

ระบบแจ้งเตือนโซเชียลมีเดียไทยสำหรับธุรกิจด้วยซัพพอร์ทเวกเตอร์แมชชีน

THAI SOCIAL MEDIA ALERT SYSTEM FOR BUSINESS BY SUPPORT VECTOR MACHINE

สุพัตรา วิริยะวิสุทธิสกุล¹ ปริญา สงวนสัตย์² พิสิษฐ์ ชาญเกียรติทอง³ และชูชาติ หล้าไชยะศักดิ์⁴

Supatta Viriyavisuthisakul¹ Parinya Sangsuansat² Pisit Charnkeitkong³ and Choochart Haruechaiyasak⁴

^{1,2,3}คณะวิศวกรรมศาสตร์และเทคโนโลยี สถาบันการจัดการปัญญาภิวัฒน์

⁴ศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ

^{1,2,3}Faculty of Engineering and Technology, Panyapiwat Institute of Management

⁴National Electronics and Computer Technology Center

บทคัดย่อ

ในปัจจุบันสื่อสังคมออนไลน์ได้รับความนิยมมากขึ้น เนื่องด้วยเป็นช่องทางที่รวดเร็วและเข้าถึงได้ง่าย ทำให้ผู้ใช้เข้ามาแสดงความคิดเห็น แลกเปลี่ยนประสบการณ์ต่างๆ รวมไปถึงธุรกิจที่ใช้ช่องทางสื่อสังคมออนไลน์ในการสื่อสารกับลูกค้าเช่นกัน หากกล่าวถึงสื่อสังคมออนไลน์ในประเทศไทยแล้ว เว็บไซต์ Pantip ได้รับความนิยมมากในสังคมไทย ดังนั้นหากข้อมูลที่ปรากฏเป็นข้อมูลเชิงลบที่เกี่ยวข้องกับสินค้าหรือบริการอาจก่อให้เกิดความเสียหายกับธุรกิจได้ ดังนั้นเพื่อให้ธุรกิจจัดการข้อมูลบนสื่อสังคมออนไลน์ได้อย่างมีประสิทธิภาพ ธุรกิจควรจะต้องมีระบบแจ้งเตือนความคิดเห็นเชิงลบบนสื่อสังคมออนไลน์ เพื่อให้ธุรกิจสามารถจัดการกับปัญหาได้อย่างรวดเร็ว จะสามารถแก้ไขได้ทันเวลาที่ในบทความวิจัยนี้จะเริ่มจากการสำรวจเครื่องมือสื่อสังคมออนไลน์ ตั้งแต่ปี 2548 จนถึงปัจจุบัน ซึ่งพบว่าแต่ละระบบมีข้อดีแตกต่างกันไป แต่ทุกระบบยังไม่ตรงความต้องการของธุรกิจและพบว่ามีภาระงานการประมวลผลข้อมูลที่มีผิดพลาดและไม่แจ้งเตือน ดังนั้นในงานวิจัยนี้จะนำเสนอระบบแจ้งเตือนธุรกิจบนสื่อสังคมออนไลน์บนพื้นฐานความต้องการของธุรกิจโดยใช้เทคนิคการเรียนรู้ของเครื่อง ระบบจะดึงข้อมูลจากเว็บไซต์ Pantip ตามคำสำคัญ (Keywords) จากนั้นข้อมูลจะผ่านกระบวนการก่อนหน้า การสกัดคุณลักษณะโดย Term Frequency – Inverse Document Frequency (TF-IDF) และเข้าสู่ขั้นตอนการจำแนกประเภทด้วย Support Vector Machine (SVM) จากการทดลองพบว่าระบบมีประสิทธิภาพดีทั้งด้านความถูกต้องและด้านความไวและความจำเพาะ เมื่อเทียบกับผู้เชี่ยวชาญ

คำสำคัญ: ระบบแจ้งเตือน สื่อสังคมออนไลน์ไทย การจำแนกความรู้สึก ซัพพอร์ทเวกเตอร์แมชชีน

Abstract

Nowadays, social media has exponential growth of information generated by social interactions. In this way, social media has a huge impact on how businesses connect with customers. Pantip.com is the one of the most popular webboard in Thailand in which the customers post their feedbacks about products and services. Among these feedbacks, some of them affect the

Corresponding Author

E-mail: supattavir@pim.ac.th

business in a negative way. To solve this problem, businesses should have a system that can alert their customer's negative feedbacks on social media. If the business can quickly deal with these negative feedbacks, the problems will be managed in timely fashion. Therefore, many businesses can take advantage of this market analysis information that can increase their opportunities in marketing. In this paper, the social media tools, that can utilized the data from social medias, are surveyed from 2005 to present. Each system provides different features, but none of them meet the business requirements, especially when the business cannot deal with the problem immediately because the negative sentiments are incorrectly classified and not alert. This paper proposes the social media alert system based on the business requirements. The machine learning technique is applied to determine which information should be alerted to business owners. The system monitors the data posted in Pantip.com that relates to the business by keywords. After the data were collected and preprocessed, feature vectors are extracted by Term Frequency-Inversed Document Frequency (TF-IDF) before feeding to Support Vector Machine (SVM). Experimental results show that it can achieved the good accuracy rate and also in terms of the sensitivity and specificity.

