การตรวจสอบข่าวปลอมภาษาไทยด้วยเทคนิคการประมวลผลภาษาธรรมชาติ

ธัชพิชญ์ ชำนาญกิจ[1] และ ฐิติรัตน์ ศิริบวรรัตนกุล[2*]
สถาบันบัณฑิตพัฒนบริหารศาสตร์ ถนนเสรีไทย คลองจั่น บางกะปิ กรุงเทพฯ 10240

** Corresponding Author: thitirat@as.nida.ac.th*

[1] นักศึกษาระดับปริญญาโท สาขาวิชาวิทยาการคอมพิวเตอร์และระบบสารสนเทศ คณะสถิติประยุกต์

[2] ผู้ช่วยศาสตราจารย์ คณะสถิติประยุกต์

ข้อมูลบทความ

ประวัติบทความ :
รับเพื่อพิจารณา : 27 พฤษภาคม 2564
แก้ไข : 20 มิถุนายน 2565
ตอบรับ : 23 มิถุนายน 2565
DOI : 10.14456/kmuttrd.2022.16

คำสำคัญ :
ข่าวปลอม / การตรวจจับข่าวปลอม / การประมวลผลภาษาธรรมชาติ

บทคัดย่อ

งานวิจัยนี้มีจุดประสงค์เพื่อพัฒนาระบบต้นแบบที่เข้าถึงง่าย และสามารถตรวจสอบได้อย่างรวดเร็ว ว่าข้อความหรือบทความภาษาไทยใด ๆ นั้นเป็นข่าวปลอมหรือบทความปลอมหรือไม่ ในการนี้ ผู้วิจัยทำการสร้างฐานข้อมูล Google BigQuery เพื่อเก็บข่าวปลอมภาษาไทยที่รวบรวมมาจากเว็บไซต์และสื่อออนไลน์ ซึ่งเป็นแหล่งรวมการแจ้งเตือนข่าวปลอมในประเทศไทยเอาไว้ จากนั้น จึงใช้เทคนิคการประมวลผลภาษาธรรมชาติในการตัดแบ่งคำในหัวข้อข่าว ตัดประโยค และพิจารณาแยกชนิดของคำ เพื่อสกัดเฉพาะคุณลักษณะสำคัญของข่าวปลอมนั้น ๆ ออกมาเก็บไว้ในส่วนของการใช้งาน สำหรับผู้ใช้ทั่วไปนั้น อินพุตข่าวภาษาไทยจากผู้ใช้จะถูกนำไปผ่านกระบวนการประมวลผลภาษาธรรมชาติ และผลของการสกัดข้อมูลที่ได้จะถูกนำไปเปรียบเทียบกับข้อมูลในฐานข้อมูลข่าวปลอม โดยระบบจะแสดงผลลัพธ์ข้อมูลในฐานข้อมูล 3 อันดับแรกที่คล้ายกับข้อความอินพุตข่าวจากผู้ใช้มากที่สุด จากผลการทดลองในข้อมูลชุดทดสอบ พบว่า ในจำนวนอินพุตที่เป็นข่าวจริงและข่าวปลอมรวมทั้งหมด 120 ข่าว ระบบต้นแบบสามารถระบุอินพุตที่เป็นข่าวปลอมได้ถูกต้องด้วย precision = 91.84% และ recall = 75.00%

Thai Fake News Detection Using Natural Language Processing

Thatchapit Chumnankit[1] and Thitirat Siriborvornratanakul[2*]
National Institute of Development Administration (NIDA), Serithai Road, Klong-Chan, Bangkapi, Bangkok 10240

** Corresponding Author: thitirat@as.nida.ac.th*

[1] Master's Student, Computer Science and Information Systems Program, Graduate School of Applied Statistics.

[2] Assistant Professor, Graduate School of Applied Statistics.

ArticleInfo

Article History:
Received: May 27, 2021
Revised: June 20, 2022
Accepted: June 23, 2022
DOI : 10.14456/kmuttrd.2022.16

Keywords :
Fake News / Fake News Detection / Thai Natural Language Processing

Abstract

In this research, a lightweight prototype of a Thai fake news detection system as an alternative for users looking for fast and accessible Thai fake news detectors is proposed. The prototype is based on a Thai fake news database (Google BigQuery) where the news is collected from online resources and communities regarding Thai fake news warnings. Using techniques based on natural language processing, each news is processed and the extracted feature is stored in the database. For each user query, the query is transformed into a feature vector; such a vector is compared to the vectors of fake news stored in the database. Top three news (in the database) with maximum matching scores are then displayed to the user. Based on the experimental results on the test set, the current prototype can identify 120 news with the precision of 91.84% and recall of 75.00%.

1. บทนำ

ในปัจจุบันด้วยวิทยาการความก้าวหน้าทางเทคโนโลยีและเครื่องมืออำนวยความสะดวกต่าง ๆ ทำให้ผู้คนสามารถเข้าถึงข้อมูลข่าวสารรวมถึงส่งต่อข้อมูลข่าวสารหนึ่งได้อย่างง่ายดาย รวดเร็วเพียงแค่ปลายนิ้วคลิก อีกทั้งยังมีประสิทธิภาพในการแพร่กระจายข่าวสูง ลักษณะเด่นประการหนึ่งของการรับส่งข่าวสารในยุคปัจจุบันคือการรับส่งข่าวสารเป็นไปในลักษณะบุคคลต่อบุคคล ไม่ถูกผูกติด คั่นกลาง เซ็นเซอร์ หรือคัดกรองด้วยสื่อมวลชนอย่างโทรทัศน์ วิทยุ หรือหนังสือพิมพ์เหมือนเช่นอดีต แม้การรับส่งข้อมูลข่าวสารในยุคปัจจุบันนี้จะมีข้อดีหลายประการ แต่ก็มีข้อเสียที่เด่นชัดคือข่าวสารที่ถูกส่งต่อกันนั้นจะไม่ผ่านการตรวจสอบจากกลุ่มสื่อสารมวลชนมืออาชีพหรือหน่วยงานที่น่าเชื่อถือใด ๆ มาก่อน

ปัญหา “ข่าวปลอม (Fake News)” ถือเป็นปัญหาหนึ่งของสังคมไทยและสังคมโลกที่สามารถส่งผลกระทบได้เป็นวงกว้างจากจุดเริ่มต้นเล็ก ๆ วัตถุประสงค์ของการสร้างข่าวปลอมนั้นมีอยู่หลากหลายไม่ว่าจะเป็นการสร้างข่าวปลอมเพื่อโจมตีบุคคลหรือฝ่ายตรงข้าม เพื่อทำให้ผู้คนจำนวนมากแตกตื่น เพื่อหวังผลประโยชน์ทางการค้า เพื่อเพิ่มยอดผู้ติดตามและยอดการกดไลก์ (Like) รวมไปถึงการสร้างข่าวปลอมเพื่อให้เกิดความเชื่อผิด ๆ เป็นต้น แต่ไม่ว่าจะด้วยเหตุผลใด ก็เป็นที่ยอมรับกันในระดับสากลว่าการสร้างข่าวปลอมนั้นไม่มีผลดีที่ยั่งยืนต่อสังคมในระยะยาว อีกทั้งด้วยความรวดเร็วของสื่อสังคมออนไลน์ที่เป็นส่วนสำคัญสำหรับการติดต่อสื่อสารในปัจจุบันยังทำให้การจำกัดขอบเขตความเสียหายของข่าวปลอมหนึ่งนั้นเป็นไปได้ยาก จากการศึกษาวิจัยของ Vosoughi และคณะ [1] จากห้องปฏิบัติการ Media Lab, Massachusetts Institute of Technology (MIT) พบว่า บนแพลตฟอร์มทวิตเตอร์ (Twitter) นั้น ข้อมูลข่าวปลอมนอกจากจะสามารถแพร่กระจายได้ไวกว่าแล้ว เนื้อหาของข้อมูลที่ถูกเผยแพร่บนแพลตฟอร์มนี้บางข้อมูลอาจไม่เคยปรากฏขึ้นจริงมาก่อนแต่เป็นการสร้างข้อมูลดังกล่าวขึ้นเอง เพื่อวัตถุประสงค์ใดวัตถุประสงค์หนึ่ง และยังขาดส่วนวิเคราะห์ความน่าเชื่อถือของข้อมูลอีกด้วย

