การออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะเพื่อติดแท็กปัญหาการให้บริการ

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

[1]สาขาวิชาเทคโนโลยีสารสนเทศ สำนักวิชาเทคโนโลยีสังคม มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา, ประเทศไทย*
[2]โครงการจัดรูปแบบการบริหารวิชาการด้านเทคโนโลยีดิจิทัลรูปแบบใหม่มหาวิทยาลัยเทคโนโลยีสุรนารี, นครราชสีมา, ประเทศไทย
[3]สาขาวิชาระบบสารสนเทศคอมพิวเตอร์ คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล, นครราชสีมา, ประเทศไทย

*ผู้ประพันธ์บรรณกิจ อีเมล: d6200701@g.sut.ac.th

รับต้นฉบับ: 28 สิงหาคม 2564; รับบทความฉบับแก้ไข: 13 กันยายน 2564; ตอบรับบทความ: 11 ตุลาคม 2564
เผยแพร่ออนไลน์: 13 ธันวาคม 2564

บทคัดย่อ

องค์การขนส่งมวลชนกรุงเทพ (ขสมก.) มีช่องทางในการร้องเรียนรถโดยสารสาธารณะผ่านเว็บบอร์ด ที่ผู้ใช้งานสามารถแสดงความคิดเห็นได้อย่างอิสระ ดังนั้นข้อมูลบนเว็บบอร์ดจึงนับว่ามีประโยชน์ และมีบทบาทในการเพิ่มประสิทธิภาพการให้บริการ เนื่องจากมาจากผู้ใช้บริการจริง แต่จำนวนข้อร้องเรียนที่เพิ่มมากขึ้น และความหลากหลายของข้อความ ตลอดจนความผิดพลาดในการใช้ภาษาของผู้ใช้ส่งผลต่อความถูกต้องในการจำแนกประเภทข้อร้องเรียน เนื่องจากผู้รับผิดชอบต้องวิเคราะห์ข้อมูลด้วยตนเอง ดังนั้นหากมีกระบวนการจำแนกข้อร้องเรียนแบบอัตโนมัติจะเป็นแนวทางหนึ่งในการแก้ไขปัญหาดังกล่าว งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อ 1) ออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ จากข้อร้องเรียนผ่านเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพด้วยกระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม แล้วคัดเลือกคำศัพท์ด้วยการวิเคราะห์น้ำหนักของคำ มาสร้างเป็นคลังคำศัพท์ แบ่งเป็น 4 คลาส ได้แก่ คลาสการขับขี่ คลาสผู้ขับขี่และพนักงานผู้ให้บริการ คลาสยานพาหนะและอุปกรณ์ให้บริการ และคลาสเวลาและการเดินรถ วัดผลด้วยการเรียนรู้ของเครื่องเพื่อจำแนกข้อความ พบว่า อยู่ในระดับดีมาก 2) ประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ จากการจับคู่กับคำศัพท์ในคลังคำศัพท์กับข้อร้องเรียนจากผู้ใช้เพิ่มความถูกต้องของผลลัพธ์ด้วยการเปรียบเทียบความคล้ายคลึงของคำด้วยเทคนิคการวัดระยะทางเลเวนชเตย์น ในกรณีที่พบคำศัพท์ที่เขียนไม่ถูกต้องตามที่กำหนดไว้ในคลังคำศัพท์ เพื่อติดแท็กจำแนกปัญหาในการให้บริการของรถโดยสารสาธารณะ ประเมินความถูกต้องโดยผู้เชี่ยวชาญ ผลการประเมินอยู่ในระดับดีมาก โดยเฉพาะข้อร้องเรียนประเด็นเดียว แสดงถึงการตัดคำภาษาไทยแบบอิงพจนานุกรมเหมาะกับคลังคำศัพท์ที่กำหนดขอบเขตได้แน่นอน ความถูกต้องและครอบคลุมของคำศัพท์ส่งผลโดยตรงต่อผลลัพธ์ในการจำแนก อย่างไรก็ดี พบปัญหาคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส เช่น เสียงดัง อาจมีที่มาจากคลาสผู้ขับขี่และพนักงานผู้ให้บริการ หรือจากคลาสยานพาหนะและอุปกรณ์ให้บริการก็ได้ตามต้นกำเนิดการเกิดเสียง ตลอดจนปัญหาการประสมคำในภาษาไทยที่มีลักษณะเฉพาะตัวส่งผลต่อความถูกต้องของผลการตัดคำ ผลลัพธ์จากงานวิจัยจะนำเสนอแนวทางการจำแนกข้อมูล ก่อนจะนำเสนอสารสนเทศเป็นภาพข้อมูลต่อไป และเป็นประโยชน์แก่ผู้รับผิดชอบ เพื่อนำไปปรับปรุงการให้บริการให้เหมาะสมกับความต้องการของผู้ใช้บริการต่อไป

คำสำคัญ : การจำแนกข้อมูล รถโดยสารสาธารณะ การตัดคำภาษาไทย การเรียนรู้ของเครื่อง ระยะทางเลเวนชเตย์น

Journal of Engineering and Digital Technology (JEDT)
Vol.9 No.2 July - December 2021

The Design and Development of Public Bus Complaint Classification Process for Service Problem Tagging

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

[1*]*School of Information Technology, Institute of Social Technology, Suranaree University of Technology, Nakhon Ratchasima, Thailand*

[2]*A New Paradigm in the Digital Technology Academic Administrator Project (DIGITECH), Suranaree University of Technology, Nakhon Ratchasima, Thailand*

[3]*Department of Computer Information System, Faculty of Business Administrator, Vongchavalitkul University, Nakhon Ratchasima, Thailand*

*Corresponding Author. E-mail address: d6200701@g.sut.ac.th

Received: 28 August 2021; Revised: 13 September 2021; Accepted: 11 October 2021
Published online: 13 December 2021

Abstract

Bangkok Mass Transit Authority (BMTA) has a webboard for service user complaints on public buses. This information is important and improving service efficiency. Now, the complaints are increasing, more different and lexical error problem affect the accuracy of the complaints classification. Because the person has to analyze the data by self. This research aims 1) to the design and development of public bus complaint classification process from BMTA webboard by Thai dictionary-based approach and keywords selection by TF-IDF to create a corpus-based divided into 4 classes: driving class, person class, vehicle class and schedule class. The result of text classification algorithms, found that it was at a very good level. 2) to evaluation the accuracy of the public bus complaint classification. Then, keywords matching with user complaints again and increase the accuracy of keywords by Levenshtein Distance. In case that found the incorrectly keywords for create complaint classification tag. The accuracy assessment of complaint classification is a very good level, especially one complaint issue. The results show performance of dictionary-based approach on Thai word segmentation is suitable for the terminology that has definite scope. However, found problems in same keywords to be duplicated in some class. For example, loudness can be attribute from the person class or the vehicle-equipment class according to the origin of the sound. As well as the unique compounding of Thai language affects to the accuracy of word tokenization. These results will present an approach to data classification and will benefit to those responsible to improve the service for customers.

Keywords: Data classification, Public bus, Thai word segmentation, Machine learning, Levenshtein distance

1) บทนำ

ระบบขนส่งสาธารณะถือว่ามีบทบาทสำคัญอย่างมากต่อการดำเนินชีวิตประจำวันตั้งแต่อดีตจนถึงปัจจุบัน ประกอบด้วยการขนส่งสาธารณะทางถนน ทางราง และทางน้ำ ที่ครอบคลุมทุกพื้นที่ทั่วประเทศ ซึ่งระบบขนส่งสาธารณะที่เข้าถึงและถูกใช้บริการมากที่สุดคือระบบขนส่งสาธารณะทางถนน โดยเฉพาะพื้นที่กรุงเทพมหานครและปริมณฑลที่เดินทางด้วยรถโดยสารประจำทางเป็นหลัก ภายใต้การดำเนินงานขององค์การขนส่งมวลชนกรุงเทพ (ขสมก.) [1] มีจำนวนเส้นทางรถโดยสารประจำทางรวมทุกประเภท จำนวน 456 เส้นทาง ครอบคลุมรถองค์กร รถเอกชนร่วมบริการ รถเล็กวิ่งในซอย รถตู้โดยสาร และรถตู้เชื่อมต่อท่าอากาศยานสุวรรณภูมิ [2] จากคุณสมบัติการให้บริการที่มีความคล่องตัวสูง สะดวก และสามารถให้บริการได้ทุกจุดตลอดระยะของการเดินทาง ตลอดจนอัตราค่าโดยสารยังอยู่ในระดับต่ำสอดรับกับรายได้ของประชากรมากที่สุด เมื่อเทียบระบบขนส่งมวลชนทางบกประเภทอื่น ๆ ในพื้นที่เดียวกัน แต่หากพิจารณาย้อนหลัง พบว่า มีจำนวนผู้ใช้บริการรถโดยสารประจำทางขสมก. ลดลงอย่างต่อเนื่องจากปี 2558 อยู่ที่ประมาณ 1,060,000 คนต่อวัน แต่ในปี 2560 กลับพบว่ามีผู้ใช้บริการเพียง 830,859 คนต่อวันเท่านั้น ลดลงเฉลี่ยร้อยละ 6.15 ต่อปี [3] ซึ่งหนึ่งในปัจจัยหลักมาจากคุณภาพการให้บริการ อันจะเห็นได้จากผลการสำรวจของสำนักงานปลัดสำนักนายกรัฐมนตรี [4] เกี่ยวกับจำนวนเรื่องร้องทุกข์ ในช่วงปีงบประมาณ 2563 ไตรมาสที่ 1 พบว่า องค์การขนส่งมวลชนกรุงเทพเป็นหน่วยงานรัฐวิสาหกิจในสังกัดกระทรวงคมนาคมที่มีประชาชนร้องเรียนโดยเฉลี่ยมากที่สุดเป็นอันดับหนึ่ง โดยส่วนใหญ่เป็นเรื่องร้องทุกข์/เสนอความคิดเห็น ใน 3 ประเด็น ได้แก่ 1) ขอให้เพิ่มเที่ยว/รอบการเดินรถขยายเส้นทางการเดินรถและเพิ่มจำนวนรถโดยสารประจำทาง 2) ขอให้ปรับปรุงการให้บริการของพนักงานขับรถและพนักงานเก็บค่าโดยสาร และ 3) ขอให้ปรับปรุงการให้บริการของรถโดยสารประจำทาง รถโดยสารสาธารณะ รถโดยสารปรับอากาศประจำทาง และรถโดยสารปรับอากาศร่วมบริการสอดคล้องกับสถิติการรับเรื่องร้องเรียนรถโดยสารประจำทางขสมก.ตามกฎหมายว่าด้วยการขนส่งทางบก ปี 2563 ที่พบว่า 3 อันดับเรื่องร้องเรียน ได้แก่ ไม่หยุดรับส่งผู้โดยสารที่ป้าย ขับรถประมาทหวาดเสียว และสภาพรถไม่สมบูรณ์ (ควันดำ) [5]

