การใช้โมเดลการทำนายข่าวปลอมจากสื่อสังคมออนไลน์ด้วยเทคนิคการทำเหมืองข้อมูล
Prediction Model of the Fake News from Online Social Media with Data Mining

พิเชษฎ์ จุลรอด[1],
ผุสดี นนทคำจันทร์[2]

บทคัดย่อ

ข่าวปลอม (Fake News) มีจำนวนเพิ่มมากขึ้นบนสื่อสังคมออนไลน์โดยข่าวปลอมดังกล่าวมีผลทำให้ทัศนคติและการตัดสินใจในเรื่องใดเรื่องหนึ่งมีผลกระทบรุนแรงต่อการอยู่ร่วมกันในสังคม โดยพื้นฐานของข่าวปลอมจะเน้นเนื้อหาที่ตื่นเต้นและกระตุ้นอารมณ์ของผู้บริโภคจึงทำให้เกิดการแพร่กระจายได้ง่ายและรวดเร็วกว่าข่าวจริง ในการตรวจสอบหาต้นตอและแหล่งข่าวเพื่อระบุว่าเป็นข่าวจริงหรือปลอมนั้นจำเป็นต้องใช้เวลาในการค้นหา การใช้โมเดลในการทำนายข่าวปลอมจากสื่อสังคมออนไลน์ด้วยเทคนิคการทำเหมืองข้อมูลนั้นจึงเป็นแนวทางในการช่วยไม่ให้ข่าวปลอมแพร่กระจายเป็นวงกว้าง และสามารถใช้เพื่อการกระตุ้นในการยับยั้งการเผยแพร่ต่อของข่าวได้ งานวิจัยเรื่องนี้มีแนวคิดเพื่อค้นหาคุณสมบัติที่สำคัญของของข่าวปลอม และเปรียบเทียบโมเดลการทำนายของข่าวปลอม ด้วยเทคนิคการทำเหมืองข้อมูล ซึ่งหาความถูกต้องของการทำนายด้วยการใช้วิธีโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมชนิดเพอร์เซปตรอนแบบหลายชั้น (Neuron Network) วิธีต้นไม้ตัดสินใจ (Decision tree) วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbors) และ วิธีนาอีฟเบย์ (Naïve Bays) โดยการวัดค่าความถูกต้อง (Accuracy) ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ย (MAE) และค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) จากการวิจัยพบว่า วิธีที่ดีที่สุดคือวิธีโครงข่ายประสาทเทียม ซึ่งมีค่าความถูกต้องเท่ากับ 95.78% ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.2011 และค่าความคลาดเคลื่อนกำลังสองเท่ากับ 0.1915 วิธีที่ได้ผลลัพธ์ดีเป็นอันดับสองคือวิธีความใกล้เคียงมากที่สุดมีค่าความถูกต้องเท่ากับ 90.51% ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.2051 และค่าความคลาดเคลื่อนกำลังสองเท่ากับ 0.2315 ตามลำดับ งานวิจัยนี้สามารถนำไปเป็นต้นแบบเพื่อสร้างระบบตรวจสอบการตรวจหาข่าวปลอมด้วยระบบอัตโนมัติต่อไปได้
คำสำคัญ : สื่อสังคมออนไลน์, ข่าวปลอม, วิธีโครงข่ายประสาทเทียม, วิธีต้นไม้ตัดสินใจ, วิธีความใกล้เคียงกันมากที่สุด, วิธีนาอีฟเบย์

Abstract

Fake news is on the rise on social media, with fake news are influencing attitudes and decisions on any matter to have a serious impact on social coexistence. The base of fake news focus on content that excites and stimulates the consumer's emotions, so it spreads more easily and faster than real news. Verification the informed source of the news to determine whether it is real or fake news takes time to find. Using a model to predict fake news from social media through data mining techniques is a way to help prevent the spread of fake news. Also can be used to stimulate the dissemination of news. This research focuses on the essential properties of fake news and compare the prediction models of fake news with data mining techniques, as well as find the correctness of classification by using perceptron algorithm of Neuron Network, Decision tree method, the K-Nearest Neighbors and Naïve Bays method. The measurement using the accuracy, the mean absolute error (MAE) and the mean square error (MSE). The results from the efficiency of predicting, the method of classifying information of four algorithms found signifies the competitive accuracy degrees of neuron network is 95.78%, the MAE is 0.2011 and the MSE is 0.1915. However, the second proficiently accuracy, MAE and MSE of the K-Nearest Neighbors is 90.51%, 0.2051, 0.2315 respectively. This research could be used as a prototype for the further construction of an automated fake news detection system.
Keywords : Social Media, Fake News, Neuron Network, Decision tree, K-Nearest Neighbors, Naïve Bays.

