PULINET Journal

การสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การสร้างฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้

มัรวาน จูแซ

หอสมุดจอห์น เอฟ เคนเนดี้ สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี

181 ถนนเจริญประดิษฐ์ ตำบลรูสะมิแล อำเภอเมืองปัตตานี จังหวัดปัตตานี 94000

Automated Data Extraction Using Web Scraping Techniques and AI for Creating News and Events Database in Thailand's Southern Border Provinces

Marwan Jusea

John F. Kennedy Library, Office of Academic Resources,

Prince of Songkla University, Pattani Campus

181 Charoen Pradit Road, Rusamilae, Mueang Pattani, Pattani 94000

E-mail: marwan.j@psu.ac.th

▶ รับบทความ 30 มิถุนายน 2568 ▶ แก้ไขบทความ 08 ธันวาคม 2568 ▶ ตอบรับบทความ 11 ธันวาคม 2568

บทคัดย่อ

บทความวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาและประยุกต์ใช้ระบบสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และเทคโนโลยีปัญญาประดิษฐ์ (AI) สู่การสร้างฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ของประเทศไทย และประเมินประสิทธิภาพของระบบที่พัฒนาขึ้นทั้งในเชิงเทคนิคและปริมาณ เพื่อยกระดับประสิทธิภาพการบริหารจัดการสารสนเทศของหอสมุดจอห์น เอฟ เคนเนดี้ สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี กระบวนการดำเนินงานประกอบด้วย 4 ขั้นตอนหลัก ได้แก่ (1) การรวบรวมข้อมูลด้วย Web Scraping จากแหล่งข่าวเป้าหมายทั้งระดับชาติและท้องถิ่นกว่า 15 แหล่ง (2) การใช้ AI ในการประมวลผลและจำแนกประเภทข้อมูลด้วยอัลกอริทึม Naive Bayes Classifier (3) การบูรณาการและลดความซ้ำซ้อนของข้อมูล และ (4) การออกแบบและจัดเก็บข้อมูลในฐานข้อมูล ผลการวิจัยพบว่า ระบบสามารถลดระยะเวลาในการจัดเก็บข้อมูลจาก 1–2 ชั่วโมงต่อวัน เหลือเพียง 10–15 นาทีต่อวัน ลดลงประมาณร้อยละ 85 พร้อมทั้งมีความถูกต้องในการดึงข้อมูลสูงถึงร้อยละ 97 ตรวจจับและป้องกันการบันทึกข่าวซ้ำได้ร้อยละ 98 นอกจากนี้ ระบบยังมีความแม่นยำในการจำแนกประเภทสูงถึงร้อยละ 92 พร้อมทั้งเพิ่มความครอบคลุมของแหล่งข่าวให้ครอบคลุมทั้งระดับชาติและท้องถิ่นรวมมากกว่า 15 แหล่งข่าว และลดอัตราความผิดพลาดในการประมวลผลเหลือเพียงร้อยละ 2 ระบบได้รับการประเมินจากผู้เชี่ยวชาญว่ามีประสิทธิภาพดีมาก ($\bar{X}$ = 4.85, SD = 0.18) โดยเฉพาะในด้านความแม่นยำและความรวดเร็วในการประมวลผล ผลลัพธ์ที่ได้สร้างประโยชน์ต่อการสนับสนุนการวางแผนนโยบาย การติดตามสถานการณ์ และการตัดสินใจเชิงยุทธศาสตร์ในพื้นที่ พร้อมทั้งวางรากฐานฐานข้อมูลที่เชื่อถือได้ ซึ่งสามารถต่อยอดเพื่อการใช้งานในอนาคตอย่างยั่งยืน

คำสำคัญ

ฐานข้อมูลจังหวัดชายแดนภาคใต้, Web Scraping, AI, การประมวลผลข้อมูลอัตโนมัติ, การสกัดข้อมูล

ผลงานนี้ได้ผ่านการนำเสนอในการประชุมวิชาการระดับชาติ PULINET ครั้งที่ 14 วันที่ 15-17 มกราคม 2568
จัดโดยสำนักวิทยบริการ มหาวิทยาลัยมหาสารคาม และข่ายงานห้องสมุดมหาวิทยาลัยส่วนภูมิภาค (PULINET)

Abstract

This research aims to develop and apply an automated data extraction system utilizing Web Scraping techniques and Artificial Intelligence (AI) to establish a comprehensive database of news and events in Thailand's Southern Border Provinces. Furthermore, the study evaluates the system's technical and quantitative performance to enhance information management efficiency at the John F. Kennedy Library, Prince of Songkla University, Pattani Campus.

The implementation process consists of four main stages: (1) data collection via Web Scraping from over 15 targeted national and local news sources; (2) data processing and classification using the Naive Bayes Classifier algorithm; (3) data integration and redundancy reduction; and (4) database design and storage.

The findings indicate that the system significantly reduced data archiving time from 1–2 hours to merely 10–15 minutes per day, representing a reduction of approximately 85%. The system demonstrated a high extraction accuracy of 97% and a duplicate detection rate of 98%. Additionally, the classification accuracy reached 92%, while the processing error rate was reduced to only 2%. The system successfully expanded coverage to include both national and local levels. Expert evaluation rated the system's overall efficiency as "very good" ($\bar{x}$ = 4.85, SD = 0.18), particularly highlighting its precision and processing speed. These results significantly contribute to supporting policy planning, situational monitoring, and strategic decision-making in the region, while establishing a reliable database foundation for sustainable future application.

Keywords

Southern Border Provinces Database, Web Scraping, AI, Automated Data Processing, Data Extraction

บทนำ (Introduction)

ในปัจจุบัน ข้อมูลข่าวสารจากสื่อออนไลน์มีบทบาทสำคัญอย่างยิ่งต่อการวิเคราะห์สถานการณ์และการตัดสินใจเชิงนโยบายในหลายมิติ โดยเฉพาะอย่างยิ่งในพื้นที่จังหวัดชายแดนภาคใต้ของประเทศไทย ซึ่งเป็นพื้นที่ที่เผชิญกับความขัดแย้งและเหตุการณ์ความไม่สงบมาอย่างต่อเนื่อง ส่งผลกระทบทั้งด้านเศรษฐกิจ สังคม และการเมือง รวมถึงภาพลักษณ์ของพื้นที่ในสายตาสาธารณชน การมีฐานข้อมูลข่าวสารที่ทันสมัย ถูกต้อง และครอบคลุมจึงเป็นปัจจัยสำคัญในการสนับสนุนการติดตามสถานการณ์ การวิจัย และการกำหนดนโยบายเชิงยุทธศาสตร์ในพื้นที่ดังกล่าว

เดิมทีการจัดเก็บข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ของหอสมุดจอห์น เอฟ เคนเนดี้ สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี ดำเนินการโดยเจ้าหน้าที่ติดตามข่าวจากเว็บไซต์ต่าง ๆ แล้วคัดลอกและบันทึกลงฐานข้อมูลด้วยตนเอง กระบวนการดังกล่าวใช้เวลาเฉลี่ย 1–2 ชั่วโมงต่อวัน และเผชิญปัญหาหลายประการ เช่น ความล่าช้าในการปรับปรุงข้อมูล ความคลาดเคลื่อนจากการป้อนข้อมูลด้วยมือ (Human Error) การเกิดข้อมูลซ้ำเมื่อมีข่าวเดียวกันจากหลายแหล่ง และข้อจำกัดด้านเวลาทำการของห้องสมุดที่ไม่สามารถติดตามข่าวได้ตลอด 24 ชั่วโมง ปัจจัยเหล่านี้ทำให้ประสิทธิภาพการให้บริการข้อมูลข่าวสารแก่ผู้ใช้ เช่น นักวิจัย นักศึกษา และหน่วยงานภายนอก ลดลงทั้งในด้านความรวดเร็ว ความครอบคลุม และการเข้าถึงข้อมูลในช่วงเวลาที่ต้องการ

เทคนิค Web Scraping ได้รับการยอมรับอย่างแพร่หลายว่าเป็นเครื่องมือที่มีประสิทธิภาพสำหรับดึงข้อมูลจากเว็บไซต์โดยอัตโนมัติ สามารถลดเวลาและภาระงานในการรวบรวมข้อมูลขนาดใหญ่จากหลายแหล่งและช่วยลดการพึ่งพา

กำลังคน ซึ่งสอดคล้องกับผลการศึกษาของ Mitchell (2018), Pant et al. (2024) และ Bhatt et al. (2023) ที่พบว่า การนำ Web Scraping มาใช้ในงานด้านข่าวสารและงานวิจัยช่วยให้การรวบรวมและจัดเตรียมข้อมูลมีความรวดเร็วและแม่นยำมากขึ้น นอกจากนี้ Valova et al. (2023) ยังได้ทบทวนองค์ความรู้และแนวโน้มปัจจุบันของ Web Scraping ในฐานะเทคโนโลยีระดับ State of the Art โดยสรุปเทคนิคและแนวทางสำคัญ เช่น การดึงข้อมูลจากหน้าเว็บแบบคงที่และแบบไดนามิก การจัดการเนื้อหาที่เรนเดอร์ด้วย JavaScript การรับมือกับกลไกป้องกันบอทของเว็บไซต์ รวมถึงการคำนึงถึงข้อจำกัดด้านกฎหมายและจริยธรรมในการเก็บข้อมูลจากเว็บ งานดังกล่าวชี้ให้เห็นว่า การออกแบบระบบ Web Scraping ที่ดีจำเป็นต้องพิจารณาทั้งมิติด้านเทคนิค (ความยืดหยุ่นต่อการเปลี่ยนแปลงโครงสร้างเว็บ ความเสถียรของตัวดึงข้อมูล และประสิทธิภาพในการประมวลผลข้อมูลจำนวนมาก) และมิติด้านจริยธรรม/นโยบายควบคู่กัน ซึ่งเป็นกรอบคิดสำคัญที่ถูกนำมาประยุกต์ใช้ในการออกแบบระบบของการวิจัยครั้งนี้

งานวิจัยในประเทศ เช่น จักรินทร์ สันติรัตนภักดี (2022), ศตวรรษ รามไชย และผุสดี พรผล (2022) และ Farias et al. (2024) ยังชี้ให้เห็นถึงศักยภาพของ Web Scraping ในการจัดการข้อมูลจากหลายแหล่งในหลากหลายบริบท ทั้งด้านการคมนาคม การท่องเที่ยว การตลาด และการศึกษา อย่างไรก็ตาม งานส่วนใหญ่เน้นบริบทเชิงพาณิชย์หรือการประยุกต์ใช้ในระดับทั่วไป ยังไม่พบการพัฒนาระบบที่มุ่งรองรับการจัดการข้อมูลข่าวสารในพื้นที่เฉพาะที่มีความเปราะบางเชิงสถานการณ์ เช่น จังหวัดชายแดนภาคใต้ อย่างเป็นระบบและต่อเนื่อง