งานวิจัยชิ้นนี้มีจุดประสงค์เพื่อพัฒนาระบบต้นแบบ (Prototype) สำหรับช่วยตรวจสอบข่าวปลอมให้กับผู้รับส่งข่าวสารในประเทศไทยคล้ายคลึงกับโครงการโคแฟค (Collaborative Fact Checking, https://cofact.org/about) อันเป็นสังคมออนไลน์บนเว็บไซต์สำหรับให้ผู้ใช้ร่วมด้วยช่วยกันตรวจสอบที่มาและข้อเท็จจริงของข่าวแต่ละข่าว ทั้งนี้งานวิจัยชิ้นนี้จะเป็นการนำเอาเทคนิคปัญญาประดิษฐ์ (Artificial Intelligence, AI) ในแขนงของการประมวลผลภาษาธรรมชาติ (Natural Language Processing) มาประยุกต์ใช้กับข้อความหรือหัวข้อข่าวภาษาไทยซึ่งเป็นอินพุตจากผู้ใช้ จากนั้นจึงทำการเปรียบเทียบว่าอินพุตจากผู้ใช้มีความใกล้เคียงกับข้อมูลใดที่ถูกเก็บอยู่ในฐานข้อมูลข่าวปลอมของระบบต้นแบบบ้างหรือไม่ โดยผู้วิจัยหวังว่าผลลัพธ์ที่ได้จากระบบต้นแบบนี้จะสามารถเป็นข้อมูลเพิ่มเติมให้ผู้ใช้สามารถเลือกที่จะเชื่อหรือไม่เชื่อข่าวนั้น ๆ ก่อนจะตัดสินใจเผยแพร่ข่าวต่อไปในวงกว้าง

2. ทบทวนวรรณกรรม

งานวิจัยของ Chulrod และ Nontakhamchan [2] เสนอการใช้เทคนิคการทำเหมืองข้อมูล (Data Mining) และการเรียนรู้ของเครื่อง (Machine Learning) เพื่อสร้างแบบจำลองสำหรับทำนายข่าวปลอมจากสื่อสังคมออนไลน์ โดยงานวิจัยนี้เลือกใช้ Application Programming Interface (API) ดึงข้อมูลข่าวจากทวิตเตอร์และเฟซบุ๊ก (Facebook) มาเป็นจำนวน 600 ข่าว จากนั้นจึงใช้เทคนิคการประมวลผลภาษาธรรมชาติมาตีความข่าวและสร้างเป็นแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 4 แบบ ได้แก่ Naïve Bays, Decision tree, K-Nearest Neighbors และ Multi-layer Perceptron โดยผลการทดลองพบว่าแบบจำลองที่ให้ผลลัพธ์ดีที่สุดคือแบบจำลอง Multi-layer Perceptron ซึ่งให้ค่าความถูกต้องในการทำนายข่าวปลอมอยู่ที่ 95.78% ด้วยแนวคิดที่คล้ายคลึงกันงานวิจัยของ Srivastava [3] ก็นำเสนอวิธีการตรวจสอบข่าวปลอมแบบเรียลไทม์โดยใช้การเรียนรู้ของเครื่องร่วมกับการประมวลผลภาษาธรรมชาติ โดยมีการทดลองแบบจำลองการเรียนรู้ของเครื่องทั้งหมด 5 แบบ ได้แก่ Support Vector Machine (SVM), Naïve Bayes, Random Forest Classifier, Logistic Regression และ Stochastic Gradient Descent ผลสรุปจากงานวิจัยนี้พบว่าแบบจำลอง SVM สามารถทำนายข่าวจริงหรือข่าวปลอมได้ด้วยความแม่นยำถึง 95%

ด้วยความก้าวหน้าของเทคนิคการเรียนรู้ของเครื่องจักรเชิงลึก (Deep Machine Learning) หรือที่รู้จักกันในชื่อ “การเรียนรู้เชิงลึก (Deep Learning)” โดยเฉพาะภายหลังจากที่ Google Research นำเสนอสถาปัตยกรรมชื่อ Transformer [4] ขึ้นมา ก็ทำให้เกิดการพัฒนาแบบก้าวกระโดดในแวดวงการประมวลผลภาษาธรรมชาติ เป็นเหตุให้ระยะหลังมีผลงานวิจัยจำนวนไม่น้อยที่นำแบบจำลองและเทคนิคการเรียนรู้เชิงลึกมาใช้ร่วมกับการประมวลผลภาษาธรรมชาติ เช่น งานวิจัยของ Thota และคณะ [5] ที่นำแบบจำลองการเรียนรู้เชิงลึกชนิด Dense Neural Network มาใช้ตรวจสอบข่าวปลอม โดยมีกระบวนการเตรียมข้อมูลในขั้นต้นประกอบด้วยการลบคำที่ไม่จำเป็น (Stop word removal) การลบสัญลักษณ์และเครื่องหมาย (Punctuation removal) และการลบคำเติมหน้าหรือเติมท้ายในภาษาอังกฤษ (Stemming) จากนั้นจึงนำผลลัพธ์ข้อความตัวอักษรที่ได้แปลงให้กลายเป็นตัวเลขที่พร้อมต่อการป้อนเป็นอินพุตให้กับแบบจำลองการเรียนรู้เชิงลึก ทั้งนี้ขั้นตอนการแปลงจากข้อความตัวอักษรให้กลายเป็นตัวเลขนั้นประกอบไปด้วย 2 ขั้นตอนย่อย ได้แก่ การสร้างคลังคำศัพท์ (Bag of words) สำหรับเปลี่ยนคำแต่ละคำให้กลายเป็นตัวเลขที่ไม่ซ้ำกัน และการใช้ Term Frequency-Inverse Document Frequency (TF-IDF) เพื่อพิจารณาองค์ประกอบของคำภายในประโยคโดยไม่นำลำดับของคำภายในเอกสารมาใช้วิเคราะห์ประกอบ ผลลัพธ์จากงานวิจัยนี้สามารถตรวจสอบข่าวปลอมได้แม่นยำถึง 94.2% บนชุดข้อมูลสำหรับทดสอบซึ่งแบบจำลองไม่เคยเห็นหรือรู้จักมาก่อน

