
**รายงานการวิจัย**

**เรื่อง**

# การพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง

# DEVELOPING AND EFFECTIVE EMOTION CLASSIFICATION OF THAI TEXT USING ADJUST TERM WEIGHING TECHNIQUE AND MACHINE LEARNING ALGORITHMS

**นิเวศ จิระวิชิตชัย**

**งานวิจัยนี้ ได้รับทุนอุดหนุนการวิจัยจากมหาวิทยาลัยศรีปทุม**

**ปีการศึกษา 2557**

หัวข้อวิจัย : การพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง
ผู้วิจัย : ดร.นิเวศ จิระวิชิตชัย
หน่วยงาน : คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม
ปีที่พิมพ์ : พ.ศ. 2558

---

## บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการให้ค่าดัชนีแบบใหม่เพื่อสะท้อนอำนาจจำแนกของคำ โดยนำเสนอวิธีการคำนวณค่าดัชนีชื่อ Term Occurrence Ratio Weighting (TOW) เพื่อพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง ผลจากการทดลองเมื่อวัดประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) เมื่อทำการลดคุณลักษณะด้วยวิธีค่าสถิติไคสแควร์ (Chi – Square) และประมวลผลด้วยเครื่องจักรเรียนรู้ สามารถสรุปได้ว่า วิธีให้ค่าน้ำหนัก TOW ร่วมกับเรียนรู้ด้วยอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ประสิทธิภาพด้านการจำแนกที่ดีสุดที่ระดับ 82.60% ผลการทดลองพบว่า วิธีการให้ค่าน้ำหนักกับดัชนีด้วย TOW ที่พัฒนาขึ้นนั้น ส่งผลให้ประสิทธิภาพของอัลกอริทึมสามารถจำแนกข้อมูลได้ถูกต้องกว่าการคำนวณดัชนีแบบอื่นอย่างชัดเจน

คำสำคัญ: **การจำแนกอารมณ์ ดัชนีของคำ การเรียนรู้ของเครื่อง**

Research Title : Developing and Effective Emotion Classification of Thai Text Using Adjust Term Weighing Technique and Machine Learning Algorithms
Name of Researcher : Dr. Nivet Chirawichitchai
Name of Institution : Faculty of Information Technology, Sripatum University
Year of Publication : B.E. 2558

---

## ABSTRACT

The objective of this research is to develop a method for term indexing that reflects the actual discrimination of isolated words. The method is called Term Occurrence Ratio Weighting (TOW), this will improve efficiency of emotion classification framework. The experimental results showed that reducing the feature by Chi – Square method and process Term Occurrence Ratio Weighting (TOW) with Decision Tree will yielded a very high classification with accuracy equaling 82.60%. The TOW indexing with classification algorithms yielded the best performance with the accuracy over all term weighting.

**Keywords:** emotion classification, indexing, machine learning

# กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยศรีปทุม ที่ให้ทุนอุดหนุนสำหรับการทำโครงการวิจัยนี้ และขอขอบคุณ ผศ.ดร.ปริญญา สงวนสัตย์ ที่ปรึกษาโครงการวิจัย ที่เสียสละเวลาให้คำปรึกษาและคำแนะนำในการทำโครงการวิจัยในครั้งนี้เป็นอย่างดี

นิเวศ จิระวิชิตชัย

30 พฤษภาคม 2557

# สารบัญ

**บทที่** **หน้า**

# สารบัญ (ต่อ)

**บทที่** **หน้า**

# สารบัญภาพประกอบ

**ภาพประกอบ** **หน้า**

# สารบัญตาราง

**ตาราง** **หน้า**

# บทที่ 1

# บทนำ

## ความเป็นมาและความสำคัญของปัญหา

การขยายตัวด้านการใช้งานระบบเครือข่ายคอมพิวเตอร์ อินเทอร์เน็ตและ พาณิชย์อีเล็กทรอนิกส์ ตลอดช่วงระยะเวลาที่ผ่านมา มีแนวโน้มในการใช้งานเพิ่มมากขึ้นอย่างรวดเร็ว ส่งผลให้ผู้ใช้งานสามารถที่จะเข้าถึงข้อมูลได้อย่างสะดวกสบาย และแพร่หลายอย่างมาก ทำให้เกิดการรับรู้และแลกเปลี่ยนข้อมูลข่าวสาร ตลอดจนแสดงความคิดเห็นผ่านสื่อในรูปแบบอิเล็กทรอนิกส์มีมากขึ้นและ จากสถิติการถึงอินเตอร์เน็ตความเร็วสูงในประเทศไทยในยุคปัจจุบันมีอัตราการเข้าถึงสูงมากเมื่อเทียบกับในอดีต ทำให้ประชากรได้รับข้อมูลข่าวสารที่มีปริมาณมากและ มีเนื้อหาที่ท่วมท้นและหลากหลายมากขึ้น ส่งผลให้ผู้บริโภคได้รับความรู้จากข่าวสารปริมาณมากขึ้นมากกว่าเดิม และจำนวนปริมาณข้อมูลข่าวสารที่เพิ่มมากขึ้นนั้น ส่วนใหญ่ล้วนมากจากการปฏิสัมพันธ์ระหว่างผู้ใช้งานกับเว็บเครือข่ายสังคมออนไลน์ (Social Media) ที่เปิดให้บริการในรูปแบบต่างๆ มากมาย ในยุคที่เราเรียกว่า เว็บ 2.0

ยุคเว็บ 2.0 ที่เครือข่ายสังคมได้เข้ามามีบทบาท การเกิดขึ้นของเนื้อหาสาระใหม่ๆ โดยเฉพาะส่วนที่เป็นข้อความแสดงความคิดเห็น กำลังกลายเป็นข้อมูลสำคัญที่สะท้อนถึงความคิดความเป็นไป และการเปลี่ยนแปลงที่เกิดขึ้นในสังคมและสามารถนำมาใช้วิเคราะห์สำหรับการพัฒนา ปรับปรุงสินค้าและบริการต่างๆ เพื่อตอบสนองความต้องการของผู้บริโภค หากแต่การเข้าถึงข้อมูลเหล่านี้ในปัจจุบัน กลับยังไม่สามารถค้นหาได้ด้วยเทคโนโลยีการสืบค้นแบบเดิม โดยปกติข้อมูลบนเว็บแบ่งออกเป็น 2 ประเภท ได้แก่ ข้อมูลทั่วไป (Facts) เช่น ข้อมูลเกี่ยวกับองค์กรและบริษัท ข้อมูลเกี่ยวกับสินค้าและบริการ หรือรายงานข่าว และข้อความแสดงความคิดเห็น (Opinions) เช่น กระทู้ในเว็บบอร์ด(Webboard) บล็อก (Blog) หรือเว็บข้อความวิจารณ์ทั่วไป (Reviews and Comments) ซึ่งข้อมูลทั่วไปจะสามารถสืบค้นได้ด้วยเทคโนโลยีการสืบค้น (Search Engine) ที่มีการพัฒนาและใช้อยู่ในปัจจุบัน โดยผู้ใช้สามารถระบุการค้นหาด้วยคำสำคัญ (Keywords) ที่ตรงกับหัวข้อ และระบบจะสืบค้นและนำเสนอผลลัพธ์ที่ตรงตามความต้องการในการสืบค้นครั้งนั้นๆ แต่ลักษณะทางเทคนิคที่ระบุการสืบค้นด้วยคำสำคัญ กลับยังไม่เหมาะสำหรับการสืบค้นข้อความแสดงความคิดเห็นได้ เนื่องจากในการสืบค้นหาความคิดเห็นต่อสินค้าและบริการหรือต่อเรื่องใดเรื่องหนึ่ง ข้อมูลความคิดเห็นจะมาจากคนจำนวนมากและจากหลากหลาย

แหล่ง ทำให้ต้องเสียเวลาอ่านและสรุปข้อความจำนวนมหาศาลนั้น ดังนั้นการนำระบบเหมืองข้อมูลมาใช้ จะช่วยรวบรวม สกัดและวิเคราะห์ความคิดเห็นของสินค้าหรือบริการในมุมมองต่างๆ เช่น ราคา การออกแบบ ฟังก์ชันการใช้งาน เพื่อทำการสรุปความคิดเห็นที่หลากหลายให้กับผู้ใช้เข้าใจได้ง่ายและรวดเร็ว และเมื่อนำมาผสานกับเทคโนโลยีการสืบค้น จะทำให้ผู้ใช้สามารถค้นหาข้อมูลแสดงความคิดเห็นที่ตรงตามความต้องการ

แต่ปัจจุบันมีข้อมูลข่าวสารมากมาย ส่งผลให้ผู้บริโภคต้องใช้เวลามากขึ้นในการคัดเลือกจำแนกข้อมูลให้ตรงตามความสนใจของตน ตัวอย่างเช่น ปัจจุบันมีหลายเว็บไซต์ที่ผู้ขายสินค้าและบริการ ทำการขายสินค้าและบริการผ่านสื่ออินเทอร์เน็ตและมีการใช้ เครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Webboard) หรือ บล็อก (Blog) ในการปฏิสัมพันธ์กับลูกค้าด้านการตลาด โดยให้ลูกค้าเข้ามาติชมสินค้าหรือบริการผ่านทางเว็บไซต์ ทำให้เพิ่มความสะดวกสบายแก่ผู้ที่เข้ามาซื้อสินค้าและบริการโดยผู้ซื้อสามารถอ่านคำติชม ที่มีผู้ใช้บริการก่อนหน้าแสดงความคิดเห็น เพื่อประกอบการตัดสินใจการสั่งซื้อสินค้าและบริการของตนเอง แต่ด้วยบางรายการสินค้าที่มีผู้ใช้เข้ามาติชมจำนวนมาก ทำให้ผู้เข้ามาอ่านอาจไม่อยากอ่านข้อมูลข่าวสารทั้งหมด ทำให้มีการคิดค้นพัฒนากระบวนการในการสรุปความคิดเห็นแบบอัตโนมัติ ซึ่งได้ข้อมูลมาจากการจำแนกความคิดเห็นจากผู้ใช้สินค้าและบริการจำนวนมากที่ให้คำติชมเข้ามา ผ่านระบบเครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Webboard) หรือ บล็อก (Blog) (Pang & Lee, 2002; Pang & Lee, 2008)

จึงสรุปได้ว่าเทคโนโลยีเหมืองข้อความแสดงความคิดเห็น (Opinion Mining) จึงเป็นเครื่องมือทางเทคโนโลยีรูปแบบใหม่ที่เข้ามาช่วยหน่วยงานและองค์กรต่างๆ ในการค้นหาข้อมูลความคิดเห็นเพื่อการบริหารจัดการลูกค้าสัมพันธ์ (CRM: Customer Relationship Management) โดยองค์กรหรือหน่วยงานธุรกิจสามารถดึงข้อมูลหลากหลายบนเครือข่ายสังคมมาใช้เพื่อประเมินความพึงพอใจของลูกค้า (Customer Satisfaction) ต่อสินค้าและการให้บริการ ทั้งนี้ ข้อความที่ถ่ายทอดถึงอารมณ์และความรู้สึกของลูกค้าเหล่านี้ ถือเป็นข้อมูลเชิงจิตวิทยา (Psychological Data) ที่สามารถนำมาใช้วิเคราะห์พฤติกรรมผู้บริโภคและพฤติกรรมทางสังคม เพื่อปรับปรุง พัฒนา สินค้าและบริการของบริษัทให้มีประสิทธิภาพมากขึ้น ตรงกับความต้องการของผู้บริโภคมากขึ้น (Pang & Lee, 2002)

การจำแนกเหมืองความคิดเห็น (Opinion Mining) จัดเป็นกระบวนการศึกษา และวิเคราะห์พฤติกรรมผู้บริโภค ถึงความรู้สึกของผู้ซื้อที่มีต่อสินค้าและบริการนั้นๆ โดยระบบจะทำการสรุปความคิดเห็นเป็นสัดส่วนจากผู้ใช้บริการทั้งหมดว่า มีความคิดเห็น หรือมมุมองต่อสินค้าและบริการนั้นๆในทิศทางของเชิงบวก หรือ เชิงลบ หรือ เฉยๆ อย่างละเท่าไร ซึ่งข้อมูลสรุปดังกล่าว จะทำให้

เกิดประโยชน์ต่อผู้ซื้อสินค้าและบริการภายหลัง ที่สามารถเข้ามาอ่านสรุปผลความคิดเห็นว่าคนส่วนใหญ่มีความเห็นไปในทิศทางใด แล้วใช้ประโยชน์จากข้อมูลความคิดเห็น (Opinion Mining) ดังกล่าว เพื่อพิจารณาประกอบการตัดสินใจสั่งซื้อสินค้าและบริการนั้นต่อไป ส่วนในมุมของผู้ขายสินค้า(หน่วยงานหรือองค์กร) ก็สามารถรับทราบถึงมุมมองของผู้บริโภคส่วนใหญ่ ว่ามีความรู้สึกอย่างไรต่อสินค้าและบริการของเขาอย่างไร (Pang & Lee, 2008)

แต่เนื่องจากในปัจจุบันสภาวะการแข่งขันด้านการตลาดสูงขึ้น การวิเคราะห์ข้อมูลความคิดเห็นในแง่มุมของเชิงบวก หรือ เชิงลบ หรือเฉยๆ ต่อสินค้าหรือผลิตภัณฑ์ต่างๆ ของบริษัท เริ่มไม่เพียงพอต่อการตัดสินใจการกำหนดกลยุทธ์ด้านการตลาดของบริษัท ที่เรียกว่าการบริหารจัดการลูกค้าสัมพันธ์ (CRM: Customer Relationship Management) เนื่องจากการวิเคราะห์ขั้นพื้นฐานดังกล่าว ไม่สามารถลงลึกถึงระดับอารมณ์ของลูกค้าที่มีต่อ การใช้สินค้าและบริการของบริษัทได้ ดังนั้นข้อมูลเชิงจิตวิทยา (Psychological Data) ของลูกค้าเหล่านี้จึงมีความสำคัญ ที่จะสามารถนำมาใช้วิเคราะห์พฤติกรรมผู้บริโภคและ พฤติกรรมทางสังคม เพื่อนำข้อมูลมาปรับปรุงสินค้าและบริการให้ตรงตามความต้องการของลูกค้าได้อย่างมีประสิทธิภาพมากขึ้น

ด้วยการเล็งเห็นถึงความสำคัญของปัญหาดังกล่าว ผู้วิจัยจึงได้พัฒนาแบบจำลองเหมืองข้อความแสดงความคิดเห็น (Opinion Mining) โดยมุ่งเน้นที่จะพัฒนาประสิทธิภาพของแบบจำลองการจำแนกอารมณ์ (Emotion Classification) ในข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง ซึ่งเป็นการคิดค้นวิธีการให้ค่าน้ำหนักของคำแบบใหม่ ที่วิเคราะห์จากความน่าจะเป็นการกระจายตัวของคำในฐานข้อมูล ซึ่งจะส่งผลให้การเรียนรู้ของเครื่องสามารถเรียนรู้ได้มีประสิทธิภาพมากขึ้น ร่วมกับวิธีการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ตลอดจนนำการลดคุณลักษณะของคำ มาประยุกต์ใช้ในการสร้างแบบจำลองการจำแนก ซึ่งส่งผลให้ค่าความแม่นยำในการจำแนกสูงขึ้น ในขณะที่ใช้ทรัพยากรที่เหมาะสม โดยองค์ความรู้ที่ได้จากงานวิจัยนี้ สามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการวิเคราะห์อารมณ์ของลูกค้าที่ซื้อสินค้าออนไลน์แบบอัตโนมัติ (Alm, et al., 2005; Gill, et al., 2008)

## วัตถุประสงค์ของการวิจัย

1. เพื่อสร้างแบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง

2. เพื่อเปรียบเทียบประสิทธิภาพแบบจำลองจำแนกอารมณ์จากข้อความภาษาไทย ที่พัฒนาขึ้นกับแบบจำลองที่ใช้วิธีคำนวณค่าน้ำหนักแบบอื่นที่ใช้ในปัจจุบัน

## สมมุติฐานการวิจัย

1. แบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทยที่สร้างขึ้นมีประสิทธิภาพอยู่ในระดับดี โดยวัดจากค่า ความถูกต้อง (Accuracy) มากกว่า 75 %

2. แบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทยที่สร้างขึ้น โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง สามารถจำแนกข้อมูลถูกต้องมากกว่าแบบจำลองที่สร้างขึ้นจากกำหนดค่าดัชนีแบบ Boolean-Weighting (Boolean), TF-Weighting (Term Frequency), TFIDF-Weighting (Term Frequency–Inverse Document Frequency)

## นิยามศัพท์

การเรียนรู้ของเครื่อง (Machine Learning) คือ การทำให้เครื่องเรียนรู้ได้จากข้อมูลตัวอย่างหรือจากสภาพแวดล้อม จุดมุ่งหมายคือการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น เมื่อเรียนรู้แล้วความรู้ที่เรียนได้จะเก็บไว้ในฐานความรู้ด้วยรูปแบบการแทนความรู้อย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชัน

ปัญญาประดิษฐ์ (Artificial Intelligence) หมายถึง ความฉลาดเทียมที่สร้างขึ้นให้กับสิ่งที่ไม่มีชีวิต ปัญญาประดิษฐ์เป็นสาขาหนึ่งในด้านวิทยาการคอมพิวเตอร์และวิศวกรรม ซึ่งสาขาปัญญาประดิษฐ์เป็นการเรียนรู้เกี่ยวกับกระบวนการการคิด การกระทำ การให้เหตุผล การปรับตัว หรือการอนุมาน และการทำงานของสมอง

การทำเหมืองข้อมูล (Data Mining) หรือการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery In Databases - KDD) หมายถึง เทคนิคเพื่อค้นหารูปแบบ (Pattern) จากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และการรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหารูปแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์

การเรียนรู้แบบมีผู้สอน (Supervised Learning) หมายถึง เทคนิคหนึ่งของการเรียนรู้ของเครื่องซึ่งสร้างฟังก์ชันจากข้อมูลสอน (Training Data) ข้อมูลสอนประกอบด้วยวัตถุเข้าและผลที่ต้องการ ผลจากการเรียนรู้จะเป็นฟังก์ชันที่อาจจะให้ค่าต่อเนื่อง (Regression) หรือ ใช้ทำนายประเภทของวัตถุ (Classification) ภารกิจของเครื่องเรียนรู้แบบมีผู้สอนคือการทำนายค่าของฟังก์ชันจากวัตถุเข้าที่ถูกต้องโดยใช้ตัวอย่างสอนจำนวนน้อย โดยเครื่องเรียนรู้จะต้องวางนัยทั่วไป (Generalize) จากข้อมูลที่มีอยู่ไปยังกรณีที่ไม่เคยพบอย่างมีเหตุผล

การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) หมายถึง เทคนิคหนึ่งของการเรียนรู้ของเครื่อง โดยการสร้างโมเดลที่เหมาะสมกับข้อมูล การเรียนรู้แบบนี้แตกต่างจากการเรียนรู้แบบมีผู้สอน คือ จะไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน การเรียนรู้แบบนี้จะพิจารณาวัตถุเป็นเซตของตัวแปรสุ่ม แล้วจึงสร้างโมเดลความหนาแน่นร่วมของชุดข้อมูล

การจำแนกประเภทข้อมูล (Classification) หมายถึง การปัญหาพื้นฐานของการเรียนรู้แบบมีผู้สอน โดยปัญหาคือการทำนายประเภทของวัตถุจากคุณสมบัติต่าง ๆ ของวัตถุ ซึ่งการเรียนรู้แบบมีผู้สอนจะสร้างฟังก์ชันเชื่อมโยงระหว่างคุณสมบัติของวัตถุกับประเภทของวัตถุจากตัวอย่างสอน แล้วจึงใช้ฟังก์ชันนี้ทำนายประเภทของวัตถุที่ไม่เคยพบ เครื่องมือหรือขั้นตอนวิธีที่ใช้สำหรับการแบ่งประเภทข้อมูลเช่น โครงข่ายประสาทเทียม ต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน เนอีฟเบย์ เป็นต้น

การประมวลผลภาษาธรรมชาติ (Natural Language Processing) หมายถึง กระบวนการรับข้อมูลที่เป็นข้อความต่อเนื่อง กระบวนการวิเคราะห์ระดับคำ วิเคราะห์ชนิดคำวิเคราะห์โครงสร้างไวยากรณ์และความหมายของข้อความโดยใช้ฐานความรู้ เพื่อจัดเก็บเข้าคอมพิวเตอร์และเรียกใช้ฐานข้อมูลเพื่อแสดงผลเป็นไปตามเป้าหมายที่ต้องการ

คำหยุด (Stop-Word) หมายถึง คำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง ตัวอย่างเช่น คำบุพบท เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับคำ คำสรรพนามเป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เป็นต้น

รากศัพท์ (Stemming) หมายถึง การหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อทำให้สามารถลดขนาดของดัชนีลงและเพิ่มประสิทธิภาพในการค้นคืนหรือ การจำแนกหมวดหมู่

การสร้างดัชนี (Indexing) หมายถึง การสร้างตัวแทนเนื้อหาของเอกสาร สำหรับใช้ในขั้นตอนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือ การหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร หรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (Term Weighting) ของคุณลักษณะ การสร้างดัชนีโดยทั่วไปที่นิยมใช้กันนั้น จะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสาร

การสกัดคุณลักษณะ (Feature Extraction) หมายถึง การดึงคุณลักษณะ (Feature) ของเอกสารออกมา ซึ่งการดึงคุณลักษณะออกมานั้นพบว่า ส่วนใหญ่จะใช้คำ เป็นตัวแทนคุณลักษณะของเอกสาร ในภาษาไทยนอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้ วลี ประโยค เป็นตัวแทนคุณลักษณะของเอกสารได้เช่นกัน

การเลือกคุณลักษณะ (Feature Selection) หมายถึง กระบวนการลดขนาดของคุณลักษณะที่จะเรียนรู้ โดยมีขั้นตอนวิธีการเลือกคุณลักษณะที่ดีและสัมพันธ์กับกลุ่มเป้าหมาย และขจัดคุณลักษณะที่ไม่ดีออกไป

## ขอบเขตการวิจัย

1. ขอบเขตด้านข้อมูลทดสอบ ข้อมูลข้อความภาษาไทย ที่นำมาใช้ในงานวิจัยนี้ประกอบด้วย ข้อมูลจากเว็บประเภท เว็บท่า (Web Portal) เว็บเครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Web board) เว็บซื้อขายสินค้า (Web Shopping) หรือ เว็บบล็อก (Blog) ซึ่งรวบรวมมาจากความคิดเห็นที่แสดงบนเว็บไซต์ 10 อับดับแรกที่ได้รับความนิยมของไทย ตัวอย่างเช่น ww.facebook.com, www.teenee.com, www.mthai.com, www.pantip.com, www.dek-d.com, www.siamphone.com, www.kapook.com, www.weloveshopping.com, ww.bloggang.com และ www.sanook.com. โดยจัดเก็บมาเป็นคลังข้อมูลขนาดใหญ่ รวมครอบคลุม 6 กลุ่มอารมณ์ อย่างละ 1,500 เอกสาร รวมจำนวนเอกสารในคลังข้อมูลทั้งสิ้น 9,000 เอกสาร

2. ขอบเขตด้านเครื่องมือทดลอง เครื่องคอมพิวเตอร์ หน่วยประมวลผลความเร็วสูง Quad-Core Processor 3.0 GHz หน่วยความจำ 16 GB ระบบปฏิบัติการวินโดวส์ Windows 7 (64-Bit) โปรแกรมจาวา Java Development Kit 8 (64-Bit) ฐานข้อมูล MySQL โปรแกรม Weka 3.70 (64-Bit) และโปรแกรม Matlab R2010 (64-Bit)

3. ขอบเขตด้านการทดสอบระบบ งานวิจัยนี้มุ่งเน้นพัฒนาเทคนิคเทคนิคปรับปรุงดัชนีของคำ ทดสอบเปรียบเทียบประสิทธิภาพในการจำแนกด้านความถูกต้อง โดยเปรียบเทียบประสิทธิภาพกับแบบจำลองวิธีการคำนวณค่าน้ำหนักอื่น ๆ ที่นิยมใช้ในปัจจุบันซึ่งประกอบด้วย กำหนดค่าดัชนีแบบ Boolean-Weighting (Boolean), TF-Weighting (Term Frequency), TFIDF-Weighting (Term Frequency–Inverse Document Frequency)

4. ขอบเขตด้านฐานข้อมูลที่ใช้ในการวิจัย งานวิจัยนี้จะจำแนกผลลัพธ์ของฐานข้อมูลความคิดเห็นออกมาเป็นกลุ่มประเภทอารมณ์ขั้นพื้นฐาน 6 แบบ ซึ่งประกอบด้วย ความโกรธ (Anger) ความกลัว (Fear) ความเศร้า (Sadness) ความรัก (Love) ความสนุก (Joy) และความประหลาดใจ (Surprise) ส่วนอารมณ์อื่นๆที่เกิดขึ้น ล้วนเป็นผลมาจากการเกิดจากการผสมอารมณ์ขั้นพื้นฐานอารมณ์ใด อารมณ์หนึ่งข้างต้นเป็นตัวชี้นำ

## ประโยชน์ของการวิจัย

1.ได้แบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่องที่มีประสิทธิภาพ ทั้งในด้านความถูกต้องในการจำแนกหมวดหมู่และความถูกต้องในภาพรวมด้านการค้นคืนสารสนเทศ ที่ใช้ทรัพยากรของระบบและหน่วยความจำอย่างเหมาะสม แต่สามารถจำแนกข้อความภาษาไทย ได้อย่างมีประสิทธิภาพและประสิทธิผล ซึ่งสามารถนำแบบจำลองนี้ไปประยุกต์ใช้กับงานด้าน การพัฒนาระบบงานการวิเคราะห์อารมณ์ของลูกค้าที่ซื้อสินค้าออนไลน์แบบอัตโนมัติ การวิเคราะห์อารมณ์ของมวลชนในการใช้งานโซเชียลมีเดีย เป็นต้น

2. ได้รับตีพิมพ์ลงในวารสารวิชาการระดับชาติหรือนานาชาติ เพื่อตอบตัวชี้วัดด้านการพัฒนางานวิจัยของมหาวิทยาลัย ซึ่งกำหนดโดยสำนักงานคณะกรรมการการอุดมศึกษา (สกอ.)

# บทที่ 2

# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้มีวัตถุประสงค์ที่จะศึกษาออกแบบขั้นตอนวิธีการเพื่อพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์จากข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง ผู้วิจัยได้ศึกษาค้นคว้าข้อมูลที่เกี่ยวข้องเพื่อให้ได้ข้อมูลสำหรับนำมาพัฒนางานวิจัย โดยรายละเอียดหัวข้อในด้านต่าง ๆ ดังนี้

## เหมืองข้อมูล

การทำเหมืองข้อมูล (Data Mining) หรือที่เรียกว่าการค้นหาความรู้ในฐานข้อมูล (Knowledge Discovery In Databases - KDD) เป็นเทคนิคเพื่อค้นหารูปแบบ (Pattern) จากข้อมูลจำนวนมหาศาลโดยอัตโนมัติ จัดเป็นขบวนการของการดึงเอาความรู้ออกมาจากข้อมูลขนาดใหญ่ โดยใช้ขั้นตอนวิธีจากวิชาสถิติ การเรียนรู้ของเครื่อง และ การรู้จำแบบ หรือในอีกนิยามหนึ่ง การทำเหมืองข้อมูล คือ กระบวนการที่กระทำกับข้อมูลจำนวนมาก เพื่อค้นหาภาพแบบ แนวทาง และความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลนั้น โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่อง และหลักคณิตศาสตร์ (Han and Kamber, 2006) การทำเหมืองข้อมลู เปรียบเสมือนวิวัฒนาการหนึ่งในการจัดเก็บและตีความหมายข้อมูลจากเดิมที่มีการจัดเก็บข้อมูลอย่างง่ายๆ มาสู่การจัดเก็บในรูปฐานข้อมูลที่สามารถดึงข้อมูลสารสนเทศมาใช้จนถึงการทำเหมืองข้อมูลที่สามารถค้นพบความรู้ที่ซ่อนอยู่ในข้อมูล โดยมีผู้ให้นิยามของการทำเหมืองข้อมลู ไว้หลายท่าน เช่น การทำเหมืองข้อมลู คือ การวิเคราะห์ข้อมูลเพื่อหาความสัมพันธ์และสรุปผลจากข้อมูลจานวนมากเพื่อทำความเข้าใจและให้เกิดประโยชน์ต่อเจ้าของธุรกิจ หรือ การทำเหมืองข้อมลู คือ กระบวนการคัดเลือกและสำรวจข้อมูล ตลอดจนเป็นการสร้างแบบจำลองของข้อมูลเพื่อค้นหารูปแบบและค้นหาความสัมพันธ์จากข้อมูลจำนวนมากเพื่อให้ได้ผลลัพธ์ที่เป็นประโยชน์ ซึ่งสรุปได้ว่าการทำเหมืองข้อมูล เป็นการนำข้อมลู มาวิเคราะห์ เพื่อให้ได้ความรู้ใหม่ออกมาเพื่อนำมาช่วยในการตัดสินใจ (David Hand, 2001) เราสามารถสรุปขั้นตอนของการค้นหาความรู้ใหม่จากกระบวนการทำเหมืองข้อมูลได้ดังนี้

1. เรียนรู้และศึกษาเกี่ยวกับฐานข้อมูลและโปรแกรมที่จะใช้ในการทำเหมืองข้อมูล (Data Preprocessing)

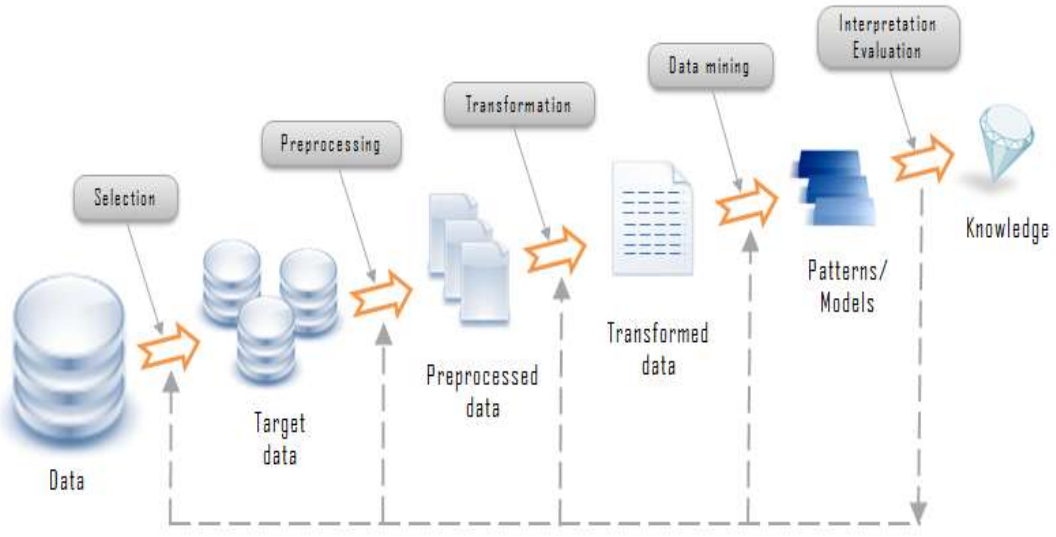
2. การกรองข้อมูลและประมวลผล (Data Cleaning and Integration) เป็นขั้นตอนสำหรับการคัดข้อมูลที่ไม่เกี่ยวข้องออกไป เพราะข้อมูลที่เก็บรวมรวมมามีเป็นจำนวนมาก จึงต้องนำมากรองเพื่อเลือกข้อมูลที่ตรงประเด็นเพราะบางข้อมูลอาจจะไม่เป็นประโยชน์กับเรา ตลอดจนเป็นขั้นตอนการรวมข้อมูลที่มีหลายแหล่งให้เป็นข้อมูลชุดเดียวกัน ซึ่งขั้นตอนนี้เป็นขั้นตอนที่เราจะได้มาซึ่งคุณภาพของข้อมูล ที่พร้อมจะนำไปวิเคราะห์

3. คัดเลือกข้อมูล (Data Selection) เป็นการระบุถึงแหล่งข้อมูลที่จะนำมาทำเหมืองข้อมูล รวมถึงการนำข้อมูลที่ต้องการออกจากฐานข้อมูล เพื่อสร้างกลุ่มข้อมูลสำหรับพิจารณาในเบื้องต้น และทำการแปลงภาพแบบข้อมูล (Data Transformation) ลดภาพและจัดข้อมูลให้อยู่ในภาพแบบเดียวกันมีรูปแบบ (Format) ที่เป็นมาตรฐานและเหมาะสมที่จะนำไปใช้กับอัลกอริทึมและแบบจำลองที่ใช้ทำเหมืองข้อมูล

4. เลือกรูปแบบของการทำเหมืองข้อมูล (Data Mining) เช่น Summarization, Classification, Regression, Association และ Clustering เป็นต้น และเลือกอัลกอริทึมที่เหมาะสมกับลักษณะของงาน จัดเป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์จากข้อมูลที่มีอยู่

5. ขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล (Pattern Evaluation) และ เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้เข้าใจ (Knowledge Representation) ในขั้นตอนนี้จะเป็นการวิเคราะห์ผลลัพธ์ที่ได้และแปลความหมาย และประเมินผลว่าผลลัพธ์นั้นเหมาะสมหรือตรงวัตถุประสงค์หรือไม่

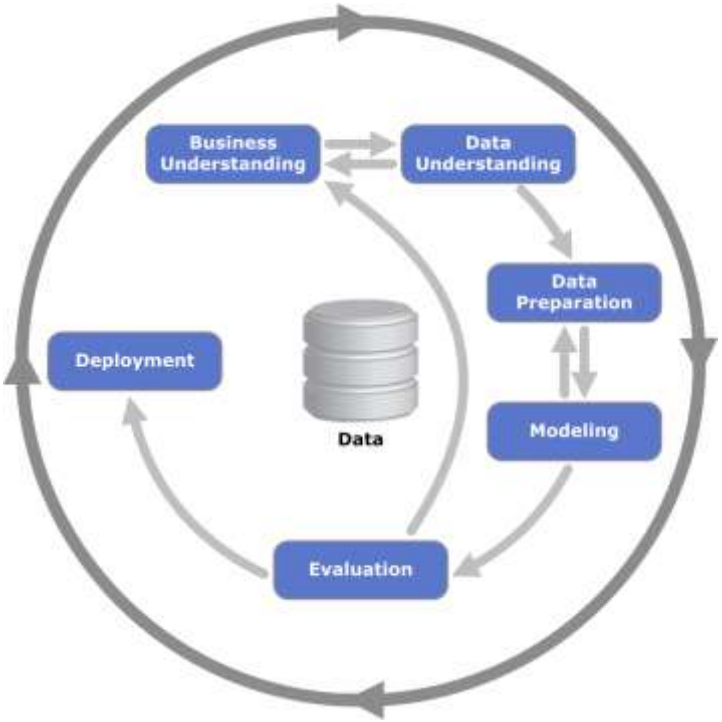
6. ใช้องค์ความรู้ที่ค้นพบขึ้น (Use of Discovered Knowledge) ซึ่งสามารถแสดงขั้นตอนในการทำเหมืองข้อมูล ระบุรายละเอียดขั้นตอนทั้งหมดเป็นแผนภาพได้ดังภาพ




ภาพประกอบ 1 ขั้นตอนกระบวนการทำเหมืองข้อมูล

**กระบวนการการทำงานแบบ CRISP-DM**

ในการวิเคราะห์ข้อมูล ด้วยเทคนิคการทำเหมืองข้อมูล มีกระบวนการมาตรฐานที่เรียกว่า “Cross-Industry Standard Process for Data Mining” หรือชื่อย่อ “CRISP-DM” ได้ถูกจัดคิดค้นเมื่อปลายปี 1996 แบบจำลองกระบวนการที่ใช้ในการเตรียมข้อมูล ทาง Data Mining จะมองตลอด อายุของโครงการซึ่งประกอบด้วยขั้นตอนต่างๆ ของโครงการ ตามลำดับงาน และความสัมพันธ์ระหว่างงาน ในการอธิบายในแต่ละระดับจะเป็นการยากที่จะบอกได้ถึงทุกความสัมพันธ์ ระหว่างงานแต่ละงานทั้งการทำเหมืองข้อมูลขึ้นกับวัตถุประสงค์ที่ตั้งไว้ด้วยกระบวนการในการทำเหมืองข้อมูล (Data Mining) ที่เป็นมาตรฐาน (Shearer C., 2000) CRISP-DM มีขั้นตอนดังนี้




ภาพประกอบ 2 กระบวนการ CRISP-DM

1. ขั้นตอน Business Understanding เป็นขั้นตอนแรกในกระบวนการ CRISP-DM ซึ่งเน้นไปที่การเข้าใจปัญหาและแปลงปัญหาที่ได้ให้อยู่ในรูปโจทย์ของการวิเคราะห์ข้อมูลทางการทำเหมืองข้อมูล พร้อมทั้งวางแผนในการดำเนินการคร่าวๆ ตัวอย่างการนำเทคนิคการทำเหมืองข้อมูลไปใช้ในการวิเคราะห์ด้านต่างๆ

2. ขั้นตอน Data Understanding ขั้นตอนนี้เริ่มจากการเก็บรวบรวมข้อมูล หลังจากนั้นจะเป็นการตรวจสอบข้อมูลที่ได้ทำการรวบรวมมาได้เพื่อดูความถูกต้องของข้อมูล และพิจารณาว่าจะใช้ข้อมูลทั้งหมดหรือจำเป็นต้องเลือกข้อมูลบางส่วนมาใช้ในการวิเคราะห์

3. ขั้นตอน Data Preparation ขั้นตอนนี้เป็นขั้นตอนที่ทำการแปลงข้อมูลที่ได้ทำการเก็บรวบรวมมา (raw data) ให้กลายเป็นข้อมูลที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ โดยการแปลงข้อมูลนี้อาจจะต้องมีการทำข้อมูลให้ถูกต้อง (data cleaning) เช่น การแปลงข้อมูลให้อยู่ในช่วง (scale) เดียวกัน หรือการเติมข้อมูลที่ขาดหายไป เป็นต้น โดยขั้นตอนนี้จะเป็นขั้นตอนที่ใช้เวลามากที่สุดของกระบวนการ CRISP-DM

4. ขั้นตอน Modeling ขั้นตอนนี้จะเป็นขั้นตอนการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูลที่ได้แนะนำไปแล้ว เช่น การจำแนกประเภทข้อมูล หรือ การแบ่งกลุ่มข้อมูล ซึ่งในขั้นตอนนี้หลายเทคนิคจะถูกนำมาใช้เพื่อให้ได้คำตอบที่ดีที่สุด ดังนั้นในบางครั้งอาจจะต้องมีการย้อนกลับไปที่ขั้นตอนที่ 3 Data Preparation เพื่อแปลงข้อมูลบางส่วนให้เหมาะสมกับแต่ละเทคนิคด้วย ตัวอย่างเทคนิคในการวิเคราะห์ข้อมูลต่างๆ เช่น การจำแนกประเภทข้อมูล (Classification) การแบ่งกลุ่มข้อมูล (Clustering) หรือ การหากฎความสัมพันธ์ (Association Rules)

5. ขั้นตอน Evaluation ในขั้นตอนนี้เราจะได้ผลการวิเคราะห์ข้อมูลด้วยเทคนิคการทำเหมืองข้อมูลแล้วแต่ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานต่อไปก็จะต้องมีการวัดประสิทธิภาพของผลลัพธ์ที่ได้ว่าตรงกับวัตถุประสงค์ที่ได้ตั้งไว้ในขั้นตอนแรก หรือ มีความน่าเชื่อถือมากน้อยเพียงใด ซึ่งอาจจะย้อนกลับไปยังขั้นตอนก่อนหน้าเพื่อเปลี่ยนแปลงแก้ไขเพื่อให้ได้ผลลัพธ์ตามที่ต้องการได้ สำหรับการสร้างโมเดลด้วยเทคนิค Classification มีการทดสอบประสิทธิภาพของโมเดลอยู่ 3 แบบใหญ่ คือ วิธี Self-consitency test วิธี Split test หรือวิธี Cross-validation test ในขั้นตอนนี้จะทำการประเมินแต่ละโมเดลด้วยว่ามีส่วนดีส่วนด้อยอย่างไร และควรเลือกใช้โมเดลใด การทำงานในส่วนนี้ต้องอาศัยทักษะในการวิเคราะห์ข้อมูล และธุรกิจ เพื่อช่วยให้การวิเคราะห์ทำได้สะดวกและรวดเร็วขึ้น

6. ขั้นตอน Deployment ในกระบวนการทำงานของ CRISP-DM นั้นไม่ได้หยุดเพียงแค่ผลลัพธ์ที่ได้จากการวิเคราะห์ข้อมูลด้วยเทคนิคทางการทำเหมืองข้อมูลเท่านั้น แม้ว่าผลลัพธ์ที่ได้จะแสดงถึงองค์ความรู้ที่มีประโยชน์ แต่จะต้องนำองค์ความรู้ที่ได้เหล่านี้ไปใช้ได้จริงในองค์กรหรือบริษัท ตัวอย่างเช่น การสร้างรายงานเพื่อให้ผู้บริหารหรือนักการตลาดเข้าใจได้ง่ายและสามารถนำไปออกโปรโมชั่นได้ เป็นต้น

## เทคนิคการทำเหมืองข้อมูล

เนื่องจากการทำเหมืองข้อมูลเป็นเทคนิคในการค้นความรู้จากข้อมูลขนาดใหญ่ การทำเหมืองข้อมูลจึงเป็นการรวมเอาศาสตร์ต่าง ๆ หลายแขนงมารวมไว้ด้วยกันโดยไม่จำกัดวิธีการที่จะใช้ ตัวอย่างศาสตร์ที่ใช้ เช่น เทคโนโลยีฐานข้อมูล (Database Technology) วิทยาศาสตร์สารสนเทศ (Information Science) สถิติ (Statistics) และระบบการเรียนรู้ (Machine Learning) เป็นต้น ซึ่งศาสตร์ต่างๆ เหล่านี้จะทำให้เกิดกระบวนการค้นความรู้ในแบบต่าง ๆ โดยภาพแบบการค้นความรู้หลักมีดังนี้ (Han and Kamber, 2006; Ian and Eibe, 2005)

