การจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง

Sentiment Classification Using Machine Learning Techniques

นิเวศ จิระวิชิตชัย
บัณฑิตวิทยาลัย
มหาวิทยาลัยศรีปทุม วิทยาเขตชลบุรี
nivet99@hotmail.com

นรินทร์ พนาวาส
สาขาวิชาคอมพิวเตอร์ธุรกิจ คณะเทคโนโลยีสารสนเทศ
มหาวิทยาลัยศรีปทุม วิทยาเขตชลบุรี
narin.pa@east.spu.ac.th

บทคัดย่อ -- งานวิจัยนี้ได้เสนอวิธีการจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยทำการทดสอบประสิทธิภาพการจำแนกด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ เนอีฟเบย์ และเคเนียเรสเนเบอร์ จากการทดลองพบว่า เมื่อลดมิติของข้อมูลด้วยค่า Information Gain แล้วส่งเข้าประมวลผลด้วยเครื่องจักรการเรียนรู้ โดยวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) จากจำนวนมิติของข้อมูลที่ส่งผลดีที่สุดพบว่า เมื่อแทนค่าคุณลักษณะด้วยค่าความถี่ของคำ (Word Frequency) และเรียนรู้ด้วยอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ที่จำนวน 300 คุณลักษณะให้ประสิทธิภาพในการจำแนกดีที่สุด โดยให้ค่าความถูกต้อง (Accuracy) เท่ากับ 86.30 %

คำสำคัญ: *การจำแนกความคิดเห็น, การลดคุณลักษณะ, ซัพพอร์ทเวกเตอร์แมชชีน*

Abstract *-- This research presented Sentiment classification with Support Vector Machine, Decision Tree, Naive-Bayes and K-Nearest Neighbor algorithms. The experimental results showed that reducing the dimension by Information Gain technique and process with machine learning algorithms. The performance of accuracy found Support Vector Machine algorithms yielded a very high Sentiment classification with the accuracy equal to 86.30%.*

Keywords: *Sentiment Classification, Data Reduction, Support Vector Machine*

1. บทนำ

การขยายตัวด้านการใช้งานระบบคอมพิวเตอร์ และอินเทอร์เน็ตตลอดช่วงระยะเวลาที่ผ่านมา มีแนวโน้มในการใช้งานเพิ่มมากขึ้นอย่างรวดเร็ว ส่งผลให้เกิดการรับรู้และแลกเปลี่ยนข้อมูลข่าวสารตลอดจนแสดงความคิดเห็นผ่านสื่อในรูปแบบอิเล็กทรอนิกส์มากขึ้นและสามารถที่จะเข้าถึงข้อมูลได้อย่างสะดวกสบาย และแพร่หลายอย่างมาก จากการเข้าถึงข้อมูลข่าวสารผ่านสื่อในยุคปัจจุบัน ทำให้มีปริมาณและเนื้อหาที่หลากหลายมากขึ้น ส่งผลให้ผู้บริโภคได้รับความรู้จากข่าวสารปริมาณมากขึ้นกว่าเดิมมาก และจำนวนปริมาณข้อมูลข่าวสารที่มากขึ้นนั้น ส่งผลให้ผู้บริโภคต้องใช้เวลามากขึ้นในการคัดเลือกจำแนกข้อมูล ให้ตรงตามความสนใจของตน ปัจจุบันมีหลายเว็บไซต์ที่ผู้ขายสินค้าและบริการ ทำการขายสินค้าและบริการผ่านสื่ออินเทอร์เน็ตและปฏิสัมพันธ์กับลูกค้า โดยให้ลูกค้าเข้ามาติชมสินค้าหรือบริการผ่านทางเว็บไซต์ ทำให้เพิ่มความสะดวกสบายแก่ผู้ที่เข้ามาซื้อสินค้าและบริการโดยผู้ซื้อสามารถอ่านคำติชม ที่มีผู้ใช้บริการก่อนหน้าแสดงความคิดเห็น เพื่อประกอบการตัดสินใจการสั่งซื้อสินค้าและบริการของตนเอง แต่ด้วยบางรายการสินค้าที่มีผู้ใช้เข้ามาติชมจำนวนมาก ทำให้ผู้เข้ามาอ่านอาจไม่อยากอ่านข้อมูลติชมทั้งหมด ทำให้มีการคิดค้นพัฒนากระบวนการในการจำแนกความคิดเห็นแบบอัตโนมัติ โดยการจำแนกความคิดเห็นจากผู้ใช้สินค้าและบริการนั้น (Sentiment Classification) จัดเป็นงานจำแนกตามความรู้สึกและอารมณ์ของผู้ใช้ที่มีต่อสินค้าและบริการนั้นๆ การสรุปความคิดเห็นจากผู้ใช้ทั้งหมดว่าชื่นชอบในสินค้าและบริการดังกล่าวหรือไม่นั้น โดยสรุปเป็นคำตอบสั้นๆจะทำให้เกิดความสะดวกสบายและรวดเร็วต่อผู้เข้ามาอ่านคำติชม เพื่อใช้ประโยชน์จากข้อมูลติชมเพื่อพิจารณาการสั่งซื้อสินค้าให้มีประสิทธิภาพสูงสุด [1-3]