นอกเหนือจากข้อร้องเรียนดังกล่าวที่ดำเนินการร้องทุกข์ผ่านช่องต่าง ๆ ของรัฐบาลแล้ว องค์การขนส่งมวลชนกรุงเทพเองยังมีช่องทางในการรับแสดงความคิดเห็นหรือรับเรื่องราวร้องเรียนรถองค์การ และร้องเรียนรถเอกชนร่วมบริการผ่านเว็บบอร์ด บนเว็บไซต์ http://www.bmta.co.th เป็นอีกหนึ่งช่องทางที่ใช้ในการติดตามตรวจสอบการให้บริการ และเป็นสื่อกลางในการแลกเปลี่ยนความคิดเห็นต่าง ๆ ที่ผู้ใช้งานสามารถตั้งกระทู้ถามตอบเพื่อแลกเปลี่ยนความคิดเห็นกันได้อย่างอิสระ ดังนั้นข้อมูลบนเว็บบอร์ดจึงนับว่ามีประโยชน์ และมีบทบาทในการเพิ่มประสิทธิภาพในการให้บริการ แต่เมื่อเวลาผ่านไปจำนวนความคิดเห็นหรือข้อร้องเรียนในการใช้บริการมีเพิ่มมากขึ้นในแต่ละวัน และความหลากหลายของข้อความที่แตกต่างกันตามบริบทของผู้ใช้งาน ตลอดจนความผิดพลาดในการใช้ภาษาที่เกิดจากความตั้งใจหรือแม้แต่ไม่ตั้งใจของผู้ใช้ ที่ก่อให้เกิดปัญหาในการตีความหมาย โดยเฉพาะอย่างยิ่งข้อความที่ไม่มีการจำแนกหมวดหมู่ไว้อย่างชัดเจน อาจส่งผลต่อการตอบคำถาม การให้ข้อมูลคืนที่ถูกต้องแก่ผู้ใช้ ตลอดจนการสรุปสถิติและการจัดประเภทข้อร้องเรียนนั้นทำได้ยากใช้เวลานาน และมีโอกาสเกิดความผิดพลาดสูง เนื่องจากผู้ดูแลระบบจะต้องนำข้อร้องเรียนมาวิเคราะห์เพื่อจำแนกหมวดหมู่ด้วยตนเอง ดังนั้นหากมีกระบวนการจำแนกข้อร้องเรียนตามหมวดหมู่ที่ต้องการจะเป็นแนวทางหนึ่งในการแก้ไขปัญหาดังกล่าว งานวิจัยชิ้นนี้มุ่งออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะบนเว็บบอร์ด เพื่อจำแนกปัญหาการให้บริการด้วยการติดแท็กอัตโนมัติ และนำเสนอเป็นสารสนเทศแก่ผู้รับผิดชอบในการนำไปปรับปรุงคุณภาพการให้บริการให้สอดคล้องกับความต้องการของผู้ใช้บริการต่อไป

การจำแนกข้อมูล (Classification) [6] เป็นเทคนิคการสร้างโมเดลหรือตัวจำแนกข้อมูล (Classifier) เพื่อทำนายหมวดหมู่ของข้อมูล (Categories/Class) โดยชุดข้อมูลที่เป็นอินพุตสำหรับสร้างตัวจำแนกข้อมูลเรียกว่าชุดข้อมูลฝึกสอน (Training Data) หากในชุดข้อมูลมีแอททริบิวต์หมวดหมู่ข้อมูลสำหรับจำแนกจะเรียกว่าการเรียนรู้แบบมีผู้สอน (Supervised Learning) ที่สร้างตัวจำแนกข้อมูลจะถูกสอนโดยแอททริบิวต์หมวดหมู่ข้อมูลต่าง ๆ ตรงข้ามกับการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning หรือ Clustering) ที่จะไม่ทราบถึงหมวดหมู่ของข้อมูล

ปัจจุบันมีผู้เสนอเทคนิคต่าง ๆ ได้แก่ 1) การจำแนกข้อมูลด้วยต้นไม้ตัดสินใจ (Decision Tree Classifier) [7] เป็นกระบวนการสร้างต้นไม้ขึ้นเพื่อใช้ในการตัดสินใจจากข้อมูลที่มีหมวดหมู่ข้อมูลระบุอยู่ 2) การจำแนกข้อมูลแบบเบย์เซียน

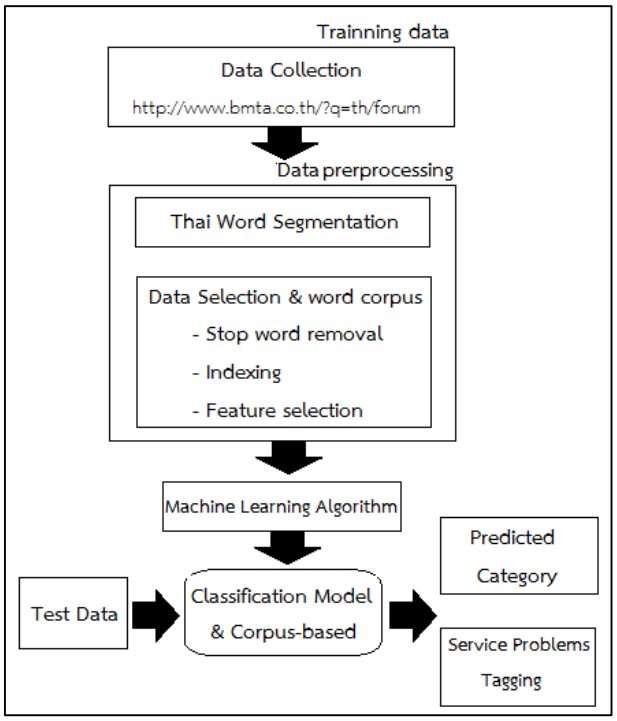
(Bayesian Classifier) [7] เป็นการสร้างตัวจำแนกข้อมูลด้วยการประยุกต์ใช้ค่าทางสถิติที่สามารถบ่งบอกถึงความน่าจะเป็นของข้อมูลเรคคอร์ดหนึ่งที่จะอยู่ในหมวดหมู่ของข้อมูลหนึ่ง ๆ โดยประยุกต์ใช้ทฤษฎีของเบย์ (Bayes' Theorem) ที่ช่วยให้สามารถจำแนกข้อมูลได้อย่างรวดเร็วและมีความแม่นยำสูง 3) การจำแนกข้อมูลด้วยฐานกฎ (Rule-based Classifier) [7] เป็นโมเดลที่จะแสดงผลด้วยเซตของกฎที่มีลักษณะแบบ IF-THEN โดยแต่ละกฎจะอยู่ในรูปฟอร์ม 4) การจำแนกข้อมูลจากกฎความสัมพันธ์ของข้อมูล (Association Rule Classifier) [6] มีที่มาจากแนวคิดกฎความสัมพันธ์ของข้อมูลที่ประกอบด้วยรายการ (Item) ต่าง ๆ ที่ปรากฏร่วมกันบ่อย ๆ เพื่อใช้อธิบายถึงรูปแบบที่แอบแฝงอยู่ในชุดข้อมูล 5) การค้นหาเพื่อนบ้านใกล้สุด k อันดับ (k-nearest neighbor: K-NN Classifier) [6] เป็นการเรียนรู้โดยการเปรียบเทียบกันระหว่างข้อมูลที่ต้องการจำแนกทั้งหมดในชุดข้อมูลฝึกสอนที่มีลักษณะเหมือนกันหรือใกล้เคียงกัน ด้วยการพิจารณาข้อมูล แอททริบิวต์ต่าง ๆ โดยจะมีข้อมูลเรคคอร์ดหนึ่งที่ถูกกำหนดเป็นจุดหนึ่งในระนาบ พิจารณาจากค่า Euclidean Distance 7) ซัพพอร์ทเวกเตอร์แมชชีน (support vector machine: SVM) [7] เป็นอัลกอริทึมในการวิเคราะห์ข้อมูลและจำแนกข้อมูล โดยอาศัยหลักการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการฝึกฝนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูลที่ดีที่สุด และ 8) การจำแนกข้อมูลด้วยโครงข่ายประสาทเทียม (Neural Network Classifier) [6], [7] ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ โดยเชื่อมต่อกันเป็นโหนด (Node) หลายชั้นเรียกว่าโครงข่ายประสาทเทียมแบบหลายชั้น (multilayer perceptron: MLP) ค่าฟังก์ชันของแต่ละโหนดเกิดจากฝึกฝน (Train) กระบวนการฝึกฝน เริ่มจากการป้อนอินพุต (Input Signal) เข้าโครงข่าย คำนวณค่าอินพุตคูณกับค่าน้ำหนักในโครงข่าย แล้วปรับค่าน้ำหนักประสาท (Weight) และไบอัส (Bias) ตามกฎการเรียนรู้เพื่อให้เอาต์พุตของโครงข่ายให้ค่าผลลัพธ์ใกล้เคียงเป้าหมายมากที่สุด แล้วส่งผ่านข้อมูลสู่ฟังก์ชันกระตุ้นและส่งออกเป็นเอาต์พุต (Output Signal) เพื่อเปรียบเทียบกับค่าเป้าหมายซึ่งเป็นค่าผิดพลาด และจะถูกส่งค่าย้อนกลับไปปรับค่าน้ำหนัก (Backpropagation Algorithm) ให้พอดี (Fit) กับข้อมูลต่อไป ด้วยจุดเด่นคือมีความมั่นคงสูงต่อสิ่งรบกวน (Noise) ทำงานได้ดีกับข้อมูลอินพุตและเอาต์พุตที่มีค่าแบบไม่ต่อเนื่อง อีกทั้งสามารถจำแนกข้อมูลได้ทั้งแบบมีผู้สอนและไม่มีผู้สอน รวมถึงมีประสิทธิภาพในการจำแนกข้อมูล แม้พบความสัมพันธ์ระหว่างแอททริบิวต์กับหมวดหมู่ต่าง ๆ เพียงเล็กน้อย จากข้อดีดังกล่าวจึงถูกนำมาถูกประยุกต์ใช้อย่างแพร่หลาย

2) วัตถุประสงค์การวิจัย

1. เพื่อออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ
2. เพื่อประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

3) วิธีดำเนินการวิจัย

งานวิจัยชิ้นนี้เป็นงานวิจัยเชิงประยุกต์ โดยใช้แนวคิดแบบจำลองการพัฒนาแอปพลิเคชันแบบเร่งรัด (Rapid Application Development Model: RAD) [8] มาเป็นกรอบในการดำเนินงาน ดังรูปที่ 1



รูปที่ 1 : กรอบการดำเนินงาน

3.1) มอดูลการเก็บรวบรวมข้อมูล (Data Collection)

งานวิจัยชิ้นนี้เก็บรวบรวมข้อร้องเรียนรถโดยสารสาธารณะขององค์การขนส่งมวลชนกรุงเทพที่เผยแพร่เป็นสาธารณะบนเว็บบอร์ด http://www.bmta.co.th/?q=th/forum ด้วย

เทคนิคการขูดเว็บ (Web Scraping) [9] ด้วยการเขียนโปรแกรมภาษา R ในการเข้าถึงเว็บไซต์ และคัดลอกข้อมูลในตำแหน่งที่ระบุมาจัดเก็บในรูปของไฟล์เอกสารธรรมดา (.txt) ที่รอการนำไปประมวลผล อย่างไรก็ดี ต้องระวังมิให้สร้างภาระในการทำงานของเครื่องแม่ข่าย จนอาจก่อให้เกิดความเสียหาย และการอนุญาตให้เข้าถึงและคัดลอกข้อมูลบนเว็บไซต์ในหมวดร้องเรียนรถองค์การที่ตั้งกระทู้ขึ้นภายในวันที่ 1 มกราคม 2564 ถึง 31 กรกฎาคม 2564 รวมทั้งสิ้น 1,275 ข้อความ ดังตารางที่ 1 แยกเป็นส่วนที่ 1 นำมาตัดคำภาษาไทยแบบ อิงพจนานุกรมเพื่อสร้างคลังคำศัพท์จำนวน 1,020 ข้อความ คิดเป็นร้อยละ 80 และส่วนที่ 2 จำนวน 255 ข้อความนำมาเป็นข้อมูลทดสอบ คิดเป็นร้อยละ 20 เพื่อประเมินผลความถูกต้องของผลการจำแนกปัญหาการให้บริการ

