วารสารสารสนเทศศาสตร์
Journal of Information Science
Journal of the Department of Information Science, Khon Kaen University, Thailand

iSchool KKU

*Research Articles*

# การจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก

# The Public Bus Complaint Classification by Deep Learning Model

**จักรินทร์ สันติรัตนภักดี [1,*], ศุภกฤษฎิ์ นิวัฒนากูล [2,3]**

**Chakkarin Santirattanaphakdi[1,*] , Suphakit Niwattanakul[2,3]**

[1] สาขาวิชาระบบสารสนเทศคอมพิวเตอร์ คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล; Department of Computer Information System, Faculty of Business Administrator, Vongchavalitkul University, Thailand.

[2] สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี; School of Information Technology, Institute of Social Technology, Suranaree University of Technology, Thailand.

[3] โครงการจัดรูปแบบการบริหารวิชาการด้านเทคโนโลยีดิจิทัลรูปแบบใหม่ มหาวิทยาลัยเทคโนโลยีสุรนารี; A New Paradigm in the Digital Technology Academic Administrator Project (DIGITECH), Suranaree University of Technology.

* Corresponding author email: chakkarin_san@vu.ac.th

**บทคัดย่อ**

**วัตถุประสงค์:** เพื่อออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก และเพื่อประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

**วิธีการศึกษา:** ออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก เพื่อสร้างโมเดลการจำแนกแบบไบนารี่ โดยแบ่งคำด้วยอัลกอริทึม deepcut จากนั้นจะแปลงข้อมูลทั้งหมดไปอยู่ในถุงคำ สำหรับนำมาสร้างดัชนีคำศัพท์ที่สามารถคำนวณค่าน้ำหนักคำสำคัญสำหรับการจำแนกข้อความ และลดปัญหาความซ้ำซ้อนของข้อมูลด้วย t-SNE เพื่อนำผลการกระจายตัวของกลุ่มที่มีความคล้ายคลึงกันมาสร้างคลาสการร้องเรียน แบ่งออกเป็น 7 คลาส

**ข้อค้นพบ:** การจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึกด้วยอัลกอริทึม deepcut ฝึกฝนโมเดลด้วยการวิเคราะห์การถดถอยโลจิสติกส์ พบว่า ทุกคลาสมีความถูกต้องมากกว่าร้อยละ 90 อย่างไรก็ดี ค่าความถูกต้องในคลาสผู้ให้บริการ มีความถูกต้องเพียงร้อยละ 83 ตามจำนวนข้อร้องเรียนในคลาสที่มีจำนวนมากที่สุด ส่งผลให้เกิดความหลากหลายของข้อมูลที่แตกต่างกันตามบริบทของผู้ใช้ เมื่อเปรียบเทียบกับอัลกอริทึม fastText พบว่า ค่าความถูกต้องของทั้งสองอัลกอริทึมมีค่าความถูกต้องอยู่ในระดับสูง และเป็นไปในทิศทางเดียวกัน ผลการประเมินผลความถูกต้องจากการติดแท็กปัญหาการให้บริการ พบว่า อยู่ในระดับดีมาก แสดงถึงการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก ให้ผลลัพธ์ความถูกต้องในระดับสูง โดยเฉพาะข้อร้องเรียนแบบ 1 ประเด็นต่อ 1 ข้อร้องเรียน เนื่องจากเหมาะกับคำศัพท์ที่บริบทสามารถกำหนดขอบเขตได้แน่นอน ในทางกลับกัน ปัญหาที่พบส่วนใหญ่ คือการมีคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส

*Received: October 1, 2021 Revised: November 11, 2021 Accepted: November 15, 2021* https://doi.org/10.14456/jiskku.2022.2

**การประยุกต์ใช้จากการศึกษานี้:** ผลลัพธ์จากงานวิจัยจะนำเสนอแนวทางการจำแนกข้อความด้วยการเรียนรู้เชิงลึก และเป็นประโยชน์แก่ผู้รับผิดชอบ เพื่อนำไปปรับปรุงการให้บริการให้เหมาะสมกับความต้องการของผู้ใช้บริการต่อไป

**คำสำคัญ:** การจำแนกข้อความ การเรียนรู้ของเครื่อง การเรียนรู้เชิงลึก รถโดยสารสาธารณะ

## Abstract

**Purpose:** The objectives of this study were to design and develop the public bus complaint classification using the deep learning approach and to assess the complaint classification accuracy.

**Methodology:** This investigation began with the design and development of public bus complaint classification using deep learning approach to create a binary model for dividing words by deep-cut algorithm. After that, all words were transformed into a word bag for word indexing which could calculate keyword weights for classifying statements to mitigate the data duplication problem with t-SNE method in order to make distribution of similar words for formulating 7 classes of public bus complaint classification.

**Findings:** The findings revealed that all classes were found to be accurate at 90%. However, the service provider class which received the largest number of the public bus complaints achieved its accuracy only 83%. This resulted in a variety of different information based on the user context. The comparison between the deep-cut algorithm and the fast-text algorithm showed that they both attained the high accuracy in the same direction. When measuring the accuracy of the service provider problem tagging, it was found that the accuracy reached good level. This indicated that the public bus complaint classification with deep-cut learning approach provided accurate results, especially the accuracy of one complaint per one issue due to the suitability of the issue terms which could be defined by the context. On the contrary, many problems found in the study were caused by duplication of terms in some classes.

**Application of the study:** The study findings are of benefit to responsible individuals for improving their services to meet their user service needs.

**Keyword:** Text classification, Machine learning, Deep learning, Public bus

## 1. บทนำ

การเดินทางด้วยรถโดยสารสาธารณะในพื้นที่กรุงเทพฯ และปริมณฑล ดำเนินการภายใต้การบริหารขององค์การขนส่งมวลชนกรุงเทพ (ขสมก.) จากจำนวนเส้นทางเดินรถโดยสารประจำทางรวมทุกประเภท จำนวน 456 เส้นทาง ครอบคลุมรถองค์กร รถเอกชนร่วมบริการ รถเล็กวิ่งในซอย รถตู้โดยสาร และรถตู้เชื่อมต่อท่าอากาศยานสุวรรณภูมิ (Bangkok Mass Transit Authority, 2019) ด้วยคุณสมบัติการให้บริการที่มีความคล่องตัวสูง สะดวก และสามารถให้บริการได้ทุกจุดตลอดระยะของการเดินทาง ตลอดจนอัตราค่าโดยสารยังสอดรับกับรายได้ของประชากรส่วนใหญ่ ในปี 2560 มีผู้ใช้บริการเฉลี่ย 830,859 คนต่อวัน (Office of Transport and Traffic Policy and Planning, 2018) แต่หากพิจารณาย้อนหลัง พบว่า มีจำนวนผู้ใช้บริการ ลดลงอย่างต่อเนื่อง ซึ่งหนึ่งในปัจจัยหลักมาจากคุณภาพการให้บริการ อันจะเห็นได้จากการแสดงความคิดเห็นหรือรับเรื่องราวร้องเรียนรถองค์การ และรถเอกชนร่วมบริการผ่านเว็บบอร์ดบนเว็บไซต์ http://www.bmta.co.th เป็นอีกหนึ่งช่องทางที่ใช้ในการติดตามตรวจสอบการให้บริการ และเป็นสื่อกลางในการแลกเปลี่ยนความคิดเห็นต่างๆ ที่ผู้ใช้งานสามารถตั้งกระทู้ถามตอบเพื่อแลกเปลี่ยนความคิดเห็นกันได้อย่างอิสระระหว่างผู้ให้บริการและผู้รับบริการ ดังนั้นข้อมูลบนเว็บบอร์ดจึงนับว่ามีประโยชน์ และมีบทบาทในการเพิ่มประสิทธิภาพในการให้บริการ แต่จากจำนวนข้อร้องเรียนรถโดยสารสาธารณะมีเพิ่มมากขึ้น ส่งผลต่อความหลากหลายของข้อความที่แตกต่างกันจากบริบทของผู้ใช้งาน ตลอดจนความผิดพลาดในการใช้ภาษาที่เกิดจากความตั้งใจหรือแม้แต่ไม่ตั้งใจของผู้ใช้ ที่ก่อให้เกิดปัญหาในการตีความหมาย โดยเฉพาะอย่างยิ่งข้อความที่ไม่มีการจำแนกหมวดหมู่ไว้อย่างชัดเจน อาจส่งผลต่อการตอบคำถาม การให้ข้อมูลคืนที่ถูกต้องแก่ผู้ใช้ ตลอดจนการสรุปสถิติการจัดประเภทข้อร้องเรียนนั้นทำได้ยาก ใช้เวลานาน และมีโอกาสเกิดความผิดพลาดสูง เนื่องจากต้องอาศัยการวิเคราะห์เพื่อจำแนกหมวดหมู่ด้วยตนเอง ดังนั้นหากมีระบบการจำแนกข้อร้องเรียนตามหมวดหมู่ที่ต้องการแบบอัตโนมัติจะเป็นแนวทางหนึ่งในการแก้ไขปัญหาดังกล่าว อย่างไรก็ดี ข้อร้องเรียนรถโดยสารสาธารณะ ยังมีปัญหาความถูกต้องของการวิเคราะห์และประมวลผลเนื่องจากความซับซ้อนในการผสมคำของภาษาไทย (Tapsai, Unger, & Meesad, 2021) ที่มีตัวอักษรหลายประเภท การผสมสระและวรรณยุกต์เพื่อสร้างคำ และการที่เขียนต่อเนื่องกันเป็นประโยค โดยไม่มีการเว้นวรรคหรือตัวคั่นใด ๆ อย่างภาษาอังกฤษ แม้ว่าการศึกษาจำนวนมากได้ดำเนินการเกี่ยวกับการแบ่งส่วนคำภาษาไทยด้วยเทคนิค ต่าง ๆ แต่ยังคงพบปัญหาดังนี้ 1) ปัญหาเกี่ยวกับประสิทธิภาพของอัลกอริทึมในการแบ่งคำภาษาไทย อันเนื่องมาจากความครอบคลุมของคลังคำศัพท์ ส่งผลอย่างยิ่งต่อความถูกต้องของผลลัพธ์ในการแบ่งคำภาษาไทย อย่างไรก็ดี หากจะให้คลังคำศัพท์ครอบคลุมครบถ้วนจำเป็นต้องมีคลังคำศัพท์ขนาดใหญ่และยากที่จะระบุได้ว่าเท่าไรจึงจะครอบคลุมครบถ้วน 2) ปัญหาเกี่ยวกับคำที่สะกดผิด ส่งผลให้แบ่งคำภาษาไทยที่ยึดตามพจนานุกรมไม่ถูกต้อง หรือไม่สามารถวิเคราะห์ความหมายได้เลย ข้อผิดพลาดในการสะกดผิดมี 7 ประเภท ได้แก่ ตัวอักษรเกิน เช่น “ไม่จอด” เป็น “ไม่จอดด” ตัวอักษรหายไป เช่น “มารยาท” เป็น “มายาท”, ตัวอักษรซ้ำ เช่น “กด | ออด” หรือ “กดด | ออด”, การพิมพ์ผิด เช่น “ขับรถเร็ว” เป็น “ขัยรถ

