การทำนายข้อความบนโซเชียลมีเดียเพื่อตรวจจับภาวะซึมเศร้า
Predictive text on social media to detect Depression

นายกฤตณัฐ แสงศรี
Krittanut Sangsri

นางสาวริตาภรณ์ ธรรมวัติ
Ritaporn Thammawat

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตร์บัณฑิต
ภาควิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ ปีการศึกษา พ.ศ. 2563

คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ

ชื่อหัวข้อโครงงาน การทำนายข้อความบนโซเชียลมีเดียเพื่อตรวจจับภาวะซึมเศร้า
Predictive text on social media to detect Depression

นิสิต นายกฤตณัฐ แสงศรี 60102010328
นางสาวริตาภรณ์ ธรรมวัติ 60102010344

ปริญญา วิทยาศาสตร์บัณฑิต (วท.บ.)
ภาควิชา วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาโครงงาน อ.ดร. วีรยุทธ เจริญเรืองกิจ

ลงชื่อ.....
(อ.ดร. วีรยุทธ เจริญเรืองกิจ)
อาจารย์ที่ปรึกษาโครงงาน

บทคัดย่อ

เนื่องจากปัจจุบันมีผู้คนมากมายใช้โซเชียลมีเดียในในการสื่อสาร และระบายความรู้สึกหรือสิ่งที่ไม่สามารถพูดออกมาในชีวิตได้ โครงการนี้จึงมีความสนใจในการนำข้อมูลจากโซเชียลมีเดียมาวิเคราะห์ โดยมีปัจจัยที่เกี่ยวข้องกับโรคซึมเศร้า เช่น เวลา คำที่เป็นสัญญาณของโรคซึมเศร้า คำเชิงบวกและเชิงลบ อีโมจิ การวิเคราะห์อารมณ์ และอื่นๆ เป็นต้น จากนั้นนำมาพัฒนาการเรียนรู้โดยใช้ pre-train model ในการสร้าง 300 เวกเตอร์มิติ จากข้อความและสิ่งบ่งชี้ด้านภาษามาเป็นคุณสมบัติในการสร้าง Machine learning โมเดลเพื่อใช้ในการตรวจจับประโยคที่บ่งชี้ถึงสัญญาณของโรคซึมเศร้า งานวิจัยนี้ได้ดึงข้อมูลภาษาไทยจาก Twitter จำนวน 21,905 ประโยค มาใช้ในการสร้างโมเดลผู้วิจัยได้นำเสนอการสำรวจข้อมูลและการทดสอบโดยการแบ่งข้อมูล 30% สำหรับการทดสอบประสิทธิภาพ โดยวัดจากค่า ความถูกต้องซึ่งผลลัพธ์ในการทำนายข้อความที่ดีที่สุดมีค่าความถูกต้อง 94%

Abstract

Nowadays the people use social media for communication and convey feelings that can't speak openly in real life. This project uses the data from social media to analyze several factors associated with depression such as time, words that show signs of depression, sentiment word, and emoji. This project uses pre-train models to create 300 - dimensional vectors from text and language indication to develop models to identify the sign of depression. This research retrieves Thai sentences from Twitter 21,905 sentences to create models. Several exploratory data analysis is presented. The test data 30 % is used to evaluate the model. The best result of a depression detection is at 94%

กิตติกรรมประกาศ

โครงงานการทำนายข้อความบนโซเชียลมีเดียเพื่อตรวจจับภาวะซึมเศร้า (Predictive text on social media to detect Depression) สามารถสำเร็จลุล่วงได้ด้วยความกรุณาจากอาจารย์ ดร. วีรยุทธ เจริญเรืองกิจ ซึ่งเป็นที่ปรึกษาของโครงงานนี้ อาจารย์ที่ปรึกษาได้ให้คำแนะนำ ข้อเสนอแนะ และข้อคิดเห็นต่างๆ ที่เป็นประโยชน์อย่างยิ่งในการทำงานวิจัย อีกทั้งยังช่วยในขั้นตอนการทำงานบางส่วนที่ติดขัด ตลอดจนการปรับปรุงแก้ไขข้อบกพร่องของการทำโครงงานเป็นกลุ่ม ที่เกิดขึ้นระหว่างการดำเนินงานทำโครงงาน รวมทั้งให้กำลังใจแก่คณะผู้จัดทำเสมอมา คณะผู้จัดทำขอกราบขอบพระคุณเป็นอย่างสูง

ขอขอบคุณท่านอาจารย์ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศรีนครินทรวิโรฒทุกท่าน ที่กรุณาให้คำแนะนำ ข้อคิดเห็น ข้อเสนอแนะและแนวทางการแก้ไขปัญหาในการจัดทำโครงงาน และรวมถึงให้ความรู้ที่เป็นประโยชน์อื่นๆ ต่อการทำโครงงานการทำนายข้อความบนโซเชียลมีเดียเพื่อตรวจจับภาวะซึมเศร้า (Predictive text on social media to detect Depression)

ขอขอบคุณเพื่อนๆ ทั้งภายในภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยศรีนครินทรวิโรฒและเพื่อนๆ ของกลุ่มผู้จัดทำที่แนะนำวิธีการ หรือเทคนิคในเรื่องต่างๆ และให้คำปรึกษาเพื่อการทำโครงงานการทำนายข้อความบนโซเชียลมีเดียเพื่อตรวจจับภาวะซึมเศร้า (Predictive text on social media to detect Depression)

สุดท้ายนี้ขอขอบคุณครอบครัวของกลุ่มผู้จัดทำที่คอยสนับสนุนในด้านต่างๆ รวมทั้งการให้กำลังใจ และเป็นที่ปรึกษาที่มีส่วนช่วยในการทำโครงงานการทำนายข้อความบนโซเชียลมีเดียเพื่อตรวจจับภาวะซึมเศร้า (Predictive text on social media to detect Depression)

สารบัญ

สารบัญรูปภาพ

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญ

จากข้อมูลสถิติขององค์การอนามัยโลก(WHO) ในปีค.ศ.2017 ระบุว่า มีผู้ป่วยโรคซึมเศร้าทั่วโลกประมาณ 322 ล้านคน หรือคิดเป็นร้อยละ 4.4 ของประชากรโลก ซึ่งโรคซึมเศร้า มีสาเหตุหลักที่เกิดจากความผิดปกติของสารเคมีในสมอง และสาเหตุอื่นๆ เช่น ภาวะเศรษฐกิจการเงิน ความผิดหวัง ความรัก ความสูญเสีย เป็นต้น หากไม่สามารถจัดการกับปัญหาได้จะทำให้ทุกปัญหากลายเป็นสาเหตุที่ทำให้เกิดความเครียด วิตกกังวล จนพัฒนาไปสู่ภาวะซึมเศร้าและคิดทำร้ายตัวเองหรือฆ่าตัวตายได้

ในปัจจุบันมีผู้คนมากกว่า 50% ของประชากรโลกสามารถเข้าถึงอินเตอร์เน็ตและใช้โซเชียลมีเดียได้ โซเชียลมีเดียจึงเป็นช่องทางที่หลายๆ คนสามารถระบายความรู้สึกแท้จริงของตนเองที่ไม่สามารถระบายออกมาในโลกความเป็นจริงได้ อีกทั้งยังไม่ต้องเปิดเผยตัวตนและมีการสนับสนุนจากกลุ่มคนที่มีสภาวะเดียวกันหรือเข้าใจซึ่งกันและกัน จึงมีการเลือกใช้โซเชียลมีเดียเพื่อเป็นช่องทางในการสื่อสาร

โดยกลุ่มผู้จัดทำเห็นว่าข้อมูลที่อยู่บนโซเชียลมีเดียมีความน่าสนใจ ค่อนข้างมีความเพียงพอต่อการนำข้อมูลมาศึกษา วิเคราะห์หาข้อความที่มีความเสี่ยงของภาวะซึมเศร้าจึงสนใจหาแนวทางจัดทำโมเดลที่สามารถทำนายข้อความบนโซเชียลมีเดียที่มีความเสี่ยงต่อภาวะซึมเศร้า เพื่อเป็นการช่วยให้ตระหนักถึงการใช้คำต่างๆ บนโซเชียลมีเดีย

1.2 วัตถุประสงค์ของโครงงาน

1.2.1 เพื่อประยุกต์การใช้ Machine Learning เพื่อวิเคราะห์ข้อความที่บ่งบอกถึงการเป็นโรคซึมเศร้า หรือมีความเสี่ยงต่อการเป็นโรคซึมเศร้า

1.2.2 เพื่อเรียนรู้และทำความเข้าใจเกี่ยวกับผู้ป่วยโรคซึมเศร้า

1.2.3 เพื่อเป็นแนวทางให้แก่ผู้ที่สนใจ อาจจะนำไปพัฒนาต่อยอดได้

1.3 ขอบเขตของโครงงาน

1.3.1 ใช้ข้อมูลภาษาอังกฤษซึ่งเป็นชุดข้อมูลจากงานวิจัยที่อ้างอิงรวบรวมโดยงานวิจัยแรกใช้ Twitter API ในการรวบรวมข้อมูลและจัดการซึ่งข้อมูลมีจำนวน 3,457 แถว และงานวิจัยที่สองรวบรวมมาจาก Reddit และ Blog มีจำนวน 3,165 แถว

1.3.2 ข้อมูลภาษาไทยที่รวบรวมจาก Twitter จำนวน 21,905 แถว

1.3.3สามารถวิเคราะห์ข้อความที่มีความเสี่ยงต่อการเป็นโรคซึมเศร้าได้อย่างถูกต้อง

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 ช่วยวิเคราะห์ข้อความที่บ่งบอกถึงการเป็นโรคซึมเศร้า หรือมีความเสี่ยงต่อการเป็นโรคซึมเศร้า