**1. การแบ่งประเภทและการทำนาย (Classification & Prediction)**

จัดเป็นกระบวนการที่ใช้ในการหาภาพแบบของชุดข้อมูลที่มีความใกล้เคียงกัน หรือเหมือนกันมากที่สุด เพื่อใช้ในการทำนายชุดข้อมูลว่าอยู่ในประเภทใดของชุดข้อมูลที่ได้ทำการแบ่งไว้แล้ว ซึ่งชุดข้อมูลที่แบ่งไว้เกิดจากการเรียนรู้จากชุดข้อมูลที่มีอยู่แล้ว (Training Data) แบบจำลองที่เกิดจากการเรียนรู้สามารถแสดงได้หลายภาพแบบ เช่น กฎการแบ่ง (Classification Rules, IF-THEN) การคำนวณแบบต้นไม้วิเคราะห์ (Decision Tree) การใช้สูตรทางคณิตศาสตร์ (Mathematical Formula) หรือโครงข่ายใยประสาทเทียม เป็นต้น ในส่วนของการทำต้นไม้วิเคราะห์จะแสดงออกมาในลักษณะของแผนภูมิโครงสร้างต้นไม้ ซึ่งก้านของต้นไม้จะแสดงถึงความรู้ที่ได้และใบไม้จะแสดงถึงประเภทชุดข้อมูลที่ถูกแบ่งออกมา แผนภูมิต้นไม้สามารถแปลงเป็นกฎการแบ่งได้ง่ายเพราะลักษณะของแผนภูมิสามารถเข้าใจได้ง่าย ในส่วนของโครงข่ายใยประสาทเทียมนั้น จะแสดงในลักษณะของการเชื่อมต่อระหว่างหน่วยที่เกิดขึ้น การทำการแบ่งประเภทนั้นมักจะใช้ประโยชน์ร่วมกับการทำนายโดยเฉพาะข้อมูลที่เป็นตัวเลข เราจึงอาจมองได้ว่าการทำนายเป็นการบอกถึงค่าตัวเลขและการบ่งบอกประเภทของข้อมูลนั้นในลักษณะของการดูแนวโน้ม (Trends) ที่จะเกิดขึ้น ตัวอย่างเทคนิคของการแบ่งประเภทและการทำนายได้แก่ การคำนวณแบบพันธุกรรม (Genertic Algorithm) การคำนวณแบบต้นไม้วิเคราะห์ และโครงข่ายใยประสาทเทียม (Neural Network) เป็นต้น

**2. การวิเคราะห์เพื่อจัดกลุ่ม (Clustering Analysis)**

การวิเคราะห์เพื่อจัดกลุ่ม จะแตกต่างกับการทำการแบ่งประเภทและการทำนายซึ่งวิเคราะห์กลุ่มข้อมูลที่มีความคล้ายกันมากที่สุด ซึ่งจะเป็นการจัดกลุ่มที่แบ่งประเภทโดยไม่มีการระบุชื่อกลุ่มในช่วงของการสอน แบบจำลองโดยทั่วไปแล้ววิธีแบบนี้จะใช้กับการจัดการแบ่งข้อมูลที่ไม่รู้ว่าจะ

จัดประเภทไว้ด้วยกันอย่างไรดี และการทำการวิเคราะห์นี้จะสามารถทำการบ่งบอกถึงชื่อของกลุ่มที่แบ่งขึ้นได้ด้วย ในการทำการวิเคราะห์เพื่อจัดกลุ่มนั้นจะอาศัยพื้นฐานของความเหมือนกันมากที่สุดและความเหมือนกันน้อยที่สุดของกลุ่ม คือ ข้อมูลที่ถูกจัดไว้ในกลุ่มเดียวกันจะมีความคล้ายกันสูงมาก แต่จะแตกต่างกันกับข้อมูลที่ถูกจัดไว้คนละกลุ่ม และตัวอย่างของการวิเคราะห์เพื่อจัดกลุ่มได้แก่ การหาค่าเฉลี่ย K (K-mean algorithm) การรวมและการแบ่งกลุ่มโดยจัดลำดับชั้น (Agglomerative And Divisive Hierarchical Clustering) และการลำดับตำแหน่งเพื่อแสดงโครงสร้างการจัดกลุ่ม (Ordering Points To Identify The Clustering Structure) เป็นต้น

**3. การวิเคราะห์ความสัมพันธ์ (Association Analysis)**

การวิเคราะห์ความสัมพันธ์เป็นภาพแบบการค้นความรู้โดยการหาสิ่งที่เรียกว่า “กฎความของสัมพันธ์ (Association Rules)” ซึ่งจะแสดงความสัมพันธ์ของค่าที่มีความสัมพันธ์และมีเงื่อนไขที่ตรงกับข้อกำหนดและลักษณะของข้อมูลที่มีการเรียนรู้ในภาพของตะกร้าจ่ายตลาด (Market Basket) หรือการซื้อขาย (Transaction) ในการทำการวิเคราะห์ความสัมพันธ์กฎที่เกิดขึ้นนี้จำเป็นที่จะต้องกำหนดค่าสนับสนุน (Support) และค่าความมั่นใจ (Confidence) ซึ่งจะเป็นตัวกำหนดว่ากฎที่เกิดขึ้นนั้นมีความสัมพันธ์กันในระดับใด และยังเป็นการช่วยยับยั้งการเกิดกฎที่ไม่จำเป็นหรือกฎที่มีความเกี่ยวข้องกันน้อยมาก ตัวอย่างเทคนิคของการวิเคราะห์ความสัมพันธ์ได้แก่ การวิเคราะห์แบบตะกร้าสินค้า (Market Basket Analysis) การคำนวณแบบแอพพริออรี (The Apriori Algorithm) และกฎความสัมพันธ์แบบหลายระดับ (Multilevel Association Rules) เป็นต้น

## การสกัดคุณลักษณะ (Feature Extraction)

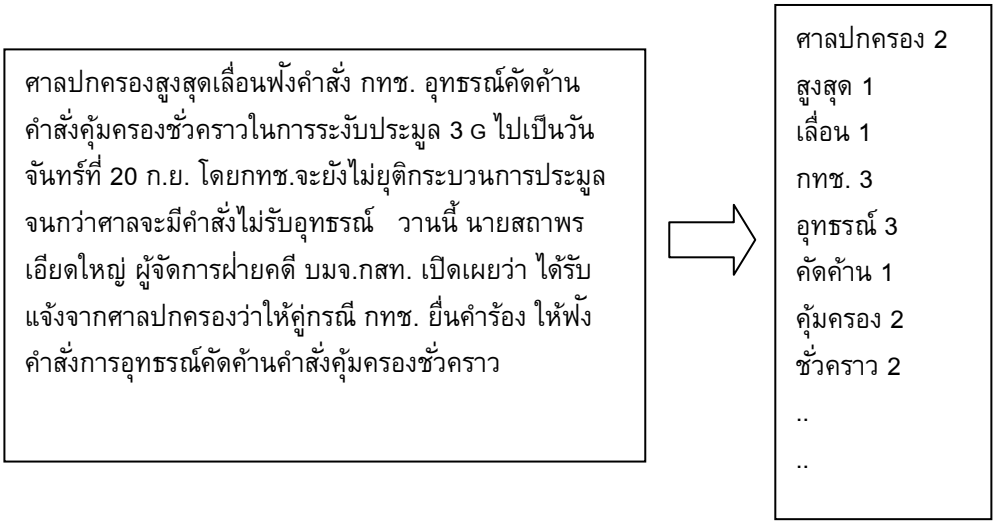
วัตถุประสงค์ของขั้นต้อนการสกัดคุณลักษณะคือ การดึงคุณลักษณะ (Feature) ของเอกสารออกมา ซึ่งการดึงคุณลักษณะออกมานั้น ก่อนอื่นต้องกำหนดก่อนว่าจะใช้อะไรเป็นตัวแทนคุณลักษณะของเอกสารและใช้ค่าใดแทนคุณลักษณะเอกสารนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศและต่างประเทศพบว่า ส่วนใหญ่จะใช้คำ เป็นตัวแทนคุณลักษณะของเอกสารและใช้พื้นฐานค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้ วลีหรือกลุ่มของคำ ประโยค เป็นตัวแทนคุณลักษณะของเอกสารได้เช่นกัน

การสกัดคุณลักษณะของเอกสารคือ การนำเอกสารจากระบบต่าง ๆ เช่น เอกสารข่าวจากเว็บไซต์ จดหมายอิเล็กทรอนิกส์ มาแปลงเพื่อให้ได้เป็นเอกสารในรูปแบบเดียวกัน มีรูปแบบองค์ประกอบที่เหมือนกัน ซึ่งในที่นี้ คือข้อความ (Plain Text) โดยสกัดแท็กทางโครงสร้างในการออกแบบภาษาทิ้งไป และแทนคุณลักษณะในเอกสารแบบใด เช่น ใช้ระดับคำเดี่ยว วลี ฯลฯ เป็น

ตัวแทนคุณลักษณะของเว็บเพจ และการคำนวณหาค่าน้ำหนัก เป็นต้น เทคนิคการสกัดคุณลักษณะอาจใช้เทคนิคการวิเคราะห์ทางภาษาเข้ามาช่วย (Language Analysis) เพื่อตัดคำที่ไม่จำเป็นออก เช่น การตัดคำที่ไม่มีนัยสำคัญออก (Stop-Word List Removal) การทำรากศัพท์ (Word Stemming) หรือตัดเอาคำที่ไม่จำเป็นออกตามสถิติของคลังประโยค การใช้หลัก N-gram ประยุกต์ในการลดทอนคำลงเพื่อลดมิติของขนาดเอกสารลง (Jaruskulchai, 1998; Pang and Lee, 2002; Pang and Lee, 2008)

