การจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะโดยใช้วิธีการสกัดข้อความและเทคนิคการเรียนรู้ของเครื่อง

Classification for Bus Driver's Behaviors Using Text Extraction and Machine Learning Technique

สถิตย์โชค โพธิ์สอาด (Satidchoke Phosaard)* และ ปิติภูมิ โพสาวัง (Pitiphum Posawang)**

Received : March 15, 2018
Revised : July 8, 2018
Accepted : July 8, 2018

บทคัดย่อ

พฤติกรรมการขับขี่รถโดยสารมีความสำคัญต่อคุณภาพและความปลอดภัยของระบบขนส่งสาธารณะ การรวบรวมและวิเคราะห์พฤติกรรมของผู้ขับขี่รถโดยสารสาธารณะเพื่อให้ผู้ประกอบการนำมากำหนดนโยบายและจัดการกับความผิดพลาดของผู้ขับขี่ เพื่อรับประกันความปลอดภัยและคุณภาพในการให้บริการ ลักษณะของเว็บไซต์สนทนาข้อร้องเรียนเป็นที่นิยมของผู้โดยสารเพื่อยื่นข้อร้องเรียนในการใช้บริการ การคัดกรองและจัดหมวดหมู่ข้อร้องเรียนเกี่ยวกับบริการจึงเป็นงานที่ซับซ้อนและต้องใช้ทรัพยากรโดยไม่จำเป็น งานวิจัยนี้เสนอเทคนิคการสกัดข้อความและเทคนิคการเรียนรู้ของเครื่อง เพื่อจำแนกประเภทข้อร้องเรียนแบบอัตโนมัติที่เกี่ยวกับพฤติกรรมการขับขี่และคุณภาพการบริการ จำนวน 4 คลาส ได้แก่ 1) คลาสการจอดรับส่งผู้โดยสาร 2) คลาสลักษณะการขับขี่ 3) คลาสการเดินรถ และ 4) คลาสคุณภาพการบริการ สำหรับการสกัดข้อความใช้หลักการตัดคำจากพจนานุกรมคำศัพท์เทียบกับคำที่ยาวที่สุดที่พบในพจนานุกรม จากนั้นสร้างดัชนีเอกสารโดยหาค่าน้ำหนักของคำเดี่ยวและนำมาจัดให้อยู่ในรูปแบบของเวกเตอร์เอกสาร เพื่อใช้เป็นชุดข้อมูลในการสร้างตัวแบบจำแนกข้อมูลโดยใช้เทคนิคการเรียนรู้ของเครื่อง ได้แก่ ต้นไม้ตัดสินใจ นาอีฟเบย์ การหาเพื่อนบ้านใกล้ที่สุด ซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายประสาทเทียม ผลการทดลองพบว่าตัวแบบโครงข่ายประสาทเทียมสามารถจำแนกข้อความพฤติกรรมการขับรถโดยสารสาธารณะได้ค่าความแม่นตรงสูงสุดซึ่งมีค่าเท่ากับ 90.23% ค่าความแม่นยำเท่ากับ 91.9% ค่าความระลึกเท่ากับ 90.2% และค่าเอฟเมเชอร์เท่ากับ 90.5% งานวิจัยนี้สามารถนำไปประยุกต์ใช้สร้างระบบจำแนกหมวดหมู่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะและระบบจำแนกหมวดหมู่เอกสารอัตโนมัติได้

คำสำคัญ: การจำแนกพฤติกรรมการขับขี่ การสกัดข้อความ การเรียนรู้ของเครื่อง

Abstract

Bus drivers' behaviors play important roles in the quality and safety of this public transportation mode. By collecting and analyzing bus drivers' behaviors, service operators could outline policies and actions needed to properly handle bus drivers' misbehaviors to guarantee safety and good quality of services. The direct and transparency characteristics of complaint web boards make them popular among passengers to lodge service complaints. Filtering and categorizing service complaints is a complicated task and requires unnecessary resources. This work applies text extraction and machine learning techniques to automatically classification complaints of bus driver behavior and quality services into five categories: 1) bus stops 2) bus driver behavior 3) timetable and

* สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี
* School of Information Technology, Institute of Social Technology, Suranaree University of Technology.
** สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยวงษ์ชวลิตกุล
** Department of Computer Science, Faculty of Engineering, Vongchavalitkul University.

4) services quality. The text extraction employed longest matching algorithm with domain-specific augmented dictionary. The document representation of the extracted words is in the form of word vector of TF-IDF weightings. Five prominent techniques, the J48 decision tree, Naïve Bayes, K-Nearest Neighbour, Support Vector Machine and the Artificial Neural Network (ANN), were used to create classification models. The results suggested that the ANN model is superior to the other model. The ANN model's accuracy is as high as 90.23% with 91.9% precision, 90.2% recall and 90.5% F-measure. The text processing practice and the model could be used to implement a real-world system to automatically classify bus drivers' behaviors.

Keywords: Bus Drivers' Behavior Classification, Text Extraction, Machine Learning.

1. บทนำ

ข้อมูลพฤติกรรมการขับขี่รถโดยสารสาธารณะมีความสำคัญต่อการพัฒนาพฤติกรรมการขับขี่ให้ปลอดภัย และควบคุมพฤติกรรมการขับขี่ที่ไม่ปลอดภัยของผู้ขับรถ นอกจากนี้ข้อมูลพฤติกรรมการขับขี่รถโดยสารสาธารณะยังเป็นส่วนสำคัญต่อการบริหารจัดการระบบขนส่งและจราจรอัจฉริยะหรือไอทีเอส (Intelligent Transportation System: ITS) เพื่อให้ผู้ใช้รถใช้ถนนได้รับความสะดวกรวดเร็ว ความปลอดภัยในการเดินทางและหลีกเลี่ยงอุบัติเหตุที่อาจเกิดจากพฤติกรรมการขับขี่รถโดยสารสาธารณะ แต่อย่างไรก็ตาม ก่อนที่จะนำข้อมูลมาใช้ประโยชน์ในการบริหารจัดการได้นั้น จำเป็นต้องมีการศึกษาพฤติกรรมเสี่ยงของผู้ขับขี่รถโดยสารสาธารณะ

พฤติกรรมเสี่ยงของผู้ขับขี่จำแนกเป็น 4 ประเภท ได้แก่ 1) การเบรกกะทันหัน 2) การออกตัวกะทันหัน 3) การเลี้ยวรถด้วยความเร็วสูง และ 4) การเปลี่ยนช่องจราจรกะทันหัน [1] พฤติกรรมเสี่ยงเหล่านี้เป็นสาเหตุสำคัญที่ทำให้เกิดอุบัติเหตุและความไม่ปลอดภัยของผู้ใช้รถใช้ถนน ที่ผ่านมามีผู้ศึกษาวิจัยพฤติกรรมของผู้ขับขี่รถโดยสารสาธารณะ เพื่อเพิ่มความปลอดภัยให้กับผู้เดินทาง เช่น ศูนย์ความเป็นเลิศด้านโลจิสติกส์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี [2] ศึกษา การเพิ่มประสิทธิภาพการกำกับดูแลรถโดยสารประจำทางโดยใช้เทคโนโลยี พบว่า หน่วยงานที่มีการนำระบบเทคโนโลยีไปใช้ในการกำกับดูแลการเดินรถ มีทั้งหน่วยงานภาครัฐ รัฐวิสาหกิจ และภาคเอกชน เช่น องค์การขนส่งมวลชนกรุงเทพ (ขสมก.) ได้ดำเนินการโครงการตรวจสอบและติดตามการปฏิบัติการเดินรถ (Tracking and Tracing) โดยใช้เทคโนโลยี 2 ระบบ ได้แก่ ระบบการกำหนดตำแหน่งบนโลก (Global Positioning Systems: GPS) และระบบการระบุข้อมูลสิ่งต่าง ๆ โดยใช้คลื่นความถี่วิทยุ (Radio Frequency Identification: RFID)

