

การจำแนกข่าวโควิด-19 โดยใช้อัลกอริทึมความคล้ายคลึงของโคไซน์

COVID-19 news classification using cosine similarity algorithm

อุกฤษฏ์วิวัฒน์ พิมพ์พา¹, สุธิดา ชัยชมชื่น² และวาทีนี นัยเพียร²

Aukkritthiwat Phimpha¹, Suthida Chaichomchuen² and Vatinee Nui pian²

Received: 18 February 2025 ; **Revised:** 21 April 2025 ; **Accepted:** 18 August 2025

บทคัดย่อ

สถานการณ์การแพร่ระบาดของโรคโควิด-19 ที่ผ่านมามีทำให้เกิดความเสียหายในหลากหลายมิติ ผู้คนทั่วโลกต้องเผชิญกับผลกระทบอย่างร้ายแรงของโรคระบาดดังกล่าว ภายใต้ความไม่ชัดเจนของแนวทางการปฏิบัติที่เหมาะสมจากหน่วยงานรัฐที่เกี่ยวข้องผนวกกับคุณลักษณะสำคัญของยุคดิจิทัลที่เปิดโอกาสให้บุคคลทั่วไปสามารถนำเสนอข้อมูลข่าวสารต่าง ๆ ผ่านช่องทางอย่างสื่อสังคมออนไลน์ได้อย่างเสรี ทำให้ข้อมูลข่าวสารเกี่ยวกับการรับมือต่อสถานะการระบาดเต็มไปด้วยข้อมูลแนวทางที่หลากหลายและมีการส่งต่อข้อมูลข่าวสารแบบไร้ระเบียบ โดยงานวิจัยนี้นำเสนอการวัดความคล้ายคลึงของโคไซน์ เพื่อมาใช้ในการจัดหมวดหมู่ระหว่างข่าวด้านสุขภาพ และข่าวที่เกี่ยวข้องกับ โควิด-19 ซึ่งจะช่วยให้ผู้ใช้งานสามารถเข้าใจถึงสถานการณ์ดังกล่าวได้เพิ่มขึ้น และสามารถนำข้อมูลมาใช้ประโยชน์ในการบรรเทาความรุนแรงที่อาจเกิดขึ้นจากสถานการณ์โรคระบาดของคนไทย โดยพิจารณาจากความถูกต้องในการทำนาย สำหรับการประเมินประสิทธิภาพงานวิจัยนั้น ได้ใช้ชุดข้อมูล เว็บไซต์ข่าวสด มติชน อสมท ซึ่งผลการทดลองแสดงให้เห็นว่าการสร้างแบบจำลองการจำแนกบทความข่าวโดยใช้ความคล้ายคลึงของโคไซน์มีความแม่นยำสูงสุดถึง 82.50 เปอร์เซ็นต์

คำสำคัญ: การจำแนกประเภทข้อความ, เหมืองข้อความ, ความคล้ายคลึงของโคไซน์, โควิด-19

Abstract

The COVID-19 pandemic caused severe, multifaceted global impacts. Amidst unclear guidelines from government agencies, the digital age has allowed individuals to freely disseminate information via social media, leading to a flood of diverse and unregulated content regarding pandemic management. This research presents the use of cosine similarity to categorize health news and COVID-19-related news. This classification assists users in better understanding the situation and utilizing information to mitigate potential pandemic-related risks in Thailand. The research effectiveness was assessed using datasets from news websites such as Khaosod, Matichon, and MCOT. Experimental results demonstrate that the classification model using cosine similarity achieved an accuracy of up to 82.50%.

Keywords: Text classification, text mining, cosine similarity, COVID-19

¹ ภาควิชาคอมพิวเตอร์ศึกษา, คณะครุศาสตร์อุตสาหกรรม, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

² อาจารย์, ภาควิชาคอมพิวเตอร์ศึกษา, คณะครุศาสตร์อุตสาหกรรม, มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ

¹ Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok

² Lecturer, Department of Computer Education, Faculty of Technical Education, King Mongkut's University of Technology North Bangkok

* Corresponding author: Aukkritthiwat Phimpha, E-mail: s6202042910039@email.kmutnb.ac.th

บทนำ

ในยุคดิจิทัลที่ข้อมูลข่าวสารมีจำนวนมากและเพิ่มขึ้นอย่างรวดเร็ว การเข้าถึงและการคัดเลือกข่าวสารที่ตรงกับความต้องการของผู้ใช้งานเป็นเรื่องที่ทำหาย เทคโนโลยีและอัลกอริทึมการคาดการณ์ที่มีประสิทธิภาพจึงมีความสำคัญอย่างยิ่ง การวิเคราะห์ความคล้ายคลึงของเนื้อหาข่าวสารเป็นวิธีหนึ่งที่สามารถนำมาใช้ในจำแนกและแนะนำข่าวสารที่เกี่ยวข้องกับความสนใจของผู้ใช้งานได้อย่างแม่นยำจากการติดเชื้อไวรัสโคโรนา 2019 หรือโรคโควิด-19 ที่มีการระบาดในประเทศไทยและทั่วโลก (เอกพล ดวงศรี, 2566) ซึ่งประเทศไทยได้ออกประกาศเป็นโรคติดต่ออันตราย ตามพระราชบัญญัติโรคติดต่อ พ.ศ. 2558 (กรมควบคุมโรค กระทรวงสาธารณสุข, 2563) โดยปีที่มีความรุนแรงพบอัตราป่วยตายด้วยสายพันธุ์เดลต้า สูงสุดช่วงปี 2564 ร้อยละ 0.98 และในปี 2565 พบผู้ป่วยสูงสุด 2,492,054 ราย ที่เกิดจากการระบาดของสายพันธุ์โอมิครอน (ดรุณี โพธิ์ศรี และคณะ, 2566) ซึ่งการแพร่ระบาดของเชื้อไวรัสทำให้เกิดผลกระทบต่อชีวิต สุขภาพ และความเป็นอยู่ของประชาชน จึงต้องมีการเผยแพร่ข่าวสารที่ถูกต้อง รัฐบาลจึงได้จัดตั้งศูนย์บริหารสถานการณ์โควิด-19 (ศบค.) เพื่อการสื่อสารข้อมูลและคำแนะนำเกี่ยวกับโควิด-19