Keywords: Alert system, Thai social media, Sentiment Classification, Support Vector Machine

บทนำ

สื่อสังคมออนไลน์เป็นช่องทางการสื่อสารที่ใช้กันอย่างกว้างขวาง ในการแลกเปลี่ยนข้อมูล ทศวรรษต่างๆ ผู้ใช้จะแบ่งปันข้อมูล เช่น ข้อความ รูปภาพ เสียง หรือ วิดีโอ เป็นต้น สำหรับสื่อสังคมออนไลน์ที่ได้รับความนิยม ได้แก่ Facebook, Google+, Twitter, MySpace, YouTube และ Microblog สื่อสังคมออนไลน์เหล่านี้เป็นเสมือนตัวกลางที่ทำหน้าที่เผยแพร่ข้อมูลออกไป ด้วยอุปกรณ์ที่ผู้ใช้มีอยู่ เช่น โทรศัพท์มือถือ แท็บเล็ต และคอมพิวเตอร์ ด้วยเหตุผลนี้สื่อสังคมออนไลน์จึงเป็นอีกช่องทางที่ธุรกิจให้ความสนใจ เพราะลูกค้าจะแบ่งปันประสบการณ์ในการบริการหรือผลิตภัณฑ์ต่างๆ ซึ่งข้อมูลเหล่านี้มีความสำคัญต่อธุรกิจมาก ถ้าธุรกิจสามารถที่จะเก็บรวบรวมความคิดเห็นของลูกค้าได้อย่างครบถ้วน จะทำให้ธุรกิจสามารถแก้ไขปัญหาและปิดช่องโหว่ได้อย่างรอบด้าน แต่ในความเป็นจริงแล้วไม่สามารถทำได้ เนื่องจากข้อมูลที่มีจำนวนมากและขยายตัวอย่างรวดเร็วในเวลาอันสั้น และในบางข้อมูลก็ไม่มีความเป็นต่อ

ธุรกิจ เพื่อแก้ไขปัญหาเหล่านี้ธุรกิจควรจะต้องมีระบบแจ้งเตือนที่สามารถจัดการกับปัญหาได้อย่างรวดเร็ว

ระบบแจ้งเตือนบนสื่อสังคมออนไลน์ระบบแรกเกิดขึ้นในปี 2548 เนื่องจากในปี 2547 นั้น Facebook เริ่มเปิดให้บริการ การปรากฏขึ้นของ Facebook ได้เปลี่ยนแปลงรูปแบบการสื่อสารข้อมูลไปจากเดิม จำนวนผู้ใช้สื่อสังคมออนไลน์มีปริมาณเพิ่มมากขึ้น ดังนั้นธุรกิจเล็งเห็นความสำคัญของข้อมูลเหล่านี้ ระบบแจ้งเตือนบนสื่อสังคมออนไลน์จึงเกิดขึ้น โดยบทความวิจัยนี้ได้ศึกษาระบบที่ได้รับความนิยมตั้งแต่ปี 2548 ถึงปัจจุบัน ซึ่งประกอบไปด้วย 9 ระบบ ได้แก่ Social Mention¹, Netvibes Dashboard Everything², TOPSY³, Google Alerts⁴, Talk walker: Social Search Brand⁵, Talk walker Alerts⁶, Whir social⁷, Pin Alerts⁸ และ

¹socialmention.com, ²netvibes.com, ³topsy.com, ⁴google.com/alerts, ⁵talkwalker.com/en/social-media-intelligence, ⁶talkwalker.com/alerts, ⁷whirsocial.com, ⁸pinalerts.com, ⁹mention.com

Mention⁹ ในแต่ละระบบจะมีคุณสมบัติที่ต่างกันออกไป แต่อย่างไรก็ตามทุกระบบจะใช้วิธีการดึงข้อมูลตามคำสำคัญที่ผู้ใช้เพิ่มเข้าไป บางระบบสามารถจำแนกความรู้สึกของแต่ละความคิดเห็นได้ โดยส่วนมากจะจำแนกความรู้สึกออกเป็น 3 ประเภท ได้แก่ เชิงบวก (Positive) เชิงลบ (Negative) และปกติ (Neutral) ซึ่งมีความสำคัญต่อธุรกิจอย่างมากโดยเฉพาะความคิดเห็นเชิงลบ (Negative) จะสามารถส่งผลกับธุรกิจได้ แต่ก็ยังพบว่าระบบมีการจำแนกที่ผิดพลาด และในบางระบบก็ไม่สามารถแจ้งเตือนได้จริง

สำหรับงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ความรู้สึกบนสื่อสังคมออนไลน์ ส่วนมากจะอาศัยวิธีการเรียนรู้ของเครื่อง (Machine Learning) ในการจำแนกประเภท เช่น Naïve Bayes, Maximum Entropy หรือ Support Vector Machine (SVM) เป็นต้น ซึ่งวิธีการ SVM ได้ถูกนำเสนอขึ้นเพื่อการจำแนกอารมณ์พื้นฐาน ได้รับค่าความถูกต้องที่ดีที่สุดคือ 96.43% (Binali, Wu & Potdar, 2013) นอกจากนี้ SVM ที่ถูกนำมาใช้งานทางด้านนี้แล้ว ยังพบว่า Naïve Bayes ก็ถูกนำมาใช้ในการจำแนกข้อมูลจาก Facebook และ Twitter เกี่ยวกับการวิเคราะห์ผลิตภัณฑ์ประเภทยาและเครื่องสำอาง จำแนกเป็น 2 ประเภท ได้แก่ Positive และ Negative ผลการวิจัยระบุว่า Naïve Bayes ให้ค่าความถูกต้องที่ 83% เมื่อเปรียบเทียบกับผลการจำแนกด้วยผู้เชี่ยวชาญ แต่เมื่อเปรียบเทียบประสิทธิภาพระหว่าง SVM และ Naïve Bayes แล้ว SVM สามารถให้ประสิทธิภาพที่ดีกว่า (Isah, Trundle & Neagu, 2014) ซึ่งสอดคล้องกับผลการวิจัยที่ได้นำเสนอการจัดกลุ่มอารมณ์ของผู้ใช้ Twitter เกี่ยวกับ Hashtag พายุเฮอริเคนแซนดี้ (Sandy Hurricane) จำแนกเป็น 4 ประเภท ได้แก่ Positive, Anger, Fear และอื่นๆ ผลการวิจัยสรุปว่า SVM สามารถให้ค่าความถูกต้องที่ 75.9% เมื่อนำข้อมูลของกลุ่มอื่นๆ ออกไป ในขณะที่ Naïve Bayes ให้ค่าความถูกต้องได้แค่ 69.1% (Brynielsson et al., 2014) เนื่องจากผู้ใช้งานสื่อสังคมออนไลน์มีอยู่ทั่วโลก จึงจำเป็นต้อง