แต่อย่างไรก็ตาม การได้มาซึ่งข้อมูลพฤติกรรมของผู้ขับขี่รถโดยสารสาธารณะนอกเหนือจากการใช้เทคโนโลยีในการตรวจสอบแล้ว ข้อมูลพฤติกรรมของผู้ขับขี่รถโดยสารสาธารณะที่ได้จากผู้โดยสารแสดงความคิดเห็นหรือร้องเรียนผ่านกระดานสนทนา (Web Board) บนเว็บไซต์ก็เป็นอีกช่องทางหนึ่งที่ใช้ในการติดตามตรวจสอบพฤติกรรมของผู้ขับขี่ได้เช่นกัน ยิ่งไปกว่านั้นกระดานสนทนายังใช้เป็นสื่อกลางในการแลกเปลี่ยนบทสนทนาหรือความคิดเห็นต่าง ๆ ผ่านระบบเครือข่ายอินเทอร์เน็ตที่ได้รับความนิยมและแพร่หลายในปัจจุบัน ผู้ใช้งานสามารถตั้งกระทู้ถามตอบเพื่อแลกเปลี่ยนความคิดเห็นกันได้อย่างอิสระ ดังนั้นกระดานสนทนาจึงถือได้ว่าเป็นอีกช่องทางหนึ่งในการสื่อสารเพื่อแลกเปลี่ยนความคิดเห็นที่ก่อให้เกิดประโยชน์และเพิ่มประสิทธิภาพในการให้บริการของหน่วยงาน

เมื่อข้อความคิดเห็นหรือข้อร้องเรียนในการใช้บริการมีเพิ่มมากขึ้น และความหลากหลายของข้อความที่อาจเกิดปัญหาในการตีความหมาย โดยเฉพาะอย่างยิ่งข้อความที่ไม่มีการจำแนกหมวดหมู่ไว้อย่างชัดเจน อาจส่งผลต่อการตอบคำถาม การให้ข้อมูลที่ถูกต้องและการสรุปสถิติการจัดประเภทข้อร้องเรียน ถึงแม้ว่าจะมีการจำแนกหมวดหมู่ข้อร้องเรียนในกระดานสนทนาไว้อย่างชัดเจนแล้วก็ตาม แต่ในบางครั้ง ผู้ใช้งานอาจตั้งกระทู้ข้อร้องเรียนผิดหมวดหมู่ เนื่องจากผู้ใช้งานเป็นผู้เลือกหมวดหมู่ด้วยตนเอง หมวดหมู่ที่เลือกอาจไม่สอดคล้องกับข้อร้องเรียนส่งผลต่อการตอบข้อร้องเรียนที่อาจได้รับคำตอบที่ไม่ชัดเจน การนำข้อร้องเรียนมารวบรวมสรุปเป็นข้อมูลทางสถิติทำได้ยากและใช้เวลาในการจัดทำเป็นเวลานานเนื่องจากผู้ดูแลระบบจะต้องนำข้อร้องเรียนมาจำแนกหมวดหมู่ด้วยตนเองอีกครั้ง หากมี

ระบบการจำแนกข้อร้องเรียนตามหมวดหมู่ที่ต้องการแบบอัตโนมัติน่าจะเป็นแนวทางหนึ่งในการแก้ไขปัญหาดังกล่าว ดังนั้น งานวิจัยนี้ จึงศึกษาวิธีการจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะจากข้อมูลข้อร้องเรียนผ่านกระดานสนทนาโดยใช้วิธีการสกัดข้อความและเทคนิคการเรียนรู้ของเครื่อง เพื่อนำผลที่ได้มาเป็นข้อมูลวิเคราะห์พฤติกรรมการขับขี่รถโดยสารสาธารณะต่อไป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การเติบโตของสารสนเทศในรูปแบบข้อความ (Textual Information) มีอัตราที่เพิ่มมากขึ้น ทั้งจากเครือข่ายอินเทอร์เน็ตและเอกสารที่องค์กรจัดเก็บไว้ จึงได้มีเทคโนโลยีการสืบค้น (Search Engine) ที่สามารถค้นหาข้อมูลด้วยคำสำคัญ (Keywords) และให้ผลลัพธ์ได้อย่างรวดเร็ว แต่อย่างไรก็ตามวิธีดังกล่าวยังไม่สามารถสรุปความ ประมวลความหมายหรือความสัมพันธ์ของคำได้อย่างตรงประเด็น ทำให้การค้นคืนเอกสารไม่ตรงกับความต้องการของผู้ใช้ ดังนั้น จึงได้มีการพัฒนาเทคนิคเหมืองข้อความ เพื่อช่วยจัดการปัญหาทั้งในส่วนการตรวจสอบหัวข้อ การติดตามและการแสดงแนวโน้ม ส่งผลให้สารสนเทศในรูปแบบข้อความสามารถเข้าถึง วิเคราะห์และประมวลผลภายในเวลาอันรวดเร็ว ผู้ใช้งานสามารถเข้าใจหรือใช้ประโยชน์จากข้อความเอกสารที่จัดเก็บไว้เป็นจำนวนมากได้อย่างมีประสิทธิภาพ [3] งานวิจัยนี้จึงใช้แนวคิดของการทำเหมืองข้อความมาจัดการข้อความจากชุดข้อมูลที่ไม่มีโครงสร้างหรือกึ่งโครงสร้าง เพื่อสกัดสารสนเทศเกี่ยวกับพฤติกรรมการขับขี่รถโดยสารสาธารณะจากข้อมูลข้อร้องเรียนผ่านกระดานสนทนาโดยมีทฤษฎีและงานวิจัยที่เกี่ยวข้อง ดังนี้

2.1 การเรียนรู้ของเครื่อง (Machine Learning)

เป็นสาขาหนึ่งของปัญญาประดิษฐ์ที่ศึกษาและสร้างขั้นตอนวิธีที่ทำให้เครื่องคอมพิวเตอร์เรียนรู้ได้จากข้อมูลตัวอย่างหรือจากสภาพแวดล้อม เพื่อสร้างตัวแบบหรือขั้นตอนวิธีและนำไปใช้หรือทำนายข้อมูลใหม่ได้ โดยมีจุดมุ่งหมายเพื่อการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น เมื่อเรียนรู้แล้วความรู้ที่เรียนได้จะถูกเก็บไว้ในฐานความรู้ด้วยรูปแบบการแทนความรู้อย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชัน ฯลฯ [4], [5]

ประเภทของการเรียนรู้ของเครื่อง ประกอบด้วย 1) การเรียนรู้แบบมีผู้สอน (Supervised Learning) เป็นการเรียนรู้จากลักษณะของข้อมูลตัวอย่างที่มีการระบุผลที่ต้องการหรือประเภทไว้แล้วนำไปทำนายข้อมูลอื่นที่ไม่รู้คำตอบ ทั้งนี้การเรียนรู้แบบมีผู้สอนสามารถนำไปประยุกต์ใช้ในการประมาณค่าของข้อมูล (Estimation) การจัดประเภทของข้อมูล (Classification) การพยากรณ์ข้อมูลในอนาคต (Prediction) และ 2) การเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) เป็นการสร้างตัวแบบที่เหมาะสมกับข้อมูล โดยไม่มีการระบุผลที่ต้องการหรือประเภทไว้ก่อน ทั้งนี้การเรียนรู้แบบไม่มีผู้สอนสามารถนำไปใช้ในการจัดกลุ่มข้อมูลได้ (Clustering) [5], [6]

