
4) CHAID (Chi-square automatic interaction detection Performs multi-level splits when computing classification trees)

ซึ่งขั้นตอนวิธี CART และ CHAID เป็นขั้นตอนวิธีที่ใช้ในการสร้างแบบจำลองที่ใช้สำหรับงานทางด้านธุรกิจ ในงานวิจัยนี้ผู้วิจัยใช้ขั้นตอนวิธี C4.5 ในการสร้างแบบจำลองต้นไม้การตัดสินใจ

2.4.3.1 ขั้นตอนวิธี ID3 (Iterative Dichotomiser 3)

ขั้นตอนวิธี ID3 เป็นต้นไม้การตัดสินใจที่นำเสนอโดย Quinlan (J. R. Quinlan, 1986) โดยการใช้ค่า Information Gain มาช่วยในการเลือกลักษณะประจำที่เหมาะสมในแต่ละโหนด ซึ่งเป็นการอาศัยทฤษฎีข่าวสาร (Information Theory) ได้กล่าวถึงการวัดปริมาณข่าวสารของข้อมูลด้วยค่าเอนโทรปี (Entropy) ซึ่งค่าเอนโทรปีเป็นปริมาณที่บ่งบอกความไม่แน่นอน (Uncertainty) ความไม่เป็นระเบียบ (Disorder) ความไม่บริสุทธิ์ (Impurity) ของข้อมูล กล่าวคือ หากค่าเอนโทรปีมากจะมีข่าวสารมาก โดยค่าเอนโทรปี ***H(X)*** สามารถคำนวณได้จากสมการที่ 2.4

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_b P(x_i) \tag{2.4}$$

โดย $P(X_i)$ คือความน่าจะเป็นที่เหตุการณ์ x_i จะเกิดขึ้นและมีเหตุการณ์ทั้งหมด n เหตุการณ์

ค่าเอนโทรปีจะถูกนำมาใช้ในการคำนวณค่า Information Gain (IG) เพื่อใช้สำหรับการเลือกลักษณะประจำที่ดีที่สุด a จากเซ็ตของลักษณะประจำทั้งหมด A $(a \in A)$ ในเซ็ตของตัวอย่างทั้งหมด S สำหรับแต่ละโหนด โดยแสดงได้จากสมการที่ 2.5 และสมการที่ 2.6

$$IG(S,a) = H(S) - H(S|a) \tag{2.5}$$

$$IG(S,a) = H(S) - \sum_{v \in Values(a)} \frac{|S_v|}{|S|} H(S_v) \tag{2.6}$$

โดย v หมายถึง ค่าของลักษณะประจำ $Values(a)$ และ $S_v = \{s|s \in S, Value(s,a) = v\}$ โดยเครื่องหมาย หมายถึงจำนวนสมาชิกในเซต s

ขั้นตอนวิธี ID3 สามารถสร้างต้นไม้การตัดสินใจจากชุดตัวอย่างได้โดยอาศัยหลักการทางสถิติในการเลือก จึงทำให้ขั้นตอนวิธี ID3 สามารถรองรับข้อมูลที่มีความผิดพลาดได้ดีขึ้น แต่มีข้อเสียคือ ใช้การค้นหาด้วยขั้นตอนวิธีแบบละโมบ (Greedy Algorithm) ทำให้ต้นไม้ที่ได้อาจไม่ใช่ต้นไม้ที่ดีที่สุด เนื่องจากไม่มีการตรวจสอบย้อนหลัง คือเมื่อเลือกลักษณะประจำที่ใช้ในแต่ละโหนดแล้ว จะไม่มีการกลับมาตรวจสอบซ้ำอีก

2.4.3.2 ขั้นตอนวิธี C4.5

ขั้นตอนวิธี C4.5 (J. Ross Quinlan, 1993) เป็นขั้นตอนวิธีการสร้างต้นไม้การตัดสินใจที่ปรับปรุงต่อเนื่องจากขั้นตอนวิธี ID3 โดยเพิ่ม Generalization เพื่อเลี่ยงปัญหา Overfitting และการรองรับข้อมูลในรูปแบบที่หลากหลายมากขึ้น เช่น ข้อมูลที่มีค่าต่อเนื่อง และไม่ต่อเนื่อง ข้อมูลที่มีค่าถ่วงน้ำหนักหรือมูลค่าที่แตกต่างกัน ข้อมูลที่มีค่าสูญหายบางลักษณะประจำ เป็นต้น โดยขั้นตอนวิธี C4.5 ได้ใช้ Gain Ratio ในการปรับเปลี่ยน Information Gain เพื่อลดความโน้มเอียง (Bias) ซึ่ง Gain ratio สามารถคำนวณได้ดังสมการที่ 2.7

$$Gain\ Ratio = \frac{Entropy(before) - \sum_{j=1}^{K} Entropy(j, after)}{\sum_{j=1}^{K} w_j log_2\, w_j} \quad (2.7)$$

โดย K คือตัวเลขของซับเซ็ต (Subsets) ที่สร้างจากการแบ่งแยก และ $(j, after)$ คือซับเซ็ตของ j หลังจากแบ่งแยกแล้ว

Generalization เป็นการทำให้ต้นไม้การตัดสินใจที่ได้นั้นสามารถนำไปใช้กับข้อมูลทั่วไปได้อย่างมีประสิทธิภาพมากขึ้น

ปัญหาโอเวอร์ฟิตติง (Overfitting) คือ ปัญหาที่เกิดขึ้นเมื่อประสิทธิภาพที่วัดจากกลุ่มทดลองมีค่าลดลงเมื่อความซับซ้อนมีมากขึ้น เนื่องจากมีจำนวนตัวอย่างที่ใช้ในการเรียนรู้จริงอย่างจำกัดเมื่อเทียบกับจำนวนตัวอย่างทั้งหมด การแก้ปัญหา Overfitting ทำได้โดยการเพิ่มจำนวนตัวอย่างฝึกฝนให้ครอบคลุมข้อมูลทั้งหมด หรือมากที่สุดเท่าที่จะทำได้ หรือการลดมิติของข้อมูล ทางแก้ไขที่เป็นที่นิยมอีกทางหนึ่งคือ การแบ่งข้อมูลที่มีอยู่ออกเป็น 2 ส่วน ได้แก่ ชุดข้อมูลสำหรับฝึกฝน (Training Set) และชุดข้อมูลสำหรับตรวจสอบ (Validation Set) สำหรับต้นไม้การตัดสินใจการลดความซับซ้อนทำได้โดยการตัดทอนโหนด (Pruning) หรือลดจำนวนกฎลง

การตัดทอน (Pruning) ทำได้ 2 รูปแบบคือ การตัดทอนจำนวนโหนด และการลดจำนวนกฎ โดยการตัดทอนนั้นสามารถทำได้ 2 ลักษณะคือ

1) Pre-Pruning หรือการตัดทอนก่อนที่ต้นไม้จะสมบูรณ์ เมื่อพิจารณาแล้วว่าตัวอย่างที่มีไม่เพียงพอจะใช้ในการตัดสินใสที่น่าเชื่อถือได้

2) Post-Pruning หรือการตัดทอนภายหลังจากต้นไม้สมบูรณ์แล้ว

โดยทั่วไปแล้ววิธีการที่นิยมที่สุดและง่ายต่อการใช้งาน คือ วิธีการตรวจสอบไขว้ (Cross-Validation) โดยใช้การแบ่งกลุ่มข้อมูลที่มี (D) อยู่เป็นกลุ่มย่อย (Fold) จำนวน K กลุ่ม หรือเรียกว่า K-fold Cross-Validation

2.4.3.3 ขั้นตอนวิธี CART (Classification And Regression Tree)

เป็นขั้นตอนวิธีที่คิดค้นโดย Breiman ในปี ค.ศ. 1984 ซึ่งเป็นต้นไม้การตัดสินใจแบบไบนารี (Binary) ประกอบด้วย กิ่งหรือแขนง (Branch) 2 กิ่งในแต่ละโหนด (Node) โดยขั้นตอนวิธี CART จะทำการแบ่งเรคคอร์ดในชุดข้อมูลทดสอบ (Training Data Set) ออกเป็นเรคคอร์ดย่อยที่ให้ค่าเป้าหมาย (Target) ที่เหมือนกัน และใช้วิธีการจีนี (Gini method) เพื่อสร้างจุดแยก ประกอบด้วย Gini Index (Gini Impurity) และ Gini Gain โดย Gini Index คือ ความน่าจะเป็นของการกำหนดป้ายกำกับที่ไม่ถูกต้องให้กับตัวอย่างโดยการเลือกป้ายกำกับแบบสุ่ม และยังใช้เพื่อวัดความสำคัญของฟีเชอร์ (feature importance) ในต้นไม้ (Tree) การคำนวณหา Gini Index และ Gini Gain แสดงได้ดังสมการที่ 2.8 และ 2.9

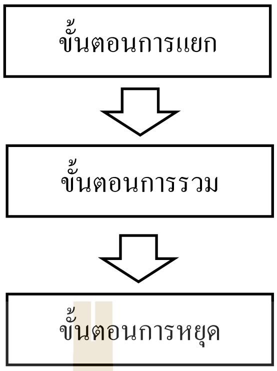
$$Gini = 1 - \sum_j p_j^2 \tag{2.8}$$

$$Gini\ Gain = \text{Gini(parent node)} - \text{Gini(children node)} \tag{2.9}$$

หลังจากคำนวณ Gini Gain สำหรับทุกแอตทริบิวต์แล้ว จะเลือกแอตทริบิวต์ที่มี Gini Gain มากที่สุดเป็นโหนดราก (Root Node) กิ่ง (Branch) ที่มี Gini เป็น 0 เป็นโหนดใบ (Leaf) ในขณะที่กิ่ง (Branch) ที่มี Gini มากกว่า 0 จำเป็นต้องแยกย่อยเพิ่มเติม โหนดจะเติบโตแบบเรียกซ้ำจนกว่าข้อมูลทั้งหมดจะถูกจัดประเภท โดยตัวแปรเอาท์พุต (Output Variable) เป็นตัวเลข (Numerical) และตัวแปรอินพุต (Input Variable) สามารถผสมระหว่างตัวแปรแบบต่อเนื่อง (Continuous variables) และแบบหมวดหมู่ (Categorical variables) ได้

2.4.3.4 ขั้นตอนวิธี CHAID (Chi-square Automatic Interaction Detection)

ขั้นตอนวิธี CHAID พัฒนาโดย Kass ในปี 1980 เป็นขั้นตอนวิธีใช้ในการจำแนกกลุ่มและหาความสัมพันธ์ของกลุ่มข้อมูล สามารถจำแนกกลุ่มโครงสร้างต้นได้ได้มากกว่า 2 โหนดขึ้นไป การสร้างแบบจำลองจะใช้ค่าการทดสอบความเป็นอิสระระหว่างตัวแปรจากสถิติไคสแควร์ในการหาค่า p-value เพื่อใช้ในการคัดเลือกตัวแปรในการจำแนกโหนด โดยมีขั้นตอนการทำงาน 3 ขั้นตอน คือ การแยก การรวม และการหยุด แสดงได้ดังภาพที่ 2.27 โดยในการใช้ขั้นตอนวิธี CHAID สามารถกำหนดพารามิเตอร์ต่าง ๆ เช่น ระดับนัยสำคัญในขั้นตอนแต่ละขั้นตอนต่างกันได้ และสามารถกำหนดความลึก และขนาดของโหนดได้ตามความต้องการ



ภาพที่ 2.27 ขั้นตอนการทำงานของขั้นตอนวิธี CHAID

1) ขั้นตอนการแยก เป็นการคัดเลือกตัวแปรในการทำนาย เพื่อใช้แยกโหนดโดยเปรียบเทียบค่า p-value และเลือกตัวแปรที่มีค่า p-value น้อยที่สุด เปรียบเทียบกับระดับนัยสำคัญ ถ้าค่า p-value น้อยกว่าหรือเท่ากับระดับนัยสำคัญให้แยกโหนดโดยใช้ตัวแปรในการทำนายนั้น แต่ถ้ามากกว่าให้กำหนดเป็นโหนดปลายทาง

2) ขั้นตอนการรวม จะทำการรวมตัวแปรในการทำนาย เพื่อลดจำนวนกลุ่มตัวแปร โดยพิจารณาจากค่า p-value

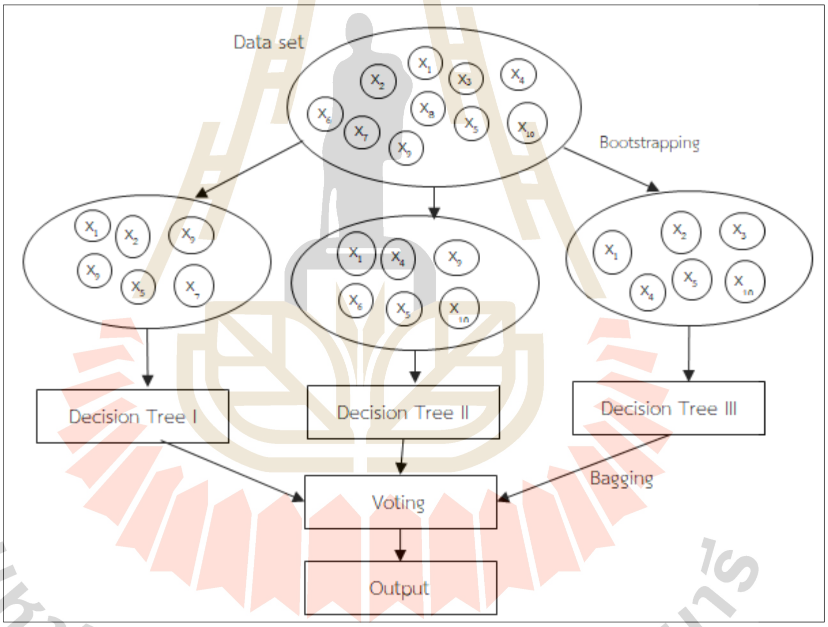
3) ขั้นตอนการหยุด จะตรวจสอบว่าการหยุดสร้างและการแยกโหนดจะหยุดเมือไร โดยมีเงื่อนไขดังนี้ ถ้าความลึกของการจำแนกกลุ่มถึงค่าความลึกมากสุด และ ถ้าขนาดของโหนดน้อยกว่าค่าขนาดของโหนดต่ำสุด

งานวิจัยนี้ใช้ขั้นตอนวิธี C4.5 เนื่องจาก ขั้นตอนวิธี C4.5 พัฒนามาจากขั้นตอนวิธี ID3 ที่เพิ่มประสิทธิภาพด้านการคำนวณ และจัดการข้อมูลที่เป็นข้อมูลค่าต่อเนื่อง (Continuous Values) และข้อมูลค่าสูญหาย (Missing Values) ได้ดี ในส่วนของขั้นตอนวิธี CART นั้นมีโครงสร้างเป็นแบบไบนารีเท่านั้น ทำให้แบ่งคลาสได้เพียง 2 คลาส และโหนดใบ (Leaf node) ในขั้นตอนวิธี CART ต้องเป็นข้อมูลค่าต่อเนื่อง (Continuous Values) เท่านั้น และขั้นตอนวิธี CHAID เป็นขั้นตอนวิธีที่ไม่สนับสนุนการการตัดทอนโหนด (Pruning) หรือลดจำนวนกฎลง ในขณะที่ขั้นตอนวิธี C4.5 มีความสามารถในการตัดทอนโหนด (Pruning) และจัดการกับปัญหาโอเวอร์ฟิตติง (Overfitting) ได้

2.4.4 การสุ่มป่าไม้ (Random Forest)

การสุ่มป่าไม้ (Random Forest: RF) เป็นขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนซึ่งใช้สำหรับปัญหาการจำแนกประเภท (Classification) และปัญหาการวิเคราะห์การถดถอย (Regression) แต่ส่วนใหญ่แล้ว จะใช้สำหรับปัญหาการจำแนกประเภท

ขั้นตอนวิธีการสุ่มป่าไม้ ใช้วิธีการสร้างตัวแบบต้นไม้การตัดสินใจ (Decision Trees) หลาย ๆ ตัวแบบ จากข้อมูลตัวอย่างจำนวนมาก โดยแต่ละตัวแบบจะได้รับข้อมูลตัวอย่างไม่เหมือนกัน นำมาประมวลผลในแต่ละต้นไม้ และเลือกต้นไม้ที่ดีที่สุด โดยการลงคะแนน (Voting) ผลลัพธ์การทำนาย หรือการหาค่าเฉลี่ยจากผลลัพธ์ของแต่ละต้นไม้การตัดสินใจ ลักษณะของต้นไม้การตัดสินใจที่อยู่ในป่าของขั้นตอนวิธีการสุ่มป่าไม้จะถูกควบคุมด้วย 3 ปัจจัย (ภูริพัทร์ ทองคำ, 2559) ดังแสดงในภาพที่ 2.28



ภาพที่ 2.28 หลักการทำงานของขั้นตอนวิธีการสุ่มป่าไม้

โดยมีรายละเอียดดังนี้

1) ต้นไม้แต่ละต้นจะถูกสอน (Train) โดยการใช้เซตย่อยจากข้อมูลตัวอย่าง (Bootstrapping) จากเซตข้อมูล (Data set) ทั้งหมด ได้ข้อมูลออกมา N ชุดที่ไม่เหมือนกัน ตามจำนวนต้นไม้การตัดสินใจ เช่น เซตข้อมูลตั้งต้นมีอยู่ 12 คุณลักษณะ (Feature) ($X_1, X_2, \ldots, X_{10}$) แต่ละต้นไม้การตัดสินใจจะได้คุณลักษณะไปไม่เหมือนกัน

มหาวิทยาลัยเทคโนโลยีสุรนารี

2) สร้างตัวแบบต้นไม้การตัดสินใจสำหรับแต่ละชุดข้อมูล เมื่อต้นไม้โตขึ้น จะสามารถค้นหาโหนด (Node) แต่ละโหนดที่อยู่ในกิ่งที่ดีที่สุดของต้นไม้โดยใช้การสุ่มเลือกคุณลักษณะจาก N คุณลักษณะ

3) ต้นไม้แต่ละต้นจะไม่มีการตัดออก แต่จะปล่อยให้ต้นไม้โตขึ้นไปเรื่อย ๆ จะได้ผลลัพธ์ที่ดีที่สุดโดยการลงคะแนน (Voting) ผลลัพธ์การทำนาย หรือการหาค่าเฉลี่ยจากผลลัพธ์ของแต่ละต้นไม้การตัดสินใจ

ขั้นตอนวิธีการสุ่มป่าไม้เป็นการสุ่มเลือกคุณสมบัติ (Feature) ต่าง ๆ ออกมาเป็นหลาย ๆ ชุด และสร้างตัวแบบต้นไม้การตัดสินใจหลาย ๆ ตัวแบบ ขึ้นอยู่กับการเก็บตัวแปรสุ่ม (Random Variables) โดยให้การแจกแจงแบบปกติหลายตัวแปรของเวกเตอร์สุ่ม p มิติ (p-dimensional random vector) $X = (X_1, \ldots, X_p)^T$ แทนค่าจริงที่นำเข้า หรือตัวแปรที่ทำนาย และตัวแปรสุ่ม Y แทนค่าจริงที่เป็นผลลัพธ์ โดยการแจกแจงร่วม (Join Distribution) ของ X และ Y คือ $P_{XY}(X,Y)$ เป้าหมายของคือการค้นหาฟังก์ชันการทำนาย (Prediction function) f(X) สำหรับการทำนายค่า Y ซึ่งฟังก์ชันการทำนายถูกกำหนดโดยฟังก์ชันการสูญเสีย (Loss function) L(Y, f(X)) และกำหนดเพื่อลดค่าความคาดหวังของการสูญเสีย (Cluter, 2012) ดังแสดงในสมการที่ 2.10

$$E_{xy}(L(Y, f(X))) \tag{2.10}$$

เมื่อ E_{XY} แสดงถึงการแจกแจงร่วมของ X และ Y โดยที่ $L(Y,f(X))$ คือการวัดความใกล้ระหว่าง $f(X)$ และ $\boldsymbol{Y}$

ขั้นตอนวิธีการสุ่มป่าไม้ใช้ต้นไม้ $h_j(X,$□□□$_j)$ เป็นฐานความรู้ (Based Learners) สำหรับการฝึกฝนข้อมูล (Training data) D = $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, เมื่อ $x_i = (x_i,1, \ldots, x_i,p)^T$ แสดงถึงตัวทำนาย p (p predictors) และ Y_i แสดงถึงผลลัพธ์

ในทางปฏิบัติการสุ่มองค์ประกอบของ □□$_j$ ใช้หลักในการสุ่ม 2 แบบคือ

1) การห่อหุ้ม (Bagging) โดยแต่ละต้นไม้จะขึ้นกับตัวอย่างบูตสแตรป (Bootstrap Sample) จากข้อมูลต้นฉบับซึ่งเป็นส่วนหนึ่งของ □□$_j$

2) การแยกโหนด (Splitting a node) เมื่อแยกโหนดพบการแยกที่ดีที่สุดผ่านชุดย่อยของตัวทำนาย m (m predictors) แทนตัวทำนาย p (p predictors) ทั้งหมดโดยแยกกันที่แต่ละโหนด การสุ่มใช้เพื่อเลือกตัวทำนายในส่วนที่เหลือของ □□$_j$

มหาวิทยาลัยเทคโนโลยีสุรนารี

ข้อดีของขั้นตอนวิธีการสุ่มป่าไม้

1) ขั้นตอนวิธีการสุ่มป่าไม้สามาถใช้ได้ทั้งกับปัญหาจัดหมวดหมู่ (Classification) และการวิเคราะห์การถดถอย (Regression)

2) ขั้นตอนวิธีการสุ่มป่าไม้ใช้ได้ทั้งกับข้อมูลแบบมีโครงสร้าง (Structured Data) เช่น ฐานข้อมูล และข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) เช่น รูปภาพ ข้อความ

เทคนิคการสุ่มป่าไม้เป็นเทคนิคการสุ่มเลือกใช้ข้อมูลและคุณลักษณะต้นไม้การตัดสินใจซึ่งถูกสร้างจากการนำข้อมูลไปสุ่มเลือกตัวอย่างแบบเลือกแล้วใส่กลับ (Sampling with Replacement) แล้วนำมาสร้างเป็นต้นไม้ ซึ่งจะมีตัวอย่างบางส่วนที่ไม่ถูกเลือก ข้อมูลส่วนนี้เรียกว่า Out-of-Bag (OOB) ซึ่งจะถูกนำมาใช้ในการทดสอบ วิธีการดังกล่าวเรียกว่า การห่อหุ้ม (Bagging) ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้การตัดสินในในแต่ละต้นจะถูกนำมาคิดเป็นผลการลงคะแนน ผลคะแนนที่มากที่สุดจะใช้ระบุสถานะของคลาส (พัฒนพงษ์ ดลรัตน์ และจารี ทองคำ, 2561) ขั้นตอนวิธีการสุ่มป่าไม้ไม่จำเป็นต้องมีข้อมูลทดสอบเพื่อประมาณความผิดพลาด เนื่องจากข้อมูล Out-of-Bag (OOB) นั้นถูกนำมาใช้ทดสอบต้นไม้การตัดสินใจแล้ว (Breiman, 2001; วัชรีวรรณ จิตต์สกุล, 2560)

2.4.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) จัดเป็นการเรียนรู้ ของเครื่องประเภทแบบการเรียนรู้โดยอาศัยตัวอย่างประเภทหนึ่ง ซึ่งมีความสามารถในการจัดหมวดหมู่ (Classification) และการวิเคราะห์การถดถอย (Regression) ซึ่งสามารถนำไปใช้กับงานต่าง ๆ (Vapnik, 1982) เช่นการรู้จำลายมือ การจำแนกวัตถุ และการจำแนกข้อความ โดยซัพพอร์ตเวกเตอร์แมชชีนจะมีการคำนวณแบบเชิงเส้น (Linear) ซึ่งจัดอยู่ในประเภทมุ่งหาผลลัพธ์ที่ดีที่สุดของการเรียนรู้ (Discriminative Training) บนการเรียนรู้จากสถิติของข้อมูล ซึ่งทำงานโดยการหาค่าระยะขอบที่มากที่สุด (Maximum Margin) ของระนาบตัดสินใจ (Decision Hyperplane) ในการแบ่งแยกกลุ่มข้อมูลที่ใช้ฝึกฝนออกจากกัน วิธีการนี้มีชื่อเรียกอีกชื่อหนึ่งว่า จัดหมวดหมู่โดยค่าระยะขอบที่มากที่สุด (Maximum Margin Classifier)

ซัพพอร์ตเวกเตอร์แมชชีนเป็นเครื่องมือที่ใช้ในการจำแนกประเภทที่นิยมมากในงานหลากหลาย เช่น การจัดหมวดหมู่เอกสาร การรู้จำใบหน้า โดยให้ประสิทธิภาพที่ดี เนื่องจากมีการจดจำข้อมูลที่อยู่บริเวณขอบของคลาสแต่ละคลาส อีกทั้งยังรองรับปริภูมิที่มีข้อมูลเป็นแบบเชิงเส้นและไม่เชิงเส้น โดยการอาศัยฟังก์ชันเคอร์เนล (Kernel) ในการลดความซับซ้อนของข้อมูลโดยการส่ง (Mapping) ไปยังปริภูมิพิเศษที่มีความเป็นเชิงเส้น ทั้งนี้ปริภูมิพิเศษดังกล่าวอาจสร้างได้จริงหรือเป็นเพียงนามธรรมก็ได้ ด้วยอาศัยกลวิธีเชิงเคอร์เนล (Kernel Trick) ภายในปริภูมิผลคูณภายใน (ปริญญา สงวนสัตย์, 2562)

หลักการทำงานของซัพพอร์ตเวกเตอร์แมชชีนข้อมูลจะถูกเขียนอยู่ในรูปสมาชิกคู่อันดับดังสมการที่ 2.11

$$\{(x_1,y_1),(x_2,y_2),(x_3,y_3),\dots,(x_n,y_n)\}$$
$$D = \{(x_i,y_i); i = 1,2,\dots,n\} \tag{2.11}$$

เมื่อ $x_i = (x_{i1},x_{i2},\dots,x_{im}) \in R^m$

y_i มีค่าเป็น 1 หรือ -1 โดย 1 คือ ข้อมูลกลุ่ม 1 และ -1 คือข้อมูลกลุ่ม 2

ซึ่งเป็นการกำหนดกลุ่มเป้าหมายให้ซัพพอร์ตเวกเตอร์แมชชีน โดยที่มุ่งเป้าเพื่อหาฟังก์ชันการตัดสินใจที่สามารถแบ่งแยกค่าที่ไม่ทราบได้ดังสมการที่ 2.12

$$f(x) = sign\left\{\sum_{k=1}^{n_v} w_k\varphi_k(x)\varphi_k(x_k) + b\right\} \tag{2.12}$$

$$\varphi(x) = [\varphi_1(x_1),\varphi_2(x_2),\dots,\varphi_n(x_{n_v})]^T \tag{2.13}$$

กลุ่มข้อมูล x จากสมการที่ 2.13 ไม่สามารถแบ่งแยกได้ด้วยสมการเส้นตรงแต่จะถูกแปลงให้อยู่ในรูปแบบที่สามารถใช้สมการเส้นตรงแบ่งแยกได้ โดยใช้เคอร์เนลฟังก์ชัน (kernel function) ดังสมการที่ 2.14

$$K(x,x_k) = \varphi(x)\varphi(x_k) \tag{2.14}$$

เมื่อ $\varphi(x)$ ฟังก์ชันสำหรับแปลงข้อมูลที่ไม่เป็นเชิงเส้นให้เป็นข้อมูลที่อยู่ในรูปแบบเชิงเส้นสามารถแบ่งแยกได้

w_k ค่าน้ำหนักที่เชื่อมโยงจาก Feature space ไปยัง Output space

b ค่าโน้มเอียง (Bias)

x_k ซัพพอร์ตเวกเตอร์โดย $k=1,2,\dots,n_v$

n_v จำนวนซัพพอร์ตเวกเตอร์

เส้นแบ่งกลุ่มข้อมูลที่ ถูกสร้างขึ้นซึ่งเป็นสมการเส้นตรงทั่วไปคือ $y = mx+b$ โดยในที่นี้จะแทน m ด้วย w^T เพื่อกำหนดเป็นสมการระนาบเกินในรูปแบบเวกเตอร์ ดังสมการที่ 2.15

มหาวิทยาลัยเทคโนโลยีสุรนารี

$$w^T x + b = 0 \qquad (2.15)$$

เมื่อ w^T คือ เวกเตอร์ปกติซึ่งตั้งฉากกับระนาบเกิน

b คือ ค่าโน้มเอียง (Bias) หรือระยะเลื่อนขนานจากจุดกำเนิด

วิธีการที่ใช้ในการหาเส้นแบ่งที่ดีที่สุดคือการเพิ่มเส้นขอบ (margin) ให้กับเส้นแบ่งทั้งสองข้างและสร้างเส้นขอบที่สัมผัสกับค่าของข้อมูลใน Feature space ที่ใกล้ที่สุด ดังนั้นเส้นแบ่งที่มีเส้นขอบกว้างที่สุดจึงเป็นเส้นแบ่งที่ดีที่สุด และเรียกตำแหน่งการสัมผัสข้อมูลที่ใกล้ที่สุดจากการเพิ่มขอบนี้ว่า “ซัพพอร์ตเวกเตอร์” (Support Vector) เนื่องจากในบางกรณีการแบ่งแยกกลุ่มไม่สามารถทำได้ถูกต้องโดยสมบูรณ์ ดังนั้นจึงต้องมีการกำหนดตัวแปรสำหรับยอมรับค่าความผิดพลาดโดยการเพิ่มตัวแปร ξ (slack variable) ดังสมการที่ 2.16 และ 2.17 ดังนี้

$$w^T x + b \geq y - \xi_i \qquad (2.16)$$

$$w^T x + b \leq y + \xi_i \qquad (2.17)$$

จากการกำหนดค่า $\xi_i > 0$ ทำให้โครงสร้างของซัพพอร์ตเวกเตอร์แมชชีนสามารถเพิ่มระยะแบ่งแยกให้ได้มากที่สุด และลดข้อผิดพลาดในการทำนายให้ต่ำที่สุดได้ ดังสมการที่ 2.18

$$Minimize \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N}\xi_i \qquad (2.18)$$

โดยที่ $y_i(w^T\varphi(x) + b) + \xi_i - 1 \geq 0$

$\xi_i \geq 0, i = 1,2, \dots, N$

แนวความคิดของขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีนเป็นขั้นตอนวิธีที่มีความรวดเร็ว และเป็นเทคนิคที่สามารถนำมาช่วยแก้ปัญหาในการจำแนกข้อมูล โดยอาศัยหลักของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้โดยเน้นไปยังเส้นแบ่งแยกแยะกลุ่มข้อมูลที่แบ่งได้ดีที่สุด โดยข้อมูลที่ถูกป้อนเข้าสู่กระบวนการส่วนใหญ่เป็นข้อมูลที่ไม่เป็นเชิงเส้น ซึ่งสามารถแก้ปัญหาดังกล่าวด้วยการนำเคอร์เนลฟังก์ชัน (Kemal Function) มาใช้ในการจำแนกข้อมูลบนระนาบหลายมิติ โดยนิยมใช้อยู่ 3 ชนิดด้วยกันคือ

มหาวิทยาลัยเทคโนโลยีสุรนารี

โพลิโนเมียล (Polynomial) ดังสมการที่ 2.19

$$K(x_i, x_j) = (x_i^T x_j + r)^\gamma \,; \gamma > 0 \tag{2.19}$$

เรเดียลเบสิสฟังก์ชัน (Radial Basis Function : RBF) ดังสมการที่ 2.20

$$K(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right) ; \gamma > 0 \tag{2.20}$$

ซิกมอยด์ (Sigmoid) ดังสมการที่ 2.21

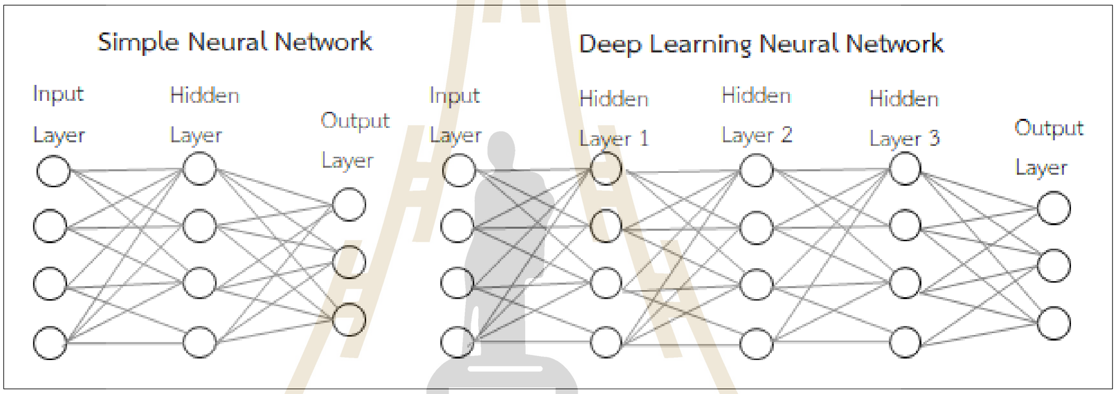
$$K(x_i, x_j) = \tanh\left(\gamma x_i^T x_j - r\right) \tag{2.21}$$

การแก้ปัญหาด้วยเคอร์เนลฟังก์ชันในการจำแนกข้อมูลบนระนาบหลายมิติจะใช้ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า โครงสร้างในการคัดเลือก ซึ่งโครงสร้างในการคัดเลือกมาจากข้อมูลที่สอนให้ระบบเรียนรู้จำนวนเซตของโครงสร้างที่ใช้อธิบายในกรณีใดกรณีหนึ่ง เรียกว่า เวกเตอร์ ดังนั้นจุดมุ่งหมายของขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน คือแบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วยหนึ่งกลุ่มของตัวแปรของเป้าหมายที่อยู่ข้างหนึ่งของระนาบ และกรณีของกลุ่มอื่นที่อยู่ระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า ซัพพอร์ตเวกเตอร์ ซึ่งวิธีการนี้เหมาะสำหรับข้อมูลที่มีมิติของข้อมูลสูง (พัฒนพงษ์ ดลรัตน์ และจารี ทองคำ, 2561)

2.4.6 การเรียนรู้เชิงลึก (Deep Learning)

การเรียนรู้เชิงลึก เป็นวิธีการหนึ่งของการเรียนรู้ของเครื่องที่เรียนรู้วิธีการแทนข้อมูลอย่างมีประสิทธิภาพ โดยมีรูปแบบการคำนวณที่ประกอบด้วยการประมวลผลหลายชั้นเพื่อเรียนรู้ในหลายระดับ ตัวอย่างการแทนข้อมูลอย่างมีประสิทธิภาพ เช่น รูปภาพหนึ่งภาพ สามารถแทนได้เป็นเวกเตอร์ของความสว่างต่อจุดพิกเซล หรือมองในระดับสูงขึ้นเป็นเซ็ตของขอบของวัตถุต่าง ๆ การแทนความหมายดังกล่าวจะทำให้การเรียนรู้ในด้านต่าง ๆ ทำได้ง่ายขึ้น เช่น ด้านการรู้จำใบหน้า (Face Recognition) การแสดงออกทางสีหน้า การเรียนรู้เชิงลึกถือว่าเป็นวิธีการที่มีศักยภาพสูงในการจัดการกับฟีเจอร์สำหรับการเรียนรู้แบบไม่มีผู้สอนหรือการเรียนรู้แบบกึ่งมีผู้สอน (นันทิพัฒน์ พลบดี, 2561)

เทคนิคการเรียนรู้เชิงลึกพัฒนามาจากความรู้ด้านโครงข่ายประสาทเทียม (Artificial Neuron Networks: ANN) โดยโครงข่ายประสาทเทียมจะอาศัยเทคนิคการทำงานของระบบโครงข่ายใยประสาทในระบบประสาทของมนุษย์ โดยจำลองการทำงานเหมือนกับกลุ่มเซลล์ประสาทที่เชื่อมโยงกันเป็นระบบประสาทที่สามารถรับรู้หลาย ๆ สิ่งในเวลาเดียวกันด้วยการประมวลผลแบบขนาน (Parallel Network) ช่วยให้ระบบสามารถตัดสินใจได้ใกล้เคียงกับมนุษย์ ทั้งการเรียนรู้เชิงลึก และโครงข่ายประสาทเทียมถูกสร้างขึ้นมาเพื่อการเรียนรู้ของเครื่อง โดยการเรียนรู้เชิงลึกจะเกิดจากโครงข่ายประสาทเทียมหลายชั้นเรียนรู้ที่จะสร้างข้อมูลตัวแทนขึ้นมาเพื่อใช้ควบคู่กับขั้นตอนวิธีต่าง ๆ ดังแสดงในภาพที่ 2.29



ภาพที่ 2.29 ความแตกต่างของโครงข่ายประสาทเทียมและการเรียนรู้เชิงลึก

ในการเรียนรู้เชิงลึก กระบวนการสกัดข้อมูลตัวแทน (Feature Extraction) จะรวมอยู่ในกระบวนการจัดหมวดหมู่ (Classification) ซึ่งในการประมวลผลภาษาธรรมชาติจะใช้วิธีการสร้างตัวแทนข้อความในรูปแบบต่าง ๆ เช่น แบบจำลองถุงคำ (Bag-of-Words) แบบจำลองทีเอฟไอดีเอฟ (Term Frequency – Inverse Document Frequency model: TF-IDF model) หรือแบบจำลองเอ็นแกรม (N-Gram model) ซึ่งการสร้างเวกเตอร์ในลักษณะนี้ทำให้เกิดข้อจำกัดขึ้น คือ 1) แบบจำลองเหล่านี้ไม่สนใจลำดับของคำในข้อความทั้งข้อความให้สูญเสียบริบทของข้อความส่งผลให้ตัวแทนข้อความที่ได้นั้นยากต่อการวิเคราะห์ความหมายที่แท้จริง 2) ถ้าเอกสารทั้งหมดที่ใช้เป็นชุดข้อมูลสอน (Training Set) มีจำนวนคำในคลังศัพท์ (Vocabulary) มาก เวกเตอร์ของตัวแทนข้อความจะมีขนาดใหญ่มาก ทำให้ใช้พื้นที่หน่วยความจำมาก (นันทิพัฒน์ พลบดี, 2561) แต่สำหรับการเรียนรู้เชิงลึก ตัวแทนข้อความจะเกิดจากการส่งข้อความเข้าไปในโครงข่ายประสาทเทียม และให้โครงข่ายนั้นเรียนรู้เองว่าจะสร้างเวกเตอร์ตัวแทนข้อความอย่างไรให้สะท้อนข้อความนั้นได้มากที่สุด หากตัวแทนข้อความนั้นสามารถจำแนกออกเป็นหลายระดับได้ ก็สามารถเพิ่มชั้นของหน่วยย่อยของโครงข่ายประสาทเทียมเพื่อให้การเรียนรู้เชิงลึกทำได้ลึกขึ้น