ตารางที่ 1 : ตัวอย่างข้อร้องเรียนรถโดยสารสาธารณะ

ลำดับ	ข้อความ
1.	บัตร Prompt Card เติมเงินแล้วใช้ไม่ได้ เครื่องไม่อ่านแตะบัตรแล้วเออเร่อ เพิ่งเติมเงินด้วยค่ะ ต้องทำอย่างไรบ้างคะ ลองไปติดต่อที่ธนาคารกรุงไทยบอกแก้ไขอะไรไม่ได้ เพราะบัตรไม่มีข้อมูล แสดงว่าถ้าเติมเงินแล้วบัตรเกิดใช้งานไม่ได้ ก็ต้องเสียเงินไปฟรีๆ เลยหรอคะ รบกวนนตรวจสอบด้วยค่ะ ขอบคุณค่ะ
2.	วันที่ 29 ธันวาคม 2563 เวลา 17.40 น ป้าย ถนนสายลวดรถสาย 25 11-9321 3-40252 ขับเร็ว วิ่งขา ไม่จอดรับผู้โดยสาร
...	...
1275.	ยืนรอรถเมล์สาย205แถวแบงค์กรุงศรี รอตั้ง 19.45 จนถึง 20.30 รถเมล์ไปเดอะมอลล์ท่าพระไม่มีเลยในช่วงเร่งรีบแบบนี้ทำไมถึงไม่มีรถ

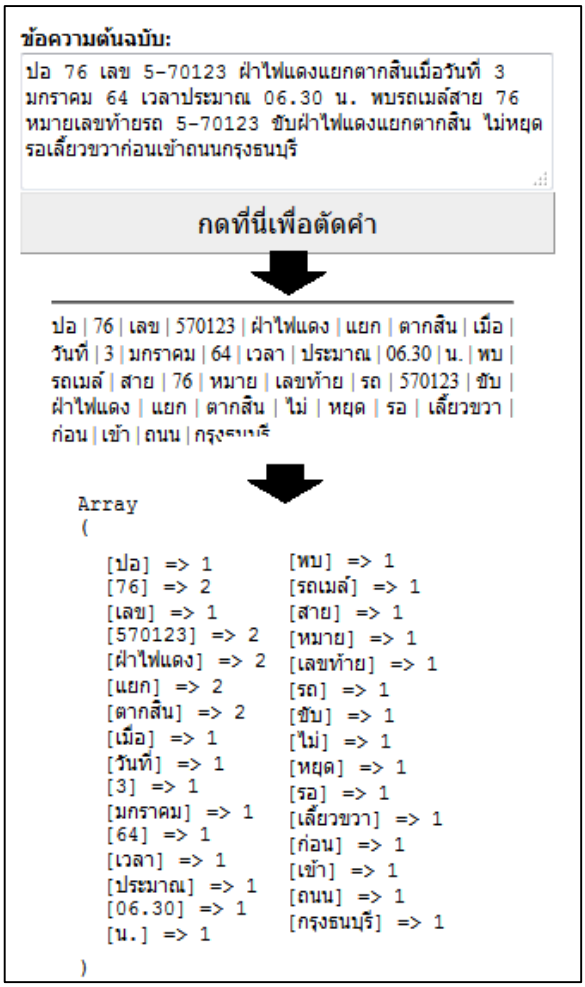
3.2) มอดูลการตัดคำภาษาไทย (Thai Word Segmentation)

เป็นหนึ่งในขั้นตอนที่สำคัญของการวิเคราะห์ข้อความ โดยนำข้อความทั่วไปซึ่งอยู่ในรูปแบบประโยคมาแบ่งออกเป็นคำหรือคุณลักษณะ (Term/Feature) เพื่อแยกส่วนของข้อความออกจากกันก่อนนำไปประมวลผลในขั้นต่อไป [10] แบ่งตามกระบวนการทำงานออกเป็น 3 กลุ่ม ได้แก่ 1) การตัดคำโดยใช้กฎ (Rule-Based Approach) โดยใช้วิธีเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร 2) การตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) ที่เก็บคำศัพท์ไว้ในพจนานุกรม แล้วนำข้อความป้อนเข้าไปค้นหาและเปรียบเทียบสายอักขระกับคำศัพท์ในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใด และประกอบด้วยคำใดบ้าง อย่างไรก็ดี การตัดคำโดยใช้พจนานุกรมก็มีข้อจำกัดบางประการ เนื่องจากมีความเป็นไปได้ที่คำที่ปรากฏในเอกสาร อาจจะไม่ปรากฏในพจนานุกรม จึงเป็นที่มาของเอ็นแกรม (N-gram) ที่นำบางส่วนของข้อความออกมาเป็นตามค่า N และ 3) การตัดคำโดยใช้คลังคำศัพท์ (Corpus-Based Approach) โดยเตรียมคลังคำศัพท์ที่มีการตัดคำและการกำกับหน้าที่ของคำไว้ล่วงหน้า

อย่างไรก็ดี การตัดคำในภาษาไทยยังพบปัญหาในการตัดคำเนื่องจากลักษณะของภาษาไทยมีการเขียนติดกันแบบไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำที่ชัดเจน แตกต่างจากภาษาอังกฤษที่มีช่องว่างแสดงให้เห็นถึงขอบเขตของแต่ละคำ [10] ปัจจุบันมีผู้คิดค้นและเสนอแนะวิธีการต่าง ๆ ในการตัดคำภาษาไทย เพื่อเพิ่มประสิทธิภาพในการตัดคำโดยใช้พจนานุกรมหรือคลังคำศัพท์ที่กำหนดไว้ล่วงหน้า ได้แก่ 1) การตัดคำแบบยาวที่สุด (Longest Matching) หลักการทำงานเริ่มจากตัวอักษรซ้ายสุดของข้อความไปยังตัวอักษรถัดไปจนกว่าจะพบคำที่กำหนดไว้ จากนั้นจะค้นหาคำถัดไปจนกว่าจะจบข้อความ ในกรณีที่พบคำที่มาจากจุดเริ่มต้นเดียวกันจะเลือกคำที่ยาวที่สุด เช่น ประโยค “ฉันนั่งตากลมที่หน้าบ้าน” จะเริ่มจากตัวอักษร “ฉ” และคำแรกที่แบ่งได้คือ “ฉัน” หลังจากนั้นก็ค้นหาตัวอักษรถัดไปและนำมาเปรียบเทียบคำในพจนานุกรมก็จะพบคำว่า “นั่ง” เป็นคำต่อไป ตัวอักษรถัดไปคือ “ต” จากตัวอักษรนี้จะได้คำว่า “ตา” กับ “ตาก” แนวคิดนี้ให้เลือกคำยาวที่สุดที่พบจึงเลือกคำว่า “ตาก” หลังจากนั้นจะค้นหาและเปรียบเทียบต่อไปจนครบซึ่งจะได้ผลลัพธ์นั่นคือ “ฉัน | นั่ง | ตาก | ลม | ที่ | หน้า | บ้าน” เช่น โปรแกรมการตัดคำเล็กซ์โต (LexToPlus: A Thai Lexeme Tokenization and Normalization Tool) ที่พัฒนาโดยเนคเทค อ้างอิงจากพจนานุกรมเล็กซิตรอน (LEXiTRON) โดยใช้เทคนิคการแบ่งคำแบบยาวที่สุด และผู้ใช้สามารถเพิ่มรายการคำศัพท์ เพื่อให้เหมาะสมกับการนำไปใช้งาน 2) การตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) เป็นวิธีการตัดคำตามรูปแบบที่สามารถจะเป็นไปได้ทั้งหมด เช่น เมื่อมีข้อความว่า “ไปหามเหสี” ก็จะตัดคำได้ 2 แบบ คือ “ไป | หาม | เห | สี” มีจำนวนคำที่ตัดได้เท่ากับ 4 และ “ไป | หา | มเหสี” ที่มีจำนวนคำที่ตัดได้เท่ากับ 3 วิธีการนี้จะให้เลือกข้อความที่หลังจากแบ่งแล้วมีจำนวนคำน้อยที่สุด ส่วนในกรณีที่มีจำนวนคำที่เท่ากันจะเลือกวิธีที่มีการตัดคำแบบยาวที่สุด 3) การตัดคำแบบคำนวณเชิง

สถิติเพื่อหาความเป็นไปได้ (Probabilistic Model) โดยนำสถิติการเกิดของคำและหน้าที่ของคำ (Part of Speech) เข้ามาช่วยในการคำนวณหาความน่าจะเป็น เพื่อเลือกคำที่มีโอกาสเกิดมากที่สุด วิธีการนี้สามารถจะตัดคำได้ดีกว่าการตัดคำแบบยาวที่สุดและการตัดคำแบบสอดคล้องมากที่สุด แต่มีข้อจำกัดคือต้องมีฐานข้อมูลการตัดคำและกำหนดหน้าที่ของคำที่ถูกต้อง เพื่อใช้สร้างสถิติ และ 4) การตัดคำแบบคุณลักษณะ (Feature-Based Approach) ที่พิจารณาจากบริบท (Context Words) และการเกิดร่วมกันของคำหรือหน้าที่ของคำ (Collocation) เข้ามาช่วยในการตัดคำ เช่น คำว่า “ตากลม” ถ้าพบคำว่า “แป้ว” หรือ “โต” ในบริบทก็จะสามารถตัดคำได้ว่า “ตา | กลม” ในทางกลับกันหากในบริบทพบคำว่า “หน้าบ้าน” จะเลือกตัดคำว่า “ตาก | ลม” เป็นผลลัพธ์ในการตัดคำ แต่วิธีการนี้จำเป็นต้องมีฐานข้อมูลขนาดใหญ่และต้องมีการเรียนรู้การสร้างคำในบริบทหรือการเกิดร่วมกันของคำแต่ละคำ เพื่อนำมาใช้ในการตัดคำ เช่น โปรแกรมตัดคำภาษาไทยแบบอิงการเรียนรู้ของเครื่องทีเล็กซ์ (TLexPlus: Thai Lexeme Analyser) ที่ใช้เทคนิคการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ (Machine Learning) โดยอาศัยหลักการ Conditional Random Fields (CRFs) ร่วมกับคลังคำศัพท์ของ BEST2009 ขนาด 9 ล้านคำในการเรียนรู้ จึงสามารถแบ่งคำที่เกิดขึ้นใหม่ คำในภาษาต่างประเทศ หรือคำแสลงใหม่ ได้อย่างถูกต้อง โดยไม่ต้องอาศัยพจนานุกรม เหมาะสำหรับนำไปแบ่งคำเพื่อหานิพจน์ระบุนาม (Named Entities) หรือชื่อเฉพาะต่าง ๆ

งานวิจัยนี้ใช้หลักการตัดคำโดยใช้พจนานุกรม เพื่อตัดคำภาษาไทยให้อยู่ในรูปคำเดี่ยว (Term) ตามที่กำหนดไว้แบบยาวที่สุด เช่น ข้อความ “ปอ 76 เลข 5–70123 ฝ่าไฟแดงแยกตากสิน เมื่อวันที่ 3 มกราคม 64 เวลาประมาณ 06.30 น. พบรถเมล์สาย 76 หมายเลขท้ายรถ 5–70123 ขับฝ่าไฟแดงแยกตากสิน ไม่หยุดรอเลี้ยวขวาก่อนเข้าถนนกรุงธนบุรี” จะได้ผลลัพธ์ว่า “ปอ | 76 | เลข | 570123 | ฝ่าไฟแดง | แยก | ตากสิน | เมื่อ | วันที่ | 3 | มกราคม | 64 | เวลา | ประมาณ | 06.30 | น. | พบ | รถเมล์ | สาย | 76 | หมาย | เลขท้าย | รถ | 570123 | ขับ | ฝ่าไฟแดง | แยก | ตากสิน | ไม่ | หยุด | รอ | เลี้ยวขวา | ก่อน | เข้า | ถนน | กรุงธนบุรี” ดังรูปที่ 2



รูปที่ 2 : ผลลัพธ์การตัดคำภาษาไทยแบบอิงพจนานุกรม

จากรูปที่ 2 จะนำผลลัพธ์จากการตัดคำในแต่ละครั้งมาเพิ่มจำนวนคำในอาร์เรย์ (Array) ที่สร้างขึ้นแบบวนลูป เพื่อหาความถี่ของคำทั้งหมดที่ปรากฏในเอกสาร จนครบ 1,020 ข้อความ