เร็ว", ตัวอักษรผิดตำแหน่ง เช่น "สแกน" เป็น "แสกน", คำสแลง เช่น "โดนรถเมล์เท" และอื่น ๆ ที่ส่งผลต่อความผิดพลาดในการประมวลผลภาษาธรรมชาติ และ 3) การประสมคำภาษาไทยที่สร้างขึ้นจากคำพื้นฐานเพื่อใช้ในวัตถุประสงค์ที่แตกต่างกัน เช่น คำศัพท์แสลง (Slang word) และคำทับศัพท์ภาษาอังกฤษ (Transliteration word) ภายใต้รูปแบบการเกิด ระยะเวลา และปริมาณ ดังนั้นจึงไม่สามารถจัดเก็บคำทั้งหมดลงในพจนานุกรมได้ เนื่องจากมีคำประเภทนี้จำนวนมาก ส่งผลต่อประสิทธิภาพของการแบ่งคำภาษาไทย

งานวิจัยชิ้นนี้มุ่งจำแนก (Classification) ข้อร้องเรียนรถโดยสารสาธารณะที่เป็นเทคนิคการสร้างโมเดลหรือตัวจำแนกข้อมูล (Classifier) (Aggarwal, 2015) เพื่อทำนายหมวดหมู่ของข้อมูล (Categories/Class) โดยชุดข้อมูลที่นำเข้าสำหรับสร้างตัวจำแนกข้อมูลเรียกว่าชุดข้อมูลฝึกสอน (Training data) หากในชุดข้อมูลมีแอททริบิวต์หมวดหมู่ข้อมูลสำหรับจำแนกจะเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised learning) ที่สร้างตัวจำแนกข้อมูลจะถูกสอนโดยแอททริบิวต์หมวดหมู่ข้อมูลต่าง ๆ ตรงข้ามกับการเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning หรือ Clustering) ที่จะไม่ทราบถึงหมวดหมู่ของข้อมูล ปัจจุบันมีผู้เสนอเทคนิคต่าง ๆ ในการจำแนกข้อมูลที่ได้รับความนิยม เช่น 1) การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ (Decision tree classifier) เป็นกระบวนการสร้างต้นไม้ขึ้นเพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลระบุอยู่ 2) การจำแนกข้อมูลแบบเบย์เซียน (Bayesian classifier) เป็นการสร้างตัวจำแนกข้อมูลด้วยการประยุกต์ใช้ค่าทางสถิติที่สามารถบ่งบอกถึงความน่าจะเป็นของข้อมูลเรคคอร์ดหนึ่งที่จะอยู่ในหมวดหมู่ของข้อมูลหนึ่ง ๆ โดยประยุกต์ใช้ทฤษฎีของเบย์ (Bayes' Theorem) 3) การจำแนกข้อมูลด้วยฐานกฎ (Rule-based classifier) เป็นโมเดลที่จะแสดงผลด้วยกลุ่มของกฎที่มีลักษณะแบบ IF-THEN โดยแต่ละกฎจะอยู่ในรูปฟอร์ม 4) การจำแนกข้อมูลจากกฎความสัมพันธ์ของข้อมูล (Association rule classifier) มาจากแนวคิดกฎความสัมพันธ์ของข้อมูลที่ประกอบด้วยรายการ (Item) ต่าง ๆ ที่ปรากฏร่วมกันบ่อย ๆ เพื่อใช้อธิบายถึงรูปแบบที่แอบแฝงอยู่ในชุดข้อมูล 5) การค้นหาเพื่อนบ้านใกล้สุด k อันดับ (k-Nearest neighbor: k-NN classifier) เป็นการเรียนรู้โดยการเปรียบเทียบกันระหว่างข้อมูลที่ต้องการจำแนกทั้งหมดในชุดข้อมูลฝึกสอนที่มีลักษณะเหมือนกันหรือใกล้เคียงกัน (k อันดับ) ด้วยการพิจารณาแอททริบิวต์ต่าง ๆ โดยจะมีข้อมูลเรคคอร์ดหนึ่งที่ถูกกำหนดเป็นจุดหนึ่งในระนาบ พิจารณาจากค่า Euclidean distance 7) ซัพพอร์ทเวกเตอร์แมชชีน (Support vector machine: SVM) เป็นอัลกอริทึมในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการฝึกฝนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลที่ดีที่สุด และ 8) การจำแนกข้อมูลด้วยโครงข่ายประสาทเทียม (Artificial neural network: ANN classifier) (Wozniak, 2014) ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ โดยเชื่อมต่อกันเป็นโหนด (Node) หลายชั้นเรียกว่าโครงข่ายประสาทเทียมแบบหลายชั้น (Multilayer perceptron: MLP) ค่าฟังก์ชันของแต่ละโหนดเกิดจากฝึกฝน กระบวนการฝึกฝนเริ่มจากการป้อนสัญญาณรับเข้า (Input signal) เข้าโครงข่ายคำนวณค่านำเข้าคูณกับค่าน้ำหนักในโครงข่ายแล้วปรับค่าน้ำหนักประสาท และไบอัส (Bias) ตามกฎการเรียนรู้ เพื่อให้ค่าผลลัพธ์ใกล้เคียงเป้าหมายมากที่สุด

แล้วส่งผ่านข้อมูลสู่ฟังก์ชันกระตุ้นและส่งออกเป็นสัญญาณออก (Output signal) เพื่อเปรียบเทียบกับค่าเป้าหมายซึ่งเป็นค่าผิดพลาด และจะถูกส่งค่าย้อนกลับไปปรับค่าน้ำหนัก (Backpropagation algorithm) ให้พอดี (Fit) กับข้อมูลต่อไป ด้วยจุดเด่นคือมีความมั่นคงสูงต่อสิ่งรบกวน (Noise) ทำงานได้ดีกับข้อมูลที่มีค่าแบบไม่ต่อเนื่อง อีกทั้งสามารถจำแนกข้อมูลได้ทั้งแบบ มีผู้สอนและไม่มีผู้สอน รวมถึงประสิทธิภาพในการจำแนกข้อมูลแม้พบความสัมพันธ์ระหว่างแอททริบิวต์กับหมวดหมู่ต่าง ๆ เพียงเล็กน้อย จากข้อดีดังกล่าวจึงถูกนำมาประยุกต์ใช้อย่างแพร่หลาย ปัจจุบันได้พัฒนาเป็นวิธีการเรียนรู้เชิงลึก (Deep learning) (Vasilev et al., 2019) ซึ่งเป็นอัลกอริทึมที่สามารถทำให้เครื่องจักรสามารถตัดสินใจได้เช่นเดียวกับมนุษย์ โดยการเรียนรู้ของเครื่องเป็นการประยุกต์ใช้ความรู้ทางด้านสถิติในการวิเคราะห์ข้อมูลและสร้างแบบจำลองสำหรับทำนายผลลัพธ์จากข้อมูล

การเรียนรู้เชิงลึกปัจจุบันได้รับการยอมรับว่ามีประสิทธิภาพในการวิเคราะห์และทำนายผลได้ดีกว่าอัลกอริทึมการเรียนรู้ของเครื่องแบบเดิมเป็นอย่างมาก ทั้งในด้านของจำนวนข้อมูลที่เพิ่มขึ้นอย่างก้าวกระโดดและหน่วยประมวลผลของคอมพิวเตอร์ที่มีประสิทธิภาพสูงขึ้น ทำให้การเรียนรู้เชิงลึกนั้นสามารถใช้ตัวแปรจำนวนมากในการวิเคราะห์ (Rosebroke, 2017) เพื่อเพิ่มประสิทธิภาพในการวิเคราะห์ได้ ไม่เพียงแต่การเพิ่มจำนวนชั้นเท่านั้น ยังสามารถปรับความสามารถในการเรียนรู้ (Learning rate) ของอัลกอริทึม โดยผลจากการเรียนรู้แบบย้อนกลับ (Backpropagation) ได้นำมาปรับปรุง แบ่งเป็นฟังก์ชันการสูญเสีย (Loss function) เพื่อคำนวณหาค่าความผิดพลาดจากการเปรียบเทียบระหว่างผลที่ได้จากแบบจำลอง และผลลัพธ์ที่ใช้ในการฝึกสอน จากนั้นนำค่าสูญเสีย (Loss) ที่ได้มาใช้กับฟังก์ชันการหาจุดสมดุลของเงื่อนไข (Optimize function) ซึ่งเป็นฟังก์ชันการปรับค่าพารามิเตอร์ที่ใช้ในการเรียนรู้ของแบบจำลองที่สร้างขึ้นมา งานวิจัยชิ้นนี้มุ่งจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึกจากอัลกอริทึม deepcut เพื่อติดแท็กปัญหาการให้บริการ และนำเสนอเป็นสารสนเทศแก่ผู้รับผิดชอบในการนำไปพัฒนาคุณภาพการให้บริการต่อไป

## 2. วัตถุประสงค์

1) เพื่อออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก
2) เพื่อประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

## 3. วิธีการศึกษา

งานวิจัยชิ้นนี้เป็นงานวิจัยเชิงประยุกต์ โดยใช้แนวคิดแบบจำลองการพัฒนาแอปพลิเคชันแบบเร่งรัด (Rapid Application Development model: RAD model) (McConnell, 1996) มาเป็นกรอบในการดำเนินงานแยกการพัฒนาออกเป็นมอดูล ก่อนจะนำมาประกอบเป็นชิ้นงานที่สมบูรณ์ ประกอบด้วย 5 มอดูล ได้แก่ มอดูลการเก็บรวบรวมข้อมูล มอดูลการเตรียมข้อมูล มอดูลการแบ่งคำภาษาไทย มอดูลการสร้างดัชนีคำศัพท์ และมอดูลการลดมิติของข้อมูลด้วย t-SNE มีรายละเอียดดังนี้

### 3.1. มอดูลการเก็บรวบรวมข้อมูล (Data Collection)

งานวิจัยชิ้นนี้เก็บรวบรวมข้อร้องเรียนรถโดยสารสาธารณะขององค์การขนส่งมวลชนกรุงเทพจากผู้ใช้ที่เผยแพร่เป็นสาธารณะบนเว็บบอร์ด http://www.bmta.co.th/?q=th/forum ในหมวดร้องเรียนรถองค์การที่ตั้งกระทู้ขึ้นภายในวันที่ 1 มกราคม 2564 ถึง 31 กรกฎาคม 2564 รวมทั้งสิ้น 1,275 ข้อความ เก็บข้อมูลเฉพาะหัวข้อร้องเรียนเท่านั้น ดังตารางที่ 1 มาจัดเก็บในรูปของไฟล์เอกสารธรรมดา (.txt) แยกเป็น 2 ส่วน ส่วนที่ 1 นำเป็นชุดข้อมูลฝึกฝน (Training dataset) เพื่อสร้างโมเดลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึกด้วย deepcut จำนวน 1,020 ข้อความ คิดเป็นร้อยละ 80 และส่วนที่ 2 จำนวน 255 ข้อความ คิดเป็นร้อยละ 20 นำมาเป็นข้อมูลทดสอบ (Test dataset) ในลักษณะของข้อมูลที่ไม่ปรากฏมาก่อนในการเรียนรู้ (Unseen dataset) เพื่อประเมินผลความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

**ตารางที่ 1** ข้อร้องเรียนรถโดยสารสาธารณะ

| ลำดับ | ข้อความ |
|---|---|
| 1. | บัตร Prompt Card เติมเงินแล้วใช้ไม่ได้ เครื่องไม่อ่าน |
| 2. | 11 25-3 9321-ไม่จอดรับผู้โดยสาร 40252 |
| ... | ... |
| 1275. | ปอ.16 ไม่รับผู้โดยสาร |

