

การจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง

นายธีระชัย เกษมพงศ์จรัส

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์ข้อมูลเพื่อนวัตกรรม
บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2568
ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

การจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง

นายธีระชัย เกษมพงศ์จรัส



การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาหลักสูตร
วิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์ข้อมูลเพื่อนวัตกรรม
บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ปีการศึกษา 2568
ลิขสิทธิ์ของมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ



ใบรับรองการค้นคว้าอิสระ

บัณฑิตวิทยาลัย มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

เรื่อง การจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง

โดย นายธีระชัย เกษมพงศ์จรัส

ได้รับอนุมัติให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาศาสตร์ข้อมูลเพื่อนวัตกรรม

คณบดีคณะเทคโนโลยีสารสนเทศและนวัตกรรมดิจิทัล

.....

(ผู้ช่วยศาสตราจารย์ ดร.สุนันฑา สดสี)

คณะกรรมการสอบการค้นคว้าอิสระ

..... ประธาน

(ดร.ชูชาติ หฤไชยะศักดิ์)

..... อาจารย์ที่ปรึกษา

(ผู้ช่วยศาสตราจารย์ ดร.มาลีรัตน์ มะลิแย้ม)

..... กรรมการ

(ดร.กาญจนา วิริยะพันธ์)

ชื่อ : นายธีระชัย เกษมพงศ์จรัส
ชื่อการค้นคว้าอิสระ : การจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง
สาขาวิชา : วิทยาศาสตร์ข้อมูลเพื่อนวัตกรรม
มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
อาจารย์ที่ปรึกษาการค้นคว้าอิสระหลัก : ผู้ช่วยศาสตราจารย์ ดร.มาลีรัตน์ มะลิแย้ม
ปีการศึกษา : 2568

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาวิธีการจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง ให้มีประสิทธิภาพในการจัดการข้อมูลในองค์กรและลดเพิ่มประสิทธิภาพการทำงาน โดยให้สามารถจำแนก ประเภทของอีเมลได้ 6 วัตถุประสงค์ ได้แก่ งานตอบคำถามทั่วไป งานส่งใบเสนอสินค้า งานบริการติดตั้ง งานตรวจสอบเอกสารข้อกำหนด งานร่างเอกสารข้อกำหนด และงานส่งเอกสารอ้างอิงข้อกำหนด โดยการรวบรวมอีเมลกลุ่มตัวอย่างจำนวน 9,569 ฉบับ แบ่งเป็นจำนวนตัวอย่างที่ใช้ในการเรียนรู้ 7,655 ฉบับและตัวอย่างที่ใช้ในการทดสอบและวัดผล 1,914 ฉบับ และสร้างแบบจำลองเปรียบเทียบ 3 วิธีการเรียนรู้ ได้แก่ เบย์สแบบง่าย, เพื่อนบ้านใกล้ที่สุดและหน่วยความจำระยะยาวแบบสั้น งานวิจัยจะแบ่งเป็น 2 ขั้นตอนดังนี้ ขั้นตอนแรกคือการจำแนกประเภทอีเมลตามวัตถุประสงค์หลัก โดยเปรียบเทียบ 3 แบบจำลอง ผลการทดลองพบว่าวิธีการเรียนรู้หน่วยความจำระยะยาวแบบสั้นให้ค่าความถูกต้องร้อยละ 96 และเพิ่มสูงขึ้นเป็นร้อยละ 99 หลังผ่านการเพิ่มประสิทธิภาพด้วยวิธีการค้นหาแบบผสมผสานและการสกัดวลีสำคัญ ขั้นตอนที่สองเป็นการพัฒนาแบบจำลองให้รองรับการจำแนกพร้อมกันสองวัตถุประสงค์สำหรับอีเมลที่มีวัตถุประสงค์ซ้อนทับกัน ผลการทดลองพบว่า วิธีการจำแนกสองวัตถุประสงค์ให้ค่าความถูกต้องร้อยละ 96.77 นอกจากนี้ผู้วิจัยได้นำแบบจำลองมาพัฒนาเป็นโปรแกรมประยุกต์เพื่อทดลองใช้ในการปฏิบัติงานจริงซึ่งให้ผลเป็นที่น่าพอใจอีกด้วย

(มีจำนวนทั้งสิ้น 84 หน้า)

คำสำคัญ : การเรียนรู้ของเครื่อง, โมเดลภาษาธรรมชาติ, การจำแนกอีเมล

_____ อาจารย์ที่ปรึกษาการค้นคว้าอิสระหลัก

Name : Mr. Theerachai Kasemphongjarat
Independent Study Title : E-Mail Objective Classification using Machine Learning
Major Field : Data Science for Innovation
King Mongkut's University of Technology North Bangkok
Independent Study Advisor : Assistant Professor Dr. Maleerat Maliyam
Academic Year : 2025

ABSTRACT

This research aims to develop a machine learning approach for email objective classification to improve organizational efficiency. Emails are classified into six categories: general inquiry response, product quotation, installation service, requirement document inspection, requirement document drafting, and reference document submission. A total of 9,569 email samples were collected, with 7,655 used for training and 1,914 for testing. Three algorithms such as Naïve Bayes, K-Nearest Neighbors, and Long Short-Term Memory were compared. The study consists of two phases. In the first phase, single objective classification was performed, where LSTM achieved the highest accuracy of 96%, improving to 99% after optimization using hybrid search and key phrase extraction. In the second phase, the model was extended to support dual objective classification for overlapping email purposes, achieving an accuracy of 96.77%. The model was further implemented as an application for real-world use, demonstrating satisfactory performance. The results indicate that the proposed approach is effective for enhancing email management in organizations.

(Total 84 Pages)

Keywords: Machine Learning, Natural Language Processing, Email Classification

_____ Advisor

กิตติกรรมประกาศ

ข้าพเจ้าขอขอบพระคุณอาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.มาลีรัตน์ มะลิแย้ม สำหรับคำแนะนำ ข้อเสนอแนะ และความเอาใจใส่ในการดำเนินงานวิจัยครั้งนี้ ความรู้และประสบการณ์ของท่านเป็นแรงสนับสนุนสำคัญที่ทำให้ข้าพเจ้าสามารถผ่านอุปสรรคต่าง ๆ ในการทำงานวิจัยได้อย่างราบรื่น ข้าพเจ้าขอขอบพระคุณคณะกรรมการสอบประเมินผลงานและอาจารย์ผู้ทรงคุณวุฒิทุกท่าน สำหรับคำแนะนำ ข้อเสนอแนะ และความคิดเห็นอันทรงคุณค่า ซึ่งช่วยให้การวิจัยครั้งนี้สมบูรณ์และมีคุณภาพยิ่งขึ้น ข้าพเจ้าขอขอบคุณเจ้าหน้าที่ และผู้ที่เกี่ยวข้องทุกท่านที่ให้ความช่วยเหลือในด้านข้อมูล เครื่องมือ และการดำเนินงานวิจัยตลอดระยะเวลา

ข้าพเจ้าขอขอบคุณครอบครัวและบุคคลอันเป็นที่รักสำหรับกำลังใจ การสนับสนุน และความเข้าใจในทุกช่วงเวลาของการศึกษา ข้าพเจ้าหวังเป็นอย่างยิ่งว่าผลงานวิจัยนี้จะเป็นประโยชน์ต่อการศึกษา การประยุกต์ใช้ความรู้ด้านเทคโนโลยี และเป็นแรงบันดาลใจให้ผู้สนใจต่อยอดงานวิจัยในอนาคต ข้าพเจ้าขอขอบพระคุณทุกท่านมา ณ โอกาสนี้

ธีระชัย เกษมพงศ์จรัส

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ง
บทคัดย่อภาษาอังกฤษ	จ
กิตติกรรมประกาศ	ฉ
สารบัญ	ช
สารบัญตาราง	ซ
สารบัญภาพ	ญ
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์งานวิจัย	2
1.3 ขอบเขตงานวิจัย	2
1.4 นิยามศัพท์เฉพาะ	2
1.5 ประโยชน์ที่ได้รับ	3
บทที่ 2 แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง	4
2.1 การเรียนรู้ของเครื่อง	4
2.2 การเรียนรู้เชิงลึก	18
2.3 การประมวลผลภาษาธรรมชาติ	24
2.4 งานวิจัยที่เกี่ยวข้อง	29
บทที่ 3 วิธีดำเนินการ	33
3.1 ขั้นตอนการดำเนินการ	33
3.2 ชุดข้อมูลและการเตรียมข้อมูล	34
3.3 การสร้างและเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่อง	46
3.4 การออกแบบแอปพลิเคชั่นจำแนกวัตถุประสงค์อีเมล	51
บทที่ 4 ผลการวิจัย	52
4.1 การประเมินผลการเรียนรู้ของเครื่อง (Model Evaluation)	52
4.2 การปรับปรุงแบบจำลองการเรียนรู้ของเครื่อง (Model Optimization)	57
4.3 การปรับปรุงแบบจำลองการเรียนรู้ของเครื่องสำหรับทำนาย 2 วัตถุประสงค์	61
4.4 การสร้างแอปพลิเคชั่นจำแนกวัตถุประสงค์อีเมล	64

สารบัญ (ต่อ)

	หน้า
บทที่ 5 สรุป อภิปรายผล และข้อเสนอแนะ	68
5.1 สรุป	68
5.2 อภิปรายผล	69
5.3 ข้อเสนอแนะ	69
บรรณานุกรม	70
ประวัติผู้เขียน	73

มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

สารบัญตาราง

ตารางที่		หน้า
2-1	กรณีการใช้งานของการเรียนรู้ของเครื่องจำแนกตามประเภทของปัญหา	7
2-2	เมทริกซ์ผู้ใช้–รายการ (User / Item Matrix)	11
2-3	เมทริกซ์ความไม่คล้ายคลึง (Dissimilarity Matrix)	12
3-1	ตัวอย่างไฟล์ซีเอสวี (CSV) เก็บบันทึกตัวอย่างอีเมล	38
3-2	ผลลัพท์ในขั้นตอนทำความสะอาดข้อมูลทั้งหมด	45
3-3	ตัวอย่างไฟล์ซีเอสวี (CSV) หลังผ่านกระบวนการทำความสะอาดข้อมูล	46
4-1	ผลประเมินแบบจำลองแบบหน่วยความจำระยะยาวแบบสั้น (LSTM)	52
4-2	ผลประเมินแบบจำลองเบย์สแบบง่าย (Naïve Bayes)	53
4-3	ผลประเมินแบบจำลองแบบหน่วยความจำระยะยาวแบบสั้น (KNN)	54
4-4	การเปรียบเทียบภาพรวมของทุกแบบจำลอง	55
4-5	ถ้อยคำที่บ่งบอกถึงลักษณะของประเภทงาน	57
4-6	ผลประเมินแบบจำลองแบบหน่วยความจำระยะยาวแบบสั้น (LSTM) ปรับปรุง	60
4-7	ผลประเมินการทดสอบตัวอย่าง แบบจำลองสำหรับทำนาย 2 ป้าย	63

สารบัญภาพ

ภาพที่		หน้า
2-1	กระบวนการเรียนรู้ของเครื่องในการจำแนกภาพสุนัขและแมว	5
2-2	ขั้นตอนการทำงานของระบบการเรียนรู้ของเครื่อง	6
2-3	การแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเพื่อใช้ในการประเมินประสิทธิภาพของแบบจำลอง	14
2-4	โครงสร้างของเมทริกซ์ความสับสน (Confusion Matrix)	16
2-5	ตัวชี้วัดประสิทธิภาพ Sensitivity, Specificity, Fall-out และ Miss rate	16
2-6	กราฟกระจายแสดงความสัมพันธ์ระหว่างค่าที่ทำนายได้กับค่าจริงของตัวแปรเป้าหมาย	16
2-7	การคำนวณค่า RMSE จากระยะห่างระหว่างค่าทำนายและค่าจริงของแต่ละตัวอย่าง	17
2-8	การเปรียบเทียบความลึกของแบบจำลองการเรียนรู้	19
2-9	ตัวอย่างฟังก์ชันกระตุ้นที่ใช้ในโครงข่ายประสาทเทียม Sigmoid, ReLU และ Tanh	20
2-10	โครงสร้างของโครงข่ายประสาทเทียมและกราฟการคำนวณของเครือข่าย	20
2-11	โครงสร้างของ Recurrent Neural Network	21
2-12	การกระจายตัวของเวกเตอร์คำในพื้นที่เวกเตอร์	26
2-13	ตัวอย่างความสัมพันธ์ของคำจากการคำนวณเวกเตอร์	27
2-14	โครงสร้างโครงข่ายประสาทเทียมสำหรับระบุชนิดคำโดยใช้ Word Embeddings	28
3-1	แผนภาพ Flow Chart แสดงขั้นตอนการดำเนินการ	33
3-2	หน้าใช้งานของ Microsoft Graph	35
3-3	ตัวอย่างไฟล์ข้อมูลอีเมลในรูปแบบไฟล์สัญกรณ์วัตถุจาวาสคริปต์ (JSON)	37
3-4	กราฟแท่ง (Bar Chart) แสดงการกระจายของป้ายชื่อ	39
3-5	ตัวอย่างเนื้อหาอีเมลที่ยังไม่ผ่านกระบวนการ	40
3-6	ตัวอย่างข้อความหลังผ่านขั้นตอน Tokenization	41
3-7	ตัวอย่างข้อความหลังการผ่านลบคำไม่สำคัญและสัญลักษณ์	42
3-8	ตัวอย่างข้อความหลังการผ่านลบชื่อบุคคลและหน่วยงานในข้อความอีเมล	43
3-9	ตัวอย่างข้อความหลังการผ่านลบลบตัวอักษรเดี่ยวและตัวเลข	44
3-10	Flow Chart การทำงานของแอพลิเคชั่นจำแนกวัตถุประสงค์อีเมล	51
3-11	โครงร่างหน้าผู้ใช้ (GUI) สำหรับแอพลิเคชั่นจำแนกวัตถุประสงค์อีเมล	51
4-1	การเปรียบเทียบ Confusion Matrix ของเทคนิค Long Short-Term Memory	53
4-2	การเปรียบเทียบ Confusion Matrix ของเทคนิค Naïve Bayes	54

สารบัญภาพ (ต่อ)

ภาพที่		หน้า
4-3	การเปรียบเทียบ Confusion Matrix ของเทคนิค K-Nearest Neighbors	55
4-4	การเปรียบเทียบ Confusion Matrix ของเทคนิค LSTM ปรับปรุง	61
4-5	สัญลักษณ์ Icon ของ Application	66
4-6	หน้าต่างโปรแกรมพยากรณ์วัตถุประสงค์ของ Email	66
4-7	การทำงานของโปรแกรมพยากรณ์วัตถุประสงค์ของอีเมล์	67
4-8	ตัวอย่าง Output ผลพยากรณ์แสดงผ่านโปรแกรม Microsoft Excel	67

พระจอมเกล้าพระนครเหนือ มหาวิทยาลัยเทคโนโลยี

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในยุคที่เทคโนโลยีก้าวหน้าอย่างรวดเร็ว องค์กรต่าง ๆ กำลังขับเคลื่อนธุรกิจด้วยข้อมูลในปริมาณมหาศาล โดยเฉพาะข้อมูลดิจิทัลในรูปแบบของเอกสารและข้อความ ซึ่งเป็นปัจจัยสำคัญและมีการเพิ่มขึ้นอย่างต่อเนื่องในแต่ละวัน การจัดการและประมวลผลข้อมูลเหล่านี้จึงมีความสำคัญอย่างยิ่งต่อการดำเนินงานและสร้างความได้เปรียบทางธุรกิจ เพื่อเพิ่มประสิทธิภาพและลดภาระงานที่ไม่จำเป็น หลายองค์กรจึงมุ่งเน้นการพัฒนาระบบอัตโนมัติเพื่อจัดการกับข้อมูลจำนวนมากนี้ โดยเฉพาะอย่างยิ่งในภาคส่วนบริการลูกค้า การจัดประเภทอีเมลโดยอัตโนมัติสามารถช่วยจัดลำดับความสำคัญของข้อความ เพิ่มความรวดเร็วในการตอบสนอง และส่งผลให้ความพึงพอใจของลูกค้าเพิ่มขึ้น ซึ่งเป็นการใช้ประโยชน์จากข้อมูลเพื่อขับเคลื่อนองค์กรไปข้างหน้าอย่างแท้จริง

อย่างไรก็ตาม แม้อีเมลยังคงเป็นช่องทางการสื่อสารหลักที่สำคัญในองค์กรและสถาบันต่าง ๆ แต่ข้อมูลที่ส่งผ่านอีเมลเหล่านี้ส่วนใหญ่เป็นข้อมูลที่ไม่มีโครงสร้าง การสกัดข้อมูลอย่างเช่นการปฏิบัติงานหรือข้อมูลสำคัญอื่น ๆ จากอีเมลจำนวนมหาศาลที่ส่งหากันอยู่ทุกวัน จึงกลายเป็นความท้าทายอย่างมาก ปัญหาหลักคือปริมาณอีเมลที่มีอยู่เป็นจำนวนมาก ประกอบกับข้อจำกัดด้านเวลาของผู้ปฏิบัตงาน ทำให้ไม่สามารถไล่อ่านอีเมลทั้งหมดเพื่อสกัดและบันทึกข้อมูลเชิงลึกได้อย่างมีประสิทธิภาพด้วยตนเอง ซึ่งไม่เพียงแต่ใช้เวลานานและมีค่าใช้จ่ายสูง แต่ยังส่งผลกระทบโดยตรงต่อประสิทธิภาพการทำงานของพนักงาน และอาจนำไปสู่ความเครียดจากการทำงานได้ ซึ่งส่งผลให้องค์กรไม่สามารถดึงข้อมูลเชิงลึกที่เป็นประโยชน์ออกมาใช้ในการตัดสินใจได้อย่างเต็มที่

จากปัญหาดังกล่าว เราจึงเล็งเห็นถึงแนวทางในการประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) เพื่อแก้ไขข้อจำกัดในการจัดการข้อมูลอีเมลเหล่านี้ โดยมีเป้าหมายหลักในการสร้างโมเดลการเรียนรู้ของเครื่องที่สามารถอ่านและจัดประเภทเนื้อหาอีเมลได้โดยอัตโนมัติ กระบวนการเริ่มต้นด้วยการเตรียมชุดข้อมูลอีเมลในอดีต ซึ่งจะต้องผ่านขั้นตอนการการติดป้ายหมวดหมู่ (Class Labeling) ให้กับข้อมูลตามลักษณะงานที่วิเคราะห์จากเนื้อหาอีเมลด้วยมนุษย์ ตามมาด้วยการแปลงข้อความดิบให้เป็นรูปแบบที่คอมพิวเตอร์เข้าใจได้ผ่านการสกัดคุณสมบัติ (Feature Extraction) เช่น การใช้เทอมความถี่-ความถี่ผกผันของเอกสาร (Term Frequency-Inverse Document Frequency หรือการฝังคำ (Word Embeddings) และขั้นตอนสุดท้ายคือ

การสร้างแบบจำลองการเรียนรู้ (Model Training) โดยใช้อัลกอริทึมการเรียนรู้ของเครื่องที่เหมาะสม เช่นตัวจำแนกแบบเบย์สแบบง่าย (Naïve Bayes) ตัวจำแนกแบบเพื่อนบ้านใกล้ (K-Nearest Neighbors) และหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory) เมื่อได้แบบจำลองการเรียนรู้ ที่มีความเสถียรและแม่นยำตามการประเมินด้วยเมตริกต่าง ๆ เช่น ค่าความแม่นยำ (Accuracy) ค่าความเที่ยง (Precision) ค่าการเรียกคืน (Recall) และคะแนนเอฟ 1 (F1-score) จะสามารถนำไปพัฒนาต่อยอดเป็นแอปพลิเคชันเพื่อบันทึกข้อมูลงานจากอีเมลที่ส่งไปแล้วในอดีตและอีเมลที่จะมีการส่งในอนาคตได้อย่างอัตโนมัติ ซึ่งจะช่วยลดภาระงานและสร้างคุณค่าจากข้อมูลที่เคยเข้าถึงได้ยากเพื่อช่วยประกอบการตัดสินใจทางธุรกิจอย่างมีประสิทธิภาพ

1.2 วัตถุประสงค์งานวิจัย

เพื่อศึกษาการนำข้อมูลจากอีเมลมาใช้ให้เกิดประโยชน์สูงสุด

1.3 ขอบเขตงานวิจัย

การค้นคว้าอิสระมีการกำหนดขอบเขตของงานวิจัย ดังนี้

1.3.1 ใช้ข้อมูลอีเมลแบบดิบ (Raw Data) ที่บันทึกจากการฐานข้อมูลอีเมลคลาวด์จำนวน 9,569 ฉบับ เป็นข้อมูล ณ วันที่ 1 มกราคม พ.ศ. 2568 ถึงวันที่ 31 ธันวาคม พ.ศ. 2568 ของกลุ่มพนักงานในบริษัทเอกชนแห่งหนึ่ง ลักษณะข้อมูลเป็นรายละเอียดของอีเมล 1 ฉบับและเป็นรูปแบบสัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) เนื่องจากข้อมูลอีเมลมีส่วนที่เป็นข้อมูลส่วนบุคคลดังนั้นในงานวิจัยนี้จะนำมาเฉพาะข้อมูลที่เกี่ยวข้อง ได้แก่ ชื่อไฟล์และเนื้อหาอีเมล

1.3.2 แฟลตฟอร์มที่ใช้ในการศึกษา ปัจจุบันในภาคธุรกิจมีตัวเลือกในการใช้แฟลตฟอร์มอีเมลคลาวด์ที่หลากหลาย โดยแฟลตฟอร์มที่งานวิจัยนี้ใช้คือไมโครซอฟต์ออฟฟิต 365 (Office 365)

1.4 นิยามศัพท์เฉพาะ

เนื่องจากงานวิจัยนี้มีการกำหนดชื่อประเภทงานที่เจาะจงตามลักษณะงานของผู้วิจัย ซึ่งปฏิบัติงานอยู่ในบริษัทตัวแทนนำเข้าและจัดจำหน่ายสินค้าแห่งหนึ่ง ดังนั้นจะอธิบายความหมายของแต่ละประเภทงานดังนี้

งานตอบคำถามทั่วไป (Normal) หมายถึงการตอบกลับอีเมลภายใต้เงื่อนไข การให้คำตอบกรณีก่อนหน้ามีการถามเกี่ยวกับสินค้า การออกแบบหรือการร้องขอการตรวจสอบและยืนยัน หรือการสนธนา ที่ไม่ตรงตามเงื่อนไขของประเภทงานอื่นๆ โดยงานนี้เป็นงานที่ไม่ได้ใช้เวลามาก

งานส่งใบเสนอสินค้า (Quote job) หมายถึง การตอบกลับอีเมลพร้อมแนบใบเสนอสินค้าหรือ Bill of Material (BoM) เนื่องจากก่อนหน้ามีการร้องข้อให้นำเสนอสินค้าพร้อมระบุความต้องการ

พนักงานจึงต้องสร้างเอกสารระบุถึงรุ่นสินค้า,ข้อกำหนดเฉพาะ,จำนวนและราคาด้วยเครื่องมือที่บริษัทกำหนด ดังนั้นงานนี้จึงต้องมีความรู้ความเข้าใจในสินค้าและเครื่องมือ ซึ่งเป็นคุณค่าของผู้ปฏิบัติงาน

งานบริการติดตั้ง (Service) หมายถึง การตอบกลับอีเมลพร้อมแนบราคาบริการติดตั้งและขอบเขตงาน เนื่องจากบริษัททำธุรกิจจำหน่ายสินค้าเทคโนโลยีสารสนเทศสำหรับองค์กร ในบางสินค้ามีความซับซ้อนและมีข้อบังคับเช่นต้องมีใบอนุญาตในการติดตั้งเพื่อให้ใช้งานได้อย่างสมบูรณ์และอยู่ในเงื่อนไขในการรับประกันสินค้า ทางบริษัทจึงมีบริการดังกล่าวเพื่อช่วยให้ผู้ใช้สามารถใช้สินค้าได้ตามวัตถุประสงค์ ดังนั้นงานนี้จะเกี่ยวข้องกับการออกแบบขอบเขตงาน (Scope of Work) จึงต้องมีความเข้าใจในความสามารถและข้อจำกัดในการให้บริการที่บริษัทสามารถทำได้ รวมถึงการประสานงานกับแผนกบริการเพื่อกำหนดราคาตามขอบเขตงานที่กำหนดและส่งกลับไปยังผู้ร้องขอต่อไป

งานตรวจสอบเอกสารข้อกำหนด (TOR Review) หมายถึง การตอบกลับอีเมลเป็นคำตอบ ในบางครั้งลักษณะความต้องการที่มีการส่งมาในรูปแบบเอกสาร TOR (Terms of Reference) โดยเฉพาะงานจัดซื้อจัดจ้างจากภาครัฐ เพื่อให้ตรวจสอบว่าข้อกำหนดที่เกี่ยวข้องกับสินค้านั้นสามารถปฏิบัติตามได้หรือไม่ การตอบกลับจะระบุว่าผ่านข้อกำหนดหรือไม่ พร้อมระบุชื่อสินค้าและรุ่น

งานร่างเอกสารข้อกำหนด (TOR Draft) หมายถึง การตอบกลับอีเมลพร้อมแนบเอกสารคุณสมบัติเฉพาะของสินค้าที่นำเสนอ เขียนเป็นเอกสารข้อกำหนดตามรูปแบบทางการหรือร่าง TOR สำหรับหน่วยงานภาครัฐหรือเอกชนบางแห่ง งานประเภทนี้จะต้องใช้ความรู้ในการเขียนเอกสารทางการที่ถูกต้องและต้องเขียนตามคุณสมบัติเฉพาะของสินค้าที่นำเสนอ แล้วส่งกลับไปยังผู้ร้องขอ

งานส่งเอกสารอ้างอิงข้อกำหนด (TOR Comply) หมายถึง การตอบกลับอีเมลพร้อมแนบเอกสาร กรณีเป็นงานจัดซื้อจัดจ้างที่ต้องการเอกสารของสินค้าที่สื่อถึงข้อกำหนดตามเอกสาร TOR งานประเภทนี้จะใช้เวลามากในการรวบรวมเอกสารและประกอบเป็นรูปเล่ม ซึ่งจะเป็นงานที่ใช้เวลาปฏิบัติมากที่สุด

1.5 ประโยชน์ที่ได้รับ

1.5.1 เพิ่มประสิทธิภาพในการจัดการอีเมล

1.5.2 ประหยัดเวลาและเพิ่มประสิทธิภาพการทำงาน

1.5.3 รองรับการใช้งานในองค์กรช่วยให้การทำงานเป็นระบบมากขึ้น

บทที่ 2
แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานค้นคว้าอิสระเรื่องการจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่อง ซึ่งวิจัยในอดีตมีหลากหลายแนวคิดและวิธีการในการประยุกต์ใช้เรียนรู้ของเครื่องกับอีเมลในการคัดแยกตามวัตถุประสงค์ที่แตกต่างกัน เพื่อให้เข้าใจถึงหลักการผู้วิจัยจึงได้ศึกษาแนวคิดและทฤษฎีที่จำเป็นเพื่อใช้เป็นแนวทางในการศึกษา รวมถึงงานวิจัยที่เกี่ยวข้องและสรุปผลการศึกษา โดยในบทนี้จะกล่าวถึงหัวข้อดังนี้

2.1 การเรียนรู้ของเครื่อง

2.2 การเรียนรู้เชิงลึก

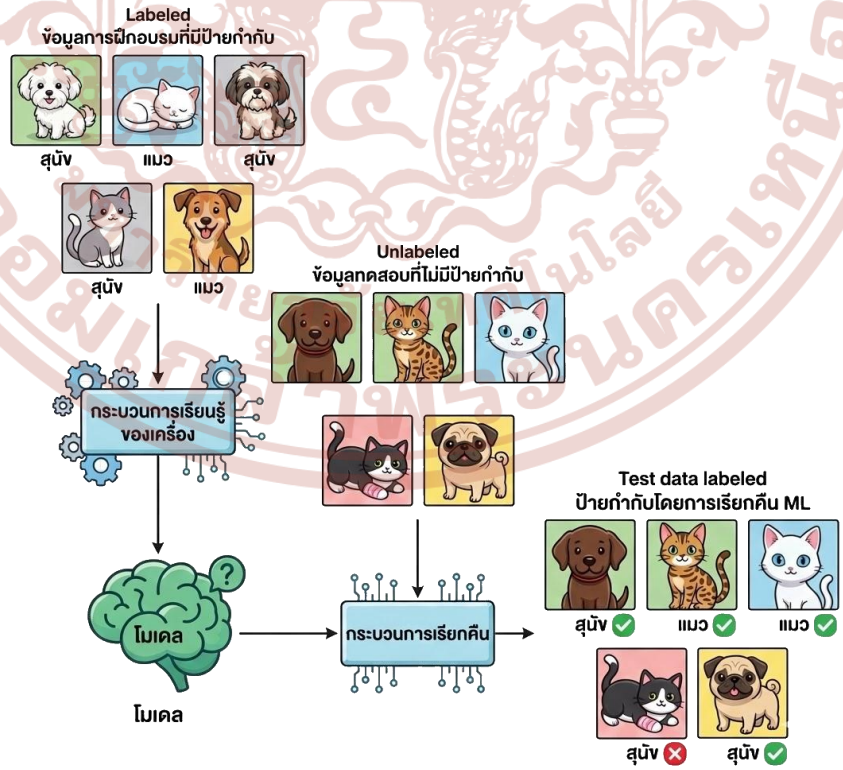
2.3 การประมวลผลภาษาธรรมชาติ

2.4 งานวิจัยที่เกี่ยวข้อง

2.1 การเรียนรู้ของเครื่อง

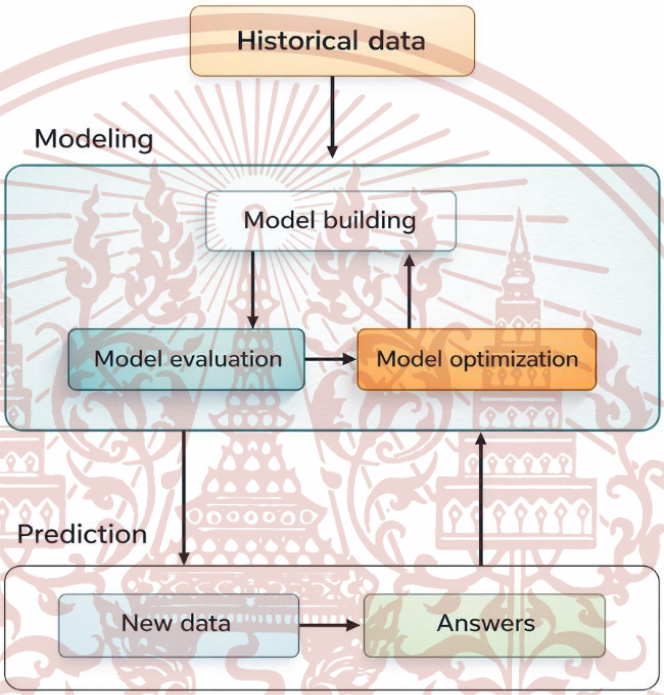
การเรียนรู้ของเครื่อง (Machine Learning) เป็นสาขาหนึ่งที่ได้รับความสนใจอย่างกว้างขวางในปัจจุบัน หากพิจารณาในบริบทของการเรียนรู้ของมนุษย์ โดยทั่วไปสามารถจำแนกการเรียนรู้ได้เป็นสองลักษณะหลัก ได้แก่ การเรียนรู้แบบท่องจำ (Rote Learning) และการเรียนรู้ที่สะท้อนถึงความเข้าใจเชิงลึกหรือสติปัญญา การจดจำข้อมูลบางประเภท เช่น หมายเลขโทรศัพท์หรือชุดคำสั่ง สามารถถือได้ว่าเป็นรูปแบบหนึ่งของการเรียนรู้ที่อาศัยกระบวนการจดจำเป็นหลัก อย่างไรก็ตาม เมื่อกล่าวถึงการเรียนรู้ในความหมายที่กว้างขึ้น มักหมายถึงความสามารถในการทำความเข้าใจรูปแบบ ความสัมพันธ์ และประสบการณ์ที่เกิดขึ้นในชีวิตประจำวัน ตัวอย่างเช่น เมื่อเด็กมีปฏิสัมพันธ์กับผู้อื่นในกลุ่ม เด็กจะค่อย ๆ เรียนรู้รูปแบบการตอบสนองของบุคคลรอบข้างต่อการกระทำของตนเอง ประสบการณ์ดังกล่าวจะมีบทบาทสำคัญในการกำหนดพฤติกรรมทางสังคมของเด็กในอนาคต แม้ว่าเด็กจะไม่ได้ย้อนกลับไปทบทวนเหตุการณ์ในอดีตอย่างเป็นระบบ แต่สมองของมนุษย์สามารถจดจำลักษณะสำคัญของสถานการณ์ต่าง ๆ ได้ เช่น สภาพแวดล้อม บทบาทของบุคคลที่เกี่ยวข้อง หรือบริบทของกิจกรรมที่เกิดขึ้น ลักษณะร่วมของประสบการณ์เหล่านี้ช่วยให้มนุษย์สามารถประเมินสถานการณ์ใหม่ได้โดยเปรียบเทียบกับประสบการณ์ที่ผ่าน เพื่ออธิบายแนวคิดดังกล่าวให้เห็นภาพชัดเจนยิ่งขึ้น สามารถพิจารณาตัวอย่างของการสอนเด็กให้จำแนกสุนัขและแมวโดยใช้บัตรภาพ (Flashcards) ในกระบวนการนี้ ผู้สอนจะแสดงบัตรภาพให้เด็กพิจารณา จากนั้นให้เด็กตัดสินใจว่า

ภาพดังกล่าวเป็นสุนัขหรือแมว หากเด็กตอบถูกต้อง บัตรภาพจะถูกจัดไว้ในกลุ่มคำตอบที่ถูกต้อง แต่หากตอบไม่ถูกต้อง บัตรภาพจะถูกจัดไว้ในอีกกลุ่มหนึ่งเพื่อสะท้อนความผิดพลาด เมื่อเด็กได้รับการฝึกฝนซ้ำหลายครั้ง ประสิทธิภาพในการจำแนกภาพจะค่อย ๆ เพิ่มขึ้นอย่างต่อเนื่อง สิ่งที่น่าสนใจ คือ เด็กไม่จำเป็นต้องได้รับการสอนกฎเกณฑ์เฉพาะเกี่ยวกับการจำแนกสุนัขและแมวตั้งแต่แรก เนื่องจากกระบวนการรับรู้ของมนุษย์มีความสามารถในการจัดประเภทข้อมูลโดยธรรมชาติ ความสามารถในการนำความรู้ที่ได้จากประสบการณ์ไปประยุกต์ใช้กับข้อมูลใหม่ที่ไม่เคยพบมาก่อนนี้ เรียกว่า ความสามารถในการทั่วไป (Generalization) ซึ่งถือเป็นคุณลักษณะสำคัญของทั้งการเรียนรู้ของมนุษย์และการเรียนรู้ของเครื่อง แม้ว่าการเรียนรู้ของมนุษย์จะมีความซับซ้อนและมีมิติที่ลึกซึ้งกว่าระบบการเรียนรู้ของเครื่องในปัจจุบัน แต่คอมพิวเตอร์มีข้อได้เปรียบที่สำคัญในด้านความสามารถในการจัดเก็บข้อมูลจำนวนมาก การเรียกคืนข้อมูลอย่างรวดเร็ว และการประมวลผลข้อมูลในปริมาณมหาศาลได้อย่างมีประสิทธิภาพ ประสบการณ์ของระบบคอมพิวเตอร์จึงมักอยู่ในรูปแบบของข้อมูลในอดีต (Historical data) ซึ่งถูกนำมาประมวลผลด้วยเทคนิคทางการคำนวณเพื่อสร้างแบบจำลอง (model) และปรับปรุงอัลกอริทึมให้มีประสิทธิภาพสูงขึ้นตามประสบการณ์ที่ได้รับภาพที่ 2-1 จะแสดงให้เห็นถึงกระบวนการของการเรียนรู้ของเครื่อง



ภาพที่ 2-1 กระบวนการเรียนรู้ของเครื่องในการจำแนกภาพสุนัขและแมว

โดยทั่วไปกระบวนการพัฒนาแบบจำลองการเรียนรู้ของเครื่องจะประกอบด้วยขั้นตอนสำคัญ ได้แก่ การสร้างแบบจำลอง (Model Building) การประเมินแบบจำลอง (Model Evaluation) และการปรับปรุงแบบจำลอง (Model Optimization) ซึ่งดำเนินการในลักษณะเป็นวงจรเพื่อปรับปรุงประสิทธิภาพแบบจำลองให้เหมาะสมกับข้อมูลและวัตถุประสงค์ของงานวิจัย ดังแสดงในภาพที่ 2-2



ภาพที่ 2-2 ขั้นตอนการทำงานของระบบการเรียนรู้ของเครื่อง

อัลกอริทึมที่ใช้ใน Machine Learning ได้รับการพัฒนาอย่างต่อเนื่องโดยนักวิจัยจากหลากหลายสาขา ไม่ว่าจะเป็นสถิติ วิทยาการคอมพิวเตอร์ คณิตศาสตร์ประยุกต์ตลอดจนภาคอุตสาหกรรมที่เกี่ยวข้องกับเทคโนโลยีดิจิทัล เช่น ระบบค้นหาข้อมูลออนไลน์ สื่อบันเทิงดิจิทัล การโฆษณาออนไลน์ และการแปลภาษา อัลกอริทึมเหล่านี้มีความหลากหลายและถูกออกแบบมาเพื่อตอบสนองต่อปัญหาที่แตกต่างกัน แต่สิ่งที่เหมือนกันคือความสามารถในการเรียนรู้จากตัวอย่างข้อมูลและการนำความรู้ที่ได้ไปประยุกต์ใช้กับกรณีใหม่ที่ไม่เคยปรากฏมาก่อน

ในปัจจุบัน Machine Learning กำลังมีบทบาทสำคัญมากขึ้นในการขับเคลื่อนเศรษฐกิจยุคใหม่ที่อาศัยข้อมูลเป็นศูนย์กลาง (data-driven economy) ทั้งนี้ เทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอนที่ได้รับความนิยมและตัวอย่างการประยุกต์ใช้งานในทางปฏิบัติสามารถสรุปได้ในตารางที่ 2-1