3.3) มอดูลการคัดเลือกและสร้างคลังคำศัพท์ (Data Selection and Word Corpus)

ผลลัพธ์จากการตัดคำจากข้อความภาษาไทยประกอบด้วยคำย่อย ๆ จำนวนมาก ส่งผลให้ยากต่อการคัดเลือกคำศัพท์เพื่อนำมาสร้างคลังคำศัพท์ ผู้วิจัยจึงเตรียมข้อมูล (Data Preprocessing) ก่อนการคัดเลือกคำศัพท์เข้าคลังคำศัพท์ แบ่งเป็น 3 ขั้นตอน ดังนี้

3.3.1) การกำจัดคำหยุด (Stop Words Removal) เป็นการนำคำที่ไม่มีนัยสำคัญต่อข้อความออก โดยที่ความหมายของคำหรือข้อความไม่เปลี่ยนแปลง คำหยุดมักจะปรากฏอยู่ทุกข้อความ และมีความถี่สูง ส่วนใหญ่จะอยู่ในรูปคำสรรพนาม

คำสันธาน คำบุพบท ถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการจำแนกหมวดหมู่ ดังนั้นจึงสามารถตัดคำดังกล่าวทิ้งได้เลย โดยที่ไม่ส่งผลกระทบต่อใจความหลัก การกำจัดคำหยุดเป็นกระบวนการที่ช่วยให้ขนาดของดัชนีลดลง อีกทั้งลดขนาดพื้นที่และเวลาในการประมวลผลลงด้วย ตัวอย่างคำหยุดดังตารางที่ 2 ร่วมกับการกำจัดตัวเลขและเครื่องหมายวรรคตอน การกำจัดช่องว่าง (Spacing) การกำจัดคำหยาบ (Swear Word) และการเปลี่ยนตัวอักษรภาษาอังกฤษให้เป็นตัวพิมพ์เล็กทั้งหมด เพื่อให้ข้อมูลอยู่ในสภาพที่พร้อมนำไปประมวลผลในขั้นต่อไป

ตารางที่ 2 : ตัวอย่างคำหยุดในภาษาไทย

คำบุพบท	คำสันธาน	คำสรรพนาม	คำวิเศษณ์	คำอุทาน
กับ	กว่า	กระผม	ใกล้	555
ของ	คือ	ข้าพเจ้า	ครับ	กำ
ซึ่ง	จึง	ฉัน	ค่ะ	ขอโทษ
...	...	...	...	...
โดย	ถ้า	นาย	ที่สุด	เยี่ยม

อย่างไรก็ตาม ยังมีการเตรียมข้อมูลรูปแบบอื่นที่สามารถทำได้อีกหลายหลายวิธีไม่ว่าจะเป็นการทำวจีวิภาค (part of speech tagging: POS), การระบุชื่อเฉพาะ (name entity recognition: NER) การจับกลุ่มคำที่มักปรากฏด้วยกัน การหาความสัมพันธ์ของคำ รวมถึงการเลือกกำจัดคำบางคำที่ไม่มีความสำคัญกับการวิเคราะห์ ซึ่งสามารถเลือกใช้แตกต่างกันไปตามบริบทของข้อมูล

3.3.2) การสร้างดัชนีคำศัพท์ (Indexing) เป็นกระบวนการแปลงเอกสารที่เป็นภาษาธรรมชาติ ให้คอมพิวเตอร์สามารถเข้าใจและประมวลผลได้ การสร้างดัชนีเป็นการสร้างตัวแทนเอกสาร (Document Representation) ให้อยู่ในรูปของเวกเตอร์เอกสาร (Word Vector) ที่สามารถคำนวณค่าน้ำหนักของคำ (Term Weighting) เช่น คำเดี่ยว (Single Word), รากศัพท์ (Stem), วลี (Phrase), ชุดลำดับ (N-Gram) หรือประโยค (Sentence) เป็นต้น

หลังจากนำข้อร้องเรียนจำนวน 1,020 รายการ เข้าสู่กระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม และการกำจัดคำหยุด จะได้อาร์เรย์คำศัพท์หนึ่งชุด งานวิจัยนี้สร้างดัชนีเอกสารโดยใช้การหาค่าน้ำหนักรูปแบบคำเดี่ยวด้วยการวิเคราะห์น้ำหนักของคำ (term frequency–inverse document frequency: TF-IDF) เป็นเทคนิคที่พิจารณาองค์ประกอบของคำภายในประโยคและเอกสารเป็นหลัก โดยจะไม่นำลำดับของคำภายในเอกสารมาวิเคราะห์ [11] ด้วยวิธีการทางสถิติ เพื่อประเมินความสำคัญของคำต่อเอกสารที่เป็นสัดส่วนโดยตรงกับจำนวนครั้งที่คำ ๆ นั้นปรากฏในเอกสาร แต่จะถูกลดความสำคัญโดยความถี่ของคำนั้นในกลุ่มเอกสารทั้งหมด โดยการให้น้ำหนักคำในแต่ละคำจากใช้ 2 ปัจจัยที่มีความสัมพันธ์กัน คือค่าความถี่ของคำ (TF) เพื่อหาว่าแต่ละคำนั้นปรากฏบ่อยแค่ไหนในแต่ละเอกสาร ดังสมการที่ 1

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

จากนั้นหาค่าความผกผันในความถี่ของเอกสาร (IDF) โดยการให้ค่าน้ำหนักกับคุณลักษณะที่ใช้เป็นตัวแทนของเอกสาร ซึ่งควรจะปรากฏอยู่เป็นจำนวนมากในเนื้อหาของเอกสารเฉพาะฉบับนั้น และปรากฏอยู่น้อยในชุดของเอกสารที่เหลือทั้งหมด จากแนวคิดที่ว่าการแทนข้อความด้วยค่าความถี่การปรากฏของคุณลักษณะเพียงอย่างเดียว ไม่สามารถจำแนกข้อความได้ดีพอ เพราะถ้าคุณลักษณะนั้นเกิดขึ้นเป็นจำนวนมากในทุก ๆ เอกสาร แสดงว่าคุณลักษณะดังกล่าวไม่สามารถใช้เป็นตัวแทนของเอกสารได้ จึงต้องหาค่าความถี่ผกผันด้วย ดังสมการที่ 2

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

กำหนดให้จำนวนเอกสารทั้งหมดที่ใช้ในการพิจารณา (n) เท่ากับ 1,020 และจำนวนเอกสารที่มีคำคำนั้นปรากฏอยู่ ($DF_{(t)}$) สูงสุดไม่เกิน 1,020 จะพบว่าคำที่ปรากฏไม่บ่อยครั้งนั้นจะมีค่าความถี่ของเอกสาร (DF) เข้าใกล้ 1 มากขึ้น เกือบทั้งหมดจะเป็นชื่อเฉพาะ เช่น สายรถประจำทาง ชื่อบุคคล ชื่อสถานที่ เป็นต้น ดังนั้นผู้วิจัยจึงพิจารณาตัดข้อมูลดังกล่าวออก เนื่องจากเป็นข้อมูลที่ไม่มีผลต่อการบ่งชี้ประเด็นในการร้องเรียนการให้บริการรถโดยสารสาธารณะ คงเหลือเพียงคำศัพท์ที่สำคัญ และอีกส่วนหนึ่งเป็นคำศัพท์สำคัญที่เขียนผิดจากคำที่ปรากฏในพจนานุกรม อันเกิดจากความตั้งใจหรือไม่ตั้งใจของผู้ใช้ ตลอดจนคำทับศัพท์ (Transliterated Word) และคำศัพท์แสลง (Slang Word) หลังจากนั้นจะนำค่าความถี่ของคำศัพท์ (TF) และค่าความผกผัน

ในความถี่ของเอกสาร (IDF) ของแต่ละคำ มาคูณกันเพื่อหาค่าน้ำหนักที่สามารถแยกคำศัพท์สำคัญออกมาได้ ดังสมการที่ 3

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

จากผลลัพธ์จะพบคำศัพท์ที่มีค่า TF-IDF จากเอกสารทั้งหมด ซึ่งเป็นคำศัพท์ที่มีแนวโน้มที่จะเป็นใจความสำคัญของเอกสารตามแนวคิดคำถูกกล่าวถึงบ่อยที่สุดและไม่ได้ปรากฏอยู่หลายเอกสารเกินจนไป เพื่อนำมาคัดเลือกคุณลักษณะสำหรับสร้างคลังคำศัพท์ในการจำแนกปัญหาในการให้บริการ

3.3.3) การเลือกคุณลักษณะ (Feature Selection) เนื่องจากจำนวนคุณลักษณะมีผลต่อประสิทธิภาพการจำแนกหมวดหมู่ข้อความ จึงเป็นที่มาของแนวคิดการสร้างคุณลักษณะใหม่จากคุณลักษณะเดิม [12] เช่น การหาความถี่เอกสาร (Document Frequency: DF) การหาค่าสารสนเทศ (Information Gain: IG) หรือการหาค่าสถิติไคสแควร์ (Chi-Square) เป็นต้น

งานวิจัยนี้จำแนกปัญหาการให้บริการตามลักษณะของข้อร้องเรียนความถี่เอกสาร (DF) ที่พบบ่อยในเว็บบอร์ดองค์การขนส่งมวลชนกรุงเทพ คัดเลือกคำศัพท์จากกระบวนการตัดคำภาษาไทยจากข้อร้องเรียนจำนวน 1,020 ข้อความ คิดเป็นร้อยละ 80 แล้วคัดเลือกโดยผู้วิจัย ได้คำศัพท์จำนวน 118 คำ จากนั้นให้ผู้เชี่ยวชาญจำนวน 3 ท่าน จำแนกคำศัพท์ที่เกี่ยวข้องกับการร้องเรียนตามบริบทการให้บริการรถโดยสารสาธารณะ แบ่งออกเป็น 4 คลาส ตามบริบทการให้บริการรถโดยสารสาธารณะ ได้แก่ 1) คลาสการขับขี่ 2) คลาสผู้ขับขี่และพนักงานผู้ให้บริการ 3) คลาสยานพาหนะและอุปกรณ์ให้บริการ และ 4) คลาสเวลาและการเดินรถ จากนั้นผู้วิจัยจัดเก็บคำศัพท์ในลักษณะถุงคำ (Bag of words) [12] ดังตารางที่ 3 ก่อนจะนำเข้า (Import) สู่ฐานข้อมูล MySQL สำหรับนำไปสร้างคลังคำศัพท์เพื่อใช้ในการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

ตารางที่ 3 : การจัดเก็บคำศัพท์ในคลังคำศัพท์

คลาสที่ 1 คลาสการขับขี่ 46 คำ			
ขับรถเร็ว	ไม่จอดรับ	เบรกกระชั้นชิด	ขับกระชาก
เลยป้าย	ฝ่าไฟแดง	ขับเลนขวา	จอดไม่ชิด
วิ่งเลนขวา	ขับแข่ง	ปาดหน้า	แซงซ้าย
ออกรถก่อน	จอดไม่สนิท	...	ตัดหน้า

ตารางที่ 3 : การจัดเก็บคำศัพท์ในคลังคำศัพท์ (ต่อ)

