

การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน

Thai Sentiment Analysis of Product Review Online Using Support Vector Machine

วิสุตา เทศเมือง¹ และ นิเวศ จิระวิจิตรชัย²

คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม

2410/2 ถนนพหลโยธิน เขตจตุจักร กรุงเทพมหานคร 10900^{1, 2}

บทคัดย่อ

งานวิจัยนี้ได้นำเสนอวิธีการจำแนกความคิดเห็นโดยการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ด้านการบริการห้องพัก โรงแรม รีสอร์ท จาก Agoda Thailand และ Twitter Thailand ที่จดทะเบียนหลักทรัพย์ โดยเก็บข้อมูลจำนวน 2,890 ข้อมูล โดยใช้เทคนิคเหมืองข้อความวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ และสร้างแบบจำลองด้วยอัลกอริทึม 4 วิธี ได้แก่ ซัพพอร์ตเวกเตอร์แมทชีน ต้นไม้ตัดสินใจ นาอูฟเบย์ และเคเนี่ยเรสเนเบอร์ เพื่อเปรียบเทียบประสิทธิภาพการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ จากการทดลองพบว่าคุณลักษณะที่ดีที่สุดคือ ซัพพอร์ตเวกเตอร์แมทชีน ระดับรองลงมาเป็น นาอูฟเบย์ , ต้นไม้ตัดสินใจ และเคเนี่ยเรสเนเบอร์ ตามลำดับ

คำสำคัญ: เหมืองข้อความ, ซัพพอร์ตเวกเตอร์แมทชีน, นาอูฟเบย์, ต้นไม้ตัดสินใจ, เคเนี่ยเรสเนเบอร์

Abstract

This research purposes the methods to distinguish comments by analyzing online comments in Thai about online hotel service

product reviews from Agoda Thailand and Twitter Thailand, both of which registered for realty registration in Thailand. The research is done by collecting 2,890 samples, then analyze them by text mining and create 4 algorithms; Support Vector Machine, Decision Tree, Naïve-Bayes, K-Nearest Neighbor to compare the effectiveness of each methods in analyzing Thai comments of online products. The result suggests that Support Vector Machine is the most reliable method while Naïve-Bayes, Decision Tree, and K-Nearest Neighbor follow respectively.

Keywords: Text Mining, Support Vector Machine, Naïve-Bayes, Decision Tree, K-Nearest Neighbor

1. บทนำ

ปัจจุบันเว็บไซต์และ Social Media มีบทบาทมากขึ้นในการเผยแพร่ข้อมูล และการแสดงความคิดเห็นที่ให้บริการต่างๆ[1] เกี่ยวกับการรีวิวสินค้าออนไลน์ในหัวข้อต่างๆไม่ว่าจะสินค้า หรือการบริการบนเว็บไซต์ เช่น ความคิดเห็นรีวิวเกี่ยวกับการบริการด้านห้องพัก โรงแรม รีสอร์ท เว็บไซต์ Agoda Thailand จัดเป็นเว็บไซต์ที่ได้รับความนิยมเป็นอย่างมากในด้าน

การบริการห้องพัก รีสอร์ท โรงแรม ซึ่งได้มีผู้ที่เข้ามาใช้บริการการจองห้องพัก รีสอร์ท โรงแรม และเข้ามาติชมหรือแสดงความคิดเห็นต่อการให้บริการและการใช้บริการ หรือรีวิวสินค้าผ่านทางเว็บไซต์จำนวนมาก อีกทั้ง Twitter Thailand เป็น Social Media ในรูปแบบหนึ่งที่มีบทบาทสำคัญอย่างมากในการเผยแพร่ข่าวสารได้โดยตรงผ่าน Smartphones หรือ Mobile devices ซึ่งรองรับผู้ใช้งานได้เป็นจำนวนมาก อีกทั้งยังสามารถเข้าถึงข้อความ รีวิวได้โดยตรงจากผู้เข้ารับบริการ ผู้วิจัยจึงนำข้อมูล Agoda Thailand และ Twitter Thailand จำนวน 2,890 ข้อมูล จากการแสดงความคิดเห็นโดยการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ทำให้เพิ่มความสะดวกสบายแก่ผู้ที่เข้ามาใช้บริการหรือกำลังตัดสินใจที่จะใช้บริการ โดยผู้ที่เข้ารับบริการสามารถอ่านความคิดเห็น รีวิวที่มีผู้ใช้บริการก่อนหน้านี้ เพื่อประกอบการตัดสินใจในการใช้บริการของตนเอง [2, 3] จากความสัมพันธ์ของปัญหาดังกล่าว ผู้วิจัยจึงเกิดแนวคิดในการเพิ่มมูลค่าและสร้างประโยชน์ให้แก่ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ โดยวิธีการจำแนกความคิดเห็นตามเทคนิค ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ เนอฟเบย์ และเคเนียร์เนสเนเบอร์ เพื่อทำการทดสอบประสิทธิภาพแบบจำลองจากความสำคัญของปัญหาดังกล่าว เพื่อเป็นการเพิ่มมูลค่าและประโยชน์ให้แก่ผู้ที่เข้ามาใช้บริการหรือกำลังตัดสินใจที่จะใช้บริการได้ตัดสินใจง่ายขึ้น

2. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

2.1 เหมืองข้อความ

เหมืองข้อความหรือการค้นหาคำค้นหาคำรู้จากข้อความ (Knowledge-Discovery in Text - KDT) คือ กระบวนการที่กระทำ กับข้อความ (โดยส่วนใหญ่จะใช้กับข้อความที่มีจำนวนมาก) เพื่อค้นหารูปแบบแนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความนั้น โดยใช้หลักการค้นคืนสารสนเทศ (Information retrieval), เหมืองข้อมูล (Data mining), การเรียนรู้ของเครื่อง (Machine learning), สถิติ (Statistics) และหลักการประมวลผลทางภาษาศาสตร์ชาติ (Computational linguistics) การทำเหมืองข้อความคล้ายกับการทำ เหมืองข้อมูล แต่แตกต่างกันตรงที่การทำเหมืองข้อมูลมักจะใช้กับข้อมูล หรือฐานข้อมูลที่มีโครงสร้าง (Structured) ส่วนการทำเหมืองข้อความจะเน้นไปที่ข้อความ ที่ไม่มีโครงสร้าง (Unstructured) หรือกึ่งโครงสร้าง (Semi-structured) [4]

2.2 นาอ็ฟเบย์

นาอ็ฟเบย์ มีพื้นฐานมาจากกฎของเบย์ เป็นทฤษฎีทางด้านสถิติโดยนำความน่าจะเป็นมาใช้ประเมินความไม่แน่นอนให้เป็นตัวเลขได้ กล่าวถึงความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้น (A) ถ้ามีเหตุการณ์อีกเหตุการณ์หนึ่งเกิดมาแล้ว (B) สามารถเขียนให้อยู่ในรูปอย่างง่าย ดังสมการที่ 1 [5]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (1)$$

$P(A|B)$ คือ ความน่าจะเป็นที่เหตุการณ์ A จะเกิดขึ้น

ถ้าเหตุการณ์ B เกิดขึ้นแล้ว

$P(B|A)$ คือ ความน่าจะเป็นที่เหตุการณ์ B จะเกิดขึ้น

ถ้าเหตุการณ์ A เกิดขึ้นแล้ว

$P(A)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ A
 $P(B)$ คือ ความน่าจะเป็นที่จะเกิดเหตุการณ์ B
 วิธีการจำแนกหมวดหมู่โดยใช้หลัก การความน่าจะเป็น เป็นการแก้ปัญหาแบบ Classification สามารถคาดการณ์ผลลัพธ์และสามารถอธิบายได้ ทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ เป็นวิธีการจำแนกประเภท ข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยใช้ในการจำแนกหมวดหมู่เอกสารข้อความ (Text Classification) ได้ดี การทำงานไม่ซับซ้อนเหมาะกับกรณีของเซตตัวอย่างมีจำนวนมาก และคุณสมบัติ (Attribute) ไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็น ดังสมการที่ 2 [6]

$$P(A_1, A_2, \dots, A_n | C_j) = \prod_{i=1}^n P(A_i | C_j) \quad (2)$$

กลุ่ม C_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{A_1, A_2, \dots, A_n\}$ หรือ ใช้ สัญลักษณ์ ว่า $P=(A_1, A_2, \dots, A_n | C_j)$ โดยที่ \prod หมายถึงผลคูณของค่า $P=(A_i | C_j)$ ทั้งหมด $i = 1, 2, 3, \dots, n$ และ $j = 1, 2, 3, \dots, n$ ดังนั้นจะได้วิธีการจำแนกประเภทแบบเบย์อย่างง่ายดังสมการที่ 3

$$V_{NB} = \operatorname{argmax}_{C_j} \prod_{i=1}^n P(A_i | C_j) \quad (3)$$

2.3 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ (Decision Tree) [5] เป็นการเรียนรู้โดยการจำแนก (Classification) ชุดข้อมูลออกเป็นกลุ่ม (Class) ต่างๆ โดยใช้คุณสมบัติ (Attribute) ของข้อมูลในการแยกแยะ ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ทำให้ทราบคุณสมบัติใดของข้อมูลที่เป็นตัวกำหนดการจำแนก และคุณสมบัติแต่ละตัวของข้อมูลนั้นมีความสำคัญมากน้อยอย่างไร ซึ่งเป็น

ประโยชน์ที่จะช่วยให้ผู้ใช้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องยิ่งขึ้น ผลลัพธ์ของการเรียนรู้ต้นไม้ตัดสินใจ ประกอบด้วย 1. โหนดภายใน (Internal Node) 2. กิ่ง (Branch, Link) คือ ค่าคุณสมบัติของคุณสมบัติในโหนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนค่าคุณสมบัติของโหนดภายในนั้น 3. โหนดใบ (Leaf Node) คือ กลุ่ม (Class) ต่างๆ ซึ่งเป็นผลลัพธ์ในการจำแนกข้อมูล ซึ่งเป็นขั้นตอนวิธีหนึ่งที่ใช้ในการสร้างต้นไม้ตัดสินใจจะใช้ค่าที่วัดที่เรียกว่าค่าแแกน (Gain) เป็นตัวตัดสินใจว่าจะใช้แอตทริบิวต์ใดในการแบ่งข้อมูลเพื่อใช้ในการตัดสินใจ โดยวิธีการกำหนดโครงสร้างต้นไม้ตัดสินใจจะเป็นตามลำดับของค่าตัวชี้วัดหรือค่าเกินของแอตทริบิวต์ที่มีค่าสูงที่สุด ซึ่งเริ่มหาค่าสารสนเทศ (Information) แต่ละแอตทริบิวต์ซึ่งบ่งบอกความสามารถของแอตทริบิวต์ในการแยกแต่ละคลาส (Class) โดยที่ ตามสมการที่ 4 ดังนี้

$$I(s_1, s_2, \dots, s_n) = -\sum_{i=1}^n \frac{s_i}{S} \log_2 \frac{s_i}{S} \quad (4)$$

