โครงการ
การเรียนการสอนเพื่อเสริมประสบการณ์

ชื่อโครงการ	ระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์ The sentiment analysis system for political messages on Twitter
ชื่อนิสิต	นาย จิตรทิวัส แจ้งจันทร์ 5933611123 นาย ชลวิทย์ ก้อนทอง 5933617023
ภาควิชา	คณิตศาสตร์และวิทยาการคอมพิวเตอร์ สาขาวิชา วิทยาการคอมพิวเตอร์
ปีการศึกษา	2562

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์

นายจิตรทิวัส แจ้งจันทร์
นายชลวิทย์ ก้อนทอง

โครงงานนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

The sentiment analysis system for political messages on Twitter.

Mr.Jittiwat Jangjun
Mr.Chonlawit Gonthong

A Project Submitted in Partial Fulfillment of the Requirements
for the Degree of Bachelor of Science Program in Computer Science
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2019

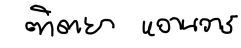
หัวข้อโครงงาน	ระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์
โดย	นายจิตรทิวัส แจ้งจันทร์
	นายชลวิทย์ ก้อนทอง
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาโครงงานหลัก	ผู้ช่วยศาสตราจารย์ ดร.ภควรรณ ปักษี

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

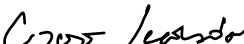
..... หัวหน้าภาควิชาคณิตศาสตร์
(ศาสตราจารย์ ดร.กฤษณะ เนียมมณี) และวิทยาการคอมพิวเตอร์

คณะกรรมการสอบโครงงาน

..... อาจารย์ที่ปรึกษาโครงงานหลัก
(ผู้ช่วยศาสตราจารย์ ดร.ภควรรณ ปักษี)



..... กรรมการ
(ผู้ช่วยศาสตราจารย์.ดร.ฑิตยา หวานวารี)



..... กรรมการ
(ผู้ช่วยศาสตราจารย์.ดร.อาธร เหลืองสดใส)

จิตรทิวัส แจ้งจันทร์,ชลวิทย์ ก้อนทอง :ระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์ (The sentiment analysis system for political messages on Twitter) อ.ที่ปรึกษาโครงงานหลัก: ผู้ช่วยศาสตราจารย์ ดร.ภควรรณ ปักษี, 69 หน้า.

ระบบวิเคราะห์ความรู้สึกของข้อความถูกพัฒนาขึ้นเนื่องจากในปัจจุบันประชากรส่วนใหญ่ใช้สื่อออนไลน์เป็นสถานที่ในการแลกเปลี่ยนความคิดเห็นระหว่างกัน ทำให้การอ่านความคิดเห็นซึ่งมีจำนวนข้อความจำนวนมากเป็นเรื่องที่ยาก โดยเฉพาะความคิดเห็นทางด้านการเมืองที่มีประเด็นหลากหลาย ผู้พัฒนาจึงพัฒนาโครงการนี้ขึ้นเป็นระบบเว็บแอปพลิเคชันที่จะช่วยให้ผู้ใช้งานสามารถเข้ามาดูสรุปผลการวิเคราะห์ข้อความการแสดงความคิดเห็นด้านการเมืองได้ง่ายและรวดเร็วขึ้น วัตถุประสงค์ของโครงการเพื่อวิเคราะห์ความรู้สึกของข้อความในการแสดงความคิดเห็นด้านการเมืองจากทวิตเตอร์ การวิเคราะห์ความคิดเห็นจะแบ่งออกเป็น ความคิดเห็นด้านบวก ความคิดเห็นด้านลบ และความคิดเห็นที่เป็นกลาง โดยใช้คำสำคัญเป็นตัวค้นหารวบรวมความคิดเห็น ซึ่งโครงการนี้ผู้พัฒนาจะเก็บรวบรวมข้อความแสดงความคิดเห็นด้านการเมืองที่เป็นภาษาไทย โดยอาศัยเทคนิคการประมวลผลภาษาธรรมชาติ การเรียนรู้ของเครื่องด้วยการถดถอยโลจิสติกส์ และการเรียนรู้เชิงลึกด้วยโครงข่ายประสาทแอลเอสทีเอ็ม เพื่อวิเคราะห์ข้อความแสดงความคิดเห็นเหล่านั้น ผลการทดสอบระบบการวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองพบว่า ความถูกต้องและความแม่นยำอยู่ที่ประมาณ 70% ซึ่งเพียงพอในการใช้งานได้แต่ยังคงต้องมีการปรับปรุงพัฒนาต่อไป

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต จิตรทิวัส แจ้งจันทร์

ลายมือชื่อนิสิต ชลวิทย์ ก้อนทอง

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาโครงงานหลัก ภควรรณ ปักษี

ปีการศึกษา 2562

5933611123, 5933602523: MAJOR COMPUTER SCIENCE
KEYWORDS: SENTIMENT ANALYSIS / POLITICAL / ULMFIT
JITTIWAT JANGJUN, CHONLAWIT GONTHONG: THE SENTIMENT ANALYSIS SYSTEM FOR POLITICAL MESSAGES ON TWITTER. ADVISOR: ASSIST. PROF. PAKAWAN PUGSEE, Ph.D., 69 pp.

The sentiment analysis system was developed because today people uses online media as a place to exchange their ideas, which make difficult to read a lot of comments, especially the political opinions. Therefore, the project was implemented to be a web application system that will allow users to view the summary of the analysis of the political opinions quickly and easily. The objective of this project is to analyze the sentiment of political comments from Twitter. The analyzed opinions are divided into positive, negative, and neutral comments. The keywords are used to collect the Thai statements about political comments. The natural language processing (NLP), machine learning with logistical regression, and deep learning with LSTM neural network are used to analyze those comment messages. The testing result of the sentiment analysis system for political messages found that the accuracy and the precision are around 70% which is enough to be used but still needs to be improved.

Department: Mathematics and Computer Science Student's Signature Jittiwat Jangjun
Student's Signature Chonlawit Gonthong
Field of Study: Computer Science Advisor's Signature Pakawan Pugsee
Academic Year: 2019

กิตติกรรมประกาศ

การจัดทำโครงงานระบบวิเคราะห์ความคิดเห็นต่อละครไทยบนทวิตเตอร์ สามารถลุล่วงไปได้ด้วยดีเนื่องจากได้รับความอนุเคราะห์ เนื่องจากได้รับความอนุเคราะห์และช่วยเหลือจากคณาจารย์และบุคลาการต่าง ๆ ดังนี้

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.ภควรรณ ปักษี อาจารย์ที่ปรึกษาโครงการ ที่คอยให้คำปรึกษา ข้อเสนอแนะทางวิชาการ อีกทั้งยังช่วยแนะนำ และชี้แนะตลอดการดำเนินการโครงงาน

ขอขอบพระคุณคณะกรรมการสอบ ได้แก่ ผู้ช่วยศาสตราจารย์.ดร.ฑิตยา หวานวารี และผู้ช่วยศาสตราจารย์.ดร.อาธร เหลืองสดใส ที่ช่วยให้คำแนะนำและข้อเสนอแนะสำหรับการพัฒนาโครงงานนี้ให้มีความถูกต้องและสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณทีมงานทาง pythainlp ที่ช่วยให้คำแนะนำและข้อเสนอแนะสำหรับเทคนิคการวิเคราะห์ข้อมูลต่าง ๆ ที่ช่วยพัฒนาโครงงานนี้ให้มีความถูกต้องและสมบูรณ์ยิ่งขึ้น

ขอขอบพระคุณ คุณพ่อและคุณแม่ที่คอยช่วยสนับสนุน ให้กำลังใจตลอดการทำโครงงาน

ขอขอบคุณเพื่อน ๆ ภาควิชาคณิตศาสตร์ สาขาวิชาวิทยาการคอมพิวเตอร์ ที่คอยช่วยเหลือและให้คำปรึกษาเกี่ยวกับโครงงาน

ท้ายที่สุดนี้ ขอขอบพระคุณทุกความกรุณาจากทุกท่านที่กล่าวมา รวมถึงบุคคลที่ไม่ได้กล่าวถึงไว้ ณ ที่นี้อีกครั้งหนึ่ง สำหรับความช่วยเหลือและคำแนะนำต่าง ๆ ซึ่งทำให้โครงงานนี้ประสบสำเร็จลุล่วงไปด้วยดี

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและเหตุผลของโครงงาน

เนื่องจากในปัจจุบันมนุษย์ได้ใช้ประโยชน์สื่อสังคมออนไลน์เป็นจำนวนมาก มีการพัฒนาของอินเทอร์เน็ตเกิดขึ้นอย่างรวดเร็ว ทำให้การสื่อสารด้วยข้อความได้รับความนิยมมากขึ้น สื่อสังคมออนไลน์ที่ได้รับความนิยมในปัจจุบัน เช่น เฟซบุ๊ก ทวิตเตอร์ ไลน์ และอินสตาแกรม สื่อเหล่านี้ได้กลายเป็นสถานที่ในการแลกเปลี่ยนความคิดเห็นของผู้คน ผู้ใช้สื่อออนไลน์จำนวนมากได้ร่วมสร้างข้อมูลหลากหลายรูปแบบ เช่น ข้อความ รูปภาพ ฯลฯ ทำให้มีข้อมูลจำนวนมหาศาลบนโลกออนไลน์ โดยเฉพาะด้านการเมืองที่ได้รับความสนใจอยู่ในปัจจุบัน ซึ่งข้อมูลเหล่านี้สามารถเก็บรวบรวม คัดลอก หรือประมวลผลด้วยคอมพิวเตอร์ได้ เพื่อนำไปใช้ประโยชน์ในด้านการวิเคราะห์ความรู้สึก (sentiment analysis) การทำเหมืองข้อมูล (data mining) หรือการสกัดข้อมูล (data extraction) ได้
ระบบวิเคราะห์ความรู้สึกของข้อความ คือระบบที่สามารถวิเคราะห์ และคัดกรองข้อความได้ว่า มีการแสดงความรู้สึกหรือไม่ และถ้ามีการแสดงความรู้สึก จะแบ่งกลุ่มข้อความออกเป็น ข้อความด้านบวก ข้อความด้านลบ หรือข้อความที่สื่อความรู้สึกแต่ไม่แสดงถึงด้านบวกหรือด้านลบอย่างชัดเจน ในงานวิจัยนี้ขอเรียกว่าข้อความที่เป็นกลาง หลังจากที่ได้ผลการวิเคราะห์ข้อความ ขั้นต่อมาจะนำไปแสดงผลให้เห็นว่า ข้อความด้านใดมีเป็นจำนวนเท่าไร ในรูปแบบที่ผู้รับสารสามารถทำความเข้าใจได้ง่าย จากการค้นคว้าพบว่ามีงานวิจัยอยู่หลากหลายที่เกี่ยวข้องกับการวิเคราะห์ความรู้สึก ผู้พัฒนาโครงการจึงได้ศึกษางานวิจัยที่ใช้เทคนิค การประมวลผลภาษาธรรมชาติ และการเรียนรู้ของเครื่อง (Machine Learning Techniques) ในการระบุความรู้สึกของข้อความ โดยมักจำแนกออกเป็นสองด้าน คือ ด้านบวก และด้านลบ ซึ่งได้งานวิจัยของ Kusrini และ Mochamad [1] ทำการวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์ โดยใช้ซัพพอร์ตเวกเตอร์แมชชีน และนาอีฟเบย์ในการทำนาย ซึ่งมีการใช้วิธีเทียบคลังคำศัพท์ (Lexicon Based Approach) [2] สำหรับการสกัดข้อมูล ผลที่ได้แสดงให้เห็นว่าการวิเคราะห์ความรู้สึกของข้อความโดยใช้ซัพพอร์ตเวกเตอร์แมชชีน มีความแม่นยำในการจำแนกอยู่ที่ 79% และวิธีนาอีฟเบย์อยู่ที่ 84% อีกงานวิจัยของ Sheeba Naz และคณะ [3] ซึ่งวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์ โดยใช้เทคนิค n-gram [4] สำหรับการสกัดข้อมูลและซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกความรู้สึก ผลที่ได้แสดงให้เห็นว่า ค่าความแม่นยำในการจำแนกความรู้สึกอยู่ที่ 81% ซึ่งค่อนข้างใกล้เคียงกับงานวิจัยแรก อย่างไรก็ตาม งานวิจัยเหล่านี้ล้วนแต่วิเคราะห์ข้อความที่เป็นภาษาอังกฤษ ซึ่งการวิเคราะห์ข้อความภาษาไทยนั้นมีขั้นตอนการประมวลผลที่ยากกว่า ตั้งแต่ขั้นตอนการตัดคำ (word segmentation) [5] เนื่องจากข้อความ

ภาษาอังกฤษมีการใช้เว้นวรรคในการบอกขอบเขตของคำ แต่ภาษาไทยเป็นภาษาที่ไม่มีเว้นวรรคระหว่างคำ ทำให้มีความกำกวม ดังตัวอย่างคลาสสิก เช่น ตากลม ซึ่งสามารถตัดคำได้ทั้ง 2 แบบคือ ตาก-ลม หรือ ตา-กลม ซึ่งต้องดูบริบทของข้อความด้วยว่าจะสื่อความหมายไปในทางใด สำหรับการวิเคราะห์ความรู้สึกของข้อความภาษาไทยมีแหล่งข้อมูลของ Wongnai-Corpus [6] ที่รวบรวมคำภาษาไทยที่สื่อถึงอารมณ์ด้านบวกและด้านลบ สำหรับการวิเคราะห์บทวิจารณ์และให้คะแนนร้านอาหารที่อยู่บนเว็บ Wongnai และ Wisesight Sentiment Corpus [7] ที่รวบรวมข้อความภาษาไทยจากสื่อสังคมออนไลน์พร้อมทั้งมีการกำกับความรู้สึกออกเป็นข้อความด้านบวก ด้านลบ เป็นกลาง และคำถาม

จากที่กล่าวมาข้างต้นทำให้เห็นแนวทางวิธีการวิเคราะห์ความรู้สึกของข้อความ ผู้พัฒนาโครงการจึงจะพัฒนาระบบวิเคราะห์ความรู้สึกของข้อความภาษาไทย เพื่อช่วยในการคัดกรองข้อความด้านบวก และข้อความด้านลบ ในการสรุปประเด็นความคิดเห็นด้านการเมือง เพื่อให้ผู้ใช้สามารถมองเห็นภาพรวมของประเด็นสถานการณ์ปัจจุบันทางการเมืองที่เกิดขึ้น

1.2 วัตถุประสงค์ของโครงงาน

1. เพื่อวิเคราะห์ความรู้สึกของข้อความที่เกี่ยวข้องกับการเมืองบนทวิตเตอร์
2. เพื่อสรุปความคิดเห็นที่มีต่อประเด็นทางการเมืองของประเทศไทย

1.3 ขอบเขตของโครงงาน

1. ระบบสามารถวิเคราะห์ข้อความที่ใช้ภาษาไทยเป็นภาษาหลัก
2. ข้อมูลที่นำมาวิเคราะห์ดึงข้อมูลจากทวิตเตอร์ จำนวนไม่ต่ำกว่า 10,000 ข้อความ
3. ระบบสามารถคัดกรองข้อความแสดงความรู้สึก และแบ่งออกเป็น ด้านบวก ด้านลบ และข้อความที่เป็นกลาง

1.4 วิธีการดำเนินงาน

1. ศึกษาค้นคว้าข้อมูล
2. กำหนดขอบเขตละวิธีการดำเนินงาน
3. เก็บรวบรวมข้อมูลที่เกี่ยวกับการเมืองจากสื่อออนไลน์ทวิตเตอร์
4. การจำแนกข้อความและระบุชนิดของคำ
5. พัฒนาระบบสำหรับการวิเคราะห์ความคิดเห็นทางด้านการเมือง
6. ตรวจสอบความถูกต้องของการจำแนกข้อมูล
7. จัดทำเอกสารและสรุปผล

ตารางเวลาการดำเนินงาน

การพัฒนาระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์ เริ่มดำเนินงานตั้งแต่เดือนกันยายน 2562 ถึง เดือนเมษายน 2563 รวมระยะเวลา 8 เดือน โดยมีตารางเวลาการดำเนินงาน ดังนี้

ขั้นตอนดำเนินงาน	2562				2563			
	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1. ศึกษาค้นคว้าข้อมูล	■	■						
2. กำหนดขอบเขตและวิธีดำเนินการ		■						
3. เก็บรวบรวมข้อมูลที่เกี่ยวกับการเมืองจากสื่อออนไลน์ทวิตเตอร์			■	■				
4.การจำแนกข้อความและระบุชนิดของคำ				■				
5.พัฒนาระบบสำหรับการวิเคราะห์ความคิดเห็นทางด้านการเมือง				■	■	■	■	
6.ตรวจสอบความถูกต้องของการจำแนกข้อมูล						■	■	
7.สรุปผลและจัดทำเอกสาร			■	■	■	■	■	■

1.5 ประโยชน์ที่ได้รับ

1. ประโยชน์ต่อผู้พัฒนา

- ได้พัฒนาโมเดลสำหรับการวิเคราะห์ข้อความและอยากพัฒนาโมเดลให้มีประสิทธิภาพที่ดีขึ้น
- ได้พัฒนาทักษะการสร้างระบบตามขั้นตอนวิธีด้านวิศวกรรมซอฟต์แวร์ ซึ่งสามารถนำไปใช้พัฒนาระบบอื่นในอนาคตได้

2. ประโยชน์ต่อผู้ใช้และสังคม

- ข้อมูลที่ได้จากการวิเคราะห์ช่วยให้เห็นแนวโน้มว่า ในปัจจุบันผู้คนในสื่อสังคมออนไลน์มีความคิดเห็นทางการเมืองในปัจจุบันเป็นด้านบวก หรือด้านลบ เกี่ยวข้องกับประเด็นใดบ้าง
- รัฐบาล หน่วยงานที่เกี่ยวข้อง หรือองค์กรที่เกี่ยวข้องกับการเมือง จะสามารถรับรู้ความคิดเห็นจากสื่อสังคมออนไลน์ เป็นไปในด้านบวก หรือด้านลบอย่างรวดเร็ว

1.6 โครงสร้างของรายงาน

บทที่ 2 จะกล่าวถึงบทความและทฤษฎีที่เกี่ยวข้องกับโครงงาน

บทที่ 3 จะกล่าวถึงการรวบรวมข้อมูล และการวิเคราะห์ข้อมูล

บทที่ 4 จะกล่าวถึงการออกแบบและพัฒนาโมเดล

บทที่ 5 จะกล่าวถึงการออกแบบและพัฒนาเว็บแอปพลิเคชัน

บทที่ 6 จะกล่าวถึงผลการทดสอบระบบการวิเคราะห์ข้อความแสดงความคิดเห็นด้านการเมือง ซึ่งจะทดสอบการจำแนกข้อความแสดงความรู้สึกที่เป็นบวก เป็นลบ และเป็นกลาง

บทที่7 จะกล่าวถึงข้อสรุป และข้อเสนอแนะทั้งหมดของโครงงานนี้

บทที่ 2
งานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงวิธีการวิเคราะห์อารมณ์และความรู้สึกของข้อความ เครื่องมือและไลบรารีต่าง ๆ ที่เกี่ยวข้องในการวิเคราะห์ความรู้สึกของข้อความ ดังรายละเอียดต่อไปนี้

2.1 การวิเคราะห์อารมณ์และความรู้สึก (Sentiment Analysis)

การวิเคราะห์ความรู้สึกของข้อความเป็นกระบวนการที่ใช้การทำงานของการประมวลผลภาษาธรรมชาติ (Natural Language Processing) การวิเคราะห์ข้อความ (text analysis) และการทำงานด้านภาษาศาสตร์เชิงคำนวณ (computational linguistics) โดยเป้าหมายของการวิเคราะห์แนวคิดทางด้านอารมณ์และความรู้สึก คือการแยกแยะความคิดเห็นหรือทัศนคติของบทความหัวข้อต่าง ๆ ในภาษาธรรมชาติของมนุษย์ ซึ่งทัศนคติสามารถตัดสินหรือประเมินผลได้ในด้านอารมณ์ของการสื่อสาร โดยโครงงานนี้สนใจวิธีการจำแนกข้อมูล 2 วิธี คือ การเรียนรู้ของเครื่องด้วยการถดถอยโลจิสติกส์ (Logistic regression) และการเรียนรู้เชิงลึก (deep learning) ด้วยโครงข่ายประสาทเทียมแอลเอสทีเอ็ม (LSTM neural network)

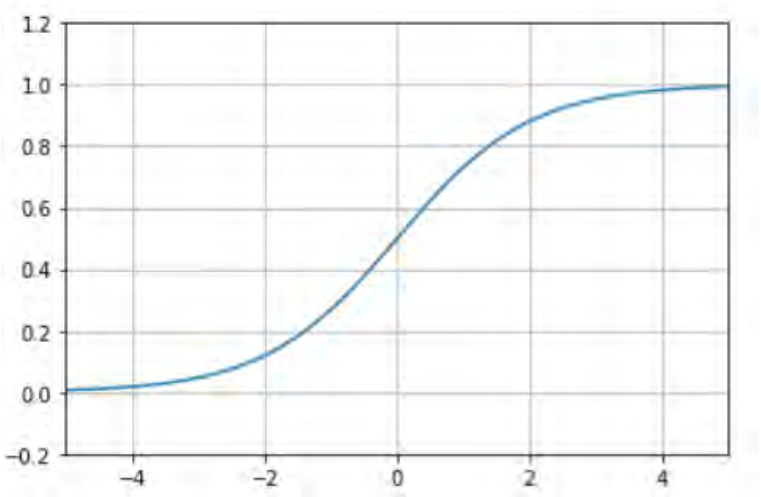
2.2 การเรียนรู้ของเครื่องด้วยการถดถอยโลจิสติกส์

การจำแนกข้อความด้วยการเรียนรู้ของเครื่องโดยใช้การถดถอยโลจิสติกส์ เป็นการวิเคราะห์ที่มีเป้าหมายเพื่อทำนายโอกาสที่จะเกิดเหตุการณ์ที่สนใจ โดยอาศัยสมการโลจิสติกส์ที่เป็นฟังก์ชันซิกมอยด์ (sigmoid) สำหรับการแยกคลาสออกเป็นสองคลาสดังสมการด้านล่าง

$$sigmoid(z) = \frac{e^z}{1 + e^z}$$

Sigmoid (z) คืออัตราส่วน ความน่าจะเป็นของการเกิดเหตุการณ์ ที่ศึกษา (e) กับความน่าจะเป็นที่จะไม่เกิดเหตุการณ์ที่ศึกษา (1-e)

