วารสาร มทร.อีสาน
ฉบับวิทยาศาสตร์และเทคโนโลยี
ISSN 1906-215X

การยกระดับการจำแนกข้อความป้อนกลับของสินค้าโดยตัวจำแนกข้อความร่วมกับการวัดความสำคัญของคำ

Enhancing Classification of Item's Feedback using Text Classifier with Word Centrality Measures

วัชรีวรรณ จิตต์สกุล[1*] และสุนันทา สดสี[1]

Received: May, 2017; Accepted: July, 2017

บทคัดย่อ

งานวิจัยฉบับนี้ นำเสนอกระบวนการจำแนกข้อความป้อนกลับ (Feedback) ของสินค้าในพาณิชย์อิเล็กทรอนิกส์ โดยการใช้ตัวจำแนกข้อความ (Text Classifier) ร่วมกับการวัดความสำคัญของคำ (Word Centrality Measures) ในงานวิจัยฉบับนี้ข้อความป้อนกลับของสินค้า คือ ข้อความแสดงความคิดเห็นต่อสินค้าจากลูกค้าเมื่อซื้อสินค้าดังกล่าว แบ่งออกเป็นความคิดเห็นเชิงบวก และความคิดเห็นเชิงลบ ตัวจำแนกข้อความที่เหมาะสมได้มาจากการทดสอบประสิทธิภาพเปรียบเทียบระหว่างอัลกอริทึมที่อยู่ในพื้นฐาน 3 รูปแบบคือ พื้นฐานกฎ (Rule-Based) พื้นฐานโครงสร้างต้นไม้ (Tree Structure-Based) และพื้นฐานการเรียนรู้ (Learning-Based) ได้แก่ Conjunctive Rule, Random Forest และ Support Vector Machine ตามลำดับ ตัวจำแนกข้อความทำหน้าที่ระบุค่าความน่าจะเป็นในการแพร่กระจายของข้อความ (Probability Distribution) มีค่าอยู่ในช่วง [0, 1] ในส่วนของการวัดความสำคัญของคำ ใช้ทฤษฎีกราฟเพื่อแทนข้อความป้อนกลับของสินค้าด้วยกราฟข้อความ (Text Graph) และกำหนดความสำคัญของคำด้วยการวัดค่าความเป็นศูนย์กลาง (Centrality Measures) ในรูปแบบความน่าจะเป็นของความเป็นศูนย์กลางของข้อความ (Probability Centrality) ซึ่งมีค่าอยู่ในช่วง [0, 1] เช่นกัน ทั้งนี้ค่าความน่าจะเป็นในการแพร่กระจายของข้อความและความน่าจะเป็นของความเป็นศูนย์กลางของข้อความจะถูกนำมาใช้ในการจำแนกข้อความ ผลการดำเนินงานวิจัยแสดงให้เห็นว่ากระบวนการจำแนกข้อความป้อนกลับที่เสนอขึ้นในงานวิจัยฉบับนี้ สามารถจำแนกข้อความทดสอบจำนวน 3 ชุดข้อความเปรียบเทียบกับตัวจำแนกอื่น ๆ ได้อย่างมีประสิทธิภาพด้วยค่าเฉลี่ยความถูกต้องในการจำแนกร้อยละ 80.9

คำสำคัญ : พาณิชย์อิเล็กทรอนิกส์; การจำแนกข้อความป้อนกลับของสินค้า; ตัวจำแนกข้อความ; การวัดความสำคัญของคำ

[1] *Faculty of Information Technology, King Mongkut's University of Technology North Bangkok*
[*] *Corresponding Author E - mail Address: watchareewan.j@it.kmutnb.ac.th*

Abstract

This paper presents a novelty of item's feedback classification in e-commerce systems. This proposed work is developed based on a combination between a text classifier and word centrality measures. Herein, the item's feedback means comments written by customers to the purchased items, which are classified into positive or negative comments. In this work, the suitable text classifier is selected from three major types of classification: Rule-based, Tree structure-based, and Learning-based, which are Conjunctive Rule, Random Forest, and Support Vector Machine, respectively. In this work, the classifier is used for identifying the feedbacks in the probability distribution value [0, 1]. On the other hand, items' feedbacks are also represented by a graph, which is presenting a relationship among words. As well as, centrality measures are applied to determine each contained word centrality, and finalize to a probability centrality in [0, 1]. Both probability distribution and probability centrality, here, are applied to classify the item's feedback to positive or negative comments. The simulation results showed that the proposed classification method was efficient to classify three benchmark datasets, compared to other existing approaches with an average of classification accuracy 80.9 %.

Keywords: Electronic Commerce; Classification of Item's Feedback; Text Classifier; Word Centrality Measure

บทนำ

ณ ปัจจุบัน อินเตอร์เน็ตถือเป็นปัจจัยสำคัญในการสื่อสาร สามารถเชื่อมต่อผู้ใช้งานทั่วโลกให้สามารถ พูดคุยหรือแลกเปลี่ยนข้อมูลข่าวสารระหว่างกัน ด้วยเทคโนโลยีการให้บริการอินเตอร์เน็ตความเร็วสูงอย่างทั่วถึง จึงทำให้เกิดพาณิชย์อิเล็กทรอนิกส์ (E-Commerce) การซื้อ - ขายสินค้า หรือบริการผ่านอินเตอร์เน็ตขึ้น ซึ่งพาณิชย์อิเล็กทรอนิกส์ สามารถทำการติดต่อสื่อสาร และซื้อขายกันได้ทันทีลดระยะเวลาในการจัดซื้อหรือส่งสินค้า ส่งผลให้พาณิชย์อิเล็กทรอนิกส์ มีปริมาณเพิ่มขึ้นและมีการแข่งขันสูง โดยเป้าหมายพาณิชย์อิเล็กทรอนิกส์นั้น ผู้ขายต้องการขายสินค้า หรือบริการให้ได้จำนวนมาก ดังนั้นเพื่อทำการเพิ่มยอดขายให้สูงขึ้น จึงมีการใช้กลยุทธ์ต่าง ๆ เข้ามาใช้ เช่น การแนะนำสินค้า หรือการบริการตามความสนใจของลูกค้า และเมื่อลูกค้าซื้อสินค้าหรือบริการนั้น ๆ ไปใช้งาน ข้อมูลป้อนกลับ (Feedback) หรือความคิดเห็นของลูกค้าหลังจากซื้อสินค้า หรือบริการนั้น ๆ จะถูกนำมาใช้ในการปรับปรุงการแนะนำสินค้า หรือบริการต่าง ๆ ให้ตรงตามความสนใจของลูกค้ามากยิ่งขึ้น ซึ่งข้อมูลป้อนกลับนั้นมีทั้งการให้คะแนนเป็นตัวเลข เช่น 5 หมายถึง พอใจมาก หรือ 1 หมายถึงไม่พอใจ และมีการแสดงความคิดเห็นเป็นข้อความ ซึ่งต้องใช้กระบวนการที่ชาญฉลาด (Intelligent Methods) เพื่อประมวลผล เช่น การประมวลภาษาธรรมชาติ (Natural