[1,2]ภาควิชาบรรณารักษศาสตร์และสารสนเทศศาสตร์ คณะมนุษยศาสตร์ มหาวิทยาเชียงใหม่ จังหวัดเชียงใหม่

บทนำ

สื่อสังคมออนไลน์ถูกนำมาใช้เป็นแหล่งในการกระจายข่าวอย่างกว้างขวาง (Mass media) ในยุคการเชื่อมต่อที่ไร้พรมแดน ซึ่งเป็นประโยชน์ที่ผู้บริโภคจะได้รับรู้ข่าวสารได้ทันท่วงที ในขณะเดียวกันก็ได้ถูกใช้เป็นเครื่องมือในการกระจายข่าวปลอมก่อให้เกิดความตื่นตระหนก ซึ่งสามารถส่งผลกระทบไปถึงความมั่นคงระดับประเทศได้ จากการศึกษาเกี่ยวกับการแพร่กระจายข่าวปลอมในสื่อสังคมออนไลน์ พบว่าข่าวปลอมถูกแบ่งปัน (Share) มากกว่าข่าวจริงถึง 1.7 เท่า โดยมีคนรับข่าวปลอมจากการส่งต่อมากกว่าถึง 100 เท่า และแพร่กระจายได้เร็วกว่าถึง 6 เท่า (Soroush, Deb and Sinan, 2018) จากพฤติกรรมในการบริโภคข่าวสารที่ตื่นเต้นและการใช้คำที่ดึงดูดความสนใจสูง (ชุลีพร ธานีรัตน์ และพัชนี เชยจรรยา, 2557; นันทิกา หนูสมวิโรจน์ สุทธิสีมา, 2562) ทำให้พฤติกรรมดังกล่าวสุ่มเสียงต่อการเข้าใจข้อมูลอย่างแท้จริง แนวคิดในการค้นหาคุณลักษณะที่สำคัญของหัวข้อข่าวและเนื้อข่าวของการวิจัยครั้งนี้ เพื่อใช้สำหรับการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ (Twitter) และเฟซบุ๊ก (Facebook) โดยทำการเปรียบเทียบโมเดลการทำนายการเรียนรู้ของเครื่อง (Machine Learning Model) ด้วยเทคนิคทำเหมืองข้อมูล โดยเลือกอัลกอริทึมที่เหมาะสมสำหรับการทำนายจากการเลือกอัลกอริทึมที่เหมาะสมในการตรวจสอบข่าวปลอมภาษาไทย บนสื่อสังคมออนไลน์ทวิตเตอร์และเฟซบุ๊ก จากอัลกอริทึม 4 แบบที่เลือกมาเท่านั้น โดยจะกล่าวถึงวิธีนาอีฟเบย์ (Naïve Bays) วิธีต้นไม้ตัดสินใจ (Decision tree) วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbors) และอัลกอริทึมโครงข่ายปราสาทเทียม (Neuron Network) ชนิดเพอร์เซปตรอนแบบหลายชั้น ซึ่งแสดงไว้ในส่วนที่ 3 การสร้างตัวแบบและผลลัพธ์การวิจัยแสดงไว้ในส่วนที่ 4 ผลการวิจัยในส่วนที่ 5 รวมทั้งข้อสรุปและอภิปรายผลการวิจัยในส่วนที่ 6 กับ 7 ตามลำดับต่อไป

วัตถุประสงค์

เพื่อเปรียบเทียบโมเดลการเรียนรู้ของเครื่องที่เหมาะสมด้วยเทคนิคทำเหมืองข้อมูลในการตรวจสอบข่าวปลอมภาษาไทย และค้นหาคุณลักษณะสำคัญต่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์และเฟซบุ๊ก

แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง

พฤติกรรมการในการบริโภคและใช้สื่อสังคมออนไลน์ในการรับรู้และเชื่อข่าวสารโดยขาดวิจารณญาณหรือการรู้เท่าทันนั้นเป็นเรื่องสำคัญที่ไม่สามารถมองข้ามข่าวปลอมที่เผยแพร่ทางสื่อสังคมออนไลน์นั้น อาจจะส่งผลกระทบต่อทัศนะคติความเชื่อสู่สังคมโดยรวมได้เพราะการกระจายข่าวในสื่อสังคมออนไลน์เป็นการส่งไปยังเป้าหมายขนาดใหญ่และจำนวนมาก หากผู้ที่รับข่าวสารนั้นขาดความรู้เท่าทันในการเชื่อข่าวนั้นโดยไม่ตระหนักถึงผลที่อาจจะเกิดขึ้น เช่น ข่าวปลอมที่เกี่ยวข้องกับพายุหรือภัยพิบัติธรรมชาติที่ร้ายแรง หรือข่าวและข้อมูลที่ถูกบิดเบือนเกี่ยวกับความมั่นคงของชาติที่อาจมีผู้ไม่หวังดีจงใจเผยแพร่เพื่อสร้างสถานการณ์ให้เกิดความสับสนวุ่นวายในสังคมได้ (สุกัญญา บูรณเดชาชัย, 2560)

1. วิธีนาอีฟเบย์ (Naïve Bays)

การเรียนรู้แบบเบย์อาศัยหลักการของการคำนวณความน่าจะเป็นของแต่ละสมมติฐาน หรือเป้าหมายของการทำนายผลลัพธ์ โดยการเรียนรู้แบบเบย์เป็นการเรียนรู้เพิ่มได้ เนื่องจากตัวอย่างใหม่ที่ได้มาจะถูกนำมาปรับเปลี่ยนการแจกแจงซึ่งมีผลต่อการเพิ่ม หรือ ลดความน่าจะเป็นในการทำนาย ทำให้มีการเรียนรู้ที่เปลี่ยนไป วิธีการนี้ตัวแบบจะถูกปรับเปลี่ยนไปตามตัวอย่างใหม่ที่ได้โดยผนวกกับความรู้เดิมที่มี ซึ่งการทำนายค่าคลาสเป้าหมายของตัวอย่าง เราสามารถคำนวณความน่าจะเป็นของสมมติฐานต่าง ๆ โดยใช้ตัวแปรตามสมการดังนี้