การแบ่งกลุ่มของข้อร้องเรียนอย่างอัตโนมัติต้องอาศัยหลักการเรียนรู้ของเครื่อง เพื่อจัดกลุ่มข้อความ โดยเทคนิคหลักที่ใช้ในการเรียนรู้แบบมีผู้สอนซึ่งมีอัลกอริทึมแบ่งประเภท (Classification Algorithm) ที่เป็นที่นิยมและแพร่หลายในงานวิจัยต่างๆ [5], [6] ได้แก่

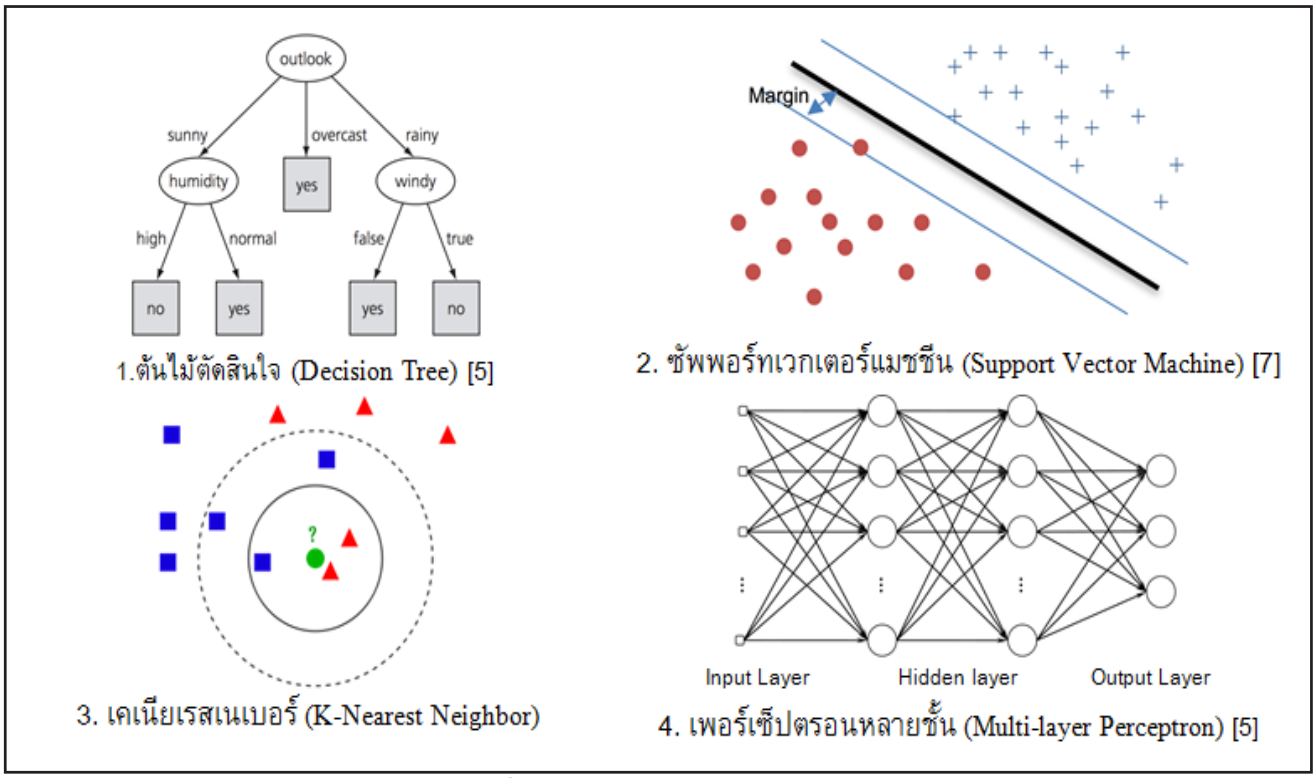
1) ต้นไม้ตัดสินใจ (Decision Tree) เป็นตัวแบบในการแบ่งประเภทโดยนำคุณลักษณะต่างๆ ของข้อมูลนำเข้ามาเป็นน้ำหนักในแต่ละโหนดของโครงสร้างข้อมูลต้นไม้ โหนดล่างสุดหรือใบ จะเป็นผลลัพธ์หรือประเภทที่จัดแบ่ง [5] ดังภาพที่ 1-1

2) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) ใช้หลักการของสมการเส้นตรงในการแบ่งกลุ่มข้อมูล [5] ดังภาพที่ 1-2

3) เนอีฟเบย์ (Naïve-Bayes) เป็นตัวแบบที่ให้ความน่าจะเป็นกับข้อมูลนำเข้าเพื่อจัดแบ่งประเภทข้อมูลนำเข้าเป็นรูปแบบที่ง่ายที่สุดรูปแบบหนึ่งของอัลกอริทึมการแบ่งประเภท [5]

4) เคเนียเรสเนเบอร์ (K-Nearest Neighbor) เป็นการจำแนกกลุ่มข้อมูลที่มีคุณสมบัติใกล้เคียงกันที่สุด K กลุ่ม โดยการคำนวณค่าระยะห่างระหว่างจุด (Euclidean Distance) ซึ่งเป็นการวัดระยะห่างระหว่างข้อมูลสองข้อมูล ถ้าข้อมูลมีความคล้ายคลึงกันมากจะมีระยะห่างน้อย [5] ดังภาพที่ 1-3

5) เพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron) เป็นอัลกอริทึมที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ หรือโครงข่ายประสาทเทียม โดยเป็นการเชื่อมกันของโหนดหลายชั้น ค่าฟังก์ชันของแต่ละโหนดเกิดจากฝึกฝน (Train) โดยส่งค่าย้อนกลับ (Back Propagation) ทำให้ได้



ภาพที่ 1 *ตัวแบบการแบ่งประเภทข้อมูล*

ตัวแบบที่สามารถให้น้ำหนักกับข้อมูลนำเข้า เพื่อให้ได้ประเภทที่ถูกจำแนกตามโหนดปลายทาง [5] ดังภาพที่ 1-4

2.2 เหมืองข้อความ (Text Mining)

เหมืองข้อความ เป็นการใช้แขนงความรู้ในด้านต่าง ๆ เช่น การทำเหมืองข้อมูล (Data Mining) การเรียนรู้ของเครื่อง การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NPL) การค้นคืนสารสนเทศ (Information Retrieval) และการจัดการความรู้ (Knowledge Management) โดยอาศัยกระบวนการรวบรวมเอกสารการวิเคราะห์ข้อความเอกสาร และการนำเสนอผลลัพธ์ให้มีความชัดเจนมากยิ่งขึ้น [8] จากการ ให้คำนิยามเหมืองข้อความของ Mahesh และคณะ [9] Singh และ Tripathi [10] สรุปได้ว่า เหมืองข้อความ หมายถึง การใช้เทคนิคการค้นหาความรู้ใหม่จากข้อมูลประเภทข้อความที่มีปริมาณมาก โดยใช้วิธีการสกัดคำ ค้นหารูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อความเอกสาร เพื่อให้เกิดความหมายและสามารถนำมาใช้ประโยชน์ได้ สำหรับกระบวนการทำเหมืองข้อความเป็นกระบวนการที่คล้ายคลึงกับกระบวนการค้นหาความรู้จากฐานข้อมูล โดยมี 5 ขั้นตอนที่สำคัญดังนี้ [11]

ขั้นตอนที่ 1 การเลือกข้อมูล (Selection) เป็นการระบุแหล่งที่มาของข้อมูลที่นำมาใช้ในการทำเหมืองข้อความ รวมถึงการนำข้อมูลที่ต้องการออกมาจากฐานข้อมูล เพื่อทำการพิจารณาในเบื้องต้นตามขอบเขตที่ต้องการศึกษา