พร้อมกันนี้ ความก้าวหน้าของเทคโนโลยีปัญญาประดิษฐ์ (AI) และการประมวลผลภาษาธรรมชาติ (NLP) เปิดโอกาสให้สามารถนำ AI มาใช้ในการคัดกรอง วิเคราะห์ และจำแนกประเภทข่าวสารได้อย่างมีประสิทธิภาพ เช่น การจำแนกข่าวทั่วไปออกจากข่าวเหตุการณ์ความไม่สงบ การตรวจจับข่าวซ้ำจากหลายแหล่ง และการลดความซ้ำซ้อนของข้อมูลผ่านการกำหนดระเบียนหลักของ “ชุดเหตุการณ์” เดียวกัน การบูรณาการ Web Scraping กับ AI จึงเป็นแนวทางสำคัญในการสร้างระบบฐานข้อมูลข่าวที่สามารถดึงข้อมูลอัตโนมัติ กรองข้อมูลที่เกี่ยวข้องอย่างแม่นยำ และลดภาระงานเชิงปฏิบัติของเจ้าหน้าที่ในขณะเดียวกันก็ยกระดับคุณภาพข้อมูลให้ตอบโจทย์การใช้งานเชิงวิเคราะห์ของหน่วยงานต่าง ๆ ทั้งภายในและภายนอกมหาวิทยาลัย

จากเหตุผลดังกล่าว ผู้วิจัยจึงพัฒนาระบบสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping ร่วมกับเทคโนโลยี AI เพื่อสร้างฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ โดยออกแบบให้ระบบสามารถดึงข่าวจากแหล่งข่าวออนไลน์ทั้งระดับชาติและท้องถิ่นมากกว่า 15 แหล่ง ประมวลผลและจำแนกประเภทข่าวด้วยโมเดล Naïve Bayes ตรวจจับและจัดการข่าวซ้ำในระดับ “ชุดเหตุการณ์” และจัดเก็บข้อมูลในฐานข้อมูลที่มีโครงสร้างรองรับการสืบค้นอย่างเป็นระบบ งานวิจัยนี้จึงมีเป้าหมายทั้งในเชิงพัฒนาระบบและเชิงประเมินประสิทธิภาพ เพื่อให้ได้ต้นแบบระบบฐานข้อมูลข่าวที่มีความน่าเชื่อถือสามารถใช้งานจริงในบริบทของหอสมุดจอห์น เอฟ เคนเนดี้ และต่อยอดเป็นโครงสร้างพื้นฐานสำหรับการติดตามสถานการณ์และการวางนโยบายในพื้นที่จังหวัดชายแดนภาคใต้ในระยะยาว

วัตถุประสงค์ (Objective)

1. เพื่อพัฒนาและประยุกต์ใช้การสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การพัฒนาฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้

2. เพื่อประเมินประสิทธิภาพของระบบการสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การพัฒนาฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้

วิธีดำเนินการวิจัย (Methodology)

การวิจัยนี้เป็นการวิจัยและพัฒนา (Research and Development: R&D) โดยมีวัตถุประสงค์หลักเพื่อพัฒนาและประเมินประสิทธิภาพของระบบการสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การสร้างฐานข้อมูลข่าวและ

เหตุการณ์ในจังหวัดชายแดนภาคใต้ ระบบที่พัฒนาขึ้นได้ถูกนำไปทดลองใช้งานจริงที่หอสมุดจอห์น เอฟ เคนเนดี้ 0สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี และมีการประเมินผลประสิทธิภาพในสองมิติหลัก คือ มิติเชิงคุณภาพ (วัดจากความคิดเห็นของผู้เชี่ยวชาญ) และมิติเชิงปริมาณ (วัดด้วยตัวชี้วัดประสิทธิภาพ หรือ KPIs)

1. ประชากรและกลุ่มตัวอย่าง

ประชากร (Population): ข้อมูลข่าวสารและเหตุการณ์ทั้งหมดที่เกี่ยวข้องกับพื้นที่จังหวัดชายแดนภาคใต้ของประเทศไทย ซึ่งเผยแพร่ผ่านแหล่งข่าวออนไลน์ต่าง ๆ ทั้งในระดับชาติและท้องถิ่น

กลุ่มตัวอย่างที่ใช้ในการประเมินประสิทธิภาพของระบบ (Sample for System Evaluation) แบ่งออกเป็น 2 ส่วน ดังนี้:

- กลุ่มผู้เชี่ยวชาญ: ผู้เชี่ยวชาญทางด้านการพัฒนาระบบสารสนเทศและเทคโนโลยีสารสนเทศ จำนวน 5 ท่าน ประเมินประสิทธิภาพโดยรวมและความเหมาะสมของระบบ
- ชุดข้อมูลสำหรับทดสอบระบบ (Test Dataset for System Evaluation): ข่าวสารที่ถูกรวบรวมจากแหล่งข่าวเป้าหมายมากกว่า 15 แหล่ง โดยกำหนดขนาดกลุ่มตัวอย่างสำหรับวัดตัวชี้วัดประสิทธิภาพเชิงเทคนิค (KPIs) จำนวน 100 ข่าว (N=100) ซึ่งได้มาจากการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling) เพื่อให้ข้อมูลเป็นตัวแทนที่ดีของประชากรข่าวทั้งหมด

ขั้นตอนการคัดเลือกกลุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling Procedure):

- ขั้นตอนที่ 1 การกำหนดกรอบตัวอย่าง (Sampling Frame Definition): ดำเนินการคัดกรองข้อมูลขั้นต้นโดยการตัดข่าวที่มีเนื้อหาซ้ำซ้อน หรือมีความคล้ายคลึงกันสูงออกจากประชากร เพื่อให้ได้กรอบตัวอย่างที่มีคุณภาพ
- ขั้นตอนที่ 2 การจัดชั้นภูมิ (Stratification): จำแนกประชากรข่าวออกเป็นชั้นภูมิ (Strata) โดยใช้เกณฑ์ 3 ด้าน ได้แก่ (1) แหล่งข่าวทั้ง 15 แหล่ง, (2) หมวดหมู่ข่าวหลัก (เช่น ข่าวเหตุการณ์ความไม่สงบ และข่าวทั่วไป), และ (3) ช่วงเวลาที่เผยแพร่ (เช่น สัปดาห์ที่ 1-4 ของเดือน)
- ขั้นตอนที่ 3 การจัดสรรขนาดตัวอย่าง (Allocation): กำหนดโควตาจำนวนตัวอย่างที่จะสุ่มจากแต่ละชั้นตามสัดส่วนปริมาณข่าวจริงที่เกิดขึ้นในชั้นนั้น ๆ พร้อมทั้งกำหนดจำนวนตัวอย่างขั้นต่ำสำหรับชั้นที่มีปริมาณข่าวน้อย (เช่น แหล่งข่าวท้องถิ่น) เพื่อให้มั่นใจว่าทุกกลุ่มได้ถูกรวมอยู่ในการประเมิน
- ขั้นตอนที่ 4 การสุ่มตัวอย่าง (Randomization): ดำเนินการสุ่มตัวอย่างแบบง่าย (Simple Random Sampling) ภายในแต่ละชั้นภูมิ จนครบตามจำนวนโควตาที่กำหนดไว้

2. เครื่องมือวิจัย

เครื่องมือวิจัยที่ใช้ในการศึกษาครั้งนี้สามารถแบ่งออกเป็น 2 ประเภทหลัก คือ 1) เครื่องมือที่ใช้ในการพัฒนาระบบ และ 2) เครื่องมือสำหรับการประเมินและวัดผลประสิทธิภาพของระบบ

2.1 เครื่องมือที่ใช้ในการพัฒนาระบบ

เครื่องมือหลักในส่วนนี้ คือ ตัวระบบที่ถูกพัฒนาขึ้นเพื่อใช้ในการสกัด รวบรวม และจัดการข้อมูลโดยอัตโนมัติ ซึ่งประกอบด้วยเทคโนโลยีสำคัญ 2 กลุ่ม:

- เทคโนโลยีสำหรับการสกัดและรวบรวมข้อมูล (Web Scraping Technology) ผู้วิจัยพัฒนาระบบดึงข้อมูลอัตโนมัติด้วยภาษา Python โดยประยุกต์ใช้เครื่องมือและไลบรารีหลักหลายตัวเพื่อให้สามารถทำงานได้อย่างมีประสิทธิภาพ เริ่มจากการใช้ Requests ในการเข้าถึงเว็บไซต์เป้าหมายเพื่อดึงข้อมูลจากแหล่งข่าวกว่า 15 แหล่ง จากนั้นนำ BeautifulSoup มาใช้ในการจัดการและวิเคราะห์โครงสร้าง HTML เพื่อระบุตำแหน่งของข้อมูลที่ต้องการ เช่น หัวข้อข่าว วันที่เผยแพร่ และเนื้อหาข่าว โดยอาศัยการวิเคราะห์โครงสร้างเว็บไซต์ผ่าน Browser DevTools ประกอบ สำหรับเว็บไซต์ที่มี

ความซับซ้อนและต้องการเพิ่มประสิทธิภาพในการดึงข้อมูลจากหลายแหล่งพร้อมกัน ผู้วิจัยได้ใช้ Scrapy ซึ่งรองรับการทำงานแบบ Concurrent Crawling เพื่อให้การเก็บข้อมูลเป็นไปอย่างรวดเร็วยิ่งขึ้น นอกจากนี้ ยังได้ตั้งค่า Cron Job บนระบบปฏิบัติการ Linux เพื่อกำหนดตารางเวลาในการทำงานอัตโนมัติ โดยระบบจะดำเนินการดึงข้อมูลตามรอบเวลาที่กำหนดไว้ เช่น ทุกวันเวลา 03:00 น. เป็นต้น

- เทคโนโลยีปัญญาประดิษฐ์สำหรับการประมวลผลและการจำแนกข้อมูล (AI/NLP Technology) ระบบประมวลผลข้อมูลเริ่มต้นด้วยขั้นตอนการเตรียมข้อมูล (Data Preprocessing) โดยประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ผ่านกระบวนการตัดคำ (Tokenization) และการทำความสะอาดข้อความ (Text Normalization) เพื่อกำจัดส่วนประกอบที่ไม่จำเป็น เช่น โฆษณา หรือลิงก์ต่าง ๆ ที่ไม่เกี่ยวข้องกับเนื้อหาข่าวสาร สำหรับการจำแนกประเภทข่าว ผู้วิจัยได้พัฒนาโมเดลการจำแนกข้อมูล (Classification Model) โดยใช้อัลกอริทึม Naive Bayes Classifier ในการทำนายและจัดหมวดหมู่ข่าวสารออกเป็น 2 กลุ่มหลัก ได้แก่ "ข่าวทั่วไป" และ "ข่าวเหตุการณ์ความไม่สงบ" ทั้งนี้ โมเดลได้ถูกฝึกสอนด้วยชุดข้อมูลข่าวที่ผ่านการระบุประเภท (Labeled Data) โดยผู้เชี่ยวชาญ เพื่อให้ระบบสามารถเรียนรู้คุณลักษณะสำคัญของข้อความ (Features) ต่าง ๆ เช่น คำสำคัญ (Keywords) และความถี่ของคำที่ปรากฏในข่าวแต่ละประเภท