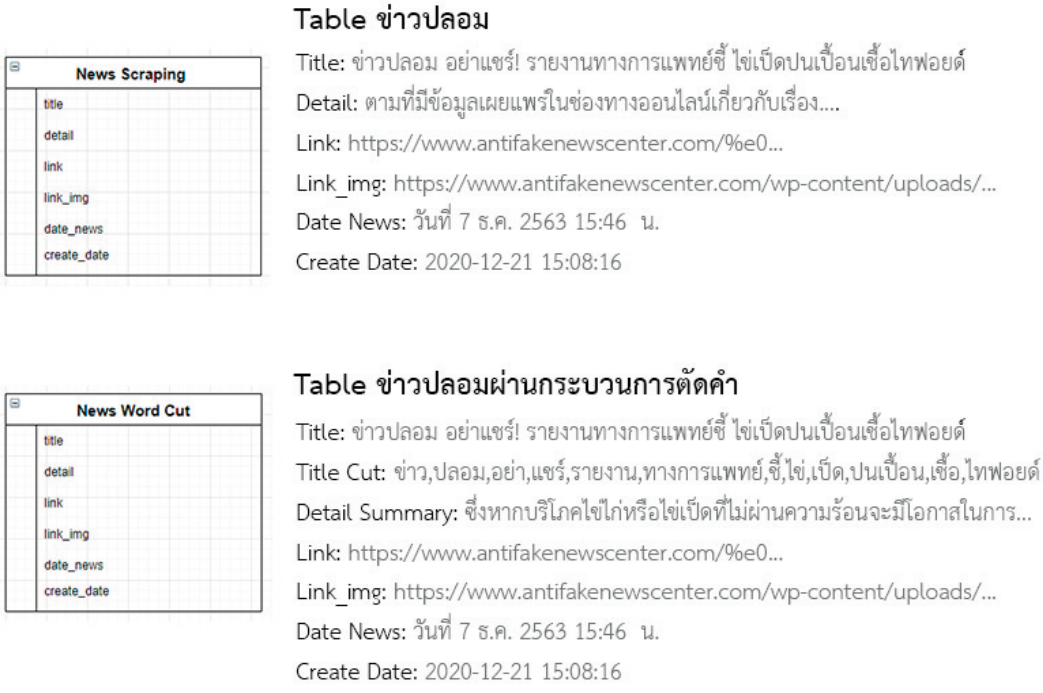
จากตัวอย่างงานวิจัยที่กล่าวไปข้างต้นจะเห็นว่าการนำเทคนิคการเรียนรู้ของเครื่องมาประกอบกับการประมวลผลภาษาธรรมชาติเป็นแนวทางการวิจัยและพัฒนาที่ได้รับความนิยมในปัจจุบัน เนื่องด้วยความพร้อมของข้อมูลใหญ่ (Big Data) สำหรับใช้สอนและทดสอบแบบจำลอง ประกอบกับประสิทธิภาพที่เหนือชั้นกว่าอย่างเด่นชัดในปัญหาที่ซับซ้อน อาทิ การแปลภาษา (Machine Translation) อย่างไรก็ตามข้อจำกัดของเทคนิคการเรียนรู้ของเครื่องโดยเฉพาะชนิดที่เป็นการเรียนรู้เชิงลึกนั้น คือ ต้องการทรัพยากรต่าง ๆ เป็นจำนวนมาก ทั้งจำนวนข้อมูลสำหรับสอน (Train dataset) ระยะเวลารอสอนแบบจำลองให้ค่อย ๆ เรียนรู้จากข้อมูล และในแบบจำลองบางชนิดก็ต้องการฮาร์ดแวร์หน่วยประมวลผลกราฟิกส์ (Graphics Processing Unit, GPU) เฉพาะทางด้วยจึงจะสามารถทำงานได้รวดเร็วแบบเรียลไทม์ ไม่เพียงเท่านี้ระบบจำพวกการเรียนรู้ของเครื่องยังมีลักษณะพื้นฐานที่ทำให้ปรับตัวต่อการเปลี่ยนแปลงสำหรับข้อมูลใหม่ ๆ ข่าวปลอมใหม่ ๆ ที่ไม่เคยเห็นมาก่อนได้ยาก (เนื่องจากระบบไม่เคยถูกสอนมาด้วยข้อมูลเหล่านี้) และการอัปเดตระบบให้เท่าทันข่าวปลอมปัจจุบันในบางครั้งอาจจำเป็นต้องนำข้อมูลทั้งหมดมาเริ่มต้นสอนแบบจำลองใหม่ตั้งแต่ศูนย์ จึงไม่สะดวกนักเมื่อพิจารณาถึงสถานการณ์ปัจจุบันที่การรับส่งข่าวสารเป็นไปอย่างรวดเร็ว มาไวไปไว และข่าวสารต่าง ๆ มีช่วงอายุที่ค่อนข้างสั้น (โดยเฉพาะเมื่อเทียบกับระยะเวลาที่ผู้พัฒนาต้องใช้ในการรวบรวมข้อมูล ตระเตรียมข้อมูล สอนและทดสอบแบบจำลองตัวใหม่)

จากข้อจำกัดของระบบการเรียนรู้ด้วยเครื่องที่กล่าวไป รวมถึงลักษณะของข่าวสารในปัจจุบันที่มีพลวัติสูง หลายข่าวมีความกำกวมในข้อเท็จจริงเป็นอย่างมาก จนทำให้ไม่มีใครหรือหน่วยงานใดหน่วยงานหนึ่งสามารถชี้ขาดข่าวจริงหรือข่าวปลอมได้อย่างเบ็ดเสร็จ เนื่องจากข่าวหนึ่ง ๆ สามารถเปลี่ยนสถานะไปมาระหว่างการเป็นข่าวจริงและข่าวปลอมได้ตามแต่เวลาและเหตุปัจจัย ณ ขณะหนึ่ง ๆ ในงานนี้ผู้วิจัยจึงเลือกจะทำการทดลองสร้างระบบช่วยคัดกรองข่าวปลอมที่ไม่พึ่งพาเทคนิคการเรียนรู้ของเครื่อง แต่จะอาศัยการรวบรวมข่าวปลอมจากแหล่งข้อมูลออนไลน์ต่าง ๆ มาเก็บไว้ในฐานข้อมูล แล้วใช้เทคนิคการประมวลผลภาษาธรรมชาติมาช่วยเปรียบเทียบระหว่างข่าวปลอมในฐานข้อมูลและข้อมูลอินพุตจากผู้ใช้ เพื่อแสดงผลลัพธ์ซึ่งไม่ใช่การตัดสินชี้ขาดว่าข่าวนั้นจริงหรือปลอมแต่เป็นการนำเสนอข้อความและลิงก์ที่เกี่ยวข้องสำหรับเป็นข้อมูลเพิ่มเติมให้ผู้ใช้ใช้ในการประกอบการตัดสินใจด้วยตนเองว่าข่าวนี้น่าจะเป็นข่าวจริงหรือข่าวปลอม ลักษณะนี้จะทำให้ได้ระบบต้นแบบที่ไม่ซับซ้อนมาก ประมวลผลไว ไม่ต้องการทรัพยากรฮาร์ดแวร์การคำนวณใดเป็นพิเศษ และสามารถนำไปเริ่มใช้งานได้จริงทันที อีกทั้งสามารถปรับเปลี่ยนผลลัพธ์การตรวจสอบข่าวปลอมไปตามข้อมูลข่าวปลอมใหม่ ๆ ที่ผู้วิจัยใส่เพิ่มเข้าไปในฐานข้อมูลได้ทันทีด้วย