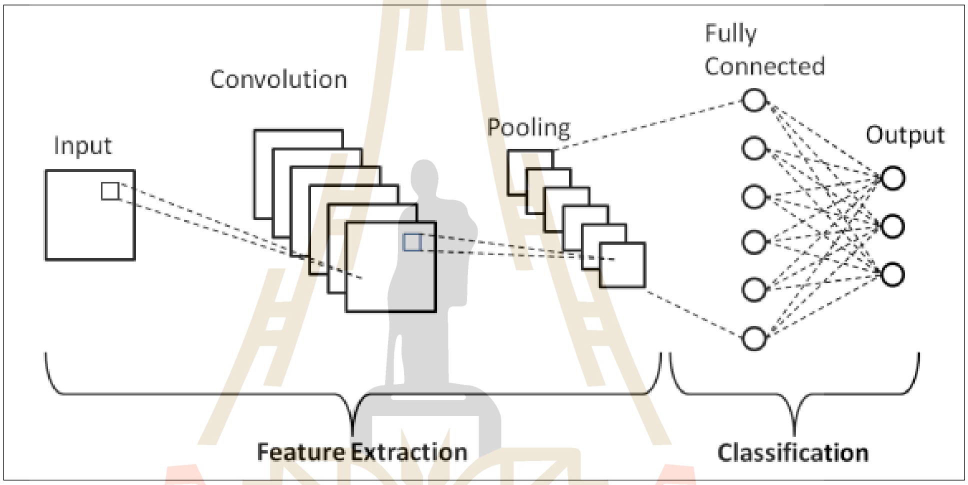
การเรียนรู้เชิงลึกประกอบด้วย ชั้นนำเข้า ชั้นซ่อนตัว และชั้นข้อมูลออก โดยมีชั้นซ่อนตัวมากกว่า 1 ชั้น ซึ่งการที่เครื่องคอมพิวเตอร์จะสามารถเข้าใจสิ่งต่าง ๆ ได้จำเป็นต้องมีองค์ความรู้ (Knowledge) ซึ่งทำหน้าที่เป็นชั้นนำเข้า (Input layer) จากนั้นก็จะประเมินชุดข้อมูลในชั้นซ่อนตัว (Hidden layer) และนำเสนอหรือแทนองค์ความรู้นั้นในชั้นข้อมูลออก (Output layer)

การเรียนรู้เชิงลึกสามารถแบ่งสถาปัตยกรรมการเรียนรู้ บนหลักการของการเรียนรู้เชิงลึกได้ หลายรูปแบบ เช่น โครงข่ายประสาทเทียมแบบลึก (Deep Artificial Neural Networks: DNN) โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า (Feed-forward neural networks) หรือเพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks: CNN) โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network: RNN) และโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ซึ่งมีการนำมาใช้งานอย่างแพร่หลายในทางคอมพิวเตอร์วิทัศน์ การรู้จำเสียงพูด การประมวลผลภาษาธรรมชาติ การรู้จำเสียง และชีวสารสนเทศศาสตร์ โดยมีรายละเอียดในแต่ละแบบดังนี้

1) โครงข่ายประสาทเทียมแบบลึก (Deep Artificial Neural Networks: DNN) เป็นเทคนิคการเรียนรู้ของเครื่องชนิดหนึ่งโดยแนวคิดของโครงข่ายประสาทเทียมนั้นมีแบบอย่างมาจากการทำงานของสมองมนุษย์ ซึ่งโครงข่ายประสาทเทียมแบบลึกก็คือโครงข่ายประสาทเทียมซึ่งมีชั้นระหว่างชั้นนำเข้า (Input Layer) และชั้นข้อมูลออก (Output Layer) จำนวนหลายชั้น โครงข่ายประสาทเทียมแบบลึกใช้สมการทางคณิตศาสตร์ในการเปลี่ยนข้อมูลเข้าให้เป็นข้อมูลออกไม่ว่าจะเป็นความสัมพันธ์เชิงเส้น (Linear relationship) หรือที่ไม่ใช่ความสัมพันธ์เชิงเส้น (Non-liner relationship) โดยเส้นเครือข่าย (Network) จะเคลื่อนที่ผ่านชั้นข้อมูลเพื่อคำนวณความน่าจะเป็นของแต่ละข้อมูลออก โครงข่ายประสาทเทียมแบบลึกสามารถสร้างแบบจำลองความสัมพันธ์ที่ไม่ใช่เชิงเส้นที่ซับซ้อน โดยชั้นข้อมูลพิเศษ หรือชั้นข้อมูลซ่อน (Extra Layer/Hidden Layer) ช่วยให้สามารถจัดองค์ประกอบของคุณสมบัติจากชั้นข้อมูลที่ต่ำกว่า ซึ่งอาจสร้างแบบจำลองข้อมูลที่ซับซ้อนได้ดี โดยโครงข่ายประสาทเทียมแบบลึกจะเป็นโครงข่ายประสาทเทียมแบบป้อนไปหน้า (Feed-forward Neural Networks) ซึ่งข้อมูลจะเคลื่อนที่จากชั้นข้อมูลเข้าไปยังชั้นข้อมูลออก โดยส่งผ่านชั้นข้อมูลซ่อนโดยไม่มีการย้อนกลับ

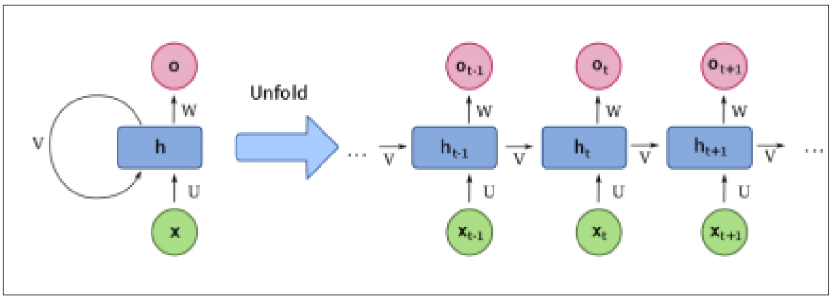
2) โครงข่ายประสาทเทียมแบบสังวัตนาการ (Convolutional Neural Networks: CNN) คือโครงข่ายประสาทเทียมหลายชั้นถูกออกแบบมาเพื่อเพิ่มความสามารถในการสกัดคุณลักษณะ (Feature) ที่มีความซับซ้อนมากยิ่งขึ้นจากข้อมูล และเป็นรูปแบบการเรียนรู้เชิงลึกที่ได้รับความนิยมในการจำแนกรูปภาพ (Image Classification) และการรู้จำรูปภาพ (Image Recognition) ดังแสดงในภาพที่ 2.30 โดยโครงข่ายประสาทเทียบแบบสังวัฒนาการนั้นตอบโจทย์

ปัญหาประเภทการรับรู้ (Perceptual Tasks) อย่างมาก ซึ่งสามารถประมวลผลข้อมูลที่อยู่ในรูปแบบของอาร์เรย์หลายชั้น (Multiple arrays) การใช้เพื่อการสกัดคุณลักษณะจากข้อมูลประเภทที่ไม่เป็นระเบียบหรือไม่มีโครงสร้าง (Unstructured data) เช่น ในรูปแบบอาร์เรย์ 1 มิติ (1 Dimension) สำหรับข้อมูลที่มีลักษณะเป็นสัญญาณ ลำดับ รวมถึงภาษาต่าง ๆ ในรูปแบบอาร์เรย์ 2 มิติ (2 Dimension) สำหรับข้อมูลที่เป็นรูปภาพหรือข้อมูลเสีย และในรูปแบบอาร์เรย์ 3 มิติ (3 Dimension) สำหรับข้อมูลที่เป็นวีดีโอ โดยมีแนวคิดหลัก 4 ประการสำหรับโครงข่ายประสาทเทียมแบบสังวัตนา (LeCun et al., 2015) การประกอบด้วย การเชื่อมต่อ (Local connections) ค่าน้ำหนักร่วมกัน (Shared weights) พูลลิ่ง (Pooling) และการใช้งานในหลายชั้น (Many Layers)



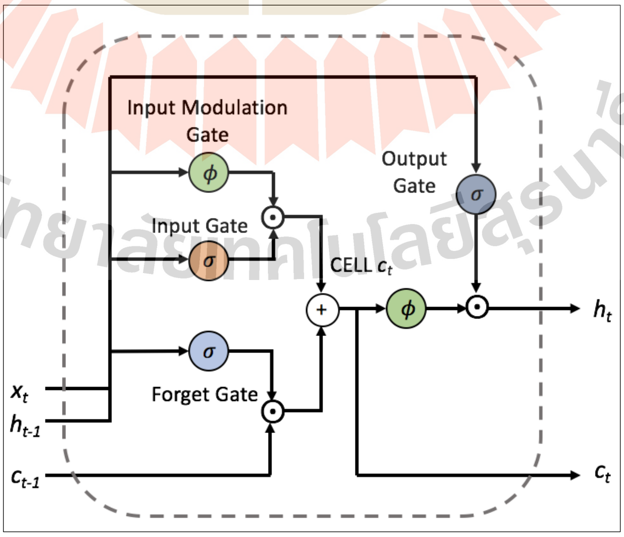
ภาพที่ 2.30 หลักการทำงานของโครงข่ายประสาทเทียมแบบสังวัตนาการ

3) โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network: RNN) คือโครงข่ายประสาทเทียมแบบหลายชั้นที่สามารถเก็บข้อมูล (Store information) ไว้ที่โหนด (Node) จึงทำให้สามารถรับข้อมูลเป็นลำดับ (Data Sequences) และให้ผลลัพธ์ออกเป็นลำดับของข้อมูลได้ โดยใช้โครงข่ายประสาทเทียมเชื่อมต่อกันหลาย ๆ อันและสามารถต่อกันเป็นวงวน (Loop) ดังแสดงในภาพที่ 2.31 จึงทำให้โครงข่ายประสาทเทียมแบบวนซ้ำเหมาะกับการประมวลผลข้อมูลที่เป็นลำดับอย่างมาก ซึ่งโครงข่ายประสาทเทียมแบบวนซ้ำ ออกแบบมาเพื่อแก้ปัญหาสำหรับงานที่ข้อมูลที่มีลำดับ โดยใช้หลักการนำสถานะภายในของโมเดล กลับมาเป็นข้อมูลเข้าใหม่คู่กับข้อมูลเข้าแบบปกติ เรียกว่า สถานะซ่อน (Hidden State) หรือสถานะภายใน (Internal State) ช่วยให้โมเดลรู้จำรูปแบบ (Pattern) ของลำดับข้อมูลเข้า (Input Sequence)



ภาพที่ 2.31 โครงสร้างของโครงข่ายประสาทเทียมแบบวนซ้ำ (ไกรศักดิ์ เกษร, 2565)

4) โครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ได้พัฒนามาจากโครงข่ายประสาทเทียมแบบวนซ้ำ ซึ่งได้แก้ปัญหาการสูญหายของเกรเดียน (Vanishing Gradient Problem: VGP) ซึ่งเป็นปัญหาหลักของโครงข่ายประสาทเทียมแบบวนซ้ำที่เกิดจากค่าเกรเดียนที่เริ่มน้อยลงในข้อมูลที่มีความยาวมากขึ้นเนื่องจาก โครงข่ายประสาทเทียมแบบวนซ้ำสามารถย้อนกลับได้เพียงช่วงระยะเวลาสั้น ๆ เท่านั้น โครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาวประกอบด้วยบล็อกหน่วยความจำ (Memory blocks) ต่าง ๆ ที่เรียกว่า เซลล์ (Cells) และเกจ (Gate) ที่เป็นเสมือนตัวควบคุมการไหลของข้อมูล ซึ่งประกอบด้วย ข้อมูลไหลเข้า (Input Gate) ข้อมูลไหลออก (Output Gate) และข้อมูลที่หายไป (Forget Gate) ดังแสดงในภาพที่ 2.32 ปัจจุบันโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว ได้ถูกนำมาใช้สำหรับการรู้จำเสียง การแต่งเพลง และการพัฒนาด้านเภสัชกรรม



ภาพที่ 2.32 โครงสร้างของโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว

มหาวิทยาลัยเทคโนโลยีสุรนารี

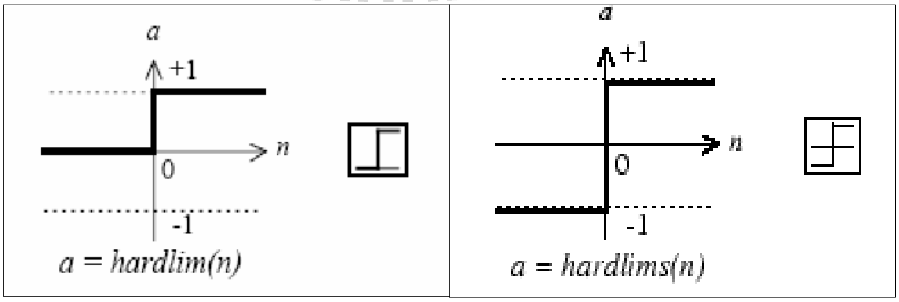
การเรียนรู้เชิงลึกถูกนำมาประยุกต์ใช้ในงานวิจัยต่าง ๆ มากมาย เช่น การศึกษาเปรียบเทียบระหว่างวิธีการหาคุณลักษณะพิเศษเฉพาะพื้นที่ และโครงข่ายประสาทเทียมแบบคอนโวลูชัน (CNN) สำหรับการค้นคืนรูปภาพลายผ้าไหมไทย (นัธทวัฒน์ รักสะอาด และคณะ, 2560) การประยุกต์ใช้การเรียนรู้เชิงลึกเพื่อจำแนกภาพถ่ายพื้นที่ด้านเกษตรกรรมจากข้อมูลการสำรวจระยะไกล (วุฒิชัย บุญพุก และคณะ, 2560) การตรวจจับภาวะซึมเศร้าโดยใช้การเรียนรู้เชิงลึก เพื่อเพิ่มประสิทธิภาพในการจัดหมวดหมู่ของข้อความที่แสดงความคิดเห็นผ่านเครือข่ายสังคมออนไลน์ โดยเปรียบเทียบ โครงข่ายประสาทเทียมแบบสังวัตนาการ โครงข่ายประสาทเทียมแบบวนซ้ำ และโครงสร้างของโครงข่ายประสาทเทียมแบบหน่วยความจำระยะสั้นแบบยาว (Orabi et al., 2018) การตรวจสอบความผิดปกติของอารมณ์ โดยใช้โครงข่ายประสาทแบบคอนโวลูชัน (Convolutional neural network long short-term memory: CNN-LSTM) (Sun, Zhang, and Li, 2019)

2.4.7 เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron)

เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) เป็นคลาสหนึ่งของโครงข่ายประสาทเทียมแบบไปข้างหน้า (Feedforward Artificial Neural Network: ANN) ซึ่งเป็นพื้นฐานของการเรียนรู้เชิงลึก (Deep Learning) โครงข่ายประสาทเทียม (Artificial Neural Network) ใช้หลักการเบื้องต้นมาจากการจำลองการทำงานของเซลล์ประสาท (Neuron) ของมนุษย์ โดยแต่ละเซลล์ประสาทจะมีหลักการสื่อสารกันผ่านการกระตุ้นด้วยศักย์ไฟฟ้า

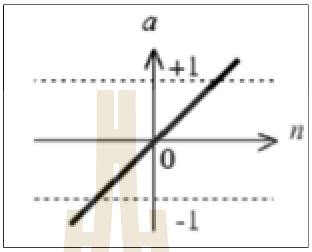
แบบจำลองเซลล์ประสาทเทียมประกอบด้วย เซลล์ประสาทหรือนิวรอน (Neuron) โหนด (Node) และการเชื่อมต่อไซแนปส์ (Synapse) ด้วยเส้นเชื่อม (Edge) และมีการกำกับด้วยค่าน้ำหนัก (Weight)

โครงสร้างเพอร์เซ็ปตรอน (Perceptron) มีลักษณะคล้ายแบบจำลองเซลล์ประสาทเทียมแต่มีการเปลี่ยนฟังก์ชันกระตุ้น (Activation Function) จากเดิมที่ใช้ฟังก์ชันฮาร์ดลิมิต (Hard-Limit) เป็นฟังก์ชันฮาร์ดลิมิตแบบสมมาตร (Symmetric Hard-Limit) ดังแสดงในภาพที่ 2.33



ภาพที่ 2.33 ฟังก์ชันฮาร์ดลิมิต และฟังก์ชันฮาร์ดลิมิตแบบสมมาตร (University of Toronto, 2007)

โครงสร้างนิวรอนแบบหน่วยเชิงเส้น (Linear Unit) มีลักษณะเหมือนเพอร์เซปตรอน แต่แตกต่างกันที่มีการเปลี่ยนฟังก์ชันการกระตุ้น (Activation Function) จากเดิมที่ใช้ฟังก์ชันฮาร์ดลิมิตแบบสมมาตร (Symmetric Hard-Limit) เป็นฟังก์ชันเชิงเส้น (Linear) ดังแสดงในภาพที่ 2.34



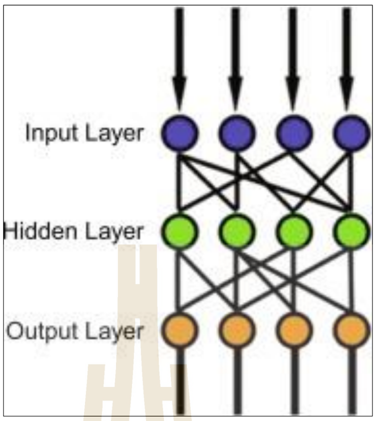
ภาพที่ 2.34 ฟังก์ชันเชิงเส้น (Linear) (Prathana Buranajun et al., 2007)

การฝึกนิวรอนเชิงเส้นจะใช้ขั้นตอนวิธี Gradient Descent ในการคำนวณค่าน้ำหนักสำหรับโครงข่ายประสาทเทียมเนื่องจากค่าน้ำหนักเป็นส่วนที่สำคัญที่สุด หากค่าน้ำหนักมีค่าเหมาะสมโครงข่ายประสาทเทียมจะมีคุณภาพสูงตามไปด้วย ขั้นตอนวิธี Gradient Descent เป็นการใช้ค่าเกรเดียนต์ของฟังก์ชัน $f(X)$ ในการกำหนดการเคลื่อนที่ของจุดจากจุดเริ่มต้นที่กำหนดขึ้นให้เป็น X_0 ไปสู่จุดต่ำสุดสัมพัทธ์ (Relative Minimum) จุดแรกที่พบจากจุดเริ่มต้นนั้น แสดงได้ดังสมการที่ 2.22

$$X_{n+1} = X_n - \eta \nabla f(X_n) \tag{2.22}$$

โดย η เป็นค่าน้อย ๆ ที่มีมากกว่าศูนย์ ซึ่งในเรื่องนี้นิยมเรียกว่า “อัตราการเรียนรู้” (Learning Rate) หรือ ขนาดขั้น (Step Size) และ $\eta \geq 0$ โดยค่าของฟังก์ชันของจุดต่อไปควรจะลดลงเสมอ หรือ $f(X_{n+1}) \leq f(X_n)$

แบบจำลองเพอร์เซ็ปตรอนหลายชั้นประกอบด้วยชั้น (Layers) 3 ประเภท คือ ชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นนำออก (Output Layer) ดังแสดงในภาพที่ 2.35



ภาพที่ 2.35 ชั้นของแบบจำลองเพอร์เซ็ปตรอนหลายชั้น (Abirami and Chitra, 2020)

ซึ่งชั้นซ่อน (Hidden Layer) จะเป็นส่วนที่สามารถเพิ่มจำนวนชั้นได้ โดยชั้นซ่อนจะทำหน้าที่คล้ายการสกัดคุณลักษณะ (Feature Extraction) ชั้นใหม่จากชั้นนำเข้า (Input Layer) ส่งต่อให้ชั้นซ่อน (Hidden Layer) ถัดไป หรือชั้นนำออก (Output Layer) ซึ่งในกระบวนการนี้สามารถสร้างคุณลักษณะใหม่ที่แบ่งแยกได้แบบเชิงเส้น โดยอาจมีคุณลักษณะที่สามารถแปลความหมายได้ เช่น ในการประมวลผลภาพคุณลักษณะในชั้นนี้อาจเป็นขอบภาพ (Edge) เป็นต้น นอกจากนั้นแล้วในชั้นซ่อน (Hidden Layer) นี้ซึ่งเป็นชั้นของการเรียนรู้ข้อมูล โดยจะสุ่มค่าน้ำหนัก (Weight) ของข้อมูลนั้น ๆ เพื่อเป็นการให้ความสำคัญของข้อมูลแต่ละข้อมูล และสุ่มค่าความโน้มเอียง (Bias) เพื่อเป็นตัวกำหนดทิศทางการเรียนรู้ของแบบจำลอง และหากมีชั้นซ่อนมากกว่า 1 ชั้น ชั้นซ่อนที่เหลือจะมีการปรับค่าน้ำหนัก (Weight) และค่าความโน้มเอียง (Bias) ในขณะเรียนรู้ให้ได้ค่าที่เหมาะสม ซึ่งการคำนวณในแต่ละชั้นซ่อนจะประกอบไปด้วยส่วนคำนวนเชิงเส้น และตามด้วยฟังก์ชันกระตุ้น (Activation Function) แสดงได้ดังสมการที่ 2.23 และ 2.24

$$a_i = h_{i-1}w_i + b_i^T \quad (2.23)$$

$$h_i = \emptyset(a_i) \quad (2.24)$$

โดย i คือ ชั้นแต่ละชั้น

w คือ ค่าน้ำหนัก (Weight) ของข้อมูลนั้น ๆ

b คือ ค่าความโน้มเอียง (Bias)

$\emptyset$ คือ ฟังก์ชันกระตุ้น (Activation Function)

การฝึกฝนเซลล์ประสาทในเพอร์เซ็ปตรอนหลายชั้นใช้ขั้นตอนวิธีการแพร่กลับของค่าความผิดพลาด (Backpropagation) เนื่องจากการฝึกฝนด้วยขั้นตอนวิธี Gradient Descent ต้องใช้ค่าความผิดพลาดเป็นหลักในการเปลี่ยนแปลงค่า ในกรณีโครงข่ายประสาทเทียมแบบหลายชั้น จึงจำเป็นต้องมีการส่งผ่านค่าความผิดพลาดจากชั้นนำออก (Output Layer) กลับมาชั้นก่อนหน้าเพื่อสังเกตค่าผิดพลาดในแต่ละชั้น

ความถูกต้องของเพอร์เซ็ปตรอนหลายชั้นขึ้นอยู่กับการกำหนดพารามิเตอร์ต่าง ๆ เช่น จำนวนชั้นซ่อน (Hidden Layer) จำนวนนิวรอนในแต่ละชั้น การกำหนดอัตราการเรียนรู้ การกำหนดค่าโมเมนตัม รวมไปถึงการสุ่มเพื่อกำหนดค่าเริ่มต้นของน้ำหนักในแต่ละครั้งที่ไม่เหมือนกัน ทำให้การประมวลผลแต่ละครั้งได้ผลลัพธ์ไม่เหมือนกัน

เพอร์เซ็ปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) หรือ โครงข่ายประสาทเทียมแบบป้อนไปข้างหน้า (Deep Feedforward Neural Network) เป็นรากฐานของแบบจำลองการเรียนรู้เชิงลึก (Deep Learning Model) (Goodfellow, Bengio, and Courville, 2016)

2.5 การประเมินผลแบบจำลอง

2.5.1 เทคนิคการประเมินแบบจำลอง (Model Evaluation Techniques)

การประเมินผลแบบจำลองเพื่อตรวจสอบว่าแบบจำลองมีความน่าเชื่อถือ และสามารถใช้งานได้จริงหรือไม่ ในหัวข้อนี้จึงอธิบายถึงเทคนิคต่าง ๆ ที่ใช้ในการประเมินว่าแบบจำลองที่ได้นั้นมีประสิทธิภาพมากน้อยเพียงใด การประเมินแบบจำลองมีวัตถุประสงค์เพื่อประเมินความถูกต้องของแบบจำลองที่ใช้กับข้อมูลที่ไม่เคยเห็นมาก่อน (unseen/out-of-sample data)

วิธีการประเมินแบบจำลองประกอบด้วย 2 วิธี คือ Holdout และ Cross-validation โดยทั้งสองวิธีใช้ข้อมูลทดสอบที่เป็นข้อมูลที่ไม่เคยเห็นมาก่อน (Test set) เพื่อใช้ในการประเมินประสิทธิภาพแบบจำลอง

2.5.1.1 โฮลด์เอาท์ (Holdout) วัตถุประสงค์ของการประเมินผลแบบโฮลด์เอาท์เพื่อทดสอบแบบจำลองในข้อมูลที่แตกต่างจากที่ได้รับการฝึกฝน โดยวิธีนี้ข้อมูลจะถูกแบ่งออกเป็น 3 ส่วนย่อย คือ

1) ชุดข้อมูลเรียนรู้ (Training set) เป็นชุดข้อมูลย่อยของชุดข้อมูล (Data set) ที่ใช้ในการสร้างแบบจำลอง

2) ชุดการตรวจสอบ (Validation set) เป็นชุดข้อมูลย่อยของชุดข้อมูลที่ใช้ในการประเมินประสิทธิภาพของตัวแบบที่สร้างขึ้นในขั้นตอนการฝึกอบรม ใช้สำหรับการปรับพารามิเตอร์ของแบบจำลอง ซึ่งไม่ใช่ทุกแบบจำลองที่ต้องการ Validation set

3) ชุดข้อมูลทดสอบ (Test set) หรือ ชุดข้อมูลใหม่ (unseen data) ซึ่งเป็นชุดข้อมูลที่ไม่เคยเห็นมาก่อน ใช้ในการประเมินประสิทธิภาพแบบจำลอง ถ้าแบบจำลองที่เหมาะกับชุดข้อมูลเรียนรู้ (Training set) มากกว่าชุดข้อมูลทดสอบ (Test set) อาจจะเป็นสาเหตุของการเกิด Overfitting

การประเมินแบบจำลองด้วยวิธี โฮลด์เอาท์เป็นวิธีการที่มีความเร็ว เรียบง่าย และมีความยืดหยุ่นสูง อย่างไรก็ตามเทคนิคนี้มักจะเกี่ยวข้องกับความแปรปรวนสูง (high variability) เนื่องจากความแตกต่างในชุดข้อมูลการฝึกอบรมและการทดสอบอาจส่งผลให้เกิดความแตกต่างที่มีความหมายในการประเมินความถูกต้อง

2.5.1.2 การตรวจสอบแบบไขว้ (Cross-Validation) เป็นเทคนิคที่เกี่ยวข้องกับการแบ่งชุดข้อมูลเพื่อใช้ในการฝึกฝน (Train) แบบจำลอง ซึ่งเทคนิคที่พบมากที่สุดคือ *k*-fold cross-validation โดยชุดข้อมูลดั้งเดิมถูกแบ่งเป็น *k* และมีขนาดย่อย ๆ เท่ากันเรียกว่า folds โดย *k* คือ เลขที่ผู้ใช้ระบุส่วนใหญ่จะเป็นเลข 5 หรือ 10 ตามที่ผู้ใช้ต้องการ กระบวนการคือการทำซ้ำ *k* ครั้ง ซึ่งแต่ละครั้งจะใช้ชุดข้อมูลย่อย *k* ชุด โดยใช้ชุดใดชุดหนึ่งเป็นชุดทดสอบ หรือตรวจสอบความถูกต้อง และชุดข้อมูลย่อย *k*-1 อื่น ๆ จะถูกใช้สำหรับการฝึกฝนแบบจำลอง การคำนวณประสิทธิภาพโดยรวมของแบบจำลองทำได้โดยการหาค่าเฉลี่ยข้อผิดพลาดที่เกิดขึ้นทั้งหมด *k* ครั้ง ตัวอย่างเช่น หากใช้การประเมินด้วย 5-fold cross-validation ชุดข้อมูลจะถูกแบ่งเป็น 5 ชุดข้อมูลย่อย ในรอบแรกชุดข้อมูลย่อยที่ 1 จะใช้เป็นชุดข้อมูลทดสอบ (Test set) ที่เหลือจะเป็นชุดข้อมูลเรียนรู้ (Training set) และจะทำวนไปจนครบ 5 รอบ แล้วหาค่าเฉลี่ย โดยข้อมูลที่อยู่ในชุดข้อมูลทดสอบ (Test set) จะถูกนำไปใช้ในชุดข้อมูลเรียนรู้ (Training set) จำนวน *k*-1 ครั้ง ซึ่งสิ่งนี้จะช่วยลดอคติ (Bias) อย่างมีนัยสำคัญและยังช่วยลดความแปรปรวนได้เนื่องจากชุดข้อมูลส่วนใหญ่ยังถูกทำไปใช้ในชุดข้อมูลทดสอบ (Test set) โดยในงานวิจัยนี้ได้ใช้การตรวจสอบแบบไขว้ โดยวิธีการไขว้ทบ 10 ส่วน (10-fold cross validation)

2.5.2 การประเมินแบบจำลองด้วยเมตริกซ์ (Model Evaluation Metrics)

การประเมินแบบจำลองด้วยเมตริกซ์ถูกนำมาใช้ในการวัดประสิทธิภาพของแบบจำลอง ซึ่งการเลือกใช้วิธีการประเมินนั้นขึ้นอยู่กับขั้นตอนวิธีที่ใช้ เช่น การใช้การเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised learning) สำหรับการจำแนกประเภท (Classification) นั้นส่วนใหญ่ใช้การประเมินแบบจำลองด้วยค่าความแม่นยำ (Precision) และค่าความระลึก (Recall)

ในงานวิจัยนี้ใช้การประเมินผลแบบจำลองด้วยการประเมินความถูกต้องของแบบจำลองโดยใช้วิธีวัดค่ามาตรฐานในการประเมินความถูกต้อง 4 ค่า ได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าประสิทธิภาพโดยรวม (F-measure) ซึ่งมีรายละเอียดดังนี้

ค่าความแม่นยำ (Precision) คือ อัตราส่วนระหว่าง จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม C_j และแบบจำลองทำนายว่าอยู่ในกลุ่ม C_j ต่อจำนวนกลุ่มตัวอย่างทั้งหมดที่แบบจำลองทำนายว่าอยู่ในกลุ่ม C_j ดังสมการที่ 2.25

$$Precision = \frac{TP}{(TP + FP)} \times 100 \tag{2.25}$$

ค่าความระลึก (Recall) คือ อัตราส่วนระหว่าง จำนวนกลุ่มตัวอย่างที่มีผลการทำนายว่าอยู่ในกลุ่ม C_j ต่อจำนวนกลุ่มตัวอย่างทั้งหมดที่อยู่ในกลุ่ม C_j ดังสมการที่ 2.26

$$Recall = \frac{TP}{(TP + FN)} \times 100 \tag{2.26}$$

ค่าความถูกต้อง (Accuracy) คือ อัตราส่วนระหว่าง จำนวนกลุ่มตัวอย่างที่มีผลการทำนายถูกต้อง ต่อจำนวนกลุ่มตัวอย่างทั้งหมด ดังสมการที่ 2.27

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \tag{2.27}$$

ค่าประสิทธิภาพโดยรวม (F-measure) คือ การเปรียบเทียบค่าความแม่นยำ และค่าความระลึกของแต่ละคลาสเป้าหมาย จึงเปรียบเสมือนค่าวัดความแม่นยำโดยรวม มีสูตรในการคำนวณดังสมการที่ 2.28

$$F - Measure = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)} \tag{2.28}$$

โดยที่

TP (True Positive) คือ จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม C_j และแบบจำลองทำนายว่าอยู่ในกลุ่ม C_j

TN (True Negative) คือ จำนวนกลุ่มตัวอย่างที่ไม่อยู่ในกลุ่ม C_j และแบบจำลองทำนายว่าไม่อยู่ในกลุ่ม C_j

FP (False Positive) คือ จำนวนกลุ่มตัวอย่างที่ไม่อยู่ในกลุ่ม C_j แต่แบบจำลองทำนายว่าอยู่ในกลุ่ม C_j

มหาวิทยาลัยเทคโนโลยีสุรนารี

FN (False Negative) คือ จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม C_j แต่แบบจำลองทำนายว่าไม่อยู่ในกลุ่ม C_j

C_j คือ กลุ่มของระดับคะแนนจากผลการประเมินภาวะซึมเศร้าที่ทดสอบด้วยแบบทดสอบภาวะซึมเศร้า 9 คำถาม (9Q) 4 ระดับคะแนน เมื่อ $0 \leq j \leq 3$

2.6 งานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ ได้ทบทวนงานวิจัยที่เกี่ยวข้องกับการวิจัยเรื่องการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยจำแนกงานวิจัยที่เกี่ยวข้องตามวัตถุประสงค์ออกเป็น 3 ส่วนได้แก่ การวิเคราะห์ความรู้สึกจากข้อมูลในเครือข่ายสังคมออนไลน์ การวิเคราะห์ภาวะซึมเศร้าจากเครือข่ายสังคมออนไลน์ และการพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า โดยมีรายละเอียดดังต่อไปนี้

2.6.1 การวิเคราะห์ความรู้สึกจากข้อมูลในเครือข่ายสังคมออนไลน์

การวิเคราะห์ความรู้สึกจากข้อมูลในเครือข่ายสังคมออนไลน์ มีวัตถุประสงค์เพื่อใช้การวิเคราะห์ความรู้สึกในกระบวนการต่าง ๆ จากข้อมูลในเครือข่ายสังคมออนไลน์ เช่นการตรวจสอบความรู้สึกของผู้ใช้งาน การตรวจสอบคำพูด ผลกระทบ และอารมณ์ และการประเมินภาวะซึมเศร้า โดยมีงานวิจัยที่มุ่งเน้นการพัฒนาขั้นตอนวิธีในการวิเคราะห์ความรู้สึกจากข้อมูลในเครือข่ายสังคมออนไลน์ที่แตกต่างกัน โดยมีรายละเอียดดังนี้

Barhan and Shakhomirov (2012) ได้พัฒนาแบบจำลองเพื่อจำแนกความรู้สึกจากข้อมูลในทวิตเตอร์ โดยใช้การสกัดคำ ประกอบด้วยเอ็นแกรม (n-Gram) เป็นทฤษฎีที่ใช้ในการคำนวณหาความน่าจะเป็นของคำต่าง ๆ โดยพิจารณาจากคำก่อนหน้า n-1 คำจากคำเป้าหมายและสัญรูปอารมณ์ (Emoticons) ผลการทดลองพบว่าขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพดีกว่า ขั้นตอนวิธีนาอีฟเบย์ โดยมีค่าความแม่นยำ 81% และค่าความระลึก 74% สอดคล้องกับงานวิจัยของ Hutto and Gilbert (2014) ได้นำเสนอวิธีการเปรียบเทียบประสิทธิภาพของขั้นตอนวิธีในการวิเคราะห์ความรู้สึกของข้อความในเครือข่ายสังคมออนไลน์ ประกอบด้วยขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน นาอีฟเบย์ และ เอนโทรปีสูงสุด (Maximum Entropy) โดยใช้ข้อมูลจากทวิตเตอร์จำนวน 4,000 ข้อความ ผลการวิจัยพบว่า ซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องสูงที่สุดคือ 91% ในขณะที่ Vateekul and Koomsubha (2016) ได้นำเสนอวิธีการใหม่ คือได้ประยุกต์ใช้การวิเคราะห์ความรู้สึกร่วมกับการเรียนรู้เชิงลึก (Deep learning) ในการจำแนกความรู้สึกของข้อมูลในทวิตเตอร์ของไทย ผลการวิจัยพบว่า เทคนิคการเรียนรู้เชิงลึกให้ผลลัพธ์ด้านความถูกต้องที่ดีกว่าเทคนิคดั้งเดิม เช่น นาอีฟย์เบย์ ซัพพอร์ตเวกเตอร์แมชชีน โดยมีค่าความถูกต้องอยู่ที่ 75% ส่วนงานวิจัยของ Sood, Hooda, Dhirn, and Bhatia (2018) ได้นำเสนอวิธีการในการ

วิเคราะห์ความรู้สึกเพื่อระบุภาวะซึมเศร้า โดยใช้โปรแกรม R Studio ในการดึงข้อมูลจากทวิตเตอร์และวิเคราะห์ความรู้สึกโดยระบุคะแนนความรู้สึกจากข้อความที่แสดงความคิดเห็นในทวิตเตอร์และระบุว่าบุคคลนั้นมีความสุขหรือไม่ ผลการศึกษาพบว่า การวิเคราะห์ความรู้สึกผ่านทวิตเตอร์ช่วยในการเพิ่มประสิทธิภาพในการกลั่นกรองความรู้สึกของประชาชนทั่วไป นอกจากนั้นแล้วงานวิจัยของ Islam et al. (2018) ได้ทำการวิเคราะห์ความรู้สึกของผู้ใช้เครือข่ายสังคมออนไลน์เพื่อตรวจสอบอารมณ์และทัศนคติเมื่อสื่อสารผ่านเครื่องมือออนไลน์เหล่านี้ โดยประยุกต์ใช้การวิเคราะห์ความรู้สึกร่วมกับการเรียนรู้ของเครื่อง ผลการวิจัยพบว่าเทคนิคการเรียนรู้ของเครื่องมีประสิทธิภาพสูงในการระบุวิธีแก้ปัญหาสุขภาพจิตในผู้ใช้เครือข่ายสังคมออนไลน์ ในขณะที่ Kashyap (2014) ได้ทำการวิจัยเรื่องการวิเคราะห์ความรู้สึกสำหรับการติดตามด้านสุขภาพ โดยใช้ข้อมูลจากผู้ใช้งานทวิตเตอร์ในการวิเคราะห์ข้อมูลด้านสุขภาพ โดยสามารถจำแนกกลุ่มของผู้รักสุขภาพ และกลุ่มผู้ที่ไม่ดูแลสุขภาพได้จากข้อความที่ผู้ใช้ทวิตเตอร์ รวมถึงสามารถวัดระดับคะแนนการดูแลสุขภาพได้ โดยมีค่าความถูกต้องอยู่ที่ 65%

2.6.2 การวิเคราะห์ภาวะซึมเศร้าจากข้อมูลในเครือข่ายสังคมออนไลน์

การวิเคราะห์โรคซึมเศร้าจากเครือข่ายสังคมออนไลน์ มีวัตถุประสงค์เพื่อสกัดความรู้ หรือทฤษฎีจากข้อมูลจำนวนมากที่อยู่ในเครือข่ายสังคมออนไลน์ เพื่อใช้ในการวิเคราะห์ภาวะซึมเศร้าจากเครือข่ายสังคมออนไลน์ โดยมีงานวิจัยที่มุ่งเน้นการพัฒนาขั้นตอนวิธีเพื่อสกัดองค์ความรู้จากข้อมูลจำนวนมากนั้นด้วยขั้นตอนวิธีที่แตกต่างกัน โดยมีรายละเอียดดังนี้

ในงานวิจัยของ Wang et al. (2013) ได้นำเสนอวิธีการวิเคราะห์ข้อมูลเพื่อตรวจสอบภาวะโรคซึมเศร้าของผู้ใช้งานเครือข่ายสังคมออนไลน์ โดยใช้การวิเคราะห์ความรู้สึกและใช้กลุ่มตัวอย่างจำนวน 180 คน ผลการทดลองพบว่า ค่าความแม่นยำที่ได้คือ 80% ส่วนงานวิจัยของ Park, Cha, and Cha (2012) ได้นำเสนอผลลัพธ์การสร้างเฟรมเวิร์คในการวิเคราะห์ความรู้สึกโดยผู้ใช้งานทวิตเตอร์ และค้นข้อความที่ระบุถึงภาวะโรคซึมเศร้า ในขั้นตอนแรกได้วิเคราะห์ข้อความคิดเห็นจากทวิตเตอร์ในระยะเวลา 2 เดือนเพื่อค้นหาว่ามีการพูดถึงโรคซึมเศร้าอย่างไรในทวิตเตอร์ ข้อความในทวิตเตอร์ส่วนใหญ่มีข้อมูลเกี่ยวกับความรู้สึกซึมเศร้า สถานะ ประวัติการรักษา จากการศึกษากลุ่มตัวอย่างจำนวน 69 คนพบว่า การใช้ข้อความด้านลบ ข้อความที่แสดงออกถึงความซึมเศร้า หรือสัญรูปอารมณ์ด้านลบผ่านทางทวิตเตอร์ที่เพิ่มขึ้นนั้นมีนัยสำคัญกับการเกิดอาการโรคซึมเศร้า อย่างไรก็ตามจากการศึกษาพบว่า ไม่มีความแตกต่างในการใช้งานสัญรูปอารมณ์ด้านบวกระหว่าง 2 กลุ่ม ในขณะที่งานวิจัยของ Choudhury, Gamon, Counts, and Horvitz (2013) ได้นำเสนอการทำนายภาวะซึมเศร้าผ่านเครือข่ายสังคมออนไลน์ โดยการรวบรวมข้อมูลการใช้งานทวิตเตอร์จำนวน 476 คน พบว่าเทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ให้ผลลัพธ์ที่ดีที่สุด โดยให้ค่าความถูกต้อง (Accuracy) ร้อยละ 70 ส่วนงานวิจัยของ Jiang et al. (2017) ได้ศึกษาถึงความสัมพันธ์ทางเสียงของกลุ่มตัวอย่าง 170 คน (ผู้ป่วยโรคซึมเศร้า 85 คน)