Language Processing) [1] ออนโทโลยี (Ontology) [2] และวิธีการจำแนก (Classification) [3] เป็นต้น สำหรับการจำแนกข้อความมีผู้วิจัยจำนวนมากนำเสนองานวิจัยที่มีประสิทธิภาพ เช่น การจำแนกความคิดเห็นของข้อมูลคอมพิวเตอร์พกพา และร้านอาหารด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) [4] การจำแนกความคิดเห็นของผลิตภัณฑ์ด้วยฐานกฎ (Rule-Based) [5] เป็นต้น เมื่อพิจารณาข้อความความคิดเห็นซึ่งประกอบด้วยคำจำนวนมาก พบว่า คำที่ปรากฏมีความสัมพันธ์ซึ่งกันและกัน เช่น มีความหมายเหมือนหรือคล้ายกัน หรือคำดังกล่าวบ่อยครั้งจะปรากฏในความคิดเห็นด้วยกันเสมอ จึงทำให้ทฤษฎีกราฟ (Graph Theory) [6] ถูกนำมาใช้ในการวิเคราะห์ความสัมพันธ์ของคำในรูปแบบของกราฟข้อความ (Text Graph) [7] และสามารถวิเคราะห์หาความสำคัญของคำด้วยวิธีการวัดค่าความเป็นศูนย์กลาง (Centrality Measures) เช่น วิธี Dependency Graph [8] วิธี Graph Structure Model [9] Degree Centrality, In-degree Centrality, Out-degree Centrality และ Closeness Centrality [10] จากงานวิจัยดังกล่าวข้างต้น งานวิจัยฉบับนี้จึงมุ่งเน้นการยกระดับการจำแนกข้อความป้อนกลับของสินค้าในพาณิชย์อิเล็กทรอนิกส์โดยนำการวิเคราะห์ความสำคัญของคำร่วมกับตัวจำแนกข้อความ

วัตถุประสงค์การวิจัย

งานวิจัยฉบับนี้ถูกพัฒนาขึ้นโดยมีวัตถุประสงค์ดังนี้

1. เพื่อนำเสนอกระบวนการจำแนกข้อความป้อนกลับของสินค้าในพาณิชย์อิเล็กทรอนิกส์โดยใช้ตัวจำแนกข้อความ (Text Classifier) ร่วมกับการวัดความสำคัญของคำ (Word Centrality Measures)
2. เพื่อเปรียบเทียบความถูกต้องการจำแนกข้อความป้อนกลับของสินค้าในพาณิชย์อิเล็กทรอนิกส์

การทบทวนวรรณกรรม

การวิจัยเรื่องการยกระดับการจำแนกข้อความป้อนกลับของสินค้าโดยตัวจำแนกข้อความร่วมกับการวัดความสำคัญของคำครั้งนี้ ผู้วิจัยได้ทบทวน ศึกษาทฤษฎี และงานวิจัยที่เกี่ยวข้อง โดยแบ่งออกเป็น 4 หัวข้อหลัก คือ

1. พาณิชย์อิเล็กทรอนิกส์ (E-Commerce) และระบบแนะนำ (Recommendation Systems)

พาณิชย์อิเล็กทรอนิกส์ คือ ธุรกรรมทุกประเภทที่เกี่ยวข้องกับกิจกรรมเชิงพาณิชย์ทั้งในระดับองค์กรและส่วนบุคคล บนพื้นฐานของการประมวล และการส่งข้อมูลดิจิทัลที่มีทั้งข้อความ เสียง และภาพ ซึ่งพาณิชย์อิเล็กทรอนิกส์ สามารถทำการติดต่อสื่อสาร และซื้อขายกันได้ทันทีผ่านระบบอินเทอร์เน็ต ลดเวลาในการจัดซื้อ/ส่งมอบสินค้า เพิ่มยอดขายให้กับผู้ขาย และผู้ซื้อมีร้านค้าให้เลือกมากขึ้น [11] ในพาณิชย์อิเล็กทรอนิกส์ส่วนใหญ่จะนำเทคโนโลยีหรือกระบวนการที่ชาญฉลาด (Intelligent Methods) มาช่วยเพิ่มประสิทธิภาพในการขายสินค้า เช่น การผนวกระบบให้คำแนะนำสินค้าในกระบวนการขายสินค้า

โดยระบบให้คำแนะนำ (Recommender Systems) เป็นเทคโนโลยีหรือเครื่องมือในการแนะนำสินค้า (Item) ให้กับผู้ใช้งานระบบหรือลูกค้า โดยเสนอข้อมูลสินค้าที่ถูกกลั่นกรองให้แก่ผู้ใช้แต่ละคนเพื่อตอบสนองความต้องการของผู้ใช้งานระบบ ณ ขณะนั้น [12] เป็นระบบที่ช่วยแนะนำข้อมูลที่คาดว่าผู้ใช้จะให้ความสนใจ โดยเปรียบเทียบหรือวิเคราะห์คุณลักษณะของผู้ใช้งาน หรือสินค้าเพื่อสร้างคำแนะนำให้เหมาะสมกับผู้ใช้แต่ละบุคคล ระบบแนะนำพื้นฐานประกอบด้วย 3 องค์ประกอบหลัก คือ ผู้ใช้ หรือลูกค้า (User) สินค้า (Item) และข้อมูลป้อนกลับ (Feedback) โดยผู้ใช้สามารถแสดงความพึงพอใจของตนที่มีต่อสินค้าผ่านข้อมูลป้อนกลับ (Feedback) ซึ่งในที่นี้อาจอยู่ในรูปแบบของคะแนนหรือข้อความแสดงความคิดเห็น

2. การจำแนกข้อความ (Text Classification)