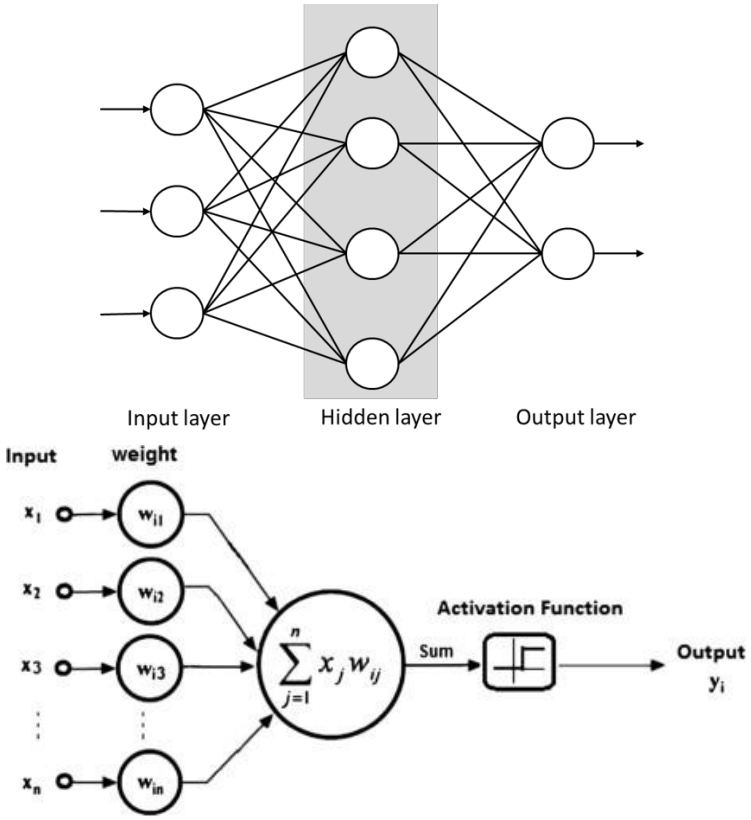
จะเห็นได้ว่าฟังก์ชันซิกมอยด์คล้ายสมการเชิงเส้นแต่ถูกพัฒนาโดยใช้ สมการเอกซ์โพเนนเชียล (e) เพื่อเปลี่ยนสมการเส้นตรงให้เป็นสมการที่ไม่ใช่เส้นตรง ซึ่งจะได้รูปคล้ายตัวเอส ดังภาพที่ 2.1 สำหรับงานที่จำแนกคลาสมากกว่าสองคลาสสามารถทำได้โดยใช้ฟังก์ชันซิกมอยด์จำนวน X ฟังก์ชัน โดยที่ X คือจำนวนคลาสที่ต้องการบอกว่าเป็นคลาสนั้นหรือไม่



ภาพที่ 2.1 สมการการถดถอยโลจิสติกส์

2.3 การเรียนรู้เชิงลึกด้วยโครงข่ายประสาทเทียมแอลเอสทีเอ็ม

การเรียนรู้เชิงลึกเป็นการทำงานของโครงข่ายประสาทเทียม (Artificial Neural Network) ซึ่งเป็นอัลกอริทึมการเรียนรู้ของเครื่องที่ได้แนวคิดมาจากระบบโครงข่ายประสาท (Neural Network) ในสมองของมนุษย์ ทำงานโดยอาศัยการซ้อนกันของโครงข่ายประสาทจำนวนหลาย ๆ ชั้น เพื่อสกัดคุณลักษณะบางอย่างออกมาจากข้อมูล โครงข่ายประสาทเทียมจะประกอบไปด้วย 3 ส่วน คือ ชั้นส่วนข้อมูลเข้า (input layer) ทำหน้าที่รับข้อมูลที่เข้าสู่โครงข่ายประสาทเทียม ชั้นส่วนซ่อน (hidden layer) ซึ่งมีแต่ละโหนด (node) ทำหน้าที่ช่วยในการคำนวณค่าและกำหนดน้ำหนักให้กับแต่ละเส้นที่ส่งสัญญาณข้อมูลโยงถึงกัน และส่วนท้ายคือ ชั้นส่วนข้อมูลออก (output layer) ทำหน้าที่แสดงผลข้อมูลคำตอบ โดยที่แต่ละโหนดในโครงข่ายประสาทเทียมจะทำหน้าที่คือ คำนวณผลบวกของผลคูณระหว่างข้อมูลที่ออกจากชั้นก่อนหน้ากับค่าน้ำหนักของเส้นแล้วนำไปผ่านฟังก์ชันกระตุ้น (activation function) และส่งเป็นข้อมูลออก ดังภาพที่ 2.2



ภาพที่ 2.2 ภาพตัวอย่างโครงข่ายประสาทเทียม(บน) และการทำงานในแต่ละโหนด(ล่าง)

ที่มา:(https://medium.com/@tongkornkitt/ml-lstms [บน],
https://becominghuman.ai/artificial-neuron-networks-basics-introduction-to-neural-networks-3082f1dcca8c [ล่าง])

2.3.1 แอลเอสทีเอ็ม (Long Short-Term Memory: LSTM)

แอลเอสทีเอ็ม (LSTM) คือ โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network) ใช้สำหรับการแก้ไขปัญหาของข้อมูลที่เป็นลักษณะลำดับที่ยาวมาก ๆ โดยมีหน้าที่จดจำข้อมูลที่เข้ามาก่อนหน้า มาวิเคราะห์กับข้อมูลที่รับเข้ามาต่อ การทำงานของแอลเอสทีเอ็ม สามารถบอกได้ว่า การทำงานที่เลือกว่าข้อมูลใดจะถูกเก็บเข้ามาในหน่วยความจำ ข้อมูลใดจะถูกลบออกจากหน่วยความจำ ข้อมูลใดจะถูกส่งออกภายนอก และการปรับปรุงข้อมูลที่ถูกเก็บในหน่วยความจำ ซึ่งประกอบด้วย 4 ส่วนหลักคือ ส่วนรับเข้า (Input gate : $\boldsymbol{i_t}$) ส่วนปรับค่ารับเข้า (Input modulation gate : $\boldsymbol{g_t}$) ส่วนคัดออก (Forget gate : $\boldsymbol{f_t}$) ข้อมูลค่าที่ถูกปรับปรุงและเก็บในหน่วยความจำ ($\boldsymbol{c_t}$) และส่วนส่งออก (Output gate : $\boldsymbol{o_t}$) ตัวอย่างของแอลเอสทีเอ็มแสดงดังภาพที่ 2.3

การทำงานของแอลเอสทีเอ็ม

1. การทำงานที่เลือกว่าข้อมูลใดจะถูกเก็บเข้ามาจากค่าส่วนรับเข้าและค่าส่วนปรับค่ารับเข้า โดยข้อมูลนำเข้า (x_t) จะมีค่าเป็น 0 หรือ 1 เท่านั้น

ค่าส่วนรับเข้า (i_t) คำนวณจากข้อมูลนำเข้า (x_t) ประกอบกับข้อมูลส่งออกก่อนหน้า (h_{t-1}) และใช้ฟังก์ชันซิกมอยด์เป็นตัวตัดสิน ตามสมการด้านล่าง

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

ค่าส่วนปรับค่ารับเข้า (g_t) คำนวณจากข้อมูลนำเข้า (x_t) ประกอบกับข้อมูลส่งออกก่อนหน้า (h_{t-1}) และใช้ฟังก์ชันฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ (Hyperbolic tangent : tanh) เป็นตัวตัดสินแทนฟังก์ชันซิกมอยด์ ตามสมการด้านล่าง โดยเมื่อผลลัพธ์ค่าส่วนปรับค่ารับเข้า (g_t) มีค่าเป็น 1 จะเก็บข้อมูลค่าส่วนรับเข้า (i_t) ลงในหน่วยความจำ

$$g_t = tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

2. การทำงานที่เลือกว่าข้อมูลใดจะถูกลบออกจากหน่วยความจำจากค่าส่วนคัดออก (f_t)

ค่าส่วนคัดออกคำนวณจากข้อมูลนำเข้าประกอบกับข้อมูลส่งออกก่อนหน้า และใช้ฟังก์ชันซิกมอยด์เป็นตัวตัดสิน ตามสมการด้านล่าง โดยเมื่อผลลัพธ์ค่าส่วนคัดออกมีค่าเป็น 0 ข้อมูลในหน่วยความจำจะถูกลบออกไป แต่ถ้าค่าส่วนคัดออกมีค่าเป็น 1 ข้อมูลในหน่วยความจำจะยังคงเก็บไว้

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

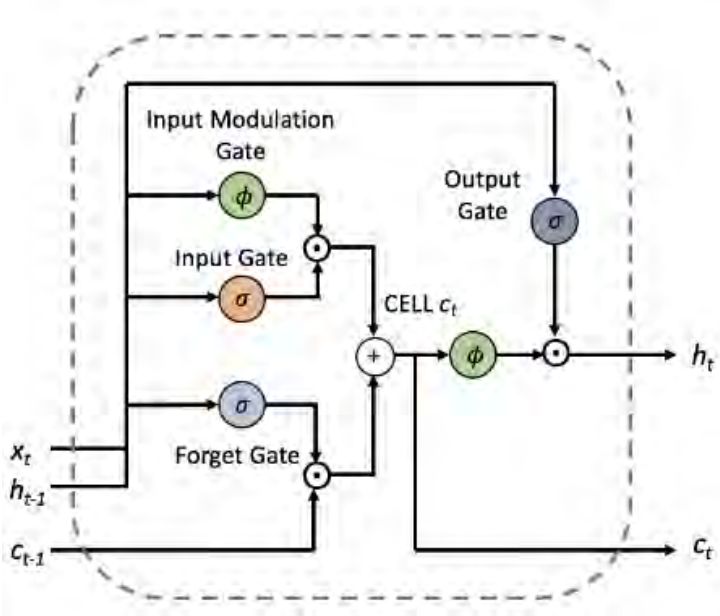
3. การทำงานที่เลือกว่าข้อมูลใดจะส่งออกไปภายนอกจากค่าส่วนส่งออก (o_t)

ค่าส่วนส่งออกคำนวณจากข้อมูลนำเข้าประกอบกับข้อมูลส่งออกก่อนหน้า และใช้ฟังก์ชันซิกมอยด์เป็นตัวตัดสิน ตามสมการด้านล่าง โดยเมื่อผลลัพธ์ค่าส่วนส่งออกมีค่าเป็น 1 จะอนุญาตให้ส่งข้อมูลออกภายนอกได้

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

4. การปรับปรุงข้อมูลที่ถูกเก็บในหน่วยความจำ เมื่อส่วนรับเข้า ส่วนปรับค่ารับเข้า และส่วนคัดออก มีการทำงานแล้ว จะมีการปรับปรุงข้อมูลที่ถูกเก็บในหน่วยความจำได้ ตามสมการ

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t$$

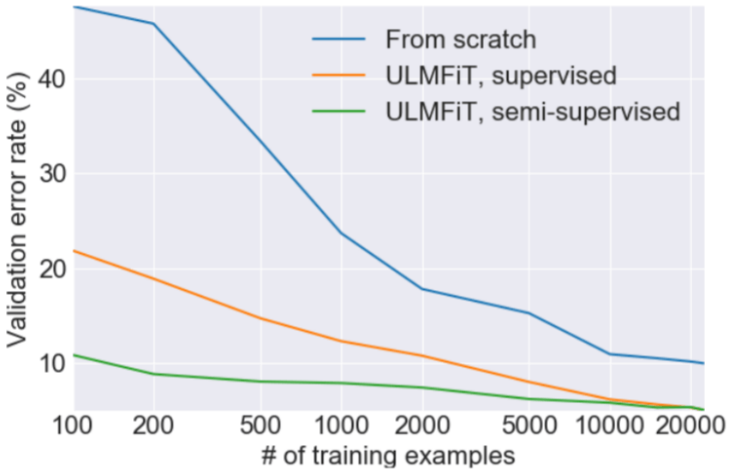


ภาพที่ 2.3 ตัวอย่างของแอลเอสทีเอ็ม

ที่มา: https://medium.com/@sinart.t/long-short-term-memory-lstm-e6cb23b494c6

2.3.2 การจำแนกข้อความด้วยยูแอลเอ็มฟิตโมเดล (ULMFiT model)

จากงานวิจัย [9] พบว่าการใช้ยูแอลเอ็มฟิต (Universal Language Model Fine-tuning for Text Classification: ULMFiT) ช่วยให้ให้การเรียนรู้มีประสิทธิภาพดีขึ้น และใช้ข้อมูลในการเรียนรู้น้อยกว่า ดังภาพที่ 2.4

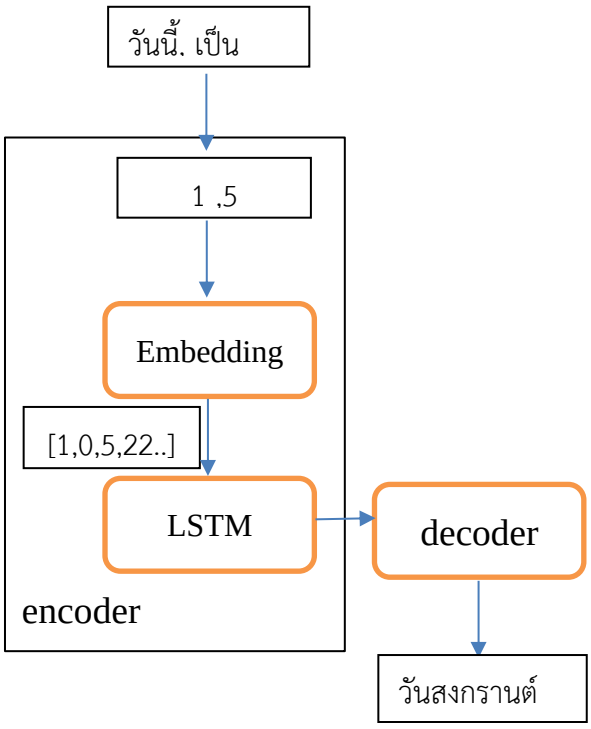


ภาพที่ 2.4 ตัวอย่างการเรียนรู้ด้วยยูแอลเอ็มฟิต

ยูแอลเอ็มฟิตเป็นการปรับค่าพารามิเตอร์ต่าง ๆ ในโครงข่ายประสาทเทียมด้วยการเรียนรู้โมเดลภาษา (Language model) คือ การสอนให้คอมพิวเตอร์สามารถเข้าใจโครงสร้างภาษาและลักษณะลำดับของคำในข้อความของภาษานั้น ตัวอย่างเช่น สมมติว่ามีข้อมูลนำเข้าเป็นคำว่า “วันนี้”

“เป็น” โมเดลภาษาจะคาดการณ์คำต่อท้ายของคำว่า “วันนี้” “เป็น” ว่าเป็นคำใดบ้าง ซึ่งจะพบว่าคำที่ควรต่อท้าย คือ “วันสงกรานต์” วิธีที่ทำให้เครื่องคอมพิวเตอร์เรียนรู้โมเดลภาษาได้ คือ ให้เครื่องเรียนรู้ข้อความเข้าไปจำนวนมาก โดยต้องมีการแปลงข้อมูลของคำเป็นเวกเตอร์ตัวเลขของคำ (embedding) จากค่าดัชนี (Index) ที่แตกต่างกันของแต่ละคำ เช่น วันนี้ มีดัชนีเป็น 1 และดัชนีนั้นแปลงเป็นเวกเตอร์ [1,0,5,22...] ที่ยาวมาก ๆ โดยขึ้นอยู่กับการกำหนดขนาดเวกเตอร์ จากนั้นนำเวกเตอร์คำที่ได้ส่งเข้าโครงข่ายประสาทแบบแอลเอสทีเอ็มสำหรับการเรียนรู้ลำดับของคำ โดยขั้นตอนการทำงานในส่วนการแปลงข้อมูลของคำเป็นเวกเตอร์ตัวเลขของคำและโครงข่ายประสาทเทียมแอลเอสทีเอ็ม จะเรียกรวมกันว่า ส่วนการเข้ารหัส (encoder) หลังจากทำงานในส่วนนี้เสร็จแล้ว ข้อมูลจะถูกส่งต่อไปยังส่วนการถอดรหัส (decoder) ซึ่งเป็นชั้นที่แปลงข้อมูลที่ได้ให้เป็นคำตอบของผลลัพธ์ ดังภาพที่ 2.5

สำหรับการวิเคราะห์ข้อความจะใช้ส่วนการเข้ารหัสที่มีการปรับค่าพารามิเตอร์ต่าง ๆ ตามโมเดลภาษานั้น ๆ แล้ว ไปต่อกับส่วนการถอดรหัสของการจำแนกข้อมูลอื่น เช่น การจำแนกความรู้สึกข้อความเป็นด้านบวกหรือด้านลบ



ภาพที่ 2.5 ตัวอย่างการจำแนกข้อความด้วยยูแอลเอ็มฟิตโมเดล

2.4 ตัวตัดคำภาษาไทยนิวเอ็มเอ็ม (newmm) [8]

เป็นหนึ่งในไลบรารีของทาง pythainlp ซึ่งใช้วิธีตัดคำของข้อความนั้นที่สามารถเป็นไปได้ทั้งหมดก่อนและเลือกข้อความที่ตัดแล้วให้ได้จำนวนคำน้อยที่สุดหรือคำยาวที่สุดในการตัดคำ

(Maximum Matching algorithm) เนื่องจากภาษาไทยมีการสร้างคำจากการประสมคำอยู่เป็นจำนวนมาก การเลือกผลที่มีจำนวนคำน้อยสุดจึงมีโอกาสถูกต้องมากกว่า ตัวอย่างเช่น

อากาศร้อนต้องเปิดพัดลม ตัดได้เป็น

อา|กาศ|ร้อน|ต้อง|เปิด|พัด|ลม กรณีเลือกให้ได้คำมากที่สุด

อากาศ|ร้อน|ต้อง|เปิด|พัดลม| กรณีเลือกให้ได้คำน้อยที่สุด

ซึ่งผลลัพธ์อย่างหลังก็จะถูกมากกว่า ดังภาพที่ 2.5

```
word_tokenize(text, engine='newmm')
['อากาศ', 'ร้อน', 'ต้อง', 'เปิด', 'พัดลม']
```

ภาพที่ 2.6 ตัวอย่างการตัดคำโดยใช้ newmm

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

TF-IDF คือ การหาคำสำคัญของข้อมูลที่เป็นข้อความ โดยการให้น้ำหนักคำในแต่ละคำ โดยใช้ 2 ปัจจัย คือ TF (Term Frequency) และ IDF (Inverse Document Frequency)

TF คือ การหาความถี่ของคำ (term) ที่ปรากฏในข้อความ (document) ซึ่งการหาค่า TF หาได้จากสูตร:

$\mathrm{TF}(t, d)$ = จำนวนคำที่ปรากฏ/จำนวนคำทั้งหมดในข้อความ

t = คำที่ปรากฏ d = เซตข้อมูลนั้น

ตัวอย่าง ยาย กิน ลำไย น้ำลาย ยาย ไหล ย้อย

ก็จะได้ค่า TF ของแต่ละคำ คือ

TF (“ยาย”, จำนวนคำทั้งหมดในข้อความ) = 2/7= 0.28

TF (“กิน”, จำนวนคำทั้งหมดในข้อความ) = 1/7= 0.14

TF (“ลำไย”, จำนวนคำทั้งหมดในข้อความ) = 1/7= 0.14

IDF คือ การวัดความสำคัญของคำที่ปรากฏในข้อความ ซึ่งถ้าคำนั้นปรากฏบ่อยในเอกสารย่อมมีความสำคัญลดลง

ตัวอย่างข้อความ 1 : ยาย กิน ลำไย น้ำลาย ยาย ไหล ย้อย

ตัวอย่างข้อความ 2 : ยาย กิน ส้มตำ น้ำลาย ยาย ไหล ย้อย

จะเห็นได้ว่า ข้อความที่ 1 และ 2 ที่เหมือนกัน คือ ยาย กิน น้ำลาย ไหล ย้อย

ข้อความที่ 1 และ 2 ที่แตกต่างกัน คือ ลำไย ส้มตำ

ดังนั้นจะเห็นได้ว่า คำที่แตกต่างกันจะมีความสำคัญมากกว่า เพราะเป็นคำที่ใช้บอกว่าข้อความนี้กล่าวถึงข้อมูลเรื่องอะไร ซึ่งหาได้จากสูตร

$$\text{IDF}(\text{t}, \text{allDocument}) = log\frac{N}{df(t)}$$

N คือ จำนวน ข้อความทั้งหมด $df(t)$ คือ จำนวนคำที่ปรากฏอยู่ในทุกข้อความ

ตัวอย่างผลลัพธ์ที่ได้คือ IDF(“กิน”,จากข้อความที่1) = log(2/2) = 0

ตัวอย่างผลลัพธ์ที่ได้คือ IDF(“ส้มตำ,จากข้อความที่2)=log(2/1) = 0.3

TF-IDF คือการนำผลลัพธ์ของ TF มาคูณกับ IDF ก็จะได้ค่าน้ำหนักของคำ (TF-IDF) ออกมา โดยถ้าค่าน้ำหนักของคำ ถ้า TF-IDF มีค่ามาก เกิดจากความถี่ของคำนั้นปรากฏในข้อความสูง แต่คำนั้นปรากฏน้อยในข้อความอื่น ๆ ดังนั้นคำนั้นจึงมีความสำคัญในการวิเคราะห์

2.6 ภาษาและไลบรารีที่ใช้

ภาษาที่ใช้สำหรับพัฒนาโมเดลนี้คือ ไพทอน (Python) เป็นภาษาสคริปต์ ทำให้ใช้เวลาในการเขียนและคอมไพล์ไม่มาก และยังถูกออกแบบให้มีโครงสร้างที่ไม่ซับซ้อน และภาษาที่ใช้ในการแสดงผลหน้าเว็บให้แก่ผู้ใช้ คือ เอชทีเอ็มแอล5 (HTML5) ซีเอสเอส (CSS) และ จาวาสคริปต์ (JAVASCRIPT)

ในโครงงานนี้ได้เลือกใช้ไลบรารีสำหรับพัฒนาระบบ ได้แก่

1. Tweepy เป็นไลบรารีซึ่งช่วยในการเชื่อมต่อไปยังทวิตเตอร์ หรือที่เรียกกันว่าทวิต-เตอร์เอพีไอ สำหรับการเก็บรวบรวมข้อมูลจากทวิตเตอร์ เช่น ชื่อผู้ใช้ทวิต วันที่ที่ทวิต ข้อความที่ทวิต ฯลฯ

2. pythainlp.umlftit เป็นไลบรารีของทาง pythainlp เพื่อช่วยขั้นตอนของการเตรียมข้อมูล (preprocess) ซึ่งได้แก่ ฟังก์ชัน word_tokenize ที่ช่วยในการตัดคำในที่นี้คือ newmm ฟังกชัน ungroup_emoji สำหรับการแยกอิโมจิ และฟังชัน replace_rep_nonum สำหรับแปลงคำซ้ำให้เหลือเพียงหนึ่งคำ

3. Pandas เป็นไลบรารีของภาษาไพทอน ที่มีความสามารถสำหรับจัดเตรียมข้อมูลให้พร้อมสำหรับการวิเคราะห์ โดยทำข้อมูลให้เป็นแถวเป็นแนวเพื่อสะดวกต่อการจัดการข้อมูล

4. Sklearn logisticregression เป็นไลบรารีสำหรับการจำแนกโดยใช้สมการของการถดถอยโลจิสติกส์

5. tfidfvectorizer ของ sklearn คือการหาค่าน้ำหนักของคำแต่ละคำอิงตามความสำคัญหรือความถี่ที่ปรากฏบนเอกสารโดยแปลงออกมาเป็นตัวเลข (tfidf) เพื่อที่จะนำตัวเลขค่าน้ำหนักของคำไปใช้ประมวลผลต่อไป

6. regular expression (re) เป็นไลบรารีสำหรับการกำหนดแพตเทิร์นของข้อความไว้สำหรับการทำความสะอาดข้อมูล

7. ไลบรารี fast.ai เป็นไลบรารีสำหรับการทำโมเดลยูแอลเอ็มฟิต

8. Flask web development คือ เว็บเฟรมเวิร์ก (web framework) สำหรับการพัฒนาเว็บแอปพลิเคชันด้วยภาษาไพทอน ซึ่ง Flask เป็นเฟรมเวิร์กขนาดเล็ก (micro-framework) ที่สะดวกและง่ายต่อการใช้งานสำหรับผู้พัฒนาเว็บ (web developer)

9. ไลบรารี Pymysql ใช้ในการจัดการฐานข้อมูล เพื่อนำข้อมูลที่อยู่ในฐานข้อมูลมาแสดงผล และนำมาวิเคราะห์

10. ไลบรารี Torch เป็นไลบรารีสำหรับทำเฟรมเวิร์กสำหรับสร้างโครงข่ายประสาทเทียม (artificial neural network) ซึ่ง Facebook พัฒนาบนภาษาไพทอนรวมถึงทำให้เรียกใช้โมเดลมาทำงานร่วมกับเว็บแอปพลิเคชันได้

บทที่ 3
การรวบรวมและวิเคราะห์ข้อมูล

ในบทนี้จะกล่าวถึงวิธีการรวบรวมข้อความแสดงความคิดเห็น เพื่อที่จะนำมาใช้วิเคราะห์และจำแนกข้อความแสดงความคิดเห็น

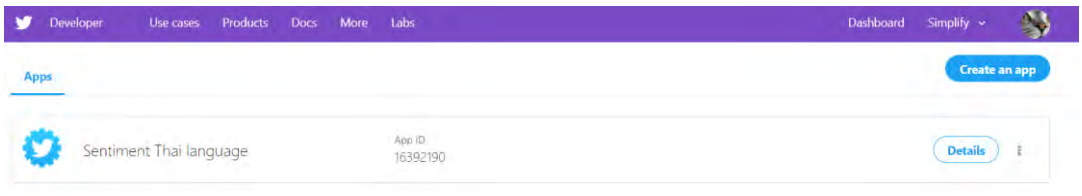
3.1 การรวบรวมข้อมูล

ผู้พัฒนาได้รวบรวมข้อความแสดงความคิดเห็นทางด้านการเมืองจาก 2 แหล่งข้อมูล ได้แก่ ข้อความบนทวิตเตอร์ (Twitter) จาก www.twitter.com โดยใช้ทวิตเตอร์ เอพีไอ (Twitter API) และเซตข้อมูลทวิตเตอร์ (Twitter Dataset) จาก archive.org ซึ่งข้อความที่เก็บมาได้อยู่ช่วงประมาณเดือนมิถุนายน ถึง เดือนกันยายน ผ่านแฮชแท็กทางการเมืองคือ #ประชุมสภา, #เพื่อไทย, #อนาคตใหม่, #พลังประชารัฐ ซึ่งจะอธิบายรายละเอียดในหัวข้อ 3.1.1 และหัวข้อ 3.1.2 ตามลำดับ

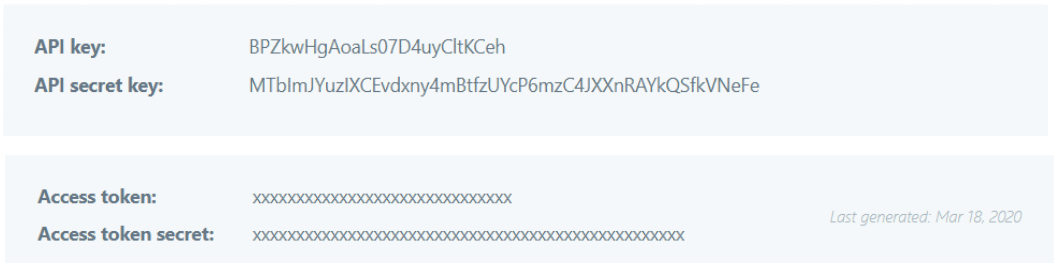
3.1.1 ทวิตเตอร์

ระบบเว็บทวิตเตอร์ เป็นเหมือนบริการเครือข่ายสังคมออนไลน์จำพวกไมโครบล็อก โดยผู้ใช้สามารถส่งข้อความได้ไม่เกิน 280 ตัวอักขระเท่านั้น (เท่ากันทุกภาษา ยกเว้นภาษาญี่ปุ่น จีน และเกาหลี ที่ยังส่งข้อความได้แค่ 140 ตัวอักขระคงเดิม) โดยเรียกข้อความที่ส่งนี้ว่า “ทวีต” (tweet) ทวิตเตอร์มีฟังก์ชันการทำงานต่าง ๆ ที่ใช้กับข้อความ เช่น การค้นหา การตอบข้อความ และการโพสต์ทวีต ดังนั้นการเก็บข้อความแสดงความคิดเห็นทางด้านการเมืองจากทวิตเตอร์ จะใช้ทวิตเตอร์เอพีไอในการดึงข้อมูลที่เป็นข้อความจากทวิตเตอร์ ซึ่งยังคงดึงได้ไม่เกิน 140 ตัวอักขระ ทำให้การดึงข้อมูลภาษาไทยในโครงงานนี้ เก็บข้อความได้เพียง 140 ตัวอักขระเท่านั้น

ในการใช้ทวิตเตอร์เอพีไออย่างแรกเลย ต้องมีการสร้างแอป (Apps) ผ่านทางบัญชีของผู้พัฒนาที่ https://developer.twitter.com ดังภาพที่ 3.1 ซึ่งการสร้างแอปจะทำให้ได้คีย์ (key) สำหรับระบุการใช้งาน ดังภาพที่ 3.2

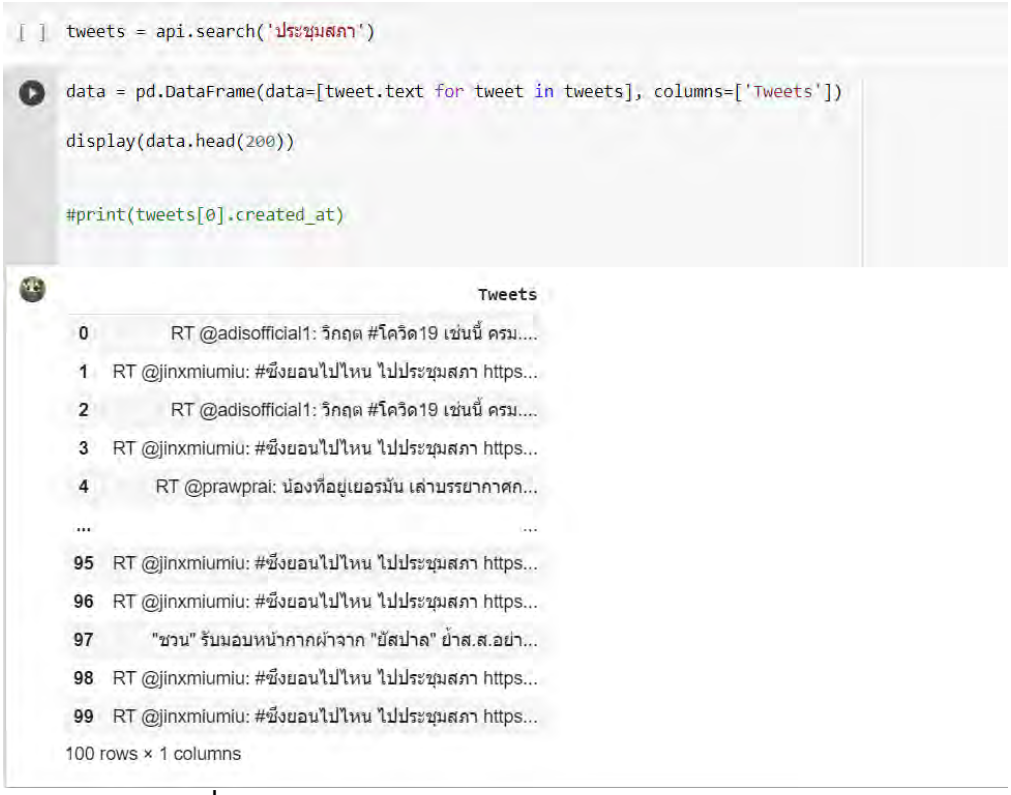


ภาพที่ 3.1 หน้าจอของแอปทวิตเตอร์



ภาพที่ 3.2 หน้ารหัสสำหรับการเรียกใช้ทวิตเตอร์เอพีไอ

หลังจากได้คีย์สำหรับการดึงข้อมูลแล้ว ต่อมาเป็นส่วนของการเขียนโค้ดสำหรับดึงข้อมูลจากทวิตเตอร์ โดยภาษาที่ใช้ในที่นี้คือ ภาษาไพทอน ไลบรารีหลักที่ใช้คือ ทวีไพ (Tweepy) ใช้สำหรับการดึงข้อความจากทวิตเตอร์ และใช้ไลบรารีแพนดาส (Pandas) ในการจัดการข้อมูลที่ดึงมาได้ให้อยู่ในรูปตาราง บางส่วนของโค้ดและผลลัพธ์การทำงานแสดงดังภาพที่ 3.3



```
[ ] tweets = api.search('ประชุมสภา')

data = pd.DataFrame(data=[tweet.text for tweet in tweets], columns=['Tweets'])

display(data.head(200))

#print(tweets[0].created_at)
```

	Tweets
0	RT @adisofficial1: วิกฤต #โควิด19 เช่นนี้ ครม....
1	RT @jinxmiumiu: #ซึงยอนไปไหน ไปประชุมสภา https...
2	RT @adisofficial1: วิกฤต #โควิด19 เช่นนี้ ครม....
3	RT @jinxmiumiu: #ซึงยอนไปไหน ไปประชุมสภา https...
4	RT @prawprai: น้องที่อยู่เยอรมัน เล่าบรรยากาศก...
...	...
95	RT @jinxmiumiu: #ซึงยอนไปไหน ไปประชุมสภา https...
96	RT @jinxmiumiu: #ซึงยอนไปไหน ไปประชุมสภา https...
97	"ชวน" รับมอบหน้ากากผ้าจาก "ยัสปาล" ย้ำส.ส.อย่า...
98	RT @jinxmiumiu: #ซึงยอนไปไหน ไปประชุมสภา https...
99	RT @jinxmiumiu: #ซึงยอนไปไหน ไปประชุมสภา https...

100 rows × 1 columns

ภาพที่ 3.3 ผลลัพธ์จากการดึงข้อมูลด้วยทวิตเตอร์เอพีไอ

3.1.2 เซตข้อมูลทวิตเตอร์ จาก archive.org

เนื่องจากการดึงข้อมูลด้วยทวิตเตอร์ เอพีไอนั้นมีข้อจำกัดบางอย่างอยู่ คือ ไม่สามารถดึงข้อความย้อนหลังได้เกิน 30 วัน หากไม่ได้สมัครแบบพรีเมียม ซึ่ง archive.org เป็นเว็บไซต์ที่เก็บข้อมูลเก่าของข้อมูลสื่อต่าง ๆ ที่ปรากฏบนอินเทอร์เน็ต และมีข้อความจากทวิตเตอร์ เก่า ๆ ถูกเก็บไว้ตามลิงค์ https://archive.org/details/twitterstream?&sort=-date&page=2 ผู้พัฒนาจึงดึงข้อมูลในส่วนนี้มาใช้ในโครงงาน เครื่องมือที่ช่วยในการพัฒนาที่ใช้ในที่นี้ คือ google colab โดยใช้คำสั่ง ! wget [option]... [URL] เพื่อดาวน์โหลดไฟล์จากเว็บไซต์ ดังภาพที่ 3.4

```
!wget https://ia803103.us.archive.org/7/items/archiveteam-twitter-stream-2019-06/twitter_stream_2019_06_30.tar -P  /content/
```

ภาพที่ 3.4 ตัวอย่างคำสั่งในการดาวน์โหลดไฟล์

โดยไฟล์ที่ได้จะเป็นนามสกุล .tar จากนั้นเรียกใช้ไลบรารี wfile เพื่อที่จะแยกไฟล์แต่ละไฟล์มาประมวลผล ดังภาพที่ 3.5

```
[ ]  import tarfile, bz2, json, csv
     from tqdm.auto import tqdm

[ ]  tar = tarfile.open("twitter_stream_2019_06_30.tar")
     names = sorted(nm for nm in tar.getnames() if nm.endswith('.bz2'))

[ ]  wfile = open('thai-2019-06-30.csv','w')
     writer = csv.writer(wfile)
     writer.writerow(['id', 'timestamp_ms', 'user_id', 'text'])
```

ภาพที่ 3.5 ส่วนของการแยกไฟล์

เนื่องจากไฟล์ที่เก็บจะรวมข้อความของทุกภาษา ดังนั้นผู้พัฒนาจะดึงเฉพาะไฟล์ข้อความที่เป็นภาษาไทยมาบันทึกเก็บไว้ประมวลผล ดังภาพที่ 3.6

```
[ ]  for name in tqdm(names):
       f = tar.extractfile(name)
       for line in bz2.open(f, 'rt'):
         d = json.loads(line)
         if d.get('lang') != 'th':  # thai only
           continue
         row = [d['id_str'], d['timestamp_ms'], d['user']['id_str'], d['text'] ]
         writer.writerow(row)
     wfile.close()
```

ภาพที่ 3.6 การกรองข้อความเฉพาะภาษาไทยและบันทึกไฟล์

จากนั้นนำข้อความภาษาไทยมาค้นเฉพาะข้อความที่เกี่ยวข้องด้านการเมืองตามแฮชแท็กทางการเมืองที่กำหนดไว้

```
[ ] import pandas as pd

    ##ใส่ชื่อไฟล์เป็น list ไว้
    dataset = ['thai-2019-06-01.csv',
               'thai-2019-06-02.csv',
               'thai-2019-06-03.csv',
               'thai-2019-06-04.csv',
               'thai-2019-06-05.csv',
               'thai-2019-06-06.csv',
               'thai-2019-06-07.csv',
               'thai-2019-06-08.csv',
               'thai-2019-06-09.csv',
               'thai-2019-06-10.csv',
               'thai-2019-06-11.csv',
               'thai-2019-06-12.csv',
               'thai-2019-06-13.csv',
               'thai-2019-06-14.csv',
               'thai-2019-06-15.csv',
               'thai-2019-06-16.csv',
               'thai-2019-06-17.csv',
               'thai-2019-06-18.csv',
               'thai-2019-06-19.csv',
               'thai-2019-06-20.csv',
               'thai-2019-06-21.csv',
               'thai-2019-06-22.csv',
               'thai-2019-06-23.csv',
               'thai-2019-06-25.csv',
               'thai-2019-06-23.csv',
               'thai-2019-06-25.csv',
               'thai-2019-06-26.csv',
               'thai-2019-06-27.csv',
               'thai-2019-06-28.csv',
               'thai-2019-06-29.csv',
               'thai-2019-06-30.csv']
    all_dataset = []

    hashtag = '#ประชุมสภา'
    hashtag2 = 'เพื่อไทย'
    hashtag3= 'อนาคตใหม่'
    hashtag4 = 'พลังประชารัฐ'

    ##ลูปอ่านทีละไฟล์
    for file in dataset:
        df = pd.read_csv(file,engine='python',error_bad_lines=False)
        all_dataset.append(df)

    ##concat เป็น pandas
    result = pd.concat(all_dataset)

    ##เลือก column ที่สนใจ
    result = result[['text']]

      ##กรอง hashtag
      result = result[(result['text'].str.contains(hashtag,na=False)) |
                      (result['text'].str.contains(hashtag2,na=False)) |
                      (result['text'].str.contains(hashtag3,na=False)) |
                      (result['text'].str.contains(hashtag4,na=False))]

      ##สร้าง dataFrame
      result = pd.DataFrame(result)

      ##ตกแต่งต่างๆ
      pd.set_option('display.max_colwidth',-1) ## ให้ข้อความเต็มแสดง column
      result.reset_index(inplace=True) ##ให้เริ่มที่ index 0

      ##print
      result
```

ภาพที่ 3.7 การกรองข้อความภาษาไทยเฉพาะแฮชแท็กทางการเมืองของเดือนมิถุนายน

สรุปการเก็บข้อมูลที่ได้จากทางทวิตเตอร์ เอพีไอ และเซตข้อมูลของ archive.org ตั้งแต่เดือนมิถุนายน ถึง เดือนกันยายน โดยรวมจากทั้ง 2 แหล่ง จากทวิตเตอร์ เป็นจำนวน 2,955 ทวีต จาก archive.org เป็นจำนวน 7,173 ทวีต โดย

เดือน มิถุนายน เก็บได้เป็นจำนวน 4,023 ทวีต

เดือน กรกฎาคม เก็บได้เป็นจำนวน 3,150 ทวีต

เดือน สิงหาคม เก็บได้เป็นจำนวน 1,200 ทวีต

เดือน กันยายน เก็บได้เป็นจำนวน 1,755 ทวีต

ผลการเก็บข้อมูลได้ข้อความทั้งหมด เป็นจำนวน 10,128 ทวีต

3.2 การวิเคราะห์ข้อมูล

จากนั้นทีมผู้พัฒนาแบ่งกลุ่มข้อความทั้งหมดที่ได้เก็บรวบรวมมา ออกเป็นข้อความด้านบวก ด้านลบ และเป็นกลาง โดยใช้ผู้อ่านจำนวน 3 คนตามเสียงส่วนใหญ่ว่าข้อความอยู่ในกลุ่มใด ตามเกณฑ์ดังนี้

1. คำที่แสดงถึงความรู้สึก

ด้านบวก เช่น ดี เยี่ยม รัก เก่ง น่ารัก สวย เท่ อร่อย สุดยอด เป็นต้น

ด้านลบ เช่น แย่ ไม่ไหว ไม่ดี ทำไมทำแบบนี้ คิดอะไรอยู่ ทำไปเพื่ออะไร เป็นต้น

2. คำไม่สุภาพ ซึ่งคำทั้งหมดจะสื่อไปในด้านลบ เช่น กู มึง เหี้ย ส้นตีน ห่า แม่ง เป็นต้น

3. คำประชดประชัน ซึ่งคำเหล่านี้จะสื่อไปในทางลบเช่นกัน เช่น ปัญหาระดับนี้ผู้นำทำได้ดีมากเลยสุดยอดจริง ๆ เป็นต้น

4. อีโมจิ

ด้านบวก 😀😃😄😁😆😍😋😊🙂😊🤗 เป็นต้น

ด้านลบ 😐😑😒🙄😠🤔😯😞😕😠😡🤬 เป็นต้น

จะนำเกณฑ์นี้มาตัดสินว่าข้อความนั้นเป็นข้อความในด้านใด

ผลลัพธ์ที่ได้จากการแบ่งกลุ่มข้อความจากการอ่านของทีมผู้พัฒนา แบ่งได้เป็น ข้อความที่เป็นบวกมีจำนวน 2,089 ทวีต ข้อความที่เป็นลบมีจำนวน 3,947 ทวีต ข้อความที่เป็นกลางมีจำนวน 4,052 ทวีต ข้อความที่มีมากกว่าหนึ่งความรู้สึกในทวีตมีจำนวน 40 ทวีต ดังตารางที่ 3.1

ตารางที่ 3.1 ผลลัพธ์จากการแบ่งกลุ่มข้อความ

ข้อความด้านบวก	2,089 ทวีต
ข้อความด้านลบ	3,947 ทวีต
ข้อความเป็นกลาง	4,052 ทวีต
ข้อความที่มีมากกว่า 1 ความรู้สึก	40 ทวีต

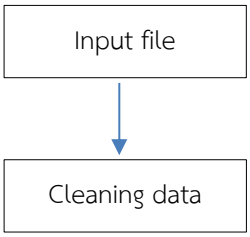
ลักษณะข้อความที่เก็บรวบรวมได้เป็นประโยคที่สามารถบอกความรู้สึกของข้อความว่าเป็นด้านบวก ด้านลบ หรือเป็นกลาง โดยจะไม่พิจารณาข้อความที่มีมากกว่า 1 ความรู้สึกในทวีต เนื่องจากข้อจำกัดของทวิตเตอร์ คือสามารถทวีตได้ไม่เกิน 140 ตัวอักษร ทำให้ส่วนมากประโยคที่เจอในทวิตเตอร์นั้นเป็นประโยคที่สั้น กระชับ และได้ใจความ จบในประโยคเดียว ข้อความที่มีมากกว่าหนึ่งความรู้สึกจึงพบได้น้อยมาก และไม่ค่อยเจอปัญหาที่มีข้อความมากกว่าหนึ่งความรู้สึกในทวีต

บทที่ 4
การสร้างโมเดลการจำแนกข้อความ

ในบทนี้จะกล่าวถึงการออกแบบและพัฒนาโมเดลสำหรับการวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมือง

4.1 การสร้างโมเดลในการวิเคราะห์ความรู้สึกทางการเมือง

การทำงานของระบบการวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมือง ประกอบด้วย 3 ส่วนหลัก คือ ส่วนการเตรียมข้อมูล (Preprocessing) ส่วนการวิเคราะห์ข้อมูล (Analysis) และส่วนวิเคราะห์และจำแนกข้อความแสดงความคิดเห็น (Classification) โดยในส่วนการเตรียมข้อมูลก่อนที่จะนำไปสร้างโมเดลสำหรับสำหรับการจำแนกจะผ่านขั้นตอนการรับข้อมูล (Input file) และการทำความสะอาดข้อมูล (Cleaning data) ดังภาพที่ 4.1



ภาพที่ 4.1 ส่วนการเตรียมข้อมูล

ขั้นตอนการเตรียมข้อมูล

ขั้นตอนที่ 1 นำข้อความจากทวิตเตอร์ ที่เก็บรวบรวมในบทที่ 3 มาเก็บเป็นเฟรมข้อมูล (Data frame) ผ่านไลบรารีแพนดาส สำหรับจัดเตรียมข้อมูลไว้สำหรับการสร้างโมเดล ผลลัพธ์ของขั้นตอนนี้ดังภาพที่ 4.2 และตัวอย่างการเรียกดูข้อความในเฟรมข้อมูลดังภาพที่ 4.3

	text	label
0	ให้ตายสิ เอาไป 17 พรรคแล้ว ยังมีจำนวน ส.ส หนอ...	ลบ
1	สุชาติ ตันเจริญรองปธ.สภาคนที่ 1 มาจากพรรคพลังป...	ลบ
2	#พรรคพลังประชารัฐ ไปลักกันค่า\nช่วงซิทคอม #มงค...	ลบ
3	พลังประชารัฐ มีคนไปยื่นคำร้อง ตั้งแต่เรื่องโต้...	ลบ
4	สำนักข่าว Deutsche Welle ของเยอรมนีทำคลิปสั้นเร...	กลาง
...	...	...
10082	พลังประชารัฐร่องเหี้ยดังมากกก 😂#ประชุมสภา	ลบ
10083	เฮ้ยๆ พระบิดาแห่งการยกเว้น ตื่นก่อน #ประชุมสภา...	กลาง

ภาพที่ 4.2 ตัวอย่างข้อความเมื่อใช้แพนดาสสำหรับเตรียมข้อมูล

```
[30] df['text'][10082]

'พลังประชารัฐร่องเหี้ยดังมากกก 😂#ประชุมสภา '
```

ภาพที่ 4.3 ตัวอย่างการเรียกดูข้อความในเฟรมข้อมูล

ขั้นตอนที่ 2 นำข้อความในที่เก็บไว้เฟรมข้อมูลมาทำความสะอาดข้อมูลโดยจะลบส่วนที่ไม่จำเป็นออก ได้แก่ สัญลักษณ์หรือตัวอักขระพิเศษ ลิงก์ แฮชแท็ก และรีทวีตออกไป แต่ยังเก็บสัญลักษณ์ที่แสดงอารมณ์ความรู้สึก หรือที่เรียกว่า อิโมจิ ไว้สำหรับการวิเคราะห์ โดยใช้ไลบรารี regular expression(re) ในการจับแพตเทิร์นของข้อความ ดังต่อไปนี้

text = re.sub(r'<.*?>','', text)	แทนที่ข้อความที่อยู่ใน <> ด้วยสตริงว่าง
text = re.sub(r'#','',text)	แทนที่ # ด้วยสตริงว่าง
text = re.sub(r'^[""RT]+','',text)	แทนที่ RT ด้วยสตริงว่าง
text = re.sub(r'@\S+', '', text)	แทนที่ข้อความที่อยู่ใน @ ด้วยสตริงว่าง
text = ' '.join(text.split())	ลบเว้นวรรคหรือบรรทัดใหม่
text = re.sub(r'http\S+', '', text)	แทนที่ข้อวามที่อยู่ใน http ด้วยสตริงว่าง

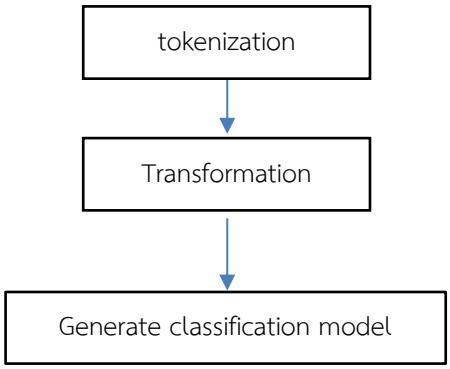
```
df['text'][555]

'ดรีมทีมมาแว้ววววว จัดจ้านในย่านกทม©\n\nฝากเป็นปากเป็นเสียงแทนชาวทวิตภพในสภาด้วยนะครับ🙏\n\n#เพื่อไทย \n#ตั้งรัฐบาล '

[61] df['text'][555]

'ดรีมทีมมาแว้ววววว จัดจ้านในย่านกทม© ฝากเป็นปากเป็นเสียงแทนชาวทวิตภพในสภาด้วยนะครับ🙏 เพื่อไทย ตั้งรัฐบาล'
```

ภาพที่ 4.4 ตัวอย่างของข้อมูลก่อนและหลังทำความสะอาด

เมื่อเสร็จส่วนการเตรียมข้อมูลแล้ว หลังจากนั้นจะทดลองใช้วิธีการจำแนกความรู้สึกของข้อความแสดงความคิดเห็น 2 วิธี คือ โมเดลจากวิธีการทางสถิติการถดถอยโลจิสติกส์ และโมเดลจากการเรียนรู้เชิงลึกด้วยยูแอลเอ็มฟิต รายละเอียดอธิบายในหัวข้อที่ 4.2 และ 4.3 ตามลำดับ

4.2 การสร้างโมเดลเพื่อจำแนกข้อความแสดงความคิดเห็นด้วยการถดถอยโลจิสติกส์

หลังจากผ่านส่วนการเตรียมข้อมูลแล้ว ต่อมาจะกล่าวถึงส่วนการวิเคราะห์และจำแนกข้อความแสดงความคิดเห็นของโมเดลที่สร้างด้วยวิธีการถดถอยโลจิสติกส์ ซึ่งประกอบไปด้วย การตัดคำ การแปลงคำเหล่านั้นให้เป็นตัวเลขเพื่อให้โมเดลสามารถคำนวณได้ และการสร้างโมเดลการจำแนกความรู้สึกของข้อความ ดังภาพที่ 4.5



ภาพที่ 4.5 ขั้นตอนการสร้างโมเดลการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์

ขั้นตอนวิเคราะห์และจำแนกข้อความด้วยการถดถอยโลจิสติกส์

ขั้นตอนที่ 1 นำข้อความที่ผ่านการทำความสะอาดแล้วมาตัดคำด้วยไลบรารี pythainlp.ulmfit.processthai ซึ่งมีฟังก์ชัน word_tokenize ในการตัดคำโดยเอนจินนิวเอ็มเอ็ม (รายละเอียดอธิบายในหัวข้อ 2.4) ซึ่งมีวิธีการตัดคำในข้อความให้ได้คำที่ยาวที่สุดและจำนวนคำในข้อความน้อยที่สุด มีการแปลงคำซ้ำให้เหลือเพียงหนึ่งคำ เช่น หายหายหาย เป็น “หาย” กับ “xxrep” ซึ่งหมายถึงคำซ้ำ โดยใช้ฟังชัน replace_rep_nonum และใช้ฟังก์ชัน ungrop_emoji ในการแยกอีโมจิ หลังจากนั้นใช้แพนดาสในการช่วยนับจำนวนคำในข้อความ ผลลัพธ์จากขั้นตอนนี้ได้ดังภาพที่ 4.6 และภาพที่ 4.7

	text	label	processed	wc
0	ให้ตายสิ เอาไป 17 พรรคแล้ว ยังมีจำนวน ส.ส หน่อ...	ลบ	ให้|ตาย|สิ|เอา|ไป|17|พรรค|แล้ว|ยัง|มี|จำนวน|ส....	25
1	สุชาติ ตันเจริญรองปธ.สภาคนที่ 1 มาจากพรรคพลังป...	ลบ	สุ|ชาติ|ตัน|เจริญ|รอง|ปธ.|สภา|คน|ที่|1|มาจาก|...	32
2	พรรคพลังประชารัฐ ไปสักกันค่า ชวงชิทคอม มงคลกิ...	ลบ	พรรค|พลัง|ประชา|รัฐ|ไป|สัก|กัน|ค่า|ชวง|ชิทคอม...	18
3	พลังประชารัฐ มีคนไปยื่นค่าร้อง ตั้งแต่เรื่องโด้...	ลบ	พลัง|ประชา|รัฐ|มี|คน|ไป|ยื่น|ค่าร้อง|ตั้งแต่|เรื่...	25
4	สำนักข่าว Deutsche Well ของเยอรมนีทำคลิปสั้นเร...	กลาง	สำนักข่าว|deutsche|well|ของ|เยอรมนี|ทำ|คลิป|สั้...	29
...	...	...	...	...
10082	พลังประชารัฐร้องเพี้ยดังมากกก 😂ประชุมสภา	ลบ	พลัง|ประชา|รัฐ|ร้อง|เพี้ย|ดัง|มาก|กก|😂|ประชุม|สภา	11
10083	เฮ้ยๆ พระบิดาแห่งการยกเว้น ตื่นก่อน ประชุมสภา ...	กลาง	เฮ้ย|ๆ|พระ|บิดา|แห่ง|การยกเว้น|ตื่น|ก่อน|ประชุ...	12
10084	อนาคตใหม่คืออิลูมินาติ คอนเฟิร์ม555555 อีหยังวะ	กลาง	อนาคต|ใหม่|คือ|อิลู|มิ|นา|ติ|คอนเฟิร์ม|5|xxrep...	13
10085	ไอดอลในดวงใจ ชูป้ายไฟค้บ ☢☢ ประชุมสภา	บวก	ไอดอล|ใน|ดวงใจ|ชู|ป้าย|ไฟ|ค้บ|xxwrep|☢|ประชุม|สภา	11
10086	ผู้นำที่ดีต้องฟังเสียงทักท้วงติดิงแต่นี่อะไรนอ...	ลบ	ผู้นำ|ที่|ดี|ต้อง|ฟัง|เสียง|ทักท้วง|ติดิง|แต่|...	21

ภาพที่ 4.6 ตัวอย่างผลลัพธ์หลังจากการตัดคำและนับคำ

```
df['processed'][5]
'สส.|ขยะ|ตลาด|ล่าง|สด|ต้อง|ฝาก|ประเทศ|ไว้|กับ|คน|แบบนี้|จริงๆ|หรอ|😔|😭|rip|พรรค|พลัง|ประชา|รัฐ|สส|ขอ|ส่วนบุญ|สส|ตลาด|ล่าง'
```

ภาพที่ 4.7 ตัวอย่างของข้อความ 1 ทวีต ที่ผ่านขั้นตอน tokenization

ขั้นตอนที่ 2 นำข้อความผลลัพธ์ที่ได้จากขั้นตอนที่ 1 มาใช้กับไลบรารี TfidfVectorizer ของ scikitlearn (รายละเอียดอธิบายในหัวข้อ 2.5) เพื่อสร้างเมทริกซ์ความสัมพันธ์ระหว่างคำกับเอกสาร (Term-Document Matrix) พร้อมกับตั้งค่าพารามิเตอร์การกำหนดจำนวนของคำด้วยฟังก์ชัน ngram_range(1,2) ซึ่งหมายถึง นับจำนวนคำต่ำสุดหนึ่งคำและคำสูงสุดสองคำ ดังนั้นการนับจำนวนคำที่ปรากฏในเอกสารจะพิจารณาแต่ละคำ (แบบ unigram) และพิจารณาคำสองคำที่ปรากฏติดกันเสมอ (แบบ bigram) ผลลัพธ์ที่ได้จากขั้นตอนนี้ คือ คำหนึ่งคำและคำที่ได้จากสองคำติดกัน โดยแต่ละคำจะมีค่าตัวเลขเฉพาะที่บ่งบอกถึงค่า TF-IDF ของคำนั้น ตัวอย่างดังภาพที่ 4.8 ในคอลัมน์ score จะเป็นค่า TF-IDF ของคำว่า “ประชุม” “สภา” และ “ประชุม สภา”

	rank	feature	score	ngram	label
0	0	ประชุม	0.029972	1	ลบ
1	1	สภา	0.029269	1	ลบ
2	2	ประชุม สภา	0.028074	2	ลบ
3	3	ไม่	0.023164	1	ลบ

ภาพที่ 4.8 ตัวอย่างคุณลักษณะของคำในคลาสลบที่ใช้ในการจำแนกกลุ่ม

จากภาพที่ 4.8 คอลัมน์ rank จะบอกถึงอันดับค่า TF-IDF ของคำเรียงจากค่ามากไปค่าน้อย โดยคำใน rank 0 หมายถึง มีคะแนนมากที่สุดในคลาส (label) แต่ละคลาส เช่น “ประชุม” เป็นคำที่มีอันดับสูงสุดในคลาสลบ “ดี” เป็นคำที่มีอันดับสูงสุดในคลาสบวก และ “พรรค” เป็นคำที่มีอันดับสูงสุดในคลาสกลาง ดังภาพที่ 4.9 และภาพที่ 4.10

	rank	feature	score	ngram	label
0	0	ดี	0.033606	1	บวก
1	1	xxrep	843	1	บวก
2	2	ดี xxrep	0.028816	2	บวก
3	3	มาก	0.025744	1	บวก
4	4	เพื่อ	0.024701	1	บวก

ภาพที่ 4.9 ตัวอย่างคุณลักษณะของคำในคลาสบวกที่ใช้ในการจำแนกกลุ่ม

	rank	feature	score	ngram	label
0	0	พรรค	0.022400	1	กลาง
1	1	ส.ส.	0.021412	1	กลาง
2	2	พรรค อนาคต	0.019532	2	กลาง
3	3	รัฐ	0.018937	1	กลาง
4	4	ไทย	0.018507	1	กลาง

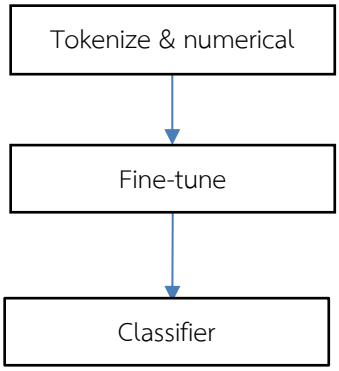
ภาพที่ 4.10 ตัวอย่างคุณลักษณะของคำในคลาสกลางที่ใช้ในการจำแนกกลุ่ม

ขั้นตอนที่ 3 ใช้ค่า TF-IDF (score) ของคำทั้งหมดที่ได้จากขั้นตอนที่ 2 (คำแต่ละคำและคำที่ได้จากสองคำติดกัน) และจำนวนคำทั้งหมดในข้อความ (wc) ที่ได้จากขั้นตอนที่ 1 เป็นคุณลักษณะที่ใช้ในการเรียนรู้ของเครื่องในการจำแนกความรู้สึกของข้อความด้วยไลบรารี LogisticRegression ของ scikitlearn ซึ่งใช้การถดถอยโลจิสติกส์ในการกำหนดค่าซิกมอยด์และพารามิเตอร์ต่าง ๆ สำหรับการจำแนกกลุ่มของข้อความ

4.3 การสร้างโมเดลเพื่อจำแนกข้อความแสดงความคิดเห็นด้วยยูแอลเอ็มฟิต

หลังจากผ่านส่วนการเตรียมข้อมูลแล้ว ต่อมาเป็นส่วนของการสร้างโมเดลภาษา โดยทำการดาวน์โหลดโมเดลภาษาไทยจากคลังข้อมูลขนาดใหญ่ ในที่นี้คือ วิกิพีเดียภาษาไทย จากนั้นนำเซตข้อมูลที่ได้ผ่านการทำความสะอาด นำมาผ่านฟังก์ชันในการตัดคำและกำหนดดัชนีระบุคำ โดยค่าของดัชนีจะถูกเปลี่ยนเป็นเวกเตอร์ของคำนั้น และนำเซตข้อมูลเวกเตอร์มาเพิ่มในโมเดลภาษา โดยผ่านไลบรารี fast.ai จากนั้นทำการเรียนรู้เพื่อให้โมเดลภาษาสามารถเข้าใจคุณลักษณะ หรือโครงสร้างภาษาของคำที่ใช้กันในทวิตเตอร์ จากนั้นนำเฉพาะส่วนการเข้ารหัส ดังภาพที่ 2.5 (ในบทที่ 2) ซึ่งได้แก่ ส่วนการแปลงข้อมูลของคำเป็นเวกเตอร์ตัวเลขของคำและโครงข่ายประสาทเทียมแอลเอสทีเอ็ม ที่เรียนรู้โมเดลภาษาจากวิกิพีเดียภาษาไทยและข้อความด้านการเมืองจากทวิตเตอร์ที่เก็บรวบรวมในโครงงานนี้ มาต่อกับส่วนการถอดรหัส ซึ่งเป็นการจำแนกความรู้สึกของข้อความแสดงความคิดเห็นเป็น

ด้านบวก ลบ เป็นกลาง ภาพที่ 4.11 แสดงขั้นตอนการสร้างโมเดลการจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิต

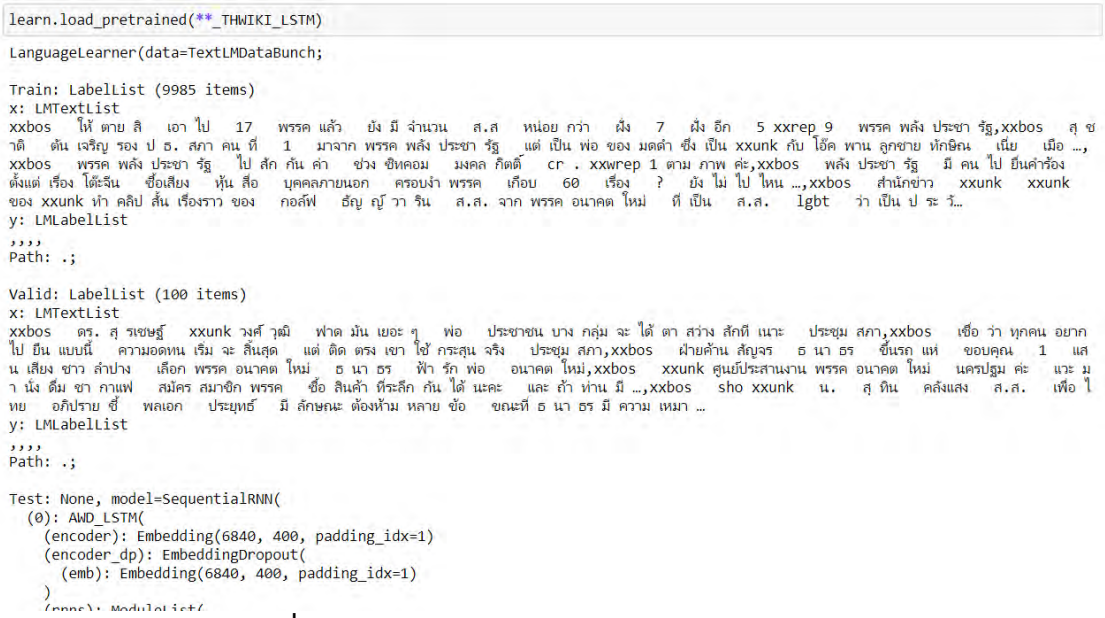


ภาพที่ 4.11 ขั้นตอนการสร้างโมเดลการจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิต

ขั้นตอนวิเคราะห์และจำแนกข้อความ

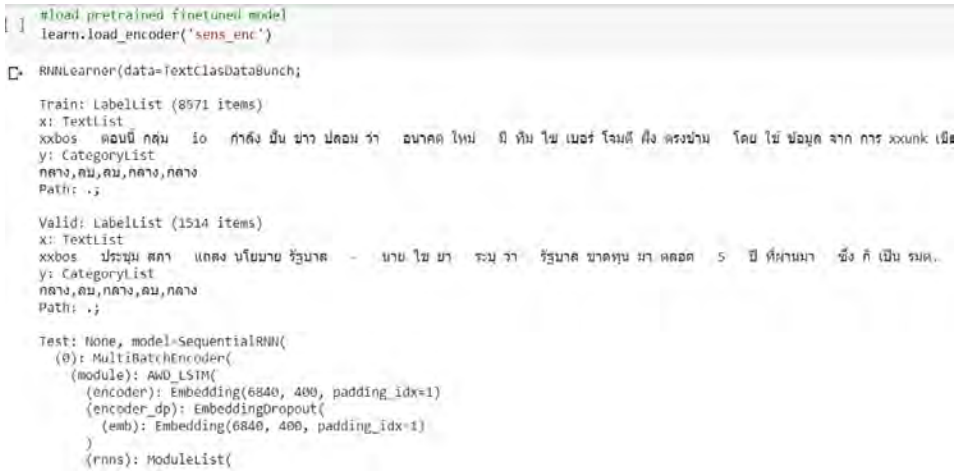
ขั้นตอนที่ 1 มีการทำงานเหมือนขั้นตอนแรกในการสร้างสร้างโมเดลการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์ แต่การแทนค่าคำด้วยตัวเลขเฉพาะที่บ่งบอกถึงคำนั้น ใช้ไลบรารีของทาง fast.ai ในการสร้างพจนานุกรมสำหรับคลังคำศัพท์สำหรับระบุคำแต่ละคำ เป็นค่าดัชนี

ขั้นตอนที่ 2 การแปลงคำเป็นเวกเตอร์คำ และการปรับจูนค่าพารามิเตอร์ต่าง ๆ ของโครงข่ายประสาทเทียม โดยใช้โมเดลภาษาที่สร้างจากการดาวน์โหลดข้อความภาษาไทยจากวิกิพีเดียภาษาไทยรวมกับข้อความแสดงความคิดเห็นที่เก็บรวบรวมได้ในโครงงาน การทำในส่วนนี้ไม่จำเป็นต้องกำหนดคลาส แต่ให้โครงข่ายประสาทเทียมเรียนรู้ให้เข้าใจข้อความภาษาไทยที่ใช้สื่อสารกันบนทวิตเตอร์ให้มากที่สุด ดังภาพที่ 4.12 หลังจากการเรียนรู้เพื่อให้เครื่องเข้าใจโมเดลภาษา จะบันทึกโครงข่ายประสาทเทียมในส่วนการเข้ารหัสเก็บไว้



ภาพที่ 4.12 การป้อนข้อความจากทวิตเตอร์ให้โมเดลภาษา

ขั้นตอนที่ 3 ดึงส่วนการเข้ารหัสที่บันทึกไว้ มาเพิ่มชั้นสำหรับการจำแนกความรู้สึกของข้อความ ที่เป็นส่วนการถอดรหัส เพื่อสร้างเป็นโมเดลที่ใช้จำแนกว่า ข้อความที่ปรากฏอยู่นี้ เป็นด้านบวก ด้านลบ หรือเป็นกลางดังภาพที่ 4.13 โดยเซตข้อมูลที่เรียนรู้ในรอบนี้จะต้องมีการกำหนดคลาสของคำตอบ เพื่อบ่งบอกกลุ่มของข้อความที่ปรากฏในหนึ่งทวิตนั้นว่ามีความรู้สึกด้านใด เพื่อให้โมเดลเกิดการเรียนรู้สำหรับการจำแนกความรู้สึกของข้อความแสดงความคิดเห็น



ภาพที่ 4.13 ตัวอย่างการเรียนรู้ข้อมูลสำหรับการแยกความรู้สึกของข้อความ

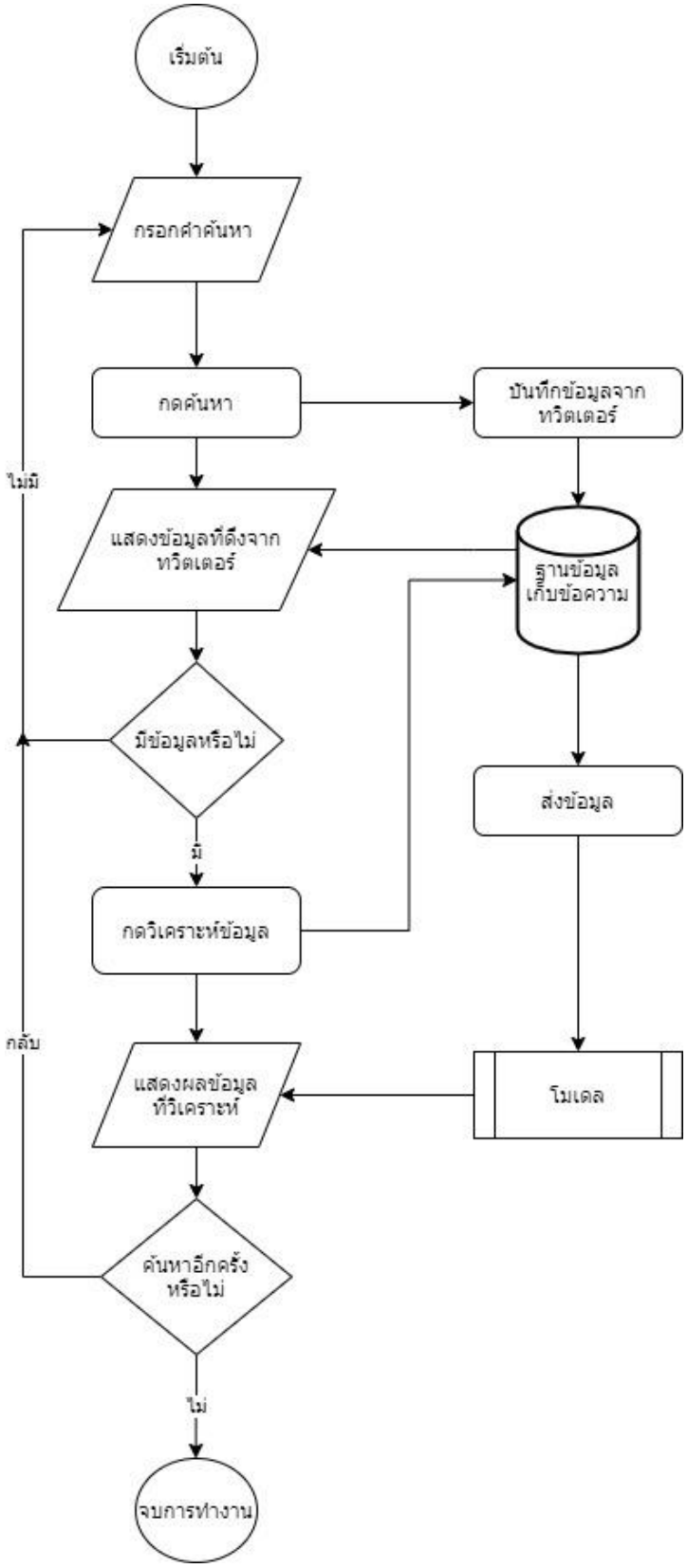
บทที่ 5
การออกแบบและพัฒนาเว็บแอปพลิเคชัน

ในบทนี้จะกล่าวถึง การออกแบบและการพัฒนาเว็บแอปพลิเคชัน เพื่อแสดงผลจากการวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์ได้แบบเรียลไทม์ (Realtime) และสามารถค้นหาข้อความบนทวิตเตอร์ตามแฮชแท็ก โดยจะกล่าวถึงการออกแบบระบบทั้งหมด ได้แก่ การออกแบบภาพรวมการทำงานเว็บแอปพลิเคชัน แผนภาพยูสเคส (Use case diagram) และแผนภาพแพ็กเกจ (Package diagram) นอกจากนี้จะกล่าวถึงการพัฒนาเว็บแอปพลิเคชันทั้งหมด คือ ภาษาและโปรแกรมที่ใช้ในการพัฒนาเว็บแอปพลิเคชันและส่วนต่อประสานผู้ใช้ (User interface) ดังรายละเอียดต่อไปนี้

5.1 การออกแบบภาพรวมการทำงานเว็บแอปพลิเคชัน

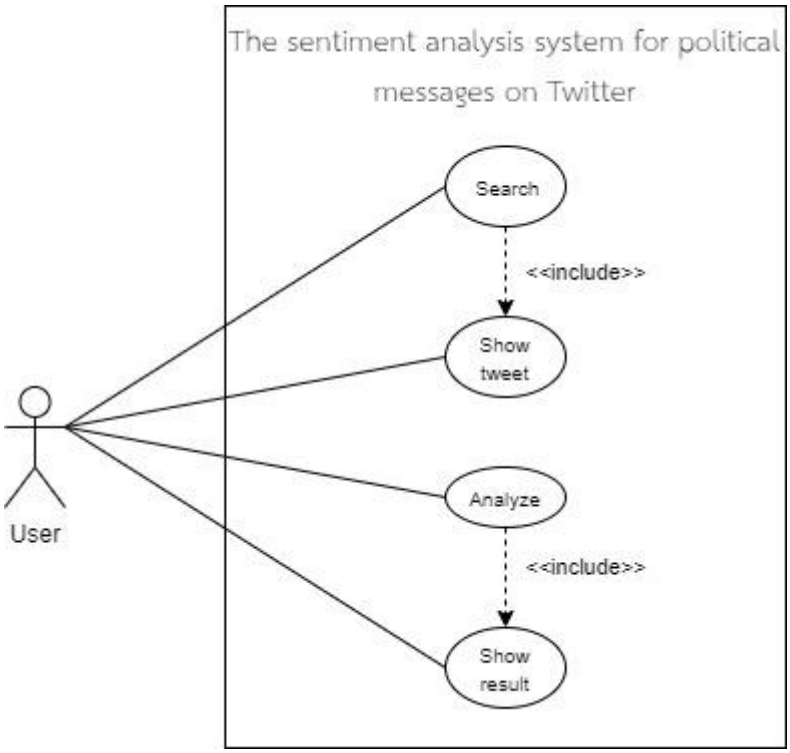
การทำงานของเว็บแอปพลิเคชันระบบวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์ประกอบด้วย การรับคำค้นหาจากผู้ใช้ การนำคำค้นหาไปดึงข้อมูลจากข้อความบนทวิตเตอร์ การนำข้อความที่ดึงมาไปบันทึกไว้ในฐานข้อมูล การนำข้อมูลจากฐานข้อมูลไปวิเคราะห์ด้วยโมเดลจำแนกข้อความที่สร้างขึ้น และนำผลการวิเคราะห์ที่ได้แสดงผลแก่ผู้ใช้ ดังภาพที่ 5.1

ภาพที่ 5.1 แสดงภาพรวมการทำงานของเว็บแอปพลิเคชันระบบวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์ ระบบเริ่มจากการรับคำค้นหาจากผู้ใช้ ระบบจะดึงข้อมูลจากทวิตเตอร์เฉพาะข้อความที่มีแฮชแท็กตรงกับคำค้นหา เนื่องจากการค้นหาข้อมูลตามแฮชแท็กจะทำให้ผู้ใช้ได้รับข้อมูลที่ตรงกับความสนใจมากกว่าการที่ไม่ได้ใช้แฮชแท็กซึ่งอาจมีข้อมูลเรื่องอื่น ๆ ที่ผู้ใช้ไม่สนใจรวมอยู่ด้วย (ในการใส่ข้อมูลคำค้นหา ผู้ใช้ไม่จำเป็นต้องเขียนแฮชแท็ก เพราะระบบจะมีการเพิ่มแฮชแท็กก่อนที่จะดึงข้อมูลจากทวิตเตอร์) ระบบจะบันทึกข้อความที่ดึงมาได้ลงในฐานข้อมูล ต่อมาจะมีการเรียกดึงข้อมูลจากฐานข้อมูล ซึ่งก็คือ ข้อความจากทวิตเตอร์ไปวิเคราะห์ด้วยโมเดลจำแนกข้อความ และแสดงผลการวิเคราะห์ข้อความที่ได้ให้แก่ผู้ใช้ (การดึงข้อมูลจากทวิตเตอร์อธิบายในบทที่ 3 และรายละเอียดโมเดลการจำแนกข้อความอธิบายในบทที่ 4)



ภาพที่ 5.1 ภาพรวมการทำงานของเว็บแอปพลิเคชันระบบวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์

5.2 แผนภาพยูสเคส



ภาพที่ 5.2 แผนภาพยูสเคสของเว็บแอปพลิเคชันวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์

ตารางที่ 5.1 คำอธิบายยูสเคส Search

Use case name	ค้นหาและดึงข้อความจากทวิเตอร์ (Search)
Participating actor	ผู้ใช้งานเว็บแอปพลิเคชัน (User)
Entry condition	เปิดเว็บแอปพลิเคชัน
Flow of events	1. ผู้ใช้พิมพ์คำค้นหาลงสู่หน้าเว็บ 2. ผู้ใช้กดปุ่มค้นหา 3. ระบบดึงข้อความจากฐานข้อมูลทวิตเตอร์ ตามแฮชแท็กคำค้นหา (ระบบจะดึงข้อความได้ไม่เกิน 100 ทวีต) 4. ระบบนำข้อความที่ดึงได้มาเก็บลงฐานข้อมูลที่สร้างไว้
Exit condition	บันทึกข้อความลงฐานข้อมูลเสร็จสิ้น

ตารางที่ 5.2 คำอธิบายยูสเคส Show tweet

Use case name	แสดงผลลัพธ์การค้นหา (Show tweet)
Participating actor	ผู้ใช้งานเว็บแอปพลิเคชัน (User)
Entry condition	ทำงานต่อจากยูสเคส Search
Flow of events	1. ระบบจะแสดงข้อความทวีตที่เก็บไว้ในฐานข้อมูล
Exit condition	ผู้ใช้งานเว็บแอปพลิเคชันเลือกได้ 2 กรณี - ผู้ใช้เลือกพิมพ์คำค้นหาใหม่ และ กดค้นหาอีกครั้ง ระบบจะแสดงข้อความทวีตที่รวบรวมจากทวิตเตอร์ใหม่ - ผู้ใช้กดวิเคราะห์ข้อมูลเพื่อนำข้อความไปวิเคราะห์

ตารางที่ 5.3 คำอธิบายยูสเคส Analyze

Use case name	วิเคราะห์ข้อความ (Analyze)
Participating actor	ผู้ใช้งานเว็บแอปพลิเคชัน (User)
Entry condition	ผู้ใช้งานกดวิเคราะห์
Flow of events	1. ระบบเรียกข้อมูลข้อความจากฐานข้อมูลเพื่อประมวลผล 2. ระบบทำความสะอาดข้อมูล (เช่น ลบลิงก์เว็บไซต์ ลิงก์รูปภาพ และลบตัวอักขระพิเศษที่ไม่สามารถวิเคราะห์ได้) 3. ระบบนำข้อความมาตัดคำ และกำหนดค่าตัวเลขให้กับคำแต่ละคำเพื่อใช้ในการวิเคราะห์ 4. ระบบสร้างชุดข้อมูลทดสอบเพื่อนำไปวิเคราะห์ด้วยโมเดลจำแนกข้อความ 5. โมเดลจำแนกข้อความตอบผลการวิเคราะห์ที่ได้ 6. ระบบเก็บผลการวิเคราะห์ลงในฐานข้อมูล
Exit condition	บันทึกผลการวิเคราะห์ลงฐานข้อมูลเสร็จสิ้น

ตารางที่ 5.4 คำอธิบายยูสเคส Show result

Use case name	แสดงผลลัพธ์การวิเคราะห์ (Show result)
Participating actor	ผู้ใช้งานเว็บแอปพลิเคชัน (User)
Entry condition	ทำงานต่อจากยูสเคส Analyze
Flow of events	1. ระบบจะดึงข้อมูลจากฐานข้อมูลไปประมวลผลเป็นแผนภูมิวงกลม 2. ระบบแสดงผลแผนภูมิวงกลม
Exit condition	ปิดเว็บแอปพลิเคชันหรือผู้ใช้เริ่มการค้นหาใหม่

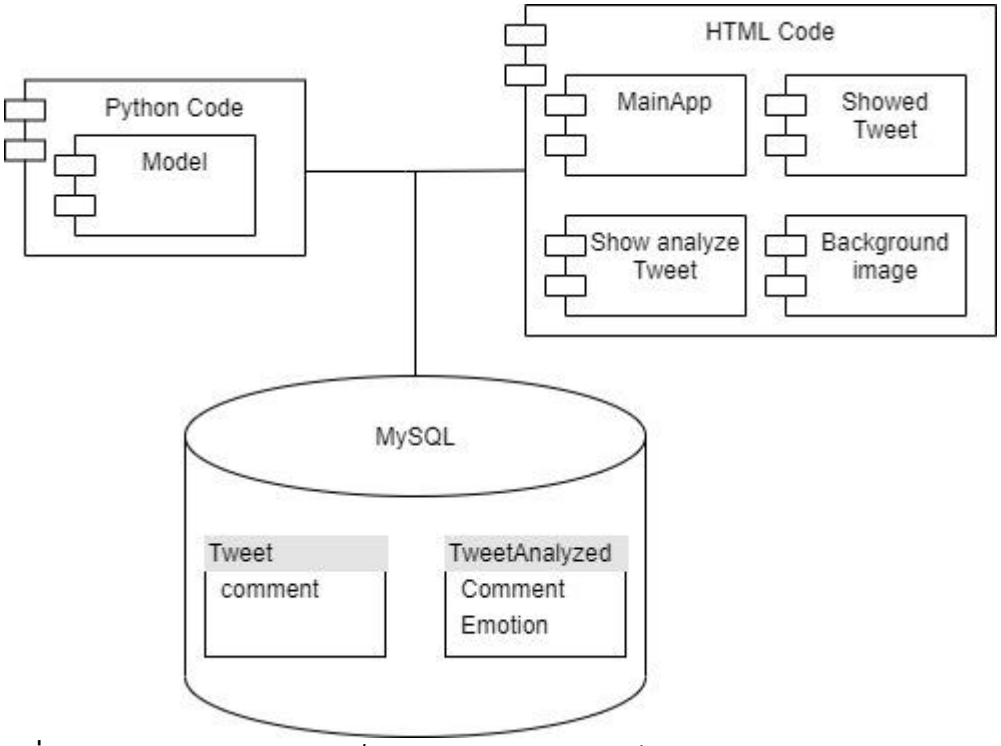
5.3 มอดูลการทำงานของเว็บแอปพลิเคชัน

มอดูลการทำงานของเว็บอปพลิเคชันประกอบการทำงานทั้งหมด 3 ส่วน คือ

1. Python Code เป็นการทำงานของระบบแอปพลิเคชัน มีส่วนประกอบข้างในเป็นมอดูลที่ชื่อว่า Model ทำงานสำหรับการวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็น

2. HTML Code เป็นส่วนแสดงผลของเว็บแอปพลิเคชัน ประกอบด้วย MainApp คือส่วนหน้าหลักของเว็บแอปพลิเคชัน Showed Tweet คือส่วนหน้าแสดงผลข้อความที่ดึงมาจากทวิตเตอร์ Show analyze Tweet คือส่วนแสดงผลการวิเคราะห์ข้อความจากทวิตเตอร์ และ Background image คือส่วนที่เก็บภาพที่ใช้แสดงผลในส่วนต่าง ๆ ของเว็บแอปพลิเคชัน

3. MySQL เป็นส่วนฐานข้อมูลของระบบซึ่งประกอบด้วย Tweet เป็นฐานข้อมูลที่ไว้เก็บข้อความที่ดึงจากทวิตเตอร์ โดยกำหนดไว้ว่า จะมี 1 หลัก ซึ่งเก็บข้อความที่ดึงจากทวิตเตอร์ จำนวนไม่เกิน 100 ข้อความ และ TweetAnalyzed เป็นฐานข้อมูลที่ไว้จัดเก็บข้อมูลที่ได้รับวิเคราะห์แล้ว โดยกำหนดไว้ว่า จะมี 2 หลัก โดยหลักแรกจะเก็บข้อความที่ดึงจากทวิตเตอร์ หลักที่สองจะเก็บความรู้สึกของข้อความ



ภาพที่ 5.2 มอดูลการทำงานของเว็บแอปพลิเคชันวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์

5.4 ภาษาและโปรแกรมที่ใช้ในการพัฒนาระบบ

ภาษาหลักที่ใช้ในการพัฒนาโปรแกรมประกอบด้วย 2 ส่วน

1. ส่วนแสดงหน้าเว็บแอปพลิเคชัน จะใช้ภาษาเอชทีเอ็มแอล5 ซีเอสเอส และจาวาสคริปต์ ในการแสดงผล

2. ส่วนระบบเว็บแอปพลิเคชัน จะใช้ภาษาไพทอนในการเขียนระบบของเว็บแอปพลิเคชัน

ส่วนของไลบรารีที่ใช้พัฒนาเว็บแอปพลิเคชันอธิบายแล้วในบทที่ 2 หัวข้อ 2.6

โปรแกรมที่ใช้ในการพัฒนาระบบมีดังนี้

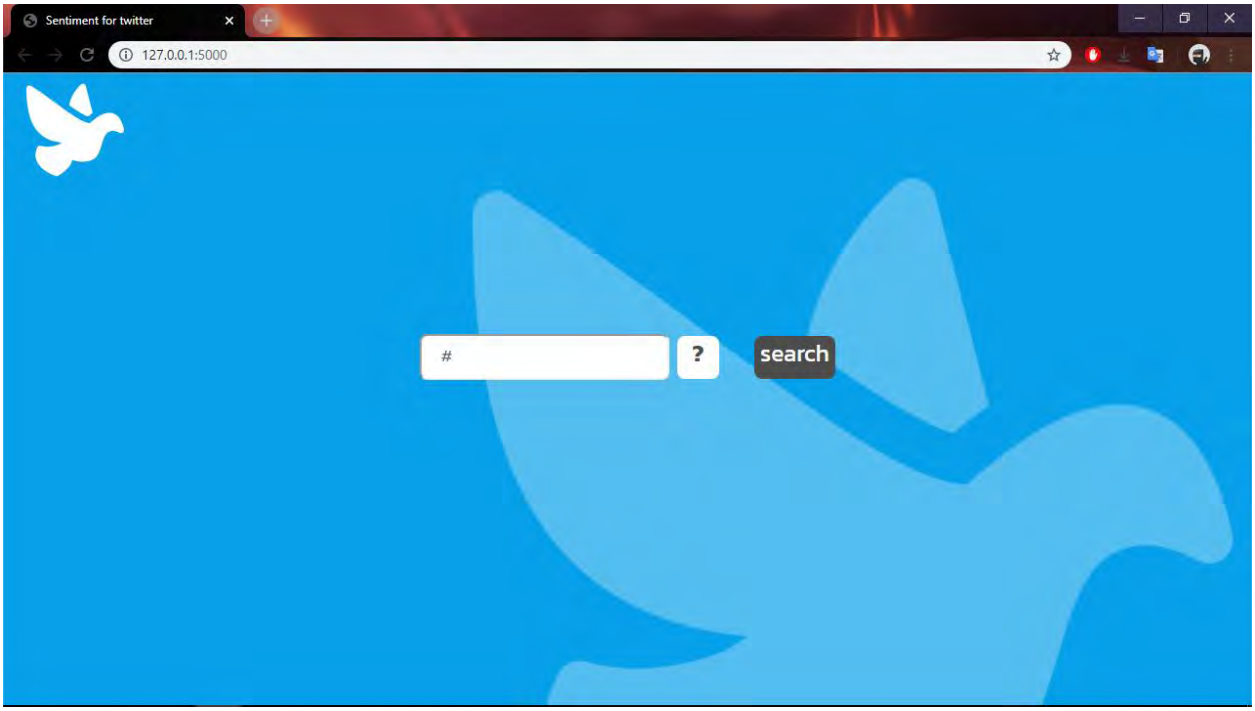
1. โปรแกรม Visual Studio Code เป็นโปรแกรมที่ใช้ในการเขียนภาษาเอชทีเอ็มแอล5 ซีเอสเอส จาวาสคริปต์ และ ไพทอน

2. โปรแกรม XAMPP ใช้ในการสร้างเซิร์ฟเวอร์จำลองเพื่อใช้ในการทดสอบระบบ

3. Draw.io ใช้ในการเขียนแผนภาพการทำงานของเว็บแอปพลิเคชัน

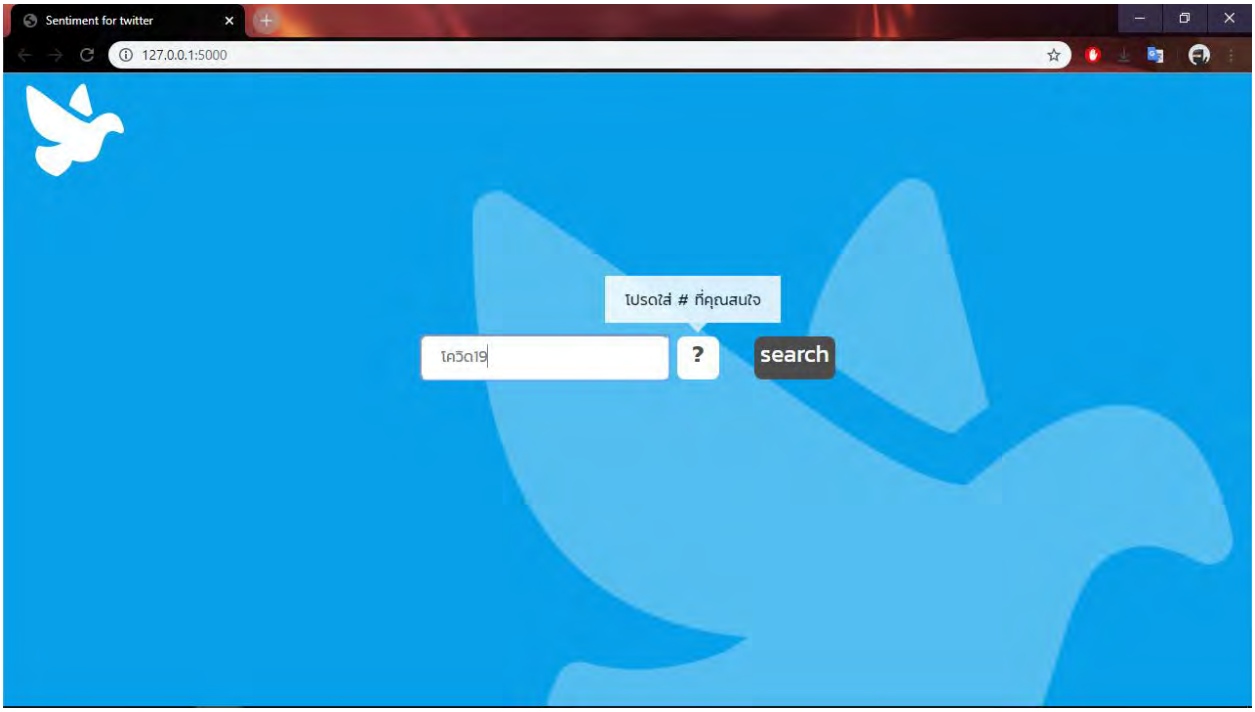
5.5 ส่วนต่อประสานผู้ใช้ (User Interface)

ในการออกแบบส่วนต่อประสานผู้ใช้จะอยู่ในแบบที่เข้าใจง่าย ใช้ได้สะดวกสบายมีการแสดงผลที่ชัดเจน ตัวหนังสือไม่ใหญ่ หรือเล็กไปจนดูไม่ดี และมีความสวยงามและสบายตา



ภาพที่ 5.3 หน้าหลักเว็บแอปพลิเคชัน

หน้าหลักเว็บแอปพลิเคชัน ดังภาพที่ 5.3 จะประกอบด้วย ช่องสำหรับใส่คำค้นหาจากผู้ใช้ซึ่งผู้ใช้สามารถใส่คำค้นหาเป็นแฮชแท็กของทวิตเตอร์ที่ผู้ใช้สนใจ โดยไม่ต้องใส่เครื่องหมายแฮชแท็ก ส่วนต่อไปคือเครื่องหมายปรัศนีจะเป็นคำอธิบายการใส่คำค้นหา เมื่อเอาเม้าส์ไปวางจะมีข้อความขึ้นมาว่า “โปรดใส่ # ที่คุณสนใจ” ดังภาพที่ 5.4 ส่วนสุดท้าย คือ ปุ่มค้นหาจะเป็นการค้นหาข้อความแสดงความคิดเห็นจากทวิตเตอร์

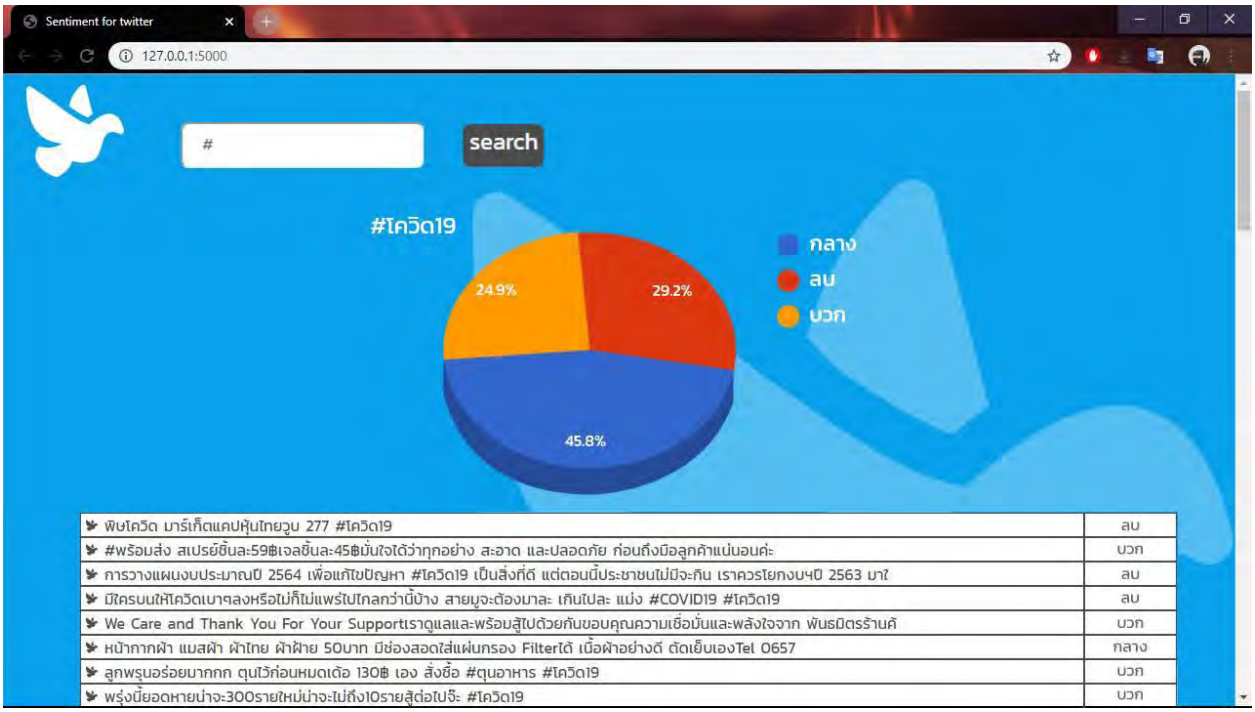


ภาพที่ 5.4 หน้าหลักเว็บแอปพลิเคชัน พร้อมการทำงานส่วนต่าง ๆ

เมื่อผู้ใช้กดปุ่มค้นหา จะแสดงข้อความที่ดึงได้จากทวิตเตอร์แบบเรียลไทม์ว่ามีผู้ใช้ทวิตเตอร์ส่งข้อความทวีตใดอยู่บ้าง หรือถ้าผู้ใช้ต้องการจะใส่คำค้นหาทวิตเตอร์ใหม่ก็สามารถใส่คำค้นหาใหม่และกดค้นหาอีกครั้ง ถ้าผู้ใช้ต้องการดูว่าในข้อความคิดเห็นเหล่านี้มีมีความรู้สึกเป็นอย่างไร สามารถกดปุ่มเพื่อวิเคราะห์ข้อความความแสดงคิดเห็นดังภาพที่ 5.5

ภาพที่ 5.5 ผลลัพธ์การดึงข้อความที่ได้จากทวิตเตอร์

เมื่อผู้ใช้กดปุ่มวิเคราะห์ จะแสดงผลการวิเคราะห์ โดยประกอบด้วย แผนภูมิวงกลมซึ่งจะบอกสัดส่วนของความรู้สึกว่าข้อความนั้นมีความรู้สึกอย่างไร ส่วนด้านล่างจะเป็นข้อความที่ดึงจากทวิตเตอร์พร้อมกับความรู้สึกของข้อความ ดังภาพที่ 5.6



ภาพที่ 5.6 ผลลัพธ์การวิเคราะห์ข้อความด้วยโมเดลจำแนกข้อความ

บทที่ 6
การทดสอบระบบ

6.1 บทนำ

ในบทนี้จะกล่าวถึง ผลการทดสอบโมเดลการวิเคราะห์ข้อความแสดงความคิดเห็นด้านการเมือง ซึ่งผู้พัฒนาจะทดสอบการจำแนกข้อความแสดงความคิดเห็นที่แสดงความรู้สึกที่เป็นบวก ความรู้สึกที่เป็นลบ และความรู้สึกที่เป็นกลาง ว่าประสิทธิภาพในการวิเคราะห์ความรู้สึกของข้อความมากน้อยเพียงใด โดยพิจารณาจากค่าความถูกต้อง ค่าความแม่นยำ และค่าเรียกคืนในการจำแนกความรู้สึกของข้อความ โดยคิดเป็นเปอร์เซ็นต์ รายละเอียดดังต่อไปนี้

การทดสอบมีการแบ่งข้อมูลเป็น 2 ชุด คือ ชุดการเรียนรู้และชุดสอบโดยการวัดประสิทธิภาพของโมเดลการจำแนกความรู้สึกของข้อความ จะอาศัยคอนฟิวชันเมทริกซ์ (confusion matrix) ในการพิจารณาจำนวนของข้อมูลที่เป็นคลาสข้อมูลจริง และคลาสข้อมูลที่ได้จากการจำแนกด้วยระบบโดยอัตโนมัติ ซึ่งตารางจะมีขนาด m x m โดยที่ m คือ จำนวนของคลาสที่ต้องการจำแนก ในกรณีตัวอย่างนี้จะมีจำนวน 3 คลาส คือ 1. คลาสข้อความความรู้สึกที่เป็นบวก 2. คลาสข้อความความรู้สึกที่เป็นลบ และ 3. คลาสข้อความความรู้สึกที่เป็นกลาง คอนฟิวชันเมทริกซ์แสดงในตารางที่ 6.1

ตารางที่ 6.1 คอนฟิวชันเมทริกซ์สำหรับการจำแนกข้อมูล

		กลุ่มข้อความที่โมเดลทำนาย		
		กลาง	บวก	ลบ
กลุ่มจริงของข้อความ	กลาง	ก	ข	ค
	บวก	ง	จ	ฉ
	ลบ	ช	ซ	ฌ

ความหมายของค่าต่าง ๆ ในตาราง มีดังนี้

ก. จำนวนข้อความที่โมเดลทำนายได้เป็นกลางและกลุ่มจริงของข้อความเป็นกลาง

ข. จำนวนข้อความที่โมเดลทำนายได้เป็นบวกแต่กลุ่มจริงของข้อความเป็นกลาง

ค. จำนวนข้อความที่โมเดลทำนายได้เป็นลบแต่กลุ่มจริงของข้อความเป็นกลาง

ง. จำนวนข้อความที่โมเดลทำนายได้เป็นกลางแต่กลุ่มจริงของข้อความเป็นบวก

จ. จำนวนข้อความที่โมเดลทำนายได้เป็นบวกและกลุ่มจริงของข้อความเป็นบวก

ฉ. จำนวนข้อความที่โมเดลทำนายได้เป็นลบแต่กลุ่มจริงของข้อความเป็นบวก

ช. จำนวนข้อความที่โมเดลทำนายได้เป็นกลางแต่กลุ่มจริงของข้อความเป็นลบ

ซ. จำนวนข้อความที่โมเดลทำนายได้บวกแต่กลุ่มจริงของข้อความเป็นลบ

ฌ. จำนวนข้อความที่โมเดลทำนายได้เป็นลบและกลุ่มจริงของข้อความเป็นลบ

ค่าความถูกต้องทั้งหมด = (ก+จ+ฌ) / (ก+ข+ค+ง+จ+ฉ+ช+ซ+ฌ) * 100%

ค่าความแม่นยำข้อความเป็นกลาง = ก / (ก+ง+ช) * 100%

ค่าเรียกคืนข้อความเป็นกลาง = ก / (ก+ข+ค) * 100%

ค่าความแม่นยำข้อความที่เป็นบวก = จ / (จ+ข+ซ) * 100%

ค่าเรียกคืนข้อความที่เป็นบวก = จ / (ง+จ+ฉ) * 100%

ค่าความแม่นยำข้อความที่เป็นลบ = ฌ / (ค+ฉ+ ฌ) * 100%

ค่าเรียกคืนข้อความที่เป็นลบ = ฌ / (ช+ซ+ ฌ) * 100%

6.2 การทดลองเพื่อเลือกเทคนิคการเรียนรู้ของเครื่องในการจำแนกข้อมูล

โครงงานนี้มีการทดลองใช้เทคนิคการจำแนกความรู้สึกของข้อความ 2 วิธี ได้แก่ การถดถอยโลจิสติกส์ และยูแอลเอ็มฟิตโมเดล โดยข้อความทั้งหมด 10,085 ข้อความ แบ่งเป็นข้อมูลสำหรับการเรียนรู้ 8,572 ข้อความ และข้อมูลสำหรับการทดสอบ 1,513 ข้อความ อัตราส่วน 85 : 15 ซึ่งได้ผลลัพธ์ดังนี้

6.2.1 การจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์

คอนฟิวชันเมทริกซ์และประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์แสดงในตารางที่ 6.2 และตารางที่ 6.3 ตามลำดับ

ตารางที่ 6.2 คอนฟิวชันเมทริกซ์ผลการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์

		กลุ่มข้อความที่โมเดลทำนาย		
		กลาง	บวก	ลบ
กลุ่มจริงของข้อความ	กลาง	461	47	137
	บวก	89	157	57
	ลบ	109	33	423

ตารางที่ 6.3 ประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์

กลุ่มข้อความ	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าเรียกคืน
กลาง	68.80%	0.70	0.71
บวก		0.66	0.52
ลบ		0.69	0.75

จากตารางที่ 6.3 พบว่าการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์ให้ค่าความแม่นยำสูงสุดคือ 0.70 สำหรับกลุ่มข้อความที่เป็นกลาง และค่าความแม่มยำน้อยสุดสำหรับกลุ่มข้อความที่เป็นบวก คือ 0.66 เมื่อเปรียบเทียบกลุ่มข้อความทั้งหมด ค่าความแม่นยำในการจำแนกไม่ได้ต่างกันมาก ในส่วนของค่าเรียกคืนพบว่า ค่าเรียกคืนสูงสุดอยู่กลุ่มข้อความที่เป็นลบ คือ 0.75 และค่าเรียกคืนน้อยสุดอยู่กลุ่มข้อความที่เป็นบวก คือ 0.52 โดยค่าความแม่นยำน้อยสุดและค่าเรียกคืนน้อยสุดอยู่กลุ่มข้อความที่เป็นบวก

6.2.2 การจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิต (ULMFIT model)

คอนฟิวชันเมทริกซ์และประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิต แสดงในตารางที่ 6.4 และตารางที่ 6.5 ตามลำดับ

ตารางที่ 6.4 คอนฟิวชันเมทริกซ์ผลการจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิต

		กลุ่มข้อความที่โมเดลทำนาย		
		กลาง	บวก	ลบ
กลุ่มจริงของข้อความ	กลาง	481	44	114
	บวก	64	181	53
	ลบ	109	46	422

ตารางที่ 6.5 ประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิต

กลุ่มข้อความ	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าเรียกคืน
กลาง	71.64%	0.74	0.75
บวก		0.67	0.61
ลบ		0.72	0.73

จากตารางที่ 6.5 พบว่า การจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิตมีค่าความแม่นยำสูงสุดคือกลุ่มข้อความที่เป็นกลางอยู่ที่ 0.74 และค่าความแม่นยำน้อยสุดอยู่ที่กลุ่มข้อความที่เป็นบวก คือ 0.67 เมื่อเปรียบเทียบกลุ่มข้อความทั้งหมด ค่าความแม่นยำในการจำแนกไม่ได้ต่างกันมาก ในส่วนของค่าเรียกคืนพบว่า ค่าเรียกคืนสูงสุดอยู่กลุ่มข้อความที่เป็นกลาง คือ 0.75 ซึ่งเป็นกลุ่มข้อความเดียวกับค่าความแม่นยำมากสุด และค่าเรียกคืนน้อยสุดอยู่ในกลุ่มข้อความที่เป็นบวก คือ 0.61 ดังนั้น การจำแนกความรู้สึกของข้อความด้วยยูแอลเอ็มฟิตให้ประสิทธิภาพสูงสุดในกลุ่มข้อความที่เป็นกลาง และประสิทธิภาพต่ำสุดในกลุ่มข้อความที่เป็นบวก

6.3 การอภิปรายผลการทดสอบ

จะเห็นได้ว่าการจำแนกความรู้สึกของข้อความด้วยโมเดลการจำแนกข้อมูลด้วยการถดถอยโลจิสติกส์และโมเดลการจำแนกข้อมูลด้วยยูแอลเอ็มฟิต ค่าความแม่นยำและค่าเรียกคืนนั้นมีคะแนนใกล้เคียงกัน แต่ค่าเรียกคืนในส่วนโมเดลการจำแนกข้อมูลด้วยยูแอลเอ็มฟิตของข้อความที่เป็นนวกจำแนกออกมาได้ดีกว่า ซึ่งมีค่าเท่ากับ 0.61 จึงทำให้ค่าความถูกต้องของการจำแนกข้อมูลด้วยยูแอลเอ็มฟิตอยู่ที่ 71.64% ซึ่งดีกว่าการถดถอยโลจิสติกส์ และจะเห็นได้ว่าค่าการเรียกคืนของกลุ่มข้อความด้านบวกที่จำแนกด้วยการถดถอยโลจิสติกส์ออกมาได้ประสิทธิภาพน้อยสุดอยู่ที่ 0.52 โอกาสที่โมเดลสามารถจำแนกข้อมูลด้านบวกได้ครบถ้วนมีเพียงครึ่งเดียวเท่านั้น อาจจะเป็นเพราะมีจำนวนข้อความที่เป็นบวกน้อยที่สุดในการเรียนรู้ของเครื่อง เมื่อใช้อัลกอริทึมการถดถอยโลจิสติกส์ในการจำแนกจึงได้ประสิทธิภาพน้อยกว่ายูแอลเอ็มฟิตที่มีการเรียนรู้โมเดลภาษาก่อนการจำแนกความรู้สึกของข้อความด้วยเหตุผลดังกล่าวทางผู้พัฒนาจึงเลือกโมเดลยูแอลเอ็มฟิตในการพัฒนาระบบเว็บแอปพลิเคชัน

6.4 การทดสอบเว็บแอปพลิเคชัน

1. การทดสอบในการใส่คำค้นหา เนื่องจากโมเดลที่พัฒนาสามารถใช้ได้แต่ภาษาไทย ดังนั้นคำค้นหาควรเป็นภาษาไทย และมีการทดสอบดึงข้อมูลจากทวิตเตอร์จาก 3 แฮชแท็กที่เหมือนกับแฮชแท็กที่ใช้ในการทดสอบกับโมเดล และ 2 แฮชแท็ก ที่เป็นแฮชแท็กที่ได้รับความนิยม ณ ตอนนี้ ดังต่อไปนี้

ประชุมสภา ผลลัพธ์จะแสดงดังภาพที่ 6.1

ภาพที่ 6.1 ผลการค้นหาของ #ประชุมสภา

พลังประชารัฐ ผลลัพธ์จะแสดงดังภาพที่ 6.2

ภาพที่ 6.2 แสดงผลการค้นหาของ #พลังประชารัฐ

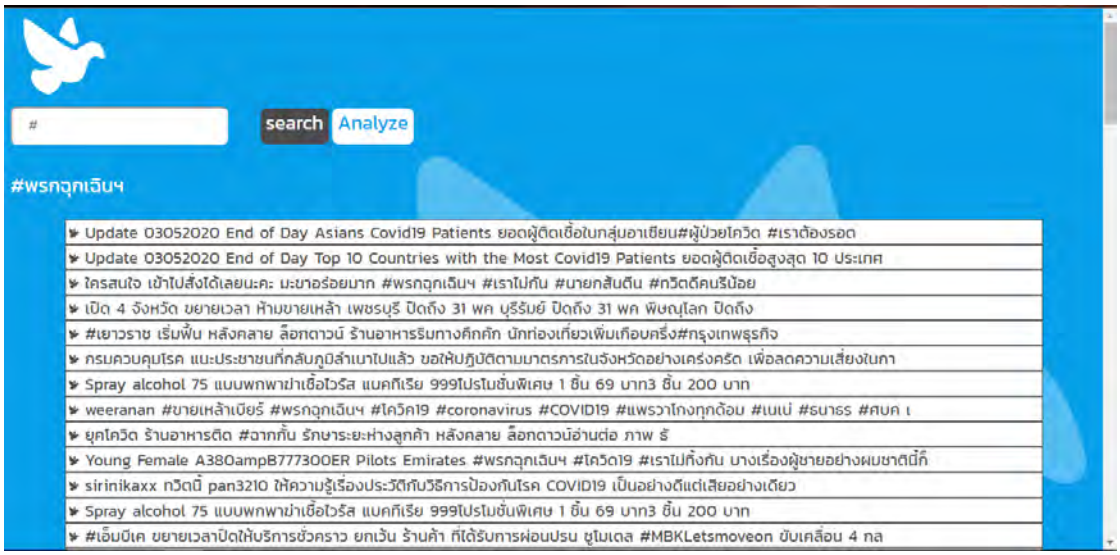
เพื่อไทย ผลลัพธ์จะแสดงดังภาพที่ 6.3

ภาพที่ 6.3 แสดงผลการค้นหาของ #เพื่อไทย

โควิด19 ผลลัพธ์จะแสดงดังภาพที่ 6.4

ภาพที่ 6.4 แสดงผลการค้นหาของ #โควิด19

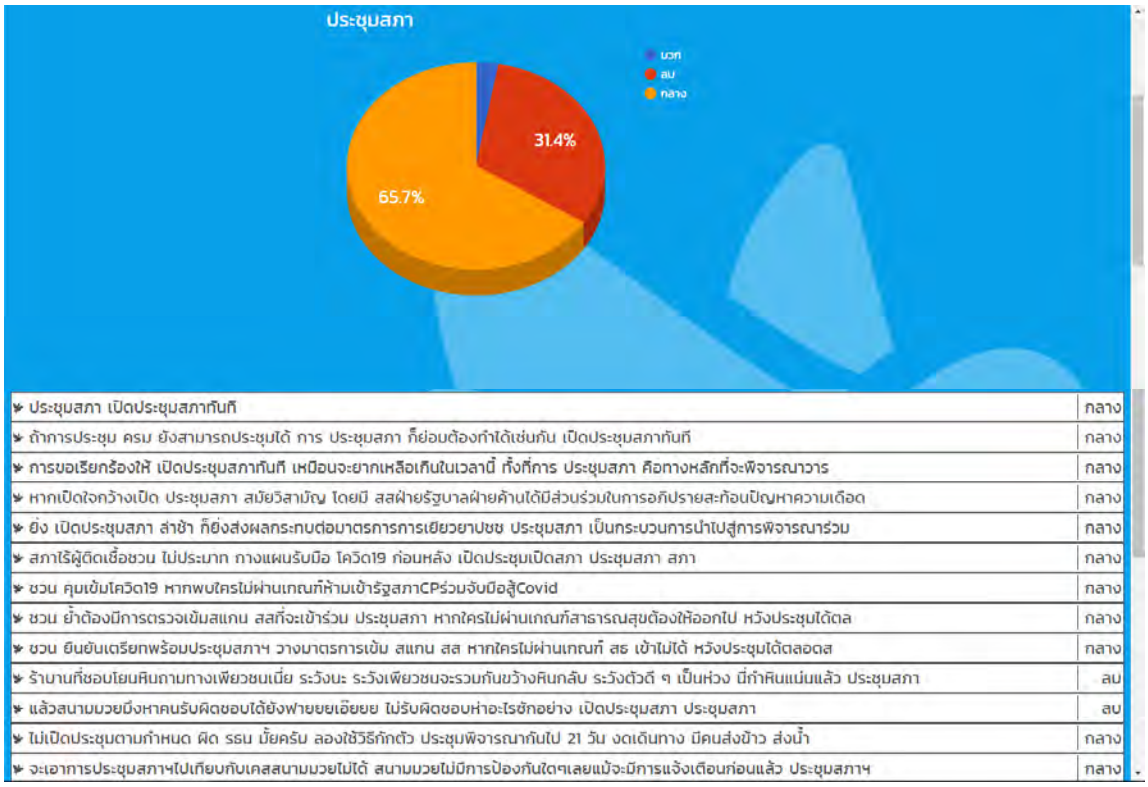
พรกฉุกเฉินฯ ผลลัพธ์จะแสดงดังภาพที่ 6.5



ภาพที่ 6.5 แสดงผลการค้นหาของ #พรกฉุกเฉินฯ

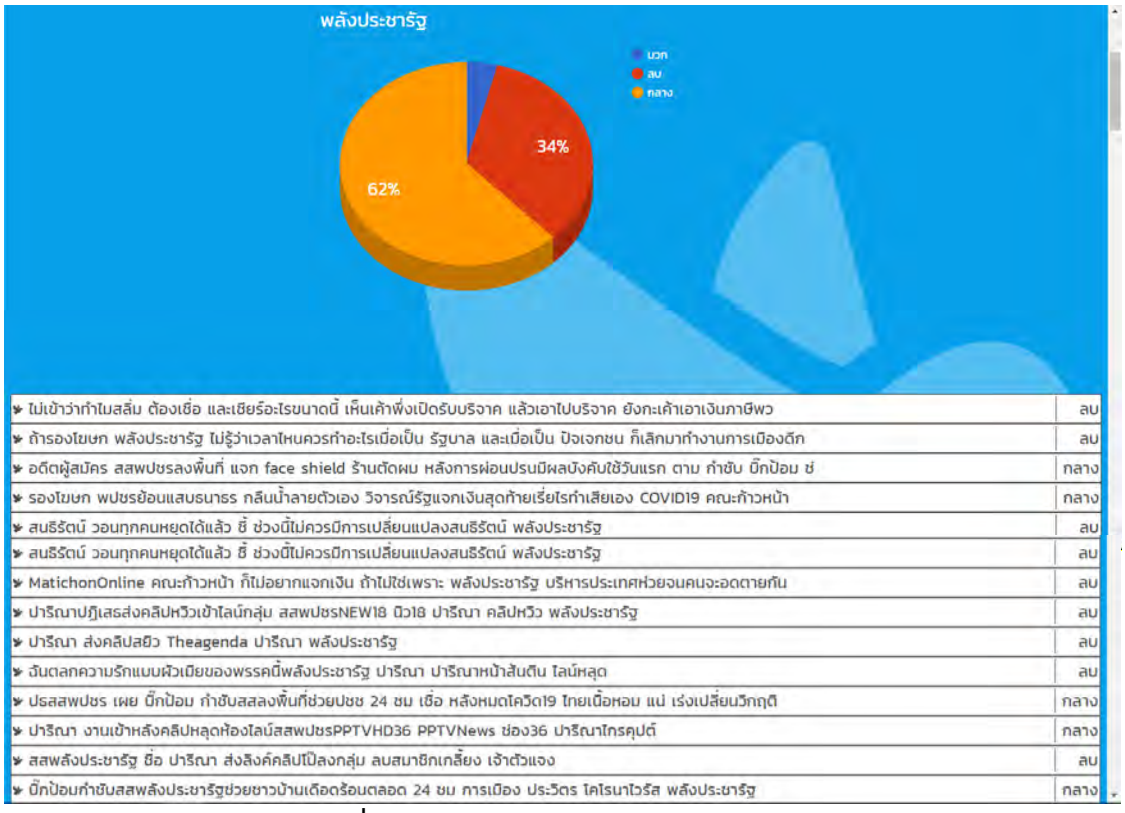
2. การทดสอบสามารถนำข้อมูลที่ดึงจากทวิตเตอร์ไปวิเคราะห์ด้วยโมเดลจำแนกข้อความแล้วได้ผลเป็นอย่างไร ผลปรากฏว่าสามารถวิเคราะห์ความรู้สึกข้อความ โดยแยกเป็นข้อความด้านบวก ด้านลบ และกลาง พร้อมทั้งแสดงแผนภูมิวงกลม ได้ผลลัพธ์ดังต่อไปนี้

ประชุมสภา ผลลัพธ์จะแสดงดังภาพที่ 6.6



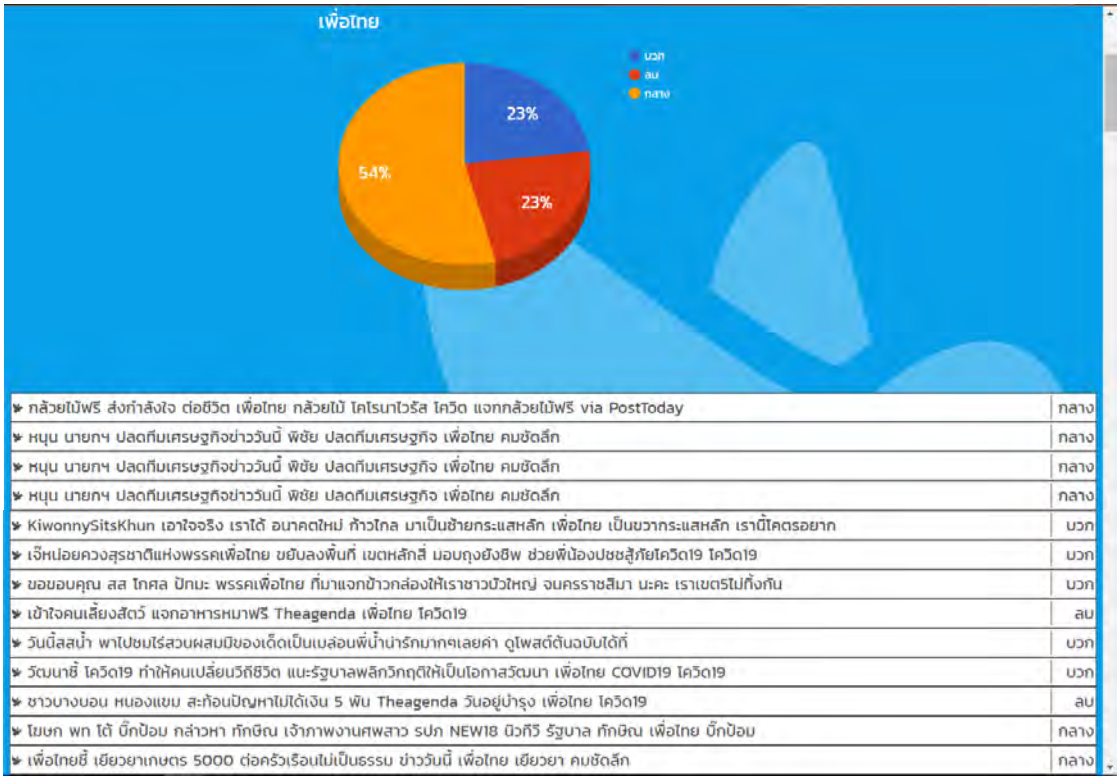
ภาพที่ 6.6 ผลการวิเคราะห์ของ #ประชุมสภา

พลังประชารัฐ ผลลัพธ์จะแสดงดังภาพที่ 6.7



ภาพที่ 6.7 ผลการวิเคราะห์ของ #พลังประชารัฐ

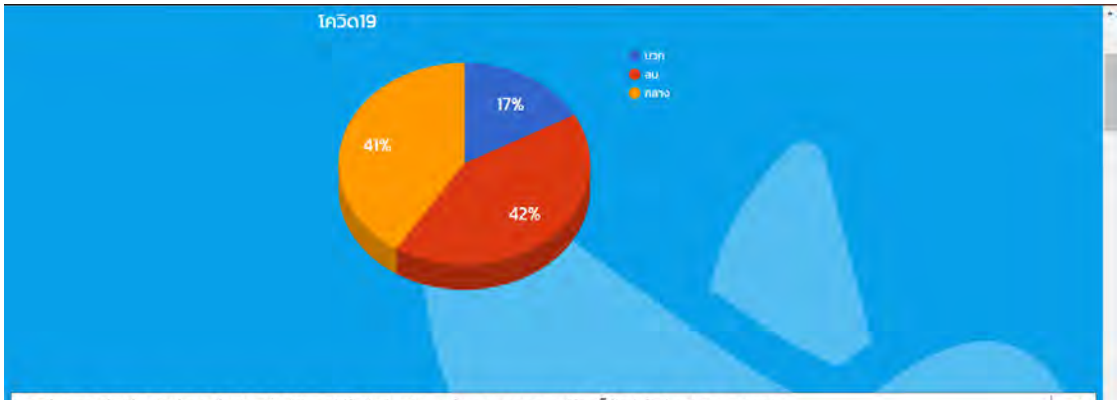
เพื่อไทย ผลลัพธ์จะแสดงดังภาพที่ 6.8



กล้วยไม้ฟรี ส่งกำลังใจ ต่อชีวิต เพื่อไทย กล้วยไม้ โคโรนาไวรัส โควิด แจกกล้วยไม้ฟรี via PostToday	กลาง
หนุน นายกฯ ปลดทีมเศรษฐกิจข่าววันนี้ พิชัย ปลดทีมเศรษฐกิจ เพื่อไทย คมชัดลึก	กลาง
หนุน นายกฯ ปลดทีมเศรษฐกิจข่าววันนี้ พิชัย ปลดทีมเศรษฐกิจ เพื่อไทย คมชัดลึก	กลาง
หนุน นายกฯ ปลดทีมเศรษฐกิจข่าววันนี้ พิชัย ปลดทีมเศรษฐกิจ เพื่อไทย คมชัดลึก	กลาง
KiwonnySitsKhun เอาใจจริง เราได้ อนาคตใหม่ ก้าวไกล มาเป็นช้ายกระแสหลัก เพื่อไทย เป็นขวากระแสหลัก เรานี้โคตรอยาก	บวก
เจ๊หน่อยควงสุรชาติแห่งพรรคเพื่อไทย ขยับลงพื้นที่ เขตหลักสี่ มอบถุงยังชีพ ช่วยพี่น้องปชชสู้ภัยโควิด19 โควิด19	บวก
ขอขอบคุณ สส โกศล ปักษะ พรรคเพื่อไทย ที่มาแจกข้าวกล่องให้เราชาวบัวใหญ่ จมครราชสีมา นะคะ เราเขต5ไม่ทิ้งกัน	บวก
เข้าใจคนเลี้ยงสัตว์ แจกอาหารหมาฟรี Theagenda เพื่อไทย โควิด19	ลบ
วันนี้สสน้ำ พาไปชมไร่สวนผสมมิของเด็ดเป็นเมล่อนพี่น้ำน่ารักมากๆเลยค่า ดูโพสต์ต้นฉบับได้ที่	บวก
วัฒนาชี้ โควิด19 ทำให้คนเปลี่ยนวิถีชีวิต แนะรัฐบาลพลิกวิกฤติให้เป็นโอกาสวัฒนา เพื่อไทย COVID19 โควิด19	บวก
ชาวบางบอน หนองแขม สะท้อนปัญหาไม่ได้เงิน 5 พัน Theagenda วันอยู่บ้ารุง เพื่อไทย โควิด19	ลบ
โฆษก พท โต้ บิ๊กป้อม กล่าวหา ทักษิณ เจ้าภาพงานศพลาว รปภ NEW18 นิวกีวี รัฐบาล ทักษิณ เพื่อไทย บิ๊กป้อม	กลาง
เพื่อไทยชี้ เยียวยาเกษตร 5000 ต่อครัวเรือนไม่เป็นธรรม ข่าววันนี้ เพื่อไทย เยียวยา คมชัดลึก	กลาง

ภาพที่ 6.8 ผลการวิเคราะห์ของ #เพื่อไทย

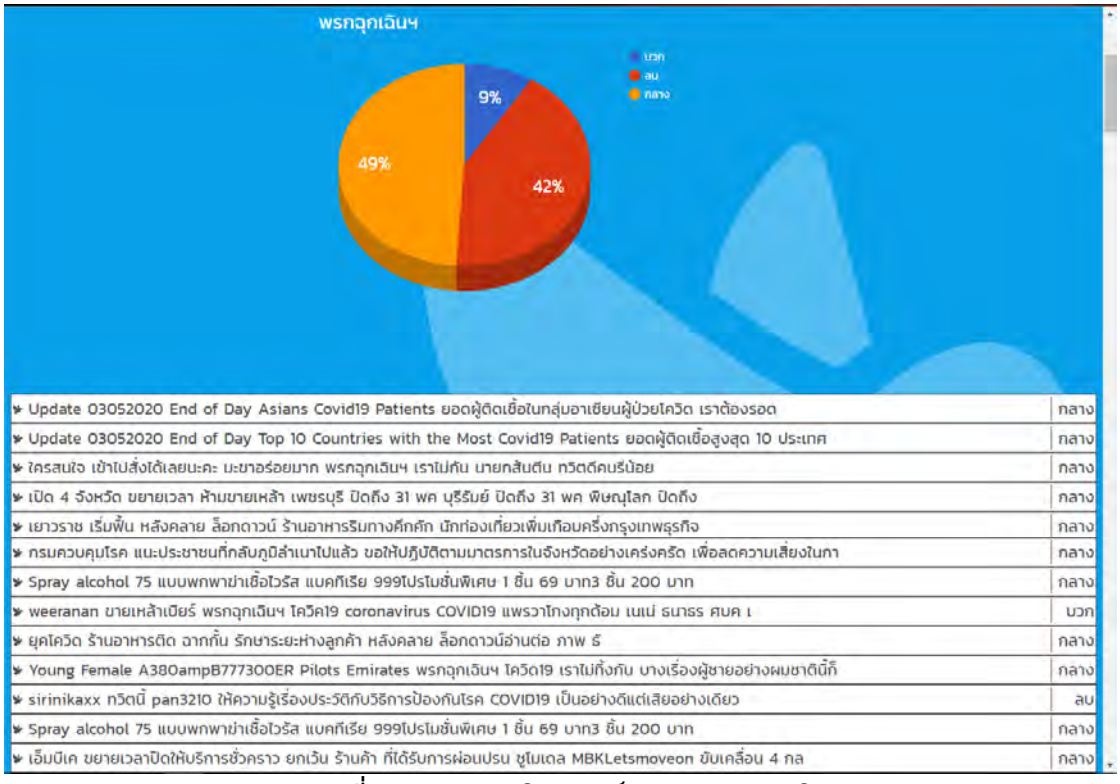
โควิด19 ผลลัพธ์จะแสดงดังภาพที่ 6.9



จะทำกินเองก้อได้ หรือทำขายก้อน่าจะรุ่ง ไม่สองไม่รู้ ดีกว่าอยู่เฉยๆน้าขอบอกกกก อยู่บ้านก็ช้อปได้ ผ่านช่องทางออนไล	ลบ
Update 03052020 End of Day Asians Covid19 Patients ยอดผู้ติดเชื้อในกลุ่มอาเซียนผู้ป่วยโควิด เราต้องรอด	กลาง
Update 03052020 End of Day Top 10 Countries with the Most Covid19 Patients ยอดผู้ติดเชื้อสูงสุด 10 ประเทศ	กลาง
ช่วยทำแบบสอบถามหน่อยนะคะ แจกเงิน 100 บาท 10 รางวัลด้วยยยdek63 dek62 dek61 dek60 dek59 ทีมมธ tcassอบ3 มธ	กลาง
นี่แหละคือทีเครียดจริงๆ ถ้าวัคซีนไม่มาเราจะทำยังไง ต้องอยู่แบบnew normalไปอีกนานแค่ไหน โควิด19	ลบ
คือวันนี้ออกไปซื้อกับข้าวแถวบ้าน เริ่มมีคนไม่ใส่แมสแล้วค่ะ รอบ 2 ล็อคดาวน์ที่นี่นานเลยนะ ประหม่ากันเกินไปหรือเปล่า โควิด19	กลาง
พร้อมส่ง หน้ากากอนามัยโรงงานไทย หนามากไม่จกตา แผ่นกรอง 3 ชั้น กล่องละ 100 ชิ้น ราคา 1150 บาทส่งฟรี สนใจทักค่ะ	กลาง
แค่สองวันเพจ The773Mission ช่วยคนที่กำลังลำบากไปแล้ว 20 ครอบครัว เข้าไปขอความช่วยเหลือ หรือเข้าไปสนับสนุนกันได้นะ	ลบ
ผู้ค้าแผงลอยตลาดลาวคลองเตยขอสู้ไม่สน เคอร์ฟิว ปักหลักประท้วงหลังมีข่าวกทมจ่อรื้อย้ายร้องขออย่าซ้ำเ	บวก
SET A ป้ายคล้องคอตกแต่งห้อง 100 แถมเซ็ตไฟโต้การ์ด SET B ป้ายหัวโตเคลือบ 130 แถมเซ็ตการ์ดค่าส่ง2545	กลาง
หางาน โควิด19 อยู่บ้านช่วงนี้ไม่มีอะไรทำ แนะนำให้ลองมาม่าเผ็ด เผ็ดกว่ามาม่าเกาหลีนะทุกคนนน แม่ค้ากินแล้วววว กระป๋อ	กลาง
5 ชีวิต ไร้งานเพราะโควิด 19 ถูกปัดตก เราไม่ทิ้งกัน สู้เดินเท้า 15 กม หวังไปตายเอาดาบหน้า	ลบ
ไม่ต้องไปอิจฉาคนอื่นหรอกแค่อยู่กับสิ่งที่มีและมีความสุขให้ได้ก็พอไม่มีใครสมบูรณ์แบบที่สุดเราเองก็แค่เป็นคนธรร	ลบ

ภาพที่ 6.9 ผลการวิเคราะห์ของ #โควิด19

พรกฉุกเฉินฯ ผลลัพธ์จะแสดงดังภาพที่ 6.10



ภาพที่ 6.10 ผลการวิเคราะห์ของ #พรกฉุกเฉินฯ

บทที่ 7
สรุปและข้อเสนอแนะ

7.1 สรุปผล

ระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์ โดยวิเคราะห์อารมณ์และความรู้สึกของข้อความแสดงความคิดเห็นออกเป็นด้านบวก,ด้านลบและด้านที่เป็น กลางการทำงานของแต่ละโมเดลประกอบด้วย 3 ส่วนหลักที่เหมือนกัน คือ ส่วนการเตรียมข้อมูล ซึ่งประกอบไปด้วยการทำความสะอาดข้อมูล,การตัดคำ ส่วนการวิเคราะห์ข้อมูล คือ การเลือกคุณลักษณะสำหรับการจำแนก และส่วนการจำแนก จะจำแนกด้วยอัลกอริทึมของแต่ละโมเดล โดยจากการทดสอบพบว่าการจำแนกความรู้สึกของข้อความแสดงความคิดเห็นด้วยโมเดลยูแอลเอ็ม ดีกว่าการถดถอยโลจิสติกส์ อยู่ที่ 71.64% :

เว็บแอปพลิเคชันสามารถดึงข้อมูลจากทวิตเตอร์ได้แบบเรียลไทม์ และกำหนดให้ดึงข้อความไม่เกิน 100 ข้อความ เพราะว่าถ้ามีการดึงข้อมูลมากกว่านี้อาจทำให้การทำงานของระบบเกิดความไม่เสถียรได้ สามารถเชื่อมต่อกับโมเดลของระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์ได้ บอกความรู้สึกของข้อความ และสามารถนำเสนอในแผนภูมิวงกลม

7.2 ผลที่ได้รับ

1. ประโยชน์ต่อผู้พัฒนา

- ผู้พัฒนาได้พัฒนาโมเดลสำหรับการวิเคราะห์ข้อความและอยากพัฒนาโมเดลให้มีประสิทธิภาพที่ดีขึ้น
- ผู้พัฒนาได้พัฒนาทักษะการสร้างระบบตามขั้นตอนวิธีด้านวิศวกรรมซอฟต์แวร์ ซึ่งสามารถนำไปใช้พัฒนาระบบอื่นในอนาคตได้
- ผู้พัฒนาได้เรียนรู้ขั้นตอนการพัฒนาระบบวิเคราะห์ความรู้สึกจากข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์
- ผู้พัฒนาได้เรียนรู้ขั้นตอนการใช้ไลบรารีที่ติดต่อกับฐานข้อมูลขนาดใหญ่ และสามารถนำข้อมูลออกมาใช้ได้อย่างเต็มประสิทธิภาพ
- ผู้พัฒนาได้เรียนรู้การเขียนภาษาไพทอน สามารถประยุกต์ และใช้ไลบรารีมาทำงานร่วมกันได้อย่างชำนาญ ด้านการเขียนเว็บแอปพลิเคชัน ผู้พัฒนาได้เรียนรู้ภาษาเอชทีเอ็มแอล5 ซีเอสเอส และจาวาสคริปต์ ที่ใช้ในการพัฒนาระบบของเว็บแอปพลิเคชันให้นำเสนอในรูปที่สวยงามและดูน่าสนใจยิ่งขึ้น

- ผู้พัฒนาได้เรียนรู้การทำงานอย่างมีระบบ และมีการประยุกต์จากการเรียนที่ได้ศึกษาเรียนรู้ที่ผ่านมา นำมาใช้ในการพัฒนาระบบ ได้เรียนรู้การทำงานเป็นทีม ได้ทำงานร่วมกับผู้อื่น

2. ประโยชน์ต่อผู้ใช้และสังคม

- ข้อมูลที่ได้จากการวิเคราะห์ช่วยให้เห็นแนวโน้มว่า ในปัจจุบันผู้คนในสื่อสังคมออนไลน์มีความคิดเห็นทางการเมืองในปัจจุบันเป็นด้านบวก หรือด้านลบ เกี่ยวข้องกับประเด็นใดบ้าง
- รัฐบาล หน่วยงานที่เกี่ยวข้อง หรือองค์กรที่เกี่ยวข้องกับการเมือง จะสามารถรับรู้ความคิดเห็นจากสื่อสังคมออนไลน์ เป็นไปในด้านบวก หรือด้านลบอย่างรวดเร็ว

7.3 ปัญหาและอุปสรรค

1. การเก็บข้อมูลจากทวิตเตอร์ โดยใช้เอพีไอ ต้องเก็บข้อมูลอย่างสม่ำเสมอ แต่เนื่องจากผู้พัฒนาได้เกิดความผิดพลาด ทำให้ข้อมูลที่เก็บมาได้สูญหาย ทำให้ข้อมูลที่ใช้ไม่เพียงพอ
2. ในการทำเฉลยข้อความ แม้จะประกอบด้วยคน 3 คนในการทำเฉลยข้อความ แต่ก็ยังมีส่วนผิดพลาดที่เกิดขึ้น เช่น การเฉลยที่ตกหล่น หรือการเขียนเฉลยผิด
3. ในการเขียนเว็บแอปพลิเคชันในการเชื่อมโยงระหว่างภาษา HTML กับ ไพทอน ยังมีการติดขัดอยู่บ้าง
4. ด้านการทำงานและการคุยงาน เนื่องจากช่วงเวลาที่พัฒนาโปรแกรมเป็นช่วงที่มีไวรัส COVIC-19 ระบาดอยู่ทำให้การคุยงานไม่สะดวกและมีการเข้าใจที่ไม่ตรงกันหลายประการ ทำให้งานล่าช้าลงไปบ้าง

7.4 วิธีการแก้ปัญหา

1. ในการเก็บข้อมูลต้องขอบคุณ archive.org ที่ได้เก็บข้อมูลทวิตเตอร์ ย้อนหลังเอาไว้และได้ปล่อยออกมาแบบสาธารณะทำให้ผู้พัฒนาสามารถดำเนินการไปกระบวนการถัดไปได้
2. ในการเก็บข้อมูล ผู้พัฒนาต้องมีความตั้งใจ มีสมาธิและความรอบคอบ เพื่อให้ข้อมูลที่ใช้ใน การทดลองมีความผิดพลาดน้อยที่สุด
3. ศึกษาเพิ่มและประยุกต์ใช้กับไลบรารีอื่น ๆ จนสามารถแก้ปัญหาได้
4. ใช้วิดีโอคอลในการทำงาน พึ่งเทคโนโลยีในการติดต่อสื่อสารให้มากขึ้น

เอกสารอ้างอิง

[1] Kusrini Magister and Mochamad Mashuri , “Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication” in proceedings of the 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, pp.365 – 368, March 2019.

[2] S. Akter and M. T. Aziz, “Sentiment Analysis On Facebook Group Using Lexicon Based Approach,” in proceeding 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, pp. 8–11, Sept 2016.

[3] Sheeba Naz,Aditi Sharan and Nidhi Malik, “sentiment classification on twitter data using support vector machine” in proceedings 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, pp.676-679, Dec. 2018.

[4] T. Jhansi rani, K.anuradha and P.vijayapal reddy, “Sentiment Classification on Twitter data using Word N Gram Model” International Journal of Technology and Engineering Science (IJTES) Vol4, pp.7032-7035, 2016.

[5] Pattarawat Chormai ,“AttaCut” [ออนไลน์] แหล่งที่มา : https://pythainlp.github.io/attacut สืบค้นเมื่อ 22 พฤศจิกายน พ.ศ. 2562

[6] Ekkalak Thongthanomkul, “wongnai-corpus” [ออนไลน์] แหล่งที่มา : https://github.com/wongnai/wongnai-corpus สืบค้นเมื่อ 25 พฤศจิกายน พ.ศ. 2562

[7] Wisesight (Thailand) Co., Ltd., “wisesight-sentiment” [ออนไลน์] แหล่งที่มา: https://github.com/PyThaiNLP/wisesight-sentiment สืบค้นเมื่อ 25 พฤศจิกายน พ.ศ. 2562

[8] PyThaiNLP, “PyThaiNLP” [ออนไลน์] แหล่งที่มา: https://www.thainlp.org/pythainlp/tutorials/ สืบค้นเมื่อ 28 พฤศจิกายน พ.ศ. 2562

[9] NLPJeremy Howard, and “Sebastian Ruder”, “ULMFiT” [ออนไลน์] แหล่งที่มา: https://arxiv.org/abs/1801.06146/ สืบค้นเมื่อ 18 ธันวาคม พ.ศ. 2562

ภาคผนวก

ภาคผนวก ก
แบบเสนอหัวข้อโครงงาน รายวิชา 2301399 Project Proposal
ปีการศึกษา 2560

ชื่อโครงงาน (ภาษาไทย)	การวิเคราะห์ข้อความแสดงความคิดเห็นของผู้ใช้บริการต่อร้านอาหารนานาชาติในกรุงเทพมหานคร
ชื่อโครงงาน (ภาษาอังกฤษ)	User comment analysis for the international restaurants in Bangkok
อาจารย์ที่ปรึกษา	อาจารย์ ดร.ภควรรณ ปักษี
ผู้ดำเนินการ	นายจิตรทิวัส แจ้งจันทร์ เลขประจำตัวนิสิต 5933611123
	นายชลวิทย์ ก้อนทอง เลขประจำตัวนิสิต 5933617023
	สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
	คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

หลักการและเหตุผล

เนื่องจากในปัจจุบันมนุษย์ได้ใช้ประโยชน์สื่อสังคมออนไลน์เป็นจำนวนมาก มีการพัฒนาของอินเทอร์เน็ตเกิดขึ้นอย่างรวดเร็ว ทำให้การสื่อสารด้วยข้อความได้รับความนิยมมากขึ้น สื่อสังคมออนไลน์ที่ได้รับความนิยมในปัจจุบัน เช่น เฟซบุ๊ค ทวิตเตอร์ ไลน์ และอินสตราแกรม สื่อเหล่านี้ได้กลายเป็นสถานที่ในการแลกเปลี่ยนความคิดเห็นของผู้คน เป็นแหล่งในการสืบค้นข้อมูล การซื้อขายสินค้า ผู้ใช้สื่อออนไลน์จำนวนมากได้ร่วมสร้างข้อมูลหลากหลายรูปแบบ เช่น ข้อความ รูปภาพ ฯลฯ ทำให้มีข้อมูลจำนวนมหาศาลบนโลกออนไลน์โดยเฉพาะด้านการเมืองที่ได้รับความสนใจอยู่ในปัจจุบัน ซึ่งข้อมูลเหล่านี้สามารถเก็บรวบรวม คัดลอก หรือประมวลผลด้วยคอมพิวเตอร์ได้ เพื่อนำไปใช้ประโยชน์ในด้านการวิเคราะห์ความรู้สึก (sentiment analysis) การทำเหมืองข้อมูล (data mining) หรือการสกัดข้อมูล (data extraction) ได้

ระบบวิเคราะห์ความรู้สึกของข้อความ คือระบบที่สามารถวิเคราะห์ และคัดกรองข้อความได้ว่า มีการแสดงความรู้สึกหรือไม่ และถ้ามีการแสดงความรู้สึก จะแบ่งกลุ่มข้อความออกเป็น ข้อความในด้านบวก ข้อความด้านลบ หรือข้อความที่เป็นกลาง หลังจากที่ได้ผลการวิเคราะห์ข้อความ ขั้นต่อมาจะนำไปแสดงผลให้เห็นว่า ข้อความด้านใดมีเป็นจำนวนเท่าไร ในรูปแบบที่ผู้รับสารสามารถทำความเข้าใจได้ง่าย จากการค้นคว้าพบว่ามีงานวิจัยอยู่หลากหลายที่เกี่ยวข้องกับการวิเคราะห์ความรู้สึก ผู้พัฒนาโครงการจึงได้ศึกษางานวิจัยที่ใช้เทคนิค การประมวลผลภาษาธรรมชาติ และการเรียนรู้ด้วยเครื่อง (Machine Learning Techniques) ในการระบุความรู้สึกของข้อความ โดยมักจำแนกออกเป็นสองด้าน คือ ด้านบวก และด้าน

ลบ ซึ่งได้งานวิจัยของ Kusrini และ Mochamad [1] ทำการวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์โดยใช้ซัพพอร์ตเวกเตอร์

แมชชีน และนาอีฟเบย์ในการทำนาย ซึ่งมีการใช้วิธีเทียบคลังคำศัพท์ (Lexicon Based Approach) [2] สำหรับการสกัดข้อมูล ผลที่ได้แสดงให้เห็นว่าการวิเคราะห์ความรู้สึกของข้อความโดยใช้ซัพพอร์ตเวกเตอร์

แมชชีน มีความแม่นยำในการจำแนกอยู่ที่ 79% และวิธีนาอีฟเบย์อยู่ที่ 84% อีกงานวิจัยของ Sheeba Naz และคณะ [3] ซึ่งวิเคราะห์ความรู้สึกของข้อความบนทวิตเตอร์โดยใช้เทคนิค n-gram [4] สำหรับการสกัดข้อมูล และซัพพอร์ตเวกเตอร์แมชชีนในการจำแนกความรู้สึก ผลที่ได้แสดงให้เห็นว่า ค่าความแม่นยำในการจำแนกความรู้สึกอยู่ที่ 81% ซึ่งค่อนข้างใกล้เคียงกับงานวิจัยแรก อย่างไรก็ตามงานวิจัยเหล่านี้ล้วนแต่วิเคราะห์ข้อความที่เป็นภาษาอังกฤษ ซึ่งการวิเคราะห์ข้อความภาษาไทยนั้นมีขั้นตอนการประมวลผลที่ยากกว่า ตั้งแต่ขั้นตอนการตัดคำ (word segmentation) [5] เนื่องจากข้อความภาษาอังกฤษมีการใช้เว้นวรรคในการบอกขอบเขตของคำ แต่ภาษาไทยเป็นภาษาที่ไม่มีเว้นวรรคระหว่างคำทำให้มีความกำกวม ดังตัวอย่างคลาสสิค เช่น ตากลม ซึ่งสามารถตัดคำได้ทั้ง 2 แบบคือ ตาก-ลม หรือ ตา-กลม ซึ่งต้องดูบริบทของข้อความด้วยว่าจะสื่อความหมายไปในทางใด

จากที่กล่าวมาข้างต้นทำให้เห็นแนวทางวิธีการวิเคราะห์ความรู้สึกของข้อความ ผู้พัฒนาโครงการจึงจะพัฒนาระบบวิเคราะห์ความรู้สึกของข้อความภาษาไทย เพื่อช่วยในการคัดกรองข้อความด้านบวก และข้อความด้านลบ ในการสรุปประเด็นความคิดเห็นด้านการเมือง เพื่อให้ผู้ใช้สามารถมองเห็นภาพรวมของประเด็นสถานการณ์ปัจจุบันทางการเมืองที่เกิดขึ้น

วัตถุประสงค์

1. เพื่อวิเคราะห์ความรู้สึกของข้อความที่เกี่ยวข้องกับการเมืองบนทวิตเตอร์
2. เพื่อสรุปความคิดเห็นที่มีต่อประเด็นทางการเมืองของประเทศไทย

ขอบเขตของโครงงาน

1. ระบบสามารถวิเคราะห์ข้อความที่ใช้ภาษาไทยเป็นภาษาหลัก
2. ข้อมูลที่นำมาวิเคราะห์ดึงข้อมูลจากทวิตเตอร์ จำนวนไม่ต่ำกว่า 10,000 ข้อความ
3. การวิเคราะห์ข้อความต้องการคลังคำศัพท์ภาษาไทยช่วยในการทำงาน
4. ระบบสามารถคัดกรองข้อความแสดงความรู้สึก และแบ่งออกเป็น ด้านบวก ด้านลบ และข้อความที่เป็นกลาง
5. ข้อมูลที่นำมาใช้สร้างวิธีการในการวิเคราะห์ข้อมูลจะเก็บข้อมูลในเดือนตุลาคมถึงเดือนพฤศจิกายน 2560

วิธีการดำเนินงาน

1. ศึกษาค้นคว้าข้อมูล
2. กำหนดขอบเขตและวิธีการดำเนินงาน
3. เก็บรวบรวมข้อมูลที่เกี่ยวกับการเมืองจากสื่อออนไลน์ทวิตเตอร์
4. วิเคราะห์การจำแนกข้อความ และระบุชนิดของคำ
5. ออกแบบ และพัฒนาระบบสำหรับการวิเคราะห์ความคิดเห็นทางด้านการเมือง
6. ตรวจสอบความถูกต้องของการจำแนกข้อมูล
7. จัดทำเอกสารและสรุปผล

ตารางเวลาการดำเนินงาน

การพัฒนาระบบวิเคราะห์ความรู้สึกของข้อความแสดงความคิดเห็นด้านการเมืองบนทวิตเตอร์ เริ่มดำเนินงานตั้งแต่เดือนกันยายน 2562 ถึง เดือนเมษายน 2563 รวมระยะเวลา 8 เดือน โดยมีตารางเวลาการดำเนินงาน ดังนี้

ขั้นตอนดำเนินงาน	2562				2563			
	ก.ย.	ต.ค.	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1. ศึกษาค้นคว้าข้อมูล	■	■						
2. กำหนดขอบเขตและวิธีดำเนินการ		■						
3. เก็บรวบรวมข้อมูลที่เกี่ยวกับการเมืองจากสื่อออนไลน์ทวิตเตอร์			■	■				
4. วิเคราะห์การจำแนกข้อความ และระบุชนิดของคำ				■				
5. ออกแบบและพัฒนาระบบสำหรับการวิเคราะห์ความคิดเห็นทางด้านการเมือง				■	■	■	■	
6. ตรวจสอบความถูกต้องของการจำแนกข้อมูล						■	■	
7. จัดทำเอกสารและสรุปผล			■	■	■	■	■	■

ประโยชน์ที่คาดว่าจะได้รับ

1. ประโยชน์ต่อผู้พัฒนา

- ได้พัฒนาโมเดลสำหรับการวิเคราะห์ข้อความ
- ได้พัฒนาทักษะการสร้างระบบตามขั้นตอนวิธีด้านวิศวกรรมซอฟต์แวร์ ซึ่งสามารถนำไปใช้พัฒนาระบบอื่นในอนาคตได้

2. ประโยชน์ต่อผู้ที่ใช้ระบบ

- ข้อมูลที่ได้จากการวิเคราะห์จะแสดงให้เห็นว่า ในปัจจุบันผู้คนในสื่อสังคมออนไลน์มีความคิดเห็นทางการเมืองในปัจจุบันเป็นด้านบวก หรือด้านลบ เกี่ยวข้องกับประเด็นใดบ้าง
- รัฐบาล หน่วยงานที่เกี่ยวข้อง หรือองค์กรที่เกี่ยวข้องกับการเมือง จะสามารถรับรู้ความคิดเห็นจากสื่อสังคมออนไลน์ เป็นไปในด้านบวก หรือด้านลบอย่างรวดเร็ว

อุปกรณ์และเครื่องมือที่ใช้

1. ฮาร์ดแวร์

เครื่องคอมพิวเตอร์	ระบบปฏิบัติการ Windows® แบบ 64 บิต
	หน่วยประมวลผล Intel® Core™ i5-6200U
	หน่วยความจำ DDR3 SDRAM ความจุ 8 กิกะไบต์
	พื้นที่ฮาร์ดดิสก์ มากกว่า 512 กิกะไบต์ ขึ้นไป

2. ซอฟต์แวร์
   - โปรแกรมภาษา Python 3.6 ใช้สำหรับการพัฒนาระบบ
   - ชุดซอฟต์แวร์ และไลบรารีของ Anaconda Distribution ใช้เป็นแพลตฟอร์ม (platform) สำหรับ พัฒนาระบบ

งบประมาณ

1. แบตเตอรี่โน๊ตบุ๊ครุ่น Acer Aspire V3-574G Hi-Power	จำนวน 1 ชิ้น ราคา 1,800 บาท
2. คีย์บอร์ด SIGNO KB-728 INVEGO (OPTICAL BLUE SWITCH)	จำนวน 1 ชิ้น ราคา 1,090 บาท
3. เมาส์ ASUS ROG GLADIUS II CORE	จำนวน 1 ชิ้น ราคา 1,490 บาท
4. หน่วยความจำ DDR3L คววามจุ 8 กิกะไบต์ ยี่ห้อ Corsair	จำนวน 2 ชิ้น ราคา 2,620 บาท
5. ฮาร์ดดิส SSD ความจุ 960 กิกะไบต์	จำนวน 1 ชิ้น ราคา 3,000 บาท
	รวมราคา 10,000 บาท

หมายเหตุ : ทั้งนี้ งบประมาณที่ตั้งไว้ขอถัวเฉลี่ยทุกรายการ

เอกสารอ้างอิง

[1] Kusrini Magister and Mochamad Mashuri , “Sentiment Analysis In Twitter Using Lexicon Based and Polarity Multiplication” in proceedings of the 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, pp.365 – 368, March 2019.

[2] S. Akter and M. T. Aziz, “Sentiment Analysis On Facebook Group Using Lexicon Based Approach,” in proceeding 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), Dhaka, Bangladesh, pp. 8–11, Sept 2016.

[3] Sheeba Naz,Aditi Sharan and Nidhi Malik, “sentiment classification on twitter data using support vector machine” in proceedings 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), Santiago, Chile, pp.676-679, Dec. 2018.

[4] T. Jhansi rani, K.anuradha and P.vijayapal reddy, “Sentiment Classification on Twitter data using Word N Gram Model” International Journal of Technology and Engineering Science (IJTES) Vol4, pp.7032-7035, 2016.

[5] ผศ.ดร.กานดา รุณนะพงศา และนางสาวปโยธร อุราธรรมกุล, “การตัดคำภาษาไทยโดยการปรับปรุงกฎและพจนานุกรมแบบใหม่” รายงานการวิจัยฉบับสมบูรณ์ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยขอนแก่น, 2549. [ออนไลน์] แหล่งที่มา: https://gear.kku.ac.th/~krunapon/research/thaiWordsSegmentation/FinalReportThaiWordsSegmentation.pdf

ภาคผนวก ข
ตัวอย่างโค้ดที่ใช้ในการพัฒนาระบบ

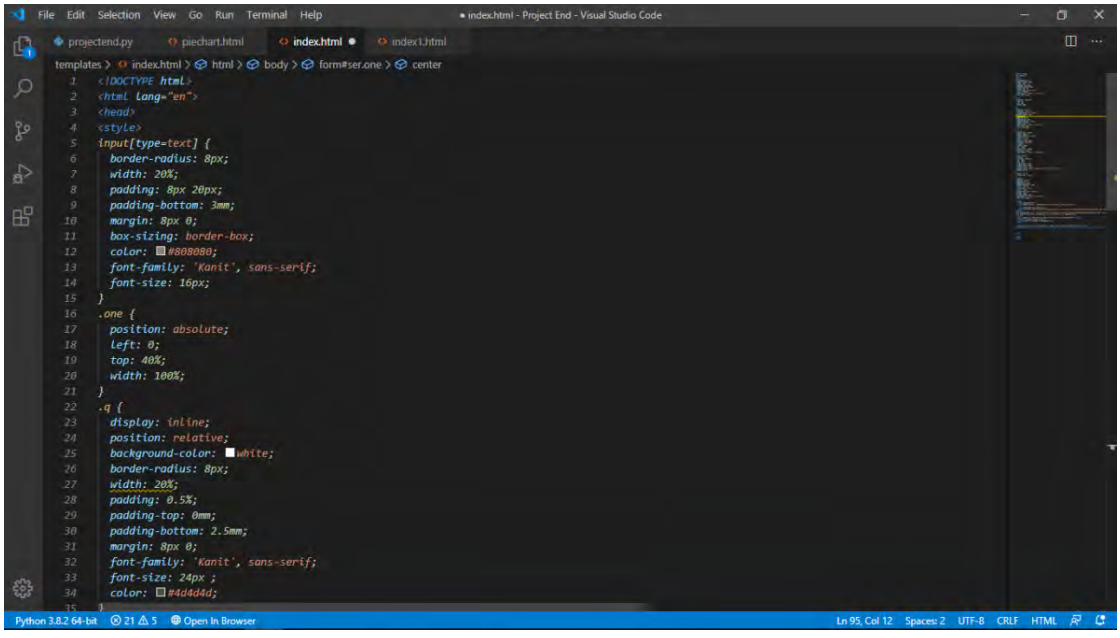
1. ตัวอย่างโค้ด tweepy

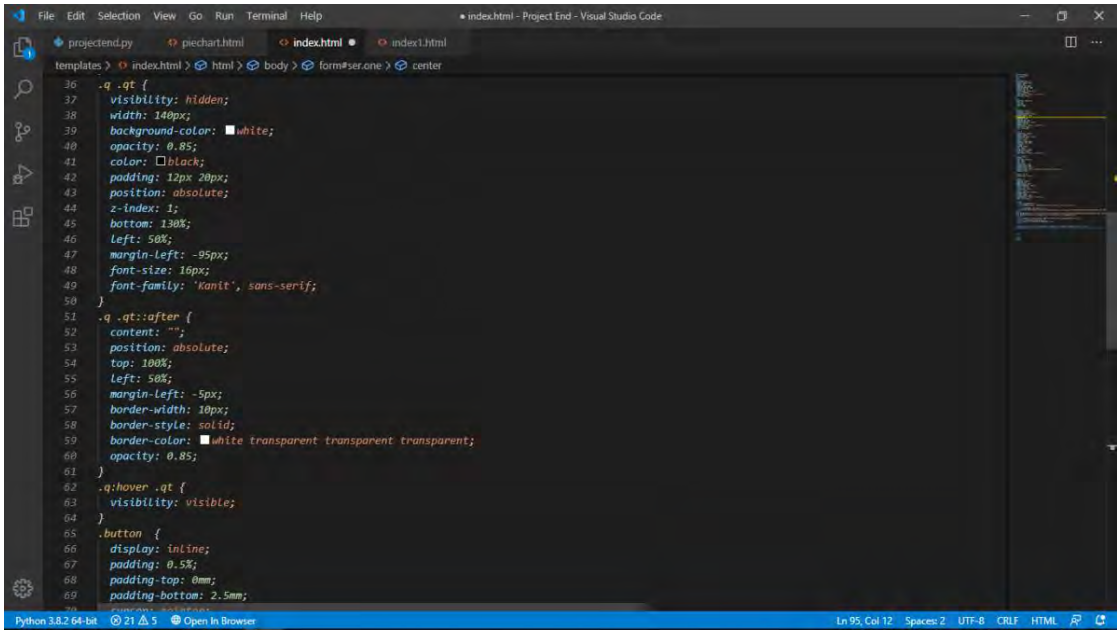
```
import tweepy
import pandas as pd
                            print(textwrite)
                            start += 1
                        elif text == "?":
                            start = count
                            textwrite = ""
                        end += 1
                    ID += 1
            text_file.close()
        consumer_key = "BPZkwHgAoaLs07D4uyCltKCeh"
        consumer_secret = "MTbImJYuzIXCEvdxny4mBtfzUYcP6mzC4JXXnRAYkQSfk
VNeFe"
        access_token = "4091669414-
zup9ASD2GD5SBJMmiDlUzJiwSzGa1wtqu5lqBe6"
        access_secret = "fLLQAqoJAiRLuA6HKYnPs7xsPMC0wIVc9QXvfUjvDzogP"

        auth = OAuthHandler("BPZkwHgAoaLs07D4uyCltKCeh", "MTbImJYuzIXCEvdxn
y4mBtfzUYcP6mzC4JXXnRAYkQSfkVNeFe")
        auth.set_access_token("4091669414-
zup9ASD2GD5SBJMmiDlUzJiwSzGa1wtqu5lqBe6", "fLLQAqoJAiRLuA6HKYnPs7xsPMC0
wIVc9QXvfUjvDzogP")
        api=tweepy.API(auth)
        tweets = api.search('ประชุมสภา')
        data = pd.DataFrame(data=[tweet.text for tweet in tweets], columns=['Twee
ts'])
```

2. **ตัวอย่างโค้ดเมท็อด** clean_tweet(text):

```
import re
import string
def clean_tweet(text):
    # ลบ text ที่อยู่ในวงเล็บ <> ทั้งหมด
    text = re.sub(r'<.*?>','', text)
    # ลบ hashtag
    text = re.sub(r'#','',text)
    text=re.sub(r'^[""RT]+','',text)#replace RT-tags
    # ลบ เครื่องหมายคำพูด (punctuation)
    #for c in string.punctuation:
    #    msg = re.sub(r'\{}'.format(c),'',text)
    text = re.sub(r'@\S+', '',text)
    # ลบ separator เช่น \n \t
    text = ' '.join(msg.split())
    text = re.sub(r'http\S+', '', text)
```

3. **ตัวอย่างโค้ดหน้าหลักของเว็ปแอปพลิเคชัน**




File Edit Selection View Go Run Terminal Help
index.html - Project End - Visual Studio Code
projectend.py
piechart.html
index.html
index1.html
templates > index.html > html > body > form#ser.one > center
.q .qt {
visibility: hidden;
width: 140px;
background-color: white;
opacity: 0.85;
color: black;
padding: 12px 20px;
position: absolute;
z-index: 1;
bottom: 130%;
left: 50%;
margin-left: -95px;
font-size: 16px;
font-family: 'Kanit', sans-serif;
}
.q .qt::after {
content: "";
position: absolute;
top: 100%;
left: 50%;
margin-left: -5px;
border-width: 10px;
border-style: solid;
border-color: white transparent transparent transparent;
opacity: 0.85;
}
.q:hover .qt {
visibility: visible;
}
.button {
display: inline;
padding: 0.5%;
padding-top: 0mm;
padding-bottom: 2.5mm;
Python 3.8.2 64-bit
21 5
Open In Browser
Ln 95, Col 12
Spaces: 2
UTF-8
CRLF
HTML

ประวัติผู้เขียน

Mr.Jittiwat Jangjun
นายจิตรทิวัส แจ้งจันทร์ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ที่อยู่ 4/2 ม.5 ต.บางกรวย อ.บางกรวย จ.นนทบุรี 11130
มือถือ 0632691155
email jame.simplify@gmail.com

Mr.Chonlawit Gonthong
นายชลวิทย์ ก้อนทอง ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ที่อยู่ 19/20 ถนน ประชาสงเคราะห์ ดินแดง กรุงเทพฯ 10400
มือถือ 0859725550
email saguraisama@gmail.com