การตรวจจับหัวข้ออัตโนมัติบนข้อมูลทวิตเตอร์โดยการใช้คุณลักษณะจากตัวชี้วัดของหุ้น

นายเอกภพ วีระสกุลวงศ์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

Automatic Topic Detection on Twitter Data Using Stock Indicator Based Features

Mr. Ekapop Verasakulvong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

หัวข้อวิทยานิพนธ์	การตรวจจับหัวข้ออัตโนมัติบนข้อมูลทวิตเตอร์โดยการใช้คุณลักษณะจากตัวชี้วัดของหุ้น
โดย	นายเอกภพ วีระสกุลวงศ์
สาขาวิชา	วิศวกรรมคอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.พีรพล เวทีกูล
อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม	ดร.อภิวดี ปิยธรรมรงค์

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.วีระ เหมืองสิน)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.พีรพล เวทีกูล)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ดร.อภิวดี ปิยธรรมรงค์)

..... กรรมการภายนอกมหาวิทยาลัย
(รองศาสตราจารย์ ดร.อานนท์ รุ่งสว่าง)

เอกภพ วีระสกุลวงศ์ : การตรวจจับหัวข้ออัตโนมัติบนข้อมูลทวิตเตอร์โดยการใช้คุณลักษณะจากตัวชี้วัดของหุ้น. (Automatic Topic Detection on Twitter Data Using Stock Indicator Based Features) อ.ที่ปรึกษาหลัก : ผศ. ดร.พีรพล เวทีกูล, อ.ที่ปรึกษาร่วม : ดร.อภิวดี ปิยธรรมรงค์

สื่อสังคมออนไลน์เป็นหนึ่งในการสื่อสารที่สำคัญและรวดเร็วที่สุดในปัจจุบัน การสังเกตการณ์ข้อมูลทวิตเตอร์ทำให้สามารถตรวจจับเหตุการณ์ที่กำลังเป็นที่สนใจแบบใกล้ทันกาลหรือหัวข้อเกิดใหม่ได้ โดยหัวข้อเกิดใหม่แต่ละหัวข้อจะประกอบด้วยกลุ่มของคำที่เกี่ยวข้องหรือกลุ่มของคำเกิดใหม่ งานวิจัยหลายงานนำเสนอวิธีการตรวจจับกลุ่มคำเหล่านี้โดยใช้คุณลักษณะที่สร้างจากสถิติของคำที่อยู่ในข้อความทวิตเตอร์ ซึ่งบางคุณลักษณะมีความคล้ายคลึงกับตัวชี้วัดของหุ้น แต่อย่างไรก็ตามวิธีเหล่านี้ใช้เพียงคุณลักษณะเดียว ซึ่งเป็นการยากที่จะตรวจจับคำเกิดใหม่ได้หลากหลายรูปแบบ แม้จะมีบางงานวิจัยพยายามใช้หลายคุณลักษณะด้วยตัวจำแนกประเภท แต่ด้วยข้อจำกัดของการสร้างตัวแปรผลเฉลยของข้อมูลที่ใช้ในการสอนตัวจำแนกประเภท ทำให้ยากต่อการนำไปใช้ นอกจากนี้ในงานวิจัยที่เกี่ยวข้องกับการตรวจจับหัวข้อเกิดใหม่ ไม่มีชุดผลเฉลยที่ชัดเจน และไม่มีการวัดประสิทธิภาพที่เป็นมาตรฐาน ในงานวิจัยนี้จึงเสนอการตรวจจับหัวข้อเกิดใหม่ด้วยคุณลักษณะจากตัวชี้วัดของหุ้นที่นิยมใช้ในปัจจุบันและมีการปรับปรุงคุณลักษณะดังกล่าวให้ดียิ่งขึ้น อีกทั้งตัวจำแนกที่ได้ประสิทธิภาพสูงสุด ซึ่งได้แก่ป่าไม้แบบสุ่ม ถูกนำมาใช้ในการตรวจจับคำเกิดใหม่โดยไม่มีข้อจำกัดในการสร้างตัวแปรผลเฉลยของข้อมูล สุดท้ายเพื่อให้สามารถวัดประสิทธิภาพในการตรวจจับคำและหัวข้อเกิดใหม่ จึงทำการสร้างชุดผลเฉลยรายวันและใช้ชุดผลเฉลยดังกล่าวในการวัดประสิทธิภาพด้วยมาตรวัดประสิทธิภาพของหัวข้อแบบแมโครที่สามารถวัดประสิทธิภาพในแง่มุมของคำและหัวข้อเกิดใหม่พร้อมกัน จากการทดลองพบว่าประสิทธิภาพของวิธีที่นำเสนอในงานวิจัยนี้ สามารถตรวจจับคำและหัวข้อเกิดใหม่ได้อย่างมีประสิทธิภาพดีกว่าวิธีการในปัจจุบัน ได้แก่ SigniTrend และ TopicSketch นอกจากนี้ยังพบว่าวิธีที่นำเสนอในงานวิจัยนี้สามารถตรวจจับคำและหัวข้อเกิดใหม่ได้ก่อนงานวิจัยอื่น

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

ลายมือชื่อ อ.ที่ปรึกษาร่วม .....

5870284521 : MAJOR COMPUTER ENGINEERING
KEYWORD: emerging topic, emerging keyword, twitter, random forest, stock indicator, trend, social network

Social media is one of the most impactful and fastest communication methods. By monitoring Twitter streams, we are able to detect an interesting event, “emerging topic”, in near real-time. Each emerging topic contains a group of related keywords or “emerging keyword”. Some prior attempts aim to detect topics on Twitter based on word’s statistic, where some of them are similar to stock indicators. However, they only use univariate feature. Therefore, it is hard to detect various types of emerging topics. Although some research uses multivariate features with a classifier, its use case is very limited due to constraints in the data labeling process. Moreover, there are no standard answer set and no standard performance evaluation method in this research field. In this research, we propose an algorithm to detect emerging topics on Twitter streams. Many stock indicators are applied and improved. To capture event patterns, various classifiers are compared and RandomForest is selected. Moreover, there are no limitations in the data labeling process. We create a daily answer set and new standard evaluation metric called “Macro topic measurement” which can evaluate the performance of both keyword and topic detection. The experiment shows that our method outperforms other baselines: SigniTrend, and TopicSketch. Furthermore, our method can detect emerging keyword earlier than other baselines.

Field of Study: Computer Engineering Student's Signature .....
Academic Year: 2018 Advisor's Signature .....
Co-advisor's Signature .....

กิตติกรรมประกาศ

ขอบพระคุณ ผศ. ดร. วีระ เหมืองสิน และ รศ. ดร. อานนท์ รุ่งสว่าง ผู้ที่ให้เกียรติเป็นประธานและกรรมการสอบวิทยานิพนธ์ของข้าพเจ้า

ขอบพระคุณ ผศ. ดร. พีรพล เวทีกูล และ ดร. อภิวดี ปิยธรรมรงค์ อาจารย์ที่ปรึกษาวิทยานิพนธ์ รวมถึง คุณ ชัชวาล สังคีตตระการ ที่ให้คำปรึกษาในด้านต่าง ๆ ไม่ว่าจะเป็นด้านงานวิจัย ด้านการเรียน ด้านการทำงาน รวมถึงมอบประสบการณ์ต่าง ๆ ให้แก่ผู้วิจัย

ขอบพระคุณ ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่เชื่อมั่นและมอบทุนอุดหนุนการศึกษา "ทุนอัจฉริยะคืนรัง" ให้แก่ผู้วิจัย และขอบคุณ อาจารย์และเจ้าหน้าที่ประจำภาควิชาฯ ที่ช่วยเหลือผู้วิจัยในด้านต่างๆเสมอมา

ขอบพระคุณ สมาชิกห้องปฏิบัติการเหมืองข้อมูล (DataMind Lab) โดยเฉพาะพี่นัท พี่ตง พี่แคน พี่บอส พี่อิฐ พี่ปลา ชมพู่ และเก้า ที่คอยช่วยเหลือ และเป็นกำลังใจให้ผู้วิจัยเสมอมา

สุดท้ายนี้ ขอบพระคุณ คุณพ่อ คุณแม่ และครอบครัว ที่เป็นกำลังใจและสนับสนุนผู้วิจัยในทุก ๆ ด้าน ขอบคุณครับ

เอกภพ วีระสกุลวงศ์

สารบัญ

หน้า

สารบัญรูป

สารบัญตาราง

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันข้อมูลข่าวสารต่าง ๆ สามารถกระจายออกไปอย่างรวดเร็ว เนื่องจากความสะดวกในการเข้าถึงอินเทอร์เน็ตผ่านอุปกรณ์หลากหลายชนิดทั่วทุกมุมของโลก ซึ่งผู้ใช้อินเทอร์เน็ตมักมีพฤติกรรมการใช้งานร่วมกันคือ รับและเผยแพร่ข้อมูลที่ตนสนใจหรือเหตุการณ์ต่าง ๆ ที่อยู่รอบตัวแก่บุคคลอื่นผ่านช่องทางต่าง ๆ เช่น เว็บบล็อก (Web blog), เครือข่ายสังคมออนไลน์ (Social network), วีดีโอออนไลน์ (Online video) โดยเฉพาะอย่างยิ่งบล็อกสั้น (Micro blog) ซึ่งมีลักษณะเป็นเว็บไซต์ที่เปิดให้บุคคลทั่วไปแชร์ข้อความขนาดสั้น ในปัจจุบันทวิตเตอร์ (Twitter) จัดเป็นบล็อกสั้นที่ได้รับความนิยมสูงสุด ด้วยเหตุนี้จึงมีความพยายามที่จะวิเคราะห์พฤติกรรมการใช้งานทวิตเตอร์เพื่อใช้ประโยชน์ในด้านต่าง ๆ เช่น การหาความสัมพันธ์ของผู้ใช้ (Social relationship) การหาผู้ใช้งานที่มีอิทธิพล (Influencer) การหาเส้นทางการกระจายข้อมูล (Spreading path) และ การตรวจจับหัวข้อเกิดใหม่ (Emerging topic) เป็นต้น

การตรวจจับหัวข้อเกิดใหม่ เป็นหนึ่งในการวิเคราะห์ข้อมูลทวิตเตอร์ที่น่าสนใจ เนื่องจากสามารถนำมาใช้ในการสังเกตการณ์หัวข้อที่ผู้คนกำลังสนใจในทวิตเตอร์ได้ตลอดเวลา โดยจะทำการตรวจจับหากลุ่มคำที่กำลังเป็นที่สนใจ หรือ “กลุ่มคำเกิดใหม่” จากนั้นหาความสัมพันธ์ระหว่างคำเพื่อจับกลุ่มคำที่เกี่ยวข้องเป็นหัวข้อเดียวกัน หรือ “หัวข้อเกิดใหม่” การตรวจจับหัวข้อเกิดใหม่สามารถนำไปต่อยอดในด้านอื่น ๆ ได้ เช่น ใช้ในการโฆษณาร่วมกับผลิตภัณฑ์อื่น ดังตัวอย่างในรูปที่ 1 เป็นโพสต์โฆษณาถุงยางอนามัยร่วมกับภาพยนตร์ก็อดซิลล่า เนื่องจากเป็นวันที่ภาพยนตร์ก็อดซิลล่าเข้าฉายในโรงภาพยนต์วันแรก หรือโฆษณาที่ใช้ประโยชน์จากกระแสของ Mac Pro ที่เพิ่งเปิดตัวในรูปที่ 2 โดยการนำเสนอผลิตภัณฑ์ที่มีรูปร่างคล้ายกับ Mac Pro ในรูปที่ 3

ปัจจุบันมีงานวิจัยเกี่ยวกับการตรวจจับหัวข้อเกิดใหม่จำนวนมาก [1-8] โดยมีบางงานวิจัยเสนอคุณลักษณะ (Feature) ที่คล้ายกับตัวชี้วัดของหุ้น MACD (Moving Average Convergence Divergence) [7] แต่งานวิจัยเหล่านี้ใช้เพียงคุณลักษณะเดียว (Univariate feature) ในการตรวจจับคำเกิดใหม่ ซึ่งเป็นการยากที่จะตรวจจับคำเกิดใหม่ได้หลากหลายรูปแบบ แม้จะมีบางงานวิจัย [9] พยายามใช้หลายคุณลักษณะ (Multivariate features) ในการตรวจจับคำเกิดใหม่ด้วยตัวจำแนกประเภท แต่ด้วยวิธีการกำกับข้อมูลหรือการสร้างตัวแปรผลเฉลย (Data labeling) อาศัย

ค่าพารามิเตอร์หลายตัว ทำให้ยากต่อการหาค่าพารามิเตอร์และการนำไปใช้ เนื่องจากหากใช้ค่าพารามิเตอร์ไม่เหมาะสม อาจทำให้บางหัวข้อเกิดใหม่ไม่ปรากฏในชุดข้อมูล หรือ ทำให้เกิดการกำกับข้อมูลส่วนที่ไม่ใช่เหตุการณ์ได้ นอกจากนี้ในงานวิจัยเกี่ยวกับการตรวจจับหัวข้อเกิดใหม่ ไม่มีชุดผลเฉลยที่ชัดเจน และไม่มีการวัดประสิทธิภาพที่เป็นมาตรฐาน ทำให้การวัดผลลัพธ์ของงานวิจัยส่วนมากจะมุ่งเน้นไปที่เหตุการณ์สำคัญและเด่นชัดเท่านั้น เช่น แผ่นดินไหว การแข่งขันฟุตบอลโลก เหตุการณ์ก่อการร้าย เป็นต้น

ในงานวิจัยฉบับนี้ต้องการความแม่นยำและความครอบคลุมเพิ่มขึ้นกว่างานในอดีต จึงเสนอการตรวจจับคำและหัวข้อเกิดใหม่ด้วยคุณลักษณะจากตัวชี้วัดของหุ้นที่หลากหลายมากขึ้น ได้แก่ rsi, stoch และ wr เนื่องจากตัวชี้วัดของหุ้นดังกล่าวมักถูกใช้ในการหาจังหวะเข้าซื้อ หรือขายออก โดยการหาสถานะราคาหุ้น ได้แก่ สถานะซื้อมากเกินไป (Overbought), สถานะขายมากเกินไป (Oversold) หรือสถานะปกติ ซึ่งสถานะซื้อมากเกินไปหรือราคาสูงผิดปกตินั้น สามารถมองเป็นสถานะที่จำนวนคำปรากฏมากผิดปกติ หรือเป็นคำที่มีโอกาสเป็นคำเกิดใหม่ นอกจากนี้ในงานวิจัยนี้ยังปรับปรุงการสร้างคุณลักษณะดังกล่าวให้เหมาะสมกับงานวิจัยนี้ยิ่งขึ้น ทำให้จำนวนคุณลักษณะในงานวิจัยนี้มีจำนวนมาก ตัวจำแนกประเภทจึงถูกนำมาประยุกต์ใช้ในการตัดสินใจในการหากลุ่มคุณลักษณะและเส้นแบ่งที่ดีที่สุดด้วยกลุ่มของต้นไม้ตัดสินใจ (Decision tree) หรือป่าไม้แบบสุ่ม (Random forest) นอกจากนี้ชุดข้อมูลสอนที่ใช้ในการสร้างตัวจำแนกประเภท ในส่วนของการกำกับข้อมูลหรือการสร้างตัวแปรผลเฉลย (Data labeling) ในงานวิจัยนี้ไม่จำเป็นต้องใช้ค่าพารามิเตอร์และสามารถกำกับข้อมูลได้ครอบคลุมทุกหัวข้อในชุดผลเฉลย ทำให้ตัวจำแนกประเภทสามารถตรวจจับคำได้ครอบคลุม สุดท้ายเพื่อให้สามารถวัดประสิทธิภาพในการตรวจจับคำและหัวข้อเกิดใหม่ จึงทำการสร้างชุดผลเฉลยรายวันอันประกอบด้วย “วัน”, “หัวข้อ”, “กลุ่มคำที่เกี่ยวข้อง” และใช้ชุดผลเฉลยดังกล่าวในการวัดประสิทธิภาพด้วยมาตรวัดประสิทธิภาพของหัวข้อแบบแมโคร (Macro Topic) ที่สามารถวัดประสิทธิภาพในแง่มุมของคำ และหัวข้อเกิดใหม่พร้อมกัน

รูปที่ 1 การนำหัวข้อของภาพยนตร์ที่เป็นกระแสมาโฆษณาร่วมกับถุงยางอนามัย [1]

รูปที่ 2 Mac Pro ที่เปิดตัวในงาน WWDC 2019 [2]

[1] อ้างอิงจาก https://www.facebook.com/OkamotoThailandOfficial

[2] อ้างอิงจาก https://www.soyacincau.com/2019/06/04/mac-pro-wwdc-2019-apple/

รูปที่ 3 การนำหัวข้อ Mac Pro ที่เป็นกระแสมาโฆษณาร่วมกับของในร้านค้า [3]

1.2 วัตถุประสงค์ของการวิจัย

เพื่อพัฒนาวิธีการตรวจจับหัวข้อเกิดใหม่ ที่สามารถตรวจจับหัวข้อเกิดใหม่ได้แม่นยำและครอบคลุมมากขึ้น กล่าวคือสามารถตรวจจับเรื่องราวที่เกิดขึ้นจริงและมีคนส่วนหนึ่งให้ความสนใจได้ ซึ่งควรมีประสิทธิภาพดีกว่าวิธีการปัจจุบัน ได้แก่ SigniTrend [6] และ TopicSketch [7]

1.3 ขอบเขตของการวิจัย

งานวิจัยนี้ทำการตรวจจับคำและหัวข้อเกิดใหม่บนชุดข้อความทวิตเตอร์ภาษาไทย ที่เก็บรวบรวมโดยใช้ทวิตเตอร์เอพีไอแบบเรส ระหว่างวันที่ 12 มิถุนายน ค.ศ. 2018 ถึง 20 มิถุนายน ค.ศ. 2018 รวมเป็นระยะเวลา 9 วัน โดยมีจำนวนข้อความประมาณ 4.8 ล้านข้อความ

ชุดผลเฉลยที่ใช้ในงานวิจัยนี้เป็นชุดผลเฉลยรายวัน ถูกสร้างขึ้นบนช่วงเวลาเดียวกับชุดข้อความทวิตเตอร์ โดยมีโครงสร้างประกอบด้วย “วัน”, “หัวข้อ”, “กลุ่มคำที่เกี่ยวข้อง” และมีจำนวนคำและหัวข้อในชุดผลเฉลย จำนวน 1,629 คำ และ 201 หัวข้อ ตามลำดับ

ผลลัพธ์สุดท้ายของการตรวจจับคำและหัวข้อเกิดใหม่ เป็นหัวข้อเกิดใหม่รายวัน โดยในแต่ละหัวข้อเกิดใหม่ประกอบด้วยกลุ่มของคำเกิดใหม่ที่เกี่ยวข้องกับหัวข้อดังกล่าว

---

[3] อ้างอิงจาก https://www.facebook.com/wellmartthailand/

1.4 ประโยชน์ที่ได้รับ

1. สามารถตรวจจับคำและหัวข้อเกิดใหม่ได้แม่นยำและครอบคลุม
2. สามารถตรวจจับคำและหัวข้อเกิดใหม่ได้อย่างรวดเร็ว
3. สามารถตรวจจับคำและหัวข้อเกิดใหม่แบบใกล้ทันกาล
4. สามารถอธิบายเหตุการณ์ต่าง ๆ จากผลลัพธ์ของการตรวจจับคำและหัวข้อเกิดใหม่ได้

1.5 ผลงานตีพิมพ์จากการวิจัย

Verasakulvong, E., Vateekul, P., Piyatumrong, A., & Sangkeettrakarn, C. (2018, July). Online Emerging Topic Detection on Twitter Using Random Forest with Stock Indicator Features. In 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.

บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

ในบทที่ 2 จะนำเสนอเกี่ยวกับทฤษฎีที่ใช้ในงานวิจัย ได้แก่

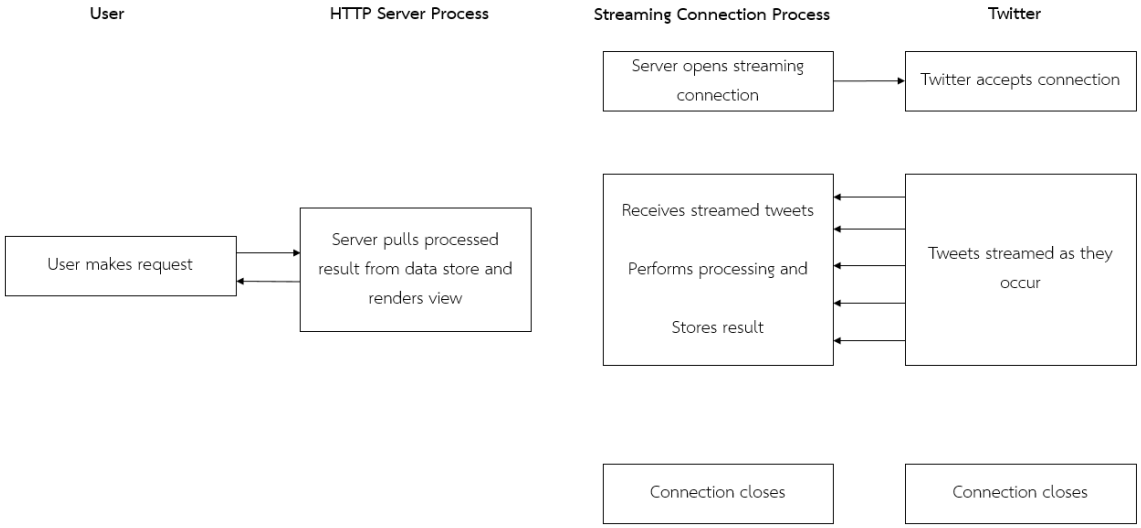
1. ทวิตเตอร์เอพีไอ (Twitter API) : ใช้ในการเก็บรวบรวมข้อความทวิตเตอร์ภาษาไทย
2. ตัวชี้วัดของหุ้น (Stock Indicator) : ใช้ในการสร้างคุณลักษณะในการตรวจจับคำเกิดใหม่ ในบทที่ 5.2.2 ถึง 5.2.5
3. ป่าไม้แบบสุ่ม (Random Forest) : เป็นตัวจำแนกประเภทที่ใช้ในการทำนายคำเกิดใหม่ ในบทที่ 5.3.2
4. การตรวจสอบแบบไขว้ (Cross Validation) : ใช้ในการตรวจสอบและเลือกพารามิเตอร์ที่ดีที่สุดในการสร้างตัวจำแนกประเภท รวมถึงเส้นแบ่ง (Threshold) ของแต่ละคุณลักษณะ
5. การกระจายเมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Vector Decomposition หรือ SVD) : ใช้ในการแตกองค์ประกอบเมทริกซ์ความสัมพันธ์ระหว่างคำกับข้อความ ในการทำดัชนีความหมายแฝง (Latent Semantic Indexing หรือ LSI)
6. การทำดัชนีความหมายแฝง (Latent Semantic Indexing หรือ LSI) : ใช้ในการหาเวกเตอร์ของคำแต่ละคำในกลุ่มของข้อความ เพื่อหาความคล้ายของแต่ละคู่ของคำในขั้นตอนการตรวจจับหัวข้อเกิดใหม่ ในบทที่ 5.4.1
7. การวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภททั่วไป (Classification Performance Evaluation) : เป็นพื้นฐานในการวัดประสิทธิภาพ
8. การวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภทที่มีหลายคลาส (Multiclass Classification Performance Evaluation) : ใช้ในการสร้างมาตรวัดประสิทธิภาพของหัวข้อแบบแมโคร (Macro Topic) ที่สามารถวัดประสิทธิภาพในมุมมองของคำและหัวข้อได้พร้อมกัน ในบทที่ 4.3.3

2.1 ทวิตเตอร์เอพีไอ (Twitter API)

ทวิตเตอร์เอพีไอ เป็นเอพีไอที่ใช้สำหรับเปิดช่องทางให้ผู้ใช้งานเชื่อมต่อกับทวิตเตอร์แทนการใช้งานผ่านเว็บไซต์ โดยสามารถโพสต์ข้อความ, ค้นหาข้อความทวิตเตอร์, ร้องขอรายชื่อผู้ใช้งานที่ติดตามหรือถูกติดตาม ซึ่งในงานวิจัยนี้ทวิตเตอร์เอพีไอแบบเรสถูกใช้ในการร้องขอข้อความทวิตเตอร์ภาษาไทย

2.1.1 ทวิตเตอร์เอพีไอแบบกระแสข้อมูล (Twitter Streaming API)

ทวิตเตอร์เอพีไอแบบกระแสข้อมูล หรือทวิตเตอร์เอพีไอแบบสตรีมมิ่ง เป็นเอพีไอที่ใช้สำหรับเปิดช่องทางให้ผู้ใช้งานเชื่อมต่อ และจะได้รับการตอบสนองกลับมาเป็นชุดข้อความอย่างต่อเนื่องจนกว่าการเชื่อมต่อจะถูกปิดลง โดยข้อความที่ส่งกลับมานั้นเป็นข้อความที่เพิ่งเกิดขึ้นเสมอ โดย รูปที่ 4 แสดงการทำงานของทวิตเตอร์เอพีไอแบบกระแสข้อมูล

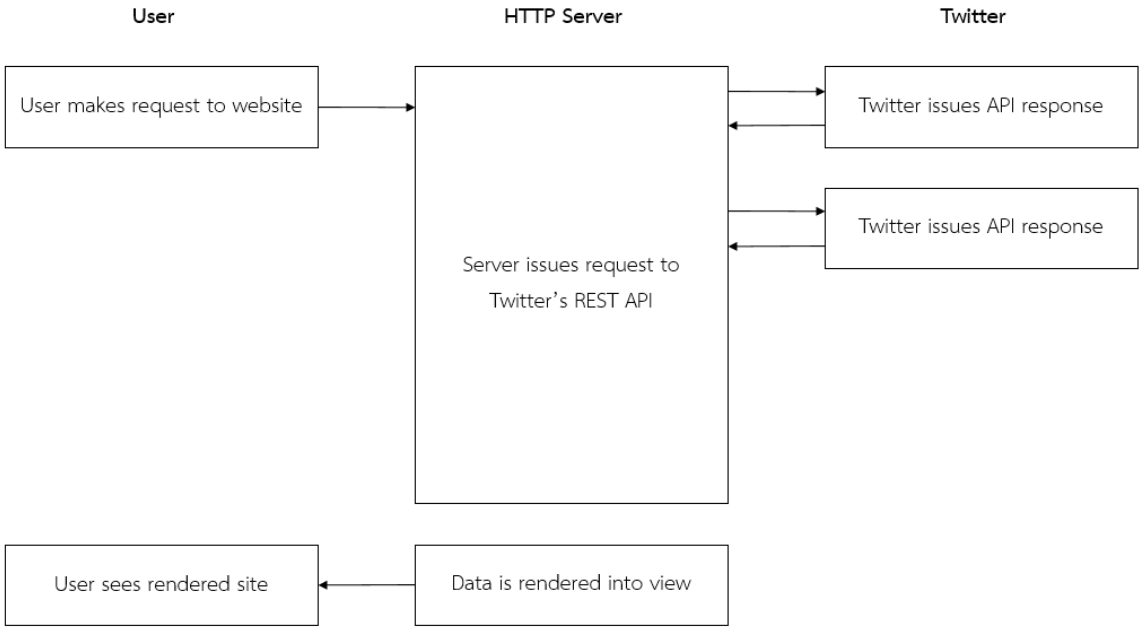


รูปที่ 4 การทำงานของทวิตเตอร์เอพีไอแบบกระแสข้อมูล [4]

2.1.2 ทวิตเตอร์เอพีไอแบบเรส (Twitter REST API)

ทวิตเตอร์เอพีไอแบบเรส (REST หรือ Representational State Transfer) เป็นเอพีไอที่ใช้สำหรับร้องขอกลุ่มของข้อความทวิตเตอร์ โดยคำร้องที่ส่งไปจะได้รับการตอบสนองเป็นชุดของข้อมูลแล้วตัดการเชื่อมต่อ โดยข้อความที่สามารถร้องขอได้นั้นเป็นข้อความในอดีตที่เพิ่งเกิดไม่เกิน 2 สัปดาห์ ดังนั้น สามารถประยุกต์การใช้งานทวิตเตอร์เอพีไอแบบเรส มาเก็บข้อความทวิตเตอร์แบบทันกาล (Real-time) ได้โดยการร้องขอทวิตเตอร์ผ่านทวิตเตอร์เอพีไอแบบเรสทุก ๆ 10 วินาที โดยมีการส่งไอดีของข้อความที่มากที่สุดของข้อความทวิตเตอร์ได้รับตอนร้องขอข้อความทวิตเตอร์ เพื่อไม่ให้เซิร์ฟเวอร์ทวิตเตอร์ส่งข้อความที่เคยส่งมาแล้วซ้ำ รูปที่ 5 แสดงการทำงานของทวิตเตอร์เอพีไอแบบเรส

[4] อ้างอิงจาก https://code.tutsplus.com/tutorials/building-with-the-twitter-api-using-real-time-streams--cms-22194



รูปที่ 5 การทำงานของทวิตเตอร์เอพีไอแบบเรส [5]

2.2 ตัวชี้วัดของหุ้น (Stock Indicator)

ในงานวิจัยฉบับนี้ ได้มีการประยุกต์ใช้ตัวชี้วัดของหุ้นประเภทโมเมนตัมเพื่อใช้ในการตรวจจับคำเกิดใหม่ โดยมีรายละเอียดตัวชี้วัดของหุ้นที่เกี่ยวข้องกับงานวิจัยดังนี้

1. SMA (Simple Moving Average)
2. EMA (Exponential Moving Average)
3. MACD (Moving Average Convergence Divergence)
4. RSI (Relative Strength Index)
5. STOCH (Stochastic Oscillator)
6. WR (William Percent Range)

โดยตัวชี้วัดของหุ้น SMA และ EMA เป็นการหาค่าเฉลี่ยของราคาหุ้นในกรอบเวลา w วันย้อนหลัง MACD เป็นตัวชี้วัดของหุ้นที่คล้ายกับคุณลักษณะที่ถูกนำเสนอโดยงานวิจัยอื่น ในงานวิจัยนี้จึงประยุกต์ตัวชี้วัดของหุ้นที่นิยมและมีประสิทธิภาพสูง ได้แก่ RSI, STOCH และ WR ในการตรวจจับคำเกิดใหม่

[5] อ้างอิงจาก https://code.tutsplus.com/tutorials/building-with-the-twitter-api-getting-started--cms-22192

2.2.1 SMA (Simple Moving Average)

ตัวชี้วัดของหุ้น SMA หรือ ค่าเฉลี่ยของราคาหุ้นในกรอบเวลา w วันย้อนหลัง เป็นการนำราคาหุ้นย้อนหลัง w วันมาหาค่าเฉลี่ย ซึ่งวิธีการนี้จะทำให้ได้ค่าเฉลี่ยที่ให้ความสำคัญกับราคาหุ้นทุกจุดเวลาเท่ากัน

กำหนดให้

$close_d$ คือ ราคาหุ้นในวันที่ d

w คือ กรอบเวลา

$$SMA_{w,d} = \frac{\sum_{i=0}^{w-1} close_{d-i}}{w} \quad (1)$$

2.2.2 EMA (Exponential Moving Average) หรือ EWMA (Exponential Weight Moving Average)

ตัวชี้วัดของหุ้น EMA หรือ $EWMA$ หรือค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก w วันย้อนหลัง เป็นการหาค่าเฉลี่ยโดยให้ความสำคัญและน้ำหนักกับราคาหุ้นของเวลาปัจจุบันมากกว่าราคาหุ้นในอดีต ซึ่งเป็นการพยายามแก้ไขข้อจำกัดที่เกิดจาก SMA ที่ให้ความสำคัญกับราคาหุ้นทุกจุดเวลาเท่ากัน โดยมีสมการในการหาคำนวณ EMA ดังนี้

กำหนดให้

$close_d$ คือ ราคาหุ้นในวันที่ d

w คือ กรอบเวลา

$$EMA_{w,d} = \frac{1}{w} * close_d + \left(1 - \frac{1}{w}\right) * EMA_{w,d-1} \quad (2)$$

ทำการเปลี่ยนจาก $\frac{1}{w}$ เป็น $\frac{1+1}{w+1}$ หรือ $\frac{2}{w+1}$ ได้สมการในการคำนวณใหม่เป็น

$$EMA_{w,d} = \frac{2}{w+1} * close_d + \left(1 - \frac{2}{w+1}\right) * EMA_{w,d-1} \quad (3)$$

$$EMA_{w,d} = (close_d - EMA_{w,d-1}) * \frac{2}{w+1} + EMA_{w,d-1} \quad (4)$$

2.2.3 MACD (Moving Average Convergence Divergence)

ตัวชี้วัดของหุ้น $MACD$ เป็นหนึ่งในตัวชี้วัดของหุ้นที่นิยมใช้มากที่สุดในปัจจุบัน ถูกคิดค้นโดย Appel [10] เป็นตัวชี้วัดที่คำนวณมาจากส่วนต่างของค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก EMA ของ 2 กรอบเวลา w ที่แตกต่างกัน ซึ่งส่วนต่างของค่าเฉลี่ยของ 2 กรอบเวลา w ที่นิยมมากที่สุดในตลาดหุ้น คือ ส่วนต่างของค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก EMA ของ w 12 วัน และ w 26 วันย้อนหลัง

จุดเด่นของตัวชี้วัดของหุ้น $MACD$ คือ สามารถใช้วิเคราะห์ราคาหุ้นได้ 2 มุมมองพร้อมกัน ได้แก่ ทิศทางแนวโน้มของราคาหุ้น (Trend) และแรงส่งของราคาหุ้น (Momentum) ทำให้แม้ว่าตัวชี้วัดของหุ้น $MACD$ จะถูกพัฒนามานานก็ยังเป็นที่ยอมรับและนิยมใช้จนถึงปัจจุบัน โดยถ้าค่าของตัวชี้วัดของหุ้น $MACD$ มีค่าเป็นบวก หมายความว่า EMA บนกรอบเวลาสั้น มีค่ามากกว่า EMA บนกรอบเวลากว้าง หรือราคาหุ้นในช่วงเวลาปัจจุบันมีอัตราการเพิ่มสูงขึ้นเร็วกว่าราคาหุ้นในอดีต โดยมีตัวอย่างของค่าตัวชี้วัดของหุ้น $MACD$ ของราคาหุ้น KKP ในรูปที่ 6 ถูกแสดงในรูปที่ 7

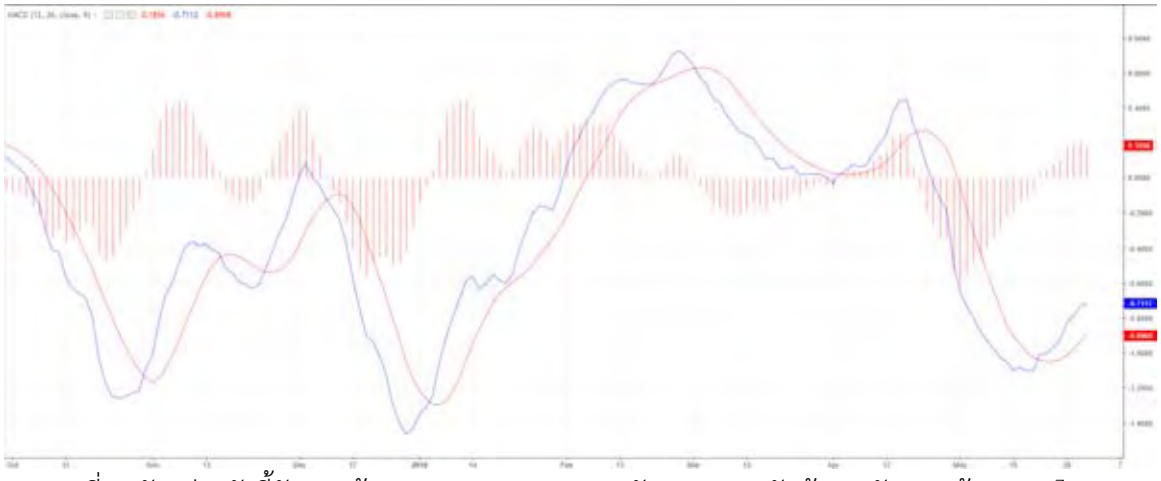
กำหนดให้

$w1$ คือ กรอบเวลาที่ 1

$w2$ คือ กรอบเวลาที่ 2 ซึ่งมีค่ามากกว่า $w1$

$$MACD_{w1,w2,d} = EMA_{w1,d} - EMA_{w2,d} \quad (5)$$

รูปที่ 6 ตัวอย่างราคาของหุ้น KKP [6]



รูปที่ 7 ตัวอย่างตัวชี้วัดของหุ้น MACD บน EMA 12 วัน และ 26 วันย้อนหลัง ของหุ้น KKP [7]

[6] อ้างอิงจาก https://www.investing.com/

[7] อ้างอิงจาก https://www.investing.com/

2.2.4 RSI (Relative Strength Index)

ตัวชี้วัดของหุ้น rsi เป็นตัวชี้วัดของหุ้นที่ใช้ดูโมเมนตัมของราคาหุ้น ถูกคิดค้นโดย W. Jr. Wilder [11] เพื่อหาว่าหุ้นตัวดังกล่าวอยู่ในสถานะใด ระหว่าง สถานะซื้อมากเกินไป (Overbought) หรือสถานะขายมากเกินไป (Oversold) หรือสถานะปกติ ซึ่งค่าของตัวชี้วัดของหุ้น rsi มีค่าอยู่ระหว่าง 0 ถึง 100 โดยถ้ามีค่ามากกว่า 70 จะถือว่าหุ้นตัวดังกล่าวอยู่ในสถานะซื้อมากเกินไป มีโอกาสปรับราคาลง แต่ถ้ามีค่าน้อยกว่า 30 จะถือว่าหุ้นตัวดังกล่าวอยู่ในสถานะขายมากเกินไป มีโอกาสปรับราคาขึ้น โดยมีตัวอย่างของค่าตัวชี้วัดของหุ้น rsi ของราคาหุ้น KKP ในรูปที่ 6 ถูกแสดงในรูปที่ 8

ตัวชี้วัดของหุ้น rsi หาได้จากขนาดของราคาที่เพิ่มขึ้นเทียบกับขนาดของราคาที่ลดลงในกรอบเวลา w ย้อนหลัง โดยถ้าราคาหุ้นของวันปัจจุบันเทียบกับเมื่อวานแล้วมีราคาสูงขึ้น จะถือว่า $gain$ มีค่าเท่ากับราคาที่เพิ่มขึ้นและ $loss$ มีค่าเท่ากับ 0 แต่ถ้าราคาหุ้นของวันปัจจุบันเทียบกับเมื่อวานแล้วมีราคาต่ำลง จะถือว่า $gain$ มีค่าเท่ากับ 0 และ $loss$ มีค่าเท่ากับราคาที่ลดลง

กำหนดให้

$close_d$ คือ ราคาหุ้นในวันที่ d

w คือ กรอบเวลา

$$gain_d = MAX(close_d - close_{d-1}, 0) \tag{6}$$

$$loss_d = MAX(close_{d-1} - close_d, 0) \tag{7}$$

$$avg\_gain_{w,d} = \frac{\sum_{i=0}^{w-1} gain_{d-i}}{w} \tag{8}$$

$$avg\_loss_{w,d} = \frac{\sum_{i=0}^{w-1} loss_{d-i}}{w} \tag{9}$$

$$rs_{w,d} = \frac{avg\_gain_{w,d}}{avg\_loss_{w,d}} \tag{10}$$

$$rsi_{w,d} = 100 - \frac{100}{1 + rs_{w,d}} \tag{11}$$

ทำการเปลี่ยนการคำนวณค่าเฉลี่ยของ $gain$ กับ $loss$ ในสมการที่ (8) และ (9) เป็นค่าเฉลี่ยถ่วงน้ำหนักด้วยสมการที่ (4) ดังนั้นจะได้สมการในการคำนวณค่าเฉลี่ยของ $gain$ กับ $loss$ ใหม่ดังนี้

$$avg\_gain_{w,d} = (close_d - avg\_gain_{w,d-1}) * \frac{2}{w+1} + avg\_gain_{w,d-1} \quad (12)$$

$$avg\_loss_{w,d} = (close_d - avg\_loss_{w,d-1}) * \frac{2}{w+1} + avg\_loss_{w,d-1} \quad (13)$$

รูปที่ 8 ตัวอย่างตัวชี้วัดของหุ้น RSI ของหุ้น KKP [8]

2.2.5 STOCH (Stochastic Oscillator)

ตัวชี้วัดของหุ้น $stoch$ เป็นตัวชี้วัดของหุ้นที่ถูกคิดค้นและพัฒนาโดย Lane [12] เป็นตัวชี้วัดของหุ้นที่ใช้ดูโมเมนตัมของราคาหุ้นคล้ายกับ rsi เพื่อหาว่าหุ้นตัวดังกล่าวอยู่ในสถานะใด ระหว่างสถานะซื้อมากเกินไป (Overbought) หรือสถานะขายมากเกินไป (Oversold) หรือสถานะปกติ ซึ่งค่าของตัวชี้วัดของหุ้น $stoch$ มีค่าอยู่ระหว่าง 0 ถึง 100 โดยถ้ามีค่ามากกว่า 80 จะถือว่าหุ้นตัวดังกล่าวอยู่ในสถานะซื้อมากเกินไป มีโอกาสปรับราคาลง แต่ถ้ามีค่าน้อยกว่า 20 จะถือว่าหุ้นตัวดังกล่าวอยู่ในสถานะขายมากเกินไป มีโอกาสปรับราคาขึ้น นอกจากนี้ $stoch$ ยังสามารถใช้งานได้อีกวิธีหนึ่งคือการดูการตัดกันของ $stoch(\%K)$ และ $stoch(\%D)$ โดยมีตัวอย่างของค่าตัวชี้วัดของหุ้น $stoch$ ของราคาหุ้น KKP ในรูปที่ 6 ถูกแสดงในรูปที่ 9

[8] อ้างอิงจาก https://www.investing.com/

ตัวชี้วัดของหุ้น $stoch$ หาได้จากราคาปัจจุบัน, ราคาต่ำสุดในกรอบเวลา w ย้อนหลัง และราคาสูงสุดในกรอบเวลา w ย้อนหลัง

กำหนดให้

$close_d$ คือ ราคาหุ้นในวันที่ d

w คือ กรอบเวลา

$close\_lowest_{w,d}$ คือ ราคาหุ้นต่ำสุดในกรอบเวลา w วันย้อนหลัง นับจากวันที่ d

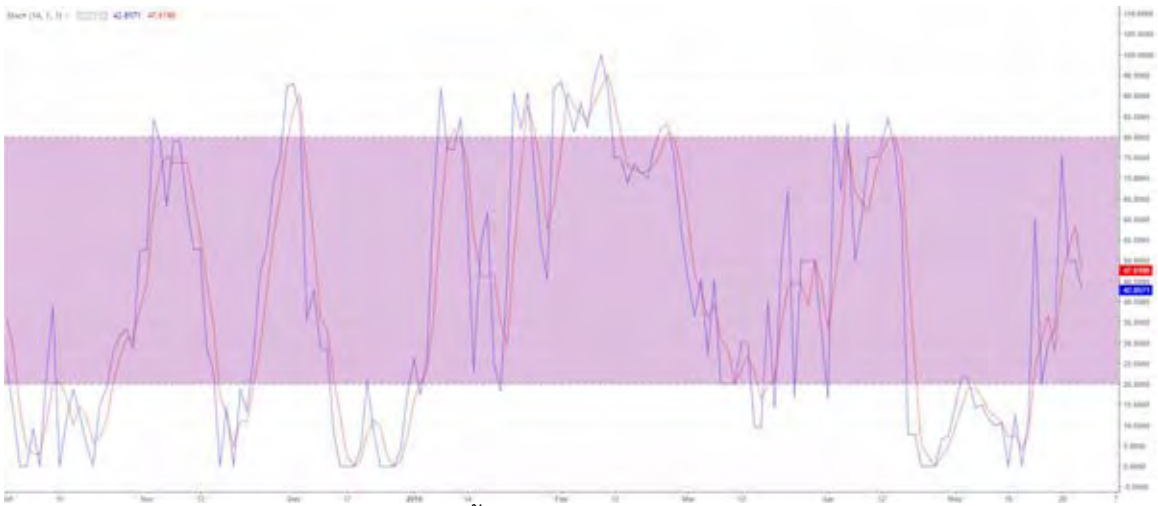
$close\_highest_{w,d}$ คือราคาหุ้นสูงสุดในกรอบเวลา w วันย้อนหลังนับจากวันที่ d

$stoch(\%K)$ คือ $stoch$ ที่คำนวณจากกรอบเวลาย้อนหลัง K หรือ $stoch_{w,d}$ ในสมการด้านล่าง เมื่อกำหนดให้ $K = w$

$stoch(\%D)$ คือ ค่าเฉลี่ยของ $stoch(\%K)$ ในของกรอบเวลา D ย้อนหลัง หรือ $stoch\_slow_{w2,w,d}$ ในสมการด้านล่าง เมื่อกำหนดให้ $K = w$ และ $D = w2$

$$stoch_{w,d} = 100 * \frac{close_d - close\_lowest_{w,d}}{close\_highest_{w,d} - close\_lowest_{w,d}} \quad (14)$$

$$stoch\_slow_{w2,w,d} = \frac{\sum_{i=0}^{w2-1} stoch_{w,d-i}}{w2} \quad (15)$$



รูปที่ 9 ตัวอย่างตัวชี้วัดของหุ้น STOCH ของหุ้น KKP [9]

[9] อ้างอิงจาก https://www.investing.com/

2.2.6 WR (Williams Percent Range)

ตัวชี้วัดของหุ้น wr เป็นตัวชี้วัดของหุ้นที่ถูกคิดค้นและพัฒนาโดย Larry Williams เป็นตัวชี้วัดของหุ้นที่ใช้ดูโมเมนตัมของราคาหุ้น คล้ายกับ rsi และ $stoch$ เพื่อหาว่าหุ้นตัวดังกล่าวอยู่ในสถานะใด ระหว่าง สถานะซื้อมากเกินไป (Overbought) หรือสถานะขายมากเกินไป (Oversold) หรือสถานะปกติ ซึ่งค่าของตัวชี้วัดของหุ้น wr มีค่าอยู่ระหว่าง 0 ถึง 100 โดยถ้ามีค่ามากกว่า 80 จะถือว่าหุ้นตัวดังกล่าวอยู่ในสถานะซื้อมากเกินไป มีโอกาสปรับราคาลง แต่ถ้ามีค่าน้อยกว่า 20 จะถือว่าหุ้นตัวดังกล่าวอยู่ในสถานะขายมากเกินไป มีโอกาสปรับราคาขึ้น นอกจากนี้ค่าของตัวชี้วัดของหุ้น wr มักมีเครื่องหมายลบดังแสดงในรูปที่ 10 ในการวิเคราะห์จึงนิยมมองข้ามเครื่องหมายลบโดยการเพิ่มค่า 100 ทำให้มีค่าตั้งแต่ 0 ถึง 100 ตัวชี้วัดของหุ้น wr มักถูกใช้ในการคาดการณ์การกลับตัวของราคา เนื่องจากมักทำจุดสูงสุดและกลับตัวเป็นทิศทางขาลงไม่กี่วันก่อนที่ราคาจะขึ้นถึงจุดสูงสุดและกลับตัวลง

กำหนดให้

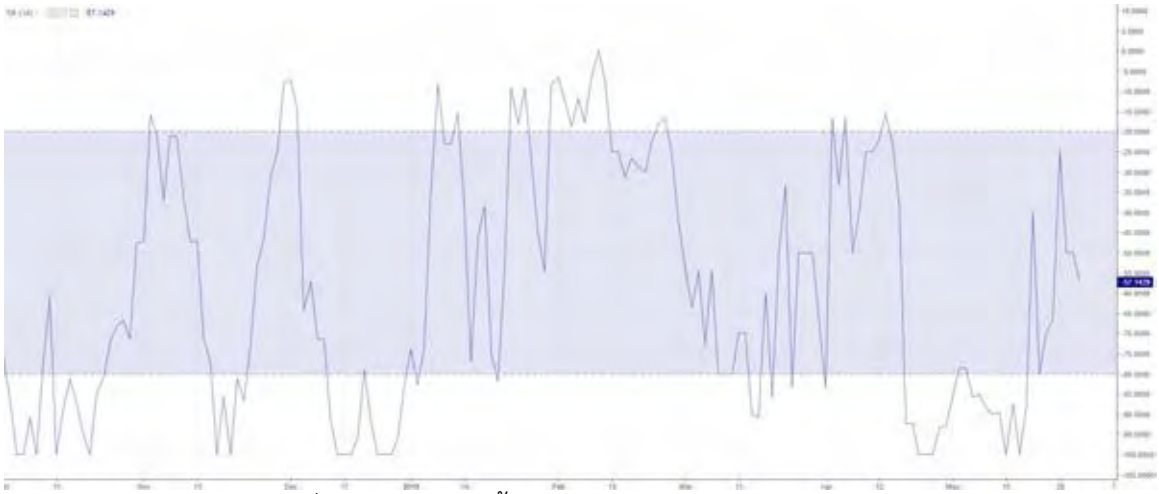
$close_d$ คือ ราคาหุ้นในวันที่ d

w คือ กรอบเวลา

$close\_lowest_{w,d}$ คือ ราคาหุ้นต่ำสุดในกรอบเวลา w วันย้อนหลัง นับจากวันที่ d

$close\_highest_{w,d}$ คือราคาหุ้นสูงสุดในกรอบเวลา w วันย้อนหลังนับจากวันที่ d

$$wr_{w,d} = 100 - 100 * \frac{close\_highest_{w,d} - close_d}{close\_highest_{w,d} - close\_lowest_{w,d}} \quad (16)$$



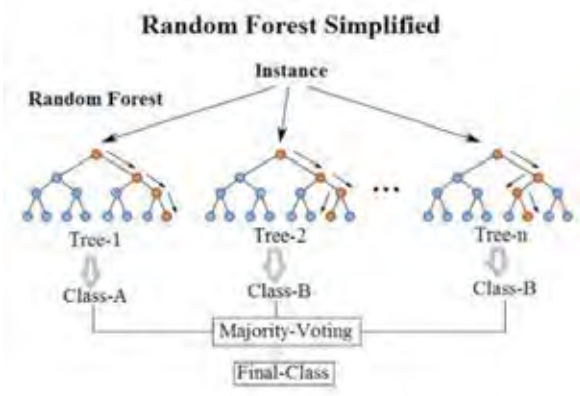
รูปที่ 10 ตัวอย่างตัวชี้วัดของหุ้น WR ของหุ้น KKP [10]

[10] อ้างอิงจาก https://www.investing.com/

2.3 ป่าไม้แบบสุ่ม (Random Forest)

ตัวจำแนกประเภทต้นไม้ตัดสินใจ (Decision tree) เป็นตัวจำแนกประเภทที่สร้างกฎขึ้นมาในการตัดสินใจ โดยในแต่ละชั้นของต้นไม้ตัดสินใจ จะทำการคำนวณหาคุณลักษณะ (Feature) และเส้นแบ่ง (Threshold) ที่ดีที่สุดที่สามารถแบ่งข้อมูลออกเป็น 2 ส่วนแล้วมีการกระจายตัวของคลาสที่สนใจน้อยที่สุด

ป่าไม้แบบสุ่ม (Random forest) เป็นตัวจำแนกประเภทที่ใช้ต้นไม้ตัดสินใจ (Decision Tree) หลายต้นที่แตกต่างกันมาช่วยในการหาคำตอบ เนื่องจากต้นไม้ตัดสินใจแม้จะเป็นตัวจำแนกประเภทที่สามารถอธิบายเหตุผลในการตัดสินใจได้ดีแต่มีความแม่นยำในการทำนายต่ำ ดังนั้นเพื่อเพิ่มความแม่นยำในการทำนาย ต้นไม้ตัดสินใจหลายต้นจึงถูกสร้างขึ้นโดยแต่ละต้นจะได้รับข้อมูลและคุณลักษณะที่ใช้ในการสอนแตกต่างกันเล็กน้อย จากนั้นนำผลลัพธ์การทำนายของทุกต้นไม้ตัดสินใจทุกต้นมารวมกันด้วยการโหวต ซึ่งวิธีการนี้สามารถเพิ่มประสิทธิภาพในการทำนาย และลดโอกาสเกิดปัญหาที่เรียกว่าการจดจำรูปแบบของข้อมูลมากเกินไป (Overfitting) โดยรูปที่ 11 แสดงตัวอย่างโครงสร้างของป่าไม้แบบสุ่ม



รูปที่ 11 ตัวอย่างโครงสร้างของป่าไม้แบบสุ่ม [11]

