การเลือกคุณลักษณะและขจัดคุณลักษณะซ้ำซ้อนสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์

วิทยานิพนธ์

ของ

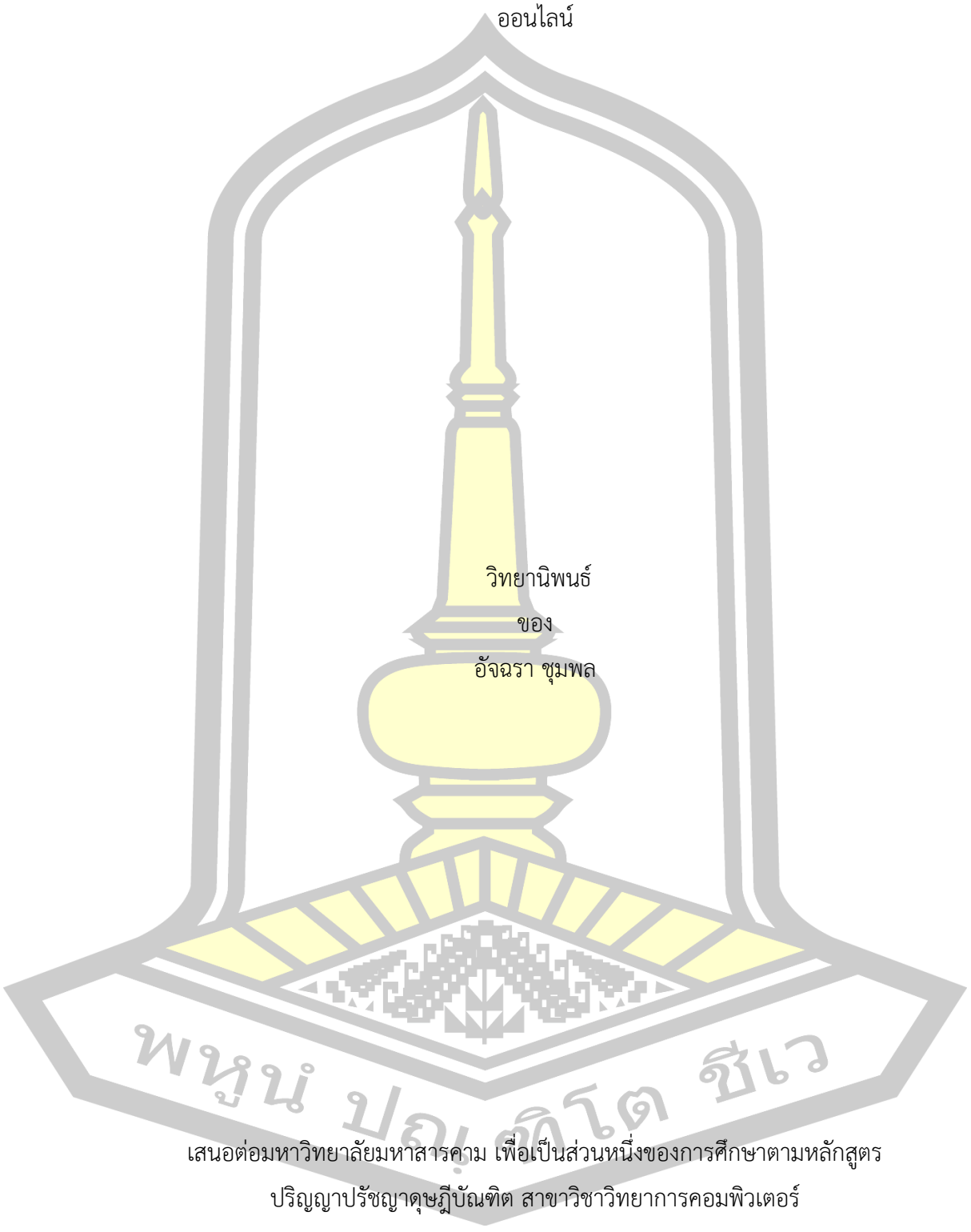
อัจฉรา ชุมพล

เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

กรกฎาคม 2562

การเลือกคุณลักษณะและขจัดคุณลักษณะซ้ำซ้อนสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์



วิทยานิพนธ์

ของ

อัจฉรา ชุมพล

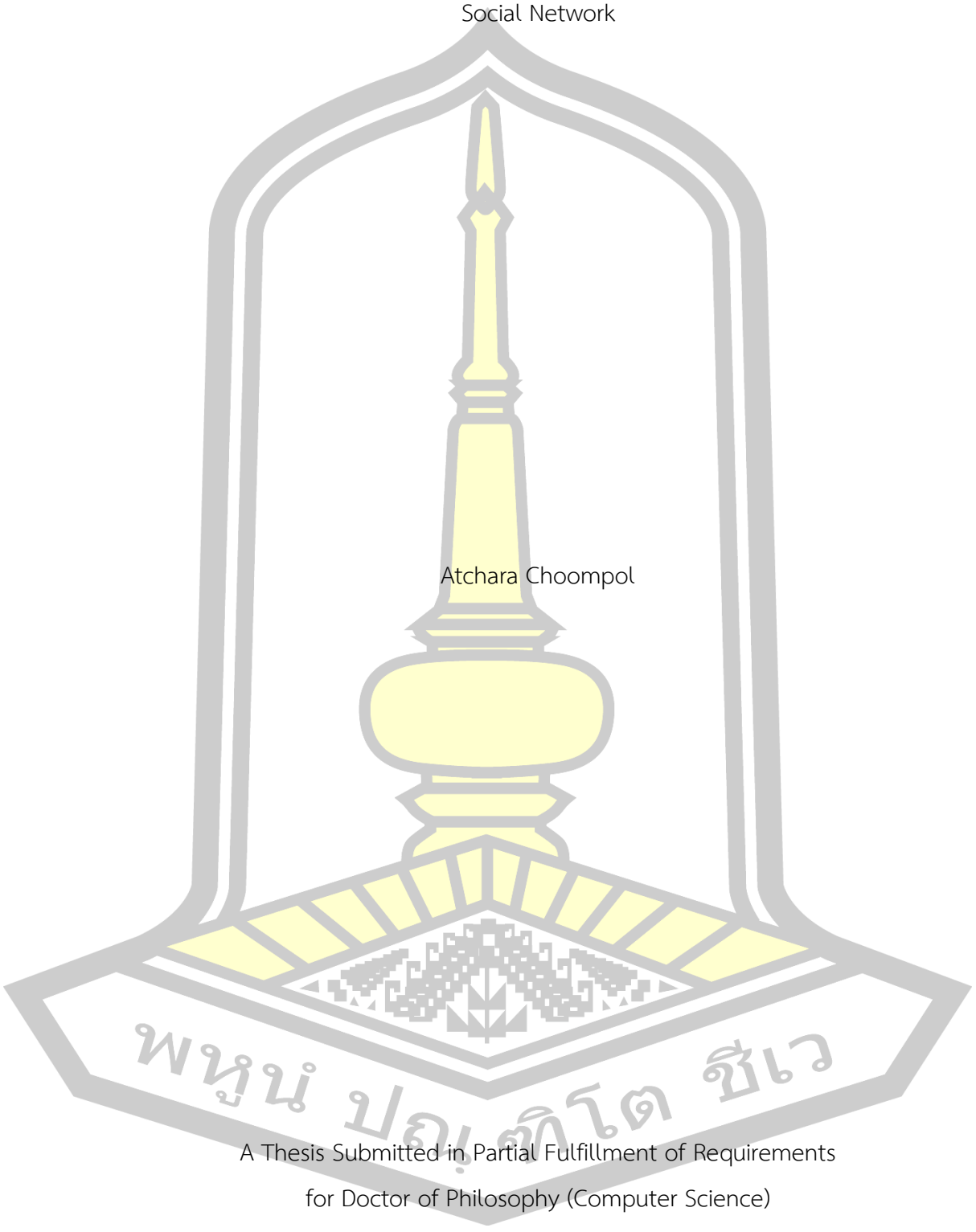
เสนอต่อมหาวิทยาลัยมหาสารคาม เพื่อเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์

กรกฎาคม 2562

Feature Selection and Redundant Feature Elimination for Opinion Classification on Social Network

Atchara Choompol



A Thesis Submitted in Partial Fulfillment of Requirements

for Doctor of Philosophy (Computer Science)

July 2019

คณะกรรมการสอบวิทยานิพนธ์ ได้พิจารณาวิทยานิพนธ์ของนางอัจฉรา ชุมพล แล้ว เห็นสมควรรับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

คณะกรรมการสอบวิทยานิพนธ์

.....ประธานกรรมการ
(ผศ. ดร. วรารัตน์ สงฆ์แป้น)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผศ. ดร. พนิดา ทรงรัมย์)

.....อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม
(ผศ. ดร. พัฒนพงษ์ ชมภูวิเศษ)

.....กรรมการ
(ผศ. ดร. มนัสวี แก่นอำพรพันธ์)

.....กรรมการ
(ผศ. ดร. ฉัตรเกล้า เจริญผล)

มหาวิทยาลัยอนุมัติให้รับวิทยานิพนธ์ฉบับนี้ เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา ปรัชญาดุษฎีบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์ ของมหาวิทยาลัยมหาสารคาม

.....
(ผศ. ศศิธร แก้วมั่น)
คณบดีคณะวิทยาการสารสนเทศ

.....
(ผศ. ดร. กริสน์ ชัยมูล)
คณบดีบัณฑิตวิทยาลัย

ชื่อเรื่อง การเลือกคุณลักษณะและขจัดคุณลักษณะซ้ำซ้อนสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์

ผู้วิจัย อัจฉรา ชุมพล

อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร. พนิดา ทรงรัมย์

ผู้ช่วยศาสตราจารย์ ดร. พัฒนพงษ์ ชมภูวิเศษ

ปริญญา ปรัชญาดุษฎีบัณฑิต **สาขาวิชา** วิทยาการคอมพิวเตอร์

มหาวิทยาลัย มหาวิทยาลัยมหาสารคาม **ปีที่พิมพ์** 2562

บทคัดย่อ

งานวิจัยนี้จึงได้นำเสนอขั้นตอนวิธีในการคัดเลือกคุณลักษณะและการลดคุณลักษณะที่ซ้ำซ้อนสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์ การคัดเลือกคุณลักษณะอาศัยหลักการผสมผสานแนวคิดวิธีฟิลเตอร์โมเดลร่วมกับแนวคิดวิธีการกฎความสัมพันธ์ นำค่าสนับสนุนและค่าความเชื่อมั่นมาพิจารณาร่วมกันเพื่อให้ค่าน้ำหนักของคุณลักษณะ โดยทำการปรับค่าสนับสนุนให้อยู่ในช่วง 0-1 เพื่อไม่ให้มีค่าสนับสนุนของแต่ที่มากเกินไป และนำค่าพารามิเตอร์ที่เรียกว่า p มาใช้เพื่อถ่วงน้ำหนักระหว่างค่าสนับสนุนและค่าความเชื่อมั่น นอกจากนี้งานวิจัยนี้ยังได้นำเสนอการขจัดคุณลักษณะที่ซ้ำซ้อนโดยพิจารณาจากคุณลักษณะที่เกิดร่วมกันในเอกสารเดียวกัน แล้วทำการเลือกคุณลักษณะที่มีค่าน้ำหนักสูงสุดและตัดคุณลักษณะที่เหลือออก จากการทดลองแสดงให้เห็นว่า วิธีการคัดเลือกคุณลักษณะที่นำเสนอให้ประสิทธิภาพในการจำแนกสูงเมื่อข้อมูลมีขนาดใหญ่ที่ ค่า p = 0.8 และให้ค่าความถูกต้องสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05 เมื่อเปรียบเทียบกับขั้นตอนวิธีการคัดเลือกคุณลักษณะแบบฟิลเตอร์โมเดล 3 วิธี ได้แก่ วิธีการ Information Gain วิธีการ Chi-Square วิธีการ Gini Index และใช้เวลาในการคัดเลือกคุณลักษณะน้อยที่สุด การขจัดคุณลักษณะที่ซ้ำซ้อนด้วยวิธีการที่นำเสนอทำให้จำนวนคุณลักษณะลดลง แต่ไม่ได้ลดประสิทธิภาพในการจำแนก

คำสำคัญ : การวิเคราะห์ความคิดเห็น, การคัดเลือกคุณลักษณะ, การขจัดคุณลักษณะซ้ำซ้อน

TITLE Feature Selection and Redundant Feature Elimination for Opinion Classification on Social Network

AUTHOR Atchara Choompol

ADVISORS Assistant Professor Panida Songram , Ph.D.

Assistant Professor Phatthanaphong Chompoowises , Ph.D.

DEGREE Doctor of Philosophy **MAJOR** Computer Science

UNIVERSITY Mahasarakham University **YEAR** 2019

ABSTRACT

This research therefore presents methods for selecting features and eliminating redundant features for opinion classification on social networks. In feature selection method, it selects features based on the concept of filter model together with the concept of association rules. Support and confidence values are used to calculated weight of feature. The support is normalized to 0-1 to remove outlier support. The parameter *p* is adapted to weight between the support and confidence values. In addition, this research presents the elimination of redundant features. If features are in the same documents, the feature having the highest weight is keep and the remaining features are eliminated. From the experiment results in feature selection, they show that the proposed method provides high classification efficiency on big dataset when $p = 0.8$. It gives higher accuracy than Information Gain, Chi-Square, and Gini Index with significance at 0.05. Moreover, it outperforms information Gain, Chi-Square, and Gini Index in computation time. For experimental results in redundant feature elimination, they show that the proposed method can reduce the number of features without efficiency of classification losses.

Keyword : Opinion Mining, Feature Selection, Eliminating Redundant Features

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ สำเร็จสมบูรณ์ได้ด้วยความกรุณาเป็นอย่างสูงจาก ผู้ช่วยศาสตราจารย์ ดร. พนิดา ทรงรัมย์ อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก และผู้ช่วยศาสตราจารย์ ดร.พัฒนพงษ์ ชมภูวิเศษ อาจารย์ที่ปรึกษาวิทยานิพนธ์ร่วม ท่านได้เมตตาให้ความรู้ ช่วยเหลือและให้แนวคิดตลอดจนคำแนะนำในการปรับปรุงแก้ไขข้อบกพร่องต่าง ๆ ตั้งแต่เริ่มต้นจนสำเร็จลุล่วงสมบูรณ์ ผู้วิจัยซาบซึ้งในความกรุณาและขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณ ผู้ช่วยศาสตราจารย์ ดร.วรารัตน์ สงฆ์แป้น ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ฉัตรเกล้า เจริญผล และผู้ช่วยศาสตราจารย์ ดร.มนัสวี แก่นอำพรพันธ์ กรรมการควบคุมวิทยานิพนธ์ ที่กรุณาให้คำปรึกษา คำแนะนำและแนวคิด ตลอดจนแนวทางแก้ไขข้อบกพร่องในการทำวิทยานิพนธ์

ขอขอบพระคุณอาจารย์สรายุทธ กรวิรัตน์ ที่ให้คำแนะนำและคอยช่วยเหลือในการพัฒนาและปรับปรุงเครื่องมือที่ใช้ในการวิจัยในครั้งนี้จนงานสำเร็จลุล่วงตามวัตถุประสงค์

ขอขอบพระคุณเพื่อนร่วมงานคณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์ และพี่น้องชาวสำนักคอมพิวเตอร์มหาวิทยาลัย มหาวิทยาลัยมหาสารคามทุกท่าน ที่หมั่นไต่ถามความก้าวหน้าด้วยความห่วงใย ให้กำลังและคอยช่วยเหลือแก่ผู้วิจัยตลอดระยะเวลาที่เขียนวิทยานิพนธ์

ขอขอบพระคุณนางพรลภัส โรมานชิค นายอมต ชุมพล นางสาวภัสสรนวรัตน์ ภูมิชูชิต นางสาวปริยาวดี ภูมิชูชิต เด็กหญิงเอวารินทร์ ชุมพล ตลอดจนญาติพี่น้องทุกคน ที่คอยส่งเสริมสนับสนุนให้ความห่วงใย และเป็นกำลังใจให้กับผู้วิจัยมาโดยตลอด

ขอกราบขอบพระคุณคุณพ่อประกาศิต ภูมิชูชิต คุณแม่จอน ภูมิชูชิต ผู้ที่มอบทรัพย์สมบัติอันมีค่ายิ่งคือการศึกษาและคอยเป็นกำลังใจให้ผู้วิจัยเสมอมา ทุกคำสอนของท่านเป็นพลังยิ่งใหญ่ที่ทำให้ผู้วิจัยมีแรงผลักดันในการทำให้การวิจัยครั้งนี้ให้สำเร็จลุล่วงไปด้วยดี

ขอขอบพระคุณมหาวิทยาลัยกาฬสินธุ์ที่ได้กรุณามอบทุนอุดหนุนในการศึกษาต่อระดับปริญญาเอกแก่ผู้วิจัยในครั้งนี้

อัจฉรา ชุมพล

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 หลักการและเหตุผล

ปัจจุบันเว็บไซต์เครือข่ายสังคมออนไลน์ (Social Networking Websites) ได้รับความนิยมเป็นอย่างมาก ผู้คนใช้เป็นช่องทางแบ่งปันข้อคิดเห็น ประสบการณ์หรือเหตุการณ์มากขึ้น เนื่องจากเป็นช่องทางการสื่อสารที่มีความสะดวกรวดเร็ว ในแต่ละวันมีข้อความบนเว็บไซต์เครือข่ายสังคมออนไลน์เพิ่มขึ้นเป็นจำนวนมากและมีแนวโน้มจะเพิ่มขึ้นทวีคูณ ข้อความบนเว็บไซต์เครือข่ายสังคมออนไลน์ ประกอบด้วย ข้อเท็จจริง (Facts) และ ความคิดเห็น (Opinions) ข้อเท็จจริงกล่าวถึงวัตถุประสงค์ คุณลักษณะสินค้าหรือบริการ สถานการณ์หรือเหตุการณ์ที่เกิดขึ้นจริงและไม่ได้บ่งบอกถึงความคิดเห็นใด ๆ ในขณะที่ข้อความคิดเห็นจะบ่งบอกทัศนคติที่มีต่อสินค้าหรือบริการและประเด็นต่าง ๆ มีทั้งความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ นักวิจัยและองค์กรหลายแห่งให้ความสนใจนำความคิดเห็นจำนวนมหาศาลที่อยู่บนเว็บไซต์เครือข่ายสังคมออนไลน์มาใช้ประโยชน์ในด้านต่าง ๆ เช่น ด้านการตลาด เพื่อติดตามทัศนคติของผู้บริโภคที่มีต่อสินค้าหรือบริการ เพื่อให้ทราบความต้องการที่แท้จริงของผู้บริโภค ด้านการเมือง เพื่อสำรวจทัศนคติของประชาชนที่มีต่อพรรคหรือนักการเมืองหรือเพื่อทำนายผลการเลือกตั้ง ด้านการศึกษา เพื่อติดตามทัศนคติของผู้เรียนเพื่อนำไปปรับปรุงการจัดการเรียนการสอนให้มีประสิทธิภาพยิ่งขึ้น เป็นต้น เนื่องจากข้อความคิดเห็นที่อยู่บนเว็บไซต์เครือข่ายสังคมออนไลน์มีจำนวนมหาศาล หากต้องการสำรวจความคิดเห็นของผู้คนส่วนมากที่มีต่อผลิตภัณฑ์หรือเหตุการณ์ต่าง ๆ อาจจะต้องเสียเวลาและงบประมาณจำนวนมาก นักวิจัยจึงได้คิดค้นวิธีการเพื่อวิเคราะห์หาประเด็นสำคัญที่ซ่อนอยู่ในข้อความความคิดเห็นจำนวนมาก วิธีหนึ่งที่นิยมใช้ คือ การทำเหมืองความคิดเห็น (Opinion Mining) หรือเรียกอีกอย่างหนึ่งว่า การวิเคราะห์ความรู้สึก (Sentiment Analysis) เป็นศาสตร์แขนงหนึ่งของการประมวลผลภาษาธรรมชาติและการทำเหมืองข้อความ บุคคลและองค์กรจำนวนมากให้ความสนใจนำข้อความคิดเห็นเหล่านั้นมาทำเหมืองความคิดเห็นเพื่อประยุกต์ใช้ในด้านต่าง ๆ เช่น ด้านการตลาด Troussas และ Virvou [1] จำแนกความคิดเห็นจากข้อความที่อยู่บนเฟสบุค (Facebook) โดยใช้วิธีการนาอีฟเบย์ (Naïve Bayes) Akaichi และคณะ [2] จำแนกความคิดเห็นจากข้อความและสัญลักษณ์แสดงอารมณ์ (Emotion) ที่อยู่บนเฟสบุค โดยวิธีการใช้คลังคำ (Lexicon Based) ร่วมกับวิธีการนาอีฟเบย์ (Naïve Bayes) ด้านการเมือง Anjaria และ Guddeti [3] ได้ใช้ข้อมูลทวิตเตอร์ในการทำนายผลการเลือกตั้งประธานาธิบดี โดยใช้วิธีการเรียนรู้ของเครื่อง (Machine Learning) Ortigosa [4] วิเคราะห์ความ

คิดเห็นของผู้เรียนจากข้อความที่อยู่บนเฟสบุค เพื่อประยุกต์ใช้กับการปรับปรุงระบบการเรียนการสอนออนไลน์ (e-Learning) ให้เหมาะสมกับคุณลักษณะของผู้เรียนแต่ละคน เป็นต้น

เนื่องจากข้อความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์ส่วนมากเป็นประโยคสั้น ๆ และค่อนข้างกำกวม (Implicit Sentence) ไม่ได้ระบุถึงคุณลักษณะของสิ่งที่กล่าวไว้ชัดเจน [5] ในการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์จะมีคำคุณลักษณะ (Feature) จำนวนมาก เนื่องจากความหลากหลายของข้อความ ดังนั้นจึงต้องมีการคัดเลือกคุณลักษณะเพื่อนำไปใช้ในการจำแนกความคิดเห็น ซึ่งเป็นกระบวนการสำคัญกระบวนการหนึ่ง ที่ใช้ในลดจำนวนคุณลักษณะที่มีจำนวนมากและทำให้ประสิทธิภาพในการประมวลผลดีขึ้น ปัจจุบันได้มีการนำเสนอวิธีการคัดเลือกคุณลักษณะหลายแบบ [6] [7] [8] [9] เช่น วิธีฟิลเตอร์ (Filter Model) วิธีแรปเปอร์ (Wrapper Model) วิธีฝังตัว (Embedded Model) และวิธีผสมผสาน (Hybrid Model) การเลือกคุณลักษณะสำหรับงานด้านการทำเหมืองข้อความส่วนใหญ่เป็นแบบ Filter Model [10] เนื่องจากมีความง่ายและมีประสิทธิภาพ ซึ่งงานวิจัยนี้ได้ทำการหาค่าน้ำหนักของแต่ละคุณลักษณะเพื่อจัดลำดับความสำคัญของคุณลักษณะโดยใช้แนวคิดวิธีฟิลเตอร์โมเดลผสมผสานกับแนวคิดวิธีการใช้กฎความสัมพันธ์ เพื่อหาค่าน้ำหนักของแต่ละคุณลักษณะและทำการคัดเลือกคุณลักษณะที่สำคัญเพื่อใช้ในการจำแนก นอกจากนี้ยังได้นำเสนอวิธีการลดคุณลักษณะที่ซ้ำซ้อนโดยที่ไม่ส่งผลกระทบต่อประสิทธิภาพในการจำแนก

1.2 วัตถุประสงค์ของการวิจัย

พัฒนาขั้นตอนวิธีในการคัดเลือกคุณลักษณะและขจัดคุณลักษณะซ้ำซ้อนสำหรับการจำแนกความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์

1.3 ความสำคัญของการวิจัย

1. นำเสนอวิธีการใหม่ในการคัดเลือกคุณลักษณะ โดยใช้แนวคิดวิธีฟิลเตอร์โมเดลผสมผสานกับแนวคิดวิธีการใช้กฎความสัมพันธ์

2. คุณลักษณะสำคัญที่ถูกคัดเลือกโดยวิธีการที่นำเสนอสามารถส่งผลให้ประสิทธิภาพในการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์สูง เมื่อเปรียบเทียบกับวิธีการ Information Gain (IG) Chi-Square (Chi2) และ Gini Index

3. วิธีการที่นำเสนอใช้เวลาในการคำนวณค่าน้ำหนักน้อยกว่าวิธีการ Information Gain (IG) Chi-Square (Chi2) และ Gini Index

4. นำเสนอวิธีการใหม่ในการขจัดคุณลักษณะที่ซ้ำซ้อนออกโดยที่ไม่ส่งผลกระทบต่อประสิทธิภาพการจำแนก

1.4 ขอบเขตของการวิจัย

1. ทำการคัดเลือกคุณลักษณะโดยใช้แนวคิดวิธีฟิลเตอร์โมเดลผสมผสานกับแนวคิดวิธีการใช้กฎความสัมพันธ์

2. ทำการขจัดคุณลักษณะที่ไม่จำเป็นสำหรับการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์

3. ใช้วิธีการนาอีฟเบย์ในการจำแนกความคิดเห็นบนเครือข่ายสังคมออนไลน์

4. เปรียบเทียบวิธีการคัดเลือกคุณลักษณะที่นำเสนอกับวิธี Information Gain (IG) Chi-Square (Chi2) และ Gini Index

5. เปรียบเทียบประสิทธิภาพในการคัดเลือกคุณลักษณะ โดยพิจารณาจากประสิทธิภาพการจำแนก และเวลาในการคำนวณค่าน้ำหนักของคุณลักษณะ

6. วัดประสิทธิภาพวิธีการขจัดคุณลักษณะที่ซ้ำซ้อน โดยพิจารณาจากประสิทธิภาพการจำแนก และเวลาในการขจัดคุณลักษณะที่ซ้ำซ้อน

7. ข้อคิดเห็นที่ใช้ในงานวิจัยนี้เป็นข้อคิดเห็นจาก 5 แหล่งข้อมูลมาตรฐาน ได้แก่ 1) Stadford Twitter Sentiment Data [11] 2) SemEval-2017 Task4A Dataset (SemEval) [12] 3) Sentiment Strength Twitter Dataset (SS-Tweet) [13] 4) Health Care Reform (HCR) [14] 5) Sanders Twitter Dataset [15]

8. จำแนกประเภทความคิดเห็นเป็น 2 กลุ่ม คือ เชิงบวก (Positive) และเชิงลบ (Negative)

9. ทำการสุ่มเลือกข้อความคิดเห็นที่เป็นข้อความคิดเห็นเชิงบวกและข้อความคิดเห็นเชิงลบจำนวนเท่ากันเพื่อใช้ในการทดลอง

1.5 นิยามศัพท์เฉพาะ

1. เว็บไซต์เครือข่ายสังคมออนไลน์ คือ เว็บไซต์ที่ผู้ใช้สามารถแบ่งบันข้อความถึงเพื่อนจำนวนมากผ่านผู้ให้บริการด้านเครือข่ายสังคมออนไลน์ (Social Network) บนอินเตอร์เน็ต ซึ่งเว็บไซต์เครือข่ายสังคมออนไลน์ ที่ใช้ในงานวิจัยนี้ คือ เว็บไซต์ Twitter

2. การทำเหมืองความคิดเห็น (Opinion Mining) คือ การวิเคราะห์หาความรู้สึกที่ซ่อนอยู่ในข้อความแสดงความคิดเห็น

3. ประสิทธิภาพในการจำแนก (Efficient) คือ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-Measure) และประสิทธิภาพด้านเวลา

4. การจำแนกความคิดเห็น (Opinion Classification) คือ การจำแนกประเภทความคิดเห็นตามลักษณะการแสดงทัศนคติ ซึ่งแบ่งเป็น 2 ระดับ ได้แก่ เชิงบวก (Positive) และ เชิงลบ (Negative)

5. คุณลักษณะที่ซ้ำซ้อน คือ คุณลักษณะที่ปรากฏในเอกสารเดียวกัน เช่น คุณลักษณะที่ 1 ปรากฏในเอกสาร 1 และ เอกสาร 5 คุณลักษณะที่ 2 ปรากฏในเอกสาร 1 และ เอกสาร 5 แสดงว่า คุณลักษณะที่ 1 และ คุณลักษณะที่ 2 เป็นคุณลักษณะที่ซ้ำซ้อนกัน เป็นต้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้ดำเนินการศึกษาเอกสาร แนวคิดและทฤษฎีต่าง ๆ ที่เกี่ยวข้อง ได้แก่ การทำเหมืองความคิดเห็น (Opinions Mining) การประมวลผลภาษาธรรมชาติ (Natural Language Processing) การคัดเลือกคุณลักษณะ (Feature Selection) รูปแบบข้อมูลแนวตั้ง (Vertical Data Format) การวัดประสิทธิภาพในการจำแนกข้อมูล (Evaluation) และงานวิจัยที่เกี่ยวข้อง เพื่อเป็นแนวทางในการตัดคุณลักษณะที่ซ้ำซ้อนและเพิ่มประสิทธิภาพสำหรับการจำแนกความคิดเห็นบนเว็บไซต์เครือข่ายสังคมออนไลน์

2.1 การทำเหมืองความคิดเห็น (Opinion Mining)

การทำเหมืองความคิดเห็น หรือการวิเคราะห์ความรู้สึก (Sentiment Analysis) [16] เป็นการวิเคราะห์ความคิดเห็น ความรู้สึก ประเมินทัศนคติและอารมณ์ของผู้คน ที่มีต่อสิ่งต่าง ๆ เช่น สินค้า บริการ องค์กร ประเด็น เหตุการณ์ เป็นต้น โดยเก็บรวบรวมข้อมูลและตรวจสอบอารมณ์ของผู้คนจากข้อความคิดเห็นที่อยู่ในสื่อสังคมออนไลน์ เช่น Facebook, LinkedIn, Twitter, Flickr และ YouTube เป็นต้น ประเด็นหลักในการทำเหมืองความคิดเห็น คือ จำแนกความคิดเห็นเพื่อให้ทราบถึงความพึงพอใจของบุคคลเหล่านั้นว่ามีความรู้สึกในเชิงบวกหรือเชิงลบ การทำเหมืองความคิดเห็นได้นำหลักการทำเหมืองข้อความและการประมวลผลภาษาธรรมชาติมาประยุกต์ใช้ ปัจจุบันมีการนำวิธีการทำเหมืองความคิดเห็นมาใช้เพื่อประโยชน์หลายด้าน ได้แก่ ด้านการตลาด ใช้เพื่อติดตามทัศนคติของผู้บริโภคที่มีต่อสินค้าหรือบริการ เพื่อให้เข้าใจถึงความต้องการที่แท้จริงของผู้บริโภค ด้านการเมือง ใช้เพื่อสำรวจทัศนคติของประชาชนที่มีต่อพรรคหรือนักการเมืองหรือเพื่อทำนายผลการเลือกตั้ง ด้านการศึกษา ใช้เพื่อติดตามทัศนคติของผู้เรียนเพื่อนำไปปรับปรุงการจัดการเรียนการสอนให้มีประสิทธิภาพยิ่งขึ้น เป็นต้น การทำเหมืองความคิดเห็นจะใช้ข้อมูลที่ได้จากข้อความแสดงความคิดเห็น ซึ่งลักษณะข้อความคิดเห็นอยู่ในรูปแบบไม่มีโครงสร้าง ในการนำข้อมูลไปเข้าสู่กระบวนการจำแนกความคิดเห็นจึงจำเป็นต้องแปลงข้อความที่ไม่มีโครงสร้างให้อยู่ในรูปแบบที่มีโครงสร้างก่อน เรียกว่ากระบวนการเตรียมข้อมูล (Data Preprocessing) เพื่อนำข้อมูลไปสร้างตัวจำแนกความคิดเห็น

การทำเหมืองความคิดเห็น แบ่งเป็น 3 ระดับ ได้แก่

1) ระดับเอกสาร (Document Level) เป็นการสรุปความคิดเห็นโดยรวมจากข้อมูลเอกสารที่รวบรวม โดยไม่สนใจประเด็นที่อยู่ในเอกสารนั้น ผลการจำแนกความคิดเห็นจะสรุปในภาพรวมของแต่ละเอกสารเป็นทัศนคติเชิงบวก ลบ หรือเป็นกลาง ตัวอย่างเช่น การรวบรวมบทวิจารณ์เกี่ยวกับผลิตภัณฑ์ (Product Review) แล้วนำมาเข้าระบบเพื่อตรวจสอบว่าบทวิจารณ์นั้นแสดงความคิดเห็น

โดยรวมเกี่ยวกับผลิตภัณฑ์เชิงบวกหรือเชิงลบ การวิเคราะห์ในระดับนี้ถือว่าแต่ละเอกสารแสดงความคิดเห็นต่อเอนทิตีเดียว (Single Entity) ดังนั้น จึงไม่ได้ใช้กับเอกสารที่ประเมินหรือเปรียบเทียบหลายเอนทิตี

2) ระดับประโยค (Sentence Level) เป็นการนำเอกสารข้อความมาแบ่งเป็นประโยค เพื่อวิเคราะห์ความคิดเห็นที่อยู่ในประโยคว่ามีความคิดเห็นเชิงบวก เชิงลบ หรือเป็นกลาง ประโยคที่เป็นกลางคือข้อความที่ไม่บ่งบอกถึงความคิดเห็น งานวิจัยส่วนมากทำการวิเคราะห์ความคิดเห็นในกับการงานด้านจำแนกประเภทของความคิดเห็น (Subjectivity Classification) ซึ่งเป็นการแยกระหว่างประโยคบอกเล่า (Objective Sentence) ที่แสดงข้อมูลความจริง กับประโยคที่แสดงมุมมองและความคิดเห็นส่วนตัว (Subjective Sentences) [17]

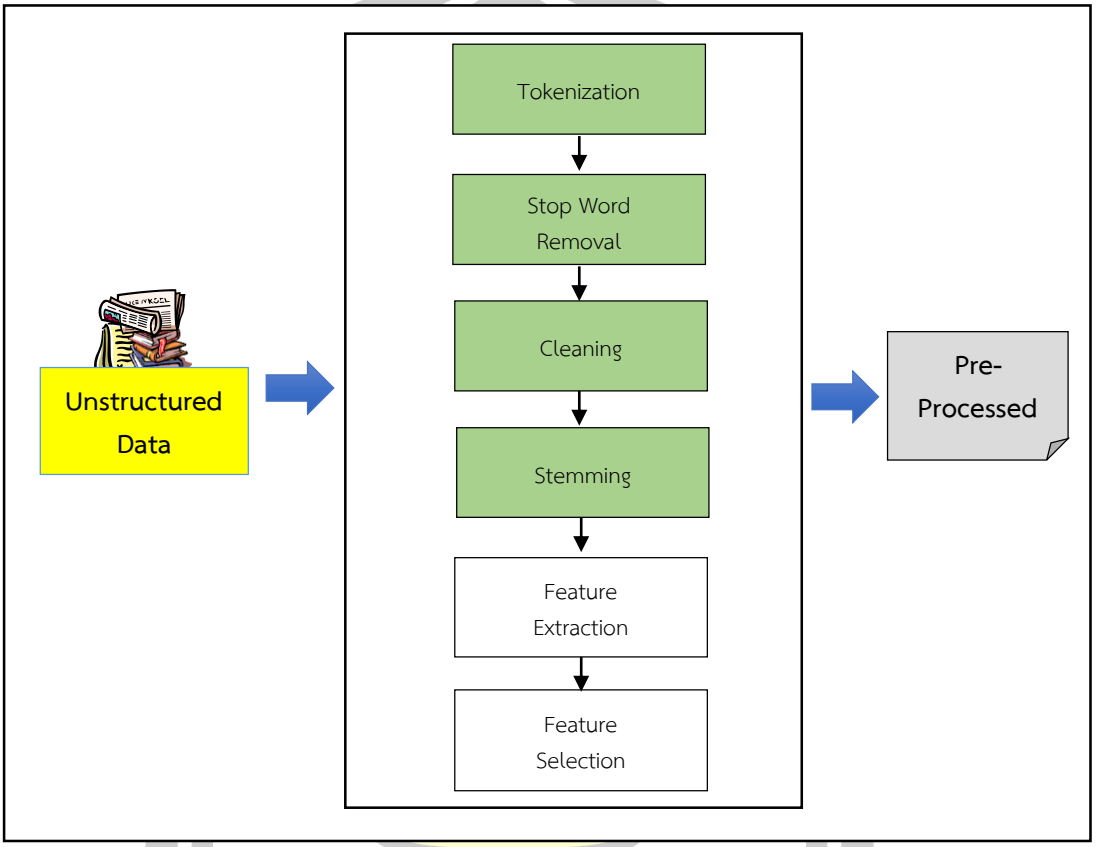
3) ระดับคำหรือวลี (Word or Phase Level) เป็นการวิเคราะห์ความคิดเห็นแบบละเอียดมากขึ้น โดยแบ่งตามคุณลักษณะที่สนใจพิจารณาที่อยู่ในข้อความ แล้วนำมาจำแนกความคิดเห็นเพื่อให้ทราบว่าผู้คนมีทัศนคติอย่างไรต่อคุณลักษณะที่พิจารณา ตัวอย่างเช่น ประโยค “แม้ว่าบริการจะไม่ดี แต่ฉันก็ชอบบรรยากาศของร้านอาหารนี้” จะเห็นว่าผู้กล่าวถึงร้านอาหารมีความคิดเห็นเชิงลบต่อบริการของร้าน แต่มีความคิดเห็นเชิงบวกต่อบรรยากาศของร้าน เป็นต้น จะเห็นว่าใน 1 ประโยคมีการพูดถึง 2 เอนติตี้ ซึ่งก็คือ บริการ กับ บรรยากาศ การวิเคราะห์ระดับคำหรือวลี จะค้นหาสิ่งที่ผู้คนกล่าวถึง และวิเคราะห์ว่าเป็นการกล่าวถึงในด้วยความรู้สึกบวก ลบ หรือเป็นกลาง

กระบวนการทำเหมืองความคิดเห็นในปัจจุบัน แบ่งออกเป็น 3 วิธี คือ 1) วิธีการใช้คลังคำ เป็นการจำแนกความคิดเห็นโดยพิจารณาจากพจนานุกรมคำ ซึ่งมีคลังคำที่ได้ระบุขั้วความคิดเห็นไว้แล้ว 2) วิธีการเรียนรู้ของเครื่อง เป็นวิธีการที่อาศัยหลักการเรียนรู้ของมนุษย์ในการพัฒนากระบวนการเรียนรู้ของเครื่อง เพื่อให้เครื่องคอมพิวเตอร์สามารถทำงานได้อย่างมีประสิทธิภาพ และ 3) วิธีการผสมผสาน ใช้หลักการผสมผสานระหว่างวิธีการใช้คลังคำและวิธีการเรียนรู้ของเครื่อง โดยการทำเหมืองความคิดเห็น 3 วิธีการข้างต้น มีกระบวนการเตรียมข้อมูลและขั้นตอนวิธีการจำแนกความคิดเห็นที่แตกต่างกัน ซึ่งจะกล่าวในหัวข้อถัดไป

2.2 ขั้นตอนการเตรียมข้อมูล (Data Preprocessing)

เนื่องจากข้อความคิดเห็นเป็นข้อมูลที่อยู่รูปแบบที่ไม่มีโครงสร้าง จำเป็นต้องแปลงข้อมูลให้อยู่ในรูปแบบที่มีโครงสร้างก่อน โดยใช้กระบวนการเตรียมข้อมูล นอกจากนี้แล้วพบว่า ข้อความคิดเห็นบนเว็บไซต์เครือข่ายสังคมออนไลน์ มีข้อมูลกำกวม (Inconsistent) และมีข้อมูลรบกวน (Noisy) ค่อนข้างมาก ซึ่งถ้าข้อมูลกำกวมหรือมีข้อมูลรบกวนมาก เป็นไปได้ว่าจะนำไปสู่ความสับสนและผลลัพธ์ที่ผิดพลาด กระบวนการเตรียมข้อมูลจะช่วยให้ข้อมูลมีรูปแบบเดียวกัน (Consistency) และมีถูกต้อง (Accuracy) มากขึ้น กระบวนการเตรียมข้อมูล [18] [19] โดยทั่วไปประกอบด้วย 6

กระบวน ได้แก่ การตัดคำ (Tokenization) การกำจัดคำหยุด (Stop Word Removal) การทำความสะอาด (Cleaning) การหารากคำศัพท์ (Stemming) การสกัดคุณลักษณะ (Feature Extraction) และการเลือกคุณลักษณะ (Feature Selection) ดังแสดงในรูปที่ 1



รูปที่ 1 ขั้นตอนการเตรียมข้อมูล

จากรูปที่ 1 แสดงให้เห็นขั้นตอนการเตรียมข้อมูลที่อยู่ในรูปแบบที่ไม่มีโครงสร้างให้มีโครงสร้างเพื่อนำไปใช้ในกระบวนการจำแนกความคิดเห็นต่อไป ซึ่งไม่จำเป็นต้องมีครบทุกกระบวนการ อาจจะเลือกทำบางกระบวนการขึ้นอยู่กับข้อมูลที่นำมาใช้และวิธีการวิเคราะห์ข้อมูล งานวิจัยบางส่วนที่ใช้วิธีการจำแนกความคิดเห็นด้วยวิธีการใช้คลังคำ กระบวนการที่ใช้มีเพียงกระบวนการสกัดคุณลักษณะ (Feature Extraction) โดยใช้คลังคำความคิดเห็นมาเปรียบเทียบกับคำที่อยู่ในเอกสารเพื่อสกัดคุณลักษณะที่สำคัญออกมาจากเอกสาร ส่วนกระบวนการเลือกคุณลักษณะ (Feature Selection) ใช้เพื่อลดจำนวนคุณลักษณะที่ได้จากกระบวนการสกัดคุณลักษณะ การเลือกคุณลักษณะที่ดีจะช่วยเพิ่มประสิทธิภาพในการวิเคราะห์ความคิดเห็นทั้งด้านเวลาและความถูกต้อง นิยมใช้กับวิธีการเรียนรู้ของเครื่อง การเตรียมข้อมูลแต่ละขั้นตอนอธิบายได้ดังนี้

2.2.1 การตัดคำ

การตัดคำ เป็นกระบวนการนำเอกสารข้อความมาแบ่งเป็นประโยค (Sentence) หรือ คำ (Word) โดยทั่วไปการตัดคำในข้อความภาษาอังกฤษนิยมใช้ ช่องว่าง (White Space) คอมมา (Comma: ,) จุดทศนิยม (Point: .) เครื่องหมายอัฒภาค (Semicolon: ;) เครื่องหมายคำถาม (Question Mark: ?) หรือสัญลักษณ์ต่าง ๆ [20] กระบวนการตัดคำ เริ่มจากสแกนข้อความทั้งหมด เพื่อหาขอบเขตคำและประโยค ซึ่งคำในภาษาอังกฤษจะใช้ช่องว่างเพื่อแบ่งคำ และใช้จุดทศนิยมหลังจบประโยค [21] ดังแสดงตัวอย่างในตาราง 1

ตาราง 1 ตัวอย่างการตัดคำภาษาอังกฤษ

INPUT	OUTPUT
This car is really great, latest technologies are included.	This | car | is | really | great | latest | technologies | are | included
This week is not going as i had hoped	This | week | is | not | going | as | i | had | hoped
I hate when I have to call and wake people up	I | hate | when | I | have | to | call | and | wake | people | up
I need a hug	I | need | a | hug

วิธีการตัดคำที่ได้รับความนิยมในงานด้านเหมืองข้อความ มี 3 วิธี [11] [22] ได้แก่ วิธีการ Unigram วิธีการ Bigram และ วิธีการ Unigram + Bigram แต่ละวิธีการมีรายละเอียด ดังนี้

1) วิธีการ Unigram [23] เป็นวิธีการหนึ่งที่ง่ายและนิยมใช้มากที่สุดในการทำเหมืองข้อความ วิธีการนี้จะทำการตัดคำ 1 คำ เพื่อแทน 1 คุณลักษณะ เช่น คำว่า “This car is really great” หากทำการตัดคำด้วย วิธีการ Unigram จะได้คุณลักษณะ คือ | This | car | is | really | great | ตัวอย่างงานวิจัยที่ใช้วิธีการ Unigram เช่น Read [24] ทำการจำแนกความคิดเห็นบนเว็บไซต์ทวิตเตอร์ที่รวบรวมจาก Usenet News Group โดยใช้ Unigram Feature เป็นตัวแทนคุณลักษณะ ผลการวัดประสิทธิภาพ พบว่า การจำแนกความคิดเห็นด้วยวิธีการ Unigram ร่วมกับวิธีการซัพพอร์ตเวกเตอร์แมชชีนมีประสิทธิภาพสูงกว่าการใช้วิธีการ Unigram ร่วมกับวิธีการนาอีฟเบย์ การตัดคำด้วยวิธีการ Unigram มีข้อเสีย คือ จะได้คุณลักษณะจำนวนมาก ทำให้การประมวลผลใช้เวลานาน และมีปัญหาเกี่ยวกับการแปลความหมายของคำเมื่อมีคำที่มีความหมายเชิงปฏิเสธอยู่ร่วมด้วย เช่น good ความหมายคือ ดี แต่เมื่อเติมคำว่า no good จะมีความหมายตรงกันข้าม คือ ไม่ดี หากทำการตัดคำโดยวิธีการ Unigrams คำว่า good จะถูกจัดไว้เป็นอีกหนึ่ง

คุณลักษณะ เมื่อทำการแทนค่าในเอกสาร จะทำให้ผลการจำแนกคลาดเคลื่อนได้ โดยสรุป วิธีการ Unigrams เป็นวิธีการที่ง่ายและได้รับความนิยมเป็นอย่างมาก แต่ยังมีข้อเสียอยู่มากเช่นกัน

2) วิธีการ Bigrams เป็นวิธีการตัดคำโดยการผสม 2 คำ แทนด้วย 1 คุณลักษณะ เช่น คำว่า “This car is really great” หากทำการตัดคำด้วย วิธีการ Bigrams จะได้คุณลักษณะ คือ |This car| |car is| |is really| |really great| งานวิจัยของ Tripathy และคณะ [22] นำเสนอผลการทดลองจำแนกความคิดเห็นโดยใช้หลักการ N-gram ได้แก่ Unigram, Bigram, Trigram, Unigram ร่วมกับ Bigram, Bigram ร่วมกับ Trigram และ Unigram ร่วมกับ Bigram ร่วมกับ Trigram พบว่า โดยภาพรวม วิธีการ Unigram ร่วมกับ Bigram มีประสิทธิภาพสูงสุดเมื่อเทียบกับวิธีการอื่น

3) วิธีการ Unigrams ร่วมกับ Bigrams เป็นวิธีการผสมผสานระหว่างวิธีการ Unigrams และวิธีการ Bigrams โดยข้อความที่มีคำปฏิเสธอยู่ก่อนหน้า เช่น no, not จะใช้หลักการแบ่งข้อความด้วย Bigrams ส่วนข้อความอื่น ๆ จะใช้หลักการแบ่งข้อความด้วย Unigrams งานวิจัยของ Go และคณะ [11] แสดงให้เห็นว่าการจำแนกข้อความคิดเห็นโดยการตัดคำด้วยวิธีการ Unigrams ร่วมกับ Bigrams มีประสิทธิภาพการจำแนกสูง โดยนำเสนอการจำแนกความคิดเห็นที่อยู่บนเว็บไซต์ทวิตเตอร์ (Twitter Data) โดยใช้ Unigram, Bigrams, Unigrams ร่วมกับ Bigrams และ Unigrams ร่วมกับ Part of Speech Tags. ผลการวิจัยแสดงให้เห็นว่าการจำแนกความคิดเห็นด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีนร่วมกับ ร่วมกับ การตัดคำโดยวิธีการ Unigrams ร่วมกับ Bigrams มีประสิทธิภาพการจำแนกสูงที่สุด

2.2.2 การกำจัดคำหยุด

คำหยุด เป็นคำที่พบมากในเอกสารข้อความ การกำจัดคำหยุดซึ่งพบมากในเอกสารข้อความมีวัตถุประสงค์เพื่อเพิ่มประสิทธิภาพในการประมวลผล เพราะคำหยุดเป็นคำที่ไม่สื่อความหมาย เป็นคำสิ้นเปลืองในการทำเหมืองข้อความ หากมีเป็นจำนวนมากจะทำให้การประมวลผลช้า ไม่มีประสิทธิภาพ การทำเหมืองข้อความภาษาอังกฤษ มีคลังคำหยุดที่ถูกรวบรวมและเผยแพร่ให้นักวิจัยหรือองค์กรนำไปใช้งานอย่างแพร่หลาย ตัวอย่างกลุ่มคำหยุด เช่น คำบุพบท (Prepositions) คำสรรพนาม (Pronouns) คำเชื่อม (Conjunctions) คำนำหน้านาม (Articles) [18] [25] การกำจัดคำหยุดทำได้หลายวิธี วิธีที่ง่ายที่สุดคือการสร้างคลังคำหยุดไว้ แล้วนำไปวิเคราะห์เปรียบเทียบกับคำที่อยู่ในเอกสาร หากปรากฏคำที่ตรงกับคลังคำหยุด ให้ทำการตัดคำเหล่านั้นออกจากเอกสาร นอกจากคำหยุดข้างต้นแล้ว ยังมีคำหยุดที่เป็นข้อความเฉพาะ วิธีการหาคำหยุดที่เป็นข้อความเฉพาะทำได้โดยการหาค่าความถี่ของคำทั้งหมดในเอกสาร จากนั้นทำการจัดเรียงลำดับของคำตามค่าความถี่ที่ปรากฏในเอกสารจากน้อยไปมาก ซึ่งจะพบว่าข้อความเฉพาะจะพบเป็นจำนวนมากในเอกสาร การกำจัดคำเหล่านี้ควรทำการตัดคำที่ตรงกับคลังคำหยุดออกก่อน จากนั้นหาค่าความถี่ของแต่ละคำในเอกสาร

แล้วกำหนดค่าความถี่สูงสุดของคำที่ปรากฏในเอกสารไว้ หากคำใดมีค่าความถี่มากกว่าที่กำหนด จะถือว่าเป็นคำหยุด ให้ลบออกไปจากเอกสาร แล้วนำส่วนที่เหลือในเอกสารไปประมวลผลขั้นตอนต่อไป [26] ตัวอย่างคำหยุดในภาษาอังกฤษ ดังแสดงในตาราง 2

ตาราง 2 ตัวอย่างคำหยุดในภาษาอังกฤษ

a	doing	if	she	very
about	don't	i've	she'd	was
above	down	in	she'll	wasn't
after	during	into	she's	we
again	each	is	should	we'd
against	few	isn't	shouldn't	we'll
all	for	it	so	we're
am	from	itself	some	we've
an	further	let's	such	were
and	had	me	than	weren't
any	hadn't	more	that	what
are	has	most	that's	what's
aren't	hasn't	mustn't	the	when
as	have	my	their	when's
at	haven't	myself	theirs	where
be	having	no	them	where's
because	he	nor	themselves	which
been	he'd	not	then	while
before	he'll	of	there	who
being	he's	off	there's	who's
below	her	on	these	whom
between	here	once	they	why
both	here's	only	they'd	why's
but	hers	or	they'll	with

ตาราง 2 ตัวอย่างคำหยุดในภาษาอังกฤษ (ต่อ)

by	herself	other	they're	won't
can't	him	ought	they've	would
cannot	himself	our	this	wouldn't
could	his	ours	those	you
couldn't	how	ourselves	through	you'd
did	how's	out	to	you'll
didn't	i	over	too	you're
do	i'd	own	under	you've
does	i'll	same	until	your
doesn't	i'm	shan't	up	yourself
				yourselves

2.2.3 การทำความสะอาดข้อมูล

การทำความสะอาดข้อมูล เป็นการตรวจสอบและแก้ไขคำที่สะกดผิดให้ถูกต้อง แก้ไขคำที่ไม่สมบูรณ์ให้เป็นคำที่สมบูรณ์ ข้อมูลที่อยู่บนเครือข่ายสังคมออนไลน์ โดยส่วนมากมักจะเขียนผิด หรือมีตัวอักษรซ้ำ ๆ กระบวนการนี้ทำเพื่อลดจำนวนคุณลักษณะที่ซ้ำซ้อน รวมคำที่มีความหมายเดียวกันให้อยู่ในรูปแบบเดิมที่ถูกต้อง ตัวอย่างข้อมูลการพิมพ์อักษรซ้ำ ๆ ที่พบบ่อยบนเครือข่ายสังคมออนไลน์ เช่น คำว่า love พิมพ์เป็น looooooooove, looooove หรือ loveeeeee เป็นต้น ขั้นตอนการทำความสะอาดข้อมูลจะเริ่มจากการตรวจสอบตัวอักษรที่ปรากฏซ้ำในคำ หากพบอักษรนั้นปรากฏเกิน 2 ครั้ง ให้ทำการลบอักษรที่ซ้ำออกเหลือไว้ไม่เกิน 2 แล้วนำไปเข้ากระบวนการตรวจสอบแก้ไขคำผิดด้วยพจนานุกรมต่อไป เช่น “This car is too smalllll” จากข้อความข้างต้น คำว่า “smalllll” จะถูกแก้ไขเป็น “small” เป็นต้น

2.2.4 การหารากคำศัพท์

การหารากคำศัพท์ เป็นวิธีการลดจำนวนคำ โดยการจัดกลุ่มคำที่มีความหมายเหมือนกันไว้ด้วยกัน [18] เช่น Agreed, Agreeing, Agreement เป็นคำที่มาจากรากคำศัพท์เดียวกัน จะถูกจัดไว้ในกลุ่มเดียวกัน คือ Agree ในการหารากคำศัพท์โดยส่วนมากจะใช้คลังคำเข้ามาช่วย ประโยชน์ของกระบวนการนี้คือ จะทำให้จำนวนคุณลักษณะลดลง ช่วยเพิ่มความเร็วในประมวลผล วิธีการหาราก