จากความสำคัญของปัญหาดังกล่าว ผู้วิจัยจึงมีแนวคิดที่จะทดสอบประสิทธิภาพการจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยนำเสนอแบบจำลองการจำแนกความคิดเห็นแบบอัตโนมัติ ทดสอบประสิทธิภาพแบบจำลองการจำแนกความคิดเห็นด้วยอัลกอริทึม ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ เนอีฟเบย์ และเคเนียเรสเนเบอร์ ในบทความนี้ประกอบด้วยส่วนต่าง ๆ ดังนี้ ส่วนที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้อง ส่วนที่ 3 กล่าวถึงวิธีการวิจัย ส่วนที่ 4 ผลการทดลอง และส่วนที่ 5 สรุปผลและข้อเสนอแนะ

2. ทฤษฎีที่เกี่ยวข้อง

2.1 การสกัดคุณลักษณะ (Feature Extraction)

วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะความคิดเห็น คือ การดึงคุณลักษณะ (Feature) ของความคิดเห็นออกมา ซึ่งการดึงคุณลักษณะออกมานั้น ต้องกำหนดก่อนว่าจะใช้อะไร เป็นตัวแทนคุณลักษณะของความคิดเห็น และใช้ค่าใดแทนคุณลักษณะความคิดเห็นนั้น จากการสำรวจงานวิจัยต่างประเทศพบว่า ส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของความคิดเห็น และใช้พื้นฐานค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้ วลี หรือกลุ่มของคำ ประโยค แทนคุณลักษณะของความคิดเห็นได้เช่นกัน ตัวแทนคุณลักษณะของเอกสารที่นิยมใช้ในการจัดหมวดหมู่เอกสารประเภทข้อความคือ ถุงคำ (Bag of words) ซึ่งจะเก็บอยู่ในรูปแบบของเวกเตอร์ โดยองค์ประกอบของเวกเตอร์อาจจะแทนด้วยคุณลักษณะของค่าความจริง (Boolean) หรือแทนด้วยค่าความถี่ของคำ (Word Frequency) ซึ่งงานวิจัยนี้ใช้การเลือกคุณลักษณะแบบคำเดียว (Single word) ผลลัพธ์ที่ได้จากการสกัดคุณลักษณะจะได้เป็นคำเดียวจำนวนมาก เพื่อมาใช้เป็นตัวแทนความคิดเห็นในการเรียนรู้ [1-2]

2.2 การเลือกคุณลักษณะ (Feature Selection)

การลดขนาดข้อมูล (Data Reduction) จัดเป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลง โดยสูญเสียลักษณะสำคัญของข้อมูลน้อยที่สุดและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุด เนื่องจากคุณลักษณะของความคิดเห็นแต่ละตัวจะมีความสำคัญต่อการจำแนกไม่เท่ากัน ดังนั้นด้วยเทคนิคการเลือกข้อมูลที่ดีจะทำให้สามารถเลือกข้อมูลที่มีความสำคัญและสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ และในความเป็นจริงมักจะเกิดเหตุการณ์ที่ เรียกกันว่าปัญหาของมิติข้อมูล (Curse of Dimensionality) ขึ้นเสมอ นั้นหมายความว่าจำเป็นต้องลดขนาดมิติของข้อมูลลง (Dimensionality Reduction) เพื่อให้ตัวจำแนกประเภทสามารถทำงานได้ถูกต้องมากขึ้น ซึ่งงานวิจัยนี้ใช้ค่าการเพิ่มของข้อมูล (Information Gain) [4] เป็นตัวลดคุณลักษณะของเอกสาร ซึ่งค่า IG จะคำนวณจากจำนวนบิตที่ได้รับสำหรับการทำนายกลุ่มโดยการดูจากการมีอยู่หรือไม่มีอยู่ของคำในเอกสาร ให้ $C_i, \ldots, C_K$ แทนเซ็ตที่เป็นไปได้ของกลุ่ม ค่า IG ของคำ w นิยามโดย