[11] อ้างอิงจาก https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d

2.4 การตรวจสอบแบบไขว้ (Cross Validation)

การตรวจสอบประสิทธิภาพเพื่อเลือกพารามิเตอร์ที่ดีที่สุดของตัวจำแนกประเภท ทำได้โดยการแบ่งข้อมูลออกเป็น 2 ส่วน ได้แก่ ข้อมูลสอน (Training dataset) กับ ข้อมูลตรวจสอบ (Validation dataset) จากนั้นใช้ข้อมูลสอนในการสร้างตัวจำแนกประเภท และใช้ข้อมูลตรวจสอบในการตรวจสอบประสิทธิภาพของตัวจำแนกประเภท แต่บางครั้งการแบ่งข้อมูลสอนและข้อมูลตรวจสอบเกิดความลำเอียงขึ้น ทำให้ตัวจำแนกประเภทที่สร้างขึ้นมีประสิทธิภาพที่ดีบนชุดข้อมูลตรวจสอบ แต่ได้ประสิทธิภาพไม่ดีเมื่อนำไปใช้งานจริง

ดังนั้นเพื่อไม่ให้เกิดความความลำเอียงในการแบ่งข้อมูล การตรวจสอบประสิทธิภาพของตัวจำแนกประเภทจึงมักทำด้วยการตรวจสอบแบบไขว้ โดยการแบ่งข้อมูลออกเป็น k ส่วน จากนั้นนำข้อมูล k-1 ส่วนเป็นข้อมูลสอน และอีก 1 ส่วนเป็นข้อมูลตรวจสอบ และวนทำทั้งหมด k รอบ ดังนั้นข้อมูลทุกส่วนจะถูกใช้ในการสอนทั้งหมด k-1 ครั้ง และถูกใช้ในการตรวจสอบ 1 ครั้งเสมอ ดังแสดงในตารางที่ 1 โดยกำหนดค่า k = 3 จะพบว่าข้อมูลทุกส่วนจะถูกใช้เป็นข้อมูลสอนทั้งหมด 2 ครั้ง และถูกใช้เป็นข้อมูลตรวจสอบทั้งหมด 1 ครั้ง

ตารางที่ 1 การแบ่งข้อมูลออกเป็นข้อมูลสอนและข้อมูลตรวจสอบของการตรวจสอบแบบไขว้ 3 รอบ

รอบของการทดสอบ	ข้อมูลส่วนที่1	ข้อมูลส่วนที่2	ข้อมูลส่วนที่3
รอบที่ 1	ข้อมูลสอน	ข้อมูลสอน	**ข้อมูลตรวจสอบ**
รอบที่ 2	ข้อมูลสอน	**ข้อมูลตรวจสอบ**	ข้อมูลสอน
รอบที่ 3	**ข้อมูลตรวจสอบ**	ข้อมูลสอน	ข้อมูลสอน

การตรวจสอบแบบไขว้เชิงเวลา (Temporal Cross Validation)

การแบ่งข้อมูลสอน (Training dataset) และข้อมูลตรวจสอบ (Validation dataset) ของการตรวจสอบแบบไขว้กับข้อมูลที่มีความสัมพันธ์กับเวลา จะแตกต่างกับการตรวจสอบแบบไขว้ปกติ เนื่องจากโดยปกติการตรวจสอบแบบไขว้จะแบ่งข้อมูลออกเป็น k ส่วนด้วยการสุ่ม ทำให้เมื่อนำมาใช้กับข้อมูลที่มีความสัมพันธ์กับเวลา จะทำให้ข้อมูลของทุกเวลา ปรากฏทั้งในข้อมูลสอนและข้อมูลตรวจสอบ

ดังนั้นการแบ่งข้อมูลสอนและข้อมูลตรวจสอบของข้อมูลที่มีความสัมพันธ์กับเวลา มักจะใช้เวลาในการแบ่งข้อมูลแทนการสุ่ม โดยในรอบแรกจะใช้ข้อมูลของเวลาในช่วงแรกเป็นข้อมูลสอน และข้อมูลของเวลาถัดมาเป็นข้อมูลตรวจสอบ จากนั้นในรอบถัดมาจะนำข้อมูลสอนและข้อมูลตรวจสอบ

ของรอบก่อนหน้ามาเป็นข้อมูลสอนของรอบปัจจุบัน และใช้ข้อมูลของเวลาถัดมาเป็นข้อมูลตรวจสอบ ดังนั้นเวลาของข้อมูลตรวจสอบจะมากกว่าเวลาของข้อมูลสอนเสมอ ดังแสดงในตารางที่ 2

ตารางที่ 2 การแบ่งข้อมูลออกเป็นข้อมูลสอนและข้อมูลตรวจสอบของการตรวจสอบแบบไขว้เชิงเวลา 3 รอบ

รอบของการทดสอบ	**ข้อมูลส่วนที่1**	**ข้อมูลส่วนที่2**	**ข้อมูลส่วนที่3**	**ข้อมูลส่วนที่4**	**ข้อมูลส่วนที่5**
รอบที่ 1	ข้อมูลสอน	ข้อมูลสอน	**ข้อมูลตรวจสอบ**		
รอบที่ 2	ข้อมูลสอน	ข้อมูลสอน	ข้อมูลสอน	**ข้อมูลตรวจสอบ**	
รอบที่ 3	ข้อมูลสอน	ข้อมูลสอน	ข้อมูลสอน	ข้อมูลสอน	**ข้อมูลตรวจสอบ**

2.5 การกระจายเมทริกซ์ด้วยวิธีการแยกค่าแบบเดี่ยว (Singular Vector Decomposition หรือ SVD)

SVD คืออัลกอริทึมที่ใช้ในการแยกส่วนประกอบของเมทริกซ์ เพื่อทำการลดขนาดของเมทริกซ์ให้มีขนาดที่เล็กลง แต่ยังรักษาความสัมพันธ์ที่คล้ายเมทริกซ์เดิมไว้ได้ ดังแสดงในรูปที่ 12 ที่ต้องการลดรูปของเมทริกซ์ A ที่มีขนาด $n \times d$ ให้เหลือขนาด $n \times r$ โดยทำการแยกส่วนประกอบของเมทริกซ์ A เป็นเมทริกซ์ $U\Sigma V^T$ แล้วทำการลดรูปเมทริกซ์ทั้ง 3 เมทริกซ์ แต่ยังคงสามารถคูณเมทริกซ์กลับแล้วมีค่าใกล้เคียงกับเมทริกซ์เดิม

โดยการแตกองค์ประกอบเมทริกซ์ A ขนาด $n \times d$ เป็น 3 เมทริกซ์คูณกัน $U\Sigma V^T$

1. U คือ เมทริกซ์เชิงตั้งฉาก (Orthogonal matrix) ที่มีขนาด $n \times n$
2. Σ คือ เมทริกซ์ทแยงมุม (Diagonal matrix) ที่มีขนาด $n \times d$ หรือก็คือเมทริกซ์ที่มีค่าไม่ใช่ 0 เฉพาะในเส้นทแยงมุม โดยตัวเลขที่อยู่ในเส้นทแยงมุมเรียกว่าค่าแบบเดี่ยว (Singular value)
3. V^T คือ เมทริกซ์เชิงตั้งฉาก (Orthogonal matrix) ที่มีขนาด $d \times d$

โดยสามารถคำนวณหาค่าของเมทริกซ์ U, Σ และ V^T ได้ดังนี้

$$A = U\Sigma V^T \tag{17}$$

$$A^T A = (V\Sigma U^T)U\Sigma V^T \tag{18}$$

$$A^T A = U\Sigma V^T(V\Sigma U^T) \tag{19}$$

$U^T U = V^T V = I$ หรือเมทริกซ์เอกลักษณ์ (Identity matrix) ดังนั้นจะได้สมการดังนี้

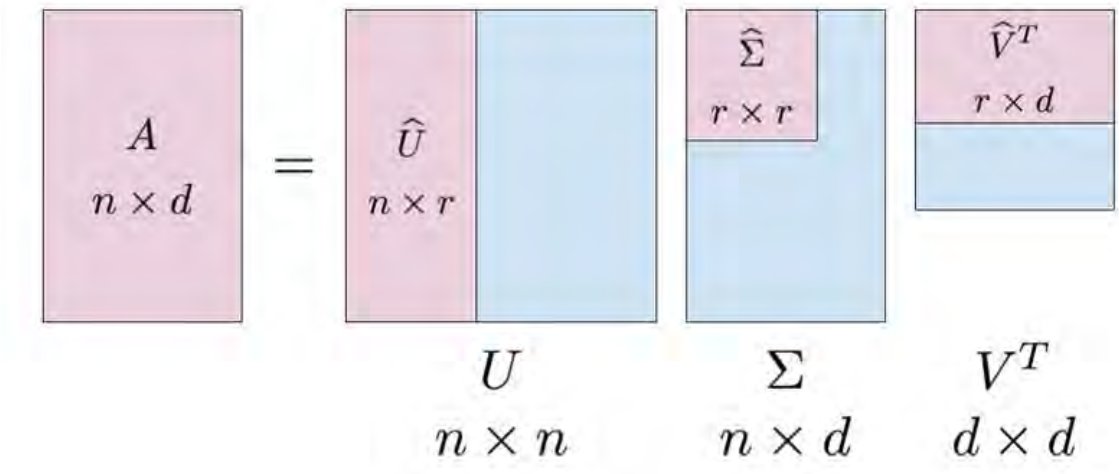
$$A^T A = V\Sigma\Sigma V^T = V\Sigma^2 V^T \tag{20}$$

$$AA^T = U\Sigma\Sigma U^T = U\Sigma^2 U^T \tag{21}$$

จาก 2 สมการดังกล่าวสามารถนำไปคำนวณหาค่าเมทริกซ์ U และ V ได้โดยใช้การแยกค่าไอเก้น (Eigen decomposition) ของ $A^T A$ และ AA^T

เมื่อสามารถแตกองค์ประกอบของเมทริกซ์ A ออกเป็น $U\Sigma V^T$ แล้วจะทำการเรียงค่าแบบเดี่ยว (Singular Value) ในเมทริกซ์ Σ จากมากไปน้อยตามเส้นทแยงมุมซ้ายบนไปขวาล่าง โดยการสลับแถวและหลักของเมทริกซ์ U กับ V^T เพื่อให้ 3 เมทริกซ์ $U\Sigma V^T$ คูณกันแล้วยังมีค่าเท่ากับเมทริกซ์ A เหมือนเดิม

เมื่อค่าแบบเดี่ยว (Singular value) ถูกเรียงจากมากไปน้อยแล้ว จะทำการย่อขนาดเมทริกซ์ของ U, Σ และ V^T โดยเมทริกซ์ A เดิมมีขนาด $n \times n$ จะถูกเลือกมาเพียง r หลักแรก จะได้เมทริกซ์ที่มีขนาดเหลือเพียง $n \times r$ ส่วนเมทริกซ์ Σ เดิมมีขนาด $n \times d$ จะถูกเลือกมาเพียง r แถวและ r หลักแรก จะได้เมทริกซ์ที่มีขนาดเหลือเพียง $r \times r$ และเมทริกซ์ V^T เดิมมีขนาด $d \times d$ จะถูกเลือกมาเพียง r แถวแรก จะได้เมทริกซ์ที่มีขนาดเหลือเพียง $r \times d$ ดังแสดงในรูปที่ 12



รูปที่ 12 การแยกส่วนประกอบและลดรูปของเมทริกซ์ด้วย SVD [12]

ตัวอย่างการย่อขนาดของเมทริกซ์ด้วย SVD

กำหนดให้ $A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 2 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 1 & 0 & 2 & 2 \end{bmatrix}$ โดยมีขนาด 7 x 5

นำเมทริกซ์ A ไปแยกส่วนองค์ประกอบเมทริกซ์จะได้ $U\Sigma V^T$ มีค่าดังแสดงในรูปที่ 13

$$\underset{U}{\begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix}} \underset{\Sigma}{\begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix}} \underset{V^T}{\begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.4 & -0.8 & 0.4 & 0.09 & 0.09 \end{bmatrix}}$$

รูปที่ 13 ค่าของเมทริกซ์หลังแยกส่วนองค์ประกอบเมทริกซ์

[12] อ้างอิงจาก https://intoli.com/blog/pca-and-svd/

เมื่อค่าแบบเดี่ยว (Singular value) ในเมทริกซ์ Σ ถูกเรียงลำดับจากมากไปน้อยแล้ว จะทำการเลือกมาเพียง r ลำดับแรก ได้ดังรูปที่ 14 และ รูปที่ 15 นอกจากนี้จะพบว่าเมื่อทำการคูณเมทริกซ์ $U\Sigma V^T$ หลังจากย่อขนาด จะได้เมทริกซ์ที่มีความคล้ายกับเมทริกซ์ A เดิมดังแสดงในรูปที่ 16

$$\begin{bmatrix} 0.13 & 0.02 & -0.01 \\ 0.41 & 0.07 & -0.03 \\ 0.55 & 0.09 & -0.04 \\ 0.68 & 0.11 & -0.05 \\ 0.15 & -0.59 & 0.65 \\ 0.07 & -0.73 & -0.67 \\ 0.07 & -0.29 & 0.32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \\ 0.4 & -0.8 & 0.4 & 0.09 & 0.09 \end{bmatrix}$$

U Σ V^T

รูปที่ 14 ค่าของเมทริกซ์เมื่อกำลังย่อขนาด

$$\begin{bmatrix} 0.13 & 0.02 \\ 0.41 & 0.07 \\ 0.55 & 0.09 \\ 0.68 & 0.11 \\ 0.15 & -0.59 \\ 0.07 & -0.73 \\ 0.07 & -0.29 \end{bmatrix} \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix} \begin{bmatrix} 0.56 & 0.59 & 0.56 & 0.09 & 0.09 \\ 0.12 & -0.02 & 0.12 & -0.69 & -0.69 \end{bmatrix}$$

U Σ V^T

รูปที่ 15 ค่าของเมทริกซ์หลังย่อขนาด

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{3} & \mathbf{3} & \mathbf{3} & 0 & 0 \\ \mathbf{4} & \mathbf{4} & \mathbf{4} & 0 & 0 \\ \mathbf{5} & \mathbf{5} & \mathbf{5} & 0 & 0 \\ 0 & \mathbf{2} & 0 & \mathbf{4} & \mathbf{4} \\ 0 & 0 & 0 & \mathbf{5} & \mathbf{5} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{2} \end{bmatrix} \approx \begin{bmatrix} \mathbf{0.92} & \mathbf{0.95} & \mathbf{0.92} & 0.01 & 0.01 \\ \mathbf{2.91} & \mathbf{3.01} & \mathbf{2.91} & -0.01 & -0.01 \\ \mathbf{3.9} & \mathbf{4.04} & \mathbf{3.9} & 0.01 & 0.01 \\ \mathbf{4.82} & \mathbf{5} & \mathbf{4.82} & 0.03 & 0.03 \\ 0.7 & \mathbf{0.53} & 0.7 & \mathbf{4.11} & \mathbf{4.11} \\ -0.69 & 1.34 & -0.69 & \mathbf{4.78} & \mathbf{4.78} \\ 0.32 & \mathbf{0.23} & 0.32 & \mathbf{2.01} & \mathbf{2.01} \end{bmatrix}$$

รูปที่ 16 ค่าของเมทริกซ์เดิมกับเมทริกซ์ใหม่ที่เกิดจากการนำเมทริกซ์หลังย่อขนาดคูณกัน

2.6 การทำดัชนีความหมายแฝง (Latent Semantic Indexing หรือ LSI)

ในงานวิจัยนี้ อัลกอริทึม LSI ถูกใช้ในการหาเวกเตอร์ของคำ และใช้เวกเตอร์ดังกล่าวในการหาความคล้ายของแต่ละคำด้วยสัมประสิทธิ์ความคล้ายโคไซน์ เพื่อจับกลุ่มคำที่เกี่ยวข้องเป็นหัวข้อเดียวกัน โดยอัลกอริทึม LSI ถูกนำเสนอในปี ค.ศ. 1990 Deerwester, Dumais [13] ซึ่งเป็นอัลกอริทึมทางภาษาศาสตร์ที่พิจารณาเกี่ยวกับการปรากฏร่วมของคำต่าง ๆ โดยพิจารณา

ความสัมพันธ์ระหว่างข้อความและคำที่ปรากฏในข้อความ โดยมีสมมติฐานว่าคำที่มีความหมายใกล้เคียงกันมักปรากฏในข้อความที่คล้ายกัน โดยทำการเปลี่ยนเมทริกซ์ความสัมพันธ์ระหว่างคำกับข้อความหรือเมทริกซ์ A เป็นเมทริกซ์ความสัมพันธ์ระหว่างคำกับหัวข้อหรือเมทริกซ์ U ด้วยอัลกอริทึม SVD

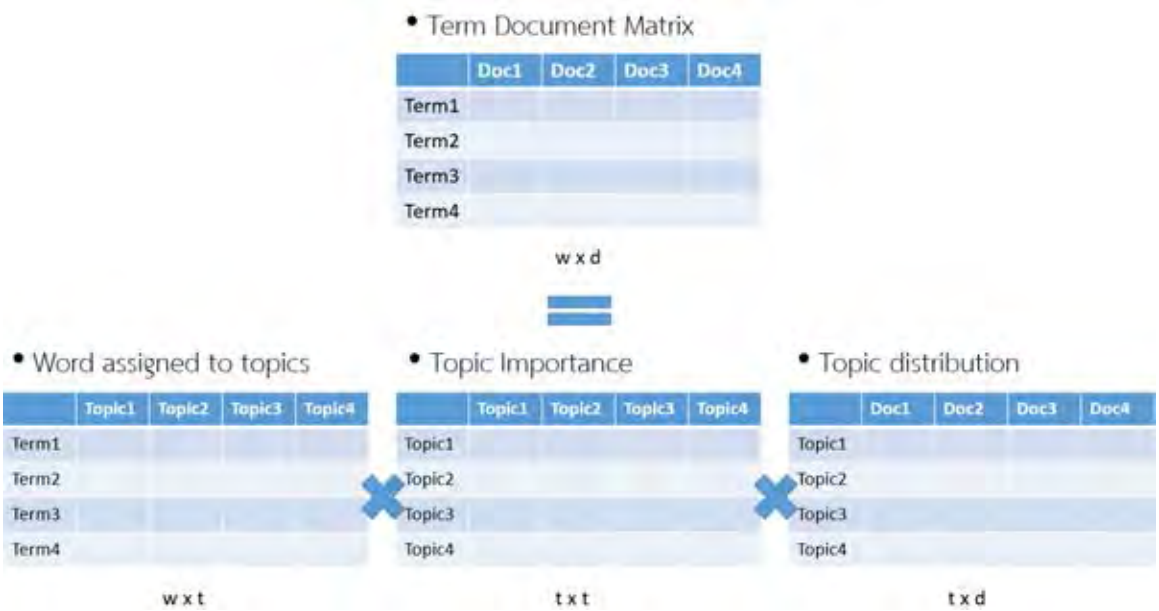
กำหนดให้

w คือ จำนวนคำที่แตกต่างกันในข้อความ d

t คือ จำนวนหัวข้อที่ต้องการแบ่ง โดย t มีค่าน้อยกว่า d มาก

ทำการสร้างเมทริกซ์ของความสัมพันธ์ระหว่างคำในกับข้อความซึ่งมีขนาดเท่ากับ w x d จากนั้นทำการแตกส่วนประกอบของเมทริกซ์ A ด้วย SVD ได้เป็นเมทริกซ์ $U\Sigma V^T$ และทำการย่อขนาดของเมทริกซ์ $U\Sigma V^T$ จะได้เมทริกซ์ U ซึ่งเป็นเมทริกซ์ที่แสดงความสัมพันธ์ระหว่างคำกับหัวข้อ โดยมีขนาดเปลี่ยนไปจาก w x w เป็น w x t ดังแสดงในรูปที่ 17

แถว 1 แถวของเมทริกซ์ U เป็นเวกเตอร์ของคำ 1 คำ ที่แสดงถึงความสัมพันธ์ของคำนั้นกับหัวข้อต่าง ๆ จำนวน t หัวข้อ ดังนั้นสามารถหาความคล้ายของแต่ละคู่ของคำได้จากการนำเวกเตอร์ของคำมาเปรียบเทียบกันด้วยสัมประสิทธิ์ความคล้ายโคไซน์ (Cosine similarity)



รูปที่ 17 การแยกส่วนประกอบและย่อขนาดของเมทริกซ์ด้วย SVD ในอัลกอริทึม LSI

2.7 การวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภททั่วไป (Classification Performance Evaluation)

การวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภท ได้มีการนิยาม $True\ Positive\ (TP)$, $False\ Positive\ (FP)$ และ $False$ Negatives (FN) ซึ่งเป็นค่าที่เกิดจากการเปรียบเทียบผลลัพธ์จากการทำนายกับผลลัพธ์ที่เป็นผลเฉลยดังแสดงในตารางที่ 1 โดยมีนิยามดังต่อไปนี้

TP คือ จำนวนแถวของข้อมูลที่ทำนายว่าใช่ และถูกต้อง

FP คือ จำนวนแถวของข้อมูลที่ทำนายว่าใช่ และ**ไม่**ถูกต้อง

FN คือ จำนวนแถวของข้อมูลที่ทำนายว่า**ไม่**ใช่ และ**ไม่**ถูกต้อง

TN คือ จำนวนแถวของข้อมูลที่ทำนายว่า**ไม่**ใช่ และถูกต้อง

		Predict	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

ตารางที่ 3 ผลลัพธ์จากการทำนายเปรียบเทียบกับผลลัพธ์ที่เป็นผลเฉลย (Confusion Matrix)

จากนั้นนำค่า TP, FP และ FN ไปคำนวณเป็น ค่าความแม่นยำในการทำนายหรือ $Precision\ (Pr)$, ความครอบคลุมในการทำนายหรือ $Recall\ (Rc)$ และค่าเฉลี่ยฮาร์มอนิกของ $Precision$ กับ $Recall$ หรือ F_1 โดยมีสมการดังต่อไปนี้

$$Pr = \frac{TP}{TP\ +\ FP} \tag{22}$$

$$Rc = \frac{TP}{TP + FN} \tag{23}$$

$$F_1 = \frac{2 * Pr * Rc}{Pr + Rc} \tag{24}$$

2.8 การวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภทที่มีหลายคลาส (Multiclass Classification Performance Evaluation)

การวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภทที่มีหลายคลาส มี 2 วิธี ได้แก่ วิธีแบบไมโคร (Micro) และวิธีแบบแมโคร (Macro) ซึ่งการวัดประสิทธิภาพของงานวิจัยนี้ ประยุกต์มาจากการวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภทที่มีหลายคลาสด้วยวิธีแบบแมโคร (Macro) โดยกำหนดให้ 1 หัวข้อในชุดผลเฉลยเท่ากับ 1 คลาสของข้อมูล

วิธีแบบไมโคร (Micro) เป็นวิธีการวัดประสิทธิภาพที่ให้ความสำคัญกับข้อมูลทุกแถวเท่ากัน ดังนั้น จะคำนวณหาค่า Pr และ Rc ของทุกคลาสจากผลรวมของ TP และ FP ดังสมการต่อไปนี้

$$Pr = \frac{\sum TP}{\sum TP \ + \ \sum FP} \tag{25}$$

$$Rc = \frac{\sum TP}{\sum TP \ + \ \sum FN} \tag{26}$$

วิธีแบบแมโคร (Macro) เป็นวิธีการวัดประสิทธิภาพที่ให้ความสำคัญกับคลาสทุกคลาสเท่ากัน หรือการให้ความสำคัญของคลาสที่มีจำนวนข้อมูลมากกับคลาสที่มีจำนวนข้อมูลน้อยเท่ากัน ทำได้โดยการหา Pr และ Rc ของแต่ละคลาส จากนั้นหาค่าเฉลี่ยของ Pr และ Rc ของทุกคลาส ดังสมการต่อไปนี้

$$Pr = \frac{\sum Pr}{n} \tag{27}$$

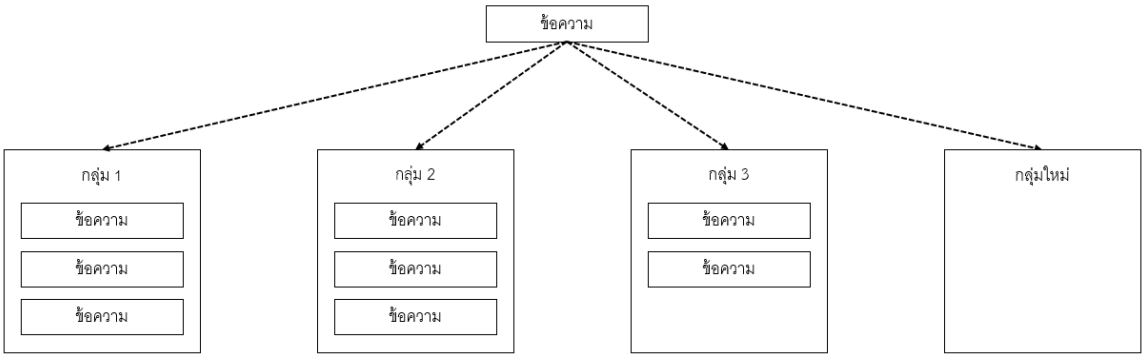
$$Rc \ = \frac{\sum Rc}{n} \tag{28}$$

บทที่ 3 งานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงงานวิจัยที่เกี่ยวข้องกับการตรวจจับหัวข้อเกิดใหม่ซึ่งสามารถแบ่งจากวิธีที่ใช้ในการตรวจจับหัวข้อเกิดใหม่ออกเป็น 2 ประเภท ได้แก่ 1) การตรวจจับหัวข้อเกิดใหม่จากการจับกลุ่มของข้อความ เป็นวิธีที่นำข้อความที่เข้ามาใหม่ไปเปรียบเทียบกับข้อความหรือกลุ่มของข้อความในฐานข้อมูล เพื่อหาข้อความที่ไม่คล้ายกับข้อความเดิมหรือข้อความที่กำลังพูดถึงหัวข้อใหม่ 2)การตรวจจับหัวข้อเกิดใหม่จากการสกัดคำเกิดใหม่ เป็นวิธีที่หาคำที่กำลังถูกพูดถึงมากหรือคำเกิดใหม่ โดยดูจากการปรากฏของคำนั้นเปรียบเทียบกับพฤติกรรมปกติของคำนั้นในฐานข้อมูล จากนั้นหาความสัมพันธ์ของกลุ่มของคำเกิดใหม่และจับกลุ่มคำที่เกี่ยวข้องเป็นหัวข้อเกิดใหม่ นอกจากนี้ในบทนี้ยังกล่าวถึงการประเมินผลการทดลองของวิจัยในปัจจุบันในส่วนที่ 3 ของบทนี้

3.1 การตรวจจับหัวข้อเกิดใหม่จากการจับกลุ่มของข้อความ

งานวิจัยในช่วงแรกถูกนำเสนอโดย Yang, Pierce [14] ในปี ค.ศ. 1998 ซึ่งพยายามตรวจจับหัวข้อเกิดใหม่โดยการจับกลุ่มกระแสของข้อความที่เกี่ยวข้อง เข้าเป็นกลุ่มเดียวกัน ข้อความจะถูกแทนที่ด้วยเวกเตอร์น้ำหนักของคำในข้อความนั้น โดยจะเก็บรักษาไว้ k คำแรกที่มีค่าน้ำหนักสูงที่สุดในแต่ละเวกเตอร์ จากนั้นนำเวกเตอร์ของข้อความไปเปรียบเทียบกับเวกเตอร์ของแต่ละกลุ่มข้อความ โดยเวกเตอร์ของกลุ่มคือจุดศูนย์กลางของกลุ่มข้อความซึ่งหาได้จากการทำนอร์มัลไลซ์ (Normalize) เวกเตอร์ทุกเวกเตอร์ในกลุ่ม การเปรียบเทียบเวกเตอร์จะใช้ค่าสัมประสิทธิ์ความคล้ายโคไซน์ในการหาความคล้ายระหว่างเวกเตอร์ของข้อความกับเวกเตอร์ของกลุ่ม โดยข้อความจะถูกรวมเข้าไปอยู่ในกลุ่มที่มีความคล้ายมากที่สุดและความคล้ายนั้นจะต้องมีค่ามากกว่าเส้นแบ่งที่กำหนดไว้ล่วงหน้า (Predefined Threshold) ในทางกลับกันถ้าความคล้ายระหว่างเวกเตอร์ของคำกับเวกเตอร์ของกลุ่มที่มากที่สุดนั้นต่ำกว่าเส้นแบ่งที่กำหนดไว้ล่วงหน้า แสดงว่าข้อความที่กำลังสนใจมีความแตกต่างกับข้อความในอดีตมาก แสดงว่าเป็นข้อความที่กล่าวถึงสิ่งใหม่ และจะทำการสร้างเป็นกลุ่มใหม่แทน ดังแสดงในรูปที่ 18



รูปที่ 18 การหาความสัมพันธ์ระหว่างข้อความใหม่ กับกลุ่มของข้อความ

ถัดมาในปี ค.ศ. 2000 Allan, Lavrenko [15] ได้นำเสนอวิธีการจับกลุ่มกระแสของข้อความใหม่ โดยการหาข้อความที่เข้ามาใหม่มีความคล้ายกับข้อความในฐานข้อมูลข้อความใดด้วยอัลกอริทึมค้นหาเพื่อนบ้านใกล้สุด (Nearest Neighbor) แต่ถ้าข้อความใหม่นั้นมีความคล้ายต่ำกว่าเส้นแบ่งที่กำหนดไว้ล่วงหน้า (Predefined threshold) ถือว่าเป็นข้อความใหม่ แต่เนื่องจากการคำนวณหาความคล้ายของทุกข้อความใหม่ เทียบกับทุกข้อความในฐานข้อมูล ถือเป็นภาระในการคำนวณที่หนักมาก ดังนั้น Petrović, Osborne [16] จึงพัฒนาต่อยอดวิธีการนี้ด้วยการนำ Locality Sensitive Hashing (LSH) [17] มาช่วยในการคำนวณการค้นหาเพื่อนบ้านใกล้สุด

การตรวจจับหัวข้อเกิดใหม่จากการนำข้อความที่เข้ามาใหม่ไปเปรียบเทียบกับข้อความหรือกลุ่มของข้อความในฐานข้อมูลได้ประสิทธิภาพไม่ค่อยดีนัก เนื่องจากข้อความของหัวข้อเกิดใหม่หลายหัวข้อมีลักษณะคล้ายกับข้อความเดิมในฐานข้อมูล เพียงแต่ไม่ได้มีปริมาณที่มากเพียงพอ

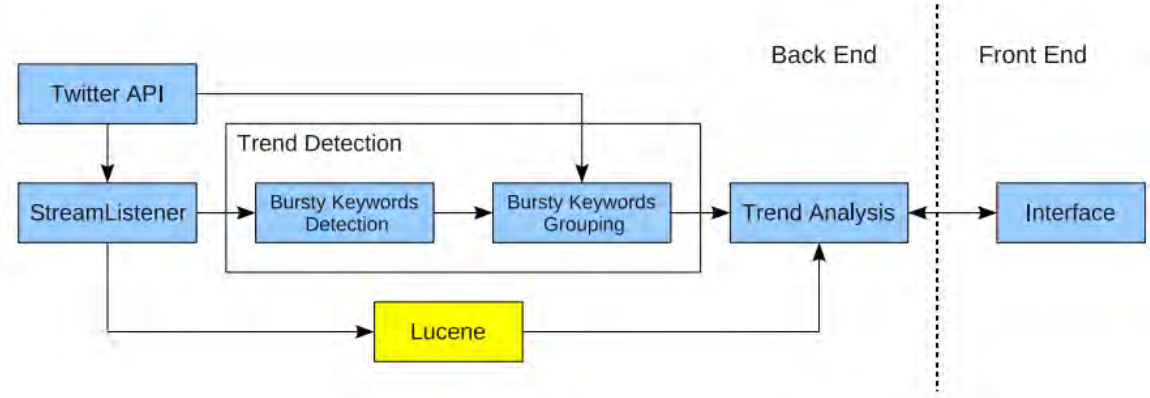
3.2 การตรวจจับหัวข้อเกิดใหม่จากการสกัดคำเกิดใหม่

ในช่วงระยะหลัง มีงานวิจัยหลายงานที่เปลี่ยนแนวคิดของการตรวจจับหัวข้อเกิดใหม่ จากการจับกลุ่มของข้อความที่เกี่ยวข้อง โดยข้อความที่ไม่สามารถจับกลุ่มได้ถือว่าเป็นข้อความของหัวข้อเกิดใหม่ เป็นการหาคำที่น่าจะเป็นคำเกิดใหม่โดยพิจารณาจากพฤติกรรมของคำที่เปลี่ยนไป แล้วนำคำเกิดใหม่ที่เกี่ยวข้องมาจับกลุ่มรวมกันเป็นหัวข้อเกิดใหม่

ในปี ค.ศ. 2010 Mathioudakis and Koudas [1] ได้นำเสนอการตรวจจับหัวข้อเกิดใหม่ในชื่อของ TwitterMonitor โดยมีสมมติฐานว่าคำเกิดใหม่คือ “คำที่มีปริมาณสูงเมื่อเทียบกับช่วงเวลาปกติ” เช่น คำว่า “NBA” ปกติปรากฏโดยเฉลี่ย 5 ข้อความต่อนาที แต่เมื่อมันปรากฏ 100 ข้อความต่อนาทีจะถือว่ามีปริมาณที่สูงผิดปกติ และเป็นคำเกิดใหม่ ต่อจากนั้นนำกลุ่มคำเกิดใหม่ที่ตรวจจับได้

ไปหาความสัมพันธ์กัน โดยดูจากการปรากฏร่วมของคำในข้อความของกรอบเวลาในอดีตสั้น ๆ ที่พึ่งผ่านมา โดยถ้าคำไหนมีความสัมพันธ์กันมากจะถูกจับกลุ่มรวมกันเป็นหัวข้อเกิดเดียวกัน ซึ่งความสัมพันธ์ระหว่างคำของงานวิจัยนี้ ใช้อัลกอริทึม LSI ในการหาเวกเตอร์ของแต่ละคำ จากนั้นนำเวกเตอร์เหล่านี้มาหาความสัมพันธ์กันด้วยสัมประสิทธิ์ความคล้ายโคไซน์ โดยได้อธิบายวิธีการหาเวกเตอร์ของคำด้วยอัลกอริทึม LSI ไว้ในบทที่ 2.6

รูปที่ 19 แสดงการทำงานของระบบตรวจจับหัวข้อเกิดใหม่ของงานวิจัยนี้ โดยรับกระแสของข้อความที่วิ่งเข้ามาตลอดเวลา จากนั้นนำข้อความนั้นไปหาคำเกิดใหม่ และนำคำเกิดใหม่ไปจับกลุ่มเป็นหัวข้อเกิดใหม่

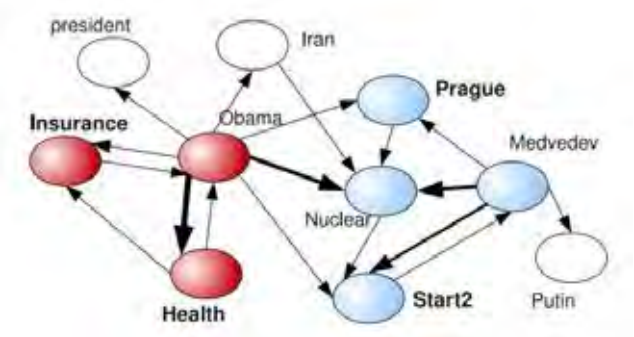


รูปที่ 19 สถาปัตยกรรมในการตรวจจับหัวข้อเกิดใหม่ของงานวิจัย TwitterMonitor [1] [13]

ในปีเดียวกัน Cataldi, Caro [2] ได้นำเสนอคุณลักษณะใหม่ที่ใช้ในการตรวจจับคำเกิดใหม่ โดยมีสมมติฐานว่าคำเกิดใหม่คือ “คำที่ปรากฏบ่อยในปัจจุบัน แต่ปรากฏน้อยในอดีต” ซึ่งคล้ายกับงานวิจัย TwitterMonitor โดยในงานวิจัยนี้เสนอคุณลักษณะใหม่ที่เรียกว่า “พลังงาน” ซึ่งเป็นคุณลักษณะที่เกิดจาก 2 แง่มุมผสมกัน แง่มุมแรกคือน้ำหนัก ซึ่งเกิดจากน้ำหนักของคำแต่ละคำเทียบกับน้ำหนักสูงสุดในข้อความนั้น และแง่มุมที่สองคือความสัมพันธ์ของผู้ใช้งานในเครือข่ายสังคมออนไลน์ จากนั้นใช้เส้นแบ่งที่กำหนดไว้ล่วงหน้า (Predefined threshold) ในการคัดเลือกคำที่น่าจะเป็นคำเกิดใหม่ จากนั้นนำกลุ่มคำเกิดใหม่ที่ตรวจจับได้ไปจับกลุ่มเป็นหัวข้อเกิดใหม่ โดยมีขั้นตอนในการสร้างหัวข้อเกิดใหม่ 3 ขั้นตอน ดังนี้

[13] อ้างอิงจาก Fig. 1

1. การหาความสัมพันธ์ของแต่ละคู่ของคำเกิดใหม่ โดยพิจารณาจากกลุ่มของข้อความที่มีทั้งสองคำปรากฏร่วมกันเดียวกัน และกลุ่มของข้อความที่ปรากฏคำใดคำหนึ่ง
2. การสร้างกราฟของหัวข้อเกิดใหม่ จากความสัมพันธ์ของแต่ละคู่ของคำเกิดใหม่ โดยมีโหนด (Node) คือคำเกิดใหม่ และมีเส้นเชื่อมระหว่างโหนด (Edge) เป็นความสัมพันธ์ของแต่ละคู่ของคำเกิดใหม่
3. การสร้างหัวข้อเกิดใหม่ โดยพิจารณาจากกราฟย่อย (Sub Graph) กล่าวคือกราฟย่อย 1 กราฟ คือตัวแทนของหัวข้อเกิดใหม่ 1 หัวข้อ ดังแสดงในรูปที่ 20



รูปที่ 20 กราฟของหัวข้อเกิดใหม่ 2 หัวข้อ [14]

ในปีถัดมาปี ค.ศ. 2011 Weng, Yao [3] ได้นำเสนอการตรวจจับหัวข้อเกิดใหม่แบบใหม่โดยมีชื่อว่า EDCoW ซึ่งย่อมาจาก Event Detection with Clustering of Wavelet-based Signals โดยการใช้เทคนิคของการวิเคราะห์คลื่นมาใช้ในการตรวจจับคำเกิดใหม่ แทนการใช้ความถี่ในการปรากฏของคำ จากนั้นดูการเปลี่ยนแปลงของสัญญาณในการตรวจจับคำเกิดใหม่ จากนั้นทำการจับกลุ่มคำเกิดใหม่ที่ตรวจจับได้เป็นหัวข้อเกิดใหม่ โดยใช้อัลกอริทึม Modularity-based graph partitioning เพื่อสร้างกราฟของคำและความสัมพันธ์ของคำ จากนั้นทำการตัดกราฟให้เป็นกราฟย่อย โดยกราฟย่อย 1 กราฟคือตัวแทนของหัวข้อเกิดใหม่ 1 หัวข้อ

ในปีเดียวกัน Alvanaki, Sebastian [4] ได้นำเสนอการตรวจจับหัวข้อเกิดใหม่บนบล็อคและข้อความทวิตเตอร์ ที่มีชื่อว่า enBlogue และถูกพัฒนาอีกครั้งโดย Alvanaki, Michel [5] ในปี ค.ศ. 2012 โดยมีสมมติฐานในการตรวจจับคำหรือหัวข้อเกิดใหม่คือ “ยิ่งเบี่ยงเบนมากยิ่งมีความฉุกเฉินมาก” หรือมีโอกาสเกิดเป็นคำเกิดใหม่มาก โดยในงานวิจัยนี้สนใจการเกิดร่วมกันของแต่ละคู่ของคำ

[14] อ้างอิงจาก Fig. 5

มากกว่าการเกิดของคำเพียงคำเดียว ดังนั้นงานวิจัยนี้จึงสร้างคุณลักษณะบนคู่ของคำแทนการสร้างคุณลักษณะของคำแต่ละคำเหมือนในงานวิจัยก่อนหน้า ซึ่งคุณลักษณะของคู่ของคำที่งานวิจัยนี้เสนอคือความสัมพันธ์ของคู่ของคำที่ดูจากการปรากฏร่วมกันในกลุ่มของข้อความ ซึ่งประกอบด้วยคุณลักษณะ 2 ตัว ได้แก่ local importance และ global importance โดย local importance คือ อัตราส่วนระหว่างจำนวนข้อความที่ทั้งสองคำปรากฏร่วมกับจำนวนข้อความที่คำใดคำหนึ่งปรากฏ และ global importance คือ อัตราส่วนระหว่างจำนวนข้อความที่ทั้งสองคำปรากฏร่วมกับจำนวนข้อความทั้งหมด จากนั้นนำความสัมพันธ์ของคู่ของคำดังกล่าวไปคัดเลือกเพื่อหาคู่ของคำที่เป็นคู่ของคำเกิดใหม่ โดยมีสมมติฐานว่าคู่ของคำไหนที่มีค่าของคุณลักษณะที่เสนอที่เกิดขึ้นจริงมีขนาดมากกว่าที่ทำนายไว้ด้วยสมการค่าเฉลี่ยถ่วงน้ำหนักจากข้อมูลในอดีตมาก ถือว่าเป็นคำคู่ของคำนั้นเป็นคู่ของคำเกิดใหม่ แต่ในขณะเดียวกัน การหาความสัมพันธ์ของแต่ละคู่ของคำ เป็นการเพิ่มภาระในการคำนวณมาก ดังนั้นเพื่อลดภาระในการคำนวณจึงทำการคัดเลือกคู่ของคำที่สนใจแทนการสนใจทุกคู่ของคำ

ต่อมาในปี ค.ศ. 2014 Schubert, Weiler [6] ได้พัฒนาการตรวจจับคำและหัวข้อเกิดใหม่แบบใหม่ในชื่อของ SigniTrend โดยมีแนวคิดมาจากงานวิจัยเกี่ยวกับการรักษาค่าทางสถิติบนกระแสของข้อมูลของ Datar, Gionis [18] และ Babcock, Datar [19] โดยทำการเสนอคุณลักษณะที่สามารถเพิ่มค่าได้ (Incremental feature) กล่าวคือ เมื่อมีกระแสของข้อมูลใหม่เข้ามา คุณลักษณะดังกล่าวจะถูกปรับปรุงค่าจากค่าเดิมเป็นค่าใหม่ได้โดยไม่จำเป็นต้องคำนวณใหม่ ในงานวิจัยนี้เสนอคุณลักษณะที่สามารถเพิ่มค่าได้ 2 ตัว คือ ค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก $EWMA$ (Exponential Weight Moving Average) และค่าความแปรปรวนเคลื่อนที่ถ่วงน้ำหนัก $EWMVar$ (Exponential Weight Moving Variance) โดยใช้สมการในการคำนวณคุณลักษณะทั้งสองตัวนี้จาก Welford [20], West [21] และ Finch [22] โดยมีสมการดังต่อไปนี้

กำหนดให้

$count_t$ คือ จำนวนคำที่ปรากฏในเวลา t

w คือ กรอบเวลา

t คือ เวลา

α คือ อัตราการเรียนรู้

$$\Delta_{w,t} = count_t - EWMA_{w,t} \quad (29)$$

$$EWMA_{w,t} = EWMA_{w,t-1} + \alpha * \Delta_{w,t} \quad (30)$$

$$EWMVar_{w,t} = (1-\alpha) * (EWMVar_{w,t-1} + \alpha * \Delta_{w,t}{}^2) \quad (31)$$

จากนั้นนำคุณลักษณะค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก $EWMA$ และคุณลักษณะค่าความแปรปรวนเคลื่อนที่ถ่วงน้ำหนัก $EWMVar$ ไปคำนวณหาค่าเบี่ยงเบนมาตรฐาน z ซึ่งเป็นคุณลักษณะสุดท้ายที่ใช้ในงานวิจัยนี้ ที่ใช้ในการเลือกคู่ของคำที่จะเป็นคู่ของคำเกิดใหม่ ดังสมการต่อไปนี้

$$z_{w,t} = \frac{count_t - EWMA_{w,t}}{\sqrt{EWMVar_{w,t}}} \quad (32)$$

แต่เนื่องจาก $EWMVar$ ของบางคำมีขนาดที่น้อยมาก ซึ่งทำให้ค่าเบี่ยงเบนมาตรฐาน z ที่คำนวณได้มีค่าสูงเกินไป จึงทำการใส่ค่าความลำเอียง β ลงไป และได้สมการในการคำนวณค่าเบี่ยงเบนมาตรฐานใหม่ z ดังนี้

$$z_{w,t} = \frac{count_t - MAX(EWMA_{w,t}, \beta)}{\sqrt{EWMVar_{w,t}} + \beta} \quad (33)$$

โดยในงานวิจัยนี้มีลักษณะคล้ายกับ [4, 5] คือการหาคุณลักษณะของคู่ของคำ แทนการหาคุณลักษณะของคำเดี่ยว และแก้ปัญหาเรื่องภาระในการคำนวณโดยการทำแฮช (Hash) ด้วยตัวกรองของบลูม (Bloom filter) และแฮชต่ำสุด (MinHash) แทนการคัดเลือกคู่ของคำที่สนใจ เนื่องจากการคัดเลือกคู่ของคำที่สนใจทำให้คู่ของคำที่ควรจะเป็นคู่ของคำเกิดใหม่สูญหายจำนวนมาก จากนั้นนำคู่ของคำที่เป็นคู่ของคำเกิดใหม่ไปจับกลุ่มเป็นหัวข้อเกิดใหม่โดยใช้อัลกอริทึมการจับกลุ่มแบบมีลำดับชั้น (Hierarchical clustering) เมื่อสิ้นวัน

ในทำนองเดียวกัน Xie, Zhu [7] ได้นำเสนอการตรวจจับคำและหัวข้อเกิดใหม่แบบใหม่ในชื่อของ TopicSketch ในปี ค.ศ. 2013 ซึ่งถูกพัฒนาต่อยอดอีกครั้งในปี ค.ศ. 2016 โดย Xie, Zhu [8] งานวิจัยนี้ได้นำเสนอคุณลักษณะใหม่ที่สามารถเพิ่มค่าได้ (Incremental feature) และสามารถคำนวณแบบขี้เกียจได้ (Lazy computing) กล่าวคือไม่จำเป็นต้องคำนวณคุณลักษณะของทุกคู่ของคำใหม่ทุกช่วงเวลา ช่วงเวลาใดไม่มีคู่ของคำไหนปรากฏ แสดงว่าคู่ของคำนั้นไม่สำคัญในช่วงเวลานั้น สามารถละเว้นการคำนวณคุณลักษณะของคู่ของคำนั้นได้ แต่เมื่อคู่ของคำนั้นปรากฏก็สามารถอัพเดทค่าได้โดยไม่จำเป็นต้องคำนวณใหม่ โดยคุณลักษณะใหม่ที่สามารถเพิ่มค่าได้แบบขี้เกียจของงานวิจัยนี้ คือ ความเร็ว v และความเร่ง acc ซึ่งพัฒนาสมการในการคำนวณคุณลักษณะมาจากงานวิจัยของ He and Parker [23]

กำหนดให้

$count_t$ คือ จำนวนคำที่ปรากฏในเวลา t

w คือ กรอบเวลา

t คือ เวลา

$t-1$ คือ เวลาครั้งล่าสุดที่ปรากฏ

ΔT คือ ระยะห่างของเวลาที่ปรากฏครั้งล่าสุด

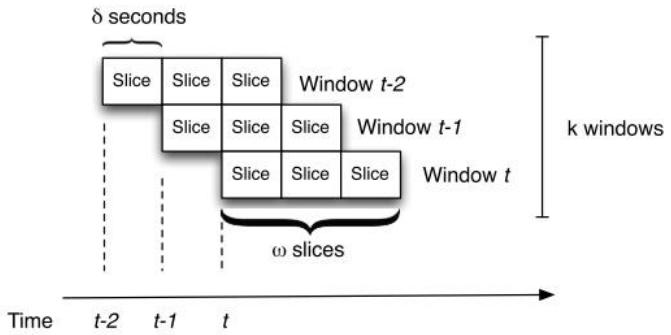
$$v_{w,t} = v_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{count_t}{w} \tag{34}$$

จากนั้นนำ v ไปสร้างเป็นสมการความต่างของ v หรือความเร่ง acc ดังนี้

$$acc_{w1,w2,t} = \frac{v_{w1,t} - v_{w2,t}}{w2 - w1} \tag{35}$$

ซึ่งในมุมมองของตัวชี้วัดของหุ้นคุณลักษณะ v มีความคล้ายกับตัวชี้วัดของหุ้น $EWMA$ ในบทที่ 2.2.2 และคุณลักษณะ acc มีความคล้ายกับตัวชี้วัดของหุ้น $MACD$ ในบทที่ 2.2.3

ในปี ค.ศ. 2015 Buntain, Lin [9] ได้พัฒนาการตรวจจับคำเกิดใหม่โดยใช้หลายคุณลักษณะ (Multivariate features) ด้วยตัวจำแนกประเภทในชื่อ LABurst วิธีของงานวิจัยนี้เริ่มจากหาคำที่เป็นคำสำคัญของแต่ละเหตุการณ์ เช่น เหตุการณ์ของแข่งขันกีฬา จะหาคำสำคัญจากนิตยสารกีฬา คะแนนของการแข่งขัน บล็อค ข้อความจากเครือข่ายสังคมออนไลน์ เป็นต้น จากนั้นนำคำสำคัญนี้เหล่าไปสร้างตัวแปรผลเฉลย (Data labeling) หรือตัวแปรเป้าหมาย (Target variable) เพื่อใช้ในการเรียนรู้ของตัวจำแนกประเภท โดยถ้าข้อมูลในกรอบเวลา w มีจำนวนคำสำคัญที่แตกต่างกันมากกว่าเส้นแบ่งที่กำหนดไว้ล่วงหน้า ρ แล้ว จะกำกับข้อมูลที่มีคำเหล่านี้ในกรอบเวลาดังกล่าวว่าเป็นข้อมูลที่สำคัญ และนำไปใช้สร้างตัวจำแนกประเภท ซึ่งการกำกับข้อมูลหรือการสร้างตัวแปรผลเฉลย (Data labeling) ของวิธีการนี้จำเป็นต้องใช้พารามิเตอร์หลายตัว อาทิเช่น เส้นแบ่งของจำนวนคำสำคัญที่แตกต่างกัน ρ, ค่าความล่าช้า (Delay) τ, จำนวนวินาทีในแต่ละช่องของเวลา δ โดยมีค่าความกว้างกรอบเวลา w เท่ากับ $(\tau + 1) * \delta$ ดังแสดงในรูปที่ 21 ซึ่งการหาค่าที่ดีที่สุดของแต่ละพารามิเตอร์ทำได้ยากและแตกต่างกันไปตามปัจจัยหลายอย่าง เช่น ลักษณะหัวข้อเกิดใหม่ที่สนใจ จำนวนคำในชุดผลเฉลยที่สนใจ จำนวนข้อความทวิตเตอร์ที่ใช้ในการสอน เป็นต้น โดยในงานวิจัยนี้ใช้ค่า δ เท่ากับ 60 และค่า w เท่ากับ 180 โดยไม่สามารถระบุเหตุผลถึงการเลือกพารามิเตอร์ดังกล่าว



รูปที่ 21 ความสัมพันธ์ของแต่ละพารามิเตอร์ในการสร้างตัวแปรผลเฉลยและสร้างคุณลักษณะของงานวิจัย LABurst [9] [15]

[15] อ้างอิงจาก Fig. 1

การตรวจจับหัวข้อเกิดใหม่จากการสกัดคำเกิดใหม่ โดยการสร้างคุณลักษณะของแต่ละคำและคัดเลือกคำที่มีพฤติกรรมโดดเด่นหรือแตกต่างจากพฤติกรรมในอดีต มีหลายงานวิจัย [1-8] ที่นำเสนอคุณลักษณะใหม่ในการหาคำเกิดใหม่ โดยคุณลักษณะที่น่าสนใจคือคุณลักษณะ z และ acc เนื่องจากเป็นคุณลักษณะที่สามารถเพิ่มค่าได้ (Incremental feature) แต่การใช้คุณลักษณะดังกล่าวเพียงคุณลักษณะเดียว (Univariate feature) เป็นการยากที่จะตรวจจับคำเกิดใหม่ได้หลากหลายรูปแบบ แม้จะมีบางงานวิจัย [9] พยายามใช้หลายคุณลักษณะ (Multivariate features) ในการตรวจจับคำเกิดใหม่ด้วยตัวจำแนกประเภท แต่ด้วยการกำกับหรือการสร้างตัวแปรผลเฉลย (Data labeling) ของข้อมูลที่ใช้สร้างตัวนำแจกประเภทจำเป็นต้องใช้ค่าพารามิเตอร์หลายตัว ทำให้ยากต่อการหาค่าพารามิเตอร์และการนำไปใช้ เนื่องจากหากใช้ค่าพารามิเตอร์ไม่เหมาะสม อาจทำให้บางหัวข้อเกิดใหม่ไม่ปรากฏในชุดข้อมูล หรือ ทำให้เกิดการกำกับข้อมูลส่วนที่ไม่ใช่เหตุการณ์ได้