1.4.2 ได้รับรู้และตระหนักถึงการใช้ข้อความต่างๆ ที่มีผลต่อโรคซึมเศร้า

1.4.3 ได้นำความรู้จากการเรียนทั้งหมดมาประยุกต์ใช้ทำงานร่วมกัน

บทที่ 2
องค์ความรู้ที่เกี่ยวข้อง

2.1 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF เป็นวิธีการหาคำสำคัญขอเอกสาร ซึ่งเกิดจากผลคูณของสองค่า คือ Term Frequency (TF) และ Inverse Document Frequency (IDF)

- Term Frequency (TF) คือ ความถี่ของคำหรือเทอม วิธีการคำนวณ TF คือ การนำจำนวนคำที่เลือกทั้งหมดหารด้วยจำนวนคำทั้งหมดใน document
    - Tf (term, document) = f (term, document) / (all term in document)
    - F (term, document) = ความถี่ของจำนวนคำที่เลือกใน document
    - all term in document = จำนวนคำทั้งหมดใน document

- Inverse Document Frequency (IDF) คือ การวัดความสำคัญของคำในเอกสารทั้งหมด ถ้าคำที่พบในหลายเอกสาร หมายความว่าคำนั้นมีความสำคัญลดลง วิธีการคำนวณ IDF คือ การนำจำนวนเอกสารทั้งหมดหารด้วยจำนวนเอกสารที่มี Term นั้น ๆ
    - Idf (term, allDoc) = log(N/df(t))
        N = จำนวนเอกสารทั้งหมด
        df(t) = จำนวนเอกสารที่มี term t

ดังนั้นการหา TF-IDF คือ TF คูณกับ IDF จะได้ TF-IDF weight ของ termนั้น ๆ ซึ่งจะช่วยกรองคำที่มีการใช้ร่วมกันในหลายๆเอกสารออกไป และระบุได้ว่าเอกสารแต่เอกสารมีความเป็นเรื่องอะไร

2.2 Sentiment Analysis

Sentiment Analysis (การวิเคราะห์ความรู้สึก) เป็นการวิเคราะห์อารมณ์และความรู้สึกจากข้อความ เพื่อบ่งบอกความรู้สึกของผู้คนที่มีต่อบางสิ่งบางอย่าง เช่น ความรู้สึกเชิงบวก (Positive) หรือ ความรู้สึกที่เชิงลบ (Negative)

ในโครงงานส่วนของข้อมูลภาษาอังกฤษมีการทำงานโดยใช้ VADER Sentiment Analysis หรือ Valence Aware Dictionary and Sentiment Reasoner เป็นการวิเคราะห์ที่ถูกปรับให้เหมาะสมต่อการวิเคราะห์ความรู้สึกบนโซเชียลมีเดีย มีการคำนวณโดยการรวมคะแนนความจุของแต่ละคำศัพท์ ค่าเกณฑ์ทั่วไปสำหรับการจำแนกประโยคว่าเป็นเชิงบวก เชิงลบ หรือว่าเป็นกลาง คือ

- เชิงบวก (Positive sentiment) มีค่า compound >= 0.05
- กลาง (Neutral sentiment) มีค่า compound อยู่ระหว่าง -0.05 ถึง 0.05
- เชิงลบ (Negative sentiment) มีค่า compound <=0.05

ในส่วนของภาษาไทยใช้ PythaiNLP ในการวิเคราะห์ความรู้สึก โดยจะมีคลังคำศัพท์ที่ถูกแบ่งเป็นคำเชิงลบ และคำเชิงบวก

2.3 Bag of Words (BOW)

โมเดลที่ใช้กันแพร่หลายในงานจัดแบ่งประเภทข้อความ ในโมเดลของ BOW กลุ่มของคำจะถูกอธิบายด้วยกระเป๋าคำ Bag of words หรือกลุ่มรวมของคำ โดยไม่ได้คำนึงถึงหลักไวยากรณ์ ความถี่ที่พบ และลำดับของคำ โดยนำมาใช้เป็น Feature ในการเทรนตัวจัดแบ่งข้อความ Classifier

2.4 Word2Vec

Word2Vec คือ โมเดลที่ใช้สร้าง word embedding พัฒนาโดยทีมนักวิจัยของ Google นำโดย Tomas Mikolov จะใช้วิธีการคำนวนตัวเลขของคำนั้น ๆ จาก context รอบๆ คำนั้น (ไอเดียมาจาก language model)

ปัญหาที่ทำให้เกิดการทำ Word2Vec

1. Coverage - คำที่เกิดร่วมกันหรือเกี่ยวข้องกัน อาจไม่ได้อยู่ติดกัน
2. Space - หากมีคำมากก็จะใช้พื้นที่มาก
3. Speed - เมื่อใช้พื้นที่มาก ความเร็วในการคำนวณมากขึ้น

Coverage สามารถแก้ไขปัญหาด้วยการ Word embedding ที่ชื่อว่า CBOW(Continuous Bag-of-Words) และ Skip-Gram สามารถใช้ Neural Network 3 Layers สร้างขึ้นมาได้

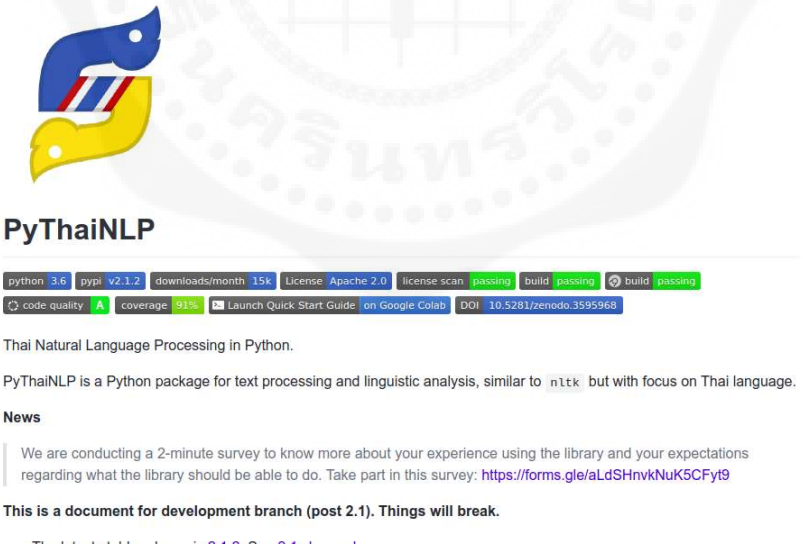
- CBOW(Continuous Bag-of-Words) คือ การใช้ context หลายคำเพื่อหา next word 1 คำ
- Skip-Gram คือ การใช้ Context 1 คำ เพื่อหา next word หลายคำ

2.5 Fasttext

Fasttext เป็นวิธีการ word embedding ที่เป็นส่วนขยายของโมเดล Word2Vec โดย Fasttext จะแทนแต่ละคำเป็นอักขระ ซึ่งFasttext จะแปลงแต่ละคำแบบ n-gram ของแต่ละตัวอักษร ตัวอย่างเช่น คำว่า “artificial” โดยกำหนดเป็น n=3 Fasttext จะแปลงค่าคำนี้เป็น ar, art, rti, tif, ifi, fic, ici, ial, al เป็นต้น วิธีการ Fasttext ช่วยให้สามารถครอบคลุมถึงความหมายของคำสั้นๆและ ทำให้ระบบเข้าใจถึง Prefixes และ Subfixes

2.5 PyThaiNLP

PyThaiNLP คือ Library package ของ Python ใช้ในการประมวลผลข้อความวิเคราะห์ทางภาษา ซึ่งใช้ในภาษาไทย มีฟังก์ชันการทำงานที่หลากหลาย เช่น ตัดคำทางภาษาไทย วิเคราะห์ชนิดของคำทางไวยากรณ์ เป็นต้น



รูปภาพที่ 2.1 *รูปภาพ PyThaiNLP*
ที่มา : https://1th.me/tQzP3

2.6 Twint

Twint เป็นเครื่องมือที่ถูกเขียนมาด้วยภาษา Python ใช้สำหรับการดึงข้อมูลจากโปรไฟล์ทวิตเตอร์โดยไม่จำเป็นต้องใช้ Twitter API

รูปภาพที่ 2.2 *รูปภาพ Twint*
ที่มา : *https://www.les-faunes.ch/en/twint-now-available-on-les-faunes-ch/*

2.7 Ngrok

Ngrok เป็นเครื่องมืออำนวยความสะดวกที่พัฒนาโดย Github ให้บุคคลอื่นสามารถเข้าใช้แอพพลิเคชั่นหรือเว็บไซต์ที่กำลังทำงานจากเครื่อง localhost ได้โดยที่บุคคลอื่นจะเข้าใช้ได้จาก url ที่ ngrok ทำการสุ่มสร้างขึ้นมาในทุกครั้งที่ปิดและเปิดใช้งานใหม่

รูปภาพที่ 2.3 *รูปภาพ ngrok*
ที่มา: https://bit.ly/3fWx7n6

2.8 Facebook API

ทาง Facebook มีการเปิด Library Service ให้นักพัฒนาเข้าถึงข้อมูลต่างๆของ Facebook ด้วย Facebook API โดยมีการให้บริการหลากหลายรูปแบบ เช่น Facebook messenger API เพื่อเข้าถึงการส่งข้อความต่างๆ เป็นต้น

