Thai Journal of Pharmacy Practice

บทความวิจัย

การใช้การเรียนรู้ของเครื่องสำหรับตรวจจับข้อความโฆษณาอาหารที่ผิดกฎหมาย

วรรณกัญญ์ นิธิโรจน์ศุภภัค[1], วีรยุทธ์ เลิศนที[2]

[1]ศูนย์จัดการเรื่องร้องเรียนและปราบปรามฯ สำนักงานคณะกรรมการอาหารและยา
[2]ภาควิชาสารสนเทศศาสตร์ทางสุขภาพ คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร

บทคัดย่อ

วัตถุประสงค์: เพื่อหาแบบจำลองที่เหมาะสมจากเทคนิคการเรียนรู้ของเครื่องสำหรับจำแนกข้อความโฆษณาอาหารเป็นข้อความที่ถูกกฎหมายและผิดกฎหมาย **วิธีการ:** ผู้วิจัยเตรียมชุดข้อความโฆษณาอาหาร จำนวน 200 ตัวอย่าง แบ่งเป็นข้อความถูกกฎหมาย 100 ตัวอย่างและผิดกฎหมาย 100 ตัวอย่าง ในขั้นตอนการเตรียมข้อมูล ข้อมูลที่ไม่เกี่ยวข้องที่สามารถเชื่อมโยงไปยังเจ้าของผลิตภัณฑ์ เช่น เลขที่ใบอนุญาตโฆษณา ชื่อการค้า และชื่อบริษัท ถูกลบออกไป หลังจากนั้นตัดคำภาษาไทยด้วยอัลกอริทึมที่เลือกคำยาวที่สุด เพื่อใช้แบ่งคำในประโยค/วลี ขั้นต่อไป นำรายการคำหยุดภาษาไทยมาใช้เพื่อลบคำที่ไม่สำคัญออก จากนั้นใช้ชุดคำแบบยูนิแกรม และแบบไบแกรมมาจัดทำคุณลักษณะในเวกเตอร์เอกสาร คุณลักษณะทั้งหมดและบางส่วนถูกนำมาใช้สร้างและทดสอบแบบจำลอง คุณลักษณะบางส่วนถูกเลือกโดยวิธีเคที่ดีที่สุด โปรแกรมภาษา PHP และชุดไลบรารี PHP-ML สำหรับการเรียนรู้ด้วยเครื่องถูกใช้เพื่อสร้างชุดโปรแกรม เทคนิคการเรียนรู้แบบมีผู้สอน 3 ชนิดถูกนำมาใช้ในการจัดทำแบบจำลอง ได้แก่ ซัพพอร์ทเวกเตอร์แมชชีน เคเนียเรสเนเบอร์ และนาอีฟเบย์ส ทำโดยใช้การสุ่มตัวอย่างแบบแบ่งชั้นร้อยละ 80 ของข้อมูลด้วยสัดส่วนที่เท่ากันของกลุ่มข้อความที่ถูกและผิดกฎหมายเพื่อนำมาใช้สร้างแบบจำลอง และร้อยละ 20 ที่เหลือใช้ทดสอบแบบจำลอง แต่ละการทดสอบทำ 10 ครั้ง ใช้ค่าเฉลี่ยของคะแนน F1 ในการบอกประสิทธิภาพของแบบจำลอง จากนั้นนำแบบจำลองที่มีคะแนน F1 เฉลี่ยมากที่สุดของแต่ละเทคนิคของการเรียนรู้ มาสร้างโปรแกรมตรวจจับข้อความโฆษณาที่ผิดกฎหมาย และทดสอบด้วยข้อความโฆษณา 40 ข้อความ **ผลการวิจัย:** ซัพพอร์ทเวกเตอร์แมชชีนเป็นตัวจำแนกข้อความโฆษณาอาหารที่มีประสิทธิภาพมากที่สุด ด้วยคะแนน F1 คือ 0.987 เมื่อใช้คุณลักษณะทั้งหมดแบบยูนิแกรม หลังตัดคำหยุดออก **สรุป:** เทคนิคเรียนรู้ของเครื่องสามารถใช้สำหรับจำแนกข้อความโฆษณาอาหารที่ถูกหรือผิดกฎหมายได้อย่างมีประสิทธิภาพ

คำสำคัญ: ข้อความโฆษณาอาหาร การเรียนรู้ของเครื่อง กฎหมาย การจัดประเภทเอกสาร

รับต้นฉบับ: 19 ก.พ. 2563, **ได้รับบทความฉบับปรับปรุง**: 1 เม.ย. 2563, **รับลงตีพิมพ์**: 7 เม.ย. 2563

ผู้ประสานงานบทความ: วรรณกัญญ์ นิธิโรจน์ศุภภัค ศูนย์จัดการเรื่องร้องเรียนและปราบปรามฯ สำนักงานคณะกรรมการอาหารและยา อำเภอเมือง จังหวัดนนทบุรี 11000 **E-mail**: nitirotsuphapha_w@su.ac.th

PSU PRINCE OF SONGKLA UNIVERSITY
Thai Journal of Pharmacy Practice

Research article

Using Machine Learning for Detection of Illegal Food Advertising Text

Wannakan Nitirotsuphaphak[1], Verayuth Lertnattee[2]

[1]Complaint and Enforcement Management Center, Food and Drug Administration
[2]Department of Health-related Informatics, Faculty of Pharmacy, Silpakorn University

Abstract

Objective: To find the appropriate model from machine learning techniques for classifying food advertising texts to legal and illegal texts. **Methods:** A set of 200 food advertising texts was prepared with 100 legal texts and 100 illegal texts. In preprocessing steps, irrelevant information that could be linked to product owners, such as advertising license numbers, trade names, and company names was removed from original texts. Then, the Thai word segmentation with the longest matching algorithm was used to separate words in sentences/phrases. In the next step, a list of Thai stopwords was applied to remove unimportant words. Then, unigram words and bigram words were used as features in document vectors. The full set and subset of features were utilized for creating and testing models. The subset of features was selected using select k best method. The PHP language with the PHP-ML library for machine learning was used to construct a set of programs. Three techniques of supervised learning were applied to create the models, i.e., support vector machine, k-nearest neighbors, and naïve Bayes. By using the stratified random technique, 80% of the collection with equal portions of legal and illegal texts was used for creating models and the rest of 20% was used for testing models. Each test was performed 10 times. The average score of F1 was used as a performance indicator. Then, models that obtained the highest average F1 for each learning technique were used to create a web application for detecting illegal food advertising text. The performance of each model was tested by 40 food advertising texts. **Results:** The support vector machine is the most effective classifier for categorizing food advertising text with the highest F1-score of 0.987 when the model was created with full features of unigrams after removing stop words. **Conclusion:** Machine learning techniques could be efficiently applied for classifying legal/illegal food advertising texts.