คำศัพท์ 2 วิธีการที่นิยมนำมาใช้งาน ได้แก่ การสร้างตาราง (Table Lookup) และการลบคำเติม (Affix Removal) แต่ละวิธีการมีรายละเอียดดังนี้

1) การสร้างตาราง เป็นวิธีการสร้างตารางรากศัพท์สำหรับแต่ละเทอม จัดเป็นวิธีการที่ง่ายไม่มีความซับซ้อน การค้นหาแต่ละเทอมทำได้สะดวก แต่ปัญหาคือ การสร้างให้ครอบคลุมทุกเทอมส่งผลให้สิ้นเปลืองพื้นที่ในการจัดเก็บ การประมวลผลล่าช้า [26] ตัวอย่างวิธีการสร้างตาราง เช่น Engineering, Engineered, Engineer, Engineers เป็นคำที่ใช้รากคำศัพท์ตัวเดียวกัน คือ Engineer แต่ละเทอมจะถูกจัดเก็บรากคำศัพท์ไว้ดังตาราง 3

ตาราง 3 ตัวอย่างของเทอมที่มีรากคำศัพท์เป็น Engineer

Term	Stem
Engineering	Engineer
Engineered	Engineer
Engineer	Engineer
Engineers	Engineer

2) การลบคำเติม เป็นการลบคำนำหน้า (Prefix) และคำต่อท้าย (Suffix) ซึ่งส่วนมากกระบวนการลบคำเติมในภาษาอังกฤษนิยมลบคำต่อท้าย เนื่องจากคำต่อท้ายจะเป็นคำที่แปลงรูปของคำตามหน้าที่ แต่ยังคงความหมายเดิม เช่น การแปลงพหูพจน์เป็นเอกพจน์ เปลี่ยน sses เป็น ss เปลี่ยน ies เป็น y เปลี่ยน es เป็น e หรือตัด s ออกจากท้ายคำ เป็นต้น ปัจจุบันมีขั้นตอนวิธีการหารากคำศัพท์เผยแพร่อยู่เป็นจำนวนมาก และที่ได้รับความนิยมเป็นอย่างมาก คือ Porter’s Stemmer

2.2.5 การสกัดคุณลักษณะ

การสกัดคุณลักษณะ เป็นขั้นตอนดึงคุณลักษณะของข้อความออกมา เพื่อใช้เป็นคุณลักษณะตัวแทน ซึ่งสามารถเลือกคุณลักษณะของข้อความได้หลายวิธี เช่น คำหรือหน้าที่ของคำ (Part of Speech) ความสัมพันธ์ของคำในประโยค (Syntax) คำปฏิเสธ (Negation) เป็นต้น งานวิจัยส่วนมากใช้วิธีการวิเคราะห์จากคำ เรียกว่า การแทนข้อความด้วยถุงคำ (Bag-of-Word) คือ สกัดคำแสดงความคิดเห็นที่อยู่ในเอกสารเพื่อเป็นตัวแทนคุณลักษณะ และวิธีการระบุหน้าที่ของคำด้วยชุดหน้าที่คำ (POS Tag Set) ซึ่งจะระบุหน้าที่ของแต่ละคำในเอกสารว่าเป็นคำนาม (Noun) คำกริยา (Verb) คำสรรพนาม (Pronoun) คำคุณศัพท์ (Adjective) คำวิเศษณ์ (Adverb) และอื่น ๆ การเลือกคำคุณลักษณะจากหน้าที่ของคำเพื่อการจำแนกความคิดเห็น โดยส่วนมากเลือกคำที่มีหน้าที่เป็นคำกริยา

คำคุณศัพท์ และคำวิเศษณ์ เนื่องจากเป็นคำที่บอกถึงความคิดเห็นในประโยคได้ หลังจากได้ตัวแทนคุณลักษณะแล้วจะมีการแทนค่าให้อยู่ในรูปแบบเวกเตอร์ วิธีที่ใช้ในการคำนวณค่าเพื่อแทนค่าคุณลักษณะที่นิยมใช้ [27] ได้แก่

1) การแทนคำด้วยค่าการเกิดขึ้นหรือไม่เกิดขึ้นของคำ (Boolean Weighting) เป็นวิธีการแทนข้อความด้วยคำที่รวบรวมไว้แล้วในถุงคำ หากมีคำในถุงคำเกิดขึ้นในข้อความที่นำมาวิเคราะห์ จะแทนค่าด้วย 1 แต่หากในข้อความไม่มีคำที่กำหนดไว้ในถุงคำเกิดขึ้น จะแทนค่าด้วย 0 ดังสมการ (2.1)

$$B_{td} = \begin{cases} 1, \ for\ term\ \ present\ in\ document \\ 0, \ otherwise \end{cases} \qquad (2.1)$$

โดยที่ B_{td} คือ ค่าการเกิดคุณลักษณะ t ในเอกสาร d

2) การแทนข้อความด้วยค่าความถี่การเกิดคำ (Term Frequency: TF) วิธีการนี้จะสนใจจำนวนครั้งของคำที่เกิดขึ้นในเอกสาร เก็บข้อมูลเป็นจำนวนความถี่ของการเกิดคุณลักษณะหรือเทอม (Term) ที่ปรากฏในเอกสาร หากคุณลักษณะใดปรากฏบ่อยในเอกสารค่าความถี่ย่อมมีค่าสูง หากไม่ปรากฏเลยค่าความถี่จะมีค่าเป็น 0 คำนวณได้แสดงดังสมการ (2.2)

$$tf_{td} = freq(t,d) \qquad (2.2)$$

เมื่อ $freq(t,d)$ คือ ค่าความถี่ของการเกิดคุณลักษณะ t ในเอกสาร d

3) การแทนค่าความถี่และค่าความถี่ผกผัน (Term Frequency – Inverse Document Frequency: TF-IDF) เป็นวิธีการให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของเอกสาร ซึ่งควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้นและปรากฏอยู่น้อยในชุดของเอกสารที่เหลือทั้งหมด วิธีการนี้ได้จากแนวความคิดว่าการแทนข้อความด้วยค่าความถี่การปรากฏของคุณลักษณะเพียงอย่างเดียว ไม่สามารถจำแนกข้อความได้ดีพอ เพราะถ้าคุณลักษณะนั้นเกิดขึ้นเป็นจำนวนมากในทุก ๆ เอกสาร แสดงว่าคุณลักษณะดังกล่าวไม่สามารถใช้เป็นตัวแทนของเอกสารได้ จึงต้องหาค่าความถี่ผกผันด้วย คำนวณได้ดังสมการ (2.3)

$$idf_{td} = \log(\frac{N}{D_t}) \quad (2.3)$$

โดยที่ N คือ จำนวนเอกสารทั้งหมด

D_t คือ จำนวนเอกสารทั้งหมดที่มีคุณลักษณะ t ปรากฏอยู่

การหาค่าความถี่และค่าความถี่ผกผัน จะคำนึงถึงความถี่ของการปรากฏคุณลักษณะในเอกสาร และค่าความถี่ผกผัน คำนวณจากผลคูณของค่าความถี่การเกิดคำและค่าความถี่ผกผัน ดังสมการ (2.4)

$$tfidf_{td} = tf_{td} \times idf_{td} \quad (2.4)$$

จากการศึกษางานวิจัยที่ผ่านมา พบว่า การเตรียมข้อมูลในการทำเหมืองความคิดเห็น ขึ้นอยู่กับวิธีการจำแนกความคิดเห็นและคุณลักษณะที่ใช้ เช่น การเตรียมข้อมูลสำหรับวิธีการใช้คลังคำส่วนมากสกัดคำความคิดเห็นจากคลังคำและหน้าที่ของคำ ในบางงานมีการตัดคำ กำจัดคำหยุด ทำความสะอาดข้อความ และหารากคำศัพท์ก่อนทำการสกัดคำและหน้าที่ของคำ ทั้งนี้ การเตรียมข้อมูลไม่จำเป็นต้องทำทุกกระบวนการข้างต้น บางงานทำแค่บางกระบวนการ บางงานทำกระบวนการกำจัดคำหยุดและหารากคำศัพท์ก่อนการตัดคำ บางงานไม่มีการตัดคำ ไม่กำจัดคำหยุด ไม่มีกระบวนการทำความสะอาด และไม่มีกระบวนการหารากคำศัพท์ มีเพียงกระบวนการสกัดคำจากคลังคำความคิดเห็น แล้วใช้วิธีการเปรียบเทียบกับคลังคำความคิดเห็นที่มีอยู่เพื่อนำไปประเมินและสรุปขั้วความคิดเห็น เช่น Karamibekr [5] ใช้ POS tagging ระบุหน้าที่ของคำ จากนั้นสกัดคำที่มีหน้าที่เป็นกริยาช่วยและคำกริยาเป็นตัวแทนคุณลักษณะเพื่อเปรียบเทียบกับคลังคำแสดงความคิดเห็นแล้วนำไปประเมินและสรุปความคิดเห็น จากงานวิจัยข้างต้นพบว่า วิธีการใช้คลังคำเน้นการใช้คลังคำความคิดเห็นมาช่วยในการประเมินขั้วความคิดเห็นเป็นหลัก ไม่มีกระบวนการคำนวณค่าน้ำหนักคุณลักษณะแต่พิจารณาว่ามีคำที่เป็นความคิดเห็นเชิงบวกในประโยคกี่คำ มีคำที่เป็นความคิดเห็นเชิงลบกี่คำจากนั้นนำมาสรุปขั้วความคิดเห็นของประโยคหรือเอกสาร ส่วนการเตรียมข้อมูลสำหรับวิธีการเรียนรู้ของเครื่อง ส่วนมากเตรียมข้อมูลโดยการตัดคำ กำจัดคำหยุด ทำความสะอาดและหารากคำศัพท์ และคัดเลือกตัวแทนคุณลักษณะด้วยวิธีการสกัดคุณลักษณะ มีการคำนวณค่าน้ำหนักของคำที่ใช้เป็นตัวแทนคุณลักษณะ ซึ่งวิธีการที่นิยมในการคำนวณค่าน้ำหนัก คือ การแทนคำด้วยค่าการเกิดขึ้น

หรือไม่เกิดขึ้นของคำ (Boolean Weighting) ส่วนการเตรียมข้อมูลเพื่อการจำแนกความคิดเห็นด้วยวิธีการผสมผสาน ใช้คุณลักษณะที่ได้จากกระบวนการใช้คลังคำร่วมกับการเลือกตัวแทนคุณลักษณะจากการสกัดคุณลักษณะที่เหมาะสม

2.3 วิธีการจำแนกประเภทความคิดเห็น (Opinion Classifier Algorithm)

การจำแนกความคิดเห็น มี 3 วิธีการหลักที่ได้รับความนิยม ได้แก่ วิธีการใช้คลังคำ วิธีการเรียนรู้ของเครื่อง และวิธีการผสมผสาน รายละเอียดแต่ละวิธีการมีดังต่อไปนี้

2.3.1 วิธีการใช้คลังคำ (Lexical Based)

วิธีการใช้คลังคำ เป็นการจำแนกความคิดเห็นโดยใช้พจนานุกรมคำที่ระบุขั้วความรู้สึกของคำไว้แล้ว ในงานวิจัยด้านการทำเหมืองความคิดเห็น เรียกว่า คำแสดงความรู้สึก (Sentiment Word) คำแสดงความคิดเห็น (Opinion Words) คำระบุขั้วความเห็น (Polar Word) หรือ คำที่เป็นความคิดเห็น (Opinion Bearing Words) [28] พจนานุกรมที่ใช้ในการทำเหมืองความคิดเห็นส่วนมากจะมีคำแสดงความรู้สึก 2 ด้าน คือ คำแสดงความรู้สึกเชิงบวก (Positive Sentiment Words) และ คำแสดงความรู้สึกเชิงลบ (Negative Sentiment Words) คำแสดงความรู้สึกเชิงบวก ใช้ในการแสดงความรู้สึกในแง่ดี เช่น สวย (Beautiful) ดี (Good) เยี่ยม (Excellent) คำแสดงความรู้สึกในเชิงลบใช้ในการแสดงความรู้สึกที่ไม่ดี เช่น แย่ (Bad) เบื่อ (Bored) เหนื่อย (Tired) เป็นต้น วิธีการรวบรวมคำแสดงความคิดเห็น มี 3 วิธีการหลัก คือ การรวบรวมด้วยตนเอง (Manual Approach) การใช้พจนานุกรม (Dictionary-based Approach และ การใช้คลังคำศัพท์ (Corpus-based Approach) [28] แต่ละวิธีการแต่ละวิธีการมีรายละเอียดดังนี้

1) การรวบรวมด้วยตนเอง เป็นวิธีการรวบรวมคำศัพท์โดยวิธีการใช้คนทำเป็นหลัก ส่วนมากจะให้ผู้เชี่ยวชาญเข้ามามีส่วนร่วมในการรวบรวมคำศัพท์และสร้างความสัมพันธ์ระหว่างคำศัพท์ไว้ด้วยกัน วิธีการนี้อาจจะต้องใช้ผู้เชี่ยวชาญจำนวนมากและใช้เวลานาน งานวิจัยส่วนมากจึงนิยมใช้วิธีการนี้ร่วมกับวิธีการอื่นที่เป็นอัตโนมัติ เช่น การรวบรวมคำศัพท์พื้นฐานในขั้นตอนแรก หรือการตรวจสอบคำศัพท์ในขั้นตอนสุดท้ายที่วิธีการแบบอัตโนมัติอาจจะมีข้อผิดพลาดเกิดขึ้น วิธีการนี้มีขั้นตอนทำงานดังนี้ [26]

1.1) กำหนดขอบเขตของหัวข้อ เป็นการกำหนดขอบเขตให้กับหัวข้อหลักและหัวข้อย่อยที่จะทำการรวบรวม ซึ่งแต่ละหัวข้ออาจจะมีความสำคัญไม่เท่ากัน

1.2) การสะสมคำศัพท์ เป็นการรวบรวมคำศัพท์ที่มีความสัมพันธ์กัน ซึ่งคำศัพท์เหล่านั้นอาจจะได้มาจากแหล่งข้อมูลที่มีอยู่แล้ว เช่น วารสาร หนังสือ หรือพจนานุกรม เป็นต้น

1.3) การประมวลผลคำศัพท์ เป็นการรวบรวมคำศัพท์ที่ได้จากการสะสมคำศัพท์มาเก็บไว้ด้วยกันอย่างเป็นระบบ เพื่อกำจัดคำซ้ำและเพื่อรวมคำที่มีความหมายเหมือนกันหรือใกล้เคียงกันมาไว้ด้วยกัน

1.4) การจัดกลุ่มคำศัพท์

1.5) การจัดลำดับความสัมพันธ์ของคำศัพท์ในแต่ละกลุ่ม

1.6) การปรับเพิ่มคำศัพท์อื่น ๆ ที่เกี่ยวข้อง

1.7) การตรวจสอบคำศัพท์

1.8) การตรวจสอบและปรับแต่งรายการคำศัพท์แต่ละชุดให้ถูกต้องสมบูรณ์ จัดคำศัพท์ที่มีความสัมพันธ์กันไว้ด้วยกัน และจัดรูปแบบการแสดงรายการคำศัพท์ให้อยู่ในรูปแบบตามที่ต้องการ

2) การใช้พจนานุกรม เป็นวิธีการรวบรวมศัพท์พื้นฐาน โยงไปยังคำพ้อง (Synonyms) และคำตรงข้าม (Antonyms) โดยรวบรวมจากพจนานุกรมคำที่ได้รับการยอมรับ เช่น WordNet เพราะพจนานุกรมคำเหล่านั้นจะมีรายการคำพ้อง และคำตรงข้าม วิธีการนี้มีขั้นตอนทำงานดังนี้

2.1) รวบรวมคำแสดงความรู้สึกพื้นฐานที่บอกได้ว่าเป็นคำแสดงทัศนคติในเชิงบวกหรือเชิงลบ ขั้นตอนนี้ใช้วิธีการเก็บรวบรวมด้วยตนเองแบบง่าย ๆ

2.2) เพิ่มเติมชุดของคำโดยการค้นหาใน WordNet หรือ พจนานุกรมออนไลน์อื่น ๆ เพื่อค้นหาคำพ้องและคำตรงข้าม

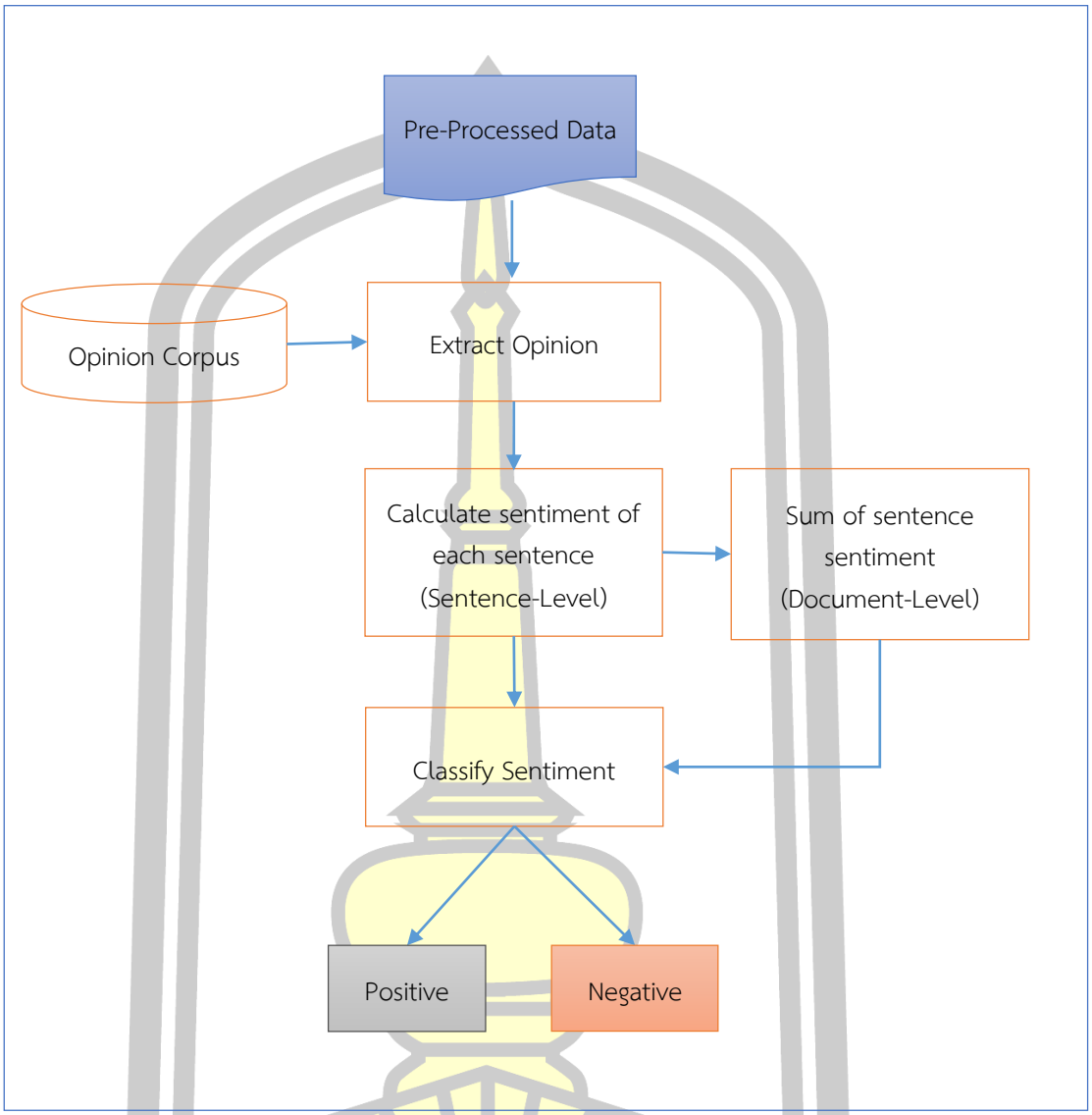
2.3) เพิ่มคำใหม่ที่ค้นพบลงในชุดคำที่สร้างขึ้นในกระบวนการแรก

2.4) ทำซ้ำทุกกระบวนการจนกระทั่งไม่พบคำใหม่

2.5) ตรวจสอบความถูกต้องของคำทั้งหมดและข้อผิดพลาดอื่น ๆ

3) การใช้คลังคำศัพท์ เป็นวิธีการที่นำมาใช้เพื่อแก้ปัญหาของคำที่ไม่มีในพจนานุกรม เช่น คำเฉพาะ คำกำกวม เป็นต้น จัดเป็นวิธีการที่มีความถูกต้องมากกว่า 2 วิธีการแรก แต่มีข้อจำกัด คือ ฐานข้อมูลต้องมีขนาดใหญ่มากพอ เพื่อให้ได้สถิติข้อมูลที่มีความน่าเชื่อถือ

การจำแนกความคิดเห็นด้วยวิธีการใช้คลังคำมี 3 กระบวนการหลัก ได้แก่ การสกัดความรู้สึกจากคลังคำความคิดเห็น คำนวณขั้วความคิดเห็น และสรุปขั้วความคิดเห็น ซึ่งสามารถสรุปในระดับประโยค (Sentence Level) หรือระดับเอกสาร (Document Level) ได้ ขั้นตอนการจำแนกความคิดเห็นแสดงดังรูปที่ 2



รูปที่ 2 การจำแนกความคิดเห็นด้วยวิธีการใช้คลังคำ

จากรูปที่ 2 เมื่อได้ข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลแล้ว จะนำคุณลักษณะที่ได้จากขั้นตอนการสกัดคุณลักษณะมาตรวจสอบขั้วความคิดเห็นโดยใช้คลังคำความคิดเห็นและแทนค่าคะแนนแต่ละคุณลักษณะตามขั้วความคิดเห็น เพื่อใช้ในการประเมินความคิดเห็น เช่น ถ้าคุณลักษณะอยู่ในกลุ่มความคิดเห็นเชิงบวก ให้มีค่าเท่ากับ +1 และคุณลักษณะที่อยู่ในกลุ่มความคิดเห็นเชิงลบ ให้มีค่าเท่ากับ -1 เป็นต้น จากนั้นคำนวณหาค่าผลรวมของคุณลักษณะที่ปรากฏในเอกสาร เพื่อพิจารณาว่าทั้งเอกสารเป็นความคิดเห็นเชิงบวกหรือเชิงลบ โดยกำหนดให้เอกสารข้อความประกอบด้วยหลายประโยค เขียนแทนด้วย $D = \{S_1, S_2, ..., S_m\}$ โดยที่ $S_1, S_2, ..., S_m$ คือ ประโยคในเอกสาร และ m คือ จำนวนประโยค ความคิดเห็นโดยรวมของเอกสารได้มาจากผลรวมความคิดเห็นของประโยคในเอกสาร ซึ่งแต่ละประโยคประกอบด้วยคุณลักษณะที่ใช้เป็นตัวแทน เขียนแทนด้วย

$S = \{o_1, o_2, ..., o_n\}$ โดยที่ $o_1, o_2, ..., o_n$ คือ คุณลักษณะในประโยค และ n คือ จำนวนคุณลักษณะ การคำนวณความคิดเห็นทั้งในระดับประโยคและระดับเอกสาร แสดงดังสมการ (2.5) และสมการ (2.6) ตามลำดับ

$$S_s = \sum_{i=1}^{n} SO_i \quad (2.5)$$

เมื่อ SO_i คือ คะแนนขั้วความคิดเห็นของคุณลักษณะที่ i

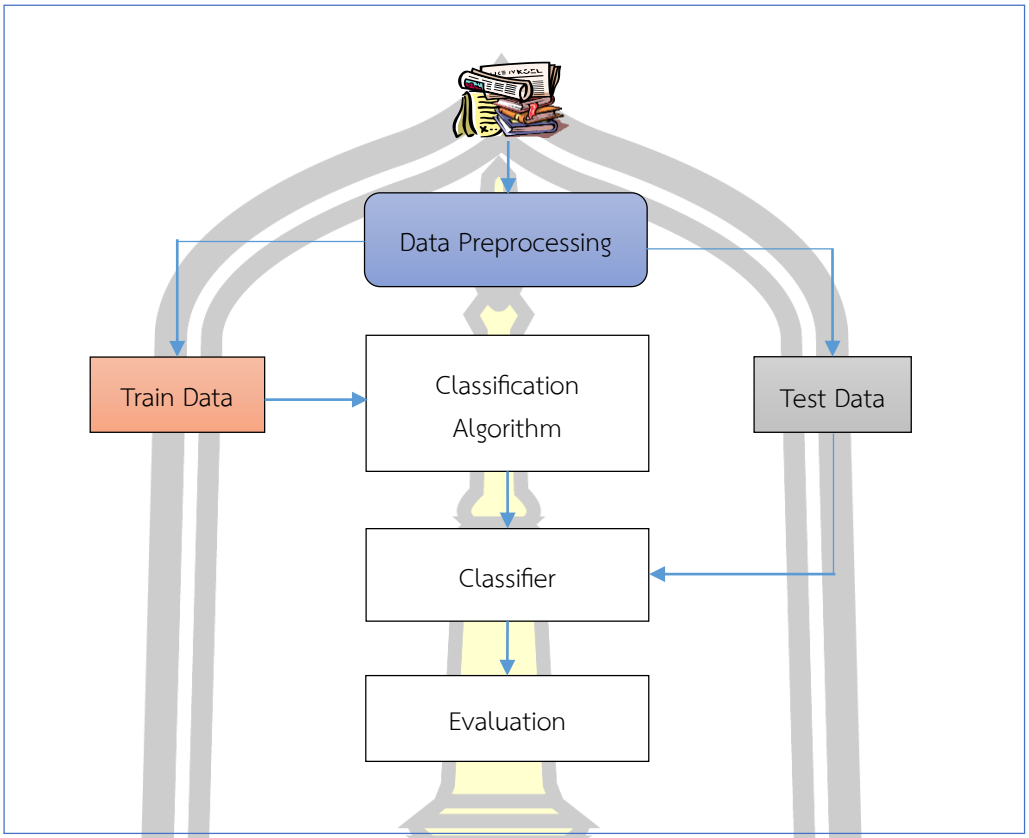
$$S_D = \sum_{j=1}^{m} SS_j \quad (2.6)$$

เมื่อ SS_j คือ คะแนนความคิดเห็นของประโยคที่ j

2.3.2 วิธีการเรียนรู้ของเครื่อง (Machine Learning)

วิธีการเรียนรู้ของเครื่อง เป็นวิธีการพัฒนากระบวนการเรียนรู้เพื่อให้เครื่องคอมพิวเตอร์สามารถทำงานได้อย่างมีประสิทธิภาพโดยอาศัยหลักการเรียนรู้ของมนุษย์ แบ่งเป็น 2 เทคนิคหลัก ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) แต่ละวิธีการอธิบายได้ดังนี้

1) การเรียนรู้แบบมีผู้สอน (Supervised Learning) หรือเรียกว่า การสร้างตัวแบบในการทำนาย (Predictive Modeling) เป็นเทคนิคที่เน้นการเรียนรู้จากข้อมูลที่มีอยู่ในอดีต เพื่อนำมาสร้างสมการหรือรูปแบบของชุดข้อมูลสอน (Training Set) สำหรับหาคำตอบให้กับข้อมูลชุดใหม่ (Test Set) การใช้ข้อมูลสอนจำนวนมากจะช่วยให้แบบจำลองการจำแนกความคิดเห็นมีความถูกต้องสูง แต่ระยะเวลาที่เครื่องใช้ในการสร้างแบบจำลองก็จะมากตามไปด้วย เทคนิคการเรียนรู้แบบมีผู้สอน [29] เกี่ยวข้องกับ 2 กระบวนการ คือ กระบวนการเรียนรู้รูปแบบจากข้อมูลชุดสอน (Training Data) และกระบวนการจำแนกประเภทข้อมูลชุดทดสอบ (Test Data) ดังแสดงในรูปที่ 3



รูปที่ 3 ขั้นตอนการจำแนกความคิดเห็นด้วยวิธีการเรียนรู้ด้วยเครื่อง

1.1) นาอีฟเบย์ (Naïve Bayes) เป็นวิธีการเรียนรู้แบบมีผู้สอนวิธีการหนึ่งที่จัดว่าง่ายและได้รับความนิยมเป็นอย่างมาก ในการจำแนกความคิดเห็น ใช้หลักการความน่าจะเป็นบนพื้นฐานทฤษฎีของเบย์ (Bayes' Theorem) การสร้างตัวจำแนกนาอีฟเบย์ จะพิจารณาจากความเป็นไปได้ของคุณลักษณะที่เกิดขึ้นในคลาสเปรียบเทียบจำนวนคุณลักษณะทั้งหมดที่เกิดขึ้นของข้อมูลชุดสอน [30] การจำแนกข้อมูลชุดทดสอบจะอาศัยค่าความน่าจะเป็นจากข้อมูลชุดสอน มาทำนายผลของข้อมูลชุดทดสอบ โดยใช้วิธีการคำนวณค่าความน่าจะเป็นของชุดข้อมูลที่อยู่ในแต่ละคลาส จนครบทุกคลาส ค่าความน่าจะเป็นของคลาสใดที่มากที่สุด คลาสนั้นจะถูกเลือกเป็นคลาสคำตอบ วิธีการคำนวณแสดงดังสมการ (2.7)

$$P(C) = \frac{Count(W \mid C)}{Count(W)} \qquad (2.7)$$

โดยที่ C คือ คลาสของความคิดเห็น เช่น คลาสความคิดเห็นเชิงบวก (Positive Class) คลาสความคิดเห็นที่เป็นเชิงลบ (Negative Class) เป็นต้น

$W|C$ คือ จำนวนคุณลักษณะที่อยู่ในคลาส

W คือ จำนวนคุณลักษณะที่อยู่ในเอกสาร

การจำแนกข้อความด้วยวิธีนาอีฟเบย์จะกำหนดให้แต่ละคุณลักษณะที่เกิดขึ้นเป็นอิสระต่อกัน ดังนั้น การคำนวณหาค่าความน่าจะเป็นของของคุณลักษณะที่เกิดขึ้นในคลาสจะเกิดจากผลรวมของคุณลักษณะ (W_i) ที่พบในคลาสนั้น แสดงดังสมการ (2.8)

$$P(C|W_1,W_2,...W_n) = \frac{P(C)\prod_{i=1}^{n} P(W_i|C)}{\prod_{i=1}^{n} P(W_i)} \quad (2.8)$$

โดยที่ $P(C|W_1,W_2,...W_n)$ คือ ความน่าจะเป็นของกลุ่มคุณลักษณะ W_i จำนวน n ตัว ที่เกิดขึ้นในคลาส C

$\prod_{i=1}^{n} P(W_i|C)$ คือ ผลคูณของความน่าจะเป็นของคุณลักษณะที่มีอยู่ในคลาส C

$\prod_{i=1}^{n} P(W_i)$ คือ ผลคูณของความน่าจะเป็นของคุณลักษณะทั้งหมดในเอกสาร

การหาคำตอบจากค่าความน่าจะเป็นที่แต่ละเอกสารที่อยู่ในคลาส จะใช้วิธีการคำนวณค่าความน่าจะเป็นของชุดของมูลที่อยู่ในแต่ละคลาส จนครบทุกคลาส แล้วทำการเปรียบเทียบค่าและเลือกคลาสคำตอบจากค่าความน่าจะเป็นของคลาสที่มีค่ามากที่สุด แสดงดังสมการ (2.9)

$$\text{argmax } P(C|W_i) \alpha \text{ argmax}[P(C)\prod_{i=1}^{n} P(W_i|C)] \quad (2.9)$$

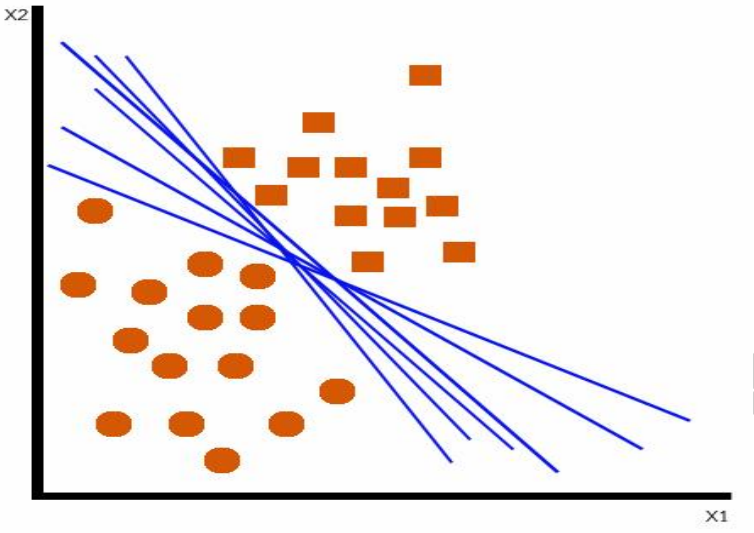
สำหรับคุณลักษณะที่พบในข้อมูลชุดสอนของคลาส ค่าความน่าจะเป็นจะเท่ากับ 0 ส่งผลให้ความน่าจะเป็นทั้งหมดจะมีค่าเท่ากับ 0 ด้วย จึงนิยมใช้วิธีการประมาณค่าความน่าจะเป็นของลาพลาเซียนสมูทติ้ง (Laplacian Smoothing) มาช่วยแก้ไขปัญหานี้ โดยการเพิ่มค่าความน่าจะเป็น เท่ากับ 1 หรือ α ให้แต่ละคุณลักษณะ จะทำให้ค่าความน่าจะเป็นไม่เป็น 0 แสดงดังสมการ (2.10)

$$P_{laplace}(W_i \mid W_{i-1}) = \frac{\alpha + C(w_{i-1}w_i)}{\alpha \mid V \mid + \sum_{w_i} C(w_{i-1}w_i)} \qquad (2.10)$$

โดยที่ $0 < \alpha \leq 1$ และ V คือ จำนวนคุณลักษณะทั้งหมดในเอกสารข้อมูลเรียนรู้

ข้อดีของวิธีนาอีฟเบย์ [30] คือ วิธีการคำนวณง่ายต่อการนำไปใช้ ได้ผลลัพธ์ที่สามารถนำไปประยุกต์ใช้ได้ดี แต่ข้อเสียคือ ใช้ได้กับแอนทริบิวที่เป็นอิสระกันเท่านั้น

1.2) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machines: SVM) [31] [27] เป็นวิธีการเรียนรู้แบบมีผู้สอนที่นิยมใช้ในการจัดกลุ่มข้อมูล ใช้แนวคิดการแบ่งข้อมูลด้วยการหาเส้นแบ่งที่เหมาะสม (Optimal Hyper Plane)



รูปที่ 4 ตัวอย่างการสร้างเส้นแบ่งกลุ่มข้อมูลของซัพพอร์ทเวกเตอร์แมชชีน

จากรูปที่ 4 จะเห็นว่าข้อมูลสามารถถูกแบ่งด้วยเส้นแบ่งมากกว่า 2 เส้น จะต้องหาเส้นแบ่งที่เหมาะสม คือ เส้นที่ทำให้ชุดข้อมูลทั้งสองกลุ่มมีระยะห่างระหว่างกันมากที่สุด (Maximum Margin Hyperplane: MMH) กำหนดให้ชุดข้อมูลเรียนรู้ $D = \{(\vec{x_i}, y_i)\}$ โดยที่ $\vec{x}_i(w_{i1}, w_{i2}, w_{i3}, .., w_{im})$ เป็นเวกเตอร์ของตัวแทนข้อความที่นำเข้า และแต่ละ $\vec{x}_i$ ถูกกำหนดไว้ด้วยคลาส y_i เมื่อ y_i เป็นค่า

จำนวนจริง มีค่าตั้งแต่ -1 ถึง +1 เส้นที่ถูกเลือกเพื่อใช้เป็นเส้นไฮเปอร์เพลน คือ เส้นที่ทำให้สมการ y_i มีค่าเท่ากับ 0 โดยมีเวกเตอร์ [27] ดังสมการ (2.11)

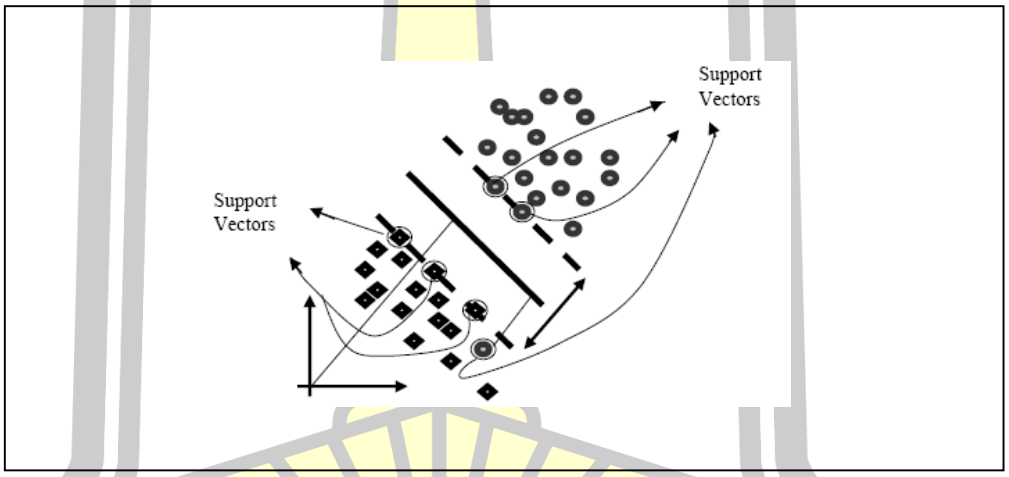
$$y = \begin{cases} +1, \vec{w} \cdot \vec{x} + b > 0 \\ -1, \vec{w} \cdot \vec{x} + b < 0 \end{cases} \quad (2.11)$$

โดยที่ w คือ เวกเตอร์ที่ตั้งฉากกับเส้นไฮเปอร์เพลน

$\vec{x}_i$ คือ ค่าเวกเตอร์ข้อมูล

b คือ ค่าโน้มเอียง (Bias)

เมื่อมีข้อมูล $\vec{x}$ เข้ามาใหม่ จะทำนายหาคลาส y จากชุดข้อมูลเรียนรู้ ($\vec{x}_i, y_i$) ที่มีค่าใกล้เคียงที่สุด



รูปที่ 5 การหาค่าซัพพอร์ทเวกเตอร์

ที่มา [31]

ข้อดีของซัพพอร์ทเวกเตอร์แมชชีน [27] คือ สามารถรองรับคุณลักษณะจำนวนมากได้ เนื่องจากใช้การแทนข้อมูลแบบเวกเตอร์และพิจารณาเส้นแบ่งกลุ่มข้อมูลจากซัพพอร์ตเวกเตอร์ (Support Vector) แต่ข้อเสีย คือ ต้องทดลองเพื่อปรับค่าพารามิเตอร์ให้เหมาะสมสำหรับแต่ละเคอร์เนล (Kernel) ที่เลือกใช้ บางครั้งอาจจะใช้เวลานาน

1.3) ต้นไม้ตัดสินใจ (Decision Tree) [29] เป็นอีกวิธีการหนึ่งที่ได้รับความนิยม เนื่องจากเป็นอัลกอริทึมที่เข้าใจง่าย ความผิดพลาดค่อนข้างน้อย การจำแนกข้อมูลทำโดยการสร้างตัวจำแนกเพื่อคัดเลือกแอตทริบิวต์ที่มีความสัมพันธ์กับคลาสมากที่สุดขึ้นมาเป็นโหนดบนสุดของ

ต้นไม้ หลังจากนั้นก็จะหาแอตทริบิวต์ถัดไปเรื่อย ๆ โครงสร้างต้นไม้ตัดสินใจประกอบด้วย โหนดภายใน (Internal Node) ที่ใช้ในการตัดสินใจ และโหนดใบ (Leaf Node) ที่เป็นกลุ่มคลาส (Class) ที่ได้จากการจำแนกข้อมูลตามคุณสมบัติ อัลกอริทึมที่นิยมใช้ในการสร้างต้นไม้ตัดสินใจ คือ ID3 และ C4.5 อัลกอริธึม C4.5 ได้พัฒนาเพิ่มเติมจากอัลกอริธึม ID3 เพื่อให้มีประสิทธิภาพมากขึ้น เช่น สามารถใช้กับข้อมูลแบบต่อเนื่อง (Continuous Data) และแบบไม่ต่อเนื่อง (Discrete Data) สามารถใช้งานได้กับข้อมูลที่เกิดการขาดหายได้ (Missing Data) และสามารถลดขนาดของต้นไม้โดยการตัดกิ่งที่ไม่จำเป็นออกไป โดยไม่ทำให้ความถูกต้องลดลง เป็นต้น

ในการหาความสัมพันธ์ของคุณลักษณะ จะใช้ค่าสารสนเทศ (Information Gain) เป็นตัววัด สามารถคำนวณดังสมการ (2.12)

$$Gain(S,A) = E(S) - \sum_{v=value(A)} \frac{|S_v|}{|S|} E(S_v) \qquad (2.12)$$

โดยที่ S คือ เซตของเอกสารทั้งหมด

A คือ คุณลักษณะที่พิจารณา

$value(A)$ คือ ค่าที่เป็นไปได้ของคุณลักษณะ A

S_v คือ เซตของเอกสารทั้งหมดที่มีคุณลักษณะ A ที่มีค่า v

E คือ ค่าเอนโทรปีของข้อมูล คำนวณได้ดังสมการ (2.13)

$$E(S) = -\sum_{i=1}^{n} P(V_i) \log_2 P(V_i) \qquad (2.13)$$

โดยที่ $P(V_i)$ คือ ความน่าจะเป็นของเอกสารที่จะอยู่ในคลาส i

n คือ จำนวนคลาสทั้งหมด

ในการจำแนกความคิดเห็นด้วยวิธีต้นไม้ตัดสินใจ โครงสร้างของต้นไม้จะประกอบไปด้วยโหนด (Node) หมายถึง คุณลักษณะ โดยอาจจะหมายถึง คำ ประโยค หรือวลี ก็ได้ และโหนดสุดท้ายคือคลาสของเอกสาร การจำแนกข้อมูลจะทดสอบจากค่าน้ำหนักของคุณลักษณะไปเรื่อยๆ จนกว่าจะถึงโหนดคลาส โดยทั่วไปจะใช้ค่าน้ำหนักแบบ Boolean Weighting

1.4) เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor: K-NN) การจำแนกประเภทข้อความด้วยวิธีเพื่อนบ้านใกล้ที่สุดจะขึ้นกับข้อมูลที่มีคุณสมบัติใกล้เคียง K ตัว หลักการ

ทำงาน คือ ทำการวัดระยะห่างระหว่างข้อมูลที่ต้องการทำนายกับข้อมูลที่อยู่ใกล้เคียง จำนวน K ตัว คำตอบที่ทำนายได้ คือ คลาสที่พบมากที่สุดของข้อมูลที่เป็นเพื่อนบ้านทั้ง K ตัว [29] การวิธีการ K-NN มาใช้กับการจำแนกความคิดเห็นโดยทั่วไปจะกำหนดค่า K จะถูกกำหนดให้เป็นเลขคี่ วิธีการ K-NN เหมาะสำหรับการจำแนกข้อมูลที่เป็นตัวเลขและไม่เหมาะกับข้อมูลที่มีแอททริบิวจำนวนมาก เนื่องจากเสียเวลาในการคำนวณหาค่าความห่างเพื่อจำแนกข้อมูล สำหรับการจำแนกข้อมูล มักจะใช้วิธีการวัดระยะทางแบบ Euclidean Distance คือ การหาค่ารากที่สองของผลต่างระหว่างคุณลักษณะแต่ละตัวยกกำลังสอง ดังสมการ (2.14)

$$d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2.14)$$

โดยที่ x_1 คือ ค่าของคุณลักษณะที่ 1 ของข้อมูลที่ต้องการจำแนก

y_1 คือ ค่าของคุณลักษณะที่ 1 ของข้อมูลเพื่อนบ้านที่นำมาพิจารณา

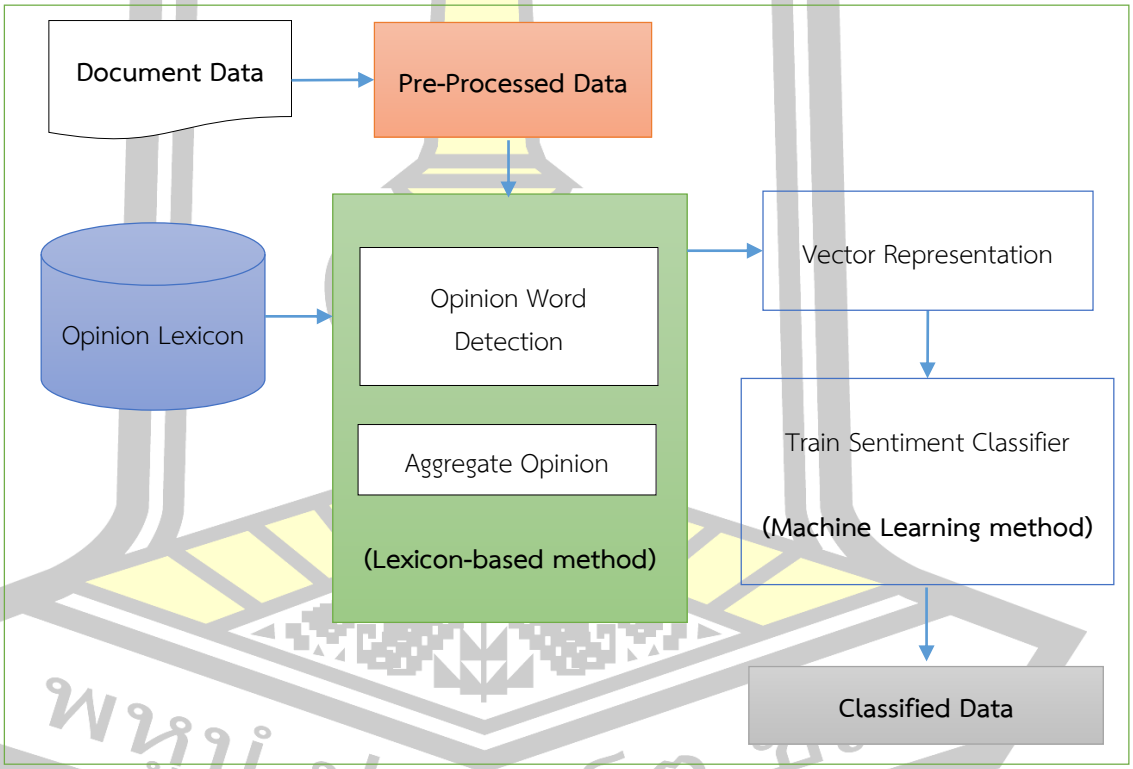
n คือ จำนวนคุณลักษณะทั้งหมด

งานวิจัยที่จำแนกความคิดเห็นโดยใช้วิธีการเรียนรู้แบบมีผู้สอน เช่น Troussas และ Virvou [1] จำแนกความคิดเห็นจากข้อความสถานะ (Status) ที่อยู่บน Facebook ด้วยวิธีการนาอีฟเบย์เปรียบเทียบกับวิธีการ Rocchio และ Perceptron พบว่า วิธีการนาอีฟเบย์ ให้ค่าความถูกต้องในการจำแนกความคิดเห็นสูงกว่าวิธีการ Perceptron และวิธีการอื่น ๆ Akaichi และคณะ [2] จำแนกความคิดเห็นจากข้อความและข้อความที่เป็นสัญลักษณ์พิเศษ (Emotion) ที่อยู่บนเฟสบุค โดยใช้วิธีการคลังคำร่วมกับนาอีฟเบย์ แล้วเปรียบเทียบกับวิธีการซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) พบว่า ค่าความถูกต้องในการจำแนกความคิดเห็นของวิธีการนาอีฟเบย์สูงกว่าวิธีการซัพพอร์ตเวกเตอร์แมชชีน เมื่อจำนวนคุณลักษณะ (Feature) ไม่มาก แต่ประสิทธิภาพการจำแนกความคิดเห็นของวิธีการนาอีฟเบย์น้อยกว่าวิธีการซัพพอร์ตเวกเตอร์แมชชีน เมื่อจำนวนคุณลักษณะเพิ่มขึ้น Ortigosa [4] นำเสนอการจำแนกความคิดเห็นของผู้เรียนเพื่อนำมาประยุกต์ใช้สำหรับปรับปรุงระบบการเรียนการสอนออนไลน์ (e-Learning) ให้เหมาะสมกับผู้เรียน โดยเปรียบเทียบการจำแนกความคิดเห็นด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีนร่วมกับการใช้คลังคำ กับ 3 วิธีการ คือ 1) ใช้คลังคำอย่างเดียว 2) ใช้คลังคำร่วมกับนาอีฟเบย์ 3) ใช้คลังคำร่วมกับต้นไม้ตัดสินใจ พบว่า การจำแนกความคิดเห็นโดยวิธีการใช้คลังคำร่วมกับซัพพอรต์เวกเตอร์แมชชีนให้ค่าความถูกต้องสูงกว่าวิธีการอื่น Anjaria และ Guddeti [3] จำแนกความคิดเห็นที่อยู่บนทวิตเตอร์ เพื่อทำนายผลการเลือกตั้ง ซึ่งได้จำแนกความคิดเห็นด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีน เปรียบเทียบกับวิธีการนาอีฟเบย์ แม็กซิมัมเอ็นโทรปี (Maximum Entropy) และอาร์ทิฟิเชียลนิวรอลเน็ตเวิร์ค

(Artificial Neural Networks) พบว่า วิธีการซัพพอร์ตเวกเตอร์แมชชีนมีความถูกต้องในการทำนายสูงกว่าวิธีการอื่น Basari และคณะ [32] จำแนกความคิดเห็นเกี่ยวกับภาพยนตร์ (Movie Review) ด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีน เปรียบเทียบกับ วิธีซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธีพาทิเคิลสวอร์มออฟติไมเซชั่น (Particle Swarm Optimization) พบว่า การจำแนกความคิดเห็นด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธีพาทิเคิลสวอร์มออฟติไมเซชั่นมีค่าความถูกต้องสูงกว่าการจำแนกความคิดเห็นด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีนอย่างเดียว

2.3.3 วิธีการผสมผสาน (Hybrids Methodology)

การจำแนกความคิดเห็นด้วยวิธีการผสมผสาน เป็นการผสมผสานระหว่างวิธีการใช้คลังคำกับวิธีการเรียนรู้ของเครื่อง โดยใช้วิธีการสกัดคำตัวแทนคุณลักษณะจากคำในเอกสารและคลังคำความคิดเห็นที่มีอยู่ เพื่อจำแนกความคิดเห็นด้วยวิธีการเรียนรู้ของเครื่อง ดังรูปที่ 6



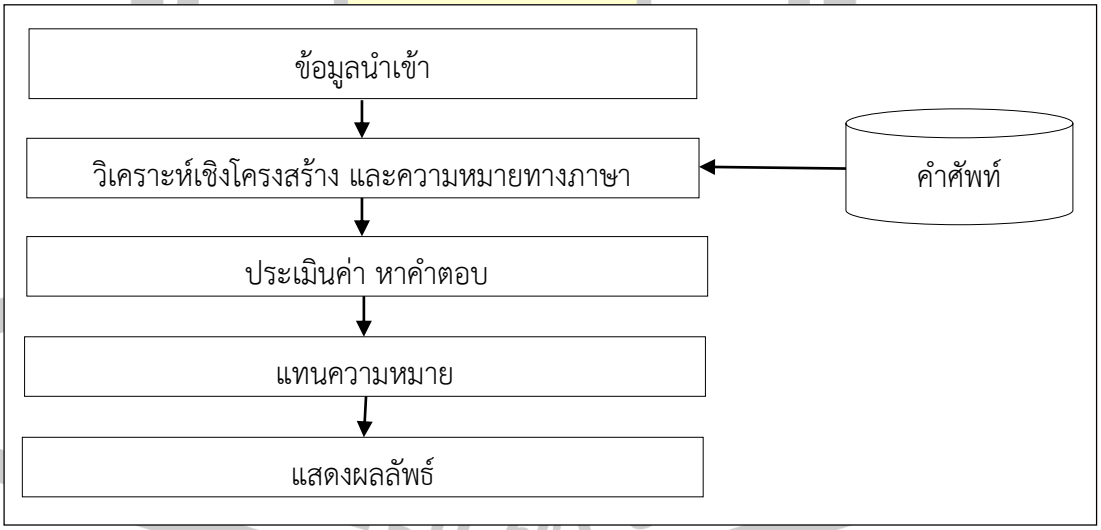
รูปที่ 6 ขั้นตอนวิธีการผสมผสาน

จากรูปที่ 6 แสดงขั้นตอนวิธีการจำแนกความคิดเห็นโดยวิธีการผสมผสานระหว่างการใช้คลังคำและเรียนรู้ของเครื่อง โดยการเตรียมข้อมูลเพื่อสกัดคุณลักษณะจากเอกสารนำเข้า เพื่อใช้เป็น

ตัวแทนคุณลักษณะและใช้ร่วมกับคลังคำความคิดเห็นที่มีอยู่ จากนั้นคำนวณขั้วความคิดเห็นของคุณลักษณะตัวแทนข้างต้น เพื่อแทนค่าในรูปแบบเวกเตอร์แล้วจึงจำแนกด้วยวิธีการเรียนรู้ของเครื่อง

