การออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรมเพื่อจำแนกปัญหาการให้บริการ

The Design and Development of Public Bus Complaint Extraction Process By Dictionary Based Approach for Service Problem Classification

จักรินทร์ สันติรัตนภักดี[1] และศศิธร อิ่มวุฒิ[1]
Chakkarin Santirattanaphakdi[1] and Sasithon Imvut[1]
[1]สาขาวิชาระบบสารสนเทศคอมพิวเตอร์ คณะบริหารธุรกิจ มหาวิทยาลัยวงษ์ชวลิตกุล
chakkarin_san@vu.ac.th

บทคัดย่อ

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อ 1) ออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรม จากข้อร้องเรียนจำนวน 389 ข้อความ พบว่า มีคำศัพท์ 128 คำที่ถูกคัดเลือกมาสร้างเป็นคลังคำศัพท์ แบ่งเป็น4 คลาส ได้แก่ คลาสผู้ขับขี่และพนักงานผู้ให้บริการ คลาสการขับขี่ คลาสยานพาหนะและอุปกรณ์ให้บริการ และคลาสเวลาและการเดินรถจากนั้นนำข้อความชุดทดสอบเข้ากระบวนการตัดคำภาษาไทยแบบอิงพจนานุกรมอีกครั้งแล้วจับคู่ผลลัพธ์จากการตัดคำกับคำสำคัญในคลังคำศัพท์แต่ละคลาส เพื่อติดแท็กจำแนกปัญหาการให้บริการและใช้เทคนิคการวัดระยะทางเลเวนชเตย์นในการเปรียบเทียบความคล้ายคลึงของคำในกรณีที่พบคำศัพท์ที่เขียนไม่ถูกต้องก่อนจะนำเสนอสารสนเทศเป็นภาพข้อมูลต่อไป 2) ผลการประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการ อยู่ในระดับดีมากโดยเฉพาะข้อร้องเรียนประเด็นเดียว แสดงถึงการตัดคำภาษาไทยแบบอิงพจนานุกรม เหมาะกับคลังคำศัพท์ที่กำหนดขอบเขตได้แน่นอน อย่างไรก็ดีพบปัญหาคำศัพท์ที่ซ้ำซ้อนกันในบางคลาสตลอดจนปัญหาการประสมคำในภาษาไทยผลลัพธ์จากงานวิจัยจะเป็นประโยชน์แก่ผู้รับผิดชอบเพื่อนำไปปรับปรุงการให้บริการกับประชาชนผู้ใช้บริการต่อไป

คำสำคัญ : การสกัดข้อมูล, รถโดยสารสาธารณะ, การตัดคำแบบอิงพจนานุกรม

Abstract

This research aims to 1) design and development of public bus complaint extraction process by dictionary-based approach. From 389 complaint text found 128 keywords that have been selected to create a corpus-based divided into 4 classes: person's class, driving class, vehicle class and bus schedule class. Then, input the test textinto extraction process by dictionary-based approachagainto matchingbetween the results of the wrapping text with the keywords in each class for create service problem classification tag. And use the Levenshte in edit distance method to comparing the similarity of words in case of incorrectly keywords. Last, information presentationto user by data visualization. 2) The accuracy assessment of service problem classification is a very good level, especially one complaint issue. The result show performance of dictionary-based approach on Thai wrappingis suitable for the terminology that has definite scope.However, found problems in same keywords to be duplicated in some

classes and compounding words in Thai. This results will benefit to those responsible to improve the service forcustomer.

Keyword: Data Extraction, Public Bus, Dictionary Based Approach

1. บทนำ

ระบบขนส่งสาธารณะมีบทบาทสำคัญอย่างมากต่อการดำเนินชีวิตประจำวันของประชาชนตั้งแต่อดีตจนถึงปัจจุบัน ประกอบด้วย ระบบขนส่งสาธารณะทางถนน ทางราง และทางน้ำ ที่ครอบคลุมทุกพื้นที่ทั่วประเทศ ซึ่งระบบขนส่งสาธารณะที่เข้าถึงและถูกใช้บริการมากที่สุด คือระบบขนส่งสาธารณะทางถนน โดยเฉพาะพื้นที่กรุงเทพมหานครและปริมณฑลที่เดินทางด้วยรถโดยสารประจำทางเป็นหลักภายใต้การดำเนินงานขององค์การขนส่งมวลชนกรุงเทพ [1] มีจำนวนเส้นทางรถโดยสารประจำทางรวมทุกประเภท จำนวน 456 เส้นทาง ครอบคลุมรถองค์กร รถเอกชนร่วมบริการ รถเล็กวิ่งในซอย รถตู้โดยสาร และรถตู้เชื่อมต่อท่าอากาศยานสุวรรณภูมิ โดยมีผู้ใช้บริการเฉลี่ยมากกว่า 438,414 คนต่อวัน มากกว่าระบบขนส่งมวลชนทุกประเภทที่ให้บริการในพื้นที่ [2] จากคุณสมบัติการให้บริการที่มีความคล่องตัวสูง สะดวก และสามารถให้บริการได้ทุกจุดตลอดระยะของการเดินทางตลอดจนอัตราค่าโดยสารยังอยู่ในระดับต่ำเมื่อเทียบระบบขนส่งมวลชนทางบกประเภทอื่น ๆ ในพื้นที่เดียวกันอีกด้วย