Keywords: food advertising text, machine learning, law, text classification

บทนำ

ตามพระราชบัญญัติอาหาร พ.ศ.2522 อาหาร หมายความว่า "ของกินหรือเครื่องค้ำจุนชีวิต ได้แก่ 1) วัตถุทุกชนิดที่คนกิน ดื่ม อม หรือนำเข้าสู่ร่างกายไม่ว่าด้วยวิธีใด ๆ หรือในรูปลักษณะใด ๆ แต่ไม่รวมถึงยา วัตถุออกฤทธิ์ต่อจิตและประสาท หรือยาเสพติดให้โทษตามกฎหมายว่าด้วยการนั้น แล้วแต่กรณี 2) วัตถุที่มุ่งหมายสำหรับใช้หรือใช้เป็นส่วนผสมในการผลิตอาหาร รวมถึงวัตถุเจือปนอาหาร สี และเครื่องปรุงรส" (1) สำหรับการโฆษณาอาหารนั้น ต้องได้รับอนุญาตจากสำนักงานคณะกรรมการอาหารและยา (อย.) จึงสามารถโฆษณาได้ การโฆษณาอาหารต้องใช้ข้อความที่เป็นไปตามประกาศสำนักงานคณะกรรมการอาหารและยา เรื่อง หลักเกณฑ์การโฆษณาอาหาร (2)

อุตสาหกรรมอาหารและเครื่องดื่มในสหรัฐอเมริกามองว่า เด็กและวัยรุ่นเป็นตลาดหลัก เนื่องจากมีอำนาจในการใช้จ่าย อิทธิพลต่อการซื้อ และจะเติบโตเป็นผู้บริโภคในอนาคต จึงมีการใช้เทคนิคและช่องทางที่หลากหลาย เพื่อเข้าถึงผู้บริโภคตั้งแต่วัยเด็กเพื่อสร้างความผูกพันต่อสินค้าและอิทธิพลต่อพฤติกรรมการซื้ออาหาร เช่น การโฆษณาทางโทรทัศน์ (3) ในประเทศออสเตรเลียได้วิเคราะห์เนื้อหาโฆษณาอาหารทางโทรทัศน์ พบว่า ร้อยละ 31 ของการโฆษณาเป็นอาหาร และร้อยละ 81 ของการโฆษณาอาหารเป็นอาหารที่ไม่ดีต่อสุขภาพหรือไม่ใช่อาหารหลัก (4) งานวิจัยที่ศึกษาผลกระทบของการโฆษณาอาหารทางโทรทัศน์ต่อพฤติกรรมการกินพบว่า เด็กมีการบริโภคเพิ่มมากขึ้นร้อยละ 45 เมื่อได้รับโฆษณาอาหาร เห็นได้ว่า การโฆษณาอาหารเพิ่มการบริโภคผลิตภัณฑ์ และผลกระทบเหล่านี้ไม่เกี่ยวข้องกับความหิวหรืออิทธิพลอื่น ๆ (5)

การเรียนรู้ของเครื่อง คือ การนำเข้าข้อมูลไปยังคอมพิวเตอร์ โดยคอมพิวเตอร์มีกระบวนการเปลี่ยนข้อมูลเหล่านั้นให้กลายเป็นองค์ความรู้ การเรียนรู้ของเครื่องสามารถนำมาใช้แก้ไขปัญหาใน 2 ประเภท ประเภทแรก คือ งานที่มีความซับซ้อนที่ไม่สามารถเขียนเป็นโปรแกรมออกมาได้ ประเภทที่สอง คือ งานมีการเปลี่ยนแปลงอยู่ตลอด ทั้งด้วยปัจจัยด้านเวลาและบุคลากร ทำให้งานอยู่ในลักษณะที่ไม่อยู่นิ่ง (6) การจำแนกประเภทข้อความ (text classification หรือ text categorization) มีพื้นฐานมาจากการเรียนรู้ของเครื่อง ในปัจจุบันเนื่องจากข้อมูลข่าวสารในเว็บไซต์มีแนวโน้มเพิ่มขึ้นเรื่อย ๆ การจำแนกประเภทข้อความจึงกลายเป็นวิธีการจัดการข้อมูลที่สำคัญ จากการทบทวนวรรณกรรม พบว่า จากงานวิจัย 132 ฉบับ มี 113 ฉบับ ใช้เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกข้อความ นอกจากนี้ยังพบว่า ผู้วิจัยนิยมใช้เทคนิคการเรียนรู้ ซัพพอร์ทเวกเตอร์แมชชีน (support vector machine: SVM) มากที่สุด รองลงมาคือ เคเนียเรสเนเบอร์ (k-nearest neighbors: k-NN) และนาอีฟเบย์ส (naïve Bayes: NB) ตามลำดับ (7) ในประเทศไทยมีงานวิจัย การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ โดยจัดหมวดหมู่ออกเป็น 10 กลุ่ม มีจำนวนตัวอย่างทั้งหมด 12,000 เอกสาร โดยใช้เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกข้อความ 6 เทคนิค พบว่า SVM ให้ประสิทธิภาพดีที่สุดเมื่อวัดจากค่าเฉลี่ยคะแนน F1 ได้ร้อยละ 91.3 (8) งานวิจัยเกี่ยวกับแบบจำลองการจัดหมวดหมู่สถานที่ท่องเที่ยวโดยใช้เทคนิคการเรียนรู้ของเครื่อง ได้พัฒนาแบบจำลองจัดหมวดหมู่สถานที่ท่องเที่ยว ซึ่งแบ่งออกเป็น 5 หมวดหมู่ โดยใช้ 4 เทคนิค เมื่อเปรียบเทียบเทคนิคการเรียนรู้ของเครื่อง พบว่า NB มีประสิทธิภาพมากที่สุด (9) งานวิจัยเกี่ยวกับการจำแนกกลุ่มคำถามอัตโนมัติบนกระดานสนทนา ได้เปรียบเทียบเทคนิคการเรียนรู้ของเครื่อง 3 เทคนิค พบว่า k-NN ให้ประสิทธิภาพมากที่สุด โดยคะแนน F1 เท่ากับ 0.892 (10)

ปัจจุบันปริมาณข้อความโฆษณาอาหารที่เผยแพร่ตามช่องทางต่าง ๆ มีจำนวนมาก หากข้อความโฆษณามีการบิดเบือน อาจทำให้เกิดผลกระทบต่อผู้บริโภค สถิติของมูลนิธิเพื่อผู้บริโภคในปี 2561 ได้ให้คำปรึกษาและรับเรื่องร้องเรียน จำนวน 4,545 เรื่อง เรื่องที่มีผู้ร้องเรียนมากที่สุด คือ เรื่องอาหาร ยา และผลิตภัณฑ์สุขภาพ คิดเป็นร้อยละ 33.11 ปัญหาอันดับหนึ่งของหมวดนี้ คือ เรื่องโฆษณาอันเป็นเท็จหรือหลอกลวง โดยมักเป็นโฆษณาผลิตภัณฑ์สุขภาพที่อาจก่อให้เกิดอันตรายและอาหารเสริม (11) จากข้อมูลดังกล่าวข้างต้น แสดงให้เห็นว่า ในปัจจุบันยังมีข้อความโฆษณาอาหารที่ทำให้เกิดความเสียหายต่อผู้บริโภค และจากการทบทวนวรรณกรรมยังไม่พบงานวิจัยที่นำการเรียนรู้ของคอมพิวเตอร์มาประยุกต์ใช้กับงานตรวจจับข้อความโฆษณา ดังนั้น เพื่อช่วยให้เจ้าหน้าที่ของรัฐสามารถตรวจจับข้อความโฆษณาที่ผิดกฎหมายจากข้อความโฆษณาที่มีเป็นจำนวนมาก ผู้ประกอบการสามารถตรวจสอบข้อความโฆษณาที่ถูกต้องได้ และป้องกันผู้บริโภคไม่ให้หลงเชื่อข้อความโฆษณาที่เข้าข่ายผิดกฎหมาย งานวิจัยนี้จึงนำการเรียนรู้ด้วยคอมพิวเตอร์เข้ามาประยุกต์เพื่อตรวจจับข้อความโฆษณาอาหารที่เข้าข่ายผิดกฎหมาย