สร้างสื่อสังคมออนไลน์ให้เหมาะสมกับผู้ใช้ในแต่ละภูมิภาค ดังนั้นภาษาเป็นอีกกลไกหนึ่งที่จะช่วยให้การใช้งานระหว่างผู้ใช้ในภูมิภาคเดียวกันสามารถทำได้ง่ายขึ้น เช่นงานวิจัยที่นำภาษาถิ่นเขมามาวิเคราะห์พฤติกรรมของผู้ใช้บนสื่อสังคมออนไลน์ในประเทศตุนีเซีย โดยการสุ่มบัญชี Facebook ขึ้นมา ทดสอบโดยวิธีการของ Naïve Bayes และ SVM ด้วยชุดข้อมูลเดียวกัน ซึ่งมีการแบ่งข้อมูลฝึกหัดไว้ที่ 60% ส่วนอีก 40% เป็นชุดข้อมูลทดสอบ ผลการวิจัยสรุปว่า SVM สามารถให้ประสิทธิภาพความถูกต้องดีกว่า Naïve Bayes (Akaichi et al., 2013) นอกจากนี้งานวิจัยเกี่ยวกับสื่อสังคมออนไลน์ในภาษาอินโดนีเซีย ได้มีการจำแนกความคิดเห็นของลูกค้า โดยแบ่งออกเป็น 4 กลุ่ม ได้แก่ Positive, Negative, Neutral และ Question ซึ่งผลวิจัยสนับสนุน SVM สามารถให้ค่าความถูกต้องดีที่สุดที่ 83.5% (Naradhipa & Purvarianti, 2011)

สำหรับงานวิจัยนี้จะกำหนดเป้าหมายไปที่ข้อมูลจากสื่อสังคมออนไลน์ในประเทศไทย โดยจะดึงเฉพาะข้อมูลจากเว็บบอร์ด Pantip (“Pantip - Learn, Share & Fun,” n.d.) เพราะเป็นเว็บบอร์ดที่ได้รับความนิยมและมีกลุ่มผู้ใช้งานร่วมกันมาก (Pantip.com Site Info, n.d.) ดังนั้นธุรกิจจึงให้ความสนใจเพราะลูกค้ามักจะแสดงความคิดเห็นเกี่ยวกับสินค้าและบริการ ซึ่งพฤติกรรมเหล่านี้เป็นโอกาสที่ดีที่ธุรกิจจะมีโอกาสได้ทราบถึงความคิดเห็นแบบเชิงลึก ในกรณีที่ลูกค้าให้ความคิดเห็นในเชิงลบ ธุรกิจควรที่จะต้องเข้าไปดูแลและจัดการปัญหาอย่างทันท่วงที อย่างไรก็ตามการจำแนกอารมณ์ของข้อความยังมีความซับซ้อน เนื่องจากโครงสร้างของประโยคติดกันไม่มีช่องว่างระหว่างคำทำให้การตัดคำทำได้ยากกว่าภาษาอังกฤษ อีกทั้งปัจจุบันนิยมใช้คำสแลงค่อนข้างมากรวมไปถึงการพิมพ์อักษรซ้ำๆ เช่น “ชอบมากกกกกก” เป็นต้น ซึ่งงานวิจัยที่ผ่านมาได้เสนอวิธีการทดสอบการจำแนกอารมณ์โดยการดึงข้อมูลจากเว็บบอร์ด Pantip จำนวน 1,635 ชุดข้อมูล จำแนกเป็น 4 ประเภท ได้แก่ Pain Gain Need และ Neutral ซึ่งกำหนดโดย

ผู้เชี่ยวชาญทางด้านการตลาด โดยใช้ k-NN ร่วมกับระยะทาง 10 แบบ ผลการวิจัยสรุปว่า ระยะทางแบบ Bray-Curtis ให้ค่าความถูกต้องที่ดีที่สุดที่ 58.62% ด้วยค่าถ่วงน้ำหนักแบบ TF-IDF (Viriyavisuthisakul et al., 2015) นอกจากการจำแนกอารมณ์ในเว็บบอร์ดแล้ว YouTube ก็เป็นสื่อสังคมออนไลน์อีกประเภทที่เกี่ยวข้องกับการแบ่งปันวิดีโอคลิป ซึ่งได้นำเสนอวิธีการดึงความคิดเห็นภาษาไทยจาก YouTube เพื่อจำแนก โดยข้อมูลประกอบไปด้วยกลุ่ม Thai Music Video (MV) จำนวน 2,771 ชุดข้อมูล และกลุ่ม Thai commercial advertisement video (AD) จำนวน 3,077 ชุดข้อมูล ซึ่งจำแนกตามทฤษฎีของ Ekman's six basic emotions ได้แก่ Anger, Disgust, Fear Happiness, Sadness และ Surprise โดยเปรียบเทียบประสิทธิภาพจากหลายๆ วิธีการ ได้แก่ Multinomial Naïve Bayes (MNB) Decision Tree และ SVM ซึ่งผลการวิจัยสรุปว่าการจำแนกความรู้สึกของความคิดเห็นในชุดข้อมูล AD นั้น SVM สามารถให้ความถูกต้องได้ดีที่ 76.14% และ MNB สามารถให้ความถูกต้องดีที่สุดที่ 84.48% (Sarakit et al., 2015) ในชุดข้อมูล MV

