วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์ ปีที่ 4 ฉบับที่ 5 กันยายน – ตุลาคม 2569
Journal of Engineering and Industrial Technology, Kalasin University, Vol. 4 No. 5 September – October 2026

บทความวิจัย (Research Article)

การตรวจสอบความสอดคล้องของข้อมูลระดับฟิลด์ในเอกสารเรือพาณิชย์ ด้วยกรอบการทำงานแบบหลายขั้นตอน

ณรงค์ศักดิ์ ยินดีสุข[1], ชัยพร เขมะภาตะพันธ์[1,*], รัฐศิลป์ รานอกภานุวัชร์[1] และ อ้อมเดือน เจียมอ่อน[1]

[1] สาขาวิชาวิศวกรรมคอมพิวเตอร์ วิทยาลัยวิศวกรรมศาสตร์และเทคโนโลยี มหาวิทยาลัยธุรกิจบัณฑิตย์ กรุงเทพฯ ประเทศไทย

*ผู้ประสานงานบทความต้นฉบับ: chaiyaporn@dpu.ac.th, โทรศัพท์: 089-1731525
(รับบทความ: 17 พฤษภาคม 2569; แก้ไขบทความ: 7 สิงหาคม 2569; ตอบรับบทความ: 10 สิงหาคม 2569)

บทคัดย่อ

ปัจจุบันการตรวจสอบความสอดคล้องของเอกสารเรือพาณิชย์ 2 ชุดที่มีเนื้อหาเกี่ยวข้องกันแต่จัดทำโดยผู้ปฏิบัติงานต่างคนกัน จำเป็นต้องใช้เจ้าหน้าที่เป็นผู้ตรวจพิสูจน์ ซึ่งค่อนข้างใช้เวลาในการตรวจพิสูจน์เป็นอย่างมาก ดังนั้นงานวิจัยนี้จึงมุ่งเน้นออกแบบและพัฒนากระบวนการเปรียบเทียบความสอดคล้องของเอกสาร 2 ชุดที่มีความเกี่ยวพันกันโดยตรงด้วยกรอบการทำงานแบบหลายขั้นตอนที่ประกอบไปด้วยเทคนิคเชิงกฎเกณฑ์ที่เสริมด้วยการประมวลผลภาษาธรรมชาติในสภาพแวดล้อมแบบออฟไลน์เพื่อเพิ่มประสิทธิภาพและความรวดเร็ว รวมทั้งป้องกันข้อมูลรั่วไหลด้วย โดยกระบวนการตรวจสอบมี 4 ขั้นตอนคือ 1) การจับคู่แบบตรงกันทุกประการ (Exact Matching) 2) การจับคู่แบบคลุมเครือ (Fuzzy Matching) 3) ความคล้ายคลึงเชิงความหมาย (Semantic Similarity) และ 4) การอนุมานภาษาธรรมชาติ (Natural Language Inference) การทดลองในงานวิจัยนี้ดำเนินการบนชุดข้อมูล 4 ประเภท จำนวนรวม 2,000 ไฟล์ โดยแยกเป็น เอกสารรายงาน (.docx) จำนวน 500 ไฟล์ เอกสารอ้างอิงเชิงข้อความ (Text-based .pdf) จำนวน 500 ไฟล์ เอกสารอ้างอิงเชิงรูปภาพ (Image-based .pdf) จำนวน 500 ไฟล์ และป้ายกำกับข้อมูลที่ถูกต้อง (Gold Label .csv) ของ 10 ฟิลด์เป้าหมายจำนวน 500 ไฟล์ เพื่อใช้ในการประเมินประสิทธิภาพของเครื่องมือด้วยตัวชี้วัดที่หลากหลาย ผลบนชุดข้อมูลเดียวกับที่ใช้ปรับค่าเกณฑ์แสดงให้เห็นว่าอัตราการผ่านเกณฑ์ (Pass rate) เพิ่มขึ้นโดยเอกสารอ้างอิงเชิงข้อความเพิ่มขึ้นจาก 87.90% เป็น 96.50% และเอกสารอ้างอิงเชิงรูปภาพเพิ่มขึ้นจาก 72.40% เป็น 85.10% ในขณะที่คะแนนเอฟวัน แบบไบนารี (Binary F1-score) สูงสุดอยู่ที่ 0.982 และ 0.920 ตามลำดับ

คำสำคัญ: การตรวจสอบแบบหลายขั้นตอน การจับคู่แบบคลุมเครือ ความคล้ายคลึงเชิงความหมาย การอนุมานภาษาธรรมชาติ

การอ้างอิงบทความ: ณรงค์ศักดิ์ ยินดีสุข, ชัยพร เขมะภาตะพันธ์, รัฐศิลป์ รานอกภานุวัชร์ และอ้อมเดือน เจียมอ่อน, "การตรวจสอบความสอดคล้องของข้อมูลระดับฟิลด์ในเอกสารเรือพาณิชย์ด้วยกรอบการทำงานแบบหลายขั้นตอน,“*วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์*, ปีที่ 4, ฉบับที่ 5, หน้า 83-96, 2569.

วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์ ปีที่ 4 ฉบับที่ 5 กันยายน – ตุลาคม 2569
Journal of Engineering and Industrial Technology, Kalasin University, Vol. 4 No. 5 September – October 2026

บทความวิจัย (Research Article)

Field-Level Data Consistency Verification in Merchant Vessel Documents Using a Multi-Stage Framework

Narongsak Yindeesuk[1], Chaiyaporn Khemapatapan[1]*, Ratthaslip Ranokphanuwat[1] and Aomduan Jeamaon[1]

[1] Department of Computer Engineering, College of Engineering and Technology, Dhurakij Pundit University, Bangkok, Thailand

* Corresponding Author: chaiyaporn@dpu.ac.th, Tel: 089-1731525
(Received: May 17, 2026; Revised: August 7, 2026; Accepted: August 10, 2026)

Abstract

At present, the verification of consistency between two merchant vessel documents that address the same subject matter and are prepared or processed by different operators remains dependent on manual checks and is considerably time-consuming. Accordingly, this study aims to design and develop a document consistency comparison framework for two interrelated document sets by a multi-stage process consisting of rule-based techniques assisted by Natural Language Processing in an offline environment. The proposed framework is intended to enhance verification efficiency and processing speed while mitigating the risk of data leakage. The verification procedure comprises four sequential stages: 1) Exact Matching, 2) Fuzzy Matching, 3) Semantic Similarity, and 4) Natural Language Inference. The experiment was conducted using four categories of data comprising 2,000 files in total: 500 report files (.docx), 500 text-based reference files (.pdf), 500 image-based reference files (.pdf), and 500 gold-label files (.csv) containing ground-truth annotations for 10 target fields. These datasets were employed to assess the performance of the proposed framework using multiple evaluation metrics. The results on the same dataset used to tune the thresholds indicate that the proposed framework improves the Pass rate. For Text-based documents, the Pass rate increased from 87.90% to 96.50%, whereas for Image-based documents, it increased from 72.40% to 85.10%. The highest binary F1-scores were 0.982 and 0.920, respectively.

Keywords: Multi-Stage Verification, Fuzzy Matching, Semantic Similarity, Natural Language Inference

Please cite this article as: N. Yindeesuk, C. Khemapatapan, R. Ranokphanuwat and A. Jeamaon, "Field-Level Data Consistency Verification in Merchant Vessel Documents Using a Multi-Stage Framework," *The Journal of Engineering and Industrial Technology, Kalasin University*, vol. 4, no. 5, pp. 83-96, 2026.