วิธีการวิจัย

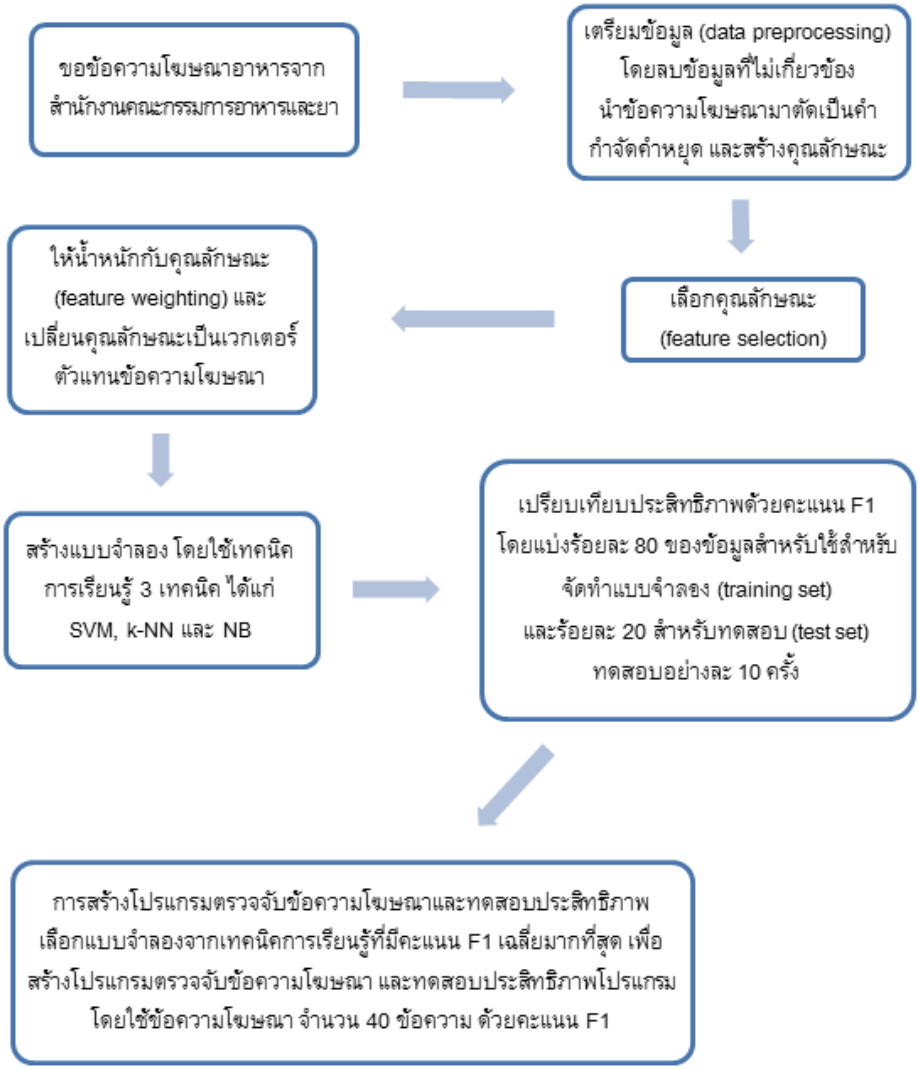
การวิจัยนี้เป็นการวิจัยเชิงทดลอง ซึ่งได้ผ่านการพิจารณาจากคณะกรรมการจริยธรรมวิจัยในมนุษย์ มหาวิทยาลัยศิลปากรก่อนเริ่มการวิจัย รูปที่ 1 สรุปขั้นตอนการวิจัย ซึ่งมีรายละเอียดดังนี้

ตัวอย่าง

ตัวอย่าง คือ ข้อความโฆษณาอาหารที่ได้รับการพิจารณาจากสำนักงานคณะกรรมการอาหารและยาในปี 2560 – 2561แล้วว่า เป็นข้อความโฆษณาที่ถูกกฎหมายหรือผิดกฎหมายอย่างละ 100 ข้อความ รวมทั้งหมด 200 ข้อความ ในงานวิจัยนี้ ข้อความโฆษณาอาหาร หมายถึง ข้อความที่การกระทำด้วยวิธีการใด ๆ ให้ประชาชนเห็นหรือทราบ ข้อความเกี่ยวกับอาหาร ส่วนประกอบของอาหาร เพื่อประโยชน์ในทางการค้า โดยเป็นข้อความได้จากโทรทัศน์ สื่อต่าง ๆ บนอินเทอร์เน็ต เช่น เว็บไซต์ เฟซบุ๊ก ส่วนข้อความโฆษณาอาหารที่ผิดกฎหมาย หมายถึง ข้อความที่โฆษณาคุณประโยชน์ คุณภาพหรือ สรรพคุณของอาหารที่เป็นเท็จหรือหลอกลวงให้เกิดความหลงเชื่อ

การเตรียมข้อมูล

ผู้วิจัยนำโฆษณาอาหารที่ได้รับจากสำนักงานคณะกรรมการอาหารและยา มาคัดเลือกเฉพาะข้อความโฆษณาเท่านั้นโดยลบข้อมูลที่ไม่เกี่ยวข้อง เช่น เลขที่ใบอนุญาตโฆษณา ชื่อการค้า ชื่อบริษัท อักขระพิเศษ หลังจากนั้น นำข้อความโฆษณาที่ได้มาผ่านการตัดเป็นโทเคน (token) ซึ่งหมายถึง คำ โดยใช้โปรแกรมตัดคำภาษาไทยแบบอิงพจนานุกรม (dictionary based) โดยใช้



รูปที่ 1. ขั้นตอนการดำเนินการ

เทคนิคการเลือกคำที่ยาวที่สุด (longest matching) ในการตัดเป็นคำจากเล็กซ์โต (LexTo) และกำจัดคำหยุด ตัวอย่างคำหยุด เช่น ก็ การ ความ จะ ทั้งนี้ เป็นต้น โดยกำหนดให้ข้อความ 1 ข้อความ อยู่ใน 1 กลุ่มเท่านั้น (single-label) การศึกษาแบ่งข้อความออกเป็น 2 กลุ่ม ได้แก่ กลุ่มที่ถูกกฎหมายและกลุ่มที่ผิดกฎหมายตามการพิจารณาของสำนักงานคณะกรรมการอาหารและยา จากนั้นนำชุดคำที่ได้ มาสร้างคุณลักษณะ (feature) 2 แบบ คือ แบบยูนิแกรม (1-grams) คือ ใช้คุณลักษณะแบบคำเดียวที่ได้จากการตัดคำ (โดยไม่รวมคำหยุด) และแบบไบแกรม (2-grams) หรือการนำชุดคำที่อยู่ติดกัน มาสร้างเป็นชุดคำใหม่ โดยกำหนดให้ใช้ชุดคำมากที่สุดเพียง 2 ชุดคำซึ่ง คุณลักษณะแบบยูนิแกรม จะถูกนำมารวมในแบบไบแกรมด้วย (ตารางที่ 1)

ตารางที่ 1. ตัวอย่างของการเตรียมข้อมูล

ข้อความโฆษณาที่ผ่านการลบข้อมูลที่ไม่เกี่ยวข้อง	ทำลายเซลล์เนื้องอกและเซลล์มะเร็ง
คุณลักษณะแบบยูนิแกรม	“ทำลาย” “เซลล์” “เนื้องอก” “เซลล์” “มะเร็ง”
คุณลักษณะแบบไบแกรม	“ทำลาย” “เซลล์” “เนื้องอก” “เซลล์” “มะเร็ง” “ทำลาย เซลล์” “เซลล์ เนื้องอก” “เนื้องอก เซลล์” “เซลล์ มะเร็ง”