การจำแนกข้อความ (Text Classification) เป็นกระบวนการหนึ่งในการทำเหมืองข้อความ (Text Mining) โดยจะทำการสร้างระบบการจำแนกข้อความอัตโนมัติด้วยเทคนิคการเรียนรู้ด้วยตนเอง เช่น วิเคราะห์ข้อมูลการให้บริการ หรือการแนะนำสินค้า เป็นต้น [3] การจำแนกข้อความถูกนำมาใช้ในพาณิชย์อิเล็กทรอนิกส์ เพื่อเพิ่มประสิทธิภาพการขายสินค้าโดยประยุกต์ใช้ร่วมกับระบบให้คำแนะนำสินค้า เมื่อข้อมูลป้อนกลับของผู้ใช้งานหรือลูกค้าที่มีต่อสินค้า อยู่ในรูปแบบของข้อความแสดงความคิดเห็น ซึ่งผลลัพธ์ของการจำแนกข้อความแสดงความคิดเห็น จะทำให้ระบบทราบว่าผู้ใช้งานมีความพึงพอใจต่อสินค้านั้น ๆ หรือไม่ และส่งผลต่อการแนะนำข้อมูลสินค้าครั้งต่อไป

ในงานวิจัยฉบับนี้ ผู้วิจัยทำการศึกษาอัลกอริทึมเกี่ยวกับตัวจำแนกข้อความ (Text Classifier) ที่มีอยู่ในปัจจุบันโดยทำการศึกษาอัลกอริทึมที่มีลักษณะพื้นฐานที่แตกต่างกัน ในที่นี้ศึกษา 3 พื้นฐานหลัก ได้แก่ ตัวจำแนกข้อความบนพื้นฐานกฎ (Rule-Based) ตัวจำแนกข้อความบนพื้นฐานโครงสร้างต้นไม้ (Tree Structure-Based) และตัวจำแนกข้อความบนพื้นฐานการเรียนรู้ (Learning-Based) ดังนี้

ตัวจำแนกข้อความบนพื้นฐานกฎ ได้แก่ อัลกอริทึม Conjunctive Rule เป็นอัลกอริทึมประเภทหนึ่งของกฎความสัมพันธ์ ซึ่งกฎที่ใช้นั้นเป็นลักษณะของการให้เหตุผลแบบอุปนัย ซึ่งเป็นการสรุปผลจากการค้นหาความเป็นจริงที่ได้จากการสังเกต หรือการทดลองซ้ำหลาย ๆ ครั้ง จากการแบ่งเป็นกรณีย่อย ๆ หลังจากนั้นนำมาสรุปให้ได้ผลลัพธ์ [13]

ตัวจำแนกข้อความบนพื้นฐานโครงสร้างต้นไม้ ได้แก่ อัลกอริทึม Random Forest เป็นอัลกอริทึมที่มีลักษณะแบบไม่ตัดแต่งกิ่ง (Un-Pruned) หรือต้นไม้ถดถอย (Regression Trees) ซึ่งถูกสร้างจากการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูล และคุณลักษณะข้อมูลแล้วนำมาสร้างเป็นต้นไม้ตัดสินใจ ซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือกจะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ เรียกข้อมูลส่วนนี้ว่า Out-of-Bag (OOB) ซึ่งวิธีการนี้เรียกว่า Bagging ผลลัพธ์ที่ได้อย่างอิสระจากต้นไม้ตัดสินใจแต่ละต้นถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด อัลกอริทึม Random Forest ไม่จำเป็นต้องมีข้อมูลทดสอบเพื่อประมาณความผิดพลาดเพราะข้อมูล OOB นั้นถูกนำมาใช้ทดสอบต้นไม้ตัดสินใจแล้ว [14]

ตัวจำแนกข้อความบนพื้นฐานการเรียนรู้ ได้แก่ อัลกอริทึม Support Vector Machine เป็นอัลกอริทึมหนึ่งที่ใช้พื้นฐานการเรียนรู้ ซึ่งใช้วิธีการจำแนกกลุ่มข้อมูลที่อาศัยระนาบการตัดสินใจมาใช้ในแบ่งกลุ่มข้อมูลออกเป็นสองส่วน โดยใช้หลักการสร้างเส้นแบ่งกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด โดยที่ Support Vector Machine จะใช้แม็ปฟังก์ชัน (Mapping Function) สำหรับแปลงข้อมูลจากโดเมนเดิม Input Space ไปยังโดเมนที่เรียกว่า Feature Space และ

สร้างเคอเนลฟังก์ชัน (Kernel Function) บน Feature Space เพื่อใช้ในการวัดความคล้ายกันของข้อมูล สำหรับเคอเนลฟังก์ชันได้แก่ Linear, Polynomial Kernel, Radial Basic Function และ Sigmoid [15]

3. ทฤษฎีกราฟ (Graph Theory)

กราฟ (Graph) เป็นแบบจำลองทางคณิตศาสตร์ ถูกคิดค้นโดยนักคณิตศาสตร์ชาวสวิตเซอร์แลนด์ เลออนฮาร์ดออยเลอร์ (Leonhard Euler) [6] กราฟสามารถใช้แทนปัญหาในโลกของความเป็นจริง โดยจำลองปัญหาด้วยแผนภาพที่ประกอบด้วยจุด (Point) หรือเรียกว่า โหนด (Node) และเส้นที่เชื่อม ระหว่างจุด 2 จุด หรือเส้นเชื่อม (Edge) ตัวอย่างเช่น แผนภาพแสดงเส้นทางการบิน แผนภาพแสดงเส้นทาง รถไฟฟ้า และวงจรไฟฟ้า เป็นต้น อีกทั้งปัจจุบันยังมีการนำไปประยุกต์ใช้ในด้านต่าง ๆ อย่างกว้างขวาง อาทิเช่น ปัญหาทางด้านจิตวิทยา ภาษาศาสตร์ เทคโนโลยีคอมพิวเตอร์ และเศษฐศาสตร์ เป็นต้น ซึ่งในงานวิจัยฉบับนี้ กราฟถูกนำมาใช้ในการจำลองความสัมพันธ์ของคำในข้อความแสดงความคิดเห็น ต่อสินค้า โดยกราฟ (G) ประกอบด้วยโหนดสมาชิก (V) และเส้นเชื่อมระหว่างโหนด (E) [6] แสดง ดังสมการที่ (1)

$$G = (V, E) \tag{1}$$

โดยที่

V เป็นเซตจำกัดที่ไม่เป็นเซตว่างของสมาชิกที่เรียกว่า จุดยอด หรือโหนด (Node)