2.9 Flask

Flask คือ Web Framework ที่ถูกพัฒนาขึ้นมาสำหรับ Python เพื่อใช้ร่วมกับ Webserver โดย Flask เป็น micro framework เนื่องจากไม่ต้องการเครื่องมือหรือ Library มากมาย และไม่ต้องการฐานข้อมูล

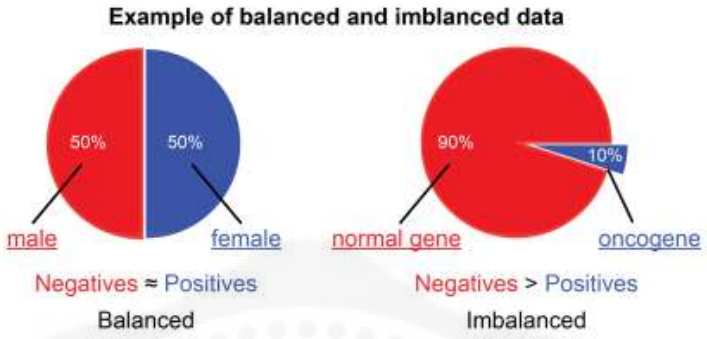
รูปภาพที่ 2.4 รูปภาพ Flask
ที่มา : https://saixiii.com/python-flask-web-application/

2.10 Imbalance Data

Imbalance Data คือ ปัญหาจำนวนชุดข้อมูลแต่ละชุดมีปริมาณที่แตกต่างกันมากหรือเรียกว่าไม่สมดุล โดยการปรับปรุงข้อมูลที่มีความไม่สมดุลให้กลายเป็นข้อมูลที่มีความสมดุลด้วยเทคนิคการสุ่มเลือกข้อมูล (Data Sampling Technique) โดยแบ่งเป็น 2 วิธีหลัก คือ

- Under sampling เป็นวิธีการสุ่มลดจำนวนข้อมูลกลุ่มส่วนใหญ่ให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนกลุ่มข้อมูลส่วนน้อยโดยการลดจำนวนข้อมูลจากคลาสส่วนมากลง
- Over sampling เป็นวิธีการที่ใช้ในการเพิ่มข้อมูลส่วนน้อยให้มีจำนวนใกล้เคียงหรือเท่ากับจำนวนข้อมูลส่วนมาก โดยการสุ่มเกินข้อมูลเพื่อเพิ่มข้อมูลอาจเพิ่มจากข้อมูลเดิมหรือสร้างข้อมูลขึ้นใหม่จากตัวอย่างข้อมูลเดิม ในที่จะใช้เทคนิค SMOTE จะทำการสุ่มสร้างข้อมูลจากข้อมูลส่วน

น้อยตามจำนวนที่กำหนด โดยจะสร้างข้อมูลสังเคราะห์เพิ่มจากข้อมูลตัวอย่างด้วยการวัดระยะห่างจากข้อมูลตัวอย่างไปยังจุดข้อมูลใกล้เคียง แล้วสุ่มสร้างข้อมูลสังเคราะห์



รูปภาพที่ 2.5 รูปภาพแสดงถึง imbalance data
ที่มา : https://medium.com/analytics-vidhya/what-is-balance-and-imbalance-dataset-89e8d7f46bc5

2.11 LogisticRegression (LR)

Logistic Regresstion Analysis ถูกนำมาใช้เพื่อทำนายว่า จะเกิดเหตุการณ์หนึ่งขึ้นหรือไม่หรือมี โอกาสเกิดขึ้นมากน้อยเพียงใด โดยมีการกำหนดค่าตัวแปรตัวหนึ่งหรือหลายตัวที่คาดว่าจะส่งผลต่อการเกิดเหตุการณ์นั้นๆ และในที่สุดก็จะทำให้เข้าใจสาเหตุการเกิดเหตุการณ์นั้นๆ ได้ในที่สุด

จะมีการทำ classification โดย output ออกมาเป็น regression จากนั้นนำไป mapping กับ Sigmoid function (หลักคณิตศาสตร์) จะได้ว่าข้อมูลนั้นเป็นคลาสใด(เป็นไปได้แค่ 2 คลาสเท่านั้น)

2.12 Multinomial naïve bayes (MultinomialNB)

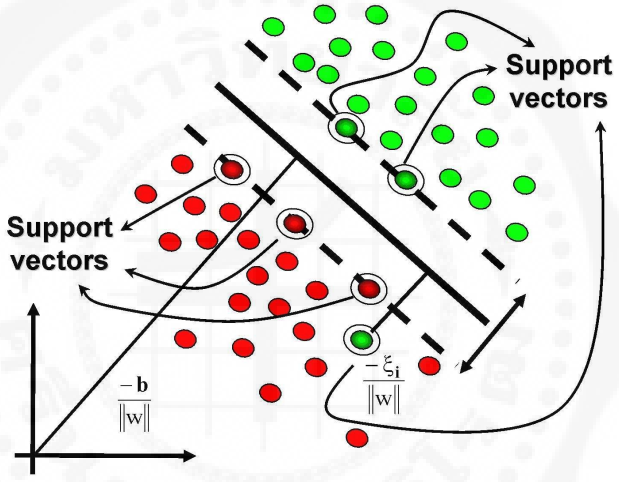
สามารถใช้ประยุกต์ได้กับงานแยกประเภท (Classify) เช่น ใช้วิเคราะห์ความรู้ให้ทำการแยกประเภทความรู้สึกโดยมีสมมติฐานด้านคุณสมบัติอย่างง่ายๆ ซึ่งทำให้เรียกการแบบประเภทแบบนี้วา นาอีฟเบย์ แนวคิดการแยกประเภทเริ่มจากคิดว่าจำนวนคำทั้งหมดนำมารวมกัน เหมือนเปรียบได้ดั่งถุงใส่คำ (bag of words) ซึ่งหมายความว่าจะไม่มีการเรียงลำดับคำ มีเพียงความถี่ของคำที่ใช้ซ้ำที่จะนำไปใช้ประโยชน์ได้

สำหรับงานแยกประเภทความรู้สึกมักใช้ Multinomial naïve bayes โดยคิดว่าประเภทหนึ่งๆ (ที่แยกประเภท) มีคุณสมบัติได้มาก จัดอยู่ในรูปเวกเตอร์ สำหรับแต่ละ

คลาส y เมื่อให้ n เป็นจำนวนคุณสมบัติ แต่ในงานแยกประเภทเอกสารใช้แทนจำนวนคำหรือจำนวนคำแทนแต่ละคุณสมบัติและความน่าจะเป็นของแต่ละคุณสมบัติ

2.13 Support Vector Machine (SVM)

Support Vector Machine (SVM) เป็นอัลกอริทึมช่วยแก้ปัญหาการจำแนกข้อมูลโดยอาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลเข้าสู่กระบวนการเทรนให้ระบบเรียนรู้ หลักการของ SVM คือ การให้ input ที่ใช้ฝึกเป็นเวคเตอร์ใน Space N มิติ จากนั้นทำการสร้างไฮเปอร์เพลน (Hyperplane) เพื่อแยกกลุ่มเวคเตอร์ของ input ออกเป็นประเภทต่างๆ



รูปภาพที่ 2.6 รูปภาพแสดงวิธีการทำงานของ SVM
ที่มา :http://kokzard.blogspot.com/2011/10/jfjkdshfkjsldf.html

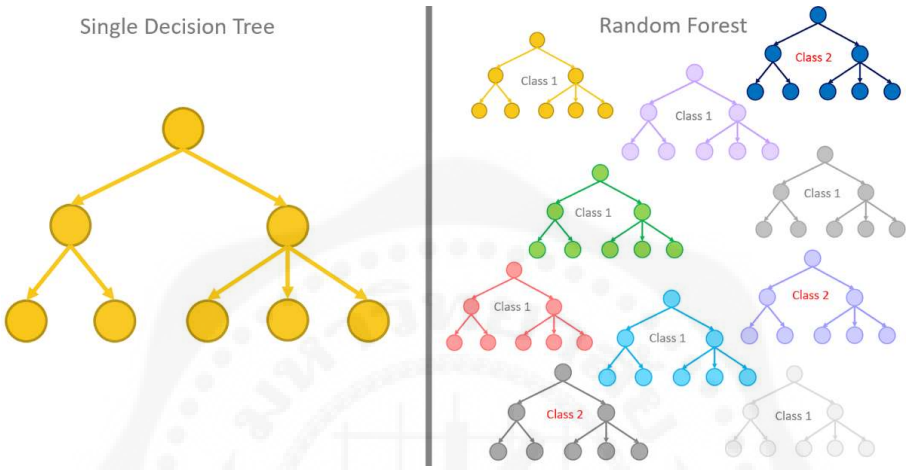
2.14 Random Forest

Random Forest เป็นโมเดลประเภทหนึ่งของ Machine Learning ถูกพัฒนาขึ้นจาก Decision Tree ซึ่ง Random Forest เป็นการเพิ่มจำนวน Tree เป็นหลายๆ ต้น ทำให้ประสิทธิภาพในการทำงานสูงขึ้น โดยมี Hyper-parameters ที่น่าสนใจดังนี้

- N_estimators คือ จำนวน tree ใน Random Forest จำนวน tree ที่มากขึ้น จะทำให้โมเดลมีประสิทธิภาพดีขึ้น
- Oob_score คือ Out of bag score เป็นการระบุว่าจะใช้ข้อมูลในส่วนที่ไม่ถูก sample ไปทำ training set ซึ่งเทียบเท่ากับการทำ Validation จาก

training set โดยไม่ต้องทำการแบ่งข้อมูลสำหรับ Validation มีประโยชน์สำหรับชุดข้อมูลเล็กๆ