จากการศึกษาเกี่ยวกับระบบแจ้งเตือนบนสื่อสังคมออนไลน์ที่ใช้งานในปัจจุบันพบว่า มีข้อจำกัดในเรื่องภาษาและแหล่งข้อมูล ในบางระบบสามารถรองรับภาษาไทยได้แต่ไม่สามารถจำแนกข้อความได้ โดยเฉพาะข้อความเชิงลบที่มีผลกับธุรกิจ ดังนั้นงานวิจัยชิ้นนี้จะมุ่งเน้นไปในความคิดเห็นภาษาไทยเชิงลบที่มีผลกับธุรกิจ เพื่อสร้างระบบแจ้งเตือนต่อไป ข้อมูลที่ใช้จะถูกดึงมาจากเว็บบอร์ด Pantip ตั้งแต่มีนาคม 2556 ถึงตุลาคม 2558 เข้าสู่กระบวนการก่อนหน้าเพื่อตัดคำ กำจัดคำหยุด การสกัดคุณลักษณะ โดยแต่ละความคิดเห็นจะถูกจำแนกออกเป็น 2 ประเภท โดยผู้เชี่ยวชาญ ข้อมูลที่ผ่านกระบวนการนี้แล้วจะพร้อมเข้าสู่การจำแนกโดยวิธีการ SVM ที่เหมาะสม ซึ่งสืบเนื่องมาจากการทบทวนวรรณกรรมตั้งแต่ปี 2543 ถึงปี 2558 ในทุกงานวิจัยที่เกี่ยวข้องกับการจำแนกข้อความบนสื่อสังคมออนไลน์สอได้รับไปในทิศทางเดียวกัน

คือ SVM ให้ประสิทธิภาพดีที่สุด โดย SVM ที่ใช้จะได้รับการปรับพารามิเตอร์ที่เหมาะสมแล้ววิเคราะห์ประสิทธิภาพการทำงานของการทำงานของการแจ้งเตือนข้อมูลเชิงลบและไม่เชิงลบ โดยเปรียบเทียบความถูกต้องกับผู้เชี่ยวชาญในการประเมินการจำแนก และการวัดความไว (Sensitivity) และความจำเพาะ (Specificity) โดยเส้นโค้ง Receiver Operating Characteristic (ROC)

งานวิจัยที่เกี่ยวข้อง

ระบบการแจ้งเตือนบนสื่อสังคมออนไลน์ได้เกิดขึ้นในปี 2548 ซึ่งอาจเป็นผลมาจากในปี 2547 Facebook ได้เปิดให้ดำเนินการ (Phillips, n.d.) ซึ่ง Facebook ได้เปลี่ยนแปลงการสื่อสารไปอย่างรวดเร็ว (Facebook Reports First Quarter 2015 Results, 2015) ระบบที่เกิดขึ้นนั้นเพื่อตอบสนองความต้องการของธุรกิจในการนำข้อมูลมาใช้ในการวิเคราะห์ทางการตลาดสำหรับระบบการแจ้งเตือนข้อมูลบน Social Media ที่มีผลต่อแบรนด์ธุรกิจ ในต่างประเทศนั้นก็ได้รับความสนใจมาตั้งแต่ปี 2548 ซึ่งมี 2 ผู้ให้บริการ ได้แก่ *Netvibes Dashboard Everything* ซึ่งมีการพัฒนาอย่างต่อเนื่องเรื่อยมาจนปัจจุบันสามารถรองรับได้ 81 ภาษา ซึ่งรวมภาษาไทยด้วย อีกทั้งยังกำหนด Zone location ของเนื้อหาข้อมูลที่น่าสนใจได้ จากแหล่ง Social Media ที่หลากหลาย ในขณะที่ทาง *Social Mention* สามารถจำแนกความรู้สึกของแต่ละโพสต์และมีการแบ่งสัดส่วนของข้อความทั้งหมด ซึ่งเป็นข้อได้เปรียบกว่า รวมทั้งการจัดคำสำคัญ 10 อันดับแรกของคำที่ค้นหาโดยผู้ใช้งานสูงสุด หรือผู้ที่มีอิทธิพลบนสื่อสังคมออนไลน์ แต่ระบบนี้มีข้อเสียที่ไม่รองรับภาษาไทย และแหล่งข้อมูลมี 8 แหล่งซึ่งไม่เพียงพอต่อความต้องการ

ปี 2549 ได้มีระบบใหม่ที่เกิดขึ้น 1 ระบบ คือ *TOPSY* ประกอบไปด้วย 3 ส่วนหลัก ส่วนแรก Social Search ซึ่งมีลักษณะคล้ายกับ Search Engine ของ Twitter สามารถแสดงผลข้อมูล Re-tweet ได้อย่างครบถ้วนและรองรับภาษาไทย ส่วนที่สอง Social Analytics

แสดงจำนวนโพสต์ที่เกี่ยวข้องกับคำสำคัญที่สนใจ นำมาเปรียบเทียบได้พร้อมกันมากที่สุด ครึ่งละ 3 คำสำคัญ และส่วนที่สาม Social Trends แสดงข้อมูลที่ได้รับ ความนิยมส่วนมาก จากการวิเคราะห์ระบบนี้พบว่า แหล่งข้อมูลยังไม่เพียงพอ และไม่สามารถแจ้งเตือนได้

ปี 2551 ทาง Google ได้เสนอระบบชื่อว่า *Google Alerts* ซึ่งสามารถรองรับได้ทุกภาษา จากการทดสอบพบว่า ข้อมูลการแจ้งเตือนที่ได้รับจาก *Google Alerts* ค่อนข้างน้อย และเป็นข้อมูลเชิงประชาสัมพันธ์มากกว่า ข้อมูลความคิดเห็นของผู้บริโภค และต่อมาในปี 2552 ระบบ *Talkwalker Alerts* ได้ถูกพัฒนาขึ้นรองรับ 22 ภาษา สามารถแสดงข้อมูลที่เป็นภาษาไทยได้ โดยเน้น ข้อมูลจากเว็บไซต์ประชาสัมพันธ์ แบ่งออกเป็น 3 ประเภท ได้แก่ News, Blogs และ Discussion แต่ไม่สามารถ จำแนกความคิดเห็นได้โดยระบบ และยังพบปัญหาที่ เหมือนกับ *Google Alerts* คือ พบข้อมูลที่แจ้งเตือน ไม่เกี่ยวข้องกับธุรกิจจำนวนมาก