ขั้นตอนที่ 2 การเตรียมข้อมูล (Preprocessing) เป็นกระบวนการที่ทำให้เกิดความมั่นใจในคุณภาพของข้อมูลที่นำมาใช้วิเคราะห์ว่ามีความถูกต้อง โดยการนำข้อมูลที่ไม่ถูกต้องออก หรือเป็นขั้นตอนที่อาจต้องแก้ไขข้อมูลก่อนนำไปใช้งาน

ขั้นตอนที่ 3 การจัดทำดัชนีข้อมูล (Indexing) เป็นการจัดข้อมูลให้เหมาะสมและตรงกับรูปแบบที่สามารถประมวลผลต่อไปได้ เช่น ตัดบางคอลัมน์ที่ไม่จำเป็นออก

ขั้นตอนที่ 4 การทำเหมืองข้อมูล (Data Mining) เป็นขั้นตอนประมวลผล โดยใช้อัลกอลิทึมต่าง ๆ เพื่อแก้ไขปัญหาหรือหารูปแบบของข้อมูล การจัดหมวดหมู่ (Classification) การค้นหากฎความสัมพันธ์ (Association Rule Discovery) การแบ่งกลุ่มข้อมูล (Data Clustering) และการสร้างจินตทัศน์ (Visualuzation)

ขั้นตอนที่ 5 การแปลผลและการประเมินผล (Interpretation/Evaluation) เป็นขั้นตอนการแปลความหมาย การตีความและการประเมินผลลัพธ์ว่ามีความเหมาะสมหรือตรงกับวัตถุประสงค์ที่ต้องการหรือไม่ ซึ่งควรนำเสนอผลการวิเคราะห์ในรูปแบบที่ผู้ใช้งานสามารถเข้าใจได้ง่าย

2.3 การประมวลผลข้อความ (Text Preprocessing)

เป็นการจำแนกเอกสารภาษาไทย ประกอบด้วยขั้นตอนการสกัดคุณลักษณะด้วยการตัดคำเพื่อให้ได้คุณลักษณะของเอกสาร จากนั้นกำจัดคำหยุดและหารากศัพท์จากฐานข้อมูลภาษาไทยที่กำหนดขึ้น หลังจากนั้นทำการให้ค่าน้ำหนักดัชนีของคำในเอกสาร (Term Frequency Inverse Document Frequency) แล้วลดขนาดคุณลักษณะของเอกสาร และทำการเรียนรู้แบบมีผู้สอน (Supervised Learning) [5], [6] มีรายละเอียดดังนี้

2.3.1 การตัดคำ (Word Segmentation) เป็นขั้นตอนในการประมวลผลเพื่อจำแนกหมวดหมู่ข้อความ เพื่อให้การจำแนกมีประสิทธิภาพ สำหรับการตัดคำในภาษาไทยยังพบปัญหาในการตัดคำเนื่องจากลักษณะของภาษาไทยมีการเขียนติดต่อกับแบบไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำที่ชัดเจน ไม่มีช่องว่างคั่นที่แสดงให้เห็นถึงขอบเขตของแต่ละคำ จึงมีผู้คิดค้นวิธีการตัดคำภาษาไทยโดยมีวิธีการ [12], [13] ดังนี้

1) การตัดคำโดยใช้กฎ (Rule-Based Approach) เป็นการตัดคำโดยตรวจสอบจากกฎเกณฑ์ทางอักขระวิธีที่อาศัยหลักไวยากรณ์ภาษาไทย เริ่มจากการตัดพยางค์ เนื่องจากพยางค์มีรูปแบบที่แน่นอนมากกว่าคำ จากนั้นนำพยางค์มาเป็นเกณฑ์ในการกำหนดขอบเขตคำ ข้อดี คือ การทำงานมีความรวดเร็วและใช้ทรัพยากรในการประมวลผลน้อย แต่มีข้อจำกัดคือ ผลของการตัดคำอาจได้เป็นกลุ่มคำที่สามารถตัดคำออกไปได้อีก

2) การตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) ใช้การเปรียบเทียบคำกับคำที่จัดเก็บในพจนานุกรมร่วมกับการใช้กฎในการตัดคำ คำทั้งหมดจะถูกจัดเก็บไว้ในพจนานุกรม แล้วนำข้อความที่ป้อนเข้ามาไปเปรียบเทียบสายอักขระกับคำในพจนานุกรม

3) การตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) เป็นการนำหลักทางสถิติและกลไกการเรียนรู้มาใช้ในการประมวลผลทางภาษา มี 2 วิธี คือ วิธีความน่าจะเป็นโดยใช้ตัวแบบไตรแกรมกำกับหน้าที่ของคำ ในการหารูปแบบของการตัดคำและลำดับหมวดหมู่ ซึ่งต้องเตรียมคลังคำที่มีการตัดคำและกำกับหน้าที่ของคำไว้ล่วงหน้าและวิธีคุณลักษณะของคำที่ช่วยแก้ไขความผิดพลาดของการตัดคำที่อาศัยความน่าจะเป็น โดยนำคุณลักษณะคำที่มีความกำกวมมาช่วยเลือกการตัดคำที่ถูกต้องให้ได้จำนวนคำที่ไม่พบในพจนานุกรมน้อยที่สุด โดยการสร้างตัวแบบกลไกการเรียนรู้

งานวิจัยนี้ใช้คลาสลองเล็กซ์โต (LongLexTo) ของโปรแกรมเล็กซ์โต (Thai Lexeme Tokenizer: LexTo) ซึ่งเป็นโปรแกรมที่ใช้ตัดคำโดยใช้พจนานุกรมคำศัพท์เทียบกับคำที่ยาวที่สุดที่พบในพจนานุกรม (Longest Matching) ที่ถูกพัฒนาโดยศูนย์เทคโนโลยีอิเล็กทรอนิกส์และคอมพิวเตอร์แห่งชาติ (เนคเทค) [14]

2.3.2 การกำจัดคำหยุด (Stop-Word Removal) เป็นการนำคำที่ไม่มีนัยสำคัญออก โดยที่ความหมายของคำหรือข้อความไม่เปลี่ยนแปลง คำหยุดจะปรากฏอยู่ในข้อความทุกข้อความ มีความถี่สูง ถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการจำแนกหมวดหมู่ การกำจัดคำหยุดเป็นกระบวนการที่ช่วยให้ขนาดของดัชนีลดลง อีกทั้งลดขนาดพื้นที่และเวลาในการประมวลผลด้วย

2.3.3 การหารากศัพท์ (Stem) เป็นการหารูปแบบเดิมของคำหรือหาคำที่มีความหมายคล้ายกันมารวมเป็นคำเดียวกัน แต่ยังมีข้อจำกัด คือ ยังไม่มีอัลกอริทึมสำหรับการหารากศัพท์ภาษาไทยได้ครอบคลุม เนื่องจากไวยากรณ์ภาษาไทยมีความซับซ้อน ต้องอาศัยผู้เชี่ยวชาญด้านภาษาไทยเป็นผู้จัดทำคลังคำสำหรับรากศัพท์ภาษาไทย สำหรับงานวิจัยนี้ไม่ได้นำการหารากศัพท์มาพิจารณาเป็นการตัดคำและดึงคำมาใช้เฉพาะโดเมนที่สนใจเท่านั้น

2.3.4 การสร้างดัชนีคำสำคัญ (Indexing) เป็นกระบวนการแปลงเอกสารที่เป็นภาษาธรรมชาติให้คอมพิวเตอร์สามารถเข้าใจและประมวลผลได้ การสร้างตัวแทนเอกสารโดยทั่วไปอยู่ในรูปแบบเวกเตอร์ค่าน้ำหนักคำซึ่งเป็นค่าคุณลักษณะของเอกสาร สำหรับงานวิจัยนี้สร้างดัชนีเอกสารโดยใช้ค่าน้ำหนักรูปแบบคำเดี่ยว