สมการที่ 1

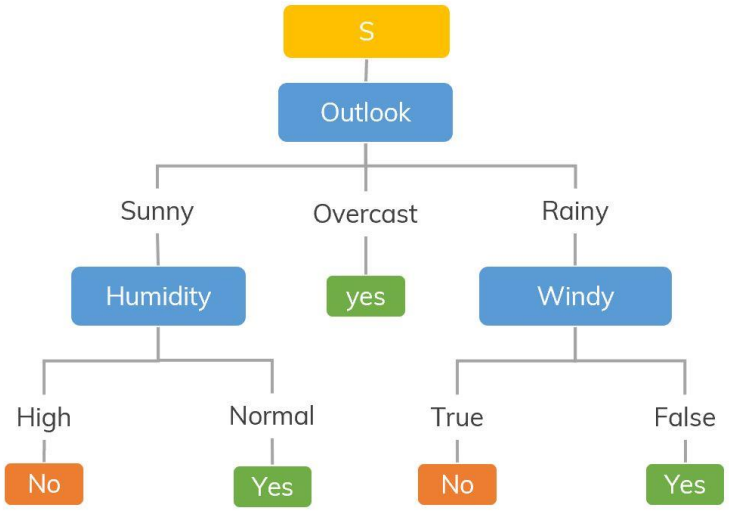
$$P(h|D) = \frac{P(D|h)*P(h)}{P(D)}$$

โดย D แทนข้อมูลที่นำมาใช้ในการคำนวณการแจกแจงความน่าจะเป็น posteriori probability ของสมมติฐาน h คือ P(h|D) ตามทฤษฎี P(h) คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h ส่วน P(D) คือ ความน่าจะเป็นก่อนหน้าของชุดข้อมูลตัวอย่าง D สำหรับ P(h|D) คือ ความน่าจะเป็นของ h เมื่อรู้ D และ P(D|h) คือ ความน่าจะเป็นของ D เมื่อรู้ h (Charoenvorakiat, 2016)

2. ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจเป็นการนำข้อมูลมาสร้างแบบจำลองการพยากรณ์ใน รูปแบบโครงสร้างต้นไม้ และมี

การทำงานแบบการเรียนรู้ของโมเดลแบบมีผู้สอน (Supervised Learning) สามารถสร้างแบบจำลองการจัดกลุ่มได้จากตัวอย่างข้อมูลที่กำหนดไว้ล่วงหน้า และพยากรณ์กลุ่มของรายการที่ยังไม่เคยจำมามาจัดกลุ่มมาก่อนได้ด้วยตัวแบบของ Tree โครงสร้างประกอบด้วย Root Node, Child และ Leaf Node วิธีการสร้างต้นไม้การตัดสินใจเป็นการรับชุดข้อมูลที่ระบุคลาสของข้อมูลเหล่านั้นเป็นข้อมูลนำเข้าหรือเรียกอีกอย่างหนึ่งว่าชุดข้อมูลฝึกการการเรียนรู้ (Training Set) โดยจะได้ผลลัพธ์เป็นตัวแบบการทำนายที่มีลักษณะเป็นโครงสร้างคล้ายต้นไม้ซึ่งตัวแบบการทำนายจะใช้ในการทำนายคลาสของชุดข้อมูลนำเข้าที่ไม่เคยรู้จักมากก่อน เช่น ตัวอย่างการทำนายการเล่นเทนนิว่าเล่นได้หรือไม่ ดังภาพที่ 1



ภาพที่ 1 ต้นไม้ตัดสินใจเลือกความสามารถในการเล่นเทนนิส (Dinh, 2020)

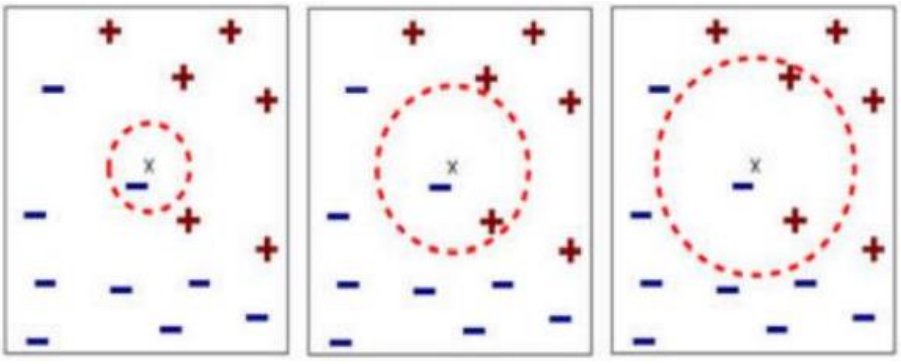
3. วิธีความใกล้เคียงกันมากที่สุด (K-Nearest Neighbors: K-NN)

วิธีความใกล้เคียงกันมากที่สุดจะดูความใกล้กันของข้อมูลที่ใกล้ที่สุด (K-NN) คือ วิธีการจัดกลุ่มข้อมูลที่ใช้การเรียนรู้ตามจำนวนข้อมูลสำหรับการทำนายโดยใช้ค่า K (Sinsomboonthong, 2015) อัลกอริทึมจะใช้ฟังก์ชันวัดความคล้ายกันของข้อมูลตามระยะทางจากตัวแทนของข้อมูล (Centroid) ไปยังข้อมูลตัวอื่น ๆ ระยะทางที่ง่ายที่สุดคือค่า Euclidean เพื่อกำหนดความคล้ายคลึงกันของข้อมูล ขั้นตอนของ K-NN มีดังนี้