ตารางที่ 2-1 กรณีการใช้งานของการเรียนรู้ของเครื่องจำแนกตามประเภทของปัญหา

ประเภทปัญหา (Problem)	ตัวอย่างการประยุกต์ใช้งาน (Example Use Cases)
การจำแนกประเภท (Classification) การกำหนดว่าหน่วยข้อมูลแต่ละรายการควรอยู่ในหมวดหมู่ใด โดยพิจารณาจากข้อมูลนำเข้า	การกรองอีเมลขยะ, การตรวจจับเนื้อหาที่ไม่เหมาะสมบนอินเทอร์เน็ต, การตรวจจับการทุจริต, การแบ่งกลุ่มลูกค้าและการตรวจจับข้อบกพร่องในกระบวนการผลิต
การถดถอย (Regression) การคาดการณ์ค่าผลลัพธ์ที่เป็นค่าตัวเลขต่อเนื่อง สำหรับแต่ละหน่วยข้อมูลจากข้อมูลนำเข้า	การพยากรณ์ราคา, การพยากรณ์ความต้องการสินค้า, การประมาณราคา, การพยากรณ์อากาศ และการคาดการณ์ยอดขายสินค้า
ระบบแนะนำ (Recommendation) การคาดการณ์ทางเลือกหรือรายการที่ผู้ใช้มีแนวโน้มจะชื่นชอบ	การแนะนำสินค้า, การคัดเลือกบุคลากร, ระบบแนะนำภาพยนตร์ เช่น เน็ตฟลิกซ์ (Netflix), การออกเดทออนไลน์และการแนะนำเนื้อหา
การประมาณค่าข้อมูลที่ขาดหาย (Imputation) การอนุมานหรือประมาณค่าของข้อมูลที่สูญหายในชุดข้อมูล	การเติมข้อมูลเวชระเบียนผู้ป่วยที่ไม่สมบูรณ์, การประมาณค่าข้อมูลลูกค้าที่ขาดหาย, และการเติมข้อมูลสำมะโนประชากรที่ไม่ครบถ้วน

ที่มา : (Brink et al., 2017:7)

2.1.1 ประเภทการเรียนรู้ของเครื่อง

ในการศึกษาด้านการเรียนรู้ของเครื่อง (Machine Learning) ปัญหาที่เกี่ยวข้องกับการวิเคราะห์ข้อมูลสามารถแบ่งออกได้เป็นสองประเภทหลัก ได้แก่ การเรียนรู้แบบมีผู้สอน (Supervised Learning) และการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ความแตกต่างที่สำคัญระหว่างแนวทางทั้งสองอยู่ที่การมีอยู่ของตัวแปรเป้าหมาย (Target Variable) ในชุดข้อมูล สำหรับปัญหาแบบมีผู้สอน ผู้พัฒนาระบบจะมีข้อมูลตัวอย่างที่ประกอบด้วยตัวแปรนำเข้า (Input Features) และค่าผลลัพธ์ที่ทราบล่วงหน้า ซึ่งเรียกว่าตัวแปรเป้าหมาย ข้อมูลดังกล่าวถูกนำมาใช้เป็นชุดข้อมูลสำหรับการฝึกสอน (Training Data) เพื่อให้แบบจำลองสามารถเรียนรู้ความสัมพันธ์ระหว่างตัวแปรนำเข้าและผลลัพธ์ที่ต้องการทำนาย ตัวอย่างของปัญหาประเภทนี้ เช่น การทำนายว่าผู้โดยสารในเหตุการณ์เรือไททานิกจะรอดชีวิตหรือไม่ การคาดการณ์ว่าลูกค้าจะยกเลิกการใช้บริการหรือไม่

หรือการประมาณค่าประสิทธิภาพการใช้เชื้อเพลิงของยานพาหนะ (Miles Per Gallon: MPG) ในกรณีเหล่านี้ แบบจำลองจะถูกฝึกให้เรียนรู้จากข้อมูลที่มีคำตอบที่ถูกต้องกำกับอยู่แล้ว ทำให้สามารถนำความรู้ที่ได้ไปใช้ทำนายผลลัพธ์ของข้อมูลใหม่ที่ไม่เคยปรากฏมาก่อน ทั้งนี้ ปัญหาส่วนใหญ่ในงานด้าน Machine Learning มักอยู่ในกลุ่มของ Supervised Learning และเทคนิคจำนวนมากที่ได้รับการพัฒนาขึ้นก็ถูกออกแบบมาเพื่อแก้ไขปัญหาประเภทนี้โดยเฉพาะ

ในทางตรงกันข้าม การเรียนรู้แบบไม่มีผู้สอนเป็นแนวทางที่ใช้เมื่อชุดข้อมูลมีเพียงตัวแปรนำเข้าเท่านั้น โดยไม่มีตัวแปรเป้าหมายที่ระบุผลลัพธ์ที่ถูกต้องไว้ล่วงหน้า กล่าวคือระบบจะต้องพยายามค้นหารูปแบบ โครงสร้าง หรือความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลด้วยตนเอง แม้ว่าจะไม่มีคำตอบที่ชัดเจนให้ใช้เป็นแนวทางในการฝึกสอนก็ตาม ภายใต้แนวทางการเรียนรู้แบบไม่มีผู้สอนสามารถแบ่งวิธีการวิเคราะห์ข้อมูลออกได้เป็นสองกลุ่มสำคัญ ได้แก่ การจัดกลุ่มข้อมูล (Clustering) และการลดมิติของข้อมูล (Dimensionality Reduction) การจัดกลุ่มข้อมูลเป็นกระบวนการที่ใช้ตัวแปรนำเข้าเพื่อค้นหากลุ่มหรือโครงสร้างตามธรรมชาติที่ปรากฏอยู่ภายในข้อมูล จากนั้นจึงแบ่งข้อมูลออกเป็นกลุ่มที่มีลักษณะคล้ายคลึงกัน วิธีการที่ได้รับความนิยมในกลุ่มนี้ ได้แก่ อัลกอริทึมเคมีนส์ (k-means), แบบจำลองการผสมแบบเกาส์เซียน (Gaussian Mixture Models) และการจัดกลุ่มแบบลำดับชั้น (Hierarchical Clustering) ในขณะที่การลดมิติของข้อมูลเป็นเทคนิคที่มุ่งลดจำนวนตัวแปรหรือคุณลักษณะของข้อมูลให้อยู่ในรูปของพิกัดใหม่จำนวนไม่มาก แต่ยังคงสามารถอธิบายความแปรปรวนของข้อมูลส่วนใหญ่ได้ วิธีการที่ใช้กันอย่างแพร่หลาย เช่น การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) การจัดสเกลหลายมิติ (Multidimensional Scaling) และการเรียนรู้โครงสร้างแมนิโฟลด์ (Manifold Learning)

2.1.2 อัลกอริทึมเบย์สแบบง่าย (Naïve Bayes)

อัลกอริทึมเบย์สแบบง่าย (Naïve Bayes)เป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ที่ได้รับความนิยมอย่างแพร่หลายในการวิเคราะห์และจำแนกข้อมูล เนื่องจากมีโครงสร้างของแบบจำลองที่เรียบง่าย แต่สามารถให้ประสิทธิภาพในการจำแนกประเภทข้อมูลได้อย่างมีประสิทธิผล แนวคิดพื้นฐานของอัลกอริทึมดังกล่าวมีรากฐานมาจาก ทฤษฎีบทของเบย์ (Bayes' Theorem) ซึ่งเป็นแนวคิดสำคัญในสถิติเชิงอนุมาน (Inferential Statistics) และเป็นพื้นฐานของแบบจำลองเชิงความน่าจะเป็นในงานด้านการเรียนรู้ของเครื่องหลายประเภท (Berrar, 2019) หลักการสำคัญของ Naïve Bayes คือการสร้างแบบจำลองความน่าจะเป็นแบบเบย์เซียน (Bayesian Probabilistic Model) เพื่อคำนวณค่าความน่าจะเป็นภายหลัง (Posterior Probability) ของ

คลาส (y_j) เมื่อพิจารณาจากคุณลักษณะของข้อมูลอินสแตนซ์ (x_i) จากนั้นจึงนำค่าความน่าจะเป็นดังกล่าวมาใช้ในการตัดสินใจจำแนกประเภทข้อมูล โดยอินสแตนซ์ข้อมูลจะถูกจัดให้อยู่ในคลาสที่มีค่าความน่าจะเป็นภายหลังสูงที่สุด แนวทางดังกล่าวทำให้ Naïve Bayes เป็นวิธีการจำแนกประเภทข้อมูลที่มีพื้นฐานทางสถิติที่ชัดเจนและสามารถตีความผลลัพธ์ได้ง่ายคำว่า “Naïve” ในชื่อของอัลกอริทึมเกิดจากสมมติฐานที่สำคัญของแบบจำลอง ซึ่งระบุว่าคุณลักษณะ (Attributes) หรือคุณสมบัติต่าง ๆ ของข้อมูลแต่ละตัวมีความเป็นอิสระต่อกัน (Conditional Independence) ภายใต้เงื่อนไขที่กำหนดคลาส (y_j) แล้ว กล่าวอีกนัยหนึ่งคือ เมื่อทราบคลาสของข้อมูลแล้ว การเกิดขึ้นของคุณลักษณะแต่ละตัวจะไม่ส่งผลกระทบต่อกัน สมมติฐานดังกล่าวถือเป็นสมมติฐานที่ค่อนข้างเข้มงวดและมักไม่สอดคล้องกับข้อมูลจริงในหลายสถานการณ์ เนื่องจากในข้อมูลจริงคุณลักษณะต่าง ๆ มักมีความสัมพันธ์หรือมีการพึ่งพากันในระดับหนึ่ง อย่างไรก็ตาม จากผลการศึกษาทางประจักษ์จำนวนมากพบว่า แม้สมมติฐานความเป็นอิสระของคุณลักษณะจะถูกละเมิดในทางปฏิบัติ แต่อัลกอริทึม Naïve Bayes ยังคงสามารถให้ผลลัพธ์ในการจำแนกประเภทข้อมูลที่มีความแม่นยำในระดับที่น่าพึงพอใจ และในบางกรณีสามารถแข่งขันกับแบบจำลองที่มีความซับซ้อนสูงกว่าได้

ในกระบวนการจำแนกประเภทข้อมูล ตัวจำแนก Naïve Bayes ใช้กฎการตัดสินใจที่เรียกว่ากฎความน่าจะเป็นภายหลังสูงสุด (Maximum A Posteriori Rule หรือ MAP) เพื่อกำหนดคลาสที่เหมาะสมที่สุดสำหรับอินสแตนซ์ข้อมูล โดยจะเลือกคลาส (y_j) ที่ทำให้ค่าความน่าจะเป็นภายหลัง $P(y_j|x)$ มีค่ามากที่สุด ดังนั้น กฎการจำแนกประเภทสามารถเขียนได้ในรูปแบบตามสมการที่ (2-1)

$$\hat{y} = \arg\max_{y_j} P\left(y_j|x\right) \qquad (2\text{-}1)$$

โดยที่ $\hat{y}$ แทนคลาสที่แบบจำลองทำนายสำหรับอินสแตนซ์ข้อมูล x หากสมมติฐานความเป็นอิสระของคุณลักษณะเป็นจริง ตัวจำแนก Naïve Bayes จะถือว่าเป็นวิธีการที่เหมาะสมที่สุดในแง่ของการลดอัตราความผิดพลาดในการจำแนกประเภท (classification error rate) ให้น้อยที่สุด
พื้นฐานทางคณิตศาสตร์ของ Naïve Bayes เริ่มต้นจากทฤษฎีบทของเบย์ ซึ่งสามารถเขียนในรูปแบบทั่วไปได้ตามสมการที่ (2-2)

$$P\left(y_j \mid x_i\right) = \frac{P\left(x_i|y_j\right)P(y_j)}{P(x_i)} \qquad (2\text{-}2)$$

โดยที่ $P\left(y_j \mid x_i\right)$ คือค่าความน่าจะเป็นภายหลังของคลาส (y_j) เมื่อทราบข้อมูลอินสแตนซ์ (x_i) ส่วน $P\left(x_i \mid y_j\right)$ คือค่าความน่าจะเป็นตามเงื่อนไข (Conditional Probability) ของ

การสังเกตข้อมูล (x_i) เมื่อทราบคลาส y_j และ $P(y_j)$ คือค่าความน่าจะเป็นล่วงหน้า (Prior Probability) ของคลาสนั้น ในกรณีที่ข้อมูลอินสแตนซ์ x ประกอบด้วยชุดคุณลักษณะหลายตัว เช่น $x = (x_1, x_2, \dots, x_p)$สามารถนำสมมติฐานความเป็นอิสระของคุณลักษณะมาใช้แยกความน่าจะเป็นร่วมออกเป็นผลคูณของความน่าจะเป็นตามเงื่อนไขของคุณลักษณะแต่ละตัวได้ตามสมการที่ (2-3)

$$P(x|y_j)P(y_j) = P(y_j)\prod_{k=1}^{p} P(x_k|y_j) \qquad (2\text{-}3)$$

ดังนั้น ค่าความน่าจะเป็นภายหลังสำหรับตัวจำแนก Naïve Bayes สามารถเขียนได้ตามสมการที่ (2-4)

$$P(y_j|x) = \frac{P(y_j)\prod_{k=1}^{p} P(x_k|y_j)}{P(x)} \qquad (2\text{-}4)$$

อย่างไรก็ตาม เนื่องจากตัวส่วน $P(x)$ เป็นค่าคงที่ที่ไม่ขึ้นอยู่กับคลาส (y_j) จึงทำหน้าที่เพียงเป็นตัวปรับสเกลของค่าความน่าจะเป็นเท่านั้น ในการจำแนกประเภทข้อมูลโดยใช้กฎ MAP จึงสามารถพิจารณาเฉพาะตัวเศษของสมการได้ ซึ่งทำให้กฎการจำแนกประเภทสามารถเขียนได้ตามสมการที่ (2-5)

$$\hat{y} = \arg\max_{y_j}\left(P(y_j)\prod_{k=1}^{p} P(x_k|y_j)\right) \qquad (2\text{-}5)$$

อัลกอริทึม Naïve Bayes ถูกนำไปประยุกต์ใช้อย่างแพร่หลายในหลายสาขา โดยเฉพาะงานที่เกี่ยวข้องกับการจำแนกประเภทข้อมูล (Classification) และการจัดอันดับข้อมูล (Ranking) ในงานจำแนกประเภท ระบบจะทำการกำหนดอินสแตนซ์ข้อมูลให้อยู่ในคลาสที่เหมาะสมตามค่าความน่าจะเป็นภายหลังที่คำนวณได้ ขณะที่ในงานจัดอันดับสามารถนำค่าความน่าจะเป็นภายหลังมาใช้เป็นคะแนนในการจัดลำดับความสำคัญของข้อมูล นอกจากนี้ ในสาขา ชีวสารสนเทศศาสตร์ (Bioinformatics) อัลกอริทึม Naïve Bayes ยังถูกนำไปใช้ในการจำแนกประเภทเนื้องอก (Tumor Classification) โดยอาศัยข้อมูลการแสดงของยีน (Gene Expression Profiling) เป็นข้อมูลชีวภาพที่มีมิติสูงและมีความซับซ้อน ข้อดีที่สำคัญของ Naïve Bayes คือความทนทานต่อการละเมิดสมมติฐาน (Robustness to Assumption Violation) กล่าวคือ แม้สมมติฐานความเป็นอิสระของคุณลักษณะจะไม่เป็นจริงในข้อมูลส่วนใหญ่ แต่แบบจำลองยังคงสามารถให้ผลลัพธ์ที่ดีได้ นอกจากนี้ อัลกอริทึมยังมีความสามารถในการจัดการกับชุดข้อมูลที่มีลักษณะตัวแปรแบบผสม (Mixed-Variable Data Sets) ได้อย่างมีประสิทธิภาพ รวมถึงสามารถจัดการกับค่าข้อมูลที่ขาดหายไป (Missing Values) ได้โดยการละเว้นคุณลักษณะที่ไม่มีข้อมูลในการคำนวณความน่าจะเป็นภายหลัง อีกทั้ง Naïve Bayes ยังมีประสิทธิภาพเชิงคำนวณ (Computational Efficiency) สูง เนื่องจากการคำนวณความน่าจะเป็น

ตามเงื่อนไขของแต่ละคุณลักษณะสามารถดำเนินการแบบขนาน (Parallel Processing) ได้ ทำให้เหมาะสมสำหรับการประมวลผลแบบกระจาย (Distributed Processing) เช่น ในสภาพแวดล้อมของ MapReduce และการวิเคราะห์ข้อมูลขนาดใหญ่ (Big Data Analytics)

อย่างไรก็ตาม ในบางสถานการณ์อาจเกิดปัญหาที่เรียกว่า ปัญหาความน่าจะเป็นศูนย์ (Zero Probability Problem) ซึ่งเกิดขึ้นเมื่อค่าคุณลักษณะของอินสแตนซ์ทดสอบไม่ปรากฏในข้อมูลฝึกสอนของคลาสใดคลาสหนึ่ง ส่งผลให้ค่าความน่าจะเป็นตามเงื่อนไขมีค่าเป็นศูนย์และทำให้ผลการคำนวณทั้งหมดเป็นศูนย์ตามไปด้วย เพื่อแก้ไขปัญหาดังกล่าวจึงนิยมใช้เทคนิค การปรับเรียบของลาปลาส (Laplace Smoothing) ซึ่งเป็นวิธีการปรับค่าความน่าจะเป็นให้ไม่เป็นศูนย์ โดยการเพิ่มค่าคงที่ขนาดเล็กเข้าไปในความถี่ของข้อมูล วิธีการดังกล่าวช่วยให้แบบจำลองสามารถทำงานได้อย่างเสถียรและลดผลกระทบจากข้อมูลตัวอย่างที่มีขนาดเล็กได้อย่างมีประสิทธิภาพ ด้วยคุณสมบัติด้านความเรียบง่าย ความรวดเร็วในการคำนวณ และความสามารถในการให้ผลลัพธ์ที่มีประสิทธิภาพ

2.1.3 อัลกอริทึมเพื่อนบ้านใกล้ (K-Nearest Neighbors)

การทำงานของอัลกอริทึมเพื่อนบ้านใกล้ (K-Nearest Neighbors) เป็นแนวทางหนึ่งที่ได้รับความนิยมในการสร้างแบบจำลองสำหรับการทำนายค่าเป้าหมาย วิธีการดังกล่าวมีหลักการสำคัญคือการพิจารณาความใกล้เคียงของตัวอย่างข้อมูล (instances) เพื่อใช้ข้อมูลของตัวอย่างที่มีลักษณะคล้ายกันมากที่สุดมาช่วยในการทำนายค่าของตัวอย่างใหม่ ในบริบทของระบบแนะนำ สามารถพิจารณาความสัมพันธ์ระหว่างผู้ใช้ (user) และรายการข้อมูล (item) เช่น บทความ ผลิตภัณฑ์ หรือเว็บไซต์ โดยแทนข้อมูลดังกล่าวในรูปของตารางที่ 2-2 ซึ่งเป็นเมทริกซ์ที่บ่งชี้ว่าผู้ใช้แต่ละรายมีปฏิสัมพันธ์กับรายการใดบ้างเช่น ค่า 1 คือผู้ใช้มีปฏิสัมพันธ์กับรายการนั้น และค่า 0 คือไม่มี

ตารางที่ 2-2 เมทริกซ์ผู้ใช้–รายการ (User / Item Matrix)

ผู้ใช้	รายการ1	รายการ2	รายการ3	รายการ4	รายการ5	รายการ6	รายการ7
ผู้ใช้ 1	0	0	1	1	0	1	1
ผู้ใช้ 2	1	1	0	0	1	1	0
ผู้ใช้ 3	1	0	0	0	0	1	0
ผู้ใช้ 4	0	0	0	0	0	0	0
ผู้ใช้ 5	1	1	1	1	1	1	1

ที่มา : (Brink et al., 2017:224)

เมื่อสร้างเมทริกซ์ผู้ใช้–รายการแล้ว ขั้นตอนถัดไปคือการคำนวณค่าความไม่คล้ายคลึงกัน (dissimilarity) หรือระยะห่าง (distance) ระหว่างรายการแต่ละคู่ โดยพิจารณาจากรูปแบบการโต้ตอบของผู้ใช้ในเมทริกซ์ดังกล่าว หากรายการสองรายการมีรูปแบบผู้ใช้ที่โต้ตอบคล้ายกัน ค่าระยะห่างจะมีค่าต่ำ แต่หากมีรูปแบบการใช้งานแตกต่างกันมาก ค่าระยะห่างจะสูงขึ้น ในการศึกษานี้เลือกใช้ Euclidean distance เป็นตัววัดระยะห่าง เนื่องจากสามารถสะท้อนระดับความแตกต่างของเวกเตอร์ข้อมูลได้อย่างมีประสิทธิภาพแสดงตามตารางที่ 2-3

ตารางที่ 2-3 เมทริกซ์ความไม่คล้ายคลึง (Dissimilarity Matrix)

	รายการ1	**รายการ2**	**รายการ3**	**รายการ3**	**รายการ5**	**รายการ6**	**รายการ7**
รายการ1	0.0	1.0	1.7	1.7	1.0	1.0	1.7
รายการ2	1.0	0.0	1.4	1.4	0.0	1.4	1.4
รายการ3	1.7	1.4	0.0	0.0	1.4	1.4	0.0
รายการ4	1.7	1.4	0.0	0.0	1.4	1.4	0.0
รายการ5	1.0	0.0	1.4	1.4	0.0	1.4	1.4
รายการ6	1.0	1.4	1.4	1.4	1.4	0.0	1.4
รายการ7	1.7	1.4	0.0	0.0	1.4	1.4	0.0

ที่มา : (Brink et al., 2017:224)

ภายหลังจากการคำนวณระยะห่างระหว่างรายการทุกคู่แล้ว ขั้นตอนสำคัญของอัลกอริทึมเพื่อนบ้านใกล้ คือการเลือกจำนวนเพื่อนบ้านที่ใกล้ที่สุด หรือค่าพารามิเตอร์ k เพื่อใช้ในการทำนายค่าเป้าหมายของรายการที่ต้องการพิจารณา หลักการพื้นฐานคือ การค้นหารายการที่มีค่าระยะห่างน้อยที่สุดจำนวน k รายการ จากนั้นจึงนำค่าของตัวแปรเป้าหมายของรายการเหล่านั้นมาคำนวณค่าเฉลี่ยเพื่อใช้เป็นค่าทำนาย อย่างไรก็ตาม ในทางปฏิบัติมักมีการกำหนดน้ำหนักให้กับเพื่อนบ้านแต่ละรายการเพื่อสะท้อนความสำคัญของความใกล้เคียง โดยรายการที่มีระยะห่างน้อยกว่าจะได้รับน้ำหนักมากกว่า วิธีการให้น้ำหนักที่นิยมใช้ ได้แก่ การให้น้ำหนักเท่ากันทั้งหมด (Equal weights) การให้น้ำหนักผกผันกับระยะห่าง $\frac{1}{d}$ และการให้น้ำหนักผกผันกับกำลังสองของระยะห่าง $\frac{1^2}{d}$ ซึ่งช่วยให้รายการที่มีความใกล้เคียงมากมีอิทธิพลต่อผลการทำนายมากขึ้น ในการประเมินประสิทธิภาพของแบบจำลอง จะทำการทดลองเปรียบเทียบค่าพารามิเตอร์ k หลายค่า โดยคำนวณค่าความคลาดเคลื่อนของการทำนายด้วยตัวชี้วัด Root Mean Square Error (RMSE) ที่ถูกปรับให้เป็น

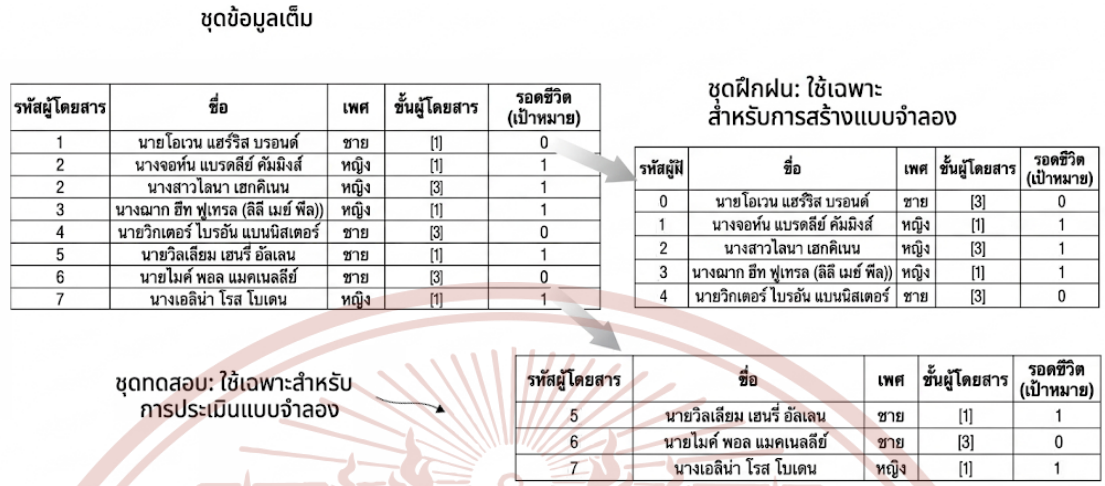
ค่ามาตรฐาน (Normalized RMSE) การทดลองจะคำนวณค่า RMSE สำหรับช่วงค่าขอ k หลายค่า เพื่อวิเคราะห์ว่าจำนวนเพื่อนบ้านที่เหมาะสมควรมีค่าเท่าใด ผลการทดลองโดยทั่วไปพบว่า เมื่อค่าของ k มีขนาดเล็ก เช่น 1 หรือ 2 เพื่อนบ้าน แบบจำลองมักให้ผลการทำนายที่มีความแม่นยำสูงกว่า เนื่องจากข้อมูลในระบบแนะนำมักมีลักษณะกระจัดกระจาย (sparse data) ทำให้รายการที่อยู่ใกล้กันจริงมีจำนวนไม่มาก การเพิ่มจำนวนเพื่อนบ้านมากเกินไปอาจทำให้ข้อมูลที่มีความแตกต่างกันถูกนำมารวมในการคำนวณค่าเฉลี่ย ส่งผลให้ความแม่นยำลดลงเล็กน้อย นอกจากนี้ยังพบว่าการให้น้ำหนักตามระยะห่าง เช่น $\frac{1}{d}$ หรือ $\frac{1}{d}^2$ สามารถช่วยปรับปรุงประสิทธิภาพของแบบจำลองได้ดีกว่าการให้น้ำหนักเท่ากัน เนื่องจากสะท้อนระดับความคล้ายคลึงของข้อมูลได้เหมาะสมกว่า

จากการวิเคราะห์ดังกล่าวสามารถสรุปได้ว่าอัลกอริทึมเพื่อนบ้านใกล้เป็นวิธีที่มีความเรียบง่าย แต่มีประสิทธิภาพสำหรับการทำนายค่าในระบบแนะนำ โดยอาศัยแนวคิดของความคล้ายคลึงกันระหว่างรายการผ่านเมทริกซ์ผู้ใช้–รายการและการคำนวณระยะห่างระหว่างข้อมูล อย่างไรก็ตาม การเลือกค่าพารามิเตอร์ k และรูปแบบการให้น้ำหนักมีผลต่อประสิทธิภาพของแบบจำลองอย่างมีนัยสำคัญ ดังนั้นการทดลองปรับค่าพารามิเตอร์และเปรียบเทียบผลลัพธ์ด้วยตัวชี้วัดที่เหมาะสม เช่น RMSE จึงเป็นขั้นตอนสำคัญในการพัฒนาแบบจำลองให้มีความแม่นยำสูงสุด

2.1.4 การวัดประสิทธิภาพแบบจำลอง (Mode Evaluation)

การวัดประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model Evaluation) เป็นขั้นตอนสำคัญในกระบวนการพัฒนาแบบจำลอง เนื่องจากแบบจำลองที่ถูกสร้างขึ้นจากข้อมูลฝึกสอนจำเป็นต้องได้รับการตรวจสอบว่าสามารถทำนายผลลัพธ์ได้อย่างถูกต้องและมีความสามารถในการนำไปใช้กับข้อมูลใหม่ที่ไม่เคยปรากฏมาก่อน กระบวนการประเมินแบบจำลองจึงมีบทบาทสำคัญในการพิจารณาความแม่นยำ ความสามารถในการอธิบายข้อมูล และความสามารถทั่วไป (Generalization) ของแบบจำลอง

ก่อนดำเนินการประเมินแบบจำลอง ข้อมูลทั้งหมดมักถูกแบ่งออกเป็นสองส่วน ได้แก่ ชุดข้อมูลสำหรับการฝึกแบบจำลอง (Training Set) และชุดข้อมูลสำหรับการทดสอบแบบจำลอง (Testing Set) การแบ่งข้อมูลในลักษณะดังกล่าวช่วยให้สามารถประเมินความสามารถของแบบจำลองในการทำนายข้อมูลที่ไม่เคยใช้ในการฝึกมาก่อน ซึ่งช่วยลดความเสี่ยงของปัญหาโอเวอร์ฟิตติ้ง (Overfitting) ที่เกิดขึ้นเมื่อแบบจำลองสามารถเรียนรู้รายละเอียดของข้อมูลฝึกสอนได้ดีเกินไป แต่ไม่สามารถนำความรู้ดังกล่าวไปใช้กับข้อมูลใหม่ได้อย่างมีประสิทธิภาพ ภาพที่ 2-3 แสดงแนวคิดการแบ่งชุดข้อมูลออกเป็น Training Set และ Testing Set โดยแบบจำลองจะถูกสร้างจากข้อมูลใน Training Set และนำไปทดสอบกับข้อมูลใน Testing Set เพื่อประเมินผลการทำนาย



รหัสผู้โดยสาร	ชื่อ	เพศ	ชั้นผู้โดยสาร	รอดชีวิต (เป้าหมาย)
1	นายโอเวน แฮร์ริส บรอนด์	ชาย	[1]	0
2	นางจอห์น แบรดลีย์ คัมมิงส์	หญิง	[1]	1
2	นางสาวไลนา เฮกคิเนน	หญิง	[3]	1
3	นางฌาก ฮีท ฟูเทรล (ลิลี เมย์ พีล))	หญิง	[1]	1
4	นายวิกเตอร์ ไบรอัน แบนนิสเตอร์	ชาย	[3]	0
5	นายวิลเลียม เฮนรี่ อัลเลน	ชาย	[1]	1
6	นายไมค์ พอล แมคเนลลีย์	ชาย	[3]	0
7	นางเอลิน่า โรส โบเดน	หญิง	[1]	1

รหัสผู้ฝึ	ชื่อ	เพศ	ชั้นผู้โดยสาร	รอดชีวิต (เป้าหมาย)
0	นายโอเวน แฮร์ริส บรอนด์	ชาย	[3]	0
1	นางจอห์น แบรดลีย์ คัมมิงส์	หญิง	[1]	1
2	นางสาวไลนา เฮกคิเนน	หญิง	[3]	1
3	นางฌาก ฮีท ฟูเทรล (ลิลี เมย์ พีล))	หญิง	[1]	1
4	นายวิกเตอร์ ไบรอัน แบนนิสเตอร์	ชาย	[3]	0

รหัสผู้โดยสาร	ชื่อ	เพศ	ชั้นผู้โดยสาร	รอดชีวิต (เป้าหมาย)
5	นายวิลเลียม เฮนรี่ อัลเลน	ชาย	[1]	1
6	นายไมค์ พอล แมคเนลลีย์	ชาย	[3]	0
7	นางเอลิน่า โรส โบเดน	หญิง	[1]	1

ภาพที่ 2-3 การแบ่งชุดข้อมูลออกเป็นชุดข้อมูลเพื่อใช้ในการประเมินประสิทธิภาพของแบบจำลอง

สำหรับปัญหาการจำแนกประเภท (Classification) การประเมินประสิทธิภาพของแบบจำลองมักเริ่มต้นจากการพิจารณาค่าความแม่นยำ (Accuracy) ซึ่งเป็นสัดส่วนของจำนวนตัวอย่างที่แบบจำลองทำนายได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด ค่าความแม่นยำสามารถคำนวณได้จากสมการดังต่อไปนี้

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2\text{-}6)$$

โดยที่

TP (True Positive) หมายถึงจำนวนตัวอย่างที่ทำนายเป็นคลาสบวกได้ถูกต้อง

TN (True Negative) หมายถึงจำนวนตัวอย่างที่ทำนายเป็นคลาสลบได้ถูกต้อง

FP (False Positive) หมายถึงจำนวนตัวอย่างที่ทำนายเป็นคลาสบวกทั้งที่เป็นคลาสลบ

FN (False Negative) หมายถึงจำนวนตัวอย่างที่ทำนายเป็นคลาสลบทั้งที่เป็นคลาสบวก

นอกจากตัวชี้วัดค่าความแม่นยำ (Accuracy) แล้ว ในงานจำแนกประเภทข้อมูล (Classification) ยังมีตัวชี้วัดที่สำคัญอีกหลายตัวที่ช่วยสะท้อนประสิทธิภาพของแบบจำลองได้อย่างละเอียดมากขึ้น โดยเฉพาะในกรณีที่ชุดข้อมูลมีความไม่สมดุลระหว่างจำนวนตัวอย่างของแต่ละคลาส ตัวชี้วัดที่นิยมใช้ร่วมกัน ได้แก่ Precision, Recall และ F1-score ซึ่งสามารถคำนวณได้จากค่าที่ปรากฏใน Confusion Matrix

ค่า Precision เป็นตัวชี้วัดที่แสดงถึงสัดส่วนของตัวอย่างที่แบบจำลองทำนายว่าเป็นคลาสบวกและมีความถูกต้องจริงเมื่อเทียบกับจำนวนตัวอย่างทั้งหมดที่ถูกทำนายว่าเป็นคลาสบวก โดยสามารถคำนวณได้จากสมการดังต่อไปนี้

$$Precision = \frac{TP}{TP+FP} \quad (2\text{-}7)$$

ในขณะที่ค่า Recall หรือที่เรียกว่า True Positive Rate หรือ Sensitivity เป็นตัวชี้วัดที่แสดงถึงความสามารถของแบบจำลองในการตรวจจับตัวอย่างที่เป็นคลาสบวกจริง โดยคำนวณจากสัดส่วนของจำนวนตัวอย่างที่ถูกทำนายเป็นคลาสบวกอย่างถูกต้องเมื่อเทียบกับจำนวนตัวอย่างคลาสบวกทั้งหมด ดังแสดงในสมการต่อไปนี้

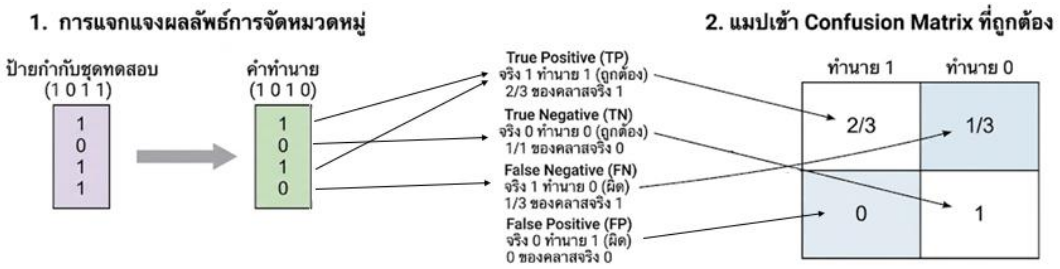
$$Recall = \frac{TP}{TP+FN} \quad (2\text{-}8)$$

อย่างไรก็ตาม ในบางกรณีค่า Precision และ Recall อาจให้ผลลัพธ์ที่แตกต่างกัน กล่าวคือ แบบจำลองอาจมีค่า Precision สูงแต่มีค่า Recall ต่ำ หรือในทางกลับกัน ดังนั้นจึงมีการนำตัวชี้วัดทั้งสองมารวมกันในรูปของค่า F1-score ซึ่งเป็นค่าเฉลี่ยฮาร์มอนิก (Harmonic Mean) ของค่า Precision และ Recall เพื่อให้สามารถประเมินประสิทธิภาพของแบบจำลองได้อย่างสมดุลมากขึ้น ค่า F1-score สามารถคำนวณได้จากสมการดังต่อไปนี้

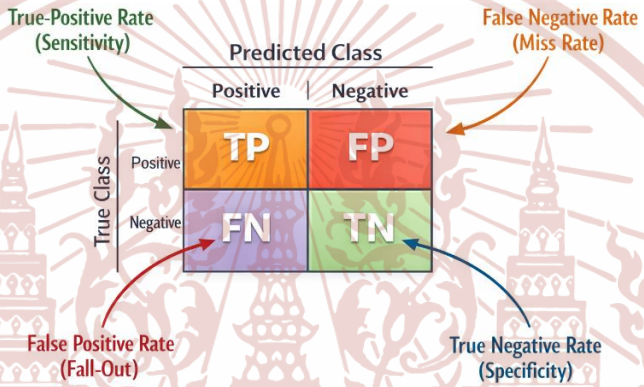
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2\text{-}9)$$

ค่า F1-score มีค่าระหว่าง 0 ถึง 1 โดยค่าที่เข้าใกล้ 1 แสดงว่าแบบจำลองมีประสิทธิภาพสูงทั้งในด้านความแม่นยำของการทำนายคลาสบวกและความสามารถในการตรวจจับตัวอย่างคลาสบวกทั้งหมด ตัวชี้วัดนี้จึงถูกนำมาใช้อย่างแพร่หลายในงานวิจัยด้านการเรียนรู้ของเครื่อง โดยเฉพาะในกรณีที่ชุดข้อมูลมีความไม่สมดุล เช่น งานตรวจจับสแปม การตรวจจับการทุจริต หรือการวินิจฉัยโรค ดังนั้น ในการประเมินประสิทธิภาพของแบบจำลองสำหรับปัญหาการจำแนกประเภท จึงควรพิจารณาตัวชี้วัดหลายประเภทประกอบกัน ได้แก่ Accuracy, Precision, Recall และ F1-score เพื่อให้สามารถวิเคราะห์ความสามารถของแบบจำลองได้อย่างครอบคลุมและสะท้อนประสิทธิภาพที่แท้จริงของแบบจำลองได้ดียิ่งขึ้น