โดยการตรวจสอบการจำแนกของการพูดใน 3 ประเภทที่แตกต่างกัน (การสัมภาษณ์ คำบรรยายภาพ และการอ่าน) และอารมณ์การพูดที่แตกต่างกัน 3 อารมณ์ (บวก กลาง และลบ) เพื่อตรวจจับความซึมเศร้า ผลการศึกษาพบว่าคำบรรยายภาพให้ผลการจำแนกที่ดีกว่าประเภทอื่น ๆ แสดงให้เห็นว่าประเภทของคำพูด และอารมณ์ เป็นวิธีการใหม่สำหรับการตรวจจับภาวะซึมเศร้า ในขณะที่ Orabi et al. (2018) ได้ประยุกต์ใช้การเรียนรู้เชิงลึกในการตรวจสอบภาวะซึมเศร้าของผู้ใช้งานทวิตเตอร์ เพื่อระบุถึงสถาปัตยกรรมเชิงลึกของการเรียนรู้เชิงลึก ที่มีประสิทธิภาพมากที่สุดในการประมวลผลภาษาธรรมชาติ เพื่อใช้ในการตรวจจับภาวะซึมเศร้าจากข้อมูลที่ไม่มีโครงสร้างในเครือข่ายสังคมออนไลน์ โดยแบ่งชุดข้อมูลเป็น 2 ชุดคือ ชุดข้อมูลจาก CLPsych และ ชุดข้อมูลจาก Bell Let's Talk และสร้างแบบจำลองด้วย CNN และ RNN ผลการทดลองพบว่า CNN ให้ประสิทธิภาพสูงกว่า RNN และสุดท้ายคืองานวิจัยของ Guntuku, Yaden, Kern, Ungar, and Eichstaedt (2017) ได้ศึกษางานวิจัยที่เกี่ยวข้องกับการวิเคราะห์และการทำนายภาวะสุขภาพจิตโดยใช้เครือข่ายสังคมออนไลน์ จากการศึกษาพบว่า อาการซึมเศร้าสามารถตรวจพบได้จากแหล่งออนไลน์ต่าง ๆ นอกจากนั้นแล้วความก้าวหน้าในการประมวลผลภาษาธรรมชาติ และการเรียนรู้ของเครื่องจะช่วยให้การคัดกร อง สื่อสังคมออนไลน์สำหรับบุคคลที่มีความเสี่ยงสามารถเป็นไปได้ในอนาคต

2.6.3 การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า

การสร้างแบบจำลองการวิเคราะห์ภาวะโรคซึมเศร้า มีวัตถุประสงค์เพื่อสร้างแบบจำลองเพื่อใช้ในการวิเคราะห์ภาวะโรคซึมเศร้า ซึ่งแบบจำลองนั้นมีกระบวนการในการสร้างและการทำงานที่แตกต่างกัน โดยมีรายละเอียดดังนี้

Aldarwish and Ahmad (2017) ได้ทำวิจัยเรื่องการพัฒนาแบบจำลองเพื่อทำนายระดับภาวะซึมเศร้าโดยใช้ข้อความแสดงความคิดเห็นจากเครือข่ายสังคมออนไลน์ โดยการทำเหมืองข้อมูล ซึ่งช่วยในการทำนายระดับสุขภาพจิต และภาวะซึมเศร้า โดยข้อมูลส่วนตัว (Profile) ในเครือข่ายสังคมออนไลน์สามารถบ่งบอกอารมณ์ หรือความรู้สึกด้านลบได้ ในงานวิจัยนี้ มีวัตถุประสงค์เพื่อตรวจสอบโพสต์ของผู้ใช้เครือข่ายสังคมออนไลน์ การจำแนกประเภทโดยใช้ปัญญาประดิษฐ์ในการวิเคราะห์โดยสร้างแบบจำลองของการจำแนกประเภทด้วยขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน และขั้นตอนวิธีนาอีฟเบย์ โดยมีค่าความถูกต้องอยู่ที่ 57% และ 63% ตามลำดับ ในขณะที่ Zhu, Li, Li, and Zhu (2016) ได้ทำวิจัยเรื่องการพัฒนาแบบจำลองเพื่อทำนายภาวะโรคซึมเศร้าจากพฤติกรรมการใช้อินเทอร์เน็ต งานวิจัยนี้มีวัตถุประสงค์เพื่อสร้างวิธีการใหม่ในการป้องกันตรวจสอบการเกิดโรคซึมเศร้าโดยใช้การวิเคราะห์ความถี่ของเวลาที่ใช้อินเทอร์เน็ต โดยการเก็บข้อมูลนักศึกษาจำนวน 728 คน และทำแบบทดสอบภาวะโรคซึมเศร้า และข้อมูลการใช้อินเทอร์เน็ต โดยสร้างโมเดลการจำแนกข้อมูลสำหรับหาความแตกต่างระหว่างกลุ่มที่มีค่าคะแนนภาวะซึมเศร้าสูง ต่ำ และสร้างแบบจำลองการทำนายสถานะโรคซึมเศร้า ผลที่ได้จากแบบจำลองการจำแนก และการทำนายอยู่ในระดับดี และ ความถี่ของการใช้งานอินเทอร์เน็ตมีผลกระทบต่อ

การเปลี่ยนแปลงระดับสุขภาพจิต นอกจากการพัฒนาแบบจำลองจากข้อความในเครือข่ายสังคมออนไลน์แล้ว Reece and Danforth (2017) ได้สร้างแบบจำลองการทำนายภาวะโรคซึมเศร้าจากรูปภาพในอินสตราแกรม (Instagram) โดยใช้ ขั้นตอนวิธี การวิเคราะห์การถดถอยโลจิสติแบบเบย์ (Bayesian Logistic Regression) กลุ่มตัวอย่างจำนวน 166 คน และใช้รูปภาพจำนวน 43,950 ภาพในการวิเคราะห์สี (Color analysis) และการตรวจสอบใบหน้า (Face detection) ผลการทดลองพบว่าแบบจำลองมีประสิทธิภาพดีเมื่อเทียบกับการวินิจฉัยแบบทั่วไป โดยมีค่าประสิทธิภาพโดยรวมอยู่ที่ร้อยละ 61 และงานวิจัยของ Reece et al. (2017) ได้พัฒนาแบบจำลองการพยากรณ์อาการป่วยทางจิตด้วยข้อมูลทวิตเตอร์ โดยใช้ข้อมูลจากกลุ่มตัวอย่างจำนวน 204 คน แบ่งเป็น คนที่มีภาวะซึมเศร้าจำนวน 105 คน และคนที่ไม่มีภาวะซึมเศร้าจำนวน 99 คน โดยใช้เทคนิคการสุ่มป่าไม้ (Random Forest) และใช้ข้อความจากทวิเตอร์ของกลุ่มตัวอย่างในการพัฒนาแบบจำลอง ผลการทดลองพบว่าการเริ่มมีภาวะซึมเศร้าอาจตรวจพบได้จากข้อมูลทวิตเตอร์หลายเดือนก่อนการวินิจฉัย และ O'Dea et al. (2015) ได้นำเสนอการตรวจจับการฆ่าตัวตายบนทวิตเตอร์ โดยรวบรวมทวีตที่เกี่ยวข้องกับการฆ่าตัวตาย 14,701 รายการและแบ่งออกเป็นสองชุดเท่า ๆ กัน (ชุด A และ B) และใช้เทคนิคในการจำแนกประเภทคือ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และเทคนิคการวิเคราะห์การถดถอยโลจิสติค (Logistic Regression) ซึ่งผลการทดลองแสดงให้เห็นว่าสามารถแยกระดับความกังวลระหว่างทวีตที่เกี่ยวข้องกับการฆ่าตัวตาย โดยใช้ทั้งมนุษย์และตัวแยกประเภทอัตโนมัติ (Automatic Machine Classifier) ได้ และยืนยันถึงการใช้งานทวิตเตอร์เพื่อแสดงความเกี่ยวข้องกับการฆ่าตัวตายและควรเฝ้าระวังหรือตรวจสอบโพสต์ดังกล่าว ส่วน Hu, Li, Heng, Li, and Zhu (2015) ได้นำเสนอปัจจัยที่เกี่ยวข้องเพิ่มเติมนอกเหนือจากข้อมูลในเครือข่ายสังคมออนไลน์ คือ ได้สร้างโมเดลการจำแนก และการพยากรณ์โดยใช้ข้อมูลด้านพฤติกรรมและภาษาศาสตร์จากผู้ใช้งานเครือข่ายสังคมออนไลน์จำนวน 10,102 คน และเปรียบเทียบค่าความถูกต้องระหว่างแบบจำลองที่สร้างขึ้นจากช่วงเวลาการเก็บข้อมูลที่แตกต่างกัน ผลลัพธ์พบว่าสามารถใช้ข้อมูลจากเครือข่ายสังคมออนไลน์ในการทำนายการเกิดภาวะโรคซึมเศร้า ผลลัพธ์ที่ดีที่สุดคือข้อมูลที่เก็บระหว่าง ครึ่งเดือน ถึง 2 เดือน ในขณะที่ Sun, Zhang, and Li (2019) ได้นำเสนอแบบจำลองไฮบริด (Hybrid model) ซึ่งเป็นการรวมกันระหว่างโครงข่ายประสาทแบบคอนโวลูชัน (Convolutional neural network long short-term memory : CNN-LSTM) และวิธีมอนติคาร์โลลูกโซ่มาร์คอฟ Markov chain Monte Carlo (MCMC) ใช้ในการระบุความรู้สึกของผู้ใช้งาน การเปลี่ยนแปลงของอารมณ์ และการตรวจสอบความผิดปกติของอารมณ์ โดยพัฒนามาจากวิธีมอนติคาร์โลลูกโซ่มาร์คอฟ และการคำนวณค่าความเหมือนระหว่างเมตริกซ์การเปลี่ยนแปลงปกติกับเมตริกซ์การเปลี่ยนแปลงปัจจุบันของผู้ใช้งาน ผลการทดลองสามารถแสดงอารมณ์ให้สอดคล้องกับคุณลักษณะเฉพาะบุคคลของผู้ใช้งาน และสามารถตรวจสอบความผิดปกติของบทสนทนาได้จากวิธีการที่นำเสนอ และโมเดลที่นำเสนอสามารถใช้กับบทสนทนาอัจฉริยะเช่น แบบจำลองการเปลี่ยนแปลงทางอารมณ์ และการตรวจจับความผิดปกติทางอารมณ์ได้ ในส่วนงานวิจัยของ Wen

(2021) ได้พัฒนาแบบจำลองการจัดหมวดหมู่สำหรับการตรวจหาภาวะซึมเศร้า โดยใช้ข้อความการแสดงความคิดเห็นผ่านทวิตเตอร์ร่วมกับเทคนิคโครงข่ายประสาทแบบป้อนไปข้างหน้า (Feed-Forward Neural Network) ผลการทดลองพบว่าแบบจำลองให้ค่าประสิทธิภาพโดยรวม (F-Measure) สูงถึง ร้อยละ 99.04 ในขณะที่งานวิจัยของ Cohn, Kruez, Matthews, and Yang (2009) ได้พัฒนาการตรวจจับอาการซึมเศร้าจากการกระทำของใบหน้าและเสียงพากย์ โดยเปรียบเทียบการวินิจฉัยทางคลินิกของภาวะซึมเศร้ากับใบหน้าและเสียงพูดในผู้ป่วยที่รับการรักษาภาวะซึมเศร้า โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) กับใบหน้า และใช้เทคนิคการวิเคราะห์การถดถอยโลจิสติค (Logistic Regression) สำหรับเสียง ผลการทดลองพบว่าการใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนสำหรับใบหน้าให้ค่าความถูกต้อง (Accuracy) ร้อยละ 88 และการใช้เทคนิคการวิเคราะห์การถดถอยโลจิสติคสำหรับเสียง ให้ค่าความถูกต้อง (Accuracy) ร้อยละ 79 เช่นเดียวกับงานวิจัยของ Gou et al. (2021) ได้นำเสนอโครงข่ายประสาทเทียมแบบลึก (Deep Neural Networks) เพื่อการรู้จำอาการซึมเศร้าโดยอาศัยการแสดงออกทางสีหน้า 2 มิติและ 3 มิติ ภายใต้ภารกิจกระตุ้นอารมณ์ โดยการเก็บรวบรวมเสียงและวีดีโอของอาสาสมัคร โดยนำเสนอแนวทางใหม่สำหรับการรับรู้ความเสี่ยงจากภาวะซึมเศร้าที่อาจเกิดขึ้นตามแบบจำลองเครือข่ายความเชื่อแบบลึก (Deep Belief Network) สองแบบ คือแบบ 2 มิติ และ 3 มิติ ผลการทดลองพบว่าการรวมคุณลักษณะ 2 มิติ และ 3 มิติ ในการสร้างแบบจำลองเครือข่ายความเชื่อแบบลึก (Deep Belief Network) นั้นสมเหตุสมผลและเป็นสากลมากกว่าวิธีอื่น ๆ

จากการทบทวนงานวิจัยพบว่า งานวิจัยส่วนใหญ่มุ่งเน้นไปที่การนำข้อมูลในสื่อสังคมออนไลน์มาวิเคราะห์การเกิดภาวะโรคซึมเศร้า โดยใช้เทคนิคที่แตกต่างกันซึ่งสามารถสรุปได้ดังตารางที่ 2.3 นอกจากนั้นยังพบว่างานวิจัยบางส่วนไม่ได้มีการเปรียบเทียบเทคนิคอื่น ๆ เพื่อหาวิธีที่ดีที่สุด อีกทั้งงานวิจัยส่วนใหญ่เน้นไปที่การใช้ข้อความในเครือข่ายสังคมออนไลน์มาวิเคราะห์เพียงอย่างเดียว โดยไม่มีการใช้ปัจจัยอื่น หรือคุณลักษณะอื่น ๆ ร่วมในการวิเคราะห์ ได้แก่คุณลักษณะด้านกายภาพ เช่น เพศ อายุ จำนวนเพื่อนในเครือข่ายสังคมออนไลน์ จำนวนผู้ติดตาม จำนวนผู้กดไลค์ ช่วงเวลาที่ใช้งานเครือข่ายสังคมออนไลน์ และคุณลักษณะของข้อความ เช่น การใช้งานอีโมติคอนในข้อความ และความยาวของข้อความแสดงความคิดเห็น รวมถึงการใช้งานรูปภาพในการแสดงความคิดเห็น และยังไม่มีงานวิจัยใดที่พัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ที่สามารถบ่งบอกระดับคะแนนภาวะซึมเศร้าได้

จึงนำมาสู่แนวคิดในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ซึ่งข้อมูลที่นำมาใช้ในการพัฒนาแบบจำลอง ได้แก่ ข้อความแสดงความคิดเห็น คุณลักษณะทางกายภาพ และคุณลักษณะของรูปภาพ และเทคนิคที่นำมาใช้ในการพัฒนาแบบจำลองนี้ได้แก่ เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) การประมวลผลรูปภาพ (Image Processing) การวิเคราะห์

ความรู้สึก (Sentiment Analysis) เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) โดยแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์นี้สามารถวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ซึ่งบ่งบอกถึงระดับคะแนนของความเสี่ยงที่จะเกิดภาวะซึมเศร้าที่เหมาะสมกับบริบทในสังคมไทย และมีค่าความถูกต้องในระดับที่ยอมรับได้

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 2.3 งานวิจัยที่เกี่ยวข้อง

ลำดับ	งานวิจัย	วัตถุประสงค์				ที่มาของข้อมูล	ปัจจัย					วิธีการที่นำมาใช้						ผลลัพธ์	
		การวิเคราะห์ความรู้สึก	การวิเคราะห์ภาวะซึมเศร้า	การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า	การสำรวจงานวิจัยที่เกี่ยวข้อง		**Sentiment Analysis**	**Text**	**Emoticon**	**Behavior**	**Image**	**Support Vector Machine**	**Naïve Bayes**	**Decision Tree**	**Regression**	**Random forest**	**Deep learning**	มีภาวะซึมเศร้า/ไม่มีภาวะซึมเศร้า	แสดงระดับคะแนนภาวะซึมเศร้า
1	Cohn et al. (2009)			X		Questionnaires & Interviews				X								X	
2	Park et al. (2012)		X			Twitter	X	X	X		X				X			X	
3	Barhan and Shakhomirov (2012)	X				Twitter	X	X	X			X	X					X	
4	Wang et al. (2013)		X			Sina Micro-blog	X	X	X	X			X	X				X	
5	De Choudhury et al. (2013)			X		Twitter	X	X			X	X						X	
6	Hutto and Gilbert (2014)	X				Twitter	X	X				X	X		X			X	

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 2.3 งานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	งานวิจัย	วัตถุประสงค์				ที่มาของข้อมูล	ปัจจัย					วิธีการที่นำมาใช้						ผลลัพธ์	
		การวิเคราะห์ความรู้สึก	การวิเคราะห์ภาวะซึมเศร้า	การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า	การสำรวจงานวิจัยที่เกี่ยวข้อง		Sentiment Analysis	Text	Emoticon	Behavior	Image	Support Vector Machine	Naïve Bayes	Decision Tree	Regression	Random forest	Deep learning	มีภาวะซึมเศร้า/ไม่มีภาวะซึมเศร้า	แสดงระดับคะแนนภาวะซึมเศร้า
7	Kashyap (2014)	X				Twitter	X	X										X	
8	Hu et al. (2015)			X		Sina Weibo		X	X	X	X				X			X	
9	O'Dea et al. (2015)			X		Twitter	X	X				X						X	
10	Vareekul and Koomsubha (2016)	X				Twitter	X	X									X	X	
11	Zhu et al. (2016)			X		Log file				X			X		X			X	
12	Guntuku et al. (2017)				X	-													
13	Reece and Danforth (2017)			X		Instagram					X				X			X	

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 2.3 งานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	งานวิจัย	วัตถุประสงค์				ที่มาของข้อมูล	ปัจจัย					วิธีการที่นำมาใช้						ผลลัพธ์	
		การวิเคราะห์ความรู้สึก	การวิเคราะห์ภาวะซึมเศร้า	การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า	การสำรวจงานวิจัยที่เกี่ยวข้อง		**Sentiment Analysis**	**Text**	**Emoticon**	**Behavior**	**Image**	**Support Vector Machine**	**Naïve Bayes**	**Decision Tree**	**Regression**	**Random forest**	**Deep learning**	มีภาวะซึมเศร้า/ไม่มีภาวะซึมเศร้า	แสดงระดับคะแนนภาวะซึมเศร้า
14	Reece et al. (2017)			X		Twitter		X								X		X	
15	Jiang et al. (2017)		X			Twitter	X	X										X	
16	Aldarwish (2017)			X		Facebook	X	X				X	X					X	
17	Islam et al. (2018)	X				Facebook	X	X				X						X	
18	Sood et al. (2018)	X				Twitter	X	X										X	
19	Orabi et al.(2018)		X			Twitter		X									X	X	
20	Sun et al. (2019)			X		Twitter	X	X									X	X	

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 2.3 งานวิจัยที่เกี่ยวข้อง (ต่อ)

ลำดับ	งานวิจัย	วัตถุประสงค์				ที่มาของข้อมูล	ปัจจัย					วิธีการที่นำมาใช้						ผลลัพธ์	
		การวิเคราะห์ความรู้สึก	การวิเคราะห์ภาวะซึมเศร้า	การพัฒนาแบบจำลองการเกิดภาวะซึมเศร้า	การสำรวจงานวิจัยที่เกี่ยวข้อง		**Sentiment Analysis**	**Text**	**Emoticon**	**Behavior**	**Image**	**Support Vector Machine**	**Naïve Bayes**	**Decision Tree**	**Regression**	**Random forest**	**Deep learning**	มีภาวะซึมเศร้า/ไม่มีภาวะซึมเศร้า	แสดงระดับคะแนนภาวะซึมเศร้า
21	Skaik and Inkpen (2020)				X	-													
22	Esteban A. Rissola et al. (2021)				X	-													
23	Jiin Kim et al. (2021)				X	-													
24	Wen (2021)			X		Twitter		X									X	X	
25	Gou et al. (2021)			X		Questionnaires & Interviews					X						X	X	
26	งานวิจัยนี้	X	X	X		Twitter	X	X	X	X	X	X	X	X		X	X	X	X

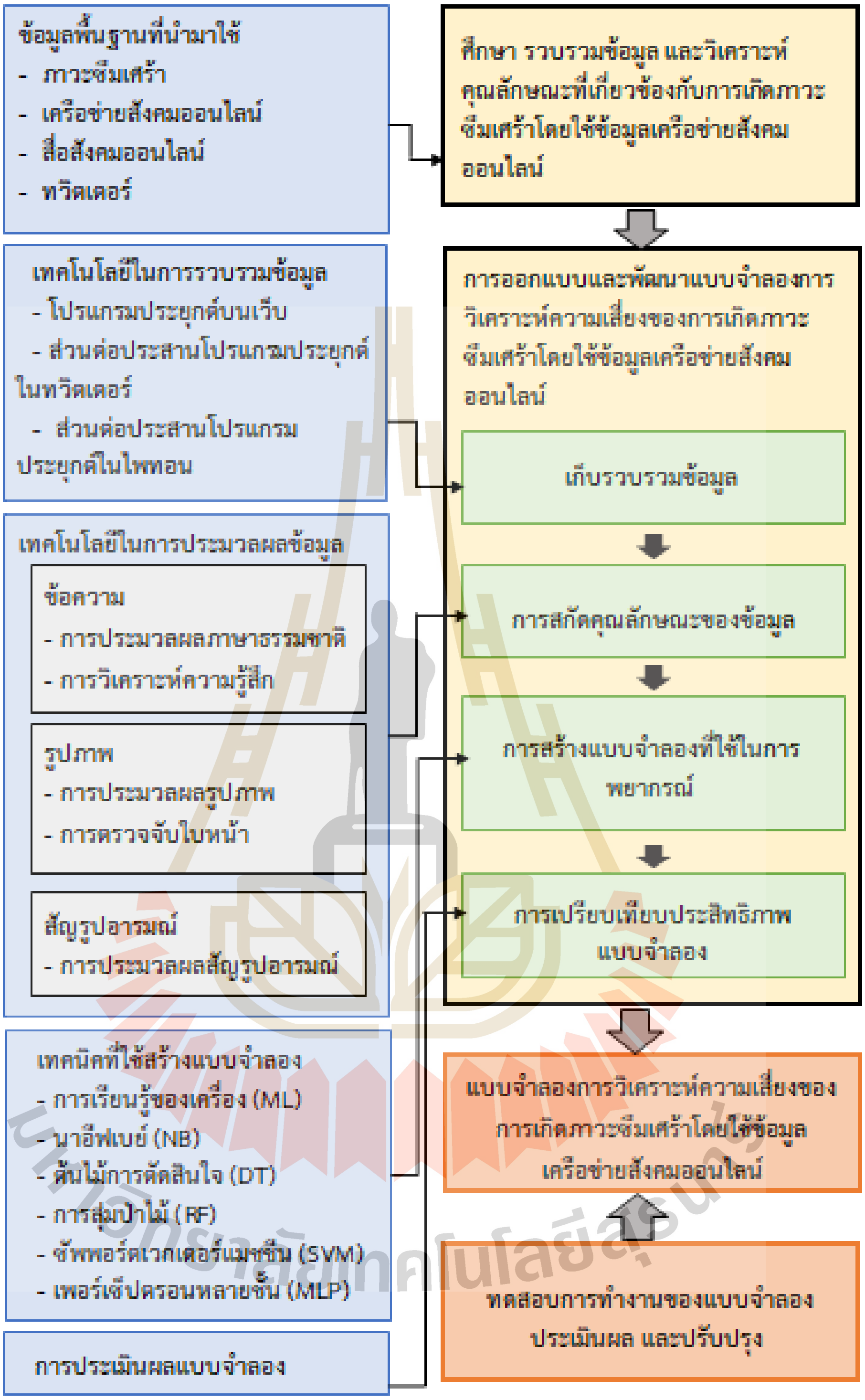
มหาวิทยาลัยเทคโนโลยีสุรนารี

2.7 กรอบแนวคิดการวิจัย

จากการปริทัศน์วรรณกรรม และศึกษาเอกสารงานวิจัยที่เกี่ยวข้อง งานวิจัยนี้ได้นำแนวคิดทฤษฎีที่เกี่ยวข้องมาใช้ในการศึกษา ประกอบด้วย

1. ข้อมูลพื้นฐานที่นำมาใช้ในการสร้างแบบจำลองในการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าซึ่งได้จากการศึกษาทฤษฎีเกี่ยวกับ ภาวะซึมเศร้า และเครือข่ายสังคมออนไลน์
2. เทคโนโลยีในการรวบรวมข้อมูล ใช้ในการรวบรวมข้อมูลจากแบบประเมินภาวะซึมเศร้า และข้อมูลจากผู้ใช้งานทวิตเตอร์ ประกอบด้วย โปรแกรมประยุกต์บนเว็บ ส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ และส่วนต่อประสานโปรแกรมประยุกต์ในไพทอน
3. เทคโนโลยีในการประมวลผลข้อมูล ได้แก่ การประมวลผลภาษาธรรมชาติ การวิเคราะห์ความรู้สึก และ การประมวลผลรูปภาพ เป็นต้น
4. เทคนิคที่นำมาใช้ในการสร้างแบบจำลองโดยใช้เทคนิคการเรียนรู้ของเครื่องในรูปแบบต่าง ๆ
5. การประเมินผลแบบจำลอง เพื่อนำมาใช้สำหรับการพัฒนากรอบแนวคิดในการวิจัยเรื่อง การพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์ ดังภาพที่ 2.36

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 2.36 กรอบแนวคิดการวิจัย

บทที่ 3
วิธีดำเนินการวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยมุ่งเน้นการออกแบบและพัฒนาขั้นตอนวิธีที่ใช้ในการสร้างแบบจำลองในการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าจากเครือข่ายสังคมออนไลน์ สำหรับรายละเอียดในบทที่ 3 มีดังต่อไปนี้

3.1 วิธีวิจัย

งานวิจัยนี้เป็นการวิจัยเชิงประยุกต์ (Applied Research) ในหัวข้อนี้ได้กล่าวถึงวิธีวิจัยที่ใช้ประกอบด้วย 3 ขั้นตอน ได้แก่ 1) ศึกษา รวบรวมข้อมูล และวิเคราะห์ปัจจัยที่เกี่ยวข้องกับการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลจากเครือข่ายสังคมออนไลน์ 2) ออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า และ 3) ทดสอบการทำงานของแบบจำลอง ประเมินผล และปรับปรุง โดยมีรายละเอียดดังนี้

3.1.1 ศึกษา รวบรวมข้อมูล และวิเคราะห์คุณลักษณะที่เกี่ยวข้องกับการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

จากการศึกษาคุณลักษณะที่เกี่ยวข้องกับปัญหาภาวะซึมเศร้าในงานวิจัยต่าง ๆ และจากแบบประเมินโรคซึมเศร้า 9 คำถาม (Marra, 2004; Pfizer, 2005; Parekh, 2017; WHO, 2018; ประชาสัมพันธ์กรมสุขภาพจิต, 2561; บุษราคัม จิตอารีย์, 2555; นันทิรา หงส์ศรีสุวรรณ์, 2559) พบว่า มีคุณลักษณะที่เกี่ยวข้องดังแสดงรายละเอียดในตารางที่ 3.1

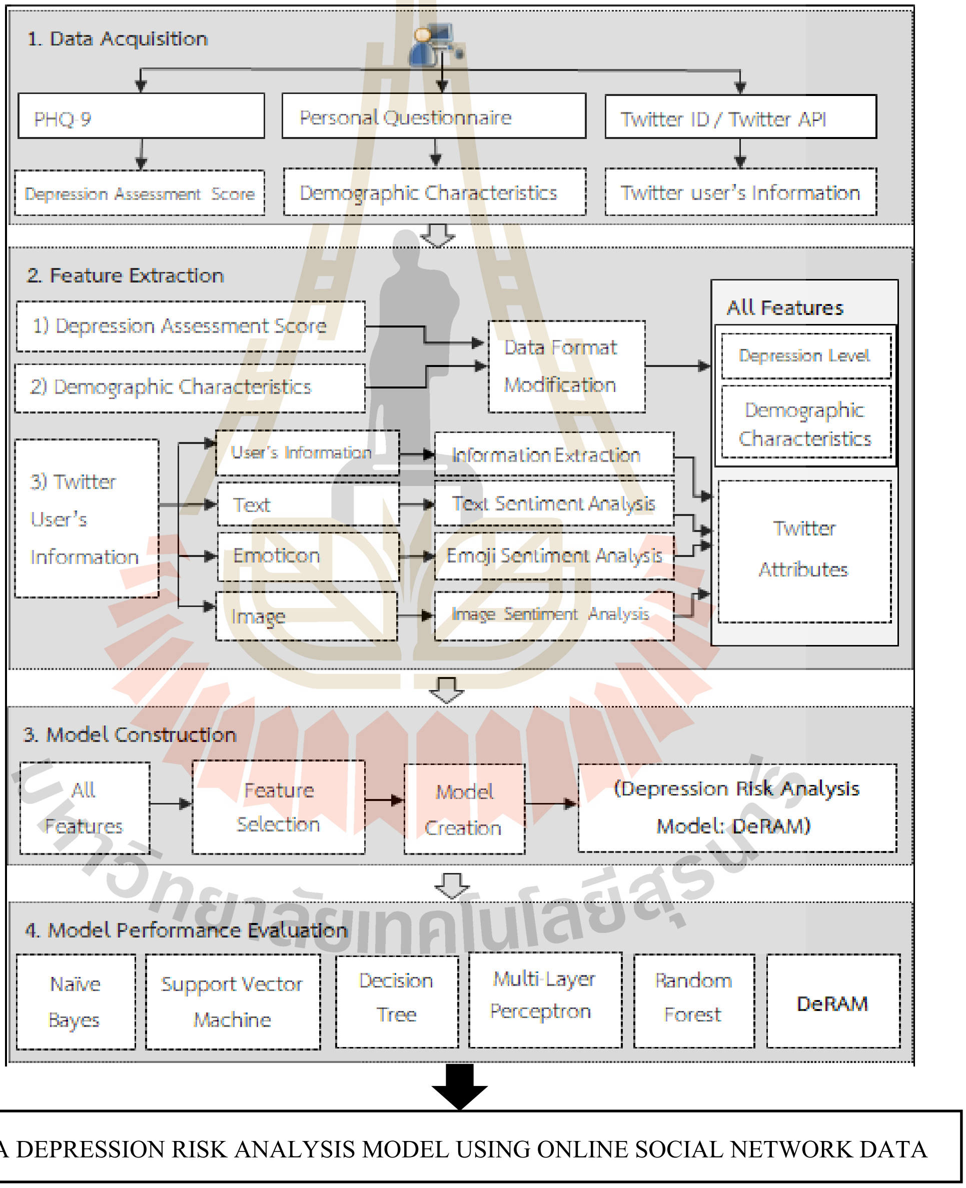
ตารางที่ 3.1 คุณลักษณะที่เกี่ยวข้องกับภาวะซึมเศร้า

คุณลักษณะของข้อมูลในทวิตเตอร์	อาการที่บ่งบอกถึงภาวะซึมเศร้า
- ข้อความ (Text)	- อาการเบื่อ ไม่สนใจอยากทำอะไร
- สัญรูปอารมณ์ (Emoticon)	- ไม่สบายใจ ซึมเศร้า ท้อแท้
- รูปภาพ (Image)	- หลับยาก หรือหลับ ๆ ตื่น ๆ หรือหลับมากเกินไป
- จำนวนผู้ติดตาม (Number of Followers)	- เหนื่อยง่ายหรือไม่ค่อยมีแรง
- จำนวนทวีต (Number of Tweet)	- เบื่ออาหารหรือกินมากเกินไป
- จำนวนรีทวีต (Number of Retweet)	- รู้สึกไม่ดีกับตัวเอง คิดว่าตัวเองล้มเหลวหรือครอบครัวผิดหวัง
- จำนวนแฮชแท็ก (Number of Hashtag)	- สมาธิไม่ดี
- วันที่โพสต์ข้อความ (Date of Post)	- พูดช้า หรือทำอะไรได้ช้าลง
- เวลาที่โพสต์ข้อความ (Time of Post)	- คิดทำร้ายตนเองหรือคิดว่าถ้าตายไปคงจะดี

มหาวิทยาลัยเทคโนโลยีสุรนารี

3.1.2 ออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ (Depression Risk Analysis Model: DeRAM) ที่พัฒนาขึ้นในงานวิจัยนี้ มุ่งเน้นการวิเคราะห์ และทำนายความน่าจะเป็นในการเกิดภาวะซึมเศร้า โดยมีกรอบการทำงานของระบบดังภาพที่ 3.1 ซึ่งมีกระบวนการทำงานประกอบด้วย 4ขั้นตอน ซึ่งในแต่ละขั้นตอนมีรายละเอียดดังต่อไปนี้



ภาพที่ 3.1 กรอบการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

3.1.2.1 ขั้นตอนการเก็บรวบรวมข้อมูล (Data Acquisition)

ในขั้นตอนนี้เป็นการเก็บข้อมูลจากผู้ใช้งานเครือข่ายสังคมออนไลน์ ในการทดลองนี้ใช้กลุ่มตัวอย่างจากผู้ใช้งานทวิตเตอร์ โดยสร้างโปรแกรมประยุกต์บนเว็บสำหรับเก็บข้อมูลคุณลักษณะทางกายภาพ และผลการประเมินภาวะซึมเศร้า ตามแบบประเมินภาวะซึมเศร้า 9 คำถามจากกรมสุขภาพจิตดังแสดงในภาพที่ 3.2 และเก็บข้อมูลการใช้งานทวิตเตอร์ของผู้ใช้ โดยใช้ทวิตพายเอพีไอ (Tweepy API) ร่วมกับภาษาไพทอน (Python) ในการดึงข้อมูลการแสดงความคิดเห็นในทวิตเตอร์ ประกอบด้วย ข้อความ รูปภาพ และสัญรูปอารมณ์ โดยสืบค้นจากคำค้น (Keyword) ที่เกี่ยวข้องกับภาวะซึมเศร้า เช่น โรคซึมเศร้า ซึมเศร้า อยากตาย เสียใจ ผิดหวัง ฯลฯ หลังจากนั้นจึงเข้าสู่กระบวนการเตรียมข้อมูล

ภาพที่ 3.2 โปรแกรมประยุกต์บนเว็บสำหรับเก็บข้อมูลตามแบบประเมินภาวะซึมเศร้า 9 คำถาม

ในขั้นตอนการเก็บรวบรวมข้อมูลนี้ สามารถแบ่งการเก็บรวบรวมข้อมูลออกเป็น 3 ส่วนได้แก่

มหาวิทยาลัยเทคโนโลยีสุรนารี

1) การเก็บรวบรวมข้อมูลจากแบบประเมินภาวะซึมเศร้า 9 คำถาม (PHQ-9) ซึ่งเป็นข้อมูลการตอบแบบประเมินภาวะซึมเศร้า 9 คำถาม ข้อคำถามประกอบด้วย

(1) เบื่อไม่อยากทำอะไรเลย

(2) ไม่สบายใจ ซึมเศร้า ท้อแท้

(3) หลับยาก หรือหลับ ๆ ตื่น ๆ หรือหลับมากเกินไป

(4) เหนื่อยง่ายหรือไม่ค่อยมีแรง

(5) เบื่ออาหารหรือกินมากเกินไป

(6) รู้สึกไม่ดีกับตัวเอง คิดว่าตัวเองล้มเหลว หรือทำให้ตนเองหรือครอบครัวผิดหวัง

(7) สมาธิไม่ดีเวลาทำอะไร เช่น ดูโทรทัศน์ ฟังวิทยุ หรือทำงานที่ต้องใช้ความตั้งใจ

(8) พูดช้า ทำอะไรช้าลงจนคนอื่นสังเกตได้ หรือกระสับกระส่ายไม่สามารถอยู่นิ่งได้เหมือนที่เคยเป็น

(9) คิดทำร้ายตนเอง หรือคิดว่าตายไปคงจะดี

ผลลัพธ์ที่ได้จากขั้นตอนนี้คือ ระดับคะแนนภาวะซึมเศร้า ซึ่งได้จาก เกณฑ์การให้คะแนนของคำตอบในแต่ละข้อคือ ไม่มีเลย = 0 เป็นบางวัน = 1 เป็นบ่อย = 2 เป็นทุกวัน = 3

2) การเก็บรวบรวมข้อมูลการตอบแบบสอบถามข้อมูลพื้นฐาน (Personal Questionnaire) โดยข้อคำถามประกอบด้วย เพศ อายุ น้ำหนัก ระดับการศึกษา โรคประจำตัว อาชีพ รายได้ จำนวนสมาชิกในครอบครัว สถานภาพ (ตนเอง) และสถานภาพ (บิดา-มารดา)

3) การเก็บรวบรวมข้อมูลจากทวิตเตอร์ซึ่งได้จากการใช้งานทวิตเตอร์ (Twitter ID / Twitter API) ผลลัพธ์ที่ได้คือ ข้อมูลจากทวิตเตอร์ (Twitter user's Information) ซึ่งประกอบด้วย ข้อมูลทั่วไป (User's Information) ข้อมูลประเภทข้อความ (Text) ข้อมูลสัญรูปอารมณ์ (Emoticon) และรูปภาพ (Image)

3.1.2.2 ขั้นตอนการสกัดคุณลักษณะของข้อมูล (Feature Extraction)

ในขั้นตอนนี้เป็นการสกัดคุณลักษณะจากข้อมูล 3 ส่วน คือ ข้อมูลระดับคะแนนภาวะซึมเศร้า (Depression Assessment Score) ข้อมูลด้านกายภาพส่วนบุคคล (Demographic Characteristics) ซึ่งเป็นข้อมูลที่ได้จากการทำแบบสอบถามข้อมูลพื้นฐาน และแบบประเมินภาวะซึมเศร้า 9 คำถาม และข้อมูลจากทวิตเตอร์ (Twitter user's Information) โดยมีรายละเอียดดังนี้

1) การสกัดคุณลักษณะจากข้อมูลระดับคะแนนภาวะซึมเศร้า (Depression Assessment Score) เป็นระดับคะแนนภาวะซึมเศร้าที่คำนวณได้จากการทำแบบประเมินภาวะซึมเศร้า 9 คำถาม โดยเปลี่ยนรูปแบบข้อมูล (Data Format Modification) จาก Numeric เป็น