(1) กำหนดค่า K ของจำนวนที่ใกล้กันมากที่สุดของข้อมูล

(2) คำนวณระยะทางแบบ Euclidean แต่ละจุดของข้อมูลกับตัวแทนของข้อมูล

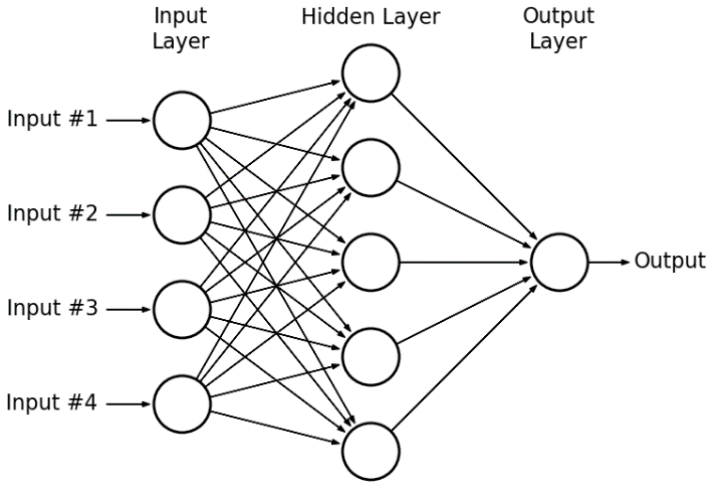
(3) เลือกข้อมูลตามจำนวน K ที่มีความคล้ายกันตามระยะทางแบบต่ำสุดดังภาพที่ 2



ภาพที่ 2 วิธีความใกล้เคียงกันมากที่สุดที่แสดงการจัดกลุ่มตามการเลือกจำนวน K

4. โครงข่ายประสาทเทียมเพอร์เซปตรอนหลายชั้น (MultiLayer Perceptron: MLP)

เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network) มีรูปแบบโครงสร้างและการทำงานของการประมวลผลเหมือนกับสมองของสิ่งมีชีวิตซึ่งมีปรับเปลี่ยนตัวเองต่อการตอบสนองของข้อมูลนำเข้า ประกอบไปด้วย 3 ชั้น ตามรูปที่ 3 โดย ในชั้นที่ 1 เป็นชั้นรับข้อมูลเข้า (Input) เพื่อทำการเรียนรู้ขั้นที่ 2 ขั้นการเรียนรู้ที่มีการเรียนรู้ข้อมูลโดยจะสุ่มค่าน้ำหนัก (Weight) ของแต่ละชุดข้อมูล เพื่อนำไปใช้บันทึกค่าความสำคัญของข้อมูลแต่ละชุดจะมีไม่เท่ากัน และสุ่มเลือกค่าความโน้มเอียง (Bias) เพื่อเป็นตัวกำหนดแนวทางการเรียนรู้ของโมเดล ส่วนในชั้นอื่น ๆ ของ Hidden Layer จะปรับค่าความโน้มเอียง (Bias) และค่าน้ำหนัก (Weight) ในขณะกำลังเรียนรู้เพื่อให้ได้ตัวแบบที่เหมาะสมที่สุด และในชั้นที่ 3 ชั้นของข้อมูลผลลัพธ์ (Output Layer) เป็นจากการแสดงผลการเรียนรู้ของโมเดล ซึ่งผลลัพธ์ที่ได้นั้นจะอยู่ในลักษณะของคลาส (Class) (พยุง มีสัจ, 2553)



ภาพที่ 3 โครงข่ายประสาทเทียมแบบ Multilayer Perceptron (Hassan, Abdelazim, Mohamed & Oliver, 2015)

5. การวัดค่าความถูกต้องของตัวแบบเป็นการวัดค่าความถูกต้อง (Accuracy) โดยการใช้การคำนวณจากจำนวนครั้งที่คำณวณได้จากการทำนายได้ถูกต้องแล้วนำมาเทียบกับจำนวนครั้งทั้งหมดที่นำไปให้ตัวแบบทำนาย ดังสมการที่ (2) และสำหรับงานวิจัยในครั้งนี้ผลลัพธ์ที่ได้จากการทำนายคือข่าวปลอมที่ไม่สามารถคำนวณค่าความถูกต้องของตัวแบบได้ตามสมการที่ (3) ซึ่งค่า True Positive (TP) คือทำนายว่าเป็นข่าวปลอม และผลจากการคำนวณคือเป็นข่าวปลอม ส่วนค่า True Negative (TN) คือค่าที่ทำนายว่าเป็นข่าวจริง และผลคือจากการคำนวณเป็นข่าวจริง สำหรับค่า False Positive (FP) เมื่อทำนายว่าเป็นข่าวปลอมและผลที่ได้จากการทำนายคือเป็นข่าวจริง และค่า False Negative (FN) เมื่อทำนายว่าเป็นข่าวจริง และผลคือจากการคำนวณเป็นข่าวปลอมโดยค่าความถูกต้องที่คำนวณได้จะมีค่าอยู่ระหว่า 0-1 (ค่าเป็น 1 หมายความว่าถูกต้อง 100%) (Wang, Xu, Fujita, & Liu, 2016)