การเลือกคุณลักษณะ

งานวิจัยนี้ใช้การเลือกคุณลักษณะ (feature selection) ด้วยวิธีเคที่ดีที่สุด (select k best) โดยกำหนดให้ k คือ จำนวนคุณลักษณะ เท่ากับร้อยละ 10, 20, 30, 40, 50, 60, 70, 80, และ 90 ของคุณลักษณะ แบบยูนิแกรม หรือแบบไบแกรม

การให้น้ำหนักกับคุณลักษณะ

การให้น้ำหนักของคุณลักษณะ (feature weighting) ที่ใช้เป็นแบบ tf-idf (term frequency-inverse document frequency) เพื่อใช้จำแนกตามความสำคัญ (12) โดย

$$w_{i,j} = tf_{i,j} \times idf_i$$

$$idf_i = \log\left(\frac{|D|}{df_i}\right)$$

$w_{i,j}$ คือ น้ำหนักของคุณลักษณะ (แบบยูนิแกรม หรือแบบไบแกรม) t ลำดับที่ i ในข้อความโฆษณา d ลำดับที่ j

$tf_{i,j}$ คือ จำนวนครั้งที่คุณลักษณะ t ลำดับที่ i เกิดขึ้นในข้อความโฆษณา d ลำดับที่ j

idf_i คือ สัดส่วนของจำนวนข้อความโฆษณาทั้งหมดต่อจำนวนข้อความโฆษณาที่ปรากฏคุณลักษณะ t ลำดับที่ i

$|D|$ คือ จำนวนข้อความโฆษณาทั้งหมด

df_i คือ จำนวนข้อความโฆษณา ซึ่งปรากฏคุณลักษณะ t ลำดับที่ i

เวกเตอร์ตัวแทนข้อความโฆษณา

เนื่องจากข้อความโฆษณามีความยาวแตกต่างกัน ข้อความที่ยาวกว่ามีจำนวนคุณลักษณะมากกว่า ในขณะที่ข้อความสั้นประกอบด้วยไม่กี่คุณลักษณะ น้ำหนักของคุณลักษณะจึงต้องทำให้เป็นมาตรฐาน โดยเทียบเคียงความยาวที่แตกต่างกัน ให้มีความยาวของเอกสารที่มีขนาดเท่ากันและเป็นเวกเตอร์ขนาด 1 หน่วย (unit vector) ในงานวิจัยนี้ใช้ L2-normalization โดยคำนวณระยะทางของพิกัดเวกเตอร์จากจุดกำเนิดหรือเรียกอีกชื่อว่า Euclidean norm ดังแสดงในสมการดังนี้

$$|d_j| = \sqrt{\sum_i |w_{i,j}|^2}$$

$|d_j|$ คือ ขนาดเวกเตอร์ข้อความโฆษณา d ลำดับที่ j

$w_{i,j}$ คือ น้ำหนักของคุณลักษณะ t ลำดับที่ i ในข้อความโฆษณา d ลำดับที่ j

ดังนั้น $\widehat{w_{i,j}}$ ซึ่งเป็นน้ำหนักของคุณลักษณะ t ลำดับที่ i ในข้อความโฆษณา d ลำดับที่ j ที่ถูกปรับขนาดมาตรฐานแบบเวกเตอร์หนึ่งหน่วยสามารถแสดงได้ในสมการ

$$\widehat{w_{i,j}} = \frac{w_{i,j}}{|d_j|}$$

ขนาดของเวกเตอร์ที่ได้ ถูกนำมาจัดรูปแบบเป็นเวกเตอร์ตัวแทนข้อความโฆษณา โดยประเภทข้อความที่ถูกกฎหมายกำหนดเป็น true และผิดกฎหมายกำหนดเป็น false (ตารางที่ 2)

การสร้างแบบจำลองจำแนกข้อความ

งานวิจัยนี้ใช้การเรียนรู้ของเครื่องแบบมีผู้สอน (supervised learning) โดยการสร้างและทดสอบแบบจำลอง

ตารางที่ 2. เวกเตอร์ตัวแทนข้อความโฆษณาที่มีขนาดเป็น 1 หน่วย

คุณลักษณะ	ข้อความโฆษณา			
	ข้อความโฆษณา$_1$	ข้อความโฆษณา$_2$	..	ข้อความโฆษณา$_n$
คุณลักษณะ$_1$	$\widehat{w_{1,1}}$	$\widehat{w_{1,2}}$	...	$\widehat{w_{1,J}}$
คุณลักษณะ$_2$	$\widehat{w_{2,1}}$	$\widehat{w_{2,2}}$	...	$\widehat{w_{2,J}}$
...	...	...	...	...
คุณลักษณะ$_m$	$\widehat{w_{m,1}}$	$\widehat{w_{m,2}}$	...	$\widehat{w_{m,n}}$
ประเภทข้อความ	true	false	...	true

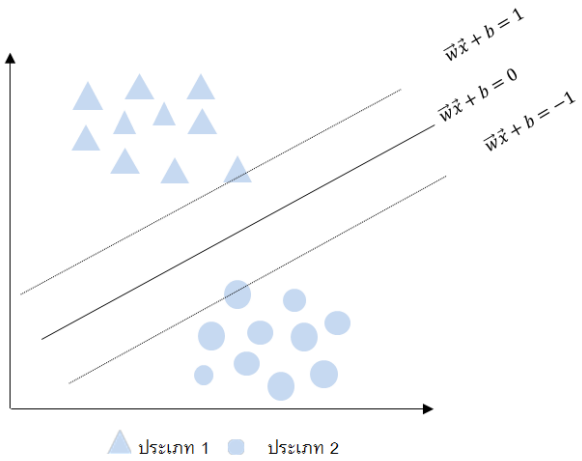
ข้อมูลด้วยโปรแกรมภาษา PHP และชุดไลบรารี PHP-ML สำหรับการเรียนรู้ด้วยเครื่องถูกใช้เพื่อสร้างชุดโปรแกรมโดยจะใช้ข้อความโฆษณา 200 ข้อความ สุ่มข้อความที่ใช้เป็นชุดสำหรับสร้างแบบจำลอง (training set) และ ชุดสำหรับทดสอบแบบจำลอง (test set) โดยกำหนดสัดส่วนให้ร้อยละ 80 ของข้อความใช้สำหรับสร้างแบบจำลอง และอีกร้อยละ 20 ใช้สำหรับทดสอบแบบจำลอง โดยมีสัดส่วนที่เท่ากันของกลุ่มข้อความที่ถูกกฎหมายและที่ผิดกฎหมาย (stratified random spilt) การสร้างแบบจำลองสำหรับจำแนกข้อความใช้ 3 เทคนิคการเรียนรู้ ได้แก่ SVM, k-NN และ NB

SVM: SVM อาศัยหลักการของการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นแบ่งข้อมูลทั้ง 2 ประเภทออกจากกัน โดยหาเส้นตรงที่ดีที่สุดในการจำแนกประเภท (13) ในงานวิจัยนี้ใช้สมการเส้นตรงในการจัดประเภท เนื่องจากในการทดลองเบื้องต้น สมการเส้นตรงให้ผลการจัดประเภทข้อความในงานวิจัยนี้ได้ดี สมการเส้นตรงที่ดีที่สุดในการจำแนกประเภท คำนวณได้จาก $g(\vec{x}) = \vec{w}\vec{x} + b$ โดย