อย่างไรก็ตามจากผลการสำรวจของสำนักงานปลัดสำนักนายกรัฐมนตรี [3] เกี่ยวกับจำนวนเรื่องร้องทุกข์ในช่วงปีงบประมาณ 2563 ไตรมาศที่ 1 พบว่า องค์การขนส่งมวลชนกรุงเทพเป็นหน่วยงานรัฐวิสาหกิจในสังกัดกระทรวงคมนาคมที่มีประชาชนร้องเรียนโดยเฉลี่ยมากที่สุดเป็นอันดับหนึ่งเช่นเดียวกับหลายปีที่ผ่านมาโดยส่วนใหญ่เป็นเรื่องร้องทุกข์/เสนอความคิดเห็น ใน 3 ประเด็น ได้แก่ 1) ขอให้เพิ่มเที่ยว/รอบการเดินรถขยายเส้นทางการเดินรถและเพิ่มจำนวนรถโดยสารประจำทาง 2) ขอให้ปรับปรุงการให้บริการของพนักงานขับรถ พนักงานเก็บค่าโดยสาร และพนักงานขับรถตู้โดยสารปรับอากาศร่วมบริการ และ 3) ขอให้ปรับปรุงการให้บริการของรถโดยสารประจำทางรถโดยสารสาธารณะ รถโดยสารปรับอากาศประจำทาง และรถโดยสารปรับอากาศร่วมบริการนอกเหนือจากข้อร้องเรียนดังกล่าวที่ดำเนินการร้องทุกข์ผ่านช่องต่าง ๆ ของรัฐบาลแล้ว องค์การขนส่งมวลชนกรุงเทพเองยังมีช่องทางในการรับแสดงความคิดเห็นหรือรับเรื่องราวร้องเรียนรถองค์การ และร้องเรียนรถเอกชนร่วมบริการผ่านกระดานสนทนา (Web Board) บนเว็บไซต์ http://www.bmta.co.th เป็นอีกหนึ่งช่องทางที่ใช้ในการติดตามตรวจสอบการให้บริการและเป็นสื่อกลางในการแลกเปลี่ยนความคิดเห็นต่างๆ ที่ผู้ใช้งานสามารถตั้งกระทู้ถามตอบเพื่อแลกเปลี่ยนความคิดเห็นกันได้อย่างอิสระ ดังนั้นกระดานสนทนาจึงมีประโยชน์และมีบทบาทในการเพิ่มประสิทธิภาพในการให้บริการ แต่เมื่อเวลาผ่านไปจำนวนความคิดเห็นหรือข้อร้องเรียนในการใช้บริการมีเพิ่มมากขึ้นในแต่ละวัน และความหลากหลายของข้อความที่แตกต่างกันตามบริบทของผู้ใช้งาน ตลอดจนความผิดพลาดในการใช้ภาษาที่เกิดจากความตั้งใจหรือแม้แต่ไม่ตั้งใจของผู้ใช้ ที่ก่อให้เกิดปัญหาในการตีความหมาย โดยเฉพาะอย่างยิ่งข้อความที่ไม่มีการจำแนกหมวดหมู่ไว้อย่างชัดเจน อาจส่งผลต่อการตอบคำถาม การให้ข้อมูลคืนที่ถูกต้องแก่ผู้ใช้ ตลอดจนการสรุปสถิติการจัดประเภทข้อร้องเรียนนั้นทำได้ยาก ใช้เวลานาน และมีโอกาสเกิดความผิดพลาดสูง เนื่องจากผู้ดูแลระบบจะต้องนำข้อร้องเรียนมาวิเคราะห์เพื่อจำแนกหมวดหมู่ด้วยตนเอง ดังนั้นหากมีระบบการจำแนกข้อร้องเรียนตามหมวดหมู่ที่ต้องการแบบอัตโนมัติน่าจะเป็นแนวทางหนึ่งในการแก้ไขปัญหาดังกล่าวงานวิจัยชิ้นนี้มุ่งออกแบบและพัฒนากระบวนสกัดข้อร้องเรียนรถโดยสารสาธารณะทั้งรถองค์กร และรถเอกชนร่วมบริการผ่านกระดานสนทนา ซึ่งถือเป็นระบบขนส่งมวลชนสาธารณะหลักของกรุงเทพมหานครและปริมณฑล ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรมเพื่อจำแนกปัญหาการให้บริการ

ด้วยการติดแท็กอัตโนมัติ และนำเสนอเป็นสารสนเทศแก่ผู้รับผิดชอบเพื่อนำไปปรับปรุงและพัฒนาคุณภาพการให้บริการ ให้มีความสอดคล้องกับความต้องการของประชาชนผู้ใช้บริการ

2. วัตถุประสงค์ของการวิจัย

1. เพื่อออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรม
2. เพื่อประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการ

3. วิธีดำเนินการวิจัย

การเก็บรวบรวมข้อมูล

งานวิจัยชิ้นนี้เก็บรวบรวมข้อร้องเรียนของผู้ใช้จากเว็บไซต์ http://www.bmta.co.th/?q=th/forum ขององค์การขนส่งมวลชนกรุงเทพแบ่งเป็นข้อมูลตั้งแต่วันที่ 1 มกราคม 2564 ถึง 28 กุมภาพันธ์ 2564 จำนวน 389 ข้อความ ดังตารางที่ 1 นำมาตัดคำภาษาไทยแบบอิงพจนานุกรม เพื่อนำมาสร้างคลังคำศัพท์ และข้อมูลตั้งแต่วันที่ 1 ถึง 31 มีนาคม 2564 จำนวน 252 ข้อความ นำมาเป็นข้อมูลทดสอบ เพื่อประเมินผลความถูกต้องของผลการจำแนกปัญหาการให้บริการ

ตารางที่ 1 ตัวอย่างข้อความคิดเห็นหรือข้อร้องเรียน

ลำดับ	ข้อความ
1	ร้องเรียนรถองค์การขนส่งมวลชน ร้องเรียนรถเมล์สาย 63 ขาไปนนทบุรี วันอาทิตย์ที่ 3 มกราคม 2564 เวลาเกือบ ๆ 7 โมงเช้า ป้ายรถเมเจอร์รัชโยธิน รถไม่จอดป้าย และวิ่งเลนกลางวิ่งเร็วไม่รับคนเลย
2	ปอ 76 เลข 5-70123 ฝ่าไฟแดงแยกตากสิน เมื่อวันที่ 3 มกราคม 64 เวลาประมาณ 06.30 น. พบรถเมล์สาย 76 หมายเลขท้ายรถ 5-70123 ขับฝ่าไฟแดงแยกตากสิน ไม่หยุดรอเลี้ยวขวาก่อนเข้าถนนกรุงธนบุรี
...	...
389	รถเมล์ครีมแดง สาย 29 ยกเลิกวิ่งไปได้นานแล้ว, รถที่พึ่งได้คือ ครีมแดง สาย 59 (ช่วงค่ำคืนกลับไม่มีวิ่ง) ช่วงรอยต่อค่ำคืนของวันที่ 27 ถึง 28 ก.พ.64 ปกติน่าจะมีรถวิ่ง 2 คัน, ตรวจสอบโดย viabus ไม่ปรากฏและเมื่อรอรถนอก GPS. ไม่ปรากฏ เช่น กันตกลงของผมเสียหาย (ต้องนำไปแช่ตู้เย็น) เพราะปกติก็กลับบ้านเวลานี้บ่อย พอที่จะทราบการมาของรถสายนี้โดยอาจไม่จำเป็นต้องใช้ viabus สรุปคือ รออยู่บริเวณ รพ.โรคไตภูมิราชนครินทร์ กลับบ้านโซนหลักสี่ ***จึงขอความกรุณา ขสมก. อย่าสร้างความเดือดร้อนให้ผดส.ยุบสาย29 ไม่พอยังไม่มีรถ 59 วิ่งช่วงค่ำคืนอีก****

เครื่องมือที่ใช้ในการวิจัย

เครื่องมือที่ใช้ในการศึกษาวิจัยครั้งนี้ คือ แบบประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการที่ผ่านกระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรมที่ส่งออกมาจำนวน 252 รายการ

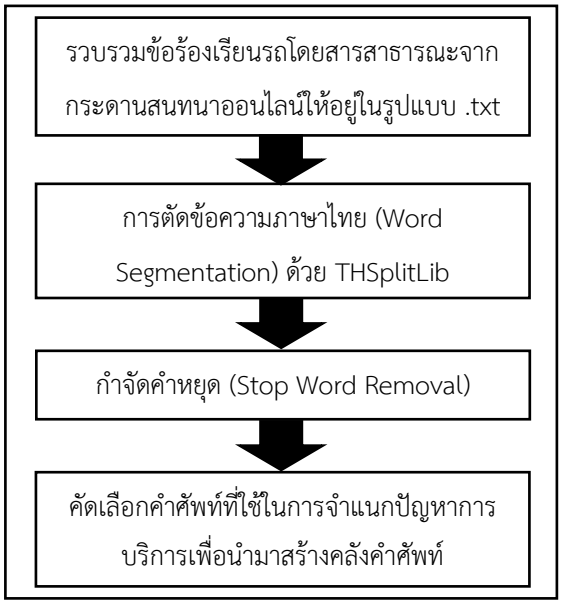
สถิติที่ใช้ในการวิจัย

สถิติที่ใช้ในการวิจัยครั้งนี้ คือ ค่าร้อยละความถูกต้อง กำหนดการแปลผลเป็น 5 ระดับ เพื่อให้ง่ายต่อการทำความเข้าใจ โดยดัดแปลงจากวิธีการใช้ร้อยละของคะแนนในการกำหนดช่วง [4]ดังนี้ ต่ำกว่าร้อยละ 50 หมายถึง ต้องปรับปรุง, ร้อยละ 50-59 หมายถึง พอใช้, ร้อยละ 60-69 หมายถึง ปานกลาง, ร้อยละ 70-79 หมายถึง ดี และร้อยละ 80-100 หมายถึง ดีมาก

ขั้นตอนการดำเนินการวิจัย

งานวิจัยชิ้นนี้มุ่งออกแบบและพัฒนากระบวนสกัดข้อร้องเรียนรถโดยสารสาธารณะที่รวบรวมข้อมูลจากกระดานสนทนาออนไลน์ ด้วยการตัดคำภาษาไทยแบบอิง

พจนานุกรม เพื่อจำแนกปัญหาการให้บริการด้วยการติดแท็กอัตโนมัติ และนำเสนอเป็นสารสนเทศแก่ผู้รับผิดชอบ โดยมีกรอบในการดำเนินงานดังภาพที่ 1



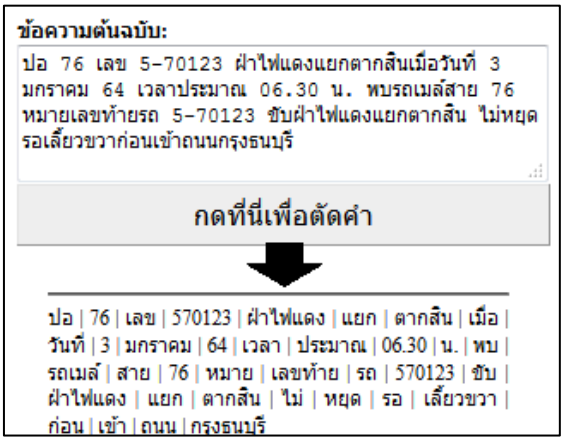
ภาพที่ 1 กรอบการดำเนินงานการกระบวนสกัดข้อร้องเรียนรถโดยสารสาธารณะ

จากภาพที่ 1 กรอบการดำเนินงานการกระบวนสกัดข้อร้องเรียนรถโดยสารสาธารณะแบ่งเป็น 6 ขั้นตอน โดยมีรายละเอียดดังนี้

1. รวบรวมข้อมูลข้อร้องเรียนรถโดยสารสาธารณะบนกระดานสนทนาออนไลน์จากเว็บไซต์ http://www.bmta.co.th/?q=th/forum ขององค์การขนส่งมวลชนกรุงเทพด้วยเทคนิคการขูดเว็บ (Web Scraping) [5] ที่เสมือนกับว่าเข้าใช้เว็บไซต์ด้วยตนเอง แต่ในทางกลับกันจะเขียนคำสั่งหรือใช้โปรแกรมสำเร็จรูปในการเข้าถึงเว็บไซต์ด้วยการเขียนโปรแกรมภาษา R เพื่อเข้าถึงข้อมูลแบบวนลูปแล้วนำมาบันทึกให้อยู่ในรูปไฟล์เอกสารธรรมดา (.txt)

2. การตัดข้อความ (Word Segmentation) เป็นหนึ่งในขั้นตอนที่สำคัญของการวิเคราะห์ข้อความ เป็นการนำข้อความทั่วไปซึ่งอยู่ในรูปแบบประโยคมาแบ่งออกเป็นคำหรือคุณลักษณะ (Term/Feature) มีจุดประสงค์เพื่อแยกส่วนของข้อความออกจากกันก่อนนำไปประมวลผลในขั้นต่อไป อย่างไรก็ดีการตัดคำในภาษาไทยยังพบปัญหาในการตัดคำเนื่องจากลักษณะของภาษาไทยมีการเขียนติดต่อกับแบบไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำที่ชัดเจน แตกต่างจากภาษาอังกฤษ ที่มีช่องว่างแสดงให้เห็นถึงขอบเขตของแต่ละคำ จึงเป็นที่มาของการตัดคำภาษาไทยก่อนนำไปวิเคราะห์ [6] แบ่งตามกระบวนการทำงานเป็น 3 กลุ่ม ได้แก่ 1) การตัดคำโดยใช้กฎ (Rule-Based Approach) เป็นการตัดคำโดยใช้วิธีเกณฑ์ทางอักขรวิธีที่กำหนดลักษณะของการประสมอักษร 2) การตัดคำโดยใช้พจนานุกรม (Dictionary-Based Approach) ที่เก็บคำศัพท์ไว้ในพจนานุกรม แล้วนำข้อความป้อนเข้าไปค้นหาและเปรียบเทียบสายอักขระกับคำศัพท์ในพจนานุกรม เพื่อหาว่าข้อความดังกล่าวควรตัดคำในบริเวณใด และประกอบด้วยคำใดบ้าง อย่างไรก็ดีการตัดคำโดยใช้พจนานุกรมก็มีข้อจำกัดบางประการ เนื่องจากมีความเป็นไปได้ที่คำที่ปรากฏในเอกสาร อาจจะไม่ปรากฏในพจนานุกรม จึงเป็นที่มาของเอ็นแกรม (N-gram) ที่นำบางส่วนของข้อความออกมาเป็นตามค่า N และ 3) การตัดคำโดยใช้คลังข้อมูล (Corpus-Based Approach) โดยเตรียมคลังข้อมูลที่มีการตัดคำและการกำกับหน้าที่ของคำไว้ล่วงหน้า

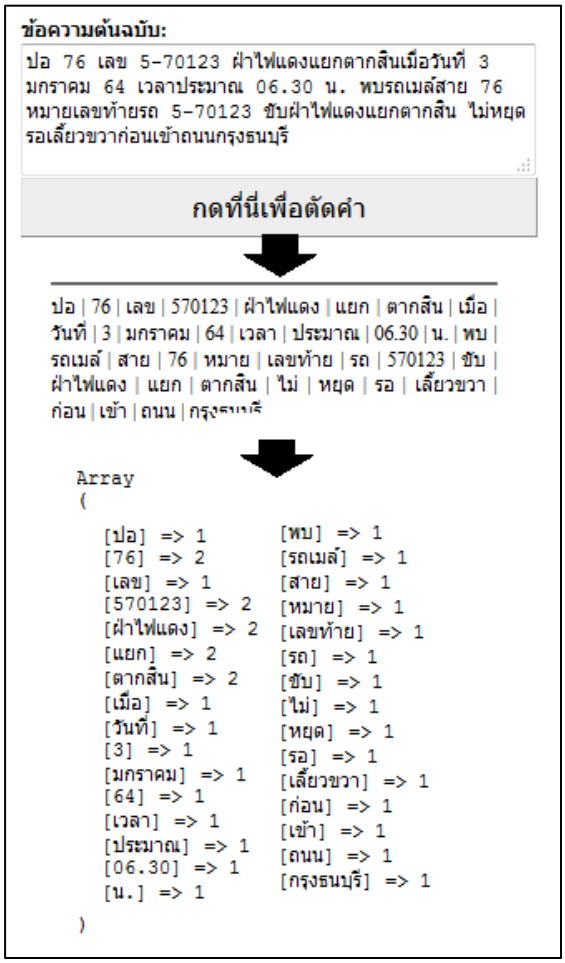
งานวิจัยนี้ใช้โปรแกรมตัดคำ THSplitlib [7] ที่พัฒนาด้วยภาษา PHP ซึ่งเป็นโปรแกรม Open Source สำหรับการตัดคำภาษาไทย ใช้หลักการตัดคำโดยใช้พจนานุกรมที่ปรับปรุงโดยผู้วิจัย ในการเปรียบเทียบการตัดคำกับคำที่จัดเก็บในพจนานุกรม เช่น ข้อความ “ขับฝ่าไฟแดงแยกตากสิน ไม่หยุดรอเลี้ยวขวาก่อนเข้าถนนกรุงธนบุรี” จะได้ผลลัพธ์ว่า “ขับ | ฝ่าไฟแดง | แยก | ตากสิน | ไม่ | หยุด | รอ | เลี้ยวขวา | ก่อน | เข้า | ถนน | กรุงธนบุรี” ดังภาพที่ 2



ภาพที่ 2 ผลลัพธ์การตัดคำภาษาไทยแบบอิงพจนานุกรม

3. การกำจัดคำหยุด (Stop-Word Removal) เป็นการนำคำที่ไม่มีนัยสำคัญต่อข้อความออก โดยที่ความหมายของคำหรือข้อความไม่เปลี่ยนแปลง คำหยุดจะปรากฏอยู่ในข้อความทุกข้อความ และมีความถี่สูง มักเป็น คำสรรพนาม คำสันธาน คำบุพบท เช่น ไว้, ไม่, ไป, ได้, ให้ เป็นต้น ถือได้ว่าคำหยุดเป็นคุณลักษณะที่ไม่เกี่ยวข้องหรือไม่มีประโยชน์ในการจำแนกหมวดหมู่ ดังนั้นจึงสามารถตัดคำดังกล่าวทิ้งได้เลย โดยที่ไม่ส่งผลกระทบต่อใจความหลัก การกำจัดคำหยุดเป็นกระบวนการที่ช่วยให้ขนาดของดัชนีลดลง อีกทั้งลดขนาดพื้นที่และเวลาในการประมวลผลด้วย ในงานวิจัยชิ้นนี้ใช้ข้อมูลคำหยุดของภาษาไทยจำนวน 115 คำที่รวบรวมจากเว็บไซต์ https://www.ranks.nl/ stopwords/thai-stopwords

4. คัดเลือกคำศัพท์ที่ใช้ในการจำแนกปัญหาการบริการเพื่อนำมาสร้างคลังคำศัพท์งานวิจัยนี้จำแนกปัญหาการให้บริการตามลักษณะของข้อร้องเรียนที่พบบ่อยในกระดาษสนทนาโดยทำการตัดคำจากข้อร้องเรียนจำนวน 389 ข้อความ นำมาตัดคำภาษาไทยโดยใช้พจนานุกรมแล้วนำมาหาความถี่ของคำที่พบมากที่สุด ดังภาพที่ 3



ภาพที่ 3 ความถี่ของคำศัพท์จากการตัดข้อความ

จากภาพที่ 3 จะเห็นว่าใจความหลักของข้อร้องเรียนดังกล่าวคือ การ “ฝ่าไฟแดง” ที่ถูกคัดเลือกเป็นคำศัพท์ที่จะนำมาใช้ในการจำแนกปัญหาการให้บริการต่อไป ซึ่งมีจำนวนความถี่เท่ากับ 2 เช่นเดียวกับ “570123”, “แยก” และ “ตากสิน” แต่ 3 คำดังกล่าวไม่ถูกคัดเลือก เนื่องจากคำแรกเป็นหมายเลขประจำเส้นทางของรถโดยสารสาธารณะเท่านั้น และอีกสองคำที่เหลือเป็นการระบุตำแหน่ง ซึ่งทั้ง 3 คำที่ไม่ถูกคัดเลือกนั้นไม่มีผลต่อการจำแนกปัญหาการให้บริการ

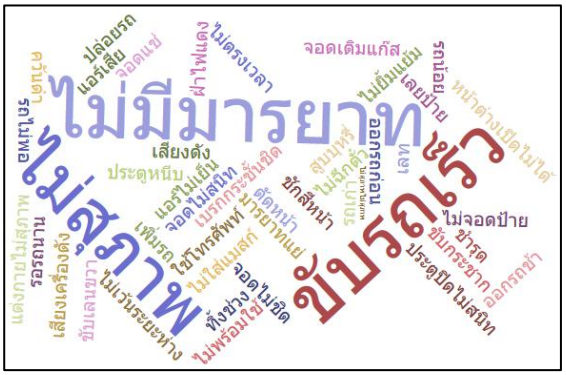
จากนั้นจะทำการตัดข้อความข้อร้องเรียนไปเรื่อย ๆ ทีละรายการแบบวนลูป จนครบ389 ข้อความตามที่กำหนดไว้ โดยจะนำผลลัพธ์จากการตัดคำนั้นมารวมเป็นความถี่ทั้งหมด แล้วคัดเลือกโดยคณะผู้วิจัย ได้คำศัพท์

จำนวน 128 คำ ที่สามารถจำแนกปัญหาการให้บริการได้ จากนั้นให้ผู้ทรงคุณวุฒิจำนวน 3 ท่าน จำแนกกลุ่มคำที่เกี่ยวข้องกับปัญหาการให้บริการรถโดยสารสาธารณะออกเป็น 4 คลาส ตามบริบทการให้บริการได้แก่ 1) คลาสผู้ขับขี่และพนักงานผู้ให้บริการ 2) คลาสการขับขี่ 3) คลาสยานพาหนะและอุปกรณ์ให้บริการ และ 4) คลาสเวลาและการเดินรถ จากนั้นผู้วิจัยจัดเก็บคำสำคัญในลักษณะคลังคำศัพท์ โดยใช้ MySQL เป็นฐานข้อมูลดังตารางที่ 2

ตารางที่ 2 การจัดเก็บคำสำคัญในคลังคำศัพท์

คลาสที่ 1 คลาสผู้ขับขี่และพนักงานผู้ให้บริการ จำนวน 32 คำ			
ไม่สุภาพ	ไม่มีมารยาท	เสียงดัง	ไม่ยิ้มแย้ม
ไม่ฉีกตั๋ว	แต่งกายไม่สุภาพ	ใช้โทรศัพท์	มารยาทแย่
ไม่ใส่แมสก์	สูบบุหรี่	...	ชักสีหน้า
คลาสที่ 2 คลาสการขับขี่ จำนวน 48 คำ			
ขับรถเร็ว	ขับกระชาก	เบรกกระชั้นชิด	ไม่จอดป้าย
เลยป้าย	ฝ่าไฟแดง	ขับเลนขวา	จอดไม่ชิด
ออกรถก่อน	จอดไม่สนิท	...	ตัดหน้า
คลาสที่ 3 คลาสยานพาหนะและอุปกรณ์ให้บริการ จำนวน 25 คำ			
แอร์เสีย	แอร์ไม่เย็น	ประตูหนีบ	รถเก่า
ไม่เว้นระยะห่าง	เสียงเครื่องดัง	ควันดำ	หน้าต่างเปิดไม่ได้
ประตูปิดไม่สนิท	ชำรุด	...	ไม่พร้อมใช้
คลาสที่ 4 คลาสเวลาและการเดินรถจำนวน 23 คำ			
ออกรถช้า	รอรถนาน	รถน้อย	จอดเติมแก๊ส
จอดแช่	ทิ้งช่วง	รถไม่พอ	ไม่ตรงเวลา
เลท	ปล่อยรถ	...	เพิ่มรถ

จากนั้นนำคำจากผลลัพธ์จากการตัดคำมาบันทึกไว้ในไฟล์ข้อความธรรมดา (.txt) มาแสดงผลด้วย Wordcloud ที่เป็นกลุ่มคำที่จับตัวกันเหมือนก้อนเมฆ เป็นเทคนิคที่ใช้สำหรับจับกลุ่มคำในภาษาใด ๆ โดยนับจำนวนจากมากไปหาน้อย แล้วทำการแสดงผล คำที่พบเจอมากก็จะกลุ่มกันเป็นก้อนเมฆที่ใหญ่ และไล่ลงมาเป็นก้อนเมฆเล็ก ๆ ตามลำดับ เป็นประโยชน์ในการทำรายงานข้อความ เพื่อให้มองเห็นคำที่ถูกใช้มากที่สุดได้ง่ายขึ้นด้วยการโปรแกรมมิ่งภาษา R โดยขนาดของคำขึ้นอยู่กับความถี่ที่พบจากผลลัพธ์จากการตัดคำดังภาพที่ 4



ภาพที่ 4 การแสดงความถี่ของคำด้วย Wordcloud

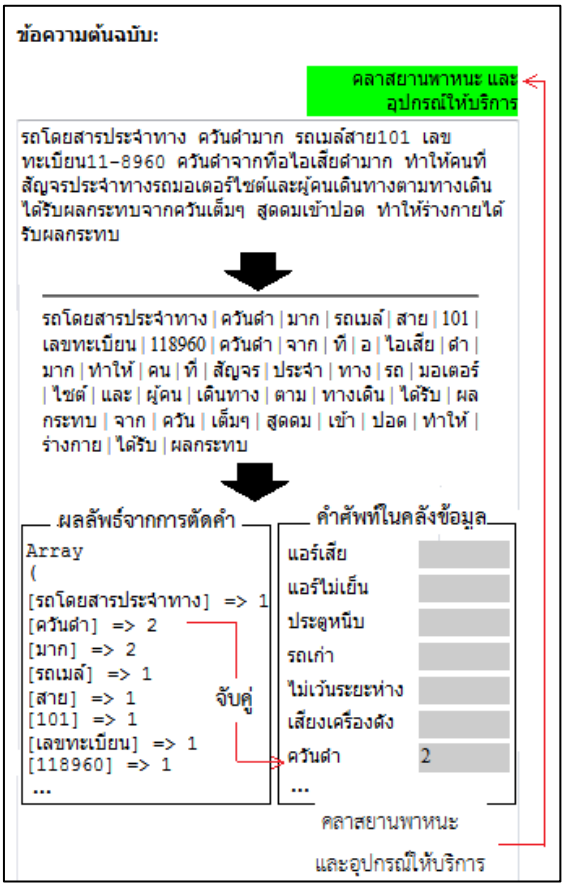
4. ผลการศึกษาและการอภิปรายผล

งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรม และเพื่อประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการแบ่งผลการศึกษาและการอภิปรายผล ดังนี้

ผลการออกแบบและพัฒนากระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรม เพื่อจำแนกปัญหาการให้บริการโดยเริ่มจากการนำข้อร้องเรียนที่เป็นชุดทดสอบจำนวน 252 ข้อความเข้าสู่เว็บแอพพลิเคชันทีละรายการ เพื่อทำการตัดข้อความตามกระบวนการที่ได้ออกแบบไว้ จากนั้นจะทำการจับคู่ผลลัพธ์จากการตัดคำกับคำสำคัญในคลังคำศัพท์ที่อยู่ในฐานข้อมูลดังภาพที่ 5

เมื่อตรวจสอบผลรวมของการจับคู่ผลลัพธ์จากการตัดคำกับคำสำคัญในคลังคำศัพท์ที่อยู่ในฐานข้อมูลของแต่ละคลาส หากคลาสไหนมีผลรวมมากกว่า 0 จะทำการติดแท็กด้วยชื่อของคลาสนั้น ๆ

การจับคู่ที่เกิดขึ้นในแต่ละครั้งนั้นจะทำการตรวจสอบเปรียบเทียบคำศัพท์จากทั้ง 4 คลาส ดังนั้นผลลัพธ์จำแนกปัญหาการให้บริการแต่ละรายการจึงสามารถมีได้มากกว่า 1 แท็ก จากนั้นจะบันทึกลงฐานข้อมูล เพื่อนำเสนอสารสนเทศเป็นภาพข้อมูลต่อไป

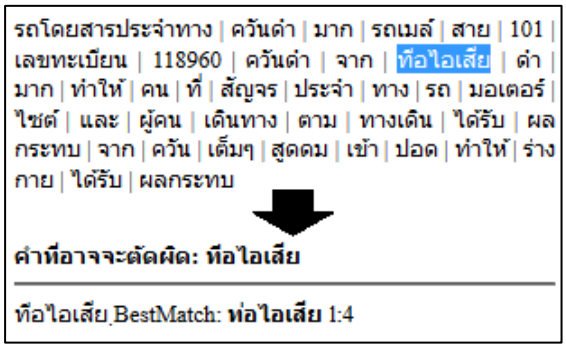


ภาพที่ 5 ผลการจำแนกปัญหาการให้บริการจากกระบวนสกัดข้อร้องเรียนรถโดยสารสาธารณะ

อย่างไรก็ดี เนื่องจากเป็นการให้ข้อร้องเรียนในการใช้บริการจากผู้ใช้ จึงมักเกิดปัญหาความหลากหลายของข้อความที่แตกต่างกันตามบริบทของผู้ใช้งาน ตลอดจนความผิดพลาดในการใช้ภาษาที่เกิดจากความตั้งใจหรือแม้แต่ไม่ตั้งใจของผู้ใช้ส่งผลให้ผลลัพธ์จากการตัดข้อความหากไม่ตรงกับคำสำคัญในคลังคำศัพท์ที่อยู่ในฐานข้อมูล จะไม่สามารถทำการจับคู่ได้ เช่น “ทีอไอเสีย” ซึ่งหมายถึงคำว่า “ท่อไอเสีย” เป็นต้น

ดังนั้นจะเห็นได้ว่า ความครอบคลุมของคลังคำศัพท์ส่งผลอย่างยิ่งต่อความถูกต้องของผลลัพธ์จากการสกัดข้อมูล อย่างไรก็ดี หากจะให้คลังคำศัพท์มีประสิทธิภาพจำเป็นเพิ่มคำศัพท์ที่สะกดผิด พิมพ์ไม่ครบถ้วน พิมพ์เกิน อันจะส่งผลให้คำศัพท์ในคลังข้อมูลมีจำนวนมากเกินความจำเป็น และยากที่จะระบุได้ว่าเท่าไรจึงจะครอบคลุมครบถ้วน

ผู้วิจัยจึงเห็นความสำคัญของปัญหาดังกล่าว จึงทำการโปรแกรมมิ่งด้วยภาษา PHP อีกครั้งด้วยฟังก์ชั่น levenshtein() เพื่อเปรียบเทียบข้อความ โดยจะคืนค่าของจำนวนตัวอักษรที่แตกต่างกัน ในการสร้างเงื่อนไขในการตรวจสอบความคล้ายคลึงของคำ โดยตรวจสอบผลลัพธ์จากการตัดข้อความมาเปรียบเทียบกับคำศัพท์ในฐานข้อมูล วิเคราะห์ออกมาเป็นเปอร์เซ็นต์แล้วเลือกแสดงคำที่มีเปอร์เซ็นต์ความคล้ายคลึงที่สุดออกมา ดังภาพที่ 6

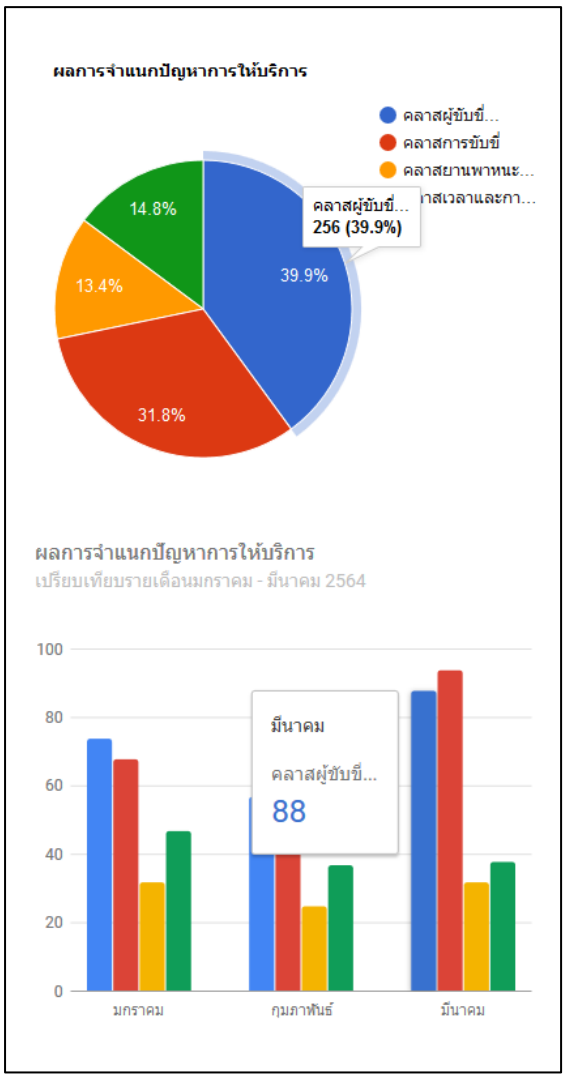


ภาพที่ 6 การเปรียบเทียบความคล้ายคลึงกันของอักขระ

จากภาพที่ 6 ผลการเปรียบเทียบความคล้ายคลึงกันของอักขระ (String similarity) ด้วยการหาค่าความต่างกันของสายอักขระสองชุดด้วยเทคนิคการวัดระยะทางเลเวนชเตย์น (Levenshtein edit distance) [8] ที่แก้ปัญหาโดยใช้กำหนดการพลวัตโดยค่าความต่างกันจะวัดจากจำนวนของอักขระที่จะต้องทำการตัดออก แทรก และ

แทนที่กับอักขระในชุดที่นำมาเปรียบเทียบจนกระทั่งมีลักษณะเหมือนชุดอักขระที่เป็นต้นแบบทุกประการ

การนำเสนอสารสนเทศในงานวิจัยชิ้นนี้จะใช้เทคนิคการนำเสนอภาพข้อมูลด้วยGoogle Charts [9] ที่สามารถแสดงผลจำนวนปัญหาการให้บริการจากการจำแนกข้อร้องเรียนที่เป็นข้อมูลเชิงปริมาณให้ง่ายต่อการทำความเข้าใจด้วยการเปรียบเทียบกันด้วยแผนภูมิวงกลม (Pie Chart) และสถิติเปรียบเทียบข้อร้องเรียนเป็นรายเดือนด้วยแผนภูมิแท่ง (Bar Chart) ดังภาพที่ 7 การนำเสนอสารสนเทศด้วยภาพข้อมูลผ่านเว็บไซต์ที่มีปฏิสัมพันธ์กับผู้ใช้แบบแอคทีฟ ด้วยการนำเมาส์ไปวางบนตามตำแหน่งที่ต้องการแล้วจะแสดงผลข้อมูลให้แก่ผู้ใช้



ภาพที่ 7 การนำเสนอภาพข้อมูลด้วย Google Charts

ผลการประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการประเมินผลโดยผู้ทรงคุณวุฒิ จำนวน 3 ท่าน ท่านละ 84 รายการ กำหนดผลการจำแนกปัญหาบริการแต่ละข้อร้องเรียน ดังนี้ หากแท็กการจำแนกปัญหาทั้งหมดถูกต้องได้ 1 คะแนน ในทางตรงกันข้าม หากแท็กผลการจำแนกปัญหาของข้อร้องเรียนใดจำแนกไม่ถูกต้องหรือถูกต้องแต่ไม่ครอบคลุมประเด็นปัญหาทั้งหมด ถือว่าได้ 0 คะแนน ผลการประเมินดังตาราง 3

ตารางที่ 3 ผลการประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการ

ผลการประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการ							
ท่านที่ 1		**ท่านที่ 2**		**ท่านที่ 3**		**ค่าเฉลี่ย**	
✓	ร้อยละ	✓	ร้อยละ	✓	ร้อยละ	ร้อยละ	แปลผล
75	89.28	77	91.67	62	73.81	84.92	ดีมาก

จากตารางที่ 3 พบว่า ผลการประเมินความถูกต้องของผลการจำแนกปัญหาการให้บริการอยู่ในระดับดีมาก แสดงถึงกระบวนการสกัดข้อร้องเรียนรถโดยสารสาธารณะออนไลน์ด้วยการตัดคำภาษาไทยแบบอิงพจนานุกรมให้ผลลัพธ์ความถูกต้องในระดับสูง โดยเฉพาะข้อร้องเรียนแบบ 1 ประเด็นต่อ 1 ข้อร้องเรียน เนื่องจากเหมาะกับคลังคำศัพท์ที่กำหนดขอบเขตได้แน่นอน ในทางกลับกันปัญหาที่พบส่วนใหญ่เนื่องมาจากมีคำศัพท์ที่ซ้ำซ้อนกันในบางคลาส เช่น เสียงดัง ที่จำเป็นต้องดูบริบทของข้อความประกอบ เนื่องจากเสียงดังเป็นกิริยาที่อาจเกิดจากบุคคลในคลาสผู้ขับขี่และพนักงานผู้ให้บริการ หรืออาจเป็นเสียงที่เกิดจากเครื่องยนต์จากคลาสยานพาหนะและอุปกรณ์ให้บริการดังนั้นการรวบรวมข้อมูลคำศัพท์ต้องมีกระบวนการที่น่าเชื่อถือ [10] และปรับปรุงให้ครอบคลุมครบถ้วนอยู่เสมอ ร่วมถึงอาจต้องจำแนกคลาสโดยคำนึงถึงบริบทของข้อความเป็นหลัก

เช่นเดียวกับการเปรียบเทียบความคล้ายคลึงกันของอักขระ ด้วยเทคนิคการวัดระยะทางเลเวนชเตย์นที่แม้จะสามารถทำงานได้อย่างมีประสิทธิภาพเป็นที่น่าพอใจ แต่อาจจะไม่เหมาะกับบริบทของภาษาไทยมากนัก เนื่องจากคุณลักษณะของภาษาไทยมิได้จำกัดเพียงจากการประสมกันของตัวอักษรเหมือนภาษาอังกฤษ แต่ภาษาไทยยังมีคุณลักษณะเฉพาะที่นอกจากเกิดจากการประสมกันของตัวอักษรแล้ว ยังมีสระ และวรรณยุกต์ที่เป็นคุณลักษณะของภาษาไทย และเป็นจุดมักเกิดข้อผิดพลาดในการใช้งานอีกด้วย อีกทั้งค่าความคลายคลึงของคำดังกล่าวเป็นผลมาจากการวิเคราะห์สายอักขระเท่านั้น มิได้เป็นผลมาจากความหมายของคำแต่อย่างใด อย่างไรก็ดีการตัดคำภาษาไทยจัดว่าเป็น NP-hard Problem เพราะไม่มีผิดถูกชัดเจน เช่น ตากลม แปลได้เป็น ตาก-ลมหรือ ตา-กลม เป็นต้น ดังนั้นการวัดความถูกต้องนอกจากจะวัดในเชิงความหมายแล้ว ยังต้องวัดในบริบทของการใช้งานด้วย

ข้อเสนอแนะในการวิจัย

เพื่อให้ข้อความมีความถูกต้องมากยิ่งขึ้นก่อนนำไปวิเคราะห์และประมวลผลข้อความ งานวิจัยในครั้งต่อไปควรให้ความสำคัญกับการกำจัดคำผิด โดยทำการลบ Noise ข้อความ เช่น รถไม่จอดดดดดดดดดดดดดดด, รถม่ายจอด ที่หมายความว่ารถไม่จอด เป็นต้นก่อนเข้ากระบวนการ Text Classification เช่น LSTM, ULMFiT หรือ GPT เป็นต้น

เอกสารอ้างอิง

[1] องค์การขนส่งมวลชนกรุงเทพ. (2562). *รายงานประจำปี 2562.* กรุงเทพฯ: องค์การขนส่งมวลชนกรุงเทพ.

[2] มหาวิทยาลัยมหิดล. (2561). *รายงานฉบับสมบูรณ์โครงการสำรวจความพึงพอใจของผู้ใช้บริการ ปี 2561.* กรุงเทพฯ: องค์การขนส่งมวลชนกรุงเทพ.

[3] สำนักงานปลัดสำนักนายกรัฐมนตรี. (2564). *ผลการดำเนินการเรื่องร้องทุกข์/เสนอความคิดเห็น ไตรมาสที่ 1 ปีงบประมาณ พ.ศ. 2564.* กรุงเทพฯ: สำนักงานปลัดสำนักนายกรัฐมนตรี.

[4] W. Wannaratn. (2016). “Test Score and Grading,”. *Journal of Humanities and Social Sciences, 2*(3), 1-11.

[5] R. Mitchell. (2015). *Web Scraping with Python.* (2nd ed). Sebastopol: O'Reilly MediaInc.

[6] C.Tapsai, H. Unger, and P.Meesad. (2021). *Thai Natural Language Processing Word Segmentation, Semantic Analysis, and Application.* Cham: Springer,

[7] สุวิชา เผือกอิ่ม. *THSplitLib.* สืบค้น เม.ย. 2564, จาก https://github.com/moohooooo/thsplitlib.

[8] F. P., Miller, A. F., Vandome, and J. McBrewster. (2009). *Levenshte in distance.* Alpha script Publishing.

[9] C. O., Wilke. (2009). *Fundamentals of Data Visualization A Primer on Making Informative and Compelling Figures.* Sebastopol: O’Reilly Media In.

[10] C. Santirattanaphakdi. (2020). “The Design and Development of a Knowledge Extraction and Retrieval System via Data Visualization Case Study: Road Major Accidents on Website,”. APHEIT journal science & technology, *9*(2), 17-32.