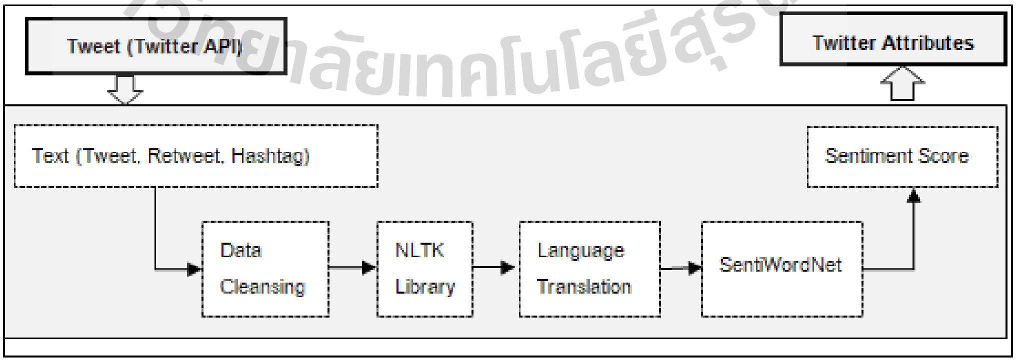
Nominal ซึ่งใช้เกณฑ์การเปลี่ยนข้อมูลตามแบบสอบถามภาวะซึมเศร้า 9 คำถาม คือ คะแนนรวมน้อยกว่า 7 ไม่มีอาการของภาวะซึมเศร้า คะแนนรวม 7 – 12 มีภาวะซึมเศร้าระดับน้อย คะแนนรวม 13 – 18 มีภาวะซึมเศร้าระดับปานกลาง คะแนนรวมมากกว่า 19 คะแนน มีภาวะซึมเศร้าระดับรุนแรง

2) การสกัดคุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics) ได้จากการทำแบบสอบถามข้อมูลส่วนบุคคล ประกอบด้วย 10 คุณลักษณะ ได้แก่ เพศ อายุ น้ำหนัก ระดับการศึกษา โรคประจำตัว อาชีพ รายได้ จำนวนสมาชิกในครอบครัว สถานภาพ (ตนเอง) และสถานภาพ (บิดา-มารดา) หลังจากนั้น จึงมีการปรับเปลี่ยนรูปแบบข้อมูล (Data Format Modification) เพื่อให้เหมาะสมกับเทคนิคที่นำมาใช้ในการสร้างแบบจำลอง

3) การสกัดคุณลักษณะจากทวิตเตอร์ (Twitter user's Information Extraction) ซึ่งข้อมูลในทวิตเตอร์ที่นำมาสกัดคุณลักษณ์ แบ่งเป็น 4 ส่วน ได้แก่ ข้อมูลทั่วไป ข้อมูลประเภทข้อความ ข้อมูลสัญรูปอารมณ์ (Emoticon) และข้อมูลรูปภาพ (Image)

3.1) การสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป 4 คุณลักษณะ (User's Information) ได้แก่ จำนวนเพื่อน จำนวนผู้ติดตาม จำนวนทวีตที่อยู่ระหว่างเวลา 6.00 น.-00.00 น. และจำนวนทวีตที่อยู่ระหว่างเวลา 0.00 น.- 6.00 น. โดยใช้ทวิตพายเอพีไอ (Tweepy API) ร่วมกับภาษาไพทอน (Python) ในการดึงข้อมูลซึ่งข้อมูลจะถูกจัดเก็บในรูปแบบไฟล์เอ็กเซล (Excel)

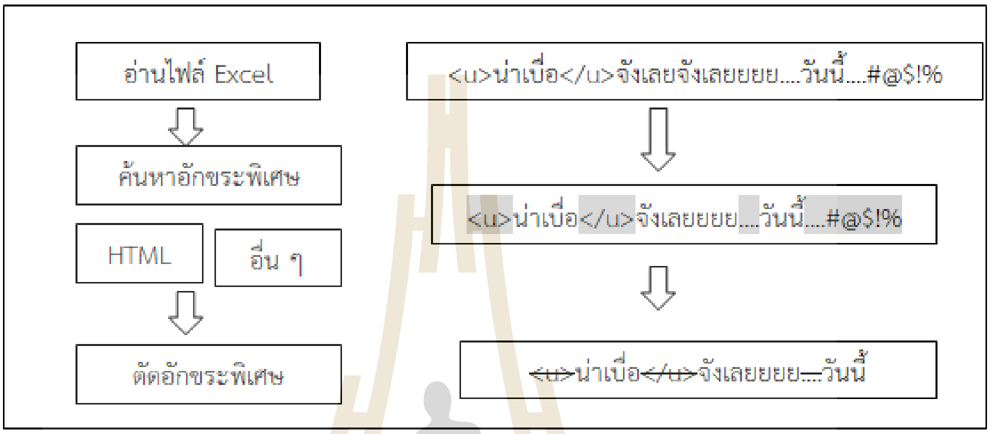
3.2) การสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) ได้แก่ ทวีต (Tweet) รีทวีต (Retweet) และแฮชแท็ก (Hashtag) เพื่อใช้การวิเคราะห์ความรู้สึก (Sentiment Analysis) หรือจำแนกขั้วอารมณ์ เป็นการนำข้อความที่ได้จากขั้นตอนการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ด้วย NLTK Library และใช้ตัวแปลภาษา (Language Translation) เปลี่ยนจากภาษาไทยเป็นภาษาอังกฤษ จากนั้นนำผลลัพธ์ที่ได้มาค้นหาคำศัพท์ในฐานข้อมูลคำศัพท์ที่บอกระดับคำเป็นบวก เป็นลบ และเป็นกลาง (SentiWordNet) เพื่อคำนวณหาค่าคะแนนความรู้สึก (Sentiment Score) ดังแสดงในภาพที่ 3.3



ภาพที่ 3.3 การสกัดคุณลักษณะจากข้อมูลที่เป็นประเภทข้อความในทวิตเตอร์

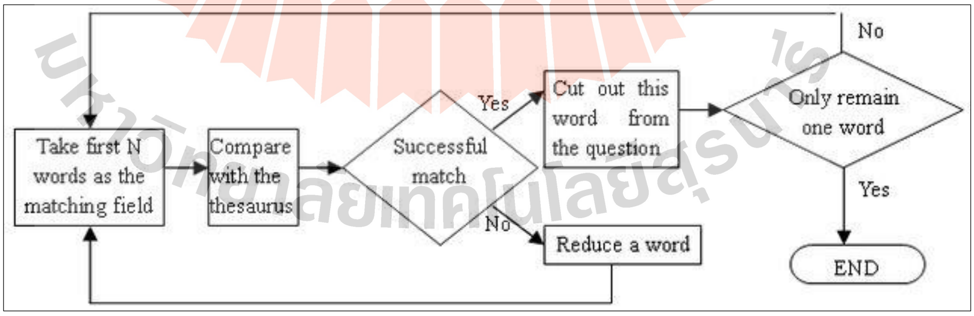
โดยมีรายละเอียดในแต่ละขั้นตอน ดังต่อไปนี้

- การทำความสะอาดข้อมูล (Data Cleansing) เนื่องจากข้อมูลที่ได้มาเป็นข้อความ (Text) ที่จำเป็นต้องใช้กระบวนการเพื่อตัดข้อความหรืออักขระที่ไม่เกี่ยวข้องออก และจัดการข้อมูลให้เหมาะสมก่อนการทำการวิเคราะห์ความรู้สึก โดยมีขั้นตอนดังแสดงในภาพที่ 3.4



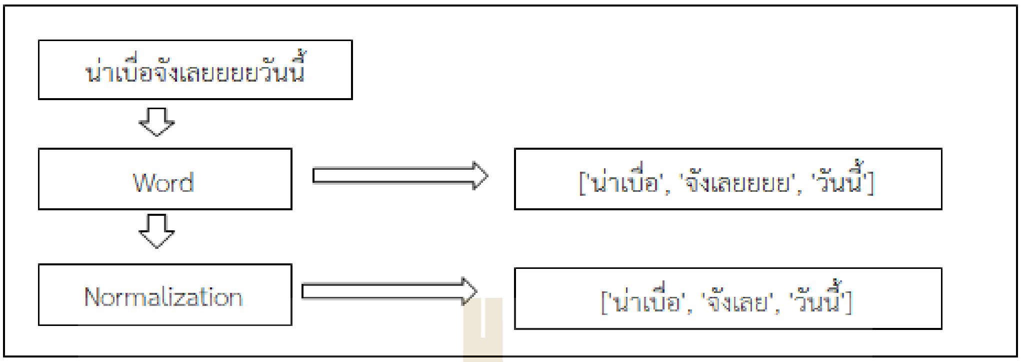
ภาพที่ 3.4 ขั้นตอนการตรวจสอบข้อมูล

- การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ด้วย NLTK Library หลังจากทำความสะอาดข้อมูลแล้วจึงนำข้อมูลที่ได้มาผ่านกระบวนการประมวลผลภาษาธรรมชาติ โดยใช้ กระบวนการตัดพยางค์ย่อย (Tokenization) เพื่อแยกข้อความในรูปประโยค (Sentence) ออกเป็นคำ (Word) หรือเรียกว่าการตัดคำ (Word Segmentation) โดยใช้ขั้นตอนวิธีการตัดคำแบบเหมือนมากที่สุด (Maximum matching Algorithm) ดังแสดงในภาพที่ 3.5



ภาพที่ 3.5 ขั้นตอนวิธีการตัดคำแบบเหมือนมากที่สุด (Zhenqiu, 2012)

หลังจากนั้นจึงเข้าสู่กระบวนการทำให้เป็นมาตรฐาน (Normalization) คือการเปลี่ยนคำที่มีความหมายเดียวกันแต่มีวิธีการเขียนต่างกันให้อยู่ในรูปแบบเดียวกัน เช่น เสียใจ เสียจัย เสียจาย เป็นต้น ดังแสดงในภาพที่ 3.6



ภาพที่ 3.6 กระบวนการการประมวลผลภาษาธรรมชาติ (Natural Language Processing)

กระบวนการทำให้เป็นมาตรฐานนี้ใช้คลังคำเหมือน (Synonym Word Corpus) ซึ่งเป็นไฟล์ที่เก็บข้อมูลคำศัพท์ในรูปแบบไฟล์ CSV ดังแสดงในตารางที่ 3.2 ในการตรวจสอบวิธีการเขียนที่แตกต่างกัน เพื่อให้อยู่ในรูปแบบเดียวกัน โดยใช้คำสั่งในการค้นหาคำศัพท์ในคลังคำเหมือนดังแสดงในภาพที่ 3.7

ตารางที่ 3.2 ตัวอย่างคลังคำเหมือน (Synonym Word Corpus)

คำศัพท์	คำเหมือน
เสียใจ	เสียจัย|เสียจาย
จังเลย	จังเลยย|จังเลยยย|จังเลยยยย|จังเลยยยยยย
น่าเบื่อ	น่าเบื่ออ|น่าเบื่อออ|น่าเบื่ออออ|น่าเบื่อออออออ
ร้องไห้	ร้องห้าย|ร้องไห้
เศร้า	โศกเศร้า|กำสรวล|คร่ำครวญ|โศก|มลาน|เศร้าโศก|ซึมเศร้า|เศร้าใจ

```
import csv
from pandas import *
data = read_csv('synonym_word.csv').to_dict()
synonym_dic=data["synonym"]
keys = [k for k, v in synonym_dic.items() if 'เสียจัย' in v]
print('Word: ',data["word"][keys[0]])
```

ภาพที่ 3.7 คำสั่งในการค้นหาคำศัพท์ในคลังคำเหมือน

- การแปลภาษา (Language Translation) เพื่อให้สามารถใช้งาน SentiWordNet ซึ่งเป็นคลังคำศัพท์ภาษาอังกฤษที่ใช้ในการวิเคราะห์ความรู้สึกของคำ ที่มีจำนวนมากถึง 117,695 คำ เนื่องจากการแปลคำศัพท์จากภาษาไทยเป็นภาษาอังกฤษให้ความถูกต้องสูงกว่าการแปลในรูปประโยค โดยการแปลคำศัพท์ใช้ตัวแปลภาษา เปลี่ยนจากภาษาไทยเป็นภาษาอังกฤษด้วย translate 3.6.1 Python Library ใน PyPi (The Python Package Index) ซึ่งเป็นแหล่งรวมชุดคำสั่งของ Python และเป็นเครื่องมือแปลภาษาที่ใช้งานง่าย รองรับผู้ให้บริการที่หลากหลาย เช่น Microsoft Translation API MyMemory API LibreTranslate และ DeepL APIs ซึ่งในงานวิจัยนี้ใช้ผู้ให้บริการในการแปลภาษาคือ MyMemory API โดยมีขั้นตอนแสดงดังภาพที่ 3.8

```
from translate import Translator
word = 'เสียใจ'
translator= Translator(from_lang="thai",to_lang="english")
translation = translator.translate(word)
print(translation)
```

ภาพที่ 3.8 คำสั่งที่ใช้สำหรับแปลภาษาด้วย MyMemory API Translate

- การใช้งาน SentiWordNet เพื่อวิเคราะห์ความรู้สึก (Sentiment Analysis) เมื่อได้ข้อมูลที่ผ่านกระบวนการประมวลผลภาษาธรรมชาติและการแปลภาษาแล้ว แล้วจึงนำมาใช้ในการวิเคราะห์ความรู้สึก (Sentiment Analysis) เพื่อหาค่าระดับคะแนนความรู้สึก (Sentiment Score) โดยใช้ SentiWordNet เวอร์ชั่น 3.0 ซึ่งมีคลังคำศัพท์มากที่สุด โดยมีมากถึง 117,695 คำ (Musto, 2014) มาใช้ร่วมกับเอพีไอในภาษาไพทอน ในการจำแนกขั้วอารมณ์สำหรับข้อความที่เป็นภาษาไทย และทำการจัดเก็บข้อมูลผลลัพธ์ที่ได้ในรูปแบบไฟล์เอ็กเซล (Excel) โดยมีขั้นตอนวิธีดังแสดงในภาพที่ 3.9

มหาวิทยาลัยเทคโนโลยีสุรนารี

```
f = open("sentiwordnet3.0.txt")
count =0.0, totalpositive =0.0, totalnegative =0.0
for line in f:
   if not line.startswith("#"):
      cols = split_line(line)
      words = get_words(cols)
      for word in sentiword:
         if word in words:
            totalpositive = totalpositive + float(get_positive(cols))
            totalnegative = totalnegative + float(get_negative(cols))
            count =count + 1
      if count >0:
   if totalpositive > totalnegative :
          print("Positive word : 1")
          print("Positive value : ",totalpositive)
          print("Negative value : ",totalnegative)
   else :
          print("Negative : -1")
          print("Positive value : ",totalpositive)
          print("Negative value : ",totalnegative)
totalscore=(totalpositive/count) - (totalnegative/count)
```

ภาพที่ 3.9 ขั้นตอนวิธีการหาค่าคะแนนความรู้สึก (Sentiment Score)

จากภาพที่ 3.9 เป็นการนำข้อความที่ได้จากขั้นตอนการประมวลผลภาษาธรรมชาติมาค้นหาคำศัพท์ใน SentiWordNet (ไฟล์ sentiwordnet3.0.txt) ดังแสดงในภาพที่ 3.10

# POS	ID	PosScore	NegScore	SynsetTerms	Gloss¶
a	00001740	0.125	0	able#1	(usually followed by `to') having the necessary means
a	00002098	0	0.75	unable#1	(usually followed by `to') not having the nece
a	00002312	0	0	dorsal#2 abaxial#1	facing away from the axis of an organ
a	00002527	0	0	ventral#2 adaxial#1	nearest to or facing toward the axis c
a	00002730	0	0	acroscopic#1	facing or on the side toward the apex¶
a	00002843	0	0	basiscopic#1	facing or on the side toward the base¶
a	00002956	0	0	abducting#1 abducent#1	especially of muscles; drawing away fr
a	00003131	0	0	adductive#1 adducting#1 adducent#1	especially of muscles;
a	00003356	0	0	nascent#1	being born or beginning; "the nascent chicks";
a	00003553	0	0	emerging#2 emergent#2	coming into existence; "an emergent re
a	00003700	0.25	0	dissilient#1	bursting open with force, as do some ripe seed
a	00003829	0.25	0	parturient#2	giving birth; "a parturient heifer"¶

ภาพที่ 3.10 ตัวอย่างไฟล์ sentiwordnet3.0.txt

ซึ่งในไฟล์ sentiwordnet3.0.txt ประกอบด้วยคำศัพท์ และค่าคะแนนต่าง ๆ โดยคำศัพท์แต่ละคำ จะประกอบด้วย ค่าคะแนนด้านบวก (Positive score: ***Pos(s)***) ค่าคะแนนด้านลบ (Negative score: ***Neg(s)***) และค่าคะแนนเป้าหมาย (Objective Score: ***Obj(s)***) ดังแสดงในสมการที่ 3.1

$$Pos(s) + Neg(s) + Obj(s) = 1 \tag{3.1}$$

โดยค่าเฉลี่ยคะแนนความรู้สึก (Sentiment Score) ของทวีตทั้งหมด $\boldsymbol{\theta}$ ได้จากคะแนนความรู้สึก (Sentiment Score) ของแต่ละทวีต $Tweet(s)_i$ ดังแสดงในสมการที่ 3.2

$$\theta = \frac{\sum_i^n Tweet(s)_i}{n} \tag{3.2}$$

ซึ่งคะแนนความรู้สึกของแต่ละทวีต $Tweet(s)_i$ คำนวณได้จากคะแนนด้านบวกและคะแนนด้านลบของจำนวนคำ m ในทวีต ดังแสดงในสมการที่ 3.3

$$Tweet(s) = \frac{\sum_{j=1}^m Pos(s)_j}{m} - \frac{\sum_{j=1}^m Neg(s)_j}{m} \tag{3.3}$$

จากสมการดังกล่าว ทำให้สามารถคำนวณหาจำนวนทวีตด้านบวกได้จาก ผลรวมของจำนวนทวีตที่มีค่าคะแนนความรู้สึกของแต่ละทวีต ***Tweet(s) >0*** และคำนวนจำนวนทวีตด้านลบได้จากผลรวมของจำนวนทวีตที่มีค่าคะแนนความรู้สึกของแต่ละทวีต ***Tweet(s)*** $\leq$***0*** และจำนวนทวีตที่แสดงถึงภาวะซึมเศร้าได้จาก ผลรวมของทวีตที่มีคำที่แสดงถึงภาวะซึมเศร้า (Depression term corpus) (ภาคผนวก ก) ซึ่งจำนวนรีทวีตและจำนวนแฮชแท็ก ใช้ขั้นตอนในการคำนวนเหมือนกับการคำนวนหาจำนวนทวีต

โดยการคำนวณคะแนนจะทำให้สามารถจำแนกคุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ 10 คุณลักษณะ ได้แก่ จำนวนทวีตด้านบวก จำนวนทวีตด้านลบ จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า จำนวนรีทวีตด้านบวก จำนวนรีทวีตด้านลบ จำนวนรีทวีตที่แสดงถึงภาวะซึมเศร้า จำนวนแฮชแท็กด้านบวก จำนวนแฮชแท็กด้านลบ จำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า และค่าเฉลี่ยคะแนนความรู้สึกของทวีต

3.3) การสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon) เพื่อใช้จำแนกอารมณ์และความรู้สึก โดยใช้การเปรียบเทียบข้อมูลจากฐานข้อมูล Emoji Sentiment Ranking ดังภาพที่ 3.11 ซึ่งฐานข้อมูล Emoji Sentiment Ranking จะมีค่าคะแนนความรู้สึก (Sentiment score) ของแต่ละสัญรูปอารมณ์ เพื่อใช้ในการจำแนกสัญรูปอารมณ์ด้านบวก และด้านลบ โดยค่าคะแนนความรู้สึก ***>0*** คือสัญรูปอารมณ์ด้านบวก ค่าคะแนนความรู้สึก $\leq$***0*** คือสัญรูปอารมณ์ด้านลบ

มหาวิทยาลัยเทคโนโลยีสุรนารี

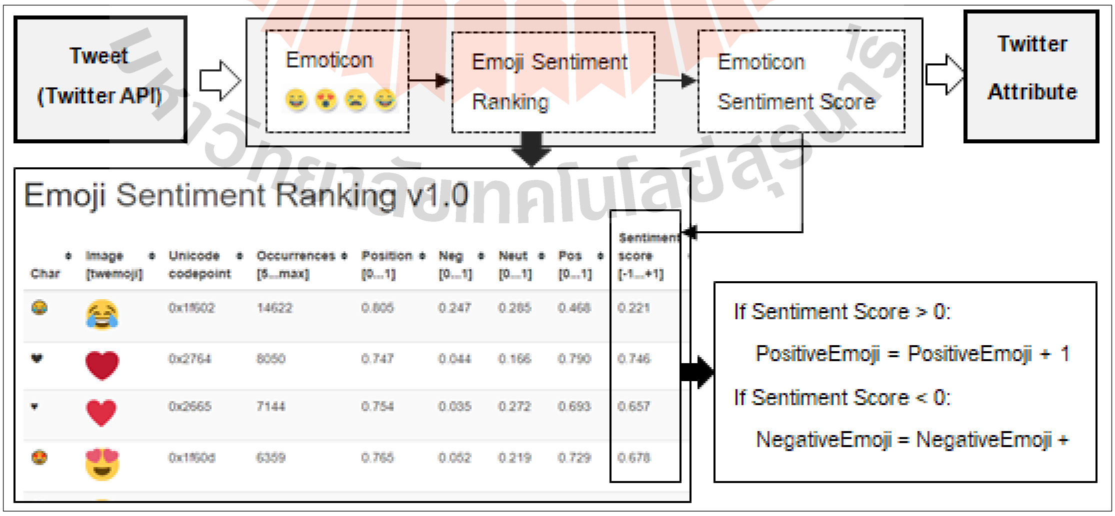
Emoji Sentiment Ranking v1.0

Char	Image [twemoji]	Unicode codepoint	Occurrences [5...max]	Position [0...1]	Neg [0...1]	Neut [0...1]	Pos [0...1]	Sentiment score [-1...+1]	Sentiment bar (c.i. 95%)	Unicode name	Unicode block
😂	😂	0x1f602	14622	0.805	0.247	0.285	0.468	0.221		FACE WITH TEARS OF JOY	Emoticons
❤	❤	0x2764	8050	0.747	0.044	0.166	0.790	0.746		HEAVY BLACK HEART	Dingbats
♥	♥	0x2665	7144	0.754	0.035	0.272	0.693	0.657		BLACK HEART SUIT	Miscellaneous Symbols
😍	😍	0x1f60d	6359	0.765	0.052	0.219	0.729	0.678		SMILING FACE WITH HEART-SHAPED EYES	Emoticons
😭	😭	0x1f62d	5526	0.803	0.436	0.220	0.343	-0.093		LOUDLY CRYING FACE	Emoticons
😘	😘	0x1f618	3648	0.854	0.053	0.193	0.754	0.701		FACE THROWING A KISS	Emoticons
😊	😊	0x1f60a	3186	0.813	0.060	0.237	0.704	0.644		SMILING FACE WITH SMILING EYES	Emoticons

ภาพที่ 3.11 ฐานข้อมูล Emoji Sentiment Ranking

ซึ่งขั้นตอนในการสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์แสดงได้ดังภาพที่ 3.12 และคุณลักษณะที่สกัดได้จากสัญรูปอารมณ์ประกอบด้วย 3 คุณลักษณะ ได้แก่ ค่าเฉลี่ยคะแนนความรู้สึกของสัญรูปอารมณ์ จำนวนสัญรูปอารมณ์ด้านบวก และจำนวนสัญรูปอารมณ์ด้านลบ โดยค่าเฉลี่ยคะแนนความรู้สึกของสัญรูปอารมณ์ *Average Setiment Score* คำนวณได้จากค่าเฉลี่ยของ *Sentiment Score* ดังสมการที่ 3.4 โดย n คือจำนวนสัญรูปอารมณ์

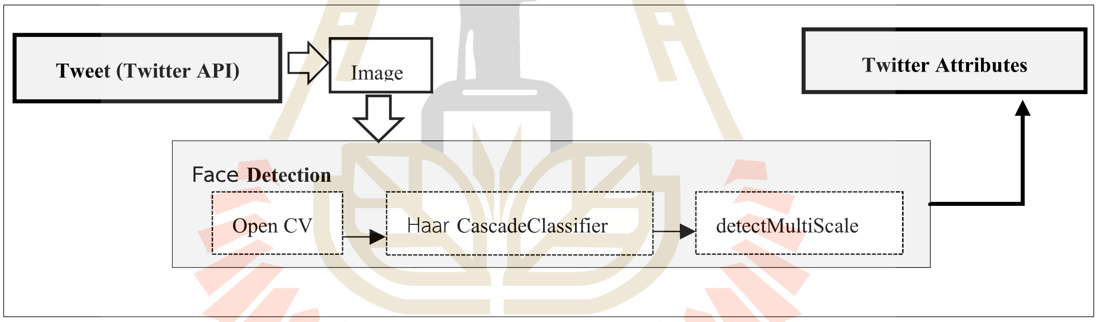
$$Average\ Setiment\ Score = \frac{\sum Sentiment\ Score}{n} \quad (3.4)$$



ภาพที่ 3.12 การสกัดคุณลักษณะจากข้อมูลที่เป็นสัญรูปอารมณ์ในทวิตเตอร์

3.4) การสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) เป็นการสกัดคุณลักษณะจากรูปภาพที่ได้จากการโพสต์ในทวิตเตอร์ โดยแบ่งเป็น 2 ส่วนคือ 1) การนับจำนวนบุคคลในภาพ โดยใช้ซอฟต์แวร์การตรวจจับใบหน้า (Face Detection) และ 2) การวิเคราะห์ค่าเฉลี่ยระดับพิกเซล (Pixel-level averages) ที่ได้จากการคำนวณค่าสี 3 รูปแบบคือ เฉดสี (Hue) ความอิ่มตัวของสี (Saturation) และความสว่างของสี (Value) ในการทำการวิเคราะห์รูปภาพ (Image Analysis) ซึ่งเฉดสี (Hue) อธิบายถึงสีของภาพ ความอิ่มตัวของสี (Saturation) แสดงถึงความสดใส และ ความสว่างของสี (Value) บ่งบอกถึงความสว่างของรูปภาพ (Reece and Danforth, 2017)

- การตรวจจับใบหน้า (Face Detection) ใช้สำหรับสกัดคุณลักษณะ 3 คุณลักษณะ ได้แก่ จำนวนรูปภาพที่ไม่มีคน จำนวนรูปภาพที่มีคน 1 คน จำนวนรูปภาพที่มีคนมากกว่า 1 คน โดยใช้การตรวจจับใบหน้าด้วยตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ (Haar feature-based cascade classifiers) (Viola and Jones, 2001) เป็นวิธีการตรวจจับวัตถุที่มีประสิทธิภาพซึ่งเป็นส่วนหนึ่งของไลบรารีโอเพนซีวี (OpenCV) ซึ่งขั้นตอนการสกัดคุณลักษณะจำนวนรูปภาพจากการตรวจจับใบหน้าแสดงได้ดังภาพที่ 3.13



ภาพที่ 3.13 ขั้นตอนการสกัดคุณลักษณะจำนวนรูปภาพจากการตรวจจับใบหน้า

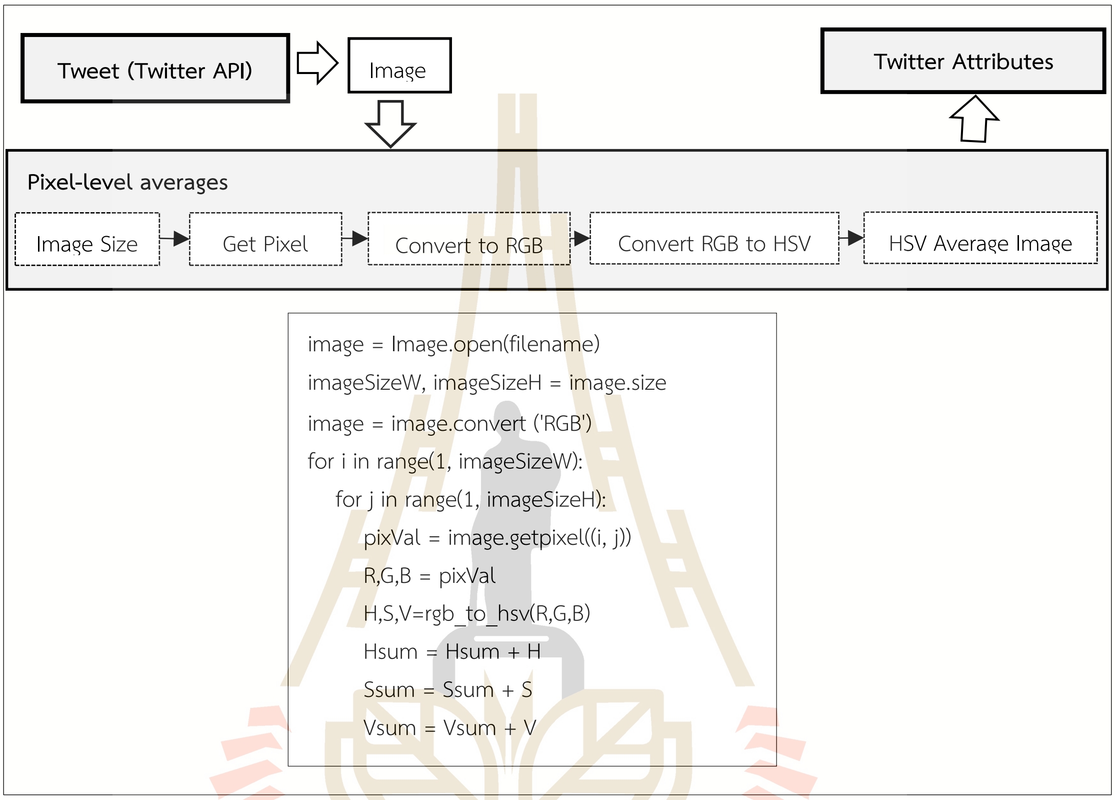
กระบวนการตรวจจับใบหน้า (Face Detection) ประกอบด้วย 3 ส่วน ได้แก่

1) การใช้งานโอเพนซีวี (Open CV) สำหรับการนำเข้ารูปภาพเพื่อใช้ในการประมวลผล

2) การรวมตัวจำแนกกลุ่มแบบต่อเรียงของฮาร์ (Haar CascadeClassifier) คือการเรียกใช้ตัวแบบการตรวจจับใบหน้าด้วยตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ (Haar feature-based cascade classifiers) ผ่านไฟล์ haarcascade_frontalface_default.xml

3) ฟังชันก์ ดีเทคมัลติสเกล (detectMutiScale) ซึ่งเป็นการตรวจหาตำแหน่งของใบหน้าด้วยฟังชันก์ detectMutiScale จากรูปภาพที่นำเข้ามา

- การวิเคราะห์ค่าเฉลี่ยระดับพิกเซล (Pixel-level averages) ใช้สำหรับสกัดคุณลักษณะ 2 คุณลักษณะ ได้แก่ จำนวนรูปภาพสีสดใส และจำนวนรูปภาพสีหม่น โดยใช้ไพทอนอิมเมจไลบรารี (Python Image Library: PIL Image) ในการหาค่าเฉลี่ยของเฉดสี (Hue: H) ความอิ่มตัวของสี (Saturation: S) และความสว่างของสี (Value: V) ซึ่งขั้นตอนวิธีแสดงได้ดังภาพที่ 3.14



ภาพที่ 3.14 ขั้นตอนวิธีการวิเคราะห์ค่าเฉลี่ยระดับพิกเซล

เมื่อได้ค่าเฉดสี (Hue: H) ความอิ่มตัวของสี (Saturation: S) และความสว่างของสี (Value: V) ในระดับพิกเซลแล้ว จึงนำไปคำนวณหาค่าสีเฉลี่ยของภาพ จากสมการที่ 3.5 ถึง 3.7

$$Pixel\ Average\ Hue = \frac{\sum Hue}{Image\ Size} \quad (3.5)$$

$$Pixel\ Average\ Saturation = \frac{\sum Saturation}{Image\ Size} \quad (3.6)$$

$$Pixel\ Average\ Value = \frac{\sum Value}{Image\ Size} \quad (3.7)$$

มหาวิทยาลัยเทคโนโลยีสุรนารี

หลังจากนั้นจึงนำค่าเฉลี่ยความอิ่มตัวของสี (*Pixel Average Saturation*) และค่าเฉลี่ยความสว่างของสี (*Pixel Average Value*) ไปใช้สำหรับวิเคราะห์รูปภาพสีสดใสและรูปภาพสีหม่น โดยมีเงื่อนไขดังนี้

```
If Pixel Average Value >=50 and Pixel Average Saturation >=50:
    Image is Colorful image
Else:
    Image is Dull image
```

หากค่าเฉลี่ยความอิ่มตัวของสี (*Pixel Average Saturation*) และค่าเฉลี่ยความสว่างของสี (*Pixel Average Value*) มีค่ามากกว่าหรือเท่ากับ 50 แสดงว่าภาพนั้นเป็นภาพสีสดใน ถ้าค่าเฉลี่ยความอิ่มตัวของสี (*Pixel Average Saturation*) หรือค่าเฉลี่ยความสว่างของสี (*Pixel Average Value*) มีค่าน้อยกว่า 50 แสดงว่าภาพนั้นเป็นภาพสีหม่น

จากขั้นตอนการสกัดคุณลักษณะของข้อมูล (Feature Extraction) ซึ่งเป็นการสกัดคุณลักษณะจากข้อมูล 3 ส่วน คือ 1) ข้อมูลด้านกายภาพส่วนบุคคล (Demographic Characteristics) ซึ่งเป็นข้อมูลที่ได้จากการทำแบบสอบถามข้อมูลพื้นฐาน และแบบประเมินภาวะซึมเศร้า 9 คำถาม 2) ข้อมูลจากทวิตเตอร์ (Twitter user’s Information) และ 3) ข้อมูลระดับคะแนนภาวะซึมเศร้า (Depression Assessment Score) โดยข้อมูลคุณลักษณะที่ได้จากการขั้นตอนการสกัดคุณลักษณะของข้อมูลทั้งหมด แสดงได้ดังตารางที่ 3.3

ตารางที่ 3.3 คุณลักษณะของข้อมูลในไฟล์ข้อมูลที่ใช้สำหรับการสร้างแบบจำลอง

ID	**ชื่อแอทริบิวต์**	**ชนิดข้อมูล**	**คำอธิบาย**
คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics)			
X_1	Gender	Nominal	เพศ 0: หญิง 1: ชาย 2: อื่นๆ
X_2	Age	Nominal	อายุ 1: <20, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: ≥60 ปี
X_3	Weight	Nominal	น้ำหนัก 1: <40, 2: 40-59, 3: 60-79, 4: 80-99, 5: ≥100 กิโลกรัม
X_4	Education	Nominal	ระดับการศึกษา 1: มัธยม 2: ปริญญาตรี 3: สูงกว่าปริญญาตรี
X_5	CongenitalDisease	Nominal	โรคประจำตัว 1:ไม่มี 2:มี

ตารางที่ 3.3 คุณลักษณะของข้อมูลในไฟล์ข้อมูลที่ใช้สำหรับการสร้างแบบจำลอง (ต่อ)

ID	ชื่อแอทริบิวต์	ชนิดข้อมูล	คำอธิบาย
X_6	Career	Nominal	อาชีพ 1:ราชการ 2:รัฐวิสาหกิจ 3: ค้าขาย/เอกชน 4: นักเรียน/นักศึกษา 5: แม่บ้าน/พ่อบ้าน 6:ว่างงาน
X_7	Income	Nominal	รายได้ 1: เพียงพอ 2: ไม่เพียงพอ
X_8	FamilyMember	Numeric	จำนวนสมาชิกในครอบครัว
X_9	CoupleStatus	Nominal	สถานภาพ 1:โสด 2: อยู่ด้วยกัน 3: แยกกันอยู่
X_{10}	ParentStatus	Nominal	สถานภาพ (บิดา-มารดา) 1:โสด 2: อยู่ด้วยกัน 3: แยกกันอยู่
คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information)			
X_{11}	Friends	Numeric	จำนวนเพื่อน
X_{12}	Follower	Numeric	จำนวนผู้ติดตาม
X_{13}	TweetPeriodAM	Numeric	จำนวนการทวีตที่อยู่ระหว่างเวลา 6.00 น.-00.00 น.
X_{14}	TweetPeriodPM	Numeric	จำนวนการทวีตที่อยู่ระหว่างเวลา 0.00 น.- 6.00 น.
คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text)			
X_{15}	PositiveTweets	Numeric	จำนวนทวีตด้านบวก
X_{16}	NegativeTweets	Numeric	จำนวนทวีตด้านลบ
X_{17}	DepressionTweets	Numeric	จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า
X_{18}	PositiveReTweets	Numeric	จำนวนรีทวีตด้านบวก
X_{19}	NegativeReTweets	Numeric	จำนวนรีทวีตด้านลบ
X_{20}	DepressionReTweets	Numeric	จำนวนรีทวีตที่แสดงถึงภาวะซึมเศร้า
X_{21}	PositiveHashtags	Numeric	จำนวนแฮชแท็กด้านบวก
X_{22}	NegativeHashtags	Numeric	จำนวนแฮชแท็กด้านลบ
X_{23}	DepressionHashtags	Numeric	จำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า
X_{24}	SentimentScore	Numeric	ค่าเฉลี่ยคะแนนความรู้สึกของทวีต
คุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon)			
X_{25}	EmojiSentimentScore	Numeric	ค่าเฉลี่ยคะแนนความรู้สึกของสัญรูปอารมณ์
X_{26}	PositiveEmoji	Numeric	จำนวนสัญรูปอารมณ์ ด้านบวก
X_{27}	NegativeEmoji	Numeric	จำนวนสัญรูปอารมณ์ด้านลบ
คุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image)			
X_{28}	ColorfulImage	Numeric	จำนวนรูปภาพสีสดใส
X_{29}	DullImage	Numeric	จำนวนรูปภาพสีหม่น
X_{30}	NoPersonImage	Numeric	จำนวนรูปภาพที่ไม่มีคน

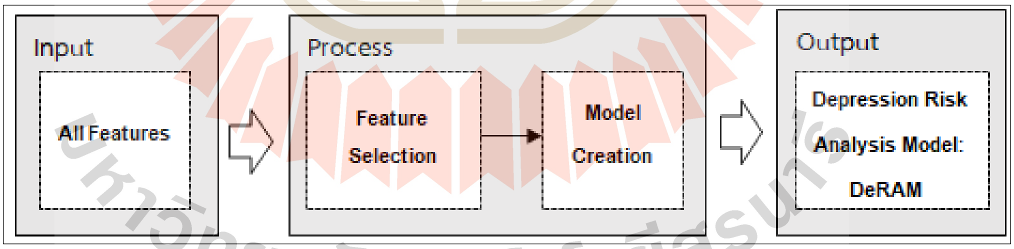
มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 3.3 คุณลักษณะของข้อมูลในไฟล์ข้อมูลที่ใช้สำหรับการสร้างแบบจำลอง (ต่อ)

ID	**ชื่อแอทริบิวต์**	**ชนิดข้อมูล**	**คำอธิบาย**
X_{31}	OnePersonImage	Numeric	จำนวนรูปภาพที่มีคน 1 คน
X_{32}	ManyPersonImage	Numeric	จำนวนรูปภาพที่มีคนมากกว่า 1 คน
ผลการประเมินภาวะซึมเศร้า (Depression Assessment Score)			
$\boldsymbol{Y}$	depression	Nominal	ภาวะซึมเศร้า Level0: ไม่มีภาวะซึมเศร้า Level1: ภาวะซึมเศร้าระดับน้อย Level2 : ภาวะซึมเศร้าระดับปานกลาง Level3 : ภาวะซึมเศร้าระดับรุนแรง

3.1.2.3 การสร้างแบบจำลองที่ใช้ในการพยากรณ์ (Model Construction)

