การจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย โดยใช้เทคนิคการทำเหมืองข้อมูล

The Classification of Credibility of Thai News Source Websites Using Data Mining Techniques

องอาจ อุ่นอนันต์[1] และพยุง มีสัจ[1]
Aongart Aun-a-nan[1] and Phayung Meesad[1]
[1]คณะเทคโนโลยีสารสนเทศ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
[1]Faculty of Information Technology, King Mongkut's University of Technology North Bangkok
Received: July 18, 2019
Revised: September 3, 2019
Accepted: September 10, 2019

บทคัดย่อ

การเพิ่มขึ้นของแหล่งข่าวที่ไม่น่าเชื่อถือในสื่อออนไลน์ที่เข้าถึงได้ในทุกวัน เช่น สื่อสังคมออนไลน์ บล็อกข่าว และเว็บไซต์หนังสือพิมพ์ออนไลน์ สร้างความเข้าใจผิดให้กับผู้ได้รับข่าวสารนั้น ซึ่งทำให้การระบุแหล่งข่าวที่น่าเชื่อถือเป็นเรื่องที่ท้าทาย เป้าหมายของการวิจัยนี้คือการสร้างโมเดลการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย มีวัตถุประสงค์เพื่อศึกษาข้อมูลปัจจัยที่เกี่ยวข้องกับความน่าเชื่อถือของเว็บไซต์ และเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้ในการจำแนกประเภท โดยการรวบรวมข้อมูลปัจจัยทางเทคนิคของเว็บไซต์แหล่งข่าวและสื่อสังคมออนไลน์ของแหล่งข่าวแล้วทำการจัดกลุ่มข้อมูลเว็บไซต์แหล่งข่าวเพื่อกำหนดป้ายกำกับกลุ่มของแหล่งข่าว โดยจัดกลุ่มที่มีประสิทธิภาพดีที่สุดแบ่งออกเป็น 5 กลุ่ม จากนั้นทำการวิเคราะห์ข้อมูลด้วยเทคนิคการจำแนกประเภทประกอบด้วย 5 เทคนิค ดังนี้ Decision Tree--C4.5, Naïve Bayes, K-Nearest Neighbor--K-NN, Multilayer Perceptron และ Support Vector Machine--SVM แล้วเปรียบเทียบค่าประสิทธิภาพพบว่าเทคนิค K-Nearest Neighbor--K-NN ที่มีค่า K เท่ากับ 5 6 และ 7 มีค่าประสิทธิภาพมากที่สุดเท่ากัน (Accuracy=96.03%, Precision=0.962, Recall=0.960, F-measure=0.959) ซึ่งผู้วิจัยเลือกใช้เทคนิค K-Nearest Neighbor--K-NN เมื่อ K เท่ากับ 6 เนื่องจากทำให้มีอำนาจจำแนกได้ดีกับจำนวน 5 กลุ่ม

คำสำคัญ: เหมืองข้อมูล, การจัดกลุ่มข้อมูล, การจำแนกประเภท, ความน่าเชื่อถือ, ข่าวออนไลน์

Abstract

The increase of unreliability in the online social media nowadays, such as social networking platforms, news blogs and online newspaper websites, causes misunderstanding for the receivers which challenge the news referencing progress. The purpose of this research is to make a model to classify the reliability of Thai-based news references websites. In these terms, this research aims to study the factors and consider the data's reliability related to each website. Also, the performance comparison of the model, used for classifying progress by collecting the website s' essential technical data factors and sorting them into groups of data's references to state the news references' category, placed the effective ones into 5 groups. Moreover, the data was analyzed by using these 5 analytical techniques: Decision Tree C4.5, Naïve Bayes, K-Nearest Neighbor--K-NN, Multilayer Perceptron and Support Vector Machine--SVM. After being analyzed in performance comparison progress, the K-Nearest Neighbor technique--K-NN, which has the max performance value of 5, 6 and 7, of its K is even (Accuracy=96.03%, Precision =0.962, Recall=0.960, F-measure=0.959). Hence, the researcher chose the K-Nearest Neighbor--KNN technique-when K equals 6 since it would make the analyzation of the 5 groups most effective.

Keywords: data mining, data clustering, classification, credibility, online news

บทนำ

ในช่วงไม่กี่ปีที่ผ่านมาสังคมได้ก้าวไปสู่การเป็นสังคมแห่งข้อมูล ปัจจุบันข้อมูลและข่าวสารบนอินเทอร์เน็ต (Internet) มีการเติบโตไปอย่างรวดเร็วและต่อเนื่อง ทำให้มีปริมาณเอกสารและข้อมูลเพิ่มขึ้นอย่างมากมายมหาศาล ทั้งในรูปแบบของข้อมูลที่เป็นตัวอักษร ตัวเลข รูปภาพ หรือคลิปวิดีโอ ในขณะที่พฤติกรรมของผู้ใช้งานอินเทอร์เน็ตเปลี่ยนไปเช่นกัน ซึ่งในประเทศไทยสถิติจำนวนผู้ใช้งานอินเทอร์เน็ตระหว่างปีพุทธศักราช 2555-2561 มีจำนวนผู้ใช้งานเพิ่มขึ้นจากร้อยละ 44.90 เป็นร้อยละ 66.10 (National Statistical Office, 2018) และมีแนวโน้มเพิ่มขึ้นอย่างต่อเนื่องทุกปีโดยเฉลี่ยร้อยละ 5.00 ต่อปี โดยพฤติกรรมการใช้งานเน้นไปเพื่อการบริโภคข้อมูลข่าวสารและใช้งานเครือข่ายสังคมออนไลน์ (social network) ทำให้เป็นอีกหนึ่งช่องทางที่ถูกนำมาใช้ในการเผยแพร่ข้อมูลข่าวสาร ซึ่งมีความสะดวกในการส่งต่อ (share) ข้อมูลข่าวสาร จึงเป็นสาเหตุให้ข้อมูลข่าวสารนั้นมีการแพร่กระจายไปอย่างรวดเร็วโดยมีทั้งข้อมูลข่าวสารที่เป็นจริงและข้อมูลข่าวสารที่เป็นเท็จถูกส่งต่อไปยังผู้ใช้งานคนอื่น ถ้าข้อมูลข่าวสารที่เป็นจริงถูกส่งต่อก็จะเป็นประโยชน์กับผู้ใช้งานที่ได้รับข้อมูล แต่ถ้าหากข้อมูลข่าวสารที่เป็นเท็จถูกส่งต่อก็จะไม่เกิดประโยชน์กับผู้ใช้งานที่ได้รับข้อมูลแล้วผู้ใช้งานยังทำการส่งต่อข้อมูลข่าวสารนั้น ไปให้กับผู้ใช้งานรายอื่น ๆ ยิ่งทำให้ข้อมูลข่าวสารที่เป็นเท็จแพร่กระจายไปเป็นวงกว้างอย่างรวดเร็ว อีกทั้งเป็นเรื่องยุ่งยากที่ผู้ใช้งานจะสามารถตรวจสอบข้อมูลข่าวสารที่ได้รับมานั้นว่าเป็นข้อมูลข่าวสารที่เป็นจริง หรือข้อมูลข่าวสารที่เป็นเท็จ เนื่องจากยังขาดเครื่องมือที่ใช้ในการตรวจสอบข้อมูลข่าวสารออนไลน์ โดยข้อมูลข่าวสารที่เป็นภาษาไทยนั้นมีความซับซ้อนของรูปประโยค เนื่องจากภาษาไทยมีลักษณะเขียนติดกันเป็นประโยคยาวต่อเนื่องไม่มีสัญลักษณะบอกการสิ้นสุดประโยคอย่างชัดเจนเหมือนภาษาอังกฤษที่ใช้จุดในการจบประโยค (Chumwatana, 2013) ซึ่งงานวิจัยในด้านการเรียนรู้ของเครื่อง (machine leaning) ที่เกี่ยวข้องกับภาษาไทยยังคงมีการพัฒนาอย่างต่อเนื่องจนถึงปัจจุบัน เพื่อให้การวิเคราะห์ข้อความภาษาไทยมีประสิทธิภาพเพิ่มขึ้น นอกจากที่จะวิเคราะห์