อย่างไรก็ตาม กรณีที่ข้อมูลมีความไม่สมดุลระหว่างจำนวนตัวอย่างในแต่ละคลาส ดังนั้นจึงมีการใช้เครื่องมือที่เรียกว่าเมทริกซ์ความสับสน (Confusion Matrix) เพื่อแสดงจำนวนตัวอย่างที่แบบจำลองทำนายถูกต้องและผิดพลาดในแต่ละประเภทของผลลัพธ์ ดังแสดงในภาพที่ ภาพที่ 2-4 ซึ่งช่วยให้นักวิจัยสามารถวิเคราะห์ประสิทธิภาพของแบบจำลองได้อย่างละเอียดมากขึ้น จากค่าที่ปรากฏใน Confusion Matrix สามารถนำไปคำนวณตัวชี้วัดอื่น ๆ ที่สำคัญ เช่น Sensitivity (True Positive Rate) และ Specificity (True Negative Rate) ดังแสดงในภาพที่ 2-5 ซึ่งเป็นตัวชี้วัดที่ช่วยประเมินความสามารถของแบบจำลองในการตรวจจับเหตุการณ์ที่สนใจได้อย่างมีประสิทธิภาพ

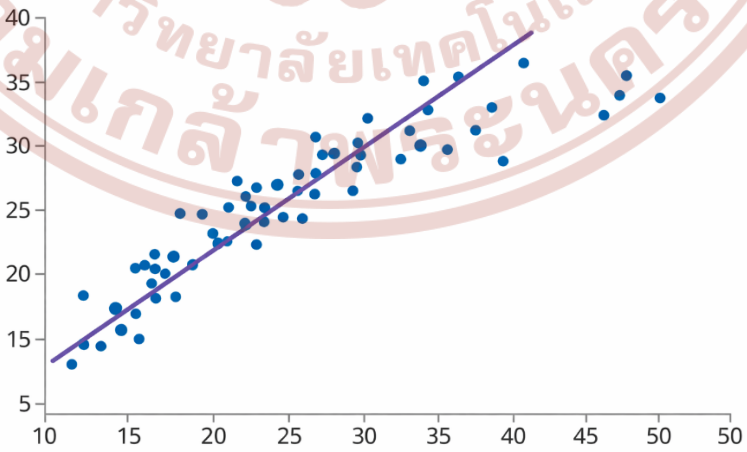


ภาพที่ 2-4 โครงสร้างของเมทริกซ์ความสับสน (Confusion Matrix)



ภาพที่ 2-5 ตัวชี้วัดประสิทธิภาพ Sensitivity, Specificity, Fall-out และ Miss rate

ในส่วนของการประเมินประสิทธิภาพของแบบจำลอง Regression สามารถใช้กราฟกระจาย (Scatter Plot) เพื่อแสดงความสัมพันธ์ระหว่างค่าที่แบบจำลองทำนายได้กับค่าจริง ดังแสดงในภาพที่ 2-6 หากแบบจำลองมีประสิทธิภาพสูง จุดข้อมูลส่วนใหญ่จะกระจายตัวอยู่ใกล้กับเส้นทแยงมุมซึ่งแสดงถึงกรณีที่ค่าทำนายเท่ากับค่าจริง อย่างไรก็ตาม การประเมินผลด้วยภาพเพียงอย่างเดียวอาจไม่เพียงพอ จึงจำเป็นต้องใช้ตัวชี้วัดเชิงปริมาณเพิ่มเติม

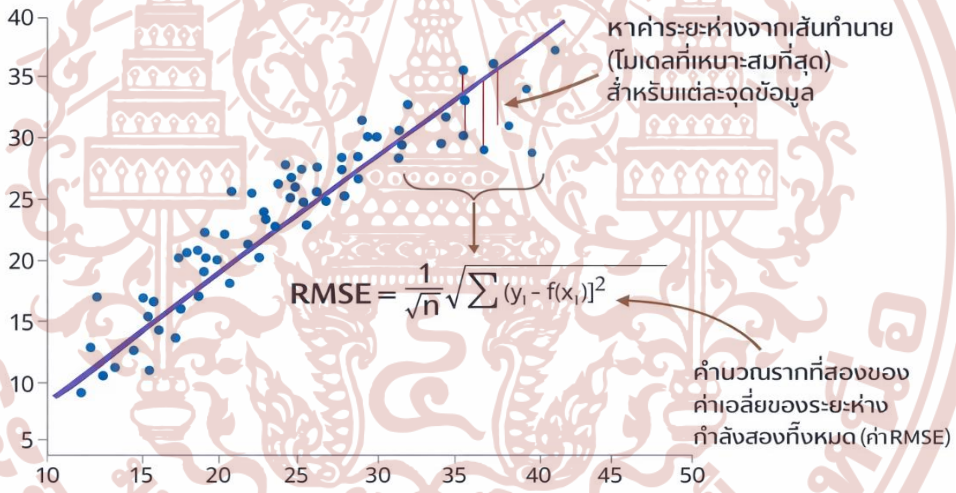


ภาพที่ 2-6 กราฟกระจายแสดงความสัมพันธ์ระหว่างค่าที่ทำนายได้กับค่าจริงของตัวแปรเป้าหมาย

หนึ่งในตัวชี้วัดที่นิยมใช้สำหรับปัญหา Regression คือค่า Root Mean Square Error (RMSE) ซึ่งเป็นตัวชี้วัดที่แสดงค่าความคลาดเคลื่อนเฉลี่ยของการทำนาย โดยพิจารณาจากกำลังสองของความแตกต่างระหว่างค่าจริงและค่าที่แบบจำลองทำนายได้ ค่า RMSE สามารถคำนวณได้จากสมการดังต่อไปนี้

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2\text{-}10)$$

โดยที่ y_i หมายถึงค่าจริงของตัวแปรเป้าหมาย $\hat{y}_i$ หมายถึงค่าที่แบบจำลองทำนายได้ และ n หมายถึงจำนวนตัวอย่างทั้งหมด ดังแสดงแนวคิดของการคำนวณค่า RMSE ในภาพที่ 2-7 ซึ่งแสดงระยะห่างระหว่างค่าทำนายและค่าจริงของแต่ละตัวอย่างในกราฟกระจาย ค่า RMSE ที่มีค่าน้อยแสดงว่าแบบจำลองสามารถทำนายค่าได้ใกล้เคียงกับค่าจริงมากขึ้น



ภาพที่ 2-7 การคำนวณค่า RMSE จากระยะห่างระหว่างค่าทำนายและค่าจริงของแต่ละตัวอย่าง

นอกจาก RMSE แล้ว ยังมีตัวชี้วัดที่สำคัญอีกตัวหนึ่งสำหรับการประเมินแบบจำลอง Regression คือค่าสัมประสิทธิ์การกำหนด หรือ R^2 (Coefficient of Determination) ซึ่งใช้วัดความสามารถของแบบจำลองในการอธิบายความแปรปรวนของตัวแปรเป้าหมาย ค่า R^2 สามารถคำนวณได้จากสมการดังต่อไปนี้

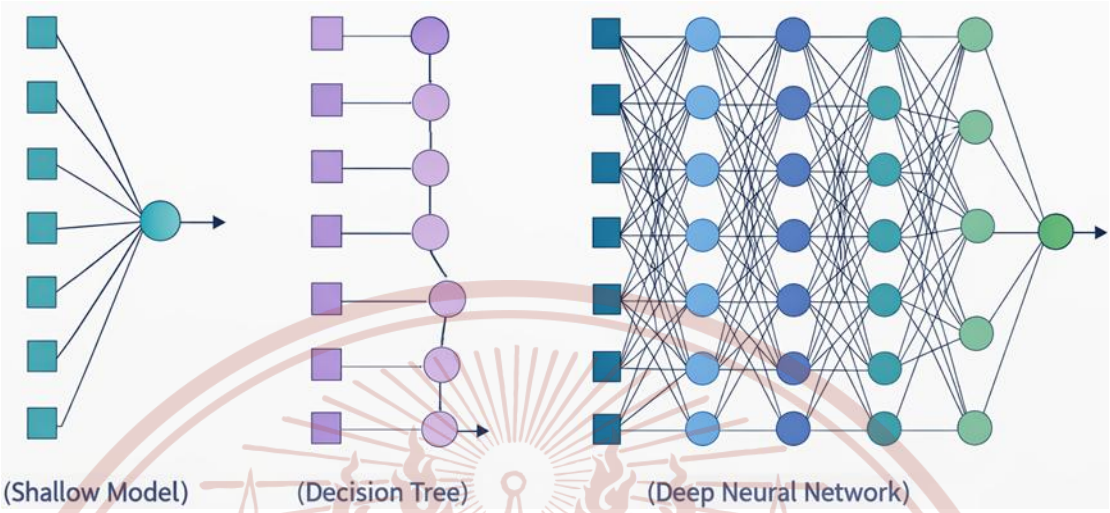
$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (2\text{-}11)$$

โดยที่ $\bar{y}$ หมายถึงค่าเฉลี่ยของค่าจริงทั้งหมด ค่า R^2 มีค่าระหว่าง 0 ถึง 1 โดยค่าที่เข้าใกล้ 1 แสดงว่าแบบจำลองสามารถอธิบายความสัมพันธ์ของข้อมูลได้ดี

โดยสรุป การประเมินประสิทธิภาพของแบบจำลองเป็นขั้นตอนสำคัญที่ช่วยให้สามารถตรวจสอบความเหมาะสมของแบบจำลองก่อนนำไปใช้งานจริง การเลือกใช้ตัวชี้วัดที่เหมาะสมกับลักษณะของปัญหา รวมถึงการแบ่งข้อมูลสำหรับการฝึกและการทดสอบอย่างถูกต้อง จะช่วยให้ผลการประเมินมีความน่าเชื่อถือและสะท้อนความสามารถที่แท้จริงของแบบจำลองได้อย่างมีประสิทธิภาพ

2.2 การเรียนรู้เชิงลึก

การเรียนรู้เชิงลึก (Deep Learning) เป็นแขนงหนึ่งของการเรียนรู้ของเครื่อง (Machine Learning) ที่มุ่งเน้นการสร้างแบบจำลองเชิงคำนวณซึ่งมีโครงสร้างคล้ายโครงข่ายประสาทเทียมหลายชั้น (Artificial Neural Networks) โดยแนวคิดหลักของวิธีการนี้คือการเรียนรู้ความสัมพันธ์ที่ซับซ้อนของข้อมูลผ่านการจัดเรียงหน่วยประมวลผลจำนวนมากเป็นลำดับชั้น (layers) ตั้งแต่ชั้นรับข้อมูล (input layer) ชั้นซ่อน (hidden layers) ไปจนถึงชั้นผลลัพธ์ (output layer) การเพิ่มจำนวนชั้นของเครือข่ายทำให้แบบจำลองสามารถเรียนรู้รูปแบบเชิงนามธรรมของข้อมูลได้ลึกยิ่งขึ้น จึงเป็นที่มาของคำว่า “Deep” ใน Deep Learning ซึ่งสะท้อนถึงความลึกของโครงสร้างเครือข่ายที่ใช้ในการคำนวณโดยทั่วไป Deep Learning สามารถอธิบายได้ว่าเป็นแบบจำลองเชิงคณิตศาสตร์ที่ประกอบด้วยหน่วยประมวลผลจำนวนมากเชื่อมต่อกันผ่านค่าน้ำหนัก (weights) ซึ่งเป็นพารามิเตอร์ที่ถูกปรับค่าในระหว่างกระบวนการฝึกสอน (training process) เพื่อให้แบบจำลองสามารถประมาณค่าความสัมพันธ์ระหว่างตัวแปรนำเข้าและผลลัพธ์ได้อย่างเหมาะสม แนวคิดพื้นฐานของโครงข่ายประสาทเทียมมีต้นกำเนิดจากความพยายามในการจำลองการทำงานของเซลล์ประสาทในสมองมนุษย์ โดยแต่ละหน่วยประมวลผลในเครือข่ายจะรับข้อมูลจากหน่วยก่อนหน้า คำนวณผลรวมถ่วงน้ำหนักและส่งค่าที่ผ่านฟังก์ชันไม่เชิงเส้นไปยังหน่วยถัดไป กระบวนการดังกล่าวทำให้เครือข่ายสามารถสร้างแบบจำลองที่มีความสามารถในการแสดงความสัมพันธ์เชิงซับซ้อนของข้อมูลได้ดีกว่าวิธีเชิงเส้นทั่วไป ในเชิงแนวคิด สามารถเปรียบเทียบรูปแบบของแบบจำลองการเรียนรู้ได้หลายลักษณะ โดยแบบจำลองอย่างง่าย เช่น การถดถอยเชิงเส้น (Linear Regression) จะมีเส้นทางการคำนวณระหว่างข้อมูลนำเข้าและผลลัพธ์ที่สั้นมาก เนื่องจากตัวแปรนำเข้าแต่ละตัวมีผลต่อผลลัพธ์โดยตรงผ่านพารามิเตอร์เพียงตัวเดียว ส่งผลให้ความสามารถในการแสดงความสัมพันธ์ของข้อมูลมีข้อจำกัด ขณะที่แบบจำลองประเภทต้นไม้ตัดสินใจ (Decision Tree) สามารถสร้างเส้นทางการคำนวณที่ยาวขึ้นได้ในบางกรณี แต่ก็ยังมีข้อจำกัดในด้านจำนวนเส้นทางที่ครอบคลุมพื้นที่ข้อมูลทั้งหมด ในทางตรงกันข้าม โครงข่าย Deep Learning สามารถสร้างเส้นทางการคำนวณที่ยาวและซับซ้อนมากขึ้น ทำให้ตัวแปรนำเข้าหลายตัวสามารถมีปฏิสัมพันธ์กันผ่านหลายชั้นของการคำนวณ ซึ่งช่วยเพิ่มความสามารถในการเรียนรู้รูปแบบที่ซับซ้อนของข้อมูลได้อย่างมีประสิทธิภาพ ดังแสดงแนวคิดการเปรียบเทียบโครงสร้างแบบจำลองในภาพที่ 2-8



ภาพที่ 2-8 การเปรียบเทียบความลึกของแบบจำลองการเรียนรู้

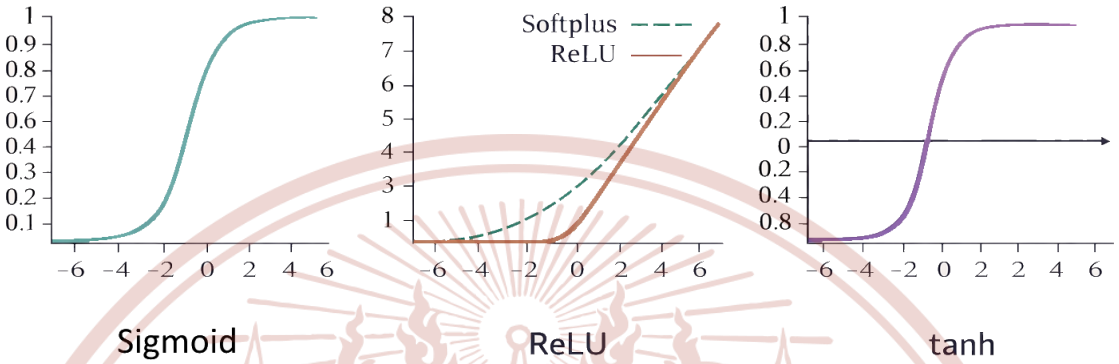
หนึ่งในโครงสร้างพื้นฐานของ Deep Learning คือโครงข่ายประสาทเทียมแบบป้อนข้อมูลไปข้างหน้า (Feedforward Neural Network) ซึ่งเป็นเครือข่ายที่มีการเชื่อมต่อของข้อมูลในทิศทางเดียวจากชั้นนำเข้าไปยังชั้นผลลัพธ์โดยไม่มีการย้อนกลับของสัญญาณ โครงสร้างดังกล่าวสามารถอธิบายได้ในรูปแบบกราฟแบบไม่มีวงจร (Directed Acyclic Graph) โดยแต่ละโหนดในเครือข่ายจะคำนวณผลรวมถ่วงน้ำหนักของข้อมูลจากโหนดก่อนหน้า จากนั้นจึงนำผลลัพธ์ดังกล่าวผ่านฟังก์ชันกระตุ้น (Activation Function) เพื่อสร้างค่าผลลัพธ์ของโหนดนั้น

ในเชิงคณิตศาสตร์ หากกำหนดให้ x เป็นเวกเตอร์ของข้อมูลนำเข้า และ w เป็นเวกเตอร์ของค่าน้ำหนักที่เชื่อมต่อเข้าสู่หน่วยประมวลผลหนึ่งหน่วย การคำนวณของหน่วยประมวลผลสามารถเขียนได้ในรูป

$$a_j = g(w^T x) \tag{2-12}$$

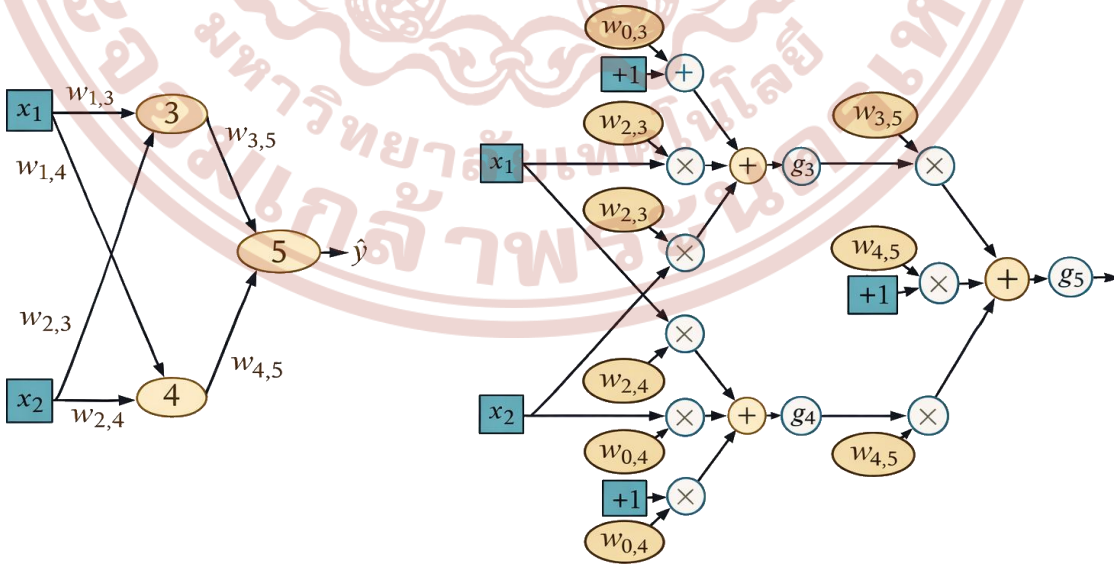
โดยที่ $g(\cdot)$ แทนฟังก์ชันกระตุ้น ซึ่งมีบทบาทสำคัญในการเพิ่มความไม่เชิงเส้น (nonlinearity) ให้กับแบบจำลอง หากไม่มีฟังก์ชันกระตุ้น การเชื่อมต่อของหลายชั้นในเครือข่ายจะยังคงเทียบเท่ากับแบบจำลองเชิงเส้นเพียงชั้นเดียว ซึ่งไม่สามารถอธิบายความสัมพันธ์ที่ซับซ้อนของข้อมูลจริงได้ฟังก์ชันกระตุ้นที่นิยมใช้ในระบบ Deep Learning มีหลายประเภท โดยแต่ละประเภทมีคุณสมบัติทางคณิตศาสตร์และลักษณะการตอบสนองที่แตกต่างกัน ตัวอย่างเช่น ฟังก์ชันซิกมอยด์ (Sigmoid Function) ซึ่งให้ค่าผลลัพธ์อยู่ในช่วงระหว่าง 0 ถึง 1 เหมาะสำหรับการประมาณค่าความน่าจะเป็น ฟังก์ชัน Rectified Linear Unit หรือ ReLU ซึ่งให้ผลลัพธ์เป็นศูนย์เมื่อค่าป้อนเข้าติดลบ และให้ค่าเท่ากับอินพุตเมื่อเป็นค่าบวก ทำให้สามารถลดปัญหาการหายไปของกราเดียนต์ (vanishing gradient) ในเครือข่ายลึกได้ นอกจากนี้ยังมีฟังก์ชัน tanh ซึ่งมีช่วงค่าระหว่าง -1 ถึง 1 และมีสมมาตร

รอบศูนย์ ทำให้เหมาะกับงานบางประเภทที่ต้องการการกระจายของค่าข้อมูลแบบสมดุล ดังแสดงลักษณะของฟังก์ชันกระตุ้นในภาพที่ 2-9



ภาพที่ 2-9 ตัวอย่างฟังก์ชันกระตุ้นที่ใช้ในโครงข่ายประสาทเทียม Sigmoid, ReLU และ Tanh

เมื่อหน่วยประมวลผลหลายหน่วยถูกเชื่อมต่อเข้าด้วยกันเป็นชั้นต่าง ๆ จะเกิดเป็นโครงสร้างเครือข่ายที่สามารถคำนวณฟังก์ชันที่ซับซ้อนมากขึ้นได้ โดยการคำนวณทั้งหมดสามารถอธิบายได้ในรูปแบบของกราฟการคำนวณ (Computation Graph) ซึ่งแสดงลำดับขั้นของการดำเนินการทางคณิตศาสตร์ภายในเครือข่าย โครงสร้างดังกล่าวช่วยให้สามารถวิเคราะห์กระบวนการคำนวณและการไหลของข้อมูลภายในแบบจำลองได้อย่างชัดเจน ตัวอย่างของโครงข่ายประสาทเทียมที่ประกอบด้วยชั้นนำเข้า ชั้นซ่อน และชั้นผลลัพธ์ รวมถึงกราฟการคำนวณของเครือข่าย แสดงไว้ในภาพที่ 2-10

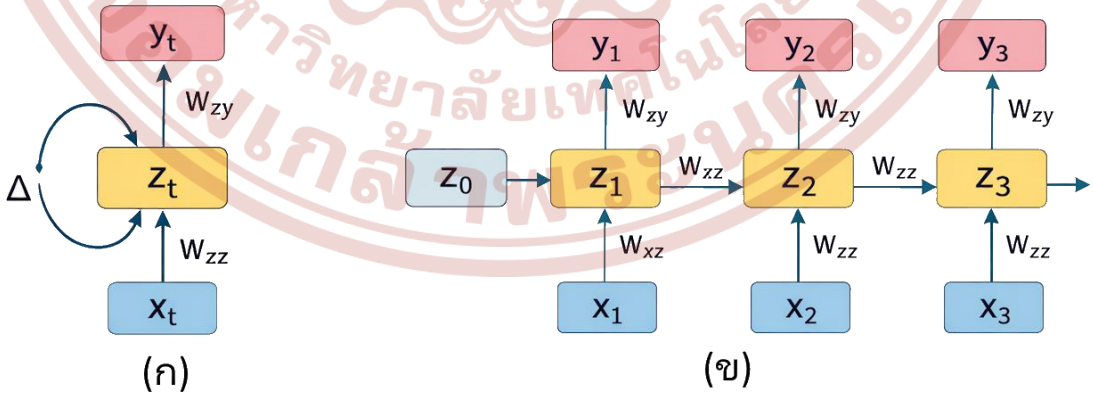


ภาพที่ 2-10 โครงสร้างของโครงข่ายประสาทเทียมและกราฟการคำนวณของเครือข่าย

2.2.1 โครงสร้างประสาทเทียมแบบวนซ้ำ (Recurrent Neural Networks)

โครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Networks หรือ RNN) เป็นแบบจำลองการเรียนรู้ของเครื่องที่ถูกออกแบบมาเพื่อจัดการกับข้อมูลที่มีลักษณะเป็นลำดับ (Sequential Data) ซึ่งแตกต่างจากโครงข่ายประสาทเทียมแบบป้อนตรง (Feedforward Neural Networks) ที่การไหลของข้อมูลเกิดขึ้นในทิศทางเดียวโดยไม่มีการย้อนกลับของสถานะภายใน โครงสร้างของ RNN มีการเชื่อมต่อแบบวนซ้ำภายในชั้นซ่อนเร้น ทำให้แบบจำลองสามารถรับข้อมูลจากผลลัพธ์ในช่วงเวลาก่อนหน้าและนำมาใช้ร่วมกับข้อมูลปัจจุบันในการคำนวณผลลัพธ์ได้ คุณสมบัตินี้ทำให้ RNN มีความสามารถในการจดจำข้อมูลจากอดีต หรือที่เรียกว่า “หน่วยความจำภายใน” (internal memory) ซึ่งมีความสำคัญ อย่างยิ่งสำหรับการประมวลผลข้อมูลที่มีลำดับเวลา เช่น การประมวลผลภาษา การรู้จำเสียง และการวิเคราะห์ข้อมูลอนุกรมเวลา

โครงสร้างพื้นฐานของ RNN ประกอบด้วยชั้นนำเข้า (Input layer) ชั้นซ่อนเร้น (Hidden Layer) และชั้นผลลัพธ์ (output layer) โดยชั้นซ่อนเร้นจะมีการเชื่อมต่อแบบวนซ้ำ (Recurrent Connection) ซึ่งทำให้สถานะของโหนดในช่วงเวลาปัจจุบันขึ้นอยู่กับทั้งข้อมูลนำเข้าในปัจจุบันและสถานะจากช่วงเวลาก่อนหน้า แนวคิดของโครงสร้างดังกล่าวสามารถอธิบายได้ดังแสดงในภาพที่ 2-11(ก) ซึ่งแสดงหน่วยประสาทที่มีการป้อนค่ากลับผ่านตัวหน่วงเวลา (Delay) เพื่อแทนค่าจากสถานะก่อนหน้า ทั้งนี้ เพื่อให้เข้าใจการทำงานของโครงข่ายได้ชัดเจนยิ่งขึ้น โครงข่าย RNN สามารถถูกแปลงให้อยู่ในรูปของโครงข่ายแบบป้อนตรงโดยการแตกออกตามเวลา (Unrolling Through Time) ดังแสดงใน ภาพที่ 2-11(ข) ซึ่งแต่ละช่วงเวลาจะถูกแทนด้วยชั้นของโครงข่ายที่มีโครงสร้างเหมือนกัน และใช้พารามิเตอร์ร่วมกันในทุกช่วงเวลา (shared weights)



ภาพที่ 2-11 โครงสร้างของ Recurrent Neural Network

ในเชิงคณิตศาสตร์ การทำงานของ RNN ในแต่ละช่วงเวลา t สามารถอธิบายได้ผ่านสมการของสถานะซ่อนเร้นและสมการของผลลัพธ์ โดยให้ x_t แทนเวกเตอร์ข้อมูลนำเข้าในช่วงเวลา t และ z_t แทนสถานะซ่อนเร้นของโครงข่าย สมการสำหรับการคำนวณสถานะซ่อนเร้นสามารถเขียนได้ดังนี้

$$z_t = g_z\left(W_{z,z}z_{t-1} + W_{x,z}x_t + w_{0,z}\right) \quad (2\text{-}13)$$

โดยที่ $W_{z,z}$ คือเมทริกซ์น้ำหนักของการเชื่อมต่อแบบวนซ้ำจากสถานะก่อนหน้า z_{t-1} ไปยังสถานะปัจจุบัน z_t ขณะที่ $W_{x,z}$เป็นเมทริกซ์น้ำหนักระหว่างอินพุตและชั้นซ่อนเร้น และ $w_{0,z}$ คือค่าความเอนเอียง (Bias) ของชั้นซ่อนเร้น ฟังก์ชัน $g_z(\cdot)$ เป็นฟังก์ชันกระตุ้นของชั้นซ่อนเร้น ซึ่งทำหน้าที่เพิ่มความไม่เชิงเส้นให้กับแบบจำลอง เมื่อได้สถานะซ่อนเร้นแล้ว ผลลัพธ์ของโครงข่ายในช่วงเวลา t สามารถคำนวณได้จากสมการ

$$\hat{y}_t = g_y\left(W_{z,y}z_t + w_{0,y}\right) \quad (2\text{-}14)$$

โดยที่ $W_{z,y}$ เป็นเมทริกซ์น้ำหนักที่เชื่อมต่อระหว่างชั้นซ่อนเร้นกับชั้นผลลัพธ์ และ $w_{0,y}$ เป็นความเอนเอียงของชั้นผลลัพธ์ ส่วน $g_y(\cdot)$ เป็นฟังก์ชันกระตุ้นของชั้นผลลัพธ์ ซึ่งอาจเลือกใช้ให้เหมาะสมกับลักษณะของปัญหา เช่น SoftMax สำหรับปัญหาการจำแนกประเภท

2.2.2 หน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory)

โครงข่ายประสาทเทียมแบบหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory หรือ LSTM) เป็นสถาปัตยกรรมหนึ่งของโครงข่ายประสาทเทียมแบบวนซ้ำ ที่ได้รับการพัฒนาเพื่อเพิ่มประสิทธิภาพในการเรียนรู้ข้อมูลลำดับที่มีความสัมพันธ์ระยะยาว (long-term dependencies) ภายในข้อมูลอนุกรมเวลา (sequential data) โดยสถาปัตยกรรมดังกล่าวเป็นส่วนหนึ่งของแนวทางการเรียนรู้เชิงลึก (Deep Learning) ที่ถูกออกแบบมาเพื่อแก้ไขข้อจำกัดสำคัญของ RNN แบบดั้งเดิม ได้แก่ ปัญหาการเลือนหายของเกรเดียนต์ (vanishing gradient) และปัญหาเกรเดียนต์ระเบิด (exploding gradient) ซึ่งมักเกิดขึ้นเมื่อมีการแพร่ย้อนกลับของความคลาดเคลื่อนผ่านช่วงเวลาที่มีความยาวมาก โครงสร้างของ LSTM ถูกออกแบบให้มีหน่วยความจำพิเศษที่เรียกว่าเซลล์หน่วยความจำ (memory cell) ซึ่งสามารถเก็บรักษาข้อมูลบริบทสำคัญไว้ได้เป็นระยะเวลานาน พร้อมทั้งมีกลไกควบคุมการไหลของข้อมูลผ่านโครงข่ายด้วยชุดของประตูควบคุม (gating mechanisms) ส่งผลให้แบบจำลองสามารถเลือกเก็บหรือทิ้งข้อมูลที่ไม่จำเป็นได้อย่างมีประสิทธิภาพ โดยทั่วไป โครงสร้างของ LSTM ประกอบด้วยกลไกประตูหลักสามประเภท ได้แก่ Forget Gate, Input Gate และ Output Gate ซึ่งทำหน้าที่ควบคุมการไหลของข้อมูลเข้าและออกจากเซลล์หน่วยความจำ กลไกของ Forget Gate มีหน้าที่ตัดสินใจว่าข้อมูลใดจากสถานะเซลล์ในช่วงเวลาก่อนหน้าควรถูกลบหรือคงไว้ ขณะที่ Input Gate ทำหน้าที่กำหนดว่าข้อมูลใหม่จากอินพุตควรถูก

เพิ่มเข้าสู่เซลล์หน่วยความจำในระดับใด ส่วน Output Gate จะควบคุมว่าข้อมูลส่วนใดของสถานะเซลล์ควรถูกนำไปใช้เป็นสถานะซ่อนเร้น (hidden state) และถูกส่งต่อไปยังชั้นถัดไปหรือช่วงเวลาถัดไปของโครงข่าย กลไกดังกล่าวช่วยให้ LSTM สามารถควบคุมการไหลของข้อมูลได้อย่างยืดหยุ่นและสามารถเรียนรู้ความสัมพันธ์ระยะยาวในข้อมูลลำดับได้ดีกว่า RNN แบบพื้นฐาน

ในเชิงคณิตศาสตร์ เครือข่าย LSTM ทำการแปลงลำดับอินพุต $x = (x_1, \dots, x_T)$ ไปเป็นลำดับผลลัพธ์ $y = (y_1, \dots, y_T)$ ผ่านการคำนวณซ้ำของหน่วยเครือข่ายตั้งแต่ช่วงเวลา $t = 1$ ถึง T โดยสมการการทำงานของหน่วย LSTM มาตรฐานสามารถอธิบายได้ดังนี้

สมการของ Input Gate

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \tag{2-15}$$

สมการของ Forget Gate

$$f_t = \sigma\left(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f\right) \tag{2-16}$$

สมการการอัปเดต สถานะเซลล์ (Cell State)

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \tag{2-17}$$

สมการของ Output Gate

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) \tag{2-18}$$

สมการของ Hidden State หรือ Output Activation

$$m_t = o_t \odot h(c_t) \tag{2-19}$$

และสมการของ เอาต์พุตของเครือข่าย

$$y_t = \varphi\left(W_{ym}m_t + b_y\right) \tag{2-20}$$

ในสมการข้างต้น W แทนเมทริกซ์น้ำหนักของโครงข่าย (Weight Matrices) ส่วน W_{ic}, W_{fc}, W_{oc} เป็นเมทริกซ์น้ำหนักแบบแนวทแยงที่ใช้สำหรับการเชื่อมต่อแบบ peephole connections ซึ่งช่วยให้เกตสามารถเข้าถึงข้อมูลจากสถานะเซลล์ได้โดยตรง ตัวแปร b แทนเวกเตอร์ไบแอส (Bias Vectors) ขณะที่ σ เป็นฟังก์ชันโลจิสติกซิกมอยด์ (Logistic Sigmoid Function) สำหรับควบคุมค่าการเปิด-ปิดของเกต ตัวแปร i, f, o, c แทนเวกเตอร์แอคติเวชันของเกตอินพุต เกตลืม เกตเอาต์พุต และสถานะเซลล์ตามลำดับ ส่วน m เป็นเวกเตอร์เอาต์พุตของเซลล์ สัญลักษณ์ $\odot$ หมายถึงการคูณแบบองค์ประกอบต่อองค์ประกอบ (Element-Wise Product) ขณะที่ฟังก์ชัน g และ h มักกำหนดเป็นฟังก์ชันไฮเพอร์โบลิกแทนเจนต์ (Tanh) และ φ เป็นฟังก์ชันแอคติเวชันของชั้นเอาต์พุต เช่น SoftMax สำหรับงานจำแนกประเภท

2.3 การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ (Natural Language Processing หรือ NLP) เป็นสาขาหนึ่งของวิทยาการคอมพิวเตอร์ที่ผสานองค์ความรู้จากด้านปัญญาประดิษฐ์ ภาษาศาสตร์คอมพิวเตอร์ และการเรียนรู้ของเครื่อง เพื่อพัฒนาเทคนิคที่ทำให้คอมพิวเตอร์สามารถเข้าใจ วิเคราะห์ และสื่อสารด้วยภาษามนุษย์ได้อย่างมีประสิทธิภาพ ภาษาธรรมชาติถือเป็นเครื่องมือสำคัญที่มนุษย์ใช้ในการสื่อสาร ถ่ายทอดความรู้ และสร้างความเข้าใจร่วมกันในสังคมมาอย่างยาวนาน นับตั้งแต่มนุษย์เริ่มใช้ภาษาพูดเมื่อหลายหมื่นปีก่อน และพัฒนาระบบการเขียนในเวลาต่อมา ภาษาจึงเป็นกลไกสำคัญในการถ่ายทอดความรู้ทางวัฒนธรรม วิทยาศาสตร์ และเทคโนโลยีจากรุ่นสู่รุ่น แนวคิดสำคัญของ NLP คือการทำให้คอมพิวเตอร์สามารถประมวลผลข้อมูลข้อความหรือภาษาพูดในลักษณะที่ใกล้เคียงกับการรับรู้ของมนุษย์ โดยครอบคลุมกระบวนการตั้งแต่การวิเคราะห์โครงสร้างของภาษา (Syntactic analysis) การตีความความหมายของข้อความ (Semantic Analysis) ไปจนถึงการทำความเข้าใจบริบทและเจตนาของผู้สื่อสาร (Pragmatic Understanding) การพัฒนาระบบดังกล่าวมีความสำคัญอย่างยิ่งในยุคที่ข้อมูลจำนวนมหาศาลถูกสร้างขึ้นในรูปแบบของข้อความบนอินเทอร์เน็ต ไม่ว่าจะเป็นบทความ ข่าวสาร เอกสารทางวิชาการ หรือข้อมูลบนสื่อสังคมออนไลน์ ซึ่งส่วนใหญ่ถูกเขียนในรูปแบบภาษาธรรมชาติที่มนุษย์เข้าใจได้ง่าย แต่คอมพิวเตอร์ต้องอาศัยเทคนิคเฉพาะในการแปลงข้อมูลเหล่านี้ให้เป็นรูปแบบที่สามารถประมวลผลได้ ในระยะแรกของการพัฒนา NLP แนวทางที่ใช้มักอาศัยกฎเกณฑ์ทางภาษาศาสตร์และแบบจำลองเชิงสัญลักษณ์ (Rule-Based Approach) ซึ่งต้องอาศัยความรู้จากผู้เชี่ยวชาญในการกำหนดกฎไวยากรณ์และโครงสร้างภาษา อย่างไรก็ตามวิธีการดังกล่าวมีข้อจำกัดในการจัดการกับความหลากหลายและความซับซ้อนของภาษามนุษย์ เนื่องจากภาษาในโลกจริงมักมีความคลุมเครือ มีบริบท และมีรูปแบบการใช้ที่แตกต่างกันตามสถานการณ์ ดังนั้น งานวิจัยในระยะต่อมาจึงเริ่มนำเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) มาใช้ในการพัฒนาแบบจำลองที่สามารถเรียนรู้รูปแบบของภาษาได้จากข้อมูลจำนวนมากแทนที่จะอาศัยกฎที่ถูกกำหนดไว้ล่วงหน้าเพียงอย่างเดียวการพัฒนาเทคนิคการเรียนรู้เชิงลึก (Deep Learning) ได้สร้างความก้าวหน้าอย่างมากให้กับงานวิจัยด้าน NLP โดยเฉพาะการใช้โครงข่ายประสาทเทียม (Neural Networks) ที่สามารถเรียนรู้ตัวแทนข้อมูล (Representation) ของคำหรือข้อความในรูปแบบเชิงตัวเลขที่มีมิติสูง แนวทางนี้ช่วยให้ระบบสามารถจับความสัมพันธ์ระหว่างคำและบริบทของภาษาได้ดีกว่าวิธีการดั้งเดิม

2.3.1 ความถี่คำ–ความถี่ผกผันเอกสาร (Term Frequency-Inverse Document Frequency)