โดย S คือ จำนวนข้อมูลทั้งหมด ซึ่งมีจำนวนข้อมูล S_i ของ CLASS _{i}

n คือ จำนวน Class ทั้งหมด

จากนั้นหาค่าเอนโทรปี (Entropy) เป็นค่าของผลรวมแต่ละแอตทริบิวต์ A ที่สามารถแยกประเภทข้อมูลสารสนเทศ (Information) แต่ละคลาสของแต่ละค่าของแอตทริบิวต์ m จำนวน ดังสมการที่ 5

$$E(A) = \sum_{j=1}^m \frac{S_{1j} + \dots + S_{nj}}{S} I(S_{1j}, \dots, S_{2j}) \quad (5)$$

ดังนั้นค่าแแกน (Gain) ของแอตทริบิวต์ที่ถูกเลือกหาได้จากการแยกประเภทข้อมูลสารสนเทศทั้งหมดของแอตทริบิวต์นั้นได้ดังสมการที่ 6

$$Gain(A) = I(s_1, s_2, \dots, s_n) - E(A) \quad (6)$$

2.4 เคเนียร์สเนเบอร์ (K-Nearest Neighbor)

เป็นวิธีการในการจัดแบ่งคลาส หลักการของวิธีการนี้ [7] จะจำแนก ประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด k ตัวจากข้อมูลบนชุดข้อมูล ตัวอย่างทางงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่ หรือข้อมูลที่ป้อนถาม (input query instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้เคียงที่สุด k ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด k ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่สุดในกลุ่ม k ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง k ตัว ประกอบด้วยแอททริบิวต์ หลายตัวแปร X_i ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ k ซึ่งค่า นี้จะเป็นตัวบอกจำนวนของกรณี (case) ที่จะต้องค้นหาในการทำนายหากกรณีใหม่ อัลกอริทึมแบบ KNN ได้แก่ 1-NN , 2-NN , 3-NN , k -NN โดยค่า k ต้องระบุในการสร้างโมเดล [7, 8]

มาตรวัดความถูกต้อง (Distance Measure) การหาความยาวระหว่างจุดที่ต้องการ โดยใช้เครื่องมือและวิธี ต่างๆ งานวิจัยได้เลือกวิธีการหามาตรวัดความแม่นยำ โดย Euclidean Distance ระยะทาง ระหว่าง 2 จุด จุดที่จะวัดนั้นมีเงื่อนไขมีหลายค่าจากหลายมิติหรือขนาดขึ้นกับรูปแบบ ซึ่งสามารถพิสูจน์หาค่าได้ด้วยทฤษฎีของ Pythagorean เมื่อมีการใช้สูตรเพื่อหา ระยะทาง ขนาดของ Euclidean ระยะทางระหว่างจุด $P=(p_1,p_2,\dots,p_n)$ และ $Q=(q_1,q_2,\dots,q_n)$ ใน Euclidean หลายขนาดระบุได้เป็น ดังสมการที่ 7

$$\sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (7)$$

2.5 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

นำเสนอทฤษฎีโดย Cortes and Vapnik (1995) เพื่อลดความผิดพลาดจากการทำนาย (Minimize error) จัดเป็นเทคนิคที่ใช้ในการแก้ปัญหาทางด้านการรู้จำรูปแบบข้อมูล อาศัยหลักการจำแนกหมวดหมู่ข้อมูลด้วยการหาระนาบตัดสินใจและแบ่งข้อมูลออกเป็น 2 ส่วน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะระหว่างขอบเขตทั้ง 2 กลุ่มมากที่สุด (Optimal separating hyperplane) [9, 10] เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูล โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มี ระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด โดยใช้ฟังก์ชันแม่บสำหรับย้ายข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชันวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชัน (Kernel Function) Kernal Function ในสื่อตีพิมพ์เกี่ยวกับ SVM [11] จะเรียกตัวแปรในการตัดสินใจว่าคุณสมบัติและตัวแปรที่เปลี่ยนแปลงใช้ในการกำหนดระนาบหลายมิติ เรียกว่า คุณลักษณะ (feature) ส่วนการเลือกที่มีความเหมาะสมที่สุด เรียกว่า การคัดเลือกคุณลักษณะ (feature selection) จำนวนเซตของคุณลักษณะที่ใช้อธิบายในกรณีหนึ่ง (เช่น แถวของการค่าคาดการณ์) เรียกว่า เวกเตอร์ (vector) ดังนั้นจุดมุ่งหมายของตัวแบบ SVM [7] คือการได้ประโยชน์สูงสุดจากระนาบหลายมิติที่แบ่งแยกกลุ่มของเวกเตอร์ในกรณีนี้ด้วย Feature Space เหมาะใช้สำหรับข้อมูลที่มีมิติของข้อมูลสูง กำหนดให้ $(x_1,y_1), \dots, (x_n,y_n)$ เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูล

ตัวอย่าง m คือ จำนวนมิติข้อมูลเข้า และ y คือ ผลลัพธ์ มีค่า $+1$ หรือ -1 ดังสมการที่ 8

$$(xR\{+1,-1\} (5) i,yi),\dots,(xn,yn) \text{ เมื่อ } x m , y \in \mathbb{E} \quad (8)$$