3.3 การประเมินผลการทดลองของงานวิจัย

การประเมินประสิทธิภาพของงานวิจัยเกี่ยวกับการตรวจจับหัวข้อเกิดใหม่ ค่อนข้างไม่ชัดเจน เนื่องจากไม่มีชุดผลเฉลยที่ชัดเจนหรือวิธีการวัดประสิทธิภาพที่เป็นมาตรฐาน และไม่สามารถนิยามได้ว่าคำหรือหัวข้อที่จะเป็นคำเกิดใหม่นั้นจะต้องถูกสนใจขนาดไหนถึงจะเรียกว่าคำเกิดใหม่หรือหัวข้อเกิดใหม่ ทำให้งานวิจัยส่วนใหญ่มุ่งเน้นไปที่การตรวจจับเหตุการณ์ที่สำคัญและดังมาก ๆ เท่านั้น เช่น เหตุการณ์แผ่นดินไหว เหตุการณ์ก่อการร้าย เหตการณ์แข่งขันกีฬา เป็นต้น

ในงานวิจัยของ [1] และ [2] ไม่มีการประเมินประสิทธิภาพของการตรวจจับคำหรือหัวข้อเกิดใหม่ มีเพียงผลลัพธ์สุดท้ายของการตรวจจับหัวข้อเกิดใหม่ ดังแสดงในรูปที่ 22 กับรูปที่ 23 และมีการแสดงค่าของคุณลักษณะที่เสนอของคำที่เป็นคำเกิดใหม่ ดังแสดงในรูปที่ 24

รูปที่ 22 ผลลัพธ์การตรวจจับหัวข้อเกิดใหม่ของงานวิจัย TwitterMonitor [1] [16]

Date	Emerging Topics (s=100)
20-04-2010	{**morning**,early, tuesday, sleep}
20-04-2010	{**music**, album ,video, stereo}
20-04-2010	{**laundry**, citrus ,urgent, liquid}
20-04-2010	{**profile**, facebook ,post, link}

(a)

Date	Emerging Topics (s=200)
20-04-2010	{**activist**,dorothy, height, death}
20-04-2010	{**rockies**, president, team, dead}

(b)

รูปที่ 23 ผลลัพธ์การตรวจจับหัวข้อเกิดใหม่ของงานวิจัย Cataldi, Caro [2] [17]

Date	Emerging Terms	Energy value ($total\ avg$ = 5.6517)
15-04-2010	**eyjafjallajökull**	7773.7575
15-04-2010	**bieber**	147.1661
15-04-2010	**wellington**	115.3432
15-04-2010	**betezy**	76.7339
15-04-2010	**diaper**	55.3219

รูปที่ 24 ค่าของคุณลักษณะที่เสนอของคำที่เป็นคำเกิดใหม่ของงานวิจัย Cataldi, Caro [2] 18

ในงานวิจัย [3] มีแสดงผลลัพธ์จากการตรวจจับหัวข้อเกิดใหม่ในรูปที่ 25 และมีการเปรียบเทียบกับงานวิจัยอื่น แต่เป็นการเปรียบเทียบเชิงบรรยาย กล่าวคือเป็นการบรรยายข้อดีและ

[16] อ้างอิงจาก Fig. 2

[17] อ้างอิงจาก Table 2

[18] อ้างอิงจาก Table 3

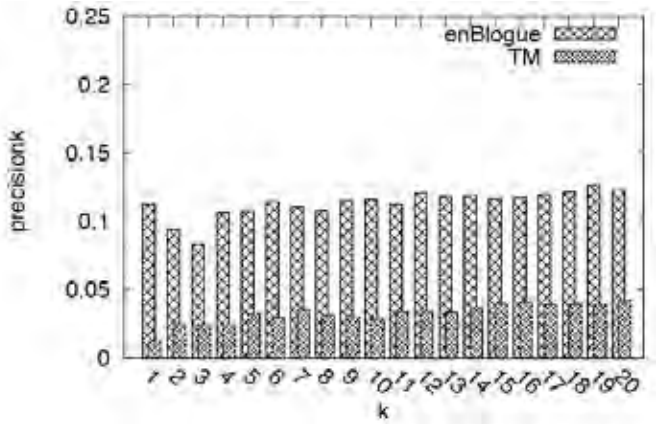
ข้อเสียของงานวิจัยนี้เทียบกับงานวิจัยอื่นโดยไม่มีหลักฐานชัดเจน หรือมีตัวเลขที่สามารถชี้วัดประสิทธิภาพของงานวิจัยว่าดีกว่าจริงหรือไม่และดีกว่าเท่าไหร่

Day	Event	ε value	Event Description
1–3	No event detected		
4	1. democrat, naoto	0.417	Ruling Democratic Party of Japan elected Naoto Kan as chief.
	2. ss501, suju	0.414	Korean popular bands Super Junior's and SS501's performance on mubank.
	3. music, mubank	0.401	Related to Event 2. mubank is a popular KBS entertainment show.
	4. shindong, youngsaeng	0.365	Related to Event 2. Shindong and Youngsaeng are member of the two bands.
	5. junior, eunhyuk	0.124	Related to Event 2. Eunhyuk is a member of super junior.
5	6. robben, break	0.404	No clear corresponding real-life even
6	No event detected		
7	7. kobe, kristen	0.417	Two events: Kristen Stewart won some MTV awards, and Kobe Bryant in a NBA match.
	8. #iphone4, ios4, iphone	0.416	iPhone 4 released during WWDC 2010
8	9. reformat, hamilton	0.391	No clear corresponding real-life event
	10. avocado, commence, ongoing	0.124	No clear corresponding real-life event
9	11. #fail whale, twitter	0.360	A number of users complained they could not use twitter due to over-capacity. A logo with whale is usually used to denote over-capacity.
10	12. vuvuzela, soccer	0.387	People started to talk about world cup.
11	13. #svk, #svn	0.418	#svk and #svn represent Team Slovakia and Slovenia in World Cup 2010.
12	14. #kor, greec, #gre	0.102	A match between South Korea and Greece in World Cup 2010.
13	15. whale, twitter	0.417	Similar as Event 10.
14	16. lippi, italy	0.326	Italy football team coach Marcello Lippi made some comments after a match in World Cup 2010.
15	17. drogba, ivory	0.417	Football player Drogba from Ivory Coast is given special permission to play in World Cup 2010.
	18. #prk, #bra, north	0.114	A match between North Korea and Brazil in World Cup 2010.
16	19. orchard, flood	0.357	Flood in Orchard Road.
17	20. greec, #gre, nigeria	0.122	A match between Greece and Nigeria in World Cup 2010.
18	21. #svk, podolski	0.403	A match between Germany and Serbia in World Cup 2010. Podolski is a member of Team Germany in World Cup 2010.
19-30	No event detected		

รูปที่ 25 ผลลัพธ์การตรวจจับหัวข้อเกิดใหม่ของงานวิจัย EDCoW [3] [19]

ในงานวิจัย [4] ไม่ได้แสดงผลลัพธ์หรือประสิทธิภาพของการตรวจจับหัวข้อเกิดใหม่ แต่ในงานวิจัยที่ต่อยอด [5] มีการวัดประสิทธิภาพเปรียบเทียบกับ TwitterMonitor [1] โดยใช้ $Precision@K$ และ $NDCG@k$ (Normalized discounted cumulative gain) ในรูปที่ 26 และรูปที่ 27 ซึ่งมีแนวคิดมาจากงานวิจัยของ Järvelin and Kekäläinen [24] ซึ่งการวัดประสิทธิภาพของงานวิจัยนี้ให้อาสาสมัครนักศึกษาในมหาวิทยาลัยจำนวน 80 คนช่วยกันประเมิน โดยการนำหัวข้อเกิดใหม่ดีที่สุด k อันดับแรกของทั้งสองวิธีมาตรวจสอบโดยอาสาสมัคร และทำการกำกับว่าหัวข้อเกิดใหม่แต่ละหัวข้อที่ตรวจจับได้เป็นหัวข้อเกิดใหม่ใช่หรือไม่ โดยดูจากข้อความทวิตเตอร์ตัวอย่างที่มีคำเกิดใหม่ปรากฏ

[19] อ้างอิงจาก Table 1



รูปที่ 26 ผลลัพธ์ของการวัดประสิทธิภาพเปรียบเทียบกับ TwitterMonitor [1] ของงานวิจัย [5] [20]

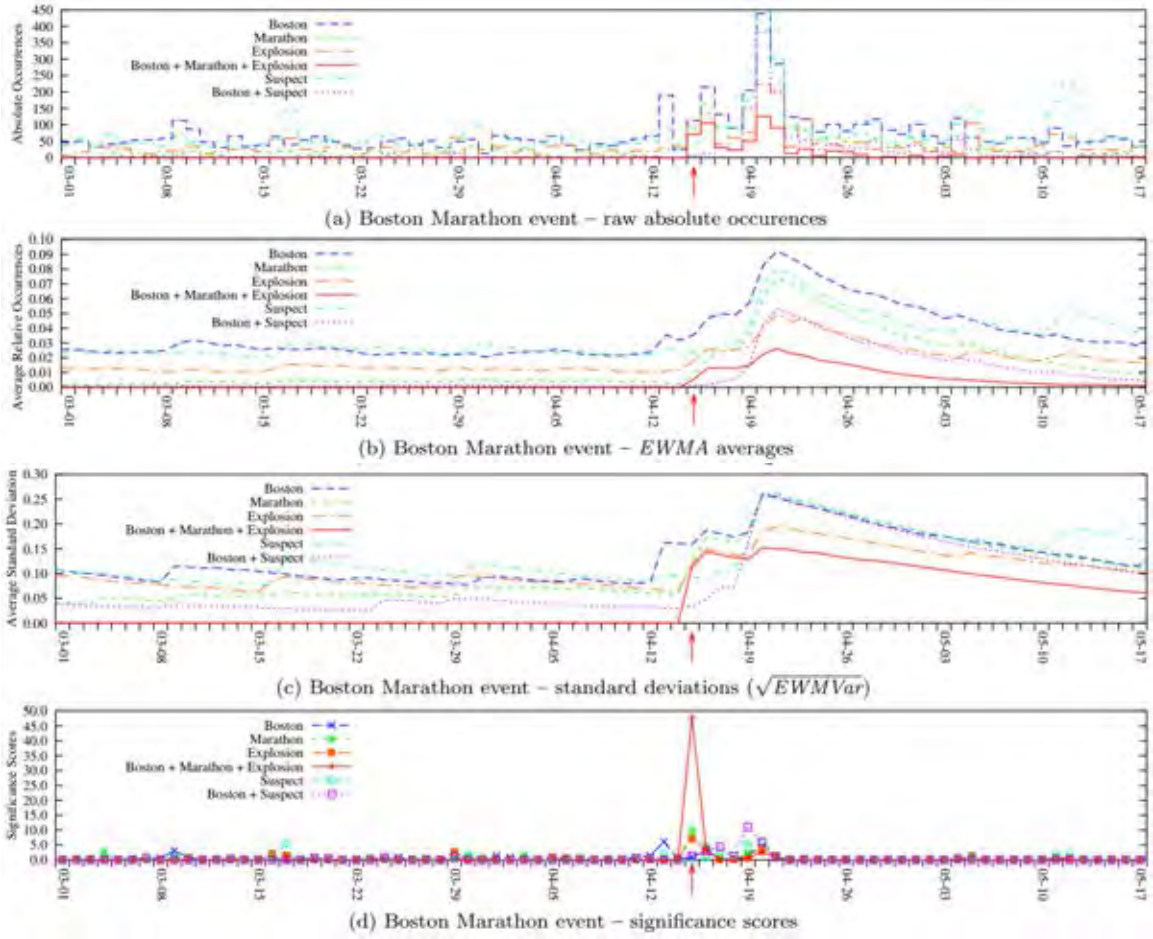
	Precision@k			NDCG@k	
k	enBlogue	TM	k	enBlogue	TM
1	0.112	0.012	1	0.112	0.012
2	0.094	0.025	2	0.094	0.025
3	0.083	0.025	3	0.089	0.026
4	0.106	0.025	4	0.115	0.028
5	0.108	0.032	5	0.123	0.033
6	0.115	0.029	6	0.137	0.033
7	0.111	0.036	7	0.151	0.041
8	0.108	0.031	8	0.160	0.041
9	0.115	0.031	9	0.177	0.044
10	0.116	0.029	10	0.191	0.045
11	0.112	0.034	11	0.198	0.053
12	0.121	0.034	12	0.220	0.056
13	0.118	0.034	13	0.229	0.059
14	0.119	0.038	14	0.239	0.068
15	0.117	0.04	15	0.246	0.073
16	0.118	0.041	16	0.258	0.077
17	0.120	0.040	17	0.270	0.078
18	0.122	0.040	18	0.285	0.081
19	0.126	0.039	19	0.3	0.083
20	0.122	0.042	20	0.304	0.09

รูปที่ 27 ผลลัพธ์ของการวัดประสิทธิภาพเปรียบเทียบกับ TwitterMonitor [1] ของงานวิจัย [5] [21]

[20] อ้างอิงจาก Fig. 6

[21] อ้างอิงจาก Table 2

ในงานวิจัย [6] มีการแสดงการเคลื่อนที่ของคุณลักษณะที่เสนอมาในงานวิจัยนี้ในรูปที่ 28 และแสดงผลลัพธ์จากการตรวจจับหัวข้อเกิดใหม่ในรูปที่ 29 ของคำเกิดใหม่จากเหตุการณ์ระเบิดในงานวิ่งที่เมืองบอสตัน แต่ไม่มีการเปรียบเทียบผลลัพธ์กับงานวิจัยอื่น



รูปที่ 28 ค่าของคุณลักษณะที่เสนอของคำเกิดใหม่ที่เลือกมาแล้ว ของเหตุการณ์ระเบิดในงานวิ่งที่เมืองบอสตัน SigniTrend [6] [22]

[22] อ้างอิงจาก Fig. 1

Table 3: Excerpt of top 50 trends on news data set 2013 (dominated by economy, sports and politics)

Score	Date	Stemmed Keywords (excerpt, edited for readability)	Explanation
58	11-18	thomson text summary 3000xtra alert outperform eikon	Reuters artifact – 233 research alerts
13	10-09	janet yellen ben bernank vice barack obama nomin	Obama nominates Yellen as U.S. Fed Chief
10	02-14	heinz buffett gdp hj merger shrank berkshir hathaway warren	Berkshire Hathaway, 3G Capital buy Heinz
9.7	07-03	turmoil armi unrest egyptian lisbon egypt mursi portug bailout	Yen rises because of turmoil in Egypt, Portugal
9.4	04-19	boston search bomb suspect marathon manhunt	Boston Marathon suspect manhunt
9.3	04-15	boston explos marathon	Boston Marathon bombing
9.3	01-28	durabl caterpillar pend	Four-month high oil, strong durable goods
8.8	09-19	inning era	Baseball end of season reports
8.8	02-26	ben bernank testimoni defend deadlock stalem	Bernanke defends bond-buying, Italian stalemate
8.6	07-11	ben bernank minut accommod forese dovish	Bernanke reassures euro bonds markets
		Economical, financial and sports news largely omitted outside the top 10 for brevity	
8.6	04-16	finish sharp metal boston marathon rebound terror bomb explos injur	Boston Marathon attack details
8.6	03-25	rescu cyprus bailout eurogroup guarante relief dutch jeroen dijsselbloem	EU cyprus bailout
8.5	01-17	hostag desert milit algeria algerian	Algerian gas plant attack
8.5	09-03	finnish wireless handset nokia smartphon microsoft	Microsoft buys Nokia's handset business
8.3	01-23	davo forum switzerland cameron referendum	Cameron promises referendum to leave EU
7.8	11-24	aton reactor iranian geneva enrich breakthrough lift uranium	Breakthrough on Iran nuclear activity
7.8	08-26	chemic weapon secretari holiday durabl	Kerry comments on Syria over chemical weapons
7.7	05-21	oklahoma moor tornado	Moore, Oklahoma hit by deadly tornados
7.6	09-23	merkel angela coalit victori william dudley shutdown	German elections succes for Angela Merkel
7.3	10-01	deadlock shutdown midnight began trillion unpaid barack obama	U.S. government partial shutdown
7.2	08-27	syrian chemic weapon strike assad tension kerri secretari	Escalations in Syria
6.9	04-20	dzhokhar tamerlan tsarnaev brother suspect captur dead watertown injur	Boston Marathon suspects captured
6.8	07-24	appl iphon smartphon markit flash beat faster	Surge in iPhone sales
6.8	12-06	nelson mandela die apartheid african africa	Nelson Mandela died
6.7	02-15	moscow communiqu	G20 finance ministers in Moscow
6.7	04-08	thatcher iron margaret	Margaret Thatcher died
6.6	07-12	airport dreamlin ethiopian heathrow boe	Boeing Dreamliner catches fire

รูปที่ 29 ผลลัพธ์การตรวจจับหัวข้อเกิดใหม่ของงานวิจัย SigniTrend [6] [23]

[23] อ้างอิงจาก Table 3

ในงานวิจัย [7] มีการแสดงผลลัพธ์จากการตรวจจับหัวข้อเกิดใหม่เทียบกับงานวิจัยอื่นในรูปที่ 30 และรูปที่ 31, การอธิบายเชิงบรรยายในข้อดีและข้อเสีย, ค่าของคุณลักษณะที่เสนอในรูปที่ 32 และนอกจากนี้ยังมีการเปรียบเทียบกับทวิตเตอร์เทรนด์ในเรื่องของความเร็วในการตรวจจับคำเกิดใหม่ในรูปที่ 33

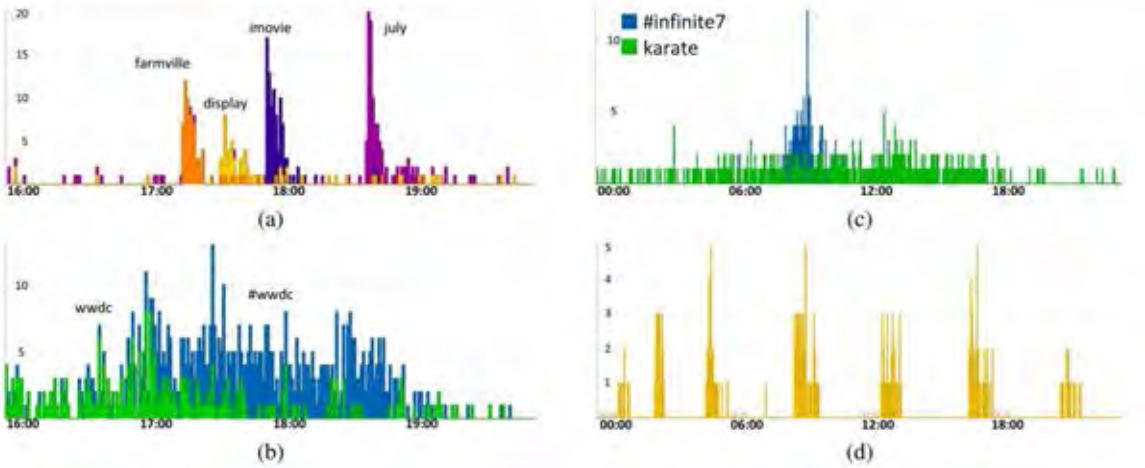
Date	Event	Sub-Event	TopicSketch	Twevent
7	MTV Moive Awards 2010	Kristen Stewart won the Best Female Performance	stewart,kristen,female,mtv	mtv movie awards, mtv, new moon, twilight, robe
		Sandra Bullock won the Genération Award	sandra,bullock,mtv,movie	
		Best Movie Award :"The Twilight Saga: New Moon"	movie,moon,mtv,awards	
	Super Junior's Yesung(@shfly3424) created his Twitter account.		yesung, twitter, @shfly3424	**None**
	Fans celebrated 3 year anniversary for boy band "T.T. Island" in Twitter.		tisland, love, <3	**None**
	Steve Jobs released iPhone 4 during WWDC 2010	Farmville client for iPhone 4 was demonstrated.	#wwdc, iphone, farmville	steve jobs, imovie, wwdc, iphone, wtf
		Retina display of iPhone 4 was introduced.	iphone,4,#wwdc,display,retina	
		iMovie for iPhone 4 was demonstrated.	iphone, 4, imovie, #wwdc	
		New iPhone 4 was available in Singapore in July.	iphone, 4, singapore, july	
8	Fans asked pop musician Justin Bieber to follow them in Twitter.		@justinbieber, follow, please	**None**
	Fans celebrated 5 year anniversary for boy band "SS501" in Twitter.		#5yrss501, ss501, #5yearss501	**None**
	The music video "Alejandro" by Lady GaGa was premiered		alejandro, video, gaga	lady gaga, alejandro music video, gaga, mv
9	Korean pop singer Eunhyuk(@allrisesilver) posted his new photo and fans said good afternoon to him.		afternoon, @allrisesilver, http://twitpic.com/1v6sc2	**None**
	Fans tried to trend hashtag #loveisalexander(Alexander is a member of Korean pop boy band U-KISS).		#loveisalexander, <3, @alexander_0729	**None**
	A number of users complained they could not use twitter due to over-capacity. A logo with whale is usually used to denote over-capacity.		**None**	twitter, whale, stupid, capacity, over again
	The season finale of American TV series Glee was broadcasted on June 8, 2010.		glee, yeah, watching	watching glee,glee,season season finale,channel
10	Korean pop boy group Infinite (consists of 7 members) had a performance.		#infinite7, huh	**None**
	Korean actor Ok Taec-Yeon(@taeccool) opened twitter account.		@taeccool,twitter,taecyeon,taec	**None**
	The movie The Karate Kid was released on June 10, 2010 in Singapore.		**None**	karate kid, movie watch movie
	Super Junior's Yesung posted a photo about his pet turtles.		**None**	yesung, tweeted
11	Super Junior was performing "Bonamana" on Music Bank.		super, junior, bonamana	**None**
	SS501 won the K-Chart on Music Bank.		ss501, won, congrats	**None**
	South Africa vs Mexico in World Cup 2010.	Match began.	south,world,cup,africa,mexico	south africa, vs mexico, mexico, goal, first goal
		South Africa first goaled. (1-0)	africa,south,goal,mexico,1-0	
		Mexico goaled. (1-1)	mexico, goal, 1-1	
		At last draw.	mexico, africa, draw, south	
	Uruguay vs France in World Cup 2010.		**None**	uruguay vs france, uruguay, france, vs
12	South Korea vs Greece in World Cup 2010.	Match began.	korea, south, greece, #kor	south korea, greece, korea vs greece, korea won, korea
		South Korea first goaled.	korea, goal, scored, 1	
		Park Ji-Sung from Korea goaled.	korea, goal, ji, 2-0	
		South Korea won the match.	korea, won, #kor, win	
	Argentina vs Nigeria in World Cup 2010.	Argentina first goaled.	argentina, goal, 1-0	arg, argentina, nigeria, argentina vs nigeria,messi
		Argentina won the match.	argentina, 1-0, nigeria, won	
	England vs USA in World Cup 2010.	Match began.	england, vs, usa, match	usa, england, eng, vs
		Steven Gerrard from England goaled.	gerrard, england, steven, goal	
		GoalKeeper Robert Green didn't catch a ball.	green, robert, wtf	
		USA goaled.	usa, england, goal, 1	

รูปที่ 30 ผลลัพธ์การตรวจจับหัวข้อเกิดใหม่เปรียบเทียบกับงานวิจัยอื่นของ TopicSketch [7] [24]

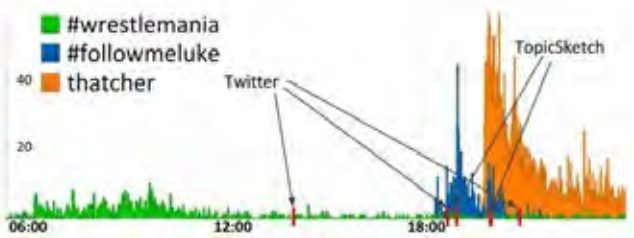
[24] อ้างอิงจาก Table 2

Event	First detected by TopicSketch	First appeared in Twitter
The popular program WrestleMania was discussed.	None	#wrestlemania **(14:34:12)**
Fans asked Luke Brooks (@luke_brooks) to follow them.	@luke_brooks, #followmeluke, follow (19:06:51)	#followmeluke **(18:54:12)**
The sudden demise of Margaret Thatcher.	thatcher, margaret **(20:05:51)**	margaret thatcher (20:49:12)

รูปที่ 31 เปรียบเทียบคำที่ตรวจจับได้ในงานวิจัย เทียบกับทวิตเตอร์เทรนด์ในหัวข้อเดียวกัน พร้อมเวลาที่ตรวจพบของหัวข้อนั้น ของงานวิจัย TopicSketch [7] [25]



รูปที่ 32 ค่าของคุณลักษณะที่เสนอของคำเกิดใหม่ที่ตรวจจับได้ในงานวิจัย TopicSketch [7] [26]



รูปที่ 33 เปรียบเทียบเวลาของคำที่ตรวจจับได้ในงานวิจัยกับคำในทวิตเตอร์เทรน [27]

งานวิจัย [8] มีการเปรียบเทียบกับงานวิจัยอื่นหลากหลายแง่มุม ได้แก่ มีการวิเคราะห์การตรวจจับหัวข้อเกิดใหม่โดยใช้ข้อมูลสังเคราะห์ที่ถูกสร้างขึ้น มีเปรียบเทียบในเชิงบรรยายข้อดีและข้อเสียกับงานวิจัยของ Li, Sun [25] และ Schubert, Weiler [6] นอกจากนี้ยังมีประเมินประสิทธิภาพในการตรวจจับหัวข้อเกิดใหม่โดยใช้ข้อมูลทวิตเตอร์เปรียบเทียบกับวิธีการอื่นในรูปที่ 34 และรูปที่ 35 โดยในงานวิจัยนี้สร้างชุดผลเฉลยโดยการนับจำนวนที่ปรากฏต่อวันของแต่ละคู่ของคำ

[25] อ้างอิงจาก Table 3

[26] อ้างอิงจาก Fig. 6

[27] อ้างอิงจาก Fig. 5

จากนั้นหาคู่ของคำที่ปรากฏต่อวันมากกว่าค่าเฉลี่ยของวันก่อนหน้า คู่ของคำไหนที่มีค่าเบี่ยงเบนมาตรฐาน z มากกว่า 3 ถือว่าเป็นคู่ของคำเกิดใหม่ในวันนั้น และจะถูกสร้างเป็นชุดผลเฉลย

กำหนดให้

$count_d$ คือ จำนวนคำที่ปรากฏในวัน d

$mean_{w,d}$ คือ ค่าเฉลี่ยของจำนวนคำที่ปรากฏในกรอบเวลา w นับจากวันที่ d

$\text{variance}_{w,d}$ คือ ค่าแปรปรวนของจำนวนคำที่ปรากฏในกรอบเวลา w นับจากวันที่ d

β คือ ค่าความลำเอียง

$$z = \frac{count_d - MAX(mean_{w,d}, \beta)}{\sqrt{\text{variance}_{w,d}} + \beta} \tag{36}$$

จากนั้นใช้คำในชุดผลเฉลยนี้ไปวัดประสิทธิภาพ Pr และ Rc โดยมีการนิยาม Pr และ Rc ใหม่ดังนี้

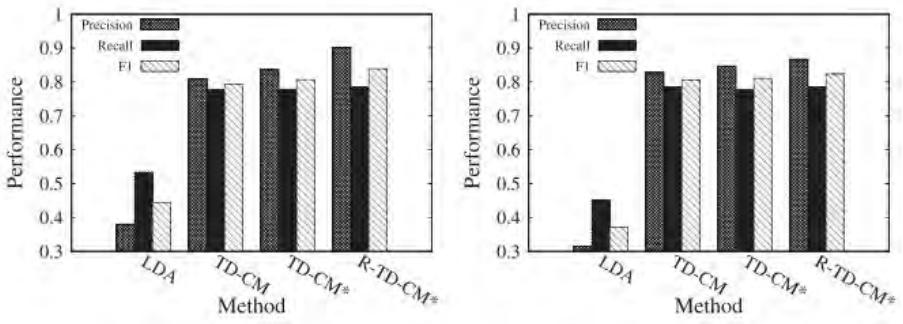
Pr คือ อัตราส่วนของจำนวนหัวข้อที่ตรวจจับได้ ที่มีคำอย่างน้อย 1 คำจาก 10 คำที่ดีที่สุดของหัวข้อนั้น ปรากฏในชุดผลเฉลย กับ จำนวนหัวข้อที่ตรวจจับได้

Rc คือ อัตราส่วนของ 1,000 คำที่มีค่าเบี่ยงเบนมากสุดต่อวัน ที่ปรากฏใน 10 คำที่ดีที่สุดของหัวข้อที่ตรวจจับได้

การเลือกคู่ของคำในชุดผลเฉลยจากค่าเบี่ยงเบนมาตรฐานนั้นมีข้อจำกัดหลายอย่าง เช่น คำที่มีการปรากฏน้อยมากต่อวัน จะเป็นคำเกิดใหม่ในวันถัดไปได้ง่าย ในขณะเดียวกันคำที่มีการปรากฏมากอยู่เป็นประจำทุกวัน การที่จะเป็นคำเกิดใหม่ได้ จำเป็นต้องมีจำนวนมากขึ้นหลายเท่า ซึ่งเป็นไปได้ยาก อีกกรณีหนึ่งคือเหตุการณ์บางอย่างที่เกิดทุกวัน แต่เกิดเพียงไม่กี่ชั่วโมงต่อวัน จะไม่ถูกนับว่าเป็นคำเกิดใหม่จากการสร้างชุดผลเฉลยของวิธีการนี้ เช่น ในช่วงแข่งขันฟุตบอลโลก มีการแข่งหลายวัน แต่คำว่า “สวัสดีบอลโลก” เกิดขึ้นมากทุกวันแต่เกิดขึ้นมากในช่วงเวลาที่มีการแข่ง ทำให้เมื่อนับจำนวนคำต่อวันแล้ว ค่าเบี่ยงเบนมาตรฐานมีค่าไม่สูงนัก แต่คำดังกล่าวถือว่าเป็นคำของหัวข้อเกิดใหม่บอลโลก

ข้อจำกัดอีกประการของการเลือกคู่ของคำในชุดผลเฉลยจากค่าเบี่ยงเบน คือ อาศัยพารามิเตอร์หลายตัว เช่น ค่า β หรือค่าความลำเอียง, กรอบเวลา w ที่ใช้การคำนวณหาค่าเฉลี่ยและ

ค่าความแปรเปรวน และเส้นแบ่งของค่าเบี่ยงเบนมาตรฐาน z ที่ในงานวิจัยนี้ใช้ค่าเบี่ยงเบนมาตรฐาน z ที่ 3 นอกจากนี้นิยามของ Pr และ Rc ไม่มีการลดค่าประสิทธิภาพเมื่อตรวจจับคำเกิดใหม่เกินกล่าวคือมาตรวัดนี้พยายามหลบเลี่ยงการตรวจจับเกิน จากข้อจำกัดที่กล่าวมาเบื้องต้น ทำให้การวัดประสิทธิภาพด้วยวิธีการดังกล่าวทำได้ยากและไม่น่าเชื่อถือ



รูปที่ 34 ประสิทธิภาพในการตรวจจับหัวข้อเกิดใหม่ ของ TopicSketch [8] [28]

Event	Sub-Event	TopicSketch	Twevent
WWDC2010	Farmville client for iPhone 4 was demonstrated.	#wwdc, farmville, iphone, zynga, netflix, god, ipad, wwdc, comes, soon	
	Retina display of iPhone 4 was introduced.	iphone, #wwdc, retina, display, pixels, crystal, clear, 326, space, interesting	steve jobs, imovie,
	iMovie for iPhone 4 was demonstrated	iphone, #wwdc, imovie, 720p, sensor 3gs, apple, 30fps, gonna, illuminated	wwdc, iphone,
	New iPhone 4 was available in Singapore in July.	iphone, singapore, july, launching, coming, 4g, available, release, #wwdc, early	wifi,

รูปที่ 35 ผลลัพธ์การตรวจจับหัวข้อเกิดใหม่เปรียบเทียบกับงานวิจัยอื่น ของ TopicSketch [8] [29]

ในงานวิจัย [9] มีการวัดผลโดยใช้ AUC (Areas Under the Curves) บนข้อมูลที่ถูกสร้างตัวแปรผลเฉลย(Data labeling) ด้วยพารามิเตอร์หลายตัว อาทิเช่น เส้นแบ่งของจำนวนคำสำคัญที่แตกต่างกัน ρ, ค่าความล่าช้า (Delay) $\boldsymbol{\tau}$, จำนวนวินาทีในแต่ละช่องของเวลา δ (ค่าความกว้างกรอบเวลา w เท่ากับ $(\tau + 1) * \delta$) ทำให้ประสิทธิภาพที่ตรวจจับได้เปลี่ยนไปตามพารามิเตอร์ที่กำหนด โดยในงานวิจัยนี้ใช้ค่า δ เท่ากับ 60 และค่า w เท่ากับ 180 โดยไม่สามารถระบุเหตุผลถึงการเลือกพารามิเตอร์ดังกล่าว นอกจากนี้ข้อมูลทวิตเตอร์ที่ใช้ในงานวิจัยนี้มีความลำเอียง เนื่องจากใช้ข้อมูลเฉพาะก่อนหรือหลังเหตุการณ์ที่สนใจ 1 ชั่วโมงเท่านั้น กล่าวคือ ข้อความที่นำมาใช้ในงานวิจัยทั้ง

[28] อ้างอิงจาก Fig. 9

[29] อ้างอิงจาก Table 2

ข้อมูลสอน (Training dataset) และข้อมูลทดสอบ (Testing dataset) เป็นข้อมูลที่ผ่านการกรองมาแล้วเชิงเวลา ทำให้ผลลัพธ์จากการตรวจจับหัวข้อเกิดใหม่ได้ประสิทธิภาพค่อนข้างดี ดังนั้นวิธีการวัดประสิทธิภาพของงานวิจัยนี้จึงไม่น่าเชื่อถือ โดยในรูปที่ 36 แสดงเหตุการณ์และจำนวนคำสำคัญของแต่ละเหตุการณ์ที่งานวิจัยนี้สนใจ โดยมีจำนวนข้อมูลของแต่ละเหตุการณ์ดังแสดงในรูปที่ 37

Sport	Key Moments
Training Data	
2010 NFL Division Championship	13
2012 Premier League Soccer Games	21
2014 NHL Stanley Cup Playoffs	24
2014 NBA Playoffs	3
2014 Kentucky Derby Horse Race	3
2014 Belmont Stakes Horse Race	3
2014 FIFA World Cup Stages A+B	80
Testing Data	
2013 MLB World Series Game 5	7
2013 MLB World Series Game 6	8
2014 NFL Super Bowl	13
2014 FIFA World Cup Third Place	11
2014 FIFA World Cup Final	7
Total	193

รูปที่ 36 แสดงจำนวนคำของแต่ละเหตุการณ์ของงานวิจัย LABurst [9] [30]

Event	Tweet Count
Training Data	
2010 NFL Division Championship	109,809
2012 Premier League Soccer Games	1,064,040
2014 NHL Stanley Cup Playoffs	2,421,065
2014 NBA Playoffs	500,170
2014 Kentucky Derby Horse Race	233,172
2014 Belmont Stakes Horse Race	226,160
2014 FIFA World Cup Stages A+B	5,867,783
Testing Data	
2013 MLB World Series Game 5	1,052,852
2013 MLB World Series Game 6	1,026,848
2013 Honshu Earthquake	444,018
2014 NFL Super Bowl	1,024,367
2014 FIFA World Cup Third Place	809,426
2014 FIFA World Cup Final	1,166,767
2014 Iwaki Earthquake	358,966
Total	16,305,443

รูปที่ 37 แสดงจำนวนข้อความทวิตเตอร์ของแต่ละเหตุการณ์ของงานวิจัย LABurst [9] [31]

[30] อ้างอิงจาก Table 2

[31] อ้างอิงจาก Table 4

การวัดประสิทธิภาพในการตรวจจับคำและหัวข้อเกิดใหม่ของงานวิจัยในปัจจุบันโดยสรุปสามารถแบ่งได้เป็น 8 วิธี ดังนี้

1. การแสดงผลลัพธ์ในการตรวจจับหัวข้อเกิดใหม่ [1,2,3,6,7]
2. การแสดงค่าของคุณลักษณะที่นำเสนอในงานวิจัย [2,6,7]
3. การเปรียบเทียบผลลัพธ์ในการตรวจจับหัวข้อเกิดใหม่กับวิธีการอื่น [7,8]
4. การเปรียบเทียบเชิงบรรยายกับวิธีการอื่น [3,6,7]
5. การเปรียบเทียบกับวิธีการอื่นด้วยความเห็นของอาสาสมัคร [5]
6. การเปรียบเทียบประสิทธิภาพด้วยข้อมูลจำลองที่ถูกสร้างขึ้นด้วยกฎบางอย่าง [8]
7. การเปรียบเทียบประสิทธิภาพโดยการนิยามความแม่นยำ (Precision) และความครอบคลุมใหม่ (Recall) ใหม่ [8]
    7.1. จากนิยามของ Pr และ Rc ที่พยายามหลบเลี่ยงการตรวจจับเกิน กล่าวคือไม่มีการลดค่าประสิทธิภาพเมื่อตรวจจับคำเกิดใหม่เกิน
    7.2. การหาคำที่ปรากฏในชุดผลเฉลยจากส่วนเบี่ยงเบนมาตรฐานรายวันของจำนวนคำที่อาศัยทำให้คำเกิดใหม่ที่ปรากฏมากต่อวันปรากฏในชุดผลเฉลยยาก และคำทั่วไปที่ปรากฏน้อยมากต่อวัน มีโอกาสปรากฏในชุดผลเฉลยง่าย
    7.3. การหาคำในชุดผลเฉลยด้วยวิธีการนี้อาศัยพารามิเตอร์หลายตัว เช่น ค่า β หรือค่าความลำเอียง, จำนวนวันก่อนหน้าที่ใช้การคำนวณหาค่าเฉลี่ยและค่าความแปรเปรวน และเส้นแบ่งของค่าเบี่ยงเบนมาตรฐาน z ทำให้ประสิทธิภาพที่ตรวจจับได้เปลี่ยนไปตามพารามิเตอร์ที่กำหนด
8. การเปรียบเทียบประสิทธิภาพโดยใช้ค่า AUC (Areas Under the Curves) [9] บนข้อมูลที่ถูกสร้างตัวแปรผลเฉลย (Data labeling) ด้วยพารามิเตอร์หลายตัว
    8.1. เนื่องจากการวัดประสิทธิภาพบนข้อมูลที่ถูกสร้างตัวแปรผลเฉลย Data labeling) ด้วยพารามิเตอร์หลายตัว ทำให้ประสิทธิภาพที่ตรวจจับได้เปลี่ยนไปตามพารามิเตอร์ที่กำหนด โดยในงานวิจัยนี้ใช้ค่า δ เท่ากับ 60 และค่า w เท่ากับ 180 โดยไม่สามารถระบุเหตุผลถึงการเลือกพารามิเตอร์ดังกล่าว
    8.2. ข้อมูลทวิตเตอร์ที่ใช้ในงานวิจัยนี้มีความลำเอียง เนื่องจากใช้ข้อมูลเฉพาะก่อนหรือหลังเหตุการณ์ที่สนใจ 1 ชั่วโมงเท่านั้น กล่าวคือ ข้อความที่นำมาใช้ในงานวิจัยทั้งข้อมูลสอนและข้อมูลทดสอบ เป็นข้อมูลที่ผ่านการกรองมาแล้วเชิงเวลา ทำให้ผลลัพธ์จากการตรวจจับหัวข้อเกิดใหม่ได้ประสิทธิภาพค่อนข้างดี

บทที่ 4
การสร้างชุดข้อมูล ชุดผลเฉลย และการประเมินผล

ในบทนี้จะนำเสนอเกี่ยวกับ 1) ชุดข้อความทวิตเตอร์และชุดข้อมูล 2) การสร้างชุดผลเฉลย และ 3) มาตรวัดประสิทธิภาพ

4.1 ชุดข้อความทวิตเตอร์และชุดข้อมูล (Twitter Data and Data)

ข้อมูลที่ใช้ในงานวิจัยนี้ประกอบด้วยข้อมูล 2 ชุด ได้แก่

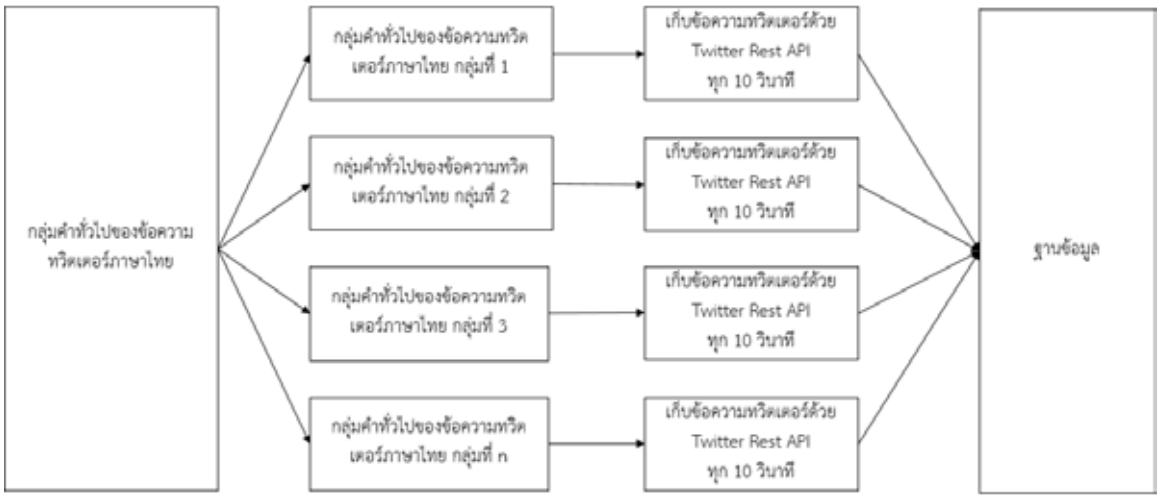
1. ชุดข้อความทวิตเตอร์ (Twitter Data) : ข้อความทวิตเตอร์ภาษาไทยที่เก็บรวบรวมด้วยทวิตเตอร์เอพีไอแบบเรส
2. ชุดข้อมูล (Data) : ชุดข้อมูลที่สร้างมาจากชุดข้อความทวิตเตอร์ โดย 1 แถวของชุดข้อมูล คือ ลักษณะของคำ 1 คำ ต่อ 1 นาที

4.1.1 ชุดข้อความทวิตเตอร์ (Twitter Data)

ชุดข้อความทวิตเตอร์ ถูกเก็บโดยการขอข้อความทวิตเตอร์ภาษาไทยผ่านทวิตเตอร์เอพีไอแบบเรสด้วยหลายโปรแกรม (Multi-thread) เพื่อให้ได้จำนวนข้อความทวิตเตอร์จำนวนมากที่สุด โดยในแต่ละโปรแกรมจะใช้กลุ่มคำทั่วไป (Stop words) ของข้อความทวิตเตอร์ภาษาไทยที่แตกต่างกันในการค้นหาข้อความทวิตเตอร์ โดยกลุ่มคำทั่วไปของข้อความทวิตเตอร์ภาษาไทยที่ใช้ในการค้นหาคือคำที่มักปรากฏบ่อยในข้อความทวิตเตอร์ภาษาไทย หรือคำที่สามารถช่วยในการค้นหาข่าวได้ดัง แสดงตัวอย่างในตารางที่ 4 ชุดข้อความทวิตเตอร์ในงานวิจัยนี้ถูกเก็บรวบรวมด้วยใช้ทั้งหมด 7 โปรแกรม และทำการขอข้อความทวิตเตอร์ภาษาไทยทุก 10 วินาทีต่อ 1 โปรแกรมดังแสดงในรูปที่ 38 ซึ่งจำนวนข้อความทวิตเตอร์ภาษาไทยในแต่ละวันที่ได้จากวิธีการนี้ถูกแสดงในตารางที่ 5

ตารางที่ 4 ตัวอย่างคำทั่วที่ใช้ในการร้องขอข้อความทวิตเตอร์ภาษาไทยผ่านทวิตเตอร์เอพีไอแบบเรส

ข่าว, นี้, นั่น, วัน, นะ, ครับ, ก็, ไทย, ค่ะ, คะ, จ้า, ไหน, งี้, เมื่อ, น่า, ใคร, ฮ่า, ท่า, ๆ, คือ, ค่า, คับ, ใน, จะ, ไม่, ใช่, ไร, ทำ, พี่, ถ้า, แต่, ไป, แล้ว, มี, 555, 55+, เฮ, มา, ทำ, จ๊ะ, มัน, นี่, อ่ะ, อะ, คุณ, เอา, ตอน, แค่, เธอ, เทอ, เขา, เรา, รูป, แบบ, บอก, หรอ, บ้าง, นึง, ซึ่ง, จัง, ฮือ, มาก, เป็น, พวก, อยาก, อย่า, เลย, ออ, ว่า, แหละ, ละ, ล่ะ, ยิ่ง, เมื่อ, หรอ, หรอก, ที่, ป่ะ, ปะ, มัน, นี่, นี้, บอก, คะ, ค่ะ, จะ, มัน, พี่, นี่, บอก, ค่ะ, คะ, คุณ, แค่, อะ, อ่ะ, ใคร, ครับ, ชอบ, คับ, เอา, นะ คะ, นะ ค่ะ, ชอบ, พูด, ตอน, น่ารัก, ไหน, ได้, งั้น, ง่ะ, อือ



รูปที่ 38 แสดงกระบวนการเก็บข้อความทวิตเตอร์ภาษาไทยด้วยหลายโปรแกรม

ตารางที่ 5 สถิติข้อความทวิตเตอร์ภาษาไทยในแต่ละวัน

วัน	จำนวนข้อความทวิตเตอร์
2018-06-12	664,204
2018-06-13	629,912
2018-06-14	568,471
2018-06-15	510,773
2018-06-16	536,679
2018-06-17	504,792
2018-06-18	504,762
2018-06-19	473,431
2018-06-20	459,288
รวม	4,852,312

4.1.2 ชุดข้อมูล (Data)

ชุดข้อมูล คือ ลักษณะของคำแต่ละคำในแต่ละนาทีของข้อความทวิตเตอร์ โดยการนำข้อความทวิตเตอร์ไปผ่านกระบวนการประมวลผลก่อน (Pre-Processing) ในบทที่ 5.1 และการสร้างข้อมูล (Data Construction) ในบทที่ 5.2 โดย 1 แถวของข้อมูล คือลักษณะของคำ 1 คำ ต่อ 1 นาที ดังนั้นคำ 1 คำซึ่งสามารถปรากฏซ้ำได้หลายช่วงเวลา จึงปรากฏในชุดข้อมูลหลายแถว จากตารางที่ 6 คำว่า “dafbama2018got7” ปรากฏจำนวน 13 คำในช่วงเวลา 2018-6-18 00:00:00 ถึง 00:00:59 ซึ่งสถิติจำนวนข้อมูลทวิตเตอร์ภาษาไทยในแต่ละวันถูกแสดงในตารางที่ 7 และสถิติจำนวนคำในข้อความทวิตเตอร์ในแต่ละวันถูกแสดงใน

ตารางที่ 8 โดยจำนวนคำที่แสดงในตารางเป็นจำนวนคำที่ผ่านกระบวนการทำความสะอาดและกำจัดคำไม่สำคัญแล้ว

ตารางที่ 6 ตัวอย่างชุดข้อมูล

เวลา	คำ	count	v_5	v_10	v_15	v_30	v_60
2018-6-18 00:00:00	dafbama2018got7	13	2.600	1.3	0.867	0.433	0.217
2018-6-18 00:01:00	dafbama2018got7	18	5.729	2.976	2.011	1.019	0.513
2018-6-18 00:02:00	dafbama2018got7	6	5.890	3.293	2.281	1.186	0.604
2018-6-18 00:03:00	dafbama2018got7	16	8.026	4.580	3.201	1.680	0.861

ตารางที่ 7 สถิติชุดข้อมูลทวิตเตอร์ภาษาไทยในแต่ละวัน (เฉพาะแถวที่มีคุณลักษณะ “จำนวนคำ” มากกว่า 2 คำต่อ 1 นาที)

วัน	จำนวนข้อมูล	จำนวนคำศัพท์
2018-06-12	320,347	6,406
2018-06-13	305,179	6,270
2018-06-14	271,967	5,810
2018-06-15	245,580	5,572
2018-06-16	264,224	5,666
2018-06-17	241,638	5,542
2018-06-18	239,134	5,535
2018-06-19	230,051	5,426
2018-06-20	221,858	5,218
รวม	2,339,978	

ตารางที่ 8 สถิติจำนวนคำเฉลี่ยในข้อความทวิตเตอร์

วัน	จำนวนคำเฉลี่ยในข้อความทวิตเตอร์
2018-06-12	6.609
2018-06-13	6.643
2018-06-14	6.552
2018-06-15	6.632
2018-06-16	6.667
2018-06-17	6.656
2018-06-18	6.589
2018-06-19	6.792
2018-06-20	6.748

4.2 การสร้างชุดผลเฉลย (Answer Set Construction)

โครงสร้างของชุดผลเฉลยของงานวิจัยนี้ประกอบด้วย “วัน”, “หัวข้อหรือหัวข้อเกิดใหม่” และ “กลุ่มของคำที่เกี่ยวข้องหรือกลุ่มของคำเกิดใหม่” ซึ่งกระบวนการสร้างชุดผลเฉลย แบ่งออกเป็น 3 ขั้นตอน ได้แก่

1. การสร้างคำในชุดผลเฉลย (Answer Keyword Construction) : ขั้นตอนการหาค้นหาคำที่เป็นคำเกิดใหม่
2. การสร้างหัวข้อในชุดผลเฉลย (Answer Topic Construction) : ขั้นตอนการสร้างหัวข้อเกิดใหม่จากการจับกลุ่มคำเกิดใหม่ที่เกี่ยวข้องกันเป็นหัวข้อเดียวกัน
3. การเพิ่มคำในแต่ละหัวข้อของชุดผลเฉลย (Answer Keyword Expansion) : ขั้นตอนการเพิ่มคำที่เกี่ยวข้องในแต่ละหัวข้อเกิดใหม่ เพื่อเติมเต็มหัวข้อเกิดใหม่ให้สมบูรณ์ยิ่งขึ้น

4.2.1 การสร้างคำในชุดผลเฉลย (Answer Keyword Construction)

1. การสร้างคำในชุดผลเฉลยขึ้นมาจากทวิตเตอร์เทรนด์

สร้างคำในชุดผลเฉลยจากทวิตเตอร์เทรนด์ (Twitter trend) ในช่วงวันที่ 12 มิถุนายน ค.ศ. 2018 ถึง 20 มิถุนายน ค.ศ. 2018 ซึ่งคำที่ได้จากทวิตเตอร์เทรนด์มักเป็นคำที่มาจากแฮชแท็ก ดังแสดงตัวอย่างของคำที่จะนำมาเป็นชุดผลเฉลยในตารางที่ 9

ตารางที่ 9 ตัวอย่างคำที่นำมาเป็นชุดผลเฉลยที่ได้จากทวิตเตอร์เทรนด์

วันที่	ทวิตเตอร์เทรนด์	คำที่ถูกสร้างเป็นชุดผลเฉลย
2018-6-20	#เมีย2018	เมีย2018
	#AWCBootCamp2018	awcbootcamp2018
	#AntManandTheWasp	antmanandthewasp
	#itcitybacon	itcitybacon
	#ลิขิตรักTheCrownPrincess	ลิขิตรักthecrownprincess
	#ICanSeeYourVoice	icanseeyourvoice