จากข้อความแล้วยังสามารถวิเคราะห์ความน่าเชื่อถือจากแกล่งที่มีของข่าวได้อีกด้วยซึ่งจะสามารถทำการวิเคราะห์ได้ง่ายกว่าการวิเคราะห์ข้อความจากเนื้อหาข่าว

ผู้วิจัยจึงมีแนวคิดที่จะพัฒนาการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย โดยใช้การเก็บข้อมูลของปัจจัยที่มีต่อความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย ที่มีการเผยแพร่และถูกส่งต่อในเครือข่ายสังคมออนไลน์ แล้วนำข้อมูลมาทำการวิเคราะห์โดยมีเป้าหมายของงานวิจัยเพื่อให้สามารถจัดกลุ่มข้อมูลแล้วทำการกำหนดกลุ่มของเว็บไซต์แหล่งข่าวได้ และสามารถจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวได้ โดยใช้เทคนิคการทำเหมืองข้อมูล จากนั้นนำโมเดลที่ได้นำไปพัฒนาเป็นเครื่องมือตรวจสอบความน่าเชื่อถือของเว็บไซต์แหล่งข่าวที่ผู้ใช้งานได้รับมาจากช่องทางต่าง ๆ เพื่อจำแนกว่าเว็บไซต์มีความน่าเชื่อถือหรือไม่

ในเอกสารงานวิจัยนี้ได้แบ่งออกเป็นหัวข้อต่อไปนี้ (1) วัตถุประสงค์การวิจัย (2) แนวคิดทฤษฎีที่เกี่ยวข้อง (3) กรอบแนวคิดการวิจัย (4) วิธีดำเนินการวิจัย (5) ผลการวิจัย (6) การอภิปรายผล และ (7) ข้อเสนอแนะ

วัตถุประสงค์การวิจัย

1. เพื่อจัดกลุ่มความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย

2. เพื่อเปรียบเทียบประสิทธิภาพของเทคนิคที่ใช้ในการจัดกลุ่มข้อมูล โดยใช้ค่า Calinski-Harabaz Score

3. เพื่อพัฒนาโมเดลที่ใช้ในการจำแนกประเภทจากข้อมูลปัจจัยที่เกี่ยวข้องกับความน่าเชื่อถือของเว็บไซต์

4. เพื่อเปรียบเทียบประสิทธิภาพของโมเดลที่ใช้ในการจำแนกประเภทความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย โดยใช้ค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าการค้นคืน (recall) และค่าประสิทธิภาพโดยรวม (F-measure)

แนวคิดทฤษฎีที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ได้ใช้แนวคิด ทฤษฎีและหลักการต่าง ๆ ที่เกี่ยวข้อง เพื่อนำมาทบทวนให้เกิดองค์ความรู้แล้วนำไปประยุกต์ใช้ในการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย โดยใช้เทคนิคการทำเหมืองข้อมูล ดังนี้

1. ความน่าเชื่อถือของข่าว

ข่าวสารปลอมสามารถเกิดจากบทความที่เขียนในรูปแบบของข่าวสารที่เป็นเท็จซึ่งอาจจะเขียนด้วยความตั้งใจที่จะหลอกลวงหรือทำให้เข้าใจผิด ซึ่งรวมถึงการใช้ภาพที่มีการดัดแปลงหรือการรายงานข่าวสารเพียงด้านเดียว การใช้บัญชีปลอม บทความที่ไม่ปรากฏแหล่งอ้างอิงที่ชัดเจน โดยมีปัจจัยที่มีอิธิพลต่อความน่าเชื่อถือและปัจจัยที่มีอิทธิพลต่อการยอมรับ (Fairbanks et al., 2018, Toommanon & Whattananarong, 2012) ดังนี้

1.1 ปัจจัยที่มีอิทธิพลต่อความน่าเชื่อถือ

1.1.1 ด้านความบริบูรณ์ของเนื้อหา

สื่อออนไลน์มีหน้าที่ในการนำเสนอเนื้อหาสาระในรูปแบบของข้อมูล สารสนเทศ ข่าวสาร ความบันเทิง ความรู้ และอื่น ๆ ซึ่งทั้งหมดนั้นเรียกว่า “เนื้อหา (contents)” ซึ่งอยู่ในรูปแบบของตัวอักษร ภาพนิ่ง ภาพเคลื่อนไหว คลิปวิดีโอ ผสมผสานกันโดยสื่อมีหน้าที่ในการนำเสนอเพียงเท่านั้น ทำให้ความบริบูรณ์ของเนื้อหาเป็นปัจจัยที่มีอิทธิพลมากที่สุด โดยปัจจัยที่มีอิทธิพลต่อความน่าเชื่อถือด้านความบริบูรณ์ของเนื้อหามีปัจจัยย่อยได้แก่ ผู้นำเสนอ ความถูกต้องแม่นยำ ชื่อเสียงของสำนักข่าว ความครบถ้วนของข่าว แหล่งที่มาของข่าว และความทันสมัยของข่าว

1.1.2 ด้านจริยธรรม

สื่อออนไลน์ถือเป็นสื่อสารมวลชนที่มุ่งเน้นไปที่กลุ่มคนทั่วไปเป็นจำนวนมาก ซึ่งแตกต่างจากสื่อสารระหว่างบุคคล การสื่อสารกลุ่มใหญ่ และการสื่อสารองค์กรที่มุ่งเน้นไปที่กลุ่มคนเฉพาะกลุ่ม ดังนั้นการสื่อสารกับคนจำนวนมากจำเป็นจะต้องมีวิธีการสื่อที่มีความเฉพาะสำหรับมวลชน อีกทั้งต้องมีจริยธรรมเนื่องจากความหลากหลายของผู้รับสารมีมากเช่นกัน ในบางครั้งมีการเลือกฝ่ายของสื่อ (Fairbanks et al., 2018) การสร้างข่าวปลอมเพื่อใส่ร้ายป้ายสีฝ่ายตรงข้ามเกิดขึ้นมาก ส่งผลให้ปัจจุบันประชาชน สถาบันทางการเมือง และสังคมถูกปลุกปั่นเพิ่มมากขึ้น