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจ ซึ่งคำนวณได้ดังสมการที่ 9

$$(w * x)+b = 0 \quad (9)$$

เมื่อ w คือ ค่าน้ำหนักและ b คือค่า bias สมการ ใช้สำหรับจำแนกประเภทของข้อมูล ดังสมการที่ 10

$$(w*x)+b>0 \text{ ถ้า } yi = +1 \text{ และ}$$

$$(w*x)+b<0 \text{ ถ้า } yi = -1 \quad (10)$$

ในการวิเคราะห์ความคิดเห็นภาษาไทย เกี่ยวกับการรีวิวสินค้าออนไลน์ที่รวบรวมมาได้ และใช้แบบจำลองในการจำแนกข้อความ 4 อัลกอริทึม คือ โดยใช้ขั้นตอนซัพพอร์ตเวกเตอร์แมชชีน นาอ์ฟเบย์ ต้นไม้ตัดสินใจ เคเนียร์เนสเนเบอร์ เพื่อเปรียบเทียบประสิทธิภาพในการวิเคราะห์ความคิดเห็นภาษาไทย เกี่ยวกับการรีวิวสินค้าออนไลน์

2.6 วรรณกรรมที่เกี่ยวข้อง

จากการทำเหมืองข้อมูลความคิดเห็นได้มีการประยุกต์เข้ากับงานวิจัย [12] ซึ่งได้สกัดความคิดเห็นของสินค้าอิเล็กทรอนิกส์โดยแบ่งกลุ่มความคิดเห็นออกเป็นความคิดเห็นด้านบวกและด้านลบตามประเภทของสินค้า และแบ่งประเภทเชิงความหมาย (Semantic) ของคำที่แสดงความคิดเห็นด้านบวก (Positive) และลบ (Negative) ที่นิยมแบบ 1, 2 และ 3 คำ และนำความคิดเห็นด้านบวกและด้านลบเหล่านั้นไปสอนกับประโยคคำแนะนำจากผู้ให้คำแนะนำว่า

เป็นประเภทด้านบวกหรือด้านลบ โดยใช้เทคนิควิธีของเบย์อย่างง่าย (Naïve Bayes) แสดงผลเป็นค่าความน่าจะเป็นด้านบวกหรือด้านลบ ข้อเสียของงานวิจัยนี้คือข้อมูลที่ใช้ในการสอนประโยคหากมีการให้ค่าน้ำหนักผิดพลาดก็จะทำให้ผลผิดพลาดไปด้วย เช่นเดียวกับงานวิจัย[13] ได้ทำการทำเหมืองความคิดเห็นจากการแสดงความคิดเห็นจากเอกสารที่เกี่ยวข้องกับบทความวิจัย ซึ่งได้สืบค้นจากฐานข้อมูลอิเล็กทรอนิกส์และเว็บไซต์ โดยงานวิจัยนี้ได้ทำการใช้เทคนิควิธีซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) และเปรียบเทียบกับวิธีของเบย์อย่างง่าย และเทคนิควิธีอื่นๆ ซึ่งแตกต่างจากงานวิจัย [14] การวิเคราะห์ข้อความแสดงความคิดเห็นเกี่ยวกับโดเมนของโรงแรมซึ่งสรุปผลจากความคิดแยกประเภทตามคุณลักษณะคือ การบริการ อาหาร สภาพโรงแรม สถานที่ตั้ง ห้องพักและอื่นๆ เพื่อเพิ่มความสะดวกให้กับผู้ใช้งานที่ต้องการทราบข้อมูลเพียงอย่างเดียวหนึ่งเท่านั้น โดยผู้ใช้งานสามารถเลือกคุณลักษณะที่ต้องการได้ และเลือกโรงแรมเพื่อเปรียบเทียบคุณลักษณะที่เลือกกว่าโรงแรมใดดีกว่ากัน โดยจะแสดงในรูปแบบกราฟแท่งเปรียบเทียบกันระหว่างความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ จากความคิดเห็นทั้งหมดซึ่งได้จากการประมวลผลภาษา (Natural Processing Language) ซึ่งแยกคำจากการแสดงความรู้สึกของคำที่เก็บไว้ในฐานข้อมูล อีกทั้งงานวิจัยนี้ยังไม่แสดงรายละเอียดคำแนะนำเกี่ยวกับความคิดเห็นของผู้ใช้บริการแต่ละคน ซึ่งมีความจำเป็นในการแสดงผลความคิดเห็นทั้งหมด ดังนั้นงานวิจัยของเราจึงได้นำเสนอการวิเคราะห์ข้อความแนะนำจากผู้ให้บริการจากโดเมนของโรงแรม

เช่นเดียวกับงานวิจัย [14]

งานวิจัยครั้งนี้แตกต่างกับงานวิจัยที่อ้างอิงมาคือได้เปรียบเทียบเทคนิคการทำเหมืองข้อมูลโดยต้นไม้ตัดสินใจ (Decision Tree) และการเรียนรู้แบบเบย์อย่างง่าย (Naïve Bayes) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เคเนียร์เนสเนเบอร์ (K-Nearest Neighbor) เพื่อสรุปการแสดงความคิดเห็นในด้านบวก ด้านลบ ทำการเปรียบเทียบกับค่าคะแนนของผู้ใช้งานของบริการนั้นๆ ที่มีความขัดแย้งกันซึ่งแตกต่างจากงานวิจัยดังกล่าว ซึ่งงานวิจัยนี้ผู้วิจัย จะสามารถสรุปภาพรวมของการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์หรือบริการนั้นๆ ได้เลยว่าเป็นด้านบวกหรือด้านลบจากข้อความโดยแสดงเป็นกราฟคำแนะนำทั้งหมดเพื่อง่ายต่อการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ได้ง่าย