อัลกอริทึม TF-IDF เป็นกลไกสำคัญที่นำมาใช้เพื่อแก้ไขปัญหาการวิเคราะห์น้ำหนักของคำในคลังข้อมูลตัวอักษร โดยมีหลักการพื้นฐานจากการคำนวณผลคูณระหว่างค่าความถี่ของคำ (Term

Frequency) และค่าส่วนกลับของความถี่ในเอกสาร (Inverse Document Frequency) เพื่อระบุความสำคัญของคำแต่ละคำในบริบทของเอกสารหนึ่ง ๆ เมื่อเปรียบเทียบกับเอกสารทั้งหมดที่มีอยู่ในชุดข้อมูล ในส่วนของการคำนวณค่าความถี่ของคำนั้น สามารถดำเนินการได้หลายรูปแบบตามวัตถุประสงค์ของการใช้งาน โดยวิธีการที่เรียบง่ายที่สุดคือการนับจำนวนครั้งที่คำนั้นปรากฏขึ้นในเอกสารหนึ่ง ๆ อย่างไรก็ตาม ยังมีวิธีการคำนวณในรูปแบบอื่นที่เป็นที่นิยม เช่น การคำนวณแบบ Binary ซึ่งจะกำหนดค่าเป็น 1 หากคำนั้นปรากฏในเอกสารและเป็น 0 หากไม่ปรากฏ หรือการคำนวณแบบ Logarithmic โดยใช้สูตร $1 + \log(\text{“}tf\text{”})$ เพื่อปรับลดทอนผลกระทบจากคำที่มีความถี่สูงเกินไปให้อยู่ในระดับที่เหมาะสม สำหรับค่าส่วนกลับของความถี่ในเอกสารหรือ IDF นั้นคำนวณได้จากการหาค่าลอการิทึมของจำนวนเอกสารทั้งหมดหารด้วยจำนวนเอกสารที่มีคำนั้นปรากฏอยู่ กระบวนการนี้มีวัตถุประสงค์เพื่อให้คำที่มีความถี่ในการปรากฏตัวต่ำหรือคำที่พบน้อยในชุดข้อมูลได้รับค่าน้ำหนักที่สูงขึ้น ซึ่งสะท้อนถึงความสามารถในการจำแนกเอกสารที่โดดเด่นกว่าคำทั่วไป โดยโครงสร้างสมการขั้นพื้นฐานของ TF-IDF สามารถแสดงผลได้ดังนี้

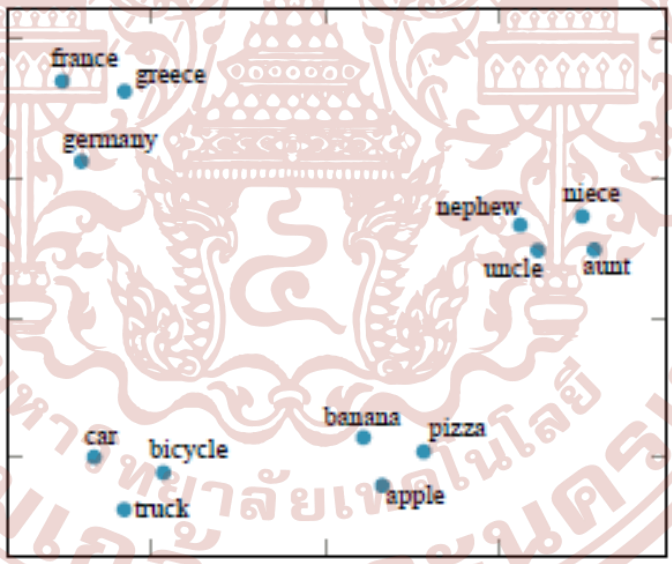
$$tf - idf(term, doc, docs) = count(term\ in\ doc) \cdot \frac{count(docs)}{count(docs\ with\ term)} \qquad (2\text{-}21)$$

การใช้ TF-IDF ถือเป็นเครื่องมือที่มีประสิทธิภาพสูงในการสร้างคุณลักษณะ (Feature Engineering) สำหรับการเรียนรู้ของเครื่อง (Machine Learning) จากข้อมูลข้อความ นอกจากนี้ยังมีประโยชน์อย่างยิ่งในระบบการสืบค้นข้อมูล (Information Retrieval) เนื่องจากการเปลี่ยนเอกสารให้อยู่ในรูปแบบของเวกเตอร์ตัวเลขช่วยให้สามารถคำนวณ "ระยะห่าง" (Distances) ระหว่างเวกเตอร์ของเอกสารแต่ละฉบับได้ หากพิจารณาว่าคำค้นหาของผู้ใช้คือเอกสารชุดหนึ่ง ระบบจะสามารถวัดระยะห่างระหว่างคำค้นหาดังกล่าวกับเอกสารอื่น ๆ ในฐานข้อมูล เพื่อจัดลำดับความเกี่ยวข้องและนำเสนอผลลัพธ์ที่มีความสอดคล้องกับความต้องการของผู้ใช้มากที่สุด

2.3.2 การฝังคำ (Word embeddings)

การฝังคำ (Word embeddings) เป็นเทคนิคที่ใช้แทนคำในรูปแบบเวกเตอร์เชิงตัวเลขเพื่อให้คอมพิวเตอร์สามารถเรียนรู้ความสัมพันธ์เชิงความหมายของคำในภาษาได้อย่างมีประสิทธิภาพ เทคนิคดังกล่าวถูกพัฒนาขึ้นเพื่อแก้ไขข้อจำกัดของวิธีการแทนคำแบบดั้งเดิม เช่น การเข้ารหัสแบบ One-hot Encoding ซึ่งแม้ว่าจะสามารถระบุคำแต่ละคำในพจนานุกรมได้อย่างชัดเจน แต่ไม่สามารถสะท้อนความสัมพันธ์หรือความคล้ายคลึงกันทางความหมายระหว่างคำได้ ดังนั้น Word Embeddings จึงถูกเสนอขึ้นเพื่อสร้างตัวแทนเชิงตัวเลขของคำในลักษณะที่สามารถสะท้อนความหมายและบริบทของคำได้ดียิ่งขึ้น โดยแนวคิดพื้นฐานของ Word Embeddings มีรากฐานมาจากสมมติฐานทางภาษาศาสตร์ของ John Rupert Firth ที่ระบุว่า “You shall know a word by the company it keeps” (Russell

and Norvig, 2021:907) ซึ่งหมายถึง ความหมายของคำสามารถเรียนรู้ได้จากบริบทของคำอื่น ๆ ที่ปรากฏร่วมกันในประโยคหรือเอกสาร แนวคิดนี้นำไปสู่การสร้างแบบจำลองที่เรียนรู้เวกเตอร์แทนคำจากการกระจายตัวของคำในคลังข้อความ (Corpus) ขนาดใหญ่ โดยแทนที่จะเก็บสถิติการปรากฏร่วมกันของคำในรูปแบบเมทริกซ์ขนาดใหญ่ที่มีมิติสูงและมีค่าศูนย์จำนวนมาก (Sparse Representation) วิธี Word Embeddings จะทำการลดมิติข้อมูลลงให้เป็นเวกเตอร์ขนาดเล็กที่มีค่าจริงต่อเนื่อง (Dense Vector Representation) ซึ่งมักมีจำนวนมิติอยู่ในช่วงหลักสิบถึงหลักร้อยมิติ ผลลัพธ์ของการเรียนรู้การฝังคำคือการกำหนดเวกเตอร์ให้กับคำแต่ละคำในพจนานุกรม โดยค่าของแต่ละมิติในเวกเตอร์ไม่ได้มีความหมายเชิงภาษาศาสตร์ที่สามารถตีความได้ แต่เมื่อพิจารณาเชิงโครงสร้างของพื้นที่เวกเตอร์ จะพบว่าคำที่มีความหมายหรือบริบทใกล้เคียงกันจะถูกจัดวางให้อยู่ใกล้กันในพื้นที่ดังกล่าว ตัวอย่างเช่น กลุ่มคำที่เกี่ยวข้องกับประเทศ เมืองหลวง หรือสมาชิกในครอบครัว มักปรากฏเป็นกลุ่ม (Cluster) ภายในพื้นที่เวกเตอร์เดียวกัน ดังแสดงในภาพที่ 2-12 ซึ่งเป็นการฉายภาพเวกเตอร์คำจากมิติสูงลงสู่พื้นที่สองมิติ เพื่อให้สามารถสังเกตความสัมพันธ์ระหว่างคำได้ง่ายขึ้น



ภาพที่ 2-12 การกระจายตัวของเวกเตอร์คำในพื้นที่เวกเตอร์ (Russell and Norvig, 2021:909)

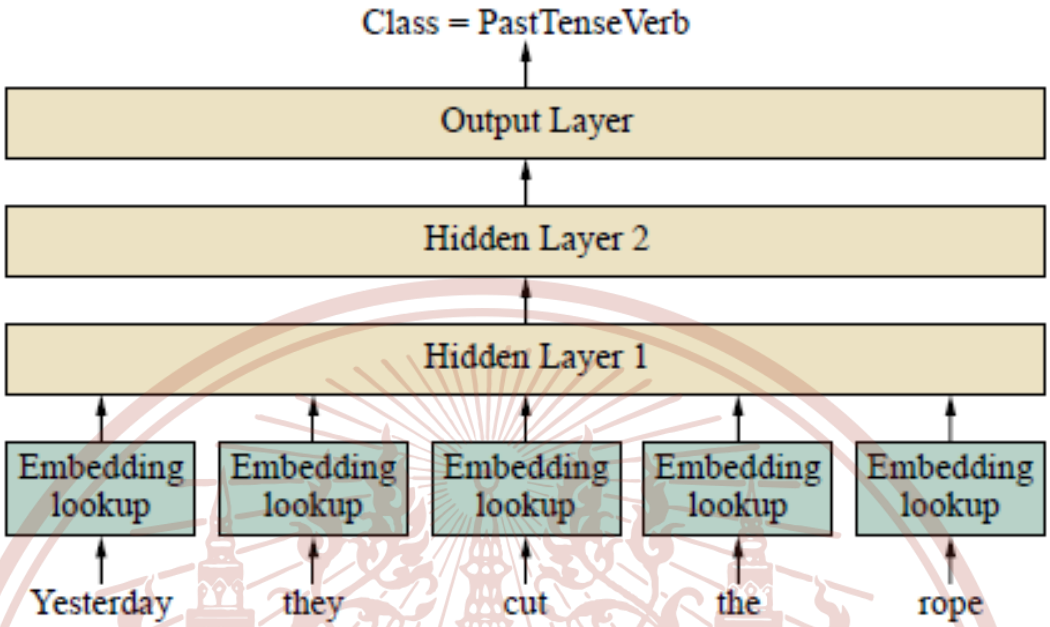
นอกจากนี้การฝังคำยังมีคุณสมบัติที่สำคัญอีกประการหนึ่ง คือความสามารถในการสะท้อนความสัมพันธ์เชิงความหมายผ่านการคำนวณเชิงเวกเตอร์ (Vector Arithmetic) กล่าวคือความแตกต่างระหว่างเวกเตอร์ของคำสองคำสามารถแสดงถึงความสัมพันธ์เชิงความหมายบางประเภทได้ เช่น ความสัมพันธ์ระหว่างประเทศและเมืองหลวง หรือความสัมพันธ์ระหว่างเพศชายและเพศหญิง ตัวอย่างเช่น เมื่อคำนวณเวกเตอร์ในลักษณะ $B - A + C$ โดยที่ A แทนคำว่า “Athens”, B แทนคำว่า “Greece” และ C แทนคำว่า “Oslo” เวกเตอร์ผลลัพธ์ D มักจะมีค่าที่ใกล้เคียงกับ

เวกเตอร์ของคำว่า “Norway” ซึ่งสะท้อนความสัมพันธ์ประเภท “เมืองหลวง–ประเทศ” ได้อย่างมีประสิทธิภาพ คุณสมบัตินี้ทำให้การฝังคำสามารถนำไปใช้แก้ปัญหาเชิงอนาลอจี (Analogy Problem) ในงานประมวลผลภาษาธรรมชาติได้ ดังแสดงตัวอย่างความสัมพันธ์ต่าง ๆ ในภาพที่ 2-13

A	B	C	D = C + (B − A)	Relationship
Athens	Greece	Oslo	Norway	*Capital*
Astana	Kazakhstan	Harare	Zimbabwe	*Capital*
Angola	kwanza	Iran	rial	*Currency*
copper	Cu	gold	Au	*Atomic Symbol*
Microsoft	Windows	Google	Android	*Operating System*
New York	New York Times	Baltimore	Baltimore Sun	*Newspaper*
Berlusconi	Silvio	Obama	Barack	*First name*
Switzerland	Swiss	Cambodia	Cambodian	*Nationality*
Einstein	scientist	Picasso	painter	*Occupation*
brother	sister	grandson	granddaughter	*Family Relation*
Chicago	Illinois	Stockton	California	*State*
possibly	impossibly	ethical	unethical	*Negative*
mouse	mice	dollar	dollars	*Plural*
easy	easiest	lucky	luckiest	*Superlative*
walking	walked	swimming	swam	*Past tense*

ภาพที่ 2-13 ตัวอย่างความสัมพันธ์ของคำจากการคำนวณเวกเตอร์ (Russell and Norvig, 2021:909)

ในเชิงการประยุกต์ใช้งานการฝังคำมักถูกนำมาใช้เป็นชั้นข้อมูลนำเข้า (Input Representation) สำหรับแบบจำลองการเรียนรู้เชิงลึกในงานประมวลผลภาษาธรรมชาติ เช่น การจำแนกข้อความ การแปลภาษา การวิเคราะห์ความรู้สึกของข้อความ และการตอบคำถาม โดยเวกเตอร์คำสามารถเรียนรู้ได้พร้อมกับการฝึกโมเดลหลัก หรือใช้เวกเตอร์ที่ผ่านการฝึกมาก่อนจากคลังข้อความขนาดใหญ่ (Pretrained Embeddings) ซึ่งช่วยลดเวลาในการฝึกโมเดลและเพิ่มประสิทธิภาพของระบบ ตัวอย่างอัลกอริทึมที่ใช้สร้าง Word Embeddings ที่ได้รับความนิยม ได้แก่ Word2Vec, GloVe และ FastText ซึ่งแต่ละวิธีมีแนวทางในการเรียนรู้เวกเตอร์คำจากข้อมูลข้อความแตกต่างกันไปเพื่อให้เข้าใจการนำ Word Embeddings ไปใช้ในโมเดลเชิงลึก ตัวอย่างหนึ่งคือการนำเวกเตอร์คำมาใช้ในงานระบุชนิดคำ (Part-of-Speech Tagging) โดยโมเดลจะรับคำหลายคำในบริบทเดียวกันเป็นหน้าต่างข้อความ (Context Window) จากนั้นทำการค้นหาเวกเตอร์คำจาก Embedding Matrix และนำไปประมวลผลผ่านโครงข่ายประสาทเทียมแบบ Feed Forward เพื่อทำนายชนิดคำของคำที่อยู่ตรงกลางของหน้าต่างดังกล่าว กระบวนการนี้ช่วยให้โมเดลสามารถเรียนรู้ทั้งความหมายของคำและบทบาททางไวยากรณ์ของคำภายในประโยคได้พร้อมกัน ดังแสดงในภาพที่ 2-13 ซึ่งอธิบายโครงสร้างของโมเดลที่ใช้ Word Embeddings เป็นตัวแทนข้อมูลนำเข้า



ภาพที่ 2-14 โครงสร้างโครงข่ายประสาทเทียมสำหรับระบุชนิดคำโดยใช้ Word Embeddings (Russell and Norvig, 2021:910)

2.3.3 การประมวลผลภาษาธรรมชาติภาษาไทย (Thai NLP)

ภาษาไทยเป็นภาษาประจำชาติของประเทศไทย และจัดอยู่ในกลุ่มภาษาไท (Tai linguistic group) ภายใต้ตระกูลภาษาขร้า–ไท (Kra–Dai language family) โดยมีจำนวนผู้ใช้ภาษาไทยทั่วโลกประมาณ 85.7 ล้านคน แม้ว่าภาษาไทยจะเป็นภาษาที่มีผู้ใช้งานจำนวนมากในระดับสากล แต่การพัฒนาเทคโนโลยีด้าน การประมวลผลภาษาธรรมชาติ (Natural Language Processing) สำหรับภาษาไทยยังคงมีข้อจำกัดเมื่อเปรียบเทียบกับภาษาที่มีทรัพยากรสูง เช่น ภาษาจีน ภาษาอังกฤษ และภาษาเยอรมัน ซึ่งมีชุดข้อมูล เครื่องมือ และงานวิจัยรองรับจำนวนมากกว่า ความแตกต่างดังกล่าวสะท้อนถึงสถานะของภาษาไทยในฐานะภาษาที่มีทรัพยากรด้านภาษาศาสตร์คอมพิวเตอร์ค่อนข้างจำกัด (low-resource language) ส่งผลให้การพัฒนาแบบจำลองภาษาธรรมชาติที่มีประสิทธิภาพสูงสำหรับภาษาไทยยังคงเป็นความท้าทายสำคัญของนักวิจัยในสาขานี้ ลักษณะเฉพาะของภาษาไทยยังเป็นปัจจัยสำคัญที่ทำให้การพัฒนาแบบจำลองภาษาธรรมชาติมีความซับซ้อนมากขึ้น หนึ่งในลักษณะที่สำคัญคือรูปแบบการเขียนแบบต่อเนื่อง (Scriptio Continua) ซึ่งเป็นรูปแบบการเขียนที่ไม่มีการเว้นวรรคเพื่อแบ่งขอบเขตของคำอย่างชัดเจน ดังนั้นข้อความภาษาไทยจำนวนมากจึงปรากฏในลักษณะของตัวอักษรที่เรียงต่อกันอย่างต่อเนื่องโดยไม่มีเครื่องหมายกำกับขอบเขตของคำหรือประโยค ลักษณะดังกล่าวส่งผลให้ขั้นตอนการประมวลผลข้อความจำเป็นต้องอาศัยกระบวนการวิเคราะห์เพิ่มเติมเพื่อระบุหน่วยคำที่ถูกต้อง นอกจากนี้ ภาษาไทยยังมีลักษณะของความกำกวมทางภาษา

(Ambiguity) ซึ่งเกิดจากการที่ขอบเขตของคำหรือวลีไม่สามารถระบุได้อย่างชัดเจนจากรูปแบบการเขียนเพียงอย่างเดียว ความกำกวมดังกล่าวไม่สามารถแก้ไขได้ด้วยกฎทางไวยากรณ์เพียงอย่างเดียว แต่จำเป็นต้องอาศัยบริบทของประโยคและความรู้ทางสถิติจากข้อมูลจำนวนมากในการช่วยตัดสินใจด้วยเหตุนี้ กระบวนการพื้นฐานในงานประมวลผลภาษาไทย เช่น การตัดคำ หรือ Word Tokenization จึงถือเป็นขั้นตอนสำคัญที่ต้องดำเนินการก่อนกระบวนการวิเคราะห์ขั้นสูงอื่น ๆ เช่น การแท็กชนิดคำ การวิเคราะห์โครงสร้างไวยากรณ์ หรือการทำความเข้าใจความหมายของข้อความ เนื่องจากหากการแบ่งคำเกิดข้อผิดพลาด อาจส่งผลกระทบต่อประสิทธิภาพของแบบจำลองในขั้นตอนถัดไปอย่างมีนัยสำคัญ ดังนั้น การพัฒนาเครื่องมือสำหรับการตัดคำและการประมวลผลภาษาไทยในระดับพื้นฐานจึงเป็นประเด็นสำคัญที่ได้รับความสนใจจากนักวิจัยในสาขา NLP

หนึ่งในเครื่องมือสำคัญที่ถูกพัฒนาขึ้นเพื่อสนับสนุนงานด้าน NLP สำหรับภาษาไทยคือ PyThaiNLP (Worawet, 2023) โดยเป็นไลบรารีโอเพนซอร์สที่พัฒนาด้วยภาษา Python และออกแบบมาเพื่ออำนวยความสะดวกในการพัฒนาแอปพลิเคชันด้านการประมวลผลภาษาไทย ไลบรารีดังกล่าวมีโครงสร้างอินเทอร์เฟซที่เป็นมาตรฐานและใช้งานได้ง่าย พร้อมทั้งเปิดให้ใช้งาน ปรับแก้ และนำไปใช้ในเชิงพาณิชย์ได้อย่างเสรี ภายในไลบรารีได้รวมเครื่องมือและแบบจำลองที่หลากหลายตั้งแต่กระบวนการพื้นฐาน เช่น ตัวตัดคำ (Tokenizers) และการแท็กลำดับคำ (Sequence Tagging) ไปจนถึงแบบจำลองภาษาที่ได้รับการฝึกมาก่อน (Pre-trained Language Models) รวมถึงการรองรับการทำงานร่วมกับแบบจำลองภาษาขนาดใหญ่ในงานวิจัยสมัยใหม่ การมีชุดเครื่องมือแบบบูรณาการดังกล่าวช่วยให้นักวิจัยสามารถดำเนินการทดลองและพัฒนาแบบจำลองขั้นสูงสำหรับภาษาไทยได้อย่างมีประสิทธิภาพมากยิ่งขึ้น ตัวอย่างเช่น งานวิเคราะห์โครงสร้างความสัมพันธ์ทางไวยากรณ์ด้วยวิธี Universal Dependencies หรือการพัฒนาระบบที่สามารถเรียนรู้การแทนค่าภาษาในลักษณะข้ามภาษา (Cross-lingual Representation Learning) ซึ่งมีบทบาทสำคัญต่อการพัฒนาแบบจำลองที่สามารถทำงานร่วมกับหลายภาษาได้ในระบบเดียวกัน ดังนั้น การพัฒนาและสนับสนุนเครื่องมือโอเพนซอร์สสำหรับภาษาไทยจึงมีความสำคัญอย่างยิ่งต่อการยกระดับศักยภาพของงานวิจัยด้านการประมวลผลภาษาธรรมชาติ และช่วยส่งเสริมให้ภาษาไทยสามารถถูกนำไปประยุกต์ใช้ในเทคโนโลยีปัญญาประดิษฐ์ได้อย่างกว้างขวางในอนาคต

2.4 งานวิจัยที่เกี่ยวข้อง

การศึกษาเกี่ยวกับการจำแนกข้อความอีเมลด้วยเทคนิคการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกได้รับความสนใจอย่างต่อเนื่องในช่วงหลายปีที่ผ่านมา เนื่องจากอีเมลเป็นหนึ่งในช่องทางการสื่อสารที่สำคัญในองค์กรและสังคมดิจิทัล การจำแนกประเภทอีเมลอย่างถูกต้องและรวดเร็วสามารถช่วยเพิ่มประสิทธิภาพในการจัดการข้อมูล ลดภาระงานของมนุษย์ และสนับสนุนกระบวนการตัดสินใจในองค์กร งานวิจัยที่เกี่ยวข้องจำนวนมากจึงมุ่งเน้นการประยุกต์ใช้เทคนิคด้านการ

ประมวลผลภาษาธรรมชาติ (Natural Language Processing) ร่วมกับอัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning) และการเรียนรู้เชิงลึก (Deep Learning) เพื่อพัฒนาแบบจำลองสำหรับการจำแนกประเภทข้อความอีเมลให้มีความแม่นยำสูงขึ้น

หนึ่งในเทคนิคพื้นฐานที่ถูกนำมาใช้ในการจำแนกข้อความคือการแปลงข้อความให้เป็นคุณลักษณะเชิงตัวเลขผ่านกระบวนการสกัดคุณลักษณะ (Feature Extraction) โดยเทคนิคที่ได้รับความนิยมอย่างแพร่หลายคือวิธี TF-IDF ซึ่งช่วยสะท้อนความสำคัญของคำในเอกสารเมื่อเปรียบเทียบกับเอกสารอื่นในชุดข้อมูล เทคนิคดังกล่าวถูกนำมาใช้ร่วมกับอัลกอริทึมการเรียนรู้ของเครื่องหลายประเภทเพื่อสร้างแบบจำลองสำหรับการจำแนกข้อความ ตัวอย่างเช่น งานวิจัยของ สิรภพ และ ฐิติรัตน์ (2564) ได้ศึกษาการเพิ่มประสิทธิภาพการจำแนกตัวงานสำหรับระบบจัดการบริการด้านไอที (IT Service Management) โดยใช้เทคนิคการเรียนรู้ของเครื่องร่วมกับการประมวลผลภาษาธรรมชาติ งานวิจัยดังกล่าวนำเทคนิค TF-IDF มาใช้ในการสกัดคุณลักษณะจากข้อความของตัวงานที่แจ้งผ่านอีเมล ก่อนนำข้อมูลเข้าสู่กระบวนการฝึกสอนแบบจำลองด้วยอัลกอริทึมการเรียนรู้ของเครื่องหลายประเภท ผลการทดลองพบว่าแบบจำลอง XGBoost สามารถให้ผลลัพธ์ที่ดีที่สุด โดยมีค่า F1-score เท่ากับ 81.21% อีกทั้งยังช่วยลดระยะเวลาในการจำแนกประเภทตัวงานให้เหลือน้อยกว่า 1 วัน ซึ่งสะท้อนให้เห็นถึงศักยภาพของการประยุกต์ใช้ Machine Learning ในการปรับปรุงประสิทธิภาพของกระบวนการจัดการอีเมลและตัวงานในองค์กร

นอกจากการใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) แบบดั้งเดิมแล้ว งานวิจัยในระยะหลังยังให้ความสำคัญกับการประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึก ซึ่งมีความสามารถในการเรียนรู้รูปแบบเชิงลำดับของข้อมูลข้อความได้ดีกว่า โดยเฉพาะโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network) และสถาปัตยกรรมที่ได้รับการพัฒนาเพิ่มเติมอย่างหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory) ซึ่งสามารถจัดการกับปัญหาการพึ่งพาระยะยาวของข้อมูลลำดับ (long-term dependencies) ได้อย่างมีประสิทธิภาพ งานวิจัยของ Hina และคณะ (2021) ได้นำเสนอแนวทางที่เรียกว่า SeFACED (Semantic-Based Forensic Analysis and Classification of E-Mail Data Using Deep Learning) ซึ่งใช้การผสมผสานระหว่าง LSTM และ GRU สำหรับการจำแนกอีเมลแบบหลายคลาส โดยแบ่งประเภทอีเมลออกเป็นสี่ประเภท ได้แก่ อีเมลปกติ อีเมลฉ้อโกง อีเมลข่มขู่ และอีเมลน่าสงสัย วิธีการดังกล่าวมุ่งเน้นการวิเคราะห์ความหมายของข้อความอีเมลในเชิงลึกเพื่อสนับสนุนการวิเคราะห์ทางนิติวิทยาศาสตร์ดิจิทัล (Digital Forensics) ผลการทดลองแสดงให้เห็นว่าโมเดลสามารถให้ความแม่นยำสูงถึง 95% ซึ่งสะท้อนถึงศักยภาพของเทคนิคการเรียนรู้เชิงลึก (Deep Learning) ในการวิเคราะห์ข้อความอีเมลที่มีโครงสร้างซับซ้อนและมีความยาวของข้อความที่แตกต่างกัน

อีกแนวทางหนึ่งที่ได้รับความสนใจคือการผสมผสานเทคนิคการสกัดคุณลักษณะกับสถาปัตยกรรมของโครงข่ายประสาทเทียมหลายรูปแบบ เพื่อเปรียบเทียบประสิทธิภาพในการจำแนกประเภทข้อความ งานวิจัยของ Adrales และ Sison (2024) ได้ศึกษาแนวทางการจำแนกอีเมลในบริบทของสถาบันการศึกษา โดยแบ่งอีเมลออกเป็นสองประเภทหลัก ได้แก่ อีเมลประเภทการสอบถามทั่วไป และอีเมลประเภทการตรวจสอบ งานวิจัยดังกล่าวใช้เทคนิคการเรียนรู้เชิงลึก (Deep Learning) หลายสถาปัตยกรรม ได้แก่ CNN, RNN และ LSTM ร่วมกับวิธีการแปลงข้อความเป็นเวกเตอร์ด้วย TF-IDF และการฝังคำ (Word Embedding) แบบ GloVe ผลการทดลองพบว่าโมเดล CNN ที่ใช้คุณลักษณะจาก TF-IDF สามารถให้ความแม่นยำสูงสุดถึง 98% ซึ่งแสดงให้เห็นว่าแม้เทคนิคการฝังคำจะสามารถจับความสัมพันธ์เชิงความหมายของคำได้ดี แต่การใช้ TF-IDF ร่วมกับโครงข่ายประสาทเทียมบางประเภทก็ยังคงให้ผลลัพธ์ที่มีประสิทธิภาพสูงในงานจำแนกข้อความอีเมล

ในบริบทของการใช้งานจริงในภาคอุตสาหกรรม งานวิจัยของ Borg และคณะ (2021) ได้ศึกษาการปรับปรุงระบบจำแนกอีเมลสำหรับการสนับสนุนลูกค้าในบริษัทโทรคมนาคมขนาดใหญ่ โดยมีวัตถุประสงค์เพื่อแทนที่ระบบเดิมที่ใช้กฎที่กำหนดด้วยมือ (Rule-Based Algorithm) ซึ่งมักมีข้อจำกัดด้านความยืดหยุ่นและความสามารถในการปรับตัวต่อข้อมูลใหม่ งานวิจัยดังกล่าวได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องทั้งแบบลำดับ (Sequential Models) และแบบไม่ลำดับ (Non-sequential Models) สำหรับการจำแนกอีเมลที่มีป้ายกำกับจำนวนมากถึง 33 ประเภท ผลการทดลองพบว่าแบบจำลอง LSTM สามารถให้ผลลัพธ์ที่ดีที่สุด โดยมีค่า F1-score เท่ากับ 0.91 ซึ่งสูงกว่าแบบจำลองแบบดั้งเดิม เช่น SVM และ ANN ผลการศึกษานี้แสดงให้เห็นว่าโมเดลที่สามารถเรียนรู้ลำดับของคำในข้อความได้มีแนวโน้มที่จะให้ประสิทธิภาพที่ดีกว่าในการจำแนกอีเมลที่มีโครงสร้างข้อความซับซ้อน

นอกจากนี้ ปัญหาที่สำคัญอีกประการหนึ่งในการจำแนกข้อความอีเมลคือปัญหาความไม่สมดุลของชุดข้อมูล (Imbalanced Data) ซึ่งเกิดขึ้นเมื่อจำนวนตัวอย่างในแต่ละคลาสมีความแตกต่างกันอย่างมาก ปัญหาดังกล่าวอาจส่งผลให้แบบจำลองมีแนวโน้มเรียนรู้คลาสที่มีจำนวนตัวอย่างมากได้ดีกว่า งานวิจัยของ Bhatti และคณะ (2021) ได้ศึกษาการจำแนกอีเมลที่เกี่ยวข้องกับอาชญากรรมไซเบอร์โดยใช้เทคนิคหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory) และได้วิเคราะห์ผลกระทบของปัญหาข้อมูลไม่สมดุลต่อประสิทธิภาพของแบบจำลอง งานวิจัยดังกล่าวเสนอการใช้เทคนิค Over-sampling เพื่อเพิ่มจำนวนตัวอย่างของคลาสที่มีจำนวนน้อยในชุดข้อมูล ผลการทดลองแสดงให้เห็นว่าการปรับสมดุลชุดข้อมูลด้วยวิธีดังกล่าวสามารถเพิ่มความแม่นยำของการจำแนกประเภทจากประมาณ 80% เป็นมากกว่า 95% ซึ่งชี้ให้เห็นถึงความสำคัญของการเตรียมข้อมูลก่อนกระบวนการฝึกสอนแบบจำลอง

จากการศึกษางานวิจัยที่เกี่ยวข้องข้างต้นสามารถสรุปได้ว่า การจำแนกข้อความอีเมลเป็นปัญหาที่ได้รับการศึกษาอย่างกว้างขวาง โดยมีการประยุกต์ใช้เทคนิคหลากหลายตั้งแต่การเรียนรู้ของเครื่องแบบดั้งเดิมไปจนถึงการเรียนรู้เชิงลึก เทคนิคการสกัดคุณลักษณะ เช่น TF-IDF และการฝังคำ (Word Embedding) ถูกนำมาใช้เพื่อแปลงข้อมูลข้อความให้สามารถนำเข้าสู่แบบจำลองได้ ขณะที่สถาปัตยกรรมของแบบจำลอง เช่น CNN, RNN และ LSTM ได้รับการพัฒนาเพื่อเพิ่มความสามารถในการเรียนรู้รูปแบบของข้อมูลข้อความที่มีความซับซ้อน นอกจากนี้ งานวิจัยหลายชิ้นยังชี้ให้เห็นถึงความสำคัญของปัจจัยอื่น ๆ เช่น การจัดการข้อมูลไม่สมดุล และการเลือกสถาปัตยกรรมของแบบจำลองที่เหมาะสมกับลักษณะของข้อมูล อย่างไรก็ตาม แม้ว่างานวิจัยก่อนหน้าจะสามารถแสดงให้เห็นถึงประสิทธิภาพของเทคนิคต่าง ๆ ในการจำแนกข้อความอีเมลได้เป็นอย่างดี แต่ยังคงมีความจำเป็นในการศึกษาวิธีการที่เหมาะสมกับบริบทของข้อมูลและปัญหาเฉพาะในแต่ละงานวิจัย เพื่อพัฒนาแบบจำลองที่สามารถนำไปประยุกต์ใช้ได้อย่างมีประสิทธิภาพในสถานการณ์จริงต่อไป

พระจอมเกล้าพระนครเหนือ มหาวิทยาลัยเทคโนโลยี

บทที่ 3
วิธีดำเนินการ

งานค้นคว้าอิสระเรื่องการจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง ทำเพื่อพัฒนาแบบจำลองการเรียนรู้ของเครื่อง เพื่อเพิ่มประสิทธิภาพการจัดการข้อมูลในองค์กรและลดภาระงาน โดยมีวิธีดำเนินการวิจัยดังต่อไปนี้

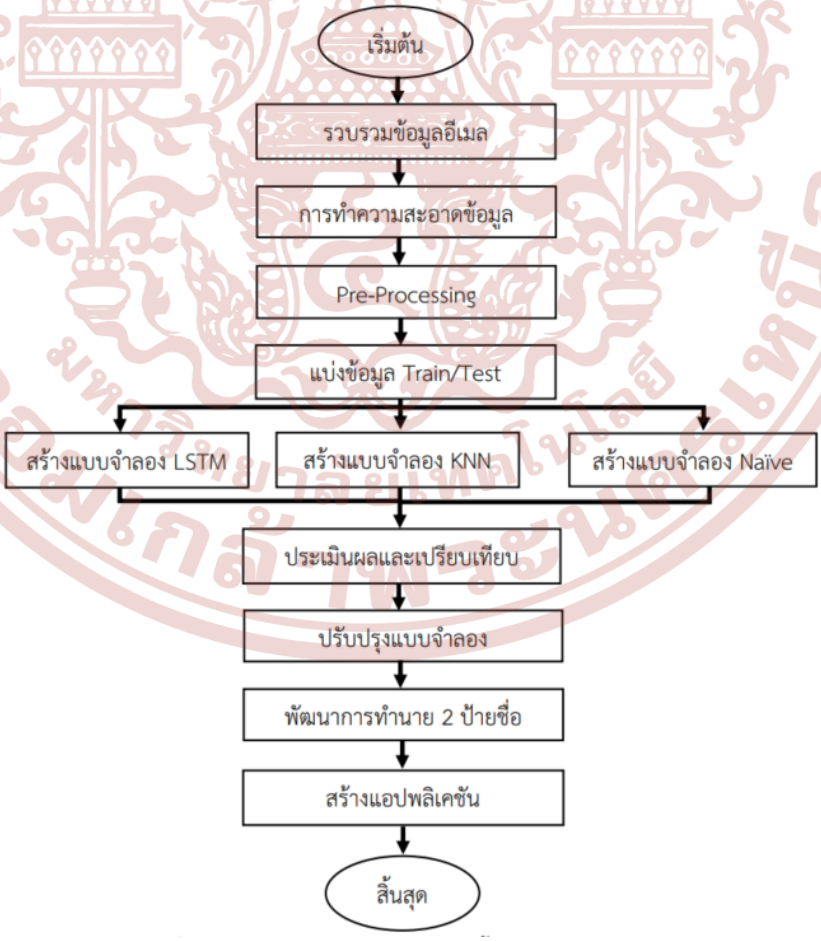
3.1 ขั้นตอนการดำเนินการ

3.2 ชุดข้อมูลและการเตรียมข้อมูล

3.3 การสร้างและเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่อง

3.4 การออกแบบแอปพลิเคชั่นจำแนกวัตถุประสงค์อีเมล

3.1 ขั้นตอนการดำเนินการ



ภาพที่ 3-1 แผนภาพ Flow Chart แสดงขั้นตอนการดำเนินการ

3.2 ชุดข้อมูลและการเตรียมข้อมูล

การวิจัยจะเริ่มต้นจากกระบวนการเก็บรวบรวมข้อมูลแฟลตฟอร์มอีเมลคลาวด์ เนื่องจากข้อมูลที่ได้มานั้นยังไม่สามารถนำมาใช้ในการสร้างแบบจำลองการเรียนรู้ของเครื่องได้ จึงจำเป็นต้องมีการเตรียมข้อมูลและทำป้ายชื่อข้อมูล (Label) ให้สมบูรณ์ก่อนนำไปใช้ ดังนั้นขั้นตอนในปฏิบัติจะเป็นไปตามหัวข้อดังต่อไปนี้