E เป็นเซตของเส้นเชื่อม (Link) ระหว่างโหนด

การวัดค่าความเป็นศูนย์กลาง (Centrality Measure) หรือความสำคัญของโหนด เป็นการวิเคราะห์การเชื่อมต่อของกราฟ ซึ่งมีการวัดอยู่หลายวิธี ในงานวิจัยนี้ศึกษาการวัด 3 รูปแบบ ดังนี้

1. วัดค่าความเป็นศูนย์กลางจากการคั่นกลาง (Betweenness Centrality: *BC*) เป็นการ กำหนดความสำคัญของโหนด โดยพิจารณาว่าโหนดที่มีสถานะเป็นสะพานเพื่อเชื่อมกลุ่มของโหนดต่าง ๆ ที่อยู่ห่างกันให้สามารถเข้าหากัน หรือเป็นตัวกลางในการติดต่อเชื่อมโยงระหว่างสมาชิกอื่น ๆ ดังนี้ [16]

$$BC(i) = \sum j \neq i \neq k \frac{\sigma_{jk}(i)}{\sigma_{jk}} \tag{2}$$

โดยที่

$BC(i)$ คือ ค่าความเป็นศูนย์กลางโดยวัดจากการคั่นกลางของโหนด i

σ_{jk} คือ จำนวนเส้นเชื่อมโยงที่สั้นที่สุดจากโหนด j ไปยังโหนด k

$\sigma_{jk}(i)$ คือ จำนวนเส้นเชื่อมโยงที่สั้นที่สุดจากโหนด j ไปยังโหนด k ที่ต้องผ่านโหนด i

2. วัดค่าความเป็นศูนย์กลางจากความใกล้ชิด (Closeness Centrality: *CC*) เป็นการกำหนด ความสำคัญของโหนดโดยพิจารณาจากความสามารถในการเข้าถึงโหนดอื่น ๆ ด้วยระยะทางที่สั้นที่สุด [17]

$$CC(i) = \frac{1}{\sum_j d(i,j)} \tag{3}$$

โดยที่

$CC(i)$ คือ ค่าความเป็นศูนย์กลางโดยวัดจากความใกล้ชิดของโหนด i

$d(i,j)$ คือ จำนวนเส้นเชื่อมโยงในเส้นทางที่สั้นที่สุดจากสมาชิกหนึ่งไปยังอีกสมาชิกหนึ่ง

3. วัดค่าความเป็นศูนย์กลางจากดีกรี (Degree Centrality: *DC*) เป็นการกำหนดความสำคัญของโหนดโดยพิจารณาจากจำนวนเส้นเชื่อมโยงทั้งหมดที่โหนดนั้น ๆ เชื่อมต่อกับโหนดอื่น ๆ [17]

$$DC(i) = \frac{\sum_j m_{ij}}{N} \tag{4}$$

โดยที่

$DC(i)$ คือ ค่าความเป็นศูนย์กลางโดยวัดจากระดับของโหนด i

m_{ij} = 1 ถ้ามีการเชื่อมต่อระหว่างโหนด

m_{ij} = 0 ถ้าไม่มีการเชื่อมต่อระหว่างกัน

N คือ จำนวนโหนดทั้งหมด

จากทฤษฎีกราฟ ถูกนำมาประยุกต์ใช้ในการสร้างกราฟข้อความ (Text Graph) ของข้อความแสดงความคิดเห็นต่อสินค้า โดยโหนด (Node) หมายถึง คำ (Word) และเส้นเชื่อมโยง (Link) ระหว่างโหนดแสดงถึงการปรากฏของคำนั้น ๆ ร่วมกับคำอื่น ๆ และการวัดค่าความเป็นศูนย์กลาง (Centrality Measure) ถูกนำมาใช้ในการระบุความสำคัญของคำ (Word Centrality) ในข้อความแสดงความคิดเห็นต่อสินค้า

4. งานวิจัยที่เกี่ยวข้อง

งานวิจัยฉบับนี้ ผู้วิจัยมุ่งเน้นการศึกษากระบวนการหรือตัวจำแนกข้อความในปัจจุบันเพื่อใช้ในการยกระดับประสิทธิภาพการจำแนกข้อความที่พัฒนาขึ้น โดยพิจารณา 3 พื้นฐานหลัก ได้แก่ ตัวจำแนกข้อความบนพื้นฐานกฎ (Rule-Based) ตัวจำแนกข้อความบนพื้นฐานโครงสร้างต้นไม้ (Tree Structure-Based) และตัวจำแนกข้อความบนพื้นฐานการเรียนรู้ (Learning-Based) โดยมีงานวิจัยที่เกี่ยวข้องสรุปได้ดังตารางที่ 1