$$IG(w) = -\sum_{j=1}^{K} P(c_j)\log P(c_j) + P(w)\sum_{j=1}^{K} P(c_j \mid w)\log P(c_j \mid w) + P(\overline{w})\sum_{j=1}^{K} P(c_j \mid \overline{w})\log P(c_j \mid \overline{w}) \quad (1)$$

ค่า $P(C_j)$ คำนวณได้จากจำนวนเอกสารที่อยู่กลุ่ม C_j กับจำนวนเอกสารทั้งหมด

ค่า P(w) คำนวณได้จากจำนวนเอกสารที่มีคำ w กับจำนวนเอกสารทั้งหมด

ค่า $P(C_j| w)$ คำนวณได้จากจำนวนเอกสารกลุ่ม C_j ที่มีคำ w กับเอกสารทั้งหมด

ค่า $P(C_j| \overline{w})$ คำนวณได้จากจำนวนเอกสารกลุ่ม C_j ที่ไม่มีคำ w กับเอกสารทั้งหมด

2.3 อัลกอริทึมการจำแนกประเภท (Classifier Algorithm)

การจำแนกข้อมูลจัดเป็นการทำเหมืองข้อมูล ที่สำคัญเทคนิคหนึ่ง เป็นกระบวนการสร้างโมเดลจัดการข้อมูลให้อยู่ในกลุ่มที่กำหนดมาให้ โดยการสร้างกฎเพื่อช่วยในการตัดสินใจจากข้อมูลที่มีอยู่ เพื่อใช้ทำนายแนวโน้มการเกิดขึ้นของข้อมูลที่ยังไม่เกิดขึ้น การจำแนกข้อมูลประเภทเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถแบ่งขั้นตอนวิธีการจำแนกประเภทได้เป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มต้นแบบและ จำแนกประเภทของกลุ่มตัวอย่างที่สนใจ โดยการตรวจสอบหาความคล้ายกับกลุ่มตัวอย่างต้นแบบ

2.3.1 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) [5-6] แนวคิดหลักของวิธีการนี้ ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยจะพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มี ระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด SVM จะใช้ฟังก์ชั่นแม็ปสำหรับย้ายข้อมูลจาก Input Space

ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) บน Feature Space เหมาะใช้สำหรับข้อมูลที่มีมิติของข้อมูลสูง กำหนดให้ $(x_i,y_i),\ldots,(x_n,y_n)$ เป็นตัวอย่างที่ใช้สำหรับการสอน n คือ จำนวนข้อมูลตัวอย่าง m คือ จำนวนมิติข้อมูลเข้า และ y คือ ผลลัพธ์มีค่า + 1 หรือ -1 ดังสมการ

$$(x_i,y_i),\ldots,(x_n,y_n) \text{ เมื่อ } x \in R^m, y \in \{+1,-1\} \quad (2)$$

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจ ซึ่งคำนวณได้ดังสมการ

$$(w * x)+b = 0 \quad (3)$$

เมื่อ w คือ ค่าน้ำหนักและ b คือค่า bias สมการ ใช้สำหรับจำแนกประเภทของข้อมูล

$$(w*x)+b>0 \text{ ถ้า } y_i = +1 \text{ และ } (w*x)+b<0 \text{ ถ้า } y_i = -1 \quad (4)$$

โดย SVM มีเคอร์เนลฟังก์ชั่น (Kernel Function) ให้ผู้ใช้สามารถประยุกต์ใช้ในการแก้ปัญหาได้หลายวิธี เช่น Linear, Polynomial, Radial Basis Function โดยงานวิจัยได้ใช้ Polynomial Kernel ในการทดลอง