### 3.2. มอดูลการเตรียมข้อมูล (Data Preprocessing)

เนื่องจากข้อมูลที่นำมาจากอินเทอร์เน็ตด้วยวิธีการขูดเว็บ (Mitchell, 2015) มักจะปรากฏคำสั่งจัดรูปแบบข้อความ (HTML tag) ตลอดจนมีเครื่องหมายควบคุมการแสดงผลเป็นส่วนประกอบ ดังนั้นก่อนจะนำไปประมวลผลจำเป็นต้องจัดการกับข้อมูลที่ไม่เป็นระเบียบก่อน โดยการใช้การกำหนดรูปแบบอักขระ (Regular expression) เพื่อกำหนดรูปแบบโครงสร้างของข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการนำไปประมวลผล

งานวิจัยชิ้นนี้ถือว่าทุกคำมีผลต่อการกำหนดคุณลักษณะของข้อความ จึงไม่มีกระบวนการกำจัดคำหยุด เป็นไปในทิศทางเดียวกับแนวโน้มการกำจัดคำหยุดที่มีจำนวนคำลดลงเรื่อย ๆ จนในปัจจุบันแนวคิดการเรียนรู้เชิงลึก ไม่มีการกำจัดคำหยุดเลย อย่างไรก็ดี กระบวนการเตรียมข้อมูล ยังคงกำจัดตัวเลขและเครื่องหมายวรรคตอน การกำจัดช่องว่าง (Space) การกำจัดคำหยาบ (Swear word) และการเปลี่ยนตัวอักษรภาษาอังกฤษให้เป็นตัวพิมพ์เล็กทั้งหมด ก่อนจะนำเข้ากระบวนการแบ่งคำภาษาไทยด้วยอัลกอริทึม deepcut ในมอดูลต่อไป

### 3.3. มอดูลการแบ่งคำภาษาไทย (Thai Word Segmentation)

ลักษณะของภาษาไทยนั้นมีความแตกต่างกับภาษาอังกฤษ เนื่องจากภาษาอังกฤษจะมีช่องว่างในการระบุคำแต่ละคำ อีกทั้งรูปแบบของคำเกิดจากการประสมกันของตัวอักษร สระ และวรรณยุกต์เข้ามา

ประกอบเพื่อบ่งชี้ความหมายของคำ อันเป็นคุณลักษณะของภาษาไทย จึงทำให้โครงสร้างของภาษาไทยมีความซับซ้อน โดยผลของการแบ่งคำที่ถูกต้องมีความสำคัญต่อการนำข้อมูลคำศัพท์ไปประมวลผลในขั้นต่อไป (Tapsai, Unger, & Meesad, 2021) อย่างไรก็ตาม ยังคงเกิดปัญหาสำคัญ 4 ประเด็น คือ 1) ประสิทธิภาพการแบ่งคำ 2) การตรวจสอบและแก้ไขคำสะกดผิด 3) รูปแบบการสะกดคำที่หลากหลาย และ 4) การแบ่งกลุ่มคำประสม ปัญหาเหล่านี้ทำให้เกิดผลลัพธ์ที่ไม่ถูกต้องและเกิดคำจำนวนมากที่กระทบต่อการประมวลผลภาษาธรรมชาติสำหรับภาษาไทย

การแบ่งคำสำหรับภาษาไทย แบ่งออกเป็น 3 ประเภท ได้แก่ 1) การแบ่งคำโดยใช้กฎ (Rules-Based Word Segmentation: RBWS) 2) การแบ่งคำตามพจนานุกรม (Dictionary-Based Word Segmentation: DBWS) และ 3) การแบ่งคำตามการเรียนรู้ (Learning-Based Word Segmentation: LBWS) งานวิจัยชิ้นนี้ใช้การแบ่งคำภาษาไทยตามการเรียนรู้เชิงลึกด้วยการเขียนโปรแกรมภาษาไพธอน (Python) ผ่าน Google Colab หรือชื่อเต็มคือ Google Colaboratory ที่เป็นบริการโฮสต์โปรแกรม Jupyter Notebook โดยบริษัทกูเกิล (Google) สำหรับเขียนและเรียกใช้ภาษาไพธอน จากเดิมที่ทำงานบนคอมพิวเตอร์มาให้บริการบนคลาวด์ (Software as a Service: Saas) ที่มีอัลกอริทึมการแบ่งคำภาษาไทยให้เลือกใช้งานมากมาย เช่น newmm, longest, pyicu, attacut, PyThaiNLP และ deepcut ที่พัฒนาโดยบริษัททรู คอร์ปอเรชั่น (True Corporation) เป็นระบบแบ่งคำแบบเรียนรู้เชิงลึก (Deep learning) โดยพัฒนาด้วยโมเดลแบบ CNN (Convolutional Neural Network) แบบ 1 มิติ มาพยากรณ์ว่าตัวอักษรตัวนี้เป็นตัวเริ่มต้นของคำหรือไม่ด้วยการจำแนกข้อมูลแบบไบนารี่ (Binary classification) โดยการเรียกใช้งานฟังก์ชัน deepcut.tokenize เพื่อแบ่งคำจากหัวข้อร้องเรียนรถโดยสารสาธารณะจำนวน 1,020 ข้อร้องเรียน ให้อยู่ในรูปคำเดี่ยว (Term) จากนั้นจะแปลงข้อมูลทั้งหมดไปอยู่ในรูปลิสต์ของคำในลักษณะถุงคำ (Bag of word) (Jo, 2019) ดังภาพที่ 1

```
['ทำไม', 'รถ', 'แดง', 'ไม่', 'มี', 'การ', 'เว้น', 'ที่', 'นั่ง', 'ช่วง', 'โควิดระบาด', 'พนักงาน', 'ขับ', 'รถ',
'เล่น', 'ไลน์', 'คุย', 'โทรศัพท์', 'ตลอดทาง', 'รถ', 'สาย', 'สาย', 'ขับ', 'เร็ว', 'และ', 'อันตราย', 'มาก',
'รถ', 'เมย์ก', 'ไม่', 'จอด', 'ป้าย', 'สาย', 'ขับ', 'รถ', 'ประมาท', 'และ', 'ไม่', 'จอด', 'ป้าย',
'ร้องเรียน', 'รถ', 'เมมาช้า', 'ขับ', 'รถ', 'โดย', 'ความ', 'ประมาท', 'ร้องเรียน', 'รถ', 'ปอ.สาย', 'ก.',
'รถ', 'เมล์', 'สาย', 'ไม่', 'จอด', 'ป้าย', 'กระเป๋า', 'รถเมย์', 'ใช้', 'วาจา', 'และ', 'กริยา', 'ไม่',
'สุภาพ', 'กับ', 'ผู้', 'โดยสาร', 'รถ', 'เมล์', 'สาย', 'กด', 'กริ่ง', 'แล้ว', 'ไม่', 'จอด', 'ป้าย', 'เลย',
'ไป', 'อีก', 'ป้าย', 'ถึง', 'จอด', ..., 'hino', 'ยูโรทูแอร์', 'ไม่', 'เย็น']
```

**ภาพที่ 1** การแบ่งคำภาษาไทยด้วยอัลกอริทึม deepcut

### 3.4. มอดูลการสร้างดัชนีคำศัพท์ (Indexing)

เป็นกระบวนการแปลงเอกสารที่เป็นภาษาธรรมชาติ ให้คอมพิวเตอร์สามารถเข้าใจและประมวลผลได้ การสร้างดัชนีเป็นการสร้างตัวแทนเอกสาร (Document representation) ให้อยู่ในรูปของเวกเตอร์เอกสาร (Word vector) ที่สามารถคำนวณค่าน้ำหนักคำ (Term weighting) หลักการคือหาคำที่

ปรากฏทั้งหมดก่อน จากข้อร้องเรียนจำนวน 1,020 ข้อความ แล้วนำไปเก็บไว้ในพจนานุกรม (Dictionary) เพื่อใช้แปลงคำศัพท์ที่พบให้อยู่ในรูปดัชนี

อย่างไรก็ดี คำที่ปรากฏบ่อย ๆ ในหลายเอกสาร อาจจะมีน้ำหนักหรือความสำคัญน้อยกว่าคำสำคัญ สำหรับการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ จึงต้องทำการหาค่าน้ำหนักรูปแบบคำเดี่ยวด้วยการวิเคราะห์น้ำหนักของคำ (Term Frequency–Inverse Document Frequency: TF-IDF) (Manning, Raghavan, & Schütze, 2008) เพื่อประเมินความสำคัญของคำต่อเอกสารจาก 2 ปัจจัยที่มีความสัมพันธ์กัน คือ ค่าความถี่ของคำ (TF) เพื่อหาว่าแต่ละคำนั้นปรากฏบ่อยแค่ไหนในแต่ละเอกสาร ดังสมการที่ 1 จากนั้นหาค่าความผกผันในความถี่ของเอกสาร (IDF) โดยการให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของเอกสารซึ่งควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยในชุดของเอกสารที่เหลือทั้งหมด ดังสมการที่ 2 และจะนำค่าความถี่ของคำศัพท์ (TF) และค่าความผกผันในความถี่ของเอกสาร (IDF) ของแต่ละคำ มาคูณกันเพื่อหาค่าน้ำหนักที่สามารถแยกคำศัพท์สำคัญออกมาได้ ดังสมการต่อไปนี้

$$TF = \frac{\text{จำนวนคำที่ปรากฏในเอกสาร}}{\text{จำนวนคำทั้งหมดในเอกสาร}} \quad (1)$$

$$IDF = \log\left(\frac{\text{จำนวนเอกสารทั้งหมดที่ใช้พิจารณา}}{\text{จำนวนเอกสารที่มีคำคำนั้นปรากฏอยู่}}\right) \quad (2)$$

$$TFIDF = TF * DF \quad (3)$$

จากผลลัพธ์จะพบคำศัพท์ที่มีค่า TF-IDF จากเอกสารทั้งหมด ซึ่งเป็นคำศัพท์ที่มีแนวโน้มที่จะเป็นใจความสำคัญของเอกสาร ตามแนวคิดคำศัพท์ถูกกล่าวถึงบ่อยที่สุดและไม่ได้ปรากฏอยู่หลายเอกสารจนเกินไป เพื่อนำมาคัดเลือกคุณลักษณะสำหรับสร้างคลังคำศัพท์ในการจำแนกปัญหาในการให้บริการ

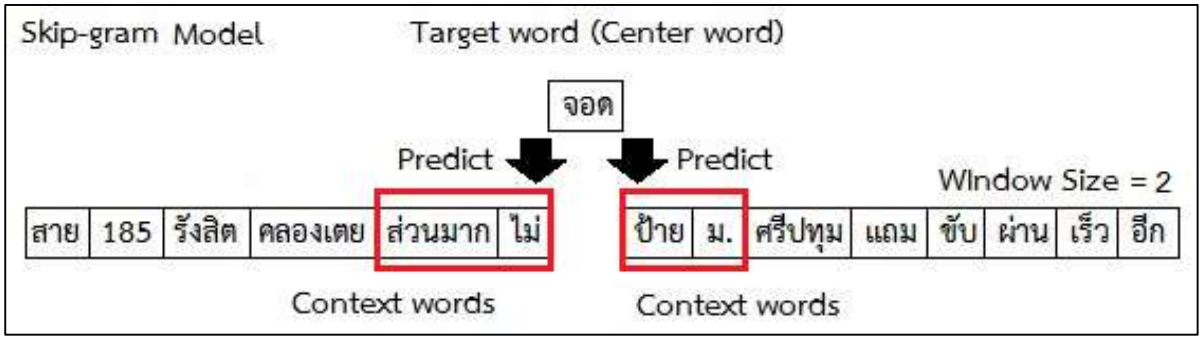
งานวิจัยชิ้นนี้ใช้ TfidfTransformer ในการหาค่าน้ำหนักของคำ (Raschka & Mirjalili, 2017) แต่อาจพบปัญหาข้อมูลที่มีความซ้ำซ้อนกัน อันเนื่องมาจากจำนวนมิติข้อมูลจำนวนมากที่ทับซ้อนกันในชุดข้อมูล