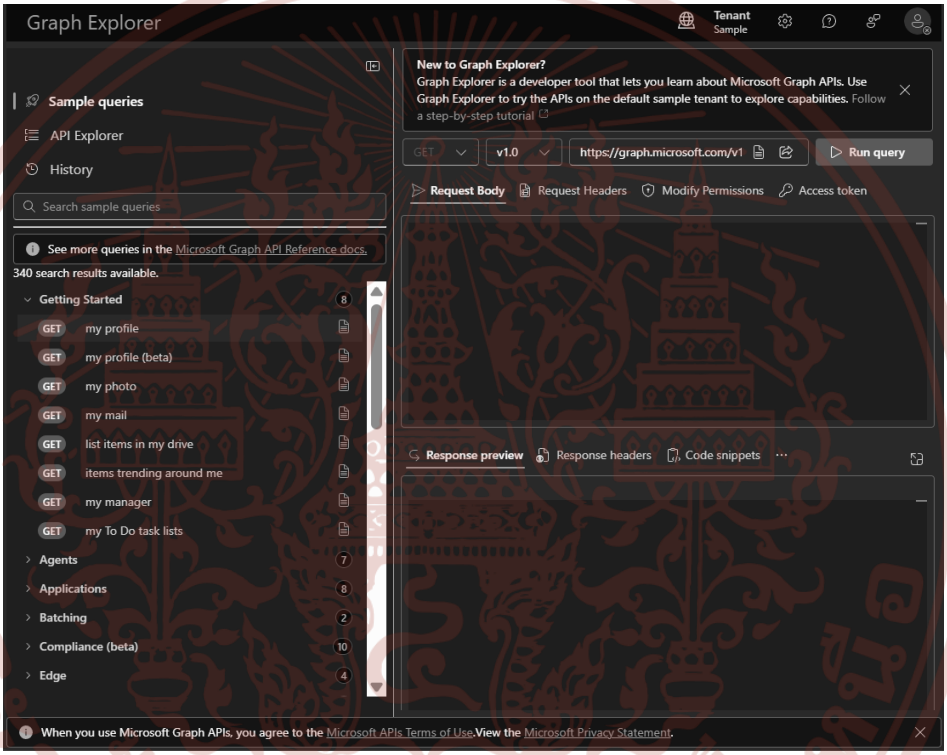
3.2.1 การรวบรวมข้อมูลอีเมล

ในการวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการเก็บรวบรวมข้อมูลอีเมลจากแฟลตฟอร์มอีเมลคลาวด์ภายใต้ ไมโครซอฟต์ออฟฟิต 365 (Microsoft Office 365) เพื่อใช้เป็นข้อมูลต้นทาง (Raw Data) สำหรับการวิเคราะห์และพัฒนาแบบจำลองการเรียนรู้ของเครื่อง ขอบเขตข้อมูลชุดข้อมูลคืออีเมลทุกฉบับของกลุ่มผู้ใช้ตัวอย่างตั้งแต่วันที่ 1 มกราคม-31 ธันวาคม พ.ศ. 2568 โดยมีจำนวนข้อมูลรวมทั้งสิ้น 9,569 ฉบับ โดยข้อมูลที่ดึงมาได้จะสะท้อนลักษณะการสื่อสารภายในองค์กร เช่น เนื้อหาอีเมล ผู้ส่ง ผู้รับ วันที่และเวลาที่ส่ง และหัวข้ออีเมล ซึ่งจะถูกจัดเก็บในรูปแบบสัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) ภายในคอมพิวเตอร์ส่วนตัว เพื่อความสะดวกในการเตรียมข้อมูลและนำไปใช้ในขั้นตอนถัดไปของการเรียนรู้ของเครื่อง

ในการดำเนินงานวิจัยครั้งนี้ ผู้วิจัยได้ประยุกต์ใช้เครื่องมือในการจัดการข้อมูลจดหมายอิเล็กทรอนิกส์จากระบบอีเมลคลาวด์ชื่อ “Outlook” ภายใต้แพลตฟอร์มไมโครซอฟต์ออฟฟิต 365 โดยมีอินเตอร์เฟสโปรแกรมประยุกต์ (Application Programming Interface) ชื่อ “Microsoft Graph” ทำหน้าที่เป็นช่องทางหลักในการเข้าถึงและแลกเปลี่ยนข้อมูลระหว่างระบบ โดยออกแบบมาเพื่อให้สามารถเรียกใช้ข้อมูลบริการต่าง ๆ ภายใต้สภาพแวดล้อมของไมโครซอฟต์ออฟฟิต 365 ได้อย่างเป็นมาตรฐานและปลอดภัย สำหรับกระบวนการพัฒนา ผู้วิจัยเลือกใช้ภาษาไพทอน (Python) เป็นภาษาหลักในการพัฒนาสคริปต์ (Script) สำหรับการดึงข้อมูล การจัดการการรับรองสิทธิ์ (Authentication) และการประมวลผลข้อมูลภายหลังการร้องขอข้อมูล โดยอาศัยไลบรารี (Library) ที่สำคัญได้แก่ไลบรารี “Requests” สำหรับส่งคำร้องขอผ่านโปรโตคอลการถ่ายโอนไฮเปอร์เท็กซ์ (Hypertext Transfer Protocol) ไปยังอินเตอร์เฟสโปรแกรมประยุกต์และไลบรารี “JSON” สำหรับการจัดการข้อมูลในรูปแบบสัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) และ ไลบรารี “Beautiful Soup” สำหรับการแปลงและสกัดข้อความจากโครงสร้างภาษาเอชทีเอ็มแอล (HyperText Markup Language) นอกจากนี้ ผู้วิจัยได้กำหนดรูปแบบการจัดเก็บข้อมูลในหน่วยจัดเก็บภายในเครื่อง (Local Storage) โดยบันทึกข้อมูลที่ผ่านการประมวลผลแล้วในรูปแบบไฟล์

สัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) เพื่อให้สามารถนำไปวิเคราะห์ต่อในขั้นตอนถัดไปได้อย่างเป็นระบบและตรวจสอบย้อนกลับได้

กระบวนการดึงข้อมูลอีเมลจากระบบอีเมลคลาวด์ประกอบด้วยหลายขั้นตอน โดยในขั้นตอนแรกเป็นการเชื่อมต่อกับระบบไมโครซอฟต์ออฟฟิต 365 ผู้วิจัยเข้าสู่ระบบผ่านหน้าใช้งานของ Microsoft Graph แสดงตามภาพที่ 3-2



ภาพที่ 3-2 หน้าใช้งานของ Microsoft Graph

ซึ่งเป็นเครื่องมือสำหรับทดสอบและตรวจสอบคำร้องขออินเตอร์เฟสโปรแกรมประยุกต์ จากนั้นดำเนินการลงชื่อเข้าใช้ (Sign in) เพื่ออนุญาตให้เข้าถึงข้อมูลอีเมลภายใต้สิทธิ์ของผู้ใช้ได้ โดยการเข้าถึงจำเป็นต้องใช้โทเค็นการเข้าถึง (Access Token) ซึ่งเปรียบเสมือนเป็นรหัสสำหรับพิสูจน์ตัวตนและการอนุญาตการเข้าถึงทรัพยากรตามกระบวนการรับรองสิทธิ์และการกำหนดสิทธิ์ โดยผู้วิจัยคัดลอกโทเค็นการเข้าถึงจากแท็บการตรวจสอบสิทธิ์ (Authentication Tab) และนำไปกำหนดเป็นตัวแปรชื่อ “ACCESS_TOKEN” ภายในสคริปต์ภาษาไพทอน เพื่อใช้เป็นส่วนหัวคำร้องขอ (Header) แบบโทเค็นผู้ถือ (Bearer Token) ในการเรียกใช้งานอินเตอร์เฟสโปรแกรมประยุกต์

ในขั้นตอนที่สอง ผู้วิจัยกำหนดขอบเขตของข้อมูลที่ต้องการดึง โดยระบุช่วงเวลาเริ่มต้นและสิ้นสุดของอีเมลที่สนใจ ผ่านการกำหนดปี เดือน วัน และเวลาอย่างชัดเจน เพื่อใช้เป็นเงื่อนไขกรอง

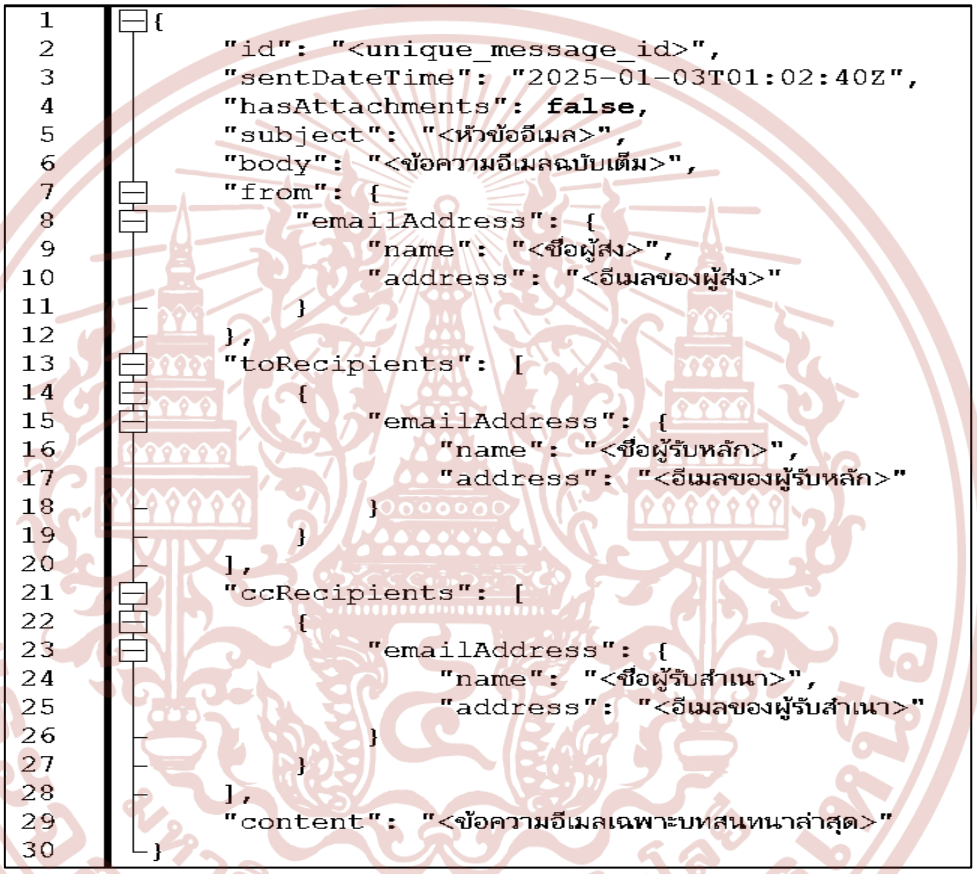
ข้อมูลผ่านพารามิเตอร์ “$filter” ในคำร้องขออินเตอร์เฟสโปรแกรมประยุกต์นอกจากนี้ ยังมีการกำหนดโฟลเดอร์ (Folder) เป้าหมาย เช่น “Sent Items” สำหรับอีเมลที่ส่งแล้ว หรือ “Inbox” สำหรับอีเมลขาเข้า เพื่อจำกัดขอบเขตข้อมูลให้สอดคล้องกับวัตถุประสงค์ของการวิจัย อันเป็นการลดปริมาณข้อมูลส่วนเกินและเพิ่มประสิทธิภาพในการประมวลผล

ขั้นตอนที่สามเป็นการร้องขออีเมลจากอินเตอร์เฟสโปรแกรมประยุกต์ Microsoft Graph โดยสคริปต์จะสร้างโครงสร้าง URL ตามรูปแบบมาตรฐานและกำหนดพารามิเตอร์ “$select” เพื่อระบุฟิลด์ข้อมูลที่ต้องการ เช่น รหัสระบุชื่อ (ID) หัวข้ออีเมล (Subject) เวลาส่ง (Sent Date Time) ข้อมูลผู้ส่ง (From) ข้อมูลผู้รับ (to Recipients) และเนื้อหาอีเมล (Body) จากนั้นส่งคำร้องขอแบบเรียกรับ (Get) พร้อมแนบส่วนหัวที่ประกอบด้วยโทเค็นผู้ถือ (Bearer Token) เพื่อยืนยันสิทธิ์การเข้าถึง เมื่อระบบตอบกลับ จะได้รับข้อมูลในรูปแบบสัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) ซึ่งมีโครงสร้างข้อมูลเชิงลำดับชั้นที่สามารถนำไปประมวลผลต่อได้โดยตรงในภาษาไพทอน

ขั้นตอนที่สี่เป็นการประมวลผลข้อความในเนื้อหาอีเมล (Body) เนื่องจากเนื้อหาดั้งเดิมจะมีส่วนเกินที่เป็นประวัติการสนธนาก่อนหน้า ที่อยู่ในหัวข้ออีเมลเดียวกัน ซึ่งนอกจากจะไม่เป็นประโยชน์ต่อการตัดสินใจแล้วยังมีปริมาณมากอีกด้วยขึ้นอยู่กับจำนวนครั้งในการสนธนา จึงจำเป็นต้องตัดส่วนเกินนี้ออก โดยเริ่มจากการตรวจสอบประเภทของเนื้อหาในฟิลด์เนื้อหาอีเมล (Body) ว่าเป็นข้อความล้วน (text) หรือภาษาเอชทีเอ็มแอล (HyperText Markup Language) หากเป็นภาษาเอชทีเอ็มแอล ผู้วิจัยใช้ไลบรารี “Beautiful Soup” แปลงโครงสร้างดังกล่าวให้เป็นข้อความล้วน (Plain text) เพื่อให้เหมาะสมต่อการวิเคราะห์ข้อความ (Text Processing) ในลำดับถัดไป นอกจากนี้ได้กำหนดเงื่อนไขในการสกัดเฉพาะเนื้อหาหลักของอีเมล โดยตัดข้อความตั้งแต่คำว่า “Best Regards” เป็นต้นไป เนื่องจากเป็นคำปิดท้ายบทสนธนาพื้นฐานที่คนในองค์กรผู้วิจัยใช้ เพื่อแยกส่วนลายเซ็นออกจากเนื้อหาสาระสำคัญ อันเป็นการลดสัญญาณรบกวน (Noise) ในข้อมูล แล้วเก็บอยู่ในคีย์ (Key) ชื่อ “Content” เพื่อให้สามารถนำไปใช้ในการประมวลผลภาษาธรรมชาติ (NLP)

ในขั้นตอนสุดท้าย ผู้วิจัยดำเนินการบันทึกข้อมูลอีเมลที่ผ่านการประมวลผลแล้วในรูปแบบไฟล์สัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) โดยจัดเก็บแยกเป็นรายฉบับเพื่อความสะดวกในการจัดการข้อมูลภายหลัง ไฟล์แต่ละรายการตั้งชื่อรูปแบบ “<รหัสระบุชื่อ (ID) >.json” เพื่อสร้างความเป็นเอกลักษณ์ของข้อมูลทั้งหมดในชุดข้อมูล ไฟล์ทั้งหมดถูกจัดเก็บไว้ในโฟลเดอร์ที่กำหนดภายในหน่วยจัดเก็บภายในเครื่อง (Local Storage) เพื่อรองรับการนำไปใช้ในขั้นตอนการพัฒนาแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model) ในลำดับถัดไป กระบวนการ

ดังกล่าวส่งผลให้ข้อมูลอีเมลจากระบบอีเมลคลาวด์ถูกแปลงให้อยู่ในรูปแบบกึ่งโครงสร้าง (Semi-Structure Data) โดยผลลัพธ์การบันทึกข้อมูลในรูปแบบไฟล์สัญกรณ์วัตถุจาวาสคริปต์ (JavaScript Object Notation) แสดงดังภาพที่ 3-3 ซึ่งสะท้อนให้เห็นถึงความครบถ้วนและความเป็นระบบของกระบวนการจัดการข้อมูลที่พัฒนาขึ้นในงานวิจัยนี้



ภาพที่ 3-3 ตัวอย่างไฟล์ข้อมูลอีเมลในรูปแบบไฟล์สัญกรณ์วัตถุจาวาสคริปต์ (JSON)

3.2.2 การทำความสะอาดข้อมูล

เนื่องจากแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model) ไม่สามารถพัฒนาได้โดยตรงจากไฟล์สัญกรณ์วัตถุจาวาสคริปต์ จึงต้องมีกระบวนการแปลงชุดข้อมูลอีเมลจากรูปแบบไฟล์ดังกล่าวไปเป็นรูปแบบค่าที่คั่นด้วยเครื่องหมายจุลภาค (Comma-Separated Values) หรือไฟล์ซีเอสวี (CSV) มีวัตถุประสงค์เพื่อจัดระเบียบข้อมูลที่มีโครงสร้างเชิงลำดับชั้นให้กลายเป็นข้อมูลเชิงตาราง โดยในขั้นตอนแรก ผู้วิจัยดำเนินการโหลดไฟล์สัญกรณ์วัตถุจาวาสคริปต์ จากโฟลเดอร์ต้นทางภายในหน่วยจัดเก็บข้อมูล และอ่านข้อมูลอีเมลแต่ละฉบับเข้าสู่ระบบประมวลผลด้วยโปรแกรมภาษาไพทอน (Python) จากนั้นทำการดึงฟิลด์ (Field) เนื้อความหลักของอีเมลที่ตัด

ส่วนเกินแล้ว (Content) เพื่อให้คงเหลือเฉพาะข้อความที่มีความหมายเชิงเนื้อหาอย่างแท้จริง ภายหลังจากกระบวนการทำความสะอาดแล้ว ผู้วิจัยจึงดำเนินการบันทึกข้อมูลทั้งหมดให้เป็นหนึ่งระเบียน (Record) ในไฟล์ซีเอสวี ซึ่งมีลักษณะเป็นข้อมูลเชิงตาราง จากนั้นผู้วิจัยนำไฟล์ซีเอสวีที่สมบูรณ์แล้วไปดำเนินการสร้างป้ายชื่อกำกับข้อมูล (Labeling) โดยอาศัยการอ่านและพิจารณาวัตถุประสงค์ของอีเมลแต่ละฉบับผ่านโปรแกรมเอ็กซ์เซล (Excel) เพื่อกำหนดหมวดหมู่หรือประเภทของอีเมลตามบริบทของแต่ละประโยค โดยผลลัพธ์ที่ได้คือชุดข้อมูลอีเมลในรูปแบบตารางดังแสดงในตารางที่ 3-1 อันเป็นรูปแบบที่สะดวกต่อการประมวลผลเชิงปริมาณและการจัดการข้อมูลในขั้นตอนต่อไป

ตารางที่ 3-1 ตัวอย่างไฟล์ซีเอสวี (CSV) เก็บบันทึกตัวอย่างอีเมล

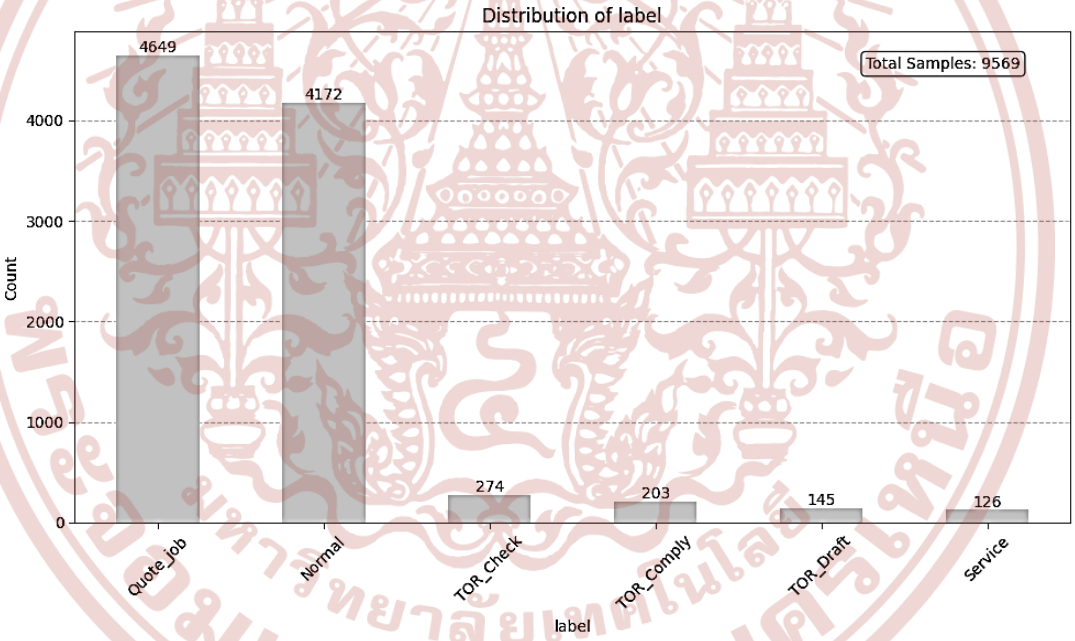
ชื่อไฟล์	**เนื้อหาอีเมล (Body)**	**ป้ายชื่อ (Label)**
AAA	Dear All E-Quotes ตามแนบครับ 684395 Quote Number : 37000220584.1 Thank you	Quote Job
BBB	เรียน ทุกท่าน ขอนำส่งเอกสารตัวอย่างการสาทิธสินค้าในวันงาน ให้ตามไฟล์แนบ	Normal
CCC	เรียน หัวหน้า เนื่องจากมีอาการ ท้องเสียและอาเจียน คาดว่าอาหารเป็นพิษครับ จึงขออนุญาตลา 1 วันครับ	Normal
DDD	เรียน คุณเอ นำส่งเอกสาร Comply Datasheet ตามไฟล์แนบ	ToR Comply
EEE	เรียน ทีมติดตั้ง อ้างอิงอุปกรณ์ตามไฟล์แนบช่วยคำนวนค่าติดตั้งและการดูแลหลังการขายให้ด้วยครับ	Service

หมายเหตุ ตารางนี้เป็นการจำลองตัวอย่างเพื่อแสดงให้เห็นลักษณะข้อมูลอีเมล มิใช้ข้อมูลจริง

หลังจากได้ชุดข้อมูลที่พร้อมนำไปใช้งานแล้ว ซึ่งมีครบทุกองค์ประกอบได้แก่ ตัวระบุเอกลักษณ์ (Uniqu ID) คือชื่อไฟล์ เนื้อหาอีเมลและป้ายชื่อสำหรับสำหรับเป็นข้อมูลในการใช้เรียนรู้แบบจำลองการเรียนรู้ของเครื่อง ถัดมาจะนำชุดข้อมูลดังกล่าวมาการวิเคราะห์การกระจายของป้ายชื่อ (Label Counting) มีวัตถุประสงค์เพื่อประเมินโครงสร้างเชิงปริมาณของชุดข้อมูลในมิติของความสมดุลระหว่างกลุ่ม (Class Distribution) ทั้งนี้ เนื่องจากความไม่สมดุลของข้อมูล (Imbalanced Data) อาจส่งผลกระทบโดยตรงต่อประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model) โดยเฉพาะในงานจำแนกประเภท (Classification) ซึ่งแบบจำลองอาจเอนเอียงไปยังกลุ่มที่มี

จำนวนตัวอย่างมากกว่า ดังนั้นการตรวจสอบสัดส่วนของข้อมูลในแต่ละป้ายชื่อจึงเป็นขั้นตอนสำคัญก่อนเข้าสู่กระบวนการฝึกสอนแบบจำลอง

ผลลัพธ์ที่ได้จะแสดงในรูปแบบค่าความถี่เชิงตัวเลข ซึ่งสะท้อนจำนวนข้อมูลในแต่ละกลุ่มอย่างชัดเจน จากนั้นนำผลลัพธ์ดังกล่าวไปแสดงในรูปแบบกราฟแท่ง (Bar Chart) เพื่อให้สามารถเปรียบเทียบสัดส่วนของแต่ละป้ายชื่อได้อย่างชัดเจนในเชิงภาพ การนำเสนอทั้งในรูปแบบกราฟช่วยให้ผู้วิจัยเข้าใจลักษณะการกระจายของข้อมูลได้อย่างครบถ้วน และสามารถประเมินได้ว่าจำเป็นต้องใช้เทคนิคการปรับสมดุลข้อมูล เช่น การเพิ่มข้อมูลกลุ่มส่วนน้อยหรือการสุ่มลดข้อมูลกลุ่มส่วนมากหรือไม่ ผลการวิเคราะห์ดังกล่าวแสดงไว้ในภาพที่ 3-4 ซึ่งเป็นพื้นฐานสำคัญในการวางแผนขั้นตอนการพัฒนาแบบจำลองในลำดับถัดไปของงานวิจัยนี้



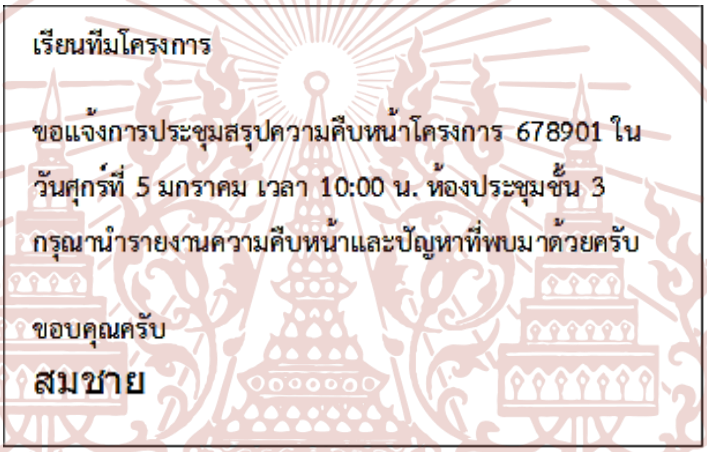
ภาพที่ 3-4 กราฟแท่ง (Bar Chart) แสดงการกระจายของป้ายชื่อ

จากกราฟการกระจายของป้ายชื่อ พบว่าชุดข้อมูลมีจำนวนทั้งหมด 9,569 ตัวอย่าง โดยกลุ่มงานตอบคำถามทั่วไปและงานส่งใบเสนอสินค้า มีสัดส่วนสูงสุดและใกล้เคียงกัน คิดเป็นสัดส่วนหลักของข้อมูลทั้งหมด ขณะที่งานบริการติดตั้ง งานตรวจสอบเอกสารข้อกำหนด งานร่างเอกสารข้อกำหนด และงานส่งเอกสารอ้างอิงข้อกำหนด มีจำนวนค่อนข้างน้อยกว่า

เนื่องจากแบบจำลองไม่สามารถพยากรณ์ได้จากข้อความโดยตรงจึงต้องมีกระบวนการแปลงข้อมูลให้อยู่ในรูปแบบตัวเลขก่อน ซึ่งมีขั้นตอนและวิธีการดังนี้

3.2.2.1. การตัดแบ่งคำภาษาไทย (Thai Text Tokenization Process)

เป็นกระบวนการเตรียมข้อมูลข้อความภาษาไทยจากไฟล์ซีเอสวี (CSV) มีวัตถุประสงค์เพื่อปรับสภาพข้อมูลให้เหมาะสมต่อการวิเคราะห์เชิงภาษาธรรมชาติ (Natural Language Processing) โดยเฉพาะในขั้นตอนการสร้างคุณลักษณะ (Feature Extraction) จากข้อความเนื่องมาจากภาษาไทยมีลักษณะเฉพาะที่ไม่มีการเว้นวรรคระหว่างคำอย่างชัดเจนแสดงตามภาพที่ 3-5 ซึ่งยากต่อการแปลงข้อความเป็นตัวเลข



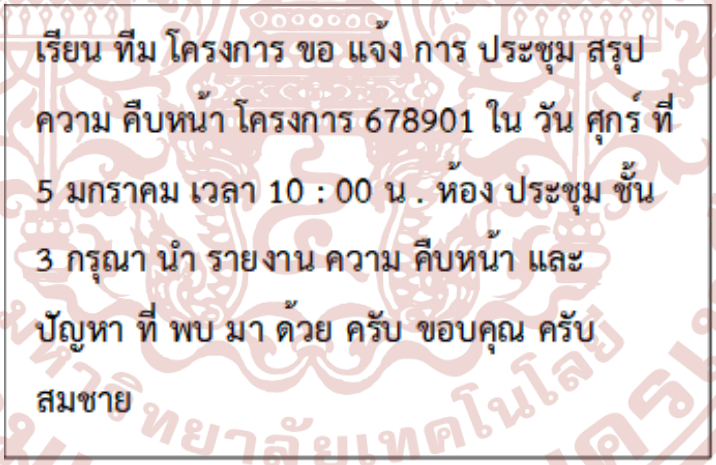
ภาพที่ 3-5 ตัวอย่างเนื้อหาอีเมลที่ยังไม่ผ่านกระบวนการ

ดังนั้นการตัดคำ (Text Tokenization) จึงเป็นขั้นตอนพื้นฐานที่มีผลโดยตรงต่อคุณภาพของคุณลักษณะที่สร้างขึ้น ในขั้นตอนแรก ผู้วิจัยดำเนินการอ่านข้อมูลจากไฟล์ซีเอสวี ที่ได้จากกระบวนการก่อนหน้า โดยใช้คำสั่ง “pd.read_csv()” ภายใต้ไลบรารีแพนดาส (Pandas) เพื่อโหลดข้อมูลเข้าสู่โครงสร้างข้อมูลแบบดาต้าเฟรม (DataFrame) ซึ่งเอื้อต่อการจัดการและประมวลผลข้อมูลเชิงตาราง จากนั้นทำการกำหนดคอลัมน์ที่ต้องการประมวลผล โดยเลือกคอลัมน์ชื่อ “Body” ซึ่งเป็นส่วนเนื้อหาหลักของอีเมลและเป็นแหล่งข้อมูลข้อความที่ใช้ในการวิเคราะห์

ในลำดับถัดมา ผู้วิจัยได้พัฒนาฟังก์ชันย่อยชื่อ “process_text()” เพื่อใช้ในการจัดการข้อความแต่ละระเบียน (Record) อย่างเป็นระบบ โดยฟังก์ชันดังกล่าวถูกออกแบบให้สามารถตรวจสอบและข้ามค่าที่เป็นค่าว่างหรือค่าไม่ปรากฏ (Not a Number) เพื่อป้องกันข้อผิดพลาดระหว่างการประมวลผล ทั้งนี้ ภายในฟังก์ชันมีการเรียกใช้คำสั่ง “word_tokenize(text, engine='newmm')” ซึ่งเป็นกลไกการตัดคำภาษาไทยที่อาศัยพจนานุกรมและอัลกอริทึมแบบการจับคู่สูงสุด (Maximum Matching) ภายใต้เอนจิน (Engine) ชื่อ “newmm” เพื่อแยกข้อความต่อเนื่องออกเป็นรายการคำ (List of Tokens) ตัวอย่างเช่น ข้อความ “การประมวลผล

ภาษาธรรมชาติ” จะถูกแยกออกเป็น ['การ', 'ประมวลผล', 'ภาษา', 'ธรรมชาติ'] ซึ่งสะท้อนหน่วยคำที่มีความหมายอย่างชัดเจน ภายหลังการตัดคำ ผู้วิจัยดำเนินการเชื่อมคำที่อยู่ในรูปแบบรายการให้กลับมาเป็นสตริง (String) เดียว โดยใช้คำสั่ง “" ".join(tokens)” เพื่อคั่นแต่ละคำด้วยช่องว่าง ทั้งนี้การเพิ่มช่องว่างระหว่างคำมีความสำคัญอย่างยิ่งต่อขั้นตอนการแปลงข้อความเป็นเวกเตอร์ (Vectorization) เช่น วิธีถุงคำ (Bag-of-Words) หรือวิธีถ่วงน้ำหนักทีเอฟ-ไอดีเอฟ (Term Frequency–Inverse Document Frequency) ซึ่งต้องอาศัยช่องว่างเป็นตัวแบ่งหน่วยคำ

เมื่อได้ฟังก์ชันที่สามารถประมวลผลข้อความรายระเบียน (Record) แล้ว ผู้วิจัยจึงนำไปประยุกต์ใช้กับข้อมูลทั้งคอลัมน์ของดาต้าเฟรม เพื่อให้ทุกแถวในคอลัมน์ “Body” ถูกแปลงผ่านกระบวนการตัดคำอย่างสอดคล้องและเป็นมาตรฐานเดียวกัน วิธีการดังกล่าวช่วยให้สามารถประมวลผลข้อมูลจำนวนมากได้อย่างมีประสิทธิภาพและลดความคลาดเคลื่อนจากการจัดการแบบรายตัวอย่าง ลักษณะผลลัพธ์ดังกล่าวแสดงตัวอย่างไว้ในภาพที่ 3-6 ซึ่งสะท้อนความพร้อมของข้อมูลสำหรับขั้นตอนถัดไปของงานวิจัย



ภาพที่ 3-6 ตัวอย่างข้อความหลังผ่านขั้นตอน Tokenization

3.2.2.2. การลบคำไม่สำคัญและสัญลักษณ์ในข้อความอีเมล

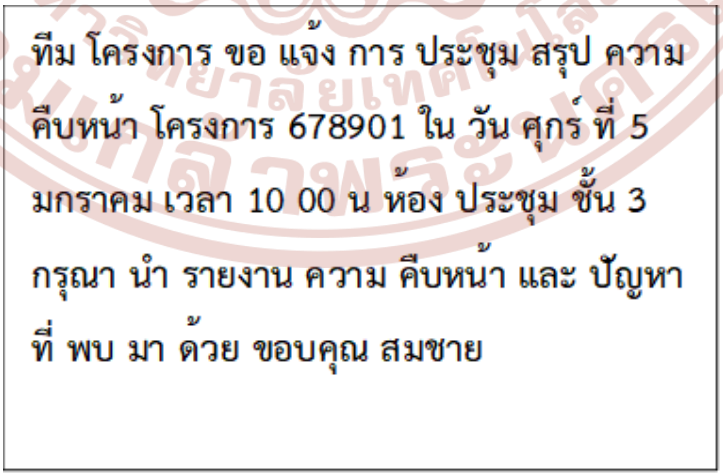
กระบวนการนี้มีวัตถุประสงค์เพื่อปรับปรุงคุณภาพข้อมูลข้อความให้อยู่ในรูปแบบที่เป็นมาตรฐาน มุ่งเน้นการกำจัดคำหรือสัญลักษณ์ที่ไม่ก่อให้เกิดสาระสำคัญเชิงความหมาย (Non-informative Tokens) ออกจากชุดข้อมูล ทั้งนี้ เนื่องจากข้อมูลอีเมลในบริบทการสื่อสารภาษาไทยมักประกอบด้วยคำสุภาพ คำเรียกขาน คำขึ้นต้นจดหมาย และสัญลักษณ์ที่หลากหลาย ซึ่งอาจส่งผลให้แบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) เรียนรู้รูปแบบที่ไม่เกี่ยวข้องกับเนื้อหา

สาระหลัก ดังนั้น การทำความสะอาดข้อมูลในขั้นตอนนี้จึงเป็นกระบวนการสำคัญในการลดสัญญาณรบกวน (Noise Reduction) และเพิ่มคุณภาพของคุณลักษณะ (Feature Quality)

ในขั้นตอนแรก ผู้วิจัยดำเนินการอ่านข้อมูลจากไฟล์ซีเอสวี (CSV) ที่ได้จากกระบวนการก่อนหน้า โดยใช้คำสั่ง “pd.read_csv()” ภายใต้ไลบรารีแพนดาส (Pandas) เพื่อโหลดข้อมูลเข้าสู่โครงสร้างดาต้าเฟรม (DataFrame) ซึ่งเอื้อต่อการจัดการข้อมูลแบบเชิงตาราง จากนั้นจึงกำหนดคอลัมน์ที่ต้องการทำความสะอาด โดยเลือกคอลัมน์ชื่อ “Body” ซึ่งเป็นส่วนที่ผ่านกระบวนการตัดคำ (Text Tokenization) และเป็นองค์ประกอบหลักที่ใช้ในการวิเคราะห์เชิงความหมาย

ลำดับถัดมาเป็นการกำหนดรูปแบบคำหรือกลุ่มคำที่ต้องการลบออกจากข้อมูล โดยแบ่งออกเป็นหลายประเภท ได้แก่ คำสุภาพ (Polite Words) เช่น “ครับ”, “ค่ะ”, “ka”, “kub” ซึ่งเป็นคำลงท้ายที่แสดงมารยาททางภาษาแต่ไม่มีผลต่อสาระสำคัญของเนื้อหา คำนำหน้าชื่อหรือคำเรียกแทนบุคคล (Prefix or Title) เช่น “พี่”, “น้อง”, และ “คุณ” รวมถึงรูปแบบการเขียนภาษาอังกฤษหรืออักษรย่อ เช่น “Mr.”, “Ms.”, “p.”, “n.”, “k.”, “p'”, และ “khun” ซึ่งมักใช้ในบริบทการสื่อสารกึ่งทางการ คำขึ้นต้นประโยคหรือคำเปิดอีเมล เช่น “เรียน”, “Dear”, “Hi”, และ “Hello” ซึ่งโดยทั่วไปทำหน้าที่เป็นเพียงโครงสร้างพิธีการของจดหมายและสัญลักษณ์พิเศษทั้งหมด

ภายหลังการกำหนดรายการคำที่ต้องการลบ ผลลัพธ์ที่ได้คือเนื้อหาอีเมลที่ถูกกำจัดคำไม่สำคัญ คำเชิงพิธีการและสัญลักษณ์ที่ไม่จำเป็นออกไป เหลือเฉพาะข้อความที่มีสาระเชิงเนื้อหาอย่างชัดเจน ดังตัวอย่างที่แสดงในภาพที่ 3-7 ซึ่งแสดงให้เห็นถึงความกระชับและความเป็นมาตรฐานของข้อมูลที่พร้อมนำไปใช้ในขั้นตอนต่อไปของงานวิจัย

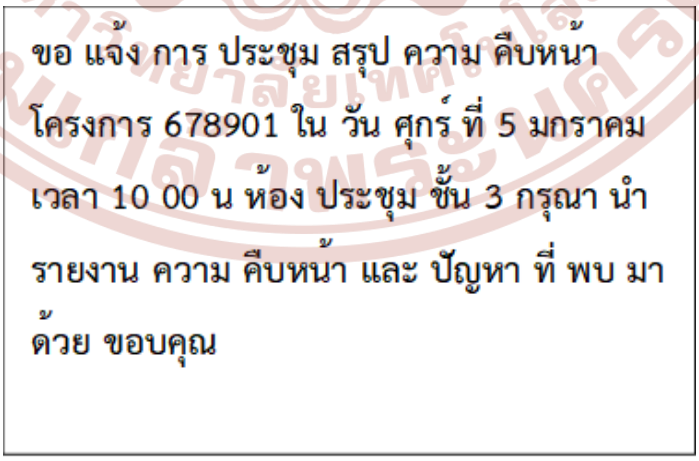


ภาพที่ 3-7 ตัวอย่างข้อความหลังการผ่านลบคำไม่สำคัญและสัญลักษณ์

3.2.2.3. การลบชื่อบุคคลและหน่วยงานในข้อความอีเมล

กระบวนการนี้มีวัตถุประสงค์เพื่อปรับปรุงข้อมูล โดยมุ่งเน้นการลบชื่อบุคคลและชื่อหน่วยงานออกจากเนื้อหา เนื่องจากองค์ประกอบดังกล่าวมิได้มีผลโดยตรงต่อการจำแนกวัตถุประสงค์ของอีเมล และอาจก่อให้เกิดอคติของแบบจำลอง (Model Bias) ในกรณีที่ระบบเรียนรู้ความถี่ของชื่อเฉพาะแทนการเรียนรู้สาระสำคัญของข้อความ ในขั้นตอนแรก ผู้วิจัยอ่านข้อมูลจากไฟล์ซีเอสวี (CSV) ที่ได้จากกระบวนการก่อนหน้า โดยใช้คำสั่ง “pd.read_csv()” ภายใต้ไลบรารีแพนดาส (Pandas) เพื่อโหลดข้อมูลเข้าสู่โครงสร้างดาต้าเฟรม (DataFrame) จากนั้นกำหนดคอลัมน์ที่ต้องการประมวลผล คือคอลัมน์ “Body” ซึ่งเป็นส่วนข้อความหลักของข้อมูล

ถัดมา ผู้วิจัยกำหนดรายการชื่อที่ต้องการลบ โดยสร้างตัวแปรชนิดรายการ (List) สำหรับจัดเก็บรายชื่อที่ตรวจพบจากการอ่านอีเมลทีละฉบับ รายชื่อดังกล่าวประกอบด้วยชื่อจริง นามสกุล ชื่อเล่น เฉพาะกรณีที่เป็นวลีซึ่งไม่ก่อให้เกิดความหมายเชิงเนื้อหา ตลอดจนชื่อบริษัทและชื่อกลุ่ม ทั้งนี้เพื่อให้การลบข้อมูลเป็นไปอย่างครอบคลุมและสอดคล้องกับบริบทของชุดข้อมูล ภายหลังจากจัดเตรียมรายการชื่อแล้ว ผู้วิจัยพัฒนาฟังก์ชันสำหรับลบคำตามเงื่อนไขที่กำหนด โดยกำหนดลำดับการลบเป็นสองรอบ ได้แก่ รอบที่หนึ่งสำหรับลบชื่อจริง นามสกุล ชื่อบริษัท และชื่อกลุ่ม และรอบที่สองสำหรับลบชื่อเล่น เพื่อป้องกันความคลาดเคลื่อนในการแทนที่ข้อความที่อาจซ้อนทับกัน เมื่อดำเนินการประมวลผลครบถ้วนแล้ว ผลที่ได้คือเนื้อหาอีเมลที่ถูกตัดข้อมูลชื่อเฉพาะออก เหลือเฉพาะข้อความที่สะท้อนวัตถุประสงค์ของการสื่อสารอย่างแท้จริง ดังแสดงในภาพที่ 3-8 ซึ่งแสดงถึงความเหมาะสมของข้อมูลที่พร้อมนำไปใช้ในขั้นตอนต่อไปของงานวิจัย

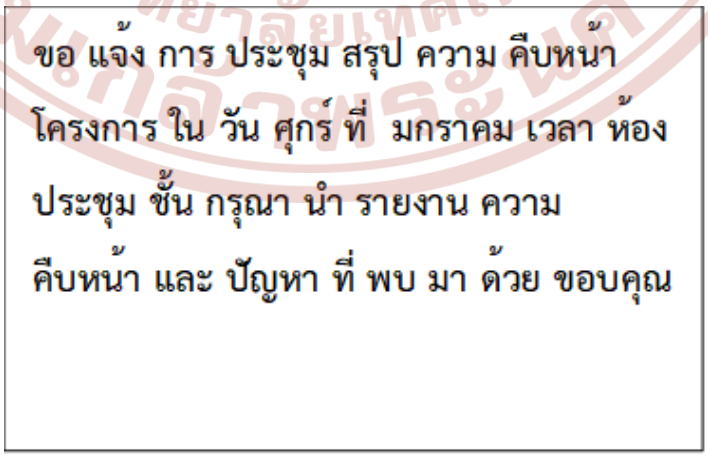


ภาพที่ 3-8 ตัวอย่างข้อความหลังการผ่านลบชื่อบุคคลและหน่วยงานในข้อความอีเมล

3.2.2.4. การลบตัวอักษรเดี่ยวและตัวเลข

การทำความสะอาดข้อความอีเมลรอบที่ 3 มีวัตถุประสงค์เพื่อปรับปรุงคุณภาพข้อมูลให้มีความเป็นมาตรฐาน โดยมุ่งเน้นการลบตัวเลขและตัวอักษรเดี่ยวทั้งภาษาไทยและภาษาอังกฤษ ซึ่งโดยทั่วไปมักไม่สะท้อนสาระสำคัญของเนื้อหา และอาจรบกวนกระบวนการสร้างคุณลักษณะ (Feature Extraction) ในการวิเคราะห์เชิงภาษาธรรมชาติ (Natural Language Processing) ขั้นตอนแรก ผู้วิจัยดำเนินการอ่านข้อมูลจากไฟล์ซีเอสวี (CSV) ที่ได้จากกระบวนการก่อนหน้า โดยใช้คำสั่ง “pd.read_csv()” ภายใต้ไลบรารีแพนดาส (Pandas) เพื่อโหลดข้อมูลเข้าสู่โครงสร้างดาต้าเฟรม (DataFrame) จากนั้นกำหนดคอลัมน์ที่ต้องการประมวลผล คือคอลัมน์ “Body” ซึ่งเป็นข้อมูลข้อความที่ผ่านการทำความสะอาดในรอบก่อนหน้าแล้ว

ลำดับถัดมา ผู้วิจัยได้พัฒนาฟังก์ชันสำหรับตรวจจับและลบตัวเลขทั้งหมด รวมถึงตัวอักษรที่มีความยาวเพียงหนึ่งตัวอักษร ไม่ว่าจะเป็นอักษรภาษาไทยหรือภาษาอังกฤษ โดยอาศัยเงื่อนไขเชิงรูปแบบ (Pattern-based Condition) เพื่อคัดกรองเฉพาะหน่วยข้อความที่ไม่ก่อให้เกิดความหมายเมื่อปรากฏเดี่ยว ๆ เช่น “a”, “b”, “ก”, “ข” หรือเลขลำดับที่ไม่เกี่ยวข้องกับบริบท ทั้งนี้เพื่อให้คงเหลือเฉพาะคำที่มีความสมบูรณ์เชิงความหมายและเหมาะสมต่อการนำไปแปลงเป็นเวกเตอร์ข้อความ (Text Vectorization) ภายหลังการประมวลผลเสร็จสิ้น ผู้วิจัยได้บันทึกผลลัพธ์เป็นไฟล์ CSV ฉบับใหม่ชื่อ “Dataset_Cleaned” เพื่อแยกสถานะของข้อมูลหลังผ่านกระบวนการทำความสะอาดขั้นสุดท้าย ผลลัพธ์ที่ได้คือเนื้อหาอีเมลที่ถูกตัดตัวเลขและตัวอักษรเดี่ยวที่ไม่มีนัยสำคัญออกอย่างครบถ้วน ดังแสดงในภาพที่ 3-9 ซึ่งสะท้อนถึงความพร้อมของชุดข้อมูลสำหรับการวิเคราะห์และการพัฒนาแบบจำลองในขั้นตอนถัดไปของงานวิจัยนี้



ภาพที่ 3-9 ตัวอย่างข้อความหลังการผ่านลบลบตัวอักษรเดี่ยวและตัวเลข

จากกระบวนการทำความความสะอาดข้อมูลทั้งหมดตารางที่ 3-2 จะแสดงให้เห็นว่าแต่ละขั้นตอนจะส่งผลอย่างไรกับเนื้อหาในอีเมลตั้งแต่เริ่มจนจบกระบวนการทำความสะอาดข้อมูล และมีข้อความอะไรที่ถูกตัดออกในแต่ละขั้นตอน

ตารางที่ 3-2 ผลลัพท์ในขั้นตอนทำความสะอาดข้อมูลทั้งหมด

กระบวนการ	เนื้อหา	คำที่ถูกตัดออก
ต้นฉบับ	เรียนทีมโครงการ ขอแจ้งการประชุมสรุปความคืบหน้าโครงการ 678901 ในวันศุกร์ที่ 5 มกราคม เวลา 10:00 น. ห้องประชุมชั้น 3 กรุณานำรายงานความคืบหน้าและปัญหาที่พบมาด้วยครับ ขอบคุณครับ สมชาย	ไม่มี
การตัดแบ่งคำภาษาไทย	เรียน ทีม โครงการ ขอ แจ้ง การ ประชุม สรุป ความ คืบหน้า โครงการ 678901 ใน วัน ศุกร์ ที่ 5 มกราคม เวลา 10 : 00 น . ห้อง ประชุม ชั้น 3 กรุณา นำ รายงาน ความ คืบหน้า และ ปัญหา ที่ พบ มา ด้วย ครับ ขอบคุณ ครับ สมชาย	ไม่มี
ลบคำไม่สำคัญและสัญลักษณ์	ทีม โครงการ ขอ แจ้ง การ ประชุม สรุป ความ คืบหน้า โครงการ 678901 ใน วัน ศุกร์ ที่ 5 มกราคม เวลา 10 00 น ห้อง ประชุม ชั้น 3 กรุณา นำ รายงาน ความ คืบหน้า และ ปัญหา ที่ พบ มา ด้วย ขอบคุณ สมชาย	“เรียน”, “ครับ”, “เครื่องหมายทวิภาค (:)”, “เครื่องหมายมหัพภาค (.)”
ลบชื่อบุคคลและหน่วยงาน	ขอ แจ้ง การ ประชุม สรุป ความ คืบหน้า โครงการ 678901 ใน วัน ศุกร์ ที่ 5 มกราคม เวลา 10 00 น ห้อง ประชุม ชั้น 3 กรุณา นำ รายงาน ความ คืบหน้า และ ปัญหา ที่ พบ มา ด้วย ขอบคุณ	“ทีม”, “โครงการ”, “สมชาย”
ลบตัวอักษรเดี่ยวและตัวเลข	ขอ แจ้ง การ ประชุม สรุป ความ คืบหน้า โครงการ ใน วัน ศุกร์ ที่ มกราคม เวลา ห้อง ประชุม ชั้น กรุณา นำ รายงาน ความ คืบหน้า และ ปัญหา ที่ พบ มา ด้วย ขอบคุณ	“678901”, “5”, “10”, “00”, “น”, “3”

หลังจากผ่านกระบวนการทำความสะอาดข้อมูลทุกขั้นตอน ผลลัพธ์ที่ได้คือเนื้อหาอีเมลที่ถูกลดคำเหลือเฉพาะคำที่มีความหมายและสามารถสื่อถึงบริบทของบทสนธนาได้ แสดงผลลัพท์ตามตัวอย่างในตารางที่ 3-3 จะเห็นว่าจำนวนคำมีขนาดที่ลดลงอย่างเห็นได้ชัดและแต่ละคำมีการวรรคตอนอย่างถูกต้องและมีความหมายเฉพาะในแต่ละคำ ไม่ใช่วลี และไม่ใช่ชื่อ ซึ่งไม่มีความจำเป็นต้องเก็บคำเหล่านั้นไว้ ทั้งนี้เมื่อจำนวนคำลดลงการในข้อความไปสร้างคุณลักษณะ (Feature Extraction) ในการวิเคราะห์เชิงภาษาธรรมชาติ (Natural Language Processing) ได้อย่างมีประสิทธิภาพ สามารถลดเวลาประมวลผลและขนาดของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ได้

ตารางที่ 3-3 ตัวอย่างไฟล์ซีเอสวี (CSV) หลังผ่านกระบวนการทำความสะอาดข้อมูล

ชื่อไฟล์	เนื้อหาอีเมล (Body)	ป้ายชื่อ (Label)
AAA	สอบถาม ข้อมูล เห็น เลือก Liquid Cooling เข้ามา ต้องใช้ หรือไม่ เพราะ จะต้อง ขอ approve ถึงจะ ทำราคา ได้	Normal
BBB	DSC ตาม ไฟล์แนบ Link DSC	Quote job
CCC	นำส่ง ราคา และ ขอบเขต Install Onsite และ Implement	Service
DDD	ไม่สามารถ เสนอ รุ่น ตาม TOR ได้ โดยจะ ติดที่ ข้อ ดังนี้ น้ำหนัก ตัว เครื่อง คอมพิวเตอร์ รวม Battery แล้ว มี น้ำหนัก ไม่เกิน	TOR Review
EEE	Datasheet Comply ตาม ลิงค์	TOR Comply
FFF	Draft TOR ตาม แนบ	TOR Draft

3.3 การสร้างและเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่อง

ในขั้นตอนการพัฒนาแบบจำลองเพื่อจำแนกประเภทข้อความ ผู้วิจัยได้นำเทคนิคการฝังคำ (Word embedding) มาใช้ในการแปลงข้อมูลรูปประโยคให้อยู่ในรูปแบบเวกเตอร์เชิงความหมายเพื่อเป็นข้อมูลนำเข้าให้แก่ระบบ โดยทำการเปรียบเทียบเพื่อหาเทคนิคที่มีประสิทธิภาพสูงสุดจาก 3 อัลกอริทึมหลัก ได้แก่ตัวจำแนกแบบเบย์สแบบง่าย (Naïve Bayes) ตัวจำแนกแบบเพื่อนบ้านใกล้ (K-Nearest Neighbors) และหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory) ซึ่งการเลือกใช้อัลกอริทึมเหล่านี้มีวัตถุประสงค์สำคัญเพื่อคำนวณค่าความน่าจะเป็น (Probability) ของแต่ละคลาสเป้าหมาย โดยผู้วิจัยจะนำค่าความน่าจะเป็นดังกล่าวไปประยุกต์ใช้ในเกณฑ์การตัดสินใจแบบมีเงื่อนไขเพื่อทำนายคลาสลำดับที่สองจากค่าความน่าจะเป็นที่รองลงมา ในกรณีที่ข้อความมีความก้ำกึ่งระหว่างสองประเภท เพื่อเพิ่มความแม่นยำในการจำแนกประเภทข้อมูลให้มีประสิทธิภาพ

3.3.1 การสร้างแบบจำลองตัวจำแนกแบบเบย์สแบบง่าย (Naïve Bayes)

พัฒนาแบบจำลองการจำแนกประเภทข้อความ (Text Classification) โดยประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) แบบดั้งเดิม (Traditional Machine Learning) ร่วมกับเทคนิคการแทนคำในรูปแบบเวกเตอร์เชิงความหมาย ทั้งนี้ ผู้วิจัยเลือกใช้สถาปัตยกรรมตัวจำแนกประเภทเชิงความน่าจะเป็นแบบเบย์สแบบง่าย (Gaussian Naive Bayes) ควบคู่กับเทคนิคการฝังคำ (Word embedding) โดยใช้ชุดคำสั่งชื่อ “Word2Vec” เพื่อสร้างตัวแทนเชิงตัวเลขของข้อมูลข้อความ ซึ่งแนวทางดังกล่าวมีความเหมาะสมกับลักษณะของข้อมูลอีเมลที่ผ่านกระบวนการทำความสะอาดและตัดคำมาแล้วในขั้นตอนก่อนหน้า การผสานวิธีการเชิงสถิติเข้ากับการเรียนรู้ความสัมพันธ์เชิงบริบทของคำ ช่วยให้แบบจำลองสามารถเรียนรู้รูปแบบของภาษาได้อย่างมีประสิทธิภาพมากยิ่งขึ้น ผู้วิจัยดำเนินการแบ่งชุดข้อมูล (Data Splitting) ออกเป็นสองส่วน ได้แก่ ชุดฝึกฝน (Training Set) ร้อยละ 80 และชุดทดสอบ (Test Set) ร้อยละ 20 โดยใช้เทคนิค Stratified Split ซึ่งเป็นวิธีการแบ่งข้อมูลที่คงสัดส่วนของแต่ละคลาสให้ใกล้เคียงกับข้อมูลต้นฉบับมากที่สุด วิธีการดังกล่าวมีความสำคัญอย่างยิ่งในกรณีที่ชุดข้อมูลมีลักษณะไม่สมดุล (Imbalanced Data) เนื่องจากช่วยลดความเอนเอียงของแบบจำลองและทำให้ผลการประเมินสะท้อนประสิทธิภาพที่แท้จริงมากขึ้นเนื่องจากแบบจำลองตัวจำแนกแบบเบย์สแบบง่าย (Naive Bayes) ไม่สามารถประมวลผลข้อความดิบ (Raw Text) ได้โดยตรง ผู้วิจัยจึงนำเทคนิคการฝังคำ (Word embedding) มาใช้ในการแปลงข้อความให้เป็นเวกเตอร์เชิงตัวเลขที่สะท้อนความหมายของคำ ด้วยไลบารี่ Word2Vec เป็นอัลกอริทึมที่เรียนรู้การแทนคำในรูปแบบเวกเตอร์โดยอาศัยบริบทของคำที่ปรากฏร่วมกันในประโยค ส่งผลให้คำที่มีความหมายใกล้เคียงกันมีตำแหน่งใกล้กันในปริภูมิเวกเตอร์ (Vector Space) ในการกำหนดค่าพารามิเตอร์ ผู้วิจัยตั้งค่าขนาดเวกเตอร์ (Vector Size) เท่ากับ 100 มิติ เพื่อให้มีมิติข้อมูลเพียงพอต่อการแทนความหมายโดยไม่ก่อให้เกิดความซับซ้อนเกินความจำเป็น กำหนดค่าขอบเขตพิจารณา (Window Size) เท่ากับ 5 เพื่อพิจารณาคำบริบทก่อนหน้าและถัดไปภายในระยะ 5 คำ และเลือกใช้สถาปัตยกรรมสกิปแกรม (Skip-gram) ซึ่งเหมาะสมกับชุดข้อมูลที่มีขนาดไม่ใหญ่มากและสามารถเรียนรู้ความสัมพันธ์ของคำที่ไม่ปรากฏบ่อยได้ดี นอกจากนี้ กำหนดจำนวนรอบการฝึกฝน (Epochs) ทั้งหมด 20 รอบ เพื่อให้แบบจำลองเรียนรู้ได้อย่างเพียงพอ สำหรับขั้นตอนการสร้างแบบจำลองผู้วิจัยเลือกใช้ตัวจำแนกประเภทเชิงความน่าจะเป็นแบบเบย์สแบบง่าย (Gaussian Naive Bayes) ซึ่งเป็นแบบจำลองเชิงความน่าจะเป็น (Probabilistic Model) ที่อาศัยทฤษฎีบทของเบย์ส (Bayes’ Theorem) ภายใต้สมมติฐานว่าแต่ละคุณลักษณะมีความเป็นอิสระต่อกัน (Conditional

Independence Assumption) เหตุผลที่เลือกใช้เนื่องจากข้อมูลที่แปลงเป็นเวกเตอร์เชิงตัวเลขแล้วมีลักษณะเป็นค่าต่อเนื่อง (Continuous Values) และสามารถมีค่าติดลบได้ ซึ่งสอดคล้องกับสมมติฐานการแจกแจงแบบปกติ (Gaussian Distribution) ของแบบจำลองประเภทนี้ และมีความรวดเร็วในการฝึกสอน เหมาะกับจำนวนข้อมูลที่มีไม่มาก

3.3.2 การสร้างแบบจำลองตัวจำแนกแบบเพื่อนบ้านใกล้ (K-Nearest Neighbors)

เป็นแนวทางการพัฒนาแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning) ที่มุ่งเน้นการแปลงข้อมูลข้อความซึ่งมีลักษณะไม่เป็นโครงสร้าง (Unstructured Data) ให้เป็นข้อมูลเชิงตัวเลขที่สามารถประมวลผลได้ทางคณิตศาสตร์ กระบวนการดังกล่าวมีความสำคัญอย่างยิ่งในงานจำแนกประเภทข้อความ (Text Classification) เนื่องจากข้อความในรูปแบบดิบไม่สามารถนำเข้าสู่แบบจำลองเชิงสถิติได้โดยตรง ผู้วิจัยจึงออกแบบขั้นตอนการดำเนินงานอย่างเป็นระบบ ตั้งแต่การเตรียมข้อมูล การสร้างเวกเตอร์แทนข้อความ การกำหนดสถาปัตยกรรมของแบบจำลอง ตลอดจนการประเมินประสิทธิภาพของแบบจำลอง เพื่อให้ผลการทดลองมีความถูกต้อง ผู้วิจัยได้ดำเนินการแบ่งชุดข้อมูลออกเป็นสองส่วน ได้แก่ ชุดข้อมูลสำหรับการฝึกสอน (Training Set) ร้อยละ 80 และชุดข้อมูลสำหรับการทดสอบ (Test Set) ร้อยละ 20 โดยใช้เทคนิคการแบ่งแบบคงสัดส่วนของแต่ละกลุ่ม (Stratified Split) เพื่อรักษาสัดส่วนของแต่ละประเภทข้อมูลให้ใกล้เคียงกับชุดข้อมูลต้นฉบับ อันเป็นการลดอคติของแบบจำลอง (Model Bias) ภายหลังจากการเตรียมข้อมูลเรียบร้อยแล้ว ขั้นตอนถัดมาคือการแปลงข้อความให้อยู่ในรูปแบบเวกเตอร์เชิงความหมายด้วยเทคนิคการฝังคำ (Word embedding) ผู้วิจัยกำหนดค่าพารามิเตอร์ที่สำคัญ ได้แก่ ขนาดเวกเตอร์ (Vector Size) เท่ากับ 100 มิติ เพื่อให้ได้ตัวแทนคำที่มีรายละเอียดเชิงความหมายเพียงพอ ค่าขอบเขตพิจารณา (Window Size) เท่ากับ 5 คำ เพื่อพิจารณาบริบทโดยรอบ และเลือกใช้สถาปัตยกรรมแบบสกิปแกรม (Skip-gram) ซึ่งเหมาะสมกับชุดข้อมูลที่มีขนาดไม่ใหญ่มาก ทั้งนี้ได้กำหนดจำนวนรอบการฝึกสอน (Epochs) เท่ากับ 20 รอบ เพื่อให้แบบจำลองสามารถเรียนรู้รูปแบบทางภาษาได้อย่างเพียงพอ สำหรับสถาปัตยกรรมของแบบจำลอง ผู้วิจัยเลือกใช้เทคนิคตัวจำแนกแบบเพื่อนบ้านใกล้ (K-Nearest Neighbors) ซึ่งเป็นอัลกอริทึมเชิงระยะทาง (Distance-based Algorithm) ที่อาศัยหลักการวัดความใกล้เคียงของข้อมูลในปริภูมิเวกเตอร์ (Vector Space) ในการจำแนกประเภท โดยผู้วิจัยกำหนดจำนวนเพื่อนบ้าน เท่ากับ 5 โดยพิจารณาจากหลักการลดความผันผวนของผลลัพธ์ (Variance) และลดผลกระทบจากสัญญาณรบกวน (Noise) ในข้อมูล จากนั้นพิจารณาประเภทของข้อมูลส่วนใหญ่ในกลุ่มเพื่อนบ้านที่ใกล้ที่สุดเพื่อกำหนดผลลัพธ์สุดท้าย

3.3.3 การสร้างแบบจำลองหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory)

ภายหลังจากการทำความสะอาดข้อมูล ผู้วิจัยแบ่งชุดข้อมูลออกเป็นสองส่วน ได้แก่ ชุดข้อมูลสำหรับการฝึกสอน (Training Set) ร้อยละ 80 และชุดข้อมูลสำหรับการทดสอบ (Test Set) ร้อยละ 20 โดยใช้เทคนิคการแบ่งข้อมูลแบบคงสัดส่วน (Stratified Sampling) เพื่อรักษาการกระจายตัวของข้อมูลในแต่ละคลาสให้ใกล้เคียงกับชุดข้อมูลต้นฉบับ วิธีการนี้ช่วยลดอคติของแบบจำลองและเพิ่มความน่าเชื่อถือของผลการประเมิน เนื่องจากแบบจำลอง LSTM ไม่สามารถประมวลผลข้อความดิบ (Raw Text) ได้โดยตรง จึงจำเป็นต้องแปลงข้อความให้อยู่ในรูปแบบลำดับตัวเลข (Numerical Sequence) จากข้อความที่ผ่านกระบวนการตัดคำ (Tokenization) เรียบร้อยแล้ว โดยผู้วิจัยกำหนดขนาดคลังคำศัพท์สูงสุด (Maximum Vocabulary Size) เท่ากับ 20,000 คำ เพื่อจำกัดจำนวนคำที่ใช้ในการเรียนรู้ และลดความซับซ้อนของแบบจำลอง จากนั้นกำหนดความยาวลำดับสูงสุดของข้อความ (Maximum Sequence Length) เท่ากับ 100 คำ เพื่อให้ข้อมูลทุกตัวอย่างมีขนาดเท่ากัน กระบวนการปรับความยาวลำดับดำเนินการโดยการเติมศูนย์ (Padding) และการตัดคำส่วนเกิน (Truncating) ที่ส่วนท้ายของข้อความ (Post-padding และ Post-truncating) ซึ่งช่วยให้โครงข่ายสามารถประมวลผลข้อมูลในลักษณะเป็นเมทริกซ์ที่มีมิติแน่นอนได้ เพื่อเพิ่มประสิทธิภาพในการเรียนรู้ความหมายของคำ ผู้วิจัยได้สร้างเมทริกซ์ตัวแทนคำ (Embedding Matrix) โดยใช้เทคนิคการฝังคำ (Word embedding) แบบ Word2Vec ซึ่งเป็นวิธีการเรียนรู้การแทนคำในปริภูมิเวกเตอร์ (Vector Space Representation) เทคนิคดังกล่าวช่วยให้คำที่มีความหมายใกล้เคียงกันมีตำแหน่งใกล้กันในปริภูมิเวกเตอร์ ส่งผลให้แบบจำลองสามารถเข้าใจความสัมพันธ์เชิงความหมายได้ดียิ่งขึ้น ในการตั้งค่าพารามิเตอร์ของ Word2Vec ผู้วิจัยกำหนดขนาดเวกเตอร์ (Vector Dimension) เท่ากับ 100 มิติ และกำหนดขนาดหน้าต่างบริบท (Context Window) เท่ากับ 5 คำ เพื่อให้แบบจำลองพิจารณาความสัมพันธ์ของคำในบริบทใกล้เคียง จากนั้นจึงนำเวกเตอร์คำที่ได้มาสร้างเป็นค่าน้ำหนักเริ่มต้น (Pre-trained Weights) ในชั้นตัวแทนคำ (Embedding Layer) ของโครงข่ายประสาทเทียม ทั้งนี้ได้กำหนดให้ชั้น Embedding สามารถปรับจูนค่า (Fine-tuning) ระหว่างการฝึกฝน แบบจำลองที่พัฒนาขึ้นมีโครงสร้างแบบสองทิศทาง (Bidirectional LSTM) ซึ่งสามารถประมวลผลข้อมูลลำดับได้ทั้งในทิศทางไปข้างหน้า (Forward Pass) และย้อนกลับ (Backward Pass) การประมวลผลสองทิศทางช่วยให้แบบจำลองรับรู้บริบทของคำจากทั้งคำก่อนหน้าและคำถัดไป ทำให้การสกัดคุณลักษณะ (Feature Extraction) มีความสมบูรณ์มากขึ้น โดยโครงสร้างของแบบจำลองประกอบด้วยชั้นต่าง ๆ ได้แก่ ชั้นแรก “Embedding Layer” ทำหน้าที่รับลำดับตัวเลขที่ได้จาก

กระบวนการ Tokenization และแปลงเป็นเวกเตอร์เชิงความหมายโดยใช้ค่าน้ำหนักจากเวกเตอร์เริ่มต้น ชั้นที่ 2 “Bidirectional LSTM Layer” กำหนดจำนวนหน่วยความจำ (Units) เท่ากับ 64 หน่วย เพื่อเรียนรู้รูปแบบความสัมพันธ์เชิงลำดับของข้อความ ชั้นที่ 3 “Dropout Layer” กำหนดอัตราการปิดหน่วยประสาท (Dropout Rate) เท่ากับ 0.3 เพื่อลดการพึ่งพาหน่วยประสาทบางหน่วยมากเกินไปและป้องกันปัญหาการจดจำข้อมูลเฉพาะชุดฝึก ชั้นที่ 4 “Dense Layer” เป็นชั้นประสานเต็ม (Fully Connected Layer) ขนาด 64 หน่วย ใช้ฟังก์ชันกระตุ้นแบบเรลู (Rectified Linear Unit) เพื่อเพิ่มความไม่เชิงเส้น (Non-linearity) ให้กับแบบจำลอง ชั้นที่ 5 “Output Layer” ใช้ฟังก์ชันกระตุ้นแบบซอฟต์แมกซ์ (Softmax Function) เพื่อคำนวณค่าความน่าจะเป็นแต่ละป้าย ชื่อการฝึกฝนแบบจำลองกำหนดค่าตัวปรับเหมาะ (Optimizer) เป็นแบบอดัม (Adaptive Moment Estimation) ซึ่งมีข้อดีในการปรับอัตราการเรียนรู้ (Learning Rate) แบบปรับตัวอัตโนมัติ ทำให้กระบวนการเรียนรู้มีเสถียรภาพและรวดเร็วมากขึ้น ฟังก์ชันความสูญเสีย (Loss Function) ที่ใช้คือเอนโทรปีไขว้เชิงจัดหมวดหมู่ (Categorical Cross-Entropy) ผู้วิจัยกำหนดจำนวนรอบการฝึก (Epochs) เท่ากับ 10 รอบ และกำหนดขนาดกลุ่มข้อมูลย่อย (Batch Size) เท่ากับ 32 ตัวอย่างต่อหนึ่งรอบการปรับปรุงค่าน้ำหนัก

3.3.4 การวัดผล (Evaluate) และเปรียบเทียบแบบจำลองการเรียนรู้ของเครื่อง

ในการประเมินประสิทธิภาพของแบบจำลองทั้งสาม ผู้วิจัยกำหนดแนวทางการวัดผลภายใต้กรอบเดียวกัน เพื่อให้สามารถเปรียบเทียบผลได้อย่างเป็นธรรมและมีความน่าเชื่อถือ โดยใช้ชุดข้อมูลทดสอบ (Test Set) เดียวกันเป็นเกณฑ์มาตรฐาน ตัวชี้วัดหลัก (Evaluation Metrics) ที่ใช้ประกอบด้วย ค่าความถูกต้องโดยรวม (Accuracy) ค่าความแม่นยำ (Precision) ค่าความครอบคลุม (Recall) และค่าดัชนีเอฟหนึ่ง (F1-score) ซึ่งคำนวณจากรายงานผลการจำแนก (Classification Report) นอกจากนี้ยังใช้เมทริกซ์ความสับสน (Confusion Matrix) เพื่อวิเคราะห์ความถูกต้องและความผิดพลาดของการจำแนกในแต่ละคลาสอย่างละเอียด

ในเชิงการพิจารณาผลลัพธ์ ผู้วิจัยให้ความสำคัญกับค่า F1-score เป็นตัวชี้วัดหลัก เนื่องจากเป็นค่าที่สะท้อนสมดุลระหว่าง Precision และ Recall โดยเฉพาะในกรณีที่ข้อมูลมีความไม่สมดุลระหว่างจำนวนตัวอย่างของแต่ละคลาส ขณะเดียวกันค่า Accuracy ถูกใช้เป็นตัวชี้วัดภาพรวมของประสิทธิภาพทั้งหมด ส่วนค่า Precision และ Recall จะถูกวิเคราะห์เพิ่มเติมในแต่ละคลาสเพื่อพิจารณาความสามารถของแบบจำลองในการลดความผิดพลาดประเภทบวกเทียม (False Positive) และลบเทียม (False Negative) ตามลำดับ

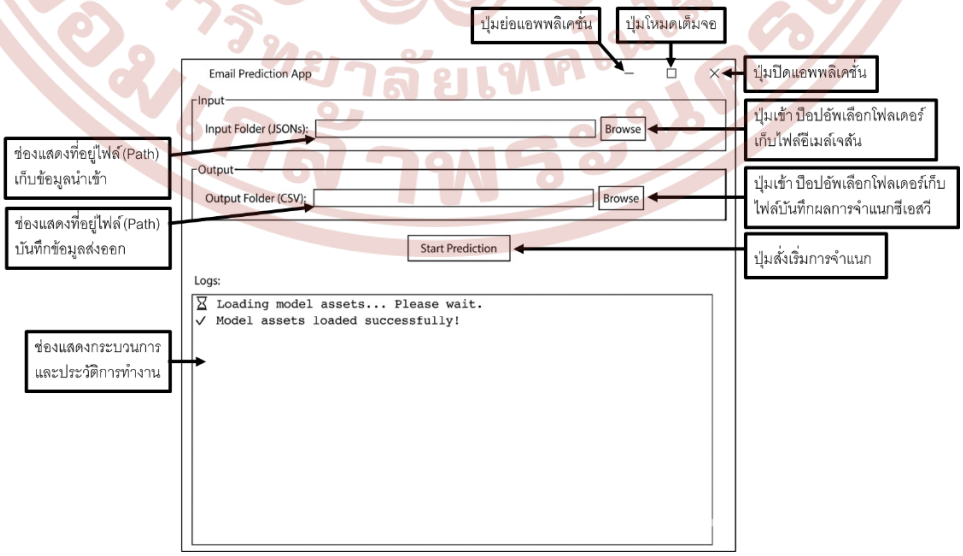
3.4 การออกแบบแอปพลิเคชั่นจำแนกวัตถุประสงค์อีเมล

หลังจากเสร็จขั้นตอนการสร้างแบบจำลองการเรียนรู้ของเครื่องแล้ว เพื่อให้สามารถใช้งานได้อย่างสะดวกจำเป็นต้องรวบรวมซอร์สโค้ดทั้งหมดประยุกต์เป็นหน้าผู้ใช้ (GUI) ที่พร้อมสำหรับนำเข้าข้อมูลแล้วประมวลผลตามลำดับแบบอัตโนมัติเพื่อให้ได้ผลการจำแนกในรูปแบบไฟล์ซีเอสวี (CSV) เพื่อให้สามารถนำไปวิเคราะห์เชิงลึกได้ โดยภาพที่ 3-10 จะอธิบายลำดับการทำงานของแอปพลิเคชั่น



ภาพที่ 3-10 Flow Chart การทำงานของแอพลิเคชั่นจำแนกวัตถุประสงค์อีเมล

แอปพลิเคชั่นที่พัฒนาขึ้นจะเป็นไฟล์พร้อมใช้งานในลักษณะไฟล์ปฏิบัติการ (Executable File) ที่สามารถเปิดใช้งานได้บนระบบปฏิบัติการวินโดว์และรองรับการแก้ไขโมเดลในอนาคตได้ โดยภาพที่ 3-11 จะแสดงให้เห็นเป็นโครงร่างหน้าผู้ใช้ (GUI) สำหรับแอพลิเคชั่นจำแนกวัตถุประสงค์อีเมล



ภาพที่ 3-11 โครงร่างหน้าผู้ใช้ (GUI) สำหรับแอพลิเคชั่นจำแนกวัตถุประสงค์อีเมล

บทที่ 4
ผลการวิจัย

งานค้นคว้าอิสระเรื่องการจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง นำเสนอผลการวิจัยจากการพัฒนาและประเมินประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องสำหรับการจำแนกวัตถุประสงค์ของอีเมล โดยมีลำดับขั้นตอนดังต่อไปนี้

4.1 การประเมินผลการเรียนรู้ของเครื่อง (Model Evaluation)

4.2 การปรับปรุงแบบจำลองการเรียนรู้ของเครื่อง (Model Optimization)

4.3 การปรับปรุงแบบจำลองการเรียนรู้ของเครื่องสำหรับทำนาย 2 วัตถุประสงค์

4.4 การสร้างแอปพลิเคชั่นจำแนกวัตถุประสงค์อีเมล

4.1 การประเมินผลการเรียนรู้ของเครื่อง (Model Evaluation)

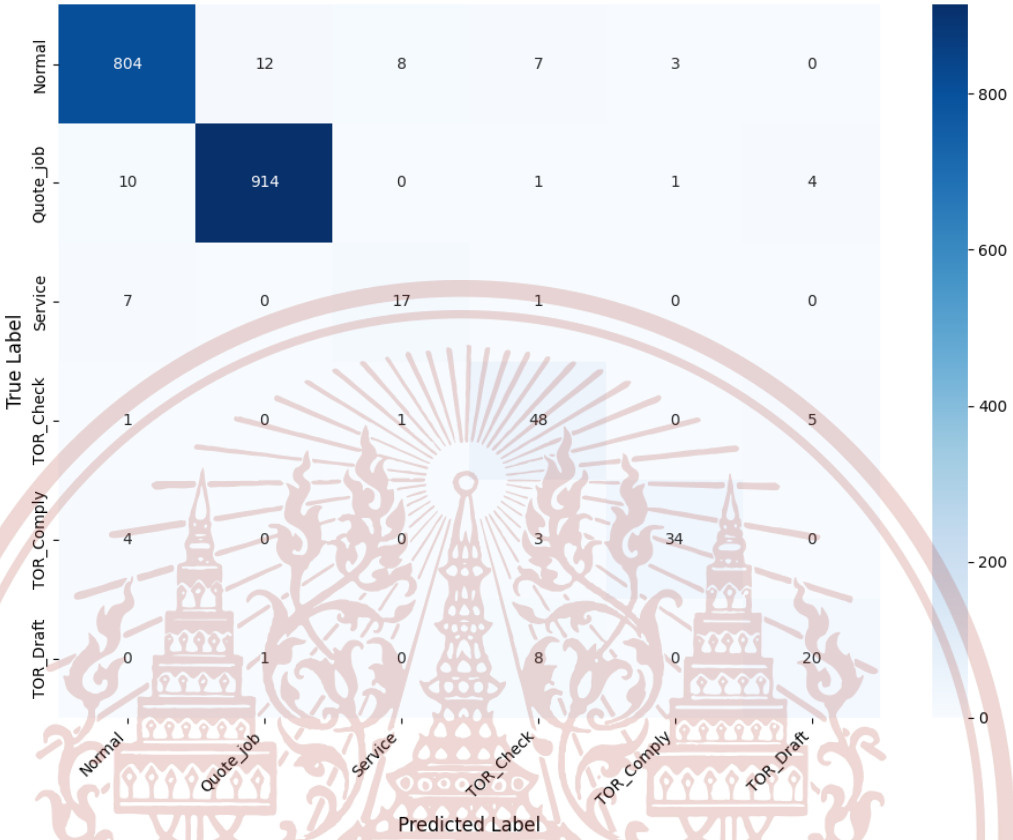
4.1.1 แบบจำลองหน่วยความจำระยะยาวแบบสั้น (Long Short-Term Memory)