2.4 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องกับการจัดหมวดหมู่ภาษาไทย เช่น ราชวิทย์ และคณะ [12] ได้เสนอ การจำแนกกลุ่มคำถาม

อัตโนมัติบนกระดานสนทนาโดยใช้เทคนิคเหมืองข้อความเพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกข้อความ 3 เทคนิควิธี ได้แก่ การหาเพื่อนบ้านใกล้สุด ต้นไม้ตันสินใจการเรียนรู้เบย์อย่างง่าย ผลการทดลองพบว่า เทคนิคการหาเพื่อนบ้านใกล้สุด มีประสิทธิภาพในการจำแนกกลุ่มคำถามได้ดีที่สุด มีค่าความถูกต้องเท่ากับ 89% ค่าความเที่ยงเท่ากับ 90% และค่าความระลึก 89% ส่วนงานวิจัยของนิเวศและคณะ [15] ได้เสนอการพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ ซึ่งได้ลดคุณลักษณะข้อมูลโดยใช้ค่าเกนความรู้ (Information Gain) และเทคนิคการเรียนรู้ของเครื่อง ผลการทดลอง พบว่า อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนให้ประสิทธิภาพสูงสุด คือ 94.3% ด้านการลดขนาดคุณลักษณะสามารถลดลงได้ 90% โดยไม่ส่งผลต่อประสิทธิภาพการจัดหมวดหมู่ ในงานวิจัยของ Choochart และคณะ [16] เสนอการให้ค่าน้ำหนักดัชนีเอกสารด้วยวิธีการให้น้ำหนักของคำของหนังสือพิมพ์ไทยรัฐบนเว็บไซต์ โดยใช้การตัดคำด้วยคำเดี่ยวและใช้อัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีนเพื่อจำแนกกลุ่มข้อความ ซึ่งวิธีการให้น้ำหนักคำเป็นวิธีที่นำมาใช้สร้างตัวแบบเอกสารที่มีประสิทธิภาพในการจำแนกได้ดี

สำหรับงานวิจัยนี้ผู้วิจัยนำงานวิจัยของ [12], [15] และ [16] มาใช้เป็นแนวทางในการศึกษาและประยุกต์ใช้วิธีการสกัดข้อความและเปรียบเทียบวิธีการเรียนรู้ของเครื่องเพื่อจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะจากข้อมูลข้อร้องเรียนผ่านกระดานสนทนา

3. วิธีดำเนินการวิจัย

3.1 การเก็บรวบรวมข้อมูล

งานวิจัยนี้เก็บรวบรวมข้อมูลความคิดเห็นหรือข้อร้องเรียนและรายละเอียดการใช้บริการจากกระดานสนทนาบนเว็บไซต์ขององค์การขนส่งมวลชนกรุงเทพ (http://www.bmta.co.th/?q=th/forum) มีข้อมูลเดือนมกราคม 2558 ถึงเดือนธันวาคม 2559 จำนวน 1,314 ข้อความ ดังตารางที่ 1

3.2 การประมวลผลข้อความ

เนื่องจากข้อความที่ได้อยู่ในรูปแบบภาษาธรรมชาติจึงต้องแปลงข้อความให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเรียนรู้ได้ เพื่อสร้างตัวแทนเนื้อหาของเอกสาร ซึ่งกระบวนการประมวลผลข้อความสำหรับงานวิจัยนี้มีขั้นตอนดังภาพที่ 2

ตารางที่ 1 *ตัวอย่างข้อความคิดเห็นหรือข้อร้องเรียน*

ลำดับ	ข้อความตัวอย่าง
1	สาย 29 หมายเลขทะเบียน 15-8570 กรุงเทพ เมื่อเวลา 8.20 น. จอดรถนอกป้ายและ ไม่จอดรถเมื่อถึงป้ายประจำทาง
2	รถสาย 8 39-130(11-5848) ขับรถเร็วมาก ปาดหน้ารถคันอื่น เวลาจอดรถส่งผู้โดยสารดูรีบ ๆ
3	รถร่วมสาย 1 สาย 35 ขับรถไม่สุภาพสุดสุด ขับกระชาก เบรคทีหัวทิ่มหัวตำ เด็กเล็กคนแก่ หกล้มระเนระนาดเป็นประจำ แถมชอบจอดขวางจอดคา รอแย่งผู้โดยสารแถวหน้าโรบินสัน
...	...
1314	รถเมลล์ 558 ขาดระยะ 1.20 นาทีโดยที่ การจราจรก็ไม่ติด รถขาดระยะจากเส้นบางนาตราดวิ่งไปพระราม 2 รอรถ ตั้งแต่ 19.45 จนถึง 21.08 รถยังไม่ผ่านสักคันไม่ทราบว่า มีนโยบาย ปรับปรุงแก้ไขยังไงครับ

ภาพที่ 2 *กระบวนการประมวลผลข้อความ*

ขั้นตอนการประมวลผลข้อความดังภาพที่ 2 เริ่มจากนำข้อร้องเรียนเกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะจากเว็บไซต์องค์การขนส่งมวลชนกรุงเทพ มาแปลงเป็นไฟล์ข้อความที่มีนามสกุล .txt จากนั้นใช้คลาสลองเล็กซ์โตของโปรแกรมเล็กซ์โตมาปรับปรุงโปรแกรมเพิ่มเติมเพื่อใช้ในการตัดคำข้อความภาษาไทยสำหรับโดเมนที่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะ

ในการตัดคำจากข้อความภาษาไทย และการกำจัดคำหยุด เป็นกระบวนการที่ต้องอาศัยพจนานุกรมคำศัพท์มาช่วยในกระบวนการตัดคำ สำหรับงานวิจัยนี้ผู้วิจัยใช้

พจนานุกรมคำศัพท์เล็กซิตรอน (Lexitron) ของเนคเทค ในการตัดคำและได้เพิ่มคำศัพท์ที่เกี่ยวข้องกับพฤติกรรมการขับรถโดยสารสาธารณะในพจนานุกรมคำศัพท์ เล็กซิตรอน เพื่อให้โปรแกรมสามารถตัดคำได้ตรงตามคำศัพท์ที่ผู้วิจัยกำหนด จากนั้นใช้คลาสลองเล็กซ์โตซึ่งเป็นคลาสที่ใช้หลักการตัดคำจากพจนานุกรมคำศัพท์เทียบกับคำที่ยาวที่สุดที่พบในพจนานุกรม โดยมีหลักการทำงาน คือ เมื่อมีการส่งข้อความเข้าในรูปแบบไฟล์ข้อความ .txt โปรแกรมจะทำการอ่านคำทั้งหมดแล้วนำไปเก็บไว้ในโปรแกรม แล้วนำคำเหล่านั้นไปเทียบกับคำที่ยาวที่สุดที่พบในพจนานุกรมของโปรแกรม เมื่อพบแล้วจึงทำการตัดคำนั้นออกมา แล้วบันทึกคำที่ถูกตัดเป็นไฟล์เอกสารเว็บเพจที่มีนามสกุล .html สำหรับนำไปใช้งานต่อไป

งานวิจัยนี้จำแนกคลาสข้อมูลพฤติกรรมการขับรถโดยสารสาธารณะตามลักษณะของคำถามที่พบบ่อยในกระดาษสนทนา โดยทำการตัดคำจากข้อร้องเรียน แล้วนำมาหาความถี่ของคำที่พบมากที่สุดในกระดานสนทนาจากนั้นให้ผู้เชี่ยวชาญทางด้านระบบขนส่งและการจราจรอัจฉริยะจำนวน 3 ท่าน ทำการประเมินเพื่อจัดกลุ่มคำที่เกี่ยวข้องกับพฤติกรรมการขับขี่ที่อาจก่อให้เกิดอันตรายและคุณภาพในการใช้บริการ จำนวน 4 คลาส ได้แก่ 1) คลาสการจอดรับส่งผู้โดยสาร 2) คลาสลักษณะการขับขี่ 3) คลาสการเดินรถ และ 4) คลาสคุณภาพการบริการ มีคำสำคัญรวมทั้งหมด 115 คำ จากนั้นผู้วิจัยจัดเก็บคำสำคัญในลักษณะของถุงคำ ดังตารางที่ 2