$\vec{w}$ คือ เวกเตอร์น้ำหนัก

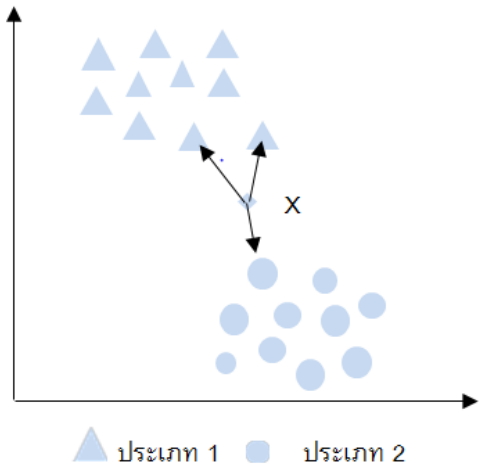
$\vec{x}$ คือ เวกเตอร์ของข้อความที่ต้องการจำแนกประเภท และ b คือ ค่า bias

ถ้า $g(\vec{x}) \geq 1$ ข้อความโฆษณานั้น จัดเป็นประเภท 1 แต่ถ้า $g(\vec{x}) \leq -1$ ข้อความโฆษณานั้นจัดเป็น ประเภท 2 (รูปที่ 2)



รูปที่ 2. ตัวอย่างการจำแนกประเภทโดย SVM

k-NN: หลักการของเทคนิค k-NN คือ หากกำหนดให้ข้อมูลที่ทดสอบ คือ x และ k คือจำนวนเพื่อนบ้านที่ใกล้ที่สุด เทคนิค k-NN จะค้นหาเพื่อนบ้านที่ใกล้ที่สุดจากชุดฝึกอบรมทั้งหมด หาก x อยู่ใกล้กับข้อมูลประเภทใดมากที่สุด ข้อมูล x จะถูกจำแนกเป็นประเภทนั้น (14) ในงานวิจัยนี้กำหนดให้ k = 3 และใช้ค่าความเหมือนโคไซน์ (cosine similarity) ระหว่างเอกสารในชุดเรียนรู้กับเอกสารทดสอบ (รูปที่ 3)



รูปที่ 3. ตัวอย่างการจำแนกประเภทโดย k-NN

NBB: การจำแนกประเภทข้อความโดยเทคนิค NBB ถูกใช้อย่างกว้างขวาง โดยพื้นฐานมาจากทฤษฎีของเบย์ (Bayes' theorem) ที่อาศัยความน่าจะเป็นในการคำนวณ เพื่อจำแนกประเภท (13, 15) สามารถคำนวณได้จาก

$$P(c_k|d_j) = argmax \frac{P(d_j|c_k)P(c_k)}{P(d_j)}$$

โดย $P(c_k|d_j)$ คือ ความน่าจะเป็นของประเภทข้อความ c ลำดับที่ k ของข้อความโฆษณา d ลำดับที่ j

$P(c_k)$ ความน่าจะเป็นของประเภทข้อความ c ที่ k ต่อประเภทข้อความทั้งหมด

$P(d_j)$ ความน่าจะเป็นของข้อความโฆษณา d ลำดับที่ j ต่อข้อความโฆษณาทั้งหมด

$P(d_j|c_k)$ ความน่าจะเป็นของข้อความโฆษณา d ลำดับที่ j เมื่อกำหนดประเภทข้อความ c ลำดับที่ k

เนื่องจากข้อความโฆษณา d ลำดับที่ j ประกอบด้วย กลุ่มอักษร t ลำดับที่ 1 จนถึงลำดับที่ n ดังนั้น $P(d_j|c_k)$ จึงคำนวณได้จาก $P(d_j|c_k) = \prod_{i=1}^{n} P(t_i|c_k)$ สามารถเขียนเป็นสมการใหม่ได้ ดังนี้

$$P(c_k|d_j) = argmax \frac{P(c_k)\prod_{i=1}^{n} P(t_i|c_k)}{P(d_j)}$$

หากความน่าจะเป็นของประเภทข้อความใดมากกว่า ข้อความโฆษณาจะจัดอยู่ในประเภทนั้น

การประเมินประสิทธิภาพแบบจำลอง

จากการเลือกคุณลักษณะและสร้างแบบจำลองการจำแนกข้อความ การศึกษาเปรียบเทียบประสิทธิภาพดังนี้1) เปรียบเทียบประสิทธิภาพของ 3 เทคนิคการเรียนรู้ระหว่างกำจัดคำหยุดและไม่กำจัดคำหยุด 2) เปรียบเทียบประสิทธิภาพของ 3 เทคนิคการเรียนรู้ที่มีการกำจัดคำหยุดระหว่างการสร้างคุณลักษณะแบบยูนิแกรมและแบบไบแกรม และ 3) เปรียบเทียบประสิทธิภาพของ 3 เทคนิคการเรียนรู้ที่มีการกำจัดคำหยุด โดยเลือกคุณลักษณะเดที่ดีที่สุดระหว่างการสร้างคุณลักษณะแบบยูนิแกรมและแบบไบแกรม โดยทำซ้ำอย่างละ 10 ครั้ง แล้วหาเฉลี่ยคะแนน F1 (14) ซึ่งคำนวณได้จาก

$$F1 = 2 \times \left(\frac{precision \times recall}{precision + recall}\right)$$

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

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

โดย TP คือ ข้อความโฆษณาที่อยู่ในประเภท p และแบบจำลองทำนายว่า อยู่ในประเภท p

FP คือ ข้อความโฆษณาที่ไม่อยู่ในประเภท p และแบบจำลองทำนายว่า อยู่ในประเภท p

FN คือ ข้อความโฆษณาที่อยู่ในประเภท p และแบบจำลองทำนายว่า ไม่อยู่ในประเภท p

TN คือ ข้อความโฆษณาที่ไม่อยู่ในประเภท p และแบบจำลองทำนายว่า ไม่อยู่ในประเภท p

ในงานวิจัยนี้ ประเภท p คือประเภทที่ถูกกฎหมาย ดังนั้นถ้าไม่อยู่ในประเภท p แสดงว่า เข้าข่ายผิดกฎหมาย

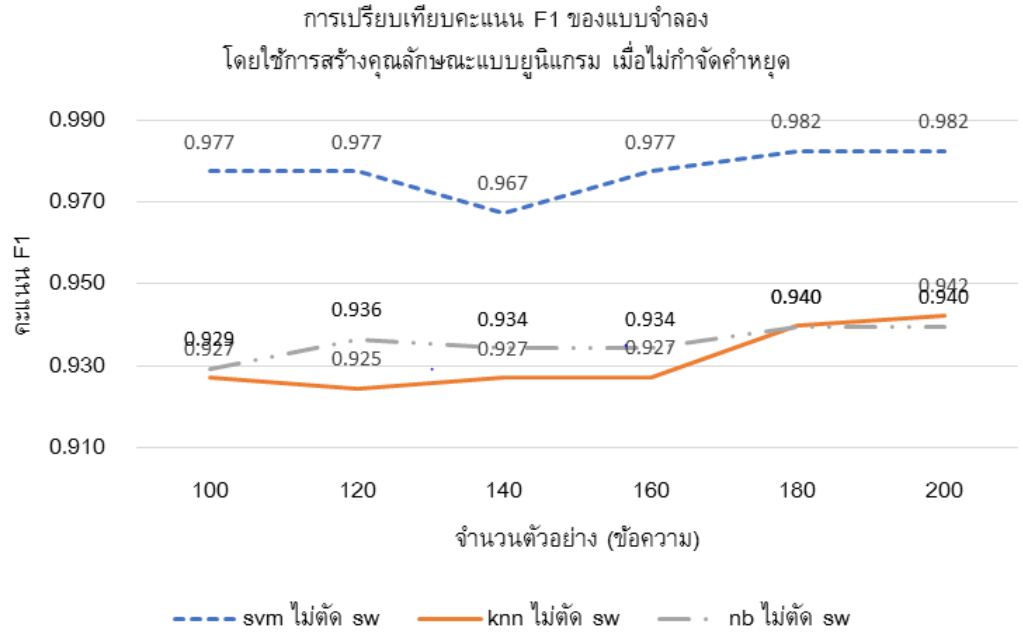
โปรแกรมตรวจจับข้อความโฆษณา

จากการประเมินประสิทธิภาพแบบจำลองการศึกษาเลือกแบบจำลองจากเทคนิคการเรียนรู้ที่มีคะแนน F1 เฉลี่ยมากที่สุดมาสร้างโปรแกรมตรวจจับข้อความโฆษณา และทดสอบประสิทธิภาพโปรแกรม โดยใช้ข้อความโฆษณา จำนวน 40 ข้อความ ซึ่งเป็นข้อความโฆษณาที่ถูกกฎหมายหรือผิดกฎหมาย อย่างละ 20 ข้อความ การเปรียบเทียบประสิทธิภาพทำโดยคำนวณคะแนน F1