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN">
<HTML>
 <HEAD>
  <TITLE> บทวิจารณ์ iPad Mini: ไม่ใช่ไอแพดราคาถูก </TITLE>
  <META NAME="Generator" CONTENT="EditPlus">
  <META NAME="Author" CONTENT="">
  <META NAME="Keywords" CONTENT="">
  <META NAME="Description" CONTENT="">
 </HEAD>
 <BODY>
  <TABLE>
  <TR>
    <TD>คุณเอ็มวิเคราะห์ได้ดีมีเหตุผลชัดเจนสมเป็นแฟนพันธุ์แท้
    หลังจากที่วันสองวันนี้เห็นแต่บทวิเคราะห์วิจารณ์ที่ตื่นเขินตามน้ำตรามาเต็มไปหมดทั้งไทยทั้งเทศ.</TD>
  </TR>
  <TR>
    <TD>แหม คมคาย อ่านแล้วกิเลสขึ้นปรี๊ดๆๆๆ :)</TD>
  </TR>
  <TR>
    <TD>ชอบการวิเคราะห์แบบมีเหตุมีผลแบบนี้มากครับ ช่วงหลังรู้สึกว่ามีแต่คนเปรียบเทียบสินค้ากันแบบ คนละ
    categories บ้าง คนละวัตถุประสงค์บ้าง ตรามาสัก 90% เรียกเรตติ้ง อ่านแล้วยังกับว่า Google จ้างคนมาติ iOS
    ยังไงอย่างนั้น เลยพาลให้รู้สึกไปอีกว่า พวก compensate กับ iOS นี่มันเยอะจริงๆ.</TD>
  </TR>
  <TR>
    <TD>คม ชัด ลึก.. ตามสไตล์คาโจชิครับ อยากรู้เรื่อง iPad mini ต้องอ่าน!</TD>
  </TR>
  </TABLE>
 </BODY>
</HTML>
```

ภาพประกอบ 3 ตัวอย่างฐานข้อมูลของเอกสาร




ภาพประกอบ 4 การสกัดคุณลักษณะของเอกสาร

## การตัดคำภาษาไทย (Thai Word Segmentation)

การประมวลผลจำแนกข้อความภาษาไทยได้อย่างมีประสิทธิภาพนั้น ในส่วนของการประมวลผลเบื้องต้นคือการสกัดคุณลักษณะของข้อความในเอกสารภาษาไทย ซึ่งขั้นตอนดังกล่าวมีความจำเป็นต้องการตัดคำในภาษาไทย (Word Segmentation) แต่พบปัญหาเนื่องจากภาษาไทยไม่มีสัญลักษณ์ที่สามารถบ่งบอกถึงขอบเขตของคำ ที่เรียงต่อเนื่องกันทั้งประโยคเหมือนกับภาษาอังกฤษที่ใช้ช่องว่าง(Space) คั่นระหว่างขอบเขตของคำ ซึ่งเป็นอุปสรรคอย่างหนึ่งที่เพื่อตัดสายอักขระภาษาไทยออกเป็นคำๆ ได้อย่างถูกต้อง จึงได้มีผู้คิดค้นพัฒนาวิธีการตัดคำภาษาไทย (Thai Word Segmentation) ซึ่งงานวิจัยด้านการตัดคำภาษาไทยมีการพัฒนาขึ้นด้วยกันหลายวิธี โดยแต่ละวิธีก็มีข้อดีและข้อเด่นของตนเอง และให้ผลลัพธ์ที่แตกต่างกันในด้านของความถูกต้อง ความรวดเร็วในการทำงาน และปริมาณการใช้ทรัพยากร โดยในปัจจุบันยังมีการวิจัยและพัฒนาการตัดคำภาษาไทยอยู่เนืองๆ ซึ่งสามารถแบ่งตามลักษณะฐานข้อมูลที่นำมาใช้ในการตัดคำ ออกเป็น 3 กลุ่มหลัก คือ หลักการตัดคำโดยใช้กฎ (Rule-Based Approach) หลักการตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) และหลักการตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) (ยืน, 2529 ; วิรัช, 2536 ; ไพศาล, 2541)

การตัดคำโดยใช้พจนานุกรม เป็นการตัดคำโดยใช้การเปรียบเทียบคำกับคำที่จัดเก็บอยู่ในพจนานุกรมร่วมกับการใช้กฎในการตัดคำด้วย ซึ่งเป็นการแก้ปัญหาการตัดคำของการใช้กฎเนื่องจากการใช้กฎเพียงอย่างเดียวไม่สามารถหาขอบเขตของคำได้ถูกต้องทั้งหมด โดยวิธีการนี้จะต้องมีการจัดเก็บคำทั้งหมดไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้ามาไปค้นหาและเปรียบเทียบสายอักขระกับคำในพจนานุกรม ปัญหาที่พบในวิธีการตัดคำโดยใช้พจนานุกรม คือไม่สามารถจัดเก็บคำทั้งหมดลงในพจนานุกรมได้ เนื่องจากมีคำใหม่ๆ เกิดขึ้นอยู่ตลอดเวลา ดังนั้นความถูกต้องของวิธีนี้จึงขึ้นอยู่กับปริมาณของคำในพจนานุกรม และต้องสูญเสียพื้นที่ในการจัดเก็บคำตามปริมาณคำศัพท์ในพจนานุกรมด้วย วิธีการตัดคำโดยพจนานุกรมนี้ สามารถแบ่งออกได้อีก 2 วิธี คือ (ยืน, 2529 ; วิรัช, 2536)

วิธีการเทียบคำที่ยาวที่สุด (Longest Matching) โดยวิธีการเทียบคำที่ยาวที่สุดนั้นจะทำการเทียบคำที่ป้อนเข้ามากับคำที่อยู่ในพจนานุกรม ถ้าไม่พบคำที่สามารถเทียบกับคำที่อยู่ในพจนานุกรมได้ ระบบจะทำการลดความยาวของข้อความลงทีละตัวอักษรตามหลักทางอักขระวิธีจนสามารถที่จะเทียบคำกับคำในพจนานุกรมได้ ซึ่งวิธีการนี้การตัดคำมีความถูกต้องมาก แต่มีข้อด้อยคือ การเทียบกับคำที่มีความยาวมากเกินไปตั้งแต่แรกทำให้การตัดคำที่มีตามมามีความผิดพลาดได้ เนื่องจากคำในภาษาไทยบางคำเป็นคำประสมที่เกิดจากหลายคำรวมกัน

วิธีการตัดคำให้ได้จำนวนคำและคำที่ไม่พบในพจนานุกรมน้อยที่สุด (Maximal Matching) จัดเป็นวิธีที่ใช้ในการแก้ปัญหาจากการตัดคำด้วยวิธีการเทียบคำที่ยาวที่สุด ที่เกิดความผิดพลาดจากการเทียบกับคำที่มีความยาวมากเกินไปตั้งแต่ต้น วิธีการนี้จะใช้การตัดคำโดยวิธีการเทียบคำที่ยาวที่สุดก่อนจากนั้นจะทำการย้อนกลับ (Backtracking) เพื่อทำการตัดคำที่สามารถเป็นไปได้อีกครั้ง โดยใช้วิธีพิจารณาทางเลือกของการตัดคำที่เป็นไปได้ทั้งหมดก่อนที่จะเลือกการตัดคำ และผลลัพธ์ของการตัดคำวิธีนี้ จะเลือกจากการตัดคำที่ค่าต้นทุนน้อยที่สุดและได้คำที่ไม่พบในพจนานุกรมให้น้อยที่สุด

| ประโยค | คำที่ยาวที่สุด |
|---|---|
| ความก้าวหน้าทางด้านวิทยาศาสตร์มีบทบาทสำคัญ | - |
| ทางด้านวิทยาศาสตร์มีบทบาทสำคัญ | ความก้าวหน้า |
| ด้านวิทยาศาสตร์มีบทบาทสำคัญ | ทาง |
| วิทยาศาสตร์มีบทบาทสำคัญ | ด้าน |
| มีบทบาทสำคัญ | วิทยาศาสตร์ |
| บทบาทสำคัญ | มี |
| สำคัญ | บทบาท |
| - | สำคัญ |

ภาพประกอบ 5 การตัดคำแบบเทียบคำยาวที่สุด

## การกำจัดคำหยุด (Stop-Word List Removal)

เป็นการนำคำที่ไม่มีนัยสำคัญออกโดยที่ไม่ทำให้ความหมายของเอกสารเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญในที่นี้หมายถึงคำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเอกสาร เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเอกสารเปลี่ยนแปลง เช่น คำบุพบทเป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธานเป็นคำที่ทำหน้าที่เชื่อมคำกับคำ คำสรรพนามเป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เป็นต้น คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเว็บเพจ และปรากฏในเว็บเพจเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกหมวดหมู่ ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์และลดขนาดของดัชนีลง ซึ่งจะช่วยประหยัดทั้งพื้นที่และเวลาในการประมวลผล (Haruechaiyasak, et al., 2008)

ที่ ใน ว่า ได้ เป็น มี จะ และ การ ให้ ของ นี้ มา ไม่ ไป กับ โดย จาก ความ ซึ่ง แต่ แล้ว ยัง ต้อง
วัน ก็ ผู้ ถึง กัน ขึ้น อยู่ คน ด้วย ส่วน อย่าง เมื่อ ทำ รับ งาน เพื่อ กล่าว มาก เพราะ ทาง ต่อ นั้น อีก เข้า ทั้ง
นาย หรือ เรื่องใช้ ปี ดี ออก ผ่าน ตาม ทำ ให้ ขณะ ตัว เวลา กว่า ก่อน นำ ครั้ง ผล ร่วม รวม สามารถ ด้าน

ภาพประกอบ 6 ตัวอย่างคำหยุด

## การหารากศัพท์ (Stemming)

เป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกหมวดหมู่ การหารากศัพท์ของคำภาษาไทยนั้น จะใช้วิธีการรวบรวมคำศัพท์ที่มีความหมายคล้ายกัน หรือมีรากศัพท์เดียวกันไว้เป็นรายการคำศัพท์ หรือจัดเก็บในคลังคำ เพื่อใช้ในการเปรียบเทียบหารากศัพท์ ซึ่งวิธีการนี้ต้องอาศัยมนุษย์เป็นผู้กำหนดไว้ก่อนว่า คำแต่ละคำมีรากศัพท์เป็นคำใด วิธีการนี้ต้องอาศัยผู้เชี่ยวชาญทางภาษาและใช้เวลาในการเก็บรวบรวมและจัดทำรายการคำศัพท์

| คำศัพท์ | รากศัพท์ |
|---|---|
| พิจารณา | พิจาร |
| พิจารณ์ | พิจาร |
| กาพิจารณา | พิจาร |
| พิจาร | พิจาร |

ภาพประกอบ 7 ตัวอย่างรากศัพท์

## การสร้างดัชนี (Indexing)

เนื่องจากคอมพิวเตอร์ไม่สามารถจำแนกหมวดหมู่ของเอกสารซึ่งเป็นภาษาธรรมชาติโดยตรงได้ดังนั้นจึงต้องแปลงเอกสารให้อยู่ในรูปแบบ ที่คอมพิวเตอร์สามารถใช้ในการเรียนรู้ได้ ขั้นตอนในการแปลงเอกสารเรียกว่า การทำดัชนี (Indexing) เพื่อสร้างตัวแทนเนื้อหาของเว็บเพจ (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือการคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสารหรืออาจจะเรียกได้ว่าการหาค่าน้ำหนัก (Term Weighting) การสร้างดัชนี

โดยทั่วไปที่นิยมใช้กันจะเริ่มจากการสร้างเวกเตอร์ตัวแทนเอกสารจากนั้นสร้างเมตริกซ์ของกลุ่มเอกสารขึ้นจากเวกเตอร์เอกสารทั้งหมดในกลุ่ม ซึ่งงานวิจัยนี้ใช้การทดลองคำนวณค่า

น้ำหนักให้กับดัชนีดังต่อไปนี้ โดยให้ (Salton, 1988; Jaruskulchai, 1998; Pang and Lee, 2002; Pang and Lee, 2008)

$Tf$ = เป็นความถี่ของคำ $i$ ในเอกสาร $k$
N = จำนวนเอกสารทั้งหมด
$n_i$ = จำนวนเว็บเพจทั้งหมดที่มีคำ $i$ เกิดขึ้น
$w_{ik}$ = น้ำหนักของคำ

Boolean-Weighting (Boolean)

ค่านี้จะพิจารณาจากการมีอยู่ของคำในแตละเว็บเพจ ถ้ามีคำปรากฏความถี่มากกว่า หรือเท่ากับ 1 ก็จะได้ค่าเป็น 1 แต่ถ้าไม่มีคำนั้น ๆ ปรากฏอยู่เลยจะมีค่าเท่ากับ 0 เรียกอีกอย่างหนึ่งว่าค่าคุณลักษณะความจริง

$$w_{ik} = \begin{cases} 1 \text{ if } tf_{ik} > 0 \\ 0 \; otherwise \end{cases}$$

Term Frequency-Weighting (TF)

ค่านี้จะพิจารณาจากความถี่ของคำที่ปรากฏในแต่ละเว็บเพจโดยตรง ถ้าคำใดมีความถี่มากก็จะได้ค่าน้ำหนักที่มีค่าสูง

$$w_{ik} = tf_{ik}$$

Term Frequency–Inverse Document Frequency -Weighting (TFIDF)

ค่านี้จะพิจารณาจากความถี่ของคำในเว็บเพจ คูณกับฟังก์ชัน log ของเว็บเพจทั้งหมด หารด้วยจำนวนเว็บเพจที่ปรากฏคำนั้นอยู กล่าวคือถ้าคำไหนมีอยู่ในทุกเว็บเพจ ก็จะทำให้ค่า log ของจำนวนเว็บเพจทั้งหมดหารด้วยจำนวนเว็บเพจที่ปรากฏคำนั้นอยูมีค่าเท่ากับ 0 ทำให้ได้ค่าน้ำหนักเท่ากับ 0 ซึ่งเป็นวิธีการให้น้ำหนักแบบมาตรฐานที่ได้รับความนิยม

$$tfidf_{ik} = tf_{ik} * \log\left(\frac{N}{n_i}\right)$$

## การเลือกคุณลักษณะ (Feature Selection)

จากที่กล่าวมาจะพบว่าเอกสารอีเล็กทรอนิกส์นั้น มีแนวโน้มในการที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้เอกสารมีจำนวนคุณลักษณะมากขึ้น ทำให้วิธีเบื้องต้นในการลดขนาดเอกสารคือการใช้การนำคำที่ไม่มีนัยสำคัญออก กับการทำรากศัพท์แล้วยังไม่เพียงพอ ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เอกสาร เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่ โดยทั่วไปไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี การลดขนาดเอกสารจึงเป็นขั้นตอนหนึ่งที่จะต้องทำก่อน การสร้างตัวจำแนกเอกสารแต่การลดขนาดของเอกสารต้องพิจารณาด้วยความระมัดระวัง เนื่องจากมีความเสี่ยงในการที่จะกำจัดคุณลักษณะที่สำคัญต่อการจำแนกหมวดหมู่ออกไป จากการศึกษาพบว่าวิธีการลดคุณลักษะเอกสารประกอบด้วยการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม อาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่ที่มีระดับสูงขึ้นค่าที่นิยมใช้ได้แก่ค่า ค่าสารสนเทศ (Information Gain) ค่าสถิติไคสแควร์ (Chi-square) (Haruechaiyasak, et al., 2008; Liao, et al., 2007; Nicholls, et al., 2010; Sharma, et al., 2012)

## อัลกอริทึมการจำแนกประเภท (Classifier Algorithm)

อัลกอริทึมในการจำแนกประเภทการเรียนรู้แบบมีผลเฉลย (Supervised Learning) สามารถแบ่งขั้นตอนวิธีการจำแนกเอกสารได้เป็น 2 ขั้นตอนคือ การเรียนรู้เพื่อสร้างกลุ่มเอกสารต้นแบบและแยกหมวดหมู่ของเอกสารที่สนใจ โดยการตรวจสอบหาความคล้ายกับกลุ่มเอกสารต้นแบบ

### 1. ต้นไม้ตัดสินใจ (Decision Tree)

เป็นการนำข้อมูลมาสร้างแบบจำลอง (Sui, et al., 2003; Quinlan, 1993) มีลักษณะเป็นผังงาน (Flowchart) เหมือนโครงสร้างต้นไม้ เป็นการเรียนรู้แบบมีผู้สอน (Supervised Learning) คือสร้างแบบจำลองขึ้นมาจากข้อมูลที่นำมาใช้เรียนรู้ โดยต้นไม้ตัดสินใจสามารถนำมาใช้ในการทำนายค่าต่าง ๆ ได้ โดยผลการทำนายจะขึ้นอยู่กับตัวแปรต้น รูปแบบของต้นไม้ ประกอบด้วยโหนดแรกสุดที่เรียกว่า โหนดราก (Root Node) จากโหนดรากก็จะแตกออกเป็นโหนดลูกซึ่งมีกิ่งในการเชื่อมระหว่างโหนด และโหนดลูกก็จะมีลูกของตัวเองซึ่งที่โหนดระดับสุดท้ายจะเรียกว่าโหนดใบ (Leaf Node) แต่ละโหนดของโหนดรากและโหนดลูกจะแสดงค่าคุณลักษณะ (Attribute) ที่ใช้ทดสอบข้อมูล ส่วนโหนดใบจะแสดงกลุ่ม (Class) ที่กำหนดไว้ เมื่อมีข้อมูลที่ต้องการทำนาย การทำงานจะเริ่มต้นจากโหนดราก ซึ่งจะนำค่าคุณลักษณะต่าง ๆ ของข้อมูลนั้นไปเปรียบเทียบกับ

คุณลักษณะของโหนด และทำการตัดสินใจว่าจะเดินทางไปตามกิ่งใด หลังจากนั้นจะเดินทางผ่านโหนดลูก และทำการเปรียบเทียบคุณลักษณะไปเรื่อย ๆ จนกระทั่งสุดท้ายไปถึงโหนดใบ ก็จะได้กลุ่มที่ถูกกำหนด

วิธีการสร้างต้นไม้ตัดสินใจ อันดับแรกจะหาคุณลักษณะที่สำคัญที่สุดมาแบ่งข้อมูลโดยคุณลักษณะนี้จะถูกตั้งให้เป็นโหนดราก จากโหนดรากจะสร้างเส้นทางเชื่อมหรือกิ่งไปยังโหนดลูก โดยจำนวนเส้นทางเชื่อมจะเท่ากับจำนวนค่าที่เป็นไปได้ของคุณลักษณะของโหนดราก ถ้าโหนดลูกเป็นกลุ่มของข้อมูลที่อยู่ในกลุ่มเดียวกันทั้งหมดให้หยุดการสร้างต้นไม้ แต่ถ้าโหนดลูกมีข้อมูลของหลายกลุ่มปะปนกัน ต้องสร้างโหนดลูกเพื่อจำแนกข้อมูลต่อไป โดยวนกลับไปทำขั้นตอนแรก ซ้ำเพื่อเลือกคุณลักษณะที่สำคัญที่สุดมาเป็นตัวแบ่งข้อมูลต่อไป สำหรับความสำคัญของคุณลักษณะสามารถหาได้จากการคำนวณค่าการเพิ่มสารสนเทศ (Information Gain)

ต้นไม้ตัดสินใจ สร้างกิ่งสาขาจะพิจารณาจากค่าความจริงของคุณลักษณะ โดยค่าที่ใช้จะมาจากการคำนวณจากค่าการเพิ่มสารสนเทศ การสร้างต้นไม้ตัดสินใจ C4.5 ใช้ค่ามาตรฐานอัตราส่วนเกน (Gain Ratio) เพื่อเลือกคุณลักษณะที่จะใช้เป็นรากหรือโหนด ถ้าให้ชุดข้อมูล M ประกอบด้วยค่าที่เป็นไปได้ คือ {m1, m2, …, mn} และให้ความน่าจะเป็นที่จะเกิดค่า m1 มีค่าเท่ากับ P(mi) จะได้ว่าค่าเกนสารสนเทศของ M เขียนแทนด้วย I(M) คำนวณได้ดังสมการ

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

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

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

$$Gain(x) = \mathrm{I(T)} = \mathrm{I_x(T)}$$

จากนั้นคำนวณค่าสารสนเทศของการแบ่งแยก (Split Information) ของคุณลักษณะแต่ละตัว ถ้าให้ T คือ ชุดของตัวอย่าง เมื่อแบ่งตัวอย่างนี้ตามคุณลักษณะ x จะได้ชุดของตัวอย่างย่อยในแต่

ละกิ่ง คือ $\{t_1, t_2, \ldots, t_n\}$ จำนวน n ชุด ตามค่าที่เป็นไปได้ในคุณสมบัติ x เมื่อคำนวณค่าสารสนเทศของการแบ่งแยกได้ดังสมการ

$$\text{Split Information} = \sum_{i=1}^{n} \frac{|t_i|}{|\mathrm{T}|} \log_2 \frac{|t_i|}{|\mathrm{T}|}$$