ปี 2554 *thirdRiver* ได้เสนอระบบชื่อว่า *whirSocial* ซึ่งในปัจจุบันรองรับภาษาอังกฤษเป็นหลัก และดึงข้อมูล จาก 3 แหล่ง ได้แก่ Facebook, Twitter และ LinkedIn. โดยสามารถเลือกรับข้อมูลเมื่อมีส่วนเกี่ยวข้องหรือผู้ที่มีอิทธิพลกับธุรกิจนั้นได้ทันทีผ่านอีเมล แต่ระบบนี้ยังมี ข้อจำกัดทางด้านแหล่งข้อมูลและภาษาที่น้อยเกินไป

ปี 2555 บริษัท Pinterest ผู้ให้บริการข้อมูลตาม ที่ผู้ใช้สนใจ มีลักษณะคล้ายๆ Bookmark Online ได้ เสนอระบบชื่อว่า *Pin Alert* ที่สามารถแจ้งเตือนเมื่อพบ Keyword ที่สนใจแบบทันทีหรือเลือกความถี่เองตาม ที่ผู้ใช้สนใจ จากแหล่งข้อมูลทั้ง 6 แหล่ง ได้แก่ Facebook, Twitter, LinkedIn, Google+, YouTube และ Pinterest ผ่านอีเมล แต่ยังพบข้อจำกัดในเรื่องของภาษาซึ่งยังไม่ รองรับภาษาไทย มีแหล่งข้อมูลที่น้อยเกินไป และข้อมูล ส่วนมากมีลักษณะเป็นข่าวประกาศ ซึ่งมีประโยชน์น้อย ในการวิเคราะห์ด้านธุรกิจ

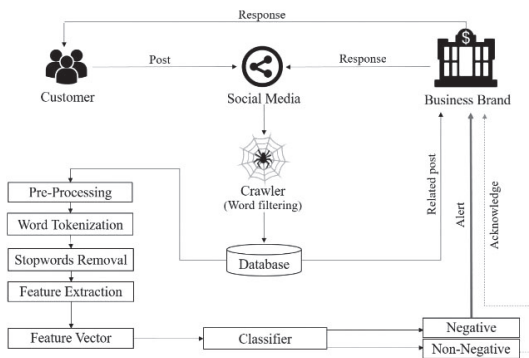
ปี 2556 ระบบใหม่ได้ถูกนำเสนอขึ้นมาภายใต้ชื่อว่า *Mention* รองรับภาษาได้หลากหลายกว่า 44 ภาษา

ซึ่งรวมไปถึงภาษาไทยด้วย โดยสามารถเลือกภาษา ที่ต้องการรับข้อมูลได้มากที่สุดถึงครึ่งละ 10 ภาษา เลือกรับข้อมูลโดยแบบ Location based นอกจากนี้ ยังพบว่า ระบบนี้สามารถดึงข้อมูลจากเว็บบอร์ด Pantip ซึ่งเป็นแหล่งข้อมูลที่สนใจได้ พร้อมทั้งมีการจำแนก ความคิดเห็นที่ตั้งเข้ามาเป็น 3 ประเภท ได้แก่ Positive, Neutral และ Negative รวมถึงมีฟังก์ชันในการสรุป ผลรวมตามความถี่ที่กำหนดเองได้ผ่านทางอีเมล และ แสดง Inferences Score ได้ ถึงแม้ว่าระบบจะสามารถ ทำงานได้มีประสิทธิภาพแต่ยังพบว่าข้อมูลข่าวโฆษณา หรือข้อความที่ไม่เกี่ยวข้องนั้นถูกดึงเข้ามาค่อนข้างมาก

จุดสำคัญที่น่าสนใจทั้งหมดของระบบแจ้งเตือน สื่อสังคมออนไลน์มีรายละเอียดตามตารางที่ 1 และตาราง ที่ 2 ซึ่งสามารถสรุปคร่าวๆ ได้ดังนี้ ระบบที่สามารถ รองรับภาษาไทยมี 6 ระบบ ระบบสามารถแจ้งเตือนผ่าน อีเมลได้มี 5 ระบบ และมี 3 ระบบที่สามารถวิเคราะห์ ความคิดเห็นได้ Mention เป็นระบบแจ้งเตือนเพียง ระบบเดียวที่สามารถดึงข้อมูลจากเว็บบอร์ด Pantip ได้ แต่ก็ไม่สามารถทำงานได้ตรงตามความต้องการของ ธุรกิจอย่างแท้จริง เพราะธุรกิจจำเป็นต้องมีการพิจารณา ปัจจัยต่างๆ โดยเฉพาะอย่างยิ่งความคิดเห็นเชิงลบที่มี การแพร่กระจายไปอย่างรวดเร็ว นอกจากนี้ระบบของ Mention ยังไม่สามารถติดตามความคิดเห็นที่สนใจได้ อีกครั้งหลังจากรับการแจ้งเตือนครั้งแรก เนื่องจากระบบ ไม่มีการติดตามความคิดเห็นนั้นๆ

วิธีการวิจัย

งานวิจัยนี้ได้เสนอวิธีการจำแนกความคิดเห็นบนสื่อ สังคมออนไลน์ ตามภาพที่ 1 โดยระบบจะตรวจสอบ ความคิดเห็นของลูกค้า ถ้าระบบพบข้อความที่ใกล้เคียง ในเชิงลบจะแจ้งเตือนไปยังธุรกิจ เพื่อให้ธุรกิจสามารถ จัดการแก้ไขปัญหาและตอบสนองไปยังลูกค้าโดยตรง หรือผ่านสื่อสังคมออนไลน์ หรือทั้งสองทาง ซึ่งจะก่อให้เกิด ผลดีต่อธุรกิจในการสร้างภาพลักษณ์ที่ดีกับตราสินค้า



ภาพที่ 1 ระบบแจ้งเตือนบนสื่อสังคมออนไลน์ที่นำเสนอ

ขั้นตอนวิจัยมีดังต่อไปนี้

1. ทำการเก็บข้อมูลความคิดเห็นภาษาไทยตามคำสำคัญที่ภาคธุรกิจสนใจจากเว็บบอร์ด Pantip โดยสร้างชุดคำสั่งในการดึงข้อมูลผ่าน HTML สำหรับส่วนหัวกระทู้และเนื้อกระทู้ และ AJAX สำหรับความคิดเห็นตอบกลับ ตั้งแต่เดือนมกราคม ปี 2556 ถึงเดือนตุลาคม ปี 2558 มาเก็บไว้ยังฐานข้อมูล MongoDB