คลาสที่ 2 คลาสผู้ขับขี่และพนักงานผู้ให้บริการ 31 คำ			
ไม่สุภาพ	ไม่มีมารยาท	เสียงดัง	ไม่ยิ้มแย้ม
ไม่ฉีกตั๋ว	แต่งกาย	ใช้โทรศัพท์	มารยาทแย่
หยาบคาย	ไม่มีสมาธิ	เงินทอน	รับเงิน
ไม่ใส่แมสก์	สูบบุหรี่	...	ชักสีหน้า
คลาสที่ 3 คลาสยานพาหนะและอุปกรณ์ให้บริการ 26 คำ			
แอร์เสีย	แอร์ไม่เย็น	เออเร่อ	รถเก่า
ท่อไอเสีย	เสียงดัง	ควันดำ	เปิดไม่ได้
แอพ	viabus	เครื่องอ่านบัตร	ไม่เว้นระยะห่าง
ปิดไม่สนิท	ชำรุด	...	ไม่พร้อมใช้
คลาสที่ 4 คลาสเวลาและการเดินรถ 15 คำ			
ออกรถช้า	รอนาน	เพิ่มรถ	จอดเติมแก๊ส
จอดแช่	ทิ้งช่วง	รถไม่พอ	ไม่ตรงเวลา
ไม่มีรถ	เพิ่มเที่ยว	เวลาเร่งด่วน	รถหมด
เลท	ปล่อยรถ	...	รถน้อย

จากนั้นนำผลลัพธ์จากการตัดคำ มาแสดงผลด้วย Wordcloud [13] ที่มีลักษณะเป็นกลุ่มคำที่จับตัวกันเหมือนก้อนเมฆ เพื่อแสดงให้เห็นถึงคำศัพท์สำคัญได้ง่ายขึ้น โดยคำที่มีความถี่สูงก็จะแสดงผลเป็นข้อความขนาดใหญ่ และไล่ลงมาเป็นข้อความขนาดเล็กลงมาตามลำดับ โดยขนาดของคำขึ้นอยู่กับความถี่ของผลลัพธ์จากการตัดคำ ดังรูปที่ 3

รูปที่ 3 : การแสดงความถี่ของคำด้วย Wordcloud

4) ผลการวิจัย

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ และเพื่อ

ประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ แบ่งผลการศึกษาและอภิปรายผล รายละเอียดดังนี้

4.1) ผลการออกแบบและพัฒนากระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

งานวิจัยนี้แบ่งคลาสข้อมูลตามข้อร้องเรียนที่พบบนเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพ แบ่งออกเป็น 4 คลาสตามบริบทการให้บริการ ได้แก่ 1) คลาสการขับขี่ (Driving Class) จำนวน 504 รายการ 2) คลาสผู้ขับขี่และพนักงานผู้ให้บริการ (Person Class) จำนวน 275 รายการ 3) คลาสยานพาหนะและอุปกรณ์ให้บริการ (Vehicle Class) จำนวน 155 รายการและ 4) คลาสเวลาและการเดินรถ (Schedule Class) จำนวน 86 รายการ โดยมีคำศัพท์ที่ถูกคัดเลือกจำนวน 118 คำจัดเก็บภายในคลังคำศัพท์ในลักษณะถุงคำ

การแปลงดัชนีคำศัพท์ (Term Weighting) [12] เพื่อสร้างตัวแทนเนื้อหาของเอกสาร (Document Representation) สำหรับใช้ในกระบวนการเรียนรู้ ให้อยู่ในรูปแบบของเวกเตอร์เอกสาร โดยการเปรียบเทียบระหว่างคำศัพท์ที่ปรากฏในข้อร้องเรียนรถโดยสารสาธารณะตามที่ระบุไว้ในคลาสกับคำศัพท์ที่เก็บอยู่ในถุงคำ แล้วนับจำนวนคำศัพท์ที่ค้นพบจากแต่ละข้อร้องเรียน เพื่อนำความถี่มาคำนวณหาค่าน้ำหนักของคำศัพท์แต่ละคำตามลำดับ ดังตารางที่ 4

ตารางที่ 4 : การแปลงดัชนีคำศัพท์ในรูปแบบของเวกเตอร์เอกสาร

ข้อร้องเรียน	คำศัพท์				คลาส
	w_1	w_2	w_n	w_{118}	
1. บัตร Prompt Card ...	0	0	...	0	Vehicle
. ...**ขับรถเร็ว** วิ่งขา **ไม่จอดรับ** ผู้โดยสาร	0.1	0.1	...		Driving
...	...	...	...	...	...
. รถเมล์สาย 26 **รถน้อย**มาก...	0	0	...	0.1	Schedule

จากตารางที่ 4 นำค่าน้ำหนักของคำศัพท์มาสร้างตัวแบบในการจำแนกกลุ่มข้อความ ในรูปแบบเอกสาร .arff [14] สำหรับนำเข้าโปรแกรมเวกา (wailato environment for knowledge analysis: Weka) เพื่อวิเคราะห์และจำแนกข้อมูล (Classification) ดังตารางที่ 5

ตารางที่ 5 : คุณลักษณะของข้อมูลที่ใช้จำแนก

ตัวแปร	ชนิดข้อมูล	คำอธิบาย
w_1	numeric	ค่าน้ำหนักคำที่ 1 (ขับรถเร็ว)
w_2	numeric	ค่าน้ำหนักคำที่ 2 (ไม่จอดรับ)
w_n	numeric	ค่าน้ำหนักคำที่ n (...)
w_{118}	numeric	ค่าน้ำหนักคำที่ 118 (รถน้อย)
class	numeric	1) Driving class 2) Person class 3) Vehicle class 4) Schedule class

งานวิจัยชิ้นนี้จะใช้การตรวจสอบแบบไขว้กัน 10 ชุด (10-Fold Cross Validation) [14] โดยแบ่งข้อมูลออกเป็น 10 ชุด ข้อมูลชุดที่ 1 ใช้เป็นชุดข้อมูลทดสอบ (Test Data) และอีก 9 ชุดที่เหลือใช้เป็นชุดข้อมูลฝึกฝน (Training Data) จากนั้นเปลี่ยนชุดข้อมูลถัดไปเป็นชุดข้อมูลทดสอบส่วนที่เหลือเป็นชุดข้อมูลฝึกฝน วนซ้ำไปจนครบทั้ง 10 ชุดข้อมูล วิธีการนี้เป็นวิธีที่นิยมเพื่อใช้ลดความผิดพลาดของผลลัพธ์จากการสุ่มเลือกชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบ การจำแนกข้อร้องเรียนรถโดยสารสาธารณะในงานวิจัยชิ้นนี้ใช้อัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning Algorithm) ในการจำแนกข้อความ (Text Classification) แบบมีผู้สอน (Supervised Learning) 3 ชนิด [14] ได้แก่

1) ซัพพอร์ทเวกเตอร์แมชชีน (support vector machine: SVM) ใช้หลักการของสมการเส้นตรงในการแบ่งกลุ่มข้อมูล แล้ววัดระยะห่างจากเส้นตรงเพื่อเปรียบเทียบความถูกต้อง

2) เคเนียเรสเนเบอร์ (k-nearest neighbor: KNN) เป็นการจำแนกกลุ่มข้อมูลที่มีคุณสมบัติใกล้เคียงกันที่สุด K กลุ่ม ด้วยคำนวณค่าระยะห่างระหว่างจุด (Euclidean Distance) ของข้อมูล ข้อมูลที่มีความคล้ายคลึงกันมากจะมีระยะห่างน้อย งานวิจัยชิ้นนี้กำหนด K เท่ากับ 3, 5 และ 7 ตามลำดับ

3) เพอร์เซ็ปตรอนหลายชั้น (Multi-layer Perceptron) เป็นอัลกอริทึมโครงข่ายประสาทเทียม (artificial neural network: ANN) ที่เลียนแบบการทำงานของโครงข่ายประสาทของมนุษย์ โดยเชื่อมต่อกันเป็นโหนด (Node) หลายชั้น ค่าฟังก์ชันของแต่ละโหนดเกิดจากฝึกฝน (Train) โดยส่งค่าย้อนกลับทำให้ได้ตัวแบบที่สามารถให้น้ำหนักกับข้อมูลนำเข้า เพื่อให้ได้ประเภทที่ถูกจำแนกตามโหนดปลายทาง

การวัดประสิทธิภาพการจำแนกคลาสข้อร้องเรียนรถโดยสารสาธารณะแบบประเมินประสิทธิภาพเอกสารที่ถูกเลือก แต่มิได้เรียงลำดับความคล้ายคลึง [11] ประกอบด้วย 4 ค่า ได้แก่ 1) ค่าความถูกต้อง (Accuracy) 2) ค่าความแม่นยำ (Precision) 3) ค่าความระลึก (Recall) และ 4) ค่าประสิทธิภาพโดยรวม (F-measure) หรือ F1 score ดังสมการที่ 4-7

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

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

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

$$\text{F-measure} = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (7)$$

โดย TP คือ จำนวนข้อมูลที่ถูกต้องที่ถูกดึงออกมา
FP คือ จำนวนข้อมูลที่ผิดพลาดที่ถูกดึงออกมา
TN คือ จำนวนข้อมูลที่ถูกต้อง แต่ไม่ถูกดึงออกมา
FN คือ จำนวนข้อมูลที่ผิดพลาดแต่ไม่ถูกดึงออกมา

ผลการจำแนกข้อร้องเรียนรถ (Classification Model) โดยสารสาธารณะจาก 3 อัลกอริทึม ด้วยซัพพอร์ทเวกเตอร์แมชชีน (SVM) เคเนียเรสเนเบอร์ (K-NN) กำหนด K เท่ากับ 3, 5 และ 7 ตามลำดับ และโครงข่ายประสาทเทียม (ANN) ได้ผลลัพธ์ค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และ ค่าประสิทธิภาพโดยรวม ดังตารางที่ 6

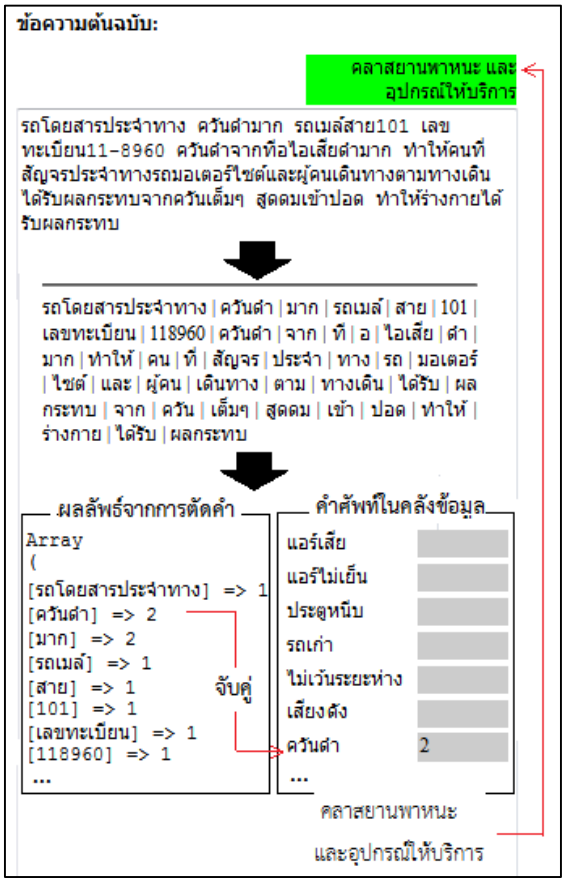
ตารางที่ 6 : ผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะด้วยเทคนิคการเรียนรู้ของเครื่อง

อัลกอริทึม	Accuracy	Precision	Recall	F-Measure
SVM	87.380	.9010	.8740	.881
K-NN (3)	87.210	.9000	.8720	.879
K-NN (5)	87.030	.8890	.8700	.875
K-NN (7)	86.860	.8980	.8690	.876
ANN	91.090	.9310	.9110	.915

จากตารางที่ 6 ผลการจำแนก (Predicted Category) ข้อร้องเรียนรถโดยสารสาธารณะด้วยเทคนิคการเรียนรู้ของเครื่อง พบว่า โครงข่ายประสาทเทียม มีค่าความถูกต้องในการจำแนกสูงที่สุด ถึง 91.09% ค่าความแม่นยำ 93.10% ค่าความระลึก 91.10% และค่าประสิทธิภาพโดยรวม 91.5% ถือว่าคลังคำศัพท์ที่สร้างขึ้นสามารถนำไปจำแนกข้อร้องเรียนรถโดยสารสาธารณะได้อย่างมีประสิทธิภาพ