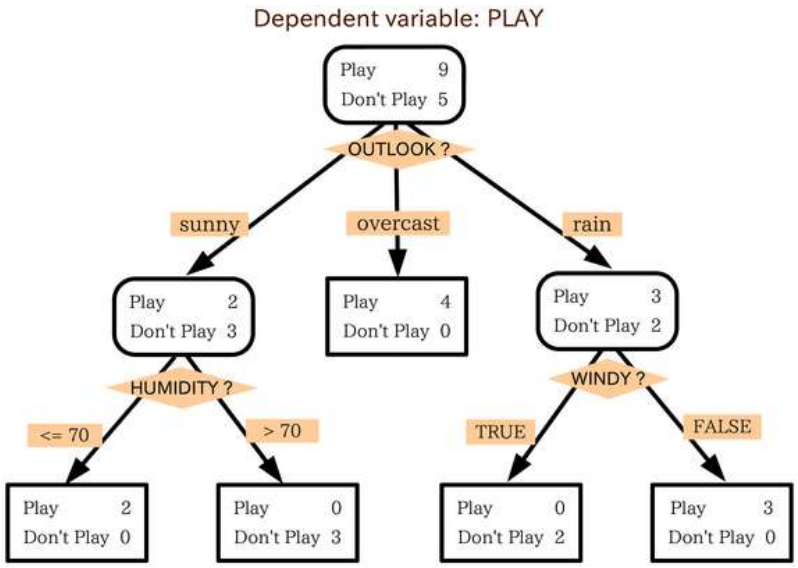
คำนวนค่ามาตรฐานอัตราส่วนเกน (Gain ratio) ได้จาก Gain Ratio = Gain – Split Information ท้ายสุดจึงเลือกค่า Gain ratio สูงสุดเป็นคุณลักษณะเริ่มต้น และเลือกคุณสมบัติถัดไปตามค่า Gain ratio น้อยลงตามลำดับ ลักษณะของปัญหาที่เหมาะกับการใช้ต้นไม้ตัดสินใจถึงแม้ว่าวิธีการเรียนรู้ของต้นไม้ตัดสินใจจะถูกพัฒนาขึ้นมาในหลายรูปแบบเพื่อให้เหมาะสมกับรูปแบบของปัญหาต่างๆ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

1. ข้อมูลแสดงอยู่ในรูปของชุดของคุณลักษณะ (เช่น อุณหภูมิ) และค่าของคุณลักษณะ (เช่น ร้อน) สำหรับรูปแบบที่ง่ายที่สุดสำหรับการเรียนรู้ของต้นไม้ตัดสินใจคือ เมื่อคุณลักษณะมีกลุ่มอยู่เพียงไม่กี่ตัว (เช่น ร้อน เย็น ปกติ) อย่างไรก็ตามมีอัลกอริทึมที่พัฒนาขึ้นมาเพื่อรองรับค่าของคุณลักษณะที่เป็นจำนวนจริง

2. ฟังก์ชันเป้าหมาย (Target Function) มีเอาต์พุตเป็นแบบไม่ต่อเนื่อง (Discrete Output Value) แต่ต้นไม้ตัดสินใจก็มีอัลกอริทึมเสริมซึ่งอนุญาตให้เอาต์พุตของฟังก์ชันเป้าหมายสามารถเป็นจำนวนจริงได้

3. อนุญาตให้มีข้อมูลที่ผิดพลาด (Error) อยู่ในข้อมูลที่ใช้ในการสอนได้ เนื่องจากวิธีการเรียนรู้ของต้นไม้ตัดสินใจนั้นทนทาน (Robust) ต่อความผิดพลาด โดยยอมให้มีความผิดพลาดได้ทั้งค่าคุณลักษณะ และการแบ่งกลุ่ม

4. ข้อมูลที่ใช้ในการเรียนรู้ไม่จำเป็นต้องมีค่าของคุณลักษณะอย่างครบถ้วน นั่นคือวิธีการของต้นไม้ตัดสินใจสามารถใช้ได้ ถึงแม้ข้อมูลนั้นจะมีค่าบางค่าของคุณลักษณะบางตัวหายไปหรือไม่ทราบค่าก็ตาม




ภาพประกอบ 8 ต้นไม้ตัดสินใจ

**2. ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine)**

หลักการของวิธีการนี้ (Sui, et al., 2003; Yang, et al., 2012) ใช้เพื่อหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน ซัพพอร์ตเวกเตอร์แมชชีนจัดเป็นการเรียนรู้ของเครื่องประเภทที่ต้องมีตัวอย่างในการเรียนรู้ (Supervised Learning) ประเภทหนึ่งซึ่งมีความสามารถในการคัดแยก (Classification) และการทำนาย (Regression) โดยเอสวีเอ็มมีพื้นฐานจากการคำนวณแบบ Linear Classifier ซึ่งจัดอยู่ในประเภทมุ่งหาผลลัพธ์ที่ดีที่สุดของการเรียนรู้ (Discriminative Training) บนการเรียนรู้จากสถิติของข้อมูล ซึ่งทำงานโดยการหาค่าระยะขอบที่มากที่สุด (Maximum Margin) ของระนาบตัดสินใจ (Decision Hyperplane) ในการแบ่งแยกกลุ่มข้อมูลที่ใช้ฝึกฝนออกจากกัน โดยจะใช้ฟังก์ชั่นแม็ปข้อมูลจาก Input Space ไปยัง Feature Space และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) บน Feature Space โดยมีวัตถุประสงค์ที่จะพยายามที่จะทำการลดความผิดพลาดจากการทำนาย (Minimize Error) พร้อมกับเพิ่มระยะแยกแยะให้มากที่สุด (Maximized Margin) ซึ่งต่างจากเทคนิคโดยทั่วไปเช่น โครงข่ายประสาทเทียม (Artificial Neural Network: ANN) ที่มุ่งเพียงทำให้ความผิดพลาดจากการทำนายให้ต่ำที่สุดเพียงอย่างเดียว เหมาะใช้สำหรับข้อมูลที่มีลักษณะมิติของข้อมูลที่มีปริมาณมาก หลักการทำงานของเอสวีเอ็ม ข้อมูลจะถูกเขียนในรูปสมาชิกคู่อันดับดังนี้

$$\{(x_1, c_1), (x_2, c_2), (x_3, c_3), (x_n, c_n)\}$$

เมื่อ $c_i$ มีค่าเป็น 1 หรือ -1 ซึ่งกำหนดให้เป็นข้อมูลแบ่งกลุ่มของข้อมูล $x_i$ ที่มีค่า $c_i$ เป็น 1 และ $x_i$ ที่มีค่า $c_i$ เป็น -1 โดยที่แต่ละ $x_i$ เป็นค่าข้อมูลมิติของเวกเตอร์จริง เมื่อกำหนดให้ข้อมูลนี้เป็นข้อมูลสำหรับฝึกฝน ซึ่งหมายความว่าการแบ่งกลุ่มของข้อมูลมีความถูกต้อง ดังนั้นเส้นแบ่งกลุ่มข้อมูลที่ถูกสร้างขึ้นซึ่งเป็นสมการเส้นตรงทั่วไปคือ y = mx+b โดยในที่นี้จะแทน m ด้วย $w^T$ เพื่อกำหนดเป็นสมการ (2-2) แสดงการแบ่งข้อมูลดังภาพ

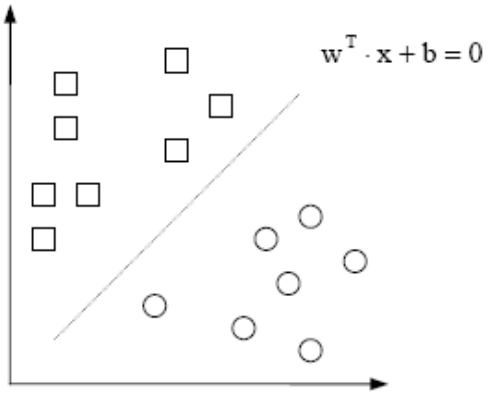
$$W^T \cdot X + b = 0$$

เมื่อ $W^T$ เป็นเวกเตอร์ตั้งฉากของค่าความชัน m ของเส้นแบ่ง

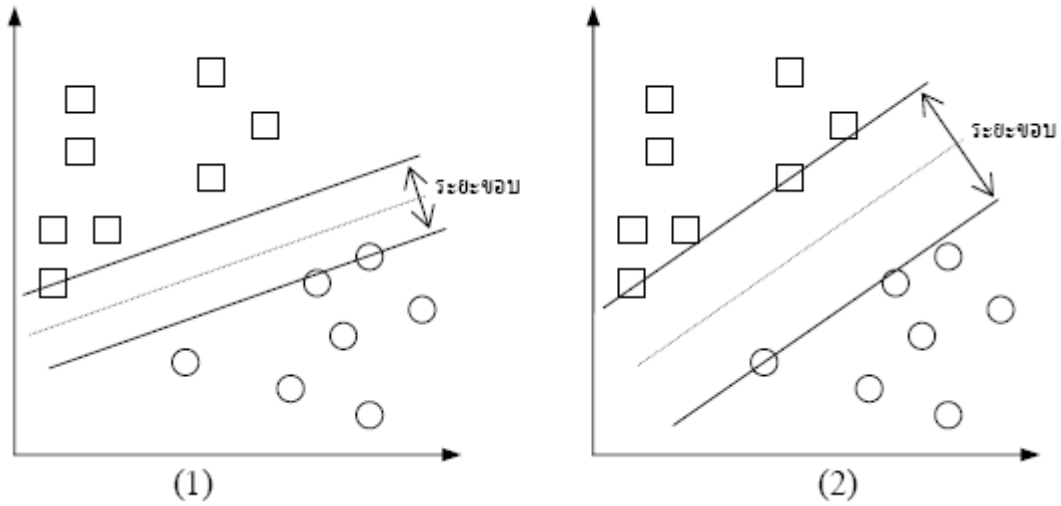
b เป็นค่าคงที่ที่ได้จากค่าของแกน y ของแต่ละข้อมูล x

ดังนั้นในทางอุดมคติ เส้นแบ่งกลุ่มที่ดีที่สุดคือ เส้นแบ่งกลุ่มที่ทำให้มีระยะขอบ (Margin) มากที่สุดของเส้นคู่ขนานที่ขยายออกจากเส้นแบ่งไปสัมผัสจุดข้อมูลของทั้งสองกลุ่มที่ใกล้ที่สุด โดยเรียกเส้นขนานนั้นว่า Parallel Hyper plane หรือกล่าวได้ว่าเป็นเส้นแบ่งที่มีระยะขอบกว้างที่สุด ดังภาพ เรียกจุดข้อมูลอย่างน้อยหนึ่งจุดจากทั้งสองกลุ่มที่สัมผัสกับเส้นขนานที่ขยายออกได้มากที่สุดว่าซัพพอร์ตเวกเตอร์ โดยค่าของเส้นขอบทั้งสองที่ใช้แบ่งกลุ่มข้อมูลเป็นสองกลุ่มคำนวณได้จากสมการ

$$W^T \cdot X + b = 1$$
$$W^T \cdot X + b = -1$$




ภาพประกอบ 9 ข้อมูลสองกลุ่มที่ถูกแบ่งด้วยเส้นตรง




ภาพประกอบ 10 การปรับความชันของเส้นแบ่งแล้วทำให้ได้ระยะขอบที่มากที่สุด

ดังนั้นเมื่อข้อมูลที่ใช้ฝึกฝนเป็นข้อมูลที่สามารถแบ่งกลุ่มได้ด้วยเส้นตรงใด ๆ ระยะที่กว้างที่สุดของ Hyper plane ทั้งสองที่ขยายออกไปจนกว่าจะพบจุดข้อมูลของทั้งสองกลุ่มคือ 2/|w| โดยค่า |w| มีค่าน้อยที่สุด ค่าข้อมูลแต่ละ $x_i$ จัดจำแนกอยู่ในกลุ่มใดสามารถพิจารณาได้จากเงื่อนไขดังนี้

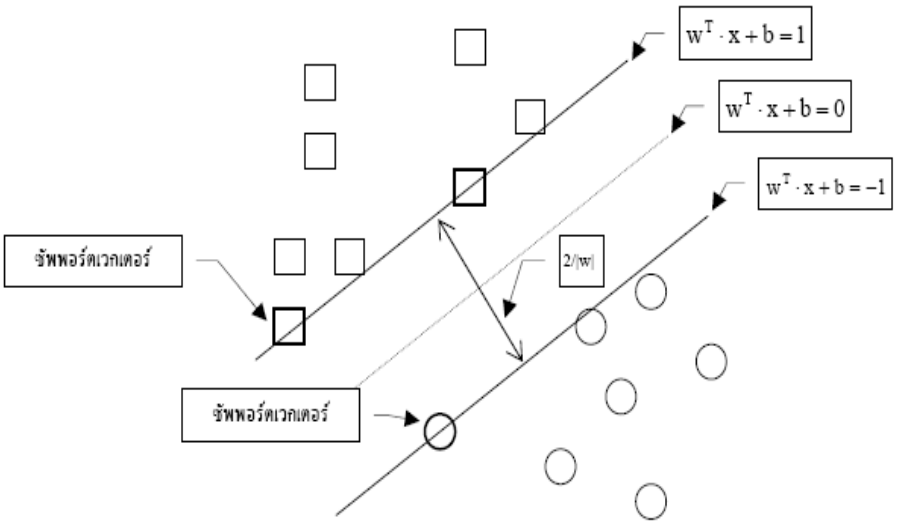
ถ้า $W^T \cdot X_i + b \geq 1$ แสดงว่า $x_i$ เป็นกลุ่มที่ 1
และ ถ้า $W^T \cdot X_i + b \leq -1$ แสดงว่า $x_i$ เป็นกลุ่มที่ 2

ดังนั้นในการคัดแยกจุดข้อมูลใด ๆ ที่นำมาฝึกฝนเพื่อกรองว่าเป็นกลุ่ม 1 สามารถตรวจสอบได้จากสมการ

$$c_i(W \cdot X_i - b) \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n$$

ดังนั้นสามารถเขียนเป็นรูปสั้นเพื่อหาระยะขอบที่น้อยที่สุด โดยที่สามารถคำนวณการแบ่งกลุ่มได้ดังสมการ

$$Minimize_{w,b} |W| \text{ โดยตรวจสอบ } c_i(W \square X_i - b) \geq 1 \quad \text{เมื่อ } 1 \leq i \leq n$$




ภาพประกอบ 11 ระยะขอบที่กว้างที่สุดเมื่อสัมผัสกับซัพพอร์ตเวกเตอร์

**3. เนอีฟเบย์ (Naïve-Bayes)**

การเรียนรู้แบบเบย์ (Narayanan, et al., 2013) เป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น ซึ่งมีพื้นฐานมาจากทฤษฎีของเบย์ (Bayes Theorem) เข้ามาช่วยในการเรียนรู้ จุดมุ่งหมายก็เพื่อต้องการสร้างโมเดลที่อยู่ในรูปของความน่าจะเป็น ซึ่งเป็นค่าที่บันทึกได้จากการสังเกต จากนั้นนำโมเดลมาหาว่าสมมติฐานใดถูกต้องที่สุดโดยใช้ความน่าจะเป็นเข้ามาช่วย ความรู้ก่อนหน้า หมายถึง ความรู้ที่เรามีเกี่ยวกับสมมติฐานแต่ละตัวก่อนที่เราจะเก็บข้อมูล เมื่อใช้งานเราจะนำความน่าจะเป็นของข้อมูลที่เก็บได้มาปรับสมมติฐานซ้ำอีกครั้ง ข้อดีของวิธีการเรียนรู้แบบนี้ คือ สามารถใช้ข้อมูลและความรู้ก่อนหน้า (Prior Knowledge)

วิธีการเรียนรู้เบย์อย่างง่าย (Naïve Bayesian Learning) การเรียนรู้เบย์อย่างง่าย เป็นวิธีจำแนกประเภทข้อมูลที่มีประสิทธิภาพวิธีหนึ่ง โดยที่ใช้งานได้ดี เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไม่ขึ้นต่อกัน มีการนำจำแนกประเภทเบย์อย่างง่ายไปประยุกต์ใช้งานในด้านการจำแนกประเภทข้อความ (Text Classification) การวินิจฉัย (Diagnosis) และพบว่าใช้งานได้ดีไม่ต่างจากการจำแนกประเภทวิธีการอื่น เนื่องจากเป็นวิธีการจำแนกข้อมูลได้อย่างมีประสิทธิภาพและอัลกอริทึมในการทำงานที่ไม่ซับซ้อน

หลักการของวิธีการนี้ ใช้การคำนวณความน่าจะเป็นซึ่งถูกใช้ในการทำนายผล จัดเป็นเทคนิคในการแก้ปัญหาแบบที่สามารถคาดการณ์ผลลัพธ์ได้และสามารถอธิบายได้ด้วย จะทำการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปร เพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละ

ความสัมพันธ์ เหมาะกับกรณีของเซตตัวอย่างมีจำนวนมากและคุณสมบัติ (Attribute) ของตัวอย่างไมขึ้นต่อกัน โดยกำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็น ดังสมการ กำหนดให้ความน่าจะเป็นของข้อมูลที่จะเป็นกลุ่ม $v_j$ สำหรับข้อมูลที่มีคุณสมบัติ n ตัว

X = $\{a_1, a_2, a_3, a_n\}$ หรือ ใช้สัญลักษณ์ว่า $P(a_1, a_2..., a_3 | v_j)$ คือ

$$P(a_1, a_2..., a_3 | v_j) = \prod_{i=1}^{n} P(a_1 | v_j)$$

$P(a_i | v_j)$

$P(v_j)$

$V_{NB}$

โดยที่ $\prod$ หมายถึง ผลคูณของค่า $P(a_i | v_j)$ ทั้งหมด i = 1,2,3,…,n และ j = 1,2,3,…,n ทำการหาค่าความน่าจะเป็นของคำที่พบในแต่ละกลุ่มโดยนำค่า $P(a_1, a_2..., a_3 | v_j)$ จากสมการ มาคูณกับค่าความน่าจะเป็นของกลุ่มนั้น ๆ คือ $P(v_j)$ ได้เท่ากับ $V_{NB}$ นำค่าที่ได้ มาเปรียบเทียบกัน กลุ่มที่มีค่าความน่าจะเป็นสูงสุด คือ คำตอบ ดังนั้นจะได้ว่า วิธีการจำแนกประเภทแบบเบย์อย่างง่าย ดังสมการ

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

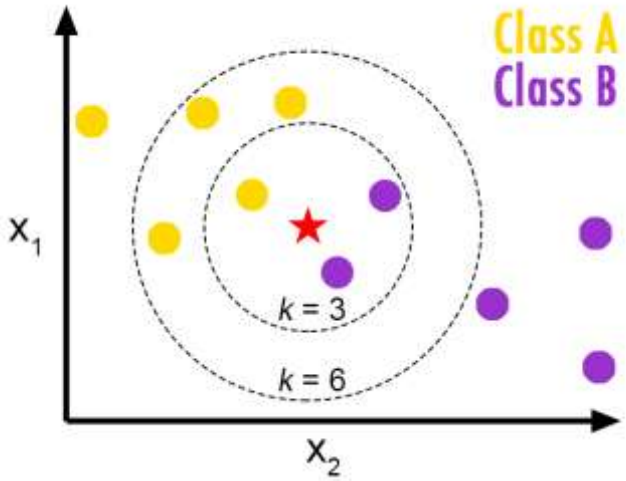
จากข้อมูลการเล่นเทนนิส สามารถคำนวณการเล่นเทนนิสได้ดังนี้ สมมติต้องการทราบว่า ณ สภาพอากาศ Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong แล้วจะเล่นเทนนิสหรือไม่? คำตอบก็คือ “no” เพราะว่าค่าความน่าจะเป็นในการเกิด no มากกว่า yes ดังตัวอย่าง

$$v_{NB} = \arg\max_{v_k \in [yes, no]} P(v_k) P(Outlook = sunny | v_k) P(Temp = cool | v_k)$$
$$P(Humidity = high | v_k) P(Wind = strong | v_k)$$
$$P(PlayTennis = yes) = 9/14 = 0.64$$
$$P(PlayTennis = no) = 5/14 = 0.36$$
$$P(Wind = strong | PlayTennis = yes) = 3/9 = 0.33$$
$$P(Wind = strong | PlayTennis = no) = 3/5 = 0.60$$
$$...$$
$$P(yes) P(sunny | yes) P(cool | yes) P(high | yes) P(strong | yes) = 0.0053$$
$$P(no) P(sunny | no) P(cool | no) P(high | no) P(strong | no) = \mathbf{0.0206}$$
$$\Rightarrow answer : PlayTennis(x) = no$$

**4. เคเนียเรสเนเบอร์ (K-Nearest Neighbor)**

เคเนียเรสเนเบอร์ (K-Nearest Neighbor) (Han and Kamber, 2006) หลักการของวิธีการนี้จะจำแนกประเภทข้อมูลโดยขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียงที่สุด 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

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



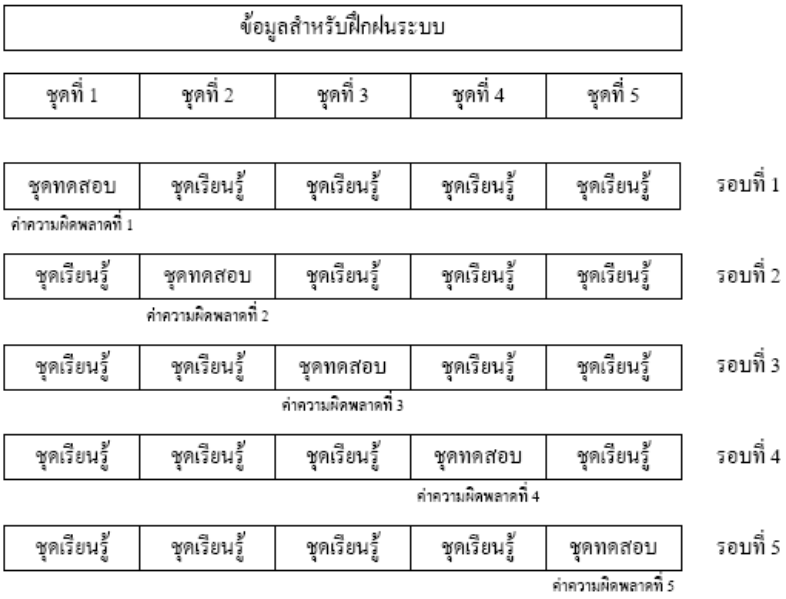

ภาพประกอบ 12 เคเนียเรสเนเบอร์

## การประเมินผลโมเดล

ครอสวาลิเดชั่น (cross validation) (Han and Kamber, 2006; Ian and Eibe, 2005; Haruechaiyasak, et al., 2008) คือ วิธีการในการคาดการค่าความผิดพลาดของโมเดลหรือ วิธีการที่เรานำเสนอโดยพื้นฐานของวิธีการ ครอสวาลิเดชั่นคือ การสุ่มตัวอย่าง โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วน ๆ และนำบางส่วนจากชุดข้อมูลนั้นมาตรวจสอบ ผลลัพธ์จากการทำครอสวาลิเดชั่นมักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล เช่น สถาปัตยกรรมเครือข่ายการสื่อสาร (Network Architecture) โมเดลในการคัดแยกประเภท (Classification Model) นั้นจะต้องมีการแบ่งข้อมูลออกเป็นชุดสอนและชุดทดสอบ แต่ในบางครั้งอาจเกิดปัญหาจากการเลือกข้อมูลที่ดีและง่ายมาเป็นข้อมูลชุดทดสอบทำให้ผลการ Classify นั้นดีเกินจริง ดังนั้นจะมีการคิดวิธี K-Fold Cross Validation ขึ้นมาแก้ปัญหา การตรวจสอบความถูกต้องของการเรียนรู้ สามารถสังเกตได้จากค่าความแม่นยำหรือค่าความผิดพลาดที่ได้จากการทำ ครอสวาลิเดชั่น ระหว่างการแบ่งชุดทดสอบในการเรียนรู้ ในที่นี้ผู้วิจัยเลือกพิจารณาความแม่นยำ ซึ่งเป็นสัดส่วนร้อยละของการคัดแยกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด โดยใช้การครอสวาลิเดชั่นแบบ k-fold ซึ่งพื้นฐานของวิธีการครอสวาลิเดชั่นคือการสุ่มตัวอย่างข้อมูลไปเป็นชุดฝึกฝนและชุดทดสอบแบบสลับกัน โดยเริ่มจากแบ่งชุดข้อมูลออกเป็นส่วนๆ และนำบางส่วนจากชุดข้อมูลนั้นมาเป็นชุดทดสอบเพื่อจำลองผลลัพธ์จากการทำครอสวาลิเดชั่น การแบ่งชุดทดสอบแบบนี้มักถูกใช้เป็นตัวเลือกในการกำหนดโมเดล โดยสังเกตจากผลของการทดสอบในแต่ละครั้งของการปรับค่าพารามิเตอร์แต่ละโมเดล การทำครอสวาลิเดชั่น

วิธี k-fold ครอสวาลิเดชั่นเป็นการแบ่งข้อมูลออกเป็น k ชุด เท่า ๆ กัน และทำการคำนวณค่าความผิดพลาด k รอบ โดยแต่ละรอบการคำนวณ ข้อมูลชุดหนึ่งจากข้อมูล k ชุดจะถูกเลือกออกมาเพื่อเป็นข้อมูลทดสอบ และข้อมูลอีก k-1 ชุดจะถูกใช้เป็นข้อมูลสำหรับการเรียนรู้ แล้วนำผลความถูกต้องหรือค่าความผิดพลาดของแต่ละรอบมารวมกันและหาค่าเฉลี่ยเพื่อเป็นค่าสะท้อนประสิทธิภาพของการฝึกฝน ดังตัวอย่างในภาพประกอบ 2-9 เป็นการกำหนดค่า k=5 หมายถึง การนำข้อมูลทั้งหมดมาแบ่งเป็น 5 ชุด และจะดำเนินการเรียนรู้และทดสอบจำนวน 5 รอบ โดยในแต่ละรอบจะเลือกข้อมูลมา 1 ชุดไม่ซ้ำกันมาเป็นชุดทดสอบ ด้วยการเรียนรู้ของชุดเรียนรู้ที่เหลือ เมื่อทำครบ 5 รอบ หาผลรวมของค่าความผิดพลาดหรือความแม่นยำจากชุดทดสอบในแต่ละรอบแล้วหารด้วยจำนวนรอบ ผลที่ได้เป็นค่าเฉลี่ยของความผิดพลาดในการเรียนรู้ของโมเดล ข้อดีของวิธีการนี้คือ ข้อมูลในแต่ละชุดที่ทำการแบ่งจะถูกทดสอบอย่างน้อย 1 ครั้ง และถูกเรียนรู้ทั้งหมด k-1 ครั้ง โดยในขั้นตอนเหล่านี้สามารถกำหนดได้ว่าต้องการขนาดข้อมูลขนาดใด และต้องการทำการ

คำนวณเป็นจำนวนรอบเท่าใด ซึ่งเหมาะสำหรับการประมวลผลทดสอบกับข้อมูลที่มีมิติขนาดจำนวนมาก




ภาพประกอบ 13 ตัวอย่างการแบ่งชุดข้อมูลของ k-fold ครอสวาลิเดชั่นเมื่อ k = 5

การประเมินความสามารถของระบบการจำแนกหมวดหมู่นั้น เน้นความสามารถในการตัดสินใจหรือการทำนายหมวดหมู่ที่ถูกต้อง วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพของโมเดลการพัฒนาประสิทธิภาพการจัดหมวดหมู่เว็บเพจภาษาไทยอัตโนมัตินั้น สามารถพิจารณาได้จากค่าความถูกต้อง โดยวัดที่ประสิทธิผลของการจำแนกข้อมูลตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) และค่าความระลึก (Recall) ซึ่งคำนวณได้ดังตารางการณ์ (Contingency Table) ได้ดังนี้

| การจำแนกหมวดหมู่ | อยู่ในหมวดหมู่ $c_j$ | ตัดสินโดยผู้เชี่ยวชาญ | |
|---|---|---|---|
| | | ใช่ | ไม่ใช่ |
| ตัดสินโดยตัวจำแนกอัตโนมัติ | ใช่ | $TP_j$ | $FP_j$ |
| | ไม่ใช่ | $FN_j$ | $TN_j$ |

ภาพประกอบ 14 Confusion matrix

$TP$ = จำนวนตัวอย่างที่อยู่ในกลุ่ม $C_j$ และตัวจำแนกทำนายว่าอยู่ในกลุ่ม $C_j$

$FP$ = จำนวนตัวอย่างที่ไม่อยู่ในกลุ่ม $C_j$ และตัวจำแนกทำนายว่าอยู่ในกลุ่ม $C_j$