ในขั้นตอนนี้เป็นขั้นตอนการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง ประกอบด้วย 2 ส่วน ได้แก่ 1) การคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection) และ 2) การสร้างแบบจำลอง (Model Creation) โดยมีข้อมูลนำเข้าที่ใช้ในการสร้างแบบจำลองคือ คุณลักษณะของข้อมูลทั้งหมด (All Features) ที่ได้จากขั้นตอนการสกัดคุณลักษณะของข้อมูล (Feature Extraction) และผลลัพธ์จากกระบวนการนี้คือ แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (Depression Risk Analysis Model: DeRAM) ดังแสดงในภาพที่ 3.15



ภาพที่ 3.15 ขั้นตอนการสร้างแบบจำลองที่ใช้ในการพยากรณ์ (Model Construction)

1) การคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection) ในงานวิจัยนี้ใช้เทคนิคในการคัดเลือกคุณลักษณะที่เหมาะสม 3 เทคนิค ได้แก่ 1) การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (Recursive Feature Elimination and Support Vector Machine: RFE-SVM) 2) การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA) และ 3) การคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) หลังจากนั้นจึงทำการประเมินประสิทธิภาพเพื่อหาเทคนิคใน

การคัดเลือกคุณลักษณะที่เหมาะสมที่สุด โดยใช้การเปรียบเทียบค่าประสิทธภาพโดยรวม (F-Measure) จากเทคนิคการเรียนรู้ของเครื่อง 5 เทคนิค ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes: NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) 3) ต้นไม้การตัดสินใจ (Decision Tree (J48): DT) 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest: RF)

2) การสร้างแบบจำลอง (Model Creation) โดยใช้ข้อมูลการใช้งานทวิตเตอร์ของผู้ใช้ โดยมีตัวแปรนำเข้า (Input Variables) คือ $X_1 - X_{32}$ และตัวแปรค่าเป้าหมาย (Target Variable) คือ Y ด้วยวิธีการไขว้ทบ 10 ส่วน (10-fold cross validation) เนื่องจากข้อมูลกลุ่มตัวอย่างมีความไม่สมดุลกันของข้อมูล ในชุดข้อมูลจำนวน 405 คน แบ่งเป็น 4 กลุ่มตามตัวแปรค่าเป้าหมาย ได้แก่ ผู้ที่ไม่มีภาวะซึมเศร้า (level0) จำนวน 154 คน มีภาวะซึมเศร้าระดับน้อย (level1) จำนวน 149 คน มีภาวะซึมเศร้าระดับปานกลาง (level2) จำนวน 70 คน และมีภาวะซึมเศร้าระดับรุนแรง (level3) จำนวน 32 คน ดังนั้นจึงนำเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีสุ่ม (Oversampling) มาใช้ในการจัดการความไม่สมดุลของข้อมูล (Imbalanced data) ซึ่งเกิดขึ้นเมื่อข้อมูลที่ใช้งานมีจำนวนข้อมูลในแต่ละกลุ่มแตกต่างกัน ทำให้ผลลัพธ์จากการจำแนกข้อมูลมีความโน้มเอียง ซึ่งเทคนิคการปรับเพิ่มข้อมูลด้วยวิธีการสุ่มในงานวิจัยนี้ใช้เทคนิค SMOTE (Synthetic Minority Over-Sampling Technique) เป็นเทคนิคที่ใช้ในการแก้ปัญหาการจำแนกข้อมูลที่ไม่สมดุล โดยเป็นการเพิ่มจำนวนข้อมูลกลุ่มน้อยให้มีจำนวนเพิ่มขึ้น ด้วยการสุ่มค่าข้อมูลที่อยู่ในกลุ่มข้อมูลน้อยขึ้นมา 1 ค่า หลังจากนั้นจึงพิจารณาค่าข้อมูลข้างเคียงอีกจำนวน K ค่า (K-Nearest Neighbor) แล้วคำนวณระยะทาง (Euclidean Distance) ระหว่างค่าที่สุ่มกับค่าข้อมูลใกล้เคียงแต่ละค่าเพื่อหาระยะทางที่น้อยที่สุด ระหว่างค่าสุ่มกับค่าข้อมูลใกล้เคียงหลังจากนั้นจึงสร้างข้อมูลเทียมระหว่างค่าข้อมูลที่สุ่มกับค่าข้อมูลใกล้เคียงที่ให้ค่าระยะทางน้อยที่สุด เพื่อประเมินแบบจำลอง ซึ่งการทดลองนี้ได้ใช้ Python Library สำหรับการประมวลผลการทดลอง และใช้ Scikit-learn Library ร่วมกับ Keras Library และ Tensorflow Library สำหรับการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง 5 เทคนิค ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes: NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) โดยใช้เคอร์เนลฟังก์ชัน (Kernel Function) คือเรเดียลเบสิสฟังก์ชัน (Radial Basis Function: RBF) 3) ต้นไม้การตัดสินใจ (Decision Tree: DT) โดยใช้ขั้นตอนวิธี C4.5 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) โดยมีชั้นซ่อน (Hidden Layers) คือ 100 และฟังก์ชั้นกระตุ้น (Activation Function) คือ ReLu และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest: RF) โดยใช้จำนวนต้นไม้ (Tree) คือ 100 ซึ่งเทคนิคการเรียนรู้ของเครื่องทั้ง 5 เทคนิคเป็นเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Machine Learning) และการทดลองนี้ทำงานบนระบบปฏิบัติการวินโดวส์ 10

3.1.2.4 ขั้นตอนการเปรียบเทียบประสิทธิภาพแบบจำลอง (Model Performance Evaluation)

ในขั้นตอนนี้เป็นการเปรียบเทียบประสิทธิภาพของแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (Depression Risk Analysis Model: DeRAM) ที่งานวิจัยพัฒนาขึ้นกับเทคนิคอื่น ๆ ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine (SVM)) 3) ต้นไม้การตัดสินใจ (Decision Tree (J48)) 4) เทคนิคการเรียนรู้เชิงลึก (Deep Learning) และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest)

3.1.3 ทดสอบและปรับปรุงประสิทธิภาพของแบบจำลองที่พัฒนา

การออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ในงานวิจัยนี้ใช้การทดสอบประสิทธิภาพของแบบจำลองในประเด็นความถูกต้องของแบบจำลอง ด้วยการวัดค่ามาตรฐาน 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure)

3.2 ประชากรและกลุ่มตัวอย่าง

3.2.1 ประชากร

ประชากรที่ใช้ในการเก็บรวมรวมข้อมูลเครือข่ายสังคมออนไลน์ทวิตเตอร์ และแบบประเมินภาวะซึมเศร้า 9 คำถาม คือ ผู้ใช้งานเครือข่ายสังคมออนไลน์ทวิตเตอร์ในประเทศไทย จำนวน 6,545,000 ผู้ใช้ (Mahittivanicha, 2020)

3.2.2 กลุ่มตัวอย่าง

กลุ่มตัวอย่างที่ใช้ในการเก็บรวมรวมข้อมูลเครือข่ายสังคมออนไลน์ และแบบประเมินภาวะซึมเศร้า 9 คำถาม คือ ผู้ใช้งานเครือข่ายสังคมออนไลน์ทวิตเตอร์ในประเทศไทย โดยกำหนดขนาดของกลุ่มตัวอย่างจากการคำนวณด้วยวิธิการสุ่มตัวอย่างของทาโร่ ยามาเน่ (Taro Yamane, 1967) ดังสมการที่ 3.8

$$n = \frac{N}{1+Ne^2} \tag{3.8}$$

เมื่อ n คือ ขนาดของกลุ่มตัวอย่าง

N คือ ขนาดของประชากร

e คือ สัดส่วนความคาดเคลื่อนของการสุ่มตัวอย่าง

ซึ่งงานวิจัยนี้กำหนดสัดส่วนความคาดเคลื่อนของการสุ่มตัวอย่างไว้ที่ 5% ดังนั้น จะได้ขนาดของกลุ่มตัวอย่างโดยประมาณ 400 คน โดยในงานวิจัยนี้เก็บข้อมูลจากผู้ใช้งานเครือข่ายออนไลน์ทวิตเตอร์จากกลุ่มตัวอย่างไม่ต่ำกว่า 400 คน

3.3 เครื่องมือที่ใช้ในการวิจัย

เครื่องมือและเทคโนโลยีที่ใช้ในการพัฒนา มีดังต่อไปนี้

3.3.1 ด้านฮาร์ดแวร์ มีคุณสมบัติดังนี้

- หน่วยประมวลผลกลางชนิด: Intel(R) Core(TM) i5 4258U ความถี่ 2.4 GHz
- หน่วยความจำสำรองขนาด 1,000 กิกะไบต์ (1,000 GB)
- หน่วยความจำหลัก 8.0 กิกะไบต์ (8.0 GB)
- อุปกรณ์เสริมอื่น ๆ เช่น เมาส์ แป้นพิมพ์ เครื่องพิมพ์ ฯลฯ

3.3.2 ด้านซอฟต์แวร์ ระบบปฏิบัติการและโปรแกรมประยุกต์สำหรับพัฒนาขั้นตอนวิธีประกอบด้วย

- ระบบปฏิบัติการ: Window 10 Home 64 bit operating system
- เว็บเบราว์เซอร์: Google Chrome
- ภาษาการเขียนโปรแกรม: Python PHP Script Language Version 7.2.5 HTML5 jQuery JSON Python
- เว็บเซิร์ฟเวอร์: Apache Web Server 2.2.8
- ระบบจัดการฐานข้อมูล: MySQL Database Version 5.0.51b และ phpMyAdmin Version 2.10.3
- Google Colaboratory (Colab)

มหาวิทยาลัยเทคโนโลยีสุรนารี

3.4 การเก็บรวบรวมข้อมูล

การเก็บรวบรวมข้อมูลในงานวิจัยครั้งนี้ เก็บรวบรวมด้วยกัน 2 ส่วน คือ ส่วนของการเก็บข้อมูลเพื่อนำมาใช้ในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้า และส่วนของการเก็บข้อมูลเพื่อวัดค่าความถูกต้องของแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า โดยมีรายละเอียดดังนี้

3.4.1 การเก็บข้อมูลเพื่อนำมาใช้ในการสร้างแบบจำลอง

ส่วนของการเก็บข้อมูลเพื่อนำมาใช้ในการสร้างตัวแบบขั้นตอนวิธีการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ในส่วนนี้ได้ใช้ข้อมูลที่มีการเก็บรวบรวมอยู่ในเครือข่ายสังคมออนไลน์ คือ ทวิตเตอร์และข้อมูลจากแบบสอบถามซึ่งสามารถนำมาใช้ในการวิเคราะห์และพัฒนาเครื่องมือได้ โดยข้อมูลที่ได้จากแบบสอบถาม ประกอบด้วยแบบสอบถามข้อมูลส่วนบุคคล และแบบประเมินภาวะซึมเศร้า 9 คำถาม ร่วมกับข้อมูลการใช้งานทวิตเตอร์ โดยใช้การสุ่มตัวอย่างแบบเครือข่าย (Network Sampling หรือ Snowballing) การได้มาซึ่งตัวอย่างชนิดนี้ จะขึ้นอยู่กับเครือข่ายสังคม เช่นการเป็นเพื่อน ความสนิทสนม หรือมีคุณสมบัติที่คล้ายคลึงกันในเครือข่ายนั้น โดยในการสุ่มตัวอย่างสำหรับการเก็บข้อมูลในงานวิจัยนี้ ใช้การเก็บข้อมูลในรูปแบบออนไลน์ผ่านเว็บแอปพลิเคชัน ด้วยการแบ่งปันลิงก์ (Link) ผ่านเครือข่ายสังคมออนไลน์ (Social Network) ซึ่งสอดคล้องกับวัตถุประสงค์ของงานวิจัยนี้

3.4.2 การเก็บข้อมูลเพื่อวัดค่าความถูกต้องของแบบจำลอง

ส่วนของการเก็บข้อมูลค่าความถูกต้องของแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าจากการทดสอบแบบจำลอง

3.5 การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลใช้สำหรับการประเมินความถูกต้องของแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยมีรายละเอียดดังนี้

การประเมินความถูกต้องของแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยใช้วิธีวัดค่ามาตรฐานในการประเมินความถูกต้อง 4 ค่า ได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าประสิทธิภาพโดยรวม (F-measure) ซึ่งมีรายละเอียดดังนี้

ค่าความแม่นยำ (Precision) คือ อัตราส่วนระหว่าง จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม C_j และแบบจำลองทำนายว่าอยู่ในกลุ่ม C_j ต่อจำนวนกลุ่มตัวอย่างทั้งหมดที่แบบจำลองทำนายว่าอยู่ในกลุ่ม C_j ดังสมการที่ 3.9

$$Precision = \frac{TP}{(TP + FP)} \times 100 \tag{3.9}$$

ค่าความระลึก (Recall) คือ อัตราส่วนระหว่าง จำนวนกลุ่มตัวอย่างที่มีผลการทำนายว่าอยู่ในกลุ่ม C_j ต่อจำนวนกลุ่มตัวอย่างทั้งหมดที่อยู่ในกลุ่ม C_j ดังสมการที่ 3.10

$$Recall = \frac{TP}{(TP + FN)} \times 100 \tag{3.10}$$

ค่าความถูกต้อง (Accuracy) คือ อัตราส่วนระหว่าง จำนวนกลุ่มตัวอย่างที่มีผลการทำนายถูกต้อง ต่อจำนวนกลุ่มตัวอย่างทั้งหมด ดังสมการที่ 3.11

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \times 100 \tag{3.11}$$

ค่าประสิทธิภาพโดยรวม (F-measure) คือ การเปรียบเทียบค่าความแม่นยำ และค่าความระลึกของแต่ละคลาสเป้าหมาย จึงเปรียบเสมือนค่าวัดความแม่นยำโดยรวม มีสูตรในการคำนวณดังสมการที่ 3.12

$$F - Measure = \frac{2 \times (Recall \times Precision)}{(Recall + Precision)} \tag{3.12}$$

โดยที่

TP (True Positive) คือ จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม C_j และแบบจำลองทำนายว่าอยู่ในกลุ่ม C_j

TN (True Negative) คือ จำนวนกลุ่มตัวอย่างที่ไม่อยู่ในกลุ่ม C_j และแบบจำลองทำนายว่าไม่อยู่ในกลุ่ม C_j

FP (False Positive) คือ จำนวนกลุ่มตัวอย่างที่ไม่อยู่ในกลุ่ม C_j แต่แบบจำลองทำนายว่าอยู่ในกลุ่ม C_j

มหาวิทยาลัยเทคโนโลยีสุรนารี

FN (False Negative) คือ จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม C_j แต่แบบจำลองทำนายว่าไม่อยู่ในกลุ่ม C_j

C_j คือ กลุ่มของระดับคะแนนจากผลการประเมินภาวะซึมเศร้าที่ทดสอบด้วยแบบทดสอบภาวะซึมเศร้า 9 คำถาม (9Q) 4 ระดับคะแนน เมื่อ $0 \leq j \leq 3$

มหาวิทยาลัยเทคโนโลยีสุรนารี

บทที่ 4
ผลการวิจัยและอภิปรายผล

งานวิจัยนี้ได้แบ่งผลการวิจัยออกเป็น 2 ส่วน ได้แก่ ผลการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์ และผลการประเมินแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์

4.1 ผลการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์

4.1.1 การเก็บรวบรวมข้อมูลจากเครือข่ายสังคมออนไลน์และแบบประเมินภาวะซึมเศร้า 9 คำถาม

4.1.2 การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

4.1.3 การหาเทคนิคในการคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection Techniques)

4.1.4 การคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection)

4.2 ผลการประเมินแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์

4.2.1 การเปรียบเทียบค่าประสิทธิภาพแบบจำลองที่พัฒนา กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) อื่น ๆ

4.2.2 การอภิปรายผล

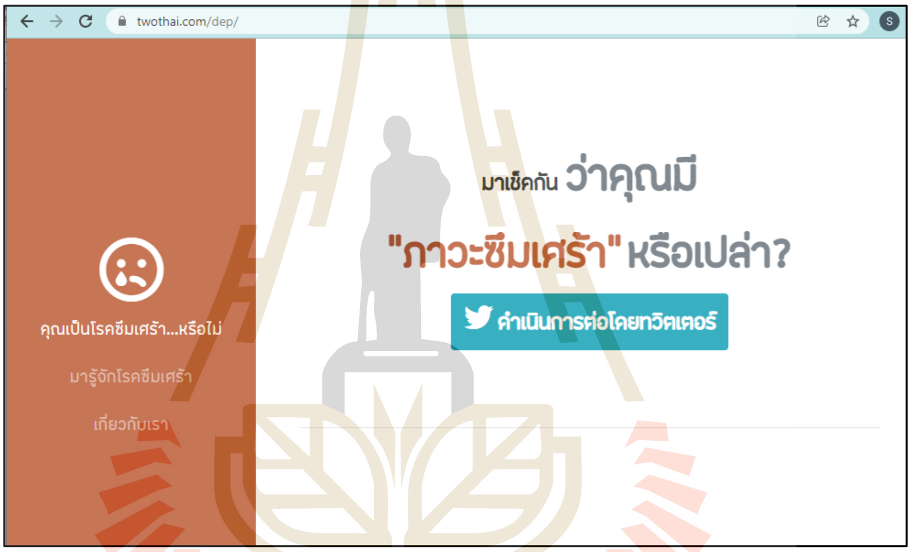
4.1 ผลการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

การพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์มุ่งเน้นที่ขั้นตอนวิธีในการออกแบบแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ที่เกิดจากชุดข้อมูลจำนวน 405 คน แบ่งเป็น 4 กลุ่ม ได้แก่ ผู้ที่ไม่มีภาวะซึมเศร้า (level0) จำนวน 154 คน มีภาวะซึมเศร้าระดับน้อย (level1) จำนวน 149 คน มีภาวะซึมเศร้าระดับปานกลาง (level2) จำนวน 70 คน และมีภาวะซึมเศร้าระดับรุนแรง (level3) จำนวน 32 คน (ภาคผนวก ข) โดยมีข้อมูลการแสดงความเห็นบนทวิตเตอร์

จำนวน 9,417 ครั้ง ประกอบด้วยการทวีต (Tweet) จำนวน 1,099 ครั้ง การรีทวีต (Retweet) จำนวน 8,318 ครั้ง และการแฮชแท็ก (Hashtag) จำนวน 137 แฮชแท็ก โดยผลการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์แบ่งเป็น 5 ส่วน ดังนี้

4.1.1 การเก็บรวบรวมข้อมูลจากเครือข่ายสังคมออนไลน์และแบบประเมินภาวะซึมเศร้า 9 คำถาม

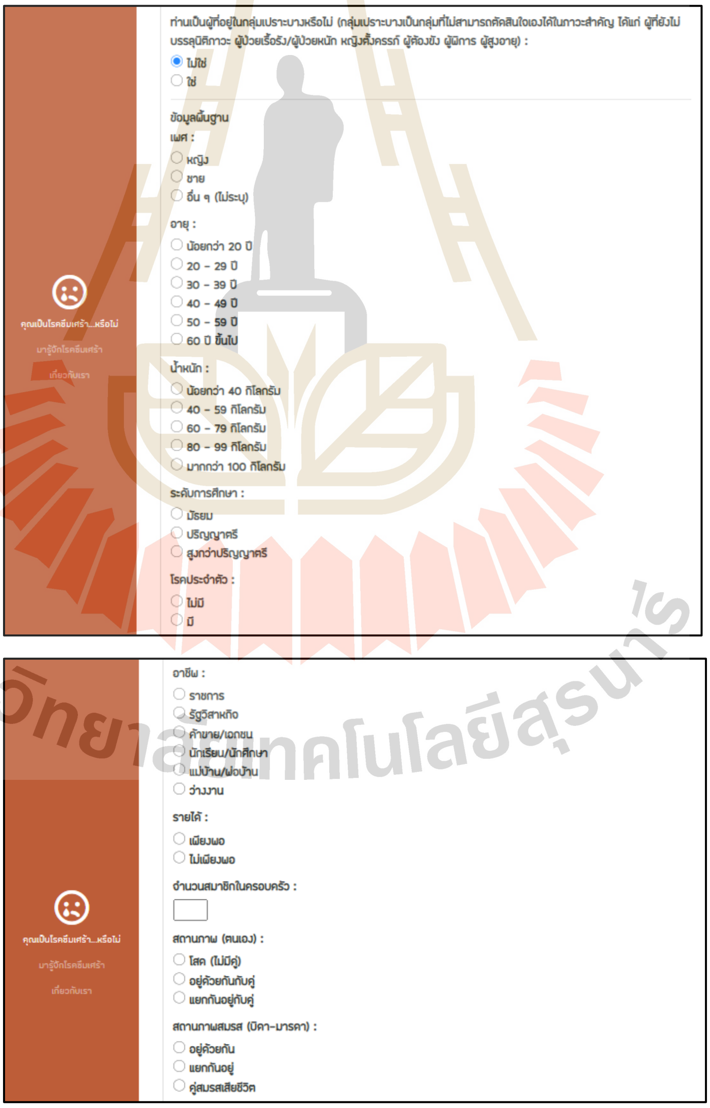
กระบวนการเก็บรวบรวมข้อมูลนั้นได้มีการจัดทำโปรแกรมประยุกต์บนเว็บสำหรับจัดเก็บข้อมูลจากผู้ใช้งานทวิตเตอร์ ดังแสดงในภาพที่ 4.1



ภาพที่ 4.1 โปรแกรมประยุกต์บนเว็บสำหรับจัดเก็บข้อมูล

โดยโปรแกรมประยุกต์บนเว็บที่ใช้สำหรับจัดเก็บข้อมูลจากผู้ใช้งานทวิตเตอร์ แบ่งส่วนจัดเก็บข้อมูลออกเป็น 5 ส่วน ประกอบด้วย ส่วนการเชื่อมต่อกับทวิตเตอร์เอพีไอ ดังแสดงในภาพที่ 4.2 ส่วนของแบบสอบถามข้อมูลพื้นฐาน ดังแสดงในภาพที่ 4.3 ส่วนของแบบประเมินภาวะซึมเศร้า 9 คำถาม ดังแสดงในภาพที่ 4.4 และส่วนของผลจากแบบประเมินภาวะซึมเศร้า 9 คำถาม ดังแสดงในภาพที่ 4.5

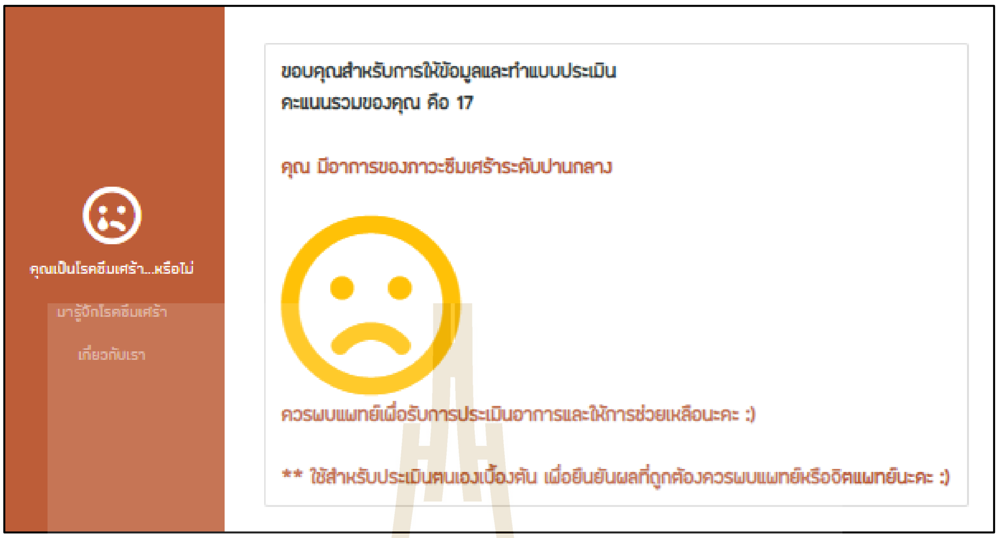
ภาพที่ 4.2 ส่วนการเชื่อมต่อกับทวิตเตอร์เอพีไอ



ภาพที่ 4.3 ส่วนของแบบสอบถามข้อมูลพื้นฐาน

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาพที่ 4.4 ส่วนของแบบประเมินภาวะซึมเศร้า 9 คำถาม



ภาพที่ 4.5 ส่วนของผลจากแบบประเมินภาวะซึมเศร้า 9 คำถาม

4.1.2 การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

จากกระบวนการเก็บรวบรวมข้อมูลจากผู้ใช้งานเครือข่ายสังคมออนไลน์ทวิตเตอร์และแบบประเมินภาวะซึมเศร้า 9 คำถามเมื่อได้ข้อมูลจากการจัดเก็บด้วยโปรแกรมประยุกต์ดังกล่าวแล้ว จึงนำข้อมูลเหล่านั้นมาเข้าสู่กระบวนการสกัดคุณลักษณะซึ่งคุณลักษณะที่สกัดได้ประกอบด้วย 5 ส่วน ดังนี้

1) คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics) ได้จากการทำแบบสอบถามข้อมูลส่วนบุคคลผ่านโปรแกรมประยุกต์บนเว็บ โดยคุณลักษณะด้านกายภาพส่วนบุคคลสามารถแสดงได้ดังตารางที่ 4.1

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.1 คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics)

ID	ชื่อแอทริบิวต์	ชนิดข้อมูล	คำอธิบาย
X_1	Gender	Nominal	เพศ 0: หญิง 1: ชาย 2: อื่นๆ
X_2	Age	Nominal	อายุ 1: <20, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: ≥60 ปี
X_3	Weight	Nominal	น้ำหนัก 1: <40, 2: 40-59, 3: 60-79, 4: 80-99, 5: ≥100 กิโลกรัม
X_4	Education	Nominal	ระดับการศึกษา 1: มัธยม 2: ปริญญาตรี 3: สูงกว่าปริญญาตรี
X_5	CongenitalDisease	Nominal	โรคประจำตัว 1:ไม่มี 2:มี
X_6	Career	Nominal	อาชีพ 1:ราชการ 2:รัฐวิสาหกิจ 3: ค้าขาย/เอกชน 4: นักเรียน/นักศึกษา 5: แม่บ้าน/พ่อบ้าน 6:ว่างงาน
X_7	Income	Nominal	รายได้ 1: เพียงพอ 2: ไม่เพียงพอ
X_8	FamilyMember	Numeric	จำนวนสมาชิกในครอบครัว
X_9	CoupleStatus	Nominal	สถานภาพ 1: โสด 2: อยู่ด้วยกัน 3: แยกกันอยู่
X_{10}	ParentStatus	Nominal	สถานภาพ (บิดา-มารดา) 1: โสด 2: อยู่ด้วยกัน 3: แยกกันอยู่

2) คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information) ได้จากข้อมูลผู้ใช้งานที่ปรากฏในทวิตเตอร์ แสดงได้ดังตารางที่ 4.2

ตารางที่ 4.2 คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information)

ID	ชื่อแอทริบิวต์	ชนิดข้อมูล	คำอธิบาย
X_{11}	Friends	Numeric	จำนวนเพื่อน
X_{12}	Follower	Numeric	จำนวนผู้ติดตาม
X_{13}	TweetPeriodAM	Numeric	จำนวนการทวีตที่อยู่ระหว่างเวลา 6.00 น.-00.00 น.
X_{14}	TweetPeriodPM	Numeric	จำนวนการทวีตที่อยู่ระหว่างเวลา 0.00 น.- 6.00 น.

3) คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) ได้จากการสกัดคุณลักษณะจากการแสดงความคิดเห็นที่เป็นข้อความผ่านทวิตเตอร์ โดยใช้การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) ร่วมกับการวิเคราะห์ความรู้สึก (Sentiment Analysis) หรือจำแนกขั้วอารมณ์ ผลลัพธ์จากการสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความสามารถแสดงได้ดังตารางที่ 4.3

ตารางที่ 4.3 คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text)

ID	**ชื่อแอทริบิวต์**	**ชนิดข้อมูล**	**คำอธิบาย**
X_{15}	PositiveTweets	Numeric	จำนวนทวีตด้านบวก
X_{16}	NegativeTweets	Numeric	จำนวนทวีตด้านลบ
X_{17}	DepressionTweets	Numeric	จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า
X_{18}	PositiveReTweets	Numeric	จำนวนรีทวีตด้านบวก
X_{19}	NegativeReTweets	Numeric	จำนวนรีทวีตด้านลบ
X_{20}	DepressionReTweets	Numeric	จำนวนรีทวีตที่แสดงถึงภาวะซึมเศร้า
X_{21}	PositiveHashtags	Numeric	จำนวนแฮชแท็กด้านบวก
X_{22}	NegativeHashtags	Numeric	จำนวนแฮชแท็กด้านลบ
X_{23}	DepressionHashtags	Numeric	จำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า
X_{24}	SentimentScore	Numeric	ค่าเฉลี่ยคะแนนความรู้สึกของทวีต

4) คุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon) ได้จากการสกัดคุณลักษณะที่เป็นสัญรูปอารมณ์ ผ่านทวิตเตอร์ โดยใช้การเปรียบเทียบข้อมูลจากฐานข้อมูล Emoji Sentiment Ranking ในการจำแนกสัญรูปอารมณ์ด้านบวก และด้านลบ ผลลัพธ์จากการสกัดคุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์สามารถแสดงได้ดังตารางที่ 4.4

ตารางที่ 4.4 คุณลักษณะจากทวิตเตอร์ที่เป็นเป็นสัญรูปอารมณ์ (Emoticon)

ID	ชื่อแอทริบิวต์	ชนิดข้อมูล	คำอธิบาย
X_{25}	EmojiSentimentScore	Numeric	ค่าเฉลี่ยคะแนนความรู้สึกของสัญรูปอารมณ์
X_{26}	PositiveEmoji	Numeric	จำนวนสัญรูปอารมณ์ ด้านบวก
X_{27}	NegativeEmoji	Numeric	จำนวนสัญรูปอารมณ์ด้านลบ

5) คุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) ได้จากการสกัดคุณลักษณะของรูปภาพจากทวิตเตอร์ ด้วยการนับจำนวนบุคคลในภาพ โดยใช้ซอฟต์แวร์การตรวจจับใบหน้า (Face Detection) (Cohn et al., 2009) และการวิเคราะห์ค่าเฉลี่ยระดับพิกเซล (Pixel-level averages) ที่ได้จากการคำนวณค่าสี 3 รูปแบบคือ เฉดสี (Hue) ความอิ่มตัวของสี (Saturation) และความสว่างของสี (Value) โดยใช้ไซพาย (SciPy) ในภาษาไพธอน (Reece and Danforth, 2017)ในการทำการวิเคราะห์รูปภาพ (Image Analysis) ซึ่งเฉดสี (Hue) อธิบายถึงสีของภาพ ความอิ่มตัวของสี (Saturation) แสดงถึงความสดใส และความสว่างของสี (Value) บ่งบอกถึงความสว่างของรูปภาพ ซึ่งผลลัพธ์จากการสกัดคุณลักษณะจากทวิตเตอร์ที่รูปภาพสามารถแสดงได้ดังตารางที่ 4.5

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.5 คุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image)

ID	ชื่อแอทริบิวต์	ชนิดข้อมูล	คำอธิบาย
X_{28}	ColorfulImage	Numeric	จำนวนรูปภาพสีสดใส
X_{29}	DullImage	Numeric	จำนวนรูปภาพสีหม่น
X_{30}	NoPersonImage	Numeric	จำนวนรูปภาพที่ไม่มีคน
X_{31}	OnePersonImage	Numeric	จำนวนรูปภาพที่มีคน 1 คน
X_{32}	ManyPersonImage	Numeric	จำนวนรูปภาพที่มีคนมากกว่า 1 คน

4.1.3 การหาเทคนิคในการคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection Techniques)

ในงานวิจัยนี้ใช้เทคนิคในการคัดเลือกคุณลักษณะที่เหมาะสม 3 เทคนิค ได้แก่ 1) การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (Recursive Feature Elimination and Support Vector Machine: RFE-SVM) 2) การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA) และ 3) การคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection)

4.1.3.1 การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (Recursive Feature Elimination and Support Vector Machine: RFE-SVM) คือใช้วิธีการวนรอบจำกัดเอาคุณลักษณะ (Feature) ออกไป (กอบเกียรติ สระอุบล, 2563) โดยคุณลักษณะที่สกัดได้จากขั้นตอนที่แล้วมีทั้งหมด 32 คุณลักษณะ ($X_1 - X_{32}$) จากนั้นตัดออกทีละ 1 คุณลักษณะ เหลือ 31 คุณลักษณะ (RFE-SVM(31)) จนตัดครบ 17 คุณลักษณะ เหลือ 15 คุณลักษณะ (RFE-SVM(15)) เพื่อหารอบการตัดที่ให้ค่าประสิทธิภาพสูงที่สุด เมื่อตัดคุณลักษณะเหลือ 21 คุณลักษณะ (RFE-SVM(21)) พบว่าค่าประสิทธิภาพมีแนวโน้มลดลงเรื่อย ๆ จึงตัดคุณลักษณะให้เหลือ 15 คุณลักษณะ (RFE-SVM(15)) เพื่อให้ตรงกับการคัดเลือกคุณลักษณะด้วยเทคนิควิธีอื่น ๆ โดยคุณลักษณะที่ตัดออกในแต่ละรอบแสดงได้ดังตารางที่ 4.6 หลังจากตัดคุณลักษณะออกแล้วจึงทำการประเมินประสิทธิภาพเพื่อหารอบการตัดคุณลักษณะที่เหมาะสมที่สุด โดยใช้การเปรียบเทียบค่าประสิทธภาพโดยรวม (F-Measure) จากเทคนิคการเรียนรู้ของเครื่อง 5 เทคนิค ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes: NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine : SVM) 3) ต้นไม้การตัดสินใจ (Decision Tree (J48): DT) 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest: RF)

จากกระบวนการนี้พบว่า เทคนิคการสุ่มป่าไม้ให้ค่าประสิทธิภาพโดยรวมสูงที่สุดในทุกรอบการตัดคุณลักษณะ ดังแสดงในตารางที่ 4.7 และเมื่อพิจารณาการตัดคุณลักษณะในรอบที่ 3 ซึ่งเป็นการตัดคุณลักษณะออก 3 คุณลักษณะ (RFE-SVM(29)) พบว่าให้ค่าประสิทธิภาพโดยรวมสูงที่สุด (F-Measure=0.8868) เมื่อใช้งานร่วมกับเทคนิคการสุ่มป่าไม้ ดังแสดงในภาพที่ 4.6 ซึ่งคุณลักษณะที่ตัดออก ประกอบด้วย จำนวนเพื่อน (X_{11} : Friend) จำนวนผู้ติดตาม (X_{12} : Followers) และจำนวน Emoticon ด้านบวก (X_{26} : PositiveEmoji)

ตารางที่ 4.6 ผลการคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (RFE-SVM)

รอบที่	**จำนวนคุณลักษณะ**	**คุณลักษณะที่ตัดออก**
1	RFE-SVM(31)	['Friends']
2	RFE-SVM(30)	['Friends', 'Followers']
3	RFE-SVM(29)	['Friends', 'Followers', 'PositiveEmoji']
4	RFE-SVM(28)	['Friends', 'Followers', 'Negative ReTweets', 'PositiveEmoji']
5	RFE-SVM(27)	['Friends', 'Followers', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji']
6	RFE-SVM(26)	['Friends', 'Followers', 'TweetPeriodAM', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji']
7	RFE-SVM(25)	['Friends', 'Followers', 'TweetPeriodAM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji']
8	RFE-SVM(24)	['Friends', 'Followers', 'TweetPeriodAM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage']
9	RFE-SVM(23)	['Friends', 'Followers', 'TweetPeriodAM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage', 'DullImage']
10	RFE-SVM(22)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage', 'DullImage']
11	RFE-SVM(21)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage', 'DullImage']
12	RFE-SVM(20)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'OnePersonImage', 'ManyPersonImage', 'DullImage']
13	RFE-SVM(19)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']

ตารางที่ 4.6 ผลการคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (RFE-SVM) (ต่อ)

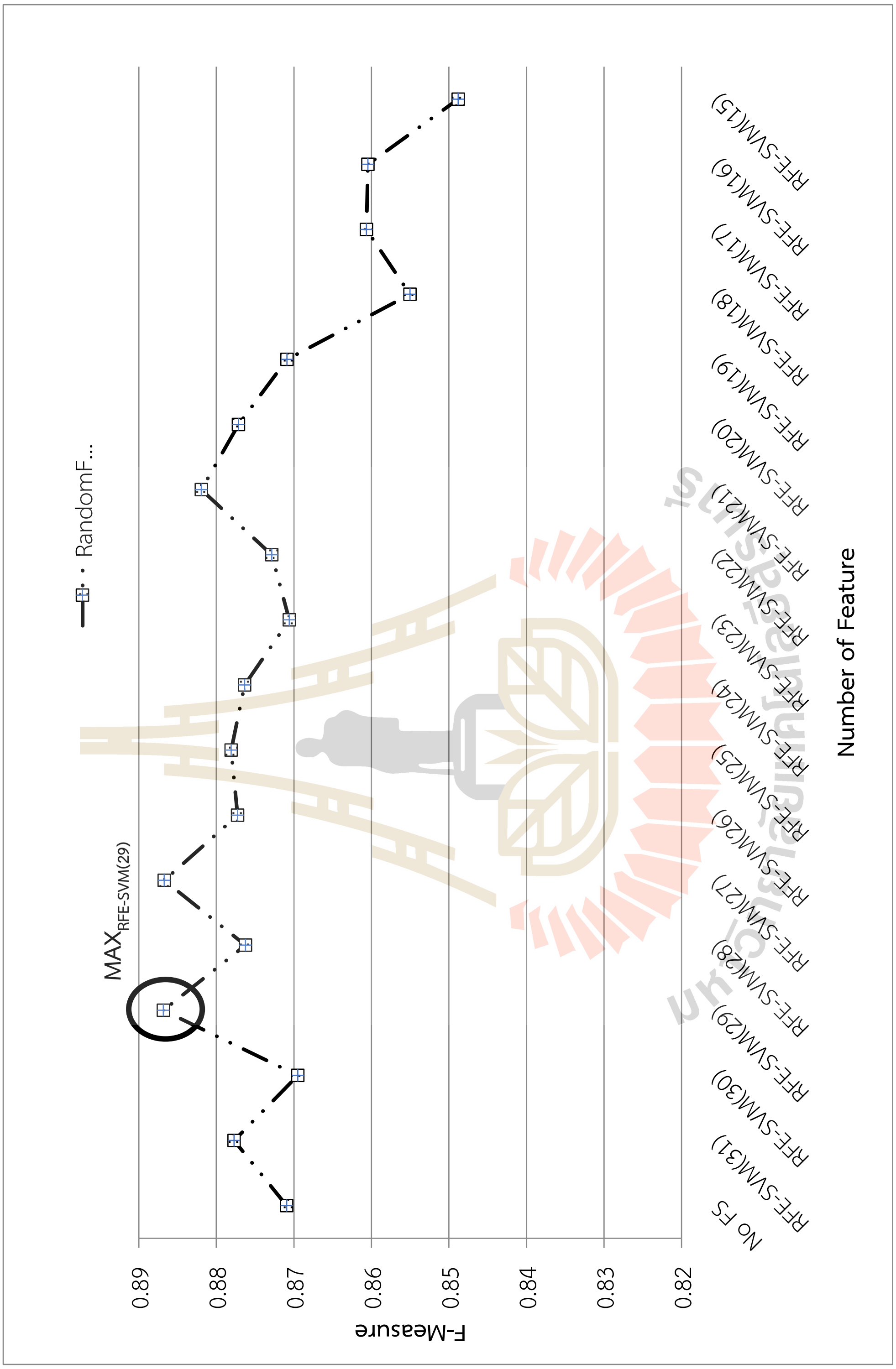
รอบที่	จำนวนคุณลักษณะ	คุณลักษณะที่ตัดออก
14	RFE-SVM(18)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
15	RFE-SVM(17)	['FamilyMember', 'Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
16	RFE-SVM(16)	['Carrer', 'FamilyMember', 'Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
17	RFE-SVM(15)	['Carrer', 'FamilyMember', 'Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Depression Tweets', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'F2', 'DullImage', 'ColorfulImage']