4.2) ผลการประเมินความถูกต้องของผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

เป็นการนำผลจากการวิจัยไปประยุกต์ใช้งานจริง ผู้วิจัยได้สร้างคลังคำศัพท์ (Corpus-based) จำนวน 118 คำ มาทดสอบจำแนกข้อร้องเรียนรถโดยสารสาธารณะกับชุดทดสอบ (Test Data) จำนวน 255 ข้อความ เข้าสู่ด้วยกระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม แล้วจึงจับคู่ระหว่างคำศัพท์กับข้อร้องเรียนจากผู้ใช้ เพื่อติดแท็กจำแนกปัญหาการให้บริการ (Service Problem Tagging) แบ่งเป็น 4 คลาส ได้แก่ คลาสการขับขี่ คลาสผู้ขับขี่และพนักงานผู้ให้บริการ คลาสยานพาหนะและอุปกรณ์ให้บริการ และคลาสเวลาและการเดินรถ ดังรูปที่ 4



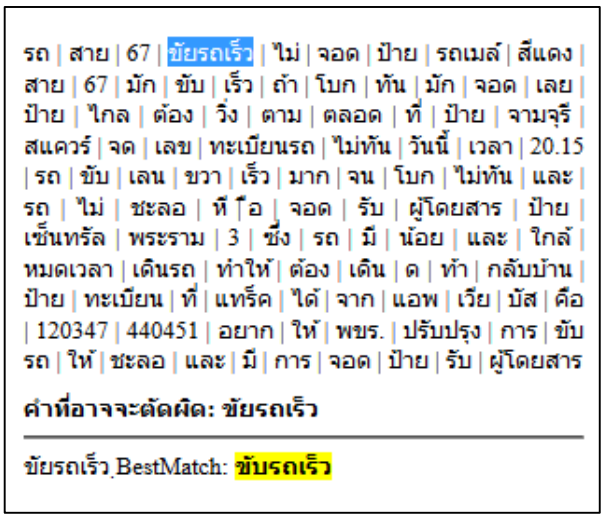
รูปที่ 4 : ผลการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ

จากรูปที่ 4 เมื่อตรวจสอบผลรวมของการจับคู่ผลลัพธ์จากการตัดคำกับคำศัพท์ในคลังคำศัพท์ที่อยู่ในฐานข้อมูลของแต่ละคลาส หากคลาสไหนพบการปรากฏของคำที่กำหนดไว้ในคลังคำศัพท์ จะทำการติดแท็กด้วยชื่อของคลาสนั้น ๆ โดยการจับคู่ที่เกิดขึ้นในแต่ละครั้งนั้นจะทำการตรวจสอบเปรียบเทียบคำศัพท์จากทั้ง 4 คลาสแบบวนลูป ดังนั้นผลลัพธ์การจำแนกข้อร้องเรียนแต่ละรายการจึงสามารถมีได้มากกว่า 1 แท็ก จากนั้นจะบันทึกลงฐานข้อมูลอีกครั้ง เพื่อนำเสนอสารสนเทศเป็นภาพข้อมูลต่อไป

อย่างไรก็ดี เนื่องจากเป็นการให้ข้อร้องเรียนในการใช้บริการจากผู้ใช้ จึงมักเกิดปัญหาความหลากหลายของข้อความที่แตกต่างกันตามบริบทของผู้ใช้งาน ตลอดจนความผิดพลาดในการใช้ภาษาที่เกิดจากความตั้งใจหรือแม้แต่ไม่ตั้งใจของผู้ใช้ ส่งผลให้ผลลัพธ์จากการตัดข้อความหากไม่ตรงกับคำศัพท์ในคลังคำศัพท์ที่อยู่ในฐานข้อมูลจะไม่สามารถทำการจับคู่ได้

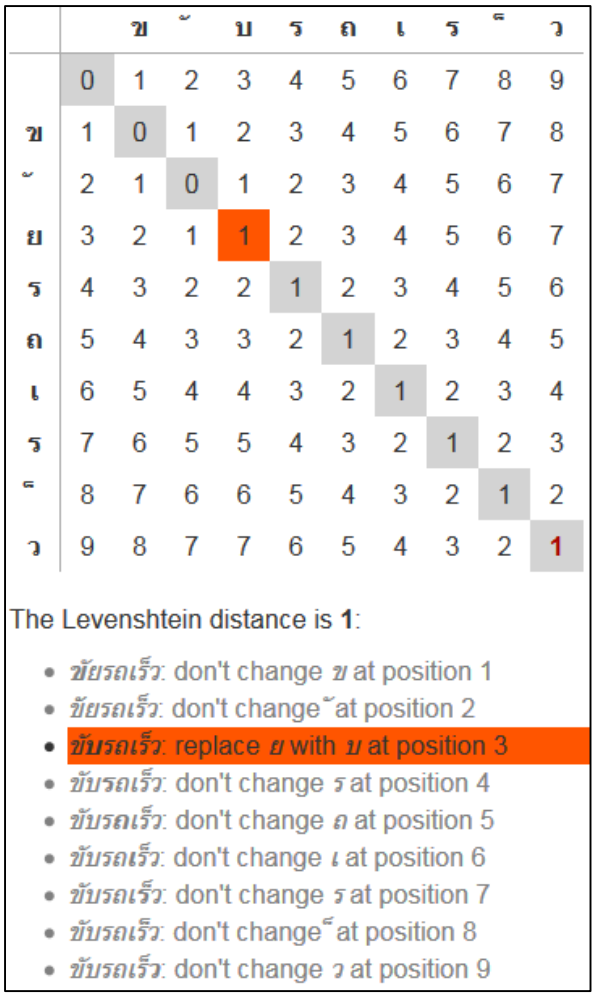
ดังนั้นจะเห็นได้ว่าความครอบคลุมของคลังคำศัพท์ส่งผลอย่างยิ่งต่อความถูกต้องของผลลัพธ์ในการจำแนกข้อร้องเรียนรถโดยสารสาธารณะ อย่างไรก็ดี หากจะให้คลังคำศัพท์ครอบคลุมครบถ้วนจำเป็นเพิ่มคำศัพท์มาจากความผิดพลาด 4 ชนิด [15] ได้แก่ พิมพ์เกิน เช่น “ไม่จอด” เป็น “ไม่จอดด” พิมพ์ผิด เช่น “ขับรถเร็ว” เป็น “ขัยรถเร็ว” พิมพ์ตก เช่น “มารยาท” เป็น “มายาท” และพิมพ์สลับ เช่น “สแกน” เป็น “แสกน” เป็นต้น อันจะส่งผลให้คำศัพท์ในคลังคำศัพท์มีจำนวนมากเกินความจำเป็น และยากที่จะระบุได้ว่าเท่าไรจึงจะครอบคลุมครบถ้วน เนื่องจากการพิมพ์ผิดพลาดถือได้ว่าเป็นพฤติกรรมพื้นฐานของมนุษย์ โดยผู้ที่พิมพ์ที่ไม่ชำนาญมักผิดพลาดจากการพิมพ์ผิด และการพิมพ์ผิดนั้นมักเกิดจากแป้นที่อยู่ติดกัน [16] หรือแม้แต่ผู้ที่พิมพ์ชำนาญก็มักจะมีความผิดพลาดจากการพิมพ์เกินใน 2 แป้นที่อยู่ติดกัน แต่อย่างไรก็ดี มีเพียงส่วนน้อยมากของความผิดพลาดที่เกิดจากอักขระตัวแรกของคำ [17] ดังนั้นการพิมพ์ผิด (Correction of spelling errors) ถือว่าเป็นพฤติกรรมอันเป็นไปตามกลไกทางธรรมชาติของมนุษย์ โดยเฉพาะอย่างยิ่งในสภาพแวดล้อมที่ถูกรบกวน

ผู้วิจัยเล็งเห็นความสำคัญของปัญหาดังกล่าว จึงทำการเปรียบเทียบความคล้ายคลึงกันของอักขระ (String Similarity) ด้วยการหาค่าความต่างกันของสายอักขระสองชุดด้วยเทคนิคการวัดระยะทางเลเวนชเตย์น (Levenshtein Distance) [18] ที่แก้ปัญหาโดยใช้กำหนดการพลวัต โดยค่าความต่างกันจะวัดจากจำนวนของอักขระที่จะต้องทำการแทรกเพิ่ม (Insertion) การลบ (Deletion) และการแทนที่ (Substitute) ของตัวอักษรหนึ่งตัว ระยะทางจะคำนวณจากจำนวนน้อยที่สุดของระยะทางที่แก้ไข (Edit Operation) [19] เพื่อแสดงความคล้ายคลึงที่มากที่สุด โดยนำผลลัพธ์จากการตัดข้อความที่ไม่ปรากฏในพจนานุกรมมาเปรียบเทียบกับคำศัพท์จำนวน 118 คำที่ดึงมาจากคลังคำศัพท์และเก็บไว้ในรูปแบบอาร์เรย์ แล้วทำการเปรียบเทียบทีละคู่แบบวนลูป ดังรูปที่ 5



รูปที่ 5 : การเปรียบเทียบความคล้ายคลึงกันของอักขระด้วยเทคนิคการวัดระยะทางเลเวนชเตย์น

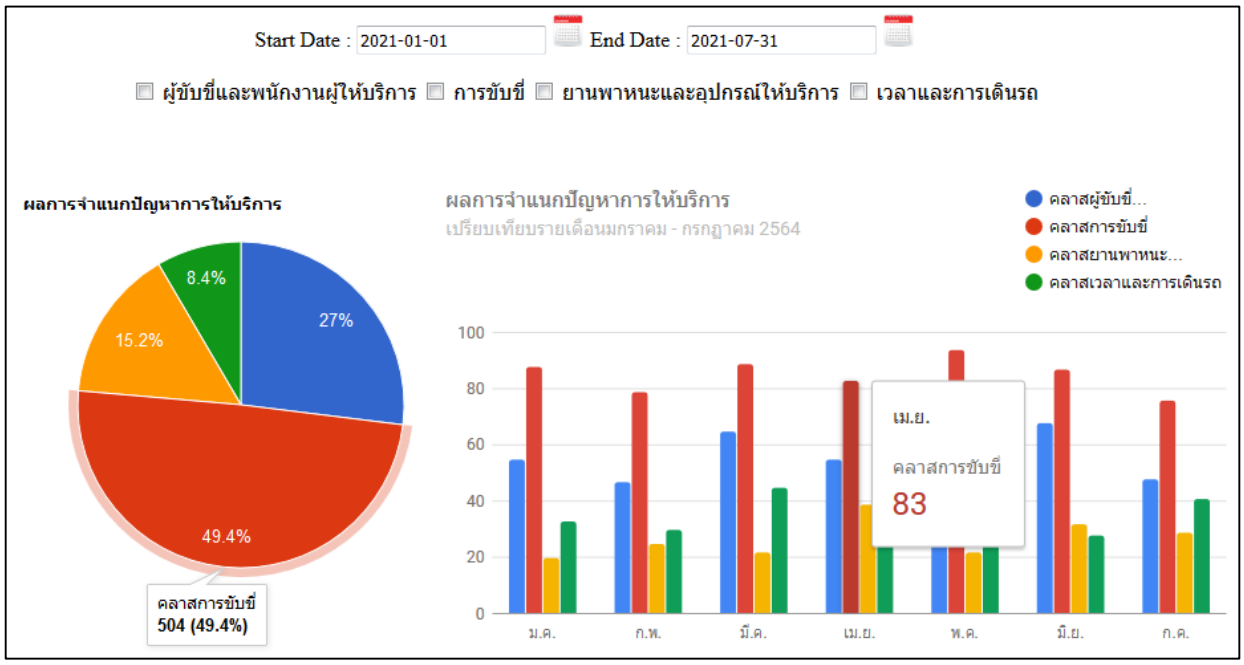
จากรูปที่ 5 การเปรียบเทียบความคล้ายคลึงกันระหว่างคำที่มาจากตัดข้อมูลคือ “ขัยรถเร็ว” กับคำว่า “ขับรถเร็ว” ที่กำหนดไว้ในคลังคำศัพท์ ซึ่งทั้งสองมีจำนวน 9 อักขระเท่ากัน เริ่มจากแยกอักขระแต่ละตัวออกจากกัน เพื่อเปรียบเทียบอักขระทีละคู่ พบว่า อักขระตัวที่ 1 และ 2 คือ “ข” และ “ั” ของทั้งสองข้อความตรงกัน ขณะที่อักขระตัวที่ 3 ของผลลัพธ์จากการตัดข้อความคือ “ย” เมื่อเปรียบเทียบกับคำที่กำหนดไว้ในคลังคำศัพท์ คือ “บ” จึงต้องทำการแทนที่อักขระ “ย” ด้วย “บ” และตรวจสอบอักขระคู่ต่อไปด้วยกระบวนการเดียวกันจนถึงอักขระตัวสุดท้าย จากนั้นจะวนลูปเปรียบเทียบจนครบทุกคำในคลังคำศัพท์ เพื่อดูว่าผลการเปรียบเทียบชุดอักขระของคำคู่ใดมีจำนวนระยะทางแก้ไขน้อยที่สุด ดังนั้นคำว่า “ขัยรถเร็ว” จึงมีความคล้ายคลึงกับคำว่า “ขับรถเร็ว” มากที่สุดเมื่อเทียบกับคำศัพท์ทั้งหมด เนื่องจากมีระยะทางแก้ไขน้อยที่สุดคือ 1 ตามผลลัพธ์ของเมตริกซ์ ดังรูปที่ 6