เรื่อย ๆ (Kokkeadtikul & Danphaibun, 2018)

1.1.3 ด้านกระบวนการผลิต

สื่อออนไลน์ที่เน้นผลลัพธ์ในการนำเสนอสารเพียงอย่างเดียวจะทำให้คุณภาพของสารที่นำเสนอออกมาไม่มีคุณภาพ ซึ่งสื่อที่มีความน่าเชื่อถือจะต้องมีกระบวนการผลิตที่ทันสมัย สามารถควบคุมคุณภาพได้ เนื่องจากปัจจุบันบุคคลทั่วไปก็สามารถผลิตและนำเสนอสื่อได้ จึงทำให้ความน่าเชื่อถือของสื่อออนไลน์นั้นมีน้อย ซึ่งในการอ้างอิงทางวิชาการก็ไม่ยอมรับการอ้างอิงแหล่งข้อมูลออนไลน์ ดังนั้นกระบวนการผลิตที่ดีจะทำให้มีความน่าเชื่อถือกับสื่อที่ผลิตออกมาได้ โดยปัจจัยที่มีอิทธิพลต่อความน่าเชื่อถือด้านกระบวนการผลิตมีปัจจัยย่อย ได้แก่ องค์กรของผู้ผลิต เทคโนโลยีที่ใช้ในการผลิต อัตลักษณ์ของผู้ผลิต และมาตรฐานการผลิตข่าว

1.2 ปัจจัยที่มีอิทธิพลต่อการยอมรับ

1.2.1 ปัจจัยด้านความหลากหลายของเนื้อหา

เนื่องจากจำนวนของแหล่งข่าวที่มีมากและสามารถค้นหาได้ง่ายทำให้ความหลากหลายของเนื้อหา ความถูกต้องของเนื้อหา และการนำเสนอเนื้อหาที่มีความละเอียดลึก มีรายละเอียดที่ครอบคลุมส่งผลให้เกิดความยอมรับในสื่อนั้น

1.2.2 ปัจจัยด้านรูปแบบการนำเสนอ

ปัจจัยนี้เกี่ยวข้องกับการออกแบบเว็บไซต์ให้มีความสะดวกต่อการใช้งานของผู้รับสาร เช่น การนำเสนอข่าวสารแบบเบ็ดเสร็จในหน้าเดียว การมีปฏิสัมพันธ์กับข่าวที่นำเสนอ มีการนำเสนอข่าวเด่น ดังนั้นควรคำนึงถึงความง่ายในการใช้งานของผู้รับสาร

1.2.3 ปัจจัยด้านชื่อเสียงและจรรยาบรรณ

ปัจจัยนี้ต้องอาศัยระยะเวลาในการสะสมชื่อเสียงโดยเกิดจากการนำเสนอข่าวสารที่มีความถูกต้อง มีจรรยาบรรณของนักสื่อสารมวลชนในการนำเสนอข่าวจะทำให้ค่อย ๆ ได้รับการยอมรับจากผู้รับข่าวจนมีชื่อเสียง

1.2.4 ปัจจัยด้านความกระชับและเชื่อมโยง

เนื้อหาของข่าวสารควรมีความกระชับและชัดเจน มีการเชื่อมโยงไปยังข่าวอื่นที่มีความเกี่ยวข้องหรือเชื่อมโยงไปยังแหล่งที่มาของข่าว ทำให้ผู้รับข่าวสามารถค้นหาที่มาของข่าวหรืออ่านรายละเอียดของข่าวเพิ่มเติมจากแหล่งอื่นได้อย่างสะดวก

1.2.5 ปัจจัยด้านความบันเทิง

เว็บไซต์ข่าวออนไลน์ในปัจจุบันไม่ได้นำเสนอเพียงแค่เนื้อหาของข่าวสารเพียงอย่างเดียว แต่ยังต้องมีการสอดแทรกความบันเทิงหรือข่าวสารที่ไม่เป็นทางการ เนื่องจากผู้รับข่าวในบางครั้งยังต้องการรับข่าวด้านความบันเทิง หรือไม่มีสาระอยู่บ้าง แต่ไม่ควรสอดแทรกเนื้อหาประเภทนี้มากเกินไปจนเป็นการรบกวนการใช้งานของผู้รับข่าว

2. เหมืองข้อมูล (data mining)

เหมืองข้อมูล (data mining) เป็นกระบวนการค้นหาองค์ความรู้ที่น่าสนใจ เช่น รูปแบบ (patterns) ความสัมพันธ์ (associations) การเปลี่ยนแปลง (changes) ความผิดปกติ (anomalies) หรือโครงสร้างที่มีนัยสำคัญ (significant structures) จากข้อมูลจำนวนมากที่เก็บไว้ในฐานข้อมูลหรือคลังข้อมูล สำหรับเว็บไซต์ข่าวออนไลน์มีการนำเสนอข่าวภายใต้หมวดหมู่ต่าง ๆ เช่น ข่าวระดับชาติ (national news) ข่าวระดับนานาชาติ (international news) ข่าวการเมือง (politics news) ข่าวการเงิน (finance news) ข่าวกีฬา (sports news) และ ข่าวบันเทิง (entertainment news) เป็นต้น ซึ่งมีความคล้ายคลึงกับการจำแนกข้อมูล (data classification) ของการทำเหมืองข้อมูลโดยขั้นตอนการจำแนกข้อมูลเริ่มต้นด้วย ชุดข้อมูลที่ใช้สำหรับการสอนคอมพิวเตอร์ (training set) ซึ่งจะต้องมีคลาส (Class) คำตอบเฉลยเอาไว้แล้วเพื่อให้คอมพิวเตอร์ทำการเรียนรู้ที่จะจำแนกข้อมูลตามคลาสที่กำหนดเอาไว้ในแต่ละข้อมูล แต่ในกรณีที่ข้อมูลไม่มีการกำหนดคลาสมาก่อนสามารถทำการกำหนดคลาสให้กับข้อมูลได้โดยการเรียนรู้ของเครื่องโดยใช้การจัดกลุ่มข้อมูล (data clustering)

2.1 ขั้นตอนการเตรียมข้อมูล (data preparation)

สำหรับขั้นตอนนี้จะเป็นเตรียมข้อมูลให้เหมาะสมกับการทำเหมืองข้อมูล โดยการนำข้อมูลปัจจัยที่ส่งผลต่อความน่าเชื่อถือของเว็บไซต์ ที่ผ่านการทำให้ค่าอยู่ในช่วงปกติ (normalization) โดยให้อยู่ในช่วงใหม่ ด้วยสมการที่ 1

$$v' = \frac{v\text{-}min_A}{max_A\text{-}min_A}(new\_max_A\text{-}new\_min_A) + new\_min_A \quad (1)$$