บทความวิจัย (Research Article)

1. บทนำ

การตรวจสอบความสอดคล้องกันของเอกสาร 2 ชุดที่เกี่ยวพันกันโดยตรง แต่มาจากผู้ดำเนินการคนละคน เช่น การจัดทำรายงานการตรวจสอบสินค้าระหว่างประเทศขององค์กรธุรกิจซึ่งประกอบด้วยข้อมูลจำนวนมากไม่ว่าจะเป็น รายละเอียดเรือพาณิชย์ รายละเอียดสินค้า ขั้นตอนวิธีการ และผลของการตรวจสอบสินค้า เป็นต้น โดยงานวิจัยนี้จะมุ่งเน้นไปที่การตรวจสอบรายละเอียดเรือพาณิชย์ ซึ่งปรกติจะมีการนำข้อมูลดังกล่าวมาจากฐานข้อมูลที่เชื่อถือได้ของสมาคมจัดชั้นเรือระหว่างประเทศ การตรวจสอบความสอดคล้องของข้อมูลในเอกสารรายงานกับเอกสารอ้างอิงก่อนที่จะทำการอนุมัติและส่งมอบไปยังคู่ค้าทางธุรกิจยังคงเป็นกระบวนการที่ต้องอาศัยเจ้าหน้าที่ดำเนินการ ซึ่งใช้เวลาและมีโอกาสเกิดความผิดพลาดด้วยปัจจัยต่าง ๆ เช่น สภาพร่างกายและจิตใจของตัวผู้ตรวจ สภาพแวดล้อมที่ควบคุมไม่ได้ หรือความเร่งรีบ เป็นต้น

แม้ในปัจจุบันจะสามารถนำเครื่องมือที่มีอยู่มากมายในอินเทอร์เน็ตเข้ามาช่วยแก้ปัญหานี้ได้ แต่เนื่องจากข้อมูลที่ต้องตรวจสอบมีลักษณะเป็นข้อมูลระดับฟิลด์ที่มีลักษณะแตกต่างกัน เช่น ข้อมูลที่เป็นรหัสเฉพาะ ตัวเลข วันที่ ชื่อเรือ ประเภทเรือ ชื่อบริษัท และชื่ออู่ต่อเรือที่มักมีคำขยายของสถานที่ตั้งต่อท้าย จึงทำให้องค์กรต้องพิจารณาเลือกใช้เครื่องมือที่สามารถรองรับความหลากหลายดังกล่าวได้ ในขณะเดียวกันก็ยังต้องสามารถตรวจสอบย้อนกลับและช่วยลดความกังวลของผู้บริหารองค์กรที่ต้องกำกับดูแลป้องกันข้อมูลขององค์กรและของคู่ค้าทางธุรกิจไม่ให้เกิดการรั่วไหลด้วย

งานวิจัยนี้จึงมุ่งเน้นออกแบบและพัฒนาเครื่องมืออัตโนมัติที่ใช้เปรียบเทียบค่าตัวเลขและตัวอักษรภาษาอังกฤษของ 10 ฟิลด์เป้าหมายในเอกสาร รายงานกับเอกสารอ้างอิงด้วยกรอบการทำงานแบบหลายขั้นตอนที่ประกอบไปด้วยเทคนิคเชิงกฎเกณฑ์ที่เสริมด้วยการประมวลผลภาษาธรรมชาติในสภาพแวดล้อมแบบออฟไลน์เพื่อเพิ่มประสิทธิภาพของผู้ตรวจและเพิ่มความมั่นใจให้กับผู้บริหารองค์กรเรื่องการตรวจสอบย้อนกลับได้และเรื่องการป้องกันข้อมูลรั่วไหล ซึ่งในบริบทของงานวิจัยนี้คือ 4 ขั้นตอนดังนี้

1) การจับคู่แบบตรงกันทุกประการ
2) การจับคู่แบบคลุมเครือ
3) ความคล้ายคลึงเชิงความหมาย
4) การอนุมานภาษาธรรมชาติ

2. วัตถุประสงค์ของการวิจัย

1) เพื่อออกแบบและพัฒนาเครื่องมืออัตโนมัติสำหรับตรวจสอบความสอดคล้องของข้อมูลในเอกสารรายงานกับเอกสารอ้างอิงแบบออฟไลน์ด้วยแนวคิดการตรวจสอบแบบหลายขั้นตอน

2) เพื่อเปรียบเทียบประสิทธิภาพของเครื่องมือที่ใช้ตรวจสอบเอกสารอ้างอิงเชิงข้อความและเชิงรูปภาพที่ระดับฟิลด์

3) เพื่อศึกษาข้อผิดพลาดของการรู้จำอักขระด้วยแสง (Optical Character Recognition) และขั้นตอนต่าง ๆ ที่ส่งผลกระทบต่อประสิทธิภาพของเครื่องมือ

3. การทบทวนวรรณกรรมและช่องว่างของงานวิจัย

3.1 กลุ่มงานด้านการสกัดข้อมูล

เน้นการแปลงข้อมูลจากเอกสารให้อยู่ในรูปแบบที่ระบบสามารถนำไปประมวลผลต่อได้ โดยงานของ Rombach and Fettke [1], Atagong et al. [2], Zheng et al. [3], Silva [4] และ Rombach et al. [5] ชี้ให้เห็นว่า การสกัดข้อมูลจากเอกสารจริงยังมีความ

ท้าทายจากรูปแบบเอกสารที่หลากหลาย และคุณภาพของข้อมูล อย่างไรก็ตาม งานส่วนใหญ่ยังเน้นการสกัดข้อมูลมากกว่าการตรวจสอบความสอดคล้องของข้อมูลในเอกสารสองชุด งานวิจัยนี้จึงเพิ่มเติมกระบวนการตรวจสอบความสอดคล้องในระดับฟิลด์และระดับไฟล์ของเอกสารสองชุด

3.2 กลุ่มงานด้านการประมวลผลเอกสารแบบหลายขั้นตอน

ใช้หลายเทคนิคร่วมกัน โดยงานของ Arroyo et al. [6], Ed-Daoudi et al. [7], Yan et al. [8] และ Wang et al. [9] แสดงให้เห็นความสำคัญของการทำงานแบบหลายขั้นตอนเพื่อประมวลผลเอกสารที่ซับซ้อน อย่างไรก็ตาม งานส่วนใหญ่ยังเน้นการเพิ่มประสิทธิภาพการสกัดข้อมูลมากกว่าการจัดลำดับเครื่องมือเพื่อตรวจสอบความสอดคล้องตามชนิดข้อมูล งานวิจัยนี้จึงออกแบบให้มี การจับคู่แบบตรงกันทุกประการ การจับคู่แบบคลุมเครือ ความคล้ายคลึงเชิงความหมาย และการอนุมานภาษาธรรมชาติ ร่วมกับกฎตัวเลขและวันที่ (Numeric and Date Rule) เพื่อให้ข้อมูลแต่ละประเภทถูกตรวจด้วยเกณฑ์ที่เหมาะสม

3.3 กลุ่มงานด้านการตรวจความคล้ายคลึงเชิงความหมาย

วัดความใกล้เคียงของข้อความเชิงความหมาย โดยงานของ Mysore et al. [10] และ Hassan et al. [11] ชี้ให้เห็นว่า การเปรียบเทียบข้อความเชิงทรานส์ฟอร์เมอร์ (Transformer-Based Text Comparison) สามารถใช้วิเคราะห์ความหมายของข้อความได้ อย่างไรก็ตาม งานกลุ่มนี้ส่วนใหญ่เน้นตรวจสอบข้อความทั่วไปหรือข้อความระดับประโยคมากกว่าข้อมูลสั้น ๆ ในระดับฟิลด์ที่ต้องใช้เกณฑ์การตรวจสอบต่างกัน งานวิจัยนี้จึงใช้การตรวจสอบความคล้ายคลึงเชิงความหมายเฉพาะฟิลด์ข้อความที่เหมาะสม และแยกข้อมูลรหัสเฉพาะ ตัวเลข และวันที่ ให้ถูกควบคุมด้วยกฎตัวเลขและวันที่