2. นำข้อมูลเข้าสู่กระบวนการก่อนหน้าซึ่งประกอบไปด้วยการตัดคำภายในประโยค (Word Tokenization) โดยไลบรารี KUCUT (Satayamas, 2008) จากนั้นทำการกำจัดคำหยุด

3. ข้อมูลที่ผ่านการตัดคำและกำจัดคำหยุดแล้วจะถูกนำมาผ่านขั้นตอนของการสกัดคุณลักษณะโดยอาศัยน้ำหนักของคำนำมาคำนวณค่า Term Frequency (TF) และ Inverse Documents Frequency (IDF) ในขั้นตอนนี้จะแปลงจากคำให้อยู่ในรูปแบบของเวกเตอร์

4. เวกเตอร์คุณลักษณะ (Feature Vector) ที่ได้ นำไปฝึกฝนตัวจำแนกประเภท ในงานวิจัยนี้เลือกใช้ Support Vector Machine (SVM) ในการจำแนกข้อความออกเป็น 2 ประเภท ประกอบด้วยความคิดเห็นเชิงลบ และความคิดเห็นไม่เชิงลบ

5. ข้อความที่ถูกจำแนกด้วย SVM ว่าเป็นความคิดเห็นเชิงลบจะทำการแจ้งเตือนไปยังธุรกิจเพื่อแก้ไขปัญหา ซึ่งจะแทนด้วยเส้น Alert ดังภาพที่ 1

6. ข้อความที่ถูกจำแนกด้วย SVM ว่าเป็นความคิดเห็นเชิงลบจะทำการส่งไปเพื่อเป็นข้อมูลในการวิเคราะห์ทางธุรกิจต่อไป แทนด้วยเส้น Acknowledge ดังภาพที่ 1 เพื่อให้ธุรกิจสามารถนำไปวิจัยได้ในภายหลัง

7. การวัดประสิทธิภาพ ความคิดเห็นเชิงลบจะต้องมีความถูกต้องสูง รวมไปถึงการวัดประสิทธิภาพความไวและความจำเพาะ

ผลการวิจัย

การจำแนกความคิดเห็นสำหรับระบบแจ้งเตือนธุรกิจบนสื่อสังคมออนไลน์ ในครั้งนี้ใช้ไลบรารี SVM จาก Scikit-learn (Scikit-learn, 2010) โดยข้อมูลจะถูกแบ่งเป็นชุดฝึกฝน (Training set) 60% และชุดทดสอบ (Test set) 40% ตามลำดับ สำหรับพารามิเตอร์ของ SVM ที่ถูกปรับทำให้ได้รับค่าความถูกต้องสูงสุดที่พบในการทดลอง คือที่ 95% ได้แก่ ค่า $C = 50000$, ค่า $\text{degree} = 1$, Kernel = rbf และ $\text{gamma} = 0.1$ ซึ่งตาราง Confusion matrix แสดงให้เห็นว่า อัตราการจำแนกประเภทผิดของพารามิเตอร์ชุดนี้ประมาณ 5% แสดงดังตารางที่ 3 ซึ่งจำนวน 18045 แทนค่า True Positive, จำนวน 891 แทนค่า False Negative, จำนวน 397 แทนค่า False Positive และจำนวน 7666 แทนค่า True Negative

สำหรับการวัดค่าความไว (Sensitivity) และความจำเพาะ (Specificity) ของระบบสามารถวัดได้จากพื้นที่ใต้เส้นโค้ง Receiver Operating Characteristic (ROC) แสดงดังภาพที่ 2 โดยเส้นโค้งนี้จะหาความสัมพันธ์ระหว่างค่า True Positive (TP) และค่า False Positive (FP) ซึ่งค่า TP จะแทนค่าความไว และ FP จะแทนค่าความจำเพาะ สามารถคำนวณได้ตามสมการที่ (1.1) และ (1.2) ตามลำดับ

$$\text{sensitivity} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}} \quad (1.1)$$

และ

$$\text{specificity} = \frac{\text{true negative}}{\text{true negative} + \text{false positive}} \quad (1.2)$$

เมื่อ

- ค่า True positive (TP) คือ ความคิดเห็นเชิงลบ แจ้งเตือนอย่างถูกต้องเป็นเชิงลบ
- ค่า False positive (FP) คือ ความคิดเห็นไม่เชิงลบ แจ้งเตือนไม่ถูกต้องเป็นเชิงลบ
- ค่า True negative (TN) คือ ความคิดเห็นไม่เชิงลบ แจ้งเตือนอย่างถูกต้องเป็นไม่เชิงลบ
- ค่า False negative (FN) คือ ความคิดเห็นเชิงลบ แจ้งเตือนไม่ถูกต้องเป็นไม่เชิงลบ

โดยทั่วไปมีความเป็นไปได้ค่าของ TP และ FP จะสูงสุดพร้อมกัน เพราะเมื่อความไวสูงจะส่งผลให้ความจำเพาะต่ำ และความไวต่ำจะส่งผลให้ความจำเพาะสูง ซึ่งถ้าต้องการให้ค่าความจำเพาะสูง ค่า False positive ต้องลดลง เส้นโค้ง ROC จะพิจารณาจากพื้นที่ใช้เส้นโค้ง (AUC) ซึ่งจะสามารถบ่งบอกประสิทธิภาพของการจำแนกความคิดเห็นของลูกค้าในระบบแจ้งเตือน ซึ่งค่า AUC อยู่ที่ 85% ดังภาพที่ 1 ซึ่งเป็นค่าที่ค่อนข้างสูง ซึ่งพื้นที่นี้หมายถึงความสมดุลระหว่างความไวและความจำเพาะ หากมีค่ามากหมายถึงความไวและความจำเพาะมีค่าสูง โดยคิดเป็นอัตราความผิดพลาดเพียง 3 ใน 20 กรณีเท่านั้น