### 3.5. มอดูลการลดมิติของข้อมูลด้วย t-SNE

เป็นหนึ่งในขั้นตอนการสำรวจข้อมูล (Data exploration) เพื่อทำความเข้าใจรายละเอียดภายในของข้อมูล ก่อนที่จะเข้าสู่กระบวนการวิเคราะห์ข้อมูลในลำดับต่อไป เนื่องจากชุดข้อมูลดิบ (Dataset) ที่ได้จากการแบ่งคำภาษาไทยนั้น อาจจะมีความซับซ้อน ส่งผลต่อความถูกต้องและระยะเวลาในการวิเคราะห์ข้อมูล จึงต้องทำการลดมิติลง ให้สามารถวิเคราะห์ข้อมูลออกมาง่ายขึ้น โดย t-SNE (t-Distributed Stochastic Neighbor Embedding) พัฒนาโดย van der Maaten, L. และ Hinton, G. (2008) เป็นวิธีที่นิยมใช้ในการแสดงผลข้อมูลที่ไม่ใช่เชิงเส้น เหมาะกับข้อมูลที่มีมิติสูง โดยจะคำนวณความคล้ายคลึงกันระหว่างคู่ของจุดข้อมูลในพื้นที่มิติสูงและพื้นที่ในมิติต่ำ เพื่อปรับความคล้ายคลึงในมิติ 2 ระดับนี้ด้วยค่าฟังก์ชันการสูญเสีย (Loss function) จากการกระจายตัวของความน่าจะเป็น (Probability distribution) โดยใช้การกระจายตัว

แบบเกาส์เซียน (Gaussian distribution) หมายถึง โอกาสการกระจายซึ่งมีลักษณะต่อเนื่อง มีรูปร่างคล้ายระฆังสองด้านเหมือนกัน เป็นตัวแทนของความผิดพลาดจากการสุ่ม ในการกำหนดความสัมพันธ์ระหว่างจุดข้อมูลในมิติสูง และใช้สัดส่วนความเบี่ยงเบนของค่าเฉลี่ยของกลุ่มตัวอย่าง (Student t-distribution) เพื่อสร้างความน่าจะเป็นแบบกระจายตัวในมิติต่ำ เพื่อป้องกันการกระจุกตัวหรือทับซ้อนของจุดข้อมูลในมิติต่ำ ซึ่งเป็นผลมาจากปัญหามิติของข้อมูล (Curse of dimensionality)

งานวิจัยชิ้นนี้ใช้แบบจำลอง Skip-gram ที่จะเลือกคำหนึ่ง ๆ ของบริบทในการทำนายคำทุกคำที่อยู่ในบริบทของคำนั้น (Zong, Xia & Zhang, 2021) การฝึกฝนโมเดลนี้จะนำเอาเวกเตอร์ของคำเป้าหมายมาใช้เป็นคำกลาง (Center word) และทำนายการกระจายตัว (Probability distribution) ของคำที่น่าจะเป็นบริบทของคำคำนี้ ดังภาพที่ 2 แสดงลักษณะการทำงานโดยใช้กับประโยคตัวอย่างคือ “สาย 185 รังสิต-คลองเตย ส่วนมากไม่จอดป้าย ม.ศรีปทุม แถมขับผ่านเร็วอีก”




**ภาพที่ 2** การทำงานของแบบจำลอง Skip-gram

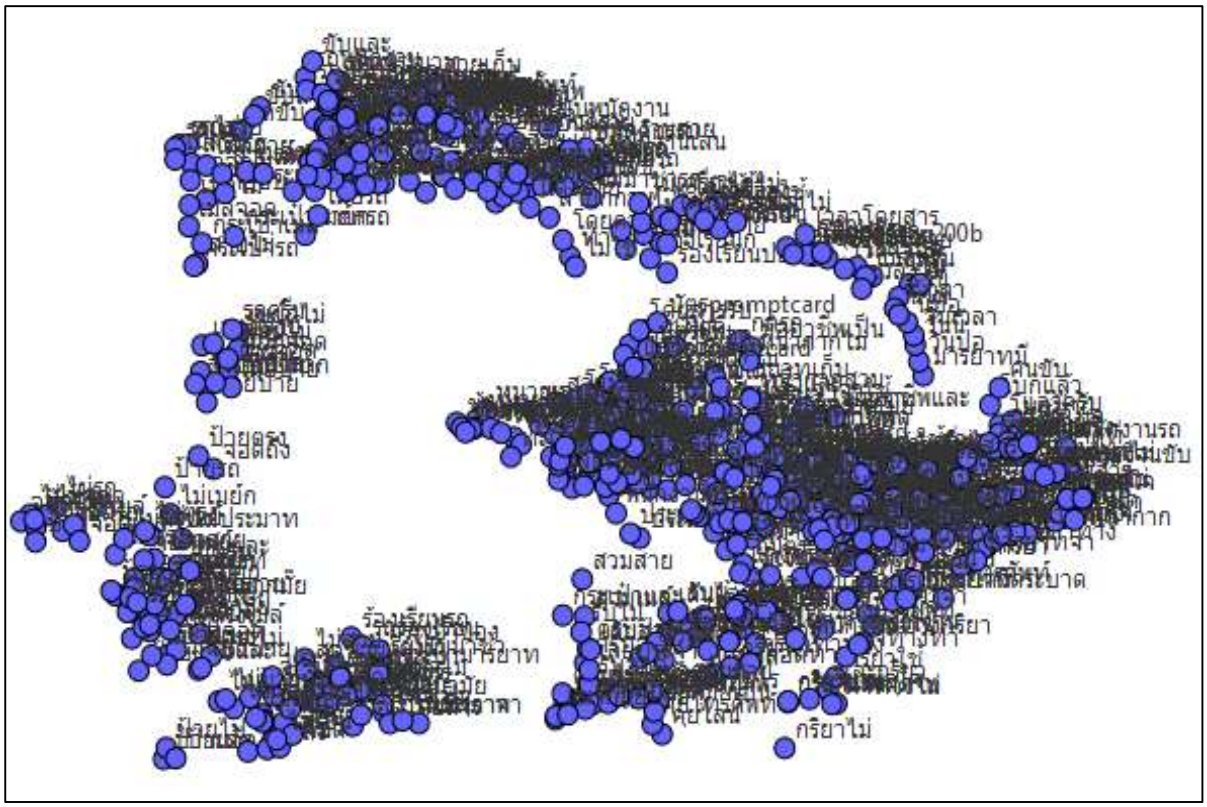
จากนั้นกำหนด WINDOW_SIZE=2 เพื่อหาชุดของคำของการตัดข้อความแต่ละครั้งจะได้ชุดคาบเกี่ยว (Overlapping) (Zhai & Massung, 2016) ของคำเป้าหมายในลักษณะก่อนหน้าและหลังคำตามจำนวนที่กำหนดของคำเป้าหมาย ดังตารางที่ 2

**ตารางที่ 2** ผลลัพธ์จากการตัดข้อความกำหนด WINDOW_SIZE=2

| No. | Input | Output |
|---|---|---|
| 1. | สาย | 185 |
| 2. | สาย | รังสิต |
| 3. | 185 | สาย |
| .. | .. | .. |
| 54. | อีก | เร็ว |

จากตารางที่ 2 ประโยคตัวอย่างคือ “สาย 185 รังสิต-คลองเตย ส่วนมากไม่จอดป้าย ม.ศรีปทุม แถมขับผ่านเร็วอีก” เมื่อกำหนด WINDOW_SIZE=2 จะได้ชุดของคำจำนวน 54 รายการ จากนั้นระบุ 3 พารามิเตอร์ที่จำเป็นเพื่อเรียกใช้ t-SNE แบบ Skip-gram (scikit-learn developers (BSD License), 2020) ตามค่าเริ่มต้นดังนี้ 1) กำหนดค่าอัตราการเรียนรู้ (Learning rate) เท่ากับ 200 เป็นตัวเลขที่ใช้ในการปรับระดับความเร็วการเรียนรู้ของโครงข่ายประสาทเทียม สำหรับ t-SNE มักอยู่ในช่วง 10 ถึง 1,000 2) กำหนด

ความสับสน (Perplexity) เท่ากับ 30 เพื่อดูการกระจายตัวของความน่าจะเป็น (Probability) มักอยู่ในช่วง 5 ถึง 50 และ 3) กำหนดจำนวนการทำซ้ำ (Iteration) เท่ากับ 1,000 สูงสุดสำหรับการเพิ่มประสิทธิภาพ แต่ไม่ควรต่ำกว่า 250 ผลลัพธ์ดังภาพที่ 3




**ภาพที่ 3** การลดมิติของข้อมูลด้วย t-SNE

จากภาพที่ 3 พบว่า ผลลัพธ์จากการแบ่งคำภาษาไทยจะประกอบไปด้วยคำย่อย ๆ จำนวนมาก เมื่อผ่านกระบวนการลดมิติของข้อมูลด้วย t-SNE แล้วจะแสดงผลข้อมูลที่มีความซ้ำซ้อนกัน อันเนื่องมาจากจำนวนมิติข้อมูลจำนวนมากที่ทับซ้อนกันในชุดข้อมูล แต่หากพิจารณาลงไปในรายละเอียดจะพบว่า มีการกระจายตัวของกลุ่มที่มีความคล้ายคลึงกันในลักษณะคลัสเตอร์ (Cluster) อันมีรูปแบบในการเกิดจากบริบทของเวกเตอร์ที่ เป็นไปในทิศทางเดียวกัน (Euclidean vector) ดังภาพที่ 4 กลุ่มที่มีรูปแบบการเกิดจากบริบทของเวกเตอร์ที่เป็นไปในทิศทางเดียวกัน เช่น “ไม่จอด” มักปรากฏร่วมกับ “ป้าย” และเป็นไปในทิศทางเดียวกับคำว่า “จอดไม่สนิท” “เลยป้าย” “แช่ขวา” และ “ไม่เข้าป้าย” เป็นต้น จึงเป็นที่มาของการสร้างคลาสการจอดรับส่ง

การจัดกลุ่มคลัสเตอร์เพื่อจำแนกข้อร้องเรียนรถโดยสารสาธารณะในงานวิจัยชิ้นนี้ กำหนดจากแนวคิดระดับความพึงพอใจของผู้รับบริการ (Customer satisfaction) มาเป็นกรอบในการจัดกลุ่ม ซึ่ง Verma & Ramanayya (2015) กล่าวว่า เกี่ยวข้องกับการบริการขนส่งสาธารณะในสององค์ประกอบหลัก ได้แก่ ความครอบคลุมของเส้นทาง (Route) และความถี่ในการเดินรถ (Frequency) ดังนั้น ผู้ประกอบการธุรกิจจำเป็นต้องประเมินความต้องการในการเดินทางและทรัพยากรที่มีกับผู้ให้บริการ โดยพิจารณาจากคุณภาพ