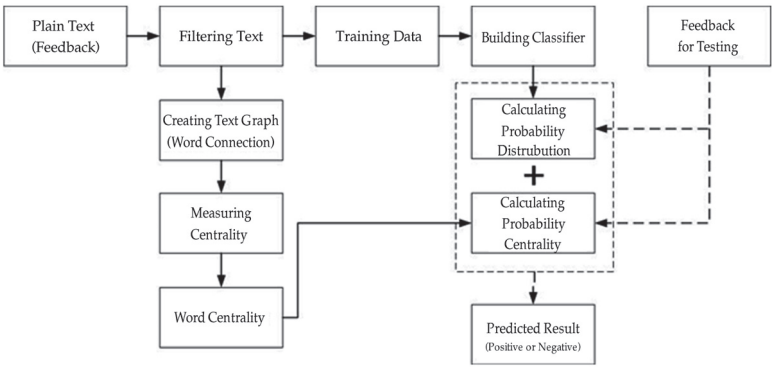
ตารางที่ 1 งานวิจัยที่เกี่ยวข้อง

งานวิจัย	ปี	อัลกอริทึม	ผลงานวิจัย
Sentiment analysis using product review data [18]	2015	- Random Forest - Support Vector Machine - Naive Bayes	ผลการเปรียบเทียบการจำแนกความคิดเห็นของข้อมูลหนังสือ เครื่องสำอาง อุปกรณ์อิเล็กทรอนิกส์ และอุปกรณ์ในบ้าน โดยจำแนกออกเป็น 5 ระดับ พบว่า อัลกอริทึม Random Forest ให้ค่า ROC สูงที่สุด โดยมีค่าเท่ากับ 0.92
Amazon Review Classification and Sentiment Analysis [5]	2015	- Rule-Based	ทำการจำแนกข้อความความคิดเห็นของข้อมูลผลิตภัณฑ์ โดยจำแนกออกเป็น 5 ระดับ คือ 1-5 จากนั้นนำ 5 ระดับมาแบ่งออกเป็น 3 กลุ่ม คือ Good(5, 4), Average(3) และ Bad(2, 1) แสดงผลของกลุ่มอยู่ในรูปแผนภาพ
Emotion Classification of Restaurant and Laptop Review Dataset: Semeval 2014 Task 4 [4]	2015	- Support Vector Machine - K-Nearest Neighbor	ผลการเปรียบเทียบการจำแนกความคิดเห็นของข้อมูลคอมพิวเตอร์พกพา และร้านอาหาร พบว่า อัลกอริทึม Support Vector Machine ให้ค่าความถูกต้องที่สูงกว่าอัลกอริทึม K-Nearest Neighbor
Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches [19]	2015	- Support Vector Machine - Naive Bayes	ผลการเปรียบเทียบการจำแนกความคิดเห็นของข้อมูลโทรศัพท์มือถือพบว่า อัลกอริทึม Support Vector Machine ให้ค่าความถูกต้องที่สูงกว่าอัลกอริทึม Naive Bayes
Sentiment Learning from Imbalanced Dataset: An Ensemble Based Method [20]	2014	- Support Vector Machine - Bagged Ensemble - SVM based Modified Bagging	ผลการเปรียบเทียบการจำแนกความคิดเห็นของข้อมูลกล้องถ่ายภาพ โทรศัพท์มือถือ และคอมพิวเตอร์พกพา พบว่า อัลกอริทึม SVM based Modified Bagging ให้ค่า ROC สูงที่สุด เท่ากับ 0.83, 0.795 และ 0.77 ตามลำดับ
IITP: Supervised Machine Learning for Aspect based Sentiment Analysis [21]	2014	- Random Forest	ทำการจำแนกความคิดเห็นของข้อมูลคอมพิวเตอร์พกพา และร้านอาหาร โดยมีค่าความถูกต้อง เท่ากับ 67.07% และ 67.37% ตามลำดับ

จากงานวิจัยที่เกี่ยวข้องในตารางที่ 1 ตัวจำแนกข้อความบนพื้นฐานกฎ ตัวจำแนกข้อความบนพื้นฐานโครงสร้างต้นไม้ และตัวจำแนกข้อความบนพื้นฐานการเรียนรู้ มุ่งเน้นการพัฒนา หรือประยุกต์ใช้ตัวจำแนกดังกล่าวเพียงอย่างเดียว ในงานวิจัยฉบับนี้ ผู้วิจัยประยุกต์ใช้การวัดความสำคัญของคำ (Word Centrality Measures) โดยนำเสนอในรูปแบบกราฟข้อความ (Text Graph) เพื่อเพิ่มประสิทธิภาพความถูกต้องในการจำแนก

กรอบแนวคิดในการวิจัย

จากการกำหนดวัตถุประสงค์งานวิจัย การทบทวน ศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้อง ผู้วิจัยจึงออกแบบกรอบแนวคิดในการยกระดับการจำแนกข้อความป้อนกลับของสินค้าโดยนำตัวจำแนกข้อความทำงานร่วมกับการวัดความสำคัญของคำ แสดงดังรูปที่ 1



รูปที่ 1 กรอบแนวคิดงานวิจัย

ในงานวิจัยฉบับนี้ ข้อมูลป้อนกลับของสินค้า หมายถึง ข้อความแสดงความคิดเห็นต่อสินค้า ซึ่งผู้วิจัยมีเป้าหมายในการจำแนกความเห็นดังกล่าวออกเป็น ความคิดเห็นเชิงบวก หรือความคิดเห็นเชิงลบ เพื่อแสดงความพึงพอใจของลูกค้าต่อสินค้านั้น ๆ และส่งผลต่อกระบวนการแนะนำสินค้าในอนาคตได้ โดยมีขั้นตอนการทำงานหลักดังนี้

1. ข้อมูลป้อนกลับของสินค้า หรือข้อความแสดงความคิดเห็นต่อสินค้า จะถูกกรองเพื่อคัดแยกออกเป็นคำ (Word)

1.1 คำที่ได้จะถูกนำมาสร้างเป็นกราฟข้อความ (Text Graph) เพื่อแสดงความสัมพันธ์ระหว่างคำ โดยกราฟที่ได้อยู่ในรูปแบบของกราฟแบบไม่มีทิศทาง (Undirected Graph)

วัดความสำคัญของคำในกราฟโดยการใช้การวัดค่าความเป็นศูนย์กลาง (Centrality Measure) ซึ่งค่าความสำคัญของคำนี้ จะถูกนำไปคำนวณค่า Probability Centrality ของข้อความแสดงความคิดเห็นออกเป็น Probability Centrality$_p$ และ Probability Centrality$_n$ คือ ความน่าจะเป็นของคำภายในข้อความดังกล่าว ส่งผลให้ข้อความเป็นเชิงบวก หรือเชิงลบ ซึ่งค่าดังกล่าวจะอยู่ในช่วง [0, 1]

1.2 คำที่ได้จะถูกนำมาเรียนรู้ และสร้างเป็นโมเดลของตัวจำแนกข้อความ (Text Classifier)

ตัวจำแนกข้อความทำหน้าที่คำนวณค่าการกระจายของค่าความน่าจะเป็น (Probability Distribution) โดยในงานวิจัยฉบับนี้เป็นแบบการกระจายที่ข้อมูลไม่ต่อเนื่อง (Discrete Probability Distribution) ซึ่งแบ่งความน่าจะเป็นของข้อความแสดงความคิดเห็นออกเป็น Probability Distribution$_p$ และ Probability Distribution$_n$ คือ ความน่าจะเป็นที่ข้อความดังกล่าวจะเป็นเชิงบวกหรือเชิงลบ ซึ่งค่าดังกล่าวจะอยู่ในช่วง [0, 1]

1.3 ในการจำแนกข้อความแสดงความคิดเห็นจะพิจารณาค่า Probability Centrality ร่วมกับค่า Probability Distribution เพื่อระบุว่าข้อความดังกล่าวเป็นข้อความเชิงบวก หรือข้อความเชิงลบ

1.3.1 ข้อความเชิงบวก เมื่อ (Probability Centrality$_p$> Probability Centrality$_n$) AND (Probability Distribution$_p$> Probability Distribution$_n$) == TRUE