3. วิธีการดำเนินการวิจัย

วิธีการดำเนินงานของงานวิจัยนี้มุ่งเน้นเพื่อพัฒนาการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ จากแหล่งข้อมูลสาธารณะภาษาไทย มีกรอบแนวคิดในการพัฒนาการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ภาษาไทย ดังรูปที่ 1



รูปที่ 1 กรอบแนวคิดการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน

3.1 การเตรียมข้อมูล

เป็นขั้นตอนการรวบรวมการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ที่นำมาสร้างแบบจำลองเพื่อการเรียนรู้จำนวน 2,890 ข้อความ จากแหล่งข่าวสาธารณะที่ได้ทำการศึกษามา ดังตารางที่ 1

ตารางที่ 1 ปริมาณข้อมูลการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์

| แหล่งข่าวสาธารณะภาษาไทย | จำนวนข้อความ |
|-------------------------|--------------|
| Agoda Thailand | 1,573 |
| Twitter Thailand | 1,317 |

3.2 Document Parser

ขั้นตอนการดึงข้อมูลจากแหล่งข้อมูลโดยใช้โปรแกรม Import.io¹ จากนั้นทำการคัดกรองข้อมูลโดยกำจัดข้อมูลที่ซ้ำกันออก การแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสม สำหรับการสร้างแบบจำลอง โดย

การนำข้อ ความมาตัดคำผ่านโปรแกรมการตัดคำ ภาษาไทย (Thai Lexeme Tokenizer: LexTo)² และ ติดป้ายข้อมูลเพื่อใช้สำหรับการเรียนรู้ (Training Set) ทำการระบุมวลหมู่ของข้อความด้วยมือ ดังตารางที่ 2

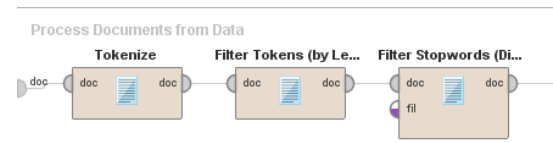
ตารางที่ 2 การติดป้ายข้อมูลเพื่อใช้สำหรับการเรียนรู้ (Training Set)

| Detail | Class |
|---|-------|
| การ บริการ ดี พนักงาน ดู น้อย ไป น้อย ส วน โรงแรม ดู สะอาด แต่ ก็ ยังมี สิ่งอำนวยความสะดวก ความสะดวก ใน ห้อง ยัง น้อย | 1 |
| สถานที่ วิว สวย ติด หาด สระ ว่ายน้ำ มี เยอะ มาก สระ ใหญ่ โรงแรม คอนเซิร์จ กว้าง | 2 |
| เป็น โรงแรม ที่ ดีเยี่ยม มากๆ สระ ว่ายน้ำ น้ำ สะอาด สวยงาม ห้องพัก สะอาด กว้างขวาง | 3 |
| บรรยากาศ ดีมาก วิว ทะเล สวย มาก สระ น้ำ ดีมาก คน ไม่ เยอะ ห้อง ไฟ สว่าง | 4 |
| โรงแรม นี้ อยู่ ใน ที่ตั้ง ดีมาก การ บริการ ดี เยี่ยม ห้อง สวย สะอาด อาหารเช้า เยอะ ดี | 5 |

3.3 การสกัดคุณลักษณะ (Feature extraction)

วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะ ความคิดเห็น คือการดึงคุณลักษณะ (Feature) ของ ความคิดเห็นออกมา ซึ่งการดึงคุณลักษณะออกมานั้น ต้องกำหนดก่อนว่าจะใช้อะไร เป็นตัวแทนคุณลักษณะ ของความคิดเห็น และใช้ค่าใดแทนคุณลักษณะความ คิดเห็นนั้น โดยการสร้างค่าจากข้อความที่เตรียมไว้ (Tokenize) กรองจำนวนคำตัวอักษรของคำ (Filter Tokenize) โดยกำหนดคำที่มีความยาวตั้งแต่ 2 ถึง 25 ตัวอักษรและกรอกคำหยุด (Filter Stopword) เป็นการ นำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมาย ของข่าวสารเปลี่ยนแปลง ในขั้นตอนนี้โปรแกรม Rapidminer 3 มาช่วยในการสกัดคุณลักษณะของ

ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ดังรูปที่ 2



รูปที่ 2 การสกัดคุณลักษณะของข้อความความคิดเห็น ภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ โดย โปรแกรม Rapidminer

3.4 การกำจัดคำหยุด (Stop-Word List Removal)

การนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ ความหมายของเอกสารเปลี่ยนแปลง คำ ที่ไม่มี นัยสำคัญ ในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มี ความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสาร แล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำ ให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับ คำ คำสรรพนามเป็นคำที่ใช้แทนคำนามที่กล่าวถึง มาแล้วในประโยค เป็นต้น จึงถือได้ว่าคำหยุดเป็น คุณลักษณะที่ไม่มีประโยชน์ในการจำแนก [2-3]