การบริการที่ได้รับตามความคาดหวังของลูกค้า ส่งผลให้การรับรู้ถึงคุณภาพของบริการที่ลูกค้าแต่ละรายจะได้รับจากการดำเนินการเดียวกัน อาจจะมีความพึงพอใจของผู้รับบริการแตกต่างกัน เช่น ความสะดวกสบาย กำหนดการและการดำเนินงาน พฤติกรรมผู้ให้บริการ ต้นทุน และปัจจัยอื่น ๆ เป็นต้น อย่างไรก็ดี แต่ละองค์ประกอบเหล่านี้สามารถมีคุณลักษณะย่อย ๆ ได้หลายอย่าง ดังตารางที่ 3

**ตารางที่ 3** คุณลักษณะที่เกี่ยวข้องกับคุณภาพการให้บริการขนส่งสาธารณะ[1]

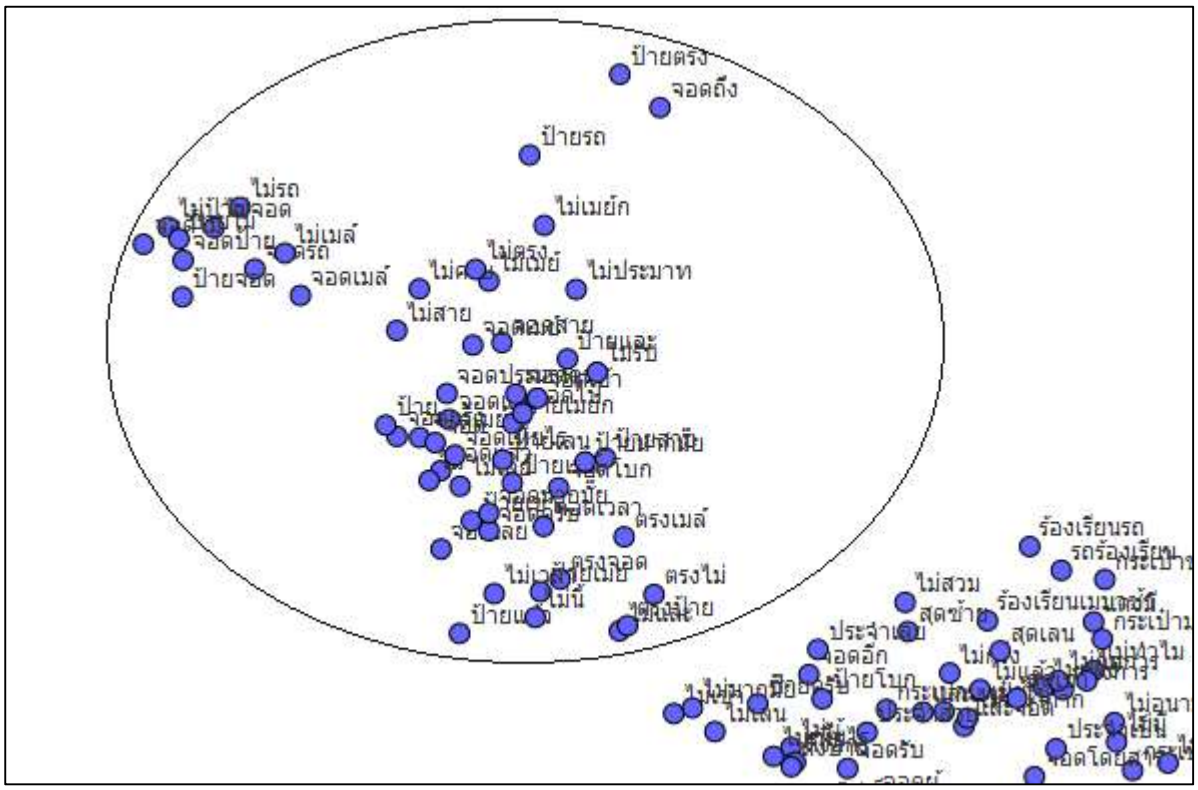
| Comfort and Convenience | Schedule and Operations | Crew Behavior | Cost and others Aspects |
|---|---|---|---|
| Overloading | Notification of Schedules | Courteousness with Passengers | Notification of Fares |
| Boarding and Alighting | Following the Schedule | Helping Children and Old Age People | Returning Small Changes |
| Seating Arrangement | Prompt Service During Break Down | Rash and Negligent | Adequacy of Fares |
| Movement within the Bus | Maintenance of Vehicles | Appearance of the Crew | Charges for Luggage |
| Driving Comfort | Cancellation of Schedules | Neatness and Professionalism | |
| Travel Time | Arrival/Departure Timings | Attitude of the Crew in General | |
| Luggage Allowance | | | |
| Stopping at the Bus Stops | | | |

จากตารางที่ 3 ผู้วิจัยทำการสังเคราะห์เพื่อสรุปคุณลักษณะทั้งหมดแยกตามบริบทการให้บริการรถโดยสารสาธารณะ เพื่อนำมาสร้างเป็นคลาสในการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ ดังตารางที่ 4

[1] หมายเหตุ. จาก Public Transport Planning and Management in Developing Countries (หน้า 109), by Verma & Ramanayya, 2015, Boca Raton, FL, USA: CRC Press

**ตารางที่** 4 การสังเคราะห์เพื่อสรุปคุณลักษณะตามบริบทการให้บริการรถโดยสารสาธารณะ

| **คลาส** | **คุณลักษณะ** | **คำอธิบาย** |
|---|---|---|
| 1) คลาสการขับขี่ | Rash and Negligent | ขับรถประมาท หวาดเสียว ไม่ปฏิบัติตามกฎจราจร ใช้ความเร็วเกินกฎหมายกำหนด |
| 2) คลาสการจอดรับส่ง | Stopping at the Bus Stops Boarding and Alighting | การจอดรถรับส่งไม่ตรงป้าย เลยป้าย จอดไม่สนิท ไม่เข้าป้าย ไม่จอดรับ |
| 3) คลาสผู้ให้บริการ | Crew Behavior<br>Courteousness with Passengers<br>Helping Children and Old Age People Appearance of the Crew<br>Neatness and Professionalism<br>Attitude of the Crew in General | กิริยามารยาท การใช้วาจา น้ำเสียงความสุภาพและความเป็นมืออาชีพของพนักงานขับและพนักงานให้บริการ |
| 4) คลาสการเดินรถ | Schedule and Operations<br>Notification of Schedules<br>Following the Schedule<br>Cancellation of Schedules<br>Travel Time Arrival/Departure Timings | การเดินรถ และสู่จุดหมายตามเวลาที่กำหนดตลอดจนปริมาณและความเพียงพอของรถโดยสารสาธารณะแต่ละเส้นทาง |
| 5) คลาสยานพาหนะ | Maintenance of Vehicles<br>Prompt Service During Break Down | สภาพและความพร้อมในการให้บริการของรถโดยสาร รวมถึงเสียงของเครื่องยนต์การปล่อยมลพิษ |
| 6) คลาสอุปกรณ์ให้บริการ | Comfort and Convenience<br>Driving Comfort | สภาพและความพร้อมของอุปกรณ์การให้บริการ เครื่องปรับอากาศ ที่นั่ง เครื่องจ่ายเงิน บัตรเติมเงิน ตลอดจนแอปพลิเคชันเสริมต่าง ๆ |
| 7) คลาสมาตรการป้องกันโรคระบาด | Overloading<br>Seating Arrangement<br>Movement within the Bus | การป้องกันโรคระบาด การเว้นระยะห่าง ความเพียงพอของที่นั่ง และปริมาณผู้โดยสารที่เหมาะสม |




**ภาพที่ 4** การกระจายตัวของกลุ่มที่มีความคล้ายคลึงกัน

จากตารางที่ 4 นำมาสร้างคลาสที่เกี่ยวข้องกับการร้องเรียนตามบริบทการให้บริการรถโดยสารสาธารณะ ตามลักษณะการกระจายตัวของกลุ่มที่มีความคล้ายคลึงกัน ดังภาพที่ 4 แบ่งออกเป็น 7 คลาส ได้แก่ 1) คลาสการขับขี่ (Driving class) จำนวน 191 ข้อร้องเรียน 2) คลาสการจอดรับส่ง (Shuttle stop class) จำนวน 190 ข้อร้องเรียน 3) คลาสผู้ให้บริการ (Person class) จำนวน 262 ข้อร้องเรียน 4) คลาสการเดินรถ (Bus schedule class) จำนวน 126 ข้อร้องเรียน 5) คลาสยานพาหนะ (Vehicle class) จำนวน 125 ข้อร้องเรียน 6) คลาสอุปกรณ์ให้บริการ (Equipment class) จำนวน 87 ข้อร้องเรียน และ 7) คลาสมาตรการป้องกันโรคระบาด (Prevent epidemics class) จำนวน 39 ข้อร้องเรียน ตามบริบทการให้บริการรถโดยสารสาธารณะ จากนั้นทำการแปลงแต่ละคลาสให้อยู่ในรูปแบบไบนารี่ (Binary) ประกอบด้วย 0 และ 1

## 4. ผลการศึกษา

งานวิจัยนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก และประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ แบ่งผลการศึกษาและการอภิปรายผล มีรายละเอียดดังนี้

### 4.1. ออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก

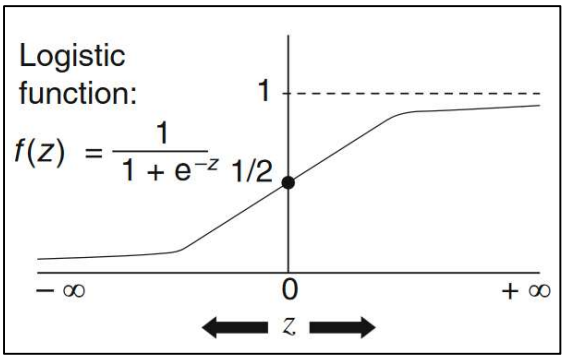
ทำนายประเภทจากการจำแนกข้อมูล (Classification model) คือการพยายามจัดข้อมูลที่กระจัดกระจายให้เข้ากับหมวดหมู่ที่ถูกกำหนดไว้ล่วงหน้า หนึ่งในรูปแบบการจำแนกข้อมูลที่ได้รับความนิยม เรียกว่าการจัดจำแนกแบบไบนารี (Binary classification) (Albon, 2018) เป็นการพิจารณาจัดหมวดหมู่ที่เข้าเกณฑ์

และไม่เข้าเกณฑ์ หรือกลุ่มที่ใช่และกลุ่มที่ไม่ใช่ ตลอดจนเป็นข้อมูลในกลุ่มและนอกกลุ่ม หากการจำแนกประเภทดังกล่าวมีการกำหนดหลักเกณฑ์กำหนดไว้ชัดเจน จะใช้หลักการพิจารณาข้อมูลด้วยฐานกฎ (Rule-based classification) แต่หากการจำแนกกลุ่มข้อมูลใดไม่มีการกำหนดกฎเกณฑ์ที่ชัดเจน อาจจำเป็นต้องใช้เทคนิคของการเรียนรู้ของเครื่อง (Machine learning) เข้ามาใช้ช่วยพิจารณาข้อมูลนั้น ๆ ว่ามีแนวโน้มที่จะอยู่ในหมวดหมู่ใด ซึ่งเทคนิคดังกล่าวจำเป็นต้องอาศัยความแตกต่างกันของข้อมูลระหว่างสองกลุ่มที่กำลังพิจารณา