โดยกำหนดให้

v' คือ ค่าที่ต้องการทำให้อยู่ในช่วงใหม่
min_A คือ ค่าน้อยที่สุดของแอททริบิว
max_A คือ ค่ามากที่สุดของแอททริบิว
new_min_A คือ ค่าน้อยที่สุดของช่วงใหม่
new_max_A คือ ค่ามากที่สุดของช่วงใหม่

2.2 ค่าประสิทธิภาพการจัดกลุ่มข้อมูล

เป็นการหาค่าอัตราส่วนของความแปรปรวนระหว่างกลุ่มกับความแปรปรวนภายในกลุ่ม เมื่อค่าที่ได้ออกมามีค่ามากแสดงถึงมีการจัดกลุ่มที่ดี สำหรับฟังก์ชันการคำนวณค่า Calinski-Harabasz Index แสดงดังสมการ (2)

$$s(k) = \frac{Tr(B_k)}{Tr(W_k)} \times \frac{N-k}{k-1} \quad (2)$$

โดย

N คือ จำนวนข้อมูลทั้งหมด
k คือ จำนวนกลุ่มที่ทำการแบ่ง
B_k คือ เมทริกซ์การกระจายตัวระหว่างกลุ่ม
W_k คือ เมทริกซ์การกระจายภายในกลุ่ม

2.3 ค่าประสิทธิภาพการจำแนกข้อมูล

ค่าความถูกต้อง (accuracy) เป็นการวัดค่าความถูกต้องโดยพิจารณาจากทุกคลาสร่วมกัน แสดงดังสมการที่ 3

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

ค่าความแม่นยำ (precision) เป็นการวัดความแม่นยำของโมเดลโดยพิจารณาแยกทีละคลาสแล้วนำมาหาค่าเฉลี่ย แสดงดังสมการที่ 4

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

ค่าเรียกค้นคืน (recall) เป็นการวัดการค้นคืนของโมเดลโดยพิจารณาแยกทีละคลาสแล้วนำมาหาค่าเฉลี่ย แสดงดังสมการที่ 5

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

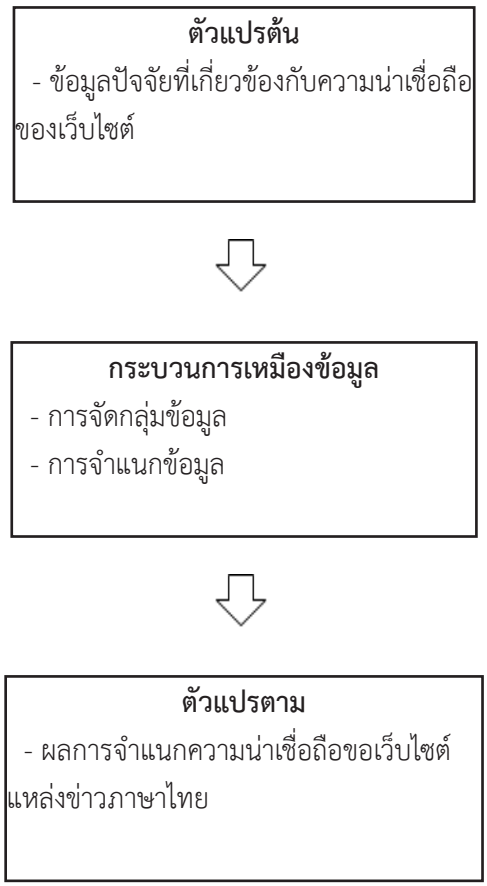
โดยที่กำหนดให้ TP = จำนวนค่าที่ทำนายว่าจริงแล้วเป็นจริง, FP = จำนวนค่าที่ทำนายว่าจริงแล้วเป็นเท็จ, TN = จำนวนค่าที่ทำนายว่าเท็จแล้วเป็นเท็จ และ FN = จำนวนค่าที่ทำนายว่าเท็จแล้วเป็นจริง

3. งานวิจัยที่เกี่ยวข้อง

ในการจำแนกหมวดหมู่ของข่าวมีการเปรียบเทียบประสิทธิภาพของอัลกอริธึมการเรียนรู้แบบอัตโนมัติหลายเทคนิค ดังนี้ Find Similar, Decision Trees, Naïve Bayes, Bayes Nets, Support Vector Machines–SVM และ Neural Networks กันสำหรับการจัดหมวดหมู่ข้อความ โดยสามารถพิจารณาในแง่ของความเร็วในการเรียนรู้ (terms of learning speed) ความเร็วการจำแนกแบบเรียลไทม์ (Real Time classification speed) และความแม่นยำในการจัดหมวดหมู่ (Dumais et al., 1998; Shahi & Pant, 2018) อีกทั้งยังมีการเปรียบเทียบประสิทธิภาพของตัวกรอง (filters) และใช้อัลกอริธึม ดังนี้ Naïve Bayes, C45, Decision Tree, Max Entropy, Winnow และ Balanced Winnow สำหรับงานการจำแนกข้อมูลความเชื่อมั่นของข่าว (Li et al., 2017) ซึ่งในการจัดหมวดหมู่บทความข่าวจากแหล่งข่าวออนไลน์และติดป้ายกำกับโดยอัตโนมัติตามโดเมนของข่าวงานนี้ใช้แนวคิดเรื่องการประมวลผลภาษาธรรมชาติและการเรียนรู้ด้วยเครื่อง โดย

การเลือกใช้เทคนิค K-Means และ Incremental Clustering ความถูกต้องของผลลัพธ์ของรูปแบบที่นำเสนอเป็นที่น่าพึงพอใจ ผลลัพธ์เหล่านี้สามารถใช้เพื่อเปรียบเทียบความสนใจของผู้ชมที่แตกต่างกันในแต่ละแหล่งข่าวได้ (Krishnamoorthy et al., 2018)

กรอบแนวคิดการวิจัย



ภาพ 1 กรอบแนวคิดงานวิจัย

จากการศึกษาทฤษฎีและงานวิจัยที่เกี่ยวข้องและนำทฤษฎีที่ได้มาประยุกต์ใช้เพื่อพัฒนาการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย โดยมีกรอบแนวคิดงานวิจัย แสดงดังภาพ 1

จากภาพข้างต้นกรอบแนวคิดของงานวิจัยประกอบไปด้วย ตัวแปรต้นซึ่งใช้ข้อมูลปัจจัยที่เกี่ยวข้องกับความน่าเชื่อถือของเว็บไซต์แหล่งข่าว แล้วนำมาเก็บลงฐานข้อมูล จากนั้นผู้วิจัยจึงใช้กระบวนการทำเหมืองข้อมูลโดยใช้การจัดกลุ่มข้อมูลก่อน เนื่องจากเว็บไซต์แหล่งข่าวนั้นยังไม่เคยมีการกำหนด Class ว่ามีจำนวนกี่ Class ถึงจะดีที่สุด เมื่อทำการกำหนด Class ให้กับแหล่งข่าวแล้วจึงดำเนินการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทยโดยใช้ข้อมูลและ Class ที่ผ่านการกำหนดมาจากขั้นตอนการจัดกลุ่มข้อมูล จากนั้นทำการเปรียบเทียบค่าประสิทธิภาพของเทคนิคการจำแนกข้อมูล โดยตัวแปรตามผู้วิจัยกำหนดไว้ว่าผลการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทยมีประสิทธิภาพดี ผู้วิจัยได้วางแผนการดำเนินงานในการศึกษาและพัฒนาการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย แสดงดังภาพ 2 และภาพ 3