$FN$ = จำนวนตัวอย่างที่อยู่ในกลุ่ม $C_j$ และตัวจำแนกทำนายว่าไม่อยู่ในกลุ่ม $C_j$

$TN$ = จำนวนตัวอย่างที่ไม่อยู่ในกลุ่ม $C_j$ และตัวจำแนกทำนายว่าไม่อยู่ในกลุ่ม $C_j$

$C_j$ = กลุ่มประเภทของเว็บเพจภาษาไทยที่สนใจวัดประสิทธิภาพ

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

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

## ประเภทของกลุ่มอารมณ์

อารมณ์ มาจากภาษาอังกฤษ "Emotion" มีความหมายว่าการเกิดการเคลื่อนไหว หรือภาวะที่ตื่นเต้น มันเป็นการยากที่จะบอกว่า อารมณ์คืออะไร แต่มีแนวคิดหนึ่ง ที่ให้ความเข้าใจได้ง่ายกล่าวไว้ว่า อารมณ์เป็นความรู้สึกภายในที่เร้า ให้บุคคลกระทำ หรือเปลี่ยนแปลงภายในตัว ของเขาเอง ซึ่งความรู้สึก เหล่านี้จะเป็นความรู้สึกที่พึงพอใจ ไม่พึงพอใจ หรือรวมกันทั้งสองกรณี อารมณ์เป็นสิ่งที่ไม่คงที่มีการแปรเปลี่ยน อยู่ตลอดเวลา ในแต่ละวันบุคคลจะมีอารมณ์ต่าง ๆ เกิดขึ้นมากมาย อาจจะเป็นความพึงพอใจ ความโกรธ ความร่าเริง ความเจ็บปวด ความผิดหวัง เพราะตลอดเวลาที่บุคคล อยู่ในสถานการณ์ใดสถานการณ์หนึ่ง บุคคลจะอยู่ภายใต้สิ่งเร้า (Stimulus) และประสบการณ์ (Experience) ที่เขามีอยู่ทำให้อารมณ์แปรเปลี่ยนไปมา ซึ่งอารมณ์ในลักษณะดังกล่าวนี้จะมีอิทธิพลต่อพฤติกรรม ของบุคคล อารมณ์เป็นปฏิกิริยาที่เกิดขึ้นภายใน เป็นพฤติกรรมที่เกิดขึ้นจากการเรียนรู้ และเป็นเหมือนตัวกระตุ้นให้เกิดแรงจูงใจ ที่จะนำไปสู่พฤติกรรมนั้น อารมณ์และแรงจูงใจจึงเป็นปรากฏการณ์ที่เชื่อมโยงกันอย่างใกล้ชิด (Gill, et al., 2008; Alm, et al., 2005; Gerrod, et al., 2001) อารมณ์มีอยู่มากมายหลายชนิดซึ่งเราอาจเรียกมันว่าอะไรก็ตาม แต่ว่าอารมณ์เหล่านั้น ก็มีความเด่นชัดและเป็นอิสระ นักจิตวิทยาได้จำแนก อารมณ์ โดยคำนึง สิ่งเร้าที่มาเป็นตัวกระตุ้น และรูปแบบการตอบสนองพฤติกรรมที่มีต่อสิ่งเร้านั้น ในวิจัยนี้จะจำแนกผลลัพธ์ของความคิดเห็นออกมาเป็นกลุ่มประเภทอารมณ์พื้นฐาน 6 แบบ ซึ่งประกอบด้วย ความโกรธ (Anger) ความกลัว (Fear) ความเศร้า (Sadness) ความรัก (Love) ความสนุก (Joy) และความประหลาดใจ (Surprise) ส่วนอารมณ์อื่นๆที่เกิดขึ้น ล้วนเป็นผลมาจากการเกิดจากการผสมอารมณ์ขั้นพื้นฐานอารมณ์ใดอารมณ์หนึ่งข้างต้นเป็นตัวชี้นำ

# บทที่ 3

# วิธีดำเนินงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง เพื่อการจำแนกความรู้สึกทางอารมณ์จากฐานข้อมูลข้อความภาษาไทย วัดประสิทธิภาพในด้านความถูกต้องในการจำแนก (Accuracy) ซึ่งผู้วิจัยได้แบ่งวิธีการดำเนินงานออกเป็น 5 ขั้นตอน ได้แก่

## การสกัดคุณลักษณะ (Feature Extraction)

วัตถุประสงค์ของขั้นตอนการสกัดคุณลักษณะจากฐานข้อมูลข้อความภาษาไทย คือ การดึงคุณลักษณะ (Feature) ของข้อความออกมา ซึ่งการดึงคุณลักษณะออกมานั้น ก่อนอื่นต้องกำหนดก่อนว่าจะใช้อะไรเป็นตัวแทนคุณลักษณะของข้อความ และใช้ค่าใดแทนคุณลักษณะข้อความนั้น จากการสำรวจงานวิจัยที่ผ่านมาทั้งในประเทศและต่างประเทศพบว่า ส่วนใหญ่จะใช้คำเป็นตัวแทนคุณลักษณะของข้อความ และใช้พื้นฐานค่าความถี่ของคำเป็นค่าของคุณลักษณะ นอกจากการใช้คำเดี่ยวแล้ว ยังสามารถใช้ วลี หรือกลุ่มของคำ ประโยค เป็นตัวแทนคุณลักษณะของข้อความได้เช่นกัน การสกัดคุณลักษณะจากฐานข้อมูลข้อความภาษาไทย ที่นำมาใช้ในงานวิจัยนี้ ประกอบด้วย ข้อมูลจากเว็บประเภท เว็บท่า (Web Portal) เว็บเครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Web board) เว็บซื้อขายสินค้า (Web Shopping) หรือ เว็บบล็อก (Blog) ซึ่งรวบรวมมาจากข้อมูลข้อความภาษาไทย ที่นำมาใช้ในงานวิจัยนี้ ประกอบด้วย ข้อมูลจากเว็บประเภท เว็บท่า (Web Portal) เว็บเครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Web board) เว็บซื้อขายสินค้า (Web Shopping) หรือ เว็บบล็อก (Blog) ซึ่งรวบรวมมาจาก ข้อความคิดเห็นที่แสดงบนเว็บไซต์ 10 อับดับแรกที่ได้รับความนิยมของไทย ตัวอย่างเช่น ww.facebook.com, www.teenee.com, www.mthai.com, www.pantip.com, www.dek-d.com, www.siamphone.com, www.kapook.com, www.weloveshopping.com, ww.bloggang.com และ www.sanook.com. โดยจัดเก็บมาเป็นคลังข้อมูลขนาดใหญ่ รวมครอบคลุม 6 กลุ่มอารมณ์ อย่างละ 1,500 เอกสาร รวมจำนวนเอกสารในคลังข้อมูลทั้งสิ้น 9,000 เอกสาร มาแปลงเพื่อให้ได้เป็นเอกสารในรูปแบบเดียวกัน มีรูปแบบองค์ประกอบที่เหมือนกัน ซึ่งในที่นี้คือ ข้อความ (Plain Text) โดยสกัดแท็กทางโครงสร้างในการออกแบบภาษาเว็บทิ้งไป และจะแทนคุณลักษณะในเอกสาร เช่น ใช้ระดับคำเดี่ยว วลี ฯลฯ

เป็นตัวแทนคุณลักษณะของข้อความ และทำการคำนวณหาค่าน้ำหนัก งานวิจัยใช้การตัดคำแบบพจนานุกรมวิธีการเทียบคำที่ยาวที่สุด (Longest Matching)

## การกำจัดคำหยุด (Stop-word) และการหารากศัพท์ (Stemming)

การกำจัดคำหยุดเป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ไม่ทำให้ความหมายของข้อความเปลี่ยนแปลง คำที่ไม่มีนัยสำคัญในที่นี้หมายถึง คำที่ใช้กันโดยทั่วไปไม่มีความหมายสำคัญต่อเนื้อหา เมื่อตัดออกจากเอกสารแล้วไม่ทำให้ใจความของเนื้อหาเปลี่ยนแปลงไป ประเภทของคำที่จัดว่าเป็นคำหยุดมีดังต่อไปนี้ เช่น คำบุพบท เป็นคำที่ใช้เชื่อมคำหรือกลุ่มคำให้สัมพันธ์กัน คำสันธาน เป็นคำที่ทำหน้าที่เชื่อมคำกับคำ คำสรรพนาม เป็นคำที่ใช้แทนคำนามที่กล่าวถึงมาแล้วในประโยค เพื่อไม่ต้องกล่าวคำนามนั้นซ้ำอีก คำวิเศษณ์ เป็นคำที่ใช้ขยายคำอื่น เช่น คำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ เพื่อให้มีความหมายชัดเจนและมีรายละเอียดมากยิ่งขึ้น คำอุทาน เป็นคำที่แสดงอารมณ์ อาการ หรือความรู้สึกของผู้พูด คำหยุดมักเป็นคำที่ปรากฏขึ้นบ่อยครั้งในเนื้อหา และปรากฏในเนื้อหาเกือบทุกฉบับ จึงถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการค้นคืนหรือการจำแนกเอกสาร ดังนั้นการกำจัดคำหยุดจึงเป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี เพื่อกำจัดคุณลักษณะที่ไม่เป็นประโยชน์ และลดขนาดของดัชนีลง รวมถึงการหารากศัพท์ จึงเป็นการหารูปเดิมของคำ หรือหาคำที่มีความหมายคล้ายกัน เพื่อปรับรวมให้เป็นคำเดียวกัน การหารากศัพท์เป็นกระบวนการที่ควรทำก่อนการจัดทำดัชนี ทำให้สามารถลดขนาดของดัชนีลง และเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนก ซึ่งงานวิจัยนี้ใช้ฐานข้อมูลดิกชันนารีในการกำจัดคำหยุดและรากศัพท์

## การสร้างดัชนีของคำ (Indexing)

การคำนวณค่าดัชนีของคำในเอกสาร (Term Weighting) เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ของเครื่องจักรการเรียนรู้ วัตถุประสงค์ของการสร้างดัชนีคือ การคำนวณหาค่าที่จะมาใช้เป็นค่าคุณลักษณะของเอกสาร ปัจจุบันมีวิธีการคำนวณค่าน้ำหนักหลากหลายแบบ ซึ่งแต่ละวิธีก็เหมาะสมกับอัลกอริทึมที่แตกต่างกัน ซึ่งวิธีที่ได้รับความนิยมสูงที่สุดในงานวิจัยด้านการจำแนกเอกสาร (Document Categorization) และการค้นคืนสารสนเทศ (Information Retrieval) คือ วิธี TFIDF (Term frequency–inverse document frequency –weighting) แต่ในงานวิจัยนี้จะนำเสนอวิธีการให้น้ำหนักแบบใหม่ ที่เรียกว่า Term Occurrence Ratio Weighting (TOW) ซึ่งเป็นเทคนิคปรับปรุงดัชนีของคำแบบใหม่ มาทำการทดสอบเปรียบเทียบประสิทธิภาพในการจำแนกด้านความถูกต้อง โดยใช้วิธีการประเมิน

ความสามารถของแบบจำลอง โดยวัดที่ประสิทธิผลของการจำแนกเอกสารตามแนวคิดทางด้าน การค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) โดยเปรียบเทียบประสิทธิภาพกับแบบจำลองวิธีการคำนวณค่าน้ำหนักแบบมาตรฐานที่ใช้ในปัจจุบัน คือ TFIDF-weighting (Term frequency–inverse document frequency) และที่นิยมใช้เป็นค่าน้ำนักในอดีต Boolean-weighting (binary), TF-weighting (Term frequency) โดยนิยามของ Term Occurrence Ratio Weighting (TOW) ที่พัฒนาขึ้นใหม่นี้ จะเป็นการพิจารณาถึงการกระจายตัวของคำในแต่ละกลุ่ม โดยวัดอัตราส่วนระหว่างจำนวนเอกสารที่ปรากฏคำนั้นและอยู่ในกลุ่ม กับจำนวนเอกสารที่ปรากฏคำนั้นแต่ไม่อยู่ในกลุ่มนั้น ซึ่งอัตราส่วนดังกล่าวแสดงถึงความสัมพันธ์ระหว่างคำกับกลุ่มเอกสาร (Between Group) ว่ามีความสัมพันธ์กับกลุ่มอย่างเด่นชัดหรือไม่ หรืออาจกล่าวได้ว่าเป็นการให้ค่าน้ำหนักกับคำที่มีอำนาจจำแนกสูงที่ขึ้น สามารถบ่งบอกลักษณะของกลุ่มได้อย่างชัดเจนกว่าเดิม ผนวกกับการนิยามค่าคงที่ 2 ขึ้นมาใหม่ เพื่อเพิ่มอำนาจในการจำแนกเข้าไปในสมการทั้งสองข้าง ดังสมการ

$$TOW_{ik} = \log(2 \times tf_{ik}) \times \log\left( 2 + \frac{\begin{array}{c}the\,number\,of\,\,documents\\ include\,\,a\,term\,\,\,in\,class\end{array}}{\begin{array}{c}the\,number\,of\,\,documents\\ include\,\,a\,term\,\,in\,not\,class\end{array}} \right)$$

*The number of documents include a term in class* คือ คำที่ปรากฏในเอกสารเหล่านั้นและอยู่ในกลุ่ม

*The number of documents include a term in not class* คือ คำที่ปรากฏในเอกสารเหล่านั้นและไม่อยู่ในกลุ่ม

## การลดคุณลักษณะ (Feature Selection)

เนื่องจากเอกสารในปัจจุบันนั้น มีแนวโน้มในการที่จะเพิ่มปริมาณสูงขึ้นทุกวัน ทำให้เมื่อสกัดคุณลักษณะของเอกสารออกมามีจำนวนคุณลักษณะจำนวนมาก ทำให้การลดขนาดเอกสารด้วยการใช้การนำคำที่ไม่มีนัยสำคัญออกจากการทำรากศัพท์แล้วยังไม่เพียงพอ ซึ่งจำนวนคุณลักษณะมีผลต่อประสิทธิภาพของการจำแนกหมวดหมู่เว็บเพจ เนื่องจากอัลกอริทึมที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกหมวดหมู่โดยทั่วไปไม่สามารถรองรับ การทำงานกับจำนวนคุณลักษณะของเอกสารที่สูงมากได้ดี และเอกสารที่มีจำนวนคุณลักษณะมากอาจก่อให้เกิดปัญหา Overfitting ซึ่งเป็น

ปรากฏการณ์ที่ตัวจำแนกหมวดหมู่ค้นพบลักษณะโดยบังเอิญของเอกสารตัวอย่าง แทนที่จะค้นพบลักษณะพื้นฐานที่จำเป็นของเว็บเพจตัวอย่าง ทำให้ตัวจำแนกหมวดหมู่ทำงานผิดพลาด ซึ่งการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม อาจจะนำคุณลักษณะพื้นฐานเหล่านี้มารวมกันเพื่อให้เป็นคุณลักษณะใหม่ที่มีระดับสูงขึ้น ซึ่งงานวิจัยนี้ได้ใช้ค่าสถิติไคสแควร์ มาวัดความเป็นอิสระต่อกันระหว่างตัวแปรคุณลักษณะกับกลุ่มของเอกสาร โดยค่าสถิติไคสแควร์ที่ใช้ในงานวิจัยนี้จะพิจารณาจากค่าสูงสุดลดหลั่นลดลง สามารถนิยามได้โดย

$$\chi^2(w,c_j)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)}$$

A คือ จำนวนเอกสารจากกลุ่ม $C_j$ ซึ่งมีคำ $w$

B คือ จำนวนเอกสารซึ่งมีคำ $w$ แต่ไม่ได้อยู่ในกลุ่ม $C_j$

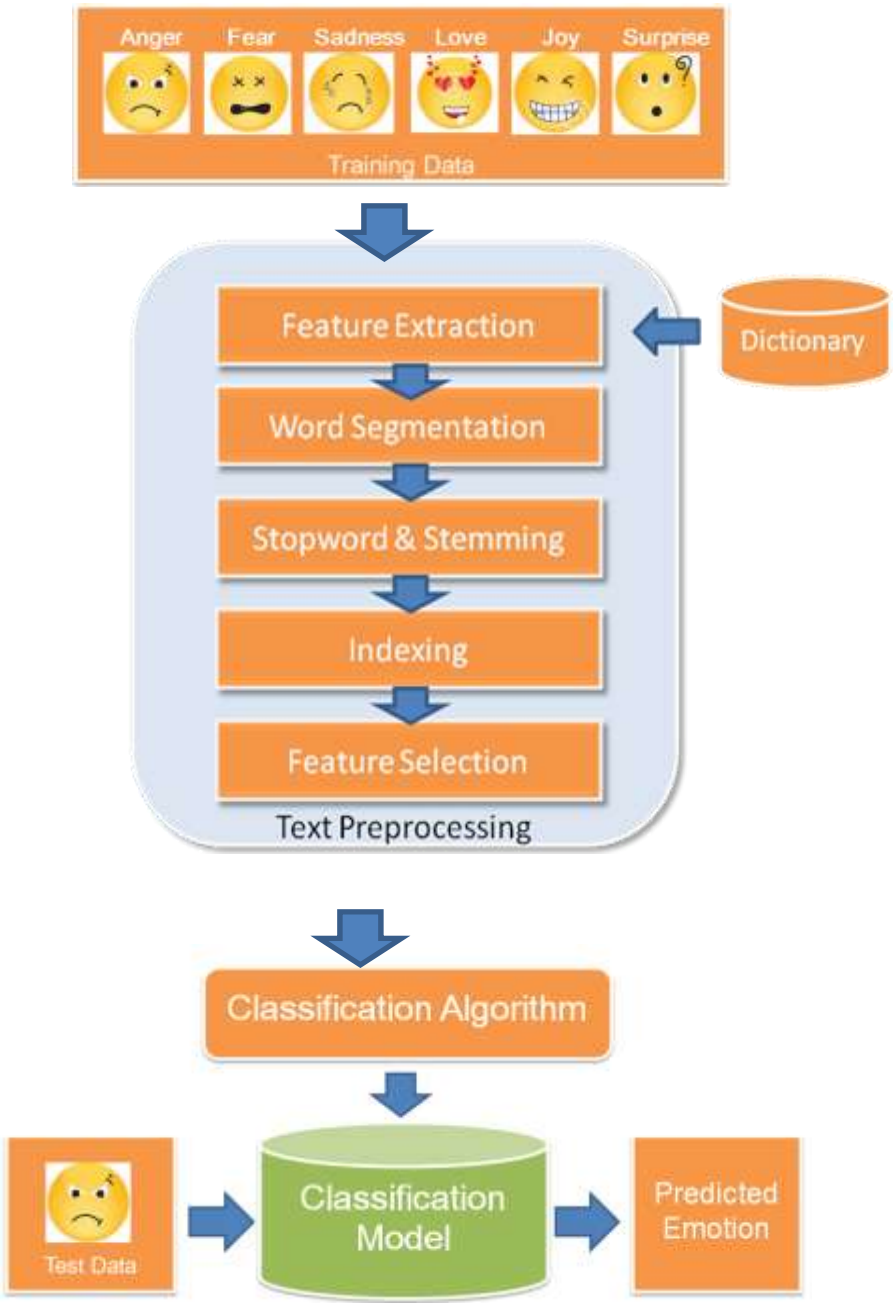
C คือ จำนวนเอกสารจากกลุ่ม $C_j$ ซึ่งไม่มีคำ $w$

D คือ จำนวนเอกสารที่ไม่ได้อยู่ในกลุ่ม $C_j$ หรือไม่มีคำ $w$

N คือ จำนวนกลุ่มเอกสารทั้งหมด

## ขั้นตอนวิธีการสร้างแบบจำลองการจำแนกอารมณ์

จากการศึกษาและวิเคราะห์ปัญหางานวิจัยที่เกี่ยวข้องกับการจำแนกเอกสารดังกล่าว ในส่วนนี้ผู้วิจัยขอนำเสนอขั้นตอนวิธีการสร้างแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่องโดยใช้เทคนิคปรับปรุงดัชนีของคำ เพื่อการจำแนกความรู้สึกทางอารมณ์จากฐานข้อมูลข้อความภาษาไทย เพื่อเพิ่มประสิทธิภาพในการค้นคืนหรือการจำแนกอารมณ์ โดยแบบจำลองที่นำเสนอในงานวิจัยนี้ ใช้คุณลักษณะแบบคำเดียว (Single Word) ที่ได้จากสกัดคุณลักษณะและทำการกำจัดคำหยุดและทำรากศัพท์ หลังจากนั้นทำการให้ค่าน้ำหนักของคำในเอกสาร (Indexing) ด้วยวิธี Term Occurrence Ratio Weighting (TOW) ซึ่งเป็นเทคนิคปรับปรุงดัชนีของคำแบบใหม่ ที่นำเสนอไปดังกล่าว จากนั้นทำการลดขนาดคุณลักษณะของเอกสารลงโดยพิจารณาคุณลักษณะที่มีค่าไคสแควร์สูง ส่งเข้าเครื่องจักรการเรียนรู้แบบมีผลเฉลยด้วยอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน ( Support Vector Machine ) ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) โดยอัลกอริทึมทั้งหมดใช้ค่าพารามิเตอร์มาตรฐานมาทำการเรียนรู้ แล้วทำการทดสอบเปรียบเทียบประสิทธิภาพด้านความถูกต้อง (Accuracy) โดยทำการทดสอบด้วยวิธี 10-ครอสวาลิเดชั่น (10-Cross Validation)




ภาพประกอบ 15 แบบจำลองการจำแนกอารมณ์

## การประเมินผลการทดลอง

การประเมินความสามารถของแบบจำลองแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่องนั้น เน้นความสามารถในการตัดสินใจหรือ การจำแนกอารมณ์ที่ถูกต้อง วิธีการทดสอบเพื่อเปรียบเทียบประสิทธิภาพของโมเดล สามารถพิจารณา

ได้จากการจำแนกข้อมูลตามแนวคิดทางด้านการค้นคืนสารสนเทศ ซึ่งก็คือการวัดค่าความถูกต้อง (Accuracy) ของแบบจำลอง

# บทที่ 4

# ผลการทดลอง

งานวิจัยนี้ มีวัตถุประสงค์เพื่อพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง มุ่งเน้นความถูกต้องการการจำแนกอารมณ์จากคลังข้อความภาษาไทย โดยที่จำลองนี้ใช้ทรัพยากรของระบบและหน่วยความจำอย่างเหมาะสม ทำการทดสอบกับกลุ่มคลังข้อความภาษาไทย ซึ่งข้อมูลทดสอบที่นำมาใช้ในงานวิจัยนี้ประกอบด้วย ข้อมูลจากเว็บประเภท เว็บท่า (Web Portal) เว็บเครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Web board) เว็บซื้อขายสินค้า (Web Shopping) ซึ่งรวบรวมมาจากความคิดเห็นที่แสดงบนเว็บไซต์ ที่ได้รับความนิยมของไทย โดยทำการทดลองการลดคุณลักษณะด้วยวิธีค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเทคนิคการเรียนรู้ของเครื่อง (Machine Learning Technique) ซึ่งประกอบด้วยซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เนอีฟเบย์ (Naïve-Bayes) ต้นไม้ตัดสินใจ (Decision Tree) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) สามารถสรุปผลการทดลอง ดังนี้

## การวิเคราะห์กลุ่มตัวอย่าง

งานวิจัยนี้ทดสอบกับคลังข้อมูลข้อความภาษาไทย ซึ่งประกอบด้วย ข้อมูลจากเว็บประเภท เว็บท่า (Web Portal) เว็บเครือข่ายสังคมออนไลน์ (Social Media) เว็บบอร์ด (Web board) เว็บซื้อขายสินค้า (Web Shopping) หรือ เว็บบล็อก (Blog) ซึ่งรวบรวมมาจากความคิดเห็นที่แสดงบนเว็บไซต์ที่ได้รับความนิยมของไทย ตัวอย่างเช่น www.facebook.com, www.teenee.com, www.mthai.com, www.pantip.com, www.dek-d.com, www.kapook.com www.siamphone.com, www.weloveshopping.com, ww.bloggang.com และ www.sanook.com โดยจัดเก็บมาเป็นคลังข้อมูลขนาดใหญ่ รวมครอบคลุม 6 กลุ่มอารมณ์ อย่างละ 1,500 เอกสาร รวมจำนวนเอกสารในคลังข้อมูลทั้งสิ้น 9,000 เอกสาร ในงานวิจัยนี้จะจำแนกผลลัพธ์ของความคิดเห็นออกมาเป็นกลุ่มประเภทอารมณ์ 6 แบบ ซึ่งประกอบด้วย ความโกรธ (Anger) ความกลัว (Fear) ความเศร้า (Sadness) ความรัก (Love) ความสนุก (Joy) และความประหลาดใจ (Surprise) เป็นกลุ่มตัวอย่างเรียนรู้ และ ทำการทดสอบด้วยวิธี 10-fold cross validation