2. สร้างคำในชุดผลเฉลยจากผลลัพธ์ของงานวิจัยปัจจุบัน

สร้างคำในชุดผลเฉลยจากผลลัพธ์ของงานวิจัยปัจจุบัน โดยการสร้างคุณลักษณะจากงานวิจัยที่มีในปัจจุบัน และเลือกคำที่แตกต่างกัน 100 คำแรกที่โดดเด่นมากที่สุดจากแต่ละคุณลักษณะมาสร้างผลเฉลยเพิ่มเติม โดยผ่านการตรวจสอบและคัดกรองจากคน ดังแสดงในตารางที่ 10

ตารางที่ 10 ตัวอย่างคำที่นำมาเป็นชุดผลเฉลย ที่โดดเด่นมากที่สุดจากคุณลักษณะความเร่งของ [7]

วันที่	คำเกิดใหม่จากงานวิจัยอื่น	ค่าของคุณลักษณะ	การตรวจสอบจากคน
2018-6-20	หนึ่งด้าวฟ้าเดียว	2.935	**ผ่าน**
	ละคร	2.280	**ผ่าน**
	เรื่อง	1.734	ไม่ผ่าน
	ตัว	1.162	ไม่ผ่าน
	ปล้ำ	1.121	**ผ่าน**
	เจ้า	1.065	ไม่ผ่าน
	ขอบคุณ	0.960	ไม่ผ่าน
	รัก	0.907	ไม่ผ่าน
	aisnextgxpeckbambam	0.882	**ผ่าน**

4.2.2 การสร้างหัวข้อในชุดผลเฉลย (Answer Topic Construction)

1. การจับกลุ่มคำในชุดผลเฉลยที่เกี่ยวข้องเป็นหัวข้อเกิดใหม่

การจับกลุ่มคำในชุดผลเฉลยที่เกี่ยวข้องเป็นหัวข้อเกิดใหม่โดยจับกลุ่มคำในชุดผลเฉลยที่ปรากฏในวันเดียวกันและปรากฏร่วมกันมากในข้อความทวิตเตอร์ โดยใช้มาตรวัด 3 ตัวในการพิจารณา จากนั้นตรวจสอบและคัดกรองจากคน

1) อัตราส่วนของ จำนวนทวิตเตอร์ที่คำทั้ง 2 คำปรากฏร่วม ต่อ จำนวนทวิตเตอร์ที่ปรากฏคำใดคำหนึ่ง
2) อัตราส่วนของ จำนวนทวิตเตอร์ที่คำทั้ง 2 คำปรากฏร่วม ต่อ จำนวนทวิตเตอร์ที่คำแรกปรากฏ
3) อัตราส่วนของ จำนวนทวิตเตอร์ที่คำทั้ง 2 คำ ปรากฏร่วม ต่อ จำนวนทวิตเตอร์ที่คำสองปรากฏ

จากตารางที่ 11 แสดงตัวอย่างของความสัมพันธ์ระหว่างคู่ของคำ จะพบว่าสามารถจับกลุ่มของคำที่เกี่ยวข้องได้เป็น 1 กลุ่ม โดยประกอบด้วยคำ “หนึ่งด้าวฟ้าเดียว”, “ละคร”, “ปล้ำ” และ “aisnextgxpeckbambam” แต่เมื่อผ่านการตรวจสอบและคัดกรองจากคนจะพบว่า “ละคร” กับ “aisnextgxpeckbambam” ไม่ได้เป็นหัวข้อเดียวกันจริง จากตัวอย่างข้อความที่ปรากฏร่วมของคำสองคำนี้ในตารางที่ 12 แสดงให้เห็นว่าข้อความที่มีคำสองคำนี้ปรากฏพร้อมกันนั้นเป็นกิจกรรมของผู้ใช้งานทวิตเตอร์บางกลุ่มที่ช่วยกันปั่นแท็ก “aisnextgxpeckbambam” ในช่วงก่อนละครฉาย

ดังนั้น การจับกลุ่มของคำที่เกี่ยวข้องในตัวอย่างข้างต้น สามารถจับกลุ่มได้ทั้งหมด 2 กลุ่ม โดยกลุ่มแรกประกอบด้วยคำ 3 คำ ได้แก่ “หนึ่งด้าวฟ้าเดียว”, “ละคร”, “ปล้ำ” และกลุ่มสองประกอบด้วยคำ 1 คำ ได้แก่ “aisnextgxpeckbambam” ดังแสดงในตารางที่ 13

ตารางที่ 11 ตัวอย่างความสัมพันธ์ของคำในชุดผลเฉลยที่ปรากฏในวันเดียวกัน

วันที่	คำเกิดใหม่	คำเกิดใหม่	มาตรวัดที่ 1	มาตรวัดที่ 2	มาตรวัดที่ 3
2018-6-20	หนึ่งด้าวฟ้าเดียว	ละคร	0.797	0.806	0.986
	หนึ่งด้าวฟ้าเดียว	ปล้ำ	0.199	0.199	1.0
	หนึ่งด้าวฟ้าเดียว	aisnextgxpeckbambam	0.001	0.001	0.06
	ละคร	ปล้ำ	0.004	0.006	0.023
	ละคร	aisnextgxpeckbambam	0.015	0.015	1.0

ตารางที่ 12 ตัวอย่างคำที่ปรากฏร่วมกันมากแต่ไม่ใช่หัวข้อเดียวกัน

วันที่	คำเกิดใหม่จากงานวิจัยอื่น	คำเกิดใหม่จากงานวิจัยอื่น	ตัวอย่างข้อความ
2018-6-20	ละคร	aisnextgxpeck bambam	รีบช่วยกันบูมแท็กเดียวก่อน**ละคร**จะมา เร็วนุช! เร็ว! 💨💨💨😂 #**AISNEXTGxPECKBAMBAM**
			@FqJ7y2uZAeFXBbr แทคเดียวก่อน**ละคร**มา #**AISNEXTGxPECKBAMBAM**
			@PALITLOV รับปั่นหนี**ละคร**เลย มาแน่นอน 555 #**AISNEXTGxPECKBAMBAM**

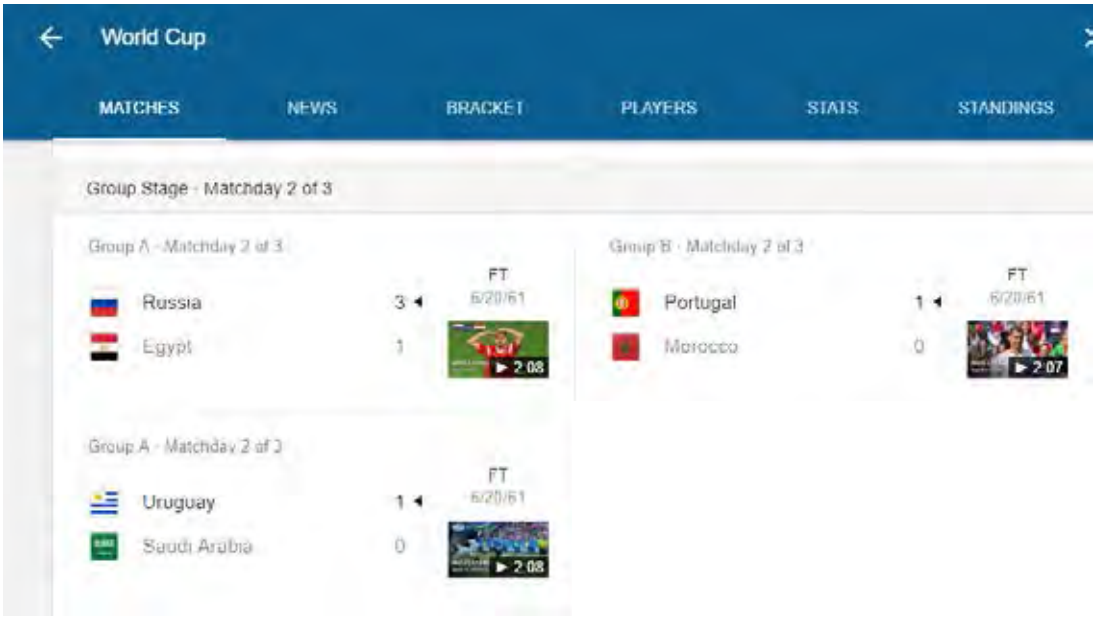
ตารางที่ 13 ตัวอย่างผลลัพธ์จากการจับกลุ่มคำที่เกี่ยวข้องที่ผ่านการตรวจสอบและคัดกรองด้วยคน

วันที่	หัวข้อของชุดผลเฉลย	คำที่เกี่ยวข้อง
2018-6-20	หนึ่งด้าวฟ้าเดียว	หนึ่งด้าวฟ้าเดียว, ละคร, ปล้ำ
	aisnextgxpeckbambam	aisnextgxpeckbambam

4.2.3 การเพิ่มคำในแต่ละหัวข้อของชุดผลเฉลย (Answer Keyword Expansion)

1. การเพิ่มคำในชุดผลเฉลยจากเว็บไซต์

การเพิ่มคำในชุดผลเฉลย โดยการนำคำในชุดผลเฉลยของแต่ละหัวข้อไปค้นหาในเว็บไซต์เพื่อหาคำที่เกี่ยวข้องมาเพิ่มในชุดผลเฉลย เนื่องจากหัวข้อเกิดใหม่หลายหัวข้อเป็นเรื่องราวที่ได้รับความสนใจจากคนส่วนมาก และถูกสื่อนำเสนอในรูปของข่าว เช่น ข่าวการแข่งขันฟุตบอลโลก สามารถหาชื่อทีมที่แข่งในแต่ละวันมาเพิ่มในชุดเฉลยได้ ดังแสดงในรูปที่ 39รูปที่ 39 ภาพจากข่าวของการแข่งขันฟุตบอลโลก ประจำวันที่ 20 มิถุนายน ค.ศ. 2018 และได้คำที่เพิ่มเข้ามาในผลเฉลยดังแสดงในตารางที่ 14



รูปที่ 39 ภาพจากข่าวของการแข่งขันฟุตบอลโลก ประจำวันที่ 20 มิถุนายน ค.ศ. 2018

ตารางที่ 14 ตัวอย่างคำที่เพิ่มเข้าในชุดผลเฉลยของหัวข้อของชุดผลเฉลย “ฟุตบอลโลก”

วันที่	หัวข้อของชุดผลเฉลย	คำที่เพิ่มเข้ามาในคำที่เกี่ยวข้อง
2018-6-20	ฟุตบอลโลก	russia, รัสเซีย, egypt, อียิป, อียิปต์, rusegy
		portugal, โปรตุเกส, morocco, โมร็อกโก, โมร็อคโก, pormor
		uruguay, อุรุกวัย, saudiarabia, ซาอุดิอาระเบีย, ซาอุ, urusau

2. การเพิ่มคำในชุดผลเฉลยจากข้อความทวิตเตอร์

การเพิ่มคำในชุดผลเฉลยให้แต่ละหัวข้อเกิดใหม่ในชุดผลเฉลย โดยการนำคำที่เกี่ยวข้องของแต่ละหัวข้อของชุดผลเฉลยไปค้นหาในข้อความทวิตเตอร์ เพื่อหาคำปรากฏร่วมมากมาเพิ่มในแต่ละหัวข้อในชุดผลเฉลย โดยทำการคัดเลือกข้อความทวิตเตอร์ที่มีคำในชุดผลเฉลยปรากฏอย่างน้อย 1 คำ และนำข้อความทวิตเตอร์ดังกล่าวมาหาคำที่ปรากฏมากที่สุด 10 คำ หากกลุ่มคำที่พบมีบางคำไม่ปรากฏในชุดผลเฉลย จะถูกเพิ่มเข้าไปในคำที่เกี่ยวข้องของชุดผลเฉลยของหัวข้อนั้น จากนั้นทำการหาข้อความทวิตเตอร์ที่เกี่ยวข้องและคำที่ปรากฏมากที่สุดซ้ำ จนไม่พบคำใหม่ที่จะสามารถนำไปเพิ่มในคำที่เกี่ยวข้องของชุดผลเฉลยของหัวข้อนั้นได้ จากนั้นทำการตรวจสอบและคัดกรองจากคนเป็นอันเสร็จสิ้นกระบวนการสร้างผลเฉลย ซึ่งผลลัพธ์และจำนวนของชุดผลเฉลยถูกแสดงในตารางที่ 15 และตารางที่ 16

ตารางที่ 15 ตัวอย่างชุดผลเฉลยสุดท้าย

วันที่	หัวข้อของชุดผลเฉลย	กลุ่มของคำที่เกี่ยวข้อง
2018-6-20	ฟุตบอลโลก	เล่น, ยิง, fifaworldcup2018, mancity, worldcup, ฟุตบอลโลก 2018, ฟุตบอลต่างประเทศ, egypt, morocco, pormor, portugal, rusegy, โปรตุเกส, ฟุตบอลโลก, russia, saudiarabia, uruguay, urusau, โมร็อกโก, โมร็อคโก, โลก, ซาอุ, ซาอุดิอาระเบีย, บอล, บอลโลก, บอลโลก2018, รัสเซีย, สวัสดี, สวัสดีบอล, สวัสดีบอลโลก, สวัสดีบอลโลก2018, อียิป, อียิปต์, อุรุกวัย
	aisnextgxpeckbambam	เป๊กแบม, peckpalitchoke, ais, supol, peckpalit, peckpalitbell, bellsupol, aisnextgxpeckbambam, เป๊กผลิตโชค, เป๊ก, bambam, ผลิตโชค

ตารางที่ 16 สถิติจำนวนหัวข้อและจำนวนคำของชุดผลเฉลยในแต่ละวัน

เวลา	จำนวนหัวข้อ	จำนวนคำ
2018-06-12	23	176
2018-06-13	20	146
2018-06-14	18	139
2018-06-15	26	171
2018-06-16	22	167
2018-06-17	26	216
2018-06-18	27	238
2018-06-19	21	207
2018-06-20	18	169
TOTAL	201	1,629

4.3 มาตรวัดประสิทธิภาพ (Measurement)

ในงานวิจัยนี้ ผู้วิจัยนำเสนอมาตรวัดประสิทธิภาพแบบใหม่ที่สามารถวัดประสิทธิภาพของการตรวจจับคำเกิดใหม่, ประสิทธิภาพของการตรวจจับหัวข้อเกิดใหม่ และ ประสิทธิภาพการตรวจจับคำและหัวข้อเกิดใหม่พร้อมกัน โดยมีรายละเอียดดังนี้

1. มาตรวัดประสิทธิภาพในมุมมองของคำ (Keyword Measurement) : เป็นการวัดประสิทธิภาพของจำนวนคำเกิดใหม่ที่ตรวจจับได้
2. มาตรวัดประสิทธิภาพในมุมมองของหัวข้อ (Topic Measurement) : เป็นการวัดประสิทธิภาพของจำนวนหัวข้อเกิดใหม่ที่ตรวจจับได้
3. มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโคร (Macro Topic Measurement) : เป็นการวัดประสิทธิภาพของคำและหัวข้อเกิดใหม่ที่ตรวจจับได้พร้อมกัน

4.3.1 มาตรวัดประสิทธิภาพในมุมมองของคำ (Keyword Measurement)

เป็นมาตรวัดประสิทธิภาพที่มองในมุมมองของ**จำนวนคำ**ที่ตรวจจับได้ โดยมีการนิยาม TP, FP, และ FN ใหม่ ดังต่อไปนี้

TP คือ จำนวนของคำที่แตกต่างกันที่ถูกทำนายว่าเป็นคำเกิดใหม่ และอยู่ในชุดผลเฉลย หรือคำที่ทำนายถูกต้อง

FP คือ จำนวนของคำที่แตกต่างกันที่ถูกทำนายว่าเป็นคำเกิดใหม่ แต่**ไม่**อยู่ในชุดผลเฉลย หรือคำที่ถูกทำนายเกิน

FN คือ จำนวนของคำที่แตกต่างกันที่**ไม่**ถูกทำนายว่าเป็นคำเกิดใหม่ แต่อยู่ในชุดผลเฉลย หรือคำที่ถูกทำนายขาด

จากตัวอย่างคำและหัวข้อเกิดใหม่ที่ถูกทำนายว่าเป็นคำเกิดใหม่ในตารางที่ 17 กับตัวอย่างคำและหัวข้อเกิดใหม่ของชุดผลเฉลยในตารางที่ 18 พบว่า

1. คำที่เป็น TP หรือ คำที่ถูกทำนายว่าเป็นคำเกิดใหม่ และอยู่ในชุดผลเฉลย มีจำนวน 3 คำ ได้แก่ “aisnextgxpeckbambam”, “เป๊กผลิตโชค” และ “peckpalitchoke”
2. คำที่เป็น FP หรือ คำที่ถูกทำนายว่าเป็นคำเกิดใหม่ แต่**ไม่**อยู่ในชุดผลเฉลย มีจำนวน 1 คำ ได้แก่ “ฝันดีจ้า”

3. คำที่เป็น FN หรือ คำที่**ไม่**ถูกทำนายว่าเป็นคำเกิดใหม่ แต่อยู่ในชุดผลเฉลย มีจำนวน 9 คำ ได้แก่ “ผลิตโชค”, “เป๊กแบม”, “bellsupol”, “peckpalit”, “supol”, “peckpalitbell”, “ais”, “bambam”, และ “เป๊ก”

จากสมการที่ (22), (23) และ (24) ในบทที่ 2.7 จะได้ค่าของ Pr, Rc, และ F_1 ของมาตรวัดประสิทธิภาพในมุมมองของคำ ดังนี้

$$Pr = \frac{3}{3+1} = 0.75$$

$$Rc = \frac{3}{3+9} = 0.25$$

$$F_1 = \frac{2 * 0.75 * 0.25}{0.75 + 0.25} = 0.375$$

ตารางที่ 17 ตัวอย่างคำและหัวข้อเกิดใหม่ที่ถูกทำนายว่าเป็นคำเกิดใหม่

หัวข้อ	กลุ่มของคำที่ถูกทำนายว่าเป็นคำเกิดใหม่	หัวข้อในผลเฉลยที่คล้ายมากที่สุด
หัวข้อที่ 1	aisnextgxpeckbambam, เป๊กผลิตโชค, peckpalitchoke, ฝันดีจ้า	aisnextgxpeckbambam
หัวข้อที่ 2	aisnextgxpeckbambam, เป๊กผลิตโชค	aisnextgxpeckbambam
หัวข้อที่ 3	อร่อย	-

ตารางที่ 18 ตัวอย่างคำและหัวข้อเกิดใหม่ของชุดผลเฉลย

หัวข้อของชุดผลเฉลย	กลุ่มของคำที่เกี่ยวข้อง
aisnextgxpeckbambam	ผลิตโชค, เป๊กแบม, bellsupol, **เป๊กผลิตโชค**, **aisnextgxpeckbambam**, peckpalit, **peckpalitchoke**, supol, peckpalitbell, bambam, เป๊ก, ais
smtmthailand	เป้, ด่า, ทีม, วดฟ, smtmthailand, ดิส, คนดู

4.3.2 มาตรวัดประสิทธิภาพในมุมมองของหัวข้อ (Topic Measurement)

เป็นมาตรวัดประสิทธิภาพที่มองในมุมมองของ**จำนวนหัวข้อ**ที่ตรวจจับได้ โดยมีการนิยาม TP, FP, และ FN ใหม่ ดังต่อไปนี้

TP คือ จำนวนของหัวข้อของชุดผลเฉลยที่แตกต่างกันที่ถูกทำนายว่าเป็นหัวข้อเกิดใหม่ หรือหัวข้อที่ทำนายถูกต้อง

FP คือ จำนวนของหัวข้อที่แตกต่างกันที่ถูกทำนายว่าเป็นหัวข้อเกิดใหม่ แต่**ไม่**อยู่ในชุดผลเฉลย หรือหัวข้อที่ถูกทำนายเกิน

FN คือ จำนวนของหัวข้อของชุดผลเฉลยที่แตกต่างกันที่**ไม่**ถูกทำนายว่าเป็นหัวข้อเกิดใหม่ หรือหัวข้อที่ถูกทำนายขาด

จากตัวอย่างคำและหัวข้อเกิดใหม่ที่ถูกทำนายว่าเป็นคำเกิดใหม่ในตารางที่ 17 กับตัวอย่างคำและหัวข้อเกิดใหม่ของชุดผลเฉลยในตารางที่ 18 พบว่า

1. หัวข้อที่เป็น TP หรือ หัวข้อที่อยู่ในชุดผลเฉลยและถูกทำนายว่าเป็นหัวข้อเกิดใหม่ มีจำนวน 1 หัวข้อ ได้แก่ หัวข้อของ”aisnextgxpeckbambam”
2. หัวข้อที่เป็น FP หรือ หัวข้อที่ถูกทำนายว่าเป็นหัวข้อเกิดใหม่แต่ไม่อยู่ในชุดผลเฉลย มีจำนวน 1 หัวข้อ ได้แก่ หัวข้อที่ 3 ในหัวข้อที่ถูกทำนาย
3. หัวข้อที่เป็น FN หรือ หัวข้อที่อยู่ในชุดผลเฉลยแต่ไม่ถูกทำนายว่าเป็นหัวข้อเกิดใหม่ มีจำนวน 1 หัวข้อ ได้แก่ หัวข้อของ”smtmthailand”

จากสมการที่ (22), (23) และ (24) ในบทที่ 2.7 จะได้ค่าของ Pr, Rc, และ F_1 ของมาตรวัดประสิทธิภาพในมุมมองของหัวข้อ ดังนี้

$$Pr = \frac{1}{1+1} = 0.5$$

$$Rc = \frac{1}{1+1} = 0.5$$

$$F_1 = \frac{2 * 0.5 * 0.5}{0.5 + 0.5} = 0.5$$

4.3.3 มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโคร (Macro Topic Measurement)

เนื่องจากข้อจำกัดของมาตรวัดประสิทธิภาพในมุมมองของคำที่สนใจเพียงคำที่แตกต่างกัน โดยไม่สนใจว่าคำที่ตรวจจับมาได้นั้นอยู่ในหัวข้อที่ถูกต้องหรือไม่ และข้อจำกัดของมาตรวัดประสิทธิภาพในมุมมองของหัวข้อที่ไม่สนใจจำนวนคำที่เกี่ยวข้องในหัวข้อ มีอย่างน้อย 1 คำที่เกี่ยวข้องกับหัวข้อ ก็นับว่าหัวข้อนั้นตรวจจับได้ มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครจึงถูกนำเสนอขึ้นมาเพื่อเป็นมาตรวัดที่สามารถวัดประสิทธิภาพในมุมมองของคำและหัวข้อที่ได้พร้อมกัน โดยพัฒนามาจากการวัดประสิทธิภาพการทำนายสำหรับงานจำแนกประเภทที่มีหลายคลาส (Multiclass Classification Performance Evaluation) แบบแมโคร (Macro) ในบทที่ 2.8 ซึ่งเป็นการวัดประสิทธิภาพที่ให้ความสำคัญกับทุกคลาสเท่ากัน โดยไม่สนใจจำนวนสมาชิกในแต่ละคลาส

กำหนดให้หัวข้อของชุดผลเฉลย 1 หัวข้อ คือ 1 คลาส จากนั้นทำการคำนวณหาประสิทธิภาพ Pr, Rc ในมุมมองของคำในแต่ละคลาส และทำการคำนวณหาค่าเฉลี่ยของประสิทธิภาพ Pr, Rc ของทุกคลาสเพื่อนำมาคำนวณหาประสิทธิภาพ F_1 เฉลี่ยของทุกคลาส แต่ด้วยการคำนวณที่กำหนดให้หัวข้อของชุดผลเฉลย 1 หัวข้อ คือ 1 คลาส ทำให้หัวข้อที่ถูกทำนายเกินนั้นไม่ถูกนำมาคำนวณ ดังนั้น 1 หัวข้อที่ถูกทำนายเกินจะถูกเพิ่มเข้าไปเป็น 1 คลาส และมีประสิทธิภาพในมุมมองของคำ Pr, Rc เป็น 0 ดังแสดงในตารางที่ 19

ตารางที่ 19 ตัวอย่างการคำนวณหาประสิทธิภาพในมุมมองของคำของแต่ละหัวข้อของชุดผลเฉลยและหัวข้อที่ถูกทำนายเกิน

หัวข้อของชุดผลเฉลย	$\boldsymbol{Pr}$	$\boldsymbol{Rc}$
aisnextgxpeckbambam	$Pr1$	$Rc1$
smtmthailand	$Pr2$	$Rc2$
หัวข้อที่ถูกทำนายเกิน	$Pr3 = 0$	$Rc3 = 0$

จากนั้นนำไปคำนวณหาประสิทธิภาพในมุมมองของคำโดยเฉลี่ยทุกหัวข้อเกิดใหม่ หรือประสิทธิภาพในมุมมองของหัวข้อแบบแมโคร (Macro topic) ได้ดังนี้

$$Pr = \frac{\sum Pr}{n} = \frac{Pr1 + Pr2 + Pr3}{3}$$
$$Rc = \frac{\sum Rc}{n} = \frac{Rc1 + Rc2 + Rc3}{3}$$
$$F_1 = \frac{2 * Pr * Rc}{Pr + Rc}$$

แต่บางครั้งหัวข้อของชุดผลเฉลย 1 หัวข้อ ถูกแสดงด้วยหัวข้อที่ถูกทำนายได้มากกว่า 1 หัวข้อ ทำให้เกิดปัญหาในการหาค่าประสิทธิภาพในมุมมองของคำของหัวข้อชุดผลเฉลยดังกล่าว ดังตารางที่ 20 ที่หัวข้อของชุดผลเฉลย “aisnextgxpeckbambam” ถูกแสดงด้วยหัวข้อที่ถูกทำนายมากกว่า 1 หัวข้อ ซึ่งผู้วิจัยได้นำเสนอวิธีการจัดการปัญหานี้ทั้งหมด 2 วิธี ได้แก่ 1) มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro average หรือ MA) และ 2) มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม (Macro merge หรือ MM)

1. มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro Average หรือ MA)

เป็นมาตรวัดประสิทธิภาพที่จะทำการหาค่า Pr และ Rc ของหัวข้อของชุดผลเฉลยที่ถูกแสดงด้วยหัวข้อที่ถูกทำนายมากกว่า 1 หัวข้อโดยใช้ค่าเฉลี่ยของ Pr และ Rc ของหัวข้อที่ถูกทำนาย

จากตารางที่ 20 หัวข้อที่ 1 และหัวข้อที่ 2 นั้นแสดงถึงหัวข้อในชุดผลเฉลยเดียวกัน ดังนั้นประสิทธิภาพในมุมมองของคำของหัวข้อของชุดผลเฉลย “aisnextgxpeckbambam” จะถูกแสดงด้วย ค่าเฉลี่ย Pr และ Rc ของประสิทธิภาพในมุมมองของคำของหัวข้อที่ 1 และหัวข้อที่ 2 ดังแสดงในตารางที่ 21

ตารางที่ 20 ตัวอย่างการคำนวณหาประสิทธิภาพในมุมมองคำของแต่ละหัวข้อที่ถูกทำนายและหัวข้อที่ทำนายขาด

หัวข้อ	**กลุ่มของคำที่ถูกทำนายว่าเป็นคำเกิดใหม่**	**หัวข้อในผลเฉลยที่คล้ายมากที่สุด**	TP	FP	FN	Pr	Rc	F_1
หัวข้อที่ 1	aisnextgxpeckbambam, เป๊กผลิตโชค, peckpalitchoke, ฝันดีจ้า	aisnextgxpeck bambam	3	1	9	0.75	0.25	0.38
หัวข้อที่ 2	aisnextgxpeckbambam, เป๊กผลิตโชค	aisnextgxpeck bambam	2	0	10	1	0.17	0.29
หัวข้อที่ 3	อร่อย	-	0	1	0	0	0	0
-	-	smtmthailand	0	0	6	0	0	0

ตารางที่ 21 ตัวอย่างการคำนวณหาประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย

หัวข้อ	**กลุ่มของคำที่ถูกทำนายว่าเป็นคำเกิดใหม่**	**หัวข้อในผลเฉลยที่คล้ายมากที่สุด**	Pr	Rc
หัวข้อที่ 1	aisnextgxpeckbambam, เป๊กผลิตโชค, peckpalitchoke, ฝันดีจ้า	aisnextgxpeck bambam	$\frac{0.75+1}{2}$	$\frac{0.25+0.17}{2}$
หัวข้อที่ 2	aisnextgxpeckbambam, เป๊กผลิตโชค	aisnextgxpeck bambam		
หัวข้อที่ 3	อร่อย	-	0	0
-	-	smtmthailand	0	0

2. มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม (Macro Merge หรือ MM)

เป็นมาตรวัดประสิทธิภาพที่จะทำการหาค่า Pr และ Rc ของหัวข้อของชุดผลเฉลยที่ถูกแสดงด้วยหัวข้อที่ถูกทำนายมากกว่า 1 หัวข้อ จากหัวข้อที่เกิดจากการรวมหัวข้อที่ถูกทำนาย

จากตารางที่ 20 หัวข้อที่ 1 และหัวข้อที่ 2 นั้นแสดงถึงหัวข้อในชุดผลเฉลยเดียวกัน ดังนั้นประสิทธิภาพในมุมมองของคำของหัวข้อของชุดผลเฉลย “aisnextgxpeckbambam” จะถูกแสดงด้วย ประสิทธิภาพในมุมมองของคำของหัวข้อที่เกิดจากการรวมหัวข้อที่ 1 และหัวข้อที่ 2 เข้าด้วยกัน ดังแสดงในตารางที่ 22

ตารางที่ 22 ตัวอย่างการคำนวณหาประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม

หัวข้อ	**กลุ่มของคำที่ถูกทำนายว่าเป็นคำเกิดใหม่**	**หัวข้อในผลเฉลยที่คล้ายมากที่สุด**	Pr	Rc
หัวข้อที่ 1	aisnextgxpeckbambam, เป๊กผลิตโชค, peckpalitchoke, ฝันดีจ้า	aisnextgxpeck bambam	0.75	0.25
หัวข้อที่ 2				
หัวข้อที่ 3	อร่อย	-	0	0
-	-	smtmthailand	0	0

บทที่ 5
แนวคิดและวิธีการดำเนินงาน

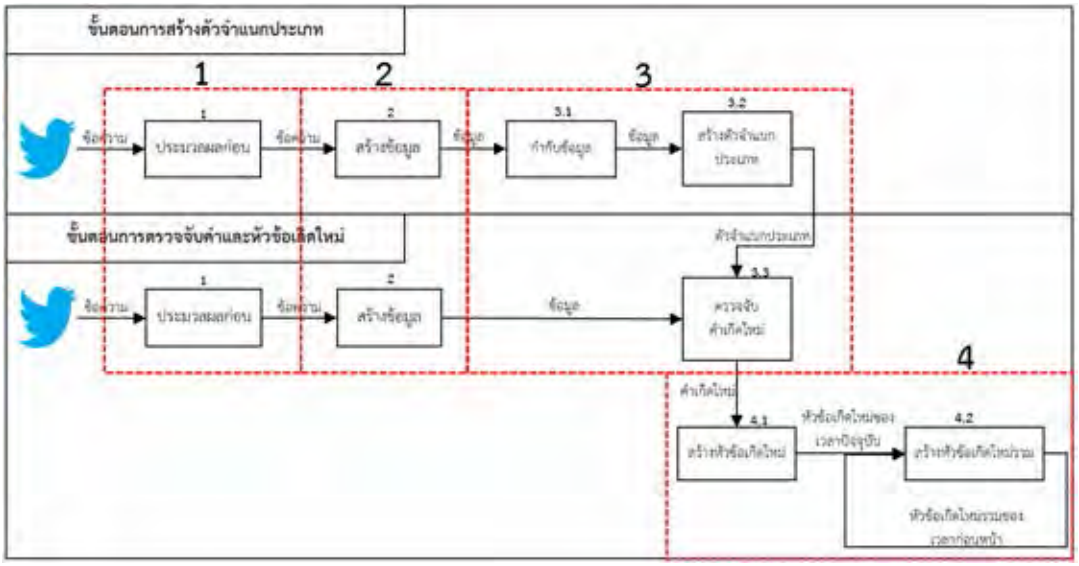
แนวคิดที่เสนอในงานวิจัย

1. การสร้างคุณลักษณะจากตัวชี้วัดของหุ้นประเภทโมเมนตัม rsi, $stoch$, wr ในบทที่ 5.2.2
   1.1. การปรับปรุงคุณลักษณะใหม่ ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก ในบทที่ 5.2.3
   1.2. การปรับปรุงคุณลักษณะใหม่ ด้วยค่าความต่างของกรอบเวลา ในบทที่ 5.2.4
   1.3. การปรับปรุงคุณลักษณะใหม่ ด้วยการชะลอการเปลี่ยนแปลง ในบทที่ 5.2.5
2. การสร้างตัวจำแนกประเภทเพื่อใช้ในการตัดสินใจในการหากลุ่มคุณลักษณะและเส้นแบ่งที่ดีที่สุด
   2.1. การกำกับข้อมูลหรือการสร้างตัวแปรผลเฉลย (Data Labeling) ที่ครอบคลุมคำในชุดผลเฉลยมากที่สุด และไม่อาศัยค่าพารามิเตอร์ ในบทที่ 5.3.1 และ 5.3.2
3. มาตรวัดประสิทธิภาพของหัวข้อแบบแมโคร (Macro Topic) ที่สามารถวัดประสิทธิภาพในแง่มุมของคำ และหัวข้อเกิดใหม่พร้อมกัน ในบทที่ 4.3.3
   3.1. สร้างชุดผลเฉลยรายวันอันประกอบด้วย “วัน”, “หัวข้อ”, “กลุ่มคำที่เกี่ยวข้อง” ในบทที่ 4.2

ขั้นตอนของการตรวจจับคำและหัวข้อเกิดใหม่

ขั้นตอนในการตรวจจับคำและหัวข้อเกิดใหม่ในรูปที่ 40 แบ่งออกเป็น 4 ขั้นตอน ได้แก่

1. การประมวลผลก่อน (Pre-processing) – ขั้นตอนการปรับปรุงข้อความทวิตเตอร์ให้พร้อมสำหรับการใช้งาน เช่น กำจัดอักขระพิเศษ, กำจัดคำไม่สำคัญ, ตัดคำ เป็นต้น
2. การสร้างข้อมูล (Data Construction) – ขั้นตอนการสร้างข้อมูลจากข้อความทวิตเตอร์ โดยการนับการปรากฏของคำแต่ละคำในแต่ละนาที จากนั้นนำไปคำนวณคุณลักษณะต่าง ๆ เช่น rsi, $stoch$, wr เป็นต้น ดังนั้น 1 แถวของข้อมูล คือ ลักษณะของคำ 1 คำ ต่อ 1 นาที
3. การตรวจจับคำเกิดใหม่ (Emerging Keyword Detection) – เป็นการนำชุดข้อมูลมากำกับข้อมูล (Data labeling) เพื่อสร้างตัวจำแนกประเภท และใช้ตัวจำแนกประเภทดังกล่าวทำนายหาคำเกิดใหม่ในแต่ละนาที
4. การตรวจจับหัวข้อเกิดใหม่ (Emerging Topic Detection) – เป็นการนำคำเกิดใหม่ที่ทำนายได้มาหาความสัมพันธ์ระหว่างคู่ของคำเกิดใหม่ เพื่อนำคำเกิดใหม่ที่เกี่ยวข้องจับกลุ่มเป็นหัวข้อเกิดใหม่ และสุดท้ายทำการรวมหัวข้อเกิดใหม่ที่เกี่ยวข้องกันในช่วงเวลาที่ใกล้เคียงกันเป็นหัวข้อเดียวกัน



รูปที่ 40 แสดงภาพรวมการทำงานของการตรวจจับคำและหัวข้อเกิดใหม่

5.1 การประมวลผลก่อน (Pre-Processing)

เป็นการปรับปรุงข้อความทวิตเตอร์ให้พร้อมใช้งาน โดยการกำจัดข้อความและคำไม่สำคัญ ซึ่งมี 2 ขั้นตอน ได้แก่

1. การทำความสะอาดข้อความบนข้อความทวิตเตอร์ (Twitter Data Cleansing)
2. การตัดคำและการกำจัดคำไม่สำคัญ (Text Tokenization and Stop Word Removal)

5.1.1 การทำความสะอาดข้อความทวิตเตอร์ (Twitter Data Cleansing)

ข้อความทวิตเตอร์มีลักษณะเป็นข้อความสั้น ๆ โดยในข้อความทวิตเตอร์บางข้อความจะมีที่อยู่ของเว็บไซต์, อักขระพิเศษ, หรือ ชื่อของผู้ใช้งานอื่นที่ถูกรีทวีต ซึ่งจำเป็นต้องนำไปทำความสะอาดก่อนนำข้อความไปใช้งาน เนื่องจากเป็นคำที่ไม่มีความหมายในตัวเอง โดยมีขั้นตอนดังนี้

1. การกำจัดชื่อผู้ใช้งานออกจากข้อความทวิตเตอร์

จากการศึกษาพบว่า ผู้ใช้งานทวิตเตอร์มีพฤติกรรมรีทวีตข้อความหรือแท็กผู้ใช้งานที่มีชื่อเสียงหรือมีอิทธิพล (Influencer) ซึ่งหากไม่กำจัดชื่อผู้ใช้งานทิ้ง ชื่อผู้ใช้งานที่มีชื่อเสียงหรือมีอิทธิพลจะปรากฏบ่อยเป็นคำเกิดใหม่ตลอดทั้งวันเป็นช่วง ๆ โดยคำดังกล่าวไม่ได้เป็นเรื่องราว หรือเหตุการณ์ที่น่าสนใจ อาทิเช่น พฤติกรรมสวัสดีตอนเช้าให้กับดารานักร้องที่ชื่นชอบ

2. การกำจัดที่อยู่ของเว็บไซต์ออกจากข้อความทวิตเตอร์

เนื่องจากที่อยู่ของเว็บไซต์บนข้อความทวิตเตอร์ไม่สามารถนำมาตัดคำเป็นคำที่มีความหมาย

3. การกำจัดอักขระพิเศษ

กำจัดอักขระพิเศษก่อนนำข้อความทวิตเตอร์ไปตัดคำ เพื่อลดจำนวนคำศัพท์ที่เป็นไปได้ ส่งผลต่อปริมาณทรัพยากรที่ต้องใช้ในการคำนวณ

4. การกำจัดตัวอักษรซ้ำ

เนื่องจากบางครั้งผู้ใช้งานทวิตเตอร์ต้องการโพสต์ข้อความที่สื่ออารมณ์มากขึ้น จึงนิยมพิมพ์อักษรตัวสุดท้ายซ้ำเพื่อเป็นการย้ำถึงอารมณ์ อาทิเช่นคำว่า “น่ารักกกกกกกกกกกกกกกกกกกก” ในตารางที่ 23 ซึ่งส่งผลทำให้ตัวตัดคำภาษาไทยตัดคำได้ไม่ดี จึงจำเป็นต้องทำการลดรูปของอักษรซ้ำให้เหลือน้อยที่สุดที่เป็นไปได้โดยไม่กระทบกับคำปกติ โดยการลดรูปคำที่มีตัวอักษรปรากฏซ้ำต่อเนื่องกันมากกว่า 3 ครั้งขึ้นไป ให้เหลือเพียง 3 ครั้ง ดังนั้นคำว่า “น่ารักกกกกกกกกกกกกกกกกกกกก” จะถูกลดรูปเหลือเพียง “น่ารักกก”

ตารางที่ 23 แสดงตัวอย่างข้อความทวิตเตอร์ที่ปรากฏอักษรซ้ำมากกว่า 3 ครั้ง

ข้อความทวิตเตอร์
อย่าว่าแต่ด้อมอื่นเอนดูวนว.เรย ทั้งด้อมและเมมเบอร์วงชั้นก้เอนดู วนว. เจ้าพวกน่า**รักกกกกก**
โฮปดื่มนิดเดียวตัวแดงนั่งจิตหลุดลอยออกนอกโลก**55555555** น่า**รักกกกกกก**
น้องเฮย์เปลี่ยนสี**ผมมมมมมมมมมมมมม หืออออออออออออออออออ** ไม่ว่าแบบไหนก็ น่ารัก**นะะะะะะะ** แต่แบบ**หือออออออ** น่า**รักกกกกกก** น่ารักทุก**แบบบบบบบ คืออออออออออ น้องงงงงงงงงง** แพ้น้องทุก**ทางงงงงงงงงง** แต่น้องจะเปลี่ยนสีผมทำ**มายยยยยยยย** เอาเถอะถ้าบอกว่าแบบไหน น่ารัก น้องทำอะไรก็น่ารักหมด**แหละะะะ**
ว้ายๆๆๆๆๆๆๆร้องผิด งื้อ น่า**รักกกกกกกก**ลิ้นพันน**แล้วววววววว** #KaewBNK48 https://t.co/LYoLKU5BCn
น่ารักก ลุค18เว่อร์ ใส ใสอ่ะ **น่ารักกกกกกกกกกกกกกกกกกกกกกกกกกกกกกกกก** น้อง**แทๆๆๆๆๆๆๆๆๆๆๆๆๆๆ น้อนนนนนนนนนนนนนนนนนนนนนนนนนนนนนนนน** https://t.co/tPxByEWUSw

5.1.2 การตัดคำและการกำจัดคำไม่สำคัญ (Text Tokenization and Stop Word Removal)

ทำการตัดคำบนข้อความทวิตเตอร์ภาษาไทยที่ผ่านการทำความสะอาดจากขั้นตอนก่อนหน้าด้วยไลบรารี่ pythainlp เวอร์ชั่น 1.7.4, deepcut เวอร์ชั่น 0.6.1.0 และ tensorflow เวอร์ชั่น 1.13.1 จากนั้นทำการตัดคำไม่สำคัญ โดยคำไม่สำคัญในงานวิจัยนี้ คือ คำที่ปรากฏเป็นประจำบนข้อความภาษาไทยบนทวิตเตอร์ ซึ่งได้จากการหาคำที่ปรากฏบ่อยในข้อความทวิตเตอร์และไม่สามารถแสดงเรื่องราวได้ ซึ่งจำเป็นต้องผ่านการตรวจสอบและคัดกรองด้วยคน ผลลัพธ์แสดงในตารางที่ 24

ตารางที่ 24 ตัวอย่างผลลัพธ์สุดท้ายของขั้นตอนการประมวลผลข้อมูลก่อน

เวลา	ข้อความ	กลุ่มคำที่ถูกประมวลผลแล้ว
2018-6-12 10:29:00	@stpd1011 @GOT7Official ถ้าเอาแท็ก dafbama2018got7 ไปรวมกับแท็กอื่น มันจะไม่นับนะ ถ้าจะเอาแท็กนี้ให้พิมพ์แค่ #GOT7 #dafBAMA2018Got7 @GOT7Official แล้วก็ใส่คำพูดอะไรไปก็ได้จ้าาา	[แท็ก, dafbama2018got, แท็ก, แท็ก, พิมพ์, แค่, แล้ว, ใส่, จ้าาา, got7, dafbama2018got7]
2018-6-12 21:32:00	เป็นติ่งมันไม่มีไรมากหรอก ไม่ค่ะ..หยุด!!! คุณต้องมีเงินที่พร้อมจะเปย์หลัวของคุณ@@ ไปละค่ะจะนอนzzzzzz #GOT7 #dafbama2018got7 @GOT7Official	[ติ่ง, หยุด, เงิน, เปย์, หลัว, นอน, got7, dafbama2018got7]
2018-6-15 23:56:00	บุญตามากกกก นานๆจะเห็นพวกเค้ามาไลฟ์ #GOT7 #dafbama2018got7	[บุญตา, มากกกก, เค้า, ไลฟ์, got7, dafbama2018got7]
2018-6-19 16:29:00	เมื่อเพื่อนถามว่าคนนี้คือใคร ความแม่ต้องอวยลูก อวยให้สุด แล้วตีหัวเพื่อนเข้าด้อม #GOT7 #Yugyoem #EyesOnYou #GOT7WORLDTOUR #dafbama2018got7 https://t.co/rbOxC5tzRj	[เพื่อน, ถาม, แม่, อวย, ลูก, อวย, หัว, เพื่อน, ด้อม, got7, yugyoem, eyesonyou,

		got7worldtour, dafbama2018got7]

5.2 การสร้างข้อมูล (Data Construction)

เป็นการนำข้อความทวิตเตอร์ไปสร้างเป็นข้อมูล โดยทำการนับการปรากฏของคำแต่ละคำในแต่ละนาที จากนั้นนำไปคำนวณคุณลักษณะต่าง ๆ เช่น $rsi, stoch, wr$ เป็นต้น ดังนั้น 1 แถวของข้อมูล คือ ลักษณะของคำ 1 คำ ต่อ 1 นาที โดยมีขั้นตอนในการสร้างข้อมูล 5 ขั้นตอน ได้แก่

1. การสร้างข้อมูลเชิงเวลา (Temporal Data Construction) : การสร้างข้อมูลของแต่ละคำในแต่ละนาที
2. การสร้างคุณลักษณะใหม่จากตัวชี้วัดของหุ้น (Stock Indicator Feature Extraction) : การสร้างคุณลักษณะ $rsi, stoch, wr$
3. การปรับปรุงคุณลักษณะใหม่จากตัวชี้วัดของหุ้น ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก v (Improving Stock Indicator Feature using Exponential Moving Average) : การสร้างคุณลักษณะตัวชี้วัดของหุ้นจากค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนักแทนจำนวนคำที่ปรากฏ ได้แก่ $rsi2, stoch2, wr2$
4. การปรับปรุงคุณลักษณะใหม่จากตัวชี้วัดของหุ้น ด้วยความต่างของกรอบเวลา w (Improving Stock Indicator Feature using Different Time Period) : การสร้างคุณลักษณะตัวชี้วัดของหุ้นโดยมีแนวคิดจาก $MACD$ ได้แก่ $rsi2\_cd, stoch\_cd, wr\_cd$
5. การปรับปรุงคุณลักษณะใหม่จากตัวชี้วัดของหุ้น ด้วยการชะลอการเปลี่ยนแปลง (Improving Stock Indicator Feature using Delaying Change) : การสร้างคุณลักษณะตัวชี้วัดของหุ้นโดยมีแนวคิดจาก $stoch(\%D)$ ได้แก่ $rsi2\_slow, stoch2\_slow, wr2\_slow$

5.2.1 การสร้างข้อมูลเชิงเวลา (Temporal Data Construction)

เป็นการนับคำในข้อความทวิตเตอร์ในแต่ละนาที โดยคำที่ปรากฏในช่วงเวลา 2018-6-18 00:00:00 ถึง 00:00:59 จะถูกกำกับว่าเกิดในช่วงเวลา 2018-6-18 00:00:00 ทั้งหมด ดังแสดงตัวอย่างในตารางที่ 25 คำว่า “dafbama2018got7” ปรากฏในช่วงเวลา 2018-6-18 00:00:00 ถึง 00:00:59 เป็นจำนวน 13 คำ

ตารางที่ 25 ตัวอย่างผลลัพธ์จากการสร้างข้อมูลเชิงเวลา

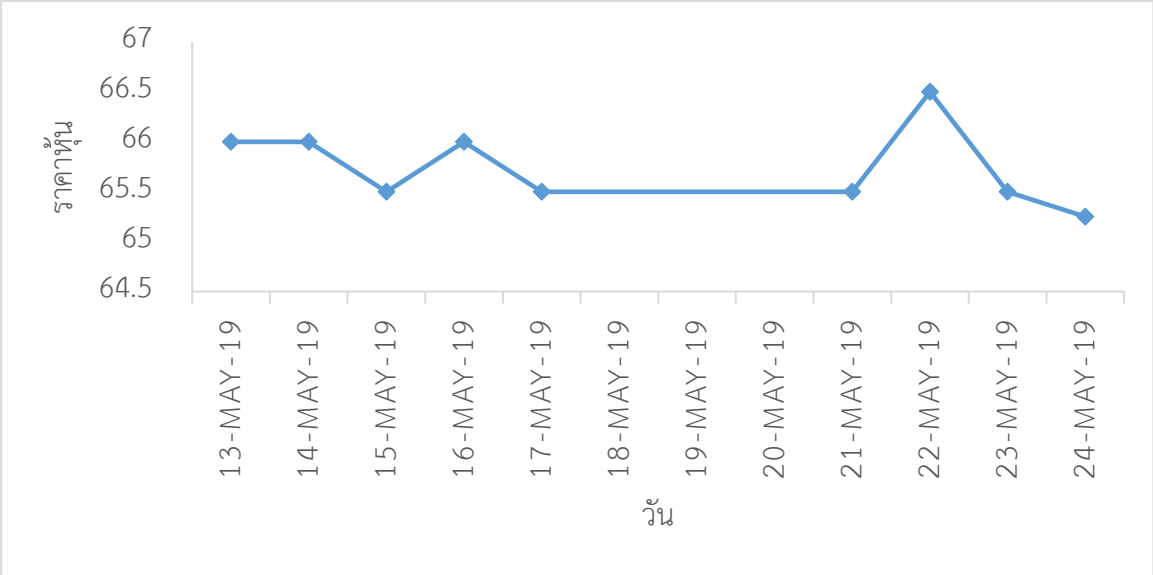
เวลา	คำ	จำนวน
2018-6-18 00:00:00	dafbama2018got7	13
2018-6-18 00:01:00	dafbama2018got7	18
2018-6-18 00:02:00	dafbama2018got7	6

5.2.2 การสร้างคุณลักษณะใหม่จากตัวชี้วัดของหุ้น (Stock Indicator Feature Extraction)

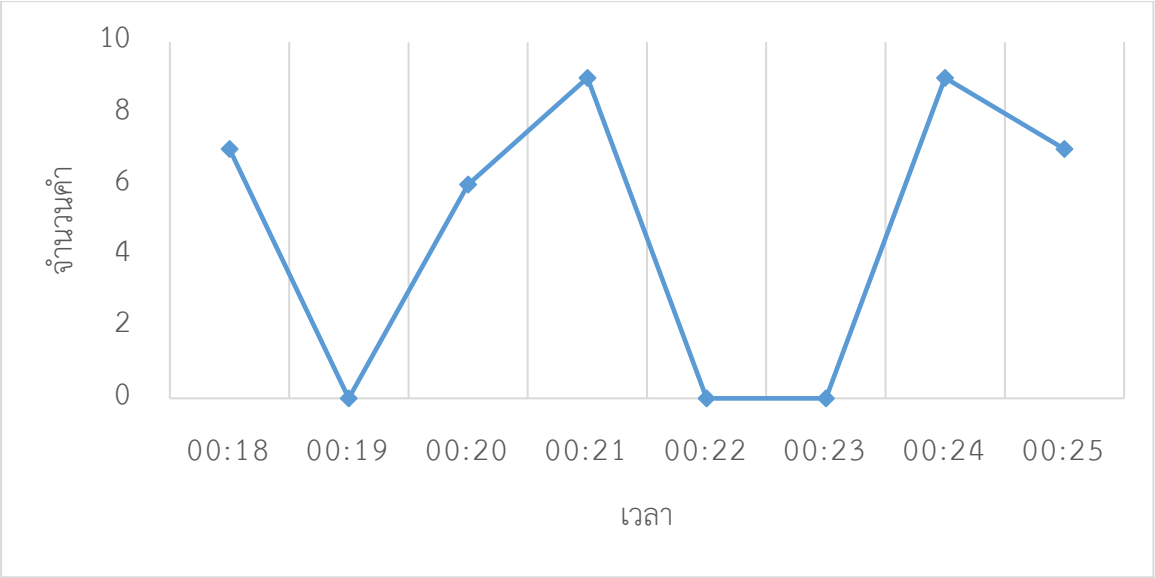
จากคุณลักษณะของงานวิจัย [7] ที่มีความคล้ายกับตัวชี้วัดของหุ้น $MACD$ แสดงให้เห็นถึงการพยายามนำตัวชี้วัดของหุ้นมาประยุกต์ใช้ในการตรวจจับคำเกิดใหม่ ดังนั้น ผู้วิจัยจึงประยุกต์ตัวชี้วัดของหุ้นที่นิยมใช้ในปัจจุบันมาเป็นหนึ่งในคุณลักษณะของงานวิจัยนี้ โดยเฉพาะตัวชี้วัดของหุ้นประเภทโมเมนตัมที่เป็นตัวชี้วัดของหุ้นสนใจเฉพาะราคาที่เปลี่ยนแปลง กล่าวคือไม่สนใจปริมาณการซื้อขาย โดยตัวชี้วัดของหุ้นประเภทโมเมนตัมนิยมใช้ในการหาสถานะของราคาหุ้น ระหว่าง สถานะซื้อมากเกินไป (Overbought), สถานะขายมากเกินไป (Oversold) หรือสถานะปกติ ซึ่งสถานะซื้อมากเกินไปหรือราคาของหุ้นสูงผิดปกตินั้น สามารถมองเป็นสถานะที่จำนวนคำปรากฏมากผิดปกติ หรือเป็นคำที่มีโอกาสเป็นคำเกิดใหม่ได้ ดังนั้นตัวชี้วัดของหุ้นประเภทโมเมนตัมสามารถนำมาประยุกต์ใช้โดยการแทนที่ “ราคาของหุ้นของแต่ละวัน” ด้วย “จำนวนคำในแต่ละนาที” ดังแสดงตัวอย่างของราคาหุ้นของแต่ละวันและจำนวนคำในแต่ละนาทีไว้ในรูปที่ 41 และ รูปที่ 42