วิธีดำเนินการวิจัย

ผู้วิจัยได้วางแผนการดำเนินงานในการศึกษาและพัฒนาการจำแนกความน่าเชื่อถือของเว็บไซต์แหล่งข่าวภาษาไทย แสดงดังภาพ 2 และภาพ 3

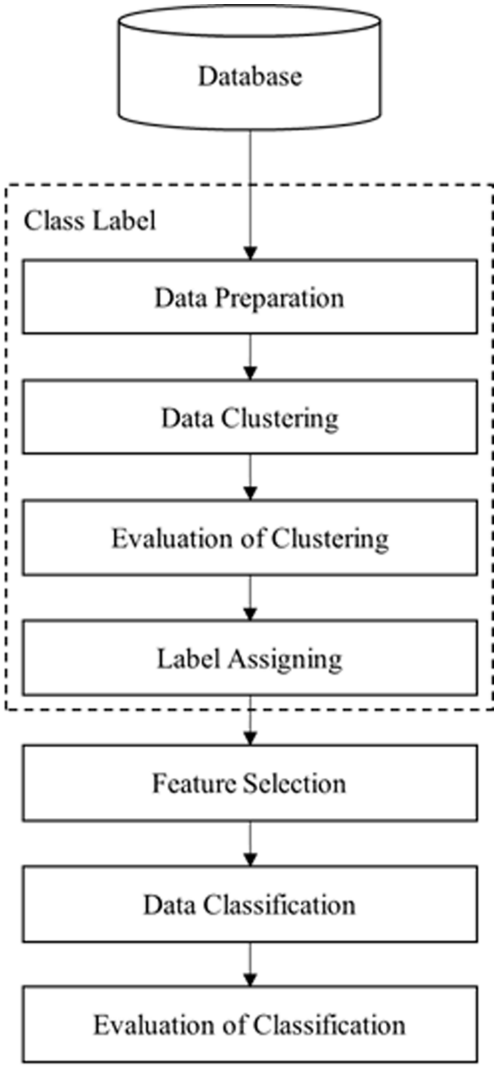
1. ขั้นตอนการกำหนดคลาส

1.1 ขั้นตอนการเตรียมข้อมูล

ขั้นตอนนี้จะทำการดึงข้อมูลจากฐานข้อมูลที่ทำการรวบรวมข้อมูลปัจจัยที่ส่งผลต่อความน่าเชื่อถือของเว็บไซต์จำนวน 35 ตัวชี้วัด มีรายละเอียดแสดงดังตาราง 1 จากเว็บไซต์แหล่งข่าวภาษาไทยจำนวน 168 เว็บไซต์ ประกอบด้วย เว็บไซต์สำนักข่าวหลักในประเทศ เว็บไซต์สำนักข่าวออนไลน์และเว็บไซต์ข่าวที่นิยมถูกส่งต่อกันในสื่อสังคม เมื่อทำการดึงข้อมูลจากฐานข้อมูลแล้วนำมาผ่านกระบวนการทำข้อมูลให้เหมาะสมกับการจัดกลุ่มและจำแนกข้อมูลทำให้ค่าอยู่ในช่วงปกติ (normalization) ให้อยู่ในค่าช่วง 0-1 เช่น ข้อมูล Total Visits ค่าปกติจะอยู่ในช่วง 0–770,570,000 และข้อมูลจำนวนปี พ.ศ. ที่จด Domain Name ค่าปกติอยู่ในช่วง 0–30 ซึ่งจะเห็นว่าข้อมูลมีความแตกต่างกันมากทำให้ต้องทำให้อยู่ในช่วงปกติ

1.2 ขั้นตอนการจัดกลุ่มข้อมูล

ผู้วิจัยได้เลือกเทคนิคการจัดกลุ่มข้อมูลเนื่องจากเว็บไซต์แหล่งข่าวยังไม่มีการกำหนดป้ายกำกับที่ชัดเจนว่าแบ่งออกเป็นกี่กลุ่ม เพื่อให้สามารถจำแนกข้อมูลเว็บไซต์แหล่งข่าวได้จึงทำการจัดกลุ่มข้อมูลก่อนด้วยเทคนิค K-Means Clustering และ Hierarchical Clustering (single linkage, average linkage และ complete linkage) แล้วนำผลลัพธ์ในการจัดกลุ่มข้อมูลมาทำการเปรียบเทียบว่าการจัดกลุ่มด้วยเทคนิคแบบใดให้ประสิทธิภาพการจัดกลุ่มได้ดีที่สุด



ภาพ 2 วิธีดำเนินการวิจัย

ตาราง 1

ข้อมูลปัจจัยที่ส่งผลต่อความน่าเชื่อถือ

	ปัจจัยย่อย
1	ข้อมูลการอ้างอิงแหล่งข่าว
2	ข้อมูลลายน้ำในรูปภาพประกอบข่าว
3	ข้อมูลผู้รายงานข่าว/ผู้เขียนข่าว
4	ข้อมูลการติดต่อกับแหล่งข่าว (ที่อยู่)
5	ข้อมูลการติดต่อกับแหล่งข่าว (เบอร์โทร)
6	ข้อมูลการติดต่อกับแหล่งข่าว (อีเมลล์)
7	คะแนนการเข้าถึง (accessibility score)
8	คะแนนการทำตาม Best Practice ของ Modern Web Development (Best Practice Score)
9	คะแนนการทำ Search Engine Optimization--SEO Score
10	คะแนน Progressive Web App--PWA Score
11	การเป็นสมาชิกของชมรมผู้ผลิตข่าวออนไลน์ (Society of Online News Providers: SONP)
12	การเป็นสมาชิกของสมาคมนักข่าวนักหนังสือพิมพ์แห่งประเทศไทย (Thai Journalist Association--TJA)
13	จำนวนปี พ.ศ. ที่จด Domain Name
14	ข้อมูลนโยบายความเป็นส่วนตัว
15	ข้อมูล Alexa Traffic Rank
16	ข้อมูล Google Pages Indexed
17	ข้อมูล Bing Pages Indexed
18	ข้อมูล Global Rank
19	ข้อมูล Country Rank
20	ข้อมูล Total Visits
21	ข้อมูล Referring Sites
22	ข้อมูล Destination Sites
23	ข้อมูล Internal Links
24	ข้อมูล External Links
25	ข้อมูลการลิงค์ไปยัง Facebook
26	ข้อมูลจำนวนผู้ติดตาม Facebook
27	ข้อมูลการลิงค์ไปยัง Twitter
28	ข้อมูลจำนวนผู้ติดตาม Twitter
29	ข้อมูลการลิงค์ไปยัง Instagram
30	ข้อมูลจำนวนผู้ติดตาม Instagram