ในยุคดิจิทัลเปิดโอกาสให้บุคคลทั่วไปสามารถนำเสนอข้อมูลข่าวสารผ่านสื่อสังคมออนไลน์ได้อย่างเสรี ทำให้รับมือต่อสภาวะการระบาดของโรคได้หลากหลายแนวทาง เนื่องจากการส่งต่อข้อมูลข่าวสารที่สร้างความเข้าใจผิดให้กับประชาชน (ทองพูล มุขรักษ์ และคณะ, 2566) เช่น การรักษาระยะห่างทางสังคม การกักตัว อาการของโรคที่เปลี่ยนไป

การนำความรู้เกี่ยวกับเทคนิคการทำเหมืองข้อความ (Text Mining) มาใช้ในการจำแนกและการพยากรณ์ในงานด้านต่าง ๆ ไม่ว่าจะเป็น รัชพิชญ์ ชำนาญกิจ และจิตติรัตน์ ศิริบรรรัตนกุล (2565) งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาระบบต้นแบบที่เข้าถึงง่ายและสามารถตรวจสอบได้อย่างรวดเร็วว่าข้อความหรือบทความภาษาไทยใด ๆ เป็นข่าวปลอม หรือบทความปลอมหรือไม่ ระบบที่พัฒนาขึ้นอาศัยฐานข้อมูล Google BigQuery ที่รวบรวมข่าวปลอมภาษาไทยจากเว็บไซต์และสื่อออนไลน์ที่เกี่ยวข้องกับการแจ้งเตือนข่าวปลอม โดยใช้เทคนิคในการตัดแบ่งคำ วิเคราะห์โครงสร้าง และแยกชนิดของคำเพื่อสกัดคุณลักษณะสำคัญของข่าวปลอม ระบบนี้ได้รับอินพุตเป็นข้อความจากผู้ใช้ ซึ่งจะถูกนำไปเปรียบเทียบกับฐานข้อมูลข่าวปลอมโดยใช้วิธีการจับคู่คุณลักษณะ ข้อมูลที่มีความคล้ายคลึงกันมากที่สุดสามอันดับแรกจะแสดงผลเป็นผลลัพธ์ เพื่อให้ผู้ใช้สามารถพิจารณาข่าวที่ใกล้เคียงกับอินพุตได้ จากผลการทดลองในชุดข้อมูลทดสอบ ซึ่งประกอบด้วยข่าวจริงและ

ข่าวปลอมรวม 120 ข่าว ระบบสามารถระบุข่าวปลอมได้อย่างถูกต้อง โดยมี ค่า precision เท่ากับ 91.84% และ recall เท่ากับ 75.00%

องอาจ อุ่นนันต์ และพยุ่ง มีสัจ (2563) สร้างแบบจำลองการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย มีวัตถุประสงค์เพื่อศึกษาข้อมูลปัจจัยที่เกี่ยวข้องกับความน่าเชื่อถือของเว็บไซต์ และเปรียบเทียบประสิทธิภาพของแบบจำลองที่ใช้ในการจำแนกประเภท โดยการรวบรวมข้อมูลปัจจัยทางเทคนิคของเว็บไซต์แหล่งข่าวและสื่อสังคมออนไลน์ของแหล่งข่าวแล้วทำการจัดกลุ่มข้อมูลเว็บไซต์แหล่งข่าวเพื่อกำหนดป้ายกำกับกลุ่มของแหล่งข่าว โดยจัดกลุ่มที่มีประสิทธิภาพดีที่สุดแบ่งออกเป็น 5 กลุ่ม จากนั้นทำการวิเคราะห์ข้อมูลด้วยเทคนิคการจำแนกประเภทประกอบด้วย 5 เทคนิค ดังนี้ Decision Tree (C4.5), Naïve Bayes, K-NN, Multilayer Perceptron และ Support Vector Machine (SVM) แล้วเปรียบเทียบค่าประสิทธิภาพพบว่าเทคนิค K-NN ที่มีค่า K เท่ากับ 5, 6 และ 7 มีค่าประสิทธิภาพมากที่สุดเท่ากัน (Accuracy=96.03%, Precision=0.962, Recall=0.960, F-measure=0.959)

กชกร เทียมทัต (2565) นำเสนอการตรวจจับข่าวปลอมภาษาไทยด้วยกระบวนการเรียนรู้ของเครื่องสามแบบ จำลองได้แก่ Naïve Bayes, Neural Network และ SVM โดยเก็บข้อมูลข่าวภาษาไทยจากเว็บไซต์ศูนย์ต่อต้านข่าวปลอม (ประเทศไทย) ในช่วงเดือนกันยายน 2561 ถึงเดือนกันยายน 2565 ข้อมูลข่าวผ่านกระบวนการเลือกคุณลักษณะ การแปลงค่าคุณลักษณะ การแบ่งค่าข้อมูลเป็นช่วง และการปรับช่วงของค่าข้อมูลให้อยู่ในช่วงเดียวกัน นอกจากนี้ยังผ่านกระบวนการสุ่มเกินเพื่อทำให้ข้อมูลข่าวมีจำนวนข่าวจริงและข่าวปลอมเท่ากัน ผลการวิจัยพบว่า แบบจำลอง SVM มีความถูกต้องร้อยละ 94 มากกว่าค่าความถูกต้องของแบบจำลอง Naïve Bayes และ Neural Network