ตัวชี้วัดของหุ้นประเภทโมเมนตัมที่นำมาประยุกต์ใช้ในงานวิจัยนี้ ได้แก่

1. RSI หรือ Relative Strength Index ในบทที่ 2.2.4
2. STOCH หรือ Stochastic Oscillator ในบทที่ 2.2.5
3. WR หรือ William Percent Range ในบทที่ 2.2.6



รูปที่ 41 แสดงราคาของหุ้น KKP ในแต่ละช่องของเวลา (วัน)



รูปที่ 42 แสดงจำนวนคำของ “dafbama2018got7” ในแต่ละช่องของเวลา (นาที)

1. การสร้างคุณลักษณะของตัวชี้วัดของหุ้น rsi

จากสมการการสร้างตัวชี้วัดของหุ้น rsi ใน (6), (7), (8), (9), (10), (11), (12) และ (13) ในบทที่ 2.2.4 สามารถนำมาประยุกต์ใช้กับจำนวนคำในแต่ละนาทีได้ดังนี้

กำหนดให้

$count_t$ คือ จำนวนคำที่ปรากฏในเวลา t

w คือ กรอบเวลา

t คือ เวลา

$t-1$ คือ เวลาครั้งล่าสุดที่ปรากฏ

$$gain_t = MAX(count_t - count_{t-1}, 0) \quad (37)$$

$$loss_t = MAX(count_{t-1} - count_t, 0) \quad (38)$$

$$avg\_gain_{w,t} = (close_d - avg\_gain_{w,t-1}) * \frac{2}{w+1} + avg\_gain_{w,t-1} \quad (39)$$

$$avg\_loss_{w,t} = (close_d - avg\_loss_{w,t-1}) * \frac{2}{w+1} + avg\_loss_{w,t-1} \quad (40)$$

$$rs_{w,t} = \frac{avg\_gain_{w,t}}{avg\_loss_{w,t}} \tag{41}$$

$$rsi_{w,t} = 100 - \frac{100}{1 + rs_{w,t}} \tag{42}$$

แต่ในการคำนวณหา $avg\_gain_{w,t}$ และ $avg\_loss_{w,t}$ ในสมการที่ (39) และ (40) นั้น เป็นการคำนวณหาค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก $EWMA$ ที่ต้องคำนวณหาทุกช่องของเวลา แต่ข้อมูลในงานวิจัยนี้อาจจะไม่ปรากฏในทุกช่องของเวลา ดังนั้นจึงนำสมการการคำนวณหาค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนักที่สามารถเพิ่มค่า (Incremental feature) และสามารถคำนวณแบบขี้เกียจได้ (Lazy computing) จากสมการที่ (34) ของงานวิจัย [7] มาใช้เพื่อลดภาระในการคำนวณ ซึ่งจะได้สมการในการคำนวณหา $avg\_gain_{w,t}$ และ $avg\_loss_{w,t}$ ใหม่ดังนี้

กำหนดให้ ΔT คือ ระยะห่างของเวลาที่ปรากฏครั้งล่าสุด

$$avg\_gain_{w,t} = avg\_gain_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{gain_t}{w} \tag{43}$$

$$avg\_loss_{w,t} = avg\_loss_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{loss_t}{w} \tag{44}$$

2. การสร้างคุณลักษณะของตัวชี้วัดของหุ้น $stoch$

จากสมการการสร้างตัวชี้วัดของหุ้น $stoch$ ใน (14) ในบทที่ 2.2.5 สามารถนำมาประยุกต์ใช้กับจำนวนคำในแต่ละนาทีได้ ดังนี้

กำหนดให้

$count_t$ คือ จำนวนคำที่ปรากฏในเวลา t

w คือ กรอบเวลา

t คือ เวลา

$count\_lowest_{w,t}$ คือจำนวนคำที่น้อยที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$count\_highest_{w,t}$ คือจำนวนคำที่มากที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

β คือ ค่าความลำเอียง

$$stoch_{w,t} = \frac{count_t - count\_lowest_{w,t} + \beta}{count\_highest_{w,t} - count\_lowest_{w,t} + \beta} \quad (45)$$

3. การสร้างคุณลักษณะของตัวชี้วัดของหุ้น wr

จากสมการการสร้างตัวชี้วัดของหุ้น wr ใน (16) ในบทที่ 2.2.6 สามารถนำมาประยุกต์ใช้กับจำนวนคำในแต่ละนาทีได้ ดังนี้

กำหนดให้

$count_t$ คือ จำนวนคำที่ปรากฏในเวลา t

w คือ กรอบเวลา

t คือ เวลา

$count\_lowest_{w,t}$ คือจำนวนคำที่น้อยที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$count\_highest_{w,t}$ คือจำนวนคำที่มากที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$$wr_{w,t} = 100 - 100 * \frac{count\_highest_{w,t} - count_t + \beta}{count\_highest_{w,t} - count\_lowest_{w,t} + \beta} \quad (46)$$

5.2.3 การปรับปรุงคุณลักษณะใหม่จากตัวชี้วัดของหุ้น ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก (Improving Stock Indicator Feature using Exponential Moving Average)

ตัวชี้วัดของหุ้น $rsi, stoch$ และ wr สร้างจากราคาของหุ้นซึ่งมีค่าในทุกช่องของเวลา(*วัน*) ต่อเนื่องกัน โดยค่อย ๆ ปรับราคาขึ้นลงจากค่าของเวลาก่อนหน้า ดังรูปที่ 41 แต่จำนวนคำในแต่ละนาทีของงานวิจัยนี้ ไม่ได้มีค่าในทุกช่องของเวลาต่อเนื่องกัน กล่าวคือมีบางเวลาที่มีค่าเป็น 0 ดังภาพดังรูปที่ 42 ทำให้เกิดปัญหาในการคำนวณ $gain_t, loss_t$ และ $count\_lowest_{w,t}$ และทำให้คุณลักษณะ $rsi, stoch$ และ wr มีค่าผิดปกติได้ง่าย

ดังนั้น ผู้วิจัยจึงใช้คุณลักษณะที่มีค่าต่อเนื่องกันในทุกช่องของเวลาแทนการใช้จำนวนคำในแต่ละนาที ซึ่งคุณลักษณะที่จะถูกนำมาใช้แทนจำนวนคำ คือ v ในสมการที่ (34) ซึ่งเป็นค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนักที่สามารถเพิ่มค่า (Incremental feature) และสามารถคำนวณแบบขี้เกียจได้ (Lazy computing) ซึ่งจะได้สมการในการคำนวณใหม่ดังนี้

1. การปรับปรุงคุณลักษณะของตัวชี้วัดของหุ้น rsi ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก v

กำหนดให้

w คือ กรอบเวลา

t คือ เวลา

$v_{w,t}$ คือ ค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนักของจำนวนคำในกรอบเวลา w วันย้อนหลังนับจากเวลา t

ΔT คือ ระยะห่างของเวลาที่ปรากฏครั้งล่าสุด

$$gain2_t = MAX(v_{w,t} - v_{w,t-1}, 0) \tag{47}$$

$$loss2_t = MAX(v_{w,t-1} - v_{w,t}, 0) \tag{48}$$

$$avg\_gain2_{w,t} = avg\_gain2_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{gain2_t}{w} \tag{49}$$

$$avg\_loss2_{w,t} = avg\_loss2_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{loss2_t}{w} \tag{50}$$

$$rs2_{w,t} = \frac{avg\_gain2_{w,t}}{avg\_loss2_{w,t}} \tag{51}$$

$$rsi2_{w,t} = 100 - \frac{100}{1 + rs2_{w,t}} \tag{52}$$

2. การปรับปรุงคุณลักษณะของตัวชี้วัดของหุ้น $stoch$ ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก v

กำหนดให้

$v\_lowest_{w,t}$ คือ v_t ที่น้อยที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$v\_highest_{w,t}$ คือ v_t ที่มากที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$$stoch2_{w,t} = \frac{v_{w,t} - v\_lowest_{w,t} + \beta}{v\_highest_{w,t} - v\_lowest_{w,t} + \beta} \tag{53}$$

3. การปรับปรุงคุณลักษณะของตัวชี้วัดของหุ้น wr ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก v

กำหนดให้

$v\_lowest_{w,t}$ คือ v_t ที่น้อยที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$v\_highest_{w,t}$ คือ v_t ที่มากที่สุดในกรอบเวลา w วันย้อนหลัง นับจากเวลา t

$$wr2_{w,t} = 100 \; - \; 100 \; * \frac{v\_highest_{w,t} - v_{w,t} + \beta}{v\_highest_{w,t} - v\_lowest_{w,t} + \beta} \tag{54}$$

5.2.4 การปรับปรุงคุณลักษณะใหม่จากตัวชี้วัดของหุ้น ด้วยความต่างของกรอบเวลา (Improving Stock Indicator Feature using Different Time Period)

จากสมการการสร้างความต่างของค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนักของ 2 กรอบเวลาที่แตกต่างกันของ (35) ที่คล้ายกับ $MACD$ สามารถนำมาประยุกต์ใช้กับตัวชี้วัดของหุ้น rsi, $stoch$ และ wr ได้ดังนี้

กำหนดให้

$$rsi2\_cd_{w1,w2,t} = \frac{rsi2_{w1,t} - rsi2_{w2,t}}{w2 - w1} \tag{55}$$

$$stoch2\_cd_{w1,w2,t} = \frac{stoch2_{w1,t} - stoch2_{w2,t}}{w2 - w1} \tag{56}$$

$$wr2\_cd_{w1,w2,t} = \frac{wr2_{w1,t} - wr2_{w2,t}}{w2 - w1} \tag{57}$$

5.2.5 การปรับปรุงคุณลักษณะใหม่จากตัวชี้วัดของหุ้น ด้วยการชะลอการเปลี่ยนแปลง (Improving Stock Indicator Feature using Delaying Change)

ในบางตัวชี้วัดของหุ้น มีการชะลอการเปลี่ยนแปลงของตัวชี้วัดของหุ้นเพื่อไม่ให้ค่าของคุณลักษณะเปลี่ยนแปลงเร็วเกินไป ดังแสดงในสมการ (15) ในบทที่ 2.2.5 ซึ่งสามารถนำมาประยุกต์ใช้กับตัวชี้วัดของหุ้น rsi, $stoch$ และ wr ได้ดังนี้

กำหนดให้

w คือ กรอบเวลา

ΔT คือ ระยะห่างของเวลาที่ปรากฏครั้งล่าสุด

$$rsi2\_slow_{w,t} = rsi2\_slow_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{rsi2_{w,t}}{w} \quad (58)$$

$$stoch2\_slow_{w,t} = stoch2\_slow_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{stoch2_{w,t}}{w} \quad (59)$$

$$wr2\_slow_{w,t} = wr2\_slow_{w,t-1} * e^{\frac{-\Delta T}{w}} + \frac{wr2_{w,t}}{w} \quad (60)$$

สรุปคุณลักษณะที่ถูกนำเสนอในงานวิจัยนี้

คุณลักษณะที่ถูกนำเสนอในงานวิจัยนี้มีทั้งหมด 12 คุณลักษณะ ดังแสดงในตารางที่ 26 และมีตัวอย่างผลลัพธ์สุดท้ายของการสร้างข้อมูลแสดงในตารางที่ 27

ตารางที่ 26 สรุปคุณลักษณะที่ถูกนำเสนอในงานวิจัย

สร้างจากตัวชี้วัดของหุ้น	ปรับปรุงด้วย v	ปรับปรุงด้วย $\mathbf{w}$	ปรับปรุงด้วยการชะลอ
$rsi_{w,t}$	$rsi2_{w,t}$	$rsi2\_cd_{w1,w2,t}$	$rsi2\_slow_{w,t}$
$stoch_{w,t}$	$stoch2_{w,t}$	$stoch2\_cd_{w1,w2,t}$	$stoch2\_slow_{w,t}$
$wr_{w,t}$	$wr2_{w,t}$	$wr2\_cd_{w1,w2,t}$	$wr2\_slow_{w,t}$

ตารางที่ 27 ตัวอย่างผลลัพธ์สุดท้ายของขั้นตอนการสร้างข้อมูล

เวลา	คำ	จำนวนคำ	rsi_5	rsi_{10}	rsi_{15}
2018-6-18 00:00:00	dafbama2018got7	13	86.111	69.697	61.207
2018-6-18 00:01:00	dafbama2018got7	18	92.009	79.392	69.727
2018-6-18 00:02:00	dafbama2018got7	6	90.852	79.583	70.583
2018-6-18 00:03:00	dafbama2018got7	16	92.635	83.076	74.504
2018-6-18 00:04:00	dafbama2018got7	10	92.045	83.609	75.634

5.3 การตรวจจับคำเกิดใหม่ (Emerging Keyword Detection)

เป็นการนำชุดข้อมูลมาทำนายหาคำเกิดใหม่ที่แตกต่างกันในแต่ละนาที โดยมีขั้นตอนในการตรวจจับคำเกิดใหม่ 3 ขั้นตอน ได้แก่

1. การกำกับข้อมูล (Data Labeling) : การสร้างตัวแปรผลเฉลยในข้อมูล เพื่อใช้ในการสร้างตัวจำแนกประเภท
2. การสร้างตัวจำแนกประเภท (Model Construction) : การสร้างตัวจำแนกประเภท เพื่อเรียนรู้พฤติกรรมของคำเกิดใหม่
3. การทำนายคำเกิดใหม่ (Emerging Keyword Prediction) : การนำตัวจำแนกประเภทมาทำนายหาคำเกิดใหม่

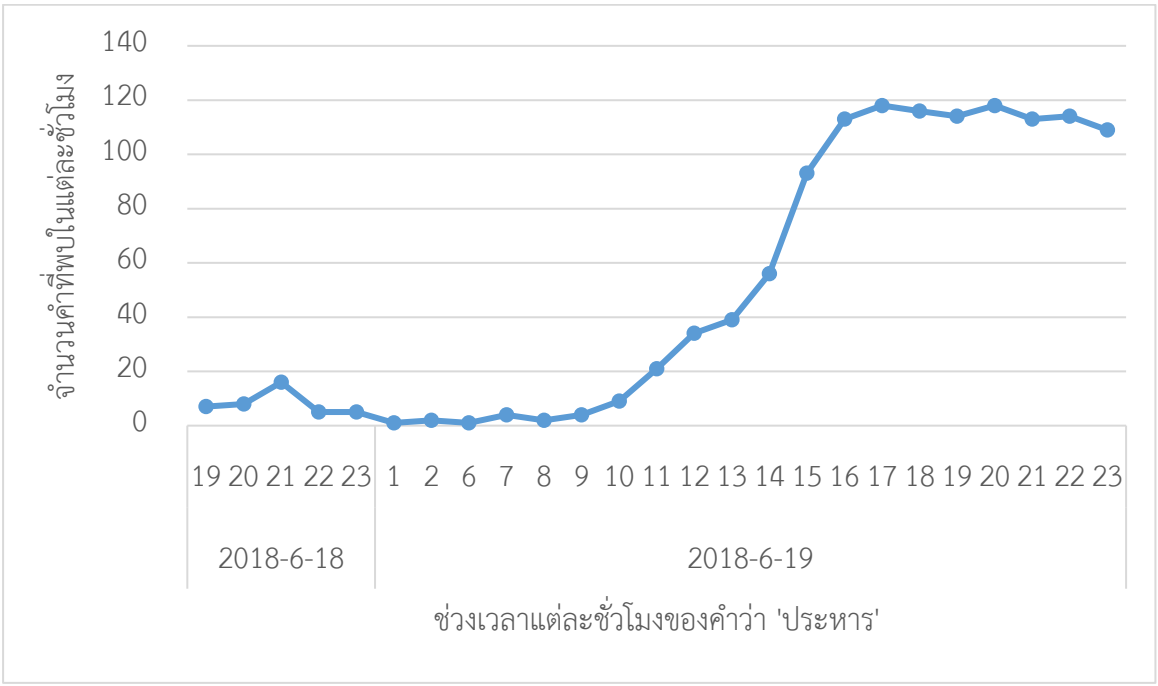
5.3.1 การกำกับข้อมูล (Data Labeling)

การกำกับข้อมูลหรือการสร้างตัวแปรผลเฉลยเพื่อใช้ในการสร้างตัวจำแนกประเภททำได้ยาก เนื่องจากไม่สามารถกำหนดได้ว่าหัวข้อเกิดใหม่แต่ละหัวข้อ เกิดในช่วงเวลาใดของแต่ละวัน แม้จะเป็นหัวข้อที่มีเวลากำกับเด่นชัดก็ตาม ตัวอย่างที่เห็นได้ชัดคือ การแข่งขันกีฬาที่แข่งในเวลากลางคืน 02.00 ถึง 04.00 ของเวลาไทย ซึ่งเป็นช่วงเวลาที่คนส่วนมากกำลังพักผ่อน ดังนั้นช่วงเวลาดังกล่าวหัวข้อของการแข่งขันฟุตบอลจะไม่ถูกพูดถึงในข้อความทวิตเตอร์มากนัก แต่จะถูกพูดถึงในช่วงเช้าแทน อีกตัวอย่างคือหัวข้อที่ไม่มีเวลาเกิดขึ้นชัดเจน เช่น หัวข้อของ “การประหารชีวิตนักโทษ” ที่เริ่ม

ออกข่าวในวันที่ 18 มิถุนายน ค.ศ. 2018 แต่คำว่า “ประหาร” ปรากฏครั้งแรกในข้อความทวิตเตอร์วันที่ 18 มิถุนายน ค.ศ. 2018 เวลา 19.00 โดยปรากฏเพียงไม่กี่คำต่อชั่วโมง แล้วเริ่มปรากฏบ่อยในข้อความทวิตเตอร์ในช่วงกลางวันของวันที่ 19 มิถุนายน ค.ศ. 2018 ดังแสดงในรูปที่ 43 จากตัวอย่างข้างต้นแม้ว่าหัวข้อของ “การประหารชีวิตนักโทษ” จะเกิดในวันที่ 18 มิถุนายน ค.ศ. 2018 แต่ไม่สามารถกำกับข้อมูลว่าเกิดในวันดังกล่าวได้

ดังนั้น การกำกับข้อมูลในงานวิจัยนี้จึงทำการกำกับข้อมูลทุกแถวที่มีคำปรากฏใน**ชุดผลเฉลยในวันเดียวกัน** โดยไม่พิจารณาว่าข้อมูลดังกล่าวจะมีคุณลักษณะหรือสัญญาณเด่นหรือไม่ เพื่อให้สามารถกำกับข้อมูลของคำเกิดใหม่ได้ครบถ้วนและไม่อาศัยพารามิเตอร์ในการกำกับข้อมูล แต่ในขณะเดียวกันการกำกับข้อมูลวิธีทำให้เกิดการกำกับข้อมูลเกิน (Over labeling) ซึ่งอาจส่งผลทำให้ตัวจำแนกประเภททำนายผิดพลาด โดยในงานวิจัยนี้ทำการแก้ไขปัญหาผลกระทบของการกำกับข้อมูลเกินในบทการสร้างตัวแนกประเภทหรือบทที่ 5.3.2

จากรูปที่ 43 ข้อมูลที่มีคำว่า “ประหาร” ในช่วงเช้าของวันที่ 19 มิถุนายน ค.ศ. 2018 ปรากฏน้อยจนเป็นไปได้ยากที่จะเป็นคำเกิดใหม่ แต่ด้วยวิธีการกำกับข้อมูลของงานวิจัยนี้ ทำให้ข้อมูลแถวดังกล่าวถูกกำกับว่าเป็นคำเกิดใหม่ ซึ่งก็คือการกำกับข้อมูลเกิน



รูปที่ 43 จำนวนคำในแต่ละชั่วโมงของคำว่า “ประหาร” ในวันที่ 19 มิถุนายน ค.ศ. 2018

ในตารางที่ 28 แถวของข้อมูลที่มีคำว่า “ประหาร” ทั้งหมดจะถูกกำกับข้อมูลว่า “ใช่” ในตัวแปรของ “คำเกิดใหม่” เนื่องจากคำว่า “ประหาร” ปรากฏในชุดผลเฉลยของ วันที่ 19 มิถุนายน ค.ศ. 2018 ในขณะที่คำว่า “ฝันดี” ซึ่งไม่ปรากฏในชุดผลเฉลยของวันที่ 19 มิถุนายน ค.ศ. 2018 จะถูกกำกับข้อมูลว่า “ไม่” ในตัวแปรของ “คำเกิดใหม่” โดยสถิติจำนวนข้อมูลที่ถูกกำกับว่า “ไม่ใช่” และ “ใช่” ถูกแสดงในตารางที่ 29

ตารางที่ 28 ตัวอย่างข้อมูลที่ผ่านการกำกับข้อมูล

เวลา	คำ	คุณลักษณะ	คำเกิดใหม่
2018-6-19 00:00:00	ประหาร	...	ใช่
2018-6-19 00:01:00	ประหาร	...	ใช่
2018-6-19 00:01:00	ฝันดี	...	ไม่ใช่

ตารางที่ 29 สถิติของข้อมูลที่ถูกกำกับว่า “ไม่ใช่” และ “ใช่”

วัน	จำนวนข้อมูลที่กำกับด้วย “ไม่ใช่”	จำนวนข้อมูลที่กำกับด้วย “ใช่”
2018-06-12	284,663	35,684
2018-06-13	278,828	26,351
2018-06-14	258,085	13,882
2018-06-15	215,824	29,756
2018-06-16	237,773	26,451
2018-06-17	217,625	24,013
2018-06-18	210,830	28,304
2018-06-19	202,577	27,474
2018-06-20	200,996	20,862
รวม	4,214,402	465,554

5.3.2 การสร้างตัวจำแนกประเภท (Model Construction)

การสร้างตัวจำแนกประเภทจำเป็นต้องใช้ชุดข้อมูลสอน (Training dataset) และชุดข้อมูลตรวจสอบ (Validation dataset) ที่มีตัวแปรผลเฉลย(Target variable) โดยใช้ชุดข้อมูลสอนในการสร้างตัวจำแนกประเภท และใช้ชุดข้อมูลตรวจสอบในการตรวจสอบประสิทธิภาพของตัวจำแนกประเภทที่สร้างขึ้น โดยการแบ่งข้อมูลออกเป็นชุดข้อมูลสอนและชุดข้อมูลตรวจสอบมักทำโดยการตรวจสอบแบบไขว้ (Cross validation) แต่เนื่องจากข้อมูลในงานวิจัยนี้มีเวลากำกับ ดังนั้นจึงใช้การตรวจสอบแบบไขว้เชิงเวลา (Temporal cross validation) ในบทที่ 2.4

นอกจากนี้การกำกับข้อมูลเกิน (Over labeling) จากบทที่ 5.3.1 หรือบทการกำกับข้อมูล (Data labeling) ทำให้ตัวจำแนกประเภทที่ถูกเลือกจากการตรวจสอบแบบไขว้เชิงเวลาทำนายผิดพลาด เนื่องจากในแต่ละรอบของการตรวจสอบแบบไขว้ หากข้อมูลตรวจสอบมีแถวของข้อมูลที่ถูกกำกับข้อมูลเกิน จะทำให้ประสิทธิภาพของตัวจำแนกประเภทคลาดเคลื่อน จากรูปที่ 43 ข้อมูลที่มีคำว่า “ประหาร” ในช่วงเช้าถูกกำกับข้อมูลเกินว่าเป็นคำเกิดใหม่ ซึ่งในความเป็นจริงข้อมูลดังกล่าวไม่ใช่คำเกิดใหม่ ดังนั้น ถ้าตัวจำแนกประเภททำนายข้อมูลแถวดังกล่าวว่า “เป็นคำเกิดใหม่” จะทำให้ประสิทธิภาพในการทำนายของการตรวจสอบมีค่าสูงขึ้น ซึ่งผิดไปจากความเป็นจริงที่ตัวจำแนกประเภทนี้ควรได้คะแนนที่น้อยลง ในทางกลับกัน ถ้าตัวจำแนกประเภททำนายข้อมูลแถวดังกล่าวว่า “ไม่เป็นคำเกิดใหม่” จะทำให้ประสิทธิภาพในการทำนายของการตรวจสอบมีค่าน้อยลง ซึ่งผิดไปจากความเป็นจริงที่ตัวจำแนกประเภทนี้ควรได้คะแนนที่มากขึ้น นอกจากนี้การที่ตัวจำแนกประเภทพยายามทำนายแถวของข้อมูลที่ถูกกำกับข้อมูลเกิน จะทำให้ข้อมูลทั่วไปมีโอกาสถูกตอบว่าเป็นคำเกิดใหม่มากขึ้น และทำให้ประสิทธิภาพในการทำนายลดลง

ปัญหาดังกล่าวเกิดจากการกำกับข้อมูลเกินและการให้ความสำคัญกับข้อมูลตรวจสอบทุกแถวมากเกินไป ดังนั้นผู้วิจัยจึงเสนอมาตรวัดประสิทธิภาพในมุมมองของคำ (Keyword Measurement) ในบทที่ 4.3.1 ที่ให้คะแนนตามจำนวนคำที่แตกต่างกันที่ทำนายถูกต้องต่อวัน แทนการใช้มาตรวัดประสิทธิภาพทั่วไปที่ให้คะแนนตามจำนวนแถวของข้อมูลที่ถูกทำนายถูกต้อง

ดังนั้น ถ้าในวันดังกล่าวมีแถวของข้อมูลที่มีคำว่า “ประหาร” ถูกทำนายว่าเป็นคำเกิดใหม่แล้ว แถวของข้อมูลที่มีคำว่า “ประหาร” ที่ถูกกำกับข้อมูลเกินในข้อมูลตรวจสอบ ไม่ว่าจะถูกทำนายว่าเป็นคำเกิดใหม่หรือไม่ ประสิทธิภาพของการตรวจสอบจะไม่ลดลงหรือเพิ่มขึ้น ทำให้การวัดประสิทธิภาพในมุมมองของคำทนต่อการกำกับข้อมูลเกินมากกว่าการวัดประสิทธิภาพทั่วไป

นอกจากนี้ เพื่อเพิ่มประสิทธิภาพในการทำนาย การสร้างตัวจำแนกประเภทในงานวิจัยนี้จึงใช้คุณลักษณะหลายชนิด ได้แก่ 1) คุณลักษณะที่นำเสนอในงานวิจัยนี้, 2) คุณลักษณะที่นำเสนอในงานวิจัยอื่น และ 3) ตัวแปรที่ใช้ในการสร้างคุณลักษณะต่าง ๆ

จากการทดลองตัวจำแนกประเภทชนิดต่าง ๆ ได้แก่ ต้นไม้ตัดสินใจ (Decision Tree), ป่าไม้แบบสุ่ม (Random Forest), เพื่อนบ้านใกล้ที่สุด K ตัว (K Nearest Neighbor), นาอีฟเบย์ (Navie Bayes) และ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) พบว่าป่าไม้แบบสุ่มให้ประสิทธิภาพในมุมมองของคำดีที่สุด

5.3.3 การทำนายคำเกิดใหม่ (Emerging Keyword Prediction)

เป็นการนำตัวจำแนกประเภทที่ได้จากขั้นตอนก่อนหน้า ไปใช้ในการทำนายข้อมูลในแต่ละนาทีว่าคำใดเป็นคำเกิดใหม่ จากนั้นนำคำที่ถูกนายว่าเป็นคำเกิดใหม่เป็นผลลัพธ์สุดท้ายของขั้นตอนการตรวจจับคำเกิดใหม่ ส่งให้กับระบบตรวจจับหัวข้อเกิดใหม่ ดังแสดงในตารางที่ 30

ตารางที่ 30 ตัวอย่างผลลัพธ์สุดท้ายของขั้นตอนการตรวจจับคำเกิดใหม่

เวลา	คำที่ถูกทำนายว่าใช่ที่แตกต่างกัน
2018-6-19 00:00:00	“dafbama2018got7” “got7” “คุย” “ติด” “นอน” “นั่ง” “น้อง” “รู้” “สู้” “ไหม” “ทีม” “ลิขิตรักthecrownprincess” “popezaap”
2018-6-19 00:01:00	“dafbama2018got7” “ขนาด” “คุย” “ติด” “นั่ง” “น้อง” “รู้” “สงสาร” “สู้” “โลก” “ไหม” “ทีม” “ลิขิตรักthecrownprincess” “ฉาก” “จูบ”
2018-6-19 00:02:00	“dafbama2018got7” “eyesonyou” “got7worldtour” “ขนาด” “คุย” “นอน” “นั่ง” “น้อง” “สาว” “สู้” “แฟน” “ไหม” “ทีม” “ลิขิตรักthecrownprincess” “popezaap”
2018-6-19 00:03:00	“นอน” “นั่ง” “น้อง” “รัก” “รู้” “ร้อง” “สงสาร” “สู้” “เดิม” “แฟน” “ไหม” “ทีม” “ลิขิตรักthecrownprincess” “ฉาก”

5.4 การตรวจจับหัวข้อเกิดใหม่ (Emerging Topic Detection)

ขั้นตอนในการตรวจจับหัวข้อเกิดใหม่มี 2 ขั้นตอน ได้แก่

1. การสร้างหัวข้อเกิดใหม่ในแต่ละนาที (Emerging Topic Construction) : การสร้างหัวข้อเกิดใหม่จากการจับกลุ่มคำเกิดใหม่ที่ทำนายได้ในแต่ละนาที
2. การสร้างหัวข้อเกิดใหม่รวม (Stateful Emerging Topic Construction) : การสร้างหัวข้อเกิดใหม่รวม จากการรวมหัวข้อเกิดใหม่รวมในนาทีก่อนหน้า กับ หัวข้อเกิดใหม่ในนาทีปัจจุบัน

ตัวอย่างโครงสร้างการทำงานของการสร้างหัวข้อเกิดใหม่ในแต่ละนาที และการสร้างหัวข้อเกิดใหม่รวม ถูกแสดงอยู่ในรูปที่ 44

รูปที่ 44 ตัวอย่างของการตรวจจับหัวข้อเกิดใหม่

5.4.1 การสร้างหัวข้อเกิดใหม่ในแต่ละนาที (Emerging Topic Construction)

การสร้างหัวข้อเกิดใหม่แต่ละนาที สร้างได้โดยการนำคำที่ถูกทำนายว่าเป็นคำเกิดใหม่ที่แตกต่างกันที่ได้จากขั้นตอนการตรวจจับคำเกิดใหม่ มาหาความสัมพันธ์ระหว่างคู่ของคำ โดยคู่ของคำไหนที่มีความคล้ายกันมาก จะถูกจับกลุ่มเข้าด้วยกันเป็นหัวข้อเกิดใหม่ในนาทีนั้น จากรูปที่ 44 สามารถตรวจจับหัวข้อเกิดใหม่ในนาทีที่ 1 ได้ 1 หัวข้อ โดยหัวข้อดังกล่าวมีคำเกิดใหม่ 2 คำ ได้แก่ คำว่า “เยอรมัน” และ “ฟุตบอล”

การสร้างหัวข้อเกิดใหม่ในแต่ละนาที แบ่งเป็น 3 ขั้นตอน ได้แก่ 1) การสร้างเวกเตอร์ของคำ, 2) การหาความสัมพันธ์ของแต่ละคู่ของคำ และ 3) การสร้างหัวข้อเกิดใหม่

1. การสร้างเวกเตอร์ของคำ

การสร้างเวกเตอร์ของคำใช้อัลกอริทึม LSI ในบทที่ 2.6 โดยการนำเมทริกซ์ความสัมพันธ์ของคำ w คำ กับข้อความทวิตเตอร์ย้อนหลัง d ข้อความ ไปแตกองค์ประกอบของเมทริกซ์และย่อขนาดด้วย SVD จนเหลือ t หัวข้อ ซึ่งเมทริกซ์ U จะเป็นเมทริกซ์ความสัมพันธ์ของคำ w คำกับหัวข้อ t หัวข้อ ดังนั้น แถวแต่ละแถวของเมทริกซ์ U จะเป็นเวกเตอร์ที่ขนาด t ซึ่งเป็นเวกเตอร์ความสัมพันธ์ของแต่ละคำกับหัวข้อ t หัวข้อ

2. การหาความสัมพันธ์ของแต่ละคู่ของคำ

นำเวกเตอร์ของคำแต่ละคำ มาหาความคล้ายด้วยสัมประสิทธิ์ความคล้ายโคไซน์ (Cosine similarity) ดังแสดงในตารางที่ 31

ตารางที่ 31 ตัวอย่างความสัมพันธ์ระหว่างคู่ของคำโดยใช้สัมประสิทธิ์ความคล้ายโคไซน์

เวลา	คำที่ 1	คำที่ 2	ความคล้าย
2018-6-19 00:01:00	eyesonyou	got7worldtour	0.910488
2018-6-19 00:01:00	dafbama2018got7	got7	0.743079
2018-6-19 00:01:00	dafbama2018got7	got7worldtour	0.710819

3. การสร้างหัวข้อเกิดใหม่

นำค่าความคล้ายโคไซน์ (Cosine similarity) ของแต่ละคู่ของคำมาคัดกรองด้วยเส้นแบ่งความคล้ายที่กำหนดไว้ล่วงหน้า (Predefined threshold)

จากตารางที่ 31 หากกำหนดเส้นแบ่งของความคล้ายไว้ที่ 0.9 จะพบว่าในเวลา 2018-6-19 00:01:00 จะได้หัวข้อเกิดใหม่จำนวน 1 หัวข้อ ซึ่งประกอบด้วยคำ 2 คำ ได้แก่ “eyesonyou” และ “got7worldtour”

หากกำหนดเส้นแบ่งของความคล้ายไว้ที่ 0.72 พบว่าในเวลา 2018-6-19 00:01:00 จะได้หัวข้อเกิดใหม่จำนวน 2 หัวข้อ โดยหัวข้อแรกประกอบด้วยคำ 2 คำ ได้แก่ “eyesonyou” และ “got7worldtour” ในขณะที่หัวข้อที่สองประกอบด้วยคำ 2 คำ ได้แก่ “dafbama2018got7” และ “got7” ดังแสดงในตารางที่ 32

หากกำหนดเส้นแบ่งของความคล้ายไว้ที่ 0.7 พบว่าในเวลา 2018-6-19 00:01:00 จะได้หัวข้อเกิดใหม่จำนวน 1 หัวข้อ ซึ่งประกอบด้วยคำ 4 คำ ได้แก่ “eyesonyou”, “got7worldtour”, “dafbama2018got7” และ “got7”

ตารางที่ 32 ตัวอย่างหัวข้อเกิดใหม่ในเวลา 2018-6-19 00:01:00 เมื่อกำหนดเส้นแบ่งของความคล้ายไว้ที่ 0.72

เวลา	กลุ่มของคำที่ถูกทำนายและมีความสัมพันธ์กัน
2018-6-19 00:01:00	“eyesonyou”, “got7worldtour”
2018-6-19 00:01:00	“dafbama2018got7”, “got7”

5.4.2 สร้างหัวข้อเกิดใหม่รวม (Stateful Emerging Topic Construction)

การสร้างหัวข้อเกิดใหม่รวม คือ การนำหัวข้อเกิดใหม่ในแต่ละช่วงเวลาที่ใกล้เคียงกันมารวมกัน เพื่อสร้างหัวข้อเกิดใหม่ที่สมบูรณ์ยิ่งขึ้น โดยพิจารณาจากการปรากฏร่วมของคำในหัวข้อเกิดใหม่ที่พึ่งถูกสร้างในเวลาปัจจุบัน กับคำในหัวข้อเกิดใหม่รวมของเวลาก่อนหน้า โดยถ้ามีคำปรากฏร่วมกันจะทำการรวมหัวข้อเกิดใหม่เข้าเปเดียวกัน

กำหนดให้

t คือ นาทีของเวลาปัจจุบัน

$t-1$ คือ นาทีก่อนหน้าของเวลาปัจจุบัน

$topic_t$ คือ กลุ่มของหัวข้อเกิดใหม่ในแต่ละนาทีของนาทีที่ t

$topic_{t,i}$ คือ หัวข้อเกิดใหม่ในแต่ละนาทีของนาทีที่ t หัวข้อที่ i

$Topic_t$ คือ กลุ่มของหัวข้อเกิดใหม่**รวม**ของนาทีที่ t

$Topic_{t,i}$ คือ หัวข้อเกิดใหม่**รวม**ของนาทีที่ t หัวข้อที่ i

จากตัวอย่างหัวข้อเกิดใหม่ของเวลาปัจจุบัน $topic_t$ ในตารางที่ 32 และจากตัวอย่างหัวข้อเกิดใหม่รวมของเวลาก่อนหน้า $Topic_{t-1}$ ในตารางที่ 33 พบว่าหัวข้อเกิดใหม่แรกของหัวข้อเกิดใหม่ปัจจุบัน $topic_{t,1}$ ประกอบด้วยคำ 2 คำ ได้แก่ “eyesonyou” และ “got7worldtour” ซึ่งคำ “eyesonyou” นั้นปรากฏในหัวข้อเกิดใหม่แรกของหัวข้อเกิดใหม่รวมของเวลาก่อนหน้า $Topic_{t-1,1}$ ดังนั้นจะทำการเพิ่มคำว่า “got7worldtour” เข้าไปในหัวข้อเกิดใหม่รวม $Topic_{t-1,1}$

หัวข้อเกิดใหม่ที่สองของหัวข้อเกิดใหม่ปัจจุบัน $topic_{t,2}$ ประกอบด้วยคำ 2 คำ ได้แก่ “dafbama2018got7” และ “got7” ซึ่งทั้งสองคำในปรากฏในหัวข้อเกิดใหม่รวมของเวลาก่อนหน้าคนละหัวข้อกัน ได้แก่หัวข้อ $Topic_{t-1,1}$ และ หัวข้อ $Topic_{t-1,2}$ ดังนั้นจะทำการรวมหัวข้อเกิดใหม่

แรก $Topic_{t-1,1}$ และหัวข้อเกิดใหม่ที่สอง $Topic_{t-1,2}$ ของหัวข้อเกิดใหม่รวมของเวลาก่อนหน้าเข้าด้วยกัน สุดท้ายจะได้ผลลัพธ์ของการสร้างหัวข้อเกิดใหม่รวมดังแสดงในตารางที่ 34 และจะได้ผลลัพธ์ของการสร้างหัวข้อเกิดใหม่จากข้อมูลจริงดังแสดงในตารางที่ 35

ตารางที่ 33 ตัวอย่างหัวข้อเกิดใหม่รวมของเวลาก่อนหน้า

เวลาเริ่ม	เวลาจบ	กลุ่มของคำที่ถูกทำนายและมีความสัมพันธ์กัน
2018-6-19 00:00:00	2018-6-19 00:00:00	“eyesonyou”, “dafbama2018got7”
2018-6-19 00:00:00	2018-6-19 00:00:00	“bringgot7togdny”, “got7”

ตารางที่ 34 ตัวอย่างผลลัพธ์จากการรวมหัวข้อเกิดใหม่รวมของเวลาก่อนหน้า กับ หัวข้อเกิดใหม่ของเวลาปัจจุบัน

เวลาเริ่ม	เวลาจบ	กลุ่มของคำที่ถูกทำนายและมีความสัมพันธ์กัน
2018-6-19 00:00:00	2018-6-19 00:01:00	“eyesonyou”, “dafbama2018got7”, “got7worldtour”, “bringgot7togdny”, “got7”

ตารางที่ 35 ตัวอย่างผลลัพธ์สุดท้ายของขั้นตอนการตรวจจับหัวข้อเกิดใหม่

เวลาเริ่ม	เวลาจบ	ระยะเวลา	กลุ่มของคำที่ถูกทำนายและมีความสัมพันธ์กัน
2018-6-19 12:31:00	2018-6-19 13:18:00	00:48:00	ประหาร, โทษ
2018-6-19 18:18:00	2018-6-19 18:18:00	00:01:00	ฆ่า, ประหารชีวิต
2018-6-19 19:06:00	2018-6-19 20:36:00	01:31:00	ญี่ปุ่น, สวัสดีบอลโลก, เชียร์, แดง
2018-6-19 20:50:00	2018-6-19 21:41:00	00:52:00	สายรักสายสวาท, หญิง, เปลว, โฉม
2018-6-19 21:00:00	2018-6-20 13:32:00	16:33:00	จูบ, ฉาก, ตัด, ทหาร, ท่าน, ปู่, พัน, ฟิน, มือ, ละคร, ลิขิตรักthecrownprincess, วัง, สงสาร, สอง, อลัน, เคท, เคธ, เจ้าหญิง, เภตรา, แผน

บทที่ 6

การทดลองและผลการทดลอง

การทดลองของงานวิจัยนี้แบ่งออกเป็น 2 การทดลองหลัก และ 1 บทวิเคราะห์ ได้แก่

1. การทดลองและผลการทดลองของการตรวจจับคำเกิดใหม่ (Emerging Keyword Experiment and Result) : การทดลองและผลการทดลองที่สนใจเพียงจำนวนคำเกิดใหม่ที่ตรวจจับได้ โดยไม่สนใจการจับกลุ่มคำเกิดใหม่เป็นหัวข้อเกิดใหม่
2. การทดลองและผลการทดลองของการตรวจจับหัวข้อเกิดใหม่ (Emerging Topic Experiment and Result) : การทดลองและผลการทดลองที่สนใจทั้งจำนวนคำเกิดใหม่ที่ตรวจจับได้ และการจับกลุ่มของคำเกิดใหม่หรือหัวข้อเกิดใหม่
3. การวิเคราะห์ข้อมูลเพิ่มเติม (Other Experiment) : การวิเคราะห์ผลลัพธ์ของการตรวจจับหัวข้อเกิดใหม่ของวิธีที่นำเสนอในงานวิจัย

6.1 การทดลองและผลการทดลองของการตรวจจับคำเกิดใหม่ (Emerging Keyword Experiment and Result)

การทดลองและผลการทดลองของการตรวจจับคำเกิดใหม่ เป็นการวัดประสิทธิภาพเฉพาะส่วนของคำเกิดใหม่ที่ตรวจจับได้ โดยไม่สนใจการจับกลุ่มคำเป็นหัวข้อ โดยแบ่งออกเป็น 4 การทดลอง ได้แก่

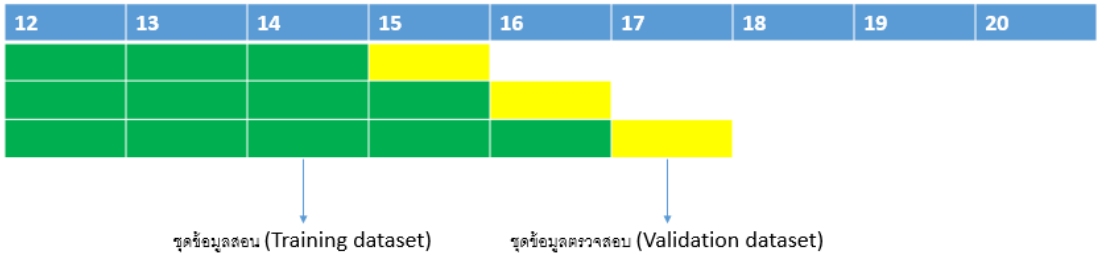
1. การทดลองประสิทธิภาพของคุณลักษณะใหม่
2. การทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากคุณลักษณะต่าง ๆ
3. การทดลองประสิทธิภาพของการตรวจจับคำเกิดใหม่
4. การทดลองความเร็วในการตรวจจับคำเกิดใหม่

การเลือกพารามิเตอร์ของแต่ละวิธีในการตรวจจับคำเกิดใหม่

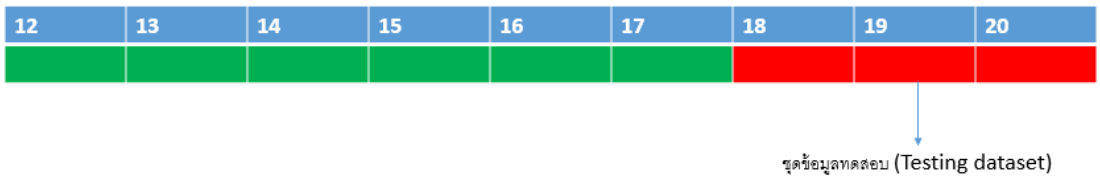
เลือกพารามิเตอร์ที่ดีที่สุดของแต่ละวิธีบนการตรวจสอบแบบไขว้เชิงเวลา (Temporal cross validation) 3 รอบ

1. รอบที่ 1 ใช้ข้อมูล 3 วันแรกเป็นชุดข้อมูลสอน (Training dataset) และใช้ข้อมูล 1 วันถัดมาหรือข้อมูลวันที่ 4 เป็นชุดข้อมูลตรวจสอบ (Validation dataset)
2. รอบที่ 1 ใช้ข้อมูล 4 วันแรกเป็นชุดข้อมูลสอน (Training dataset) และใช้ข้อมูล 1 วันถัดมาหรือข้อมูลวันที่ 5 เป็นชุดข้อมูลตรวจสอบ (Validation dataset)
3. รอบที่ 1 ใช้ข้อมูล 5 วันแรกเป็นชุดข้อมูลสอน (Training dataset) และใช้ข้อมูล 1 วันถัดมาหรือข้อมูลวันที่ 6 เป็นชุดข้อมูลตรวจสอบ (Validation dataset)

จากนั้นทำการทดสอบประสิทธิภาพในมุมมองของคำบนข้อมูล 3 วันถัดมา หรือข้อมูล 3 สุดท้าย ดังแสดงในรูปที่ 45 และรูปที่ 46



รูปที่ 45 การแบ่งข้อมูลสอนและตรวจสอบบนการตรวจสอบแบบไขว้เชิงเวลา



รูปที่ 46 การแบ่งข้อมูลทดสอบ

6.1.1 การทดลองประสิทธิภาพของคุณลักษณะใหม่

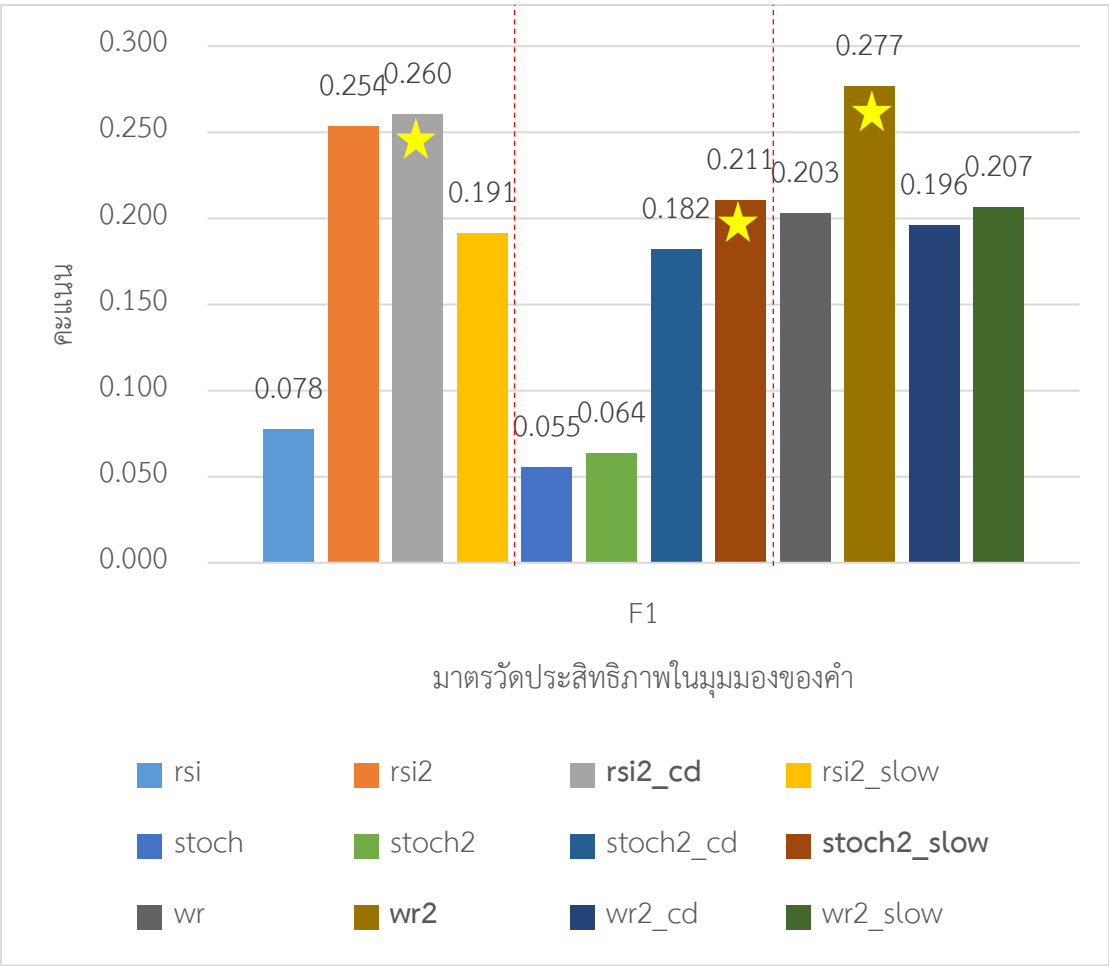
1. วัตถุประสงค์ของการทดลอง

- ทดลองประสิทธิภาพของคุณลักษณะตัวชี้วัดของหุ้นที่ถูกนำเสนอในงานวิจัยนี้ ในบทที่ 5.2.2 ถึง 5.2.5 ได้แก่
  - $rsi, stoch$ และ wr
  - $rsi2, stoch2$ และ $wr2$
  - $rsi2\_cd, stoch2\_cd$ และ $wr2\_cd$
  - $rsi2\_slow, stoch2\_slow$ และ $wr2\_slow$
- ทดสอบหาค่ากรอบเวลา w ที่ดีที่สุดของแต่ละคุณลักษณะในการตรวจสอบแบบไขว้ ได้แก่ 5, 10, 15, 30 และ 60
- ทดสอบหาคู่ของกรอบเวลา $w1, w2$ ที่ดีที่สุดของคุณลักษณะ $rsi2\_cd, stoch2\_cd$ และ $wr2\_cd$ ในการตรวจสอบแบบไขว้ ได้แก่ (5,10), (5,15), (5,15), (10,15), (10,30), (15,30), (15,60) และ (30,60)
- ทดสอบหาค่าจุดแบ่งเปอร์เซ็นต์ไทล์ (Percentile) ที่ดีที่สุดที่ใช้เป็นเส้นแบ่ง (Threshold) ของแต่ละคุณลักษณะในการตรวจสอบแบบไขว้ ได้แก่ 0.1, 0.2, ..., 0.8, 0.9, 0.95, 0.96, 0.97, 0.98 และ 0.99

2. ผลการทดลอง

- พบว่าในกลุ่มคุณลักษณะ rsi ได้แก่ คุณลักษณะ $rsi, rsi2, rsi2\_cd$ และ $rsi2\_slow$ คุณลักษณะที่ดีที่สุดคือ $rsi2\_cd$ โดยมีคู่ของกรอบเวลา $w1$ และ $w2$ เท่ากับ 30 และ 60 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ 0.98 หรือมีค่าเท่ากับ 0.163153
- พบว่าในกลุ่มคุณลักษณะ $stoch$ ได้แก่ คุณลักษณะ $stoch, stoch2$, $stoch2\_cd$ และ $stoch2\_slow$ คุณลักษณะที่ดีที่สุดคือ $stoch2\_slow$ โดยมีค่ากรอบเวลา w เท่ากับ 10 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ 0.9 หรือมีค่าเท่ากับ 67.21507
- พบว่าในกลุ่มคุณลักษณะ wr ได้แก่ คุณลักษณะ $wr, wr2, wr2\_cd$ และ $wr2\_slow$ คุณลักษณะที่ดีที่สุดคือ $wr2$ โดยมีค่ากรอบเวลา w เท่ากับ 30 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ 0.98 หรือมีค่าเท่ากับ 95.105489
- พบว่าการปรับปรุงคุณลักษณะ ด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก v ได้ประสิทธิภาพดีกว่าก่อนการปรับปรุงในทุกกลุ่มคุณลักษณะ
  - $rsi2$ ได้ประสิทธิภาพดีกว่า rsi

- $stoch2$ ได้ประสิทธิภาพดีกว่า $stoch$
- $wr2$ ได้ประสิทธิภาพดีกว่า wr



รูปที่ 47 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมอง ของคำของการทดลองประสิทธิภาพของคุณลักษณะใหม่