2.2 เครื่องมือสำหรับการประเมินและวัดผลประสิทธิภาพของระบบ

การประเมินและวัดผลประสิทธิภาพของระบบในการวิจัยครั้งนี้ดำเนินการในสองมิติหลัก ได้แก่ การประเมินเชิงคุณภาพ และการประเมินเชิงเทคนิค โดยใช้เครื่องมือที่ได้รับการออกแบบมาอย่างเหมาะสม ดังนี้

- การประเมินเชิงคุณภาพโดยผู้เชี่ยวชาญ (Qualitative Evaluation): ผู้วิจัยได้พัฒนาแบบประเมินสำหรับผู้เชี่ยวชาญเพื่อประเมินประสิทธิภาพและความเหมาะสมของการนำเทคนิค Web Scraping และปัญญาประดิษฐ์มาใช้ในการรวบรวมและจัดเก็บข้อมูล โดยผู้เชี่ยวชาญทางด้านการพัฒนาระบบสารสนเทศและเทคโนโลยีสารสนเทศ จำนวน 5 ท่าน ทำการประเมิน ประเด็นการประเมินครอบคลุม 7 ด้านสำคัญ ได้แก่ ประสิทธิภาพในการรวบรวมข้อมูล ความแม่นยำในการประมวลผล ความรวดเร็ว การใช้งานปัญญาประดิษฐ์ ความง่ายและความยืดหยุ่นในการใช้งาน ความปลอดภัยและการปฏิบัติตามกฎหมาย และการรองรับการพัฒนาต่อเนื่อง ผลการประเมินถูกนำมาวิเคราะห์โดยการคำนวณหาค่าเฉลี่ย ($\bar{x}$) และส่วนเบี่ยงเบนมาตรฐาน (SD)

- การประเมินเชิงปริมาณด้วยตัวชี้วัดประสิทธิภาพ (Quantitative Evaluation - KPIs): ผู้วิจัยได้กำหนดตัวชี้วัดประสิทธิภาพหลัก (Key Performance Indicators: KPIs) จำนวน 5 ด้าน เพื่อวัดผลการทำงานของระบบเชิงปริมาณ ประกอบด้วย 1) ความถูกต้องในการดึงข้อมูล (Data Extraction Accuracy) 2) อัตราการตรวจจับข่าวซ้ำ (Duplicate Detection Rate: DDR) 3) ความแม่นยำในการจำแนกประเภทข่าว (Categorization Accuracy) 4) ความครอบคลุมของแหล่งข่าว (Source Coverage) และ 5) อัตราความผิดพลาด (Error Rate)

ตารางที่ 1 รายละเอียดตัวชี้วัดประสิทธิภาพ (KPIs)

ตัวชี้วัดประสิทธิภาพ (KPI)	วัตถุประสงค์การวัด	ชุดข้อมูลทดสอบ	สูตรการคำนวณ
1. ความถูกต้องในการดึงข้อมูล (Data Extraction Accuracy)	ตรวจสอบว่าข้อมูลที่ดึงมา (เช่น หัวข้อ, วันที่, เนื้อหา) ตรงกับต้นฉบับ	**N=100 ข่าว** สุ่มด้วยวิธี การคัดเลือกกลุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling) เพื่อเปรียบเทียบกับต้นฉบับ	$\boldsymbol{Accuracy_{article}} = \frac{\#\text{ข่าวที่ถูกต้องครบทุกฟิลด์}}{\boldsymbol{N}} \times \mathbf{100\%}$ • N=100 ข่าว
2. อัตราการตรวจจับข่าวซ้ำ (Duplicate Detection Rate: DDR)	วัดความสามารถในการระบุและป้องกันการบันทึก "ชุดเหตุการณ์ซ้ำจริง"	สร้าง Gold Standard โดยผู้เชี่ยวชาญและวัดค่าตามสูตร	$\boldsymbol{DDR} = \frac{\boldsymbol{TP_{dup}}}{\boldsymbol{FP_{dup}} + \boldsymbol{FN_{dup}}} \times \mathbf{100\%}$ • $\boldsymbol{TP_{dup}}$: ชุดข่าวซ้ำที่ระบบตรวจจับถูกต้อง • $\boldsymbol{FN_{dup}}$: ชุดข่าวซ้ำที่มีจริงแตะระบบพลาด • $\boldsymbol{FP_{dup}}$: ชุดที่ระบบบอกว่าซ้ำแต่จริง ๆ แล้วไม่ซ้ำ
3. ความแม่นยำในการจำแนกประเภทข่าว (Categorization Accuracy)	วัดความถูกต้องของการทำนายประเภทข่าว (เหตุการณ์ความไม่สงบ/ข่าวทั่วไป) โดยโมเดล Naive Bayes	ใช้ **N=100 ข่าว** สุ่มด้วยวิธี การคัดเลือกกลุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling)	$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \times \mathbf{100\%}$ • $\boldsymbol{TP}$**:** โมเดลทำนายว่า เหตุการณ์ความไม่สงบ และความจริงก็เป็นเหตุการณ์ความไม่สงบ (ทายถูกฝั่ง Positive) • $\boldsymbol{TN}$**:** โมเดลทำนายว่า เหตุการณ์ทั่วไป และความจริงก็เป็น เหตุการณ์ทั่วไป (ทายถูกฝั่ง Negative) • $\boldsymbol{FP}$**:** โมเดลทำนายว่า เหตุการณ์ความไม่สงบ แต่ความจริงเป็น ข่าวทั่วไป (เตือนเกินเหตุ) • $\boldsymbol{FN}$**:** โมเดลทำนายว่า ข่าวทั่วไป แต่ความจริงก็เป็น ข่าวเหตุการณ์ความไม่สงบ (พลาดเหตุสำคัญ)
4. ความครอบคลุมของแหล่งข่าว (Source Coverage)	ตรวจสอบความสามารถในการดึงข่าวจากแหล่งข่าว	นับจำนวนแหล่งข่าวที่ระบบดึงได้จริง ($\boldsymbol{S}_{\text{จริง}}$) เทียบกับ	$\boldsymbol{Coverage} = \frac{\boldsymbol{S}_{\text{จริง}}}{\boldsymbol{S}_{\text{เป้าหมาย}}} \times \mathbf{100\%}$

ตัวชี้วัดประสิทธิภาพ (KPI)	วัตถุประสงค์การวัด	ชุดข้อมูลทดสอบ	สูตรการคำนวณ
	หลากหลายทั้งระดับชาติและท้องถิ่น	บัญชีเป้าหมายที่กำหนดไว้ 15 แห่ง	• $S_{\text{จริง}}$: จำนวนแหล่งข่าวที่ระบบดึงได้จริง • $S_{\text{เป้าหมาย}}$: จำนวนแหล่งข่าวตามบัญชีเป้าหมาย
5. อัตราความผิดพลาด (Error Rate)	วัดสัดส่วนข่าวที่มีปัญหา (เนื้อหาขาด/ผิดรูปแบบ/เมตาดาต้าผิด/ไฟล์เสียหาย)	**N=100 ข่าว** สุ่มด้วยวิธี การคัดเลือกกลุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling)	$\boldsymbol{Error\ Rate = \frac{E}{N} \times 100\%}$ • E: จำนวนข่าวที่พบความผิดพลาด • N: จำนวนข่าวที่สุ่มตรวจ

***หมายเหตุ** การคัดเลือกกลุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling) ถูกใช้เพื่อคัดเลือกกลุ่มตัวอย่างทดสอบ N=100 ข่าวสำหรับ KPI เชิงปริมาณ เพื่อให้มั่นใจว่าผลลัพธ์เป็นตัวแทนที่เที่ยงตรงของประชากรข่าวทั้งหมด

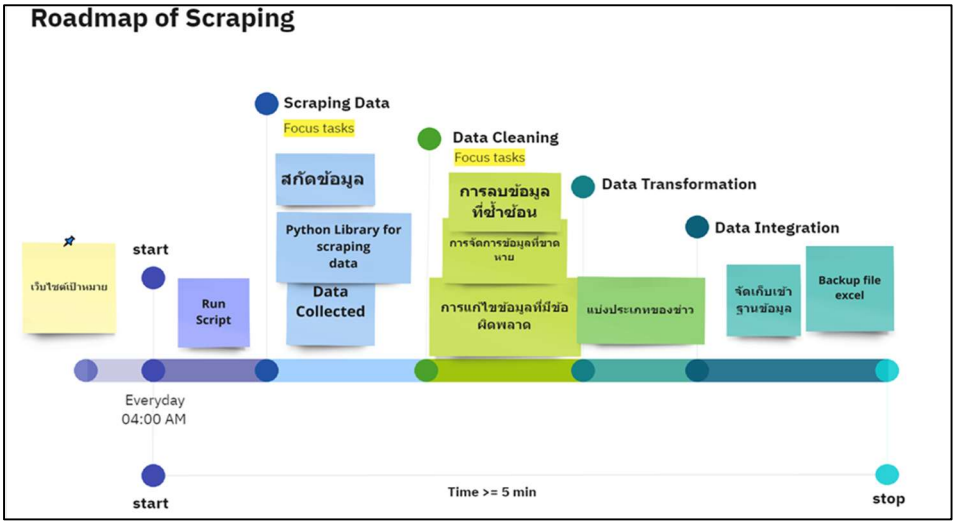
3. ขั้นตอนการดำเนินการวิจัย

ขั้นตอนการดำเนินการวิจัยครั้งนี้ได้ออกแบบให้เป็นกระบวนการที่เป็นลำดับขั้นและเชื่อมโยงกัน ประกอบด้วย 4 ขั้นตอนหลัก คือ 1) การเก็บรวบรวมข้อมูลด้วย Web Scraping 2) การใช้ AI ในการประมวลผลและจำแนกข้อมูล 3) การบูรณาการข้อมูลจากแหล่งข่าวต่าง ๆ และ 4) การออกแบบและจัดเก็บข้อมูลในฐานข้อมูล

ขั้นตอนที่ 1: การเก็บรวบรวมข้อมูลด้วย Web Scraping

การเก็บรวบรวมข้อมูลด้วย Web Scraping ในขั้นตอนนี้มีเป้าหมายหลัก คือ การดึงข้อมูลจากแหล่งข่าวออนไลน์ต่าง ๆ แบบเรียลไทม์ ซึ่งข้อมูลที่เก็บได้จะถูกดึงในรูปแบบที่สอดคล้องกับโครงสร้างเฉพาะของแต่ละเว็บไซต์ ทำให้สามารถเข้าถึงข้อมูลได้อย่างสะดวกและรวดเร็ว กระบวนการนี้ประกอบด้วยรายละเอียดเชิงลึก ดังนี้

- **การวิเคราะห์โครงสร้างเว็บไซต์เป้าหมาย** การคัดเลือกแหล่งข่าวเริ่มจากการกำหนดหลักเกณฑ์ที่ชัดเจนเพื่อให้ได้ข้อมูลที่น่าเชื่อถือและมีคุณภาพ ซึ่งหลักเกณฑ์ในการคัดเลือกแหล่งข่าวประกอบด้วย การพิจารณาความน่าเชื่อถือของเว็บไซต์ ความสม่ำเสมอในการอัปเดตข่าว ความครอบคลุมของเนื้อหาเฉพาะในพื้นที่จังหวัดชายแดนภาคใต้ รวมถึงความเข้ากันได้กับระบบ Web Scraping ที่จะนำมาใช้งาน จากการคัดเลือกนี้ ทีมพัฒนาสามารถรวบรวมแหล่งข่าวที่ตรงตามหลักเกณฑ์ได้ทั้งหมด 15 แหล่งข่าว โดยแต่ละแหล่งข่าวมีโครงสร้าง HTML ที่แตกต่างกัน ทำให้ทีมพัฒนาต้องเริ่มต้นด้วยการวิเคราะห์โครงสร้างภายในของเว็บไซต์แต่ละแห่งอย่างละเอียด เพื่อทำความเข้าใจการจัดวางเนื้อหาข่าว เช่น การจัดตำแหน่งของหัวข้อข่าว วันที่ ผู้เขียน และเนื้อหาข่าวหลัก ทั้งนี้ เพื่อให้การดึงข้อมูลสามารถระบุตำแหน่งข้อมูลได้อย่างถูกต้องและรวดเร็ว ในขั้นตอนนี้ ทีมพัฒนาจะใช้เครื่องมือสำหรับตรวจสอบโครงสร้างเว็บ เช่น DevTools ของเบราว์เซอร์ เพื่อช่วยในการตรวจสอบและแยกแยะโครงสร้าง HTML ที่เกี่ยวข้อง ทำให้การออกแบบโค้ด Web Scraping เป็นไปได้อย่างมีประสิทธิภาพ
- **การพัฒนาชุดคำสั่งในการ Scraping ข้อมูล** เป็นขั้นตอนสำคัญที่ต้องทำอย่างละเอียดดังภาพที่ 2 โค้ดที่เขียนจะต้องสามารถดึงข้อมูลได้อย่างแม่นยำในหลากหลายรูปแบบ เพื่อระบุองค์ประกอบต่าง ๆ ที่ต้องการดึงข้อมูล เช่น หัวข้อข่าว ลิงก์ วันที่เผยแพร่ และเนื้อหาข่าว ขึ้นอยู่กับลักษณะการจัดวางของแต่ละเว็บไซต์ โดยภาษา Python จะเป็นภาษาหลักที่ใช้ในการเขียนโค้ด ร่วมกับไลบรารีสำคัญ เช่น BeautifulSoup สำหรับการจัดการ HTML และ Requests สำหรับการเข้าถึงเว็บไซต์ รวมถึง Scrapy หากต้องการประสิทธิภาพในการเก็บข้อมูลจากหลายแหล่งพร้อมกัน (Mitchell, 2018; Pant et al., 2024) โดยมีการเชื่อมโยงข้อมูลระบบ ดังนี้:



ภาพที่ 1 แสดง Roadmap of Scraping

○ **การตั้งค่าชุดคำสั่ง Web Scraping** ชุดคำสั่งถูกพัฒนาขึ้นด้วยภาษา Python โดยใช้ไลบรารี เช่น BeautifulSoup และ Requests สำหรับการดึงข้อมูลจากเว็บไซต์เป้าหมาย โดยกำหนดเงื่อนไขการดึงข้อมูล เช่น URL ของแหล่งข่าว โครงสร้าง HTML และข้อมูลที่ต้องการดึง (หัวข้อข่าว, วันที่เผยแพร่ และเนื้อหาข่าว)

```
hdr = {'User-agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/110.0.0.0 Safari/537.36'}
pages = np.arange(0,10,10)
for page_num in pages:
    url = "https://www.israneus.org/article/south-news/other-news.html?start=" +str(page_num)
    print(url)
    page = requests.get(url,headers=hdr)
    soup = BeautifulSoup(page.text,'html.parser')
    sleep(randint(2,8))
    for store in soup.findAll('h3',class_='contentheading'):
        header = store.a.text
        base_link = "https://www.israneus.org"
        link = store.find('a')['href']
        url = urllib.parse.urljoin(base_link,link)
        page_url = requests.get(url,headers=hdr)
        soup_url = BeautifulSoup(page_url.content,'html.parser')
        web_data = soup_url.findAll('div', class_='flexicontent')
        sleep(randint(2,8))
        for d in web_data:
            name = d.h1.span.text.strip()
            news_name.append(name)
            url_news.append(url)
            date = d.find('div',class_='flexi value field_created').text.strip()
            date_time.append(date)
            created = d.find('div', class_='flexi value field_created_by').text.strip()
            created_by.append(created)
            categories = d.find('div',class_='flexi value field_categories').text.strip()
            news_categories.append(categories)
            tags = d.find('div',class_='flexi value field_tags').text.strip()if d.find('div',class_='flexi value field_tags') else '*****'
            news_tags.append(tags)
            value = d.findAll('div',class_='desc-content field_text')
```

ภาพที่ 2 แสดงตัวอย่างชุดคำสั่ง Web Scraping

○ **การตั้งค่า Cron Job** สำหรับการทำงานอัตโนมัติ ตั้งเวลา Cron Job บนเซิร์ฟเวอร์ Linux โดยเพิ่มคำสั่งในไฟล์ Crontab เพื่อให้สคริปต์ทำงานอัตโนมัติ ดังนี้:

```
0 3 * * * /usr/bin/python3 /path/to/scraping_script.py >> /path/to/logs/scraping.log 2>&1
```

ภาพที่ 3 แสดงตัวอย่างชุดคำสั่งการเชื่อมโยงระบบ

- 0 3 * * * หมายถึง ทำงานทุกวันเวลา 03:00 น.
- /usr/bin/python3 ระบุตำแหน่งของ Python
- /path/to/scraping_script.py ระบุที่อยู่ของสคริปต์หลัก
- ข้อมูล Log จะถูกบันทึกไว้ในไฟล์ scraping.log เพื่อตรวจสอบการทำงาน

- **การควบคุมและตรวจสอบการดึงข้อมูลอย่างต่อเนื่อง** เพื่อให้การดึงข้อมูลเป็นไปอย่างราบรื่น ทีมพัฒนาได้ตั้งค่าให้ Web Scraping สามารถทำงานได้แบบอัตโนมัติและมีการกำหนดรอบการดึงข้อมูลตามความถี่ที่เหมาะสม เช่น การตั้งเวลาทำงาน (Scheduler) เพื่อดึงข้อมูลทุก ๆ วัน เพื่อป้องกันการใช้ทรัพยากรมากเกินไป การตั้งเวลาจะช่วยควบคุมปริมาณข้อมูลที่ดึงมาและลดความเสี่ยงจากการถูกปิดกั้นจากเว็บไซต์ นอกจากนี้ ระบบยังถูกออกแบบให้สามารถส่งอีเมลแจ้งเตือนไปยังเจ้าหน้าที่ทุกวันหลังจากการทำงานของระบบเสร็จสิ้น เพื่อป้องกันปัญหาที่อาจเกิดขึ้น เช่น ข้อมูลดึงมาไม่ครบถ้วนหรือโครงสร้างเว็บไซต์มีการเปลี่ยนแปลง โดยเจ้าหน้าที่สามารถตรวจสอบและดำเนินการแก้ไขได้ทันท่วงที

- **การจัดการปัญหาการบล็อกการเข้าถึงเว็บไซต์** เนื่องจากเว็บไซต์บางแห่งอาจมีการป้องกันการดึงข้อมูลอัตโนมัติด้วยเทคนิคต่าง ๆ เช่น การจำกัดจำนวนครั้งในการเข้าถึง ทีมพัฒนาได้มีการวางแผนแก้ปัญหาไว้ล่วงหน้า โดยการใช้ Proxies และการหมุนเวียน IP เพื่อไม่ให้ถูกตรวจจับและจำกัดการเข้าถึง

- **การจัดการกับการเปลี่ยนแปลงของโครงสร้างเว็บไซต์** เว็บไซต์ข่าวออนไลน์อาจเปลี่ยนแปลงโครงสร้าง HTML ของหน้าเว็บตามการอัปเดต ทีมพัฒนาได้ออกแบบโค้ด Web Scraping ให้ยืดหยุ่นต่อการเปลี่ยนแปลง โดยมีการตั้งค่าตรวจสอบความถูกต้องของข้อมูลที่ดึงมา หากพบความผิดปกติ (เช่น ข้อมูลที่ดึงมาไม่ครบ หรือไม่มีการอัปเดต) ระบบจะส่งการแจ้งเตือนผ่านอีเมลทีมพัฒนา เพื่อให้ทีมพัฒนาสามารถปรับปรุงโค้ดให้สอดคล้องกับโครงสร้างที่เปลี่ยนแปลงได้อย่างรวดเร็ว

- **การบันทึกและจัดเก็บข้อมูลเบื้องต้นก่อนการประมวลผล** ข้อมูลที่ดึงมาได้จะถูกจัดเก็บในรูปแบบไฟล์ CSV เบื้องต้น เพื่อเตรียมสำหรับการนำเข้าสู่ระบบ AI ในขั้นตอนถัดไป ไฟล์เหล่านี้ช่วยให้สามารถตรวจสอบข้อมูลดิบได้อย่างง่ายดาย และเป็นแหล่งข้อมูลสำรองในกรณีที่ต้องย้อนกลับมาตรวจสอบข้อมูลที่ได้จาก Web Scraping

ขั้นตอนที่ 2: การใช้ AI ในการประมวลผลและจำแนกข้อมูล

หลังจากที่ข้อมูลข่าวสารดิบถูกดึงมาจากเว็บไซต์ผ่านกระบวนการ Web Scraping แล้ว ข้อมูลเหล่านี้จะถูกส่งไปยังระบบปัญญาประดิษฐ์ (AI) เพื่อประมวลผลและจำแนกหมวดหมู่ โดย AI ที่นำมาใช้ในระบบนี้มีวัตถุประสงค์หลักเพื่อช่วยจำแนกและกรองข่าวสารที่เกี่ยวข้องสำหรับพื้นที่จังหวัดชายแดนภาคใต้โดยเฉพาะ (Bhujbal et al., 2023) โดยมีรายละเอียดดังนี้

- **การเตรียมข้อมูลและการทำความสะอาดข้อมูลดิบ** ข้อมูลข่าวสารที่ได้จาก Web Scraping อาจประกอบด้วยข้อความ รูปภาพ และเมตาดาต้าที่ไม่ได้ถูกจัดรูปแบบในแบบเดียวกันทั้งหมด ดังนั้น ขั้นตอนแรกของการประมวลผลด้วย AI คือ การทำความสะอาดข้อมูล เพื่อกำจัดข้อมูลที่ไม่เกี่ยวข้อง เช่น โฆษณา, แบนเนอร์, ลิงก์ที่ไม่เกี่ยวข้อง และข้อมูลเสียง หรือวิดีโอ ทำให้ข้อมูลมีความสอดคล้องกันและเหมาะสำหรับการนำเข้า AI ขั้นตอนนี้ใช้เครื่องมือทาง NLP (Natural Language Processing) เช่น Tokenization และ Text Normalization เพื่อให้ข้อมูลถูกแปลงเป็นรูปแบบที่เหมาะสมสำหรับการประมวลผล