รูปที่ 6 : การเปรียบเทียบความคล้ายคลึงกันของอักขระ

จากรูปที่ 6 เมื่อพิจารณาคำที่มาจากความผิดพลาด 4 ชนิด [15] ได้แก่ พิมพ์เกิน เช่น “ไม่จอด” เป็น “ไม่จอดด” พิมพ์ผิด เช่น “ขับรถเร็ว” เป็น “ขัยรถเร็ว” พิมพ์ตก เช่น “มารยาท” เป็น “มายาท” และพิมพ์สลับ เช่น “สแกน” เป็น “แสกน” เป็นต้น พบว่า กระบวนการเปรียบเทียบความคล้ายคลึงกันของอักขระสองชุดด้วยเทคนิคการวัดระยะทาง เลเวนชเตย์นยังสามารถให้ผลลัพธ์เปรียบเทียบกับคำศัพท์ที่ถูกต้องได้อย่างมีประสิทธิภาพ

การนำเสนอสารสนเทศ ในงานวิจัยชิ้นนี้จะใช้เทคนิคการนำเสนอภาพข้อมูลที่นิยมนำมาใช้ประกอบการรายงาน เพื่อวิเคราะห์ และสรุปผล ปัจจุบันมีหลากหลายเครื่องมือที่สามารถนำมาใช้ในการแสดงผลด้วยการนำเสนอภาพข้อมูลที่สามารถแสดงผลข้อมูลเชิงปริมาณให้ง่ายต่อการทำความเข้าใจด้วย Google Charts [20] ที่ทำงานผ่านส่วนต่อประสานโปรแกรมประยุกต์ (application program interface : API) โดยจะส่งข้อมูลไปประมวลผลที่เครื่องแม่ข่าย และรับข้อมูลมาแสดงผลบนเว็บไซต์ เพื่อนำเสนอภาพข้อมูล [21] ด้วยแผนภูมิวงกลม (Pie Chart) ในการเปรียบเทียบคลาสการจำแนกข้อร้องเรียน และแผนภูมิแท่ง (Bar Chart) ในการแสดงผลเปรียบเทียบข้อร้องเรียนปัญหาการให้บริการในแต่ละเดือน ตามเงื่อนไขในการค้นหาระหว่างวันเริ่มต้น (Start Date) และวันสิ้นสุด (End Date) ดังรูปที่ 7



รูปที่ 7 : การนำเสนอภาพข้อมูลด้วย Google Charts

จากรูปที่ 7 ผลการจำแนกข้อร้องเรียนรถสาธารณะจากเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพ (ขสมก.) ระหว่างเดือนมกราคมถึงกรกฎาคม 2564 เกือบครึ่งเป็นการร้องเรียนพฤติกรรมการขับขี่ ซึ่งแทบทั้งหมดเป็นการร้องเรียนรถโดยสารสาธารณะไม่จอดรับ และไม่จอดป้ายถึงร้อยละ 70 จากจำนวนข้อร้องเรียนทั้งหมดในคลาสการขับขี่ อย่างไรก็ดี ประเด็นสำคัญจากข้อร้องเรียนด้านพฤติกรรมการขับขี่ เช่น ขับรถเร็ว ขับแข่ง และฝ่าไฟแดง แม้จะมีสัดส่วนไม่มากนักเมื่อเทียบกับข้อร้องเรียนทั้งหมด แต่ถือเป็นพฤติกรรมสำคัญที่อาจก่อให้เกิดอุบัติเหตุที่ส่งผลต่อชีวิตและทรัพย์สินบนท้องถนน ผลการจำแนกข้อร้องเรียนผ่านเว็บบอร์ดดังกล่าวสอดคล้องกับสถิติการรับเรื่องร้องเรียนรถโดยสารประจำทางขสมก. ตามกฎหมายว่าด้วยการขนส่งทางบก ปี 2563 ที่พบว่า 3 อันดับเรื่องร้องเรียนได้แก่ ไม่หยุดรับส่งผู้โดยสารที่ป้าย ขับรถประมาทหวาดเสียว และสภาพรถไม่สมบูรณ์ (ควันดำ) [5] ที่นำมาสู่ข้อเสนอแนะเชิงนโยบายในการติดอุปกรณ์ควบคุมความเร็วด้วย GPS การติดกล้องหน้ารถควบคู่ไปกับอบรมและติดตามตรวจสอบพฤติกรรมการขับของผู้ขับขี่อย่างสม่ำเสมอ

การประเมินผลความถูกต้องกับข้อมูลชุดทดสอบจำนวน 255 ข้อความโดยผู้เชี่ยวชาญ 3 ท่าน ท่านละ 85 รายการ กำหนดผลการจำแนกปัญหาบริการแต่ละข้อร้องเรียน โดยหากแท็กการจำแนกปัญหาทั้งหมดถูกต้องได้ 1 คะแนน ในทางตรงกันข้ามหากแท็กผลการจำแนกปัญหาของข้อร้องเรียนใดจำแนกไม่ถูกต้อง หรือถูกต้องแต่ไม่ครอบคลุมประเด็นปัญหาทั้งหมด ถือว่าได้ 0 คะแนน ผลการประเมินดังตาราง 7

ตารางที่ 7 : ผลประเมินความถูกต้องของผลการจำแนกข้อร้องเรียน

ท่านที่ 1 N=85		ท่านที่ 2 N=85		ท่านที่ 3 N=85		ค่าเฉลี่ย N=255	
✓	ร้อยละ	✓	ร้อยละ	✓	ร้อยละ	ร้อยละ	แปลผล
77	90.59	72	84.71	68	80.00	85.10	ดีมาก

จากตารางที่ 7 พบว่า ผลการประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการ อยู่ในระดับดีมาก (ร้อยละ 85.10) แสดงถึงกระบวนการจำแนกข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรมและการจับคู่ผลลัพธ์จากการตัดคำกับคำศัพท์ในคลังคำศัพท์ที่อยู่ในฐานข้อมูล ให้ผลลัพธ์ความถูกต้องในระดับสูง โดยเฉพาะข้อร้องเรียนแบบ 1 ประเด็นต่อ 1 ข้อร้องเรียน เนื่องจากเหมาะกับคลังคำศัพท์ที่กำหนดขอบเขตได้แน่นอน ในทางกลับกัน ปัญหาที่พบส่วนใหญ่เนื่องมาจากมีคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส เช่น เสียงดัง ที่จำเป็นต้องดูบริบทของข้อความประกอบ เนื่องจากเสียงดังเป็นกิริยาที่อาจเกิดจากบุคคลในคลาสผู้ขับขี่และพนักงานผู้ให้บริการ หรืออาจเป็นเสียงที่เกิดจากเครื่องยนต์จากคลาสยานพาหนะและอุปกรณ์ให้บริการ ดังนั้นการรวบรวมข้อมูลคำศัพท์ต้องมีกระบวนการที่น่าเชื่อถือ และปรับปรุงให้ครอบคลุมครบถ้วนอยู่เสมอ ร่วมถึงอาจต้องจำแนกคลาสโดยคำนึงถึงบริบทของข้อความเป็นหลัก

เช่นเดียวกับการเปรียบเทียบความคล้ายคลึงกันของอักขระด้วยเทคนิคการวัดระยะทางเลเวนชเตย์นที่แม้จะสามารถทำงานได้อย่างมีประสิทธิภาพเป็นที่น่าพอใจ แต่อาจจะไม่เหมาะกับบริบทของภาษาไทยมากนัก เนื่องจากคุณลักษณะของภาษาไทยมิได้จำกัดเพียงจากการประสมกันของตัวอักษรเหมือนภาษาอังกฤษ แต่ภาษาไทยยังมีคุณลักษณะเฉพาะตัว ที่ไม่เพียงเกิดจากการประสมกันของตัวอักษรเท่านั้น ยังมีสระ และวรรณยุกต์ที่เป็นคุณลักษณะของภาษาไทยเป็นจุดมักเกิดข้อผิดพลาดในพิมพ์ข้อความอีกด้วย อีกทั้งค่าความคลายคลึงของคำดังกล่าวเป็นผลมาจากการวิเคราะห์สายอักขระเท่านั้น มิได้เป็นผลมาจากความหมายของคำแต่อย่างใด อย่างไรก็ดี การตัดคำภาษาไทยจัดว่าเป็น NP-hard Problem เพราะไม่มีผิดถูกชัดเจน เช่น “ตากลม” สามารถตัดได้เป็น “ตาก | ลม” หรือ “ตา | กลม” เป็นต้น ดังนั้นการวัดความถูกต้องนอกจากจะวัดในเชิงความหมายแล้ว ยังต้องวัดในบริบทของการใช้งานด้วย

จะเห็นได้ว่าการแบ่งคำภาษาไทยยังมีปัญหาความถูกต้องเนื่องจากความซับซ้อนในการผสมคำของภาษาไทย [10] ที่มีตัวอักษรหลายประเภท การผสมสระและวรรณยุกต์เพื่อสร้างคำ และการที่เขียนต่อเนื่องกันเป็นประโยค โดยไม่มีการเว้นวรรคหรือตัวคั่นใด ๆ อย่างภาษาอังกฤษ แม้ว่าการศึกษาจำนวนมากได้ดำเนินการเกี่ยวกับการแบ่งส่วนคำภาษาไทยด้วยเทคนิคต่าง ๆ แต่ยังคงพบปัญหาดังนี้

1. ปัญหาเกี่ยวกับประสิทธิภาพของอัลกอริทึมในการแบ่งคำภาษาไทย อันเนื่องมาจากความครอบคลุมของคลังคำศัพท์ส่งผลอย่างยิ่งต่อความถูกต้องของผลลัพธ์ในการแบ่งคำภาษาไทย อย่างไรก็ดี หากจะให้คลังคำศัพท์ครอบคลุมครบถ้วนจำเป็นต้องมีคลังคำศัพท์ขนาดใหญ่ และยากที่จะระบุได้ว่าเท่าไรจึงจะครอบคลุมครบถ้วน