1.3.2 ข้อความเชิงลบ เมื่อผลลัพธ์ในข้อ 1.3.1 เป็น FALSE

ผลการดำเนินงานวิจัย

ในส่วนนี้กล่าวถึงการประเมินประสิทธิภาพกระบวนการที่นำเสนอในงานวิจัยนี้ โดยแบ่งออกเป็น 3 หัวข้อ คือ 1) การคัดเลือกตัวจำแนกข้อความที่เหมาะสม 2) การวิเคราะห์กราฟข้อความ และ 3) การเปรียบเทียบประสิทธิภาพในการจำแนก

ในการประเมินประสิทธิภาพกระบวนการที่นำเสนอ ผู้วิจัยทำการทดสอบกับชุดข้อมูลจากฐานข้อมูล UCI Machine Learning Repository จำนวน 3 ชุดข้อมูล [22] ได้แก่ ข้อมูลแสดงความคิดเห็นเกี่ยวกับภาพยนตร์ รวบรวมจากเว็บไซต์ www.imdb.com ข้อมูลแสดงความคิดเห็นเกี่ยวกับร้านอาหารรวบรวมจากเว็บไซต์ www.yelp.com และข้อมูลแสดงความคิดเห็นเกี่ยวกับสินค้า รวบรวมจากเว็บไซต์ www.amazon.com โดยทั้ง 3 เว็บไซต์เป็นตัวแทนของพาณิชย์อิเล็กทรอนิกส์ในปัจจุบัน ในส่วนการวัดและประเมินผล Cross-validation 10 folds ถูกนำมาใช้กับข้อมูลทั้ง 3 ชุด และทดสอบซ้ำจำนวน 5 ครั้ง

1. การคัดเลือกตัวจำแนกข้อความที่เหมาะสม

จากการศึกษาตัวจำแนกข้อความพื้นฐาน 3 รูปแบบ คือ พื้นฐานกฎ (Rule-Based) พื้นฐานโครงสร้างต้นไม้ (Tree Structure-Based) และพื้นฐานการเรียนรู้ (Learning-Based) ได้แก่ Conjunctive Rule, Random Forest และ Support Vector Machine ผู้วิจัยทำการคัดเลือกตัวจำแนกข้อความที่เหมาะสม โดยประเมินผลลัพธ์ประสิทธิภาพการจำแนกด้าน [23] ความถูกต้อง (Accuracy) ค่าความผิดพลาด (Mean Absolute Error: MAE) ค่าความเที่ยงตรง (Precision) ค่า F-measure และ ROC ตามลำดับในข้อมูลทดสอบ จำนวน 3 ชุด ดังตารางที่ 2

ตารางที่ 2 การเปรียบเทียบประสิทธิภาพการจำแนก

Datasets	Classifier	Accuracy (%)	MAE	Precision	F-Measure	ROC Area
www.imdb.com	Random Forest	78.333	0.408	0.783	0.783	0.827
	Conjunctive Rule	56.333	0.481	0.595	0.523	0.563
	Support Vector Machine	76.667	0.233	0.768	0.766	0.767
www.yelp.com	Random Forest	80.333	0.338	0.819	0.801	0.915
	Conjunctive Rule	56.667	0.472	0.697	0.481	0.567
	Support Vector Machine	74.667	0.253	0.753	0.745	0.747
www.amazon.com	Random Forest	83.000	0.340	0.830	0.830	0.898
	Conjunctive Rule	55.000	0.4738	0.711	0.444	0.550
	Support Vector Machine	74.667	0.253	0.753	0.745	0.747

จากตารางที่ 2 แสดงให้เห็นว่า ตัวจำแนกข้อความประเภทพื้นฐานโครงสร้างต้นไม้ Random Forest แสดงประสิทธิภาพการจำแนกข้อความได้ดีกว่าตัวจำแนกข้อความอื่น ๆ ด้วยค่าเฉลี่ยความถูกต้อง (Accuracy) ร้อยละ 80 ค่าเฉลี่ยความผิดพลาด (Mean Absolute Error: MAE) 0.36 ค่าเฉลี่ยความเที่ยงตรง (Precision) 0.81 ค่าเฉลี่ย F-measure 0.8 และค่าเฉลี่ย ROC 0.88 แสดงดังตารางที่ 3 ดังนั้น Random Forest จึงเหมาะสมที่จะถูกคัดเลือกเป็นตัวจำแนกข้อมูลเพื่อคำนวณค่า Probability Distribution ในงานวิจัยฉบับนี้

ตารางที่ 3 ค่าเฉลี่ยประสิทธิภาพการจำแนกข้อความ

Classifier	Accuracy (%)	MAE	Average Precision	F-Measure	ROC Area
Random Forest	80.55533	0.362	0.810667	0.804667	0.88
Conjunctive Rule	56	0.4756	0.667667	0.482667	0.56
Support Vector Machine	75.33367	0.246333	0.758	0.752	0.753667

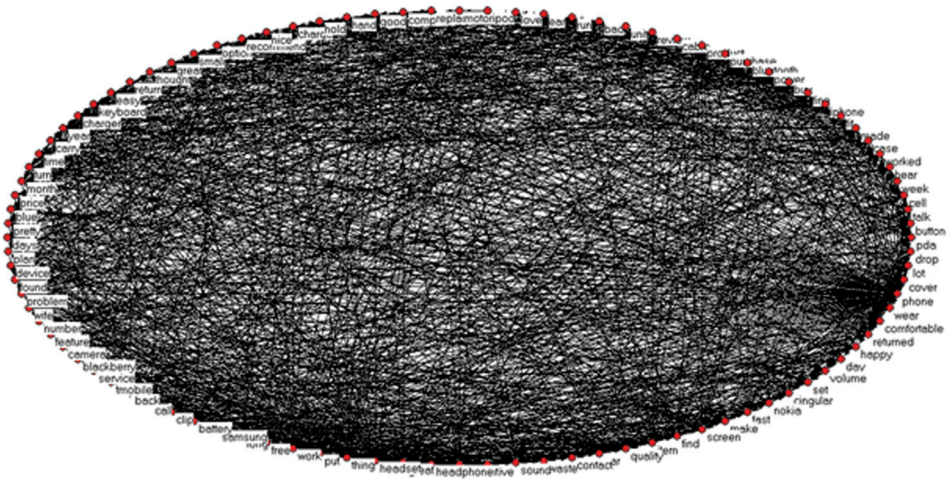
2. การวิเคราะห์กราฟข้อความ ในส่วนของการวัดความสำคัญของคำ งานวิจัยฉบับนี้ใช้ทฤษฎีกราฟเพื่อแทนข้อความป้อนกลับของสินค้าด้วยกราฟข้อความ (Word Graph) และกำหนดความสำคัญของคำด้วยการวัดค่าความเป็นศูนย์กลาง (Centrality Measures) ด้วย การวัดค่าความเป็นศูนย์กลางจากการคั่นกลาง (Betweenness Centrality: *BC*) การวัดค่าความเป็นศูนย์กลางจากความใกล้ชิด (Closeness Centrality: *CC*) และการวัดค่าความเป็นศูนย์กลางจากดีกรี (Degree Centrality: *DC*) เมื่อนำชุดข้อความทดสอบทั้ง 3 ชุด มาสร้างกราฟข้อความ (Text Graph)