- **การใช้ AI สำหรับการจำแนกประเภทข่าวสาร** ข้อมูลข่าวสารจะถูกส่งไปยังโมเดลการจำแนกประเภทข่าว ซึ่งเป็นกระบวนการสำคัญในการกรองข้อมูลให้เหมาะสมกับเป้าหมาย โดยในงานวิจัยนี้ได้นำ Naive Bayes Classifier มาใช้เป็นเครื่องมือในการจำแนกประเภทของข่าวสารระหว่าง "ข่าวทั่วไป" และ "ข่าวเหตุการณ์ความไม่สงบ" เนื่องจากโมเดล Naive Bayes มีความเหมาะสมสำหรับการประมวลผลข้อมูลข้อความที่มีความซับซ้อนต่ำและสามารถให้ผลลัพธ์ที่แม่นยำ ทั้งนี้หากพบข้อมูลที่มีความซับซ้อนสูง ระบบจะพิจารณาจากค่าความเชื่อมั่นของโมเดล และส่งต่อให้เจ้าหน้าที่ตรวจสอบก่อนยืนยันผลการจำแนกประเภท พร้อมบันทึกผลการทบทวนกลับมาเป็นข้อมูลฝึก เพื่อปรับปรุงโมเดลอย่างต่อเนื่องและลดความคลาดเคลื่อน กระบวนการทำงานมี ดังนี้:

○ **การเตรียมชุดข้อมูลสำหรับการฝึกโมเดล** ชุดข้อมูลตัวอย่างถูกเตรียมไว้โดยการจัดหมวดหมู่ข่าวเป็น "ข่าวทั่วไป" และ "ข่าวเหตุการณ์ความไม่สงบ" พร้อมกับการทำความสะอาดข้อมูล เช่น การลบคำซ้ำ การตัดคำที่ไม่จำเป็น และการแปลงข้อความให้อยู่ในรูปแบบที่เหมาะสม เช่น การใช้ Tokenization และการทำ Stemming เพื่อให้ข้อมูลพร้อมสำหรับการนำเข้าโมเดล

○ **การฝึกโมเดล Naive Bayes** จะดำเนินการด้วยการใช้ชุดข้อมูลที่เตรียมไว้ โดยอาศัยคุณสมบัติข้อความ (Text Features) เช่น คำสำคัญ (Keywords) และความถี่ของคำ (Word Frequency) เพื่อคำนวณความน่าจะเป็นของข้อความในแต่ละประเภท อัลกอริทึม Naive Bayes จะพิจารณาความสัมพันธ์ระหว่างคำและหมวดหมู่เพื่อสร้างแบบจำลองที่สามารถทำนายประเภทข่าวได้อย่างมีประสิทธิภาพ

○ **การทำนายประเภทข่าว** เมื่อได้รับข้อมูลข่าวใหม่ โมเดล Naive Bayes จะวิเคราะห์ข้อความโดยพิจารณาบริบทโดยรวมร่วมกับคำสำคัญ แทนการตัดสินจากคำเดียวโดด ๆ ระบบจะเปรียบเทียบคำที่มักพบในข่าว "เหตุการณ์ความไม่สงบ" เช่น "ระเบิด" "ยิงปะทะ" "ความรุนแรง" กับคำในข่าว "ข่าวทั่วไป" เช่น "การพัฒนา" "เทศกาล" "งานประชุม" จากนั้นคำนวณความน่าจะเป็นของแต่ละคลาสเพื่อจัดประเภทให้เหมาะสม ตอนที่พบคำเสี่ยงสูงอย่าง "ระเบิด" แต่บริบทส่วนใหญ่เกี่ยวกับการประชุมหรือบริหารจัดการ ระบบอาจจัดประเภทเป็น "ข่าวทั่วไป" เว้นแต่จะมีหลักฐานเชิงเหตุการณ์ที่ชัดเจน เช่น "เกิดเหตุ" "บาดเจ็บ" สำหรับกรณีที่คะแนนการจำแนกก้ำกึ่งหรือไม่มั่นใจ ระบบจะตั้งสถานะตรวจทานและส่งต่อให้เจ้าหน้าที่ตรวจสอบก่อนยืนยันผลการจัดประเภท พร้อมทั้งบันทึกผลการทบทวนกลับสู่ระบบเพื่อปรับปรุงเกณฑ์และน้ำหนักคุณลักษณะในรูปแบบการเรียนรู้ต่อเนื่อง ทำให้การจำแนกประเภทมีความแม่นยำเพิ่มขึ้นในอนาคต

○ **การปรับปรุงความแม่นยำของโมเดล** ระบบมีการปรับปรุงความแม่นยำของโมเดลโดยการตรวจสอบผลการจำแนกโดยผู้เชี่ยวชาญ และนำข้อผิดพลาดที่พบกลับมาเป็นข้อมูลเพื่อปรับปรุงและฝึกโมเดลใหม่ด้วยเทคนิค Feedback Loop ด้วยเทคนิค Naive Bayes Classifier นี้ ทำให้ระบบสามารถจำแนกประเภทข่าวสารได้อย่างรวดเร็วและมีประสิทธิภาพ ช่วยลดภาระของเจ้าหน้าที่ในการตรวจสอบข่าวสารจำนวนมาก (Slamet et al., 2018)

ขั้นตอนที่ 3: การบูรณาการข้อมูลจากแหล่งข่าวต่าง ๆ

การบูรณาการข้อมูลจากแหล่งข่าวต่าง ๆ ให้เป็นระบบเดียวกัน เป็นขั้นตอนสำคัญที่ช่วยให้ฐานข้อมูลมีคุณภาพสูงและสามารถตอบสนองความต้องการของผู้ใช้งานในการเข้าถึงข้อมูลได้อย่างแม่นยำและรวดเร็ว การดำเนินการบูรณาการข้อมูล จึงประกอบด้วยการทำงานหลายขั้นตอนย่อย ดังนี้

- **การสร้างโครงสร้างฐานข้อมูลที่มีมาตรฐานกลาง (Unified Data Schema)** ก่อนที่จะบูรณาการข้อมูลที่ดึงมาจากแหล่งข่าวต่าง ๆ จำเป็นต้องออกแบบโครงสร้างฐานข้อมูลกลางที่สามารถรองรับและจัดเก็บข้อมูลจากแหล่งต่าง ๆ ได้โดยไม่เกิดความขัดแย้ง โครงสร้างนี้กำหนดมาตรฐานกลาง เช่น การใช้ชื่อฟิลด์และประเภทข้อมูลที่สอดคล้องกันสำหรับข้อมูลประเภทต่าง ๆ เช่น หัวข้อข่าว วันที่เผยแพร่ ผู้เขียน ประเภทข่าว หรือเนื้อหาหลักของข่าว (Content) มาตรฐานกลางนี้ช่วยให้ข้อมูลจากแหล่งต่าง ๆ สามารถถูกนำมารวมและจัดเก็บได้อย่างเป็นระบบ โดยลดปัญหาความไม่สอดคล้องของข้อมูล
- **การแปลงข้อมูล (Data Transformation)** เนื่องจากข้อมูลที่ดึงมาแต่ละแหล่งมีโครงสร้างและรูปแบบที่ต่างกัน หรือใช้ชื่อฟิลด์ที่ต่างกัน การแปลงข้อมูลจึงมีบทบาทสำคัญในการปรับข้อมูลให้เข้ากับโครงสร้างฐานข้อมูลกลาง กระบวนการนี้จะทำการแปลงรูปแบบวันที่ เปลี่ยนประเภทข้อมูลให้สอดคล้องกัน และปรับชื่อฟิลด์ให้เป็นมาตรฐาน ให้เข้ากับฐานข้อมูลกลาง รวมถึงการใช้เทคนิคการทำความสะอาดข้อมูล (Data Cleaning) เพื่อลบข้อมูลซ้ำและแก้ไขข้อมูลที่ผิดพลาด

ขั้นตอนที่ 4: การออกแบบและจัดเก็บข้อมูลในฐานข้อมูล

การออกแบบและจัดเก็บข้อมูลในฐานข้อมูลมีความสำคัญอย่างยิ่ง เนื่องจากระบบ Web Scraping จะดึงข้อมูลจำนวนมากจากแหล่งข่าวออนไลน์ที่หลากหลายอย่างต่อเนื่อง จำเป็นต้องมีการวางโครงสร้างที่รองรับทั้งปริมาณและความ

ซับซ้อนของข้อมูลจาก Web Scraping เพื่อให้ฐานข้อมูลมีประสิทธิภาพในการจัดเก็บข้อมูลจำนวนมาก และสามารถรองรับการเข้าถึงข้อมูลที่สะดวก รวดเร็ว และเป็นปัจจุบัน

4. การวิเคราะห์ข้อมูล

การวิเคราะห์ข้อมูลในงานวิจัยนี้มีวัตถุประสงค์เพื่อประเมินประสิทธิภาพของระบบการสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การสร้างฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ โดยอาศัยผลลัพธ์จากการวัดผลในสองมิติหลัก ได้แก่ มิติเชิงคุณภาพ และมิติเชิงปริมาณ การดำเนินการวิเคราะห์มีรายละเอียด ดังต่อไปนี้

4.1 การวิเคราะห์ข้อมูลเชิงคุณภาพ (Qualitative Data Analysis)

การวิเคราะห์ข้อมูลเชิงคุณภาพมุ่งเน้นการประเมินประสิทธิภาพและความเหมาะสมของระบบจากมุมมองของผู้เชี่ยวชาญ โดยมีขั้นตอนการดำเนินงาน ดังนี้:

- **กลุ่มเป้าหมายและเครื่องมือ:** ดำเนินการเก็บรวบรวมข้อมูลจากแบบประเมินความคิดเห็นของกลุ่มผู้เชี่ยวชาญด้านการพัฒนาระบบสารสนเทศและเทคโนโลยีสารสนเทศ จำนวน 5 ท่าน
- **เกณฑ์การวิเคราะห์:** พิจารณาผลการประเมินครอบคลุม 7 ประเด็นสำคัญ ได้แก่ ประสิทธิภาพในการรวบรวมข้อมูล, ความแม่นยำในการประมวลผล, ความรวดเร็ว, การใช้งาน AI, ความง่ายในการใช้งาน, ความปลอดภัย และการรองรับการพัฒนาต่อเนื่อง
- **สถิติที่ใช้ในการวิเคราะห์:** ข้อมูลที่ได้จะถูกนำมาวิเคราะห์ทางสถิติด้วยการหาค่าเฉลี่ย (Mean: $\bar{X}$) และส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation: SD) เพื่อสรุประดับความพึงพอใจและประสิทธิภาพโดยรวมของระบบตามเกณฑ์ที่กำหนด

4.2 การวิเคราะห์ข้อมูลเชิงปริมาณ (Quantitative Data Analysis)

การวิเคราะห์ข้อมูลเชิงปริมาณมีวัตถุประสงค์เพื่อวัดสมรรถนะทางเทคนิคของระบบ (System Performance) ผ่านตัวชี้วัดประสิทธิภาพหลัก (KPIs) โดยมีกระบวนการ ดังนี้:

- **การเตรียมชุดข้อมูลทดสอบ (Test Dataset):** ใช้ชุดข้อมูลข่าวสารที่รวบรวมจากแหล่งข่าวเป้าหมายมากกว่า 15 แหล่ง โดยกำหนดขนาดกลุ่มตัวอย่าง จำนวน 100 ข่าว (N=100) กระบวนการคัดเลือกกลุ่มตัวอย่างใช้วิธีการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling) ตามสัดส่วนของหมวดหมู่ข่าวและช่วงเวลา เพื่อให้มั่นใจว่าข้อมูลทดสอบเป็นตัวแทนที่ดีและปราศจากความเอนเอียง (Bias) ของประชากรข่าวทั้งหมด
- **ตัวชี้วัดและการคำนวณ:** ข้อมูลจากชุดทดสอบจะถูกนำมาประมวลผลเพื่อวัดค่าตามตัวชี้วัดประสิทธิภาพหลัก 5 ด้าน ได้แก่ ความถูกต้องในการดึงข้อมูล (Data Extraction Accuracy), อัตราการตรวจจับข่าวซ้ำ (Duplicate Detection Rate: DDR), ความแม่นยำในการจำแนกประเภทข่าว (Categorization Accuracy), ความครอบคลุมของแหล่งข่าว (Source Coverage) และอัตราความผิดพลาด (Error Rate)
- **การแปลผลข้อมูล:** ผลลัพธ์จากการคำนวณจะถูกนำเสนอในรูปแบบร้อยละ (%) เพื่อสะท้อนถึงเสถียรภาพและความน่าเชื่อถือของระบบ ซึ่งความน่าเชื่อถือของผลลัพธ์เหล่านี้ได้รับการยืนยันผ่านความรัดกุมของกระบวนการสุ่มตัวอย่างแบบแบ่งชั้นที่ได้ดำเนินการในขั้นตอนการเตรียมข้อมูล

ผลการวิจัยและอภิปรายผลการวิจัย (Result and Discussion)

การวิจัยเรื่อง การสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การสร้างฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพของระบบ โดยผู้วิจัยขอนำเสนอผลการวิจัยแบ่งออกเป็น 3 ส่วนหลักตามลำดับการวิเคราะห์ ได้แก่ 1) ผลการประเมินประสิทธิภาพเชิงคุณภาพโดยผู้เชี่ยวชาญ 2) ผลการประเมิน

ตัวชี้วัดประสิทธิภาพหลัก (KPIs) ในเชิงปริมาณ และ 3) ผลการดำเนินงานของระบบในเชิงเทคนิคและการปฏิบัติงานจริง ดังรายละเอียดต่อไปนี้

1. ผลการประเมินเชิงคุณภาพโดยผู้เชี่ยวชาญ

ระบบฐานข้อมูลข่าวและเหตุการณ์จังหวัดชายแดนภาคใต้ที่พัฒนาขึ้น ได้รับการประเมินโดยผู้เชี่ยวชาญด้านการพัฒนาระบบสารสนเทศและเทคโนโลยีสารสนเทศ จำนวน 5 ท่าน เพื่อพิจารณาความเหมาะสมของการนำเทคนิค Web Scraping และ AI มาประยุกต์ใช้ ผลการประเมินพบว่า ในภาพรวมระบบมีประสิทธิภาพอยู่ในระดับดีมาก โดยมีค่าเฉลี่ยรวมอยู่ที่ 4.85 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 0.18 สิ่งนี้สะท้อนให้เห็นว่า ผู้เชี่ยวชาญมีความคิดเห็นที่สอดคล้องกันและเป็นไปในทิศทางบวกต่อระบบที่พัฒนาขึ้น

จุดแข็งที่โดดเด่นที่สุดของระบบปรากฏในด้านความแม่นยำในการประมวลผลข้อมูล และด้านความรวดเร็วในการประมวลผล โดยทั้งสองด้านได้รับค่าเฉลี่ยสูงสุด ($\bar{x}$ = 5.00, SD = 0.00) ซึ่งแสดงให้เห็นว่า ผู้เชี่ยวชาญยอมรับและมีความเห็นพ้องต้องกันในสมรรถนะของอัลกอริทึม AI ที่พัฒนาขึ้น โดยระบบสามารถทำงานได้อย่างถูกต้องและรวดเร็วอย่างมีนัยสำคัญ

ด้านความง่ายและความยืดหยุ่นในการใช้งาน ($\bar{x}$ = 4.67, SD = 0.42) และด้านการรองรับการพัฒนา ($\bar{x}$ = 4.84, SD = 0.24) แม้ว่าอยู่ในระดับดีมากแล้ว แต่มีค่าส่วนเบี่ยงเบนมาตรฐานที่สูงกว่าด้านอื่นเล็กน้อย สิ่งนี้สะท้อนให้เห็นว่ามีโอกาสและความจำเป็นในการปรับปรุงส่วนติดต่อผู้ใช้ (User Interface) ให้มีความเป็นมิตรมากขึ้น และตอบสนองต่อความต้องการที่หลากหลายของผู้ใช้งานต่าง ๆ เพื่อให้ระบบเข้าถึงผู้ใช้ได้อย่างแพร่หลายยิ่งขึ้นในอนาคต ทั้งนี้ ผู้เชี่ยวชาญยังกล่าวเสริมว่ากลไกการสกัดและรวบรวมข้อมูลนี้มีประสิทธิภาพสูงและเป็นประโยชน์อย่างมากต่อหน่วยงาน

ตารางที่ 2 แสดงตารางผลการประเมินประสิทธิภาพของระบบจากผู้เชี่ยวชาญ จำนวน 5 ท่าน

ประเด็นวัดประสิทธิภาพของระบบ	สรุปผลการประเมิน		
	$\bar{x}$	SD	แปลผล
1. ด้านประสิทธิภาพในการรวบรวมข้อมูล	4.83	0.23	**ดีมาก**
2. ด้านความแม่นยำในการประมวลผลข้อมูล	5.00	0.00	**ดีมาก**
3. ด้านความรวดเร็วและประสิทธิภาพในการประมวลผล	5.00	0.00	**ดีมาก**
4. ด้านการใช้งานของ AI ในการจำแนกและวิเคราะห์ข้อมูล	4.82	0.21	**ดีมาก**
5. ด้านความง่ายและความยืดหยุ่นในการใช้งาน	4.67	0.42	**ดีมาก**
6. ด้านความปลอดภัยและการปฏิบัติตามกฎหมาย	4.81	0.22	**ดีมาก**
7. ด้านการรองรับการพัฒนาต่อเนื่อง	4.84	0.24	**ดีมาก**
เฉลี่ยรวม	**4.85**	**0.18**	**ดีมาก**

2. ผลการประเมินเชิงปริมาณด้วยตัวชี้วัดประสิทธิภาพของระบบ (Key Performance Indicators: KPIs)

เพื่อยืนยันประสิทธิภาพของระบบในเชิงประจักษ์ ผู้วิจัยได้ดำเนินการทดสอบกับชุดข้อมูลกลุ่มตัวอย่าง จำนวน 100 ข่าว (N=100) ซึ่งได้จากการสุ่มตัวอย่างแบบแบ่งชั้นภูมิ (Stratified Random Sampling) เพื่อให้มั่นใจว่าผลลัพธ์เป็นตัวแทนที่ดีของข้อมูลทั้งหมด ผลการประเมินตามตัวชี้วัดประสิทธิภาพหลัก (KPIs) ทั้ง 5 ด้าน ปรากฏดังตารางที่ 3

ตารางที่ 3 แสดงผลการประเมินประสิทธิภาพของระบบ

ตัวชี้วัด	เกณฑ์การประเมิน	วิธีการวัดผล	ผลการประเมิน
ความถูกต้องในการดึงข้อมูล (Data Extraction Accuracy)	ข้อมูลที่ดึงต้อง **ตรงกับต้นฉบับทุกฟิลด์** (หัวข้อ/วันที่/แหล่งที่มา/เนื้อหา)	สุ่มข่าว **N=100** ข่าว เทียบค่าที่ระบบดึงกับต้นฉบับแบบ **ต่อข่าว** (Exact-Match) สูตร: Accuracy Article = (ข่าวถูกครบทุกฟิลด์/N) × 100%	**97%** (ถูก 97/100 ข่าว)
อัตราการตรวจจับข่าวซ้ำ (Duplicate Detection Rate: DDR)	ระบุ “**ชุดเหตุการณ์ซ้ำจริง**” ได้ถูกต้อง (วัดเป็น **ชุดเหตุการณ์** ไม่ใช่บทความเดี่ยว)	สร้าง Gold Standard โดยผู้เชี่ยวชาญ นับ **TP_dup, FN_dup** สูตร: DDR = TP_dup/(TP_dup + FN_dup) × 100%	**98%** (TP=98, FN=2 จาก 100 ชุด)
ความแม่นยำในการจำแนกประเภท (Categorization Accuracy)	ทำนายตรงกับป้ายกำกับผู้เชี่ยวชาญ (2 คลาส: เหตุการณ์ความไม่สงบ/ข่าวทั่วไป)	ชุดทดสอบ **100 ข่าว** โดยโมเดล **Naive Bayes**; สูตร: Accuracy = (TP+TN)/(TP+TN+FP+FN) × 100%	**92%** (จำแนกถูก 92/100)
ความครอบคลุมของแหล่งข่าว (Source Coverage)	ดึงได้ **ครอบคลุมบัญชีแหล่งข่าวเป้าหมาย** ทั้งระดับชาติและท้องถิ่น	นับจำนวนแหล่งข่าวที่ดึงได้จริง **S_จริง** เทียบกับ **S_เป้าหมาย**; สูตร: Coverage = S_จริง/S_เป้าหมาย × 100%	**>15 แหล่ง** (ชาติ+ท้องถิ่น)
อัตราความผิดพลาด	สัดส่วนข่าวที่ **เนื้อหาขาด/ผิดรูปแบบ/เมตาดาต้าผิด/ไฟล์เสียหาย**	สุ่มตรวจ **N=100** ข่าว; สูตร: Error Rate = E/N × 100%)	**2%** (พบผิดพลาด 2/100)

หมายเหตุ TP = True Positive, FN = False Negative (ศึกษาคำนิยามตัวชี้วัดฉบับสมบูรณ์ได้จากตารางที่ 1)

จากตารางที่ 3 ผลการประเมินสามารถวิเคราะห์นัยสำคัญของผลการทดสอบในแต่ละด้านได้ ดังนี้

ด้านความถูกต้องในการดึงข้อมูล (Data Extraction Accuracy) ผลลัพธ์ที่ร้อยละ 97 สะท้อนถึงเสถียรภาพของอัลกอริทึม Web Scraping ที่สามารถจัดการกับโครงสร้างเว็บไซต์ที่หลากหลายได้ดี

ด้านอัตราการตรวจจับข่าวซ้ำ (Duplicate Detection Rate) ระบบแสดงความสามารถที่โดดเด่นในการจัดกลุ่มเนื้อหา โดยพบกรณีผิดพลาดเพียง 2 กรณีจาก 100 ชุดเหตุการณ์ (ร้อยละ 98) ซึ่งหมายความว่า ระบบสามารถตรวจจับและกรองข่าวที่ซ้ำกันออกได้อย่างมีประสิทธิภาพ ช่วยให้ผู้ใช้งานไม่ต้องเห็นข้อมูลที่ซ้ำซ้อน

ด้านความแม่นยำในการจำแนกประเภท (Categorization Accuracy) โมเดล Naive Bayes พิสูจน์ให้เห็นถึงความเหมาะสมในการใช้งานกับบริบทข่าวสามจังหวัดชายแดนภาคใต้ โดยสามารถแยกแยะความแตกต่างระหว่าง "ข่าวทั่วไป" และ "เหตุการณ์ความไม่สงบ" ได้อย่างแม่นยำถึงร้อยละ 92 แม้จะมีคำกำกวมปรากฏในเนื้อหา เช่น “ไฟไหม้” “เจ้าหน้าที่” และ “ทหาร”

ด้านความครอบคลุมของแหล่งข่าว (Source Coverage) ระบบประสบความสำเร็จในการเข้าถึงแหล่งข่าวเป้าหมายได้ครบถ้วนทั้ง 15 แหล่ง คิดเป็นร้อยละ 100 ตามเกณฑ์ที่กำหนด จุดเด่นสำคัญ คือ ความสามารถในการดึงข้อมูลที่ครอบคลุมทั้งสื่อระดับชาติและสื่อท้องถิ่นในพื้นที่จังหวัดชายแดนภาคใต้ การบรรลุเป้าหมายนี้มีความสำคัญอย่างยิ่งต่อการสร้างฐานข้อมูลที่

สมบูรณ์ เนื่องจากระบบต้องจัดการกับโครงสร้างเว็บไซต์ (HTML Structure) ที่มีความหลากหลายและซับซ้อนของแต่ละแหล่งข่าว ซึ่งการครอบคลุมแหล่งข่าวท้องถิ่นช่วยให้มั่นใจได้ว่าเหตุการณ์ระดับพื้นที่ซึ่งอาจไม่ปรากฏในสื่อหลักจะถูกรวบรวมเข้าสู่ระบบอย่างไม่ตกหล่น

ด้านอัตราความผิดพลาด (Error Rate) ผลการทดสอบพบอัตราความผิดพลาดเพียงร้อยละ 2 ซึ่งส่วนใหญ่เกิดจากข้อมูลต้นทางไม่สมบูรณ์ เช่น เนื้อหาขาดหายหรือไฟล์เสียหาย อัตราความผิดพลาดขั้นต่ำนี้แสดงว่าระบบมีเสถียรภาพในการทำงานกับข้อมูล และกระบวนการทำความสะอาดข้อมูลทำงานได้อย่างมีประสิทธิภาพ ทำให้ระบบพร้อมสำหรับการใช้งานจริง และรักษาคุณภาพข้อมูลไว้ในระดับสูง

ผลลัพธ์ทั้งหมดนี้ยืนยันได้ว่า ระบบที่พัฒนาขึ้นสามารถบรรลุวัตถุประสงค์ของงานวิจัยอย่างสมบูรณ์ ทั้งในมิติของความถูกต้องทางข้อมูล และประสิทธิภาพเชิงปฏิบัติการ เพื่อสนับสนุนการตัดสินใจเชิงนโยบายต่อไป

3. ผลการดำเนินงานของระบบในเชิงเทคนิค (Technical Operational Performance)

จากการทดสอบระบบในสภาพแวดล้อมจริงแสดงให้เห็นว่า ระบบสามารถดึงข้อมูลข่าวจากแหล่งข่าวออนไลน์มากกว่า 15 แหล่ง ซึ่งครอบคลุมทั้งสำนักข่าวระดับชาติและสื่อท้องถิ่นในจังหวัดชายแดนภาคใต้ ระบบทำงานต่อเนื่องตลอด 24 ชั่วโมง และมีผลต่อการเปลี่ยนแปลงประสิทธิภาพการปฏิบัติงานที่มีนัยสำคัญในหลายมิติ ผลที่เห็นได้ชัดที่สุด คือ การลดเวลาในการจัดเก็บข้อมูลข่าว ซึ่งเดิมต้องใช้เวลา 1–2 ชั่วโมงต่อวัน แต่หลังจากการใช้ระบบแล้ว ลดเหลือเพียง 10–15 นาทีต่อวัน คิดเป็นการลดลงประมาณร้อยละ 85 การช่วยเหลือนี้ทำให้เจ้าหน้าที่สามารถเปลี่ยนบทบาทจากการป้อนข้อมูลไปยังงานที่มีมูลค่าสูงกว่า เช่น การตรวจสอบคุณภาพของข้อมูล การวิเคราะห์เนื้อหา และการให้บริการสารสนเทศแก่ผู้ใช้ได้มากขึ้น

ระบบมีกลไกการจัดการ “ชุดเหตุการณ์” ที่มีประสิทธิภาพสูง โดยพิจารณาความคล้ายคลึงของหลายองค์ประกอบ ได้แก่ หัวข้อ เนื้อหา เวลาที่เผยแพร่ และสถานที่ของเหตุการณ์ หลังจากนั้น ระบบจะกำหนด "ระเบียนหลัก" (Master Record) เพียงรายการเดียวสำหรับเหตุการณ์หนึ่ง ๆ ในขณะที่ข่าวจากแหล่งอื่นจะถูกบันทึกเป็นแหล่งอ้างอิงเพิ่มเติม ผลจากการจัดการนี้ทำให้ฐานข้อมูลมีความสะอาด (Clean Data) เนื่องจากลดความซ้ำซ้อน และในขณะเดียวกัน ยังรักษาความโปร่งใสในการตรวจสอบย้อนกลับว่าข้อมูลมาจากแหล่งใดบ้าง

ตารางที่ 4 แสดงการเปรียบเทียบประสิทธิภาพการปฏิบัติงานระหว่างระบบเดิมและระบบใหม่

ด้านการปฏิบัติงาน (Operational Aspect)	**ระบบเดิม (Manual)**	**ระบบใหม่ (Automated with AI)**
ระยะเวลาในการรวบรวมข้อมูล/วัน	1–2 ชั่วโมง	10–15 นาที (ลดลงประมาณ 85%)
กระบวนการจำแนกประเภทข่าว	เจ้าหน้าที่คัดแยกด้วยตนเอง	AI จำแนกอัตโนมัติ (ความแม่นยำ 92%)
การจัดการข้อมูลซ้ำ	พบความซ้ำซ้อนและ Human Error	ตรวจจับและป้องกันข่าวซ้ำได้ 98%
การจัดเก็บและสืบค้น	ไม่เป็นหมวดหมู่ ค้นหายาก	จัดเก็บเป็นระเบียบ ค้นหาง่ายตามโครงสร้างมาตรฐาน

อภิปรายผลการวิจัย (Discussion)

ผลการวิจัยชี้ให้เห็นว่า การพัฒนาระบบสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สามารถยกระดับประสิทธิภาพการบริหารจัดการข้อมูลข่าวสารในจังหวัดชายแดนภาคใต้ได้อย่างมีนัยสำคัญ โดยประเด็นสำคัญจากการเปรียบเทียบประสิทธิภาพของโมเดลและการดำเนินงาน มีดังนี้

ในด้านประสิทธิภาพการจำแนกประเภทข่าวสาร โมเดล Naive Bayes ที่พัฒนาขึ้นมีความแม่นยำร้อยละ 92 ซึ่งเมื่อเปรียบเทียบกับเทคโนโลยีระดับสูงในปัจจุบันพบว่า เป็นระดับที่ยอมรับได้และคุ้มค่าเชิงปฏิบัติการ แม้งานวิจัยของ Minaee et al.

(2021) ที่ทบทวนวรรณกรรมเกี่ยวกับโมเดล Deep Learning สำหรับการจำแนกข้อความ จะชี้ให้เห็นว่าโมเดลขั้นสูง เช่น BERT และ LSTM สามารถทำความแม่นยำได้ถึงร้อยละ 95-98 ในชุดข้อมูลข่าวมาตรฐาน แต่โมเดลเหล่านี้ต้องแลกด้วยต้นทุนทางคอมพิวเตอร์ที่สูงและระยะเวลาประมวลผลที่นาน สอดคล้องกับงานศึกษาของ Li et al. (2022) และ Zhang (2004) ที่ระบุว่าแม้ Naive Bayes จะมีข้อจำกัดเรื่องข้อสมมติความเป็นอิสระของตัวแปรซึ่งอาจส่งผลต่อความสามารถในการจับบริบทที่ซับซ้อน แต่กลับมีจุดเด่นด้านประสิทธิภาพการใช้ทรัพยากรและความรวดเร็วในการฝึกฝนโมเดลที่เหนือกว่า Deep Learning หลายเท่า ในบริบทของงานวิจัยนี้ที่มุ่งเน้นการใช้งานจริงในหอสมุดซึ่งมีข้อจำกัดด้านทรัพยากรเซิร์ฟเวอร์และต้องการความรวดเร็วในการประมวลผลข้อมูลรายวัน การเลือกใช้ Naive Bayes ที่มีความแม่นยำร้อยละ 92 จึงเป็นจุดสมดุลที่เหมาะสมระหว่างความถูกต้องและความคล่องตัวในการปฏิบัติงาน

นอกจากนี้ ในมิติของการปฏิบัติงาน ระบบสามารถลดภาระงานของเจ้าหน้าที่ได้ถึงร้อยละ 85 โดยลดเวลาทำงานจาก 1–2 ชั่วโมงเหลือเพียง 10–15 นาที ซึ่งสอดคล้องกับผลการศึกษาของ Mitchell (2018) ที่พบว่า Web Scraping เป็นเครื่องมือสำคัญในการจัดการ Big Data บนอินเทอร์เน็ต เมื่อผนวกกับความสามารถในการตรวจจับข่าวซ้ำของระบบที่ทำได้ถึงร้อยละ 98 ซึ่งสูงกว่าการตรวจสอบด้วยมือ ช่วยยืนยันว่าการนำระบบอัตโนมัติมาใช้ไม่เพียงเพิ่มความรวดเร็ว แต่ยังช่วยลดความผิดพลาดจากมนุษย์ได้อย่างมีประสิทธิภาพตามแนวทางที่ Pant et al. (2024) ได้เสนอแนะ

ข้อจำกัดของการวิจัย (Limitations)

แม้งานวิจัยนี้จะบรรลุวัตถุประสงค์หลัก แต่ยังมีข้อจำกัดที่ควรพิจารณาเพื่อการพัฒนาต่อยอดในอนาคต ประการแรก คือ ขอบเขตของแหล่งข้อมูลซึ่งจำกัดเฉพาะเว็บไซต์ข่าวออนไลน์ จำนวน 15 แหล่ง ยังไม่ครอบคลุมถึงแพลตฟอร์มโซเชียลมีเดียหรือสื่อมัลติมีเดียอื่น ๆ ซึ่งอาจทำให้ขาดมิติของข้อมูลข่าวสารที่หมุนเวียนในช่องทางที่ไม่เป็นทางการ ประการต่อมา คือ ความละเอียดในการจำแนกประเภทข่าวยังคงเป็นแบบทวิภาคระหว่างข่าวเหตุการณ์ความไม่สงบและข่าวทั่วไป ซึ่งอาจไม่เพียงพอสำหรับการวิเคราะห์เชิงลึกที่ต้องการความเฉพาะเจาะจง เช่น ข่าวเศรษฐกิจ หรือข่าวการศึกษา นอกจากนี้ การเลือกใช้โมเดล Naive Bayes แม้จะประมวลผลได้รวดเร็ว แต่ในกรณีที่ข้อความมีความซับซ้อนสูงหรือมีความกำกวมทางภาษา โมเดลอาจให้ผลลัพธ์ที่คลาดเคลื่อนเมื่อเทียบกับโมเดลที่ซับซ้อนกว่า รวมถึงขนาดตัวอย่างในการประเมินผล KPI ที่ใช้การสุ่มตัวอย่างข่าวจำนวน 100 รายการ ซึ่งแม้วิธีการสุ่มตัวอย่างแบบแบ่งชั้นจะช่วยลดความคลาดเคลื่อนได้ แต่ปริมาณดังกล่าวอาจน้อยเกินไปเมื่อเทียบกับปริมาณข่าวสะสมในระยะยาว ซึ่งอาจส่งผลต่อความเชื่อมั่นทางสถิติในบางมิติ