สมการที่ 2

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

สมการที่ 3

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

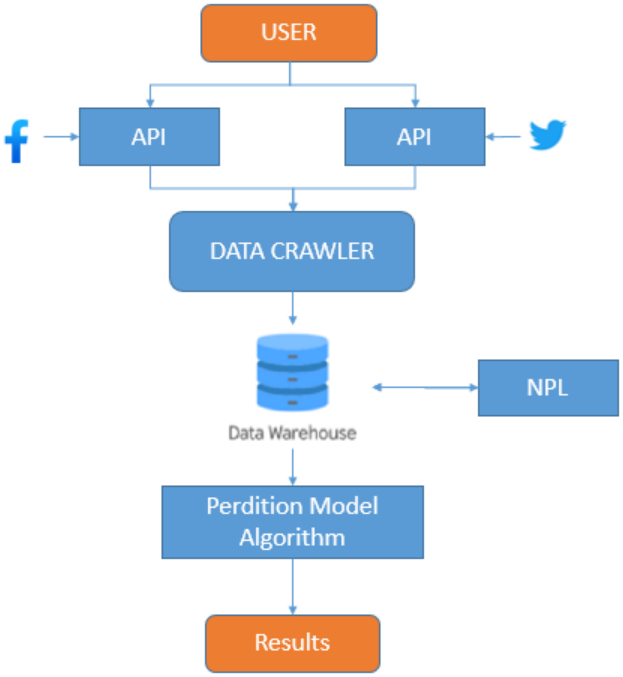
6. งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่ามีการนำเทคนิคเหมืองข้อมูลและการเรียนรู้ของเครื่อง มาช่วยในการแยกแยะข่าวปลอม ไม่ว่าจะเป็นผลงานที่เป็นการนำเอาการประมวลผลภาษาธรรมชาติมาช่วยในการหาคะแนนความรู้สึกในเนื้อหา ร่วมกับคุณลักษณะอื่นของข่าว และใช้เทคนิคต้นไม้ตัดสินใจ เทคนิคซัปพอร์ตเวกเตอร์แมชชีน ในการตรวจสอบข่าวปลอม (Krishnan & Chen, 2018) หรืองานวิจัยที่ได้นำเอาเทคนิคโครงข่ายประสาทเทียม แบบ Long Short Term Memory ซึ่งเป็นการ

ประยุกต์ใช้โครงข่ายประสาทเทียมที่มีความรู้หรือความจำจากข้อมูลก่อนหน้ามาใช้ในการคำนวณร่วมกับโครงข่ายประสาทเทียมแบบคอนโวลูชัน ประกอบด้วยชั้นคอนโวลูชันและชั้นพูลลิงในโครงข่ายเพื่อเปรียบเทียบตัวแบบที่เหมาะสมในการแยกแยะข่าวปลอม (Oluwaseun, Deepayan & Shahrzed, 2018) หรืองานวิจัยของ ชวัล วัฒนากิจจากุล และอนันท์ ชกสุริวงค์ (2563) ที่ได้นำเอาเทคนิคต้นไม้ตัดสินใจ เทคนิคซัปพอร์ตเวกเตอร์แมชชีน เทคนิคโครงข่ายประสาทเทียมมาทำการแยกแยะข่าวปลอม อีกทั้งงานวิจัยที่รวมเอาเทคนิคทางด้านการเรียนรู้ของเครื่องมาผสมผสานกันออกเป็นตัวแบบ (Ruchansky, Seo, & Liu, 2017) ที่ได้ทำการพัฒนาตัวแบบชื่อ CSI (CaptureScore and Integrate) เพื่อใช้ในการสอนและทดสอบประสิทธิภาพการแยกแยะข่าวปลอมเปรียบเทียบกับเทคนิคอื่น

วิธีดำเนินการวิจัย

การวิจัยในครั้งนี้ผู้วิจัยได้ใช้ข้อมูลจากสื่อสังคมออนไลน์ทวิตเตอร์และเฟซบุ๊ก โดยการสร้างโปรแกรมในการติดต่อกับสื่อสังคมออนไลน์ทวิตเตอร์และเฟซบุ๊กผ่านทาง Application Programming Interface (API) สำหรับใช้ในการดึงข้อมูลข่าว มาอย่างละ 300 รวมข้อมูลข่าว 600 ชุดข้อมูลและมีความแตกต่างประเภทของข่าว โดยประกอบด้วยข่าวจริงจำนวน 300 ข่าว (50%) และข่าวปลอม 300 ข่าว (50%) จากผู้ใช้สื่อสังคมได้แบ่งข้อมูลแต่ละส่วนออกเป็นร้อยละ70 เพื่อนำไปใช้ให้อัลกอริทึมแต่ละตัวได้ทำการเรียนรู้ก่อนและจะได้มาเป็นตัวโมเดลและนำข้อมูลที่เหลืออีกร้อยละ 30 เป็นข้อมูลทดสอบ จะต้องให้ข้อมูลว่าเป็นข่าวปลอมหรือข่าวจริงก่อนดังแผนภาพตามรูปที่ 4



ภาพที่ 4 แผนภาพในการทำงานของการวิจัย

1. การสร้างชุดข้อมูล

เมื่อรวบรวมข้อมูลและได้แบ่งสัดส่วนการคำนวนและเรียนรู้ของแต่ละอัลกอริทึมแล้ว ขั้นตอนต่อไปคือการรวบรวมคุณสมบัติและคุณลักษณะต่าง ๆ ของข้อมูลเพื่อเปลี่ยนให้เป็นระบบตัวเลขที่สามารถนำไปคำนวณและเปรียบเทียบค่าทางสถิติได้ตามรูปแบบของแต่ละอัลกอริทึม ดังนี้ ความยาวของชื่อผู้ใช้ ชื่อผู้ใช้มีตัวเลขหรือไม่ จำนวนเครื่องหมายตกใจ จำนวนผู้ติดตาม การโพสต์เป็นวันที่เท่าไหร่ของสัปดาห์ จำนวนวนการเชื่อมโยง จำนวนการโพสต์ของผู้ใช้ ช่วงเวลาการโพสต์ ความยาวของเนื้อหา จำนวนเครื่องหมายคำถาม จำนวนเพื่อน จำนวนแฮชแท็ก จำนวนการกดชื่นชอบ จำนวนการแบ่งปัน คะแนนความรู้สึก อัตราส่วนเพื่อนและผู้ติดตามหลังจากนั้นผู้วิจัยได้ดำเนินการการแปลงข้อมูล (Data Preprocessing) เพื่อให้ข้อมูลอยู่

ในมาตรฐานเดียวกัน (Standardized) ด้วยวิธีการทำให้อยู่ในช่วงค่าเลข 0-1 (Standard Scaler) โดยสามารถอธิบายการแปลงข้อมูลได้ดังสมการที่ (4)

สมการที่ 4

$$X_i = \frac{x_i - \mu}{\sigma}$$

โดยที่

X_i คือ ข้อมูลนำเข้า

μ คือ ค่าเฉลี่ยของข้อมูลนำเข้าทั้งหมด

σ คือ ค่าเบี่ยงเบนมาตรฐานของข้อมูลนำเข้าทั้งหมด

2. การประมวลผลภาษาธรรมชาติในส่วนของเทคนิคการประมวลผลภาษาธรรมชาติ(Natural Language Process: NLP) เป็นเทคนิคการประมวลภาษา เพื่อให้คอมพิวเตอร์ได้เข้าใจภาษาของมนุษย์และสามารถตีความได้ในลักษณะของคะแนนความรู้สึก ซึ่งในงานวิจัยฉบับนี้ได้นำทฤษฎีส่วนนี้มาใช้ในการเพิ่มคุณลักษณะด้านการตีความด้านคะแนนความรู้สึกให้กับข้อมูล และได้ดำเนินการให้เครื่องสามารถเรียนรู้ของข้อมูลได้จากข้อมูลตัวอย่างที่ผู้วิจัยได้ดำเนินการจัดเตรียม โดยความรู้ที่เครื่องได้เรียนรู้และแยกแยะได้เก็บไว้ในฐานความรู้ซึ่งอยู่ในรูปแบบหลากหลายด้วยการสร้างกฎ ฟังก์ชันความจำของเทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจ ความใกล้เคียงกันมากที่สุด โครงข่ายประสาทเทียมเพอร์เซปตรอนหลายชั้น มาใช้ในการสร้างตัวแบบเพื่อเปรียบเทียบความถูกต้องของการแยกข่าว

3. การแบ่งข้อมูลเพื่อการเรียนรู้และทดสอบ ผู้วิจัยได้ใช้โมเดลจำนวน 4 แบบ ที่ใช้เทคนิคนาอีฟเบย์ ต้นไม้ตัดสินใจ ความใกล้เคียงกันมากที่สุด โครงข่ายประสาทเทียมเพอร์เซปตรอนหลายชั้น ตามลำดับและนำข้อมูลจากหัวข้อ 1 ที่รวบรวมได้แบ่งเป็น 2 ส่วน คือส่วนที่ใช้ให้แต่ละอัลกอริทึมได้เรียนรู้และส่วนที่ใช้ทดสอบโมเดลตามตารางที่ 1 และตารางที่ 2

ตารางที่ 1 การแบ่งข้อมูลเพื่อใช้ในเรียนรู้ของแต่ละอัลกอริทึม

	Naïve Bays		Decision tree		K-Nearest Neighbors		Neuron Network	
	ข่าวจริง	**ข่าวปลอม**	**ข่าวจริง**	**ข่าวปลอม**	**ข่าวจริง**	**ข่าวปลอม**	**ข่าวจริง**	**ข่าวปลอม**
Facebook	210	210	210	210	210	210	210	210
Twitter	210	210	210	210	210	210	210	210

ตารางที่ 2 การแบ่งข้อมูลเพื่อใช้ในการทดสอบของแต่ละอัลกอริทึม

	Naïve Bays		Decision tree		K-Nearest Neighbors		Neuron Network	
	ข่าวจริง	**ข่าวปลอม**	**ข่าวจริง**	**ข่าวปลอม**	**ข่าวจริง**	**ข่าวปลอม**	**ข่าวจริง**	**ข่าวปลอม**
Facebook	90	90	90	90	90	90	90	90
Twitter	90	90	90	90	90	90	90	90