3. ระเบียบวิธีการวิจัย

3.1 รวบรวมชุดข้อมูลข่าวปลอมภาษาไทย

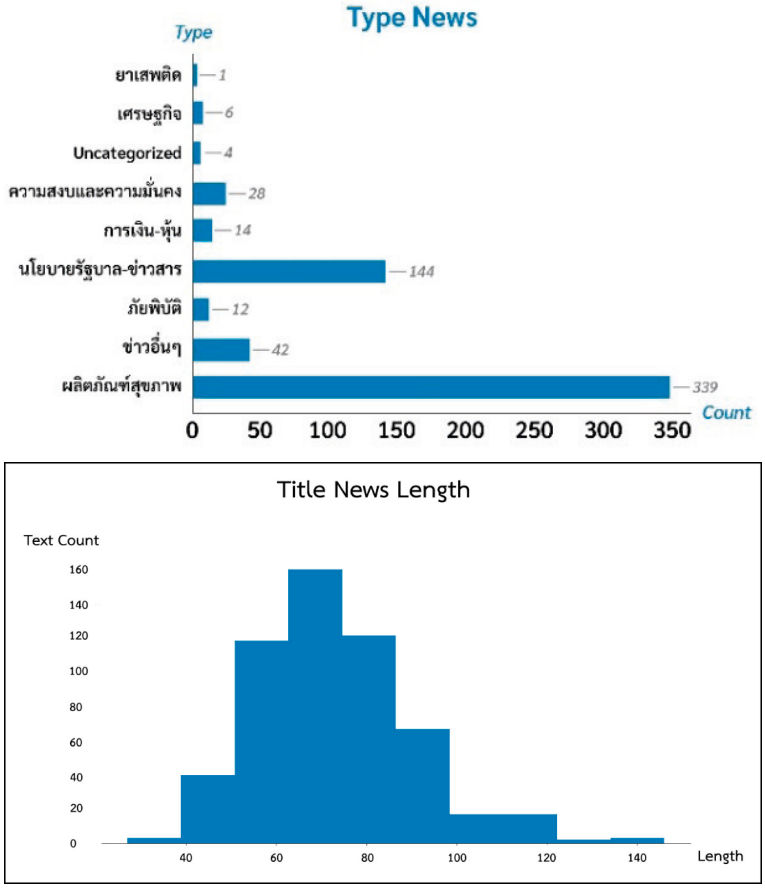
ฐานข้อมูลข่าวปลอมของงานวิจัยชิ้นนี้เลือกใช้ Google BigQuery ซึ่งเป็นบริการของ Google Cloud Platform ที่แม้จะรองรับการทำงานกับภาษา SQL แต่เบื้องหลังมีการออกแบบที่เป็นเอกลักษณ์ของ NoSQL ทำให้การประมวลผลกับข้อมูลจำนวนมากของ Google BigQuery มีความรวดเร็วมากขึ้นอย่างมีนัยสำคัญ สำหรับข้อมูลข่าวปลอมภาษาไทยที่ใช้ในการสร้างฐานข้อมูลของงานวิจัยนี้ ผู้วิจัยทำการรวบรวมมาจากเว็บไซต์ศูนย์ต่อต้านข่าวปลอมประเทศไทย (www.antifakenewscenter.com) ภายในช่วงระยะเวลาของการนำเสนอข่าวตั้งแต่ 10 ธันวาคม ค.ศ. 2017 ถึง 5 ธันวาคม ค.ศ. 2020 รวมเป็นปริมาณข้อมูลทั้งสิ้น 15 เมกะไบท์ โดยจะเลือกเฉพาะข่าวที่ถูกจัดอยู่ในหมวดข่าวปลอมมาเก็บลงในฐานข้อมูลเท่านั้น อีกทั้งผู้วิจัยจะทำการเพิ่มข่าวปลอมใหม่ ๆ ลงฐานข้อมูลทุกวันในช่วงเวลา 8:00 นาฬิกา ทั้งนี้จำนวนข่าวปลอมที่ถูกบันทึกเพิ่มเติมต่อวันจะมีปริมาณมากน้อยแตกต่างกันไปตามปริมาณข้อมูลที่พบจากเว็บไซต์ต้นทางดังกล่าว รูปที่ 1 แสดงตัวอย่างของข่าวปลอมที่ถูกเก็บอยู่ในฐานข้อมูลของระบบต้นแบบ โดยแต่ละข่าวจะประกอบไปด้วย หัวข้อข่าว รายละเอียดข่าวทั้งหมด วันที่ที่ข่าวนี้ถูกนำเสนอในสื่อต้นทาง ตำแหน่งที่ตั้งไฟล์รูปภาพข่าว หัวข้อข่าวแบบแยกคำ และรายละเอียดข่าวแบบสรุป



รูปที่ 1 แผนภาพแสดงการออกแบบฐานข้อมูลข่าวปลอม และตัวอย่างข่าวปลอมที่ถูกเก็บไว้ในฐานข้อมูล

สำหรับการวิเคราะห์ข่าวปลอมของงานวิจัยนี้ ผู้วิจัยเลือกเฉพาะ “หัวข้อข่าว” ดังที่ปรากฏอยู่ในเว็บไซต์มาวิเคราะห์เท่านั้น ไม่ได้นำเนื้อหาทั้งหมดของข่าวปลอมแต่ละข่าวมาวิเคราะห์ด้วยทั้งหมด โดยจากการสำรวจข้อมูลเบื้องต้น จำนวนข่าวปลอมที่ผู้วิจัยรวบรวมมาได้คือ 725 ข่าว แบ่งเป็น 9 หมวดหมู่ดังแสดงในรูปที่ 2 (บน) จากรูปจะเห็นได้ชัดว่าบางหมวดหมู่มีจำนวนข่าวปลอมที่แตกต่างกันอย่างมีนัยสำคัญ เพื่อให้ได้ข้อมูลในปริมาณที่เพียงพอต่อการพัฒนาและ

ทดสอบระบบ และเพื่อหลีกเลี่ยงปัญหาความไม่สมดุลกันของปริมาณข่าวในหมวดต่าง ๆ ซึ่งอาจนำไปสู่ผลลัพธ์การทำงานที่ไม่เป็นธรรม ในงานวิจัยนี้ผู้วิจัยจึงเลือกข่าวปลอมในหมวด “ผลิตภัณฑ์สุขภาพ” ซึ่งมีจำนวนมากที่สุดมาใช้เป็นหลัก นอกจากนี้เมื่อวิเคราะห์ข้อมูลข่าวปลอมที่รวบรวมมาลึกลงไปอีก ผู้วิจัยก็พบว่าหัวข้อข่าวของแต่ละข่าวปลอมนั้นมีความยาวของหัวข้อข่าว (นับเป็นตัวอักษร) แตกต่างกันดังรายละเอียดในรูปที่ 2 (ล่าง) โดยความยาวเฉลี่ยของหัวข้อข่าวปลอมทั้งหมดในฐานข้อมูลคือ 72 ตัวอักษร



รูปที่ 2 ผลลัพธ์การสำรวจข่าวปลอมทั้งหมดที่รวบรวมมาสำหรับเก็บในฐานข้อมูล บนคือกราฟแสดงจำนวนของข่าวปลอมภาษาไทยที่รวบรวมมาได้ในแต่ละหมวดหมู่ และ ล่างคือกราฟแสดงความยาว (หน่วย: ตัวอักษร) ของหัวข้อข่าวทั้งหมด

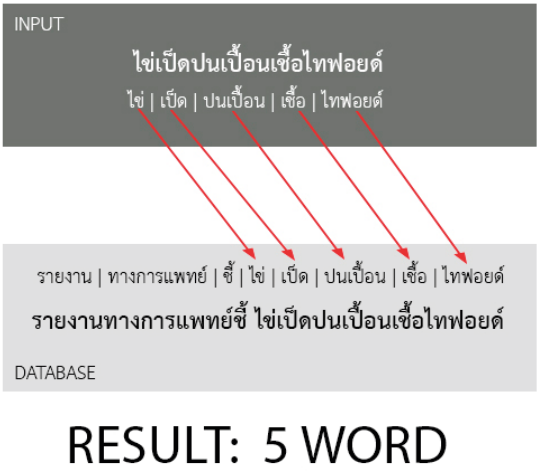
3.2 ตรวจสอบข่าวปลอมโดยใช้รูปประโยคทั้งหมด