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.7 การเปรียบเทียบผลการคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ

F-Measure	**Model**				
	SVM	DT	NB	**RF**	MLP
No FS	0.79877	0.8608	0.39805	**0.87094**	0.84433
RFE-SVM(31)	0.83785	0.83373	0.39572	**0.87773**	0.866
RFE-SVM(30)	0.84475	0.82543	0.35445	**0.86951**	0.84785
RFE-SVM(29)	0.85943	0.84187	0.35382	**0.88684**	0.85146
RFE-SVM(28)	0.85887	0.8538	0.35262	**0.87624**	0.8545
RFE- SVM(27)	0.86413	0.85635	0.35538	**0.88671**	0.8292
RFE- SVM(26)	0.85594	0.86715	0.34619	**0.87725**	0.85899
RFE- SVM(25)	0.84243	0.85466	0.33022	**0.8781**	0.83597
RFE- SVM(24)	0.83806	0.84497	0.31006	**0.87637**	0.79422
RFE-SVM(23)	0.81834	0.85528	0.2936	**0.87058**	0.80727
RFE- SVM(22)	0.83072	0.85748	0.29022	**0.87285**	0.81162
RFE-SVM(21)	0.84451	0.86139	0.29448	**0.88193**	0.79725
RFE- SVM(20)	0.82765	0.85542	0.29699	**0.87715**	0.77307
RFE- SVM(19)	0.75171	0.85466	0.29468	**0.87089**	0.73444
RFE- SVM(18)	0.69119	0.80666	0.30013	**0.85501**	0.73384
RFE- SVM(17)	0.59707	0.8158	0.30046	**0.86062**	0.67546
RFE- SVM(16)	0.60187	0.82919	0.29961	**0.86047**	0.67493
RFE- SVM(15)	0.61425	0.81838	0.29961	**0.84882**	0.65024

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 4.6 การประเมินประสิทธิภาพเพื่อหาคุณลักษณะที่ควรตัดออกเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (RFE-SVM)

4.1.3.2 การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA) เป็นการทดสอบสมมติฐานทางสถิติเชิงพาราเมตริกเพื่อพิจารณาว่าค่าเฉลี่ยจากตัวอย่างข้อมูลตั้งแต่สามตัวอย่างขึ้นไป F-statistic หรือ F-test เป็นคลาสของการทดสอบทางสถิติที่คำนวณอัตราส่วนระหว่างค่าความแปรปรวน เช่น ความแปรปรวนจากสองตัวอย่างที่ต่างกัน หรือความแปรปรวนที่อธิบายและไม่ได้อธิบายโดยการทดสอบทางสถิติ การวิเคราะห์ความแปรปรวนเป็นวิธีการหนึ่งในสถิติเอฟ (F-Statistic) หรือที่อ้างถึงในที่นี้ว่า ANOVA F-test โดยจะใช้เมื่อตัวแปรหนึ่งตัวเป็นตัวเลขและอีกตัวหนึ่งเป็นแบบจัดหมวดหมู่ เช่น ตัวแปรอินพุตที่เป็นตัวเลขและตัวแปรเป้าหมายการจัดหมวดหมู่ในงานจำแนกประเภท ผลลัพธ์ของการทดสอบนี้สามารถใช้สำหรับการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งคุณลักษณะเหล่านั้นที่ไม่ขึ้นอยู่กับตัวแปรเป้าหมาย (Target Variable) สามารถลบออกจากชุดข้อมูลได้

โดยในงานวิจัยนี้ใช้ไลบรารี Scikit-learn (Pedregosa et al., 2011) มีการใช้งาน ANOVA F-test ในฟังก์ชัน f_classif() ฟังก์ชันนี้สามารถใช้ในการคัดเลือกคุณลักษณะ เช่น การคัดเลือกคุณลักษณะที่เกี่ยวข้องมากที่สุด *k* อันดับแรก โดยพิจารณาจากค่าคะแนน (Scores) ที่มากที่สุด ผ่านคลาส SelectKBest ผลคะแนนการสกัดคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test) แสดงได้ดังตารางที่ 4.8

ตารางที่ 4.8 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test)

Feature	Scores
Weight	0.435978
Carrer	0.566803
OnePersonImage	0.764148
ColorfulImage	0.866694
EmojiSentimentScore	0.88613
NegativeEmoji	1.073621
Depression Tweets	1.75525
DullImage	1.896132
CoupleStatus	2.063366
Depression ReTweets	2.271948

ตารางที่ 4.8 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test) (ต่อ)

Feature	**Scores**
TweetPeriodAM	2.341295
Positive Hashtag	2.351542
PositiveEmoji	2.729806
Friends	2.73714
Income	2.851039
CongenitalDisease	3.170981
NoPersonImage	3.229499
Negative Hashtag	3.25853
ManyPersonImage	3.709526
Genderx	4.507929
Followers	5.585183
ParentStatus	6.22048
Depression Hashtag	7.907006
FamilyMember	9.573841
TweetPeriodPM	9.621609
SentimentScore	10.931365
Negative Tweets	11.610215
Negative ReTweets	13.687593
Positive Tweets	13.891461
Positive ReTweets	19.930093
Education	25.756057
Age	34.790212

มหาวิทยาลัยเทคโนโลยีสุรนารี

จากตารางที่ 4.8 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test) แสดงจากน้อยไปมาก โดยคุณลักษณะที่มีคะแนนน้อยที่สุดคือ น้ำหนัก (Weight) รองลงมาคือ อาชีพ (Career) และจำนวนรูปภาพที่มีคน 1 คน (OnePersonImage) ตามลำดับ จากนั้นจึงตัดคุณลักษณะทีละ 1 คุณลักษณะ เหลือ 31 คุณลักษณะ (ANOVA(31)) จนตัดครบ 17 คุณลักษณะ เหลือ 15 คุณลักษณะ (ANOVA(15)) (เท่ากับเทคนิควิธีเวียนเกิด) โดยคุณลักษณะที่ตัดออกในแต่ละรอบแสดงได้ดังตารางที่ 4.9

หลังจากตัดคุณลักษณะออกแล้วจึงทำการประเมินประสิทธิภาพเพื่อหารอบการตัดคุณลักษณะที่เหมาะสมที่สุด โดยใช้การเปรียบเทียบค่าประสิทธภาพโดยรวม (F-Measure) จากเทคนิคการเรียนรู้ของเครื่อง 5 เทคนิค ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) 3) ต้นไม้การตัดสินใจ (Decision Tree (J48): DT) 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest: RF) จากกระบวนการนี้พบว่า เทคนิคการสุ่มป่าไม้ให้ค่าประสิทธิภาพโดยรวมสูงที่สุดในทุกรอบการตัดคุณลักษณะ ดังแสดงในตารางที่ 4.10

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.9 การตัดคุณลักษณะตามลำดับคะแนนจากเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test)

รอบที่	จำนวนคุณลักษณะ	คุณลักษณะที่ตัดออก
1	ANOVA(31)	['Weight']
2	ANOVA(30)	['Weight', 'Carrer']
<u>3</u>	ANOVA(29)	['Weight', 'Carrer', 'OnePersonImage']
4	ANOVA(28)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage']
5	ANOVA(27)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore']
6	ANOVA(26)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji']
7	ANOVA(25)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets']
8	ANOVA(24)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage']
9	ANOVA(23)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus']
10	ANOVA(22)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets']
11	ANOVA(21)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM']
12	ANOVA(20)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM', 'Positive Hashtag']

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.9 การตัดคุณลักษณะตามลำดับคะแนนจากเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA F-test) (ต่อ)

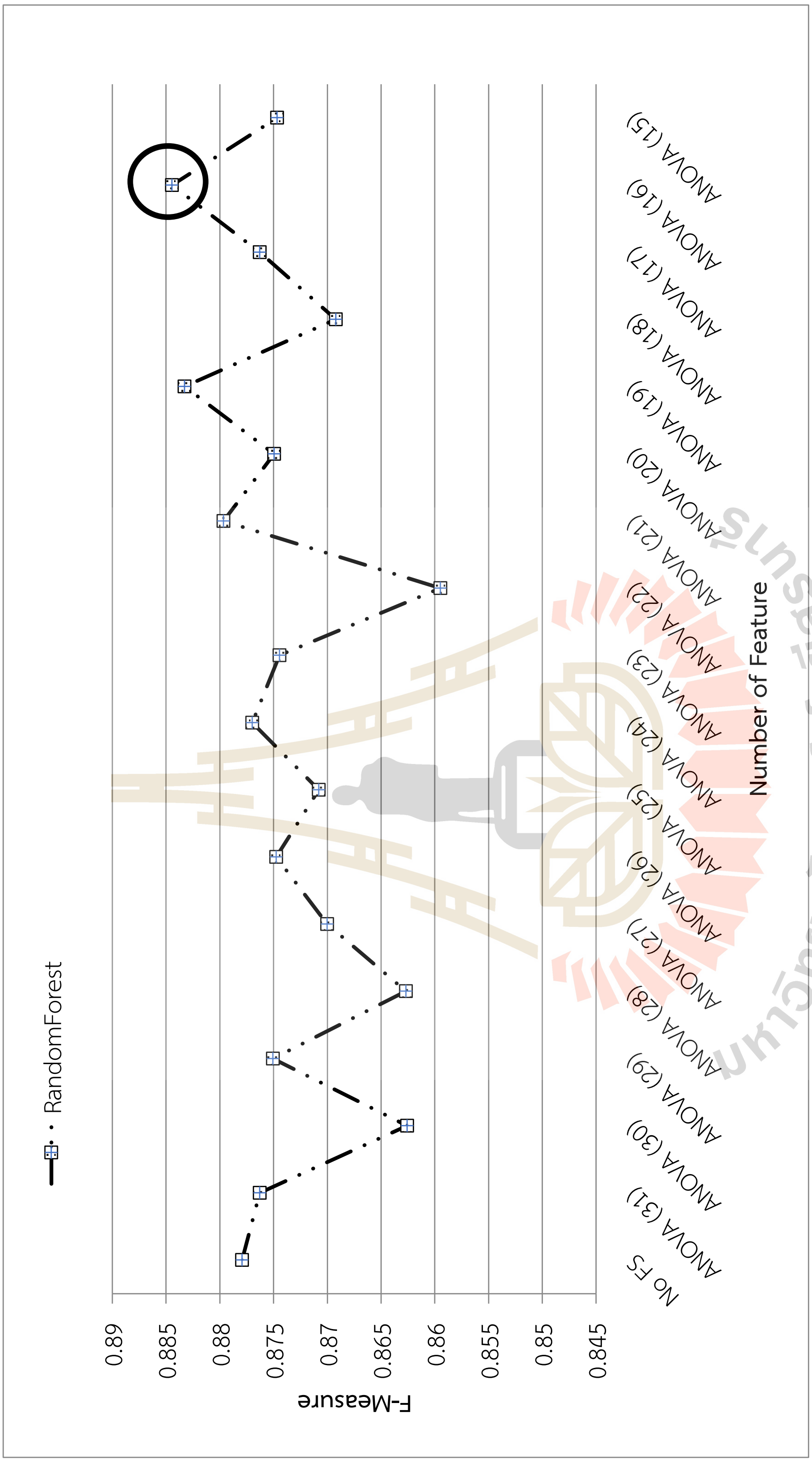
รอบที่	**จำนวนคุณลักษณะ**	**คุณลักษณะที่ตัดออก**
13	ANOVA(19)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM', 'Positive Hashtag', 'PositiveEmoji']
14	ANOVA(18)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM', 'Positive Hashtag', 'PositiveEmoji', 'Friends']
15	ANOVA(17)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM', 'Positive Hashtag', 'PositiveEmoji', 'Friends', 'Income']
16	**ANOVA(16)**	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM', 'Positive Hashtag', 'PositiveEmoji', 'Friends', 'Income', 'CongenitalDisease']
17	ANOVA(15)	['Weight', 'Carrer', 'OnePersonImage', 'ColorfulImage', 'EmojiSentimentScore', 'NegativeEmoji', 'Depression Tweets', 'DullImage', 'CoupleStatus', 'Depression ReTweets', 'TweetPeriodAM', 'Positive Hashtag', 'PositiveEmoji', 'Friends', 'Income', 'CongenitalDisease', 'NoPersonImage']

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.10 การเปรียบเทียบผลการคัดเลือกคุณลักษณะด้วยเทคนิค ANOVA F-test Feature Selection

F-Measure	**Model**				
	SVM	DT	NB	**RF**	MLP
No FS	0.79877	0.85645	0.39805	**0.87794**	0.86005
ANOVA (31)	0.79887	0.8668	0.40012	**0.87628**	0.84328
ANOVA (30)	0.79887	0.85655	0.42269	**0.86258**	0.8489
ANOVA (29)	0.79898	0.84786	0.42269	**0.87505**	0.80388
ANOVA (28)	0.79646	0.85079	0.42269	**0.86267**	0.84171
ANOVA (27)	0.79301	0.84985	0.42229	**0.87001**	0.82553
ANOVA (26)	0.79301	0.8496	0.4243	**0.87474**	0.84295
ANOVA (25)	0.79378	0.86018	0.42804	**0.87079**	0.8098
ANOVA (24)	0.79771	0.84643	0.41069	**0.87697**	0.78703
ANOVA (23)	0.79633	0.87035	0.41846	**0.87446**	0.82652
ANOVA (22)	0.79433	0.85443	0.41554	**0.85948**	0.82163
ANOVA (21)	0.79369	0.8631	0.40741	**0.87966**	0.79517
ANOVA (20)	0.79369	0.85227	0.41249	**0.87495**	0.82331
ANOVA (19)	0.79665	0.84332	0.39126	**0.88329**	0.7823
ANOVA (18)	0.8258	0.86637	0.39048	**0.8692**	0.78039
ANOVA (17)	0.82314	0.86254	0.36577	**0.87628**	0.79209
ANOVA (16)	0.82102	0.86106	0.36616	**0.88445**	0.78771
ANOVA (15)	0.80352	0.83824	0.36335	**0.87468**	0.7618

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 4.7 ประสิทธิภาพโดยรวมจากการคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวนเมื่อใช้งานร่วมกับเทคนิคการสุ่มป่าไม้

เมื่อพิจารณาการตัดคุณลักษณะในรอบที่ 16 (ANOVA(16)) ซึ่งเป็นการตัดคุณลักษณะออก 16 คุณลักษณะ พบว่าให้ค่าประสิทธิภาพโดยรวมสูงที่สุด (F-Measure=0.884449) เมื่อใช้งานร่วมกับเทคนิคการสุ่มป่าไม้ ดังแสดงในภาพที่ 4.7 ซึ่งคุณลักษณะที่ตัดออก ประกอบด้วย น้ำหนัก (X_3 : Weight) อาชีพ (X_6 : Carrer) จำนวนรูปภาพที่มีคน 1 คน (X_{31} : OnePersonImage) จำนวนรูปภาพสีสดใส (X_{28} : ColorfulImage) ค่าเฉลี่ยคะแนนความรู้สึกของสัญรูปอารมณ์ (X_{25} : EmojiSentimentScore) จำนวนสัญรูปอารมณ์ด้านลบ (X_{27} : NegativeEmoji) จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า (X_{17} : DepressionTweets) จำนวนรูปภาพสีหม่น (X_{29} : DullImage) สถานภาพ (X_9 : CoupleStatus) จำนวนรีทวีตด้านบวก (X_{20} : Depression ReTweets) จำนวนการทวีตที่อยู่ระหว่างเวลา 0.00 น.- 6.00 น. (X_{14} : TweetPeriodAM) จำนวนแฮชแท็กด้านบวก (X_{21} : Positive Hashtag) จำนวนสัญรูปอารมณ์ด้านบวก (X_{26} : PositiveEmoji) จำนวนเพื่อน (X_{11} : Friends) รายได้ (X_7 : Income) โรคประจำตัว (X_5 : CongenitalDisease)

4.1.3.3 การคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) เป็นการหาความสำคัญของคุณลักษณะ (Feature Importance) จากเทคนิคการสุ่มป่าไม้ โดยความสำคัญของคุณลักษณะคือสิ่งที่แสดงให้เห็นว่าคุณลักษณะใดที่เกี่ยวข้อง และนำไปสู่การปรับปรุงแบบจำลองโดยใช้การคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งในงานวิจัยนี้ใช้การหาความสำคัญของคุณลักษณะจากเทคนิคการสุ่มป่าไม้ด้วยไลบรารี scikit-learn โดยใช้แอททริบิวต์ feature_importances_ ในคลาส RandomForestClassifier ซึ่งเป็นตัวกำหนดค่าความสำคัญ (Scores) ให้คุณลักษณะในแต่ละคุณลักษณะ ซึ่งผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) แสดงได้ดังตารางที่ 4.11

ตารางที่ 4.11 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection)

Feature	Scores
Carrer	0.002834
CoupleStatus	0.004431
Negative Hashtag	0.005155
Positive Hashtag	0.005838
Depression Hashtag	0.005883
CongenitalDisease	0.007033
Income	0.008618
Depression Tweets	0.008923

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.11 ผลคะแนนการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) (ต่อ)

Feature	Scores
ParentStatus	0.012184
Gender	0.015384
Positive Tweets	0.026225
Education	0.026246
Depression ReTweets	0.026844
Negative Tweets	0.028754
Weight	0.030732
DullImage	0.032637
ManyPersonImage	0.033421
NegativeEmoji	0.033774
FamilyMember	0.036704
Age	0.037554
OnePersonImage	0.037564
NoPersonImage	0.040565
PositiveEmoji	0.041967
EmojiSentimentScore	0.042065
Negative ReTweets	0.045239
Positive ReTweets	0.046693
TweetPeriodAM	0.049305
ColorfulImage	0.049673
TweetPeriodPM	0.05557
Followers	0.064806
SentimentScore	0.067891
Friends	0.069487

จากตารางที่ 4.11 ผลคะแนนการคัดเลือกคุณลักษณะด้วยการสุ่มป่าไม้ (Random Forest Feature Selection) แสดงจากน้อยไปมาก โดยคุณลักษณะที่มีคะแนนน้อยที่สุดคือ อาชีพ (Carrer) รองลงมาคือ เพศ (Gender) และข้อความทวีตด้านบวก (Positive Tweets) ตามลำดับ จากนั้นจึงตัดคุณลักษณะทีละ 1 คุณลักษณะ (Random(31)) จนครบ 17 คุณลักษณะ (Random(15)) (เท่ากับเทคนิควิธีเวียนเกิด และเทคนิควิธีการวิเคราะห์ความแปรปรวน) โดยคุณลักษณะที่ตัดออกในแต่ละรอบแสดงได้ดังตารางที่ 4.12 หลังจากตัดคุณลักษณะออกแล้วจึงทำการประเมินประสิทธิภาพเพื่อหารอบการตัดคุณลักษณะที่เหมาะสมที่สุด โดยใช้การเปรียบเทียบค่าประสิทธิภาพโดยรวม (F-Measure)

จากเทคนิคการเรียนรู้ของเครื่อง 5 เทคนิค ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes: NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) 3) ต้นไม้การตัดสินใจ (Decision Tree (J48): DT) 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest: RF) จากกระบวนการนี้พบว่า เทคนิคการสุ่มป่าไม้ให้ค่าประสิทธิภาพโดยรวมสูงที่สุดในทุกรอบการตัดคุณลักษณะ ดังแสดงในตารางที่ 4.13

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.12 การตัดคุณลักษณะตามลำดับคะแนนจากเทคนิคการสุ่มป่าไม้

รอบที่	จำนวนคุณลักษณะ	คุณลักษณะที่ตัดออก
1	RFE-SVM(31)	['Friends']
2	RFE-SVM(30)	['Friends', 'Followers']
3	RFE-SVM(29)	['Friends', 'Followers', 'PositiveEmoji']
4	RFE-SVM(28)	['Friends', 'Followers', 'Negative ReTweets', 'PositiveEmoji']
5	RFE-SVM(27)	['Friends', 'Followers', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji']
6	RFE-SVM(26)	['Friends', 'Followers', 'TweetPeriodAM', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji']
7	RFE-SVM(25)	['Friends', 'Followers', 'TweetPeriodAM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji']
8	RFE-SVM(24)	['Friends', 'Followers', 'TweetPeriodAM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage']
9	RFE-SVM(23)	['Friends', 'Followers', 'TweetPeriodAM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage', 'DullImage']
10	RFE-SVM(22)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage', 'DullImage']
11	RFE-SVM(21)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'ManyPersonImage', 'DullImage']

มหาวิทยาลัยเทคโนโลยีสุรนารี

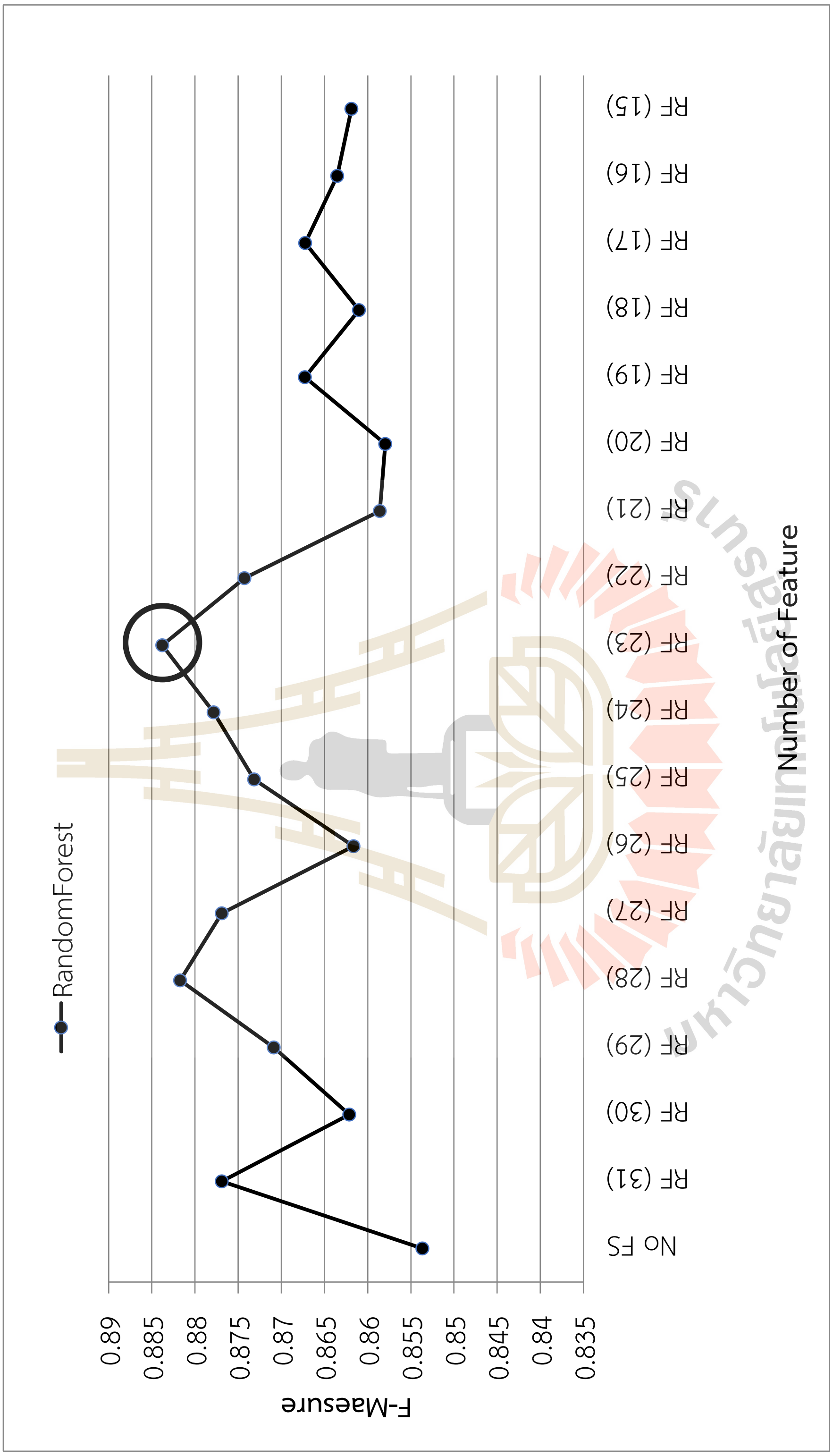
ตารางที่ 4.12 การตัดคุณลักษณะตามลำดับคะแนนจากเทคนิคการสุ่มป่าไม้ (ต่อ)

รอบที่	จำนวนคุณลักษณะ	คุณลักษณะที่ตัดออก
12	RFE-SVM(20)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'OnePersonImage', 'ManyPersonImage', 'DullImage']
13	RFE-SVM(19)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
14	RFE-SVM(18)	['Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
15	RFE-SVM(17)	['FamilyMember', 'Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
16	RFE-SVM(16)	['Carrer', 'FamilyMember', 'Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'ManyPersonImage', 'DullImage', 'ColorfulImage']
17	RFE-SVM(15)	['Carrer', 'FamilyMember', 'Friends', 'Followers', 'TweetPeriodAM', 'TweetPeriodPM', 'Depression Tweets', 'Positive ReTweets', 'Negative ReTweets', 'Depression ReTweets', 'PositiveEmoji', 'NegativeEmoji', 'NoPersonImage', 'OnePersonImage', 'F2', 'DullImage', 'ColorfulImage']

ตารางที่ 4.13 เปรียบเทียบผลการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้

F-Measure	**Model**				
	SVM	DT	NB	**RF**	MLP
No FS	0.79877	0.8579	0.39805	**0.85364**	0.87702
RF **(31)**	0.80155	0.85175	0.41867	**0.87691**	0.85148
RF **(30)**	0.80155	0.87278	0.43011	**0.86211**	0.84558
RF **(29)**	0.79898	0.85272	0.45554	**0.87089**	0.84745
RF **(28)**	0.79547	0.86086	0.43291	**0.88174**	0.84301
RF **(27)**	0.79818	0.8638	0.44149	**0.8769**	0.83883
RF **(26)**	0.79818	0.85512	0.44342	**0.86162**	0.85046
RF **(25)**	0.79953	0.85922	0.46494	**0.87316**	0.83931
RF **(24)**	0.79766	0.85764	0.47298	**0.87785**	0.81556
RF **(23)**	0.79713	0.86094	0.45057	**<u>0.88379</u>**	0.84034
RF **(22)**	0.79497	0.8497	0.44805	**0.87428**	0.83264
RF **(21)**	0.79261	0.86281	0.43895	**0.85861**	0.82542
RF **(20)**	0.80059	0.84375	0.40895	**0.85797**	0.84024
RF **(19)**	0.79805	0.85619	0.41563	**0.86726**	0.802
RF **(18)**	0.79601	0.83706	0.39769	**0.86101**	0.79084
RF **(17)**	0.79601	0.84081	0.39152	**0.86725**	0.80429
RF **(16)**	0.79308	0.84926	0.36902	**0.86353**	0.77173
RF **(15)**	0.78532	0.82855	0.36254	**0.86189**	0.74501

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 4.8 ประสิทธิภาพโดยรวมจากการคัดเลือกคุณลักษณะด้วยการสุ่มป่าไม้

เมื่อพิจารณาการตัดคุณลักษณะในรอบที่ 9 (RF(23)) ซึ่งเป็นการตัดคุณลักษณะออก 9 คุณลักษณะ พบว่าให้ค่าประสิทธิภาพโดยรวมสูงที่สุด (F-Measure=0.883788) เมื่อใช้งานร่วมกับเทคนิคการสุ่มป่าไม้ ดังแสดงในภาพที่ 4.8 ซึ่งคุณลักษณะที่ตัดออก ประกอบด้วย อาชีพ ($\boldsymbol{X_6}$: Carrer) สถานภาพ ($\boldsymbol{X_9}$: CoupleStatus) จำนวนแฮชแท็กด้านลบ ($\boldsymbol{X_{22}}$: Negative Hashtag) จำนวนแฮชแท็กด้านบวก ($\boldsymbol{X_{21}}$: Positive Hashtag) จำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า ($\boldsymbol{X_{23}}$: Depression Hashtag) โรคประจำตัว ($\boldsymbol{X_5}$: CongenitalDisease) รายได้ ($\boldsymbol{X_7}$: Income) จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า ($\boldsymbol{X_{17}}$: DepressionTweets) สถานภาพ (บิดา-มารดา) ($\boldsymbol{X_{10}}$: 'ParentStatus)

จากการคัดเลือกคุณลักษณะด้วยเทคนิคทั้ง 3 เทคนิค ประกอบด้วย 1) การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (Recursive Feature Elimination and Support Vector Machine: RFE-SVM) 2) การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (Analysis of Variance : ANOVA) และ 3) การคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) เมื่อนำมาใช้งานร่วมกับการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องอื่น ๆ ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes: NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) 3) ต้นไม้การตัดสินใจ (Decision Tree (J48): DT) 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest: RF) พบว่าการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ (Random Forest: RF) ให้ค่ามาตรฐานในการประเมินประสิทธิภาพ 4 ค่า ได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าประสิทธิภาพโดยรวม (F-measure) สูงที่สุดในทุกเทคนิควิธีในการคัดเลือกคุณลักษณะ ดังแสดงในตารางที่ 4.14 ถึงตารางที่ 4.17

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.14 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง โดยไม่ใช้เทคนิคการคัดเลือกคุณลักษณะ (No FS)

Model	Accuracy	Precision	Recall	F-Measure
SVM	0.777222	0.871008	0.77472	0.78262
DT	0.844921	0.84073	0.83611	0.83978
NB	0.479048	0.474612	0.47521	0.40483
RF	**0.873254**	**0.88174**	**0.86576**	**0.86477**
MLP	0.872937	0.844493	0.84583	0.81363

ตารางที่ 4.15 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ ที่ 29 คุณลักษณะ (RFE-SVM(29))

Model	Accuracy	Precision	Recall	F-Measure
SVM	0.867698	0.890818	0.86063	0.85943
DT	0.865079	0.870981	0.85875	0.85515
NB	0.433651	0.535724	0.43063	0.35382
RF	**0.890079**	**0.897964**	**0.88458**	**0.88956**
MLP	0.853492	0.860201	0.87632	0.85572

ตารางที่ 4.16 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน ที่ 15 คุณลักษณะ (ANOVA(15))

Model	Accuracy	Precision	Recall	F-Measure
SVM	0.822698	0.872883	0.81979	0.82102
DT	0.870476	0.87622	0.87299	0.87485
NB	0.45381	0.407404	0.45007	0.36616
RF	**0.879048**	**0.894197**	**0.88174**	**0.87928**
MLP	0.797302	0.811275	0.81507	0.78715

มหาวิทยาลัยเทคโนโลยีสุรนารี

ตารางที่ 4.17 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องโดยใช้การคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ ที่ 23 คุณลักษณะ (RF(23))

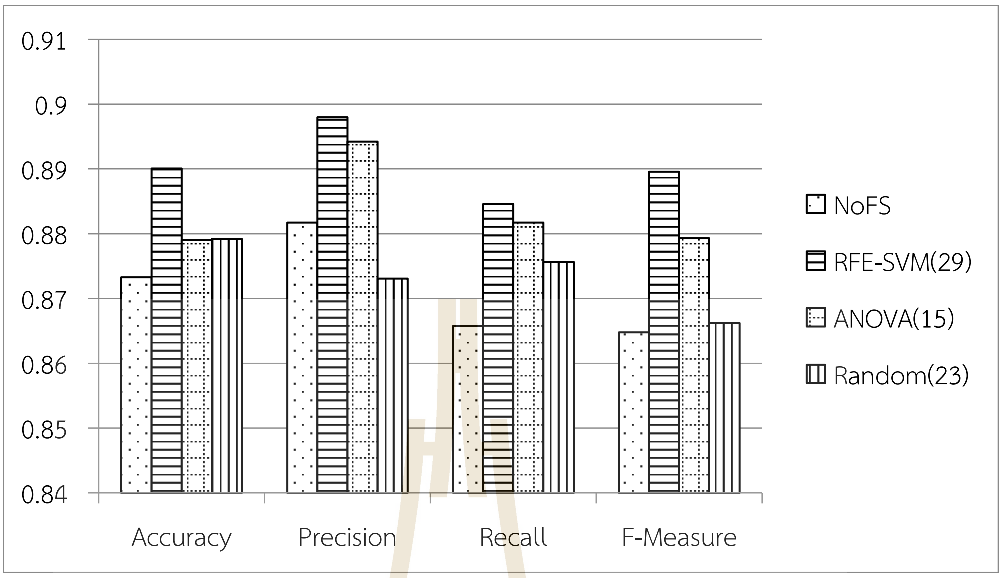
Model	**Accuracy**	**Precision**	**Recall**	**F-Measure**
SVM	0.794603	0.873053	0.790625	0.797131
DT	0.879127	0.871176	0.868403	0.858637
NB	0.48746	0.488022	0.489583	0.450572
RF	**0.879206**	**0.875652**	**0.866181**	**0.873065**
MLP	0.825317	0.833017	0.841319	0.78576

เมื่อพิจารณาค่าประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ (RF) ซึ่งให้ค่าประสิทธิภาพสูงที่สุดในทุกเทคนิควิธีในการคัดเลือกคุณลักษณะพบว่า การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ ที่ 29 คุณลักษณะ (RFE-SVM(29)) ให้ค่ามาตรฐานในการประเมินประสิทธิภาพ 4 ค่า ได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าประสิทธิภาพโดยรวม (F-measure) สูงที่สุด และการไม่ใช้เทคนิคการคัดเลือกคุณลักษณะ (No FS) ให้ค่ามาตรฐานในการประเมินประสิทธิภาพ 4 ค่า ต่ำที่สุด ดังแสดงในตารางที่ 4.18

ตารางที่ 4.18 การเปรียบเทียบค่าประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้

Feature Selection Technique	**Accuracy**	**Precision**	**Recall**	**F-Measure**
No FS	0.873254	0.88174	0.86576	0.86477
RFE-SVM(29)	**0.890079**	**0.897964**	**0.88458**	**0.88956**
ANOVA(15)	0.879048	0.894197	0.88174	0.87928
RF(23)	0.879206	0.873065	0.875652	0.866181

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 4.9 ประสิทธิภาพการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ร่วมกับการคัดเลือกคุณลักษณะด้วยวิธีการต่าง ๆ

จากแผนภูมิในภาพที่ 4.9 แสดงให้เห็นถึงการเปรียบเทียบประสิทธิภาพในการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ ร่วมกับการคัดเลือกคุณลักษณะด้วยวิธีการต่าง ๆ ได้แก่ การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิด โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (RFE-SVM) การคัดเลือกคุณลักษณะด้วยเทคนิคการวิเคราะห์ความแปรปรวน (ANOVA) และการคัดเลือกคุณลักษณะด้วยเทคนิคการสุ่มป่าไม้ (Random Forest Feature Selection) พบว่า การคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ ที่ 29 คุณลักษณะ (RFE-SVM(29)) ร่วมกับการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ ซึ่งในงานวิจัยนี้ เรียกว่าแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องแบบไฮบริดจ์ (Hybrid Support Vector Machine and Random Forest: Hybrid SVM-RF) โดยให้ค่ามาตรฐานในการประเมินประสิทธิภาพ 4 ค่า ได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าประสิทธิภาพโดยรวม (F-measure) สูงที่สุด (F-measure=0.88956) ซึ่งคุณลักษณะที่ตัดออก ประกอบด้วย จำนวนเพื่อน (X_{11} : Friend) จำนวนผู้ติดตาม (X_{12} : Followers) จำนวนสัญรูปอารมณ์ด้านบวก (X_{26} : PositiveEmoji)

4.1.4 การคัดเลือกคุณลักษณะที่เหมาะสม (Feature Selection)

คุณลักษณะที่เป็นตัวแปรนำเข้า (Input Variables) ประกอบด้วย คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics) (**X_1 - X_{10}**) คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information) (**X_{12} - X_{14}**) คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) (**X_{15} – X_{24}**) คุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon) (**X_{25} , X_{27}**) และคุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) (**X_{28} - X_{32}**) โดยสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่องแบบไฮบริดจ์ (Hybrid Support Vector Machine and Random Forest: Hybrid SVM-RF) ซึ่งเป็นการคัดเลือกคุณลักษณะด้วยเทคนิควิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (Recursive Feature Elimination and Support Vector Machine: RFE-SVM) (RFE-SVM(29)) ร่วมกับการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ (Random Forest) โดยผลการเปรียบเทียบประสิทธิภาพ แสดงได้ดังตารางที่ 4.19 และภาพที่ 4.10

มหาวิทยาลัยเทคโนโลยีสุรนารี

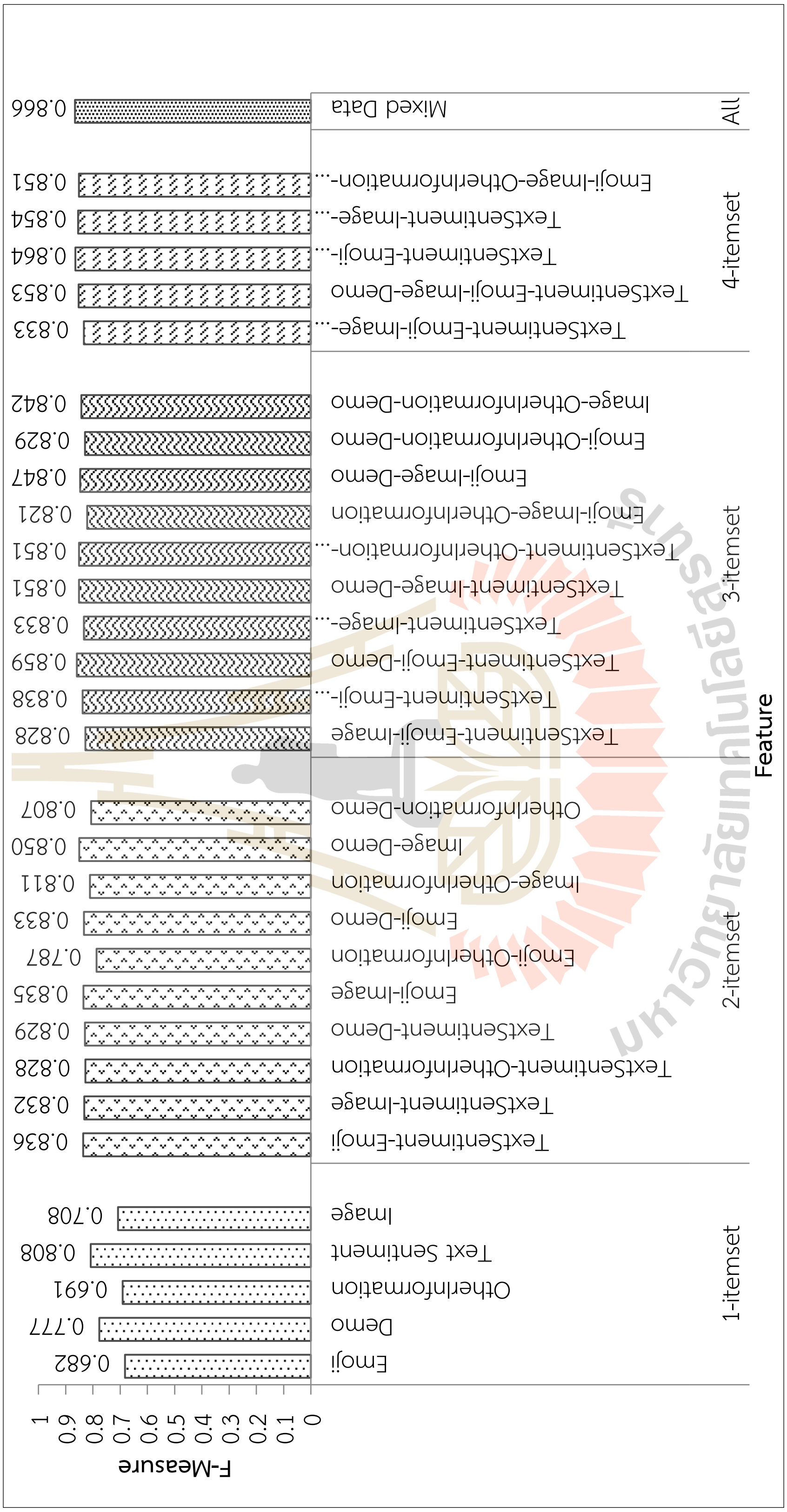
ตารางที่ 4.19 การเปรียบเทียบค่าประสิทธิภาพโดยรวม (F-measure) ของแบบจำลองจากการใช้ตัวแปรนำเข้า (Input Variables) ที่แตกต่างกัน