2.4 การประมวลผลภาษาธรรมชาติ

การทำเหมืองความคิดเห็นอาศัยหลักการประมวลผลภาษาธรรมชาติในการเตรียมข้อมูล เนื่องจากความคิดเห็นอยู่ในรูปแบบข้อความที่คอมพิวเตอร์ไม่สามารถเข้าใจได้ จึงต้องใช้การประมวลผลภาษาธรรมชาติมาช่วยในการแปลงข้อความที่มนุษย์สื่อสารหรือที่เรียกว่าภาษาธรรมชาติ (Natural Language) ซึ่งมีรูปแบบที่ไม่แน่นอน ให้อยู่ในรูปแบบของภาษาที่คอมพิวเตอร์เข้าใจได้ มีโครงสร้างแน่นอน มีรูปแบบไวยากรณ์และการตีความหมายที่ชัดเจน (Formal Language) การที่จะทำให้คอมพิวเตอร์เข้าใจภาษาธรรมชาติของมนุษย์นั้น อาศัยหลักการปัญญาประดิษฐ์ (Artificial Intelligence : AI) และ หลักภาษาศาสตร์ (Linguistics) การประมวลผลภาษาธรรมชาติสามารถนำไปประยุกต์ใช้ในงานด้านต่าง ๆ ที่เกี่ยวกับภาษาได้ เช่น นำมาช่วยในการวิเคราะห์เอกสารต่างๆ ว่าเกี่ยวข้องกับเรื่องใด เพื่อช่วยในการค้นคืนข้อมูล (Information Retrieval) หรือ นำมาช่วยให้คอมพิวเตอร์สามารถสรุปประเด็นสำคัญ (Information Summarization) ที่อยู่ในเอกสารได้ การประมวลผลภาษาธรรมชาติประกอบด้วย 2 ส่วน [33] คือ การทำความเข้าใจภาษา (Natural Language Understanding: NLU) และการสร้างภาษา (Natural Language Generation: NLG) ขั้นตอนการทำงานของภาษาธรรมชาติสามารถแสดงได้ดังรูปที่ 7



รูปที่ 7 ขั้นตอนการทำงานของภาษาธรรมชาติ

ที่มา : [33]

2.4.1 ระดับการวิเคราะห์ภาษาธรรมชาติ

1) การวิเคราะห์ระดับการผสมคำ (Morphological Level) เป็นการนำคำหรือวลีมาวิเคราะห์ เพื่อแบ่งย่อยหรือต่อเข้ากันให้เกิดเป็นคำที่มีความหมาย

2) การวิเคราะห์ระดับไวยากรณ์ (Syntactic Analysis) เป็นการสำรวจหน้าที่ของคำในประโยค รวมถึงความสัมพันธ์ของคำในประโยคเพื่อให้ทราบว่าคำใดทำหน้าที่เป็น ประธาน (Subject) กริยา (Verb) หรือ กรรม (Object) ของประโยค

3) การวิเคราะห์ระดับความหมายของคำ (Semantic Level) เป็นการสำรวจความหมายของคำที่นำมาใช้ในประโยค เนื่องจากคำบางคำสามารถมีความหมายได้หลายความหมายขึ้นอยู่กับบริบทที่ถูกใช้ในประโยคนั้น ๆ

4) การวิเคราะห์ระดับโครงสร้าง (Discourse Level) เป็นการสำรวจเกี่ยวกับโครงสร้างของข้อความและโครงสร้างของเอกสาร โดยดูจากคำหรือประโยคข้างเคียง เช่น คำที่มีความหมายเชิงลบเมื่อตามด้วยคำที่มีความหมายเชิงบวก จะกลายเป็นความหมายเชิงลบ เช่น ไม่ (No) เมื่อรวมกับคำว่า ดี (Good) มีความหมายในเชิงลบ คือหมายถึงไม่ดี ส่วนคำที่มีความหมายเชิงลบเมื่อรวมกับความที่มีความหมายเชิงลบ จะหมายถึงความหมายในเชิงบวก เช่น คำว่า ไม่ เมื่อรวมกับคำว่า เลว มีความหมายในเชิงบวก คือ ไม่เลว เป็นต้น

2.4.2 เทคนิคการประมวลผลภาษาธรรมชาติ

เทคนิคการประมวลผลภาษาธรรมชาติ เป็นการวิเคราะห์คำที่อยู่ในประโยคเพื่อหาคำที่จะใช้เป็นคำหลัก โดยการเปรียบเทียบกับคำที่มีอยู่แล้วในคลังคำศัพท์ แต่ถ้าคำหลักที่มีอยู่แล้วไม่ตรงกับความต้องการ ผู้ใช้สามารถเพิ่มคำหลักลงในคลังคำศัพท์ได้เพื่อให้ระบบสามารถวิเคราะห์ได้ตรงกับการใช้งานมากที่สุด เทคนิคการประมวลผลภาษาธรรมชาติที่ได้รับความนิยมในปัจจุบันมี 2 วิธีการ ได้แก่ การวิเคราะห์คำหลัก (Keyword Analysis) และการวิเคราะห์ไวยากรณ์ มีรายละเอียดดังต่อไปนี้ [33]

1) การวิเคราะห์คำหลัก (Keyword Analysis) เป็นวิธีการวิเคราะห์คำจากประโยคเพื่อให้ทราบว่าคำใดที่จะใช้เป็นคำหลัก โดยการเปรียบเทียบกับคำที่ถูกจัดเก็บไว้ในฐานความรู้ แต่ถ้าคำหลักที่มีอยู่ไม่ตรงกับความต้องการของผู้ใช้นั้น ผู้ใช้สามารถเพิ่มคำหลักลงในฐานความรู้ได้เพื่อช่วยให้ระบบสามารถวิเคราะห์ได้ตรงกับผู้ใช้มากที่สุด โดยขั้นตอนพื้นฐานของกระบวนการวิเคราะห์โดยใช้คำหลักในการประมวลผลภาษาธรรมชาติจะเกี่ยวข้องกับขั้นตอนการแบ่งคำแต่ละคำในประโยค (Parsing) และ ขั้นตอนการจับคู่รูปแบบของคำ (Pattern Matching)

2) การวิเคราะห์ไวยากรณ์ (Syntactic Analysis) วิธีการนี้เป็นวิธีการวิเคราะห์ประโยคแบบละเอียด ดำเนินการเกี่ยวกับการวิเคราะห์ไวยากรณ์ ความหมาย และความสัมพันธ์ระหว่างคำในประโยค เพื่อนำไปประมวลผล และแสดงผลลัพธ์ให้กับผู้ใช้ วิธีการนี้มีการวิเคราะห์รูปประโยค วิเคราะห์ความหมายของประโยค หาความหมายของคำในประโยค วิเคราะห์หาความหมายที่แท้จริงของประโยค และทำการทดลองเชื่อมประโยคหลายประโยคเข้าด้วยกัน

2.4.3 องค์ประกอบของภาษาธรรมชาติ

1) ตัววิเคราะห์ (Parser) เป็นองค์ประกอบหนึ่งที่ทำหน้าที่วิเคราะห์ไวยากรณ์ของประโยคที่ถูกป้อนเข้าสู่ระบบ ซึ่งแบ่งประโยคออกเป็นคำ โดยคำแต่ละคำจะถูกนำไปจับคู่กับโครงสร้างที่ใช้ในการวิเคราะห์ชนิดของคำ ทั้งนี้โครงสร้างดังกล่าวถูกออกแบบไว้ในลักษณะของต้นไม้เรียกว่า Parser Tree

2) พจนานุกรม (Lexicon) เป็นพจนานุกรมที่ใช้ประกอบการวิเคราะห์ความหมาย ซึ่งประกอบด้วยการสะกดคำที่ถูกต้อง ความหมาย และหน้าที่ของคำนั้น ๆ ในประโยค กรณีที่คำมีความหมายมากกว่าหนึ่งความหมาย พจนานุกรมจะแสดงความหมายทั้งหมดออกมาให้

3) ส่วนการทำความเข้าใจ (Under Stander) เป็นส่วนที่ทำงานร่วมกับฐานความรู้เพื่อกำหนดความหมายของประโยคทั้งประโยค โดยจะใช้พาร์เซอร์ทรีในการค้นหาและอ้างอิงข้อมูลในฐานความรู้

4) ฐานความรู้ (Knowledge Based) เป็นที่จัดเก็บความรู้ต่าง ๆ ที่เกี่ยวข้องกับภาษาธรรมชาติที่มนุษย์ใช้ในการสื่อสาร ซึ่งฐานความรู้จะช่วยอำนวยความสะดวกให้กับระบบในการกำหนดความหมายของประโยคที่ผู้ใช้ป้อนเข้าสู่ระบบ

5) ส่วนจัดทำโครงสร้างข้อมูล (Generator) เมื่อส่วนทำความเข้าใจทำการค้นหาความหมายของประโยคได้แล้วจะสร้างโครงสร้างข้อมูลของประโยคได้ และจะถูกจัดเก็บไว้ในหน่วยความจำของระบบ

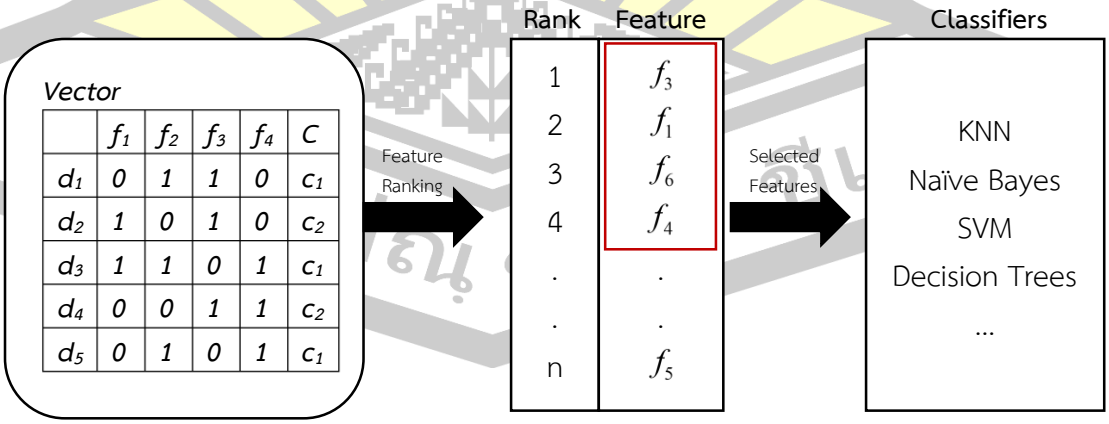
2.5 การคัดเลือกคุณลักษณะสำหรับทำเหมืองข้อความ (Feature Selection for Text Mining)

กระบวนการสำคัญในการทำเหมืองข้อความ คือ กระบวนการเตรียมข้อมูลเพื่อให้ข้อมูลที่อยู่ในรูปแบบข้อความซึ่งไม่มีโครงสร้างเป็นข้อมูลที่อยู่ในรูปแบบที่มีโครงสร้างเพื่อนำไปประมวลผลได้ แต่ด้วยปริมาณข้อมูลที่มีอยู่เป็นจำนวนมาก หากจำนวนคุณลักษณะที่ใช้มีมาก จะส่งผลต่อประสิทธิภาพในการประมวลผล เนื่องจากอัลกอริทึมในการสร้างแบบจำลองการวิเคราะห์ข้อมูลโดยทั่วไป ไม่สามารถรองรับการทำงานกับจำนวนคุณลักษณะที่มีจำนวนมากได้ดี [34] เวกเตอร์ที่แปลงจากรูปแบบข้อความโดยทั่วไปพบว่ามีคุณลักษณะจำนวนมาก เนื่องจากคุณลักษณะได้มาจากคำ

ที่อยู่ในเอกสารข้อความทั้งหมด ถึงแม้ในการเตรียมข้อมูลจะมีกระบวนการกำจัดคำหยุด การหารากศัพท์อยู่แล้ว แต่พบว่าคุณลักษณะยังมีจำนวนมากเมื่อเทียบกับจำนวนเอกสาร และเนื่องจากคุณลักษณะแต่ละตัวมีความสำคัญไม่เท่ากัน คุณลักษณะจำนวนมากที่ได้จากเอกสารมีทั้งคุณลักษณะที่สำคัญ ซึ่งจำเป็นต่อการจำแนก และอาจจะมีคุณลักษณะที่ไม่จำเป็น (Less Informative) หรือเป็นคุณลักษณะที่ซ้ำซ้อน (Redundant Features) ซึ่งอาจทำให้ตัวจำแนกเรียนรู้ผิดพลาดและส่งผลต่อประสิทธิภาพการจำแนกได้ ดังนั้น การคัดเลือกคุณลักษณะจึงเป็นกระบวนการที่สำคัญอย่างหนึ่งเพื่อให้ได้คุณลักษณะของข้อมูลที่ดีเพื่อใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้ก่อนจะสร้างตัวจำแนก ลดจำนวนคุณลักษณะที่มีจำนวนมากและทำให้ประสิทธิภาพในการประมวลผลดีขึ้น ปัจจุบันได้มีการนำเสนอวิธีการคัดเลือกคุณลักษณะหลายแบบ [6] [7] [8] [9] เช่น วิธีฟิลเตอร์ (Filter Model) วิธีแรปเปอร์ (Wrapper Model) วิธีฝังตัว (Embedded Model) และวิธีผสมผสาน (Hybrid Model) การเลือกคุณลักษณะสำหรับงานด้านการทำเหมืองข้อความส่วนใหญ่เป็นแบบ Filter Model เนื่องจากมีความง่ายและมีประสิทธิภาพ [10]

2.5.1 วิธีฟิลเตอร์ (Filter models)

วิธีฟิลเตอร์เป็นการคัดเลือกเลือกคุณลักษณะโดยใช้ค่าน้ำหนักของคุณลักษณะและคลาส โดยการคำนวณหาค่าน้ำหนักของแต่ละคุณลักษณะหรือค่าน้ำหนักที่เกิดจากความสัมพันธ์ระหว่างคุณลักษณะกับคลาสต่าง ๆ จากนั้นจะทำการเรียงคุณลักษณะตามค่าน้ำหนักจากมากไปน้อย โดยคุณลักษณะที่มีค่าน้ำหนักมากที่สุดแสดงว่ามีความสำคัญมากที่สุด และจะเลือกคุณลักษณะที่มีค่าน้ำหนักมากไปใช้ในการสร้างตัวจำแนก เทคนิคในการคำนวณค่าน้ำหนักของคุณลักษณะมีหลายวิธี เช่น Mutual Information (MI) Information Gain (IG) Chi-Square (Chi2) Gini Index Fisher Score และ ReliefF เป็นต้น ภาพประกอบแสดงวิธีการทำงานของ Filter Model แสดงดังรูปที่ 8



	f_1	f_2	f_3	f_4	C
d_1	0	1	1	0	c_1
d_2	1	0	1	0	c_2
d_3	1	1	0	1	c_1
d_4	0	0	1	1	c_2
d_5	0	1	0	1	c_1

Rank	Feature
1	f_3
2	f_1
3	f_6
4	f_4
.	.
.	.
n	f_5

รูปที่ 8 แสดงขั้นตอนของ Filter Model

จากรูปที่ 8 แสดงขั้นตอนของวิธีการ Filter Model โดยเริ่มจากนำเข้าเวกเตอร์ข้อมูล ซึ่งแต่ละแถวในเวกเตอร์แทนด้วยเอกสารและแต่ละคอลัมน์แทนด้วยคุณลักษณะ ทำการคำนวณค่าน้ำหนักของคุณลักษณะจากการปรากฏในเอกสารและพิจารณาความสัมพันธ์ระหว่างแต่ละคุณลักษณะกับคลาสต่าง ๆ จากนั้นจะได้คุณลักษณะที่เรียงลำดับตามค่าน้ำหนักมากสุดไปน้อยสุดเพื่อนำไปใช้งานในการจำแนกข้อมูลต่อไป

เทคนิคในการคำนวณค่าน้ำหนักของคุณลักษณะมีหลายวิธี เช่น ค่าสารสนเทศร่วม (Mutual Information : MI) Information Gain (IG) Chi-Square (Chi2) Fisher Score, Gini index และ ReliefF เป็นต้น ในงานวิจัยนี้ผู้วิจัยพัฒนาขั้นตอนวิธีการใหม่สำหรับคำนวณค่าน้ำหนักของคุณลักษณะและเลือกคุณลักษณะที่มีความจำเป็นต่อการจำแนกความคิดเห็นเปรียบเทียบกับวิธีการเลือกคุณลักษณะพื้นฐานที่ได้รับความนิยมในงานทางด้านเหมืองข้อความ ได้แก่ ค่าการเพิ่มของข้อมูล (Information Gain: IG) ค่าสถิติไคสแคว์ (Chi-Square) และค่า Gini Index แต่ละเทคนิคมีรายละเอียดดังนี้

1) ค่าการเพิ่มสารสนเทศ (Information Gain : IG) [35] เป็นเครื่องมือวัดค่าความสำคัญของคุณลักษณะอีกวิธีการหนึ่งที่ได้รับความนิยมเป็นอย่างมากในงานด้านเหมืองข้อความ ค่าการเพิ่มสารสนเทศได้นำไปประยุกต์ใช้ในอัลกอริทึม ID3 และ C4.5 โดยทำการเลือกคุณลักษณะสำหรับแบ่งข้อมูลจากคุณลักษณะที่มีค่าการเพิ่มสารสนเทศสูงที่สุด การหาค่าการเพิ่มสารสนเทศใช้การปรากฏของคำที่มีในเอกสาร [36] [37] คำนวณได้ดังสมการ (2.16)

$$\begin{aligned} IG(t_i \mid C) = & -\sum_k p(c_k)\log_2 p(c_k) \\ & + p(t_i)\sum_k p(c_k \mid t_i)\log_2 p(c_k \mid t_i) \\ & + p(\bar{t_i})\sum_k p(c_k \mid \bar{t_i})\log_2 p(c_k \mid \bar{t_i}) \end{aligned} \qquad (2.16)$$

เมื่อ $p(t_i)$ คือ ความน่าจะเป็นของการพบเทอม t_i ในเอกสาร

$p(\bar{t_i})$ คือ ความน่าจะเป็นของการไม่พบเทอม t_i ในเอกสาร

$p(c_k)$ คือ ค่าความน่าจะเป็นของคลาส k

$p(c_k \mid t_i)$ คือ ค่าความน่าจะเป็นของคลาส k เมื่อพบคำ t_i

ขั้นตอนการหาค่าการเพิ่มสารสนเทศแสดงได้ดังรูปที่ 9

Algorithm 1: Information Gain Feature Ranking

1. S = 0
2. For each $c_k \in C$ do:
3. calculate $p(c_k)$;
4. $H_c = S + p(c_k) \times \log 2(p(c_k))$
5. $S \leftarrow H_c$
6. End For
7. For each $e_i \in E$
8. Calculate $p(e_i)$;
9. $Sum = S + P(e_i) \times \log 2(p(e_i))$;
10. $S \leftarrow Sum$;
11. End For
12. For each class $c_k \in C$ do:
13. For each term $e_i \in E$ do:
14. Calculate $p(c_k \mid e_i)$;
15. $M = S + p(c_k \mid e_i) \times \log_2 p(c_k \mid e_i)$;
16. $S \leftarrow M$;
17. End For
18. End For
19. $H(C \mid E) = (-1) \times Sum \times (-1) \times M$;
20. $IG = H_c - H(C \mid E)$

รูปที่ 9 ขั้นตอนการหาค่าการเพิ่มสารสนเทศ

2) ค่าสถิติไคสแคว์ (Chi-Square) นอกจากเทคนิคการเลือกคุณลักษณะโดยการหาค่าการเพิ่มของข้อมูลข้างต้นแล้ว การหาค่าสถิติไคสแคว์เป็นอีกเทคนิคหนึ่งที่ได้รับความนิยมในการทำเหมืองความคิดเห็นเช่นกัน เป็นการวัดความสัมพันธ์ระหว่างคุณลักษณะ (t) และ คลาส (c) [38] [39] [39] โดยค่าไคสแคว์สามารถคำนวณได้จากสมการ (2.17)

$$\chi^2(t_i, c_k) = \sum \frac{(observed(t_i, c_k) - \exp ected(t_i, c_k))^2}{\exp ected(t_i, c_k)} \qquad (2.17)$$

เมื่อ $observed(t_i, c_k)$ คือ จำนวนเอกสารที่ปรากฏคุณลักษณะ t_i ในคลาส c_k

$\exp ected(t_i, c_k)$ คือ ความน่าจะเป็นของการเกิดคุณลักษณะ t_i ในคลาส c_k หาได้จากสมการ (2.18)

$$\exp ected(t_i, c_k) = prob(c_k) \times observed(t_i, c_k) \quad (2.18)$$

เมื่อ $prob(c_k)$ คือ ค่าความน่าจะเป็นของการเกิดคลาส k ในเอกสารทั้งหมด หาได้จากสมการ (2.19)

$$prob(c_k) = \frac{n(d_j, c_k)}{N} \quad (2.19)$$

ขั้นตอนการหาค่าไคสแคว์ แสดงได้ดังรูปที่ 10

Algorithm 2: Chi-Square Feature Ranking

1. for each class $c_k \in C$ do:
2. for document $d_j \in D$ do:
3. if d_j in c_k do:
4. $n_k \mathrel{+}= 1$;
5. for each term $t_i \in T$ do:
6. if t_i in c_k do:
7. $Observed(t_i, c_k) \mathrel{+}= 1$;
8. end for
9. end for
10. $Chi^2(t_i, c_k) = ((Observed(t_i, c_k) - Expected(t_i, c_k))^2) / Expected(t_i, c_k)$;
11. $prob(c_k) = n_2 / (n_1 + n_2)$;
12. end for
13. for each class $t_i \in T$ do:
14. for each term $c_k \in C$ do:
15. $Expected(t_i, c_k) = \Pr ob(c_k) \times Observed(t_i, c_k)$;
16. $Chi^2(t_i, c_k) = ((Observed(t_i, c_k) - Expected(t_i, c_k))^2) / Expected(t_i, c_k)$;
17. end for
18. $Chi^2(t_i) = Chi^2(t_i, c_1) + Chi^2(t_i, c_2)$;
19. end for

รูปที่ 10 ขั้นตอนการหาค่าไคสแคว์

3) Gini index [40] [41] เป็นค่าที่บ่งบอกว่าคุณลักษณะนั้นว่าสมควรนำมาใช้เป็นคุณลักษณะในการแบ่งข้อมูลหรือไม่ คุณลักษณะที่มีค่า Gini index น้อยแสดงว่าเป็นคุณลักษณะที่มีความสำคัญและสามารถแบ่งข้อมูลได้ดี แนวคิดการใช้ Gini index คือ สมมติ D คือ ข้อมูลตัวอย่างซึ่งอยู่คลาสแตกต่างกัน k คลาส ข้อมูลใน D สามารถแบ่งเป็นข้อมูลย่อยได้ k ชุด สมมติ d_i คือชุดข้อมูลย่อยที่อยู่ในคลาส C_i ดังนั้น Gini index ของ D แสดงได้ดังสมการ (2.20)

$$Gini(D) = 1 - \sum_{i-1}^{k} p_i^2 \qquad (2.20)$$

โดยที่ P_i คือโอกาสที่ D จะเกิดข้อมูลอยู่ในคลาส C_i ถ้า Gini Index เท่ากับ 0 แสดงว่าข้อมูลทุกตัวที่อยู่ใน D อยู่ในคลาสเดียวกัน จากนั้นทำการคำนวณหาค่า Gini Sprit ของแต่ละคุณลักษณะเพื่อพิจารณาว่าคุณลักษณะใดสามารถแยกแยะข้อมูลได้ดีที่สุด ซึ่งถ้า Gini Split น้อยแสดงว่าสามารถแยกแยะข้อมูลได้ดี วิธีการคำนวณค่า Gini Split แสดงดังสมการ (2.21)

$$GiniSplit_T(D) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2) \qquad (2.21)$$

โดยที่ $|D_1|$ คือ จำนวนค่าน้ำหนักของคุณลักษณะ t_i ที่มีค่าเท่ากับ 1

$|D_2|$ คือ จำนวนค่าน้ำหนักของคุณลักษณะ t_i ที่มีค่าเท่ากับ 0

ขั้นตอนการหาค่า Gini Index แสดงได้ดังรูปที่ 11

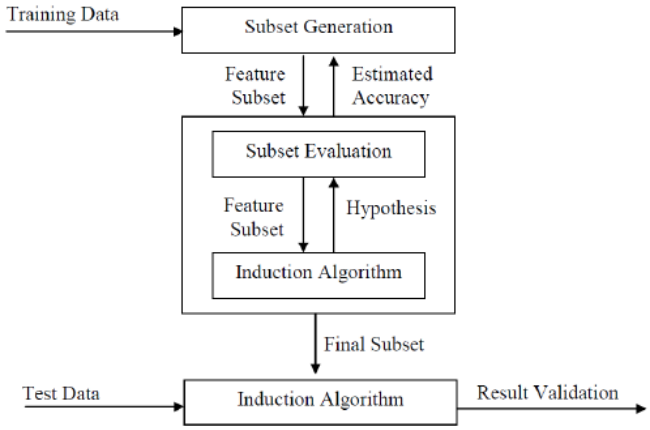
Algorithm 3: Gini Index Feature Ranking

1. for each class $c_k \in C$ do:
2. for document $d_j \in D$ do:
3. if d_j in c_k do:
4. $n_k += 1$;
5. for each term $t_i \in T$ do:
6. if t_i in c_k do:
7. $n(t_i, c_k) += 1$;
8. end for
9. end for
10. $p_k = n_k / N$
11. $Gini(D) = 1 - \sum p_k{}^2$;
12. $Gini(t_i, D_j) = 1 - [n(t_i, c_k) / \sum n(t_i, c_k)]$
13. end for

รูปที่ 11 *ขั้นตอนการหาค่า* Gini

2.5.2 วิธี Wrapper Models

การคัดเลือกคุณลักษณะวิธี Wrapper Model เป็นการคัดเลือกคุณลักษณะจากตัวจำแนกที่สร้างขึ้นมาจากเซตของคุณลักษณะที่กำหนดไว้แล้วทำการวัดประสิทธิภาพการทำงานของตัวจำแนก จากนั้นทำการเลือกคุณลักษณะที่ทำให้ตัวจำแนกมีประสิทธิภาพในการจำแนกมากที่สุด [42] ภาพประกอบแสดงวิธีการทำงานของ Wrapper Model แสดงดังรูปที่ 12



รูปที่ 12 *วิธีการทำงานของวิธี* Wrapper Model

ที่มา [42]

จากรูปที่ 9 แสดงวิธีการทำงานของวิธี Wrapper Model โดยวิธีการคัดเลือกคุณลักษณะมี 3 ขั้นตอนหลัก คือ ค้นหาซับเซต (Subset) ของคุณลักษณะ แล้วประเมินผลซับเซตดังกล่าวจากประสิทธิภาพของตัวจำแนก และทำซ้ำไปเรื่อย ๆ จนกว่าจะได้ตัวจำแนกที่มีประสิทธิภาพดีเป็นที่พึงพอใจ จึงนำซับเซตที่ได้ไปใช้กับข้อมูลชุดทดสอบ ซึ่งการคัดเลือกคุณลักษณะด้วยวิธีการแบบ Wrapper models ให้ประสิทธิภาพในการจำแนกดีกว่าวิธีการแบบ Filter models แต่ใช้เวลาในการประมวลผลนานเมื่อเปรียบเทียบกับวิธีการคัดเลือกคุณลักษณะแบบ Filter models

2.5.3 วิธี Embedded Models

การคัดเลือกคุณลักษณะวิธี Embedded Models เป็นวิธีการคัดเลือกคุณลักษณะที่ให้ประสิทธิภาพในการจำแนกสูงแต่ใช้เวลาในการประมวลผลน้อย โดยรวมเอาข้อดีของ Wrapper models และ Filter models เข้าด้วยกัน ค่าน้ำหนักของคุณลักษณะสามารถเรียนรู้ได้จากตัวจำแนกที่สร้างขึ้น วิธีการคัดเลือกคุณลักษณะที่มีประสิทธิภาพในการคัดเลือกคุณลักษณะที่สกัดจากข้อความ คือ Support Vector Machine (SVM) ซึ่งสร้างตัวจำแนกจากคุณลักษณะทั้งหมดจากนั้นทำการคำนวณน้ำหนักของคุณลักษณะเพื่อคัดเลือกคุณลักษณะที่มีความสำคัญ คุณลักษณะที่มีค่าน้ำหนักเป็น 0 จะถูกกำจัดทิ้งไป โดยที่ประสิทธิภาพของตัวจำแนกยังคงเดิม คุณลักษณะที่มีค่าน้ำหนักสูง

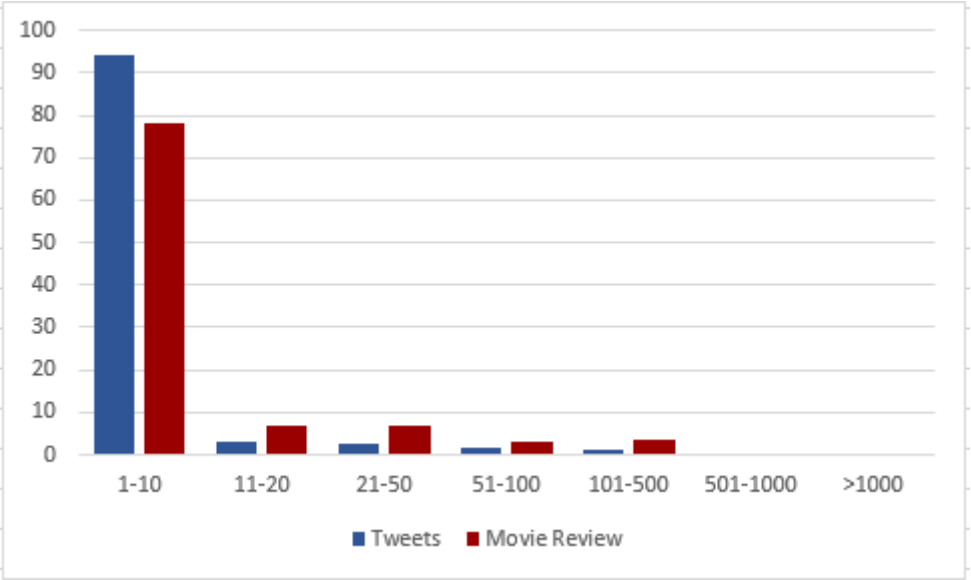
แสดงว่ามีความสำคัญและจะถูกเลือกนำไปใช้ ตัวจำแนก SVM มีประสิทธิภาพในการจำแนกข้อมูลที่เป็นข้อความซึ่งเป็นข้อมูลเบาบางและมีจำนวนคุณลักษณะที่สูง [43] ดังนั้นงานวิจัยนี้จึงเลือกวิธีการคัดเลือกคุณลักษณะจากตัวจำแนกที่สร้างขึ้นโดยใช้ SVM และใช้ Linear kernel ซึ่งให้ประสิทธิภาพที่ดีเมื่อจำแนกข้อมูลสองชุด การจำแนกข้อมูลข้อความ x โดยใช้ Linear kernel ใน SVM สามารถทำได้โดยใช้สมการ (2.22)

$$prediction(x) = \text{sgn}[b + w^T x] \text{ for } w = \sum_i \alpha_i x_i \qquad (2.22)$$

โดยที่ $\boldsymbol{w}^T$ คือ เวกเตอร์ที่ตั้งฉากกับเส้นไฮเปอร์เพลน b คือ ค่าโน้มเอียง (Bias) $\boldsymbol{x}$ คือ ค่าเวกเตอร์ข้อมูลที่จะจำแนก $x_i = (x_{i1},x_{i2},...,x_{id})$ คือ ซัพพอร์ตเวกเตอร์ และ d คือ จำนวนคุณลักษณะในตัวจำแนก น้ำหนักของเวกเตอร์ $w = (w_1, w_2,...,w_d)$ สามารถคำนวณได้จากตัวจำแนก SVM ดังสมการที่ 2.6 สำหรับการคัดเลือกข้อมูล $|w_j|$ หมายถึงค่าน้ำหนักของคุณลักษณะ j ถ้าค่าน้ำหนักของคุณลักษณะ j มีค่าใกล้ 0 แสดงว่าคุณลักษณะ j มีผลกระทบน้อยในการจำแนกข้อมูล แต่ถ้าค่าน้ำหนักของคุณลักษณะ j มีค่าสูงแสดงว่าคุณลักษณะ j มีผลกระทบสูงในการจำแนกข้อมูล ดังนั้นคุณลักษณะที่มีค่าน้ำหนักน้อยจะถูกกำจัดออกไปเพราะไม่มีความสำคัญสำหรับการจำแนกข้อมูล

2.6 รูปแบบข้อมูลแนวตั้ง (Vertical Data Format)

วิธีการจำแนกข้อความโดยส่วนมากจะสร้างตัวจำแนกจากรูปแบบเวกเตอร์แนวนอน (Horizontal Vector) ชุดข้อมูลจะถูกจัดเก็บในรูปแบบของแถว แต่ละแถวแทนด้วยเอกสาร และประกอบด้วยค่าน้ำหนักของคุณลักษณะที่เป็นตัวแทนของเอกสารทั้งหมด ถ้าไม่ปรากฏคุณลักษณะในเอกสารค่าน้ำหนักจะมีค่าเป็น 0 ในงานวิจัยของ Saif และคณะ [44] ได้ทำการวิเคราะห์ความถี่ของการเกิดคำแต่ละคำในเอกสาร โดยเปรียบเทียบข้อมูลทวิตเตอร์ (Twitter Data) กับข้อมูลวิจารณ์ภาพยนตร์ (Movie Review Data) พบว่า ค่าความถี่ของคำที่เกิดขึ้นในเอกสารทั้งหมดที่มีจำนวนน้อยกว่า 10 ครั้ง มีมากถึงร้อยละ 93 สำหรับข้อมูลทวิตเตอร์ และ ร้อยละ 78 สำหรับข้อมูลวิจารณ์ภาพยนตร์ ดังรูปที่ 13



รูปที่ 13 *แสดงความถี่ของการเกิดคำ*

ที่มา [44]

จากรูปที่ 10 แสดงให้เห็นว่าข้อความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์มีการใช้คำศัพท์ที่หลากหลายและเป็นไปได้ว่าคำศัพท์จำนวนมากเหล่านั้นอาจเป็นข้อมูลที่เป็นทางการ คำที่สะกดผิด อักษรย่อ หรือคำเฉพาะที่ใช้กับบุคคลบางกลุ่ม [45] คำศัพท์ที่หลากหลายเหล่านั้นส่งผลให้ขนาดของคำคุณลักษณะมีจำนวนมาก ส่งผลให้ใช้เวลาในการประมวลผลมากและประสิทธิภาพในการจำแนกความคิดเห็นลดลง ตัวอย่างข้อมูลในรูปแบบเวกเตอร์สำหรับใช้ในการจำแนกความคิดเห็นโดยทั่วไปจะเป็นรูปแบบเวกเตอร์ในแนวนอน โดยที่แถวประกอบด้วย เซตของเอกสาร $D=\{d_1,d_2,..,d_n\}$ เมื่อ n = จำนวนเอกสารทั้งหมด และแต่ละคอลัมน์ประกอบด้วย เซตของคุณลักษณะ $T=\{t_1,t_2,..,t_m\}$ เมื่อ m = จำนวนคุณลักษณะทั้งหมด และ w เป็นค่าน้ำหนัก (Weight) ของการแทนค่าการเกิดคุณลักษณะในเอกสาร แสดงดังตาราง 4

ตาราง 4 *รูปแบบเวกเตอร์แนวนอน* (Horizontal Vector)

ลำดับเอกสาร (Document ID)	คุณลักษณะที่ 1 (t_1)	คุณลักษณะที่ 2 (t_2)	...	คุณลักษณะที่ n (t_n)
d_1	w_{11}	w_{12}	...	w_{1m}
d_2	w_{21}	w_{22}	...	w_{2m}
...	...	...	...	...
d_n	w_{n1}	w_{n2}		w_{nm}

วิธีการแปลงข้อมูลเวกเตอร์แนวนอนเป็นข้อมูลแนวตั้ง เป็นอีกหนึ่งวิธีการที่ช่วยให้ การวิเคราะห์ ช่วยลดระยะเวลาในการประมวลผล และง่ายต่อการวิเคราะห์ความสัมพันธ์ของคุณลักษณะ ข้อมูลแนวตั้ง ประกอบด้วยแถวซึ่งเป็นตัวแทนคุณลักษณะ แต่ละคุณลักษณะ ประกอบด้วยเอกสาร ที่ปรากฏคุณลักษณะนั้น เช่น แถวที่ 1 คือ คุณลักษณะ t_1 ประกอบด้วยเอกสาร d_1, d_5 แสดงว่า ในเอกสาร d_1 และ d_5 มีคุณลักษณะ t_1 ปรากฏอยู่ ดังตาราง 5

ตาราง 5 ตัวอย่างรูปแบบข้อมูลแนวตั้ง (Vertical Data Format)

คุณลักษณะ (Feature)	**ชุดของเอกสาร (Set of Documents)**
t_1	d_1, d_5
t_2	d_2, d_3
...	...
t_n	d_1, d_3, d_4

ตัวอย่างงานวิจัยที่ใช้รูปแบบข้อมูลแนวตั้ง เพื่อเพิ่มประสิทธิภาพของขั้นตอนวิธี เช่น Zaki และ Gouda [12] แสดงให้เห็นว่าการจัดรูปแบบข้อมูลแนวตั้ง สามารถลดขนาดของหน่วยความจำ และมีประสิทธิภาพดีกว่าการใช้รูปแบบข้อมูลแนวนอน Viger และ Gomariz [46] แสดงให้เห็นว่าการทำเหมืองรูปแบบลำดับ (Sequential Pattern Mining) โดยใช้รูปแบบข้อมูลแนวตั้งช่วยให้ประสิทธิภาพดีเยี่ยม และช่วยลดเวลาในการคำนวณ งานวิจัยนี้ ผู้วิจัยได้ใช้รูปแบบข้อมูลแนวตั้งมาใช้ในการวิเคราะห์คุณลักษณะ ซึ่งข้อดีของวิธีการนี้คือช่วยลดเวลาในการประมวลผลได้ดีมาก

2.7 การวัดประสิทธิภาพในการจำแนกข้อมูล (Evaluation)

การวัดประสิทธิภาพสำหรับงานวิจัยด้านเหมืองความคิดเห็นโดยทั่วไปจะหาค่าความถูกต้อง (Accuracy) ความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure) การวัดประสิทธิภาพจำเป็นต้องแบ่งข้อมูลออกเป็น 2 ส่วน คือ ข้อมูลชุดสอนกับข้อมูลชุดทดสอบ

2.7.1 การแบ่งข้อมูล

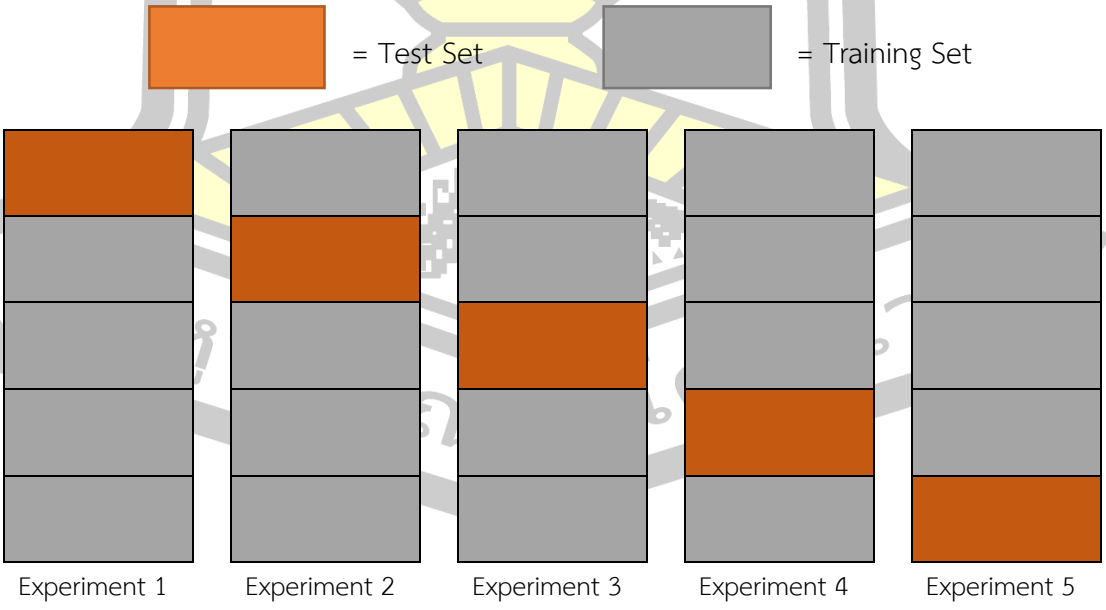
การแบ่งข้อมูลเพื่อใช้วัดประสิทธิภาพการจำแนก แบ่งเป็น 3 วิธีการหลัก [29] ได้แก่

1) วิธี Self Consistency Test เป็นวิธีการที่ง่ายที่สุด คือ ข้อมูลที่ใช้สร้างโมเดลและข้อมูลที่ใช้ทดสอบเป็นข้อมูลชุดเดียวกัน วิธีการนี้ได้ผลการวัดประสิทธิภาพที่ค่อนข้างสูง แต่ไม่ค่อยได้รับความ

นิยมและไม่เหมาะสมที่จะนำไปใช้รายงานผลในงานวิจัยต่าง ๆ เนื่องจากการวัดประสิทธิภาพมีความลำเอียงจากการใช้ข้อมูลชุดเดียวกันในการสร้างตัวจำแนกและตัวทดสอบ

2) วิธี Split Test เป็นวิธีการแบ่งข้อมูลด้วยการสุ่ม ซึ่งจะแบ่งออกเป็น 2 ส่วน เช่น 60% ต่อ 40% หรือ 70% ต่อ 30% โดยข้อมูลส่วนที่ใช้สำหรับสร้างโมเดล คือ ข้อมูลส่วนที่ 1 (60%,70%) และ ข้อมูลที่ใช้ในการทดสอบประสิทธิภาพของโมเดล คือ ข้อมูลส่วนที่ 2 (40%,30%) การทดสอบด้วยวิธีนี้ จะทำการสุ่มข้อมูลเพียงครั้งเดียว ถ้าสุ่มข้อมูลที่ใช้สำหรับการทดสอบที่มีลักษณะที่ใกล้เคียงกับข้อมูลที่ใช้สร้างโมเดล ก็จะทำให้ผลการวัดประสิทธิภาพออกมาดี แต่ถ้าสุ่มข้อมูลทดสอบที่มีลักษณะแตกต่างกับข้อมูลที่ใช้สำหรับสร้างโมเดลมาก ก็จะทำให้ผลของการวัดประสิทธิภาพออกมาไม่ดี ดังนั้น หากจะใช้วิธี Split Test ควรจะมีการสุ่มหลายครั้ง วิธีการนี้จะใช้เวลาในการสร้างโมเดลน้อย เหมาะกับชุดข้อมูลที่มีขนาดใหญ่มาก

3) วิธี Cross-Validation Test เป็นวิธีที่นิยมใช้ในการวัดประสิทธิภาพของโมเดล เนื่องจากผลที่ได้มีความน่าเชื่อถือ การวัดประสิทธิภาพด้วยวิธีการนี้ จะทำการเลือกข้อมูลออกมา K ชุด เพื่อประเมินผล จากข้อมูลทั้งหมด โดยการทดลองครั้งแรกจะใช้ชุดข้อมูลชุดแรกเป็นตัวทดสอบและข้อมูลชุดที่เหลือเป็นชุดข้อมูลเรียนรู้ ทำไปเรื่อย ๆ จนครบจำนวน K ชุด เช่น 5-fold cross-validation คือ จะทำการแบ่งข้อมูลออกเป็น 5 ชุด โดยที่แต่ละชุดมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลชุดแรกเป็นข้อมูลทดสอบประสิทธิภาพของโมเดลและข้อมูลชุดที่เหลือเป็นชุดเรียนรู้ ทำวนไปเช่นนี้จนครบ 5 รอบ แล้วนำค่าประสิทธิภาพที่ได้ในแต่ละรอบมาคำนวณค่าเฉลี่ย การแบ่งข้อมูลด้วยวิธีการ Cross-validation Test ดังแสดงในรูปที่ 14



รูปที่ 14 *ตัวอย่างขั้นตอนการทำงานของ* K-Fold Cross Validation

วิธีการประเมินผลด้วย K-Fold Cross Validation มีข้อดีคือ ข้อมูลทุกชุดจะถูกนำมาทดสอบประเมินผล แต่มีข้อเสียคือ ใช้เวลานานในการทดสอบ โดยขึ้นอยู่กับจำนวนชุดที่จะนำมาทดสอบ

2.7.2 การวัดประสิทธิภาพการจำแนก

งานวิจัยด้านการทำเหมืองความคิดเห็น (Opinion Mining) โดยทั่วไปใช้วิธีการวัดประสิทธิภาพ [29] ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเฉลี่ยประสิทธิภาพโดยรวม (F-Measure) อธิบายโดยใช้ตาราง Confusion Matrix ซึ่งเป็นตารางแบบจัตุรัส มีจำนวนแถวเท่ากับจำนวนคอลัมน์ และเท่ากับจำนวนคลาส เช่น มีคลาสคำตอบของข้อมูลชุดสอน 2 คลาส คือ ความคิดเห็นเชิงบวก (Positive) และ ความคิดเห็นเชิงลบ (Negative) ทำให้ตารางนี้ถูกสร้างเป็นตารางขนาด 2x2 โดยที่ข้อมูลด้านคอลัมน์ เป็นคลาสที่อยู่ในข้อมูลชุดสอน (Actual) และข้อมูลด้านแถว เป็นคลาสที่ทำนายได้ (Predicted) ดังตาราง 6

ตาราง 6 Confusion Matrix

Predicted / Actual	Positive	Negative
Positive	***TP***	***FN***
Negative	***FP***	***TN***

โดยที่ ***TP*** (True Positive) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Positive

FN (False Negative) คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส Negative แต่คำตอบคือ Positive

FP (False Positive) คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส Positive แต่คำตอบคือ Negative

TN (True Negative) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Negative

1) การวัดค่าความถูกต้อง (Accuracy)

การวัดค่าความถูกต้องของการจำแนกความคิดเห็นเป็นการวัดความถูกต้องของวิธีการหรือรูปแบบการจำแนกประเภทข้อมูลโดยจะพิจารณารวมทุกคลาส คำนวณจากผลรวมของค่าที่ทำนายคลาสได้ถูกต้องหารด้วยผลรวมของค่าที่ทำนายทั้งหมด ดังสมการ (2.23)

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (2.23)$$

2) การวัดค่าความแม่นยำ (Precision)

การวัดค่าความแม่นยำของการจำแนกความคิดเห็นเป็นการวัดประสิทธิภาพของวิธีการหรือรูปแบบการจำแนกประเภทข้อมูลโดยจะพิจารณาแยกทีละคลาส คำนวณจากค่าที่ทำนายถูกต้องว่าเป็นคลาสที่พิจารณา หารด้วยผลรวมของค่าที่ทำนายถูกว่าเป็นคลาสที่พิจารณาและค่าที่ทำนายว่าเป็นคลาสอื่นแต่ในความเป็นจริงอยู่ในคลาสที่พิจารณา ดังสมการ (2.24) สมการ (2.25) ตามลำดับ

$$\Pr ecision_{positive} = \frac{TP}{TP + FP} \quad (2.24)$$

$$\Pr ecision_{negative} = \frac{TN}{FN + TN} \quad (2.25)$$

3) การวัดค่าความระลึก (Recall)

การวัดค่าความระลึกของวิธีการจำแนกความคิดเห็นเป็นการวัดความถูกต้องของวิธีการโดยจะพิจารณาแยกทีละคลาส คำนวณจากค่าที่ทำนายถูกต้องว่าเป็นคลาสที่พิจารณา หารด้วยผลรวมของค่าที่ทำนายถูกต้องว่าเป็นคลาสที่พิจารณาและค่าที่ทำนายว่าเป็นคลาสที่พิจารณาแต่คำตอบอยู่ในคลาสอื่น ดังสมการ (2.26) สมการ (2.27) ตามลำดับ

$$\Re call_{positive} = \frac{TP}{TP + FN} \quad (2.26)$$

$$\Re call_{negative} = \frac{TN}{FP + TN} \quad (2.27)$$

4) ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure)

ค่าเฉลี่ยประสิทธิภาพโดยรวมซึ่งจะพิจารณาแยกทีละคลาส เป็นการนำค่าความระลึกและค่าความแม่นยำมาพิจารณาร่วมกัน ระบบที่มีประสิทธิภาพดีจะต้องมีค่าความระลึกและค่าความแม่นยำสูงใกล้เคียงกัน ดังสมการ (2.28) สมการ (2.29) ตามลำดับ

$$F-measure_{positive}=2\times\frac{\Pr ecision_{positive}\times \operatorname{Re} call_{positive}}{\Pr ecision_{positive}+\operatorname{Re} call_{positive}} \quad (2.28)$$

$$F-measure_{negative}=2\times\frac{\Pr ecision_{negative}\times \operatorname{Re} call_{negative}}{\Pr ecision_{negative}+\operatorname{Re} call_{negative}} \quad (2.29)$$

2.8 งานวิจัยที่เกี่ยวข้อง

2.8.1 งานวิจัยที่ใช้วิธีการใช้คลังคำ

Karamibekr และ Ghorbani [5] วิเคราะห์ความแตกต่างของการทำเหมืองความคิดเห็นที่มีต่อสินค้า (Product Review) และ ความคิดเห็นประเด็นทั่วไปทางสังคม (Social Issue) เนื่องจากการจำแนกความคิดเห็นประเด็นทางสังคมมีประสิทธิภาพในการจำแนกไม่สูงมาก เมื่อเปรียบเทียบกับการจำแนกความคิดเห็นที่มีต่อสินค้า จากผลการศึกษาพบว่า ความคิดเห็นที่มีต่อสินค้าจะมีคำคุณลักษณะที่บ่งบอกถึงเรื่องนั้นๆ ปรากฏอยู่ในประโยค แต่ความคิดเห็นในประเด็นทางสังคมมีคำคุณลักษณะที่บอกถึงประเด็นน้อยมาก และยังพบว่าความคิดเห็นที่มีต่อสินค้าเป็นประโยคที่ชัดเจน ไม่กำกวม (Explicit Mention s) ในขณะที่ความคิดเห็นประเด็นทั่วไปของสังคมส่วนมากเป็นประโยคกำกวม (Implicit Mentions) ดังนั้น การจำแนกความคิดเห็นด้วยวิธีการเดียวกับการวิเคราะห์ความคิดเห็นในประเด็นทางสังคมจึงยังไม่เพียงพอ ผู้วิจัยจึงได้นำเสนอวิธีการใหม่ในการจำแนกความคิดเห็นประเด็นทางสังคม โดยใช้วิธีการหาคำกริยาที่บ่งบอกถึงความรู้สึก (Opinion Verb) ในประโยค และใช้เป็นคำหลักในการหาหน้าที่ของคำอื่น ๆ งานวิจัยนี้ใช้คลังคำจาก Stanford POS Tagger ในการหาคำกริยา คำนาม คำคุณศัพท์ คำวิเศษณ์ เมื่อทราบหน้าที่ของคำแต่ละคำแล้วให้กำหนดค่าน้ำหนัก (Strange) ของคำ คือ คำที่เป็นความคิดเห็นเชิงลบมาก (Strongly Negative) เท่ากับ -2, คำที่เป็นความคิดเห็นเชิงลบ (Negative) เท่ากับ –1, คำที่เป็นความคิดเห็นเชิงบวก (Positive) เท่ากับ 1, คำที่เป็นความคิดเห็นเชิงบวกมาก (Strong Positive) เท่ากับ 2 ชุดข้อมูลที่ใช้สำหรับงานวิจัย เป็นข้อมูลความคิดเห็นเกี่ยวกับประเด็นทางสังคม จำนวน 1,016 ความคิดเห็น ประกอบด้วยความคิดเห็นเชิงบวก จำนวน 588 ความคิดเห็น และความคิดเห็นเชิงลบ จำนวน 428 ความคิดเห็น ขั้นตอนวิธีการจำแนกความคิดเห็นเริ่มจากการแบ่งเอกสารออกเป็นประโยค แล้วสกัดคำกริยาและระบุหน้าที่คำอื่นๆ ที่บ่งบอกถึงความรู้สึก โดยใช้ Stanford POS Tagger แล้วให้ค่าคะแนนแต่ละคำ จากนั้นคำนวณค่าความคิดเห็นในประโยคจากค่าผลรวมของคะแนนความคิดเห็นแต่ละคำ และหาค่าความ

คิดเห็นทั้งเอกสารโดยการนำค่าความคิดเห็นแต่ละประโยคมารวมกัน การประเมินประสิทธิภาพการจำแนกใช้วิธีการประเมินความถูกต้อง พบว่า วิธีการจำแนกความคิดเห็นในประเด็นทางสังคม มีความถูกต้องในการจำแนก เท่ากับ 65%

Mostafa [47] นำเสนอการเหมืองข้อความสำหรับวิเคราะห์ความรู้สึกของผู้บริโภคที่มีต่อสินค้ายี่ห้อดัง ได้แก่ Nokia, T-Mobile, IBM, KLM และ DHL โดยรวบรวมข้อความที่อยู่ในเว็บไซต์ทวิตเตอร์ (Twitter) มาประเมินความรู้สึกของผู้บริโภค เพื่อประเมินว่าผู้บริโภคมีความคิดเห็นต่อสินค้าแต่ละยี่ห้อในเชิงบวก (Positive) หรือ เชิงลบ (Negative) วิธีการที่ผู้วิจัยใช้ คือ พจนานุกรมคำแสดงความคิดเห็น (Sentiment Corpus) จำนวน 6,800 คำ ประกอบด้วยความคิดเห็นเชิงบวก และความคิดเห็นเชิงลบ จากนั้นกำหนดให้คำความคิดเห็นเชิงบวกมีค่าน้ำหนักเท่ากับ +1 และคำความคิดเห็นเชิงลบมีค่าน้ำหนักเท่ากับ -1 ประเมินความคิดเห็นจากคำที่พบในประโยค และสรุปผลเพื่อแสดงให้เห็นในรูปแบบกราฟ โดยใช้ StreamGraph Software แสดงให้เห็นความรู้สึกโดยรวมของผู้บริโภคที่มีต่อสินค้ายี่ห้อเหล่านั้น