หัวข้อ 3.2 นี้เป็นการทดลองแรกสำหรับขั้นตอนวิธีที่ผู้วิจัยจะใช้ในการตรวจสอบว่าข่าวภาษาไทยหนึ่ง ๆ ซึ่งเป็นอินพุตที่ได้รับจากผู้ใช้นั้นมีแนวโน้มที่จะเป็นข่าวปลอมหรือไม่ ทั้งนี้โดยจะไม่มีการจำกัดความยาวของอินพุตหรือก็คือผู้ใช้สามารถใส่อินพุตเข้ามายาวหรือสั้นแค่ไหนก็ได้ แล้วระบบจะนำข้อความทั้งหมดที่ผู้ใช้อินพุตเข้ามาไปผ่านกระบวนการเตรียมข้อมูล (pre-processing) อันได้แก่ กระบวนการตัดคำจากประโยค (Word tokenization) ดังตัวอย่างในรูปที่ 3 จากนั้นจึงตามด้วยกระบวนการลบคำซ้ำ และกระบวนการลบช่องว่างของคำ ตามลำดับ เมื่อผ่านขั้นตอนการเตรียมข้อมูลแล้วผลที่ได้คือคำภาษาไทยจำนวน N คำใด ๆ ที่แตกต่างกัน

ทั้งหมด (ค่า N นี้อาจเหมือนหรือแตกต่างกันไปได้ในแต่ละข่าว) โดยหัวข้อข่าวปลอมทุกหัวข้อในฐานข้อมูล เช่นเดียวกับอินพุตจากผู้ใช้ ล้วนแต่ถูกนำมาผ่านกระบวนการนี้เหมือนกันทั้งสิ้น

รูปที่ 3 ตัวอย่างการนำ PyThaiNLP มาใช้ในการตัดคำ

เมื่อข้อมูลในฐานข้อมูลข่าวปลอมและข้อมูลอินพุตจากผู้ใช้ถูกเปลี่ยนให้อยู่ในรูปเซตของคำดังที่กล่าวไปข้างต้นแล้ว ผู้วิจัยจึงใช้วิธีการของ Term Frequency (แบบ Raw Count) เพื่อนับว่าข่าวปลอมแต่ละข่าวในฐานข้อมูลมีจำนวนคำซ้ำกับคำในอินพุตเป็นจำนวนกี่คำ รูปที่ 4 แสดงตัวอย่างการค้นหาข่าวปลอมในฐานข้อมูลด้วยเทคนิค Term Frequency โดยในตัวอย่างนี้อินพุตจากผู้ใช้ถูกนำไปผ่านกระบวนการต่าง ๆ จนเหลือเพียงเซตของคำ 5 คำซึ่งแตกต่างกันทั้งหมด ในทำนองเดียวกันหัวข้อข่าวในฐานข้อมูลก็ผ่านกระบวนการเดียวกันจนเหลือเพียงเซตของคำ 8 คำที่แตกต่างกัน ลูกศรสีแดงแสดงคำในอินพุตที่ปรากฏอยู่ในข่าวปลอม กล่าวคือในตัวอย่างนี้สามารถนับคะแนนความเหมือนของข่าว (เทียบกับอินพุต) ได้เท่ากับ 5 คะแนนเนื่องจากมีคำ 5 คำที่ตรงกัน



RESULT: 5 WORD

รูปที่ 4 ตัวอย่างการค้นหาข่าวปลอมในฐานข้อมูลด้วยเทคนิค Term Frequency

การประมวลผลภาษาไทยทั้งหมดที่ผู้วิจัยกล่าวถึงในสองย่อหน้าก่อนนี้ เป็นการประยุกต์ใช้ฟังก์ชันที่อยู่ในไลบรารี PyThaiNLP [6] เวอร์ชัน 2.25 ทั้งสิ้น ผลลัพธ์การเปรียบเทียบอินพุตจากผู้ใช้และข่าวปลอมในฐานข้อมูลจะแสดงออกมาในรูปแบบของข่าวปลอม 3 อันดับในฐานข้อมูลที่มีคะแนนความเหมือนเรียงลำดับจากมากไปน้อยดัง

แผนภาพสรุปในรูปที่ 5 อย่างไรก็ตามเนื่องจากระบบต้นแบบ ณ ปัจจุบันยังไม่มีความสามารถในการชี้ขาดว่าข่าวปลอมจากฐานข้อมูลที่ถูกคัดเลือกมา 3 อันดับนี้ใช่หรือไม่ใช่ข่าวเดียวกันกับที่ผู้ใช้ป้อนอินพุตเข้ามา ณ จุดนี้ยังต้องอาศัยผู้ใช้พิจารณาหัวข้อของข่าวปลอมทั้ง 3 อันดับเอง โดยในกรณีที่อินพุตเป็นข่าวที่ไม่มีเก็บอยู่ในฐานข้อมูลข่าวปลอมของระบบ ผลลัพธ์ข่าวปลอม 3 อันดับแรกที่ถูกระบบเลือกออกมาก็มักจะเป็นข่าวที่ไม่มีความเกี่ยวข้องกันกับข่าวอินพุตอย่างเห็นได้ชัด ผู้ใช้จึงสามารถอนุมานได้ด้วยตนเองว่าระบบไม่พบข่าวอินพุตนี้อยู่ในฐานข้อมูลข่าวปลอม

รูปที่ 5 แผนภาพสรุปการทำงานของระบบในการตรวจสอบอินพุตเทียบกับข่าวปลอมในฐานข้อมูล

3.3 ตรวจสอบข่าวปลอมจากรูปประโยคโดยใช้เพียงคำนามและลักษณนามเท่านั้น

ขั้นตอนวิธีการที่นำเสนอในหัวข้อ 3.2 นั้น ในกรณีที่อินพุตเป็นข้อความยาว ๆ จนทำให้ข้อความถูกตัดแบ่งออกเป็นเซตขนาดใหญ่ที่มีคำแตกต่างกันอยู่เป็นจำนวนมาก คำจำนวนมากเหล่านี้จะส่งผลกระทบทำให้ระบบไม่สามารถค้นหาได้แบบเจาะจง เนื่องจากคำเหล่านี้สามารถไปซ้ำกับคำในหัวข้อข่าวในฐานข้อมูลข่าวปลอมได้ด้วยความถี่ (term frequency) สูง จนทำให้ระบบเข้าใจผิดไปว่าอินพุตนั้น ๆ เป็นข่าวเดียวกันกับข่าวปลอมในฐานข้อมูล ดังตัวอย่างแสดงในรูปที่ 6 จะเห็นว่าคำว่า 'อาหาร' 'แล้ว' 'ของ' 'โดย' 'ได้' 'ไม่' 'คน' เป็นคำพบบ่อยที่ไม่สื่อความหมายถึงใจความสำคัญของข่าว แต่ด้วยวิธีการในหัวข้อ 3.2 จึงทำให้คำเหล่านี้ถูกนำไปนับรวมเป็นคะแนนความเหมือนระหว่างอินพุตและข่าวปลอมในฐานข้อมูลได้ ทั้งนี้ผู้วิจัยพบว่าประเภทของคำที่ส่งผลให้เกิดการค้นหาผิดเพี้ยนลักษณะนี้มากที่สุดคือ คำสันธาน (Conjunction) และคำบุพบท (Preposition)

INPUT

บอกต่อ..เผื่อได้บุญ กินมะนาว 2 ลูก ต่อ โซดา 1 ขวด ผสมกัน กินเช้าหรือกลางวันด้วย ก็ดีคับ...กรดของมะนาวจะฆ่าเซล โรคมะเร็งโดยตรง องค์กรณ์อนามัยโลก

'เป็นความลับ' 'เนื้อสัตว์' 'โรคมะเร็ง' 'วิทยาทาน' 'เป็นหลัก' 'กลางวัน' 'บอกต่อ' 'มะเร็ง' 'ส่งต่อ' 'องค์กร' 'อนามัย' 'เรื่อง' 'โดยตรง' 'มะนาว' 'อาหาร'....'แล้ว' 'ของ' 'โดย' 'ได้' 'ไม่' 'คน'

DATABASE

เก็บภาษีที่ดินไม่บอก...ดินทางไปจ่าย	[ต้อง, หรือ, ของ, และ, ไม่, ไป]
คนแรก...โดยไม่ผิดกฎหมายได้แล้ว	[แล้ว, ของ, โดย, ได้, ไม่, คน]
กินอาหารค้างคืนที่นำไปอุ่น..มะเร็ง	[โรคมะเร็ง, อาหาร, เป็น, กิน, ที่, ไป]
....	[... , ... , ...]