3.5 การหารากศัพท์ (Stemming)

การหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการ จัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลงและเพิ่ม ประสิทธิภาพในการค้นคืนหรือการจำแนกมวลหมู่ เช่น ใช้การหารากศัพท์ด้วยวิธี Porter Stemming Algorithm

3.6 การสร้างดัชนี (Indexing)

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนก มวลหมู่ของเอกสารซึ่งเป็นภาษารวมชาติโดยตรงได้

ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสาร เรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้วัตถุประสงค์ของการสร้างดัชนีคือการคำนวณค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร การสร้างดัชนีโดยทั่วไปที่นิยมใช้กัน จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร จากนั้นจะสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม [3] ซึ่งงานวิจัยนี้ใช้การทดลองให้ค่าน้ำหนักให้กับดัชนีดังต่อไปนี้ โดยให้

- f_{ik} = เป็นความถี่ของคำ i ในเอกสาร k
- N = จำนวนเอกสารทั้งหมดรวมของทุกกลุ่ม
- n_i = จำนวนเอกสารทั้งหมดที่มีคำ i เกิดขึ้น

2.4.3 Term Frequency-Inverse Document Frequency (TFIDF)

ค่านี้จะพิจารณาจากความถี่ของคำในเอกสารคูณกับฟังก์ชัน \log ของเอกสารทั้งหมดหารด้วยจำนวนเอกสารที่ปรากฏค่านั้นอยู่ ซึ่งเป็นวิธีการให้น้ำหนักแบบมาตรฐานที่ได้รับความนิยม ดังสมการที่ 11

$$a_{ik} = f_{ik} * \log\left(\frac{N}{n_i}\right) \quad (11)$$

3.7 การเลือกคุณลักษณะ (Feature Selection)

การคัดเลือกคุณลักษณะ (Feature Selection) คือการลดขนาดของข้อมูล โดยการทำให้ข้อมูลเดิมมีขนาดลดลงและสูญเสียลักษณะสำคัญของข้อมูลที่น้อยที่สุด โดยการใช้นิยามการคัดเลือกที่ต่างกันทำให้ได้คุณลักษณะที่ต่างกันด้วย การคัดเลือกคุณลักษณะ (Feature Selection) มีชื่อเรียกหลาย

อย่าง ในงานด้านสถิติมักเรียกว่า variable selection เพราะมองว่าคุณลักษณะแต่ละอัน คือตัวแปรแบบสุ่ม (random variable) บางทีเรียกว่า subset selection เพราะการเลือกของจำนวนหนึ่งออกจากของทั้งหมด ก็คือการเลือก subset นั้นเอง ซึ่งงานวิจัยนี้ใช้ค่าไครสแควร์ เป็นตัวลดคุณลักษณะของข้อมูล [2-3]

การทดสอบค่าไครสแควร์ (Chi-Square Test) เป็นวิธีการทดสอบเพื่อเปรียบเทียบข้อมูลที่อยู่ในรูปของความถี่หรือในรูปของสัดส่วน เช่น การศึกษาความคิดเห็น ความสนใจ หรือการยอมรับ ไม่สามารถวัดค่าออกมาเป็นตัวเลขที่แน่นอน แต่สามารถจำแนกออกเป็นหมวดหมู่ได้ เช่น มากที่สุด มาก ปานกลาง น้อย น้อยที่สุด หรือไม่ดี ค่าบวก ค่าลบ เป็นต้น ข้อมูลจากการเก็บรวบรวม จากตัวแปรที่เกี่ยวข้องแล้ว จำแนกออกมาเป็นความถี่หรือสัดส่วน ถ้าหากต้องการศึกษาว่าการแจกแจงความถี่ของข้อมูลที่ได้จากตัวแปรหนึ่ง เป็นไปลักษณะใด หรือถ้าหากต้องการเปรียบเทียบตัวแปร 2 กลุ่มหรือมากกว่า 2 กลุ่มว่ามีความสัมพันธ์กันหรือไม่

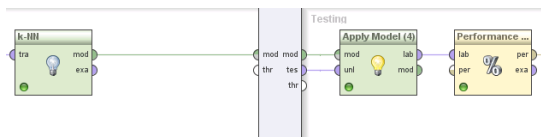
การคัดเลือกคุณลักษณะ คือการลดขนาดของข้อมูล โดยการทำให้ข้อมูลเดิม ที่มีขนาดลดลงและสูญเสียลักษณะสำคัญ ของข้อมูลน้อยที่สุด เทคนิคการคัดเลือกที่ต่างกันทำให้ได้คุณลักษณะต่างกันด้วย การคัดเลือกคุณลักษณะ (Feature Selection) มีชื่อเรียกหลายอย่าง ในงานด้านสถิติ มักเรียกว่า การคัดเลือกตัวแปร เพราะมองว่าคุณลักษณะแต่ละอัน คือตัวแปรแบบสุ่ม (random variable) บางทีเรียกว่า การคัดเลือกชุดข้อมูลย่อย เพราะการเลือกของจำนวนหนึ่งออกจากของทั้งหมดก็คือการเลือกชุดข้อมูลย่อยนั่นเอง

การประเมินค่าของ แอททริบิวต์ โดยคำนวณค่าไครสแควร์ (Chi-Square) ทางสถิติ ดังแสดงในสมการที่ (12)

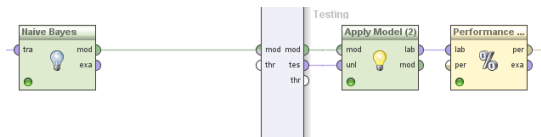
$$x^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \quad (12)$$