การใช้ความคล้ายคลึงของโคไซน์เป็นเครื่องมือที่มีความสำคัญในหลายงานวิจัย เช่น การค้นหาข้อมูล ใช้ในการจัดอันดับเอกสารที่เกี่ยวข้องกับคำค้นหา (Nakata & Nezu, 2013) การจัดหมวดหมู่เอกสารหรือบทความข่าวตามความคล้ายคลึงกับหมวดหมู่ที่กำหนด (Kowsari *et al.*, 2019) การแนะนำเนื้อหาที่เกี่ยวข้อง เช่น การแนะนำบทความ ข่าว หรือผลิตภัณฑ์ให้กับผู้ใช้ตามความสนใจ (Nuipian & Chuaykhun, 2023) แม้ว่าอัลกอริทึมความคล้ายคลึงของโคไซน์จะถูกนำมาใช้อย่างแพร่หลายในงานวิจัยต่าง ๆ แต่ยังคงงานวิจัยที่ประเมินผลการใช้งานในบริบทของภาษาไทย ซึ่งมีลักษณะเฉพาะที่แตกต่างจากภาษาอื่น ๆ เช่น ลักษณะเฉพาะของภาษาไทย การไม่มีการเว้นวรรคระหว่างคำ

งานวิจัยนี้มีวัตถุประสงค์เพื่อพัฒนาแบบจำลองสำหรับจำแนกบทความข่าวภาษาไทย โดยประยุกต์ใช้เทคนิคการวัดความคล้ายคลึงของโคไซน์ (Cosine Similarity) ซึ่งเป็นกระบวนการหนึ่งในงานประมวลผลภาษาธรรมชาติและเหมืองข้อความ เทคนิคดังกล่าวช่วยให้สามารถประเมินระดับความคล้ายคลึงกันของเนื้อหาข่าวได้อย่างมีประสิทธิภาพ และสามารถประยุกต์ใช้ในการค้นหา จัดหมวดหมู่ข่าว แนะนำเนื้อหาที่เกี่ยวข้อง และดึงข้อมูลที่ตรงตามบริบท เพื่อสนับสนุนการสื่อสารข้อมูลที่ถูกต้อง ลดการเผยแพร่ข้อมูลที่บิดเบือนในยุคดิจิทัล และเพิ่มความสามารถในการเข้าถึงข่าวสารที่น่าเชื่อถือ เพื่อประเมินความเหมาะสมของเทคนิค Cosine Similarity ในบริบทของภาษาไทย งานวิจัยนี้จึงได้เปรียบเทียบประสิทธิภาพกับ Jaccard Similarity ซึ่งเป็นอัลกอริทึมพื้นฐานที่มักใช้กับการวัดความคล้ายคลึงของคำในรูปแบบเซต (set-based comparison) โดยงานวิจัยก่อนหน้านี้ เช่น Chaudhary and Khurana (2021) แสดงให้เห็นว่า Cosine Similarity มักให้ผลลัพธ์ที่แม่นยำกว่า Jaccard ในงานวิเคราะห์ข้อความ โดยเฉพาะเมื่อนำมาใช้กับข้อมูลที่มีโครงสร้างภาษาเฉพาะเช่นภาษาไทย ผู้วิจัยจึงมุ่งศึกษาประสิทธิภาพของ Cosine และ Jaccard Similarity ในบริบทของการจำแนกข่าวเกี่ยวกับโรคโควิด-19 เพื่อหาข้อเสนอแนะในการเลือกเทคนิคที่เหมาะสมต่อไป

วัตถุประสงค์ของการวิจัย

เพื่อพัฒนาแบบจำลองการจำแนกข่าวโควิด-19 โดยอัลกอริทึมการวัดความคล้ายคลึงของโคไซน์

วิธีดำเนินการวิจัย

ขั้นตอนการสร้างแบบจำลองประกอบไปด้วย การรวบรวมข้อมูล การทำความสะอาดข้อมูล การประมวลผลข้อความ การคัดเลือกคำสำคัญ การคำนวณค่าความคล้ายคลึงของโคไซน์ การสร้างแบบจำลอง และการประเมินประสิทธิภาพแบบจำลอง ดัง Figure 1

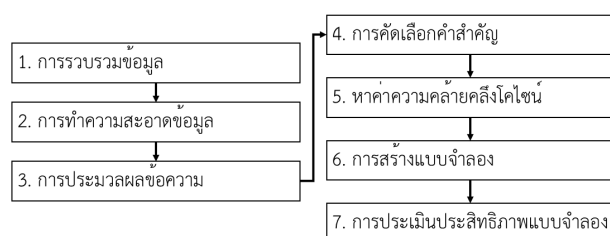


Figure 1 Research Process.

1. การรวบรวมข้อมูล (Data Collection)

งานวิจัยฉบับนี้ได้รับรวบรวมข่าวโรคโควิด-19 บนเว็บไซต์ข่าวสด มติชน อสมท ตั้งแต่วันที่ 26 ตุลาคม 2564 ถึงวันที่ 31 มกราคม 2566 จำนวน 20,000 ข่าว โดยใช้ Web Scraper ซึ่งเป็นส่วนเสริมของ Google Chrome การกำหนดหมวดหมู่ของข่าวในงานวิจัยนี้อ้างอิงจากหมวดหมู่ที่เว็บไซต์ข่าวต้นทางได้จัดไว้แล้ว ซึ่งมีอยู่ในเว็บไซต์ข่าวสด มติชน และอสมท โดยระบบจัดการเนื้อหา (Content Management System) ของแต่ละเว็บไซต์จะมีการจัดกลุ่มเนื้อหาข่าวตามหัวข้อที่เกี่ยวข้อง จากนั้นผู้วิจัยได้ตรวจสอบข่าวแต่ละบทความเพิ่มเติมเพื่อยืนยันความถูกต้องของหมวดหมู่ โดยพิจารณาจากเนื้อหาหลักของข่าวว่ามีการกล่าวถึงโรคโควิด-19 โดยตรงหรือไม่ เพื่อให้ได้ชุดข้อมูลที่มีความถูกต้องและสอดคล้องกับวัตถุประสงค์ของการวิจัย

2. การทำความสะอาดข้อมูล (Data Cleansing)

ขั้นตอนนี้เป็นเตรียมข้อมูลให้สะอาดและมีความสมบูรณ์ก่อนนำไปใช้กับแบบจำลอง โดยการลบข่าวที่ซ้ำซ้อน และกำจัดอักขระที่ไม่จำเป็น เช่น การลบสัญลักษณ์ และอักขระพิเศษ ด้วย Regular Expression การลบลิงก์ URL HTML tag และอีโมจิ

3. การประมวลผลข้อความ

3.1 การตัดคำ (Word Segmentation)

ใช้ไลบรารีของ PyThaiNLP เลือกใช้เอ็นจิน (Engine) แบบ newmm (Phatthiyaphaibun *et al.*, 2023) เป็นการอ้างอิงจากพจนานุกรมโดยใช้เทคนิค Maximal Matching Algorithm และ Thai Character Cluster (TCC) ร่วมกับพจนานุกรมที่กำหนดเอง (Custom Dictionary) ผลลัพธ์ใน Figure 2

News content
กรมวิทยาศาสตร์การแพทย์,แถลงข่าวการเฝ้า, ระวัง, เชื้อ,โควิด-19,กลายพันธุ์,พบ,เชื้อ, ...
โควิด-19,::ผว,จังหวัด,ชัยนาท,ส่ง,ปิด,ชุมชน,สุข ใจ,หลัง,พบ,คลัสเตอร์,ใหม่,ตรวจ,พบ,เชื้อ,...
คลัสเตอร์,โรงพัก,โคราชลาม,ไม่,หยุด,ติด,เชื้อ ,เพิ่ม,อีก,3,ราย,รวม,แล้ว,33,ราย,ไม่,เร่ง,ด่วน, ขอให้,งด,ติดต่อ,ราชการ,...

Figure 2 Example of the result of word cutting.

3.2 การลบคำหยุด (Stop Word Removal)

ใช้ไลบรารีของ PyThaiNLP ที่มีการสร้างคลังคำหยุดสำหรับภาษาไทย เช่น คำว่า “ที่” “ถึง” และ “นั้น” เป็นต้น รวมทั้งคำที่มีอักขระตัวเดียว เช่น “ก” “ข” และ “ด” เป็นต้น

3.3 น้ำหนักคำของคำ (Term Weighting)

การให้น้ำหนักคำตามความถี่คำ – ความถี่ผกผันของคำในเอกสาร (Term frequency-inverse document frequency : TF-IDF) เป็นหนึ่งในคุณลักษณะ (Feature) ที่ได้รับความนิยมอย่างแพร่หลาย ซึ่ง TF-IDF เป็นวิธีการทางสถิติที่ใช้ในการประเมินความสำคัญของคำต่อเอกสาร ความสำคัญของคำเป็นสัดส่วนโดยตรงกับจำนวนครั้งที่คำคำนั้นปรากฏในเอกสารนั้น คำนวณได้จากสมการที่ 1 (Manning *et al.*, 2008) คำ TF-IDF ที่แสดงใน Figure 3 คือคำ TF-IDF ของคำในเอกสารทั้งหมดที่มีค่ามากที่สุด 50 คำ

$$TF - IDF_{(t,d)} = f_{(t,d)} \times \log\left(\frac{N}{df_t}\right)$$

เมื่อ

$f_{(t,d)}$ = จำนวนครั้งที่คำ t ปรากฏในเอกสาร d

N = จำนวนเอกสารทั้งหมดในคลังข้อมูล

df_t = จำนวนเอกสารที่มีคำ t ปรากฏ

คำ	TF-IDF	คำ	TF-IDF
ติดเชื้อ	0.087	โดส	0.029
โควิด	0.087	เดินทาง	0.029
คน	0.079	สถานการณ์	0.028
วัคซีน	0.072	ล้าน	0.027
จำนวน	0.064	ข้อมูล	0.027
ผู้ป่วย	0.064	บ้าน	0.027
วันที่	0.063	เดือน	0.024
โรงพยาบาล	0.060	ที่จะ	0.024
ตรวจ	0.054	เรื่อง	0.024
ไทย	0.054	แพร่ระบาด	0.024
ประเทศ	0.053	เวลา	0.023
จังหวัด	0.049	ทำ	0.023
โรค	0.049	ระบบ	0.023
ประชาชน	0.047	ป้องกัน	0.022
ฉีดวัคซีน	0.045	เสี่ยง	0.022
พื้นที่	0.045	สะสม	0.021
เข็ม	0.042	อายุ	0.021
ปี	0.040	ดูแล	0.020
เชื้อ	0.037	ควบคุม	0.020
อาการ	0.036	บริการ	0.020
กล่าวว่า	0.036	ระบาด	0.020
สำหรับ	0.034	ยา	0.019
ฉีด	0.033	รายงาน	0.019
นพ.	0.032	รักษา	0.019
มาตรการ	0.031	เสียชีวิต	0.019

Figure 3 Word Weights by frequency.

4. การคัดเลือกคำสำคัญ

นำคำที่ได้จาก Figure 3 ซึ่งผ่านการคำนวณค่าความถี่ และการกรองคำที่ไม่สำคัญแล้วมาพิจารณาความสัมพันธ์ระหว่างคำแต่ละคู่โดยใช้หลักการ co-occurrence หรือการพบคำร่วมกันในข่าวเดียวกัน เกณฑ์ที่ใช้ในการคัดเลือกคำคู่สำคัญ คือ คู่คำใดที่มีค่านับสนับสนุน (Support) หรือความถี่ในการปรากฏร่วมกันสูงกว่า 6,000 ครั้ง จะถือว่าคู่นั้นมีความสัมพันธ์ในเชิงความสำคัญ ซึ่งแสดงให้เห็นถึงความเกี่ยวข้องระหว่างคำทั้งสองคำ มีผลใน Figure 4

ลำดับ	คู่คำ	ค่าสหสัมพันธ์
1	ฉีดวัคซีน, เข็ม	16,783
2	เข็ม, โดส	15,291
3	จำนวน, เข็ม	10,807
4	ติดเชื้อ, วันที่	10,554
5	ประชากร, เข็ม	10,342
6	ฉีดวัคซีน, โดส	9,634
7	จำนวน, ฉีดวัคซีน	7,671
8	วัคซีน, เข็ม	7,131
9	จำนวน, โดส	6,957
10	ผู้ป่วย, โรงพยาบาล	6,687
11	ฉีดวัคซีน, วัคซีน	6,604
12	ติดเชื้อ, อายุ	6,129
13	ติดเชื้อ, โรงพยาบาล	6,048

Figure 4 The effects of common words in COVID-19 news.

5. ความคล้ายคลึงของโคไซน์

เป็นแนวคิดทางคณิตศาสตร์ที่ใช้วัดความคล้ายคลึงระหว่างเวกเตอร์สองตัวในพื้นที่ n มิติ โดยพิจารณาจากมุมระหว่างเวกเตอร์ (Angle) มากกว่าความยาวของเวกเตอร์ (Magnitude) เทคนิคนี้นิยมใช้ในหลากหลายสาขา เช่น การประมวลผลภาษาธรรมชาติ การทำเหมืองข้อมูล การค้นคืนข้อมูล และระบบแนะนำ (Schütze *et al.*, 2008)

การคำนวณค่าจากความต่างของมุม 2 มุม ดัง Figure 5 และในสมการที่ 2 ซึ่งผลลัพธ์จะมีค่าอยู่ระหว่าง 0 ถึง 1 เท่านั้น หากมีค่าเท่ากับ 0 แปลว่าคู่มิมีความคล้ายคลึงกัน แต่หากค่าเท่ากับ 1 แปลว่าคู่มิเหมือนกันทุกประการ (Chaudhary & Khurana, 2021)

$$\text{Cosine similarity} = \frac{A \cdot B}{||A|| ||B||} \quad (2)$$

เมื่อ

A และ B คือเวกเตอร์สองตัวที่ต้องการวัดความคล้ายคลึง $A \cdot B$ คือ ผลคูณจุด (dot product) ระหว่างเวกเตอร์ทั้งสอง สามารถคำนวณได้จากสมการที่ 3

$$A \cdot B = \sum_{i=1}^n A_i \cdot B_i \quad (3)$$

$||A||$ และ $||B||$ คือความยาว (magnitude) ของเวกเตอร์ A และ B สามารถคำนวณได้จากสมการที่ 4 และสมการที่ 5

$$||A|| = \sqrt{\sum_{i=1}^n A_i^2} \quad (4)$$

$$||B|| = \sqrt{\sum_{i=1}^n B_i^2} \quad (5)$$

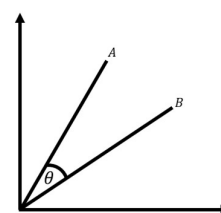


Figure 5 Similarity of cosine between A and B.

6. การสร้างแบบจำลอง

ในการสร้างแบบจำลองสำหรับการทดลองครั้งนี้ ชุดข้อมูลจำนวน 20,000 ข่าว ได้รับการประเมินโดยใช้เทคนิค 5-fold cross-validation เพื่อให้ผลการประเมินมีความน่าเชื่อถือ และสามารถสะท้อนประสิทธิภาพของแบบจำลองในเชิงทั่วไป (generalization) ได้ดียิ่งขึ้น (Kohavi, 1995) ผู้วิจัยได้นำค่าที่มีค่าสนับสนุนสูง จาก Figure 4 มาสร้างกฎความสัมพันธ์ จำนวนทั้งสิ้น 13 กฎ โดยแต่ละค่าจะถูกใช้ในการคำนวณค่าความคล้ายคลึงของโคไซน์ (Cosine Similarity) หากผลรวมของค่าความคล้ายคลึงทั้งหมดมากกว่า Threshold = 0.5 ซึ่งเป็นค่ามาตรฐานที่นิยมใช้ในงานจำแนกประเภทแบบไบนารี (Sebastiani, 2002) บทความข่าวนั้นจะถูกจัดเป็นข่าวที่เกี่ยวข้องกับโควิด-19 แบบจำลองได้รับการพัฒนาโดยใช้ภาษา Python ร่วมกับแพลตฟอร์ม Google Colab ในการพัฒนาและทดสอบประสิทธิภาพอย่างเป็นระบบ

7. การประเมินประสิทธิภาพแบบจำลอง (Evaluation model)

การประเมินประสิทธิภาพของแบบจำลองสำหรับการวิจัยใช้เครื่องมือที่ชื่อว่าตาราง Confusion matrix ดัง Table 1 ประเมินผลลัพธ์ของการทำนาย โดยประกอบไปด้วยการประเมินค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก ค่าความสมดุล (Alpaydin, 2021)

Table 1 Confusion matrix.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TP)	False Positives (FP)
Predicted Negative (0)	False Negative (FN)	True Negative (TN)

เมื่อ

TP = True Positive คือ คลาสเป้าหมายเท่ากับ Yes และพยากรณ์ว่า Yes

TN = True Negative คือ คลาสเป้าหมายเท่ากับ Yes และพยากรณ์ว่า No

FP = False Positive คือ คลาสเป้าหมายเท่ากับ No และพยากรณ์ว่า No

FN = False Negative คือ คลาสเป้าหมายเท่ากับ No และพยากรณ์ว่า Yes

ค่าความถูกต้อง (Accuracy) คำนวณจากสัดส่วนของจำนวนการทำนายที่ถูกต้องต่อจำนวนตัวอย่างทั้งหมดในชุดข้อมูล ดังสมการที่ 6

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

ค่าความแม่นยำ (Precision) คือ การทำนายคลาสบวกอย่างถูกต้อง โดยคำนวณจากสัดส่วนของจำนวนการทำนายที่ถูกต้องเมื่อเปรียบเทียบกับจำนวนทั้งหมดที่แบบจำลองทำนายเป็นบวก ดังสมการที่ 7

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

ค่าความระลึก (Recall) คือ เป็นค่าที่บ่งบอกว่า แบบจำลองที่ทำนายถูกต้องมีกี่ค่า ซึ่งสามารถหาได้จากสมการดังสมการที่ 8

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

ค่าความสมดุล (F1-Score) คือ ค่าความสมดุลระหว่างความแม่นยำในการทำนายในกลุ่มเป้าหมาย และค่าความระลึก ซึ่งสามารถหาดังสมการที่ 9

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

ผลการทดลองและวิจารณ์

หลังจากพัฒนาแบบจำลองจึงดำเนินการประเมินประสิทธิภาพของแบบจำลองโดยใช้เทคนิค 5-fold cross-validation เพื่อให้ผลลัพธ์มีความน่าเชื่อถือและลดอคติที่อาจเกิดจากการแบ่งข้อมูลเพียงครั้งเดียว ผลการทดลองสรุปได้ว่าแบบจำลองที่ใช้ Cosine Similarity ให้ค่าความถูกต้องเฉลี่ย 87.20%

News	Cosine	Class	Prediction
The Department of Disease Control (DDC) reported that Thailand ranked 26th worldwide in cumulative COVID-19 cases, with a cumulative vaccination coverage of 41.6%. The DDC stated that 19 deaths were reported, including 1 case in Bangkok, 2 in Nakhon Ratchasima, 2 in Yasothon, 2 in Udon Thani, 1 in Khon Kaen, 1 in Kamphaeng Phet, 1 in Chanthaburi, 1 in Phitsanulok, 1 in Nakhon Si Thammarat, 1 in Chumphon, 2 in Rayong, 1 in Chanthaburi, 1 in Nakhon Nayok, 1 in Prachin Buri, 1 in Suphan Buri, and 1 in Phetchaburi. Among the fatalities, 13 were male and 6 were female. Thirteen individuals were 60 years of age or older, while six were under 60 years of age. Six had underlying medical conditions, whereas none had a documented history of chronic disease. The DDC also reported that, as of 13 June, a total of 15,350 vaccine doses had been administered. Since 28 February 2021, cumulative vaccinations reached 138,713,285 doses, comprising 56,860,010 first doses (81.7% of the population), 52,923,673 second doses (76.1%), and 28,929,602 third doses (41.6%).	0.758	COVID-19	COVID-19
The first four countries in the world have officially declared COVID-19 an endemic disease, while Thailand is still considering the transition. The Ministry of Public Health is preparing guidelines for classifying COVID-19 as endemic. On 15 June, the Ministry of Public Health stated that the Omicron BA.2 variant generally causes milder symptoms, with declining infection and mortality rates, although infections among high-risk groups continue to increase. Dr. Opas Kamkawinpong, Director-General of the Department of Disease Control, emphasized the importance of receiving a booster (third) vaccine dose. For the current situation, 23,618 new RT-PCR-confirmed cases were reported, with an additional 42,138 positive ATK results, bringing the total number of new cases to 65,756. Among these patients, 1,131 developed pneumonia and 325 required mechanical ventilation. Cumulative confirmed cases reached 734,727, with 18,939 patients hospitalized, 543,184 cumulative recoveries (since 1 January), 223,414 patients still receiving treatment, and 49 new deaths.	0.713	COVID-19	COVID-19

Figure 6 Results from forecasting COVID-19 news data and other health news.

ค่าความแม่นยำ 85.40% ค่าความระลึก 89.10% และค่าความสมดุล 87.20% แบบจำลองที่ใช้ Jaccard Similarity ค่าความถูกต้องเฉลี่ย 78.20% ค่าความแม่นยำ 74.10% ค่าความระลึก 85.20% และค่าความสมดุล 79.30% สามารถแสดงตัวอย่างดัง Figure 6

หลังจากทดสอบประสิทธิภาพแบบจำลองแล้วจึงนำข้อมูลข่าวจำนวน 200 ข่าว ซึ่งเป็นข้อมูลใหม่ที่ไม่เคยใช้ในการพัฒนาแบบจำลอง (Unseen Data) โดยแบ่งเป็นข่าวที่เกี่ยวข้องกับโควิด-19 จำนวน 100 ข่าว ข่าวอื่น ๆ ที่เกี่ยวข้อง กับสุขภาพจำนวน 100 ข่าว เพื่อนำมาเปรียบเทียบ ประสิทธิภาพระหว่าง Cosine Similarity กับ Jaccard Similarity และตรวจสอบความถูกต้อง

การทดสอบกับแบบจำลองทำการวัดค่าความถูกต้อง ค่าความแม่นยำ ค่าความระลึก และค่าความสมดุล ซึ่งผลจากหาประสิทธิภาพ พบว่า แบบจำลองให้ Cosine Similarity ค่าความถูกต้อง 82.50% ค่าความแม่นยำ 80.40% ค่าความระลึก 86.00% และค่าความสมดุล 83.10% แบบจำลองให้ Jaccard Similarity ค่าความถูกต้อง 70.00% ค่าความแม่นยำ 64.90% ค่าความระลึก 87.00% และค่าความสมดุล 74.40% ดัง Table 2

Table 2 Results of model performance evaluation.

Confusion Matrix for Cosine Similarity				
	Covid-19		Health	
Covid-19	86		14	
Health	21		79	
Confusion Matrix for Jaccard Similarity				
	Covid-19		Health	
Covid-19	79		21	
Health	33		67	
	Accuracy	Precision	Recall	F1-score
Cosine	0.825	0.804	0.860	0.831
Jaccard	0.700	0.649	0.870	0.744

สรุปผลและข้อเสนอแนะ

สำหรับจำแนกข่าวโควิด-19 โดยกระบวนการเริ่มตั้งแต่การรวบรวมข้อมูลจากเว็บไซต์ข่าวจำนวน 20,000 ข่าว การทำความสะอาดข้อมูล การประมวลผลข้อความ และการ

สร้างแบบจำลองผ่านการวิเคราะห์ความคล้ายคลึงของข้อความ ใน Figure 6 พบว่าข่าวที่เกี่ยวข้องกับโรคโควิด-19 แต่คำว่า “โควิด” กลับไม่ปรากฏเป็นคำที่พบบ่อยที่สุดในลำดับต้น ๆ ซึ่งสะท้อนให้เห็นถึงลักษณะเฉพาะของการใช้ภาษาไทยที่ใช้ในงานข่าว ในกรณีของโรคระบาดที่มีการใช้คำหลากหลายเพื่ออ้างถึงสิ่งเดียวกัน ในข่าวมักใช้คำศัพท์ที่หลากหลายในการกล่าวถึงโควิด-19 เช่น “ไวรัส”, “โรคระบาด”, “โคโรนา”, “โคโรนาไวรัส”, “โควิด-19” หรือ “COVID-19” แบบจำลองที่พัฒนาขึ้นในงานวิจัยนี้ไม่อาศัยเพียงการตรวจสอบคำหลัก (Keyword Matching) อย่างคำว่า “โควิด” เท่านั้น แต่พิจารณาความสัมพันธ์เชิงเนื้อหาโดยรวมของข่าว ซึ่งช่วยให้สามารถเข้าใจและจัดประเภทข่าวจาก บริบทและความหมายร่วมของคำทั้งชุด มากกว่าการพึ่งพาคำใดคำหนึ่งเพียงอย่างเดียว ผลการทดลองแสดงให้เห็นว่าแบบจำลองสามารถจำแนกข่าวได้อย่างมีประสิทธิภาพ โดยผลลัพธ์การประเมินประสิทธิภาพมีค่าความถูกต้องตอนสร้างแบบจำลองอยู่ที่ 87.20% และ 82.50% ในชุดข้อมูลใหม่ที่ไม่เคยใช้ในการพัฒนาแบบจำลอง

นอกจากนี้ เมื่อนำมาทดสอบเปรียบเทียบระหว่าง Cosine Similarity กับ Jaccard Similarity แบบจำลองที่พัฒนาจากความคล้ายคลึงของโคไซน์นั้นมีประสิทธิภาพที่ดีกว่าราว ๆ 12.5% ซึ่งสอดคล้องกับงานวิจัยของ ปรีนทร์ วงษ์ทอง (2568) ได้เปรียบเทียบระหว่าง Cosine Similarity และ Jaccard Similarity ในระบบแนะนำรายวิชาสำหรับการเทียบโอน พบว่า Cosine Similarity ให้ค่าความแม่นยำ (Precision) และค่าความสมดุล (F1-score) สูงกว่า Jaccard อย่างมีนัยสำคัญ

ผลการทดลองชี้ให้เห็นว่า Cosine Similarity เป็นเครื่องมือที่มีประสิทธิภาพในการจัดหมวดหมู่ข่าวสารด้านสุขภาพ และสามารถประยุกต์ใช้เป็นเครื่องมือในการคัดกรองหรือนำข่าวที่เกี่ยวข้องได้อย่างมีประสิทธิภาพในยุคข้อมูลข่าวสารที่แพร่กระจายรวดเร็ว ลดความเสี่ยงจากข่าวปลอมหรือข้อมูลที่บิดเบือน ซึ่งเป็นประโยชน์ต่อการสื่อสารข้อมูลสาธารณสุขอย่างถูกต้องและแม่นยำ

ข้อเสนอแนะในการทำวิจัยในอนาคตควรมีการวิเคราะห์เพิ่มเติม เช่น การสร้างกราฟ ROC Curve เพื่อประเมินประสิทธิภาพของแบบจำลองในเชิงลึก และการทดสอบ sensitivity analysis เพื่อค้นหาค่า threshold ที่เหมาะสมที่สุดในบริบทของภาษาไทย