รูปที่ 6 ตัวอย่างปัญหาที่พบในขั้นการตรวจสอบข่าวปลอมโดยใช้รูปประโยคทั้งหมด อักษรสีแดงตัวหนาหมายถึงคำที่มักพบได้บ่อยในประโยค ซึ่งทำให้เกิดปัญหาในการตรวจสอบข่าวปลอมด้วยเทคนิคในหัวข้อ 3.2

เพื่อแก้ข้อจำกัดของขั้นตอนวิธีในหัวข้อ 3.2 ในหัวข้อที่ 3.3 นี้ผู้วิจัยจึงทำการพัฒนาต่อยอดไปอีก โดยภายหลังจากการนำอินพุตไปผ่านกระบวนการตัดคำ ลบคำซ้ำ และลบช่องว่างตามปกติแล้ว ผู้วิจัยจะนำเซตของคำที่ได้ไปผ่านกระบวนการแยกประเภทของคำด้วยเทคนิค Part of Speech Tagging (POS Tags) ใน PyThaiNLP อีกดังตัวอย่างในรูปที่ 7 โดยจุดประสงค์ของขั้นตอนที่เพิ่มมานี้ก็เพื่อจะเลือกเอาเฉพาะคำนาม (noun) และลักษณนาม (classifier) มาใช้ในการนับค่า Term Frequency ต่อไป เหตุผลที่ผู้วิจัยเลือกเฉพาะคำนามและลักษณนามมาใช้ในการค้นหาข่าวในฐานข้อมูลนั้น เนื่องจากประโยคภาษาไทยมีการใช้คำนามและลักษณนามที่หลากหลายแตกต่างกัน และคำสองประเภทนี้มีการซ้ำของคำน้อยกว่าคำประเภทชนิดอื่น หรืออาจกล่าวได้ว่าคำนามและลักษณนามในภาษาไทยนั้นมีเอกลักษณ์ที่ค่อนข้างชัดเจน เหมาะสมที่จะนำคำมาใช้เพื่อจำกัดขอบเขตของการค้นหาข่าวให้แคบลงและลดการค้นหาด้วยคำที่ไม่จำเป็นหรือคำที่ขาดเอกลักษณ์ออกไป

รูปที่ 7 ตัวอย่างการแบ่งประเภทของคำ

4. ผลการทดลอง

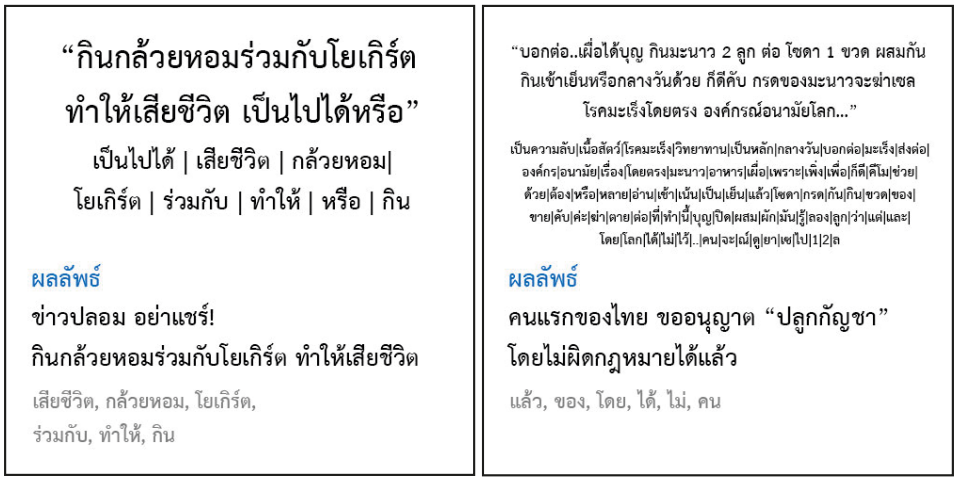
สำหรับการประเมินความถูกต้องในการทำงานของระบบต้นแบบที่สร้างขึ้นนั้น ด้วยปัจจุบันยังไม่พบว่ามีตัวชี้วัดมาตรฐานหรือชุดข้อมูลมาตรฐานสำหรับให้นักวิจัยและพัฒนาใช้ในการทดลองระบบตรวจสอบข่าวจริงหรือข่าวปลอมในภาษาไทย ในการนี้ผู้วิจัยจึงทำการเก็บรวบรวมข้อมูลข่าวและบทความเพิ่มเติมสำหรับใช้ทดสอบระบบต้นแบบนี้ขึ้นมาเอง

โดยเฉพาะ โดยเป็นข่าวจริงและข่าวปลอมจำนวนอย่างละ 60 ข่าวเท่ากัน ทั้งนี้ข่าวสำหรับทดสอบที่รวบรวมมาได้นั้นมีตั้งแต่ข่าวที่มีจำนวนตัวอักษรน้อยระหว่าง 100 – 300 ตัวอักษรและมีจำนวนตัวอักษรมากระหว่าง 600 – 1,000 ตัวอักษร โดยข่าวสำหรับทดสอบทั้งหมดนำมาจากเว็บไซต์โคแฟค (https://cofact.org/) และ จากเฟซบุ๊กแฟนเพจ “อ๋อ มันเป็นอย่างนี้นี่เอง by อาจารย์เจษฎ์” ซึ่งแม้จะเป็นคนละแหล่งข่าวกันกับที่ผู้วิจัยใช้ในการสร้างฐานข้อมูล (ในหัวข้อ 3.1) แต่ก็ปรากฏมีข้อมูลที่ซ้ำกันอยู่ประมาณ 5%

นอกจากนี้ในส่วนของการประเมินความสามารถของระบบด้วย Confusion Matrix ด้วยข้อจำกัดของระบบปัจจุบันที่ยังไม่สามารถชี้ขาดได้ว่าข่าวอินพุตนั้นใช่หรือไม่ใช่ข่าวเดียวกันกับข่าวปลอม 3 อันดับที่ระบบเลือกมาจากฐานข้อมูล ดังนั้นผู้วิจัยจึงจะใช้วิธีการประเมินโดยบุคคลช่วยในขั้นตอนนี้ โดยหากผู้ใช้มีความเห็นว่าข่าวที่อินพุตเข้าไปไม่มีความสอดคล้องหรือไม่ใช่ข่าวเดียวกันกับข่าวปลอม 3 อันดับที่ได้จากระบบเลย ในกรณีนี้จะถือว่าระบบทายว่าข่าวนี้เป็น “ข่าวจริง” แต่ในทางตรงกันข้าม หากผู้ใช้เห็นว่าข่าวที่อินพุตเข้าไปนั้นอ่านดูแล้วเป็นข่าวเดียวกันกับหนึ่งในสามข่าวปลอมที่ระบบค้นหามาให้ ในกรณีนี้จะถือว่าระบบทายว่าข่าวนี้เป็น “ข่าวปลอม”