สรุปผลการวิจัย

ผลจากการแบ่งข้อมูลเพื่อให้แต่ละอัลกอริทึมได้เรียนรู้ข้อมูลข่าวร้อยละ 70 และได้ตัวแบบเพื่อนำมาทดลองใช้ในการทำนายเพื่อหาข่าวจริงและข่าวปลอมทั้ง 4 ตัวแบบ ผลที่ออกมาสำหรับในกระบวนการทดสอบ นั้นวิธีที่ดีที่สุดคือวิธีโครงข่ายประสาทเทียม ซึ่งมีค่าความถูกต้องเท่ากับ 95.78% ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.2011 และค่าความคลาดเคลื่อนกำลังสองเท่ากับ 0.1915 วิธีที่ได้ผลลัพธ์ดีเป็นอันดับสองคือวิธีความใกล้เคียงมากที่สุดมีค่าความถูกต้องเท่ากับ 90.51% ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.2051 และค่าความคลาดเคลื่อนกำลังสองเท่ากับ 0.2315 ตามลำดับ แต่สำหรับในกระบวนเรียนรู้นั้นค่าความถูกต้องของแต่ละอัลกอริทึม ได้แสดงไว้ในตารางที่ 3 ซึ่งค่าที่ลดลงเล็กน้อยโดยจากการทดสอบ อาจจะเกิดจากข้อมูลที่แตกต่างกันในรายละเอียด เพราะโดยปกติแล้วการทำเหมืองข้อมูลให้มีประสิทธิภาพมากที่สุดนั้นจะต้องทดสอบด้วยกันหลายวิธี เนื่องจากความถูกต้องที่ทำนายได้นั้นขึ้นอยู่กับข้อมูลที่จะนำมาทำนายอีกด้วย โดยในการวิจัย

ครั้งนี้ มีข้อสรุปได้ว่า ตัวแบบ โครงข่ายประสาทเทียมเพอร์เซปตรอนหลายชั้นนั้นมีค่าความแม่นยำ มากกว่าตัวแบบความใกล้เคียงกันมากที่สุด เทคนิคนาอีฟเบย์ แบะต้นไม้ตัดสินใจ ดังนั้นแล้ว ตัวแบบโครงข่ายประสาทเทียมเพอร์เซปตรอนหลายชั้น จึงเหมาะสมที่จะนำไปไปทำนายผลการแยกข่าวจริงและข่าวปลอมได้ และสามารถชี้แนะในการรับตระหนักรู้รวมถึงพฤติกรรมในการบริโภคข่าวและสื่อในสังคมออนไลน์ ในขณะที่ได้เห็นหัวข้อข่าวในเบื้องต้นได้

ตารางที่ 3 ผลลัพธ์ที่ได้จากการเรียนรู้และการทดสอบแต่ละอัลกอริทึม

การวัดค่าของผลลัพธ์ที่ได้	ผลที่ได้จากการทดสอบแต่ละอัลกอริทึม				ผลที่ได้จากการเรียนรู้แต่ละอัลกอริทึม			
	Naïve Bays	Decision tree	K-Nearest Neighbors	Neuron Network	Naïve Bays	Decision tree	K-Nearest Neighbors	Neuron Network
Accuracy	89.01%	89.55%	90.51%	**95.78%**	91.22%	90.89%	93.68%	**96.75%**
MAE	0.2051	**0.2011**	0.2051	**0.2011**	0.2051	0.1998	0..2117	**0.1987**
MSE	0.2105	0.2022	0.2315	**0.1915**	0.2105	0.2018	0.2265	**0.1922**

การอภิปรายผลการวิจัย

จากการเตรียมข้อมูลเพื่อให้เครื่องเรียนรู้และทดสอบจากตัวแบบต่าง ๆ ของการแยกข่าวปลอมจากสื่อสังคมออนไลน์นั้น เริ่มด้วยการศึกษาเรื่องการนิยามคุณลักษณะของข่าวออกเป็น 3 ด้านคือ ด้านผู้กระจายข่าว (User Based) โดยจะสนใจในบริบทของผู้ใช้ที่กระจายข่าว โดยดูจำนวนเพื่อน จำนวนผู้ติดตาม จำนวนเผยแพร่และแชร์ข้อความหรือการแบ่งปันต่อ ระยะเวลาที่สร้างบัญชีของผู้กระจาย สำหรับด้านเนื้อหาข่าว (Content Based) ซึ่งจะสนใจในรายละเอียดของเนื้อหาข่าว เช่น จำนวนแฮชแท็ก (Hash Tag) จำนวนการเชื่อมโยง และจำนวนด้านการมีส่วนร่วม (Social Based) คุณลักษณะของข่าวปลอมบนสื่อสังคมออนไลน์สภาพแวดล้อมของการแยกแยะข่าวปลอม การแสดงความคิดเห็น การกดปุ่มชื่นชอบ การแบ่งปันต่อ โดยการนำคุณสมบัติของข่าวและผู้ใช้ทั้งหมด มาหาตัวแบบหรับการตรวจสอบข่าวปลอมที่เป็นภาษาไทยบนสื่อสังคมออนไลน์ ที่ใช้เทคนิคโครงข่ายประสาทเทียมโดยให้ค่าความถูกต้องมากที่สุดถึง 95.78% % ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.2011 และค่าความคลาดเคลื่อนกำลังสองเท่ากับ 0.1915 โดยอยู่ในค่าความคลาดเคลื่อนที่น้อย ซึ่งผลการทดสอบที่ได้ค่าน้อยกว่าจากกระบวนการเรียนรู้ที่ได้มากกว่าโดยที่ค่าความถูกต้องอยู่ที่ 96.75% ค่าความคลาดเคลื่อนสัมบูรณ์เฉลี่ยเท่ากับ 0.1987 และค่าความคลาดเคลื่อนกำลังสองเท่ากับ 0.1922 อาจเกิดจากการใช้ข้อมูลที่มีการแบ่งชุดข้อมูล 70% ในการเรียนรู้และ 30% ในการทดสอบ งานวิจัยฉบับนี้สามารถนำไปใช้ในการประยุกต์เพื่อสร้างระบบตรวจสอบข่าวปลอมในแบบทันเวลาจริง (Real-Time Verification) และสามารถเก็บรวบรวมข้อมูลข่าว เพื่อใช้ในการเรียนรู้และทดสอบตัวแบบเพื่อให้ได้ความถูกต้องที่มากยิ่งขึ้น

สรุปและข้อเสนอแนะ

การใช้โมเดลการทำนายข่าวปลอมจากสื่อสังคมออนไลน์ด้วยเทคนิคการทำเหมืองข้อมูลนั้น ผู้วิจัยมีข้อเสนอแนะในการวิจัยครั้งต่อไป ดังนี้

1. การเลือกใช้วิธีในการพยากรณ์เพื่อหาโมเดลในอนาคตจะศึกษาหลักการทำงานของวิธีอื่น ๆ ที่สามารถทำนายผลลัพธ์ที่มีความหลากหลายมากขึ้นและเลือกตัวแบบที่เหมาะสมมากขึ้น รวมถึงการปรับเปลี่ยนค่าการใช้พารามิเตอร์ให้เหมาะสมกับข้อมูลที่มีและเลือกวิธีการพยากรณ์นั้นให้เกิดประสิทธิภาพสูงสุด

2. ศึกษาวิธีการอย่างอื่นที่มีผลต่อการกระจายข่าว รวมทั้งการเก็บข้อความที่มีความหมายแฝงและหาวิธีประมวลผลภาษาธรรมชาติที่มีประสิทธิภาพสูงมาช่วยในการหาความหมายของคำแฝงเหล่านั้น

เอกสารอ้างอิง

1. ชวัล วัฒนากิจจากุล และอนันท์ ชกสุริวงค์. (2563). **ตัวแบบการเรียนรู้ของเครื่องเพื่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์.** ใน การประชุมวิชาการวิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยี และสถาปัตยกรรมศาสตร์ ครั้งที่ 11, 1289-1295.

2. ชุลีพร ธานีรัตน์ และพัชนี เชยจรรยา. (2557). รูปแบบการดำเนินชีวิต การเปิดรับข่าวสารและพฤติกรรมการท่องเที่ยว ตลาดย้อนยุคของนักท่องเที่ยวชาวไทย. **วารสารนิเทศศาสตร์และนวัตกรรม นิด้า, 1(1) ฉบับปฐมฤกษ์,** 131-145.
3. นันทิกา หนูสม และวิโรจน์ สุทธิสีมา. (2562). ลักษณะของข่าวปลอมในประเทศไทยและระดับความรู้เท่าทันข่าวปลอมบนเฟซบุ๊กของผู้รับสารในเขตกรุงเทพมหานคร. **วารสารนิเทศศาสตร์, 37(1),** 37-45.
4. พยุง มีสัจ. (2553). **ระบบฟัซซีและโครงข่ายประสาทเทียม.** กรุงเทพฯ: ศูนย์ปผลิตตำรา เรียนมหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ.
5. สุกัญญา บูรณเดชาชัย. (2560). **ไม่ชัวร์แชร์ไป สังคมวุ่นวาย, สังคมไทย ชัวร์ก่อนแชร์.** กรุงเทพฯ:
6. กองทุนพัฒนาสื่อปลอดภัยและสร้างสรรค์. สืบค้นจาก http://imgs.mcot.net/images/2018/05/1525684457247.pdf
7. Charoenvorakiat, P. (2016). **Bayes' theorem. Artificial Intelligence, Data Mining, Feature, Inspiration, Quantum Computer.** Available from: http://www.parinya.net/node/2295.
8. Dinh, A. T. (2020). **Decision Tree Classifier.** Available from https://dinhanhthi.com/decision-tree-classifier.
9. Hassan, H., Abdelazim, N., Mohamed, Z., & Oliver, S. (2015). Assessment of Artificial Neural Network for Bathymetry Estimation Using High Resolution Satellite Imagery in Shallow Lakes: Case Study El Burullus Lake. **International Water Technology Journal, 5(4),** pp. 248-259.
10. Krishnan, S. & Chen, M. (2018). **Identify Tweet with Fake news.** IEEE International Conference on Information Reuse and Integration for Data Science, Utah, USA, 7-9 July 2018, pp. 460–464.
11. Oluwaseun, A., Deepayan, B. & Shahrzed, Z. (2018). **Fake News Identification on Twitter with Hybrid CNN and RNN Models.** International Conference on Social Media & Society, Copenhagen, Denmark, 18-20 July 2018, pp. 226-230.
12. Ruchansky, N., Seo, S., & Liu, Y. (2017). **CSI: A Hybrid Deep Model for Fake News Detection.** The 26th ACM International Conference on Information and Knowledge Management (CIKM 2017), Singapore, 6-10 November 2017, pp. 797-806.
13. Sinsomboonthong, S. (2015). **Data Mining.** Bangkok: Chamchuree products.
14. Soroush, V., Deb, R., & Sinan, A. (2018). **The spread of true and false news online.** Science, 359, pp. 1146-1151.
15. Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). **Towards felicitous decision making: An overview on challenges and trends of Big Data.** Information Sciences, 367-368, pp. 747–765.