Marrese-Taylor และคณะ [48] นำเสนอขั้นตอนวิธีในการทำเหมืองความคิดเห็น เพื่อประยุกต์ใช้กับข้อความคิดเห็นเกี่ยวกับการท่องเที่ยว โดยรวบรวมข้อมูลจากเว็บไซต์ TripAdvisor ขั้นตอนการจำแนกความคิดเห็น ผู้วิจัยพัฒนาเพิ่มเติมจากขั้นตอนของ Bing Liu’s ซึ่งเป็นขั้นตอนที่พัฒนาขึ้นเพื่อจำแนกความคิดเห็นเกี่ยวกับสินค้าทั่วไป ไม่ได้เจาะจงความคิดเห็นด้านการท่องเที่ยวโดยตรง งานวิจัยนี้มีขั้นตอนการจำแนกความคิดเห็น 3 ขั้นตอน คือ 1) สกัดประเด็น (Aspect) ที่กล่าวถึงในข้อความ แล้วแยกแต่ละประเด็นออกเป็นประโยค (Sentence) 2) นำประโยคมาวิเคราะห์เพื่อจำแนกประเภทความคิดเห็น โดยแบ่งเป็น 3 ด้าน ได้แก่ ความคิดเห็นเชิงบวก ความคิดเห็นเชิงลบ และความคิดเห็นที่เป็นกลาง 3) สรุปความคิดเห็นภาพรวมว่านักท่องเที่ยวส่วนมากมีความคิดเห็นเกี่ยวกับแต่ละประเด็นเป็นอย่างไร ผลการวัดประสิทธิภาพขั้นตอนวิธีจำแนกความคิดที่นำเสนอเปรียบเทียบวิธีการของ Liu’s พบว่า วิธีการที่นำเสนอมีค่าความถูกต้องสูงกว่าวิธีการของ Liu’s

Somasundaran [49] จำแนกความคิดเห็นด้วยวิธีการใช้คลังคำ ข้อมูลที่ใช้ในการวิจัยรวบรวมจากคลังข้อมูลความคิดเห็น MPQA Corpus จำนวน 2,232 ความคิดเห็น ประกอบด้วย ข้อความคิดเกี่ยวกับ Gun Right จำนวน 306 ข้อความ ข้อความเกี่ยวกับ Gay Rights จำนวน 846 ข้อความ ข้อความเกี่ยวกับ Abortion 550 ข้อความ และข้อความเกี่ยวกับ Creationism จำนวน 530 ข้อความ โดยสร้างคลังคำศัพท์ความรู้สึกและคำตรงข้าม มีขั้นตอนการดำเนินการ 4 ขั้นตอน คือ 1) สร้าง Candidate Set จากคลังคำศัพท์ที่สกัดจากข้อความ ด้วยวิธี N-gram เช่น “can only rise to meet it by making some radical changes” จะได้ Candidate Set คือ [can, can only, can only rise] 2) ลบคำที่ไม่มีขั้วความคิดเห็นออก 3) หาค่าความคิดเห็นของข้อความใน Candidate

Set 4) บันทึกลงในคลังคำศัพท์ คุณลักษณะที่ใช้ในการจำแนกความคิดเห็น คือ Model Verbs และ Syntactic Rules ผลการประเมินประสิทธิภาพ พบว่า มีความถูกต้องในการจำแนก เท่ากับ 63.96

Cruz [50] นำเสนอวิธีการอัตโนมัติสำหรับสร้างพจนานุกรม Lemma Level ซึ่งแรงจูงใจในการทำวิจัยเกิดจากปัญหาของการวิเคราะห์ประโยคจากคำในพจนานุกรม เช่น คำว่า “Good” ในพจนานุกรมมีความหมายในเชิงบวก แต่หากนำไปใช้กับบางประโยค เช่น “His second album is not so good” จะเห็นว่าประโยคนี้มีคำว่า Good หากใช้การวิเคราะห์ด้วยพจนานุกรม ข้อความนี้จะจัดเป็นข้อความเชิงบวก แต่หากเราวิเคราะห์ตามรูปประโยคแล้วจะพบว่าข้อความนี้เป็นเชิงลบ วิธีการสร้างพจนานุกรม Lemma Level ผู้วิจัยได้สร้างพจนานุกรม Synset Level สำหรับภาษาอังกฤษ ที่พัฒนาเพิ่มเติมจาก SentiWordNet 3.0 ซึ่งจัดเป็นหนึ่งในพจนานุกรมที่นิยมมากที่สุดในปัจจุบัน จากนั้นทำการพัฒนาพจนานุกรม Lemma Level จำนวน 8 Layer แต่ละ Layer จะถูกจัดเรียงในลักษณะที่เป็นชั้นต่อกันโดยกำหนดให้ Layer ถัดไปมี Lemmas ทั้งหมดของ Layer ก่อนหน้า ผลการวิจัยพบว่า วิเคราะห์ความคิดเห็นโดยใช้พจนานุกรม Lemma Level สำหรับข้อความภาษาอังกฤษ ได้ค่าความถูกต้องมากกว่า 80% ใน Layer ที่ 1-7 และสำหรับข้อความภาษาสเปนและอีก 3 ภาษา ที่ใช้อย่างเป็นทางการในสเปน ได้ค่าความถูกต้องมากกว่า 80% ใน Layer ที่ 1-6 เช่นกัน ส่วนใน Layer อื่นๆ ให้ค่าความถูกต้องที่ 60%

Terrana [51] นำเสนอการวิเคราะห์ข้อความที่อยู่บน Facebook ของผู้ใช้ โดยสร้าง Facebook Page เพื่อให้เข้าถึงข้อมูลของที่กดถูกใจเพจนั้น แล้วทำการดึงข้อมูลเพื่อตรวจสอบว่าใครกล่าวถึงอะไร กล่าวถึงใคร และกล่าวในเชิงบวก เชิงลบ หรือเป็นกลาง การเข้าถึงข้อมูลที่อยู่บน Facebook ผู้วิจัยใช้ Facebook Graph APIs และ Facebook Query Language (FQL) Table Reference สำหรับดึงข้อมูล Text Messages, Comments และ Likes และทำการวิเคราะห์ต้นฉบับเนื้อหาโดยใช้ Linguistic Inquiry and Word Count (LIWC) ซึ่งเป็นซอฟต์แวร์ในการคำนวณระดับที่ User Root เพื่อใช้ระบุหมวดหมู่ของคำที่แตกต่าง แล้วจำแนกความรู้สึกตามเนื้อหาทางอารมณ์ของพวกเขา จากนั้นนำมาสร้างเป็นกราฟความสัมพันธ์ของผู้ใช้

2.8.2 งานวิจัยที่ใช้วิธีการเรียนรู้ของเครื่อง

Go และคณะ [11] นำเสนอวิธีการจำแนกความคิดเห็นที่อยู่บนเว็บไซต์ทวิตเตอร์ (Twitter Data) โดยใช้วิธีการเรียนรู้ของเครื่อง ได้แก่ นาอีฟเบย์, แม็กซิมัม เอ็นโทรปี (Maximum Entropy) และซัพพอร์เวกเตอร์แมชชีน ใช้ Unigram, Bigrams, Unigrams ร่วมกับ Bigrams และ Unigrams ร่วมกับ Part of Speech Tags. ข้อมูลที่ใช้ในการสร้างรูปแบบการจำแนก คือ ข้อความที่รวบรวมโดยใช้ Twitter API จำนวน 1,600,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 800,000 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 800,000 ข้อความ ในกระบวนการเตรียมข้อมูลชุดสอน

ผู้วิจัยได้ลบข้อความแสดงอารมณ์ (Emoticons) ออก จากนั้นทำการแทนค่าข้อมูล โดยข้อความที่ขึ้นต้นด้วยสัญลักษณ์ @ ถูกแทนที่ด้วยข้อความ USERNAME ข้อความที่เป็นที่อยู่เว็บไซต์ ถูกแทนค่าด้วยข้อความ HTTP นอกจากนี้ยังมีการตรวจสอบตัวอักษรที่ซ้ำกันมากกว่า 2 ตัว เพื่อแทนที่ด้วยตัวอักษรที่ซ้ำจำนวน 2 ตัว ทำการสกัดคุณลักษณะด้วยวิธีการต่างๆ ข้างต้น ร่วมกับการจำแนกด้วยวิธีการเรียนรู้ของเครื่อง ข้อมูลที่ใช้ในการทดสอบเป็นข้อความที่รวบรวมโดย Twitter API ประกอบด้วยข้อความคิดเห็นเชิงบวกจำนวน 182 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 177 ข้อความ ผลการวิจัยพบว่า วิธีการซัพพอร์ตเวกเตอร์แมชชีนร่วมกับ ร่วมกับ Unigrams และ Bigrams มีประสิทธิภาพการจำแนกสูงที่สุด คือ 82.20%

Anjaria และ Guddeti [3] นำเสนอการศึกษาปัจจัยที่มีอิทธิพลต่อการทำเหมืองความคิดเห็นของข้อมูลบนทวิตเตอร์ โดยใช้วิธีการเรียนรู้แบบมีผู้สอน ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน, นาอีฟเบย์, แม็กซิมัม เอ็นโทรปี, อาร์ทิฟิเชียล นิวรอล เน็ตเวิร์ค (Artificial Neural Networks) และรวมหลักการวิเคราะห์องค์ประกอบกับซัพพอร์ตเวเตอร์แมชชีนในการที่จะลดจำนวนมิติ (Dimensionality Reduction) ของข้อมูล ในการทดสอบผู้วิจัยได้ทดลอง 2 กรณีที่แตกต่างกัน คือ การทำนายผลเลือกประธานาธิบดีประเทศสหรัฐอเมริกา ในปี ค.ศ.2012 และ การเลือกตั้งที่ Karnataka ในปี ค.ศ.2013 ผลการประเมินประสิทธิภาพการทำนาย พบว่า วิธีการซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องสูงกว่าวิธีการอื่น คือ การทำนายผลการเลือกตั้งประธานาธิบดี ประเทศสหรัฐอเมริกา ได้ค่าความถูกต้องเท่ากับ 88% และผลการทำนายการเลือกตั้งที่ Kanataka ได้ค่าความถูกต้องเท่ากับ 58%

Akaichi และคณะ [2] นำเสนอการจำแนกความรู้สึกจากข้อความที่โพสต์บน Facebook เพื่อวิเคราะห์ความคิดเห็นและพฤติกรรมของผู้ใช้เฟสบุค งานวิจัยนี้มีขั้นตอนการดำเนินการ 5 ขั้นตอน คือ ขั้นตอนที่ 1) รวบรวมข้อมูลที่โพสต์บนเว็บไซต์เฟสบุคของผู้ใช้ชาวตูนิเซียน (Tunisian) จำนวน 260 คน ขั้นตอนที่ 2) พัฒนาคลังคำศัพท์สำหรับเก็บคำที่อยู่ในรูปแบบภาษาที่ไม่เป็นทางการ (Informal Language) แต่ใช้บ่อยบนเครือข่ายสังคมออนไลน์ โดยผู้วิจัยได้พัฒนาคลังคำศัพท์ 3 ประเภท คือ คลังคำศัพท์สำหรับเก็บอักษรย่อ (Acronyms) เช่น LOL (Positive), GR8 (Positive), CU (Neutral) เป็นต้น คลังคำศัพท์สำหรับเก็บสัญลักษณ์แสดงอารมณ์ (Emoticon) เช่น :), :-), :p, :’(, <3 เป็นต้น และคลังคำศัพท์สำหรับเก็บคำอุทาน (Interjections) เช่น Wow, Haha, Hihi, Oh dear เป็นต้น ซึ่งในคลังคำศัพท์ทั้งสามประเภทนี้จะระบุความหมายไว้ด้วยว่าคำดังกล่าวมีความหมายในเชิงบวก (Positive) เชิงลบ (Negative) หรือ เป็นกลาง (Neutral) ขั้นตอนที่ 3) สกัดคุณลักษณะของข้อความที่รวบรวมมาจากเฟสบุคให้อยู่ในรูปแบบของข้อมูลที่มีโครงสร้าง โดยใช้ Unigrams, Bigrams, Trigrams และ Part-of-Speech เป็นคุณลักษณะในการจำแนกข้อความ ขั้นตอนที่ 4) สร้างโมเดลการเรียนรู้ (Training Model) กำหนดคลาสคำตอบ แล้วแบ่งข้อมูล 2 กลุ่ม คือ ข้อมูลชุดสอน 60% ข้อมูลชุดทดสอบ 40% ขั้นตอนสุดท้าย ทำการวัดประสิทธิภาพของโมเดล งานวิจัยนี้

กำหนดคุณลักษณะสำหรับจำแนกความคิดเห็น 7 รูปแบบ ได้แก่ Unigrams, Bigrams, Trigrams, Unigrams + Bigrams, Unigrams + Trigrams, Bigrams + Trigrams และ Unigrams + Bigrams + Trigrams และทำการเปรียบเทียบการใช้วิธีการเรียนรู้ของเครื่อง 2 วิธี คือ นาอีฟเบย์ และซัพพอร์ตเวกเตอร์แมชชีน ผลการวิจัยพบว่า เมื่อใช้คุณลักษณะ Unigrams วิธีการซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้อง เท่ากับ 72.78% ซึ่งสูงกว่าวิธีนาอีฟเบย์ แต่เมื่อใช้คุณลักษณะแบบ Bigrams พบว่า วิธีนาอีฟเบย์ให้ค่าความถูกต้องสูงกว่าวิธีซัพพอร์ตเวกเตอร์แมชชีน คือ 69.42%

Basari และคณะ [32] นำเสนอการทำเหมืองความคิดเห็นจากข้อความคิดเห็นเกี่ยวกับหนัง (Movie Review) มีขั้นตอนการดำเนินการ 6 ขั้นตอน คือ ขั้นตอนที่ 1) ใช้ชุดข้อมูลทวิตเตอร์ (Twitter) ที่เว็บไซต์ Stanford ได้รวบรวมไว้ ขั้นตอนที่ 2) เตรียมข้อมูล (Preprocessing) ประกอบด้วยการกรองข้อมูล ในการกรองข้อมูลนี้จะใช้คำสำคัญ (Keyword) เป็นชื่อของหนัง เช่น Transformers, Star Trek, X-Men, The Hangover เป็นต้น จากนั้นทำการลบคำที่ไม่มีความหมายออก เช่น @, Url, Hashtag (#) เป็นต้น ขั้นตอนที่ 3) สกัดและเลือกคุณลักษณะ (Feature Selection and Extraction) ขั้นตอนที่ 4) กำหนดค่าน้ำหนักคุณลักษณะ (Feature Weighting) ขั้นตอนที่ 5) จำแนกความคิดเห็นด้วยวิธีการเรียนรู้ของเครื่อง ซึ่งงานวิจัยนี้ได้ใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธี Particle Swarm Optimization เปรียบเทียบกับการใช้วิธีการซัพพอร์ตเวกเตอร์แมชชีนอย่างเดียว ขั้นตอนที่ 6) ตรวจสอบและประเมินผล ผลการวิจัย พบว่า การใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนอย่างเดียว มีค่าความถูกต้องเท่ากับ 71.87% ส่วนการใช้วิธีซัพพอร์ตเวกเตอร์แมชชีนร่วมกับวิธี Particle Swarm Optimization มีค่าความถูกต้องสูงขึ้นเป็น 77.00%

Pak และ Paroubek [52] นำเสนอการจำแนกความคิดเห็นที่อยู่บนเว็บไซต์ทวิตเตอร์ (Twitter) มีขั้นตอนคือ สกัดคุณลักษณะ โดยใช้ n-gram การเตรียมข้อมูลประกอบด้วย 4 ขั้นตอน คือ 1) การกรองข้อความ (Filtering) หรือเรียกอีกอย่างว่าการทำความสะอาดข้อความ โดยลบข้อความที่เป็น URL (เช่น http://example.com) ชื่อผู้ใช้ทวิตเตอร์ (เช่น @alex) สัญลักษณ์พิเศษที่ใช้ในทวิตเตอร์ เช่น สัญลักษณ์ รีทวิต (RT) และอีโมติคอน 2) การตัดคำ งานวิจัยนี้ใช้การตัดคำโดยใช้ช่องว่าง (Space) และสัญลักษณ์พิเศษ (Punctuation Marks) 3) ลบคำที่ไม่มีนัยสำคัญออก เช่น “a” “an” “the” เป็นต้น 4) สร้าง n-gram feature โดยคำที่เป็นคำปฏิเสธ (Negation Word) จะถูกรวมเป็นคำเดียวกับคำถัดไป เช่น I do not like fish เมื่อสร้างรูปแบบ bigrams จะได้ “I do+not”, “do+not like”, “not+like fish” เป็นต้น งานวิจัยนี้ใช้ตัวจำแนก 3 วิธีการ คือ Naïve Bayes, SVM และ CRF ซึ่งผลการวิจัยพบว่า Naïve Bayes มีประสิทธิภาพความถูกต้องในการจำแนกสูงที่สุด

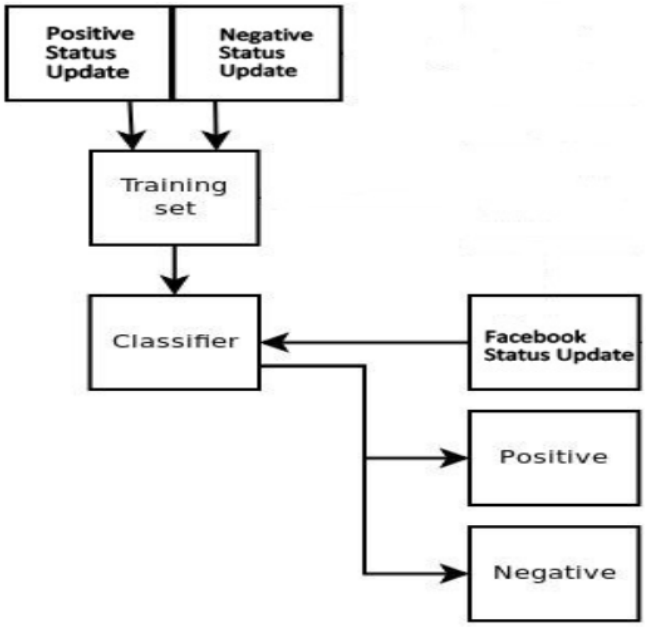
Yang และคณะ [53] จำแนกความคิดเห็นที่อยู่บนเว็บบล็อกด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีนและคอนดิชั่นนอลแรนดอมฟิลด์ (Conditional Random Field : CRF) วิธีการคือ รวมรวม

ข้อมูลโพสต์ที่อยู่บนเว็บบล็อกแบ่งเป็นข้อมูลชุดสอนและข้อมูลชุดทดสอบ จากนั้นสร้างคลังคำศัพท์และตัวจำแนกจากข้อมูลชุดสอน งานวิจัยนี้จำแนกความคิดเห็นในระดับประโยคและใช้ผลการวิเคราะห์แต่ละประโยคไปสรุปความคิดเห็นระดับเอกสาร ผลการวิจัย พบว่า การจำแนกความคิดเห็นในระดับประโยค วิธีการคอนดิชั่นนอลแรนดอมฟิลด์ มีประสิทธิภาพดีกว่าวิธีการซัพพอร์ตเวกเตอร์แมชชีน และพบว่าประโยคสุดท้ายของเอกสารจะเป็นข้อสรุปความคิดเห็นในระดับเอกสารได้

2.8.3 งานวิจัยที่ใช้วิธีการใช้คลังคำร่วมกับเรียนรู้ของเครื่อง

Ortigosa และคณะ [4] นำเสนอวิธีการสกัดความรู้สึกและวิเคราะห์การเปลี่ยนแปลงทางอารมณ์ของผู้เรียน จากข้อความที่อยู่บน Facebook เพื่อนำมาประยุกต์ใช้สำหรับปรับปรุงระบบการเรียนการสอนออนไลน์ (e-Learning) ให้เหมาะสมกับผู้เรียน งานวิจัยนี้ใช้วิธีการที่เรียกว่า Sentbuk ในการดึงข้อความที่อยู่บน Facebook แล้วนำมาวิเคราะห์ถึงความรู้สึกของผู้เรียนด้วยการผสมผสานกันระหว่างวิธีการใช้คลังคำและเรียนรู้ของเครื่อง ทำการจำแนกความรู้สึกเป็น 3 กลุ่ม คือ เชิงบวก เชิงลบ และเป็นกลาง และทำการทดสอบวิธีการจำแนกความรู้สึก 4 วิธีการ คือ 1) ใช้ Lexical-base อย่างเดียว 2) ใช้ Tree decision (J48-C4.5) ร่วมกับ Lexicon-based tagging 3) ใช้ Naïve-Bayes ร่วมกับ Lexicon-based tagging และ 4) ใช้ Support Vector Machine ร่วมกับ Lexicon-based tagging ผลการวิจัยพบว่า การวิเคราะห์ความรู้สึกจากข้อความบน Facebook ด้วยวิธีการ Support Vector Machine ร่วมกับ Lexicon-based tagging ให้ค่าความถูกต้อง (Accuracy) สูงที่สุด คือ 83.27% ผลการวิเคราะห์นำไปประยุกต์ใช้เป็นข้อมูลพื้นฐานสำหรับผู้สอนและใช้ปรับปรุงระบบ e-learning ได้

Troussas และ คณะ [1] นำเสนอการจำแนกความคิดเห็นจากข้อความที่อยู่บนเฟสบุค ด้วยวิธีนาอีฟเบย์ ขั้นตอนแรกทำการรวบรวมข้อมูลสถานะที่อยู่บนเฟสบุค ซึ่งประกอบด้วย 2 กลุ่ม คือ ความคิดเห็นเชิงบวก (Positive) และ ความคิดเห็นเชิงลบ (Negative) จากนั้นแบ่งข้อมูลชุดสอนและชุดทดสอบเป็น 50% - 50% แล้วสร้างตัวจำแนกด้วยข้อมูลชุดสอนและทำการจำแนกความคิดเห็นข้อมูลชุดทดสอบ ดังรูปที่ 15



รูปที่ 15 *ขั้นตอนการวิเคราะห์ความคิดเห็นของ* Troussas

ผู้วิจัยได้เปรียบเทียบการจำแนกความคิดเห็น 3 วิธีการ คือ นาอีฟเบย์, โรชิโน (Rocchino), เปอร์เซปทรอน (Perceptron) พบว่า วิธีการนาอีฟเบย์มีความแม่นยำในการจำแนกมากกว่าวิธีการอื่น คือ 77.00%

Mudinas และคณะ [54] นำเสนอการวิเคราะห์ความคิดเห็นที่ชื่อว่า pSenti เป็นการจำแนกความคิดเห็นด้วยวิธีการใช้คลังคำร่วมกับวิธีการเรียนรู้ของเครื่อง และนำเสนอผลการเปรียบเทียบการจำแนกความคิดเห็นด้วยวิธีการ pSenti กับการจำแนกความคิดเห็นด้วยวิธีการใช้คลังคำอย่างเดียว และการจำแนกความคิดเห็นด้วยวิธีการเรียนรู้ของเครื่องอย่างเดียว 5 กระบวนการหลัก คือ 1) การเตรียมข้อมูล 2) การสกัดคุณลักษณะ 3) การค้นหาคำคุณลักษณะที่เป็นคำแสดงความคิดเห็นด้วยวิธีการใช้คลังคำ 4) จำแนกความคิดเห็นด้วยกระบวนการเรียนรู้ของเครื่อง 5) การปรับค่าน้ำหนักความคิดเห็น และ 6) สรุปผล ข้อมูลที่ใช้มี 2 ชุด คือ ข้อมูลวิจารณ์ซอฟต์แวร์ (Software Reviews) ของ CNET และข้อมูลวิจารณ์ภาพยนตร์ (Movie Reviews) ของ IMDB ในการจำแนกความคิดเห็นด้วยวิธีการใช้คลังคำ โดยใช้คลังความคิดเห็นเดียวกับ pSenti ส่วนการจำแนกความคิดเห็นด้วยวิธีการเรียนรู้ของเครื่องอย่างเดียวใช้คลังความคิดเห็นเดียวกับ pSenti และใช้ถุงคำเป็นคุณลักษณะ ผลการจำแนกความคิดเห็น พบว่า pSenti มีความถูกต้องในการจำแนกเท่ากับ 82.30% ซึ่งสูงกว่าวิธีการใช้คลังคำอย่างเดียวที่มีค่าความถูกต้องเท่ากับ 66.00% แต่น้อยกว่าวิธีการเรียนรู้ของเครื่องอย่างเดียวที่มีค่าความถูกต้องเท่ากับ 86.85%

Fang และ Chen [55] นำเสนอวิธีการผสมผสานการใช้คลังคำกับวิธีการซัพพอร์ตเวกเตอร์แมชชีน เพื่อพัฒนาการจำแนกความคิดเห็น และแสดงให้เห็นถึงปัญหาในการใช้วิธีการใช้คลังคำอย่าง

เดียวไม่สามารถครอบคลุมการจำแนกความคิดเห็นได้ทุกโดเมน เช่น คำว่า Long เป็นความคิดเห็นเชิงบวกสำหรับอายุการใช้งานแบตเตอรี่ แต่หากกล่าวถึงโดเมนของกล้องอาจจะหมายถึงความล่าช้าของชัตเตอร์ ซึ่งเป็นความคิดเห็นเชิงลบ การใช้วิธีการเรียนรู้ของเครื่องจะช่วยเพิ่มประสิทธิภาพการจำแนกได้ เนื่องจากจะทำการเรียนรู้จากประโยคนำเข้า แต่อาจจะยังไม่ใช่วิธีการที่ดีที่สุด งานวิจัยนี้ได้ทำการรวบรวมความคิดเห็นสำหรับแต่ละโดเมน และรวบรวมคำศัพท์ที่เป็นความคิดเห็นเชิงบวกและเชิงลบให้กับแต่ละโดเมนจากคลังคำศัพท์ความคิดเห็นและรวบรวมรูปแบบการใช้ภาษาจากเว็บไซต์ คลังคำแบ่งเป็น 2 รูปแบบคือ คลังคำคุณลักษณะ และคลังคำความคิดเห็นที่มีต่อคุณลักษณะ การสร้างตัวจำแนกได้ผสมผสานวิธีการใช้คลังคำกับวิธีการ SVM คือ คัดเลือกตัวแทนคุณลักษณะนำประโยคนำเข้าจากหน้าที่ของคำและผสมผสานความรู้ที่อยู่ในคลังคำศัพท์ที่สร้างขึ้นและใส่คุณลักษณะเพิ่มเติมลงในคุณลักษณะที่ได้จากการคัดเลือกโดยหน้าที่ของ เช่น ประโยค “The case is rigid so it gives the camera extra nice protection.” ขั้นตอนแรกคัดเลือกคุณลักษณะจากหน้าที่ของคำ ได้แก่ คำนาม (Nouns) คำกริยา (Verb) คำคุณศัพท์ (Adjective) และ คำกริยาวิเศษณ์ (Adverb) ผลที่ได้คือ [case, rigid, give, camera, extra, nice, protection] จากนั้นจำแนกคุณลักษณะจากคำที่ได้ ซึ่งคำว่า case ในประโยคจะมีความหมายเกี่ยวกับอุปกรณ์กล้อง ในขั้นตอนนี้จะทำการเพิ่มคำว่า Accessory เพิ่มเข้าไป จะได้ [case, rigid, give, camera, extra, nice, protection, accessory] ขั้นตอนต่อไปคือ การจำแนกขั้วความคิดเห็น โดยผสมผสานคลังคำที่สร้างขึ้นกับคลังคำความคิดเห็นของ MPQA ในการทดสอบประสิทธิภาพการจำแนก งานวิจัยนี้ได้เปรียบเทียบประสิทธิภาพการจำแนกความคิดเห็น 4 วิธีการ ได้แก่ SVM, MPQA + SVM, DomainLexicons + SVM และ DomainLexicons + MPQA + SVM พบว่า วิธีการจำแนกความคิดเห็นโดยใช้ DomainLexicons + MPQA + SVM มีประสิทธิภาพความถูกต้องของการจำแนกสูงที่สุด คือ 66.80%

Zhang และคณะ [56] พัฒนาวิธีการใหม่ชื่อว่า LMS ในการจำแนกความคิดเห็นบนเว็บไซต์ทวิตเตอร์ ประกอบด้วย 3 กระบวนการหลัก ได้แก่ กระบวนการเตรียมข้อมูล กระบวนการใช้คลังคำและกระบวนการเรียนรู้ของเครื่อง การเตรียมข้อมูลประกอบด้วย 4 กระบวนการย่อย คือ 1) การตัดคำด้วยช่องว่าง 2) ทำความสะอาดข้อมูลโดยตัดลิงค์และสัญลักษณ์พิเศษที่ไม่มีความหมายทิ้งไป เช่น RT, @, url เป็นต้น 3) การหารากคำศัพท์ เป็นการหารากคำศัพท์เดิมจากซึ่งข้อความบนเว็บไซต์ทวิตเตอร์ส่วนมากมักใช้คำย่อ เช่น wknd หมายถึง weekend กระบวนการนี้ใช้คลังคำมาช่วย 4) ระบุหน้าที่ของคำในประโยคให้อยู่ในรูปแบบ Part of Speech Tagging (POS) กระบวนการถัดไปคือการหาประเภทของประโยค แบ่งเป็น 3 ประเภท คือ ประโยคบอกเล่า (Declarative Sentence) ประโยคคำสั่ง (Imperative Sentence) และประโยคคำถาม (Interrogative Sentence) ซึ่งประโยคคำถามจัดเป็นประโยคที่ไม่บอกขั้วความคิดเห็น ในบางประโยคที่ขึ้นต้นด้วยคำสรรพนามแสดงว่าอ้าง

ถึงประโยคแรก เช่น I bought this iPhone yesterday. It is awesome. จากประโยคข้างต้น It หมายถึง iPhone ในประโยคแรก เป็นต้น นอกจากการใช้คลังคำความคิดเห็นแล้วยังมีกฎความคิดเห็น (Opinion Rules) ได้แก่ 1) กฎคำปฏิเสธ (Negation Rules) เมื่ออยู่หน้าคำจะถูกเปลี่ยนความหมายเป็นตรงข้าม เช่น this cellphone is not good จะเห็นว่า good เป็นคำเชิงบวก เมื่อมีคำว่า not อยู่หน้าคำจะความหมายจะถูกเปลี่ยนเป็นเชิงลบ 2) กฎประโยคที่มีคำว่าแต่ (But-Clause Rules) ประโยคที่อยู่ก่อนและหลังคำว่าแต่ จะมีมีความตรงข้ามกัน 3) กฎคำที่มีความหมายลดลงและเพิ่มขึ้น (Decreasing and Increasing Rules) คำที่มีความหมายเกี่ยวกับการเพิ่มขึ้นหรือลดลงจะเกี่ยวข้องกับคำแสดงความคิดเห็น เช่น ยาช่วยบรรเทาความเจ็บปวดอย่างมาก (The drug eases my pain greatly) ซึ่งคำว่าเจ็บปวด (Pain) เป็นคำเชิงลบในคลังคำความคิดเห็น แต่เมื่อดูรูปประโยคแล้วพบว่าประโยคนี้มีความหมายในเชิงบวก จึงมีกฎคำที่มีความหมายลดลงและเพิ่มขึ้น คือ คำที่มีความหมายลดลงรวมกับคำเชิงลบจะมีความหมายเชิงบวก ส่วนคำที่มีความหมายลดลงรวมกับคำที่มีความหมายเชิงบวกจะมีความหมายเชิงลบ หลังจากผ่านกระบวนการใช้คลังคำแล้ว ทำการแปลงข้อมูลในอยู่ในรูปแบบเวกเตอร์ งานวิจัยนี้ใช้การสกัดคุณลักษณะโดยใช้ไคสแคว์ และใช้ซัพพอร์ตเวกเตอร์แมชชีนเป็นตัวจำแนก และเปรียบเทียบการจำแนกความคิดเห็นที่พัฒนากับวิธีการเรียนรู้ของเครื่องอย่างเดียว วิธีการใช้คลังคำอย่างเดียว ผลการวิจัยพบว่า วิธีการ LMS ซึ่งใช้วิธีการใช้คลังคำร่วมกับการเรียนรู้ของเครื่องมีประสิทธิภาพการจำแนกสูงสุด คือ มีค่าความถูกต้องโดยเฉลี่ยเท่ากับ 85.40

Read [24] นำเสนอการจำแนกความคิดเห็นบนเว็บไซต์ทวิตเตอร์ โดยใช้ข้อมูลอีโมติคอน (Emoticons) ที่รวบรวมจาก Usenet news groups ประกอบด้วยข้อความอีโมติคอนที่บ่งบอกความคิดเห็นในด้านบวก และอีโมติคอนที่บ่งบอกความคิดเห็นในด้านลบ ใช้ Unigram Feature เป็นตัวแทนคุณลักษณะและแทนค่าด้วยความถี่ของการเกิดคำ วิธีการที่ใช้เป็นตัวจำแนก ประกอบด้วย 2 วิธีการ คือ ซัพพอร์ตเวกเตอร์แมชชีน และนาอีฟเบย์ การทดสอบใช้การแบ่งข้อมูลด้วยวิธี 3-fold cross validation ผลการวัดประสิทธิภาพ พบว่า การจำแนกความคิดเห็นด้วยวิธีการนาอีฟเบย์ มีค่าความถูกต้อง เท่ากับ 78.90 % ส่วนการจำแนกความคิดเห็นด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีน มีค่าความถูกต้อง 81.50%

Hamouda และคณะ [57] นำเสนอวิธีการจำแนกความคิดเห็นโดยใช้คลังคำศัพท์ที่สร้างด้วยวิธีการเรียนรู้ของเครื่อง ประกอบด้วย 2 กระบวนการหลัก คือ 1) กระบวนการเตรียมข้อมูล (Data Preparation Phase) ข้อมูลที่ใช้คือข้อความคิดเห็นเกี่ยวกับสินค้าบนเว็บไซต์อเมซอน (Amason Product Review Data Set) ซึ่งมีหลากหลายโดเมน เช่น หนังสือ กล้อง เครื่องเล่น mp3 เป็นต้น ข้อความคิดเห็นมีจำนนทั้งหมด 5,000,000 ข้อความ ประกอบด้วยข้อความและคะแนนที่ผู้แสดงความคิดเห็นกำหนดไว้ (Rating Score) มีค่าตั้งแต่ 1-5 ผู้วิจัยทำการคัดเลือกและแบ่งข้อมูลเป็น 2

กลุ่ม คือ ค่าคะแนนเท่ากับ 1-2 จัดอยู่ในกลุ่มข้อความคิดเห็นเชิงลบ และค่าคะแนนเท่ากับ 4-5 จัดอยู่ในกลุ่มข้อความคิดเห็นเชิงบวก หลังจากจัดกลุ่มแล้วได้ข้อความคิดเห็นจำนวนทั้งหมด 756,958 ความคิดเห็น ผู้วิจัยได้ทำการสุ่มมาจำนวน 25,000 ความคิดเห็น เลือกความคิดเห็นที่มีตัวอักษรไม่เกิน 500 จัดไว้ในกลุ่มของข้อความคิดเห็นทั่วไป (Normal Review) จากนั้นทำการตัดคำโดยใช้ตัวเลข สัญลักษณ์พิเศษที่ไม่ใช่คำที่ใช้ทั่วไป จากนั้นวิเคราะห์คำหรือวลี (Morphological Analysis) เพื่อหารากคำศัพท์ เช่น run คือรากคำศัพท์ของ runs, ran และ running เป็นต้น และกระบวนการที่ 2) คือ กระบวนการสร้างคลังคำศัพท์ (Lexicon Development Phase) ในกระบวนการนี้มีการเรียนรู้ข้อมูลชุดสอนโดยใช้วิธีการซัพพอร์ตเวกเตอร์แมชชีน จากนั้นทำการประเมินค่าน้ำหนักของคำที่จะเป็นตัวแทนคุณลักษณะโดยใช้ค่าความถี่ผกผัน (TF-IDF) ในการคำนวณหาค่าการจำแนกแล้วเก็บไว้ในคลังคำศัพท์สำหรับใช้ในการจำแนกต่อไป และหลังจากได้สร้างคลังคำศัพท์แล้ว ผู้วิจัยทดสอบการจำแนกข้อมูลโดยใช้ข้อมูลความคิดเห็นที่อยู่บนเว็บไซต์อเมซอน จำนวน 4,000 ความคิดเห็น ประกอบด้วย ความคิดเห็นเชิงบวก 2,000 ความคิดเห็น และความคิดเห็นเชิงลบ 2,000 ความคิดเห็น จากนั้นวัดประสิทธิภาพการจำแนกด้วย MLBSL กับ SentiWordNet, คลังคำศัพท์ที่รวบรวมเอง และคลังคำศัพท์ที่มาจากหลายแหล่งข้อมูล พบว่า การจำแนกความคิดเห็นจากคลังคำศัพท์ MLBSL มีประสิทธิภาพความถูกต้องสูงที่สุด คือ 71.75%

Lu และ Tsou [58] นำเสนอวิธีการจำแนกความคิดเห็นโดยการผสมผสานวิธีการใช้คลังคำกับวิธีการเรียนรู้ของเครื่อง โดยขั้นตอนแรกจะใช้คลังคำศัพท์ความคิดเห็นขนาดใหญ่จะถูกจัดไว้ในข้อมูลชุดสอน เรียกว่าวิธีการเรียนรู้แบบมีผู้สอน คือ คลังคำศัพท์โดยทั่วไปจะถูกรวบรวมโดยไม่คำนึงถึงบริบทหรือโดเมน งานวิจัยนี้ได้ปรับปรุงวิธีการสร้างคลังคำด้วยวิธีการเรียนรู้ของเครื่องเรียกว่า SVM-Lexicon โดยลบคำที่มีค่าความเชื่อมั่นต่ำในคลังคำศัพท์ออกรวมถึงวิเคราะห์คำกริยาที่ระบุความคิดเห็นเพิ่มในคลังคำความคิดเห็น ใช้ตัวจำแนกความคิดเห็น 3 วิธีการ ได้แก่ นาอีฟเบย์ แมกซิมัมเอ็นโทรปี และซัพพอร์ตเวกเตอร์แมชชีน ผลการวิจัยพบว่า การจำแนกความคิดเห็นด้วยวิธีการ SVM ร่วมกับ SVM-Lexicon มีประสิทธิภาพความถูกต้องและความแม่นยำสูงที่สุด คือ 74.20% และ 71.20% ตามลำดับ

วาทินี นุ้ยเพียร และ พยุง มีสัจ [59] นำเสนอการคัดเลือกคุณลักษณะเพื่อให้ได้คำที่เหมาะสมในการแทนเอกสารและเพิ่มประสิทธิภาพในการจำแนกเอกสารให้มีความถูกต้องมากขึ้น โดยทำการเปรียบเทียบการคัดเลือกคุณลักษณะเพื่อลดมิติของข้อมูลแบบการกรอง 3 วิธี ได้แก่ อินฟอร์เมชันเกน เกนเรโช และไคสแคร์ วิธีการดำเนินการประกอบด้วย 4 ขั้นตอน คือ ขั้นตอนที่ 1) การเตรียมข้อมูล ซึ่งข้อมูลที่ใช้ในการทดลองเป็นข้อมูลบทคัดย่อภาษาอังกฤษจากฐานข้อมูล ACM Digital Library ขั้นตอนที่ 2) การคัดเลือกคุณลักษณะเพื่อลดมิติข้อมูล การคัดเลือกคุณลักษณะใช้ 2 วิธี ได้แก่ การกรอง และการควบรวม ขั้นตอนที่ 3) การจำแนกประเภทข้อมูล ในการจำแนก

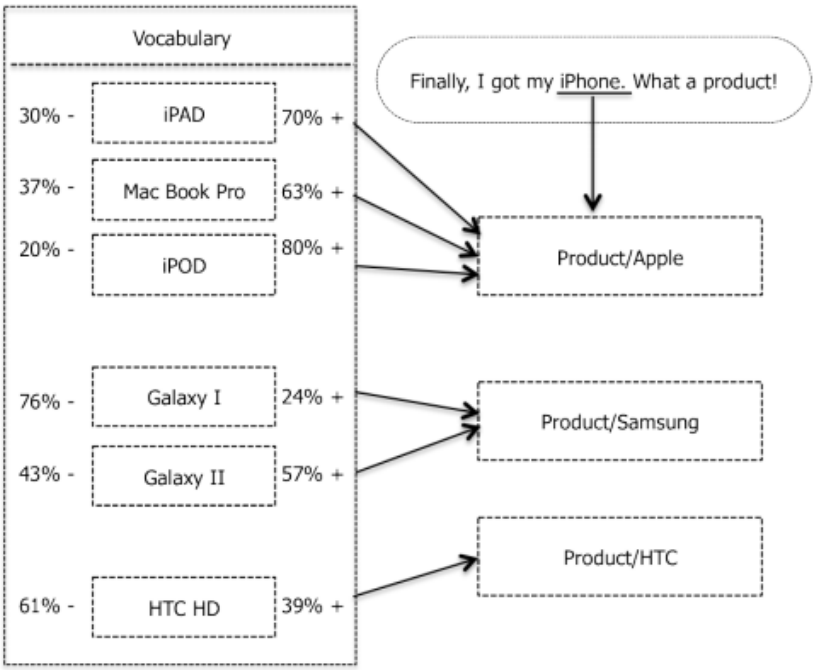
ข้อความมใช้วิธี นาอีฟเบย์ (NB) เบย์เซียนเน็ต (BN) เคเนียเรสต์เนเบอร์ (KNN) และ ซัพพอร์ตเวกเตอร์แมชชีน ขั้นตอนที่ 4) การประเมินประสิทธิภาพ ใช้วิธีการวัดค่าความแม่นจำ ค่าเรียกคืน และค่าประสิทธิภาพโดยรวม (F-measure) ผลการทดลอง พบว่า วิธีคัดเลือกแบบไคสแคร์ให้ผลดีที่สุด วัดประสิทธิภาพโดยรวม ได้เท่ากับ 82.20% และ การควบรวมใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) ร่วมกับการค้นหาด้วยวิธีเชิงพันธุกรรม (SVMG) และการค้นหาด้วยวิธีละโมบ (SVMGD) โดยวิธีการคัดเลือกแบบ SVMGD ให้ผลดีที่สุด วัดประสิทธิภาพโดยรวมได้เท่ากับ 94.00% ซึ่งการจำแนกข้อความทั้งสองวิธีใช้ขั้นตอนวิธีแบบซัพพอร์ตเวกเตอร์แมชชีนโดยใช้เคอร์เนลแบบเรเดียลเบสิสฟังก์ชัน (SVMR) เมื่อเปรียบเทียบประสิทธิภาพทั้งวิธีการกรองและการควบรวม สรุปได้ว่าประสิทธิภาพโดยรวมของการควบรวมมีค่ามากกว่าการกรอง 1.80% ซึ่งทำให้นักวิจัยสามารถนำเทคนิคของการควบรวมไปใช้เพิ่มประสิทธิภาพการจำแนกข้อความ

ฐิติมา เกษมศรีธนาวัฒน์ และ ธันสนี เพียตระกูล [60] นำเสนอวิธีการสรุปความเห็นจากทัศนคติที่เกี่ยวข้องกับหนังสือ โดยแบ่งทัศนคติเป็น 2 ด้าน คือ ด้านบวก (Positive) และด้านลบ (Negative) ข้อมูลนำมาใช้ในการทดสอบ เป็นข้อความคิดเห็นภาษาอังกฤษ รวบรวมมาจากความเห็นเกี่ยวกับหนังสือคอมพิวเตอร์ด้านโปรแกรมมิ่งบนเว็บไซต์ Amazon ในขั้นตอนการเตรียมข้อมูล ผู้วิจัยทำการเลือกคุณลักษณะ (Feature Selection) โดยใช้วิธีการ Principle Components Analysis และ Relief Algorithm การจำแนกความคิดเห็นใช้หลักการ Machine Learning ได้แก่ Naïve Bayes, Decision Tree (J48) และ Multi-Layer Perceptron การประเมินความถูกต้องของการจำแนกความคิดเห็น ได้ใช้การทดสอบแบบ 5-Fold Cross-Validation ทำการเปรียบเทียบผลการจำแนกแต่ละวิธี พบว่า Naïve Bayes ให้ความถูกต้องในการจำแนกสูงที่สุด คือ 68%

2.8.4 งานวิจัยที่นำเสนอวิธีการลดคุณลักษณะ

Saif และคณะ [45] นำเสนอวิธีการจัดกลุ่มคำเพื่อลดคุณลักษณะและแก้ไขปัญหาข้อมูลเบาบาง โดยคำที่มีความหมายเหมือนกันจะถูกจัดกลุ่มไว้ด้วยกัน ในงานวิจัยนี้มี 2 วิธีการ ที่ใช้ในการจัดกลุ่มคำ คือ Semantic Smoothing และ วิธีการ Automatic Sentiment-Topic Extraction วิธีการ Semantic Smoothing เป็นการสกัดความหมายที่ซ่อนอยู่ในข้อความทวิต และนำมารวบรวมไว้ในตัวจำแนกข้อมูลชุดสอน (Classifier Training) ที่ผ่านการแก้ไขแล้ว ตัวอย่างเช่น คำว่า “iPad”, “iPod” และ “Mac Book Pro” พบบ่อยมากในข้อความทวิตและเมื่อนำมาวิเคราะห์แล้วพบว่าส่วนใหญ่มีความคิดเห็นในเชิงบวก (Positive Polarity) จึงทำการสร้างรูปแบบเกี่ยวกับความหมายของคำเหล่านี้ไว้ในกลุ่ม “Product/Apple” เพื่อเก็บไว้เป็นผลของข้อมูลทดสอบที่จะเข้ามาใหม่ เช่น ข้อมูลทดสอบระบุว่า “I got my iPhone. What a product!” ผลการจำแนกการทดสอบจะกลายเป็น

ความคิดเห็นเชิงบวก ตามที่ได้รวบรวมไว้ในข้างต้น รายละเอียดวิธีการ Semantic Smoothing ดังรูปที่ 16



รูปที่ 16 *แสดงแนวคิดวิธีการ* Semantic Smoothing

อีกวิธีการคือ Automatic Sentiment-Topic Extraction เป็นรูปแบบที่ใช้การตรวจสอบทั้งหัวข้อและความคิดเห็นพร้อมกัน ประกอบด้วย 3 ขั้นตอน คือ เลือก 1 ความคิดเห็นจากเอกสาร จากนั้นเลือกหัวข้อที่เกี่ยวข้อง แล้วสร้างคลังหัวข้อและความคิดเห็นเป็นคลังคำความคิดเห็น โดยข้อความในแต่ละเทอมจะถูกจัดกลุ่มไว้ตามขั้วความคิดเห็นและอยู่ภายใต้หัวข้อที่กำหนด ดังแสดงในรูปที่ 17

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
Positive	dream	bought	song	eat	movi
	sweet	short	listen	food	show
	train	hair	love	coffe	award
	angel	love	music	dinner	live
	love	wear	play	drink	night
	goodnight	shirt	album	yummi	mtv
	free	dress	band	chicken	concert
	club	photo	guitar	tea	vote
Negative	feel	miss	rain	exam	job
	today	sad	bike	school	hard
	hate	cry	car	week	find
	sick	girl	stop	tomorrow	hate
	cold	gonna	ride	luck	interview
	suck	talk	hit	suck	lost
	weather	bore	drive	final	kick
	headache	feel	run	studi	problem

รูปที่ 17 *แสดงแนวคิดวิธีการ* Automatic Sentiment-Topic Extraction

ข้อมูลที่ใช้ในการทดสอบรวบรวมจาก Stanford Twitter Sentiment Dataset โดยสุ่มข้อความคิดเห็นจำนวน 1000 ข้อความ เป็นข้อมูลทดสอบ ตัวจำแนกที่ใช้ทำสอบคือ นาอีฟเบย์ ผลการทดลองพบว่า การเลือกชุดคุณลักษณะทั้งสองแบบ จะมีประสิทธิภาพดีกว่าการเลือกคุณลักษณะแบบพื้นฐาน โดยการเลือกคุณลักษณะโดยใช้วิธีการ Semantic Smoothing มีประสิทธิภาพความถูกต้องในการจำแนก เท่ากับ 84.0 % และ การเลือกคุณลักษณะโดยวิธีการ Automatic Sentiment-Topic Extraction ประสิทธิภาพความถูกต้องในการจำแนก เท่ากับ 86.3% 2 ข้อเสียของวิธีการนี้คือ จะต้องมีการระบุจำนวนหัวข้อไว้ล่วงหน้า ซึ่งเป็นเรื่องยากที่จะคาดเดาจำนวนที่เหมาะสมของหัวข้อ

Sayfullina [61] นำเสนอวิธีการเลือกคุณลักษณะที่มีความสำคัญและรวมคำคุณลักษณะที่มีความหมายเหมือนกันไว้ด้วยกัน ข้อมูลที่ใช้ในการวิจัยรวบรวมจาก KDD Cup และ SemEval 2013 ประกอบด้วย 5 ขั้นตอน คือ 1) เลือกเทอมจากจำนวนการเกิดของคำ ใช้รูปแบบ Unigram และ Bigrams พจนานุกรมความคิดเห็น ซึ่งได้ระบุสีที่สื่ออารมณ์ให้แต่ละคำ สำหรับคำศัพท์ของชุดข้อมูลทวิตเตอร์ ประกอบด้วย คำปฏิเสธ คำแสลง และอีโมติคอน 2) กระบวนการเตรียมข้อความ กำจัดคำซ้ำ 3) หารากคำศัพท์ 4) จัดอันดับคุณลักษณะ ด้วยค่า Tf-Idf 5) รวมคำที่มีความหมายคล้ายกันโดยใช้อัลกอริทึมของ Word Clustering ร่วมกับการใช้คลังคำใน WordNet มาตรวจสอบความหมายว่าใกล้เคียงกันหรือไม่ ถ้าหากเป็นคำที่มีความหมายใกล้เคียงกันให้รวมเป็นคุณลักษณะเดียวกัน ซึ่งการใช้วิธีการนี้ช่วยลดปัญหาข้อมูลเบาบาง จำนวนคุณลักษณะลดลงจาก 5029 เหลือ 3182 คุณลักษณะ นำไปทดสอบการจำแนกความคิดเห็นโดยวิธีการนาอีฟเบย์ พบว่าการลดคุณลักษณะด้วยวิธีการที่นำเสนอช่วยให้ประสิทธิภาพการจำแนกดีขึ้น

Saif และคณะ [62] นำเสนอวิธีการลดคุณลักษณะ 2 วิธีการ คือ Shallow Semantic Smoothing เป็นวิธีการลดจำนวนคำศัพท์โดยการวิเคราะห์จากประโยคว่าอยู่ในกลุ่มคำศัพท์ใด แล้วจะจัดกลุ่มไว้ในกลุ่มคำศัพท์ที่ได้ตั้งค่าไว้ เช่น ประโยค “Downloading apps for my iPhone! So much fun :)” จะถูกจัดอยู่ในกลุ่มคำ “Product” วิธีการนี้จะช่วยลดขนาดคำศัพท์ของข้อมูลชุดสอนลงและช่วยลดปัญหาข้อมูลเบาบาง อีกวิธีการหนึ่ง คือ วิธีการ Semantic Smoothing for Naïve Bayes Classifier เป็นการใช้หลักการข้างต้นเพื่อมาประยุกต์ใช้ร่วมกับวิธีการนาอีฟเบย์ โดยดูค่าความสัมพันธ์ของคำด้วย ข้อมูลที่ใช้ในการวิจัยครั้งนี้เป็นข้อมูลทวิตเตอร์ ซึ่งประกอบด้วย 1.6 ล้าน ข้อความคิดเห็น รวบรวมโดย Go และคณะ ในการทดลองผู้วิจัยทำการเลือกข้อมูลมาจำนวน 60,000 ข้อความคิดเห็น สำหรับเป็นชุดสอนและข้อมูลชุดทดสอบ ตัวจำแนกที่ใช้คือ นาอีฟเบย์ ผลการวิจัย พบว่า การเลือกคุณลักษณะด้วยวิธีการ Shallow Semantic Smoothing ให้ประสิทธิภาพความถูกต้องเท่ากับ 76.30 ส่วนการเลือกคุณลักษณะด้วยวิธีการ Semantic Smoothing for Naïve Bayes Classifier ช่วยให้ประสิทธิภาพการจำแนกเพิ่มขึ้นเป็น 81.30% ซึ่งจากการวิเคราะห์งานวิจัยนี้ จะเห็นว่าการเลือกคุณลักษณะโดยการวิเคราะห์รูปประโยค อาจจะทำให้สูญเสียคุณลักษณะที่สำคัญซึ่งมีผลต่อตัวจำแนก ส่วนการปรับปรุงคุณลักษณะสำหรับตัวจำแนกโดยตรงจะช่วยเพิ่มประสิทธิภาพการจำแนกได้