ตาราง 1 (ต่อ)

	ปัจจัยย่อย
31	ข้อมูลการลิงค์ไปยัง YouTube
32	ข้อมูลจำนวนผู้ติดตาม YouTube
33	ข้อมูลการอัพเดทข่าวสาร
34	ข้อมูลการแสดงวันที่ลงข่าวในเนื้อหาข่าว

1.3 ขั้นตอนการเปรียบเทียบค่าประสิทธิภาพการจัดกลุ่มข้อมูล

ในขั้นตอนการวัดประสิทธิภาพของข้อมูลที่ได้ทำการจัดกลุ่ม ซึ่งจะพิจารณาจากค่า Calinski-Harabaz Index แล้วทำการเปรียบเทียบค่าที่มากที่สุดในการจัดกลุ่ม

1.4 ขั้นตอนการกำหนดติดป้ายกำกับ

เมื่อได้วิธีการจัดกลุ่มและจำนวนกลุ่มที่มีค่า Calinski-Harabaz Index มากที่สุดแล้ว ผู้วิจัยได้ทำการแปลความหมายของแต่ละกลุ่มเพื่อกำหนดป้ายกำกับให้แต่ละกลุ่ม

2. ขั้นตอนการจำแนกข้อมูล

ขั้นตอนนี้จะทำการดึงข้อมูลของเว็บไซต์แหล่งข่าวออนไลน์ที่ทำการกำหนดป้ายกำกับให้กับเว็บไซต์แหล่งข่าวแล้วนำมาทำเหมืองข้อมูล ด้วยเทคนิคการจำแนกข้อมูล (data classification) 5 เทคนิค ได้แก่ Decision Tree--C4.5, Naïve Bayes, K-Nearest Neighbor--K-NN, Multilayer Perceptron และ Support Vector Machine--SVM โดยผู้วิจัยได้ใช้ภาษา Python ในการพัฒนาโมเดล และทำการทดสอบโมเดลทั้ง 5 เทคนิค

3. ขั้นตอนการเปรียบเทียบค่าประสิทธิภาพการจำแนกข้อมูล

หลังจากทำการสร้างโมเดลจากเทคนิคการจำแนกข้อมูลแล้ว ผู้วิจัยทำการเปรียบเทียบค่าประสิทธิภาพของโมเดลที่ได้จากเทคนิคต่าง ๆ ด้วยค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าการค้นคืน (recall) และค่าประสิทธิภาพโดยรวม (F-measure) เพื่อหาโมเดลที่มีประสิทธิภาพมากที่สุด

ผลการวิจัย

ผู้วิจัยได้นำเสนอผลการทดสอบประสิทธิภาพตามขั้นตอนการดำเนินการวิจัย ซึ่งได้ผลการวิจัยดังนี้

1. ผลการเปรียบเทียบประสิทธิภาพการจัดกลุ่มข้อมูล

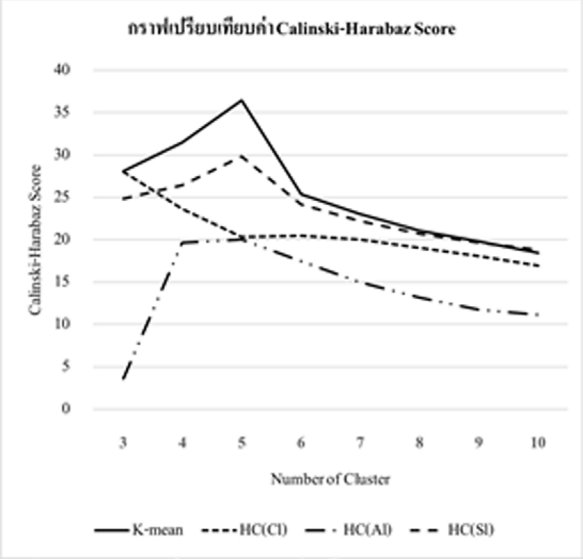
จากการนำข้อมูลปัจจัยที่ส่งผลต่อความน่าเชื่อถือของเว็บไซต์แหล่งข่าวมาทำการจัดกลุ่ม โดยเทคนิคการจัดกลุ่ม 2 เทคนิค ได้แก่ K-Means Clustering และ Hierarchical Clustering (Complete Linkage--HC(Cl), Average Linkage--HC(Al), Single Linkage--HC(Sl)) โดยได้กำหนดจำนวนการแบ่งกลุ่มไว้ตั้งแต่ 3 ถึง 10 กลุ่ม แล้วทำการเปรียบเทียบประสิทธิภาพของการจัดกลุ่ม ซึ่งพิจารณาจากค่า Calinski-Harabaz Index เพื่อดูการกระจายตัวของกลุ่ม ซึ่งผลการประเมินการกระจายตัวของข้อมูลเพื่อการจัดกลุ่ม แสดงดังตาราง 2

ตาราง 2

ผลค่า Calinski-Harabaz Index ของการจัดกลุ่ม

Cluster	K-means	HC(Cl)	HC(Al)	HC(Sl)
3	28.06	28.05	3.63	24.87
4	31.49	23.66	19.68	26.44
5	36.47	20.32	20.03	29.84
6	25.40	20.48	17.46	24.15
7	23.08	20.05	14.97	22.24
8	21.08	19.07	13.17	20.69
9	19.79	18.06	11.74	19.64
10	18.43	16.92	11.11	18.81

นำค่าประสิทธิภาพของการจัดกลุ่มข้อมูลใน Calinski-Harabaz Index แสดงผลดังภาพ 3 ตารางข้างต้นทำการสร้างกราฟเพื่อเปรียบเทียบค่า



ภาพ 3 กราฟเปรียบเทียบค่า Calinski-Harabaz Index

จากตารางและกราฟเปรียบเทียบค่าประสิทธิภาพการจัดกลุ่มเว็บไซต์แหล่งข่าวพบว่า เทคนิค K-Means Clustering มีค่าประสิทธิภาพมากที่สุดเมื่อ K=5 (Calinski-Harabaz Index=36.47) เทคนิค Hierarchical Clustering ใช้การเชื่อมต่อแบบ Complete Linkage มีค่าประสิทธิภาพมากที่สุดเมื่อ K=3 (Calinski-Harabaz Index=28.05) เทคนิค Hierarchical Clustering ใช้การเชื่อมต่อแบบ Average Linkage มีค่าประสิทธิภาพมากที่สุดเมื่อ K=5 (Calinski-Harabaz Index=20.03) เทคนิค Hierarchical Clustering ใช้การเชื่อมต่อแบบ Single Linkage มีค่าประสิทธิภาพมากที่สุดเมื่อ K=5 (Calinski-Harabaz Index=29.84) จะเห็นได้ว่าการแบ่งกลุ่มที่ K=5 มีประสิทธิภาพดีที่สุดจำนวน 3 เทคนิค ส่วน K=3 มีเพียง 1 เทคนิค ดังนั้นผู้วิจัยจึงทำการแบ่งกลุ่มของเว็บไซต์แหล่งข่าวออกเป็น 5 กลุ่ม โดยใช้เทคนิค K-Means Clustering