3.4 กลุ่มงานด้านการประมวลผลด้วยภาษาธรรมชาติ

ใช้การอนุมานภาษาธรรมชาติวิเคราะห์ความสัมพันธ์ระหว่างสองข้อความ โดยงานของ Perikos and Souli [12], Gubelmann et al. [13], Negru et al. [14] และ Madaan et al. [15] แสดงให้เห็นบทบาทของการอนุมานภาษาธรรมชาติในการวิเคราะห์ความสัมพันธ์เชิงความหมายและการประเมินความเข้าใจภาษาของโมเดล ขณะเดียวกัน งานของ ภวิกา อ้วนละมัย และคณะ [16] ศึกษาระบบแนะนำหนังสือด้วยวิธีการแบบอิงเนื้อหา สะท้อนแนวคิดการใช้คุณลักษณะของเนื้อหาเพื่อเปรียบเทียบความเกี่ยวข้องของข้อมูล อย่างไรก็ตาม งานส่วนใหญ่ไม่ได้มุ่งตรวจสอบความสอดคล้องของข้อมูลระดับฟิลด์ในเอกสารโดยตรง โดยเฉพาะข้อมูลชื่อเฉพาะที่มีคำขยาย งานวิจัยนี้จึงใช้การอนุมานภาษาธรรมชาติเป็นส่วนเสริมในขั้นตอนที่ 4 เฉพาะกรณีที่มีความกำกวมเชิงความหมาย เช่น ชื่อเต็มกับชื่อย่อ หรือชื่อบริษัทที่มีคำขยาย

โดยสรุป งานวิจัยที่นำมาอ้างอิงเน้นการสกัดข้อมูลจากเอกสาร การปรับปรุงคุณภาพการอ่านเอกสาร และการเปรียบเทียบข้อความเชิงความหมายเป็นหลัก มากกว่าการตรวจสอบความสอดคล้องของข้อมูลระดับฟิลด์ระหว่างเอกสารสองชุดโดยตรง อีกทั้งข้อมูลระดับฟิลด์ที่มีลักษณะแตกต่างกัน เช่น รหัสเฉพาะ ตัวเลข วันที่ และชื่อองค์กรที่มีคำขยาย ซึ่งไม่ควรใช้เกณฑ์ตรวจสอบแบบเดียวกันทั้งหมด

งานวิจัยนี้จึงเติมเต็มช่องว่างดังกล่าวด้วยการเสนอกรอบการทำงานแบบหลายขั้นตอนเชิงกฎเกณฑ์ที่เสริมด้วยการประมวลผลภาษาธรรมชาติแบบออฟไลน์ โดยใช้การจับคู่แบบตรงกันทุกประการในขั้นตอนที่ 1 เพื่อตรวจสอบข้อมูลที่เป็น รหัสเฉพาะ

บทความวิจัย (Research Article)

ตัวเลข วันที่ ใช้การจับคู่แบบคลุมเครือในขั้นตอนที่ 2 เพื่อตรวจสอบข้อมูลที่อาจมีการพิมพ์ผิดเล็กน้อย ใช้ความคล้ายคลึงเชิงความหมายในขั้นตอนที่ 3 เพื่อตรวจสอบข้อมูลที่เขียนไม่เหมือนกันแต่มีความหมายเดียวกัน และใช้การอนุมานภาษาธรรมชาติในขั้นตอนที่ 4 เพื่อตรวจสอบข้อมูลที่กำกวมว่าสอดคล้องกันหรือไม่ โดยมีการนำกฎตัวเลขและวันที่มาใช้ในขั้นตอนที่ 2 ถึง 4 เพื่อป้องกันไม่ให้เครื่องมือใช้เกณฑ์การตรวจสอบความใกล้เคียงของตัวอักษรหรือความคล้ายคลึงเชิงความหมายกับฟิลด์ที่เป็นรหัสเฉพาะ ตัวเลข และวันที่ พร้อมรายงานผลเชิงอรรถาธิบายในระดับฟิลด์และระดับไฟล์ซึ่งสอดคล้องกับแนวคิดการค้นหาข้อมูลและตรวจสอบความถูกต้องก่อนนำไปใช้งานจริง (Retrieval Then Verify)

4. วิธีดำเนินการวิจัย

4.1 ภาพรวมของเครื่องมือ

เครื่องมือถูกออกแบบให้ตรวจสอบความสอดคล้องของข้อมูลระหว่างเอกสารรายงานกับเอกสารอ้างอิงแบบหลายขั้นตอนตั้งแต่เปรียบเทียบความตรงกันของอักขระจนถึงวิเคราะห์ความสอดคล้องเชิงความหมายและตรรกะ โดยงานวิจัยนี้ใช้เครื่องคอมพิวเตอร์ทั่วไปที่มีหน่วยประมวลผลกลางอินเทลคอร์ไอ7 ความเร็ว 2.8 กิกะเฮิร์ตซ์ หน่วยความจำหลัก 16 กิกะไบต์ หน่วยจัดเก็บข้อมูลหลัก 1 เทระไบต์ และไม่ต้องเชื่อมต่ออินเทอร์เน็ตเพื่อป้องกันข้อมูลรั่วไหล ยกเว้นช่วงต้นที่ต้องทำการดาวน์โหลดโมเดลมาเก็บไว้ในหน่วยจัดเก็บข้อมูลหลักก่อนนำเข้าเอกสาร โดยมีกระบวนการทำงานหลักดังนี้

1) นำเข้าชุดเอกสารเพื่อทำการสกัดข้อความดิบด้วยเครื่องมือต่าง ๆ จากนั้นนำไปบันทึกไว้เพื่อรอประมวลผลในขั้นตอนต่อไป

2) นำข้อความดิบที่ทำการบันทึกไว้มาผ่านขั้นตอนการทำความสะอาดข้อมูล จากนั้นทำการสกัดข้อความดิบจาก 10 ฟิลด์เป้าหมายด้วยไลบรารีนิพจน์ปรกติ (Regex)

3) นำข้อความดิบที่ผ่านการทำความสะอาดข้อมูลแล้วมาเข้ากระบวนการทำการตรวจสอบแบบหลายขั้นตอน

4) นำผลลัพธ์การตรวจสอบแบบหลายขั้นตอนมาเปรียบเทียบกับป้ายกำกับข้อมูลที่ถูกต้องเพื่อทำการประเมินประสิทธิภาพด้วยค่าความแม่นยำ และคะแนนเอฟวันแบบไบนารีรวมถึงวิเคราะห์ค่าต่าง ๆ ในระดับฟิลด์และแสดงผลแบบเชิงอรรถาธิบาย