ตารางที่ 1 จำนวนกลุ่มตัวอย่างคลังข้อมูลอารมณ์

| ลำดับ | ชื่อกลุ่ม | จำนวนเอกสาร |
|---|---|---|
| 1 | ความโกรธ (Anger) | 1,500 |
| 2 | ความกลัว (Fear) | 1,500 |
| 3 | ความเศร้า (Sadness) | 1,500 |
| 4 | ความรัก (Love) | 1,500 |
| 5 | ความสนุก (Joy) | 1,500 |
| 6 | ความประหลาดใจ (Surprise) | 1,500 |
| | รวม | 9,000 |

## ผลการทดลองจำแนกคลังข้อมูลอารมณ์

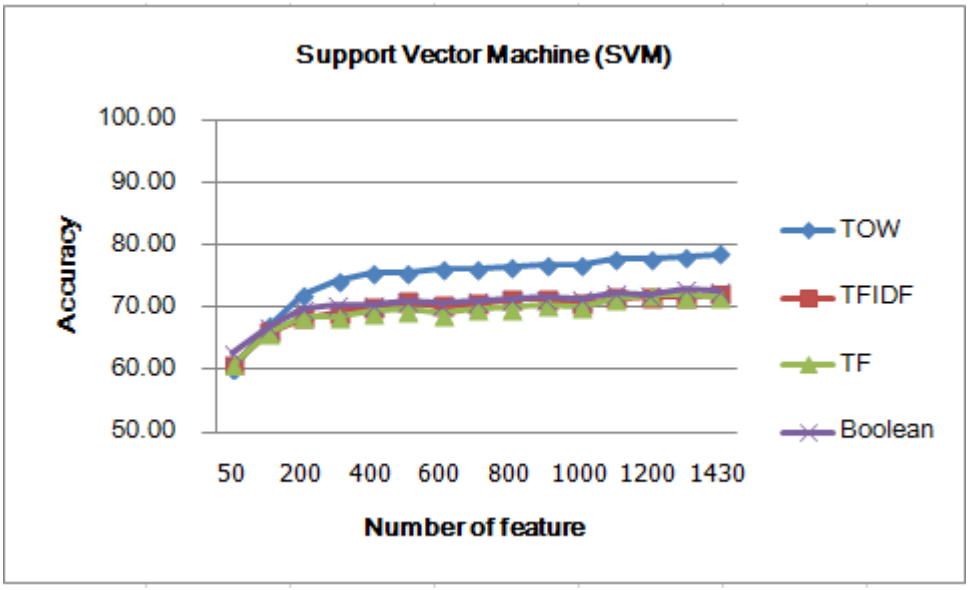
เมื่อลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) และเรียนรู้ด้วย ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เนอีฟเบย์ (Naïve-Bayes) ต้นไม้ตัดสินใจ (Decision Tree) เคเนียเรสเนเบอร์ (K-Nearest Neighbor)

### 4.1 ผลการทดลองเมื่อจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน

จากการทดลองให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด ด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึม Support Vector Machine พบว่าวิธี TOW Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 75.58% รองลงมาเป็นวิธี Boolean Weighting ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 72.91% วิธี TFIDF Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 72.04% และสุดท้าย วิธี TF Weighting ที่จำนวน 1200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 71.95% ตามลำดับ

ตารางที่ 2 ผลการทดลอง เมื่อเรียนรู้ด้วยซัพพอร์ตเวกเตอร์แมชชีน

| | SVM | | | |
|---|---|---|---|---|
| **Feature** | **TOW** | **TFIDF** | **TF** | **Boolean** |
| 50 | 60.12 | 60.79 | 60.91 | 62.70 |
| 100 | 67.04 | 66.04 | 66.00 | 66.87 |
| 200 | 72.04 | 68.25 | 68.50 | 69.87 |
| 300 | 74.16 | 69.00 | 68.62 | 70.41 |
| 400 | 75.45 | 70.00 | 69.33 | 70.54 |
| 500 | 75.58 | 70.95 | 69.70 | 71.12 |
| 600 | 76.08 | 70.16 | 69.04 | 70.58 |
| 700 | 76.20 | 70.54 | 69.79 | 71.00 |
| 800 | 76.50 | 71.37 | 70.00 | 71.37 |
| 900 | 76.83 | 71.37 | 70.45 | 71.54 |
| 1000 | 76.79 | 70.54 | 70.20 | 71.41 |
| 1100 | 77.91 | 71.41 | 71.33 | 72.41 |
| 1200 | 77.87 | 71.70 | 71.95 | 72.16 |
| 1300 | 78.04 | 71.62 | 71.87 | 72.91 |
| 1430 | 78.58 | 72.04 | 71.79 | 72.75 |



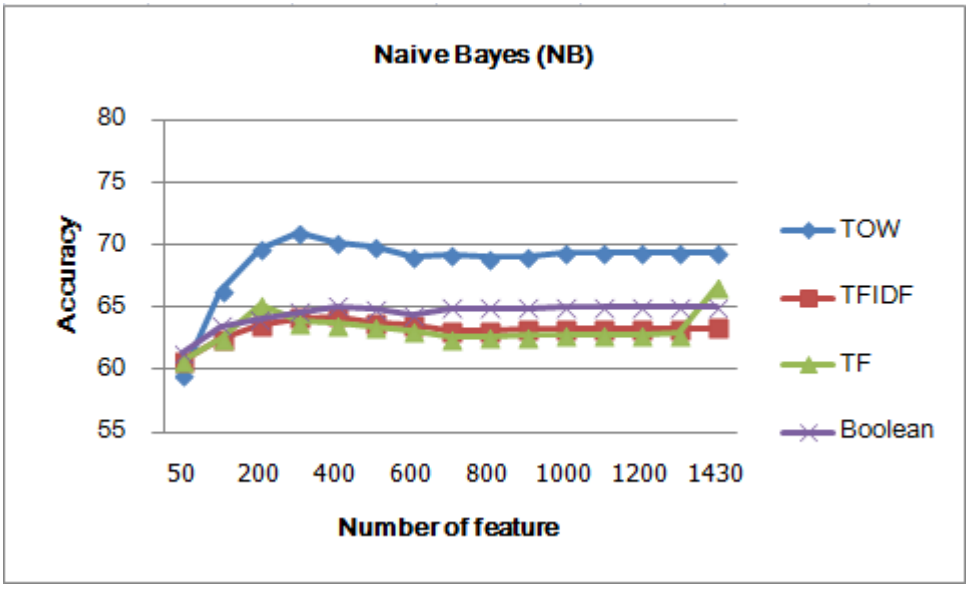

ภาพประกอบ 16 กราฟเปรียบเทียบประสิทธิภาพอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน

**4.2 ผลการทดลองเมื่อจำแนกข้อมูลด้วยเนอีฟเบย์**

จากการทดลองให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด ด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมเนอีฟเบย์ พบว่าวิธี TOW Weighting ที่จำนวน 300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 70.95% รองลงมาเป็น วิธี TF Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 66.62% วิธี Boolean Weighting ที่จำนวน 400 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 65.08% และสุดท้ายวิธี TFIDF Weighting ที่จำนวน 400 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 64.29% ตามลำดับ

ตารางที่ 3 ผลการทดลองเมื่อเรียนรู้ด้วยเนอีฟเบย์

| | NB | | | |
|---|---|---|---|---|
| **Feature** | **TOW** | **TFIDF** | **TF** | **Boolean** |
| 50 | 59.58 | 60.66 | 60.75 | 61.33 |
| 100 | 66.41 | 62.41 | 62.45 | 63.50 |
| 200 | 69.70 | 63.62 | 65.25 | 64.04 |
| 300 | 70.95 | 64.16 | 63.87 | 64.58 |
| 400 | 70.25 | 64.29 | 63.70 | 65.08 |
| 500 | 69.91 | 63.70 | 63.45 | 64.83 |
| 600 | 69.04 | 63.62 | 63.12 | 64.37 |
| 700 | 69.25 | 63.12 | 62.58 | 64.91 |
| 800 | 69.00 | 63.12 | 62.70 | 64.91 |
| 900 | 69.08 | 63.20 | 62.75 | 64.95 |
| 1000 | 69.41 | 63.20 | 62.87 | 65.00 |
| 1100 | 69.41 | 63.20 | 62.87 | 65.00 |
| 1200 | 69.41 | 63.25 | 62.87 | 65.00 |
| 1300 | 69.41 | 63.29 | 62.91 | 65.00 |
| 1430 | 69.41 | 63.33 | 66.62 | 65.00 |




ภาพประกอบ 17 กราฟเปรียบเทียบประสิทธิภาพของอัลกอริทึมเนอีฟเบย์

**4.3 ผลการทดลองเมื่อจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ**

จากการทดลองให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด ด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมต้นไม้ตัดสินใจ พบว่าวิธี TOW Weighting ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 82.60% รองลงมาเป็น วิธี TFIDF Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 63.00% วิธี Boolean Weighting ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 62.58% และสุดท้ายวิธี TF Weighting ที่จำนวน 500 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 62.29% ตามลำดับ

# บทที่ 5

# สรุปผล อภิปรายผลและข้อเสนอแนะ

การนำเสนอในบทนี้ เป็นการรวบรวมสาระสำคัญจากการพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง ที่มีประสิทธิภาพทั้งในด้านความถูกต้องในการจำแนกประเภท และความถูกต้องในภาพรวมด้านการค้นคืนสารสนเทศ ที่ใช้ทรัพยากรของระบบ และหน่วยความจำอย่างเหมาะสม แต่สามารถจำแนกประเภทได้อย่างมีประสิทธิภาพและประสิทธิผล ซึ่งสามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการวิเคราะห์อารมณ์ของลูกค้าที่ซื้อสินค้าออนไลน์แบบอัตโนมัติ รวมทั้งสรุป การอภิปรายผลและข้อเสนอแนะ ตลอดจนแนวทางในการพัฒนางานวิจัย

## สรุปผลการวิจัย

งานวิจัยนี้ได้นำเสนอวิธีการคำนวณค่าน้ำหนักดัชนีแบบใหม่ให้กับคำ ในกลุ่มโดเมนตัวอย่างการแสดงความรู้สึกผ่านข้อความในเอกสารออนไลน์บนเว็บโซเชียลมีเดีย เว็บขายสินค้าออนไลน์ต่างๆ ซึ่งสามารถแปลผลถึงอารมณ์ของกลุ่มเป้าหมายนั้นๆ การค้นหาข้อมูลความคิดเห็นเพื่อการบริหารจัดการลูกค้าสัมพันธ์ (CRM: Customer Relationship Management) โดยองค์กรหรือหน่วยงานธุรกิจสามารถดึงข้อมูลหลากหลายบนเครือข่ายสังคมมาใช้เพื่อประเมินความพึงพอใจของลูกค้า (Customer Satisfaction) ต่อสินค้าและการให้บริการ ทั้งนี้ ข้อความที่ถ่ายทอดถึงอารมณ์และความรู้สึกของลูกค้าเหล่านี้ ถือเป็นข้อมูลเชิงจิตวิทยา (Psychological Data) ที่สามารถนำมาใช้วิเคราะห์พฤติกรรมผู้บริโภคและพฤติกรรมทางสังคม เพื่อปรับปรุง พัฒนา สินค้าและบริการของบริษัทให้มีประสิทธิภาพสูงขึ้น ตรงกับความต้องการของผู้บริโภคมากขึ้น

โดยงานวิจัยนี้มุ่งเน้นที่จะพัฒนาประสิทธิภาพของแบบจำลองการจำแนกอารมณ์ (Emotion Classification) ในข้อความภาษาไทย โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่อง ซึ่งเป็นการคิดค้นวิธีการให้ค่าน้ำหนักของคำแบบใหม่ ที่วิเคราะห์จากความน่าจะเป็นการกระจายตัวของคำในฐานข้อมูลที่มีประสิทธิภาพ โดยทำการสกัดคุณลักษณะของคำภาษาไทยโดยใช้วิธีการการตัดคำแบบพจนานุกรมข้อมูล ตลอดจนให้ค่าดัชนีที่สกัดได้ด้วยวิธีTerm Occurrence Ratio Weighting (TOW) ที่พัฒนาขึ้น และทำการลดคุณลักษณะที่สกัดได้จำนวนมากนั้นด้วยวิธีค่าสถิติไคสแควร์มาวัดความเป็นอิสระต่อกันระหว่างตัวแปรคุณลักษณะกับกลุ่มของเอกสาร ก่อน

นำไปประมวลผลด้วยเครื่องจักรการเรียนรู้ ซึ่งประกอบด้วยอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine ) ต้นไม้ตัดสินใจ (Decision Tree) เนอีฟเบย์ (Naïve-Bayes) และเคเนียเรสเนเบอร์ (K-Nearest Neighbor) โดยมีข้อสรุปและข้อค้นพบที่ได้จากการวิจัยดังนี้

1. เมื่อจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน จากผลการทดลองพบว่า การให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เมื่อเปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด ด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึม Support Vector Machine พบว่าวิธี TOW Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 75.58% รองลงมาเป็นวิธี Boolean Weighting ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 72.91% วิธี TFIDF Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 72.04% และสุดท้าย วิธี TF Weighting ที่จำนวน 1200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 71.95% ตามลำดับ

2. เมื่อจำแนกข้อมูลด้วยเนอีฟเบย์ จากผลการทดลองพบว่า การให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุดด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมเนอีฟเบย์ พบว่าวิธี TOW Weighting ที่จำนวน 300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 70.95% รองลงมาเป็น วิธี TF Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 66.62% วิธี Boolean Weighting ที่จำนวน 400 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 65.08% และสุดท้ายวิธี TFIDF Weighting ที่จำนวน 400 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 64.29% ตามลำดับ

3. เมื่อจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ จากผลการทดลองพบว่า การให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW

Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุดด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมต้นไม้ตัดสินใจ พบว่าวิธี TOW Weighting ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 82.60% รองลงมาเป็น วิธี TFIDF Weighting ที่จำนวน 1430 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 63.00% วิธี Boolean Weighting ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 62.58% และสุดท้ายวิธี TF Weighting ที่จำนวน 500 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 62.29% ตามลำดับ

4. เมื่อจำแนกข้อมูลด้วยเคเนียเรสเนเบอร์ (K-Nearest Neighbor) จากผลการทดลองพบว่าการให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึม เคเนียเรสเนเบอร์ (K-Nearest Neighbor) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด ด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมเคเนียเรสเนเบอร์ พบว่าวิธี TOW Weighting ที่จำนวน 300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 68.25% รองลงมาเป็น วิธี TF Weighting ที่จำนวน 200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 66.95% วิธี TFIDF Weighting ที่จำนวน 200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 66.12% และสุดท้ายวิธี Boolean Weighting ที่จำนวน 200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 64.62% ตามลำดับ

5. เมื่อเปรียบเทียบประสิทธิภาพในการจำแนกข้อมูลจากอัลกอริทึมทั้งหมด จากผลการทดลองพบว่าการสร้างดัชนีด้วยวิธี TOW Weighting และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) สามารถสรุปได้ว่าอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ประสิทธิภาพด้านการจำแนกที่ดีสุด รองลงมาเป็น อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) อัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) และอัลกอริทึมเคเนียเรสเนเบอร์ ตามลำดับ เมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุดในทุกอัลกอริทึม พบว่าวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 82.60%

6. เมื่อพิจารณาด้านการลดคุณลักษณะ โดยเปรียบเทียบด้านการลดคุณลักษณะและวิธีการให้ค่าน้ำหนัก TOW Weighting โดยวัดค่าประสิทธิภาพด้วยค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า เมื่อลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) พบว่าทุกอัลกอริทึม สามารถลดคุณลักษณะลงได้เป็นจำนวนมาก แต่ยังคงให้ประสิทธิภาพในการจำแนกเอกสารที่แม่นยำ ซึ่งสามารถสรุปได้ว่า การลดคุณลักษณะด้วยวิธีค่าดังกล่าว สามารถลดคุณลักษณะได้เป็นจำนวนมาก แต่ไม่ส่งผลให้ประสิทธิภาพในการจำแนกลดลง

7. จากสรุปผลการทดลองดังกล่าว พบว่า วิธีการให้ค่าน้ำหนัก TOW Weighting ที่พัฒนาขึ้นใหม่นั้น สามารถสร้างอำนาจจำแนกให้กับคุณลักษณะที่สกัดได้ เพื่อใช้ในการเรียนรู้เพื่อสร้างแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคความน่าจะเป็นของคำร่วมกับการเรียนรู้ของเครื่องได้อย่างมีประสิทธิภาพและประสิทธิผล อันเนื่องมาจากลักษณะการให้ค่าถ่วงน้ำหนักของคุณลักษณะที่สกัดได้ด้วยวิธีอัตราส่วนดังกล่าวนั้น พิจารณาจากพฤติกรรมการกระจายตัวของคำที่ปรากฏในแต่ละกลุ่มเอกสารที่นำมาเรียนรู้ โดยนำมาสร้างเป็นตัวแปรที่เก็บอัตราส่วนของคำที่ปรากฏในกลุ่มและระหว่างกลุ่มเอกสาร ซึ่งเป็นตัวแปรดังกล่าวเป็นปัจจัยสำคัญที่สามารถแก้ปัญหาด้านความกำกวมที่พบบ่อยในการจำแนกเอกสารออนไลน์ รวมถึงสะท้อนอำนาจจำแนกที่แท้จริงของคำที่สกัดได้จากเอกสาร ซึ่งปัจจัยดังกล่าวนั้นส่งผลให้ค่าน้ำหนัก TOW Weighting ที่พัฒนาขึ้นมีประสิทธิภาพในการจำแนกมากกว่าวิธีคำนวณค่าน้ำหนักให้กับดัชนีแบบอื่น ๆ อย่างเด่นชัด

## อภิปรายผลการวิจัย

จากข้อสรุปผลการวิจัยพบว่า ปัจจัยหลักที่ส่งผลให้ประสิทธิภาพในการจำแนกประเภทด้วยวิธีการให้ค่าน้ำหนัก TOW Weighting มีประสิทธิภาพสูงนั้น เกิดจากปัจจัยหลักด้านพฤติกรรมการกระจายตัวของคำในเอกสารที่นำมาเรียนรู้ ซึ่งวิธีการให้ค่าน้ำหนัก TOW Weighting พัฒนามาจากการปรับปรุงจากวิธีการคำนวณค่าน้ำหนักให้กับดัชนีแบบเดิมคือ สมการ $tfidf_{ik}$ ซึ่งเป็นวิธีการคำนวณค่าน้ำหนักที่เป็นมาตรฐานและได้รับความนิยม ในการนำไปประยุกต์ใช้ในการจำแนกเอกสารจำนวนมาก การคำนวณค่าน้ำหนักให้กับดัชนีนั้นมีหลายวิธี แต่โดยส่วนใหญ่จะคำนวณบนพื้นฐานที่ว่า ถ้ายิ่งมีการเกิดขึ้นของคำนั้นในเอกสารมากเท่าไหร่ คำนั้นก็ยิ่งน่าจะมีความเกี่ยวข้องกับการจัดกลุ่มของเอกสารนั้นมากยิ่งขึ้น และถ้าคำใดปรากฏบนทุกเอกสารมากเท่าใด คำนั้นน่าจะไม่มีผลต่อการจัดกลุ่ม