4.1 การทดลองตรวจสอบข่าวปลอมโดยใช้รูปประโยคทั้งหมด

สำหรับระบบที่สร้างขึ้นด้วยขั้นตอนวิธีในหัวข้อ 3.2 นั้นได้ผลลัพธ์การประเมิน Confusion Matrix คือ true positive=40 (จำนวนข่าวปลอมที่ถูกระบบทำนายว่าเป็นข่าวปลอม), false positive=4 (จำนวนข่าวจริงที่ถูกระบบทำนายผิดไปเป็นข่าวปลอม), true negative=56 (จำนวนข่าวจริงที่ถูกระบบทำนายว่าเป็นข่าวจริง) และ false negative=20 (จำนวนข่าวจริงที่ถูกระบบทำนายผิดไปเป็นข่าวปลอม) คิดเป็นค่า accuracy=80%, precision=90.91%, recall=66.67% และ F1score=76.92% หรือก็คือในกรณีที่ระบบทำนายว่าอินพุตหนึ่ง ๆ เป็นข่าวปลอมนั้น คำทำนายนี้จะเชื่อถือได้มากถึงประมาณ 90.91% แต่จะมีข่าวปลอมเพียงประมาณ 66.67% ของข่าวปลอมทั้งหมดเท่านั้นที่ระบบจะทำนายได้ถูกต้องว่าเป็นข่าวปลอม (ระบบหาอินพุตนี้เจอในฐานข้อมูลข่าวปลอม) ในขณะที่ข่าวปลอมที่เหลืออีกจำนวนกว่า 33.33% นั้นระบบจะทำนายผิดไปเป็นข่าวจริง (หรือก็คือระบบหาอินพุตนี้ในฐานข้อมูลข่าวปลอมไม่พบนั่นเอง) รูปที่ 8 แสดงผลการทดลองใน 2 กรณีทั้งกรณีที่อินพุตเป็นข้อความสั้นและอินพุตเป็นข้อความยาว โดยจะเห็นว่าในกรณีที่อินพุตเป็นข้อความยาวนั้นหัวข้อข่าวที่เป็นผลลัพธ์การค้นหาอันดับแรกแทบไม่มีความเกี่ยวข้องใด ๆ กับข้อความในข่าวอินพุตเลย หรือก็คือระบบที่สร้างขึ้นด้วยขั้นตอนวิธีในหัวข้อ 3.2 ทำนายว่าข่าวมะนาวโชดารักษามะเร็งในรูปขวานั้นเป็น “ข่าวจริง” ซึ่งถือเป็นการทำนายที่ผิดพลาดอันเกิดจากข้อจำกัดของขั้นตอนวิธีนี้ดังที่ผู้วิจัยอธิบายไปแล้ว (ในตอนต้นของหัวข้อ 3.3)



รูปที่ 8 ผลลัพธ์ข่าวปลอมอันดับ 1 ที่ได้จากการตรวจสอบข่าวปลอมโดยใช้รูปประโยคทั้งหมด ซ้ายคือกรณีที่อินพุตเป็นข้อความสั้น และ ขวาคือกรณีที่อินพุตเป็นข้อความยาว

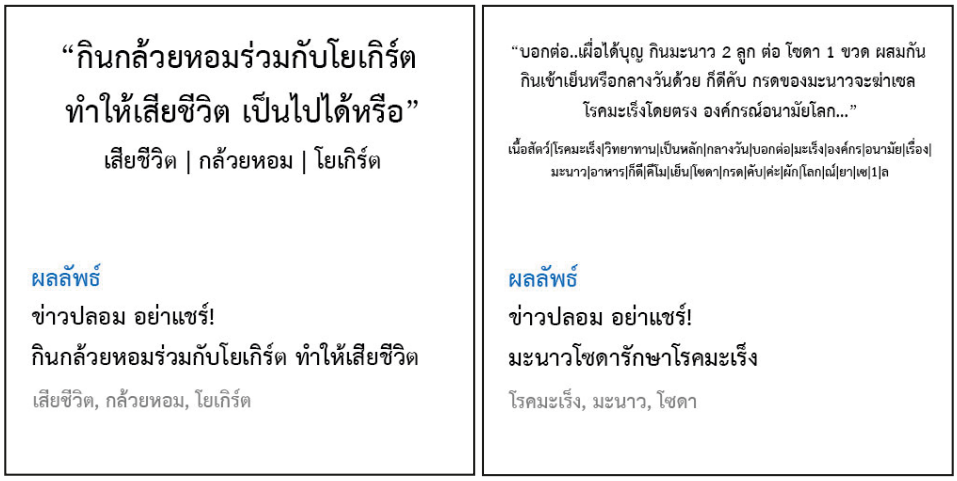
4.2 การทดลองตรวจสอบข่าวปลอมจากรูปประโยคโดยใช้เฉพาะคำนามและลักษณนามเท่านั้น

สำหรับระบบที่สร้างขึ้นด้วยขั้นตอนวิธีปรับปรุงในหัวข้อ 3.3 นั้นได้ผลลัพธ์การประเมิน Confusion Matrix คือ true positive=45 (จำนวนข่าวปลอมที่ถูกระบบทำนายว่าเป็นข่าวปลอม), false positive=4 (จำนวนข่าวจริงที่ถูกระบบทำนายผิดไปเป็นข่าวปลอม), true negative=56 (จำนวนข่าวจริงที่ถูกระบบทำนายว่าเป็นข่าวจริง) และ false negative=15 (จำนวนข่าวจริงที่ถูกระบบทำนายผิดไปเป็นข่าวปลอม) คิดเป็นค่า accuracy=84.17%, precision=91.84%, recall=75.00% และ F1score=82.57% โดยพิจารณาจากตารางที่ 1 จะเห็นว่าจำนวน false positive ลดลงทำให้ค่า recall ของขั้นตอนวิธีในหัวข้อ 3.3 สูงขึ้นกว่าของขั้นตอนวิธีในหัวข้อ 3.2 (recall=66.67%) กล่าวคือ เทคนิคการคัดเลือกเฉพาะคำนามและลักษณนามมาใช้เพื่อค้นหาข่าวปลอมในฐานข้อมูลนั้น ทำให้ระบบต้นแบบมีประสิทธิภาพเพิ่มขึ้น โดยในกรณีที่ระบบทำนายว่าอินพุตหนึ่ง ๆ เป็นข่าวปลอมนั้น คำทำนายนี้จะเชื่อถือได้มากถึง 91.84% และจะมีข่าวปลอมเพียงประมาณ 25% ของข่าวปลอมทั้งหมดเท่านั้นที่ระบบจะทำนายผิดไปว่าเป็นข่าวจริง ในขณะที่ข่าวปลอมที่เหลืออีกกว่า 75% นั้นระบบจะทำนายได้ถูกต้องทั้งหมด

ตารางที่ 1 เปรียบเทียบผลการทดลอง โดยตัวหนาหมายความถึงผลลัพธ์ที่ดีที่สุดสำหรับตัวชี้วัดในแถวนั้น ๆ