2.3.2 ต้นไม้ตัดสินใจ (Decision Tree) [6-7] ต้นไม้จะประกอบด้วยโหนดแทนคุณลักษณะ และโหนดล่างสุดแทนหมวดหมู่ การสร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าความจริงที่ใช้จะมาจากการคำนวณหาค่า Information Gain คุณลักษณะใดที่มีค่า Information Gain มากที่สุดจะถูกเลือกเป็นโหนดลูก การคำนวณค่า Information Gain การสร้างต้นไม้ตัดสินใจ C4.5 จะใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ $\{m_1, m_2, \ldots, m_n\}$ และให้ความน่าจะเป็นที่จะเกิดค่า m_1 มีค่าเท่ากับ $P(m_1)$ จะได้ว่าค่าเกนสารสนเทศ (Information Gain) ของ M เขียนแทนด้วย I(M) คำนวณได้ดังสมการ

$$I(M) = \sum_{i=1}^{n} -P(m_i)\log_2 P(m_i) \quad (5)$$

ถ้าให้ข้อมูลสอน คือ T และคุณลักษณะที่เป็นโหนด คือ x และมีค่าทั้งหมดที่เป็นไปได้ n ค่า โหนดปัจจุบันจะแบ่งตัวอย่าง T ออกตามกิ่งเป็น $\{t_1, t_2, \ldots, t_n\}$ ตามค่าที่เป็นไปได้ของ x ดังนั้นจึงสามารถคำนวณค่าเกนสารสนเทศหลังจากแบ่งตามคุณลักษณะ x และค่ามาตรฐานเกน (GAIN) ของคุณลักษณะ x ได้ดังสมการ

$$I_x(T) = \sum_{i=1}^{n} \frac{|t_i|}{|T|} I(t_i) \quad (6)$$

$$Gain(x) = I(T) = I_x(T) \quad (7)$$

จากนั้นคำนวณค่าสารสนเทศของการแบ่งแยก (Split Information) ของคุณลักษณะแต่ละตัว ถ้าให้ T คือ ชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณลักษณะ x จะได้ชุดของตัวอย่างย่อยในแต่ละกิ่ง คือ $\{t_1, t_2, \ldots, t_n\}$ จำนวน n ชุด ตามค่าที่เป็นไปได้ในคุณสมบัติ x เมื่อคำนวณค่าสารสนเทศของการแบ่งแยกได้ดังสมการ

$$\text{Split Information} = \sum_{i=1}^{n} \frac{|t_i|}{|T|} \log_2 \frac{|t_i|}{|T|} \quad (8)$$

คำนวนค่ามาตรฐานอัตราส่วนเกน (Gain ratio) ได้จาก Gain Ratio = Gain – Split Information ท้ายสุดจึงเลือกค่า Gain ratio สูงสุดเป็นคุณลักษณะเริ่มต้น และเลือกคุณสมบัติถัดไปตามค่า Gain ratio น้อยลงตามลำดับ

2.3.3 เนอีฟเบย์ (Naïve-Bayes) [8] หลักการของวิธีการนี้ใช้การคำนวณความน่าจะเป็นซึ่งถูกใช้ใน การทำนายผล เป็นเทคนิคในการแก้ปัญหาแบบ การจำแนกประเภทที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย มันจะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ การเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง อัลกอริทึมในการทำงานที่ไม่ซับซ้อน เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลเท่ากับ

$$P(a_1, a_2 \ldots, a_3 \mid v_j) = \prod_{i=1}^{n} P(a_1 \mid v_j) \quad (9)$$

กลุ่ม v_j สำหรับข้อมูลที่มีคุณสมบัติ n ตัว $X = \{ a_1, a_2, \ldots, a_n \}$ หรือใช้สัญลักษณ์ว่า $P(a_1, a_2, \ldots, a_n \mid v_j)$ โดยที่ $\prod$ หมายถึง ผลคูณของค่า $(a_i \mid v_j)$ ทั้งหมด $i = 1,2,3,\ldots,n$ และ $j = 1,2,3,\ldots,n$ ดังนั้นเราจะได้ว่าวิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการ

$$V_{NB} = \arg\max_{v_j \in V} P(v_j) \times \prod_{i=1}^{n} P(a_1 \mid v_j) \quad (10)$$