ผลการวิจัย

งานวิจัยนี้ได้มีการทดลองเพื่อหาจำนวนตัวอย่างที่เหมาะสมก่อนเริ่มการทำวิจัย โดยเริ่มจากจำนวนข้อความ 100 ข้อความ เพิ่มขึ้นทีละ 20 ข้อความ จนถึง 200 ข้อความ โดยสุ่มตัวอย่างให้ร้อยละ 80 ของข้อความใช้สำหรับจัดทำแบบจำลอง และอีกร้อยละ 20 ใช้สำหรับทดสอบการใช้งาน จากการทดสอบเปรียบเทียบประสิทธิภาพ 3 เทคนิคการเรียนรู้ ในการสร้างแบบจำลองโดยใช้คุณลักษณะแบบยูนิแกรมที่ไม่กำจัดคำหยุด พบว่า การเพิ่มจำนวนตัวอย่าง ทำให้ประสิทธิภาพของแบบจำลองเพิ่มขึ้นไม่มากนัก ตามรูปที่ 4 ผู้วิจัยจึงใช้ตัวอย่างข้อความโฆษณาอาหารจำนวน 200 ข้อความ ในการวิจัย

การใช้เทคนิคการเรียนรู้สำหรับจำแนกข้อความโฆษณาอาหาร ได้ใช้ตัวอย่างข้อความโฆษณาอาหารที่ถูกกฎหมายจำนวน 100 ข้อความและที่ผิดกฎหมาย จำนวน 100 ข้อความ รวม 200 ข้อความ โดยสุ่มตัวอย่างให้ร้อยละ 80 ของข้อความใช้สำหรับจัดทำแบบจำลอง และอีกร้อยละ 20 ใช้สำหรับทดสอบการใช้งาน จากการทดสอบเปรียบเทียบประสิทธิภาพ 3 เทคนิคการเรียนรู้ ในการสร้างแบบจำลองแบบยูนิแกรมที่ไม่กำจัดคำหยุด และแบบจำลองกำจัดคำหยุด พบว่า ใช้ชุดคำในการสร้างแบบจำลอง 2,640 หน่วยเมื่อกำจัดคำหยุดแล้ว เหลือชุดคำในการสร้างแบบจำลอง 2,532 หน่วย แบบจำลองที่กำจัดคำหยุดทั้ง 3 เทคนิคการเรียนรู้ให้ประสิทธิภาพในการจำแนกข้อความดีกว่าแบบจำลองที่ไม่กำจัดคำหยุด ซึ่ง SVM มีประสิทธิภาพในการจำแนกดีที่สุด โดยได้คะแนน F1 เท่ากับ 0.987 (รูปที่ 5)



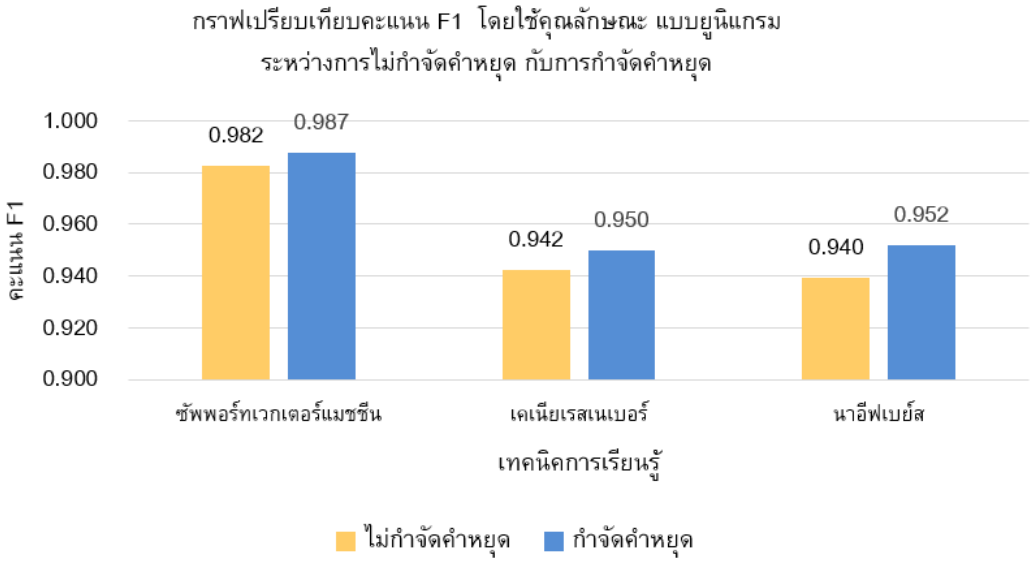
รูปที่ 4. คะแนน F1 ของแบบจำลองระหว่างการใช้จำนวนตัวอย่าง 100, 120, 140, 160, 180 และ 200 ข้อความ

จากนั้นการศึกษาเลือกใช้แบบจำลองที่กำจัดคำหยุด เพื่อทดสอบเปรียบเทียบประสิทธิภาพ 3 เทคนิคการเรียนรู้ ระหว่างการสร้างคุณลักษณะแบบยูนิแกรมและแบบไบแกรม พบว่า SVM ด้วยการสร้างคุณลักษณะแบบยูนิแกรม ยังคงมีประสิทธิภาพในการจำแนกดีที่สุด ตามรูปที่ 6

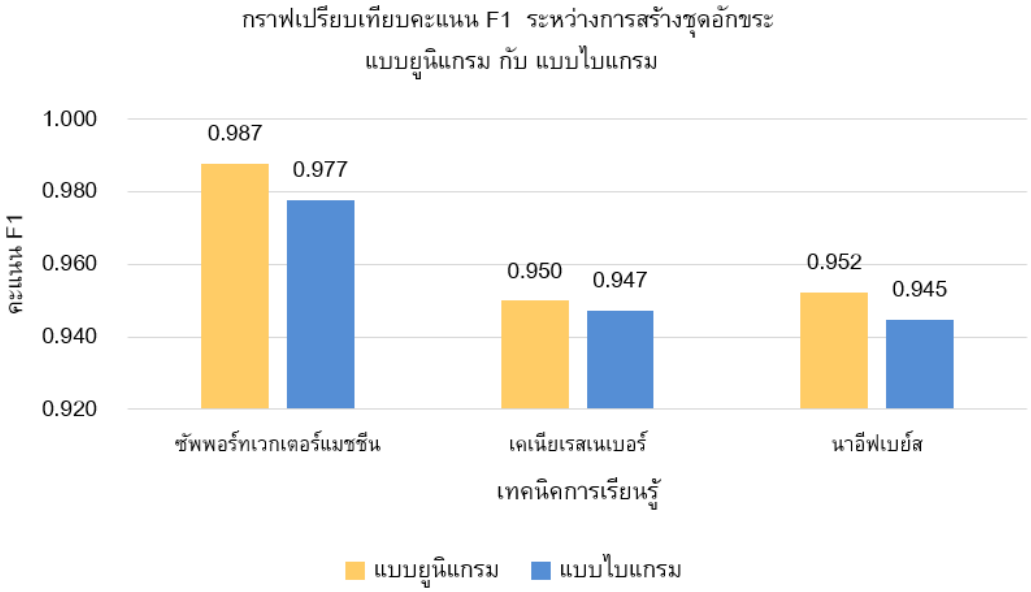
เมื่อนำแบบจำลองที่กำจัดคำหยุดที่มีการสร้างคุณลักษณะแบบยูนิแกรมและแบบไบแกรม มาเพิ่มการเลือกใช้คุณลักษณะเคที่ดีที่สุด จะได้จำนวนคุณลักษณะในการสร้างแบบจำลองตามตารางที่ 3

เมื่อทดสอบแบบจำลองที่กำจัดคำหยุด ที่มีการสร้างคุณลักษณะ แบบยูนิแกรม และเลือกคุณลักษณะ เคที่ดีที่สุด พบว่า ซัพพอร์ทเวกเตอร์แมชชีนให้ประสิทธิภาพในการจำแนกดีที่สุด ที่ k เท่ากับร้อยละ 80 คะแนน F1 เท่ากับ 0.985 ตามตารางที่ 4

เมื่อทดสอบแบบจำลองที่กำจัดคำหยุดที่มีการสร้างคุณลักษณะแบบไบแกรมและเลือกคุณลักษณะ เคที่ดีที่สุด พบว่า SVM ให้ประสิทธิภาพในการจำแนกดีที่สุด ที่ k เท่ากับร้อยละ 90 คะแนน F1 เท่ากับ 0.987 ตามตารางที่ 5



รูปที่ 5. คะแนน F1 เมื่อใช้คุณลักษณะแบบยูนิแกรมระหว่างการไม่กำจัดคำหยุดกับการกำจัดคำหยุด



รูปที่ 6. คะแนน F1 ระหว่างการสร้างคุณลักษณะแบบยูนิแกรมและแบบไบแกรม

แบบจำลองที่มีคะแนน F1 เฉลี่ยมากที่สุดของทั้งยูนิแกรมและไบแกรม ได้แก่ SVM แบบยูนิแกรมที่ใช้ทุกคุณลักษณะที่มีการกำจัดคำหยุด (SVM-U-100) โดยคะแนน F1 เท่ากับ 0.987 และ SVM แบบไบแกรมที่ k เท่ากับร้อยละ 90 (SVM-B-90) โดยคะแนน F1 เท่ากับ 0.987 เมื่อนำทั้งสองแบบจำลองมาสร้างโปรแกรมตรวจจับข้อความ และทดสอบประสิทธิภาพโปรแกรมโดยใช้ข้อความโฆษณา จำนวน 40 ข้อความ

ตารางที่ 3. จำนวนคุณลักษณะที่ใช้สร้างแบบจำลองจากการเลือกคุณลักษณะเคที่ดีที่สุด

เคที่ดีที่สุด (ร้อยละ)	จำนวนคุณลักษณะที่ใช้สร้างแบบจำลอง	
	แบบยูนิแกรม (2,561 หน่วย)	แบบไบแกรม (13,264 หน่วย)
10	256 หน่วย	1,326 หน่วย
20	512 หน่วย	2,653 หน่วย
30	768 หน่วย	3,979 หน่วย
40	1,024 หน่วย	5,306 หน่วย
50	1,281 หน่วย	6,632 หน่วย
60	1,537 หน่วย	7,958 หน่วย
70	1,793 หน่วย	9,285 หน่วย
80	2,049 หน่วย	10,611 หน่วย
90	2,305 หน่วย	11,938 หน่วย

พบว่า SVM เมื่อสร้างคุณลักษณะแบบยูนิแกรม สามารถจำแนกข้อความได้ถูกต้องจำนวน 39 ข้อความ ส่วน SVM เมื่อสร้างคุณลักษณะแบบไบแกรม และกำหนดค่าเคที่ดีที่สุดที่ร้อยละ 90 สามารถจำแนกข้อความได้ถูกต้องจำนวน 36 ข้อความ ตามตารางที่ 6

การอภิปรายผลและสรุป

งานวิจัยนี้ได้ทดสอบประสิทธิภาพเทคนิคการเรียนรู้สำหรับจำแนกข้อความโฆษณาอาหาร 3 เทคนิค โดยเปรียบเทียบการกำจัดคำหยุด การสร้างคุณลักษณะ และการเลือกคุณลักษณะ เพื่อให้ได้แบบจำลองที่สามารถจำแนกข้อความได้ดีที่สุด ผลการทดลองพบว่า SVM ให้ประสิทธิภาพในการจำแนกดีที่สุดเมื่อสร้างคุณลักษณะแบบไบแกรม และกำหนดค่าเคที่ดีที่สุดที่ร้อยละ 90 หรือเมื่อสร้างคุณลักษณะแบบยูนิแกรมที่ใช้ทุกคุณลักษณะ เมื่อนำแบบจำลองมาสร้างโปรแกรมตรวจจับข้อความทดสอบประสิทธิภาพโปรแกรมโดยใช้ข้อความโฆษณา จำนวน 40 ข้อความ พบว่า SVM เมื่อสร้างคุณลักษณะแบบยูนิแกรม มีประสิทธิภาพดีกว่า โดยสามารถจำแนกข้อความได้ถูกต้องจำนวน 39 ข้อความ จากการทดลองจะเห็นได้ว่า การกำจัดคำหยุดมีผลต่อประสิทธิภาพในการจำแนกข้อความ เนื่องจากเมื่อทำการกำจัดคำหยุด ทั้ง 3 เทคนิคการเรียนรู้ คะแนน F1 มีค่าสูงขึ้น นั่นคือสามารถจำแนกข้อความได้มีประสิทธิภาพมากขึ้น

ตารางที่ 4. คะแนน F1 โดยใช้สร้างคุณลักษณะแบบยูนิแกรม

เทคนิคการเรียนรู้[1]	คะแนน F1 เมื่อกำหนดค่า k ที่ระดับต่าง ๆ (ร้อยละ)								
	10	20	30	40	50	60	70	80	90
SVM	0.930	0.930	0.947	0.957	0.937	0.952	0.975	0.985	0.980
k-NN	0.937	0.952	0.955	0.950	0.972	0.955	0.945	0.947	0.935
NB	0.909	0.932	0.915	0.947	0.950	0.967	0.957	0.927	0.942

1: support vector machine (SVM); k-nearest neighbors: k-NN; naïve Bayes: NB

ตารางที่ 5. คะแนน F1 โดยใช้สร้างคุณลักษณะแบบไบแกรม

เทคนิคการเรียนรู้[1]	คะแนน F1 เมื่อกำหนดค่า k ที่ระดับต่าง ๆ (ร้อยละ)								
	10	20	30	40	50	60	70	80	90
SVM	0.942	0.965	0.950	0.960	0.970	0.982	0.975	0.985	0.987
k-NN	0.940	0.972	0.972	0.957	0.970	0.977	0.947	0.957	0.945
NB	0.915	0.917	0.945	0.957	0.955	0.950	0.965	0.952	0.945

1: support vector machine (SVM); k-nearest neighbors: k-NN; naïve Bayes: NB

ในแบบจำลองที่ใช้คุณลักษณะแบบไบแกรม จำนวนคุณลักษณะมีมากขึ้น เมื่อใช้ร่วมกับการกำหนดค่าเคที่ดีที่สุด ทำให้ได้เฉพาะคุณลักษณะที่มีความสำคัญมาก มาจำแนกข้อความ ซึ่งประสิทธิภาพที่ได้นั้นใกล้เคียงกับแบบจำลองที่ใช้คุณลักษณะแบบยูนิแกรม ในงานนี้จะเห็นได้ว่า แบบจำลองที่ใช้คุณลักษณะแบบไบแกรมไม่ได้ทำให้แบบจำลองที่มีประสิทธิภาพดีขึ้นกว่าคุณลักษณะแบบยูนิแกรม อาจเนื่องมาจากกลุ่มคำที่มีความสำคัญที่ใช้ในการแยกประเภทเอกสารนั้น เป็นกลุ่มคำที่อยู่ในคุณลักษณะแบบยูนิแกรม และเนื่องจากการใช้คุณลักษณะแบบยูนิแกรมมีคุณลักษณะน้อย ทำให้การประมวลผลทั้งขั้นตอนของการสร้างเรียนรู้เพื่อสร้างแบบจำลอง และการทดสอบข้อความโฆษณาทำได้เร็วขึ้น ในการใช้งานจริงจึงสร้างแบบจำลองโดยใช้คุณลักษณะแบบยูนิแกรม ซึ่งคุณลักษณะแบบยูนิแกรมสามารถทำให้ประสิทธิภาพของ k-NN และ NB ดีเช่นกัน โดยการเลือกจำนวนคุณลักษณะที่เหมาะสมกับแต่ละเทคนิคของการเรียนรู้ จากผลการวิจัยนี้สามารถนำแบบจำลองไปพัฒนาเป็นแอปพลิเคชัน เพื่อช่วยงานเจ้าหน้าที่ในภาครัฐเพื่อตรวจสอบข้อความโฆษณาอาหารที่ผิดกฎหมาย ในภาคเอกชน โปรแกรมช่วยผู้ประกอบการให้สามารถตรวจสอบข้อความโฆษณาได้ถูกต้องก่อนยื่นคำขออนุญาต และช่วยผู้บริโภคให้สามารถตรวจสอบข้อความโฆษณาในเบื้องต้นได้ อีกทั้งยังสามารถประยุกต์ใช้กับการโฆษณาผลิตภัณฑ์สุขภาพประเภทอื่น ๆ ได้

ตารางที่ 6. ค่าที่ทดสอบของโปรแกรมตรวจจับข้อความโฆษณา

ค่าที่ทดสอบ	โปรแกรมตรวจจับข้อความโฆษณา	
	SVM-U-100	SVM-B-90
TP	19	16
FP	0	0
FN	1	4
TN	20	20
precision	1	1
recall	0.950	0.800
คะแนน F1	0.974	0.889

ข้อเสนอแนะ

สำหรับข้อความโฆษณานั้น ผู้โฆษณาสามารถปรับเปลี่ยนข้อความได้หลากหลาย อาจใช้คำที่มีความหมายเหมือนกัน (synonym) เพื่อให้ข้อความโฆษณามีความหมายเช่นเดิม ซึ่งงานวิจัยนี้ไม่ได้ทดลองคำที่มีความหมายเหมือนกัน ดังนั้น หากเป็นข้อความที่ใช้คำที่มีความหมายเดียวกัน อาจทำให้จำแนกข้อความได้ไม่ถูกต้อง นอกจากนี้การโฆษณาอาหารยังพบในหลากหลายสื่อเป็นจำนวนมาก ไม่ว่าจะเป็นสื่อสิ่งพิมพ์ สื่อกระจายเสียง สื่อออนไลน์ การที่จะดำเนินการตรวจสอบโฆษณาทั้งหมดได้

นั้น ต้องใช้ทั้งบุคลากร และระยะเวลา เป็นจำนวนมาก โดยการโฆษณานอกจากใช้ข้อความในการสื่อสารแล้ว ยังมีการใช้ รูปภาพ เสียง วิดีทัศน์ ซึ่งงานวิจัยนี้ทำการทดลองกับข้อความเท่านั้น ดังนั้น เพื่อเป็นการคุ้มครองผู้บริโภค เจ้าหน้าที่สามารถบังคับใช้กฎหมายได้อย่างรวดเร็วทันกับสถานการณ์ จึงควรนำเทคโนโลยีเข้ามาประยุกต์ใช้ในการตรวจจับโฆษณาที่ผิดกฎหมายในสื่ออื่น ๆ

กิตติกรรมประกาศ

การวิจัยครั้งนี้สำเร็จได้ด้วยความอนุเคราะห์จากสำนักงานคณะกรรมการอาหารและยา ที่ให้ความอนุเคราะห์ข้อมูล และทุนอุดหนุนการวิจัยประเภทนักศึกษาระดับบัณฑิตศึกษาจากกองทุนวิจัยและสร้างสรรค์ คณะเภสัชศาสตร์ มหาวิทยาลัยศิลปากร

เอกสารอ้างอิง

1. Food Act, B.E. 2522 Royal Gazette No.96, Part 79A special (May 13, 1979).
2. Announcement of the Food and Drug Administration Re: criteria for food advertisement B.E. 2561. Royal Gazette No.135, Part 322D special (December 17, 2018).
3. Story M, French S. Food advertising and marketing directed at children and adolescents in the US. Int J Behav Nutr Phys Act 2004; 1: 3.
4. Chapman K, Nicholas P, Supramaniam R. How much food advertising is there on Australian television?. Health Promot Int 2006; 21: 172-80.
5. Harris JL, Bargh JA, Brownell KD. Priming effects of television food advertising on eating behavior. Health Psychol 2009; 28: 404-13.
6. Shalev-Shwartz S, Ben-David S. Introduction. Understanding machine learning: From theory to algorithms. New York: Cambridge university press; 2014. p. 19-23.
7. Jindal R, Malhotra R, Jain A. Techniques for text classification: Literature review and current trends. Webology 2015; 12: 1-28.
8. Chirawichitchai N, Sa-nguansat P, Meesad P. Developing and effective automatic Thai document categorization. NIDA Development Journal 2011; 51: 187-205.
9. Chatcharaporn K, Angskun T, Angskun J. Tourist attraction categorization models using machine learning techniques. Suranaree Journal of Science and Technology 2012; 6: 35-58.
10. Tipsena R, Jareanpon C, Somprasertsri G. Automatic question classification on webboard using text mining techniques. Journal of Science and Technology Mahasarakham University. 2014; 33: 493-502.
11. Foundation for Consumers. Foundation for Consumers reveals the situation of consumers in 2018, found number 1 exaggerated ads [online]. 2019 [cited Sep 26, 2019]. Available from: www.consumerthai.org/news-consumerthai/ffc-news/4302-620124comsumerstat.html.
12. Havrlant L, Kreinovich V. A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). Int J Gen Syst 2017; 46: 27-36.
13. Mohammad AH, Alwada'n T, Al-Momani O. Arabic text categorization using support vector machine, Naïve Bayes and neural network. GSTF Int J Comput 2016; 5: 108.
14. Khamar K. Short text classification using kNN based on distance function. Int J Adv Res Comput Commun Eng 2013; 2: 1916-9.
15. Al-Khurayji R, Sameh A. An effective arabic text classification approach based on kernel naive bayes classifier. Int J Artif Intell Appl 2017; 8: 1-10.