3.8 การสร้างแบบจำลอง (Model Builder)

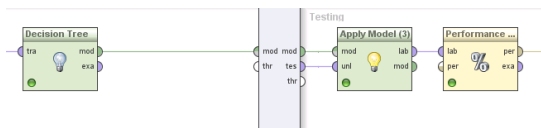
การสร้างแบบจำลองสำหรับการเรียนรู้ด้วยอัลกอริทึมการจำแนก 4 วิธี คือ Naïve Bayes ดังรูปที่ 3 และ SVM ดังรูปที่ 4, K-Nearest Neighbor ดังรูปที่ 5 และ Decision Tree ดังรูปที่ 6 เพื่อเปรียบเทียบประสิทธิภาพการจำแนกหมวดหมู่



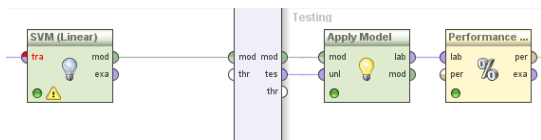
รูปที่ 2 สร้างแบบจำลองโดยใช้อัลกอริทึม K-Nearest Neighbor



รูปที่ 3 สร้างแบบจำลองโดยใช้อัลกอริทึม Naive Bayes



รูปที่ 4 สร้างแบบจำลองโดยใช้อัลกอริทึม Decision Tree



รูปที่ 5 สร้างแบบจำลองโดยใช้อัลกอริทึม SVM

3.9 การจำแนกหมวดหมู่ (Classifier)

การจำแนกหมวดหมู่ความคิดเห็นภาษาไทย เกี่ยวกับการรีวิวสินค้าออนไลน์ ในงานวิจัยนี้ จำแนกออกเป็น 5 หมวด คือ 1) ความสะอาด 2) สิ่งอำนวยความสะดวก

ความสะอาด 3) ความสะดวกสบายและคุณภาพห้องพัก 4) การให้บริการ 5) คุ่มค่ากับเงินที่จ่าย

3.10 การวัดประสิทธิภาพ (Testing and Evaluation)

การทดสอบแบบจำลองการจำแนกหมวดหมู่ ข้อมูลความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์แหล่งข้อมูลสาธารณะภาษาไทย พิจารณาจากค่าความถูกต้อง โดยใช้วิธีการประเมินความสามารถของแบบจำลอง วัดที่ประสิทธิภาพของการจำแนกข้อมูลตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือ การวัดค่าความถูกต้อง (Accuracy) คำนวณได้ดังสมการที่ 13 [15, 16]

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad (13)$$

C_j คือ หมวดหมู่ข้อมูลความคิดเห็นภาษาไทย เกี่ยวกับการรีวิวสินค้าออนไลน์ที่สนใจวัดประสิทธิภาพ

TP คือ จำนวนตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FP คือ จำนวนตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

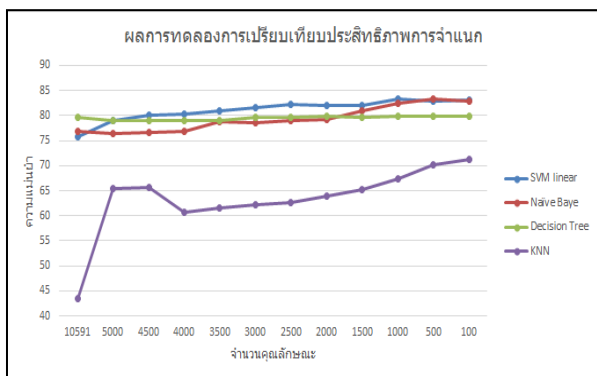
FN คือ จำนวนตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

TN คือ จำนวนตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

4. ผลการดำเนินงาน

จากผลการทดลองการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมทชีน และใช้วิธีการลดขนาดมิติของข้อมูลด้วยค่าไครสแควร์ (Chi-Square)

โดยพิจารณาเรียงลำดับจากค่าสูงสุด ซึ่งผลที่ได้จากการลดขนาดของมิติดังกล่าว จะถูกส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลยเรียนรู้ และทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง (Accuracy) เมื่อแทนค่าดัชนีด้วยวิธี TFIDF และทำการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ 4 อัลกอริทึมพบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ประสิทธิภาพในการจำแนกคุณลักษณะที่ดีที่สุดคือ ซัพพอร์ทเวกเตอร์แมชชีน (SVM linear) ที่คุณลักษณะที่ 1,000 มีค่าความแม่นยำ 83.38% ระดับรองลงมาเป็น เนออีฟเบย์ (Naive-Bayes) ที่คุณลักษณะที่ 500 มีค่าความแม่นยำ 83.31% ระดับรองลงมาเป็น ต้นไม้ตัดสินใจ (Decision Tree) ที่คุณลักษณะที่ 500 มีค่าความแม่นยำ 79.92% ระดับรองลงมาเป็น และเคเน็ยเรสเนเบอร์ (K-Nearest Neighbor) ที่คุณลักษณะที่ 100 มีค่าความแม่นยำ 71.26% ตามลำดับ ดังรูปที่ 6



รูปที่ 6 ผลการทดลองเปรียบเทียบแต่ละขั้นตอนวิธี

5. สรุป

งานวิจัยนี้ผู้วิจัยได้นำเสนอการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ทเวกเตอร์แมชชีน จาก Agoda