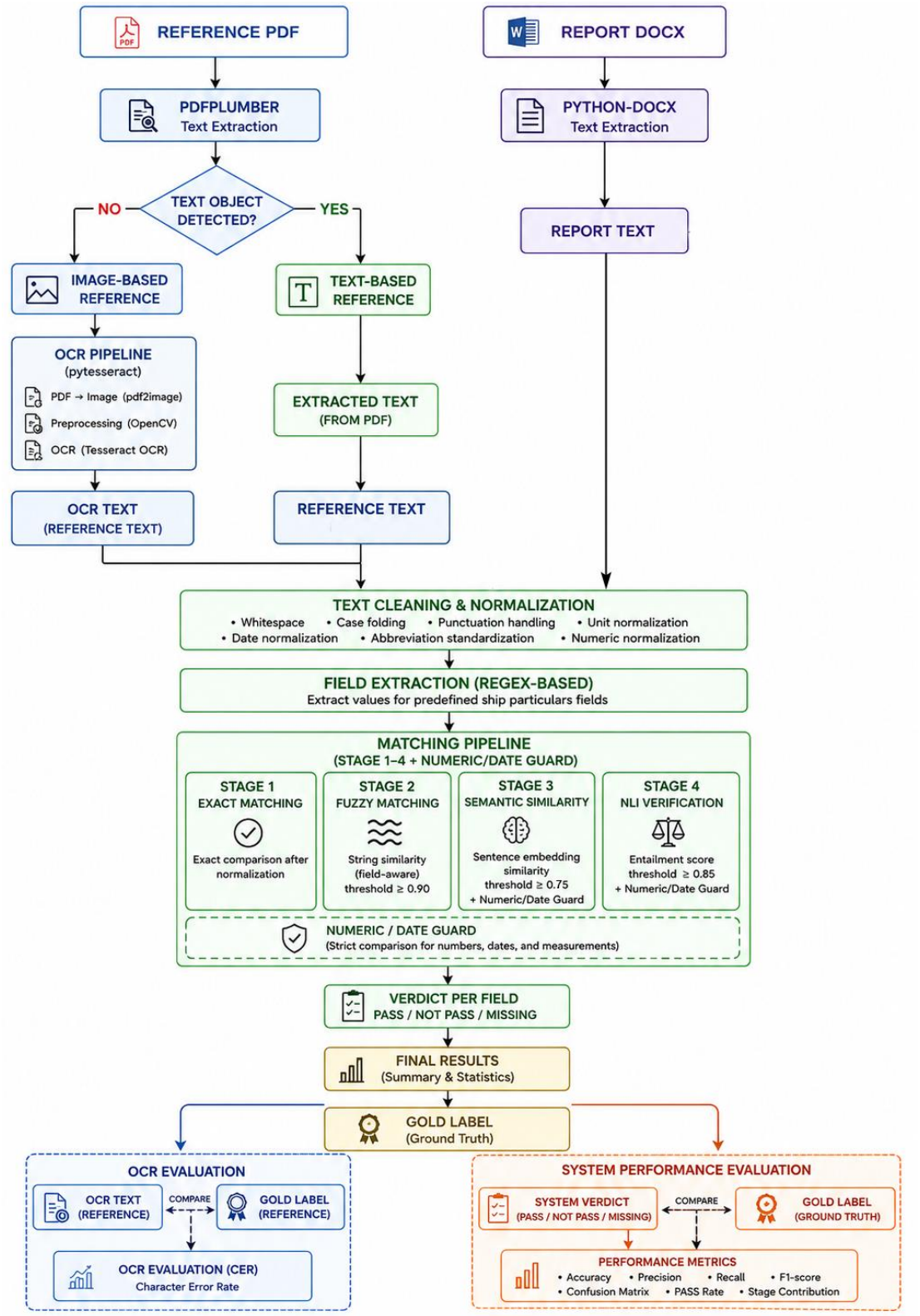
4.2 ชุดข้อมูล

การทดลองในงานวิจัยนี้ดำเนินการบนชุดข้อมูล 4 ประเภท จำนวนรวม 2,000 ไฟล์ โดยมีรายละเอียดของแต่ละประเภทดังนี้

1) “เอกสารรายงาน” 500 ไฟล์
2) “เอกสารอ้างอิงเชิงข้อความ” 500 ไฟล์
3) “เอกสารอ้างอิงเชิงรูปภาพ” 500 ไฟล์ และ
4) “ป้ายกำกับข้อมูลที่ถูกต้อง” 500 ไฟล์

เอกสารอ้างอิงเชิงข้อความและเชิงรูปภาพเป็นเอกสารเดียวกัน แต่ต่างกันที่วิธีการเข้าถึงข้อความ โดยแบบแรกสามารถสกัดข้อความได้โดยตรง ส่วนแบบหลังต้องผ่านขั้นตอนการรู้จำอักขระด้วยแสง ขณะที่เอกสารรายงานเป็นเอกสารต้องการตรวจสอบไม่ใช่ไฟล์เดียวกันกับเอกสารอ้างอิงแต่มีข้อมูลระดับฟิลด์ที่สอดคล้องกัน ชุดข้อมูลใน 3) ได้มาจากการพิมพ์ชุดข้อมูลใน 2) ออกมาเป็นกระดาษแล้วทำการสแกนเป็นไฟล์ภาพเพื่อปรับให้เป็นรูปแบบเชิงรูปภาพเพื่อจำลองสถานการณ์ที่เครื่องมือต้องทำงานกับเอกสารซึ่งไม่สามารถสกัดข้อความได้โดยตรง

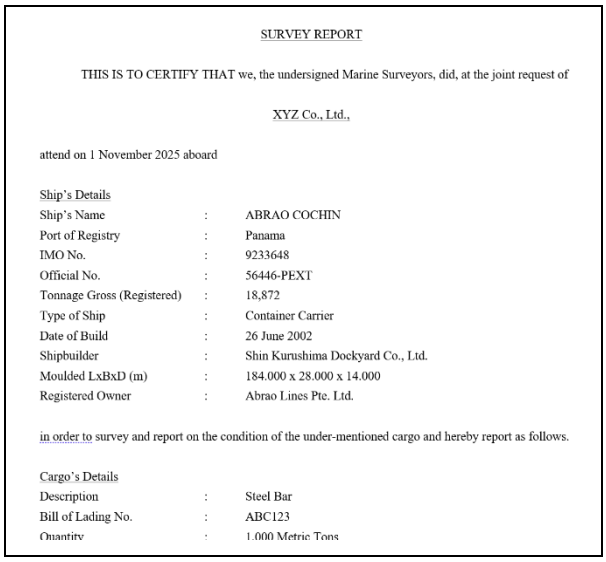
บทความวิจัย (Research Article)



รูปที่ 1 แผนผังกระบวนการทำงานของเครื่องมือ

บทความวิจัย (Research Article)

อย่างไรก็ตาม เอกสารดังกล่าวยังไม่รวมปัจจัยที่อาจส่งผลต่อการประมวลผล เช่น ความเอียงของกระดาษ รอยยับ รอยขีดเขียน ความไม่ชัด เป็นต้น จึงยังไม่สามารถใช้แทนภาพสแกนจริงทุกประเภทได้



SURVEY REPORT

THIS IS TO CERTIFY THAT we, the undersigned Marine Surveyors, did, at the joint request of

XYZ Co., Ltd.,

attend on 1 November 2025 aboard

Ship's Details

Ship's Name	:	ABRAO COCHIN
Port of Registry	:	Panama
IMO No.	:	9233648
Official No.	:	56446-PEXT
Tonnage Gross (Registered)	:	18,872
Type of Ship	:	Container Carrier
Date of Build	:	26 June 2002
Shipbuilder	:	Shin Kurushima Dockyard Co., Ltd.
Moulded LxBxD (m)	:	184.000 x 28.000 x 14.000
Registered Owner	:	Abrao Lines Pte. Ltd.

in order to survey and report on the condition of the under-mentioned cargo and hereby report as follows.

Cargo's Details

Description	:	Steel Bar
Bill of Lading No.	:	ABC123
Quantity	:	1,000 Metric Tons

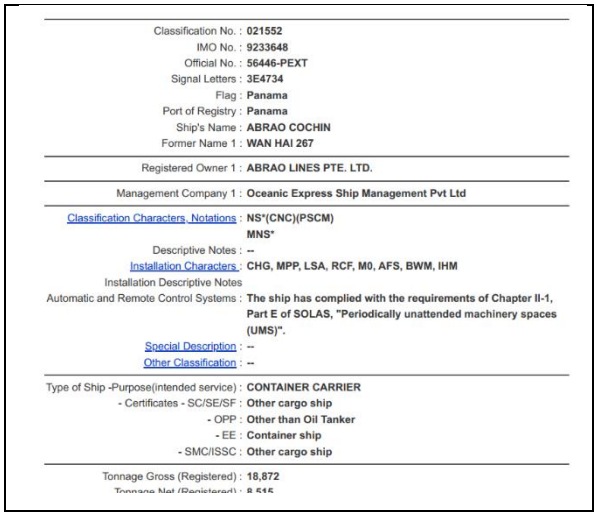
รูปที่ 2 ลักษณะเนื้อหาในเอกสารรายงาน

4.3 รายละเอียดขั้นตอนการทำงานของเครื่องมือ

4.3.1 การนำเข้าข้อมูลและสกัดข้อความดิบ

นำชุดเอกสาร 4 ประเภทดังที่กล่าวถึงก่อนหน้านี้เข้าสู่กระบวนการเพื่อตรวจสอบและทำการสกัดข้อมูลดิบออกมาด้วยเครื่องมือต่าง ๆ ดังนี้

1) เอกสารรายงาน ใช้ python-docx
2) เอกสารอ้างอิงเชิงข้อความ ใช้ pdfplumber
3) เอกสารอ้างอิงเชิงรูปภาพ ใช้ Tesseract
4) ส่วนป้ายกำกับข้อมูลที่ถูกต้องจะถูกบันทึกไว้เพื่อใช้สำหรับการเปรียบเทียบและประเมินประสิทธิภาพ



Classification No. :	021552
IMO No. :	9233648
Official No. :	56446-PEXT
Signal Letters :	3E4734
Flag :	Panama
Port of Registry :	Panama
Ship's Name :	ABRAO COCHIN
Former Name 1 :	WAN HAI 267
Registered Owner 1 :	ABRAO LINES PTE. LTD.
Management Company 1 :	Oceanic Express Ship Management Pvt Ltd
Classification Characters, Notations :	NS*(CNC)(PSCM) MNS*
Descriptive Notes :	--
Installation Characters :	CHG, MPP, LSA, RCF, M0, AFS, BWM, IHM
Installation Descriptive Notes	
Automatic and Remote Control Systems :	The ship has complied with the requirements of Chapter II-1, Part E of SOLAS, "Periodically unattended machinery spaces (UMS)".
Special Description :	--
Other Classification :	--
Type of Ship -Purpose(intended service) :	CONTAINER CARRIER
- Certificates - SC/SE/SF :	Other cargo ship
- OPP :	Other than Oil Tanker
- EE :	Container ship
- SMC/ISSC :	Other cargo ship
Tonnage Gross (Registered) :	18,872

รูปที่ 3 ลักษณะเนื้อหาในเอกสารอ้างอิง

จากนั้นทำการสกัด “ข้อความดิบ” จาก 10 ฟิลด์เป้าหมายในเอกสารรายงานและเอกสารอ้างอิงตามรายการด้านล่างนี้ด้วยไลบรารีนิพจน์ปรกติ

ชื่อฟิลด์

1. Ship’s Name
2. Port of Registry
3. IMO No.
4. Official No.
5. Tonnage Gross (Registered)
6. Type of Ship
7. Date of Build
8. Shipbuilder
9. Moulded LxBxD (m)
10. Registered Owner

โดยผลลัพธ์จะถูกนำไปเก็บไว้ในรูปแบบดังตัวอย่างนี้

ฟิลด์	เอกสารรายงาน	เอกสารอ้างอิง
1.	ข้อความดิบ	ข้อความดิบ

เพื่อนำไปประมวลผลในขั้นตอนต่อไป

บทความวิจัย (Research Article)

4.3.2 การทำความสะอาดข้อมูล

ก่อนไปสู่ขั้นตอนการตรวจสอบแบบหลายขั้นตอน ข้อความดิบที่สกัดมาได้จะถูกนำไปทำความสะอาดเพื่อลดสิ่งรบกวนด้วยฟังก์ชันเสริมต่าง ๆ ดังนี้

ฟังก์ชัน clean_ocr_field() จะทำการลบอักขระที่ไม่เกี่ยวข้อง ปรับรูปแบบเว้นวรรค กรองเฉพาะตัวเลขในฟิลด์ที่เป็นตัวเลข เป็นต้น

ฟังก์ชัน normalize_field_name() จะทำการรวมชื่อฟิลด์หลากหลายที่มีความหมายเดียวกันให้เป็นมาตรฐาน เช่น “ship’s name”, “vessel name” เป็น “ship name” เป็นต้น

ฟังก์ชัน normalize_value() จะทำการปรับรูปแบบข้อมูล เช่น แปลงเป็นตัวพิมพ์เล็ก ลบเครื่องหมายวรรคตอน เป็นต้น

4.3.3 กระบวนการตรวจสอบแบบหลายขั้นตอน

ขั้นที่ 1: การจับคู่แบบตรงกันทุกประการขั้นตอนแรกนี้ใช้การตรวจสอบอักขระแบบหนึ่งต่อหนึ่ง

ขั้นที่ 2: การจับคู่แบบคลุมเครือ ขั้นตอนนี้ใช้การวัดระยะทางเลเวนชเตย์น (Levenshtein Edit Distance) เพื่อวัดความต่างเชิงตัวอักษรระหว่างข้อความ s_1 และ s_2 โดยให้ d(i,j) เป็นระยะทางระหว่างของข้อความ ดังสมการที่ 1

$$d(i,j)=\begin{cases} i, & \text{if } j=0 \\ j, & \text{if } i=0 \\ \min\begin{cases} d(i-1,j)+1 \\ d(i,j-1)+1 \\ d(i-1,j-1)+1[s_1[i]\neq s_2[j]] \end{cases}, & \text{otherwise} \end{cases} \quad (1)$$

โดย

1[s1[i]≠s2[j]] = 0 ถ้าอักขระเท่ากัน และ 1 ถ้าไม่เท่า

d(|s1|,|s2|) คือ ระยะทางเลเวนชเตย์น

อย่างไรก็ตาม ผลลัพธ์ที่ได้ในขั้นตอนนี้มีค่าเป็นจำนวนเต็มและไม่สามารถใช้เทียบข้อความสั้นกับยาวได้เนื่องจากมาตราส่วนไม่เท่ากัน จึงต้องนำค่าดังกล่าวมาทำให้เป็นมาตรฐานเดียวกันด้วยความยาวของข้อความเพื่อให้ค่าอยู่ในช่วง 0-1 ดังสมการที่ 2

$$\text{FuzzySim}(s_1,s_2)=1-\frac{d(|s_1|,|s_2|)}{\max(|s_1|,|s_2|)} \quad (2)$$

โดย

$|s_1|$, $|s_2|$ คือ ความยาวของข้อความ ค่าเข้าใกล้ 1 หมายถึงคล้ายกันมาก เกณฑ์การตัดสินเป็นผ่าน ถ้า FuzzySim ≥0.90

ขั้นที่ 3: ความคล้ายคลึงเชิงความหมาย ขั้นตอนนี้เป็นการคำนวณหาค่าความคล้ายคลึง (Similarity Score) โดยใช้โมเดล Sentence Transformer “all-mpnet-base-v2” สร้างค่าเวกเตอร์ของชุดข้อความ (Sentence Embedding) จากนั้นนำไปคำนวณหาค่าความคล้ายคลึงโคไซน์ (Cosine Similarity) และค่าความคล้ายคลึง ดังสมการที่ 3

$$\text{CosSim}(\mathbf{A},\mathbf{B}) = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\ \|\mathbf{B}\|} \quad (3)$$

โดย

A,B = เวกเตอร์ของข้อความ (เช่น 768 มิติ)

บทความวิจัย (Research Article)

A•B = dot product (คูณแล้วบวกรวมทุกมิติ)

||A||, ||B|| = ความยาวเวกเตอร์

เกณฑ์การตัดสินเป็นผ่าน ถ้า CosSim≥0.75

เหตุผลที่เลือกใช้โมเดลนี้เนื่องจากเป็นโมเดลคุณภาพสูง (768 มิติ) ที่ผ่านการฝึกบนข้อมูลชุดใหญ่ที่ประกอบด้วยประโยคและข้อความสั้น ๆ กว่า 1 พันล้านคู่เพื่อใช้เปรียบเทียบเชิงความหมายโดยเฉพาะ รองรับการถอดความข้อความสั้นได้ดี อีกทั้งยังมีขนาดเล็ก (ประมาณ 400 เมกะไบต์) สามารถดาวน์โหลดลงมาทำงานบนเครื่องคอมพิวเตอร์ทั่วไปได้ Model card: all-mpnet-base-v2.

ขั้นที่ 4: การอนุมานภาษาธรรมชาติ ขั้นตอนสุดท้ายนี้เป็นการคำนวณหาค่าความน่าจะเป็น (P) โดยการใช้โมเดล NLP “roberta-large-mnli” ตรวจสอบความสัมพันธ์ของข้อความ 2 ข้อความ คือ ข้อความตั้งต้น (Premise) กับ ข้อความที่ต้องตรวจ (Hypothesis) โดยโมเดลจะให้ค่าความน่าจะเป็นออกมา 3 จำพวก คือ สนับสนุน (Entailment) ขัดแย้ง (Contradiction) และ ไม่แน่ใจ (Neutral) โดยมีเกณฑ์การตัดสินเป็นผ่าน ถ้า สนับสนุน ≥0.85

เหตุผลที่เลือกใช้โมเดลนี้เนื่องจากเป็นโมเดลคุณภาพสูงที่ผ่านการฝึกบนข้อความภาษาอังกฤษแบบไม่มีป้ายกำกับขนาดประมาณ 160 กิกะไบต์ และยังถูกปรับแต่งบนชุดข้อมูล MNLI (Multi-Genre Natural Language Inference) ซึ่งประกอบด้วยคู่ประโยคประมาณ 433,000 คู่ เพื่อใช้สำหรับงานจัดประเภทข้อความและการอนุมานภาษาธรรมชาติ อีกทั้งยังมีขนาดไม่ใหญ่มาก (ประมาณ 1.6 กิกะไบต์) สามารถดาวน์โหลดลงมาทำงานบนเครื่องคอมพิวเตอร์ทั่วไปได้ Model card: roberta-large-mnli.

4.4 การกำหนดค่าเกณฑ์ (Threshold)

การกำหนดค่าเกณฑ์ในงานวิจัยนี้ดำเนินการผ่านกระบวนการสอบเทียบค่าเกณฑ์ (Threshold Calibration) โดยใช้ผลลัพธ์ระดับฟิลด์จากผลการทดลอง โดยเครื่องมือได้ทดลองค่าเกณฑ์หลายระดับสำหรับใช้ในขั้นตอน การจับคู่แบบคลุมเครือ ความคล้ายคลึงเชิงความหมาย และ การอนุมานภาษาธรรมชาติ แล้วเปรียบเทียบผลด้วย ค่าความแม่นยำ (Accuracy) ค่าความเที่ยงตรง (Precision) ค่าความระลึก (Recall) คะแนนเอฟวันแบบไบนารี ผลบวกปลอม (False Positive) และผลลบปลอม (False Negative)

อย่างไรก็ดี ค่าเกณฑ์ดังกล่าวเป็นค่าที่ได้มาจากการปรับจูนบนชุดข้อมูลทดลอง (2,000 ไฟล์) ชุดเดียวกันกับที่ใช้รายงานผล

การเลือกค่าเกณฑ์ดังกล่าวให้ความสำคัญกับการรักษาค่าความเที่ยงตรงให้อยู่ในระดับสูงก่อน เนื่องจากบริบทของงานตรวจสอบเอกสารมีความเสี่ยงจากผลบวกปลอมมากกว่าผลลบปลอม

ผลการทดลองพบว่าค่าเกณฑ์การจับคู่แบบคลุมเครือ ≥0.90, ค่าเกณฑ์ความคล้ายคลึงเชิงความหมาย ≥0.75 และค่าเกณฑ์การอนุมานภาษาธรรมชาติ ≥0.85 ร่วมกับกฎตัวเลขและวันที่ (Numeric and Date Rule) เป็นการตั้งค่าที่เหมาะสม เนื่องจากสามารถรักษาค่าความเที่ยงตรงไว้ที่ 1.000 ในชุดทดลอง

5. ผลการทดลอง

ผลการทดลองแสดงให้เห็นว่าอัตราการผ่านเกณฑ์เพิ่มขึ้นอย่างต่อเนื่องจากขั้นที่ 1 ถึง 4 โดยเอกสารอ้างอิงเชิงข้อความเพิ่มขึ้น 8.60 จุดร้อยละ จาก 87.90% เป็น 96.50% และเอกสารอ้างอิงเชิงรูปภาพเพิ่มขึ้น 12.70 จุดร้อยละ จาก 72.40% เป็น

85.10% ในขณะที่คะแนนเอฟวันแบบไบนารีสูงสุดอยู่ที่ 0.982 และ 0.920 ตามลำดับ

นอกจากนี้ยังพบว่าประสิทธิภาพของการรู้จำอักขระด้วยแสงมีบทบาทอย่างมีนัยสำคัญเนื่องจากเป็นส่วนทำหน้าที่สกัดข้อความดิบก่อนนำไปประมวลผล โดยค่าเฉลี่ยและมัธยฐานของอัตราส่วนความผิดพลาดของอักขระ (Character Error Rate) ในเอกสารอ้างอิงเชิงรูปภาพอยู่ที่ 0.1528 และ 0.1336 ตามลำดับ

จากผลการทดลองพบว่าแต่ละขั้นมีบทบาทแตกต่างกันอย่างชัดเจน กล่าวคือ

ขั้นที่ 1: การจับคู่แบบตรงกันทุกประการ ให้ค่าความเที่ยงตรงสูง แต่มีค่าความระลึกต่ำ เนื่องจากเป็นการตรวจสอบแบบหนึ่งต่อหนึ่งและไม่รองรับความคลาดเคลื่อนของข้อมูล

ขั้นที่ 2: การจับคู่แบบคลุมเครือ มีบทบาทในการจัดการความแตกต่างด้านรูปแบบของข้อความซึ่งเป็นกรณีที่ข้อความสองชุดเขียนต่างกันแต่มีความหมายเหมือนกัน เช่น

Co Limited กับ Co Ltd

26 Jun 2002 กับ 26 June 2002

Co., Ltd. กับ Co Ltd โดยสามารถเพิ่มอัตราการผ่านเกณฑ์ของเอกสารอ้างอิงเชิงข้อความและเอกสารอ้างอิงเชิงรูปภาพจาก 87.90% เป็น 91.40% และจาก 72.40% เป็น 78.90% ตามลำดับ

ขั้นที่ 3: ความคล้ายคลึงเชิงความหมาย มีบทบาทในการจัดการความแตกต่างเชิงความหมายในกรณีที่ข้อความสองชุดเขียนต่างกันเกินกว่าที่ขั้นที่ 2 จะจัดการได้แต่ยังมีความหมายเหมือนกัน เช่น Mitsui E&S Shipbuilding Co Ltd Tamano Shipyard กับ Mitsui E&S Shipbuilding โดยสามารถเพิ่มอัตราการผ่านเกณฑ์ของเอกสารอ้างอิงเชิงข้อความและเชิงรูปภาพจาก 91.40% เป็น 95.80% และจาก 78.90% เป็น 84.40% ตามลำดับ โดยในขั้นนี้มีการตั้งค่ากฎตัวเลขและวันที่ไว้เพื่อป้องกันความผิดพลาด

ขั้นที่ 4: การอนุมานภาษาธรรมชาติ มีบทบาทเป็นขั้นสุดท้ายในการตรวจสอบว่าข้อความสองชุดมีความสัมพันธ์กันแบบสนับสนุน, ขัดแย้ง และไม่แน่ใจ โดยในขั้นนี้มีการตั้งค่า กฎตัวเลขและวันที่ไว้เพื่อป้องกันความผิดพลาดเช่นกัน

แม้ขั้นนี้จะเพิ่มอัตราการผ่านเกณฑ์จากขั้นที่ 3 ได้จำกัด จาก 95.80% เป็น 96.50% และจาก 84.40% เป็น 85.10% ตามลำดับ แต่ก็ยังคงมีบทบาทสำคัญในเชิงความน่าเชื่อถือของเครื่องมือในกรณีที่ต้องตรวจสอบความสอดคล้องเชิงตรรกะมากกว่าเชิงความหมาย

ตารางที่ 1 ผลการวิเคราะห์อัตราร้อยละ (%) ของจำนวนฟิลด์ที่ตรวจสอบ

ขั้นที่	ฟิลด์รวม	เอกสารอ้างอิงเชิงข้อความ				เอกสารอ้างอิงเชิงรูปภาพ			
		ผ่าน		ไม่ผ่าน		ผ่าน		ไม่ผ่าน	
1	5,000	4,395	87.90%	605	12.10%	3,620	72.40%	1,380	27.60%
2	5,000	4,570	91.40%	430	8.60%	3,945	78.90%	1,055	21.10%
3	5,000	4,790	95.80%	210	4.20%	4,220	84.40%	780	15.60%
4	5,000	4,825	96.50%	175	3.50%	4,255	85.10%	745	14.90%

บทความวิจัย (Research Article)

ตารางที่ 2 ผลการวิเคราะห์ประสิทธิภาพของเครื่องมือจากเมทริกซ์ความสับสน (Confusion Matrix)

ขั้นที่	เอกสารอ้างอิงเชิงข้อความ				เอกสารอ้างอิงเชิงรูปภาพ			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
1	87.90%	1.000	0.879	0.936	72.40%	1.000	0.724	0.840
2	91.40%	1.000	0.914	0.955	78.90%	1.000	0.789	0.882
3	95.80%	1.000	0.958	0.979	84.40%	1.000	0.844	0.915
4	96.50%	1.000	0.965	0.982	85.10%	1.000	0.851	0.920

ตารางที่ 3 ตัวอย่างการตัดสินใจของเครื่องมือโดยใช้ค่าคะแนนเทียบกับค่าเกณฑ์

ขั้นที่	ข้อความดิบที่สกัดได้จากเอกสารอ้างอิง ข้อความดิบที่สกัดได้จากเอกสารรายงาน	ค่าคะแนน	ค่าเกณฑ์	คำตัดสิน
1	Bulk Carrier Bulk Carrier	1	1	ผ่าน
2	African Bateleur Shipping Co Limited African Bateleur Shipping Co Limit	0.94	≥0.90	ผ่าน
3	Mitsui E&S Shipbuilding Co Ltd Tamano Shipyard Mitsui E&S Shipbuilding	0.77	≥0.75	ผ่าน
4	Oil Carrier Bulk Carrier	0.18*	≥0.85	ไม่ผ่าน

ผลการทดลองจากตารางที่ 1 แสดงให้เห็นว่าจำนวนฟิลด์ที่เครื่องมือสามารถตัดสินให้ผ่าน เพิ่มขึ้นอย่างต่อเนื่องในแต่ละขั้นตอน และมีความแตกต่างอย่างชัดเจนระหว่างรูปแบบของข้อมูลทั้งสองประเภท (5,000 ฟิลด์ = 500 ไฟล์ x 10 ฟิลด์)

ผลการทดลองในตารางที่ 2 แสดงให้เห็นว่าประสิทธิภาพโดยรวมเพิ่มขึ้นอย่างต่อเนื่องในแต่ละขั้นตอน และมีความแตกต่างอย่างชัดเจนระหว่างรูปแบบของข้อมูลทั้งสองรูปแบบ โดยค่าความเที่ยงตรงที่รายงานเป็นค่าภายใต้เงื่อนไขของชุดข้อมูลทดลองนี้เท่านั้น

ตารางที่ 3 แสดงตัวอย่าง ขั้นที่ 2 ตัดสินเป็นผ่านเนื่องจากมีค่าคะแนน 0.94 ซึ่งสูงกว่าค่าเกณฑ์ ≥0.90, ขั้นที่ 4 ตัดสินเป็นไม่ผ่านเนื่องจากมีค่าคะแนน 0.18* (Entailment) ซึ่งต่ำกว่าค่าเกณฑ์ ≥0.85 เป็นต้น

นอกจากการรายงานผลแบบนับจำนวนฟิลด์แล้ว งานวิจัยนี้ได้เพิ่มรายงานผลแบบนับจำนวนไฟล์เพื่อสะท้อนการนำเครื่องมือไปใช้จริง โดยกำหนดให้ผลการตรวจสอบไฟล์เป็นผ่าน ก็ต่อเมื่อทั้ง 10 ฟิลด์เป้าหมายในไฟล์นั้นผ่านเกณฑ์การตรวจสอบทั้งหมด แต่หากพบว่าอย่างน้อยหนึ่งฟิลด์ในไฟล์นั้นเป็นไม่ผ่าน ก็จะตัดสินว่าไฟล์นั้นเป็นไม่ผ่าน (ไฟล์ที่ผ่านในขั้นที่ 4 เป็นไฟล์ที่มีฟิลด์ Shipbuilder ซึ่งมีชื่อบริษัทพร้อมคำขยาย) แสดงได้ดังตารางที่ 4

บทความวิจัย (Research Article)

ตารางที่ 4 ผลการวิเคราะห์อัตราร้อยละ (%) ของจำนวนไฟล์ที่ตรวจสอบ

ขั้นที่	ไฟล์รวม	เอกสารอ้างอิงเชิงข้อความ				เอกสารอ้างอิงเชิงรูปภาพ			
		ผ่าน		ไม่ผ่าน		ผ่าน		ไม่ผ่าน	
1	500 ไฟล์	110	22.00%	390	78.00%	15	3.00%	485	97.00%
2	500 ไฟล์	165	33.00%	335	67.00%	50	10.00%	450	90.00%
3	500 ไฟล์	310	62.00%	190	38.00%	85	17.00%	415	83.00%
4	500 ไฟล์	335	67.00%	165	33.00%	100	20.00%	400	80.00%

ตารางที่ 5 อัตราส่วนความผิดพลาดของอักขระ

ค่า	อัตราความผิดพลาด
ค่าเฉลี่ย	0.1528
ค่ามัธยฐาน	0.1336
ค่าสูงสุด	0.6475

ผลการประเมินความคลาดเคลื่อนของข้อความดิบที่สกัดได้จากกระบวนการรู้จำอักขระด้วยแสงโดยใช้ อัตราส่วนความผิดพลาดของอักขระ

6. สรุปและอภิปรายผลการวิจัย

งานวิจัยนี้นำเสนอเครื่องมือตรวจสอบความสอดคล้องของข้อมูลของเอกสารสองชุดแบบหลายขั้นตอน ได้แก่ การจับคู่แบบตรงกันทุกประการ, การจับคู่แบบคลุมเครือ, ความคล้ายคลึงเชิงความหมาย และการอนุมานภาษาธรรมชาติ เพื่อรองรับลักษณะข้อมูลระดับฟิลด์ที่แตกต่างกัน โดยผลบนชุดข้อมูลเดียวกับที่ใช้ปรับค่าเกณฑ์แสดงให้เห็นว่าเครื่องมือสามารถเพิ่มประสิทธิภาพได้อย่างต่อเนื่อง โดยเฉพาะขั้นตอนการจับคู่แบบคลุมเครือ และความคล้ายคลึงเชิงความหมาย ขณะที่การอนุมานภาษาธรรมชาติในขั้นตอนที่ 4 ช่วยเพิ่มอัตราการผ่านเกณฑ์ในระดับฟิลด์ได้จำกัดเพียง 0.70 จุดร้อยละแต่ยังมีประโยชน์ในบางกรณี เช่น ฟิลด์ Shipbuilder ที่มีรูปแบบชื่อเต็ม ชื่อย่อ หรือคำขยายขององค์กร อย่างไรก็ตามงานวิจัยนี้ยังมีข้อจำกัดดังนี้