Ong [63] นำเสนอวิธีการปรับปรุงข้อมูลเบาบางด้วยค่าการเพิ่มของข้อมูล (Sparsity Adjusted Information Gain: SAIG) ซึ่งได้ปรับเปลี่ยนแมทริกซ์ค่าการเพิ่มข้อมูลแบบเดิมและปรับปรุงค่าการจัดอันดับตามค่าความเบาบางของเวกเตอร์คุณลักษณะ ทำให้คุณลักษณะลดลงแต่ประสิทธิภาพยังคงอยู่ในระดับตามเป้าหมาย โดยนำข้อมูลความถี่ของเทอม ความถี่ของเอกสาร ความเบาบาง (Sparsity) ความหนาแน่น (Density) เพื่อมาปรับปรุงการเลือกคุณลักษณะ ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลวิจารณ์ผลิตภัณฑ์และข้อมูลวิจารณ์ภาพยนตร์ ขั้นการทดลองเริ่มจากการเตรียมข้อมูล ทำการลบข้อความที่ซ้ำกันออกจากชุดข้อมูล จากนั้นลบข้อความที่พิมพ์อักษรซ้ำๆกันเกิน 3 ครั้ง เช่น “yaaaaaaayyyy!!!!!” จะถูกแก้ไขเป็น “yaaay!!!” หลังจากนั้นทำการตัดคำโดยกำหนดขอบเขตของคำจาก ช่องว่าง (Whitespace) และอักขระพิเศษ เช่น เครื่องหมายคำถามและวงเล็บ จากนั้นทำการตรวจสอบแก้ไขคำที่สะกดผิด ลบคำหยุด และหารากคำศัพท์ ผู้วิจัยใช้ขั้นตอนวิธีการจำแนก 3 วิธีการ ได้แก่ นาอีฟเบย์ ซัพพอร์ตเวกเตอร์แมชชีน และ เพื่อนบ้านใกล้ที่สุด ผลการวิจัยนำเสนอโดยการเปรียบเทียบระหว่างการเลือกคุณลักษณะโดยค่าการเพิ่มของข้อมูลแบบเดิม (Information Gain) กับการเลือกคุณลักษณะด้วยวิธีการปรับปรุงข้อมูลเบาบางด้วยค่าการเพิ่มของข้อมูล (SAIG) พบว่า เมื่อข้อมูลมีจำนวนมาก การเลือกคุณลักษณะด้วยวิธีการ SAIG ร่วมกับการจำแนกด้วยวิธีการซัพพอร์ตเวกเตอร์แมชชีน มีประสิทธิภาพการจำแนกสูงกว่าวิธีการอื่น

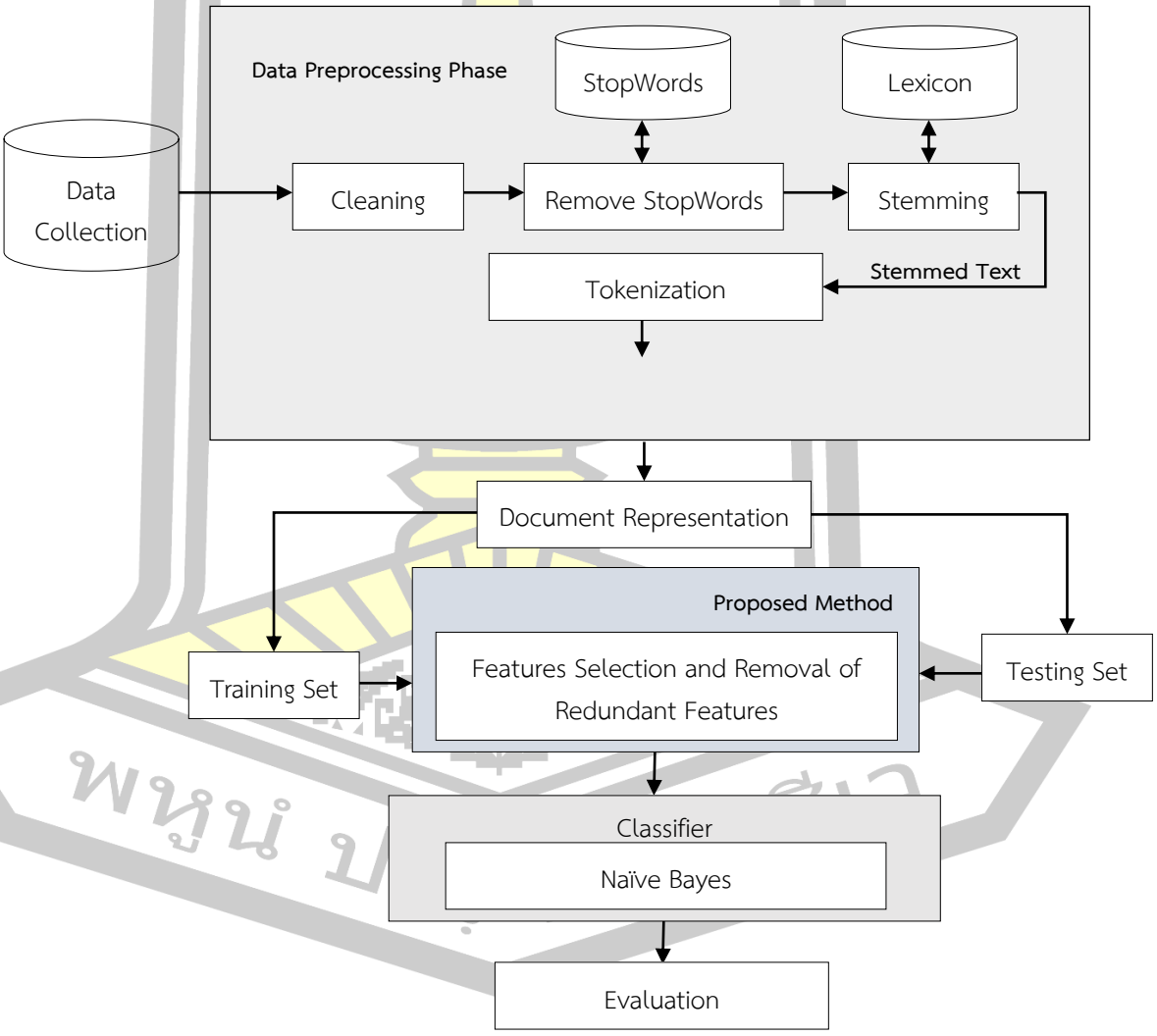
Parlar และคณะ [64] นำเสนอวิธีการใหม่ในการเลือกคุณลักษณะสำหรับการจำแนกความคิดเห็น โดยใช้วิธีการกำหนดค่าน้ำหนักของคุณลักษณะด้วยการประยุกต์ใช้วิธีการดึงข้อมูลสารสนเทศ (Information Retrieval) ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลวิจารณ์ผลิตภัณฑ์และข้อมูลวิจารณ์ภาพยนตร์ภาษาตุรกี (Turkish Languages) และภาษาอังกฤษ (English Languages) ทดสอบการคัดเลือกคุณลักษณะที่นำเสนอ เปรียบเทียบกับการคัดเลือกคุณลักษณะที่ 4 วิธีการ ได้แก่ วิธีการ Chi Square วิธีการ Information Gain วิธีการ Document Frequency Difference และ วิธีการ Optimal Orthogonal Centroid เปรียบเทียบการจำแนกความคิดเห็น 4 วิธีการ ได้แก่ Naïve Bayes Multinomial, Support Vector Machines, Maximum Entropy และ Decision Trees ผลการวิจัย พบว่า เมื่อใช้วิธีการ Naïve Bayes Multinomial ในการจำแนกความคิดเห็นทั้งข้อความที่เป็นภาษาตุรกีและข้อความภาษาอังกฤษ วิธีการคัดเลือกคุณลักษณะที่นำเสนอมีประสิทธิภาพสูงกว่าวิธีการอื่น

Pratiwi และ Adiwijaya [65] นำเสนอวิธีการลดคุณลักษณะสำหรับการจำแนกความคิดเห็นบทวิจารณ์ภาพยนตร์ ข้อมูลที่ใช้ในการวิจัย เป็นข้อมูลวิจารณ์ภาพยนตร์ Benchmark Dataset ประกอบด้วย ความคิดเห็นเชิงบวก จำนวน 1,000 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 1,000 ข้อความ การทดสอบแบ่งข้อมูลด้วยวิธีการ 10 Fold Cross-Validation Test การทดลองประกอบด้วย 10 ขั้นตอน คือ 1) อ่านข้อมูลที่รวบรวม (Reading the Dataset) 2) ลบข้อความที่ไม่ใช่ตัวอักษร (Non Alphabetic Removal) 3) ตัดข้อความ (Tokenization) 4) กำจัดคำหยุด (Stopwords Removal) 5) หาราคำศัพท์ (Stemming) 6) สร้างชุดคำศัพท์ (Initial Vocabulary Construction) 7) สร้างเมตริกซ์คุณลักษณะ (Initial Feature Matrix Construction 8) ตรวจสอบคุณลักษณะที่ปรากฏในข้อมูลชุดสอน ถ้าคุณลักษณะปรากฏในทั้งสองคลาสจำนวนเท่ากัน แสดงว่าคุณลักษณะนั้นไม่มีความสำคัญต่อการจำแนก 8) คัดเลือกคุณลักษณะด้วยวิธีการที่ IGDFFS 10) สร้างคลังคำศัพท์ (Dictionary Construction) ผลการวิจัย พบว่า การคัดเลือกคุณลักษณะด้วยวิธี IGDFFS ช่วยลดคุณลักษณะที่ไม่สำคัญได้มากกว่า 90% ส่วนวิธีการจำแนกความคิดเห็นที่นำเสนอ มีค่าประสิทธิภาพความถูกต้องในการจำแนกความคิดเห็น 96%

จากงานวิจัยที่เกี่ยวข้องแสดงให้เห็นว่าการคัดเลือกคุณลักษณะมีความสำคัญต่อประสิทธิภาพการจำแนกความคิดเห็น งานวิจัยที่ผ่านมาส่วนใหญ่ทำการคัดเลือกคุณลักษณะโดยดูลักษณะการเกิดในเอกสาร การจัดกลุ่มคำที่เหมือนกันไว้ด้วยกัน จะเห็นว่าการปรากฏของคำในเอกสารมีความสำคัญต่อการคัดเลือกคุณลักษณะ งานวิจัยนี้ผู้วิจัยได้ประยุกต์ใช้แนวคิดวิธีฟิลเตอร์โมเดลผสมผสานกับแนวคิดวิธีการใช้กฎความสัมพันธ์ เพื่อวิเคราะห์ความสำคัญของคุณลักษณะที่มีต่อคลาส ซึ่งเป็นวิธีการที่ง่าย รวดเร็วและมีประสิทธิภาพ นอกจากนี้แล้วงานวิจัยนี้ได้นำเสนอขั้นตอนวิธีการขจัดคุณลักษณะที่ซ้ำซ้อนที่ไม่ส่งผลกระทบต่อตัวจำแนกและยังช่วยลดระยะเวลาในการสร้างตัวจำแนกและทดสอบตัวจำแนกได้

บทที่ 3
วิธีดำเนินการวิจัย

การดำเนินการวิจัยให้บรรลุวัตถุประสงค์ ผู้วิจัยได้อาศัยแนวคิดและวิธีดำเนินการวิจัยตามทฤษฎี และงานวิจัยที่เกี่ยวข้องกับการจำแนกความคิดเห็นที่อยู่บนเว็บไซต์เครือข่ายสังคมออนไลน์ โดยการดำเนินงานวิจัย ประกอบไปด้วย 6 ขั้นตอน ได้แก่ 1) การรวบรวมข้อมูล (Collected Data) 2) การเตรียมข้อมูล (Data Preprocessing) 3) การเลือกคุณลักษณะ (Feature Selection) และการกำจัดคุณลักษณะที่ซ้ำซ้อน (Removal of Redundant Features) 4) การแบ่งข้อมูล (Data Partitioning) 5) การสร้างตัวจำแนก (Classifier) และ 6) การวัดประสิทธิภาพการจำแนก (Evaluation) ดังรูปที่ 18



รูปที่ 18 *ขั้นตอนการดำเนินการวิจัย*

จากรูปที่ 18 แสดงขั้นตอนการดำเนินการวิจัย โดยเริ่มจากกระบวนการรวบรวมข้อมูลความคิดเห็นบนเครือข่ายสังคมออนไลน์ ในงานวิจัยนี้ผู้วิจัยใช้ข้อมูลความคิดเห็นที่รวบรวมจากเว็บไซต์ทวิตเตอร์ จากนั้นนำเข้าสู่กระบวนการเตรียมข้อมูล ประกอบด้วย กระบวนการทำความสะอาดข้อความ โดยการลบตัวอักษรซ้ำและแก้ไขคำที่สะกดผิด กระบวนการลบคำหยุดหรือคำที่ไม่มีนัยสำคัญในการจำแนกความคิดเห็น โดยใช้คลังคำศัพท์ที่มีอยู่แล้ว จากนั้นทำการหารากคำศัพท์ โดยใช้คลังคำศัพท์ จากนั้นนำเข้าสู่กระบวนการตัดคำ งานวิจัยนี้ตัดคำด้วยวิธีการ Unigrams ร่วมกับ Bigrams โดยข้อความที่มีคำปฏิเสธ เช่น no, not จะใช้หลักการตัดคำด้วยวิธีการ Bigrams ส่วนข้อความอื่น ๆ จะใช้หลักการตัดคำด้วย Unigrams ซึ่งคุณลักษณะได้จากการตัดคำ จากนั้นเข้าสู่กระบวนการแทนค่าน้ำหนักในเอกสารด้วยวิธีการ Boolean Weighting โดยหากพบคุณลักษณะในเอกสาร จะให้มีค่า เท่ากับ 1 หาก ไม่พบให้มีค่า เท่ากับ 0 ขนาดเวกเตอร์จะมีขนาดเท่ากับขนาดเอกสาร x ขนาดของคุณลักษณะ หลังจากได้เวกเตอร์แล้วทำการแบ่งข้อมูลเป็น 2 ชุด ได้แก่ ข้อมูลชุดสอน (Training Set) และข้อมูลชุดทดสอบ (Testing Set) นำข้อมูลชุดสอนมาจัดลำดับความสำคัญของคุณลักษณะ เพื่อเข้าสู่กระบวนการเลือกคุณลักษณะและขจัดคุณลักษณะที่ซ้ำซ้อนและสร้างตัวจำแนก เพื่อนำไปเลือกคุณลักษณะและจำแนกความคิดเห็นข้อมูลชุดทดสอบ แล้วทำการประเมินประสิทธิภาพของการจำแนก วิเคราะห์ผลการทดลองเพื่อปรับปรุงประสิทธิภาพต่อไป รายละเอียดแต่ละกระบวนการมีดังต่อไปนี้

3.1 การรวบรวมข้อมูล (Data Collected)

ผู้วิจัยรวบรวมข้อมูลความคิดเห็นที่รวบรวมจากเครือข่ายสังคมออนไลน์จำนวน 5 ชุดข้อมูล ได้แก่ 1) Stadford Twitter Sentiment Data [11] 2) SemEval-2017 Task4A Dataset (SemEval) [12] 3) Sentiment Strength Twitter Dataset (SS-Tweet) [13] 4) Health Care Reform (HCR) [14] 5) Sanders Twitter Dataset [15] รายละเอียดข้อมูลแต่ละชุด ได้แก่ จำนวนข้อมูลทั้งหมด จำนวนข้อมูลในแต่ละคลาส แสดงดังตาราง 7 ในการทดลองผู้วิจัยทำการสุ่มข้อความโดยเลือกข้อความคิดเห็นที่เป็นข้อความคิดเห็นเชิงบวกและข้อความคิดเห็นเชิงลบจำนวนเท่ากัน ชุดข้อมูลที่ใช้ในการวิจัย แสดงดังตาราง 8 ส่วนรายละเอียดด้านคุณลักษณะอื่น ๆ ได้แก่ ความยาวสูงสุด ความยาวสั้นสุด และความยาวเฉลี่ย ผู้วิจัยนำข้อมูลที่รวบรวมทั้งหมดมาวิเคราะห์ ผลการวิเคราะห์ข้อมูลแสดงดังตาราง 9

ตาราง 7 จำนวนชุดข้อมูลทั้งหมด

ชุดข้อมูล	จำนวนข้อมูลทั้งหมด	จำนวนข้อมูลในแต่ละคลาส				
		Negative	Positive	Natural	Other	Irrelevant
STS	1,600,000	800,000	800,000	-	-	-
SemEval	13,975	2,186	5,349	6,440	-	-
SS-Twitter	4,242	1,037	1,252	1,953	-	-
HCR	2,516	1,381	541	470	45	79
Sanders	5,513	654	570	2,503	-	1,786

ตาราง 8 ชุดข้อมูลที่ใช้ในการวิจัย

ชุดข้อมูล	จำนวนข้อความคิดเห็น	จำนวนข้อมูลในแต่ละคลาส	
		Negative	Positive
STS	10,000	5,000	5,000
SemEval	4,000	2,000	2,000
SS-Twitter	2,600	1,300	1,300
HCR	1,000	500	500
Sanders	1,000	500	500

ตาราง 9 คุณลักษณะของชุดข้อมูลที่ใช้ในการวิจัย

ชุดข้อมูล	ความยาวสูงสุด (ตัวอักษร)	ความยาวสั้นสุด (ตัวอักษร)	ความยาวเฉลี่ย (ตัวอักษร)
STS	152	8	74
SemEval	185	11	109
SS-Twitter	164	4	97
HCR	142	23	113
Sanders	148	9	97

ข้อมูลแต่ละชุดมีรายละเอียดดังนี้

1) Stanford Twitter Sentiment Data (STS) ประกอบด้วย ข้อความคิดเห็นข้อมูลชุดสอน จำนวน 1,600,000 ข้อความ แบ่งเป็นความคิดเห็นเชิงบวก (Positive) จำนวน 800,000 ข้อความ และความคิดเห็นเชิงลบ (Negative) จำนวน 800,000 ข้อความ รูปแบบของข้อมูลเป็นไฟล์เอกสารประเภท Comma Separated Value (CSV) ประกอบด้วย 6 ฟิลด์ ได้แก่ 1) ขั้วความคิดเห็น (Polarity of the Tweet) มี 2 ด้าน คือ 0 หมายถึง ความคิดเห็นเชิงลบ และ 4 หมายถึง ความคิดเห็นที่เป็นเชิงบวก 2) ลำดับ (ID of the Tweet) มีรูปแบบเป็นตัวเลขจำนวนเต็ม 3) วัน/เวลา (Date of the Tweet) เช่น Sat May 16 23:58:44 UTC 2009 4) ประเด็นที่โพสต์ (Domain) เช่น Obama, Nike, San Francisco เป็นต้น หากไม่ระบุจะมีค่าเป็น NO_QUERY ซึ่งข้อมูลชุดสอนจะไม่ได้ระบุประเด็นที่โพส 5) ชื่อผู้โพสต์ (User that Tweeted) และ 6) ข้อความคิดเห็น (Text of the Tweet) ตามลำดับ จากการวิเคราะห์คุณลักษณะข้อมูล พบว่า ข้อมูลชุดนี้มีความยาวสูงสุดเท่ากับ 152 ตัวอักษร ความยาวสั้นสุด เท่ากับ 8 ตัวอักษร และมีความยาวเฉลี่ย เท่ากับ 74 ตัวอักษร ในการวิจัยนี้ผู้วิจัยทำการเลือกข้อมูลจำนวน 2 ฟิลด์ ได้แก่ ขั้วความคิดเห็น และข้อความคิดเห็น และทำการสุ่มข้อความคิดเห็นเพื่อการทดลอง จำนวน 10,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 5,000 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 5,000 ข้อความ ตัวอย่างชุดข้อมูล Stanford Twitter Sentiment Data ดังแสดงในรูปที่ 19

```
"0","1467810369","Mon Apr 06 22:19:45 PDT 2009","NO_QUERY","_TheSpecialOne_","@switchfoot http://twitpic.com/2y1zl - Awww, that's a bum
"0","1467810672","Mon Apr 06 22:19:49 PDT 2009","NO_QUERY","scotthamilton","is upset that he can't update his Facebook by texting it...
"0","1467810917","Mon Apr 06 22:19:53 PDT 2009","NO_QUERY","mattycus","@Kenichan I dived many times for the ball. Managed to save 50%
"0","1467811184","Mon Apr 06 22:19:57 PDT 2009","NO_QUERY","ElleCTF","my whole body feels itchy and like its on fire "
"0","1467811193","Mon Apr 06 22:19:57 PDT 2009","NO_QUERY","Karoli","@nationwideclass no, it's not behaving at all. i'm mad. why am i h
"0","1467811372","Mon Apr 06 22:20:00 PDT 2009","NO_QUERY","joy_wolf","@Kwesidei not the whole crew "
"0","1467811592","Mon Apr 06 22:20:03 PDT 2009","NO_QUERY","mybirch","Need a hug "
"0","1467811594","Mon Apr 06 22:20:03 PDT 2009","NO_QUERY","coZZ","@LOLTrish hey  long time no see! Yes.. Rains a bit ,only a bit  LOL
"0","1467811795","Mon Apr 06 22:20:05 PDT 2009","NO_QUERY","2Hood4Hollywood","@Tatiana_K nope they didn't have it "
"0","1467812025","Mon Apr 06 22:20:09 PDT 2009","NO_QUERY","mimismo","@twittera que me muera ? "
"0","1467812416","Mon Apr 06 22:20:16 PDT 2009","NO_QUERY","erinx3leannexo","spring break in plain city... it's snowing "
"0","1467812579","Mon Apr 06 22:20:17 PDT 2009","NO_QUERY","pardonlauren","I just re-pierced my ears "
"0","1467812723","Mon Apr 06 22:20:19 PDT 2009","NO_QUERY","TLeC","@caregiving I couldn't bear to watch it.  And I thought the UA loss
"0","1467812771","Mon Apr 06 22:20:19 PDT 2009","NO_QUERY","robrobbierobert","@octolinz16 It it counts, idk why I did either. you never
"0","1467812784","Mon Apr 06 22:20:20 PDT 2009","NO_QUERY","bayofwolves","@smarrison i would've been the first, but i didn't have a gun
"0","1467812799","Mon Apr 06 22:20:20 PDT 2009","NO_QUERY","HairByJess","@iamjazzyfizzle I wish I got to watch it with you!! I miss you
"0","1467812964","Mon Apr 06 22:20:22 PDT 2009","NO_QUERY","lovesongwriter","Hollis' death scene will hurt me severely to watch on film
"0","1467813137","Mon Apr 06 22:20:25 PDT 2009","NO_QUERY","armotley","about to file taxes "
"0","1467813579","Mon Apr 06 22:20:31 PDT 2009","NO_QUERY","starkissed","@LettyA ahh ive always wanted to see rent  love the soundtrack
"0","1467813782","Mon Apr 06 22:20:34 PDT 2009","NO_QUERY","gi_gi_bee","@FakerPattyPattz Oh dear. Were you drinking out of the forgotte
"0","1467813985","Mon Apr 06 22:20:37 PDT 2009","NO_QUERY","quanvu","@alydesigns i was out most of the day so didn't get much done "
"0","1467813992","Mon Apr 06 22:20:38 PDT 2009","NO_QUERY","swinspeedx","one of my friend called me, and asked to meet with her at Mid
"0","1467814119","Mon Apr 06 22:20:40 PDT 2009","NO_QUERY","cooliodoc","@angry_barista I baked you a cake but I ated it "
"0","1467814180","Mon Apr 06 22:20:40 PDT 2009","NO_QUERY","viJILLante","this week is not going as i had hoped "
"0","1467814192","Mon Apr 06 22:20:41 PDT 2009","NO_QUERY","Ljelli3166","blagh class at 8 tomorrow "
"0","1467814438","Mon Apr 06 22:20:44 PDT 2009","NO_QUERY","ChicagoCubbie","I hate when I have to call and wake people up "
"0","1467814783","Mon Apr 06 22:20:50 PDT 2009","NO_QUERY","KatieAngell","Just going to cry myself to sleep after watching Marley and M
"0","1467814883","Mon Apr 06 22:20:52 PDT 2009","NO_QUERY","gagoo","im sad now  Miss.Lilly"
"0","1467815199","Mon Apr 06 22:20:56 PDT 2009","NO_QUERY","abel209","ooooh.... LOL  that leslie.... and ok I won't do it again so lesl
"0","1467815753","Mon Apr 06 22:21:04 PDT 2009","NO_QUERY","BaptisteTheFool","Meh... Almost Lover is the exception... this track gets m
"0","1467815923","Mon Apr 06 22:21:07 PDT 2009","NO_QUERY","fatkat309","some1 hacked my account on aim  now i have to make a new one"
"0","1467815924","Mon Apr 06 22:21:07 PDT 2009","NO_QUERY","EmCDL","@alielayus I want to go to promote GEAR AND GROOVE but unfornately
"0","1467815988","Mon Apr 06 22:21:09 PDT 2009","NO_QUERY","merisssa","thought sleeping in was an option tomorrow but realizing that it
"0","1467816149","Mon Apr 06 22:21:11 PDT 2009","NO_QUERY","Pbearfox","@julieebaby awe i love you too!!!! 1 am here  i miss you"
"0","1467816665","Mon Apr 06 22:21:21 PDT 2009","NO_QUERY","jsoo","@HumpNinja I cry my asian eyes to sleep at night "
"0","1467816749","Mon Apr 06 22:21:20 PDT 2009","NO_QUERY","scarletletterm","ok I'm sick and spent an hour sitting in the shower cause
"0","1467817225","Mon Apr 06 22:21:27 PDT 2009","NO_QUERY","crosland_12","@cocomix04 ill tell ya the story later  not a good day and il
"0","1467817374","Mon Apr 06 22:21:30 PDT 2009","NO_QUERY","ajaxpro","@MissXu sorry! bed time came here (GMT+1)  http://is.gd/fNge"
"0","1467817502","Mon Apr 06 22:21:32 PDT 2009","NO_QUERY","Tmttq86","@fleurylis I don't either. Its depressing. I don't think I even w
"0","1467818007","Mon Apr 06 22:21:39 PDT 2009","NO_QUERY","Anthony_Nguyen","Bed. Class 8-12. Work 12-3. Gym 3-5 or 6. Then class 6-10.
"0","1467818020","Mon Apr 06 22:21:39 PDT 2009","NO_QUERY","itsanimesh","really don't feel like getting up today... but got to study to
```

รูปที่ 19 ตัวอย่างข้อมูลความคิดเห็นจาก Stanford Twitter Sentiment Data

2) SemEval-2017 Task4A Dataset (SemEval) เป็นข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์ จำนวน 13,975 ข้อความ รูปแบบของไฟล์ เป็นเอกสารประเภท text file (.txt) ประกอบด้วย 3 ฟิลด์ ได้แก่ 1) ลำดับของข้อความคิดเห็น (SID) 2) ขั้วความคิดเห็น (Polarity) 3) ข้อความคิดเห็น (Twitter Message) ประกอบด้วย 3 คลาส ได้แก่ ความคิดเห็นเชิงบวก (Positive) ความคิดเห็นเชิงลบ (Negative) และความคิดเห็นที่เป็นกลาง (Neutral) ข้อมูลชุดนี้ประกอบด้วยข้อความคิดเห็นเชิงบวก จำนวน 5,349 ข้อความ ข้อความคิดเห็นเชิงลบ จำนวน 2,186 ข้อความ และข้อความที่เป็นกลาง จำนวน 6,440 ข้อความ จากการวิเคราะห์คุณลักษณะข้อมูล พบว่า ข้อมูลชุดนี้มีความยาวสูงสุด เท่ากับ 185 ตัวอักษร ความยาวสั้นสุด เท่ากับ 11 ตัวอักษร และมีความยาวเฉลี่ย เท่ากับ 109 ตัวอักษร ผู้วิจัยได้ทำการสุ่มข้อความคิดเห็นเพื่อใช้ในการทดลอง จำนวน 4,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 2,000 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 2,000 ข้อความ ตัวอย่างข้อมูลชุด SemEval-2017 Task4A Dataset แสดงดังรูปที่ 17

```
    801989080477154944      neutral #ArianaGrande Ari By Ariana Grande 80% Full https://t.co/ylhCMETHHW #Singe
2   801989272341453952      positive        Ariana Grande KIIS FM Yours Truly CD listening party in Burbank ht
3   801990978424962944      positive        Ariana Grande White House Easter Egg Roll in Washington https://t.
4   801996232553963008      positive        #CD #Musics Ariana Grande Sweet Like Candy 3.4 oz 100 ML Sealed In
5   801998343442407040      neutral SIDE TO SIDE 😮 @arianagrande #sidetoside #arianagrande #musically #comunic
6   802001659970744064      positive        Hairspray Live! Previews at the Macy's Thanksgiving Day Parade! ht
7   802003380973568000      positive        #LindsayLohan Is 'Feeling Thankful' After Blasting #ArianaGrande F
8   802014830467174016      neutral I hate her but... I love her songs Dammit ._.#ArianaGrande
9   802020578609623040      neutral Ariana Grande【Right There ft. Big Sean】#アリアナ #arianagrande https://t.co/
10  802021059490934016      positive        which one would you prefer to listen to for a whole day? 😊👹; i cou
11  802021122384523008      neutral Booty Baby Ari#ArianaGrande #PrincessAri #bootybaby #DangerousWomanTour #D
12  802022296818880000      neutral #LindsayLohan backs out of a #Kettering holiday appearance, just after thr
13  802022559520673024      positive        My idols are #littlemix #justinbieber #arianagrande
14  802022965743415040      neutral Ariana Grande - The Sims 3 - SimsDomination #ArianaGrande https://t.co/l2K
15  802024079851069056      neutral #Music #ArianaGrande-THE REMIX-JAPAN ONLY #CD E78 https://t.co/zOxb3hwBpg
16  802024085777629056      positive        #Beauty #ArianaGrande-CHRISTMAS & CHILL-JAPAN ONLY #CD BONUS TRACK
17  802024981638958976      neutral【Popular Song】Ahh, I said I'm putting down my story in a popular song #Tay
18  802026442687217024      neutral Ariana Grande Private Event for Coach in Japan, August 2015 https://t.co/5
19  802028835013206016      positive        #Beauty #ArianaGrande-THE REMIX-JAPAN ONLY #CD E78 https://t.co/ga
20  802028852180462976      positive        4. One last time - #ArianaGrandehttps://t.co/0o4T0q3GsY
21  802030814246489984      positive        More #newarrivals #pentatonix #christmasalbum #arianagrande #frank
22  802032149885026048      positive        so much love for this woman, ughh 😭😭 #arianagrande https://t.co/3
23  802032523152718976      neutral Ariana lips💄#mac #ariana #arianagrande https://t.co/T4aK8LagqX
24  802041485948555008      neutral Ariana Grande - 2015 NYC Pride Dance On The Pier https://t.co/9xe1KJuPOe #
25  802042842096902016      positive        thanks God it's Friday. #Thanksgiving #arianagrande #arianator😉 ht
26  802045545992723968      neutral #MUSIC #ArianaGrande@karinrino PM16:00Ariana Grande - Baby I https://t.co/
27  802045726318587008      positive        New on @Twitter . Big fan of @NICKIMINAJ and @ArianaGrande #Ariana
28  802198774114385024      negative        Soros brainwashes & enslaves U #blacklivesmatter protesting fools.
29  802198896474803968      neutral Obama's latest #BlackLivesMatter recruiting videoWelcome to the Modesto ma
30  802198916489874944      negative        Iggy from the"Revolution Club" mad that Trump called #Chicago a "
31  802198916594887936      negative        #BlackLivesMatter #potus #HillaryClinton You called for this. Gett
32  802199118030343936      neutral @Sosa_Baby49 so you also don't even understand what #BlackLivesMatter actu
33  802199636987416064      neutral Lake County Black Lives Matter hosts Thanksgiving food distribution - Chic
34  802199759872109952      negative        #BlackLivesMatter protester hit by Black drivers https://t.co/pag7
35  802202968527252992      neutral There's not enough hair to track all the death. #BlackLivesMatter Reminds
36  802203076522364032      negative        When will #BlackLivesMatter protest the black violence that kills
37  802203170923548032      positive        @riseupnet @AnonUKRadio @IGD_News @BLM5280 We must unite ALL again
38  802203261923115008      neutral @TahjDeathstyle Donald Trump is great for so many reasons. He will label #
39  802203369007943040      neutral The latest Hasem Ben Sober's Daily Mulch! https://t.co/nOgOWZkxZ7 #blackfr
40  802203384904354944      negative        Electoral College must reject Trump https://t.co/5L8ai9piHj #rejec
41  802203403560647936      negative        Yeah police action #BlackLivesMatter that's your problem. Not the
```

รูปที่ 20 ตัวอย่างข้อมูลชุด SemEval-2017 Task4A Dataset

3) Sentiment Strength Twitter Dataset (SS-Tweet) ประกอบด้วยข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์ จำนวน 4,242 ข้อความ เป็นเอกสารประเภท text file (.txt) ประกอบด้วย 3 ฟิลด์ ได้แก่ 1) ค่าเฉลี่ยความคิดเห็นเชิงบวก (Mean Pos) 2) ค่าเฉลี่ยความคิดเห็นเชิงลบ (Mean Neg) และ 3) ข้อความคิดเห็น (Tweet) ตามลำดับ ในการระบุขั้วความคิดเห็นทำได้โดยนำค่าเฉลี่ยความคิดเห็นเชิงบวกหารกับค่าเฉลี่ยความคิดเห็นเชิงลบ หากผลหารมีค่าเท่ากับ 1 แสดงว่าเป็นข้อความคิดเห็นที่เป็นกลาง หากผลหารมีค่ามากกว่า 1.5 แสดงว่าเป็นข้อความคิดเห็นเชิงบวกและหากไม่ใช่ 2 กรณีข้างต้นแสดงว่าเป็นข้อความคิดเห็นเชิงลบ ผู้วิจัยทำการคำนวณขั้วความคิดเห็นด้วยวิธีการข้างต้น พบว่าข้อมูลที่รวบรวมประกอบด้วย ข้อความคิดเห็นเชิงบวก จำนวน 1,320 ข้อความ ข้อความคิดเห็นเชิงลบ จำนวน 2,910 ข้อความ และข้อความคิดเห็นที่เป็นกลาง จำนวน 12 ข้อความ จากการวิเคราะห์คุณลักษณะข้อมูล พบว่า ข้อมูลชุดนี้มีความยาวสูงสุด เท่ากับ 164 ตัวอักษร ความยาวสั้นสุด เท่ากับ 4 ตัวอักษร และมีความยาวเฉลี่ย เท่ากับ 97 ตัวอักษร ผู้วิจัยได้ทำการสุ่มข้อความคิดเห็นเพื่อใช้ในการทดลอง จำนวน 2,600 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 1,300 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 1,300 ข้อความ ตัวอย่างข้อมูลชุด Sentiment Strength Twitter Dataset แสดงดังรูปที่ 18

	mean pos	mean neg	Tweet
1	mean pos	mean neg	Tweet
2	3	2	?RT @justinbiebcr: The bigger the better....if you know what I mean ;)
3	3	1	Listening to the "New Age" station on @SlackerRadio ? http://slacker.com/r/nqKf
4	1	1	I favorited a YouTube video -- Drake and Josh - The Storm "We will rock you" http://youtu.be/
5	4	2	i didnt mean knee high I ment in lengt it goes down to my knees ^-^ and is so cute I love it!
6	2	1	I wana see the vid Kyan
7	1	3	if my mom went on for the love of ray J or any reality show i'd bee pissed .
8	1	1	Ok so I just got a deal for my own show "For The Love Of Deez Nuts" go to VH1/deezNuts to appl
9	3	1	@Mrhilton1985 Welcome to Twitter xx
10	2	1	@kjbmusic oh yeah... however, I'd still like to be in the midst of it all though... u know...
11	2	2	Can't say I like the new facebook layout. But just posted pics from my Super Bowl week. =)
12	2	1	I need a nice tea-drinking pic for our #Tea Club Membership page - anyone got one they'd be ha
13	3	1	@JonathanRKnight so twitpic it lol, I love Home Depot, love working w/my hands and building th
14	1	3	@BarCough it's enough to make you sick, eh? there's nothing sacred anymore
15	3	2	Hacienda is now level 80 time to get epic gear for her!!!! Oh and maybe some sleep would be go
16	1	1	DJ K-City presents "Crank of America" THE MIXTAPE. wanna be on it? DM me. http://twitpic.com/1
17	1	3	"Iran, with its unity and God's grace, will punch the arrogance (West) 22nd of Bahman (Feb 11)
18	4	3	@TiffyStarz wtf, where i come from noone likes metal and hardcore, like 5 of my mates max are
19	1	1	www.moneyhackers.org Reasons why entrepreneurs to venture in Nursing Agencies | Best ...: Reas
20	1	2	#4WordsOnObamasHand Don't Say The N-Word
21	1	3	City watchdog in chaos as chief executive Hector Sants resigns just months before general elec
22	2	1	RT @MangaUK: God it is a big news day today! By popular demand, "Ah! My Goddess Season 2" will
23	2	2	Wow my sis said she gona get a tramp stamp tat LOL I told her it's to late for that ha ha evil
24	3	2	@russmarshalek Sold! Would love to be your crazyass big sis -- how could I say no?! Cannot bel
25	2	1	@zzramesses yes; hope to release that feature next month
26	1	1	Phil Collins- You Can't Hurry Love
27	2	3	i need money! i need new car!!! jesus...somebody please buy my old car :DDD
28	2	3	In ny wif @DJWALLAH and the Heavy Hitter crew... shout to @freddyphh for looking real gay in t
29	3	1	RT @RockinGreenSoap: I Flip(in) Love @rockingreensoap! Follow them to win a free Flip Camera!
30	1	3	@ShiedhaO well damn! Renee still aint playin, is she?! and neither is Jack!! @LikasParody
31	1	2	@heydusti oh, geez, I'd have a lot more songs that way.
32	1	1	@natuhtack I will have some apple chicken sausage delivered to you.
33	1	2	@wendywave1 HAHAHAH that was worded weird. I'm reading while a candle is burning in my room
34	1	1	@KantGitRite you should follow me while ur at it u tickle my dick with ur comments
35	2	2	RT @importantdate: You can preview the exclusive Tim Burton's #Alice Boutique, opening 2/11. I
36	2	2	@iJuslisen pretty much my g..i need a banger..get it to the bloggers..lets do it.
37	2	2	Photo: Chace Crawford?. Oh man I can look at pictures of him all day BUT i have homework to ge
38	1	1	Commission Ritual.: P.S. I want to make this perfectly clear: Commission Ritual will get your
39	1	2	On Insurance and Hospitals: I wonder if there is a connection between the hospital system over
40	1	2	Chilling textin (?) ... To much info..foreal you know its los ward yung squad bitch...

รูปที่ 21 *ตัวอย่างข้อมูลชุด* Sentiment Strength Twitter Dataset

4) Health Care Reform (HCR) ประกอบด้วยข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์ เมื่อเดือนมีนาคม ปี ค.ศ. 2010 จำนวน 2,516 ข้อความ โดยผู้รวบรวมคัดเลือกจากข้อความทวิตเตอร์ที่มีแฮชแท็ก #hcr รูปแบบของข้อมูลเป็นไฟล์เอกสารประเภท Comma Separated Value (CSV) ประกอบด้วย 6 ฟิลด์ ได้แก่ 1) ลำดับ (Tweet id) 2) รหัสผู้ใช้งาน (User id) 3) ชื่อผู้ใช้งาน (Username) 4) ข้อความคิดเห็น (Content) 5) ขั้วความคิดเห็น (Sentiment) 6) ประเด็นที่โพสต์ (Target) 7) รหัสผู้อธิบาย Annotator id 8) ความคิดเห็นเพิ่มเติม (Comment) 9) ข้อโต้แย้ง (Dispute) ชุดข้อมูลนี้ประกอบด้วย 5 คลาส ได้แก่ ความคิดเห็นเชิงบวก (Positive) ความคิดเห็นเชิงลบ (Negative) ความคิดเห็นที่เป็นกลาง (Neutral) ข้อความที่ไม่ตรงประเด็น (Irrelevant) และข้อความอื่น ๆ (Other) จากการวิเคราะห์คุณลักษณะข้อมูล พบว่า ข้อมูลชุดนี้มีความยาวสูงสุดเท่ากับ 142 ตัวอักษร ความยาวสั้นสุด เท่ากับ 23 ตัวอักษร และมีความยาวเฉลี่ย เท่ากับ 113 ตัวอักษร ผู้วิจัยทำการเลือกข้อมูลจำนวน 2 ฟิลด์ ได้แก่ ข้อความคิดเห็นและขั้วความคิดเห็น และทำการสุ่มข้อความคิดเห็นเพื่อการทดลอง จำนวน 1,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 500 ข้อความ และ ข้อความคิดเห็นเชิงลบ จำนวน 500 ข้อความ ตัวอย่างข้อมูลชุด Health Care Reform แสดงดังรูปที่ 22

	A	B	C	D	E	F	G	H	I
1	tweet id	user id	username	content	sentimen	target	annotator	comment	dispute
2	10237553563	69128478		RT @angelsmomaw: #HCR is unwanted because it w	negative	hcr	aluckhardt		
3	10239984258	7713202	GOPLeade	RT @WMRepublicans President's Remarks Yesterday	negative	hcr	aluckhardt		
4	10240791063	34927577	cnsnews_	RT @johnboehner: Pelosi on #HCR: "We have to pas	negative	dems	aluckhardt		
5	10253203734	16930489	ExJon	RT @vermontaigne Cancer patient died after 58,000	negative	hcr	aluckhard	Factual, but wait time	
6	10255459398	15350894	LJSearles	RT @HealthReformNow: As the President said to Re	negative	gop	aluckhard	Also positive for hcr.	
7	10255930511	21280367	kiliffe	#tcot oppose #hcr because they fear that 'brown pec	negative	conservat	aluckhardt		
8	10268857013	45648676	Rainbow8	RT @thebighoot Barack Obama delivers remarks at S	negative	obama	aluckhardt		
9	10284982547	21964694	EzKool	Grandma is Safe with this Health Care Bill - http://bi	negative	conservat	aluckhard	Because "death pane	
10	10286758830	43977798	Liam_Fox	Ten Wrong Reasons to Oppose Health Reform -- Poli	negative	conservat	aluckhard	Not necessarily positi	
11	10287678432	71941447	Dudeman	Make no mistake, the #hcr legislation is nothing but	negative	hcr	aluckhardt		
12	10287773768	18464266	peterdaou	By what logic does pushing a #hcr bill and dropping i	negative	obama	aluckhard	Could be dems instea	
13	10294219761	35039919	Pkatt	if people werent so brainwashed we could find a sol	negative	other	aluckhard	I can't tell if the "brai	
14	10296807551	22643800	hipEchik	Ronald Reagan On Socialized Medicine 1961 : http://	negative	hcr	aluckhard	"Socialized" carried n	
15	9946829766	17388022		RW so pissed over #hcr they think ppl will vote again	negative	gop	AMahmud		
16	9898405633	17971455	toddeher	President announces he doesn't give a damn what A	negative	obama	AMahmud		
17	9915233477	91824727	BlockTax	RT @MikeBlockCPA: Can They Make Obamacare Wor	negative	hcr	AMahmud		
18	9941134669	16887628	mattklew	He repeated discredited talking points with a straigh	negative	obama	AMahmud		
19	9941460369	60994766	StopBlue	McConnell: "What we know about the health care bi	negative	hcr	AMahmud	negative toward heal	
20	9945515119	19326476	BrettR476	RT @Marnus3: Maybe Obama could trick some reput	negative	hcr	AMahmud	Sarcastic?	
21	9951361736	63065243	OsborneIr	#GOPcodered They know America needs #HCR, they	negative	gop	AMahmud		
22	9971475686	17809507	AP_Mobil	Stupak: 12 Dems ready to oppose health care bill htt	negative	hcr	AMahmud		
23	9985323762	18086787	RebelCapi	Before he was President, he was for #singlepayer &	negative	obama	AMahmud		

รูปที่ 22 ตัวอย่างข้อมูลชุด Health Care Reform

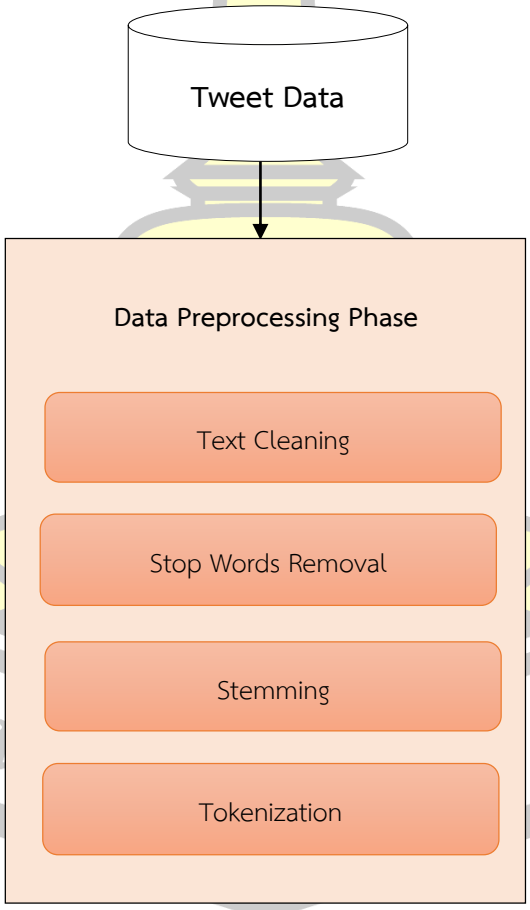
5) Sanders Twitter Dataset ประกอบด้วยข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์ จำนวน 5,512 เอกสาร แบ่งเป็น 4 หัวข้อ ได้แก่ Apple, Google, Microsoft และ Twitter รูปแบบของข้อมูลเป็นไฟล์เอกสารประเภท Comma Separated Value (CSV) ประกอบด้วย 2 ฟิลด์ ได้แก่ ขั้วความคิดเห็น (Sentiment) และ ข้อความคิดเห็น (TweetText) ประกอบด้วย 4 คลาส ได้แก่ ความคิดเห็นเชิงบวก (Positive) ความคิดเห็นเชิงลบ (Negative) ความคิดเห็นที่เป็นกลาง (Neutral) และ ความคิดเห็นอื่น ๆ (Irrelevant) ข้อมูลชุดนี้ประกอบด้วยข้อความคิดเห็นเชิงลบ จำนวน 654 ข้อความ ข้อความคิดเห็นที่เป็นกลาง จำนวน 2,503 ข้อความ และข้อความคิดเห็นเชิงบวก จำนวน 570 ข้อความ และ ข้อความคิดเห็นอื่น ๆ จำนวน 1,786 ข้อความ จากการวิเคราะห์คุณลักษณะข้อมูล พบว่า ข้อมูลชุดนี้มีความยาวสูงสุด เท่ากับ 148 ตัวอักษร ความยาวสั้นสุด เท่ากับ 9 ตัวอักษร และมีความยาวเฉลี่ย เท่ากับ 97 ตัวอักษร ผู้วิจัยได้ทำการสุ่มข้อความคิดเห็นเพื่อใช้ในการทดลอง จำนวน 1,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 500 ข้อความ ข้อความคิดเห็นเชิงลบ จำนวน 500 ข้อความ ตัวอย่างข้อมูลชุด Sanders Twitter Dataset แสดงดังรูปที่ 23

	A	B	C	D	E	F	G	H	I	J	K
1	Sentimen	TweetText									
2	positive	Now all @Apple has to do is get swype on the iphone and it will be crack. Iphone that is									
3	positive	@Apple will be adding more carrier support to the iPhone 4S (just announced)									
4	positive	Hilarious @youtube video - guy does a duet with @apple 's Siri. Pretty much sums up the love affair! http									
5	positive	@RIM you made it too easy for me to switch to @Apple iPhone. See ya!									
6	positive	I just realized that the reason I got into twitter was ios5 thanks @apple									
7	positive	I'm a current @Blackberry user, little bit disappointed with it! Should I move to @Android or @Apple @i									
8	positive	The 16 strangest things Siri has said so far. I am SOOO glad that @Apple gave Siri a sense of humor! http:/									
9	positive	Great up close & personal event @Apple tonight in Regent St store!									
10	positive	From which companies do you experience the best customer service aside from @zappos and @apple?									
11	positive	Just apply for a job at @Apple, hope they call me lol									
12	positive	RT @JamaicanIdler: Lmao I think @apple is onto something magical! I am DYING!!! haha. Siri suggested w									
13	positive	Lmao I think @apple is onto something magical! I am DYING!!! haha. Siri suggested where to find whores									
14	positive	RT @PhillipRowntree: Just registered as an @apple developer... Here's hoping I can actually do it... Any h									
15	positive	Wow. Great deals on refurbed #iPad (first gen) models. RT: Apple offers great deals on refurbished 1st-g									
16	positive	Just registered as an @apple developer... Here's hoping I can actually do it... Any help, greatly appreciate									
17	positive	ä½ å¥½ ! Currently learning Mandarin for my upcoming trip to Hong Kong. I gotta hand it to @Apple iPhor									
18	positive	Come to the dark side ðŸ“±â€œ@gretcheneclark: Hey @apple, if you send me a free iPhone, I will public									
19	positive	Hey @apple, if you send me a free iPhone (any version will do), I will publicly and ceremoniously burn m									
20	positive	Thank you @apple for Find My Mac - just located and wiped my stolen Air. #smallvictory #thievingbastarc									
21	positive	Thanks to @Apple Covent Garden #GeniusBar for replacing my MacBook keyboard/cracked wristpad duri									
22	positive	@DailyDealChat @apple Thanks!!									
23	positive	iPads Replace Bound Playbooks on Some N.F.L. Teams http://t.co/2UXAWKwf @apple @nytimes									

รูปที่ 23 ตัวอย่างข้อมูลชุด Sanders Twitter Dataset

3.2 การเตรียมข้อมูล (Data Preprocessing)

ข้อมูลที่ใช้ในการวิจัยเป็นข้อความที่อยู่บนเว็บไซต์เครือข่ายสังคมออนไลน์ ประกอบด้วยแฮชแท็ก (Hash Tag: #) ที่ใช้อธิบายสถานะที่ต้องการเน้นประเด็นหรือความรู้สึกพิเศษต่างๆ และมีข้อความแสดงอารมณ์ (Emoticons) ที่หลากหลาย บางข้อความมีที่อยู่เว็บไซต์ (URL) แทรกอยู่ด้วย นอกจากนี้ยังมีการแท็กชื่อบุคคลโดยข้อความที่แท็กชื่อบุคคลคือข้อความที่ขึ้นต้นด้วยเครื่องหมายแอด (@) ข้อความคิดเห็นที่อยู่บนเว็บไซต์ทวิตเตอร์เป็นข้อความสั้นๆ ไม่มีโครงสร้างที่แน่นอน มีคำศัพท์สแลง อักขรพิเศษ และคำสะกดผิดค่อนข้างมาก การเตรียมข้อมูลเป็นกระบวนการสำคัญกระบวนการหนึ่งที่จะช่วยลดจำนวนคุณลักษณะเพื่อเพิ่มประสิทธิภาพและลดระยะเวลาในการจำแนกความคิดเห็น ประกอบด้วย 4 ขั้นตอน ได้แก่ การทำความสะอาดข้อความ การกำจัดคำหยุด การหารากคำศัพท์ และการตัดคำ ดังแสดงในรูปที่ 24



รูปที่ 24 ขั้นตอนการเตรียมข้อมูล

3.2.1 การทำความสะอาดข้อความ (Text Cleaning)

เนื่องจากข้อความบนเครือข่ายสังคมออนไลน์เป็นข้อความที่ไม่มีโครงสร้างที่แน่นอน ประกอบด้วยคำและอักขระพิเศษที่ไม่มีผลต่อความคิดเห็นเป็นจำนวนมาก เช่น ชื่อผู้ใช้งาน ที่อยู่เว็บไซต์ การจะนำข้อมูลไปจำแนกให้มีประสิทธิภาพจะต้องดำเนินการทำความสะอาด โดยตรวจสอบและลบข้อความที่ได้ทำการวิเคราะห์แล้วว่าไม่มีผลต่อการจำแนกความคิดเห็น รวมทั้งดำเนินการตรวจสอบและแก้ไขคำที่สะกดผิด ขั้นตอนการทำความสะอาดข้อความ แสดงดังรูปที่ 25 ประกอบด้วย 6 ขั้นตอน คือ

1) การแปลงข้อความทั้งหมดให้เป็นตัวอักษรพิมพ์เล็ก (Lower Case Tokens) เช่น Http จะถูกแก้ไขเป็น http เป็นต้น

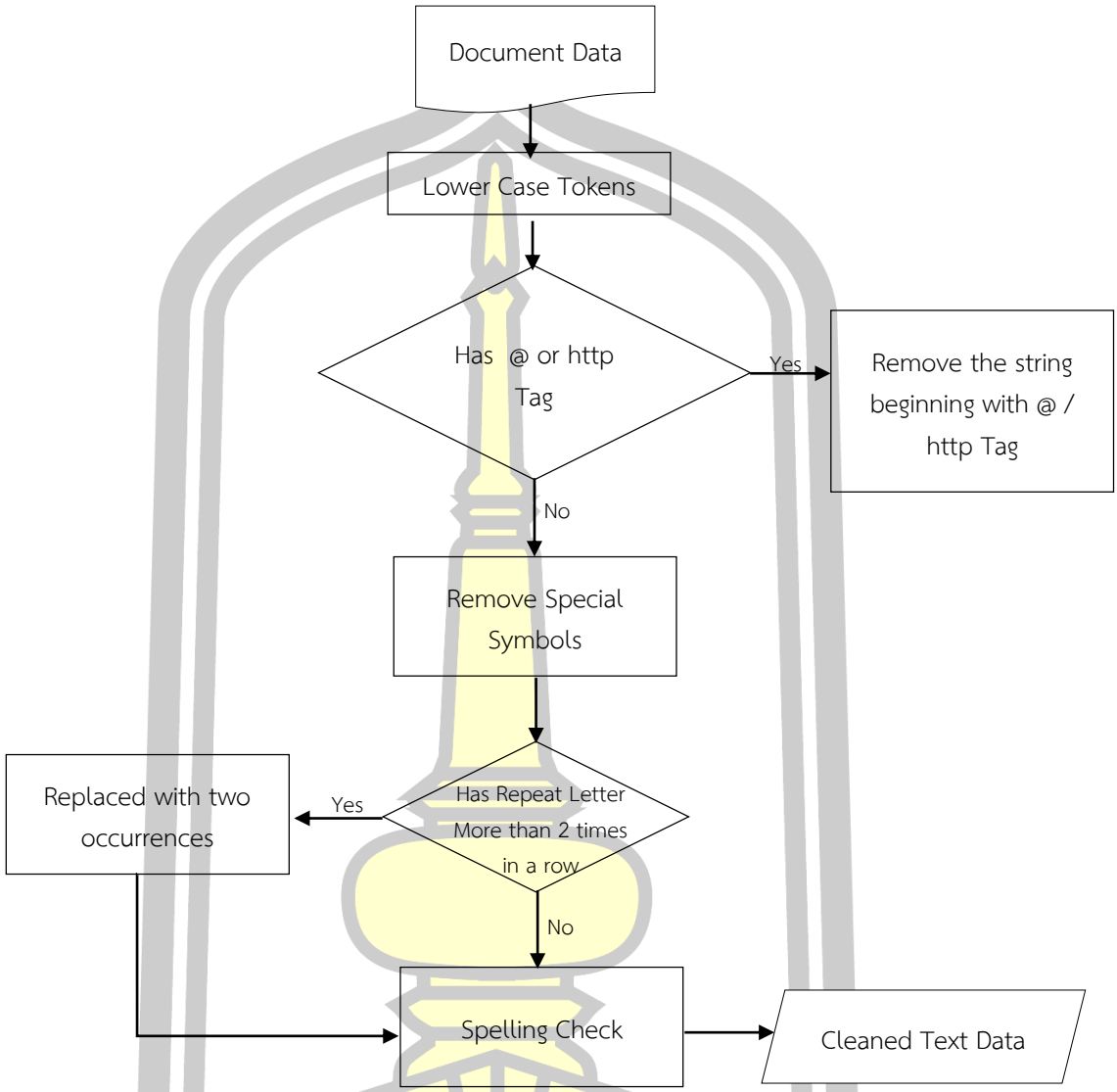
2) ลบข้อความที่ขึ้นต้นด้วย “@” ออก เนื่องจากตามหลักมาตรฐาน ข้อความที่ขึ้นต้นด้วย “@” เป็นชื่อผู้ใช้งาน ที่ผู้โพสต์อาจจะต้องการแท็กข้อความไปหาบุคคลอื่น และชื่อบุคคลไม่มีผลต่อการจำแนกความคิดเห็น

3) ลบข้อความที่ขึ้นต้นด้วย “http” ซึ่งเป็นที่อยู่เว็บไซต์ (URL) เช่น http://youtube.com/ceve5 เนื่องจากที่อยู่เว็บไซต์ไม่มีผลต่อการจำแนกความคิดเห็น

4) ลบสัญลักษณ์พิเศษและตัวเลขออกจากข้อความ ตัวอย่างอักขระพิเศษ เช่น “#” , “$” , “&” เป็นต้น

5) ตรวจสอบตัวอักษรที่พิมพ์ซ้ำ ๆ ซึ่งพบมากบนเครือข่ายสังคมออนไลน์ เช่น คำว่า “love” ผู้โพสต์อาจจะพิมพ์คำว่า looove, looooove, loooooooove ซึ่งหากมีตัวอักษรที่อยู่ติดกันและถูกพิมพ์ซ้ำกันมากกว่า 2 ตัวอักษร จะถูกแทนที่ด้วยตัวอักษรนั้นเพียง 2 ตัวอักษร เช่น ตัวอย่างข้างต้น จะถูกแปลงเป็นคำว่า loove

6) ตรวจสอบและแก้ไขคำที่สะกดผิด ส่วนมากใช้หลักการตรวจสอบกับพจนานุกรม คำผิดที่พบบ่อย โดยนำคำที่โพสต์ไปตรวจสอบกับพจนานุกรม ถ้าพบว่าเป็นคำผิดจะทำการตรวจสอบให้ถูกต้องตามพจนานุกรม เช่น คำว่า Like ผู้โพสต์พิมพ์คำว่า Lik เมื่อผ่านกระบวนการตรวจสอบและแก้ไขคำที่สะกดผิด คำว่า Lik จะถูกแทนด้วยคำที่ถูกต้อง คือ Like เป็นต้น

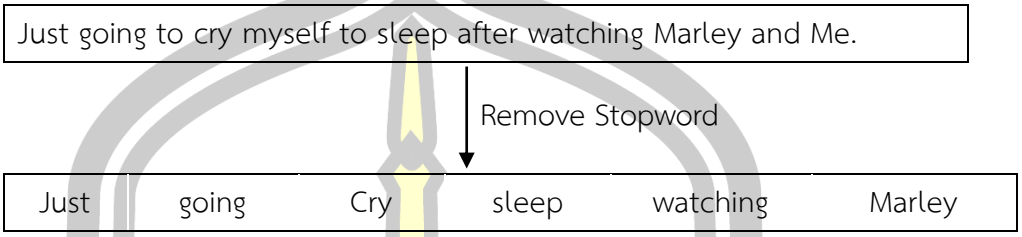


รูปที่ 25 ขั้นตอนการทำความสะอาดข้อความ

3.2.2 การกำจัดคำหยุด (Stop Word Removal)

คำหยุดเป็นคำที่ไม่สื่อความหมาย เป็นข้อความขยะที่ปรากฏในเอกสารค่อนข้างสูง และหากไม่ลบออกจะส่งผลให้ผลการทดลองเกิดความผิดพลาด คำหยุดที่พบบ่อยในข้อความภาษาอังกฤษ เช่น “a” , “an” , “the” “you” , “and” , “I” , “of” , “with” เป็นต้น งานวิจัยนี้ผู้วิจัยใช้คลังคำหยุดของ Standford Stopword List ซึ่งอยู่ในรูปแบบเอกสารข้อความ (Text File) ประกอบด้วยคำที่ไม่มีนัยสำคัญจำนวน 257 คำ มาตรวจสอบกับเอกสารที่ผ่านกระบวนการทำความสะอาด หากพบคำ

หยุดในเอกสารข้อความจะดำเนินการลบคำนั้นออกจากเอกสาร ตัวอย่างการกำจัดคำหยุด แสดงดังรูปที่ 26



รูปที่ 26 ตัวอย่างการกำจัดคำหยุด

3.2.3 การหารากคำศัพท์ (Stemming)

ในการสื่อสารเพื่อสื่อความหมายเดียวกัน อาจจะใช้คำศัพท์ได้หลากหลาย ขึ้นอยู่กับบริบทและหน้าที่ของคำ แต่คำต่าง ๆ เหล่านี้ส่วนมากจะมาจากรากคำศัพท์เดียวกัน เพื่อเป็นการลดจำนวนคำที่จะนำไปใช้เป็นตัวแทนของคำคุณลักษณะในการจำแนกความคิดเห็นให้มีประสิทธิภาพดีขึ้น จึงต้องมีกระบวนการหารากคำศัพท์เพื่อจัดกลุ่มคำที่มีความหมายเหมือนกัน ให้เป็นคำเดียวกัน เช่น คำว่า Stems, Stemmer, Stemming, Stemmed มาจากรากคำศัพท์เดียวกัน คือ คำว่า “Stem” งานวิจัยนี้ผู้วิจัยใช้วิธีการหารากคำศัพท์ของ Porter Stemmer ซึ่งประกอบด้วย 5 ขั้นตอน และ 5 กฎ [26] ได้แก่

ขั้นตอนที่ 1) แก้ไขคำที่ลงท้ายด้วยอักษรตามกฎข้อที่ 1 ดังต่อไปนี้

sses	->	ss;
ies	->	i;
ss	->	ss;
s	->	ϕ ;

ขั้นตอนที่ 2) แก้ไขคำที่ลงท้ายด้วยอักษรตามกฎข้อที่ 2 ดังต่อไปนี้

ational	->	ate;
tional	->	tion;
enci	->	ence;
anci	->	ance ;
izer	->	ize;
abli	->	able;
entli	->	ent;

eli -> e;

ousli -> ous;

ization -> ize;

ation -> ate;

ator -> ate;

alism -> al;

iveness -> ive;

fullness -> ful;

ousness -> ous;

aliti -> al;

iviti -> ive;

biliti -> ble;

ขั้นตอนที่ 3) แก้ไขคำที่ลงท้ายด้วยอักษรตามกฎข้อที่ 3 ดังต่อไปนี้

Icate -> ic;

ative -> ϕ ;

alize -> al;

iciti -> ic;

ical -> ic;

ful -> ϕ ;

ness -> ϕ ;

ขั้นตอนที่ 4) แก้ไขคำที่ลงท้ายด้วยอักษรตามกฎข้อที่ 4 ดังต่อไปนี้

al -> ϕ ;

ance -> ϕ ;

ence -> ϕ ;

er -> ϕ ;

ic -> ϕ ;

ing -> ϕ ;

able -> ϕ ;

ible -> ϕ ;

ant -> ϕ ;

ement -> ϕ ;

ment -> ϕ ;

ent -> ϕ ;

ou -> ϕ ;

ism -> ϕ ;

ate -> ϕ ;

iti -> ϕ ;

ous -> ϕ ;

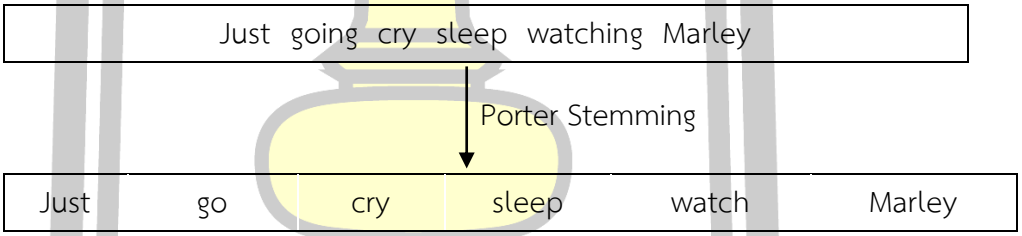
ive -> ϕ ;

ize -> ϕ ;

ขั้นตอนที่ 5) แก้ไขคำที่ลงท้ายด้วยอักษรตามกฎข้อที่ 5 ดังต่อไปนี้

e -> ϕ ;

ตัวอย่างผลของการหารากคำศัพท์ด้วยวิธีการของ Porter Stemmer แสดงดังรูปที่ 27

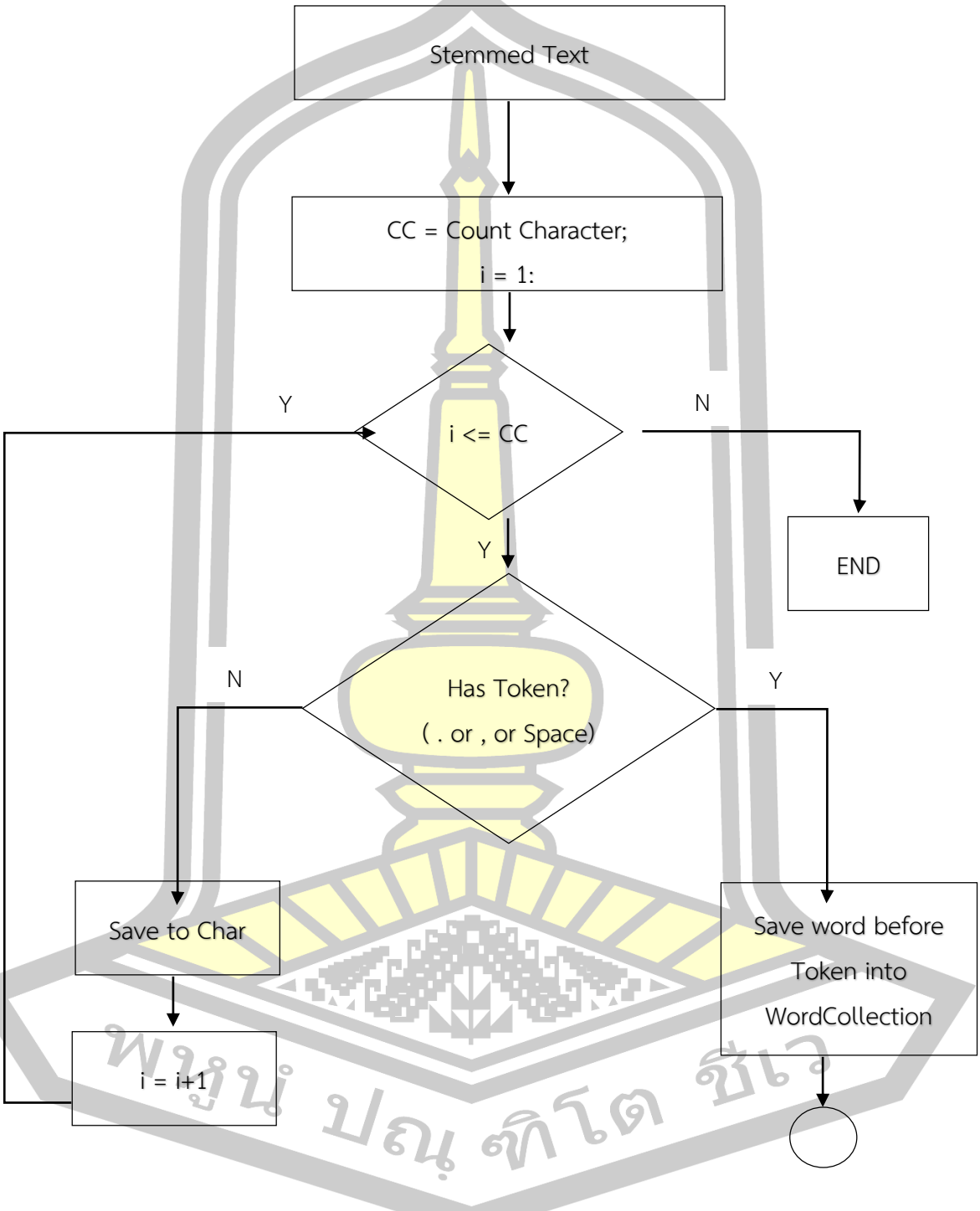


รูปที่ 27 ตัวอย่างการหารากคำศัพท์ด้วยวิธีการ Porter Stemmer

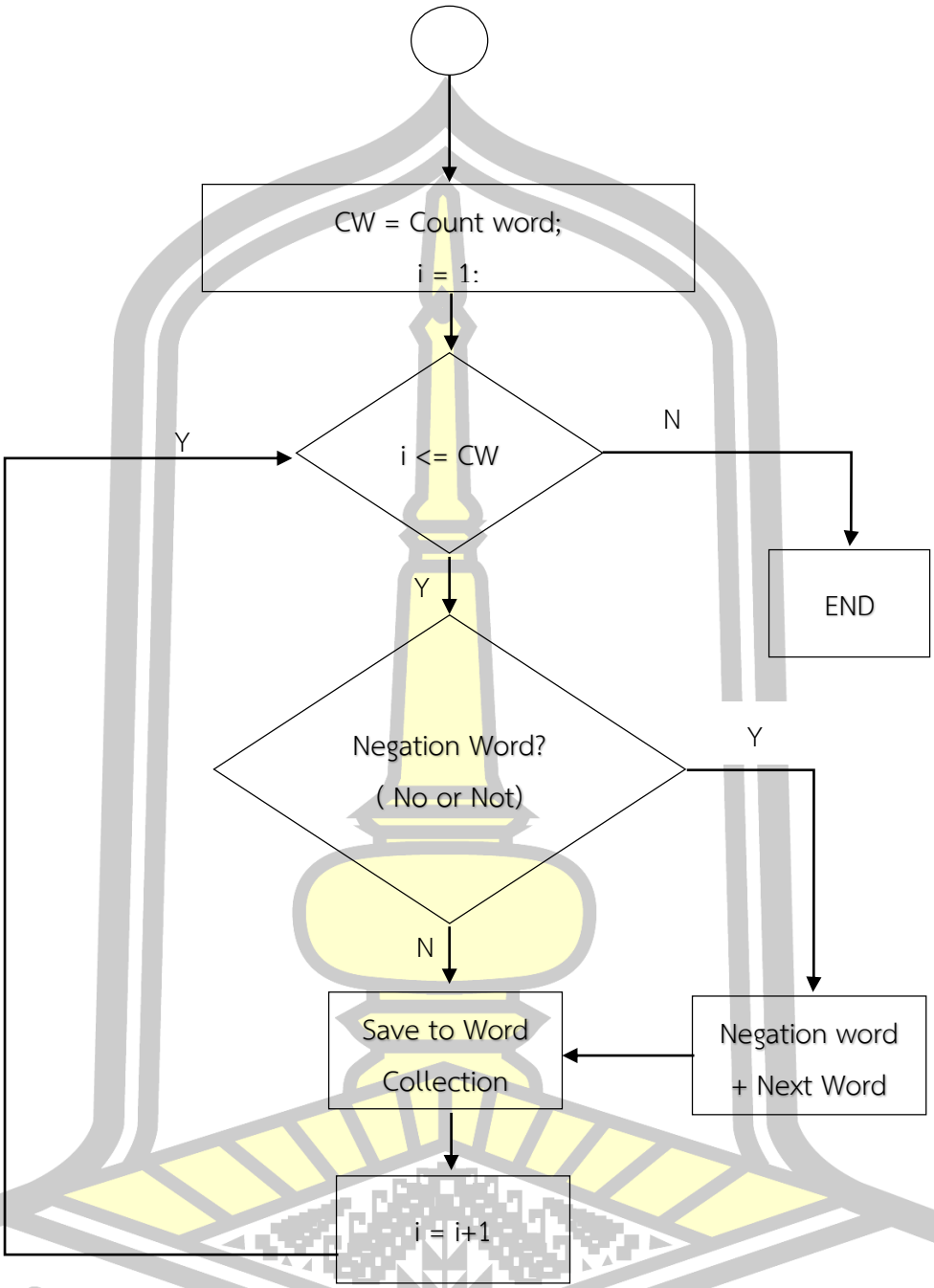
3.2.4 การตัดคำ

กระบวนการตัดคำ เป็นการนำเอกสารข้อความมาแบ่งเป็นคำ ผู้วิจัยใช้หลักการแบ่งข้อความด้วยวิธีการ Unigrams ร่วมกับ Bigrams โดยข้อความที่มีคำปฏิเสธอยู่ก่อนหน้า เช่น no, not จะใช้หลักการแบ่งข้อความด้วย Bigrams ส่วนข้อความอื่น ๆ จะใช้หลักการแบ่งข้อความด้วย Unigrams เนื่องจากคำที่มีคำปฏิเสธอยู่ก่อนหน้าจะทำให้ความหมายของคำนั้นเปลี่ยนไป เช่น คำว่า Good หมายถึง ดี เมื่อมีคำว่า No อยู่ก่อนหน้า คือ No Good ความหมายจะเปลี่ยนเป็นไม่ดี การตัดคำวิธีนี้เป็นการลดจำนวนคุณลักษณะเบื้องต้น และลดข้อผิดพลาดในการแปลความหมายของคุณลักษณะได้ ซึ่งจากงานวิจัยของ Go และคณะ [11] แสดงให้เห็นว่าวิธีการ Unigrams ร่วมกับ Bigrams มี

ประสิทธิภาพสูงที่สุด ในการจำแนกความคิดเห็นที่อยู่บนเว็บไซต์ทวิตเตอร์ (Twitter Data) ขั้นตอนการตัดคำแสดงดังรูปที่ 28 และ รูปที่ 29



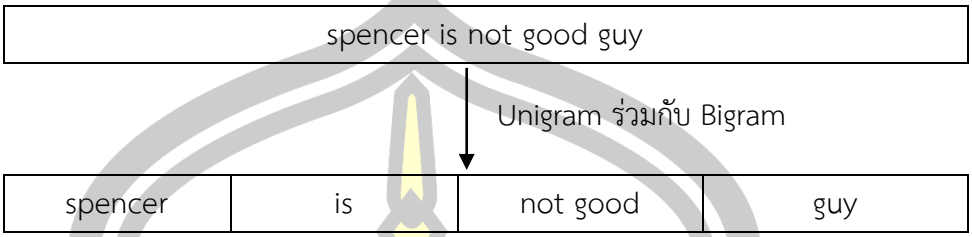
รูปที่ 28 ขั้นตอนการตัดคำ (1)



รูปที่ 29 ขั้นตอนการตัดคำ (2)

ตัวอย่างการตัดคำด้วยวิธีการ Unigram ร่วมกับ Bigram แสดงดังรูปที่ 30

.



รูปที่ 30 *ตัวอย่างการตัดคำด้วยวิธีการ* Unigram *ร่วมกับ* Bigram

3.4 การแทนค่าในเอกสาร (Document Representation)

เอกสารที่ได้หลังจากผ่านกระบวนการเตรียมข้อมูล จะแสดงในรูปแบบของเวกเตอร์ (Vector Model) โดย 1 เอกสาร จะแทนด้วย 1 เวกเตอร์ แต่ละเวกเตอร์ประกอบด้วยคุณลักษณะ ซึ่งคุณลักษณะได้จากการตัดคำด้วยวิธีการ Unigrams ร่วมกับ Bigrams งานวิจัยนี้ใช้วิธีการแทนค่าในเอกสารโดยใช้ถุงคำ (Bag of Word) และแทนค่าน้ำหนักด้วยวิธีการ Boolean Weighting ตัวอย่างการแทนค่าในเอกสารแสดงดังตาราง 10

ตาราง 10 *ตัวอย่างชุดข้อมูลนำเข้า*

Document id	Text	Class
d_1	spencer is not a good guy.	c_2
d_2	It was the best of times.	c_1
d_3	It was the worst of times.	c_2
d_4	She is a good girl.	c_1
d_5	It is a good times.	c_1

จากตาราง แสดงตัวอย่างข้อมูลนำเข้า ซึ่งอยู่ในรูปแบบข้อความ แต่ละแถวแทนด้วย 1 เอกสาร เมื่อนำมาเข้ากระบวนการเตรียมข้อมูลแล้วจะได้คุณลักษณะ คือ คำที่อยู่ในเอกสาร ซึ่งจากตารางข้างต้น ประกอบด้วย 4 เอกสาร และ คุณลักษณะแสดงดังตาราง 11

ตาราง 11 ตัวอย่างคุณลักษณะที่ได้จากการคัดเลือก

Feature_ID	Features (Word in Documents)
f_1	spencer
f_2	not good
f_3	guy
f_4	best
f_5	time
f_6	worst
f_7	good
f_8	girl

ในขั้นตอนต่อไป เป็นการแทนค่าการเกิดคุณลักษณะในเอกสาร โดยหากพบคุณลักษณะในเอกสาร จะให้มีค่า เท่ากับ 1 หาก ไม่พบให้มีค่า เท่ากับ 0 ขนาดเวกเตอร์จะมีขนาดเท่ากับ ขนาดเอกสาร (Size of Document) x ขนาดของคุณลักษณะ (Size of Feature) ซึ่งจากตัวอย่างข้างต้น ขนาดของเวกเตอร์ที่ได้ คือ ขนาด 5x8 ดังแสดงในตาราง 12

ตาราง 12 ตัวอย่างการแทนค่าในเอกสาร

Feature_ID / Document_ID	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
d_1	1	1	1	0	0	0	0	0
d_2	0	0	0	1	1	0	0	0
d_3	0	0	0	0	1	1	0	0
d_4	0	0	0	0	0	0	1	1
d_5	0	0	0	0	1	0	1	0

3.5 การคัดเลือกคุณลักษณะ (Feature Selection)

เมื่อผ่านกระบวนการแทนค่าในเอกสาร จะได้เวกเตอร์ซึ่งเป็นรูปแบบเวกเตอร์แนวนอน (Horizontal Vector) ผู้วิจัยได้ทำการหาค่าน้ำหนักของแต่ละคุณลักษณะเพื่อจัดลำดับความสำคัญของคุณลักษณะ โดยนำเวกเตอร์ที่ได้จากกระบวนการการแทนค่าในเอกสารมาหาค่าน้ำหนักของ

คุณลักษณะที่เกิดขึ้นในแต่ละคลาสและเลือกค่าสูงสุดมาเป็นค่าน้ำหนักของคุณลักษณะนั้น โดยกำหนดให้ เซตของเอกสาร $D=\{d_1,d_2,..,d_n\}$ เมื่อ n = จำนวนเอกสารทั้งหมด เซตของคุณลักษณะ $T=\{t_1,t_2,..,t_m\}$ เมื่อ m = จำนวนคุณลักษณะทั้งหมด และเซตของคลาส $C=\{c_1,c_2,..,c_k\}$ เมื่อ k = จำนวนคลาสทั้งหมด แต่ละเอกสาร $d\in D$ จะนำเสนอในรูปแบบ $(t_1,t_2,..t_k,c_i\}$ เมื่อ $t\in T$ และ $c\in C$ แสดงดังตาราง 13 ตัวอย่างข้อมูลที่อยู่ในรูปแบบเวกเตอร์แนวนอน แสดงดังตาราง 14

ตาราง 13 รูปแบบเวกเตอร์แนวนอน

Document id	t_1	t_2	...	t_m	Class
d_1	w_{11}	w_{12}	...	w_{1m}	c_1
d_2	w_{21}	w_{22}	...	w_{2m}	c_1
d_3	w_{31}	w_{32}	...	w_{3m}	c_2
...	...	...	...	...	...
d_n	w_{n1}	w_{n2}		w_{nm}	c_k

ตาราง 14 ตัวอย่างข้อมูลที่อยู่ในรูปแบบเวกเตอร์แนวนอน

Feature (F) / Document_ID	t_1	t_2	t_3	t_4	t_5	t_6	Class
d_1	1	1	1	1	1	0	c_1
d_2	0	1	0	1	0	1	c_1
d_3	1	1	1	1	0	1	c_1
d_4	0	0	1	0	0	1	c_2
d_5	1	0	0	1	1	1	c_2

ตาราง 15 *รูปแบบข้อมูลแนวตั้ง* (Vertical Data)

Transaction	Set of Documents
t_1	$\{d_1, d_3, d_5\}$
t_2	$\{d_1, d_2, d_3\}$
t_3	$\{d_1, d_3, d_4\}$
t_4	$\{d_1, d_2, d_3, d_5\}$
t_5	$\{d_1, d_5\}$
t_6	$\{d_2, d_3, d_4, d_5\}$
c_1	$\{d_1, d_2, d_3\}$
c_2	$\{d_4, d_5\}$

นิยามที่ 3.1 กำหนดให้ $S(t_i)$ คือ เซตของเอกสารที่มีคุณลักษณะ t_i

ตัวอย่างที่ 3.1 จากตาราง 15 พบว่า คุณลักษณะ t_3 ปรากฏในเอกสาร d_1, d_3 และ d_4 ดังนั้น $S(t_3) = \{d_1, d_3, d_4\}$

นิยามที่ 3.2 ค่าสนับสนุน (Support) ของคุณลักษณะ t_i หมายถึง จำนวนเอกสารที่มีคุณลักษณะ t_i แทนด้วย $|S(t_i)|$

ตัวอย่างที่ 3.2 จากตาราง 14 พบว่าคุณลักษณะ t_3 ในเอกสารทั้งหมด จำนวน 3 เอกสาร ซึ่งปรากฏใน d_1, d_3 และ d_4 ดังนั้น $|S(t_3)|$ มีค่าเท่ากับ 3

นิยามที่ 3.3 ค่าสนับสนุน (Support) ของคุณลักษณะ t_i ในคลาส c_j คือ จำนวนเอกสารที่มีคุณลักษณะ t_i และอยู่ในคลาส c_j แทนด้วย $|S(t_i, c_j)|$

โดยที่ $|S(t_i, c_j)|$ สามารถคำนวณได้ง่ายโดยการใช้สมการ (3.1) ดังนี้

$$|S(t_i, c_j)| = |S(t_i) \cap S(c_j)| \qquad (3.1)$$

ตัวอย่างที่ 3.3 จากตาราง 15 ซึ่งเป็นรูปแบบข้อมูลแนวตั้ง พบว่า $|S(t_3,c_1)|$ = $|S(t_3) \cap S(c_1)|$ = $|\{d_1,d_3,d_4\} \cap \{d_1,d_2,d_3\}|$ =$|\{d_1,d_3\}|$ = 2 ซึ่งเมื่อดูจากตาราง 14 ซึ่งเป็นเวกเตอร์แนวนอน พบว่า จำนวนของคุณลักษณะ t_3 ในคลาส c_1 มีจำนวนเท่ากับ 2 เอกสาร คือ d_1 และ d_3 ซึ่งมีค่าเท่ากัน

นิยามที่ 3.4 ค่าความเชื่อมั่น (Confident) ของการเกิดคุณลักษณะ t_i ในคลาส c_j คือ ค่าที่แสดงให้เห็นถึงโอกาสการเกิด t_i ในคลาส c_j ค่าความเชื่อมั่น คำนวณได้จากสมการ (3.2)

$$C(t_i,c_j) = \frac{S(t_i,c_j)}{S(t_i)} \qquad (3.2)$$

ตัวอย่างที่ 3.4 จากตาราง 14 ค่าความเชื่อมั่นของคุณลักษณะ t_3 ในคลาส c_1 หรือ $C(t_3,c_1)$ สามารถคำนวณได้จาก $S(t_3,c_1)/S(t_3)$ = 2/3 = 0.667 ซึ่งแสดงให้เห็นว่าเมื่อมีคุณลักษณะ t_3 ปรากฏในเอกสาร หมายถึง เอกสารนั้นมีโอกาสที่จะอยู่ในคลาส c_1 เท่ากับ 66.70% ส่วนค่าความเชื่อมั่นของคุณลักษณะ t_3 ในคลาส c_2 หรือ $C(t_3,c_2)$ สามารถคำนวณได้จาก $S(t_3,c_2)/S(t_3)$ = 1/3 = 0.333 ซึ่งแสดงให้เห็นว่าเมื่อมีคุณลักษณะ t_3 ปรากฏในเอกสาร หมายถึง เอกสารนั้นมีโอกาสที่จะอยู่ในคลาส c_2 เท่ากับ 33.30%

นิยามที่ 3.5 ค่าลำดับของคุณลักษณะ หมายถึง ลำดับของคุณลักษณะ t_i ที่อยู่ในคลาส c_j เมื่อเรียงตามค่าสนับสนุน (Support) จากมากไปหาน้อย คุณลักษณะ ในคลาส c_j มีค่าระหว่าง 0 ถึง $N-1$ เมื่อ N คือ จำนวนคุณลักษณะทั้งหมดที่ปรากฏในคลาส j แทนด้วย $R(t_i,c_j)$

ตัวอย่างที่ 3.5 จากตาราง 14 ค่าสนับสนุนของคุณลักษณะทั้งหมดที่อยู่ในคลาส c_1 มีค่าดังนี้

Feature (F)	t_1	t_2	t_3	t_4	t_5	t_6	Class
$\|S(t_i,c_1)\|$	2	3	2	3	1	2	c_1
$R(t_i,c_1)$	2	1	2	1	3	2	c_1

เมื่อพิจารณาคุณลักษณะตามค่าสนับสนุนจากมากไปน้อย ปรากฏว่าคุณลักษณะ t_1 ในคลาส c_1 มีค่าลำดับ เท่ากับ 2 หรือ $R(t_1,c_1)$ = 2 ซึ่งหมายถึง มีค่าสนับสนุนมากที่สุดเป็นอันดับ 2 ใน

คลาส c_1 กรณีที่ค่าสนับสนุนเท่ากัน จะให้ค่าลำดับคุณลักษณะที่ปรากฏก่อนไว้ลำดับแรก เช่น t_2 กับ t_4 มีค่าสนับสนุนเท่ากัน คือ 3 การจัดลำดับจะมีค่าเท่ากัน คือ 1

เนื่องจากค่าสนับสนุนของแต่ละคุณลักษณะมีค่าแตกต่างกันมากเกินไป จึงต้องทำการปรับค่าน้ำหนักของค่าสนับสนุนให้มีค่าระหว่าง 0 – 1 เพื่อให้ค่าสนับสนุนมีความสำคัญเท่ากับค่าความเชื่อมั่น ดังนิยามที่ 3.6

นิยามที่ 3.6 ค่าน้ำหนักของค่าสนับสนุน คือ ค่าที่ได้จากการจัดลำดับของค่าสนับสนุนของคุณลักษณะทั้งหมดที่อยู่ในคลาส c_j มีค่าอยู่ระหว่าง 0 ถึง 1 แทนด้วย $PS(t_i, c_j)$ ค่าความน่าจะเป็นของค่าสนับสนุนคำนวณได้จากสมการ (3.3)

$$PS(t_i, c_j) = \frac{R(t_i, c_j)}{N} \qquad (3.3)$$

เมื่อ N คือ จำนวนคุณลักษณะทั้งหมดที่อยู่ในคลาส j

ตัวอย่างที่ 3.6 จากตัวอย่าง 3.4 $R(t_1, c_1)$ = 1 สามารถปรับค่าน้ำหนักของค่าสนับสนุน ได้ดังนี้

$$PS(t_1, c_1) = \frac{R(t_1, c_1)}{N} = \frac{2}{6} = 0.33$$

นิยามที่ 3.7 ค่าน้ำหนักของคุณลักษณะ t_i ในคลาส c_j คือ ค่าที่แสดงให้เห็นถึงความสำคัญของคุณลักษณะ t_i ที่มีต่อคลาส c_j ค่าน้ำหนักของคุณลักษณะสามารถคำนวณได้จากสมการ (3.4)

$$w(t_i, c_j) = p \times PS(t_i, c_j) + (1-p) \times C(t_i, c_j) \qquad (3.4)$$

โดยที่ p คือ ค่าคงที่ มีค่ามากกว่าหรือเท่ากับ 0 และน้อยกว่าหรือเท่ากับ 1

$PS(t_i, c_j)$ คือ ค่าน้ำหนักของค่าสนับสนุน ของคุณลักษณะ t_i ในคลาส c_j

$C(t_i, c_j)$ คือ ค่าความเชื่อมั่นของคุณลักษณะ t_i ในคลาส c_j

ตัวอย่างที่ 3.7 จากตาราง 14 สมมติ p มีค่าเท่ากับ 0.9 ค่าน้ำหนักของคุณลักษณะ t_3 ในคลาส c_1 หรือ $w(t_3, c_1)$ คำนวณได้ดังนี้

$$\begin{aligned} w(t_3, c_1) &= p \times PS(t_3, c_1) + (1-p) \times C(t_3, c_1) \\ &= 0.9 \times 0.4 + (1-0.9) \times 0.667 \\ &= 0.36 + 0.667 \\ &= 0.427 \end{aligned}$$

ค่าน้ำหนักของคุณลักษณะ t_i จะพิจารณาจากค่าน้ำหนักของคุณลักษณะ t_i ที่อยู่ในแต่ละคลาส แล้วเอาค่าน้ำหนักที่มีค่ามากที่สุดมาเป็นค่าน้ำหนักของคุณลักษณะ t_i คำนวณได้ดังสมการ (3.5)

$$M(t_i) = \max(w(t_i, c_1), w(t_i, c_2), ..., w(t_i, c_j)) \quad (3.5)$$

เมื่อ $M(t_i)$ คือ ค่าน้ำหนักของคุณลักษณะ t_i

$w(t_i, c_1)$ คือ ค่าน้ำหนักของคุณลักษณะ t_i ที่อยู่ในคลาส c_1

$w(t_i, c_2)$ คือ ค่าน้ำหนักของคุณลักษณะ t_i ที่อยู่ในคลาส c_2

$w(t_i, c_j)$ คือ ค่าน้ำหนักของคุณลักษณะ t_i ที่อยู่ในคลาส c_j

ตัวอย่างที่ 3.8 ค่าน้ำหนักของคุณลักษณะ t_3 ที่อยู่ในคลาส c_1 เท่ากับ 0.427 และ สมมติค่าน้ำหนักของคุณลักษณะ t_3 ที่อยู่ในคลาส c_2 เท่ากับ 0.213 จะได้ $M(t_3)$ เท่ากับ 0.427

จากแนวคิดข้างต้น สามารถสรุปเป็นขั้นตอนการคัดเลือกคุณลักษณะ ได้ดังรูปที่ 31

Algorithm 1: Proposed Method

1. for each class $c_k \in C$
2. for each features $t_i \in T$ do:
3. $|S(t_i, c_k)| = |S(t_i) \cap S(c_k)|$
4. $Con(t_i, c_k) = \frac{|S(t_i, c_k)|}{|S(t_i)|}$
5. end for
6. $sort(T)$ in ascending order of support
7. for each features $t_i \in sort(T)$ do:
8. $PS(t_i, c_k) = \frac{R(t_i, c_k)}{|D|}$
9. $w(t_i, c_k) = p \times PS(t_i, c_k) + (1 - p) \times Con(t_i, c_k)$
10. $M(t_i) = \max(w(t_i, c_k))$
11. end for
12. end for

รูปที่ 31 แสดงขั้นตอนการคัดเลือกคุณลักษณะ

จากรูปที่ 31 บรรทัดที่ 1 – บรรทัดที 5 เป็นการหาค่าสนับสนุน และค่าความเชื่อมั่นโดยข้อมูลนำเข้าเป็นข้อมูลรูปแบบแนวตั้ง ค่าสนับสนุนคำนวณด้วยวิธีการ หาค่าจำนวนของผลการอินเตอร์เซคชั่นระหว่างเซตของคุณลักษณะและเซตของคลาส และค่าความเชื่อมั่นคำนวณจากค่าสนับสนุนของคุณลักษณะ จากนั้นบรรทัดที่ 6 เป็นการจัดลำดับคุณลักษณะโดยเรียงค่าสนับสนุนจากมากไปน้อย จากนั้นบรรทัดที่ 7 – 11 เป็นการคำนวณค่าน้ำหนักของคุณลักษณะที่ทำการจัดลำดับแล้ว โดยที่ บรรทัดที่ 8 คือ การปรับค่าน้ำหนักของค่าสนับสนุนให้มีค่าอยู่ระหว่าง บรรทัดที่ 9 เป็นการคำนวณค่าน้ำหนักของคุณลักษณะที่อยู่ในคลาส และบรรทัดที่ 10 คือการหาค่าน้ำหนักที่มีค่ามากที่สุดมาเป็นค่าน้ำหนักของคุณลักษณะ t_i

3.5 การลดคุณลักษณะที่ซ้ำซ้อน

หลังจากทำการหาค่าน้ำหนักของแต่ละคุณลักษณะเรียบร้อยแล้ว ผู้วิจัยทำการเรียงลำดับคุณลักษณะที่มีความสำคัญสูงสุดไว้ลำดับแรก

ตาราง 16 ตัวอย่างเวกเตอร์แนวนอน

Feature (F) / Document_ID	t_1	t_2	t_3	t_4	t_5	t_6	Class
d_1	1	0	1	0	1	0	c_1
d_2	0	1	0	1	0	0	c_1
d_3	0	1	1	1	0	1	c_1
d_4	0	0	1	0	0	0	c_2
d_5	1	0	0	0	1	0	c_2

ตาราง 17 รูปแบบข้อมูลแนวตั้ง (Vertical Data)

Transaction	Set of Documents
t_1	$\{d_1, d_5\}$
t_2	$\{d_2, d_3\}$
t_3	$\{d_1, d_3, d_4\}$
t_4	$\{d_2, d_3\}$
t_5	$\{d_1, d_5\}$
t_6	$\{d_3\}$

ผู้วิจัยทำการตัดคุณลักษณะที่ซ้ำซ้อนโดยใช้แนวคิด ดังต่อไปนี้

นิยามที่ 3.8 กำหนดให้ $t_i, t_j \in T$ ถ้า $s(t_i) = s(t_j)$ แสดงว่า t_i และ t_j ปรากฏอยู่ในเอกสารชุดเดียวกัน ให้ตัดตัวใดตัวหนึ่งออกได้ โดยการจะตัดตัวใดออกจะพิจารณาจากค่าน้ำหนักของคุณลักษณะที่ได้คำนวณไว้แล้ว

ตัวอย่างที่ 3.8 จากตาราง 17 พบว่า t_1, t_5 เป็นคุณลักษณะที่ปรากฏในเอกสารเดียวกัน คือ d_1 และ d_5 และ t_2, t_4 ปรากฏในเอกสารเดียวกัน คือ d_2 และ d_3 จากนั้นผู้วิจัยจะทำการตรวจสอบค่าน้ำหนักของคำคุณลักษณะเหล่านั้น เพื่อเลือกคุณลักษณะที่มีค่าน้ำหนักสูงที่สุดไว้ แล้วตัดคุณลักษณะที่เหลือออก ซึ่งจากตัวอย่างพบว่า คุณลักษณะที่มีค่าน้ำหนักสูงสุด คือ t_1 และ t_2 ดังนั้น t_5 และ t_4 จะถูกตัดออก เพราะถือว่าเป็นคุณลักษณะที่ซ้ำกับ t_1 และ t_2 และเมื่อลบออกแล้วจะเห็นว่าจำนวนคุณลักษณะลดลง แต่เอกสารยังคงปรากฏคุณลักษณะอยู่ จึงไม่ทำให้สูญเสียข้อมูลที่ส่งผลต่อตัวจำแนก เมื่อผ่านกระบวนการลดคุณลักษณะที่ซ้ำแล้ว จะคงเหลือทรานเซคชั่นของคุณลักษณะ ดังนี้

$$s(t_1) = \{d_1, d_5\}$$
$$s(t_2) = \{d_2, d_3\}$$
$$s(t_3) = \{d_1, d_3, d_4\}$$
$$s(t_6) = \{d_3\}$$

เมื่อผ่านการบวนการตัดคุณลักษณะที่ซ้ำออกแล้ว จะต้องทำการแปลงข้อมูลในแนวตั้งกลับมาเป็นเวกเตอร์แนวนอนก่อนนำไปเข้าสู่กระบวนการจำแนก ซึ่งจะเห็นว่าเมื่อแปลงข้อมูลกลับมาอยู่ในรูปแบบเวกเตอร์แนวนอน มีจำนวนคุณลักษณะลดลง แต่ทุกเอกสารยังคงปรากฏคุณลักษณะ ดังตาราง 18

ตาราง 18 ตัวอย่างข้อมูลหลังผ่านกระบวนการลดคุณลักษณะ

Feature (T) / Document (D)	t_1	t_2	t_3	t_6	Class
d_1	1	0	1	0	c_1
d_2	0	1	0	0	c_1
d_3	0	1	1	1	c_1
d_4	0	0	1	0	c_2
d_5	1	0	0	0	c_2

จากแนวคิดข้างต้น สามารถสรุปเป็นขั้นตอนการลดคุณลักษณะ ได้ดังรูปที่ 32

Algorithm 2: Reduce Redundant Features

Input: A Horizontal Vector HD, $HD=D\times T$

Output: A selected features DT

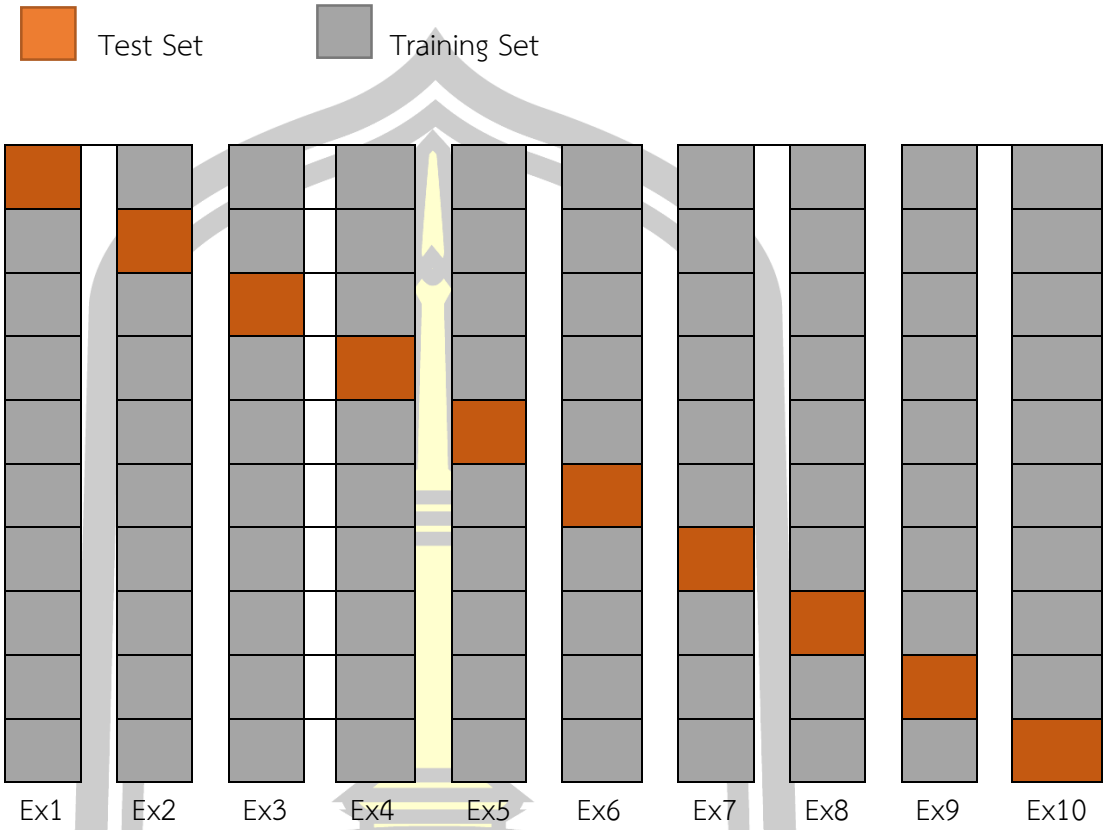
Method:

1. **for** each $t_i \in T$
2. $s(t_i)=\phi$
3. **for** $d_j \in D$
4. **if** $w_{ji}=1$
5. $s(t_i)=s(t_i)\cup d_j$
6. $s(T)=s(T_i)\cup s(t_i)$
7. $VD=T\times s(T)$
8. sort distinct feature in descending order according to number of document.
9. **for** each $t_i \in VD$
10. $s(t_i)=\phi$
11. if $W(t_i)>W(t_{i+1})$
12. $s(t_i)=s(t_i)\cup t_i$
13. else $s(t_i)=s(t_i)\cup t_{i+1}$
14. $s(T)=s(T_i)\cup s(t_i)$
15. $DF=T\times s(T)$

รูปที่ 32 แสดงขั้นตอนการตัดคุณลักษณะที่ซ้ำออก

3.6 การแบ่งข้อมูล

การจำแนกความคิดเห็นด้วยวิธีการเรียนรู้แบบมีผู้สอน จะต้องแบ่งข้อมูลเป็น 2 กลุ่ม คือ ข้อมูลชุดสอน และข้อมูลชุดทดสอบ งานวิจัยนี้แบ่งข้อมูลด้วยวิธีการ 10-Fold Cross Validation อัลกอริทึมทุกตัวจะใช้ข้อมูลชุดสอนและข้อมูลชุดทดสอบเดียวกัน ทำการแบ่งข้อมูลโดยวิธีการสุ่มแบ่งข้อมูลออกเป็น 10 ชุด เท่าๆ กัน แต่ละรอบจะมีข้อมูล 9 ชุด ถูกใช้เป็นข้อมูลชุดสอน และอีก 1 ชุด จะถูกใช้เป็นข้อมูลชุดทดสอบ การเลือกข้อมูลเพื่อนำมาทดสอบในแต่ละรอบจะไม่เลือกข้อมูลที่เคยใช้แล้ว การวัดประสิทธิภาพใช้วิธีการนำค่าประสิทธิภาพการทดสอบแต่ละรอบมาหาค่าเฉลี่ยเพื่อใช้เป็นค่าประสิทธิภาพของตัวจำแนกที่นำมาทดสอบ การแบ่งชุดข้อมูลดังแสดงในรูปที่ 33



*รูปที่ 33 การแบ่งข้อมูลด้วยวิธี 10-*Fold Cross Validation

3.7 กระบวนการจำแนกความคิดเห็น (Sentiment Classification)

การจำแนกความคิดเห็นด้วยวิธีการเรียนรู้แบบมีผู้สอน จะนำข้อมูลชุดสอนที่เตรียมไว้มาเรียนรู้เพื่อสร้างตัวจำแนกสำหรับทำการทดสอบกับข้อมูลชุดทดสอบ เรียกว่าวิธีการเรียนรู้แบบมีผลเฉลย (Supervise Learning) มี 2 ขั้นตอนหลัก คือ 1) ขั้นตอนการเรียนรู้ (Training Step) คือ นำข้อมูลชุดสอนมาเรียนรู้ เพื่อสร้างตัวแบบในการจำแนกความคิดเห็น 2) ขั้นตอนการทดสอบ (Testing Step) คือ นำแบบจำลองการจำแนกความคิดเห็นมาจำแนกข้อมูลชุดทดสอบ งานวิจัยนี้ใช้ขั้นตอนวิธีการจำแนกความคิดเห็น 2 วิธี ได้แก่ วิธีการซัพพอร์ทเวกเตอร์แมชชีน และวิธีการนาอีฟเบย์ ซึ่งจากการศึกษางานวิจัยที่ผ่านมา พบว่า ข้างต้นเป็นวิธีการที่ได้รับความนิยมและมีประสิทธิภาพสูงในการจำแนกความคิดเห็น

หลังจากทำการทดสอบจำแนกข้อมูลชุดทดสอบแล้ว ขั้นตอนต่อไปคือการวัดประสิทธิภาพของการจำแนก เพื่อประเมินความสามารถของการทำนายคลาสคำตอบของตัวจำแนก งานวิจัยนี้ใช้

วิธีการวัดประสิทธิภาพของการจำแนกความคิดเห็น โดยพิจารณาระยะเวลาในการประมวลผล ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวม การวัดประสิทธิภาพด้านเวลาในการประมวลผล ใช้วิธีการวัดตั้งแต่กระบวนการคัดเลือกคุณลักษณะ จนสิ้นสุดกระบวนการจำแนกความคิดเห็น การวัดประสิทธิภาพอธิบายโดยใช้ตาราง Confusion Matrix ขนาด 2x2 โดยที่ข้อมูลด้านคอลัมน์ เป็นคลาสที่อยู่ในข้อมูลชุดสอน (Actual) และข้อมูลด้านแถว เป็นคลาสที่ทำนายได้ (Predicted) ดังตาราง 19

ตาราง 19 Confusion Matrix

Predicted / Actual	Positive	Negative
Positive	a	b
Negative	c	d

โดยที่ a (True Positive) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Positive

b (False Negative) คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส Negative แต่คำตอบคือ Positive

c (False Positive) คือ จำนวนข้อมูลที่ทำนายว่าเป็นคลาส Positive แต่คำตอบคือ Negative

d (True Negative) คือ จำนวนข้อมูลที่ทำนายถูกว่าเป็นคลาส Negative

1) การวัดค่าความถูกต้อง (Accuracy)

การวัดค่าความถูกต้อง เป็นตรวจสอบความถูกต้องของการทำนาย โดยจะพิจารณารวมทุกคลาส ดังสมการ (3.6)

$$Accuracy = \frac{a+d}{a+d+b+c} \quad (3.6)$$

2) การวัดค่าความแม่นยำ (Precision)

การวัดค่าความแม่นยำของการจำแนกความคิดเห็น เป็นตรวจสอบความถูกต้องของการทำนายโดยพิจารณาแยกทีละคลาส เช่น ค่าความแม่นยำของการทำนายคลาสความคิดเห็นเชิงบวก จะพิจารณาจากค่าที่ทำนายถูกต้องว่าเป็นคลาสความคิดเห็นเชิงบวก หารด้วยค่าที่ทำนายว่าเป็นความคิดเห็นเชิงบวกทั้งหมด ดังสมการ (3.7) และ ค่าความแม่นยำของการทำนายคลาสความคิดเห็นเชิงลบ จะพิจารณาจากค่าที่ทำนายถูกต้องว่าเป็นคลาสความคิดเห็นเชิงลบ หารด้วยค่าที่ทำนายว่าเป็นความคิดเห็นเชิงลบทั้งหมด ดังสมการ (3.8)

$$\mathrm{Pr}ecision_{positive} = \frac{a}{a+c} \tag{3.7}$$

$$\mathrm{Pr}ecision_{negative} = \frac{d}{b+d} \tag{3.8}$$

3) การวัดค่าความระลึก (Recall)

การวัดค่าความระลึกของวิธีการจำแนกความคิดเห็นเป็นการวัดความถูกต้องของการทำนายโดยแยกพิจารณาทีละคลาส เช่น ค่าความระลึกของการทำนายคลาสความคิดเห็นเชิงบวก จะพิจารณาจากค่าที่ทำนายถูกต้องว่าเป็นคลาสความคิดเห็นเชิงบวก หารด้วยค่าที่เป็นความคิดเห็นเชิงบวกทั้งหมดในข้อมูลชุดสอน ดังสมการ (3.9) และ ค่าความระลึกของการทำนายคลาสความคิดเห็นเชิงลบ จะพิจารณาจากค่าที่ทำนายถูกต้องว่าเป็นคลาสความคิดเห็นเชิงลบ หารด้วยค่าที่เป็นความคิดเห็นเชิงลบทั้งหมดในข้อมูลชุดสอน ดังสมการ (3.10)

$$\mathrm{Re}call_{positive} = \frac{a}{a+b} \tag{3.9}$$

$$\mathrm{Re}call_{negative} = \frac{d}{c+d} \tag{3.10}$$

4) ค่าเฉลี่ยประสิทธิภาพโดยรวม (F-measure)

ค่าเฉลี่ยประสิทธิภาพโดยรวม เป็นการพิจารณาค่าความระลึกร่วมกับค่าความแม่นยำ ซึ่งจะพิจารณาแยกทีละคลาส ดังสมการ (3.11) และ สมการ (3.12)

$$F-measure_{positive}=2\times\frac{\Pr ecision_{positive}\times \mathrm{Re}call_{positive}}{\Pr ecision_{positive}+\mathrm{Re}call_{positive}} \quad (3.11)$$

$$F-measure_{negative}=2\times\frac{\Pr ecision_{negative}\times \mathrm{Re}call_{negative}}{\Pr ecision_{negative}+\mathrm{Re}call_{negative}} \quad (3.12)$$

หลังจากประเมินประสิทธิภาพตัวจำแนกทั้งหมด นำผลลัพธ์ที่ได้มาเปรียบเทียบประสิทธิภาพระหว่างการเลือกคุณลักษณะด้วยวิธีการที่นำเสนอและการเลือกคุณลักษณะด้วยวิธีการทั่วไป ได้แก่ การใช้ค่าการเพิ่มสารสนเทศ (Information Gain :IG) ค่าสถิติไคสแคว์ (Chi-Square) และ ค่า Gini Index ซึ่งจากการศึกษาพบว่า งานวิจัยทางด้านเหมืองข้อความส่วนใหญ่ใช้ 3 วิธีการนี้ในการคัดเลือกคุณลักษณะ ซึ่งเป็นวิธีการที่ง่ายและมีประสิทธิภาพ [66] งานวิจัยนี้ผู้วิจัยใช้นาอีฟเบย์เป็นตัวจำแนก ซึ่งเป็นวิธีการที่ง่ายต่อการนำไปใช้ และผลลัพธ์ที่สามารถนำไปประยุกต์ใช้ได้ดี [30] จากนั้นวัดค่าประสิทธิภาพการจำแนก ได้แก่ ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก ค่าประสิทธิภาพโดยรวม และระยะเวลาที่ใช้ในการประมวลผล

บทที่ 4
ผลการวิจัยและการอภิปราย

การวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาขั้นตอนวิธีในการลดคุณลักษณะและแก้ไขปัญหาคุณลักษณะซ้ำซ้อนสำหรับการจำแนกความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์ ประกอบด้วย การเลือกคุณลักษณะ การตรวจสอบคุณลักษณะที่ซ้ำซ้อน การจำแนกความคิดเห็นและทดสอบประสิทธิภาพ การนำเสนอผลการวิจัยแบ่งเป็น 3 ส่วน ได้แก่ ส่วนที่ 1. เครื่องมือและข้อมูลต่าง ๆ ที่ใช้ในการทดลอง ส่วนที่ 2. วิธีการทดลอง ส่วนที่ 3. ผลการประเมินประสิทธิภาพและการอภิปรายผลการทดลอง มีรายละเอียดดังต่อไปนี้

4.1 เครื่องมือและข้อมูลที่ใช้ในการทดลอง

4.1.1 เครื่องมือที่ใช้ในการทดลอง

เครื่องมือและข้อมูลต่าง ๆ ที่ใช้ในการทดลองในงานวิจัย ได้แก่ ด้านฮาร์ดแวร์ ประกอบด้วย เครื่องคอมพิวเตอร์แบบพกพา DELL ซีพียู Intel® Core™ i7-4720HQ CPU @ 2.60GHz แรม 8.00 GB ด้านซอฟต์แวร์และภาษาที่ใช้ในการเขียนการทดลอง ได้แก่ ระบบปฏิบัติการ Windows 10 Pro 64-bit พัฒนาซอฟต์แวร์โดยใช้ภาษาไพทอน (Python)

4.1.2 ผลการรวบรวมข้อมูลในการทดลอง

ผู้วิจัยรวบรวมข้อมูลความคิดเห็นบนเครือข่ายสังคมออนไลน์จากแหล่งข้อมูลต่าง ๆ จำนวน 5 ชุดข้อมูล ได้แก่ 1) Stadford Twitter Sentiment Data (STS) 2) SemEval-2017 Task4A Dataset (SemEval) 3) Sentiment Strength Twitter Dataset (SS-Tweet) 4) Health Care Reform (HCR) และ 5) Sanders Twitter Dataset (Sander) ข้อมูลที่รวบรวมทั้งหมดประกอบด้วยหลากหลายโดเมน ในการทดลองผู้วิจัยทำการสุ่มข้อความแบบไม่เจาะจงโดเมน โดยเลือกข้อความคิดเห็นที่เป็นข้อความคิดเห็นเชิงบวกและข้อความคิดเห็นเชิงลบจำนวนเท่ากัน ประกอบด้วย 1) Stadford Twitter Sentiment Data (STS) จำนวน 10,000 ข้อความ 2) SemEval-2017 Task4A Dataset (SemEval) จำนวน 4,000 ข้อความ 3) Sentiment Strength Twitter Dataset (SS-Tweet) จำนวน 2,600 ข้อความ 4) Health Care Reform (HCR) จำนวน 1,000 ข้อความ และ 5) Sanders Twitter Dataset (Sander) จำนวน 1,000 ข้อความ จำนวนข้อมูลทั้งหมดที่ใช้ในการทดลอง แสดงดังตาราง 20

ตาราง 20 ชุดข้อมูลที่ใช้ในการวิจัย

ชุดข้อมูล	จำนวนข้อความคิดเห็น	ข้อความความคิดเห็นเชิงบวก	ข้อความความคิดเห็นเชิงลบ
STS	10,000	5,000	5,000
SemEval	4,000	2,000	2,000
SS-Twitter	2,600	1,300	1,300
HCR	1,000	500	500
Sanders	1,000	500	500

4.1.3 ผลการจัดเตรียมข้อมูล

ข้อมูลที่ใช้ในการทดลองเป็นข้อความซึ่งจัดอยู่ในรูปแบบไม่มีโครงสร้าง จำเป็นต้องทำการแปลงข้อมูลให้อยู่ในรูปแบบมีโครงสร้าง ซึ่งแสดงในรูปแบบของเวกเตอร์ (Vector Model) เพื่อนำเข้ากระบวนการคัดเลือกคุณลักษณะและจำแนกความคิดเห็นต่อไป งานวิจัยนี้สกัดคุณลักษณะโดยการตัดคำที่อยู่ในเอกสารด้วยวิธีการ Unigrams และ Bigrams แทนค่าในเอกสารโดยวิธีการใช้ถุงคำ (Bag of Word) และแทนค่าน้ำหนักในเวกเตอร์ด้วยวิธีการ Boolean Weighting หลังจากผ่านกระบวนการเตรียมข้อมูล จะได้เวกเตอร์ที่มีคุณลักษณะ ดังตาราง 21

ตาราง 21 คุณลักษณะของแต่ละชุดข้อมูล

ชุดข้อมูล	จำนวนคุณลักษณะ
STS	12,772
SemEval	9,065
SS-Twitter	6,845
HCR	2,503
Sanders	1,867

4.2 วิธีการทดลอง

ผู้วิจัยแบ่งวิธีการทดลองเป็น 2 ส่วน ดังนี้

ส่วนที่ 1 การเปรียบเทียบประสิทธิภาพของวิธีการเลือกคุณลักษณะ ขั้นตอนนี้เป็นการนำข้อความคิดเห็นมาใช้ทดลองกับวิธีการตัดคำด้วยวิธีการ Unigram + Bigram จากนั้นแบ่งข้อมูลเป็นชุดสอนและชุดทดสอบ จากนั้นนำข้อมูลชุดสอนไปทำการจัดอันดับคุณลักษณะโดยการหาค่าน้ำหนักของคุณลักษณะด้วยวิธีการที่นำเสนอ เปรียบเทียบกับ วิธีการอื่น ๆ ได้แก่ Information Gain (IG)

Chi-Square (Chi2) และ Gini Index ซึ่งงานวิจัยทางด้านเหมืองข้อความส่วนใหญ่ใช้ 3 วิธีการนี้ในการคัดเลือกคุณลักษณะ ซึ่งเป็นวิธีการที่ง่ายและมีประสิทธิภาพ [10] แล้วทำการเลือกคุณลักษณะตามค่าน้ำหนักโดยเรียงลำดับจากค่าน้ำหนักจากมากไปน้อย จำนวนคุณลักษณะที่นำมาใช้ในการทดลอง คือ 10% 20% 30% 40% 50% 60% 70% 80% 90% และ 100% โดยจำนวนคุณลักษณะ 10% หมายถึง คุณลักษณะที่มีความสำคัญสูงสุด 10% จากคุณลักษณะทั้งหมด เช่น จำนวนคุณลักษณะมีจำนวน 1,000 คุณลักษณะ จะทำการเรียงคุณลักษณะทั้งหมดตามค่าน้ำหนักจากมากไปน้อย จากนั้นทำการเลือกคุณลักษณะ 100 คุณลักษณะแรก ซึ่งหมายถึง 10% ของคุณลักษณะทั้งหมด จากนั้นทำการจำแนกความคิดเห็นด้วยวิธีการ Naïve Bayes

ส่วนที่ 2 การเปรียบเทียบประสิทธิภาพการเลือกคุณลักษณะทั้งหมดกับการลดคุณลักษณะที่ซ้ำซ้อน โดยใช้การจำแนกความคิดเห็นด้วยวิธีการ Naïve Bayes

4.3 ผลการทดลอง

ผลการวัดประสิทธิภาพการทดลอง นำเสนอตามลำดับ 5 ชุดข้อมูล ได้แก่ 1) Stadford Twitter Sentiment Data (STS) 2) SemEval-2017 Task4A Dataset (SemEval) 3) Sentiment Strength Twitter Dataset (SS-Tweet) 4) Health Care Reform (HCR) และ 5) Sanders Twitter Dataset (Sander) รายละเอียดผลการวัดประสิทธิภาพการทดลอง มีดังนี้

4.3.1 ผลการเปรียบเทียบประสิทธิภาพของวิธีการเลือกคุณลักษณะ

การนำเสนอผลการเปรียบเทียบประสิทธิภาพของวิธีการเลือกคุณลักษณะ ประกอบด้วย ค่าความถูกต้อง ค่าความแม่นยำในคลาสเชิงบวก ค่าความแม่นยำในคลาสเชิงลบ ค่าความระลึกในคลาสเชิงบวก ค่าความระลึกในคลาสเชิงลบ ค่าประสิทธิภาพโดยรวมในการจำแนกความคิดเห็นคลาสเชิงบวก และค่าประสิทธิภาพโดยรวมในคลาสเชิงลบ โดยผู้วิจัยได้นำเสนอผลการวัดประสิทธิภาพตามจำนวนการเลือกคุณลักษณะ ในส่วนของคุณลักษณะที่นำมาใช้ในการทดลอง จำนวน 100% หมายถึงไม่มีการเลือกคุณลักษณะ ดังนั้น ค่าประสิทธิภาพในแต่ละวิธีการจึงมีค่าเท่ากัน รายละเอียดนำเสนอตามชุดข้อมูล ดังนี้

1) ผลการวัดประสิทธิภาพการทดลองข้อมูลชุด STS

ชุดข้อมูล Stanford Twitter Sentiment Data (STS) ประกอบด้วยข้อความคิดเห็น จำนวน 10,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 5,000 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 5,000 ข้อความ แสดงผลการทดลองได้ดังนี้

ตาราง 22 ค่าความถูกต้องของชุดข้อมูล STS

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	71.16	71.77	71.43	71.26	71.21	71.69	71.66	72.15	72.25	71.62
Chi2	71.05	71.77	71.45	71.28	71.25	71.72	71.66	72.10	72.25	71.61
IG	71.06	71.34	71.18	71.08	71.05	71.25	71.54	72.10	72.27	71.43
Proposed	**71.91**	**72.39**	**72.13**	**71.89**	**71.95**	**71.94**	**72.27**	**72.43**	**72.30**	**72.13**

ตาราง 23 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความถูกต้องของข้อมูลชุด* STS

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.514	0.254	0.085	0.319	0.709	6.083	8	0.000
Pair 2 Proposed-Chi2	0.520	0.261	0.087	0.319	0.721	5.967	8	0.000
Pair 3 Proposed-IG	0.704	0.325	0.108	0.454	0.954	6.498	8	0.000

จากตาราง 22 แสดงประสิทธิภาพความถูกต้องการจำแนกความคิดเห็นของข้อมูลชุด STS มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 70% และ 80% พบว่า วิธีการที่นำเสนอมีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการ Chi-Square ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% 30% 40% 50% และ 60% พบว่า วิธีการที่นำเสนอมีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอมีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และ วิธีการ Gini Index ตามลำดับ

จากตาราง 23 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความถูกต้องสูงกว่าวิธีการอื่นอย่างมีนัยสำคัญที่ 0.05

ตาราง 24 ค่าความแม่นยำของชุดข้อมูล STS *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	69.83	70.50	70.15	69.94	69.81	69.98	69.87	70.46	70.81	70.15
Chi2	69.72	70.50	70.19	69.97	69.86	70.03	69.87	70.41	70.81	70.15
IG	69.69	70.02	69.89	69.81	69.65	69.51	69.62	70.40	70.85	69.94
Proposed	71.01	71.24	71.33	71.93	73.14	73.68	72.64	71.63	70.86	71.94

ตาราง 25 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความแม่นยำของข้อมูลชุด* STS *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	1.790	1.239	0.413	0.838	2.742	4.335	8	0.002
Pair 2 Proposed-Chi2	1.789	1.214	0.405	0.855	2.722	4.420	8	0.002
Pair 3 Proposed-IG	2.002	1.319	0.440	0.988	3.016	4.553	8	0.002

จากตาราง 24 แสดงประสิทธิภาพความแม่นยำการจำแนกความคิดเห็นของชุดข้อมูล STS ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% 70% และ 80% พบว่า วิธีการที่นำเสนอมีความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% 40% 50% และ 60% พบว่า วิธีการที่นำเสนอมีความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอมีความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

จากตาราง 25 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงบวก สูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 26 ค่าความแม่นยำของชุดข้อมูล STS *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	73.17	73.41	73.48	73.86	74.10	74.43	74.14	74.21	73.94	73.86
Chi2	73.06	73.41	73.47	73.85	74.12	74.42	74.14	74.17	73.94	73.84
IG	73.31	73.21	73.45	73.83	73.97	74.02	74.23	74.18	73.94	73.79
Proposed	73.01	73.73	73.07	71.92	70.95	70.49	71.98	73.36	73.99	72.50

ตาราง 27 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความแม่นยำของข้อมูลชุด* STS *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-1.360	1.510	0.503	-2.521	-0.199	-2.701	8	0.027
Pair 2 Proposed-Chi2	-1.342	1.525	0.508	-2.514	-0.170	-2.641	8	0.030
Pair 3 Proposed-IG	-1.293	1.433	0.478	-2.394	-0.192	-2.708	8	0.027

จากตาราง 26 แสดงประสิทธิภาพความแม่นยำการจำแนกความคิดเห็นของชุดข้อมูล STS ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% และ 70% พบว่า วิธีการ Information Gain มีความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% 40% และ 60% พบว่า วิธีการ Gini Index มีความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการ Chi-Square มีความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% 50% 60% และ 70% พบว่า วิธีการ Gini Index มีความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

จากตาราง 27 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงลบน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 28 ค่าความระลึกของชุดข้อมูล STS *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	75.31	75.22	75.47	76.02	76.42	76.93	76.68	76.39	75.76	76.02
Chi2	75.22	75.22	75.43	75.98	76.42	76.89	76.68	76.37	75.76	76.00
IG	75.65	75.26	75.58	76.05	76.42	76.76	76.98	76.39	75.74	76.09
Proposed	74.31	75.23	74.15	71.95	69.52	68.38	71.58	74.39	75.80	72.81

ตาราง 29 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของข้อมูลชุด* STS *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-3.210	3.108	1.036	-5.599	-0.821	-3.098	8	0.015
Pair 2 Proposed-Chi2	-3.184	3.111	1.037	-5.575	-0.793	-3.071	8	0.015
Pair 3 Proposed-IG	-3.280	3.058	1.019	-5.631	-0.929	-3.218	8	0.012

จากตาราง 28 แสดงประสิทธิภาพค่าระลึกการจำแนกความคิดเห็นของชุดข้อมูล STS ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 30% และ 40% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% 60% และ 70% พบว่า วิธีการ Gini Index มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Gini Index มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

จากตาราง 29 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 30 ค่าความระลึกของชุดข้อมูล STS ในคลาสเชิงลบ

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	67.23	68.45	67.64	66.85	66.37	66.72	66.82	67.97	68.78	67.43
Chi2	67.09	68.45	67.72	66.93	66.45	66.82	66.82	67.89	68.78	67.44
IG	66.73	67.60	67.07	66.49	66.06	66.01	66.28	67.87	68.84	66.99
Proposed	69.61	69.62	70.17	71.90	74.44	75.56	73.03	70.53	68.84	71.52

ตาราง 31 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความระลึกของข้อมูลชุด* STS *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	4.097	3.088	1.029	1.723	6.470	3.980	8	0.004
Pair 2 Proposed-Chi2	4.083	3.044	1.015	1.744	6.423	4.024	8	0.004
Pair 3 Proposed-IG	4.528	3.178	1.059	2.085	6.971	4.274	8	0.003

จากตาราง 30 แสดงประสิทธิภาพค่าระลึกการจำแนกความคิดเห็นของชุดข้อมูล STS ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% 70% และ 80% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% 40% 50% และ 60% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

จากตาราง 31 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงลบสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 32 *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* STS *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	72.47	72.78	72.71	72.85	72.97	73.29	73.11	73.30	73.20	72.97
Chi2	72.37	72.78	72.72	72.85	72.99	73.30	73.11	73.27	73.20	72.95
IG	72.55	72.55	72.62	72.80	72.88	72.96	73.12	73.27	73.21	72.88
Proposed	72.63	73.18	72.71	71.94	71.28	70.93	72.11	72.99	73.25	72.34

ตาราง 33 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของข้อมูลชุด* STS *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.629	0.934	0.311	-1.347	0.089	-2.020	8	0.078
Pair 2 Proposed-Chi2	-0.619	0.951	0.317	-1.350	0.112	-1.953	8	0.087
Pair 3 Proposed-IG	-0.549	0.882	0.294	-1.227	0.129	-1.867	8	0.099

จากตาราง 32 แสดงประสิทธิภาพความโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล STS ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% และ 90% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการที่นำเสนอ และวิธีการ Information Gain ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% และ 80% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% และ 60% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

จากตาราง 33 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกไม่แตกต่างจากวิธีการอื่น

ตาราง 34 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล STS *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	70.07	70.85	70.44	70.18	70.02	70.36	70.29	70.95	71.27	70.49
Chi2	69.95	70.84	70.48	70.22	70.07	70.42	70.29	70.89	71.27	70.49
IG	69.86	70.29	70.11	69.97	69.79	69.79	70.03	70.89	71.30	70.23
Proposed	71.27	71.62	71.59	71.91	72.65	72.94	72.50	71.92	71.32	71.97

ตาราง 35 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของข้อมูลชุด* STS *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	1.477	0.875	0.292	0.804	2.149	5.063	8	0.001
Pair 2 Proposed-Chi2	1.477	0.849	0.283	0.824	2.129	5.217	8	0.001
Pair 3 Proposed-IG	1.743	0.976	0.325	0.993	2.494	5.358	8	0.001

จากตาราง 34 แสดงประสิทธิภาพความโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล STS ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% 30% 70% และ 80% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% 50% และ 60% พบว่า วิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

จากตาราง 35 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีประสิทธิภาพโดยรวมในคลาสเชิงลบ สูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

2) ผลการวัดประสิทธิภาพการทดลองข้อมูลชุด SemEval

ข้อมูลชุด SemEval เป็นข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์ ซึ่งถูกรวบรวมไว้ในข้อมูลชุด SemEval-2017 Task4A ในการทดลองผู้วิจัยได้ทำการสุ่มข้อความคิดเห็น จำนวน 4,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 2,000 ข้อความ และข้อความคิดเห็นเชิงลบ จำนวน 2,000 ข้อความ รายละเอียดผลประสิทธิภาพแสดงตามลำดับ ดังนี้

ตาราง 36 ค่าความถูกต้องของชุดข้อมูล SemEval

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	86.68	87.38	86.95	87.23	87.00	87.38	87.48	87.55	87.68	87.26
Chi2	86.68	87.38	86.95	87.23	87.00	87.38	87.50	87.55	87.68	87.26
IG	86.68	87.40	87.38	86.98	86.95	87.18	87.23	87.55	87.68	87.22
Proposed	**87.18**	87.10	**87.38**	**87.33**	**87.50**	87.33	**87.63**	87.50	87.68	87.40

ตาราง 37 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความถูกต้องของข้อมูลชุด* SemEval

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.144	0.277	0.092	-0.068	0.357	1.565	8	0.156
Pair 2 Proposed-Chi2	0.142	0.277	0.092	-0.071	0.355	1.541	8	0.162
Pair 3 Proposed-IG	0.178	0.288	0.096	-0.044	0.399	1.849	8	0.102

จากตาราง 36 แสดงประสิทธิภาพความถูกต้องของการจำแนกความคิดเห็นของชุดข้อมูล SemEval มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 40% และ 50% พบว่า วิธีการที่นำเสนอมีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการ Gini Index มีค่าความถูกต้องเท่ากัน

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 80% พบว่า วิธีการ Gini Index มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการ Gini Index มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าความถูกต้องเท่ากัน

จากตาราง 37 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพความถูกต้องไม่แตกต่างจากวิธีการอื่น

ตาราง 38 *ค่าความแม่นยำของชุดข้อมูล* SemEval *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	87.17	88.24	87.65	88.34	88.41	88.58	88.90	88.70	88.47	88.27
Chi2	87.17	88.24	87.65	88.34	88.41	88.58	88.95	88.70	88.47	88.28
IG	87.62	88.29	88.05	88.10	88.41	88.60	88.58	88.70	88.47	88.31
Proposed	87.34	87.27	87.47	88.08	89.44	89.72	89.43	88.64	88.47	88.43

ตาราง 39 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* SemEval *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.156	0.661	0.220	-0.353	0.664	0.706	8	0.500
Pair 2 Proposed-Chi2	0.150	0.658	0.219	-0.356	0.656	0.684	8	0.513
Pair 3 Proposed-IG	0.116	0.740	0.247	-0.453	0.684	0.469	8	0.652

จากตาราง 38 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล SemEval ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% และ 20% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ซึ่งมีค่าความแม่นยำในคลาสเชิงบวกเท่ากัน

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ซึ่งมีค่าความแม่นยำในคลาสเชิงบวกเท่ากัน

จากตาราง 39 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพแม่นยำในคลาสเชิงบวก ไม่แตกต่างจากวิธีการอื่น

ตาราง 40 *ค่าความแม่นยำของชุดข้อมูล* SemEval *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	86.14	86.52	86.24	86.16	85.70	86.23	86.14	86.48	86.94	86.28
Chi2	86.14	86.52	86.24	86.16	85.70	86.23	86.14	86.48	86.94	86.28
IG	85.76	86.54	86.71	85.92	85.61	85.87	85.97	86.48	86.94	86.20
Proposed	86.97	86.90	87.32	86.64	85.79	85.23	86.02	86.46	86.94	86.47

ตาราง 41 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* SemEval *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.191	0.605	0.202	-0.274	0.656	0.948	8	0.371
Pair 2 Proposed-Chi2	0.191	0.605	0.202	-0.274	0.656	0.948	8	0.371
Pair 3 Proposed-IG	0.274	0.530	0.177	-0.133	0.682	1.552	8	0.159

จากตาราง 40 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล SemEval ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 40% และ 50% พบว่า วิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% และ 30% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

จากตาราง 41 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพความแม่นยำในคลาสเชิงลบ ไม่แตกต่างจากวิธีการอื่น

ตาราง 42 ค่าความระลึกของชุดข้อมูล SemEval *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	85.94	86.23	86.00	85.79	85.22	85.82	85.64	86.06	86.67	85.93
Chi2	85.94	86.23	86.00	85.79	85.22	85.82	85.64	86.06	86.67	85.93
IG	85.43	86.22	86.50	85.55	85.12	85.37	85.48	86.06	86.67	85.82
Proposed	86.91	86.82	87.25	86.37	85.05	84.29	85.35	86.07	86.67	86.09

ตาราง 43 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* SemEval *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.157	0.823	0.274	-0.476	0.789	0.571	8	0.584
Pair 2 Proposed-Chi2	0.157	0.823	0.274	-0.476	0.789	0.571	8	0.584
Pair 3 Proposed-IG	0.264	0.737	0.246	-0.302	0.831	1.077	8	0.313

จากตาราง 42 แสดงประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็นของชุดข้อมูล SemEval ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% และ 40% พบว่า วิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% 60% และ 70% พบว่า วิธีการ Gini Index มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าระลึกในคลาสเชิงบวกสูงเท่ากัน

จากตาราง 43 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพความระลึกในคลาสเชิงบวก ไม่แตกต่างจากวิธีการอื่น

ตาราง 44 *ค่าความระลึกของชุดข้อมูล* SemEval *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	87.41	88.45	87.89	88.74	88.91	88.98	89.27	88.94	88.60	88.58
Chi2	87.41	88.45	87.89	88.74	88.91	88.98	89.32	88.94	88.60	88.58
IG	87.97	88.50	88.24	88.48	88.90	89.00	88.91	88.94	88.60	88.61
Proposed	87.39	87.33	87.39	88.19	89.83	90.23	89.79	88.84	88.60	88.62

ตาราง 45 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* SemEval *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.044	0.748	0.249	-0.531	0.620	0.178	8	0.863
Pair 2 Proposed-Chi2	0.039	0.744	0.248	-0.533	0.611	0.157	8	0.879
Pair 3 Proposed-IG	0.006	0.842	0.281	-0.642	0.653	0.020	8	0.985

จากตาราง 44 แสดงประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็นของชุดข้อมูล SemEval ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% และ 20% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 40% พบว่า วิธีการ Gini Index และวิธีการ Chi-Square มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่าวิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทั้ง 4 วิธีการมีค่าระลึกในคลาสเชิงลบเท่ากัน

จากตาราง 45 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงลบไม่แตกต่างจากวิธีการอื่น

ตาราง 46 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล SemEval *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	86.55	87.22	86.82	87.04	86.78	87.18	87.24	87.36	87.56	87.08
Chi2	86.55	87.22	86.82	87.04	86.78	87.18	87.26	87.36	87.56	87.09
IG	86.51	87.24	87.27	86.81	86.73	86.95	87.00	87.36	87.56	87.05
Proposed	87.12	87.05	87.36	87.21	87.19	86.92	87.34	87.33	87.56	87.23

ตาราง 47 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* SemEval *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.148	0.301	0.100	-0.084	0.379	1.473	8	0.179
Pair 2 Proposed-Chi2	0.146	0.301	0.100	-0.086	0.377	1.449	8	0.185
Pair 3 Proposed-IG	0.183	0.274	0.091	-0.028	0.394	2.004	8	0.080

จากตาราง 46 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล SemEval ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 40% และ 50% พบว่า วิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการ Chi-Square และวิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกเท่ากัน

จากตาราง 47 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวก ไม่แตกต่างจากวิธีการอื่น

ตาราง 48 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล SemEval *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	86.77	87.47	87.06	87.43	87.27	87.58	87.67	87.69	87.76	87.41
Chi2	86.77	87.47	87.06	87.43	87.27	87.58	87.70	87.69	87.76	87.42
IG	86.85	87.51	87.47	87.18	87.23	87.40	87.42	87.69	87.76	87.39
Proposed	87.18	87.12	87.35	87.41	87.76	87.66	87.86	87.64	87.76	87.53

ตาราง 49 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* SemEval *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.116	0.260	0.087	-0.084	0.315	1.334	8	0.219
Pair 2 Proposed-Chi2	0.112	0.259	0.086	-0.087	0.311	1.300	8	0.230
Pair 3 Proposed-IG	0.137	0.297	0.099	-0.091	0.365	1.381	8	0.205

จากตาราง 48 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล SemEval ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Chi-Square และวิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ และวิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% และ 60% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบเท่ากัน

จากตาราง 49 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบ ไม่แตกต่างจากวิธีการอื่น

3) ผลการวัดประสิทธิภาพการทดลองข้อมูลชุด SS-Twitter

ผลการวัดประสิทธิภาพการทดลองข้อมูลชุด SS-Twitter ประกอบด้วย ค่าความถูกต้อง ค่าความแม่นยำในคลาสเชิงบวก ค่าความแม่นยำในคลาสเชิงลบ ค่าความระลึกในคลาสเชิงบวก ค่าความระลึกในคลาสเชิงลบ ค่าประสิทธิภาพโดยรวมในการจำแนกความคิดเห็นคลาสเชิงบวก และค่าประสิทธิภาพโดยรวมในคลาสเชิงลบ แสดงผลประสิทธิภาพตามลำดับ ดังนี้

ตาราง 50 *ค่าความถูกต้องของชุดข้อมูล* SS-Twitter

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	59.31	60.54	60.19	60.69	61.12	60.35	60.73	60.54	60.12	60.40
Chi2	59.23	60.23	60.23	60.42	61.08	60.38	60.58	60.58	60.12	60.32
IG	58.35	60.19	60.38	60.00	60.65	60.04	59.88	60.58	60.12	60.02
Proposed	60.92	62.00	61.77	61.62	61.27	61.35	60.85	60.96	60.12	61.21

ตาราง 51 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความถูกต้องของชุดข้อมูล* SS-Twitter

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.808	0.654	0.218	0.305	1.311	3.703	8	0.006
Pair 2 Proposed-Chi2	0.890	0.695	0.232	0.356	1.424	3.841	8	0.005
Pair 3 Proposed-IG	1.186	0.788	0.263	0.579	1.792	4.511	8	0.002

จากตาราง 50 แสดงประสิทธิภาพความถูกต้องของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% 40% 50% และ 70% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และ วิธีการ Gini Index ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการ Gini Index ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า ทุกวิธีการมีค่าความถูกต้องเท่ากัน

จากตาราง 51 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพความถูกต้องสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 52 ค่าความแม่นยำของชุดข้อมูล SS-Twitter *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	58.51	59.71	59.42	59.83	59.96	59.12	59.40	59.31	58.95	59.36
Chi2	58.45	59.35	59.43	59.53	59.89	59.14	59.31	59.35	58.95	59.27
IG	57.93	59.30	59.57	59.17	59.32	58.75	58.52	59.35	58.95	58.98
Proposed	60.49	61.41	61.66	63.14	63.90	63.16	61.16	60.19	58.95	61.56

ตาราง 53 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* SS-Twitter *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	2.206	1.357	0.452	1.163	3.248	4.878	8	0.001
Pair 2 Proposed-Chi2	2.296	1.385	0.462	1.231	3.360	4.971	8	0.001
Pair 3 Proposed-IG	2.578	1.558	0.519	1.380	3.776	4.962	8	0.001

จากตาราง 52 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% 40% 50% และ 70% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และ วิธีการ Gini Index ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการ Gini Index ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด

จากตาราง 53 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงบวกสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 54 ค่าความแม่นยำของชุดข้อมูล SS-Twitter *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	62.13	62.60	62.88	64.20	64.61	63.09	63.16	62.67	62.05	63.04
Chi2	62.06	62.39	62.95	63.92	64.58	63.14	62.88	62.72	62.05	62.96
IG	61.62	62.94	64.06	64.04	64.90	63.25	62.72	62.72	62.05	63.14
Proposed	61.65	62.99	62.11	60.65	59.66	60.14	60.73	62.24	62.05	61.36

ตาราง 55 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* SS-Twitter *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-1.686	1.847	0.616	-3.106	-0.266	-2.737	8	0.026
Pair 2 Proposed-Chi2	-1.608	1.828	0.609	-3.013	-0.203	-2.639	8	0.030
Pair 3 Proposed-IG	-1.787	1.867	0.622	-3.222	-0.351	-2.871	8	0.021

จากตาราง 54 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 60% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square และวิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าความแม่นยำในคลาสเชิงบวกเท่ากัน

จากตาราง 55 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงบวกน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 56 ค่าความระลึกของชุดข้อมูล SS-Twitter *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	69.87	68.13	68.77	70.21	71.30	70.42	69.97	69.10	68.17	69.55
Chi2	69.80	68.21	68.92	70.13	71.22	70.41	69.58	69.10	68.17	69.50
IG	70.41	69.64	70.82	71.30	72.70	71.14	70.33	69.10	68.17	70.40
Proposed	64.21	65.84	63.36	57.17	53.13	55.66	60.72	66.23	68.17	61.61

ตาราง 57 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* SS-Twitter *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-7.939	6.237	2.079	-12.733	-3.145	-3.819	8	0.005
Pair 2 Proposed-Chi2	-7.894	6.189	2.063	-12.651	-3.137	-3.827	8	0.005
Pair 3 Proposed-IG	-8.791	6.479	2.160	-13.771	-3.811	-4.070	8	0.004

จากตาราง 56 แสดงประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 40% 50% 60% 70% และ 80% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% และ 30% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าระลึกในคลาสเชิงบวกเท่ากัน

จากตาราง 57 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงบวกน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 58 ค่าความระลึกของชุดข้อมูล SS-Twitter *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	49.36	53.38	52.23	51.89	51.55	50.71	51.77	52.22	52.25	51.71
Chi2	49.28	52.70	52.15	51.43	51.54	50.78	51.85	52.30	52.25	51.59
IG	47.08	51.32	50.71	49.50	49.27	49.43	49.81	52.30	52.25	50.19
Proposed	57.76	58.33	60.30	66.25	69.56	67.17	61.08	55.88	52.25	60.95

ตาราง 59 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* SS-Twitter *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	9.247	6.044	2.015	4.601	13.893	4.590	8	0.002
Pair 2 Proposed-Chi2	9.367	6.035	2.012	4.727	14.006	4.656	8	0.002
Pair 3 Proposed-IG	10.768	6.694	2.231	5.623	15.913	4.826	8	0.001

จากตาราง 58 แสดงประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Gini Index มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% 50% และ 60% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square และ วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าระลึกในคลาสเชิงลบเท่ากัน

จากตาราง 59 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงลบสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 60 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล SS-Twitter *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	63.69	63.64	63.75	64.60	65.14	64.28	64.25	63.83	63.23	64.05
Chi2	63.62	63.47	63.82	64.39	65.07	64.28	64.03	63.85	63.23	63.97
IG	63.57	64.06	64.71	64.67	65.33	64.36	63.88	63.85	63.23	64.18
Proposed	62.29	63.55	62.50	60.01	58.02	59.17	60.94	63.07	63.23	61.42

ตาราง 61 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* SS-Twitter *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-2.626	2.522	0.841	-4.564	-0.687	-3.123	8	0.014
Pair 2 Proposed-Chi2	-2.553	2.500	0.833	-4.475	-0.632	-3.064	8	0.015
Pair 3 Proposed-IG	-2.764	2.484	0.828	-4.673	-0.855	-3.339	8	0.010

จากตาราง 60 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% 50% และ 60% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square และวิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกเท่ากัน

จากตาราง 61 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 62 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล SS-Twitter *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	55.01	57.63	57.06	57.40	57.35	56.23	56.90	56.97	56.73	56.81
Chi2	54.94	57.14	57.04	57.00	57.33	56.29	56.83	57.04	56.73	56.70
IG	53.38	56.54	56.61	55.84	56.02	55.49	55.52	57.04	56.73	55.91
Proposed	59.64	60.57	61.19	63.33	64.23	63.46	60.91	58.89	56.73	60.99

ตาราง 63 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* SS-Twitter *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	4.186	2.345	0.782	2.383	5.988	5.354	8	0.001
Pair 2 Proposed-Chi2	4.290	2.361	0.787	2.475	6.105	5.451	8	0.001
Pair 3 Proposed-IG	5.087	2.807	0.936	2.929	7.244	5.436	8	0.001

จากตาราง 62 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล SS-Twitter ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 20% 30% 40 50% และ 70% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการ Gini Index ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบเท่ากัน

จากตาราง 63 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบ สูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

4) ผลการวัดประสิทธิภาพการทดลองข้อมูลชุด HCR

ข้อมูลชุด Health Care Reform (HCR) เป็นข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์เมื่อเดือนมีนาคม ปี ค.ศ. 2010 ในการทดลอง ผู้วิจัยทำการสุ่มข้อความคิดเห็น จำนวน 1,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 500 ข้อความ และ ข้อความคิดเห็นเชิงลบ จำนวน 500 ข้อความ แสดงผลประสิทธิภาพตามลำดับ ดังนี้

ตาราง 64 ค่าความถูกต้องของชุดข้อมูล HCR

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	75.80	78.40	80.10	80.40	81.20	81.70	82.90	81.60	81.90	80.44
Chi2	76.00	78.50	80.30	80.50	81.20	81.60	82.70	81.70	81.90	80.49
IG	76.50	77.90	79.50	81.20	81.10	81.80	82.20	81.50	81.90	80.40
Proposed	77.50	78.10	79.00	80.10	80.80	82.80	82.20	82.90	82.20	80.62

ตาราง 65 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความถูกต้องของชุดข้อมูล* HCR

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.178	0.977	0.326	-0.573	0.929	0.546	8	0.600
Pair 2 Proposed-Chi2	0.133	0.967	0.322	-0.610	0.877	0.414	8	0.690
Pair 3 Proposed-IG	0.222	0.806	0.269	-0.397	0.842	0.827	8	0.432

จากตาราง 64 แสดงประสิทธิภาพความถูกต้องของการจำแนกความคิดเห็นของชุดข้อมูล HCR มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Chi-Square มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการที่นำเสนอ และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Chi-Square มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการ Gini Index และ วิธีการ Chi-Square มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

9. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ซึ่งมีค่าเท่ากัน

จากตาราง 65 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพความถูกต้อง ไม่แตกต่างกับวิธีการอื่น

ตาราง 66 ค่าความแม่นยำของชุดข้อมูล HCR *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	77.09	79.60	80.54	79.88	80.24	80.22	81.65	79.85	80.06	79.90
Chi2	77.55	79.83	80.80	79.90	80.30	80.18	81.68	79.98	80.06	80.03
IG	78.55	79.20	80.47	80.63	79.67	80.16	80.62	79.71	80.06	79.90
Proposed	77.17	77.74	79.24	79.62	80.45	83.47	81.99	82.23	80.78	80.30

ตาราง 67 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* HCR *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.396	1.609	0.536	-0.842	1.633	0.737	8	0.482
Pair 2 Proposed-Chi2	0.268	1.692	0.564	-1.033	1.568	0.475	8	0.648
Pair 3 Proposed-IG	0.402	1.783	0.594	-0.968	1.772	0.677	8	0.518

จากตาราง 66 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล HCR ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% และ 30% พบว่า วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% 70% และ 80% พบว่า วิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ซึ่งมีค่าเท่ากัน

จากตาราง 68 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงบวก ไม่แตกต่างกับวิธีการอื่น

ตาราง 68 ค่าความแม่นยำของชุดข้อมูล HCR *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	76.41	78.24	80.67	81.51	82.89	84.00	84.62	83.66	84.04	81.78
Chi2	76.45	78.21	80.71	81.64	82.58	83.84	84.05	83.69	84.04	81.69
IG	76.58	77.91	79.47	82.53	83.42	84.27	84.40	83.65	84.04	81.81
Proposed	78.28	79.17	79.16	80.79	81.25	82.31	82.51	83.61	83.84	81.21

ตาราง 69 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* HCR *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.569	1.336	0.445	-1.596	0.458	-1.277	8	0.237
Pair 2 Proposed-Chi2	-0.477	1.219	0.406	-1.413	0.460	-1.174	8	0.274
Pair 3 Proposed-IG	-0.594	1.440	0.480	-1.701	0.512	-1.239	8	0.251

จากตาราง 68 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล HCR ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% และ 60% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

9. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

จากตาราง 69 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงลบ ไม่แตกต่างกับวิธีการอื่น

ตาราง 70 ค่าความระลึกของชุดข้อมูล HCR *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	75.62	77.52	80.68	82.28	83.89	85.06	85.52	85.20	85.63	82.38
Chi2	75.45	77.34	80.68	82.46	83.65	84.88	84.94	85.20	85.63	82.25
IG	75.44	77.16	79.18	82.96	84.40	85.12	85.56	85.20	85.63	82.29
Proposed	78.60	79.93	79.26	81.52	82.05	82.41	83.25	84.43	85.23	81.85

ตาราง 71 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* HCR *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.524	1.972	0.657	-2.041	0.992	-0.798	8	0.448
Pair 2 Proposed-Chi2	-0.394	1.949	0.650	-1.893	1.104	-0.607	8	0.561
Pair 3 Proposed-IG	-0.441	2.149	0.716	-2.093	1.211	-0.616	8	0.555

จากตาราง 70 แสดงประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็นของชุดข้อมูล HCR ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% และ 20% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Gini Index และ วิธีการ Chi-Square มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ และ วิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% 60% และ 70% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% และ 90% พบว่า วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain มีค่าระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

จากตาราง 71 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงบวก ไม่แตกต่างกับวิธีการอื่น

ตาราง 72 ค่าความระลึกของชุดข้อมูล HCR *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	77.20	79.75	79.68	78.73	78.74	78.61	80.37	77.91	78.01	78.78
Chi2	77.91	80.22	80.06	78.73	78.93	78.61	80.56	78.10	78.01	79.01
IG	78.85	79.21	80.15	79.54	78.06	78.63	78.94	77.72	78.01	78.79
Proposed	76.80	76.88	79.29	78.87	79.53	83.16	81.17	81.51	79.04	79.58

ตาราง 73 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* HCR *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.806	2.200	0.733	-0.885	2.497	1.099	8	0.304
Pair 2 Proposed-Chi2	0.569	2.351	0.784	-1.238	2.376	0.726	8	0.489
Pair 3 Proposed-IG	0.793	2.455	0.818	-1.094	2.681	0.969	8	0.361

จากตาราง 72 แสดงประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็นของชุดข้อมูล HCR ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% และ 30% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Chi-Square มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% 70% และ 80% พบว่า วิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการ Information Gain ตามลำดับ

จากตาราง 73 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าระลึกในคลาสเชิงลบ ไม่แตกต่างกับวิธีการอื่น

ตาราง 74 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล HCR *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	76.35	78.55	80.61	81.06	82.02	82.57	83.54	82.44	82.75	81.10
Chi2	76.48	78.56	80.74	81.16	81.94	82.46	83.28	82.51	82.75	81.10
IG	76.96	78.17	79.82	81.78	81.97	82.57	83.02	82.36	82.75	81.04
Proposed	77.88	78.82	79.25	80.56	81.24	82.94	82.62	83.32	82.95	81.06

ตาราง 75 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* HCR *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.034	0.929	0.310	-0.749	0.680	-0.111	8	0.914
Pair 2 Proposed-Chi2	-0.033	0.898	0.299	-0.724	0.657	-0.111	8	0.914
Pair 3 Proposed-IG	0.020	0.780	0.260	-0.580	0.620	0.077	8	0.941

จากตาราง 74 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล HCR ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการ Information Gain ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 80% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการ Chi-Square ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการที่นำเสนอตามลำดับ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ซึ่งมีค่าเท่ากัน

จากตาราง 75 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวก ไม่แตกต่างกับวิธีการอื่น

ตาราง 76 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล HCR *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	76.80	78.99	80.17	80.10	80.76	81.22	82.44	80.68	80.91	80.23
Chi2	77.17	79.20	80.38	80.16	80.71	81.14	82.27	80.80	80.91	80.31
IG	77.70	78.55	79.80	81.01	80.65	81.35	81.58	80.58	80.91	80.24
Proposed	77.54	78.01	79.22	79.82	80.38	82.73	81.83	82.55	81.37	80.38

ตาราง 77 ผลการวิเคราะห์ Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* HCR *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	0.153	1.048	0.349	-0.652	0.959	0.439	8	0.672
Pair 2 Proposed-Chi2	0.079	1.064	0.355	-0.739	0.897	0.222	8	0.830
Pair 3 Proposed-IG	0.147	1.001	0.334	-0.622	0.916	0.440	8	0.672

จากตาราง 76 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล HCR ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% และ 30% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Information Gain ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการ Information Gain ตามลำดับ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการ Information Gain ซึ่งมีค่าเท่ากัน

จากตาราง 77 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบ ไม่แตกต่างกับวิธีการอื่น

5) ผลการวัดประสิทธิภาพการทดลองข้อมูลชุด Sanders

ข้อมูลชุด Sander Twitter Dataset (Sander) เป็นข้อมูลที่รวบรวมจากเว็บไซต์ทวิตเตอร์ ในการทดลองผู้วิจัยทำการสุ่มข้อความคิดเห็น จำนวน 1,000 ข้อความ แบ่งเป็นข้อความคิดเห็นเชิงบวก จำนวน 500 ข้อความ ข้อความคิดเห็นเชิงลบ จำนวน 500 ข้อความ แสดงผลประสิทธิภาพตามลำดับ ดังนี้

ตาราง 78 ค่าความถูกต้องของชุดข้อมูล Sanders

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	79.20	83.80	85.20	86.50	88.30	87.60	89.30	90.40	89.80	86.68
Chi2	79.30	83.80	85.20	86.90	88.30	87.80	89.60	90.50	89.80	86.80
IG	79.30	83.70	85.50	87.00	88.40	88.70	89.40	90.40	89.80	86.91
Proposed	79.80	82.20	84.20	86.40	87.00	87.10	87.60	88.50	89.80	85.84

ตาราง 79 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความถูกต้องของชุดข้อมูล* Sanders

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.833	0.875	0.292	-1.506	-0.161	-2.858	8	0.021
Pair 2 Proposed-Chi2	-0.956	0.868	0.289	-1.622	-0.289	-3.304	8	0.011
Pair 3 Proposed-IG	-1.067	0.843	0.281	-1.714	-0.419	-3.798	8	0.005

จากตาราง 78 แสดงประสิทธิภาพความถูกต้องของการจำแนกความคิดเห็นของชุดข้อมูล Sander มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการที่นำเสนอ มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Gini Index มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 50% พบว่า วิธีการ Information Gain มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40 และ 60% พบว่า วิธีการ Information Gain มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Chi-Square มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square วิธีการ Gini Index มีค่าความถูกต้องสูงที่สุด รองลงมาคือ วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า ทุกวิธีการมีค่าความถูกต้องเท่ากัน

จากตาราง 79 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความถูกต้องน้อยกว่าวิธีการอื่นอย่างมีนัยสำคัญที่ 0.05

ตาราง 80 *ค่าความแม่นยำของชุดข้อมูล* Sanders *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	79.12	83.71	85.81	86.40	87.81	86.94	87.92	89.68	88.93	86.26
Chi2	79.30	83.86	85.81	86.77	87.81	86.98	87.99	89.87	88.93	86.37
IG	79.63	83.49	85.88	86.94	88.03	87.82	87.90	90.18	88.93	86.53
Proposed	79.53	82.43	83.58	86.01	86.26	85.65	87.01	87.64	88.83	85.22

ตาราง 81 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* Sanders *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-1.042	0.882	0.294	-1.720	-0.364	-3.545	8	0.008
Pair 2 Proposed-Chi2	-1.153	0.851	0.284	-1.807	-0.499	-4.067	8	0.004
Pair 3 Proposed-IG	-1.318	0.919	0.306	-2.024	-0.611	-4.302	8	0.003

จากตาราง 80 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล Sander ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% และ 70% พบว่า วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 50% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 400% 60% และ 80% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการ Gini Index วิธีการ Information Gain และวิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

จากตาราง 81 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงบวกน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 82 *ค่าความแม่นยำของชุดข้อมูล* Sanders *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	80.50	84.20	84.62	86.92	88.86	88.43	90.82	91.37	90.88	87.40
Chi2	80.55	84.08	84.62	87.28	88.86	88.78	91.43	91.40	90.88	87.54
IG	80.40	84.38	85.26	87.43	88.71	89.58	91.08	90.90	90.88	87.62
Proposed	80.36	82.10	84.65	86.72	87.80	88.61	88.26	89.43	90.90	86.54

ตาราง 83 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความแม่นยำของชุดข้อมูล* Sanders *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.863	1.074	0.358	-1.689	-0.037	-2.410	8	0.042
Pair 2 Proposed-Chi2	-1.006	1.132	0.377	-1.876	-0.135	-2.665	8	0.029
Pair 3 Proposed-IG	-1.088	0.956	0.319	-1.823	-0.353	-3.414	8	0.009

จากตาราง 82 แสดงประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นของชุดข้อมูล Sander ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% และ 80% พบว่า วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Gini Index และ วิธีการ Chi-Square ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% พบว่า วิธีการ Gini Index และ วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 60% พบว่า วิธีการ Information Gain มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และ วิธีการ Gini Index ตามลำดับ

7. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Chi-Square มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

8. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าความแม่นยำในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square ซึ่งมีค่าเท่ากัน

จากตาราง 83 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความแม่นยำในคลาสเชิงลบน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 84 *ค่าความระลึกของชุดข้อมูล* Sanders *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	80.89	84.24	84.25	86.64	89.15	88.80	91.47	91.84	91.46	87.64
Chi2	80.89	84.05	84.25	87.06	89.15	89.18	92.04	91.84	91.46	87.77
IG	80.60	84.52	85.16	87.31	88.94	89.97	91.59	91.12	91.46	87.85
Proposed	80.54	81.77	84.78	86.90	88.10	89.14	88.46	89.62	91.38	86.74

ตาราง 85 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* Sanders *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.894	1.350	0.450	-1.932	0.143	-1.987	8	0.082
Pair 2 Proposed-Chi2	-1.026	1.370	0.457	-2.079	0.028	-2.246	8	0.055
Pair 3 Proposed-IG	-1.109	1.132	0.377	-1.979	-0.238	-2.938	8	0.019

จากตาราง 84 แสดงประสิทธิภาพค่าความระลึกของการจำแนกความคิดเห็นของชุดข้อมูล Sander ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% 50% และ 80% พบว่า วิธีการ Chi-Square และวิธีการ Gini Index มีค่าความระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Information Gain มีค่าความระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Information Gain มีค่าความระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Gini Index และวิธีการ Chi-Square ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% และ 60% พบว่า วิธีการ Information Gain มีค่าความระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Gini Index ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Chi-Square มีค่าความระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Information Gain วิธีการ Gini Index และ วิธีการ Chi-Square มีค่าความระลึกในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ

จากตาราง 85 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความระลึกในคลาสเชิงบวก น้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 86 ค่าความระลึกของชุดข้อมูล Sanders *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	78.60	83.87	86.23	86.49	87.54	86.68	87.43	89.06	88.26	86.02
Chi2	78.84	84.08	86.23	86.89	87.54	86.68	87.43	89.26	88.26	86.13
IG	79.33	83.54	86.06	86.93	88.00	87.68	87.47	89.65	88.26	86.32
Proposed	79.31	82.62	83.62	85.99	86.04	85.32	86.80	87.36	88.31	85.04

ตาราง 87 ผลการวิเคราะห์ Paired-Sample t-test *ค่าความระลึกของชุดข้อมูล* Sanders *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.977	0.996	0.332	-1.742	-0.211	-2.942	8	0.019
Pair 2 Proposed-Chi2	-1.093	0.957	0.319	-1.829	-0.358	-3.428	8	0.009
Pair 3 Proposed-IG	-1.283	0.998	0.333	-2.050	-0.516	-3.858	8	0.005

จากตาราง 86 แสดงประสิทธิภาพค่าความระลึกของการจำแนกความคิดเห็นของชุดข้อมูล Sander ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Information Gain มีค่าความระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% และ 30% พบว่า วิธีการ Chi-Square มีค่าความระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% พบว่า วิธีการ Gini Index และ วิธีการ Chi-Square มีค่าความระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% และ 90% พบว่า วิธีการ Information Gain มีค่าความระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 50% 60% และ 70% พบว่า วิธีการ Information Gain มีค่าความระลึกในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

จากตาราง 87 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าความระลึกในคลาสเชิงลบ น้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 88 *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* Sanders *ในคลาสเชิงบวก*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	80.00	83.97	85.02	86.52	88.47	87.86	89.66	90.74	90.18	86.94
Chi2	80.09	83.95	85.02	86.92	88.47	88.07	89.97	90.84	90.18	87.06
IG	80.11	84.00	85.52	87.12	88.48	88.88	89.71	90.65	90.18	87.18
Proposed	80.03	82.10	84.18	86.46	87.17	87.36	87.73	88.62	90.09	85.97

ตาราง 89 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* Sanders *ในคลาสเชิงบวก*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.964	0.866	0.289	-1.630	-0.299	-3.340	8	0.010
Pair 2 Proposed-Chi2	-1.086	0.858	0.286	-1.745	-0.426	-3.797	8	0.005
Pair 3 Proposed-IG	-1.212	0.766	0.255	-1.801	-0.623	-4.746	8	0.001

จากตาราง 88 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล Sander ในคลาสเชิงบวก มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการที่นำเสนอ และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% 30% และ 50% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และ วิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% และ 60% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 80% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการ Gini Index มีค่าประสิทธิภาพโดยรวมในคลาสเชิงบวกสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการที่นำเสนอ ตามลำดับ

จากตาราง 89 แสดงให้เห็นว่า ขั้นตอนวิธีการที่นำเสนอมีค่าประสิทธิภาพโดยรวมน้อยกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 90 ค่าประสิทธิภาพโดยรวมของชุดข้อมูล Sanders *ในคลาสเชิงลบ*

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	79.54	84.03	85.42	86.70	88.19	87.55	89.10	90.20	89.55	86.70
Chi2	79.69	84.08	85.42	87.08	88.19	87.72	89.39	90.31	89.55	86.83
IG	79.86	83.96	85.66	87.18	88.35	88.62	89.24	90.27	89.55	86.97
Proposed	79.83	82.36	84.13	86.35	86.91	86.93	87.52	88.39	89.59	85.78

ตาราง 91 *ผลการวิเคราะห์* Paired-Sample t-test *ค่าประสิทธิภาพโดยรวมของชุดข้อมูล* Sanders *ในคลาสเชิงลบ*

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.919	0.779	0.260	-1.518	-0.320	-3.538	8	0.008
Pair 2 Proposed-Chi2	-1.047	0.772	0.257	-1.640	-0.454	-4.070	8	0.004
Pair 3 Proposed-IG	-1.187	0.737	0.246	-1.753	-0.620	-4.833	8	0.001

จากตาราง 90 แสดงประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นของชุดข้อมูล Sander ในคลาสเชิงลบ มีรายละเอียดผลการทดลอง ดังนี้

1. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการที่นำเสนอ วิธีการ Chi-Square และวิธีการ Gini Index ตามลำดับ

2. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 20% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Information Gain และวิธีการที่นำเสนอ ตามลำดับ

3. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 30% และ 50% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Gini Index วิธีการ Chi-Square และวิธีการที่นำเสนอ ตามลำดับ

4. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 40% และ 60% พบว่า วิธีการ Information Gain มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Gini Index และวิธีการที่นำเสนอ ตามลำดับ

5. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 70% และ 80% พบว่า วิธีการ Chi-Square มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Information Gain วิธีการ Gini Index และ วิธีการที่นำเสนอ ตามลำดับ

6. เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 90% พบว่า วิธีการที่นำเสนอ มีค่าประสิทธิภาพโดยรวมในคลาสเชิงลบสูงที่สุด รองลงมาคือ วิธีการ Chi-Square วิธีการ Information Gain และ วิธีการ Gini Index ซึ่งมีค่าเท่ากัน

สรุปผลประสิทธิภาพของการจำแนกความคิดเห็นโดยภาพรวม พบว่า วิธีการที่นำเสนอมีประสิทธิภาพสูงกว่าวิธีการอื่น เมื่อทดลองกับข้อมูลขนาดใหญ่ที่มีคุณลักษณะเป็นจำนวนมาก ซึ่งจากการทดลองกับข้อมูลทั้ง 5 ชุด พบว่า วิธีการที่นำเสนอมีประสิทธิภาพดีที่สุดเมื่อทดลองกับข้อมูลชุด Stadford Twitter Sentiment Data (STS) ขนาด 10,000 ข้อความ ข้อมูลชุด SemEval-2017 Task4A Dataset (SemEval) ขนาด 4,000 ข้อความ ข้อมูลชุด Sentiment Strength Twitter Dataset (SS-Tweet) ขนาด 2,600 ข้อความ ส่วนการทดสอบกับข้อมูลขนาดเล็ก พบว่า วิธีการที่นำเสนอมีประสิทธิภาพการจำแนกสูงที่สุด เมื่อเลือกคุณลักษณะที่สำคัญที่สุดจำนวน 10% ทั้งนี้ จากการทดสอบการจัดลำดับของคุณลักษณะพบว่า วิธีการที่นำเสนอมีความคล้ายคลึงกับการจัดลำดับคุณลักษณะด้วยวิธีการ Information Gain และ วิธีการ Chi Square มีข้อแตกต่างกัน คือ วิธีการที่นำเสนอจะให้ความสำคัญกับคุณลักษณะที่มีค่าความถี่สูงมากกว่า จากการพิจารณาข้อมูลเวกเตอร์ พบว่า มีคุณลักษณะที่เกิดขึ้นในเอกสารเป็นจำนวนมากกว่า 20 ครั้ง คิดเป็นร้อยละ 3 การจัดลำดับคุณลักษณะเหล่านี้ด้วยวิธีการที่นำเสนอความแตกต่างจากวิธีการอื่น เนื่องจากวิธีการที่นำเสนอพิจารณาค่าสนับสนุนหรือความถี่ของการปรากฏคุณลักษณะร่วมกับค่าความเชื่อมั่นของคุณลักษณะ

4.3.2 ผลการวัดประสิทธิภาพด้านเวลา

ผู้วิจัยใช้หน่วยวัดระยะเวลาเป็นวินาที การวัดเริ่มตั้งแต่กระบวนการคำนวณค่าน้ำหนักของคุณลักษณะ ระยะเวลาในการจัดลำดับคุณลักษณะ จนถึงระยะเวลาในการเลือกคุณลักษณะ ผลการวัดประสิทธิภาพด้านเวลาแสดงการเปรียบเทียบระยะเวลาที่ใช้แต่ละวิธีการของแต่ละชุดข้อมูล ดังตาราง 92

ตาราง 92 ประสิทธิภาพด้านเวลาที่ใช้ในการคัดเลือกคุณลักษณะ ข้อมูลชุด STS

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	2.475	2.752	2.986	3.232	3.479	3.697	3.939	4.187	4.423	3.149
Chi2	1.948	2.233	2.600	2.893	3.219	3.573	3.826	4.152	4.461	2.923
IG	2.691	2.929	3.257	3.531	3.868	4.148	4.472	4.769	5.131	3.512
Proposed	1.834	2.158	2.504	2.811	3.149	3.438	3.771	4.069	4.379	2.844

ตาราง 93 ผลการวิเคราะห์ Paired-Sample t-test *ประสิทธิภาพด้านเวลาของชุดข้อมูล* STS

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.340	0.211	0.070	-0.501	-0.178	-4.837	8	0.001
Pair 2 Proposed-Chi2	-0.088	0.024	0.008	-0.107	-0.069	-10.873	8	0.000
Pair 3 Proposed-IG	-0.743	0.050	0.017	-0.781	-0.704	-44.860	8	0.000

จากตาราง 92 แสดงประสิทธิภาพด้านเวลา เมื่อทดสอบกับชุดข้อมูล Stadford Twitter Sentiment Data (STS) พบว่า วิธีการที่นำเสนอใช้ระยะเวลาในการคัดเลือกคุณลักษณะน้อยที่สุด และเมื่อพิจารณาผลการวิเคราะห์ Paired-Sample t-test ในตาราง 93 แสดงให้เห็นว่า วิธีการที่นำเสนอมีประสิทธิภาพด้านเวลาสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 94 ประสิทธิภาพด้านเวลาที่ใช้ในการคัดเลือกคุณลักษณะ ข้อมูลชุด SemEval

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	0.733	0.785	0.834	0.874	0.914	0.953	0.992	1.028	1.088	0.911
Chi2	0.681	0.703	0.750	0.785	0.831	0.863	0.904	0.948	0.988	0.828
IG	0.781	0.788	0.831	0.875	0.912	0.951	0.988	1.028	1.080	0.915
Proposed	0.558	0.558	0.606	0.645	0.684	0.723	0.763	0.804	0.845	0.687

ตาราง 95 ผลการวิเคราะห์ Paired-Sample t-test *ประสิทธิภาพด้านเวลาของชุดข้อมูล* SemEval

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.224	0.019	0.006	-0.239	-0.209	-35.528	8	0.000
Pair 2 Proposed-Chi2	-0.141	0.007	0.002	-0.146	-0.135	-60.308	8	0.000
Pair 3 Proposed-IG	-0.227	0.004	0.001	-0.231	-0.224	-170.890	8	0.000

จากตาราง 94 แสดงประสิทธิภาพด้านเวลา เมื่อทดสอบกับชุดข้อมูล SemEval พบว่า วิธีการที่นำเสนอใช้ระยะเวลาในการคัดเลือกคุณลักษณะน้อยที่สุด และเมื่อพิจารณาผลการวิเคราะห์ Paired-Sample t-test ในตาราง 95 แสดงให้เห็นว่า วิธีการที่นำเสนอมีประสิทธิภาพด้านเวลาสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 96 ประสิทธิภาพด้านเวลาที่ใช้ในการคัดเลือกคุณลักษณะ ข้อมูลชุด SS-Tweet

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	0.386	0.412	0.431	0.456	0.471	0.483	0.504	0.528	0.543	0.468
Chi2	0.386	0.411	0.430	0.453	0.470	0.490	0.505	0.519	0.541	0.467
IG	0.395	0.419	0.445	0.462	0.484	0.496	0.516	0.532	0.552	0.478
Proposed	0.317	0.301	0.322	0.340	0.354	0.372	0.389	0.410	0.426	0.359

ตาราง 97 ผลการวิเคราะห์ Paired-Sample t-test *ประสิทธิภาพด้านเวลาของชุดข้อมูล* SS-Tweet

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.109	0.015	0.005	-0.121	-0.098	-21.688	8	0.000
Pair 2 Proposed-Chi2	-0.108	0.015	0.005	-0.120	-0.097	-21.641	8	0.000
Pair 3 Proposed-IG	-0.119	0.016	0.005	-0.131	-0.107	-22.956	8	0.000

จากตาราง 96 แสดงประสิทธิภาพด้านเวลา เมื่อทดสอบกับชุดข้อมูล SS-Tweet พบว่า วิธีการที่นำเสนอใช้ระยะเวลาในการคัดเลือกคุณลักษณะน้อยที่สุด และเมื่อพิจารณาผลการวิเคราะห์ Paired-Sample t-test ในตาราง 97 แสดงให้เห็นว่า วิธีการที่นำเสนอมีประสิทธิภาพด้านเวลาสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 98 ประสิทธิภาพด้านเวลาที่ใช้ในการคัดเลือกคุณลักษณะ ข้อมูลชุด HCR

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	0.089	0.060	0.063	0.066	0.069	0.065	0.066	0.067	0.095	0.071
Chi2	0.096	0.088	0.092	0.099	0.097	0.098	0.099	0.099	0.101	0.097
IG	0.076	0.078	0.075	0.066	0.066	0.068	0.069	0.068	0.067	0.070
Proposed	0.044	0.044	0.049	0.052	0.055	0.055	0.060	0.061	0.060	0.053

ตาราง 99 ผลการวิเคราะห์ Paired-Sample t-test *ประสิทธิภาพด้านเวลาของชุดข้อมูล* HCR

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.018	0.013	0.004	-0.028	-0.007	-3.969	8	0.004
Pair 2 Proposed-Chi2	-0.043	0.004	0.001	-0.047	-0.040	-31.595	8	0.000
Pair 3 Proposed-IG	-0.017	0.011	0.004	-0.025	-0.009	-4.761	8	0.001

จากตาราง 98 แสดงประสิทธิภาพด้านเวลา เมื่อทดสอบกับชุดข้อมูล HCR พบว่า วิธีการที่นำเสนอใช้ระยะเวลาในการคัดเลือกคุณลักษณะน้อยที่สุด และเมื่อพิจารณาผลการวิเคราะห์ Paired-Sample t-test ในตาราง 99 แสดงให้เห็นว่า วิธีการที่นำเสนอมีประสิทธิภาพด้านเวลาสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

ตาราง 100 ประสิทธิภาพด้านเวลาที่ใช้ในการคัดเลือกคุณลักษณะ ข้อมูลชุด Sander

วิธีการ	จำนวนคุณลักษณะที่สำคัญที่สุด									
	10%	20%	30%	40%	50%	60%	70%	80%	90%	ค่าเฉลี่ย
GINI	0.047	0.049	0.049	0.050	0.051	0.057	0.054	0.055	0.058	0.052
Chi2	0.061	0.058	0.071	0.068	0.068	0.069	0.070	0.071	0.072	0.068
IG	0.047	0.048	0.048	0.050	0.072	0.052	0.049	0.052	0.052	0.052
Proposed	0.038	0.040	0.042	0.048	0.042	0.043	0.045	0.044	0.047	0.043

ตาราง 101 *ผลการวิเคราะห์* Paired-Sample t-test *ประสิทธิภาพด้านเวลาของชุดข้อมูล* Sanders

	Paired Differences					t	df	Sig (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Of the Difference				
				Lower	Upper			
Pair 1 Proposed-GINI	-0.009	0.003	0.001	-0.012	-0.007	-8.458	8	0.000
Pair 2 Proposed-Chi2	-0.024	0.004	0.001	-0.027	-0.022	-20.773	8	0.000
Pair 3 Proposed-IG	-0.009	0.008	0.003	-0.015	-0.003	-3.297	8	0.011

จากตาราง 100 แสดงประสิทธิภาพด้านเวลา เมื่อทดสอบกับชุดข้อมูล Sander พบว่า วิธีการที่นำเสนอใช้ระยะเวลาในการคัดเลือกคุณลักษณะน้อยที่สุด และเมื่อพิจารณาผลการวิเคราะห์ Paired-Sample t-test ในตาราง 101 แสดงให้เห็นว่า วิธีการที่นำเสนอมีประสิทธิภาพด้านเวลาสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

สรุปผลประสิทธิภาพด้านเวลาที่ใช้ในการคัดเลือกคุณลักษณะ โดยภาพรวม พบว่า วิธีการที่นำเสนอมีประสิทธิภาพด้านเวลาสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05

4.3.3 ผลการประเมิน Big-O

การวัดประสิทธิภาพของอัลกอรึทึมในรูปแบบของเวลา อาจจะมีปัญหากรณีที่รันโปรแกรมในเครื่องคอมพิวเตอร์ที่มีคุณสมบัติไม่เท่ากัน ผู้วิจัยจึงทำการคำนวณหา Big-O เพื่อวัดประสิทธิภาพของอัลกอริทึม ผลการประเมิน Big-O ของแต่ละวิธีการ มีผลดังนี้

1) การประเมิน Big-O ของวิธีการ GINI

Algorithm 3: Gini Index Feature Ranking

1. for each class $c_k \in C$ do:
2. for document $d_j \in D$ do:
3. if d_j in c_k do:
4. $n_k \mathrel{+}= 1$;
5. for each term $t_i \in T$ do:
6. if t_i in c_k do:
7. $n(t_i, c_k) \mathrel{+}= 1$;
8. end for
9. end for
10. $p_k = n_k \,/\, N$
11. $Gini(D) = 1 - \sum p_k{}^2$;
12. $Gini(t_i, D_j) = 1 - [n(t_i, c_k) / \sum n(t_i, c_k)]$
13. end for

รูปที่ 34 Gini Index Algorithm

จากรูปที่ 34 พบว่า บรรทัดที่ 2 – 9 เป็นการหาค่าความถี่ของแต่ละคุณลักษณะที่อยู่ในคลาส ซึ่ง Big-O = O(|D|.|T|) โดยที่ |D| คือ จำนวนเอกสาร และ |T| คือ จำนวนคุณลักษณะ ซึ่งต้องหาคุณลักษณะในแต่ละคลาส ถ้าจำนวนคลาส คือ |C| จะได้ Big-O = O(|C|.|D|.|T|) บรรทัดที่ 10 คำนวณค่าความน่าจะเป็น ซึ่งใช้ Big-O = O(1) แต่ต้องคำนวณทุกคลาส ดังนั้น Big-O = O(1.|C|) บรรทัดที่ 11 คำนวณค่า Gini ของเอกสาร ซึ่งใช้ Big-O = O(1) แต่ต้องคำนวณทุกคลาส ดังนั้น Big-O = O(1.|C|) บรรทัดที่ 12 คำนวณค่า Gini Split ซึ่งใช้ Big-O = O(1) แต่ต้องคำนวณทุกคลาส ดังนั้น Big-O = O(1.|C|) เมื่อพิจารณาภาพรวมทั้งหมด จะได้ว่า Big-O ของวิธีการ Gini Index เท่ากับ O(|C|.|D|.|T|) + O(1.|C|) + O(1.|C|) + O(1.|C|) ดังนั้น จึงสรุปได้ว่า Big-O = O(|C|.|D|.|T|)

2) ผลการประเมิน Big-O ของวิธีการ Chi2

Algorithm 2: Chi-Square Feature Ranking

1. for each class $c_k \in C$ do:
2. for document $d_j \in D$ do:
3. if d_j in c_k do:
4. $n_k \mathrel{+}= 1$;
5. for each term $t_i \in T$ do:
6. if t_i in c_k do:
7. $Observed(t_i, c_k) \mathrel{+}= 1$;
8. end for
9. end for
10. $Chi^2(t_i, c_k) = ((Observed(t_i, c_k) - Expected(t_i, c_k))^2) / Expected(t_i, c_k)$;
11. $prob(c_k) = n_2 / (n_1 + n_2)$;
12. end for
13. for each class $t_i \in T$ do:
14. for each term $c_k \in C$ do:
15. $Expected(t_i, c_k) = \Pr ob(c_k) \times Observed(t_i, c_k)$;
16. $Chi^2(t_i, c_k) = ((Observed(t_i, c_k) - Expected(t_i, c_k))^2) / Expected(t_i, c_k)$;
17. end for
18. $Chi^2(t_i) = Chi^2(t_i, c_1) + Chi^2(t_i, c_2)$;
19. end for

รูปที่ 35 Chi-Square Algorithm

จากรูปที่ 35 พบว่า บรรทัดที่ 1 – 9 เป็นการหาค่าความถี่ของแต่ละคุณลักษณะที่อยู่ในคลาส ซึ่ง Big-O = O(|D|.|T|) โดยที่ |D| คือ จำนวนเอกสาร และ |T| คือ จำนวนคุณลักษณะ ซึ่งต้องหาคุณลักษณะในแต่ละคลาส ถ้าจำนวนคลาส คือ |C| จะได้ Big-O = O(|C|.|D|.|T|) บรรทัดที่ 10 คำนวณค่าไคสแคว์ของคุณลักษณะ ซึ่งใช้ Big-O = O(|T|) แต่ต้องคำนวณทุกคลาส ดังนั้น Big-O = O(|T|.|C|) บรรทัดที่ 11 คำนวณค่าความน่าจะเป็น ซึ่งใช้ Big-O = O(1) แต่ต้องคำนวณทุกคลาส ดังนั้น Big-O = O(1.|C|) บรรทัดที่ 13 – 19 เป็นการ Expected ของคุณลักษณะ ซึ่งใช้ Big-O = O(|T|) แต่ต้องคำนวณทุกคลาส ดังนั้น Big-O = O(|T|.|C|) เมื่อพิจารณาภาพรวมทั้งหมด จะได้ว่า Big-O ของวิธีการ Chi-Square เท่ากับ O(|C|.|D|.|T|) + O(|T|.|C|) + O(1.|C|) + O(|T|.|C|) ดังนั้น จึงสรุปได้ว่า Big-O = O(|C|.|D|.|T|)

3) ผลการประเมิน Big-O ของวิธีการ Information Gain

Algorithm 1: Information Gain Feature Ranking

1. S = 0
2. For each $c_k \in C$ do:
3. calculate $p(c_k)$;
4. $H_c = S + p(c_k) \times \log 2(p(c_k))$
5. $S \leftarrow H_c$
6. End For
7. For each $e_i \in E$
8. Calculate $p(e_i)$;
9. $Sum = S + P(e_i) \times \log 2(p(e_i))$;
10. $S \leftarrow Sum$;
11. End For
12. For each class $c_k \in C$ do:
13. For each term $e_i \in E$ do:
14. Calculate $p(c_k \mid e_i)$;
15. $M = S + p(c_k \mid e_i) \times \log_2 p(c_k \mid e_i)$;
16. $S \leftarrow M$;
17. End For
18. End For
19. $H(C \mid E) = (-1) \times Sum \times (-1) \times M$;
20. $IG = H_c - H(C \mid E)$

รูปที่ 36 Information Gain Algorithm

จากบรรทัดที่ 1 ถึง 6 ใช้ Big-O = O(|C|) ซึ่ง |C| คือ จำนวนคลาส จากบรรทัดที่ 7 ถึง 11 ใช้ Big-O = O(|E|) ซึ่ง |E| คือ ค่าที่เป็นไปได้ทั้งหมดของคุณลักษณะ ซึ่งงานวิจัยนี้มีค่าน้ำหนักของคุณลักษณะ 2 ค่า คือ 0 และ 1 จากบรรทัดที่ 12 ถึง 18 ใช้ Big-O = O(|E|.|C|) จากอัลกอริทึมข้างต้น จะได้ค่า Information Gain ของคุณลักษณะ จำนวน 1 คุณลักษณะ เมื่อต้องการหาค่า Information Gain ของคุณลักษณะทั้งหมด จะใช้ Big-O = O(|E|.|C|.|T|)

4) ผลการประเมิน Big-O ของวิธีการที่นำเสนอ

Algorithm 1: Proposed Method

1. for each class $c_k \in C$
2. for each features $t_i \in T$ do:
3. $|S(t_i, c_k)| = |S(t_i) \cap S(c_k)|$
4. $Con(t_i, c_k) = \frac{|S(t_i, c_k)|}{|S(t_i)|}$
5. end for
6. $sort(T)$ in ascending order of support
7. for each features $t_i \in sort(T)$ do:
8. $PS(t_i, c_k) = \frac{R(t_i, c_k)}{|D|}$
9. $w(t_i c_k) = p \times PS(t_i, c_k) + (1-p) \times Con(t_i, c_k)$
10. $M(t_i) = \max(w(t_i, c_k))$
11. end for
12. end for

รูปที่ 37 Proposed Algorithm

จากรูปที่ 37 บรรทัดที่ 1 – บรรทัดที 5 เป็นการหาค่าสนับสนุน และค่าความเชื่อมั่นโดยข้อมูลนำเข้าเป็นข้อมูลรูปแบบแนวตั้ง ค่าสนับสนุนคำนวณด้วยวิธีการ หาค่าจำนวนของผลการอินเตอร์เซคชั่นระหว่างเซตของคุณลักษณะและเซตของคลาส และค่าความเชื่อมั่นคำนวณจากค่าสนับสนุนของคุณลักษณะ ใช้ Big-O = O(|C|.|T|) จากนั้นบรรทัดที่ 6 เป็นการจัดลำดับคุณลักษณะโดยเรียงค่าสนับสนุนจากมากไปน้อย ใช้ Big-O = O(|T|log|T|) และจะต้องทำการจัดลำดับทุกคลาส ดังนั้น Big-O = O(|C|.|T|.log|T|) จากนั้นบรรทัดที่ 7 – 11 เป็นการคำนวณค่าน้ำหนักของคุณลักษณะที่ทำการจัดลำดับแล้ว ใช้ Big-O = O(|T|) เมื่อพิจารณาภาพรวมทั้งหมด จะได้ว่า Big-O ของวิธีการที่นำเสนอ เท่ากับ O(|C|.|T|) + O(|C|.|T|.log|T|) + O(|T|) ดังนั้น จึงสรุปได้ว่า Big-O = O(|C|.|T|.log|T|)

สรุปโดยภาพรวม พบว่า วิธีการที่นำเสนอมีประสิทธิภาพ Big-O ดีกว่าวิธีการอื่น

4.3.3 ผลการวัดประสิทธิภาพการลดคุณลักษณะที่ซ้ำซ้อน

การทดลองนี้เพื่อเปรียบเทียบการใช้คุณลักษณะทั้งหมดกับการลดคุณลักษณะที่ซ้ำซ้อนด้วยวิธีการที่นำเสนอ การทดลองนี้ทดสอบกับข้อมูล จำนวน 5 ชุดข้อมูล ได้แก่ 1) Stadford Twitter Sentiment Data (STS) 2) SemEval-2017 Task4A Dataset (SemEval) 3) Sentiment Strength Twitter Dataset (SS-Tweet) 4) 2) Health Care Reform (HCR) 5) Sanders Twitter Dataset (Sander) โดยใช้ตัวจำแนก คือ นาอีฟเบย์ รายละเอียดผลการวัดประสิทธิภาพการทดลอง มีดังนี้

ตาราง 102 ผลการทดสอบจำนวนคุณลักษณะที่ซ้ำซ้อน

ชุดข้อมูล	จำนวนคุณลักษณะทั้งหมด	จำนวนคุณลักษณะที่ซ้ำซ้อน	ร้อยละคุณลักษณะที่ซ้ำซ้อน
STS	12,772	3,540	27.72
SemEval	9,065	3,375	37.23
SS-Tweet	6,845	2,932	42.83
HCR	2,503	1,087	43.43
Sander	1,867	763	40.87

จากตาราง 57 แสดงผลการทดสอบจำนวนคุณลักษณะที่ซ้ำซ้อน พบว่า ข้อมูลชุด Health Care Reform (HCR) มีจำนวนคุณลักษณะมากที่สุด รองลงมาคือ ข้อมูลชุด Sentiment Strength Twitter Dataset (SS-Tweet) ข้อมูลชุด Sanders Twitter Dataset (Sander) ข้อมูลชุด SemEval-2017 Task4A Dataset (SemEval) และ ข้อมูลชุด Stadford Twitter Sentiment Data (STS) ตามลำดับ

ตาราง 103 ผลการวัดประสิทธิภาพความถูกต้อง ก่อนและหลังการลดคุณลักษณะซ้ำซ้อน

ชุดข้อมูล	ความถูกต้องก่อนลดคุณลักษณะซ้ำซ้อน	ความถูกต้องหลังลดคุณลักษณะซ้ำซ้อน	ค่าการเปลี่ยนแปลง (+/-)
STS	72.25	72.31	**+0.06**
SemEval	87.52	87.53	**+0.01**
SS-Tweet	60.69	61.85	**+1.16**
HCR	81.40	81.30	-0.10
Sander	89.20	88.20	-1.00

จากตาราง 58 แสดงผลการวัดประสิทธิภาพความถูกต้องของการจำแนกความคิดเห็น ก่อนและหลังคุณลักษณะที่ซ้ำซ้อน พบว่า ข้อมูลชุด Stadford Twitter Sentiment Data (STS) ข้อมูลชุด SemEval-2017 Task4A Dataset (SemEval) และ ข้อมูลชุด Sentiment Strength Twitter Dataset (SS-Tweet) มีประสิทธิภาพเพิ่มขึ้น เมื่อทำการลดคุณลักษณะที่ซ้ำซ้อน ส่ข้อมูลชุด Health Care Reform (HCR) และ ข้อมูลชุด Sanders Twitter Dataset (Sander) พบว่ามีประสิทธิภาพลดลงเล็กน้อย เมื่อลดคุณลักษณะที่ซ้ำซ้อน

ตาราง 104 ผลการวัดประสิทธิภาพความแม่นยำ ก่อนและหลังการลดคุณลักษณะซ้ำซ้อน

ชุดข้อมูล	ความแม่นยำในคลาสเชิงบวก		ค่าการเปลี่ยนแปลง (+/-)	ความแม่นยำในคลาสเชิงลบ		ค่าการเปลี่ยนแปลง (+/-)
	ก่อน	หลัง		ก่อน	หลัง	
STS	71.09	71.07	-0.02	73.56	73.75	**+0.18**
SemEval	87.88	88.15	**+0.26**	87.22	86.98	-0.24
SS-Tweet	61.41	60.75	-0.66	63.08	63.56	**+0.48**
HCR	79.30	79.37	**+0.08**	84.05	83.59	-0.46
Sander	87.98	88.03	**+0.04**	90.75	88.41	-2.34

จากตาราง 59 แสดงผลการวัดประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็น ก่อนและหลังคุณลักษณะที่ซ้ำซ้อน พบว่า หลังการลดคุณลักษณะซ้ำซ้อน ข้อมูลชุด SemEval-2017 Task4A Dataset (SemEval) ข้อมูลชุด Health Care Reform (HCR) และ ข้อมูลชุด Sanders Twitter Dataset (Sander) มีประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นในเชิงบวกเพิ่มขึ้น ข้อมูลชุด Stadford Twitter Sentiment Data (STS) และ ข้อมูลชุด Sentiment Strength Twitter Dataset (SS-Tweet) มีประสิทธิภาพความแม่นยำของการจำแนกความคิดเห็นในคลาสเชิงลบเพิ่มขึ้น

ตาราง 105 ผลการวัดประสิทธิภาพค่าระลึก ก่อนและหลังการลดคุณลักษณะซ้ำซ้อน

ชุดข้อมูล	ค่าระลึกในคลาสเชิงบวก		ค่าการเปลี่ยนแปลง (+/-)	ค่าระลึกในคลาสเชิงลบ		ค่าการเปลี่ยนแปลง (+/-)
	ก่อน	หลัง		ก่อน	หลัง	
STS	75.04	75.33	**+0.29**	69.49	69.35	-0.15
SemEval	87.06	86.70	-0.37	87.90	88.26	**+0.36**
SS-Tweet	65.84	68.31	**+2.46**	58.40	55.57	-2.82
HCR	85.82	85.27	-0.55	76.79	77.42	**+0.63**
Sander	91.38	88.40	-2.98	86.94	87.93	**+0.98**

จากตาราง 60 แสดงผลการวัดประสิทธิภาพค่าระลึกของการจำแนกความคิดเห็น ก่อนและหลังคุณลักษณะที่ซ้ำซ้อน พบว่า ข้อมูลชุด Stadford Twitter Sentiment Data (STS) และ ข้อมูลชุด Sentiment Strength Twitter Dataset (SS-Tweet) มีค่าระลึกของการจำแนกความคิดเห็นในคลาสเชิงบวกเพิ่มขึ้น ข้อมูลชุด SemEval-2017 Task4A Dataset (SemEval) ข้อมูลชุด Health Care Reform (HCR) และ ข้อมูลชุด Sanders Twitter Dataset (Sander) มีค่าระลึกของการจำแนกความคิดเห็นในคลาสเชิงลบเพิ่มขึ้น

ตาราง 106 ผลการวัดประสิทธิภาพโดยรวม ก่อนและหลังการลดคุณลักษณะซ้ำซ้อน

ชุดข้อมูล	ประสิทธิภาพโดยรวมในคลาสเชิงบวก		ค่าการเปลี่ยนแปลง	ประสิทธิภาพโดยรวมในคลาสเชิงลบ		ค่าการเปลี่ยนแปลง
	ก่อน	หลัง	(+/-)	ก่อน	หลัง	(+/-)
STS	72.98	73.11	**+0.13**	71.44	71.45	**+0.01**
SemEval	87.43	87.38	-0.05	87.52	87.58	**+0.06**
SS-Tweet	63.28	64.17	**+0.89**	57.59	59.11	**+1.52**
HCR	82.22	81.99	-0.23	79.99	80.15	**+0.16**
Sander	89.52	88.10	-1.42	88.65	88.06	-0.59

จากตาราง 61 แสดงผลการวัดประสิทธิภาพประสิทธิภาพโดยรวมของการจำแนกความคิดเห็น ก่อนและหลังคุณลักษณะที่ซ้ำซ้อน พบว่า หลังจากทำการลดคุณลักษณะที่ซ้ำซ้อน ข้อมูลชุด Stadford Twitter Sentiment Data (STS) และ ข้อมูลชุด Sentiment Strength Twitter Dataset (SS-Tweet) มีประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นในคลาสเชิงบวกและคลาสเชิงลบเพิ่มขึ้น ข้อมูลชุด SemEval-2017 Task4A Dataset (SemEval) และข้อมูลชุด Health Care Reform (HCR) มีประสิทธิภาพโดยรวมของการจำแนกความคิดเห็นในคลาสเชิงลบเพิ่มขึ้น

ตาราง 107 ประสิทธิภาพด้านเวลา (วินาที) ในการการสร้างโมเดลและการทดสอบโมเดลก่อนและหลังลดคุณลักษณะซ้ำซ้อน

ชุดข้อมูล	ระยะเวลาก่อนลดคุณลักษณะซ้ำซ้อน	ระยะเวลาหลังลดคุณลักษณะซ้ำซ้อน	ค่าการเปลี่ยนแปลง (+/-)
STS	0.662	0.459	-0.203
SemEval	0.202	0.116	-0.087
SS-Tweet	0.105	0.053	-0.052
HCR	0.018	0.010	-0.008
Sander	0.014	0.008	-0.007

จากตาราง 62 แสดงประสิทธิภาพด้านเวลาในการสร้างโมเดลและการทดสอบโมเดล ก่อนและหลังการลดคุณลักษณะที่ซ้ำซ้อน พบว่า การลดคุณลักษณะที่ซ้ำซ้อนช่วยให้เวลาในการสร้างโมเดลและทดสอบโมเดลลดลง

ตาราง 108 ระยะเวลา (วินาที) ที่ใช้ในการลดคุณลักษณะซ้ำซ้อน

ชุดข้อมูล	จำนวนข้อมูล	จำนวนคุณลักษณะ	ระยะเวลาที่ใช้
STS	10,000	12,772	30.274
SemEval	4,000	9,065	7.968
SS-Tweet	2,600	6,845	3.830
HCR	1,000	2,503	0.448
Sander	1,000	1,867	0.315

จากตาราง 63 แสดงประสิทธิภาพด้านเวลาที่ใช้ในการสร้างลดคุณลักษณะที่ซ้ำซ้อน พบว่า การลดคุณลักษณะที่ซ้ำซ้อนใช้ระยะเวลาเพิ่มขึ้นตามขนาดของข้อมูล ข้อมูลขนาดใหญ่ มีคุณลักษณะจำนวนมาก จะใช้เวลาในการตรวจสอบและลดคุณลักษณะซ้ำซ้อนเพิ่มขึ้น

บทที่ 5
สรุปผลการวิจัย

งานวิจัยนี้ ประกอบด้วย 2 ส่วนหลัก คือ 1) การพัฒนาขั้นตอนวิธีในการคัดเลือกคุณลักษณะ และ 2) การขจัดคุณลักษณะที่ซ้ำซ้อนสำหรับการจำแนกความคิดเห็นที่อยู่บนเครือข่ายสังคมออนไลน์ ซึ่งการในการคัดเลือกคุณลักษณะอาศัยหลักการผสมผสานแนวคิดวิธีฟิลเตอร์โมเดลร่วมกับแนวคิดวิธีการกฎความสัมพันธ์ โดยนำค่าสนับสนุนและค่าความเชื่อมั่นมาพิจารณาร่วมกันเพื่อให้ค่าน้ำหนักของคุณลักษณะ โดยสามารถปรับค่าพารามิเตอร์ที่เรียกว่า p เพื่อถ่วงน้ำหนักระหว่างค่าสนับสนุนและค่าความเชื่อมั่น โดยทำการทดลองกับชุดข้อมูลที่เป็นข้อความบนเครือข่ายสังคมออนไลน์ทั้งหมด 5 ชุดข้อมูล และพิจารณาประสิทธิภาพในการจำแนกและเวลาที่ใช้ในการคัดเลือกคุณลักษณะ ในส่วนของการขจัดคุณลักษณะที่ซ้ำซ้อน พิจารณาจากคุณลักษณะที่เกิดร่วมกันในเอกสารเดียวกัน แล้วทำการเลือกคุณลักษณะที่มีค่าน้ำหนักสูงสุดและตัดคุณลักษณะที่เหลือออก โดยทำการทดลองกับชุดข้อมูลที่เป็นข้อความบนเครือข่ายสังคมออนไลน์ทั้งหมด 5 ชุดข้อมูล และพิจารณาประสิทธิภาพในการจำแนกก่อนและหลังจากขจัดคุณลักษณะที่ซ้ำซ้อน จำนวนคุณลักษณะก่อนและหลังจากขจัดคุณลักษณะที่ซ้ำซ้อน เวลาที่ใช้ในการสร้างตัวจำแนกก่อนและหลักจากขจัดคุณลักษณะที่ซ้ำซ้อน และเวลาที่ใช้ในการขจัดคุณลักษณะที่ซ้ำซ้อน ผลการวิจัยสามารถสรุปผล อภิปรายผล และข้อเสนอแนะดังนี้

5.1 สรุปผลและอภิปราย

1. วิธีการที่นำเสนอให้ประสิทธิภาพในการจำแนกสูงเมื่อข้อมูลมีขนาดใหญ่ และให้ค่าความถูกต้องสูงกว่าวิธีการอื่น อย่างมีนัยสำคัญที่ 0.05 เมื่อเปรียบเทียบกับขั้นตอนวิธีการคัดเลือกคุณลักษณะแบบฟิลเตอร์โมเดล 3 วิธี ได้แก่ วิธีการ Information Gain วิธีการ Chi-Square วิธีการ Gini Index เนื่องจากวิธีการที่นำเสนอใช้การปรับค่าอัตราส่วนของน้ำหนักของค่าสนับสนุนกับค่าความเชื่อมั่นให้เท่ากัน จากนั้นจึงทำการปรับค่าถ่วงน้ำหนักให้กับค่าสนับสนุนและค่าความเชื่อมั่น โดยการปรับค่าถ่วงน้ำหนักจะทดสอบกับข้อมูลในแต่ละชุด ซึ่งจากการทดลองพบว่าการถ่วงน้ำหนักให้ค่าสนับสนุนมากกว่าค่าความเชื่อมั่นช่วยให้การจำแนกความคิดเห็นประสิทธิภาพมากขึ้น

2. วิธีการที่นำเสนอใช้เวลาในการคัดเลือกคุณลักษณะน้อยที่สุด เมื่อเปรียบเทียบกับขั้นตอนวิธีการคัดเลือกคุณลักษณะ 3 วิธีการข้างต้น เนื่องจากวิธีการที่นำเสนอเป็นวิธีการที่ง่าย โดยใช้หลักการคำนวณค่าน้ำหนักโดยพิจารณาจากค่าสนับสนุนร่วมกับค่าความเชื่อมั่น

3. จากการทดลองปรับพารามิเตอร์ (p) กับข้อมูลทั้ง 5 ชุด พบว่า ค่าที่เหมาะสมที่ทำให้ประสิทธิภาพการจำแนกความคิดเห็นสูงสุด คือ ค่า p = 0.8 ซึ่งเป็นการถ่วงน้ำหนักไปที่ค่าสนับสนุน

มากกว่าค่าความเชื่อมั่น แสดงให้เห็นว่า เมื่อข้อมูลมีขนาดใหญ่ค่าสนับสนุนมีความสำคัญมากกว่าค่าความเชื่อมั่น

4. การขจัดคุณลักษณะที่ซ้ำซ้อนด้วยวิธีการที่นำเสนอ ทำให้จำนวนคุณลักษณะลดลง แต่ไม่ได้ลดประสิทธิภาพในการจำแนกลง เนื่องจากมีบางข้อความบนเครือข่ายสังคมออนไลน์นิยมใช้คำคู่กัน ซึ่งผู้วิจัยได้วิเคราะห์แล้วว่าข้อความที่ใช้คู่กันนั้น สามารถใช้เป็นคุณลักษณะสำหรับการจำแนกความคิดเห็นแทนกันได้

5.2 ข้อเสนอแนะ

1. ขั้นตอนวิธีการคัดเลือกคุณลักษณะที่ได้นำเสนอ เป็นขั้นตอนวิธีการที่มีประสิทธิภาพเมื่อทดสอบกับข้อมูลบนเครือข่ายสังคมออนไลน์ขนาดใหญ่ ดังนั้น ผู้วิจัยเสนอแนะว่าควรนำแนวคิดวิธีการที่นำเสนอไปทดสอบกับข้อมูลที่หลากหลาย เช่น การจำแนกบทความวิจัย การวิเคราะห์ข้อความคิดเห็นที่เป็นบทวิจารณ์ เป็นต้น ซึ่งข้อความดังกล่าวเป็นข้อความที่มีความยาวมากกว่าข้อความบนเครือข่ายสังคมออนไลน์

2. ประสิทธิภาพของการจำแนกความคิดเห็น ขึ้นอยู่กับการปรับค่าพารามิเตอร์ (p) ในสมการหาค่าน้ำหนักของคุณลักษณะที่นำเสนอ ซึ่งในงานวิจัยนี้ได้เลือกค่าพารามิเตอร์ที่ทดสอบกับข้อมูลแต่ละชุด โดยทดลองปรับค่าพารามิเตอร์ไปเรื่อย ๆ ตั้งแต่ 0.1 ถึง 0.9 แล้วเลือกพารามิเตอร์ที่ให้ค่าประสิทธิภาพสูงสุด ดังนั้น ถ้ามีแนวคิดการปรับค่าพารามิเตอร์อัตโนมัติตามชุดข้อมูล จะช่วยให้ลดขั้นตอนการทดลองที่ใช้เลือกพารามิเตอร์ได้

3. ข้อความที่ใช้ทดสอบในงานวิจัยนี้เป็นข้อความภาษาอังกฤษ ควรนำแนวคิดขั้นตอนวิธีการที่นำเสนอไปปรับใช้กับข้อมูลที่เป็นข้อความภาษาไทย

4. ควรนำหลักการโครงสร้างข้อมูล เช่น โครงสร้างต้นไม้ (Decision Tree) การหาเซตผลต่าง (Diff Set) มาพัฒนาขั้นตอนการขจัดคุณลักษณะที่ซ้ำซ้อน เพื่อลดเวลาในการขจัดคุณลักษณะที่ซ้ำซ้อน

บรรณานุกรม

บรรณานุกรม

[1] Troussas C, Virvou M, Junshean Espinosa K, Llaguno K, Caro J. Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. Information, Intelligence, Systems and Applications (IISA), 2013 Fourth International Conference on; 10-12 July 2013; 1-6.

[2] Akaichi J, Dhouioui Z, Lopez-Huertas Perez MJ. Text mining facebook status updates for sentiment classification. System Theory, Control and Computing (ICSTCC), 2013 17th International Conference; 11-13 Oct. 2013; 640-645.

[3] Anjaria M, Guddeti RMR. Influence factor based opinion mining of Twitter data using supervised learning. Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on; 6-10 Jan. 2014; 1-8.

[4] Ortigosa A, Martín JM, Carro RM. Sentiment analysis in Facebook and its application to e-learning. Computers in Human Behavior 2014; 31[0]: 527-541.

[5] Mostafa Karamibekr AAG. Sentiment Analysis of Social Issues. 2012 International Conference on Social Informatics; Canada. IEEE; 215-221.

[6] Yang J, Liu Y, Zhu X, Liu Z, Zhang X. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Inf Process Manage 2012; 48741-754.

[7] Song Q, Ni J, Wang G. A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data. IEEE Trans on Knowl and Data Eng 2013; 25[1]: 1-14.

[8] Yang J, Liu Y, Zhu X, Liu Z, Zhang X. A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Inf Process Manage 2012; 48[4]: 741-754.

[9] Das S. Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. Morgan Kaufmann Publishers Inc., 2001.

[10] Deng X, Li Y, Weng J, Zhang J. Feature selection for text classification: A review. Multimedia Tools and Applications [journal article]2019; 78[3]: 3797-3816. https://doi.org/10.1007/s11042-018-6083-5

[11] Go A, Bhayani R, Huang L. Twitter sentiment classification using distant supervision. 2009;

[12] Rosenthal S, Farra N, Nakov P. SemEval-2017 Task 4: Sentiment Analysis in Twitter. Association for Computational Linguistics; 502-518.

[13] Thelwall M, Buckley K, Paltoglou G. Sentiment strength detection for the social web. J Am Soc Inf Sci Technol 2012; 63[1]: 163-173.

[14] Speriosu M, Sudan N, Upadhyay S, Baldridge J. Twitter polarity classification with label propagation over lexical links and the follower graph. Association for Computational Linguistics, 2011.

[15] Saif H, Fernández M, He Y, Alani H. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. 2013.

[16] Hall JC. A Linguistic Model for Improving Sentiment Analysis Systems. Master of ScienceThesis. Fargo, North Dakota: North Dakota State University; 2014.

[17] Liu B. Sentiment Analysis and Opinion Mining. Morgan \& Claypool Publishers; 2012.

[18] Kumar AA, S.Chandrasekhar. Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering. International Journal of Engineering Research & Technology July 2012; 1[5]: 1-6.

[19] C.Ramasubramanian, R.Ramya. Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm. International Journal of Advanced Research in Computer and Communication Engineering December 2013; 2[12]: 4536-4537.

[20] Blessy Selvam SA. A Survey on Opinion Mining Framework. International Journal of Advanced Research in Computer and Communication Engineering September 2013; 2[9]: 3544-3549.

[21] Mumu T. Social Network Opinion and Posts Mining for Community Preference Discovery. Master's Thesis: University of Windsor; 2013.

[22] Tripathy A, Agrawal A, Rath S. Classification of Sentiment Reviews using N-gram Machine Learning Approach. 2016.

[23] Louwerse M, Lewis G, Wu J. Unigrams, bigrams and LSA. Corpus linguistic explorations of genres in Shakespeare's plays. 2008:108-129.

[24] Read J. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. Proceedings of the ACL Student Research Workshop; Association for Computational Linguistics; 43-48.

[25] Zhang Z. Urcf: an approach to integrating user reviews into memory-based collaborative filtering. Ph.D. Dissertation: University of Maryland at Baltimore County; 2013.

[26] ตั้งวงศ์ศานต์ ศ. ระบบการจัดเก็บและการสืบค้นสารสนเทศด้วยคอมพิวเตอร์. พิมพ์ครั้งที่ 1. กรุงเทพฯ: โรงพิมพ์พิทักษ์การพิมพ์; 2551.

[27] กานดา แผ่วัฒนากุล. การวิเคราะห์เหมืองข้อเสนอแนะจากบทวิจารณ์รายการโทรทัศน์. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต. กรุงเทพฯ: สถาบันบัณฑิตพัฒนบริหารศาสตร์; 2555.

[28] Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies 2012; 5[1]: 1-167.

[29] เอกสิทธิ์ พัชรวงศ์ศักดา. การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิง เบื้องต้น. พิมพ์ครั้งที่ 1. กรุงเทพฯ: เอเชีย ดิจิตอลการพิมพ์; 2557.

[30] Phua YL. Social Media Sentiment Analysis and Topic Detection for Singapore English. Master's Thesis: Naval Postgraduate School; 2013.

[31] Singh PK, Husain MS. Methodological Study of Opinion Mining and Sentiment Analysis Techniques. International Journal on Soft Computing 2014; 5[1]: 11-21.

[32] Basari ASH, Hussin B, Ananta IGP, Zeniarja J. Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. Procedia Engineering 2013; 53[0]: 453-462.

[33] ผุสดี บุญรอด. การศึกษาปัจจัยที่มีผลต่อการย่อความภาษาไทย และการพัฒนาเทคนิคการย่อความภาษาไทยโดยใช้การประมวลผลธรรมชาติร่วมกับฐานความรู้ออนโทโลยี. วิทยานิพนธ์ปรัชญาดุษฎีบัณฑิต. กรุงเทพฯ: มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ; 2552.

[34] Chairawichitchai N. Automatic Thai Document Classification Model. The Journal of Industrial Technology January -- April 2013; 9[1]:

[35] Zheng Z, Wu X, Srihari R. Feature selection for text categorization on imbalanced data. SIGKDD Explor Newsl 2004; 6[1]: 80-89.

[36] Lee C, Lee GG. Information gain and divergence-based feature selection for machine learning-based text categorization. Information Processing & Management 2006; 42[1]: 155-165.

[37] Alhaj TA, Siraj MM, Zainal A, Elshoush HT, Elhaj F. Feature Selection Using Information Gain for Improved Structural-Based Alert Correlation. PloS one 2016; 11[11]: e0166017-e0166017. https://www.ncbi.nlm.nih.gov/pubmed/27893821 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5125592/

[38] Abdul-Rahman S, Mutalib S, Khanafi NA, Ali AM. Exploring Feature Selection and Support Vector Machine in Text Categorization. Computational Science and Engineering (CSE), 2013 IEEE 16th International Conference on; 3-5 Dec. 2013; 1101-1104.

[39] Mowafy M, Rezk A, Hm E-b. An Efficient Classification Model for Unstructured Text Document.

[40] Gini C. Variabilit‡ e mutabilita. Memorie di metodologia statistica 1912;

[41] Zhu W, Feng J, Lin Y. Using Gini-Index for Feature Selection in Text Categorization. 2014/02; Atlantis Press;

[42] Hossain MR, Oo AMT, Ali ABMS. The Combined Effect of Applying Feature Selection and Parameter Optimization on Machine Learning Techniques for Solar Power Prediction. American Journal of Energy Research 2013; 1[1]: 7-16. http://pubs.sciepub.com/ajer/1/1/2

[43] Mladeni D, #263, Brank J, Grobelnik M, Milic-Frayling N. Feature selection using linear classifier weights: interaction with classification models. ACM, 2004.

[44] Hadi We, Aburub F, Alhawari S. A new fast associative classification algorithm for detecting phishing websites. Applied Soft Computing 2016; 48[Supplement C]: 729-734. http://www.sciencedirect.com/science/article/pii/S1568494616303970

[45] Saif H, He Y, Alani H. Semantic smoothing for Twitter sentiment analysis. 2011.

[46] Mohsin M, Rayhan Ahmed M, Ahmed T. IJSRSET162366 | Closed Frequent Pattern Mining Using Vertical Data Format: Depth First Approach. 2016.

[47] Mostafa MM. More than words: Social networks' text mining for consumer brand sentiments. Expert Systems with Applications 2013; 40[10]: 4241-4251.

[48] Marrese-Taylor E, Velásquez JD, Bravo-Marquez F. A novel deterministic approach for aspect-based opinion mining in tourism products reviews. Expert Systems with Applications 2014; 41[17]: 7764-7775.

[49] Swapna Somasundaran JW. Recognizing Stances in Ideological On-Line Debates. NAACL HLT 2010; June 2010; Califonia. Association for Computational Linguistics; 116-124.

[50] Cruz FL, Troyano JA, Pontes B, Ortega FJ. Building layered, multilingual sentiment lexicons at synset and lemma levels. Expert Systems with Applications 2014; 41[13]: 5984-5994.

[51] Terrana D, Augello A, Pilato G. Facebook Users Relationships Analysis Based on Sentiment Classification. Semantic Computing (ICSC), 2014 IEEE International Conference on; 16-18 June 2014; 290-296.

[52] Pak A, Paroubek P. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. LREC; 1320-1326.

[53] Yang C, Lin KH, Chen H-H. Emotion classification using web blog corpora. Web Intelligence, IEEE/WIC/ACM International Conference on; IEEE; 275-278.

[54] Mudinas A, Zhang D, Levene M. Combining lexicon and learning based approaches for concept-level sentiment analysis. ACM, 2012.

[55] Fang J, Chen B. Incorporating lexicon knowledge into svm learning to improve sentiment classification. Google Patents, 2012.

[56] Lei Zhang RG, Mohamed Dekhil, Meichun Hsu, Bing Liu, ed. Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis: Hewlett-Packard Development Company, L.P. 2011.

[57] Hamouda A, Marei M, Rohaim M. Building machine learning based senti-word lexicon for sentiment analysis. Journal of Advances in Information Technology 2011; 2[4]: 199-203.

[58] Lu B, Tsou BK. Combining a large sentiment lexicon and machine learning for subjectivity classification. Machine Learning and Cybernetics (ICMLC), 2010 International Conference on; IEEE; 3311-3316.

[59] วาทินี นุ้ยเพียร, พยุง มีสัจ. การเปรียบเทียบเทคนิคการคัดเลือกคุณลักษณะแบบการกรองและการควบรวมของการทำเหมืองข้อความเพื่อการจำแนกข้อความ. วารสารวิชาการเทคโนโลยีอุตสาหกรรม [บทความวิจัย] กันยายน - ธันวาคม 2556; 9[3]:

[60] ฐิติมา เกษมศรีธนาวัฒน์, ธนัสนี เพียรตระกูล. การจำแนกความคิดเห็นโดยใช้ตัวจำแนกแบบเบย์ร่วมกับการเลือกคุณลักษณะด้วยอัลกอริทึมรีลีฟ. CIT 2011&UniNOMS 2011; 2-Aug-2011; Bangkok, Thailand. ThaiLIS: สำนักงานคณะกรรมการการอุมศึกษา; 1-6.

[61] Sayfullina L. Reducing Sparsity in Sentiment Analysis Data using Novel Dimentionality Reduction Approaches. Aalto University; 2014.

[62] Saif H, He Y, Alani H. Alleviating data sparsity for Twitter sentiment analysis. CEUR Workshop Proceedings (CEUR-WS.org), 2012.

[63] Ong BY, Goh SW, Xu C. Sparsity adjusted information gain for feature selection in sentiment analysis. 2015 IEEE International Conference on Big Data (Big Data); Oct. 29 2015-Nov. 1 2015; 2122-2128.

[64] Parlar T, Ozel SA, Song F. QER: a new feature selection method for sentiment analysis. Human-Centric Computing and Information Sciences [Article]2018; 819. <Go to ISI>://WOS:000431867100001

[65] Pratiwi AI, Adiwijaya. On the Feature Selection and Classification Based on Information Gain for Document Sentiment Analysis. Applied Computational Intelligence and Soft Computing [Article]2018; 5. <Go to ISI>://WOS:000426701800001

[66] Ghosh M, Sanyal G. Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis. Applied Computational Intelligence and Soft Computing [Article]2018; 12. <Go to ISI>://WOS:000447518700001

ประวัติผู้เขียน

ชื่อ	นางอัจฉรา ชุมพล
วันเกิด	วันที่ 15 กรกฎาคม พ.ศ. 2522
สถานที่เกิด	อำเภอ ยางตลาด จังหวัด กาฬสินธุ์
สถานที่อยู่ปัจจุบัน	บ้านเลขที่ 4 ซอยเทศบาลอาชา 1 ถนนเทศบาลอาชา ตำบลตลาด อำเภอเมืองมหาสารคาม จังหวัดมหาสารคาม รหัสไปรษณีย์ 44000
ตำแหน่งหน้าที่การงาน	พนักงานในสถาบันอุดมศึกษา สายวิชาการ
สถานที่ทำงานปัจจุบัน	คณะวิศวกรรมศาสตร์และเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์ ถนนเกษตรสมบูรณ์ ตำบลกาฬสินธุ์ อำเภอเมือง จังหวัดกาฬสินธุ์ รหัสไปรษณีย์ 46000
ประวัติการศึกษา	พ.ศ. 2537 มัธยมศึกษาตอนต้น โรงเรียนยางตลาดวิทยาคาร อำเภอยางตลาด จังหวัดกาฬสินธุ์
	พ.ศ. 2540 มัธยมศึกษาตอนปลาย โรงเรียนกาฬสินธุ์พิทยาสรรพ์ อำเภอเมือง จังหวัดกาฬสินธุ์
	พ.ศ. 2543 ปริญญาวิทยาศาสตรบัณฑิต (วท.บ.) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม
	พ.ศ. 2549 ปริญญาวิทยาศาสตรมหาบัณฑิต (วท.ม.) สาขาเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้า พระนครเหนือ
	พ.ศ. 2562 ปริญญาปรัชญาดุษฎีบัณฑิต (ปร.ด.) สาขาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยมหาสารคาม