การวิเคราะห์ความสนใจของนักท่องเที่ยวในกรุงเทพมหานคร จากบทวิจารณ์ในทริปแอดไวเซอร์โดยใช้การเรียนรู้เชิงลึก

ANALYSIS OF TOURIST ATTRACTIONS IN BANGKOK FROM TRIPADVISOR REVIEWS USING DEEP LEARNING

ธุวานนท์ ขวัญเกื้อ

THUWANON KHWANKUEA

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2567
KMITL-2024-SC-M-017-026

ANALYSIS OF TOURIST ATTRACTIONS IN BANGKOK FROM TRIPADVISOR REVIEWS USING DEEP LEARNING

THUWANON KHWANKUEA

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS
KMITL DIGITAL ANALYTICS AND INTELLIGENT CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2024
KMITL-2024-SC-M-017-026

หัวข้อการค้นคว้าอิสระ	การวิเคราะห์ความสนใจของนักท่องเที่ยวในกรุงเทพมหานครจากบทวิจารณ์ในทริปแอดไวเซอร์โดยใช้การเรียนรู้เชิงลึก
ชื่อนักศึกษา	นายธุวานนท์ ขวัญเกื้อ
รหัสประจำตัว	65056047
ปริญญา	วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์) ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ.	2567
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ	ผู้ช่วยศาสตราจารย์ ดร.กนกกรรณ์ ลิ้โรจนาประภา

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อแบ่งกลุ่มความสนใจของ นักท่องเที่ยวชาวต่างชาติที่เดินทางมาท่องเที่ยวในกรุงเทพมหานคร ผ่านบทวิจารณ์ออนไลน์ภาษาอังกฤษจากเว็บไซต์ TripAdvisor ในหมวด Sights & Landmarks จำนวน 14,953 บทวิจารณ์ ซึ่งเป็นบทวิจารณ์หลังจากการเข้าเยี่ยมชมสถานที่ต่างๆ ทำการทดลองในขั้นตอนการเตรียมข้อมูล ประกอบด้วย 3 ปัจจัย การลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก การลบ/ไม่ลบคำที่มีความถี่ต่ำ (กำหนดไว้ที่ 10 คำ) และการแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming หรือ Lemmatization ทำให้ได้การทดลองทั้งสิ้น 8 การทดลอง จากนั้นทำการกำหนดจำนวนกลุ่มความสนใจของนักท่องเที่ยวเป็น 2 กลุ่มจากค่า Mean Silhouette Coefficient ที่มีค่ามากที่สุด และทำการวิเคราะห์ด้วยวิธี LDA ด้วยการเตรียมข้อมูลตามการทดลองที่ 8 ทำให้สามารถกำหนดกลุ่มบทวิจารณ์เป็นกลุ่มที่เกี่ยวกับความประทับใจในสถานที่และกลุ่มที่เกี่ยวกับการเข้าชมสถานที่ ผลการวิเคราะห์ความเด่นและความแพร่หลาย พบว่านักท่องเที่ยวจะกล่าวถึงกลุ่มความประทับใจในสถานที่มากกว่ากลุ่มการเข้าชมสถานที่ และนักท่องเที่ยวรู้สึกบวกกับกลุ่มความประทับใจในสถานที่ คำที่รู้สึกบวกได้แก่ visit, buddha, beautiful, one, must, worth, amazing, reclining, river, great และนักท่องเที่ยวรู้สึกลบกับกลุ่มการเข้าชมสถานที่ คำที่รู้สึกลบ ได้แก่ get, bath, long, go, see, entrance, around, inside, place, also สำหรับการเปรียบเทียบการจำแนกความรู้สึกเชิงบวกและเชิงลบของนักท่องเที่ยวต่างชาติจากบทวิจารณ์ตามคะแนนที่ให้กับสถานที่นั้นๆ ด้วยอัลกอริทึม Bidirectional Long Short-Term Memory (BiLSTM) และ Convolution Neural Network (CNN) ร่วมกับการการทดลองการแก้ปัญหาความไม่สมดุล 2 วิธี Undersampling และ SMOTE และการทดลองการเตรียมข้อมูลทั้ง 8 การทดลอง จากนั้นทำการแบ่งข้อมูลออกเป็น 3 ส่วน ได้แก่ ข้อมูลสำหรับฝึกสอนร้อยละ 70 ข้อมูลสำหรับตรวจสอบความถูกต้องร้อยละ 15 และข้อมูลทดสอบร้อยละ 15 พร้อมกับมีการปรับจูนไฮเปอร์พารามิเตอร์ด้วย Grid Search พบว่า โมเดล CNN ที่แก้ปัญหาความไม่สมดุลกันของชุดข้อมูลด้วยวิธี SMOTE ร่วมกับการเตรียมข้อมูลตามการทดลองที่ 5 (การลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และการแปลงคำด้วยวิธี Stemming) มีประสิทธิภาพในการจำแนกสูงที่สุดในชุดข้อมูลทดสอบ โดยมีค่าความถูกต้องร้อยละ 94.94 ค่าความเที่ยงร้อยละ 95.08 และค่าความไวร้อยละ 94.95

คำสำคัญ : การจัดสรรหัวข้อแฝง การเตรียมข้อมูล จำแนกความรู้สึก บทวิจารณ์

Independent Study Title	Analysis of Tourist Attractions in Bangkok from Tripadvisor Reviews Using Deep Learning
Student Name	Thuwanon Khwankuea
Student ID	65056047
Degree	Master of Science (Data Science and Analytics)
	KMITL-Digital Analytics and Intelligent Center
Year	2024
Independent Study Advisor	Asst. Prof. Dr. Kanongkan Leerojanaprapa

Abstract

This research aims to segment the interests of international tourists visiting Bangkok through English online reviews from TripAdvisor in the Sights & Landmarks category. There are 14,953 reviews that were written post-visit. The data preparation experiments involved three factors: removing/none removing words related to name of the places, removing/none removing low frequency (setting as 10) and converting words to their root forms using Stemming or Lemmatization methods. This resulted in 8 different experiments. Next, the number of tourist interest was set to two cluster based on the highest Mean Silhouette Coefficient and then LDA analysis on 8 Experiment of text preparation into two Topics: as overall impression of the place and visitation to a place. The Salience Valence analysis revealed tourists mentioned the impressions to a place more frequently than the visiting experience and had positive feedback towards the impressions to a place, with positive words including visit, buddha, beautiful, one, must, worth, amazing, reclining, river, and great. On the other hand, tourists had negative feedback towards the visitation to a place, with negative words including get, bath, long, go, see, entrance, around, inside, place, Next the classification of positive and negative sentiments of international tourists based on review rating using BiLSTM and CNN algorithms are compared. Two imbalance resolution methods, Undersampling and SMOTE, were experimented along with the 8 experiments of data preparation. The data was split into three parts: 70% for training, 15% for validation, and 15% for testing, along with Hyperparameter tuning using Grid Search. As a result, the CNN model, which addressed data imbalance using the SMOTE method combined with data preparation with Experiment 5 (removing words related to name of the places, not removing words with low frequency, and applying Stemming), demonstrated the highest classification efficiency on the test set, with an accuracy of 94.94%, precision of 95.08%, and recall of 94.95%.

Keywords : data preparation, lda, online reviews, sentiment analysis

กิตติกรรมประกาศ

การค้นคว้าอิสระเล่มนี้ สามารถสำเร็จได้ไปได้ด้วยดี เนื่องด้วยได้รับความกรุณา คำแนะนำ ความช่วยเหลือในด้านต่างๆ ทั้งด้านงานวิชาการและการดำเนินงานวิจัย ผู้วิจัยขอกราบขอบพระคุณ บุคคล และกลุ่มบุคคลดังต่อไปนี้

ผศ.ดร.กนกกรรณ์ สิ้โรจนาประภา อาจารย์ที่ปรึกษา ที่ได้ให้คำปรึกษา ให้แนวคิดต่างๆ ในการดำเนินงานวิจัย และช่วยตรวจทานความถูกต้องของการค้นคว้าอิสระ

รศ.ดร.ละออ บุญเกษม และ ผศ.ดร.วรางคณา กิ้มปาน กรรมการสอบ ที่ได้ให้ความกรุณาให้คำแนะนำเพิ่มเติม ตลอดจนจุดบกพร่อง รวมถึงช่วยตรวจทานแก้ไขข้อผิดพลาด

ขอบคุณนักศึกษาร่วมชั้นเรียนปริญญาโท ที่ให้คำปรึกษา แลกเปลี่ยนความรู้ทั้งด้านการเรียน การทำงาน ด้านวิชาการ ด้วยดีมาตลอด

สุดท้ายนี้ขอกราบขอบพระคุณ บิดา มารดา ที่ให้กำเนิด อบรม เลี้ยงดู ให้กำลังใจ ส่งเสริม ผู้วิจัยให้พัฒนาก้าวมาถึงจุดนี้ จนสามารถประสบความสำเร็จได้

ข้าพเจ้าผู้จัดทำหวังเป็นอย่างยิ่งว่าการค้นคว้าอิสระเล่มนี้ สามารถมอบประโยชน์ให้แก่ผู้ที่สนใจจะนำศึกษาต่อได้ ไม่ว่าจะเป็นในด้านความรู้พื้นฐาน แนวคิด ทฤษฎี และวิธีการวิจัยที่นำเสนอในเล่มนี้ รวมถึงการประยุกต์ใช้ผลการวิจัยในการพัฒนางานวิชาการหรือการทำงานในอนาคต

นายธุวานนท์ ขวัญเกื้อ

สารบัญ

สารบัญ (ต่อ)

สารบัญ (ต่อ)

สารบัญตาราง

สารบัญรูป

สารบัญรูป (ต่อ)

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

การท่องเที่ยวเป็นหนึ่งในอุตสากรรมสำคัญที่ขับเคลื่อนเศรษฐกิจของประเทศกระตุ้นให้มีการนำเอาทรัพยากรของประเทศไปใช้ให้เกิดประโยชน์นอกจากจะสร้างรายได้มหาศาลเข้าประเทศแล้วยังได้แผยแพร่วัฒนธรรมไทยสู่สายตาชาวต่างชาติที่เข้ามาท่องเที่ยวในประเทศ เกิดการจ้างงานในหลายภาคส่วนที่เกี่ยวข้อง ไม่ว่าจะเป็นการใช้จ่ายในส่วนโรงแรมที่พัก การจับจ่ายซื้อสินค้าของนักท่องเที่ยวรอบๆ สถานที่ท่องเที่ยวนั้นๆ ล้วนแล้วแต่สร้างรายได้ให้กับประชาชนในพื้นที่ โดยรายได้ภาคกาคการท่องเที่ยวก่อนโควิดนั้นเคยสร้างรายได้ให้กับประเทศไทยกว่า 3 ล้านล้านบาท หรือประมาณ 18% ของ GDP แบ่งเป็นต่างชาติประมาณ 2 ใน 3 และนักท่องเที่ยวไทย 1 ใน 3 (ธนาคารแห่งประเทศไทย, 2566) และกรุงเทพมหานครเป็นหมุดหมายปลายทางแรกที่นักท่องเที่ยวเลือกเดินทางเข้ามาประเทศไทย ทำให้เป็นจังหวัดที่มีนักท่องเที่ยวเดินทางเข้ามาท่องเที่ยวมากที่สุด ในปี 2566 ตั้งแต่เดือนมกราคม – ตุลาคม กว่า 21.29 ล้านคน สร้างรายได้กว่า 453,024.43 ล้านบาท (ศูนย์วิจัยด้านการตลาดการท่องเที่ยว, 2566)

กรุงเทพมหานครเป็นเมืองที่มีความหลากหลายทางวัฒนธรรม มีการผสมผสมทั้งโบราณสถานและความเป็นเมืองสมัยใหม่ได้อย่างลงตัว เสน่ห์ของกรุงเทพฯ ที่ทำให้เป็นที่ดึงดูดนักท่องเที่ยวมากมายจากทั่วโลก เช่น สถานที่ที่มีความสำคัญทางประวัติศาสตร์หรือศาสนา การท่องเที่ยวเชิงวัฒนธรรม ห้างสรรพสินค้าชั้นนำขนาดใหญ่มากมายทั่วกรุงเทพสำหรับนักท่องเที่ยวที่ชื่นชอบการชอปปิ้ง จากข้อมูลจากเว็บไซต์ Travelness ได้จัดอันดับให้กรุงเทพฯ เป็นเมืองที่นักท่องเที่ยวต่างชาติมาเยือนกว่า 22 ล้านคน มีการใช้จ่ายเฉลี่ยต่อวันที่ 173 ดอลล่าหรือประมาณ 5,800 บาท (Bessadeg, 2024) และยังเป็นเมืองที่มีนักท่องเที่ยวมาเยือนมาที่สุดในโลกในปี 2023 อีกด้วย

เพื่อหาแนวทางการวิเคราะห์เพื่อทำความเข้าใจความรู้สึกของนักท่องเที่ยวเชิงลึกของสถานที่ท่องเที่ยวในกรุงเทพมหานคร จากประสบการณ์ที่นักท่องเที่ยวที่เคยเดินทางมาท่องเที่ยวยังสถานที่ต่างๆ ในกรุงเทพฯ ได้แสดงความคิดเห็นหรือบทวิจารณ์ ข้อเสนอแนะ และมุมมองอื่นๆ อีกมากมายลงไปในเว็บไซต์ท่องเที่ยวที่น่าเชื่อถืออย่าง TripAdvisor ที่มีบทวิจารณ์กว่า 1 พันล้านบทวิจารณ์ ในปี 2022 (Statista, 2022) เป็นประโยชน์กับนักท่องเที่ยวจากทั่วโลกที่วางแผนจะเดินทางมาท่องเที่ยวในกรุงเทพฯ ใช้ในการพิจารณาวางแผนการเดินทางท่องเที่ยวตามสถานที่ต่างๆ ซึ่งเป็นหนึ่งในตัวแปรสำคัญที่ทำให้นักท่องเที่ยวตัดสินใจเดินทางมาท่องเที่ยวหรือไม่ ซึ่งการนำข้อมูลแสดงความรู้สึกเหล่านั้นมาวิเคราะห์ เพื่อที่จะนำไปสู่การเข้าใจนักท่องเที่ยวที่เขียนบทวิจารณ์เกี่ยวกับสถานที่ต่างๆ ในกรุงเทพฯ จะสามารถนำไปปรับปรุงกับสถานที่ท่องเที่ยวต่างๆ ในกรุงเทพฯ ที่คล้ายคลึงกัน เพื่อสร้างความประทับใจให้แก่นักท่องเที่ยวที่มาเยือนได้ในอนาคตต่อไป

ดังนั้นในงานวิจัยนี้จากที่กล่าวมาข้างต้นเล็งเห็นถึงความสำคัญในการพัฒนาการท่องเที่ยวโดยจะทำการหาข้อมูลเชิงลึกจากของนักท่องเที่ยว ความพึงพอใจ อารมณ์ความรู้สึก ที่มีต่อสถานที่ท่องเที่ยวต่างๆ ในกรุงเทพมหานคร จากเว็บไซต์ TripAdvisor โดยอยู่บนพื้นฐานของฐานข้อมูลการจัดประเภทสถานที่ท่องเที่ยวยอดนิยมในกรุงเทพฯ จากการเก็บข้อมูลของตัวเว็บไซต์เอง โดยจะมีหมวดหมู่ที่นำมาพิจารณาคือหมวด Sights & Land marks 4 สถานที่ คือ 1. Wat Phra Chetuphon

2. Temple of Dawn (Wat Arun) 3. The Grand Palace 4. Temple of Emerald Buddha (Wat Phra Kaew) ที่เป็นบทความภาษาอังกฤษ มาวิเคราะห์โดยใช้ศาสตร์ด้านการประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) และการเรียนรู้เชิงลึกมาประยุกต์ใช้

ในงานวิจัยนี้ได้อ้างอิงหลักการวิเคราะห์ความเด่นและความแพร่หลาย (Salience - Valence Analysis) จากงานวิจัยการวิเคราะห์ความสนใจของนักท่องเที่ยวในจังหวัดภูเก็ตของ Taecharungroj and Mathayomchan (2019) ซึ่งเป็นเครื่องมือที่ช่วยในการตีความความสนใจของนักท่องเที่ยว มาใช้ในการวิเคราะห์ความสนใจของนักท่องเที่ยวในขอบเขตระดับปริญาโท โดยงานวิจัยนี้มีการทดลองในขั้นตอนการเตรียมข้อมูล ประกอบด้วย 3 ปัจจัย การลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก การคำลบ/ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำออก และการแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming หรือ Lemmatization ทำให้ได้การทดลองทั้งสิ้น 8 การทดลอง ในส่วนของการวิเคราะห์จะแบ่งออกเป็น 2 ส่วน โดยส่วนแรกจะเป็นโมเดลการเรียนรู้แบบไม่มีผู้สอน โดยจะใช้การจัดสรรหัวข้อแฝงหรือแอลดีเอ (Latent Dirichlet Allocation: LDA) ใช้ในการแบ่งกลุ่มความสนใจของนักท่องเที่ยว ร่วมกับการกำหนดกลุ่มความสนใจของนักท่องเที่ยวที่เหมาะสมจากค่า Mean Silhouette Coefficient ที่มีค่ามากที่สุด และส่วนที่ 2 สำหรับการเปรียบเทียบการจำแนกความรู้สึกเชิงบวกและเชิงลบของนักท่องเที่ยวต่างชาติจากบทวิจารณ์ตามคะแนนที่ให้กับสนานที่นั้นๆ ด้วยอัลกอริทึม Bidirectional Long Short-Term Memory (BiLSTM) และ Convolution Neural Network (CNN) ร่วมกับการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล 2 วิธี Undersampling และ SMOTE และการทดลองการเตรียมข้อมูลทั้ง 8 การทดลอง จากนั้นแบ่งข้อมูลออกเป็น 3 ส่วน ได้แก่ ข้อมูลสำหรับฝึกสอนร้อยละ 70 ข้อมูลสำหรับตรจสอบความถูกต้องร้อยละ 15 และข้อมูลทดสอบร้อยละ 15 พร้อมกับมีการปรับจูนไฮเปอร์พารามิเตอร์ด้วย Grid Search

1.2 วัตถุประสงค์ของงานวิจัย

1) แบ่งกลุ่มความสนใจของนักท่องเที่ยวต่างชาติที่เดินทางมาท่องเที่ยวในกรุงเทพมหานครในหมวด Sights & Landmarks จากบทวิจารณ์ออนไลน์
2) เปรียบเทียบการจำแนกความรู้สึกของนักท่องเที่ยวต่างชาติที่เดินทางมาท่องเที่ยวด้วยอัลกอริทึม Bidirectional Long Short-Term Memory (BiLSTM) และ Convolution Neural Network (CNN)

1.3 ขอบเขตของงานวิจัย

1) ข้อมูลที่เป็นบทวิจารณ์ที่ใช้ในการวิเคราะห์เป็นภาษาอังกฤษจากเว็บไซต์ TripAdvisor
2) ข้อมูลที่นำวิเคราะห์เริ่มตั้งแต่เดือน มกราคม ปี ค.ศ. 2018 จนถึงเดือน กันยายน ค.ศ.2023
3) บทวิจารณ์ที่ใช้ในการวิเคราะห์เป็นบทวิจารณ์ของนักท่องเที่ยวที่เคยไปเที่ยวกรุงเทพมหานครในหมวด Sights & Landmarks

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) รับรู้ถึงหัวข้อสำคัญที่นักท่องเที่ยวให้ความสนใจ เพื่อที่จะเป็นแนวทางในการพัฒนาปรับปรุงการท่องเที่ยว เพื่อตอบสนองต่อกลุ่มนักท่องเที่ยวต่างประเทศ
2) การพัฒนาแบบจำลองโดยใช้อัลกอรึทึมด้านการเรียนรู้เชิงลึกให้มีประสิทธิภาพในการทำนายที่สูง
3) หน่วยงานที่เกี่ยวข้องการการท่องเที่ยวสามารถนำไปใช้ในการออกแผนงานที่สามารถจับกลุ่มเป้าหมายได้ตรงเป้า จากโมเดลการจัดสรรหัวข้อแฝง
4) นำเอาการพยากรณ์ความรู้สึกที่ได้ไปใช้วิเคราะห์กับบทวิจารณ์ที่เกี่ยวข้องกับสถานที่นั้นๆ เพื่อรับทราบถึงปัญหา ปรับปรุงหรือพัฒนาให้ดีขึ้น

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการศึกษาครั้งนี้ผู้วิจัยได้นำเสนอเนื้อหาที่เน้นถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องโดยมีรายละเอียดของเนื้อหาประกอบ ดังต่อไปนี้

2.1 การเตรียมข้อมูลบทวิจารณ์

2.2 การเรียนรู้แบบไม่มีผู้สอนเพื่อจัดกลุ่มความสนใจของนักท่องเที่ยว

2.3 การเรียนรู้แบบมีผู้สอนเพื่อทำนายความรู้สึกของนักท่องเที่ยว

2.4 การเปรียบเทียบประสิทธิภาพของการทำนาย

2.5 การวิเคราะห์ความเด่นและความแพร่หลาย

2.6 งานวิจัยที่เกี่ยวข้อง

2.1 การเตรียมข้อมูลบทวิจารณ์

การเตรียมเตรียมข้อมูลในงานวิจัยนี้อ้างอิงหลักการเตรียมข้อมูลรูปแบบภาษาอังกฤษจาก Liu et al. (2023) โดยมีขั้นตอนดังนี้

2.1.1 ลบบทวิจารณ์ที่ซ้ำออก (Remove Duplicate Reviews)

ทำการลบแถวที่มีข้อมูลที่ซ้ำๆ กันออก อันเนื่องมาจากการเก็บข้อมูล (Scrapping) จากเว็บไซต์อาจมีการเก็บข้อมูลซ้ำๆ กันได้จึงต้องทำการตรวจสอบและลบออกไปให้เหลือเพียงข้อมูลเดียว

2.1.2 แปลงบทวิจารณ์ทั้งหมดให้เป็นตัวพิมพ์เล็ก (Convert Reviews to Lowercase Form)

บทวิจารณ์ทั้งหมดจะถูกแปลงให้เป็นตัวพิมพ์เล็กทั้งหมดเนื่องจากคอมพิวเตอร์นั้น จะเข้าใจความหมายของตัวอักษรตัวพิมพ์ใหญ่และตัวพิมพ์เล็กแตกต่างกัน กรณีที่เป็นคำๆ เดียวกันแต่ตัวขึ้นต้นตัวนึงเป็นตัวพิมพ์เล็ก อีกตัวเป็นตัวพิมพ์ใหญ่จะถือว่าเป็นคนละคำกัน เช่น “Temple” กับ “temple” หมายถึง วัดเหมือนกันแต่ คอมพิวเตอร์เข้าจะเข้าใจว่าเป็นคนละคำกัน จึงต้องแปลงให้ตัวพิมพ์ใหญ่ทั้งหมดเป็นตัวพิมพ์เล็กในลักษณะเดียวกันทั้งหมด คอมพิวเตอร์จึงสามารถเข้าใจลักษณะคำที่เหมือนกันได้

2.1.3 ลบลิงก์ ตัวอักษรพิเศษ รวมทั้งเครื่องหมายวรรคตอน ที่ไม่เกี่ยวข้องออก (Remove Links, Special Character, Punctuation)

ทำการลบคำอื่นๆ พวกเครื่องหมายหรือสัญลักษณ์ต่างๆ ที่ไม่เกี่ยวข้องกับบทวิจารณ์ออกไป เนื่องจากไม่มีความสำคัญหรือจุดสนใจที่จะนำมาวิเคราะห์รวมถึงตัวเลขด้วย เช่น เครื่องหมายคำถาม (?) เรื่องหมายตกใจ (!) หรือเครื่องหมายอิโมจิ เป็นต้น

2.1.4 ทำการตัดคำให้กับบทวิจารณ์ (Break Reviews into Words)

ขั้นตอนการตัดคำหรือแยกคำในบทวิจารณ์จากที่อยู่ในรูปของประโยค ให้อยู่ในรูปของคำศัพท์เดี่ยวๆ ที่เรียกว่า Token ซึ่งในภาษาอังกฤษการแบ่งคำออกเป็นคำเดี่ยวๆ สามารถทำได้ โดยการใช้ช่องว่างในแต่ละคำ ในการตัดคำออกมา จะได้เป็น Token ของชุดข้อมูล เช่น I took 2 hours to explore around the area. ตัดคำได้เป็น “I”, “took”, “2”, “hours”, “to”, “explore”, “around”, “the”, “area”, “.”

2.1.5 ลบคำที่ไม่สื่อความหมาย (Remove Stop Words)

ลบคำที่พบได้บ่อยแต่ไม่มีความหมายที่สำคัญ เช่น คำพวก a, the, is, and ที่ไม่มีความหมายเกี่ยวข้องกับเนื้อหาออกไป ช่วยลดขนาดของข้อมูลที่จะต้องใช้ในการประมวลผล เช่น We plan to go to Thailand next week. จะได้เป็น “plan”, “go”, “Thailand”, “next”, “week”

2.1.6 แปลงคำให้อยู่ในรูปของรากศัพท์ (Convert Words to Root Form)

การแปลงคำๆ เดียวกันให้อยู่ในรูปพื้นฐานหรือรากของคำศัพท์นั้นๆ โดยเลือก 2 วิธีหลักๆ ที่ใช้ในงานด้าน NLP คือ แบบ Stemming และแบบ Lemmatization สามารถดึงมาใช้ได้ จาก Library ชื่อ NLTK

Stemming คือ การตัดแบบหยาบๆ โดยจะตัดส่วนท้ายของคำออก ทำให้การสะกดคำอาจไม่ถูกต้องในหลักภาษา ทำให้คำศัพท์ที่ได้อาจไม่ได้มีอยู่จริงๆ ได้ เช่น studies จะตัดได้ studi เหมาะสำหรับการทำงานที่ต้องการความรวดเร็วและงานที่ไม่เน้นไปที่ความหมายของคำ

Lemmatization คือ กระบวนการแปลงคำ โดยมีการพิจาณาถึง ความหมายและบริบทของคำตามหลักไวยากรณ์ ด้วยรายการคำศัพท์ใน Dictionary อย่างเหมาะสม ส่วนใหญ่จะตัดส่วนท้ายของคำ เช่น “runner” จะได้ “runner” คำเดิมเมื่อพิจาณาแล้วว่าเป็นคำนาม หรือคำว่า “runs” จะได้ “run” เป็นคำกริยาแปลงเป็นรูปพื้นฐานได้โดยตัด s ออก ดังนั้น จึงมีความแม่นยำกว่าวิธี Stemming แต่ใช้เวลามากกว่า

2.1.7 ถุงคำศัพท์ (Bag of Words: BOW)

Bag of Word เป็นการแปลงข้อมูลรูปแบบข้อความ (Text) ให้อยู่ในรูปแบบที่คอมพิวเตอร์สามารถเข้าใจได้ ซึ่ง Bag of Words เป็นวิธีในการสร้าง Feature ของข้อความขึ้นมา โดยใช้หลักการ One-Hot Encoding ในการเข้ารหัสข้อมูลในทุกคำของข้อมูล โดยการเข้ารหัสนั้นจะเป็นการเข้ารหัสผ่าน Token ที่ได้ทำการตัดคำเอาไว้แล้ว โดยเริ่มจากการสร้างคลังคำศัพท์จากชุดข้อมูลที่ใช้ในการวิเคราะห์ ซึ่งจะแทนค่าเป็น 1 เมื่อ คำเหล่านั้นปรากฏขึ้นในชุดข้อมูล และเป็น 0 เมื่อ คำเหล่านั้นไม่ปรากฏในชุดข้อมูล ตัวอย่างเช่น ประโยคที่ 1 “First time in Pattaya and second time in Bangkok” ประโยคที่ 2 “First time in Thailand ” จากทั้ง 2 ประโยคจะได้พจนานุกรมของคำศัพท์ที่ว่ามีคำว่า [“First”, “time”, “in”, “Pattaya”, “and”, “second”, “Bangkok”, “Thailand”] ดังนั้น จะได้เวกเตอร์ของประโยคที่ 1 คือ [1, 2, 2, 1, 1, 1, 1,0] มีคำว่า “time” และ “in” 2 ครั้ง นอกนั้นมี 1 ครั้ง และเวกเตอร์ของประโยคที่ 2 คือ [1, 1, 1, 0, 0, 0, 0, 1] ไม่มีคำว่า “Pattaya”, “and”, “second”, “Bangkok”

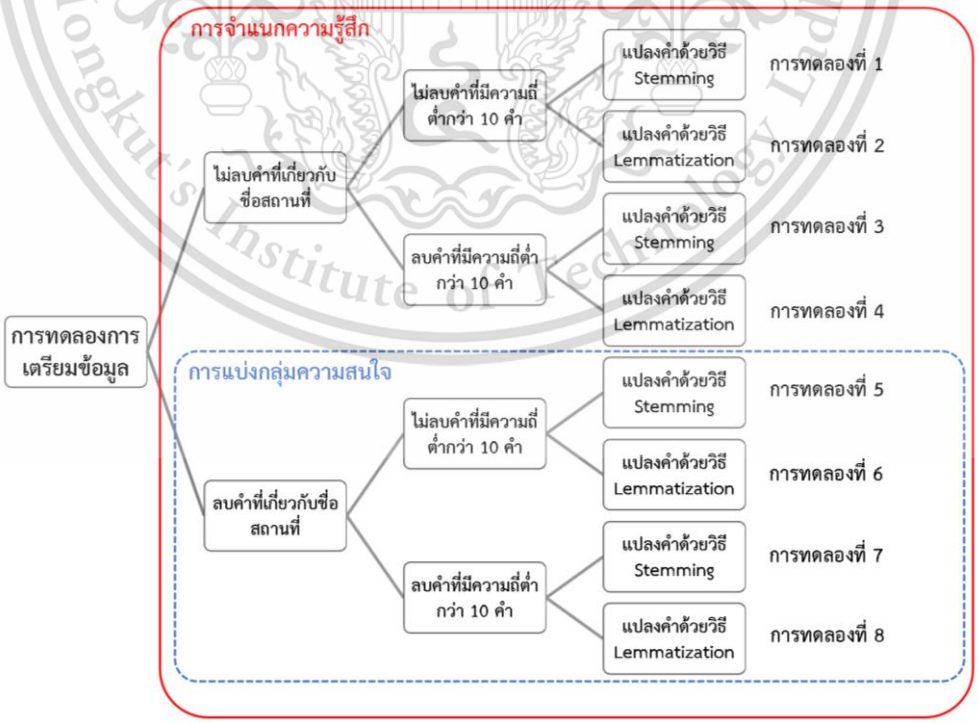
2.1.8 การทดลองการเตรียมข้อมูล

ในหัวข้อ 2.1.5 มีการเพิ่มการลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก เป็นคำไม่สื่อความหมายตามหลักการของ Taecharungroj and Mathayomchan (2019) ในการเข้าโมเดลการจัดสรรหัวข้อแฝงเกี่ยวกับสถานที่ท่องเที่ยวต้องมีการลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก เพราะจะทำให้ชื่อสถานที่ท่องเที่ยวจะมาเป็นคำศัพท์ที่มีค่าน้ำหนักสูงในกลุ่มความสนใจแทน ทำให้ไม่สามารถหาข้อมูลเชิงลึกได้ จากสถานที่ท่องเที่ยวทั้งหมดมีคำที่ต้องลบออก 13 คำ ดังนี้ “bangkok”, “temple”, “grand”, “palace”, “thailand”, “wat”, “phra”, “chetuphon”, “dawn”, “arun”, “emerald”, “kaew”, “thai”

เพิ่มขั้นตอนในการทดลองเปรียบเทียบการลบ/ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำออก เนื่องจากการวิเคราะห์ด้วยการเรียนรู้ของเครื่อง การที่บางคำที่มีความถี่น้อยมากๆ นั้นอาจจะไม่จำเป็นต่อการเรียนรู้ของเครื่อง โดยกำหนดความถี่ของคำไว้ที่ 10 คำ และช่วยลดเวลาในการเรียนรู้ของเครื่องได้ (จิระเมศร์ รุจิกรหิรัณย์, 2565)

ในหัวข้อ 2.1.6 เพิ่มการทดลองการแปลงคำศัพท์ให้อยู่ในรูปรากศัพท์ทั้ง 2 วิธี ทั้งแบบ Stemming และ Lemmatization เปรียบเทียบกัน

ดังนั้น สามารถสรุปรูปแบบการเตรียมข้อมูลได้ทั้งหมด 8 การทดลอง (8 Experiment) ดังรูปที่ 2.1 โดยแบ่งเป็น 2 ส่วน ส่วนที่เกี่ยวกับการแบ่งกลุ่มความสนใจ ประกอบด้วย การทดลองที่ 5 - 8 ตามกลักการของ Taecharungroj and Mathayomchan (2019) และส่วนที่ 2 การทดลอง ที่ 1 - 8 เนื่องจากไม่เกี่ยวข้องกับการแบ่งกลุ่มความสนใจของนักท่องเที่ยว จึงสามารถนำการเตรียมข้อมูลทั้ง 8 การทดลอง มาใช้สำหรับการจำแนกความรู้สึก



รูปที่ 2.1 รูปแบบการเตรียมข้อมูลทั้งหมด

2.1.9 การระบุประเภทของความรู้สึกในบทวิจารณ์ (Class Labeling)

ในบทวิจารณ์จากเว็บไซต์ TripAdvisor มีการให้คะแนนเรตติงประกอบคู่กับบทวิจารณ์ ต่ำสุดที่ 1 คะแนน และสูงสุดที่ 5 คะแนนโดยจากงานวิจัยของ Manurung and Lhaksmana (2023) ได้ระบุวิธีการแบ่งความรู้สึกของบทวิจารณ์โดยอิงกับคะแนนเรตติง ได้ทำการแบ่งแยกระดับเรตติงจาก 5 เรต คือ เรตระดับบน 4 - 5 คะแนน ทำการระบุเป็นความรู้สึกเชิงบวก (Positive) ส่วนเรตระดับล่าง 1 - 3 คะแนน ทำการระบุเป็นความรู้สึกเชิงลบ (Negative)

2.2 การเรียนรู้แบบไม่มีผู้สอนเพื่อจัดกลุ่มความสนใจของนักท่องเที่ยว

ใช้ 2 เทคนิค ในการวิเคราะห์ การการกำหนดจำนวนกลุ่มที่เหมาะสมด้วย Mean Silhouette Coefficient และโมเดลการจัดสรรหัวข้อแฝง (LDA)

2.2.1 การกำหนดจำนวนกลุ่มที่เหมาะสมด้วย Mean Silhouette Coefficient

โดยทั่วไปแล้วในการหาจำนวนกลุ่มความสนใจของโมเดลการจัดสรรหัวข้อแฝง (LDA) จะต้องมีการกำหนดค่า k กลุ่มความสนใจที่เหมาะสมก่อน โดยการกำหนดจำนวนกลุ่มที่เหมาะสมในงานวิจัยนี้จะอ้างอิงหลักการจาก Jiang et al. (2017) และ Panichella et al. (2013) ใช้วิธีหาจาก Mean Silhouette Coefficient ให้ทำการมองกลุ่มความสนใจ (Topics) เป็นกลุ่มของ Cluster โดยที่จุดภายใน Cluster เป็นบทวิจารณ์ ดังนั้นสามารถหาค่า Silhouette Coefficient ดังสมการที่ (2.1)

$$s(di) = \frac{b(di) - a(di)}{\max(a(di), b(di))} \tag{2.1}$$

เมื่อ $a(di)$ คือ ระยะทางมากสุดจากบทวิจารณ์ (di) ไปยังบทวิจารณ์อื่นๆ ในกลุ่ม (Cluster)

$b(di)$ คือ ระยะทางน้อยสุดที่สุดจากบทวิจารณ์ (di) ไปยัง Centroid ของกลุ่มอื่นที่ ไม่ใช่กลุ่ม (Cluster) ของบทวิจารณ์ (di) นั้น

$s(di)$ คือ Silhouette Coefficient ของบทวิจารณ์ (di)

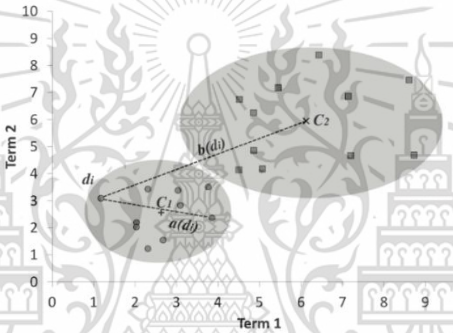
สุดท้ายจะทำการหาค่าเฉลี่ยของการแบ่งกลุ่ม (Cluster) ด้วย Mean Silhouette Coefficient จากสมการที่ (2.2) เป็นค่าเฉลี่ยของกลุ่ม (Cluster)

$$s(C) = \frac{1}{n}\sum_{i=1}^{n} s(d_i) \tag{2.2}$$

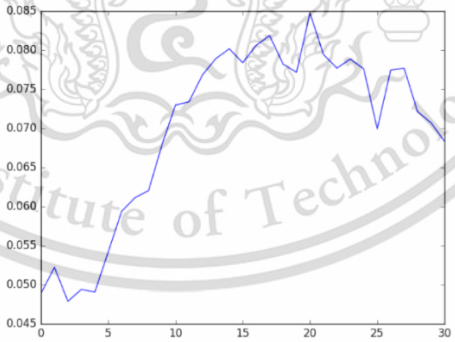
$s(C)$ คือ Mean Silhouette Coefficient โดยที่ C คือ กลุ่ม (Cluster) จากการแบ่งกลุ่มด้วยโมเดลการจัดสรรหัวข้อแฝง (LDA) และ n คือ จำนวนบทวิจารณ์ทั้งหมดใน Cluster

จากรูปที่ 2.2 แสดงตัวอย่างการคำนวณหา Silhouette Coefficient ของสมการที่ (2.1) โดยที่ C แทนกลุ่ม (Cluster) ที่มีจุด Centroid เท่ากับค่าเวกเตอร์เฉลี่ยของบทวิจารณ์ทั้งหมดและรูปที่ 2.3 แสดงตัวอย่างการหาจำนวนกลุ่มความสนที่เหมาะสม (Topic) จากกราฟ Silhouette Coefficient Score (Mean Silhouette Coefficient) ที่แตกต่างกันตั้งแต่ 0 - 30 กลุ่ม จากโมเดลการจัดสรรหัวข้อแฝง (LDA) จะเห็นได้ว่าที่กลุ่มความสนใจที่ 20 นั้นมี Silhouette Coefficient Score (Mean Silhouette Coefficient) ที่มากที่สุด ดังนั้นจะได้ว่ากลุ่มความสนใจที่เหมาะสมสำหรับโมเดลการจัดสรรหัวข้อแฝง (LDA) เท่ากับ 20 กลุ่มความสนใจ

โดยค่าที่ได้จากการคำนวณ Silhouette Coefficient มีค่าอยู่ในช่วงระหว่าง -1 ถึง 1 ยิ่งใกล้ 1 มาก คือดี โดยจะใช้ Mean Silhouette Coefficient ในการหาจำนวนกลุ่มความสนใจที่เพื่อใช้เป็นตัวกำหนดกลุ่มความสนใจของนักท่องเที่ยวในโมเดลจากการจัดสรรหัวข้อแฝง (LDA) โดยทำการแบ่งกลุ่มความสนใจตั้งแต่ 2 - 10 กลุ่ม แล้วเลือก Mean Silhouette Coefficient ในกลุ่มที่มีค่ามากที่สุด



รูปที่ 2.2 การคำนวณหา Silhouette Coefficient
(ที่มา Panichella et al., 2013)



รูปที่ 2.3 ตัวอย่างกราฟแสดงค่า Silhouette Coefficient Score
(ที่มา Jiang et al., 2017)

2.2.2 การจัดสรรหัวข้อแฝง (Latent Dirichlet Allocation: LDA)

Latent Dirichlet Allocation (LDA) หรือโมเดลการจัดสรรหัวข้อแฝงเป็นเครื่องมือที่ช่วยให้สามารถสกัดหัวข้อหรือประเด็นที่สำคัญที่ซ่อนอยู่ในข้อมูลเอกสารขนาดใหญ่ได้ โดย LDA จะทำการนำเอกสารเหล่านี้มาแยกออกเป็นโครงสร้างหรือการกระจายของคำต่างๆ ซึ่งแต่ละเอกสารจะมีการกระจายของคำที่แสดงถึงหัวข้อต่างๆ ที่เกี่ยวข้องกับเอกสารนั้นๆ ทำให้สามารถค้นหาและจัดกลุ่มคำ

ให้อยู่ในกลุ่มหัวข้อที่เกี่ยวข้องกันได้อย่างเป็นระบบ ซึ่งหัวข้อเหล่านี้มักจะเป็นโครงสร้างที่เรียกว่า “หัวข้อแฝง” หรือ “ประเด็น” ที่ซ่อนอยู่ภายในข้อมูล โดยโมเดลจะกำหนดความน่าจะเป็นของการกระจายคำในแต่ละหัวข้อด้วยการใช้ Distribution เช่น Dirichlet Distribution และจากนั้นจึงสร้างโมเดลที่สามารถสร้างข้อมูลที่สอดคล้องกับการกระจายของคำในข้อมูลต้นฉบับ สมการการจัดสรรหัวข้อแฝงสามารถเขียนได้ดังสมการที่ (2.3)

$$P(W,Z,\theta,\varphi;\alpha,\beta)=\prod_{j=1}^{M}P(\theta_j;\alpha)\prod_{i=1}^{K}P(\varphi_i;\beta)\prod_{t=1}^{N}P(Z_{j,t}\mid\theta_j)P(W_{j,t}\mid\varphi_{Z_{j,t}}) \qquad (2.3)$$

เมื่อ α คือ พารามิเตอร์ควบคุมการกระจายตัวของของหัวข้อจากการแจกแจงแบบดีรีเคล (Dirichlet Distribution)

θ_j คือ ความน่าจะเป็นของแต่บทวิจารณ์ลำดับที่ j

φ_i คือ ความน่าจะเป็นของบทวิจารณ์ที่จะเป็นหัวข้อที่ i

β คือ พารามิเตอร์ควบคุมการกระจายตัวของคำจาก ความน่าจะเป็นของคำศัพท์กับหัวข้อแฝง

$Z_{j,t}$ คือ หัวข้อแฝงของบทวิจารณ์ที่ j ของการสุ่มครั้งที่ t

$W_{j,t}$ คือ คำศัพท์ที่สุ่มได้จากกลุ่มหัวข้อแฝงที่มาจากการความน่าจะเป็นของหัวข้อแฝงของบทวิจารณ์ลำดับที่ j ครั้งที่ t

$\varphi_{Z_{j,t}}$ คือ กลุ่มหัวข้อแฝงจากความน่าจะเป็นของหัวข้อแฝง ในบทวิจารณ์ที่ j จากการสุ่มครั้งที่ t

M คือ ลำดับของบทวิจารณ์ที่ $j=1,2,3,...,M$

N คือ จำนวนคำศัพท์ที่ปรากฏทั้งหมด

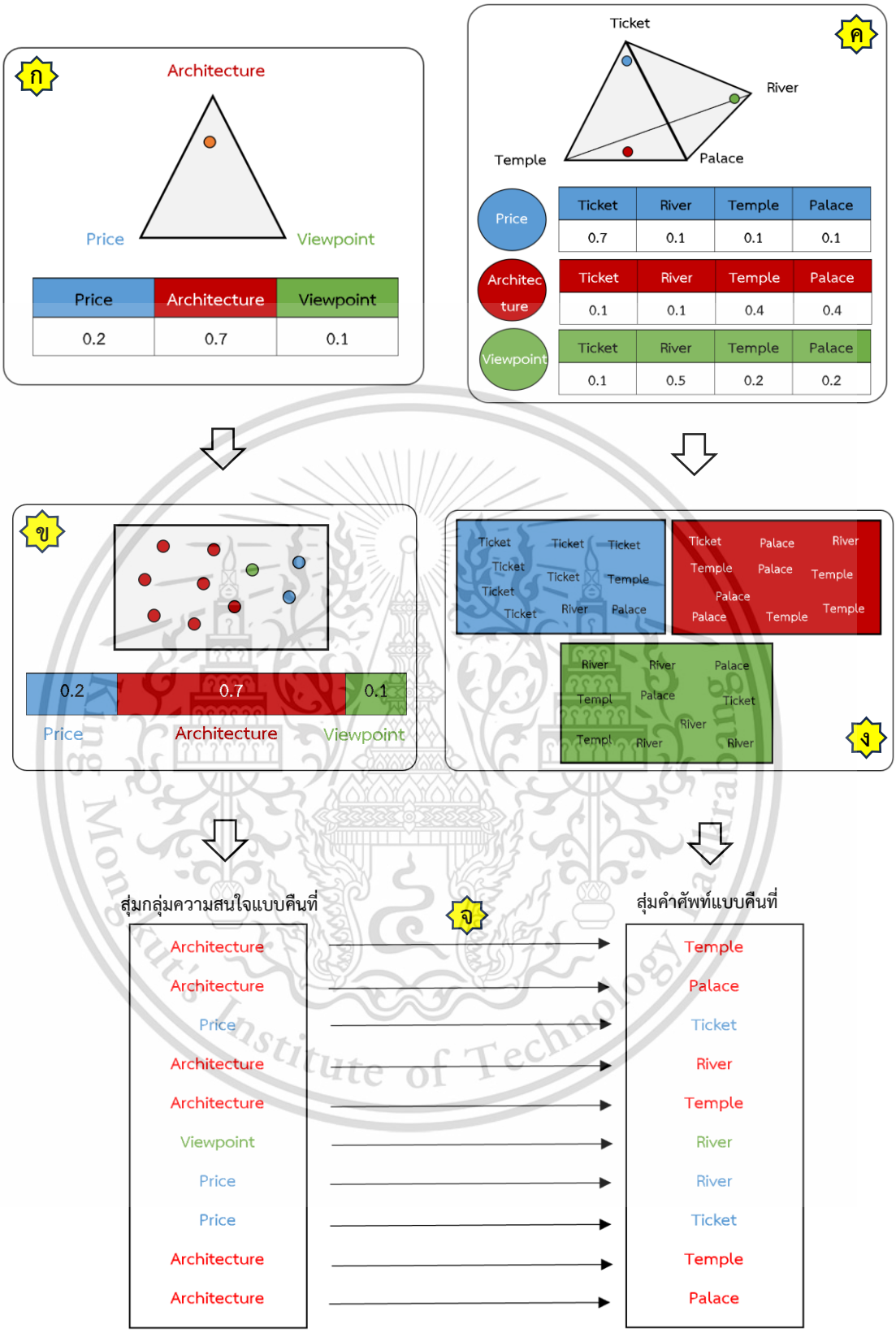
K คือ จำนวนกลุ่มที่เหมาะสม

ตัวอย่างเพื่อความเข้าในตัวโมเดลการจัดสรรหัวข้อแฝง (Latent Dirichlet Allocation: LDA) สมมติให้ k (กลุ่มความสนใจ) = 3 กลุ่มความสนใจ คือ Price, Architecture, Viewpoint บทวิจารณ์ที่ผ่านกระบวนการเตรียมข้อมูลแล้วมีความยาว 10 คำ คือ [Price Ticket Temple Palace River Price Ticket Palace Temple River] มีคำศัพท์ที่ปรากฏในบทวิจารณ์ทั้งสิ้น 4 คำ คือ [Ticket, River, Temple, Palace] จากรูปที่ 2.4 เป็นโครงสร้างการทำงานของโมเดลการจัดสรรหัวข้อแฝง (Serrano, 2020) ประกอบด้วย 3 ส่วน คือ

ส่วนที่ 1 รูปที่ 2.4ก – 2.4ข การกระจายแบบดีรีเคลแดงความสัมพันธ์ระหว่างบทวิจารณ์กับกลุ่มความสนใจ

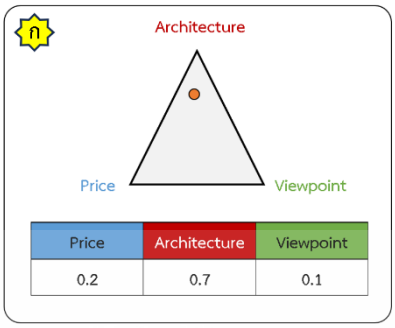
ส่วนที่ 2 รูปที่ 2.4ค – 2.4ง การกระจายแบบดีรีเคลแสดงความสัมพันธ์ระหว่างบทวิจารณ์กับคำศัพท์

ส่วนที่ 3 รูปที่ 2.4จ แสดงคำศัพท์ที่ทำการสุ่มขึ้นมาจากแต่ละกลุ่มความสนใจ



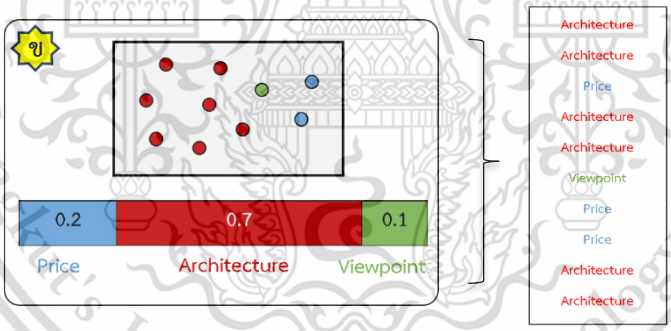
รูปที่ 2.4 ตัวอย่างโครงสร้างการทำงานของ LDA

ส่วนที่ 1 รูปที่ 2.4ก – 2.4ข แสดงการกระจายแบบดีรีเคลตามความสัมพันธ์กับบทวิจารณ์กับกลุ่มความสนใจ



รูปที่ 2.4ก การกระจายแบบดีรีเคลของกลุ่มความสนใจ

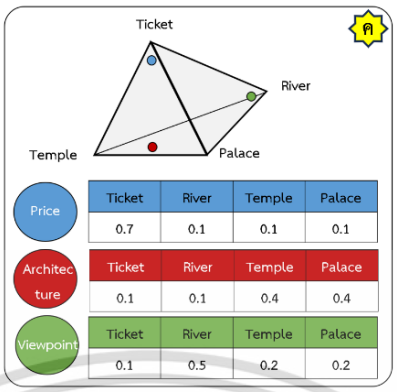
จากรูปที่ 2.4ก จุดสีส้มแสดงตัวแทนของบทวิจารณ์ที่กระจายในรูปสามเหลี่ยมโดยที่จุดเข้าไปในตำแหน่งที่มีคำที่ตรงตามกลุ่มความสนใจที่กำหนดไว้มากที่สุด ในที่นี้คือ 3 กลุ่มความสนใจ เป็นการกระจายแบบดีรีเคลสามมิติ โดยที่จุดสีส้มแทนที่แทนบทวิจารณ์นั้นจะกระจายเข้าไปตามมุมต่างๆ ของสามเหลี่ยม ยิ่งเข้าใกล้มุมใดก็แสดงว่ามีความน่าจะเป็นที่จะอยู่ในกลุ่มความสนใจนั้นๆ จากรูปที่ 2.4ก จะเห็นว่าจุดสีส้มเข้าใกล้มุมสามเหลี่ยมบน คือ กลุ่มความสนใจ Architecture มากสุด โดยมีโอกาสเป็นกลุ่มความสนใจ Price, Architecture, Viewpoint เท่ากับ 20%, 70% และ 10% ตามลำดับ



รูปที่ 2.4ข กลุ่มความสนใจที่ถูกสุ่มขึ้นมา

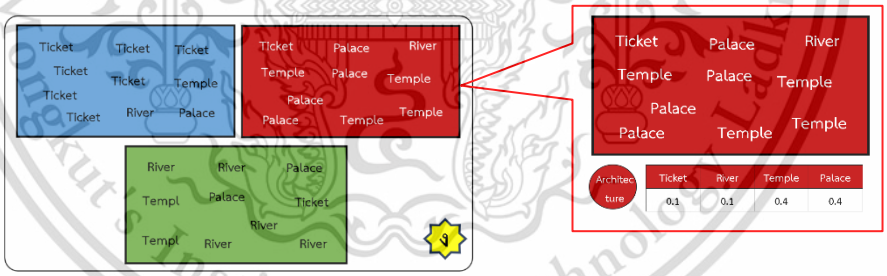
จากรูปที่ 2.4ข แสดงกล่องที่มีจุดสีต่างๆ โดยให้มองจุดสีเสมือนเป็นลูกบอลที่มีข้อความติดอยู่ว่าเป็นกลุ่มความสนใจใด เมื่อพิจารณาตามสีฟ้า แดง และเขียวจะเห็นว่าคำว่า Price เป็นลูกบอลสีฟ้ามี 2 ลูก แทนความน่าจะเป็น 20% ลูกบอลสีแดงแทนกลุ่มความสนใจ Architecture มี 7 ลูก แทนความน่าจะเป็น 70% ลูกบอลสีเขียวแทนกลุ่มความสนใจ Viewpoint มี 1 ลูก แทนความน่าจะเป็น 10% จากนั้นจะทำการสุ่มหยิบคำขึ้นมา จากความยาวของคำ คือ 10 คำ ทำการสุ่มหยิบ 10 ครั้ง (ทุกครั้งที่หยิบออกจะทำการใส่กลับคืนไปทุกครั้งก่อนที่จะทำการสุ่มหยิบใหม่) ผลลัพธ์จากการสุ่มจะเป็นไปตามรูปที่ 2.4จ (ซ้าย) พบว่าเป็นกลุ่มความสนใจ Architecture ทั้งหมด 6 ครั้ง เป็นกลุ่มความสนใจ Price ทั้งหมด 3 ครั้ง และ Viewpoint 1 ครั้ง

ส่วนที่ 2 แสดงดังรูปที่ 2.4ค – 2.4ง เป็นการกระจายแบบดีรีเคลแสดงความสัมพันธ์ระหว่างบทวิจารณ์กับคำศัพท์



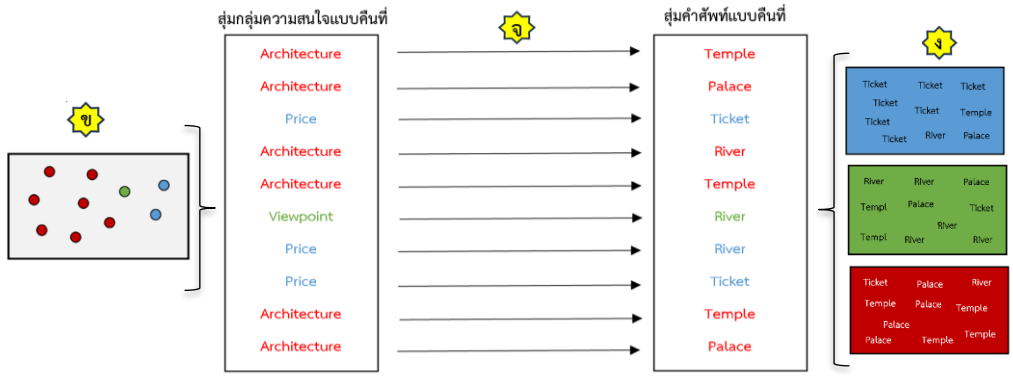
รูปที่ 2.4ค การกระจายแบบดีรีเคลของคำศัพท์

จากรูปที่ 2.4ค จากบทวิจารณ์ตัวอย่างมีคำศัพท์เกิดขึ้นทั้งหมด 4 คำ คือ Ticket, River, Temple, Palace ทำให้การกระจายแบบดีรีเคลเป็นแบบ 4 มิติ แบบพีระมิด ซึ่งมีจุดต่างๆ ภายในแทนกลุ่มความสนใจตามสี คือ จุดสีฟ้าแทนกลุ่มความสนใจ Price จุดสีแดงแทนกลุ่มความสนใจ Architecture จุดสีเขียวแทนกลุ่มความสนใจ Viewpoint การกระจายตัวของจุดต่างๆ แสดงถึงความน่าจะเป็นของคำที่จะเกิดขึ้นในกลุ่มความสนใจใด ตัวอย่างเช่น จุดสีแดงแทนกลุ่มความสนใจ Architecture อยู่กึ่งกลางระหว่างมุมของคำว่า Temple กับ Palace และห่างจากมุมอื่น ทำให้ความน่าจะเป็นของคำว่า Temple กับ Palace ที่จะอยู่ในกลุ่ม Architecture เท่าๆ กัน



รูปที่ 2.4ง ความน่าจะเป็นที่จะเกิดคำศัพท์ในแต่ละกลุ่มความสนใจ

จากรูปที่ 2.4ง จะเห็นว่ามีกล่องทั้งหมด 3 กล่อง เป็นตัวแทนของกลุ่มความสนใจทั้งสามกลุ่ม จำแนกตามสี ภายในกล่องประกอบไปด้วยคำศัพท์ที่มาจากความน่าจะเป็นที่จะเกิดคำศัพท์นั้นขึ้นในกลุ่มความสนใจนั้นๆ จากรูปในกล่องสีแดงเป็นกล่องกลุ่มความสนใจ Architecture จะเห็นว่าความน่าจะเป็นที่คำศัพท์นั้นจะเกิดขึ้นในกลุ่มความสนใจ Architecture ดังนี้ Ticket 10%, River 10%, Temple 40%, Palace 40% ซึ่งเป็นตัวแทนของคำที่อยู่ในกล่อง



รูปที่ 2.4จ คำศัพท์ที่สุ่มขึ้นมาจากแต่ละกลุ่มความสนใจ

ส่วนที่ 3 แสดงดังรูปที่ 2.4จ แสดงตัวอย่างคำศัพท์ที่ทำการสุ่มขึ้นมาจากแต่ละกลุ่มความสนใจ จากกลุ่มความสนใจที่ทำการสุ่มขึ้นมาจากรูปที่ 2.4ข จากการสุ่ม 10 ครั้ง เป็นกลุ่มความสนใจความสนใจ Architecture ทั้งหมด 6 ครั้ง เป็นกลุ่มความสนใจ Price ทั้งหมด 3 ครั้ง และกลุ่มความสนใจ Viewpoint 1 ครั้ง ดังรูปที่ 2.4จ (ซ้าย) จากนั้นแต่ละกลุ่มที่สุ่มได้จากรูปที่ 2.4ข จะสุ่มหยิบคำจากรูปที่ 2.4ง ตามสีกล่องของกลุ่มความสนใจ ดังนั้นถ้ากลุ่มความสนใจที่สุ่มได้อยู่ในกลุ่ม Architecture ก็จะทำการสุ่มคำที่อยู่ในกล่องสีแดงในรูปที่ 2.4ง จากรูปที่ 2.4จ (ขวา) จะเห็นว่ากลุ่มความสนใจ Architecture เมื่อสุ่มหยิบคำในกล่องสีแดงได้คำว่า Temple ทำการสุ่มคำศัพท์ตามกลุ่มความสนใจไปจนครบ 10 ครั้ง ก็จะได้คำศัพท์ทั้งหมด 10 คำ สำหรับบทวิจารณ์ต่อไปก็ทำตามขั้นตอนดังรูปที่ 2.4 ไปเรื่อยๆ จนครับทุกบทวิจารณ์ และสามารถหาค่าน้ำหนักของแต่ละคำศัพท์ได้โดยการนำคำศัพท์ที่เกิดขึ้นมาหารกับจำนวนคำศัพท์นั้นที่เกิดขึ้นทั้งหมด ได้เป็นค่าน้ำหนักที่เป็นผลลัพธ์ของโมเดล

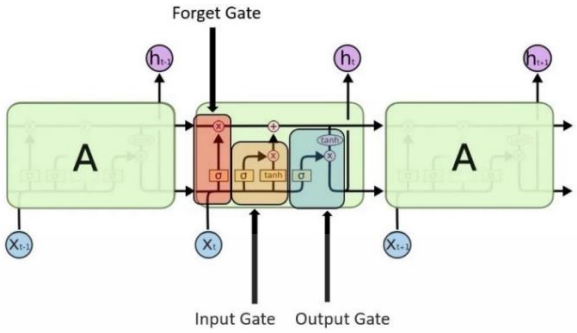
2.3 การเรียนรู้แบบมีผู้สอนเพื่อทำนายกลุ่มตามความรู้สึกของนักท่องเที่ยว

การเรียนรู้แบบมีผู้สอนโดยใช้เทคนิคการเรียนรู้เชิงลึก 2 เทคนิค คือ อัลกอริทึมหน่วยความจำระยะยาว-ระยะสั้นชนิดสองทาง (Bidirectional Long Short-Term Memory: BiLSTM) และอัลกอริทึมโครงข่ายประสาทแบบคอนโวลูชั่น (Convolution Neural Network: CNN) เพื่อทำนายกลุ่มความรู้สึกของนักท่องเที่ยวจากบทวิจารณ์

2.3.1 Bidirectional Long Short-Term Memory (BiLSTM)

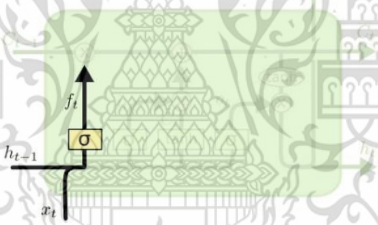
Long Short-Term Memory (LSTM) เป็นแบบจำลองชนิดหนึ่งของอัลกอริทึมนิวรัลเน็ตเวิร์-กแบบวนกลับ Recurrent Neural Network (RNN) ที่พัฒนาขึ้นมาเพื่อแก้ปัญหา Vanishing Gradient หรือ Exploding Problem ในโมเดล RNN กับข้อมูลที่มีความยาวมากๆ โมเดลแบบ LSTM นั้นประกอบไปด้วยประตู (Gate) ที่ควบคุมการไหลเข้าออกของข้อมูลมีทั้งหมดประกอบด้วย Input Gate, Forget Gate, Output Gate และ Cell State ที่จะรวมข้อมูลที่ผ่านมาและส่งผ่านข้อมูลเรียกว่า Long-Term Information Retainer (การเก็บรักษาข้อมูลระยะยาว) มี Hidden State ที่จะส่งต่อข้อมูลจาก Cell ที่ผ่านมา เรียกว่า Short-Term Memory (ความจำระยะสั้น) จากรูปที่ 2.5

แสดงตัวอย่างการทำงานภายใน LSTM Cell และ LSTM Gate และ Single LSTM Memory Cell สามารถคำนวณได้จาก สมการที่ (2.4) – (2.9) (ภูมิรพี ภูมิค้า, 2562; Lawrence, 2023)



รูปที่ 2.5 ตัวอย่างการทำงานภายใน LSTM Cell และ LSTM Gate
(ที่มา Mittal, 2019)

Forget Gate: ควบคุมการไหลของข้อมูลว่าอะไรควรจะลบออกจาก Cell State มีการรับ Hidden State จากในหน่วยเวลาก่อนหน้าและ Input ปัจจุบันและส่งผ่าน Sigmoid Activation Function โดย Output มีค่าอยู่ระหว่าง 0 กับ 1 โดยที่ 0 หมายถึง ลบออก และ 1 หมายถึงเก็บข้อมูลไว้ ดังสมการที่ (2.4)



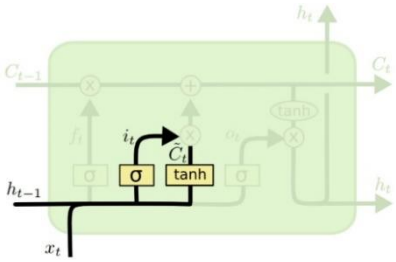
รูปที่ 2.6 การทำงานภายใน LSTM Forget Gate
(ที่มา Zhao, 2023)

$$f_t = \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{2.4}$$

เมื่อ f_t คือ ผลลัพธ์ที่ได้จาก Forget Gate ในช่วงหน่วยเวลา
σ คือ Sigmoid Function
W_{xf} คือ ค่าน้ำหนักสำหรับคำนวณ Input ใน Forget Gate
x_t คือ ค่า Input ที่นำเข้ามาคำนวณ
W_{hf} คือ ค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Forget Gate
h_{t-1} คือ ค่า Hidden State ที่ได้จาการคำนวณในหน่วยเวลาก่อนหน้า
b_f คือ ค่า Bias ที่ใช้คำนวณใน Forget Gate

Input Gate: พิจารณา Input ปัจจุบันและ Hidden State ในหน่วยเวลาก่อนหน้า โดย Input Gate ใช้สำหรับการอัพเดทค่า Cell State มี 2 ส่วนคือ ส่วนที่เป็น Sigmoid Activation Function มีหน้าที่ตัดสินว่าต้องการข้อมูลแค่ไหน ดังสมการที่ (2.5) ส่วนที่สอง เป็น Tanh

Activation Function ให้ค่าน้ำหนักระดับความสำคัญช่วง -1 ถึง 1 เพื่อให้ได้ข้อมูลที่เกี่ยวข้องที่จำเป็นสำหรับ Output ของ Tanh ดังสมการที่ (2.6)



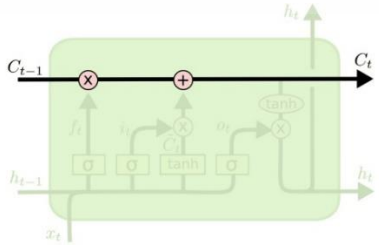
รูปที่ 2.7 การทำงานภายใน LSTM Input Gate
(ที่มา Zhao, 2023)

$$i_t = \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \tag{2.5}$$

$$\tilde{C}_t = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \tag{2.6}$$

เมื่อ i_t คือ ผลลัพธ์ที่ได้จาก Input Gate ในช่วงหน่วยเวลา
σ คือ Sigmoid Function
$\tilde{C}_t$ คือ ผลลัพธ์ที่ได้จากอัพเดทค่า Tanh Function ในช่วงหน่วยเวลา
$\tanh$ คือ Hyperbolic Tangent Function
W_{xi} คือ ค่าน้ำหนักสำหรับคำนวณ Input ใน Input Gate
W_{xc} คือ ค่าน้ำหนักสำหรับคำนวณ Input จาก Cell State Gate
x_t คือ ค่า Input ที่นำเข้ามาคำนวณ
W_{hi} คือ ค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Input Gate
W_{hc} คือ ค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Cell State Gate
h_{t-1} คือ ค่า Hidden State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า
b_i คือ ค่า Bias ที่ใช้คำนวณใน Input Gate
b_c คือ ค่า Bias ที่ใช้คำนวณใน Cell State Gate

Cell State: จาก Forget Gate และ Input Gate ทำการอัพเดท Cell State โดยตัว Cell State ก่อนหน้าจะถูกคูณด้วย Output ของ Forget Gate และถูกรวมเข้ากับ Input Gate เพื่อใช้คำนวณใช้ครั้งถัดไป ดังสมการที่ (2.7)



รูปที่ 2.8 การทำงานภายใน LSTM Cell State
(ที่มา Zhao, 2023)

$$C_t = f_t c_{t-1} + i_t \tanh\left(W_{xc} x_t + W_{hc} h_{t-1} + b_c \right) \quad (2.7)$$

เมื่อ C_t คือ ผลลัพธ์ที่ได้จาก Cell State ในช่วงหน่วยเวลา

f_t คือ ผลลัพธ์ที่ได้จาก Forget Gate

c_{t-1} คือ ค่า Cell State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า

i_t คือ ผลลัพธ์ที่ได้จาก Input Gate

$\tanh$ คือ Hyperbolic Tangent Function

W_{xc} คือ ค่าน้ำหนักสำหรับคำนวณ Input จาก Cell State Gate

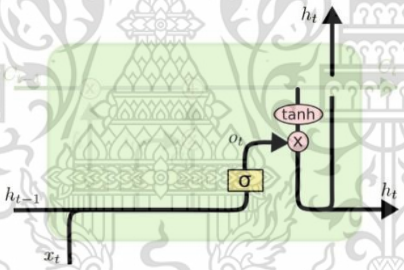
x_t คือ ค่า Input ที่นำเข้ามาคำนวณ

W_{hc} คือ ค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Cell State Gate

h_{t-1} คือ ค่า Hidden State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า

b_c คือ ค่า Bias ที่ใช้คำนวณใน Cell State Gate

Output Gate: การคำนวณ Output ของ Cell มี Output จาก Sigmoid Activation Function จากนั้นจะอัพเดท Cell State ผ่าน Tanh Function ได้ออกมาเป็น Hidden State แสดงดังรูปที่ 2.9



รูปที่ 2.9 การทำงานภายใน LSTM Output Gate
(ที่มา Zhao, 2023)

$$o_t = \sigma\left(W_{xo} x_t + W_{ho} h_{t-1} + b_o \right) \quad (2.8)$$

$$h_t = o_t \tanh\left(C_t \right) \quad (2.9)$$

เมื่อ o_t คือ ผลลัพธ์ที่ได้จาก Output Gate

σ คือ Sigmoid Function

W_{xo} คือ ค่าน้ำหนักสำหรับคำนวณ Input ใน Output Gate

x_t คือ ค่า Input ที่นำเข้ามาคำนวณ

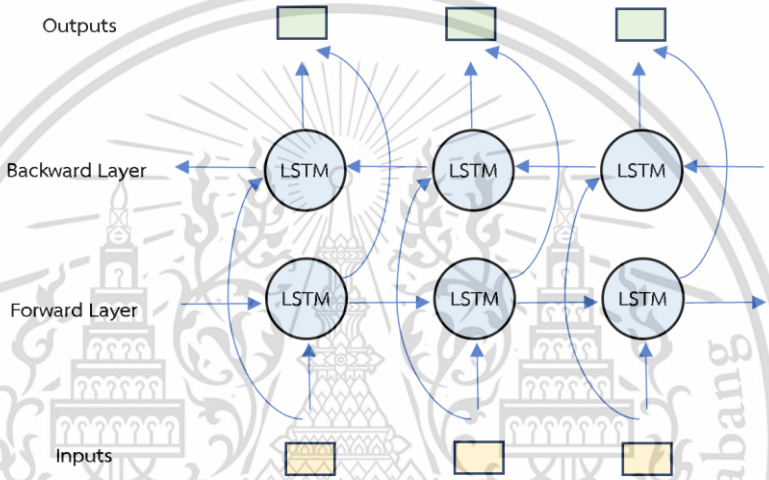
W_{ho} คือ ค่าน้ำหนักสำหรับคำนวณ Hidden State ใน Output Gate

h_{t-1} คือ ค่า Hidden State ที่ได้จากการคำนวณในหน่วยเวลาก่อนหน้า

b_o คือ ค่า bias ที่ใช้คำนวณใน Output Gate

h_t คือ ค่า Hidden State จากการคำนวณ

โดยที่ BiLSTM มี LSTM แบบคู่ 2 ทิศทาง ประกอบด้วย LSTM Cell แบบไปข้างหน้า (Forward LSTM) และ LSTM Cell แบบย้อนกลับ (Backward LSTM) ดังรูปที่ 2.10 จากเดิมที่ LSTM จะพิจารณาคำศัพท์ ก่อนหน้าทั้งหมดตามลำดับเพื่อจะพิจารณาความรู้สึกจากข้อความ ขณะที่ BiLSTM สามารถพิจารณาย้อนกลับลำดับจากหลังไปหน้าด้วยทำให้เข้าใจบริบทของข้อความได้ดีขึ้น เช่น “I do not like this movie” LSTM จะเริ่มจาก “I” ไป “movie” ส่งผ่านสถานะจากคำแรกไปจนถึงคำสุดท้าย BiLSTM จะเริ่มจาก “I” ไป “movie” และจาก “movie” ไป “I” พร้อมกัน ทำให้เข้าใจบริบทของ “not like” ได้ดีกว่า LSTM เพียงอย่างเดียว มีประโยชน์มากในงานด้าน NLP จากการทำนายคำในประโยคจากการประมวลผลข้อความแบบย้อนกลับ วิเคราะห์ความรู้สึก หรือข้อมูลที่มีลักษณะเป็นลำดับ



รูปที่ 2.10 ตัวอย่างโครงสร้างของแบบจำลอง Bidirectional Long Short-Term Memory (BiLSTM)

2.3.2 Convolution Neural Network (CNN)

CNN (Convolution Neural Network) เป็นหนึ่งในการเรียนรู้ของเครื่อง ที่มักถูกนำไปใช้กับงานที่มีลักษณะข้อมูลแบบ 2 มิติ เช่น งานด้านภาพ มีการนำ CNN มาใช้ประโยชน์กับงานด้านการทำนายความรู้สึกของนักท่องเที่ยว เพราะงานด้านรูปภาพและด้าน NLP โดย Input ของข้อความในรูปของเมทริกซ์ โดยที่แต่ละแถวของคำ อยู่ในรูปของเวกเตอร์ที่ได้มาจากการทำ Word Embeddings หรือ One-Hot Embedding และด้วย Convolution Neural Networks สามารถแยก Area Feature จากข้อมูลทั้งหมด ด้วย Convolution Operation ส่วนของคำถูกแยกออกจากกัน เป็นคุณลักษณะและถูกพิจารณาเป็นความสัมพันธ์ของกลุ่มคุณลักษณะ

จากรูปที่ 2.11 แสดงโครงสร้างของแบบจำลอง Convolution Neural Network สำหรับจำแนกความรู้สึก โดยที่ Input ของ Network แสดงถึง เมทริกซ์ ของแต่ละแถว เป็นเวกเตอร์ของคำศัพท์ และจำนวนหลักแสดงถึงจำนวน Vector Dimension ของคำศัพท์ Feature Vector ได้มาจากการทำ Filter 3 ขนาด คือ 2, 3, 4 จาก Input Matrix เมื่อได้ Feature Vector แล้ว ทำการ Max-Pooling จากข้อความเพื่อลดขนาดของ Feature สุดท้ายก็สามารถการจำแนกความรู้สึกด้วยชั้นของ Fully Connected Neural Network และ Activating Function

ตัวอย่างเพื่อความเข้าใจจากรูปที่ 2.11 และ 2.12 จาก Zhang and Wallace (2017) และ Kim (2017) จะแบ่งการอธิบายออกเป็น 5 ส่วน ได้แก่

1) Sentence: จากตัวอย่างประโยค “I love travel Bangkok very much!” มีคำทั้งหมด 6 คำ และ 1 เครื่องหมายอัศเจรีย์ ยาวทั้งหมด 7 คำในประโยค โดยให้ **s** คือ ความยาวของประโยค และทำการเลือกมา 5 Vector Dimension ของคำศัพท์ โดยให้ **d** คือ Vector Dimension ของคำศัพท์ ทำให้ได้รูปร่างเมทริกซ์ของประโยค เป็น **s x d** หรือในตัวอย่างนี้คือ 7 × 5

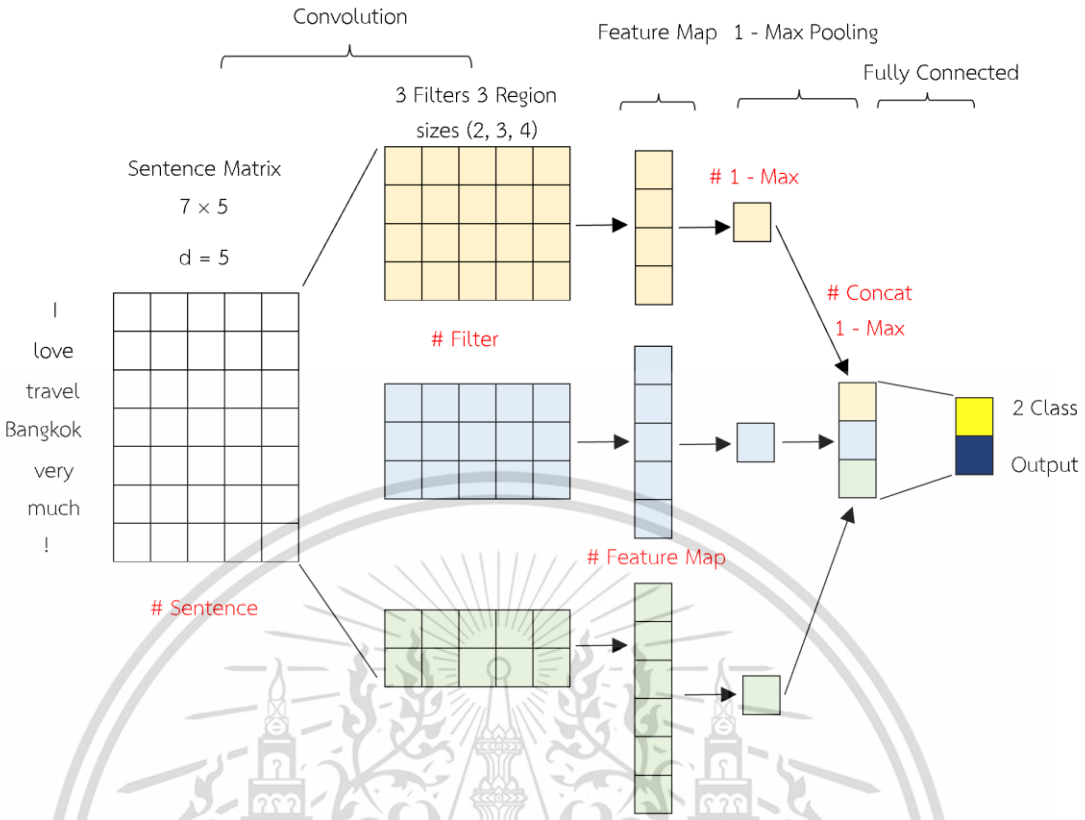
2) Filter: Filter หรือตัวกรอง จะทำการจับลักษณะเฉพาะที่แตกต่างกันของข้อมูลข้อความ โดยจะทำการจับไปตามข้อมูล Input (ข้อความ) ซึ่งจะขึ้นอยู่กับ Region size (Kernel size) หรือขนาดของ Filter ที่จะเป็นตัวบอกว่าจะจับกับข้อมูลข้อความทีละกี่คำ แล้วเลื่อน Filter ลงไปทีละคำ จนครอบคลุมทุกคำในประโยค จากรูปที่ 2.11 จะมี 3 Filter และแต่ละ Filter มีขนาดที่แตกต่างกัน 3 ขนาด คือ 2, 3, 4 จากรูปที่ 2.12 เป็นตัวอย่างการ Filter ทีละ 2 คำ จากรูปจะเริ่มต้นที่คำว่า “I love” แล้วทำการเลื่อนลงไป 1 คำไปจับกับคำว่า “love travel” ต่อเนื่องไปจนถึงคำสุดท้ายของข้อความแล้วสร้าง Feature Map ที่แตกต่างกันในแต่ละ Filter

3) Feature map: จากรูปที่ 2.12 ทำการ Filter แบบ 2 คำ เป็น Matrix (w) ขนาด 2 × 5 แล้วทำการรวม ตัวเลขระหว่างผลคูณเชิงสเกลาร์ระหว่างเมทริกซ์ของประโยคและเมทริกซ์ของ Filter จากตัวอย่างในรูป เมื่อทำการ Filter ทีละ 2 คำ ที่คำว่า “I love” จะได้ Output ของ Feature map ตัวแรก คือ (0.6 x 0.2 + 0.5 x 0.1 + … + 0.1 x 0.1) = 0.51 จากนั้นก็จะทำการเลื่อนตัว Filter ลงมา 1 คำ จาก “I love” เป็น “love travel” หลังจากรวมตัวเลขแล้วจะได้ 0.53 เป็นตัวถัดไป

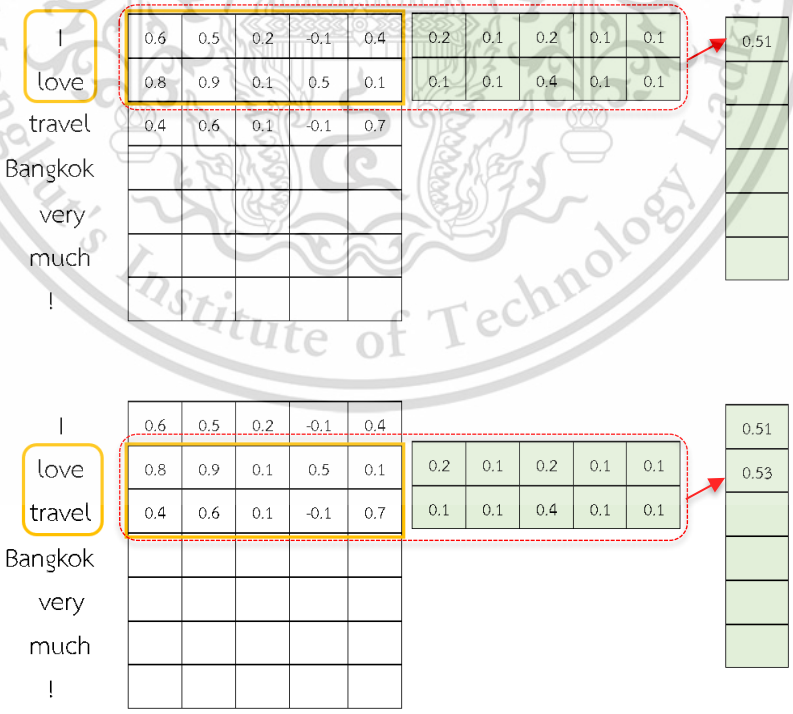
โดยสามารถหาขนาด Output ของ Feature Map ได้จาก (s – h + 1 x 1) โดย h คือ Region Size ที่แสดงถึงแถวของคำที่ถูก Filter และ s คือ ความยาวของประโยค ในกรณีตัวอย่างนี้คือ (7 – 2 + 1 x 1) = 6 x 1

4) 1 - Max: ใช้ 1 - Max Pooling เพื่อเลือกค่าที่มากที่สุดในแต่ละ Feature Map เพื่อสร้าง Feature Vector ที่มีขนาดคงที่

5) Concat 1 - Max: นำ Feature Vector ที่ได้จากทุก Filter มาต่อกันเป็น Feature Vector ที่มีขนาด 3 x 1 แล้วใช้ Fully Connected Layer เพื่อจำแนกประเภทของประโยค



รูปที่ 2.11 ตัวอย่างโครงสร้างของแบบจำลอง Convolution Neural Network ในงานด้าน NLP (ที่มา Zhang and Wallace, 2017)



รูปที่ 2.12 ตัวอย่างการ Filter ทีละ 2 คำศัพท์ (ที่มา Kim, 2017)

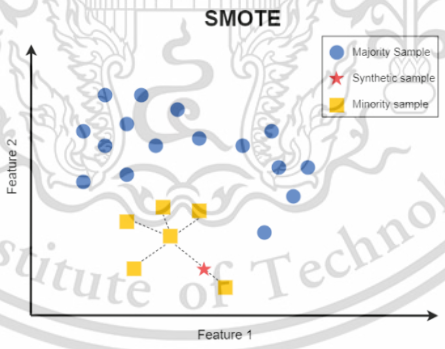
2.3.3 แก้ปัญหาความไม่สมดุลกันของชุดข้อมูล (Imbalance Dataset)

จากบทวิจารณ์ที่ได้จาก TripAdvisor พบว่าเป็นบทวิจารณ์เชิงบวกมากกว่าบทวิจารณ์เชิงลบ ทำให้เมื่อนำข้อมูลไปใช้ในการฝึกสอนโมเดลจะทำให้ประสิทธิภาพในการทำนายลดลง มีแนวโน้มในการทำนายไปทางบทวิจารณ์เชิงบวกมากว่าบทวิจารณ์เชิงลบ ซึ่งหมายความว่าโมเดลที่ฝึกสอนด้วย Imbalanced Data จะทำนายได้ไม่แม่นยำ

ดังนั้นเพื่อป้องกันการโน้มเอียงไปที่คลาสใดคลาสหนึ่ง (Prediction Bias) ซึ่งในงานวิจัยนี้จะใช้การแก้ปัญหาให้ข้อมูลทั้ง 2 กลุ่มมี จำนวนที่ใกล้เคียงกัน โดยได้เลือกวิธีการแก้ปัญหามา 2 วิธี คือ Undersampling (วิธีการสุ่มลด) และ SMOTE (การสุ่มตัวอย่างสังเคราะห์ของข้อมูลกลุ่มน้อย)

1. Undersampling เป็นการสุ่มเพื่อลดข้อมูลของคลาสที่มีจำนวนมากกว่าด้วยการลบข้อมูลของคลาสที่มีจำนวนมากเพื่อให้ใกล้เคียงกับคลาสที่มีจำนวนน้อย ลดขนาดของข้อมูลทำให้ใช้ทรัพยากรในการคำนวณน้อยลง แต่มีข้อจำกัดที่อาจทำให้สูญเสียข้อมูลสำคัญ เมื่อในคลาสข้อมูลจำนวนมากที่ถูกลบออกไป (Kulkarni et al., 2021)

2. SMOTE (Synthetic Minority Over-sampling Technique: SMOTE) เป็นการสุ่มเพิ่มจำนวนข้อมูลของคลาสข้อมูลที่มีจำนวนน้อยขึ้นมาใหม่โดยนำมาพิจารณาทีละตัวจนครบทุกตัว อาศัยหลักการ กำหนดจำนวนด้วยอัลกอริทึมเพื่อนบ้านที่ใกล้เคียงที่สุด (K-Nearest Neighbor: KNN) จำนวน k ตัวแล้วทำการสุ่มสร้างข้อมูลขึ้นมาใหม่บนพื้นที่ใดๆ บนทางที่เชื่อมโยงระหว่างจุดข้อมูลที่กำลังพิจารณาและจุดข้อมูลของเพื่อนบ้านที่ใกล้ที่สุด (วิทยา ปัญญา และ วฤษาย์ ร่มสายหยุด, 2565) แสดงดังรูปที่ 2.13



รูปที่ 2.13 ภาพลักษณะวิธีการสุ่มตัวอย่างสังเคราะห์ของข้อมูลกลุ่มน้อย (SMOTE) (ที่มา วิทยา ปัญญา และ วฤษาย์ ร่มสายหยุด, 2565)

2.3.4 Grid Search

เป็นเทคนิคในการหาค่าที่เหมาะสมที่สุดสำหรับ Hyperparameter ของแบบจำลองการเรียนรู้ของเครื่อง โดยวิธีนี้จะทำการสร้างตารางของค่าพารามิเตอร์หลายๆ ค่า และทำการทดสอบแบบจำลองกับแต่ละชุดค่าพารามิเตอร์นั้นๆ เช่น การทดสอบค่าอัตราการเรียนรู้ (Learning Rate) ในกรอบที่ต้องการทดลอง [0.1, 0.01] และ จำนวนชั้น Hidden Layer ในกรอบที่ต้องการทดลอง [2, 3] จะได้กรอบการทดลองหรือ Search Space ทั้งหมด 2 x 2 = 4 รูปแบบ ดังนี้ [0.1, 2], [0.1, 3],

[0.01, 2] และ [0.01, 3] เพื่อหาค่าที่เหมาะสมที่สุด กระบวนการนี้สามารถปรับใช้ได้กับอัลกอริธึมการเรียนรู้ของเครื่องหลายแบบ

จากการศึกษาของ Manurung and Lhaksmana (2023) มีการใช้ Random Search กับอัลกอริธึม LSTM และ CNN สำหรับอัลกอริธึม LSTM ทำการทดสอบกับพารามิเตอร์ LSTM Block (LSTM Cell) 10 - 240 หน่วย และจำนวนโหนดในชั้น Fully Connected 32 - 128 หน่วย สำหรับ CNN จะทดสอบกับ จำนวน Filter 32 – 64 และขนาดของ Filter (Kernel Size) 3 - 5 โดยงานวิจัยนี้จะประยุกต์ใช้หลักการดังกล่าวมาใช้กับการจูนด้วย Grid Search โดยรายละเอียดการตั้งค่าของแต่ละ Hyperparameter สามารถดูได้ในบทที่ 3 หัวข้อ 3.6.3 และ 3.6.4

จุดสังเกตสำหรับการใช้ปรับจูน Hyperparameter เมื่อทำงานกับข้อมูลขนาดใหญ่ซับซ้อน ทำให้การทำงานของโมเดลอาจมีค่าใช้จ่ายที่สูง และอาจให้ผลลัพธ์ที่ต่ำกว่าปกติได้ หากพารามิเตอร์ที่ตั้งไว้ไม่เหมาะสมกับโมเดล และการคำนวณมีความซับซ้อนตามจำนวนพารามิเตอร์ที่ตั้งไว้ (Shetty et al., 2024)

2.4 การเปรียบเทียบประสิทธิภาพของการทำนาย

2.4.1 เมทริกซ์ความสับสน (Confusion Matrix)

ใช้ในการวัดผลการจำแนกประเภทของข้อมูล (Classification) เพื่อที่จะวัดประสิทธิภาพของการจำแนกความรู้สึกของบทวิจารณ์ที่เป็นเชิงบวกหรือบทวิจารณ์เชิงลบ เพื่อที่จะประเมินว่าแบบจำลองสามารถนำไปใช้ได้จริงหรือไม่ โดยสามารถคำนวณได้จากเมทริกซ์ความสับสนดังตารางที่ 2.1

ตารางที่ 2.1 เมทริกซ์ความสับสน

		Actual (ค่าจริง)	
		Positive	Negative
Prediction (ค่าทำนาย)	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

จากตารางที่ 2.1 จากข้อมูลคลาสที่จะทำนายมี 2 ประเภท คือ ด้านบวก (Positive) และด้านลบ (Negative) ค่าของตัวแปรต่างๆ ใน ตารางสามารถแบ่งออกเป็น 4 กรณีดังนี้

1. True Positive (TP) หมายถึง ผลที่ได้จากการทำนายเป็นเชิงบวกและตรงกับค่าจริงของข้อมูลที่เป็นเชิงบวก

2. False Positive (FP) หมายถึง ผลที่ได้จากการทำนายเป็นเชิงบวกแต่ค่าจริงของข้อมูลเป็นเชิงลบ

3. False Negative (FN) หมายถึง ผลที่ได้จากการทำนายเป็นเชิงลบแต่ค่าจริงของข้อมูลเป็นเชิงบวก

4. True Negative (TN) หมายถึง ผลที่ได้จากการทำนายเป็นเชิงลบและตรงกับค่าจริงของข้อมูลที่เป็นเชิงลบ

ค่าความถูกต้อง (Accuracy)

ค่าทางสถิติที่ใช้เปรียบเทียบความสัมพันธ์ของข้อมูลการทำนายผลระหว่างผลลัพธ์จริงกับผลลัพธ์ที่ได้จากการทำนาย โดยอ้างอิงจากค่าในตารางเมทริกซ์ความสับสน เพื่อหาค่าความถูกต้องได้ดังสมการที่ (2.10)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2.10}$$

ค่าความเที่ยง (Precision)

ค่าที่ใช้สำหรับวัดประสิทธิภาพในการทำนายของแบบจำลอง โดยคำนวณจากค่า True Positive (TP) เทียบกับผลรวมของกลุ่มข้อมูลที่ได้จากการทำนายที่เป็นเป็นเชิงบวกทั้งหมด $(TP + FP)$ ดังสมการที่ (2.11)

$$Precision = \frac{TP}{TP + FP} \tag{2.11}$$

ค่าความไวหรือระลึก (Recall)

ค่าที่ใช้สำหรับวัดประสิทธิภาพในการทำนายของแบบจำลอง โดยคำนวณจากค่า True Positive (TP) เทียบกับผลรวมของผลลัพธ์ที่ทำนายได้เป็นเชิงบวกทั้งหมด $(TP + FN)$ ดังสมการที่ (2.12)

$$Recall = \frac{TP}{TP + FN} \tag{2.12}$$

2.5 การวิเคราะห์ความเด่นและความแพร่หลาย (Salience Valence Analysis)

Salience Valence Analysis หรือ การวิเคราะห์ความเด่น (Salience) และความแพร่หลาย (Valence) เป็นแนวคิดในการวิเคราะห์หาจุดเด่นและจุดด้อยของกลุ่มความสนใจ จากงานวิจัยของ Taecharungroj and Mathayomchan (2019) ได้มีการพัฒนาเครื่องมือ 2 ตัวที่ช่วยในการตีความความสนใจของนักท่องเที่ยวจากบทวิจารณ์ ที่ได้จาก LDA โดยที่ตัวแรกคือ ความสนใจที่นักท่องเที่ยวมีให้ในแต่ละกลุ่มความสนใจและมีความรู้สึกเชิงบวกมากแค่ไหนกับแต่ละกลุ่มความสนใจในชื่อ การวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม (Dimensional Salience Valence Analysis: DSVA) และตัวที่สองคือ ความสนใจของนักท่องเที่ยวต่อคำศัพท์จากบทวิจารณ์ ว่าส่งผลกระทบต่อความรู้สึกเชิงบวกหรือลบในชื่อ การวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ (Lexical Salience Valence Analysis: LSVA) เพื่อช่วยในการปรับปรุงจุดด้อย และเสริมในจุดเด่นการท่องเที่ยวในสถานที่นั้นๆ ให้ดียิ่งขึ้น

2.5.1 การวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม (Dimensional Salience Valence Analysis: DSVA)

ในการหาความสนใจที่นักท่องเที่ยวมีให้ในแต่ละกลุ่มความสนใจ เรียกว่า ความเด่นเชิงกลุ่ม (Dimensional Salience) หาได้จากจำนวนบทวิจารณ์เชิงบวกทั้งหมดในกลุ่มความสนใจหารด้วยจำนวนบทวิจารณ์ทั้งหมด คำนวณได้จากสมการที่ (2.13) และความรู้สึกเชิงบวกที่นักท่องเที่ยวมีให้มากแค่ไหนกับแต่ละกลุ่มความสนใจ เรียกว่า ความแพร่หลายเชิงกลุ่ม (Dimensional Valence) คำนวณได้จากสมการที่ (2.14) แล้วนำมาสร้างแผนภูมิแท่งระหว่างค่าความเด่นเชิงกลุ่ม (Dimensional Salience) และค่าความแพร่หลายเชิงกลุ่ม (Dimensional Valence) ของกลุ่มความสนใจ ว่ากลุ่มความสนใจไหนที่นักท่องเที่ยวสนใจเป็นพิเศษ และความสนใจนั้นไปในมุมมองเชิงบวกหรือลบ

$$Dimensional\ salience = \left(\frac{r_{POS}}{R}\right) \times 100 \tag{2.13}$$

$$Dimensional\ valence = \left(\frac{r_{POS} - e_{POS}}{r}\right) \times 100 \tag{2.14}$$

$$e_{POS} = \frac{r \times R_{POS}}{R} \tag{2.15}$$

เมื่อ r_{POS} คือ จำนวนบทวิจารณ์เชิงบวกในกลุ่มความสนใจ
r คือ จำนวนบทวิจารณ์ทั้งหมดในกลุ่มความสนใจ
e_{POS} คือ จำนวนบทวิจารณ์เชิงบวกที่คาดหวังในกลุ่มความสนใจ
R คือ จำนวนบทวิจารณ์ทั้งหมด
R_{POS} คือ จำนวนบทวิจารณ์เชิงบวกทั้งหมด

2.5.2 การวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ (Lexical Salience Valence Analysis: LSVA)

หลังจากได้ผลลัพธ์จากโมเดลการจัดสรรหัวข้อแฝง จะทำการทำการแปลงให้อยู่ในรูปของความเด่นเชิงคำศัพท์ (Term Salience) และความแพร่หลายเชิงคำศัพท์ (Term Valence) โดยที่ความเด่นเชิงคำศัพท์ (Term Salience) หมายถึง ความถี่ของคำศัพท์เฉพาะที่ปรากฏในบทวิจารณ์ ยิ่งมีค่าสูงแสดงว่านักท่องเที่ยวนิยมใช้คำศัพท์เฉพาะนั้นๆ สูงในการเขียนบทวิจารณ์ สามารถคำนวณได้ดังสมการที่ (2.16) ส่วนความแพร่หลายเชิงคำศัพท์ (Term Valence) หมายถึง ความรู้สึกเชิงบวกหรือลบของนักท่องเที่ยวที่มีต่อคำศัพท์เฉพาะ สามารถคำนวณได้ดังสมการที่ (2.17) ซึ่งเมื่อนำมาสร้างแผนภูมิฟองเปรียบเทียบระหว่างความเด่นเชิงคำศัพท์ (Term Salience) และความแพร่หลายเชิงคำศัพท์ (Term Valence) จะทำให้เห็นว่านักท่องเที่ยวให้ความสนใจกับคำไหนเป็นพิเศษ และมีความรู้อย่างไรต่อสิ่งนั้น

$$Term\ salience = \log_{10}(t) \tag{2.16}$$

$$Term\ valence = \frac{\bar{x}_{POS} - \bar{x}_{POS}}{\bar{x}_{POS} + \bar{x}_{POS}} \tag{2.17}$$

เมื่อ $\bar{x}_{POS}$ คือ จำนวนเฉลี่ยของคำศัพท์เฉพาะที่ปรากฏในบทวิจารณ์เชิงบวก
$\bar{x}_{NEG}$ คือ จำนวนเฉลี่ยของคำศัพท์เฉพาะที่ปรากฏในบทวิจารณ์เชิงลบ
t คือ จำนวนคำศัพท์เฉพาะที่เกิดขึ้น

ความเด่นเชิงคำศัพท์ (Term Salience): ความเด่นเชิงคำศัพท์ที่สูง หมายถึง ความถี่ของคำศัพท์เฉพาะที่ปรากฏในบทวิจารณ์ที่สูง ยิ่งมีค่าสูงแสดงว่านักท่องเที่ยวนิยมใช้คำศัพท์เฉพาะนั้นๆ สูงในการเขียนบทวิจารณ์

ความแพร่หลายเชิงคำศัพท์ (Term Valence): บอกถึงความรู้สึกเชิงบวกหรือลบของนักท่องเที่ยวที่มีต่อคำศัพท์เฉพาะ

สามารถตีความออกมาได้ 4 รูปแบบ ดังรูปที่ 2.14 ดังนี้

ความเด่น (Salience)	ความแพร่หลาย (Valence)
ความเด่น (Salience) เป็นบวกและ ความแพร่หลาย (Valence) เป็นลบ หมายถึง นักท่องเที่ยวให้ความสนใจ แต่มีความรู้สึกเชิงลบ	ความเด่น (Salience) เป็นบวกและ ความแพร่หลาย (Valence) เป็นบวก หมายถึง นักท่องเที่ยวให้ความสนใจ และมีความรู้สึกเชิงบวก
ความเด่น (Salience) เป็นลบและ ความแพร่หลาย (Valence) เป็นลบ หมายถึง นักท่องเที่ยวให้ความสนใจ น้อยและมีความรู้สึกเชิงลบ	ความเด่น (Salience) เป็นลบและ ความแพร่หลาย (Valence) เป็นบวก หมายถึง นักท่องเที่ยวให้ความสนใจ น้อยแต่มีความรู้สึกเชิงบวก

รูปที่ 2.14 การตีความความเด่นและความแพร่หลายเชิงคำศัพท์

2.6 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้องกับเทคนิคต่างๆ ที่เกี่ยวข้องกับการวิเคราะห์บทวิจารณ์ที่ได้จากเว็บไซต์ TripAdvisor

Taecharungroj and Mathayomchan (2019) ได้พัฒนาวิธีการวิเคราะห์บทวิจารณ์ภาษาอังกฤษจากเว็บไซต์ TripAdvisor ในหมวด Tourist Attraction ในจังหวัดภูเก็ตประกอบด้วยสถานที่ท่องเที่ยว 5 สถานที่ คือ ชายหาด 20 หาด จำนวน 25,458 บทวิจารณ์ เกาะต่างๆ 12 เกาะ จำนวน 12,584 บทวิจารณ์ ตลาดต่างๆ 12 ตลาด จำนวน 3,514 บทวิจารณ์ วัด 2 แห่ง จำนวน 10,519 บทวิจารณ์ และถนนคนเดิน 1 แห่ง จำนวน 13,004 บทวิจารณ์ รวมทั้งสิ้น 25,458 บทวิจารณ์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง 2 โมเดล คือ โมเดล Latent Dirichlet allocation (LDA) และโมเดล Naïve Bayes ในการหากลุ่มความสนใจจากบทวิจารณ์จะใช้โมเดล LDA เพื่อหากลุ่มความสนใจนักท่องเที่ยวในแต่ละสถานที่ โดยมีการใช้ K-Mean เพื่อจำนวนกลุ่มความสนใจในบทวิจารณ์ที่ 1-10 กลุ่มแล้วใช้ Elbow Method เข้ามาช่วยในการหากลุ่มความสนใจที่เหมาะสมในแต่ละสถานที่ หลังจากที่ได้ผลลัพธ์จาก LDA โมเดล Naïve Bayes วิเคราะห์ความรู้สึก เพื่อนำไปสร้างแผนภูมิฟองสบู่ เพื่อวิเคราะห์หาข้อมูลเชิงลึก ในจังหวัดภูเก็ต โดยงานวิจัยนี้ได้ค่าความถูกต้องจากผลลัพธ์ที่ดีที่สุดที่ 78% สำหรับงานด้านการท่องเที่ยวควรมีค่าความถูกต้องมากกว่า 70%

Ramadhani et al. (2021) เก็บข้อมูลจาก TripAdvisor 5 ชายหาดในเกาะบาหลี จำนวน 10,708 บทวิจารณ์ ประกอบด้วย 1. หาด Double Six 2. หาด Sminyak 3. หาด Nusa Dua 4. หาด Delinkinhg และ 5. หาด Cangue ทำการวิเคราะห์ Sentiment มีการใช้ระบุ Label โดยของบทวิจารณ์โดยใช้ VADER Sentiment Analysis Tool เพื่อระบุความรู้สึกของบทวิจารณ์ว่าเป็นเชิงบวกหรือเชิงลบด้วยคะแนนของรูปประโยค เพื่อเข้าโมเดลการเรียนรู้เชิงลึกแบบ LSTM เพื่อหาประสิทธิภาพในการจำแนกความรู้สึกจากบทวิจารณ์ โดยพบว่า ค่า Accuracy ของแต่ละชายหาดเกาะกลุ่มอยู่ที่ประมาณ 81 - 84%

Puh and Babac (2022) ได้ทำการเก็บข้อมูลลูกค้าที่เคยเข้าพักในโรงแรมแล้วได้เขียนบทวิจารณ์ลงในเว็บไซต์ TripAdvisor จำนวน 20,491 บทวิจารณ์ เพื่อพัฒนาและเปรียบเทียบโมเดลการเรียนรู้ของเครื่องที่แตกต่างกัน สำหรับการจำแนกบทวิจารณ์ของลูกค้าที่เคยเข้าพักว่ามีความรู้สึกเชิงบวก เชิงลบ หรือรู้สึกเฉยๆ พร้อมทั้งทำการทำนาย คะแนนของดาวที่ลูกค้าจะให้ตั้งแต่ 1 - 5 ดาว โดยโมเดลที่นำมาใช้แบ่งเป็น 2 ส่วน คือ ส่วนที่เป็นโมเดลการเรียนรู้ของเครื่อง ประกอบด้วย Naïve Bays และ SVM ส่วนที่สองจะเป็นโมเดลการการเรียนรู้เชิงลึก ประกอบด้วย CNN, LSTM และ BiLSTM โดยผลที่คือ โมเดลแบบการเรียนรู้เชิงลึกให้ประสิทธิภาพที่ดีกว่าโมเดลการเรียนรู้ของเครื่อง เพื่อพิจารณาจุดหลักในเรื่องของการการทำนายความรู้สึก ถึงอย่างไรก็ตามแม้โมเดลการเรียนรู้ของเครื่องก็ยังให้ค่า Accuracy ที่ค่อนข้างดีแม้จะน้อยกว่าโมเดลการเรียนรู้เชิงลึก เช่น SVM ได้ Accuracy ที่ 0.8 เมื่อเทียบกับ BiLSTM ที่ 0.89 และเมื่อมาพิจารณาจุดหลักในเรื่องการทำนายคะแนนของดาว ที่มีความซับซ้อนกว่าการทำนายความรู้สึก นั้นจะพบว่าโมเดลการเรียนรู้เชิงลึกให้ค่า Accuracy ที่ดีกว่าโมเดลการเรียนรู้ของเครื่องอย่างชัดเจน เช่น Naïve Bays ให้ Accuracy ที่ 46% เมื่อเทียบกับ BiLSTM ที่ 76%

Chugh and Phumchusri (2020) ทำการวิเคราะห์หาข้อมูลเชิงลึกจากบทวิจารณ์ในเว็บไซต์ TripAdvisor ประเภท Bangkok Tours/Activities ในกรุงเทพมหานคร ทั้งหมด 68,000 บทวิจารณ์ ประกอบด้วย กิจกรรม 3,827 บทวิจารณ์ ขี่จักรยานท่องเที่ยว 8,004 บทวิจารณ์ การเรียนทำอาหาร 4,899 บทวิจารณ์ ทัวร์รับประทานอาหาร 5,424 บทวิจารณ์ การเที่ยวชมสถานที่ต่างๆ 15,862 บทวิจารณ์ และ สปา 21,839 บทวิจารณ์ โดยวัตถุประสงค์หลักของงานวิจัยมี 3 ข้อ คือ 1. หาข้อมูลความชื่นชอบและแนวโน้มความชื่นชอบของนักท่องเที่ยว 2. พัฒนาโมเดลทำนายผลที่สามารถคาดการณ์จำนวนคะแนนเรตติง 5 ดาว ของบทวิจารณ์ เพื่อที่จะระบุตัวแปรสำคัญที่มีผลกระทบกับความรู้สึกเชิงบวกของนักท่องเที่ยวได้ 3. พัฒนาโมเดลทำนายผลที่สามารถคาดการณ์จำนวนคะแนนเรตติง 1 ดาว ของบทวิจารณ์ เพื่อที่จะระบุตัวแปรสำคัญที่มีผลกระทบกับ ความรู้สึกเชิงลบของนักท่องเที่ยวได้ ในส่วนของเทคนิคที่ใช้ มีการใช้ Sentiment Analysis with R Studio นับจำนวนความถี่ของคำที่เป็นบวกหรือคำที่เป็นลบในประโยค เพื่อระบุคะแนนความรู้สึกในแต่ละบทวิจารณ์ มีการใช้ NLP ทำการนับความถี่ของคำเพื่อหาว่าคำอะไรที่เป็นที่กล่าถึงมากที่สุดในทั้ง 6 Tours/Actives ทั้งด้านบวกและด้านลบ ในประเภท ขี่จักรยานท่องเที่ยวพบว่า คำว่า “guide” มีการกล่าวถึงมากสุดทั้งในด้านบวกและด้านลบ และใช้โมเดล Logistic Regression เพื่อใช้ในการหาตัวแปรสำคัญที่ใช้ในการทำนายหาคะแนนเรตติง 5 ดาว และเรตติง 1 ดาว พบว่า “sentiment” เป็นตัวแปรสำคัญที่ส่งผลในการทำนายที่จะเป็นเรตติง 5 ดาว หรือ 1 ดาว

Alharbi et al. (2022) ทำการเก็บข้อมูลบทวิจารณ์จากสถานที่ท่องเที่ยวจาก Google Map ที่เป็นสถานที่ประเภทเดียวกันทั้ง 14 เมืองในประเทศซาอุดิอาราเบีย ประกอบด้วย ประเภทพิพิธภัณฑ์ แหล่งโบราณคดี สถานที่สำคัญทางศาสนาอิสลาม ศูนย์นิทรรศการ สวนสาธารณะ และสถานที่พักผ่อน เพื่อที่จะสร้างโมเดลที่สามารถจำแนกความรู้สึกในภาษาซาอุดิอาราเบียได้อย่างแม่นยำ โดยใช้ 3 เทคนิค คือ SVM, LSTM และ RNN มีการใช้ Lexicon-Base Method เพื่อใช้จำแนกบทวิจารณ์ว่าเป็นเชิงบวก เชิงลบ หรือกลางๆ โดยถ้ามีคำที่เป็นบวกมากกว่าลบจะจำแนกว่าเป็นบวก เช่นเดียวกันถ้ามีคำที่เป็นลบมากกว่าคำที่เป็นบวกจะจำแนกว่าเป็นลบ และถ้าคำที่เป็นบวกเท่ากับคำที่เป็นลบจะจำแนกเป็นกลาง โดยผลลัพธ์ที่ได้จากทั้ง 3 โมเดล พบว่า SVM ให้ค่า Accuracy ที่สูงถึง 98% ซึ่งมากว่าทั้ง RNN และ LSTM

นิธิกร เลิศชาญวุฒิ (2564) วิเคราะห์ความสนใจของนักท่องเที่ยวที่มีต่อถนนเยาวราช ประเทศไทย จากการเก็บข้อมูลจากเว็บไซต์ TripAdvisor จำนวน 3,992 บทวิจารณ์ โดยใช้เทคนิคการเรียนรู้ของเครื่อง แบ่งการวิเคราะห์ออกเป็น 2 ส่วน คือ การจัดสรรหัวข้อแฝง (LDA) โดยใช้ K-Means ร่วมในการหากลุ่มความสนใจที่เหมาะสมและการวิเคราะห์ความรู้สึกของนักท่องเที่ยวด้วยการจำแนกความรู้สึกออกเป็นเชิงบวกและเชิงลบ โดยใช้โมเดลทั้งหมด 3 โมเดล คือ นาอีฟเบร์ (Naïve Bays), การถดถอยเชิงโลจิสติค (Logistic Regression), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) โดยจะใช้ทั้ง 3 โมเดลในการตัดสินใจร่วมกัน โดยผลลัพธ์ที่ได้จากการแบ่งกลุ่มความสนใจได้เป็น 4 กลุ่มความสนใจ คือ แหล่งช้อปปิ้ง ตลาดอาหารริมทางยามค่ำ อาหาร การท่องเที่ยวชมเมือง ในส่วนของการจำแนกความรู้สึกเมื่อร่วมกันตัดสินใจแล้วให้ค่าความถูกต้องที่ 88.62%

งานวิจัยที่เกี่ยวข้องสามารถสรุปได้ดังเทคนิคที่ใช้ดังตารางที่ 2.2

ก แทน Taecharungroj and Mathayomchan (2019)

ข แทน Ramadhani et al. (2021)

ค แทน Puh and Babac (2022)

ง แทน Chugh and Phumchusri (2020)

จ แทน Alharbi et al. (2022)

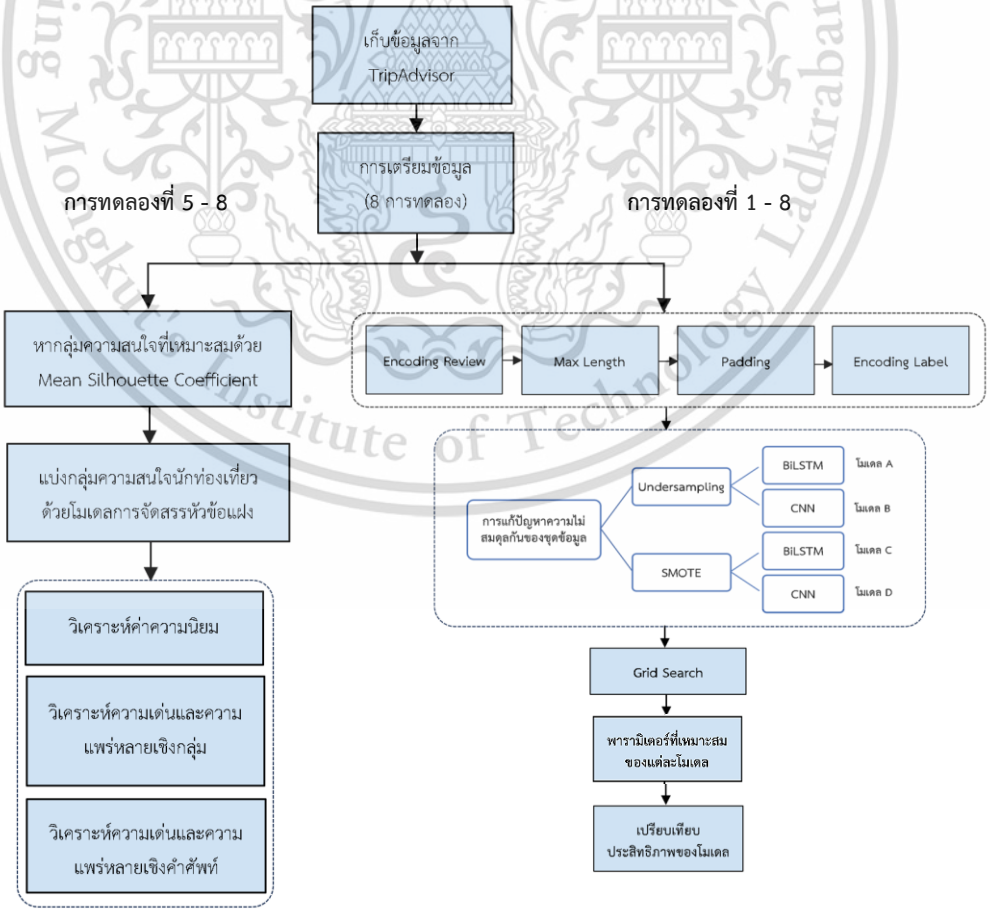
ฉ แทน นิธิกร เลิศชาญวุฒิ (2564)

ตารางที่ 2.2 สรุปงานวิจัยที่เกี่ยวข้อง

เทคนิคที่ใช้	**งานวิจัยที่เกี่ยวข้อง**					
	ก	**ข**	**ค**	**ง**	**จ**	**ฉ**
วิเคราะห์การแบ่งกลุ่ม						
K-Means	✓	-	-	-	-	✓
หาหัวข้อแฝง						
LDA	✓	-	-	-	-	✓
ระบุความรู้สึก						
VADER Sentiment Tool	-	✓	-	-	-	-
หาความถี่ของคำ						
NLP Frequency Analysis	-	-	-	✓	-	-
การจำแนกความรู้สึก						
Naïve Bayes	✓	-	✓	-	-	✓
SVM	-	-	✓	-	✓	✓
LSTM	-	✓	✓	-	✓	-
BiLSTM	-	-	✓	-	-	-
RNN	-	-	-	-	✓	-
CNN	-	-	✓	-	-	-
Logistic Regression	-	-	-	✓	-	✓

บทที่ 3
วิธีการดำเนินงานวิจัย

ในงานวิจัยนี้ได้กำหนดแนวทางการวิจัยและวิธีการจาก Taecharungroj and Mathayomchan (2019) ซึ่งเป็นการวิเคราะห์ความสนใจและความรู้สึกของนักท่องเที่ยวที่เดินทางยังจังหวัดภูเก็ต ประเทศไทย ที่พัฒนาด้วยโปรแกรม KNIME โดยผู้วิจัยได้นำมาปรับปรุงการวิเคราะห์ให้มีความสมบูรณ์มากขึ้น โดยเพิ่มการทดลอง (Experiment) วิธีการเตรียมข้อมูล รูปแบบต่างๆ ประกอบด้วย ประกอบด้วย 3 ปัจจัย การลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก การลบ/ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำออก และการแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming หรือ Lemmatization ทำให้ได้การทดลองทั้งสิ้น 8 การทดลอง จึงทำให้เกิดรูปแบบของ Bag of Words 8 รูปแบบ และผู้วิจัยได้เพิ่มการจำแนกความรู้สึกของนักท่องเที่ยวด้วยเทคนิคการเรียนรู้เชิงลึก 2 เทคนิค คือ อัลกอริทึมหน่วยความจำระยะยาว-ระยะสั้น ชนิด 2 ทาง (Bidirectional Long Short-Term Memory: BiLSTM) และอัลกอริทึมโครงข่ายประสาทแบบคอนโวลูชั่น (Convolution Neural Network: CNN) ร่วมกับการการทดลองการแก้ปัญหาความไม่สมดุลของชุดข้อมูล 2 วิธี Undersampling และ SMOTE และการทดลองการเตรียมข้อมูลทั้ง 8 การทดลอง พร้อมกับมีการปรับจูนไฮเปอร์พารามิเตอร์ด้วย Grid Search เพื่อให้ได้แบบจำลองโดยใช้อัลกอริทึมด้านการเรียนรู้เชิงลึกที่มีประสิทธิภาพการทำนายที่สูง ตามโครงสร้างงานวิจัยดังรูปที่ 3.1



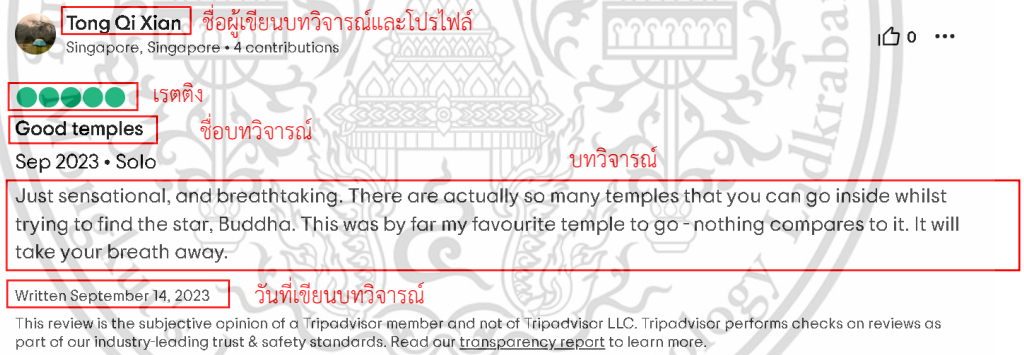
รูปที่ 3.1 โครงสร้างงานวิจัย

3.1 การเก็บข้อมูล

3.1.1 การเก็บข้อมูลจากเว็บไซต์

การเก็บข้อมูลจะเก็บจากหัวข้อ Things to Do in Bangkok จากเว็บไซต์ TripAdvisor ซึ่งเป็นสถานที่ท่องเที่ยวยอดนิยมที่มีการจัดอันดับขึ้นโดยการเก็บข้อมูลของ TripAdvisor เอง โดยอิงจากข้อมูลบทวิจารณ์ เรตติง รูปภาพ และความนิยม ในการเก็บข้อมูลผ่านเว็บไซต์ TripAdvisor จะเก็บข้อมูลด้วยส่วนขยายที่ทำการติดตั้งผ่าน Browser Google Chrome มีชื่อว่า TripAdvisor Review Scraper version 1.0.4 ดังรูปที่ 3.2 ซึ่งจะทำการเก็บข้อมูลบทวิจารณ์ที่เขียนโดยนักท่องเที่ยวต่างชาติหลังจากการเข้าเยี่ยมชมสถานที่ต่างๆในกรุงเทพมหานคร มีหัวข้อดังนี้ ชื่อผู้เขียนบทวิจารณ์และโปรไฟล์ เรตติง ชื่อบทวิจารณ์ บทวิจารณ์ และวันที่เขียนบทวิจารณ์ ดังรูปที่ 3.3 แล้วสร้างออกมาเป็นรูปแบบตาราง โดยจะเป็นไฟล์ในรูปแบบเอกเซล csv ดังรูปที่ 3.4

รูปที่ 3.2 TripAdvisor Review Scraper



รูปที่ 3.3 ตัวอย่างบทวิจารณ์

	A	B	C	D	E	F	G	H
1	User Name	User Profile	Rating	Review Title	Review	Date		
2	Superbunny1	www.tripadvi		3 Seen better an	Not the biggest reclin	Written October 2, 2023		
3	Momin A	www.tripadvi		5 A hidden gem	This ancient temple c	Written October 2, 2023		
4	Raditya A	www.tripadvi		5 Excellent	Bigger reclining Buddh	Written September 30, 2023		
5	John	www.tripadvi		5 Amazing Thail	The biggest reclining I	Written September 28, 2023		
6	Tong Qi Xian	www.tripadvi		5 Good temples	Just sensational, and	Written September 14, 2023		
7	Mei	www.tripadvi		5 Magnificent.	Cultural experience. T	Written September 12, 2023		
8	Craig H	www.tripadvi		4 This review wa	Our guide was very kr	Written August 28, 2023		
9	Janice W	www.tripadvi		5 Historical figur	Great views, holds ma	Written August 28, 2023		
10	Anamika T	www.tripadvi		4 Beautiful and c	It is also called templ	Written August 14, 2023		
11	Sharon P	www.tripadvi		5 A stunningly b	A wonderful place to v	Written August 10, 2023		
12	Anco M	www.tripadvi		5 Impressive	Last day before leavin	Written August 5, 2023		
13	Azriel_and_Smurf	www.tripadvi		4 One of truly ar	A breathtaking beautif	Written August 3, 2023		
14	Nigel D	www.tripadvi		5 Make your owr	Definitely worth going	Written August 1, 2023		
15	Jane R	www.tripadvi		5 Mesmering	Thailand is a country i	Written July 31, 2023		
16	Ching Iam	www.tripadvi		5 Nice temple	Before visiting this pla	Written July 29, 2023		
17	lifenadventure	www.tripadvi		5 Impressive	This is a must see anc	Written July 28, 2023		
18	Jessi	www.tripadvi		5 A must-see for	The temple's intricatel	Written July 23, 2023		

รูปที่ 3.4 ตัวอย่างไฟล์ที่ได้จากการเก็บข้อมูลผ่าน TripAdvisor Review Scraper

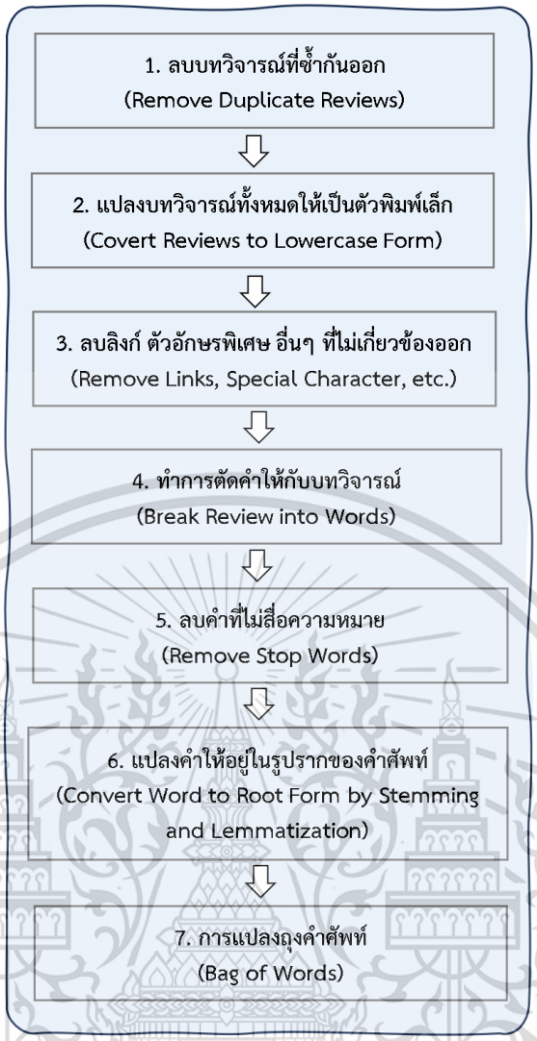
3.1.2 การเลือกข้อมูลสถานที่ในกรุงเทพฯ ที่จะนำมาพิจารณา

จากหมวด Things to Do in Bangkok จากการจัดอันดับโดย TripAdvisor โดยที่ Top 5 สถานที่ท่องเที่ยวในกรุงเทพฯ พบว่ามี 4 จาก 5 อันดับแรกอยู่ในหมวด Sights & Landmarks ได้แก่ Wat Phra Chetuphon, Temple of Dawn (Wat Arun), The Grand Palace, Temple of the Emerald Buddha (Wat Phra Kaew) ทำการเลือกทั้ง 4 สถานที่ จำนวน 15,015 บทวิจารณ์ ดังรูปที่ 3.5

รูปที่ 3.5 สถานที่หมวด Sights & Landmarks

3.2 การเตรียมข้อมูล

หลังจากได้เก็บข้อมูลบทวิจารณ์จากเว็บไซต์แล้วนำข้อความที่ได้มาทำการจัดเตรียมตามกระบวนการเตรียมข้อมูลที่มีลักษณะเป็นข้อความที่คอมพิวเตอร์สามารถนำไปประมวลผลได้ มีลักษณะเป็นข้อความตามขั้นตอนต่อไปนี้



รูปที่ 3.6 ขั้นตอนการเตรียมข้อมูล

ตัวอย่างการเตรียมข้อมูล

จากรูปที่ 3.6 ในขั้นตอนการเตรียมข้อมูล ขอยกตัวอย่างเพื่อความเข้าใจดังนี้ จากตัวอย่างข้อความ “So great to see Thailand richness and history in this beautiful facility. Worth the price (entrance fee) it was so pretty” ในหมวด Sights & Landmarks

1. ลบบทวิจารณ์ที่ซ้ำออก (Remove Duplicate Reviews)

ลบบทวิจารณ์ที่มีการซ้ำกันเนื่องจากขั้นตอนการเก็บข้อมูล การรวมข้อมูล เข้าด้วยกันที่อาจมีการเกิดบทวิจารณ์ที่ซ้ำซ้อนกัน โดยทำการตวจสอบ ใน Column Review หากมี Review (บทวิจารณ์) ที่ซ้ำกันจะทำการลบออกไป

2. แปลงบทวิจารณ์ทั้งหมดให้เป็นตัวพิมพ์เล็ก (Convert Reviews to Lowercase Form)

แปลงบทวิจารณ์ให้อยู่ในรูปตัวพิมพ์เล็กเพื่อให้เข้าใจคำศัพท์นั้นๆ ไปในทิศทางเดียวกัน ดังนั้นจากประโยคบทวิจารณตัวอย่าง แปลงเป็น ตัวพิมพ์เล็กได้ดังนี้ “so great to see thailand richness and history in this beautiful facility. worth the price (entrance fee) it was so pretty”

3. ลบลิงก์ ตัวอักษรพิเศษ อิโมจิ ตัวเลข และอื่นๆ ที่ไม่เกี่ยวข้องออก (Remove Links, Special Character, etc.)

เมื่อเป็นตัวพิมพ์เล็กหมดแล้ว พบว่ามีวงเล็บ () และจุด . ที่ไม่เกี่ยวข้องกับบทวิจารณ์จึงทำการลบออกไป จะได้ “so great to see thailand richness and history in this beautiful facility worth the price entrance fee it was so pretty”

4. ตัดคำให้กับบทวิจารณ์ (Break Reviews into Words)

['so', 'great', 'to', 'see', 'thailand', 'richness', 'and', 'history', 'in', 'this', 'beautiful', 'facility', 'worth', 'the', 'price', 'entrance', 'fee', 'it', 'was', 'so', 'pretty']

5. ลบคำที่ไม่สื่อความหมายออก (Remove Stop Words)

หลังจากตัดคำแล้วจะเห็นว่ามี คำที่ไม่สื่อความหมาย คือ so, to, and, in, this, the, it, was ดังนั้นเมื่อลบคำเหล่าออกจะได้ ดังนี้ ['great', 'see', 'thailand', 'richness', 'history', 'beautiful', 'facility', 'worth', 'price', 'entrance', 'fee', 'pretty']

6. แปลงคำให้อยู่ในรูปรากศัพท์ (Convert Words to Root Form)

หลังจากที่ลบคำที่ไม่สื่อความหมายออกแล้ว จะเห็นว่าคำศัพท์อย่าง richness, history, beautiful, facility, entrace, และ pretty ยังไม่อยู่ในรูปรากศัพท์ ด้วยวิธี Stemming จะสามารถตัดส่วนท้ายของคำออกได้ดังนี้ ['great', 'see', 'thailand', 'rich', 'histori', 'beauti', 'facil', 'worth', 'price', 'entranc', 'fee', 'pretti'] และหากใช้วิธี Lemmatization จะได้ดังนี้ ['great', 'see', 'thailand', 'richness', 'history', 'beautiful', 'facility', 'worth', 'price', 'entrance', 'fee', 'pretty']

7. การแปลงถุงคำศัพท์ (Bag of Words)

สร้างคลังคำศัพท์จากชุดข้อมูลที่ใช้ในการวิเคราะห์ ซึ่งจะใช้คลังคำศัพท์ที่สร้างขึ้นแทนค่าเป็น 1 เมื่อ คำเหล่านั้นปรากฏขึ้นในชุดข้อมูล และเป็น 0 เมื่อ คำเหล่านั้นไม่ปรากฏในชุดข้อมูล เช่น [0, 0, 0, 0, ..., 1, 0, 1, 0]

การทดลองในขั้นตอนการเตรียมข้อมูล

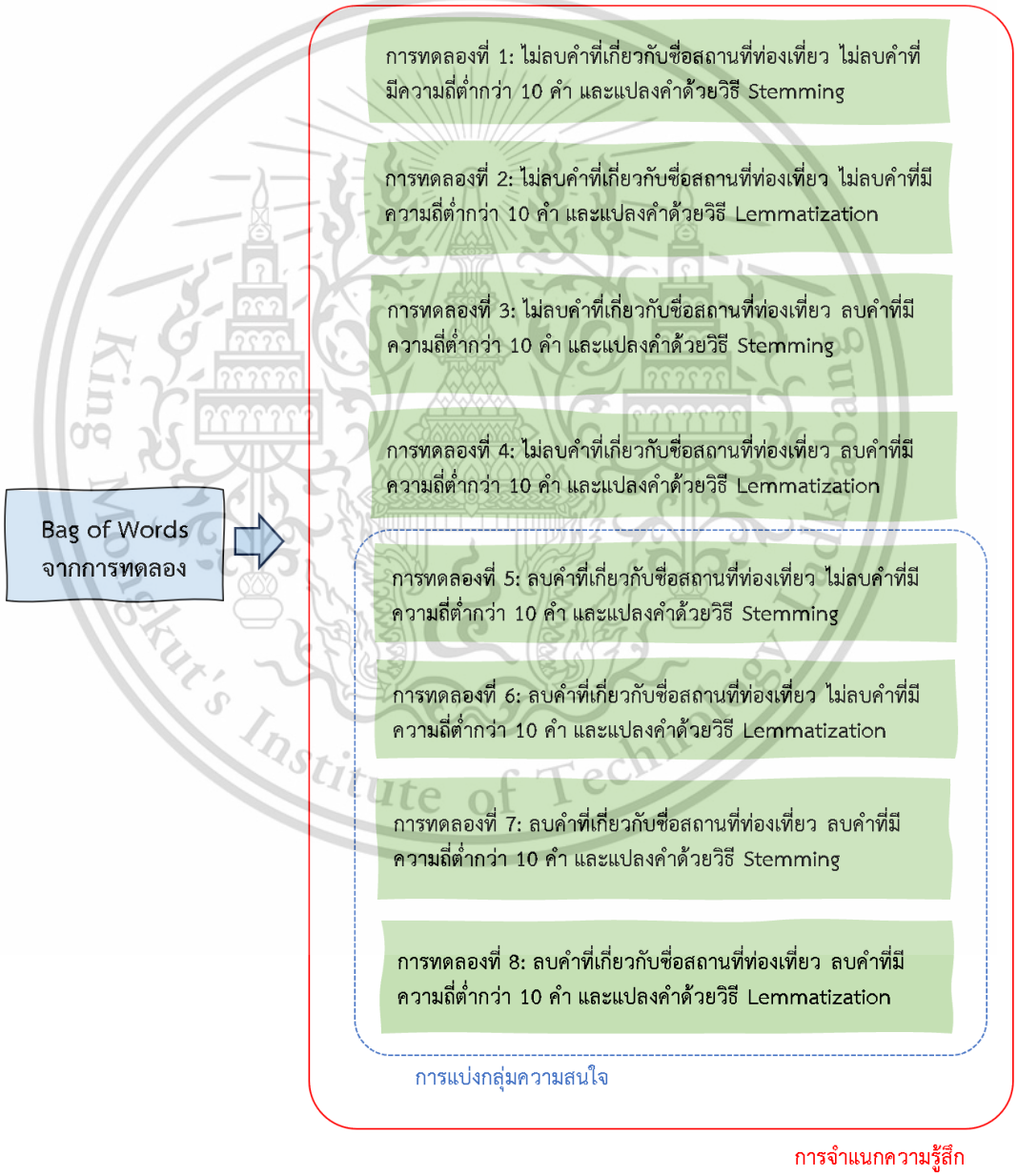
จากการเตรียมข้อมูลแบบปกติในรูปที่ 3.6 ทำการทดลองเพิ่มการเตรียมข้อมูล ในขั้นตอนที่ 5 ดังนี้ เพิ่มการลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก เพิ่มการลบ/ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำออก และเพิ่มการแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming และ Lemmatization ในขั้นตอนที่ 6 โดยมีรายละเอียดดังนี้

เพิ่มการลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก ได้แก่คำว่า “bangkok”, “temple”, “grand”, “palace”, “thailand”, “wat”, “phra”, “chetuphon”, “dawn”, “arun”, “emerald”, “kaew”, “thai” ทั้งหมด 13 คำ

ตัวอย่างเช่น จากขั้นตอนที่ 5 ['great', 'see', 'thailand', 'richness', 'history', 'beautiful', 'facility', 'worth', 'price', 'entrance', 'fee', 'pretty'] หลังจากลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก คำว่า 'thailand' จะหายไป และหากลบคำที่มีความถี่ต่ำกว่า 10 คำออก โดยสมมติว่า 'richness' เป็นคำที่มีความถี่น้อยกว่า 10 คำ ก็จะถูกลบออกไปได้ดังนี้ ['great', 'see', 'history', 'beautiful',

'facility', 'worth', 'price', 'entrance', 'fee', 'pretty'] และหากแปลงคำศัพท์ให้อยู่ในรูปรากศัพท์แบบ Stemming จะเป็น ['great', 'see', 'histori', 'beauti', 'facil', 'worth', 'price', 'entranc', 'fee', 'pretti'] ซึ่งจะเป็นคำศัพท์สุดท้ายก่อนเข้าสู่กระบวนการแปลงถุงคำศัพท์

สุดท้ายหลังจากมีการลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก การลบ/ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำออก และการแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming และ Lemmatization ทำให้ได้การทดลองทั้งสิ้น 8 การทดลอง และจะทำให้เกิด Bag of Words ที่ผ่านการเตรียมข้อมูลทั้งหมด 8 การทดลอง (8 Experiment) ได้ดังรูปที่ 3.7 และมีคำศัพท์ที่เกิดจากการเตรียมข้อมูลที่แตกต่างกันดังตารางที่ 3.1 Bag of Words โดยจะใช้ 4 รูปแบบ สำหรับการเรียนรู้แบบไม่มีผู้สอน (การแบ่งกลุ่มความสนใจ) และ 8 รูปแบบสำหรับการเรียนรู้แบบมีผู้สอน (การจำแนกความรู้สึก)



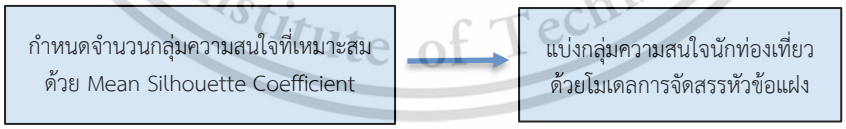
รูปที่ 3.7 Bag of Words จากการเตรียมข้อมูลทั้ง 8 การทดลอง

ตารางที่ 3.1 คำศัพท์ใน Bag of Words จากการเตรียมข้อมูลทั้ง 8 การทดลอง

การทดลอง (Experiment)	จำนวนคำศัพท์ใน Bag of Words
1. ไม่ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming	9,669
2. ไม่ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization	12,878
3. ไม่ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming	2,380
4. ไม่ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization	2,834
5. ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming	9,658
6. ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization	12,865
7. ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming	2,369
8. ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization	2,821

3.3 การแบ่งกลุ่มความสนใจของนักท่องเที่ยว

หลังจากที่ผ่านการเตรียมข้อมูลมาแล้ว บทวิจารณ์ทั้งหมดจะอยู่ในรูปแบบเวกเตอร์ ที่สามารถนำเข้ามาสู่การหาจำนวนกลุ่มที่เหมาะสม ด้วยค่า Mean Silhouette coefficient เมื่อได้จำนวนกลุ่ม (k) จะนำจำนวนกลุ่มที่เหมาะสมเหล่านี้ไปใช้ในการกำหนดกลุ่มความสนใจของนักท่องเที่ยวในโมเดลการจัดสรรหัวข้อแฝง (LDA) เพื่อหาคำศัพท์เฉพาะในแต่ละกลุ่มความสนใจที่ได้กำหนดไว้



รูปที่ 3.8 การแบ่งกลุ่มนักท่องเที่ยว

3.4 การวิเคราะห์ความนิยม (Popularity Analysis)

การวิเคราะห์ความสนใจของนักท่องเที่ยว โดยวิเคราะห์จากจำนวนบทวิจารณ์ในแต่ละกลุ่มความสนใจและคะแนนเฉลี่ยความนิยมของบทวิจารณ์ในกลุ่มนั้นๆ โดยคิดจากคะแนนเรตติงรวมของแต่ละกลุ่มความสนใจหารด้วยจำนวนบทวิจารณ์ ก็จะได้ คะแนนเรตติงเฉลี่ย ที่แสดงถึงความนิยมของนักท่องเที่ยวที่มีต่อกลุ่มความสนใจนั้น

3.5 การวิเคราะห์ความเด่นและความแพร่หลาย (Salience-Valence Analysis)

หลังจากที่ผ่านกระบวนการแบ่งกลุ่มความสนใจของนักท่องเที่ยวด้วยวิธีการจัดสรรหัวข้อแฝง (Latent Dirichlet Allocation: LDA) แล้วจะนำมาพิจาณาต่อเพื่อหาข้อมูลเชิงลึกด้วยวิธีการวิเคราะห์ความเด่นและความแพร่หลายแบบเชิงกลุ่มและเชิงคำศัพท์

3.5.1 การวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม (Dimensional Salience-Valence Analysis)

ความเด่นเชิงกลุ่ม (Dimensional Salience) มีไว้เพื่อหาว่านักท่องเที่ยวมีการกล่าวถึงกลุ่มความสนใจนั้นๆ มากแค่ไหน หาได้จากจำนวนบทวิจารณ์ทั้งหมด หารด้วยจำนวนบทวิจารณ์ทั้งหมดในกลุ่มความสนใจนั้น

ความแพร่หลายเชิงกลุ่ม (Dimensional Valence) มีไว้เพื่อหาว่านักท่องเที่ยวมีความรู้สึกเชิงบวกมากแค่ไหนต่อกลุ่มความสนใจนั้นๆ หาได้จาก สัดส่วนระหว่างจำนวนบทวิจารณ์เชิงบวกลบด้วยจำนวนบทวิจารณ์เชิงบวกคาดหวังแล้วหารด้วยจำนวนบทวิจารณ์ทั้งหมด

ตัวอย่าง Sights & Landmarks กำหนดให้มี 2 กลุ่มความสนใจ

ขั้นตอนที่ 1: เริ่มจากการนับจำนวนบทวิจารณ์ทั้งหมดในแต่ละกลุ่มความสนใจ ได้ดังนี้ กลุ่มความสนใจ 1 ทั้งหมด 7,195 กลุ่มความสนใจ 2 ทั้งหมด 7,758 รวมทั้งสิ้น 14,953

จะได้ตัวแปรดังนี้ r (กลุ่มความสนใจ 1) = 7,195 และ r (กลุ่มความสนใจ 2) = 7,758

ขั้นตอนที่ 2: ทำการหาผลรวมของบทวิจารณ์เชิงบวก ทั้งหมดของทุกๆ กลุ่มความสนใจ ได้ดังนี้

กลุ่มความสนใจ 1 มีบทวิจารณ์เชิงบวก 5,848 จากบทวิจารณ์ทั้งหมด 7,195

กลุ่มความสนใจ 2 มีบทวิจารณ์เชิงบวก 7,415 จากบทวิจารณ์ทั้งหมด 7,758

ดังนั้น ผลรวมของบทวิจารณ์ที่เป็นบทวิจารณ์ที่เป็นเชิงบวกทั้งหมด 13,263 จากบทวิจารณ์ทั้งหมด 14,953

จะได้ตัวแปร ต่างๆ ดังนี้ R_{POS} = 13,263, r_{POS} (กลุ่มความสนใจ 1) = 5,848, r_{POS} (กลุ่มความสนใจ 2) = 7,415 และ R = 14,953

ขั้นตอนที่ 3: หาจำนวนบทวิจารณ์เชิงบวกคาดหวังของ ทั้งหมดของทุกๆ กลุ่มความสนใจ ได้ดังนี้

จากขั้นตอนที่ 2 สามารถหาบทวิจารณ์เชิงบวกคาดหวังของแต่ละกลุ่มความสนใจได้ จากสมการที่ (3.1)

$$e_{POS} = \frac{r \times R_{POS}}{R} \tag{3.1}$$

$$e_{POS(\text{topic1})} = \frac{7{,}195 \times 13{,}263}{14{,}953} = 6{,}382 \tag{3.2}$$

จากสมการที่ (3.2) ตัวอย่างการคำนวณหาบทวิจารณ์เชิงบวกคาดหวังของกลุ่มความสนใจ 1 จะได้

กลุ่มความสนใจ 1 มีบทวิจารณ์เชิงบวกคาดหวัง 6,382

กลุ่มความสนใจ 2 มีบทวิจารณ์เชิงบวกคาดหวัง 6,881

ขั้นตอนที่ 4: หา Dimensional Salience ของแต่ละกลุ่มความสนใจ

จากขั้นตอนที่ 1 และ 2 ทำการแทนค่าในสมการที่ (3.3) เพื่อหา Dimensional Salience ของกลุ่มความสนใจ

$$Dimensional\ salience = \left(\frac{r_{POS}}{R} \right) \times 100 \tag{3.3}$$

$$Dimensional\ salience_{(\text{topic1})} = \left(\frac{5,848}{14,953} \right) \times 100 = 39.11\% \tag{3.4}$$

จากสมการที่ (3.4) ตัวอย่างการคำนวณหา Dimensional Salience ของกลุ่มความสนใจ 1 จะได้

กลุ่มความสนใจ 1 มีค่า Dimensional Salience 39.11%

กลุ่มความสนใจ 2 มีค่า Dimensional Salience 49.59%

ขั้นตอนที่ 5: หา Dimensional Valance ของแต่ละกลุ่มความสนใจ

จากขั้นตอนที่ 1 และ 2 ทำการแทนค่าในสมการที่ (3.5) เพื่อหา Dimensional Valence ของกลุ่มความสนใจ

$$Dimensional\ valence = \left(\frac{r_{POS} - e_{POS}}{r} \right) \times 100 \tag{3.5}$$

$$Dimensional\ valence_{(\text{topic1})} = \left(\frac{5,848 - 6,382}{7,195} \right) \times 100 = -7.42\% \tag{3.6}$$

จากสมการที่ (3.6) ตัวอย่างการคำนวณหา Dimensional Valence ของกลุ่มความสนใจ 1 จะได้

กลุ่มความสนใจ 1 มีค่า Dimensional Valence -7.42%

กลุ่มความสนใจ 2 มีค่า Dimensional Valence 6.88%

ตารางที่ 3.2 สรุปผลที่ได้จากการวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม

กลุ่มความสนใจ	บทวิจารณ์เชิงบวก	บทวิจารณ์เชิงบวกคาดหวัง	จำนวนบทวิจารณ์ทั้งหมด	Dimensional Salience	Dimensional Valence
1	5,848	6,382	7,195	39.11%	-7.42%
2	7,415	6,881	7,758	49.59%	6.88%
รวม	13,263	-	14,953	-	-

3.5.2 การวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ (Lexicon Salience-Valence Analysis)

ความเด่นเชิงคำศัพท์: เป็นการนับความถี่ของคำศัพท์ที่ปรากฏในบทวิจารณ์ ว่ามีจำนวนมากเพียงใด ยิ่งมีค่ามากแสดงว่าคำนั้นถูกกล่าวถึงมาก

ความแพร่หลายเชิงคำศัพท์: จากคำศัพท์ที่กล่าวถึงในบทวิจารณ์ว่าเป็นความรู้สึกเชิงบวกหรือลบมากแค่ไหน

ขั้นตอนที่ 1: ทำการนับความถี่ของคำที่ปรากฏในแต่ละกลุ่มความสนใจว่ามีทั้งหมดเท่าไหร่ ยกตัวอย่างคำว่า “crowd” เป็นคำศัพท์เฉพาะที่อยู่ในกลุ่มความสนใจ 1 มีจำนวนทั้งสิ้น 2,546 คำ แทนด้วยตัวแปร t เพื่อแทนในสมการที่ (3.7)

ตารางที่ 3.3 ตัวอย่างบทวิจารณ์ในกลุ่มความสนใจ 1 ที่มีคำศัพท์ “crowd”

บทวิจารณ์	ความรู้สึก	กลุ่มความสนใจ
crowd people around world place visit	Positive	1
dress propriate entranc fee crowd	Positive	1
tourist visit palace crowd area scam guide	Negative	1
.	.	.
.	.	.
crowd beauti location must visit	Positive	1

จากสมการที่ (3.8) แทนค่า t = 2,546 ก็จะได้ความเด่นเชิงคำศัพท์ (Term salience) เท่ากับ 3.41

$$Term\ salience = \log_{10}(t) \tag{3.7}$$

$$Term\ salience_{(crowd)} = \log_{10}(2{,}546) = 3.41 \tag{3.8}$$

ขั้นตอนที่ 2: นับความถี่ของคำเหล่านั้นว่า อยู่ใน Positive หรือ Negative เท่าไหร่ เพื่อหาค่า $\bar{x}_{POS}$ และ $\bar{x}_{NEG}$

โดย $\bar{x}_{POS}$ คือ จำนวนเฉลี่ยของคำศัพท์เฉพาะที่ปรากฏในบทวิจารณ์เชิงบวกหาได้จากการนับความถี่ของคำนั้นๆ ในแต่ละบทวิจารณ์ที่เป็นเชิงบวกของแต่ละกลุ่มความสนใจ หาร จำนวนบทวิจารณ์ที่เป็นเชิงบวกทั้งหมดของแต่ละกลุ่มความสนใจ และ $\bar{x}_{NEG}$ คือ จำนวนเฉลี่ยของคำศัพท์เฉพาะที่ปรากฏในบทวิจารณ์เชิงลบหาได้จากการนับความถี่ของคำนั้นๆ ในแต่ละบทวิจารณ์ที่เป็นเชิงลบของแต่ละกลุ่มความสนใจหารจำนวนบทวิจารณ์ที่เป็นเชิงลบทั้งหมดของแต่ละกลุ่มความสนใจ

จากตัวอย่างในตารางที่ 3.3 เป็นบทวิจารณ์ที่อยู่ในกลุ่มความสนใจ 1 เป็นบทวิจารณ์ที่เป็นเชิงบวกทั้งหมด 5,848 และเชิงลบทั้งหมด 1,347 ทำการนับความถี่ของคำว่า “crowd” ที่ปรากฏขึ้นในแต่ละบทวิจารณ์ทั้งที่เป็นเชิงบวกและเชิงลบ พบว่า มีคำว่า “crowd” ปรากฏขึ้นในบทวิจารณ์ที่เป็นเชิงบวกทั้งสิ้น 1,918 ครั้ง ปรากฏในบทวิจารณ์เชิงลบทั้งสิ้น 628 ครั้ง ดังนั้น สามารถหา $\bar{x}_{POS}$ และ $\bar{x}_{NEG}$ ของคำว่า “crowd” ได้ดังสมการที่ (3.9) (3.10) และ (3.11) (3.12) ตามลำดับ

$$\bar{x}_{POS_{(\text{crowd})}} = \frac{Frequency\ of\ crowd\ in\ positive\ reviews}{Total\ positive\ reviews} \tag{3.9}$$

$$\bar{x}_{POS_{(crowd)}} = \frac{1,918}{5,848} = 0.33 \tag{3.10}$$

$$\bar{x}_{NEG_{(crowd)}} = \frac{Frequency\ of\ crowd\ in\ negative\ reviews}{Total\ negative\ reviews} \tag{3.11}$$

$$\bar{x}_{NEG_{(crowd)}} = \frac{628}{1,347} = 0.47 \tag{3.12}$$

ขั้นตอนที่ 3: หาความแพร่หลายเชิงคำศัพท์ (Term valence)

จากขั้นตอนที่ 2 เมื่อได้ $\bar{x}_{POS_{(crowd)}}$ และ $\bar{x}_{NEG_{(crowd)}}$ แล้ว สามารถหาความแพร่หลายเชิงคำศัพท์ Term Valence ได้จากสมการที่ (3.13) และ (3.14)

$$Term\ valence = \frac{\bar{x}_{POS} - \bar{x}_{NEG}}{\bar{x}_{POS} + \bar{x}_{NEG}} \tag{3.13}$$

$$Term\ valence = \frac{0.33 - 0.47}{0.33 + 0.47} = -0.17 \tag{3.14}$$

ในกรณีของคำอื่นๆ ก็สามารถหาได้โดยใช้หลักการเดียวกันกับคำว่า “crowd”

ตารางที่ 3.4 สรุปผลที่ได้จากการวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์

กลุ่มความสนใจ	คำศัพท์	ความถี่ของคำทั้งหมด	ความถี่ของคำในบทวิจารณ์เชิงบวก	ความถี่ของคำในบทวิจารณ์เชิงลบ	จำนวนบทวิจารณ์เชิงบวกทั้งหมด	จำนวนบทวิจารณ์เชิงลบทั้งหมด	$\bar{x}_{POS}$	$\bar{x}_{NEG}$	Term Salience	Term Valence
1	crowd	2,546	1,918	628	5,848	1,347	0.33	0.47	3.41	- 0.17

3.6 การจำแนกความรู้สึก

จากข้อมูลที่ผ่านการเตรียมข้อมูลแล้วบทวิจารณ์แล้วจาก 15,015 บทวิจารณ์จะเหลือ 14,953 บทวิจารณ์ จากคะแนนเรตติงที่ถูกกำหนดไว้ในแต่ละบทวิจารณ์ จะทำการระบุความรู้สึกให้กับบทวิจารณ์ โดยจะกำหนดให้เป็นความรู้สึกเชิงบวก เมื่อคะแนนเรตติงเป็น 4-5 คะแนน และเป็นความรู้สึกเชิงลบเมื่อคะแนนเรตติงเป็น 1-3 คะแนน ความรู้สึกเชิงบวกและลบทั้งหมดในบทวิจารณ์หลังจากระบุด้วยคะแนนเรตติง จะได้ดังตารางที่ 3.5

ตารางที่ 3.5 ความรู้สึกเชิงบวกและลบในบทวิจารณ์ทั้งหมด

ความรู้สึก		รวม
เชิงบวก	เชิงลบ	
13,263	1,690	14,953

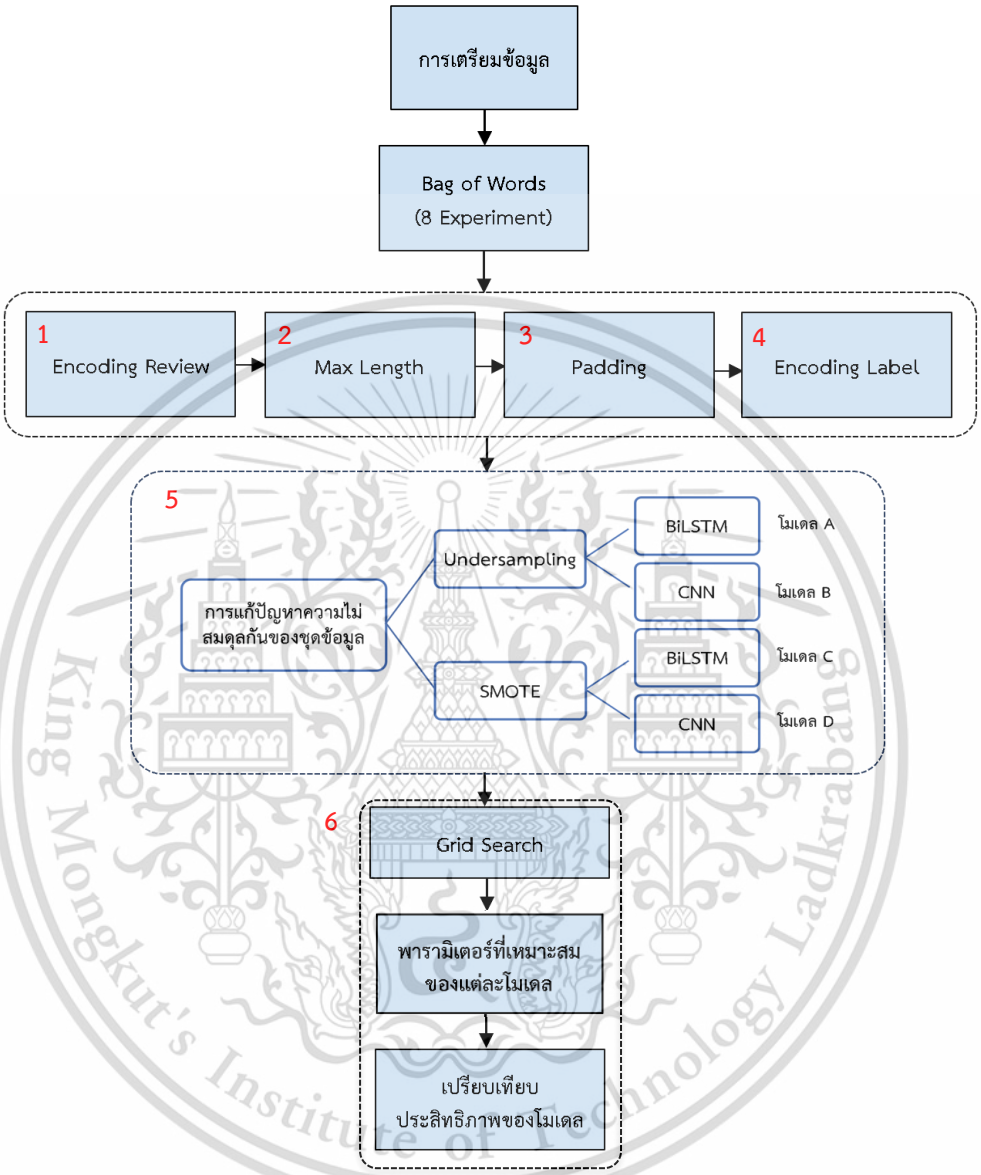
3.6.1 การแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล

จากตารางที่ 3.5 จะพบว่าบทวิจารณ์เชิงลบและเชิงบวกมีความแตกต่างกันสูง ทำให้เกิดปัญา Imbalance Data โดยการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล จะมี 2 วิธี 1) Undersampling (การสุ่มเพื่อลดข้อมูลกลุ่มมากให้เท่ากับข้อมูลกลุ่มน้อย) และ 2) SMOTE: Synthetic Minority Over-sampling Technique (การสุ่มตัวอย่างสังเคราะห์ของข้อมูลกลุ่มน้อยให้เท่ากับข้อมูลกลุ่มมาก)

3.6.2 ค่าคุณลักษณะของโมเดล (Feature of Model)

หลังจากที่ผ่านกระบวนการเตรียมข้อมูลมาแล้ว จะได้รูปแบบของ Bag of words ทั้ง 8 รูปแบบ จาก 8 การทดลอง ก่อนที่จะเข้าสู่กระบวนจำแนกความรู้สึกด้วยแบบจำลองการเรียนรู้เชิงลึกทั้งแบบ BiLSTM และ CNN ต้องมีการเตรียมคุณลักษณะ ดังรูปที่ 3.9 ประกอบด้วยขั้นตอนการเตรียมคุณลักษณะ 4 ขั้นตอน คือ Encoding Review, Max Length, Padding และ Encoding Label จากนั้น มีการทดลองในส่วนของการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล 2 วิธี คือ Undersampling และ SMOTE จากนั้นจะทำการแบ่งส่วนข้อมูล ออกเป็น 3 ส่วน คือ ข้อมูลสำหรับฝึกสอน (Training set) 70% ข้อมูลสำหรับการตรวจสอบความถูกต้อง (Validation set) 15% และข้อมูลสำหรับการทดสอบ (Testing set) 15% โดยจะนำข้อมูลฝึกสอน (Training) มาหาพารามิเตอร์ที่เหมาะสมด้วย Grid Search เมื่อได้แล้วนำมาฝึกสอนโมเดลด้วยชุดข้อมูลฝึกสอน (Training) และใช้

ชุดข้อมูลตรวจสอบความถูกต้อง (Validation set) เพื่อประเมินประสิทธิภาพว่าโมเดลฝึกสอนมีประสิทธิภาพหรือไม่ และสุดท้ายจะวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Testing set)



รูปที่ 3.9 การเตรียมคุณลักษณะ

จากรูปที่ 3.9 เป็นการเตรียมคุณลักษณะก่อนเข้าสู่โมเดลมีรายละเอียดดังต่อไปนี้

1. หลังจากที่ผ่านกระบวนการเตรียมข้อมูล การลบ/ไม่ลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก การลบ/ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำออก และการแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming และ Lemmatization จะได้รูปแบบของ Bag of Words ตามการทดลองการเตรียมข้อมูลทั้ง 8 การทดลอง ซึ่งใน Bag of Words จะมีรูปแบบของคำศัพท์ (Vocabulary) ของแต่ละการทดลอง เป็นตัวแทน Feature ของบทวิจารณ์ที่จะเข้าไปฝึกสอนโมเดล

โดยจะนำ Index ของคำศัพท์นั้นๆ ใน Bag of Words เป็นตัวแทนของคำศัพท์ในบทวิจารณ์ เช่น ประโยค [richness history beautiful facility worth price] เมื่อผ่านการแปลงข้อความด้วยวิธี One-Hot Encoding ได้เป็น ถุงคำศัพท์ ที่มี Index ของแต่ละคำระบุเอาไว้ สมมติว่า richness (Index: 9449), history (Index: 5231), beautiful (Index: 1071), facility (Index: 4145), worth (Index: 12724) , price (Index: 8598)

จากรูปแบบ Bag of Words

[richness history beautiful facility worth price] -> [0, 0, 0, 0, ..., 1, 0, 0, 0]

แทนด้วย Index ของคำศัพท์

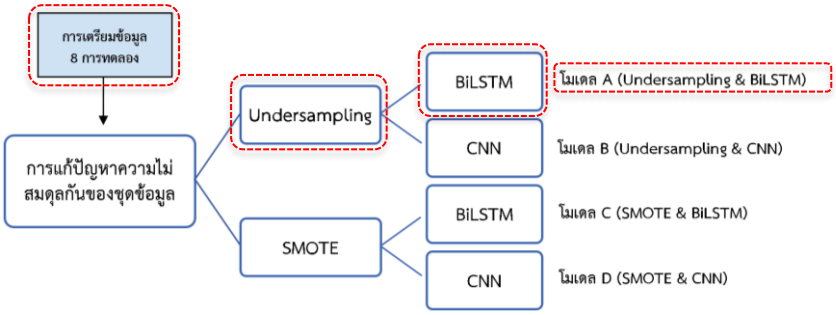
[richness history beautiful facility worth price] -> [9449, 5231, 1071, 4145, 12724, 8598]

2. ทำการหาค่า Max Length ของบทวิจารณ์ เป็นการหาความยาวของบทวิจารณ์ที่มีคำศัพท์เกิดขึ้นมากที่สุดตัวอย่างเช่น [tour much better get historical perspective place get know .....] สมมติว่าบทวิจารณ์นี้มีคำศัพท์ยาวต่อเนื่องไป จาก คำว่า know รวมจำนวนทั้งหมด 300 คำ จะได้ค่า Max Length ที่แทนความยาวของบทวิจารณ์ทั้งหมดที่ยาวที่สุดเท่ากับ 300 คำ

3. Padding เป็นการเตรียมบทวิจารณ์ทั้งหมดให้มีความยาวเท่ากัน โดยการแทนที่ด้วย 0 เพื่อจะให้มีความยาวเท่ากับบทวิจารณ์ที่ยาวที่สุดก่อนหน้านี้คือ 300 คำ ตัวอย่างเช่น จากขั้นตอนที่ 1 [richness history beautiful facility worth price] มีความยาว 6 คำ เมื่ออยู่ในรูปแบบ Encoding ด้วย Index ของคำศัพท์จาก Bag of words จะได้ [9449, 5231, 1071, 4145, 12724, 8598] เมื่อทำการ Padding จะทำการเติม 0 ต่อไปจาก 8598 ดังนี้ [9449, 5231, 1071, 4145, 12724, 8598, 0, 0, 0, ..., 0] เติม 0 ส่วนท้ายไปเรื่อยๆ จากความยาวทั้งหมดจาก 6 คำ จนครบ 300 คำ สุดท้ายบทวิจารณ์ทั้งหมดจะอยู่ในรูปแบบที่มีความยาวเท่ากันทั้งหมด

4. ทำการ Encoding Label คำตอบของบทวิจารณ์จาก Positive และ Negative จะถูกแปลงให้อยู่ในรูปแบบตัวเลข ดังนี้ [0 1] แทน Positive และ [1 0] แทน Negative

5. หลังจากที่ผ่านกระบวนการเตรียมคุณลักษณะในขั้นตอนที่ 1 - 4 ต่อไปจะทำการจำแนกความรู้สึกเชิงบวกและเชิงลบด้วยอัลกอริทึม Bidirectional Long Short-Term Memory (BiLSTM) และ Convolution Neural Network (CNN) ร่วมกับการการทดลองการแก้ปัญหาความไม่สมดุลของชุดข้อมูล 2 วิธี Undersampling และ SMOTE ประกอบด้วย โมเดล A (Undersampling & BiLSTM) โมเดล B (Undersampling & CNN) โมเดล C (SMOTE & BiLSTM) และโมเดล D (SMOTE & CNN) ดังรูปที่ 3.10



รูปที่ 3.10 โมเดล A – D จำแนกตามการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล และการเตรียมข้อมูลทั้ง 8 การทดลอง

6. จากนั้นจะทำการแบ่งส่วนข้อมูล ออกเป็น 3 ส่วน คือ ข้อมูลสำหรับฝึกสอน (Training set) 70% ข้อมูลสำหรับการตรวจสอบความถูกต้อง (Validation set) 15% และข้อมูลสำหรับการทดสอบ (Testing set) 15% โดยจะนำข้อมูลฝึกสอน (Training) มาเพื่อหาค่า Hyperparameter ที่เหมาะสมด้วย Grid Search ก่อนนำมาฝึกสอนโมเดลด้วยชุดข้อมูลฝึกสอน (Training) และใช้ชุดข้อมูลตรวจสอบความถูกต้อง (Validation set) เพื่อประเมินประสิทธิภาพว่าโมเดลฝึกสอนมีประสิทธิภาพหรือไม่ และสุดท้ายจะเปรียบเทียบประสิทธิภาพของโมเดลด้วยชุดข้อมูลทดสอบ (Testing set)

3.6.3 อัลกอริทึมหน่วยความจำระยะยาว-ระยะสั้นชนิด 2 ทาง Bidirectional Long Short-Term Memory: BiLSTM)

หลังจากที่ได้ผ่านการะบวนการเตรียมคุณลักษณะในหัวข้อ 3.6.2 นำข้อมูลมาฝึกสอนด้วยอัลกอริทึม BiLSTM ด้วยวิธี Undersampling (โมเดล A) และวิธี SMOTE (โมเดล C) ด้วยค่า Hyperparameter อ้างอิงจากหัวข้อ 2.3.4 ของ BiLSTM โดยจะมีการกำหนดช่วงของ LSTM Cell ในชั้น LSTM Layer และจำนวนโหนดในชั้น Fully Connected ที่จะทำการจูนเพื่อหาค่าที่ดีที่สุด โดยที่ค่าอื่นๆ จะคงที่ ดังตารางที่ 3.6 การกำหนดค่าพารามิเตอร์ และรายละเอียดการตั้งค่า Search Space ดูได้ที่ภาคผนวก ค

ตารางที่ 3.6 การตั้งค่า Hyperparameter ของโมเดล A และโมเดล C

Hyperparameter	ค่าที่ใช้ในแต่ละ Hyperparameter	
	โมเดล A	โมเดล C
Epoch	5	5
Batch Size	50	50
Learning Rate	0.001	0.001
Optimizer	Adam	Adam
LSTM Units (LSTM Layer)	**[176, 216, 256, 296, 336]**	**[18, 24, 30]**
Dropout Rate	0.2	0.5
Fully Connected Units	**[88, 128, 168]**	**[6, 8]**
Output Activation Function	Sigmoid	Sigmoid

3.6.4 อัลกอริทึมโครงข่ายประสาทแบบคอนโวลูชั่น (Convolution Neural Network: CNN)

หลังจากที่ได้ผ่านการระบวนการเตรียมคุณลักษณะในหัวข้อ 3.6.2 นำข้อมูลมาฝึกสอนด้วยอัลกอริทึม CNN ด้วยวิธี Undersampling (โมเดล B) และวิธี SMOTE (โมเดล D) ด้วยค่า Hyperparameter อ้างอิงจากหัวข้อ 2.3.4 ของ CNN โดยจะมีการกำหนดจำนวน Filter และ Kernel Size ในชั้น Convolution ที่จะทำการจูนเพื่อหาค่าที่ดีที่สุด โดยค่าอื่นๆ จะคงที่ แสดงดังตารางที่ 3.7 การกำหนดพารามิเตอร์ และรายละเอียดการตั้งค่า Search Space ดูได้ที่ภาคผนวก ค

ตารางที่ 3.7 การตั้งค่า Hyperparameter ของโมเดล B และโมเดล D

Hyperparameter	ค่าที่ใช้ในแต่ละ Hyperparameter	
	โมเดล B	**โมเดล D**
Epochs	20	20
Batch Size	64	64
Learning Rate	0.001	0.001
Optimizer	Adam	Adam
Filter	**[16, 32, 48, 64]**	**[8, 16, 24]**
Kernel Size	**[3, 5, 7, 9]**	**[3, 7]**
Dropout Rate	0.5	0.5
Fully Connected Units	10	10
Output Activation	Sigmoid	Sigmoid

บทที่ 4
ผลการวิจัยและการอภิปรายผล

จากบทที่ 3 ที่ได้นำเสนอขั้นตอนการดำเนินการต่างๆ ในงานวิจัย ต่อไปจะเป็น ส่วนของผลการวิจัยและอภิปรายผล จะประกอบไปด้วย การแบ่งกลุ่มความสนใจของนักท่องเที่ยว การวิเคราะห์ความนิยม การวิเคราะห์ความเด่นและความแพร่หลาย การจำแนกความรู้สึก

4.1 ชุดข้อมูลทดลอง

ชุดข้อมูลทดลองที่ได้จากการเก็บข้อมูลมาจากเว็บไซต์ TripAdvisor ในบทวิจารณ์รูปแบบภาษาอังกฤษทำการเก็บข้อมูลตั้งแต่วันที่ 1 เดือนมกราคาม ปี 2018 ถึงสิ้นเดือนกันยายน 2023 มีบทวิจารณ์จากทั้ง 4 สถานที่ ประกอบด้วย 1. วัดพระเชตุพนฯ (Wat Phra Chetuphon) 2. วัดอรุณ (Temple of Dawn (Wat Arun)) 3. พระบรมมหาราชวัง (The Grand Palace) 4. วัดพระแก้ว (Temple of the Emerald Buddha (Wat Phra Kaew)) รวม 15,015 บทวิจารณ์ หลังจากการผ่านกระบวนการเตรียมข้อมูลแล้วจะมีบทวิจารณ์ทั้งสิ้น 14,953 บทวิจารณ์ มีชุดข้อมูลทั้งหมด 8 การทดลอง จากรูปที่ 4.1 เป็นตัวอย่างชุดข้อมูลที่ผ่านการลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก ลบคำที่มีความถี่ต่ำกว่า 10 คำออกและแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming ก่อนที่จะเข้าสู่กระบวนการ Bag of Words

	cleaned_words	Label	Rating
0	great see rich histori beauti facil worth pric...	Positive	5
1	earlier decid guid tour tour last one half hou...	Positive	5
2	build amaz good take photo best experi visit	Positive	5
3	time visit one place place allow plenti time l...	Positive	5
4	absolut worth visit due histori beauti entir c...	Positive	5
...	...	...	...
15011	beauti buddha get terribl close must respect t...	Positive	5
15012	unfortun arriv around hoard tourist travel via...	Negative	3
15013	countless attract tri squeez list even quick v...	Positive	5
15014	say visit site amaz beauti creativ went creat	Positive	5
15015	nice thing sure would excit	Negative	3

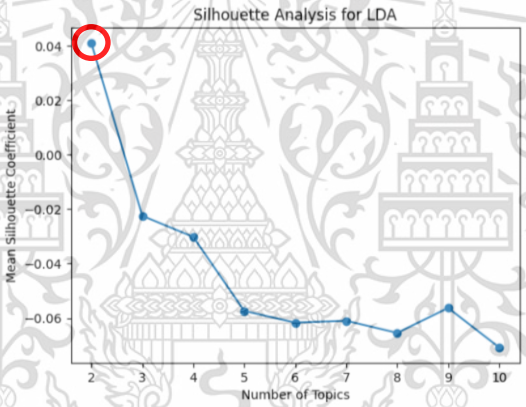
14953 rows × 4 columns

รูปที่ 4.1 ตัวอย่างชุดข้อมูลที่ทำลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออก ลบคำที่มีความถี่ต่ำออก และแปลงคำให้อยู่ในรูปรากศัพท์ด้วยวิธี Stemming (การทดลองที่ 7)

4.2 ผลของการแบ่งกลุ่มความสนใจนักท่องเที่ยว

4.2.1 ผลการวิเคราะห์ค่า Mean Silhouette Coefficient เพื่อหาจำนวนกลุ่มความสนใจที่เหมาะสม

หลังจากที่ข้อมูลผ่านกระบวนการเตรียมข้อมูลแล้วทั้ง 8 การทดลองแล้ว ทำการเลือก 4 การทดลองที่จะนำมาใช้ในการแบ่งกลุ่มความสนใจของนักท่องเที่ยว คือ การทดลองที่ 5, 6, 7 และ 8 ตามอ้างอิงในหัวข้อ 2.1.8 จะทำการหาจำนวนกลุ่มความสนใจที่เหมาะสมด้วย Mean Silhouette Coefficient ที่มีค่ามากสุด จากรูปที่ 4.2 เป็นตัวอย่างของข้อมูลที่ผ่านกระบวนการเตรียมข้อมูลแบบการทดลองที่ 5 จะเป็นกราฟความสัมพันธ์ระหว่าง Mean Silhouette Coefficient และจำนวนกลุ่มความสนใจ (Number of Topics) ที่ตั้งค่าไว้เริ่มต้นที่ 2 ถึง 10 กลุ่มความสนใจ จากรูปจะเห็นว่า ที่กลุ่มความสนใจเท่ากับ 2 มีค่า Mean Silhouette Coefficient สูงที่สุดที่ 0.04 และยิ่งจำนวนกลุ่มความสนใจเพิ่มขึ้น (Number of Topics) ค่า Mean Silhouette Coefficient มีค่าลดลงเรื่อยๆ ดังนั้น กลุ่มความสนใจของ การทดลองที่ 5 คือ 2 กลุ่มความสนใจ และผลของการทดลองอื่นๆ เท่ากับ 2 กลุ่มความสนใจเช่นกันทั้ง 4 การทดลอง ดังรูปที่ ก.1 – ก.4 ตามภาคผนวก ก สรุปดังตารางที่ 4.1



รูปที่ 4.2 ค่า Mean Silhouette Coefficient ต่อจำนวนกลุ่มความสนใจที่คำนวณจากการเตรียมข้อมูลการทดลองที่ 5

ตารางที่ 4.1 กลุ่มความสนใจที่เหมาะสมตามการเตรียมข้อมูล

การเตรียมข้อมูล	จำนวนกลุ่มความสนใจ
การทดลองที่ 5: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming	2
การทดลองที่ 6: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization	2
การทดลองที่ 7: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming	2
การทดลองที่ 8: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization	2

4.2.2 ผลของโมเดลการจัดสรรหัวข้อแฝง

หลังจากที่ได้ผลลัพธ์จากการวิเคราะห์หากลุ่มความสนใจที่เหมาะสมด้วย Mean Silhouette Coefficient แล้ว ต่อไปจะทำการแบ่งกลุ่มความสนใจของนักท่องเที่ยวด้วยโมเดลการจัดสรรหัวข้อแฝง ตามรูปแบบตามจำนวนกลุ่มที่เหมาะสมจากตารางที่ 4.1 โดยทำการวิเคราะห์ทั้ง 4 การทดลอง เพื่อแสดงผลคำศัพท์เฉพาะของแต่ละกลุ่มความสนใจโดยเลือกแสดงคำศัพท์ตามค่าน้ำหนัก 20 อันดับแรก โดยที่ค่าน้ำหนักแสดงถึงความสำคัญของคำศัพท์ในแต่ละกลุ่มความสนใจ ดังตารางในภาคผนวก ข ตามตารางที่ ข.1 – ข.4

ตารางที่ 4.2 คำศัพท์ในกลุ่มความสนใจของแต่ละการทดลอง

การทดลอง	กลุ่มความสนใจ	คำศัพท์ในกลุ่มความสนใจ
5	กลุ่ม 1	“buddha”, “visit”, “see”, “place”, “beauty”, “reclin”, “must”, “one”, “crowd”, “worth”, “amaz”, “templ”, “time”, “take”, “build”, “statu”, “mani”, “get”, “lot”, “guid”
	กลุ่ม 2	“get”, “go”, “visit”, “take”, “baht”, “river” “cover”, “place”, “boat”, “around”, “wear”, “see”, “tourist”, “peopl”, “walk” “entranc”, “time”, “dress”, “crowd”, “long”
6	กลุ่ม 1	“get”, “take”, “place”, “visit”, “go”, “see”, “baht”, “river”, “around”, “must”, “boat”, “long”, “beautiful”, “tourist”, “entrance”, “people”, “dress”, “time”, “wear”, “one”
	กลุ่ม 2	“buddha”, “visit”, “see”, “place”, “reclining”, “beautiful”, “one”, “must”, “time”, “worth”, “amazing”, “building”, “statue”, “tour”, “many”, “guide”, “well”, “take”, “really”, “around”
7	กลุ่ม 1	“get”, “go”, “visit”, “place”, “crowd”, “take”, “see”, “cover”, “tour”, “peopl”, “guid”, “tourist”, “dress”, “baht”, “wear”, “time”,“around”, “must”, “long”, “hot”
	กลุ่ม 2	“buddha” “visit” “see” “beauti” “place” “reclin” “one” “must” “worth” “amaz” “templ” “river” “time” “take” “statu” “build” “architectur” “around” “mani” “well”
8	กลุ่ม 1	“visit”, “place”, “see”, “buddha” “beautiful”, “one”, “time”, “must” “worth”, “amazing”, “tour”, “reclining” “river”, “get”, “tourist”, “guide”, “really” “great”, “many”, “building”
	กลุ่ม 2	“buddha”, “get”, “take”, “baht”, “long” “go”, “see”, “entrance”, “water”, “around”, “must”, “dress”, “inside”, “wear”, “place”, “visit”, “shoulder”, “short”, “also”, “sure”

จากคำศัพท์ที่แสดงผลในตารางที่ 4.2 เมื่อนำมาหาค่าเฉลี่ยจากค่าน้ำหนัก 20 คำแรก ที่มีความถี่มากที่สุด โดยรายละเอียดของค่าน้ำหนักนำเสนอในภาค ผนวก ข ตามตารางที่ ข.1 – ข.4 จะได้ดังตารางที่ 4.3 เป็นการหาค่าน้ำหนักเฉลี่ยของกลุ่ม 1 และกลุ่ม 2 ทั้ง 4 การทดลอง แล้วนำมาหาค่าเฉลี่ยรวมของทั้งสองกลุ่มในแต่ละการทดลอง จะพบว่า การทดลองที่ 8 มีค่าเฉลี่ยของน้ำหนักทั้งสองกลุ่มมากที่สุด ดังนั้นจะทำการเลือกพิจารณา การทดลองที่ 8: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization มาเป็นตัวหลักในการพิจารณาหัวข้อถัดไป

ตารางที่ 4.3 น้ำหนักเฉลี่ยของกลุ่มความสนใจของทุกการทดลอง

การทดลอง	**น้ำหนักเฉลี่ย**		**ค่าเฉลี่ยรวม 2 กลุ่ม**
	กลุ่ม 1	**กลุ่ม 2**	
5	0.42	0.58	0.500
6	0.61	0.36	0.485
7	0.62	0.39	0.490
8	**0.50**	**0.62**	**0560**

จาก การทดลองที่ 8: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization จะทำการตั้งชื่อกลุ่มความสนใจของนักท่องเที่ยว การตั้งชื่อกลุ่มความสนใจสามารถใช้ผู้เชี่ยวชาญลงความเห็นโดยการพิจาณาจากกลุ่มคำศัพท์ที่มีค่าน้ำหนักของคำสูง ตามการศึกษาของ Taecharungroj and Mathayomchan (2019) สำหรับงานวิจัยนี้ใช้ความคิดเห็นจากผู้วิจัยและยืนยันความคิดเห็นของการตั้งชื่อกลุ่มโดยอาจารย์ที่ปรึกษา โดยในงานวิจัยนี้ผู้วิจัยจะพิจารณาจากคำศัพท์ที่เป็นเอกลักษณ์ของแต่ละกลุ่มจากมุมมองของผู้วิจัยเอง จากตารางที่ 4.4 พบว่าสามารถตั้งชื่อกลุ่มความสนใจได้ 2 แบบ คือ กลุ่มที่ 1 ตั้งชื่อว่า กลุ่มความประทับใจในสถานที่ มีคำศัพท์ที่เป็นเอกลักษณ์ เช่น “beautiful”, “worst” , “must”, “amazing” และกลุ่มที่ 2 ตั้งชื่อว่า กลุ่มการเข้าชมสถานที่ มีคำศัพท์ที่เป็นเอกลักษณ์ เช่น “get”, “take”, “long”, “go”, “entrance”

ตารางที่ 4.4 คำศัพท์ของกลุ่มความสนใจจากโมเดลการจัดสรรหัวข้อแฝงตามการเตรียมข้อมูลแบบการทดลองที่ 8

ความประทับใจในสถานที่		การเข้าชมสถานที่	
คำศัพท์	**น้ำหนัก**	**คำศัพท์**	**น้ำหนัก**
visit	1.00	buddha	1.00
place	0.94	get	0.94
see	0.87	take	0.86
buddha	0.81	baht	0.79
beautiful	0.67	long	0.67
one	0.53	go	0.66
time	0.52	see	0.63
must	0.49	entrance	0.58
worth	0.45	water	0.56
amazing	0.40	around	0.55
tour	0.36	must	0.55
reclining	0.35	dress	0.54
river	0.34	inside	0.54
get	0.33	wear	0.53
tourist	0.31	place	0.52
guide	0.31	visit	0.51
really	0.31	shoulder	0.48
great	0.31	short	0.47
many	0.31	also	0.47
building	0.30	sure	0.47
ค่าเฉลี่ย	0.50	ค่าเฉลี่ย	0.62

4.3 ผลการวิเคราะห์ความนิยม

แสดงถึงความนิยมที่นักท่องเที่ยวให้ไว้ในแต่ละบทวิจารณ์โดยแสดงตามจำนวนบทวิจารณ์ที่อยู่ในแต่ละกลุ่มความสนใจ และคะแนนเรตติงเฉลี่ยมาพิจาณาเปรียบเทียบกัน จากผลในการทดลองที่ 8: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization แสดงให้เห็นว่านักท่องเที่ยวมีความสนใจด้านความประทับใจในสถานที่มากกว่าการเข้าชมสถานที่ ดูได้จากจำนวนบทวิจารณ์และคะแนนเรตติงที่นักท่องเที่ยวเขียนทั้งหมดไปอยู่ในกลุ่มความสนใจด้านความประทับใจในสถานที่ถึง 9,771 บทวิจารณ์ มีคะแนนเรตติงเฉลี่ย 4.55 ซึ่งมากกว่ากลุ่มความสนใจด้านการเข้าชมสถานที่ที่มีบทวิจารณ์ 5,182 บทวิจารณ์ มีคะแนนเรตติงเฉลี่ยที่ 4.26 ดังตารางที่ 4.5

ตารางที่ 4.5 คะแนนเรตติงเฉลี่ยในแต่ละกลุ่มความสนใจ

กลุ่มความสนใจ	จำนวนบทวิจารณ์	คะแนนเรตติงเฉลี่ย
ความประทับใจในสถานที่	9,771	4.55
การเข้าชมสถานที่	5,182	4.26

4.4 ผลการวิเคราะห์ความเด่นและความแพร่หลาย

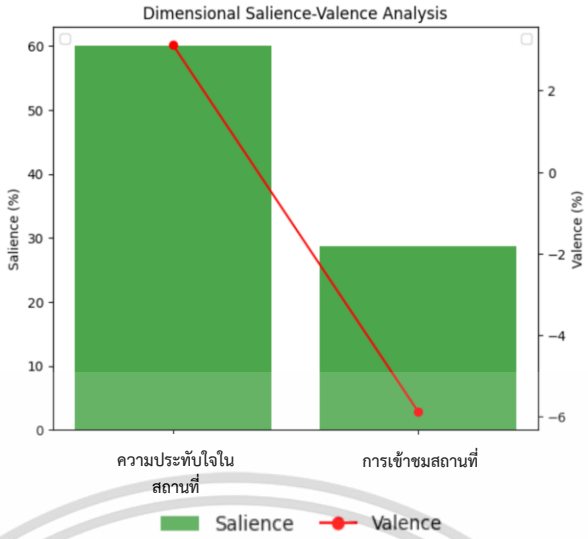
4.4.1 ผลการวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม (Dimensional Salience Valence Analysis: DSVA)

กลุ่มความสนใจของนักท่องเที่ยวในหมวด Sights & Landmarks ทั้ง 2 กลุ่ม ใน การทดลองที่ 8: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization จะนำมาวิเคราะห์ความเด่นเชิงกลุ่ม คือ การหาว่านักท่องเที่ยวมีกล่าวถึงกลุ่มความสนใจนั้นๆ มากแค่ไหน ซึ่งหาได้จากจำนวนบทวิจารณ์ทั้งหมดหารด้วยจำนวนบทวิจารณ์ทั้งหมดในกลุ่มความสนใจนั้น และความแพร่หลายเชิงกลุ่ม เพื่อหาว่านักท่องเที่ยวมีความรู้สึกเชิงบวกมากแค่ไหนต่อกลุ่มความสนใจนั้นๆ หาได้จากสัดส่วนระหว่างจำนวนบทวิจารณ์เชิงบวกลบด้วยจำนวนบทวิจารณ์เชิงบวกคาดหวังแล้วหารด้วยจำนวนบทวิจารณ์ทั้งหมด

ตารางที่ 4.6 ความเด่นและความแพร่หลายของกลุ่มความสนใจ

กลุ่มความสนใจ	จำนวนบทวิจารณ์เชิงบวก	จำนวนบทวิจารณ์เชิงบวกที่คาดหวัง	จำนวนบทวิจารณ์ทั้งหมด	ความเด่น (Salience)	ความแพร่หลาย (Valence)
ความประทับใจในสถานที่	8,971	8,667	9,771	59.99%	3.11%
การเข้าชมสถานที่	4,292	4,596	5,182	28.70%	-5.87%
รวม	**13,263**	-	**14,953**	-	-

จากกลุ่มความสนใจทั้ง 2 กลุ่ม เมื่อพิจารณาค่าความเด่นจะพบว่า กลุ่ม 1 กลุ่มความประทับใจในสถานที่มีค่าความเด่น 59.99% และกลุ่ม 2 กลุ่มการเข้าชมสถานที่มีค่าความเด่น 28.70% แสดงให้เห็นว่าจากบทวิจารณ์ทั้งหมดที่นักท่องเที่ยวเขียนลงบทเว็บไซต์ TripAdvisor นั้น ในหมวด Sights & Landmarks จะเขียนเกี่ยวกับความประทับใจในสถานที่ มากกว่าการเขียนเกี่ยวกับการเข้าชมสถานที่ และเมื่อพิจารณาค่าความแพร่หลายเพื่อดูว่าบทวิจารณ์ที่นักท่องเที่ยวเขียนนั้นมีความรู้สึกไปทางบวกมากแค่ไหน พบว่าเมื่อนักท่องเที่ยวเขียนบทวิจารณ์เกี่ยวกับความประทับใจในสถานที่นั้นจะมีความรู้สึกไปทางบวกที่ 3.11% เมื่อเขียนเกี่ยวกับกลุ่มการเข้าชมสถานที่ กลับมีค่าเป็นลบที่ -5.87% แสดงให้เห็นว่านักท่องเที่ยวเมื่อเขียนบทวิจารณ์เกี่ยวกับความประทับใจในสถานที่มีความรู้สึกไปทางบวก แต่เมื่อเขียนเกี่ยวกับการเข้าชมสถานที่ กลับมีความรู้สึกไปทางลบแทน



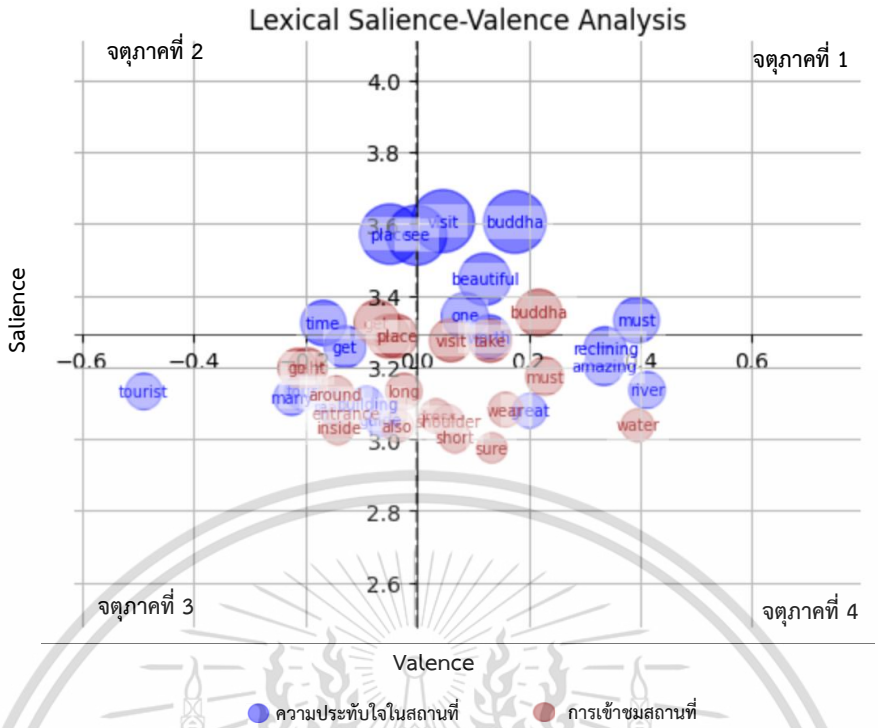
รูปที่ 4.3 แผนภาพการวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม การทดลองที่ 8

4.4.2 ผลการวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ (Lexical Salience-Valence Analysis: LSVA)

จากคำศัพท์ที่ได้มาจากโมเดลการจัดสรรหัวข้อแฝงทั้ง 20 คำ ในการทดลองที่ 8: ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization จะนำมาวิเคราะห์ความเด่นเชิงคำศัพท์ ที่มาจากการนับความถี่ของคำศัพท์ที่ปรากฏในบทวิจารณ์ว่ามีความถี่มากแค่ไหน ยิ่งมีค่ามากแสดงว่าคำศัพท์นั้นเป็นคำที่นักท่องเที่ยวกล่าวถึงมากมีความโดดเด่น และวิเคราะห์ความแพร่หลายเชิงคำศัพท์ เพื่อหาว่าจากคำศัพท์ที่นักท่องเที่ยวกล่าวถึงในบทวิจารณ์นั้นเป็นความรู้สึกเชิงบวกหรือลบมากแค่ไหน ดังตารางที่ 4.7 แสดงค่าความเด่นและความแพร่หลายเชิงคำศัพท์ และนำคำศัพท์ ค่าน้ำหนัก ค่าความเด่น และค่าความแพร่หลาย จากตารางที่ 4.7 ไปสร้างแผนภาพการวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ ดังรูปที่ 4.4

ตารางที่ 4.7 ความเด่นและความแพร่หลายเชิงคำศัพท์

ความประทับใจในสถานที่				การเข้าชมสถานที่			
คำศัพท์	**น้ำหนัก**	**ความเด่น (Salience)**	**ความแพร่หลาย (Valence)**	**คำศัพท์**	**น้ำหนัก**	**ความเด่น (Salience)**	**ความแพร่หลาย (Valence)**
visit	1.00	3.61	0.05	buddha	1.00	3.35	0.22
place	0.94	3.57	-0.05	get	0.94	3.32	-0.07
see	0.87	3.57	0.00	take	0.86	3.27	0.13
buddha	0.81	3.60	0.18	baht	0.79	3.20	-0.20
beautiful	0.67	3.45	0.12	long	0.67	3.13	-0.02
one	0.53	3.34	0.09	go	0.66	3.20	-0.21
time	0.52	3.32	-0.17	see	0.63	3.29	-0.05
must	0.49	3.33	0.39	entrance	0.58	3.07	-0.13
worth	0.45	3.29	0.13	water	0.56	3.04	0.40
amazing	0.40	3.20	0.34	around	0.55	3.13	-0.15
tour	0.36	3.13	-0.20	must	0.55	3.17	0.23
reclining	0.35	3.25	0.34	dress	0.54	3.06	0.03
river	0.34	3.14	0.41	inside	0.54	3.03	-0.14
get	0.33	3.26	-0.13	wear	0.53	3.08	0.16
tourist	0.31	3.13	-0.49	place	0.52	3.29	-0.04
guide	0.31	3.05	-0.06	visit	0.51	3.27	0.06
really	0.31	3.09	-0.15	shoulder	0.48	3.05	0.06
great	0.31	3.08	0.20	short	0.47	3.01	0.07
many	0.31	3.12	-0.22	also	0.47	3.04	-0.04
building	0.30	3.10	-0.09	sure	0.47	2.98	0.13



รูปที่ 4.4 แผนภาพการวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ การทดลองที่ 8

จากรูปที่ 4.4 สีน้ำเงินแทนกลุ่มความประทับใจในสถานที่ สีน้ำตาลแทนการเข้าชมสถานที่ ขนาดของฟองสบู่แทนด้วยค่าน้ำหนัก ในการพิจารณาจะพิจารณาทั้ง 4 จตุภาคโดยจะอ้างอิงตามหลักการวิเคราะห์ของ Taecharungroj and Mathayomchan (2019)

พิจารณาจตุภาคที่ 1 มีความโดดเด่นสูงและมีความแพร่หลายสูงแสดงให้เห็นว่านักท่องเที่ยวมีการกล่าวถึงคำเหล่านั้นสูงและรู้สึกในเชิงบวกเมื่อเขียนถึงคำเหล่านั้น เมื่อเรียงตามค่าน้ำหนักแล้วในกลุ่มความประทับใจในสถานที่ ได้แก่ “visit” “buddha” “beautiful” “one” “must” โดยมีคำว่า “see” ที่ถึงแม้จะมีความโดดเด่นสูงแต่มีความแพร่หลายเป็น 0 แสดงถึงความรู้สึกกลางๆ และในกลุ่มการเข้าชมสถานที่มี 1 คำ ได้แก่ “buddha”

พิจารณาจตุภาคที่ 2 มีความโดดเด่นสูงแต่มีความแพร่หลายต่ำแสดงให้เห็นว่านักท่องเที่ยวมีการกล่าวถึงคำเหล่านั้นสูงและรู้สึกในเชิงลบเมื่อเขียนถึงคำเหล่านั้น ในกลุ่มความประทับใจในสถานที่ ได้แก่ “place” “time” และกลุ่มการเข้าชมสถานที่ได้แก่ “get”

พิจารณาจตุภาคที่ 3 มีความโดดเด่นและความแพร่หลายต่ำแสดงให้เห็นว่ามีการกล่าวถึงไม่มากและมีความรู้สึกเชิงลบต่อคำเหล่านั้น โดยคำศัพท์ในกลุ่มความประทับใจในสถานที่ ได้แก่ “tour” “get” “tourist” “guide” “really” “many” “building” ในกลุ่มการเข้าชมสถานที่ได้แก่ “baht” “long” “go” “see” “entrance” “around” “inside” “place” “also”

พิจารณาจตุภาคที่ 4 มีความโดดเด่นต่ำแต่ความแพร่หลายสูง แสดงให้เห็นว่านักท่องเที่ยวไม่กล่าวถึงนักแต่ก็ยังมีความรู้สึกเชิงบวก ในกลุ่มความประทับใจในสถานที่ ได้แก่ “worth” “amazing” “reclining” “river” “great” ในกลุ่มการเข้าชมสถานที่ได้แก่ “take” “water” “must” “dress” “wear” “visit” “shoulder” “short” “sure”

4.5 โมเดลการจำแนกความรู้สึกนักท่องเที่ยว

หลังจากที่หากลุ่มความสนใจด้วยวิธี LDA แล้วต่อมาจะทำการจำแนกความรู้สึกของนักท่องเที่ยว ตามวัตถุประสงค์ข้อ 2 เมื่อข้อมูลทั้งหมดผ่านกระบวนการเตรียมข้อมูลทั้ง 8 การทดลอง จากบทวิจารณ์ทั้งหมด 14,953 เป็นบทวิจารณ์เชิงบวก 13,263 และบทวิจารณ์เชิงลบ 1,690 เนื่องจากประเภทของข้อมูลที่เป็นบทวิจารณ์เชิงบวกมากกว่าเชิงลบ ซึ่งจะทำแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล 2 วิธี คือ วิธี Undersampling ที่จะทำการลบประเภทของข้อมูลที่เป็นบทวิจารณ์เชิงบวกให้เท่ากับข้อมูลที่เป็นบทวิจารณ์เชิงลบ ดังนั้นบทวิจารณ์เชิงบวกจะถูกลบออกไปให้เท่ากับบทวิจารณ์เชิงลบ จาก 13,263 จะลดลงเท่ากับ 1,690 ทำให้บทวิจารณ์รวมทั้งหมดโดยวิธี Undersampling เท่ากับ 3,380 บทวิจารณ์ และวิธี SMOTE เพื่อทำให้คลาสของข้อมูลที่เป็นบทวิจารณ์เชิงลบเท่ากับข้อมูลที่เป็นบทวิจารณ์เชิงบวก ดังนั้นบทวิจารณ์เชิงลบจะมีจำนวนเพิ่มขึ้นเป็น 13,263 ทำให้บทวิจารณ์รวมทั้งหมดโดยวิธี SMOTE เท่ากับ 26,526 บทวิจารณ์ จากนั้นข้อมูลจะถูกแบ่งออกเป็น 3 ส่วน คือ ข้อมูลสำหรับฝึกสอน (Training set) 70%, ข้อมูลสำหรับตรวจสอบความถูกต้อง (Validation set) 15% และข้อมูลสำหรับทดสอบ (Testing set) 15 % ดังตารางที่ 4.8

ตารางที่ 4.8 สรุปการแบ่งข้อมูลบทวิจารณ์

วิธีแก้ปัญหาความไม่สมดุลของชุดข้อมูล	บทวิจารณ์ทั้งหมดหลังจากแก้ปัญหาความไม่สมดุลของชุดข้อมูล	แบ่งข้อมูลบทวิจารณ์		
		Training set (70%)	Validation set (15%)	Testing set (15%)
Undersampling	3,380	2,365	508	507
SMOTE	26,526	18,567	3,980	3,979

นำข้อมูลฝึกสอน (Training) มาเพื่อหาค่า Hyperparameter ที่เหมาะสม ด้วย Grid Search เมื่อได้แล้วนำมาฝึกสอนโมเดลด้วยชุดข้อมูลฝึกสอน (Training) และใช้ชุดข้อมูลตรวจสอบความถูกต้อง (Validation set) ประเมินประสิทธิภาพว่าโมเดลฝึกสอนมีประสิทธิภาพหรือไม่ และสุดท้ายจะเปรียบเทียบประสิทธิภาพของโมเดลด้วยชุดข้อมูลทดสอบ (Testing set) หลังจากการฝึกเสร็จสมบูรณ์ จากหัวข้อที่ 4.5.1 – 4.5.4 จะแสดงผลการจำแนกความรู้สึกของโมเดล A, B, C และ D ตามการแก้ปัญหาความไม่สมดุลของชุดข้อมูลและการเตรียมข้อมูลทั้ง 8 การทดลอง ดังขั้นตอนที่ 5 ในหัวข้อ 3.6.2 คุณลักษณะของโมเดล โดยผลลัพธ์ที่ได้ผ่านการหาพารามิเตอร์ที่เหมาะสมด้วย Grid Search แล้ว และวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Testing set)

4.5.1 การสร้างโมเดล A (Undersampling & BiLSTM)

ผลการจำแนกความรู้สึกด้วย BiLSTM แก้ปัญหาความไม่สมดุลกันของชุดข้อมูลด้วยวิธี Undersampling จำแนกตาม 8 การทดลอง ที่ผ่านการจูนตาม Hyperparameter ที่กำหนดในตารางที่ 4.9 และวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Testing set) ออกมา 3 ค่า คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Precision) และค่าระลึก (Recall) ดังตารางที่ 4.10

ตารางที่ 4.9 ค่า Hyperparameter ที่นำมาจูนของโมเดล A

Hyperparameter	ค่าที่ใช้ของโมเดล A จำแนกตามพารามิเตอร์และค่าที่นำมาใช้ในการจูนโมเดล
Epoch	5
Batch Size	50
Learning Rate	0.001
Optimizer	Adam
LSTM Units (LSTM Layer)	**[176, 216, 256, 296, 336]**
Dropout Rate	0.2
Fully Connected Units	**[88, 128, 168]**
Output Activation Function	Sigmoid

ตารางที่ 4.10 ประสิทธิภาพของโมเดล A (Undersampling & BiLSTM) จำแนกตาม 8 การทดลองที่ผ่านการจูนด้วย Grid Search

การทดลอง	พารามิเตอร์ที่เหมาะสม		Training set	Testing set		
	LSTM Units	Fully Connected Units	Validation Accuracy (%)	Accuracy (%)	Precision (%)	Recall (%)
1	176	88	73.03	72.38	72.48	72.38
2	216	128	72.24	73.76	73.83	73.77
3	336	168	71.06	70.41	71.39	70.44
4	216	128	72.24	73.96	73.99	73.96
5	**336**	**168**	74.40	**74.35**	**74.98**	**74.34**
6	216	128	72.83	73.76	73.81	73.77
7	256	168	69.68	68.44	68.53	68.43
8	256	88	70.66	73.76	73.78	73.77

จากตารางที่ 4.10 สังเกตได้ว่า การทดลองที่ 5 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming) ให้ค่าความถูกต้อง 74.35% ค่าความเที่ยง 74.98% และค่าความไว 74.34% ซึ่งทั้ง 3 ค่า มีค่ามากสุดในแต่ละการทดลองโดยมีพารามิเตอร์ที่เหมาะสม คือ LSTM Units = 336 ในชั้น LSTM Layer และจำนวนโหนดในชั้น Fully Connected = 168

4.5.2 การสร้างโมเดล B (Undersampling & CNN)

ผลการจำแนกความรู้สึกด้วย CNN แก้ปัญหาความไม่สมดุลกันของชุดข้อมูลด้วยวิธี Undersampling จำแนกตาม 8 การทดลอง ที่ผ่านการจูนตาม Hyperparameter ที่กำหนดในตารางที่ 4.11 และวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Testing set) ออกมา 3 ค่า คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Precision) และค่าระลึก (Recall) ดังตารางที่ 4.12

ตารางที่ 4.11 ค่า Hyperparameter ที่นำมาจูนของโมเดล B

Hyperparameter	**ค่าที่ใช้ของโมเดล B จำแนกตามพารามิเตอร์และค่าที่นำมาใช้ในการจูนโมเดล**
Epochs	20
Batch Size	64
Learning Rate	0.001
Optimizer	Adam
Filter	**[16, 32, 48, 64]**
Kernel Size	**[3, 5, 7, 9]**
Dropout Rate	0.5
Fully Connected Units	10
Output Activation	Sigmoid

ตารางที่ 4.12 ประสิทธิภาพของโมเดล B (Undersampling & CNN) จำแนกตาม 8 การทดลอง ที่ผ่านการจูนด้วย ด้วย Grid Search

การทดลอง	**พารามิเตอร์ที่เหมาะสม**		**Training**	**Testing**		
	Filter	**Kernel Size**	**Accuracy (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
1	48	5	74.40	73.96	74.11	73.96
2	16	3	74.21	76.52	77.39	76.55
3	48	5	77.55	73.57	73.58	73.57
4	32	5	74.16	74.16	74.25	74.16
5	16	3	78.54	76.13	76.14	76.14
6	**32**	**5**	73.42	**77.90**	**77.92**	**77.91**
7	16	5	75.39	76.13	76.43	76.12
8	16	5	73.42	72.58	72.69	72.59

จากตารางที่ 4.12 สังเกตได้ว่า การทดลองที่ 6 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization) ให้ค่าความถูกต้อง 77.90% ค่าความเที่ยง 77.92% และค่าความไว 77.91% ซึ่งทั้ง 3 ค่า มีค่ามากสุดในแต่ละการทดลอง โดยมีพารามิเตอร์ที่เหมาะสม คือ จำนวน Filter = 32 และมีขนาด Kernel Size = 5

4.5.3 การสร้างโมเดล C (SMOTE & BiLSTM)

ผลการจำแนกความรู้สึกด้วย BiLSTM แก้ปัญหาความไม่สมดุลกันของชุดข้อมูลด้วยวิธี SMOTE จำแนกตาม 8 การทดลอง ที่ผ่านการจูนตาม Hyperparameter ที่กำหนดในตารางที่ 4.13 และวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Testing set) ออกมา 3 ค่า คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Precision) และค่าระลึก (Recall) ดังตารางที่ 4.14

ตารางที่ 4.13 ค่า Hyperparameter ที่นำมาจูนของโมเดล C

Hyperparameter	**ค่าที่ใช้ของโมเดล C จำแนกตามพารามิเตอร์และค่าที่นำมาใช้ในการจูนโมเดล**
Epoch	5
Batch Size	50
Learning Rate	0.001
Optimizer	Adam
LSTM Units (LSTM Layer)	**[18, 24, 30]**
Dropout Rate	0.5
Fully Connected Units	**[6, 8]**
Output Activation Function	Sigmoid

ตารางที่ 4.14 ประสิทธิภาพของโมเดล C (SMOTE & BiLSTM) จำแนกตาม 8 การทดลอง ที่ผ่านการจูนด้วย Grid Search

การทดลอง	**พารามิเตอร์ที่เหมาะสม**		**Training**	**Testing**		
	LSTM Units	**Fully Connected Units**	**Validation Accuracy (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**
1	24	8	92.96	92.71	93.48	92.71
2	**30**	**8**	93.39	**93.84**	**94.19**	**93.84**
3	18	8	92.78	93.13	93.20	93.14
4	30	6	93.81	93.28	93.36	93.29
5	30	6	94.47	93.79	94.01	93.79
6	18	6	93.66	93.23	93.78	93.24
7	18	6	93.31	92.83	92.94	92.84
8	18	6	93.56	93.13	93.24	93.14

จากตารางที่ 4.14 สังเกตได้ว่า การทดลองที่ 2 (ไม่ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization) ให้ค่าความถูกต้อง 93.84% ค่าความเที่ยง 94.19% และค่าความไว 93.84% ซึ่งทั้ง 3 ค่า มีค่ามากสุดในแต่ละการทดลองโดยมี

พารามิเตอร์ที่เหมาะสม คือ LSTM Units = 30 ในชั้น LSTM Layer และจำนวนโหนดในชั้น Fully Connected = 8

4.5.4 การสร้างโมเดล D (SMOTE & CNN)

ผลการจำแนกความรู้สึกด้วย CNN แก้ปัญหาความไม่สมดุลกันของชุดข้อมูลด้วยวิธี SMOTE จำแนกตาม 8 การทดลอง ที่ผ่านการจูนตาม Hyperparameter ที่กำหนดในตารางที่ 4.15 และวัดประสิทธิภาพด้วยชุดข้อมูลทดสอบ (Testing set) ออกมา 3 ค่า คือ ค่าความถูกต้อง (Accuracy) ค่าความไว (Precision) และค่าระลึก (Recall) ดังตารางที่ 4.16

ตารางที่ 4.15 ค่า Hyperparameter ที่นำมาจูนของโมเดล D

Hyperparameter	ค่าที่ใช้ของโมเดล D จำแนกตามพารามิเตอร์และค่าที่นำมาใช้ในการจูนโมเดล
Epochs	20
Batch Size	64
Learning Rate	0.001
Optimizer	Adam
Filter	**[8, 16, 24]**
Kernel Size	**[3, 7]**
Dropout Rate	0.5
Fully Connected Units	10
Output Activation	Sigmoid

ตารางที่ 4.16 ประสิทธิภาพของโมเดล D (SMOTE & CNN) จำแนกตาม 8 การทดลอง ที่ผ่านการจูนด้วย Grid Search

การทดลอง	พารามิเตอร์ที่เหมาะสม		Training	Testing		
	Filter	Kernel Size	Accuracy (%)	Accuracy (%)	Precision (%)	Recall (%)
1	8	3	92.42	94.24	94.46	94.24
2	24	3	94.82	94.19	94.35	94.20
3	16	3	93.19	93.26	93.35	93.27
4	8	7	93.91	92.83	93.08	92.84
5	**8**	**7**	93.49	**94.94**	**95.08**	**94.95**
6	8	3	93.66	93.61	93.65	93.62
7	16	7	93.37	92.81	92.92	92.81
8	24	7	93.46	93.39	93.48	93.39

จากตารางที่ 4.16 สังเกตได้ว่า การทดลองที่ 5 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming) ให้ค่าความถูกต้อง 94.94% ค่าความเที่ยง 95.08% และค่าความไว 94.95% ซึ่งทั้ง 3 ค่ามีค่ามากสุดในแต่ละการทดลอง โดยมีพารามิเตอร์ที่เหมาะสมคือ จำนวน Filter = 8 และมีขนาด Kernel Size = 7

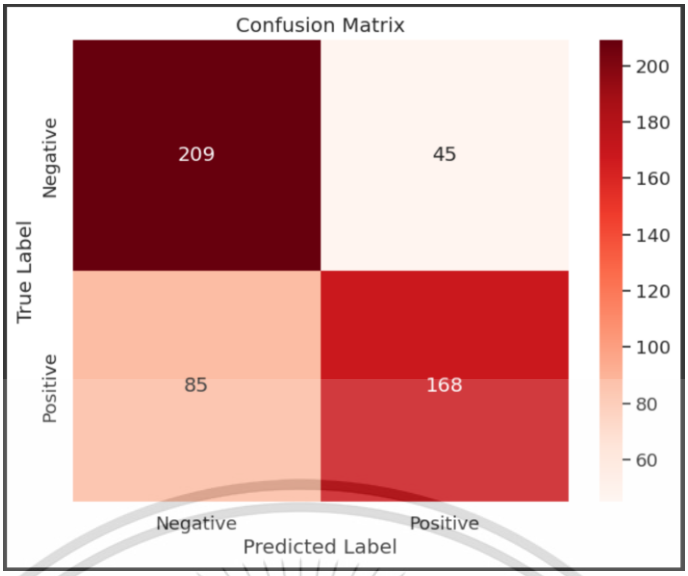
4.5.5 เปรียบเทียบโมเดล

จากหัวข้อ 4.5.1 – 4.5.4 การสร้างโมเดล A – D มีการจำแนกตามเตรียมข้อมูลทั้ง 8 การทดลองและทำการเลือกโมเดลที่มีประสิทธิภาพสูงที่สุดตามการทดลองการเตรียมข้อมูลทั้ง 8 การทดลอง จากชุดข้อมูลทดสอบ (Testing) มาทำการเปรียบเทียบกับโมเดลทั้ง 4 โมเดล

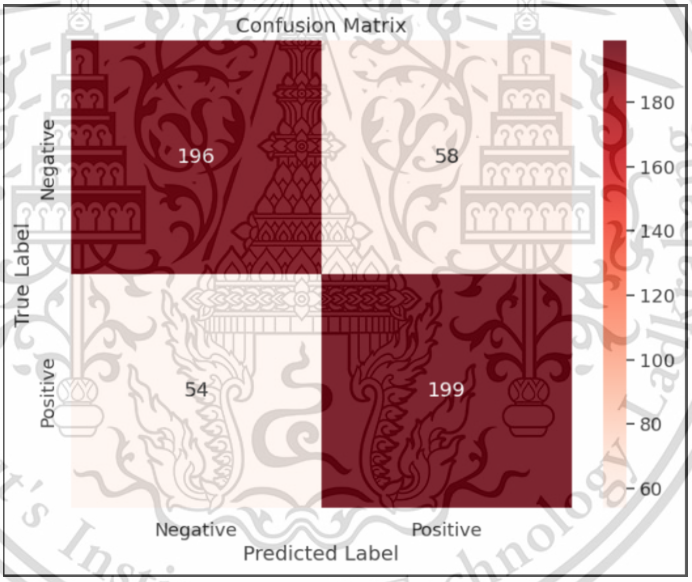
ตารางที่ 4.17 เปรียบเทียบโมเดลกับการทดลองการเตรียมข้อมูลที่ดีที่สุด

โมเดล	การทดลอง	Testing		
		Accuracy (%)	Precision (%)	Recall (%)
A	5	74.35	74.98	74.34
B	6	77.90	77.92	77.91
C	2	93.84	94.19	93.84
D	**5**	**94.94**	**95.08**	**94.95**

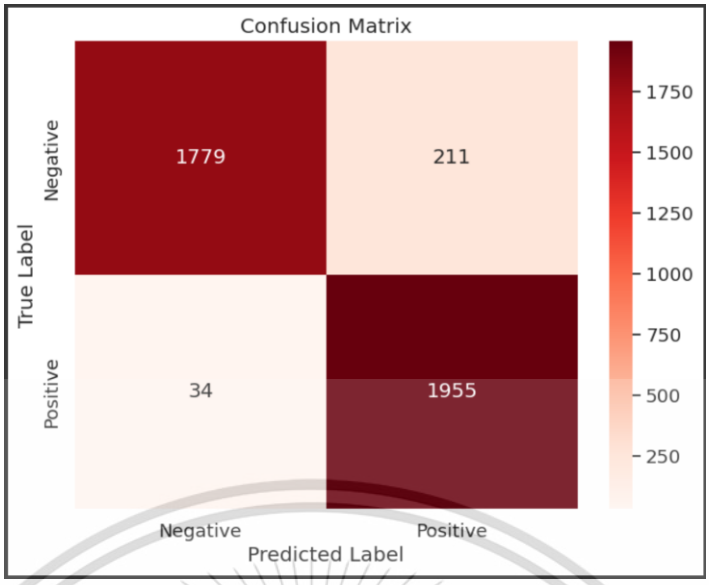
จากตารางที่ 4.17 จะเห็นว่าโมเดล D (SMOTE & CNN) ร่วมกับการเตรียมข้อมูล การทดลองที่ 5 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming) มีประสิทธิภาพสูงที่สุดจากทุกโมเดล มีค่าความถูกต้อง 94.94% ค่าความเที่ยง 95.08% และค่าความไว 94.95% จากตารางเปรียบเทียบจะเห็นว่า เมื่อเปรียบเทียบวิธีการแก้ปัญหาความไม่สมดุลของชุดข้อมูล วิธีการ SMOTE จะประสิทธิภาพที่ดีกว่า Undersampling และจะเห็นว่าการลบคำเกี่ยวข้องกับชื่อสถานที่ออกนั้นมีผลทำให้โมเดลมีประสิทธิภาพที่สูงขึ้น



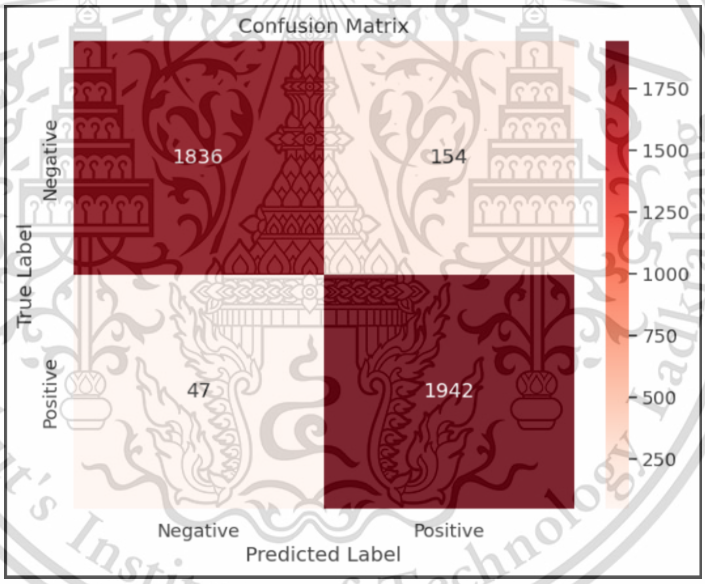
รูปที่ 4.5 Confusion Matrix ของโมเดล A ร่วมกับ การทดลองที่ 5



รูปที่ 4.6 Confusion Matrix ของโมเดล B ร่วมกับ การทดลอง 6



รูปที่ 4.7 Confusion Matrix ของโมเดล C ร่วมกับ การทดลองที่ 2



รูปที่ 4.8 Confusion Matrix ของโมเดล D ร่วมกับ การทดลองที่ 5

4.6 อภิปรายผล

4.6.1 อภิปรายผลการแบ่งกลุ่มความสนใจ ผลการวิเคราะห์ความนิยม และผลการวิเคราะห์ความเด่นและความแพร่หลาย

1) ผลการแบ่งกลุ่มความสนใจของนักท่องเที่ยวด้วย Mean Silhouette Coefficient จากการทดลองที่ 5 – 8 สามารถแบ่งกลุ่มความสนใจได้ 2 กลุ่มความสนใจเท่ากันทั้งหมด และลดลงเมื่อกลุ่มความสนใจเพิ่มขึ้น แสดงให้เห็นว่าอาจมีคำศัพท์ที่คล้ายคลึงกันในแต่ละบทวิจารณ์ เช่น คำศัพท์ในบทวิจารณ์ที่อยู่ในกลุ่ม 1 อาจกระจายในกลุ่มอื่นๆ ด้วย ทำให้เมื่อเพิ่มกลุ่มความสนใจทำให้ค่า Mean Silhouette Coefficient ลดลงเรื่อยๆ

2) ผลการวิเคราะห์ความนิยม แสดงให้เห็นว่า คะแนนเรตติงเฉลี่ยที่สูงขึ้นในกลุ่มความประทับใจในสถานที่อาจบ่งบอกว่าปัจจัยที่เกี่ยวข้องกับความงดงามและเอกลักษณ์ของสถานที่นั้นๆ มีความสำคัญมากในการสร้างความพึงพอใจให้แก่นักท่องเที่ยวมากกว่าปัจจัยที่เกี่ยวข้องกับการเข้าชมสถานที่ และจากจำนวนบทวิจารณ์ที่มากกว่ากลุ่มการเข้าชมสถานที่ ประมาณ 2 เท่า สะท้อนให้เห็นว่านักท่องเที่ยวมักจะพูดถึงและให้ความสำคัญกับความรู้สึกและประสบการณ์ที่ได้รับจากสถานที่มากกว่าขั้นตอนและกระบวนการของการเข้าชม

ดังนั้นเพื่อรักษาคะแนนเรตติงในกลุ่มความประทับใจในสถานที่ให้สูง ควรให้ความสำคัญกับการพัฒนาสภาพแวดล้อมและความสวยงามของสถานที่ให้คงอยู่และดีขึ้นอย่างต่อเนื่อง และเพื่อให้คะแนนเรตติงในกลุ่มการเข้าชมสถานที่สูงขึ้นควรพิจารณาปรับปรุงและเพิ่มประสิทธิภาพในการจัดการและบริการที่เกี่ยวข้องกับการเข้าชมสถานที่

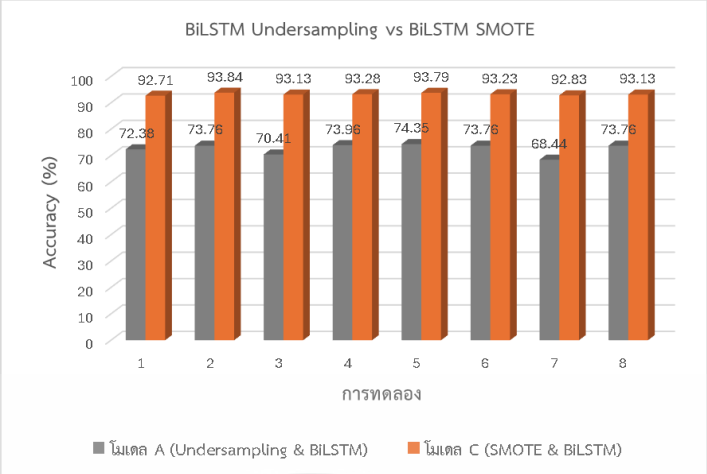
3) ผลการวิเคราะห์ความเด่นและความแพร่หลายทั้งเชิงกลุ่มและเชิงคำศัพท์ แสดงให้เห็นว่านักท่องเที่ยวให้ความสำคัญกับความประทับใจในสถานที่มากกว่าการเข้าชมสถานที่และกล่าวถึงความประทับใจในสถานที่ในแง่บวกโดยความรู้สึกเชิงบวกของนักท่องเที่ยวมักจะเกี่ยวข้องกับประสบการณ์ที่ได้รับจากสถานที่และความงดงามของสถานที่ เช่น คำว่า "beautiful," "worth," และ "amazing" ซึ่งนักท่องเที่ยวมักกล่าวถึงในเชิงบวก

ในทางกลับกัน นักท่องเที่ยวมีแนวโน้มที่จะกล่าวถึงการเข้าชมสถานที่ในแง่ลบ โดยใช้คำที่สื่อถึงประสบการณ์เชิงลบเช่น "get," "take," และ "baht" ซึ่งสะท้อนถึงปัญหาที่นักท่องเที่ยวอาจพบเจอในระหว่างการเข้าชมสถานที่ เช่น ความยากลำบากในการเดินทาง การเสียค่าใช้จ่าย หรือการบริการที่ไม่เป็นไปตามความคาดหวัง

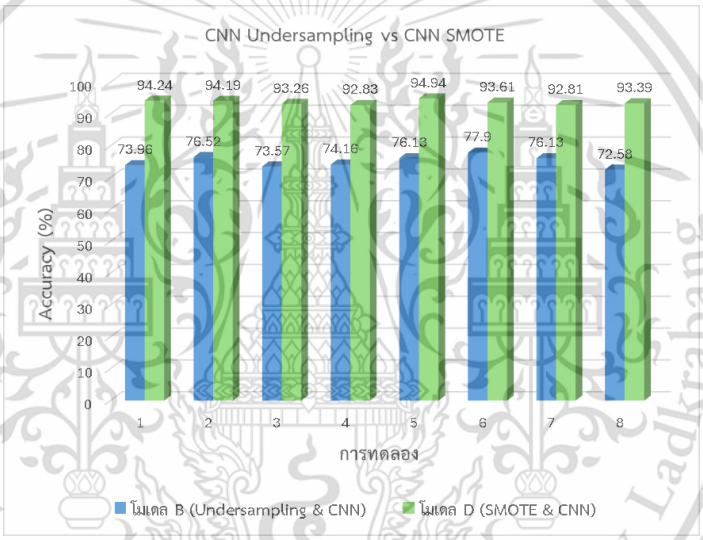
การวิเคราะห์นี้สามารถช่วยให้ผู้จัดการสถานที่ท่องเที่ยวสามารถปรับปรุงบริการและสิ่งอำนวยความสะดวกเพื่อเพิ่มความพึงพอใจและประสบการณ์ที่ดีของนักท่องเที่ยวได้ในอนาคต เช่น กรณีศึกษาการปรับปรุงการเข้าชมสถานที่ หน่วยงานการท่องเที่ยวที่ดูแลสถานที่ท่องเที่ยว เช่น วัดพระเชตุพนฯ สามารถนำข้อมูลไปใช้ในการจัดการคิวเข้าชมวัดและเพิ่มป้ายบอกทางที่ชัดเจน นอกจากนี้สามารถจัดอบรมพนักงานให้มีทักษะในการให้บริการนักท่องเที่ยว ทั้งนี้เพื่อเพิ่มความสะดวกสบายและลดความแออัดในการเข้าชมวัด

4.6.2 อภิปรายผลการจำแนกความรู้สึก

1) เมื่อทำการเปรียบเทียบประสิทธิภาพระหว่างวิธีการ Undersampling กับ SMOTE พบว่าวิธี SMOTE มีประสิทธิภาพที่สูงกว่าวิธี Undersampling จากรูปที่ 4.9 เมื่อเปรียบเทียบค่าความถูกต้องด้วยวิธี Undersampling และ SMOTE ในโมเดล BiLSTM จะเห็นว่า วิธี SMOTE มีค่าความถูกต้องเฉลี่ยที่ 93.24% ซึ่งมากกว่า Undersampling ที่เฉลี่ยอยู่ที่ 72.60% เช่นเดียวกันในโมเดล CNN จากรูปที่ 4.10 วิธี SMOTE มีค่าความถูกต้องเฉลี่ยที่ 93.66% ซึ่งมากกว่า Undersampling ที่เฉลี่ยอยู่ที่ 75.12% สรุปได้ว่าการแก้ปัญหาความไม่สมดุลของชุดข้อมูลวิธี SMOTE ดีกว่า Undersampling

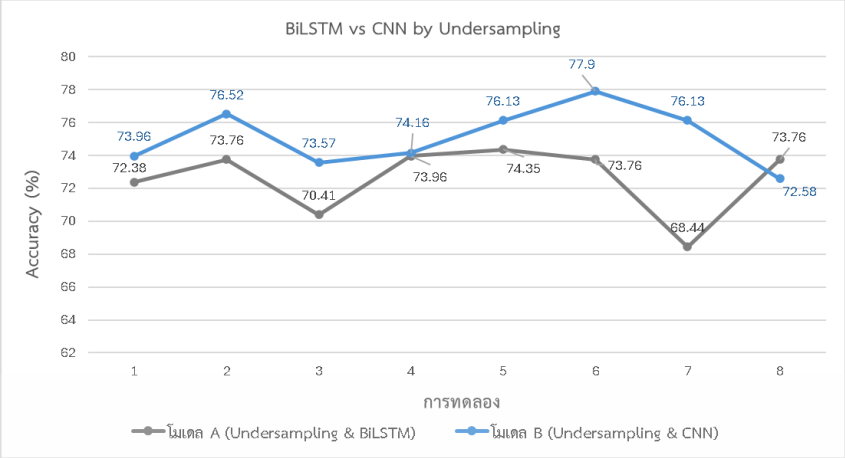


รูปที่ 4.9 เปรียบเทียบประสิทธิภาพของวิธี Undersampling และวิธี SMOTE กับโมเดล BiLSTM

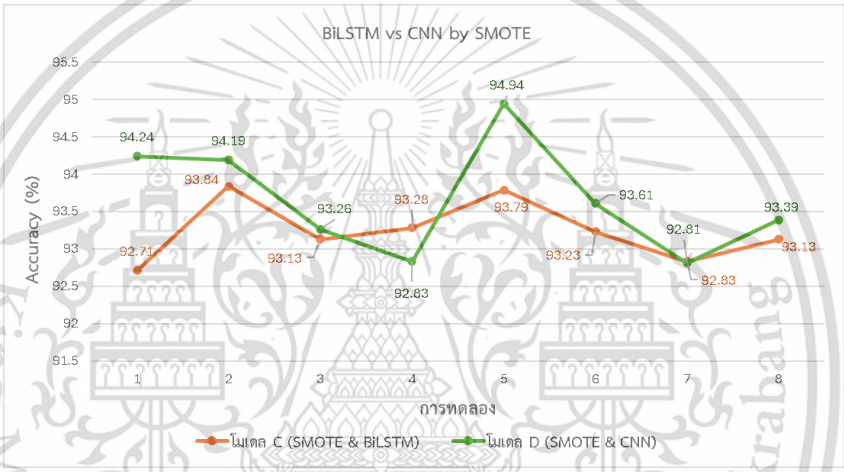


รูปที่ 4.10 เปรียบเทียบประสิทธิภาพของวิธี Undersampling และวิธี SMOTE กับโมเดล CNN

2) เมื่อเปรียบเทียบโมเดล BiLSTM และ CNN ด้วยวิธีการ Undersampling ในรูปที่ 4.11 สังเกตได้ว่า โมเดล CNN มีค่าความถูกต้องที่สูงกว่า BiLSTM ในทุกการทดลองยกเว้น การทดลองที่ 8 ที่ BiLSTM มีค่าความถูกต้องมากกว่า CNN เช่นเดียวกันกับ รูปที่ 4.12 เมื่อเปรียบเทียบโมเดล BiLSTM และ CNN ด้วยวิธีการ SMOTE จะสังเกตได้ว่า CNN มีค่าความถูกต้องสูงกว่า BiLSTM ยกเว้น การทดลองที่ 4 และ 7 แสดงให้เห็นว่าโมเดล CNN โดยรวมแล้วมีประสิทธิภาพที่ดีกว่าโมเดล BiLSTM



รูปที่ 4.11 เปรียบเทียบประสิทธิภาพของโมเดล BiLSTM และ CNN เมื่อใช้วิธี Undersampling



รูปที่ 4.12 เปรียบเทียบประสิทธิภาพของโมเดล BiLSTM และ CNN เมื่อใช้วิธี SMOTE

3) จากตารางที่ 4.17 โมเดลที่มีประสิทธิภาพสูงที่สุดตามการทดลองการเตรียมข้อมูลทั้ง 8 การทดลอง จะเห็นว่าจาก 3 ใน 4 การเตรียมข้อมูลของโมเดลที่มีประสิทธิภาพสูงสุด มีการลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยวออก ส่งผลให้ประสิทธิภาพของโมเดลสูงขึ้น เนื่องจากชื่อสถานที่นั้นจะถูกกล่าวถึงทั้งในบทวิจารณ์ที่เป็นเชิงบวกหรือลบ ดังนั้นการลบคำที่เกี่ยวกับสถานที่ออกไปช่วยเพิ่มความชัดเจนในการเรียนรู้ของโมเดลที่ปราศจากความซับซ้อนของคำที่กระจายทั้งบทวิจารณ์เชิงบวกหรือลบ

4) การลบคำที่มีความถี่ต่ำกว่า 10 คำออกมีแนวโน้มจะทำให้ประสิทธิภาพลดลง แม้ว่าจะช่วยให้การฝึกสอนโมเดลทำได้เร็วขึ้น เมื่อทำการลบคำที่มีความถี่ต่ำออก จะสังเกตว่า คำศัพท์ (Vocabulary) ก่อนลบกับหลังลบมีจำนวนที่แตกต่างกันมาก จากตารางที่ 3.1 คำศัพท์ในการทดลองที่ 6 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization) มีคำศัพท์ 12,865 คำ เมื่อเทียบการเพิ่มการลบคำที่มีความถี่ต่ำออกในการทดลองที่ 8 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization) มีคำศัพท์ 2,821 คำ จะเห็นว่ามีคำศัพท์แตกต่างกันประมาณ 4.5 เท่า ทำให้ไปลดคุณลักษณะที่ใช้ในการเรียนรู้ของโมเดลทำให้ประสิทธิภาพลดลง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

จากข้อมูลบทวิจารณ์ทั้งหมด 14,953 ที่ผ่านการเตรียมข้อมูลทั้งหมด 8 การทดลอง โดยจะใช้ 4 การทดลอง คือ การทดลองที่ 5 - 8 สำหรับการเรียนรู้แบบไม่มีผู้สอนประกอบด้วยการกำหนดจำนวนกลุ่มความสนใจของนักท่องเที่ยวด้วย Mean Silhouette Coefficient เพื่อนำไปใช้ในการจัดกลุ่มความสนใจของนักท่องเที่ยวด้วยโมเดลการจัดสรรหัวข้อแฝง และใช้ 8 การทดลอง สำหรับการเรียนรู้แบบมีผู้สอน ด้วยอัลอัลกอรึทึมหน่วยความจำระยะยาว-ระยะสั้นชนิด 2 ทาง ที่ (Bidirectional Long Short-Term Memory: BiLSTM) และอัลกอริทึมโครงข่ายประสาทแบบคอนโวลูชั่น (Convolution Neural Network: CNN) มีการใช้เทคนิค Undersampling และ SMOTE เพื่อทำให้ประเภทของข้อมูลบทวิจารณ์เชิงบวกเท่ากับบทวิจารณ์เชิงลบ พร้อมกับมีการปรับจูนไฮเปอร์พารามิเตอร์ด้วย Grid Search สามารถสรุปผลการวิเคราะเชิงลึกได้ดังนี้

ส่วนที่ 1 จากข้อมูลบทวิจารณ์ที่ผ่านการเตรียมข้อมูลทั้ง 4 รูปแบบ ประกอบด้วย การทดลองที่ 5, 6, 7, 8 เมื่อนำมากำหนดจำนวนกลุ่มความสนใจที่เหมาะสมด้วย Mean Silhouette Coefficient พบว่า ทั้ง 4 การทดลอง สามารถแบ่งได้ 2 กลุ่มความสนใจ เท่ากันทั้งหมด เมื่อพิจารณาจากน้ำหนักเฉลี่ย 20 คำ จากทั้ง 2 กลุ่มความสนใจ พบว่า การทดลองที่ 8 มีค่ามากที่สุด โดยเมื่อพิจารณาจากคำศัพท์ที่เกิดขึ้นทั้ง 2 กลุ่มความสนใจ สามารถตีความกลุ่มที่ 1 เกี่ยวกับความประทับใจในสถานที่ กลุ่มที่ 2 เกี่ยวกับการเข้าชมสถานที่ ซึ่งการเตรียมข้อมูลตามการทดลองที่ 8 จะนำมาใช้วิเคราะห์ในส่วนที่ 2 ถัดไป

ส่วนที่ 2 การวิเคราะห์เพื่อทำความเข้าใจความรู้สึกเชิงลึกของนักท่องเที่ยว ประกอบด้วยการวิเคราะห์ความนิยม และการวิเคราะห์ความเด่นและความแพร่หลาย

2.1 ผลการวิเคราะห์ความนิยมของนักท่องเที่ยวต่อกลุ่มความสนใจ จากการหาคะแนนเรตติงเฉลี่ยจากบทวิจารณ์ในกลุ่มความสนใจ ทั้ง 2 กลุ่ม จากคะแนนเต็ม 5 พบว่านักท่องเที่ยวมีความสนใจด้านความประทับใจในสถานที่ซึ่งมีคะแนนเรตติง 4.55 คะแนน มากกว่าความสนใจด้านการเข้าชมสถานที่ซึ่งมีคะแนนเรตติงที่ 4.26 คะแนน ต่างกัน 0.29 คะแนน

2.2 ผลการวิเคราะห์ความเด่นและความแพร่หลายเชิงกลุ่ม (Dimensional Salience Valence Analysis: DSVA) พบว่าสถานที่ท่องเที่ยวในหมวด Sights & Landmarks มีความโดดเด่น 2 ด้าน คือ ด้านความประทับใจในสถานที่ และมีความรู้สึกไปในเชิงบวกแสดงให้เห็นว่าจุดเด่น ด้านที่ดีของสถานที่ซึ่งสามารถชูเป็นจุดเด่นในการประชาสัมพันธ์การท่องเที่ยวได้ และด้านการเข้าชมสถานที่แต่โดดเด่นในความรู้สึกเชิงลบแสดงให้เห็นจุดด้อยในหมวด Sights & Landmarks

2.3 ผลการวิเคราะห์ความเด่นและความแพร่หลายเชิงคำศัพท์ (Lexical Salience-Valence Analysis: LSVA) พบว่าสถานที่ท่องเที่ยวในหมวด Sights & Landmarks ในกลุ่มความประทับใจในสถานที่ เมื่อพิจารณาเกี่ยวกับคำที่มีความรู้สึกเชิงบวก คำที่มีความโดดเด่นสูง ได้แก่ “visit” “buddha” “beautiful” “one” “must” สื่อให้เห็นถึงความสวยงามที่ควรค่าแก่การเยี่ยม

ชมซักครั้ง และคำที่มีความโดดเด่นต่ำ ได้แก่ “worth” “amazing” “reclining” “river” “great” สื่อให้เห็นถึง ความคุ้มค่า บรรยากาศริมน้ำ เมื่อพิจารณาเกี่ยวกับคำที่มีความรู้สึกเชิงลบ คำที่มีความโดดเด่นสูงได้แก่ “place” “time” แสดงให้เห็นถึงการใช้เวลาหรือกิจกรรมหรือสิ่งที่เกี่ยวเนื่องในการท่องเที่ยวที่มีเวลาเข้ามาเกี่ยวข้องเป็นสิ่งที่นักท่องเที่ยวไม่พึงพอใจ และคำที่มีความโดดเด่นต่ำได้แก่ “tour” “get” “tourist” “guide” “really” “many” “building” สื่อให้เห็นถึง จำนวนนักท่องเที่ยวในสถานที่นั้นๆ มากเกินไป หรือมัคคุเทศก์ที่บรรยายเกี่ยวกับสถานที่หรือสิ่งปลูกสร้างยังไม่ดีพอ

ในกลุ่มการเข้าชมสถานที่ เมื่อพิจารณาคำที่มีความรู้สึกเชิงบวก คำที่มีความโดดเด่นสูงได้แก่ “buddha” เกี่ยวกับศาสนา พระพุทธเจ้า และคำที่มีความโดดเด่นต่ำได้แก่ “take” “water” “must” “dress” “wear” “visit” “shoulder” “short” “sure” สื่อให้เห็นถึงนักท่องเที่ยวชื่นชอบการแต่งกายให้เข้ากับสถานที่ หรือการแต่งชุดไทย เมื่อพิจารณาเกี่ยวกับคำที่มีความรู้สึกเชิงลบ คำที่มีความโดดเด่นสูงได้แก่ “get” และคำที่มีความโดดเด่นต่ำได้แก่ “baht” “long” “go” “see” “entrance” “around” “inside” “place” “also” แสดงให้เห็นถึง ความไม่พึงพอในการการเข้าสถานที่ แถวที่ยาวหรือเกี่ยวกับค่าใช้จ่ายในการเข้าสถานที่

ส่วนที่ 3 การจำแนกความรู้สึกของโมเดล A, B, C และ D ตามการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล 2 วิธี และการเตรียมข้อมูลทั้ง 8 การทดลอง จากนั้นทำการแบ่งข้อมูลออกเป็น 3 ส่วน ได้แก่ ส่วนที่ใช้ฝึกสอน (Training) 70% สำหรับตรวจสอบความถูกต้อง (Validation) 15% และข้อมูลสำหรับทดสอบ (Testing) 15% และมีใช้ Grid Search เพื่อหาพารามิเตอร์ที่เหมาะสมในแต่ละอัลกอรึทึมครบทั้ง 8 การทดลองโดยการเปรียบเทียบประสิทธิภาพของโมเดล ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความระลึก (Precision) ค่าความแม่นยำ (Recall) ซึ่งจากการทดสอบด้วยชุดข้อมูลทดสอบ สามารถสรุปได้ดังนี้

1) วิธีการแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล ด้วยเทคนิค SMOTE มีประสิทธิภาพที่สูงกว่า Undersampling ทั้งโมเดล BiLSTM และ CNN

2) โมเดล CNN มีประสิทธิภาพที่สูงกว่า BiLSTM ไม่ว่าจะเป็นการใช้เทคนิค Undersampling หรือ SMOTE

3) การลบคำเกี่ยวกับชื่อสถานที่ท่องเที่ยวออกช่วยเพิ่มประสิทธิภาพของโมเดลให้สูงขึ้น และการลบคำที่มีความถี่ต่ำกว่า 10 คำออก มีแนวโน้มทำให้ประสิทธิภาพลดลง

4) การทดลองการเตรียมข้อมูลที่ดีที่สุดเมื่อใช้โมเดล A (Undersampling + BiLSTM) คือ การทดลองที่ 5 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming) มีค่าความถูกต้องเท่ากับ 74.35% ค่าความระลึกเท่ากับ 74.98% ค่าความแม่นยำเท่ากับ 74.35%

5) การทดลองการเตรียมข้อมูลที่ดีที่สุดเมื่อใช้โมเดล B (Undersampling & CNN) คือ การทดลองที่ 6 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Lemmatization) มีค่าความถูกต้องเท่ากับ 77.90% ค่าความระลึกเท่ากับ 77.92% ค่าความแม่นยำเท่ากับ 77.91%

6) การทดลองการเตรียมข้อมูลดีที่สุด เมื่อใช้โมเดล C (SMOTE & BiLSTM) คือ การทดลองที่ 2 (ไม่ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี

Lemmatization) มีค่าความถูกต้องเท่ากับ 93.84% ค่าความระลึกเท่ากับ 94.19% ค่าความแม่นยำเท่ากับ 93.84%

7) การทดลองการเตรียมข้อมูลที่ดีที่สุด เมื่อใช้โมเดล D (SMOTE & CNN) คือ การทดลองที่ 5 (ลบคำที่เกี่ยวกับชื่อสถานที่ท่องเที่ยว ไม่ลบคำที่มีความถี่ต่ำกว่า 10 คำ และแปลงคำด้วยวิธี Stemming) มีค่าความถูกต้องเท่ากับ 94.94% ค่าความระลึกเท่ากับ 95.08% ค่าความแม่นยำเท่ากับ 94.95% ซึ่งเป็นโมเดลที่ดีที่สุดจากทั้ง 4 โมเดล

5.2 ข้อเสนอแนะ

1) ทดลองครั้งต่อไปอาจพิจารณาลบคำเพิ่มเติมเพื่อให้เกิดความชัดเจนในแต่ละกลุ่มความสนใจ เนื่องจากในการทดลองลบคำที่เกี่ยวกับชื่อสถานที่ออกนั้น มีเพียง 13 คำ สามารถทดลองลบคำอื่นๆ เพิ่มเติมได้ หรือลบคำที่มีการแสดงผลทั้ง 2 กลุ่มความสนใจ

2) การตั้งชื่อกลุ่มความสนใจอาศัยตีความและพิจารณาหลายๆ มุมมอง ดังนั้นการทำความเข้าใจในเนื้อหาที่โมเดลดึงออกมาเป็นสิ่งสำคัญที่จะช่วยให้สามารถตั้งชื่อหัวข้อได้อย่างถูกต้องและเหมาะสม หรือถ้าผู้ที่สนใจศึกษาต่อมีความรู้ทางด้านภาษาศาสตร์ หรืองานวิจัยอื่นๆ อ้างอิง ที่สามารถพิจารณาคำที่ได้จากโมเดลว่าควรตั้งชื่ออย่างไร เป็นอีกหนึ่งแนวทางที่น่าสนใจ

5.3 ข้อจำกัดของงานวิจัย

1) การตั้งชื่อกลุ่มความสนใจนั้นมาจากมุมมองของผู้วิจัยซึ่งไม่ได้มีความเชี่ยวชาญหรือทำงานเกี่ยวกับด้านการท่องเที่ยว ซึ่งผู้อ่านหรือผู้เชี่ยวชาญในด้านการท่องเที่ยวอาจมีมุมมองที่แตกต่างกัน

2) เนื่องจากการเก็บข้อมูลนั้นเก็บมาแหล่งที่มาเดียวคือ TripAdvisor ดังนั้น การนำโมเดลไปใช้งานกับข้อมูลในเว็บไซต์อื่นที่ผู้เขียนบทวิจารณ์เป็นกลุ่มที่แตกต่างจากกลุ่มผู้เขียนบทวิจารณ์ใน TripAdvisor อาจให้ผลไม่ถูกต้องได้ เช่น ความต้องการของนักท่องเที่ยวแบบประหยัดนั้นมีความแตกต่างกันกับความต้องการของนักท่องเที่ยวใน TripAdvisor

3) งานวิจัยนี้รันโมเดลด้วย Google Colab จากจำนวนบทวิจารณ์กว่า 26,000 บทวิจารณ์ หลังจากใช้ SMOTE เพื่อแก้ปัญหาความไม่สมดุลกันของชุดข้อมูล และมีการเตรียมข้อมูลหลายการทดลอง ทำให้มีข้อจำกัดในเรื่องค่าใช้จ่ายในการรันโมเดลทำให้การตั้งค่า Search Space ใน Grid Search สามารถตั้งได้จำกัด หากผู้ที่สนใจ มี GPU ที่มีประสิทธิภาพสูงสามารถเพิ่มช่วงของ Search Space หรือตั้งค่าเพื่อหา Hyperparameter ตัวอื่นๆ เพิ่มได้ เช่น Learning Rate หรือ Batch Size

เอกสารอ้างอิง

การท่องเที่ยวแห่งประเทศไทย ศูนย์วิจัยด้านการตลาดการท่องเที่ยว. (2566). *สถานการณ์การท่องเที่ยวรายจังหวัด ปี 2566*. เรียกใช้เมื่อ 3 พฤศจิกายน 2566 จาก https://intelligencecenter.tat.or.th/articles/22978

จิระเมศร์ รุจิกรหิรัณย์. (2565). *การพัฒนาโมเดลการวิเคราะห์อารมณ์ครูในชั้นเรียนผ่านการรู้จำคำพูดโดยใช้การเรียนรู้เชิงลึก.* วิทยานิพนธ์ครุศาสตรมหาบัณฑิต, สาขาวิชาวิธีวิทยาการพัฒนานวัตกรรมทางการศึกษา, คณะครุศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย.

ธนาคารแห่งประเทศไทย. (2566). *การท่องเที่ยวไทยฟื้นตัวอย่างไรในเชิงพื้นที่*. เรียกใช้เมื่อ 15 พฤศจิกายน 2566 จาก https://www.bot.or.th/th/research-and-publications/articles-and-publications/articles/chaengsibia/article-2023jun13.html

นิธิกร เลิศชาญวุฒิ. (2564). *การวิเคราะห์ความสนใจของนักท่องเที่ยวด้วยเทคนิคการเรียนรู้ของเครื่อง : กรณีศึกษา ถนนเยาวราช ประเทศไทย.* วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต, สาขาวิชาสถิติและการิเคราะห์ธุรกิจ, คณะวิทยาศาสตร์, สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง.

ภูมิรพี ภูมิค้า. (2562). *เทคนิคการเรียนรู้เชิงลึกเพื่อวิเคราะห์ความรู้สึกจากผู้ใช้ผลิตภัณฑ์.* วิทยานิพนธ์วิศวกรรมศาสตรมหาบัณฑิต, สาขาวิชาวิศวกรรมโทรคมนาคมและคอมพิวเตอร์, มหาวิทยาลัยเทคโนโลยีสุรนารี.

วิทยา ปัญญา, และ วฤษาย์ ร่มสายหยุด. (2565). วิธีการสร้างแบบจำลองเชิงทำนายพฤติกรรมการผิดเงื่อนไขการปล่อยชั่วคราวของศาล จากชุดข้อมูลที่ไม่สมดุลโดยใช้เทคนิคการเรียนรู้ของเครื่อง. *วารสารวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยมหาสารคาม, 42(2)*, 47-57.

Alharbi, B. A., Mezher, M. A., & Barakeh, A. M. (2022). Tourist Reviews Sentiment Classification using Deep Learning Techniques: A Case Study in Saudi Arabia. *International Journal of Advanced Computer Science and Applications, 13(6)*.

Shetty, A. M., Aljunid, M. F., Manjaiah, D. H., & Afzal, A. M. (2024). Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis. *Conference: Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis*, 451-474.

Panichella, A., Dit, B., Oliveto, R., Penta, M. D., Poshynanyk, D., & Lucia, A. D. (2013). How to effectively use topic models for software engineering tasks? An approach based on Genetic Algorithms. *International Conference on Software Engineering (ICSE) 35th.* San Francisco, CA:18-26 May 2013, 522-531

Bessadeg, F. (2024). *Top 20 Most Visited Cities in the World*. Retrieved March 15, 2024, from Travelness: https://travelness.com/most-visited-cities-in-the-world

Chugh, N., & Phumchusri, N. (2020). Bangkok Tours and Activities Data Analysis via User-Generated Content. *2020 International Conference on Computing, Electronics & Communications Engineering (iCCECE).* Southend, UK, 17-18 August 2020, 98-102.

Manurung, K. A., & Lhaksmana, K. M. (2023). Sentiment Analysis of Tourist Attraction Review from TripAdvisor Using CNN and LSTM. *International Journal on Information and Comunication Technology 9(1)*, 73-85.

Kim, J. (2017). *Understanding how Convolutional Neural Network (CNN) perform text classification with word embeddings*. Retrieved April 3, 2024, from medium: https://towardsdatascience.com/understanding-how-convolutional-neural-network-cnn-perform-text-classification-with-word-d2ee64b9dd0b

Kulkarni, A., Chong, D., & Batarseh, F. A. (2021). *Foundations of Data Imbalance and Solutions for a Data Democracy*. Data Democracy: 1st Edition At the Nexus of Artificial Intelligence.

Lawrence, S. J. (2023). *What is LSTM? - Introduction to Long Short-Term Memory*. Retrieved April 27, 2024, from SCALER Topics: https://www.scaler.com/topics/deep-learning/lstm/

Liu, J., Hu, S., Mehraliyev, F., & Liu, H. (2023). Text classification in tourism and hospitality-a deep learning perspective. *Tourism and hospitality*.

Mittal, A. (2019). *Understanding RNN and LSTM*. Retrieved April 28, 2024, from Medium: https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e

Puh, K., & Babac , M. B. (2022). Predicting sentiment and rating of tourist reviews using machine learning. *Journal of Hospitality and Tourism Insights, 6(3)*, 1188-1204.

Ramadhani, A., Sutoyo, E., & Widartha, V. P. (2021). LSTM-based Deep Learning Architecture of Tourist Review in Tripadvisor. *2021 Sixth International Conference on Informatics and Computing (ICIC)*, Jakarta, Indonesia, 03-04 November 2021, 1-6.

Serrano, L. (2020). *Latent Dirichlet Allocation (Part 1 of 2)*. Retrieved February 22, 2020, from Youtube: https://www.youtube.com/watch?v=T05t-SqKArY&list=FLW3BpFpTi7GihkIBAlnv-lA&index=86

Statista. (2022). *Total number of user reviews and ratings on Tripadvisor worldwide from 2014 to 2023*. Retrieved November 23, 2023, from https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/

Taecharungroj, V., & Mathayomchan, B. (2019). Analysis TripAdvisor reviews of tourist attractions in Phuket, Thailand. *Tourism Management 75(6)*, 550-568.

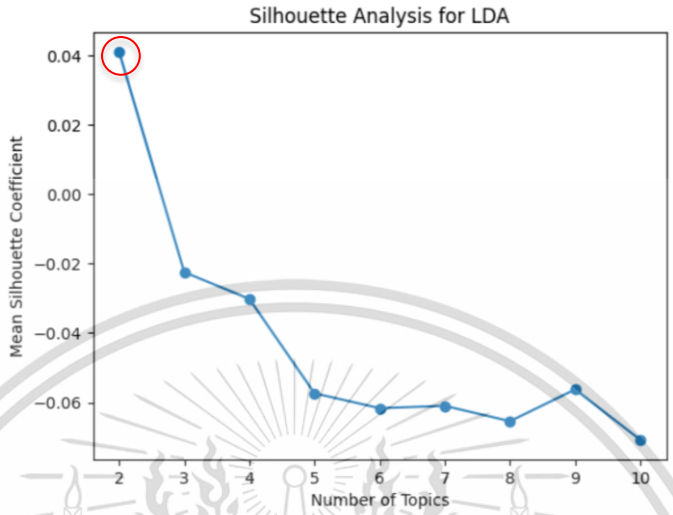
Jiang, Y., Song, X., Harrison, J., Quegan, S., & Maynard, D. (2017). Comparing Attitudes to Climate Change in the Media using sentiment analysis based on Latent Dirichlet Allocation. *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, Copenhagen, Denmark: Association for Computational Linguistics, 25-30.

Zhang, Y., & Wallace, B. C. (2017). A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, Taipei, Taiwan, 27 November – 1 December, 2017, 253-256.

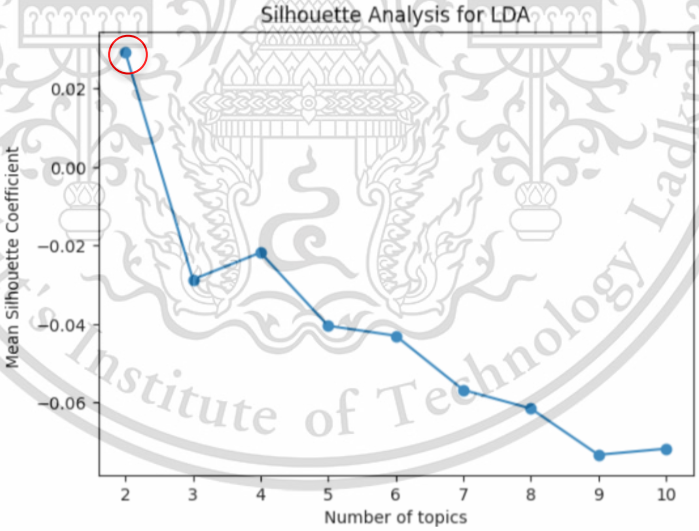
Zhao, Y. (2023). *Complete Guide to RNN, LSTM, and Bidirectional LSTM*. Retrieved April 28, 2024, from dagshub: https://dagshub.com/blog/rnn-lstm-bidirectional-lstm/

ภาคผนวก ก
กราฟความสัมพันธ์ระหว่าง Mean Silhouette Coefficient กับกลุ่มความสนใจ ของการทดลองที่ 5 - 8

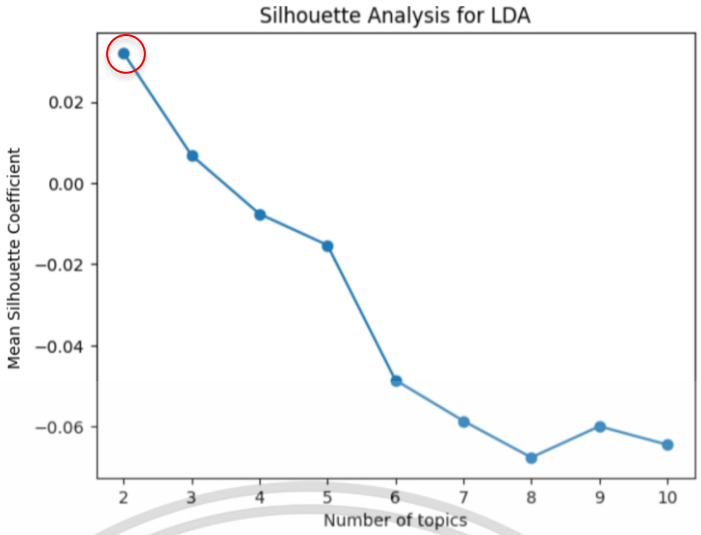
จากรูปที่ ก.1 - ก.4 จะพบว่าค่า Mean Silhouette Coefficient มีค่ามากสุดที่ จำนวนกลุ่มความสนใจ (Number of Topics) เท่ากับ 2 ทั้งหมด ดังนั้น กลุ่มความสนใจที่เหมาะสมของการทดลองที่ 5 – 8 เท่ากับ 2 กลุ่มความสนใจ



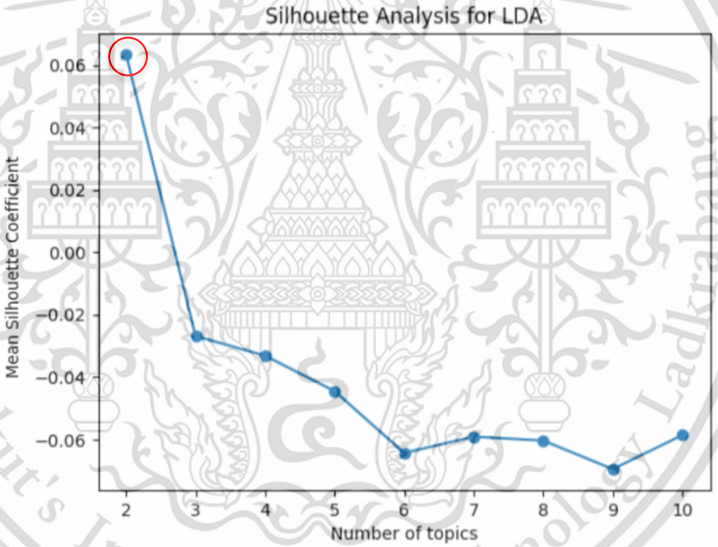
รูปที่ ก.1 ค่า Mean Silhouette Coefficient ต่อกลุ่มความสนใจที่คำนวณจากการเตรียมข้อมูล การทดลองที่ 5



รูปที่ ก.2 ค่า Mean Silhouette Coefficient ต่อกลุ่มความสนใจที่คำนวณจากการเตรียมข้อมูล การทดลองที่ 6



รูปที่ ก.3 ค่า Mean Silhouette Coefficient ต่อกลุ่มความสนใจที่คำนวณจากการเตรียมข้อมูล การทดลองที่ 7



รูปที่ ก.4 ค่า Mean Silhouette Coefficient ต่อกลุ่มความสนใจที่คำนวณจากการเตรียมข้อมูล การทดลองที่ 8

ภาคผนวก ข
คำศัพท์ของกลุ่มความสนใจจากโมเดลการจัดสรรหัวข้อแฝง

ตารางที่ ข.1 คำศัพท์ของกลุ่มความสนใจจากโมเดลการจัดสรรหัวข้อแฝง การทดลองที่ 5

กลุ่มที่ 1		กลุ่มที่ 2	
คำศัพท์	**น้ำหนัก**	**คำศัพท์**	**น้ำหนัก**
buddha	1.00	get	1.00
visit	0.97	go	0.88
see	0.78	visit	0.82
place	0.67	take	0.76
beauti	0.55	baht	0.67
reclin	0.44	river	0.65
must	0.44	cover	0.56
one	0.37	place	0.53
crowd	0.35	boat	0.52
worth	0.32	around	0.52
amaz	0.32	wear	0.50
templ	0.30	see	0.49
time	0.30	tourist	0.47
take	0.27	peopl	0.47
build	0.27	walk	0.47
statu	0.23	entranc	0.46
mani	0.22	time	0.46
get	0.21	dress	0.46
lot	0.21	crowd	0.45
guid	0.21	long	0.45
ค่าเฉลี่ย	0.42	ค่าเฉลี่ย	0.58

ตารางที่ ข.2 คำศัพท์ของกลุ่มความสนใจจากโมเดลการจัดสรรหัวข้อแฝง การทดลองที่ 6

กลุ่มที่ 1		กลุ่มที่ 2	
คำศัพท์	**น้ำหนัก**	**คำศัพท์**	**น้ำหนัก**
get	1.00	buddha	1.00
take	0.79	visit	0.64
place	0.75	see	0.61
visit	0.74	place	0.59
go	0.73	reclining	0.44
see	0.71	beautiful	0.43
baht	0.71	one	0.37
river	0.67	must	0.35
around	0.57	time	0.31
must	0.55	worth	0.29
boat	0.55	amazing	0.27
long	0.53	building	0.25
beautiful	0.53	statue	0.24
tourist	0.50	tour	0.22
entrance	0.50	many	0.21
people	0.50	guide	0.21
dress	0.49	well	0.19
time	0.48	take	0.19
wear	0.48	really	0.19
one	0.46	around	0.19
ค่าเฉลี่ย	0.61	ค่าเฉลี่ย	0.36

ตารางที่ ข.3 คำศัพท์ของกลุ่มความสนใจจากโมเดลการจัดสรรหัวข้อแฝง การทดลองที่ 7

กลุ่มที่ 1		กลุ่มที่ 2	
คำศัพท์	**น้ำหนัก**	**คำศัพท์**	**น้ำหนัก**
get	1.00	buddha	1.00
go	0.93	visit	0.92
visit	0.88	see	0.68
place	0.75	beauti	0.57
crowd	0.75	place	0.54
take	0.74	reclin	0.45
see	0.67	one	0.40
cover	0.61	must	0.35
tour	0.58	worth	0.30
peopl	0.56	amaz	0.30
guid	0.54	templ	0.29
tourist	0.53	river	0.28
dress	0.53	time	0.27
baht	0.51	take	0.25
wear	0.51	statu	0.24
time	0.50	build	0.22
around	0.48	architectur	0.21
must	0.48	around	0.20
long	0.47	mani	0.19
hot	0.42	well	0.19
ค่าเฉลี่ย	0.62	ค่าเฉลี่ย	0.39

ตารางที่ ข.4 คำศัพท์ของกลุ่มความสนใจจากโมเดลการจัดสรรหัวข้อแฝง การทดลองที่ 8

กลุ่มที่ 1		กลุ่มที่ 2	
คำศัพท์	**น้ำหนัก**	**คำศัพท์**	**น้ำหนัก**
visit	1.00	buddha	1.00
place	0.94	get	0.94
see	0.87	take	0.86
buddha	0.81	baht	0.79
beautiful	0.67	long	0.67
one	0.53	go	0.66
time	0.52	see	0.63
must	0.49	entrance	0.58
worth	0.45	water	0.56
amazing	0.40	around	0.55
tour	0.36	must	0.55
reclining	0.35	dress	0.54
river	0.34	inside	0.54
get	0.33	wear	0.53
tourist	0.31	place	0.52
guide	0.31	visit	0.51
really	0.31	shoulder	0.48
great	0.31	short	0.47
many	0.31	also	0.47
building	0.30	sure	0.47
ค่าเฉลี่ย	0.50	ค่าเฉลี่ย	0.62

ภาคผนวก ค
การตั้งค่า Search Space ของโมเดล เพื่อทำ Grid Search

ค.1 Search Space สำหรับโมเดล BiLSTM (โมเดล A และ โมเดล C)

ค.1.1 สำหรับวิธีการ Undersampling & BiLSTM (โมเดล A)

โดยเริ่มจากค่าตั้งต้น 256 ในชั้น LSTM Layer และ 128 ในชั้น Fully Connected ตาม Manurung and Lhaksmana (2023) พบว่ามี Accuracy ที่สูงจึงนำค่าดังกล่าวมาเพื่อหารูปแบบพารามิเตอร์ที่เหมาะสม โดยมีการปรับเพิ่ม/ลด ครั้งละ 40 ทั้ง 2 ชั้น ดังนั้น จึงทำการเซ็ตค่าในช่วง [176, 216, 256, 296, 336] สำหรับชั้น LSTM Layer และ [88, 128, 168] สำหรับชั้น Fully Connected

ค.1.2 สำหรับวิธีการ SMOTE & BiLSTM (โมเดล C)

เมื่อลองใช้ค่าในช่วงเดียวกับ ค.1.1 พบว่า Accuracy ที่ได้ระหว่างการ Train ลดลงเรื่อยๆ ดังนั้น จึงทำการทดลองปรับลดค่าในชั้น LSTM Layer และชั้น Fully Connected ลงเรื่อยๆ จนได้ช่วงที่มีประสิทธิภาพที่สูง คือ 24 ในชั้น LSTM Layer และ 6, 8 ในชั้น Fully Connected จึงนำค่าดังกล่าวมาเพื่อหารูปแบบพารามิเตอร์ที่เหมาะสม โดยมีการปรับเพิ่ม/ลด ครั้งละ 6 สำหรับชั้น LSTM Layer [18, 24, 30] และ [6, 8] สำหรับชั้น Fully Connected

ค.2 Search Space สำหรับโมเดล CNN (โมเดล B และโมเดล D)

ค.2.1 สำหรับวิธีการ Undersampling & CNN (โมเดล B)

ทำการทดลองค่า Filter [16, 32, 48, 64] และ Kernel Size [3, 5, 7, 9] ในช่วงตาม Manurung and Lhaksmana (2023) พบว่าได้ Accuracy ที่สูง จึงนำค่าดังกล่าวมา เพื่อหารูปแบบพารามิเตอร์ที่เหมาะสม โดยมีการปรับเพิ่ม/ลด ครั้งละ 16 สำหรับ Filter และ 2 สำหรับ Kernel Size

ค.2.2 สำหรับวิธีการ SMOTE & CNN (โมเดล D)

ทำการทดลองค่าในช่วงเดียวกับ ค.2.1 แต่พบว่ายิ่งเพิ่มจำนวน Filter และ Kernel Size ทำให้ Accuracy ลดลง จึงปรับลดค่าในช่วง ค.2.1 ลงเป็น [8, 16, 24] สำหรับ Filter และ [3, 7] สำหรับ Kernel Size

ค.3 ข้อจำกัดในการตั้งค่า Search Space

เนื่องจากวิธี SMOTE นั้น ทำให้มีข้อมูลกว่า 26,000 บทวิจารณ์ และจำแนกตาม 8 การทดลอง เนื่องจากมีงบประมาณที่จำกัด (รันโมเดลแบบเสียค่าใช้จ่ายผ่าน Google Colab) ทำให้ตั้งค่าช่วง Search Space ที่น้อย กอปรกับเรื่องเวลาและค่าใช้จ่ายในการ Train จะเพิ่มขึ้นตาม Search Space ที่มากขึ้น ถ้าตัดข้อจำกัดเหล่านี้ออกสามารถที่จะตั้งค่าในช่วงใน Search Space ที่ละเอียดขึ้นได้ เช่น ปรับเพิ่ม/ลด ค่าในแต่ละชั้น ทีละ 2 ก็ได้ [18, 20, 22, ...] ในชั้น LSTM Layer หรือ [8, 10, 12, ...] สำหรับจำนวน Filters

ค.4 จากการทดลองหลายๆ ครั้งทั้ง BiLSTM และ CNN ด้วยวิธี Undersampling และ SMOTE สามารถสรุปได้ดังนี้

ค.4.1 ใช้ค่าเริ่มต้นจาก Manurung and Lhaksmana (2023) จากนั้น ค่อยๆ ปรับค่าเพิ่มขึ้นหรือลดลง เพื่อให้สังเกตเห็นแนวโน้มของประสิทธิภาพของโมเดล

ค.4.2 ขนาดของข้อมูล

Undersampling: เป็นเทคนิคที่ทำให้ขนาดของข้อมูลลดลงโดยการตัดคลาสข้อมูลที่มีมากกว่าให้สมดุลกับคลาสที่มีน้อยกว่า ทำให้ข้อมูลที่ใช้ในการฝึกโมเดลมีขนาดเล็กลง ดังนั้น การตั้งค่าจำนวน Units อาจจะเลือกค่าที่มากขึ้นเพื่อให้โมเดลมีความสามารถในการจับลักษณะของข้อมูลที่มากขึ้น

SMOTE: เป็นเทคนิคที่เพิ่มข้อมูลที่ขาดของคลาสที่มีน้อยกว่าให้สมดุลกับคลาสที่มีมากกว่าโดยการสร้างข้อมูลสังเคราะห์ที่คล้ายกับข้อมูลจริง ทำให้ข้อมูลที่ได้มีขนาดใหญ่ขึ้น ดังนั้น จำนวน Units อาจจะเลือกค่าที่เล็กลงเพื่อให้โมเดลสามารถเรียนรู้จากข้อมูลที่มีความคล้ายคลึงกันได้อย่างมีประสิทธิภาพมากขึ้น

ค.4.3 ลักษณะของข้อมูล

Undersampling: ข้อมูลหลังการทำ Undersampling อาจมีลักษณะที่ชัดเจนและง่ายต่อการจับลักษณะ ดังนั้น การตั้งค่า Units ที่สูงขึ้นอาจช่วยให้โมเดลสามารถเรียนรู้รายละเอียดของข้อมูลได้ดีขึ้น

SMOTE: ข้อมูลที่สร้างขึ้นจาก SMOTE อาจมีความซับซ้อนและมีความคล้ายคลึงกันกับข้อมูลจริง ดังนั้น การตั้งค่า Units ที่น้อยลงอาจช่วยให้โมเดลไม่ Overfit กับข้อมูลที่สร้างขึ้น

ประวัติผู้เขียน

ชื่อ	นายธุวานนท์ ขวัญเกื้อ
วัน เดือน ปีเกิด	5 กันยายน 2538
ที่อยู่ปัจจุบัน	78 บางนาตราด 36 แขวงบางนาใต้ เขตบางนา กรุงเทพฯ 10260
ประวัติการศึกษา	(2561) วิศวกรรมศาสตร์บัณฑิต เกรดเฉลี่ย 3.58
	สาขา วิศวกรรมเครื่องกล
	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง วิทยาเขตชุมพร
	เขตรอุดมศักดิ์ จังหวัดชุมพร