สรุปและอภิปรายผล

สำหรับงานวิจัยนี้ได้นำเสนอระบบแจ้งเตือนธุรกิจบนสื่อสังคมออนไลน์ โดยเก็บความคิดเห็นภาษาไทยที่เกี่ยวข้องกับธุรกิจจากเว็บบอร์ด Pantip โดยผู้เชี่ยวชาญจะจำแนกความคิดเห็นเป็น 2 กลุ่ม ได้แก่ เชิงลบและไม่เชิงลบ การเตรียมข้อมูลก่อนการจำแนกประเภทจะต้องทำการตัดคำแบ่งออกเป็นคำเดี่ยวหรือกลุ่มคำก่อน

โดยใช้ KUCUT ซึ่งเป็นตัวตัดคำภาษาไทยและทำการกำจัดคำหยุด จากนั้นใช้วิธีการ TF-IDF ในการสกัดคุณลักษณะ และใช้ SVM ในการจำแนกความคิดเห็นเป็น 2 ประเภทดังที่กล่าวไปข้างต้น ถ้าพบความคิดเห็นเชิงลบ ระบบจะทำการแจ้งเตือนไปยังธุรกิจ แต่หากเป็นความคิดเห็นที่ไม่เป็นเชิงลบระบบจะทำการระบุและจำเก็บไว้เพื่อให้ธุรกิจสามารถนำมาวิเคราะห์ได้ในภายหลัง สำหรับค่าพารามิเตอร์ของ SVM จะถูกปรับเพื่อหาค่าที่เหมาะสมและได้ความถูกต้องมากที่สุด ซึ่งในการทดลองนี้พบว่าการกำหนดพารามิเตอร์ต่างๆ ได้แก่ ค่า $C = 50000$, $\text{degree} = 1$, $\text{kernel} = \text{rbf}$ และ $\text{gamma} = 0.1$ จะได้รับค่าความถูกต้อง 95% ส่วนอัตราความผิดพลาดในการจำแนกผิดคิดเป็น 5% ในด้านความไวและความจำเพาะพบว่า ครอบคลุมพื้นที่ใต้เส้นโค้ง ROC คิดเป็น 85% ของพื้นที่ทั้งหมดซึ่งจัดว่าอยู่ในเกณฑ์ดี

สำหรับการวิจัยในอนาคตนั้นจะมุ่งเน้นไปในเรื่องการเพิ่มความถูกต้องให้แก่ระบบแจ้งเตือนธุรกิจบนสื่อสังคมออนไลน์ โดยจะเริ่มต้นจากการเพิ่มประสิทธิภาพการเรียนรู้ของเครื่องในการจำแนกข้อความ โดยเฉพาะตัวอย่างความคิดเห็นเชิงลบ นอกจากนี้ยังพบว่า ในขั้นตอนการตัดคำภาษาไทยด้วย KUCUT ยังมีความผิดพลาดเนื่องจากภาษาไทยมีโครงสร้างประโยคที่ซับซ้อนกว่าภาษาอังกฤษ ไม่มีช่องว่างระหว่างคำในประโยค และมีคำสแลงที่สร้างขึ้นใหม่โดยผู้ใช้ค่อนข้างมาก ด้วยเหตุนี้ทำให้การตัดคำเกิดความผิดพลาด ส่งผลให้ความถูกต้องที่ได้อาจจะลดลง ซึ่งหากมีวิธีการตัดคำที่มีประสิทธิภาพมากกว่านี้จะทำให้ความถูกต้องมากขึ้น สำหรับการทำงานของระบบในอนาคตจะพัฒนาระบบให้สามารถแจ้งเตือนได้ในหลายทางเพื่อให้ธุรกิจ อาทิ ธุรกิจค้าปลีก ซึ่งมีลูกค้าจำนวนมากมาแสดงความคิดเห็นบนสื่อสังคมออนไลน์สามารถจัดการกับความคิดเห็นบนสื่อสังคมออนไลน์ได้เร็วมากขึ้น

ตารางที่ 1 การเปรียบเทียบคุณสมบัติของเครื่องมือสื่อสังคมออนไลน์

System Name Technical Information	Social Mention	Netribes Dash board Everything	TOPSY	Google Alerts	Talk walker: Social Search Brand	Talk walker Alerts	Whir social	Pin Alerts	Mention
Support Thai language	-	●	●	●	●	●	-	-	●
Free Service	●	●	●	●	-	●	-	●	-
Alerts via email	-	-	-	●	-	●	●	●	●
Sentiment	Positive Neutral Negative	-	-	-	Positive Neutral Negative	-	-	-	Positive Neutral Negative
Location based filter	-	●	-	●	●	-	-	-	-
Frequency Report	- Anytime - Last hrs. - Last 12 hrs. - Last Day - Last Week - Month	-	- Urgent - Past 1 hrs. - Past 1 day - Past 7 days - Past 22 days - Past 30 days - All Time - Specific Range (Custom range)	- Urgent - Once a day - Once a week	-	- Urgent - Once a day - Once a week	Daily Email	- Once an hour - As it happens - Once a day - Once a week	E-mail - Daily - Weekly Desktop - Instantly - Hourly
Scoring	Passion, Reach	-	Topsy Sentiment Score	-	-	-	-	-	Inference Score
Top Keyword	●	-	●	-	●	-	-	-	-
Comparable	●	-	●	-	-	-	-	-	-

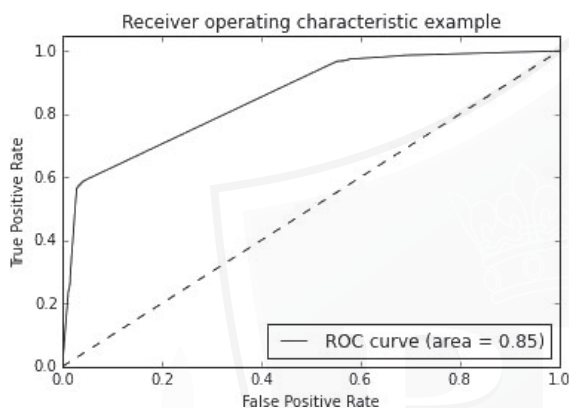
ตารางที่ 2 แหล่งข้อมูลที่ระบบแจ้งเตือนสามารถดึงได้