Thailand และ Twitter Thailand และสร้างแบบจำลองสร้างแบบจำลองด้วยอัลกอริทึม 4 วิธี ได้แก่ ซัพพอร์ทเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ นาอี่ฟเบย์ และเคเน็ยเรสเนเบอร์ เพื่อเปรียบเทียบประสิทธิภาพค่าความถูกต้อง (Accuracy) พบว่าเมื่อแทนค่าดัชนีด้วยค่า TFIDF และลดคุณลักษณะด้วย ค่าไครสแควร์ (Chi-Square) และเรียนรู้ด้วยเครื่องจักรการเรียนรู้ พบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ประสิทธิภาพในการจำแนกที่ดีที่สุด เมื่อลดมิติข้อมูลลงเหลือ 1,000 คุณลักษณะให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 83.38% ซึ่งจะให้ประสิทธิภาพการจำแนกข้อมูลความคิดเห็นที่เขียนจากความรู้สึกและอารมณ์ของ ได้เป็นอย่างดี และข้อมูลความคิดเห็นประเภทนี้สามารถลดมิติของข้อมูลด้วยวิธีไครสแควร์ (Chi-Square) ลงได้เป็นอย่างมาก ซึ่งการลดมิติของข้อมูลนั้นไม่ส่งผลกระทบต่อประสิทธิภาพในการจำแนกข้อมูลแต่อย่างใด จากกระบวนการที่น่าเสนอยังสามารถไปประยุกต์การให้บริการอื่นๆ ได้อีกด้วย

เอกสารอ้างอิง

- [1] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, "The role of Twitter during a natural disaster: Case study of 2011 Thai Flood" Proceedings of PICMET: Technology Management for Emerging Technologies Vol.12, pp. 2227-2232, 2012.
- [2] B. Pang, and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and

- Trends in Information Retrieval, Vol.2, 2008, pp. 1-135.
- [3] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of EMNLP 2002.
- [4] พัชรภรณ์ สิริคำฟู และมาลีรัตน์ โสदानิด "การจำแนกหมวดหมู่ข้อความข่าวสารภักย์พิบัติ อุทกภัยจากแหล่งข้อมูลสาธารณะภาษาไทย" The Eleventh National Conference on Computing and Information Technology, pp.124-130 ปี 2015.
- [5] A. Kongthon, C. Haruechaiyasak, J. Pailai, and S. Kongyoung, "The role of Twitter during a natural disaster: Case study of 2011 Thai Flood" Proceedings of PICMET: Technology Management for Emerging Technologies Vol.12, pp. 2227-2232, 2012.
- [6] G. Backfried, J. Gollner, G. Qirchmayr, K. Rainer, G. Kienast, G.Thallinger, C. Schmidt, and A. Peer, "Integration of Media Sources for Situation Analysis in the Different Phases of Disaster Management: The QuOIMA Project" Intelligence and Security Informatics Conference (EISIC), pp. 143-146, 2013.
- [7] Joachims, "Text categorization with support vector machines: Learning with many relevant features" Proc of the 10th European Conf. on Machine Learning, 1998.
- [8] Quinlan, J. R., C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
- [9] นิเวศ จิระวิจิตรชัย และคณะ "การจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติด้วยซัพพอร์ตเวกเตอร์แมชชีน" The 6th National Conference Computing and Information Technology, pp. 92-87, 2010.
- [10] อรทิพย์ เลื่องงาม, ชัยพร เขมะภาตะพันธ์ การจัดประเภทเอกสารด้วยวิธีเอชวีเอ็ม เพื่อการป้องกันเอกสารรั่วไหล การประชุมวิชาการ "นเรศวรวิจัย" ครั้งที่ 7, 3-12.
- [11] Joachims, "Text categorization with support vector machines: Learning with many relevant features". Proc. of the 10th European Conf. on Machine Learning, 1998.
- [12] Kushal Dave, Steve Lawrence and David M. Pennock. "Mining the peanut gallery: opinion extraction and semantic classification of product reviews" Proceeding of the 12th International Conference on World Wide Web. Budapest, Hungary, 20-24 May 2003, pp. 519-528.
- [13] KhairullahKhan, BaharumB. Baharudin, Aurangzeb Khan and Fazal-e-Malik, "MiningOpinion from Text Documents: A Survey" Proceeding of 3rd IEEE International Conference on Digital Ecosystems and Technologies. Istanbul, Turkey, 1-3 June 2009, pp. 217-222.

- [14] Alisa Kongthon, Niran Angkawattanawit, Chatchawal Sangkeettrakarn, PornpimonPalingoon and Choochart Haruechaiyasak, "Using an Opinion Mining Approach to Exploit Web Content in Order to ImproveCustomer Relationship Management Proceeding of (PICMIT'10) TechnologyManagement for Global Economic Growth,Phuket, Thailand, 18-22 July 2010, pp.1-6.
- [15] นิเวศ จิระวิชิตชัย และคณะ "การจัดหมวดหมู่เอกสารภาษาไทยด้วยเครือข่ายฟังก์ชันฐานรัศมี" National Conference on Information Technology, pp 302-307, 2010.
- [16] นิเวศ จิระวิชิตชัย การค้นหาเทคนิคเหมืองข้อมูลเพื่อสร้างโมเดลการวิเคราะห์โรคอัตโนมัติ มหาวิทยาลัยราชภัฏสวนสุนันทา, 2553.