การจัดจำแนกแบบไบนารี จะทำการแปลงข้อมูลให้อยู่ในรูปแบบไบนารี่ (Binary) ประกอบด้วย 0 และ 1 เทคนิคพื้นฐานหนึ่งที่ใช้ในการจัดหมวดหมู่ ได้แก่ เทคนิคการวิเคราะห์การถดถอยโลจิสติกส์ (Logistic regression) (Raschka & Mirjalili, 2017) ซึ่งเป็นการพยายามฟิตเส้นโค้งซิกมอยด์ (Sigmoid curve) ให้เข้ากับข้อมูลจริงมากที่สุดเท่าที่ทำได้ ด้วยหลักการที่คล้ายกับการวิเคราะห์การถดถอยเชิงเส้น (Linear regression) ซึ่งเป็นการพยายามฟิตสมการเส้นตรงเข้ากับข้อมูลจริง สิ่งที่แตกต่างคือข้อมูลที่ทำนายจากการทำการถดถอยโลจิสติกส์ นั้นเป็นการทำนายระหว่างค่า 1 (ในเกณฑ์) และ ค่า 0 (นอกเกณฑ์) ในขณะที่การทำนายจากการทำการถดถอยเชิงเส้นนั้น ค่าที่ได้จากการทำนายอาจเป็นค่าตัวเลขใด ๆ ก็ได้ เพื่อให้ผลจากการทำนายโดยการถดถอยโลจิสติกส์มีค่าเป็น 0 หรือ 1 เท่านั้น ตัวแปรที่ส่งผลต่อการทำนายจะถูกคำนวณให้กลายเป็นค่าความน่าจะเป็น (Probability) ว่าค่านั้น ๆ มีความเป็นไปได้ที่จะเป็น 1 เท่าใด (Kleinbaum & Klein, 2010) ดังสมการที่ 4

$$P(X) = \frac{1}{1+e^{-\propto+\sum \beta_i x_i}} \qquad (4)$$

จากสมการที่ 4 การนำสมการถดถอยเชิงเส้น ไปแทนค่า ของโลจิสติกส์ฟังก์ชัน (Logistic Function) ซึ่งค่าที่ได้ออกมาจะมีค่าระหว่าง 0,1 ทำให้สามารถทำนายการจำแนกข้อมูลได้ ดังภาพที่ 5




**ภาพที่ 5** โมเดลการถดถอยเชิงเส้นที่ได้หลังจากแทนค่า[2]

[2] หมายเหตุ. จาก Logistic Regression A Self-Learning Text (หน้า 345), by Kleinbaum & Klein, 2010, New York: Springer Science+Business Media LLC

งานวิจัยชิ้นนี้จำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยอัลกอริทึม deepcut จากไลบารี่ scikit-learn (Raschka & Mirjalili, 2017) แบบวนลูป เพื่อทดสอบประสิทธิภาพของโมเดลจากการวิเคราะห์การถดถอยโลจิสติกส์ของข้อร้องเรียนแต่ละคลาส การวัดประสิทธิภาพโมเดล (Manning, Raghavan, Schütze, 2008) จะวัดค่าจากความสามารถของการเรียนรู้ของเครื่อง (Machine learning) งานวิจัยชิ้นนี้เลือกใช้ค่าความถูกต้อง (Accuracy) เป็นหลักในการวัดผลในการจำแนกข้อมูลแบบไบนารี่ (Binary classification) หมายถึงโมเดลที่มีการกำหนดข้อมูลแต่ละคลาสเป็นป้ายกำกับ (Label) ที่แทนคลาสไม่ใช่ (No) เป็น 0 และคลาสใช่ (Yes) เป็น 1 ซึ่งผลลัพธ์จากการทำนายของโมเดลจะบอกว่ามีโอกาสที่จะเป็นคลาส 1 กี่เปอร์เซ็นต์ตามค่าความเชื่อมั่น ซึ่งมีค่าตั้งแต่ 0 ถึง 1 โดยกำหนดชั้นผลลัพธ์ (Output layer) แบบซิกมอยด์ (Sigmoid) หากมีค่าเชื่อมั่น (Threshold) มากกว่า 0.5 จะทำนายว่าเป็นคลาส 1 ซึ่งหาได้จากตาราง Confusion matrix (Raschka & Mirjalili, 2017) แสดงผลการประเมินผลลัพธ์การทำนายกับผลลัพธ์จริง ๆ ที่โมเดลจำแนกได้ดังตารางที่ 5

**ตารางที่ 5** ตาราง Confusion Matrix

| | **Predict class** | |
|---|---|---|
| Actual class | Class=yes | Class=no |
| Class=Yes | TP | FN |
| Class=No | FP | TN |

จากตารางที่ 5 เมื่อมีการทำนาย 2 ประเภท ผลการทำนายทั้งหมดที่เป็นไปได้จะมี 4 ชนิด ดังนี้

1) TP (True Positive) หมายถึง ค่าคลาสเป้าหมายคือ Yes แบบโมเดลทำนายว่า Yes
2) FN (False Negative) หมายถึง ค่าคลาสเป้าหมายคือ Yes แบบโมเดลทำนายว่า No
3) FP (False Positive) หมายถึง ค่าคลาสเป้าหมายคือ No แบบโมเดลทำนายว่า Yes
4) TN (True Negative) หมายถึง ค่าคลาสเป้าหมายคือ No แบบโมเดลทำนายว่า No

จากนั้นนำมาคำนวณหาค่าที่จำเป็นต่อการวัดประสิทธิภาพโมเดล ประกอบด้วย 1) ค่าความถูกต้อง (Accuracy) คือจำนวนข้อมูลที่ทำนายถูกโดยพิจารณารวมทุกคลาส 2) ความแม่นยำ (Precision) คือค่าของตัวแบบที่ทำนายให้ถูกต้อง โดยพิจารณาแยกทีละคลาส 3) ค่าความครบถ้วน (Recall) คือจำนวนการกระทำด้วยกันแบบที่ตรงกับความเป็นจริง โดยพิจารณาแยกทีละคลาส ดังสมการที่ 5, 6 และ 7 อย่างไรก็ดี ค่าความแม่นยำและค่าความครบถ้วนจะเป็นประโยชน์ต่อผู้ใช้ต่างกลุ่มกันไป โดยผู้ใช้จะให้ความสำคัญกับ ค่าความแม่นยำมากกว่าค่าความครบถ้วน เนื่องจากผู้ใช้ส่วนใหญ่ต้องการให้ข้อมูลที่เป็นผลลัพธ์ที่ตรงกับสิ่งที่ต้องการค้นหา แต่ในทางกลับกัน ผู้ที่พัฒนาระบบค้นหาข้อมูลจะเน้นไปที่ค่าความครบถ้วน เนื่องจากต้องการให้ระบบค้นหาและเลือกเอกสารที่ถูกต้องจากทั้งหมดให้ได้มากที่สุด ดังนั้นอีกค่าหนึ่งคือค่าประสิทธิภาพโดยรวม (F-measure) หรือ F1 score ที่เกิดจากการเปรียบเทียบกันระหว่างค่าความแม่นยำและ ค่าความครบถ้วนของแต่ละคลาสเป้าหมาย เพื่อเป็นมาตรวัดเดี่ยว (Single metric) ที่วัดความสามารถ ของโมเดลแทนค่าความแม่นยำและค่าความครบถ้วน ดังสมการที่ 8

Accuracy = (TP+TN )/(TP+FP+TN+FN) (5)

Precision = (TP )/(TP+FP) (6)

Recall = (TP )/(TP+FN) (7)

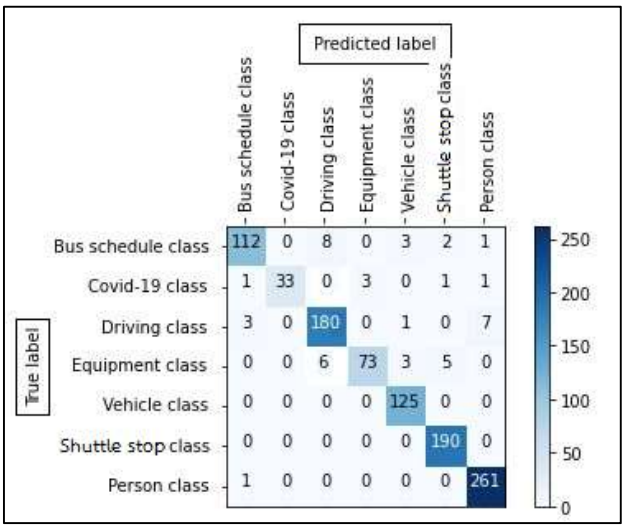
F-measure = 2 * (Precision*Recall)/(Precision+Recall) (8)

งานวิจัยชิ้นนี้ใช้การตรวจสอบแบบไขว้กัน 10 ชุด (10-Fold Cross Validation) (Zizka, Darena, & Svoboda, 2020) โดยแบ่งข้อมูลออกเป็น 10 ชุด ข้อมูลชุดที่ 1 ใช้เป็นชุดข้อมูลทดสอบ (Test dataset) และอีก 9 ชุดที่เหลือใช้เป็นชุดข้อมูลฝึกฝน (Training dataset) จากนั้นเปลี่ยนชุดข้อมูลถัดไปเป็นชุดข้อมูลทดสอบส่วนที่เหลือเป็นชุดข้อมูลฝึกฝน วนซ้ำไปจนครบทั้ง 10 ชุดข้อมูล วิธีการนี้เป็นวิธีที่นิยมเพื่อใช้ลดความผิดพลาดของผลลัพธ์จากการสุ่มเลือกชุดข้อมูลการเรียนรู้และชุดข้อมูลทดสอบ ผลลัพธ์ดังตารางที่ 6

**ตารางที่ 6** ค่าความแม่นยำการพยากรณ์ข้อร้องเรียนรถโดยสารสาธารณะ

| **ข้อร้องเรียนรถโดยสารสาธารณะ** | **Accuracy** | **Precision** | **Recall** | **F-Measure** |
|---|---|---|---|---|
| 1) คลาสการขับขี่ | 0.936417 | 0.921417 | 0.918417 | 0.919915 |
| 2) คลาสการจอดรับส่ง | 0.961841 | 0.946841 | 0.945341 | 0.946090 |
| 3) คลาสผู้ให้บริการ | 0.838597 | 0.823597 | 0.819297 | 0.821441 |
| 4) คลาสการเดินรถ | 0.914877 | 0.899877 | 0.892377 | 0.896112 |
| 5) คลาสยานพาหนะ | 0.982439 | 0.967439 | 0.958589 | 0.962993 |
| 6) คลาสอุปกรณ์ให้บริการ | 0.973587 | 0.958587 | 0.954337 | 0.956457 |
| 7) คลาสมาตรการป้องกันโรคระบาด | 0.984361 | 0.969361 | 0.959861 | 0.964588 |

จากตารางที่ 6 พบว่า ผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะทั้งหมดมีความถูกต้องมากกว่าร้อยละ 90 โดยเฉพาะคลาสมาตรการป้องกันโรคระบาด มีความถูกต้องสูงถึงร้อยละ 98 สอดคล้องกับจำนวนข้อร้องเรียนในคลาสที่มีจำนวนน้อยที่สุด ในทางกลับกันค่าความถูกต้องในคลาสผู้ให้บริการ มีความถูกต้องน้อยที่สุด เพียงร้อยละ 83 ตามจำนวนข้อร้องเรียนในคลาสที่มีจำนวนมากที่สุด ทำให้เกิดความหลากหลายของข้อมูลที่แตกต่างกันตามบริบทของผู้ใช้ ผลการประเมินดังกล่าวแสดงในตาราง Confusion matrix ดังภาพที่ 6