- Min_samples_leaf คือ การระบุจำนวนข้อมูลขั้นต่ำใน leaf node ของแต่ละ decision tree เป็นการช่วยลดการเกิด overfitting



รูปภาพที่ 2.7 รูปภาพแสดงวิธีการทำงานของ Random Forest
ที่มา : https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be65147

2.15 งานวิจัยที่เกี่ยวข้อง

2.15.1 Identifying Depression on Social Media

Kali Cornn ได้ทดลอง model หลายๆ อย่างเพื่อเปรียบเทียบว่า model ไหนจะมีประสิทธิภาพที่ดีกว่าในการแยกระหว่าง depressed comment และ non-depressed comment จากข้อมูล comment ของ Reddit โดยโมเดลที่ผู้วิจัยได้ทำการทดลองมีผลลัพธ์ออกมาได้ดี เช่น

1. BERT model ได้ accuracy 85.7% ซึ่งทำออกมาได้ดีกว่า baseline model เล็กน้อย
2. CNN model without embedding ผู้วิจัยกล่าวว่าเป็น model ที่ทำออกมาได้ดีที่สุด ได้ accuracy 92.5% การใช้ one-dimensional convolutional layers มีประโยชน์ในการ

จำแนกข้อความเฉพาะที่จะพิจารณา ในทางตรงข้าม RNN model สามารถพิจารณาข้อความทั้งหมดซึ่งอาจจะไม่ตรงประเด็นที่ต้องการจะพิจารณา

3. CNN model with embeddings เป็นโมเดลที่ทำออกมาได้แย่ที่สุดจากการวิจัยที่ผู้วิจัยทดลอง

ข้อดีของการใช้ Character-based CNN ซึ่งตรงข้ามกับ Word-based CNN คือ ข้อความบน social media นั้นไม่ซ้ำกันมากและไม่จำเป็นต้องมีความเป็นทางการ เช่น ผู้ใช้สามารถใช้อีโมจิ เป็นต้น การใช้ Word-based CNN จึงดูเหมือนจะเป็นข้อเสีย

2.15.2 Facebook Social Media for Depression Detection in the Thai community

Kantinee Katchapakirin และคณะ ได้ทำการศึกษาความเป็นไปได้ในการตรวจจับโรคซึมเศร้าในสังคมไทยผ่าน Facebook ซึ่งเป็น Social media ที่เป็นที่นิยมในไทย โดยอาสาสมัครที่เข้าร่วมโครงการวิจัยนี้จะร่วมทำแบบสำรวจสุขภาพจิต 20 ข้อ ที่เรียกว่า Thai mental health questionnaire (TMHQ) ซึ่งจะสามารถจำแนกกลุ่มอาสาสมัครได้เป็นสองกลุ่ม คือ Depress และ Non-Depress ในส่วนของกลุ่มที่ผู้วิจัยได้ใช้คือกลุ่ม Depress

หลังจากที่กลุ่มผู้วิจัยได้ข้อมูลนำมาเก็บเป็น list เรียบร้อย จากนั้นทำการแปลภาษาจากภาษาไทยเป็นภาษาอังกฤษ เนื่องจากไม่มี NLP (Natural Language Processing) resource ในภาษาไทย โดยใช้ Google Cloud Translation API และใช้ NLTK python library เพื่อจัดการกับ text ที่ถูกแปลมา

ในส่วนของ Machine Learning Algorithm กลุ่มผู้วิจัยได้ใช้หลากหลายแบบ 3 Algorithm ที่ให้ความแม่นยำ คือ

- Support Vector Machine (SVM) - กลุ่มผู้วิจัยใช้ Weka ในการ implement sequential minimal optimization algorithm (SMO) เพื่อการเทรน Support vector classification เนื่องจากชุดข้อมูลของกลุ่มผู้วิจัยมีขนาดเล็กจึงไม่สามารถแบ่งเป็น training และ test set ผลลัพธ์ความแม่นยำของ SVM model ออกมาดีกว่า

majority vote ที่ใช้เป็น baseline ของการประเมินผล

- Random Forest - กลุ่มผู้วิจัยได้ใช้ RapidMiner ด้วย Random Forest Algorithm เพื่อเทรนและประเมินโมเดล มีผลลัพธ์ความแม่นยำประมาณ 84.6%
- Deep Learning - กลุ่มผู้วิจัยได้ใช้ RapidMiner ด้วย Deep Learning Algorithm ถูกนำไปใช้กับ automatic parameters optimization เพื่อให้เหมาะสมกับชุดข้อมูล โดยถูกแบ่งออกเป็น 2 กลุ่ม คือ positive sentiment และ negative sentiment โดยผลลัพธ์ประสิทธิภาพของโมเดลได้ค่าความแม่นยำประมาณ 85%

2.15.3 Monitoring tweets for Depression to detect at-risk users

Zanaira Jamil ต้องการที่จะสามารถหาความเสี่ยงของข้อความที่โพสต์ลงในโซเชียล และในส่วนของผู้ใช้ด้วย ข้อมูลต้องทำการแก้ปัญหา imbalance เพราะเกือบจะทั้งหมดของข้อมูลในการเทรนไม่มีแนวโน้มการเป็นโรคซึมเศร้า ในส่วน feature ที่ผู้วิจัยเลือกใช้

- Polarity words คือ เพื่ออธิบายว่าคำไหนเป็นคำพูดเชิงบวกหรือเชิงลบโดยแปลงเป็นค่าตัวเลขโดยจะอยู่ระหว่าง -5 (คำพูดเชิงลบ) ไปจนถึง +5(คำพูดเชิงบวก)
- Depression words คือ นับคำที่สามารถสื่อได้ถึงโรคซึมเศร้า
- Pronouns คือ นับว่ามีการพูดถึงบุคคลที่หนึ่งและบุคคลที่สอง
- Bag of Word

จากนั้นผู้วิจัยได้ทำการนำ feature ต่าง ๆ มาทำการทดลองโดยจะใช้ feature ที่แตกต่าง หรือเพิ่มมากขึ้นในแต่ละการทดลองจากนั้นจะนำมา train โดยใช้ SVM ซึ่งยกตัวอย่างผลที่ได้คือ ใช้ Feature Depression words , Pronouns polarity จากนั้นนำมาเทรนโดยใช้ Support Vector Machine (SVM) มีการแก้ปัญหา imbalance โดยวิธี Down ค่าที่ได้ออกมา Accuracy เท่ากับ 0.6102 Precision เท่ากับ 0.1237 Recall เท่ากับ 0.8020 ซึ่งผู้วิจัยจะให้ความสำคัญกับระดับผู้ใช้มากกว่า หมายความว่าเมื่อดูในระดับข้อความที่โพสไปแล้วจะทำการดูขึ้นไปอีกขั้นถึงระดับผู้ใช้ว่ามีโอกาสที่จะเป็นโรคซึมเศร้ามากน้อยแค่ไหน

2.15.4 The Psychology of Word Use in Depression Forums in English and in Spanish: Testing Two Text Analytic Approaches

Nairan Ramirez-Esparza และคณะได้ทำการศึกษาเพื่อหาการใช้คำของภาวะซึมเศร้าในฟอรัมอังกฤษและสเปน โดยมีสองแนวทางในการศึกษา

1. สังเกตเครื่องหมายทางภาษาของภาวะซึมเศร้าในฟอรัมของอังกฤษและสเปน โดยผู้วิจัยรวบรวมโดยใช้ bulletin board systems (bbs) และใช้โปรแกรม LIWC2001 เพื่อเปรียบเทียบหมวดภาษาในกลุ่มต่างๆ ของภาษาอังกฤษและภาษาสเปน ผลแสดงให้เห็นว่าตัวชี้นำทางภาษาเกี่ยวข้องกับภาวะซึมเศร้าสูงกว่าการโพสต์แบบไม่มีการกดทับของภาษา
2. วิเคราะห์รูปแบบที่ผู้คนใช้เมื่อพูดถึงภาวะซึมเศร้าในฟอรัมของอังกฤษและสเปน โดยผู้วิจัยรวบรวมโดยใช้ bulletin board systems (bbs) ผลการวิเคราะห์คนที่มีภาวะซึมเศร้าที่เขียนภาษาสเปนมีแน้วโน้มที่จะแสดงถึงความกังวลเชิงความสัมพันธ์มากกว่า คนที่มีภาวะซึมเศร้าที่เขียนภาษาอังกฤษมีแนวโน้มที่จะพูดถึงความกังวลด้านยา

2.15.5 Early Detection of Depression: Social Network Analysis and Random Forest Techniques

การศึกษานี้ใช้ข้อมูลจากโซเชียลมีเดียมาสำรวจวิธีการในการตรวจหาโรคซึมเศร้าในระยะเริ่มต้น โดยวิเคราะห์ชุดข้อมูลเพื่อระบุลักษณะพฤติกรรมจากแง่มุมต่างๆของงานเขียนที่มีการความเกี่ยวข้องกับผู้เป็นโรคซึมเศร้า ได้แก่ จำนวนคำในประโยค, การกระจายข้อความ ,ช่องว่างของเวลา และช่วงเวลา

งานวิจัยนี้มีการสำรวจชุดข้อมูล ผู้วิจัยสังเกตได้ว่าผู้ที่มีแนวโน้มเป็นโรคซึมเศร้ามีช่วงเวลาที่โพสต์ข้อความในเวลาช่วง 05.00 - 23.00 น. และ 18.00 - 24.00 น.