ตารางที่ 4 การวิเคราะห์กราฟข้อความในชุดข้อมูลทดสอบ

Characteristics	Datasets		
	www.imdb.com	www.yelp.com	www.amazon.com
Network size	100	100	100
Number of Links (Edges)	1117	1141	1289
Density	0.1117	0.1141	0.1289
Average degree	22.3400	22.8200	25.7800

ตารางที่ 4 แสดงการวิเคราะห์กราฟข้อความในชุดข้อความทดสอบเมื่อกำหนดให้จำนวนคำ (Word) สูงที่สุด คือ 100 คำ ผลการวิเคราะห์แสดงให้เห็นว่า ความสัมพันธ์ของคำในชุดข้อมูลทดสอบ www.amazon.com คำในข้อมูลนี้มีการเชื่อมโยงกันจำนวนมากมีความสัมพันธ์กันสูงที่สุด โดยพิจารณาจากจำนวนเส้นเชื่อม (Link) ความหนาแน่น (Density) และค่าเฉลี่ยดีกรี (Average Degree) ที่มีค่าสูงกว่า www.imdb.com และ www.yelp.com

รูปที่ 2 แสดงกราฟข้อความในชุดข้อความทดสอบ www.amazon.com เมื่อกำหนดให้โหนดแสดงถึงคำ และเส้นเชื่อมแสดงถึงความสัมพันธ์ระหว่างคำ เมื่อคำ A ปรากฏแล้ว บ่อยครั้งคำ B จะปรากฏด้วย เส้นเชื่อม (Link) ระหว่างคำ A และ B จึงเกิดขึ้น



รูปที่ 2 กราฟข้อความของชุดข้อมูล www.amazon.com

3. การเปรียบเทียบค่าความถูกต้องในการจำแนกข้อมูล

ส่วนนี้แสดงการเปรียบเทียบค่าความถูกต้องในการจำแนกข้อมูลโดยการใช้ตัวจำแนกข้อความ Random Forest ร่วมกับการวัดความสำคัญของคำ (Word Centrality Measures) ทั้ง 3 รูปแบบ *BC, CC,* และ *DC* และเปรียบเทียบผลลัพธ์กับตัวจำแนกข้อความดั้งเดิม Conjunctive Rule, Support Vector Machine, และ Random Forest

ตารางที่ 5 แสดงให้เห็นว่า กระบวนการจำแนกข้อความที่นำเสนอในงานวิจัยฉบับนี้ เมื่อใช้ตัวจำแนกข้อความ Random Forest ร่วมกับการวัดความสำคัญของคำโดยพิจารณาจากความใกล้ชิด *CC* ให้ค่าเฉลี่ยความถูกต้องในการจำแนกข้อมูลทดสอบทั้ง 3 ชุด สูงกว่าตัวจำแนกข้อความอื่น ๆ

ตารางที่ 5 การเปรียบเทียบค่าความถูกต้องในการจำแนกข้อมูลทดสอบ 3 ชุดข้อมูล

Average Accuracy (%)	Conjunctive Rule	Support Vector Machine	Random Forest	Proposed Work
	56	75.33367	80.55533	80.9

ตัวอย่างการจำแนกข้อความป้อนกลับของสินค้าในพาณิชย์อิเล็กทรอนิกส์ โดยกระบวนการจำแนกข้อความที่นำเสนอในงานวิจัยฉบับนี้ ผลลัพธ์ในตารางที่ 6 แสดงให้เห็นว่า ข้อความป้อนกลับทั้ง 3 ข้อความสามารถจำแนกได้เป็นข้อความเชิงลบ ข้อความเชิงบวก และข้อความเชิงบวก ตามลำดับ โดยสรุปข้อความเชิงลบ หมายถึง ลูกค้าซื้อสินค้าดังกล่าวแสดงความคิดเห็นต่อสินค้านั้นในเชิงลบ และข้อความเชิงบวก หมายถึง ลูกค้าซื้อสินค้าดังกล่าว และแสดงความคิดเห็นต่อสินค้านั้นในเชิงบวก แสดงถึงความพึงพอใจในสินค้า

ตารางที่ 6 ตัวอย่างการจำแนกข้อความด้วยกระบวนการที่นำเสนอ

Sentence	Predicted Type of Comment	Actual Target
All in all its an insult to one's intelligence and a huge waste of money.	0	0
I ate there twice on my last visit, and especially enjoyed the salmon salad.	1	1
It was quite comfortable in the ear.	1	1

หมายเหตุ Negative = 0
Positive = 1

สรุปผลการวิจัย

งานวิจัยฉบับนี้นำเสนอการจำแนกข้อความแบบใหม่ โดยการใช้ตัวจำแนกข้อความ Random Forest ร่วมกับการวัดความสำคัญของคำโดยพิจารณาจากความใกล้ชิด *CC* เพื่อประยุกต์ใช้ในการจำแนกข้อความป้อนกลับของลูกค้าเมื่อซื้อสินค้าในพาณิชย์อิเล็กทรอนิกส์ โดยตัวจำแนกที่นำเสนอสามารถจำแนกความคิดเห็นของลูกค้าต่อสินค้าเป็นความคิดเห็นเชิงลบ หรือความคิดเห็นเชิงบวก ซึ่งทำให้ระบบพาณิชย์อิเล็กทรอนิกส์ทราบถึงความพึงพอใจของลูกค้าที่มีต่อสินค้า จากผลการทดสอบและเปรียบเทียบประสิทธิภาพแสดงให้เห็นว่ากระบวนการที่นำเสนอขึ้นสามารถจำแนกข้อความได้อย่างมีประสิทธิภาพด้วยค่าเฉลี่ยความถูกต้องร้อยละ 80.9 และสูงกว่าตัวจำแนกข้อมูลอื่น

อภิปรายผลการวิจัย

เนื่องจากกระบวนการจำแนกข้อความที่นำเสนอใช้การวัดความสำคัญของคำร่วมด้วย โดยความคิดเห็นหรือข้อมูลป้อนกลับจะถูกแทนด้วยกราฟข้อความ เพื่อแสดงความสัมพันธ์ในการปรากฏคำนั้น ๆ ร่วมกับคำอื่น ๆ จึงทำให้ประสิทธิภาพการจำแนกข้อความด้วยกระบวนการนี้จะขึ้นอยู่กับคุณลักษณะของกราฟข้อความด้วย เช่น ถ้ากราฟข้อความที่สร้างขึ้นมีความสัมพันธ์ของคำต่ำ หมายถึง มีค่าความหนาแน่นหรือจำนวนเส้นเชื่อม หรือค่าเฉลี่ยดีกรีต่ำ จะทำให้ค่า Probability Centrality มีค่าต่ำด้วย ซึ่งอาจส่งผลต่อประสิทธิภาพการจำแนกข้อความ

สำหรับงานวิจัยในอนาคต ผู้วิจัยมุ่งศึกษาเรื่องความหนาแน่นของกราฟข้อความ (Text Graph) ส่งผลหรือมีอิทธิพลต่อประสิทธิภาพการจำแนกข้อความอย่างไร และการหาค่าความหนาแน่นที่เหมาะสมต่อกระบวนการจำแนกข้อความที่นำเสนอ

References

[1] Wolfgang Himmel, Ulrich Reincke and Hans Wilhelm Michelmann. (2009). Text Mining and Natural Language Processing Approaches for Automatic Categorization of Lay Requests to Web-Based Expert Forums. **Journal of Medical Internet Research**. Vol. 11. No. 3. e25. pp. 1-6. DOI: 10.2196/jmir.1123

[2] Peng Zhou and Nora El-Gohary. (2016). Ontology-Based Multilabel Text Classification of Construction Regulatory Documents. **Journal of Computing in Civil Engineering**. Vol. 30. No. 4. pp. 1-13

[3] Vandana Korde. (2012). Text Classofication and Classifiers: A Survey. **International Journal of Artificial Intelligence & Applications**. Vol. 3. No. 2. pp. 85-99

[4] D. K. Kirange and Ratnadeep R. Deshmukh. (2014). Emotion Classification of Restaurant and Laptop Review Dataset: Semeval 2014 Task 4. **International Journal of Computer Applications**. Vol. 113. No. 6. pp. 17-20

[5] Aashutosh Bhatt, Ankit Patel, Harsh Chheda and Kiran Gawande. (2015). Amazon Review Classification and Sentiment Analysis. **International Journal of Computer Science and Information Technologies**. Vol. 6. No. 6. pp. 5107-5110

[6] Bondy, J. and Murty, U. (1976). **Graph Theory With Applications**. USA. : Elsecier Science Publishing Co., Inc.

[7] Kjetil Valle. (2011). **Graph-Based Representations for Textual Case-Based Reasoning**. Department of Computer and Information Science, Norwegian University of Science and Technology. pp. 1-10

[8] Asma Khazaal Abdulsahib and Siti Sakira Kamaruddin. (2015). Graph Based Text Representation for Document Clustering. **Journal of Theoretical and Applied Information Technology**. Vol. 76. No. 1. pp. 1-13

[9] J. Wu, Z. Xuan, and D. Pan. (2011). Enhancing Text Representation for Classification Tasks with Semantic Graph Structures. **International Journal if Innovative Computing, Information Control**. Vol. 7. No. 5. pp. 2689-2698

[10] Fragkiskos D. Malliaros and Konstantinos Skianis. (2015). Graph-Based Term Weighting for Text Categorization. **IEEE/ACM International Conference on Advances in Social Networks**. New York, NY, US. pp. 1473-1479

[11] Organisation for Economic Co-operation and Development. (2016). **Consumer Protection in E-commerce**. France : OECD Publishing

[12] Ricci, F., Rokach, L., Shapira, B., and Kantor, P.B. (2011). **Recommender Systems Handbook**. US. : Springer US Publisher

[13] Mohd Fauzi bin Othman and Thomas Moh Shan Yau. (2007). Comparison of Different Classification Techniques Using WEKA for Breast Cancer. In **Biomed 06, IFMBE Proceeding 15**. Germany : Springer Berlin Heidelberg. pp. 520-523

[14] Leo Breiman and R. (2001). Random Forests. **Machine Learning**. Vol. 45. Issue. 1. pp. 5-32. DOI: 10.1023/A:1010933404324

[15] W. Ali, S.M. Shamsuddin and A.S. Ismail. (2011). Web Proxy Cache Content Classification based on Support Vector Machine. **Journal of Artificial Intelligence**. Vol. 4. No. 1. pp. 100-109. DOI: 10.3923/jai.2011.100-109

[16] Hanneman, Robert A. and Mark Riddle. (2005). **Introduction to Social Network Methods**. Riverside, CA : University of California, Riverside

[17] M. M. Durland, K. A. Fredericks. (2006). **Social Network Analysis in Program Evaluation**. San Francisco : Jossey-Bass, Calif

[18] Xing Fang and Justin Zhan. (2015). Sentiment Analysis Using Product Review Data. **Journal of Big Data**. Vol. 2. No. 5. pp. 1-14

[19] Manvee Chauhan. (2015) Sentimental Analysis of Product Based Reviews Using Machine Learning Approaches. **Journal of Network Communications and Emerging Technologies (JNCET)**. Vol. 5. No. 2. pp. 19-25

[20] Vinodhini Gopalakrishnan, Chandrasekaran Ramaswamy. (2014). Sentiment Learning from Imbalanced Dataset: An Ensemble Based Method. **International Journal of Artificial Intelligence**. Vol. 12. No. 2. pp. 75-87

[21] Deepak Kumar Gupta and Asif Ekbal. (2014). IITP: Supervised Machine Learning for Aspect based Sentiment Analysis. In **Proceedings of the 8th International Workshop on Semantic Evaluation**. Dublin, Ireland. pp. 319-323

[22] Dimitrios Kotzias, Misha Denil, Nando De Freitas and Padhraic Smyth. (2015). From Group to Individual Labels using Deep Features. In **Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. Sydney, NSW, Australia. pp. 1-10

[23] Powers, David M W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. **Journal of Machine Learning Technologies**. Vol. 2. No. 1. pp. 37-63