**ภาพที่ 6** ตาราง Confusion Matrix

อย่างไรก็ดี ผู้วิจัยทดลองเปลี่ยนการตรวจสอบแบบไขว้กัน 10 ชุด (10-Fold cross validation) เป็น 3, 5 และ 20 ตามลำดับ พบว่า ค่าความถูกต้องเปลี่ยนแปลงเพียงเล็กน้อยตามจำนวนชุดการตรวจสอบที่เพิ่มขึ้น ส่วนค่าความแม่นยำ และค่าความระลึกเปลี่ยนแปลงไม่แตกต่างกันมากนัก

**4.2. ประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ**

เมื่อนำผลลัพธ์จากการฝึกฝนโมเดลด้วยการถดถอยโลจิสติกส์มาเป็นแบบเรียนรู้เพื่อพยากรณ์ข้อร้องเรียนรถโดยสารอัตโนมัติ โดยทำการเขียนโปรแกรมภาษาไพธอนอีกครั้ง แล้วนำเข้าข้อร้องเรียนรถโดยสารสาธารณะ จำนวน 255 รายการที่กำหนดให้เป็นชุดข้อมูลทดสอบแบบไม่ปรากฏมาก่อนในชุดข้อมูลฝึกฝน (Unseen dataset) ทีละรายการแบบวนลูป ร่วมกับการจำแนกข้อความจากโมเดลที่ใช้สร้างการฝังคำ (Word embedding) ด้วยแนวคิดการแปลงคำให้อยู่ในรูปแบบเว็กเตอร์ เพื่อให้สามารถเปรียบเทียบกันได้ โดยใช้วิธีการคำนวณตัวเลขของคำนั้น ๆ มาจากบริบทของคำรอบ ๆ โดย Tomas Mikolov (2013) ได้เสนอวิธีการสร้างตัวแทนเชิงความหมายของคำ (Vector representations of words) จากแนวคิดความหมายของคำคำหนึ่งในประโยคนั้นมีความสัมพันธ์กับความหมายของคำที่อยู่รอบข้าง ด้วยการเรียนการสร้างคำของบริบทในประโยคโดยใช้เทคนิคโครงข่ายประสาทเทียม (Neural network) เรียกว่า Word2Vec ที่แปลงคำในเอกสารให้อยู่ในรูปเวกเตอร์ เพื่อให้สามารถเปรียบเทียบกันได้ โดยคำที่มีความหมายเหมือนกันหรือคล้ายกันจะมีค่าของเวกเตอร์ที่ใกล้เคียงกัน ในทางตรงกันข้ามคำที่มีความหมายแตกต่างกันจะมีค่าของเวกเตอร์ที่แตกต่างมากตามไปด้วย ประกอบด้วย 2 แบบจำลอง ได้แก่ CBOW (Continuous bag of words) และ Skip-gram ในงานวิจัยชิ้นนี้ใช้ fastText ด้วยแนวคิดการเรียนรู้เชิงลึกความเร็วสูง (Bhattacharjee, 2018) ที่ถูกพัฒนาด้วยบนพื้นฐานภาษา C++ โดย Facebook AI Research (FAIR) เป็นเครื่องมือในการจำแนกข้อความ (Text classification) และช่วยลดจำนวนมิติของข้อความ (Text representation) ที่ถูกเผยแพร่ในลักษณะ Pre-train word vectors สำหรับใช้งานกับภาษาไทย ที่ทำ

การคำนวณ fastText จากข้อมูลในเว็บไซต์วิกิพีเดีย (Wikipedia) เพื่อเปรียบเทียบค่าความถูกต้องในการจำแนกข้อร้องเรียนรถโดยสารสาธารณะจากอัลกอริทึม deepcut แล้วทำการติดแท็กตามคลาสที่มีค่าความน่าจะเป็นมากที่สุด ผลการดำเนินงานดังตารางที่ 7

**ตารางที่ 7** ค่าความถูกต้องของการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยอัลกอริทึม deepcut เปรียบเทียบกับอัลกอริทึม fastText

| ลำดับ | ข้อร้องเรียนรถโดยสารสาธารณะ | | คลาส | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | การขับขี่ | การจอดรับส่ง | ผู้ให้บริการ | การเดินรถ | ยานพาหนะ | อุปกรณ์ให้บริการ | มาตรการป้องกันโรคระบาด |
| 1. | 543 ไม่จอดป้าย | deepcut | 0.045 | 0.805 | 0.153 | 0.040 | 0.007 | 0.015 | 0.022 |
| | | fastText | 0.048 | 0.711 | 0.059 | 0.053 | 0.086 | 0.078 | 0.071 |
| 2. | รถเมล์สาย 114 ไม่จอด 63 และรับผู้โดยสารที่ป้าย | deepcut | 0.033 | 0.777 | 0.269 | 0.058 | 0.003 | 0.024 | 0.028 |
| | | fastText | 0.060 | 0.683 | 0.013 | 0.036 | 0.278 | 0.116 | 0.253 |
| ... | ... | deepcut | ... | ... | ... | ... | ... | ... | ... |
| | | fastText | ... | ... | ... | ... | ... | ... | ... |
| 255. | ปอ.16 ไม่รับผู้โดยสาร | deepcut | 0.040 | 0.896 | 0.223 | 0.046 | 0.001 | 0.017 | 0.023 |
| | | fastText | 0.053 | 0.801 | 0.129 | 0.047 | 0.422 | 0.406 | 0.400 |

จากตารางที่ 7 พบว่า ค่าความถูกต้องของการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยอัลกอริทึม deepcut เปรียบเทียบกับอัลกอริทึม fastText มีค่าความถูกต้องในระดับสูง และเป็นไปในทิศทางเดียวกัน อย่างไรก็ดี แม้ค่าความถูกต้องจากอัลกอริทึม fastText จะไม่สูงมากนักเมื่อเทียบกับ อัลกอริทึม deepcut แต่มีการฝึกฝนข้อมูลที่รวบรวมจากเว็บไซต์วิกิพีเดียนั้นค่อนข้างคงที่ จึงไม่เกิดการขยายถุงคำออกไปอย่างไร้ขอบเขต จนส่งผลต่อประสิทธิภาพการใช้ทรัพยากรเครื่องและเวลาในการประมวลผลที่น้อยกว่าเมื่อเทียบกับอัลกอริทึม deepcut ตลอดจนข้อมูลในเว็บไซต์วิกิพีเดียภาษาไทยนั้นใช้ภาษาที่ค่อนข้างเป็นทางการ จึงส่งผลให้ข้อมูลที่อยู่ในคลังคำศัพท์มีความถูกต้องสูง และพร้อมนำไปใช้ประโยชน์ได้ทันที

ดังนั้นเพื่อประเมินผลความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะจากการติดแท็กปัญหาการให้บริการกับข้อมูลชุดทดสอบแบบไม่ปรากฏมาก่อนในชุดข้อมูลฝึกฝน (Unseen dataset) จำนวน 255 ข้อความ ในงานวิจัยชิ้นนี้ทำการประเมินผลโดยผู้เชี่ยวชาญ 3 ท่าน ท่านละ 85 รายการ กำหนดผลการติดแท็กปัญหาการให้บริการแต่ละข้อร้องเรียน โดยหากติดแท็กปัญหาการให้บริการถูกต้องทั้งหมดตามคลาสที่ร้องเรียนได้ 1 คะแนน ในทางตรงกันข้าม หากติดแท็กปัญหาการให้บริการของข้อร้องเรียนใดจำแนกไม่ถูกต้อง หรือถูกต้องแต่ไม่ครอบคลุมประเด็นปัญหาทั้งหมด ถือว่าได้ 0 คะแนน ผลการประเมินดังตารางที่ 8

**ตารางที่ 8** ผลประเมินความถูกต้องของผลการติดแท็กปัญหาการให้บริการ

| ท่านที่ 1 | | ท่านที่ 2 | | ท่านที่ 3 | | ค่าเฉลี่ย | |
|---|---|---|---|---|---|---|---|
| N=85 | | N=85 | | N=85 | | N=255 | |
| ✓ | ร้อยละ | ✓ | ร้อยละ | ✓ | ร้อยละ | ร้อยละ | แปลผล |
| 74 | 87.06 | 72 | 84.71 | 74 | 87.06 | 86.27 | ดีมาก |

จากตารางที่ 8 พบว่า ผลการประเมินความถูกต้องของผลการติดแท็กปัญหาให้บริการ อยู่ในระดับดีมาก (ร้อยละ 86.27) แสดงถึงการออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก ให้ผลลัพธ์ความถูกต้องในระดับสูง โดยเฉพาะข้อร้องเรียนแบบ 1 ประเด็นต่อ 1 ข้อร้องเรียน เนื่องจากเหมาะกับคำศัพท์ที่บริบทสามารถกำหนดขอบเขตได้แน่นอน ในทางกลับกัน ปัญหาที่พบส่วนใหญ่เนื่องมาจากมีคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส เช่น “เสียงดัง” ที่จำเป็นต้องดูบริบทของข้อความประกอบ เนื่องจากเสียงดังเป็นกิริยาที่อาจเกิดจากบุคคลในคลาสการขับขี่ หรืออาจเป็นเสียงที่เกิดจากเครื่องยนต์จากคลาสยานพาหนะ หรือคลาสอุปกรณ์ให้บริการก็เป็นได้ หรือ “ไม่สุภาพ” ที่ไม่จำกัดเพียงคลาสผู้ให้บริการ แต่อาจถูกนำไปใช้ร้องเรียนพฤติกรรมการขับขี่ ตลอดจน “รถร้อน” ที่เป็นคำเรียกรถโดยสาธารณะที่ไม่ใช่รถปรับอากาศที่เป็นส่วนหนึ่งในการจำแนกข้อร้องเรียนในคลาสยานพาหนะที่เกิดจากอุณหภูมิในรถที่ไม่ได้มาตรฐาน เป็นต้น ดังนั้น การรวบรวมข้อมูลคำศัพท์ต้องมีกระบวนการที่น่าเชื่อถือ และปรับปรุงให้ครอบคลุมครบถ้วนอยู่เสมอ รวมถึงอาจต้องจำแนกคลาสโดยคำนึงถึงบริบทของข้อความเป็นหลัก ดังนั้นการวัดความถูกต้องนอกจากจะวัดในเชิงความหมายแล้ว ยังต้องวัดในบริบทของการใช้งานด้วย

**5. อภิปรายผล**

องค์การขนส่งมวลชนกรุงเทพ (ขสมก.) มีช่องทางในการยื่นข้อร้องเรียนรถโดยสารสาธารณะผ่านเว็บบอร์ด ที่ผู้ใช้งานสามารถตั้งกระทู้เพื่อแสดงความคิดเห็นได้อย่างอิสระ ดังนั้น ข้อมูลบนเว็บบอร์ดจึงนับว่ามีประโยชน์ และมีบทบาทในการเพิ่มประสิทธิภาพการให้บริการ แต่จำนวนข้อร้องเรียนที่เพิ่มมากขึ้นและความหลากหลายของข้อความ ตลอดจนความผิดพลาดในการใช้ภาษาของผู้ใช้ส่งผลต่อความถูกต้องในการจำแนกประเภทข้อร้องเรียน เนื่องจากผู้รับผิดชอบต้องวิเคราะห์ข้อมูลด้วยตนเอง ดังนั้นหากกระบวนการการจำแนกข้อร้องเรียนแบบอัตโนมัติจะเป็นแนวทางหนึ่งในการแก้ไขปัญหาดังกล่าว งานวิจัยชิ้นนี้มีวัตถุประสงค์ เพื่อออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึกจากการเก็บรวบรวมข้อมูลข้อร้องเรียนรถโดยสารสาธารณะบนเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพ จำนวน 1,275 ข้อความ แบ่งเป็นชุดข้อมูลฝึกฝน เพื่อสร้างโมเดลจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึกด้วย deepcut และชุดข้อมูลทดสอบ เพื่อประเมินผลความถูกต้องของผลการจำแนกข้อร้องเรียนการให้บริการด้วยการติดแท็กปัญหาในการให้บริการ คิดเป็นร้อยละ 80:20 เตรียมข้อมูลก่อน