ผู้วิจัยเสนอ 2 แนวทางการที่แตกต่างกันที่ขึ้นอยู่กับ Machine Learning

1. ใช้ 1 Random forest (RF) classifier with 2 threshold
2. ใช้ 2 independent RF classifier : ตัวแปรแรกเป็นตัวระบุอาสาสมัครที่มีภาวะซึมเศร้า และอีกตัวแปรเป็นตัวระบุว่าไม่มีภาวะซึมเศร้า

2.15.6 ภาษาไทยการสื่อสารและโรคซึมเศร้า : การสำรวจเบื้องต้นเพื่อเข้าใจ

โรคซึมเศร้าและผู้มีภาวะซึมเศร้าในสังคมไทย

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาภาวะซึมเศร้าของนักศึกษามหาวิทยาลัยธรรมศาสตร์และสำรวจภาษาที่ใช้สื่อสารเกี่ยวกับโรคซึมเศร้า เพื่อค้นหาลักษณะทางภาษาที่บ่งชี้สัญญาณของโรคซึมเศร้า

ข้อมูลที่ใช้แบ่งเป็น 2 ส่วน

- ส่วนที่ 1 ข้อมูลจากการใช้แบบสอบถาม
- ส่วนที่ 2 ข้อมูลที่ได้จากการบันทึกชีวิตประจำวันและสัมภาษณ์อาสาสมัครกลุ่มที่เป็นโรคซึมเศร้า

ผลของการศึกษาด้านอุบัติการณ์ภาวะซึมเศร้าของนักศึกษามหาวิทยาลัยธรรมศาสตร์ พบว่ามีอัตราร้อยละ 23.3 ซึ่งมากกว่าอัตราของประชากรไทยที่เข้าข่ายภาวะโรคซึมเศร้า

ผลของการศึกษาภาษาซึมเศร้าพบว่ามีลักษณะเด่นที่การใช้คำบ่งชี้บุคคลที่เป็นผู้พูดเป็นผู้ถูกกระทำ ใช้ศัพท์ที่สื่ออารมณ์ทางลบ ศัพท์ที่สื่อถึงความหมดหวัง ไร้ค่า และความตาย ศัพท์ที่สื่อถึงการกระทำและพฤติกรรมก้าวร้าว การทำร้ายตัวเอง

บทที่ 3
วิธีการดำเนินโครงการ

3.1 วิธีการดำเนินงาน

3.1.1 เริ่มต้นและวางแผนโครงงาน
- ค้นคว้าหาข้อมูลศึกษาแนวคิดและทฤษฎีที่เกี่ยวข้อง
- ศึกษาเทคโนโลยีที่จะนำมาใช้ในโครงงาน

3.1.2 รวบรวมข้อมูล
- รวบรวมข้อมูลจากโซเชียลมีเดีย
- เตรียมข้อมูลเพื่อให้พร้อมต่อการนำไปใช้ในโครงงาน

3.1.3 วิเคราะห์ข้อมูล
- ทำ Pre - processing
- เทรนข้อมูลในอัลกอริทึมเพื่อสร้าง Baseline จัดการข้อมูลโดยใช้การเพิ่มฟีเจอร์ต่างๆ เข้าไป
- เทรนข้อมูล ทดสอบและปรับปรุงประสิทธิภาพ

3.1.4 สรุปและเผยแพร่งานวิจัย
- จัดทำ Application เพื่อนำเสนอตัวอย่างการทำนาย
- สรุปผลงานเพื่อนำเสนอ

3.2 แผนการดำเนินงาน

ขั้นตอนการดำเนินงาน	2564			
	มกราคม	กุมภาพันธ์	มีนาคม	เมษายน
ประชุมและวางแผนการทำงาน				
1.ศึกษาแนวคิด ทฤษฎี และเทคโนโลยี				
เก็บรวบรวมข้อมูล				
1.รวบรวมข้อมูลจาก Twitter				

2.เตรียมข้อมูล				
3.ศึกษาข้อมูล				
การพัฒนาระบบ				
1. พัฒนาและปรับปรุงประสิทธิภาพ				
บทสรุป				
1.จัดทำ Application เพื่อนำเสนอตัวอย่าง				
2.นำเสนอโครงงาน				

รูปภาพที่ 3.1 รูปภาพตารางแผนการดำเนินงาน

3.3 อุปกรณ์และเครื่องมือที่ใช้

3.3.1 ฮาร์ดแวร์ (Hardware)

- Computer
- Notebook

3.3.2 ซอร์ฟแวร์ (Software)

- Google Collaboratory
- Jupyter notebook
- Command
- Visual Studio code

3.3.3 ภาษาที่ใช้

- Python

3.4 ชุดข้อมูล (Data set) ที่ใช้งาน

ข้อมูลจากงานวิจัย Depression Sentiment Analysis with Twitter Data และ Classification of Depression on Social Media Using Text Mining ซึ่งเป็นชุดข้อมูลที่ผู้วิจัยรวบรวมโดยใช้ Twitter API ข้อมูลที่นำมาใช้มีจำนวน 3574 แถว

ข้อมูลจากงานวิจัย The psychology of word use in Depression Forums in English and in Spanish: Testing Two Text Analytic approaches ซึ่งข้อมูลเป็นไฟล์ .txt โดยเมื่อทำการอ่านและนำออกมามีจำนวน 3048 แถว

โดยรวมแล้วจะมีข้อมูลทั้งหมด 6622 แถว

ข้อมูลที่รวบรวมจากทวิตเตอร์โดยกรองผ่าน hashtag และเวลา แบ่งเป็น 2 แบบ คือ

1. Sentiment = 1 คือข้อมูลปกติที่ไม่มีแนวโน้มต่อโรคซึมเศร้า จำนวน 11673 แถว
2. Sentiment = -1 คือข้อมูลที่รวบรวมโดยใช้ Hashtag เช่น #โรคซึมเศร้า #ซึมเศร้าเพื่อนรัก #ซึมเศร้า เป็นต้น จำนวน 10232 แถว

3.5 การออกแบบและพัฒนา

3.5.1 การเตรียมชุดข้อมูล

3.5.1.1 ทำการรวบรวมข้อมูลจากทวิตเตอร์โดยใช้ Twint มีการกรองผ่าน Hashtag และเวลา ข้อมูลที่รวบรวมมาแบ่งเป็นสองชุดคือ

a) ชุดที่มีการตั้งสมมติฐานว่ามีแนวโน้มต่อภาวะซึมเศร้า จะใช้การกรองผ่าน Hashtag (#โรคซึมเศร้า #ซึมเศร้าเพื่อนรัก #ซึมเศร้า) ในช่วงเวลา 00.00 - 05.00 น. และ 18.00 -23.00 น. มีจำนวน 10232 แถว

	tweet	sentiment
0	เหนื่อย พอแล้ว เราไม่ไหวแล้ว ยอมแพ้ให้กับโรคนี้...	-1
1	ในบางครั้ง ฉันเคยคิดอยากหายไปจากโลกนี้ #โรคซึม...	-1
2	ไม่อยากให้ถึงอนาคตข้างหน้าเลยอ่ะ ไม่รู้จะเป็นย...	-1
3	ร้องไห้แทนการกินข้าวแต่ละมื้อ55555555 #โรคซึมเ...	-1
4	คือแบบคุยกับใครได้บ้าง อึดอัด อยากระบายมากๆ เห...	-1
...	...	...
10227	พยายามหาเหตุผลว่าทำไมต้องมีชีวิตอยู่ต่อ พยายาม...	-1
10228	เป็นคนอินเทรนด์นะกุเนี่ย😒 เป็นแม่งทั้ง #มะเร็ง...	-1
10229	I wanna shout out!!!! I'm Bipolar Disorder: De...	-1
10230	บางครั้งเหตุผลที่ทำร้ายตัวเองก็เพราะแค่อยากให้...	-1
10231	บางคนนี่เขาไม่รู้สึกผิดหรอที่ทำให้คนๆนึงเป็นซึ...	-1

10232 rows × 2 columns

รูปภาพที่ 3.2 ตัวอย่างชุดข้อมูลที่มีแนวโน้มภาวะซึมเศร้า

b) ชุดข้อมูลปกติที่ไม่มีแนวโน้มต่อภาวะซึมเศร้า มีจำนวน 11673 แถว

	tweet	sentiment
0	@J13SB ดีใจเหมือนกันนะคะ ขอบคุณมากๆเลยน้าาที่ต...	1
1	@creamiii25 น้องครีมมี่ ขอบคุณมากนะคะ🥲❤️ ดีใจท...	1
2	เพลงในสต้อคเยอะ ก็มีเพลงเล่นได้เยอะ ดีใจที่คนน...	1
3	ชั้นไม่ควรได้รับคำชม งานชั้นมันห่วยแตก แต่ได้ค...	1
4	@Kaew_SCM @PolarKen @PalmHyukLover นั้นง้ัย..ผ...	1
...	...	...
9995	น่ารักมากก ดีใจแทนเค้าเลย https://t.co/YkKv4X...	1
9996	พิจ้านดูมีน้ำมีนวลขึ้นเยอะเลยยยยย ดีใจจจจ ส่วน...	1
9997	@FaiFan_PRP ดีใจที่ชอบนะครับ ✨ น้องบอก มากินผ...	1
9998	เขาอุ้มกันๆๆๆๆๆ ดีใจจนนั้มตาไหล #ImperialEven...	1
9999	มึงงง แบบดีใจมากเลยอ่ะ คู่ชิปเจอกันทีนี่แทบร้อ...	1