ตัวแปรนำเข้า	F-Measure										
	1	2	3	4	5	6	7	8	9	10	AVG
Emoji	0.71948	0.68262	0.68041	0.67718	0.68317	0.68031	0.67867	0.67431	0.67344	0.673	0.682259
Demo	0.70091	0.786	0.78968	0.78374	0.79186	0.78691	0.78126	0.78745	0.78041	0.7771	0.776532
OtherInformation	0.70408	0.68514	0.69371	0.68518	0.68443	0.69488	0.68701	0.68132	0.69591	0.69713	0.690879
Text Sentiment	0.80308	0.81214	0.82028	0.80942	0.80726	0.80728	0.79731	0.8125	0.79894	0.81293	0.808114
Image	0.71915	0.70847	0.70878	0.70414	0.70559	0.70795	0.71203	0.70299	0.70945	0.70041	0.707897
TextSentiment-Emoji	0.849	0.83271	0.83674	0.82882	0.82851	0.83129	0.83763	0.83474	0.84138	0.83663	0.835745
TextSentiment-Image	0.87633	0.82913	0.83086	0.81939	0.81688	0.82066	0.82984	0.83661	0.82643	0.83184	0.831796
TextSentiment-OtherInformation	0.82683	0.82882	0.83288	0.82336	0.8318	0.83465	0.82027	0.83056	0.82423	0.82296	0.827635
TextSentiment-Demo	0.8495	0.83003	0.82277	0.82263	0.8183	0.82614	0.83215	0.83011	0.83042	0.83131	0.829336
Emoji-Image	0.85474	0.83666	0.83842	0.82248	0.8271	0.83791	0.82215	0.8302	0.83816	0.8398	0.834761
Emoji-OtherInformation	0.85596	0.78676	0.76831	0.76677	0.77618	0.79757	0.79415	0.77424	0.77956	0.76705	0.786653
Emoji-Demo	0.83567	0.82238	0.82361	0.83304	0.83917	0.84147	0.84029	0.82956	0.83153	0.83293	0.832964
Image-OtherInformation	0.81075	0.81764	0.8014	0.81969	0.7992	0.81005	0.81	0.81231	0.81009	0.81752	0.810865
Image-Demo	0.85161	0.84187	0.8469	0.85011	0.85033	0.84563	0.85558	0.8494	0.85202	0.8581	0.850155
OtherInformation-Demo	0.85665	0.80707	0.80529	0.79825	0.80408	0.80758	0.80354	0.79207	0.79089	0.80014	0.806556
TextSentiment-Emoji-Image	0.88427	0.82702	0.8297	0.82379	0.82273	0.82588	0.82922	0.81225	0.81847	0.80326	0.827658
TextSentiment-Emoji-OtherInformation	0.84164	0.83182	0.84025	0.83313	0.83	0.85351	0.83705	0.82674	0.83331	0.8508	0.837824
TextSentiment-Emoji-Demo	0.88677	0.86127	0.85643	0.85648	0.85471	0.86121	0.85056	0.86016	0.85314	0.84884	0.858957
TextSentiment-Image-OtherInformation	0.85543	0.83939	0.8243	0.83204	0.82895	0.8252	0.82591	0.8357	0.82762	0.83618	0.833071

ตารางที่ 4.19 การเปรียบเทียบค่าประสิทธิภาพโดยรวม (F-measure) ของแบบจำลองจากการใช้ตัวแปรนำเข้า (Input Variables) ที่แตกต่างกัน (ต่อ)

ตัวแปรนำเข้า	**F-Measure**										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**AVG**
TextSentiment-Image-Demo	0.88634	0.85378	0.84804	0.84022	0.84704	0.84461	0.84227	0.84505	0.85366	0.8509	0.851192
TextSentiment-OtherInformation-Demo	0.85036	0.85043	0.85626	0.85414	0.85372	0.84582	0.84432	0.85628	0.85745	0.8421	0.851087
Emoji-Image-OtherInformation	0.87801	0.81438	0.81636	0.81489	0.81683	0.82546	0.80559	0.80893	0.80717	0.82242	0.821004
Emoji-Image-Demo	0.86826	0.8421	0.84593	0.83409	0.84681	0.84591	0.83789	0.85691	0.84128	0.8485	0.846767
Emoji-OtherInformation-Demo	0.85965	0.81846	0.82897	0.8269	0.83086	0.81939	0.83493	0.81542	0.83297	0.82366	0.829122
Image-OtherInformation-Demo	0.83956	0.85612	0.83597	0.83899	0.84393	0.8422	0.84159	0.8436	0.84169	0.8407	0.842435
TextSentiment-Emoji-Image-OtherInformation	0.8598	0.83148	0.82263	0.84098	0.83397	0.83208	0.82748	0.8257	0.81968	0.8378	0.83316
TextSentiment-Emoji-Image-Demo	0.88315	0.85399	0.84411	0.8468	0.85104	0.85679	0.84455	0.84153	0.84898	0.85414	0.852507
TextSentiment-Emoji-OtherInformation-Demo	0.86764	0.87509	0.87402	0.85068	0.86497	0.87564	0.85747	0.85494	0.8654	0.85793	0.864378
TextSentiment-Image-OtherInformation-Demo	0.87726	0.84339	0.84514	0.85946	0.85048	0.85681	0.85468	0.85602	0.84548	0.85551	0.854422
Emoji-Image-OtherInformation-Demo	0.86312	0.84384	0.84885	0.85641	0.85592	0.84738	0.83415	0.85372	0.86378	0.84621	0.851339
Mixed Data	0.88909	0.85573	0.86057	0.86495	0.8825	0.85571	0.86087	0.86167	0.85739	0.86763	0.865612

มหาวิทยาลัยเทคโนโลยีสุรนารี



ภาพที่ 4.10 การเปรียบเทียบค่าประสิทธิภาพของแบบจำลองจากการใช้ตัวแปรนำเข้า (Input Variables) ที่แตกต่างกัน

เมื่อพิจารณาจากแผนภูมิในภาพที่ 4.10 ซึ่งใช้คุณลักษณะที่เป็นตัวแปรนำเข้าที่แตกต่างกัน จะเห็นได้ว่าคุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon) ให้ค่าประสิทธิภาพโดยรวมน้อยที่สุด (F-measure=0.682) รองลงมาคือ คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information) (F-measure=0.691) คุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) (F-measure=0.708) คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics) (F-measure=0.777) และคุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) ให้ค่าประสิทธิภาพโดยรวมสูงที่สุด (F-measure=0.808) ซึ่งเป็นคุณลักษณะที่อยู่ในกลุ่ม 1 รายการ (1-itemset) ในขณะที่คุณลักษณะในกลุ่ม 2 รายการ (2-itemset) 3 รายการ (3-itemset) และ 4 รายการ (4-itemset) ให้ค่าประสิทธิภาพโดยรวมสูงกว่าคุณลักษณะที่อยู่ในกลุ่ม 1 รายการ และเมื่อนำคุณลักษณะทั้งหมด (Mixed Data Sources) มาใช้ในการสร้างแบบจำลองพบว่า คุณลักษณะทั้งหมดให้ค่าประสิทธิภาพโดยรวมสูงที่สุด (F-measure=0.866) ดังนั้น ข้อมูลคุณลักษณะทั้งหมด คุณลักษณะด้านกายภาพส่วนบุคคล (Demographic Characteristics) $(X_1 - X_{10})$ คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information) $(X_{12} - X_{14})$ คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) $(X_{15} - X_{24})$ คุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon) (X_{25}, X_{27}) และคุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) $(X_{28} - X_{32})$ จึงเหมาะสมกับการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ในงานวิจัยนี้

4.2 ผลการประเมินแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์

4.2.1 การเปรียบเทียบค่าประสิทธิภาพแบบจำลองที่พัฒนา กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) อื่น ๆ

ผลการประเมินประสิทธิภาพแบบจำลองที่พัฒนากับเทคนิคการเรียนรู้ของเครื่องอื่น ๆ แสดงได้ดังตารางที่ 4.20 ซึ่งพบว่าแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (Depression Risk Analysis Model: DeRAM) ซึ่งใช้การคัดเลือกคุณลักษณะด้วยวิธีเวียนเกิด (Recursive Feature Elimination: RFE) โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ ร่วมกับเทคนิคการสุ่มป่าไม้ในการสร้างแบบจำลองในงานวิจัยนี้ ให้ค่าประสิทธิภาพโดยรวมสูงกว่าแบบจำลองที่พัฒนาด้วยเทคนิคการเรียนรู้ของเครื่องอื่น ๆ โดยให้ค่าประสิทธิภาพโดยรวมสูงที่สุด (F-measure=0.8851) รองลงมาคือ เทคนิคการสุ่มป่าไม้ (F-measure=0.8647) เทคนิคต้นไม้การตัดสินใจ (F-measure=0.8397) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (F-measure=0.8136) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (F-measure=0.7826) และเทคนิคนาอีฟเบย์ที่ให้ค่าประสิทธิภาพโดยรวมต่ำที่สุด (F-measure=0.4048)

ตารางที่ 4.20 ผลการประเมินประสิทธิภาพแบบจำลองที่พัฒนากับเทคนิคการเรียนรู้ของเครื่องอื่น ๆ

Model	F-measure	Accuracy	Precision	Recall
SVM	0.7826	0.7772	0.8710	0.7747
DT	0.8397	0.8449	0.8407	0.8361
NB	0.4048	0.4790	0.4746	0.4752
RF	0.8647	0.8732	0.8817	0.8657
MLP	0.8136	0.8729	0.8444	0.8458
DeRAM	**0.8851**	**0.8900**	**0.9015**	**0.8769**

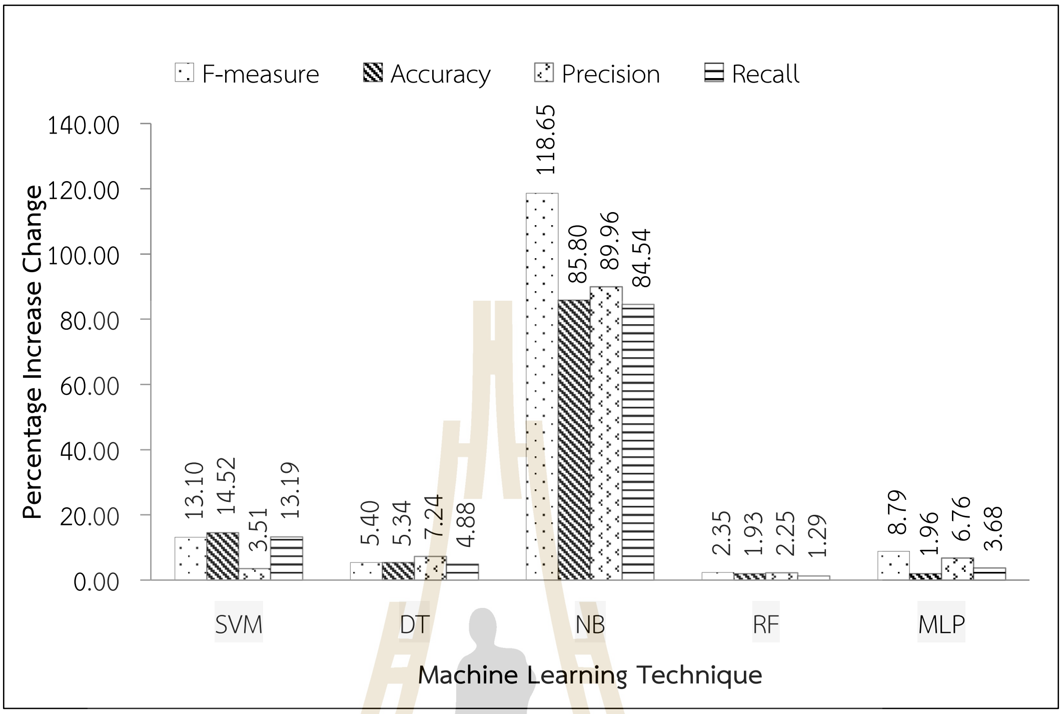
เมื่อพิจารณาผลการเปรียบเทียบค่าประสิทธิภาพแบบจำลองที่พัฒนา กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) อื่น ๆ ด้วยการเปรียบเทียบค่าร้อยละเพิ่มเปลี่ยนแปลง (Percentage Increase Change) ดังสมการที่ 4.1

$$Percentage\ Increase\ Change = \left[\frac{v_1 - v_2}{v_2}\right] \times 100 \qquad (4.1)$$

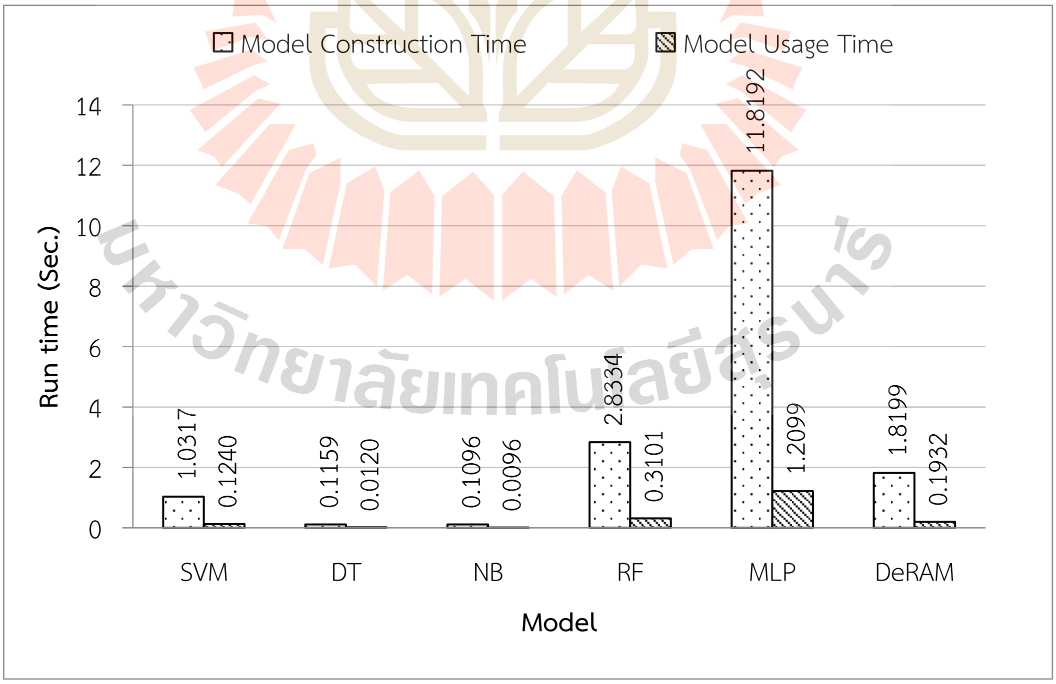
โดยที่ v_1 คือค่าประสิทธิภาพของแบบจำลองที่พัฒนา (DeRAM)

v_2 คือค่าประสิทธิภาพของแบบจำลองอื่น ๆ

เมื่อพิจารณาจากแผนภูมิในภาพที่ 4.11 พบว่าแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (DeRAM) ให้ค่ามาตรฐานในการประเมินประสิทธิภาพแบบจำลอง 4 ค่า คือ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) สูงที่สุด โดยมีค่าประสิทธิภาพโดยรวม (F-measure) สูงกว่าแบบจำลองที่พัฒนาด้วยเทคนิคนาอีฟเบย์มากที่สุดคือร้อยละ 118.65 รองลงมาคือเทคนิคซัพพอร์ตเวกเตอร์แมชชีนซึ่งเป็นบรรทัดฐานที่งานวิจัยส่วนใหญ่ใช้ ร้อยละ 13.10 นอกจากนั้นยังมีค่าประสิทธิภาพโดยรวมสูงกว่า เทคนิคเพอร์เซ็ปตรอนหลายชั้น ร้อยละ 8.79 เทคนิคต้นไม้การตัดสินใจร้อยละ 5.40 และเทคนิคการสุ่มป่าไม้ร้อยละ 2.35 ตามลำดับ



ภาพที่ 4.11 การเปรียบเทียบค่าประสิทธิภาพแบบจำลองที่พัฒนา กับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) อื่น ๆ



ภาพที่ 4.12 เวลาในการสร้างและเรียกใช้แบบจำลองรูปแบบต่าง ๆ

จากกระบวนการสร้างแบบจำลอง เมื่อพิจารณาเวลาที่ใช้ในการประมวลผลแบบจำลองดังแผนภูมิในภาพที่ 4.12 พบว่าแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (DeRAM) ให้ค่ามาตรฐานในการประเมินประสิทธิภาพแบบจำลอง 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-measure) สูงที่สุด โดยใช้เวลาในการสร้างแบบจำลอง (Model Construction Time) 1.8199 วินาที และใช้เวลาในการเรียกใช้แบบจำลอง (Model Usage Time) 0.1932 วินาที ซึ่งต่างจากแบบจำลองที่พัฒนาด้วยเทคนิคนาอีฟเบย์ (Naïve Bays: NB) ซึ่งใช้เวลาในการประมวลผลน้อยที่สุด โดยแบบจำลองที่พัฒนา (DeRAM) ใช้เวลาการสร้างและเรียกใช้แบบจำลองมากกว่าเทคนิคนาอีฟเบย์ (Naïve Bays: NB) โดยมีค่า 1.7103 วินาที และ 0.1837 วินาที ตามลำดับ แต่ให้ค่าความถูกต้อง (Accuracy) โดยรวมสูงกว่าเทคนิคนาอีฟเบย์ ถึงร้อยละ 118.65 ซึ่งแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าในงานวิจัยนี้มุ่งเน้นที่ประสิทธิภาพด้านความถูกต้องเป็นสำคัญ

4.2.2 การอภิปรายผล

จากผลการทดลองสามารถอภิปรายผลได้ดังนี้ แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า (Depression Risk Analysis Model: DeRAM) ที่นำเสนอในงานวิจัยนี้ ใช้การคัดเลือกคุณลักษณะด้วยวิธีเวียนเกิด (Recursive Feature Elimination: RFE) โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ ร่วมกับเทคนิคการสุ่มป่าไม้ในการสร้างแบบจำลอง โดยมีค่าประสิทธิภาพโดยรวม (F-measure) สูงกว่าเทคนิคการเรียนรู้ของเครื่องอื่นทั้งหมด (F-measure = 0.8851) ในขณะที่เทคนิคนาอีฟเบย์ให้ค่าประสิทธิภาพโดยรวมต่ำที่สุด (F-measure = 0.4048) และเมื่อพิจารณาที่ค่าร้อยละเพิ่มเปลี่ยนแปลง (Percentage Increase Change) พบว่าเทคนิคการเรียนรู้ของเครื่องแบบไฮบริดจ์ (Hybrid SVM-RF) ที่นำเสนอมีค่าประสิทธิภาพโดยรวมสูงกว่าเทคนิคนาอีฟเบย์มากที่สุดคือ ร้อยละ 118.65 รองลงมาคือเทคนิคซัพพอร์ตเวกเตอร์แมชชีนซึ่งเป็นบรรทัดฐานที่งานวิจัยส่วนใหญ่ใช้ ร้อยละ 13.10 (Barhan and Shakhomirov, 2012; De Choudhury et al., 2013; Hutto and Gilbert, 2014; Bridianne O'Dea et al., 2015; Aldarwish, 2017; Islam et al., 2018) นอกจากนั้นยังมีค่าประสิทธิภาพโดยรวมสูงกว่า เทคนิคเพอร์เซ็ปตรอนหลายชั้น ร้อยละ 8.79 เทคนิคต้นไม้การตัดสินใจร้อยละ 5.40 และเทคนิคการสุ่มป่าไม้ร้อยละ 2.35

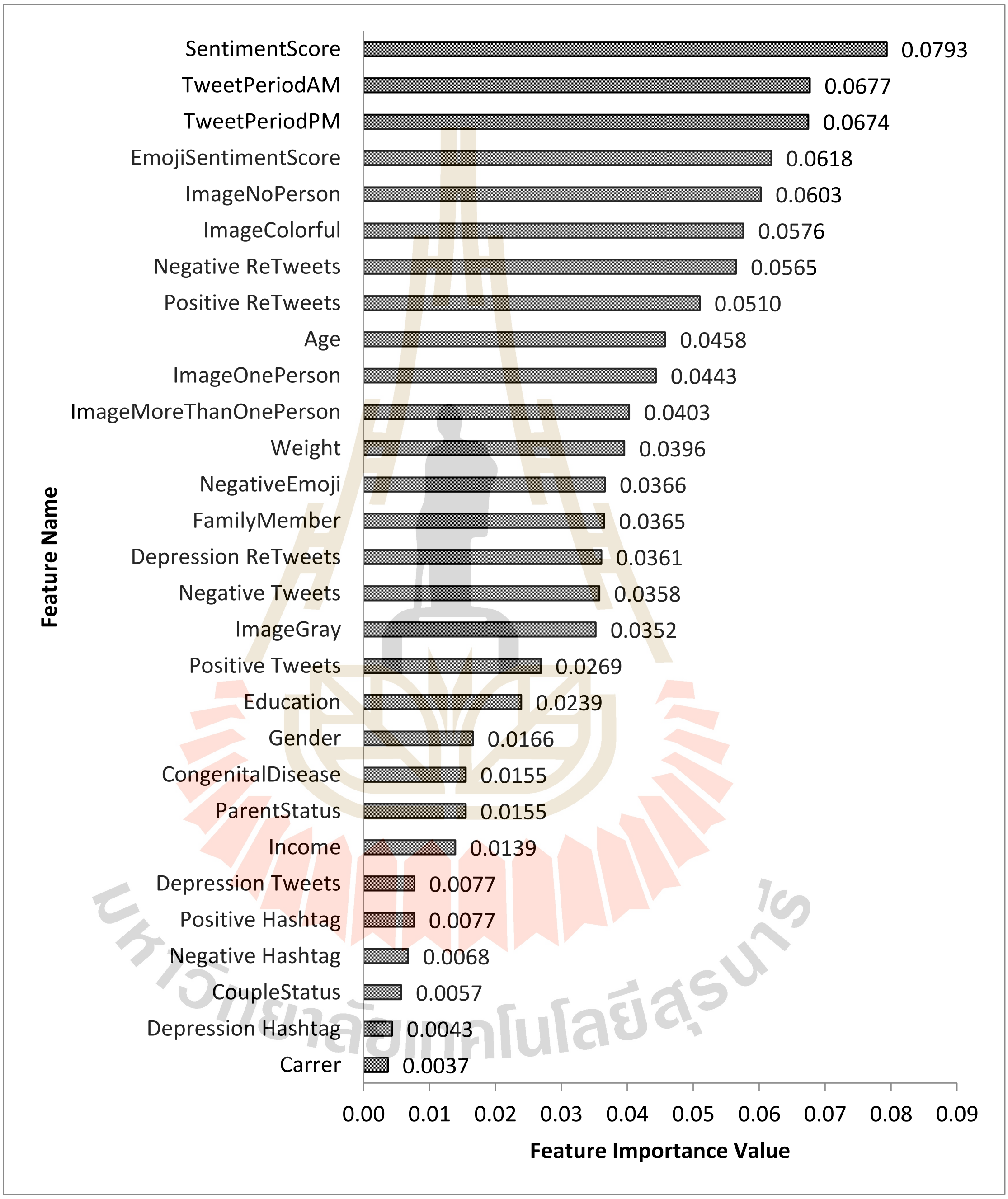
สาเหตุที่ทำให้เทคนิคที่พัฒนาขึ้นมีประสิทธิภาพสูงที่สุดในงานวิจัยนี้ คือ การนำวิธีเวียนเกิดโดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบมาใช้ในการคัดเลือกคุณลักษณะ ร่วมกับการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่าไม้ เพราะเทคนิคการสุ่มป่าไม้เป็นเทคนิคที่ใช้ได้ทั้งกับข้อมูลแบบมีโครงสร้าง (Structured Data) และข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) (ปริญญา สงวนสัตย์, 2562) อีกทั้งยังเป็นการพัฒนามาจากเทคนิคต้นไม้

การตัดสินใจ สอดคล้องกับงานวิจัยของ Reece et al. (2017) ที่สร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าด้วยเทคนิคการสุ่มป่าไม้ ซึ่งมีค่าประสิทธิภาพโดยรวมอยู่ที่ 0.772 โดยมีค่าน้อยกว่าเทคนิคที่พัฒนาขึ้นในงานวิจัยนี้ เนื่องจากความแตกต่างของข้อมูลนำเข้า (Input Data) ปริมาณของข้อมูลที่ใช้ในการสร้างแบบจำลอง และขั้นตอนวิธีในการสร้างแบบจำลองที่แตกต่างกัน ในขณะที่งานวิจัยของ Husseini et al. (2018) ได้ประยุกต์ใช้การเรียนรู้เชิงลึก (CNN) ในการตรวจสอบภาวะซึมเศร้าของผู้ใช้งานทวิตเตอร์ โดยใช้ข้อมูลจากชุดข้อมูลสาธารณะ (Public Dataset) และใช้เฉพาะข้อความในการประมวลผลเท่านั้น โดยค่าประสิทธิภาพโดยรวมที่ได้คือร้อยละ 86.967 ซึ่งมากกว่าการใช้เทคนิคเพอร์เซ็ปตรอนหลายชั้น (MLP) ในงานวิจัยนี้ เนื่องจากเป็นเทคนิคที่ต้องใช้ข้อมูลในการฝึกฝนจำนวนมาก และแต่ละเทคนิคย่อยในการเรียนรู้เชิงลึกก็เหมาะสำหรับข้อมูลรูปแบบที่แตกต่างกันออกไป ซึ่งในงานวิจัยนี้มีข้อมูลหลากหลายส่วน เช่น ข้อมูลด้านกายภาพส่วนบุคคล ข้อความ สัญรูปอารมณ์ และรูปภาพ แตกต่างจากงานวิจัยส่วนใหญ่ที่ใช้เพียงข้อความการแสดงความคิดเห็น

จากผลการทดลองจะเห็นได้ว่า เทคนิคการเรียนรู้ของเครื่องแบบไฮบริดจ์ที่พัฒนาขึ้นมีค่าประสิทธิภาพโดยรวมสูงกว่าเทคนิคการสุ่มป่าไม้ เทคนิคต้นไม้การตัดสินใจ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน และเทคนิคนาอีฟเบย์ตามลำดับ เนื่องจากมีการตัดคุณลักษณะบางส่วนออกไปจากขั้นตอนการคัดเลือกคุณลักษณะ (Feature Selection) ซึ่งเทคนิคนาอีฟเบย์แม้ว่าจะเป็นเทคนิคหนึ่งที่มีความนิยมมากที่สุด (Barhan and Shakhomirov, 2012; Wang et al., 2013; Hutto and Gilbert, 2014; Zhu et al., 2016; Aldarwish, 2017) แต่เทคนิคนาอีฟเบย์นั้นเหมาะสำหรับการจำแนกชุดข้อมูลขนาดใหญ่ (Kabir, Rahman, Hossain, and Dahal, 2011) ซึ่งในงานวิจัยนี้ข้อมูลที่นำมาใช้ในการสร้างแบบจำลองยังมีจำนวนข้อมูลไม่มากพอ ส่งผลให้เทคนิคนาอีฟเบย์ในงานวิจัยนี้ให้ค่าประสิทธิภาพโดยรวมต่ำที่สุด นอกจากนั้นแล้วข้อมูลที่ใช้ในการทดลองนี้อยู่ในบริบทที่แตกต่างกับงานวิจัยอื่น ไม่ว่าจะเป็นคุณลักษณะของกลุ่มตัวอย่าง อาทิ เชื้อชาติ และภาษาที่ใช้ในการแสดงความคิดเห็นผ่านเครือข่ายสังคมออนไลน์ คุณลักษณะที่สกัดมาใช้จากทวิตเตอร์ และด้านกายภาพ ดังนั้นเทคนิคที่พัฒนาขึ้นมาจะเหมาะสมที่สุดกับข้อมูลคุณลักษณะชุดนี้

ในส่วนของการหาคุณลักษณะที่สำคัญสำหรับการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์ เมื่อพิจารณาจากแผนภูมิในภาพที่ 4.12 พบว่า คุณลักษณะที่สำคัญที่สุดในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ในงานวิจัยนี้ คือ ค่าเฉลี่ยคะแนนความรู้สึกของทวีต(Sentiment Score) รองลงมาคือ จำนวนทวีตที่อยู่ระหว่างเวลา 6.00 น. - 00.00 น.(TweetPeriodAM) และจำนวนทวีตที่อยู่ระหว่างเวลา 0.00 น. - 6.00 น. (TweetPeriodPM) ในขณะที่ อาชีพ (Career) คือคุณลักษณะที่สำคัญน้อยที่สุดในการสร้างแบบจำลองในงานวิจัยนี้ นอกจากนั้นแล้วคุณลักษณะที่ถูกตัดออกในขั้นตอนการสร้างแบบจำลอง มีจำนวน 3 คุณลักษณะ คือ จำนวนเพื่อน (Friends) จำนวน

ผู้ติดตาม (Followers) จำนวนสัญรูปอารมณ์ด้านบวก (PositiveEmoji) ซึ่งเป็นคุณลักษณะที่ไม่ถูกนำมาพิจารณาในการสร้างแบบจำลองในงานวิจัยนี้



ภาพที่ 4.13 คุณลักษณะที่สำคัญสำหรับการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์

มหาวิทยาลัยเทคโนโลยีสุรนารี

นอกจากนั้นยังพบว่า คุณลักษณะที่เหมาะสมในการสร้างแบบจำลองในงานวิจัยนี้ คือ คุณลักษณะทั้งหมด ซึ่งประกอบด้วย คุณลักษณะด้านกายภาพ (Demographic Characteristics) คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลทั่วไป (Twitter User's Information) คุณลักษณะจากทวิตเตอร์ที่เป็นข้อมูลประเภทข้อความ (Text) คุณลักษณะจากทวิตเตอร์ที่เป็นสัญรูปอารมณ์ (Emoticon) และคุณลักษณะจากทวิตเตอร์ที่เป็นรูปภาพ (Image) ซึ่งให้ค่าประสิทธิภาพโดยรวม (F-measure) สูงที่สุด (F-measure=0.866) โดยคุณลักษณะที่สำคัญ (Feature Importance) ในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ คือ ค่าเฉลี่ยคะแนนความรู้สึกของทวีต (Sentiment Score) จำนวนการทวีตที่อยู่ระหว่างเวลา 6.00 – 00.00 น. (TweetPeriod1) และจำนวนการทวีตที่อยู่ระหว่างเวลา 00.00 – 6.00 น. (TweetPeriod2) ในขณะที่คุณลักษณะที่ไม่มีผลต่อประสิทธิภาพของแบบจำลอง คือ จำนวนเพื่อน (Friends) จำนวนผู้ติดตาม (Followers) และจำนวนสัญรูปอารมณ์ด้านบวก (PositiveEmoji)

อย่างไรก็ตาม ข้อมูลที่นำมาใช้ในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้เครือข่ายสังคมออนไลน์ในงานวิจัยนี้ ยังมีจำนวนข้อมูลไม่มากพอ และข้อมูลที่สามารถรวบรวมได้เป็นข้อมูลที่อยู่ในรูปแบบสาธารณะเท่านั้น ซึ่งในอนาคตควรมีการเพิ่มปริมาณข้อมูลที่นำมาใช้ในการสร้างแบบจำลอง และพิจารณาปัจจัยอื่น ๆ ที่เกี่ยวข้องเพิ่มเติม เช่น ข้อความในรูปภาพ ข้อความกำกับรูปภาพ รวมถึงการปรับปรุงกระบวนการ และวิธีการในการสร้างแบบจำลอง เพื่อเพิ่มประสิทธิภาพของแบบจำลอง และเลือกผลลัพธ์ที่ดีที่สุด

มหาวิทยาลัยเทคโนโลยีสุรนารี

บทที่ 5
สรุปและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

โรคซึมเศร้าเป็นหนึ่งในสาเหตุหลักที่ก่อให้เกิดปัญหาการฆ่าตัวตายตามมา โดยผู้ที่มีความเสี่ยงที่จะเกิดโรคซึมเศร้า มีจำนวนเพิ่มขึ้น และส่วนใหญ่จะไม่ได้รับการรักษาหรือไม่ทราบว่าตนเองมีภาวะซึมเศร้า ซึ่งนำไปสู่การเกิดโรคซึมเศร้า รวมถึงบุคคลในครอบครัวหรือผู้ที่ใกล้ชิดก็ไม่ทราบเช่นเดียวกัน ทำให้ไม่ได้รับการรักษาที่เหมาะสมและทันท่วงที นอกจากนั้นแล้วยังมีงานวิจัยที่พบว่า ผู้ที่มีความเสี่ยงที่จะเกิดโรคซึมเศร้าบางส่วนแสดงออกผ่านทางเครือข่ายสังคมออนไลน์ เนื่องจากเครือข่ายสังคมออนไลน์เป็นรูปแบบการสื่อสารผ่านช่องทางที่ไม่ต้องอาศัยน้ำเสียง การสบตา และการแสดงออกทางสีหน้า ซึ่งสามารถแสดงออกได้จากการโพสต์ข้อความ หรือรูปภาพในเครือข่ายสังคมออนไลน์

งานวิจัยนี้จึงนำเสนอแนวคิดในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยสามารถวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า ซึ่งบ่งบอกถึงระดับคะแนนของความเสี่ยงที่จะเกิดภาวะซึมเศร้าที่เหมาะสมกับบริบทในสังคมไทย และมีค่าความถูกต้องในระดับที่ยอมรับได้ ในส่วนของข้อมูลที่นำมาใช้ในการวิเคราะห์นั้นได้ใช้ปัจจัย หรือคุณลักษณะอื่น ๆ นอกเหนือจากข้อความ เช่น คุณลักษณะทางกายภาพส่วนบุคคล คุณลักษณะของรูปภาพ คุณลักษณะของสัญรูปอารมณ์ โดยมีวัตถุประสงค์เพื่อออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ และสมมติฐานการวิจัยคือแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์สามารถประมวลผลได้ความถูกต้องร้อยละ 80 ขึ้นไป

ความรู้ ทฤษฎี และเทคโนโลยีที่ใช้ในการออกแบบและพัฒนาแบบจำลองคือ ความรู้เกี่ยวกับโรคซึมเศร้า ภาวะซึมเศร้า เครือข่ายสังคมออนไลน์ ทวิตเตอร์ โดยใช้เทคโนโลยีในการเก็บรวบรวมข้อมูลประกอบด้วย การสร้างโปรแกรมประยุกต์บนเว็บ (Web Application) การใช้งานส่วนต่อประสานโปรแกรมประยุกต์ในทวิตเตอร์ และไพธอน (Twitter API และ Python API) ในส่วนของเทคนิคที่ใช้ในการประมวลผลข้อมูลประกอบด้วย การประมวลผลภาษาธรรมชาติ (Natural Language Processing) การวิเคราะห์ความรู้สึก (Sentiment Analysis) การประมวลผลรูปภาพ (Image Processing) การตรวจจับใบหน้า (Face Detection) การประมวลผลสัญรูปอารมณ์

(Emoticon Processing) และเทคนิคที่นำมาใช้ในการสร้างแบบจำลอง ได้แก่ การเรียนรู้ของเครื่อง (Machine Learning) เทคนิคนาอีฟเบย์ (Naïve Bayes: NB) เทคนิคการสุ่มป่าไม้ (Random Forest: RF) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) เพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron: MLP) โดยขั้นตอนการออกแบบและพัฒนาแบบจำลองประกอบด้วย การเก็บรวบรวมข้อมูล การสกัดคุณลักษณะของข้อมูล การสร้างแบบจำลอง และการประเมินความถูกต้องของแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ โดยใช้วิธีวัดค่ามาตรฐานในการประเมินความถูกต้อง 4 ค่า ได้แก่ ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าความถูกต้อง (Accuracy) และค่าประสิทธิภาพโดยรวม (F-measure)

ผลการวิจัยพบว่า แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าที่พัฒนา (Depression Risk Analysis Model: DeRAM) ซึ่งใช้การใช้การคัดเลือกคุณลักษณะด้วยวิธีเวียนเกิด (Recursive Feature Elimination : RFE) โดยใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีนเป็นฐานตัวแบบ (Model Base) ร่วมกับเทคนิคการสุ่มป่าไม้ในการสร้างแบบจำลอง ที่นำเสนอในงานวิจัยนี้ มีค่าประสิทธิภาพโดยรวม (F-measure) ร้อยละ 88.51 ซึ่งสูงกว่าเทคนิคการเรียนรู้ของเครื่องอื่นทั้งหมด และให้ค่าความถูกต้อง (Accuracy) ร้อยละ 89.00 นอกจากนี้ยังพบว่าคุณลักษณะที่สำคัญที่สุดในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ในงานวิจัยนี้ คือ ค่าเฉลี่ยคะแนนความรู้สึกของทวีต (Sentiment Score) รองลงมาคือ จำนวนทวีตที่อยู่ระหว่างเวลา 6.00 น. – 00.0 น (TweetPeriodAM) และจำนวนทวีตที่อยู่ระหว่างเวลา 0.00 น. – 6.00 น. (TweetPeriodPM) ในขณะที่อาชีพ (Career) คือคุณลักษณะที่สำคัญน้อยที่สุดในการสร้างแบบจำลองในงานวิจัยนี้ นอกจากนั้นแล้วคุณลักษณะที่ถูกตัดออกในขั้นตอนการสร้างแบบจำลอง ซึ่งเป็นคุณลักษณะที่ไม่ถูกนำมาพิจารณาในการสร้างแบบจำลองในงานวิจัยนี้ มีจำนวน 3 คุณลักษณะ คือ จำนวนเพื่อน (Friends) จำนวนผู้ติดตาม (Followers) จำนวนสัญรูปอารมณ์ด้านบวก (PositiveEmoji)

5.2 ข้อจำกัดของการวิจัย

จากการดำเนินงานวิจัย ได้พบข้อจำกัดของการวิจัย ดังนี้

1. ข้อจำกัดด้านประชากร/กลุ่มตัวอย่าง เนื่องจากการดำเนินการวิจัยในครั้งนี้ ต้องการประชากรกลุ่มตัวอย่างเป็นผู้ใช้งานเครือข่ายสังคมออนไลน์ ทวิตเตอร์ และเป็นผู้ทำแบบสอบถามภาวะซึมเศร้า 9 คำถาม จากสถิติถึงแม้ว่าผู้ใช้งานทวิตเตอร์จะมีจำนวนมาก แต่ในงานวิจัยนี้ใช้ข้อมูลการแสดงความคิดเห็นผ่านทวิตเตอร์ในส่วนของข้อมูลที่ผู้ใช้งานเปิดเป็นสาธารณะเท่านั้น ซึ่งกลุ่มตัวอย่างบางส่วนไม่ได้เปิดการแสดงความคิดเห็นในรูปแบบสาธารณะ จึงทำให้ต้องใช้ระยะเวลาในการเก็บข้อมูล และได้กลุ่มตัวอย่างในปริมาณไม่มากพอ

มหาวิทยาลัยเทคโนโลยีสุรนารี

2. ข้อจำกัดในการประมวลผลข้อมูล เนื่องจากข้อมูลที่ใช้ในงานวิจัยนี้มีหลายส่วน เช่น ข้อความ รูปภาพ สัญรูปอารมณ์ และแต่ละส่วนมีปริมาณจำนวนมาก ส่งผลต่อระยะเวลาในการประมวลผล และการเลือกเทคนิคที่นำมาใช้ในการวิจัย

5.3 การประยุกต์ผลการวิจัย

งานวิจัยนี้นำเสนอแนวคิดในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ซึ่งสามารถนำไปประยุกต์ใช้ทางด้านวิชาการ และทางด้านสังคม ดังนี้

5.3.1 การประยุกต์ใช้ผลการวิจัยทางด้านวิชาการ

1) แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ สามารถนำไปใช้เป็นตัวแบบในการพัฒนาระบบวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

2) คุณลักษณะที่สำคัญในงานวิจัยนี้สามารถนำไปเป็นข้อมูลนำเข้าในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าด้วยขั้นตอนวิธีอื่น ๆ ได้

3) ขั้นตอนวิธีในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลจากทวิตเตอร์ในงานวิจัยนี้สามารถนำไปประยุกต์ใช้กับการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลจากเครือข่ายสังคมอื่น ๆ ได้

5.3.2 การประยุกต์ใช้ผลการวิจัยทางด้านสังคม

1) แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ในงานวิจัยนี้ สามารถนำไปใช้ในการตรวจสอบภาวะซึมเศร้าในระดับเริ่มต้น ซึ่งสามารถนำไปประยุกต์ใช้ในสถานศึกษา ครอบครัว หรือสถานที่ทำงาน โดยครู-อาจารย์ ผู้ปกครอง หรือผู้บังคับบัญชา สามารถตรวจสอบระดับภาวะซึมเศร้าของนักศึกษา บุตรหลาน หรือผู้ใต้บังคับบัญชาได้จากประวัติส่วนตัว และข้อมูลการใช้งานเครือข่ายสังคมออนไลน์ โดยที่ผู้อาจมีภาวะซึมเศร้าเหล่านั้นไม่จำเป็นต้องทำแบบประเมินภาวะซึมเศร้า 9 คำถาม

2) การตรวจสอบระดับภาวะซึมเศร้าโดยใช้แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ในงานวิจัยนี้ มีส่วนช่วยให้ทราบถึงระดับภาวะซึมเศร้าในตัวบุคคล และสามารถเข้าถึงการรักษาได้ทันท่วงที

มหาวิทยาลัยเทคโนโลยีสุรนารี

5.4 ข้อเสนอแนะในการวิจัยต่อไป

1. แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ในงานวิจัยนี้ ยังมีจำนวนข้อมูลไม่มากพอ และข้อมูลที่สามารถรวบรวมได้เป็นข้อมูลที่อยู่ในรูปแบบสาธารณะเท่านั้น ควรเพิ่มปริมาณการจัดเก็บข้อมูลให้มากขึ้น

2. แบบจำลองสามารถเพิ่มประสิทธิภาพด้วยการเปรียบเทียบเทคนิคที่นำมาใช้ในการคัดเลือกคุณลักษณะ และสร้างแบบจำลองอื่น ๆ เพิ่มเติม เช่น การเรียนรู้เชิงลึกในเทคนิคต่าง ๆ

3. แบบจำลองสามารถเพิ่มปัจจัยที่เกี่ยวข้องเพื่อใช้ในกระบวนการสร้างแบบจำลอง เช่น ข้อความในรูปภาพ ระยะเวลาในการใช้งานเครือข่ายสังคมออนไลน์ การวิเคราะห์ความรู้สึก (Sentiment Analysis) ด้วยวิธีอื่น ๆ

4. แบบจำลองสามารถเพิ่มเติมกระบวนการ การให้คลาสคำตอบ (Label) ในระดับข้อความ หรือรูปภาพ เพื่อนำไปสู่การใช้เทคนิคในการสร้างแบบจำลองที่เหมาะสมกับคุณลักษณะของข้อมูลนั้น ๆ มากยิ่งขึ้น

5. แบบจำลองควรเพิ่มขั้นตอนการตรวจสอบประสิทธิภาพหรือความถูกต้องของผลการวิเคราะห์ความรู้สึก เช่น การประเมินโดยผู้เชี่ยวชาญ

มหาวิทยาลัยเทคโนโลยีสุรนารี

รายการอ้างอิง

กษิดิ์เดช ทาแป้ง. (2559). การจำแนกข้อความส่อเสียดในทวิตเตอร์ด้วยการใช้ความน่าจะเป็นของทวีต (วิทยานิพนธ์ระดับปริญญาโท). สืบค้นจาก http://cuir.car.chula.ac.th/handle/123456789/55546.