3. สรุปผลการทดลอง

- การปรับปรุงคุณลักษณะด้วยค่าเฉลี่ยเคลื่อนที่ถ่วงน้ำหนัก v ได้ประสิทธิภาพดีขึ้นเสมอ ดังนั้นในการทดลองต่อจากนี้ จะไม่พิจารณาคุณลักษณะ $rsi, stoch$ และ wr
- การปรับปรุงคุณลักษณะด้วยด้วยความต่างของกรอบเวลา w และการชะลอการเปลี่ยนแปลง ทำให้ได้คุณลักษณะที่ดีที่สุดในกลุ่มคุณลักษณะ rsi และ $stoch$ ตามลำดับ

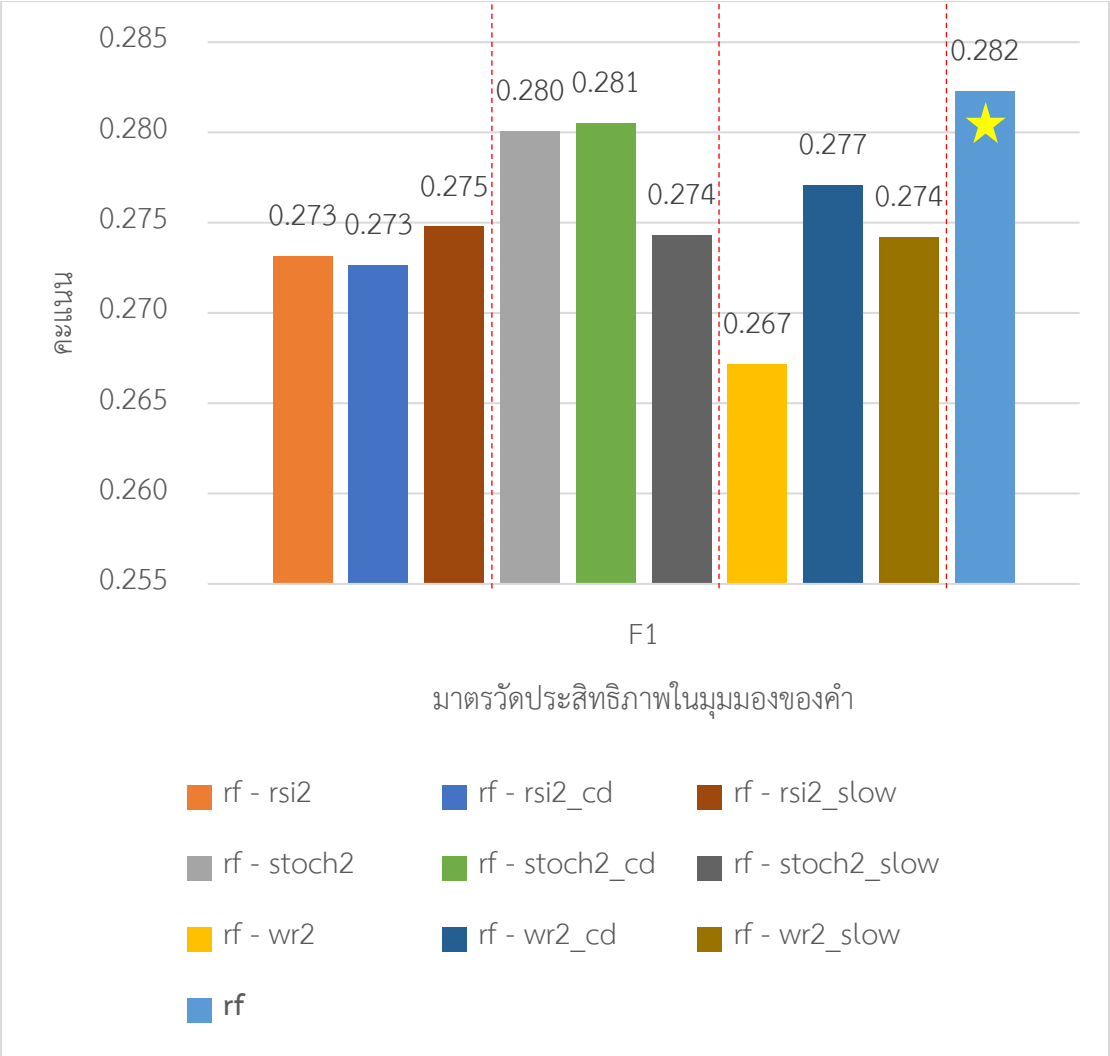
6.1.2 การทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากคุณลักษณะต่าง ๆ

1. วัตถุประสงค์ของการทดลอง

- ทดลองประสิทธิภาพของคุณลักษณะที่ใช้ในการสร้างตัวจำแนกประเภท
  - สร้างตัวจำแนกประเภทป่าไม้แบบสุ่มจากคุณลักษณะทั้งหมด rf
  - สร้างตัวจำแนกประเภทป่าไม้แบบสุ่มจากคุณลักษณะทั้งหมด ยกเว้นทีละคุณลักษณะ
    - ไม่ใช้คุณลักษณะ $rsi2$ หรือ rf – rsi2
    - ไม่ใช้คุณลักษณะ $rsi2\_cd$ หรือ rf – rsi2_cd
    - ไม่ใช้คุณลักษณะ $rsi2\_slow$ หรือ rf – rsi2_slow
    - ไม่ใช้คุณลักษณะ $stoch2$ หรือ rf – stoch2
    - ไม่ใช้คุณลักษณะ $stoch2\_cd$ หรือ rf – stoch2_cd
    - ไม่ใช้คุณลักษณะ $stoch2\_slow$ หรือ rf – stoch2_slow
    - ไม่ใช้คุณลักษณะ $wr2$ หรือ rf – wr2
    - ไม่ใช้คุณลักษณะ $wr2\_cd$ หรือ rf – wr2_cd
    - ไม่ใช้คุณลักษณะ $wr2\_slow$ หรือ rf – wr2_slow

2. ผลการทดลอง

- พบว่าเมื่อทำการสร้างตัวจำแนกประเภทจากคุณลักษณะทั้งหมดยกเว้นทีละคุณลักษณะ จะได้ประสิทธิภาพในมุมมองของคำ F_1 ด้อยกว่าการสร้างตัวจำแนกประเภทจากคุณลักษณะทั้งหมด
- พบว่าการสร้างตัวจำแนกประเภทจากคุณลักษณะทั้งหมด ยกเว้นคุณลักษณะ $wr2$ ได้ประสิทธิภาพในมุมมองของคำ F_1 ลดลงมากที่สุด บ่งบอกถึงคุณลักษณะ $wr2$ มีผลต่อการตรวจจับคำเกิดใหม่มาก



รูปที่ 48 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของคำ ของการทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากคุณลักษณะต่าง ๆ

3. สรุปผลการทดลอง

ตัวจำแนกประเภทที่สร้างจากคุณลักษณะทั้งหมด ได้ประสิทธิภาพในมุมมองของคำดีที่สุด ดังนั้นสามารถสรุปได้ว่า คุณลักษณะทั้งหมดที่ใช้มีผลต่อการตรวจจับคำเกิดใหม่ทุกคุณลักษณะ

6.1.3 การทดลองประสิทธิภาพของการตรวจจับคำเกิดใหม่

1. วัตถุประสงค์ของการทดลอง

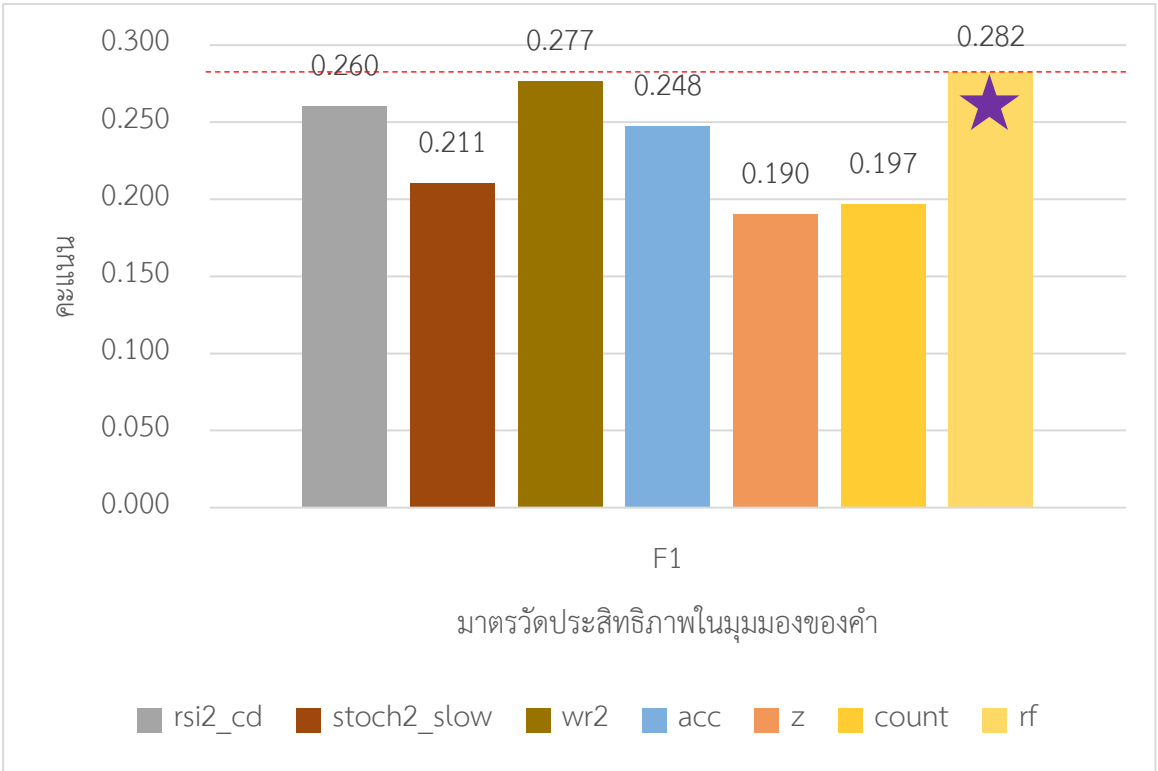
- ทดลองประสิทธิภาพในการตรวจจับคำเกิดใหม่ของตัวจำแนกประเภทและคุณลักษณะต่าง ๆ
  - ตัวจำแนกประเภทป่าไม้แบบสุ่มที่สร้างจากมาตรวัดในมุมมองของคำ rf
  - คุณลักษณะที่นำเสนอ $rsi2\_cd$, $stoch2\_slow$ และ $wr2$
    - เนื่องจากในกลุ่มคุณลักษณะ rsi พบว่าคุณลักษณะ $rsi2\_cd$ ให้ประสิทธิภาพดีที่สุด จากการผลทดลองที่ 6.1.1
    - เนื่องจากในกลุ่มคุณลักษณะ $stoch$ พบว่าคุณลักษณะ $stoch2\_slow$ ให้ประสิทธิภาพดีที่สุด จากการผลทดลองที่ 6.1.1
    - เนื่องจากในกลุ่มคุณลักษณะ wr พบว่าคุณลักษณะ $wr2$ ให้ประสิทธิภาพดีที่สุด จากการผลทดลองที่ 6.1.1
  - คุณลักษณะของงานวิจัยอื่น acc, z
- ทดสอบหาค่ากรอบเวลา w ที่ดีที่สุดของแต่ละคุณลักษณะในการตรวจสอบแบบไขว้ ได้แก่ 5, 10, 15, 30 และ 60
- ทดสอบหาคู่ของกรอบเวลา $w1, w2$ ที่ดีที่สุดของคุณลักษณะ $rsi2\_cd$, $stoch2\_cd$ และ $wr2\_cd$ ในการตรวจสอบแบบไขว้ ได้แก่ (5,10), (5,15), (5,15), (10,15), (10,30), (15,30), (15,60) และ (30,60)
- ทดสอบหาค่าจุดแบ่งเปอร์เซ็นต์ไทล์ (Percentile) ที่ดีที่สุดที่ใช้เป็นเส้นแบ่ง (Threshold) ของแต่ละคุณลักษณะในการตรวจสอบแบบไขว้ ได้แก่ 0.1, 0.2, ..., 0.8, 0.9, 0.95, 0.96, 0.97, 0.98 และ 0.99

2. ผลการทดลอง

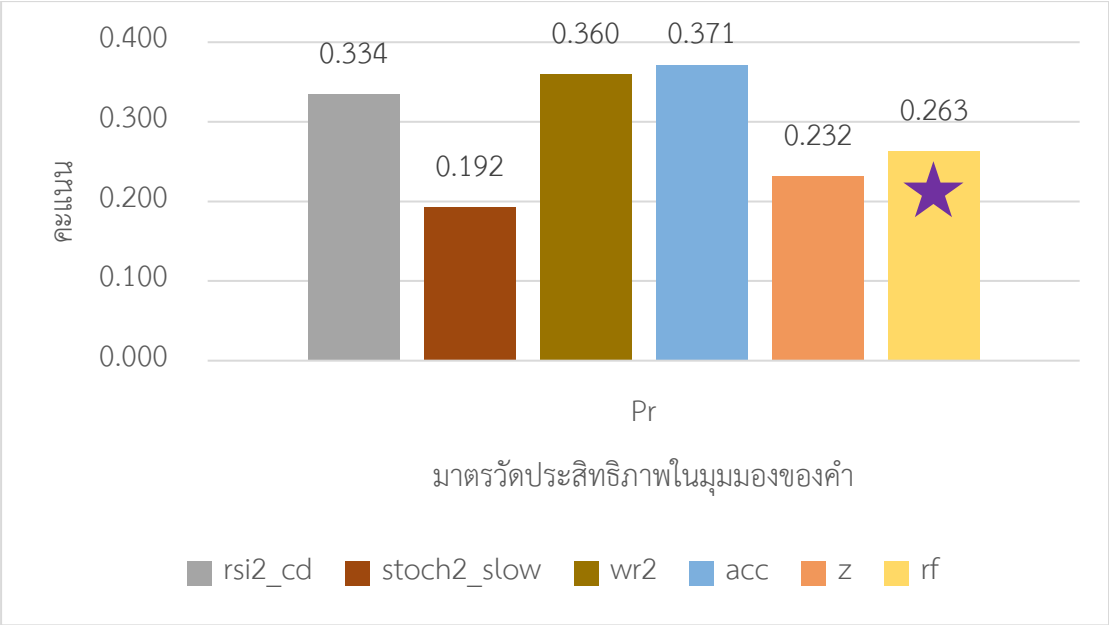
- จากผลการทดลองในเบื้องต้น พบว่าค่ากรอบเวลา w และค่าจุดแบ่งเปอร์เซ็นต์ไทล์ ที่ให้ประสิทธิภาพดีที่สุดของแต่ละคุณลักษณะมีดังนี้
  - คุณลักษณะ $rsi2\_cd$ มีคู่ของกรอบเวลา $w1$ และ $w2$ เท่ากับ 30 และ 60 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ 0.98 หรือมีค่าเท่ากับ 0.163153
  - คุณลักษณะ $stoch2\_slow$ มีค่ากรอบเวลา w เท่ากับ 10 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ 0.9 หรือมีค่าเท่ากับ 67.21507
  - คุณลักษณะ $wr2$ มีค่ากรอบเวลา w เท่ากับ 30 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ 0.98 หรือมีค่าเท่ากับ 95.105489

- คุณลักษณะ acc มีคู่ของกรอบเวลา เท่ากับ w 30 กับ 60 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ที่ 0.95 หรือมีค่าเท่ากับ 0.032824
- คุณลักษณะ z มีค่ากรอบเวลา w เท่ากับ 5 และมีค่าจุดแบ่งเปอร์เซ็นต์ไทล์ที่ 0.95 หรือมีค่าเท่ากับ 2.545816

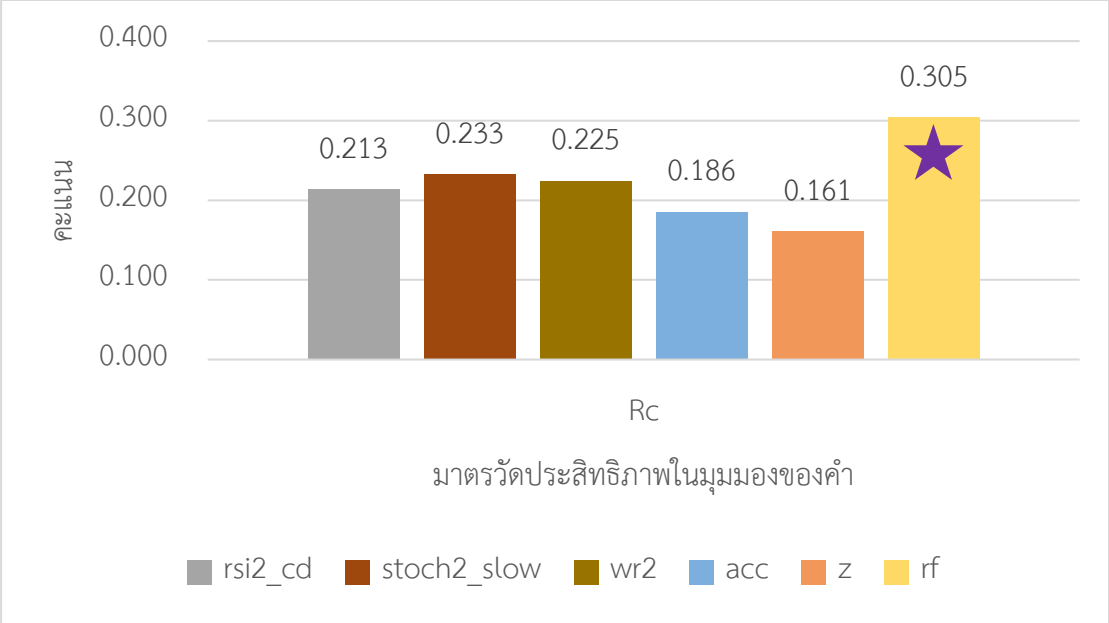
- พบว่าตัวจำแนกประเภทป่าไม้แบบสุ่ม rf ได้ประสิทธิภาพในมุมมองของคำ F_1 ดีที่สุด
- พบว่าคุณลักษณะ $wr2$ ให้ประสิทธิภาพในมุมมองของคำ F_1 เป็นอันดับ 2 และใกล้เคียงกับตัวจำแนกประเภทป่าไม้แบบสุ่ม rf มาก
  - จากรูปที่ 50 และ รูปที่ 51 เมื่อเทียบตัวจำแนกประเภท rf กับคุณลักษณะ $wr2$ พบว่าตัวจำแนกประเภท rf มี Rc สูงกว่า แต่มี Pr ต่ำกว่า
  - กล่าวคือตัวจำแนกประเภท rf สามารถตรวจจับคำเกิดใหม่ได้ครอบคลุมกว่า แต่ทำนายได้แม่นยำน้อยกว่า ซึ่งการตรวจจับคำได้ครอบคลุมกว่านั้นสำคัญกว่าความแม่นยำของทำนาย เนื่องจากการตรวจจับคำได้แม่นยำน้อย สามารถแก้ไขได้โดยการกรองเพิ่มเติมหลังจากตรวจจับคำได้ แต่การตรวจจับคำได้น้อยนั้นไม่สามารถแก้ไขให้ตรวจจับคำได้มากขึ้นได้



รูปที่ 49 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของคำ ของการทดลองประสิทธิภาพของการตรวจจับคำเกิดใหม่



รูปที่ 50 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของคำ Pr ของการทดลองประสิทธิภาพของการตรวจจับคำเกิดใหม่



รูปที่ 51 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของคำ Rc ของการทดลองประสิทธิภาพของการตรวจจับคำเกิดใหม่

3. สรุปผลการทดลอง

ตัวจำแนกประเภทที่เสนอในงานวิจัยนี้ ได้ประสิทธิภาพในการตรวจจับคำ F_1 ดีที่สุด และสามารถตรวจจับคำได้ครอบคลุมกว่าวิธีการอื่นมาก

6.1.4 การทดลองความเร็วในการตรวจจับคำเกิดใหม่

1. วัตถุประสงค์ของการทดลอง

- ทดลองวัดความเร็วในการตรวจจับคำเกิดใหม่แต่ละคำที่อยู่ในชุดผลเฉลยของตัวจำแนกประเภทป่าไม้แบบสุ่มที่สร้างจากมาตรวัดในมุมมองของคำ rf กับคุณลักษณะต่าง ๆ
  - คุณลักษณะที่นำเสนอ $rsi2\_cd$, $stoch2\_slow$ และ $wr2$
  - คุณลักษณะของงานวิจัยอื่น acc, z
- ใช้ค่ากรอบเวลา w และค่าจุดแบ่งเปอร์เซ็นต์ไทล์ ของแต่ละคุณลักษณะ ที่ได้ประสิทธิภาพดีที่สุดในการทดลองที่ 6.1.3
- ทำการวัดความเร็วในการตรวจจับคำเกิดใหม่ โดยดูจากคุณลักษณะต่อไปนี้
  - จำนวนคำที่ rf ตรวจจับได้เร็วกว่า
  - จำนวนคำที่ rf ตรวจจับได้ช้ากว่า
  - จำนวนคำที่ rf ตรวจจับได้ แต่วิธีการอื่นตรวจจับไม่ได้
  - จำนวนคำที่ rf ตรวจจับไม่ได้ แต่วิธีการอื่นตรวจจับได้

2. ผลการทดลอง

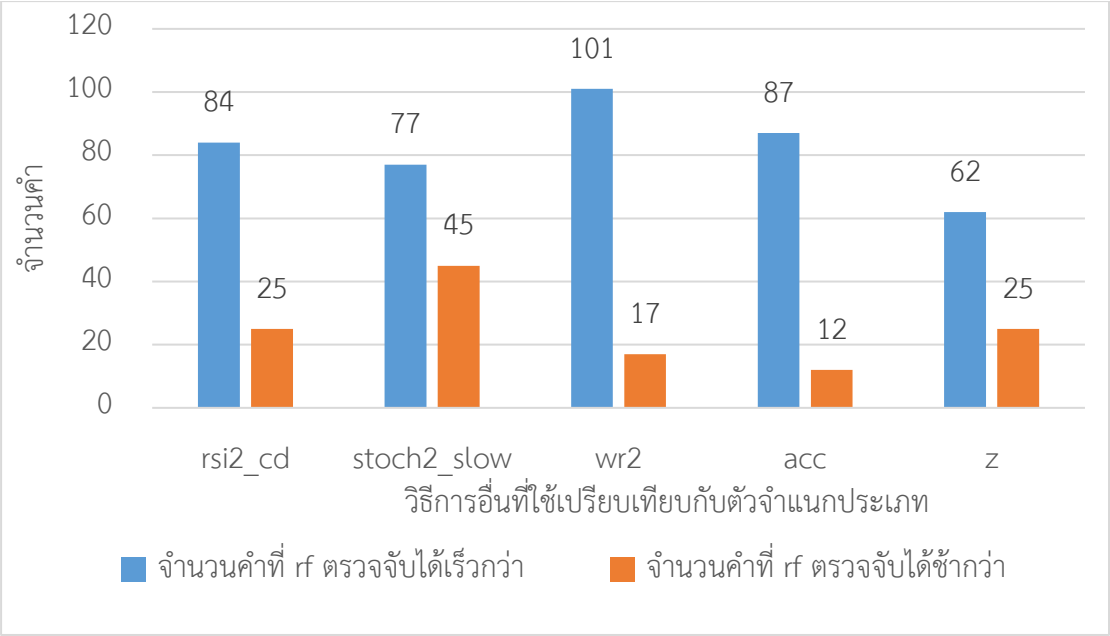
- จากรูปที่ 52 แสดงจำนวนคำที่ทั้งสองวิธีการตรวจจับพบ แต่ตรวจจับได้ไม่พร้อมกัน ซึ่งจะพบว่าตัวจำแนกประเภทป่าไม้แบบสุ่ม rf สามารถตรวจจับคำได้เร็วกว่าวิธีการอื่นทุกวิธี
- จากรูปที่ 53 แสดงจำนวนคำที่วิธีการใดวิธีการหนึ่งตรวจจับพบ แต่อีกวิธีการตรวจจับไม่พบ ซึ่งจะพบว่าตัวจำแนกประเภทป่าไม้แบบสุ่ม rf สามารถตรวจจับคำที่วิธีการอื่นตรวจจับไม่พบได้จำนวนมาก
- สาเหตุที่ทำให้ตัวจำแนกประเภทป่าไม้แบบสุ่ม rf สามารถตรวจจับคำได้เร็วกว่าวิธีการอื่น รวมถึงสามารถตรวจจับคำที่วิธีการอื่นตรวจจับไม่พบ เนื่องจากตัวจำแนกประเภทป่าไม้แบบสุ่ม rf มีความครอบคลุม (Recall) ในการตรวจจับคำที่สูงกว่าวิธีการอื่นมากในผลการทดลองใน 6.1.3

ตารางที่ 36 ตัวอย่างคำที่ทั้งสองวิธีการตรวจจับพบ แต่ตัวจำแนกประเภท rf ตรวจจับได้เร็วกว่า

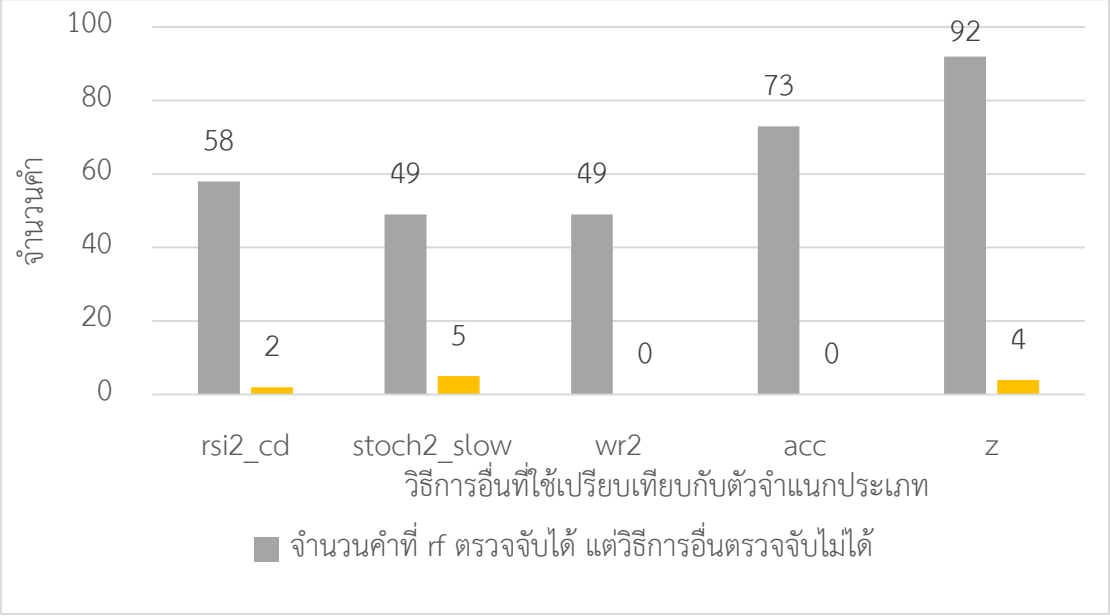
วัน	คำเกิดใหม่	เวลาที่ rf ตรวจพบ	เวลาที่ acc ตรวจพบ
2018-6-18	itcitybacon	19:06	20:02
	blackpink	00:00	10:21
	bambam	15:29	15:35
	สิง	21:32	21:34
	พี่สิง	21:29	21:32
	แทยอน	16:12	16:23
	taeyeon_somethingnew	15:05	15:27
	lazyloxy	20:46	20:53
	ปู่	20:58	21:00
	worldcup	00:00	00:01

ตารางที่ 37 ตัวอย่างคำที่ทั้งสองวิธีการตรวจจับพบ แต่ตัวจำแนกประเภท rf ตรวจจับได้ช้ากว่า

วัน	คำเกิดใหม่	เวลาที่ rf ตรวจพบ	เวลาที่ acc ตรวจพบ
2018-6-18	งาน	06:56	06:48
	เล่น	01:14	00:00
	สงสาร	21:16	20:57
2018-6-19	เพลง	10:42	07:30
	lisa	21:22	21:16
	ชนะ	03:00	02:58
	ละคร	23:06	23:05
2018-6-20	dafbama2018 peckpalitchoke	20:16	19:45
	ตาย	17:02	10:38
	รัก	06:50	06:32
	ปล้ำ	22:04	22:03
	แมงเม่า	21:21	21:15



รูปที่ 52 ผลการทดลองแสดงจำนวนคำที่ rf ตรวจจับได้เร็วกว่าและช้ากว่าเมื่อเทียบกับวิธีการอื่น ของการทดลองความเร็วในการตรวจจับคำเกิดใหม่



รูปที่ 53 ผลการทดลองแสดงจำนวนคำที่ rf ตรวจจับได้แต่อีกวิธีตรวจจับไม่ได้ และ จำนวนคำที่ rf ตรวจจับไม่ได้แต่วิธีการอื่นตรวจจับได้ ของการทดลองความเร็วในการตรวจจับคำเกิดใหม่

3. สรุปผลการทดลอง

ตัวจำแนกประเภทที่เสนอในงานวิจัยนี้สามารถตรวจจับคำได้เร็วกว่าทุกวิธี

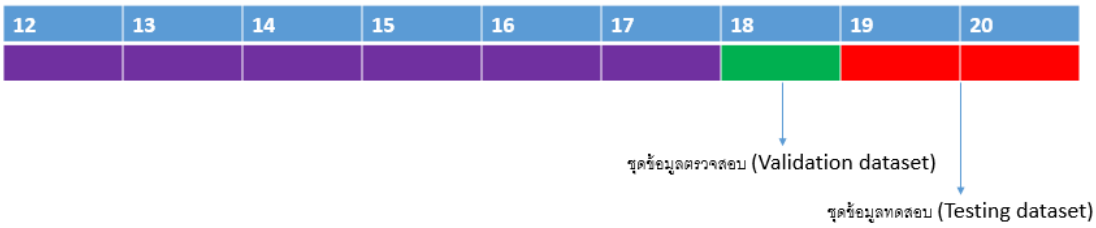
6.2 การทดลองและผลการทดลองของการตรวจจับหัวข้อเกิดใหม่ (Emerging Topic Experiment and Result)

การทดลองและผลการทดลองของการตรวจจับหัวข้อเกิดใหม่ เป็นการวัดประสิทธิภาพของคำเกิดใหม่ที่ตรวจจับได้และการจับกลุ่มคำเป็นหัวข้อเกิดใหม่พร้อมกัน โดยใช้มาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโคร (Macro Topic) โดยแบ่งออกเป็น 3 การทดลอง ได้แก่

1. การทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำ
2. การทดลองประสิทธิภาพของการตรวจจับคำและหัวข้อเกิดใหม่
3. การทดลองความเร็วในการตรวจจับหัวข้อเกิดใหม่

การเลือกพารามิเตอร์ของการตรวจจับหัวข้อเกิดใหม่

ทำการเลือกพารามิเตอร์ของการตรวจจับหัวข้อเกิดใหม่ ของแต่ละวิธีตรวจจับคำเกิดใหม่ โดยทำการตรวจสอบหาพารามิเตอร์ที่ดีที่สุดบนการข้อมูลในวันที่ 18 มิถุนายน ค.ศ. 2018 และนำไปวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครบนข้อมูล 2 วันถัดมา คือข้อมูลวันที่ 19 มิถุนายน ค.ศ. 2018 ถึง 20 มิถุนายน ค.ศ. 2018 ดังแสดงในรูปที่ 54



รูปที่ 54 การแบ่งข้อมูลของการตรวจจับหัวข้อเกิดใหม่

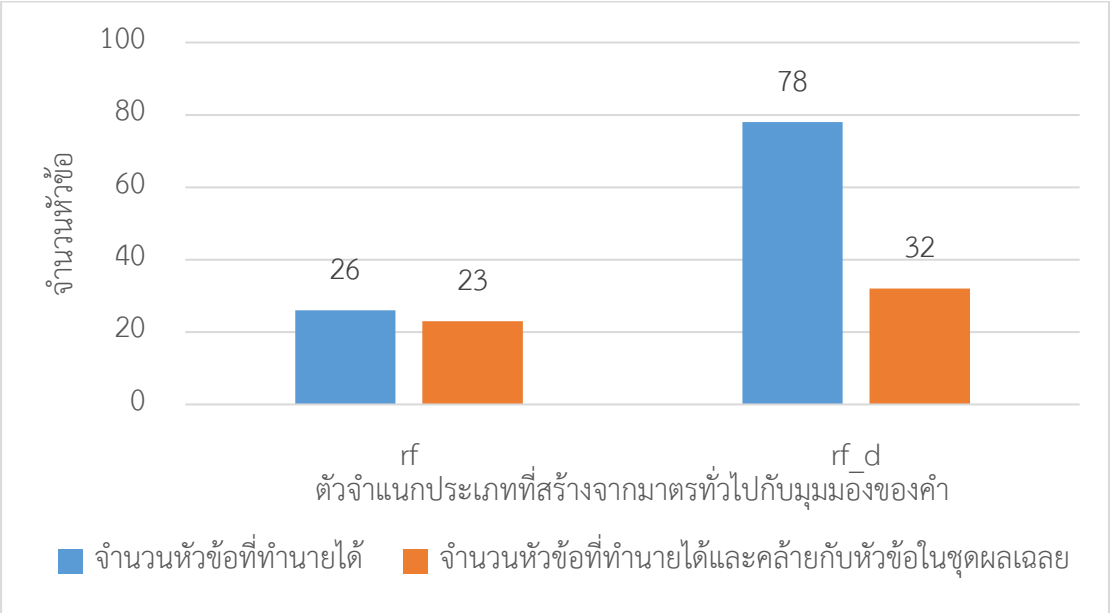
6.2.1 การทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำ

1. วัตถุประสงค์ของการทดลอง

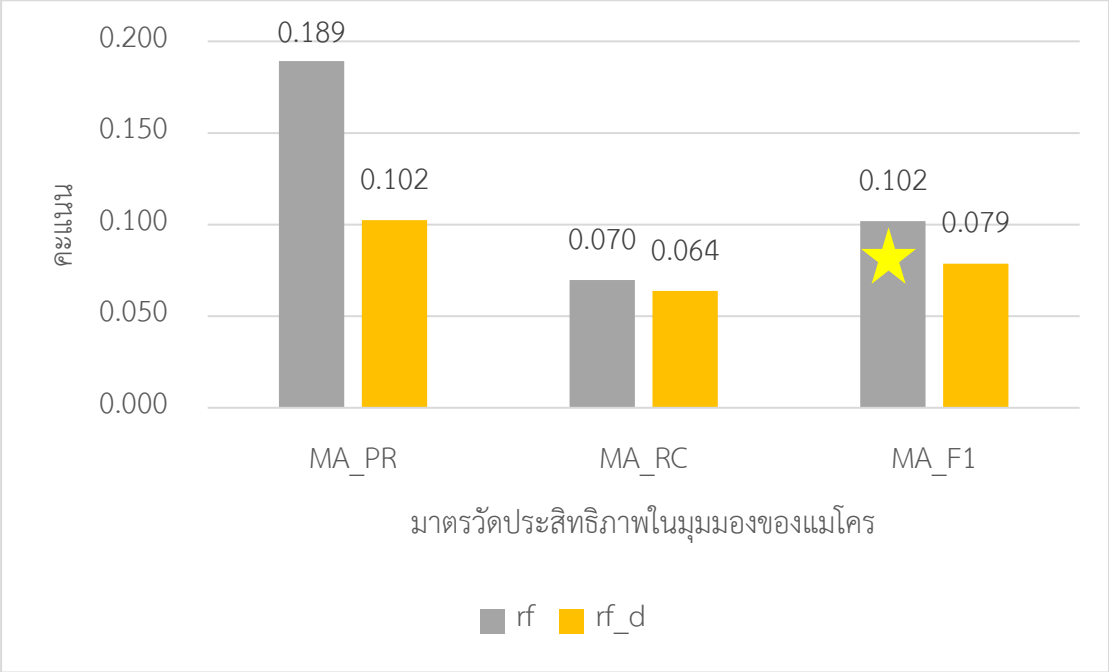
- ทดลองประสิทธิภาพของตัวจำแนกประเภทป่าไม้แบบสุ่มที่สร้างจากมาตรประสิทธิภาพวัดมาทั่วไปกับมาตรวัดในมุมมองของคำ
    - ป่าไม้แบบสุ่ม ที่สร้างจากมาตรวัดประสิทธิภาพในมุมมอง**ทั่วไป** rf_d
    - ป่าไม้แบบสุ่ม ที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของ**คำ** rf
- จับกลุ่มคำที่ถูกทำนายว่าเป็นคำเกิดใหม่ เพื่อสร้างหัวข้อเกิดใหม่ด้วย LSI
- จากผลการทดสอบหาค่ากรอบเวลา w ที่ดีที่สุดของการเลือกข้อความทวิตเตอร์ w นาทีย้อนหลังในขั้นตอนการตรวจจับหัวข้อเกิดใหม่ ในภาคผนวก ก พบว่าค่ากรอบเวลา w ที่ดีที่สุดคือ 30 ดังนั้นในการทดลองนี้จะใช้ข้อความทวิตเตอร์ย้อนหลัง 30 นาทีในการหาความสัมพันธ์ระหว่างคู่ของคำด้วยอัลกอริทึม LSI
- ทดสอบหาค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ในการคัดเลือกคู่ของคำที่มีความเกี่ยวข้องกันเป็นหัวข้อเดียวกัน ได้แก่ 0.1, 0.11, 0.12, ..., 0.97, 0.98 และ 0.99
- แบ่งเป็น 2 การทดลองย่อย ได้แก่
    - การทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro average หรือ MA) : ทำการตรวจสอบหาค่าพารามิเตอร์ที่ดีที่สุดโดยใช้ MA และทำการทดสอบประสิทธิภาพด้วย MA
    - การทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม (Macro merge หรือ MM) : ทำการตรวจสอบหาค่าพารามิเตอร์ที่ดีที่สุดโดยใช้ MM และทำการทดสอบประสิทธิภาพด้วย MM

2. ผลการทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro average หรือ MA)

- จากข้อมูลตรวจสอบ พบว่าตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองทั่วไป rf_d และ ตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf มีค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ที่ให้ประสิทธิภาพดีที่สุด เท่ากับ 0.9 และ 0.89 ตามลำดับ
- รูปที่ 55 แสดงจำนวนหัวข้อที่ทำนายได้จากทั้ง 2 วิธี
  - พบว่าจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf ทำนายออกมา 26 หัวข้อ และคล้ายกับหัวข้อในชุดผลเฉลยถึง 23 หัวข้อ กล่าวคือทำนายเกินเพียง 3 หัวข้อ
  - ในขณะที่จำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf_d ทำนายออกมา 78 หัวข้อ และคล้ายกับหัวข้อในชุดผลเฉลยเพียง 32 หัวข้อ กล่าวคือทำนายเกินถึง 46 หัวข้อ
- จากข้อมูลทดสอบในรูปที่ 56 พบว่าจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf ให้ประสิทธิภาพดีกว่าทั้งในด้านความแม่นยำและความครอบคลุม



รูปที่ 55 ผลการทดลองแสดงจำนวนหัวข้อที่ทำนายได้ และจำนวนหัวข้อที่ทำนายได้และคล้ายกับหัวข้อในชุดผลเฉลย ของการทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำโดยใช้ค่าเฉลี่ย



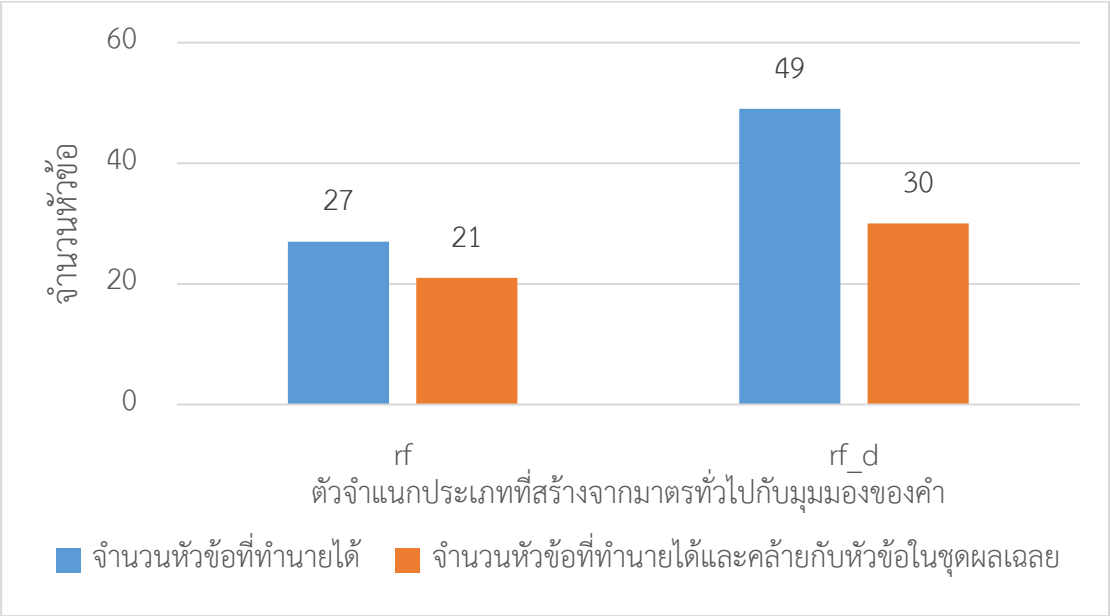
รูปที่ 56 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ยของการทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำ

3. ผลการทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม (Macro merge หรือ MM)

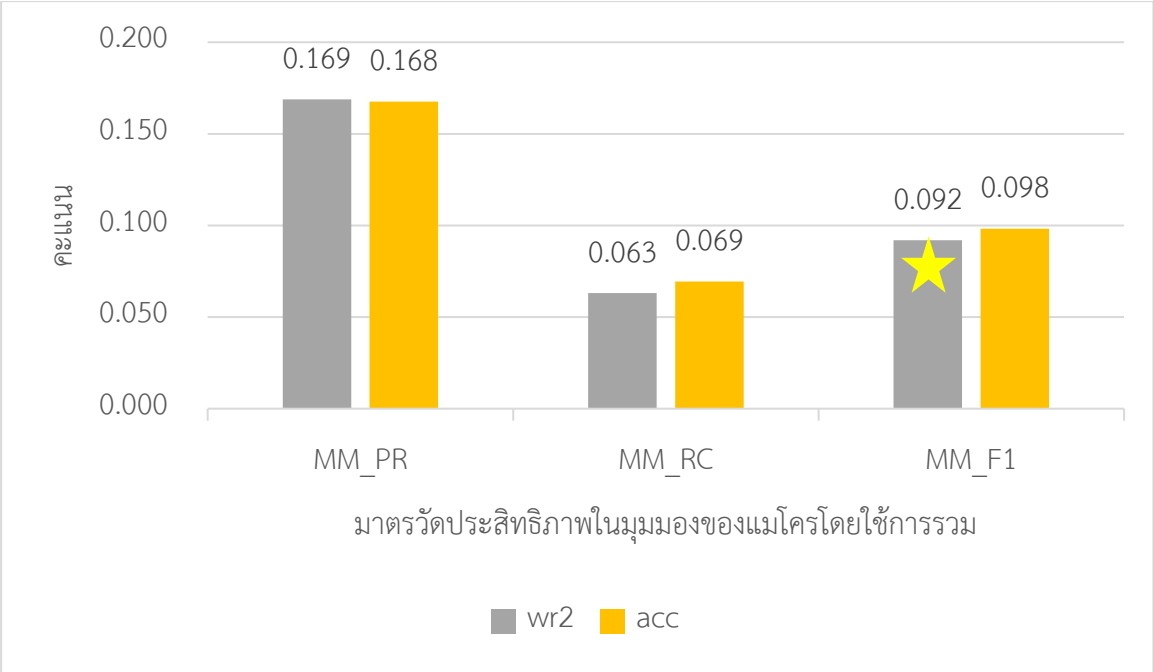
- จากข้อมูลตรวจสอบ พบว่าตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองทั่วไป rf_d และ ตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf มีค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ที่ให้ประสิทธิภาพดีที่สุด เท่ากับ 0.96 และ 0.83 ตามลำดับ
- รูปที่ 57 แสดงจำนวนหัวข้อที่ทำนายได้จากทั้ง 2 วิธี
    - พบว่าจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf ทำนายออกมา 27 หัวข้อ และคล้ายกับหัวข้อในชุดผลเฉลยถึง 21 หัวข้อ กล่าวคือทำนายเกิน 6 หัวข้อ
    - ในขณะที่จำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf_d ทำนายออกมา 49 หัวข้อ และคล้ายกับหัวข้อในชุดผลเฉลยเพียง 30 หัวข้อ กล่าวคือทำนายเกิน 19 หัวข้อ

- สาเหตุที่ทำให้ rf_d ตรวจจับหัวข้อได้เกินลดลงอย่างมาก เมื่อเทียบกับการทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม (Macro merge หรือ MM) เนื่องจากเมื่อวัดผลโดยใช้การรวมแทนการหาใช้ค่าเฉลี่ย ทำให้เส้นแบ่งความคล้ายที่ดีที่สุดของ rf_d เปลี่ยนจาก 0.9 เป็น 0.96 ทำให้คำที่ถูกทำนายเกินหาคู่ของคำที่คล้ายกันยาก และถูกคัดกรองออกไปเนื่องจากไม่สามารถจับกลุ่มเป็นหัวข้อได้

- จากข้อมูลทดสอบในรูปที่ 58 พบว่าจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf ให้ประสิทธิภาพดีกว่าทั้งในด้านความแม่นยำและความครอบคลุม



รูปที่ 57 ผลการทดลองแสดงจำนวนหัวข้อที่ทำนายได้ และจำนวนหัวข้อที่ทำนายได้และคล้ายกับหัวข้อในชุดผลเฉลย ของการทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำโดยใช้การรวม



รูปที่ 58 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม

ของการทดลองประสิทธิภาพของตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำ

4. สรุปผลการทดลอง

ตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf มีประสิทธิภาพดีกว่าตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพทั่วไป rf_d ทั้งในมาตรวัดประสิทธิภาพในมุมมองของหัวข้อโดยใช้การค่าเฉลี่ย (Macro average หรือ MA) และ มาตรวัดประสิทธิภาพในมุมมองของหัวข้อโดยใช้การรวม (Macro merge หรือ MM)

6.2.2 การทดลองประสิทธิภาพของการตรวจจับคำและหัวข้อเกิดใหม่

1. วัตถุประสงค์ของการทดลอง

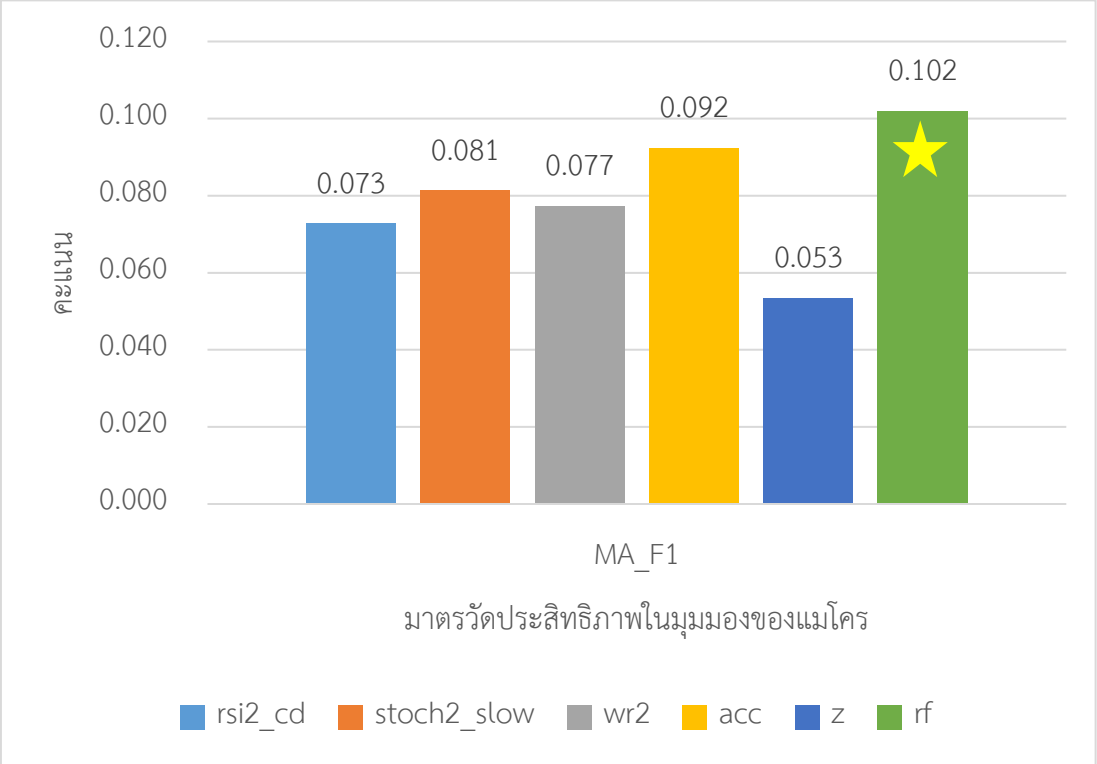
- ทดลองประสิทธิภาพของตัวจำแนกประเภท เทียบกับคุณลักษณะที่นำเสนอในงานวิจัยนี้และคุณลักษณะของงานวิจัยอื่น
  - ตัวจำแนกประเภทป่าไม้แบบสุ่มที่สร้างจากมาตรวัดในมุมมองของคำ rf
  - คุณลักษณะที่นำเสนอ $rsi2\_cd$, $stoch2\_slow$ และ $wr2$
  - คุณลักษณะของงานวิจัยอื่น acc, z
- จับกลุ่มคำที่ถูกทำนายว่าเป็นคำเกิดใหม่ เพื่อสร้างหัวข้อเกิดใหม่ด้วย LSI
- จากผลการทดสอบหาค่ากรอบเวลา w ที่ดีที่สุดของการเลือกข้อความทวิตเตอร์ w นาทีย้อนหลังในการเรียนรู้ของอัลกอริทึม LSI ได้แก่ 1, 5, 10, 15 และ 30 ในภาคผนวก ก พบว่าค่ากรอบเวลา w ที่ดีที่สุดคือ 30 ดังนั้นในการทดลองนี้จะใช้ข้อความทวิตเตอร์ย้อนหลัง 30 นาทีในการหาความสัมพันธ์ระหว่างคู่ของคำด้วยอัลกอริทึม LSI
- ทดสอบหาค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ในการคัดเลือกคู่ของคำที่มีความเกี่ยวข้องกันเป็นหัวข้อเดียวกัน ได้แก่ 0.1, 0.11, 0.12, ..., 0.97, 0.98 และ 0.99
- แบ่งเป็น 2 การทดลองย่อย ได้แก่
  - การทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro average หรือ MA) : ทำการตรวจสอบหาค่าพารามิเตอร์ที่ดีที่สุดโดยใช้ MA และทำการทดสอบประสิทธิภาพด้วย MA
  - การทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม (Macro merge หรือ MM) : ทำการตรวจสอบหาค่าพารามิเตอร์ที่ดีที่สุดโดยใช้ MM และทำการทดสอบประสิทธิภาพด้วย MM

2. ผลการทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro average หรือ MA)