2. ปัญหาเกี่ยวกับคำที่สะกดผิด ส่งผลให้แบ่งคำภาษาไทยที่ยึดตามพจนานุกรมไม่ถูกต้อง หรือไม่สามารถวิเคราะห์ความหมายได้เลย ข้อผิดพลาดในการสะกดผิดมี 7 ประเภท ได้แก่ ตัวอักษรเกิน เช่น “ไม่จอด” เป็น “ไม่จอดด”, ตัวอักษรหายไป เช่น “มารยาท” เป็น “มายาท”, ตัวอักษรซ้ำ เช่น “กดด|ออด”, การพิมพ์ผิด เช่น “ขับรถเร็ว” เป็น “ขัยรถเร็ว”, ตัวอักษรผิดตำแหน่ง เช่น “สแกน” เป็น “แสกน”, คำสแลง เช่น “โดนรถเมล์เท” และอื่น ๆ ที่เกิดจากหลายข้อผิดพลาดร่วมกัน แม้ว่าจะมีการใช้เทคนิคหลายอย่างในการแก้ไขคำที่สะกดผิด เช่น การวัดความคล้ายคลึงของคำ รูปแบบเอ็น-แกรม (N-Gram) ความน่าจะเป็นแบบเบย์ (Bayesian approach) การเรียนรู้ของเครื่อง (Machine Learning) หรือการสืบค้นคำไทยตามเสียงอ่าน (Thai Soundex) เป็นต้น แต่ก็ยังคงเกิดปัญหาความไม่ถูกต้องบางประการ และส่งผลต่อความผิดพลาดในการประมวลผลภาษาธรรมชาติ

3. การประสมคำภาษาไทยที่สร้างขึ้นจากคำพื้นฐาน เพื่อใช้ในวัตถุประสงค์ที่แตกต่างกัน เช่น คำศัพท์แสลง (Slang word) และคำทับศัพท์ภาษาอังกฤษ (Transliteration word) ภายใต้รูปแบบการเกิด ระยะเวลา และปริมาณ ดังนั้นจึงไม่สามารถจัดเก็บคำทั้งหมดลงในพจนานุกรมได้ เนื่องจากมีคำประเภทนี้จำนวนมากส่งผลต่อประสิทธิภาพของการแบ่งคำภาษาไทย

5) สรุปและอภิปรายผล

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อ 1) ออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะ จากข้อร้องเรียนผ่านเว็บบอร์ดขององค์การขนส่งมวลชนกรุงเทพด้วยกระบวนการตัดคำภาษาไทยโดยใช้พจนานุกรม จากนั้นคัดเลือกและสร้างคลังคำศัพท์ ด้วยการวิเคราะห์น้ำหนักของคำตามลักษณะของข้อร้องเรียนความถี่เอกสารที่พบบ่อย จัดเก็บคำศัพท์ในลักษณะถุงคำ แล้วคัดเลือกคำศัพท์มาสร้างเป็นคลังคำศัพท์แบ่งเป็น 4 คลาส ได้แก่ คลาสการขับขี่ คลาสผู้ขับขี่และพนักงานผู้ให้บริการ คลาสยานพาหนะและอุปกรณ์ให้บริการ และคลาสเวลาและการเดินรถ วัดผลด้วยหลักการเรียนรู้ของเครื่องเพื่อจำแนกข้อความ พบว่า อัลกอริทึมโครงข่ายประสาทเทียมแบบเพอร์เซ็ปตรอนหลายชั้น มีค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าประสิทธิภาพโดยรวมสูงที่สุด 2) ประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการจากการจับคู่กับคำศัพท์ในคลังคำศัพท์จำนวน 118 คำกับข้อร้องเรียนจากผู้ใช้ชุดทดสอบ เพิ่มความถูกต้องของผลลัพธ์ด้วยการเปรียบเทียบความคล้ายคลึงของคำด้วยเทคนิคการวัดระยะทางเลเวนชเตย์นในกรณีที่พบคำศัพท์ที่เขียนไม่ถูกต้องตามที่กำหนดไว้ในคลังคำศัพท์ เพื่อติดแท็กจำแนกปัญหาในการให้บริการของรถโดยสารสาธารณะ ประเมินความถูกต้องโดยผู้เชี่ยวชาญ ผลการประเมินอยู่ในระดับดีมาก โดยเฉพาะข้อร้องเรียนประเด็นเดียวแสดงถึงการตัดคำภาษาไทยแบบอิงพจนานุกรม เหมาะกับคลังคำศัพท์ที่กำหนดขอบเขตได้แน่นอน ความถูกต้องและครอบคลุมของคำศัพท์ส่งผลโดยตรงต่อผลลัพธ์ในการจำแนก อย่างไรก็ดีพบปัญหาคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส เช่น เสียงดัง อาจมีที่มาจากคลาสผู้ขับขี่และพนักงานผู้ให้บริการ หรือจากคลาสยานพาหนะและอุปกรณ์ให้บริการก็ได้ตามต้นกำเนิดการเกิดเสียง ตลอดจนปัญหาการประสมคำในภาษาไทยที่มีลักษณะเฉพาะตัวส่งผลต่อความถูกต้องของผลการตัดคำ ดังนั้นการวัดความถูกต้อง นอกจากจะวัดในเชิงความหมายแล้ว ยังต้องวัดในบริบทของการใช้งานด้วย

6) ข้อเสนอแนะ

การประมวลผลภาษาธรรมชาติสำหรับภาษาไทยนั้น มีการศึกษาเกี่ยวกับการแบ่งคำ 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) อย่างไรก็ตาม ยังคงเกิดปัญหาสำคัญ 4 ประเด็น คือ 1) ประสิทธิภาพการแบ่งคำ 2) การตรวจสอบและแก้ไขคำสะกดผิด 3) รูปแบบการสะกดคำที่หลากหลาย และ 4) การแบ่งกลุ่มคำประสม ปัญหาเหล่านี้ทำให้เกิดผลลัพธ์ที่ไม่ถูกต้องและเกิดคำจำนวนมากที่กระทบต่อการประมวลผลภาษาธรรมชาติสำหรับภาษาไทย ปัจจุบันมีการนำเสนออัลกอริทึมการแบ่งคำที่มีประสิทธิภาพสูงที่เข้ามาแก้ไขปัญหาสำคัญ 4 ประการ เรียกว่า การแบ่งส่วนภาษาไทยโดยการทดสอบจัดอันดับอัตโนมัติโดยสะกดผิดร่วมกับการแก้ไขคำสะกดผิด (Thai language segmentation by automatic ranking trie with misspelling correction: TLS-ART-MC) [10] ที่เข้ามาแก้ไขปัญหาสำคัญ 4 ประการ ซึ่งปรับปรุงประสิทธิภาพการประมวลผลและให้ผลลัพธ์ที่มีความแม่นยำสูง จากการรวม 3 เทคนิค ได้แก่ การจัดอันดับโครงสร้างข้อมูล (Ranking Trie) เป็นเทคนิคที่จัดเรียงคำใหม่ใน

โครงสร้างข้อมูล (Trie) เพื่อปรับปรุงประสิทธิภาพการแบ่งส่วนคำ การสืบค้นแบบตามเสียงอ่านสมบูรณ์ (Completed Soundex) เป็นอัลกอริทึมที่เข้ามาแก้ไขปัญหาการสะกดผิด และการแบ่งกลุ่มแบบสองรอบ (Two-Passes Segmentation) จะนำไปใช้กับกฎที่กำหนดไว้ล่วงหน้า เพื่อช่วยในการแก้ปัญหาการแบ่งส่วนคำประสมโดยไม่บันทึกคำประสมทั้งหมดในพจนานุกรม

อย่างไรก็ดี เพื่อเพิ่มประสิทธิภาพในการแบ่งคำหรือจำแนกข้อความทางภาษาไทย งานวิจัยในครั้งต่อไปควรให้ความสำคัญกับการกำจัดคำผิด โดยทำการลบ Noise ข้อความ เช่น “รถไม่จอดดดดดดดดดดดด” หรือ “รถม่ายจอด” เป็นต้น เพื่อเพิ่มคุณภาพของข้อมูลก่อนนำเข้ากระบวนการประมวลผลภาษาธรรมชาติ ด้วย LSTM, ULMFiT หรือ GPT เป็นต้น ตลอดจนเพิ่มประสิทธิภาพการตัดข้อความภาษาไทย ความสมบูรณ์ของคลังคำศัพท์ และนำไปประยุกต์ใช้กับการจำแนกข้อมูลจริงแบบเรียลไทม์

REFERENCES

[1] Bangkok Mass Transit Authority, “Annual Report 2019,” (in Thai), Bangkok Mass Transit Authority, Bangkok, Thailand, 2019.

[2] Mahidol University, “Final Report: User Satisfaction Survey 2018,” (in Thai), Bangkok Mass Transit Authority, Bangkok, Thailand, 2018.

[3] Office of Transport and Traffic Policy and Planning. “The number of passengers on the BMTA bus.” MISTRAN.otp.go.th. http://mistran.otp.go.th/mis/Interview_HIPubilcBus.aspx (accessed Aug. 20, 2021).

[4] The Office of the Permanent Secretary, The Prime Minister's Office, “Complaint/Opinion Processing Results Quarter 1 Fiscal Year 2021,” (in Thai), The Office of the Permanent Secretary, The Prime Minister's Office, Bangkok, Thailand, 2021.

[5] Department of Land Transport, “Transport Statistics report 2020,” (in Thai), Transp. Statist. Group, Planning Div., Dep. Land Transp., Bangkok, Thailand, 2020.

[6] C. C. Aggarwal, *Data Classification Algorithm and Applications*. Boca Raton, FL, USA: CRC Press, 2015.

[7] M. Wozniak, *Hybrid Classifiers Methods of Data, Knowledge, and Classifier Combination*. Heidelberg, Germany: Springer-Verlag, 2014.

[8] S. McConnell, *Rapid Development: Taming Wild Software Schedules*. Washington, DC, USA: Microsoft Press, 1996.

[9] R. Mitchell, *Web Scraping with Python: Collecting Data from the Modern Web*. Sebastopol, CA, USA: O'Reilly Media, 2015.

[10] C. Tapsai, H. Unger and P. Meesad, *Thai Natural Language Processing Word Segmentation, Semantic Analysis, and Application*. Cham, Switzerland: Springer International Publishing, 2021.

[11] C. D. Manning, P. Raghavan and H. Schütze, *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2018.

[12] T. Jo, *Text Mining Concepts, Implementation, and Big Data Challenge*. Cham, Switzerland: Springer International Publishing, 2019.

[13] T. Kwantler, *Text mining in practice with R*. West Sussex, UK: John Wiley & Sons, 2017.

[14] J. Zizka, F. Darena and A. Svoboda, *Text Mining with Machine Learning Principles and Techniques*. Boca Raton, FL, USA: CRC Press, 2020.

[15] F. J. Damerau, “A technique for computer detection and correction of spelling,” *Commun. ACM*, vol. 7, no. 3, pp. 171–176, Mar. 1964.

[16] J. Grudin, “Non-hierarchic specification of components in transcription typewriting,” *Acta Psychol.*, vol. 54, no. 1–3, pp. 249–262, Oct. 1983.

[17] K. Kukich, “Techniques for automatically correcting words in text,” *ACM Comput. Surv.*, vol. 24, no. 4, pp. 377–439, Dec. 1992.

[18] V. I. Levenshtein, “Binary codes capable of correcting deletions insertions, and reversals,” *Dokl. Phys.*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[19] F. P. Miller, A. F. Vandome and J. McBrewster, *Levenshtein Distance Information theory, Computer science, String (computer science), String metric, Damerau-Levenshtein distance, Spell checker, Hamming distance*. Gladbach, Germany: Alphascript Publishing, 2009.

[20] J. Martinez, *Google Charts for Institutional Research Websites*. Houston, TX, USA: University of Houston, 2018.

[21] C. O Wilke, *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. Sebastopol, CA, USA: O’Reilly Media, 2019.