เมื่อได้คำที่เกี่ยวข้องกับพฤติกรรมการขับรถโดยสารสาธารณะแล้ว ขั้นตอนต่อไปคือการสร้างดัชนีคำสำคัญ (Indexing) เป็นการสร้างตัวแทนเอกสาร (Document Representation) ให้อยู่ในรูปแบบของเวกเตอร์ค่าน้ำหนักของคำ ซึ่งเป็นการคำนวณหาค่าคุณลักษณะของเอกสารหรือการหาค่าน้ำหนักของคำ (Term Weighting) ในงานวิจัยนี้ใช้การสร้างดัชนีเอกสารโดยการหาค่าน้ำหนักรูปแบบคำเดี่ยว (Single Word) โดยใช้วิธีการให้น้ำหนักคำซึ่งเป็นวิธีที่ใช้ประเมินความสำคัญของคำหรือคุณลักษณะนั้น [17] โดยค่าน้ำหนักของคำหาได้จากการเปรียบเทียบคำถามกับคำสำคัญในถุงคำ แล้วนับจำนวนคำสำคัญที่พบในคำถาม เพื่อนำความถี่มาคำนวณหาค่าน้ำหนักของคำหรือคุณลักษณะ ดังสมการที่ 1 และ 2

ตารางที่ 2 *การจัดเก็บคำสำคัญในลักษณะของถุงคำ*

คลาสที่ 1 การจอดรับส่งผู้โดยสาร (28 คำ)			
จอดก่อนป้าย	จอดนอกป้าย	ไม่ตรงป้าย	นอกป้าย
ไม่เข้าป้าย	ไม่จอด	ไม่จอดป้าย	ไม่จอดรับ
ไม่ยอมจอด	ไม่รับ	...	จอดเลยป้าย
คลาสที่ 2 ลักษณะการขับขี่ (43 คำ)			
ขับรถเร็ว	ขับเบียด	ไล่รถคันหน้า	ขวางทาง
ขับรถอันตราย	ขับรถไว	ขับหวาดเสียว	ขับเร็ว
ตัดหน้า	เบรคกะทันหัน	...	ขับกระชาก
คลาสที่ 3 การเดินรถ (16 คำ)			
ขาดช่วง	ขาดระยะ	เว้นระยะ	นานมาก
รอนาน	รอรถ	มาช้า	ไม่มีรถ
ออกไม่เป็นเวลา	ปล่อยรถ	...	หมดระยะ
คลาสที่ 4 คุณภาพการบริการ (28 คำ)			
พูดจาไม่สุภาพ	ค่าโดยสาร	ตั๋วผิด	ไม่พอใจ
เสียมารยาท	พูดไม่ดี	มารยาทแย่	ทอนเงินผิด
สูบบุหรี่	ประตูหนีบ	...	ไม่ฉีกตั๋ว

$$W_{(f,\,d)} = TF_{(f,\,d)} \times IDF_{(f)} \quad (1)$$

$$IDF_{(f)} = \log \frac{|D|}{|DF_{(f)}|} \quad (2)$$

โดยที่

$W_{(f,\,d)}$ คือ ค่าน้ำหนักของคุณลักษณะ (*f*) ที่ปรากฏในเอกสาร (d)

$TF_{(f,\,d)}$ คือ ความถี่ของคุณลักษณะ (*f*) ที่ปรากฏในเอกสาร (d)

$|D|$ คือ จำนวนเอกสารทั้งหมด

$|DF_{(f)}|$ คือ จำนวนของเอกสารทั้งหมดที่มีคุณลักษณะ (*f*) ปรากฏอยู่

จากนั้นนำค่าน้ำหนักของคำสำคัญมาจัดให้อยู่ในรูปแบบเวกเตอร์เอกสาร (Word Vector) ดังตารางที่ 3

ตารางที่ 3 *เวกเตอร์เอกสาร*

Question	W_1	W_2	...	$W_{j=115}$	Class
Q_1	F_{11}	*F12*	...	F_{1j}	Class1
Q_2	F_{21}	*F22*	...	F_{2j}	Class2
Q_3	F_{31}	*F32*	...	F_{3j}	Class3
...	...	...	...	...	...
$Q_{i=1314}$	F_{i1}	F_{j2}	...	F_{ij}	Class4

3.3 สร้างตัวแบบในการจำแนกกลุ่มข้อความ

นำข้อมูลค่าน้ำหนักของคำสำคัญจากตารางที่ 3 มาสร้างเป็นไฟล์เอกสาร .arff ซึ่งเป็นคุณลักษณะของข้อมูลที่ใช้ในการจำแนกคลาสข้อมูลพฤติกรรมการขับรถโดยสารสาธารณะ ดังแสดงในตารางที่ 4 จากนั้นใช้โปรแกรม Weka (Wailato Environment for Knowledge Analysis) เพื่อวิเคราะห์และจำแนกข้อมูล (Classification) [5] สำหรับการสร้างและทดสอบตัวแบบใช้เทคนิควิธีตรวจสอบแบบไขว้กัน 10 ชุด (10-Fold Cross Validation) โดยแบ่งข้อมูลออกเป็น 10 ชุด ข้อมูลชุดที่ 1 ใช้เป็นชุดข้อมูลทดสอบ (Testing Dataset) และอีก 9 ชุดที่เหลือใช้เป็นชุดข้อมูลการเรียนรู้ (Training Dataset) จากนั้นเปลี่ยนชุดข้อมูลถัดไปเป็นชุดข้อมูลทดสอบส่วนที่เหลือเป็นชุดข้อมูลการเรียนรู้ ทำวนไปจนครบทั้ง 10 ชุดข้อมูล วิธีการนี้เป็นวิธีที่นิยมเพื่อใช้ลดความผิดพลาดของผลลัพธ์จากการสุ่มเลือกชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ [5]

ตารางที่ 4 *คุณลักษณะของข้อมูลที่ใช้ในการจำแนก*

ตัวแปร	ชนิด	คำอธิบาย
W_1	numeric	ค่าน้ำหนักของคำสำคัญที่ 1 (จอดก่อนป้าย)
W_2	numeric	ค่าน้ำหนักของคำสำคัญที่ 2 (จอดนอกป้าย)
W_3	numeric	ค่าน้ำหนักของคำสำคัญที่ 3 (ไม่ตรงป้าย)
...	...	
W_{115}	numeric	ค่าน้ำหนักของคำสำคัญที่ 115 (ไม่ฉีกตั๋ว)
Class	numeric	Class1, Class2, Class3, Class4

3.4 การประเมินตัวแบบ

วัดประสิทธิภาพการจำแนกคลาสข้อมูลพฤติกรรมขับรถโดยสารสาธารณะด้วยค่าความแม่นตรง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าเอฟเมเซอร์ (F-measure) ดังสมการที่ 3, 4, 5 และ 6

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

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

$$F\text{-}measure = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (6)$$

FP = คำที่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะที่ไม่อยู่ในคลาส C_j และตัวแบบทำนายว่าอยู่ในคลาส C_j

FN = คำที่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะที่อยู่ในคลาส C_j และตัวแบบทำนายว่าไม่อยู่ในคลาส C_j

TN = คำที่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะที่ไม่อยู่ในคลาส C_j และตัวแบบทำนายว่าไม่อยู่ในคลาส C_j

C_j = คลาสของคำที่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะจำนวน 4 คลาส เมื่อ $1 \leq j \leq 4$

4. ผลการทดลองและการอภิปรายผล

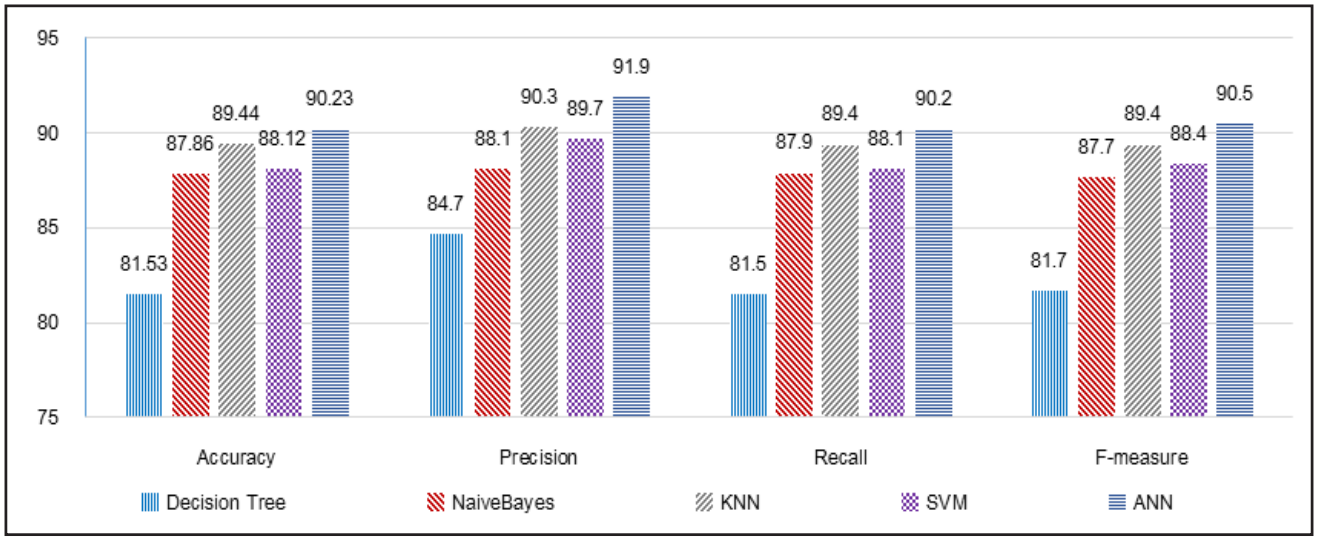
การทดลองนี้ใช้โปรแกรม Weka 3.6 สำหรับสร้างตัวแบบข้อมูลสำหรับการเรียนรู้และทดสอบ มีจำนวนทั้งสิ้น 1,314 เรคคอร์ด ประกอบด้วยคุณลักษณะ ดังนี้ 1) ค่าน้ำหนักคุณลักษณะของคำสำคัญแต่ละคำจำนวน 115 คุณลักษณะ และ 2) คลาสกลุ่มคำที่เกี่ยวกับพฤติกรรมเสี่ยงของผู้ขับขี่ที่อาจก่อให้เกิดอันตรายและคุณภาพในการใช้บริการจำนวน 4 คลาส มีจำนวนข้อมูลในแต่ละคลาสดังนี้ คลาสที่ 1 มีจำนวน 330 เรคคอร์ด คลาสที่ 2 มีจำนวน 327 เรคคอร์ด คลาสที่ 3 มีจำนวน 332 เรคคอร์ด และคลาสที่ 4 มีจำนวน 325 เรคคอร์ด ผลการจำแนกข้อความพฤติกรรมการขับรถโดยสารสาธารณะดังภาพที่ 3 มีรายละเอียด ดังนี้

4.1 ผลการจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ

การเรียนรู้และทดสอบเพื่อสร้างตัวแบบต้นไม้ตัดสินใจใช้อัลกอลิทึม J48 ได้กำหนดพารามิเตอร์ค่าความเชื่อมั่น (Confidence Factor) เท่ากับ 0.25 จำนวนของข้อมูลที่น้อยที่สุดต่อใบ (Leaf) เท่ากับ 2 ผลการเรียนรู้และทดสอบพบว่า ขนาดของต้นไม้ (Size of the Tree) มีค่าเท่ากับ 63 โหนด มีจำนวนใบ (Number of Leaves) เท่ากับ 32 ใบ ค่าความแม่นตรงมีค่าเท่ากับ 81.53% เมื่อพิจารณาในแต่ละคลาส พบว่า ค่าความแม่นยำมีค่าระหว่าง 62-94.4% ค่าความระลึกมีค่าระหว่าง 56.6-96.2% ค่าเอฟเมเซอร์มีค่าระหว่าง 67.4-88.2%

4.2 ผลการจำแนกข้อมูลด้วยนาอีฟเบย์

ผลการเรียนรู้และทดสอบของตัวแบบนาอีฟเบย์ พบว่าค่าความแม่นตรงมีค่าเท่ากับ 87.86% เมื่อพิจารณาในแต่ละคลาส พบว่า ค่าความแม่นยำมีค่าระหว่าง 76.4-90.7% ค่าความระลึกมีค่าระหว่าง 64.2-95.9% ค่าเอฟเมเซอร์มีค่าระหว่าง 73.9-94.9%



ภาพที่ 3 เปรียบเทียบค่าความแม่นตรง ความแม่นยำ ค่าความระลึกและค่าเอฟเมเซอร์ของแต่ละตัวแบบ

4.3 ผลการจำแนกข้อมูลด้วยเพื่อนบ้านใกล้ที่สุด

การเรียนรู้และทดสอบเพื่อสร้างตัวแบบเพื่อนบ้านที่ใกล้ที่สุดได้กำหนดพารามิเตอร์จำนวนเพื่อนบ้าน (k) เท่ากับ 1 พบว่า ค่าความแม่นตรงมีค่าเท่ากับ 89.44% เมื่อพิจารณาในแต่ละคลาส พบว่า ค่าความแม่นยำมีค่าระหว่าง 82.1-95.9% ค่าความระลึกมีค่าระหว่าง 75.5-95.9% ค่าเอฟเมเซอร์มีค่าระหว่าง 84.2 - 95.9%

4.4 ผลการจำแนกข้อมูลด้วยซัพพอร์ตเวกเตอร์แมชชีน

การเรียนรู้และทดสอบเพื่อสร้างตัวแบบซัพพอร์ตเวกเตอร์แมชชีนใช้อัลกอลิทึม Sequential Minimal Optimization (SMO) ได้กำหนดพารามิเตอร์ Filter Type เท่ากับ Normalize Training Data และใช้เคอร์เนล Polynomial เป็นเคอร์เนลพื้นฐานในการจำแนกข้อมูล ผลการจำแนกพบว่า ค่าความแม่นตรงมีค่าเท่ากับ 88.12% เมื่อพิจารณาในแต่ละคลาส พบว่า ค่าความแม่นยำมีค่าระหว่าง 70.2 - 97.1% ค่าความระลึกมีค่าระหว่าง 71.1-93.6% ค่าเอฟเมเซอร์มีค่าระหว่าง 78.4 - 93.2%

4.5 ผลการจำแนกข้อมูลด้วยโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น

การเรียนรู้และทดสอบเพื่อสร้างตัวแบบโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น ได้กำหนดพารามิเตอร์อัตราการเรียนรู้ (Learning Rate) เท่ากับ 0.3 ค่าความแกว่ง (Momentum) เท่ากับ 0.2 กำหนดจำนวนชั้นซ่อน (Hidden Layer) ตั้งแต่ 2-5 ชั้น เพื่อเลือกจำนวนชั้นซ่อนที่เหมาะสม และจำนวนครั้งที่ทำการฝึกสอน (Training Time) เท่ากับ 500

ผลการจำแนก พบว่า โครงสร้างตัวแบบ 115-58-4 (จำนวนโหนดชั้นนำเข้า-ชั้นซ่อน-ชั้นนำออก) มีค่าความแม่นตรงเท่ากับ 90.23% เมื่อพิจารณาในแต่ละคลาส พบว่า ค่าความแม่นยำมีค่าระหว่าง 72.4-98.5% ค่าความระลึกมีค่าระหว่าง 75.5-97.4% ค่าเอฟเมเซอร์มีค่าระหว่าง 83.1 - 95.3%

เมื่อพิจารณาค่าความแม่นตรง ความแม่นยำ ค่าความระลึกและเอฟเมเซอร์ในการจำแนกข้อมูลของตัวแบบทั้ง 5 ตัวแบบ พบว่า ตัวแบบโครงข่ายประสาทเทียม มีค่าความแม่นตรง ค่าความแม่นยำและค่าความระลึกโดยรวมของทุกคลาสสูงกว่าตัวแบบอื่นๆ ที่ใช้ในการทดลองในครั้งนี้ ดังนั้นการสร้างตัวแบบเพื่อจำแนกข้อมูลพฤติกรรมการขับขี่รถโดยสารสาธารณะด้วยเทคนิควิธีโครงข่ายประสาทเทียมจึงเป็นขั้นตอนวิธีที่เหมาะสมที่สามารถนำมาประยุกต์ใช้งานได้อย่างมีประสิทธิภาพ

5. สรุป

งานวิจัยนี้เป็นงานวิจัยเชิงประยุกต์ที่เสนอวิธีการจำแนกพฤติกรรมการขับขี่รถโดยสารสาธารณะจากข้อมูลข้อร้องเรียนของผู้โดยสารผ่านเว็บไซต์สนทนาโดยใช้วิธีการสกัดข้อความและใช้คลาสลองเล็กซ์โตของเนคเทค โดยใช้หลักการตัดคำจากพจนานุกรมคำศัพท์เทียบกับคำที่ยาวที่สุดที่พบในพจนานุกรมมาปรับปรุงโปรแกรมเพิ่มเติมเพื่อใช้ตัดคำข้อความภาษาไทยสำหรับโดเมนข้อมูลพฤติกรรมการขับขี่ที่อาจก่อให้เกิดอันตรายและคุณภาพในการบริการ จากนั้นสร้างดัชนีเอกสารโดยการหาค่าน้ำหนักรูปแบบคำเดี่ยว

จากนั้นนำมาจัดให้อยู่ในรูปแบบของเวกเตอร์เอกสาร สำหรับใช้จำแนกข้อมูลข้อร้องเรียนจำนวน 4 คลาส ได้แก่ 1) คลาสการจอดรับส่งผู้โดยสาร 2) คลาสลักษณะการขับขี่ 3) คลาสการเดินรถ และ 4) คลาสคุณภาพการบริการ จากนั้นใช้ขั้นตอนวิธีต้นไม้ตัดสินใจ นาอีฟเบย์ การหาเพื่อนบ้านใกล้ที่สุด ซัพพอร์ตเวกเตอร์แมชชีนและโครงข่ายประสาทเทียมเพื่อสร้างตัวแบบในการจำแนกข้อมูล ผลการทดลองพบว่าตัวแบบโครงข่ายประสาทเทียม มีค่าความแม่นตรง ค่าความแม่นยำ ค่าความระลึกและค่าเอฟเมเซอร์โดยรวมของทุกคลาสสูงกว่าตัวแบบอื่นๆ ที่นำมาใช้ในการจำแนกข้อมูลในครั้งนี้

งานวิจัยในอนาคต ผู้วิจัยจะดำเนินการเพิ่มคำสำคัญที่เกี่ยวกับพฤติกรรมการขับรถโดยสารสาธารณะให้ครอบคลุมที่สามารถใช้กับลักษณะข้อความหลายรูปแบบ และพิจารณาเทคนิควิธีการเรียนรู้ของเครื่องแบบอื่นๆ เพิ่มเติมที่สามารถนำมาสร้างตัวแบบเพื่อใช้ในการจำแนกข้อความให้มีประสิทธิภาพมากขึ้นโดยไม่ส่งผลต่อการจำแนกหมวดหมู่พอดีเกินไป (Over Fitting) กับข้อมูลชุดฝึกและชุดทดสอบที่อาจทำให้ไม่เหมาะกับข้อมูลที่ต้องการทำนายจริง

6. เอกสารอ้างอิง

[1] I. Han and K. S. Yang. “Characteristic analysis for cognition of dangerous driving using automobile black boxes.” *International journal of automotive technology*, Vol. 10, No. 5, pp. 597-605, 2009.

[2] Logex. Center for Logistic Excellence. “The Study of Routed-Bus Administration Efficiecy Improvement.” *Division of Land Transportation*, Bangkok, 2009.

[3] W. Wongvilaisakul. “Text Mining and its Applications.” *Panyapiwat Journal*, Vol. 4, pp. 157-165, 2013.

[4] E. Alpaydin. “Introduction to machine learning.” *MIT press,* 2009.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal. “Data Mining: Practical machine learning tools and techniques.” *Morgan Kaufmann*, 2016.

[6] M. Mohri, A. Rostamizadeh, and A. Talwalkar. “Foundations of machine learning.” *MIT press*, 2012.

[7] A. Ben-Hur and J. Weston. “A user’s guide to support vector machines.” *In Data mining techniques for the life sciences*, pp. 223-239, 2010.

[8] R. Feldman and J. Sanger. “The text mining handbook: advanced approaches in analyzing unstructured data.” *Cambridge University Press*, 2007.

[9] T. R. Mahesh, M. B. Suresh, and M. Vinayababu. “Text mining: advancements, challenges and future directions.” *International journal of reviews in Computing*, Vol. 3, pp. 61-65, 2010.

[10] V. Singh and S. P. Tripathi. “Text Mining Approaches to Extract Interesting Association Rules from Text Documents.” *International Journal of Computer Science Issues*, Vol. 9, No. 3, pp. 545-552, 2012.

[11] H. Cherfi, A. Napoli, and Y. Toussaint. “Towards a text mining methodology using association rule extraction.” *Soft Computing*, Vol. 10, No. 5, pp. 431-441, 2006.

[12] R. Tipsena, C. Jareanpon, and G. Somprasertsri. “Automatic Question Classification on Webboard Using Text Mining Techniques.” *MSU Journal of Science and Technology*, Vol. 33, No. 5, pp. 493-502, 2013.

[13] N. Jirawichitchai. “Automated Thai Document Classification Model.” *Journal of Industrial Technology*, Vol. 9, No. 1, pp. 141-149, 2013.

[14] National Electronics and Computer Technology Center. *LexTo Thai Lexeme Tokenizer*. Available Online at http://www.sansarn.com/lexto/, accessed on 7 July 2018.

[15] N. Chirawichitchai, P. Sanguansat, and P. Meesad. “Developing and Effective Automatic Thai Document Categorization.” *NIDA Development Journal*, Vol. 51, No. 3, pp 187-205, 2011.

[16] H. Choochart, W. Jitkrittum, C. Sangkeettrakarn, and C. Damrongrat. “Implementing News Article Category Browsing Based on Text Categorization Technique.” *International Conference on Web Intelligence and Intelligent Agent Technology*, pp. 143-14, 2008.

[17] L. Jing, H. Hou-Kuan, and S. Hong-Bo. “Improved feature selection approach TFIDF in text mining.” *Machine Learning and Cybernetics, 2002. Proceedings*. *International Conference*, *IEEE*, Vol. 2, 2002.