	การทดลองตรวจสอบข่าวปลอม โดยใช้รูปประโยคทั้งหมด (หัวข้อที่ 4.1)	**การทดลองตรวจสอบข่าวปลอม จากรูปประโยคโดยใช้เฉพาะคำนาม และลักษณนามเท่านั้น (หัวข้อที่ 4.2)**
True positive	40	45
False positive	4	4
True negative	56	56
False negative	20	15
Accuracy	80%	**84.17%**
Precision	90.91%	**91.84%**
Recall	66.67%	**75.00%**
F1 score	76.92%	**82.57%**

รูปที่ 9 แสดงผลการทดลองใน 2 กรณีทั้งกรณีที่อินพุตเป็นข้อความสั้นและอินพุตเป็นข้อความยาว โดยในรูปขวาจะเห็นว่าแม้ข้อความอินพุตจะยาว แต่กลุ่มของคำที่ถูกแยกและเลือกออกมาเฉพาะคำนามและลักษณนามก็เหลือเพียงไม่กี่คำ ทำให้ได้ผลลัพธ์ข่าวอันดับแรกที่เห็นได้ชัดเจนว่าเป็นข่าวเดียวกันกับอินพุตที่ได้รับมาจริง ๆ กล่าวคือข่าวมะนาวโซดารักษามะเร็งนั้น ถูกตรวจพบโดยระบบต้นแบบในงานวิจัยนี้ว่าเป็น “ข่าวปลอม”



รูปที่ 9 ผลลัพธ์ข่าวปลอมอันดับ 1 ที่ได้จากการตรวจสอบข่าวปลอมโดยใช้เฉพาะคำนามและลักษณนามเท่านั้น ซ้ายคือกรณีที่อินพุตเป็นข้อความสั้น และ ขวาคือกรณีที่อินพุตเป็นข้อความยาว

5. วิจารณ์และสรุปผลการทดลอง

จากการทดลองทั้ง 2 การทดลองในหัวข้อที่ 4 ผู้วิจัยพบว่ารูปแบบแรก (หัวข้อ 4.1) ของการตรวจสอบข่าวปลอมโดยใช้รูปประโยคทั้งหมดนั้น สามารถแสดงข่าวในฐานข้อมูลที่เกี่ยวข้องได้ถูกต้องเพียงบางส่วน แต่ไม่สามารถใช้ตรวจสอบอินพุตข่าวที่มีความยาวหลายบรรทัดได้ เนื่องจากปริมาณคำที่เหลือภายหลังการตัดคำมีจำนวนมากและมีอัตลักษณ์ต่ำจนไม่สามารถจับคู่กับข่าวเดียวกันที่มีปรากฏอยู่ในฐานข้อมูลข่าวปลอมได้ ในส่วนของรูปแบบที่สอง (หัวข้อ 4.2) ของการตรวจสอบข่าวปลอมโดยใช้เฉพาะคำนามและลักษณนามนั้น ผู้วิจัยได้ผลลัพธ์การค้นพบข่าวปลอมที่แม่นยำขึ้น แต่ก็ยังมีข้อจำกัดที่ไม่สามารถตรวจสอบประโยคอินพุตข่าวที่สั้นมากได้ อย่างไรก็ตามทั้งสองวิธีมีจุดอ่อนเดียวกันคือไม่สามารถตรวจสอบข่าวที่ไม่มีปรากฏอยู่ในฐานข้อมูลของงานวิจัยนี้ได้ เพื่อแก้ข้อจำกัดนี้ในอนาคตผู้วิจัยมีแนวคิดจะเพิ่มการเก็บข่าวในฐานข้อมูล โดยอาจเก็บทั้งข่าวจริงและข่าวปลอมแยกกันเพื่อให้ข่าวในฐานข้อมูลมีขอบข่ายกว้างขวางครอบคลุมมากขึ้น และผู้วิจัยสามารถจะนำผลการเปรียบเทียบทั้งในหมวดข่าวจริงและข่าวปลอมมาร่วมกันพิจารณาได้

ในลำดับถัดไปผู้วิจัยมีแนวคิดที่จะพัฒนาระบบให้มีความฉลาดมากขึ้นในอีกหลาย ๆ ด้านโดยใช้เทคนิคประมวลผลภาษาธรรมชาติอื่น ๆ เพื่อให้สามารถวิเคราะห์ความหมายและรูปแบบอารมณ์ของข่าวได้และที่สำคัญคือระบบในอนาคตจะต้องสามารถระบุชัดเจนได้ว่าข่าวแต่ละข่าวที่ระบบเลือกออกมาให้นั้น เป็นข่าวเดียวกับอินพุตหรือไม่ด้วยระดับความมั่นใจที่เท่าไหร่ ทั้งนี้ในส่วนของปัญหาการตรวจสอบข้อความในข่าวที่สั้นมากและมีคำนามน้อยจนอาจทำให้การตรวจสอบเฉพาะคำนามและลักษณนามไม่แม่นยำเพียงพอ ในกรณีนี้อาจพิจารณานำคำกริยาในประโยคมาพิจารณาร่วมด้วยเพื่อเพิ่มความแม่นยำ หรืออาจนำเนื้อความในข่าวบางส่วนหรือทั้งหมดมาทำเป็นเวกเตอร์ตัวแทนของเนื้อความข่าวและใช้เทคนิคการวัดความคล้ายคลึงระหว่างเวกเตอร์ตัวแทนเอกสาร (อาทิ Cosine similarity) มาใช้ประกอบกันในการพิจารณาเทียบข่าวปลอม นอกจากนี้ผู้วิจัยยังมีแนวคิดในการพิจารณาประเภทของคำในข้อความข่าวเพิ่มเติม อาทิ คำบุพบท สันธาน คำปฏิเสธ คำช่วยหน้ากริยา คำช่วยหลังกริยา คำศัพท์ที่พบบ่อยเกินไป รวมถึงกลุ่มคำในกลุ่ม stop list เพื่อทำการทดลองเปรียบเทียบกันบนชุดข้อมูลทดสอบที่มีจำนวนมากขึ้นว่าส่วนผสมของการพิจารณาแบบใดที่จะให้ประสิทธิภาพและประสิทธิผลดีที่สุดสำหรับการตรวจสอบข่าวปลอมภาษาไทย สุดท้ายผู้วิจัยมีความสนใจที่จะเพิ่มช่องทาง LINE official และเว็บไซต์สำหรับตรวจสอบข่าวปลอมให้ผู้ใช้งานคนไทยสามารถเข้าถึงและใช้งานได้สะดวกขึ้น

6. เอกสารอ้างอิง

1. Vosoughi, S., Roy, D. and Aral, S., 2018, "The Spread of True and False News Online," *Science,* 359 (6380), pp. 1146-1151. https://doi.org/10.1126/science.aap9559

2. Chulrod, P. and Nontakhamchan, P., 2020, "Prediction Model of the Fake News from Online Social Media with Data Mining," *College of Asia Journal*, 10 (4), pp. 121-128. (In Thai)

3. Srivastava, A., 2020,), "Real Time Fake News Detection Using Machine Learning and NLP," *International Research Journal of Engineering and Technology (IRJET)*, 7 (6), pp. 3683-3679.

4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, T. and Polosukhin, I., 2017, "Attention is All You Need," *The 31st International Conference on Neural Information Processing Systems (NIPS'17)*, Long Beach, California, pp. 6000-6010.

5. Thota, A., Tilak, P., Ahluwalia, S. and Lohia, N., 2018, "Fake News Detection: A Deep Learning Approach," *SMU Data Science Review*, 1 (3), Article No. 10.

6. Phatthiyaphaibun, W., Chaovavanich, K., Polpanumas, C., Suriyawongkul, A., Lowphansirikul, L. and Chormai, P., 2016, "PyThaiNLP: Thai Natural Language Processing in Python," Zenodo. http://doi.org/10.5281/zenodo.3519354