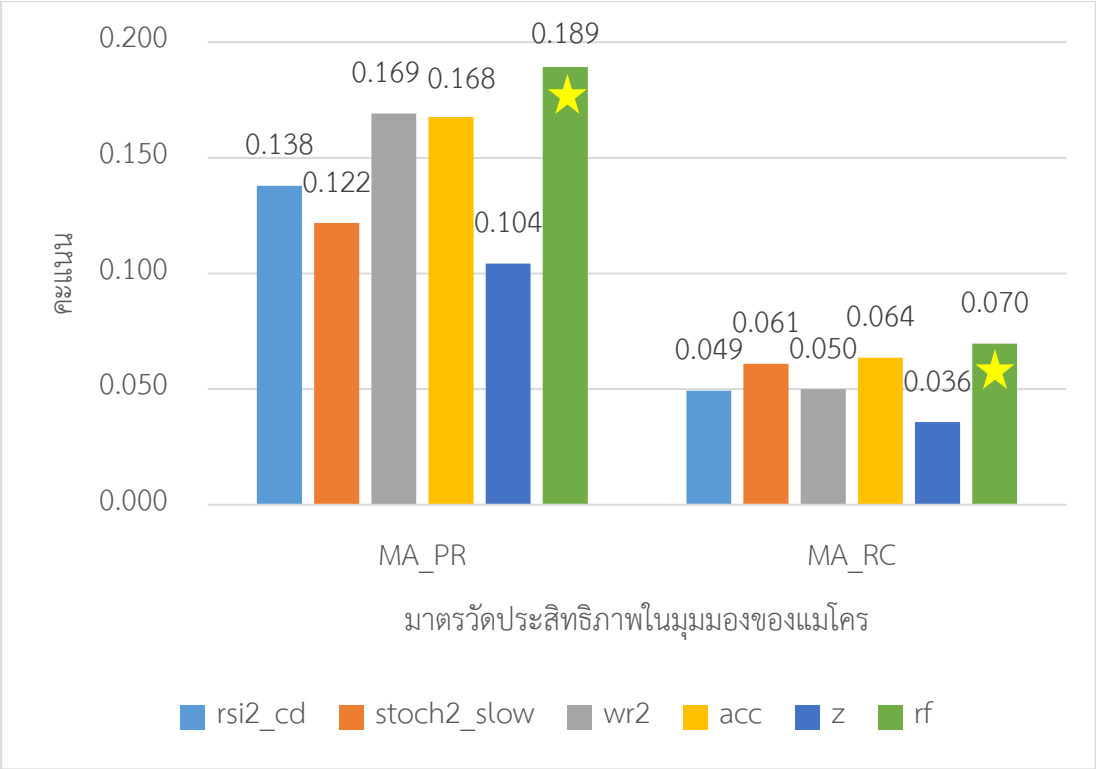
- จากข้อมูลตรวจสอบ พบว่าค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ที่ให้ประสิทธิภาพดีที่สุดของแต่ละวิธีมีค่า ดังนี้
  - ตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำ rf มีค่าเส้นแบ่งเท่ากับ 0.89
  - คุณลักษณะที่นำเสนอ $rsi2\_cd$ มีค่าเส้นแบ่งเท่ากับ 0.63

- คุณลักษณะที่นำเสนอ $stoch2\_slow$ มีค่าเส้นแบ่งเท่ากับ 0.83
- คุณลักษณะที่นำเสนอ $wr2$ มีค่าเส้นแบ่งเท่ากับ 0.76
- คุณลักษณะของงานวิจัยอื่น acc มีค่าเส้นแบ่งเท่ากับ 0.49
- คุณลักษณะของงานวิจัยอื่น z มีค่าเส้นแบ่งเท่ากับ 0.44

- จากข้อมูลทดสอบในรูปที่ 59 และ รูปที่ 60 พบว่าตัวจำแนกประเภท rf มีประสิทธิภาพในมุมมองของหัวข้อโดยใช้ค่าเฉลี่ยดีที่สุดทั้งความแม่นยำและความครอบคลุม



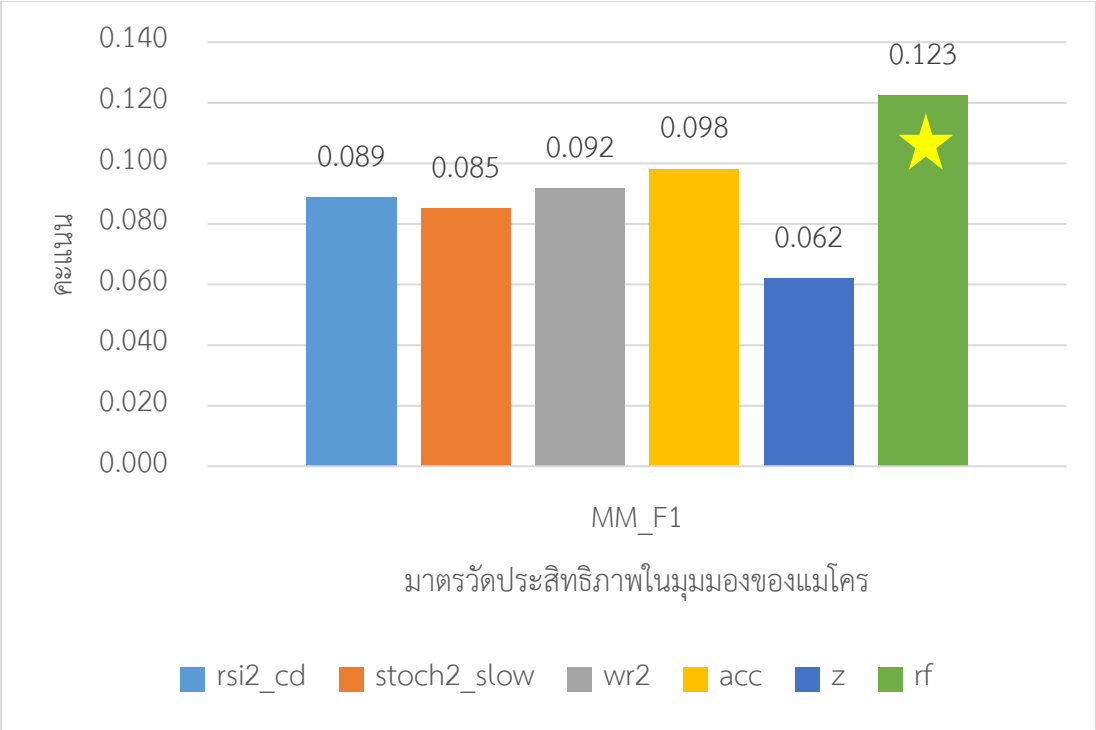
รูปที่ 59 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย F1 ของการทดลองประสิทธิภาพของการตรวจจับคำและหัวข้อเกิดใหม่



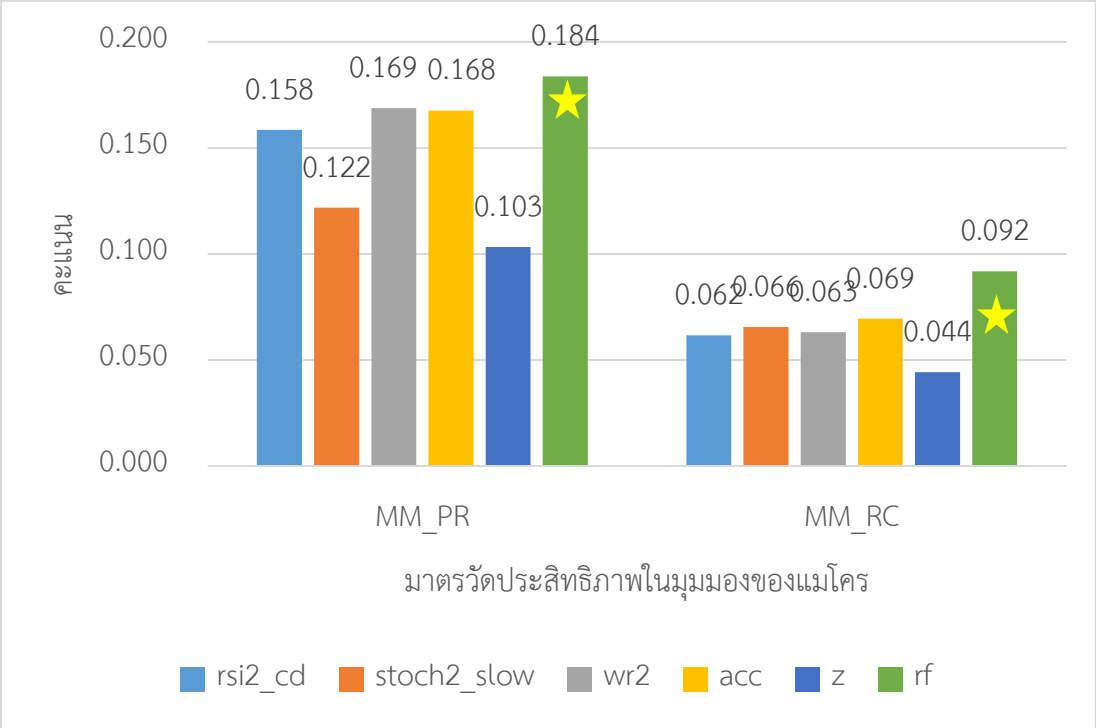
รูปที่ 60 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย Pr และ Rc ของการทดลองประสิทธิภาพของการตรวจจับคำและหัวข้อเกิดใหม่

3. ผลการทดลองเมื่อใช้ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้ค่าเฉลี่ย (Macro merge หรือ MM)

- จากข้อมูลตรวจสอบ พบว่าค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ที่ให้ประสิทธิภาพดีที่สุดของแต่ละวิธีมีค่า ดังนี้
  - ตัวจำแนกประเภทที่สร้างจากมาตรวัดในมุมมองของคำ rf มีค่าเส้นแบ่งเท่ากับ 0.83
  - คุณลักษณะที่นำเสนอ $rsi2\_cd$ มีค่าเส้นแบ่งเท่ากับ 0.79
  - คุณลักษณะที่นำเสนอ $stoch2\_slow$ มีค่าเส้นแบ่งเท่ากับ 0.83
  - คุณลักษณะที่นำเสนอ $wr2$ มีค่าเส้นแบ่งเท่ากับ 0.76
  - คุณลักษณะของงานวิจัยอื่น acc มีค่าเส้นแบ่งเท่ากับ 0.49
  - คุณลักษณะของงานวิจัยอื่น z มีค่าเส้นแบ่งเท่ากับ 0.44
- จากข้อมูลทดสอบในรูปที่ 61 และ รูปที่ 62 พบว่าตัวจำแนกประเภท rf มีประสิทธิภาพในมุมมองของหัวข้อโดยใช้การรวมดีที่สุดทั้งความแม่นยำและความครอบคลุม



รูปที่ 61 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม F1 ของการทดลองประสิทธิภาพของการตรวจจับคำและหัวข้อเกิดใหม่



รูปที่ 62 ผลการทดลองแสดงมาตรวัดประสิทธิภาพในมุมมองของหัวข้อแบบแมโครโดยใช้การรวม Pr และ Rc ของการทดลองประสิทธิภาพของการตรวจจับคำและหัวข้อเกิดใหม่

4. สรุปผลการทดลอง

ตัวจำแนกประเภทที่สร้างจากมาตรวัดประสิทธิภาพในมุมมองของคำ rf มีประสิทธิภาพดีกว่าวิธีการอื่นทั้งในมาตรวัดประสิทธิภาพในมุมมองของหัวข้อโดยใช้การค่าเฉลี่ย (Macro average หรือ MA) และ มาตรวัดประสิทธิภาพในมุมมองของหัวข้อโดยใช้การรวม (Macro merge หรือ MM)

6.2.3 การทดลองความเร็วในการตรวจจับหัวข้อเกิดใหม่

1. วัตถุประสงค์ของการทดลอง

- ทดลองวัดความเร็วในการตรวจจับหัวข้อแต่ละหัวข้อที่อยู่ในชุดผลเฉลย ของตัวจำแนกประเภทเทียบกับแต่ละวิธีอื่น ๆ
    - คุณลักษณะที่นำเสนอ $rsi2\_cd$, $stoch2\_slow$ และ $wr2$
    - คุณลักษณะของงานวิจัยอื่น acc, z
- จับกลุ่มคำที่ถูกทำนายว่าเป็นคำเกิดใหม่ เพื่อสร้างหัวข้อเกิดใหม่ด้วย LSI
- ใช้ค่ากรอบเวลา w และเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ของแต่ละวิธีที่ได้ประสิทธิภาพดีที่สุดในการทดลองที่ 6.2.2 ในมาตรวัดประสิทธิภาพในมุมมองของหัวข้อโดยใช้ค่าเฉลี่ย (Macro average หรือ MA)
- ทำการวัดความเร็วในการตรวจจับหัวข้อเกิดใหม่ โดยดูจากคุณลักษณะต่อไปนี้
    - จำนวนหัวข้อที่ rf ตรวจจับได้เร็วกว่า
    - จำนวนหัวข้อที่ rf ตรวจจับได้ช้ากว่า
    - จำนวนหัวข้อที่ rf ตรวจจับไม่ได้ หรือตรวจจับได้พร้อมกัน
- ทำการตรวจจับหัวข้อเกิดใหม่ด้วย LSI

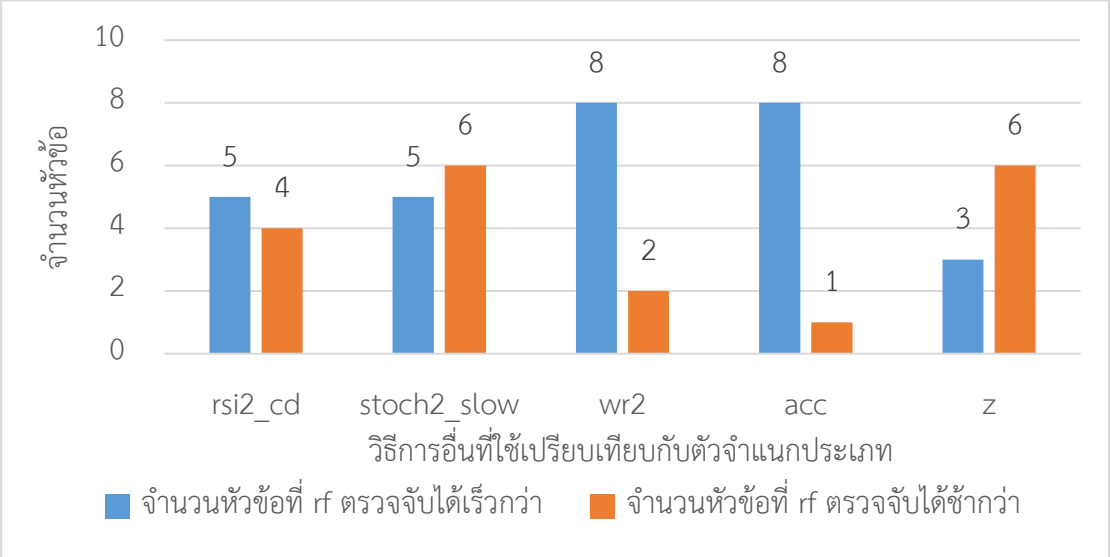
2. ผลการทดลอง

- จากรูปที่ 63 และ รูปที่ 64 เมื่อเปรียบเทียบความเร็วในการตรวจจับหัวข้อของตัวแนกประเภท rf กับคุณลักษณะ acc พบว่า
    - rf ตรวจจับหัวข้อได้เร็วกว่า 8 หัวข้อ และตรวจจับหัวข้อที่ acc ตรวจจับไม่ได้ 4 หัวข้อ รวมเป็น 12 หัวข้อ
    - acc ตรวจจับหัวข้อได้เร็วกว่า 1 หัวข้อ และตรวจจับหัวข้อที่ rf ตรวจจับไม่ได้ 2 หัวข้อ รวมเป็น 3 หัวข้อ

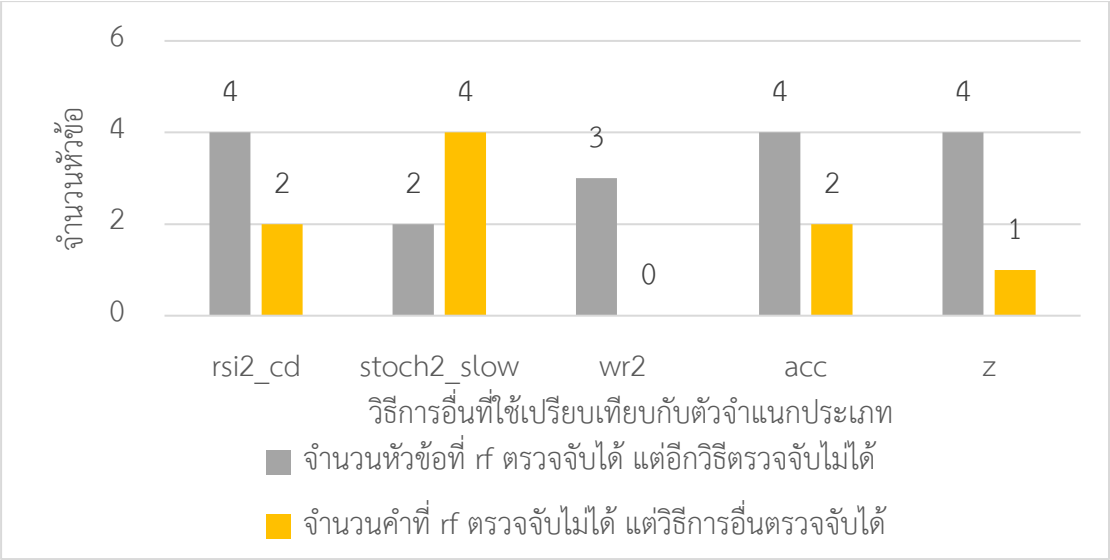
- โดยสรุป ตัวแนกประเภท rf มีความเร็วในการตรวจจับหัวข้อเร็วกว่า acc

- จากรูปที่ 63 และ รูปที่ 64 เมื่อเปรียบเทียบความเร็วในการตรวจจับหัวข้อของตัวแนกประเภท rf กับคุณลักษณะ z พบว่า
  - rf ตรวจจับหัวข้อได้เร็วกว่า 3 หัวข้อ และตรวจจับหัวข้อที่ z ตรวจจับไม่ได้ 4 หัวข้อ รวมเป็น 7 หัวข้อ
  - z ตรวจจับหัวข้อได้เร็วกว่า 6 หัวข้อ และตรวจจับหัวข้อที่ rf ตรวจจับไม่ได้ 1 หัวข้อ รวมเป็น 7 หัวข้อ
  - โดยสรุป ตัวแนกประเภท rf กับ z มีความเร็วในการตรวจจับหัวข้อใกล้เคียงกัน

- จากรูปที่ 63 และ รูปที่ 64 เมื่อเปรียบเทียบความเร็วในการตรวจจับหัวข้อของตัวแนกประเภท rf กับคุณลักษณะที่นำเสนอในงานวิจัยนี้ $rsi2\_cd$, $stoch2\_slow$ และ $wr2$ พบว่า rf ตรวจจับหัวข้อได้ช้ากว่าคุณลักษณะ $stoch2\_slow$ อย่างไรก็ตามประสิทธิภาพในมุมมองของหัวข้อแบบแมโครต่ำกว่าตัวจำแนกประเภท rf มาก จากการทดลองที่ 6.2.2



รูปที่ 63 ผลการทดลองแสดงจำนวนหัวข้อที่ rf ตรวจจับได้เร็วกว่าและช้ากว่าเมื่อเทียบกับวิธีการอื่นของการทดลองความเร็วในการตรวจจับหัวข้อเกิดใหม่



รูปที่ 64 ผลการทดลองแสดงจำนวนหัวข้อที่ rf ตรวจจับได้แต่อีกวิธีตรวจจับไม่ได้ และ จำนวนหัวข้อที่ rf ตรวจจับไม่ได้แต่วิธีการอื่นตรวจจับได้ ของการทดลองความเร็วในการตรวจจับหัวข้อเกิดใหม่

3. สรุปผลการทดลอง

ตัวจำแนกประเภท rf สามารถตรวจจับหัวข้อแต่ละหัวข้อได้เร็วกว่าเกือบทุกวิธี อย่างไรก็ตาม rf มีประสิทธิภาพในมุมมองของหัวข้อแบบแมโครดีกว่ามาก

6.3 การวิเคราะห์ข้อมูลเพิ่มเติม (Other Experiment and Result)

การวิเคราะห์ข้อมูลเพิ่มเติม แบ่งออกเป็น 3 การวิเคราะห์ ได้แก่

1. การวิเคราะห์ผลลัพธ์การตรวจจับคำและหัวข้อเกิดใหม่ของตัวจำแนกประเภท
2. คำเกิดใหม่ที่ตัวจำแนกประเภทตรวจจับได้แต่วิธีการอื่นตรวจจับไม่ได้
3. ความสำคัญของแต่ละคุณลักษณะในตัวจำแนกประเภทที่ดีที่สุด
4. ข้อความทวิตเตอร์ที่ใช้ในการปั้มโฆษณา

6.3.1 การวิเคราะห์ผลลัพธ์การตรวจจับคำและหัวข้อเกิดใหม่ของตัวจำแนกประเภท

จากหัวข้อที่ถูกทำนายของข้อมูลทดสอบ ในวันที่ 19 มิถุนายน ค.ศ. 2018 ถึง 20 มิถุนายน ค.ศ. 2018 ของตัวจำแนกประเภท rf ในตารางที่ 48 และตารางที่ 49 พบว่าคำที่ตรวจจับเกินเกือบทั้งหมดเป็นคำที่น่าสนใจและเกี่ยวข้องกับหัวข้อเกิดใหม่ แต่ไม่ปรากฏในชุดผลเฉลย ดังตัวอย่างต่อไปนี้

หัวข้อ “dafbama2018got7” ในวันที่ 19 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับรายการประกวดเพลงในเกาหลี โดยการให้แฟน ๆ โหวตให้นักร้องที่ชื่อชอบผ่านทวิตเตอร์
- มีคำที่น่าสนใจคือ “ป้าย” และ “ครีม”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 20:49:00 ถึง 2018-06-19 21:03:00
    - เกิดจากนักร้องคนหนึ่งที่เป็นที่นิยมมากโดนครีมป้ายในรายการทีวี ซึ่งนักร้องคนดังกล่าวอยู่ในวง GOT7 ที่กำลังร่วมแข่งรายการเพลง dafbama2018 ทำให้หัวข้อเกิดใหม่ที่ทำนายได้ “dafbama2018” ปรากฏคำว่า “ป้าย” และ “ครีม” ซึ่งเป็นคำที่ไม่เกี่ยวข้องกับหัวข้อนี้

ตารางที่ 38 ตัวอย่างข้อความทวิตเตอร์ของคำ “ป้าย”, “ครีม” ของหัวข้อ “dafbama2018got7” ในวันที่ 19 มิถุนายน ค.ศ. 2018

แก ถามหน่อยคนเราโดน**ครีมป้าย**หน้าคือมันต้องมีความสุขขนาดนี้เลยหรอวะ จิงๆหรอแก? #MarkBam https://t.co/vcSdqaZv02 ตอน**ป้ายครีม**ตรงคางก็จะงับมือน้องไง เลยโดนป้าบเข้าที่แก้ม แต่หาได้โกรธไม่ ยิ้มหวานละเกิ๊นนพ่อคุณ รอบสองน้องป้ายอีก โดนปากพี่ถึงกับสะดุ้ง หึหึ ถ้าเป็นในฟิคพี่ก็ดึงมือมากินเค้กตรงนิ้วไปแล้ว ความพยายามจะกิน**ครีม**จากมือน้องของพี่เค้า โดนน้องเอา**ครีมป้าย**หน้าซะเลย (กราบขอบพระคุณเจ้าของคลิปค่า 🙏🙏🙏) #MarkBam #GOT7 #GOT7WORLDTOUR #GOT7TheNewEra https://t.co/wT9SjbrabL @lovebambama1 น้อง**ป้ายครีม**โดนปากพี่สะดุ้งเลย 😆

- อีกคำที่น่าสนใจคือ “แบตหมดจ้า”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 00:22:00 ถึง 2018-06-19 00:36:00
    - เกิดจากการปั้มข้อความทวิตเตอร์ของผู้ใช้งานคนหนึ่ง ที่ปั้มข้อความเดิมถึง 20 ครั้งต่อนาที เพื่อโหวตคะแนนให้นักร้องเกาหลีที่ชื่นชอบ

ตารางที่ 39 ตัวอย่างข้อความทวิตเตอร์ของคำ “แบตหมดจ้า” ของหัวข้อ “dafbama2018got7” ในวันที่ 19 มิถุนายน ค.ศ. 2018

แบตหมดจ้า เพิ่งกู้มาได้10% #GOT7WORLDTOUR #갓세븐 #GOT7 #EyesOnYou #dafBAMA2018Got7 #GOT7WorldTour #BringGOT7toGDNY #GDNY @GOT7Official #BringGOT7onTheEllenShow @TheEllenShow #BringGOT7onTheLatelateShow @latelateshow

หัวข้อ “oreoandritzwaferroll” ในวันที่ 19 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับมินิคอนเสิร์ตของเป๊กผลิตโชคในงานเปิดตัวโอรีโอ และ ริทซ์ เวเฟอร์โรล
- มีคำที่น่าสนใจคือ “นุช”
    - เป็นคำที่นักน้องชื่อดัง เป๊กผลิตโชค ใช้เรียกแฟน ๆ ดังนั้น เวลาแฟน ๆ พูดถึงตัวเองในข้อความทวิตเตอร์ จะแทนตัวเองว่า “นุช” ทำให้หัวข้อที่เกี่ยวกับ เป๊กผลิตโชค ทุกหัวข้อ มักจะปรากฏคำว่า “นุช” เสมอ

ตารางที่ 40 ตัวอย่างข้อความทวิตเตอร์ของคำ “นุช” ของหัวข้อ “oreoandritzwaferroll” ในวันที่ 19 มิถุนายน ค.ศ. 2018

@chayut_piyawan โชคดีที่สามีให้ไปนะคะ สู้ๆเพื่อพี่โชค พี่โชคผู้เปลี่ยนชีวิต**นุช** 555 #PeckPalitchoke #เป๊กผลิตโชค #dafBAMA2018PeckPalitchoke #dafBAMA2018
@nookewwwww น่าร้ากกกกที่สุดคนนี้ จริงใจกะคนรอบข้าง #dafBAMA2018PeckPalitchoke #เป๊กผลิตโชค #Peckpalitchoke ถ้าจะให้อธิบายความน่ารักของเฮีย...**นุช**ขอกระดาษซักรีมค่ะ☺
@stfcfafah เราเป็น**นุช**หน้าจอ ยุ ตจว.จ้า ทัพหน้าซู่ๆๆ ส่งกลจ.ไปให้ทุกคน😍😍😍

รูปที่ 65 มินิคอนเสิร์ตของเป๊กผลิตโชคในงานเปิดตัวโอรีโอ และ ริทซ์ เวเฟอร์โรล วันที่ 19 มิถุนายน ค.ศ. 2018 ณ ลาน Eden ชั้น G เซ็นทรัลเวิล์ด [32]

[32] อ้างอิงจาก https://www.mypalit.net/article/1266

หัวข้อ “smtmthailand” ในวันที่ 19 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับรายการทีวี “Show Me The Money Thailand”
- มีคำที่น่าสนใจคือ “นายนะ” และ “พี่อุ๋ย”
    - “นายนะ” ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 23:00:00 ถึง 2018-06-19 23:14:00
    - “พี่อุ๋ย” ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 22:29:00 ถึง 2018-06-19 22:43:00
    - นายนะและพี่อุ๋ยเป็นชื่อคนที่อยู่ในรายการ
    - เป็นตัวอย่างคู่ของคำที่ดี เนื่องจากคำว่า “นายนะ” และ “พี่อุ๋ย” ปรากฏในข้อความเดียวกันในวันที่ 19 มิถุนายน ค.ศ. 2018 จำนวน 2 ข้อความเท่านั้น แต่คำว่า “นายนะ” มีความสัมพันธ์กับ “smtmthailand” และคำว่า “พี่อุ๋ย” มีความสัมพันธ์กับ “smtmthailand” ดังนั้น 2 คำนี้ถึงถูกรวมเข้ามาเป็นหัวข้อเดียวกันได้

ตารางที่ 41 ตัวอย่างข้อความทวิตเตอร์ของคำ “นายนะ” และ “พี่อุ๋ย” ของหัวข้อ “smtmthailand” ในวันที่ 19 มิถุนายน ค.ศ. 2018

ตลก**นายนะ**55555 ใครควรออก วดฟ เพราะเค้าโดนด่าเยอะ #smtmthailand
นั่นไง **นายนะ**มันร้ายยย ท่านผู้ชม ดูทำมาเปงแซะนิ่มๆ หน้านิ่งๆ #smtmthailand
ขำ**พี่อุ๋ย**ที่บอกว่า เห่ยกูแรปรัวว่ะ แรปเร็วว่ะ แต่คนฟังไม่รู้เรื่องเลย555555555 #SMTMThailand
ตอนแรกนึกว่าเป็นที่ห้องอัดอีกแล้ว กูเลยฟังไม่รู้เรื่อง พอ**พี่อุ๋ย**บอกฟังไม่รู้เรื่อง อ้าว ไม่เกี่ยวกับอัดแล่ววววว #SMTMThailand

หัวข้อ “บอลโลก” ในวันที่ 19 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับการแข่งขันฟุตบอลโลก
- มีคำที่น่าสนใจคือ “แดง”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 18:52:00 ถึง 2018-06-19 19:06:00
    - เกิดจากการแจกใบแดงให้นักกีฬาโคลัมเบีย นัดที่แข่งกับทีมญี่ปุ่น โดยเป็นการโดนใบแดงเร็วเป็นอันดับ 2 ในประวัติศาสตร์ฟุตบอลโลก โดย ซานเชซ ถูกใบแดงไล่ออกในนาทีที่ 3 จากการใช้แขนป้องกันลูกยิงในเขตโทษ

ตารางที่ 42 ตัวอย่างข้อความทวิตเตอร์ของคำ “แดง” ของหัวข้อ “บอลโลก” ในวันที่ 19 มิถุนายน ค.ศ. 2018

ไอเหี้ย แปปๆใบ**แดง**แร้ว... #สวัสดีบอลโลก
เสียดายโรดริเกวซไม่ได้ลง ตายแล้วโคลัมเบียใบ**แดง** #สวัสดีบอลโลก
เล่นไม่กี่นาทีก็ใบ**แดง** +จุดโทษแล้ว เย้ๆๆๆๆ **JPJPJPJPJPJPJP** #สวัสดีบอลโลก
ใบ**แดง**เลยว่ะ โคลัมเบีย เสียจุดโทษ + ใบแดง เหลือสิบคน ตั้งแต่สามนาทีแรก! เอาว่ะ ญี่ปุ่นต้องเอาหน่อย ต้องชนะนะ #สวัสดีบอลโลก

หัวข้อ “ลิขิตรักthecrownprincess” ในวันที่ 19 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับละครไทยเรื่องหนึ่ง
- มีคำที่น่าสนใจคือ “เคท”, “ฟิน”, “วัง”, “แผน”, “หัวใจ”, “พัน”, “สอง”, “เจ้าหญิงเคธ”, “ฉาก”, และ “เคธ”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 20:49:00 ถึง 2018-06-19 21:03:00
    - เป็นตอนจบของละครไทยที่กำลังเป็นที่นิยมมากในช่วงเวลานั้น ซึ่งแฟน ๆ ฟินกับฉากหวานของคู่พระนาง "ณเดชน์-ญาญ่า” ทำให้ตอนจบนั้นของละครเรื่องนี้นั้นมีเรตติ้งสูงถึง 6.2
- มีอีกคำที่น่าสนใจคือ “the rapper” ซึ่งไม่ควรจะถูกจับกลุ่มรวมกับหัวข้อนี้ แต่ถูกรวมเข้าด้วยกันเนื่องจากมีคำว่า “ฉาก” ร่วมกัน

ตารางที่ 43 ตัวอย่างข้อความทวิตเตอร์ของคำ “เคท”, “ฟิน”, “วัง”, “แผน”, “หัวใจ”, “พัน”, “สอง”, “เจ้าหญิงเคธ”, “ฉาก”, และ “เคธ” ของหัวข้อ “ลิขิตรักthecrownprincess” ในวันที่ 19 มิถุนายน ค.ศ. 2018

สงสัยอ่ะว่าทำไม**เคท**ถึงถูกคนในครอบครัวลืมได้ขนาดนี้ อยากรู้จริงนี้มากอ่ะ #ลิขิตรักTheCrownPrincess
ดิฉันไม่ไหวแล้วแต่งเลยเจ้าค่ะแต่งงงง ขออภัยในเสียงของหม่อมฉันด้วย**ฟิน**และหน่วงมากในเวลาเดียวกันนั้น ❤❤❤#ลิขิตรักTheCrownPrincess https://t.co/U5Bvd3cFaD
ความพยายามของติ่ง ในการลบแสงไฟฟอร์จูนเนอร์😂😂 #ใครขับฟอร์จูนเนอร์เข้า**วัง** #ขนลุก #ลิขิตรักthecrownprincess https://t.co/nm99aJilFs
เมื่อแผนไม่เป็นไปตามแผน...**เคธ**จึงต้องสกัดกั้น งานแต่งของวิลและอลิซทุกวิธี... ชม #ลิขิตรักTheCrownPrincess ตอน 11 ย้อนหลังก่อนใครได้แล้วที่ ดูเลย ☞ https://t.co/ujsw3V3sl7 #MelloThailand #Ch3 #ละครช่อง3 #ลิขิตรัก https://t.co/rOZZnDz8US
เจ้าหญิงอลิซ จะจัดการกับหน้าที่และ**หัวใจ**ของตัวเองอย่างไร #ลิขิตรักTheCrownPrincess EP.11

ดูย้อนหลังได้แล้วตอนนี้ที่ https://t.co/uLP3L1dAcC #Ch3Thailand #mello https://t.co/OkYue4njME

อยากซื้อ**ฉาก**ทั้งหมดของ #ลิขิตรักTheCrownPrincess อยากจะดูทั้งหมดของ**ฉาก**ที่ตัดไป อยากดูความอีโรติกที่ไม่เห็นในละครของผู้**พัน**กับเจ้าหญิง ขอมากไปมั้ย #งอแงวนไป

ชอบซีนนี้ที่สุด วันนี้ได้เห็นฝีมือการแสดงของ #ณเดชน์ญาญ่า ทั้ง**สอง**พัฒนาไปไกลมากๆทำให้เชื่อในความรักของตัวละครจริงๆ ทุกอย่างพอดีไม่มากไม่น้อย ดูละครแบบไม่คาดหวังทำสมองกลวงๆ ดูการนำเสนอของผู้จัดและผกก. มันดีมากๆ น้อยแต่มาก จากอินเนอร์ของนักแสดง 👏👏 #ลิขิตรัก TheCrownPrincess https://t.co/0JbEmpE1m9

รูปที่ 66 เรตติ้งของลิขิตรัก The Crown Princess วันที่ 19 มิถุนายน 2018 [33]

[33] อ้างอิงจาก https://pantip.com/topic/37790245

หัวข้อ “สายรักสายสวาท” ในวันที่ 19 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับละครไทยเรื่องหนึ่ง
- มีคำที่น่าสนใจคือ “สงสาร” และ “โฉม”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 21:13:00 ถึง 2018-06-19 21:27:00
    - โฉม เป็นชื่อของตัวละครในละครเรื่องนี้

ตารางที่ 44 ตัวอย่างข้อความทวิตเตอร์ของคำ “สงสาร” และ “โฉม” ของหัวข้อ “สายรักสายสวาท” ในวันที่ 19 มิถุนายน ค.ศ. 2018

โฉมนางก็ไม่ได้ร้ายตั้งแต่แรกแล้วป่ะ คือ**สงสาร**นางมาก จากคนที่หม่อมย่ามาขอแทบจะถวายหัวให้ไปเป็นลูกสะใภ้ แต่ตอนนี้กลับแทบจะถีบส่ง แล้วยังต้องมาเจอความร้ายของยัยโสมอีก เห้อออออ #สายรักสายสวาท
สงสาร**โฉม**สำอางค์ นางควรรีบอย่าขาดจากยิ่งศักดิ์ #สายรักสายสวาท
อหสงสารหญิง**โฉม**อะทำดีแค่ไหนแต่เค้าไม่เห็นค่าไรสักอย่าง #สายรักสายสวาท

หัวข้อ “บอลโลก” ในวันที่ 20 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับการแข่งขันฟุตบอลโลก
- มีคำที่น่าสนใจคือ “ชนะ”, “เมสซี่”, “โด้”, “พัง” และ “ประตู”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-20 19:39:00 ถึง 2018-06-20 19:53:00
    - นักกีฬาฟุตบอล “เมสซี่” แม้ไม่ได้มีแข่งในวันนี้ แต่มีการพาดพิงถึงบ่อยครั้ง

ตารางที่ 45 ตัวอย่างข้อความทวิตเตอร์ของคำ “เมสซี่” และ “โด้” ของหัวข้อ “โด้” ในวันที่ 20 มิถุนายน ค.ศ. 2018

@kiddevil16 โค้ชโปรตุเกสมันให้โด้แบกก็จริง แต่ให้แบกแบบมีระบบอะ เพื่อนต้องไปกับโด้มันด้วย แล้วเหมือนกับ**โด้**มันเล่นเพื่อทีมมากขึ้น มันเชื่อมเกมส์กับเพื่อนมากขึ้น ภาพตัดไปที่อาเจนฯ แม่งระบบ**เมสซี่**ชัดๆ อะไรๆก็ให้เมสซี่ มองหาเมสซี่ก่อนเลย ซึ่งคู่แข่งมันจับทางได้ง่ายกว่า อาเจนฯเป็นแบบนี้ประจำ สำหรับเรา เนย์มาร์ Overrated ถ้าวันนึงเนย์จะได้บัลลงดอร์ก็เป็นเพราะ**เมสซี่**กับโรนัล**โด้**ไม่อยู่แล้ว และเนย์มาร์ก็มีชื่อเสียงกับภาพจำมากพอว่าเป็นนักเตะที่ควรได้บัลลงดอร์ ทั้งๆที่ฝีเท้าก็ไม่ได้ขนาดนั้น **เมสซี่**อาจจะไม่ใช่นักเตะที่มีพรสวรรค์สูงที่สุดในประวัติศาสตร์ แต่ไอ้**โด้**คือนักเตะที่มีพรแสวงสูงที่สุดในประวัติศาสตร์แน่นอน https://t.co/8snsly3Ude ในระหว่างที่ฝรั่งกำลังเถียงกันว่าระหว่าง**เมสซี่** ซาล่าห์ โรนัล**โด้**ใครคือนักบอลที่เก่งที่สุดในโลก กูก็นั่งเครียดว่านัดต่อไปเยอรมนีจะชนะสวีเดนไหม😓😓😓😓😓😓😓😓😓😓😓😓😓😓

หัวข้อ “หนึ่งด้าวฟ้าเดียว” ในวันที่ 20 มิถุนายน ค.ศ. 2018

- เป็นหัวข้อเกี่ยวกับละครไทยเรื่องหนึ่ง
- มีคำที่น่าสนใจคือ “เจ้าตัว”, “เตรียม”, “ฟื้น”, “แมง”, “เจ้า”, “พระ”, “เม่า”, “ฉาก”, “ศรีเรือน”, “พี่ขันทอง”, “หมาก”, “ช้าง”, “สงสาร” และ “พี่แต้ว”
    - ปรากฏเป็นคำเกิดใหม่ครั้งแรกในข้อความทวิตเตอร์ในช่วงเวลา 2018-06-19 20:49:00 ถึง 2018-06-19 21:03:00
    - เป็นตอนจบของละครหนึ่งด้าวฟ้าเดียว

ตารางที่ 46 ตัวอย่างข้อความทวิตเตอร์ของคำ “แมง”, “ขันทอง”, “พี่แต้ว” และ “ช้าง” ของหัวข้อ “หนึ่งด้าวฟ้าเดียว” ในวันที่ 20 มิถุนายน ค.ศ. 2018

ไปอ่านตอนจบละคร #หนึ่งด้าวฟ้าเดียว ร้องไห้ไปกับเจ้า**แมง**เม่า อ่านไปเรื่อยๆ ยิ้มไปกับพ่อ**ขันทอง** และแม่**แมง**เม่า ถึงตอนสุดท้ายเราก็ยังไบโพล่าได้อีก😊 พรุ่งนี้ดูสดเราก็คงร้องไห้และยิ้มอีกเป็นแน่ รักทุกตัวละครในเรื่องนี้จัง มีความสุขใจยิ่งนัก 💘 #หนึ่งด้าวฟ้าเดียว
#หนึ่งด้าวฟ้าเดียว เจมส์เรื่องนี้เก่งมาก เล่นดีมาก ส่วน**พี่แต้ว**ไม่ต้องพูดถึงอาจารย์อยู่แล้ว
คืนนี้นอกจากเข้าหอแล้ว ซีน**แมง**เม่าจะยกขันหมากไปขอ**พี่ขันทอง**นี่ล่ะที่เรารอ55555555 #หนึ่งด้าวฟ้าเดียว https://t.co/ugnAn0w6my
ช้างศึก"แม่พังคีรีบัญชร"จะได้ทำหน้าที่สำคัญใน ปวศ.เพื่อร่วมกอบกู้เอกราชชาติบ้านเมืองกับกอง"ทัพพระเจ้าตาก"คืนวันพุธนี้ร่วมกันส่งพลังใจจะหลั่งน้ำตาไปพร้อมๆกัน #หนึ่งด้าวฟ้าเดียว https://t.co/7nHM9pRIoT

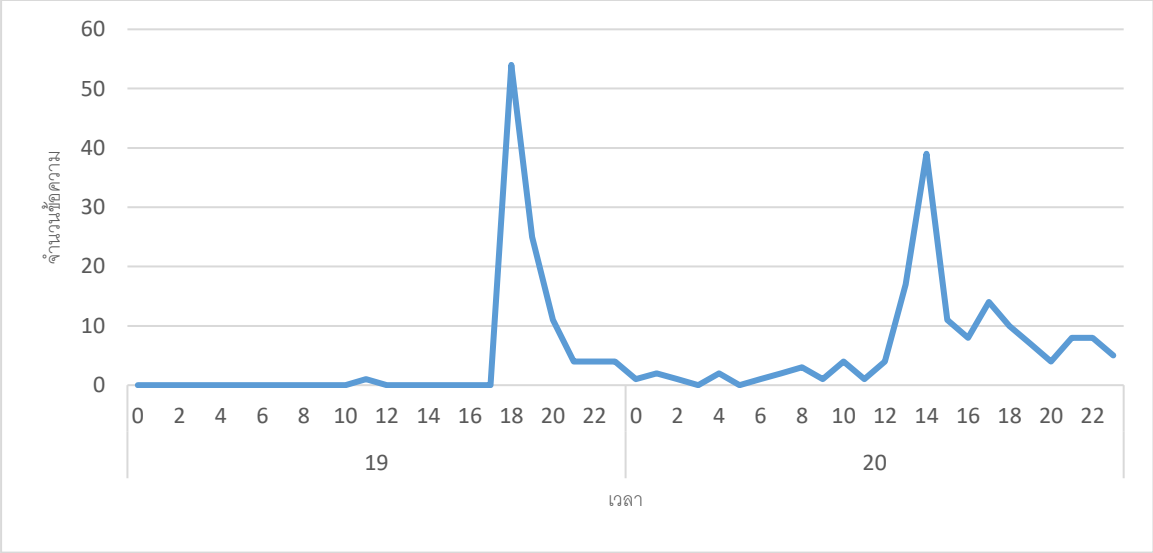
หัวข้อที่ตรวจจับเกิน ในวันที่ 19 และ 20 มิถุนายน ค.ศ. 2018

- มีคำที่น่าสนใจคือ “ควาน”, “ทอง” และ “ลิน”
    - เกิดจากนักร้องเกาหลีชื่อดังคนหนึ่ง ย้อมผมจากสีดำเป็นสีทอง
    - มีปริมาณทวิตเตอร์ที่เกี่ยวข้องเยอะมาก เมื่อเทียบกับช่วงเวลาปกติ ดังแสดงในรูปที่ 68

ตารางที่ 47 ตัวอย่างข้อความทวิตเตอร์ของคำ “ควาน”, “ทอง” และ “ลิน” ในวันที่ 19 และ 20 มิถุนายน ค.ศ. 2018

มึงเปนลมแล้ว เมื่อวานเพื่อนพึ่งส่งรูปอิดิท**ควานลิน**ผม**ทอง**มา กุพึ่งบอกเพื่อนว่ากุจะตาย อิเหี้ย เปนลมมมมมมมมมมมมม
อั่ยเหี้ยๆๆๆ ผม**ทอง**งง **ฏีควานลิน**น ผัวฝรั่ง ฮีออออ
จะร้องไห้แล้วเว้ยยยย ใจพี่สั่นไหวเลยยย ฮีอออออ **ควานลิน**ย้อมผม**ทอง ควานลิน**ย้อมผม**ทอง ควานลิน**ย้อมผม**ทอง**งง!!!!!! https://t.co/D8uw3Q3q53
ควานลินหรอ ผม**ทอง**ขาวมาก!
เพิ่งพูดกะอีอั้มว่าแบบไม่เคยเห็น**ควานลิน**กับแจฮวานย้อมผม ฮีอออออออ อยากเห็นหัว**ทอง**แบบชัดๆ เป็นบุญตา ชั้นต้องทัมอย่างไร

รูปที่ 67 ตัวอย่างของข้อความทวิตเตอร์ที่ตรวจจับเกิน



รูปที่ 68 จำนวนข้อความที่ปรากฏคำว่า “ควาน” “ทอง” และ “ลิน” พร้อมกันต่อชั่วโมง ในวันที่ 19 มิถุนายน ค.ศ. 2018 ถึง 20 มิถุนายน ค.ศ. 2018

ตารางที่ 48 ผลลัพธ์ของการตรวจจับคำและหัวข้อเกิดใหม่ของตัวจำแนกประเภท ในวันที่ 19 มิถุนายน ค.ศ. 2018

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
aisnextgxpeckbambam	[bambam, ais, peckpalit, ผลิตโชค, เป๊กผลิตโชค, supol, aisnextgxpeckbambam, peckpalitbell, peckpalitchoke, เป๊กแบม, bellsupol, เป๊ก]	[]	[]
awcbootcamp2018	[itcitybacon, aov, arenaofvalor, esports, แซล, ทีม, rov, สู้, awcbootcamp2018, daretowin, ม่อน, แซลม่อน, rovth]	[]	[]
blackpink	[lisa, blackpink_squareup, squareup, blackpink, ddu_du_ddu_du, squareup_blackpink]	[lisa, blackpink]	[]
dafbama2018got7	[bringgot7ontheellenshow, gdny, eyesonyou, แบม, gotofficial, got7worldtour, มาร์ค, got7, official, dafbama2018got7,	[[bringgot7ontheellenshow, gdny, eyesonyou, got7worldtour, got7, dafbama2018got7,	[[แบตหมดจ้า, มากมายยย, สิวะ, 갓세븐], [ป้าย, ครีม]]

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
	markbam, bringgot7togdny, bringgot7onthelatelateshow]	bringgot7togdny, bringgot7onthelatelateshow], [แบม, markbam]]	
incredibles2	[ครอบครัว, ภาค, incredibles2, หนัง, incredibles2th]	[]	[]
inuyashiki	[ลุง, คุณลุง, หนัง, movietwit, inuyashiki]	[]	[]
jibbnk48	[หน้าตา, bnk48, จิ๊บ, jibbnk48]	[]	[]
kissmeagain	[เก้า, kissmeagain, เตนิว, ซัน, petekao, taynew, พีทเก้า]	[]	[]
oreoandritzwaferroll	[ผลิตโชค, เป๊กผลิตโชค, dafbama2018, palitarea, peckpalitchoke, เป๊ก, oreoandritzwaferroll, dafbama2018peckpalitchoke]	[peckpalitchoke, เป๊กผลิตโชค, oreoandritzwaferroll, dafbama2018peckpalitchoke]	[เป๊กแบม, aisnextgxpeckbambam, aisnextg, นุช]
smtmthailand	[เป้, ดิส, ด่า, เฟรม, ร้อง, smtmthailand, therapper, ทีม, วดฟ, ร้องเพลง, คนดู, เพลง, รายการ]	[ดิส, ด่า, smtmthailand, วดฟ, รายการ]	[นายนะ, พี่อุ๋ย]

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
worldsenbatsuth	[musicbnk48, เฌอสิค, worldsenbatsuth, jibbnk48, เลือกตั้ง, bnk48, cherprangbnk48, เฌอ, พี่เฌอ, คะแนน, เลือกตั้งakb, ปราง, มิวสิค, akb48, bnk, jennisbnk48, เฌอปราง, อันดับ]	[]	[]
wwic2018	[วินเนอร์, winner, วิน, wwic2018]	[]	[]
บอลโลก	[ฟุตบอลโลก2018, coljpn, โคลอมเบีย, ตูนิเซีย, เซเนกัล, อังกฤษ, tuneng, ครึ่งแรก, samuraiblue, senegal, สวัสดีบอล, บอลโลก, worldcup, poland, โลก2018, mancity, รอบ, โลก, สวัสดี, england, บอลโลก2018, สวัสดีบอลโลก2018, polsen, สวัสดีบอลโลก, จุดโทษ, โปแลนด์, ชนะ, ญี่ปุ่น, บอล, tunisia, เคน, colombia, japan]	[[ญี่ปุ่น, บอล, สวัสดีบอลโลก], [อังกฤษ, ชนะ]]	[[แดง], []]
บุพเพสันนิวาส	[คุณพี่, โป๊ป, บุพเพสันนิวาส,	[]	[]

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
	popezaap]		
ประหารชีวิต	[คดี, ฆ่า, ประหาร, ประหารชีวิต, โทษประหาร, ยกเลิก, ตาย, นักโทษ, แอมเนสตี้]	[]	[]
รถโรงเรียน schoolrangers	[schoolrangers, รถโรงเรียน, ปี๊, ออฟ, ออฟฟิศ, รถโรงเรียน schoolrangers, ออฟกัน]	[]	[]
ลิขิตรัก thecrownprincess	[ลิขิต, จูบ, เจ้าหญิง, thecrownprincessเดชน์ญาญ่า, ทำตามหัวใจ, หมวด, รัก, หมวดแพน, ลัน, ร้องไห้, แพน, สงสาร, ดีมาก, อลัน, น่าสงสาร, ปู่, พี่แอน, ณเดชน์ญาญ่า, ลิขิตรักthecrownprincess, เล่นดีมาก, ร่าเล่น, ร่า, รัก thecrownprincess, เดชน์ญาญ่า, คู่, แอน, ร้าย, เจ้า, ผู้พัน, วิน, เจ้าชายอลัน, ตอนจบ, ละคร, thecrownprincess, เจ้าลัน, ลิขิตรัก]	[[จูบ, เจ้าหญิง, ปู่, ณเดชน์ญาญ่า, ลิขิตรักthecrownprincess, ละคร, อลัน], [แพน, เจ้าชายอลัน, หมวด], [จูบ, ลิขิตรักthecrownprincess]]	[[เคท, ฟิน, วัง, แผน, หัวใจ, พัน, สอง, เจ้าหญิงเคธ, ฉาก, เคธ], [], [therapper, ฉาก]]

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
สายรักสายสวาท	[รักสายสวาท, สาย, สายสวาท, โสม, คุณเปลว, สายรักสายสวาท, เปลว, สายรัก, ลิขิตรัก, หญิง]	[เปลว, สายรักสายสวาท, หญิง]	[สงสาร, โฉม]
เกมเสน่หา	[แต้ว, เกมเสน่หา, เกมหัวใจ, เสน่หา, เกม]	[]	[]
เมีย2018	[ทีมอรุณา, เมีย, กันยา, อรุณา, เมีย 2018]	[]	[]

ตารางที่ 49 ผลลัพธ์ของการตรวจจับคำและหัวข้อเกิดใหม่ของตัวจำแนกประเภท ในวันที่ 20 มิถุนายน ค.ศ. 2018

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
aisnextgxpeckbambam	[bambam, ais, peckpalit, ผลิตโชค, เป๊กผลิตโชค, supol, aisnextgxpeckbambam, peckpalitbell, peckpalitchoke, เป๊กแบม, bellsupol, เป๊ก]	[peckpalitchoke, เป๊กแบม, เป๊กผลิตโชค, aisnextgxpeckbambam]	[dafbama2018peckpalitchoke, oreoandritzwaferroll, aisnextg, นุช]
awcbootcamp2018	[เกมส์, itcitybacon, aov, เล่นเกม, arenaofvalor, esports, เล่น, ทีม, สู้, awcbootcamp2018, rov, daretowin, ม่อน, เกม, แซลม่อน, rovth]	[เกม, เล่น]	[]
bellsupol	[peckpalit, bell, ผลิตโชค, supol, เป๊กผลิตโชค, peckpalitbell, peckpalitchoke, เป๊ก, bellsupol, dafbama2018peckpalitchoke]	[]	[]
icanseeyourvoice	[ปู่จ๋า, ปู่, จ๋าน, จ๋า, ปู่จ๋าน, icanseeyourvoice]	[]	[]
incredibles2	[incredibles2, ครอบครัว, หนัง,	[]	[]

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
	incredibles2th]		
inuyashiki	[ลุง, คุณลุง, หนัง, movietwit, inuyashiki]	[]	[]
jibbnk48	[bnk48, จิ๊บ, jibbnk48]	[]	[]
naturerepublicxexo	[naturerepublicthailand, naturerepublicxexo, คูปอง, exo]	[]	[]
ขุนพันธ์2	[ขุนพันธ์2, movietwit, ตำรวจดียังมีอยู่ไหม, หนัง]	[]	[]
บอลโลก	[ฟุตบอลโลก2018, ฟุตบอลต่างประเทศ, urusau, โมร็อคโก, pormor, โมร็อกโก, egypt, ยิง, ซาอุดิอาระเบีย, fifaworldcup2018, อุรุกวัย, สวัสดีบอล, บอลโลก, อียิป, รัสเซีย, อียิปต์, worldcup, ซาอุ, portugal, rusegy, mancity, uruguay, โปรตุเกส, โลก, สวัสดี, england, บอลโลก2018, russia, สวัสดีบอลโลก2018, สวัสดีบอลโลก,	[[ยิง, โปรตุเกส, สวัสดีบอลโลก], [อียิปต์, รัสเซีย]]	[[ชนะ, เมสซี่, โด้, พัง, ประตู], []]