2. ผลการกำหนดคลาส

หลังจากที่ผู้วิจัยทำการแบ่งกลุ่มของเว็บไซต์แหล่งข่าวออกเป็น 5 กลุ่ม โดยเทคนิค K-Means Clustering แล้วดำเนินการแปลความหมายของแต่ละกลุ่มเพื่อจะได้กำหนดคลาสให้กับเว็บไซต์แหล่งข่าว โดยมีรายละเอียดของทั้ง 5 กลุ่มโดยเรียงจากกลุ่มที่มีความน่าเชื่อถือมากที่สุดไปจนถึงกลุ่มที่มีความน่าเชื่อถือน้อยที่สุด แสดงดังตาราง 3

ตาราง 3

รายละเอียดของพฤติกรรมแต่ละกลุ่ม

กลุ่ม	รายละเอียด
1	- มีการอ้างอิงแหล่งข่าวน้อย
	- มีการทำลายน้ำในรูปภาพประกอบข่าวมาก
	- มีข้อมูลผู้รายงานข่าว/ผู้เขียนข่าวน้อย
	- มีข้อมูลติดต่อกับแหล่งข่าวมาก
	- เว็บไซต์มีโครงสร้างที่เหมาะสมกับมือถือและ SEO
	- เป็นสมาชิกของสมาคมหรือองค์รใดที่เกี่ยวกับสื่อมากที่สุด
	- มีระยะเวลาในการจด Domain มาก
	- มีนโยบายรักษาข้อมูลส่วนตัวมากที่สุด
	- เว็บไซต์อยู่ในระดับความนิยมของ Alexa, Google และ Bing มากที่สุด
	- มีอัตรา Referring Sites มาก
	- มีการเชื่อมโยงกับ Social Media มาก
	- มีการอัพเดทข่าวสารตลอดเวลา
	- มีการแสดงวันที่ลงข่าวในเนื้อหาข่าว

ตาราง 3 (ต่อ)

กลุ่ม	รายละเอียด
2	- มีการอ้างอิงแหล่งข่าวมาก
	- มีการทำลายน้ำในรูปภาพประกอบข่าวมากที่สุด
	- มีข้อมูลผู้รายงานข่าว/ผู้เขียนข่าวมากที่สุด
	- มีข้อมูลติดต่อกับแหล่งข่าวปานกลาง
	- เว็บไซต์มีโครงสร้างที่เหมาะสมกับมือถือและ SEO
	- เป็นสมาชิกของสมาคมหรือองค์รใดที่เกี่ยวกับสื่อมาก
	- มีระยะเวลาในการจด Domain ปานกลาง
	- มีนโยบายรักษาข้อมูลส่วนตัวน้อย
	- เว็บไซต์อยู่ในระดับความนิยมของ Alexa, Google และ Bing มาก
	- มีอัตรา Referring Sites ปานกลาง
	- มีการเชื่อมโยงกับ Social Media มาก
	- มีการอัพเดทข่าวสารตลอดเวลา
	- มีการแสดงวันที่ลงข่าวในเนื้อหาข่าว
3	- มีการอ้างอิงแหล่งข่าวปานกลาง
	- มีการทำลายน้ำในรูปภาพประกอบข่าวน้อย
	- มีข้อมูลผู้รายงานข่าว/ผู้เขียนข่าวปานกลาง
	- มีข้อมูลติดต่อกับแหล่งข่าวมากที่สุด
	- เว็บไซต์มีโครงสร้างที่เหมาะสมกับมือถือและ SEO
	- เป็นสมาชิกของสมาคมหรือองค์รใดที่เกี่ยวกับสื่อปานกลาง
	- มีระยะเวลาในการจด Domain มากที่สุด
	- มีนโยบายรักษาข้อมูลส่วนตัวน้อย
	- เว็บไซต์อยู่ในระดับความนิยมของ Alexa, Google และ Bing ปานกลาง
	- มีอัตรา Referring Sites ปานกลาง
	- มีการเชื่อมโยงกับ Social Media มาก
	- มีการอัพเดทข่าวสารตลอดเวลา
	- มีการแสดงวันที่ลงข่าวในเนื้อหาข่าว

ตาราง 3 (ต่อ)

กลุ่ม	รายละเอียด
4	- มีการอ้างอิงแหล่งข่าวมากที่สุด - มีการทำลายน้ำในรูปภาพประกอบข่าวปานกลาง - มีข้อมูลผู้รายงานข่าว/ผู้เขียนข่าวมาก - มีข้อมูลติดต่อกับแหล่งข่าวน้อย - เว็บไซต์มีโครงสร้างที่เหมาะสมกับมือถือและ SEO - ไม่เป็นสมาชิกของสมาคมหรือองค์รใดที่เกี่ยวกับสื่อ - มีระยะเวลาในการจด Domain น้อย - มีนโยบายรักษาข้อมูลส่วนตัวน้อย - เว็บไซต์อยู่ในระดับความนิยมของ Alexa, Google และ Bing น้อย - มีอัตรา Referring Sites มาก - มีการเชื่อมโยงกับ Social Media น้อย - มีการอัพเดทข่าวสารปานกลาง - มีการแสดงวันที่ลงข่าวในเนื้อหาข่าว
5	- ไม่มีการอ้างอิงแหล่งข่าว - ไม่มีการทำลายน้ำในรูปภาพประกอบข่าว - ไม่มีข้อมูลผู้รายงานข่าว/ผู้เขียนข่าว - มีข้อมูลติดต่อกับแหล่งข่าวน้อย - เว็บไซต์ไม่มีโครงสร้างที่เหมาะสมหรืออาจถูกปิดไปแล้ว - ไม่เป็นสมาชิกของสมาคมหรือองค์รใดที่เกี่ยวกับสื่อ - มีระยะเวลาในการจด Domain น้อยที่สุด - ไม่มีนโยบายรักษาข้อมูลส่วนตัว - เว็บไซต์อยู่ในระดับความนิยมของ Alexa, Google และ Bing น้อยที่สุด - มีอัตรา Referring Sites น้อยที่สุด - มีการเชื่อมโยงกับ Social Media น้อยที่สุด - ไม่มีการอัพเดทข่าวสาร - ไม่มีการแสดงวันที่ลงข่าวในเนื้อหาข่าว