คุณภาพของการรู้จำอักขระด้วยแสงในเอกสารเชิงรูปภาพ ซึ่งส่งผลต่อการสกัดข้อความดิบ และขั้นตอนการจับคู่ โดยเฉพาะฟิลด์ที่ไวต่อความคลาดเคลื่อน เช่น Ship’s Name, IMO No., Official No. และ Date of Build

ระบบยังไม่สามารถแก้ไขข้อผิดพลาดจากการรู้จำอักขระด้วยแสงได้ทั้งหมด โดยเฉพาะกรณีที่ข้อมูลถูกอ่านผิดหรือสูญหายตั้งแต่ต้น

ชุดข้อมูลเอกสารอ้างอิงเชิงรูปภาพสร้างมาจากการสแกนชุดข้อมูลเอกสารอ้างอิงเชิงข้อความ ทำให้ความแตกต่างของข้อมูลมีเพียงแค่ข้อผิดพลาดจากการรู้จำอักขระด้วยแสงเท่านั้น

การใช้นิพจน์ปรกติ แม้จะรวดเร็วและตรวจสอบย้อนกลับได้ แต่ยังมีข้อจำกัดเมื่อรูปแบบเอกสารเปลี่ยนไป หรือเมื่อสกัดข้อมูลผิดตั้งแต่ต้นก็จะส่งผลให้ขั้นตอนต่อไปไม่สามารถตรวจสอบข้อมูลได้

ค่าเกณฑ์ที่ใช้นั้นได้จากการปรับจูนบนชุดข้อมูลทดลองเท่านั้น จึงยังไม่สามารถรับประกันผลลัพธ์กับเอกสารจริงที่มีความหลากหลายมากขึ้นได้ รวมถึงเอกสารอ้างอิงเชิงรูปภาพในงานวิจัยนี้ยังไม่ครอบคลุมลักษณะของเอกสารสแกนจริงทั้งหมด เช่น ความเอียง รอยยับ ความไม่ชัด เป็นต้น

ไม่มีชุดข้อมูลทดสอบที่แยกต่างหาก ทำให้ค่าความเที่ยงตรง 1.000 เกิดจากชุดข้อมูลทดลองนี้เท่านั้น ไม่สามารถรับประกันได้ว่าจะได้ 1.000 กับเอกสารชุดใหม่

โดยสรุป เครื่องมือที่นำเสนอสามารถเพิ่มประสิทธิภาพการตรวจสอบความสอดคล้องของข้อมูลได้ และเหมาะกับสภาพแวดล้อมที่ต้องการควบคุมเรื่องข้อมูลรั่วไหล เนื่องจากทำงานแบบออฟไลน์ได้ อย่างไรก็ตาม ควรพัฒนาต่อโดยทดสอบกับเอกสารจริงที่หลากหลายมากขึ้น แยกชุดปรับค่าเกณฑ์ออกจากชุดทดสอบ และปรับปรุงการรู้จำอักขระด้วยแสง, การสกัดข้อความดิบ และปรับแต่งให้เป็นมาตรฐาน เพื่อให้ทนทานมากขึ้นในอนาคต

7. เอกสารอ้างอิง

[1] A. M. Rombach and P. Fettke, “Deep learning based key information extraction from business documents: Systematic literature review,” *ACM Computing Surveys*, vol. 58, no. 2, pp. 1-37, 2026. DOI: 10.1145/3749369.

[2] S. D. Atagong, H. E. Z. Tonnang, K. Senagi, M. Wamalwa, K. M. Agboka, and J. Odindi, “A review on knowledge and information extraction from PDF documents and storage approaches,” *Frontiers in Artificial Intelligence*, vol. 8, Art. no. 1466092, 2025. DOI: 10.3389/frai.2025.1466092.

[3] H. Zheng, S. Wang, and L. Huang, “A comprehensive survey on document-level information extraction,” in *Proceedings of the Workshop on the Future of Event Detection*, Miami, FL, USA, 2024, pp. 58-72. DOI: 10.18653/v1/2024.futured-1.6.

[4] K. Silva and T. Silva, “A review on document information extraction approaches,” in *Proceedings of the Student Research Workshop Associated with RANLP 2021*, Online, 2021, pp. 174-179. DOI: 10.26615/issn.2603-2821.2021_024.

[5] A. M. Rombach, J. Lahann, T. Niesen, and P. Fettke, “Utilizing deep learning for field-level information extraction from German real estate tax notices,” *Journal of Emerging Technologies in Accounting*, vol. 22, no. 1, pp. 101-118, 2025. DOI: 10.2308/JETA-2023-028.

[6] R. Arroyo, J. Yebes, E. Martínez, H. Corrales, and J. Lorenzo, “Key information extraction in purchase documents using deep learning and rule-based corrections,” in *Proceedings of the First Workshop on Pattern-based Approaches to NLP in the Age of Deep Learning*, Gyeongju, Republic of Korea, 2022, pp. 11-20. https://aclanthology.org/2022.pandl-1.2/

[7] R. Ed-Daoudi, F. Z. Zakka, M. Ouqassou, and B. Ettaki, “A hybrid OCR-XGBoost-Transformer pipeline for resume parsing with spatial-semantic integration,” *Informatica*, vol. 49, no. 12, 2025. DOI: 10.31449/inf.v49i12.9588.

[8] Z. Yan, Z. Ye, J. Ge, J. Qin, J. Liu, Y. Cheng, and C. Gurrin, “DocExtractNet: A novel framework for enhanced information extraction from business

documents," *Information Processing & Management*, vol. 62, no. 3, Art. no. 104046, 2025. DOI: 10.1016/j.ipm.2024.104046.

[9] G. Wang, J. Yu, X. Zhang, T. Deb, X. Liu, and P. He, "A multi-stage pipeline for accurate handwritten information extraction from financial forms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2025, pp. 6030-6038.

[10] S. Mysore, A. Cohan, and T. Hope, "Multi-vector models with textual guidance for fine-grained scientific document similarity," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Seattle, WA, USA, 2022, pp. 4453-4470. DOI: 10.18653/v1/2022.naacl-main.331.

[11] E. Hassan, A. S. Talaat, and M. A. Elsabagh, "Intelligent text similarity assessment using RoBERTa with integrated chaotic perturbation optimization techniques," *Journal of Big Data*, vol. 12, no. 1, Art. no. 164, 2025. DOI: 10.1186/s40537-025-01233-3.

[12] I. Perikos and S. Souli, "Natural language inference with transformer ensembles and explainability techniques," *Electronics*, vol. 13, no. 19, Art. no. 3876, 2024. DOI: 10.3390/electronics13193876.

[13] R. Gubelmann, I. Katis, C. Niklaus, and S. Handschuh, "Capturing the varieties of natural language inference: A systematic survey of existing datasets and two novel benchmarks," *Journal of Logic, Language and Information*, vol. 33, no. 1, pp. 21-48, 2024. DOI: 10.1007/s10849-023-09410-4.

[14] V. A. Negru, R. Vacareanu, C. Lemnaru, M. Surdeanu, and R. Potolea, "MorphNLI: A stepwise approach to natural language inference using text morphing," in *Findings of the Association for Computational Linguistics: NAACL 2025*, Albuquerque, NM, USA, 2025, pp. 6953-6968. DOI: 10.18653/v1/2025.findings-naacl.385.

[15] L. Madaan, D. Esiobu, P. Stenetorp, B. Plank, and D. Hupkes, "Lost in inference: Rediscovering the role of natural language inference for large language models," in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Albuquerque, NM, USA, 2025, pp. 9229-9242. DOI: 10.18653/v1/2025.naacl-long.466.

[16] ภวิกา อ้วนละมัย และคณะ, "การพัฒนาระบบแนะนำหนังสือด้วยวิธีการแบบอิงเนื้อหา," *วารสารวิศวกรรมและเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยกาฬสินธุ์*, ปีที่ 3, ฉบับที่ 3, หน้า 36-53, 2568. DOI: 10.14456/jeit.2025.13.