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
	ฟุตบอลโลก, morocco, เล่น, บอล, saudiarabia]		
บุพเพสันนิวาส	[แม่นาย, โป๊ป, เบลล่า, บุพเพสันนิวาส รีรัน, เบล, ล่า, คุณพี่, บุพเพสันนิวาส]	[]	[]
ประหารชีวิต	[ฆ่า, ประหาร, ประหารชีวิต, โทษ ประหาร, ตาย, แอมเนสตี้]	[ประหาร]	[โทษ]
ลิขิตรัก thecrownprincess	[ลิขิต, ลิขิตรักthecrownprincess, รักthecrownprincess, รัก, thecrownprincess, ลิขิตรัก]	[ลิขิตรักthecrownprincess]	[เคท, ฟิน, จูบ, เจ้าหญิง, วัง, ปู่, ณเดชน์ญาญ่า, แผน, หัวใจ, พัน, สอง, เจ้าหญิงเคธ, ฉาก, ละคร, อลัน, เคธ]
สัมผัสรัตติกาล	[คุณศักดิ์, จันทร์จ้า, สัมผัสรัตติกาล, โสม, คุณธนา, สัมผัส, หมอ, รัตติกาล, หมอน, ไปหาหมอ, หญิง]	[]	[]
สายรักสายสวาท	[รักสายสวาท, สาย, สายสวาท, สาย รักสายสวาท, เปลว, สายรัก, รัก, ลิขิต รัก]	[เปลว, สายรักสายสวาท]	[หญิง]
หนึ่งด้าวฟ้าเดียว	[ตาก, ขันหมาก, ออกญาวัง, ทอง, เจ้า ค่ะ, เจมส์, ขันทอง, นักแสดง, แม่, แมงเม่า, เทวา, หล่อ, ป้า, ปล้ำ,	[[แมงเม่า, ปล้ำ, พี่ขัน, ขัน, ทอง, เจ้า คะ, ละคร, หนึ่งด้าวฟ้าเดียว, แม่พัง], [หล่อ, เจมส์],	[[เจ้าตัว, เตรียม, ฟื้น, แมง, เจ้า, พระ, เม่า, ฉาก, ศรีเรือน, พี่ขันทอง, หมาก, ช้าง], [],

หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
	ร้องไห้, ขัน, พี่ปล้ำ, ท่านขุน, ฟ้า, แต้ว, ปล้ำเจ้า, ออกพระศรี, ท่านขุนรักษ์, ฟ้าเดียว, ด้าวฟ้า, วัง, ขุน, พี่ขัน, แม่แมงเม่า, พ่อ, ด้าว, รักษ์เทวา, แม่เป้า, รักษ์, เจ้าคะ, ละคร, หนึ่งด้าว, หนึ่งด้าวฟ้าเดียว, พัง, แม่พัง]	[แต้ว, ร้องไห้]]	[สงสาร, พี่แต้ว]]
เกมกลมายา	[เกม, เกมกลมายา, มายา]	[]	[]
เล็บครุฑ	[เล็บ, ครุฑ, เล็บครุฑ]	[]	[]
เมีย2018	[ทีมอรุณา, เมีย, กันยา, อรุณา, เมีย 2018]	[]	[]

ตารางที่ 50 ผลลัพธ์ของการตรวจจับคำและหัวข้อเกิดใหม่เกินของตัวจำแนกประเภท ในวันที่ 19-20 มิถุนายน ค.ศ. 2018

วัน	หัวข้อผลเฉลย	กลุ่มคำที่เกี่ยวข้องในผลเฉลย	คำที่ทำนายถูกต้อง	คำที่ทำนายเกิน
6/19/2018	over	□	□	[ควาน, ทอง, ลิน]
6/19/2018	over	□	□	[ตอบ, ถาม]
6/19/2018	over	□	□	[ชัด, ได้ยิน]
6/20/2018	over	□	□	[นอน, หลับ]
6/20/2018	over	□	□	[ควาน, ลิน]
6/20/2018	over	□	□	[สวย, เจน]
6/20/2018	over	□	□	[น้ำ, ไหล]
6/20/2018	over	□	□	[bringgot7ontheellenshow, bringgot7onthelatelateshow, dafbama2018got7, got7, got7worldtour]

6.3.2 คำเกิดใหม่ที่ตัวจำแนกประเภทตรวจจับได้แต่วิธีการอื่นตรวจจับไม่ได้

ตัวจำแนกประเภทที่เสนอในงานวิจัยนี้ ใช้คุณลักษณะหลายชนิดและหลายกรอบเวลาในการเรียนรู้ ทำให้สามารถตรวจจับคำเกิดใหม่ที่มีคุณลักษณะไม่เด่นมากได้ ดังแสดงในตารางที่ 51 ที่แสดงคำที่ตัวจำแนกประเภทตรวจจับได้ แต่คุณลักษณะอื่นตรวจจับไม่ได้ โดยมีค่าในคอลัมน์ (Column) ของแต่ละคุณลักษณะคือ ความต่างของค่าคุณลักษณะที่มากที่สุดของคำและวันที่สนใจ กับ เส้นแบ่ง (Threshold) ที่ดีที่สุดของคุณลักษณะ โดยผ่านการทำนอร์มัลไลซ์ (Normalize) ด้วยค่าต่ำสุดและมากสุดของคุณลักษณะในข้อมูลทดสอบแล้ว โดยค่าที่ได้เป็นค่าที่สามารถแสดงว่าคำดังกล่าวมีค่าสูงสุดของคุณลักษณะ ต่ำกว่าหรือมากกว่าเส้นแบ่งของลักษณะมากน้อยเท่าไหร่ โดยมีการสูตรคำนวณดังต่อไปนี้ ถ้าค่าน้อยกว่า 0 แสดงว่า ค่าคุณลักษณะที่มากที่สุดของคำและวันที่สนใจ มีค่าต่ำกว่าเส้นแบ่งที่ดีที่สุดของคุณลักษณะนั้น ทำให้คุณลักษณะนั้นไม่สามารถตรวจจับคำเกิดใหม่ดังกล่าวได้ ในทางกลับกัน ถ้าค่ามากกว่า 0 แสดงว่าคุณลักษณะนั้นสามารถตรวจจับคำดังกล่าวได้

กำหนดให้

$threshold_x$ คือ เส้นแบ่ง (Threshold) ที่ดีที่สุดของคุณลักษณะ x ซึ่งได้จากการทดลองที่ 6.1.1 และ 6.1.3

$Max(x_{w,d})$ คือ ค่าของคุณลักษณะ x ที่มีคำ w ในวันที่ d ที่มากที่สุด

$Max(x)$ คือ ค่าของคุณลักษณะ x ที่มากที่สุดในข้อมูลทดสอบ

$Min(x)$ คือ ค่าของคุณลักษณะ x ที่น้อยที่สุดในข้อมูลทดสอบ

$$x\_norm_{w,d} = \frac{Max(x_{w,d}) - Min(x)}{Max(x) - Min(x)} \quad (61)$$

$$threshold\_norm_w = \frac{threshold_x - Min(x)}{Max(x) - Min(x)} \quad (62)$$

$$value = 100 * \frac{x\_norm_{w,d} - threshold_x}{threshold_x} \quad (63)$$

ตารางที่ 51 คำที่ตัวจำแนกประเภทป่าไม้แบบสุ่มตรวจจับได้ แต่คุณลักษณะอื่นตรวจจับไม่ได้

วัน	คำเกิดใหม่	count	rsi2_cd	stoch2_slow	wr2	acc	z
2018-6-18	แพน	12	-5.02	-36.08	-2.53	-5.76	-5.90
2018-6-19	อลัน	12	-3.77	-13.19	-0.67	-3.59	-5.85
2018-6-20	อียิปต์	8	-1.90	-5.88	-1.26	-4.47	-17.37
2018-6-18	kissmeagain	4	-5.02	-64.00	-92.36	-10.88	-50.32
2018-6-18	พีทเก้า	4	-9.87	-84.05	-97.45	-10.86	-50.32
2018-6-18	เก้า	4	-14.96	-86.69	-91.89	-10.65	-50.32
2018-6-20	เจมส์	15	-3.13	-14.55	-0.42	-3.91	-1.16
2018-6-18	พี่เฌอ	3	-17.71	-98.67	-99.97	-10.83	-71.07
2018-6-20	เจ้าคะ	8	-5.84	-18.26	-2.02	-5.21	-17.41
2018-6-18	เฌอปราง	8	-14.19	-88.41	-97.02	-9.87	-17.36
2018-6-18	มิวสิค	4	-13.23	-71.60	-91.50	-9.58	-50.32
2018-6-19	เฌอปราง	3	-17.86	-97.66	-99.94	-10.83	-71.07
2018-6-19	cherprangbnk48	7	-1.19	-15.49	-2.31	-4.66	-22.21
2018-6-19	พี่เฌอ	5	-16.38	-95.72	-98.05	-10.44	-37.41
2018-6-18	เจ้าหญิง	8	-2.08	-28.52	-64.42	-5.64	-17.37
2018-6-19	bnk48	7	-4.84	-5.67	-3.24	-5.88	-22.25
2018-6-19	เจ้าชายอลัน	10	-4.37	-27.49	-0.69	-4.43	-10.48

วัน	คำเกิดใหม่	count	rsi2_cd	stoch2_slow	wr2	acc	z
2018-6-19	แบม	10	-7.04	-15.75	-0.49	-4.87	-10.48
2018-6-19	มาร์ค	12	-5.38	-24.34	-1.17	-5.27	-5.85
2018-6-19	เซเนกัล	6	-10.14	-52.55	-12.70	-8.50	-28.62
2018-6-19	itcitybacon	7	-5.69	-47.43	-67.75	-6.65	-22.23
2018-6-20	ปู่	5	-15.61	-85.94	-96.60	-10.27	-37.41
2018-6-18	สวีเดน	9	-8.42	-20.41	-0.88	-4.84	-13.56
2018-6-20	bambam	8	-3.66	-19.52	-1.46	-4.42	-17.36
2018-6-20	เป๊กแบม	9	-2.60	-3.25	-0.79	-2.69	-13.55
2018-6-19	โปแลนด์	6	-9.88	-47.63	-7.40	-7.99	-28.60
2018-6-18	โป๊ปเบลล่า	7	-7.89	-35.68	-4.34	-6.81	-22.28
2018-6-18	เปลว	8	-9.58	-39.39	-3.34	-7.11	-17.36
2018-6-19	ณเดชน์ญาญ่า	7	0.09	-16.60	-1.06	-4.02	-22.21
2018-6-20	ประหารชีวิต	13	-3.15	31.22	-1.69	-3.18	-4.06
2018-6-18	ตี่นเต้น	11	1.63	-12.03	-0.69	-3.91	-7.98
2018-6-18	ดอกหญ้าในพายุ	9	-9.26	1.33	-1.74	-3.81	-13.58
2018-6-20	ฆ่า	15	-1.57	25.42	-1.09	-2.76	-1.16
2018-6-20	โลก	13	-2.44	35.98	-0.71	-0.83	-4.07
2018-6-19	therapper	8	1.51	-44.92	-50.25	-9.43	-17.36

6.3.3 ความสำคัญของแต่ละคุณลักษณะในตัวจำแนกประเภทที่ดีที่สุด

คุณลักษณะที่ใช้ในการสร้างตัวจำแนกประเภท มี 3 ประเภท ได้แก่1) คุณลักษณะที่นำเสนอในงานวิจัยนี้, 2) คุณลักษณะที่นำเสนอในงานวิจัยอื่น และ 3) ตัวแปรที่ใช้ในการสร้างคุณลักษณะต่าง ๆ โดยในแต่ละคุณลักษณะจะมีการใช้ค่ากรอบเวลา w ที่แตกต่างกัน ทำให้คุณลักษณะที่ใช้ในงานวิจัยมีจำนวนมาก โดยมีรายละเอียดแสดงในตารางที่ 52

ตารางที่ 52 จำนวนคุณลักษณะในแต่ละกลุ่มคุณลักษณะที่ใช้ในการสร้างตัวจำแนกประเภท

กลุ่มคุณลักษณะ	คุณลักษณะ	กรอบเวลา	จำนวนคุณลักษณะ
rsi	rsi2	5, 10, 15, 30, 60	5
	rsi2_cd	(5,10), (5,15), (5,15), (10,15), (10,30), (15,30), (15,60), (30,60)	8
	rsi2_slow	5, 10, 15, 30, 60	5
stoch	stoch 2	5, 10, 15, 30, 60	5
	stoch 2_cd	(5,10), (5,15), (5,15), (10,15), (10,30), (15,30), (15,60), (30,60)	8
	stoch 2_slow	5, 10, 15, 30, 60	5
wr	wr2	5, 10, 15, 30, 60	5
	wr2_cd	(5,10), (5,15), (5,15), (10,15), (10,30), (15,30), (15,60), (30,60)	8
	wr2_slow	5, 10, 15, 30, 60	5
acc	acc	(5,10), (5,15), (5,15), (10,15), (10,30), (15,30), (15,60), (30,60)	8
z	z	5, 10, 15, 30, 60	5
ตัวแปรที่ใช้ในการคำนวณคุณลักษณะต่าง ๆ			57
รวม			124

จากตารางที่ 53 คุณลักษณะของตัวจำแนกประเภทป่าไม้แบบสุ่มที่ดีที่สุด 20 อันดับแรก สามารถพิจารณาที่มาของคุณลักษณะดังกล่าวได้ ดังนี้

- เป็นคุณลักษณะสุดท้ายที่เสนอในงานวิจัยนี้ 7 คุณลักษณะ
  - เป็นคุณลักษณะของ rsi จำนวน 4 คุณลักษณะ ได้แก่
    - rsi2_60, rsi2_slow_60, rsi2_slow_30, rsi2_cd_30_60
  - เป็นคุณลักษณะของ stoch จำนวน 2 คุณลักษณะ ได้แก่
    - stoch2_slow_30, stoch2_slow_60
  - เป็นคุณลักษณะของ wr จำนวน 1 คุณลักษณะ ได้แก่
    - wr2_slow_60
- เป็นตัวแปรที่ใช้ในการสร้างคุณลักษณะของงานวิจัยนี้ 12 คุณลักษณะ
- เป็นคุณลักษณะสุดท้ายของงานวิจัยอื่น 0 คุณลักษณะ
- เป็นตัวแปรที่ใช้ในการสร้างคุณลักษณะงานวิจัยอื่น 1 คุณลักษณะ ได้แก่ v_60
- รวมเป็น 20 คุณลักษณะ

จากตารางที่ 54 คุณลักษณะของตัวจำแนกประเภทป่าไม้แบบสุ่มที่ดีที่สุด 20 อันดับแรก สามารถพิจารณากรอบเวลาของคุณลักษณะดังกล่าวได้ ดังนี้

- คุณลักษณะที่ใช้กรอบเวลา 5 นาทีย้อนหลัง จำนวน 0 คุณลักษณะ
- คุณลักษณะที่ใช้กรอบเวลา 10 นาทีย้อนหลัง จำนวน 2 คุณลักษณะ
  - ได้แก่ v_highest_10, v_lowest_10
- คุณลักษณะที่ใช้กรอบเวลา 15 นาทีย้อนหลัง จำนวน 2 คุณลักษณะ
  - ได้แก่ v_highest_15, v_lowest_15
- คุณลักษณะที่ใช้กรอบเวลา 30 นาทีย้อนหลัง จำนวน 4 คุณลักษณะ
  - v_highest_30, v_lowest_30, rsi2_slow_30, stoch2_slow_30
- คุณลักษณะที่ใช้กรอบเวลา 60 นาทีย้อนหลัง จำนวน 11 คุณลักษณะ
- คุณลักษณะที่ใช้ 2 กรอบเวลา คือ กรอบเวลา 30 และ 60 จำนวน 1 คุณลักษณะ
  - rsi2_cd_30_60
- แสดงให้เห็นว่ากรอบของเวลาที่สำคัญที่สุดคือกรอบของเวลา w 60 นาทีย้อนหลัง

ตารางที่ 53 อันดับและความเกี่ยวข้องของคุณลักษณะของคุณลักษณะในตัวจำแนกประเภทป่าไม้แบบสุ่ม

อันดับ	คุณลักษณะ	ความสำคัญ	ความเกี่ยวข้อง			ประเภทของคุณลักษณะ		
			rsi	stoch	wr	2	cd	slow
1	v_highest_60	0.038		v_highest	v_highest			
2	v_highest_30	0.032		v_highest	v_highest			
3	v_lowest_60	0.026		v_lowest	v_lowest			
4	v2_gain_60	0.026	v2_gain					
5	v_highest_15	0.026		v_highest	v_highest			
6	**stoch2_slow_60**	0.024		**stoch2_slow**				**stoch2_slow**
7	v_highest_10	0.022		v_highest	v_highest			
8	count_highest_60	0.022		count_highest	count_highest			
9	v_lowest_30	0.021		v_lowest	v_lowest			
10	**wr2_slow_60**	0.021			**wr2_slow**			**wr2_slow**
11	v2_loss_60	0.020	v2_loss					
12	**rsi2_slow_60**	0.020	**rsi2_slow**					**rsi2_slow**
13	**rsi2_60**	0.018	**rsi2**			**rsi2**		
14	rs2_60	0.018	rs2					
15	v_lowest_15	0.017		v_lowest	v_lowest			
16	**rsi2_cd_30_60**	0.017	**rsi2_cd**				**rsi2_cd**	

อันดับ	คุณลักษณะ	ความสำคัญ	ความเกี่ยวข้อง			ประเภทของคุณลักษณะ		
			rsi	stoch	wr	2	cd	slow
17	**rsi2_slow_30**	0.016	**rsi2_slow**					**rsi2_slow**
18	v_lowest_10	0.015		v_lowest	v_lowest			
19	**v_60**	0.014						
20	**stoch2_slow_30**	0.014		**stoch2_slow**				**stoch2_slow**
จำนวน			7	11	10	1	1	5

ตารางที่ 54 อันดับและกรอบเวลาของคุณลักษณะของคุณลักษณะในตัวจำแนกประเภทป่าไม้แบบสุ่ม

อันดับ	คุณลักษณะ	ความสำคัญ	กรอบของเวลา				
			5	10	15	30	60
1	v_highest_60	0.038					60
2	v_highest_30	0.032				30	
3	v_lowest_60	0.026					60
4	v2_gain_60	0.026					60
5	v_highest_15	0.026			15		
6	stoch2_slow_60	0.024					60
7	v_highest_10	0.022		10			
8	count_highest_60	0.022					60
9	v_lowest_30	0.021				30	
10	wr2_slow_60	0.021					60
11	v2_loss_60	0.020					60
12	rsi2_slow_60	0.020					60
13	rsi2_60	0.018					60
14	rs2_60	0.018					60
15	v_lowest_15	0.017			15		
16	rsi2_cd_30_60	0.017				30	60

อันดับ	คุณลักษณะ	ความสำคัญ	กรอบของเวลา				
			5	10	15	30	60
17	rsi2_slow_30	0.016				30	
18	v_lowest_10	0.015		10			
19	v_60	0.014					60
20	stoch2_slow_30	0.014				30	
จำนวน			0	2	2	5	12

6.3.4 ข้อความทวิตเตอร์ที่ใช้ในการปั๊มโฆษณา

เป็นข้อความทวิตเตอร์ที่ใช้ในการโฆษณา เช่น รับสอนพิเศษ ขายคอร์สลดน้ำหนัก ขายครีม เป็นต้น และเพื่อเพิ่มโอกาสที่ผู้ใช้งานทวิตเตอร์อื่นเห็นข้อความนี้ จึงนิยมติดแท็กที่กำลังเป็นเทรนด์ในปัจจุบันลงไปในข้อความจำนวนมาก ซึ่งจากข้อความนี้ทำให้เราสามารถตรวจจับคำที่เป็นเทรนด์ได้ง่ายขึ้นจากแท็กของข้อความกลุ่มนี้ แต่ขณะเดียวกันการหาความสัมพันธ์ระหว่างคู่ของคำทำได้ยากขึ้น เนื่องจากคำ 2 คำที่ไม่เกี่ยวข้องและไม่ควรปรากฏในข้อความเดียวกัน จะถูกทำให้ปรากฏในข้อความเดียวกันจากการแท็กของข้อความโฆษณาเหล่านี้ ตัวอย่างของข้อความที่เป็นโฆษณาถูกแสดงในตารางที่ 55

ตารางที่ 55 ตัวอย่างข้อความที่เป็นโฆษณา

#กสพท #กัปตนชลธร #กระเป๋า #หมากปริญ #หลินโฮ #หนึ่งด้าวฟ้าเดียว #หนึ่งข้อความถึงคุณพี # หน้ากากเป็ดน้อย #หิ้วเกาหลี #ฟุตบอลโลก #ฟอลฟรี #งานpart time จ่ายเป็นสัปดาห์ งานถูกกฎหมาย ขอคนจริงจังพร้อมเริ่มงาน *รับพิเศษกทม.และปริมณฑล คลิ๊ก> https://t.co/tleHxVnYXq https://t.co/QWW7nyjGEB
WorldCup #ยยขคพ #รีบจีบนะจะผอมแล้ว #สวัสดีบอลโลก #สายรักสายสวาท #สวัสดีบอลโลก 2018 #หมากปริญ #หนึ่งด้าวฟ้าเดียว #ออฟกัน #ญี่ปุ่น #ARG #BNK48 #BLACKPINK #BTOB #BamBam #BELPAN #BTS ✿กรุงเทพฯ ปริมณฑล พิจารณาเป็นพิเศษ ✿รับอายุ 18ปีขึ้นไป สนใจงานคลิ๊ก☞ https://t.co/l1s3fkzUV4 https://t.co/n813SlVpj9
หาคนสนใจอยากลดน้ำหนักปักหมุดเลยคะ #ณเดชน์ญาญ่า #เรื่องเล่าเช้านี้ #KissMeAgain #IFCAsia #expressminions #whatthefest #TheNextBoyGirlBandTH #พร้อมส่ง #มนตราลายหงส์ #WakeUpชะนี #แผ่นดินไหว #อุบัติเหตุ #ลดฟ้าหนักแบบไม่แดกยาไว้ย #ลดฟ้าหนัก #howtoperfact #ฉันจะผอม #รีบจีบนะจะผอมแล้ว https://t.co/1XMdOx3wA4
ใครอยากลดน้ำหนักปรึกษานานาได้นะค่ะ ID @piy4790e #ลดฟ้าหนักแบบไม่แดกยาไว้ยย #ฝนตก #ตลาดนัดรถไฟบังทัน #ของดีบอกต่อ #ณเดชน์ญาญ่า #JibBNK48 #ตลาดนัดWANANAONE #ถูกและดีบอกต่อ #ถูกบอกต่อ #มิ้งโป๊ะแตก #รีวิวเชเว่น #เรื่องเล่าเสาร์อาทิตย์ #ลดสิว #ลดฟ้าหนัก # ลดความอ้วน #รีบจีบนะจะผอม https://t.co/DLWiypLYlI

มอนิ่งค้าบบ~ สิ่งหนึ่งที่สำคัญ! คือการส่งกำลังใจให้กัน บิวมอบภาพนี้ให้เป็น #แรงบันดาลใจ สำหรับสาวๆ ที่กำลังฟิตหุ่นนะคะ สู้ๆจ้าา #Chimz #ลดความอ้วน #ลดน้ำหนัก #ฉันจะผอม #howtobeauty #howtoperfect #ของดีบอกต่อ #ของมันต้องมี #SanggyunKentaFMinBKK #WWIC2018 #whatthefest #KissMeAgain https://t.co/nCjuhPrLEB

มารีนคอลลาเจน100ml. ราคา650.- ส่งฟรีems - แก้ปัญหาผิว ลดเลือนริ้วรอย ฝ้า กระ จุดด่างดำที่เกิดจากสิว สนใจDMสั่งเลยจ้า LINE ID : ouiangelo #marinecollagen #มารีนคอลลาเจน #BLACKPINK #ครีมทาหน้า #import #NewProfilePic #อย #LISA #ใช้ดีบอกต่อ #ถูกและดี #ThaiPBS #JINU https://t.co/NcfERfqCCN

ครีมกันแดดที่ครองใจสาวญี่ปุ่นมาอย่างต่อเนื่อง Anessa คลิกhttps://t.co/PtwAvA33x3 #Anessa #SkincareMilk #thishop #ช้อปปิ้งออนไลน์ #ดิสช้อป #แอพช้อปปิ้ง #เรื่องเล่าเช้านี้ #ณเดชน์ ญาญ่า #เรื่องเล่าเสาร์อาทิตย์ #GERMEX #whatthefest https://t.co/KH4ULVR0xd

แล้วคุณจะหลงรักผิว มากกว่าที่เคย.. ครีมบำรุงผิว ที่ให้มากกว่าแค่การบำรุง WINK WHITE BODY LOTION #เบลล่าเลือกแล้วแล้วคุณล่ะ? https://t.co/nxDfUv8k33 #โป๊ปธนวรรธน์ #เบลล่าราณี #popezaap #bellacampen #บุพเพสันนิวาส #ทีมน้ำอ้อยร้อยล้าน #popebellarealthai #โป๊ปเบลล่า https://t.co/zmVJaYspj3

บทที่ 7
สรุปผลการวิจัยและแนวทางการวิจัยในขั้นถัดไป

7.1 สถาปัตยกรรม

รูปที่ 69 แสดงสถาปัตยกรรมของงานวิจัย โดยแบ่งออกเป็น 2 ส่วนหลัก คือ 1) ส่วนของการสร้างตัวจำแนกประเภทจากข้อความทวิตเตอร์ในฐานข้อมูล และ 2) ส่วนของการตรวจจับคำและหัวข้อเกิดใหม่

1. ส่วนของการสร้างตัวจำแนกประเภทจากข้อความทวิตเตอร์ในฐานข้อมูล

เป็นส่วนของการสร้างตัวจำแนกประเภทจากข้อความทวิตเตอร์ในอดีต เพื่อนำไปใช้ในการทำนายคำเกิดใหม่กับข้อความทวิตเตอร์ในนาทีปัจจุบันในส่วนของการตรวจจับคำและหัวข้อเกิดใหม่ โดยมีขั้นตอน ดังนี้

   1.1. การประมวลผลก่อน (Pre-Processing) ได้แก่ การทำความสะอาดข้อความ การตัดคำ และการกำจัดคำไม่สำคัญ
   1.2. การสร้างข้อมูล (Data Construction) ได้แก่ การสร้างข้อมูลเชิงเวลา การสร้างคุณลักษณะจากตัวชี้วัดของหุ้น และการปรับปรุงคุณลักษณะจากตัวชี้ของหุ้น
   1.3. การกำกับข้อมูล (Data Labeling) เป็นการสร้างตัวแปรผลเฉลยของข้อมูลเพื่อใช้ในการสร้างตัวจำแนกประเภท
   1.4. การสร้างตัวจำแนกประเภท ป่าไม้แบบสุ่ม (Random Forest Construction)

2. ส่วนของการตรวจจับคำและหัวข้อเกิดใหม่

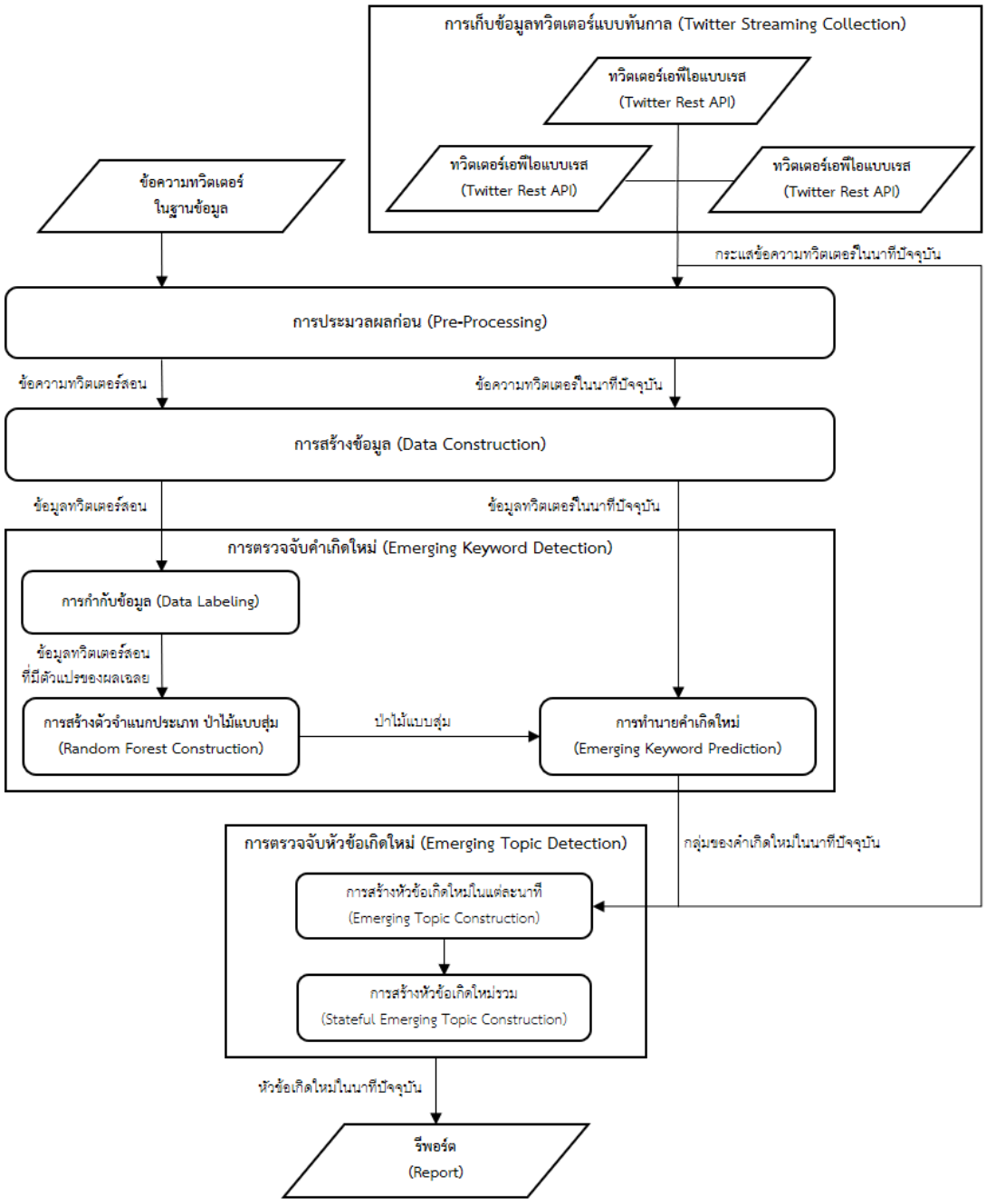
เป็นส่วนของการนำตัวจำแนกประเภทไปใช้ในการทำนายหาคำเกิดใหม่ในแต่ละนาที และนำคำเกิดใหม่เหล่านี้ไปหาความสัมพันธ์เพื่อสร้างเป็นหัวข้อเกิดใหม่ โดยมีขั้นตอน ดังนี้

   2.1. การเก็บข้อมูลทวิตเตอร์แบบทันกาล (Twitter Streaming Collection) เป็นการเก็บรวบรวมข้อความทวิตเตอร์ด้วยทวิตเตอร์เอพีไอแบบเรสหลายโปรแกรมพร้อมกัน
   2.2. การประมวลผลก่อน (Pre-Processing) ได้แก่ การทำความสะอาดข้อความ การตัดคำ และการกำจัดคำไม่สำคัญ
   2.3. การสร้างข้อมูล (Data Construction) ได้แก่ การสร้างข้อมูลเชิงเวลา การสร้างคุณลักษณะจากตัวชี้วัดของหุ้น และการปรับปรุงคุณลักษณะจากตัวชี้ของหุ้น
   2.4. การทำนายคำเกิดใหม่ (Emerging Keyword Prediction) เป็นการใช้ตัวจำแนกประเภทที่ถูกสร้างจากข้อทวิตเตอร์ในฐานข้อมูล ในการทำนายคำเกิดใหม่ในแต่ละนาที

2.5. การสร้างหัวข้อเกิดใหม่ในแต่ละนาที (Emerging Topic Construction) เป็นการหาความสัมพันธ์จากกลุ่มคำเกิดใหม่ที่ทำนายได้ในแต่ละนาที เพื่อจับกลุ่มคำที่เกี่ยวข้องเป็นหัวข้อเกิดใหม่เดียวกัน

2.6. การสร้างหัวข้อเกิดใหม่รวม (Stateful Emerging Topic Construction) เป็นการรวมหัวข้อเกิดใหม่ในแต่ละนาทีที่เกี่ยวข้องเป็นหัวข้อเกิดใหม่รวมเดียวกัน

2.7. รีพอร์ต (Report) เป็นการนำผลลัพธ์ของแต่ละหัวข้อเกิดใหม่มาแสดงผล



รูปที่ 69 สถาปัตยกรรมของงานวิจัย

7.2 สรุปการวิจัย

ในงานวิทยานิพนธ์นี้ ผู้วิจัยได้เสนอการตรวจจับคำและหัวข้อเกิดใหม่ โดยใช้คุณลักษณะจากตัวชี้วัดของหุ้นประเภทโมเมนตัม rsi, $stoch$, wr และปรับปรุงคุณลักษณะดังกล่าวให้ดียิ่งขึ้น อีกทั้งตัวจำแนกประเภทยังถูกนำมาใช้งานโดยไม่มีข้อจำกัดในการสร้างตัวแปรผลเฉลยของข้อมูลที่ใช้ในการสร้างตัวจำแนกประเภท เนื่องจากในงานวิจัยนี้ใช้มาตรวัดประสิทธิภาพของคำแทนการใช้มาตรวัดประสิทธิภาพทั่วไปในการตรวจสอบแบบไขว้ สุดท้าย งานวิจัยนี้ยังสร้างชุดผลเฉลยรายวัน และมาตรวัดประสิทธิภาพของหัวข้อแบบแมโครที่สามารถวัดประสิทธิภาพในแง่มุมของคำและหัวข้อเกิดใหม่พร้อมกัน จากการทดลองพบว่าตัวจำแนกประเภทที่สร้างขึ้นมีประสิทธิภาพในการตรวจจับคำและหัวข้อเกิดใหม่ ดีกว่าทุกวิธีการในปัจจุบันทั้งความแม่นยำและความครอบคลุม รวมถึงความเร็วในการตรวจจับ นอกจากนี้ยังสามารถตรวจจับคำเกิดใหม่ที่ไม่ปรากฏในชุดผลเฉลยได้

7.3 แนวทางวิจัยในขั้นถัดไป

การตรวจจับคำเกิดใหม่

1. การสร้างคุณลักษณะของคู่ของคำในข้อความแทนการสร้างคุณลักษณะของคำเดี่ยว จะช่วยเพิ่มประสิทธิภาพในการตรวจจับคำเกิดใหม่ให้ดียิ่งขึ้น แต่เป็นการเพิ่มภาระในการคำนวณมากขึ้นมากหลายเท่า
2. การสร้างตัวจำแนกประเภทแยกในแต่ละประเภทของคำเกิดใหม่ โดยทำการจับกลุ่มคำเกิดใหม่จากพฤติกรรมด้วยอัลกอริทึม DTW (Dynamic Time Wrapping)
3. การใช้ข้อมูลอื่นในเครือข่ายสังคมออนไลน์ (Social network) มาเพิ่มประสิทธิภาพในการทำนายคำเกิดใหม่ เช่น จำนวนชอบ (Like), จำนวนรีทวีต (Retweet), จำนวนผู้ติดตามของผู้ใช้งานที่โพสต์ข้อความ เป็นต้น

การตรวจจับหัวข้อเกิดใหม่ หรือการจับกลุ่มคำเกิดใหม่

1. การจับกลุ่มคำเกิดใหม่ด้วยอัลกอริทึมกราฟ เพื่อรักษาความสัมพันธ์ของแต่ละคู่ของคำเกิดใหม่ในหัวข้อเดียวกันในแต่ละช่วงเวลา และอาจนำมาใช้ในการกำจัดคำเกิดใหม่ที่ตรวจจับเกินโดยได้

การทดสอบประสิทธิภาพ

1. การทดสอบประสิทธิภาพการตรวจจับคำเกิดใหม่ ด้วยข้อมูลทวิตเตอร์ที่มีเวลาห่างจากชุดข้อมูลทวิตเตอร์สอนมาก เพื่อทดสอบว่าวิธีตรวจจับหัวข้อเกิดใหม่แต่ละวิธีมีความยืดหยุ่น และสามารถรองรับการใช้งานในอนาคตมากน้อยเพียงใด
2. การทดสอบประสิทธิภาพการตรวจจับหัวข้อเกิดใหม่โดยใช้การวัดประสิทธิภาพจากงานวิจัยในปัจจุบัน เช่น การทดสอบประสิทธิภาพโดยใช้อาสาสมัครจำนวนมากที่มีความหลากหลายทั้งช่วงอายุ เพศ สังคม และการศึกษา ในการให้คะแนนแต่ละหัวข้อเกิดใหม่ว่าเป็นหัวข้อเกิดใหม่หรือไม่

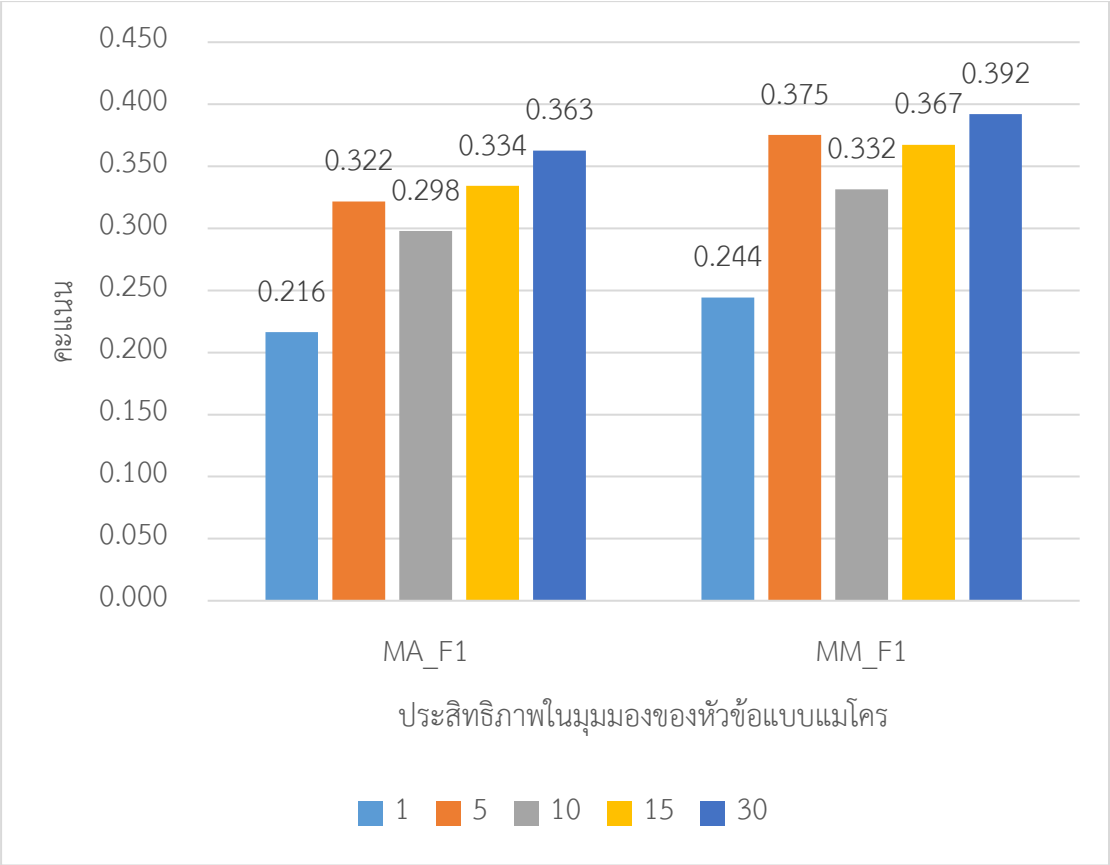
ภาคผนวก ก
การทดลองและผลการทดลองของการหาค่ากรอบเวลาที่ดีที่สุดของการเลือกข้อความทวิตเตอร์ย้อนหลังในขั้นตอนการตรวจจับหัวข้อเกิดใหม่

1. วัตถุประสงค์ของการทดลอง

- เนื่องจากขั้นตอนการตรวจจับหัวข้อเกิดใหม่ (Emerging Topic Detection) หรือการจับกลุ่มคำเกิดใหม่ที่เกี่ยวข้องเป็นหัวข้อเดียวกันนั้น เป็นขั้นตอนที่ใช้ในหาความสัมพันธ์ของคำในกลุ่มข้อความทวิตเตอร์ ดังนั้นเพื่อเพิ่มความแม่นยำในการหาความสัมพันธ์ระหว่างคู่ของคำ จึงมีการประยุกต์ใช้ข้อความทวิตเตอร์ย้อนหลัง w นาที แทนการใช้ข้อความทวิตเตอร์ในนาทีปัจจุบันเท่านั้น แต่การตรวจจับหัวข้อเกิดใหม่นั้น จำเป็นต้องใช้กลุ่มของคำที่ถูกทำนายว่าเป็นคำเกิดใหม่ ซึ่งวิธีตรวจจับคำเกิดใหม่ในงานวิจัยนี้มีจำนวนมาก ได้แก่ rf, rsi_cd, $stoch\_slow$, $wr2$, acc และ z ซึ่งทำให้ยากต่อการหาค่ากรอบเวลา w นาทีย้อนหลังที่ดีที่สุดของแต่ละวิธีตรวจจับคำเกิดใหม่ ดังนั้นในการทดลองนี้จึงใช้ กลุ่มคำเกิดใหม่ในชุดผลเฉลย แทนคำเกิดใหม่จากการทำนายของวิธีต่าง ๆ เพื่อทดลองหาค่ากรอบเวลา w นาทีย้อนหลังที่ดีที่สุดในการตรวจจับหัวข้อเกิดใหม่
- ทดสอบหาค่ากรอบเวลา w ที่ดีที่สุดของการเลือกข้อความทวิตเตอร์ w นาทีย้อนหลังในการเรียนรู้ของอัลกอริทึม LSI ได้แก่ 1, 5, 10, 15 และ 30
- ทดสอบหาค่าเส้นแบ่งสัมประสิทธิ์ความความคล้ายโคไซน์ (Cosine similarity threshold) ในการคัดเลือกคู่ของคำที่มีความเกี่ยวข้องกันเป็นหัวข้อเดียวกัน ได้แก่ 0.1, 0.11, 0.12, ..., 0.97, 0.98 และ 0.99

2. ผลการทดลอง

- จากรูปที่ 70 พบว่าค่ากรอบเวลา w 30 ให้ประสิทธิในมุมมองของหัวข้อแบบแมโครดีที่สุด ทั้งแมโครโดยใช้ค่าเฉลี่ย (Macro average) และใช้การรวม (Macro merge) โดยมีค่าสัมประสิทธิ์ความคล้ายโคไซน์ที่ 0.875



รูปที่ 70 ประสิทธิภาพในมุมมองของหัวข้อแบบแมโครเมื่อใช้กรอบเวลาของข้อความย้อนหลังต่าง ๆ ในการหาความคล้ายของคู่ของคำ

ภาคผนวก ข

ค่าสถิติของคุณลักษณะที่สร้างจากตัวชี้วัดของหุ้น

ตารางที่ 56 ค่าสถิติของคุณลักษณะที่สร้างจากตัวชี้วัดของหุ้นในกรอบเวลาต่าง ๆ

กลุ่มคุณลักษณะ	ประเภทคุณลักษณะ	กรอบเวลา	ค่าเฉลี่ย	ค่าเบี่ยงเบนมาตรฐาน	ค่าต่ำสุด	ค่าสูงสุด
rsi	rsi	5	50.257	13.733	0.877	99.372
		10	50.326	8.671	1.724	98.770
		15	50.320	6.479	2.542	98.167
		30	50.275	3.840	4.874	96.344
		60	50.221	2.240	10.403	92.837
	rsi2	5	50.223	12.512	4.098	98.279
		10	50.296	7.191	11.090	96.390
		15	50.346	5.417	13.663	94.423
		30	50.464	3.898	15.741	89.409
		60	50.581	3.020	20.487	82.938
	rsi2_slow	5	34.209	18.466	0.000	104.795
		10	30.411	18.093	0.000	101.208
		15	29.183	18.082	0.000	93.566
		30	27.874	18.193	0.000	84.193
		60	26.873	18.212	0.000	75.674
	rsi2_cd	5_10	-0.015	1.556	-9.559	8.253
		5_15	-0.012	1.026	-6.059	4.486
		10_15	-0.010	0.606	-5.633	3.368
		10_30	-0.008	0.296	-2.719	1.727
		15_30	-0.008	0.218	-2.452	1.376
		15_60	-0.005	0.107	-1.220	0.737
		30_60	-0.004	0.072	-0.984	0.642

กลุ่มคุณลักษณะ	ประเภทคุณลักษณะ	กรอบเวลา	ค่าเฉลี่ย	ค่าเบี่ยงเบนมาตรฐาน	ค่าต่ำสุด	ค่าสูงสุด
stoch	stoch	5	46.560	40.947	0.103	100.000
		10	37.982	35.733	0.083	100.000
		15	34.560	33.199	0.081	100.000
		30	30.249	29.845	0.080	100.000
		60	26.976	27.387	0.077	100.000
	stoch2	5	55.283	37.077	0.369	100.000
		10	54.921	33.820	0.424	100.000
		15	55.516	32.848	0.429	100.000
		30	58.135	32.869	0.434	100.000
		60	62.857	33.898	0.363	100.000
	stoch2_slow	5	37.420	23.684	0.174	109.929
		10	33.445	23.138	0.081	104.003
		15	32.565	23.621	0.045	102.353
		30	33.136	26.135	0.021	100.902
		60	35.564	29.437	0.010	100.606
	stoch2_cd	5_10	0.072	5.661	-19.500	19.753
		5_15	-0.023	3.568	-9.877	9.932
		10_15	-0.119	3.533	-18.836	19.051
		10_30	-0.161	1.645	-4.926	4.942
		15_30	-0.175	1.624	-6.518	6.551
		15_60	-0.163	0.808	-2.198	2.210
		30_60	-0.157	0.802	-3.237	3.264

กลุ่มคุณลักษณะ	ประเภทคุณลักษณะ	กรอบเวลา	ค่าเฉลี่ย	ค่าเบี่ยงเบนมาตรฐาน	ค่าต่ำสุด	ค่าสูงสุด
wr	wr	5	36.453	39.803	0.000	99.916
		10	32.591	35.036	0.000	99.921
		15	30.227	32.583	0.000	99.921
		30	26.655	29.011	0.000	99.921
		60	23.749	26.235	0.000	99.922
	wr2	5	42.762	37.132	0.000	99.800
		10	42.667	33.963	0.000	99.701
		15	42.772	32.968	0.000	99.720
		30	44.456	32.952	0.000	99.708
		60	48.829	34.218	0.000	99.726
	wr2_slow	5	30.767	22.410	0.000	108.197
		10	27.748	21.309	0.000	102.705
		15	26.907	21.529	0.000	101.132
		30	27.369	23.683	0.000	99.206
		60	30.030	26.842	0.000	99.686
	wr2_cd	5_10	0.640	4.668	-19.721	19.266
		5_15	0.472	2.905	-9.943	9.734
		10_15	0.168	3.153	-19.583	16.220
		10_30	0.056	1.434	-4.978	4.805
		15_30	-0.017	1.465	-6.602	6.211
		15_60	-0.068	0.712	-2.215	2.145
		30_60	-0.108	0.747	-3.292	3.055

บรรณานุกรม

1. Mathioudakis, M. and N. Koudas. *TwitterMonitor: Trend Detection over the Twitter Stream*. 2010. ACM.
2. Cataldi, M., L.D. Caro, and C. Schifanella, *Emerging topic detection on Twitter based on temporal and social terms evaluation*, in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*. 2010, ACM: Washington, D.C. p. 1-10.
3. Weng, J., et al., *Event detection in twitter*. 2011. 1-21.
4. Alvanaki, F., et al., *EnBlogue: emergent topic detection in web 2.0 streams*, in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. 2011, ACM: Athens, Greece. p. 1271-1274.
5. Alvanaki, F., et al., *See what's enBlogue: Real-time emergent topic identification in social media*. 2012.
6. Schubert, E., M. Weiler, and H.-P. Kriegel, *SigniTrend*, in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. 2014. p. 871-880.
7. Xie, W., et al., *TopicSketch: Real-Time Bursty Topic Detection from Twitter*, in *2013 IEEE 13th International Conference on Data Mining*. 2013. p. 837-846.
8. Xie, W., et al., *TopicSketch: Real-Time Bursty Topic Detection from Twitter.* IEEE Transactions on Knowledge and Data Engineering, 2016. **28**(8): p. 2216-2229.
9. Buntain, C., J. Lin, and J. Golbeck, *Learning to Discover Key Moments in Social Media Streams*. 2015.
10. Appel, G., *The Moving Average Convergence-Divergence Method.* Great Neck, NY: Signalert, 1979: p. 1647-1691.
11. W. Jr. Wilder, J., *New Concepts in Technical Trading Systems*. 1978.
12. Lane, G.C., *Lane's stochastics.* Technical Analysis of Stocks and Commodities, 1984. **2**(3): p. 80.
13. Deerwester, S.C., et al., *Indexing by Latent Semantic Analysis.* JASIS, 1990. **41**: p. 391-407.

14. Yang, Y., T. Pierce, and J. Carbonell, *A study of retrospective and on-line event detection*, in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 1998, ACM: Melbourne, Australia. p. 28-36.
15. Allan, J., et al. *Detections , Bounds , and Timelines : UMass and TDT-3*. 2000.
16. Petrović, S., M. Osborne, and V. Lavrenko, *Streaming first story detection with application to Twitter*, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 2010, Association for Computational Linguistics: Los Angeles, California. p. 181-189.
17. Indyk, P. and R. Motwani, *Approximate nearest neighbors: towards removing the curse of dimensionality*, in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. 1998, ACM: Dallas, Texas, USA. p. 604-613.
18. Datar, M., et al., *Maintaining stream statistics over sliding windows: (extended abstract)*, in *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*. 2002, Society for Industrial and Applied Mathematics: San Francisco, California. p. 635-644.
19. Babcock, B., et al., *Maintaining variance and k-medians over data stream windows*, in *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 2003, ACM: San Diego, California. p. 234-243.
20. Welford, B.P., *Note on a Method for Calculating Corrected Sums of Squares and Products.* Technometrics, 1962. **4**(3): p. 419-420.
21. West, D.H.D., *Updating mean and variance estimates: an improved method.* Commun. ACM, 1979. **22**(9): p. 532-535.
22. Finch, T., *Incremental calculation of weighted mean and variance*. 2009.
23. He, D. and D.S. Parker, *Topic dynamics: an alternative model of bursts in streams of topics*, in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, ACM: Washington, DC, USA. p. 443-452.
24. Järvelin, K. and J. Kekäläinen, *Cumulated gain-based evaluation of IR*

techniques. ACM Trans. Inf. Syst., 2002. **20**(4): p. 422-446.

25. Li, C., A. Sun, and A. Datta, *Twevent: segment-based event detection from tweets*, in *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2012, ACM: Maui, Hawaii, USA. p. 155-164.

ประวัติผู้เขียน

ชื่อ-สกุล	เอกภพ วีระสกุลวงศ์
วัน เดือน ปี เกิด	27 ตุลาคม 2536
สถานที่เกิด	กรุงเทพฯ
วุฒิการศึกษา	สำเร็จการศึกษาปริญญาวิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ จากภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2557 และเข้าศึกษาหลักสูตรวิศวกรรมศาสตรมหาบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ ที่ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ในปีการศึกษา 2558
ที่อยู่ปัจจุบัน	61/127 ซ.ประเสริฐมนูกิจ 27 ถ.ประเสริฐมนูกิจ เขตลาดพร้าว แขวงจรเข้บัว กรุงเทพฯ 10230
ผลงานตีพิมพ์	Verasakulvong, E., Vateekul, P., Piyatumrong, A., & Sangkeettrakarn, C. (2018, July). Online Emerging Topic Detection on Twitter Using Random Forest with Stock Indicator Features. In 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE) (pp. 1-6). IEEE.