System Name Technical Information	Social Mention	Netvibes Dash board Everything	TOPSY	Google Alerts	Talk walker: Social Search Brand	Talk walker Alerts	Whir social	Pin Alerts	Mention
Facebook	●	-	-	●	●	-	●	●	●
Twitter	●	●	●	●	●	-	●	●	●
PANTIP	-	RSS Feed	-	-	-	-	-	-	●
YouTube	●	●	-	-	●	-	-	●	●
Instagram	-	-	-	-	●	-	-	-	●
Google+	-	●	-	●	●	-	●	●	●
LinkedIn	-	-	-	-	●	-	●	●	-
Pinterest	-	-	-	-	-	-	-	●	●
Website	●	●	-	●	●	●	●	●	●
Other Source	-	●	-	-	-	-	-	-	-

ตารางที่ 3 Confusion Matrix ข้อมูลทดสอบ

		Predicted	
		Non-negative	Negative
Actual	Non-negative	18045	891
	Negative	397	7666

งานวิจัยนี้ได้รับทุนสนับสนุนจากโครงการพัฒนาศักยภาพบุคลากรเพื่อการวิจัยและพัฒนาสำหรับภาคอุตสาหกรรม (NUI-RC) สำนักงานพัฒนาวิทยาศาสตร์และเทคโนโลยีแห่งชาติ (สวทช.)



ภาพที่ 2 เส้นโค้ง ROC แสดงประสิทธิภาพการจำแนกของ SVM

บรรณานุกรม

- Akaichi, J., Dhouioui, Z., José, M. & Pérez, L.-H. (2013). *Text Mining Facebook Status Updates for Sentiment Classification*. Sinaia: System Theory, Control and Computing (ICSTCC), 2013 17th International Conference.
- Binali, H., Wu, C. & Potdar, V. (2013). Computational Approaches for Emotion Detection in Text. In *4th IEEE International Conference on Digital Ecosystems and Technologies - Conference Proceedings of IEEE-DEST 2010, DEST 2010* (pp. 172-177). Digital Ecosystems and Technologies (DEST), 2010 4th IEEE International Conference on.
- Brynielsson, J., Johansson, F., Jonsson, C. & Westling, A. (2014). Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Security Informatics*, 3(1), 7.
- Facebook Reports First Quarter 2015 Results. (2015). Retrieved February 20, 2015, from <http://investor.fb.com/releasedetail.cfm?ReleaseID=908022>
- Isah, H., Trundle, P. & Neagu, D. (2014). *Social Media Analysis for Product Safety using Text Mining and Sentiment Analysis*. Bradford: Computational Intelligence (UKCI), 2014 14th UK Workshop on, IEEE.

- Naradhipa, A. R. & Purwarianti, A. (2011). Sentiment classification for Indonesian message in social media. In *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics* (pp. 1-4). IEEE.
- Pantip - Learn, Share & Fun. (n.d.). Retrieved December 20, 2015, from <http://pantip.com/>
- Pantip.com Site Info. (n.d.). *Alexa Internet*. Retrieved February 17, 2015, from <http://www.alexa.com/siteinfo/pantip.com>
- Phillips, S. (n.d.). *A brief history of Facebook*. *The Guardian*. Retrieved February 19, 2015, from <http://www.theguardian.com/technology/2007/jul/25/media.newmedia>
- Sarakit, P., Theeramunkong, T., Haruechaiyasak, C. & Okumura, M. (2015). Classifying emotion in Thai youtube comments. In *2015 6th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES)* (pp. 1-5). IEEE.
- Satayamas, V. (2008). *KUCut-Python*. Retrieved February 20, 2015, from <https://veer66.wordpress.com/2008/02/20/recent-postsuser-loginlog-in-using-openidwhat-is-openidusername-password-create-new-account-request-new-passwordlog-in-using-openidcancel-openid-login/>
- Scikit-learn. (2010). *sklearn.svm.SVC — scikit-learn 0.17 documentation*. Retrieved December 27, 2015, from <http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>
- Viriyavisuthisakul, S., Sanguansat, P., Charnkeitkong, P. & Haruechaiyasak, C. (2015). A comparison of similarity measures for online social media Thai text classification. In *2015 12th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)* (pp. 1-6). IEEE.



Name and Surname: Supatta Viriyavisuthisakul

Highest Education: Bachelor Degree (Information Technology),
Thai-Nichi Institute of Technology

University or Agency: Panyapiwat Institute of Management

Field of Expertise: Sentiment Analysis, Information Technology

Address: 85/1 Chaengwattana Rd., Bang Talad, Pakkred, Nonthaburi
11120



Name and Surname: Parinya Sanguansat

Highest Education: Ph.D. (Electrical Engineering), Chulalongkorn
University

University: Panyapiwat Institute of Management

Field of Expertise: Machine Learning, Image Processing

Address: 85/1 Chaengwattana Rd., Bang Talad, Pakkred, Nonthaburi
11120



Name and Surname: Pisit Charnkeitkong

Highest Education: Ph.D. (Electrical & Computer Engineering),
Yokohama National University

University or Agency: Panyapiwat Institute of Management

Field of Expertise: Network design and computer architecture

Address: 85/1 Chaengwattana Rd., Bang Talad, Pakkred, Nonthaburi
11120



Name and Surname: Choochart Haruechaiyasak

Highest Education: Ph.D. (Electrical and Computer Engineering),
University of Miami

University or Agency: NECTEC

Field of Expertise: Text mining, Sentiment analysis, Information
Retrieval (IR), Natural Language Processing (NLP) and web intelligence

Address: 112 Thailand Science Park, Phahonyothin Rd., Khlong Nueng,
Khlong Luang, Pathum Thani 12120