ข้อเสนอแนะ (Recommendations)

เพื่อให้ระบบเกิดประโยชน์สูงสุดและรองรับการพัฒนาในอนาคต ผู้วิจัยมีข้อเสนอแนะทั้งในเชิงปฏิบัติและเชิงวิชาการ สำหรับการนำไปใช้จริง หน่วยงานความมั่นคงและสถาบันการศึกษาสามารถนำต้นแบบระบบนี้ไปประยุกต์ใช้เพื่อสร้างคลังข้อมูลข่าวสารเฉพาะด้าน โดยควรให้ความสำคัญกับการปฏิบัติตามพระราชบัญญัติคุ้มครองข้อมูลส่วนบุคคล (PDPA) อย่างเคร่งครัด ควบคู่ไปกับการต่อยอดระบบให้มีฟังก์ชันการแสดงผลข้อมูลเชิงทัศนภาพ หรือ Dashboard เพื่อสรุปแนวโน้มสถานการณ์รายวัน ซึ่งจะช่วยสนับสนุนการตัดสินใจของผู้บริหารได้อย่างรวดเร็ว สำหรับการวิจัยในอนาคต ควรศึกษาเปรียบเทียบประสิทธิภาพของการใช้โมเดล Deep Learning ในการจำแนกหมวดหมู่ข่าวที่ละเอียดและซับซ้อนยิ่งขึ้น เพื่อเพิ่มความลึกซึ้งในการวิเคราะห์เนื้อหา รวมถึงควรพัฒนาระบบ Web Scraping ให้รองรับการดึงข้อมูลจากแพลตฟอร์มที่หลากหลายขึ้น พร้อมทั้งศึกษาความพึงพอใจและประสบการณ์การใช้งานจากกลุ่มผู้ใช้จริงที่หลากหลาย เช่น นักวิจัย และเจ้าหน้าที่ภาครัฐ เพื่อนำผลป้อนกลับมาปรับปรุงระบบให้ตอบโจทย์การใช้งานจริงได้อย่างรอบด้าน

สรุปผลการวิจัย (Conclusion)

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อพัฒนาและประเมินประสิทธิภาพของระบบสกัดข้อมูลอัตโนมัติด้วยเทคนิค Web Scraping และ AI สู่การสร้างฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ของประเทศไทย โดยมุ่งแก้ไขปัญหากระบวนการจัดเก็บข่าวแบบเดิมที่ใช้เวลานาน มีความเสี่ยงต่อความผิดพลาด และเกิดข้อมูลซ้ำจำนวนมากในฐานข้อมูลข่าวของหอสมุดจอห์น เอฟ เคนเนดี้ สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี

ผลการดำเนินการวิจัยแสดงให้เห็นว่า ระบบที่พัฒนาขึ้นสามารถเปลี่ยนกระบวนการจัดเก็บข่าวจากการทำงานด้วยมือมาเป็นกระบวนการอัตโนมัติที่มีประสิทธิภาพสูง โดยลดระยะเวลาในการจัดเก็บข้อมูลจากเดิมประมาณ 1–2 ชั่วโมงต่อวัน เหลือเพียง 10–15 นาทีต่อวัน คิดเป็นการลดลงประมาณร้อยละ 85 ขณะเดียวกันยังคงรักษาคุณภาพข้อมูลในระดับสูง โดยมีความถูกต้องในการดึงข้อมูล (Data Extraction Accuracy) ร้อยละ 97 ความสามารถในการตรวจจับข่าวซ้ำ (Duplicate Detection Rate) ร้อยละ 98 และความแม่นยำในการจำแนกประเภทข่าว ร้อยละ 92 พร้อมทั้งลดอัตราความผิดพลาดในการประมวลผลเหลือเพียงร้อยละ 2

ในมุมมองของผู้เชี่ยวชาญด้านระบบสารสนเทศ ระบบได้รับการประเมินว่า อยู่ในระดับดีมากในทุกด้าน โดยมีค่าเฉลี่ยรวม ($\bar{x}$ = 4.85 และ SD = 0.18) โดยเฉพาะด้านความแม่นยำและความรวดเร็วในการประมวลผลที่ได้รับคะแนนเต็ม ($\bar{x}$ = 5.00, SD = 0.00) สะท้อนว่า ระบบที่พัฒนาขึ้นไม่เพียงแต่ตอบโจทย์ด้านเทคนิค แต่ยังมีความเหมาะสมต่อการนำไปใช้งานจริงในบริบทของหน่วยงานสารสนเทศ

ระบบสกัดข้อมูลอัตโนมัติและฐานข้อมูลข่าวที่พัฒนาขึ้นในงานวิจัยนี้ สามารถทำหน้าที่เป็นโครงสร้างพื้นฐานข้อมูลสำหรับการติดตามสถานการณ์และวิเคราะห์ข่าวในจังหวัดชายแดนภาคใต้ได้อย่างมีประสิทธิภาพ ช่วยลดภาระงานของเจ้าหน้าที่ เพิ่มความถูกต้องและความเป็นระเบียบของข้อมูล และสนับสนุนการวางแผนนโยบายและการตัดสินใจเชิงยุทธศาสตร์ของหน่วยงานที่เกี่ยวข้องได้อย่างเป็นรูปธรรม ทั้งยังมีศักยภาพสูงในการต่อยอดไปสู่การบูรณาการกับระบบสืบค้นแดชบอร์ดเชิงวิเคราะห์ และเทคโนโลยี AI ขั้นสูงในอนาคต ตามแนวทางข้อเสนอแนะของงานวิจัยนี้

กิตติกรรมประกาศ (Acknowledgements)

ผู้วิจัยขอขอบพระคุณผู้บริหาร บุคลากร หอสมุดจอห์น เอฟ เคนเนดี้ สำนักวิทยบริการ มหาวิทยาลัยสงขลานครินทร์ วิทยาเขตปัตตานี ที่ให้การสนับสนุนและอำนวยความสะดวกในการดำเนินการวิจัย และโอกาสในการนำระบบสกัดข้อมูลอัตโนมัติไปทดสอบการใช้งานจริง เพื่อพัฒนาระบบฐานข้อมูลข่าวและเหตุการณ์ในจังหวัดชายแดนภาคใต้ให้มีประสิทธิภาพ

ขอขอบพระคุณ ผู้เชี่ยวชาญด้านการพัฒนาระบบสารสนเทศและเทคโนโลยีสารสนเทศ ทั้ง 5 ท่าน ที่กรุณาสละเวลาในการตรวจสอบเครื่องมือและประเมินประสิทธิภาพของระบบ พร้อมทั้งให้ข้อเสนอแนะในการปรับปรุงแก้ไข จนทำให้งานวิจัยนี้มีความถูกต้องและสมบูรณ์ตามหลักวิชาการ

คุณค่าและประโยชน์อันพึงมีจากงานวิจัยฉบับนี้ ผู้วิจัยขอมอบเพื่อเป็นวิทยาทานแก่ผู้ที่สนใจศึกษา และเพื่อเป็นประโยชน์ต่อการพัฒนาจังหวัดชายแดนภาคใต้สืบไป

รายการอ้างอิง (References)

จักรินทร์ สันติรัตนภักดี. (2565). กระบวนการสกัดข้อมูลรายงานอุบัติเหตุทางถนนรายใหญ่และความสามารถในการนำเสนอสารสนเทศด้วยภาพข้อมูลผ่านเว็บไซต์. *วารสารศรีนครินทรวิโรฒวิจัยและพัฒนา (สาขามนุษยศาสตร์และสังคมศาสตร์), 14*(27), 14-34. https://so04.tci-thaijo.org/index.php/swurd/article/view/259751

ศตวรรษ รามไชย และ ผุสดี พรผล. (2565). การเตรียมข้อมูลจากเว็บในอุตสาหกรรมการท่องเที่ยว: กรณีที่พักในจังหวัดภูเก็ต. ใน *การประชุมวิชาการระดับชาติ ด้านวิทยาศาสตร์และเทคโนโลยี เครือข่ายสถาบันอุดมศึกษาภาคใต้ ครั้งที่ 7* (น.1-10). ฐานข้อมูลวิจัย สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏภูเก็ต.

Bhatt, C., Bisht, A., Chauhan, R., Vishvakarma, A., Kumar, M., & Sharma, S. (2023). Web scraping techniques and its applications: A review [Conference session]. In *2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT)* (pp. 1-8). https://doi.org/10.1109/cisct57197.2023.10351298

Bhujbal, M., Bibawanekar, B., & Deshmukh, P. (2023). News aggregation using web scraping news portals. *International Journal of Advanced Research in Science, Communication and Technology*, *3*(2), 275-284. https://doi.org/10.48175/IJARSCT-12138

Farias, W. A. S., Melo, D. M. A., Santos, L. M. dos, de Oliveira, Â. A. S., Medeiros, R. L. B. A., & Silva, Y. K. R. O. (2024). *Web scraping as a scientific tool for theoretical reference*. https://doi.org/10.21203/rs.3.rs-3854342/v1

Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., & He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *13*(2), 1–54. https://doi.org/10.1145/3495162

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys (CSUR)*, *54*(3), 1–40. https://doi.org/10.1145/3439726

Mitchell, R. (2018). *Web scraping with Python: Collecting data from the modern web* (2nd ed.). O'Reilly Media.

Pant, S., Yadav, E. N., Milan, Sharma, M., Bedi, Y., & Raturi, A. (2024). Web scraping using beautiful soup [Conference session]. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1-6). Chikkaballapur, India. https://doi.org/10.1109/ickecs61492.2024.10617017

Slamet, C., Andrian, R., Maylawati, D. S., Suhendar, Darmalaksana, W., & Ramdhani, M. A. (2018). *Web scraping and Naïve Bayes Classification for job search engine.* 288(1):012038-. https://doi.org/10.1088/1757-899X/288/1/012038

Valova, I., Mladenova, T., Kanev, G., & Halacheva, T. (2023). Web scraping - state of art, techniques and approaches [Conference session]. In *2023 31st National Conference with International Participation (TELECOM)* (pp. 1-4). Sofia, Bulgaria. https://doi.org/10.1109/telecom59629.2023.10409723

Zhang, H. (2004). The optimality of naive Bayes. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)* (pp. 562–567). AAAI Press.