10000 rows × 2 columns

รูปภาพที่ 3.3 ตัวอย่างชุดข้อมูลที่ไม่มีแนวโน้มภาวะซึมเศร้า

3.5.1.2 นำข้อมูลทั้งสองมารวมกันจะมีจำนวน 21905 แถว และแยกอีโมจิออกมาไว้อีกหนึ่งคอลัมน์ เพื่อเก็บไว้ใช้ในการพิจารณาภายหลัง

3.5.1.3 ทำความสะอาดข้อมูลโดย ลบอักขระพิเศษ ลิงค์ เมนชั่น แฮชแท็กและเก็บแยกออกมาไว้อีกหนึ่งคอลัมน์

	tweet	sentiment	emoji	text
0	เหนื่อย พอแล้ว เราไม่ไหวแล้ว ยอมแพ้ให้กับโรคนี...	-1	[]	เหนื่อย พอแล้ว เราไม่ไหวแล้ว ยอมแพ้ให้กับโรคนี้
1	ในบางครั้ง ฉันเคยคิดอยากหายไปจากโลกนี้ #โรคซึม...	-1	[]	ในบางครั้ง ฉันเคยคิดอยากหายไปจากโลกนี้
2	ไม่อยากให้ถึงอนาคตข้างหน้าเลยอ่ะ ไม่รู้จะเป็นย...	-1	[]	ไม่อยากให้ถึงอนาคตข้างหน้าเลยอ่ะ ไม่รู้จะเป็นย...
3	ร้องไห้แทนการกินข้าวแต่ละมื้อ55555555 #โรคซึมเ...	-1	[]	ร้องไห้แทนการกินข้าวแต่ละมื้อ
4	คือแบบคุยกับใครได้บ้าง อึดอัด อยากระบายมากๆ เห...	-1	[]	คือแบบคุยกับใครได้บ้าง อึดอัด อยากระบายมากๆ เห...
...	...	...	...	...
21900	น่ารักมากก ดีใจแทนเค้าเลย https://t.co/YkKv4X...	1	[]	น่ารักมากก ดีใจแทนเค้าเลย
21901	พิจ้านดูมีน้ำมีนวลขึ้นเยอะเลยยยยย ดีใจจจจ ส่วน...	1	[]	พิจ้านดูมีน้ำมีนวลขึ้นเยอะเลยยยยย ดีใจจจจ ส่วน...
21902	@FaiFan_PRP ดีใจที่ชอบนะครับ ✨ น้องบอก มากินผ...	1	[✨, 🥲]	ดีใจที่ชอบนะครับ น้องบอก มากินผมซี่ กรักมือรอแ...
21903	เขาอุ้มกันๆๆๆๆๆ ดีใจจนนั้มตาไหล #ImperialEven...	1	[]	เขาอุ้มกันๆๆๆๆๆ ดีใจจนนั้มตาไหล
21904	มึงงง แบบดีใจมากเลยอ่ะ คู่ชิปเจอกันทีนี่แทบร้อ...	1	[]	มึงงง แบบดีใจมากเลยอ่ะ คู่ชิปเจอกันทีนี่แทบร้อ...

21905 rows × 4 columns

รูปภาพที่ 3.4 ตัวอย่างชุดข้อมูลหลังทำ pre-process

3.5.2 ทำการสร้าง Baseline

- แปลงข้อมูลให้อยู่ในรูปเวกเตอร์โดยใช้ FastText

```
 4.26765531e-03,  5.17885925e-03, -5.28506127e-02,  2.44235229e-02,
-4.59444523e-03,  2.61836737e-04, -2.99553521e-02,  8.50612579e-02,
 1.05120900e-02, -1.59920700e-02, -1.40769803e-02,  6.80253329e-03,
 1.02433893e-04, -4.64742848e-02, -2.89157364e-02, -9.10252002e-03,
-4.16046025e-02, -1.18312858e-02,  4.57578546e-04,  4.31059545e-02,
 2.46266398e-02,  3.91795190e-02,  1.83321419e-03,  4.33298341e-02,
 1.73082391e-02,  1.40286446e-02,  2.89684079e-03, -3.33451164e-02,
 1.56326918e-02,  1.86656844e-02,  2.22096476e-02, -2.71963645e-02,
-1.83610253e-02, -4.56163349e-03, -5.87974096e-03,  1.73642846e-02,
 4.44281565e-02, -2.77683925e-03, -1.61705577e-03, -1.06567071e-02,
-1.45563074e-02, -4.99023332e-03,  1.26779866e-02,  9.09587762e-03,
 3.57038952e-03,  1.51834459e-02,  8.11387992e-03, -1.12482446e-02,
 4.94769976e-03,  1.96278399e-02, -7.23541730e-03,  2.27325451e-02,
 1.69234969e-02,  2.18236207e-02,  9.25337751e-03, -4.25133595e-03,
-8.27822048e-02, -2.53505943e-02,  2.63393027e-02, -1.18572902e-02,
-1.30889797e-02,  4.06712198e-02,  2.30828134e-02,  4.13463670e-02,
 1.87927152e-01,  6.05922834e-03, -7.36105359e-03, -1.06239521e-02,
-2.13799938e-02,  1.06800760e-02,  2.23097468e-02, -1.82750097e-02,
-7.17951657e-04, -2.00678067e-03,  9.67787134e-03,  1.27278356e-02,
 1.38937487e-02, -9.57457150e-03, -1.60723858e-02,  5.47112400e-03,
-5.35342532e-04,  2.52476250e-02, -1.06625695e-02, -6.08932817e-03])
```

รูปภาพที่ 3.5 ตัวอย่างข้อมูลที่ถูกแปลงให้อยู่ในรูปเวกเตอร์

- นำข้อมูลที่ทำความสะอาดแล้วมาทำการเรียนรู้และทดสอบประสิทธิภาพในแต่ละอัลกอริทึม

	Accuracy	Precision	Recall
Fasttext + logistic regression	0.915247	0.931825	0.907037
Fasttext + SVM	0.939288	0.956112	0.928490
Fasttext + RandomForest	0.901704	0.909954	0.904748

รูปภาพที่ 3.6 ประสิทธิภาพในแต่ละอัลกอริทึม

โดย Baseline ที่เลือกมาคือ Fasttext + Random Forest ซึ่งมีค่า Accuracy 0.90 หรือ 90%

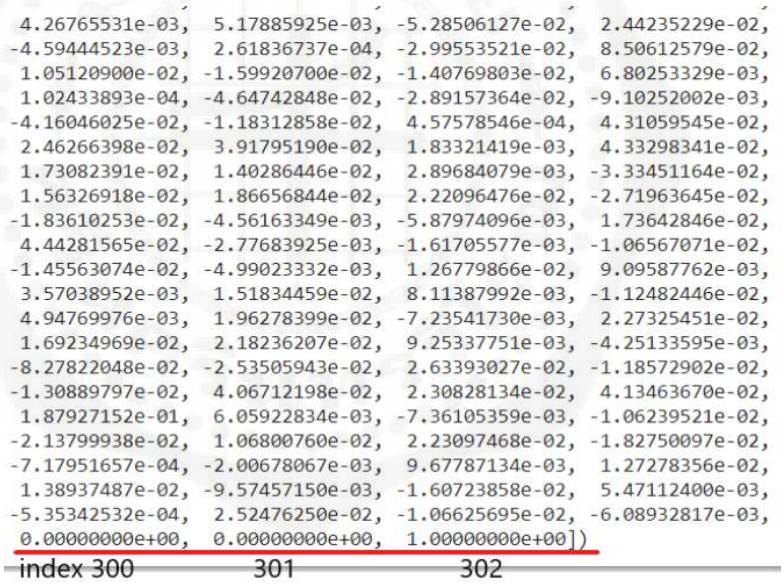
3.5.3 เพิ่มเทคนิคอื่นเพื่อเพิ่มประสิทธิภาพ

3.5.3.1 เพิ่มฟีเจอร์

- Positive/Negative Word

Positive/Negative Word คือ นำข้อมูลมาวิเคราะห์อารมณ์ (Sentiment Analysis) โดยใช้ PyThaiNLP ซึ่งมีข้อมูล Positive word, Negative word และ Swear word

- นับจำนวน Negative word , Positive word
- คำหรือข้อความที่เป็นลักษณะภาษาของภาวะซึมเศร้า ที่ถูกจำแนกด้วยความถี่ในการใช้งาน
    - ไม่พบซึมเศร้า เช่น เบื่อ หดหู่ หลับไม่ลง
    - พบซึมเศร้าน้อยถึงปานกลาง เช่น ไม่เห็นอนาคต ไม่มีสมาธิ หงุดหงิดง่าย คิดเรื่องเดิม
    - พบซึมเศร้ารุนแรง เช่น ทำร้ายตัวเอง เขียนจดหมายลาตาย ไม่รู้สึกตัว



รูปภาพที่ 3.7 ตัวอย่างข้อมูลที่ถูกแปลงให้อยู่ในรูปเวกเตอร์หลังเพิ่มฟีเจอร์

จากรูปภาพที่ 3.7 เมื่อเพิ่มฟีเจอร์คำที่เป็นลักษณะทางภาษาของภาวะซึมเศร้า คำเชิงบวก และคำเชิงลบ จะอยู่ที่ index 300, 301 และ 302 ตามลำดับ

3.5.4 สร้าง Chat (Demo)

3.5.4.1 ลดขนาดของข้อมูล

- นำข้อมูลมาลดขนาดเพื่อความสะดวกในการรันในแต่ละการทดลองโดยลดข้อมูลจาก 300-dimensions เป็น100-dimensions เนื่องจาก 300 - dimensions ใช้เวลาในการรันค่อนข้างนาน