จากการประเมินการชุดข้อมูลจำนวน 1,914 ตัวอย่างได้ผลลัพท์ความแม่นยำ (Accuracy) อยู่ที่ร้อยละ 96 ในกรณีวิเคราะห์แยกแต่ละป้ายชื่อสามารถสรุปได้ตามตารางที่ 4-1

ตารางที่ 4-1 ผลประเมินแบบจำลองแบบหน่วยความจำระยะยาวแบบสั้น (LSTM)

วัตถุประสงค์	Precision	Recall	F1-score	Sample
งานตอบคำถามทั่วไป	0.97	0.96	0.97	834
งานส่งใบเสนอสินค้า	0.99	0.98	0.98	930
งานบริการติดตั้ง	0.65	0.68	0.67	25
งานตรวจสอบเอกสารข้อกำหนด	0.71	0.87	0.78	55
งานส่งเอกสารอ้างอิงข้อกำหนด	0.89	0.83	0.86	41
งานร่างเอกสารข้อกำหนด	0.69	0.69	0.69	29
ค่าเฉลี่ย	**0.82**	**0.84**	**0.83**	

สำหรับการวิเคราะห์ Confusion Matrix ของ LSTM แสดงตามภาพที่ 4-1 โดยจะอธิบายในรูปแบบความเข้มของสี ซึ่งจะบงบอกถึงจำนวนตัวอย่างที่ถูกทำนายซึ่งในแนวตั้งคือป้ายชื่อจริงของตัวอย่างและในแนวนอนคือป้ายชื่อที่ได้จากการทำนายของแบบจำลอง



ภาพที่ 4-1 การเปรียบเทียบ Confusion Matrix ของเทคนิค Long Short-Term Memory

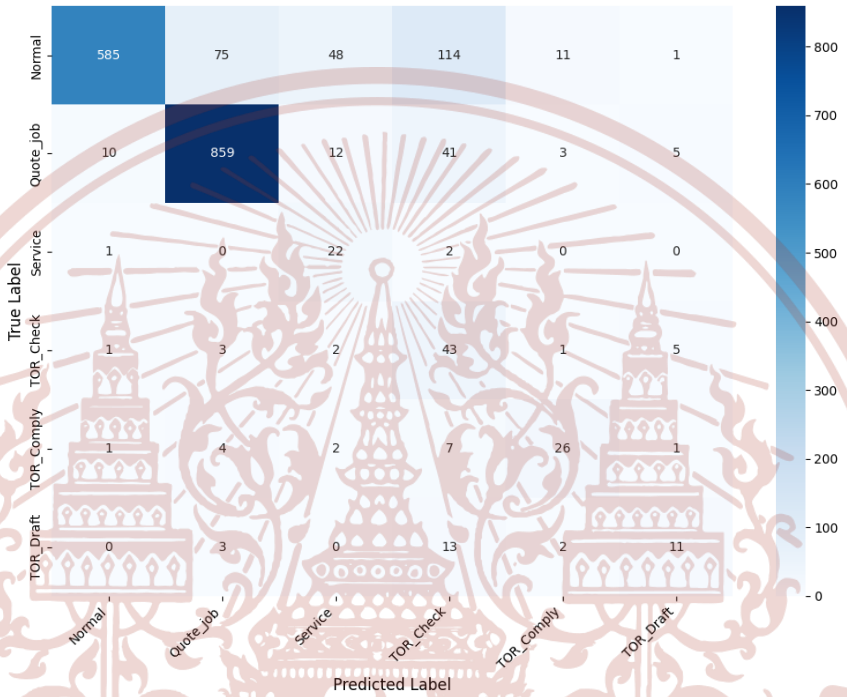
4.1.2 แบบจำลองเบย์สแบบง่าย (Naïve Bayes)

จากการประเมินการชุดข้อมูลจำนวน 1,914 ตัวอย่างได้ผลลัพท์ความแม่นยำ (Accuracy) อยู่ที่ร้อยละ 81 ในกรณีวิเคราะห์แยกแต่ละป้ายชื่อสามารถสรุปได้ตามตารางที่ 4-2

ตารางที่ 4-2 ผลประเมินแบบจำลองเบย์สแบบง่าย (Naïve Bayes)

วัตถุประสงค์	Precision	Recall	F1-score	Sample
งานตอบคำถามทั่วไป	0.98	0.70	0.82	834
งานส่งใบเสนอสินค้า	0.91	0.92	0.92	930
งานบริการติดตั้ง	0.26	0.88	0.40	25
งานตรวจสอบเอกสารข้อกำหนด	0.20	0.78	0.31	55
งานส่งเอกสารอ้างอิงข้อกำหนด	0.60	0.63	0.62	41
งานร่างเอกสารข้อกำหนด	0.48	0.38	0.42	29
ค่าเฉลี่ย	**0.57**	**0.72**	**0.58**	

สำหรับการวิเคราะห์ Confusion Matrix ของ Naïve Bayes แสดงตามภาพที่ 4-2 โดยจะอธิบายในรูปแบบความเข้มของสี ซึ่งจะบงบอกถึงจำนวนตัวอย่างที่ถูกทำนายซึ่งในแนวตั้งคือป้ายชื่อจริงของตัวอย่างและในแนวนอนคือป้ายชื่อที่ได้จากการทำนายของแบบจำลอง



ภาพที่ 4-2 การเปรียบเทียบ Confusion Matrix ของเทคนิค Naïve Bayes

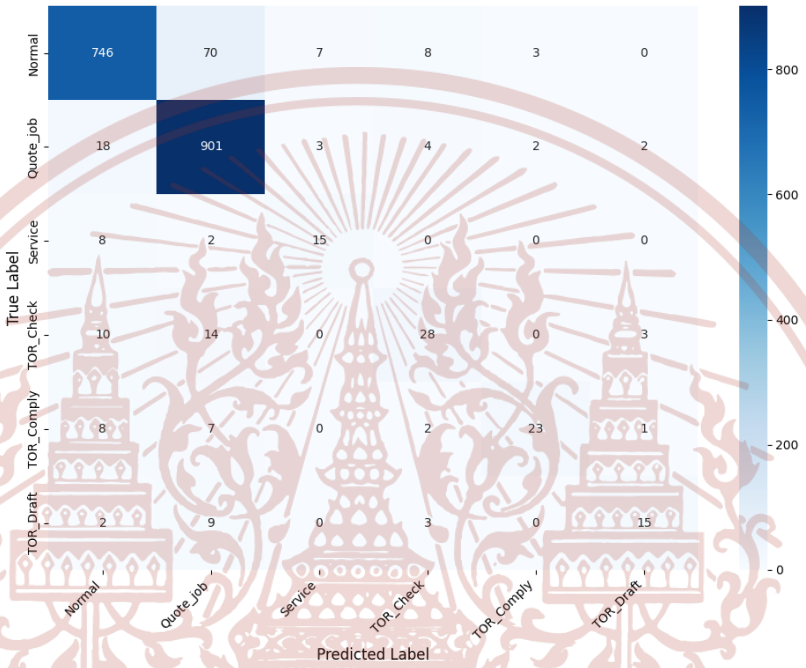
4.1.3 แบบจำลองเพื่อนบ้านใกล้ (K-Nearest Neighbors)

จากการประเมินการชุดข้อมูลจำนวน 1,914 ตัวอย่างได้ผลลัพท์ความแม่นยำ (Accuracy) อยู่ที่ร้อยละ 90 ในกรณีวิเคราะห์แยกแต่ละป้ายชื่อสามารถสรุปได้ตามตารางที่ 4-3

ตารางที่ 4-3 ผลประเมินแบบจำลองแบบหน่วยความจำระยะยาวแบบสั้น (KNN)

วัตถุประสงค์	Precision	Recall	F1-score	Sample
งานตอบคำถามทั่วไป	0.94	0.89	0.92	834
งานส่งใบเสนอสินค้า	0.90	0.97	0.93	930
งานบริการติดตั้ง	0.60	0.60	0.60	25
งานตรวจสอบเอกสารข้อกำหนด	0.62	0.51	0.56	55
งานส่งเอกสารอ้างอิงข้อกำหนด	0.82	0.56	0.67	41
งานร่างเอกสารข้อกำหนด	0.71	0.52	0.60	29
ค่าเฉลี่ย	0.77	0.68	0.71	

สำหรับการวิเคราะห์ Confusion Matrix ของ KNN แสดงตามภาพที่ 4-3 โดยจะอธิบายในรูปแบบความเข้มของสี ซึ่งจะบงบอกถึงจำนวนตัวอย่างที่ถูกทำนายซึ่งในแนวตั้งคือป้ายชื่อจริงของตัวอย่างและในแนวนอนคือป้ายชื่อที่ได้จากการทำนายของแบบจำลอง



ภาพที่ 4-3 การเปรียบเทียบ Confusion Matrix ของเทคนิค K-Nearest Neighbors

หลังจากการประเมินผลการเรียนรู้ของเครื่อง (Evaluate) ทุกเทคนิคสำเร็จ เมื่อนำผลลัพธ์ที่ได้มาแสดง จะเห็นถึงความแตกต่างกันตามตารางที่ 4-4

ตารางที่ 4-4 การเปรียบเทียบภาพรวมของทุกแบบจำลอง

ตัวชี้วัดประสิทธิภาพ (Metrics)	**LSTM**	**NVB**	**KNN**
F1-score แยกตามกลุ่ม			
งานตอบคำถามทั่วไป	0.97	0.82	0.92
งานส่งใบเสนอสินค้า	0.98	0.92	0.93
งานบริการติดตั้ง	0.67	0.40	0.60
งานตรวจสอบเอกสารข้อกำหนด	0.78	0.31	0.56
งานส่งเอกสารอ้างอิงข้อกำหนด	0.86	0.62	0.67
งานร่างเอกสารข้อกำหนด	0.69	0.42	0.60
ค่าเฉลี่ย F1-score	0.83	0.58	0.71
ความแม่นยำ (Accuracy)	0.96	0.81	0.90

จากการวิเคราะห์และเปรียบเทียบประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่องทั้ง 3 เทคนิค ได้แก่ Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN) และ Gaussian Naive Bayes พบว่าแต่ละแบบจำลองมีจุดเด่นและข้อจำกัดที่แตกต่างกันตามลักษณะของอัลกอริทึมและกลไกการประมวลผลข้อมูล อย่างไรก็ตาม เมื่อพิจารณาจากผลการทดลองโดยรวม พบว่าแบบจำลอง LSTM ให้ประสิทธิภาพสูงที่สุด โดยมีค่าความแม่นยำสูงถึงร้อยละ 96 ซึ่งสะท้อนให้เห็นถึงศักยภาพของสถาปัตยกรรมโครงข่ายประสาทเทียมเชิงลึกที่สามารถเรียนรู้รูปแบบของข้อมูลลำดับได้อย่างมีประสิทธิภาพ อัลกอริทึมดังกล่าวได้รับการออกแบบให้สามารถจดจำความสัมพันธ์ระยะยาวของข้อมูลภายในลำดับข้อความ ทำให้สามารถจับบริบทของคำและโครงสร้างของประโยคได้ดีกว่าวิธีการจำแนกแบบดั้งเดิม ส่งผลให้การทำนายประเภทของข้อความมีความถูกต้องมากยิ่งขึ้น โดยเฉพาะในกรณีของข้อมูลข้อความที่มีความซับซ้อนทางภาษาและมีความสัมพันธ์เชิงบริบทระหว่างคำภายในประโยค ในขณะที่แบบจำลอง KNN แสดงผลลัพธ์ที่มีความแม่นยำในระดับที่น่าพึงพอใจ โดยมีค่าความแม่นยำอยู่ที่ร้อยละ 90 อัลกอริทึมนี้มีลักษณะเด่นในด้านความเรียบง่ายของโครงสร้าง เนื่องจากไม่ต้องผ่านกระบวนการเรียนรู้แบบเชิงลึก แต่ใช้หลักการคำนวณระยะห่างระหว่างเวกเตอร์ของข้อมูลเพื่อค้นหาตัวอย่างที่มีความคล้ายคลึงกันมากที่สุดภายในชุดข้อมูลฝึกสอน จากนั้นจึงใช้ข้อมูลของเพื่อนบ้านที่ใกล้ที่สุดในการตัดสินใจจำแนกประเภทของข้อมูล วิธีการดังกล่าวสามารถให้ผลลัพธ์ที่ดีในกรณีที่ข้อมูลมีรูปแบบการกระจายตัวที่ชัดเจนในพื้นที่เวกเตอร์หลายมิติ อย่างไรก็ตาม ประสิทธิภาพของอัลกอริทึมอาจลดลงเมื่อจำนวนมิติของข้อมูลเพิ่มขึ้นหรือเมื่อข้อมูลมีความซับซ้อนสูง ซึ่งเป็นลักษณะที่พบได้บ่อยในข้อมูลข้อความจากภาษาธรรมชาติ สำหรับแบบจำลอง Gaussian Naive Bayes แม้ว่าจะให้ค่าความแม่นยำต่ำกว่าสองเทคนิคก่อนหน้า โดยมีค่าความแม่นยำอยู่ที่ร้อยละ 81 แต่แบบจำลองดังกล่าวยังคงมีข้อได้เปรียบที่สำคัญในด้านประสิทธิภาพการคำนวณ เนื่องจากใช้หลักการทางสถิติและทฤษฎีความน่าจะเป็นในการคำนวณความน่าจะเป็นของแต่ละคลาส โดยมีสมมติฐานว่าคุณลักษณะของข้อมูลมีความเป็นอิสระต่อกัน ส่งผลให้แบบจำลองสามารถประมวลผลข้อมูลได้อย่างรวดเร็วและใช้ทรัพยากรในการคำนวณน้อยกว่าแบบจำลองอื่น จึงเหมาะสำหรับการใช้เป็นแบบจำลองพื้นฐานหรือใช้ในกรณีที่ต้องการระบบที่สามารถประมวลผลข้อมูลจำนวนมากได้อย่างรวดเร็ว เมื่อพิจารณาผลการเปรียบเทียบโดยรวม สามารถสรุปได้ว่าแบบจำลอง LSTM เป็นเทคนิคที่มีความเหมาะสมมากที่สุดสำหรับปัญหาการจำแนกข้อความในงานวิจัยนี้ เนื่องจากไม่เพียงแต่ให้ค่าความแม่นยำสูงสุดเท่านั้น แต่ยังสามารถสร้างค่าความน่าจะเป็นของการทำนายที่มีความน่าเชื่อถือมากกว่าแบบจำลองอื่น ๆ อีกด้วย คุณสมบัติดังกล่าวมีความสำคัญอย่างยิ่งต่อการนำผลลัพธ์ไป

ประยุกต์ใช้ในกระบวนการตัดสินใจแบบมีเงื่อนไข โดยเฉพาะการวิเคราะห์ค่าความน่าจะเป็นของคลาสที่มีลำดับความสำคัญรองลงมาเพื่อใช้ในการทำนายคลาสลำดับที่สองของข้อความ ซึ่งเป็นแนวทางที่ช่วยเพิ่มความยืดหยุ่นในการตีความผลลัพธ์ของระบบและช่วยให้แบบจำลองสามารถรองรับสถานการณ์ที่ข้อความหนึ่งอาจมีวัตถุประสงค์มากกว่าหนึ่งประเภทได้อย่างมีประสิทธิภาพมากกว่าอัลกอริทึมอื่นที่นำมาเปรียบเทียบในการศึกษานี้ ทั้งนี้ ผลการวิเคราะห์ดังกล่าวแสดงให้เห็นถึงศักยภาพของเทคนิคการเรียนรู้เชิงลึกในการประมวลผลภาษาธรรมชาติ และยืนยันถึงความเหมาะสมของการเลือกใช้ LSTM เป็นแบบจำลองหลักสำหรับการพัฒนาระบบต้นแบบในงานวิจัยฉบับนี้

4.2 การปรับปรุงแบบจำลองการเรียนรู้ของเครื่อง (Model Optimization)

จากการประเมินประสิทธิภาพของแบบจำลองในระยะเริ่มต้น ผู้วิจัยได้ดำเนินการปรับปรุงสถาปัตยกรรมและกระบวนการเรียนรู้ของแบบจำลอง เพื่อเพิ่มขีดความสามารถในการจำแนกประเภทข้อความที่มีความซับซ้อนและมีบริบททางภาษาแตกต่างกัน การปรับปรุงดังกล่าวมีวัตถุประสงค์เพื่อยกระดับความแม่นยำของการทำนาย รวมทั้งลดผลกระทบจากข้อมูลรบกวนที่อาจส่งผลต่อประสิทธิภาพของแบบจำลอง โดยแนวทางการปรับปรุงครอบคลุมทั้งด้านคุณภาพของข้อมูล กระบวนการสกัดคุณลักษณะ และการปรับแต่งพารามิเตอร์ของแบบจำลองให้เหมาะสมกับลักษณะของข้อมูลข้อความภาษาธรรมชาติที่ใช้ในการศึกษา

ประการแรก ผู้วิจัยได้เลือกใช้แนวคิดของการลดขนาดชุดข้อมูลทดสอบ ซึ่งเป็นการคัดกรองตัวอย่างข้อมูลโดยการตีความจากผู้เชียวชาญถึงถ้อยคำที่บ่งบอกถึงลักษณะของประเภทงาน เพื่อให้เหลือเฉพาะข้อมูลที่มีความสมบูรณ์และมีป้ายกำกับที่ชัดเจน โดยคำของแต่ละประเภทงานจะอ้างอิงตามตารางที่ 4-5

ตารางที่ 4-5 ถ้อยคำที่บ่งบอกถึงลักษณะของประเภทงาน

วัตถุประสงค์	คำที่เกี่ยวข้อง
งานส่งใบเสนอสินค้า	bom, dsc, equote, equotes, quote, revise spec, revise ให้, share link, solution id, solution link, แชร์ เรียบร้อย, แชร์ แล้ว, ปรับ สเปค, revise bom, ปรับ dsc, revise dsc, dsc ตาม แนบ, equotes ตาม แนบ, dsc option ปรับ, เพิ่ม dsc, revise dsc ตาม แนบ, quote ตาม แนบ, dsc revise, bom ตาม แนบ, bom จาก dsc, equote option, ปรับ bom, revised dsc, ส่ง dsc
งานบริการติดตั้ง	configure, implement, install sevice, migrate, scope, setup, sevice install, ค่า ติดตั้ง, ราคา ติดตั้ง

ตารางที่ 4-5 (ต่อ)

งานตรวจสอบเอกสารข้อกำหนด	confirm tor, tor เข้า, tor เช็ค, tor ข้อ ที่, tor ตก, tor ตรวจสอบ, tor ต้องการ , tor ติด, tor ที่ เป็น file, เช็ค tor, ใน tor, จาก tor, ตก spec, ตก tor, ตก สเปค, ตรวจสอบ tor, ตาม tor, ติด tor, ผ่าน tor, ผ่าน ข้อกำหนด, ส่วน tor, ของ tor, เนื่องจาก tor, เป็น ข้อกำหนด tor, สำหรับ ข้อกำหนด tor, tor server ที่ ส่ง มา, tor เนื่องจาก, กับ tor ของ, tor คือ
งานส่งเอกสารอ้างอิงข้อกำหนด	compile, compile datasheet, compile tor compile datasheet, compile เอกสาร compile, tor comply, comply tor, comply datasheet, data sheet, data sheet comply, datasheet, datasheet comply, เอกสาร compile, เอกสาร comply, เอกสาร comply datasheet, เอกสาร datasheet bom comply
งานร่างเอกสารข้อกำหนด	draft tor, revise draft tor, revise tor, tor revise, tor revised, tor update, tor แก้, tor แก้ไข, tor ตาม แนบ, tor ปรับ, update tor, เพิ่ม tor, แก้ tor, แก้ไข tor, ปรับ tor, tor เปลี่ยน, เปลี่ยน tor , draft แนบ, tor ตาม ด้านล่าง, tor ตาม ไฟล์ แนบ, tor ของ ตาม ไฟล์, ส่ง tor แก้, tor ดังนี้, tor ขอ ปรับ
งานตอบคำถามทั่วไป	* ตัวอย่างที่ไม่มีถ้อยคำตามที่สอดคล้องกับป้ายชื่อทั้งหมด

เมื่อได้ถ้อยคำแล้วจะนำมาสร้างเป็นชุดข้อมูลสำหรับทดสอบ โดยจับคู่ถ้อยคำกับป้ายชื่อทั้งหมด ซึ่งกระบวนการนี้ช่วยลดความซ้ำซ้อนของข้อมูลและลดความกำกวมของตัวอย่างที่อาจก่อให้เกิดความคลาดเคลื่อนในกระบวนการเรียนรู้ของแบบจำลอง การคัดเลือกข้อมูลที่มีคุณภาพยังช่วยให้แบบจำลองสามารถเรียนรู้ลักษณะสำคัญของแต่ละคลาสได้อย่างมีประสิทธิภาพมากขึ้น ภายใต้ข้อจำกัดของทรัพยากรในการประมวลผล ในการศึกษานี้ผู้วิจัยได้ใช้สถาปัตยกรรมของแบบจำลองสองทิศทางหน่วยความจำระยะยาวแบบสั้น (Bi-LSTM) ที่มีจำนวนหน่วยประมวลผล 64 หน่วย ซึ่งถือว่าเป็นขนาดที่เหมาะสมสำหรับการเรียนรู้รูปแบบของข้อมูลลำดับข้อความในขณะที่ยังสามารถควบคุมภาระในการคำนวณให้อยู่ในระดับที่เหมาะสมได้ เมื่อใช้ร่วมกับชุดข้อมูลที่ผ่านการขัดเกลาแล้ว แบบจำลองจึงสามารถสร้างขอบเขตการตัดสินใจในการจำแนกประเภทข้อมูลได้อย่างมีประสิทธิภาพมากขึ้น นอกจากการปรับปรุงคุณภาพของข้อมูลแล้ว ผู้วิจัยยังได้พัฒนาวิธีการสกัดคุณลักษณะของข้อความเพื่อให้แบบจำลองสามารถมุ่งเน้นการเรียนรู้เฉพาะส่วนที่มีความสำคัญต่อการจำแนก

เดิมทีการเตรียมข้อมูลสำหรับการฝึกแบบจำลองจะใช้ข้อความทั้งประโยคเป็นข้อมูลนำเข้า อย่างไรก็ตาม วิธีการดังกล่าวอาจทำให้แบบจำลองต้องประมวลผลคำจำนวนมากที่ไม่มีความสำคัญ

ต่อการตัดสินใจ ผู้วิจัยจึงได้ปรับปรุงขั้นตอนการเตรียมข้อมูลโดยนำแนวคิดของการค้นหาแบบไฮบริดมาใช้ในการสกัดวลีสำคัญโดยเทียบกับถ้อยคำที่มีการกำหนดไว้จากกระบวนการก่อน กระบวนการนี้ทำให้สามารถระบุส่วนของข้อความที่มีนัยสำคัญต่อการจำแนกประเภทได้อย่างมีประสิทธิภาพมากขึ้น โดยระบบสามารถตรวจจับกลุ่มคำสำคัญได้ทั้งในลักษณะที่เรียงต่อกันโดยตรงและในลักษณะที่มีคำอื่นแทรกอยู่ระหว่างคำสำคัญ ซึ่งช่วยให้สามารถครอบคลุมรูปแบบของภาษาในประโยคที่มีความหลากหลายได้ดียิ่งขึ้น ผลจากกระบวนการสกัดวลีสำคัญดังกล่าวทำให้ผู้วิจัยสามารถลดความยาวของลำดับข้อมูลนำเข้าเข้าสู่แบบจำลองได้อย่างมีนัยสำคัญ จากเดิมที่กำหนดความยาวสูงสุดของลำดับคำไว้ที่หนึ่งร้อยคำ สามารถปรับลดลงเหลือเพียงสองคำที่เป็นวลีสำคัญหลัก การลดความยาวของลำดับข้อมูลนำเข้าช่วยลดสัญญาณรบกวนที่เกิดจากคำแวดล้อมที่ไม่เกี่ยวข้อง และทำให้แบบจำลองสามารถมุ่งเน้นการเรียนรู้คุณลักษณะที่มีผลต่อการตัดสินใจได้โดยตรงมากขึ้น แนวทางดังกล่าวยังช่วยลดภาระในการคำนวณของแบบจำลอง และส่งผลให้กระบวนการฝึกแบบจำลองสามารถดำเนินการได้อย่างมีประสิทธิภาพมากยิ่งขึ้น

ในส่วนของการปรับปรุงสถาปัตยกรรมและกระบวนการเรียนรู้ของแบบจำลอง ผู้วิจัยได้ดำเนินการปรับแต่งพารามิเตอร์สำคัญของโครงข่ายประสาทเทียมให้สอดคล้องกับลักษณะของข้อมูลที่ผ่านการสกัดคุณลักษณะแล้ว การปรับปรุงดังกล่าวครอบคลุมทั้งการออกแบบชั้นแสดงผลของแบบจำลองและการกำหนดค่าพารามิเตอร์ที่ใช้ในการฝึกแบบจำลอง โดยผู้วิจัยได้เลือกใช้ฟังก์ชันกระตุ้นแบบ Sigmoid ในชั้นแสดงผลร่วมกับฟังก์ชันการสูญเสียแบบ Binary Crossentropy ซึ่งเป็นแนวทางที่เหมาะสมสำหรับการเรียนรู้ลักษณะของแต่ละคลาสในลักษณะที่เป็นอิสระต่อกัน การกำหนดค่าดังกล่าวช่วยให้แบบจำลองสามารถสร้างค่าความน่าจะเป็นของแต่ละคลาสได้อย่างชัดเจน และทำให้ค่าความเชื่อมั่นของคลาสที่มีความเป็นไปได้สูงสุดสามารถสะท้อนลักษณะของข้อมูลได้อย่างแม่นยำมากยิ่งขึ้น นอกจากนี้ ผู้วิจัยยังได้ดำเนินการปรับจูนพารามิเตอร์ของกระบวนการฝึกแบบจำลองเพื่อให้สามารถเรียนรู้รูปแบบของข้อมูลได้อย่างเต็มประสิทธิภาพ โดยมีการเพิ่มจำนวนรอบของการฝึกจากเดิม 10 เป็น 100 รอบ เพื่อให้แบบจำลองมีโอกาสปรับปรุงค่าน้ำหนักภายในโครงข่ายประสาทเทียมได้มากขึ้น กระบวนการดังกล่าวช่วยให้แบบจำลองสามารถลู่เข้าสู่ค่าที่ทำให้ความผิดพลาดของการทำนายลดลงอย่างต่อเนื่อง นอกจากนี้ยังมีการปรับลดขนาดของกลุ่มข้อมูลที่ใช้ในการฝึกในแต่ละรอบจาก 32 ตัวอย่างเหลือเพียง 4 ตัวอย่าง การปรับลดขนาดของกลุ่มข้อมูลทำให้เกิดการปรับปรุงค่าน้ำหนักของแบบจำลองบ่อยครั้งมากขึ้น ซึ่งส่งผลให้แบบจำลองสามารถเรียนรู้รูปแบบของวลีสำคัญที่มีลักษณะเฉพาะได้อย่างละเอียดและแม่นยำมากขึ้นจากการดำเนินการปรับปรุงทั้งในด้านคุณภาพของข้อมูล กระบวนการสกัดคุณลักษณะ และการกำหนดค่าพารามิเตอร์ของแบบจำลอง ส่งผลให้ประสิทธิภาพของแบบจำลองสองทิศทางหน่วยความจำระยะยาวแบบสั้น (Bi-LSTM) เพิ่มขึ้นอย่างมีนัยสำคัญเมื่อเปรียบเทียบกับการตั้งค่าแบบเริ่มต้น แนวทางดังกล่าวช่วยให้

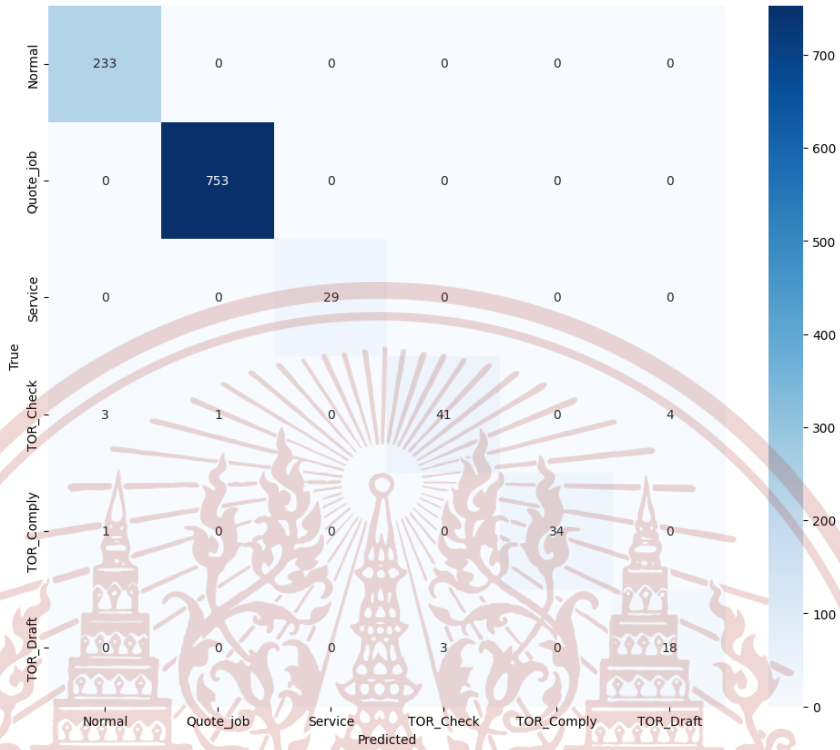
แบบจำลองสามารถมุ่งเน้นการเรียนรู้เฉพาะลักษณะสำคัญของข้อความที่เกี่ยวข้องกับการจำแนกประเภทได้อย่างมีประสิทธิภาพมากขึ้น ทั้งยังช่วยลดผลกระทบจากข้อมูลรบกวนและความซ้ำซ้อนของข้อมูลที่อาจทำให้กระบวนการเรียนรู้มีความคลาดเคลื่อน

ผลลัพธ์ที่ได้สะท้อนให้เห็นถึงความสำคัญของการออกแบบกระบวนการเตรียมข้อมูลและการปรับแต่งสถาปัตยกรรมของแบบจำลองให้สอดคล้องกับลักษณะของปัญหาที่ศึกษา ซึ่งเป็นปัจจัยสำคัญที่ช่วยให้แบบจำลองการเรียนรู้เชิงลึกสามารถแสดงศักยภาพในการประมวลผลภาษาธรรมชาติและการจำแนกข้อความได้อย่างมีประสิทธิภาพในบริบทของงานวิจัยนี้ โดยการประเมินผลแบบจำลองการเรียนรู้ของเครื่องฉบับปรับปรุงเมื่อทดสอบกับตัวอย่างจำนวน 1,120 อีเมล พบว่าค่าความแม่นยำ (Accuracy) อยู่ที่ร้อยละ 99 ในกรณีวิเคราะห์แยกแต่ละป้ายชื่อสามารถสรุปได้ตามตารางที่ 4-6

ตารางที่ 4-6 ผลประเมินแบบจำลองแบบหน่วยความจำระยะยาวแบบสั้น (LSTM) ปรับปรุง

วัตถุประสงค์	Precision	Recall	F1-score	Sample
งานตอบคำถามทั่วไป	0.98	1	0.99	233
งานส่งใบเสนอสินค้า	1	1	1	753
งานบริการติดตั้ง	1	1	1	29
งานตรวจสอบเอกสารข้อกำหนด	0.93	0.84	0.88	49
งานส่งเอกสารอ้างอิงข้อกำหนด	1	0.97	0.99	35
งานร่างเอกสารข้อกำหนด	0.82	0.86	0.84	21
ค่าเฉลี่ย	0.96	0.94	0.95	

สำหรับการวิเคราะห์ Confusion Matrix ของ LSTM แบบปรับปรุง แสดงตามภาพที่ 4-4 โดยจะอธิบายในรูปแบบความเข้มของสี ซึ่งจะบงบอกถึงจำนวนตัวอย่างที่ถูกทำนายซึ่งในแนวตั้งคือป้ายชื่อจริงของตัวอย่างและในแนวนอนคือป้ายชื่อที่ได้จากการทำนายของแบบจำลอง



ภาพที่ 4-4 การเปรียบเทียบ Confusion Matrix ของเทคนิค LSTM ปรับปรุง

ผลจากการเพิ่มประสิทธิภาพแบบจำลอง (Model Optimization) แสดงให้เห็นถึงการพัฒนาในเชิงบวกอย่างมีนัยสำคัญ โดยค่าความแม่นยำโดยรวมเพิ่มขึ้นจากร้อยละ 96 เป็นร้อยละ 99 ประสิทธิภาพที่เพิ่มขึ้นอย่างโดดเด่นปรากฏชัดในกลุ่มคลาสที่มีความซับซ้อนสูง เช่น กลุ่มงานเอกสาร ซึ่งมีค่าความถูกต้อง (Precision) เพิ่มขึ้นอย่างมาก เนื่องจากการบูรณาการเทคนิค Hybrid Search และการปรับปรุงพารามิเตอร์การฝึกฝนที่ช่วยให้แบบจำลองสามารถระบุคุณลักษณะเด่นของข้อความได้อย่างแม่นยำและลดอิทธิพลของข้อมูลรบกวนแวดล้อมได้อย่างมีประสิทธิภาพ

4.3 การปรับปรุงแบบจำลองการเรียนรู้ของเครื่องสำหรับทำนาย 2 วัตถุประสงค์

เพื่อให้แบบจำลองสามารถรองรับลักษณะของข้อมูลข้อความที่มีสถานการณ์ซ้อนทับกันมากกว่าหนึ่งประเภทภายในประโยคเดียว ผู้วิจัยได้ดำเนินการพัฒนาต่อยอดสถาปัตยกรรมของแบบจำลองสองทิศทางหน่วยความจำระยะยาวแบบสั้น (Bi-LSTM) ให้สามารถรองรับการพยากรณ์ผลลัพธ์ในลักษณะสองฉลากหรือแบบคู่ (Dual-label Prediction) แนวคิดดังกล่าวเกิดขึ้นจากข้อสังเกตที่พบในชุดข้อมูลว่า ข้อความจำนวนหนึ่งอาจมีวัตถุประสงค์หรือบริบทมากกว่าหนึ่งประเภทปรากฏอยู่ภายในรูปประโยคเดียว ซึ่งหากใช้กระบวนการจำแนกแบบคลาสเดียวเพียงคลาสเดียว อาจไม่สามารถสะท้อนความหมายเชิงบริบทของข้อความได้อย่างครบถ้วน ดังนั้น ผู้วิจัยจึงออกแบบกลไกการประมวลผลที่สามารถตรวจจับคำสำคัญหลายรายการภายในข้อความ และ

จัดลำดับความสำคัญของคำเหล่านั้นตามตำแหน่งที่ปรากฏในประโยค เพื่อนำมาใช้เป็นพื้นฐานในการพยากรณ์คลาสหลักและคลาสรองของข้อความอย่างเป็นระบบ กลไกสำคัญที่ใช้ในกระบวนการดังกล่าวคือการปรับปรุงวิธีการค้นหาคำสำคัญในขั้นตอนเตรียมข้อมูลให้สามารถตรวจพบคำหรือวลีที่มีความสำคัญได้มากกว่าหนึ่งรายการภายในข้อความเดียว โดยอาศัยแนวคิดของการค้นหาแบบไฮบริดที่ถูกพัฒนาให้รองรับการตรวจจับหลายตำแหน่งพร้อมกัน กระบวนการนี้เริ่มต้นจากการสแกนข้อความทั้งหมดเพื่อตรวจหาวลีสำคัญที่ตรงกับคลังข้อมูลคำสำคัญที่กำหนดไว้ล่วงหน้า เมื่อระบบตรวจพบวลีที่ตรงกับเงื่อนไข ระบบจะบันทึกตำแหน่งเริ่มต้นของคำหรือวลีที่ตรวจพบในรูปของดัชนีตำแหน่งภายในประโยค ซึ่งทำให้สามารถระบุได้ว่าคำสำคัญแต่ละรายการปรากฏขึ้นในตำแหน่งใดของข้อความ กระบวนการบันทึกตำแหน่งดังกล่าวมีบทบาทสำคัญต่อการจัดลำดับความสำคัญของคำสำคัญที่ถูกตรวจพบ เนื่องจากตำแหน่งของคำในประโยคมักสะท้อนลำดับของเหตุการณ์หรือความสำคัญของข้อมูลในบริบทของภาษา เมื่อได้ข้อมูลตำแหน่งของคำสำคัญทั้งหมดแล้ว ขั้นตอนถัดมาคือการจัดเรียงลำดับของคำสำคัญตามตำแหน่งการปรากฏในประโยค กระบวนการนี้ช่วยให้ระบบสามารถแยกแยะได้ว่าคำสำคัญใดเป็นข้อมูลที่ปรากฏก่อนและคำสำคัญใดปรากฏในลำดับถัดมา จากหลักการดังกล่าว ผู้วิจัยได้กำหนดนิยามของคลาสหลักให้หมายถึงวลีสำคัญที่ปรากฏเป็นลำดับแรกภายในประโยค ในขณะที่คลาสรองหมายถึงวลีสำคัญที่ปรากฏในลำดับถัดมา การกำหนดโครงสร้างในลักษณะนี้ช่วยให้แบบจำลองสามารถสะท้อนลำดับความสำคัญของข้อมูลที่ปรากฏในข้อความได้อย่างสอดคล้องกับลักษณะของภาษาในบริบทจริง อีกทั้งยังช่วยให้กระบวนการวิเคราะห์ข้อความสามารถรองรับสถานการณ์ที่มีเหตุการณ์หรือวัตถุประสงค์หลายประเภทเกิดขึ้นภายในข้อความเดียวได้อย่างมีประสิทธิภาพ