2.3.4 เคเนียเรสเนเบอร์ (K-Nearest Neighbor) [9] หลักการของวิธีการนี้ จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด K ตัวจากข้อมูลบนชุดข้อมูลตัวอย่าง เป็นอัลกอริทึ่มที่เรียบง่าย การทำงานโดยขึ้นกับระยะทางน้อยสุดจากสมาชิกใหม่

หรือข้อมูลที่ป้อนถาม (Input query instance) กับข้อมูลตัวอย่างฝึกฝน จะคำนวณหาเพื่อนบ้านที่ใกล้ที่สุด K ตัว หลังจากนั้นเราจะรวบรวมสมาชิกที่ใกล้เคียงที่สุด K ตัวแล้วเลือกคลาสที่สมาชิกส่วนใหญ่ที่ในกลุ่ม K ดังกล่าวสังกัดอยู่มากที่สุดให้กับสมาชิกใหม่ ข้อมูลการจำแนกโดยใช้ข้อมูลข้างเคียง K ตัว ประกอบด้วยแอททริบิวต์หลายตัวแปร X_i ซึ่งจะนำมาใช้ในการแบ่งกลุ่ม Y_i โดยระบุค่าตัวเลขจำนวนเต็มบวกให้กับ K ซึ่งค่านี้จะเป็นตัวบอกจำนวนของกรณี (Case) ที่จะต้องค้นหาในการทำนายกรณีใหม่ โดยงานวิจัยนี้กำหนด 1-KNN หมายถึงจะค้นหา 1 กรณีที่มีลักษณะใกล้เคียงกับกรณีใหม่ (1 Nearest Cases) การนำระยะทางที่หาได้จากสมาชิกใน ข้อมูลตัวอย่างฝึกฝน มาเรียงลำดับจากน้อยไปหามากแล้วเลือกสมาชิกที่มีระยะทาง (Distance) ใกล้เคียงที่สุดออกมา K ตัวโดยใช้การวัดระยะทางแบบ Euclidean Distance มีหลักการคือ การวัดระยะทางระหว่างสองวัตถุ ถ้าวัตถุห่างกันมากแสดงว่าวัตถุนั้นมีความคล้ายกันน้อย ถ้ามีค่าน้อยก็แสดงว่ามีความคล้ายคลึงกันมาก โดยที่ ค่า P_i แทนคุณสมบัติจากฐานข้อมูล Q_i แทนคุณสมบัติที่ผู้ใช้ระบุ ดังแสดงในสมการ

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}. \quad (11)$$

2.4 การวัดประสิทธิภาพ

วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพแต่ละอัลกอริทึมในการจำแนกประเภทของความคิดเห็นนั้น สามารถวัดที่ประสิทธิภาพของการจำแนกข้อมูล ตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ซึ่งคำนวณได้ดังตาราง Confusion matrix [10] ดังสูตร

$$\text{Accuracy} = (TP + TN)/(TP + FP + FN + TN) \quad (12)$$

TP = ตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FP = ตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าอยู่กลุ่ม C_j

FN = ตัวอย่างที่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

TN = ตัวอย่างที่ไม่อยู่กลุ่ม C_j และตัวจำแนกทำนายว่าไม่อยู่กลุ่ม C_j

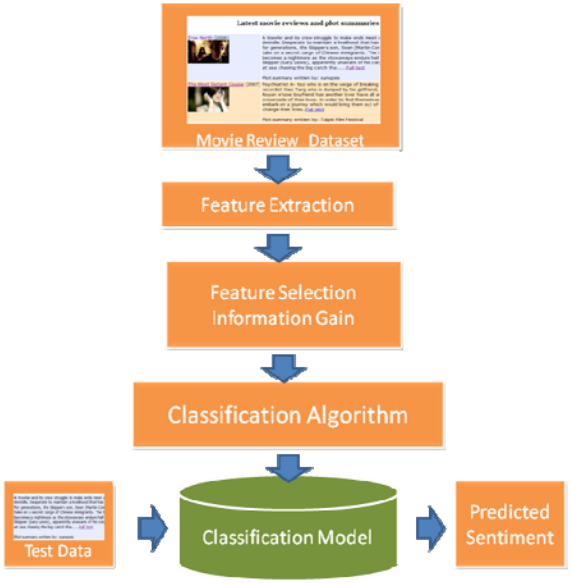
C_j = กลุ่มประเภทของความคิดเห็น

		actual value		
		p	*n*	total
prediction outcome	*p'*	True Positive	False Positive	P'
	n'	False Negative	True Negative	N'
	total	P	N	

ภาพที่ 1 Confusion matrix

3. วิธีการดำเนินการวิจัย

งานวิจัยนี้ทำการทดลองการลดมิติของข้อมูลร่วมกับอัลกอริทึมการจำแนกประเภทเรียนรู้แบบมีผลเฉลย (Supervised Learning) โดยนำเสนอแบบจำลองการจำแนกความคิดเห็นแบบอัตโนมัติ ทดสอบกับกลุ่มตัวอย่าง ความคิดเห็นผู้ชมต่อภาพยนตร์ต่างประเทศ จากฐานข้อมูล Internet Movie Database (IMDB) [3] ซึ่งข้อมูลชุดนี้ประกอบด้วย 2 ความคิดเห็น คือ ชอบภาพยนตร์ และ ไม่ชอบภาพยนตร์ ซึ่งมีจำนวนมิติของข้อมูล 1166 คุณลักษณะ จำนวน 2000 ตัวอย่าง เป็นกลุ่มตัวอย่างเรียนรู้ และทำการทดสอบด้วยวิธี 10-fold cross validation



ภาพที่ 2 ขั้นตอนการจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง

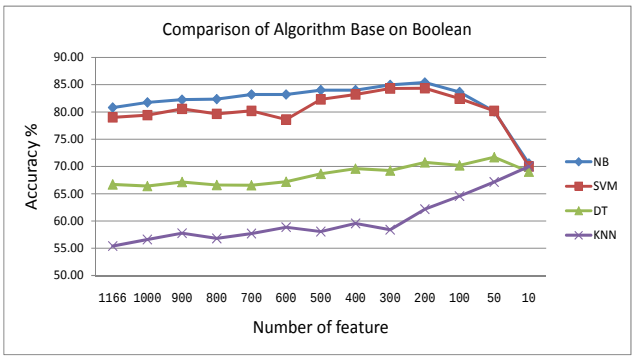
โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ จะทำการลดขนาดมิติของข้อมูลด้วยค่า Information Gain โดยพิจารณาเรียงลำดับจากค่าสูงสุด ซึ่งผลที่ได้จากการลดขนาดของมิติดังกล่าว จะถูกส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลยด้วย อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) มาเรียนรู้ เพื่อทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง (Accuracy) ในการจำแนกความคิดเห็น

4.ผลการทดลอง

การทดลองการจำแนกความคิดเห็นของผู้ชมต่อภาพยนตร์ต่างประเทศ จากฐานข้อมูล Internet Movie Database (IMDB) โดยทำการทดลองกับข้อมูล 2 ชุด คือ ชุดที่แทนค่าน้ำหนักด้วยค่าความจริง (Boolean) กล่าวคือ 0 คือไม่พบคำในความคิดเห็นนั้น ส่วน 1 คือพบคำในความคิดเห็นนั้น และชุดที่แทนค่าน้ำหนักด้วยค่าความถี่ของคำ (Word Frequency) ที่แสดงความคิดเห็น ซึ่งได้ผลการทดลองดังนี้

4.1 ชุดที่แทนค่าน้ำหนักด้วยค่าความจริง (Boolean)

ผลทดลองโดยแทนค่าคุณลักษณะด้วย ค่าความจริง (Boolean)และใช้วิธี Information Gain ลดคุณลักษณะและทำการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ 4 อัลกอริทึม พบว่าอัลกอริทึมเนอีฟเบย์ (Naïve-Bayes)ให้ประสิทธิภาพในการจำแนกดีที่สุด รองลงมาเป็นซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ต้นไม้ตัดสินใจ (Decision Tree) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ตามลำดับ ดังภาพที่ 3

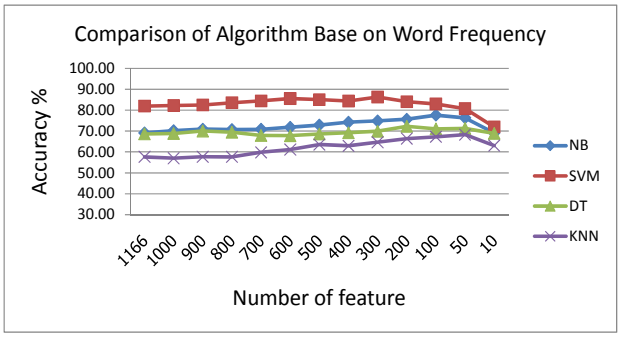


ภาพที่ 3 การเปรียบเทียบประสิทธิภาพของอัลกอริทึม เมื่อแทนค่าคุณลักษณะด้วยค่าความจริง

และเมื่อพิจารณาจุดที่ให้ประสิทธิภาพในการจำแนกดีที่สุด พบว่าอัลกอริทึมเนอีฟเบย์ เมื่อลดมิติข้อมูลลงเหลือ 200 คุณลักษณะให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 85.40% โดยสามารถลดมิติของข้อมูลได้ถึง 82.84% จากคุณลักษณะที่สกัดได้ทั้งหมด

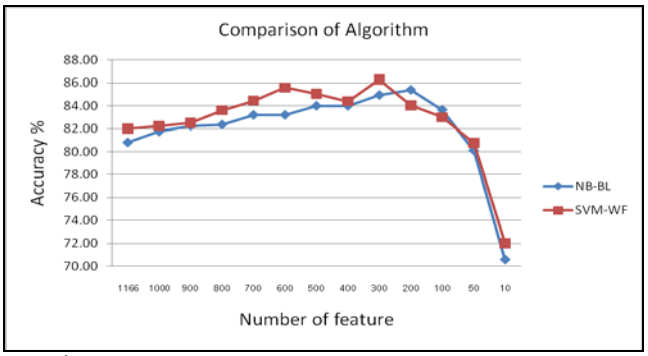
4.2 ชุดที่ค่าน้ำหนักด้วยค่าความถี่ของคำ (Word Frequency)

ผลทดลองโดยแทนค่าคุณลักษณะด้วยค่าความถี่ของคำ (Word Frequency) และใช้วิธี Information Gain ลดคุณลักษณะและทำการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ 4 อัลกอริทึม พบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ประสิทธิภาพในการจำแนกดีที่สุด รองลงมาเป็นเนอีฟเบย์ (Naïve-Bayes) ต้นไม้ตัดสินใจ (Decision Tree) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) ตามลำดับ และเมื่อพิจารณาจุดที่ให้ประสิทธิภาพในการจำแนกดีที่สุดพบว่า อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน เมื่อลดมิติข้อมูลลงเหลือ 300 คุณลักษณะให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 86.30 % โดยสามารถลดมิติของข้อมูลได้ถึง 74.27% จากคุณลักษณะที่สกัดได้ทั้งหมด ดังภาพที่ 4



ภาพที่ 4 การเปรียบเทียบประสิทธิภาพของอัลกอริทึม เมื่อแทนค่าคุณลักษณะด้วยค่าความถี่ของคำ

เมื่อทำการเปรียบเทียบประสิทธิภาพด้วยอัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) ที่แทนค่าด้วยค่าความจริง (Boolean) กับซัพพอร์ทเวกเตอร์แมชชีนที่แทนค่าด้วยค่าความถี่ของคำ (Word Frequency) พบว่า ซัพพอร์ทเวกเตอร์แมชชีน ให้ประสิทธิภาพในการจำแนกสูงกว่าอย่างเห็นได้ชัดเจน และมีประสิทธิภาพในการจำแนกเริ่มต่ำกว่าเนอีฟเบย์ เมื่อจำนวนคุณลักษณะลดลงที่ระดับ 200 คุณลักษณะ ดังภาพที่ 5



ภาพที่ 5 การเปรียบเทียบประสิทธิภาพของอัลกอริทึมอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนและเนอีฟเบย์

5.สรุปผลการวิจัย

งานวิจัยนี้ได้เสนอการจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยนำเสนอแบบจำลองการจำแนกความคิดเห็นแบบอัตโนมัติ ทดสอบประสิทธิภาพแบบจำลองด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ เนอีฟเบย์ และเคเนียเรสเนเบอร์ โดยวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) ของแบบจำลองพบว่า เมื่อแทนค่าคุณลักษณะด้วย ค่าความจริง (Boolean)และ ลดคุณลักษณะด้วย Information Gain และเรียนรู้ด้วยเครื่องจักรการเรียนรู้ 4 อัลกอริทึม พบว่าอัลกอริทึมเนอีฟเบย์ (Naïve-Bayes)ให้ประสิทธิภาพในการจำแนกดีที่สุด เมื่อลดมิติข้อมูลลงเหลือ 200 คุณลักษณะให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 85.40 % โดยสามารถลดมิติของข้อมูลได้ถึง 82.84% จากคุณลักษณะที่สกัดได้ทั้งหมด เมื่อแทนค่าด้วยค่าความถี่ของคำ (Word Frequency) และใช้วิธี Information Gain ลดคุณลักษณะและทำการเรียนรู้ด้วยเครื่องจักรการเรียนรู้ 4 อัลกอริทึม พบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) ให้ประสิทธิภาพในการจำแนกดีที่สุด เมื่อลดมิติข้อมูลลงเหลือ 300 คุณลักษณะให้ค่าความถูกต้อง (Accuracy) สูงที่สุดเท่ากับ 86.30 % โดยสามารถลดมิติของข้อมูลได้ถึง 74.27% จากคุณลักษณะที่สกัดได้ทั้งหมด

ผลการทดลองจากงานวิจัยนี้ พบว่าอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) และเนอีฟเบย์ (Naïve-Bayes) มีประสิทธิภาพในการจำแนกข้อมูลความคิดเห็นของผู้ชม ที่เขียนจากความรู้สึกและอารมณ์ของผู้ใช้ที่มีต่อการใช้สินค้าและบริการนั้นๆ ซึ่งในที่นี้คือความคิดเห็นที่มีต่อภาพยนตร์ได้เป็นอย่างดี และข้อมูลประเภทนี้สามารถลดมิติของข้อมูลด้วยวิธี Information Gain ลงได้อย่างมาก โดยไม่กระทบต่อประสิทธิภาพในการจำแนกข้อมูลแต่อย่างใดและ ยิ่งเมื่อลดมิติของข้อมูลลง ยิ่งส่งผลให้ค่าความถูกต้องในการจำแนกเพิ่มสูงขึ้นอย่างมีนัยสำคัญ ซึ่งการลดมิติของข้อมูลนี้สามารถลดทรัพยากรของระบบและลดระยะเวลาในการประมวลผลได้เป็นอย่างมาก ผลจากงานวิจัยนี้สามารถนำแบบจำลองที่นำเสนอไปประยุกต์ใช้กับการจำแนกข้อมูลกับประเภทอื่นๆเช่น การสร้างระบบจัดหมวดหมู่เอกสารอัตโนมัติ (Text Categorization) การคัดกรองเอกสาร (Document Filtering) การจัดทำดัชนีอัตโนมัติเพื่อใช้ในการค้นคืนเอกสาร (Automatic Indexing for IR System) การจัดหมวดหมู่ของเว็บเพจ (Web Page Classification) เป็นต้น

6.เอกสารอ้างอิง

[1] B. Pang, and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, 2008, pp. 1–135.

[2] Bing Liu, “Sentiment analysis: a multifaceted problem”, IEEE Intelligent System 25 ,2010, pp. 76–80.

[3] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques”, Proceedings of EMNLP 2002.

[4] นิเวศ จิระวิชิตชัย ปริญญา สงวนสัตย์ และพยุง มีสัจ. “การศึกษาทดลองเทคนิคการลดคุณลักษณะและอัลกอริทึมการจัดหมวดหมู่ของเอกสารภาษาไทย ”, วารสารวิทยาศาสตร์ลาดกระบัง ปี 2553.

[5] Joachims, “Text categorization with support vector machines: Learning with many relevant features”. Proc. of the 10 th European Conf. on Machine Learning, pp.137–142, 1998.

[6] พรพล ธรรมรงค์รัตน์ ลัดดา ปรีชาวีรกุล และวิภาดา เวทย์ประสิทธิ์. “การจำแนกประเภทเว็บเพจโดยใช้ค่าความถี่เอกสารและซัพพอร์ตเวกเตอร์แมชชีน”, The 12th National Computer Science and Engineering Conference, 2008.

[7] Quinlan, J. R., C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[8] D. Lewis, “Naive bayes at forty: The independence assumption in information retrieval”. Proc. of European Conf. on Machine Learning, 1998, pp. 4–15.

[9] D. Aha, D. Kibler, Instance-based learning algorithms. Machine Learning. Volume 6, Number 1, pp. 37-66.

[10] Max Bramer, Principles of Data Mining , Springer-Verlag. March 2007.