จากหลักการดังกล่าว ผู้วิจัยจึงได้ค้นหาข้อบกพร่องและทำการพัฒนาปรับปรุงดัชนี เพื่อเพิ่มประสิทธิภาพในการจำแนกเอกสารมากขึ้น โดยพัฒนาการให้น้ำหนักกับดัชนีใหม่ที่มีชื่อว่า TOW Weighting โดยคำนึงถึงการกระจายตัวของคำที่พบในกลุ่มตัวอย่างที่พบมาสร้างเป็น

อัตราส่วนของคำแล้วแทนที่ดัชนีด้วยวิธีการดังกล่าว ซึ่งวิธีที่พัฒนาขึ้นใหม่นี้สามารถแก้ปัญหาดั้งเดิมในการจำแนกเอกสาร ซึ่งประกอบปัจจัยด้วยด้านความถี่ของคำที่ปรากฏในเอกสาร และปัจจัยด้านความถี่เอกสารผกผัน ส่งผลให้ประสิทธิภาพในการจำแนกดีกว่าทุกวิธีที่นำมาเปรียบเทียบอย่างชัดเจน ในขณะที่ใช้จำนวนคุณลักษณะที่ใช้ในการเรียนรู้เพื่อสร้างตัวจำแนกน้อยกว่า ส่งผลให้แบบจำลองที่พัฒนาขึ้น ใช้หน่วยความจำและทรัพยากรในการประมวลผลลดลงกว่าแบบเดิม โดยแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่องที่พัฒนาขึ้นนี้ สามารถนำไปประยุกต์ใช้กับการพัฒนาระบบงานการวิเคราะห์อารมณ์ของกลุ่มลูกค้าที่ซื้อสินค้าออนไลน์แบบอัตโนมัติ ส่งผลให้องค์กรหรือหน่วยงานธุรกิจสามารถดึงข้อมูลหลากหลายบนเครือข่ายสังคมมาใช้ เพื่อประเมินความพึงพอใจของลูกค้า (Customer Satisfaction) ต่อสินค้าและการให้บริการ ทั้งนี้เพราะข้อความที่ถ่ายทอดถึงอารมณ์และความรู้สึกของลูกค้าเหล่านี้ถือเป็นข้อมูลเชิงจิตวิทยา (Psychological Data) ที่สามารถนำมาใช้วิเคราะห์พฤติกรรมผู้บริโภคและพฤติกรรมทางสังคม เพื่อปรับปรุง พัฒนา สินค้าและบริการของบริษัทให้มีประสิทธิภาพสูงขึ้น ตรงกับความต้องการของผู้บริโภคมากขึ้น อันนำมาซึ่งผลกำไรของบริษัทในระยะยาว

## ข้อเสนอแนะ

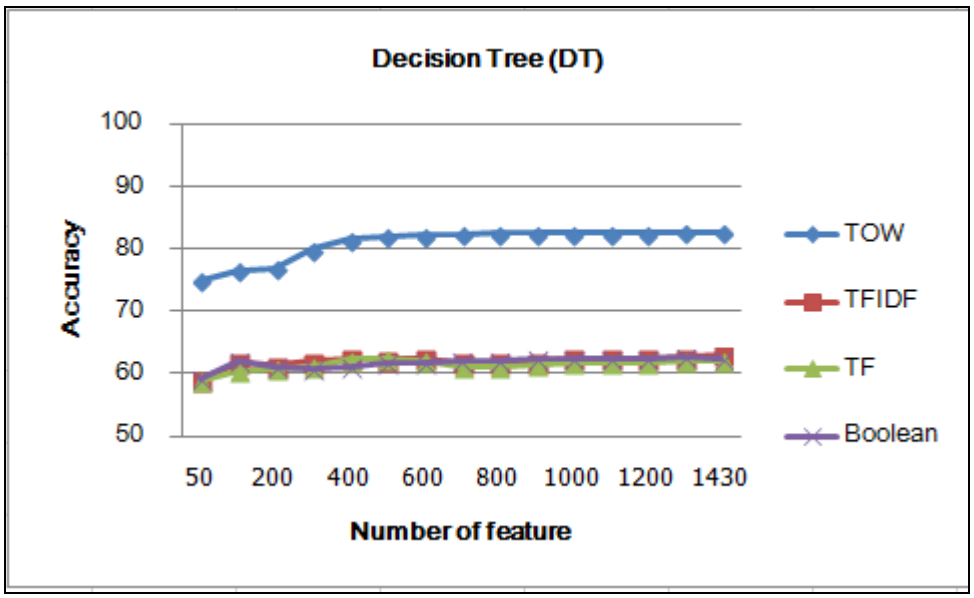
จากข้อค้นพบในการพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่องที่ได้จากงานวิจัยนี้นั้น ผู้วิจัยของเสนอแนะเพื่อพัฒนาปรับปรุงประสิทธิภาพของงานวิจัยดังต่อไปนี้

1. งานวิจัยการพัฒนาประสิทธิภาพแบบจำลองการจำแนกอารมณ์โดยใช้เทคนิคปรับปรุงดัชนีของคำร่วมกับการเรียนรู้ของเครื่องนี้ มุ่งเน้นเพื่อพัฒนาวิธีการคำนวณค่าดัชนีแบบใหม่ เพื่อพัฒนาประสิทธิภาพในการการจำแนกประเภท โดยใช้อัลกอริทึมจำแนกประเภทพื้นฐานซึ่งประกอบด้วย อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) อัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) และอัลกอริทึมเคเนียเรสเนเบอร์ (K-Nearest Neighbor) แต่มิได้ทดสอบกับอัลกอริทึมประเภทอื่น ตลอดจนการปรับแต่ง แก้ไขพามิเตอร์ของอัลกอริทึม ดังนั้นจึงควรมีการศึกษาวิธีการปรับแต่งอัลกอริทึมและพารามิเตอร์ต่างๆ เพื่อเพิ่มประสิทธิภาพในการจำแนกอารมณ์ให้แม่นยำมากขึ้น และงานวิจัยนี้ทำการทดสอบการลดคุณลักษณะด้วยวิธีการลดคุณลักษณะด้วยวิธีค่าสถิติไคสแควร์ (Chi – Square) ซึ่งส่งผลให้จำนวนคุณลักษณะลดลงอย่างมาก เมื่อส่งเข้าเครื่องจักรการเรียนรู้ ซึ่งการลดคุณลักษณะแบบนี้เป็นแบบ Filtering แต่ยังไม่ได้ทดลองกับการลดคุณลักษณะแบบ Wrapper ดังนั้นจึงควรมีการศึกษาการลดคุณลักษณะแบบ Wrapper รวมถึงการผสานการลดคุณลักษณะทั้ง 2 แบบเข้าด้วยกัน

2. งานวิจัยนี้มุ่งเน้นการวัดประสิทธิภาพด้านความถูกต้อง (Accuracy) ในการจำแนกเอกสารเป็นหลัก โดยคำนึงถึงระยะเวลาที่ใช้ในการประมวลผลเพื่อสร้างดัชนีเป็นปัจจัยรอง ดังนั้นจึงควรมีการศึกษาด้านการปรับปรุงประสิทธิภาพด้านความเร็ว และการบริหารหน่วยความจำที่ใช้ในการประมวลผลเพื่อสร้างดัชนี เพื่อให้เกิดประสิทธิภาพสูงสุด ตลอดจนการจำแนกประเภทแบบทันเวลา (Real Time)

ตารางที่ 4 ผลการทดลองเมื่อเรียนรู้ด้วยต้นไม้ตัดสินใจ

| | DT | | | |
|---|---|---|---|---|
| **Feature** | **TOW** | **TFIDF** | **TF** | **Boolean** |
| 50 | 75.00 | 58.91 | 58.87 | 59.08 |
| 100 | 76.50 | 61.87 | 60.50 | 62.16 |
| 200 | 76.90 | 61.29 | 60.87 | 61.12 |
| 300 | 80.00 | 61.95 | 61.16 | 60.87 |
| 400 | 81.50 | 62.37 | 62.25 | 61.04 |
| 500 | 82.00 | 62.33 | 62.29 | 61.83 |
| 600 | 82.20 | 62.54 | 62.20 | 61.70 |
| 700 | 82.30 | 61.79 | 61.12 | 62.00 |
| 800 | 82.50 | 61.83 | 61.20 | 62.08 |
| 900 | 82.50 | 62.00 | 61.50 | 62.25 |
| 1000 | 82.50 | 62.50 | 61.66 | 62.50 |
| 1100 | 82.50 | 62.50 | 61.83 | 62.50 |
| 1200 | 82.50 | 62.50 | 61.79 | 62.50 |
| 1300 | 82.60 | 62.62 | 62.10 | 62.58 |
| 1430 | 82.60 | 63.00 | 62.20 | 62.41 |



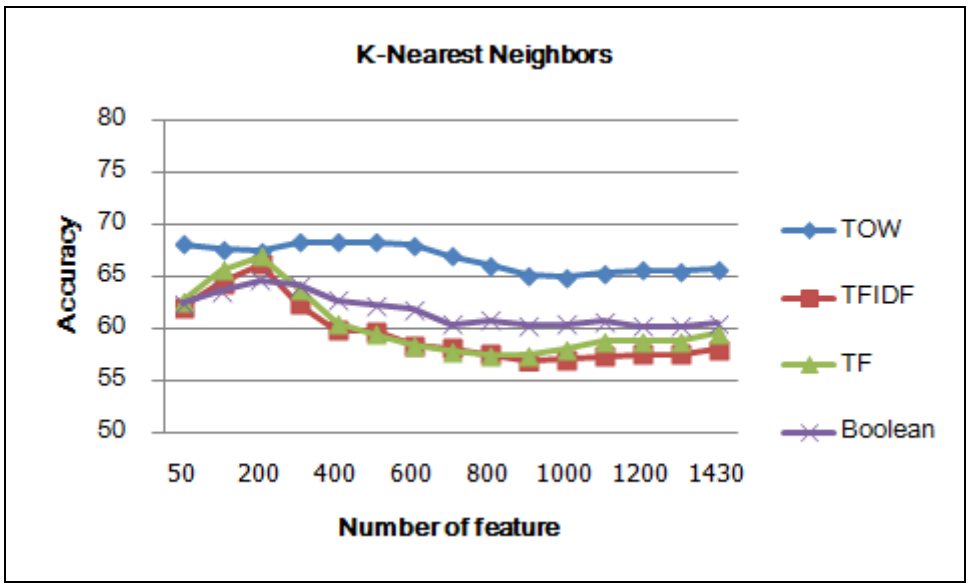

ภาพประกอบ 18 กราฟเปรียบเทียบประสิทธิภาพของอัลกอริทึมต้นไม้ตัดสินใจ

**4.4 ผลการทดลองเมื่อจำแนกข้อมูลด้วยเคเนียเรสเนเบอร์**

จากการทดลองให้ค่าน้ำหนักกับดัชนีแบบต่างๆ เปรียบเทียบกับค่าน้ำหนัก TOW และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) ร่วมกับเรียนรู้ด้วยอัลกอริทึม เคเนียเรสเนเบอร์ (K-Nearest Neighbor) ทำการวัดประสิทธิภาพจากค่าความถูกต้อง (Accuracy) สามารถสรุปได้ว่า การคำนวณค่าน้ำหนักเพื่อสร้างดัชนีด้วยวิธี TOW Weighting ให้ประสิทธิภาพการจำแนกอารมณ์จากคลังข้อมูลภาษาไทยออกมาดีที่สุดเมื่อปเรียบเทียบกับวิธีการอื่นๆ และเมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด ด้วยวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมเคเนียเรสเนเบอร์ พบว่าวิธี TOW Weighting ที่จำนวน 300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 68.25% รองลงมาเป็น วิธี TF Weighting ที่จำนวน 200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 66.95% วิธี TFIDF Weighting ที่จำนวน 200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 66.12% และสุดท้ายวิธี Boolean Weighting ที่จำนวน 200 คุณลักษณะ ให้ประสิทธิภาพเท่ากับ 64.62% ตามลำดับ

ตารางที่ 5 ผลการทดลองเมื่อเรียนรู้ด้วยเคเนียเรสเนเบอร์

| | KNN | | | |
|---|---|---|---|---|
| Feature | TOW | TFIDF | TF | Boolean |
| 50 | 68.08 | 62.04 | 62.62 | 62.37 |
| 100 | 67.62 | 64.37 | 65.62 | 63.62 |
| 200 | 67.41 | 66.12 | 66.95 | 64.62 |
| 300 | 68.25 | 62.29 | 63.79 | 64.20 |
| 400 | 68.25 | 59.83 | 60.58 | 62.70 |
| 500 | 68.25 | 59.66 | 59.45 | 62.25 |
| 600 | 67.95 | 58.37 | 58.33 | 61.87 |
| 700 | 66.91 | 58.04 | 57.79 | 60.41 |
| 800 | 66.12 | 57.50 | 57.50 | 60.79 |
| 900 | 65.08 | 56.83 | 57.37 | 60.29 |
| 1000 | 64.87 | 57.04 | 58.08 | 60.41 |
| 1100 | 65.33 | 57.29 | 58.79 | 60.66 |
| 1200 | 65.62 | 57.50 | 58.83 | 60.20 |
| 1300 | 65.50 | 57.50 | 58.83 | 60.16 |
| 1430 | 65.66 | 57.95 | 59.50 | 60.45 |

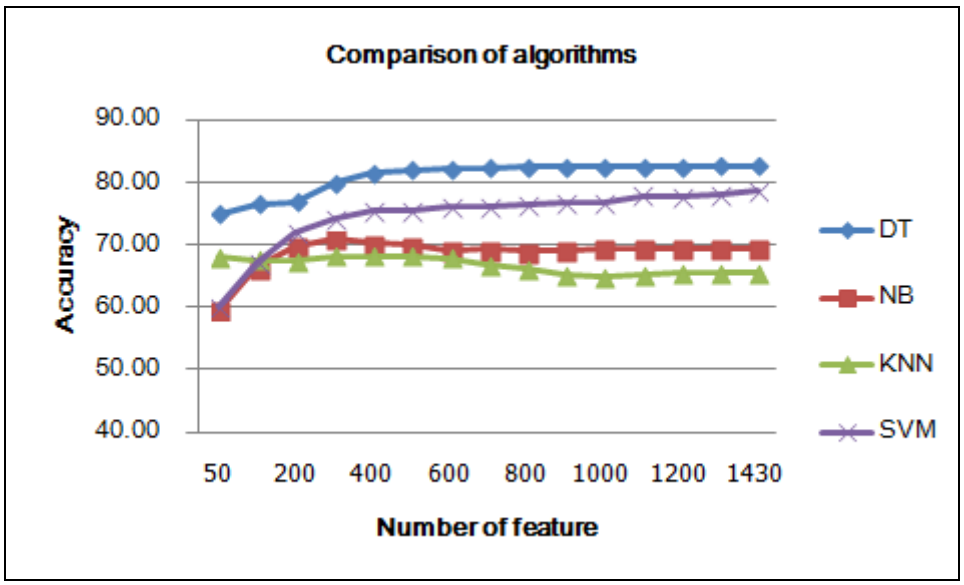



ภาพประกอบ 19 กราฟเปรียบเทียบประสิทธิภาพของอัลกอริทึมเคเนียเรสเนเบอร์

**4.5 ผลการทดลองเปรียบเทียบประสิทธิภาพอัลกอริทึมทั้งหมด**

ตารางที่ 6 ผลการทดลองประสิทธิภาพอัลกอริทึมทั้งหมด

| | TOW | | | |
|---|---|---|---|---|
| Feature | DT | NB | KNN | SVM |
| 50 | 75.00 | 59.58 | 68.08 | 60.12 |
| 100 | 76.50 | 66.41 | 67.62 | 67.04 |
| 200 | 76.90 | 69.70 | 67.41 | 72.04 |
| 300 | 80.00 | 70.95 | 68.25 | 74.16 |
| 400 | 81.50 | 70.25 | 68.25 | 75.45 |
| 500 | 82.00 | 69.91 | 68.25 | 75.58 |
| 600 | 82.20 | 69.04 | 67.95 | 76.08 |
| 700 | 82.30 | 69.25 | 66.91 | 76.20 |
| 800 | 82.50 | 69.00 | 66.12 | 76.50 |
| 900 | 82.50 | 69.08 | 65.08 | 76.83 |
| 1000 | 82.50 | 69.41 | 64.87 | 76.79 |
| 1100 | 82.50 | 69.41 | 65.33 | 77.91 |
| 1200 | 82.50 | 69.41 | 65.62 | 77.87 |
| 1300 | 82.60 | 69.41 | 65.50 | 78.04 |
| 1430 | 82.60 | 69.41 | 65.66 | 78.58 |




ภาพประกอบ 20 กราฟเปรียบเทียบประสิทธิภาพของแต่ละอัลกอริทึม

จากการทดลองสร้างดัชนีด้วยวิธี TOW Weighting และทำการลดคุณลักษณะด้วยค่าสถิติไคสแควร์ (Chi – Square) และเรียนรู้ด้วยอัลกอริทึมทั้งหมด ทำการเปรียบเทียบประสิทธิภาพจากค่าความแม่นยำ (Accuracy) สามารถสรุปได้ว่า อัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ให้ประสิทธิภาพด้านการจำแนกที่ดีสุด รองลงมาเป็น อัลกอริทึมซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) อัลกอริทึมเนอีฟเบย์ (Naïve-Bayes) และอัลกอริทึมเคเนียเรสเนเบอร์ ตามลำดับ เมื่อพิจารณาค่าพารามิเตอร์ที่ส่งผลให้ประสิทธิภาพด้านการจำแนกที่ดีสุด พบว่าวิธีให้ค่าน้ำหนัก TOW ร่วมกับอัลกอริทึมต้นไม้ตัดสินใจ (Decision Tree) ที่จำนวน 1300 คุณลักษณะ ให้ประสิทธิภาพการจำแนกดีสุดที่ระดับ 82.60%

# บรรณานุกรม


ไพศาล เจริญพรสวัสดิ์. 2541. “การตัดคำภาษาไทยโดยใช้คุณลักษณะ. ” วิทยานิพนธ์ปริญญาวิศวกรรมศาสตรมหาบัณฑิต ภาควิชาวิศวกรรมคอมพิวเตอร์ บัณฑิตวิทยาลัยจุฬาลงกรณ์มหาวิทยาลัย.

ยืน ภู่วรวรรณ และ วิวรรธน์ อิ่มอารมณ์. 2529. “การแบ่งแยกพยางค์ไทยด้วยดิกชันนารี. ” ในรายงานการประชุมทางวิชาการวิศวกรรมไฟฟ้า สถาบันอุดมศึกษาแห่งประเทศไทย ครั้งที่ 9 มหาวิทยาลัยขอนแก่น.

วิรัช ศรเลิศล้ำวาณิช. 2536. “การตัดคำภาษาไทยในระบบแปลภาษาการแปลภาษาด้วยคอมพิวเตอร์. ” ในบทความศูนย์เทคโนโลยีอีเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ กรุงเทพมหานคร.

Alm, C.O., Roth, D., and Sproat, R. 2005. **Emotions from text: machine learning for text-based emotion prediction.** Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Canada.

David J. Hand, Heikki Mannila and Padhraic Smyth 2001. **Principles of Data Mining**. MIT Press.

Gill, A.J., French, R.M., Gergle, D., and Oberlander, J. ed. 2008. **The language of emotion in short blog texts.** Proceedings of the ACM conference on Computer supported cooperative work. San Diego. CA. USA.

Gerrod P.W. 2001. **Emotions in Social Psychology**. Psychology Press.

Han, J., & Kamber, M. 2006. **Data mining: Concepts and techniques** (2nd ed.). San Francisco: Morgan Kaufmann.

Haruechaiyasak, C., Jitkrittum, W., Sangkeettrakarn, C., & Damrongrat, C. 2008. **Implementing news article category browsing based on text categorization technique.** International Conference on Web Intelligence and Intelligent Agent Technology. pp: 143-146. Sydney: IEEE Xplore.

Ian H., Witten & Eibe, Frank. 2005. **Data mining : Practical machine learning tools and techniques** (3rd ed.). Boston: Morgan Kaufman.

# บรรณานุกรม (ต่อ)

Jaruskulchai, C. 1998. **An automatic indexing for thai text retrieval**. Doctor of Philosophy Major Computer Science. Graduate School George Washington University.

Liao, C., Alpha, S., & Dixon, P. 2007. **Feature preparation in text categorization.** Technical Report, 2(1), pp: 1-8.

Nicholls, C., and Song, F. 2010. **Comparison of Feature Selection Methods for Sentiment Analysis.** Advances in Artificial Intelligence. Lecture Notes in Computer Science. Vol. 6085. pp: 286-289.

Narayanan, V., Arora, I., and Bhatia, A. 2013. **Fast and Accurate Sentiment Classification Using an Enhanced Naive Bayes Model.** Intelligent Data Engineering and Automated Learning. Lecture Notes in Computer Science. Vol. 8206. pp: 194-201.

Pang, B., Lee, L., and Vaithyanathan, S. ed. 2002. **Thumbs up?: sentiment classification using machine learning techniques.** Proceedings of the conference on Empirical methods in natural language processing.

Pang, B., and Lee, L. 2008. **Opinion Mining and Sentiment Analysis**. Foundations and Trends in Information Retrieval. Vol. 2. pp: 1-135.

Quinlan, J. R. 1993. **C4.5 Programs for machine learning.** San Francisco: Morgan Kaufmann.

Shearer C., 2000. **The CRISP-DM model: The new blueprint for data mining**. Journal of Data Warehousing. Vol.5(4). pp: 13–22.

Sharma, A., and Dey, S., 2012. **A comparative study of feature selection and machine Learning techniques for sentiment analysis.** Proceedings of the ACM Research in Applied Computation Symposium. San Antonio. Texas.

Sui, H., Khoo, C., Chan, S., and Chan, S. 2003. **Sentiment Classification of Product Reviews Using SVM and Decision Tree Induction.** 14th Annual ASIST SIG CR Workshop.

Salton, G., Buckley, C. 1988. **Term-weighting approaches in automatic text retrieval.** Information Processing and Management, 24(5), pp: 513-523.

Yang, Y., Xu, C., and Ren, G. 2012. **Sentiment Analysis of Text Using SVM.** Information Engineering and Mechatronics. Lecture Notes in Electrical Engineering. Vol. 138. pp: 1133-1139.

# ประวัติย่อผู้วิจัย

| | |
|---|---|
| ชื่อ | ดร.นิเวศ จิระวิชิตชัย |
| วัน เดือน ปีเกิด | วันที่ 30 เมษายน 2517 |
| สถานที่เกิด | กรุงเทพฯ |
| สถานที่อยู่ปัจจุบัน | บ้านเลขที่ 195 ถนนจารุเมือง อำเภอปทุมวัน กรุงเทพฯ 10330 |
| ตำแหน่งหน้าที่การงานปัจจุบัน | ผู้อำนวยการหลักสูตร วิทยาศาสตรมหาบัณฑิต |
| | สาขาวิชาระบบสารสนเทศคอมพิวเตอร์ |
| สถานที่ทำงานปัจจุบัน | คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยศรีปทุม |
| ประวัติการศึกษา | พ.ศ. 2540 บธ.บ. มหาวิทยาลัยรามคำแหง |
| | พ.ศ. 2546 คอ.ม.สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ |
| | พ.ศ. 2553 ปร.ด. สถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ |