การจำแนกประเภทข้อความโฆษณาบนเฟซบุ๊กโดยใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

นายศุภมงคล อัครดำรงค์รัตน์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

CLASSIFICATION OF ADVERTISEMENT TEXT ON FACEBOOK USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE

Mr. Suphamongkol Akkaradamrongrat

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

หัวข้อวิทยานิพนธ์	การจำแนกประเภทข้อความโฆษณาบนเฟซบุ๊กโดยใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย
โดย	นายศุภมงคล อัครดำรงค์รัตน์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(รองศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ)

ศุภมงคล อัครดำรงค์รัตน์ : การจำแนกประเภทข้อความโฆษณาบนเฟซบุ๊กโดยใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย. (CLASSIFICATION OF ADVERTISEMENT TEXT ON FACEBOOK USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุกรี สินธุภิญโญ

นักการตลาดนิยมทำการตลาดผ่านสื่อสังคมออนไลน์มากขึ้นในปัจจุบัน เนื่องจากแพลตฟอร์มโซเชียลมีเดียได้รับความนิยมอย่างมากและมีผู้ใช้งานเป็นจำนวนหลายล้านคน โดยเฉพาะเฟซบุ๊กซึ่งเป็นแพลตฟอร์มที่ได้รับความนิยมสูงสุดในประเทศไทย อย่างไรก็ตามผู้คิดค้นโฆษณาต้องมีความเข้าใจพฤติกรรมของผู้บริโภคเพื่อให้สามารถคิดค้นถ้อยคำโฆษณาที่ดี ตัวแบบ AISAS เป็นตัวแบบหนึ่งซึ่งถูกนำเสนอโดยบริษัทเดนท์สึเพื่ออธิบายพฤติกรรมของผู้บริโภค ตัวแบบดังกล่าวนิยามสถานะที่เกิดขึ้นหลังจากผู้บริโภคเห็นโฆษณาของสินค้าทั้งหมดห้าสถานะ ได้แก่ ความใส่ใจ (Attention) ความสนใจ (Interest) การค้นหา (Search) การลงมือกระทำ (Action) และการแบ่งปัน (Share) วิทยานิพนธ์นี้ได้พัฒนาตัวแบบการเรียนรู้ของเครื่องเพื่อใช้จำแนกประเภทโฆษณาภาษาไทยจากเฟซบุ๊กออกเป็นสถานะตามตัวแบบ AISAS เพื่อเป็นประโยชน์ต่อผู้ลงโฆษณา อย่างไรก็ตาม ข้อมูลที่ถูกรวบรวมมาเพื่อการเรียนรู้เป็นข้อมูลที่ไม่สมดุล เนื่องจากตัวอย่างที่เป็นคลาสบวกมีจำนวนน้อย ทำให้ตัวแบบมีประสิทธิภาพต่ำในการทำนายตัวอย่างบวก เพื่อเพิ่มประสิทธิภาพของตัวแบบ ผู้วิจัยได้นำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย เทคนิคการคัดเลือกคุณลักษณะมาใช้ อีกทั้งได้เสนอเทคนิคการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงประยุกต์ใช้ร่วมกับตัวแบบจำแนกประเภทนาอีฟเบย์ การถดถอยโลจิสติกส์ และซัพพอร์ตเวกเตอร์แมชชีน และได้นำเทคนิคการสร้างข้อความมาประยุกต์ใช้ร่วมกับตัวแบบจำแนกประเภทแอลเอสทีเอ็ม ผลการทดลองพบว่าหลังการประยุกต์ใช้เทคนิคต่าง ๆ ทุกตัวแบบจำแนกประเภทสามารถทำนายตัวอย่างคลาสบวกเป็นจำนวนมากขึ้นและถูกต้องมากขึ้นในเกือบทุกชุดข้อมูล โดยสังเกตได้จากค่าความแม่นและค่าระลึกที่เพิ่มขึ้น เทคนิคการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงทำให้บางตัวแบบมีค่าระลึกเพิ่มขึ้น เทคนิคการสร้างข้อความทำให้ตัวแบบแอลเอสทีเอ็มได้รับค่าระลึกสูงแต่ค่าความแม่นต่ำ อย่างไรก็ตามทุกเทคนิคทำให้ค่าความถูกต้องต่ำลงในชุดข้อมูลส่วนใหญ่

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6070327421 : MAJOR COMPUTER SCIENCE
KEYWORD: AISAS model, SMOTE, Feature selection, Text generation

Suphamongkol Akkaradamrongrat : CLASSIFICATION OF ADVERTISEMENT TEXT ON FACEBOOK USING SYNTHETIC MINORITY OVER-SAMPLING TECHNIQUE. Advisor: Asst. Prof. Dr. SUKREE SINTHUPINYO

Online marketing becomes popular nowadays due the number of users is very high, espicailly Facebook, which is the most popular social media platform in Thailand. However, creating of a good advertising requires understanding in consumer behavior. Dentsu's AISAS model has been proposed to describe consumer behavior. The model defines reaction when the consumer has seen advertising into five stages: attention, interest, search, action, and share. For the benefit of marketers, the purpose of this thesis is to build the machine learning classifier models to classify Thai-language advertisement text from Facebook as the stage they are. However, the collected Facebook dataset is imbalanced due to it contains low positive class. This leads to the low ability of classifier models to predict postitive class samples. To overcome this problem, synthetic minority over-sampling technique (SMOTE), feature selection technique, and the proposed technique, adding of new features which are similar words, were adopted. The results show that these techniques could increase an ability to create model predicting more positive samples in almost dataset. This can be observed in the improving of recall and precision values. Text generation techniques could create model yield high recall but low precision. However, these techniques also decreased the accuracy value in most dataset.

Field of Study: Computer Science Student's Signature .....
Academic Year: 2018 Advisor's Signature .....

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่กรุณาคอยให้คำปรึกษาจนวิทยานิพนธ์นี้สามารถสำเร็จได้ด้วยดี นอกจากนี้อาจารย์ยังได้มอบโอกาสดี ๆ หลายอย่างในชีวิตให้ผู้วิจัย ทำให้ตัวผู้วิจัยเองรู้สึกเชื่อมั่นในความสามารถของตัวเองมากขึ้น

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์ ประธานกรรมการสอบวิทยานิพนธ์ และผู้ช่วยศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ ผู้ให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ และชี้แนะแนวทางในการปรับปรุงวิทยานิพนธ์ให้ดียิ่งขึ้น

ขอขอบคุณภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย ที่กรุณามอบทุนอุดหนุนการศึกษาระดับบัณฑิตศึกษา ซึ่งสนับสนุนค่าเล่าเรียนให้กับผู้วิจัยตลอดระยะเวลาที่ศึกษาอยู่ที่แห่งนี้ ผู้วิจัยรู้สึกยินดีและเป็นเกียรติมากที่ได้รับทุนสนับสนุนนี้

ขอขอบคุณพี่นิสิตปริญญาเอก พรพิมล กะชามาศ ที่ได้ให้คำปรึกษาเกี่ยวกับข้อมูลในการทำวิทยานิพนธ์และได้รวบรวมข้อมูลที่เป็นประโยชน์ให้แก่ผู้วิจัย

ขอขอบคุณคุณแม่ที่คอยอยู่ข้างเคียงและสนับสนุนมาตลอด ทำให้ผู้วิจัยได้มาถึงจุดที่ไกลเกินกว่าที่ฝันไว้ และขอขอบคุณทุก ๆ ท่านที่เคยให้ความช่วยเหลือผู้วิจัยตลอดมา ที่ไม่ได้กล่าวถึงในที่นี้

ศุภมงคล อัครดำรงค์รัตน์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1 บทนำ

1.1. ที่มาและความสำคัญของปัญหา

ความก้าวหน้าทางเทคโนโลยีทำให้ผู้คนสามารถเข้าถึงข้อมูลข่าวสารจำนวนมากได้อย่างง่ายดาย ปัจจุบันเราไม่เพียงแต่เห็นโฆษณาจากสินค้าหลากหลายยี่ห้อผ่านทางโทรทัศน์หรือนิตยสารเท่านั้น แต่ยังรวมไปถึงช่องทางเครือข่ายสังคมออนไลน์ เช่น เฟซบุ๊กหรือทวิตเตอร์ ในประเทศไทยเฟซบุ๊กเป็นแพลตฟอร์มโซเชียลมีเดียที่ได้รับความนิยมสูงที่สุด ในปีพ.ศ. 2560 มีจำนวนผู้ใช้เฟซบุ๊กสูงถึง 49 ล้านยูสเซอร์ [1] ดังนั้นโฆษณาที่ถูกโพสลงบนเฟซบุ๊กจะถูกเห็นโดยผู้คนนับล้านคนในแต่ละวัน และเนื่องจากการแข่งขันทางการตลาดที่สูงในปัจจุบัน ผู้โฆษณาจำนวนมากพยายามเผยแพร่โฆษณาเพื่อส่งเสริมสินค้าของตนเองผ่านทางหลากหลายช่องทาง โดยเฉพาะช่องทางแพลตฟอร์มโซเชียลมีเดีย อย่างไรก็ตามการสร้างถ้อยคำโฆษณาที่ดีเพื่อนำไปโพสช่องทางแพลตฟอร์มโซเชียลมีเดียไม่ใช่เรื่องง่ายนัก ผู้คิดค้นโฆษณาต้องมีความเข้าใจในเรื่องพฤติกรรมของผู้บริโภคจึงจะสามารถสร้างถ้อยคำโฆษณาที่ดีได้

ตัวแบบ (model) หนึ่งซึ่งเป็นที่นิยมใช้ในการอธิบายพฤติกรรมของผู้บริโภคในยุคปัจจุบัน คือ ตัวแบบ AISAS [2] ซึ่งถูกนำเสนอโดยบริษัทเดนท์สึ ตัวแบบดังกล่าวนิยามสถานะต่าง ๆ ที่เกิดขึ้นหลังจากผู้บริโภคได้เห็นโฆษณาของสินค้าหรือบริการ แบ่งออกเป็นห้าสถานะ ได้แก่ ความใส่ใจ (Attention) ความสนใจ (Interest) การค้นหา (Search) การลงมือกระทำ (Action) และการแบ่งปัน (Share) ซึ่งจะเป็นประโยชน์ต่อผู้ลงโฆษณาหากสามารถรู้ได้ว่าถ้อยคำโฆษณาที่คิดค้นขึ้นจะส่งผลอย่างไรต่อพฤติกรรมของผู้บริโภค

ในงานวิจัยนี้มีจุดประสงค์เพื่อประยุกต์ใช้ขั้นตอนวิธีในด้านการเรียนรู้ของเครื่องเพื่อสร้างตัวแบบที่สามารถจำแนกประเภทของถ้อยคำโฆษณาภาษาไทยที่ถูกโพสบนเฟซบุ๊กออกเป็นสถานะต่าง ๆ ตามตัวแบบ AISAS อย่างไรก็ตาม ตัวอย่างชุดข้อมูลที่ถูกรวบรวมมาเป็นชุดข้อมูลที่ไม่สมดุล เนื่องจากข้อมูลตัวอย่างที่เป็นโฆษณาที่สามารถทำให้ผู้บริโภคเกิดสถานะต่าง ๆ อย่างน้อยหนึ่งสถานะนั้นมีจำนวนน้อย ส่งผลให้จำนวนตัวอย่างในแต่ละกลุ่มข้อมูลที่ใช้เพื่อการเรียนรู้ของเครื่องมีความแตกต่างกันมาก ในการแก้ปัญหานี้ ผู้วิจัยจึงเสนอวิธีการแก้ปัญหาสองแบบได้แก่ แบบที่ (1) ผู้วิจัยได้นำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (synthetic minority over-sampling technique: SMOTE) มาประยุกต์ใช้เพื่อให้จำนวนตัวอย่างในแต่ละกลุ่มของชุดข้อมูลฝึกสอนมีจำนวนเท่ากัน รวมถึงได้มีการนำเทคนิคการคัดเลือกคุณลักษณะ (feature selection) มาประยุกต์ใช้ร่วมด้วยเพื่อกำจัด

คุณลักษณะ (feature) ที่ไม่เกี่ยวข้องออกก่อนนำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ อีกทั้งได้เสนอเทคนิคการเพิ่มคำซึ่งเป็นคุณลักษณะใหม่ที่สร้างขึ้นโดยการหาคำคล้ายคลึงโดยพิจารณาจากเวกเตอร์ของคำ โดยผู้วิจัยได้เปรียบเทียบผลลัพธ์ที่ได้จากการนำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ร่วมกับเทคนิคการคัดเลือกคุณลักษณะบนสามขั้นตอนวิธีด้านการเรียนรู้ของเครื่องจักร ได้แก่ นาอีฟเบย์ การถดถอยโลจิสติกส์ และซัพพอร์ตเวกเตอร์แมชชีน และแบบที่ (2) นำวิธีการสร้างข้อความ (text generation) มาใช้เพิ่มจำนวนตัวอย่างกลุ่มน้อย โดยใช้วิธีการสร้างข้อความแบบลูกโซ่แบบมาร์คอฟ (Markov chains) และวิธีการสร้างข้อความด้วยแอลเอสทีเอ็ม (long short-term memory networks: LSTM) แล้วทดสอบบนตัวแบบการจำแนกประเภทแอลเอสทีเอ็ม

1.2. วัตถุประสงค์ของการวิจัย

เพื่อสร้างและพัฒนาตัวแบบที่มีประสิทธิภาพสำหรับจำแนกประเภทของข้อความโฆษณาภาษาไทยออกเป็นสถานะต่าง ๆ ตามตัวแบบ AISAS โดยการประยุกต์ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับการคัดเลือกคุณลักษณะ และวิธีการเพิ่มคุณลักษณะใหม่โดยใช้คำที่มีความหมายคล้ายคลึงกัน รวมถึงเปรียบเทียบประสิทธิภาพของวิธีการจำแนกประเภทแต่ละแบบ ได้แก่ นาอีฟเบย์ การถดถอยโลจิสติกส์ และซัพพอร์ตเวกเตอร์แมชชีน สำหรับงานจำแนกประเภทข้อความโฆษณาภาษาไทย

1.3. ขอบเขตของการวิจัย

1. งานวิจัยนี้ทำการวิเคราะห์ข้อความโฆษณาภาษาไทยซึ่งอาจมีคำภาษาอังกฤษรวมอยู่ด้วย

2. ข้อมูลที่นำมาวิเคราะห์จะคัดเลือกมาเฉพาะข้อความจากโพสต์ ไม่รวมถึงอิโมติคอนหรือรูปภาพ

3. โฆษณาแต่ละกลุ่มจะใช้ยี่ห้อสินค้าหนึ่งยี่ห้อเพื่อเป็นตัวแทนของกลุ่ม

1.4. ประโยชน์ที่คาดว่าจะได้รับจากงานวิจัย

1. สามารถสร้างตัวแบบที่มีประสิทธิภาพสำหรับจำแนกประเภทของข้อความโฆษณาภาษาไทยออกเป็นสถานะต่าง ๆ ตามตัวแบบ AISAS เพื่อเป็นประโยชน์ต่อผู้ลงโฆษณา

2. สามารถเสนอวิธีการที่เหมาะสมสำหรับช่วยให้การจำแนกประเภทข้อความบนชุดข้อมูลที่ไม่สมดุลมีประสิทธิภาพมากขึ้น

1.5. ขั้นตอนในการทำวิจัยเบื้องต้น

1. ศึกษาความรู้และทฤษฎีที่เกี่ยวข้อง

1.1 ศึกษาความรู้เกี่ยวกับตัวแบบ AISAS

1.2 ศึกษาความรู้เกี่ยวกับเทคนิคสำหรับการประมวลผลภาษาธรรมชาติ ได้แก่ การตัดคำ การทำข้อความให้เป็นเวกเตอร์ เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร

1.3 ศึกษาความรู้เกี่ยวกับวิธีการจำแนกประเภท ได้แก่ นาอีฟเบย์ การถดถอยโลจิสติกส์ และซัพพอร์ตเวกเตอร์แมชชีน แอลเอสทีเอ็ม

1.4 ศึกษาความรู้เกี่ยวกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย การสร้างข้อความ

1.5 ศึกษาความรู้เกี่ยวกับเทคนิคการคัดเลือกคุณลักษณะ การแปลงคำเป็นเวกเตอร์

2. ออกแบบแนวคิดการทดลอง

3. ออกแบบวิธีการประเมินประสิทธิภาพของแนวคิดที่นำเสนอ

4. รวบรวมและเตรียมข้อมูลสำหรับการเรียนรู้ของระบบ

5. พัฒนาระบบตามแนวคิดที่ออกแบบไว้

6. ทดสอบและประเมินประสิทธิภาพของแนวคิดที่นำเสนอ

7. วิเคราะห์ผลการทดลอง

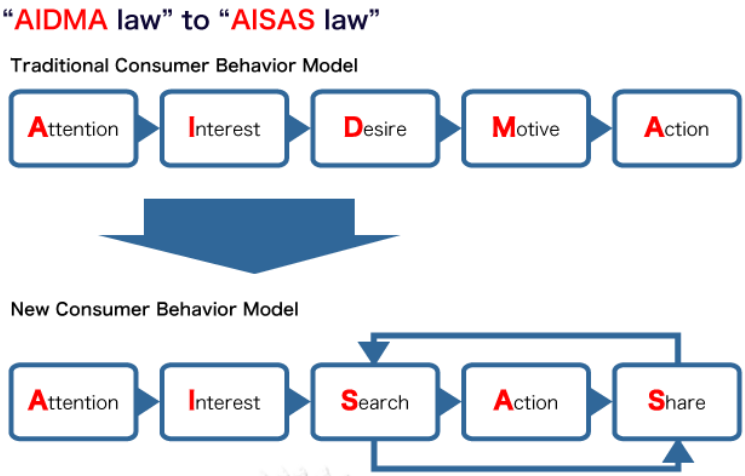
8. สรุปผลการทดลองและจัดทำเล่มวิทยานิพนธ์

บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1. ตัวแบบ AISAS

ตัวแบบ AISAS [2] เป็นหนึ่งในตัวแบบซึ่งนิยมใช้ในการอธิบายพฤติกรรมของผู้บริโภคที่ตอบสนองต่อสินค้าหรือบริการในยุคสมัยใหม่ ตัวแบบดังกล่าวถูกพัฒนามาจากตัวแบบดั้งเดิมคือ AIDMA ซึ่งนำเสนอโดย โรนัลด์ ฮอลล์ ตั้งแต่ปีค.ศ. 1920 โดยตัวแบบ AIDMA ถูกคิดค้นมาเพื่ออธิบายพฤติกรรมของผู้บริโภคซึ่งตอบสนองต่อสินค้าหรือบริการหลังจากได้รับสารจากผู้ประกอบการ โดยเมื่อผู้บริโภคได้เห็นโฆษณาของสินค้าหรือบริการจะเกิดพฤติกรรมที่เกิดขึ้นเป็นขั้นตอนตามลำดับ ได้แก่ เริ่มแรกผู้บริโภคจะใส่ใจ (Attention) ต่อสินค้าหรือบริการ ต่อมาเมื่อผู้บริโภคเริ่มพิจารณาสินค้าหรือบริการ จะเกิดความสนใจ (Interest) และทำให้เกิดความต้องการ (Desire) เกิดจนแรงกระตุ้น (Motive) ส่งผลให้ผู้บริโภคลงมือ (Action) ซื้อสินค้านั้น ตัวแบบนี้ทำให้ผู้ประกอบการสามารถระบุจุดประสงค์ของการโฆษณาว่าจะถูกสร้างขึ้นมาเพื่อให้เกิดอิทธิพลต่อผู้บริโภคที่มีพฤติกรรมในขั้นตอนใด ส่งผลให้โฆษณาสามารถสื่อสารต่อกลุ่มเป้าหมายที่เหมาะสม ซึ่งตรงกับแนวคิดพื้นฐานทางการตลาด

อย่างไรก็ตาม ในยุคปัจจุบันได้เกิดความก้าวหน้าทางเทคโนโลยีส่งผลให้ผู้บริโภคสามารถเข้าถึงข้อมูลจำนวนมากได้อย่างง่ายดายผ่านเครือข่ายอินเทอร์เน็ต รวมถึงมีสินค้ายี่ห้อใหม่เกิดขึ้นจำนวนมากทำให้เกิดการแข่งขันทางการตลาดสูงขึ้น บริษัทเดนท์สึพบว่าตัวแบบ (model) AIDMA ซึ่งอธิบายขั้นตอนการเกิดพฤติกรรมของผู้บริโภคตามลำดับแบบเส้นตรงจึงไม่เหมาะสมต่อการตลาดในยุคปัจจุบัน จึงเกิดเป็นตัวแบบใหม่คือ AISAS โดยตัวแบบดังกล่าวได้อธิบายว่าเมื่อผู้บริโภคได้เห็นโฆษณาของสินค้าหรือบริการ จะเกิดความใส่ใจ (Attention) นำไปสู่ความสนใจ (Interest) จากนั้นผู้บริโภคจะทำการค้นหาข้อมูลเพิ่มเติม (Search) เกี่ยวกับสินค้าหรือบริการทางเครือข่ายอินเทอร์เน็ต ทั้งข้อมูลจากผู้ประกอบการและคำวิจารณ์ต่อสินค้าในเครือข่ายสังคม ซึ่งในขั้นตอนนี้อาจมีการเปรียบเทียบสินค้าแต่ละยี่ห้อ เมื่อพบว่าสินค้าตรงใจจึงจะลงมือ (Action) ซื้อสินค้านั้น และบ่อยครั้งมักตามด้วยการแบ่งปัน (Share) ประสบการณ์และความคิดเห็นในการซื้อและบริโภคสินค้านั้นต่อผู้อื่น ทั้งแบบปากต่อปากในชีวิตจริงและพูดคุยผ่านเครือข่ายสังคมต่าง ๆ บนอินเทอร์เน็ต ซึ่งขั้นตอนนี้จะเห็นได้ว่าผู้บริโภคได้เปลี่ยนแปลงเป็นผู้ส่งสารอีกด้วย



ภาพที่ 1 การพัฒนาของตัวแบบ AIDMA สู่ตัวแบบ AISAS

(*เข้าถึงได้จาก https://www.pinterest.com/pin/301530137522813801/*)

2.2. เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

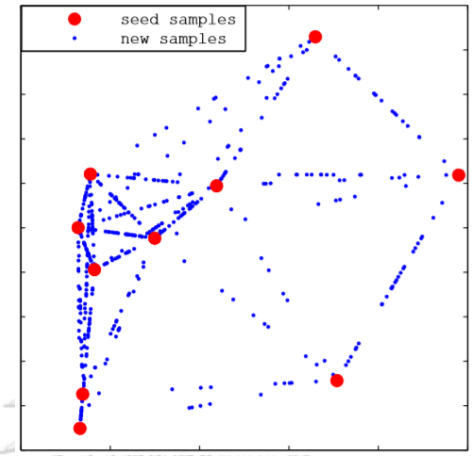
เนื่องจากชุดข้อมูลที่นำมาใช้เป็นชุดข้อมูลที่ไม่สมดุล กล่าวคือ มีกลุ่มตัวอย่างที่เป็นคลาสใดคลาสหนึ่งมากเกินไป ดังเช่น ข้อมูลโฆษณาในกลุ่มเครื่องสำอาง ถูกรวบรวมมาทั้งหมด 999 ตัวอย่าง มีตัวอย่างที่เป็นคลาสแชร์ (Share) ซึ่งเป็นคลาสบวก (positive class) อยู่เพียง 72 ตัวอย่าง และตัวอย่างที่ไม่เป็นคลาสแชร์ (Not Share) ซึ่งเป็นคลาสลบ (negative class) อยู่ถึง 927 ตัวอย่าง ทำให้การทำนายถูกบิดเบือนด้วยกลุ่มตัวอย่างที่เป็นคลาสที่มีจำนวนมาก ตัวแบบจำแนกประเภทอาจทำนายตัวอย่างใหม่เป็นข้อมูลกลุ่มที่มีจำนวนมากเพียงอย่างเดียวก็สามารถได้ค่าความถูกต้อง (accuracy) สูง ส่งผลให้การแปลผลค่าความถูกต้องในการทำนายผิดเพี้ยนไป เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย [3] เป็นวิธีในการสร้างตัวอย่างข้อมูลสังเคราะห์ที่มีความคล้ายคลึงกับตัวอย่างดั้งเดิมโดยการสุ่มตัวอย่างแบบโอเวอร์แซมปลิ้ง (oversampling) กับข้อมูลกลุ่มที่มีจำนวนน้อยให้มีจำนวนมากขึ้นจนใกล้เคียงหรือเท่ากับข้อมูลกลุ่มที่มีจำนวนมาก เพื่อให้ตัวแบบจำแนกประเภทสามารถทำนายตัวอย่างใหม่เป็นกลุ่มข้อมูลที่มีจำนวนน้อยมากขึ้น โดยการทำงานของเทคนิค SMOTE สามารถอธิบายได้ดังต่อไปนี้

(1) พิจารณาแต่ละตัวอย่างเฉพาะที่เป็นตัวอย่างกลุ่มน้อย มองหาตัวอย่างที่เป็นเพื่อนบ้านใกล้สุดกับตัวอย่างดังกล่าวจำนวน k ตัว

(2) สุ่มเลือกตัวอย่างที่เป็นเพื่อนบ้านใกล้สุดมาหนึ่งตัวอย่าง

(3) ลากเส้นเชื่อมตามระยะทางแบบยูคลิเดียนจากตัวอย่างที่กำลังพิจารณาไปยังตัวอย่างเพื่อนบ้านใกล้สุดที่สุ่มมาได้

(4) สุ่มจุดที่อยู่บนเส้นเชื่อมดังกล่าวขึ้นมาเป็นจำนวนเท่ากับจำนวนตัวอย่างใหม่ที่ต้องการ ให้จุดเหล่านั้นเป็นตัวอย่างสังเคราะห์ตัวอย่างใหม่



ภาพที่ 2 การสุ่มข้อมูลตัวอย่างด้วยเทคนิค SMOTE

2.3. การคัดเลือกคุณลักษณะแบบไคกำลังสอง

ในทางสถิติ การทดสอบไคกำลังสอง [5] เป็นการหาความขึ้นต่อกันของการเกิดสองเหตุการณ์ใด ๆ การทดสอบไคกำลังสองถูกนำมาประยุกต์ใช้ในสาขาการเรียนรู้ของเครื่องเพื่อคัดเลือกคุณลักษณะที่มีความขึ้นต่อคลาสเช่นกัน สำหรับการคัดเลือกคุณลักษณะของการจำแนกประเภทข้อความ ทั้งสองเหตุการณ์ถือเป็นการปรากฏของคำ (0, 1) และการเกิดขึ้นของคลาส (0, 1) เราสามารถคำนวณค่าไคกำลังสองของคำใด ๆ ได้ดังสมการต่อไปนี้ เมื่อ กำหนดให้คลาสแบ่งออกเป็นคลาส 0 และคลาส 1

$$\chi^2(D,t,c) = \sum_{e_t \epsilon \{1,0\}} \sum_{e_c \epsilon \{1,0\}} \frac{\left(N_{e_t e_c} - E_{e_t e_c}\right)^2}{E_{e_t e_c}} \qquad (1\text{-}1)$$

โดยกำหนดให้

e_t หมายถึง การปรากฏของคำในเอกสาร มีค่าเป็น 0, 1

e_c หมายถึง คลาสของเอกสาร มีค่าเป็น 0, 1

$N_{e_t e_c}$ หมายถึง ความถี่ที่ได้จากการสังเกต

$E_{e_t e_c}$ หมายถึง ความถี่ที่คาดหวัง

ในกรณีที่คำใด ๆ สามารถมีค่าเป็น 0 (ไม่ปรากฏ), 1 (ปรากฏ) และคลาสของเอกสาร มีค่าเป็น 0, 1 สามารถเขียนสมการให้อยู่ในรูปใหม่ดังต่อไปนี้

$$\chi^2(D,t,c) = \frac{N(N_{11}N_{00}-N_{10}N_{01})^2}{(N_{11}+N_{01})(N_{11}+N_{10})(N_{10}+N_{00})(N_{01}+N_{00})} \quad (1\text{-}2)$$

โดยกำหนดให้

N_{11} หมายถึง ความถี่ของเอกสารที่คำ t ปรากฏและเอกสารมีคลาสเป็น 1

N_{00} หมายถึง ความถี่ของเอกสารที่คำ t ไม่ปรากฏและเอกสารมีคลาสเป็น 0

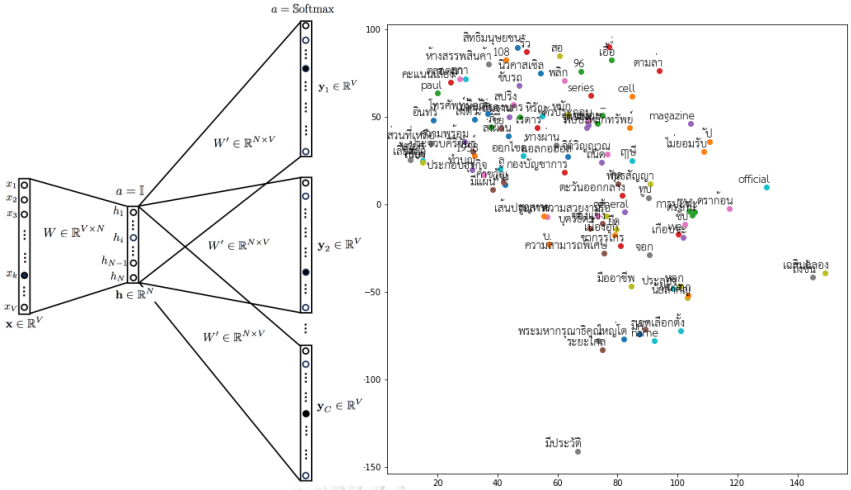
N_{10} หมายถึง ความถี่ของเอกสารที่คำ t ปรากฏแต่เอกสารมีคลาสเป็น 0

N_{01} หมายถึง ความถี่ของเอกสารที่คำ t ไม่ปรากฏแต่เอกสารมีคลาสเป็น 1

2.4. การแปลงคำเป็นเวกเตอร์

การแปลงคำเป็นเวกเตอร์ (Word2Vec) [6] เป็นการเปลี่ยนรูปแบบคำให้อยู่ในรูปแบบของเวกเตอร์ที่มีความยาวจำกัด การแปลงคำให้อยู่ในรูปแบบของเวกเตอร์แบบดั้งเดิมนั้นคือการเข้ารหัสแบบวันฮ็อท (one-hot) ซึ่งมีข้อเสียคือมีขนาดเวกเตอร์ยาวเท่ากับจำนวนคำทั้งหมดในคลังข้อความ และไม่มีความสัมพันธ์ใด ๆ ระหว่างแต่ละเวกเตอร์ของคำ การแปลงคำเป็นเวกเตอร์แบบใหม่ได้พัฒนาให้สามารถแปลงคำอยู่ในรูปแบบเวกเตอร์ที่มีความยาวสั้นลงและมีความสัมพันธ์ระหว่างแต่ละเวกเตอร์ของคำเนื่องจากเวกเตอร์ของคำในรูปแบบใหม่สร้างขึ้นจากการพิจารณาบริบทของคำที่เกิดใกล้เคียงกัน การสร้างเวกเตอร์ของคำอาศัยสมมติฐานว่าคำอยู่ในบริบทเดียวกัน (ปรากฏใกล้เคียงกัน) ในคลังข้อความจะถือว่ามีความหมายใกล้เคียงกัน โดยเมื่อคำถูกแปลงให้อยู่ในรูปเวกเตอร์แล้ว เวกเตอร์ของคำแต่ละเวกเตอร์สามารถนำมาหาความคล้ายคลึงกันในเชิงความหมายได้ ซึ่งคำที่มีความหมายใกล้เคียงกันจะพบว่าเวกเตอร์มีความคล้ายคลึงกันสูง

การสร้างเวกเตอร์ของคำทำได้โดยการใช้เทคนิคสคิบแกรม (skip-gram) ซึ่งประยุกต์ใช้วิธีการเครือข่ายประสาทเทียม (neural networks) เพื่อทำนายคำที่อยู่ในบริบทโดยรอบของคำใด ๆ โดยตัวแบบที่สร้างขึ้นจะทำการรับเข้าคำแต่ละคำในคลังข้อความ (คำที่รับเข้าจะอยู่ในรูปแบบเวกเตอร์ของคำที่ใช้วิธีการแปลงแบบวันฮ็อท) เพื่อทำนายคำที่ปรากฏก่อนและหลังคำนั้น โดยหลังจากตัวแบบดังกล่าวทำการเรียนรู้เสร็จสิ้นแล้ว ค่าน้ำหนัก (weight) ที่อยู่ในส่วนเส้นเชื่อมกับชั้นซ่อน (hidden layer) ของโครงสร้างเครือข่ายประสาทเทียมจะถูกนำไปแทนเวกเตอร์ของแต่ละคำเพื่อนำไปใช้งานต่อไป



ภาพที่ 3 เทคนิคสคิบแกรม และคำซึ่งถูกแทนด้วยเวกเตอร์ในมิติใหม่

(*เข้าถึงได้จาก https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa และ*

https://raw.githubusercontent.com/cstorm125/thai2fit/master/images/random.png)

2.5. นาอีฟเบย์

นาอีฟเบย์ (Naïve Bayes) [7] เป็นวิธีการจำแนกข้อมูลประเภทหนึ่งซึ่งมีแนวคิดมาจากทฤษฎีชองเบย์ โดยอาศัยหลักการทางสถิติมาใช้จำแนกข้อมูล วิธีการนาอีฟเบย์จะพิจารณาความน่าจะเป็นของข้อมูลที่จะเป็นคลาสใด ๆ โดยคำนวณจากความน่าจะเป็นของการเกิดแต่ละคลาสร่วมกับความน่าจะเป็นของการเกิดแต่ละคุณลักษณะของแต่ละคลาส อย่างไรก็ตามวิธีการนาอีฟเบย์มีสมมติฐานว่าคุณลักษณะแต่ละตัวของข้อมูลไม่ขึ้นต่อกัน ทำให้สามารถลดความซับซ้อนในการคำนวณลงไปมากแต่ยังคงมีประสิทธิภาพในการจำแนกประเภทข้อมูลจากที่ได้เห็นความสำเร็จจากการประยุกต์ใช้วิธีการนาอีฟเบย์ในหลากหลายสาขา รวมไปถึงการจำแนกประเภทข้อความอีกด้วย การเรียนรู้แบบนาอีฟเบย์มีขั้นตอนดังต่อไปนี้

(1) รับชุดข้อมูลสอนเข้ามา คำนวณความน่าจะเป็นในการเกิดคลาสแต่ละคลาสแล้วบันทึกค่าที่ได้เอาไว้ ดังนี้

สำหรับแต่ละคลาส C_i

บันทึกค่า $P'(C_i)$ ซึ่งใช้ประมาณ $P(C_i)$

(2) คำนวณความน่าจะเป็นในการเกิดแต่ละค่าคุณลักษณะของแต่ละคลาส แล้วบันทึกค่าที่ได้เอาไว้ ดังนี้

สำหรับแต่ละค่าคุณลักษณะ $A_j = a_j$ ของแต่ละคลาส C_i

บันทึกค่า $P'(A_j = a_j|C_i)$ ซึ่งใช้ประมาณ $P(A_j = a_j|C_i)$

เมื่อมีตัวอย่างใหม่เข้ามา สามารถใช้ความรู้ที่อยู่ในรูปแบบความน่าจะเป็นที่ได้ถูกบันทึกเอาไว้มาคำนวณเพื่อหาว่าตัวอย่างดังกล่าวมีความน่าจะเป็นสูงสุดที่จะเป็นคลาสใด แล้วจึงจำแนกข้อมูลเป็นคลาสนั้น ดังต่อไปนี้

$$C = Max_{i=1}^{m} P'(C_i) \prod_{j=1}^{n} P'(A_j = a_j|C_i) \quad (2\text{-}1)$$

โดยกำหนดให้

$C_i ; i = 1,2, \dots, m$ หมายถึง คลาสของข้อมูล ซึ่งมีได้ตั้งแต่คลาสที่ 1 จนถึง m

$A_i ; j = 1,2, \dots, n$ หมายถึง ค่าคุณลักษณะของข้อมูล ซึ่งมีได้ตั้งแต่ค่าคุณลักษณะที่ 1 จนถึง n

C หมายถึง คลาสซึ่งเป็นคำตอบของตัวอย่างใหม่

2.6. การถดถอยโลจิสติกส์

การถดถอยโลจิสติกส์ (logistic regression) [8] เป็นวิธีการทางสถิติที่นำมาประยุกต์ใช้ในการเรียนรู้ของเครื่องเพื่อจำแนกประเภทข้อมูลที่มีสองกลุ่ม ซึ่งปกติแล้วจะนิยมแทนด้วยกลุ่ม 0 และกลุ่ม 1 โดยวิธีการนี้จะใช้ตัวแปรอิสระซึ่งในที่นี้คือคุณลักษณะของข้อมูลในการทำนายโอกาสที่จะเกิดตัวแปรตามหรือคลาสของข้อมูล โดยเราสามารถคำนวณค่าความน่าจะเป็นในการเกิดเหตุการณ์ได้ดังสมการที่ (3-1)

$$h = \frac{1}{1+e^{-z}} \quad (3\text{-}1)$$

โดยที่ $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

โดยกำหนดให้

$\beta_i ; i = 1,2, \dots, p$ หมายถึง ค่าพารามิเตอร์ที่ทำการประมาณค่าได้จากคุณลักษณะของข้อมูล

$X_1, X_2, \dots, X_p$ หมายถึง ตัวแปรอิสระหรือคุณลักษณะของข้อมูลซึ่งมีทั้งหมด p ตัว

h หมายถึง ค่าสมมติฐานซึ่งแทนความน่าจะเป็นในการเกิดเหตุการณ์ มีค่าระหว่าง 0 ถึง 1

e หมายถึง ลอการิทึมธรรมชาติซึ่งมีค่าโดยประมาณเท่ากับ 2.71828

และสามารถคำนวณค่าคอสท์ฟังก์ชัน [9] ของการถดถอยโลจิสติกส์ในการหาค่าความผิดพลาด (error) เพื่อช่วยในการประมาณค่าพารามิเตอร์ (β_i) ได้ดังสมการที่ (3-2)

$$cost(h, Y) = -\log(h) \quad if\ Y = 1$$
$$= -\log(1-h) \quad if\ Y = 0 \tag{3-2}$$

โดยกำหนดให้

h หมายถึง ค่าสมมติฐานซึ่งแทนความน่าจะเป็นในการเกิดเหตุการณ์ มีค่าระหว่าง 0 ถึง 1

Y หมายถึง ตัวแปรตามหรือกลุ่มของข้อมูล มีค่าเป็น 0 หรือ 1

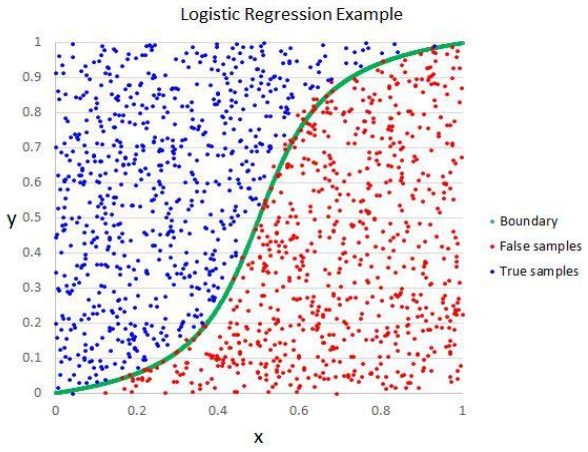
ค่าที่ได้จากการคำนวณค่าความน่าจะเป็นในการเกิดเหตุการณ์ (h) จะอยู่ในช่วง 0 ถึง 1 จากนั้นนำค่าที่ได้มาพิจารณาเปรียบเทียบกับค่าขีดแบ่ง (threshold) ที่เป็นขอบข่ายการตัดสินใจ (decision boundary) ในการทำนายค่าตัวแปรตามหรือกลุ่มของข้อมูลว่าเป็นกลุ่มใด ตัวอย่างเช่นหากเรากำหนดให้ค่าขีดแบ่งเท่ากับ 0.5 จะสามารถทำนายกลุ่มของข้อมูลได้ดังสมการที่ (3-3)

$$if \quad h \geq 0.5 \ \ predict\ Y = 1$$
$$else \ \ predict\ Y = 0 \tag{3-3}$$

โดยกำหนดให้

h หมายถึง ค่าสมมติฐานซึ่งแทนความน่าจะเป็นในการเกิดเหตุการณ์ มีค่าระหว่าง 0 ถึง 1

Y หมายถึง ตัวแปรตามหรือกลุ่มของข้อมูล มีค่าเป็น 0 หรือ 1



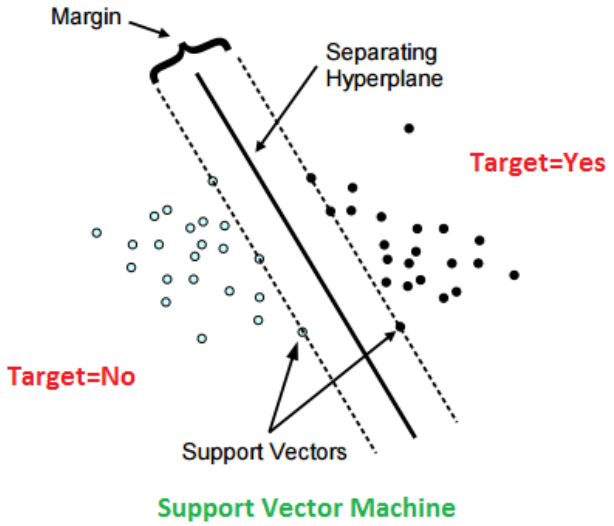
ภาพที่ 4 การแบ่งกลุ่มของข้อมูลออกเป็นสองกลุ่มได้แก่กลุ่ม 1 (TRUE) และกลุ่ม 0 (FALSE) โดยใช้การวิเคราะห์การถดถอยโลจิสติกส์

(เข้าถึงได้จาก https://qph.ec.quoracdn.net/main-qimg-92cb1d347c71244a26c63032198c5e81-c)

2.7. ซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีน [10] เป็นวิธีการจำแนกประเภทหนึ่งที่ถูกนำมาประยุกต์ใช้ในการจำแนกประเภทข้อความเนื่องจากมีประสิทธิภาพสูง แนวคิดการทำงานของซัพพอร์ตเวกเตอร์แมชชีนคือการสร้างระนาบเส้นตรง (hyperplane) ที่สามารถแบ่งแยกกลุ่มข้อมูลได้ดีที่สุด นั่นคือมีระยะห่าง (margin: $\gamma = \frac{1}{\|w\|}$) จากตัวอย่างที่เป็นซัพพอร์ตเวกเตอร์ (support vector) มากที่สุด ซึ่งสอดคล้องกับสมการดังต่อไปนี้

$$\max_{w,\gamma} \gamma \;\; such\ that\ \forall i, y_i(wx_i + b) \geq \gamma \qquad (4\text{-}1)$$



ภาพที่ 5 การสร้างระนาบเส้นตรงที่สามารถแบ่งแยกสองกลุ่มข้อมูลได้ดีที่สุดตามวิธีการซัพพอร์ตเวกเตอร์แมชชีน

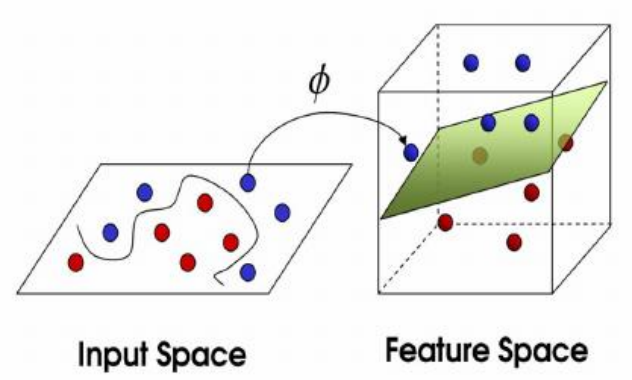
(เข้าถึงได้จาก http://dni-institute.in/blogs/wp-content/uploads/2015/09/SVM-Planes.png)

อย่างไรก็ตาม ปกติการสร้างระนาบที่สามารถแบ่งกลุ่มทุกข้อมูลตัวอย่างออกอย่างสมบูรณ์แบบมักไม่สามารถทำได้ วิธีการต่อมาจึงยอมให้มีการเกิดความผิดพลาดในการแบ่งตัวอย่างข้อมูลได้ แต่ตัวอย่างข้อมูลที่ถูกแบ่งไปอยู่ในกลุ่มที่ผิดจะต้องมีระยะห่างจากระนาบเส้นตรงน้อยที่สุด จึงมีการนำค่าซึ่งเป็นค่าการลงโทษ (ξ) ซึ่งเป็นระยะห่างจากตัวอย่างข้อมูลที่ถูกแบ่งไปอยู่ในกลุ่ม

ที่ผิดกับระนาบเส้นตรงเข้ามาคำนวณด้วย ระนาบเส้นตรงที่สร้างได้จึงสอดคล้องกับสมการใหม่ดังต่อไปนี้

$$\begin{aligned} &\min_{w,b,\xi_i \geq 0} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \\ &such\ that\ \forall i, y_i(wx_i + b) \geq 1 - \xi_i \end{aligned} \tag{4-2}$$

นอกจากนี้ซัพพอร์ตเวกเตอร์แมชชีนยังมีเคอร์เนลฟังก์ชันที่เสมือนแมปข้อมูลไปอยู่ในมิติใหม่ที่สูงขึ้นเพื่อให้ระนาบเส้นตรงสามารถแบ่งแยกข้อมูลได้อีกด้วย

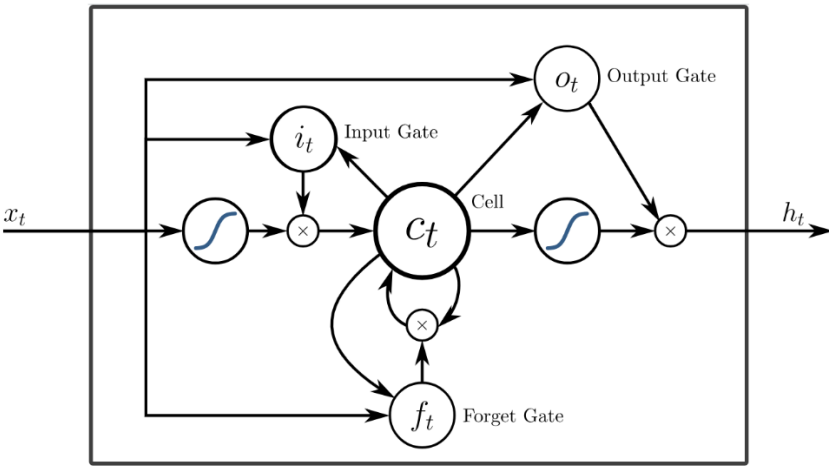


ภาพที่ 6 การแมปข้อมูลไปอยู่ในฟีเจอร์สเปซซึ่งเป็นมิติที่สูงขึ้นเพื่อให้ระนาบสามารถแบ่งแยกข้อมูลได้

(เข้าถึงได้จาก https://www.researchgate.net/figure/SVM-Feature-space_fig3_274896715)

2.8. แอลเอสทีเอ็ม

แอลเอสทีเอ็ม (Long short-term memory: LSTM) เป็นรูปแบบหนึ่งของเครือข่ายประสาทเทียมซึ่งถูกนำเสนอเพื่อแก้ปัญหาการขึ้นต่อกันในระยะยาว (Long-term dependency) ของเครือข่ายประสาทเทียมแบบอาร์เอ็นเอ็นดั้งเดิม (Recurrent neural network: RNN) เครือข่ายประสาทเทียมประเภทนี้นิยมใช้ในงานวิเคราะห์อนุกรมเวลา (Time series analysis) และการประมวลภาษาธรรมชาติ (Natural language processing) แอลเอสทีเอ็มแต่ละหน่วยประกอบด้วยอินพุตเกต (Input gate) ซึ่งมีหน้าที่ควบคุมการอัปเดตข้อมูล, เอาท์พุตเกต (Output gate) ซึ่งมีหน้าที่ควบคุมการส่งออกข้อมูล และฟอร์เก็ตเกต (Forget gate) ซึ่งมีหน้าที่ควบคุมการกำจัดข้อมูล แอลเอสทีเอ็มแต่ละหน่วยประกอบด้วยสมการทรานซิชันดังต่อไปนี้



ภาพที่ 7 โครงสร้างของแอลเอสทีเอ็มแต่ละหน่วย

(เข้าถึงได้จาก https://en.wikipedia.org/wiki/Long_short-term_memory)

$$\mathrm{i}_t = \sigma(\mathrm{W}_t\mathrm{x}_t + \mathrm{U}_i\mathrm{h}_{t-1} + \mathrm{V}_i\mathrm{c}_{t-1}) \quad (5\text{-}1)$$

$$\mathrm{f}_t = \sigma(\mathrm{W}_f\mathrm{x}_t + \mathrm{U}_f\mathrm{h}_{t-1} + \mathrm{V}_f\mathrm{c}_{t-1}) \quad (5\text{-}2)$$

$$\mathrm{o}_t = \sigma(\mathrm{W}_o\mathrm{x}_t + \mathrm{U}_o\mathrm{h}_{t-1} + \mathrm{V}_o\mathrm{c}_{t-1}) \quad (5\text{-}3)$$

$$\tilde{\mathrm{c}}_t = \tanh(\mathrm{W}_c\mathrm{x}_t + \mathrm{U}_c\mathrm{h}_{t-1}) \quad (5\text{-}4)$$

$$\mathrm{c}_t = f_t^i * \mathrm{c}_{t-1} + \mathrm{i}_t * \tilde{\mathrm{c}}_t \quad (5\text{-}5)$$

$$\mathrm{h}_t = \mathrm{o}_t * \tanh(\mathrm{c}_t) \quad (5\text{-}6)$$

โดยกำหนดให้

i_t หมายถึง อินพุตเกต

f_t หมายถึง ฟอร์เก็ตเกต

o_t หมายถึง เอาท์พุตเกต

c_t หมายถึง หน่วยความจำ (Memory cell)

h_t หมายถึง สถานะซ่อน (Hidden state)

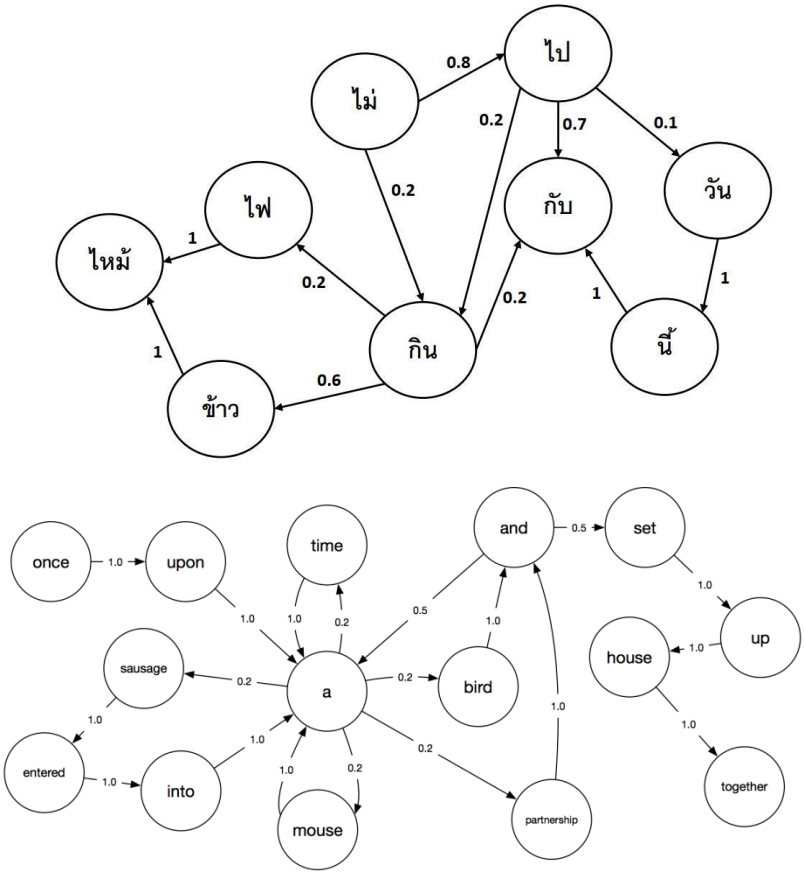
x_t หมายถึง อินพุตในเวลาปัจจุบัน

σ หมายถึง ฟังก์ชันโลจิสติกส์ซิกมอยด์ (Logistic sigmoid function)

$*$ หมายถึง การคูณเอลิเมนต์ไวซ์ (Element-wise multiplication)

2.9. การสร้างข้อความแบบลูกโซ่แบบมาร์คอฟ

ความหมายของลูกโซ่แบบมาร์คอฟ (Markov chains) [11] สามารถอธิบายได้ดังนี้ เมื่อมีการกำหนดกลุ่มของสถานะ (state) ใด ๆ $S = \{s_1, s_2, \ldots, s_r\}$ การเปลี่ยนสถานะจากสถานะหนึ่งในปัจจุบัน s_i ไปยังสถานะใด ๆ s_j จะถูกกำหนดด้วยความน่าจะเป็น p_{ij} และความน่าจะเป็นดังกล่าวไม่ขึ้นกับความน่าจะเป็นในขั้นตอนการเปลี่ยนสถานะก่อนหน้า ความน่าจะเป็น p_{ij} จะถูกเรียกว่าความน่าจะเป็นในการเปลี่ยนสถานะ (transition probabilities) สำหรับสถานะเริ่มต้นอาจถูกกำหนดโดยเฉพาะเจาะจงหรือถูกนิยามการเกิดด้วยความน่าจะเป็น p_i



ภาพที่ 8 ความน่าจะเป็นต่าง ๆ ของการเปลี่ยนสถานะของคำ

(เข้าถึงได้จาก http://www.thagomizer.com/blog/2017/11/07/markov-models.html)

ในแง่ของการสร้างข้อความแบบลูกโซ่แบบมาร์คอฟ (Markov chains text generation) แต่ละสถานะจะถูกแทนด้วยคำ การเปลี่ยนสถานะหมายถึง การเกิดคำใด ๆ ที่ตามมาเมื่อกำหนดคำก่อนหน้า สำหรับในขั้นตอนการเรียนรู้ ความน่าจะเป็นของการเกิดคำใด ๆ เมื่อกำหนดคำก่อน

หน้าจะถูกสำรวจจากคลังข้อความ เมื่อขั้นตอนการเรียนรู้เสร็จสิ้น การสร้างข้อความสามารถทำได้โดยขั้นตอนดังนี้

(1) กำหนดคำเริ่มต้น ซึ่งอาจจะกำหนดโดยเฉพาะเจาะจง หรือสุ่มคำแบบถ่วงน้ำหนัก โดยให้น้ำหนักของแต่ละคำเป็นความน่าจะเป็นในการเกิดคำนั้น ๆ ในคลังข้อความ

(2) การสร้างคำถัดมา ทำได้โดยการสุ่มคำแบบถ่วงน้ำหนักตามความน่าจะเป็นในการเกิดคำเมื่อกำหนดคำก่อนหน้า ค่าน้ำหนักของแต่ละคำ (W_{t_i}) สามารถคำนวณได้ดังสมการต่อไปนี้

$$W_{t_i} = P(t_i|t_{i-1}) \quad (5\text{-}1)$$

t_i หมายถึง คำใด ๆ ที่กำลังพิจารณาในตำแหน่งปัจจุบัน

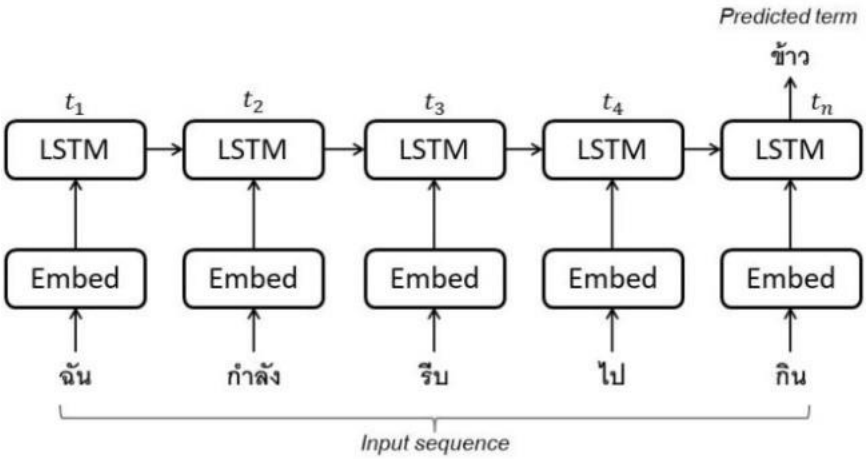
t_{i-1} หมายถึง คำก่อนหน้าคำที่กำลังพิจารณาในตำแหน่งปัจจุบัน

W_{t_i} หมายถึง ค่าน้ำหนักของคำ t_i ซึ่งมีค่าเท่ากับ ความน่าจะเป็นในการเกิดคำ t_i เมื่อกำหนดคำก่อนหน้า t_{i-1}

2.10. การสร้างข้อความด้วยแอลเอสทีเอ็ม

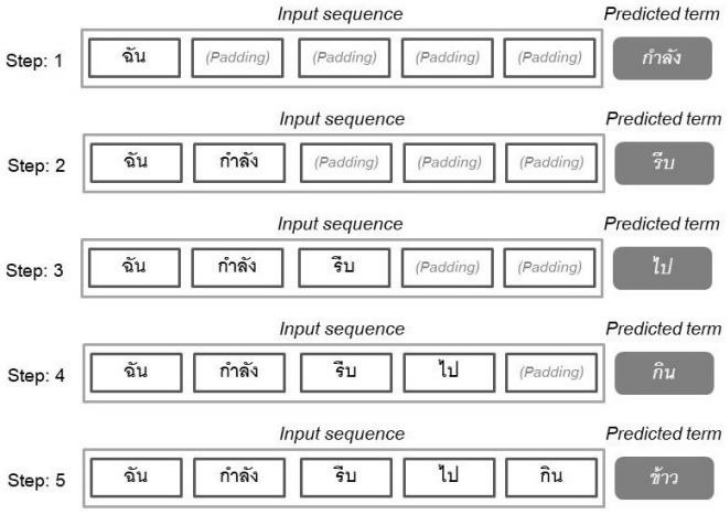
เครือข่ายประสาทเทียมแบบแอลเอสทีเอ็มประสบความสำเร็จอย่างมากในสาขาการประมวลภาษาธรรมชาติ แอลเอสทีเอ็มยังถูกนำมาประยุกต์ใช้ในการสร้างข้อความสังเคราะห์เช่นกัน ในงานวิจัยนี้ แอลเอสทีเอ็มสร้างข้อความโดยการรับเข้าชุดลำดับของคำ (Sequence of terms) เพื่อทำนายคำถัดไป โครงสร้างแอลเอสทีเอ็มรูปแบบนี้ถูกเรียกว่า โครงสร้างแบบหลายต่อหนึ่ง (Many-to-one)

ในงานวิจัยนี้ เอกสารที่เป็นคลาสบวกแต่ละเอกสารจะถูกแบ่งออกเป็นชุดลำดับของคำเพื่อเป็นข้อมูลฝึกสำหรับสร้างข้อความ ตัวอย่างเช่น เอกสารที่ประกอบด้วยคำดังนี้ “ฉัน กำลัง รีบ ไป กิน ข้าว” จะถูกแบ่งออกเป็นห้าตัวอย่างเพื่อเป็นชุดข้อความฝึกสอน ดังนี้ (1) “ฉัน” และฉลาก (Label) “กำลัง” (2) “ฉัน กำลัง” และฉลาก “รีบ” (3) “ฉัน กำลัง รีบ” และฉลาก “ไป” (4) “ฉัน กำลัง รีบ ไป” และฉลาก “กิน” (5) “ฉัน กำลัง รีบ ไป กิน” และฉลาก “ข้าว”



ภาพที่ 9 กระบวนการทำนายคำของแอลเอสทีเอ็ม

เมื่อกระบวนการเรียนรู้ของเครือข่ายแอลเอสทีเอ็มเสร็จสิ้น ข้อความสังเคราะห์สามารถถูกสร้างขึ้นได้โดยการเลือกคำหนึ่งคำ (Seed word) ขึ้นมาป้อนเข้าสู่เครือข่ายแอลเอสทีเอ็มเพื่อทำนายคำถัดไป จากนั้นใช้คำเหล่านั้นที่ได้เพื่อทำนายคำถัดไปเรื่อย ๆ จนกว่าจะครบจำนวนคำที่กำหนด จึงจะได้เอกสารสังเคราะห์ขึ้นมาหนึ่งเอกสาร



ภาพที่ 10 ชุดลำดับของคำซึ่งใช้เป็นอินพุต และคำถัดไปของลำดับของคำเหล่านั้นถูกใช้เป็นฉลาก

2.11. งานวิจัยที่เกี่ยวข้อง

2.11.1. เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยได้ถูกนำมาประยุกต์ใช้เพื่อแก้ปัญหาความไม่สมดุลของคลาสอย่างแพร่หลาย เทคนิคการสุ่มตัวอย่างแบบดั้งเดิมนั้นจะสร้างตัวอย่างใหม่โดยเพียงการ

ผลิตซ้ำตัวอย่างกลุ่มน้อยให้มีจำนวนมากขึ้น วิธีการดังกล่าวมีผลเสียคือทำให้ตัวแบบยึดติดกับข้อมูลสอนมากเกินไป (overfitting) ด้วยเหตุนี้ เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยจึงได้ถูกนำเสนอเพื่อแก้ปัญหาดังกล่าว โดยการสังเคราะห์ตัวอย่างใหม่ซึ่งมีความแตกต่างแต่คล้ายคลึงกับตัวอย่างเดิม วิธีการดังกล่าวถูกนำไปประยุกต์ใช้อย่างมากมายเนื่องจากมีประสิทธิภาพสูงในการเพิ่มประสิทธิภาพของชุดข้อมูลที่ไม่สมดุล ในเวลาต่อมามีงานวิจัยที่พยายามเสนอวิธีการใหม่ ๆ ในการพัฒนาประสิทธิภาพของเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยอย่างมากมาย เช่น เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบเส้นเขตแดน (borderline-SMOTE) เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบซัพพอร์ตเวกเตอร์แมชชีน (SVM SMOTE) เป็นต้น จึงมีงานวิจัยที่ทำการเปรียบเทียบประสิทธิภาพของเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยหลาย ๆ แบบ เช่น ในงานวิจัยของ [12] ผู้วิจัยได้เปรียบเทียบการจำแนกประเภทข้อความภาษาเกาหลีโดยการประยุกต์ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยหกแบบ ได้แก่ เทคนิคการสุ่มตัวอย่างแบบดั้งเดิม เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบปกติ เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบเส้นเขตแดน เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบเส้นเขตแดน 2 (borderline-SMOTE2) เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบซัพ-พอร์ตเวกเตอร์แมชชีน (SVM SMOTE) และการสังเคราะห์แบบปรับเปลี่ยนได้ (ADASYN) โดยทดลองบนหลายตัวแบบ ได้แก่ เพื่อนบ้านไกล้สุดเคตัว (KNN) การถอดถอยโลจิสติกส์ นาอีฟเบย์ และ ซัพพอร์ตเวกเตอร์แมชชีน ก่อนการสุ่มเพิ่มตัวอย่าง ผู้วิจัยยังได้แบ่งความไม่สมดุลของชุดข้อมูลออกเป็นหลายระดับ ตั้งแต่ ระดับความไม่สมดุล 5 เปอร์เซ็นต์ (ข้อมูลมีความไม่สมดุลสูงที่สุด) ไปจนถึงระดับความไม่สมดุล 95 เปอร์เซ็นต์ (ข้อมูลมีความไม่สมดุลต่ำที่สุด) พบว่าที่ระดับความไม่สมดุล 95 เปอร์เซ็นต์ ทุกตัวแบบให้ผลลัพธ์ที่ดีสุดและไม่มีความแตกต่างอย่างมีนัยสำคัญสำหรับแต่ละเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแต่ละแบบในแง่ของคะแนนเอฟวัน (F1 score) และงานวิจัยนี้ยังแสดงให้เห็นว่าตัวแบบนาอีฟเบย์มีความทนทานที่สุดในการทำนายตัวอย่างกลุ่มน้อยเมื่อชุดข้อมูลที่ความไม่สมดุลสูง แต่เมื่อชุดข้อมูลถูกทำให้สมดุลแล้ว ตัวแบบซัพพอร์ตเวกเตอร์แมชชีนจะมีประสิทธิภาพสูงที่สุด

ในอดีตมีงานวิจัยที่ได้นำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาประยุกต์ใช้กับการจำแนกประเภทข้อความภาษาไทยแล้วเช่นกัน งานวิจัยของ [13] ผู้วิจัยได้ประยุกต์ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยเพื่อจำแนกคลาสอารมณ์ของข้อความภาษาไทยซึ่งเป็นความคิดเห็นต่อวิดีโอบนเว็บไซต์ยูทูป โดยจำแนกออกเป็นหกคลาสอารมณ์ ได้แก่ โกรธ กลัว ขยะแขยง มีความสุข เศร้า และประหลาดใจ เนื่องจากในชุดข้อมูลมีบางคลาสอารมณ์ปรากฏน้อยมาก ผู้วิจัยจึงได้ใช้เทคนิค

การสุ่มเพิ่มตัวอย่างกลุ่มน้อยทำให้ชุดข้อมูลมีความสมดุลก่อนที่จะป้อนชุดข้อมูลดังกล่าวเพื่อฝึกสอนตัวแบบ ซึ่งมีสามตัวแบบที่ถูกนำมาทดลอง ได้แก่ ซัพพอร์ตเวกเตอร์แมชชีน มัลติโนเมียลนาอีฟเบย์ และต้นไม้ตัดสินใจ พบว่าทุกตัวแบบมีประสิทธิภาพในการทำนายตัวอย่างใหม่สูงขึ้นในแง่ค่าความถูกต้องหลังจากชุดข้อมูลถูกทำให้สมดุล และตัวแบบซัพพอร์ตเวกเตอร์แมชชีนให้ค่าความถูกต้องสูงที่สุด นอกจากนั้น มีงานวิจัยของ [14] ที่ได้นำเทคนิคนี้ไปประยุกต์ใช้ในงานทางการแพทย์ โดยผู้วิจัยได้นำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้เพื่อเพิ่มประสิทธิภาพการจำแนกชุดข้อมูลโรคเบาหวานแบบสอง ผู้วิจัยยังได้เปรียบเทียบประสิทธิภาพของตัวแบบสำหรับชุดข้อมูลที่มีความไม่สมดุลหลายระดับ พบว่าเมื่อข้อมูลมีความสมดุลมากที่สุด (ทุกคลาสมีจำนวนเท่ากัน) ตัวแบบจะมีประสิทธิภาพในการทำนายตัวอย่างใหม่สูงที่สุด อย่างไรก็ตาม ผู้วิจัยได้พบว่าหลังจากการใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย สำหรับทุกตัวแบบ ค่าระลึก (recall) เพิ่มขึ้นในขณะที่ค่าความถูกต้องลดลง อย่างไรก็ตาม ตัวแบบนาอีฟเบย์มีประสิทธิภาพสูงที่สุดในแง่การเพิ่มขึ้นของค่าระลึก

นอกเหนือจากนี้ยังคงมีงานวิจัยที่เสนอวิธีการเพิ่มประสิทธิภาพเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย เช่น งานวิจัยของ [15] เสนอเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบลำดับชั้น (hierarchical SMOTE) เพื่อแก้ปัญหาการจำแนกคำถาม (questions classification) บนชุดข้อมูลที่ไม่สมดุล ผลการทดลองแสดงให้เห็นว่าวิธีการที่เสนอสามารถเพิ่มประสิทธิภาพของตัวแบบ ในแง่ของ ค่าความถูกต้อง ค่าความแม่น (precision) ค่าระลึก และคะแนนเอฟวัน โดยทดลองบนตัวแบบนาอีฟเบย์

2.11.2. การคัดเลือกคุณลักษณะ

วิธีการคัดเลือกคุณลักษณะถูกนำมาใช้เพื่อสกัดคุณลักษณะที่เกี่ยวข้องกับคลาสและขจัดคุณลักษณะที่ไม่เกี่ยวข้องออกไป วิธีการดังกล่าวถูกนำมาประยุกต์ใช้อย่างแพร่หลายในงานจำแนกประเภทข้อความซึ่งจำนวนคุณลักษณะของตัวอย่างมีปริมาณมาก มีหลายงานวิจัยได้แสดงให้เห็นถึงความสำเร็จในการประยุกต์ใช้การคัดเลือกคุณลักษณะทั้งในแง่ของการลดเวลาในการประมวลผล และในหลายครั้งสามารถเพิ่มประสิทธิภาพในการจำแนกประเภทข้อมูลของตัวแบบได้อีกด้วย ตัวอย่างเช่น งานวิจัย [16] ได้เสนอการประยุกต์ใช้วิธีการคัดลักษณะคุณลักษณะแบบจีนีอินเด็กซ์ (Gini index) ร่วมกับตัวแบบซัพพอร์ตเวกเตอร์แมชชีนสำหรับงานจำแนกประเภทความรู้สึกในข้อความ พบว่าผลลัพธ์การจำแนกประเภทมีประสิทธิภาพมากขึ้นทั้งในแง่ของค่าความถูกต้องและการลดลงของเวลาที่ใช้ในการประมวลผล นอกจากนี้วิธีการคัดเลือกคุณลักษณะ

ที่นิยมในปัจจุบันมีหลายวิธีการ มีผู้วิจัย [17] ได้เปรียบเทียบประสิทธิภาพของวิธีการคัดเลือกคุณลักษณะสี่วิธีการ ได้แก่ ความถี่ของเอกสาร สถิติไค (CHI statistics) อินฟอร์เมชั่นเกน (information gain) และเกนเรโช (gain ratio) โดยได้ทดลองบนหลายตัวแบบ ผู้วิจัยพบว่าวิธีการเกนเรโชมีประสิทธิภาพสูงกว่าวิธีการอื่น และตัวแบบซัพพอร์ตเวกเตอร์แมชชีนให้ผลลัพธ์ดีที่สุดในแง่ของค่าความถูกต้อง และยังมีงานวิจัยที่ได้เสนอวิธีการคัดเลือกคุณลักษณะใหม่ ๆ เช่นกัน ดังเช่น งานวิจัย [18] ผู้วิจัยได้นำเสนอวิธีการใหม่โดยการประยุกต์ใช้ขั้นตอนวิธีทางพันธุกรรม (genetic algorithm) โดยวิธีการที่เสนอสามารถให้ผลลัพธ์สูงกว่าวิธีการคัดลักษณะคุณลักษณะแบบดั้งเดิมที่มีอยู่สำหรับการจำแนกประเภทความรู้สึกในข้อความ

2.11.3. การสร้างข้อความ

การสร้างข้อความ (text generation) หรือการสร้างภาษาธรรมชาติ (natural language generation) เป็นการพัฒนาวิธีเพื่อให้สามารถสร้างข้อความที่คล้ายคลึงกับข้อความที่เป็นภาษาธรรมชาติของมนุษย์ [19] ในปัจจุบันวิธีการทางด้านสถิติและเครือข่ายประสาทเทียมถูกนิยมมาประยุกต์ใช้เพื่อสร้างข้อความ การสร้างข้อความถูกประยุกต์ใช้ในหลายสาขา ดังเช่น การสร้างหุ่นยนต์ตอบโต้อัตโนมัติ (chatbots) การสร้างเนื้อเพลง (lyric generation) การตอบคำถาม (question answering) การแปลด้วยเครื่อง (machine translation) การสร้างแคปชั่นรูปภาพ (image caption) และการเล่าเรื่อง (story telling) ตัวอย่างเช่น ระบบการแปลด้วยเครื่องของกูเกิล (Google's Neural Machine Translation System) [20] ประยุกต์ใช้แอลเอสทีเอ็มและแอลเอสทีเอ็มแบบสองทางเพื่อแปลข้อความ บางงานวิจัย [21] ประยุกต์ใช้แอลเอสทีเอ็มเพื่อทำระบบโต้ตอบอัตโนมัติโดยกำหนดเงื่อนไข งานวิจัย [22] ประยุกต์ใช้แอลเอสทีเอ็มสร้างแคปชั่นรูปภาพโดยอัตโนมัติ ในงานวิจัยที่เกี่ยวกับการสร้างเนื้อเพลง งานวิจัย [23] ได้ประยุกต์ใช้วิธีการของมาร์คอฟเพื่อสร้างเนื้อเพลงโดยเรียนรู้จากเพลงของศิลปินจำนวนหลายคน งานวิจัย [24] ใช้แอลเอสทีเอ็มสร้างเนื้อเพลงจากสไตล์เพลงของนักร้องเพลงแร็ป

อย่างไรก็ตาม เครือข่ายประสาทเทียมแบบ GANs (Generative adversarial networks) ถือว่าเป็นศาสตร์แห่งศิลป์ (state-of-the-art) ในการสร้างข้อมูลเทียม GANs บางประเภท เช่น RelGAN (Relational Generative Adversarial Networks) [25] และ TextGAN [26] ถูกพัฒนาขึ้นเพื่อสร้างข้อความเทียมโดยเฉพาะ เมื่อนำข้อความที่ถูกสร้างขึ้นโดยเทคนิคเหล่านี้มาประเมินโดยมนุษย์ จะพบว่าได้รับคะแนนสูงในแง่ของความคล้ายคลึงกับข้อความธรรมชาติที่สร้างโดยมนุษย์

ในงานวิจัยนี้ ผู้วิจัยได้นำเทคนิคการสร้างข้อความมาเพื่อแก้ปัญหาความไม่สมดุลของข้อมูล โดยการสร้างข้อความเทียมซึ่งเป็นข้อมูลคลาสกลุ่มน้อยขึ้นมาเพื่อให้มีจำนวนเท่ากับข้อมูลคลาสกลุ่มมาก แม้ในปัจจุบันจะมีวิธีการมากมายที่นำเสนอเพื่อสร้างข้อความที่เป็นธรรมชาติ เช่น GANs ที่มีอยู่หลากหลายแบบ วิธีการต่าง ๆ ถูกนำเสนอเพื่อสร้างข้อความที่ดูคล้ายภาษาธรรมชาติของมนุษย์มากที่สุด อย่างไรก็ตาม จุดประสงค์ของงานวิจัยนี้ต้องการนำเทคนิคการสร้างข้อความมาแก้ปัญหาชุดข้อความไม่สมดุล ในการทดลองของเรา เทคนิคการสร้างข้อความสองแบบได้ถูกนำมาใช้ แบบแรกเป็นวิธีการทางสถิติ เรียกว่ามาร์คอฟเชน อีกวิธีการหนึ่งเป็นการประยุกต์ใช้เครือข่ายประสาทเทียมเรียกว่าแอลเอสทีเอ็ม

บทที่ 3 ระเบียบวิธีวิจัย

3.1. วิธีการวัดผล

ผู้วิจัยได้ใช้คอนฟิวชั่นเมทริกซ์ (confusion matrix) ดังตารางต่อไปนี้ ในการประเมินประสิทธิภาพของตัวแบบ

ตารางที่ 1 แสดงคอนฟิวชั่นเมทริกซ์ที่ใช้ประเมินประสิทธิภาพของตัวแบบ

	Predicted		
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

ค่าความถูกต้องและคะแนนเอฟวันถูกนำมาใช้วัดประสิทธิภาพในการจำแนกประเภทของตัวแบบ โดยค่าความถูกต้องบ่งบอกถึงความถูกต้องในการทำนายข้อมูลของตัวแบบ ซึ่งคำนวณได้จากสมการ

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (5\text{-}1)$$

อย่างไรตาม ในกรณีของการจำแนกประเภทชุดข้อมูลที่ไม่สมดุล การใช้เพียงค่าความถูกต้องในการประเมินสิทธิภาพอาจสร้างความเข้าใจผิดได้เนื่องจากอาจเกิดอคติของตัวแบบในการทำนายข้อมูล ตัวแบบสามารถได้รับค่าความถูกต้องสูงได้เพียงทำนายทุกตัวอย่างให้เป็นคลาสของกลุ่มตัวอย่างที่มีจำนวนมาก ดังนั้นค่าความแม่น ค่าระลึก รวมไปถึงคะแนนเอฟวัน จึงได้ถูกนำมาคำนวณด้วย เนื่องจากตัวแบบจะได้รับค่าความแม่นและค่าระลึกสูงก็ต่อเมื่อสามารถทำนายคลาสที่เป็นบวกได้จำนวนมากและมีความถูกต้องสูง ส่วนค่าคะแนนเอฟวันเป็นการคำนวณร่วมกันของทั้งค่าความแม่นและค่าระลึก โดยค่าความแม่น ค่าระลึก และคะแนนเอฟวัน สามารถคำนวณได้ดังสมการต่อไปนี้

$$Precision = \frac{TP}{TP+FP} \quad (5\text{-}2)$$

$$Recall = \frac{TP}{TP+FN} \quad (5\text{-}3)$$

$$F_1 = 2 * \frac{precision*recall}{precision+recall} \quad (5\text{-}4)$$

3.2. ขั้นตอนการดำเนินงาน

3.2.1. การรวบรวมข้อมูล

ข้อความโฆษณาภาษาไทยถูกรวบรวมมาทั้งหมดจำนวน 2,928 ข้อความ ซึ่งแบ่งออกเป็นสามกลุ่ม ได้แก่ เครื่องสำอาง เครื่องใช้ไฟฟ้า และสุขภัณฑ์ โฆษณากลุ่มเครื่องสำอางมีทั้งหมด 999 ข้อความ โฆษณากลุ่มเครื่องใช้ไฟฟ้ามีทั้งหมด 966 ข้อความ และโฆษณากลุ่มสุขภัณฑ์มีทั้งหมด 963 ข้อความ ในเบื้องต้นจะมีการกรองข้อความซ้ำและอีโมติคอนออกไป แต่ละข้อความโฆษณาจะถูกจำแนกประเภทออกเป็นคลาสต่าง ๆ ตามสถานะของตัวแบบ AISAS ดังนี้ : ก่อให้เกิดความใส่ใจ (Attention) หรือไม่ก่อให้เกิดความใส่ใจ (Not Attention), ก่อให้เกิดความสนใจ (Interest) หรือไม่ก่อให้เกิดความสนใจ (Not Interest), ก่อให้เกิดการค้นหา (Search) หรือไม่ก่อให้เกิดการค้นหา (Not Search), ก่อให้เกิดการลงมือกระทำ (Action) หรือไม่ก่อให้เกิดการลงมือกระทำ (Not Action), ก่อให้เกิดการแบ่งปัน (Share) หรือไม่ก่อให้เกิดการแบ่งปัน (Not Share) นิยามของคลาสต่าง ๆ แสดงดังตารางต่อไปนี้

ตารางที่ 2 แสดงนิยามของคลาสต่าง ๆ ตามสถานะของตัวแบบ AISAS

คลาส	นิยาม
Attention	ข้อความโฆษณามีแนวโน้มทำให้ผู้อ่านเกิดความใส่ใจ
Interest	ข้อความโฆษณามีแนวโน้มทำให้ผู้อ่านเกิดความสนใจ
Search	ข้อความโฆษณามีแนวโน้มทำให้ผู้อ่านค้นหาข้อมูลเพิ่มเติม
Action	ข้อความโฆษณามีแนวโน้มทำให้ผู้อ่านตัดสินใจซื้อสินค้าหรือบริการ
Share	ข้อความโฆษณามีแนวโน้มทำให้ผู้อ่านแบ่งปันข้อมูลให้กับผู้อื่น

ข้อความโฆษณาภาษาไทยในการจำแนกประเภทโฆษณา แต่ละโฆษณาจะถูกโหวตคลาสที่เหมาะสมโดยครูผู้เชี่ยวชาญด้านภาษาไทยทั้งหมดสามคน ครูแต่ละคนจะเลือกสถานะที่เหมาะสมกับโฆษณาใด ๆ กล่าวคือ เลือกว่าโฆษณานั้นก่อให้เกิดความใส่ใจหรือไม่ ก่อให้เกิดความสนใจหรือไม่ ก่อให้เกิดการค้นหาหรือไม่ ก่อให้เกิดการลงมือกระทำหรือไม่ และก่อให้เกิดการแบ่งปันหรือไม่ จากนั้นโฆษณานั้น ๆ จะถูกตัดสินว่าเป็นคลาสใด ๆ ตามหลักเสียงข้างมากตามการโหวตของครูผู้เชี่ยวชาญทั้งสามคน โฆษณาหนึ่งสามารถมีได้มากกว่าหนึ่งสถานะ เช่น ทั้งก่อให้เกิดความใส่ใจและก่อให้เกิดความสนใจ หรืออาจไม่ก่อให้เกิดสถานะใด ๆ เลยก็ได้ ซึ่งหลังจากการแบ่งกลุ่มโฆษณาเรียบร้อยแล้ว พบว่าโฆษณาที่เป็นสถานะตามตัวแบบ AISAS อย่างน้อยหนึ่งสถานะนั้นมีจำนวนน้อยมาก ส่งผลให้ชุดข้อมูลโฆษณาเป็นชุดข้อมูลที่ไม่สมดุลตัวอย่างคลาสของข้อความโฆษณากลุ่มเครื่องสำอางแสดงดังตารางต่อไปนี้

ตารางที่ 3 แสดงตัวอย่างคลาสของข้อความโฆษณากลุ่มเครื่องสำอาง

ข้อความโฆษณา	คลาส
ทานรกปลาเด้งนามุ และใช้เอสเซนท์มิสท์แล้วตามด้วยครีมบำรุงของสเนลไวท์ สวยได้ครบทั้งภายในและภายนอกเลยนะจ๊ะ	Attention
ครีมหอยทากขาวสเนลไวท์ ตบแล้วใสใช้แล้วตึง ครีมทาหน้าที่ดารานักแสดงนิยมใช้ในเวลานี้	Interest
Namulife ผู้ผลิตและจำหน่าย ผลิตภัณฑ์เมือกหอยทาก Snail White สินค้าที่ดีที่สุด การันตี ด้วยยอดขายอันดับ1 โปรดระวังสินค้าลอกเลียนแบบ	Search
กระเป๋าใส่เครื่องสำอาง ชิคๆ จาก snailwhite พกพาสะดวก ใส่เครื่องสำอาง ใส่ของส่วนตัว ใส่อะไรอะไรก็ได้คุ้มมากเพราะได้ ฟรีๆ แค่ซื้อ snailwhite gift set ที่ Watsons	Action
วันแม่ปีนี้ คุณมีของขวัญให้คุณแม่รึยังคะ มาลุ้นรับ SNAILWHITE Miracle เป็นของขวัญให้คุณแม่กัน มีจำนวนจำกัดเพียง100รางวัลเท่านั้น (มูลค่ารางวัลละ1,190บาท) เพียง 1.กดแชร์โพสนี้ 2.โพสต์รูปคู่กับคุณแม่ พร้อมตั้งคำบรรยายภาพว่า “คุณอยากทำกิจกรรมอะไรกับคุณแม่” พร้อมแฮชแท็ก #MyBeautifulMom #snailwhite	Share

จำนวนโฆษณาของแต่ละกลุ่ม ในแต่ละโฆษณาประเภทต่าง ๆ แสดงดังตารางต่อไปนี้

ตารางที่ 4 แสดงจำนวนโฆษณากลุ่มเครื่องสำอาง

คลาส	จำนวน	คลาส	จำนวน
Attention	36	Not Attention	963
Interest	193	Not Interest	806
Search	86	Not Search	913
Action	22	Not Action	977
Share	72	Not Share	927

ตารางที่ 5 แสดงจำนวนโฆษณากลุ่มเครื่องใช้ไฟฟ้า

คลาส	จำนวน	คลาส	จำนวน
Attention	34	Not Attention	932
Interest	175	Not Interest	791
Search	84	Not Search	882
Action	17	Not Action	949
Share	84	Not Share	882

ตารางที่ 6 แสดงจำนวนโฆษณากลุ่มสุขภัณฑ์

คลาส	จำนวน	คลาส	จำนวน
Attention	40	Not Attention	923
Interest	209	Not Interest	754
Search	74	Not Search	889
Action	15	Not Action	948
Share	69	Not Share	894

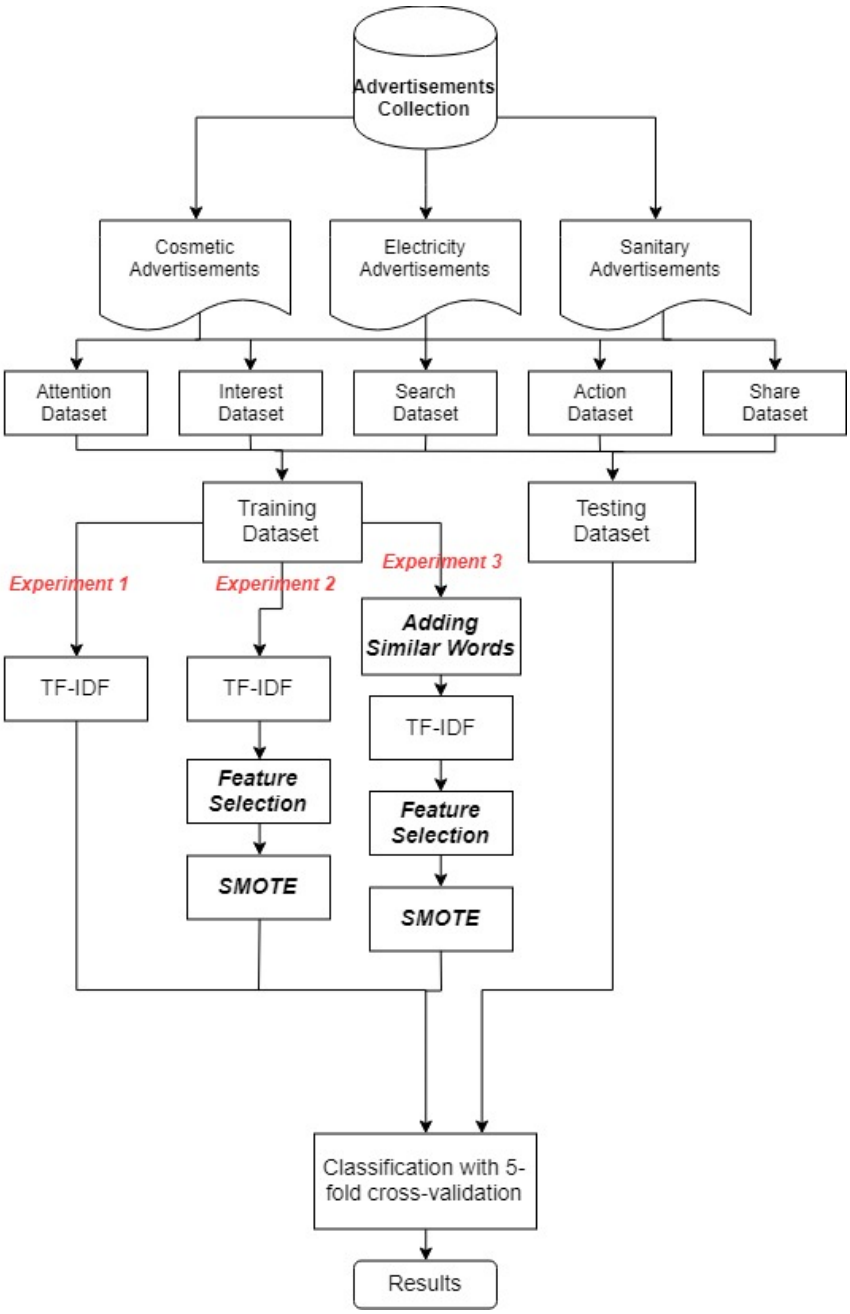
4.2.2. ข้อความโฆษณาถูกนำมาตัดคำโดยใช้ไลบรารีคัดคำ (Cutkum) [27] หลังจากตัดคำแล้ว แต่ละข้อความโฆษณาจะถูกแทนให้อยู่ในรูปแบบเวกเตอร์โดยการใช้เทคนิคเมทริกซ์คำ-เอกสาร (document-term matrix) โดยให้ค่าในแต่ละตำแหน่งของเวกเตอร์เป็นความถี่ของคำ

4.2.3.เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร (term frequency - inverse document frequency: TF-IDF) ได้ถูกนำมาใช้เพื่อให้ค่าในเวกเตอร์ถูกให้ค่าน้ำหนักใหม่ ซึ่งเทคนิคดังกล่าวได้นำความถี่ของคำและการปรากฏของคำในทุก ๆ เอกสารมาพิจารณา โดยให้ถือว่าคำที่มีความถี่ในเอกสารนั้นมากแต่ปรากฏในเอกสารอื่น ๆ น้อยถือเป็นคำที่สำคัญ

4.2.4. ข้อความโฆษณาแต่ละกลุ่ม (เครื่องสำอาง, เครื่องใช้ไฟฟ้า, สุขภัณฑ์) ถูกแยกออกเป็นห้าชุดข้อมูล ได้แก่ ชุดข้อมูล Attention ชุดข้อมูล Interest ชุดข้อมูล Search ชุดข้อมูล Action และชุดข้อมูล Share โดยแต่ละชุดข้อมูลถูกสร้างขึ้นเพื่อจำแนกว่าโฆษณาเกิดสถานะนั้น ๆ หรือไม่ ตัวอย่างเช่น สำหรับชุดข้อมูล Attention แต่ละโฆษณาในชุดข้อมูลนี้จะถูกจำแนกออกเป็นสองคลาส ได้แก่ คลาส Attention (ก่อให้เกิดความใส่ใจ) และคลาส Not Attention (ไม่ก่อให้เกิดความใส่ใจ)

4.2.5. ห้าตัวแบบสำหรับจำแนกประเภทได้ถูกสร้างขึ้นสำหรับแต่ละชุดข้อมูล ได้แก่ (1) ตัวแบบ Attention (Attention Model) สร้างขึ้นเพื่อจำแนกโฆษณาออกเป็นคลาส Attention และคลาส Not Attention ดังนั้นชุดข้อมูล Attention จะถูกป้อนเข้าสู่ตัวแบบดังกล่าว (2) ตัวแบบ Interest (Interest Model) สร้างขึ้นเพื่อจำแนกโฆษณาออกเป็นคลาส Interest และคลาส Not Interest (3) ตัวแบบ Search (Search Model) สร้างขึ้นเพื่อจำแนกโฆษณาออกเป็นคลาส Search และคลาส Not Search (4) ตัวแบบ Action (Action Model) สร้างขึ้นเพื่อจำแนกโฆษณาออกเป็นคลาส Action และคลาส Not Action และ (5) ตัวแบบ Share (Share Model) สร้างขึ้นเพื่อจำแนกโฆษณาออกเป็นคลาส Share และคลาส Not Share ในขั้นตอนนี้ผู้วิจัยได้แบ่งการทดลองออกเป็นสองส่วน ได้แก่ (1) นาอีฟเบย์ การถดถอยโลจิสติกส์ และซัพพอร์ตเวกเตอร์แมชชีน ได้ถูกนำมาประยุกต์ใช้เป็นตัวแบบสำหรับจำแนกประเภท แต่ก่อนการป้อนข้อมูลดังกล่าวเข้าสู่ตัวแบบ ผู้วิจัยได้แบ่งการทดลองออกเป็นสามแบบเพื่อทำการเปรียบเทียบผลลัพธ์ในการจำแนกประเภท ได้แก่ (1.1) ใช้ชุดข้อมูลดั้งเดิมที่ไม่สมดุล (Original Imbalanced Dataset) (1.2) ใช้ชุดข้อมูลที่ใช้เทคนิคคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยเพื่อทำให้ข้อมูลสมดุล (Balanced Dataset using CHI2+SMOTE) (1.3) เพิ่มคุณลักษณะซึ่งเป็นคำใหม่ที่คล้ายคลึงกับคำสำคัญที่ถูกคัดเลือกโดยเทคนิคคัดเลือกคุณลักษณะก่อนการใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (Balanced Dataset using Adding Similar Words+CHI2+SMOTE) โดยเทคนิคการแปลงคำเป็นเวกเตอร์จะถูกประยุกต์ใช้เพื่อหาคำที่คล้ายคลึงกับคำที่สำคัญ และ (2) นำวิธีการสร้างข้อความมาใช้เพื่อทำให้ข้อมูลสมดุล (Balanced Dataset using CHI2+SMOTE) โดยใช้วิธีการสร้างข้อความสองแบบ ได้แก่ การสร้างข้อความแบบลูกโซ่แบบมาร์คอฟ (Markov chains) และการสร้างข้อความโดยใช้แอลเอสทีเอ็ม (long short-term memory networks: LSTM) โดยใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยแบบ oversampling เป็นมาตรฐาน (baseline) และใช้แอลเอสทีเอ็มเป็นตัวแบบสำหรับจำแนกประเภท

โดยการทดลองทั้งหมดสามารถแบ่งเป็นหัวข้อได้ดังนี้



ภาพที่ 11 ขั้นตอนการทดลองสามแบบแรก (1.1, 1.2, 1.3)

4.2.5.1. ชุดข้อมูลที่ไม่สมดุล (Original Imbalanced Dataset)
ชุดข้อมูลดั้งเดิมซึ่งเป็นชุดข้อมูลที่ไม่สมดุลได้ถูกป้อนเข้าสู่ตัวแบบดังกล่าว เทคนิคไขว้ข้ามห้ากลุ่ม (5-fold cross validation) ได้ถูกนำมาใช้ประเมินตัวแบบ

4.2.5.2. ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะ (Balanced Dataset using CHI2+SMOTE)

ชุดข้อมูลต่อมาได้ใช้เทคนิคคัดเลือกคุณลักษณะแบบไคกำลังสอง โดยแบ่งจำนวนคุณลักษณะสำคัญหลายระดับ ได้แก่ 25 50 100 250 500 1000 และทุกคุณลักษณะ ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยก่อนป้อนเข้าสู่ตัวแบบ เทคนิคไขว้ข้ามห้ากลุ่มได้ถูกนำมาใช้ประเมินตัวแบบ

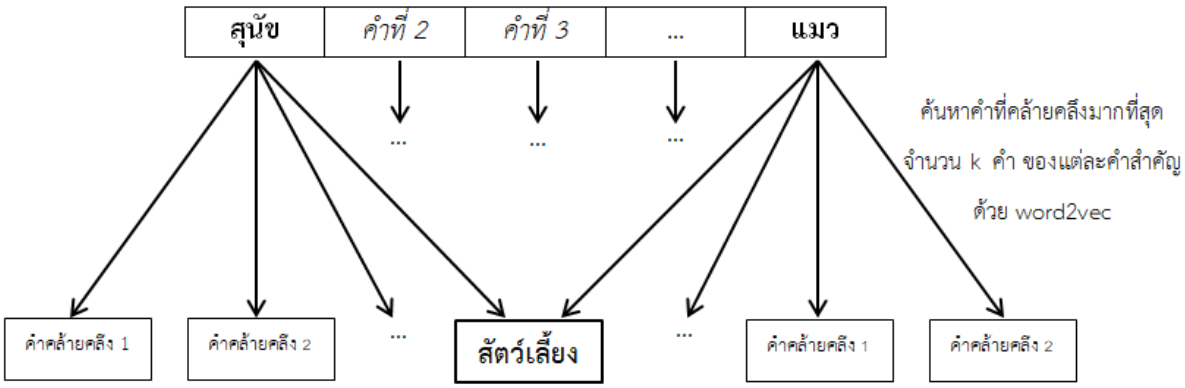
4.2.5.3. ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะและการเพิ่มคุณลักษณะที่เป็นคำคล้ายคลึง(Balanced Dataset using Adding Similar Words+CHI2+SMOTE)

ในการทดลองนี้ผู้วิจัยได้เสนอวิธีการใหม่ เนื่องจากตัวอย่างที่เป็นคลาสบวก (คลาสที่เป็นสถานะใดสถานะหนึ่งตามตัวแบบ AISAS) มีจำนวนน้อยจึงอาจทำให้คุณลักษณะที่มีผลต่อการเกิดคลาสบวกเป็นคำที่อยู่แยกกัน ทำให้ตัวแบบไม่พบแพทเทิร์นของคำที่มีผลต่อการเกิดคลาสบวก ดังตัวอย่างเวกเตอร์ของเอกสารต่อไปนี้

สุนัข	โคมไฟ	บ้าน	ประหยัด	ดีไซน์	สว่าง	แมว	คลาส
1	1	0	1	0	1	0	+
0	1	1	0	1	0	0	-
0	0	1	1	0	1	0	-
0	0	1	0	1	1	**1**	+
0	1	0	1	0	0	0	-
0	1	1	1	0	1	0	-
0	0	0	1	0	1	0	-

ภาพที่ 12 คลาสของเวกเตอร์ของเอกสารจำนวนเจ็ดตัวอย่าง

จากตัวอย่างซึ่งประกอบด้วยตัวอย่างเวกเตอร์ของคำจำนวนเจ็ดตัวอย่าง จะพบว่าคำว่า *สุนัข* และ *แมว* ซึ่งเป็นคำที่มีความคล้ายคลึงกันในเชิงความหมายน่าจะเป็นคำที่มีผลต่อการเกิดคลาสบวก แต่เนื่องจากคำดังกล่าวอยู่แยกเป็นคนละคุณลักษณะทำให้ตัวแบบจำแนกประเภทไม่พบแพทเทิร์นของคำที่มีผลต่อการเกิดคลาสบวก ผู้วิจัยจึงเพิ่มคุณลักษณะซึ่งเป็นคำใหม่ที่คล้ายคลึงกับคำสำคัญที่ถูกคัดเลือกโดยเทคนิคคัดเลือกคุณลักษณะก่อนการใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย โดยเทคนิคการแปลงคำเป็นเวกเตอร์จะถูกประยุกต์ใช้เพื่อหาคำที่คล้ายคลึงกับคำที่สำคัญ ดังต่อไปนี้



ภาพที่ 13 การสร้างคุณลักษณะซึ่งเป็นคำใหม่โดยประยุกต์ใช้วิธีการหาความคล้ายคลึงระหว่างเวกเตอร์ของคำ

แต่ละคำสำคัญจะถูกนำไปหาคำที่คล้ายคลึงมากที่สุดจำนวน k คำ ซึ่งในงานวิจัยนี้ ผู้วิจัยเลือกค่า k เท่ากับ 5 วิธีการเริ่มจากแปลงคำให้อยู่ในรูปเวกเตอร์ (ด้วยเทคนิค word2vec) ก่อน แล้วจึงใช้วิธีการหาความคล้ายคลึงระหว่างเวกเตอร์ของคำเพื่อค้นหาคำที่คล้ายคลึงมากที่สุด เมื่อได้คำคล้ายคลึงสูงสุดห้าอันดับแรกของแต่ละคำแล้วจะทำการค้นหาคำคล้ายคลึงที่เกิดซ้ำแล้วนำคำดังกล่าวมาเพิ่มเป็นคุณลักษณะใหม่

สุนัข	โคมไฟ	บ้าน	ประหยัด	ดีไซน์	สว่าง	แมว	**สัตว์เลี้ยง**	คลาส
1	1	0	1	0	1	0	**1**	+
0	1	1	0	1	0	0	0	-
0	0	1	1	0	1	0	0	-
0	0	1	0	1	1	**1**	**1**	+
0	1	0	1	0	0	0	0	-
0	1	1	1	0	1	0	0	-
0	0	0	1	0	1	0	0	-

ภาพที่ 14 คลาสของเวกเตอร์ของเอกสารจำนวนเจ็ดตัวอย่างที่ถูกเพิ่มคำซึ่งเป็นคุณลักษณะใหม่

โดยจะให้ค่าในเวกเตอร์ของคุณลักษณะใหม่เป็น 1 หากตัวอย่างนั้นมีคำที่สร้างคำใหม่นี้ขึ้นมาปรากฏอยู่ (ค่าในเวกเตอร์เป็น 1) และเป็น 0 หากตัวอย่างนั้นไม่ปรากฏคำที่สร้างคำนี้ขึ้นมา (ค่าในเวกเตอร์เป็น 0) ก่อนที่จะผ่านเข้าสู่ขั้นตอนการใช้ TF-IDF เช่นเดิมในตอนถัดไป โดยคุณลักษณะใหม่ดังกล่าวจะประกอบด้วยคุณลักษณะสำคัญที่ทำให้เกิดแพทเทิร์นของตัวอย่างซึ่งจะมีผลให้ตัวแบบจำแนกประเภทสามารถแยกแยะคลาสของตัวอย่างได้ดีขึ้น

อย่างไรก็ตามคำที่สร้างขึ้นมาใหม่อาจประกอบด้วยทั้งคุณลักษณะที่เป็นประโยชน์และไม่เป็นประโยชน์ต่อการจำแนกคลาสของข้อมูล การคัดเลือกคุณลักษณะในขั้นตอนถัดไปจึงจะเป็นการกรองคุณลักษณะที่ไม่เกี่ยวข้องออกไปอีกครั้งหนึ่งก่อนการใช้วิธีสุ่มเพิ่มตัวอย่างกลุ่มน้อย

โดยขั้นตอนวิธีการในการสร้างคำคล้ายคลึงที่เกิดซ้ำและนำคำดังกล่าวมาเพิ่มเป็นคุณลักษณะใหม่เป็นไปตามขั้นตอนวิธี (algorithm) ดังต่อไปนี้

Generate similar words;
inputs: terms: *all uqnique words*, k: *selected top k rank*
dictionary ← { }
n ← length(terms) − 1
for i ← 0 **to** n
    similar_words ← generate_top_similar_words(terms[i], k)
    for each similar_word **in** similar_words
        key ← similar_word
        if key **in** dictionary
            add i in dictionary[key] list
        else
            create empthy dictionary[key] list
            add i in dictionary[key] list
remove elements which length(element) < 2 from dictionary
Add similar words to the document vectors;
for each document_vector **in** document_vectors
    for each original_word_index_list **in** dictionary
        value ← 0
        for each word_index **in** original_word_index_list
            if terms[word_index] = 1
                value ← 1
        add value to end of document_vector

ตัวอย่างคำคล้ายคลึงที่เกิดซ้ำซึ่งถูกนำมาเพิ่มเป็นคุณลักษณะใหม่ที่สร้างได้ เป็นดังนี้

ตารางที่ 7 คำคล้ายคลึงที่เกิดซ้ำซึ่งถูกนำมาเพิ่มเป็นคุณลักษณะใหม่ ในโฆษณาประเภทเครื่องสำอาง

คำคล้ายคลึงที่เกิดซ้ำ	คำที่สร้างคำคล้ายคลึงร่วมกัน
เทิด	น้อม, เฉลิมพระเกียรติ
ทำนุบำรุง	น้อม, เสริมสร้าง
น้อยลง	น้อย, เยอะ
มีเสน่ห์	น่ารัก, รวย, อ่อนโยน, เซ็กซี่
สมอง	ประสาท, ร่างกาย, หัวใจ
เสี่ย	น้า, แก๊ง
ป้า	น้า, พี่
ไอศกรีม	น้ำหอม, เมนู
เครื่องสำอาง	น้ำหอม, ผลิตภัณฑ์, แบรนด์, แฟชั่น
ขนมหวาน	น้ำหอม, เมนู

ตารางที่ 8 คำคล้ายคลึงที่เกิดซ้ำซึ่งถูกนำมาเพิ่มเป็นคุณลักษณะใหม่ ในโฆษณาประเภทเครื่องใช้ไฟฟ้า

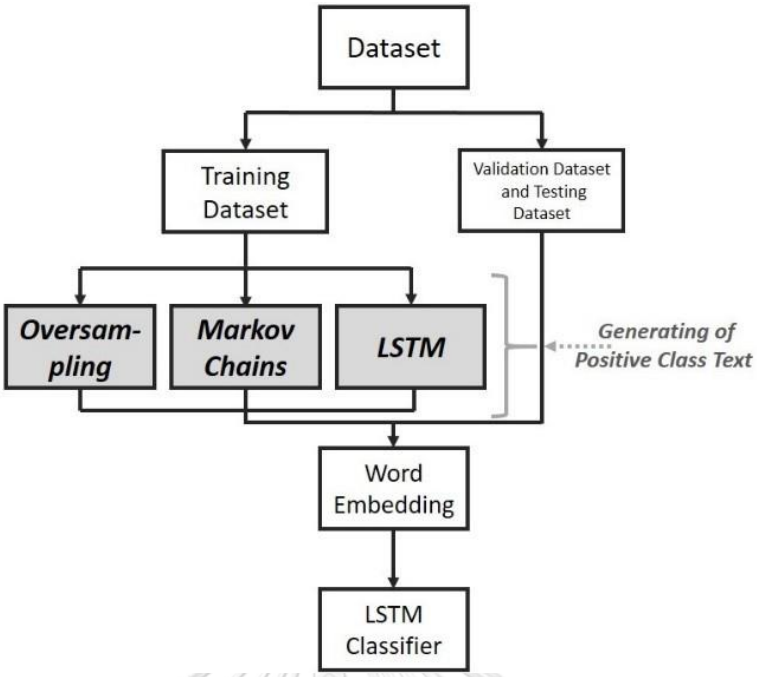
คำคล้ายคลึงที่เกิดซ้ำ	คำที่สร้างคำคล้ายคลึงร่วมกัน
เหมาะสม	ถูกต้อง, ปลอดภัย, สมควร, เหมาะ
ผิดพลาด	ถูกต้อง, ผิด
พึงพอใจ	ถูกใจ, ผ่อนคลาย, ลังเล, แปลกใจ, ไว้วางใจ
ดึงดูดใจ	ถูกใจ, สะดุดตา, สิ้นเปลือง, อวด
อิจฉา	ถูกใจ, เป็นห่วง, แปลกใจ
ถ่ายแบบ	ถ่าย, โฆษณา
ถ่ายรูป	ถ่าย, ถ่ายภาพ, อาบน้ำ
กัดกร่อน	ถ่ายเท, รั่ว, เผาไหม้, เรืองแสง
ตกตะกอน	ถ่ายเท, เผาไหม้, เรืองแสง
ทัวร์นาเมนต์	ถ้วย, นัด, รายการ, โซน

ตารางที่ 9 คำคล้ายคลึงที่เกิดซ้ำซึ่งถูกนำมาเพิ่มเป็นคุณลักษณะใหม่ ในโฆษณาประเภทสุขภัณฑ์

คำคล้ายคลึงที่เกิดซ้ำ	คำที่สร้างคำคล้ายคลึงร่วมกัน
มีประสิทธิภาพ	ทันสมัย, ทันสมัย, ประสิทธิภาพ, ปลอดภัย, รวดเร็ว
ข้าวเหนียว	กระเทียม, พริก, เนย, โรย
กะทิ	กระเทียม, พริก, มะนาว
สังกะสี	กระเบื้อง, ทองเหลือง, พลาสติก, อะลูมิเนียม
ศิลาแลง	กระเบื้อง, ปูน, หินอ่อน, อิฐ
ทาสี	กระเบื้อง, มุง
ถือปูน	กระเบื้อง, ปูน
เข็มขัด	กระเป๋า, เก้าอี้, แหวน
เก้าอี้	กระเป๋า, บันได, ป้าย, โต๊ะ
กล่อง	กระเป๋า, ขวด, ตู้, ถุง, ฝา, เครื่อง, แผ่น

4.2.5.4. ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคการสร้างข้อความ (Balanced Dataset using Text Generation)

ในการทดลองสุดท้าย ผู้วิจัยนำวิธีการสร้างข้อความมาประยุกต์ใช้เพื่อเพิ่มจำนวนตัวอย่างกลุ่มน้อยให้มีจำนวนเท่ากับตัวอย่างกลุ่มมาก ผู้วิจัยเปรียบเทียบวิธีการสร้างข้อความสามแบบ ได้แก่ (1) การสุ่มเพิ่มตัวอย่างแบบดั้งเดิม (traditional over-sampling) ซึ่งจะถูกใช้เป็นบรรทัดฐาน (baseline) (2) การสร้างข้อความแบบลูกโซ่แบบมาร์คอฟ และ (3) การสร้างข้อความด้วยแอลเอสทีเอ็ม โดยใช้แอลเอสทีเอ็มเป็นตัวแบบสำหรับจำแนกประเภท



ภาพที่ 15 ภาพรวมของขั้นตอนการทดลองแบบสุดท้าย (1.4)

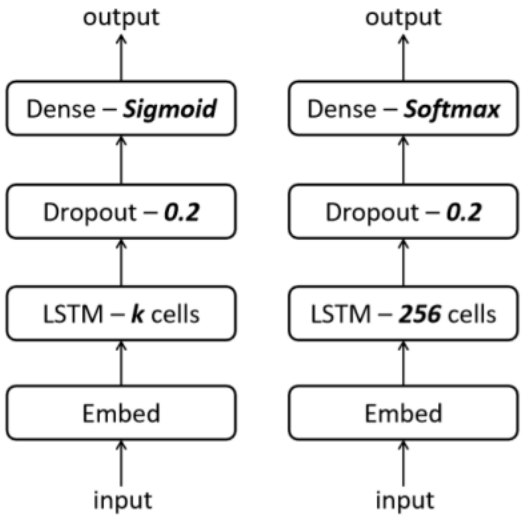
ตารางที่ 10 แสดงรายละเอียดเกี่ยวกับการแบ่งข้อมูลก่อนป้อนเข้าสู่ตัวแบบ

ชุดข้อมูล	จำนวนคำที่ไม่ซ้ำ	จำนวนตัวอย่างสำหรับสอน	จำนวนตัวอย่างสำหรับทดสอบ	จำนวนตัวอย่างสำหรับvalidate
โฆษณากลุ่มเครื่องสำอาง				
Attention	4042	600	200	199
Interest				
Search				
Action				
Share				
โฆษณากลุ่มเครื่องใช้ไฟฟ้า				
Attention	4924	600	183	183
Interest				
Search				
Action				
Share				
โฆษณากลุ่มสุขภัณฑ์				
Attention	5018	600	182	181
Interest				
Search				
Action				
Share				

โดยเริ่มแรก ชุดข้อมูลจะถูกแบ่งเป็นข้อมูลสอน, ข้อมูลสำหรับตรวจสอบ (validation dataset) และข้อมูลทดสอบ ดังแสดงในตารางก่อนหน้า จากนั้นวิธีการสร้างข้อความสามแบบ ได้แก่ (1) การสุ่มเพิ่มตัวอย่างแบบดั้งเดิม (traditional over-sampling) (2) การสร้างข้อความแบบลูกโซ่แบบมาร์คอฟ และ (3) การสร้างข้อความด้วยแอลเอสทีเอ็ม จะถูกประยุกต์ใช้กับตัวอย่างข้อความเฉพาะในข้อมูลสอนที่เป็นคลาสบวกเพื่อเพิ่มจำนวนข้อความที่เป็นคลาสบวกจนมีจำนวนเท่ากับข้อความที่เป็นคลาสลบ ซึ่งโครงสร้างของเครือข่ายแอลเอสทีเอ็มที่ใช้สร้างข้อความจะประกอบไปด้วยหน่วย

ของแอลเอสทีเอ็มทั้งหมด 256 หน่วยและใช้ฟังก์ชันซอฟต์แม็กซ์ (Softmax function) เป็นฟังก์ชันกระตุ้น (activation function)

หลังจากชุดข้อมูลถูกทำให้สมดุลด้วยเทคนิคต่าง ๆ ชุดข้อมูลจะถูกนำไปป้อนเข้าสู่ตัวแบบสำหรับจำแนกประเภทซึ่งเป็นเครือข่ายแอลเอสทีเอ็มเช่นกัน โดยโครงสร้างของเครือข่ายแอลเอสทีเอ็มที่ใช้จำแนกประเภทประกอบไปด้วยหน่วยของแอลเอสทีเอ็มทั้งหมด k หน่วยซึ่งแตกต่างกันไปในแต่ละชุดข้อมูลและใช้ฟังก์ชันซิกมอยด์ (Sigmoid function) เป็นฟังก์ชันกระตุ้น



ภาพที่ 16 (ซ้าย) โครงสร้างของเครือข่ายแอลเอสทีเอ็มสำหรับจำแนกประเภท และ (ขวา) โครงสร้างของเครือข่ายแอลเอสทีเอ็มสำหรับการสร้างข้อความ

ตารางที่ 11 จำนวนหน่วยของแอลเอสทีเอ็มและจำนวนรอบ (EPOCHS) ที่ใช้สำหรับแต่ละชุดข้อมูล

ชุดข้อมูล	จำนวนหน่วยแอลเอสทีเอ็ม	จำนวน epochs
โฆษณากลุ่มเครื่องสำอาง		
Attention	16	10
Interest	128	100
Search	128	100
Action	8	10
Share	128	100

ชุดข้อมูล	จำนวนหน่วยแอลเอสทีเอ็ม	จำนวน epochs
โฆษณากลุ่มเครื่องใช้ไฟฟ้า		
Attention	128	10
Interest	128	100
Search	32	10
Action	8	10
Share	128	100
โฆษณากลุ่มสุขภัณฑ์		
Attention	128	10
Interest	128	100
Search	128	100
Action	8	10
Share	128	100

ตัวอย่างข้อความสังเคราะห์ที่ถูกสร้างขึ้นด้วยเทคนิคการสร้างข้อความแบบลูกโซ่แบบมาร์คอฟเป็นดังนี้

“ลง ตัว ที่สุด ของ หิน อ่อน สี สัน อันดุดัน เข้มแข็ง สอด ประสาน กับ บรรยากาศ สบาย ๆ ไป กับ ความ ร่วมสมัย ง่าย ใช้ กับ รูฝักบัวแบบ ใหญ่ ค่อน มา ทาง ทิศ ตะวัน ตก แต่ง ห้อง น้ำ ต่ำ 2 ก๊อก น้ำ สวย ด้วย ดีไซน์วัสดุปิด ผิว สัมผัส ด้วย ดีไซน์วัสดุปิด ผิว ลาย เรขาคณิต และ ช่วย ให้ กับ บรรยากาศ ของ ลวดลาย นำ เอา เอกลักษณ์ เฉพาะ ตัว Italia Collection รุ่น ใหม่ ที่ 29 30 เมษายน – 1 พฤษภาคม 2559) ด้วย โทนสี ชุด ครัว สำหรับ วาง ผัง บ้าน หรือ อาคาร ก็ สวย ของ ทำ ผนัง ตาม สไตล์ คุณ”

ตัวอย่างข้อความสังเคราะห์ที่ถูกสร้างขึ้นด้วยเทคนิคการสร้างข้อความด้วยแอลเอสทีเอ็มเป็นดังนี้

“แบบ ห้องน้ำ ด้วย ความ รู้สึก เคย สังเกต ไหม ว่า ทำไม ใคร หลาย คน ชอบ แอบ ไป ร้องไห้ ใน ห้องน้ำ เหตุผล ก็ เพราะ ห้องน้ำ เป็น ทั้ง มุม สงบ และ เป็น พื้นที่ ส่วน ตัว ที่ ทำ ให้ เรา ได้ ที่ อยู่ ที่ ช่วย ความ รู้สึก ผ่อนคลาย เสมือน และ ได้ แรง บันดาลใจ ใหม่ ๆ และ หน้า ให้ ออก แบบ ส่วน โค้ง ด้าน หน้า ให้ เข้า งาน สรีระ ของ ดู สูง อายุ และ ผู้ ใช้ วีลแชร์ อย่าง ใช้ ใน งาน มุม โดย นัก ออก แบบ ห้องน้ำ ที่ โดย กลาง สามารถ บริการ ออกแบบ ห้องน้ำ หลากสไตล์ จาก เรา ได้ ที่ ลิงก์ โทร เบอร์โทร โทร เบอร์โทร ของ”

บทที่ 4 ผลการทดลอง

4.1 ชุดข้อมูลที่ไม่สมดุล (Original Imbalanced Dataset)

ชุดข้อมูลดั้งเดิมซึ่งเป็นชุดข้อมูลที่ไม่สมดุลได้ถูกป้อนเข้าสู่ตัวแบบ นาอีฟเบย์, การถดถอยโลจิสติกส์, และซัพพอร์ตเวกเตอร์แมชชีน เทคนิคไขว้ข้ามห้ากลุ่ม (5-fold cross validation) ได้ถูกนำมาใช้ประเมินตัวแบบ ได้ผลลัพธ์ของแต่ละตัวแบบตามตารางต่อไปนี้

ตารางที่ 12 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มเครื่องสำอาง

	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
ตัวแบบ	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
Attention	91.09	0	0	0	96.40	0	0	0	96.40	0	0	0
Interest	64.46	22.45	34.20	27.11	80.58	33.33	0.52	1.02	80.08	12.50	0.52	1
Search	82.08	7.34	9.30	8.21	91.39	0	0	0	91.29	0	0	0
Action	96.90	15.39	9.09	11.43	97.80	0	0	0	97.80	0	0	0
Share	91.29	40	41.67	40.82	95.40	100	36.11	53.06	95.40	100	36.11	53.06

ตารางที่ 13 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มเครื่องใช้ไฟฟ้า

	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
ตัวแบบ	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
Attention	94.62	0	0	0	96.48	0	0	0	96.48	0	0	0
Interest	71.95	20	18.29	19.10	81.78	0	0	0	81.88	50	2.29	4.37
Search	87.06	7.84	4.88	6.02	91.51	0	0	0	91.30	0	0	0
Action	97.83	0	0	0	98.24	0	0	0	98.24	0	0	0
Share	85.82	9.23	7.14	8.05	91.30	0	0	0	91.30	0	0	0

ตารางที่ 14 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มสุขภัณฑ์

	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
ตัวแบบ	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
Attention	93.35	7.41	4.88	5.88	95.74	0	0	0	95.64	0	0	0
Interest	67.83	21.82	22.97	22.38	79.81	0	0	0	80.19	64.29	4.31	8.07
Search	89.86	8.11	4.05	5.41	92.85	0	0	0	92.85	0	0	0
Action	98.36	0	0	0	98.55	0	0	0	98.55	0	0	0
Share	88.89	12	8.70	10.08	92.84	0	0	0	92.94	100	1.45	2.86

จากผลลัพธ์ของทุกตัวแบบ เห็นได้ว่าค่าความถูกต้องของชุดข้อมูลดั้งเดิมซึ่งไม่สมดุลนั้นสูง แต่ค่าความแม่น ค่าระลึก และคะแนนเอฟวันต่ำมากเนื่องจากอคติของตัวแบบที่มีโนวแน้มทำนายเพียงแต่คลาสกลุ่มมากซึ่งเป็นคลาสลบ อย่างไรก็ตามในกรณีที่ไม่ได้มีการนำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ เมื่อชุดข้อมูลฝึกไม่สมดุล นาอีฟเบย์ให้ผลลัพธ์ในแง่ของค่าความแม่น ค่าระลึก และคะแนนเอฟวัน สูงกว่าทั้งวิธีการถอดถอยโลจิสติกส์และซัพพอร์ต-เวกเตอร์แมชชีน และตัวแบบ Interest ให้ผลลัพธ์ดีที่สุดเมื่อเทียบกับตัวแบบอื่นเนื่องจากชุดข้อมูล Interest (ที่เป็นชุดข้อมูลดั้งเดิม) มีความไม่สมดุลต่ำที่สุด

4.2 ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะ (Balanced Dataset using CHI2+SMOTE)

ชุดข้อมูลต่อมาได้ใช้เทคนิคคัดเลือกคุณลักษณะแบบไคกำลังสอง โดยแบ่งจำนวนคุณลักษณะสำคัญหลายระดับ ได้แก่ 25 50 100 250 500 1000 และทุกคุณลักษณะ ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยก่อนป้อนเข้าสู่ตัวแบบ เทคนิคไขว้ข้ามห้ากลุ่มได้ถูกนำมาใช้ประเมินตัวแบบ ได้ผลลัพธ์ของแต่ละตัวแบบตามตารางต่อไปนี้

ตารางที่ 15 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มเครื่องสำอางหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ตัวแบบ Attention												
25	95.40	**18.75**	8.33	**11.54**	95.40	**29.17**	19.44	23.33	96	33.33	11.11	16.67
50	94.19	10.71	8.33	9.38	94.80	25	**22.22**	**23.53**	96.30	**46.15**	**16.67**	**24.49**
100	93.59	11.11	**11.11**	11.11	94.29	18.18	16.67	17.39	95	18.18	11.11	13.79
250	93.19	7.89	8.33	8.11	91.99	2.17	2.78	2.49	91.69	2.04	2.78	2.35
500	81.58	1.95	8.33	3.16	90.39	0	0	0	91.93	0	0	0
1000	89.79	2.86	5.56	3.77	90.89	0	0	0	91.79	0	0	0
ทั้งหมด	91.09	0	0	0	93.69	0	0	0	92.29	0	0	0
ตัวแบบ Interest												
25	77.58	13.95	3.11	5.09	61.76	20.38	33.68	25.39	71.67	20.39	16.06	17.97
50	76.18	13.11	4.15	6.30	60.96	19.12	31.61	23.83	70.47	18.90	16.06	17.37
100	74.37	16.13	7.77	10.49	62.86	21.29	34.20	26.24	70.77	22.65	21.24	21.93
250	70.17	15.23	11.92	13.37	64.46	**25.15**	**42.49**	**31.60**	68.87	25.21	31.09	27.84
500	67.87	17.68	18.13	17.90	62.36	23.01	40.41	29.32	61.46	23.18	**43.01**	30.13
1000	71.67	21.88	18.13	19.83	67.97	24.29	31.09	27.27	69.07	**26.98**	35.23	**30.56**
ทั้งหมด	64.46	**22.45**	**34.20**	**27.11**	67.97	24.90	32.64	28.25	69.07	26.03	32.64	28.97
ตัวแบบ Search												
25	89.99	15	3.49	5.66	89.19	**22.50**	10.47	14.29	90.59	25	4.65	7.84
50	89.59	**15.38**	4.65	7.14	87.29	21.13	17.44	19.11	90.39	**27.27**	6.98	11.11
100	87.49	13.21	8.14	10.07	85.59	19.79	**22.09**	**20.88**	90.09	19.05	4.65	7.48
250	84.08	10.75	11.63	11.17	76.38	9.68	20.93	13.24	78.18	8.23	15.12	10.66
500	84.79	10.71	10.47	10.59	80.38	10.71	17.44	13.27	81.38	11.54	**17.44**	**13.89**
1000	76.38	8.79	**18.60**	**11.94**	84.38	11.96	12.79	12.36	82.78	10.19	12.79	11.34
ทั้งหมด	82.08	7.34	9.30	8.21	82.98	11.82	15.12	13.27	84.08	12.37	13.95	13.12
ตัวแบบ Action												
25	97	**21.43**	13.64	**16.67**	96.50	**15.79**	**13.64**	**14.63**	97.60	**33.33**	9.09	**14.29**

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
50	96.40	15	13.64	14.29	96.20	13.64	**13.64**	13.64	97.40	25	9.09	13.33
100	95.20	11.76	**18.18**	14.29	95.60	10.71	**13.64**	12	95.30	9.68	**13.64**	11.32
250	95.20	9.38	13.64	11.11	95.30	6.90	9.09	7.84	96.40	11.11	9.09	10
500	91.99	3.23	9.09	4.76	94.90	6.06	9.09	7.27	96.70	13.33	9.09	10.81
1000	96	9.09	9.09	9.09	95	6.25	9.09	7.41	96.70	13.33	9.09	10.81
ทั้งหมด	96.90	15.38	9.09	11.43	95	6.25	9.09	7.41	97	16.67	9.09	11.77
ตัวแบบ Share												
25	93.19	**53.45**	43.06	**47.69**	93.79	60.42	40.28	48.33	95.40	**100**	36.11	53.06
50	92.19	45.59	43.06	44.29	93.79	60.87	38.89	47.46	95.50	**100**	37.50	**54.55**
100	90.09	36.63	51.39	42.78	93.49	55.93	**45.83**	50.38	95.30	90.32	38.89	54.37
250	88.49	31.93	**52.78**	39.79	93.79	58.93	**45.83**	51.56	93.89	60.38	**44.44**	51.20
500	89.99	36	50	41.86	93.39	55.36	43.06	48.44	92.99	51.72	41.67	46.15
1000	87.59	27.97	45.83	34.74	93.59	57.14	44.44	50	93.69	59.18	40.28	47.93
ทั้งหมด	91.29	40	41.67	40.82	95.60	**86.84**	**45.83**	**60**	93.89	61.70	40.28	48.74

ตารางที่ 16 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มเครื่องใช้ไฟฟ้าหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ตัวแบบ Attention												
25	94.41	0	0	0	95.24	**26.92**	**20.59**	**23.33**	95.76	**11.11**	**2.94**	**4.65**
50	93.89	0	0	0	94.72	22.58	**20.59**	21.54	95.65	0	0	0
100	85.20	**1.77**	**5.88**	**2.72**	92.24	6.38	8.82	7.41	93.79	0	0	0
250	83.44	1.54	**5.88**	2.44	94.51	0	0	0	95.55	0	0	0
500	87.16	1.09	2.94	1.59	95.55	0	0	0	95.76	0	0	0
1000	92.13	0	0	0	95.65	0	0	0	95.76	0	0	0
ทั้งหมด	94.62	0	0	0	95.24	0	0	0	95.76	0	0	0
ตัวแบบ Interest												

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
25	**80.44**	33.33	8	12.90	78.36	**31.11**	16	21.13	81.78	46.15	3.43	6.38
50	76.61	26.17	16	19.86	74.43	27.78	25.71	26.71	81.88	**50**	5.14	9.33
100	75.98	26.45	18.29	21.62	68.43	23.14	32	26.86	78.47	28	12	16.80
250	73.29	23.90	21.71	22.75	66.87	23.64	37.14	28.89	76.09	27.78	20	23.26
500	68.74	21.52	**27.43**	**24.12**	69.46	26.92	**40**	**32.18**	71.64	25.13	**28.57**	**26.74**
1000	70.19	21.32	24	22.58	72.67	25.41	26.29	25.84	72.88	24.86	24.57	24.71
ทั้งหมด	71.95	20	18.29	19.10	74.22	26.88	24.57	25.67	75.47	25.40	18.29	21.26
ตัวแบบ Search												
25	88.61	3.33	1.22	1.79	84.78	19.63	25.61	22.22	88.92	17.95	8.54	11.57
50	87.68	6.98	3.66	4.80	83.54	**19.69**	**30.49**	**23.92**	88.92	**22.22**	12.20	15.75
100	85.82	6.35	4.88	5.52	80.85	15.89	29.27	20.60	87.99	18.52	12.20	14.71
250	83.95	**9.89**	10.98	10.41	82.61	15.32	23.17	18.45	85.61	16.47	**17.07**	**16.77**
500	84.47	9.52	9.76	9.64	85.20	9.33	8.54	8.92	86.23	6.78	4.88	5.67
1000	79.50	9.72	**17.07**	**12.39**	86.03	9.23	7.32	8.16	87.58	10.42	6.10	7.69
ทั้งหมด	87.06	7.84	4.88	6.02	85.51	7.35	6.10	6.67	87.47	6.67	3.66	4.72
ตัวแบบ Action												
25	97.10	**13.33**	11.76	12.50	96.89	**11.76**	**11.76**	**11.77**	97.52	**11.11**	5.88	7.69
50	95.86	10.34	**17.65**	**13.04**	95.65	6.90	**11.76**	8.70	95.65	9.68	**17.65**	**12.50**
100	95.14	8.33	**17.65**	11.32	97.72	0	0	0	98.03	0	0	0
250	95.34	8.82	**17.65**	11.77	97.72	0	0	0	98.14	0	0	0
500	95.86	4	5.88	4.76	97.62	0	0	0	98.14	0	0	0
1000	97.62	0	0	0	97.62	0	0	0	98.14	0	0	0
ทั้งหมด	97.83	0	0	0	97.62	0	0	0	98.03	0	0	0
ตัวแบบ Share												
25	89.03	10.71	3.57	5.36	86.65	9.09	5.95	7.19	89.96	6.67	1.19	2.02
50	88.41	11.11	4.76	6.67	84.27	10.47	**10.71**	**10.59**	89.13	4.35	1.19	1.87
100	88.10	**18.37**	10.71	13.53	83.13	9.28	**10.71**	9.95	89.03	4.17	1.19	1.85
250	82.92	12.84	16.67	14.51	83.02	8.33	9.52	8.89	87.89	7.69	3.57	4.88
500	80.23	12.06	**20.24**	**15.11**	87.06	12.73	8.33	10.07	86.44	10.17	**7.14**	**8.39**

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
1000	84.06	11.11	11.90	11.49	88.72	9.68	3.57	5.22	88.92	10.34	3.57	5.31
ทั้งหมด	85.82	9.23	7.14	8.05	87.68	**14.29**	8.33	10.53	88.20	**10.53**	4.76	6.56

ตารางที่ 17 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มสุขภัณฑ์หลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ตัวแบบ Attention												
25	93.56	**8**	4.88	6.06	92.32	9.76	**9.76**	9.76	95.12	20	4.88	7.84
50	92.32	7.69	7.32	**7.50**	91.69	6.67	7.32	6.98	94.91	**21.43**	**7.32**	**10.91**
100	89.62	4.62	7.32	5.66	90.76	7.14	**9.76**	8.25	93.25	10	**7.32**	8.45
250	88.27	5	**9.76**	6.61	92.52	**10.26**	**9.76**	**10**	93.56	11.11	**7.32**	8.82
500	90.97	5.77	7.32	6.45	93.15	9.68	7.32	8.33	93.35	10.34	**7.32**	8.57
1000	90.45	5.26	7.32	6.12	93.04	9.38	7.32	8.22	92.73	8.57	**7.32**	7.90
ทั้งหมด	93.35	7.41	4.88	5.88	93.25	10	7.32	8.45	93.35	10.34	**7.32**	8.57
ตัวแบบ Interest												
25	75.36	**29.82**	16.27	21.05	34.20	22.37	**91.39**	35.94	28.02	21.85	**99.52**	35.83
50	73.33	24.81	15.79	19.30	48.70	24.28	72.73	**36.41**	29.95	22.02	97.13	**35.90**
100	71.69	25.29	20.57	22.69	61.26	**26.24**	50.72	34.58	49.95	23.86	67.46	35.25
250	69.47	23.65	22.97	23.30	57.97	23.96	49.76	32.35	57.10	23.60	50.24	32.11
500	65.80	22.22	27.75	24.68	59.52	25.58	52.63	34.43	60	**25.88**	52.63	34.70
1000	63.38	22.22	**32.54**	**26.41**	67.83	25.59	31.10	28.08	67.54	25.67	32.06	28.51
ทั้งหมด	67.83	21.82	22.97	22.38	66.86	22.98	27.27	24.95	68.31	21.53	21.53	21.53
ตัวแบบ Search												
25	90.82	13.79	5.41	7.77	89.86	17.02	10.81	13.22	91.59	0	0	0
50	90.15	18.18	10.81	13.56	88.60	**17.65**	16.22	**16.90**	91.69	12.50	2.70	4.44
100	89.18	**20.31**	17.57	**18.84**	86.18	14.43	18.92	16.37	91.78	**17.65**	4.05	6.59
250	86.28	14.58	18.92	16.47	84.83	13.27	**20.27**	16.04	90.24	13.51	**6.76**	**9.01**

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
500	85.41	14.68	**21.62**	17.49	85.99	6.17	6.76	6.45	88.31	3.92	2.70	3.20
1000	85.31	8.51	10.81	9.52	88.02	10.94	9.46	10.15	89.66	9.76	5.41	6.96
ทั้งหมด	89.86	8.11	4.05	5.41	88.12	8.47	6.76	7.52	90.24	11.43	5.41	7.34
ตัวแบบ Action												
25	97.39	0	0	0	97.39	0	0	0	98.07	0	0	0
50	96.23	**7.14**	13.33	9.30	97.49	0	0	0	97.39	0	0	0
100	94.59	6.38	**20**	**9.68**	97.49	0	0	0	97.78	0	0	0
250	95.17	2.70	6.67	3.85	97.78	0	0	0	97.87	0	0	0
500	96.62	0	0	0	97.68	0	0	0	97.87	0	0	0
1000	97.97	0	0	0	97.68	0	0	0	97.87	0	0	0
ทั้งหมด	98.36	0	0	0	97.68	0	0	0	97.87	0	0	0
ตัวแบบ Share												
25	90.55	10.71	4.35	6.19	89.93	11.11	5.80	7.62	92.42	0	0	0
50	89.41	7.69	4.35	5.56	88.79	8.51	5.80	6.90	92.21	0	0	0
100	85.98	12.50	15.94	14.01	84.94	8.70	**11.59**	9.94	91.90	9.09	1.45	2.50
250	80.58	12.18	27.54	16.89	81.93	5.88	10.14	7.45	89.20	9.30	5.80	7.14
500	78.09	14.50	**42.03**	**21.56**	82.35	6.09	10.14	7.61	84.11	6.25	**8.70**	7.27
1000	82.97	**15.33**	30.43	20.39	87.95	7.27	5.80	6.45	89.93	11.11	5.80	7.62
ทั้งหมด	88.89	12	8.70	10.08	88.06	**11.67**	10.14	**10.85**	90.65	**13.79**	5.80	**8.16**

หลังจากนำวิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ เมื่อจำนวนคุณลักษณะมีความเหมาะสม วิธีการดังกล่าวทำให้ค่าความแม่นและค่าระลึกเพิ่มสูงขึ้น เมื่อค่าความแม่นและค่าระลึกสูงขึ้นส่งผลให้คะแนนเอฟวันเพิ่มสูงขึ้นเนื่องจากคะแนนเอฟวันได้นำทั้งค่าความแม่นและค่าระลึกมาพิจารณาร่วมกัน หลังการประยุกต์ใช้วิธีการดังกล่าว คะแนนเอฟวันของวิธีการถอดถอยโลจิสติกส์และซัพพอร์ตเวกเตอร์แมชชีนเพิ่มขึ้นอย่างมีนัยสำคัญในเกือบทุกตัวแบบ อย่างไรก็ตามนาอีฟเบย์ซึ่งเป็นวิธีการที่ทนทานต่อชุดข้อมูลที่ไม่สมดุลมากที่สุด มีอัตราการเพิ่มขึ้นของคะแนนเอฟวันต่ำที่สุดเมื่อเทียบกับวิธีการอื่น แต่นาอีฟเบย์ยังให้ผลลัพธ์ที่ดีในชุดข้อมูลโฆษณากลุ่มสุขภัณฑ์

สุดท้ายเป็นการแสดงให้เห็นค่าเอ็มซีซี (MCC: Matthews correlation coefficient) หลังจากนำวิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้

ตารางที่ 18 แสดงค่าเอ็มซีซีของโฆษณากลุ่มเครื่องสำอางหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Attention			
25	0.1037	**0.2152**	0.1759
50	0.0648	0.2088	**0.2621**
100	0.0779	**0.1446**	0.1174
250	**0.0458**	-0.0169	-0.0190
500	**-0.0379**	-0.0489	-0.0444
1000	**-0.0110**	-0.0467	-0.0425
ทั้งหมด	-0.0458	**-0.0322**	-0.0400
ตัวแบบ Interest			
25	-0.0288	**0.0183**	0.0115
50	-0.0401	**-0.0034**	-0.0047
100	-0.0259	0.0335	**0.0397**
250	-0.0437	**0.1028**	0.0834
500	-0.0207	0.0670	**0.0732**
1000	0.0283	0.0722	**0.1128**
ทั้งหมด	0.0512	0.0823	**0.0961**
ตัวแบบ Search			
25	0.0326	**0.1011**	0.0746
50	0.0395	**0.1234**	0.0999
100	0.0388	**0.1300**	0.0545
250	**0.0245**	0.0182	-0.0059
500	0.0227	0.0303	**0.0404**
1000	0.0031	**0.0380**	0.0196
ทั้งหมด	-0.0158	0.0403	**0.0440**

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Action			
25	0.1562	0.1289	**0.1649**
50	0.1246	0.1169	**0.1396**
100	**0.1223**	0.0985	0.0912
250	**0.0889**	0.0553	0.0822
500	0.0179	0.0486	**0.0936**
1000	0.0704	0.0502	**0.0936**
ทั้งหมด	0.1031	0.0502	**0.1087**
ตัวแบบ Share			
25	0.4439	0.4623	**0.5865**
50	0.4011	0.4559	**0.5980**
100	0.3816	0.4720	**0.5752**
250	0.3516	**0.4874**	0.4867
500	0.3713	**0.4537**	0.4274
1000	0.2938	**0.4706**	0.4565
ทั้งหมด	0.3613	**0.6123**	0.4682

ตารางที่ 19 แสดงค่าเอ็มซีซีของโฆษณากลุ่มเครื่องใช้ไฟฟ้าหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Attention			
25	-0.0278	**0.2112**	0.0400
50	-0.0311	**0.1883**	-0.0175
100	-0.0346	**0.0351**	-0.0318
250	-0.0424	**0.0271**	-0.0185
500	-0.0428	-0.0185	**-0.0163**
1000	-0.0407	-0.0175	**-0.0163**
ทั้งหมด	-0.0263	-0.0214	**-0.0163**
ตัวแบบ Interest			

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
25	0.0842	**0.1081**	0.0850
50	0.0738	0.1126	**0.1141**
100	**0.0818**	0.0754	0.0745
250	0.0667	0.0904	**0.0972**
500	0.0485	**0.1388**	0.0927
1000	0.0421	**0.0910**	0.0817
ทั้งหมด	0.0218	**0.1013**	0.0732
ตัวแบบ Search			
25	-0.0331	**0.1410**	0.0696
50	-0.0117	**0.1563**	0.1089
100	-0.0203	**0.1144**	0.0876
250	0.0162	**0.0941**	0.0890
500	**0.0115**	0.0088	-0.0156
1000	**0.0185**	0.0072	0.0158
ทั้งหมด	**-0.0055**	-0.0112	-0.0144
ตัวแบบ Action			
25	**0.1105**	0.1018	0.0690
50	**0.1149**	0.0687	0.1096
100	**0.0984**	-0.0097	-0.0061
250	**0.1026**	-0.0097	-0.0043
500	**0.0278**	-0.0106	-0.0043
1000	-0.0106	-0.0106	**-0.0043**
ทั้งหมด	-0.0086	-0.0106	**-0.0061**
ตัวแบบ Share			
25	**0.0124**	0.0034	-0.0090
50	0.0169	**0.0196**	-0.0241
100	**0.0793**	0.0069	-0.0257
250	**0.0525**	-0.0043	-0.0073
500	**0.0493**	0.0352	0.0133

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
1000	**0.0275**	0.0063	0.0103
ทั้งหมด	0.0051	**0.0459**	0.0131

ตารางที่ 20 แสดงค่าเอ็มซีซีของโฆษณากลุ่มสุขภัณฑ์หลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Attention			
25	0.0303	0.0574	**0.0799**
50	**0.0350**	0.0264	0.1033
100	0.0048	0.0355	**0.0510**
250	0.0111	**0.0610**	0.0577
500	0.0179	0.0490	**0.0531**
1000	0.0125	**0.0470**	0.0415
ทั้งหมด	0.0265	0.0510	**0.0531**
ตัวแบบ Interest			
25	0.0844	0.1175	**0.1397**
50	0.0442	0.1260	**0.1298**
100	0.0563	**0.1205**	0.1053
250	0.0425	**0.0798**	0.0736
500	0.0293	0.1132	**0.1183**
1000	0.0327	0.0767	**0.0792**
ทั้งหมด	0.0210	**0.0390**	0.0168
ตัวแบบ Search			
25	0.0438	**0.0836**	-0.0313
50	0.0902	**0.1080**	0.0260
100	**0.1312**	0.0909	0.0526
250	**0.0922**	0.0832	0.0476
500	**0.1003**	-0.0110	-0.0285
1000	0.0167	**0.0377**	0.0205

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ทั้งหมด	0.0072	0.0126	**0.0311**
ตัวแบบ Action			
25	-0.0131	-0.0131	**-0.0084**
50	**0.0794**	-0.0126	-0.0131
100	**0.0900**	-0.0126	-0.0107
250	**0.0202**	-0.0107	-0.0100
500	-0.0170	-0.0114	**-0.0100**
1000	**-0.0093**	-0.0114	-0.0100
ทั้งหมด	**-0.0053**	-0.0114	-0.0100
ตัวแบบ Share			
25	0.0238	**0.0302**	-0.0179
50	0.0042	**0.0118**	-0.0220
100	**0.0656**	0.0193	0.0080
250	**0.0855**	-0.0187	0.0179
500	**0.1456**	-0.0154	-0.0118
1000	**0.1289**	0.0010	0.0302
ทั้งหมด	0.0439	0.0450	**0.0453**

4.3 ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะและการเพิ่มคุณลักษณะที่เป็นคำคล้ายคลึง (Balanced Dataset using Adding Similar Words+CHI2+SMOTE)

ชุดข้อมูลต่อมาได้ทำการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึง ก่อนการใช้เทคนิคคัดเลือกคุณลักษณะแบบไคกำลังสอง โดยแบ่งจำนวนคุณลักษณะสำคัญหลายระดับ ได้แก่ 25 50 100 250 500 1000 และทุกคุณลักษณะ ร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยก่อนป้อนเข้าสู่ตัวแบบ เทคนิคไขว้ข้ามห้ากลุ่มได้ถูกนำมาใช้ประเมินตัวแบบ ได้ผลลัพธ์ของแต่ละตัวแบบตามตารางต่อไปนี้

ตารางที่ 21 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มเครื่องสำอางหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ตัวแบบ Attention												
25	95.30	**23.81**	**13.89**	**17.54**	95	**28.12**	**25**	**26.47**	96.50	**57.14**	**11.11**	**18.61**
50	93.89	12.12	11.11	11.59	94.49	24.32	**25**	24.66	96.30	42.86	8.33	13.95
100	92.39	6.52	8.33	7.32	93.99	20	22.22	21.05	95.80	25	8.33	12.5
250	91.19	5.17	8.33	6.38	89.99	6.76	13.89	9.09	91.19	5.17	8.33	6.38
500	90.49	1.64	2.78	2.06	90.49	3.17	5.56	4.04	90.29	3.08	5.56	3.96
1000	85.49	0.9	2.78	1.36	90.19	0	0	0	91.59	0	0	0
ทั้งหมด	92.39	0	0	0	94.09	4	2.78	3.28	92.29	0	0	0
ตัวแบบ Interest												
25	77.58	17.02	4.15	6.67	65.07	21.32	30.05	24.95	75.48	23.47	11.92	15.81
50	76.38	17.91	6.22	9.23	64.36	22.74	35.23	27.64	74.88	**26.23**	16.58	20.32
100	74.88	17.78	8.29	11.31	64.06	22.88	36.27	28.06	74.27	25.38	17.1	20.43
250	71.17	15.83	11.4	13.25	64.46	**25.45**	**43.52**	**32.12**	71.37	25.91	25.91	25.91
500	65.27	17.09	20.73	18.74	60.16	21.45	39.9	27.9	62.26	22.46	**38.86**	28.46
1000	69.87	20.97	20.21	20.58	65.87	23.38	33.68	27.6	65.07	24	37.31	**29.21**
ทั้งหมด	66.27	**22.79**	**31.09**	**26.26**	67.47	23.6	30.57	26.64	68.87	25.21	31.09	27.84
ตัวแบบ Search												
25	90.29	**21.05**	4.65	7.62	89.29	**21.62**	9.3	13.01	90.69	11.11	1.16	2.11
50	88.59	11.11	4.65	6.56	87.09	18.84	15.12	16.77	90.59	16.67	2.33	4.08
100	87.29	14.04	9.3	11.19	86.19	21.11	22.09	**21.59**	90.69	**23.08**	3.49	6.06
250	83.68	12.62	15.12	**13.76**	77.18	13.02	**29.07**	17.99	86.19	12.86	10.47	11.54
500	83.28	10.68	12.79	11.64	76.58	11.05	24.42	15.22	75.58	10.1	**23.26**	14.09
1000	71.17	8.61	**24.42**	12.73	81.08	12.95	20.93	16	79.58	11.18	19.77	**14.29**
ทั้งหมด	83.08	7.22	8.14	7.65	83.98	13	15.12	13.98	83.68	11.88	13.95	12.83
ตัวแบบ Action												

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
25	96.70	17.65	13.64	**15.39**	96.20	**13.64**	**13.64**	**13.64**	97.80	50	9.09	15.39
50	96.10	13.04	13.64	13.33	95.80	11.54	13.64	12.5	97.90	**66.67**	9.09	**16**
100	94.29	9.3	**18.18**	12.31	95.90	12	13.64	12.77	96.60	16.67	**13.64**	15
250	93.59	8	18.18	11.11	94.09	4.88	9.09	6.35	95.80	8.33	9.09	8.7
500	92.19	4.84	13.64	7.14	94.90	6.06	9.09	7.27	96.20	10	9.09	9.52
1000	94.70	5.71	9.09	7.02	94.90	6.06	9.09	7.27	96.60	12.5	9.09	10.53
ทั้งหมด	97.10	**18.18**	9.09	12.12	95.20	6.67	9.09	7.69	96.90	15.38	9.09	11.43
ตัวแบบ Share												
25	93.09	**52.31**	47.22	**49.64**	93.99	63.04	40.28	49.15	95.40	**100**	36.11	53.06
50	91.79	43.75	48.61	46.05	94.39	**69.05**	40.28	50.88	95.50	**100**	37.5	**54.55**
100	90.49	37.36	47.22	41.72	94.29	68.29	38.89	49.56	95.50	**100**	37.5	**54.55**
250	88.09	31.2	54.17	39.59	93.13	53.33	44.44	48.49	94.39	67.39	43.06	52.54
500	86.09	27.21	**55.56**	36.53	92.89	50.82	43.06	46.62	92.39	47.06	**44.44**	45.71
1000	87.49	27.73	45.83	34.55	93.19	53.33	44.44	48.49	92.69	49.21	43.06	45.93
ทั้งหมด	92.49	47.62	41.67	44.44	94.29	65.31	**44.44**	**52.89**	93.89	61.22	41.67	49.59

ตารางที่ 22 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มเครื่องใช้ไฟฟ้าหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ตัวแบบ Attention												
25	94.41	0	0	0	94.41	**20.59**	20.59	**20.59**	96.27	0	0	0
50	88.20	2.38	5.88	3.39	88.82	9.78	**26.47**	14.29	96.17	0	0	0
100	83.85	1.59	5.88	2.5	90.17	8.22	17.65	11.22	93.27	3.03	2.94	2.99
250	82.20	1.41	5.88	2.27	95.14	15.79	8.82	11.32	94.72	5.26	2.94	3.77
500	82.30	1.42	5.88	2.29	95.24	7.14	2.94	4.17	95.34	**7.69**	**2.94**	**4.26**
1000	86.44	**2.91**	**8.82**	**4.38**	95.24	0	0	0	95.76	0	0	0

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ทั้งหมด	95.96	0	0	0	95.55	0	0	0	95.55	0	0	0
ตัวแบบ Interest												
25	78.57	22.41	7.43	11.16	76.81	27.52	17.14	21.13	81.78	**42.86**	1.71	3.3
50	76.61	26.17	16	19.86	73.29	27.07	28	27.53	81.78	40	1.14	2.22
100	75.05	**27.7**	23.43	25.39	70.08	25.43	33.71	28.99	81.06	21.43	1.71	3.18
250	73.40	26.16	25.71	25.94	66.67	23.47	37.14	28.76	77.95	25	10.86	15.14
500	67.81	21.9	30.29	25.42	68.53	25.84	**39.43**	**31.22**	74.74	29.82	29.14	**29.48**
1000	66.25	21.93	**33.71**	**26.58**	72.67	**27.64**	31.43	29.41	71.43	26.51	**32.57**	29.23
ทั้งหมด	74.95	20.87	13.71	16.55	74.43	26.62	23.43	24.92	74.74	25.18	20	22.29
ตัวแบบ Search												
25	88.20	2.94	1.22	1.72	83.75	**20**	**30.49**	**24.16**	90.17	**21.74**	6.1	9.52
50	86.65	3.92	2.44	3.01	83.44	19.53	30.49	23.81	89.13	15.15	6.1	8.7
100	85.20	5.8	4.88	5.3	81.37	16.89	30.49	21.74	89.13	17.14	7.32	10.26
250	80.85	8.13	12.2	9.76	79.71	13.46	25.61	17.65	85.40	8.45	7.32	7.84
500	79.81	**8.15**	**13.41**	**10.14**	84.37	12.9	14.63	13.71	85.20	12.35	**12.2**	**12.27**
1000	83.44	7.61	8.54	8.05	86.03	11.59	9.76	10.6	86.96	9.26	6.1	7.35
ทั้งหมด	88.61	6.25	2.44	3.51	86.23	9.52	7.32	8.28	87.79	9.09	4.88	6.35
ตัวแบบ Action												
25	96.79	**11.76**	11.76	**11.43**	96.69	10.53	11.76	11.11	97.62	12.5	5.88	8
50	95.03	**11.76**	11.76	7.69	96.27	8.7	11.76	10	96.69	5.88	5.88	5.88
100	93.89	6.25	**17.65**	9.23	96.38	**12.5**	**17.65**	**14.63**	95.96	**13.33**	**23.53**	**17.02**
250	93.48	5.77	17.65	8.7	97.83	0	0	0	98.03	0	0	0
500	94.41	4.88	11.76	6.9	97.52	0	0	0	98.14	0	0	0
1000	96.89	0	0	0	97.62	0	0	0	98.14	0	0	0
ทั้งหมด	98.24	0	0	0	97.83	0	0	0	98.14	0	0	0
ตัวแบบ Share												
25	89.03	10.71	3.57	5.36	86.13	9.68	7.14	8.22	90.68	12.5	1.19	2.17
50	89.34	**25.64**	11.9	16.26	85.61	**16.87**	**16.67**	**16.77**	90.48	10	1.19	2.13
100	88.20	22.22	14.29	17.39	83.02	10	11.9	10.87	90.27	8.33	1.19	2.08

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
250	82.09	14.4	21.43	17.23	83.33	10.31	11.9	11.05	89.44	5	1.19	1.92
500	76.92	12.02	**26.19**	16.48	84.78	12.05	11.9	11.98	84.47	9.76	9.52	9.64
1000	81.37	14.18	22.62	**17.43**	86.44	10.17	7.14	8.39	86.75	13.33	**9.52**	**11.11**
ทั้งหมด	88.51	13.51	5.95	8.26	87.58	15.38	9.52	11.77	88.92	**17.14**	7.14	10.08

ตารางที่ 23 แสดงค่าความถูกต้อง ค่าความแม่น ค่าระลึก และคะแนนเอฟวัน ของโฆษณากลุ่มสุขภัณฑ์หลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
ตัวแบบ Attention												
25	93.25	3.85	2.44	2.99	92	7.14	7.32	7.23	95.64	0	0	0
50	92.11	2.7	2.44	2.56	91.80	8.7	9.76	9.2	95.85	**66.67**	4.88	9.09
100	89.41	**5.8**	9.76	7.27	88.99	5.48	9.76	7.02	94.60	17.65	7.32	10.35
250	87.02	5.32	12.2	7.41	91.69	6.67	7.32	6.98	92.94	9.09	7.32	8.11
500	86.50	5.05	12.2	7.14	93.67	**16.67**	**12.2**	**14.09**	94.29	20.83	**12.2**	**15.39**
1000	87.33	5.49	**12.2**	**7.58**	93.56	11.11	7.32	8.82	92.94	9.09	7.32	8.11
ทั้งหมด	93.56	8	4.88	6.06	93.77	12	7.32	9.09	93.35	10.34	7.32	8.57
ตัวแบบ Interest												
25	67.15	23.48	27.75	25.44	33.91	22.35	**91.87**	**35.96**	27.92	21.88	**100**	35.91
50	73.33	**27.52**	19.62	22.91	46.67	23.24	71.29	35.06	27.92	21.83	99.52	35.8
100	70.53	24.47	22.01	23.17	51.69	23.78	63.16	34.56	32.66	22.27	93.78	**36**
250	69.76	24	22.97	23.47	53.91	23.52	56.97	33.29	38.94	22.43	82.3	35.25
500	65.80	21.79	26.79	24.03	59.42	24.82	49.76	33.12	51.98	23.23	59.81	33.47
1000	62.42	24.14	**40.19**	**30.16**	64.06	**26.24**	43.06	32.61	61.74	**25.46**	46.41	32.88
ทั้งหมด	69.28	19.21	16.27	17.62	68.41	25.62	29.67	27.49	69.18	23.81	23.92	23.87
ตัวแบบ Search												
25	90.34	11.76	5.41	7.41	88.21	**14.71**	13.51	14.09	90.27	0	0	0

จำนวนคุณลักษณะ	นาอีฟเบย์				การถดถอยโลจิสติกส์				ซัพพอร์ตเวกเตอร์แมชชีน			
	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	เอฟวัน
50	89.47	15.69	10.81	12.8	86.47	14.13	17.57	15.66	92.37	14.29	1.35	2.47
100	88.50	**16.42**	14.86	15.6	85.31	14.55	21.62	**17.39**	92.37	**27.27**	4.05	7.06
250	84.73	13.79	21.62	16.84	84.25	13.22	21.62	16.41	91.59	15.79	4.05	6.45
500	83.19	13.24	**24.32**	**17.14**	79.70	10.59	**24.32**	14.75	82.71	8.66	**14.86**	**10.95**
1000	84.83	13.91	21.62	16.93	86.67	10.98	12.16	11.54	87.63	9.09	8.11	8.57
ทั้งหมด	90.53	10	4.05	5.77	87.73	9.23	8.11	8.63	90.05	10.81	5.41	7.21
ตัวแบบ Action												
25	96.81	5	6.67	5.71	97.49	7.69	**6.67**	7.14	98.07	0	0	0
50	96.23	10	20	**13.33**	97.10	0	0	0	97.49	0	0	0
100	93.82	**7.02**	**26.67**	11.11	97.30	0	0	0	97.39	0	0	0
250	92.66	5.8	**26.67**	9.52	97.49	0	0	0	97.68	0	0	0
500	95.27	5.26	13.33	7.55	97.68	0	0	0	97.87	0	0	0
1000	97.68	0	0	0	97.49	0	0	0	97.87	0	0	0
ทั้งหมด	98.45	0	0	0	97.68	0	0	0	97.87	0	0	0
ตัวแบบ Share												
25	90.55	13.33	5.8	8.08	89.10	**10.87**	7.25	8.7	92.84	0	0	0
50	89.72	14.29	8.7	10.81	86.29	9.09	10.14	9.59	92.73	0	0	0
100	85.98	13.33	17.39	15.09	85.36	10	**13.04**	**11.32**	92.73	0	0	0
250	80.37	12.5	28.99	17.47	81.31	6.98	13.04	9.09	91.17	13.64	4.35	6.59
500	75.60	**14.83**	**50.72**	**22.95**	78.92	5.92	13.04	8.15	84.32	6.38	**8.7**	7.36
1000	75.60	13.91	46.38	21.41	84.01	8.74	13.04	10.47	87.12	7.69	7.25	7.46
ทั้งหมด	89.82	12.82	7.25	9.26	88.27	7.69	5.8	6.61	90.86	**14.81**	5.8	**8.33**

ต่อไปนี้เป็นการแสดงให้เห็นค่าเอ็มซีซี หลังจากนำวิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่มาใช้

ตารางที่ 24 แสดงค่าเอ็มซีซีของโฆษณากลุ่มเครื่องสำอางหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Attention			
25	0.1589	0.2393	**0.2413**
50	0.0845	**0.2180**	0.1769
100	0.0344	**0.1797**	0.1266
250	0.0209	**0.0479**	0.0209
500	-0.0269	**-0.0060**	-0.0075
1000	-0.0513	-0.0497	**-0.0434**
ทั้งหมด	-0.0395	**0.0034**	-0.0400
ตัวแบบ Interest			
25	-0.0129	0.0311	**0.0347**
50	-0.0096	0.0567	**0.0653**
100	-0.0123	**0.0599**	0.0594
250	-0.0356	**0.1091**	0.0816
500	-0.0312	0.0404	**0.0563**
1000	0.0200	0.0639	**0.0777**
ทั้งหมด	0.0517	0.0626	**0.0834**
ตัวแบบ Search			
25	0.0618	**0.0910**	0.0085
50	0.0173	**0.0994**	0.0317
100	0.0476	**0.1403**	0.0592
250	0.0485	**0.0767**	0.0416
500	0.0250	**0.0422**	0.0265
1000	0	**0.0622**	0.0389
ทั้งหมด	-0.0163	**0.0522**	0.0391
ตัวแบบ Action			
25	0.1385	0.1169	**0.2065**
50	0.1134	0.1040	**0.2411**

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
100	0.1026	0.1070	**0.1335**
250	**0.0907**	0.0377	0.0655
500	0.0462	0.0486	**0.0759**
1000	0.0456	0.0486	**0.0895**
ทั้งหมด	**0.1149**	0.0535	0.1031
ตัวแบบ Share			
25	0.4601	0.4744	**0.5865**
50	0.4169	0.5010	**0.5980**
100	0.3691	0.4887	**0.5980**
250	0.3509	0.4509	**0.5113**
500	0.3213	**0.4301**	0.4165
1000	0.2918	**0.4509**	0.4213
ทั้งหมด	0.4054	**0.5102**	0.4744

ตารางที่ 25 แสดงค่าเอ็มซีซีของโฆษณากลุ่มเครื่องใช้ไฟฟ้าหลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Attention			
25	-0.0278	**0.1769**	-0.0087
50	-0.0191	**0.1103**	-0.0107
100	-0.0406	**0.0729**	-0.0050
250	-0.0476	**0.0943**	0.0134
500	-0.0471	0.0238	**0.0264**
1000	**-0.0114**	-0.0214	-0.0163
ทั้งหมด	**-0.0138**	-0.0185	-0.0185
ตัวแบบ Interest			
25	0.0282	**0.0871**	0.0549
50	0.0738	**0.1117**	0.0410
100	0.1059	**0.1068**	0.0104

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
250	**0.0972**	0.0881	0.0522
500	0.0568	0.1240	**0.1410**
1000	0.0616	**0.1259**	0.1166
ทั้งหมด	0.0263	**0.0962**	0.0752
ตัวแบบ Search			
25	-0.0380	**0.1592**	0.0742
50	-0.0387	**0.1548**	0.0450
100	-0.0268	**0.1282**	0.0602
250	-0.0049	**0.0783**	-0.0004
500	-0.0049	**0.0517**	0.0419
1000	-0.0102	**0.0309**	0.0067
ทั้งหมด	-0.0149	**0.0098**	0.0047
ตัวแบบ Action			
25	**0.0980**	0.0944	0.0746
50	0.0583	**0.0824**	0.0420
100	0.0781	0.1304	**0.1576**
250	**0.0727**	-0.0086	-0.0061
500	**0.0499**	-0.0114	-0.0043
1000	-0.0156	-0.0106	**-0.0043**
ทั้งหมด	**0**	-0.0086	-0.0043
ตัวแบบ Share			
25	**0.0124**	0.0091	0.0123
50	**0.1234**	0.0889	0.0047
100	**0.1168**	0.0157	-0.0014
250	**0.0780**	0.0191	-0.0191
500	**0.0571**	0.0365	0.0115
1000	**0.0781**	0.0133	0.0424
ทั้งหมด	0.0341	0.0566	**0.0581**

ตารางที่ 26 แสดงค่าเอ็มซีซีของโฆษณากลุ่มสุขภัณฑ์หลังประยุกต์ใช้วิธีการคัดเลือกคุณลักษณะร่วมกับเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยและการเพิ่มคุณลักษณะใหม่

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
ตัวแบบ Attention			
25	-0.0034	**0.0305**	-0.0068
50	-0.0154	0.0492	**0.1728**
100	0.0212	0.0173	**0.0889**
250	0.0173	0.0264	**0.0451**
500	0.0133	0.1102	**0.1313**
1000	0.0198	**0.0577**	0.0451
ทั้งหมด	0.0303	**0.0626**	0.0531
ตัวแบบ Interest			
25	0.0459	0.1188	**0.1456**
50	0.0748	0.0970	**0.1386**
100	0.0502	0.0962	**0.1234**
250	0.0464	0.0810	**0.0941**
500	0.0229	**0.0951**	0.0788
1000	0.0699	**0.1060**	0.1001
ทั้งหมด	-0.0111	**0.0747**	0.0454
ตัวแบบ Search			
25	0.0330	**0.0778**	-0.0212
50	0.0754	**0.0846**	0.0229
100	0.0946	**0.0990**	0.0809
250	**0.0916**	0.0858	0.0459
500	**0.0919**	0.0592	0.0219
1000	**0.0928**	0.0436	0.0197
ทั้งหมด	0.0191	0.0209	**0.0274**
ตัวแบบ Action			
25	0.0417	**0.0589**	-0.0084
50	**0.1236**	-0.0147	-0.0126

จำนวนคุณลักษณะ	นาอีฟเบย์	การถดถอยโลจิสติกส์	ซัพพอร์ตเวกเตอร์แมชชีน
100	**0.1125**	-0.0137	-0.0131
250	**0.0972**	-0.0126	-0.0114
500	**0.0623**	-0.0114	-0.0100
1000	-0.0114	-0.0126	**-0.0100**
ทั้งหมด	**-0.0038**	-0.0114	-0.0100
ตัวแบบ Share			
25	**0.0429**	0.0322	0
50	**0.0590**	0.0220	-0.0090
100	**0.0768**	0.0353	-0.0090
250	**0.0923**	-0.0029	0.0384
500	**0.1693**	-0.0209	-0.0100
1000	**0.1466**	0.0211	0.0055
ทั้งหมด	0.0450	0.0049	**0.0504**

ต่อไปนี้เป็นการแสดงตัวอย่างการทำนายคลาสของตัวแบบการถอดถอยโลจิสติกส์เปรียบเทียบกับคลาสจริงของข้อมูลทดสอบซึ่งเป็นโฆษณากลุ่มเครื่องสำอาง เมื่อใช้ข้อมูลฝึกสอนด้วยจำนวนคุณลักษณะที่มีความเหมาะสมที่สุด และนำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ร่วมกับเทคนิคการคัดเลือกคุณลักษณะใหม่ที่เป็นคำคล้ายคลึง

ตารางที่ 27 แสดงตัวอย่างการทำนายคลาสของตัวแบบ ชุดข้อมูล ATTENTION

ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
หากคุณสงสัยว่าผลิตภัณฑ์ NAMULIFESNAILWHITE ของแท้หรือไม่? สามารถตรวจสอบได้ด้วยวิธีง่าย ๆ และรู้ผลทันทีนะคะกับวิธีการเช็คสินค้าจากฉลากกันปลอมด้วยนวัตกรรมล้ำสมัย ANTI–COUNTERFEITQRCODESTICKER เพื่อความมั่นใจสูงสุดของผู้บริโภค 1)แกะห่อพลาสติกออก 2)นำเหรียญขูดแถบ ScratchArea เมื่อขูดเสร็จจะพบรหัสสินค้า 3)ใช้แอพพลิเคชั่นที่สามารถสแกน QRCode ได้เลือกเข้าโปรแกรม 4) กดสแกน QRCode รหัสสินค้า QRCode จะแสดงบนหน้าจอ 5)กดสีชมพูตรวจสอบรหัสบนแถบขูด ScratchArea และบน QRCode จะตรงกัน	Attention	Attention

ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
พบกับเทปพิเศษของไดอารี่ตุ๊ดซี่ส์ตอน 30 วันมหัศจรรย์แห่งหอยกับNAMULIFE SNAILWHITE MIRACLE 30 DAYS MIRACLE สามารถติดตามชมไดอารี่ตุ๊ดซี่ส์เทปพิเศษนี้เพิ่มเติมได้ทาง Youtube NAMULIFE Channel นะคะฮามากบอกเลยอยากแชร์ ***ข้อแนะนำเพื่อเพิ่มอรรถรสในการชมอย่าลืมกด HD ตรงมุมล่างขวาของวีดีโอด้วยนะจ๊ะ	Not Attention	Attention
รกปลาเด้งนามุ อาหารเสริมของยุคนี้ทำจากเยื่อหุ้มรังไข่ปลาแซลมอนแท้ 100% ไม่มีสารตกค้างเพราะมาจากธรรมชาติแท้ๆ ผิวจะสวยเนียนใสตึงกระชับทั้งตัวสวยเด้งเด็กวันละเม็ดก่อนนอน	Not Attention	Not Attention
พรีเซ็นเตอร์คนล่าสุดคุณใหม่ดาวิกา NAMULIFE SNAILWHITE CRÈME BODY WASH ถึงเวลาผิวกายสะอาดขาวใสเด้งสุขภาพดีด้วยใหม่NAMULIFE SNAILWHITE CRÈME BODY WASH ครั้งแรกกับครีมอาบน้ำเนื้อโลชั่น เมื่อสัมผัสน้ำให้ฟองครีมละเอียดนุ่มแต่ล้างออกง่าย ความลับเพื่อสุขภาพผิวกายที่ดีที่เริ่มต้นตั้งแต่การอาบน้ำทั้งอาบและช่วยบำรุงในหนึ่งเดียว	Attention	Not Attention

ตารางที่ 28 แสดงตัวอย่างการทำนายคลาสของตัวแบบ ชุดข้อมูล INTEREST

ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
ฟื้นคืนคอลลาเจนยกกระชับผิวเพียงข้ามคืน NAMULIFE SNAILWHITE OVERNIGHT FIRMING MASK ขนาด 50mlราคา 790 บาทโปรโมชั่นซื้อ 2แถม1 ฟรีครีมมาส์กผิวหน้ายามค่ำคืน ด้วยสารสกัดจากธรรมชาติเข้มข้นให้ผิวคุณสวยเพียงข้ามคืน ช่วยปรับสมดุลการสร้างคอลลาเจนยกกระชับผิวที่หย่อนคล้อยให้เต่งตึงแก้ปัญหาการขาดความตึงกระชับของผิวหน้า Watsons ทุกสาขาวันนี้-31ธันวาคมนี้ Central Robinson Eveandboy Jeleng ตั้งแต่วันที่ 1ตุลาคม-31 ธันวาคมนี้	Interest	Interest
ลองกันเลยนะคะแล้วจะรู้ว่ามาส์กคุณภาพดีที่ทำจากผ้าคอตตอนแท้ 100% ทอผ่านน้ำและอุดมด้วยสารสกัดช่วยบำรุงผิวแบบเต็มๆต้องสเนลไวท์มาส์กชอทท์เท่านั้น	Not Interest	Interest
ลบให้ดูกันแบบชัดๆเครื่องสำอางเมคอัพติดทนขนาดไหนไม่ต้องห่วงสเนลไวท์คลีนซิ่งเมือกล้างหน้าที่สุดแห่งการทำความสะอาดผิวให้หมดจดกระจ่างใสพร้อมฟื้นฟูในขั้นตอนเดียว ไม่ทำร้ายผิวหน้าให้แห้งตึง ลดและฟื้นฟูผิวที่มีริ้วรอยให้ดูตื้นขึ้นผิวอ่อนกว่าวัยดูเด็กเมือกล้างหน้าสเนลไวท์ คลีนซิ่งขนาด151มล.ราคา590บาท	Not Interest	Not Interest
ครีมหอยขาวสเนลไวท์ตบแล้วใสใช้แล้วตึง สำหรับคนที่ต้องการฟื้นฟูสภาพผิวอย่างแท้จริง เหมาะกับคนผิวแพ้ง่าย สารสกัดจากธรรมชาติ	Interest	Not Interest

ตารางที่ 29 แสดงตัวอย่างการทำนายคลาสของตัวแบบ ชุดข้อมูล SEARCH

ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
สวัสดีวันจันทร์ค่า ตอนนี้ใครๆก็ตกใจกับครีมกันแดด snailwhite ทาแล้วเหมือนไม่ได้ทา ไม่รบกวนการแต่งหน้าทิ้งไว้แต่การปกป้องล้ำลึกและบำรุงให้คุณดูเด็กนานแสนนาน snailwhite sunscreen SPF50 PA4+ ปกป้องล้ำลึกแบบไร้ตัวตน ปกป้องผิวคุณตั้งแต่วันนี้เพื่อผิวขาวกระจ่างใส กระฝ้าลดเลือนผิวหน้าอ่อนเยาว์ ให้คุณดูดีกว่าใครด้วยเมือกหอยทากสูตรเฉพาะของ snailwhite เมือกหอยทากที่ดีที่สุด การันตีด้วยยอดขายอันดับ1 โปรดระวังสินค้าลอกเลียนแบบ	Search	Search
หนึ่งไอเทมช่วยเซฟเวลาพร้อมดูแลผิวให้ผู้หญิงอย่างเราไม่ต้องเสียเวลาเช็ดเครื่องสำอางอีกต่อไปเพราะ NAMULIFE SNAILWHITE CLEANSING มี QUICK-TECHNOLOGY ชำระได้ทั้งคราบสกปรกบนผิวหน้าและเครื่องสำอางสูตรกันน้ำแบบนี้มีเวลาบำรุงผิวเพิ่มอีกเยอะเลย	Not Search	Search
สุดพิเศษโปรโมชั่น NamuLife SNAILWHITE Whip Soap ซื้อ1แถม1 ได้ที่โรบินสัน CentralShowDCBTSSiam และ Tsuruha (ทุกสาขาที่ร่วมรายการ) วิปโฟมฟองละเอียดทำความสะอาดได้ลึกถึงรูขุมขนอ่อนโยนต่อผิวหน้าพร้อมฟื้นฟูให้ผิวสุขภาพดี หมายเหตุ*ตั้งแต่วันนี้ถึงจนกว่าสินค้าจะหมด	Not Search	Not Search
คุณรู้หรือไม่คะว่าควรมาส์กหน้าเป็นเวลากี่นาทีที่จะทำให้ผิวคุณเอิบอิ่มสวยปังที่สุด	Search	Not Search

ตารางที่ 30 แสดงตัวอย่างการทำนายคลาสของตัวแบบ ชุดข้อมูล ACTION

ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
SNAILWHITE CRÈME BODYWASH ครั้งแรกกับครีมอาบน้ำเนื้อโลชั่น เมื่อสัมผัสน้ำให้ฟองครีมละเอียดนุ่มล้างออกง่ายสุดยอด ความลับเพื่อสุขภาพผิวกายที่ดีที่เริ่มต้นตั้งแต่การอาบน้ำ ทั้งอาบและช่วยบำรุงในหนึ่งเดียว โอกาสสุดท้ายกิจกรรมอาบให้สวยกว่าใหม่ ลุ้นรางวัล100000บาทถึง5รางวัลได้ที่ **ลิงก์**	Action	Action
เราจะไม่ยอมเป็นตัวเลือกของใครเพราะเราคือผู้หญิงที่มีเสน่ห์และมีคุณค่าในแบบตัวเอง สวยในแบบที่คุณเป็น	Not Action	Action
ตอบแทนซุปเปอร์แฟน SNAILWHITE ด้วยของสมนาคุณสุดชิคกระเป๋าเดินทาง NAMULIFEขนาด18นิ้วจำนวน1ใบเมื่อซื้อสินค้าครบ 3900 บาทเฉพาะที่ **ลิงก์** ตลอดเดือนสิงหาคม สินค้ามีจำนวนจำกัดรีบมาช้อปและรับของสมนาคุณชิคๆกันได้เลยค่าสนใจสั่งซื้อ **ลิงก์**	Not Action	Not Action

ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
ผิวหน้าดีแล้วผิวกายต้องดีด้วย เพิ่มออร่าให้ผิวกายมาฟื้นฟูผิวกันเถอะด้วย NAMULIFE SNAILWHITE BODY BOOSTER ครีมบำรุงผิวกายที่มีส่วนผสมสารสกัดจากธรรมชาติเมือกหอยทากแอสตาแซนธินและเซรามายด์สูตรเข้มข้น ที่มีประสิทธิภาพในการช่วยฟื้นบำรุงผิวกายที่คล้ำเสียให้ผิวดูมีสุขภาพดีขาวกระจ่างใส อ่อนเยาว์เต่งตึงและนุ่มชุ่มชื้นในหนึ่งเดียว สามารถใช้ทาทั่วผิวกายและตบเบาๆหลังอาบน้ำทั้งเช้าเย็น	Action	Not Action

ตารางที่ 31 แสดงตัวอย่างการทำนายคลาสของตัวแบบ ชุดข้อมูล SHARE

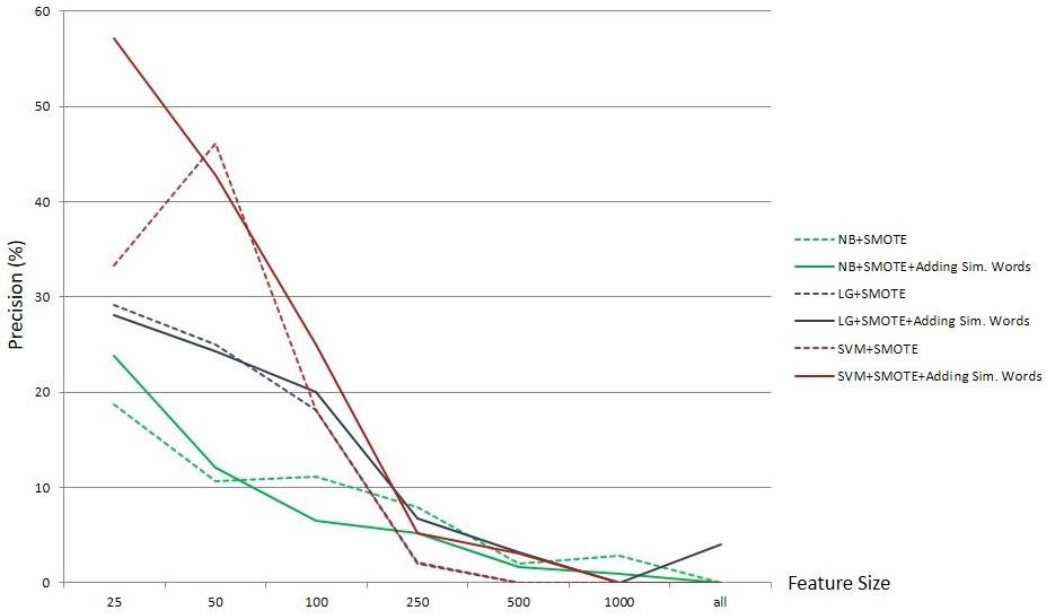
ข้อความโฆษณา	คลาสจริง	คลาสที่ทำนาย
โครงการสวย MakeSense MakeMore จาก NAMULIFE SNAILWHITE เพื่อจุดประกายผู้หญิงให้มี attitude ความสวยจากภายนอกและภายในแบ่งปันเวลามาสร้างประโยชน์ให้ผู้อื่นซึ่งทางSNAILWHITEขอความร่วมมือท่านในการ 1)ลงรูปส่วนตัวของท่านชู2นิ้ว 2)เขียนข้อความใต้ภาพสวยแล้ว?? (ในความคิดของคุณ) ที่เป็นเรื่องดีดี 3) และ #สวยMakeSenseMakeMore โดยหลังจากนี้จะมีกิจกรรมช่วยเหลือสังคมสงเคราะห์ที่เชิญชวนทุกคนมาทำร่วมกันเพื่อสร้างค่านิยมที่ว่าผู้หญิงที่สวยจะไม่หยุดสวยที่ตัวเองแต่จะแบ่งปันความสวยอย่างมีสาระประโยชน์ให้ผู้อื่นด้วย	Share	Share
สาวๆค่ะรู้ไหมค่ะว่าน้ำมะพร้าวนั้นเป็นเครื่องดื่มจากธรรมชาติอันบริสุทธิ์ที่เปี่ยมด้วยฮอร์โมนเอสโตรเจนสูง ที่มีส่วนในการเสริมสร้างคอลลาเจนและอีลาสตินทำให้ผิวพรรณเปล่งปลั่งสดใสมีความกระชับยืดหยุ่นและชะลอการเกิดริ้วรอยแห่งวัยได้ นอกจากนี้ในน้ำมะพร้าวยังมีฤทธิ์ช่วยขับปัสสาวะและขับของเสียจากร่างกาย เปรียบดั่งเครื่องดื่มดีท็อกซ์ที่ยิ่งดื่มผิวพรรณจึงยิ่งผุดผ่องมากขึ้น อีกทั้งน้ำมะพร้าวยังมีค่าความเป็นด่างที่จะช่วยปรับสมดุลให้กับร่างกายในตอนที่มีความเป็นกรดสูงได้เป็นอย่างดีจึงทำให้กลไกลในการทำงานของระบบภายในเป็นไปอย่างปกติ ส่งผลให้คุณสาวๆมีสุขภาพที่ดีทั้งภายในและภายนอกมากขึ้นค่ะ	Not Share	Share
สวยๆกับสเนลไวท์ ชิน-เอค มิสท์ใช้ฉีดพ่นบริเวณใบหน้าทุกวันเช้าเย็นหรือเวลาที่ต้องการความสดชื่นได้ด้วยนะคะ ถ้าฉีดหลังแต่งหน้าจะช่วยทำให้เครื่องสำอางค์ดูเฟรชขึ้นด้วย ฉีดแล้วตบเบาให้ให้เอสเซนท์ซึมลงบนผิว รับรองว่าตบแล้วดูใสใช้แล้วตึง	Not Share	Not Share
วันนี้ชวนมาพบกับพรีเซนเตอร์คนใหม่ของ snailwhite sunscreen สวยไม่กลัวแดดผิวไม่เสียนะจ๊ะ แบบสเนลไวท์ตบแล้วดูใสใช้แล้วตึงเย็นนี้ 18:00 น ที่paragon มาเชียร์ presenter เรากันเยอะเยอะนะคะ	Share	Not Share

4.4 เปรียบเทียบการทดลอง CHI2+SMOTE Adding Similar Words+CHI2+SMOTE

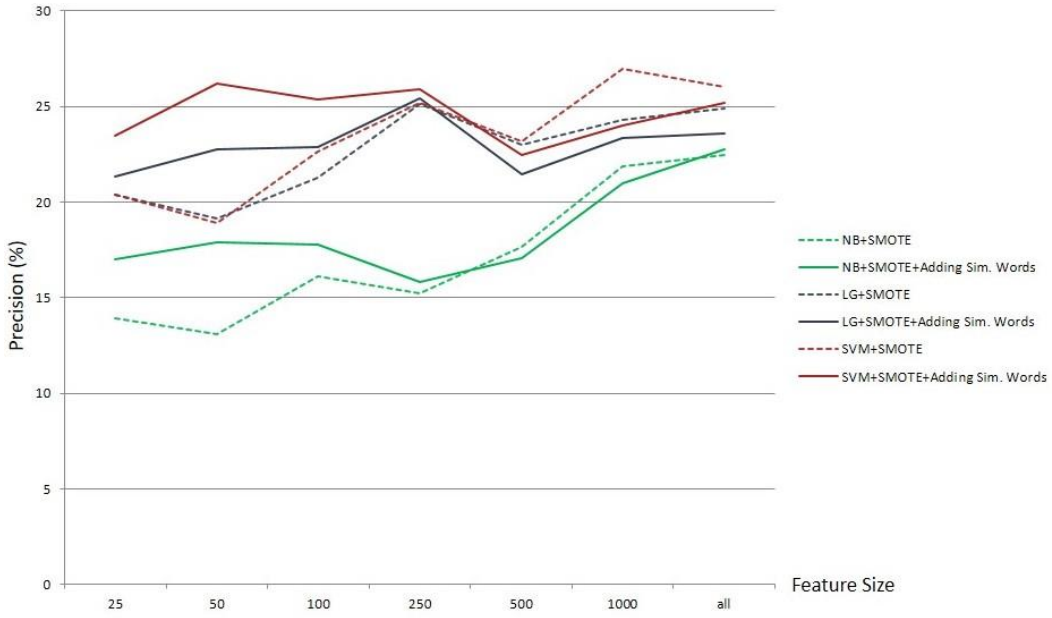
4.4.1 เปรียบเทียบในบริบทของค่าระลึกและค่าความแม่นในทุกระดับคุณลักษณะ

ในหัวข้อนี้เป็นการเปรียบเทียบผลการทดลองก่อนและหลังจากนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย โดยกราฟต่อไปนี้แสดงค่าความแม่นและค่าระลึกเมื่อมีการใช้จำนวนคุณลักษณะหลายระดับ ได้แก่ 25, 50, 100, 250, 500, 1000 และทุกคุณลักษณะ เปรียบเทียบกันระหว่างการทดลองก่อนนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ (CHI2+SMOTE) และหลังจากนำวิธีการดังกล่าวมาใช้ (CHI2+SMOTE+Adding Similar Words) บนวิธีการเรียนรู้ของเครื่องทั้งสามวิธี ได้แก่ นาอีฟเบย์ (Naïve Bayes: *NB*) การถดถอยโลจิสติกส์ (Logistic Regression: *LG*) และซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: *SVM*)

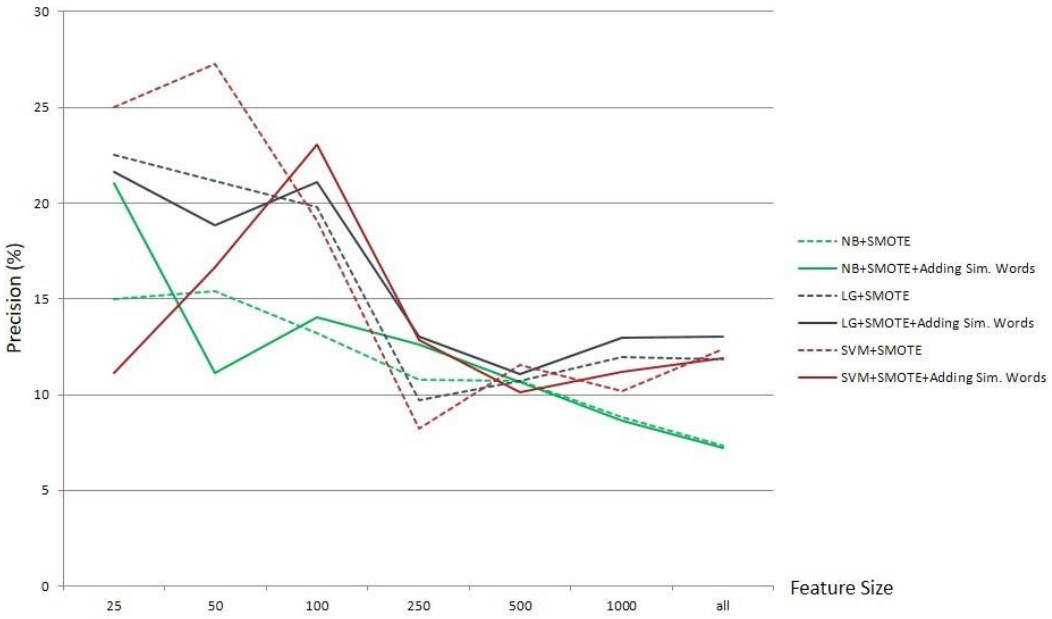
โดยกราฟแสดงค่าความแม่นของทั้ง 15 ตัวแบบ เป็นดังต่อไปนี้



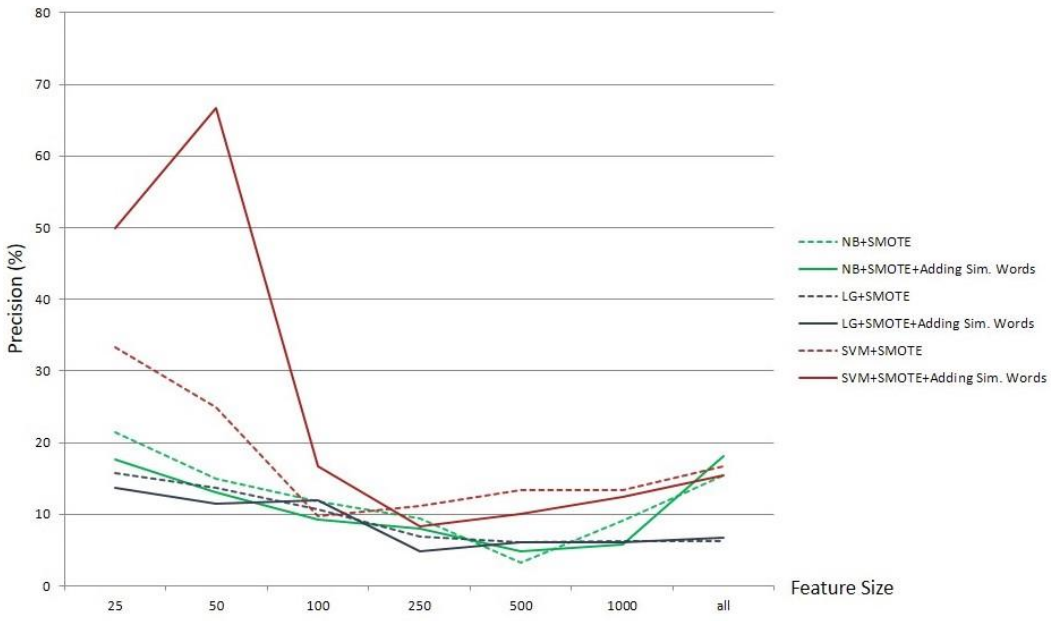
ภาพที่ 17 แสดงค่าความแม่นของตัวแบบ ATTENTION โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



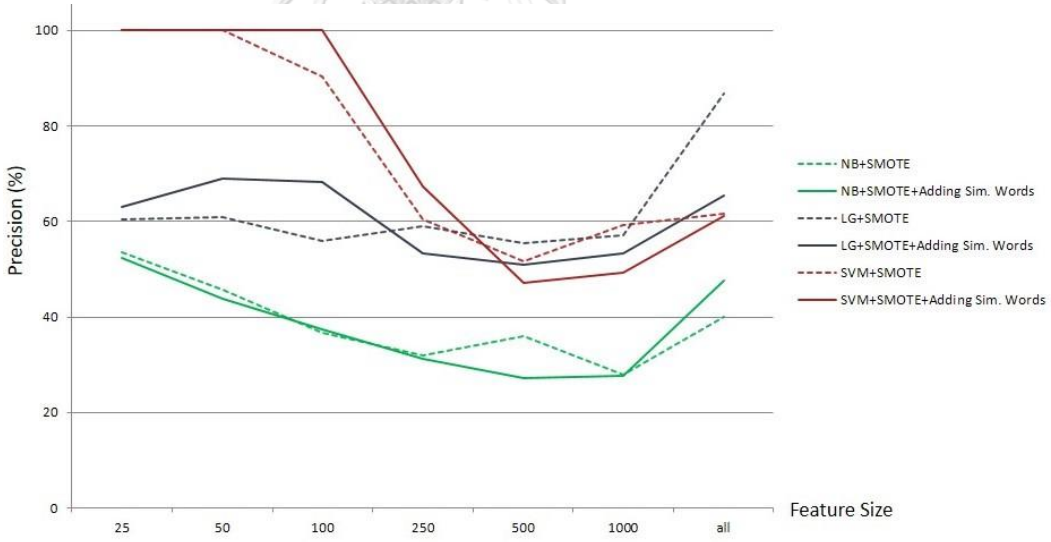
ภาพที่ 18 แสดงค่าความแม่นของตัวแบบ INTEREST โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



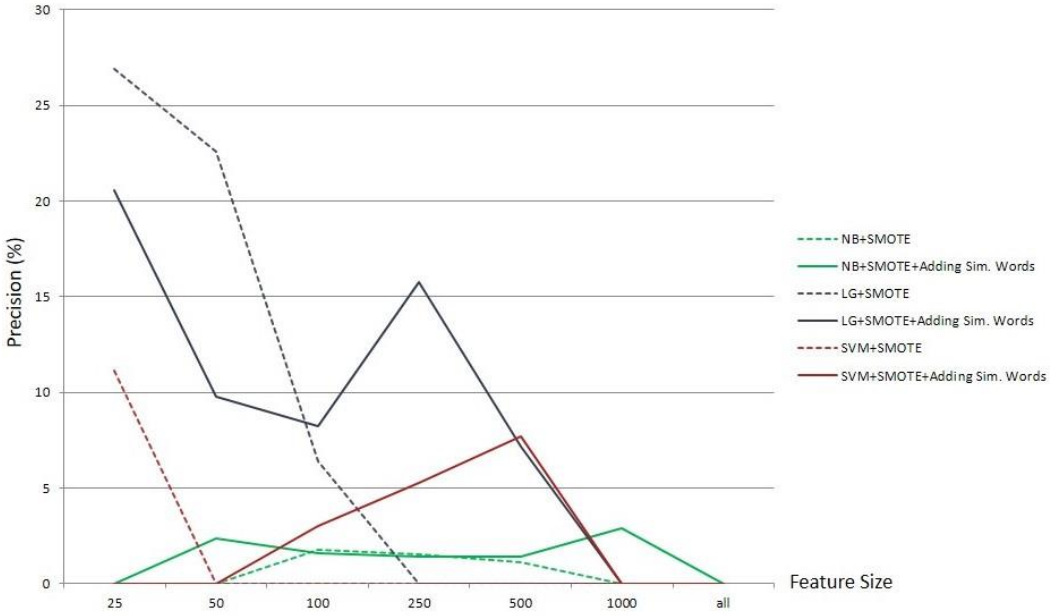
ภาพที่ 19 แสดงค่าความแม่นของตัวแบบ SEARCH โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



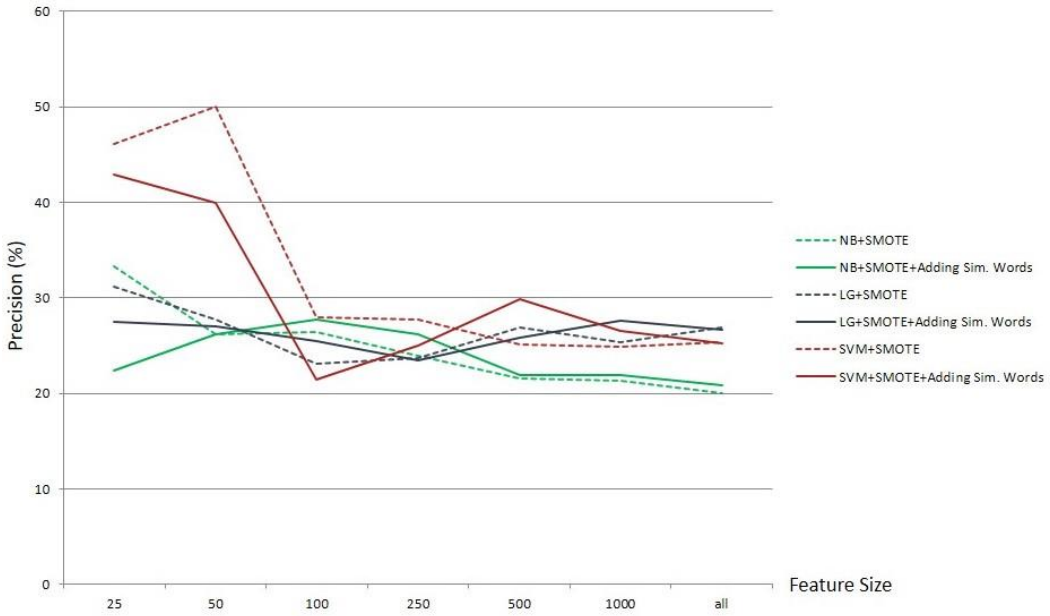
ภาพที่ 20 แสดงค่าความแม่นของตัวแบบ ACTION โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



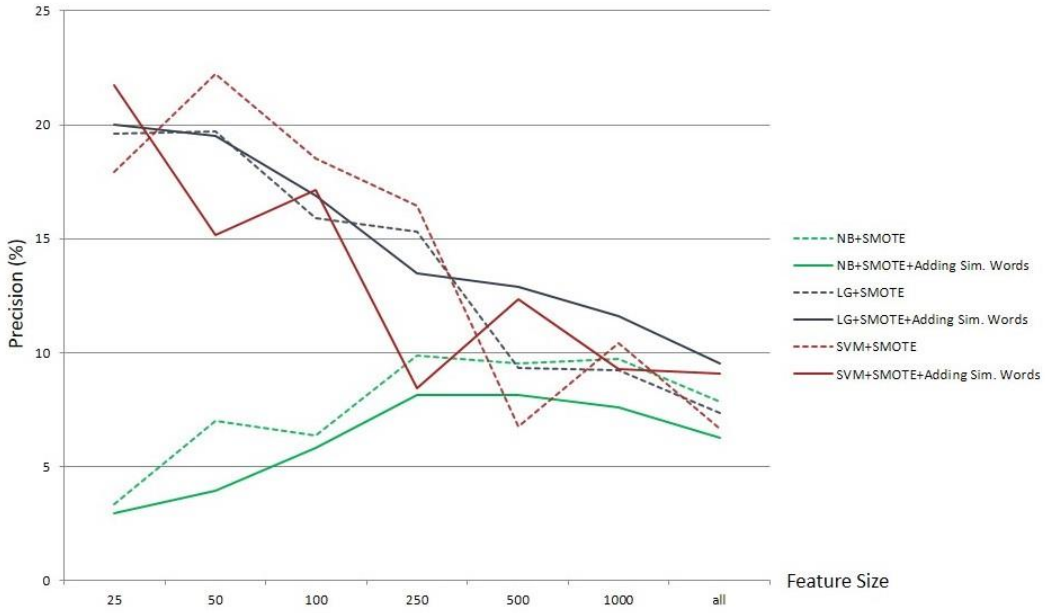
ภาพที่ 21 แสดงค่าความแม่นของตัวแบบ SHARE โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



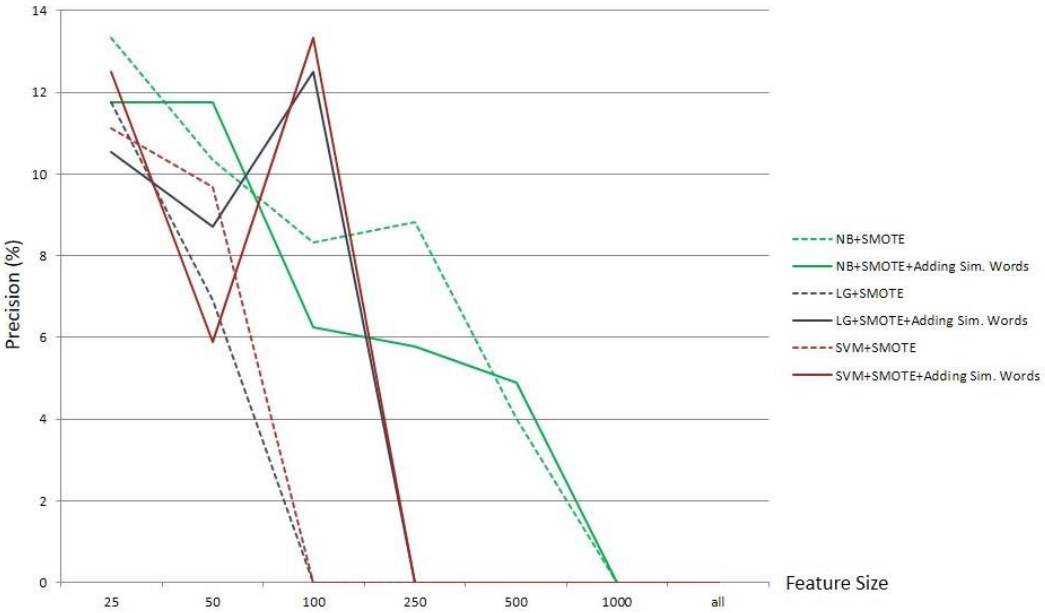
ภาพที่ 22 แสดงค่าความแม่นของตัวแบบ ATTENTION โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



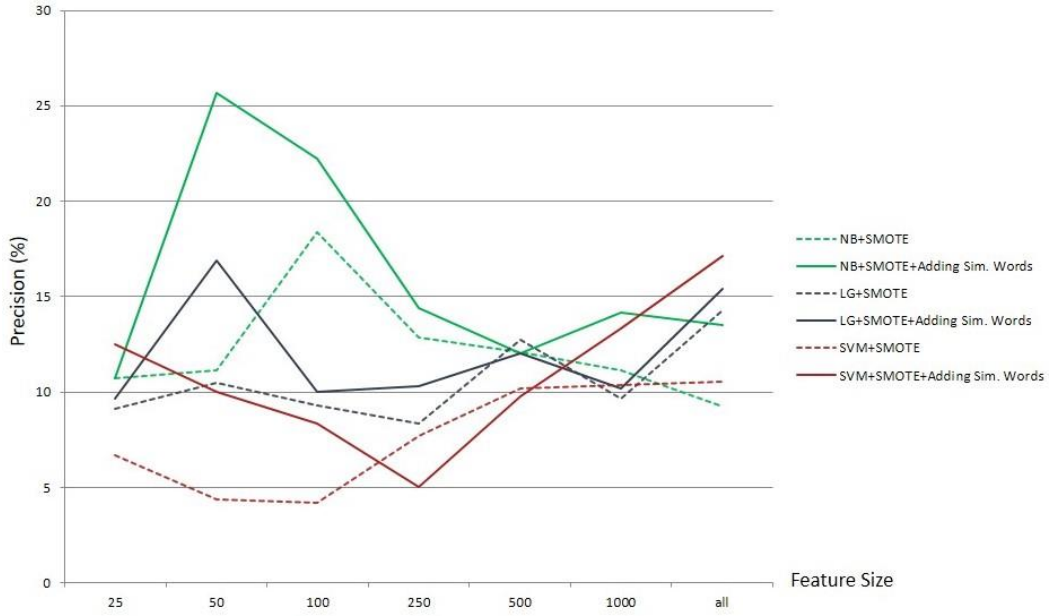
ภาพที่ 23 แสดงค่าความแม่นของตัวแบบ INTEREST โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



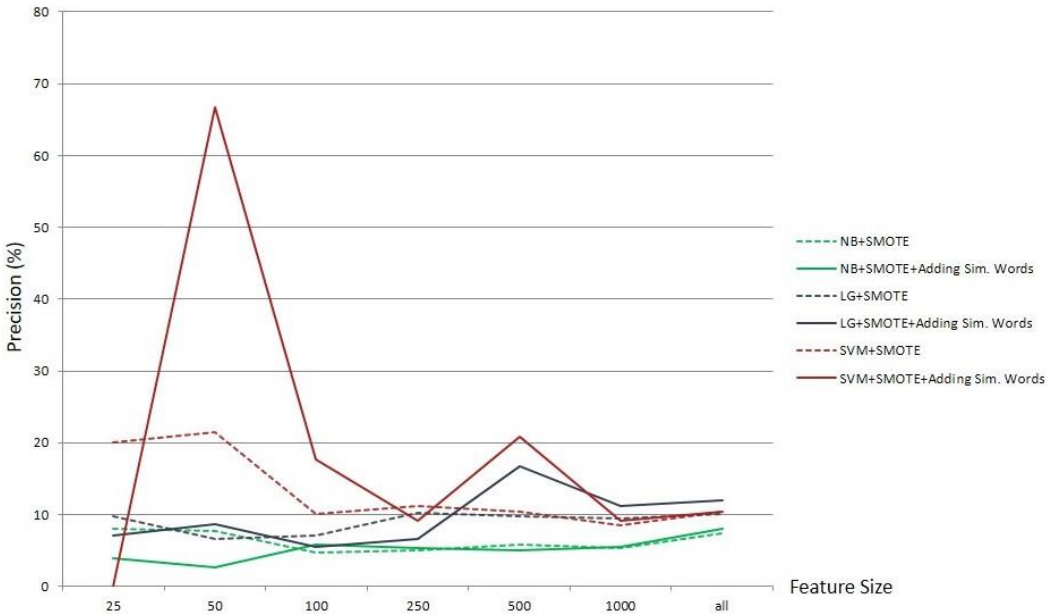
ภาพที่ 24 แสดงค่าความแม่นของตัวแบบ SEARCH โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



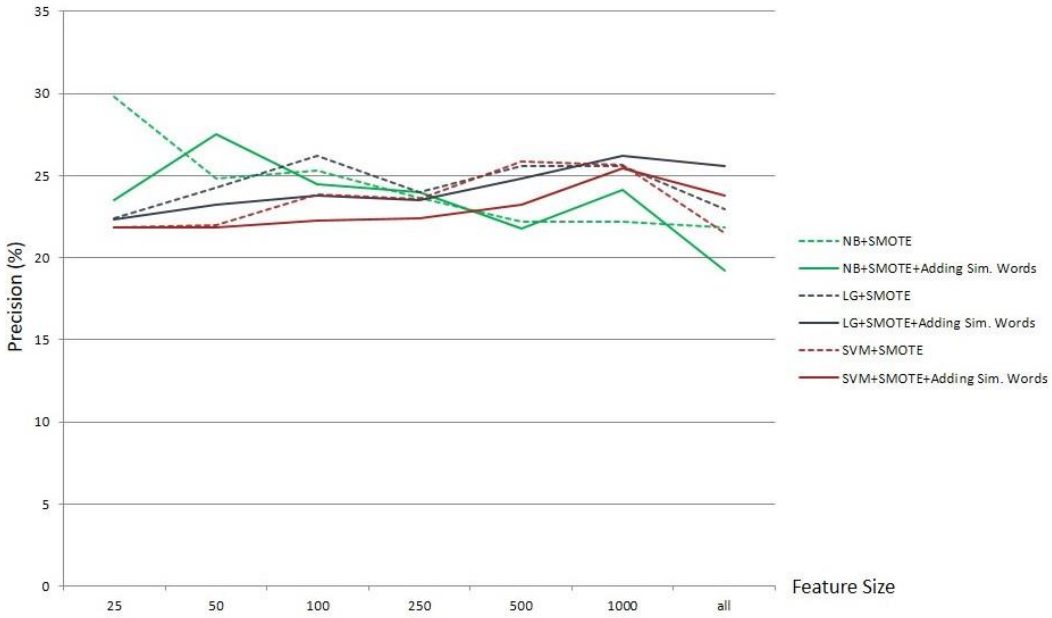
ภาพที่ 25 แสดงค่าความแม่นของตัวแบบ ACTION โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



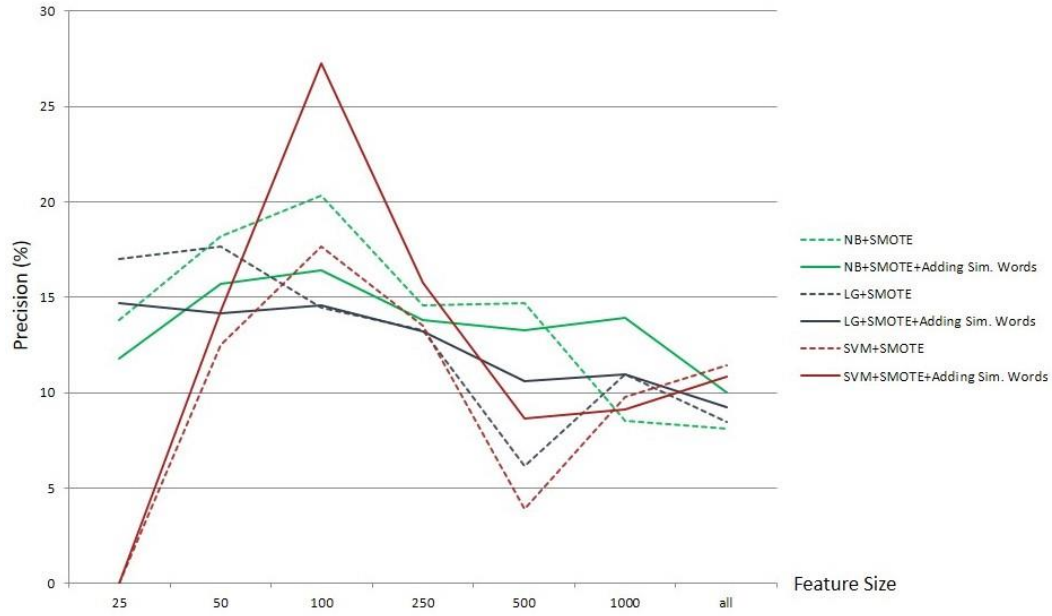
ภาพที่ 26 แสดงค่าความแม่นของตัวแบบ SHARE โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



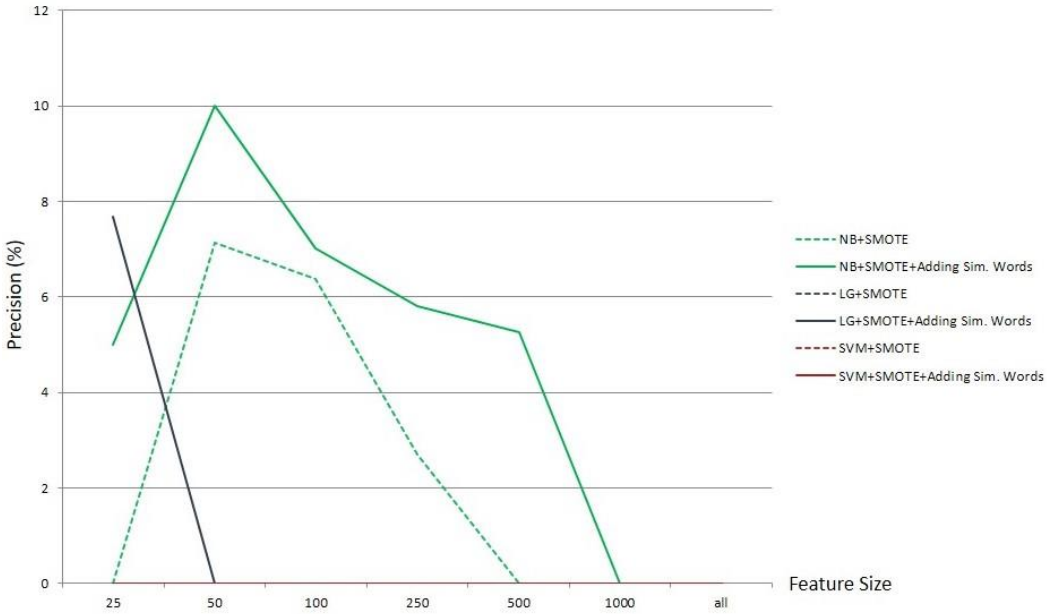
ภาพที่ 27 แสดงค่าความแม่นของตัวแบบ ATTENTION โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



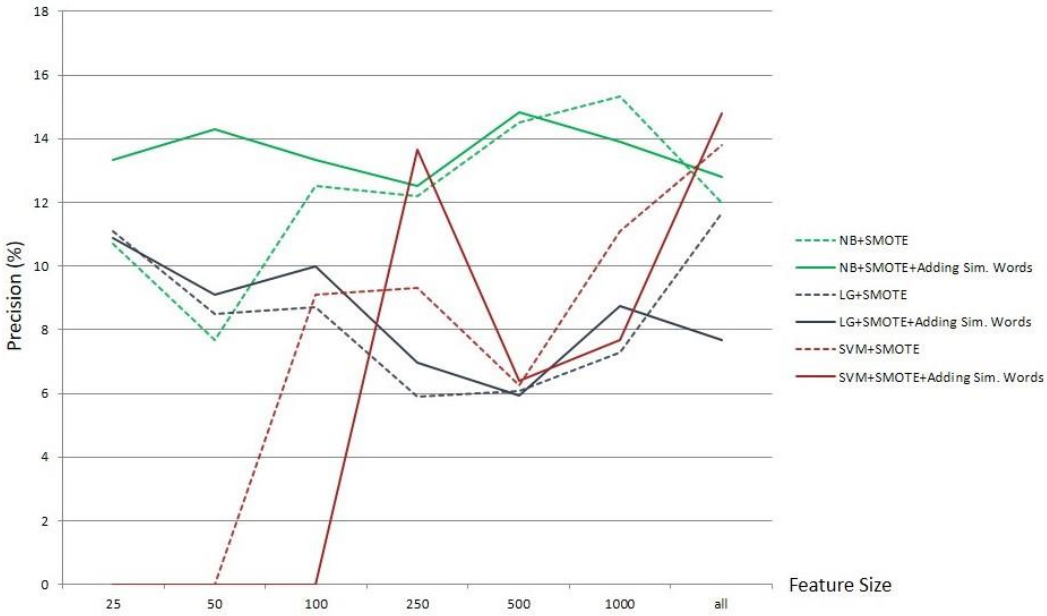
ภาพที่ 28 แสดงค่าความแม่นของตัวแบบ INTEREST โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



ภาพที่ 29 แสดงค่าความแม่นของตัวแบบ SEARCH โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

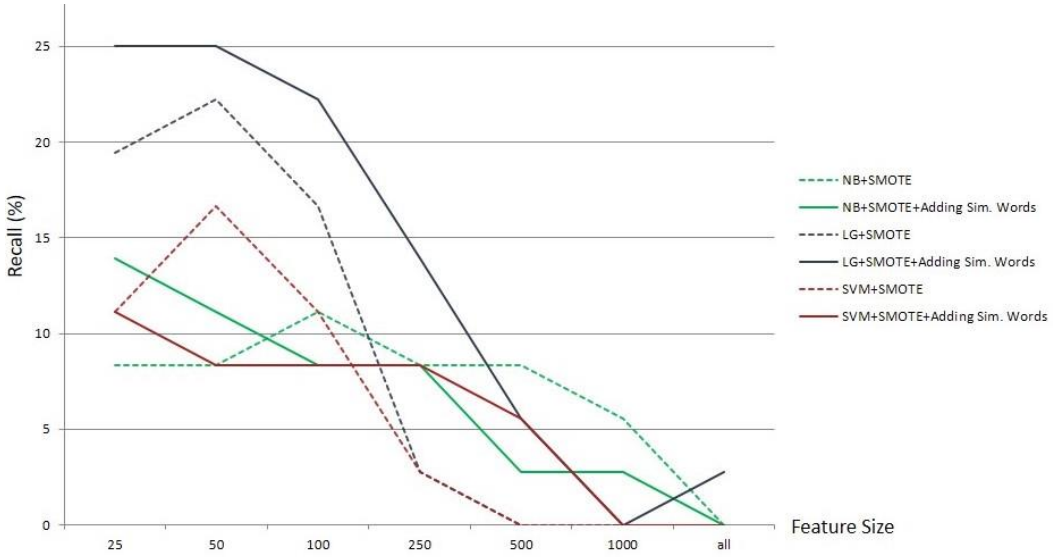


ภาพที่ 30 แสดงค่าความแม่นของตัวแบบ ACTION โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

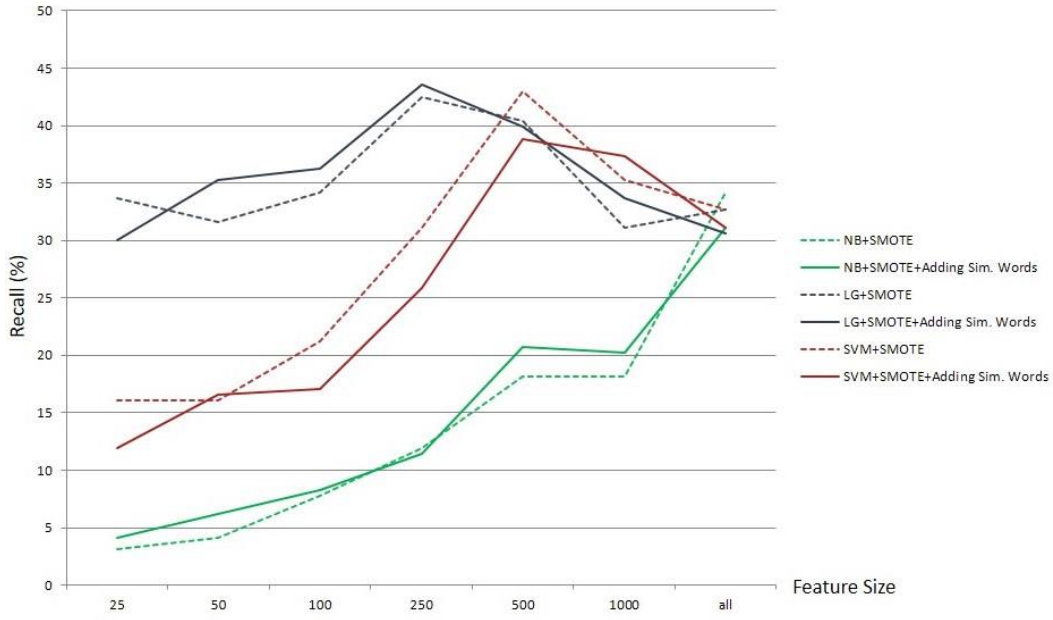


ภาพที่ 31 แสดงค่าความแม่นของตัวแบบ SHARE โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

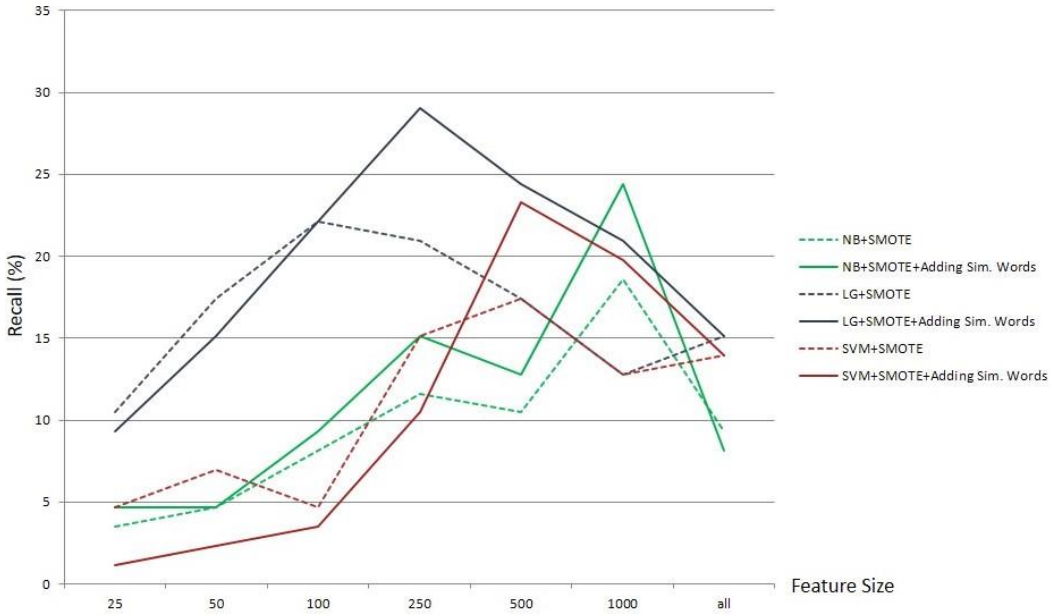
และกราฟแสดงค่าระลึกของทั้ง 15 ตัวแบบ เป็นดังต่อไปนี้



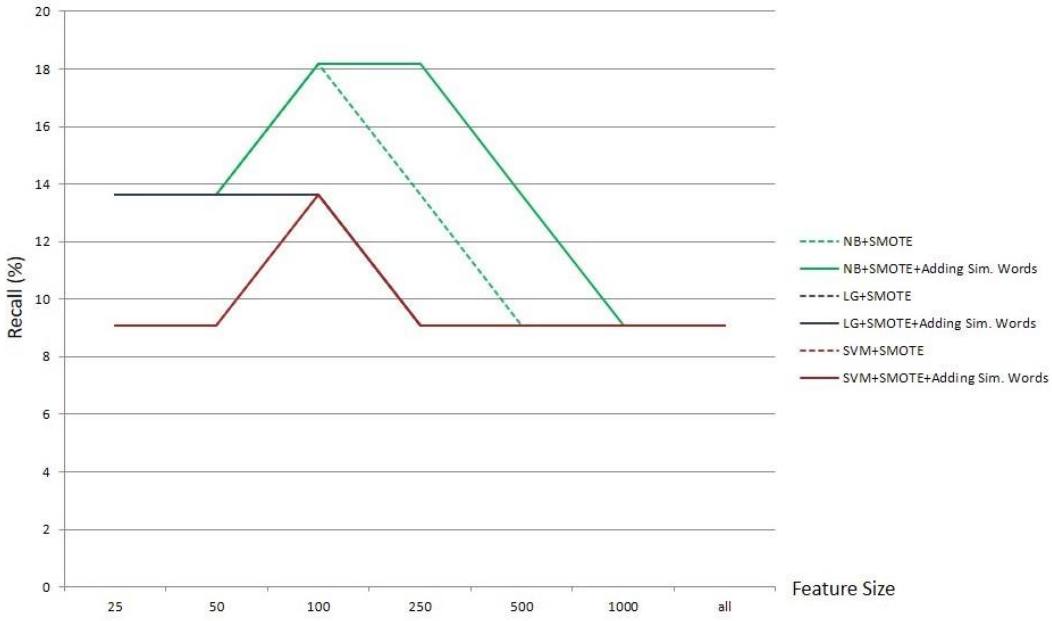
ภาพที่ 32 แสดงค่าระลึกของตัวแบบ ATTENTION โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



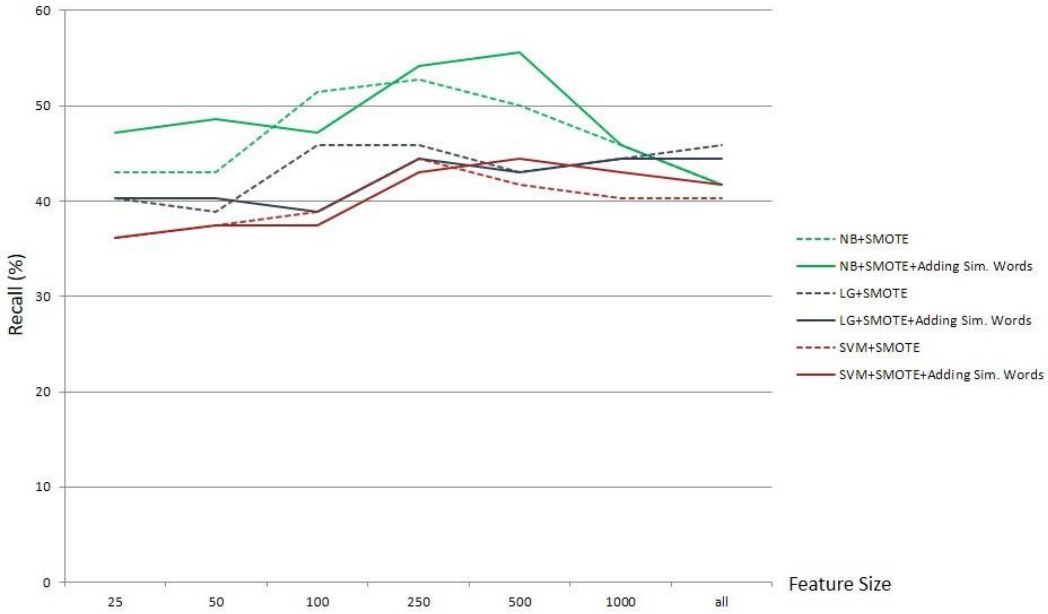
ภาพที่ 33 แสดงค่าระลึกของตัวแบบ INTEREST โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



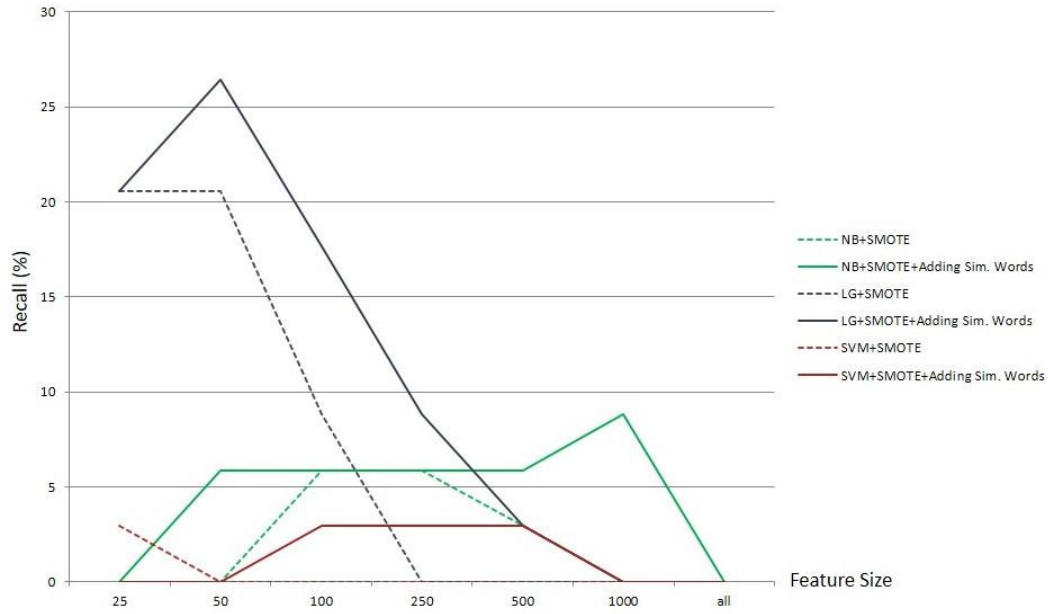
ภาพที่ 34 แสดงค่าระลึกของตัวแบบ SEARCH โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



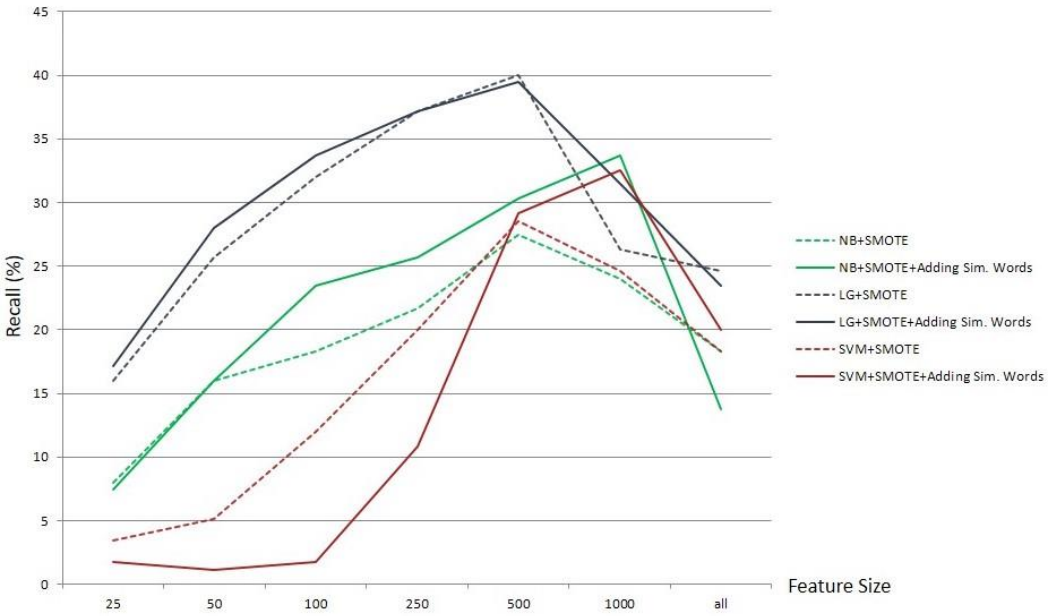
ภาพที่ 35 แสดงค่าระลึกของตัวแบบ ACTION โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



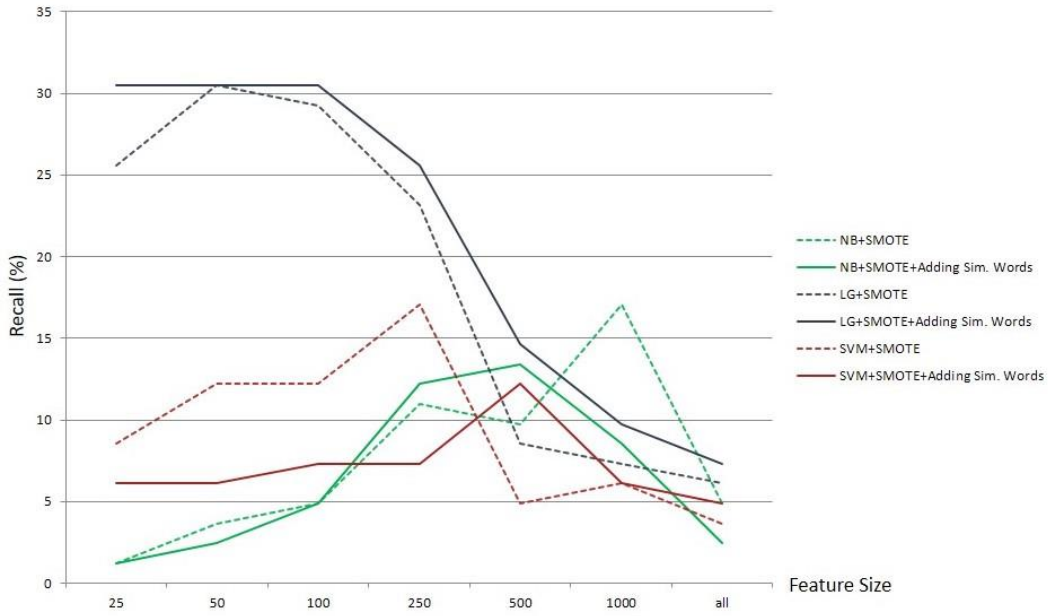
ภาพที่ 36 แสดงค่าระลึกของตัวแบบ SHARE โฆษณากลุ่มเครื่องสำอาง ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



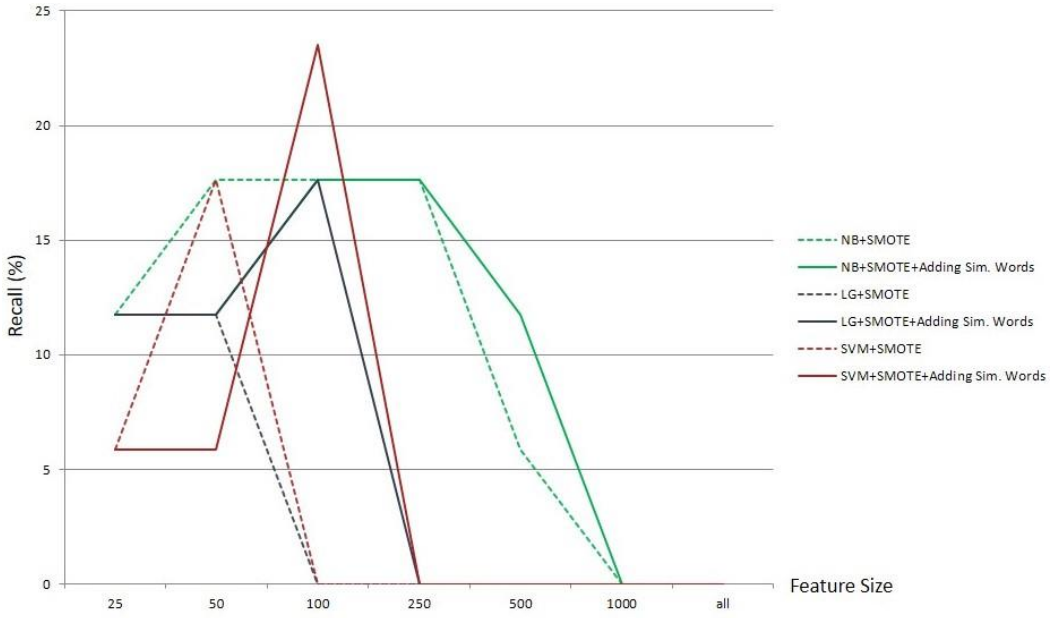
ภาพที่ 37 แสดงค่าระลึกของตัวแบบ ATTENTION โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



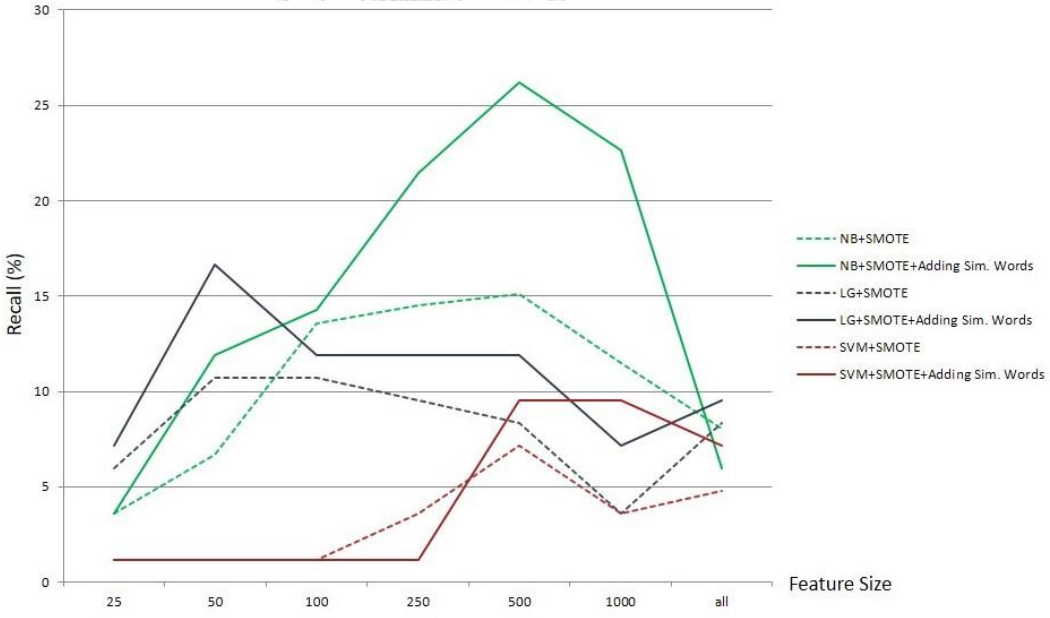
ภาพที่ 38 แสดงค่าระลึกของตัวแบบ INTEREST โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



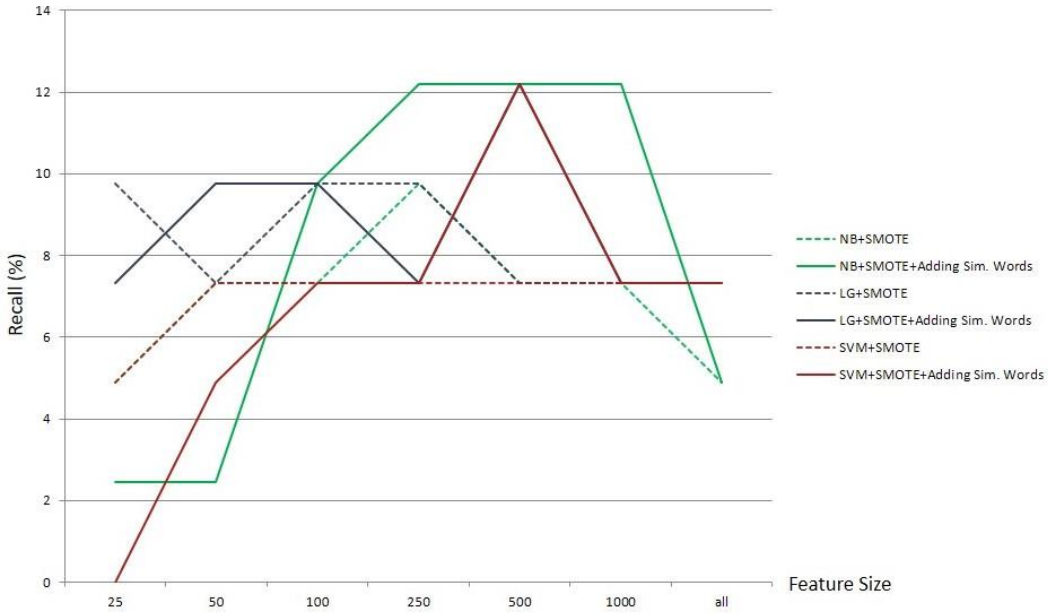
ภาพที่ 39 แสดงค่าระลึกของตัวแบบ SEARCH โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



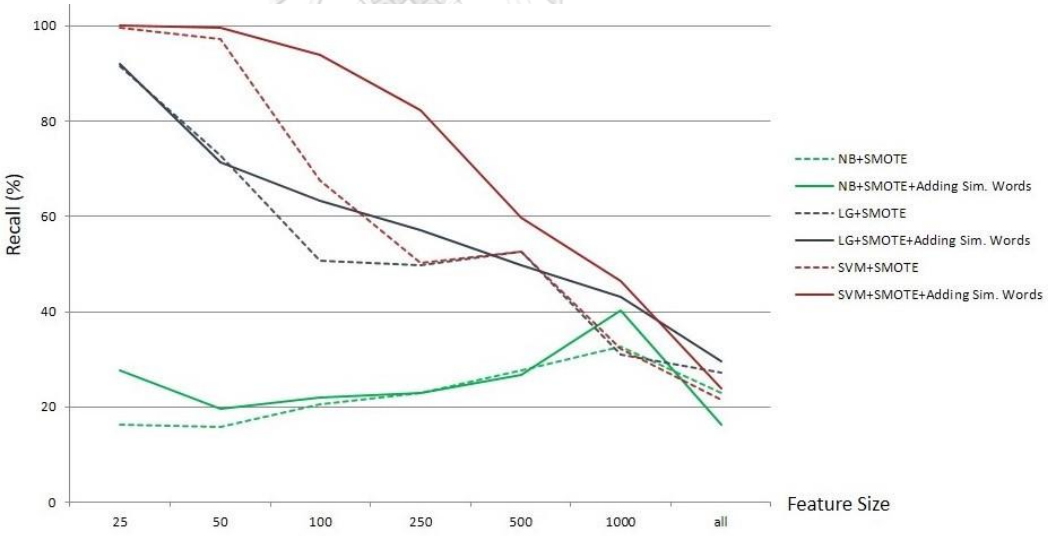
ภาพที่ 40 แสดงค่าระลึกของตัวแบบ ACTION โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



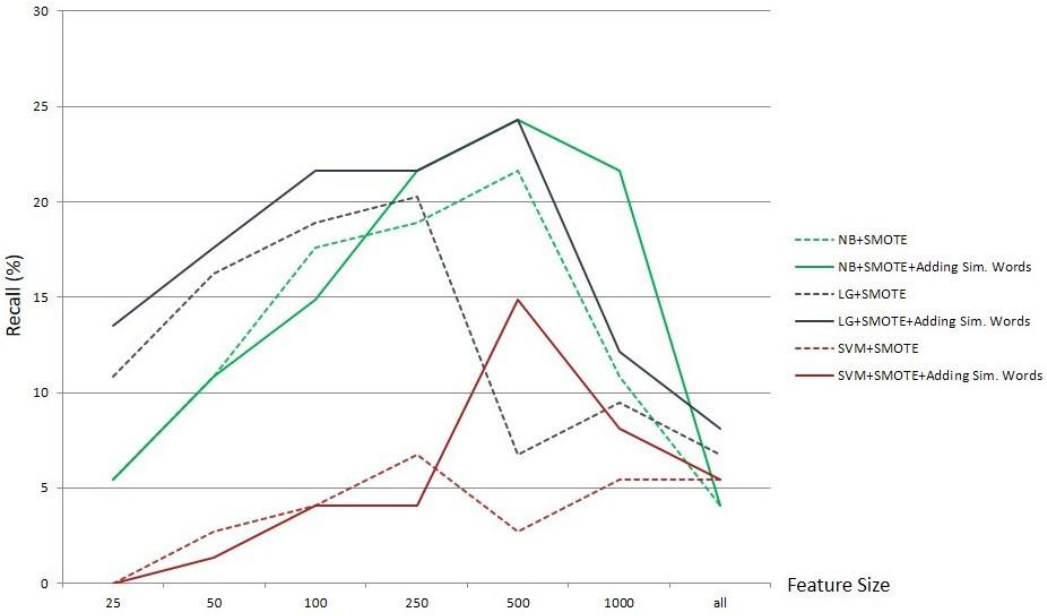
ภาพที่ 41 แสดงค่าระลึกของตัวแบบ SHARE โฆษณากลุ่มเครื่องใช้ไฟฟ้า ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



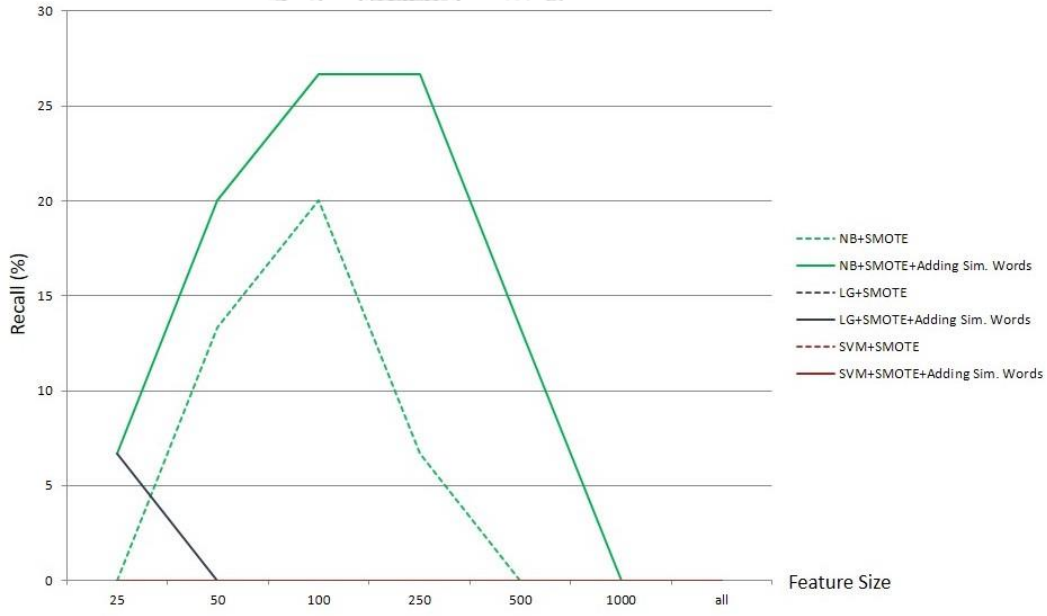
ภาพที่ 42 แสดงค่าระลึกของตัวแบบ ATTENTION โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



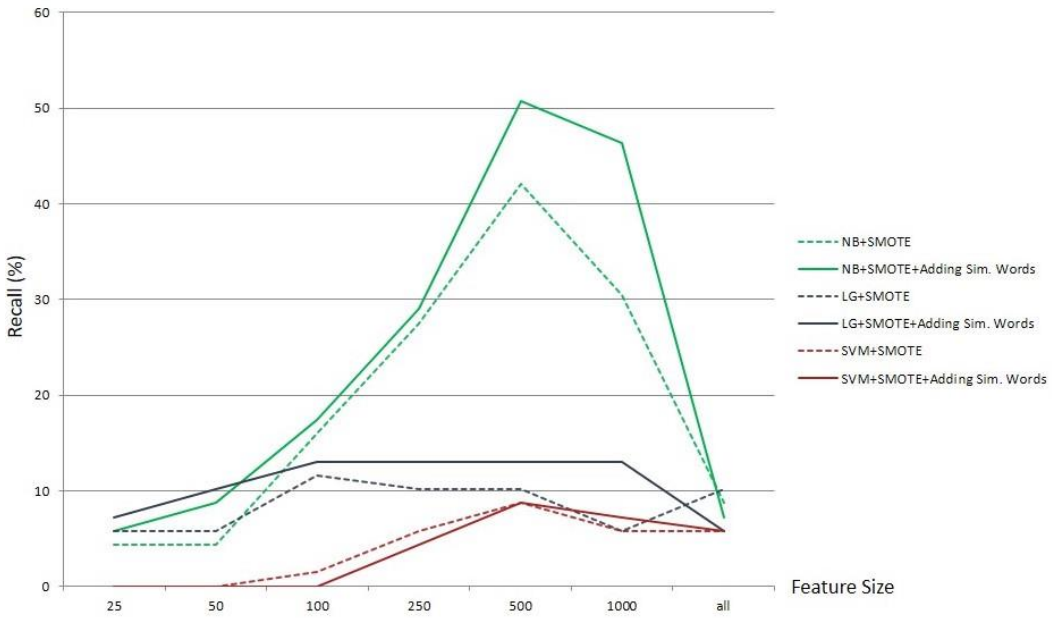
ภาพที่ 43 แสดงค่าระลึกของตัวแบบ INTEREST โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



ภาพที่ 44 แสดงค่าระลึกของตัวแบบ SEARCH โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



ภาพที่ 45 แสดงค่าระลึกของตัวแบบ ACTION โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย



ภาพที่ 46 แสดงค่าระลึกของตัวแบบ SHARE โฆษณากลุ่มสุขภัณฑ์ ก่อนและหลังการนำวิธีการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึงมาใช้ร่วมกับวิธีการคัดเลือกคุณลักษณะและการสุ่มเพิ่มตัวอย่างกลุ่มน้อย

4.4.2 เปรียบเทียบประสิทธิภาพในบริบทของค่าระลึกและค่าความแม่นโดยใช้การทดสอบที

ในหัวข้อนี้เป็นการทดสอบเปรียบเทียบตัวแบบทั้ง 15 ตัวแบบ (ตัวแบบทั้งหมดประกอบด้วย โฆษณากลุ่มเครื่องสำอาง 5 ตัวแบบ โฆษณากลุ่มเครื่องใช้ไฟฟ้า 5 ตัวแบบ และโฆษณากลุ่มสุขภัณฑ์ 5 ตัวแบบ) โดยในแต่ละตัวแบบ ใช้วิธีการเปรียบเทียบค่าระลึกของ 5-fold ระหว่างก่อนและหลังใช้วิธีการที่เสนอ (ก่อน : CHI2+SMOTE, หลัง : CHI2+SMOTE+Adding similar words) โดยใช้จำนวนค่าคุณลักษณะที่ความเหมาะสมที่สุด กล่าวคือ ให้ค่าระลึกสูงที่สุดเมื่อเปรียบเทียบทุกจำนวนคุณลักษณะ ซึ่งใช้วิธีการทดสอบที (t-test) แบบ paired sample t-test กับทุกตัวแบบและมีการใช้ระดับค่าอัลฟ่า (alpha) หลายระดับ ได้ผลลัพธ์ดังนี้

ตารางที่ 32 เปรียบเทียบจำนวนตัวแบบที่มีประสิทธิภาพเพิ่มขึ้น ลดลง และไม่แตกต่างจากเดิม ในบริบทของค่าระลึก หลังการประยุกต์ใช้วิธีการที่เสนอ

		นาอีฟเบย์			การถดถอยโลจิสติกส์			ซัพพอร์ตเวกเตอร์แมชชีน		
ค่าความเชื่อมั่น	alpha	จำนวนตัวแบบที่มีประสิทธิภาพสูงขึ้น	จำนวนตัวแบบที่มีประสิทธิภาพไม่แตกต่างจากเดิม	จำนวนตัวแบบที่มีประสิทธิภาพแย่ลง	จำนวนตัวแบบที่มีประสิทธิภาพสูงขึ้น	จำนวนตัวแบบที่มีประสิทธิภาพไม่แตกต่างจากเดิม	จำนวนตัวแบบที่มีประสิทธิภาพแย่ลง	จำนวนตัวแบบที่มีประสิทธิภาพสูงขึ้น	จำนวนตัวแบบที่มีประสิทธิภาพไม่แตกต่างจากเดิม	จำนวนตัวแบบที่มีประสิทธิภาพแย่ลง
0.95	0.05	5	9	1	1	14	0	2	13	0
0.9	0.1	7	7	1	1	14	0	5	9	1
0.85	0.15	7	7	1	1	14	0	5	9	1
0.8	0.2	10	4	1	6	9	0	8	5	2

ต่อมาใช้วิธีการเดียวกับการทดสอบข้างต้น แต่เปรียบเทียบในแง่ของค่าความแม่น ได้ผลลัพธ์ดังต่อไปนี้

ตารางที่ 33 เปรียบเทียบจำนวนตัวแบบที่มีประสิทธิภาพเพิ่มขึ้น ลดลง และไม่แตกต่างจากเดิม ในบริบทของค่าความแม่น หลังการประยุกต์ใช้วิธีการที่เสนอ

		นาอีฟเบย์			การถดถอยโลจิสติกส์			ซัพพอร์ตเวกเตอร์แมชชีน		
ค่าความเชื่อมั่น	alpha	จำนวนตัวแบบที่มีประสิทธิภาพสูงขึ้น	จำนวนตัวแบบที่มีประสิทธิภาพไม่แตกต่างจากเดิม	จำนวนตัวแบบที่มีประสิทธิภาพแย่ลง	จำนวนตัวแบบที่มีประสิทธิภาพสูงขึ้น	จำนวนตัวแบบที่มีประสิทธิภาพไม่แตกต่างจากเดิม	จำนวนตัวแบบที่มีประสิทธิภาพแย่ลง	จำนวนตัวแบบที่มีประสิทธิภาพสูงขึ้น	จำนวนตัวแบบที่มีประสิทธิภาพไม่แตกต่างจากเดิม	จำนวนตัวแบบที่มีประสิทธิภาพแย่ลง
0.95	0.05	1	13	1	0	15	0	0	14	1
0.9	0.1	2	11	2	1	11	3	1	13	1
0.85	0.15	3	10	2	2	10	3	1	12	2
0.8	0.2	3	9	3	5	7	3	2	11	2

พบว่าหลังการประยุกต์ใช้วิธีการที่นำเสนอ จากการใช้วิธีการทดสอบทีพบว่ามีตัวแบบจำนวนหนึ่งมีค่าระลึกเพิ่มขึ้น โดยนาอีฟเบย์มีจำนวนตัวแบบที่มีประสิทธิภาพเพิ่มขึ้นมากที่สุด อย่างไรก็ตามการลดระดับค่าความเชื่อมั่น (เพิ่มค่าอัลฟาขึ้น) จะพบว่าจำนวนตัวแบบที่มีค่าระลึกเพิ่มขึ้นมีจำนวนเพิ่มมากขึ้น โดยในระดับความเชื่อมั่น 0.8 จะพบว่านาอีฟเบย์มีจำนวนตัวแบบที่มีประสิทธิภาพเพิ่มขึ้นถึง 10 ตัวแบบ อย่างไรก็ตามยังประกอบด้วยตัวแบบที่ประสิทธิภาพไม่แตกต่างไปจากเดิมสี่ตัวแบบ และมีประสิทธิภาพลดลงหนึ่งตัวแบบ

สำหรับตัวแบบจำนวนหนึ่งที่มีประสิทธิภาพไม่แตกต่างจากเดิมหรือมีประสิทธิภาพลดลงเกิดขึ้นเนื่องจากวิธีการที่เสนอ นั่นคือการเพิ่มคำคล้ายคลึงเข้าไปในชุดข้อมูลดังกล่าว ไม่ได้เพิ่มคำที่ประกอบด้วยคุณลักษณะที่เกี่ยวข้องกับคลาสเข้าไปด้วยตามสมมติฐาน

ในบริบทของค่าความแม่น พบว่าหลังการใช้วิธีการเสนอ ไม่สามารถสรุปได้อย่างชัดเจน เนื่องจากตัวแบบส่วนใหญ่ไม่ได้มีประสิทธิภาพแตกต่างไปจากเดิม อีกทั้งตัวแบบที่มีประสิทธิภาพเพิ่มขึ้นและลดลงยังมีจำนวนใกล้เคียงกัน

4.4.3 สรุปการเปรียบเทียบประสิทธิภาพระหว่าง CHI2+SMOTE และ Adding Similar Words+CHI2+SMOTE

สรุปการเปรียบเทียบประสิทธิภาพก่อนและหลังใช้วิธีการที่เสนอในแง่ของค่าระลึก เมื่อคัดเลือกจำนวนคุณลักษณะที่เหมาะสมที่สุดสำหรับแต่ละวิธี เป็นดังต่อไปนี้

ตารางที่ 34 สรุปค่าระลึกเปรียบเทียบก่อนและหลังประยุกต์ใช้วิธีการที่เสนอ

	นาอีฟเบย์		การถดถอยโลจิสติกส์		ซัพพอร์ตเวกเตอร์แมชชีน	
	CHI2+SMOTE	Adding Similar Words+CHI2+SMOTE	CHI2+SMOTE	Adding Similar Words+CHI2+SMOTE	CHI2+SMOTE	Adding Similar Words+CHI2+SMOTE
โฆษณากลุ่มเครื่องสำอาง						
Attention	11.11	13.89	22.22	**25**	16.67	11.11
Interest	34.20	31.09	42.49	**43.52**	43.01	38.86
Search	18.60	24.42**	22.09	**29.07**	17.44	23.26**
Action	**18.18**	**18.18**	13.64	13.64	13.64	13.64
Share	52.78	**55.56****	45.83	44.44	44.44	44.44
โฆษณากลุ่มเครื่องใช้ไฟฟ้า						
Attention	5.88	8.82	20.59	**26.47**	2.94	2.94
Interest	27.43	33.71**	**40.00**	39.43	28.57	32.57*
Search	17.07	13.41	**30.49**	**30.49**	17.07	12.20
Action	17.65	17.65	11.76	17.65	17.65	**23.53**
Share	20.24	**26.19***	10.71	16.67**	7.14	9.52
โฆษณากลุ่มสุขภัณฑ์						
Attention	9.76	**12.20**	10.26	**12.20**	7.32	**12.20***
Interest	32.54	40.19**	91.39	91.87	99.52	**100**
Search	21.62	**24.32***	20.27	**24.32**	6.76	14.86**
Action	20.00	**26.67**	0.10	6.67	0.10	0
Share	42.03	**50.72****	11.59	13.04	8.70	8.70

*เมื่อเทียบประสิทธิภาพกับวิธีการก่อนหน้า (CHI2+SMOTE) ด้วยวิธี 1-tail paired sample T-test พบว่ามีค่าเพิ่มขึ้นที่ระดับนัยสำคัญ 0.1

**เมื่อเทียบประสิทธิภาพกับวิธีการก่อนหน้า (CHI2+SMOTE) ด้วยวิธี 1-tail paired sample T-test มีค่าเพิ่มขึ้นที่ระดับนัยสำคัญ 0.05

สรุปการเปรียบเทียบประสิทธิภาพก่อนและหลังใช้วิธีการที่เสนอในแง่ของค่าความแม่น เมื่อคัดเลือกจำนวนคุณลักษณะที่เหมาะสมที่สุดสำหรับแต่ละวิธี เป็นดังต่อไปนี้

ตารางที่ 35 สรุปค่าความแม่นเปรียบเทียบก่อนและหลังหลังประยุกต์ใช้วิธีการที่เสนอ

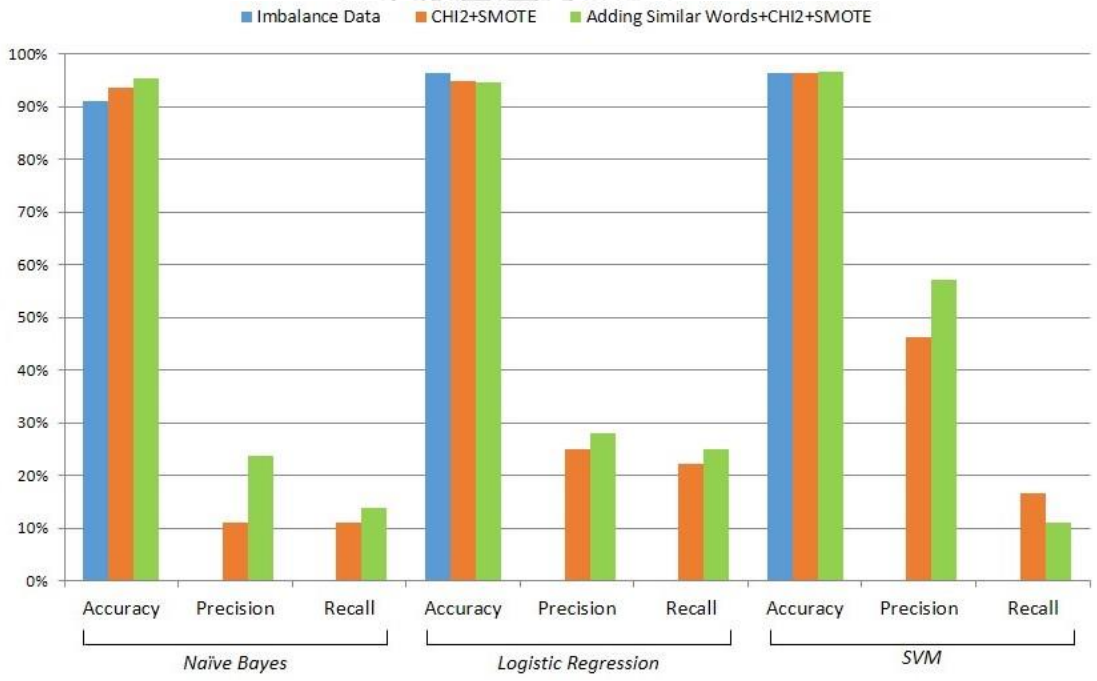
	นาอีฟเบย์		การถดถอยโลจิสติกส์		ซัพพอร์ตเวกเตอร์แมชชีน	
	CHI2+SMOTE	Adding Similar Words+CHI2+ SMOTE	CHI2+SMOTE	Adding Similar Words+CHI2+ SMOTE	CHI2+SMOTE	Adding Similar Words+CHI2+ SMOTE
โฆษณากลุ่มเครื่องสำอาง						
Attention	11.11	23.81	25.00	28.12	46.15	**57.14**
Interest	22.45	23.81	25.15	**25.45**	23.18	22.46
Search	8.79	14.04	**19.79**	13.02	11.54	10.10
Action	11.76	8	15.79	13.64	9.68	**16.67**
Share	31.39	27.21	58.93	**65.31**	60.38	47.06
โฆษณากลุ่มเครื่องใช้ไฟฟ้า						
Attention	1.77	2.91*	**26.92**	9.78	11.11	7.69
Interest	21.52	21.93	**26.92**	25.84	25.13	26.51
Search	9.72	8.15	19.69	**20**	16.47	12.35
Action	10.34	6.25	11.76	12.50	9.68	**13.33***
Share	12.06	12.02	10.50	**16.87***	10.17	13.33
โฆษณากลุ่มสุขภัณฑ์						
Attention	5.00	5.49	9.76	16.67	10.34	**20.83**
Interest	22.22	**24.14****	22.37	22.35	19.08	21.88
Search	**14.68**	13.24	13.27	10.59	13.51	8.66
Action	6.38	7.02	0.	**7.69**	0	0
Share	14.50	**14.83**	8.70	10	6.25	6.38

*เมื่อเทียบประสิทธิภาพกับวิธีการก่อนหน้า (CHI2+SMOTE) ด้วยวิธี 1-tail paired sample T-test พบว่ามีค่าเพิ่มขึ้นที่ระดับนัยสำคัญ 0.1

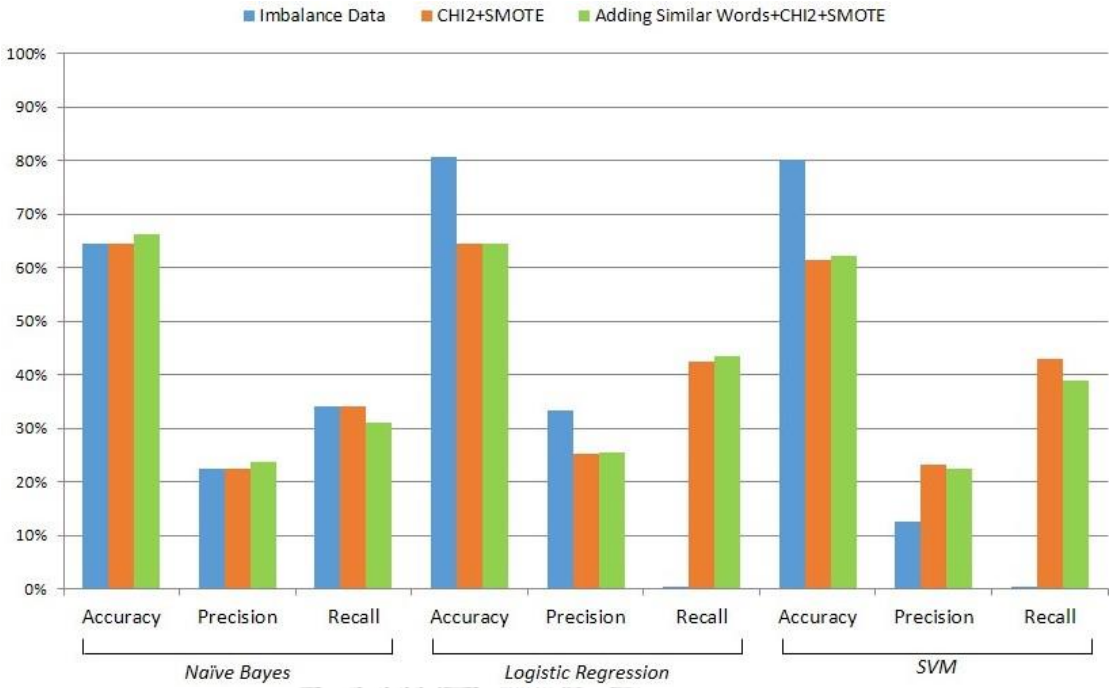
**เมื่อเทียบประสิทธิภาพกับวิธีการก่อนหน้า (CHI2+SMOTE) ด้วยวิธี 1-tail paired sample T-test มีค่าเพิ่มขึ้นที่ระดับนัยสำคัญ 0.05

4.5 เปรียบเทียบสามการทดลองแรก (Original Imbalanced Dataset, Balanced Dataset using CHI2+SMOTE, Balanced Dataset using Adding Similar Words+CHI2+SMOTE)

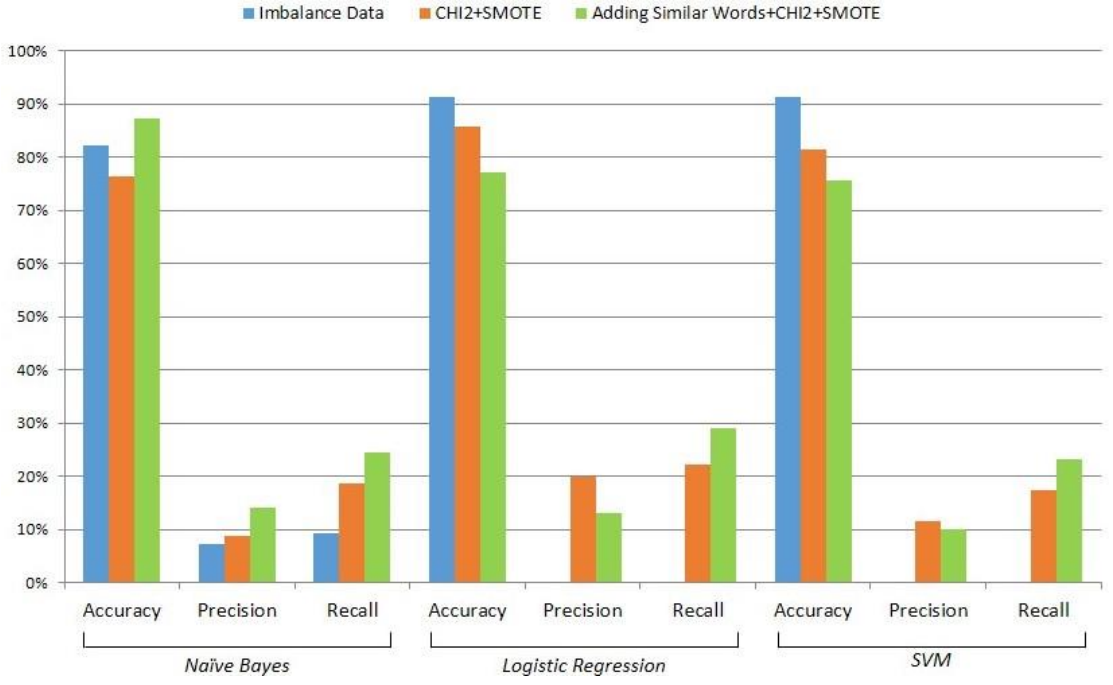
ในหัวข้อนี้ ผู้วิจัยได้เปรียบเทียบผลลัพธ์ที่ได้จากการทดลองทั้งสามการทดลองซึ่งใช้ชุดข้อมูลที่แตกต่างกัน ได้แก่ (1) ชุดข้อมูลที่ไม่สมดุล (Original Imbalanced Dataset) (2) ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะ (Balanced Dataset using CHI2+SMOTE) และ (3) ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะและการเพิ่มคุณลักษณะที่เป็นคำคล้ายคลึง (Balanced Dataset using Adding Similar Words+CHI2+SMOTE) โดยสำหรับการทดลองที่ (2) และ (3) ผู้วิจัยได้เลือกผลลัพธ์ค่าความถูกต้อง ค่าความแม่น และค่าระลึก เมื่อใช้จำนวนคุณลักษณะที่มีความเหมาะสมที่ทำให้ตัวแบบมีประสิทธิภาพในการทำนายตัวอย่างที่เป็นคลาสบวกมากที่สุด กล่าวคือเป็นจำนวนคุณลักษณะที่ทำให้ตัวแบบได้ค่าระลึกสูงที่สุด ซึ่งอาจแตกต่างกันไปในแต่ละชุดข้อมูลและแต่ละตัวแบบการเรียนรู้ของเครื่องทั้งสามวิธีการ



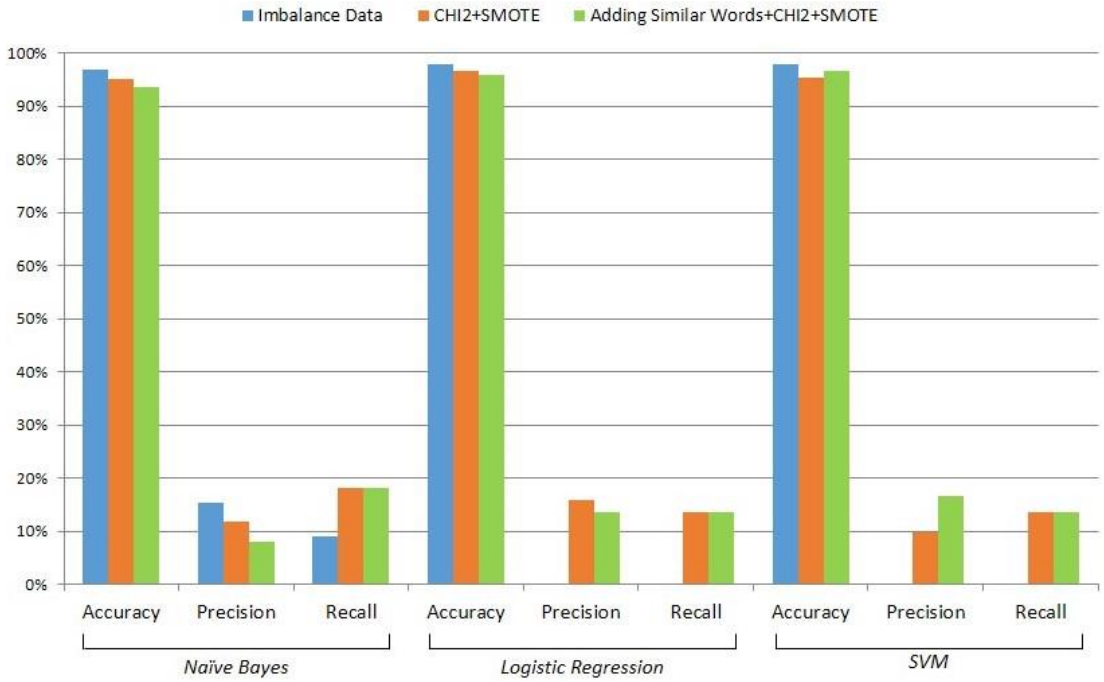
ภาพที่ 47 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องสำอาง ตัวแบบ ATTENTION



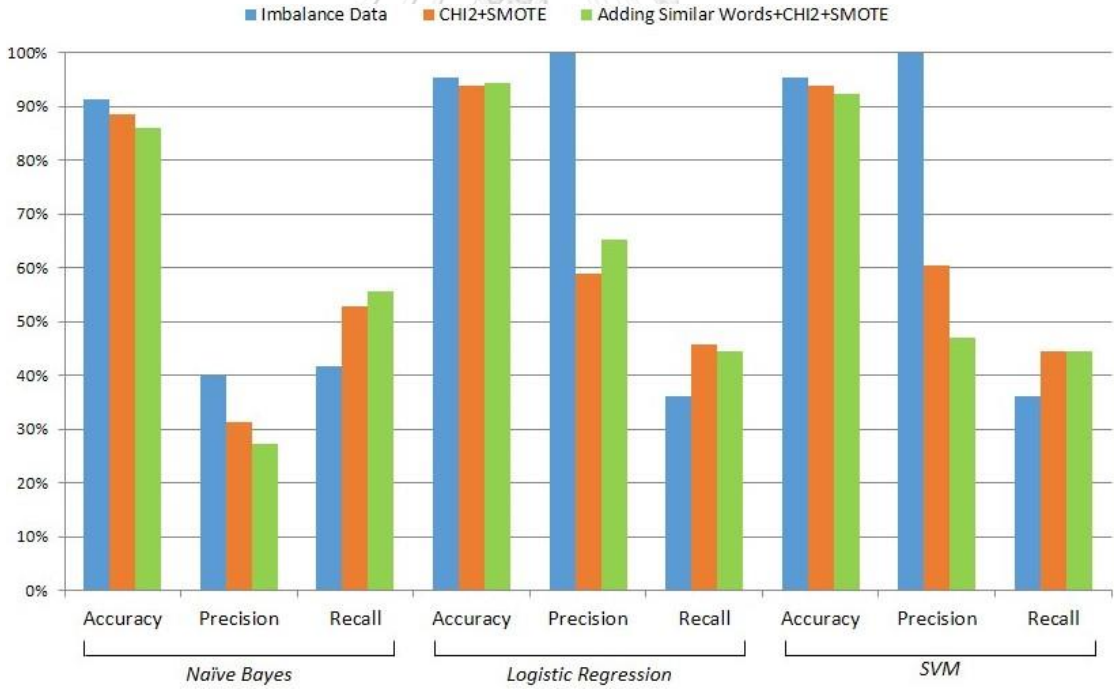
ภาพที่ 48 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องสำอาง ตัวแบบ INTEREST



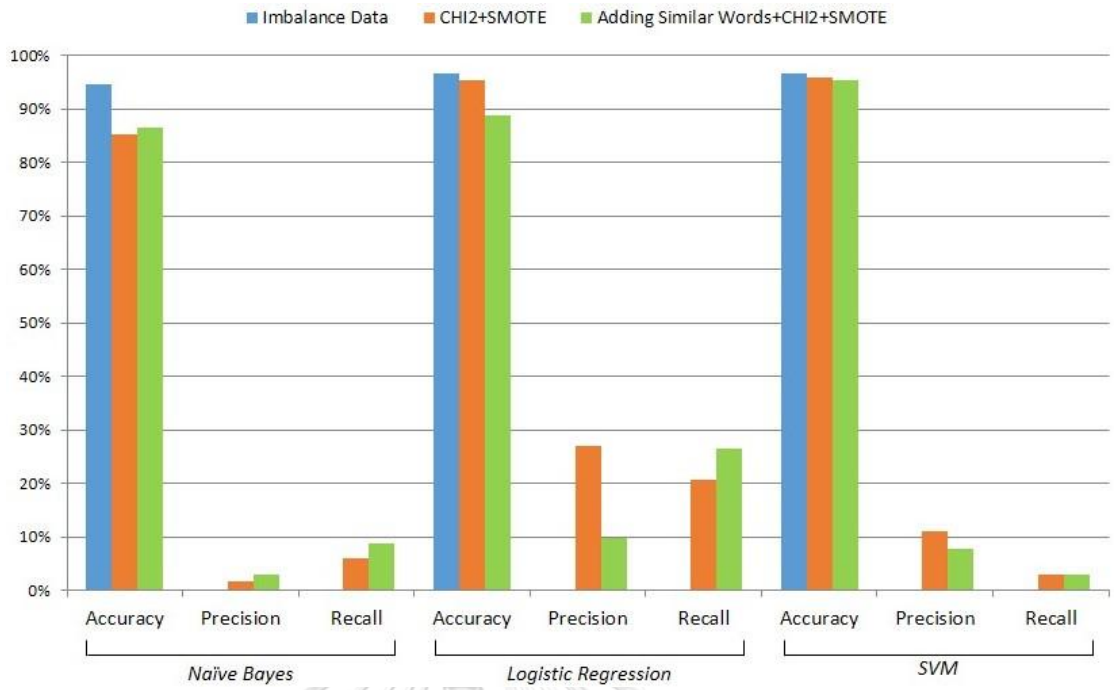
ภาพที่ 49 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องสำอาง ตัวแบบ SEARCH



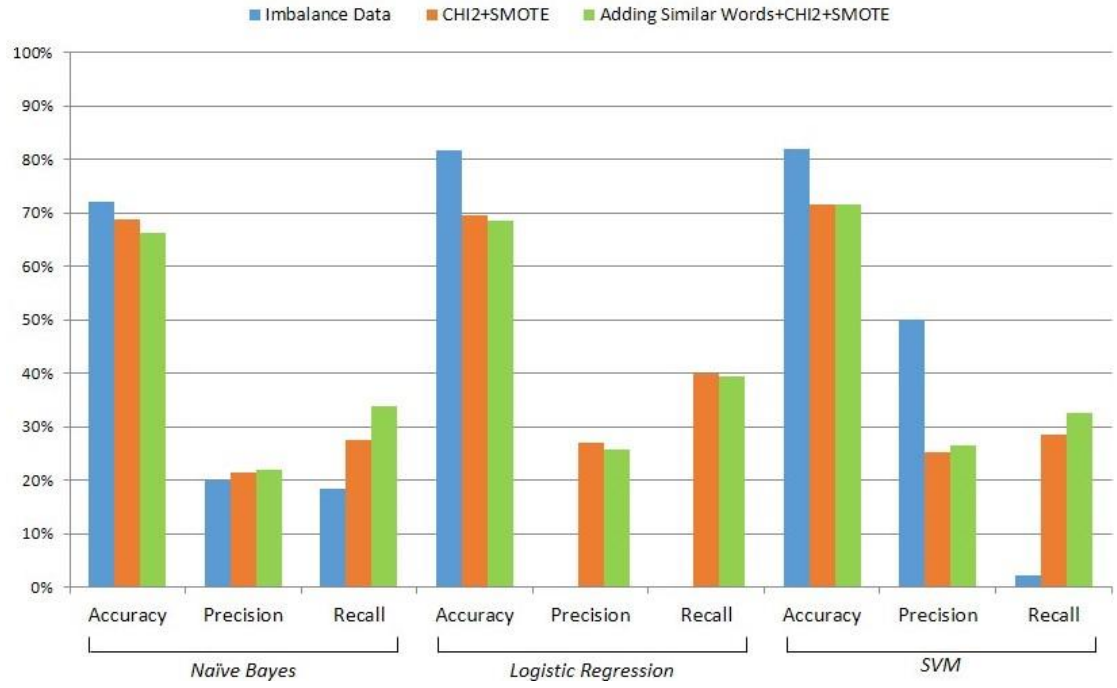
ภาพที่ 50 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องสำอาง ตัวแบบ ACTION



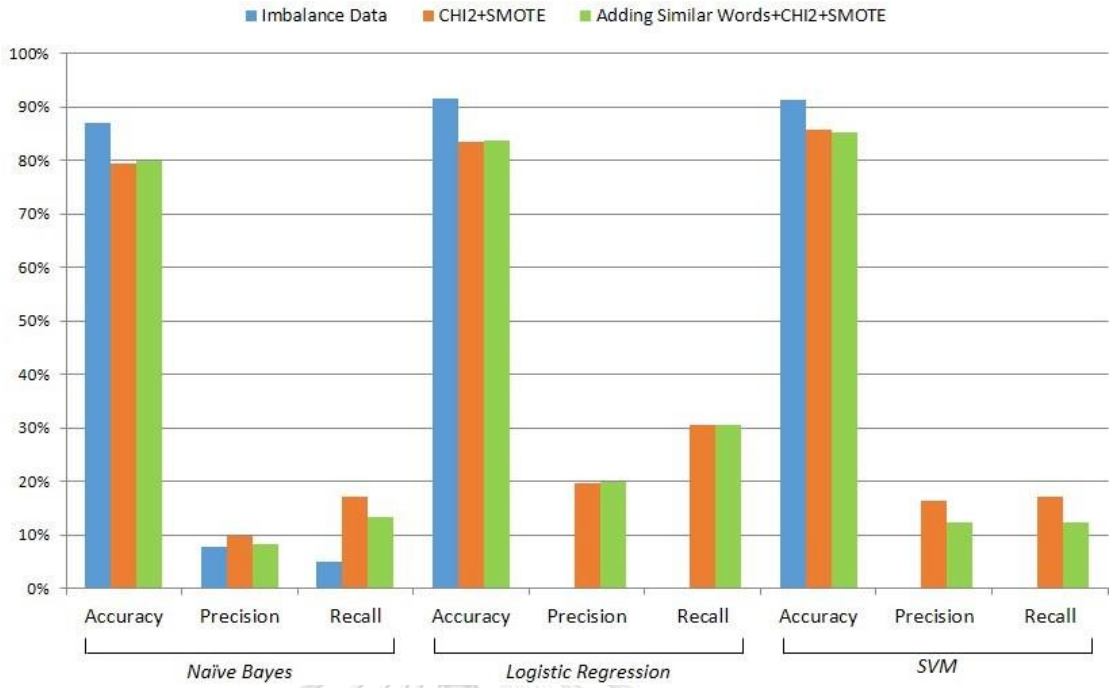
ภาพที่ 51 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องสำอาง ตัวแบบ SHARE



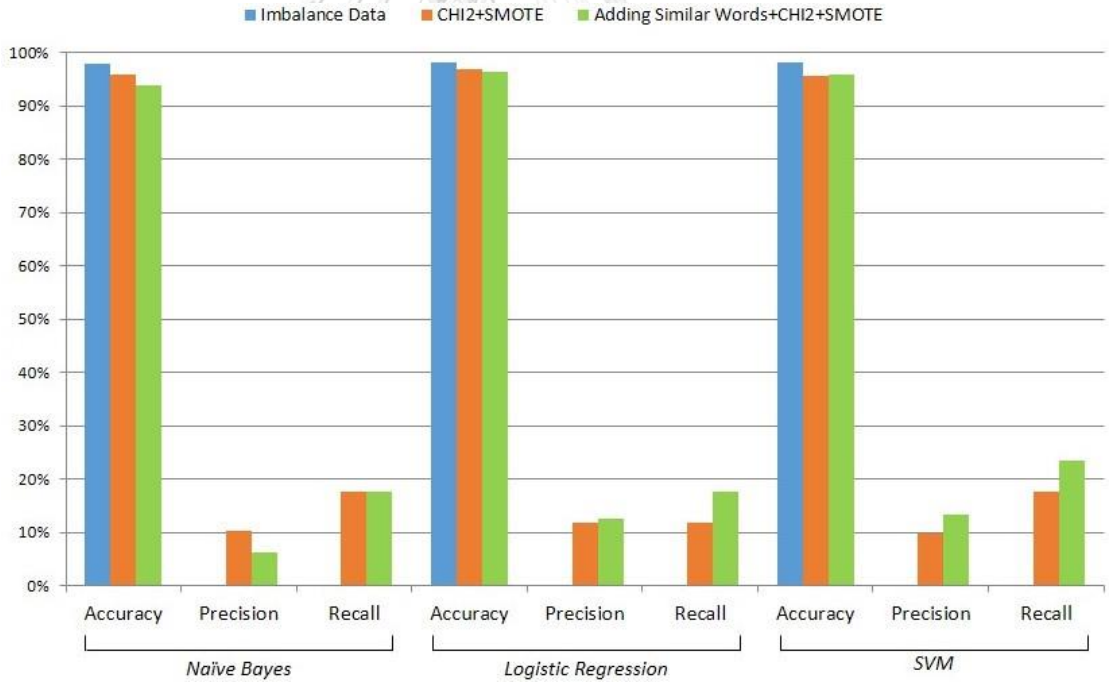
ภาพที่ 52 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องใช้ไฟฟ้า ตัวแบบ ATTENTION



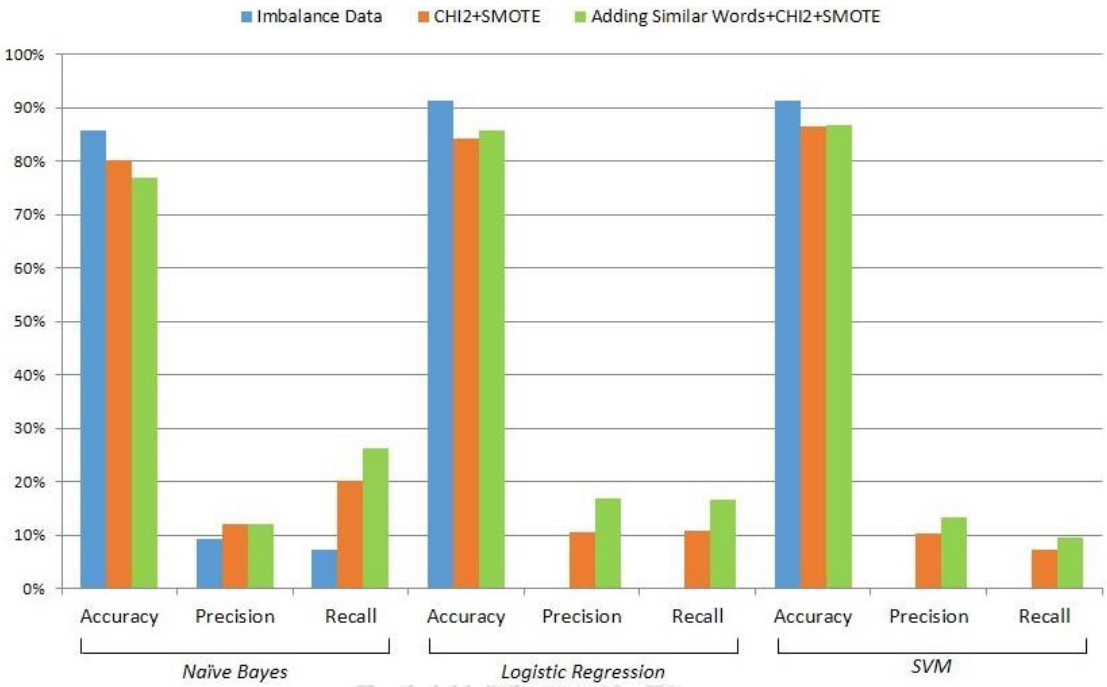
ภาพที่ 53 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องใช้ไฟฟ้า ตัวแบบ INTEREST



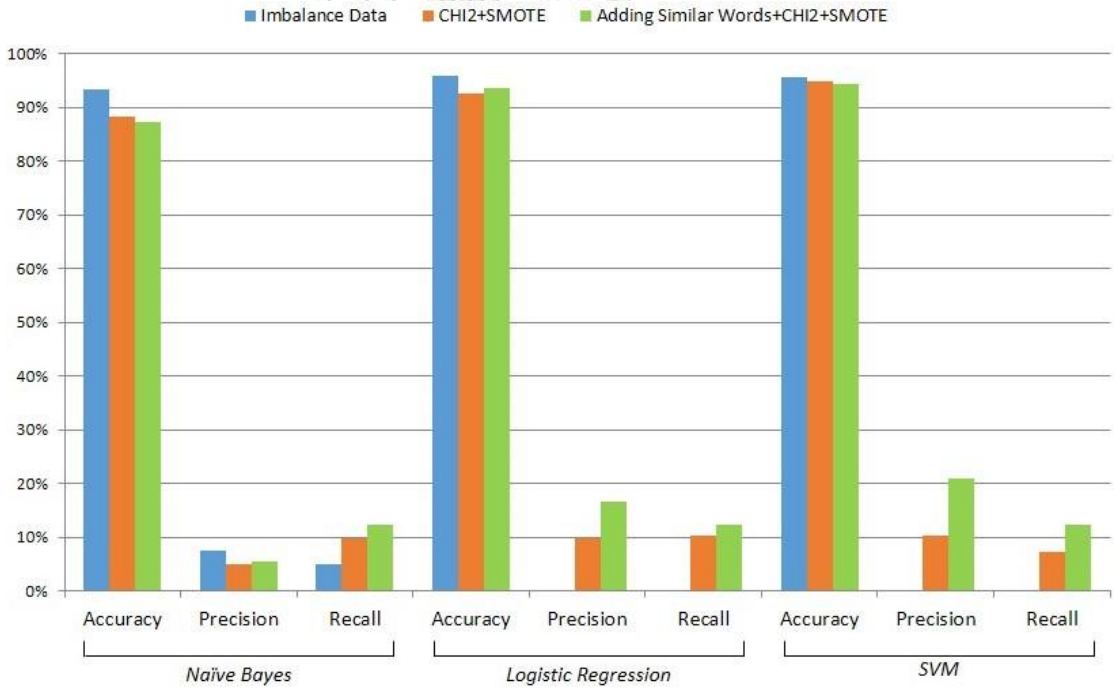
ภาพที่ 54 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องใช้ไฟฟ้า ตัวแบบ SEARCH



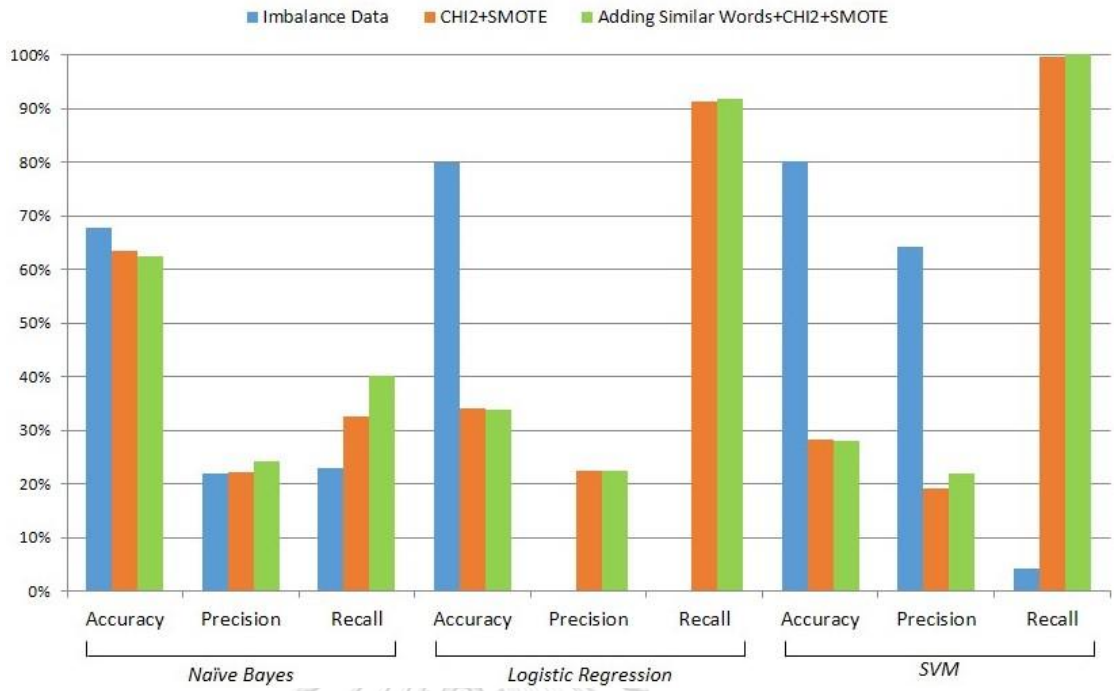
ภาพที่ 55 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องใช้ไฟฟ้า ตัวแบบ ACTION



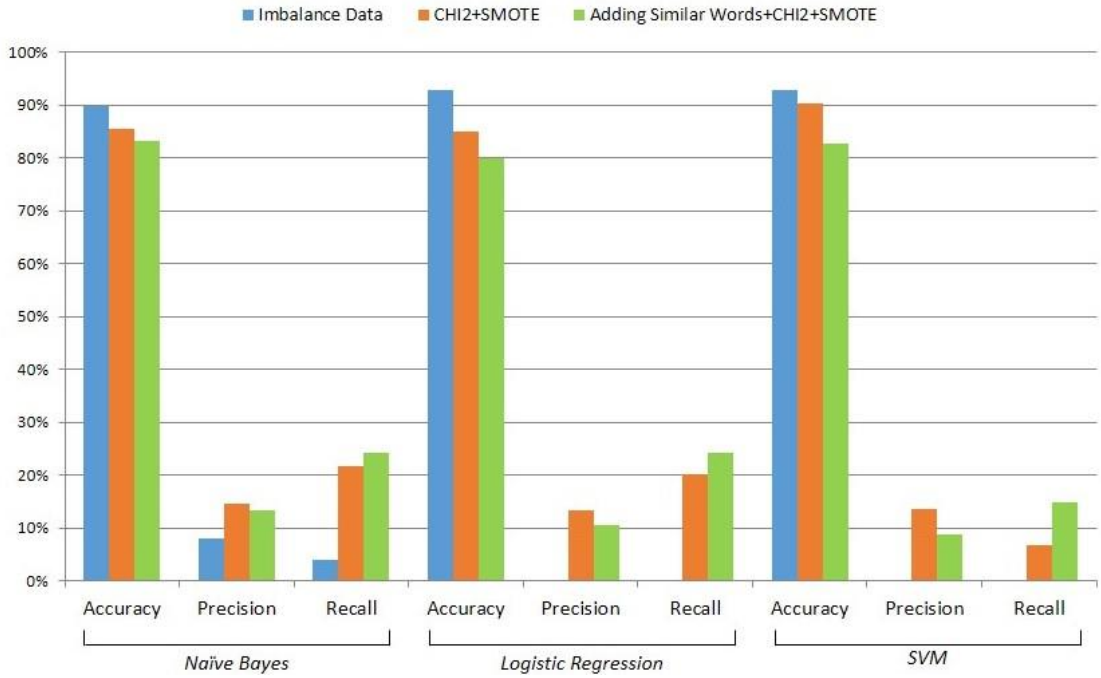
ภาพที่ 56 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มเครื่องใช้ไฟฟ้า ตัวแบบ SHARE



ภาพที่ 57 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มสุขภัณฑ์ ตัวแบบ ATTENTION

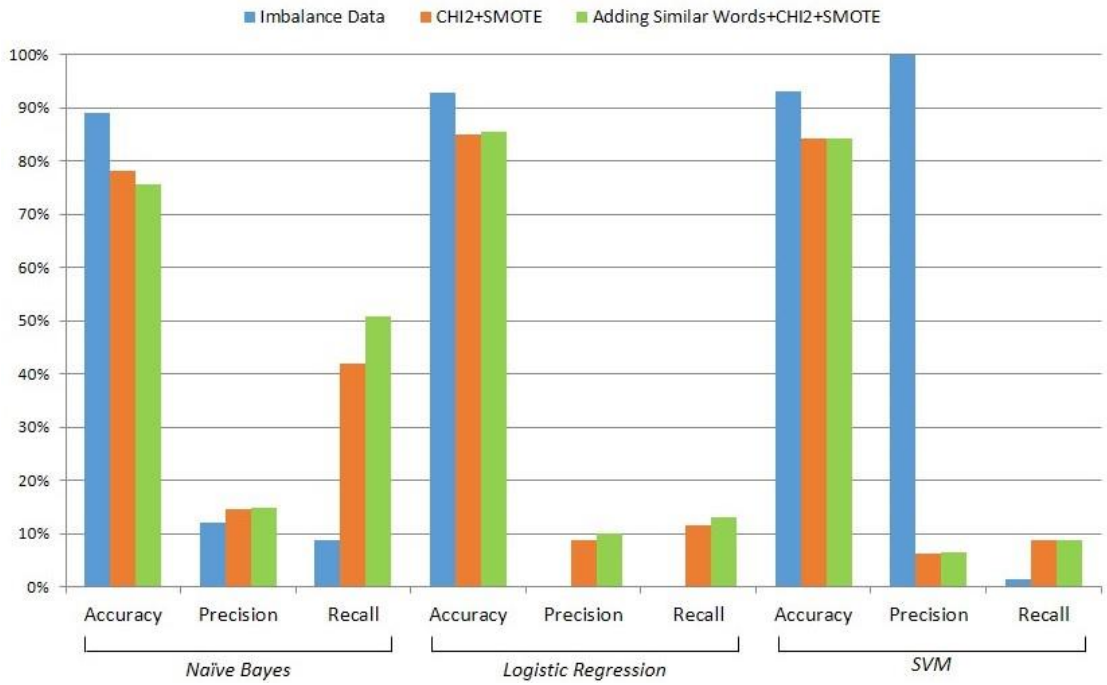


ภาพที่ 58 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มสุขภัณฑ์ ตัวแบบ INTEREST



ภาพที่ 59 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มสุขภัณฑ์ ตัวแบบ SEARCH

ภาพที่ 60 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มสุขภัณฑ์ ตัวแบบ ACTION



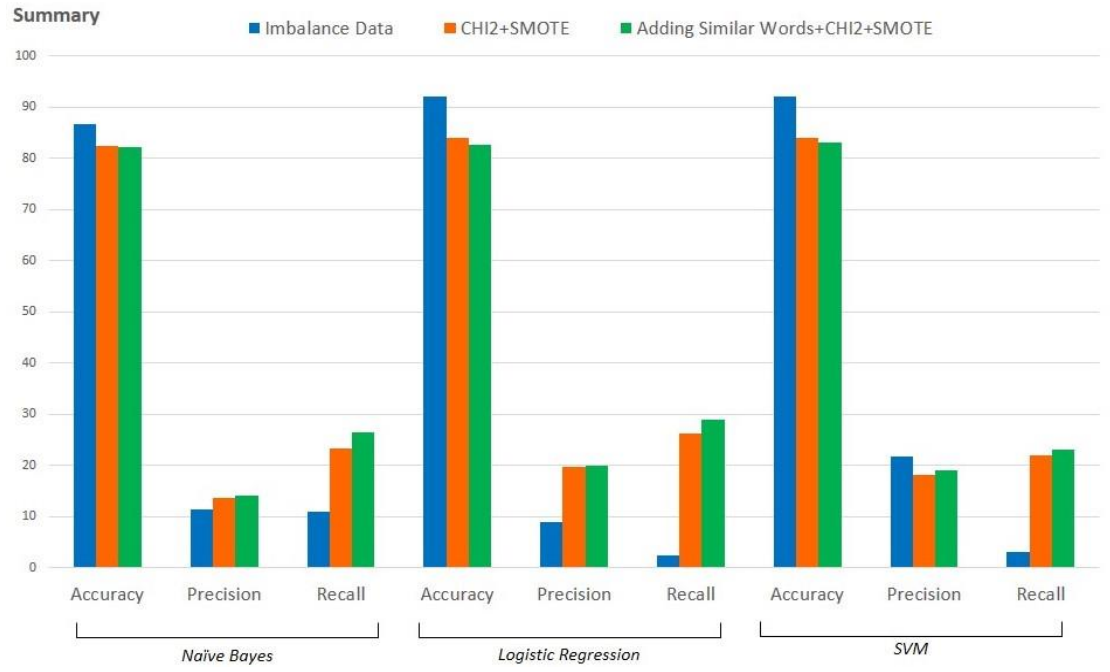
ภาพที่ 61 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง โฆษณากลุ่มสุขภัณฑ์ ตัวแบบ SHARE

ต่อไปนี้เป็นการเฉลี่ยรวมผลลัพธ์ของทั้ง 15 ตัวแบบ และเป็นการเปรียบเทียบผลลัพธ์ระหว่างชุดข้อมูลที่ไม่สมดุล (Original Imbalanced Dataset) กับชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะ (Balanced Dataset using CHI2+SMOTE)

และการเปรียบเทียบผลลัพธ์ระหว่างชุดข้อมูลที่ไม่สมดุล กับชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะและการเพิ่มคุณลักษณะที่เป็นคำคล้ายคลึง (Balanced Dataset using Adding Similar Words+CHI2+SMOTE) ซึ่งสามารถแสดงดังตารางและกราฟต่อไปนี้

ตารางที่ 36 แสดงผลลัพธ์เฉลี่ยของทุกการทดลองเปรียบเทียบกัน

ตัวแบบ	ผลลัพธ์ (%)	(1) ชุดข้อมูลที่ไม่สมดุล	(2) CHI2+SMOTE	(3) Adding Similar Words+CHI2+SMOTE	เปรียบเทียบ (1) และ (2)	เปรียบเทียบ (1) และ (3)
นาอีฟเบย์	ความถูกต้อง	86.76	82.49	82.28	-4.27	-4.48
	ค่าความแม่น	11.47	13.58	14.19	**+2.11**	**+2.72**
	ค่าระลึก	11.04	23.27	26.47	**+12.23**	**+15.43**
การถดถอยโลจิสติกส์	ความถูกต้อง	92.04	83.91	82.64	-8.13	-9.4
	ค่าความแม่น	8.98	19.64	19.86	**+10.67**	**+10.88**
	ค่าระลึก	2.53	26.23	28.97	**+23.7**	**+26.44**
ซัพพอร์ตเวกเตอร์แมชชีน	ความถูกต้อง	92.02	83.94	83.02	-8.09	-9
	ค่าความแม่น	21.85	18.18	18.96	-3.67	-2.89
	ค่าระลึก	3.04	22.06	23.19	**+19.02**	**+20.14**



ภาพที่ 62 กราฟเปรียบเทียบผลลัพธ์แต่ละการทดลอง เฉลี่ยผลลัพธ์ของทุกตัวแบบ

ในภาพรวม หลังจากการประยุกต์ใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะจะเห็นได้ว่าทั้งค่าความแม่นและค่าระลึกของทุกตัวแบบจำแนกประเภทเพิ่มขึ้น แต่ค่าความถูกต้องลดลง ค่าระลึกสามารถเพิ่มขึ้นได้อีกเล็กน้อยในบางตัวแบบหลังจากการประยุกต์ใช้เทคนิคสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วมกับเทคนิคคัดเลือกคุณลักษณะและการเพิ่มคุณลักษณะที่เป็นคำคล้ายคลึง อย่างไรก็ตาม หลังจากการประยุกต์ใช้วิธีการต่าง ๆ แล้ว ในแง่ของค่าความแม่นและค่าระลึก ตัวแบบการถดถอยโลจิสติกส์ให้ผลลัพธ์สูงสุด และมีค่าเพิ่มขึ้นมากที่สุดเมื่อเปรียบเทียบกับก่อนหน้าการประยุกต์ใช้วิธีการดังกล่าว

4.6 ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคการสร้างข้อความ (Balanced Dataset using Text Generation)

ต่อไปนี้เป็นผลลัพธ์หลังการทำให้ชุดข้อมูลสมดุลด้วยเทคนิคการสร้างข้อความแบบมาร์คอฟเชนและแบบแอลเอสทีเอ็ม โดยใช้เทคนิคการสุ่มเพิ่มตัวอย่างแบบ oversampling เป็นบรรทัดฐาน แสดงให้เห็นว่าวิธีการดังกล่าวสามารถทำให้ตัวแบบทำนายตัวอย่างคลาสบวกได้เป็นจำนวนมากขึ้น และวิธีการมาร์คอฟเชนให้ผลลัพธ์ดีที่สุดในแง่ของค่าความแม่นและค่าระลึกในชุดข้อมูลส่วนใหญ่

ตารางที่ 37 แสดงผลลัพธ์ของตัวแบบจำแนกประเภทแอลเอสทีเอ็ม เปรียบเทียบระหว่างชุดข้อมูลที่ไม่สมดุลกับชุดข้อมูลที่สมดุลด้วยเทคนิคการสร้างข้อความแต่ละวิธี

	ชุดข้อมูลที่ไม่สมดุล						ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคการสร้างข้อความ					
การสร้างข้อความ	-			การสุ่มเพิ่มตัวอย่างแบบ oversampling			มาร์คอฟเชน			LSTM		
ตัวแบบ	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	ความถูกต้อง	ค่าความแม่น	ค่าระลึก	ความถูกต้อง	ค่าความแม่น	ค่าระลึก
โฆษณากลุ่มเครื่องสำอาง												
Attention	98	0	0	92	0	0	81.5	**2.86**	**25**	97	0	0
Interest	74.5	9.09	26.67	81	10.34	20	75.5	**16**	**53.33**	74	12.24	40
Search	92	**12.5**	10	81.5	6.45	**20**	78.5	5.41	**20**	78.5	5.41	**20**
Action	99	0	0	93	0	0	76.5	**2.13**	**50**	94.5	0	0
Share	96.5	**71.43**	50	96.5	66.67	**60**	93	35.71	50	95	50	50
โฆษณากลุ่มเครื่องใช้ไฟฟ้า												
Attention	97.81	0	0	96.17	0	0	82.51	3.33	25	74	**12.24**	**40**
Interest	71.04	13.33	12.9	68.85	6.67	6.45	68.31	17.07	22.58	72.68	**22.68**	**25.81**
Search	96.17	0	0	62.84	3.08	28.57	72.68	**6.12**	**42.86**	84.15	0	0
Action	98.91	0	0	95.63	0	0	82.51	**3.13**	**50**	94.54	0	0
Share	84.7	0	0	90.16	0	0	79.23	**10**	**21.43**	82.51	0	0
โฆษณากลุ่มสุขภัณฑ์												
Attention	92.86	0	0	87.36	0	0	81.32	**8**	**15.38**	91.21	0	0
Interest	69.23	21.21	18.92	69.23	**26.83**	**29.73**	58.79	17.24	27.03	71.98	25	18.92
Search	81.32	4.17	8.33	86.26	**6.67**	8.33	74.18	5.13	**16.67**	80.22	3.85	8.33
Action	98.9	0	0	68.13	0	0	75.82	**4.35**	**100**	98.9	0	0
Share	87.91	9.1	7.69	88.46	0	0	89.56	**20**	**15.38**	83.52	0	0

ตารางที่ 38 แสดงผลลัพธ์ในแง่ของค่าเอ็มซีซีของตัวแบบจำแนกประเภทแอลเอสทีเอ็ม เปรียบเทียบระหว่างชุดข้อมูลที่ไม่สมดุลกับชุดข้อมูลที่สมดุลด้วยเทคนิคการสร้างข้อความแต่ละวิธี

	ชุดข้อมูลที่ไม่สมดุล		ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคการสร้างข้อความ	
ตัวแบบ	-	การสุ่มเพิ่มตัวอย่างแบบoversampling	มาร์คอฟเชน	LSTM
โฆษณากลุ่มเครื่องสำอาง				
Attention	0	-0.0328	**0.0304**	-0.0144
Interest	0.1238	0.1696	**0.1987**	0.0994
Search	0.1529	0.1010	**0.1816**	0.0565
Action	0	-0.0276	**0.0734**	-0.0218
Share	**0.6321**	0.5805	0.3702	0.5789
โฆษณากลุ่มเครื่องใช้ไฟฟ้า				
Attention	**0**	-0.0553	-0.0378	-0.0251
Interest	-0.0705	-0.0833	**0.0592**	-0.0646
Search	0	0.0132	**0.0158**	-0.0737
Action	0	-0.0157	**0.1988**	-0.0210
Share	-0.0891	-0.0615	**0.0769**	0.0006
โฆษณากลุ่มสุขภัณฑ์				
Attention	0	0.0280	-0.0334	**0.0446**
Interest	-0.0452	-0.0181	**0.1641**	0.0576
Search	0.0928	0.0541	**0.1099**	-0.0139
Action	0	-0.0254	**0.1812**	-0.0078
Share	**0.1335**	-0.0769	0.0510	-0.0703

บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

หลังจากการเตรียมข้อมูลซึ่งเป็นข้อความโฆษณาภาษาไทยจากเฟซบุ๊ก โดยใช้การตัดคำ การแทนเอกสารด้วยเมทริกซ์คำ-เอกสาร และใช้เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสาร การนำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ประยุกต์ใช้กับชุดข้อมูลดังกล่าวสามารถทำให้ตัวแบบจำแนกประเภทนาอีฟเบย์ การถดถอยโลจิสติกส์ และซัพพอร์ตเวกเตอร์แมชชีน มีประสิทธิภาพเพิ่มขึ้นในแง่ของค่าความแม่นและค่าระลึก โดยสังเกตจากการเพิ่มขึ้นของค่าดังกล่าวในเกือบทุกชุดข้อมูล นอกจากนี้ เมื่อนำเทคนิคการคัดเลือกคุณลักษณะมาใช้ร่วมด้วยเพื่อหาจำนวนคุณลักษณะที่เหมาะสมของแต่ละชุดข้อมูลและตัดคุณลักษณะที่ไม่เกี่ยวข้องออกไป จะพบว่าเมื่อจำนวนคุณลักษณะมีความเหมาะสม สามารถทำให้ค่าความแม่นและค่าระลึกเพิ่มขึ้นได้อีก

ต่อมาผู้วิจัยได้เสนอเทคนิคการเพิ่มคุณลักษณะใหม่ซึ่งเป็นคำคล้ายคลึง โดยตั้งสมมติฐานว่าคุณลักษณะใหม่จะประกอบด้วยคุณลักษณะซึ่งเป็นคำที่สามารถทำให้ตัวแบบพบแพทเทิร์นของแต่ละกลุ่มข้อมูลมากยิ่งขึ้น หลังจากประยุกต์ใช้เทคนิคดังกล่าวพบว่าเทคนิคดังกล่าวอาจสร้างคำทั้งที่เกี่ยวข้องและไม่เกี่ยวข้องออกมา ทำให้ในบางชุดข้อมูล เทคนิคดังกล่าวสามารถเพิ่มประสิทธิภาพในแง่ของค่าระลึกได้ แต่ก็ทำให้ค่าดังกล่าวในบางชุดข้อมูลมีค่าต่ำลงเนื่องจากคำใหม่ที่เพิ่มเข้าไปไม่ได้มีคำที่เป็นประโยชน์ต่อการจำแนกกลุ่มข้อมูลอยู่ด้วย สำหรับค่าความแม่นพบว่าในตัวแบบส่วนใหญ่ไม่ได้มีความแตกต่างไปจากเดิม

ในการทดลองสุดท้าย ผู้วิจัยเปลี่ยนแปลงวิธีการเตรียมข้อมูล โดยตัดขั้นตอนการแทนเอกสารด้วยเมทริกซ์คำ-เอกสาร และใช้เทคนิคความถี่ของคำ-ส่วนกลับความถี่ของเอกสารออกไป และใช้เทคนิคใหม่ในการทำให้ชุดข้อมูลสมดุล นั่นคือเทคนิคการสร้างข้อความ ก่อนจะแปลงคำเป็นเวกเตอร์แล้วป้อนข้อมูลเข้าสู่ตัวแบบจำแนกประเภทแอลเอสทีเอ็ม โดยเทคนิคการสร้างข้อความที่นำมาใช้ประกอบไปด้วยการสร้างข้อความแบบมาร์คอฟเชนและการสร้างข้อความด้วยแอลเอสทีเอ็ม โดยผลลัพธ์พบว่าการใช้เทคนิคการสร้างข้อความแบบมาร์คอฟเชนสามารถเพิ่มประสิทธิภาพของตัวแบบได้ดีที่สุดในแง่ของค่าความแม่นและค่าระลึก และจากการสังเกตยังพบว่าในชุดข้อมูลส่วนใหญ่ การใช้เทคนิคการสร้างข้อความแบบมาร์คอฟเชนเพื่อทำให้ข้อมูลสมดุลก่อนใช้ตัวแบบแอลเอสทีเอ็มจำแนกประเภทแอลเอสทีเอ็มทำให้ตัวแบบได้รับค่าระลึกสูงแต่ค่าความแม่นต่ำ

โดยสรุป การใช้เทคนิคแต่ละรูปแบบที่เสนอมาสามารถทำให้ตัวแบบทำนายข้อมูลเป็นกลุ่มที่เป็นคลาสบวกซึ่งเป็นคลาสกลุ่มน้อยมากขึ้นได้ อย่างไรก็ตาม เนื่องจากทั้งในชุดข้อมูลสอนและข้อมูลทดสอบประกอบด้วยข้อมูลที่เป็นคลาสบวกจำนวนน้อยมาก เมื่อตัวแบบทำนายข้อมูลเป็นกลุ่มที่เป็นคลาสบวกเป็นจำนวนครั้งมากขึ้นก็ส่งผลให้ค่าความถูกต้องต่ำลงเช่นกัน

5.2 อภิปรายผลการทดลอง

การประยุกต์ใช้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยในบริบทของการจำแนกข้อความภาษาไทย ทำให้ตัวแบบต่าง ๆ สามารถทำนายตัวอย่างที่เป็นคลาสบวกได้มากขึ้น ซึ่งสอดคล้องกับหลาย ๆ งานวิจัยก่อนหน้า อีกทั้งยังเห็นได้ว่าเทคนิคดังกล่าวสามารถทำงานได้ดีแม้ข้อมูลจะมีคุณลักษณะจำนวนมากเนื่องจากเป็นข้อความ

การใช้เทคนิคการคัดเลือกคุณลักษณะเพื่อกำจัดคุณลักษณะที่ไม่เกี่ยวข้องออกไป สามารถทำให้เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมีประสิทธิภาพดีขึ้นไปอีก อย่างไรก็ตามแต่ละชุดข้อมูลอาจมีจำนวนคุณลักษณะที่เหมาะสมไม่เท่ากัน จึงต้องมีการทดลองหาจำนวนคุณลักษณะที่เหมาะสมแตกต่างกันไปตามแต่ละชุดข้อมูล

เทคนิคการเพิ่มคุณลักษณะใหม่ที่ได้นำเสนอ จากการทดสอบทีจะเห็นว่าสามารถเพิ่มประสิทธิภาพตัวแบบได้ในบางชุดข้อมูล เนื่องจากคุณลักษณะที่เพิ่มเข้าไปประกอบด้วยคุณลักษณะที่เกี่ยวข้องกับคลาสตามสมมติฐาน แต่ในบางชุดข้อมูล เทคนิคนี้ไม่สามารถเพิ่มคุณลักษณะที่เกี่ยวข้องเข้าไปได้ จึงจะเห็นได้ว่าประสิทธิภาพของตัวแบบไม่แตกต่างจากเดิมหรือแย่ลง เทคนิคนี้จึงอาจเสนอให้เป็นทางเลือกหนึ่งในการจัดการกับคุณลักษณะของข้อความเพื่อเพิ่มประสิทธิภาพของตัวแบบในบางชุดข้อมูลได้

ในการทดลองสุดท้าย เทคนิคการสร้างข้อความแบบมาร์คอฟเชนสามารถเพิ่มประสิทธิภาพของชุดข้อมูลซึ่งเป็นข้อความที่ไม่สมดุลได้เช่นกัน โดยไม่จำเป็นต้องแปลงชุดข้อความให้อยู่ในรูปแบบของเมทริกซ์คำ-เอกสาร ข้อมูลดังกล่าวอาจถูกแทนให้อยู่ในรูปเวกเตอร์แล้วนำไปใช้กับตัวแบบที่ทำการคัดเลือกคุณลักษณะแบบอัตโนมัติอย่างแอลเอสทีเอ็มได้ อย่างไรก็ตาม เมื่อพิจารณาในหลายชุดข้อมูลจะพบว่าการเพิ่มข้อความที่เป็นคลาสบวกโดยการใช้เทคนิคการสร้างข้อความแบบมาร์คอฟเชนทำให้ตัวแบบแอลเอสทีเอ็มได้รับค่าระลึกสูงแต่ค่าความแม่นค่อนข้างต่ำ

5.3 ปัญหาและอุปสรรคในการดำเนินงาน

1. เนื่องจากข้อมูลที่รวบรวมมา เมื่อผ่านการแปะฉลากคลาสตัวอย่างข้อมูลแต่ละตัวอย่างแล้วพบว่าข้อมูลที่เป็นตัวอย่างบวกซึ่งเป็นโฆษณาที่ก่อให้เกิดสถานะต่าง ๆ มีจำนวนน้อยมาก

ส่งผลให้การค้นพบแพทเทิร์นของโฆษณาที่ดีเหล่านี้จึงน้อยเช่นกัน ทำให้ตัวแบบสำหรับทำนายคลาสของโฆษณาเหล่านี้มีประสิทธิภาพต่ำกว่าที่ควรจะเป็น จึงต้องใช้เทคนิคต่าง ๆ เข้ามาช่วย

2. การประมวลผลข้อมูลที่เป็นภาษาไทยทำได้ค่อนข้างยาก ดังเช่น การตัดคำ ต้องใช้วิธีการต่าง ๆ เข้ามาช่วยเนื่องจากภาษาไทยไม่ได้เว้นวรรคคำแบบภาษาอังกฤษ ซึ่งการตัดคำก็มักจะทำได้ไม่สมบูรณ์แบบเช่นกัน คำที่ตัดผิดพลาดจะกลายเป็นคุณลักษณะที่ไม่ดีหรือเป็นข้อมูลรบกวนทำให้ตัวแบบจำแนกประเภทมีประสิทธิภาพต่ำลง อีกทั้งหลายคำในข้อมูลโฆษณาที่รวบรวมมาเป็นคำที่ไม่เป็นทางการ ซึ่งบางครั้งยังสามารถเขียนได้หลายรูปแบบ เช่นคำว่า เลิศ เริศ เริ่ด คำเหล่านี้เมื่อกลายเป็นคุณลักษณะจะถือว่าเป็นคนละคุณลักษณะกันทั้ง ๆ ที่มีความหมายเหมือนกัน ซึ่งเมื่อนำไปเป็นข้อมูลสอนจะทำให้ตัวแบบมีประสิทธิภาพต่ำลงไปอีก
3. บางคำในโฆษณากลุ่มเดียวกันหรือต่างกลุ่มกันอาจจะสื่อความหมายต่างกันหรือลักษณะที่ต่างกันได้ เช่น คำว่า ใส ในโฆษณากลุ่มเครื่องสำอางอาจหมายถึงลักษณะของผิว แต่สำหรับโฆษณากลุ่มสุขภัณฑ์อาจจะหมายถึงลักษณะของกระเบื้อง ในงานวิจัยนี้จึงได้แยกข้อมูลเพื่อสอนตัวแบบของโฆษณาแต่ละประเภทเพื่อลดความผิดพลาดเมื่อคำถูกนำไปเป็นคุณลักษณะ
4. เนื่องจากงานวิจัยที่ศึกษาเกี่ยวกับการประมวลผลข้อความภาษาไทยยังมีจำนวนไม่มากนัก การค้นหาเทคนิคต่าง ๆ ที่ช่วยเพิ่มประสิทธิภาพการประมวลผลข้อความภาษาไทยจึงทำได้ยากขึ้นและไม่หลากหลายเท่าภาษาอังกฤษ

5.4 ข้อเสนอแนะ

1. การรวบรวมข้อมูลโฆษณาให้มีจำนวนมากขึ้นและมีความหลากหลายจะเป็นวิธีที่ดีในการค้นพบแพทเทิร์นของโฆษณาที่ดี ทำให้ตัวแบบจำแนกประเภทได้มีประสิทธิภาพขึ้น
2. มีตัวแปรหลายอย่างซึ่งเป็นพารามิเตอร์ที่สามารถปรับค่าได้เพื่อให้สามารถค้นพบค่าที่ทำให้ตัวแบบมีประสิทธิภาพมากขึ้น เช่น
    a. วิธีการที่ใช้ตัดคำภาษาไทย ซึ่งนอกจากไลบรารีคัดคำซึ่งเป็นวิธีการแบบเครือข่ายประสาทที่ได้ใช้ในงานนี้แล้ว ยังมีวิธีการอื่น ๆ อีกมายมาก ดังเช่น วิธีการเทียบคำที่สั้นที่สุด (shortest word pattern matching) วิธีการเทียบคำที่ยาวที่สุด (longest word pattern matching) วิธีการตัดคำโดยใช้วิธีการทางสถิติ เช่น เอ็นแกรม (n-gram) เป็นต้น

b. การแทนคุณลักษณะด้วยคำแบบมากกว่ายูนิแกรม เช่น ไบแกรม (bi-gram) และไตรแกรม (tri-gram)
c. ใช้วิธีการแก้ปัญหาชุดข้อมูลที่ไม่สมดุลวิธีอื่นนอกเหนือจาก SMOTE
d. ใช้วิธีการคัดเลือกคุณลักษณะแบบอื่นนอกเหนือจากวิธีไคกำลังสอง
e. ใช้วิธีการสร้างข้อความแบบอื่นนอกเหนือจากการสร้างข้อความแบบมาร์คอฟเชนและการสร้างข้อความด้วยแอลเอสทีเอ็ม
f. ใช้วิธีการสำหรับจำแนกประเภทแบบอื่นนอกเหนือจากที่ใช้ในงานวิจัยนี้
g. ในขั้นตอนการหาคำคล้ายคลึงที่เกิดร่วมกัน ผู้วิจัยได้ค้นหาคำคล้ายคลึง k อันดับแรกของแต่ละคำ ซึ่งในการทดลองได้เลือกค่า k เท่ากับ 5 เท่านั้น แต่การเลือกค่า k ซึ่งเป็นค่าอื่นนอกเหนือจากนี้มีผลทำให้ได้คำใหม่ซึ่งเป็นคุณลักษณะที่มีคุณภาพไม่เท่ากัน การทดลองเปลี่ยนค่าดังกล่าวให้หลากหลายมากขึ้นน่าจะทำให้พบค่า k ที่เหมาะสมมากกว่านี้

3. เพื่อเพิ่มประสิทธิภาพให้กับตัวแบบสำหรับจำแนกประเภท ควรหาวิธีการทำให้ข้อมูลซึ่งเป็นข้อความภาษาไทยสะอาดที่สุดก่อนที่จะใช้เป็นข้อมูลเรียนรู้สำหรับตัวแบบ เช่น การแก้การสะกดคำผิด การทำให้คำที่มีความหมายเหมือนกันอยู่ในรูปแบบเดียวกันก่อน เป็นต้น อย่างไรก็ตาม หากสามารถแยกคำที่มีหน้าตาเหมือนกันแต่มีความหมายไม่เหมือนกันหรือสื่อถึงลักษณะที่ต่างกัน (เช่นคำว่า ใส จากประโยค ผิวหน้ากระจ่างใส กระเบื้องใส ฟ้าใส) แยกออกเป็นคนละคุณลักษณะกัน น่าจะทำให้ตัวแบบมีประสิทธิภาพมากขึ้น

บรรณานุกรม

[1] S. Leesa-nguansuk. Thailand makes top 10 in social media use. Bangkok Post. [Online]. Available: https://www.bangkokpost.com/

[2] L. Amaly and H. Hudrasyah. 2012. Measuring effectiveness of marketing communication using AISAS ARCAS model. Journal of Business and Management, 1(5), pp. 352-364.

[3] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer. (Jun 2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. [Online]. pp. 321-357. Available: https://arxiv.org/pdf/1106.1813.pdf

[4] Zheng, Zhuoyuan, Yunpeng Cai, and Ye Li. "Oversampling method for imbalanced classification." *Computing and Informatics* 34.5 (2016): 1017-1037.

[5] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze*, Introduction to Information Retrieval,* Cambridge University Press. 2008.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. In Proceedings of Workshop at ICLR, 2013.

[7] T. Mitchell. *Machine Learning,* McGraw-Hill, New York (1997): 177-182.

[8] Amorninit, U. (n.d.). *Logistic Regression Analysis*. Retrieved from http://www.utcc.ac.th/public_content/files/001/P268_1.pdf

[9] Ng, A. (n.d.). *Logistic Regression*. Retrieved from http://www.holehouse.org/mlclass/06_Logistic_Regression.html

[10] A. Ng. *Support Vector Machines*. Retrieved from http://cs229.stanford.edu/notes/cs229-notes3.pdf

[11] Charles M. Grinstead and J. Laurie Snell, *Introduction to Probability,* American Mathematical Society. 1997.

[12] Y. Suh, J. Yu, J. Mo, L. Song and C.Kim. Jan 2017. A Comparison of Oversampling Methods on Imbalanced Topic Classification of Korean News Articles. Journal of Cognitive Science. [Online]. 18(4). pp.391-437. Available:http://cogsci.snu.ac.kr/jcs/issue/vol18/no4/suh.pdf

[13] P. Sarakit, T. Theeramunkong and C. Haruechaiyasak. 2015. "Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm," In the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Chonburi.

[14] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh and D. Khalili. Jan 2016. "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes," Medical decision making, pp. 137-144.

[15] A. Mohasseb, M. Bader-El-Den, M. Cocea, H. Liu. 2018. "Improving imbalanced question classification using structured smote based approach," In the 2018 International Conference on Machine Learning and Cybernetics.

[16] A. S. Manek, P. D. Shenoy, M. C. Mohan and K. R. Venugopal. 2017 "Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier," World Wide Web. 20(2), pp. 135-154.

[17] Y. Liu, J. W. Bi, Z. P. Fan. Sep 2017. "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms," Expert Systems with Applications, pp. 323-329.

[18] A. Onan, S. Korukoğlu. Feb 2017. A feature selection model based on genetic rank aggregation for text sentiment classification. Journal of Information

Science. [Online]. 43(1). pp.25-38. Available: http://journals.sagepub.com/doi/ pdf/10.1177/0165551515613226

[19] Kathleen R. McKeown. Text generation. Cambridge University Press, 1992 Jun 26.

[20] Y. Wu, M. Schuster, Z. Chen, Q.V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144. 2016 Sep 26.

[21] T.H. Wen, M. Gasic, N. Mrksic, P.H. Su, D. Vandyke, S. Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. arXiv preprint arXiv:1508.01745. 2015 Aug 7.

[22] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. InInternational conference on machine learning 2015 Jun 1 (pp. 2048-2057).

[23] G. Barbieri, F. Pachet, P. Roy, M. Degli Esposti. Markov Constraints for Generating Lyrics with Style. InEcai 2012 Aug 15 (Vol. 242, pp. 115-120).

[24] P. Potash, A. Romanov, A. Rumshisky. Ghostwriter: Using an LSTM for automatic rap lyric generation. InProceedings of the 2015 Conference on Empirical Methods in Natural Language Processing 2015 (pp. 1919-1924).

[25] W. Nie, N. Narodytska, A. Patel. 2018. RelGAN: Relational Generative Adversarial Networks for Text Generation.

[26] Y. Zhang, Z. Gan, K. Fan, Z. Chen, R. Henao, D. Shen, L. Carin. Adversarial feature matching for text generation. InProceedings of the 34th International Conference on Machine Learning-Volume 70 2017 Aug 6 (pp. 4006-4015). JMLR. org.

[27] P. Treeratpituk. *Cutkum: Thai Word-Segmentation with LSTM in Tensorflow*. Retrieved from https://github.com/pucktada/cutkum

ประวัติผู้เขียน

ชื่อ-สกุล	ศุภมงคล อัครดำรงค์รัตน์
วัน เดือน ปี เกิด	12 สิงหาคม 2536
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	วท.บ. วิทยาการคอมพิวเตอร์ มหาวิทยาลัยธรรมศาสตร์
ที่อยู่ปัจจุบัน	142/13 สตูดิโอโซน 102 ลาดพร้าว พลับพลา วังทองหลาง กรุงเทพมหานคร 10310
ผลงานตีพิมพ์	S. Akkaradamrongrat, P. Kachamas and S. Sinthupinyo. 2018. Classification of Advertisement Text on Facebook Using Synthetic Minority Over-Sampling Technique. In Proceedings of 2018 International Conference on Algorithms, Computing and Artificial Intelligence (ACAI'18). Sanya, China, 6 pages. https://doi.org/10.1145/3302425.3302471 S. Akkaradamrongrat, P. Kachamas and S. Sinthupinyo. 2019. "Text Generation for Imbalanced Text Classification," In the 16th International Joint Conference on Computer Science and Software Engineering (JCSSE2019), Chonburi, Thailand