ระบบทำนายอารมณ์ตามสถานที่จากสื่อสังคมออนไลน์

DOOME: SENTIMENTAL PREDICTION FROM SOCIAL MEDIA USING LOCATION-BASE SYSTEM

สรัล รักวิจิตรศิลป์

สิรภัทร อัชชศิริ

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2559

ระบบทำนายอารมณ์ตามสถานที่จากสื่อสังคมออนไลน์

DOOME: SENTIMENTAL PREDICTION FROM SOCIAL MEDIA USING LOCATION-BASE SYSTEM

สรัล รักวิจิตรศิลป์
สิรภัทร อัชชศิริ

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2559

ปริญญานิพนธ์ปีการศึกษา 2559

ภาควิชาวิศวกรรมคอมพิวเตอร์

คณะวิศวกรรมศาสตร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง ระบบทำนายอารมณ์ตามสถานที่จากสื่อสังคมออนไลน์

DOOME: SENTIMENTAL PREDICTION FROM SOCIAL MEDIA USING LOCATION-BASE SYSTEM

ผู้จัดทำ

1. นายสรัล รักวิจิตรศิลป์ รหัสนักศึกษา 56011278

2. นายสิรภัทร อัชชศิริ รหัสนักศึกษา 56011317

_____ อาจารย์ที่ปรึกษา

(ดร.อรทัย สังข์เพ็ชร)

_____ อาจารย์ที่ปรึกษา

(ดร.อักฤทธิ์ สังข์เพ็ชร)

ระบบทำนายอารมณ์ตามสถานที่จากสื่อสังคมออนไลน์

นายสรัล	รักวิจิตรศิลป์	56011278
นายสิรภัทร	อัชชศิริ	56011317
ดร.อรทัย	สังข์เพ็ชร	อาจารย์ที่ปรึกษา
ดร.อักฤทธิ์	สังข์เพ็ชร	อาจารย์ที่ปรึกษาร่วม

ปีการศึกษา 2559

บทคัดย่อ

การที่จะตัดสินใจเดินทางไปสถานที่ใดสถานที่หนึ่ง จำเป็นต้องใช้ปัจจัยหลายด้าน ซึ่งสิ่งที่น่าจะเป็นตัวช่วยได้ดีที่สุดก็คือความรู้สึกและประสบการณ์จากผู้ที่เคยไปมา โดยปกติผู้คนมักจะแสดงความรู้สึก อารมณ์ หรือประสบการณ์ผ่านสื่อสังคมออนไลน์ ในโครงงานจึงเลือกที่จะสร้างระบบที่มีการรวบรวมข้อมูลจากสื่อสังคมออนไลน์และนำมาทำการวิเคราะห์เพื่อทำนายอารมณ์ในภาพรวมของสถานที่นั้นนั้นว่ามีการแสดงออกมาในลักษณะใด โดยเราจะประยุกต์ใช้ความรู้ในด้านการเรียนรู้ของเครื่องและการทำเหมืองข้อมูลเพื่อทำการวิเคราะห์ ผลลัพธ์ที่ได้จะเป็นตัวบอกว่าคนที่เคยเดินทางไปยังสถานที่นั้นแสดงความรู้สึกออกมาเป็นอย่างไร ข้อมูลดังกล่าวจะเป็นตัวช่วยในการพิจารณาให้เลือกสถานที่ที่ตรงกับความรู้สึกและความต้องการของผู้ใช้ในขณะนั้น

DooMe: Sentimental Prediction From Social Media Using Location-base System

Mr. Sarun	Rakwijitsil	56011278
Mr. Sirapat	Attchasiri	56011317
Dr. Orathai	Sangpetch	Advisor
Dr. Akkarit	Sangpetch	Co-Advisor

Academic Year 2016

ABSTRACT

To make a decision where we would like to go, there are many relevant factors. One of the most important factors is experiences from people who have visited there. Nowadays people like expressing their emotions and thoughts on social networks, such as Facebook and Twitter. Thus, we plan to build a system to collect data from a variety of social networks. The data will then be analyzed in order to predict how people feel when visiting such a place. In the analysis phase, we will apply knowledge in the machine learning and data mining field to build a model where it processes words and characters posted by people to indicate the emotions people have towards the place. The result from our model can be used to help other people who have never been to such a place to decide whether they want to go.

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยดีด้วยความช่วยเหลือจากหลายฝ่ายทั้งในทางตรงและทางอ้อม โดยจะสำเร็จลงไม่ได้หากปราศจากความช่วยเหลือของบุคคลเหล่านี้

อาจารย์ที่ปรึกษาทั้งสองท่านคือ ดร.อรทัย สังข์เพ็ชร และ ดร.อักฤทธิ์ สังข์เพ็ชร ซึ่งท่านเป็นผู้ที่ให้คำแนะนำ คำปรึกษา ข้อคิดเห็นอันเป็นประโยชน์ต่อการพัฒนาผลงานรวมถึงคอยติดตามดูแลการดำเนินงานอย่างใกล้ชิดเพื่อที่จะทำให้การทำงานต่าง ๆ เป็นไปอย่างราบรื่นและออกมามีประสิทธิภาพสูงสุด

อาจารย์และบุคลากรในภาควิชาวิศวกรรมคอมพิวเตอร์ที่ได้ให้คำแนะนำและถ่ายทอดความรู้มาตลอดรวมถึงห้องปฏิบัติการ SAIG (Software and Application Interest Group) และ SOUP (System Operations Usability Parallel computing) ที่ได้เอื้อเฟื้อสถานที่และทรัพยากรต่าง ๆ ที่จำเป็นในการทำวิจัยและพัฒนาโครงงาน

ในท้ายที่สุดนี้ขอขอบคุณบิดา มารดา และครอบครัวที่ได้อบรมเลี้ยงดูและสั่งสอน พร้อมทั้งให้โอกาสในการศึกษาและให้กำลังใจเสมอมา

สรัล รักวิจิตรศิลป์

สิรภัทร อัชชศิริ

สารบัญ

หน้า

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

สารบัญตาราง (ต่อ)

ตาราง **หน้า**

สารบัญรูป

รูป หน้า

สารบัญรูป (ต่อ)

รูป หน้า

สารบัญรูป (ต่อ)

รูป **หน้า**

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

ในปัจจุบันการท่องเที่ยวนั้นเป็นรายได้หลักในหลายประเทศ โดยปกติแล้วทุกครั้งในการท่องเที่ยว ก่อนจะมีการท่องเที่ยวนั้นก็จะมีการวางแผนการท่องเที่ยวว่าสามารถจะไปที่ใดได้บ้าง โดยอาจจะดูจากทางอินเทอร์เน็ตว่ามีสถานที่ใดน่าสนใจ หรือมีสถานที่ในใจแล้วหาข้อมูลของสถานที่นั้น ๆ แต่ปัญหาคือ ข้อมูลที่หามาได้นั้นอาจจะไม่ใช่ข้อมูลในปัจจุบัน อาจจะเป็นข้อมูลเมื่อเดือนที่แล้ว หรือปีที่แล้ว ทำให้ไม่รู้ว่าสถานที่ที่เราจะไปนั้นมีสภาพเป็นอย่างไร คนที่เคยไปเมื่อไม่นานมานี้มีประสบการณ์ที่ดีหรือไม่ หรือในอินเทอร์เน็ตอาจจะไม่มีข้อมูลของสถานที่ที่จะไปอยู่เลย ทำให้อาจจะต้องสอบถามคนที่รู้จักหรือเพื่อนที่เคยไปมา แต่เพียงแค่การสอบถามข้อมูลอาจจะยังไม่พอที่จะใช้พิจารณา โดยปกติแล้วบุคคลทั่วไปมักจะแสดงความคิดเห็นต่อสถานที่ที่เคยไปมาผ่านทางสื่อสังคมออนไลน์ ซึ่งข้อมูลพวกนี้เป็นข้อมูลที่ค่อนข้างเป็นปัจจุบัน และเป็นข้อมูลที่ได้จากหลายบุคคล ข้อมูลที่ได้จากสื่อสังคมออนไลน์นี้สามารถบอกถึงประสบการณ์ที่ได้รับมาได้ เช่น ถ้าบุคคลนั้นได้รับประสบการณ์ที่ดี ก็จะทำให้บุคคลนั้นมีอารมณ์ดี หรือในทางกลับกันถ้าบุคคลนั้นได้รับประสบการณ์ที่ไม่ดี ก็จะทำให้บุคคลนั้นมีอารมณ์โกรธเป็นต้น

ดังนั้นทางผู้พัฒนาจึงเห็นว่าถ้าสามารถนำข้อความเหล่านั้นที่อยู่บนสื่อสังคมออนไลน์มาประมวลจำแนกออกเป็นอารมณ์ต่าง ๆ ได้จะช่วยให้คนอื่น ๆ สามารถนำข้อมูลเหล่านี้ไปใช้ต่อได้ โดยผู้พัฒนาประยุกต์ใช้ความรู้ทางด้านการเรียนรู้ของเครื่อง (Machine Learning) และการทำเหมืองข้อมูล (Data mining) เพื่อหาอารมณ์ของบุคคลที่พิมพ์ข้อความนั้น ๆ ก็จะทำให้เรามีข้อมูลที่ใกล้เคียงกับปัจจุบันหรือเป็นปัจจุบัน เพื่อที่จะซึ่งจะช่วยให้เราวางแผนการท่องเที่ยวได้ง่ายมากขึ้น

1.2 วัตถุประสงค์ของโครงงาน

1) เพื่อพัฒนางานด้านวิทยาศาสตร์และเทคโนโลยีให้มีการอำนวยความสะดวกให้กับสังคมปัจจุบัน
2) เพื่อใช้ในการทำนายอารมณ์ของบุคคลในบริเวณที่เราสนใจและนำข้อมูลที่ได้จากการทำนายมาช่วยให้การวางแผนกิจกรรมต่าง ๆ เช่น การท่องเที่ยว ทำได้ง่ายขึ้น
3) เพื่อสร้างระบบที่สามารถทำนายอารมณ์ด้วยวิธีการใช้ การเรียนรู้ของเครื่อง
4) เพื่อประยุกต์ใช้ประโยชน์ของข้อมูลจากสื่อสังคมออนไลน์ที่มีหลากหลายในปัจจุบัน

1.3 ประโยชน์ที่คาดว่าจะได้รับ

โครงงานนี้สามารถนำไปใช้เป็นตัวช่วยในการตัดสินใจที่จะเลือกเดินทางไปยังสถานที่ต่าง ๆ ของผู้ใช้โดยแค่ผู้ใช้เปิดใช้แอปพลิเคชันและดูอารมณ์ตามสถานที่ต่าง ๆ ที่แสดงอยู่บนแผนที่ซึ่งจะทำให้ทราบถึงประสบการณ์ที่ได้รับของคนที่เคยไปยังสถานที่ดังกล่าว

1.4 ขอบเขตของโครงงาน

1) ระบบนี้รองรับเฉพาะข้อความในส่วนที่เป็นภาษาไทย
2) ระบบนี้สามารถสร้างโมเดลที่ใช้ในการทำนายอารมณ์จากข้อความได้โดยการเลือกใช้วิธีพื้นฐานทั่วไปและอัลกอริทึมในด้านการเรียนรู้ของเครื่อง (Machine Learning) 2 เทคนิคคือ
   1.1) Rule-based method (วิธีพื้นฐานทั่วไป)
   1.2) Naïve Bayes (อัลกอริทึมในด้านการเรียนรู้ของเครื่อง)
   1.3) Multinomial Naïve Bayes (อัลกอริทึมในด้านการเรียนรู้ของเครื่อง)
   เพื่อทำการทดสอบหาประสิทธิภาพที่ดีที่สุด
3) ระบบนี้สามารถรองรับการเก็บและจัดการข้อมูลทางสื่อสังคมออนไลน์ในปริมาณมากที่จะเกิดขึ้นในอนาคตได้
4) ระบบนี้สามารถทำนายอารมณ์ได้ 8 อารมณ์ คือ joy (มีความสุข สนุกสนาน) sadness (เศร้า เสียใจ) fear (กลัว) angry (โกรธ) disgust (รังเกียจ) surprise (ประหลาดใจ) anticipation (คาดหวัง สนใจ) acceptance (ยอมรับ)
5) ระบบนี้สามารถแสดงผลผ่านโทรศัพท์มือถือได้

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้องด้านระบบจัดการข้อมูล

การจัดการข้อมูลเป็นส่วนสำคัญอย่างหนึ่งของระบบนี้เนื่องจากระบบนี้ต้องรองรับข้อมูลที่มีปริมาณมากจากสื่อสังคมออนไลน์เช่น Facebook และ Twitter ซึ่งถ้าระบบมีการจัดการข้อมูลที่ไม่ดีจะทำให้ระบบทำงานได้ช้า หรือทำงานได้ไม่เร็วเท่าที่ตั้งใจไว้ เพราะเหตุนี้ความรู้ในเรื่องของการจัดการข้อมูลจึงจำเป็น เครื่องมือที่ที่เลือกมาใช้ในการจัดการข้อมูลมีดังนี้

2.1.1 Apache Hadoop

Hadoop [16] เป็นซอฟต์แวร์ของ The Apache Software Foundation ใช้สำหรับการจัดเก็บและประมวลผลข้อมูลขนาดใหญ่ โดย Hadoop จะสามารถรองรับการขยายตัวของข้อมูลได้ดี เพราะ Hadoop ถูกออกแบบมาสำหรับการคำนวณผลแบบกระจายโดยผ่านเครื่องคอมพิวเตอร์มากมายที่อยู่ในกลุ่มเดียวกัน ซึ่งในกลุ่มสามารถมารถมีคอมพิวเตอร์ตั้งแต่ 1 เครื่องไปยังหลักพันเครื่อง แกนหลักของ Hadoop จะประกอบไปด้วยส่วนของที่เก็บข้อมูลซึ่งเรียกว่า Hadoop Distributed File System (HDFS) และส่วนของการประมวลผลซึ่งเรียกว่า MapReduce

Hadoop มี 4 modules หลักคือ Hadoop Common Hadoop Distributed File System (HDFS) Hadoop YARN และ Hadoop MapReduce

1) Hadoop Common เป็น common utilities ที่สนับสนุน module อื่น ๆ ของ Hadoop
2) Hadoop Distributed File System (HDFS) เป็น distributed file system ที่ทำให้สามารถเข้าถึงข้อมูลของแอปพลิเคชันได้อย่างรวดเร็ว
3) Hadoop YARN เป็น framework สำหรับการทำ job scheduling และ cluster management
4) Hadoop MapReduce เป็น YARN-based system ที่ใช้สำหรับการประมวลผลแบบ parallel ในข้อมูลที่มีขนาดใหญ่

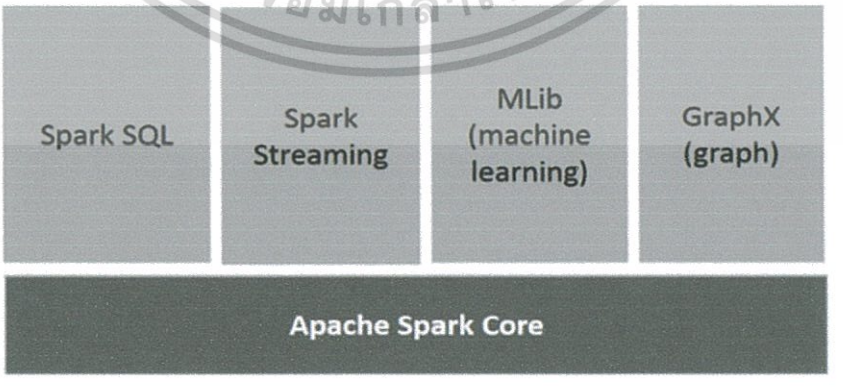
ภายในโครงงานนี้เราได้เลือกใช้ HDFS ที่เป็น distributed file system สำหรับเก็บไฟล์ Parquet ที่เป็นไฟล์ฐานข้อมูลที่ใช้ภายในโครงงาน เพราะตัวระบบต้องสามารถรองรับปริมาณข้อมูลขนาดใหญ่จากสื่อสังคมออนไลน์เช่น Twitter ซึ่งในการในการรับข้อมูล ถ้าเรามี keyword ในการค้นหามาก ข้อมูลก็จะมีปริมาณมากตามไปด้วย

2.1.2 Apache Spark

Spark [8] เป็นซอฟต์แวร์สำหรับการประมวลผลข้อมูล โดยประมวลผลเป็นแบบกระจาย โดยผ่านเครื่องคอมพิวเตอร์มากมายที่อยู่ในกลุ่มเดียวกัน ซึ่งเป็นของ The Apache Software Foundation Spark มีพื้นฐานมาจาก Hadoop MapReduce และได้พัฒนาต่อยอดทำให้มีขีดความสามารถมากขึ้น โดยปกติแล้ว Spark ใช้สำหรับการประมวลผลข้อมูลที่มีขนาดใหญ่

ซึ่งส่วนประกอบหลักของ Spark มีดังนี้คือ Spark Core Spark SQL Spark Streaming MLib และ GraphX [9]

1) Spark Core เป็น engine สำหรับ spark platform ซึ่งมีความสามารถในการประมวลผลแบบ in-memory และสามารถอ้างถึงข้อมูลจาก storage system อื่น ๆ ได้
2) Spark SQL จะมี data abstraction แบบใหม่คือ RDD ซึ่งเป็นกลุ่มของออปเจกต์ ที่ไม่สามารถเปลี่ยนได้ (immutable) และเป็นแบบกระจาย (distributed) และข้อมูลที่อยู่ใน RDD จะถูกแบ่งเป็น partition ซึ่งทำให้สามารถนำไปคำนวณบนเครื่องอื่นที่อยู่ในกลุ่ม (cluster) เดียวกันได้
3) Spark Streaming ใช้สำหรับการทำ streaming analytic โดยข้อมูลที่นำเข้ามาจะถูกจะถูกแบ่งให้เป็นชุด (batches) จากข้อมูลที่เป็นชุด ๆ จะถูกนำเข้าไปประมวลผลใน Spark engine และผลลัพธ์จะเป็นแบบชุด (batches) ข้อมูล
4) MLib เป็น distributed machine learning framework ที่ทำให้สามาถใช้งานฟังก์ชันการทำงานของ machine learning ได้
5) GraphX เป็น distributed graph-processing framework ซึ่งมี API สำหรับให้ผู้ใช้สามารถจำลองกราฟที่ใช้กำหนด ซึ่งใช้ Pregel abstraction API

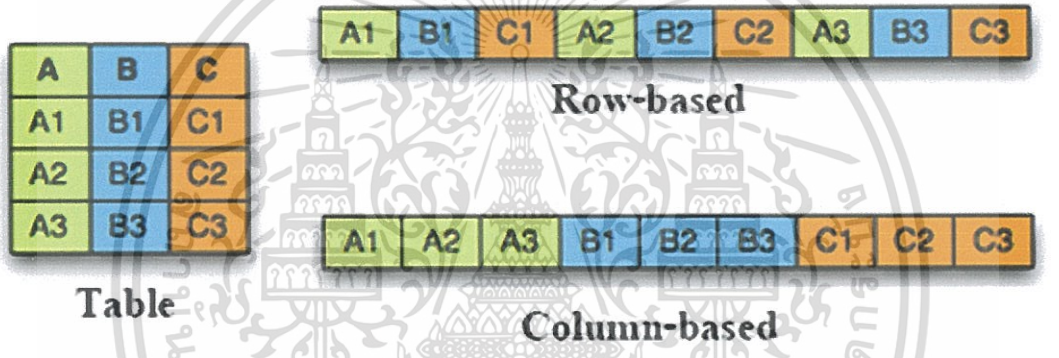


รูป 2.1 ส่วนประกอบของ Apache Spark

ภายในโครงงานนี้เราได้เลือกใช้ Spark SQL ซึ่งทำให้สามารถใช้คำสั่งแบบ SQL เพื่อดึงข้อมูลจากไฟล์ Parquet และใช้สำหรับการเขียนไฟล์ Parquet

2.1.3 Apache Parquet

Parquet [6] เป็น columnar storage format ที่ใช้ใน Hadoop ecosystem ซึ่งมีข้อได้เปรียบคือการบีบอัด (compression) ทำให้เนื้อที่การใช้งานนั้นลดลง และการเลือกข้อมูลแบบทั้ง column สามารถทำได้ดีกว่าแบบ row format โดย Parquet ใช้วิธี record shredding และ assembly algorithm ในการสร้างโครงสร้างข้อมูลที่มีความซับซ้อน [7] จากความสามารถในการเลือกข้อมูลแบบทั้ง column ทำให้การสรุปข้อมูลหรือการที่จะเข้าถึงข้อมูลเพียงแค่ column เดียวสามารถทำได้ดีกว่าการเลือกใช้ storage แบบ row format ซึ่งในการนำข้อมูลไปประมวลผลมักจะเลือกข้อมูลจาก column หนึ่ง ๆ ไปประมวลผล



รูป 2.2 ความแตกต่างระหว่าง Column-based กับ Row-based

ภายในโครงงานนี้ได้เลือกใช้ไฟล์ Parquet เพราะไฟล์ Parquet สามารถใช้งานร่วมกับ Spark และ HDFS ได้ และ Parquet มีการเก็บข้อมูลแบบ column ซึ่งทำให้สามารถดึงข้อมูล column เดี่ยวได้อย่างรวดเร็วมากกว่าฐานข้อมูลที่เก็บข้อมูลแบบ row

2.2 การแบ่งกลุ่มอารมณ์

อารมณ์ [10] มีความหมายว่า การเกิดการเคลื่อนไหว หรือภาวะที่ตื่นเต้นเป็นการยากที่จะบอกว่าอารมณ์คืออะไร แต่มีแนวคิดหนึ่งที่ช่วยให้เกิดความเข้าใจได้ง่ายกล่าวว่าอารมณ์เป็นความรู้สึกภายในที่เร้าให้บุคคลกระทำหรือเปลี่ยนแปลงภายในตัวของบุคคลนั้นเองซึ่งความรู้สึกเหล่านี้จะเป็นความรู้สึกที่พึงพอใจ ไม่พึงพอใจ หรือรวมกันทั้งสองกรณี อารมณ์เป็นสิ่งที่ไม่คงที่มีการแปรเปลี่ยนอยู่ตลอดเวลา

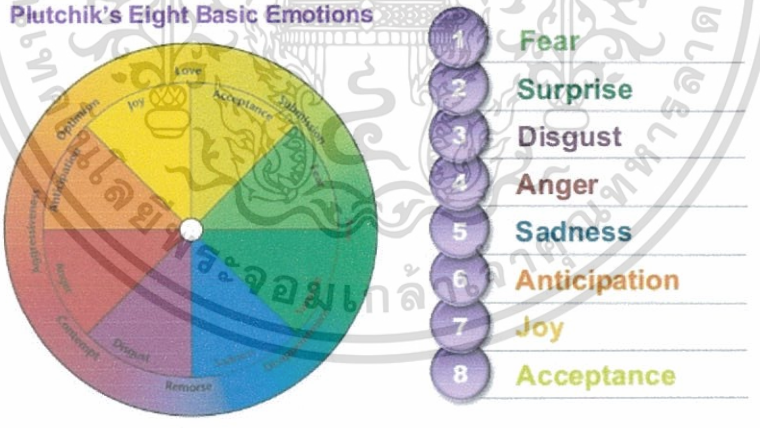
อารมณ์ที่มีการแสดงออกมาของมนุษย์นั้นมีลักษณะที่หลากหลายขึ้นอยู่กับปัจจัยที่เกิดขึ้นในช่วงเวลานั้นโดยที่บางลักษณะอารมณ์ที่ถูกนิยามขึ้นมีความใกล้เคียงกันจึงได้มีการทำการแบ่งกลุ่มของอารมณ์เพื่อให้เกิดความเข้าใจง่ายโดยการแบ่งกลุ่มทางอารมณ์นั้นได้แบ่งตามผู้เชี่ยวชาญ

ทางด้านจิตวิทยาซึ่งมีหลายทฤษฎีด้วยกัน นักจิตวิทยาแต่ละคนต่างก็มีการแบ่งอารมณ์ที่ต่างกันโดยทำการเลือกการแบ่งกลุ่มของอารมณ์ดังนี้

2.2.1 การแบ่งกลุ่มอารมณ์โดย โรเบิร์ต พูลทชิค (Robert Plutchik)

มีการจำแนกอารมณ์พื้นฐานออกเป็น 8 ชนิด [11,17] ตามรูป 2.3 คือ

1) กลัว (Fear) เป็นอารมณ์ที่แสดงออกถึงความรู้สึกว่าเป็นอันตราย
2) ประหลาดใจ (Surprise) เป็นอารมณ์ที่ก่อให้เกิดการเปลี่ยนแปลงของสิ่งเร้าในระบบประสาทอย่างฉับพลัน
3) รังเกียจ (Disgust) เป็นอารมณ์อันเกิดจากการกระทบกับสัมผัสที่ไม่พึงปรารถนา
4) โกรธ (Anger) เป็นอารมณ์ที่ไม่พึงพอใจอย่างแท้จริง
5) เศร้าเสียใจ (Sadness) เป็นอารมณ์ที่เกิดขึ้นเมื่อบุคคลต้องประสบกับความพลัดพราก
6) คาดหวัง (Anticipation) เป็นอารมณ์เมื่อเราต้องการให้มีสิ่งใดสิ่งหนึ่งเกิดขึ้นตามที่เราคิดไว้
7) รื่นเริง (Joy) เป็นอารมณ์ที่ก่อให้เกิดสภาวะของความเชื่อมั่น
8) ยอมรับ (Acceptance) เป็นอารมณ์ที่มีการตกลงที่จะรับในสิ่งใหม่ที่เห็นว่าถูกต้องหรือทำให้เกิดความพอใจ



รูป 2.3 การแบ่งกลุ่มอารมณ์โดย Robert Plutchik

ในโครงงานนี้เราได้ทำการศึกษาการแบ่งกลุ่มของอารมณ์เพื่อที่จะนำมาใช้ในจำแนกอารมณ์ของคนที่มักจะแสดงออกมาให้พบเห็นมากที่สุดบนสื่อสังคมออนไลน์โดยทำการเลือกประเภทของอารมณ์ตามการแบ่งกลุ่มอารมณ์ของ โรเบิร์ต พูลทชิค (Robert Plutchik) ที่มีการแบ่งแยกไว้แล้ว

2.3 การทำเหมืองข้อความ (Text Mining)

การทำเหมืองข้อความ [12] เป็นการวิเคราะห์ข้อมูลที่มีอยู่อยากหลากหลาย การประยุกต์ใช้เทคนิคการทำเหมืองข้อความในการแก้ปัญหาอาจถูกเรียกว่าการวิเคราะห์ข้อความ ซึ่งเป็นเทคนิคเพื่อค้นหารูปแบบ (pattern) ของจากข้อความจำนวนมากที่พบในปัจจุบัน

การทำเหมืองข้อความทำให้เราได้รับข้อมูลเชิงลึกที่มีคุณค่าในการที่จะนำไปต่อยอดในทางธุรกิจขอมูลที่นำมาทำเหมืองข้อความมักได้มาจาก เอกสารต่าง ๆ อีเมล ข้อความที่มีการแสดงความคิดเห็นจากสื่อสังคมออนไลน์ที่เป็นที่นิยมไม่ว่าจะเป็น Twitter หรือ Facebook เป็นต้น

โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง หลักการประมวลเอกสาร หลักการประมวลผลข้อความ และการประมวลผลภาษาธรรมชาติ (Natural Language Processing) เพื่อค้นหารูปแบบ แนวทางและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความนั้น

ขั้นตอนหลัก ๆ ที่มีความจำเป็นในการทำเหมืองข้อความประกอบด้วย

1) ทำความเข้าใจปัญหา
2) ทำความเข้าใจข้อมูล
3) เตรียมข้อมูล (Training set, Test set)
4) สร้างแบบจำลองจากอัลกอริทึม
5) ประเมิน
6) นำไปใช้งาน

โดยในโครงงานนี้ได้นำวิธีการของการทำเหมืองข้อความมาใช้กับความคิดเห็นที่เก็บมาจากสื่อสังคมออนไลน์ ในขั้นตอนของการเตรียมข้อมูลที่จะถูกนำไปใช้ในกระบวนการต่อไปซึ่งทำงานร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติ และการเรียนรู้ของเครื่องเพื่อที่จะสามารถทำนายอารมณ์จากข้อความที่ผู้คนได้มีการแสดงความคิดเห็นจากสื่อสังคมออนไลน์ออกมาในรูปแบบที่เข้าใจง่าย

2.4 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

การประมวลผลภาษาธรรมชาติหรือ NLP เป็นการสร้างโปรแกรมทางคอมพิวเตอร์ให้มีความสามารถในการเข้าใจภาษามนุษย์โดยที่ NLP เป็นสาขาหนึ่งของ ปัญญาประดิษฐ์ (Artificial Intelligence)

การพัฒนาแอปพลิเคชันทางด้านการประมวลผลภาษาธรรมชาตินั้นมีความท้าทายเป็นอย่างมาก โดยเป้าหมายของระบบประมวลผลภาษาธรรมชาติคือการออกแบบและสร้างซอฟต์แวร์เพื่อที่จะวิเคราะห์ทำความเข้าใจและสร้างภาษาที่มนุษย์ใช้ในการดำเนินชีวิตปกติเพื่อที่ว่าในที่สุดเราจะสามารถที่จะอยู่กับเครื่องคอมพิวเตอร์ราวกับว่าเรากำลังอยู่กับบุคคลอื่น

NLP ถูกนำมาใช้ในการวิเคราะห์ข้อความที่ช่วยให้เครื่องจักรเข้าใจวิธีการพูดหรือการเขียนของมนุษย์โดยที่การมีปฏิสัมพันธ์ระหว่างมนุษย์กับคอมพิวเตอร์จะช่วยในการใช้งานบนโลกความเป็นจริงเช่น การสรุปความอัตโนมัติ การวิเคราะห์อารมณ์จากข้อความ การสกัดหัวข้อ การสกัดความสัมพันธ์เป็นต้น ซึ่งปกติแล้ว NLP ถูกใช้งานมากในด้านการทำเหมืองข้อความ (text mining)

โดยที่ในโครงงานนี้เราได้นำวิธีการของการประมวลผลภาษาธรรมชาติมาใช้ช่วยในส่วนการตัดคำในข้อความภาษาไทยเพื่อนำไปใช้ในการสร้างคุณลักษณะของคำในการกระบวนการสร้างตัวแบบจำลองในการทำนายอารมณ์ต่อไป

2.4.1 N-Gram model

N-Gram [5] เป็นแบบจำลองที่ใช้ในการคำนวณค่าความน่าจะเป็นของชุดตัวอักษรที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของคำที่เขียนเรียงต่อกันจนกลายเป็นประโยค หน่วยที่ N-Gram ใช้ในการสร้างแบบจำลอง อาจจะเป็นเสียง คำ หรือ ตัวอักษรก็ได้ โดยที่ขนาดของ N-Gram ขึ้นอยู่กับที่เราจะเป็นผู้กำหนด ตั้งแต่ 1 จนถึง N ในแบบจำลอง N-Gram นี้ใช้ความยาวของคำที่เขียนเรียงกันซึ่งมีความแตกต่างกันได้แก่ 1-Gram (unigram) 2-Gram (bigram) 3-Gram (trigram) เป็นต้น

N-Gram ถูกนำมาประยุกต์ใช้กับงานด้านการประมวลผลธรรมชาติ ซึ่งเป็นการนำบางส่วนของข้อความนั้นออกมาเป็นหน่วยคำตามค่าของ N ที่ทำการกำหนด โดยทั่วไปแล้วคำในภาษาไทยนิยมใช้การตัดคำแบบ 2 3 และ 4 Gram

ในโครงงานนี้เราจะนำรูปแบบของ N-Gram model ใช้เป็นคุณลักษณะ(Feature) เพื่อใช้ในการสร้างแบบจำลองที่ใช้ทำนาย ซึ่งโดยปกติจะถูกใช้ในกระบวนการของการตัดคำเพื่อให้ออกมาเป็นหน่วยคำที่สื่อความหมาย

2.5 Rule-Based expert system

Rule-based expert system [3] หรือเรียกอีกชื่อหนึ่งว่า Knowledge-based system ซึ่งจะประกอบด้วย 2 ส่วนคือ knowledge และ expert ซึ่ง knowledge คือความเข้าใจในทางทฤษฎีหรือทางปฏิบัติของสิ่ง ๆ หนึ่ง และคนที่เป็นเจ้าของ knowledge เรียกว่า expert โดย knowledge จะเป็นตัวที่ทำให้ expert สามารถแก้ปัญหาต่าง ๆ ได้เช่นในเรื่องของสัญญาไฟจราจร knowledge ของ expert มีอยู่ว่าถ้าเกิดพบสัญญาไฟจราจรมีสีเขียวจะสามารถไปได้ แต่ถ้าเกิดสัญญาณไฟจราจรมีสีแดงจะหยุด ซึ่งตัว knowledge จะเป็นดังนี้

IF สัญญาณไฟจราจรมีสีเขียว

THEN ไป

IF สัญญาณไฟจราจรมีสีแดง

THEN หยุด

โดยถ้า expert เกิดพบปัญหาตาม knowledge ระบุไว้ก็จะสามารถแก้ปัญหาในส่วนนั้นได้โดยใช้ knowledge ที่มีอยู่

ซึ่งในโครงงานนี้เราได้นำในส่วนของ Rule-Based มาใช้เนื่องจากข้อความจากสื่อสังคมออนไลน์ที่เราทำการเลือกใช้คือ Twitter เป็นลักษณะข้อความสั้น ๆ การที่คนที่ใช้งาน จะทำการพิมพ์ข้อความขึ้นมานั้นจะต้องมีคำที่เป็น keyword หรือสัญลักษณ์ที่แสดงออกถึงอารมณ์ ซึ่งถ้าเราสามารถที่จะหาคำที่เป็น keyword และสัญลักษณ์ที่แสดงออกถึงอารมณ์ได้อย่างคลอบคลุมเราก็จะสามารถตัดสินได้ว่าข้อความนั้นต้องการที่จะแสดงออกถึงอารมณ์ลักษณะใดได้อย่างถูกต้อง

2.6 การเรียนรู้ของเครื่อง (Machine Learning)

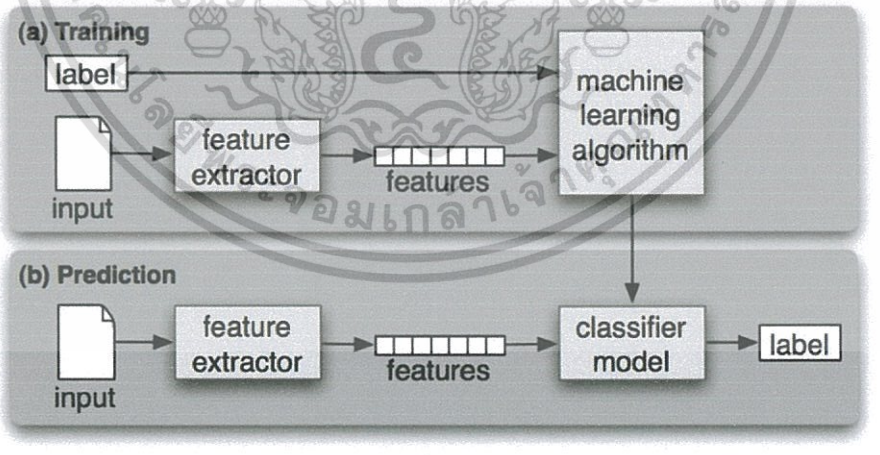
การเรียนรู้ของเครื่อง เป็นสาขาหนึ่งของปัญญาประดิษฐ์ (Artificial Intelligence)ที่เกี่ยวข้องกับการพัฒนาเทคนิควิธีเพื่อให้คอมพิวเตอร์สามารถเรียนรู้ โดยเน้นวิธีการสร้างตรรกะของคอมพิวเตอร์จากการวิเคราะห์ชุดข้อมูล การเรียนรู้จึงเกี่ยวข้องอย่างมากกับวิชาสถิติศาสตร์ การเรียนรู้ของเครื่องถูกใช้เพื่อเพิ่มประสิทธิภาพในการแก้ปัญหาด้านต่าง ๆ เช่น การสร้างให้คอมพิวเตอร์ สามารถแยกแยะวัตถุ เสียงหรือตัวอักษรได้ หรือจำแนกข้อมูลจำนวนมากที่ไม่สามารถทำได้โดยมนุษย์เป็นต้น

กล่าวได้ว่าการเรียนรู้ของเครื่องนั้นจะไม่มีการกำหนดเงื่อนไขตายตัวใด ๆ ลงในโปรแกรมเพื่อทำการจัดการข้อมูล แต่จะใช้วิธีการวิเคราะห์และหาความสัมพันธ์ของข้อมูลเหล่านั้น แล้วสร้างวิธีการจัดการตอบสนองต่อข้อมูลนั้น ๆ ขึ้นมาเอง ซึ่งเมื่อโปรแกรมมีความสามารถในการจัดการและตอบสนองกับข้อมูลที่รับเข้ามาด้วยตัวเองมนุษย์จึงไม่ต้องมาคอยศึกษา วิเคราะห์ และปรับแก้ไขโปรแกรมใหม่ทุกครั้ง เพื่อจัดการกับข้อมูลรูปแบบใหม่ ๆ ที่ถูกเพิ่มเข้ามาอีกในอนาคต

ซึ่งลักษณะทั่วไปของการเรียนรู้ของเครื่องเป็นการสร้างอัลกอริทึม หรือ โปรแกรมคอมพิวเตอร์จากการให้ข้อมูลฝึก (Training data) สำหรับสอนให้คอมพิวเตอร์เรียนรู้เพื่อให้ได้แบบจำลองในการทำนายเพื่อนำมาใช้วิเคราะห์และหาความสัมพันธ์ของข้อมูลเหล่านั้น

โดยทั่วไปแล้วขั้นตอนวิธีในการเรียนรู้ของเครื่อง [13] มีกระบวนการทำอยู่ 3 ขั้นตอนในการสร้างแบบจำลองของปัญหาที่ใช้ในการวิเคราะห์คือ

1) Feature extraction หรือการดึงคุณลักษณะเป็นการเตรียมข้อมูลก่อนที่จะใช้สร้างแบบจำลองสำหรับปัญหาที่ต้องการ โดยเป็นขั้นตอนการแปลงลักษณะเด่นของข้อมูลที่ต้องการใช้ในการทำกระบวนการเรียนรู้ของเครื่องให้อยู่ในรูปแบบที่ใช้งานได้ เช่นการแปลงข้อมูลจากรูปแบบที่เป็นข้อความหรือรูปภาพให้อยู่ในรูปแบบของตัวเลข วิธีการนี้จะช่วยลดขนาดของข้อมูลที่ต้องใช้ในขั้นตอนการประมวลผลลง ซึ่งมีความสำคัญมากหากข้อมูลที่นำเข้ามีจำนวนมหาศาล
2) Regularization เป็นการให้ค่าน้ำหนักและความสำคัญของข้อมูลที่จะใช้ในการสร้างแบบจำลอง โดยเป็นขั้นตอนที่เมื่อเราได้ทำการดึงคุณลักษณะที่เราต้องการมาจากข้อมูลแล้วมาทำการพิจารณาว่าอะไรที่เป็นคุณสมบัติที่สำคัญต่อปัญหา และส่งผลกระทบต่อการวิเคราะห์ข้อมูล ซึ่งทำให้เราลดข้อมูลที่ไม่จำเป็นออกไปเพื่อให้ได้แบบจำลองที่มีความเรียบง่าย
3) Cross-validation เป็นการตรวจสอบความถูกต้องของแบบจำลอง โดยเป็นขั้นตอนที่หลังจากเราสร้างแบบจำลองที่ใช้ในการทำนายขึ้นมาแล้วนำมาทดสอบว่าแบบจำลองที่สร้างนั้นสามารถทำนายข้อมูลได้อย่างถูกต้องแม่นยำ มีประสิทธิภาพตรงกับความต้องการของเราหรือไม่ หนึ่งในวิธีการนี้คือ Out-Of-Time (OOT) testing ซึ่งจะทดสอบการทำงานของแบบจำลองจากข้อมูลที่ไม่เคยเห็นมาก่อน หรือข้อมูลที่ไม่เคยถูกนำมาใช้ในการฝึกการเรียนรู้เพื่อสร้างแบบจำลองนั้น



รูป 2.4 การทำงานของ Machine Learning เพื่อสร้างแบบจจำลองในการวิเคราะห์ปัญหา

ในโครงงานนี้เราทำการเลือกเทคนิควิธีการในการเรียนรู้ของเครื่องเพื่อนำมาสร้างแบบจำลองของข้อมูลที่ใช้ในการทำนายอารมณ์ของข้อความในขั้นต้น 2 เทคนิคคือ

2.6.1 Naive-Bayes

Naive-Bayes [14] เป็นขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน(Supervised Learning) สร้างขึ้นจากหลักการความน่าจะเป็นซึ่งจะใช้วิเคราะห์หาความน่าจะเป็นของสิ่งที่ยังไม่เคยเกิดขึ้น เป็นเทคนิคในการแก้ปัญหาแบบการจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ การเรียนรู้แบบเบย์อย่างง่ายเป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่งมีขั้นตอนในการทำงานที่ไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมากและคุณสมบัติ(Attribute) ของตัวอย่างไม่ขึ้นต่อกัน

โดยที่เทคนิคนี้ทำงานอยู่บนพื้นฐานทฤษฎีของเบย์ (Bayes' theorem) กับสมมติฐานของความเป็นอิสระในการทำนาย โดยวิธีการคำนวณความน่าจะเป็นดังสมการดังต่อไปนี้

$$P(c|x) = \frac{P(x|c) \times P(c)}{P(x)} \qquad (2.1)$$

จากสมการ 2.1 ของ Bayes จะมี 4 ส่วนที่สำคัญ คือ

1) Posterior probability หรือ P(c|x) คือ ค่าความน่าจะเป็นที่ข้อมูลที่มีแอตทริบิวต์เป็น x จะมีคลาส c
2) Prior probability หรือ P(c) คือ ค่าความน่าจะเป็นของคลาส C
3) Likelihood หรือ P(x|c) คือ ค่าความน่าจะเป็นที่ข้อมูล training data ที่มีคลาส C และมีแอตทริบิวต์ x
4) Prior หรือ P(x) คือ ค่าความน่าจะเป็นของการทำนาย

2.6.2 Multinomial Naïve Bayes

Multinomial Naïve Bayes [4] มีการปรับปรุงและพัฒนามาจากอัลกอริทึม Naïve Bayes ใช้สำหรับข้อมูลที่มีการกระจายตัวนิยมใช้ในการทำงานเกี่ยวกับการจำแนกข้อความ (Text Classification) มีการนับความถี่ของคำที่เกิดขึ้นในเอกสาร

สมมติให้ $n_1, n_2, ..., n_k$ เป็นจำนวนของคำที่เกิดขึ้นในเอกสารของคำแต่ละคำ และให้ $P_1, P_2, ..., P_k$ เป็นค่าความน่าจะเป็นของแต่ละคำเมื่อทำการสุ่มตัวอย่างจากเอกสารทั้งหมดในหมวดหมู่เอกสาร H สมมติว่าความน่าจะเป็นเป็นอิสระจากบริบทของคำและตำแหน่งในเอกสารซึ่งสมมติฐานเหล่านี้นำไปสู่การกระจายตัวแบ multinomial สำหรับความน่าจะเป็นของเอกสาร

ในส่วนของการกระจายตัวนี้ความน่าจะเป็นของเอกสาร E ขึ้นอยู่กับคลาส H ในคำอื่น สูตรสำหรับคำนวนความน่าจะเป็น คือ

$$P(E|H) \approx N! \times \prod_{i=1}^{k} \frac{P_i^{n_i}}{n_i} \qquad (2.2)$$

ซึ่ง $N = n1 + n2 + \ldots + nk$ เป็นจำนวนรวมของคำในเอกสาร

P_i เป็นค่าประมาณโดยคำนวนจากความถี่ที่เกี่ยวข้องของคำแต่ละคำ (i) ที่พบในข้อความของเอกสารที่ใช้สำหรับการเทรนทั้งหมดที่เกี่ยวข้องกับหมวดหมู่ H

2.7 งานวิจัยที่เกี่ยวข้อง

2.7.1 Emotion Classification of Thai Text based Term weighting and Machine Learning Techniques

ในงานวิจัยนี้ [1] พูดถึงวิธีการที่จะสร้างการวิเคราะห์หาอารมณ์จากข้อความจากการแสดงความคิดเห็นต่าง ๆ ที่ปรากฏอยู่บนสื่อสังคมออนไลน์ที่ผู้คนมักจะมาแสดงความคิดเห็นกันในโดยทำการเก็บจากเว็บไซต์ของเมืองไทยซึ่งเป็นการวิเคราะห์ในส่วนของข้อความที่มาจากภาษาไทย มีการนำเรื่องการประมวลผลภาษาธรรมชาติ และเทคนิคในด้านการเรียนรู้ของเครื่องมาใช้ในการสร้างโมเดลในการทำนายมีการทดลองใช้เทคนิค 4 เทคนิคด้วยกันคือ Support Vector Machine Naive Bayes Decision Tree และ K-Nearest Neighbor เพื่อทำการทดสอบประสิทธิภาพจากการทำนายอารมณ์

ในโครงงานนี้เราได้ทำการทดลองโดยมีการนำเทคนิค Naive Bayes มาใช้และทำการเพิ่มเติมในส่วนของเทคนิค Multinomial Naive Bayes มาทดสอบเพิ่มเป็นเทคนิคที่ 2 สิ่งที่แตกต่างคือในงานวิจัยนี้ทำการตัดคำที่เก็บมาจากความคิดเห็นในสื่อสังคมออนไลน์โดยเริ่มตั้งแต่กระบวนการแรกที่จะต้องหาหน่วยคำที่เล็กที่สุดถึงจะทำการตัดเป็นคำออกมาได้ แต่ในโครงงานนี้ได้มีการประยุกต์ใช้เครื่องมือที่เป็นตัวช่วยในการตัดคำที่มีผู้พัฒนาสร้างไว้โดยในเครื่องมือที่นำมาใช้นี้อาศัยคลังข้อมูลภาษาจากทาง NECTEC เก็บรวบรวมไว้ให้ซึ่งทำให้มั่นใจได้ในระดับหนึ่งว่าจะช่วยในเรื่องของการตัดคำให้ออกมาอย่างมีประสิทธิภาพ

2.7.2 Is Naïve Bayes a Good Classifier for Document Classification?

ในงานวิจัยนี้ [2] พูดถึงการทำ classification ของเอกสารในเรื่องของการจำแนกให้อยู่ใน 4 ประเภทคือ กีฬา ธุรกิจ การเมือง และการท่องเที่ยว เนื้อหาทั้งหมดที่ใช้ในการทำโมเดลจะเป็นภาษาอังกฤษทั้งหมด ซึ่งในงานวิจัยนี้มุ่งเน้นไปยังการหาประสิทธิภาพโดยใช้วิธีของ Naïve Bayes ซึ่งเป็นวิธี classification ที่เหมาะกับการทำ classification ของเอกสาร ซึ่งผลที่ได้ออกมา วิธีของ Naïve Bayes มีประสิทธิภาพมากที่สุดเมื่อเทียบกับ classification วิธีอื่น ๆ เช่น decision tree neural network และ support vector machines (ในเรื่องของความแม่นยำ และประสิทธิภาพในการคำนวณ)

ในโครงงานนี้มีการเลือกเทคนิค Naïve Bayes มาใช้ จากงานวิจัยนี้ทำให้มั่นใจได้ว่าการใช้เทคนิค Naïve Bayes น่าจะให้ผลลัพธ์ที่มีประสิทธิภาพ โดยสิ่งที่แตกต่างคือข้อมูลที่ผู้พัฒนาใช้จะเป็นภาษาไทยและมีเป้าหมายในการจำแนกที่ต่างกัน

บทที่ 3
การวิเคราะห์และออกแบบระบบ

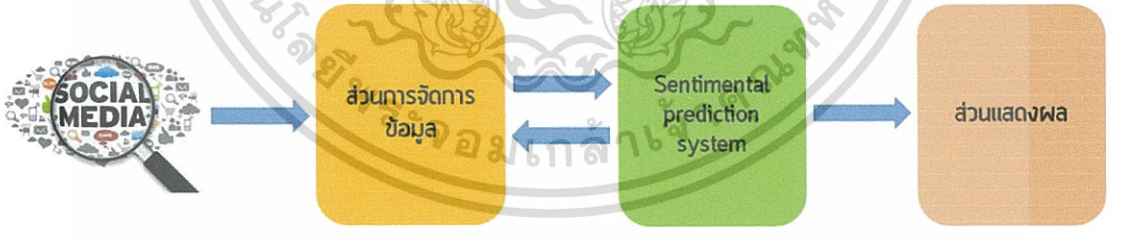
3.1 ภาพรวมของระบบ

ระบบทำนายอารมณ์บริเวณสถานที่จากสื่อสังคมออนไลน์เป็นระบบที่มีการนำข้อมูลจากสื่อสังคมออนไลน์ในรูปแบบของข้อความที่มีการแสดงความคิดเห็นเกี่ยวกับสถานที่นั้น ๆ มาทำการเก็บไปประมวลผลเพื่อที่จะมีการทำนายอารมณ์ที่แสดงต่อบริเวณหรือสถานที่ดังกล่าว เพื่อที่จะเป็นการช่วยตัดสินใจที่จะเดินทางไปหรือหลีกเลี่ยงการเดินทาง

ในการทำนายอารมณ์โดยอาศัยบริเวณที่สนใจนั้นระบบจะทำการเก็บข้อมูลความคิดเห็นในรูปแบบข้อความจากสื่อสังคมออนไลน์ก่อน โดยสร้างให้มีการรองรับข้อมูลที่มีปริมาณมากและอำนวยความสะดวกในการจัดการข้อมูลที่ได้มา หลังจากนั้นคือการนำไปเข้าสู่ขั้นตอนในการประมวลผลซึ่งต้องมีการนำข้อมูลที่เก็บในฐานข้อมูลมาทำการวิเคราะห์ข้อความด้วยกระบวนการเรียนรู้ของเครื่องเพื่อทำการทำนายอารมณ์จากสถานที่ในบริเวณที่สนใจผ่านส่วนแสดงผล

โดยภาพรวมของระบบแบ่งออกเป็น 3 ส่วนดังนี้

1) ส่วนการจัดการข้อมูล
2) ส่วนการประมวลผล (Sentimental prediction system)
3) ส่วนแสดงผล



รูป 3.1 แผนผังการทำงานโดยรวมของระบบ

3.1.1 ส่วนการจัดการข้อมูล

ทำหน้าที่ในการดึงข้อมูลจากสื่อสังคมออนไลน์คือ Facebook Twitter และ Foursquare ผ่าน API ของทางผู้ให้บริการที่มีให้เช่น Graph API และ Twitter REST API เป็นต้น โดยข้อมูลที่นำเข้ามาจะเป็นข้อความที่ผู้ใช้งานได้โพสต์บนสื่อสังคมออนไลน์ซึ่งเป็นข้อความที่สื่อถึงสถานที่หนึ่ง ๆ หรือข้อความในบริเวณที่กำหนด ซึ่งนำเข้ามาจัดเก็บในฐานข้อมูลที่ทำการออกแบบไว้เพื่อเก็บข้อมูลที่เราจำเป็นต้องใช้ในการนำไปวิเคราะห์เพื่อสร้างแบบจำลองในการทำนายอารมณ์และข้อมูลที่จะถูกนำไปทำนาย

3.1.2 ส่วนการประมวลผล (Sentimental prediction system)

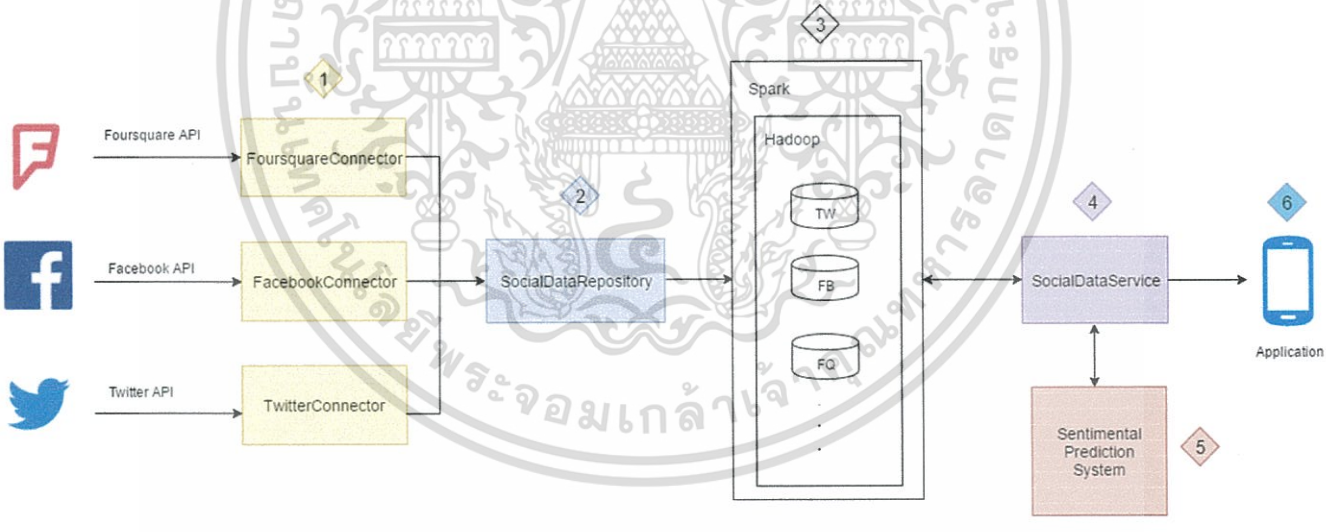
ทำหน้าที่ในการนำข้อมูลที่มีการแสดงความคิดเห็นจากสถานที่ที่ต้องการจากส่วนจัดการข้อมูลที่มีการเก็บรวบรวมไว้มาทำการสร้างคุณลักษณะ (Feature) เพื่อใช้สร้างแบบจำลองในการทำนายผ่านกระบวนการเรียนรู้ของเครื่อง (machine learning) เพื่อที่จะทำนายอารมณ์ที่คนส่วนมากแสดงออกมาในบริเวณพื้นที่ที่เราสนใจ

3.1.3 ส่วนแสดงผล

ทำหน้าที่แสดงผลบริเวณสถานที่ต่าง ๆ ณ ช่วงเวลาที่เราเข้าใช้งานซึ่งมีการแสดงออกเป็นสีแทนอารมณ์ต่าง ๆ พร้อมทั้งสัญลักษณ์ที่สื่อถึงอารมณ์นั้น โดยเลือกแสดงจากอารมณ์ที่มีค่ามากที่สุดในเวลานั้นมีการแบ่งระดับอารมณ์ออกเป็น 8 ระดับอารมณ์ตามทฤษฎีของโรเบิร์ต พูลทชิค [11,17] ซึ่งแสดงให้อยู่ในรูปแบบที่ผู้ใช้งานเข้าใจง่าย

3.2 โครงสร้างของระบบ

ระบบที่นำเสนอมีโครงสร้างดังต่อไปนี้



รูป 3.2 โครงสร้างของระบบ

จากรูป 3.2 โครงสร้างของระบบสามารถแบ่งออกได้เป็น 6 ส่วนหลักคือ

3.2.1 ส่วนที่ติดต่อกับสื่อสังคมออนไลน์

จากรูป 3.2 จะเป็นส่วนหมายเลข 1 ซึ่งในส่วนนี้จะมีบริการอยู่ 3 บริการ คือ FacebookConnector FoursquareConnector และ TwitterConnector ซึ่งมีไว้ติดต่อกับ API ของสื่อสังคมออนไลน์เพื่อดึงข้อมูลเข้ามาเก็บในฐานข้อมูล

3.2.1.1 การนำเข้าข้อมูล Facebook ซึ่งใช้ Graph API โดยเลือก API ดังนี้

1) ค้นหาข้อมูลเพจ graph.facebook.com/{page-id}
2) ค้นหาข้อมูลโพสต์ในเพจ graph.facebook.com/{page-id}/feed
3) ค้นหาข้อมูลคอมเมนต์ในเพจ graph.facebook.com/{page-id}/comment

3.2.1.2 การนำเข้าข้อมูล Foursquare ซึ่งใช้ Foursquare API โดยเลือก API ดังนี้

1) ค้นหาสถานที่ https://api.foursquare.com/v2/venues/search
2) ค้นหาจำนวนคนที่อยู่ขณะนี้
   https://api.foursquare.com/v2/venues/VENUE_ID/herenow
3) ค้นหาสถานที่ถัดไปจากสถานที่ปัจจุบัน
   https://api.foursquare.com/v2/venues/VENUE_ID/nextvenues
4) ค้นหาคำแนะนำเกี่ยวกับสถานที่
   https://api.foursquare.com/v2/venues/VENUE_ID/tips
5) ค้นหาสถานที่ที่ได้รับความนิยมในบริเวณที่กำหนด
   https://api.foursquare.com/v2/venues/trending
6) ค้นหาช่วงเวลาที่มีจำนวนคนที่มากที่สุดในสถานที่นั้น
   https://api.foursquare.com/v2/venues/VENUE_ID/hours-
7) ค้นหาข้อมูลเกี่ยวกับสถานที่
   https://api.foursquare.com/v2/venues/VENUE_ID
8) ค้นหารูปภาพที่เกี่ยวกับสถานที่
   https://api.foursquare.com/v2/venues/VENUE_ID/photos

3.2.1.3 การนำเข้าข้อมูล Twitter ผ่านทาง REST APIs ของ Twitter โดยเลือกใช้ API ดังนี้

1) ค้นหาทวีต
   https://api.twitter.com/1.1/search/

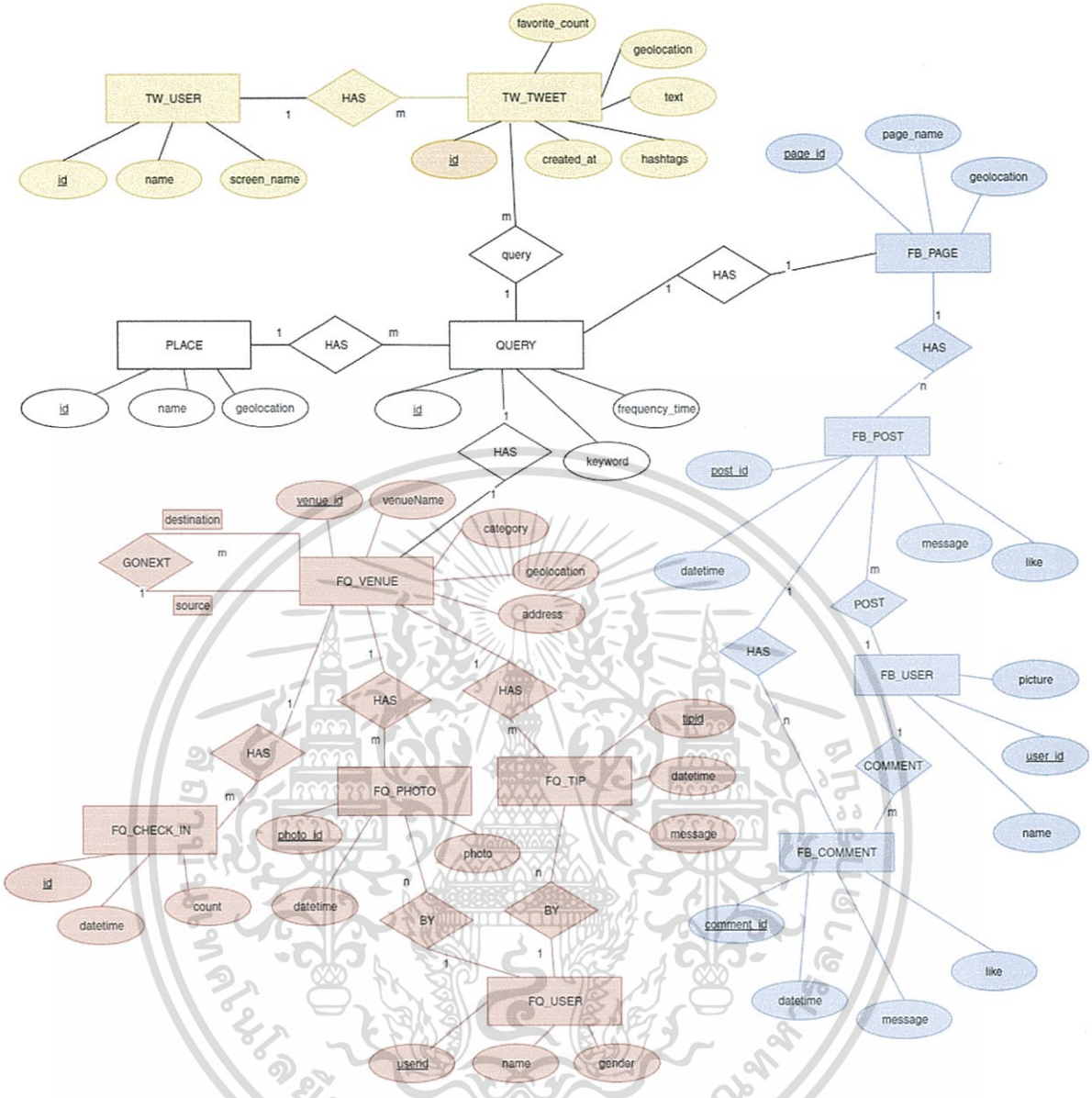
การดึงข้อมูลจากสื่อสังคมออนไลน์ในระบบนี้จะทำการดึงแบบอัตโนมัติเพื่อที่ให้ได้ข้อมูลที่เป็นปัจจุบันมากที่สุด โดยการดึงแบบอัติโนมัติจะใช้ไลบารี่ node-cron ของ kelektiv ซึ่งเป็นไลบารี่ของ Node.js ทำให้สามารถตั้งเวลาในการรันและควบคุมงาน (task) ได้ง่ายขึ้น โดยการดึงข้อมูลแบบอัตโนมัติจะตั้งเวลาไว้ทุก 5 นาที เพื่อที่จะไม่ให้เกินขอบเขตของการส่งคำขอ (API Rate Limit)

3.2.2 ส่วนจัดการเก็บข้อมูลสื่อสังคมออนไลน์

จากรูป 3.2 จะเป็นส่วนหมายเลข 2 ซึ่งในส่วนนี้จะมีตัว Social Data Repository เป็นศูนย์กลางในการจัดการข้อมูลทั้งหมดที่เข้ามาเพื่อทำให้อยู่ในรูปแบบที่เหมาะสมและเก็บข้อมูลลงสู่ฐานข้อมูล ซึ่งการทำให้อยู่ในรูปแบบที่เหมาะสมคือการทำให้ข้อมูลจากหลายแห่งเช่น Facebook และ Twitter ทำให้เป็นรูปแบบมาตรฐานกลางในการจัดเก็บข้อมูล และเก็บข้อมูลที่เป็นข้อมูลดิบด้วยเช่นกัน ในส่วนนี้จะใช้ Apache Spark ในการประมวลผลข้อมูลเพื่อที่จะทำการเก็บข้อมูลลง Hadoop ซึ่ง Spark มีความสามารถในการประมวลผลแบบกลุ่มจึงเหมาะกับการจัดการข้อมูลที่มีขนาดใหญ่

3.2.3 ส่วนฐานข้อมูล

จากรูป 3.2 จะเป็นส่วนหมายเลข 3 ซึ่งในส่วนนี้จะใช้ HDFS เป็นตัวเก็บข้อมูลเพื่อรองรับข้อมูลที่มีปริมาณมากและเพิ่มขึ้นอย่างต่อเนื่อง โดยในระบบจะเก็บข้อมูลจาก Twitter Facebook และ Foursquare ซึ่งในรูป 3.3 เป็น ER Diagram ที่ออกแบบเพื่อเป็นฐานข้อมูลที่จะใช้เก็บข้อมูล



รูป 3.3 ER Diagram ของการจัดเก็บข้อมูลสื่อสังคมออนไลน์

โดยในแต่ละ Entity Type มีคำอธิบายดังนี้

ตาราง 3.1 คำอธิบาย Entity Type ของ Twitter

Entity Type	คำอธิบาย
TW_TWEET	ทวีตที่ผู้ใช้งาน Twitter โพสต์
TW_USER	บัญชีผู้ใช้งาน Twitter

ตาราง 3.2 คำอธิบาย Entity Type ของ Facebook

Entity Type	คำอธิบาย
FB_PAGE	เพจบน Facebook
FB_POST	โพสต์ที่อยู่บนหน้าเพจ
FB_COMMENT	ความคิดเห็นในแต่ละโพสต์
FB_USER	บัญชีผู้ใช้งานหรือบัญชีเพจ Facebook

ตาราง 3.3 คำอธิบาย Entity Type ของ Foursquare

Entity Type	คำอธิบาย
FQ_VENUE	สถานที่ใน Foursquare
FQ_CHECKIN	จำนวนคนที่อยู่ในแต่ละสถานที่
FQ_POPULARHOUR	ช่วงเวลายอดนิยมของแต่ละสถานที่
FQ_PHOTO	รูปภาพที่ผู้ใช้งานถ่ายในแต่ละสถานที่
FQ_TIP	ความคิดเห็นของผู้ใช้ในแต่ละสถานที่
FQ_USER	บัญชีผู้ใช้งาน Foursquare

ตาราง 3.4 คำอธิบาย Entity Type ของข้อมูลส่วนกลาง

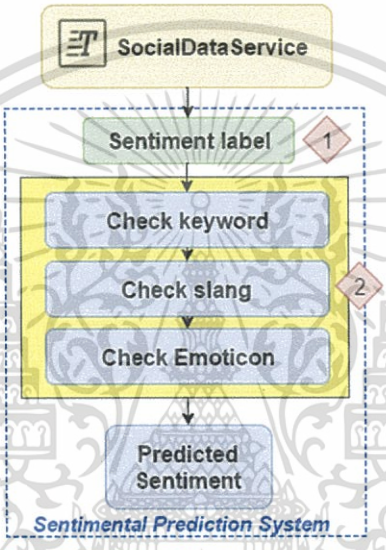
Entity Type	คำอธิบาย
QUERY	Keyword ที่ใช้หาข้อมูลบนสื่อสังคมออนไลน์
PLACE	สถานที่ตาม keyword

3.2.4 ส่วนให้บริการข้อมูลสื่อสังคมออนไลน์

จากรูป 3.2 จะเป็นส่วนหมายเลข 4 ซึ่งในส่วนนี้จะมีตัว SocialDataService ซึ่งเป็นบริการที่นำข้อมูลจากฐานข้อมูลไปให้ส่วนทำนายอารมณ์เพื่อวิเคราะห์ และทำนายอารมณ์ออกมา และเป็นบริการที่นำข้อมูลจากฐานข้อมูลไปให้ส่วนแสดงผลหรือแอปพลิเคชัน โดยในตัวบริการนี้จะทำเป็น REST API เพื่อความสะดวกในการติดต่อระหว่างตัวบริการกับแอปพลิเคชันที่เข้ามาติดต่อด้วย

3.2.5 ส่วนทำนายอารมณ์โดยใช้ข้อความจากสื่อสังคมออนไลน์

จากรูป 3.2 จะเป็นส่วนหมายเลข 5 เป็นส่วนโครงสร้างหลักที่สำคัญของระบบซึ่งใช้ในการทำนายอารมณ์ข้อความที่ได้จากสื่อสังคมออนไลน์โดยภายในประกอบด้วยโครงสร้างการทำงานดังรูป 3.3 และ 3.4 ซึ่งการทำงานจะดึงข้อมูลจากส่วนให้บริการข้อมูลจากสื่อสังคมออนไลน์ทุก ๆ 5 นาที ตามเวลาที่มีการดึงข้อมูลจากสื่อสังคมออนไลน์ลงมาเก็บไว้ในระบบซึ่งข้อมูลที่ใช้จะเป็นข้อมูลตั้งแต่ปัจจุบันที่เข้าใช้งานระบบรวมทั้งข้อมูลย้อนกลับไป 3 ชั่วโมง



รูป 3.4 กระบวนการทำงานของ Rule-based method ในส่วน Sentimental Prediction System

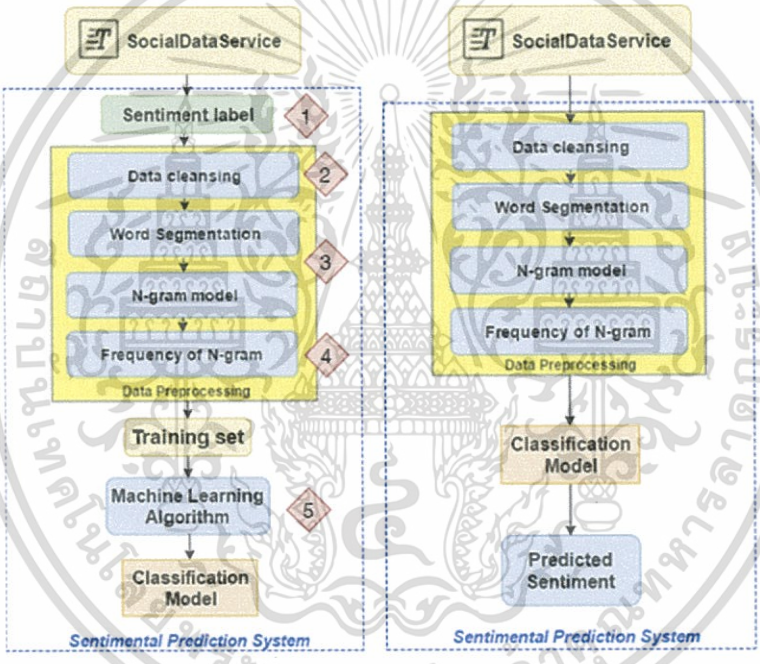
จากรูป 3.4 ประกอบด้วยส่วนของกระบวนการทำงานดังนี้

1) ส่วน Sentiment label (หมายเลข 1) ส่วนนี้คือส่วนการกำหนดอารมณ์ให้กับข้อความที่เราจะเอามาใช้ในการ training โดยในส่วนนี้จะใช้คนกำหนดอารมณ์ตามอารมณ์ที่กำหนดไว้
2) ส่วน Check keyword, Check slang และ Check Emoticon (หมายเลข 2) ส่วนนี้เป็นส่วนที่ทำการตรวจสอบว่า keyword slang หรือ emoticon ที่กำหนดไว้นั้นปรากฏอยู่ในข้อความนั้นหรือไม่ โดยเราจะมีการกำหนดกลุ่มของ keyword slang และ emoticon ของแต่ละอารมณ์ขึ้นมา ซึ่งถ้าเกิดมีปรากฏอยู่ในข้อความนั้น ข้อความนั้นก็จะถูกกำหนดอารมณ์ตามกลุ่มอารมณ์ที่ keyword slang หรือ emoticon นั้นอยู่ กรณีที่เกิดขึ้นได้มี 3 กรณีคือ

2.1) ในข้อความมีอารมณ์ปรากฏเพียงอารมณ์เดียว โดยถ้าเป็นกรณีนี้จะสามารถใช้อารมณ์นั้นได้ทันที

2.2) ในข้อความมีอารมณ์ปรากฏมากกว่า 1 อารมณ์ ถ้าเป็นกรณีนี้จะมีการเลือกอารมณ์ที่มีความถี่ของการเกิดมากที่สุดมาแสดงเป็นอารมณ์ของข้อความนั้น หรือถ้าหากความถี่ของอารมณ์เท่ากัน จะเรียงลำดับตามอารมณ์ที่มีจำนวน training set จากมากไปน้อย

2.3) ในข้อความไม่มีอารมณ์ใดปรากฏอยู่เลย ถ้าเป็นกรณีจะมีการกำหนดอารมณ์พื้นฐานไว้ ซึ่งถ้าไม่มีอารมณ์ใกปรากฏอยู่ในข้อความนั้นเลย ก็จะใช้อารมณ์ที่พบมากที่สุดเป็นคำตอบแทน



ก) ข)

รูป 3.5 กระบวนการทำงานของอัลกอริทึม Naïve Bayes และ Multinomial Naïve Bayes ในส่วน Sentimental Prediction System

ก) การสร้างโมเดล

ข) การนำโมเดลไปใช้

จากรูป 3.5 ประกอบด้วยส่วนของกระบวนการทำงานดังนี้

1) ส่วน Sentiment label (หมายเลข 1) ส่วนนี้คือส่วนการกำหนดอารมณ์ให้กับข้อความที่เราจะเอามาใช้ในการ training โดยในส่วนนี้จะเหมือนกับส่วนหมายเลข 1 ของรูป 3.4

2) ส่วน Data cleansing (หมายเลข 2) ส่วนนี้คือส่วนการทำความสะอาดข้อมูล ก่อนที่จะนำข้อความเข้าสู่กระบวนการทำ Word Segmentation จำเป็นต้องมีการทำความสะอาดข้อความก่อนเช่น ลบอักขระที่ไม่ใช่ภาษาไทยและช่องว่างออกจากข้อความให้หมด และลบตัวอักษรที่ซ้ำกันที่ไม่มีความหมายเช่น มากกกกก ลบให้เหลือ มาก เป็นต้น

3) ส่วน Word Segmentation และ N-gram (หมายเลข 3) ส่วนนี้เป็นส่วนของการตัดคำและการใช้กระบวนการ N-gram โดยการตัดคำนั้นได้ใช้ไลบารี่ของ pythainlp ซึ่งเป็นไลบารี่ภาษา python ซึ่งผลที่ได้จะเป็นอาเรย์ของคำที่ตัดมาได้ หลังจากนั้นจะนำเอาคำที่ได้จากการตัดคำมาเข้าสู่กระบวนการ N-gram โดยที่เราเลือกใช้มี 3 ประเภทคือ unigram bigram และ trigram โดยผลลัพธ์ที่ได้จะเป็นจะเป็นอาเรย์ของคำที่ผ่านกระบวนการ N-gram ซึ่งเราจะนำเอาคำเหล่านี้มาเป็นคุณลักษณะ (Feature) ที่จะนำไปใช้กับกระบวนการเรียนรู้ของเครื่อง (Machine learning)

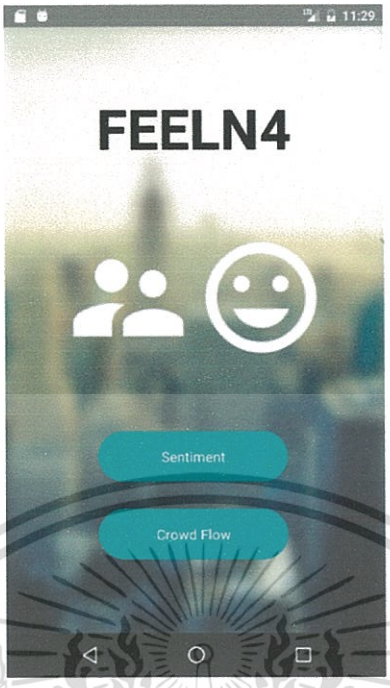
4) ส่วน Frequency of N-gram (หมายเลข 4) ส่วนนี้เป็นส่วนของการนับความถี่ของคำจากข้อความที่นำมาใช้เป็น training set ซึ่งคำต้นแบบที่ใช้นับความถี่เป็นที่ได้มาจากกระบวนการ Word Segmentation และ N-gram (หมายเลข 3)

5) ส่วน Machine Learning Algorithm (หมายเลข 5) ส่วนนี้เป็นส่วนนำข้อมูลที่ได้มาจากกระบวนการ Frequency of N-gram (หมายเลข 4) มาเข้าสู่กระบวนการนี้เพื่อสร้างแบบจำลองที่จะใช้ในการทำนายอารมณ์ ซึ่งอัลกอริทึมที่เลือกใช้คือ Naïve Bayes และ Multinomial Naïve Bayes

3.2.6 ส่วนแสดงผลสำหรับผู้ใช้งาน

จากรูป 3.2 จะเป็นส่วนหมายเลข 6 ซึ่งส่วนนี้เป็นส่วนแสดงผลให้ผู้ใช้งานได้เห็นว่าผู้คนที่อยู่ ณ ที่นั้นแสดงอารมณ์หรือความรู้สึกเป็นอย่างไร โดยเลือกแสดงจากอารมณ์ที่มีค่ามากที่สุดในช่วงเวลานั้นมีการแบ่งระดับอารมณ์ซึ่งแทนด้วยอารมณ์ 8 อารมณ์ด้วยกันคือ joy (มีความสุข สนุกสนาน) sadness (เศร้า เสียใจ) fear (กลัว) angry (โกรธ) disgust (รังเกียจ) surprise (ประหลาดใจ) anticipation (คาดหวัง สนใจ) acceptance (ยอมรับ)

ในการสร้างแอปพลิเคชันจะใช้ React Native เป็น framework ของ Facebook ซึ่งทำให้สามารถใช้ภาษา Javascript ในการสร้างแอปพลิเคชันลงบนหลายแพลตฟอร์ม (Platform) เช่น ระบบปฏิบัติการ iOS และ Android เป็นต้น

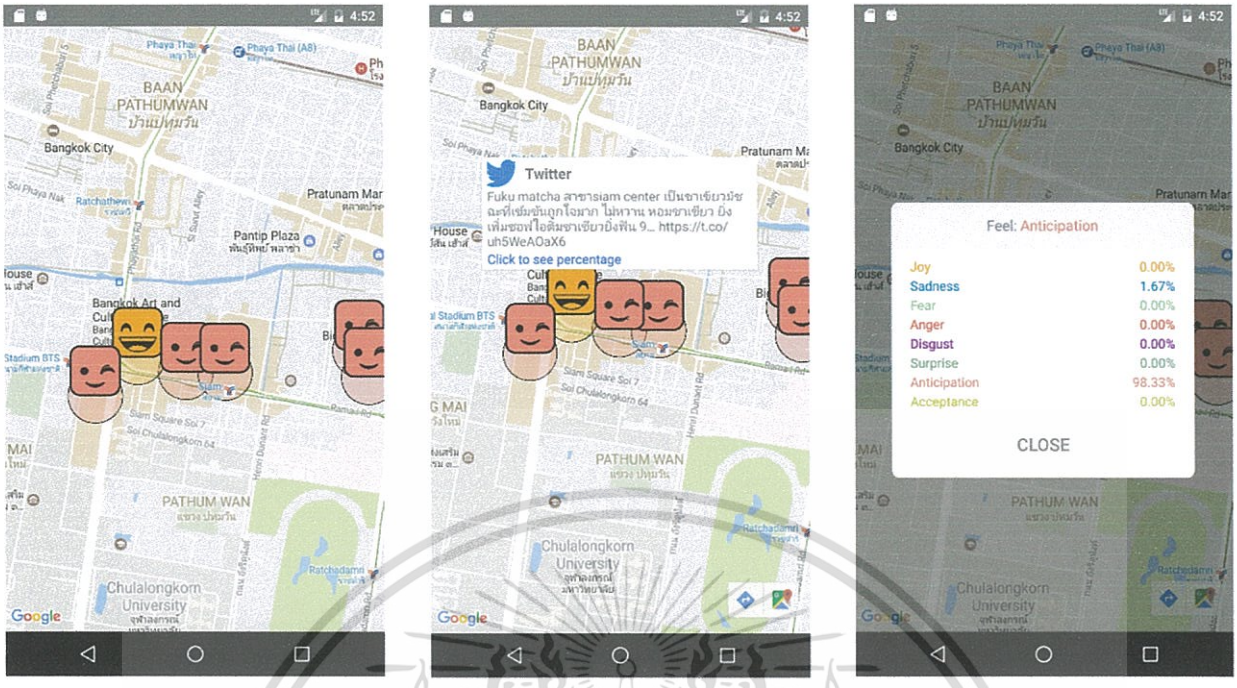


รูป 3.6 หน้าแรกของแอปพลิเคชัน

จากรูป 3.6 จะเป็นหน้าแรกเมื่อผู้ใช้งานเรียกใช้แอปพลิเคชันเมื่อผู้ใช้งานทำการเลือกปุ่มในส่วนของ Sentiment จะพบกับหน้าหลักของระบบดังรูป 3.6 ในหน้านี้จะแสดงสัญลักษณ์ของอารมณ์ในภาพรวมว่าสถานที่นั้นมีลักษณะอารมณ์เป็นอย่างไร

โดยที่สัญลักษณ์แทนอารมณ์ที่ใช้ทั้งหมดมีดังนี้

รูป 3.7 สัญลักษณ์ที่ใช้แทนอารมณ์ทั้ง 8 อารมณ์



ก) ข) ค)

รูป 3.8 หน้าหลักของแอปพลิเคชัน

ก) เมื่อเริ่มต้นใช้งาน

ข) เมื่อกดที่สัญลักษณ์

ค) เมื่อกดแสดงเปอร์เซ็นต์

จากรูป 3.8 เมื่อทำการกดเลือกไปที่สัญลักษณ์ของอารมณ์ (ก) จะปรากฏข้อความที่แสดงให้เห็นถึงที่มาของอารมณ์ในภาพรวมของสถานที่ดังกล่าว (ข) และสามารถที่จะเข้าไปดูค่าเปอร์เซ็นต์ของแต่ละอารมณ์ว่าสถานที่นั้นมีค่าอยู่ที่เท่าไร (ค)

บทที่ 4

การทดลองและผลการทดลอง

4.1 ชุดข้อมูลที่ใช้ในการทดลอง

ข้อมูลที่นำมาใช้ในการทดลองหลังจากเก็บรวบรวมมาจากแหล่งต่าง ๆ ต้องทำการเลือกข้อมูลที่สามารถนำมาใช้ได้ในขั้นต้นก่อนหลังจากนั้นนำข้อมูลที่เลือกมาทำการระบุลักษณะของข้อความนั้นว่ามีการแสดงความรู้สึกอย่างไรโดยทำการระบุอารมณ์เป็น 8 อารมณ์คือ Joy, Sadness, Fear, Anger, Disgust, Surprise, Anticipation, Acceptance ประกอบด้วยข้อมูลจำนวน 3 ชุดคือ

1) ข้อมูลชุด A จำนวน 3,429 ข้อความ
   ช่วงระยะเวลาที่ใช้ในการเก็บข้อมูลเดือนพฤศจิกายน 2559
2) ข้อมูลชุด B จำนวน 11,193 ข้อความ
   ช่วงระยะเวลาที่ใช้ในการเก็บข้อมูลเดือนพฤศจิกายน - ธันวาคม 2559
3) ข้อมูลชุด C จำนวน 14,622 ข้อความ
   ช่วงระยะเวลาที่ใช้ในการเก็บข้อมูลเดือนพฤศจิกายน 2559 - มกราคม 2560

ตาราง 4.1 จำนวนข้อความที่ระบุอารมณ์ของข้อมูลทั้ง 3 ชุด

อารมณ์พื้นฐานของ Robert Plutchik	จำนวนข้อความ		
	ข้อมูลชุด A	ข้อมูลชุด B	ข้อมูลชุด C
Joy	715	2,960	3,675
Sadness	377	1,341	1,718
Fear	52	634	686
Anger	279	1,352	1,631
Disgust	42	151	193
Surprise	238	404	642
Anticipation	1,223	3,168	4,391
Acceptance	503	1,183	1,686
รวม	**3,429**	**11,193**	**14,622**

4.2 การทดลองวัดความแม่นยำของ Rule-Based method

4.2.1 วัตถุประสงค์

เพื่อเปรียบเทียบความแม่นยำของ Rule-Based method หาค่าประสิทธิภาพต่าง ๆ คือ True Positive Rate (Recall) False Positive Rate และ Precision จากตาราง Confusion matrix และระยะเวลาที่ใช้ในการทำนายจากการทดสอบ 3 รูปแบบคือ

1) แบบมีการใช้ Keyword อย่างเดียว
2) แบบมีการใช้ Keyword ร่วมกับ Slang
3) แบบมีการใช้ Keyword ร่วมกับ Slang ร่วมกับ Emoticon

4.2.2 วิธีการทดลอง

นำข้อมูลที่ใช้ในการทดลองทั้ง 3 ชุดมาผ่านกระบวนการของ Rule-Based method ซึ่งอธิบายในหัวข้อที่ 3.2.5 และนำมาผ่านการวัดประสิทธิภาพของโมเดลเพื่อที่จะหาค่าต่าง ๆ ตามที่ได้กล่าวไว้ในวัตถุประสงค์ ซึ่ง Keyword Slang และ Emoticon ที่ใช้มีตัวอย่างดังนี้

ตาราง 4.2 ตัวอย่าง Keyword, Slang และ Emoticon ที่ใช้สำหรับ Rule-Based method

EMO	KEYWORD	SLANG	EMOTICON
Joy	ปลื้ม ดีใจ เบิกบาน รื่นเริง สนุก สบาย ตลก ความสุข	ร้าก ฟิน กริ๊ส ขรรม ตั้ลล๊าก ชิล เชิล	😀😁🤣😃😄😂😆😉😊😋😎😍😘🙂
Sadness	เสียใจ หดหู่ ผิดหวัง เศร้า สลดใจ เหงา	ฮือ ฮรือ แง้ ซึ้ด อิจ เส้า เสียจัย บรัย แง	😭😢😞😔😑😒😕😟😩😫😖😣
Fear	กลัว น่ากลัว ไม่กล้า หวั่น ไม่นะ ขนลุก สยอง กังวล	หลอน แบร่ น่ากลัวจุง น่ากลัวจรุง	😨😰😱🤯😧👻☠💩💀
Anger	โกรธ แค้น โมโห ไรวะ ไม่พอใจ มั่นหน้า ข้องใจ ฉุน	เวรเอ้ย เวรจริง เวร หัวร้อน ไงวะ แม่ง	😤☹😠😡🤬👿🔥🤘
Disgust	สกปรก ขยะแขยง รังเกียจ เกลียด หุเรศ	อี๋ หยี แหวะ	😒🤢
Surprise	ประหลาดใจ เฮ้ย อะไรกัน ตกใจ สะดุ้ง	อู้ว เงิบ ตะลึง โอ้โห แฟนตาซี โห จริงดิ	😮😯😲😳🙀
Anticipation	หวัง คาด รอ คอย มั่นใจ ขอ ได้โปรด	ป่ะล่ะ ได้มั๊ย	🤔🤷
Acceptance	ยอม ก็ได้ ตามนั้น โอเค ก็ดี พอใจ ช่างมัน แล้วแต่	อือ อืม งืม จัดไป	😐😑😶😌

4.2.3 ผลการทดลอง

4.2.3.1 ค่าความแม่นยำจากการทดสอบแต่ละรูปแบบ

โดยที่ค่าความแม่นยำคือ จำนวนข้อมูลที่ทำนายถูกของทุกคลาส

1) แบบมีการใช้ Keyword อย่างเดียว

ชุด A มีความแม่นยำ = 29.19 % และมีค่าต่าง ๆ ตามตาราง 4.3

ชุด B มีความแม่นยำ = 36.16 % และมีค่าต่าง ๆ ตามตาราง 4.4

ชุด C มีความแม่นยำ = 34.65 % และมีค่าต่าง ๆ ตามตาราง 4.5

ตาราง 4.3 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A แบบมีการใช้ Keyword อย่างเดียว ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	367	111	18	61	24	17	85	32
Sadness	86	131	17	45	9	13	56	20
Fear	7	4	21	6	2	1	8	3
Anger	43	46	12	85	15	11	47	20
Disgust	5	5	1	15	8	0	5	3
Surprise	50	36	15	55	15	11	46	10
Anticipation	385	204	52	121	48	40	317	56
Acceptance	174	102	17	48	19	20	62	61

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	51.3%	27.6%	32.9%
Sadness	34.7%	16.6%	20.5%
Fear	40.4%	3.9%	13.7%
Anger	30.5%	11.1%	19.5%
Disgust	19.0%	3.9%	5.7%
Surprise	4.6%	3.2%	9.7%
Anticipation	25.9%	14.0%	50.6%
Acceptance	12.1%	4.9%	29.8%

ตาราง 4.4 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด B แบบมีการใช้ Keyword อย่างเดียว ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1567	421	79	182	55	89	462	105
Sadness	311	546	37	120	50	43	183	51
Fear	102	81	323	43	9	11	53	12
Anger	234	179	42	540	51	37	201	68
Disgust	31	16	10	42	23	6	20	4
Surprise	81	59	28	88	19	28	83	18
Anticipation	907	530	110	367	90	119	878	167
Acceptance	364	226	39	140	37	34	200	142

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	52.9%	24.7%	43.6%
Sadness	40.7%	15.3%	26.5%
Fear	50.9%	3.3%	48.4%
Anger	39.9%	10.0%	35.5%
Disgust	15.1%	2.8%	6.9%
Surprise	6.9%	3.1%	7.6%
Anticipation	27.7%	15.0%	42.2%
Acceptance	12.0%	4.2%	25.0%

ตาราง 4.5 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด C แบบมีการใช้ Keyword อย่างเดียว ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1937	517	80	280	80	110	522	149
Sadness	413	683	58	141	44	57	251	71
Fear	108	96	333	47	16	8	64	14
Anger	291	212	28	651	62	58	236	93
Disgust	33	26	7	59	24	6	30	9
Surprise	140	114	17	147	21	41	129	33
Anticipation	1255	737	170	507	129	169	1197	227
Acceptance	534	310	77	185	70	54	254	201

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	52.7%	25.3%	41.1%
Sadness	39.8%	15.6%	25.3%
Fear	48.5%	3.1%	43.2%
Anger	39.9%	10.5%	32.3%
Disgust	12.4%	2.9%	5.4%
Surprise	6.4%	3.3%	8.2%
Anticipation	27.3%	14.5%	44.6%
Acceptance	11.9%	4.6%	25.2%

2) แบบมีการใช้ Keyword ร่วมกับ Slang

ชุด A มีความแม่นยำ = 33.97 % และมีค่าต่าง ๆ ตามตาราง 4.6

ชุด B มีความแม่นยำ = 41.20 % และมีค่าต่าง ๆ ตามตาราง 4.7

ชุด C มีความแม่นยำ = 39.56 % และมีค่าต่าง ๆ ตามตาราง 4.8

ตาราง 4.6 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A แบบมีการใช้ Keyword ร่วมกับ Slang ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	263	72	3	63	3	1	296	14
Sadness	74	97	1	33	2	2	163	5
Fear	4	4	24	7	3	1	10	2
Anger	39	32	1	93	5	2	94	4
Disgust	0	6	0	18	6	0	11	1
Surprise	38	19	2	53	3	4	121	1
Anticipation	320	133	8	94	3	10	636	22
Acceptance	130	80	7	42	1	5	196	42

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	36.8%	22.3%	30.3%
Sadness	25.7%	11.3%	21.9%
Fear	43.6%	0.7%	52.2%
Anger	34.4%	9.8%	23.1%
Disgust	14.3%	0.6%	23.1%
Surprise	1.7%	0.7%	16.0%
Anticipation	51.9%	40.4%	41.7%
Acceptance	8.3%	1.7%	46.2%

ตาราง 4.7 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด B แบบมีการใช้ Keyword ร่วมกับ Slang ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1379	242	21	188	10	19	1062	39
Sadness	301	506	6	72	5	6	428	17
Fear	92	71	353	26	6	1	79	6
Anger	189	137	13	592	22	8	353	38
Disgust	26	10	2	38	18	1	56	1
Surprise	53	44	3	82	0	15	206	1
Anticipation	772	312	21	310	8	18	1662	65
Acceptance	340	154	6	107	10	2	476	87

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	46.6%	21.5%	43.8%
Sadness	37.7%	9.8%	34.3%
Fear	55.7%	0.7%	83.1%
Anger	43.8%	8.4%	41.8%
Disgust	11.8%	0.6%	22.8%
Surprise	3.7%	0.5%	21.4%
Anticipation	52.5%	33.1%	38.5%
Acceptance	7.4%	1.7%	34.3%

ตาราง 4.8 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด C แบบมีการใช้ Keyword ร่วมกับ Slang ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1655	311	26	240	14	17	1357	55
Sadness	367	613	8	99	4	5	594	28
Fear	84	69	381	44	6	2	94	6
Anger	227	172	13	687	27	13	446	46
Disgust	25	19	3	60	23	2	61	1
Surprise	87	66	6	132	0	19	329	3
Anticipation	1098	459	27	402	9	28	2277	91
Acceptance	488	228	13	147	8	7	664	130

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	45.0%	21.7%	41.1%
Sadness	35.7%	10.3%	31.6%
Fear	55.5%	0.7%	79.9%
Anger	42.1%	8.7%	37.9%
Disgust	11.9%	0.5%	25.3%
Surprise	3.0%	0.5%	20.4%
Anticipation	51.9%	34.6%	39.1%
Acceptance	7.7%	1.8%	36.1%

3) แบบมีการใช้ Keyword ร่วมกับ Slang ร่วมกับ Emotion

ชุด A มีความแม่นยำ = 35.64 % และมีค่าต่าง ๆ ตามตาราง 4.9

ชุด B มีความแม่นยำ = 43.55 % และมีค่าต่าง ๆ ตามตาราง 4.10

ชุด C มีความแม่นยำ = 41.72 % และมีค่าต่าง ๆ ตามตาราง 4.11

ตาราง 4.9 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A แบบมีการใช้ Keyword ร่วมกับ Slang ร่วมกับ Emoticon ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	419	85	3	23	2	1	172	10
Sadness	83	111	1	28	3	1	145	5
Fear	6	5	24	5	1	2	9	2
Anger	40	41	1	86	9	1	87	7
Disgust	2	6	2	21	5	1	7	1
Surprise	46	27	2	51	2	3	106	3
Anticipation	407	155	8	87	1	10	532	23
Acceptance	164	78	7	41	1	5	165	42

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	58.6%	27.6%	35.9%
Sadness	29.4%	13.0%	21.9%
Fear	44.4%	0.7%	50.0%
Anger	31.6%	8.1%	25.1%
Disgust	11.1%	0.6%	20.8%
Surprise	1.3%	0.7%	12.5%
Anticipation	43.5%	31.3%	43.5%
Acceptance	8.3%	1.7%	45.2%

ตาราง 4.10 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด B แบบมีการใช้ Keyword ร่วมกับ Slang ร่วมกับ Emoticon ด้วย Rule-Based method

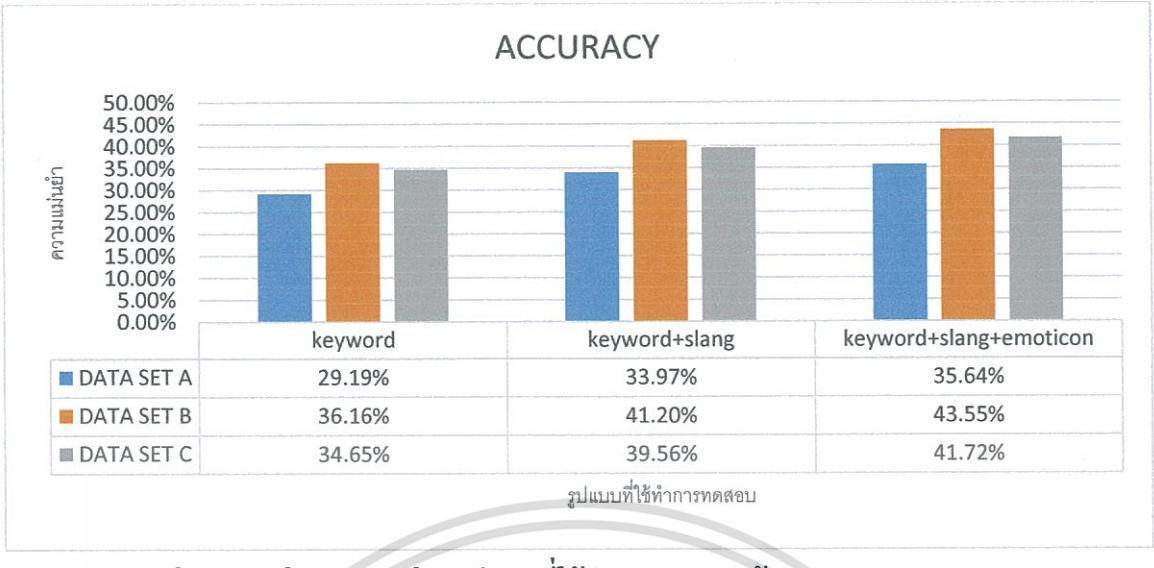
ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1909	262	15	84	5	19	630	36
Sadness	314	538	7	78	7	8	376	13
Fear	138	77	312	24	2	2	71	8
Anger	234	157	11	560	21	11	325	33
Disgust	34	12	4	35	16	2	48	1
Surprise	79	51	2	77	1	15	176	3
Anticipation	954	383	18	278	8	23	1440	64
Acceptance	394	181	3	97	5	3	415	84

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	64.5%	26.1%	47.1%
Sadness	40.1%	11.4%	32.4%
Fear	49.2%	0.6%	83.9%
Anger	41.4%	6.8%	45.4%
Disgust	10.5%	0.4%	24.6%
Surprise	3.7%	0.6%	18.1%
Anticipation	45.5%	25.4%	41.4%
Acceptance	7.1%	1.6%	34.7%

ตาราง 4.11 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด C แบบมีการใช้ Keyword ร่วมกับ Slang ร่วมกับ Emoticon ด้วย Rule-Based method

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	2329	336	22	112	7	20	806	43
Sadness	398	653	9	96	8	10	523	21
Fear	159	74	324	32	4	0	86	7
Anger	288	177	11	660	24	16	405	50
Disgust	31	21	3	56	24	1	54	4
Surprise	128	71	6	123	4	22	281	7
Anticipation	1342	553	34	376	12	28	1960	86
Acceptance	569	246	11	138	7	8	577	129

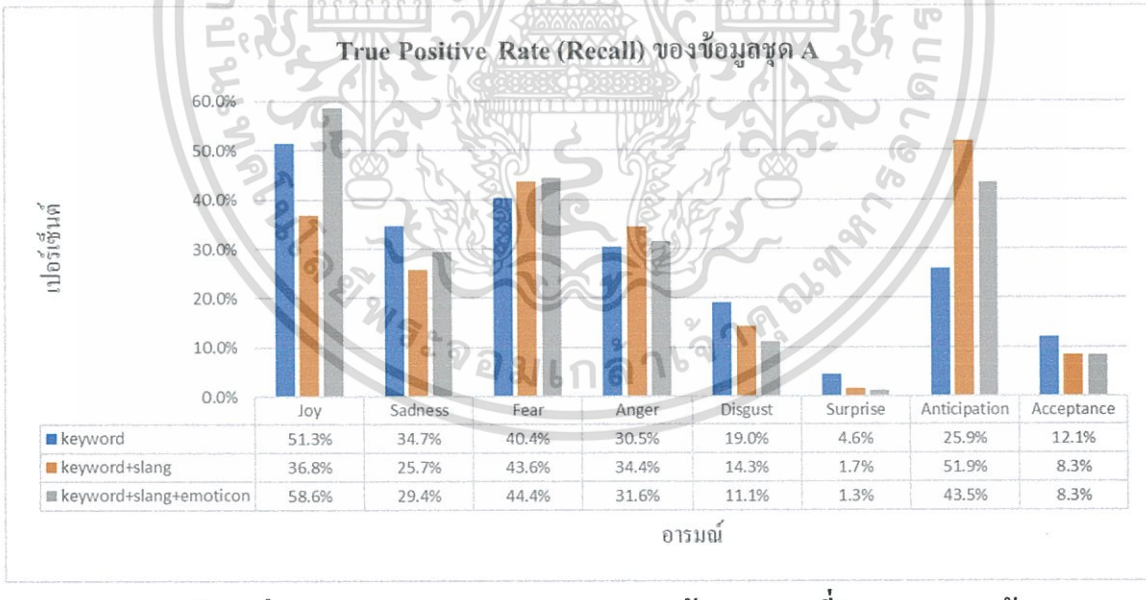
CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	63.4%	26.6%	44.4%
Sadness	38.0%	11.5%	30.6%
Fear	47.2%	0.7%	77.1%
Anger	40.5%	7.2%	41.4%
Disgust	12.4%	0.5%	26.7%
Surprise	3.4%	0.6%	21.0%
Anticipation	44.6%	26.7%	41.8%
Acceptance	7.7%	1.7%	37.2%



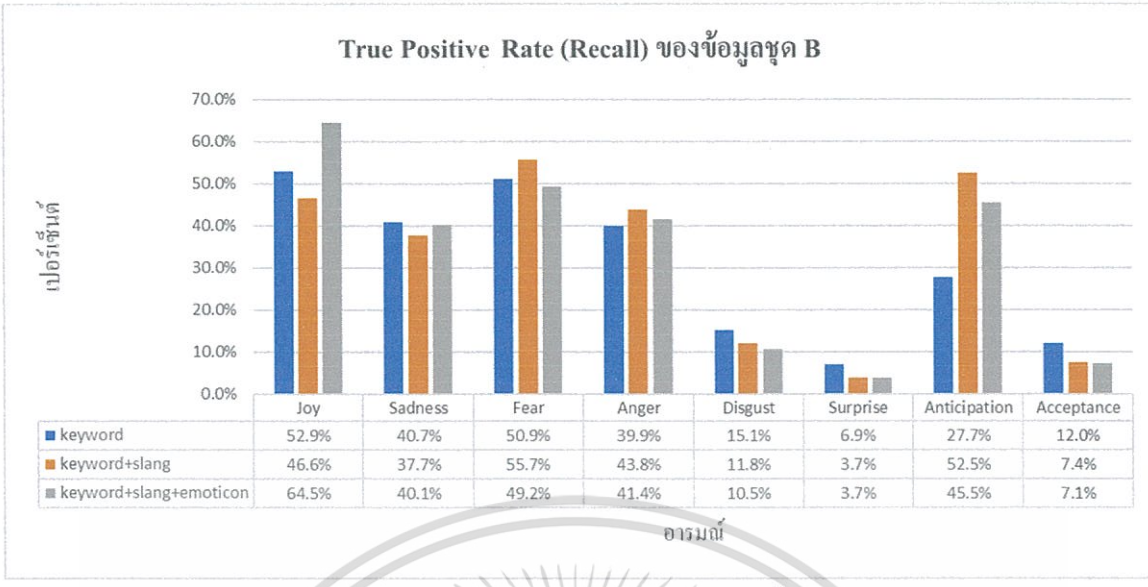
รูป 4.1 ค่าความแม่นยำของแต่ละรูปแบบที่ใช้ทำการทดสอบด้วย Rule-Based method

4.2.3.2 ค่าประสิทธิภาพต่าง ๆ ที่ได้จากการทดสอบข้อมูลแต่ละชุด

1) ค่า True Positive Rate (Recall) คือ จำนวนข้อมูลที่ทำทายถูกจากที่ถูกทั้งหมด

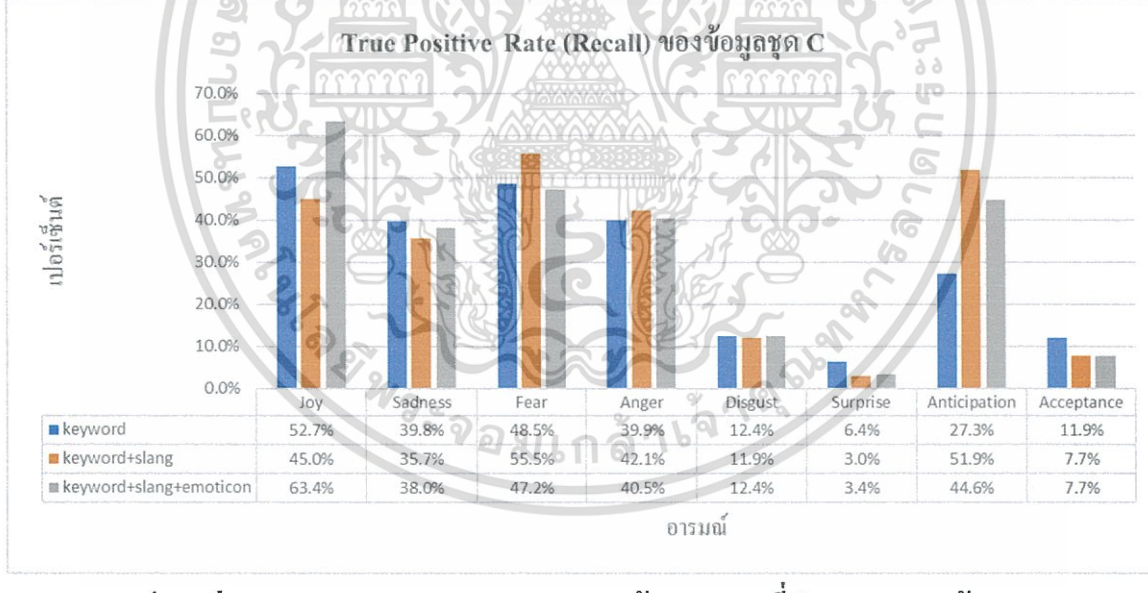


รูป 4.2 ค่า True Positive Rate (Recall) ของข้อมูลชุด B ที่ทำการทดสอบด้วย Rule-Based method



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
keyword	52.9%	40.7%	50.9%	39.9%	15.1%	6.9%	27.7%	12.0%
keyword+slang	46.6%	37.7%	55.7%	43.8%	11.8%	3.7%	52.5%	7.4%
keyword+slang+emoticon	64.5%	40.1%	49.2%	41.4%	10.5%	3.7%	45.5%	7.1%

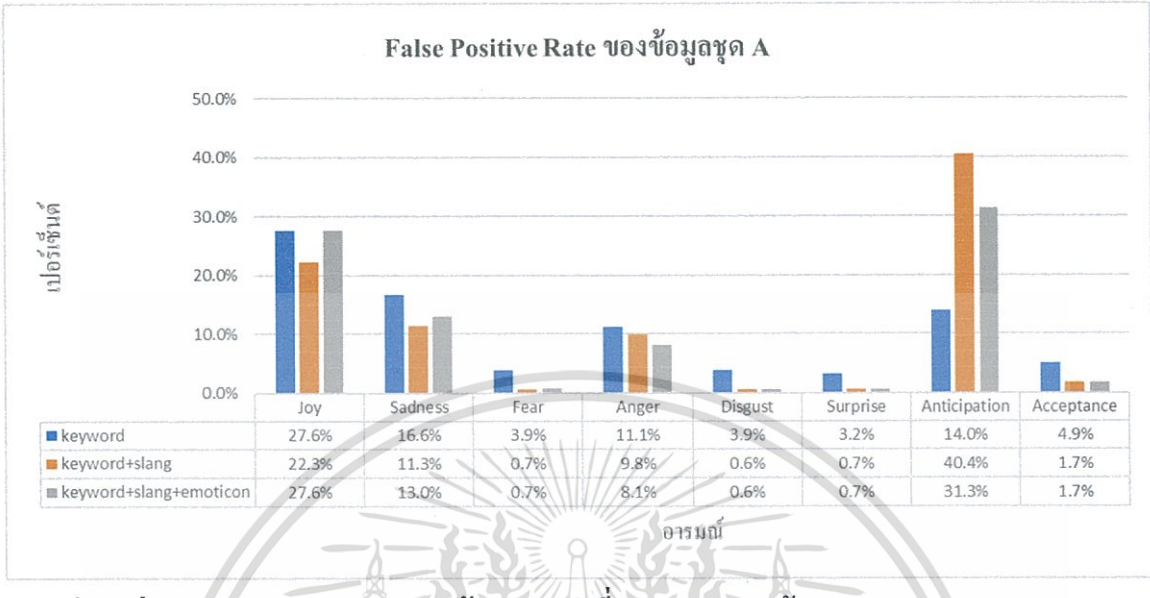
รูป 4.3 ค่า True Positive Rate (Recall) ของข้อมูลชุด B ที่ทำการทดสอบด้วย Rule-Based method



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
keyword	52.7%	39.8%	48.5%	39.9%	12.4%	6.4%	27.3%	11.9%
keyword+slang	45.0%	35.7%	55.5%	42.1%	11.9%	3.0%	51.9%	7.7%
keyword+slang+emoticon	63.4%	38.0%	47.2%	40.5%	12.4%	3.4%	44.6%	7.7%

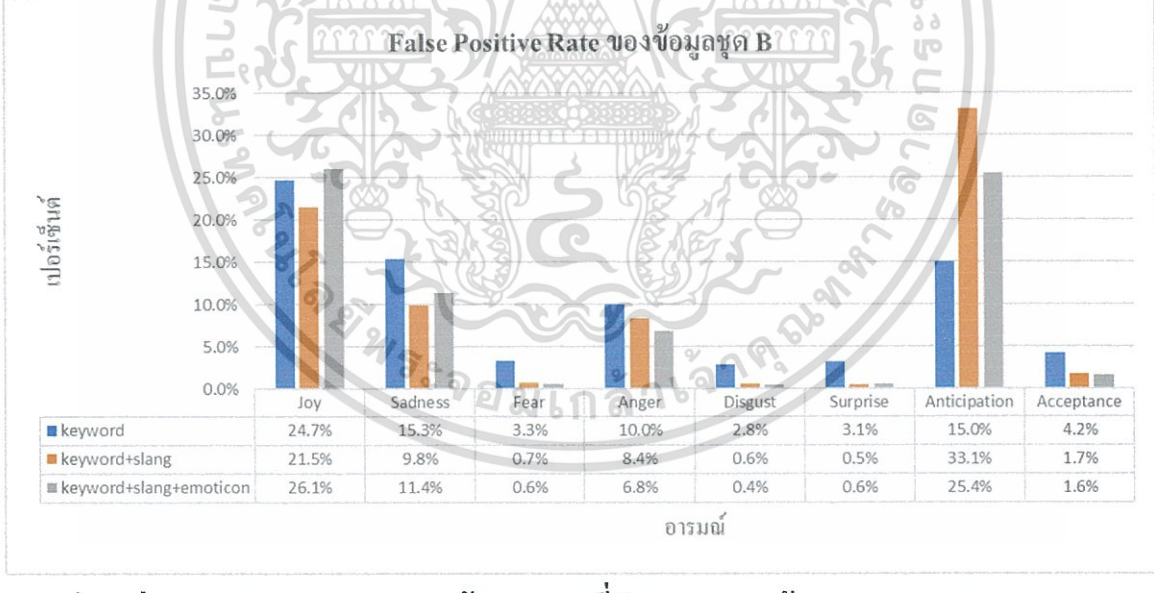
รูป 4.4 ค่า True Positive Rate (Recall) ของข้อมูลชุด C ที่ทำการทดสอบด้วย Rule-Based method

2) ค่า False Positive Rate คือ จำนวนข้อมูลที่ทำนายว่าไม่ถูกจากที่ถูกทั้งหมด



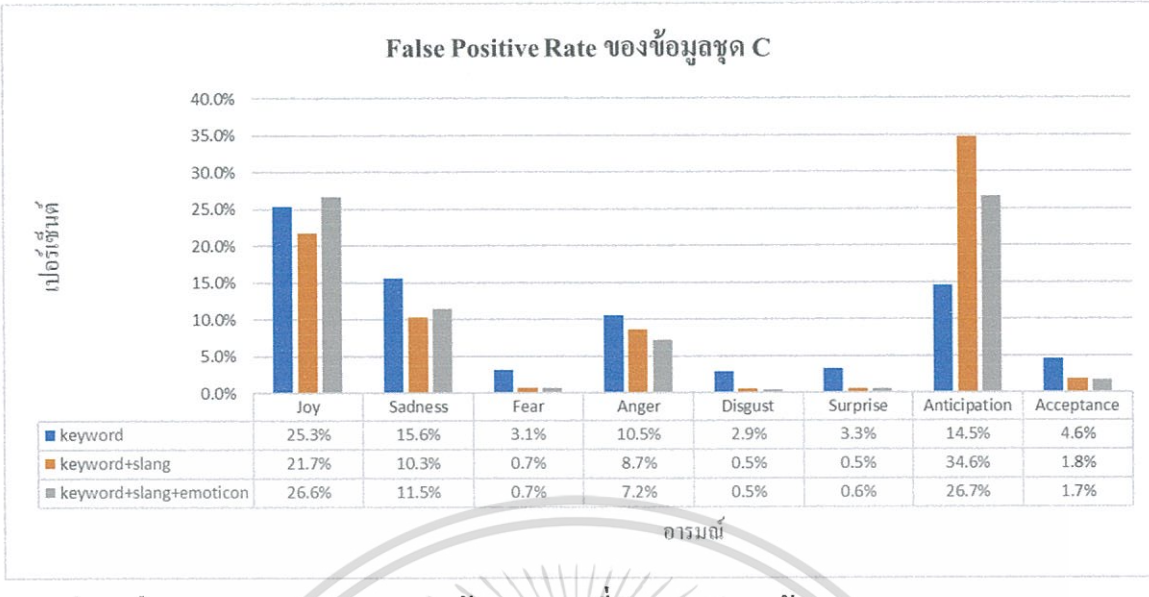
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
keyword	27.6%	16.6%	3.9%	11.1%	3.9%	3.2%	14.0%	4.9%
keyword+slang	22.3%	11.3%	0.7%	9.8%	0.6%	0.7%	40.4%	1.7%
keyword+slang+emoticon	27.6%	13.0%	0.7%	8.1%	0.6%	0.7%	31.3%	1.7%

รูป 4.5 ค่า False Positive Rate ของข้อมูลชุด A ที่ทำการทดสอบด้วย Rule-Based method



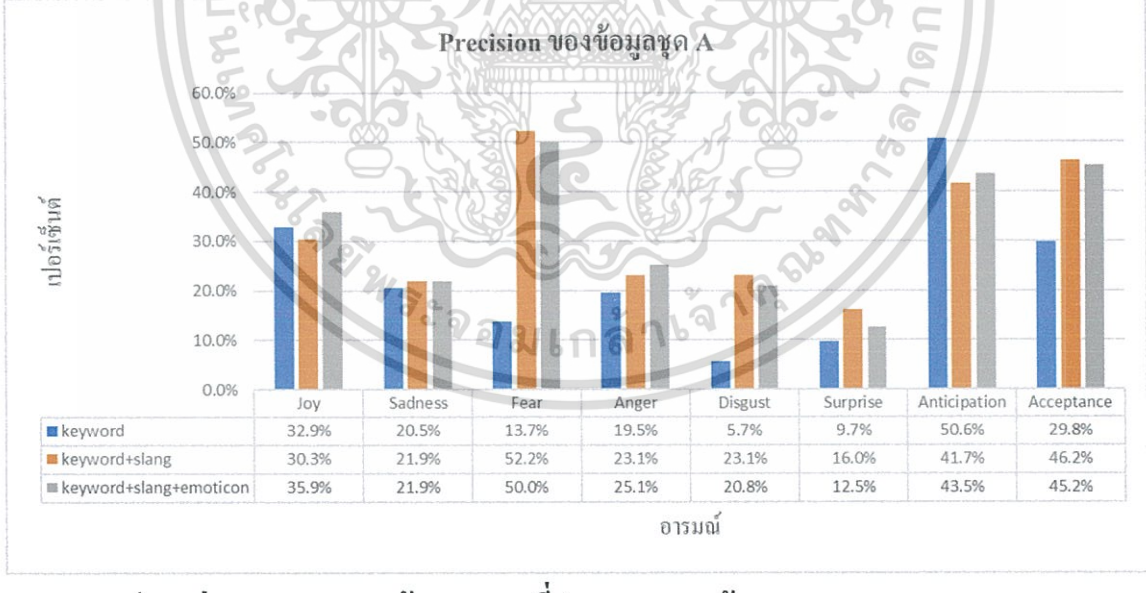
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
keyword	24.7%	15.3%	3.3%	10.0%	2.8%	3.1%	15.0%	4.2%
keyword+slang	21.5%	9.8%	0.7%	8.4%	0.6%	0.5%	33.1%	1.7%
keyword+slang+emoticon	26.1%	11.4%	0.6%	6.8%	0.4%	0.6%	25.4%	1.6%

รูป 4.6 ค่า False Positive Rate ของข้อมูลชุด B ที่ทำการทดสอบด้วย Rule-Based method

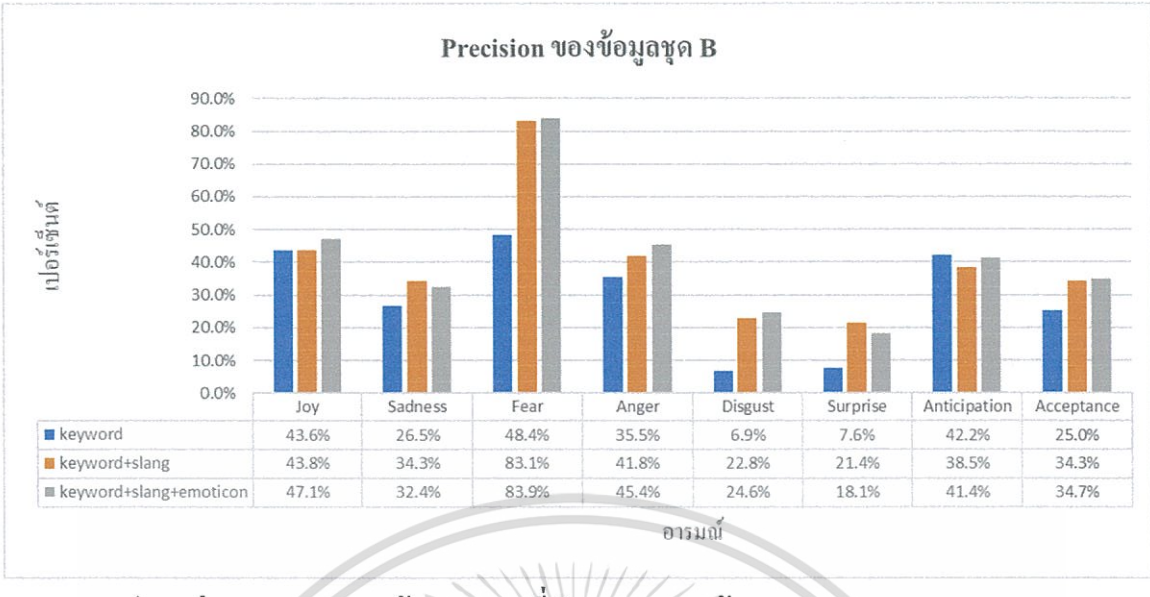


รูป 4.7 ค่า False Positive Rate ของข้อมูลชุด C ที่ทำการทดสอบด้วย Rule-Based method

3) ค่า Precision คือ จำนวนที่ทำนายถูกจากข้อมูลที่ทำนายว่าเป็นคลาสที่พิจารณาอยู่

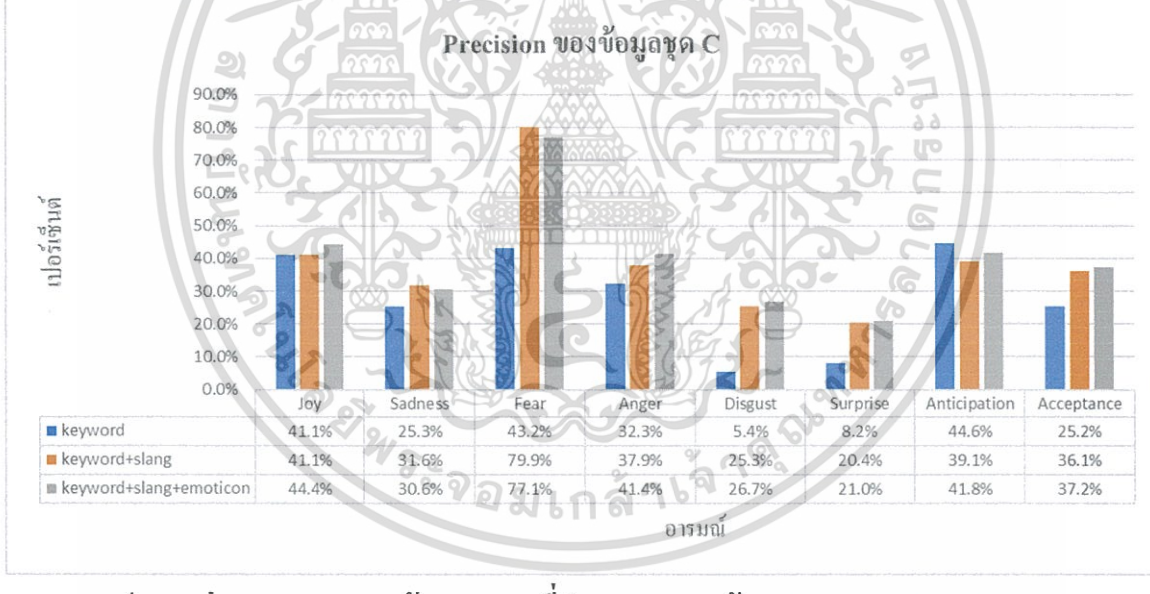


รูป 4.8 ค่า Precision ของข้อมูลชุด A ที่ทำการทดสอบด้วย Rule-Based method



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
keyword	43.6%	26.5%	48.4%	35.5%	6.9%	7.6%	42.2%	25.0%
keyword+slang	43.8%	34.3%	83.1%	41.8%	22.8%	21.4%	38.5%	34.3%
keyword+slang+emoticon	47.1%	32.4%	83.9%	45.4%	24.6%	18.1%	41.4%	34.7%

รูป 4.9 ค่า Precision ของข้อมูลชุด B ที่ทำการทดสอบด้วย Rule-Based method

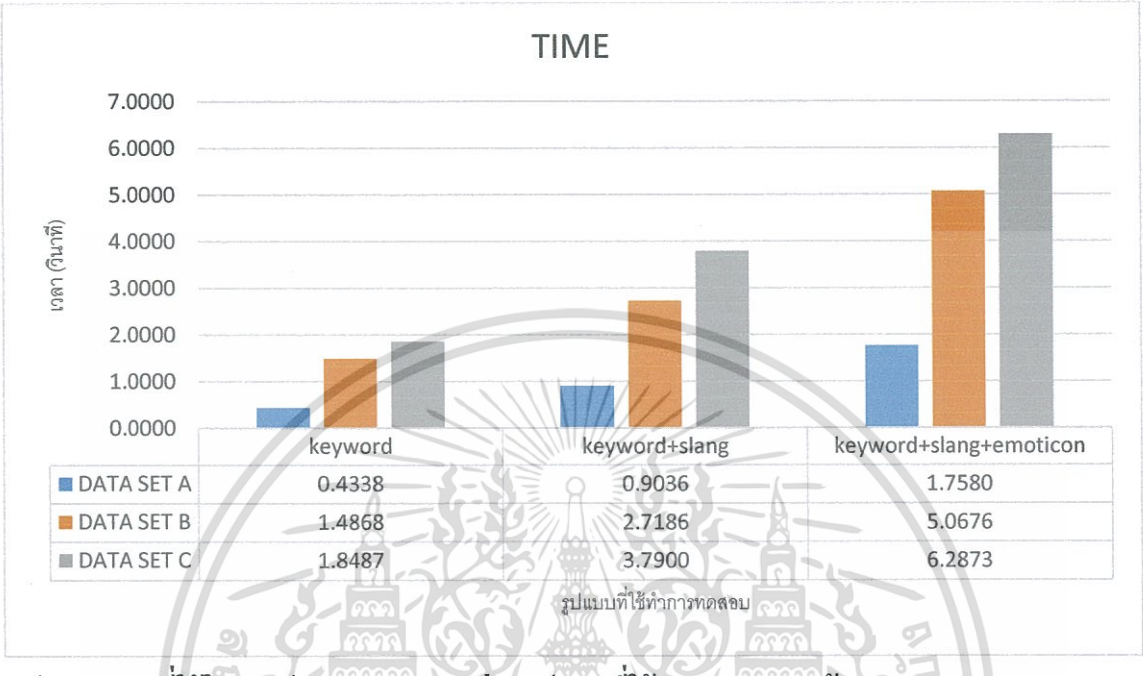


	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
keyword	41.1%	25.3%	43.2%	32.3%	5.4%	8.2%	44.6%	25.2%
keyword+slang	41.1%	31.6%	79.9%	37.9%	25.3%	20.4%	39.1%	36.1%
keyword+slang+emoticon	44.4%	30.6%	77.1%	41.4%	26.7%	21.0%	41.8%	37.2%

รูป 4.10 ค่า Precision ของข้อมูลชุด C ที่ทำการทดสอบด้วย Rule-Based method

4.2.3.3 เวลาที่ใช้ในการประมวลผลของข้อมูลแต่ละชุด

จากการทดสอบด้วยการประมวลผลของข้อมูลแต่ละชุด จำนวนชุดละ 10 ครั้ง



รูป 4.11 เวลาที่ใช้ในการประมวลผลของแต่ละรูปแบบที่ใช้ทำการทดสอบด้วย Rule-Based method

4.2.4 สรุปผลการทดลอง

จากการทดลองวัดค่าความแม่นยำด้วย Rule-Based method ทั้ง 3 รูปแบบนั้นการทดสอบด้วยรูปแบบของการใช้ keyword ร่วมกับ slang ร่วมกับ emoticon นั้นได้ผลออกมามีความแม่นยำมากที่สุดซึ่งวิธีการนี้ใช้ระยะเวลาในการประมวลผลที่มีความรวดเร็ว

4.3 การทดลองวัดความแม่นยำของอัลกอริทึม Multinomial Naïve Bayes

4.3.1 วัตถุประสงค์

เพื่อเปรียบเทียบความแม่นยำของอัลกอริทึม Multinomial Naïve Bayes หาค่าประสิทธิภาพต่าง ๆ คือ True Positive Rate (Recall) False Positive Rate และ Precision จากตาราง Confusion matrix และระยะเวลาที่ใช้ในการทำนายจากการทดสอบ 6 รูปแบบคือ

1) ใช้การตัดคำแบบ unigram
2) ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon
3) ใช้การตัดคำแบบ unigram ร่วมกับ bigram
4) ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับการใช้ Emoticon
5) ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับ trigram
6) ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับ trigram ร่วมกับการใช้ Emoticon

4.3.2 วิธีการทดลอง

1) นำข้อมูลที่ใช้ในการทดลองชุด A มาผ่านกระบวนการของอัลกอริทึม Multinomial Naïve Bayes ซึ่งอธิบายในหัวข้อที่ 3.2.5 เพื่อสร้างแบบจำลองที่ใช้ในการทำนายตามรูปแบบที่ 1) - 6)
2) นำข้อมูลที่ใช้ในการทดลองชุด B และ C มาผ่านกระบวนการของอัลกอริทึม Multinomial Naïve Bayes ซึ่งอธิบายในหัวข้อที่ 3.2.5 เพื่อสร้างแบบจำลองที่ใช้ในการทำนายตามรูปแบบที่ 1) - 2)
3) ทำการวัดประสิทธิภาพด้วยการทำ n-fold cross validation ซึ่งเลือกใช้การทำ 10-fold cross validation

4.3.3 ผลการทดลอง

4.3.3.1 ความแม่นยำจากการทดสอบแต่ละแบบ

1) ใช้การตัดคำแบบ unigram

ชุด A มีความแม่นยำ = 45.55 % และมีค่าต่าง ๆ ตามตาราง 4.12

ชุด B มีความแม่นยำ = 50.71 % และมีค่าต่าง ๆ ตามตาราง 4.13

ชุด C มีความแม่นยำ = 49.61 % และมีค่าต่าง ๆ ตามตาราง 4.14

ตาราง 4.12 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	87	6	0	9	0	4	442	30
Sadness	18	10	0	3	0	0	265	14
Fear	8	1	0	2	0	0	44	8
Anger	18	6	0	12	0	1	158	11
Disgust	2	1	0	3	0	0	23	1
Surprise	10	5	0	6	0	1	148	12
Anticipation	98	14	0	9	0	4	1394	48
Acceptance	29	5	0	1	0	1	399	58

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	15.1%	6.4%	32.2%
Sadness	3.2%	1.2%	20.8%
Fear	0.0%	0.0%	0.0%
Anger	5.8%	1.0%	26.7%
Disgust	0.0%	0.0%	0.0%
Surprise	0.5%	0.3%	9.1%
Anticipation	89.0%	79.4%	48.5%
Acceptance	11.8%	4.2%	31.9%

ตาราง 4.13 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด B ที่ใช้การตัดคำแบบ unigram ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	893	38	6	67	0	71	286	45
Sadness	86	277	5	29	1	1	732	23
Fear	40	14	240	21	0	0	328	6
Anger	86	8	1	469	1	0	513	18
Disgust	15	2	0	19	0	0	79	4
Surprise	58	2	2	25	0	1	211	7
Anticipation	433	58	12	115	0	5	3685	63
Acceptance	122	28	7	42	0	1	844	112

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	38.1%	9.5%	51.5%
Sadness	24.0%	1.5%	64.9%
Fear	37.0%	0.3%	87.9%
Anger	42.8%	3.1%	59.6%
Disgust	0.0%	0.0%	0.0%
Surprise	0.3%	0.1%	6.7%
Anticipation	84.3%	58.5%	48.0%
Acceptance	9.7%	1.7%	40.3%

ตาราง 4.14 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด C ที่ใช้การตัดคำแบบ unigram ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1051	35	7	92	0	13	1644	78
Sadness	105	303	4	32	0	3	973	44
Fear	54	14	206	27	0	0	392	19
Anger	98	17	2	495	1	0	646	43
Disgust	24	1	1	20	0	1	99	3
Surprise	83	4	0	31	0	6	345	19
Anticipation	570	82	10	156	0	8	4995	117
Acceptance	170	30	6	58	0	0	1186	199

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	36.0%	9.4%	48.8%
Sadness	20.7%	1.4%	62.3%
Fear	28.9%	0.2%	87.3%
Anger	38.0%	3.1%	54.3%
Disgust	0.0%	0.0%	0.0%
Surprise	1.2%	0.2%	19.4%
Anticipation	84.1%	60.9%	48.6%
Acceptance	12.1%	2.5%	38.1%

2) ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon

ชุด A มีความแม่นยำ = 45.14 % และมีค่าต่าง ๆ ตามตาราง 4.15

ชุด B มีความแม่นยำ = 50.94 % และมีค่าต่าง ๆ ตามตาราง 4.16

ชุด C มีความแม่นยำ = 49.62 % และมีค่าต่าง ๆ ตามตาราง 4.17

ตาราง 4.15 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	87	4	0	8	0	7	436	36
Sadness	18	8	0	5	0	0	263	16
Fear	7	1	0	1	0	0	46	8
Anger	16	6	0	12	0	1	161	10
Disgust	3	2	0	3	0	0	21	1
Surprise	9	4	0	7	0	2	146	14
Anticipation	98	16	0	11	0	4	1382	56
Acceptance	28	4	0	1	0	2	401	57

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	15.1%	6.3%	32.7%
Sadness	2.6%	1.2%	17.8%
Fear	0.0%	0.0%	0.0%
Anger	5.8%	1.1%	25.0%
Disgust	0.0%	0.0%	0.0%
Surprise	1.1%	0.4%	12.5%
Anticipation	88.2%	79.2%	48.4%
Acceptance	11.6%	4.8%	28.8%

ตาราง 4.16 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด B ที่ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	896	28	9	74	0	6	1281	48
Sadness	77	273	7	28	0	2	749	18
Fear	39	15	241	22	0	0	327	5
Anger	84	8	1	476	1	1	509	16
Disgust	15	2	0	17	0	0	82	3
Surprise	61	4	1	19	0	1	214	6
Anticipation	428	48	13	106	0	6	3708	62
Acceptance	125	28	6	45	0	1	844	107

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	38.3%	9.4%	51.9%
Sadness	23.7%	1.3%	67.2%
Fear	37.1%	0.4%	86.7%
Anger	43.4%	3.1%	60.5%
Disgust	0.0%	0.0%	0.0%
Surprise	0.3%	0.1%	5.9%
Anticipation	84.8%	58.7%	48.1%
Acceptance	9.3%	1.6%	40.4%

ตาราง 4.17 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด C ที่ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	1043	41	11	84	0	11	1658	72
Sadness	98	304	4	36	0	3	977	42
Fear	53	17	211	21	0	0	392	18
Anger	102	15	1	488	1	0	657	38
Disgust	20	2	0	21	0	0	99	7
Surprise	89	6	0	28	0	5	341	19
Anticipation	568	76	13	155	0	7	4998	121
Acceptance	173	28	8	53	0	1	1179	207

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	35.7%	9.4%	48.6%
Sadness	20.8%	1.4%	62.2%
Fear	29.6%	0.3%	85.1%
Anger	37.5%	3.0%	55.1%
Disgust	0.0%	0.0%	0.0%
Surprise	1.0%	0.2%	18.5%
Anticipation	84.2%	61.1%	48.5%
Acceptance	12.6%	2.4%	39.5%

3) ใช้การตัดคำแบบ unigram ร่วมกับ bigram

ชุด A มีความแม่นยำ = 46.01 % และมีค่าต่าง ๆ ตามตาราง 4.18

ตาราง 4.18 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ร่วมกับ bigram ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	84	1	0	3	0	2	466	22
Sadness	21	5	0	3	0	0	273	8
Fear	8	0	0	1	0	0	47	7
Anger	15	2	0	8	0	0	173	8
Disgust	2	2	0	0	0	0	25	1
Surprise	16	1	0	1	0	0	155	9
Anticipation	80	6	0	4	0	0	1439	38
Acceptance	24	2	0	1	0	1	423	42

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	14.5%	5.8%	33.6%
Sadness	1.6%	0.4%	26.3%
Fear	0.0%	0.0%	0.0%
Anger	3.9%	0.4%	38.1%
Disgust	0.0%	0.0%	0.0%
Surprise	0.0%	0.1%	0.0%
Anticipation	91.8%	83.9%	48.0%
Acceptance	8.5%	3.2%	31.1%

4) ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับการใช้ Emoticon

ชุด A มีความแม่นยำ = 45.61 % และมีค่าต่าง ๆ ตามตาราง 4.19

ตาราง 4.19 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	72	4	0	3	0	4	476	19
Sadness	17	5	0	2	0	0	274	12
Fear	8	0	0	1	0	0	48	6
Anger	18	3	0	6	0	1	172	6
Disgust	2	2	0	0	0	0	25	1
Surprise	14	2	0	2	0	0	153	11
Anticipation	86	8	0	5	0	0	1432	36
Acceptance	27	1	0	0	0	1	415	49

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	12.5%	6.0%	29.5%
Sadness	1.6%	0.6%	20.0%
Fear	0.0%	0.0%	0.0%
Anger	2.9%	0.4%	31.6%
Disgust	0.0%	0.0%	0.0%
Surprise	0.0%	0.2%	0.0%
Anticipation	91.4%	83.9%	47.8%
Acceptance	9.9%	3.1%	35.0%

5) ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับ trigram

ชุด A มีความแม่นยำ = 45.37 % และมีค่าต่าง ๆ ตามตาราง 4.20

ตาราง 4.20 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับ trigram ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	127	6	0	6	0	1	401	37
Sadness	25	10	0	1	0	0	249	25
Fear	10	1	0	1	0	0	41	10
Anger	24	5	0	7	0	0	152	18
Disgust	3	1	0	3	0	0	20	3
Surprise	21	3	0	2	0	0	140	16
Anticipation	136	12	0	6	0	0	1336	77
Acceptance	36	2	0	0	0	1	378	76

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	22.0%	8.9%	33.2%
Sadness	3.2%	1.0%	25.0%
Fear	0.0%	0.0%	0.0%
Anger	3.4%	0.6%	26.9%
Disgust	0.0%	0.0%	0.0%
Surprise	0.0%	0.1%	0.0%
Anticipation	85.3%	74.2%	49.2%
Acceptance	15.4%	6.3%	29.0%

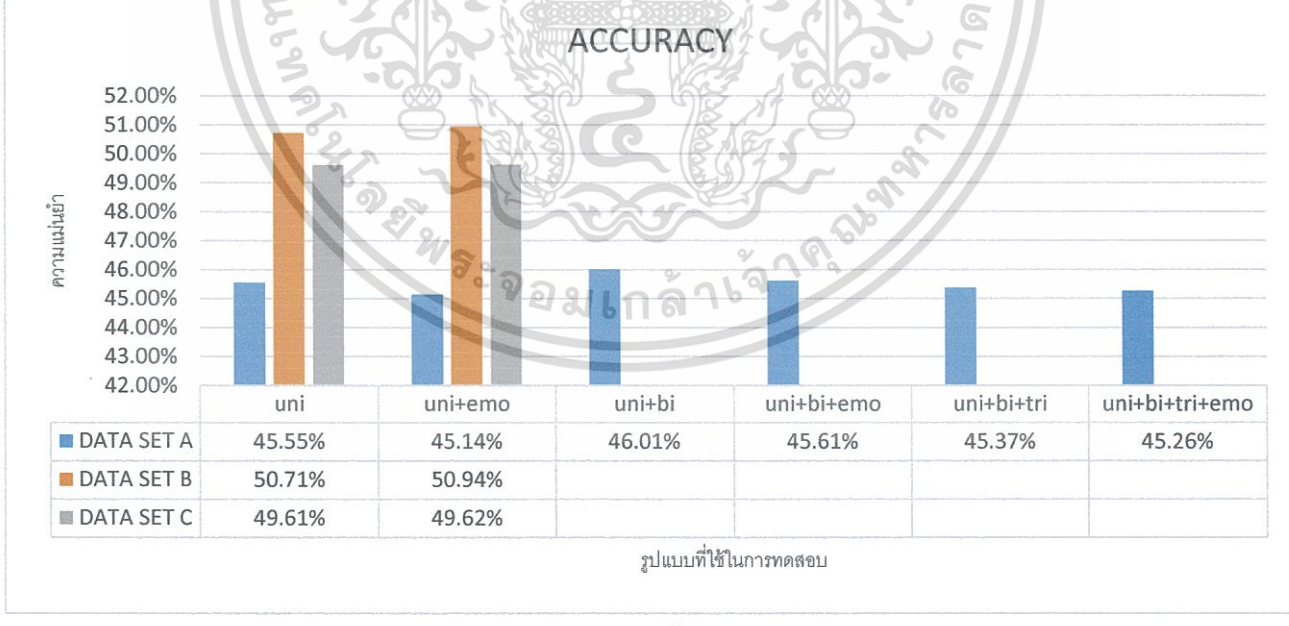
6) ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับ trigram ร่วมกับการใช้ Emoticon

ชุด A มีความแม่นยำ = 45.26 % และมีค่าต่าง ๆ ตามตาราง 4.21

ตาราง 4.21 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ร่วมกับ bigram ร่วมกับ trigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Multinomial Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	131	4	0	5	0	1	398	39
Sadness	30	5	0	3	0	0	249	23
Fear	9	1	0	1	0	0	45	7
Anger	26	6	0	7	0	0	146	21
Disgust	3	2	0	2	0	0	21	2
Surprise	23	5	0	2	0	0	139	13
Anticipation	143	8	0	7	0	0	1328	81
Acceptance	42	2	0	1	0	0	367	81

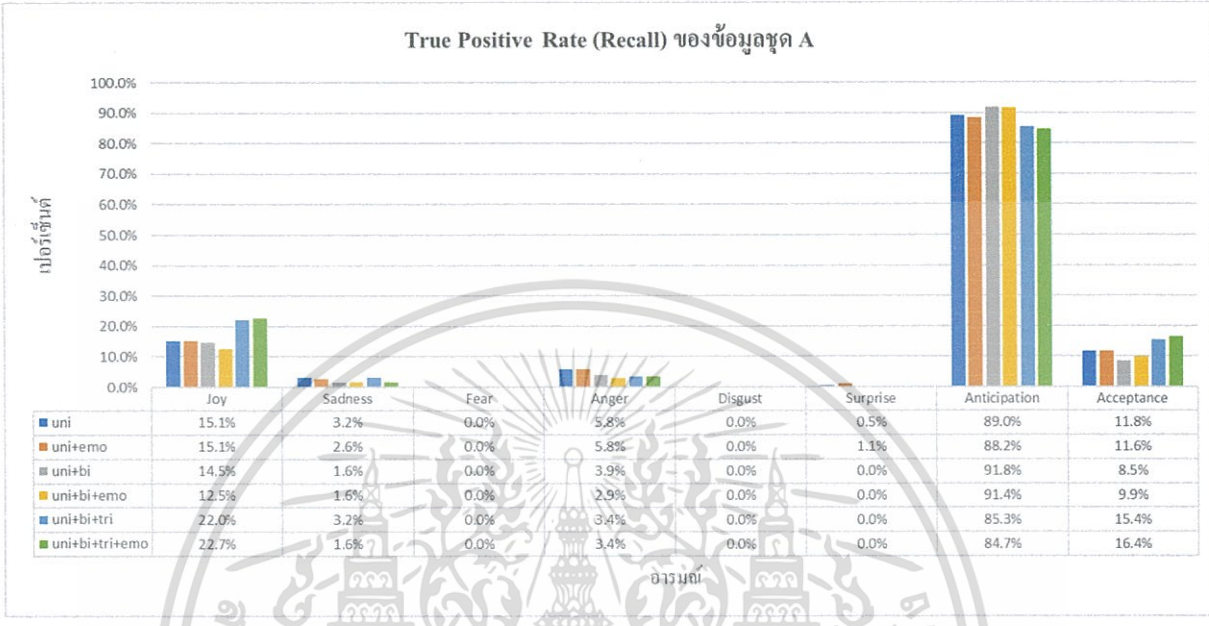
CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	22.7%	9.7%	32.2%
Sadness	1.6%	0.9%	15.2%
Fear	0.0%	0.0%	0.0%
Anger	3.4%	0.7%	25.0%
Disgust	0.0%	0.0%	0.0%
Surprise	0.0%	0.0%	0.0%
Anticipation	84.7%	73.3%	49.3%
Acceptance	16.4%	6.3%	30.3%



รูป 4.12 ค่าความแม่นยำของแต่ละรูปแบบที่ใช้ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

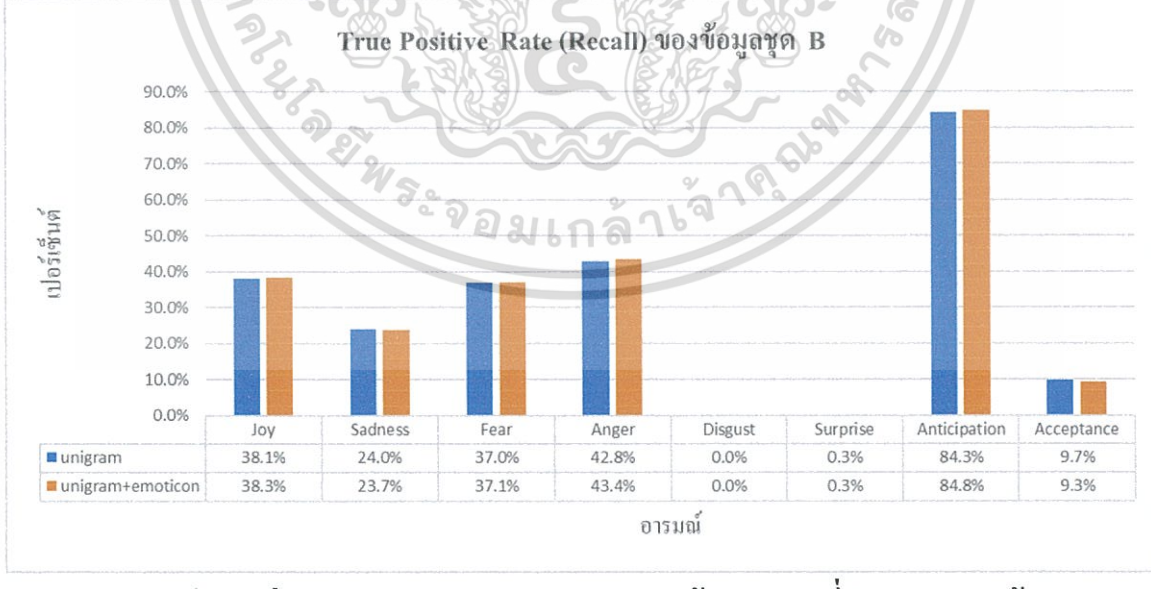
4.3.3.2 ค่าประสิทธิภาพต่าง ๆ ที่ได้จากการทดสอบข้อมูลแต่ละชุด

1) ค่า True Positive Rate (Recall) คือ จำนวนข้อมูลที่ทำทายถูกจากที่ถูกทั้งหมด



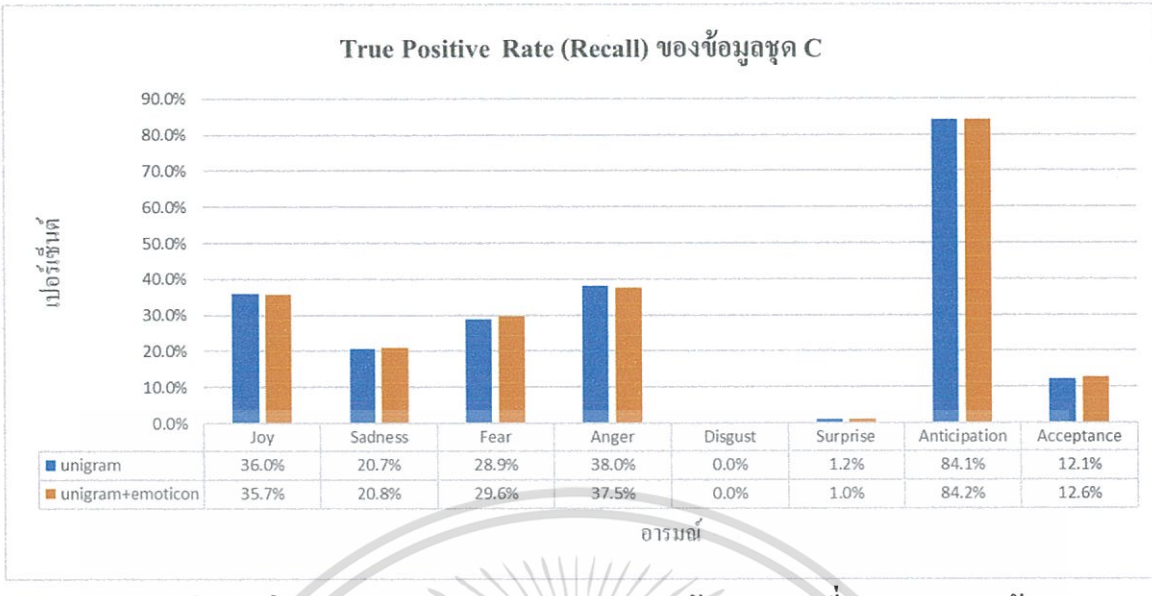
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
uni	15.1%	3.2%	0.0%	5.8%	0.0%	0.5%	89.0%	11.8%
uni+emo	15.1%	2.6%	0.0%	5.8%	0.0%	1.1%	88.2%	11.6%
uni+bi	14.5%	1.6%	0.0%	3.9%	0.0%	0.0%	91.8%	8.5%
uni+bi+emo	12.5%	1.6%	0.0%	2.9%	0.0%	0.0%	91.4%	9.9%
uni+bi+tri	22.0%	3.2%	0.0%	3.4%	0.0%	0.0%	85.3%	15.4%
uni+bi+tri+emo	22.7%	1.6%	0.0%	3.4%	0.0%	0.0%	84.7%	16.4%

รูป 4.13 ค่า True Positive Rate (Recall) ของข้อมูลชุด A ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	38.1%	24.0%	37.0%	42.8%	0.0%	0.3%	84.3%	9.7%
unigram+emoticon	38.3%	23.7%	37.1%	43.4%	0.0%	0.3%	84.8%	9.3%

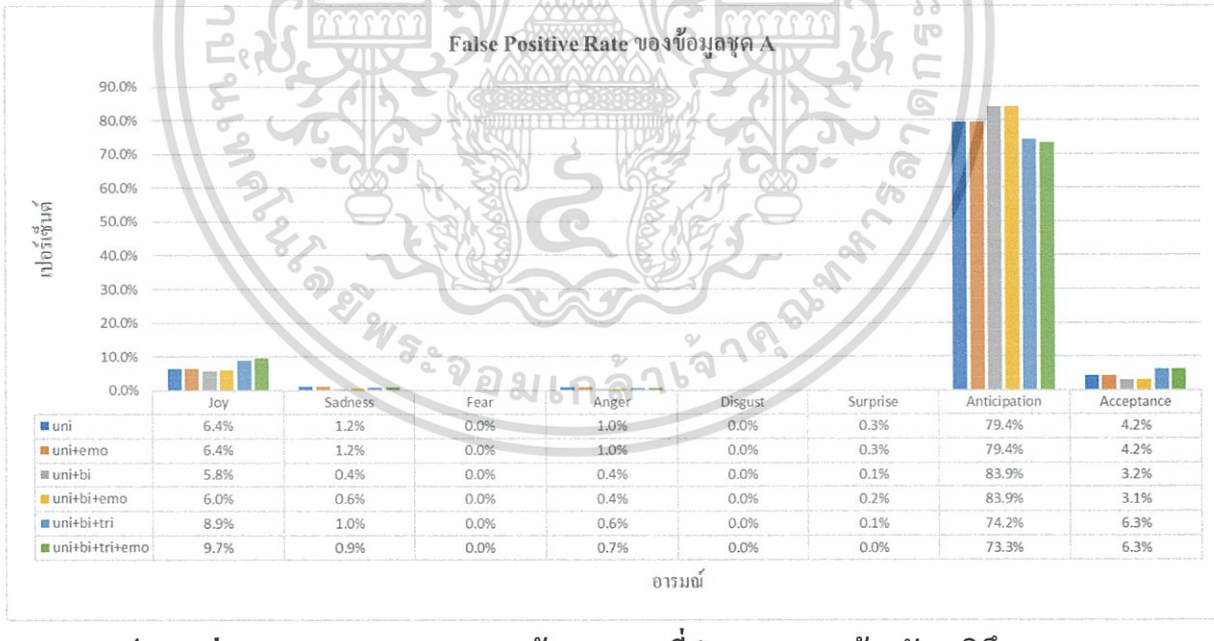
รูป 4.14 ค่า True Positive Rate (Recall) ของข้อมูลชุด B ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	36.0%	20.7%	28.9%	38.0%	0.0%	1.2%	84.1%	12.1%
unigram+emoticon	35.7%	20.8%	29.6%	37.5%	0.0%	1.0%	84.2%	12.6%

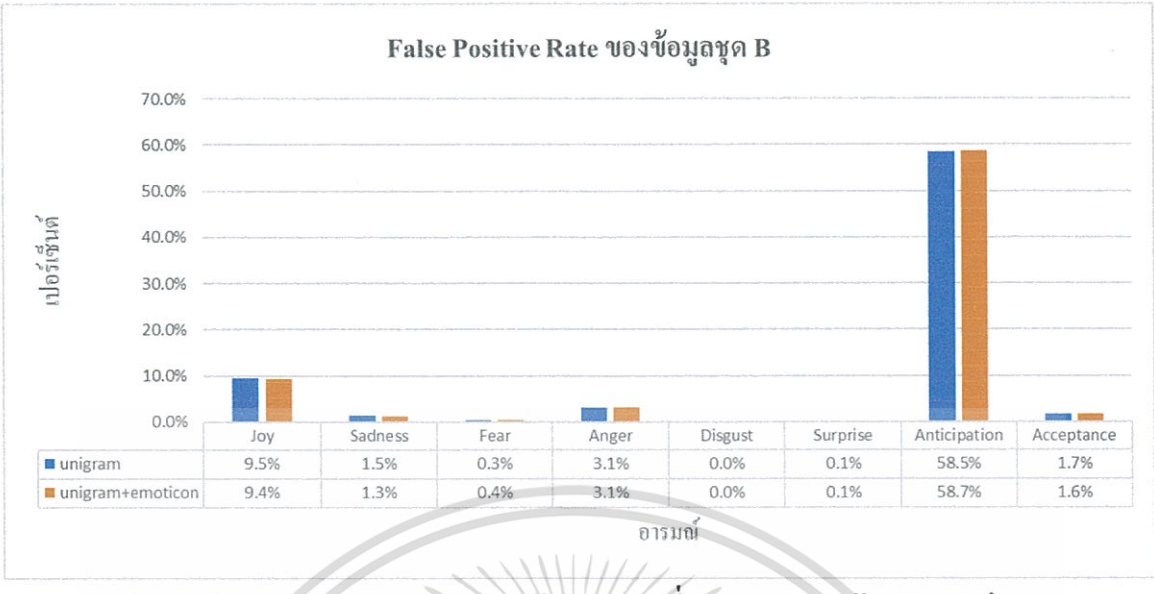
รูป 4.15 ค่า True Positive Rate (Recall) ของข้อมูลชุด C ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

2) ค่า False Positive Rate คือ จำนวนข้อมูลที่ทำนายว่าไม่ถูกจากที่ถูกทั้งหมด

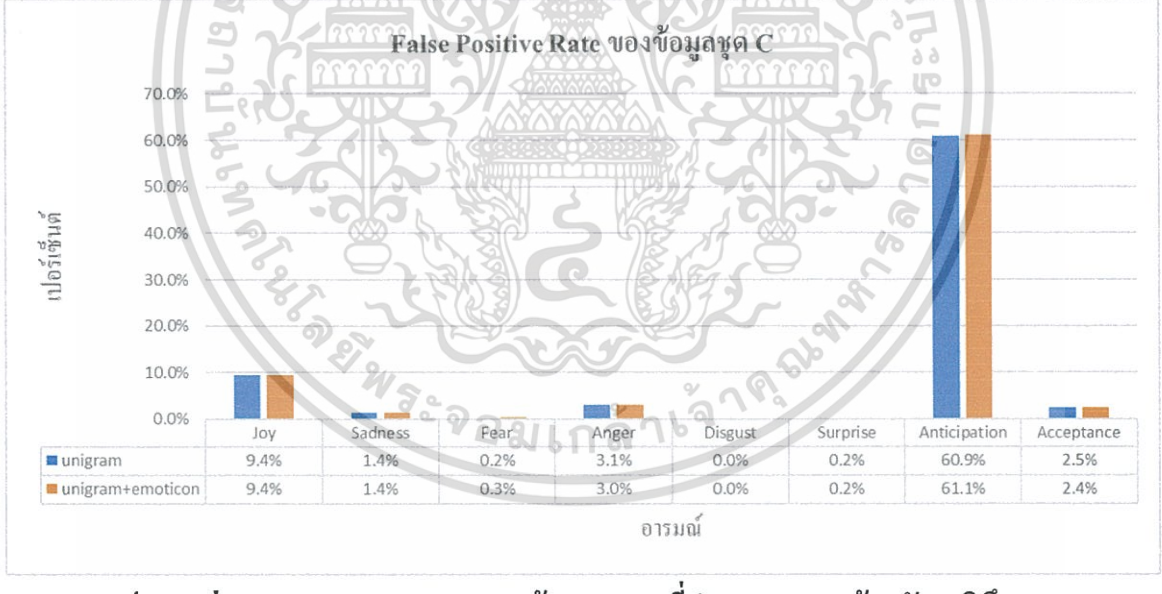


	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
uni	6.4%	1.2%	0.0%	1.0%	0.0%	0.3%	79.4%	4.2%
uni+emo	6.4%	1.2%	0.0%	1.0%	0.0%	0.3%	79.4%	4.2%
uni+bi	5.8%	0.4%	0.0%	0.4%	0.0%	0.1%	83.9%	3.2%
uni+bi+emo	6.0%	0.6%	0.0%	0.4%	0.0%	0.2%	83.9%	3.1%
uni+bi+tri	8.9%	1.0%	0.0%	0.6%	0.0%	0.1%	74.2%	6.3%
uni+bi+tri+emo	9.7%	0.9%	0.0%	0.7%	0.0%	0.0%	73.3%	6.3%

รูป 4.16 ค่า False Positive Rate ของข้อมูลชุด A ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

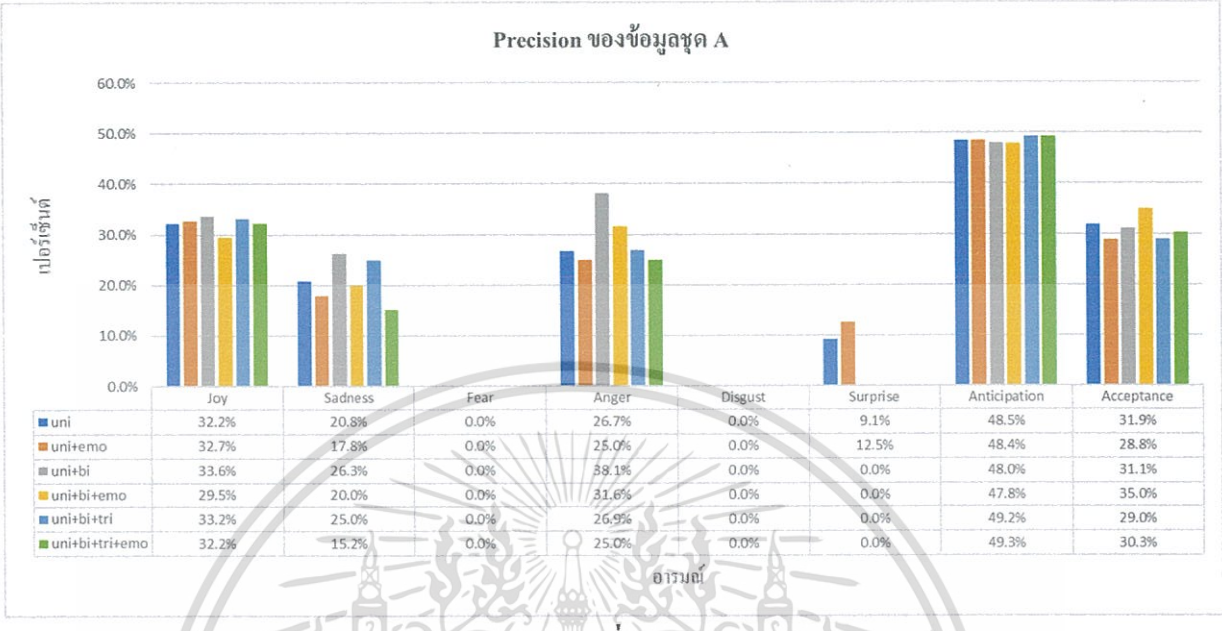


รูป 4.17 ค่า False Positive Rate ของข้อมูลชุด B ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes



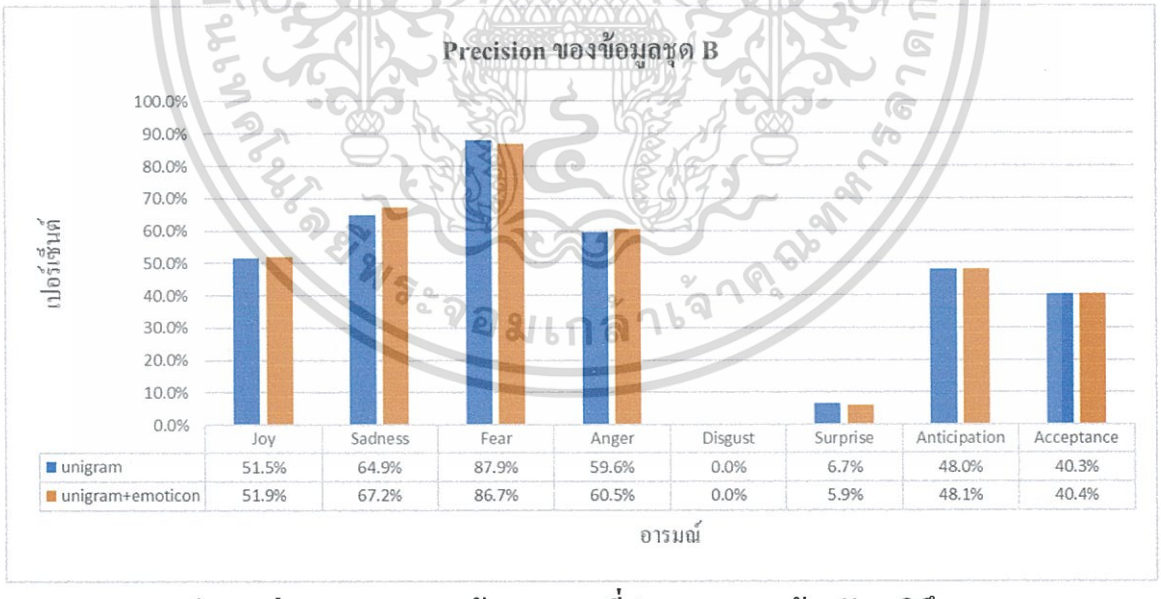
รูป 4.18 ค่า False Positive Rate ของข้อมูลชุด C ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

3) ค่า Precision คือ จำนวนที่ทำนายถูกจากข้อมูลที่ทำนายว่าเป็นคลาสที่พิจารณาอยู่



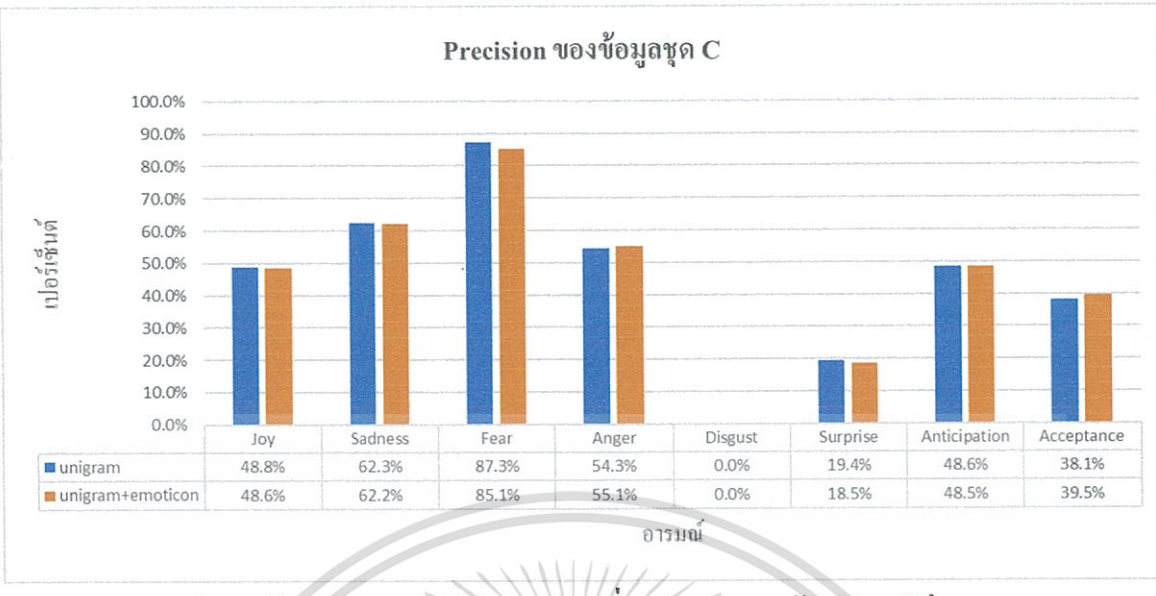
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
uni	32.2%	20.8%	0.0%	26.7%	0.0%	9.1%	48.5%	31.9%
uni+emo	32.7%	17.8%	0.0%	25.0%	0.0%	12.5%	48.4%	28.8%
uni+bi	33.6%	26.3%	0.0%	38.1%	0.0%	0.0%	48.0%	31.1%
uni+bi+emo	29.5%	20.0%	0.0%	31.6%	0.0%	0.0%	47.8%	35.0%
uni+bi+tri	33.2%	25.0%	0.0%	26.9%	0.0%	0.0%	49.2%	29.0%
uni+bi+tri+emo	32.2%	15.2%	0.0%	25.0%	0.0%	0.0%	49.3%	30.3%

รูป 4.19 ค่า Precision ของข้อมูลชุด A ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	51.5%	64.9%	87.9%	59.6%	0.0%	6.7%	48.0%	40.3%
unigram+emoticon	51.9%	67.2%	86.7%	60.5%	0.0%	5.9%	48.1%	40.4%

รูป 4.20 ค่า Precision ของข้อมูลชุด B ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

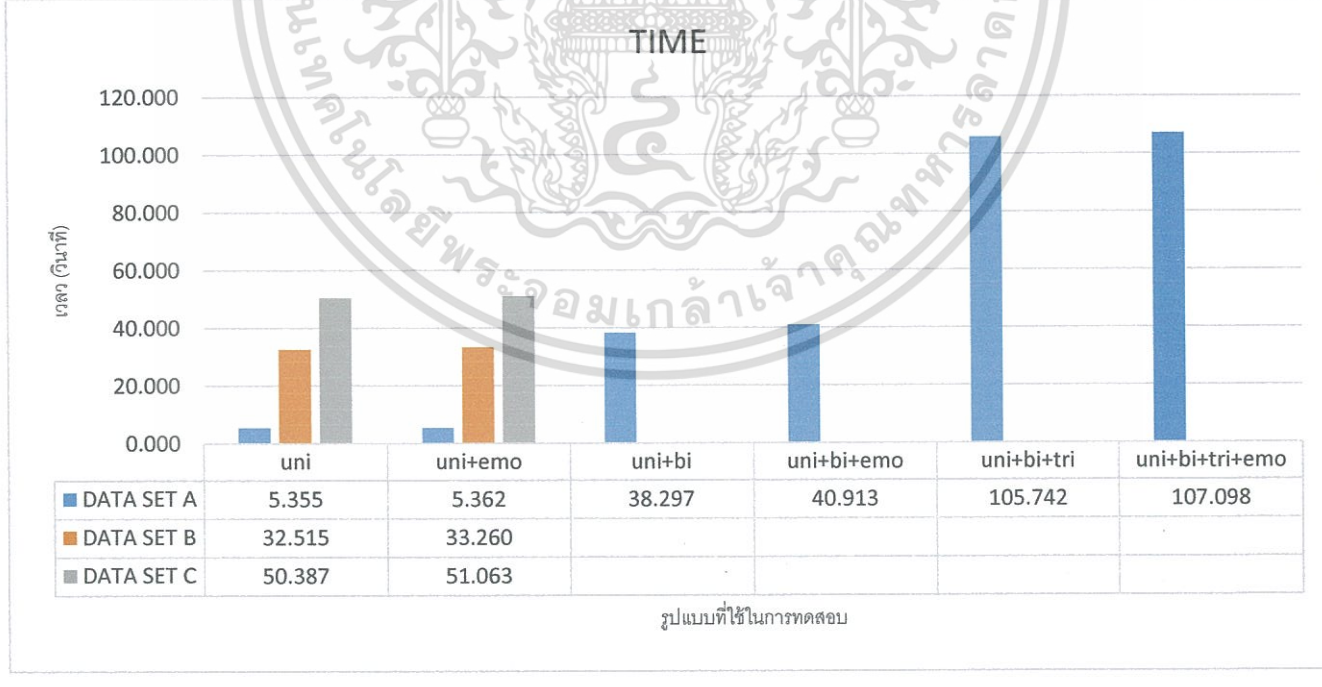


	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	48.8%	62.3%	87.3%	54.3%	0.0%	19.4%	48.6%	38.1%
unigram+emoticon	48.6%	62.2%	85.1%	55.1%	0.0%	18.5%	48.5%	39.5%

รูป 4.21 ค่า Precision ของข้อมูลชุด C ที่ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

4.3.3.3 เวลาที่ใช้ในการประมวลผลข้อมูลแต่ละชุด

จากการทดสอบด้วยการประมวลผลของข้อมูลแต่ละชุด จำนวนชุดละ 10 ครั้ง



	uni	uni+emo	uni+bi	uni+bi+emo	uni+bi+tri	uni+bi+tri+emo
DATA SET A	5.355	5.362	38.297	40.913	105.742	107.098
DATA SET B	32.515	33.260				
DATA SET C	50.387	51.063				

รูป 4.22 เวลาที่ใช้ในการประมวลผลของแต่ละรูปแบบที่ใช้ทำการทดสอบด้วยอัลกอริทึม Multinomial Naïve Bayes

4.3.4 สรุปผลการทดลอง

จากการทดลองวัดค่าความแม่นยำและเวลาที่ใช้ประมวลผลด้วยอัลกอริทึม Multinomial Naïve Bayes ด้วยการทดสอบทั้ง 6 รูปแบบนั้นจะมีเพียงข้อมูลชุด A เท่านั้นที่มีการทดสอบครบทั้ง 6 รูปแบบเนื่องจากเป็นชุดข้อมูลที่มีขนาดเล็กที่สุดและผลที่ได้ในส่วนของค่าความแม่นยำที่เกิดจากการทดสอบของทั้ง 6 รูปแบบพบว่ามีค่าใกล้เคียงกันมาก แต่พบว่าในส่วนของเวลาที่ใช้มีระยะเวลาที่นานกว่ากันมาก หากมีการนำไปใช้กับชุดข้อมูล B และ C ที่มีขนาดและจำนวนข้อมูลที่มากขึ้นผลที่ได้จากค่าความแม่นยำจะไม่ต่างกันแต่จำเป็นจะต้องใช้เวลาในการประมวลผลที่นานขึ้นเนื่องจากต้องผ่านหลายขั้นตอนกว่าจะสามารถทำนายผลออกมาได้

จากการทดลองนี้พบว่าการใช้รูปแบบของการใช้การตัดคำแบบ unigram ก็เพียงพอแล้วที่จะใช้สำหรับการทำนายผลเนื่องจากมีปัจจัยในส่วนของเวลามาเกี่ยวข้องและค่าความแม่นยำที่ออกมาพบว่ามีค่าต่างกันเพียงเล็กน้อยอยู่ที่ไม่เกิน 1%

4.4 การทดลองวัดความแม่นยำของอัลกอริทึม Naïve Bayes

4.4.1 วัตถุประสงค์

เพื่อเปรียบเทียบความแม่นยำของอัลกอริทึม Naïve Bayes หาค่าประสิทธิภาพต่าง ๆ คือ True Positive Rate (Recall) False Positive Rate และ Precision จากตาราง Confusion matrix และระยะเวลาที่ใช้ในการทำนายจากการทดสอบ 2 รูปแบบคือ

1) ใช้การตัดคำแบบ unigram
2) ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon

4.4.2 วิธีการทดลอง

1) นำข้อมูลที่ใช้ในการทดลองทั้ง 3 ชุดมาผ่านกระบวนการของอัลกอริทึม Naïve Bayes ซึ่งอธิบายในหัวข้อที่ 3.2.5 เพื่อสร้างแบบจำลองที่ใช้ในการทำนายตามรูปแบบที่ 1 และ 2
2) ทำการวัดประสิทธิภาพด้วยการทำ n-fold cross validation ซึ่งเลือกใช้การทำ 10-fold cross validation

4.4.3 ผลการทดลอง

4.4.3.1 ความแม่นยำจากการทดสอบแต่ละแบบ

1) ใช้การตัดคำแบบ unigram

ชุด A มีความแม่นยำ = 28.84 % และมีค่าต่าง ๆ ตามตาราง 4.22

ชุด B มีความแม่นยำ = 40.28 % และมีค่าต่าง ๆ ตามตาราง 4.23

ชุด C มีความแม่นยำ = 38.59 % และมีค่าต่าง ๆ ตามตาราง 4.24

ตาราง 4.22 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ด้วยอัลกอริทึม Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	158	49	20	62	18	123	99	49
Sadness	47	67	7	21	10	54	60	44
Fear	4	8	11	8	0	11	12	9
Anger	34	12	6	45	13	50	26	20
Disgust	5	0	0	12	1	10	2	0
Surprise	37	18	4	12	7	62	27	15
Anticipation	237	140	38	142	50	242	539	179
Acceptance	81	58	13	53	15	66	101	106

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	35.7%	11.8%	44.5%
Sadness	32.9%	5.1%	42.4%
Fear	64.3%	4.7%	45.6%
Anger	42.0%	7.5%	37.9%
Disgust	12.6%	3.5%	3.7%
Surprise	40.8%	14.7%	7.2%
Anticipation	45.6%	16.4%	64.0%
Acceptance	24.6%	7.6%	27.1%

ตาราง 4.23 Confusion matrix และค่าต่างๆ ของข้อมูลชุด B ที่ใช้การตัดคำแบบ unigram ด้วยอัลกอริทึม Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	836	107	120	147	79	478	432	143
Sadness	139	380	53	81	33	164	196	108
Fear	27	27	417	38	14	35	50	41
Anger	84	42	51	460	63	184	121	91
Disgust	17	5	3	22	15	25	21	11
Surprise	44	14	15	30	15	125	51	12
Anticipation	582	233	207	328	142	530	1992	357
Acceptance	148	89	48	108	43	185	251	284

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	35.7%	11.8%	44.5%
Sadness	32.9%	5.1%	42.4%
Fear	64.3%	4.7%	45.6%
Anger	42.0%	7.5%	37.9%
Disgust	12.6%	3.5%	3.7%
Surprise	40.8%	14.7%	7.2%
Anticipation	45.6%	16.4%	64.0%
Acceptance	24.6%	7.6%	27.1%

ตาราง 4.24 Confusion matrix และค่าต่าง ๆ ของข้อมูล C ที่ใช้การตัดคำแบบ unigram ด้วยอัลกอริทึม Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	967	113	139	205	80	680	536	200
Sadness	163	449	63	104	35	252	246	152
Fear	31	29	437	45	18	49	58	45
Anger	91	57	59	504	69	271	134	117
Disgust	22	6	3	29	18	36	23	12
Surprise	62	18	21	44	19	223	72	29
Anticipation	728	287	242	456	197	870	2636	522
Acceptance	211	115	60	150	64	291	349	409

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	33.1%	11.2%	42.5%
Sadness	30.7%	4.7%	41.8%
Fear	61.4%	4.2%	42.7%
Anger	38.7%	7.8%	32.8%
Disgust	12.1%	3.3%	3.6%
Surprise	45.7%	17.3%	8.3%
Anticipation	44.4%	16.3%	65.0%
Acceptance	24.8%	8.3%	27.5%

2) ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon

ชุด A มีความแม่นยำ = 28.20 % และมีค่าต่าง ๆ ตามตาราง 4.25

ชุด B มีความแม่นยำ = 40.58 % และมีค่าต่าง ๆ ตามตาราง 4.26

ชุด C มีความแม่นยำ = 38.74 % และมีค่าต่าง ๆ ตามตาราง 4.27

ตาราง 4.25 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด A ที่ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	152	45	23	63	20	126	105	44
Sadness	44	60	14	21	7	62	57	45
Fear	9	3	11	6	0	11	11	12
Anger	32	11	7	47	13	48	24	24
Disgust	5	2	1	9	0	10	3	0
Surprise	32	14	3	13	8	63	26	23
Anticipation	243	142	43	138	58	244	530	169
Acceptance	77	60	12	42	18	75	105	104

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	36.3%	11.6%	45.2%
Sadness	32.7%	5.2%	41.9%
Fear	63.8%	4.7%	45.3%
Anger	43.1%	7.5%	38.5%
Disgust	14.3%	3.7%	3.9%
Surprise	43.1%	14.4%	7.8%
Anticipation	45.8%	16.6%	63.8%
Acceptance	24.2%	7.3%	27.6%

ตาราง 4.26 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด B ที่ใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Naïve Bayes

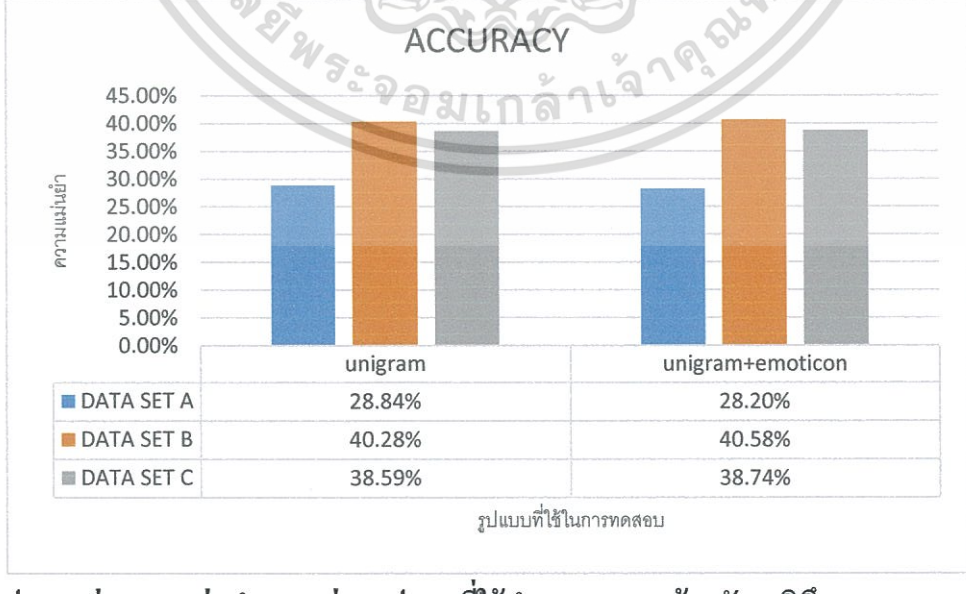
ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	849	100	124	150	84	458	438	139
Sadness	134	377	51	80	41	161	205	105
Fear	30	28	414	37	14	42	43	41
Anger	81	43	54	472	58	181	123	84
Disgust	18	4	4	18	17	22	22	14
Surprise	48	11	12	27	17	132	44	15
Anticipation	565	247	211	327	164	519	2001	337
Acceptance	152	90	44	114	36	180	260	280

CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	36.3%	11.6%	45.2%
Sadness	32.7%	5.2%	41.9%
Fear	63.8%	4.7%	45.3%
Anger	43.1%	7.5%	38.5%
Disgust	14.3%	3.7%	3.9%
Surprise	43.1%	14.4%	7.8%
Anticipation	45.8%	16.6%	63.8%
Acceptance	24.2%	7.3%	27.6%

ตาราง 4.27 Confusion matrix และค่าต่าง ๆ ของข้อมูลชุด C โดยใช้การตัดคำแบบ unigram ร่วมกับการใช้ Emoticon ด้วยอัลกอริทึม Naïve Bayes

ACTUAL	PREDICTED							
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
Joy	976	114	148	200	80	671	540	191
Sadness	162	445	59	96	40	264	242	156
Fear	30	30	437	37	25	47	62	44
Anger	87	52	59	527	65	253	142	117
Disgust	22	4	5	28	17	36	23	14
Surprise	61	19	17	45	26	229	65	26
Anticipation	736	293	232	476	176	884	2622	519
Acceptance	202	111	67	152	52	297	356	412

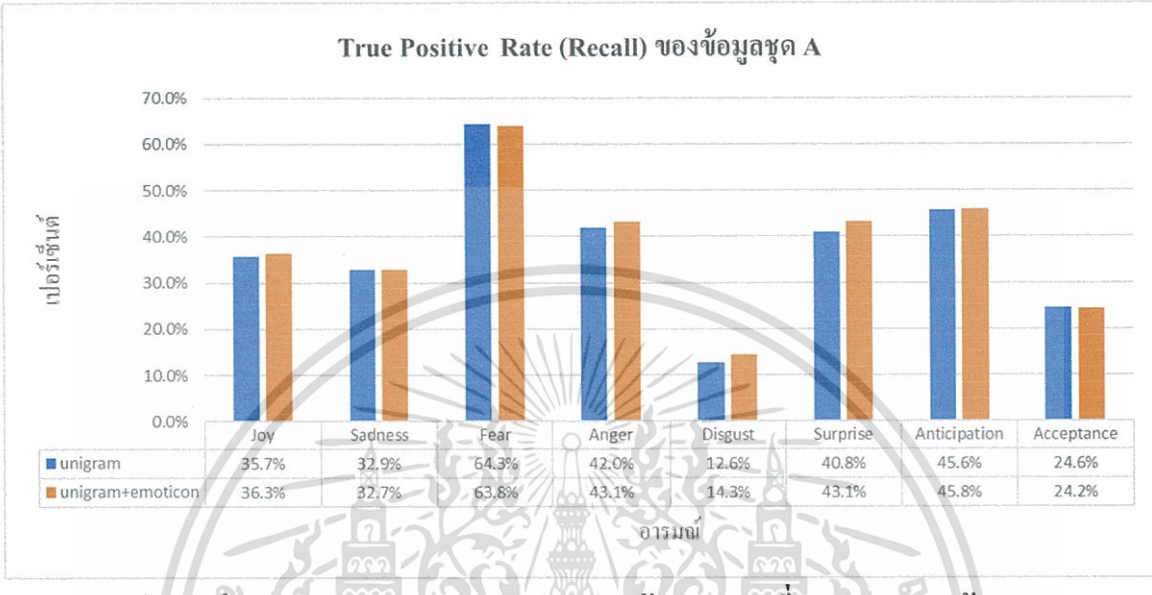
CLASS	True Positive Rate (Recall)	False Positive Rate	Precision
Joy	33.4%	11.1%	42.9%
Sadness	30.4%	4.7%	41.7%
Fear	61.4%	4.2%	42.7%
Anger	40.5%	7.8%	33.8%
Disgust	11.4%	3.2%	3.5%
Surprise	46.9%	17.3%	8.5%
Anticipation	44.2%	16.5%	64.7%
Acceptance	25.0%	8.2%	27.9%



รูป 4.23 ค่าความแม่นยำของแต่ละรูปแบบที่ใช้ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

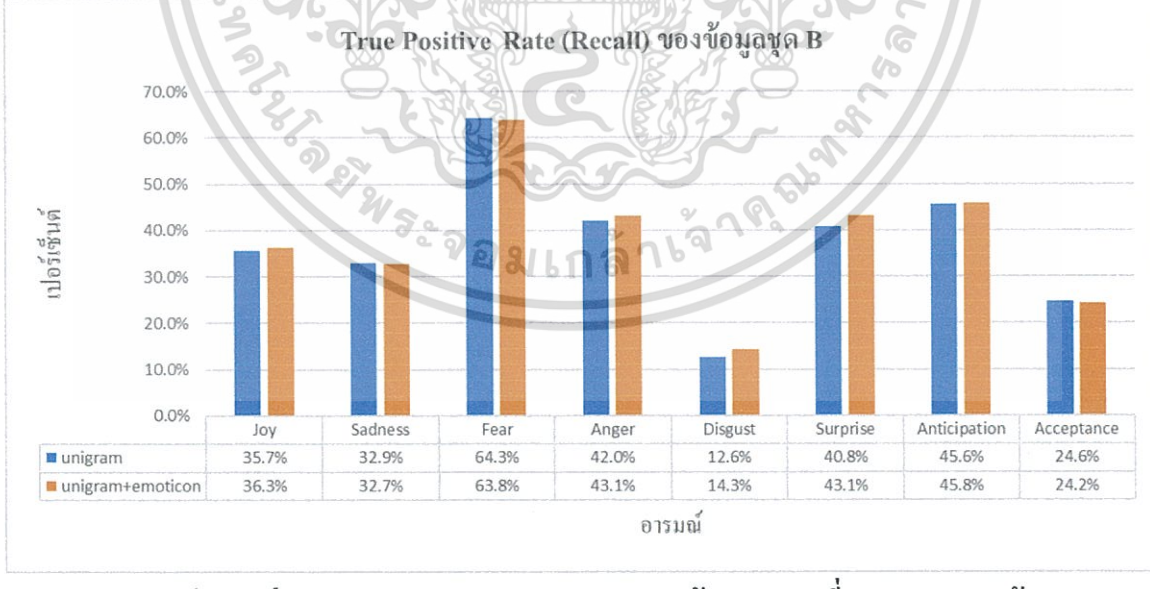
4.4.3.2 ค่าประสิทธิภาพต่าง ๆ ที่ได้จากการทดสอบข้อมูลแต่ละชุด

1) ค่า True Positive Rate (Recall) คือ จำนวนข้อมูลที่ทำทายถูกจากที่ถูกทั้งหมด



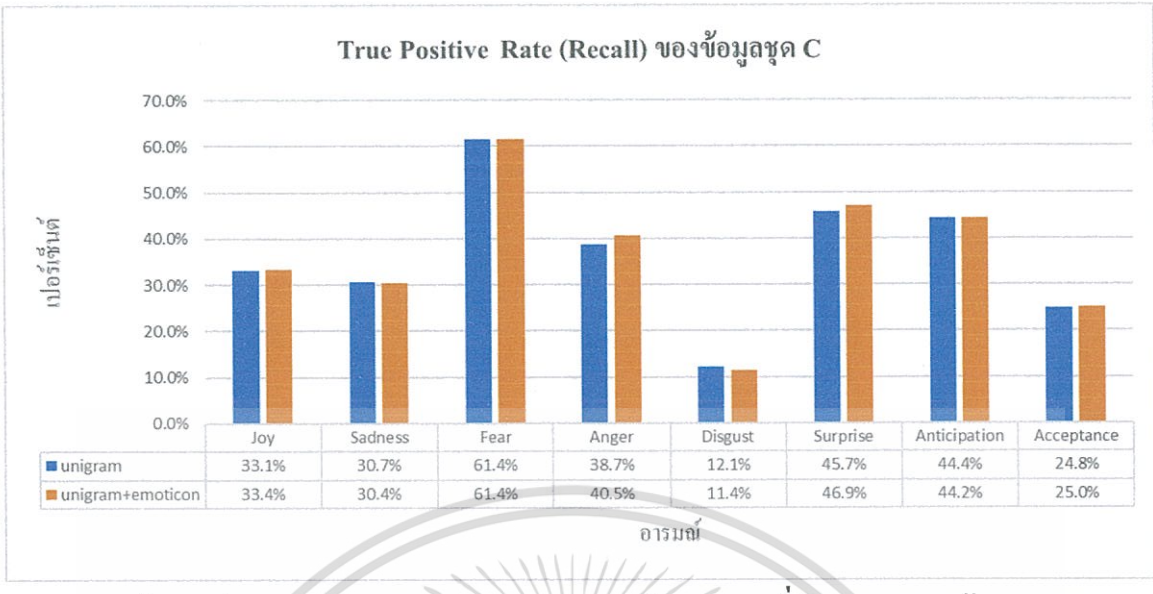
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	35.7%	32.9%	64.3%	42.0%	12.6%	40.8%	45.6%	24.6%
unigram+emoticon	36.3%	32.7%	63.8%	43.1%	14.3%	43.1%	45.8%	24.2%

รูป 4.24 ค่า True Positive Rate (Recall) ของข้อมูลชุด A ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes



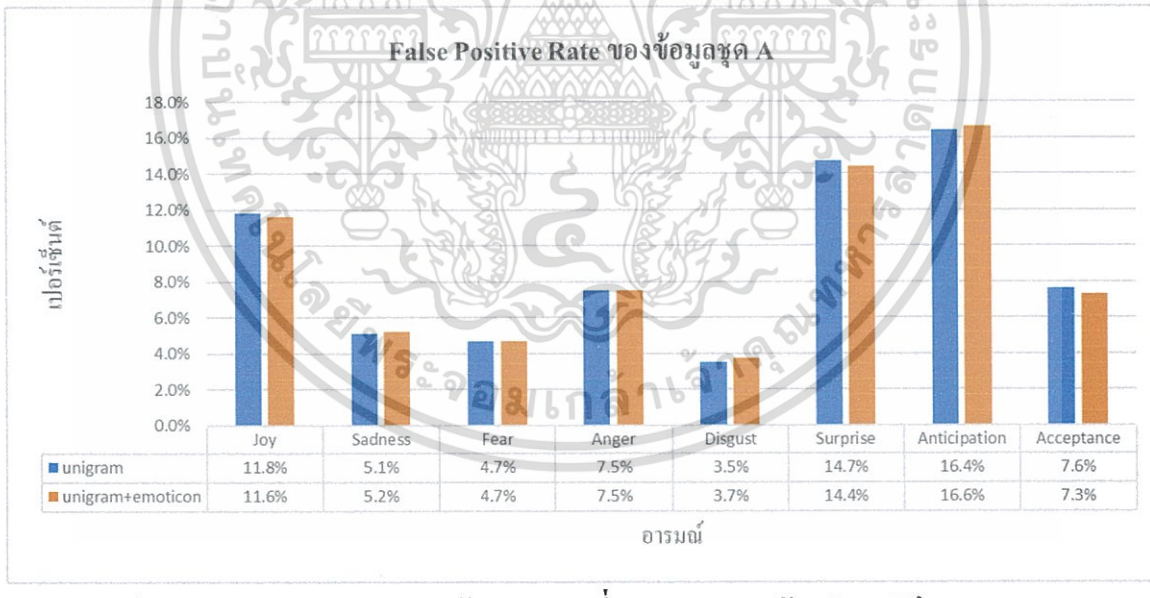
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	35.7%	32.9%	64.3%	42.0%	12.6%	40.8%	45.6%	24.6%
unigram+emoticon	36.3%	32.7%	63.8%	43.1%	14.3%	43.1%	45.8%	24.2%

รูป 4.25 ค่า True Positive Rate (Recall) ของข้อมูลชุด B ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

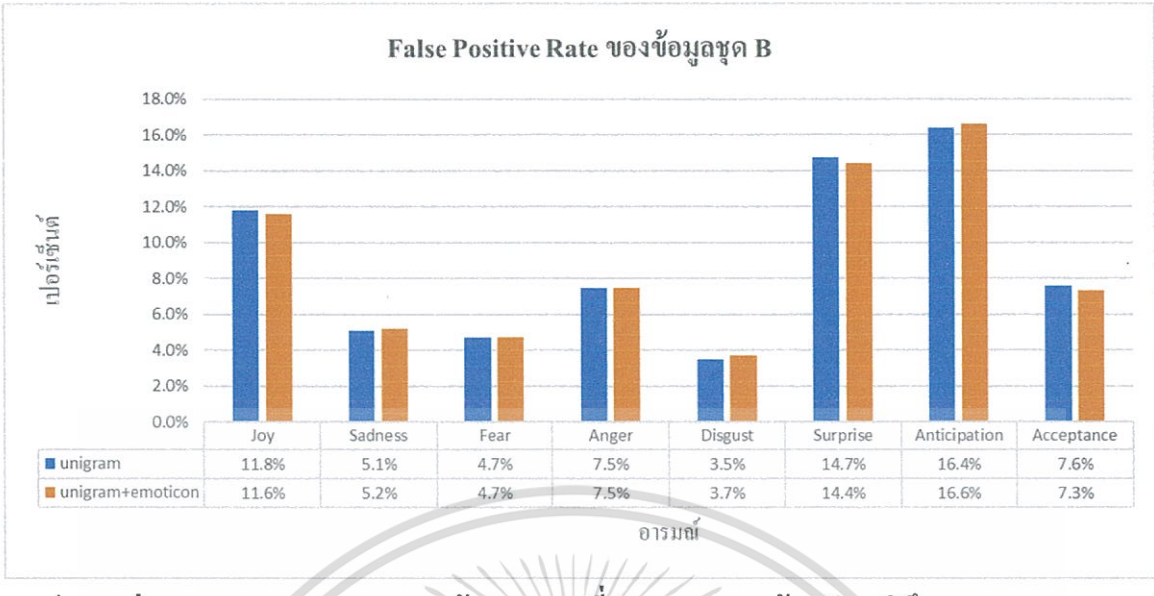


รูป 4.26 ค่า True Positive Rate (Recall) ของข้อมูลชุด C ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

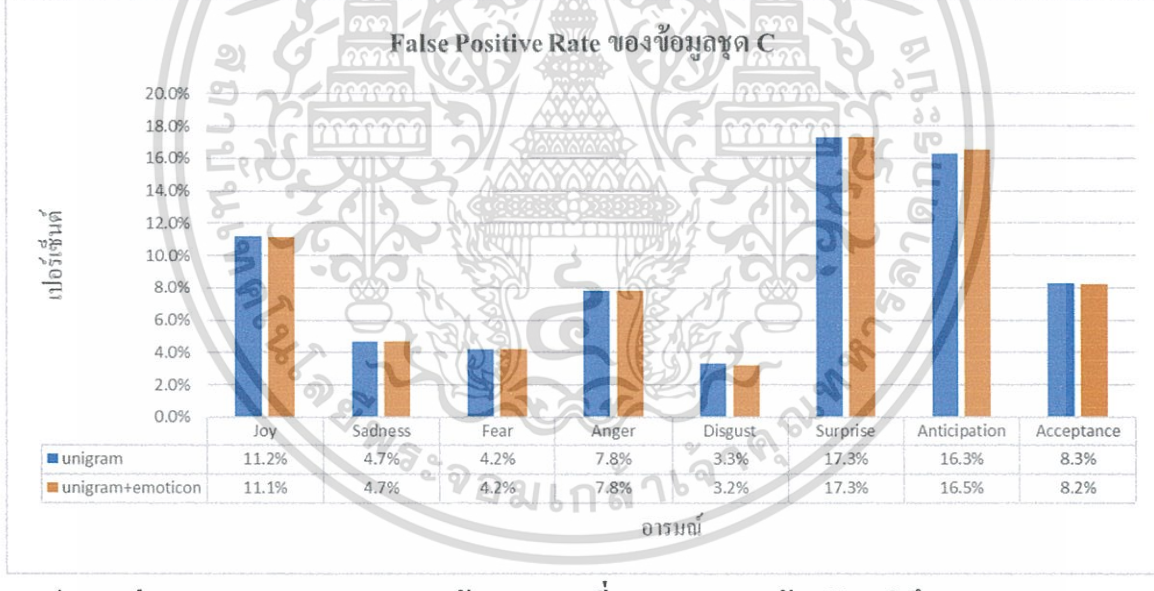
2) ค่า False Positive Rate คือ จำนวนข้อมูลที่ทำนายว่าไม่ถูกจากที่ถูกทั้งหมด



รูป 4.27 ค่า False Positive Rate ของข้อมูลชุด A ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

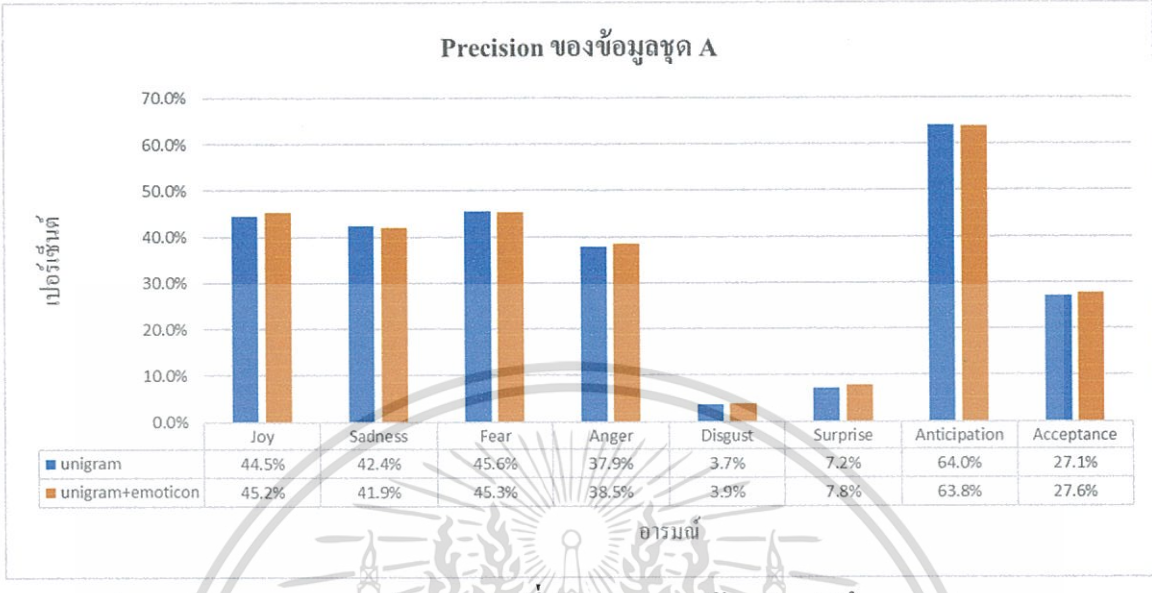


รูป 4.28 ค่า False Positive Rate ของข้อมูลชุด B ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes



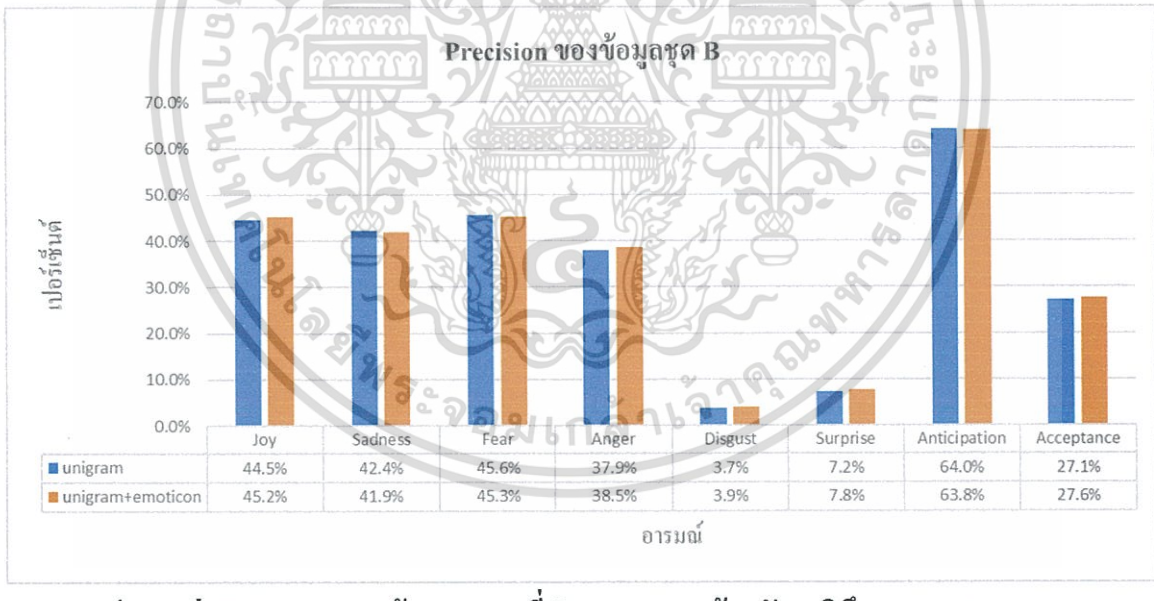
รูป 4.29 ค่า False Positive Rate ของข้อมูลชุด C ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

3) ค่า Precision คือ จำนวนที่ทำนายถูกจากข้อมูลที่ทำนายว่าเป็นคลาสที่พิจารณาอยู่



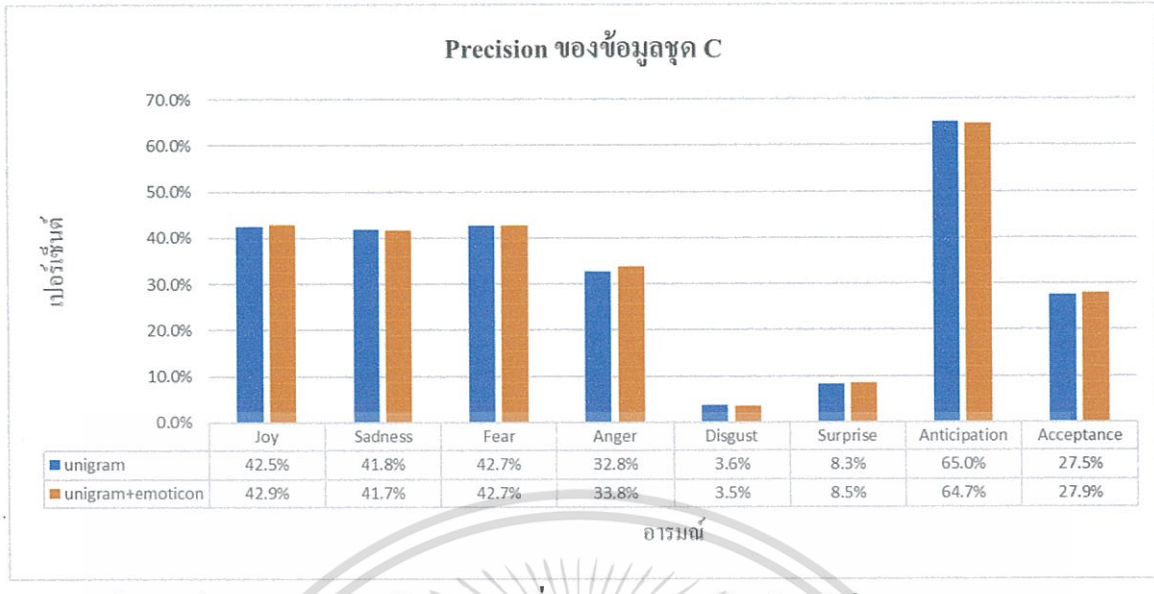
	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	44.5%	42.4%	45.6%	37.9%	3.7%	7.2%	64.0%	27.1%
unigram+emoticon	45.2%	41.9%	45.3%	38.5%	3.9%	7.8%	63.8%	27.6%

รูป 4.30 ค่า Precision ของข้อมูลชุด A ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes



	Joy	Sadness	Fear	Anger	Disgust	Surprise	Anticipation	Acceptance
unigram	44.5%	42.4%	45.6%	37.9%	3.7%	7.2%	64.0%	27.1%
unigram+emoticon	45.2%	41.9%	45.3%	38.5%	3.9%	7.8%	63.8%	27.6%

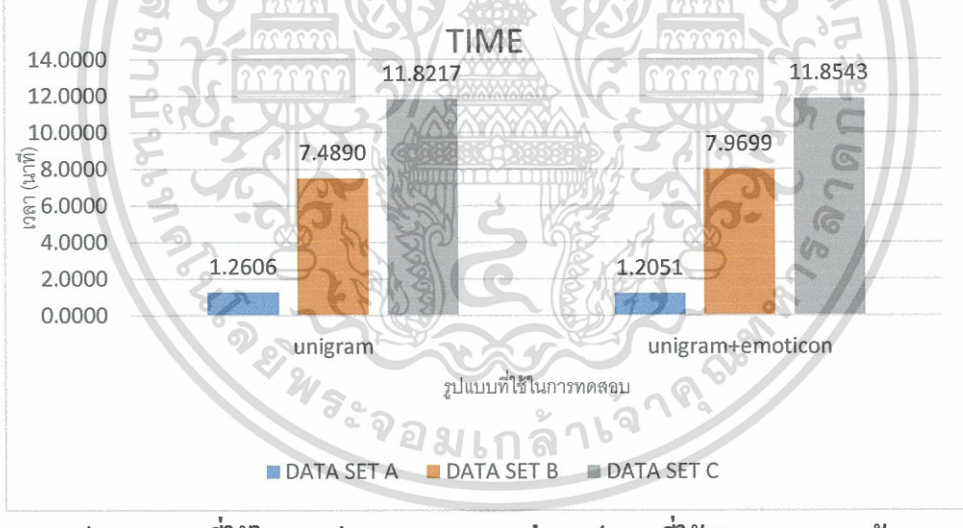
รูป 4.31 ค่า Precision ของข้อมูลชุด B ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes



รูป 4.32 ค่า Precision ของข้อมูลชุด C ที่ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

4.4.3.3 เวลาที่ใช้ในการประมวลผลข้อมูลแต่ละชุด

จากการทดสอบด้วยการประมวลผลของข้อมูลแต่ละชุด จำนวนชุดละ 10 ครั้ง



รูป 4.33 เวลาที่ใช้ในการประมวลผลของแต่ละรูปแบบที่ใช้ทำการทดสอบด้วยอัลกอริทึม Naïve Bayes

4.4.4 สรุปผลการทดลอง

จากการทดลองวัดค่าความแม่นยำและเวลาที่ใช้ประมวลผลด้วยอัลกอริทึม Naïve Bayes ด้วยการทดสอบทั้ง 2 รูปแบบนั้นให้ผลทั้งค่าความแม่นยำและเวลาออกมามีลักษณะใกล้เคียงกันมาก แต่พบว่าการใช้อัลกอริทึม Naïve Bayes นั้นให้ค่าความแม่นยำที่น้อยกว่าการใช้ Rule-Based method (ด้วยรูปแบบของการใช้ keyword ร่วมกับ slang ร่วมกับ emoticon) และอัลกอริทึม Multinomial Naïve Bayes

บทที่ 5
บทสรุปและข้อเสนอแนะ

5.1 บทสรุปของโครงงาน

โครงงานระบบทำนายอารมณ์ตามสถานที่จากสื่อสังคมออนไลน์ที่จัดทำขึ้นมานี้เป็นการนำทฤษฎีทางด้านการทำเหมืองข้อมูลและการเรียนรู้ของเครื่องมาใช้ในการวิเคราะห์ข้อความจากสื่อสังคมออนไลน์ในปัจจุบันที่มีการใช้งานกันอย่างแพร่หลายให้เกิดประโยชน์สูงสุดโดยเลือกที่จะนำมาเป็นตัวช่วยในการตัดสินใจในด้านการท่องเที่ยวโดยอาศัยประสบการณ์จากผู้ที่เคยไปยังสถานที่นั้นมาก่อน มีการแสดงผลออกมาในรูปแบบแอปพลิเคชันให้ผู้ใช้งานเข้าใจได้ง่าย ผู้ใช้งานไม่ต้องเสียเวลาในการอ่านและเข้าถึงข้อมูลจำนวนมากด้วยตนเอง

ในการทำโครงงานนี้พวกเราได้เรียนรู้ในหลายส่วนตั้งแต่เริ่มต้นศึกษาจนสิ้นสุดไม่ว่าจะเป็น

1) ในส่วนของข้อมูลจากสื่อสังคมออนไลน์ที่ตั้งใจจะนำมาใช้นั้นแต่ละประเภทมีรูปแบบที่แตกต่างกันเช่น
   - Facebook ได้ข้อมูลจาเพจแต่เพจส่วนใหญ่จะเป็นข้อความโฆษณา
   - Twitter ได้ข้อมูลที่เป็นสาธารณะและข้อความที่เป็นความคิดเห็นตามสถานที่
   - Foursquare ได้ข้อมูลสถานที่ที่มีการ check in แต่ส่วนมากไม่แสดงข้อความที่เป็นคำพูด

   ซึ่งข้อมูลที่ตรงกับความต้องการใช้งานมากที่สุดคือในส่วนของ twitter
2) ในส่วนของการตัดคำเพื่อที่จะนำมาใช้ บางครั้งคำบางคำอาจจะให้อารมณ์ที่แสดงออกมาไม่เท่ากันเช่น มาก กับ มากกก (เป็นคำที่ไม่เป็นทางการ) บางคำจะเป็นลักษณะของภาษาพูดที่นิยมใช้กันลักษณะเป็นเพียงคำสั้น ๆ แต่อาจจะเป็นตัวที่แสดงถึงอารมณ์ที่แท้จริงมากกว่าคำที่เป็นทางการที่ปรากฏใน dictionary หรือคำที่เป็นประโยค
3) อัลกอริทึมที่นำมาใช้บางอย่างไม่เหมาะกับรูปแบบของข้อมูลที่มีต้องทำการประยุกต์ดัดแปลงให้ตรงกับความต้องการการใช้งาน

ด้วยสิ่งต่าง ๆ ที่พวกเราได้เรียนตลอดการทำโครงงานนี้เราจึงจำเป็นที่จะต้องสร้างระบบให้ออกมารองรับกับความต้องการมากที่สุด

ในส่วนการวิธีการที่ได้นำมาใช้ในการทำนายอารมณ์ตามสถานที่นั้นระบบเลือกใช้ในส่วนของ Multinomial Naive Bayes มาใช้เป็นตัวหลักในการทำนายอารมณ์เนื่องจากค่าความแม่นยำที่ได้จากชุดข้อมูลที่ใช้ในการทดสอบมีค่ามากที่สุด

5.2 ปัญหาอุปสรรคและแนวทางแก้ไข

1) การประมวลผลข้อความภาษาไทยทำได้ยากเช่นการตัดคำ ซึ่งในไลบารี่ที่ใช้งานเพื่อตัดคำภาษาไทยยังมีข้อจำกัดอยู่เช่น ถ้าเกิดในข้อความมีอักขระอื่นนอกจากอักขระภาษาไทย จะไม่สามารถตัดคำออกมาได้ ซึ่งวิธีแก้ปัญหาคือต้องตัดขระอื่นที่ไม่ใช่ภาษาไทยออกให้หมดก่อนที่จะใช้ไลบารี่ในการตัดคำ
2) การนับความถี่ของคำ ยิ่งมีคำมาก เวลาที่ใช้ในการนับความถี่ยิ่งใช้มากตามไปด้วย ซึ่งการที่จะทำให้การทำงานเร็วจำเป็นต้องไปศึกษาการเขียนโปรแกรมที่ใช้ Multithreading ซึ่งจำเป็นต้องใช้เวลาศึกษาและทดสอบมากขึ้น
3) การใช้ภาษาไทยในสื่อสังคมออนไลน์โดยส่วนใหญ่จะใช้คำที่ไม่เป็นทางการในการสื่อสาร หรือมีการแปลงคำให้รูปแบบเปลี่ยนไปแต่ความหมายยังคงเดิมเช่น ฟิน จร้า น่าร้าก เป็นต้น ทำให้ยากกต่อการตัดคำ ซึ่งแก้ปัญหาโดยลองใช้ Rule-based method ในการกำหนด keyword เหล่านี้ขึ้นมาแทน และนำมาใช้ในการทำนายอารมณ์
4) สื่อสังคมออนไลน์เช่น Facebook เราไม่สามารถเข้าถึงข้อความที่เป็นสาธารณะได้ ซึ่งการที่จะเข้าถึงได้จำเป็นต้องติดต่อไปยัง Facebook เพื่อของสิทธิในการเข้าถึง ซึ่งใน Facebook เราจะเข้าถึงข้อมูลของเพจตามสถานที่ต่าง ๆ แทน
5) อารมณ์บางอารมณ์คนส่วนมากจะไม่แสดงบนสื่อสังคมออนไลน์ ทำให้ข้อมูลของอารมณ์เหล่านั้นมีไม่เพียงพอที่จะใช้ทำนาย และการดึงข้อมูลจาก Twitter ยังจำเป็นต้องค้นหาผ่าน keyword เพราะปกติแล้วข้อความที่อยู่บน Twitter มีส่วนน้อยที่ทำการเปิดให้ระบบค้นหาผ่านพิกัดค้นหาได้ ทำให้ข้อความที่ได้มาอาจจะมีความหลากหลายน้อยลง เพราะจำเป็นต้องมี keyword นั้นอยู่ในประโยค

5.3 แนวทางในการพัฒนาต่อ

1) ผู้ใช้งานสามารถเลือกสถานที่ ที่ตนเองสนใจด้วยการค้นหาผ่าน keyword ของสถานที่ได้
2) มีการนำข้อมูลจากสื่อสังคมออนไลน์ชนิดอื่นมาร่วมในการวิเคราะห์มากขึ้นไม่ใช่แค่จาก Twitter เพียงแหล่งเดียว
3) พัฒนาในส่วนที่จะนำข้อมูลอารมณ์ต่อสถานที่สนใจในอดีตรวบรวมนำมาเพื่อใช้ในการทำนายผลลัพธ์ที่จะเกิดขึ้นในอนาคตของสถานที่ดังกล่าว

บรรณานุกรม

[1] Nivet Chirawichitchai, "Emotion Classification of Thai Text based UsingTerm weighting and Machine Learning Techniques." in *11th IEEE International Joint Conference on Computer Science and Software Engineering (JCSSE)*, p 91-96, 2014.

[2] S.L. Ting, W.H. Ip, Albert H.C. Tsang, "Is Naïve Bayes a Good Classifier for Document Classification?." in *International Journal of Software Engineering and Its Applications*, p. 37-46, 2011.

[3] Michael Negnevitsky, "Artificial Intelligence A Guide to Intelligent Systems." 2nd ed. Edinburgh Gate: Pearson Education Limited, p. 25-28, 2005.

[4] Ian H. Witten & Eibe Frank, "DATA MINING Practical Machine Learning Tools and Techniques." 2nd ed. United States: Morgan Kaufmann, p. 94-97, 2005.

[5] Steven Bird, Ewan Klein, Edward Loper, "Natural Language Processing with Python." USA: O'Reilly Media, Inc, 2009.

[6] Apache Software Foundation, 2014, "Apache parquet documentation." [Online]. Available: https://goo.gl/S6kqMA

[7] Julien Le Dem, 2013, "Dremel made simple with Parquet." [Online]. Available: https://goo.gl/K9mPOJ

[8] "Apache Spark." [Online]. Available: https://goo.gl/5REXnm

[9] The Apache Software Foundation, "Apache Spark™ - Lightning-Fast Cluster Computing." [Online]. Available: https://goo.gl/7TRBpv

[10] Kendra Cherry, 2016, "Theories of Emotion" [Online]. Available:
https://goo.gl/v9TdlC

[11] Wikipedia, "Robert Plutchik, Theory of emotion." [Online].
Available: https://goo.gl/Xhww2B

[12] Margaret Rouse, "text mining (text analytics)." [Online].
Available: https://goo.gl/JBiWgq

[13] Jeffrey Strickland, Ph.D., CMSP, "Is Machine Learning about Machines Learning?."
[Online]. Available: https://goo.gl/28PDK9

[14] Sunil Ray, "Easy Steps to Learn Naive Bayes Algorithm." [Online].
Available: https://goo.gl/7cESGV

[15] Ahmet Taspinar, "Text Classification and Sentiment Analysis." [Online].
Available: https://goo.gl/cPJxGx

[16] The Apache Software Foundation, "Apache™ Hadoop®". [Online]
Available: https://goo.gl/p6Kp3D

[17] Sigma Xi Members(The Scientific Research Society), "American Scientist, Volumn 89 ",
p.344-350, 2001.