ในขั้นตอนของการพยากรณ์ผลลัพธ์ ผู้วิจัยได้ออกแบบตรรกะการทำงานของระบบให้ดำเนินการแบบเป็นลำดับขั้นตอน ซึ่งแตกต่างจากกระบวนการจำแนกแบบดั้งเดิมที่มักเลือกผลลัพธ์เพียงคลาสเดียวที่มีค่าความน่าจะเป็นสูงที่สุด ระบบที่พัฒนาขึ้นจะทำงานในลักษณะของกระบวนการประมวลผลแบบลำดับขั้น โดยเริ่มจากการนำวลีสำคัญลำดับแรกที่ถูกสกัดออกจากข้อความเข้าสู่แบบจำลอง Bi-LSTM เพื่อทำการพยากรณ์คลาสหลักของข้อความ กระบวนการนี้ช่วยให้ระบบสามารถยืนยันประเภทของสถานการณ์หลักที่ปรากฏในข้อความได้อย่างแม่นยำ หลังจากนั้นระบบจะตรวจสอบว่ามีวลีสำคัญลำดับถัดมาปรากฏอยู่ในข้อความหรือไม่ หากพบวลีสำคัญเพิ่มเติม ระบบจะนำวลีดังกล่าวเข้าสู่แบบจำลอง Bi-LSTM อีกครั้งหนึ่ง เพื่อทำการพยากรณ์คลาสรองของข้อความ กระบวนการดังกล่าวช่วยให้ระบบสามารถวิเคราะห์ข้อมูลที่มีความซับซ้อนเชิงบริบทได้ดียิ่งขึ้น เนื่องจากสามารถแยกการประเมินของแต่ละสถานการณ์ออกจากกันอย่างเป็นระบบ ในกรณีที่ระบบตรวจพบเพียงวลีสำคัญเพียงรายการเดียวภายในข้อความ ระบบจะกำหนดค่าของคลาสรองให้เป็นค่าว่างโดยอัตโนมัติ การกำหนดกลไกดังกล่าวช่วยลดความเสี่ยงของการทำนายคลาสรองโดยไม่มี

หลักฐานข้อมูลรองรับ ซึ่งอาจทำให้เกิดข้อผิดพลาดในการตีความผลลัพธ์ของแบบจำลองได้ แนวทางนี้จึงช่วยรักษาความน่าเชื่อถือของระบบในการพยากรณ์ผลลัพธ์ และทำให้การแปลผลลัพธ์ของแบบจำลองมีความชัดเจนและสอดคล้องกับลักษณะของข้อมูลจริงมากยิ่งขึ้น ในด้านสถาปัตยกรรมของแบบจำลอง ผู้วิจัยยังคงใช้โครงสร้างหลักของ Bi-LSTM ที่มีจำนวนหน่วยประมวลผล 64 หน่วย ซึ่งเป็นโครงสร้างที่เหมาะสมต่อการเรียนรู้ข้อมูลลำดับของข้อความ เนื่องจากสามารถประมวลผลบริบทของคำได้ทั้งในทิศทางไปข้างหน้าและย้อนกลับ ส่งผลให้แบบจำลองสามารถจับความสัมพันธ์ของคำภายในประโยคได้อย่างครอบคลุมมากขึ้น ในชั้นแสดงผลของแบบจำลอง ผู้วิจัยเลือกใช้ฟังก์ชันกระตุ้นแบบ Sigmoid ร่วมกับฟังก์ชันการคำนวณความสูญเสียแบบ Binary Cross Entropy ซึ่งเป็นแนวทางที่เหมาะสมสำหรับการเรียนรู้ข้อมูลในลักษณะหลายฉลาก เนื่องจากสามารถคำนวณค่าความน่าจะเป็นของแต่ละคลาสได้อย่างเป็นอิสระต่อกัน นอกจากนี้ยังมีการใช้เทคนิคการแปลงข้อมูลด้วยไลบรารี่ “MultiLabelBinarizer” เพื่อเตรียมข้อมูลป้ายกำกับให้เหมาะสมกับการเรียนรู้แบบหลายฉลากของแบบจำลอง สำหรับการประเมินประสิทธิภาพของแบบจำลอง ผู้วิจัยได้กำหนดตัวชี้วัดการประเมินผลในหลายมิติ เพื่อให้สามารถวิเคราะห์ความสามารถของระบบได้อย่างรอบด้าน ตัวชี้วัดแรกใช้สำหรับวัดความถูกต้องของการพยากรณ์คลาสหลัก ตัวชี้วัดที่สองใช้สำหรับประเมินความถูกต้องของการพยากรณ์คลาสรอง ซึ่งช่วยแสดงให้เห็นถึงศักยภาพของระบบในการตรวจจับสถานการณ์รองที่อาจปรากฏร่วมกับสถานการณ์หลักในข้อความเดียว นอกจากนี้ ผู้วิจัยยังได้กำหนดตัวชี้วัดแบบผสมที่ใช้สำหรับประเมินความถูกต้องของการพยากรณ์ในกรณีที่แบบจำลองต้องระบุคลาสทั้งสองรายการได้อย่างถูกต้องพร้อมกัน ตัวชี้วัดดังกล่าวมีความสำคัญอย่างยิ่งต่อการประเมินประสิทธิภาพของระบบในสถานการณ์ที่ข้อความมีความซับซ้อนและมีหลายบริบทปรากฏร่วมกัน ผลการทดลองจากกระบวนการประเมินประสิทธิภาพดังกล่าวได้ถูกนำเสนอในตารางที่ 4-7

ตารางที่ 4-7 ผลประเมินการทดสอบตัวอย่าง แบบจำลองสำหรับทำนาย 2 ป้าย

ประเภทความแม่นยำ (Metric)	ค่าความแม่นยำ (Accuracy)	จำนวนตัวอย่าง (Samples)
วัตถุประสงค์หลักถูกต้อง	ร้อยละ 100	217
วัตถุประสงค์รองถูกต้อง	ร้อยละ 96.77	
ถูกต้องทั้งสองวัตถุประสงค์	ร้อยละ 96.77	

ซึ่งแสดงให้เห็นถึงความสามารถของแบบจำลอง Bi-LSTM ที่ได้รับการพัฒนาให้รองรับการพยากรณ์แบบสองฉลากในการวิเคราะห์ข้อความที่มีหลายบริบทภายในประโยคเดียว ผลลัพธ์ที่ได้สะท้อนให้เห็นว่าแนวทางการออกแบบกลไกการตรวจจับคำสำคัญ การจัดลำดับตำแหน่งของข้อมูล และการพยากรณ์แบบลำดับขั้น สามารถช่วยเพิ่มความยืดหยุ่นและความแม่นยำของระบบใน

การจำแนกประเภทข้อความได้อย่างมีประสิทธิภาพมากยิ่งขึ้น ทั้งยังแสดงให้เห็นถึงศักยภาพของการประยุกต์ใช้เทคนิคการเรียนรู้เชิงลึกในการประมวลผลภาษาธรรมชาติที่ต้องเผชิญกับข้อมูลที่มีความซับซ้อนทางบริบทในงานวิจัยนี้

4.4 การสร้างแอปพลิเคชั่นจำแนกวัตถุประสงค์อีเมล

ในการนำแบบจำลองที่พัฒนาขึ้นไปประยุกต์ใช้ในสภาพแวดล้อมการทำงานจริง ผู้วิจัยได้พัฒนาแนวทางการสร้างแอปพลิเคชันต้นแบบสำหรับการประมวลผลข้อมูลอีเมลที่ดึงมาจากระบบขององค์กร โดยข้อมูลดังกล่าวอยู่ในรูปแบบไฟล์ JSON ที่ได้จากบริการของ Microsoft Graph API ซึ่งเป็นส่วนหนึ่งของแพลตฟอร์มของ Microsoft สำหรับการเข้าถึงข้อมูลจากระบบอีเมลและบริการอื่น ๆ ภายในองค์กร แนวคิดหลักของการพัฒนาแอปพลิเคชันนี้คือการนำแบบจำลองการเรียนรู้เชิงลึกที่ผ่านการปรับปรุงแล้วมาประยุกต์ใช้เพื่อช่วยวิเคราะห์และจำแนกประเภทของข้อความอีเมลโดยอัตโนมัติ โดยแบบจำลองที่นำมาใช้คือ Long Short-Term Memory (LSTM) ซึ่งได้รับการออกแบบเพิ่มเติมให้สามารถรองรับการทำนายได้มากกว่าหนึ่งคลาสภายในข้อความเดียว หรือที่เรียกว่าการทำนายแบบสองคลาส (Dual Class Prediction) แนวทางดังกล่าวมีความสำคัญต่อการประมวลผลข้อมูลจริง เนื่องจากข้อความอีเมลในบริบทของการทำงานมักมีเนื้อหาที่เกี่ยวข้องกับหลายสถานการณ์พร้อมกัน การพัฒนาแอปพลิเคชันจึงมีเป้าหมายเพื่อสร้างเครื่องมือที่สามารถนำผลลัพธ์ของงานวิจัยไปใช้งานได้จริงและช่วยเพิ่มประสิทธิภาพในการวิเคราะห์ข้อมูลข้อความขององค์กร ในด้านสถาปัตยกรรมของระบบ แอปพลิเคชันถูกพัฒนาขึ้นบนพื้นฐานของเทคโนโลยีการเรียนรู้เชิงลึกและการประมวลผลภาษาธรรมชาติ โดยใช้ภาษา Python เป็นภาษาหลักในการพัฒนา เนื่องจากเป็นภาษาที่มีเครื่องมือสนับสนุนด้านวิทยาการข้อมูลและปัญญาประดิษฐ์จำนวนมาก เพื่อให้ระบบสามารถทำงานได้อย่างครบถ้วน ผู้วิจัยได้ติดตั้งไลบรารีที่เกี่ยวข้องกับกระบวนการประมวลผลข้อมูลหลายรายการ โดยไลบรารีที่มีบทบาทสำคัญต่อการทำงานของระบบประกอบด้วย TensorFlow ซึ่งทำหน้าที่เป็นเฟรมเวิร์กหลักสำหรับการโหลดแบบจำลอง LSTM ที่ผ่านการฝึกแล้วและใช้ในการประมวลผลการทำนายผลลัพธ์ของข้อมูลใหม่ในขั้นตอน Inference เฟรมเวิร์กดังกล่าวมีความสามารถในการประมวลผลแบบโครงข่ายประสาทเทียมเชิงลึกได้อย่างมีประสิทธิภาพและสามารถรองรับการคำนวณเชิงเวกเตอร์ที่มีความซับซ้อนได้ดี

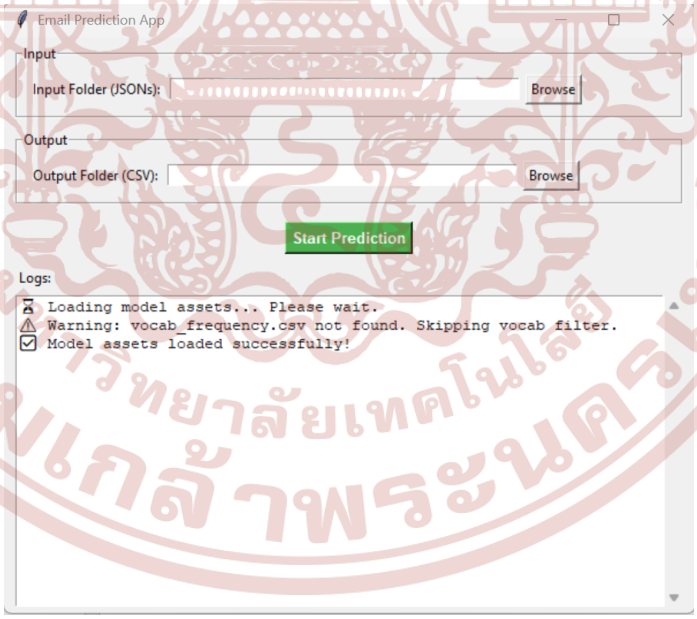
ในส่วนของการประมวลผลภาษาธรรมชาติ ผู้วิจัยได้ใช้ไลบรารี “PyThaiNLP” เพื่อทำหน้าที่ตัดคำภาษาไทยและเตรียมข้อมูลข้อความให้อยู่ในรูปแบบที่เหมาะสมสำหรับการนำเข้าสู่แบบจำลอง เนื่องจากภาษาไทยมีลักษณะเฉพาะที่แตกต่างจากภาษาอื่นตรงที่ไม่มีตัวคั่นระหว่างคำ การตัดคำจึงเป็นขั้นตอนสำคัญในการเตรียมข้อมูลก่อนการวิเคราะห์ นอกจากนี้ยังมีการใช้ไลบรารี “Gensim”

เพื่อสนับสนุนการจัดการข้อมูลคำและเวกเตอร์คำ ซึ่งมักถูกนำมาใช้ร่วมกับกระบวนการสร้างตัวแทนเชิงตัวเลขของคำในรูปแบบการฝังคำ การใช้เทคนิคดังกล่าวช่วยให้แบบจำลองสามารถเข้าใจความสัมพันธ์เชิงความหมายของคำภายในข้อความได้ดีขึ้น สำหรับการจัดการข้อมูลและการคำนวณเชิงตัวเลข ผู้วิจัยได้ใช้ไลบรารี “Pandas” เพื่อช่วยจัดการข้อมูลในรูปแบบตารางและบันทึกผลลัพธ์ของการทำนายออกมาในรูปแบบไฟล์ซีเอสวี ซึ่งสามารถนำไปวิเคราะห์ต่อในโปรแกรมตารางคำนวณได้อย่างสะดวก นอกจากนี้ยังมีการใช้ไลบรารี “NumPy” เพื่อสนับสนุนการคำนวณเชิงคณิตศาสตร์และการจัดการอาร์เรย์ของข้อมูลที่ใช้ในการประมวลผลของแบบจำลอง ส่วนการจัดการข้อมูลป้ายกำกับและการประเมินประสิทธิภาพของแบบจำลอง ผู้วิจัยได้ใช้ไลบรารี “Scikit-learn” ซึ่งมีเครื่องมือสำคัญสำหรับการทำเข้ารหัสป้ายชื่อ (Label Encoding) และการคำนวณตัวชี้วัดประสิทธิภาพของแบบจำลองในงานด้านการจำแนกข้อมูล เมื่อพัฒนาโค้ดของระบบเสร็จสมบูรณ์ ผู้วิจัยได้ดำเนินการสร้างแอปพลิเคชันที่สามารถใช้งานได้จริงโดยไม่ต้องติดตั้งสภาพแวดล้อมการพัฒนาเพิ่มเติมบนเครื่องของผู้ใช้งาน กระบวนการดังกล่าวดำเนินการโดยใช้เครื่องมือ “PyInstaller” ซึ่งเป็นเครื่องมือสำหรับรวมซอร์สโค้ดภาษา Python และไลบรารีทั้งหมดให้เป็นไฟล์ปฏิบัติการเพียงไฟล์เดียวที่สามารถรันบนระบบปฏิบัติการวินโดว์ได้โดยตรง กระบวนการสร้างไฟล์ปฏิบัติการถูกควบคุมผ่านสคริปต์อัตโนมัติที่กำหนดขั้นตอนการทำงานไว้อย่างชัดเจน เริ่มจากการตรวจสอบสภาพแวดล้อมของระบบเพื่อให้มั่นใจว่ามีการติดตั้งไลบรารีทั้งหมดที่จำเป็นสำหรับการทำงานของโปรแกรม จากนั้นจึงดำเนินการคอมไพล์ซอร์สโค้ดของแอปพลิเคชันโดยใช้คำสั่งของ “PyInstaller” เพื่อแปลงไฟล์ไพทอน (Python) ให้เป็นไฟล์ปฏิบัติการที่สามารถใช้งานได้โดยตรง ในการกำหนดค่าของกระบวนการคอมไพล์ ผู้วิจัยได้ตั้งค่าให้โปรแกรมรวมไฟล์ทั้งหมดไว้ในไฟล์ปฏิบัติการเพียงไฟล์เดียว ซึ่งช่วยให้การแจกจ่ายโปรแกรมไปยังผู้ใช้งานเป็นไปได้อย่างสะดวกและลดความยุ่งยากในการติดตั้งระบบ นอกจากนี้ยังมีการกำหนดค่าให้ซ่อนหน้าต่างคำสั่งเมื่อผู้ใช้งานเปิดโปรแกรม เพื่อให้การทำงานของแอปพลิเคชันมีลักษณะเหมือนซอฟต์แวร์ทั่วไปที่มีส่วนติดต่อผู้ใช้แบบกราฟิก หลังจากกระบวนการคอมไพล์เสร็จสมบูรณ์ ระบบจะสร้างไฟล์ปฏิบัติการของโปรแกรมในโฟลเดอร์ผลลัพธ์ โฟลเดอร์ดังกล่าวจะประกอบด้วยไฟล์แบบจำลองที่ถูกบันทึกในรูปแบบไฟล์ “.h5” รวมถึงไฟล์ข้อมูลเสริมที่ใช้ในการเตรียมข้อมูลก่อนการทำนาย เมื่อโครงสร้างของโฟลเดอร์ถูกจัดวางอย่างถูกต้อง แอปพลิเคชันจะสามารถโหลดแบบจำลองเข้าสู่หน่วยความจำและเริ่มกระบวนการวิเคราะห์ข้อมูลอีเมลได้ทันที หลังจากผ่านกระบวนการสร้างแอปพลิเคชั่นสำเร็จ จะได้ไฟล์ที่สามารถคลิกเพื่อเริ่มต้นใช้งานได้ทันที ลักษณะไฟล์แสดงตามภาพที่ 4-5



ภาพที่ 4-5 สัญลักษณ์ Icon ของ Application

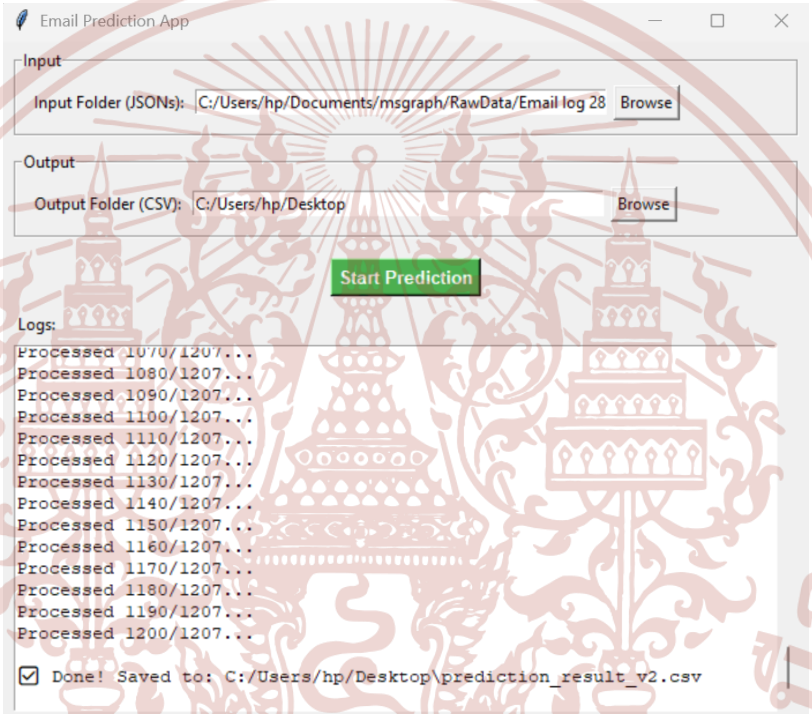
เมื่อเริ่มต้นการทำงานของแอปพลิเคชัน ระบบจะแสดงหน้าต่างโปรแกรมหลักขึ้นมาสำหรับให้ผู้ใช้งานสามารถโต้ตอบกับระบบได้ โดยลักษณะของส่วนติดต่อผู้ใช้ของโปรแกรมจะแสดงดังที่ปรากฏในภาพที่ 4-6 ซึ่งเป็นหน้าต่างควบคุมหลักของระบบสำหรับกำหนดค่าการประมวลผลข้อมูล ในหน้าต่างดังกล่าวผู้ใช้งานสามารถกำหนดตำแหน่งของแหล่งข้อมูลที่ต้องการนำเข้าสู่กระบวนการวิเคราะห์ได้ผ่านปุ่มเลือกโฟลเดอร์ที่จัดเตรียมไว้ในโปรแกรม โดยโฟลเดอร์ดังกล่าวทำหน้าที่เป็นตำแหน่งจัดเก็บข้อมูลนำเข้า (Data Input) ซึ่งประกอบด้วยไฟล์ข้อมูลอีเมลในรูปแบบ JSON ที่ได้จากกระบวนการดึงข้อมูลผ่านระบบบริการขององค์กร ข้อมูลเหล่านี้จะถูกนำเข้าสู่กระบวนการเตรียมข้อมูลและการประมวลผลด้วยแบบจำลองการเรียนรู้ของเครื่องเพื่อดำเนินการพยากรณ์วัตถุประสงค์ของอีเมลตามที่แบบจำลองได้เรียนรู้ไว้ก่อนหน้า



ภาพที่ 4-6 หน้าต่างโปรแกรมพยากรณ์วัตถุประสงค์ของ Email

เมื่อผู้ใช้งานได้กำหนดตำแหน่งของโฟลเดอร์สำหรับข้อมูลนำเข้าและโฟลเดอร์สำหรับจัดเก็บผลลัพธ์เรียบร้อยแล้ว ขั้นตอนถัดไปคือการเริ่มต้นกระบวนการประมวลผลข้อมูล โดยผู้ใช้งานสามารถดำเนินการได้ผ่านปุ่มคำสั่งที่ระบุว่า “Start Prediction” ซึ่งปรากฏอยู่ภายในหน้าต่างควบคุมหลักของแอปพลิเคชัน ปุ่มดังกล่าวทำหน้าที่เป็นตัวเริ่มต้นกระบวนการทำงานของระบบทั้งหมด ตั้งแต่ขั้นตอน

การอ่านข้อมูลอีเมลจากโฟลเดอร์ที่กำหนด การเตรียมข้อมูลข้อความสำหรับการวิเคราะห์ ตลอดจนการนำข้อมูลเข้าสู่แบบจำลองการเรียนรู้ของเครื่องเพื่อทำการพยากรณ์วัตถุประสงค์ของอีเมล เมื่อระบบดำเนินการประมวลผลข้อมูลทั้งหมดเสร็จสมบูรณ์ โปรแกรมจะแสดงข้อความแจ้งเตือนผลการทำงานภายในกรอบ “Logs” เพื่อยืนยันว่ากระบวนการพยากรณ์ได้เสร็จสิ้นลงอย่างสมบูรณ์ โดยตัวอย่างลักษณะของหน้าต่างโปรแกรมในระหว่างทำงานของระบบแสดงให้เห็นได้ตามภาพที่ 4-7



ภาพที่ 4-7 การทำงานของโปรแกรมพยากรณ์วัตถุประสงค์ของอีเมล์

เมื่อโปรแกรมทำงานเสร็จเรียบเรียบ จะได้ไฟล์ CSV ที่เก็บผลการทำงาน สามารถเข้าไปค้นหาในตำแหน่งที่ระบุไว้และสามารถใช้โปรแกรม Microsoft Excel ในการเปิดดูผลพยากรณ์ แสดงตามภาพที่ 4-8

Filename	clean_email_body_initial	Predicted_Class
AAMkAGQ3ODlmYmNiLTE5ODAtNDBjNy1iZTI0LTAzMmQ5ZmM0MjU1Y wBGAAAAAAAqsC8a07RIRKAJ60yHzag-BwCqqMHiOQ-rQpZu2XGAe5EtAAAAAAEJAACqqMHiOQ-rQpZu2XGAe5EtAACf1am-AAA=.json	Dear P'Kloyสเปค Dell Pro Tower QCT1250 ทั้ง 2 ใน เหมือนกันครับ	Normal
AAMkAGQ3ODlmYmNiLTE5ODAtNDBjNy1iZTI0LTAzMmQ5ZmM0MjU1Y wBGAAAAAAAqsC8a07RIRKAJ60yHzag-BwCqqMHiOQ-rQpZu2XGAe5EtAAAAAAEJAACqqMHiOQ-rQpZu2XGAe5EtAACf1am2AAA=.json	Dear P'KloyQuote #3700021570017.1 อันนี้ไม่สามารถทำใน primire ได้ครับเนื่องจากไม่มี no os ให้ Quote คับ	Quote_job
AAMkAGQ3ODlmYmNiLTE5ODAtNDBjNy1iZTI0LTAzMmQ5ZmM0MjU1Y wBGAAAAAAAqsC8a07RIRKAJ60yHzag-BwCqqMHiOQ-rQpZu2XGAe5EtAAAAAAEJAACqqMHiOQ-rQpZu2XGAe5EtAACf1am9AAA=.json	Dear P'Bumจากการดู TOR ที่ส่งมาครับลำดับที่ 1. จะเขย่งที่ข้อ 1.1.4 ครับ (เสนอรุ่นที่เขย่งเป็น Dell Pro Slim Plus)ลำดับที่ 2. ถ้าเสนอ Dell Pro Slim จะติดที่ข้อ 1.2.4 และถ้าจะเสนอรุ่นที่สูงกว่านี้ไม่ได้ครับ (Dell Pro Slim Plus) จะติดที่ข้อ 1.2.1 โดยเป็น TOR ของ CPU intel รุ่นเก่าครับ (i7-14700) ซึ่งถ้าไปรุ่นที่สูงกว่านี้จะไม่มี CPU ตัวนี้ให้ Config ครับลำดับที่ 3. สามารถเสนอ Dell Pro 14 ได้ครับ	TOR_Check

ภาพที่ 4-8 ตัวอย่าง Output ผลพยากรณ์แสดงผ่านโปรแกรม Microsoft Excel

หมายเหตุ : สำหรับตัวอย่างที่มีการสื่อถึง 2 ป้ายชื่อ ใน Output จะสำเนาตัวอย่างใส่ในแถวถัดไปแล้วใส่ผลพยากรณ์ลำดับที่ 2 (Second Class) ในช่องผลพยากรณ์

บทที่ 5
สรุป อภิปรายผล และข้อเสนอแนะ

งานค้นคว้าอิสระเรื่องการจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่องมีวัตถุประสงค์เพื่อพัฒนาแบบจำลองที่มีประสิทธิภาพในการจำแนกประเภทงานจากอีเมล และสร้างเครื่องมือที่ใช้งานได้จริงในรูปแบบแอปพลิเคชัน โดยสามารถแบ่งเป็นหัวข้อได้ดังนี้

5.1 สรุป

5.2 อภิปรายผล

5.3 ข้อเสนอแนะ

5.1 สรุป

5.1.1 ประสิทธิภาพของแบบจำลองการเรียนรู้ของเครื่อง (Machine Learning Model)

จากการเปรียบเทียบทั้ง 3 เทคนิคการเรียนรู้ของเครื่องพบว่า แบบจำลองหน่วยความจำระยะยาวแบบสั้น (LSTM) ให้ประสิทธิภาพสูงสุดในการจำแนกวัตถุประสงค์อีเมล โดยมีความแม่นยำ (Accuracy) สูงถึงร้อยละ 96 และเพิ่มเป็นร้อยละ 99 หลังจากผ่านการทำปรับปรุงแบบจำลอง (Model Optimization) และเมื่อเปรียบเทียบกับเทคนิคอื่นพบว่า ตัวจำแนกแบบเพื่อนบ้านใกล้ (KNN) ได้ร้อยละ 90 และตัวจำแนกแบบเบย์สแบบง่าย (Naive Bayes) ได้ร้อยละ 81 ซึ่งมีประสิทธิภาพลดหลั่นกันลงมาตามลำดับ จากผลลัพท์แสดงให้เห็นว่าการจำแนกวัตถุประสงค์อีเมลด้วยอัลกอลิทึมที่เป็นลักษณะการเรียนรู้แบบเชิงลึก (Deep Learning) จะได้เปรียบกว่าอัลกอลิทึมการเรียนรู้แบบมีผู้สอน (Supervisor Learning) ในกรณีที่ข้อมูลเป็นข้อความดิบ (Raw Text) หรือเป็นประโยคที่ไม่มีรูปแบบตายตัวอย่างเช่น บทสนทนาในอีเมลเป็นต้น

5.1.2 ประสิทธิภาพของการจำแนกแบบหลายป้าย (Multi-label Classification)

งานวิจัยประสบความสำเร็จในการพัฒนาแบบจำลองให้รองรับการจำแนกแบบหลายป้ายเพื่อแก้ปัญหาอีเมลหนึ่งฉบับที่มีหลายวัตถุประสงค์ เช่น เป็นทั้งการสอบถามและส่งเอกสารข้อกำหนดในเวลาเดียวกัน โดยผลการทดสอบพบว่าการใช้เทคนิคไบนารีความเกี่ยวข้อง (Binary Relevance) ร่วมกับหน่วยความจำระยะยาวแบบสั้น (LSTM) ให้ค่าสูญเสียแฮมมิง (Hamming Loss) ที่ต่ำและมีอัตราส่วนการจับคู่แบบตรงทั้งหมดอยู่ในเกณฑ์ดีที่ร้อยละ 96.77 ซึ่งสะท้อนว่าแบบจำลองสามารถระบุวัตถุประสงค์ที่ซ้อนทับกันได้อย่างถูกต้องแม่นยำกว่าการจำแนกแบบป้ายเดียว (Single-label) ในกรณีข้อมูลที่มีความซับซ้อน

5.2 อภิปรายผล

จากผลการศึกษา มีประเด็นที่น่าสนใจในการอภิปรายผลดังนี้

5.2.1 ความสำคัญของการจำแนกแบบหลายป้ายในบริบทงานจริง

การที่แบบจำลองสามารถทำการจำแนกแบบหลายป้ายได้นั้น ถือเป็นหัวใจสำคัญของการนำไปใช้ในองค์กร เนื่องจากพฤติกรรมการเขียนอีเมลทางธุรกิจมักจะรวมหลายความต้องการไว้ในฉบับเดียว การจำแนกแบบป้ายเดียวอาจทำให้ข้อมูลตกหล่น แต่การใช้การจำแนกแบบหลายป้ายช่วยให้ระบบสามารถจัดหมวดหมู่งานได้ครบถ้วนทุกมิติ

5.2.2 ประสิทธิภาพของเทคนิคหน่วยความจำระยะยาวแบบสั้น (LSTM)

สาเหตุที่หน่วยความจำระยะยาวแบบสั้น (LSTM) โดดเด่นกว่าเทคนิคอื่น เนื่องจากโครงสร้างที่สามารถจดจำลำดับคำและบริบทของภาษาไทยได้ดี อีกทั้งยังเป็นเทคนิคที่สามารถคำนวนค่าความน่าจำเป็น (Probability) ของทุกป้ายได้ เมื่อนำมาประยุกต์ใช้กับงานการจำแนกแบบหลายป้ายโดยการดึงป้ายชื่อที่มีความน่าจำเป็นสูงสุด 2 อันดับแรกมาใช้พยากรณ์ จึงสามารถจับความสัมพันธ์ระหว่างคำสำคัญของแต่ละกลุ่มงานที่ซ่อนอยู่ในข้อความเดียวกันได้อย่างมีประสิทธิภาพ

5.2.3 การปรับปรุงกระบวนการทำงานด้วยแอปพลิเคชัน

แอปพลิเคชันที่พัฒนาขึ้นพิสูจน์ให้เห็นว่าการนำปัญญาประดิษฐ์ (AI) มาผสานรวมกับเครื่องมือที่ใช้งานอยู่เดิม เช่น ไมโครซอฟต์เอาท์ลุค (Microsoft Outlook) ผ่านอินเตอร์เฟสโปรแกรมประยุกต์ (API) เป็นแนวทางที่สร้างผลกระทบเชิงบวกได้จริง ช่วยเปลี่ยนการทำงานด้วยมือ (Manual) มาเป็นระบบอัตโนมัติได้อย่างสมบูรณ์

5.3 ข้อเสนอแนะ

เนื่องจากข้อจำกัดด้านทรัพยากร งานวิจัยนี้จึงไม่ได้ใช้แบบจำลองภาษาขนาดใหญ่ (Large Language Models) ที่จำเป็นต้องมีเครื่องคอมพิวเตอร์ประสิทธิภาพสูงในการประมวลผล ซึ่งเทคนิคดังกล่าวจะให้ประสิทธิภาพในการเข้าใจความหมายเชิงบริบท (Contextual Meaning) ในระดับที่ลึกซึ้งยิ่งขึ้นและอาจช่วยเพิ่มความแม่นยำในกรณีที่ข้อความอีเมลมีความสั้นหรือมีความกำกวมสูงได้ และเนื่องด้วยบริบทการประยุกต์ใช้ในงานวิจัยนี้ถูกออกแบบให้ใช้เฉพาะกลุ่มงานที่ผู้วิจัยหยิบยกมาใช้ในศึกษา ถ้าจะนำไปใช้กับบริบทอื่นๆ สามารถนำกรอบทำงานนี้ไปประยุกต์ใช้ได้ แต่ต้องทำความเข้าใจลักษณะข้อมูลของงานที่ทำหรือปรึกษาผู้เชียวชาญเรื่องดังกล่าว (Domain Expert) ในการจัดกลุ่มหรือแบ่งป้ายชื่อข้อมูลให้ตรงกับวัตถุประสงค์ที่ควรจะเป็นได้

บรรณานุกรม

ภาษาไทย

นนทกุล เพชรรู้แจ้ง, ดุษฎี ประเสริฐธิติพงษ์ และวิจักษณ์ ศรีสัจจะเลิศวาจา. “โมเดลวิเคราะห์อารมณ์หลายป้ายสำหรับรีวิวภาษาไทยของร้านอาหาร.” *Procedia of Multidisciplinary Research*. 3 (2568).

บุญโย โยธนัต. “การจัดกลุ่มข้อความภาษาไทยด้วย K-Means และ TF-IDF ในไพทอนสำหรับการประยุกต์ทางการศึกษา.” *วารสารวิธีวิทยาการวิจัย*. 38 (2568) : 69–90.

สิรภพ แสนบุตร และฐิติรัตน์ ศิริบวรรัตนกุล. “การเพิ่มประสิทธิภาพการจำแนกตั๋วงานสำหรับระบบจัดการบริการด้านไอทีโดยใช้การเรียนรู้ของเครื่องจักร.” *วารสารวิจัยและพัฒนา มจธ.* 44 (2564) : 667–680.

ภาษาอังกฤษ

Abdulhamid, S. M., et al. “Comparative Analysis of Classification Algorithms for Email Spam Detection.” *International Journal of Computer Network and Information Security*. 10 (2018) : 60–67.

Adrales, L. F. and Sison, A. M. “Deep Learning Approaches to a School-based Email Classification with TF-IDF-based Vectorization.” *Proceedings of the 2024 IEEE 6th Symposium on Computers and Informatics (ISCI)*. 2024 : 97–105.

Berrar, D. “Bayes' Theorem and Naive Bayes Classifier.” *Encyclopedia of Bioinformatics and Computational Biology*. 1 (2019) : 403–412.

Bhatti, P., Jalil, Z. and Majeed, A. “Email Classification using LSTM: A Deep Learning Technique.” *Proceedings of the 2021 International Conference on Cyber Warfare and Security (ICCWS)*. 2021 : 97–105.

Borg, A., et al. “E-mail Classification with Machine Learning and Word Embeddings for Improved Customer Support.” *Neural Computing and Applications*. 33 (2021) : 1881–1902.

Brink, H., Richards, J. W. and Fetherolf, M. *Real-World Machine Learning*. Shelter Island, NY : Manning Publications, 2017.

Chuttur, M. Y. and Parianen, Y. "A Comparison of Machine Learning Models to Prioritise Emails using Emotion Analysis for Customer Service Excellence." *Knowledge Engineering and Data Science*. 5 (2022) : 41–52.

Dhina, M. M. and Sumathi, S. "An Innovative Approach to Classify Hierarchical Remarks with Multi-Class using BERT and Customized Naïve Bayes Classifier." *International Journal of Engineering, Science and Technology*. 13 (2021) : 32–45.

Dogra, V., et al. "A Complete Process of Text Classification System Using State-of-the-Art NLP Models." *Computational Intelligence and Neuroscience*. (2022) : 1–26.

Guo, G., et al. "KNN Model-Based Approach in Classification." *Lecture Notes in Computer Science*. Springer, 2003 : 986–996.

Heakl, A., et al. "ResuméAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models." *Procedia Computer Science*. (2024).

Hina, M., et al. "SeFACED: Semantic-Based Forensic Analysis and Classification of E-Mail Data Using Deep Learning." *IEEE Access*. 9 (2021) : 98398–98411.

Iqbal, F., et al. "Transfer Learning-based Forensic Analysis and Classification of E-Mail Content." *ACM Transactions on Asian and Low-Resource Language Information Processing*. (2023).

Phatthiyaphaibun, W., et al. "PyThaiNLP: Thai Natural Language Processing in Python." *arXiv preprint*. (2023).

Rizvi, S. M. H., et al. "Text Classification Using Graph Convolutional Networks: A Comprehensive Survey." *ACM Computing Surveys*. 57 (2025) : 1–38.

Russell, S. J. and Norvig, P. *Artificial Intelligence: A Modern Approach*. 4th ed. Harlow : Pearson Education Limited, 2021.

Sak, H., et al. “Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling.” *arXiv preprint*. (2014).

Xiao, W., et al. “Multi-Label Text Classification Model Based on Multi-Level Constraint Augmentation and Label Association Attention.” *ACM Transactions on Asian and Low-Resource Language Information Processing*. 23 (2024).



ประวัติผู้เขียน

ชื่อ	ธีระชัย เกษมพงศ์จรัส
ชื่อการค้นคว้าอิสระ	การจำแนกวัตถุประสงค์อีเมลด้วยวิธีการเรียนรู้ของเครื่อง
สาขาวิชา	วิทยาศาสตร์ข้อมูลเพื่อนวัตกรรม ภาควิชาเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
ประวัติ	ประวัติการศึกษา พ.ศ. 2562 สำเร็จการศึกษาระดับปริญญาตรี สาขาเทคโนโลยีอิเล็กทรอนิกส์ สถาบันการอาชีวศึกษากรุงเทพมหานคร ประวัติการทำงาน พ.ศ. 2562 ถึง 2569 บริษัท เก็ตออน เทคโนโลยี จำกัด (ปัจจุบัน) พ.ศ. 2561 ถึง 2562 บริษัท ดิจิตอล โฟกัส จำกัด พ.ศ. 2556 ถึง 2561 บริษัท กุลธรเคอร์บี้ จำกัด (มหาชน)
สถานที่ติดต่อ	เลขที่ 360/164 ร่มเกล้า 18 แขวงมีนบุรี เขตมีนบุรี กทม. 10510