นอกจากนี้ ควรพัฒนาแบบจำลองให้สามารถจำแนกข่าวได้หลากหลายหมวดหมู่มากยิ่งขึ้น หรือขยายขอบเขตของชุดข้อมูลให้ครอบคลุมข่าวเกี่ยวกับโรคอื่น ๆ เช่น โรคไข้หวัดใหญ่สายพันธุ์ใหม่, RSV, หรือโรคระบบทางเดินหายใจอื่น ๆ

ที่มีลักษณะของการรายงานข่าวคล้ายกับโรคโควิด-19 เพื่อประเมินความสามารถของแบบจำลองในการแยกแยะเนื้อหาข่าวที่มีบริบทใกล้เคียงกัน และนำไปประยุกต์ใช้ในระบบการจัดการข่าวสารด้านสุขภาพที่มีความแม่นยำ ครอบคลุม และตอบสนองต่อสถานการณ์ปัจจุบันได้ดียิ่งขึ้น

กิตติกรรมประกาศ

ขอขอบพระคุณภาคีวิชาคอมพิวเตอร์ศึกษา คณะครุศาสตร์อุตสาหกรรม มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือที่อำนวยความสะดวกสถานที่ ห้องปฏิบัติการ และอำนวยความสะดวกต่าง ๆ จนทำให้การวิจัยครั้งนี้เสร็จสมบูรณ์ตามเป้าประสงค์ของนักวิจัย

เอกสารอ้างอิง

- กรมควบคุมโรค กระทรวงสาธารณสุข. (2563). คู่มือเจ้าหน้าที่รับมือภาวะฉุกเฉินด้านสุขภาพจากการระบาดของโรคโควิด-19 ในประเทศไทย. <https://ddc.moph.go.th/viralpneumonia/>
- กษกร เทียมทัต. (2565). การตรวจจับข่าวปลอมภาษาไทยโดยใช้แบบจำลองการเรียนรู้ของเครื่อง [วิทยานิพนธ์วิทยาศาสตร์มหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์].
- ดรุณี โพธิ์ศรี, เอกลักษณ์ เอี่ยมประดิษฐ์ และเอกพล เสมอชัย. (2566). การเฝ้าระวังและควบคุมโรคโควิด-19: การมีส่วนร่วมของทุกภาคส่วนในการจัดการสถานการณ์ฉุกเฉินด้านสาธารณสุขในจังหวัดนครปฐม. *วารสารวิทยาศาสตร์สุขภาพแห่งประเทศไทย*, 32(ภาคผนวก 1), S49–S60.
- รัชพิชญ์ ข่านาญกิจ และจิตติรัตน์ ศิริบรรณรัตน์กุล. (2565). การตรวจสอบข่าวปลอมภาษาไทยด้วยเทคนิคการประมวลผลภาษาธรรมชาติ. *วารสารวิจัยและพัฒนา มจร.*, 45(2), 275–287.
- ทองพูล มุขรักษ์, บำรุง ศรีนวลปาน, ชวัลรัตน์ ศรีนวลปาน และกรกฎ จำเริญ. (2566). จริยธรรมและจรรยาบรรณของสื่อในสังคมไทยในการนำเสนอข่าวยุคดิจิทัล กรณีศึกษา: การกราดยิงหนองบัวลำภู, คดีน้องชมพู่ และคดีแดงโม ความท้าทายที่ยังมีปัญหา. *วารสารวิชาการบัณฑิตศึกษา มหาวิทยาลัยราชภัฏนครศรีธรรมราช*, 1(1), 54–67.
- ปรินทร์ วงษ์ทอง, อนเนก พุทธิเดช และสุวิทย์ สมสุภาพรุ่งยศ. (2568). การพัฒนาระบบแนะนำการเทียบโอนรายวิชาโดยหาค่าความเหมือนของข้อความ. *เอกสารประกอบการประชุมวิชาการระดับชาติด้านเทคโนโลยีสารสนเทศและการสื่อสารแห่งประเทศไทย (NCCIT 2025)*, 317–324. มหาวิทยาลัยเชียงใหม่.

- องอาจ อุ่นอนันต์ และพญง มีสัง. (2563). การจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทยโดยใช้เทคนิคการทำเหมืองข้อมูล. *วารสารวิชาการมหาวิทยาลัยอีสเทิร์นเอเซีย ฉบับวิทยาศาสตร์และเทคโนโลยี*, 14(2), 101–116.
- เอกพล ดวงศรี. (2566). สงคราม (โรค) โควิด-19: อุปลักษณะเชิงมโนทัศน์ของสังคมไทยและโลกที่สะท้อนผ่านหนังสือพิมพ์ไทยออนไลน์. *วารสารมนุษยศาสตร์มหาวิทยาลัยเชียงใหม่*, 24(1), 226–251.

- Alpaydin, E. (2021). *Machine learning*. MIT Press.
- Chaudhary, K., & Khurana, S. (2021). Performance evaluation of text similarity measures using cosine similarity in document clustering. *International Journal of Computer Applications*, 183(1), 25–30.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 14(2), 1137–1145.
- Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150. <https://doi.org/10.3390/info10040150>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Nakata, T., & Nezu, M. (2013). An overview of information retrieval. In *Information retrieval and hypertext* (pp. 1–16). Springer.
- Nuipian, V., & Chuaykhun, J. (2023, December). Book recommendation system based on course descriptions using cosine similarity. *Proceedings of the 2023 7th International Conference on Natural Language Processing and Information Retrieval* (pp. 273–277).
- Phatthiyaphaibun, W., Chaovanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L., Chormai, P., & Udomcharoenchaikit, C. (2023). PyThaiNLP: Thai natural language processing in Python. *Journal of Open Source Software*, 8(1), Article 23456. <https://doi.org/10.5281/zenodo.123456>
- Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* (Vol. 39). Cambridge University Press.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), 1–47.