จะนำไปประมวลผลจำเป็นต้องจัดการกับข้อมูลที่ไม่เป็นระเบียบ เพื่อกำหนดรูปแบบโครงสร้างของข้อมูลให้อยู่ในรูปแบบที่เหมาะสมต่อการนำไปประมวลผล จากนั้นแบ่งคำภาษาไทยแบบแบ่งคำตามการเรียนรู้เชิงลึกด้วยอัลกอริทึม deepcut เพื่อตัดหัวข้อร้องเรียนรถโดยสารสาธารณะให้อยู่ในรูปคำเดี่ยว จากนั้นจะแปลงข้อมูลทั้งหมดไปอยู่ในรูปลิสต์ของคำในลักษณะถุงคำ สำหรับนำมาสร้างดัชนีคำศัพท์ ในการแปลงเอกสารที่เป็นภาษาธรรมชาติ ให้อยู่ในรูปของเวกเตอร์เอกสาร ที่สามารถคำนวณค่าน้ำหนักคำสำคัญสำหรับการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ และแก้ปัญหาข้อมูลที่มีความซ้ำซ้อนกัน อันเนื่องมาจากจำนวนมิติข้อมูลจำนวนมากที่ทับซ้อนกันในชุดข้อมูล ด้วยการลดมิติของข้อมูลด้วย t-SNE เพื่อสร้างความน่าจะเป็นในการป้องกันการกระจุกตัวหรือทับซ้อนของจุดข้อมูลในมิติต่ำ ซึ่งเป็นผลมาจากปัญหามิติของข้อมูล นำผลการกระจายตัวของกลุ่มที่มีความคล้ายคลึงกันมาสร้างคลาสที่เกี่ยวข้องกับการร้องเรียนตามบริบทการให้บริการรถโดยสารสาธารณะ แบ่งออกเป็น 7 คลาส ได้แก่ 1) คลาสการขับขี่ 2) คลาสการจอดรับส่ง 3) คลาสผู้ให้บริการ 4) คลาสการเดินรถ 5) คลาสยานพาหนะ 6) คลาสอุปกรณ์ให้บริการ และ 7) คลาสมาตรการป้องกันโรคระบาด ผลการออกแบบและพัฒนาการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึกด้วยการทำนายประเภทจากการจำแนกข้อมูลแบบไบนารีด้วยเทคนิคการวิเคราะห์การถดถอยโลจิสติกส์ตรวจสอบแบบไขว้กัน 10 ชุด พบว่า ผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะทั้งหมดมีความถูกต้องมากกว่าร้อยละ 90 อย่างไรก็ดี ค่าความถูกต้องในคลาสผู้ให้บริการ มีความถูกต้องน้อยที่สุด เพียงร้อยละ 83 ตามจำนวนข้อร้องเรียนในคลาสที่มีจำนวนมากที่สุด ส่งผลให้เกิดความหลากหลายของข้อมูลที่แตกต่างกันตามบริบทของผู้ใช้ และการใช้คำผิดทั้งจากความตั้งใจหรือไม่ตั้งใจจากผู้ใช้ที่ส่งผลต่อความถูกต้องในการจำแนกข้อความ

ผลการประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะจากโมเดลที่ฝึกฝนด้วยการถดถอยโลจิสติกส์มาเป็นแบบเรียนรู้เพื่อพยากรณ์ข้อร้องเรียนรถโดยสารอัตโนมัติ เปรียบเทียบค่าความถูกต้องกับอัลกอริทึม fastText พบว่า ค่าความถูกต้องของทั้งสองอัลกอริทึมมีค่าความถูกต้องอยู่ในระดับสูง และเป็นไปในทิศทางเดียวกัน เมื่อประเมินผลความถูกต้องจากการติดแท็กปัญหาการให้บริการ พบว่าอยู่ในระดับดีมาก แสดงถึงการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยการเรียนรู้เชิงลึก ให้ผลลัพธ์ความถูกต้องในระดับสูง โดยเฉพาะข้อร้องเรียนแบบ 1 ประเด็นต่อ 1 ข้อร้องเรียน เนื่องจากเหมาะกับคำศัพท์ที่บริบทสามารถกำหนดขอบเขตได้แน่นอน ในทางกลับกัน ปัญหาที่พบส่วนใหญ่ เนื่องมาจากมีคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส เช่น เสียงดัง ที่จำเป็นต้องดูบริบทของข้อความประกอบ เนื่องจากเสียงดังเป็นกิริยาที่อาจเกิดจากบุคคลในคลาสการขับขี่ หรืออาจเป็นเสียงที่เกิดจากเครื่องยนต์จากคลาสยานพาหนะ หรือคลาสอุปกรณ์ให้บริการก็เป็นได้ดังนั้นการวัดความถูกต้อง อย่างไรก็ดี การตัดคำภาษาไทยจัดว่าเป็น NP-hard Problem เพราะไม่มีผิดถูกชัดเจน เช่น “ตากลม” สามารถตัดได้เป็น “ตาก | ลม” หรือ “ตา | กลม” เป็นต้น ดังนั้นการวัดความถูกต้องนอกจากจะวัดในเชิงความหมายแล้ว ยังต้องวัดในบริบทของการใช้งานด้วย

## 6. ข้อเสนอแนะ

งานวิจัยในครั้งต่อไปควรให้ความสำคัญกับประสิทธิภาพของอัลกอริทึมในการตัดคำภาษาไทย เนื่องจากภาษาไทยมีความซับซ้อน และมีลักษณะเฉพาะตัวแตกต่างจากภาษาอื่น หากแต่งานวิจัยเกี่ยวกับภาษาไทยยังมีไม่มาก และขาดคลังคำศัพท์ (Corpus) ขนาดใหญ่ที่มีคุณภาพ ปัจจุบันการตัดคำภาษาไทยส่วนใหญ่ที่ใช้กันเรียนรู้จากคลังข้อมูลของ BEST Corpus ที่ประกอบไปด้วยคำ 5 ล้านคำ และมีหมวดหมู่ต่าง ๆ เช่น บทความวิชาการ, สารานุกรม, ข่าว และนวนิยาย เป็นต้น แต่ปัญหาคือโมเดลของการตัดคำที่ใช้ในปัจจุบันยังไม่ได้ออกแบบมาให้รองรับหมวดหมู่ใหม่ หรือคำใหม่ที่เกิดขึ้นในปัจจุบัน เช่น คำที่ปรากฏในสื่อสังคมออนไลน์ หรือคำศัพท์แสลง เป็นต้น ดังนั้นแนวทางการพัฒนาการตัดคำภาษาไทยอาจไม่จำเป็นต้องเริ่มต้นจากโมเดลที่ซับซ้อน แต่ทำได้ด้วยการเพิ่มหมวดหมู่ใหม่ และเพิ่มจำนวนคำให้สมบูรณ์มากขึ้น เพื่อเพิ่มคุณภาพของข้อมูลก่อนนำเข้ากระบวนการประมวลผลภาษาธรรมชาติ

## เอกสารอ้างอิง


Aggarwal, C. C. (2015). **Data classification algorithms and applications.** Boca Raton, FL, USA: CRC Press.

Albon, C. (2018). **Machine learning with python cookbook practical solutions from pre-processing to deep learning.** Sebastopol, CA, USA: O'Reilly Media Inc.

Bangkok Mass Transit Authority. (2019). **Annual report 2019.** (In Thai), Bangkok: Bangkok Mass Transit Authority, Bangkok.

Bhattacharjee, J. (2018). **Fasttext quick start guide get started with facebook's library for text representation and classification.** Birmingham, UK: Packt Publishing.

Jo, T., (2019). **Text mining concepts, implementation, and big data challenge.** Cham, Switzerland: Springer International Publishing.

Kleinbaum, D. G. & Klein, M. (2010). **Logistic regression a self□learning text.** 3rd ed. New York, USA: Springer Science+Business Media LLC.

Manning, C.D., Raghavan, P., & Schütze, H., (2008). **Introduction to information retrieval.** Cambridge, UK: Cambridge University Press.

McConnell, S. (1996). **Rapid development: taming wild software schedules.** Washington, DC, USA: Microsoft Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. **proceedings of the 26th international conference on neural information processing systems.** (3111-3119). New York, USA: Curran Associates.

Mikolov, T., Corrado, G., Chen, K. & Dean, J. (2013). **Efficient estimation of word representa-tions in vector space. proceedings of the international conference on learning representations.** (1-12). Scottsdale, Arizona, USA: arXiv.

Mitchell, R. (2015). **Web scraping with python: collecting data from the modern web.** Sebastopol, CA, USA: O'Reilly Media Inc.

Office of Transport and Traffic Policy and Planning. (2018). **The number of passengers on the BMTA bus.** (In Thai). Retrieved 20 August 2021, from http://mistran.otp.go.th/mis/Interview_HIPubilcBus.aspx.

Raschka, S. & Mirjalili, V. (2017). **Python machine learning machine learning and deep learning with python, scikit-learn, and tensorflow.** 2nd ed. Birmingham, UK: Packt Publishing.

Rosebrock, A. (2017). **Deep learning for computer vision with python: starter bundle.** n.p., USA: PyImageSearch.

Scikit-learn developers (BSD License). (2020). **Sklearn.manifold.TSNE.** from https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html. Re-trieved 5 August 2021.

Tapsai, C., Unger, H., & Meesad, P., (2021). **Thai natural language processing word seg-mentation, semantic analysis, and application.** Cham, Switzerland: Springer Inter-national Publishing.

van der Maate, L. & Hinto, H. (2008). Visualizing data using t-SNE. **Journal of Machine Learning Research, 9**, 2579-2605.

Vasilev, I., Slater, D., Spacagna, G., Roelants, P. & Zocca, V. (2019). **Python deep learning exploring deep learning techniques and neural network architectures with pytorch, keras, and tensorflow.** 2nd ed. Birmingham, UK: Packt Publishing.

Verma, A. & Ramanayya, T.V. (2015). **Public transport planning and management in de-veloping countries.** Boca Raton, FL, USA: CRC Press.

Wozniak, M. (2014). **Hybrid classifiers methods of data, knowledge, and classifier combination.** Heidelberg, Germany: Springer-Verlag.

Zhai, C. & Massung, S. (2016). **Text data management and analysis: a practical introduction to information retrieval and text mining.** New York, USA: ACM Books.

Zizka, J., Darena, F., & Svoboda, A., (2020). **Text mining with machine learning principles and techniques.** Boca Raton, FL, USA: CRC Press.

Zong, C., Xia, R. & Zhang, J. (2021). **Text data mining.** Tsinghua, Beijing, China: Tsinghua University Press.