3.5.4.2 สร้าง Application สำหรับเชื่อมต่อ

- นำกระบวนการต่าง ๆ มาเป็น web application โดยการใช้ ngrok เปลี่ยน local server ให้อยู่ในรูปแบบ public server เช่น การทำความสะอาดข้อมูลและการเปลี่ยนข้อความให้อยู่ในรูปเวคเตอร์เพื่อรับและส่งข้อมูลไปยัง Facebook โดยใช้ Facebook API (messenger)

3.5.4.3 ประมวลผลข้อความ

- นำเวคเตอร์ที่ได้มาทำนายโดยจะมีการใช้โมเดลที่ทำการบันทึกไว้จากการเรียนรู้ก่อนหน้า ในที่นี้จะเลือกใช้โมเดลที่ได้ค่าความถูกต้องที่ได้ผลลัพธ์ดีที่สุดคือ SVM จากนั้นทำการตอบกลับโดยดูจากค่าที่ได้จากการทำนาย ถ้าทำนายว่าข้อความมีภาวะโรคซึมเศร้าจะทำการตอบกลับข้อความเชิงให้กำลังใจ ยกตัวอย่างเช่น อีกไม่นานก็จะดีขึ้นและเธอจะผ่านมันไปได้ และฉันอาจจะไม่เข้าใจเธอ แต่ฉันจะอยู่ข้างๆ เธอนะ ถ้าทำนายว่าข้อความไม่มีภาวะซึมเศร้าจะทำการตอบกลับข้อความว่า ขอให้เป็นวันที่ดี

บทที่ 4
ผลการดำเนินโครงงาน

4.1 การทำ Retrain Word2Vec ในชุดข้อมูลภาษาอังกฤษ

ผลการทดสอบประสิทธิภาพโดยวัดจากค่า Accuracy ก่อนการทำ Retrain ในแต่ละอัลกอริทึม

	Accuracy	Precision	Recall
Word2Vec + logistic regression	0.751887	0.698779	0.655216
Word2Vec + SVM	0.749371	0.734528	0.573791
Word2Vec + RandomForest	0.854051	0.866864	0.745547

รูปภาพที่ 4.1 ประสิทธิภาพการทดสอบ ก่อนการทำ Retrain

- Logistic regression ได้ Accuracy 0.75
- Support Vector Machine ได้ Accuracy 0.75
- Random Forest ได้ Accuracy 0.85

ผลการทดสอบประสิทธิภาพโดยวัดจากค่า Accuracy หลังการทำ Retrain ในแต่ละอัลกอริทึม

- Logistic regression ได้ Accuracy 0.77
- Support Vector Machine ได้ Accuracy 0.77
- Random Forest ได้ Accuracy 0.85

	Accuracy	Precision	Recall
W2v + logistic regression	0.770508	0.718833	0.689567
W2v + SVM	0.774031	0.757252	0.631043
W2v + RandomForest	0.851032	0.864583	0.739186

รูปภาพที่ 4.2 ประสิทธิภาพการทดสอบ หลังการทำ Retrain

4.2 การสำรวจข้อมูลภาษาไทย

การสำรวจข้อมูลเพื่อตอบสมมติฐานว่าเป็นข้อความที่มีแนวโน้มต่อโลกซึมเศร้า หรือไม่มีแนวโน้มต่อโลกซึมเศร้า

- ตัดคำ

พอจะตายขึ้นมาจริงๆ เป็นห่วงทำไม
['พอ', 'จะ', 'ตาย', 'ขึ้น', 'มา', 'จริงๆ', 'เป็นห่วง', 'ทำไม']

รูปภาพที่ 4.3 ตัวอย่างการตัดคำ 1

วันไหนสอบวันนั้นยิ่งดิ่ง
['วัน', 'ไหน', 'สอบ', 'วันนั้น', 'ยิ่ง', 'ดิ่ง']

รูปภาพที่ 4.4 ตัวอย่างการตัดคำ 2

- WordCloud

Mono - gram : ไม่มีแนวโน้มต่อภาวะโรคซึมเศร้า

รูปภาพที่ 4.5 Word Cloud Mono - gram 1

Mono - gram : มีแนวโน้มต่อภาวะโรคซึมเศร้า

รูปภาพที่ 4.6 Word Cloud Mono - gram 2

Bi - gram : ไม่มีแนวโน้มต่อภาวะโรคซึมเศร้า

รูปภาพที่ 4.7 Word Cloud Bi- gram 1

Bi - gram : มีแนวโน้มต่อภาวะโรคซึมเศร้า

รูปภาพที่ 4.8 Word Cloud Bi- gram 2

4.3 ผลการดำเนินงานจากการทดสอบประสิทธิภาพ

ผลการดำเนินงานจากการทดสอบประสิทธิภาพด้วยการวัดค่า Accuracy ในภาษาไทยโดยใช้ Fasttext เป็นวิธีการเปลี่ยนประโยคเป็นเวคเตอร์

- Logistic regression ได้ Accuracy 0.92
- Support Vector Machine ได้ Accuracy 0.93
- Random Forest ได้ Accuracy 0.90

	Accuracy	Precision	Recall
Fasttext + logistic regression	0.915247	0.931825	0.907037
Fasttext + SVM	0.939288	0.956112	0.928490
Fasttext + RandomForest	0.901704	0.909954	0.904748

รูปภาพที่ 4.9 ประสิทธิภาพการทดสอบ FastText ในแต่ละโมเดล (ภาษาไทย)

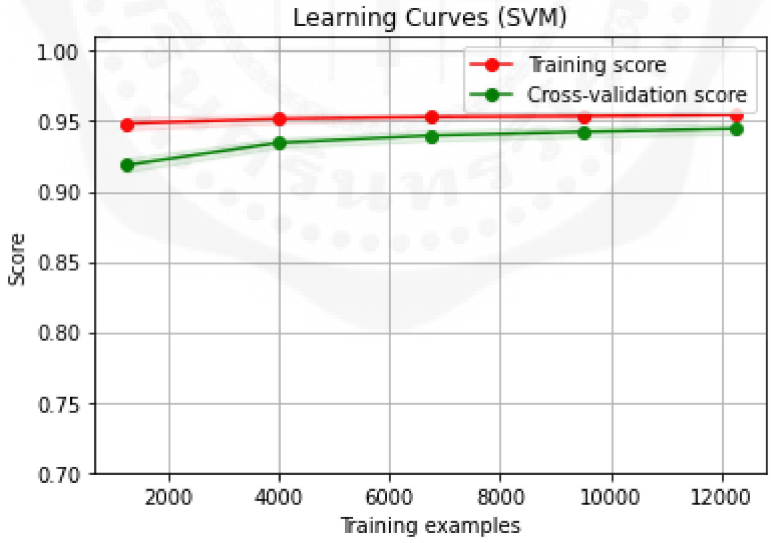
4.3.1 ทำการตรวจสอบความถูกต้องของการเรียนรู้ เนื่องจากการวัดประสิทธิภาพมีค่าค่อนข้างสูง

- ค่าความน่าจะเป็นของการทำนาย

```
prob : [0.97345912 0.02654088]    predict : -1
prob : [0.44403595 0.55596405]    predict : 1
prob : [0.15580054 0.84419946]    predict : 1
prob : [0.02888738 0.97111262]    predict : 1
prob : [0.74981939 0.25018061]    predict : -1
prob : [0.87439817 0.12560183]    predict : -1
prob : [0.80811731 0.19188269]    predict : -1
prob : [0.36430724 0.63569276]    predict : 1
prob : [0.20906828 0.79093172]    predict : 1
prob : [0.68964307 0.31035693]    predict : -1
prob : [0.8834799 0.1165201]    predict : -1
prob : [0.94405353 0.05594647]    predict : -1
prob : [0.00116283 0.99883717]    predict : 1
prob : [0.29221572 0.70778428]    predict : 1
prob : [0.95470421 0.04529579]    predict : -1
prob : [0.53428364 0.46571636]    predict : -1
prob : [0.28052025 0.71947975]    predict : 1
prob : [0.14646879 0.85353121]    predict : 1
prob : [0.40741249 0.59258751]    predict : 1
prob : [0.00380423 0.99619577]    predict : 1
prob : [0.56314403 0.43685597]    predict : -1
```

รูปภาพที่ 4.10 ความน่าจะเป็นของการทำนาย

- Learning Curve



รูปภาพที่ 4.11 Learning Curve

4.4 ผลการดำเนินงานจากการทดสอบประสิทธิภาพหลังจากเพิ่มฟีเจอร์

ผลการดำเนินงานจากการทดสอบประสิทธิภาพด้วยการวัดค่า Accuracy ในภาษาไทยหลังจากทำการเพิ่มฟีเจอร์ เช่น คำเชิงบวก/เชิงลบ คำหรือข้อความที่เป็นลักษณะทางภาษาของภาวะซึมเศร้า เป็นต้น และใช้ Fasttext เป็นวิธีการเปลี่ยนประโยคเป็นเวคเตอร์

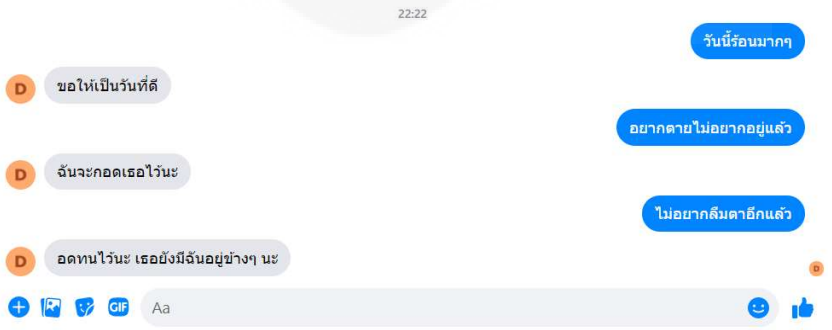
- Logistic regression ได้ Accuracy 0.92
- Support Vector Machine ได้ Accuracy 0.93
- Random Forest ได้ Accuracy 0.90

	Accuracy	Precision	Recall
Fasttext + logistic regression	0.921181	0.935531	0.914006
Fasttext + SVM	0.934267	0.962064	0.911705
Fasttext + RandomForest	0.907943	0.916957	0.908254

รูปภาพที่ 4.12 ประสิทธิภาพการทดสอบ FastText หลังการเพิ่มฟีเจอร์ในแต่ละโมเดล (ภาษาไทย)

4.5 การสร้างแอพพลิเคชั่นในรูปแบบแชท

ผลลัพธ์การสร้างแอพพลิเคชั่นในรูปแบบแชทเพื่อนำเสนอตัวอย่างการทำนายข้อความที่ได้รับจากผู้ใช้ โดยการนำโมเดลที่ได้ทำการเรียนรู้มาใช้ในการทำนาย เมื่อได้รับข้อความ ถ้าผลการทำนายเป็นภาวะซึมเศร้าระบบจะตอบกลับข้อความในเชิงการให้กำลังใจ หากผลการทำนายเป็นปกติหรือไม่มีภาวะซึมเศร้าจะตอบกลับว่าขอให้เป็นวันที่ดี



รูปภาพที่ 4.13 ตัวอย่างข้อความที่ทำนายและส่งข้อความตอบกลับหลังจากการทำนาย

บทที่ 5
สรุปผล อภิปรายผล และข้อเสนอแนะ

5.1 สรุปผลการดำเนินงาน

สรุปผลการดำเนินงานในชุดข้อมูลภาษาอังกฤษต่อเนื่องจากผลการดำเนินงานของภาคการศึกษา 1/2563 ได้มีการใช้ Pre-train FastText เพื่อทำการหาคำที่ตรงกันและแปลงเป็นเวกเตอร์ ได้มีการทดลองทำ Re-train จากชุดข้อมูลที่ใช้เพื่อเปรียบเทียบประสิทธิภาพของการทำนายซึ่งผลลัพธ์ของการทำ Re-train ได้ประสิทธิภาพจากการวัดด้วยค่า Accuracy พบว่า Logistic Regression และ Support Vector Machine มีค่าเพิ่มขึ้นเป็น 77% จากก่อนการทดลอง 75% ส่วน Random Forest มีค่าลดลงเพียงเล็กน้อยเป็น 85.1% จากก่อนการทดลอง 85.4%

ผลการดำเนินงานในชุดข้อมูลภาษาไทย โดยชุดข้อมูลภาษาไทยได้มีการเก็บรวบรวมผ่านทวิตเตอร์ แบ่งเป็น 2 ประเภท คือ ข้อมูลที่มีแนวโน้มต่อภาวะซึมเศร้าและข้อมูลที่ไม่มีแนวโน้มต่อภาวะซึมเศร้า ข้อมูลทั้งหมดมีจำนวน 21,905 แถว มีการทำ Pre-process และแปลงเป็นเวกเตอร์โดยใช้ pre-train Fasttext ได้ Baseline 90% เมื่อเพิ่มฟีเจอร์ต่างๆ ได้ค่าประสิทธิภาพที่ดีที่สุดวัดจากค่า Acccuracy คือ 93%

5.2 ปัญหาและอุปสรรค

- ในส่วนของภาษาไทยเครื่องมือในการทำ Sentiment Analysis มีอยู่น้อย
- ชุดข้อมูลที่รวบรวมยังไม่ถูกต้องมากพอ
- มีปัจจัยหลายอย่างที่มีผลต่อการวิเคราะห์ว่าผู้ใช้คนนั้นเป็นโรคซึมเศร้าหรือไม่
- ปัจจัยบางอย่างมีความขัดแย้งกัน

5.3 ข้อเสนอแนะ

ส่วนของข้อมูลอาจมีความไม่ชัดเจนในเรื่องของการแยกกลุ่มที่ถูกต้องและชัดเจน ควรเพิ่มเติมข้อมูลหรือควรมีข้อมูลที่มีความสมบูรณ์มากกว่า

ส่วนของปัจจัยควรได้รับการศึกษาเพิ่มเติม เนื่องจากบางปัจจัยเกินมีความขัดแย้งกัน เช่น ภาวะโรคซึมเศร้าอาจนิยมใช้อีโมจิในการโพสต์ข้อความหรือไม่นิยมใช้เป็นต้น และคำที่มีผลต่อภาวะซึมเศร้าหรือคำที่คนที่มีภาวะซึมเศร้านิยมใช้น่าจะมีมากขึ้น

บรรณานุกรม

[1] “Sentiment-Analysis-Python @ Python3.Wannaphong.Com.” [Online]. Available: https://python3.wannaphong.com/2017/02/sentiment-analysis-python.html.

[2] “lstm-เท่าที่เข้าใจ-75027db3167f @ medium.com.” [Online]. Available: https://medium.com/@sanparithmarukatat/lstm-เท่าที่เข้าใจ-75027db3167f.

[3] R. N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, “The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches,” ICWSM 2008 - Proc. 2nd Int. Conf. Weblogs Soc. Media, pp. 102–108, 2008, [Online]. Available: https://www.aaai.org/Papers/ICWSM/2008/ICWSM08-020.pdf.

[4] “Identifying-depression @ github.com.” [Online]. Available: https://github.com/Inusette/Identifying-depression.

[5] “Tf-Idf-คำไหนสำคัญนะ @ Lukkiddd.Com.” [Online]. Available: https://lukkiddd.com/tf-idf-คำไหนสำคัญนะ-dd1e1568312e.

[6] “a-step-by-step-guide-on-sentiment-analysis-with-rnn-and-lstm @ medium.com.” [Online]. Available: https://medium.com/@lamiae.hana/a-step-by-step-guide-on-sentiment-analysis-with-rnn-and-lstm-3a293817e314.

[7] “logistic-regression-ไม่มีอะไรเป็นไปตามอย่างที่คิดเสมอ -machine-learning-101-bba2f666234d @ medium.com.” [Online]. Available: https://medium.com/mmp-li/logistic-regression-ไม่มีอะไรเป็นไปตามอย่างที่คิดเสมอ-machine-learning-101-bba2f666234d.

[8] “Word2Vec-ทำอย่างไร-B3De9D9a38B3 @ Lukkiddd.Com.” [Online]. Available: https://lukkiddd.com/word2vec-ทำอย่างไร-b3de9d9a38b3.

[9] “logistic-regression-ไม่มีอะไรเป็นไปตามอย่างที่คิดเสมอ -machine-learning-101-bba2f666234d @ medium.com.” [Online]. Available: https://medium.com/mmp-li/logistic-regression-ไม่มีอะไรเป็นไปตามอย่างที่คิดเสมอ-machine-learning-101-bba2f666234d.

[10] “เอกสารประกอบการสอน วิชา ระบบสนับสนุนการตัดสินใจ สาขาเทคโนโลยีสารสนเทศ วิทยาลัยเซาธ์อีสท์บางกอก,” pp. 1–7, [Online]. Available: http://www.theerapone.com/sbc/courses/dss/doc/Bayes_NaivBayes.pdf.
[11] “อัลกอริทึม Support Vector Machine (SVM) @ kokzard.blogspot.com.” [Online]. Available: http://kokzard.blogspot.com/2011/10/jfjkdshfkjsldf.html.
[12] “GloVe and fastText — Two Popular Word Vector Models in NLP @ cai.tools.sap.” [Online]. Available: https://cai.tools.sap/blog/glove-and-fasttext-two-popular-word-vector-models-in-nlp/#:~:text=fastText is another word embedding,an n-gram of characters.

[13] “เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c @ medium.com.” [Online]. Available: https://medium.com/@witchapongdaroontham/เจาะลึก-random-forest-part-2-of-รู้จัก-decision-tree-random-forest-และ-xgboost-79b9f41a1c1c.

[14] “random-forest-คืออะไร-74d2a0af3d7 @ medium.com.” [Online]. Available: https://medium.com/@pradyasin/random-forest-คืออะไร-74d2a0af3d7.

[15] “13 Common Words and Phrases That May Signal Depression @ www.thehealthy.com.” [Online]. Available: https://www.thehealthy.com/mental-health/depression/words-phrases-sign-depression/.

[16] “vaderSentiment @ github.com.” [Online]. Available: https://github.com/cjhutto/vaderSentiment.

[17] จันทิมา อังคพณิชกิจ. (2563). ภาษาการสื่อสารและโรคซึมเศร้า การสำรวจเบื้องต้นเพื่อเข้าใจโรคซึมเศร้าและผู้มีภาวะซึมเศร้าในสังคมไทย.กรุงเทพฯ:อมรินทร์พริ้นติ้ง.

[18] “thai-sentiment-analysis-toolkit @ www.kaggle.com.” [Online]. Available: https://www.kaggle.com/rtatman/thai-sentiment-analysis-toolkit

[19] “product @ ngrok.com.” [Online]. Available: https://ngrok.com/product.

[20] “Early Detection of Depression: Social Network Analysis and Random Forest Techniques @ www.ncbi.nlm.nih.gov.” [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6598420/.