3. ผลการเปรียบเทียบค่าประสิทธิภาพการจำแนกข้อมูล

เป็นการนำข้อมูลปัจจัยที่ส่งผลต่อความน่าเชื่อถือที่มีการจัดกลุ่มแล้วติดป้ายกำกับ (Label) แล้วนำมาทำการจำแนกข้อมูลแหล่งข่าว โดยเปรียบเทียบเทคนิคการจำแนกข้อมูล 5 เทคนิค ได้แก่ Decision Tree--C4.5, Naïve Bayes, K-Nearest Neighbor--K-NN (กำหนดค่า K 2 ถึง 7), Multilayer Perceptron และ Support Vector Machine--SVM สำหรับการประเมินผลการจำแนกข้อมูลพิจารณาจากค่าความถูกต้อง (accuracy) ค่าความแม่นยำ (precision) ค่าค้นคืน (recall) และค่าประสิทธิภาพโดยรวม (F-measure) ซึ่งผลการประเมินการกระจายตัวของข้อมูลเพื่อการจัดกลุ่ม ดังตาราง 4

ตาราง 4

ผลการเปรียบเทียบประสิทธิภาพการจำแนกเว็บไซต์แหล่งข่าว

Model	Accuracy	Precision	Recall	F-measure
Decision Tree (C4.5)	86.09	0.861	0.861	0.861
Naïve Bayes	68.87	0.739	0.689	0.695
K-NN (K=2)	92.72	0.932	0.927	0.926
K-NN (K=3)	92.72	0.929	0.927	0.926
K-NN (K=4)	93.38	0.938	0.934	0.933
K-NN (K=5)	96.03	0.962	0.960	0.959
K-NN (K=6)	96.03	0.962	0.960	0.959
K-NN (K=7)	96.03	0.962	0.960	0.959
Multilayer Perceptron	94.70	0.948	0.947	0.946
SVM	94.70	0.947	0.947	90.47

จากตารางเปรียบเทียบประสิทธิภาพการจำแนกเว็บไซต์แหล่งข่าวพบว่า เทคนิค K-Nearest Neighbor--K-NN มีค่าประสิทธิภาพมากที่สุดเท่ากัน 3 เทคนิคโดยที่ค่า K มีค่า 5 ถึง 7 โดยมีค่า (Accuracy=96.03%, Precision=0.962, Recall=0.960, F-measure=0.959) ซึ่งผู้วิจัยเลือกใช้เทคนิค K-Nearest Neighbor--K-NN เมื่อ K เท่ากับ 6 เนื่องจากทำให้มีอำนาจจำแนกได้ดีกับจำนวน 5 กลุ่ม

การอภิปรายผล

เนื่องจากเว็บไซต์แหล่งข่าวภาษาไทยไม่มีการติดป้ายกำกับคลาส (label) มาก่อนจึงต้องมีการจัดกลุ่มข้อมูล (data clustering) ของเว็บไซต์แหล่งข่าวออนไลน์ก่อนที่จะทำการจำแนกข้อมูล (data classification) เว็บไซต์แหล่งข่าวออนไลน์ โดยในขั้นตอนของการจัดกลุ่มข้อมูลแหล่งข่าวพบว่า เทคนิค K-Means Clustering มีค่าประสิทธิภาพมากที่สุดเมื่อ K=5 (Calinski-Harabaz Index=36.47) จึงทำการติดป้ายกำกับเว็บไซต์แหล่งข่าวด้วยเทคนิค K-Means Clustering ซึ่งใช้การจัดกลุ่มเพื่อติดป้ายกำกับแหล่งข่าวและข่าวสารสอดคล้องกับงานวิจัย (Krishnamoorthy et al., 2018) ทำให้การวิเคราะห์ข้อมูลที่ไม่มีการกำหนดคลาสมาก่อนสามารถทำได้แม่นยำมากขึ้น

แล้วจึงนำไปทำการจำแนกข้อมูลแหล่งข่าว พบว่า เทคนิค K-Nearest Neighbor--K-NN (K=6) มีค่าประสิทธิภาพมากที่สุด (accuracy=96.03%, precision=0.962, recall=0.960, F-measure=0.959) ซึ่งการจำแนกข้อมูลด้วยเทคนิค Multilayer Perceptron นี้จะถูกนำไปพัฒนาเครื่องมือใช้ในการตรวจสอบความน่าเชื่อถือของข่าวออนไลน์ซึ่งสอดคล้องกับงานวิจัยของ (Shah & Ravana, 2014) ที่สามารถใช้การเทคนิคเรียนรู้ของเครื่องในการตรวจสอบความน่าเชื่อถือของเว็บไซต์

ข้อเสนอแนะ

ในการจัดกลุ่มแหล่งข่าวออนไลน์ สามารถใช้ปัจจัยอื่นที่เกี่ยวข้องเข้ามาช่วยในการจัดกลุ่มแหล่งข่าวออนไลน์ได้ เช่น ปัจจัยด้านความเป็นกลางของเนื้อหา ซึ่งจะต้องทำการวิเคราะห์อารมณ์ของเนื้อหาของข่าวว่ามีความเป็นกลาง หรือเอนเอียงไปทางใดทางหนึ่งหรือไม่ หรือปัจจัยด้านคุณภาพของเนื้อหาของข่าว ต้องทำการวิเคราะห์ความถูกต้องในการพิมพ์เนื้อหาข่าว หรือเนื้อหาข่าวใกล้เคียงกับแหล่งข่าวอื่นหรือไม่

References

Chumwatana, T. (2013). A survey of Automatic Indexing Techniques for Thai Text documents. *Information Technology Journal, 9*(1), 81-91. (in Thai)

Dumais, S., Platt, J., & Heckerman, D. (1998). Inductive learning algorithm and representation for text categorization. *In Conference of Information and Knowledge Management (CIKM)* (pp. 148-155). Maryland, USA.: CIKM. doi: 10.1145/288627.288651.

Fairbanks, J., Fitch, N., Knauf, N., & Briscoe, E., (2018). Credibility assessment in the News: Do we need to read?. *In Misinformation and Misbehavior Mining on the Web* (pp. 1-8). CA., USA.: James P. Fairbanks. http://jpfairbanks.net/publication/mis2-2018/.

Kokkeadtikul, C., & Danphaibun, T., (2018). Fake news: Fake news problems, challenge and policy action. *NBTC Journal, 3*, 173-192. (in Thai)

Krishnamoorthy, A., Patil, A. K., Vasudevan, N., & Pathari, V. (2018). News article classification with clustering using Semi-Supervised Learning. *In International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 86-91). Bangalore, India: ICACCI

Li, J., Fong, S., Zhuang, Y., & Khoury, R. (2016). Hierarchical classification in text mining for sentiment analysis of online news. *Soft Computing, 20*(9), 3411–3420.

National Statistical Office. (2018). *The 2018 household survey on the use of Information and Communication Technology*. Bangkok: Economic and Social Statistics Bureau. (in Thai)

Shahi, T., & Pant, A. (2018). Nepali news classification using Naïve Bayes, support vector machines and neural networks. *In International Conference on Communication information and Computing Technology (ICCICT)* (pp. 1-5). Mumbai, India: ICCICT.

Toommanon, T., & Whattananarong, K. (2012). Creditability and innovation adoption of Online Newspapers. *Technical Education Journal King Mongkut's University of Technology North Bangkok, 3*(2), 25-33. (in Thai)