กันยา พาณิชย์ศิริ, และเบญจพร ตันตสูติ. (2559). การติดสื่อสังคมออนไลน์และภาวะสมาธิสั้นของนักเรียนชั้นมัธยมศึกษาตอนปลายในเขตกรุงเทพมหานคร. *วารสารสมาคมจิตแพทย์แห่งประเทศไทย 2559, 61*(3), 191-204.

กานดา แผ่วัฒนากุล. (2555). *การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์.* วิทยานิพนธ์ปริญญามหาบัณฑิต คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

กานดา รุณนะพงศา, และ ปโยธร อุราธรรมกุล. (2549). *รายงานการวิจัยการตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่.* ภาควิชาวิศวกรรมศาสตร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น.

กรมวุฒิ นงนุช, อนุชา ซาเฮาะ, และสุวุฒิ ตุ้มทอง. (2559). การวิเคราะห์บทความอัตโนมัติโดยใช้กระบวนการภาษาธรรมชาติ. *ใน การประชุมวิชาการระดับชาติ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ ครั้งที่ 1* (หน้า 472-479). พระนครศรีอยุธยา: สถาบันวิจัยและพัฒนา มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ.

กรมสุขภาพจิต กระทรวงสาธารณสุขและสมาคมจิตแพทย์แห่งประเทศไทย. (2546). *มาตรฐานการปฏิบัติการพยาบาลจิตเวชและสุขภาพจิตเล่มที่ 2.* กรุงเทพฯ: โรงพิมพ์ มหาวิทยาลัยธรรมศาสตร์.

กอบเกียรติ สระอุบล. (2563). *เรียนรู้ Data Science และ AI: Machine Learning ด้วย Python.* (พิมพ์ครั้งที่ 1). กรุงเทพฯ: สำนักพิมพ์ หสม มีเดย เนทเวิร์ค.

ไกรฤกษ์ เชยชื่น, และ ปกรณ์ ทู้ไพเราะ. (2557). *รายงานการวิจัยเรื่องรถเข็นสินค้าอัตโนมัติควบคุมการทำงานด้วยการประมวลผลภาพ.* มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์.

ชัชวาล บุตรมาศ. (2650). *การบันทึกขอมูลการเขาออกอาคารดวยการรูจำใบหนากรณีศึกษาศาลจังหวัดอำนาจเจริญ.* วิทยานิพนธ์ปริญญามหาบัณฑิต คณะวิทยาศาสตร์ มหาวิทยาลัยอุบลราชาธานี.

นันทิพัฒน์ พลบดี. (2561). *การแยกพื้นที่บาดแผลจากภาพถ่ายด้วยการเรียนรู้เชิงลึกและการขยายข้อมูลแบบต่าง ๆ.* วิทยานิพนธ์ปริญญามหาบัณฑิต ภาควิชาคอมพิวเตอร์ บัณฑิตวิทยาลัย มหาวิทยาลัยศิลปากร.

นันทิรา หงษ์ศรีสุวรรณ์. (2559). ภาวะซึมเศร้า Depression. *วารสาร มฉก. วิชาการ, 19*(38), 105-118.

นัธทวัฒน์ รักสะอาด, และโอฬาริก สุรินต๊ะ. (2560). การศึกษาเปรียบเทียบระหว่างวิธีการหาคุณลักษณะเฉพาะพื้นที่และวิธีการเรียนรู้เชิงลึกสำหรับการค้นคืนรูปภาพลายผ้าไหม. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, 37*(6), 736-746.

นงเยาว์ สอนจะโปะ. (2562). รูปแบบการจำแนกกลุ่มข้อความภาษาไทยแบบอัตโนมัติ โดยใช้การเรียนรู้ของเครื่องด้วยเทคนิค Unsupervised Learning ร่วมกับการประมวลผลภาษาธรรมชาติ. *วารสารวิชาการศรีปทุม ชลบุรี, 14*(4), 95-106.

นนทรัฐ ไผ่เจริญ. (2557). *การศึกษาการใช้เครือข่ายสังคมออนไลน์ (Social Network) เพื่องานข่าว : กรณีศึกษาแผนกข่าวภาษาลาว สำนักข่าววิทยุเอชียเสรี สำนักงานประเทศไทย.* วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาเทคโนโลยีการศึกษา มหาวิทยาลัยศรีนครินทรวิโรฒ.

ณัฐพงษ์ หนูสิงห์. (2560). การพัฒนาแนวทางใหม่ในการวิเคราะห์ข้อมูลรายงานการป่วยของกระทรวงสาธารณสุข โดยวิธี Map Reduce. *Journal of the Thai Medical Informatics Association 1*, 23-30.

บุญเสริม กิจศิริกุล. (2548). *ปัญญาประดิษฐ์ เอกสารคำสอนวิชา 2110654.* กรุงเทพฯ: ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย.

บุษราคัม จิตอารีย์. (2555). *ปัจจัยที่มีอิทธิพลต่อภาวะซึมเศร้าในผู้สูงอายุในชุมชน จังหวัดนครปฐม.* วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาพยาบาลศาสตรมหาบัณฑิตสาขาวิชาการพยาบาลเวชปฏิบัติชุมชนบัณฑิตวิทยาลัย มหาวิทยาลัยคริสเตียน.

ปริญญา สงวนสัตย์. (2562). Artificial Intelligence with Machine Learning, AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง. (พิมพ์ครั้งที่ 1). นนทบุรี: ไอดีซี พรีเมียร์.

ประชาสัมพันธ์กรมสุขภาพจิต. (2561). *แบบประเมินโรคซึมเศร้า 9 คำถาม (9Q).* [ออนไลน์]: Available: https://www.dmh.go.th/test/download/files/2Q%209Q%208Q%20(1).pdf

โปรยทิพย์ สันตะพันธุ์, ศิริญพร บุสหงส์, และเชาวลิต ศรีเสริม. (2562). ภาวะซึมเศร้าในวัยรุ่น: บทบาทพยาบาล. *วารสารเกื้อการุณย์, 26*(1), 187-199.

พัฒนพงษ์ คลรัตน์, และจารี ทองคำ. (2561). การเปรียบเทียบประสิทธิภาพของแบบจำลองในการพยากรณ์ความสำเร็จการศึกษาของนักเรียนระดับประกาศนียบัตรวิชาชีพ. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, 37*(3) , 380-388.

พจนา เปลี่ยนเกิด. (2557). โรคซึมเศร้า: บทบาทพยาบาลในการดูแลรักษา. *วารสารพยาบาลทหารบก, 15*(1), 18-21.

ภูริพัทร์ ทองคำ. (2559). *อัลกอริทึมแบบรวมสำหรับการเลือกคุณสมบัติของข้อมูล.* วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์.

มาสวีร์ มาศดิศรโชติ. (2557). การทำเหมืองความคิดเห็นภาษาไทย. *วารสารศรีปทุมปริทัศน์ ฉบับวิทยาศาสตร์และเทคโนโลยี, 6*(1), 120-128.

ระพีพัฒน์ แสงสิน. (2556). *ระบบศูนย์กลางข้อมูลสถาบันกวดวิชาโดยใช้เทคโนโลยีเว็บเซอร์วิส.* วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิทยาการคอมพิวเตอร์และสารสนเทศ มหาวิทยาลัยศิลปากร.

รุสลี่ สุทธวีร์กูล, และ วิไลพร แซ่ลี้. (2554). การตรวจจับใบหน้าด้วยวิธีการพื้นฐานของการจำลองรูปแบบ Haar-like Face Detection based-on Haar-like Features. *SWU Engineering Journal, 6(2),* 34-43

วัชรีวรรณ จิตต์สกุล, และสุนันทา สดสี. (2560). การวิเคราะห์การจำแนกข้อความด้วยการเปรียบเทียบความเสถียรของอัลกอริทึม. *วารสารศรีปทุมปริทัศน์ ฉบับวิทยาศาสตร์และเทคโนโลยี, 9* , 19-31.

วุฒิชัย บุญพุก, พีรพงศ์ ต่อทีฆะ, กฤตนัย ต่อศรี, และสอนกิจจา บุญโปร่ง. (2560). การประยุกต์ใช้ระเบียบวิธีการเรียนรู้เชิงลึกเพื่อจำแนกประเภทการใช้ที่ดินด้านเกษตรกรรมในประเทศไทยโดยใช้ซอฟต์แวร์คาเฟ่. *วารสารสมาคมสำรวจข้อมูลระยะไกลและสารสนเทศภูมิศาสตร์แห่งประเทศไทย ปีที่ 19, 19,* 65-77.

สถาบันวิจัยประชากรและสังคม มหาวิทยาลัยมหิดล. (2560). *สุขภาพคนไทย 2560.* สำนักพิมพ์อมรินทร์พริ้นติ้งแอนด์พับลิชชิ่ง: กรุงเทพฯ. หน้า 88-89.

สัญญา สมัยมาก, ฌานิน หาญณรงค์, สุรศักดิ์ วิเศษทรัพย์, นิติพงษ์ เสมทับ, และพรนิภา เอี่ยมบริสุทธิ์. (2560). การระบุตำแหน่งหุ่นยนต์โดยใช้การประมวลผลภาพดิจิตอล. *ในการประชุมวิชาการระดับชาติมหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์ ครั้งที่ 2* (หน้า 248-255). นครปฐม: มหาวิทยาลัยเทคโนโลยีราชมงคลรัตนโกสินทร์.

สมนึก สินธุปวน. (2546). *การวิเคราะห์กระจายคำใน ประโยคภาษาไทย โดยการโปรแกรมเชิงเจนเนติก.* วิทยานิพนธ์ปริญญามหาบัณฑิต สถาบันบัณฑิตพัฒนบริหารศาสตร์.

สำนักงานพัฒนาระบบข้อมูลข่าวสารสุขภาพ. (2018). *อัตราการป่วยโรคซึมเศร้า* [ออนไลน์]: Available: http://www.hiso.or.th/hiso/visualize/Index.php

ศักดิ์ชัย ศรีมาภรณ์. (2560). การพัฒนาระบบสารสนเทศบนแผนที่ออนไลน์ของหอพักเครือข่ายมหาวิทยาลัยราชภัฏลำปาง. *Industrial Technology Lampang Rajabhat University Journal, 5*(2), 29-37.

มหาวิทยาลัยเทคโนโลยีสุรนารี

ศิโรรัตน์ สุนทรสุข. (2559). *การวิเคราะห์ข้อความตัวอักษาหน้ายาจากภาพอย่างอัตโนมัติ.* วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยธรรมศาสตร์.

ศุภกิจ จิระศิริโชติ, จิราพร เทียนกิ่งแก้ว, และทวีชัย อวยพรกชกร. (2553). ระบบสนับสนุนการนำเข้าข้อมูลภาพสู่ฐานข้อมูลเชิงสัมพันธ์. *ในการประชุมวิชาการ National Conference on Information Technology 2010* (หน้า 378-383). กรุงเทพฯ.

ศูนย์โรคซึมเศร้า โรงพยาบาลพระศรีมหาโพธิ์. (2560). *องค์ความรู้และสาระสำคัญเกี่ยวกับโรคซึมเศร้า.* World Health Day 2017, Depression: Let's talk:1-2.

อรรถศิษฐ์ พัฒนะศิริ และ วรลักษณ์ วงศ์โดยหวัง ศิริเจริญ. (2559). สื่อสังคม และอิทธิพลต่อการตัดสินใจซื้อของผู้บริโภค Social Media, How does it Influence Customers Purchasing Decision?. *วารสารวิชาการนวัตกรรมสื่อสารสังคม, 2(8),* 152-160.

อิสรภาพ ล้อรัตนไชยยงค์. (2560). การสร้างคลังศัพท์บอกความรู้สึกในภาษาไทยจากบทวิจารย์ออนไลน์ (วิทยานิพนธ์ระดับปริญญาโท). สืบค้นจาก http://cuir.car.chula.ac.th/bitstream/ 123456789/ 60010/1/ 5780196522.pdf.

เอมิกา เหมมินทร์. (2557). *พฤติกรรมการใช้และความคิดเห็นเกี่ยวกับผลที่ได้จากการใช้เครือข่ายสังคมออนไลน์ (Social Media) ของประชาชนในเขตกรุงเทพมหานคร.* วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

อังศุมาลี สุทธิภักติ. (2553). *รายงานการวิจัยการเปรียบเทียบประสิทธิภาพของการจำแนกหมวดหมู่ของข้อคิดเห็นในแบบสอบถามปลายเปิด โดยวิธีเนอีฟเบย์และซัพพอร์ตเวคเตอร์แมชชีน.* วิทยาลัยราชพฤกษ์.

Abirami, S. & Chitra, P. (2020). Chapter Fourteen - Energy-efficient edge based real-time healthcare support system. *Advances in Computers, 117(1),* 339-368.

Acharya, T., & Ray, A. K. (2005). *Image Processing Principles and Applications*. Hoboken, New Jersey, United States of America: John Wiley & Sons, Inc.

Aldarwish, M. M., & Ahmad, H. F. (2017). Predicting Depression Levels Using Social Media Posts. *In Proceeding of IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)* (pp 277-280). Bangkok.

Alisa, K., Niran, A., Chatchawal, S., Pornpimon, P., & Choochart, H. (2010). Using an Opinion Mining Approach to Exploit Web Contentin Order to Improve Customer Relationship Management. *In Proceedings of Technology Management for Global Economic Growth. Retrieved December10, 2011 from IEEE Xplore Digital Library.*

Alsaedi, N., Burnap, P., & Rana, O. (2017). Can We Predict a Riot? Disruptive Event Detection Using Twitter. *ACM Trans. Internet Technol, 17(2),* 18-26.

Alsaeedi, A. & Khan, M. (2019). A Study on Sentiment Analysis Techniques of Twitter Data. *International Journal of Advanced Computer Science and Applications, 10*, 361-374.

Amarouche, K., Benbrahim, H. & Kassou, I. (2015). Product Opinion Mining for Competitive Intelligence. *In Proceedings of The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)* (pp 358-365). Morocco.

Amasha, M. A., & Alkhalaf, S. (2014). The effect of using facebook markup language (fbml) for designing an e-learning model in higher education. *International Journal of Research in Computer Science, 4*(5), 1-9.

Armstrong, H. (2015). *Machines Thai learn in the wild*. London, UK: Nesta.

Aroonpiboon, B. (2012). Social Media and Social Network. Retrieved from http://www.thailibrary.in.th/2012/01/10/social-media-social-network/.

Barhan, A., & Shakhomirov, A. (2012). Methods for Sentiment Analysis of twitter messages. *In Proceeding of The 12th Conference of Fruct Association* (pp 215-222).

Bellaachia, A., & Guven, E. (2006). *Predicting Breast Cancer Survivability Using Data Mining Tech-niques*. The George Washington University.

Bollen, J., Mao,H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2(1),* 1-8.

Breiman, L. (2001). Random Forests. Machine Learning 45, 1: 5-32.

Buranajun, P., Sasananan, M., & Sasananan, S. (2007). *Prediction of Product Design and Development Success using Artificial Neural Network.* In Proceedings of International Conference on Operations and Supply Chain Management (pp 1-8). Bangkok, Thailand.

Chen, W-L., Kan, C-D., Lin, C-H., Chen, Y-S., & Mai, Y-C. (2017). Hypervolemia screening in predialysis healthcare for hemodialysis patients using fuzzy color reason analysis. *International Journal of Distributed Sensor Networks, 13(1),* 1-13.

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *In Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, Vol. 2*. (pp 128–137).

Cohn, J.F., Kruez, T.S., Matthews, I., Yang, Y., Nguyen, M.H., Padilla, M.T., ... Torre, F.D.L. (2009). *Detecting depression from facial actions and vocal prosody.* In Proceedings of Affective Computing and Intelligent Interaction. (pp 1-7). 10.1109/ACII.2009.5349358.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). *Random Forests. Ensemble Machine Learning:* Methods and Applications. (45), 157-176.

Dai, W., & Ji, W. (2014). A MapReduce Implementation of C4.5 Decision Tree Algorithm. *International Journal of Database Theory and Application, 7*(1), 49–60.

Domingosand, P., & Pazzani, M. (1997). *On the optimality of the simplp Bayesian classifier under zero-oneloss.* Machine Learning *29,* 103–130.

Lundh, F. (2011). Pillow. Retrieved from https://pillow.readthedocs.io/en/stable/index.html

Gonzalez, R. C., Woods, R. E. (2018). *Digital Image Processing* (4 th Edition). England: Pearson Education Limited.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning.* MIT Press.

Guo, W., Yang, H., Liu, Z., Xu, Y., & Hu, B. (2021). Deep Neural Networks for Depression Recognition Based on 2D and 3D Facial Expressions Under Emotional Stimulus Tasks. *Frontiers in Neuroscience, 15,* 1-19.

Guntuku, H. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). *Detecting depression and mental illness on social media: an integrative review.* Current Opinion in Behavioral Sciences. *18*, 43-49.

Hall, M. (2560). *Facebook: Encyclopædia Britannica.* [On-line]: Available: https://www.britannica.com/topic/Facebook

Hootsuite. (2019). *Digital 2019 Thailand: Social Media Audiences Quarterly Growth*. [On-line]. Available: https://www.slideshare.net/DataReportal/digital-2019-thailand-january-2019-v01

Hu, Q., Li, A., Heng, F., Li, J., & Zhu, T. (2015). *Predicting Depression of Social Media User on Different Observation Windows.* In Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology *(WI-IAT)* (pp. 361-364), Singapore.

Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *In Proceedings of the 8th International Conference on Weblogs and Social Media,* ICWSM 2014.

มหาวิทยาลัยเทคโนโลยีสุรนารี

Ibraheem, N. A., Hasan, Mokhtar M. H., Khan, R. Z., & Mishra, P. K. (2012). Understanding Color Models: A Review. *ARPN Journal of Science and Technology, 3(2),* 265-275.

Islam, M. R., Kabir, A., Ahmed, A., Raihan M. Kamal, A., Wang, H., & Ulhaq, A. (2018). Depression detection from social network data using machine learning techniques. *Health Information Science and Systems.* 6, 8.

Jackson, P., & Moulinier, I. (2007). *Natural Language Processing for Online Applications: Text retrieval, extraction and categorization.* John Benjamins Publishing: Amsterdam, Netherlands.

Jiang, H., Hu, B., Liu, Z., Yan, L., Wang, T., Liu, F., Kang, H., & Li, X. (2017*). Investigation of different speech types and emotions for detecting depression using different classifiers.* Speech Communication, *1*(90), 39-46.

Kabir, M. F., Rahman, C. M., Hossain, A., & Dahal, K. (2011). Enhanced Classification Accuracy on Naive Bayes Data Mining Models. *International Journal of Computer Applications,* 28(3), 9-16. https://doi.org/10.5120/3371-4657

Kaplan, A. M., & Haenlein M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons, 1(53),* 59-68.

Katchapakirin, K., Wongpatikaseree, K., Yomaboot, P., & Kaewpitakkun, Y. (2018). *Facebook Social Media for Depression Detection in the Thai Community.* In Proceeding of International Joint Conference on Computer Science and Software Engineering (JCSSE). Bangkok. Thailand.

Kashyap, R. (2014). *Big Data Analysis for User Health Tracking.* A thesis submitted in partial fulfillment of the requirements For the degree of Master of Science in Software Engineering. CALIFORNIA STATE UNIVERSITY, NORTHRIDGE.

King, B. (2019). *How To Use Web APIs in Python 3* [On-line]: Available: https://www.digitalocean.com/community/tutorials/how-to-use-web-apis-in-python-3

Ko, M. N., Cheek, G. P., & Shehab, M. (2010). Social-Networks Connect Services. *Journal of Computer, 43* (8), 37-43.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine, 16*(9), 606–613.

Kumar, A., Kaur, A. & Kumar M. (2018). *Face detection techniques: a review.* Artificial Intelligence Review, *52(1).*

มหาวิทยาลัยเทคโนโลยีสุรนารี

Larsson, K., Baker, S., Silins, I., Guo, Y., Stenius, U., Korhonen, A., & Berglund, M. (2017). Text mining for improved exposure assessment. *Journal Public Library of Science, 12*(3), 1-21.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning.* Insight Review. (521):436-444.

Mahittivanicha, N. (2020). *สถิติและพฤติกรรมการใช้ social media ทั่วโลก Q1 ปี 2020.* [On-line]: Available: https://www.twfdigital.com/blog/2020/02/global-social-media-usage-stats-q1-2020/

Marra, T. (2004). *Depressed and Anxious: The Dialectical Behavior Therapy Workbook for Overcoming Depression and Anxiety.* United States: New Harbinger Publications.

Meier,J.D., Homer, A., Hill, D., Taylor, J., Bonsode, P., Wall, L., Jr, R. B., & Bogawat, A. (2008). *Web Application Architecture Guide.* Microsoft Corporation. United States.

Milette, K., Hudson, M., Baron, M., Thombs, B. D., & Canadian Scleroderma Research Group. (2010). *Comparison of the PHQ-9 and CES-D depression scales in systemic sclerosis: Internal consistency reliability, convergent validity and clinical correlates.* Rheumatology (Oxford, England), *49*(4), 789–796.

Mitchell, T. M. (1997). *Machine learning.* New York, NY: McGraw-Hill.

Musto, C., Semeraro, G., Polignano, M. (2014). A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts. *CEUR Workshop Proceedings, 1314,* 59-68.

Neppelenbroek, M., Lossek, M., Janssen, R., & Boer, T. (2011). *Twitter An Architectural Review. Netherlands:* Software Architecture Faculty of Science University of Utrecht.

Nilsson, N. J. (1998). *Introduction to Machine Learning.* Department of Computer Science Stanford University. Stanford, CA.

O'Dea, B., Wan, S., Batterhamc, P.J., Calear, A.L., Paris, C., & Christensen, H. (2015). Detecting suicidality on Twitter. *Internet Interventions, 2(2015),* 183–188

OpenCV. (2022). OpenCV. Retrieved from https://opencv.org/

Orabi, H. A., Buddhitha, P., Orabi, H. M. & Inkpen, D. (2018). *Deep Learning for Depression Detection of Twitter Users.* In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp 88–97). New Orleans, Louisiana.

Pandey, M., & Kumar Sharma, V. (2013). A Decision Tree Algorithm Pertaining to the Student Performance Analysis and Prediction. *International Journal of Computer Applications, 61*(13), 1–5.

Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in Twitter. *In Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD).*

Parekh, R. (2017). *What Is Depression?.* [On-line]. Available: https://www.psychiatry.org/patients-families/depression/what-is-depression

Radloff, L.S. (1977). *The CES-D scale: A self report depression scale for research in the general population.* Applied Psychological Measurement 1, 385-401.

Reece, A.G., & Danforth, C.M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Sci, 6*(15), 1-12.

Reece, A.G., Reagan, A.J., Lix, K.L.M., Dodds, P.S., Danforth, C.M., & Langer, E.J. (2017). Forecasting the onset and course of mental illness with Twitter data. *Sci Rep 7, 13006 (2017).* https://doi.org/10.1038/s41598-017-12961-9

Boomgaard, Rein van den. (2017). *Image Processing and Computer Vision.* [On-line]: Avaliable: https://staff.fnwi.uva.nl/r.vandenboomgaard/IPCV20162017/LectureNotes/IP/index.html

Rizvi, D. Q. (2011). A Review on Face Detection Methods. *Journal of Management Development and Information Technology, 11,* 1-12.

Scassellati, B. (1998). Finding Eyes and Faces with a Foveated Vision System. Retrieved from https://www.researchgate.net/publication/239034657_Finding_Eyes_and_Faces_with_a_Foveated_Vision_System.

Shaik, R., & Inkpen, D. (2020). *Using Social Media for Mental Health Surveillance: A Review. ACM Comput. Surv.,* 53(6), Article 129 (December 2020), 1-31. doi:https://doi.org/10.1145/3422824

Siricharoen, W. V. (2012) Social Media, How does it Work for Business?. *International Journal of Innovation, Management and Technology, 3(4),* 476-479.

Sood, A., Hooda, M., Dhirn, S., & Bhatia, M. (2018). An Initiative to Identify Depression using Sentiment Analysis: A Machine Learning Approach. *Indian Journal of Science and Technology. 11*, 1–20.

Statista Research Department. (2019). *Number of Twitter users in Thailand from 2014 to 2019* [On-line]: Available: https://www.statista.com/statistics/490584/twitter-users-thailand/

Sun, X., Zhang, C., & Li, L. (2019). Dynamic emotion modelling and anomaly detection in conversation based on emotional transition tensor. *Information Fusion. 46*, 11-22.

Yamane, T. (1967). *Statistics An Introductory Analysis.* 2nd Ed. New York. Harper and Row.

Yang, X., Zhang, Q., & Li, Z. (2019). Contour Detection in Cassini ISS images based on Hierarchical Extreme Learning Machine and Dense Conditional Random Field. *Research in Astron. Astrophys, 0(2019),* 1-12.

Tjepkema, L. (2019). *Top 5 Social Media Predictions for 2019* [On-line]. Available: https://www.emarsys.com/resources/blog/top-5-social-media-predictions-2019/

Twitter. (2020). *Get started with the Twitter developer platform* [On-line]: Available: https://developer.twitter.com/en/docs/basics/getting-started

University of Toronto. (2007). Machine Learning and Data Mining Neural Network Toolbox in Matlab [On-line]: Available: http://www.cs.toronto.edu/~ruiyan/csc411/Tutorial4.pdf

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data:* Springer Series in Statistics. Springer-Verlag Berlin, Heidelberg.

Vateekul, P., & Koomsubha, T. (2016). A study of sentiment analysis using deep learning techniques on Thai Twitter data. *In Proceeding of 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)* (pp 1–6).

Viola, P., Jones, M. (2001). Rapid object detection using a Boosted cascade of simple features. Proc. Int'l Conf. Computer Vision and Pattern Recognition, pp. 1-9.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., & Bao, Z. (2013). *A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network.* In: Li J. et al. (eds) Trends and Applications in Knowledge Discovery and Data Mining. PAKDD 2013. Lecture Notes in Computer Science (Vol. 7867). Springer, Berlin, Heidelberg.

Wen, S. (2021). Detecting Depression from Tweets with Neural Language Processing. *Journal of Physics: Conference Series, 1792(2021),* 012058.

WHO. (2018). *Depression* [On-line]: Available: http://www.who.int/en/news-room/fact-sheets/detail/depression

Zhenqiu, L. (2012). *Design of Automatic Question Answering System Base on CBR.* In Proceeding of 2012 International Workshop on Information and Electronics Engineering (IWIEE) (pp. 981-985). Harbin, China.

Zhu, C., Li, B., Li, A., & Zhu, T. (2016). *Predicting Depression from Internet Behaviors by Time-Frequency Features.* In Proceeding of IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 383-390), Omaha, NE.

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาคผนวก

คลังคำที่แสดงถึงภาวะซึมเศร้า (Depression Term Corpus)

มหาวิทยาลัยเทคโนโลยีสุรนารี

คลังคำที่แสดงถึงภาวะซึมเศร้า (Depression Term Corpus)

ID	Depression Term (Thai)	Depression Term (English)	Pos(s)	Neg(s)
1	เบื่อ	bored	0.375	0.25
2	เศร้า	sad	0.125	0.75
3	เหงา	lonely	0.125	0.2
4	กังวล	worry	0.25	0.5
5	กลุ้มใจ	dismal	0	0.625
6	เจ็บปวด	painful	0.125	0.75
7	เซ็ง	sick	0.125	0.75
8	ท้อ	discouraged	0.5	0.25
9	ไร้ค่า	worthless	0	0.5
10	ไม่มีใคร	no one	0	0
11	ปัญหา	problem	0	0.625
12	กดดัน	force	0	0
13	เครียด	stressed	0.125	0.75
14	ซึมเศร้า	depressed	0	0.875
15	โรคซึมเศร้า	depression	0	0.375
16	ท้อแท้	disheartened	0.5	0.25
17	เหนื่อย	tired	0.25	0
18	เหนื่อยล้า	Tired	0.25	0
19	นอนไม่หลับ	can not sleep	0	0
20	อยากตาย	want to die	0	0
21	หลบหนี	escape	0	0.375
22	ไม่ดี	not good	0	0
23	หมดกำลัง	exhausted	0	0.25
24	หมดกำลังใจ	discouraged	0.25	0.25
25	หมดความรู้สึก	exhausted	0	0.25
26	หมดค่า	run out	0	0
27	หมดจิตหมดใจ	out of mind	0	0
28	ความขมขื่น	bitterness	0.125	0.625
29	ความขมุกขมัว	insomnia	0	0.75

มหาวิทยาลัยเทคโนโลยีสุรนารี

ID	Depression Term (Thai)	Depression Term (English)	Pos(s)	Neg(s)
30	ความขุ่นเคือง	resentment	0.125	0.625
31	ความข้องใจ	grievances	0	0
32	ความฉิบหาย	Woe	0	0
33	ความชั่วช้า	iniquity	0	0.75
34	ความทรุดโทรม	dilapidation	0.125	0
35	ฟุ้งเฟ้อ	epicurean	0.125	0
36	คลุมเครือ	vague	0.25	0.375
37	วิกฤติ	crisis	0.25	0.375
38	เจ็บปวด	painful	0.125	0.75
39	ขัดแย้ง	controversial	0.625	0.125
40	รังเกียจ	mind	0	0.5
41	อกหัก	heartbroken	0	0.625
42	มืด	dark	0.125	0.125
43	ฮือ	haha	0	0
44	ลึกลับ	mysterious	0.25	0.5
45	กลุ้มใจ	dismal	0	0.625
46	น่ากลัว	scary	0	0.75
47	เกียจคร้าน	lazy	0	0
48	ขวางโลก	perverse	0	0.375
49	เศร้าใจ	depressed	0	0.5
50	ไม่มีทางรักษา	no cure	0	0
51	สิ้นหวัง	despair	0.25	0.625
52	ไร้ผล	in vain	0	0
53	เศร้าซึม	Depressed	0	0.5
54	สลดใจ	Depressed	0	0.5
55	ท้อแท้	disheartened	0.5	0.25
56	น่าเบื่อ	boring	0.375	0.25
57	กระสับกระส่าย	restless	0	0.875
58	กระเซอะกระเซิง	aimlessly	0.25	0
59	เครียด	stressed	0.125	0.75

มหาวิทยาลัยเทคโนโลยีสุรนารี

ID	Depression Term (Thai)	Depression Term (English)	Pos(s)	Neg(s)
60	คบไม่ได้	can't date	0	0
61	โง่	stupid	0	0.75
62	จังไร	be accursed	0	0
63	ไม่สน	don't care	0	0
64	จัญไร	beastly	0.25	0
65	เจือ	alloy	0.25	0.25
66	เจือจาง	dilute	0	0.125
67	ต่ำ	low	0	0.25
68	ต่ำต้อย	humble	0	0.375
69	มีปัญหา	have a problem	0	0
70	ไม่เอาไหน	no way	0	0
71	ไม่ได้เรื่อง	don't matter	0	0
72	สับสนวุ่นวาย	chaos	0	0.25
73	สับสน	confused	0	0.75
74	ไม่รู้สึกตัว	unconscious	0	0.375
75	ไม่สบายใจ	uneasy	0	0.5
76	ลาตาย	goodbye	0	0
77	ลาก่อน	goodbye	0	0
78	ร้าวราน	angry	0	0.875

มหาวิทยาลัยเทคโนโลยีสุรนารี

ภาคผนวก ข

คุณลักษณะของกลุ่มตัวอย่างแบ่งตามระดับภาวะซึมเศร้า

มหาวิทยาลัยเทคโนโลยีสุรนารี

คุณลักษณะของกลุ่มตัวอย่างแบ่งตามระดับภาวะซึมเศร้า

ระดับภาวะซึมเศร้า	**จำนวนกลุ่มตัวอย่าง (คน)**
ผู้ที่ไม่มีภาวะซึมเศร้า (level0)	154
มีภาวะซึมเศร้าระดับน้อย (level1)	149
มีภาวะซึมเศร้าระดับปานกลาง (level2)	70
มีภาวะซึมเศร้าระดับรุนแรง (level3)	32
รวม	405

คุณลักษณะของกลุ่มตัวอย่างแบ่งตามเพศ

เพศ	**จำนวนกลุ่มตัวอย่าง (คน)**
หญิง	358
ชาย	43
อื่น ๆ	4
รวม	405

คุณลักษณะของกลุ่มตัวอย่างแบ่งตามอายุ

อายุ	**จำนวนกลุ่มตัวอย่าง (คน)**
<20	91
20-29	309
30-39	3
40-49	2
50-59	0
>60	0
รวม	405

คุณลักษณะของกลุ่มตัวอย่างแบ่งตามระดับการศึกษา

ระดับการศึกษา	**จำนวนกลุ่มตัวอย่าง (คน)**
มัธยม	22
ปริญญาตรี	378
สูงกว่าปริญญาตรี	5
รวม	405

มหาวิทยาลัยเทคโนโลยีสุรนารี

ประวัติผู้เขียน

นางสาวสุดา ทิพย์ประเสริฐ เกิดเมื่อวันที่ 16 พฤศจิกายน พ.ศ. 2526 เกิดที่อำเภอปักธงชัย จังหวัดนครราชสีมา สำเร็จการศึกษาระดับปริญญาตรีวิทยาศาสตร์บัณฑิต (วิทยาการคอมพิวเตอร์) (เกียรตินิยมอันดับสอง) จากมหาวิทยาลัยราชภัฏนครราชสีมา เมื่อ พ.ศ. 2548 ภายหลังสำเร็จการศึกษาได้เข้าทำงานที่สำนักคอมพิวเตอร์ มหาวิทยาลัยราชภัฏนครราชสีมา ในตำแหน่งนักวิชาการคอมพิวเตอร์ ทำหน้าที่พัฒนางานระบบสารสนเทศเพื่อการบริหาร ต่อมาในปี พ.ศ. 2552 ได้รับทุนการศึกษาจากมหาวิทยาลัยราชภัฏนครราชสีมาเพื่อเข้าศึกษาต่อในระดับปริญญาโท หลักสูตรวิทยาการสารสนเทศมหาบัณฑิต สาขาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีสุรนารี หลังจากสำเร็จการศึกษาในระดับปริญญาโท ได้รับตำแหน่งงานเป็นหัวหน้างานระบบสารสนเทศเพื่อการบริหาร เป็นเวลา 4 ปี หลังจากนั้นจึงได้เข้าทำงานที่คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีราชมงคลอีสาน ในตำแหน่ง อาจารย์ประจำสาขาวิชาระบบสารสนเทศ ต่อมาในปี พ.ศ. 2561 ได้รับทุนพัฒนาบุคลลากรจากมหาวิทยาลัยเทคโนโลยีราชมงคลอีสานเพื่อศึกษาต่อในระดับปริญญาเอก ในหลักสูตรปรัชญาดุษฎีบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีสุรนารี