ตัวแบบเชิงลึกเพื่อตรวจจับข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์

Deep Model for Fake Thai News Detection on Social Network

ชวัล วัฒนากิจจากุล

Chawan Watthanakitjakul

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาการจัดการเทคโนโลยีสารสนเทศ
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Management of Information Technology
Prince of Songkla University

2564

ตัวแบบเชิงลึกเพื่อตรวจจับข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์

Deep Model for Fake Thai News Detection on Social Network

ชวัล วัฒนากิจจากุล

Chawan Watthanakitjakul

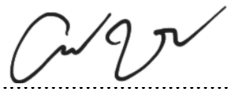
วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญา
วิทยาศาสตรมหาบัณฑิต สาขาวิชาการจัดการเทคโนโลยีสารสนเทศ
มหาวิทยาลัยสงขลานครินทร์

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science in Management of Information Technology
Prince of Songkla University

2564

ชื่อวิทยานิพนธ์ ตัวแบบเชิงลึกเพื่อตรวจจับข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์
ผู้เขียน นายชวัล วัฒนากิจจากุล
สาขาวิชา การจัดการเทคโนโลยีสารสนเทศ

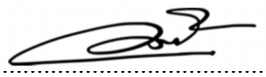
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก



.....

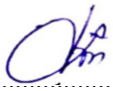
(ดร.อนันท์ ชกสุริวงค์)

คณะกรรมการสอบ

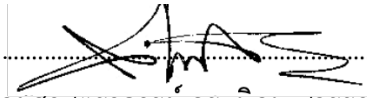


.....ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. วัชรวลี ตั้งคุปตานนท์)

.....กรรมการ
(ดร.อนันท์ ชกสุริวงค์)



.....กรรมการ
(รองศาสตราจารย์ ดร. วัฒนพงศ์ เกิดทองมี)



.....กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร. นิคม สุวรรณวร)

บัณฑิตวิทยาลัย มหาวิทยาลัยสงขลานครินทร์ อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษา ตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาการจัดการเทคโนโลยีสารสนเทศ

.....
(ศาสตราจารย์ ดร. ดำรงศักดิ์ ฟ้ารุ่งสาง)
คณบดีบัณฑิตวิทยาลัย

ขอรับรองว่า ผลงานวิจัยนี้มาจากการศึกษาวิจัยของนักศึกษาเอง และได้แสดงความขอบคุณบุคคลที่มีส่วนช่วยเหลือแล้ว

ลงชื่อ.....

(ดร.อนันท์ ชกสุริวงค์)

อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

ลงชื่อ.....

(นายชวัล วัฒนากิจจากุล)

นักศึกษา

ข้าพเจ้าขอรับรองว่า ผลงานวิจัยนี้ไม่เคยเป็นส่วนหนึ่งในการอนุมัติปริญญาในระดับใดมาก่อน และไม่ได้ถูกใช้ในการยื่นขออนุมัติปริญญาในขณะนี้

ลงชื่อ.....

(นายชวัล วัฒนากิจจากุล)

นักศึกษา

ชื่อวิทยานิพนธ์ ตัวแบบเชิงลึกเพื่อตรวจจับข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์
ผู้เขียน นายชวัล วัฒนากิจจากุล
สาขาวิชา การจัดการเทคโนโลยีสารสนเทศ
ปีการศึกษา 2564

บทคัดย่อ

ปัจจุบันข่าวปลอมบนสื่อสังคมออนไลน์ส่งผลกระทบอย่างมหาศาล เนื่องจากแพร่กระจายได้ง่ายและรวดเร็วกว่าข่าวจริง ซึ่งหากทำการตรวจสอบข่าวจำเป็นต้องใช้เวลา และใช้ทรัพยากรมนุษย์จำนวนมาก ผู้วิจัยจึงมีแนวคิดที่จะค้นหาคุณลักษณะที่สำคัญของข่าวปลอม และเปรียบเทียบตัวแบบการเรียนรู้ของเครื่องระหว่างตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม เพื่อค้นหาตัวแบบที่เหมาะสมกับการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ เพื่อช่วยลดการใช้ทรัพยากรในการตรวจสอบข่าวว่าเป็นข่าวปลอมหรือไม่ โดยผลลัพธ์ของคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมบนสื่อสังคมออนไลน์ทวิตเตอร์ได้แก่ จำนวนผู้ติดตาม คะแนนความรู้สึก ความยาวอักษรของเนื้อหา จำนวนการแบ่งปัน อัตราส่วนของเพื่อนและผู้ติดตาม จำนวนการกดชื่นชอบ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี โดยตัวแบบที่เหมาะสมกับการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ คือ ตัวแบบเทคนิคโครงข่ายประสาทเทียมซึ่งได้ความถูกต้องสูงถึง 97%

Thesis Title Deep Model for Fake Thai News Detection on Social Network
Author Mr. Chawan Watthanakitjakul
Major Program Management of Information Technology
Academic Year 2021

ABSTRACT

Nowadays, fake news on social media has caused many problems because they spread easier and faster than the real ones, while fake news detection or examination consumes high resources (human power, time, etc.). Thus, there is a need for an automatic method to examine or verify, so this research aims to find significant features of fake Thai news and an appropriate machine learning model between Decision tree, Support Vector Machine and Neural Network model to examine the fake Thai news on Twitter. The evaluation results show that the significant features of fake Thai news are the amount of follower, the sentiment score of news content, the length of content's character, the amount of retweet, the ratio of friend and follower, the amount of news favorited, the amount of post since signing up. The machine learning model that suits to examine the fake Thai news is a Neural Network model which performs 97 percent of accuracy.

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงได้ด้วยความกรุณา และการช่วยเหลือของ ดร.อนันท์ ชกสุริวงค์ อาจารย์ที่ปรึกษาหลักวิทยานิพนธ์ที่ได้ให้แรงบันดาลใจ กำลังใจ ความรู้ คำปรึกษา และให้ข้อเสนอแนะแนวทางในการดำเนินการวิจัย ตลอดจนช่วยตรวจ และแก้ไขวิทยานิพนธ์ให้มีความถูกต้องสมบูรณ์

ขอขอบคุณคุณธนัท วันวิน ผู้มีอาชีพเกี่ยวกับข่าวสาร ที่ให้ความช่วยเหลือในการตรวจสอบข่าวสาร และให้คำแนะนำ ข้อเสนอแนะ ข้อคิดเห็นต่าง ๆ ที่มีประโยชน์ต่อวิทยานิพนธ์ฉบับนี้

ขอขอบคุณคณะกรรมการสอบวิทยานิพนธ์ทุกท่าน ที่กรุณาให้คำแนะนำ และแนวทางในการปรับปรุงวิทยานิพนธ์ให้สมบูรณ์

ขอขอบคุณ ผู้ช่วยศาสตราจารย์ ดร. เพ็ชรัตน์ สุริยะไชย และผู้ช่วยศาสตราจารย์ ดร. วัชรวลี ตั้งคุปตานนท์ ที่ได้ให้กำลังใจ ความรู้ คำปรึกษา และให้ข้อเสนอแนะแนวทางในการดำเนินการวิจัยครั้งนี้ให้มีความถูกต้องสมบูรณ์ อีกทั้งขอขอบคุณคณาจารย์ในหลักสูตรทุกท่าน ที่ให้ความรู้ และคำปรึกษาทั้งในเนื้อหาวิชาและในการจัดทำวิทยานิพนธ์ฉบับนี้ให้ลุล่วงสมบูรณ์

ขอขอบคุณเจ้าหน้าที่หลักสูตรทุกท่านที่ให้ความคำแนะนำ ช่วยเหลือทั้งการดำเนินการ และการทำเอกสาร

ขอขอบคุณพี่ ๆ เพื่อน ๆ และ น้อง ๆ นักศึกษาปริญญาโท หลักสูตรการจัดการเทคโนโลยีสารสนเทศ มหาวิทยาลัยสงขลานครินทร์ ทุกท่านที่ได้ให้คำปรึกษา คำแนะนำ และเป็นกำลังใจที่ดีมาโดยตลอด

ขอขอบคุณเพื่อนกลุ่มไก่ชนของข้าพเจ้าทุกท่าน ที่ได้ให้คำปรึกษา ความช่วยเหลือ และให้กำลังใจที่ดีมาโดยตลอด

ขอขอบพระคุณผู้ช่วยศาสตราจารย์สุธน แซ่ว่องที่เป็นทั้งอาจารย์ พี่ และเพื่อนในขณะเดียวกัน ซึ่งเชื่อในตัวข้าพเจ้า และได้ให้ทั้งโอกาส ความรู้ทั้งในด้านคอมพิวเตอร์ การจัดการ และด้านการใช้ชีวิต ทำให้ข้าพเจ้ามีความรู้ความสามารถได้ถึงทุกวันนี้

สุดท้ายนี้ ข้าพเจ้าขอกราบขอบพระคุณบิดา มารดา และครอบครัวของข้าพเจ้าที่เลี้ยงดู ส่งเสริมสนับสนุน ให้โอกาส คำแนะนำ คำปรึกษา และกำลังใจแก่ข้าพเจ้ามาโดยตลอด จนทำให้ข้าพเจ้าประสบความสำเร็จได้ถึงทุกวันนี้

ชวัล วัฒนากิจจากุล

สารบัญ

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

สารบัญ(ต่อ)

หน้า

รายการตาราง

หน้า

รายการภาพประกอบ

หน้า

บทที่ 1
บทนำ

1.1 ความสำคัญและที่มาของการวิจัย

การแพร่กระจายของข่าวปลอมอย่างรวดเร็วและยากต่อการตรวจสอบ ระงับข่าว แก้ไข บนสื่อสังคมออนไลน์หรือโซเชียลมีเดียทวิตเตอร์ (Twitter) ซึ่งเป็นสื่อสังคมออนไลน์ที่นิยมอย่างมากในประเทศไทย อีกทั้งข่าวปลอมสามารถส่งผลเสียอย่างมหาศาลในภายหลัง ทั้งระดับท้องถิ่นไปจนถึงระดับประเทศได้ โดยมีผลของการศึกษาเรื่องของการแพร่กระจายของข่าวปลอมบนสื่อสังคมออนไลน์ [1] ของมหาวิทยาลัย Massachusetts Institute of Technology (MIT) ผลการศึกษาออกมาเป็นสถิติที่บ่งบอกว่าข่าวปลอมถูกแบ่งปันมากกว่าข่าวจริง แพร่กระจายได้รวดเร็วกว่าถึง 6 เท่า และมีอัตราการแบ่งปันมากกว่า 1.7 เท่า ดังตารางที่ 1-1

ตารางที่ 1-1 ผลการศึกษาเรื่องข่าวปลอมบนสื่อสังคมออนไลน์

หัวข้อ/ประเภท	ข่าวจริง	ข่าวปลอม
เวลาที่ใช้ในการแพร่กระจาย	ปกติ	เร็วกว่า 6 เท่า
จำนวนการเห็นข่าว	น้อยกว่า 1,000 คน	1,000 - 100,000 คน
อัตราการแบ่งปันให้กับผู้อื่น	ปกติ	สูงกว่า 1.7 เท่า

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่ามีการนำเทคนิคด้านการเรียนรู้ของเครื่อง มาช่วยในการแยกแยะข่าวปลอม ผลงาน S.Krishnan M.Chen [2] เป็นการนำเอากระบวนการ Netural Language Processing (NLP) มาช่วยในการหาคะแนนความรู้สึกในเนื้อหาร่วมกับคุณลักษณะของข่าวและนำไปใช้ในการจำแนกหมวดหมู่ข่าวจริงและปลอมด้วยเทคนิค Decision .Tree J48 และ Support Vector Machine (SVM) หรือผลงาน O.Ajao, D.Bhowmik, S.Zargari [3] ที่ได้นำเอาเทคนิค Recurrent Neural Network (RHN) แบบ Long Short Term Memory (LTSM) ร่วมกับ Convolutional Neural Network (CNN) เพื่อเปรียบเทียบตัวแบบพยากรณ์ที่เหมาะสมในการแยกแยะข่าวปลอม อีกทั้งงานวิจัยที่รวมเอาเทคนิคทางด้านการเรียนรู้ของเครื่องมาผสานกันออกเป็นตัวแบบ N.Ruchansky, S.Seo, Y.Liu [4] ที่ได้ทำการพัฒนาตัวแบบ CSI (Capture Score and Integrate) เพื่อทดสอบประสิทธิภาพการแยกแยะข่าวปลอมเปรียบเทียบกับเทคนิคเรียนรู้ของเครื่อง

อื่น อย่างเช่น Recurrent Neural Network (RNN) หรือ Support Vector Machine เป็นต้น และมีงานวิจัยใกล้เคียงกับการตรวจจับข่าวปลอม ซึ่งเป็นการตรวจจับข้อมูลข่าวขยะ (Spam Messages) โดยการใช้เทคนิคการเรียนรู้ของเครื่อง Fucntional Tree (FT), A Naive Bayes/Decision-Tree Hybrid (NBTree) และ Random Forest ในการนำมาพัฒนาและเปรียบเทียบผลการแยกแยะข้อความข่าวขยะและทดสอบความถูกต้อง หรืองานวิจัยของ J.Ma, W.Gao, et. al [5] ศึกษาเรื่องการแยกแยะข้อความที่เป็นข่าวลือ โดยการใช้เทคนิคโครงข่ายประสาทเทียมที่ชื่อว่า Recurrent Neural Network, Long Short-Term Memory และ Gated Recurrent Unit (GRU) เพื่อใช้ในการแยกแยะข่าวลือและเปรียบเทียบความถูกต้องกับเทคนิค อย่างไรก็ตามจากการศึกษางานวิจัยข้างต้น พบว่ายังมีข้อจำกัดดังนี้

1. งานวิจัยมีการใช้การเรียนรู้ของเครื่องมือที่มีข้อจำกัดเรื่องจำนวนคุณลักษณะของข่าวได้น้อย ส่งผลให้คุณลักษณะสำคัญของข่าวบางอย่างไม่ถูกใช้ในการพยากรณ์ซึ่งอาจส่งผลในการแยกแยะข่าวปลอม
2. การแยกแยะข่าวปลอมยังมีข้อจำกัดเรื่องเนื้อความ (Content Based) ที่สนใจเฉพาะเนื้อความที่เป็นภาษาอังกฤษเท่านั้น ซึ่งส่งผลทำให้การประยุกต์ใช้กับข่าวภาษาไทยอาจจะไม่ได้ผลความถูกต้องมากเท่าที่ควร
3. ตัวแบบการแยกแยะข่าวปลอมมีข้อจำกัดเรื่องข้อมูลที่เน้นไปทางใดทางหนึ่ง เช่น เหตุการณ์วาตภัย เฮอริเคนแซนดี้ เป็นต้น ซึ่งจะทำให้ตัวแบบอาจจะไม่ให้ผลความถูกต้องมากเท่าที่ควรในการนำไปใช้กับข้อมูลที่ไม่ขึ้นกับเหตุการณ์ เช่น ข่าวการเมือง ข่าวบันเทิง

จากปัญหาและข้อจำกัดข้างต้น ทำให้ผู้วิจัยเล็งเห็นถึงความสำคัญ โดยมีแนวคิดในการคัดเลือกคุณลักษณะที่สำคัญของข่าวที่ใช้ในการตรวจสอบข่าวปลอมและออกแบบพัฒนาตัวแบบการเรียนรู้ของเครื่องมาใช้ในการแยกแยะข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ โดยใช้เทคนิคการเรียนรู้ของเครื่องที่รองรับคุณลักษณะจำนวนมาก และมุ่งเน้นกับการแยกแยะข่าวปลอมที่เป็นภาษาไทยในสภาพแวดล้อมของผู้ใช้ในประเทศไทย โดยใช้ข้อมูลหลากหลายลักษณะ เช่น ข่าวการเมือง ข่าวบันเทิง มาใช้ในการสอนเครื่องเพื่อเรียนรู้ในการแยกแยะข่าวปลอม

โดยในประเทศไทยมีความนิยมใช้สื่อสังคมออนไลน์ทวิตเตอร์ และเฟซบุคอย่างแพร่หลาย ซึ่งจากข้อมูลของงานวิจัย [1] พบว่าความนิยมในการแพร่กระจายของข่าวปลอม และปริมาณของข่าวปลอมเกิดขึ้นบนสื่อสังคมออนไลน์ทวิตเตอร์มากกว่าสื่อสังคมออนไลน์เฟซบุคในปริมาณมาก เนื่องด้วยเฟซบุคมีระบบกรองหรือระบบจำกัดการมองเห็น ทำให้ข่าวปลอมที่แบ่งปันในเฟซบุคมีประมาณน้อย ดังนั้นผู้วิจัยจึงเลือกใช้สื่อสังคมออนไลน์ทวิตเตอร์ในการเก็บข้อมูลข่าวปลอมภาษาไทย

ผู้วิจัยมีแนวคิดในการค้นหาคุณลักษณะสำคัญที่เหมาะสมในการใช้ตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ด้วยวิธีต้นไม้ตัดสินใจแบบ Extra Trees Classifier [6] และ

การสร้างตัวแบบการเรียนรู้ของเครื่องเพื่อแยกแยะข่าวปลอมที่รองรับจำนวนคุณลักษณะของข่าวจำนวนมาก เนื้อความเป็นภาษาไทย และสามารถแยกแยะข้อมูลที่ไม่ขึ้นกับเหตุการณ์ได้ โดยการประยุกต์ใช้เทคนิค การประมวลผลภาษาธรรมชาติ (Natural Language Process) [7] เพื่อใช้ในการสกัดคุณลักษณะเชิงความรู้สึกจากข่าวภาษาไทยและแบ่งคำร่วมไปกับเทคนิคการเรียนรู้ของเครื่องต้นไม้ตัดสินใจ [8] และ ซัพพอร์ทเวกเตอร์แมชชีน [8] เปรียบเทียบกับเทคนิคการเรียนรู้ของเครื่องโครงข่ายประสาทเทียม [9,10] จากนั้นนำผลที่ได้มาทดสอบโดยการวัดค่าความถูกต้องของตัวแบบสำหรับคัดเลือกตัวแบบที่เหมาะสมในการแยกแยะข่าวปลอมภาษาไทย เพื่อให้ได้ตัวแบบการเรียนรู้ของเครื่องเชิงลึกที่สามารถแยกแยะข่าวปลอมที่มีความแม่นยำสูงกว่า 90% และนำคุณลักษณะที่ได้จากวิธีต้นไม้ตัดสินใจแบบ Extra Trees Classifier มาเปรียบเทียบกับคุณลักษณะที่สำคัญที่ได้จากตัวแบบการเรียนรู้ของเครื่องทั้งสามตัวแบบ จากนั้นจะตัดคุณลักษณะที่ไม่สำคัญออกและนำข้อมูลที่ได้ไปสอนและทดสอบตัวแบบอีกครั้งและสังเกตการเปลี่ยนแปลงของค่าความถูกต้องเพื่อยืนยันความสำคัญของคุณลักษณะ

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาคุณลักษณะที่สำคัญของข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

1.2.2 เพื่อศึกษา ออกแบบ และพัฒนาตัวแบบการใช้เทคนิคการเรียนรู้ของเครื่องเชิงลึกเพื่อตรวจจับข่าวปลอมภาษาไทย

1.3 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1.3.1 ได้คุณลักษณะที่สำคัญของข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

1.3.2 ได้ตัวแบบการใช้เทคนิคการเรียนรู้ของเครื่องเชิงลึกเพื่อตรวจจับข่าวปลอมภาษาไทย

1.4 โครงสร้างการวิจัย

งานวิจัยฉบับนี้จะทำการเก็บรวบรวมข้อมูลข่าว ทำการตรวจสอบข้อมูลข่าว เพิ่มคุณลักษณะทางด้านความรู้สึกให้กับข่าว สร้างตัวแบบและใช้ข้อมูลในการสอนตัวแบบ วัดค่าความถูกต้อง จากนั้นทำการเปรียบเทียบความถูกต้อง และเปรียบเทียบคุณลักษณะที่สำคัญ

โดยการดำเนินการวิจัยถูกแบ่งออกเป็น 10 ส่วน คือการเชื่อมต่อระบบกับสื่อสังคมออนไลน์ทวิตเตอร์เพื่อค้นคืนข้อมูล การแปลงข้อมูลและบันทึกลงสู่ฐานข้อมูล การตรวจสอบข้อมูลโดยผู้เชี่ยวชาญ การเพิ่มคุณลักษณะด้านคะแนนความรู้สึก การเตรียมข้อมูลและทำมาตรฐานข้อมูล การเลือกคุณลักษณะที่สำคัญของข้อมูล การสร้างตัวแบบและคำนวณค่าความถูกต้อง การเปรียบเทียบ

ความถูกต้องของตัวแบบ การเปรียบเทียบคุณลักษณะที่สำคัญ การทดสอบคุณลักษณะที่สำคัญ โดยมีรายละเอียดดังต่อไปนี้

1.4.1 การเชื่อมต่อระบบกับสื่อสังคมออนไลน์ทวิตเตอร์เพื่อค้นคืนข้อมูล

ทำการสร้างโปรแกรมเชื่อมโยงกับสื่อสังคมออนไลน์ทวิตเตอร์ผ่านทาง Application Programming Interface (API) ผู้วิจัยได้สร้างโปรแกรมติดต่อกับสื่อสังคมออนไลน์ทวิตเตอร์ ผ่านทาง Application Programming Interface (API) เพื่อใช้ในการค้นคืนข้อมูลข่าว

1.4.2 การแปลงข้อมูลและบันทึกลงสู่ฐานข้อมูล

หลังจากที่ได้ทำการค้นคืนข้อมูลจากสื่อสังคมออนไลน์ทวิตเตอร์แล้ว ระบบจะทำการคำนวณคุณลักษณะส่วนหนึ่ง เช่น ชื่อผู้ใช้มีตัวเลขหรือไม่ อัตราส่วนระหว่างเพื่อนและผู้ติดตามของผู้ใช้ ก่อนบันทึกลงสู่ฐานข้อมูล

1.4.3 การตรวจสอบข้อมูลโดยผู้เชี่ยวชาญ

ในงานวิจัยฉบับนี้ ดำเนินการเก็บรวบรวมข่าวสารจากสื่อสังคมออนไลน์ทวิตเตอร์ ซึ่งจะต้องมีการตรวจสอบข้อมูลของข่าวก่อนนำไปใช้สอนและทดสอบตัวแบบ งานวิจัยฉบับนี้ได้ใช้วิธีการตรวจสอบโดยผู้เชี่ยวชาญซึ่งมีคุณสมบัติเป็นผู้ที่มีอาชีพเป็นนักข่าวในการตรวจสอบข้อมูลโดยระบุข่าวที่เก็บรวบรวมได้ว่าเป็นข่าวปลอมหรือไม่ เพื่อนำข้อมูลที่ได้ไปใช้ในการสอนและทดสอบตัวแบบการเรียนรู้ของเครื่อง

1.4.4 การเพิ่มคุณลักษณะด้านคะแนนความรู้สึก

ขั้นตอนนี้ใช้เทคนิคการประมวลผลภาษาธรรมชาติ [7, 11] เพื่อประเมินค่าความรู้สึกของเนื้อหา (Sentiment Analysis) ซึ่งจะระบุออกมาได้เป็น 3 ลักษณะ คือเชิงบวก เป็นกลาง หรือเชิงลบ และนำข้อมูลที่ได้เพิ่มให้เป็นคุณลักษณะอีกอย่างหนึ่งของข่าว

1.4.5 การเตรียมข้อมูลและทำมาตรฐานข้อมูล

ก่อนที่จะนำข้อมูลที่เก็บรวบรวมไปใช้ในการสอนและทดสอบตัวแบบนั้น จำเป็นจะต้องทำข้อมูลให้อยู่ในช่วงมาตรฐานเดียวกัน (Standardization) ผู้วิจัยจึงใช้เทคนิค Standard Scalar [12,13] ในการทำให้ข้อมูลอยู่ในช่วงมาตรฐานเดียวกันก่อนจะนำไปใช้ต่อไป

1.4.6 การเลือกคุณลักษณะที่สำคัญของข้อมูล

การเลือกคุณลักษณะที่สำคัญของข้อมูลข่าวในงานวิจัยฉบับนี้ได้ใช้วิธีการใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier ในการเลือกคุณลักษณะที่สำคัญจากข้อมูลทั้งหมด

1.4.7 การสร้างตัวแบบและคำนวนค่าความถูกต้อง

เป็นการนำข้อมูลที่รวบรวมได้ทั้งหมดแบ่งออกเป็น 3 ส่วน โดยจะใช้ข้อมูล 2 ใน 3 สำหรับการสอนตัวแบบ และใช้ข้อมูลอีกส่วนในการทดสอบความถูกต้องของตัวแบบ ซึ่งจะได้ผลลัพธ์เป็นการระบุข่าวว่าเป็นข่าวจริงหรือข่าวปลอม และนำไปเปรียบเทียบกับผลลัพธ์จริงเพื่อคำนวณความถูกต้อง

โดยผู้วิจัยได้ทำการสร้างตัวแบบการเรียนรู้จำนวน 3 ตัวแบบคือ ต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม เพื่อนำข้อมูลข่าวจริงและข่าวปลอมมาใช้ในการสอนและทดสอบตัวแบบ และคำนวณค่าความถูกต้องเพื่อใช้ในการเปรียบเทียบตัวแบบต่อไป

ซึ่งในขั้นตอนนี้จะสามารถระบุคุณลักษณะที่สำคัญที่ใช้ในการตรวจระบุข่าวปลอมได้จากตัวแบบการเรียนของเครื่องในทุกตัวแบบ

1.4.8 การเปรียบเทียบความถูกต้องของตัวแบบ

เป็นการนำค่าความถูกต้องของทั้ง 3 ตัวแบบมาทำการเปรียบเทียบเพื่อค้นหาตัวแบบที่เหมาะสมในการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

1.4.9 การเปรียบเทียบคุณลักษณะที่สำคัญ

เมื่อได้คุณลักษณะที่สำคัญจากตัวแบบที่สร้างขึ้นในหัวข้อ 1.4.6 และ 1.4.7 จากนั้นนำคุณลักษณะที่ได้มาเปรียบเทียบกัน โดยเลือกคุณลักษณะที่ซ้อนทับกันจากทั้งสองขั้นตอน และนำข้อมูลที่มีเฉพาะคุณลักษณะที่ซ้อนทับกันนั้นนำไปทดสอบกับตัวแบบอีกครั้ง

1.4.10 การทดสอบคุณลักษณะที่สำคัญ

หลังจากได้คุณลักษณะที่ทับซ้อนกันจากหัวข้อ 1.4.9 และได้ค่าความถูกต้องของตัวแล้ว จะนำข้อมูลข่าวทั้งหมด เพื่อใช้ในการตัดคุณลักษณะที่ไม่ซ้อนทับกันจากนั้นนำข้อมูลที่ได้ไปทดสอบกับตัวแบบที่ได้ค่าความถูกต้องมากที่สุดและสังเกตการเปลี่ยนแปลงของค่าความถูกต้องเพื่อยืนยันความสำคัญของคุณลักษณะที่ซ้อนทับ

1.5 ขอบเขตของการวิจัย

ในส่วนของขอบเขตเขตของงานวิจัยจะแบ่งออกเป็น 6 ด้าน ประกอบด้วย ด้านข้อมูลและกลุ่มตัวอย่าง ด้านการตรวจระบุข้อมูล ด้านการเพิ่มคุณลักษณะคะแนนความรู้สึก ด้านตัวแบบสำหรับระบุข่าวปลอมภาษาไทย ด้านการประเมินตัวแบบ ด้านการค้นหาคุณลักษณะที่สำคัญของการแยกแยะข่าวปลอม

1.5.1 ด้านข้อมูลและกลุ่มตัวอย่าง

ข้อมูลข่าวที่ใช้สอนและทดสอบตัวแบบจะใช้ข้อมูลข่าวสารภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ ซึ่งจะใช้ข้อมูลอย่างน้อย 369 ข่าว และจะต้องมีอัตราส่วนของข่าวปลอมอย่างน้อยร้อยละ 40 ของข้อมูลข่าวทั้งหมด คำนวณได้จากสมการการคำนวณประชากรที่มีจำนวนไม่แน่นอน [14] ดังสมการที่ 1-1

$$n = \frac{P(1-P)z^2}{e^2} \quad (1\text{-}1)$$

เมื่อ n = จำนวนกลุ่มตัวอย่าง
P = สัดส่วนของประชากรที่ผู้วิจัยกำหนดสุ่ม
Z = ระดับความมั่นใจ ที่ผู้วิจัยกำหนด
e = สัดส่วนของความคลาดเคลื่อนที่ยอมให้เกิดขึ้นได้

ซึ่งในงานวิจัยฉบับนี้ต้องการอัตราส่วนของข่าวปลอมอย่างน้อยร้อยละ 40 ของข้อมูลข่าวทั้งหมด ความถูกต้องอย่างน้อย 95 เปอร์เซ็นต์ และมีความคลาดเคลื่อนที่ยอมรับได้ที่ 5 เปอร์เซ็นต์ ซึ่งเมื่อแทนค่าสูตรที่ 1-1 จะได้ผลลัพธ์ของจำนวนประชากรที่ 369 ข่าว

1.5.2 ด้านการตรวจระบุข้อมูล

ในการตรวจระบุข้อมูลข่าวเพื่อใช้ในการสอนและทดสอบตัวแบบจะใช้หลักการในการกำหนดความคลาดเคลื่อนของการตรวจระบุข้อมูลที่ยอมรับได้อยู่ที่ 5% โดยจะใช้วิธีการให้ผู้ที่มีความเชี่ยวชาญหรือมีอาชีพเกี่ยวกับข่าวสารตรวจระบุข้อมูลโดยจะระบุผลลัพธ์ระหว่างข่าวจริงหรือข่าวปลอม

1.5.3 ด้านการเพิ่มคุณลักษณะคะแนนความรู้สึก

งานวิจัยฉบับนี้จะใช้เทคนิคการประมวลผลภาษาธรรมชาติเพื่อทำการระบุคะแนนความรู้สึกของเนื้อหาข่าว โดยผลลัพธ์ที่ได้มี 3 แบบคือ เชิงบวก เชิงลบ และเป็นกลาง ซึ่งจะนำข้อมูลที่ได้ไปเพิ่มคุณลักษณะให้กับข้อมูลเพื่อใช้สอนและทดสอบตัวแบบการเรียนรู้ของเครื่องต่อไป

1.5.4 ด้านตัวแบบสำหรับตรวจระบุข่าวปลอมภาษาไทย

ผู้วิจัยได้เสนอวิธีการในการนำเทคนิคที่สนใจมาทำการสร้างตัวแบบเพื่อใช้ในการแยกแยะข่าวปลอมซึ่งประกอบไปด้วยเทคนิคดังนี้

- เทคนิคการเรียนรู้ของเครื่องประเภทต้นไม้ตัดสินใจ
- เทคนิคการเรียนรู้ของเครื่องประเภทซัพพอร์ทเวกเตอร์แมชชีน
- เทคนิคการเรียนของเครื่องประเภทโครงข่ายประสาทเทียม

1.5.5 ด้านการประเมินตัวแบบ

จากหัวข้องานวิจัยที่เกี่ยวข้องผู้วิจัยเสนอวิธีการในการประเมินตัวแบบเพื่อทำการแยกแยะข่าวปลอมซึ่งประกอบไปด้วยเทคนิคการประเมินดังนี้

- ค่าความถูกต้อง (Accuracy)
- ค่าความแม่นยำ (Precision)
- ค่าความอ่อนไหว (Recall)
- ค่าประสิทธิภาพของตัวแบบ (F-Measure)

โดยจะใช้ค่าความถูกต้องเป็นหลักในการเปรียบเทียบความสามารถในการตรวจระบุข่าวปลอมของตัวแบบการเรียนรู้ของเครื่อง

1.5.6 ด้านการค้นหาคุณลักษณะที่สำคัญของการแยกแยะข่าวปลอม

งานวิจัยฉบับนี้ใช้ 2 ส่วนในการค้นหาคุณลักษณะที่สำคัญของการแยกแยะข่าวปลอม คือ ส่วนที่ได้จากการใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier และส่วนที่ได้จากตัวแบบการเรียนรู้ของเครื่อง และนำทั้ง 2 ส่วนมาทำการเปรียบเทียบกันโดยเลือกเฉพาะคุณลักษณะที่ซ้ำกันจากทั้ง 2 ส่วนและนำเฉพาะคุณลักษณะดังกล่าวของข้อมูลไปทดสอบกับตัวแบบอีกครั้ง

1.6 สรุป

ในงานวิจัยฉบับนี้ได้นำเสนอการค้นหาคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ด้วยเทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier และนำมาเปรียบเทียบกับคุณลักษณะที่สำคัญที่ได้จากตัวแบบการเรียนรู้ทั้งสามตัวแบบ จากนั้นนำเฉพาะคุณลักษณะที่สำคัญไปใช้ทดสอบเพื่อดูความเปลี่ยนแปลงของค่าความถูกต้องของตัวแบบที่ได้ค่าความถูกต้องมากที่สุด และสรุปผล

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ทฤษฎีและหลักการ

คุณลักษณะของข่าวปลอม และเทคนิคต่าง ๆ ที่ใช้ในการวิจัยครั้งนี้ ประกอบไปด้วย คำนิยามของข่าวปลอม ข่าวปลอมบนสื่อสังคมออนไลน์ คุณลักษณะของข่าวปลอมบนสื่อสังคมออนไลน์ การประมวลผลภาษาธรรมชาติ การเรียนรู้ของเครื่อง เทคนิคต้นไม้ตัดสินใจ เทคนิคซัพพอร์ทเวกเตอร์แมชชีน เทคนิคโครงข่ายประสาทเทียม และการวัดค่าความถูกต้องของตัวแบบ รวมไปถึงการนิยามและสำรวจข้อมูลการตรวจระบุข่าวปลอม โดยมีรายละเอียดดังต่อไปนี้

2.1.1 คำนิยามของข่าวปลอม

ข่าวปลอมได้เกิดขึ้นมากเป็นเวลานาน โดยก่อนหน้าที่โลกจะมีอินเตอร์เน็ต หรือก่อนหน้าที่มีการใช้อินเตอร์เน็ตกันอย่างแพร่หลาย ข่าวปลอมจะอยู่ในรูปแบบของเอกสาร นิตยสาร หรือสื่อต่าง ๆ อย่างไรก็ตามไม่มีคำนิยามที่เป็นทางการสำหรับข่าวปลอม แต่ได้มีงานวิจัย [14] ได้นิยามความหมายของข่าวปลอมได้ว่า เป็นข่าวที่มีเนื้อหาที่ผิดไปจากความเป็นจริง และมีแนวโน้มว่าจะสามารถชักจูงผู้อ่านให้เข้าใจผิดได้ ซึ่งหัวใจหลักของการตั้งคำนิยามนี้คือ ความจริง และความตั้งใจ ในเรื่องของความจริงคือเนื้อหาที่นำเสนอไม่เป็นความจริง และในส่วนของความตั้งใจคือเนื้อหาตั้งใจให้ผู้อ่านได้รับข่าวสารที่จูงใจไปทิศทางใดทิศทางหนึ่ง [15, 16]

2.1.2 ข่าวปลอมบนสื่อสังคมออนไลน์ทวิตเตอร์

ข่าวปลอมในสื่อสังคมออนไลน์ทวิตเตอร์มีคุณลักษณะเฉพาะที่มากกว่าข่าวปลอมแบบดั้งเดิมหรือในรูปแบบเอกสาร นิตยสาร ฯลฯ นั่นก็คือ เรื่องของตัวตนปลอม และการกระจายข่าวสารเฉพาะกลุ่ม

ในซึ่งสื่อสังคมออนไลน์ทวิตเตอร์ ผู้ใช้สามารถสร้างบัญชีปลอมเพื่อทำการกระจายข่าวสาร ซึ่งทำได้ง่ายและขาดการตรวจสอบอย่างจริงจัง โดยมีการสร้างหุ่นยนต์สังคม (Social Bot) ในรูปแบบของบัญชีปลอมอีกด้วย โดยจากงานวิจัย [17] พบว่าหลัก ๆ แล้วหุ่นยนต์สังคมจะใช้อัลกอริทึมอัตโนมัติในการสร้างข่าวปลอม และผู้ใช้งานที่เป็นมนุษย์ซึ่งเป็นเจ้าของบัญชีปลอมจะทำหน้าที่ในการกระจายข้อมูลข่าวสารนั้น ๆ ต่อไป ซึ่งเป็นรูปแบบใหม่ในการกระจายข่าวและการบริโภคข่าวในยุคสมัยปัจจุบัน เนื่องจากสื่อสังคมออนไลน์ที่มีความฉลาดในการค้นคืนข่าวสารที่ผู้ใช้สนใจ

ปรากฏบริเวณหน้าหลักในการใช้งานของผู้ใช้ ทำให้ผู้ใช้คิดว่าข่าวนั้น ๆ เป็นเรื่องจริง ทำให้ความรุนแรงของข่าวปลอมทวีขึ้นอย่างรวดเร็ว

2.1.3 คุณลักษณะของข่าวปลอมบนสื่อสังคมออนไลน์

ในสภาพแวดล้อมของการแยกแยะข่าวปลอม ได้มีงานวิจัย [14] ศึกษาเรื่องการนิยามคุณลักษณะออกมาเป็น 3 ส่วนหลัก ๆ ดังต่อไปนี้

ผู้กระจายข่าว (User-Based) สนใจสภาพแวดล้อมของผู้ใช้ที่กระจายข่าว เช่น จำนวนเพื่อนของผู้กระจายข่าว จำนวนผู้ติดตาม ข้อมูลส่วนตัวที่เปิดเผย จำนวนการเผยแพร่ข้อความ ระยะเวลาที่สร้างบัญชีจนถึงปัจจุบัน เป็นต้น

เนื้อหา (Content-Based) สนใจในสภาพแวดล้อมของเนื้อหาข่าว เช่น จำนวนแฮชแท็ก (Hash Tag) จำนวนลิงค์ที่อยู่ในเผยแพร่ จำนวนคำในเผยแพร่ การมีรูปภาพ เป็นต้น

การมีส่วนร่วม (Social Based) สนใจในท่าทีของผู้ใช้อื่น ๆ ที่ได้รับข่าวสาร เช่น การแสดงความคิดเห็น การกดปุ่มถูกใจ การแบ่งปันต่อ เป็นต้น

2.1.4 การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ [7, 11] เป็นเทคนิคการคำนวณเกี่ยวกับภาษาเพื่อให้คอมพิวเตอร์ได้เข้าใจภาษาของมนุษย์และสามารถตีความได้ในลักษณะของคะแนนความรู้สึก ความหมายของประโยค เป็นต้น

2.1.5 การเรียนรู้ของเครื่อง

คำนิยามจากหนังสือและงานวิจัยของการเรียนรู้ของเครื่อง [9, 10] เป็นการทำให้เครื่องเรียนรู้ได้จากข้อมูลตัวอย่างที่เราสนใจ โดยมุ่งเน้นไปในการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น และเมื่อทำการเรียนรู้แล้วเครื่องจะนำความรู้ที่เรียนเก็บไว้ในฐานความรู้อยู่ในรูปแบบหลาย ๆ แบบ เช่น การสร้างกฎ

2.1.6 เทคนิคต้นไม้ตัดสินใจ

เทคนิคต้นไม้ตัดสินใจ [8, 9] เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่อง โดยใช้โครงสร้างของต้นไม้เป็นต้นแบบในการนำเสนอตัวเลขของเส้นทางความน่าจะเป็นของผลลัพธ์ในแต่ละเส้นทาง

ซึ่งในการสร้างต้นไม้ตัดสินใจนั้น เราจำเป็นต้องทราบคำถามหรือปัญหาที่เราต้องการค้นหาคำตอบเป็นลำดับขั้น แต่ละขั้นของต้นไม้จะทำการตัดเส้นทางออกไปจากความน่าจะเป็นที่มีค่าต่ำ และเลือกเส้นทางที่มีความน่าจะเป็นสูงไปจนถึงคำตอบที่ต้องการค้นหา

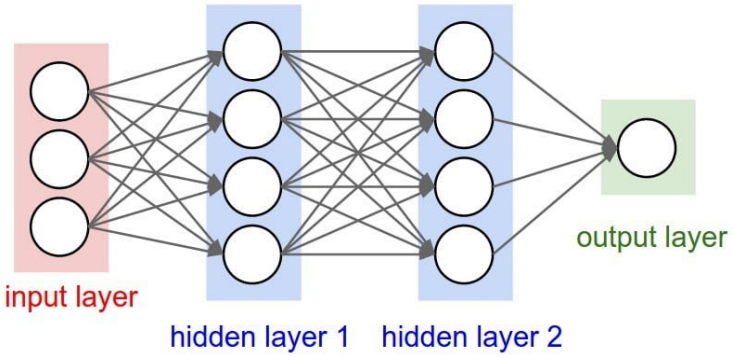
2.1.7 เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

เทคนิคซัพพอร์ทเวกเตอร์แมชชีน [8, 9] เป็นเทคนิคหนึ่งของการเรียนรู้ของเครื่องโดยใช้ระนาบของการตัดสินใจ แบ่งข้อมูลออกเป็นสองส่วนโดยใช้สมการเส้นตรงเพื่อแบ่งฝั่งของข้อมูลออกจากกัน ซึ่งจะสามารถทำการคัดแยก (Classification) และทำนาย (Regression) โดยเทคนิคนี้จะพยายามทำให้ผลความคลาดเคลื่อนของข้อมูลต่ำขณะที่พยายามทำให้ความต่างระหว่างชุดของข้อมูลมากที่สุด เทคนิคนี้มักจะถูกใช้ในลักษณะของข้อมูลที่มีคุณลักษณะค่อนข้างสูงเมื่อเทียบกับตัวอย่างข้อมูล หรือข้อมูลไม่สามารถแบ่งได้ด้วยเส้นตรง

กระบวนการทำงานของซัพพอร์ทเวกเตอร์แมชชีน คือการหาค่าระยะขอบที่มากที่สุดของระนาบตัดสินใจ โดยการแบ่งขอ้มูลออกจากกัน และมีเส้นขอบคั่นระหว่างฝั่ง ซึ่งข้อมูลใหม่ที่เข้ามาจะใช้วิธีการหาระยะทางที่ใกล้เส้นของคั่นน้อยที่สุดจึงถือว่าข้อมูลใหม่เป็นข้อมูลในกลุ่มที่ใกล้ที่สุดนั้น

2.1.8 เทคนิคโครงข่ายประสาทเทียม

โครงข่ายประสาทเทียม [8, 9, 10] เป็นตัวแบบพยากรณ์ที่ได้รับแรงบันดาลใจจากการทำงานของสมองมนุษย์ โดยประสาทเทียมนั้นจะรับเอาข้อมูลที่เป็นผลลัพธ์จากประสาทเทียมตัวอื่นควบรวมกับค่าน้ำหนักเพื่อเอาใช้คำนวณ หากการคำนวณได้ค่ามากกว่าที่ตั้งเอาไว้จะทำการส่งต่อข้อมูลนี้ไปยังประสาทเทียมตัวอื่น เพื่อใช้ในการคำนวณถัดไป จนออกมาเป็นผลลัพธ์ที่เราสนใจโดยแสดงเป็นตัวอย่างดังภาพที่ 2-1 ที่เป็นการรับข้อมูลขาเข้า (Input Layer) จำนวน 3 ชุด เพื่อนำไปคำนวณผ่านชั้นที่ซ่อนอยู่ (Hidden Layer) 2 ชั้นและได้ผลลัพธ์ออกมาเป็นข้อมูลที่ต้องการ (Output Layer)



ภาพที่ 2-1 ตัวอย่างโครงข่ายประสาทเทียม

อีกทั้งโครงข่ายประสาทเทียม ในปัจจุบันมีการทำลายสถิติเรื่องของความถูกต้องและความเร็วในการจดจำภาพและเสียง อ่านลายมือ เข้าใจความหมายของคำ การแบ่งรูป ขับรถอัตโนมัติ

และอื่น ๆ อีกมากมาย โดยสามารถรองรับทั้งในเรื่องของคุณลักษณะของข้อมูลหรือเรื่องของข้อมูลที่มีจำนวนมาก และต้องการความสามารถในการคำนวณที่มีความซับซ้อนสูง [9]

2.1.9 การวัดค่าความถูกต้องของตัวแบบ

การวัดค่าความถูกต้องของตัวแบบ สามารถคำนวณได้จากการสร้างเมทริกซ์ความสับสนหรือคอนฟิวชันเมทริกซ์ (Confusion Matrix) ของความถูกต้องของการทำนาย (Confusion Matrix) โดยอ้างอิงจากปัญหา โดยเราสามารถนิยามออกมาได้เป็นข้อมูลดังนี้

- True Positive (TP) : เมื่อพยากรณ์ว่าเป็นข่าวปลอมและผลคือเป็นข่าวปลอม
- True Negative (TN) : เมื่อพยากรณ์ว่าเป็นข่าวจริงและผลคือเป็นข่าวจริง
- False Positive (FP) : เมื่อพยากรณ์ว่าเป็นข่าวปลอมและผลที่ได้คือเป็นข่าวจริง โดยผลของการทำนาย ทำให้มีโอกาสเกิดความเสียหายได้สูงเชิงคุณภาพ
- False Negative (FN) : เมื่อพยากรณ์ว่าเป็นข่าวจริงและผลที่ได้คือเป็นข่าวปลอม โดยผลของการทำนาย ทำให้มีโอกาสเกิดความเสียหายได้สูงเชิงปริมาณ

ซึ่งสามารถนำเอาข้อมูลดังกล่าวไปใช้ในการคำนวณความถูกต้อง ของตัวแบบพยากรณ์ได้ด้วยสมการของการวัดค่าความถูกต้อง การวัดค่าความแม่นยำ การวัดค่าความอ่อนไหว และการวัดค่าประสิทธิภาพของตัวแบบ โดยการสร้างคอนฟิวชันเมทริกซ์ [18] ดังตารางที่ 2-1

ตารางที่ 2-1 การสร้างเมทริกซ์ความสับสน

		ผลที่เกิดขึ้นจริง	
		เชิงบวก	เชิงลบ
การทำนาย	เชิงบวก	TP	FP
	เชิงลบ	FN	TN

จากนั้นนำผลที่ได้จากการคำนวณคอนฟิวชันเมทริกซ์ดังตารางที่ 2-1 มาใช้ในการคำนวณ ค่าความถูกต้อง ความแม่นยำ ความอ่อนไหว และค่าประสิทธิภาพของตัวแบบโดยมีรายละเอียดดังต่อไปนี้

2.1.9.1 การวัดค่าความถูกต้อง

Accuracy [18] คือการคำนวณจำนวนคำตอบที่พยากรณ์ถูกต้องเทียบกับจำนวนคำตอบทั้งหมด ที่นำไปให้ตัวแบบทำการพยากรณ์ ซึ่งนำมาประยุกต์ใช้กับงานวิจัยฉบับนี้โดยการวัดความถูกต้องของตัวแบบพยากรณ์ ซึ่งสามารถคำนวณโดยการนำผลของคอนฟิวชันเมทริกซ์มาใช้ในสมการที่ (1)

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (2\text{-}1)$$

2.1.9.2 การวัดค่าความแม่นยำ

Precision [18] คือการคำนวณความแม่นยำหรือจำนวนที่ทำนายถูกจากข้อมูลที่ทำนายเป็นคลาสที่สนใจอยู่ โดยสามารถนำมาประยุกต์ใช้กับงานวิจัยฉบับนี้ได้โดยการวัดความแม่นยำของตัวแบบพยากรณ์ที่ทำนายผลออกมาเป็นข่าวปลอม ซึ่งสามารถคำนวณโดยการนำผลของคอนฟิวชันเมทริกซ์มาใช้ในสมการที่ (2)

$$Precision = \frac{TP + TN}{TP + FN + TN + FP} \quad (2\text{-}2)$$

2.1.9.3 การวัดค่าความอ่อนไหว

Recall [18] คือค่าความอ่อนไหว หรือจำนวนข้อมูลที่ทำนายถูกต่อข้อมูลที่ทำนายถูกรวมกับข้อมูลที่ทำนายผิดซึ่งเป็นประเภทที่สนใจ โดยการวัดค่าความอ่อนไหวของตัวแบบพยากรณ์ที่ทำนายผลออกมาเป็นข่าวปลอม ซึ่งสามารถคำนวณโดยการนำผลของคอนฟิวชันเมทริกซ์มาใช้ในสมการที่ (3)

$$Recall = \frac{TP + TN}{TP + FN + TN + FP} \quad (2\text{-}3)$$

2.1.9.4 การวัดค่าประสิทธิภาพของตัวแบบ

F-Measure [18] หรือการวัดประสิทธิภาพของตัวแบบ คำนวณได้โดยการใช้ค่าน้ำหนักเฉลี่ยของค่าความแม่นยำและค่าความอ่อนไหวของแต่ละประเภทเป้าหมาย โดยในงานวิจัยฉบับนี้ได้นำมาประยุกต์ใช้ในการวัดค่าประสิทธิภาพของตัวแบบในการระบุหรือแยกแยะข่าวปลอมซึ่งสามารถคำนวณโดยการนำผลของคอนฟิวชันเมทริกซ์มาใช้ในสมการที่ (4)

$$F\ Measure\ = \frac{TP + TN}{TP + FN + TN + FP} \quad (2\text{-}4)$$

2.1.10 การเลือกคุณลักษณะที่สำคัญ

การเลือกคุณลักษณะที่สำคัญสามารถทำได้จากการใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งมีหลากหลายเทคนิค ในงานวิจัยฉบับนี้ได้เลือกใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier เป็นหนึ่งในกลุ่มของตัวแบบที่เรียกว่า Ensemble learning ที่มีหลักการ คือ การสอนตัวแบบที่เหมือนกันตัวแบบจำนวนมากด้วยข้อมูลชุดเดียวกัน โดยทุกรอบของการสอนตัวแบบจะเลือกส่วนของข้อมูลที่ใช้สอนต่างกัน จากนั้นนำการตัดสินใจของตัวแบบเหล่านั้นมาใช้ในการประเมิน

ซึ่งในเทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier จะประกอบด้วยตัวแบบต้นไม้ตัดสินใจที่เหมือนกันจำนวนหนึ่งรวมกันเรียกว่า ป่า (Forest) โดยต้นไม้ตัดสินใจแต่ละตัวแบบที่เหมือนกันจะได้รับจำนวนคุณลักษณะที่ไม่เหมือนกันเพื่อไปใช้ในการคำนวณ

โดยเทคนิค Extra Trees Classifier คำนวณสิ่งที่เรียกว่า Feature importances [18] หรือการชี้วัด "ความสำคัญของคุณลักษณะ" ในลักษณะหาความสัมพันธ์ โดยวัดว่าต้นไม้แต่ละต้นที่ใช้ คุณลักษณะนั้นสามารถลดค่าความไม่บริสุทธิ์ (Gini Index) ได้จำนวนใด และนำมาใช้ในการประเมินคุณลักษณะที่สำคัญ

2.1.11 การได้มาซึ่งคุณลักษณะที่สำคัญจากตัวแบบ

ในการได้มาซึ่งคุณลักษณะที่สำคัญจากตัวแบบทั้งสามตัวแบบของงานวิจัยฉบับนี้มีรายละเอียดดังนี้

2.1.11.1 คุณลักษณะที่สำคัญจากตัวแบบต้นไม้ตัดสินใจ

ในการสร้างและใช้งานตัวแบบต้นไม้ตัดสินใจ จะได้มาซึ่งคุณลักษณะที่มีความสำคัญต่อการสอนและ ทดสอบตัวแบบ สามารถทำได้โดยใช้วิธีการเรียกดูค่าคุณลักษณะของตัวแบบ ที่มีชื่อว่า `feature_importances_` [19] ซึ่งผลลัพธ์ที่ได้จะเป็นคุณลักษณะที่สำคัญของการสอนและทดสอบตัวแบบ

2.1.11.2 คุณลักษณะที่สำคัญจากตัวแบบซัพพอร์ทเวกเตอร์แมชชีน

ในการสร้างตัวแบบซัพพอร์ทเวกเตอร์แมชชีนจะได้มาซึ่งคุณลักษณะที่มีความสำคัญต่อการสอนและ ทดสอบตัวแบบ สามารถทำได้โดยใช้วิธีการเรียกดูค่าคุณลักษณะของตัว

แบบ ที่มีชื่อว่า `coef_` [20] ซึ่งผลลัพธ์ที่ได้จะเป็นคุณลักษณะที่สำคัญของการสอนและทดสอบตัวแบบ

2.1.11.3 คุณลักษณะที่สำคัญจากตัวแบบโครงข่ายประสาทเทียม

ในการสร้างตัวแบบซัพพอร์ทเวกเตอร์แมชชีนจะได้มาซึ่งคุณลักษณะที่มีความสำคัญต่อการสอนและ ทดสอบตัวแบบ สามารถทำได้โดยใช้คลาสตัวช่วยที่มีชื่อว่า `PermutationImportance` และ `Eli5` [21] ซึ่งผลลัพธ์ที่ได้จะเป็นคุณลักษณะที่สำคัญของการสอนและทดสอบตัวแบบ

2.2 งานวิจัยที่เกี่ยวข้อง

ในการศึกษาวิจัยเรื่องการแยกแยะข่าวปลอมผ่านทางสื่อสังคมออนไลน์ ผู้วิจัยได้ทำการศึกษาค้นคว้าเอกสารงานวิจัยที่เกี่ยวข้องเพื่อนำมาประกอบความรู้ ศึกษาแนวทาง และอ้างอิงในการทำงานวิจัย ซึ่งสรุปสาระสำคัญได้ดังต่อไปนี้

2.2.1 เกี่ยวกับการนิยามและสำรวจข้อมูลการตรวจระบุข่าวปลอม

Kai Shu, Arny Silva, Suhang Wang, Jiliang Tang, and Huan Liu. [14] ได้ทำการสำรวจแนวคิด วิธีการเกี่ยวกับการแยกแยะข่าวปลอมผ่านทางสื่อสังคมออนไลน์ โดยมีรายละเอียดที่สำคัญดังนี้

- ความหมายของข่าวปลอม
- ลักษณะของข่าวปลอมแบบดั้งเดิมและข่าวปลอมบนสื่อสังคมออนไลน์
- การแยกแยะข่าวปลอม แบ่งมุมมองออกเป็น 3 มุมมองคือ
    - คุณลักษณะของข่าวปลอม (Feature Extraction) ซึ่งประกอบไปด้วย แหล่งที่มา หัวข้อข่าว เนื้อหาข่าว รูปภาพ วิดีโอ
        - แยกคุณลักษณะจากภาษา เช่น เนื้อหาข่าว ความรู้สึกของประโยค
        - แยกคุณลักษณะจากสิ่งที่ปรากฏ เช่น จากภาพ เสียง
    - สภาพแวดล้อมของข่าวปลอม แบ่งออกเป็น 3 รูปแบบ
        - ข้อมูลที่เกี่ยวกับผู้เผยแพร่ข่าว เช่น ชื่อ จำนวนผู้ติดตาม
        - ข้อมูลที่เกี่ยวกับข่าว เช่น เนื้อหาข่าว หัวข้อ
        - ข้อมูลที่เกี่ยวกับการมีส่วนร่วม เช่น การแสดงความคิดเห็น
    - การสร้างตัวแบบการเรียนรู้ เพื่อแยกแยะข่าวปลอม
        - อ้างอิงจากตัวข่าวและส่วนที่เกี่ยวข้องในข่าว เช่น ชื่อผู้แต่ง เนื้อหา

- อ้างอิงจากสภาพแวดล้อมของข่าว เช่น การแสดงความคิดเห็น
- การประเมินผลของการแยกแยะข่าวปลอม ซึ่งนำเสนอวิธีการออกมาเป็น 4 วิธีคือ
  - ความถูกต้อง หรือ Accuracy
  - ความแม่นยำ หรือ Precision
  - ความอ่อนไหว หรือ Recall
  - ประสิทธิภาพของตัวแบบ หรือ F-Measure

ซึ่งผู้วิจัยได้นำข้อมูลในส่วนนี้มาใช้อ้างอิง สกัดข้อมูลเป็นความรู้ และใช้เป็นแนวทางในการสร้างตัวแบบเพื่อแยกแยะข่าวปลอมบนสื่อสังคมออนไลน์ อีกทั้งยังใช้ในการอ้างอิงในขั้นตอนออกแบบระเบียบการวิจัยที่จะกล่าวถึงในหัวข้ออื่นต่อไป

2.2.2 เกี่ยวกับการใช้เทคนิคการเรียนรู้ของเครื่องเพื่อตรวจระบุข่าวปลอม

ผู้วิจัยได้ทำการศึกษาวรรณกรรมที่เกี่ยวข้อง โดยเป็นงานที่ใช้เทคนิคการเรียนรู้ของเครื่อง ทั้งแบบทั่วไป ทั้งแบบเชิงลึกเพื่อใช้สำหรับแยกแยะข่าวปลอมผ่านทางสื่อสังคมออนไลน์ โดยมีรายละเอียดดังตารางที่ 2-1 ซึ่งเป็นการเปรียบเทียบงานวิจัยในด้านการใช้เทคนิคการเรียนรู้ของเครื่องประเภทต่าง ๆ และความถูกต้องที่ได้จากการทดลอง

ตารางที่ 2-2 งานวิจัยและเทคนิคที่ใช้ในงานวิจัยจากการทบทวนวรรณกรรม

ชื่องานวิจัย	NLP	จำนวนที่พบการใช้เทคนิคประเภท				ความถูกต้อง (%)
		Decision Tree	Random Forest	SVM	Neural Network	
Identify tweet with fake news (2018) [2]	ใช้	1	-	1	-	99 *
Fake News Identification on Twitter with Hybrid CNN and RNN Models (2018) [3]	-	-	-	-	2	82
CSI : A Hybrid Deep Model for Fake News Detection (2017) [4]	-	1	-	1	1	95.4

ชื่องานวิจัย	NLP	จำนวนที่พบการใช้เทคนิคประเภท				ความถูกต้อง (%)
		Decision Tree	Random Forest	SVM	Neural Network	
Fake and Spam Messages : Detecting Misinformation during Natural Disasters on Social Media (2015) [22]	-	2	1	-	-	96.43 *
Improving Spam Detection in Online Social Network (2015) [23]	-	1	-	-	-	87.9
Event Adversarial Neural Network for Multi-Modal Fake News Detection (2018) [24]	-	-	-	-	3	82.7
Rumors Detection in Chinese via Crowd Responses (2014) [25]	ใช้	-	-	1	-	95.24
Detect rumors using time series of social context information on microblogging websites (2015) [26]	ใช้	1	1	1	-	89.6
Detecting Hoaxes Frauds and Deception in	ใช้	1	-	1	-	96.6

ชื่องานวิจัย	NLP	จำนวนที่พบการใช้เทคนิคประเภท				ความถูกต้อง (%)
		Decision Tree	Random Forest	SVM	Neural Network	
Writing Style Online (2012) [27]						
Detecting rumors from microblogs with recurrent neural networks (2016) [5]	ใช้	1	1	1	2	91
Fake news or truth using satirical cues to detect potentially misleading news (2016) [28]	ใช้	-	-	1	-	93
News credibility evaluation on microblog with a hierarchical propagation model (2014) [29]	-	-	-	1	-	88.9 *
Prominent features to rumor propagation in online social media (2013) [30]	-	1	1	1	-	93
จำนวนครั้งที่พบในวรรณกรรมที่สืบค้น คิดเป็นเปอร์เซ็นต์การถูกใช้งาน	6 46.1%	9 69.2 %	4 30.7 %	9 69.2 %	8 61.5 %	

หมายเหตุ * วรรณกรรมที่ใช้ข้อมูลเฉพาะเหตุการณ์ในการสอนและทดสอบเครื่อง เช่นเหตุการณ์เครื่องบินเฮอริเคนแซนดี้ เป็นต้น

หลังจากที่ผู้วิจัยได้ทำการศึกษาและทำการทบทวนวรรณกรรมจากข้อมูลในตารางข้างต้นแล้ว พบว่ามีการใช้เทคนิคต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน และโครงข่ายประสาทเทียมเป็นหลัก และมีบางงานที่ใช้เทคนิค Random Forest มาใช้ในการตรวจจับข่าวปลอม และมีการประยุกต์ใช้เทคนิคการประมวลผลธรรมชาติ มาช่วยในการแยกคุณลักษณะบางชนิดออกจากเนื้อข่าวเพื่อเพิ่มความถูกต้อง

ในงานวิจัย [14] ได้พูดถึงประเด็นในหัวข้อของคุณลักษณะและตัวแบบที่หากมีความซับซ้อนมากขึ้นจะทำให้ได้รับความถูกต้องมากยิ่งขึ้น และยังสามารถขยายตัวแบบให้ปรับใช้ได้หลากหลายชุดข้อมูลมากกว่า ซึ่งในงานวิจัยฉบับนี้เราตั้งเป้าหมายในการสร้างตัวแบบที่ซับซ้อนหรือตัวแบบเชิงลึกเพื่อตรวจจับข่าวปลอม โดยผู้วิจัยจึงเลือกเฉพาะงานวิจัยที่มีการใช้ตัวแบบที่เป็นไปในแนวทางของตัวแบบเชิงลึกมาศึกษาอ้างอิงดังต่อไปนี้

Oluwaseun Ajao et al. [3] นำเสนอเฟรมเวิร์คที่ใช้ในการตรวจจับข่าวปลอมจากทวิตเตอร์ โดยการรวมกันระหว่างเทคนิค Convolution Neural Network (CNN) กับ Long-Short Term Recurrent Neural Network (RNN-LTSM) เข้าด้วยกัน ซึ่งได้ผลลัพธ์ความถูกต้องได้มากถึง 82% โดยไม่ต้องใช้ความรู้ก่อนหน้าของคุณลักษณะที่น่าสนใจของข่าวจำพวกนั้น

Natali Ruchansky et al. [4] พัฒนาตัวแบบที่ใช้คุณลักษณะสามด้านนั่นก็คือ ข้อความ การตอบโต้ และแหล่งที่มาของข่าวเพื่อเพิ่มความถูกต้องของการพยากรณ์ โดยตั้งชื่อตัวแบบให้สอดคล้องกับคุณลักษณะที่สนใจคือ Capture Score Integrate ซึ่งแบ่งออกเป็นสามโมดูล โมดูลที่หนึ่งทำงานเกี่ยวกับข้อความและการตอบโต้ของผู้ใช้โดยใช้เทคนิค Recurrent Neural Network และโมดูลที่สองสนใจเกี่ยวกับแหล่งที่มาโดยอ้างอิงจากพฤติกรรมของผู้ใช้ จากนั้นนำผลลัพธ์ทั้งสองโมดูลมารวมกันที่โมดูลที่สามเพื่อแยกแยะข้อมูลว่าเป็นข่าวปลอมหรือไม่ ซึ่งเมื่อนำไปทดสอบกับข้อมูลจริงแล้วได้ค่าความถูกต้องมากกว่าตัวแบบอื่นที่ผู้วิจัยได้นำมาเปรียบเทียบ

Yaqing Wang et al. [24] ได้นำเสนอเฟรมเวิร์คที่ชื่อว่า Event Adversarial Neural Networks (EANN) ที่สามารถสกัดเอาคุณลักษณะคงที่ของข่าวปลอมที่เข้ามาใหม่ในเหตุการณ์นั้น ๆ โดยเฟรมเวิร์คนี้ประกอบไปด้วยสามส่วนคือ ส่วนสกัดคุณลักษณะจากข้อความและรูป (Multi-modal Feature Extractor) ส่วนตรวจจับข่าวปลอม (Fake news detector) และส่วนแยกแยะเหตุการณ์ (Event Discriminator) โดยส่วนตรวจจับข่าวปลอมจะทำงานร่วมกับส่วนสกัดคุณลักษณะ เพื่อใช้ข้อมูลที่ได้นำไปช่วยในการตรวจจับ และในส่วนแยกแยะเหตุการณ์จะทำหน้าที่นำคุณลักษณะเฉพาะของเหตุการณ์นั้น ๆ ออกจากคุณลักษณะทั้งหมด ให้คงเหลือเฉพาะคุณลักษณะที่ทุกเหตุการณ์มีเหมือนกัน โดยผู้แต่งได้นำข้อมูลจากสื่อสังคมออนไลน์ Weibo และ Twitter มาใช้ในการสอนเครื่องและทดสอบ ซึ่งผลที่ได้มีความถูกต้องมากกว่าเทคนิคอื่น ๆ ที่ใช้กันอยู่ในปัจจุบัน

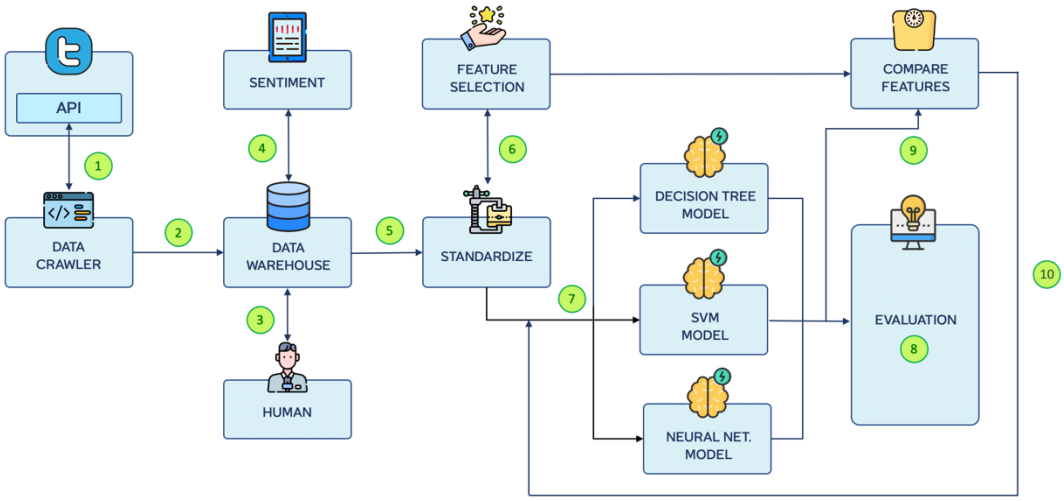
Jing Ma et al. [26] นำเสนอกระบวนการเรียนรู้แบบต่อเนื่องเพื่อระบุข่าวลือ โดยสร้างตัวแบบที่ใช้เทคนิค Recurrent Neural Network สำหรับเรียนรู้ส่วนที่ซ่อนอยู่ของข้อมูลที่ใช้สภาพแวดล้อมของข้อมูลมาช่วย ซึ่งผลจากการทดลองคือสามารถให้ความถูกต้องมากกว่าตัวแบบอื่นที่เอามาเปรียบเทียบ โดยผู้วิจัยได้ให้ข้อมูลไว้ว่าหากตัวแบบนี้สามารถปรับปรุงได้โดยการทำให้ซับซ้อนมากขึ้นโดยการเพิ่มจุดต่อ (Node) หรือเพิ่มชั้นที่ซ่อนอยู่ และการนำตัวแบบนี้ไปใช้ตรวจจับโดยนับเวลาผลที่ได้คือได้ความถูกต้องมากในเวลาที่รวดเร็วเมื่อเทียบกับการใช้เทคนิคอื่น

จากการศึกษางานวิจัยข้างต้น พบว่าช่องทางการเพิ่มความถูกต้องของตัวแบบสามารถทำได้หลากหลายวิธี และวิธีที่สามารถประยุกต์ใช้กับข้อมูลหลากหลายไม่ขึ้นกับเหตุการณ์ใดเหตุการณ์หนึ่งคือตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมแบบวนซ้ำซึ่งรองรับคุณลักษณะได้จำนวนมาก อีกทั้งสามารถปรับแต่งค่าต่าง ๆ เช่น จำนวนจุดเชื่อมต่อ จำนวนชั้นที่ซ่อนอยู่ เพื่อเพิ่มความถูกต้องได้อีกด้วย และผู้วิจัยยังพบว่าการพัฒนาตัวแบบการศึกษางานวิจัยข้างต้นอ้างอิงไปกับภาษาอังกฤษและพฤติกรรมการใช้สื่อสังคมออนไลน์ ของชาวต่างชาติ ซึ่งอาจจะไม่สามารถให้ความถูกต้องมากนักเมื่อนำไปใช้กับภาษาไทยหรือพฤติกรรมการใช้สื่อสังคมออนไลน์ ของคนไทย

ดังนั้นผู้วิจัยจึงนำเสนอแนวคิดในการสร้างตัวแบบ สำหรับตรวจจับข่าวปลอมภาษาไทยโดยใช้เทคนิคประเภทโครงข่ายประสาทเทียมแบบวนซ้ำด้วยคุณลักษณะและชั้นที่ซ่อนอยู่จำนวนมาก (ตัวแบบเชิงลึก) ทำงานร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติ เพื่อใช้ในการเรียนรู้การตรวจจับข่าวปลอมภาษาไทย

บทที่ 3
วิธีการดำเนินการวิจัย

ผู้วิจัยได้ดำเนินการศึกษา รวบรวม วิเคราะห์ และออกแบบระบบเพื่อการค้นหาคุณลักษณะที่สำคัญ และการสร้างตัวแบบการเรียนรู้ของเครื่องในการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ โดยตัวแบบการเรียนรู้ของเครื่องมีจำนวน 3 ตัวแบบ ซึ่งใช้เทคนิคการเรียนรู้ของเครื่องที่ต่างกัน โดยใช้เทคนิคต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม ตามลำดับ โดยภาพรวมของระบบจะเป็นดังภาพที่ 3-1



ภาพที่ 3-1 ภาพรวมของการดำเนินการวิจัย

ลำดับวิธีการทำงานของระบบจะแบ่งออกเป็น 10 ส่วน คือการเชื่อมต่อระบบกับสื่อสังคมออนไลน์ทวิตเตอร์เพื่อค้นคืนข้อมูล การแปลงข้อมูลและบันทึกลงสู่ฐานข้อมูล การตรวจสอบข้อมูลโดยผู้เชี่ยวชาญ การเพิ่มคุณลักษณะด้านคะแนนความรู้สึก การเตรียมข้อมูลและทำมาตรฐานข้อมูล การเลือกคุณลักษณะที่สำคัญของข้อมูล การสร้างตัวแบบและคำนวณค่าความถูกต้อง การเปรียบเทียบความถูกต้องของตัวแบบ การเปรียบเทียบคุณลักษณะที่สำคัญ การทดสอบคุณลักษณะที่สำคัญ โดยมีรายละเอียดดังต่อไปนี้

3.1 การเชื่อมต่อระบบกับสื่อสังคมออนไลน์ทวิตเตอร์เพื่อค้นคืนข้อมูล

ในส่วนของการเชื่อมต่อระบบกับสื่อสังคมออนไลน์ทวิตเตอร์นั้นผู้วิจัยได้สร้างโปรแกรมด้วยกรอบการทำงานที่ชื่อว่า Groovy on Grails [31] ซึ่งเป็นกรอบการทำงานในรูปแบบเว็บด้วยการใช้ภาษา Groovy [32] และผู้วิจัยได้ศึกษาการค้นคืนข้อมูลผ่านสื่อสังคมออนไลน์ทวิตเตอร์ผ่านทาง Application Programming Interface (API) ของสื่อสังคมออนไลน์ทวิตเตอร์ ซึ่งก่อนที่จะสามารถเชื่อมต่อได้ จำเป็นที่จะต้องสมัครใช้บริการ และได้รับการอนุมัติจากสื่อสังคมออนไลน์ทวิตเตอร์ก่อน หลังจากได้รับการอนุมัติให้สามารถค้นคืนข้อมูลได้แล้ว ผู้วิจัยจะได้รับรหัส 4 ชุด ได้แก่ Access Token, Access Secret, Consumer Key และ Consumer Secret เพื่อใช้ในการยืนยันตัวก่อนที่จะสามารถค้นคืนข้อมูลจากสื่อสังคมออนไลน์ทวิตเตอร์

โดยในงานวิจัยฉบับนี้ได้เลือกข้อมูลที่เป็นการให้ข่าวสาร ความคิดเห็นประกอบกับข่าวสารเป็นหลักใน หรือเนื้อหาที่มีแนวโน้มในการชักจูงความเข้าใจของผู้อ่าน โดยประเภทของข่าวสารมีเนื้อหาที่แตกต่างกัน เช่น ข่าวประเภทกีฬา ข่าวประเภทสังคม

ในการค้นคืนข้อมูลข่าวจากสื่อสังคมออนไลน์ทวิตเตอร์ จะต้องใช้รหัส หรือ Identification (ID) ของข่าว ๆ นั้น ในการระบุค้นคืนข้อมูล โดยข้อมูลที่ได้รับจะแบ่งออกเป็น 2 ชุด คือ ข้อมูลเกี่ยวกับเนื้อหา และข้อมูลเกี่ยวกับผู้เผยแพร่ โดยมีรายละเอียดดังต่อไปนี้

3.1.1 ข้อมูลเกี่ยวกับเนื้อหา

ในส่วนเนื้อหาของข่าวเมื่อทำการร้องขอข้อมูลแล้วจะได้ข้อมูลจากสื่อสังคมออนไลน์ทวิตเตอร์ ดังแสดงในตารางที่ 3-1

ตารางที่ 3-1 ข้อมูลเกี่ยวกับเนื้อหาที่ได้จากการค้นคืนข้อมูลผ่านสื่อสังคมออนไลน์ทวิตเตอร์

ข้อมูลที่ได้	ความหมาย
created_at	ถูกสร้างข้อมูลเมื่อวันเวลาใด
id	รหัสเฉพาะของการเผยแพร่หรือข่าวนั้น ๆ เป็นตัวเลข
id_str	รหัสเฉพาะของการเผยแพร่หรือข่าวนั้น ๆ เป็นแบบข้อความ
text	เนื้อหาของข่าว
hashtags	แฮชแท็ก หรือ การระบุกลุ่มของข้อมูล ของเนื้อหาข่าว
symbols	สัญลักษณ์ที่เกิดขึ้นในเนื้อหาข่าว
user_mentions	ผู้ใช้งานอื่นที่ถูกอ้างอิงในเนื้อหาข่าว
urls	ลิงค์ภายนอกที่เกิดขึ้นในเนื้อหาข่าว
retweet_count	จำนวนการแบ่งปันข่าว
favorite_count	จำนวนการกดชื่นชอบหรือถูกใจข่าว
favorited	ข่าวนี้ได้ถูกกดชื่นชอบหรือไม่
retweeted	ข่าวนี้ได้ถูกแบ่งปันหรือไม่
lang	ภาษาที่ใช้
geo	พื้นที่ที่ข่าวนี้ถูกสร้าง

3.1.2 ข้อมูลเกี่ยวกับผู้สร้างเนื้อหาหรือผู้เผยแพร่

ในส่วนผู้สร้างเนื้อหาเมื่อค้นคืนข้อมูลแล้วจะได้รับข้อมูลจากสื่อสังคมออนไลน์ทวิตเตอร์ดังแสดงในตารางที่ 3-2

ตารางที่ 3-2 ข้อมูลผู้สร้างเนื้อหาที่ได้จากการค้นคืนข้อมูลผ่านสื่อสังคมออนไลน์ทวิตเตอร์

ข้อมูลที่ได้	ความหมาย
location	สถานที่ของบัญชีผู้ใช้
id	รหัสเฉพาะของบัญชีผู้ใช้นั้น ๆ เป็นตัวเลข
id_str	รหัสเฉพาะของบัญชีผู้ใช้นั้น ๆ เป็นแบบข้อความ
name	ชื่อของบัญชีผู้ใช้
screen_name	ชื่อที่แสดงสาธารณะ
description	คำอธิบายของบัญชีผู้ใช้
urls	การเชื่อมโยงภายนอกที่อยู่ในบัญชี
followers_count	จำนวนผู้ติดตาม
friends_count	จำนวนเพื่อน
created_at	ถูกสร้างข้อมูลบัญชีผู้ใช้เมื่อวันเวลาใด
listed_count	จำนวนการเผยแพร่ตั้งแต่สร้างบัญชีผู้ใช้

โดยข้อมูลที่ได้รับมาจากสื่อสังคมออนไลน์ทวิตเตอร์จะถูกนำไปทำความสะอาดข้อมูลก่อนที่จะบันทึกลงฐานข้อมูลต่อไปในขั้นตอนที่ 3.2

3.2 การแปลงและบันทึกลงสู่ฐานข้อมูล

ก่อนจะนำข้อมูลไปใช้ผู้วิจัยจำเป็นต้องทำความสะอาดข้อมูล และแปลงข้อมูล เพื่อให้การนำข้อมูลไปใช้ในการสอนและทดสอบตัวแบบการเรียนรู้ของเครื่องเป็นไปอย่างมีประสิทธิภาพมากที่สุด โดยจะแบ่งงานออกเป็น 3 ส่วนคือ การคำนวณข้อมูลเนื้อหา การคำนวณข้อมูลเกี่ยวกับผู้สร้างเนื้อหา และการบันทึกข้อมูลลงสู่ฐานข้อมูล

3.2.1 การคำนวณข้อมูลเนื้อหา

ในส่วนของการคำนวณข้อมูลเนื้อหานั้นจะทำการทำความสะอาดข้อมูล โดยการแปลงข้อมูล และการเพิ่มคุณลักษณะบางส่วนเพื่อนำไปใช้สอนและทดสอบตัวแบบดังต่อไปนี้

3.2.1.1 เวลาที่ถูกสร้างข้อมูล

สำหรับข้อมูลที่ถูกสร้างจะได้รับมาเป็นรูปแบบของวันเดือนปีและเวลาของการสร้างข้อมูลซึ่งก่อนจะนำไปใช้ผู้วิจัยจำเป็นจะต้องแปลงข้อมูลให้เป็นตัวเลขโดยจะแบ่งเป็น 2 ส่วน คือข้อมูลวันที่และข้อมูลเวลาดังต่อไปนี้

ข้อมูลวันที่จะทำการแปลงข้อมูลโดยคำนวณออกมาเป็นข้อมูลวันของสัปดาห์และแปลงข้อมูลเป็นตัวเลขโดยใช้หลักเกณฑ์ดังตารางที่ 3-3

ตารางที่ 3-3 การแปลงข้อมูลวันของสัปดาห์เป็นตัวเลข

ข้อมูลวัน	ค่าตัวเลขที่ใช้
วันอาทิตย์	1
วันจันทร์	2
วันอังคาร	3
วันพุธ	4
วันพฤหัสบดี	5
วันศุกร์	6
วันเสาร์	7

ข้อมูลเวลาจะทำการแปลงข้อมูลออกเป็นไตรมาสของเวลาหรือออกเป็น 4 ประเภทซึ่งมีรายละเอียดดังตารางต่อไปนี้

ตารางที่ 3-4 การแปลงข้อมูลเวลาเป็นตัวเลข

ข้อมูลเวลา	ค่าตัวเลขที่ใช้
00:00:00 น. - 05:59:59 น.	1
06:00:00 น. - 11:59:59 น.	2
12:00:00 น. - 17:59:59 น.	3
18:00:00 น. - 23:59:59 น.	4

3.2.1.2 จำนวนคำและจำนวนอักษร

ในการคำนวณจำนวนคำและจำนวนอักษรภาษาไทย ผู้วิจัยได้ใช้ไลบรารีชื่อ PyThaiNLP [33] ในการคำนวณทั้งจำนวนคำและจำนวนอักษรภาษาไทย

3.2.1.3 จำนวนเครื่องหมายปรัศนีและเครื่องหมายอัศเจรีย์

ในส่วนของจำนวนเครื่องหมายอัศเจรีย์ และจำนวนเครื่องหมายอัศเจรีย์ ผู้วิจัยได้สร้างโปรแกรมเพิ่มเติมจากโปรแกรมค้นคืนข้อมูลด้วยภาษา Groovy เพื่อทำการคำนวณข้อมูลจำนวนเครื่องหมายปรัศนีและเครื่องหมายอัศเจรีย์ จากเนื้อหาของข่าว

3.2.1.4 แฮชแท็ก

การคำนวณแฮชแท็ก หรือการระบุกลุ่มของข้อมูล จะใช้โปรแกรมเพิ่มเติมจากโปรแกรมค้นคืนข้อมูลด้วยภาษา Groovy ในการนับจำนวนจากเครื่องหมาย “#” ที่อยู่ในเนื้อหาของข่าว

3.2.1.5 การเชื่อมโยงภายนอก

การคำนวณจำนวนการเชื่อมโยงภายนอกที่อยู่ในเนื้อหาข่าว จะใช้โปรแกรมเพิ่มเติมจากโปรแกรมค้นคืนข้อมูล ในการนับข้อมูลที่ได้จากการค้นคืนค่าผ่านทาง Application Programming Interface ของสื่อสังคมออนไลน์ โดยค้นหารูปแบบของการเชื่อมโยงที่อยู่ภายในเนื้อหาข่าว

3.2.2 การคำนวณข้อมูลเกี่ยวกับผู้สร้างเนื้อหา

ในส่วนของการทำคำนวณเกี่ยวกับข้อมูลผู้สร้างเนื้อหานั้นจะทำการทำความสะอาดข้อมูล โดยการแปลงข้อมูล และการเพิ่มคุณลักษณะบางส่วนเพื่อนำไปสอนและทดสอบตัวแบบดังต่อไปนี้

3.2.2.1 จำนวนอักษรของผู้ใช้บัญชี

ในส่วนของชื่อผู้ใช้บัญชีผู้วิจัยได้นำไปใช้โดยการค้นคืนจำนวนตัวอักษรของชื่อผู้ใช้บัญชีเป็นหนึ่งในคุณลักษณะที่นำไปใช้สอนและทดสอบตัวแบบการเรียนรู้ โดยคำนวณได้จากการใช้ไลบรารี PyThaiNLP ในการคำนวณจำนวนอักษร

3.2.2.2 ชื่อผู้ใช้มีตัวเลขหรือไม่

เป็นการคำนวณว่าชื่อผู้ใช้ประกอบด้วยตัวเลขหรือไม่ โดยสามารถคำนวณจากรูปแบบของตัวเลขภายในตัวชื่อผู้ใช้ซึ่งจะใช้เกณฑ์ดังตารางที่ 3-5 ในการพิจารณาข้อมูล

ตารางที่ 3-5 การตรวจสอบชื่อบัญชีผู้ใช้มีตัวเลขหรือไม่

คุณลักษณะจากการคำนวณ	เกณฑ์	ค่าตัวเลขที่ใช้
ชื่อบัญชีผู้ใช้มีตัวเลขหรือไม่	มี	1
	ไม่มี	0

3.2.2.3 จำนวนอักษรของชื่อที่แสดงสาธารณะ

ในส่วนของชื่อที่แสดงสาธารณะผู้วิจัยได้นำไปใช้โดยการค้นคืนจำนวนตัวอักษรของชื่อผู้ใช้บัญชีเป็นหนึ่งในคุณลักษณะที่นำไปใช้สอนและทดสอบตัวแบบการเรียนรู้ โดยคำนวณได้จากการใช้ไลบรารี PyThaiNLP ในการคำนวณจำนวนอักษร

3.2.2.4 อัตราส่วนระหว่างเพื่อนและผู้ติดตาม

ในการคำนวณอัตราส่วนระหว่างเพื่อนและผู้ติดตามจะคำนวณโดยใช้สมการที่ (1) ในการคำนวณ

$$\textit{อัตราส่วน} = \textit{จำนวนเพื่อน}/\textit{จำนวนผู้ติดตาม} \quad (3\text{-}1)$$

เมื่อได้ข้อมูลอัตราส่วนที่นำไปใช้จากสมการข้างต้นแล้ว จึงนำข้อมูลดังกล่าวบันทึกลงสู่ฐานข้อมูลเพื่อนำไปใช้สอนและทดสอบตัวแบบการเรียนรู้ต่อไป

3.2.3 การบันทึกลงสู่ฐานข้อมูล

จากนั้นบันทึกข้อมูลจากสื่อสังคมออนไลน์ และข้อมูลที่ทำการแปลงแล้วลงสู่ฐานข้อมูลส่วนกลางซึ่งใช้ฐานข้อมูล MySQL [34] ในการเก็บข้อมูล

3.3 การตรวจสอบข้อมูลโดยผู้เชี่ยวชาญ

ผู้วิจัยได้ทำการค้นคืนข้อมูลจากฐานข้อมูล และส่งข้อมูลให้กับผู้เชี่ยวชาญที่มีคุณสมบัติเป็นผู้ที่มีอาชีพเป็นนักข่าวช่วยในการตรวจสอบข้อมูลโดยระบุ และส่งกลับมาเพื่อบันทึกลงสู่ฐานข้อมูลเพื่อใช้ในการสอนและทดสอบตัวแบบการเรียนรู้ของเครื่องต่อไป

โดยผู้วิจัยกำหนดให้ผู้ตรวจระบุข้อมูลของข่าวเป็นสองลักษณะ คือ ข่าวจริง และข่าวปลอม จากนั้นส่งกลับมาให้กับผู้วิจัยเพื่อบันทึกลงสู่ฐานข้อมูลเพื่อใช้ในการสอนและทดสอบตัวแบบการเรียนรู้ของเครื่องต่อไป

3.4 การเพิ่มคุณลักษณะด้านคะแนนความรู้สึก

ในการเพิ่มคุณลักษณะทางด้านคะแนนความรู้สึกให้กับข้อมูล ผู้วิจัยได้ใช้ไลบรารีชื่อ PyThaiNLP เพื่อใช้ในการกำหนดคะแนนความรู้สึกให้กับข้อมูล ซึ่งไลบรารีได้ใช้อัลกอริทึมหรือเทคนิคชื่อ Naïve Bayes [18,19] ในการให้คะแนนความรู้สึกกับข้อมูล โดยใช้คลังคำศัพท์พื้นฐานของไลบรารี โดยกำหนดค่าความรู้สึกเป็นกลาง มีค่าเท่ากับ 0 หากเป็นเชิงบวกมีค่าเท่ากับ 1 และเชิงลบมีค่าเท่ากับ -1

3.5 การเตรียมข้อมูลและทำมาตรฐานข้อมูล

หลังจากเพิ่มคุณลักษณะทางด้านคะแนนความรู้สึกให้กับข้อมูลแล้ว ข้อมูลคุณลักษณะที่จะนำมาใช้ในการสอนและทดสอบตัวแบบจะถูกสรุปได้ตามตารางที่ 3-6

ตารางที่ 3-6 คุณลักษณะที่ใช้สอนและทดสอบตัวแบบ

คุณลักษณะที่ใช้สอนและทดสอบตัวแบบ		
ความยาวของชื่อผู้ใช้	ความยาวของชื่อแสดง	ชื่อผู้ใช้มีตัวเลขหรือไม่
จำนวนผู้ติดตาม	จำนวนเพื่อน	อัตราส่วนเพื่อนและผู้ติดตาม
คะแนนความรู้สึก	จำนวนการแบ่งปัน	จำนวนการกดชื่นชอบ
ความยาวอักษรของเนื้อหา	จำนวนคำของเนื้อหา	จำนวนการเชื่อมโยง
ช่วงเวลาการเผยแพร่	วันในสัปดาห์ของการเผยแพร่ข่าว	ชื่อผู้ใช้มีตัวเลขหรือไม่
จำนวนเครื่องหมายอัศเจรีย์	จำนวนแฮชแท็ก	จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี
จำนวนเครื่องหมายปรัศนี		

ซึ่งในขั้นตอนนี้จะนำข้อมูลที่ได้จากการตรวจระบุข้อมูลแล้ว และเพิ่มคุณลักษณะทางด้านคะแนนความรู้สึกมาทำการเตรียมข้อมูลโดยการทำให้ข้อมูลอยู่ในมาตรฐานเดียวกัน โดยในส่วนของการเตรียมข้อมูลจะแบ่งออกเป็น 2 ขั้นตอน คือการทำมาตรฐานข้อมูล และการแบ่งชุดข้อมูล โดยมีรายละเอียดดังต่อไปนี้

3.5.1 การทำมาตรฐานข้อมูล

ในส่วนนี้ผู้วิจัยทำการ Standardization [10] หรือการทำมาตรฐานข้อมูลด้วยเทคนิค Standard Scaler [12,13] ซึ่งจะทำให้ข้อมูลของงานวิจัยฉบับนี้อยู่ในมาตรฐานเดียวกันก่อนที่จะนำไปใช้สอนและทดสอบตัวแบบดังสมการ

$$x_{i(scaled)} = \frac{x_i - \mu}{\sigma} \quad (3\text{-}2)$$

โดยที่

x_i คือข้อมูลอินพุต

μ คือค่าเฉลี่ยของข้อมูลอินพุตทั้งหมด

σ คือค่าเบี่ยงเบนมาตรฐานของอินพุตทั้งหมด

3.5.2 การแบ่งชุดข้อมูล

ผู้วิจัยได้ทำการแบ่งข้อมูลข่าวทั้งหมดออกเป็น 2 กลุ่ม คือกลุ่มที่ถูกใช้สอนและกลุ่มที่ถูกใช้ทดสอบ โดยผู้วิจัยได้แบ่งข้อมูลออกเป็น 5 ชุดดังตารางที่ 3-7

ตารางที่ 3-7 ชุดข้อมูลที่ใช้สอนและทดสอบตัวแบบ

ชื่อชุดข้อมูล	ปริมาณข้อมูลที่ใช้สอน	ปริมาณข้อมูลที่ใช้ทดสอบ
ชุดข้อมูลที่ 1	50%	50%
ชุดข้อมูลที่ 2	60%	40%
ชุดข้อมูลที่ 3	70%	30%
ชุดข้อมูลที่ 4	80%	20%
ชุดข้อมูลที่ 5	90%	10%

3.6 การเลือกคุณลักษณะที่สำคัญของข้อมูล

การเลือกคุณลักษณะที่สำคัญ ในงานวิจัยฉบับนี้ได้เลือกใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier และคัดเลือกคุณลักษณะที่มีความสำคัญเพื่อเปรียบเทียบกับคุณลักษณะสำคัญที่ได้จากตัวแบบการเรียนรู้ต่อไป

3.7 การสร้างตัวแบบและคำนวณค่าความถูกต้อง

งานวิจัยฉบับนี้ได้สร้างตัวแบบทั้งหมด 3 ตัวแบบเพื่อใช้ในการตรวจสอบข่าว โดยใช้เทคนิค คือ ต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม โดยการทดสอบการปรับค่าพารามิเตอร์ของเทคนิคที่ใช้โดยมีรายละเอียดดังต่อไปนี้

3.7.1 ตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ

เป็นตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจเป็นตัวพยากรณ์ข่าว โดยทำการปรับพารามิเตอร์เป็นค่าต่าง ๆ ดังตารางที่ 3-8

ตารางที่ 3-8 การปรับค่าพารามิเตอร์ของตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ

พารามิเตอร์					
Criterion	**Splitter**	**Min sample split**	**Min sample leaf**	**Min weight fraction leaf**	**Max depth**
gini	best	2	1	0.0	none
entropy	random	3	2	0.1	1
		4	3	0.2	2
		5	4	0.5	3
			5	1	4
					5
					6

3.7.2 ตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

เป็นตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน เป็นตัวพยากรณ์ข่าว โดยทำการปรับพารามิเตอร์เป็นค่าต่าง ๆ ดังตารางที่ 3-9

ตารางที่ 3-9 การปรับค่าพารามิเตอร์ของตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

พารามิเตอร์								
C	**Class weight**	**Dual**	**Fit intercept**	**Inter-cept scaling**	**loss**	**Multi class**	**penalty**	**Random state**
1000	dict	true	true	0.1	hinge	ovr	l1	none
100	balanced	false	false	1	squared_hinge	crammer_singer	l2	1
10				10				2
1				100				3
0.1				1000				4

3.7.3 ตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียม

เป็นตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียม เป็นตัวพยากรณ์ข่าว โดยทำการปรับพารามิเตอร์เป็นค่าต่าง ๆ ดังตารางที่ 3-10

ตารางที่ 3-10 การปรับค่าพารามิเตอร์ของตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียม

พารามิเตอร์								
Random state	**Verbose**	**Solver**	**Alpha**	**Max tier**	**Activation**	**Hidden layer sizes**	**Learning rate init**	**tol**
2	10	sgd	0.0001	300	relu	50	0.01	0.0001
3	100	adam	0.00001	450	tanh	100	0.1	0.00001
4		lbfgs		500	identity	500		
				600		900		
				700		1000		
				1,000				
				10,000				
				1,000,000				

3.8 ส่วนการเปรียบเทียบความถูกต้องของตัวแบบ

การเปรียบเทียบความถูกต้องของตัวแบบจะใช้การคำนวณค่าความถูกต้อง ค่าความแม่นยำ ค่าความอ่อนไหว และค่าประสิทธิภาพของตัวแบบ เพื่อใช้ในการเปรียบเทียบและประเมินประสิทธิภาพของตัวแบบที่ทำการทดลอง

3.9 การเปรียบเทียบคุณลักษณะที่สำคัญ

ในการคัดเลือกคุณลักษณะที่สำคัญของข่าวปลอม ผู้วิจัยได้ใช้ 2 วิธีในการค้นหาคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมได้แก่ การใช้เทคนิคต้นไม้ตัดสินใจในการหาคุณลักษณะที่สำคัญและคุณลักษณะสำคัญที่ได้จากตัวแบบการเรียนรู้ จากนั้นนำผลของทั้งสองวิธีมาเปรียบเทียบโดยจะมีรายละเอียดดังต่อไปนี้

3.9.1 การใช้เทคนิคต้นไม้ตัดสินใจในการหาคุณลักษณะที่สำคัญ

ในการวิจัยนี้ผู้วิจัยได้ใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier ในการค้นหาคุณลักษณะที่สำคัญของการตรวจสอบข่าวปลอม โดยผลลัพธ์ที่ได้จะเป็นอัตราส่วนความสำคัญของคุณลักษณะที่ใช้ในการตรวจสอบข่าวปลอม โดยทำการสร้างตัวแบบที่ใช้ต้นไม้ตัดสินใจแบบ Extra Trees Classifier และนำข้อมูลที่ผ่านการทำมาตรฐานข้อมูลแล้วมาใช้ในการให้ตัวแบบเรียนรู้แล้วประเมิณ ซึ่งผลลัพธ์ที่ได้จะอยู่ในรูปแบบของอัตราส่วนของความสำคัญของคุณลักษณะ

3.9.2 การได้มาซึ่งคุณลักษณะสำคัญจากตัวแบบการเรียนรู้ของเครื่อง

ในงานวิจัยฉบับนี้ได้สร้างตัวแบบการเรียนรู้ของเครื่องด้วยเทคนิค ต้นไม้ตัดสินใจ ซัพพอร์ทเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม โดยแต่ละตัวแบบสามารถคัดเลือกคุณลักษณะเฉพาะ จากความสามารถของไลบรารี่ที่ใช้ได้ โดยมีรายละเอียดตามหัวข้อ 2.1.11

3.9.3 การเปรียบเทียบคุณลักษณะที่สำคัญ

หลังจากได้คุณลักษณะที่สำคัญจากหัวข้อ 3.6.2.1 และ 3.6.2.2 แล้ว จะนำมาทำการเปรียบเทียบและเลือกเฉพาะคุณลักษณะที่ซ้อนทับกัน (Intersection) จากทั้งสองหัวข้อเลือกเป็นคุณลักษณะที่สำคัญ

3.10 การทดสอบคุณลักษณะที่สำคัญ

ในการทดสอบคุณลักษณะที่สำคัญ จะนำเฉพาะข้อมูลในคุณลักษณะสำคัญที่ได้จากหัวข้อ 3.6.2.4 มาสอนและทดสอบตัวแบบที่ได้ความถูกต้องมากที่สุดอีกครั้ง เพื่อสังเกตการเปลี่ยนแปลงของค่าความถูกต้อง หากค่าความถูกต้องคลาดเคลื่อนไม่เกินร้อยละ 3 จะสรุปว่าคุณลักษณะที่เลือกมาเป็นคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอม

หากมีความคลาดเคลื่อนมากเกินกว่าร้อยละ 3 จะทำการทดสอบตัดคุณลักษณะที่สำคัญออกครั้งละ 1 คุณลักษณะจากนั้นนำข้อมูลไปสอนและทดสอบตัวแบบและสังเกตความแตกต่างของค่าความถูกต้องจนกว่าค่าความคลาดเคลื่อนจะน้อยกว่าร้อยละ 3

บทที่ 4
ผลการดำเนินงาน

ผลการศึกษางานวิจัยในครั้งนี้ได้แบ่งออกเป็น 6 ส่วน ได้แก่ ผลการเก็บข้อมูล ผลการตรวจระบุข้อมูล ผลการประเมินคะแนนความรู้สึกจากเนื้อหาข่าว ผลการหาคุณลักษณะที่สำคัญ ผลการสร้างและทดสอบตัวแบบ ผลการคัดเลือกคุณลักษณะที่สำคัญจากตัวแบบ ผลการประเมินประสิทธิภาพของตัวแบบ ผลการเปรียบเทียบคุณลักษณะที่สำคัญ และผลจากการทดสอบคุณลักษณะที่สำคัญ โดยมีรายละเอียดดังต่อไปนี้

4.1 ผลการเก็บข้อมูล

ผู้วิจัยได้ทำการเก็บข้อมูลข่าวจากสื่อสังคมออนไลน์ทวิตเตอร์ตามขอบเขตงานวิจัย ได้ข้อมูลข่าวทั้งหมด 386 ข่าว ผ่านทาง Application Programming Interface (API) จากการสร้างโปรแกรมด้วยกรอบการทำงานที่ชื่อว่า Groovy on Grails จากนั้นทำการแปลงข้อมูลตามหลักเกณฑ์ที่กล่าวไว้ในบทที่ 3 ก่อนจะบันทึกข้อมูลลงสู่ฐานข้อมูล

4.2 ผลการตรวจระบุข้อมูล

ผู้วิจัยได้ทำการค้นคืนข้อมูลจากฐานข้อมูล และส่งข้อมูลให้กับคุณธนัท วันวิน ซึ่งเป็นผู้ที่มีอาชีพเป็นนักข่าวช่วยในการตรวจสอบข้อมูลโดยระบุ และส่งกลับมาเพื่อบันทึกลงสู่ฐานข้อมูลเพื่อใช้ในการสอนและทดสอบตัวแบบการเรียนรู้ของเครื่องต่อไป โดยข้อมูลที่ส่งให้ผู้ตรวจจะเป็นไปตามบทที่ 3 และผลลัพธ์ที่ผู้มีอาชีพนักข่าวส่งกลับมาจะมีรายละเอียดดังตารางที่ 4-1

ตารางที่ 4-1 ประเภทข่าวหลังจากตรวจระบุข่าวจากผู้มีวิชาชีพนักข่าว

ประเภทข่าว	จำนวน(ข่าว)	คิดเป็นเปอร์เซ็นต์
ข่าวจริง	227	59%
ข่าวปลอม	159	41%

4.3 ผลการประเมินคะแนนความรู้สึกจากเนื้อหาข่าว

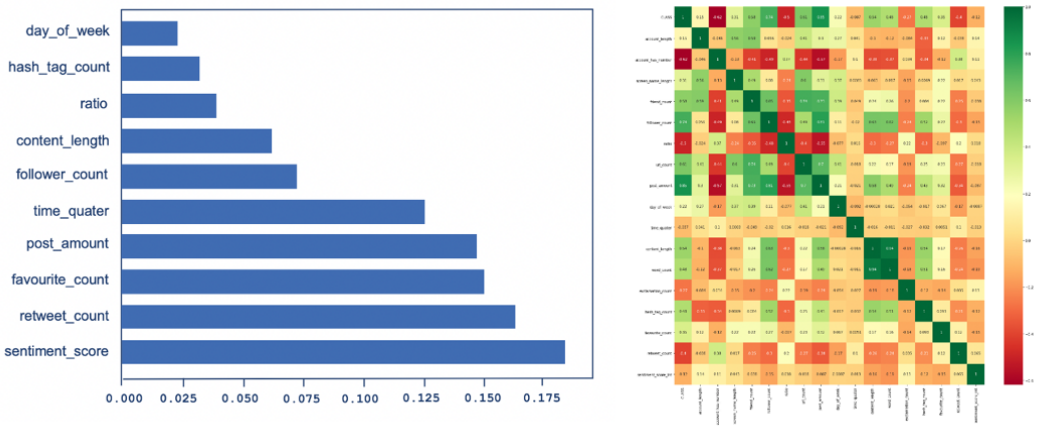
ในงานวิจัยฉบับนี้มีการเพิ่มคุณลักษณะของข่าว คือ คะแนนความรู้สึกของเนื้อหาข่าว โดยใช้เทคนิคการประมวลผลภาษาธรรมชาติประเภทการวิเคราะห์ความรู้สึก ซึ่งผลลัพธ์ที่ได้จะออกมาเป็นเชิงบวก เชิงลบ และเป็นกลาง โดยผลลัพธ์ของการหาคะแนนความรู้สึกของเนื้อหาข่าวจะเป็นดังตารางที่ 4-2

ตารางที่ 4-2 ผลลัพธ์การหาคะแนนความรู้สึกของเนื้อหาข่าว

ประเภทเนื้อหาข่าว	จำนวน(ข่าว)	คิดเป็นเปอร์เซ็นต์
เชิงบวก	158	40.9%
เป็นกลาง	44	11.4%
เชิงลบ	184	47.7%

4.4 ผลการหาคุณลักษณะที่สำคัญ

ในการหาคุณลักษณะที่สำคัญโดยใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier ได้ผลการหาคือคุณลักษณะ วันของสัปดาห์ จำนวนแฮชแท็ก อัตราส่วนเพื่อนและผู้ติดตาม ความยาวของเนื้อหา จำนวนผู้ติดตาม เวลาของการเผยแพร่ข่าว จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี จำนวนการกดถูกใจ จำนวนการแบ่งปัน คะแนนความรู้สึก ดังแสดงในภาพที่ 4-1



ภาพที่ 4-1 คุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมจากเทคนิค Extra Trees Classifier

ในการหาคุณลักษณะที่สำคัญโดยใช้เทคนิคต้นไม้ตัดสินใจแบบ Extra Trees Classifier ได้ผลลัพธ์ คือคุณลักษณะคะแนนความรู้สึกมีความสำคัญต่อการตรวจสอบข่าวปลอมมากที่สุด ซึ่งจำนวนการแบ่งปัน จำนวนการกดถูกใจ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี ช่วงเวลาของการเผยแพร่ข่าว จำนวนผู้ติดตาม ความยาวเนื้อหา อัตราส่วนเพื่อนและผู้ติดตาม จำนวนแฮชแท็ก วันของสัปดาห์มีความสำคัญรองลงมาตามลำดับ

4.5 ผลการสร้างและทดสอบตัวแบบ

ผู้วิจัยได้ทำการสร้างตัวแบบการเรียนรู้ของเครื่องโดยใช้เทคนิคต้นไม้ตัดสินใจ เทคนิคซัพพอร์ทเวกเตอร์แมชชีน และเทคนิคโครงข่ายประสาทเทียม ด้วยไลบรารี Scikit Learn [35] เพื่อใช้ในการสอนและทดสอบตัวแบบ และจากนั้นทำการใช้ข้อมูลข่าวที่เตรียมพร้อมแล้วทำมาตรฐานข้อมูลเพื่อให้ข้อมูลอยู่ในมาตรฐานเดียวกัน จากนั้นนำข้อมูลไปใช้สอนและทดสอบตัวแบบและจดบันทึกค่าความถูกต้อง ค่าความแม่นยำ ค่าความอ่อนไหว และค่าประสิทธิภาพของตัวแบบ โดยมีรายละเอียดดังต่อไปนี้

4.5.1 การทดสอบตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ

ในการทดสอบตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจจะทำการปรับค่าพารามิเตอร์ตามรายละเอียดในบทที่ 3 ซึ่งจะได้ผลการทดสอบ 3 อันดับสูงสุดดังตารางที่ 4-3

ตารางที่ 4-3 ผลลัพธ์การทดสอบตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ 3 อันดับ

อันดับ	พารามิเตอร์						ผลลัพธ์			
	Criterion	Splitter	Min sample split	Min sample leaf	Min weight fraction leaf	Max depth	Accuracy	Precision	Recall	F-measure
1	entropy	best	2	1	0.0	None	0.96	0.96	0.96	0.96
2	gini	best	3	1	0.0	None	0.95	0.94	0.96	0.95
3	gini	best	2	1	0.0	None	0.95	0.94	0.95	0.95

4.5.2 การทดสอบตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

ในการทดสอบตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีนจะทำการปรับค่าพารามิเตอร์ตามรายละเอียดในบทที่ 3 ซึ่งจะได้ผลการทดสอบ 3 อันดับสูงสุดดังตารางที่ 4-4

ตารางที่ 4-4 ผลลัพธ์การทดสอบตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน 3 อันดับ

อันดับ	พารามิเตอร์									ผลลัพธ์			
	C	Class weight	Dual	Fit intercept	Intercept scaling	loss	Multi Class	Penalty	Random State	Accuracy	Precision	Recall	F-measure
1	100	None	True	True	1	squared_hinge	Ovr	l2	3	0.94	0.89	0.98	0.93
2	100	None	True	True	10	squared_hinge	Ovr	l2	3	0.93	0.91	0.96	0.91
3	100	None	false	True	1	hinge	Ovr	l2	3	0.93	0.89	0.95	0.93

4.5.3 การทดสอบตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียม

ในการทดสอบตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมจะทำการปรับค่าพารามิเตอร์ตามรายละเอียดในบทที่ 3 ซึ่งจะได้ผลการทดสอบ 3 อันดับสูงสุดดังตารางที่ 4-5

ตารางที่ 4-5 ผลลัพธ์การทดสอบตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียม 3 อันดับ

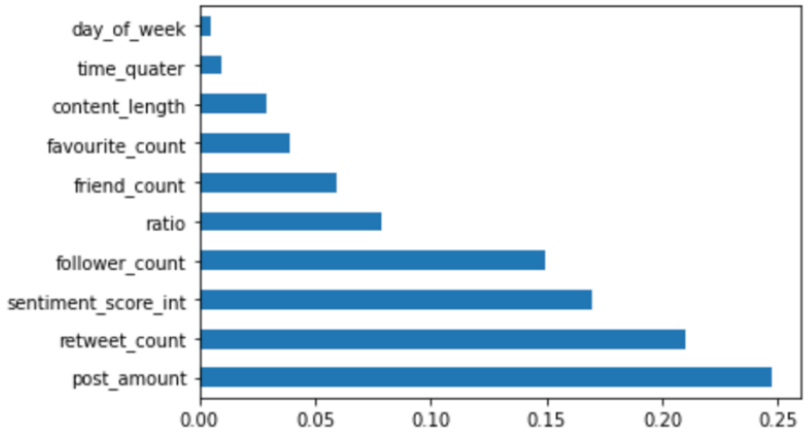
อันดับ	พารามิเตอร์									ผลลัพธ์			
	Random state	Verbose	Solver	Alpha	Max tier	Activation	Hidden layer sizes	Learning rate init	tol	Accuracy	Precision	Recall	F-measure
1	3	10	lbfgs	1e-4	600	tanh	50	0.1	0.00001	0.97	0.92	1	0.96
2	3	10	adam	1e-4	450	Identity	50	0.1	0.00001	0.97	0.92	1	0.95
3	3	10	adam	1e-4	10,000	Identity	50	0.1	0.00001	0.96	0.91	1	0.95

4.6 ผลการคัดเลือกคุณลักษณะที่สำคัญจากตัวแบบ

ในการคัดเลือกคุณลักษณะที่สำคัญจากตัวแบบซึ่งใช้ตัวช่วยในการหาคุณลักษณะสำคัญดังรายละเอียดจากบทที่ 3 ซึ่งได้ผลลัพธ์แยกตามตัวแบบการเรียนรู้ต้นไม้ตัดสินใจ ตัวแบบซัพพอร์ทเวกเตอร์แมชชีน ตัวแบบโครงข่ายประสาทเทียม โดยดังรายละเอียดต่อไปนี้

4.6.1 คุณลักษณะที่สำคัญจากตัวแบบต้นไม้ตัดสินใจ

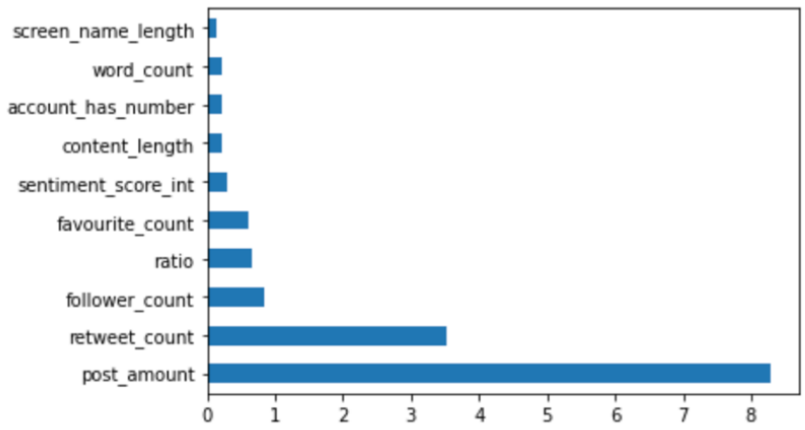
จากการสร้างตัวแบบจากเทคนิคต้นไม้ตัดสินใจ จะได้ผลลัพธ์ความสำคัญของคุณลักษณะเรียงตามความสำคัญ 10 อันดับคือ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี จำนวนการแบ่งปัน คะแนนความรู้สึก จำนวนผู้ติดตาม อัตราส่วนระหว่างเพื่อนและผู้ติดตาม จำนวนเพื่อน จำนวนการกดชื่นชอบ ความยาวเนื้อหา ช่วงเวลาการเผยแพร่ วันที่เผยแพร่ ตามลำดับดังภาพที่ 4-2



ภาพที่ 4-2 คุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมจากตัวแบบต้นไม้ตัดสินใจ

4.6.2 คุณลักษณะที่สำคัญจากตัวแบบซัพพอร์ทเวกเตอร์แมชชีน

จากการสร้างตัวแบบจากเทคนิคซัพพอร์ทเวกเตอร์แมชชีน จะได้ผลลัพธ์ความสำคัญของคุณลักษณะเรียงตามความสำคัญ 10 อันดับคือ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี จำนวนการแบ่งปัน จำนวนผู้ติดตาม อัตราส่วนระหว่างเพื่อนและผู้ติดตาม จำนวนการกดชื่นชอบ คะแนนความรู้สึก ความยาวเนื้อหา บัญชีมีตัวเลขหรือไม่ จำนวนคำ ความยาวของชื่อที่แสดง ตามลำดับดังภาพที่ 4-3



ภาพที่ 4-3 คุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมจากตัวแบบซัพพอร์ทเวกเตอร์แมชชีน

4.6.3 คุณลักษณะที่สำคัญจากตัวแบบโครงข่ายประสาทเทียม

จากการสร้างตัวแบบจากเทคนิคโครงข่ายประสาทเทียม จะได้ผลลัพธ์ความสำคัญของคุณลักษณะเรียงตามความสำคัญ 10 อันดับคือ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี จำนวนการแบ่งปัน ความยาวเนื้อหา จำนวนผู้ติดตาม คะแนนความรู้สึก อัตราส่วนระหว่างเพื่อนและผู้ติดตาม จำนวนเพื่อน จำนวนการกดชื่นชอบ จำนวนเครื่องหมายอัศเจรีย์ จำนวนคำ ตามลำดับดังภาพที่ 4-4

Weight	Feature
0.4758 ± 0.0398	post_amount
0.0369 ± 0.0039	retweet_count
0.0291 ± 0.0111	content_length
0.0286 ± 0.0080	follower_count
0.0208 ± 0.0158	sentiment_score_int
0.0166 ± 0.0084	ratio
0.0130 ± 0.0093	friend_count
0.0114 ± 0.0053	favourite_count
0.0068 ± 0.0091	exclamation_count
0.0052 ± 0.0000	word_count
0.0026 ± 0.0033	account_has_number
0.0021 ± 0.0021	hash_tag_count
0.0016 ± 0.0062	url_count
0.0010 ± 0.0078	screen_name_length
0.0005 ± 0.0039	time_quater
0 ± 0.0000	day_of_week
0 ± 0.0000	question_mark_count

ภาพที่ 4-4 คุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมจากตัวแบบโครงข่ายประสาทเทียม

4.7 ผลการประเมินประสิทธิภาพของตัวแบบ

จากหัวข้อ 4.5 ผลลัพธ์ของการสร้างและทดสอบตัวแบบจะแสดงให้เห็นว่าตัวแบบที่ให้ค่าความถูกต้องมากที่สุดคือตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียม ซึ่งมีค่าความถูกต้องสูงถึง 97% และการวัดค่าความถูกต้อง ค่าความแม่นยำ ค่าความอ่อนไหว และค่าประสิทธิภาพของตัวแบบ ค่าสูงที่สุดของการวัดค่าตัวแบบ จะมีรายละเอียดดังตารางที่ 4-6

ตารางที่ 4-6 ค่าสูงที่สุดของการวัดค่าตัวแบบ

การวัด	ตัวแบบ	ผลลัพธ์
ความถูกต้อง	โครงข่ายประสาทเทียม	0.97
ความแม่นยำ	ต้นไม้ตัดสินใจ	0.96
ความอ่อนไหว	โครงข่ายประสาทเทียม	1
ค่าประสิทธิภาพของตัวแบบ	โครงข่ายประสาทเทียมและต้นไม้ตัดสินใจ	0.96

4.8 ผลการเปรียบเทียบคุณลักษณะที่สำคัญ

จากหัวข้อ 4.4 และ 4.6 สามารถแสดงข้อมูลการเปรียบเทียบคุณลักษณะที่สำคัญได้ดังแสดงในตารางที่ 4-7

ตารางที่ 4-7 ตารางเปรียบเทียบคุณลักษณะที่สำคัญ

คุณลักษณะ	Extra Trees Classifier	ตัวแบบ			รวม
		ต้นไม้ตัดสินใจ	ซัพพอร์ทเวกเตอแมชชีน	โครงข่ายประสาทเทียม	
ความยาวของชื่อผู้ใช้					0
จำนวนผู้ติดตาม	✓	✓	✓	✓	4
คะแนนความรู้สึก	✓	✓	✓	✓	4
ความยาวอักษรของเนื้อหา	✓	✓	✓	✓	4
ความยาวของชื่อแสดง			✓		1

คุณลักษณะ	Extra Trees Classifier	ตัวแบบ			รวม
		ต้นไม้ตัดสินใจ	ซัพพอร์ทเวกเตอแมชชีน	โครงข่ายประสาทเทียม	
จำนวนเพื่อน		✓		✓	2
จำนวนการแบ่งปัน	✓	✓	✓	✓	4
จำนวนคำของเนื้อหา			✓	✓	2
วันของการเผยแพร่	✓	✓			2
จำนวนแฮชแท็ก	✓				1
ชื่อผู้ใช้มีตัวเลขหรือไม่			✓		1
อัตราส่วนเพื่อนและ	✓	✓	✓	✓	4
จำนวนการกดชื่นชอบ	✓	✓	✓	✓	4
จำนวนการเชื่อมโยง					0
ชื่อที่แสดงมีตัวเลขหรือไม่					0
จำนวนการเผยแพร่ตั้งแต่	✓	✓	✓	✓	4
ช่วงเวลาในการเผยแพร่ข่าว	✓	✓			2
จำนวนเครื่องหมายอัศเจรีย์				✓	1
จำนวนเครื่องหมายปรัศนี					0

จากตารางที่ 4-7 คุณลักษณะที่พบในการทดลองหาคุณลักษณะที่สำคัญจากทุกแหล่ง มีจำนวน 7 คุณลักษณะ คือ จำนวนผู้ติดตาม คะแนนความรู้สึก ความยาวอักษรของเนื้อหา จำนวนการแบ่งปัน อัตราส่วนของเพื่อนและผู้ติดตาม จำนวนการกดชื่นชอบ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี

4.9 ผลจากการทดสอบคุณลักษณะที่สำคัญ

เมื่อได้คุณลักษณะที่สำคัญจากการเปรียบเทียบ ในหัวข้อ 4.8 แล้ว จะนำข้อมูลเฉพาะคุณลักษณะ จำนวนผู้ติดตาม คะแนนความรู้สึก ความยาวอักษรของเนื้อหา จำนวนการแบ่งปัน อัตราส่วนของเพื่อนและผู้ติดตาม จำนวนการกดชื่นชอบ จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี มาใช้ใน

การทดสอบตัวแบบที่ได้ความถูกต้องมากที่สุด ซึ่งคือ ตัวแบบการเรียนรู้ที่ใช้เทคนิคโครงข่ายประสาทเทียม ซึ่งได้ความถูกต้อง 97 เปอร์เซ็นต์ และเมื่อนำข้อมูลเฉพาะคุณลักษณะที่คัดเลือกมาแล้วไปใช้ในการสอนและทดสอบตัวแบบพบว่าได้ค่าความถูกต้องที่ 95 เปอร์เซ็นต์

บทที่ 5
สรุปผลวิจัยและข้อเสนอแนะ

งานวิจัยนี้ได้ศึกษาการเปรียบเทียบตัวแบบการเรียนรู้ระหว่างตัวแบบการเรียนรู้ที่ใช้เทคนิคต้นไม้ตัดสินใจ เทคนิคซัพพอร์ทเวกเตอร์แมชชีน และเทคนิคโครงข่ายประสาทเทียม เพื่อหาตัวแบบที่เหมาะสมในการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

โดยการดำเนินการวิจัยเริ่มจากการเก็บรวบรวมข่าวจริงและปลอมบนสื่อสังคมออนไลน์ทวิตเตอร์ จากนั้นนำข่าวที่ได้ส่งต่อให้กับผู้มีวิชาชีพทางด้านข่าวสารในการตรวจระบุข่าวเพื่อระบุว่าข่าวใดเป็นข่าวจริงข่าวใดเป็นข่าวปลอม และนำข้อมูลเนื้อหาข่าวไปค้นหาคะแนนความรู้สึกเพื่อเพิ่มคุณลักษณะให้กับข้อมูล จากนั้นนำข้อมูลที่ครบถ้วนไปใช้สอนและทดสอบตัวแบบ และวัดค่าความถูกต้องของตัวแบบ

ในหัวข้อการสรุปผลการวิจัยและข้อเสนอแนะนี้จะแบ่งหัวข้อออกเป็น 3 หัวข้อ คือ ผลการวิจัย ปัญหาและอุปสรรค และข้อเสนอแนะ โดยมีรายละเอียดดังต่อนี้

5.1 สรุปผลการวิจัย

การสรุปผลการวิจัยของงานวิจัยฉบับนี้สามารถแบ่งออกได้เป็น 3 ด้าน คือ ด้านความถูกต้องของตัวแบบ ด้านคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอม และด้านการวิเคราะห์คุณลักษณะร่วมกับกลุ่มตัวอย่าง โดยมีรายละเอียดดังต่อไปนี้

5.1.1 ด้านความถูกต้องของตัวแบบ

ผลการวิจัยในด้านของความถูกต้องของตัวแบบสามารถสรุปได้ว่า ตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมสามารถให้ความถูกต้องในการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์สูงสุดที่ 97 เปอร์เซ็นต์ เมื่อเทียบกับตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ และตัวแบบที่ใช้เทคนิคซัพพอร์ทเวกเตอร์แมชชีน

5.1.2 ด้านคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอม

ผลการวิจัยในด้านของคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ ได้แก่ จำนวนผู้ติดตาม คะแนนความรู้สึก ความยาวอักษรของเนื้อหา จำนวนการแบ่งปัน อัตราส่วนของเพื่อนและผู้ติดตาม จำนวนการกดชื่นชอบ จำนวนการเผยแพร่

ตั้งแต่สร้างบัญชี ซึ่งสามารถแบ่งประเภทของคุณลักษณะตามทฤษฎีในหัวข้อที่ 2.1.3 ได้ดังตารางที่ 5-1

ตารางที่ 5-1 ประเภทของคุณลักษณะตามทฤษฎีในหัวข้อที่ 2.1.3

ประเภท	คุณลักษณะ
ผู้กระจายข่าว	จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี จำนวนผู้ติดตาม อัตราส่วนของเพื่อนและผู้ติดตาม
เนื้อหา	คะแนนความรู้สึก ความยาวอักษรของเนื้อหา
การมีส่วนร่วม	จำนวนการกดชื่นชอบ จำนวนการแบ่งปัน

5.1.3 ด้านการวิเคราะห์คุณลักษณะร่วมกับกลุ่มตัวอย่าง

เมื่อนำคุณลักษณะในหัวข้อ 5.1.2 ไปวิเคราะห์ร่วมกับข้อมูลกลุ่มตัวอย่างพบว่า จำนวนการเผยแพร่ตั้งแต่สร้างบัญชี มีความสำคัญสำหรับตัวแบบในการทำนายผล หากมีการเผยแพร่ตั้งแต่สร้างบัญชีปริมาณที่สูงจะมีแนวโน้มสูงในการทำนายเป็นข่าวจริง

ส่วนของจำนวนผู้ติดตามและอัตราส่วนระหว่างเพื่อนและผู้ติดตาม เมื่อวิเคราะห์ร่วมกับข้อมูลกลุ่มตัวอย่างพบว่า หากมีจำนวนเพื่อนต่ำและมีปริมาณของผู้ติดตามสูงมีแนวโน้มสูงในการทำนายเป็นข่าวปลอม

ส่วนของคะแนนความรู้สึกและความยาวอักษรของเนื้อหานั้น จะสัมพันธ์กับปริมาณของจำนวนการกดชื่นชอบและจำนวนการแบ่งปัน ซึ่งสามารถตีความได้ว่าหากคะแนนความรู้สึกเป็นเชิงลบ มีแนวโน้มที่จะทำให้มีการกดชื่นชอบและแบ่งปันสูง หากความยาวของอักษรและเนื้อหาต่ำจะมีจำนวนการกดชื่นชอบและแบ่งปันสูง ซึ่งจุดประสงค์ของการสร้างข่าวปลอมต้องการให้มีการแบ่งปันสูง ทำให้อนุมานได้ว่า หากความยาวของเนื้อหาต่ำและคะแนนความรู้สึกเป็นเชิงลบมีแนวโน้มสูงในการทำนายเป็นข่าวปลอม

5.2 ปัญหาและอุปสรรค

อุปสรรคที่เกิดขึ้นอยู่ในขั้นตอนการเก็บข้อมูลจากสื่อสังคมออนไลน์ เนื่องจากข่าวปลอมที่อยู่บนสื่อสังคมออนไลน์ทวิตเตอร์ เกิดขึ้นในเวลาจำกัด กล่าวคือจะถูกลบออกไปค่อนข้างเร็ว และเนื้อหาข่าวปลอมจะค่อนข้างถูกค้นพบได้ยาก อีกทั้งใช้เวลานานในการตรวจสอบว่าเป็นข่าวปลอมหรือไม่ก่อนทำการเก็บข้อมูล ทำให้ใช้เวลาในส่วนการเก็บข้อมูลค่อนข้างนานกว่าที่คาดไว้

5.3 ข้อเสนอแนะ

ในส่วนของข้อเสนอแนะ ผู้วิจัยได้แบ่งออกเป็น 2 หัวข้อคือ ข้อเสนอแนะทางด้านการปรับปรุงความถูกต้องและแม่นยำของตัวแบบ และข้อเสนอแนะทางด้านการพัฒนาต่อยอดและนำไปใช้จริง ซึ่งมีรายละเอียดดังต่อไปนี้

5.3.1 ข้อเสนอแนะทางด้านการปรับปรุงความถูกต้องและแม่นยำของตัวแบบ

การเพิ่มความถูกต้องและแม่นยำให้กับตัวแบบ สามารถทำได้หากสามารถเก็บข้อมูลข่าวได้มากขึ้น และปรับพารามิเตอร์ต่าง ๆ ให้กับตัวแบบอย่างเหมาะสม จะสามารถเพิ่มความถูกต้องและแม่นยำได้

อีกทั้งหากสามารถเก็บข้อมูลข่าวและการพัฒนาของข่าวตามเวลาจริงได้ เช่น จำนวนการกดชื่นชอบหรือจำนวนการแบ่งปัน ณ เวลาใด ๆ เป็นข้อมูลอนุกรมเวลาได้ อาจจะสามารถพัฒนาตัวแบบที่ถูกต้องและแม่นยำได้มากขึ้น

5.3.2 ข้อเสนอแนะทางด้านการพัฒนาต่อยอดและนำไปใช้จริง

ในการนำตัวแบบไปใช้จริง สามารถทำได้หลากหลายรูปแบบ เช่น แบบผลัก (push) กล่าวคือให้ผู้ใช้งานทั่วไปนำข่าวที่ตนเองสนใจส่งให้ระบบตรวจสอบ ซึ่งอาจจะเป็นเว็บแอพพลิเคชั่นหรือโมบายแอพพลิเคชั่น หรือเป็นแบบค้นคืน (Pull) หรือเรียกว่าระบบกรองข่าวสาร โดยให้ระบบค้นคืนข้อมูลที่เกิดขึ้นจากสื่อสังคมออนไลน์ทวิตเตอร์ตลอดเวลาเพื่อตรวจสอบข่าวสารที่เกิดขึ้น

เอกสารอ้างอิง

[1] Soroush Vosoughi, Deb Roy, Sinan Aral. (2018). “The spread of true and false news online.” Science. Vol. 359, Issue 6380, pp. 1146-1151 DOI: 10.1126/science.aap9559

[2] Saranya Krishnan, Min Chen. (2018). Identify Tweet With Fake News. Division of Computing and Software Systems, School of STEM University of Washington Bothell Bothell, USA

[3] Oluwaseun Ajao, Deepayan Bhowmik, Shahrzad Zargari. (2017). Fake News Identification on Twitter with Hybrid CNN and RNN Models. C3Ri Research Institute Sheffield Hallam University United Kingdom

[4] Natali Ruchansky, Sungyong Seo, Yan Liu. (2017). CSI: A Hybrid Deep Model for Fake News Detection. University of Southern California Los Angeles, California

[5] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong and Meeyoung Cha. (2016). Detecting rumors from microblogs with recurrent neural networks. The Chinese University of Hong Kong, Hong Kong SAR

[6] AlindGupta. (2020). “ML | Extra Tree Classifier for Feature Selection”. (Online) Available on https://www.geeksforgeeks.org/ml-extra-tree-classifier-for-feature-selection/ (26 November 2020).

[7] Atefeh Farzindar, Diana Inkpen. (2017). Natural Language Processing for Social Media: Second Edition (Synthesis Lectures on Human Language Technologies) 2nd Edition. Morgan & Claypool Publishers. United States of America.

[8] Dávid Natingga. (2017). Data Science Algorithms in a Week: Top 7 algorithms for computing, data analysis, and machine learning. Birmingham:Packt Publishing Ltd. United States of America.

[9] Nick McClure. (2018). TensorFlow Machine Learning Cookbook Second Edition. Packt Publishing Ltd. United States of America.

[10] Aurelien Geron. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow. O’Reilly Media, Inc. United States of America.

[11] Atefeh Farzindar, Diana Inkpen. (2017). Natural Language Processing for Social Media: Second Edition (Synthesis Lectures on Human Language Technologies) 2nd Edition. Morgan & Claypool Publishers. United States of America.

[12] Scikit-learn developers. (2007). “sklearn.preprocessing.StandardScaler”. (Online) Available on https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing. StandardScaler.html (26 November 2020).

[13] Jason Brownlee. (2020). “How to Use StandardScaler and MinMaxScaler Transforms in Python”. (Online) Available on https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/ (26 November 2020).

[14] บุญชม ศรีสะอาด (2535). การวิจัยเบื้องต้น. (ตีพิมพ์ครั้งที่ 3). กรุงเทพฯ: สุวีริยาสาส์น.

[14] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. (2016). Fake News Detection on Social Media: A Data Mining Perspective. Computer Science & Engineering, Arizona State University, Tempe, AZ, USA,Charles River Analytics, Cambridge, MA, USA, Computer Science & Engineering, Michigan State University, East Lansing, MI, USA

[15] Eni Mustafaraj and Panagiotis Takis Metaxas. The fake news spreading plague: Was it preventable? arXiv preprint arXiv:1703.06988, 2017.

[16] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. Proceedings of the Association for Information Science and Technology, 52(1):1–4, 2015.

[17] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. Communications of the ACM, 59(7):96–104, 2016.

[18] Joel Grus. (2017). Data Science from Scratch. O’Reilly Media, Inc. United States of America.

[19] Scikit-learn developers. (2007). “sklearn.tree.DecisionTreeClassifier”. (Online) Available on https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTree Classifier.html#sklearn.tree.DecisionTreeClassifier.feature_importances_ (26 November 2020).

[20] Scikit-learn developers. (2007). “sklearn.svm.SVC”. (Online) Available on https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html (26 November 2020).

[21] Scikit-learn developers. (2007). “Permutation feature importance”. (Online) Available on https://scikit-learn.org/stable/modules/permutation_importance.html (26 November 2020).

[22] Meet Rajdev and Kyumin Lee. (2015). Fake and Spam Messages : Detecting Misinformation during Natural Disasters on Social Media. Department of Computer Science Utah State University

[23] Arushi Gupta, Rishabh Kaushal. (2015). Improving Spam Detection in Online Social Networks.Department of Information Technology Indira Gandhi Delhi Technical University for Women Kashmere Gate, Delhi.

[24] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su and Jing Gao. (2018).Event Adversarial Neural Networks for Multi-Modal Fake News Detection. Department of Computer Science, State University of New York at Buffalo, New York, United States of America.

[25] Guoyong Cai, Hao Wu, Rui Lv. (2014). Rumors Detection in Chinese via Crowd Responses. Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology 541004 Guilin, P.R. China

[26] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu and Kam-Fai Wong. (2015). Detect rumors using time series of social context information on microblogging websites. Beijing University of Posts and Telecommunications, Beijing, China

[27] Sadia Afroz, Michael Brennan and Rachel Greenstadt. (2012). Detecting Hoaxes Frauds and Deception in Writing Style Online.Department of Computer Science Drexel University, Philadelphia, PA 19104

[28] Victoria L. Rubin, Niall J. Conroy, Yimin Chen and Sarah Cornwell. (2016). Fake news or truth using satirical cues to detect potentially misleading news. Language and Information Technology Research Lab (LIT.RL) Faculty of Information and Media Studies University of Western Ontario, London, Ontario, CANADA

[29] Zhiwei Jin, Juan Cao, Yu-Gang Jiang and Yongdong Zhang. (2014).News credibility evaluation on microblog with a hierarchical propagation model. Key

Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China

[30] Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen and Yajun Wang. (2013). Prominent features of rumor propagation in online social media. Korea Advanced Institute of Science and Technology, Republic of Korea

[31] The Grails Team. (2016). “The Grails Framework”. (Online) Available on http://docs.grails.org/3.3.11/guide/single.html (4 October 2018).

[32] Groovy Language. (2015). “Groovy Language Documentation”. (Online) Available on https://groovy-lang.org/single-page-documentation.html (4 October 2018).

[33] PyThaiNLP. (2018). “Thai natural language processing in Python”. (Online) Available on https://github.com/PyThaiNLP/pythainlp (4 October 2018).

[34] MySQL. (2014). “MySQL Documentation”. (Online) Available on https://dev.mysql. com/doc/ (1 October 2018).

[35] Scikit-learn developers. (2007). “User Guide”. (Online) Available on https://scikit-learn.org/stable/user_guide.html (26 November 2020).

ภาคผนวก

ภาคผนวก ก

ผลงานตีพิมพ์และเผยแพร่

การประชุมวิชาการวิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยี และสถาปัตยกรรมศาสตร์ ครั้งที่ 11

(ESTACON 2020)

ESTACON 11th
THE 11TH ENGINEERING, SCIENCE, TECHNOLOGY AND ARCHITECTURE CONFERENCE 2020

การประชุมวิชาการวิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยี และสถาปัตยกรรมศาสตร์ ครั้งที่ 11
(ESTACON 2020)

วันที่ 3 กรกฎาคม 2563

แจ้ง แจ้งผลการพิจารณาบทความวิจัย (ESTACON 2020)
เรียน คุณชวัล วัฒนากิจจากุล

ตามที่ท่านได้ส่งบทความเพื่อเข้าร่วมงานประชุมวิชาการ วิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยี และสถาปัตยกรรมศาสตร์ ครั้งที่ 11 ประจำปี 2563 (ESTACON 2020) ซึ่งจะจัดขึ้นในวันที่ 21 สิงหาคม 2563 ณ มหาวิทยาลัยราชภัฏนครราชสีมา ในหัวข้อเรื่อง **"(ST14) ตัวแบบการเรียนรู้ของเครื่องเพื่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์"**

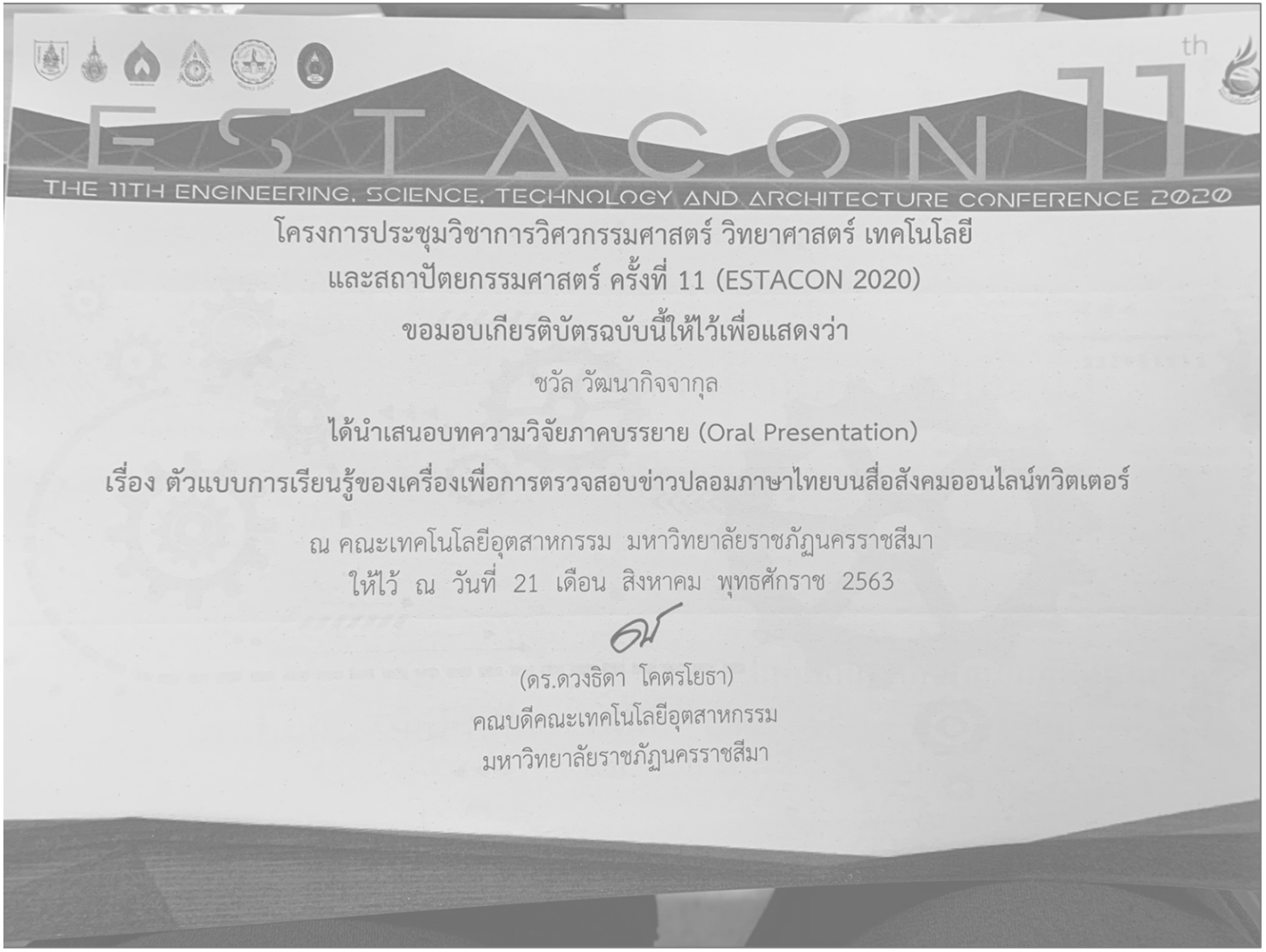
ในการนี้ คณะกรรมการดำเนินงานประชุมวิชาการ มีความยินดีที่จะเรียนให้ท่านทราบว่า บทความเรื่องดังกล่าว **"ผ่านการพิจารณา"** โดยผู้ทรงคุณวุฒิให้ทำการนำเสนอผลงานวิชาการในรูปแบบบรรยาย (Oral Presentation) ในการประชุมวิชาการวิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยี และสถาปัตยกรรมศาสตร์ ครั้งที่ 11 ประจำปี 2563 (ESTACON 2020) แล้ว

จึงเรียนมาเพื่อทราบ

ลงชื่อ

(ดร.ดวงธิดา โคตรโยธา)
ประธานการดำเนินงาน
การประชุมวิชาการ ESTACON 2020

คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏนครราชสีมา
โทร 044-255451
Email: estacon2020@nrru.ac.th


ESTACON 11th
THE 11TH ENGINEERING, SCIENCE, TECHNOLOGY AND ARCHITECTURE CONFERENCE 2020
โครงการประชุมวิชาการวิศวกรรมศาสตร์ วิทยาศาสตร์ เทคโนโลยี
และสถาปัตยกรรมศาสตร์ ครั้งที่ 11 (ESTACON 2020)
ขอมอบเกียรติบัตรฉบับนี้ให้ไว้เพื่อแสดงว่า
ชวัล วัฒนากิจจากุล
ได้นำเสนอบทความวิจัยภาคบรรยาย (Oral Presentation)
เรื่อง ตัวแบบการเรียนรู้ของเครื่องเพื่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์
ณ คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏนครราชสีมา
ให้ไว้ ณ วันที่ 21 เดือน สิงหาคม พุทธศักราช 2563
(ดร.ดวงธิดา โคตรโยธา)
คณบดีคณะเทคโนโลยีอุตสาหกรรม
มหาวิทยาลัยราชภัฏนครราชสีมา


ESTACON 11th
THE 11TH ENGINEERING, SCIENCE, TECHNOLOGY AND ARCHITECTURE CONFERENCE 2020
การประชุม
วิชาการ
วิศวกรรมศาสตร์
วิทยาศาสตร์ เทคโนโลยี
และสถาปัตยกรรมศาสตร์
ครั้งที่ 11
The 11th Engineering science Technology and Architecture Conference 2020
ก้าวสู่โลกอนาคตอย่างชาญฉลาด
ไปพร้อมกับความท้าทาย
เชิงสิ่งแวดล้อม
(Environmental Challenges and Smart Futures)
วันที่ 21 สิงหาคม 2563
คณะเทคโนโลยีอุตสาหกรรม มหาวิทยาลัยราชภัฏนครราชสีมา

ESTACON 11th
THE 11TH ENGINEERING, SCIENCE, TECHNOLOGY AND ARCHITECTURE CONFERENCE 2020

ST14

ตัวแบบการเรียนรู้ของเครื่องเพื่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

Machine Learning Models for Verifying Fake Thai News on Twitter

ชวัล วัฒนากิจจากุล[1,*] และ อนันท์ ชกสุริวงค์[2]

[1] ภาควิชาการจัดการเทคโนโลยีสารสนเทศ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ 15 ถ.กาญจนวณิชย์ อ.หาดใหญ่ จ.สงขลา 90110
[2] ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยสงขลานครินทร์ 15 ถ.กาญจนวณิชย์ อ.หาดใหญ่ จ.สงขลา 90110
*ผู้ติดต่อ: w.chawan@gmail.com, 086-748-3789

บทคัดย่อ

ปัจจุบันข่าวปลอมบนสื่อสังคมออนไลน์สามารถส่งผลกระทบรุนแรง เนื่องจากแพร่กระจายได้ง่ายและรวดเร็วกว่าข่าวจริง ซึ่งหากทำการตรวจสอบข่าวจำเป็นต้องใช้เวลานาน ส่งผลให้ข่าวปลอมแพร่กระจายเป็นวงกว้าง การระงับข่าวปลอมจึงทำได้ยาก ผู้วิจัยจึงมีแนวคิดที่จะค้นหาคุณลักษณะที่สำคัญของข่าวปลอม และเปรียบเทียบตัวแบบการเรียนรู้ของเครื่องระหว่างตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ ซัปพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม เพื่อค้นหาตัวแบบที่เหมาะสมกับการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ เพื่อเป็นแนวทางสำหรับต่อยอดในการสร้างระบบตรวจสอบข่าวปลอมอัตโนมัติต่อไป

โดยผลลัพธ์ของคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมบนสื่อสังคมออนไลน์ทวิตเตอร์ได้แก่ จำนวนครั้งของการโพสต์ข้อความตั้งแต่เริ่มสร้างบัญชีผู้ใช้ ช่วงเวลาการโพสต์ ค่าประเมินความรู้สึกของเนื้อหาข่าว จำนวนครั้งการกดชื่นชอบ และจำนวนครั้งในการแบ่งปันต่อ โดยตัวแบบที่เหมาะสมกับการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ คือ ตัวแบบเทคนิคโครงข่ายประสาทเทียมซึ่งได้ความถูกต้องสูงถึง 97%

คำสำคัญ: ข่าวปลอม ข่าวปลอมภาษาไทย การเรียนรู้ของเครื่อง ทวิตเตอร์ ตรวจสอบข่าว

Abstract

Nowadays, fake news on social media has caused many problems because they spread easier and faster than the real ones, while fake news detection or examination consumes high resources (human power, time, etc.). Thus, there is a need for an automatic method to examine or verify, so this research aims to find important features of fake Thai news and an appropriate machine learning model between Decision tree, Support Vector Machine and Neural Network model to examine the fake Thai news on Twitter for build automatic fake Thai news detection system in the future.

The evaluation results show that the significant features of fake Thai news are the amount of post since signing up, period time of the post, a sentiment of news content, the amount of news favorited, and the amount of news shared. The machine learning model that suits to examine the fake Thai news is a Neural Network model which performs 97 percent of accuracy.

Keywords: Fake news, Fake Thai news, Machine Learning, Twitter, Verify news

1. บทนำ

ปัจจุบันมีการใช้งานสื่อสังคมออนไลน์กันอย่างแพร่หลาย และได้ถูกใช้เป็นเครื่องมือในการกระจายข่าวปลอม ซึ่งสามารถส่งผลกระทบไปถึงความมั่นคงระดับประเทศได้ โดยผลการศึกษาเกี่ยวกับการแพร่กระจายข่าวปลอมบนสื่อสังคมออนไลน์ [1] บ่งบอกว่าข่าวปลอมถูกแบ่งปันมากกว่าข่าวจริงถึง 1.7 เท่า โดยมีคนรับข่าวปลอมมากกว่าถึง 100 เท่า และแพร่กระจายได้เร็วกกว่าถึง 6 เท่า

ผู้วิจัยจึงมีแนวคิดที่จะค้นหาคุณลักษณะที่สำคัญสำหรับการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ และทำการเปรียบเทียบตัวแบบการเรียนรู้ของเครื่อง (Machine Learning Model) [2] ในการตรวจสอบข่าวปลอมภาษาไทย โดยทำการเปรียบเทียบตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ [2] เทคนิคซัปพอร์ตเวกเตอร์แมชชีน [2] และเทคนิคเครื่องโครงข่ายประสาทเทียม [3-5] เพื่อค้นหาตัวแบบที่เหมาะสมในการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

2. วัตถุประสงค์

เพื่อเปรียบเทียบตัวแบบการเรียนรู้ของเครื่องที่เหมาะสมในการตรวจสอบข่าวปลอมภาษาไทย และค้นหาคุณลักษณะสำคัญต่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

3. ทฤษฎีและวรรณกรรมที่เกี่ยวข้อง

งานวิจัยฉบับนี้จะศึกษาในเรื่องของ คุณลักษณะของข่าวปลอมบนสื่อสังคมออนไลน์ การประมวลผลภาษาธรรมชาติ เทคนิคการเรียนรู้ของเครื่องที่นำมาใช้ การวัดค่าความถูกต้องของตัวแบบ และงานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดดังต่อไปนี้

3.1 คุณลักษณะของข่าวปลอมบนสื่อสังคมออนไลน์

สภาพแวดล้อมของการแยกแยะข่าวปลอม ได้มีงานวิจัย [6] ทำการศึกษาเรื่องการนิยามคุณลักษณะออกเป็น 3 ด้านดังต่อไปนี้

ด้านผู้กระจายข่าว (User Based) ซึ่งจะสนใจในสภาพแวดล้อมของผู้ใช้ที่กระจายข่าว ตัวอย่างเช่น จำนวนเพื่อน จำนวนผู้ติดตาม จำนวนการเผยแพร่ข้อความ ระยะเวลาที่สร้างบัญชีของผู้กระจายข่าว

ด้านเนื้อหาข่าว (Content Based) ซึ่งจะสนใจในรายละเอียดของเนื้อหาข่าว เช่น จำนวนแฮชแท็ก (Hash Tag) จำนวนการเชื่อมโยง หรือจำนวนคำ

ด้านการมีส่วนร่วม (Social Based) สนใจในท่าทีของผู้ใช้อื่นที่ได้รับข่าวสาร ตัวอย่างเช่น การแสดงความคิดเห็น การกดปุ่มชื่นชอบ การแบ่งปันต่อ

3.2 การประมวลผลภาษาธรรมชาติ

ในส่วนของเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Process) [2, 7] เป็นเทคนิคการประมวลภาษา เพื่อให้คอมพิวเตอร์ได้เข้าใจภาษาของมนุษย์ และสามารถตีความได้ในลักษณะของคะแนนความรู้สึก ซึ่งในงานวิจัยฉบับนี้ได้นำทฤษฎีส่วนนี้มาใช้ในการเพิ่มคุณลักษณะด้านการตีความด้านคะแนนความรู้สึกให้กับข้อมูล

3.3 การเรียนรู้ของเครื่อง

เป็นการทำให้เครื่องเรียนรู้ได้จากข้อมูลตัวอย่างที่เราสนใจ โดยความรู้ที่เรียนรู้ได้ เก็บไว้ในฐานความรู้ ซึ่งอยู่ในรูปแบบหลากหลาย เช่น การสร้างกฎ ฟังก์ชันความจำ โดยในงานวิจัยฉบับนี้ จะสนใจเทคนิคการเรียนรู้ต้นไม้ตัดสินใจ ซัปพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม มาใช้ในการสร้างตัวแบบเพื่อเปรียบเทียบความถูกต้อง โดยมีรายละเอียดดังนี้

เทคนิคต้นไม้ตัดสินใจ (Decision Tree) เป็นแบบจำลองทางคณิตศาสตร์ ในรูปแบบโครงสร้างต้นไม้ โดยโหนดแรกสุดจะเป็นรากของต้นไม้ (Root Node) แต่ละโหนดแสดงเงื่อนไขของคุณลักษณะ (Feature) ซึ่งเชื่อมต่อกันด้วยกิ่ง (Branch) โดยกิ่งที่เชื่อมจะแสดงผลของเงื่อนไขของคุณลักษณะนั้น และโหนดใบ (Leaf Node) จะแสดงคลาสที่กำหนดซึ่งเป็นผลการทำนาย

เทคนิคซัปพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เป็นเทคนิคในการคัดแยกกลุ่มเพื่อจัดประเภทโดยอาศัยระนาบการตัดสินใจที่เรียกว่าไฮเปอร์เพลน (Hyperplane) มาใช้ในการจำแนกกลุ่มข้อมูล ซึ่งไฮเปอร์เพลนที่เหมาะสมคือไฮเปอร์เพลนที่มีระยะห่างระหว่างขอบมากที่สุด (Maximum Margin)

เทคนิคโครงข่ายประสาทเทียม (Artificial Neural Network) มีรูปแบบโครงสร้างและการทำงานของการประมวลผลเหมือนกับสมองของสิ่งมีชีวิตซึ่งมีปรับเปลี่ยนตัวเองต่อการตอบสนองของอินพุตตามกฎของการเรียนรู้ โดยประกอบไปด้วย 3 ส่วนคือ ชั้นอินพุต ชั้นซ่อน และชั้นเอาต์พุต

3.4 การวัดค่าความถูกต้องของตัวแบบ

การวัดค่าความถูกต้องของตัวแบบ (Accuracy) สามารถทำได้โดยคำนวณจำนวนครั้งที่สามารถพยากรณ์ถูกต้องเทียบกับจำนวนครั้งทั้งหมดที่นำไปให้ตัวแบบทำการพยากรณ์ ดังสมการที่ (1)

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \quad (1)$$

สำหรับงานวิจัยฉบับนี้ผลลัพธ์ที่สนใจคือเป็นข่าวปลอมหรือไม่ซึ่งสามารถคำนวนค่าความถูกต้องของตัวแบบได้จากสมการที่ (2)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

ซึ่งค่า True Positive (TP) คือพยากรณ์ว่าเป็นข่าวปลอม และผลคือเป็นข่าวปลอม True Negative (TN) คือพยากรณ์ว่าเป็นข่าวจริง และผลคือเป็นข่าวจริง ค่า False Positive (FP) เมื่อพยากรณ์ว่าเป็นข่าวปลอม และผลที่ได้คือเป็นข่าวจริง ค่า False Negative (FN) เมื่อพยากรณ์ว่าเป็นข่าวจริง และผลคือเป็นข่าวปลอม

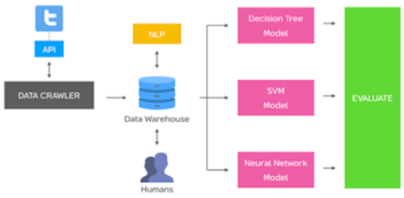
3.5 งานวิจัยที่เกี่ยวข้อง

จากการศึกษางานวิจัยที่เกี่ยวข้องพบว่ามีการนำเทคนิคการเรียนรู้ของเครื่อง มาช่วยในการแยกแยะข่าวปลอม ไม่ว่าจะเป็นผลงาน [8] ที่เป็นการนำเอาการประมวลผลภาษาธรรมชาติมาช่วยในการหาคะแนนความรู้สึกในเนื้อหา ร่วมกับคุณลักษณะอื่นของข่าว และใช้เทคนิคต้นไม้ตัดสินใจ เทคนิคซัปพอร์ตเวกเตอร์แมชชีน ในการตรวจสอบข่าวปลอม หรืองานวิจัย [9] ที่ได้นำเอาเทคนิคโครงข่ายประสาทเทียม แบบ Long Short Term Memory ซึ่งเป็นการประยุกต์ใช้โครงข่ายประสาทเทียมที่มีความรู้หรือความจำจากข้อมูลก่อนหน้ามาใช้ในการคำนวณร่วมกับโครงข่ายประสาทเทียมแบบคอนโวลูชัน ซึ่งเป็นโครงข่ายประสาทเทียมที่เพิ่มการคำนวณที่มีลักษณะแบบคอนโวลูชันเข้าไปในโครงข่ายประกอบด้วยชั้นคอนโวลูชันและชั้นพูลลิงในโครงข่ายเพื่อเปรียบเทียบตัวแบบที่เหมาะสมในการแยกแยะข่าวปลอม อีกทั้งงานวิจัยที่รวมเอาเทคนิคทางด้านการเรียนรู้ของเครื่องมาผสมผสานกันออกเป็นตัวแบบ [10] ที่ได้ทำการพัฒนาตัวแบบชื่อ CSI (Capture Score and Integrate) เพื่อใช้ในการสอนและทดสอบประสิทธิภาพการแยกแยะข่าวปลอมเปรียบเทียบกับเทคนิคอื่น

อย่างไรก็ตามจากการศึกษางานวิจัยข้างต้น พบว่ายังมีข้อจำกัดดังนี้ 1) การตรวจจับข่าวปลอมยังมีข้อจำกัดเรื่องเนื้อความ ที่สนใจเฉพาะข่าวที่เนื้อความเป็นภาษาอังกฤษ ทำให้การใช้กับข่าวภาษาไทยอาจจะไม่ได้ผลความถูกต้องมากเท่าที่ควร 2) การตรวจจับข่าวปลอมบางเทคนิคมีการใช้ข้อมูลการมีส่วนร่วมของผู้ใช้อื่น เช่นการแสดงความคิดเห็น การกดชื่นชอบ ซึ่งการมีส่วนร่วมของคนไทยอาจจะไม่เหมือนกับชาวต่างชาติ 3) ตัวแบบการแยกแยะ มีข้อจำกัดเรื่องข้อมูลที่เน้นไปทางใดทางหนึ่งเช่น พายุ ปลาบึก เป็นต้น ทำให้ตัวแบบอาจจะไม่ให้ผลความถูกต้องมากเท่าที่ควรในการตรวจสอบข่าวข้อมูลประเภทอื่น เช่น ข่าวการเมือง เป็นต้น

4. วิธีดำเนินการวิจัย

วิธีการดำเนินการวิจัย มีโครงสร้างดังรูปที่ 1



รูปที่ 1 ภาพรวมของการดำเนินการวิจัย

ซึ่งแบ่งส่วนการดำเนินการวิจัยออกเป็น 6 ส่วน โดยมีรายละเอียดดังต่อไปนี้

4.1 การเก็บรวบรวมข้อมูล

ผู้วิจัยได้สร้างโปรแกรมในการติดต่อกับสื่อสังคมออนไลน์ทวิตเตอร์ (Twitter) ผ่านทาง Application Programming Interface (API) สำหรับใช้ในการดึงข้อมูลข่าว โดยใช้ข้อมูลข่าวจำนวน 386 ชุดข้อมูลและแตกต่างประเภทของข่าว โดยประกอบด้วยข่าวจริงจำนวน 159 ข่าว (41%) และข่าวจริง 227 ข่าว (59%) จากผู้ใช้สื่อสังคมออนไลน์ทวิตเตอร์ และก่อนจะนำไปใช้เพื่อสอนและทดสอบตัวแบบ จะต้องให้ข้อมูลกำกับข่าวระบุเป็นข่าวปลอมหรือข่าวจริงก่อน โดยในงานวิจัยฉบับนี้ได้เลือกให้ผู้ที่มีวิชาชีพเกี่ยวกับด้านข่าวสาร ในการตรวจสอบข้อมูลโดยระบุข่าวที่เก็บรวบรวมมาว่าเป็นข่าวปลอมหรือไม่

4.2 การเพิ่มคุณลักษณะด้านความรู้สึกของข้อความ

หลังจากตรวจสอบข้อมูลโดยผู้ที่มีวิชาชีพทางด้านข่าวสาร จะเป็นขั้นตอนการใช้เทคนิคการประมวลผลภาษาธรรมชาติ ในการตีค่าความรู้สึกของเนื้อหาข่าว (Sentiment Analysis) จะสามารถระบุออกมาได้เป็น 3 ลักษณะ คือ เชิงบวก เป็นกลาง และเชิงลบ

4.3 การคำนวณเพื่อเพิ่มคุณลักษณะให้ข้อมูล

ผู้วิจัยได้ทำการคำนวณคุณลักษณะ ความยาวของชื่อผู้ใช้ ความยาวของชื่อ ชื่อผู้ใช้มีตัวเลขหรือไม่ จำนวนเครื่องหมายตกใจ/เครื่องหมายคำถาม จำนวนการเชื่อมโยง โพสต์เป็นวันใดของสัปดาห์ ช่วงเวลาการโพสต์ ความยาวของเนื้อหา จำนวนคำ จำนวนแฮชแท็ก อัตราส่วนเพื่อนและผู้ติดตาม โดยการสร้างโปรแกรมด้วยภาษา Python ในการคำนวณร่วมกับไลบรารี PyThaiNLP [11] ในการช่วยคำนวณเรื่อง คำและตัวอักษรภาษาไทย

4.4 การสร้างชุดข้อมูล

เมื่อรวบรวมข้อมูลและคำนวณข้อมูลตามหัวข้อข้างต้นแล้ว จะได้ข้อมูลที่จะนำไปใช้สอนและทดสอบตัวแบบดังแสดงในตารางที่ 1

ตารางที่ 1 คุณลักษณะที่นำมาใช้

ความยาวของชื่อผู้ใช้	ชื่อผู้ใช้มีตัวเลขหรือไม่
ความยาวของชื่อ	จำนวนเครื่องหมายตกใจ
จำนวนผู้ติดตาม	โพสต์เป็นวันใดของสัปดาห์
จำนวนการเชื่อมโยง	จำนวนการโพสต์ของผู้ใช้
ช่วงเวลาการโพสต์	ความยาวของเนื้อหา
จำนวนคำ	จำนวนเครื่องหมายคำถาม
จำนวนเพื่อน	จำนวนแฮชแท็ก
จำนวนการกดชื่นชอบ	จำนวนการแบ่งปัน
คะแนนความรู้สึก	อัตราส่วนเพื่อนและผู้ติดตาม

และจากนั้นผู้วิจัยได้ทำการแปลงข้อมูล (Data Preprocessing) เพื่อให้ข้อมูลอยู่ในมาตรฐานเดียวกัน (Standardized) ด้วยวิธีการ Standard Scaler โดยสามารถอธิบายการแปลงข้อมูลได้ดังสมการที่ (3)

$$x_{i(scaled)} = \frac{x_i - \mu}{\sigma} \qquad (3)$$

โดยที่

x_i คือข้อมูลอินพุต

μ คือค่าเฉลี่ยขอข้อมูลอินพุตทั้งหมด

σ คือค่าเบี่ยงเบนมาตรฐานของอินพุตทั้งหมด

4.5 การสร้างตัวแบบ และคำนวณค่าความถูกต้อง

ในขั้นตอนนี้ผู้วิจัยได้สร้างตัวแบบจำนวน 3 ตัวแบบ ที่ใช้เทคนิคต้นไม้ตัดสินใจ ซัปพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียมตามลำดับ และนำข้อมูลจากหัวข้อ 4.4 รวบรวมได้แบ่งเป็น 2 ส่วน คือส่วนที่ใช้สอนและส่วนที่ใช้ทดสอบตัวแบบ (อัตราส่วนของข่าวจริงต่อข่าวปลอมเป็น 3:2) โดยแบ่งเป็นอัตราส่วนข้อมูล 5 ชุดข้อมูลด้วยวิธีการสุ่ม ดังตารางที่ 2

ตารางที่ 2 ชุดข้อมูลที่ใช้สอนและทดสอบตัวแบบ

ชุดข้อมูล	จำนวนข้อมูลใช้สอน	จำนวนข้อมูลใช้ทดสอบ
ชุดที่ 1	50 %	50 %
ชุดที่ 2	60 %	40 %
ชุดที่ 3	70 %	30 %
ชุดที่ 4	80 %	20 %
ชุดที่ 5	90 %	10 %

ซึ่งจะได้ผลลัพธ์เป็นการระบุข่าวว่าเป็นข่าวจริงหรือข่าวปลอม และนำไปใช้เทียบกับผลลัพธ์จริงเพื่อคำนวณความถูกต้อง และจะทดสอบคุณลักษณะที่สำคัญ โดยการไม่ใช้คุณลักษณะนั้น ๆ ในการสอนและทดสอบ และสังเกตการเปลี่ยนแปลงของค่าความถูกต้อง โดยหากค่าความถูกต้องเปลี่ยนแปลงน้อยกว่า 3% จะถือว่าคุณลักษณะนั้นไม่สำคัญต่อการตรวจสอบข่าวปลอม เปรียบเทียบกับการค้นหาคุณลักษณะที่สำคัญด้วยวิธีการต้นไม้ตัดสินใจแบบ Extra Trees Classifier [5] ซึ่งเป็นอัลกอริทึมที่พัฒนาต่อยอดจากต้นไม้ตัดสินใจและใช้ในการค้นหาคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมภาษาไทย

4.6 การเปรียบเทียบค่าความถูกต้องของตัวแบบ

เป็นการนำค่าความถูกต้อง ของทั้ง 3 ตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ ซัปพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียมตามลำดับ ด้วยวิธีการตามหัวข้อ 3.4 มาทำการเปรียบเทียบตัวแบบ เพื่อทำการค้นหาตัวแบบที่เหมาะสมในการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์

5. ผลการวิจัย

ผลการวิจัยจะถูกแบ่งออกเป็น 2 ด้าน คือ ด้านความถูกต้อง และด้านคุณลักษณะของข่าวปลอม โดยมีรายละเอียดดังนี้

5.1 ด้านความถูกต้อง

ในการทดสอบตามหัวข้อวิธีการดำเนินการวิจัย ได้ผลลัพธ์เป็นค่าความถูกต้อง ของตัวแบบโครงข่ายประสาทเทียม ได้ความถูกต้องมากที่สุดถึง 97% โดยเป็นชุดข้อมูลจากข้อมูลชุดที่ 3 ดังแสดงในตารางที่ 3

ตารางที่ 3 ผลลัพธ์ความถูกต้องของตัวแบบ

ตัวแบบ	ชุดข้อมูล	ความถูกต้อง
ต้นไม้ตัดสินใจ	ชุดที่ 3	96 %
ซัปพอร์ตเวกเตอร์แมชชีน	ชุดที่ 3	95 %
โครงข่ายประสาทเทียม	ชุดที่ 3	97 %

5.2 ด้านคุณลักษณะของข่าวปลอม

จากการดำเนินการวิจัยเพื่อค้นหาคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอมภาษาไทย ตามหัวข้อที่ 4.5 พบว่าคุณลักษณะสำคัญของข่าวบนสื่อสังคมออนไลน์ทวิตเตอร์ที่สำคัญ และมีผลต่อการตรวจสอบข่าวปลอมได้แก่

ด้านผู้กระจายข่าว (User-Based) ได้แก่ จำนวนการโพสต์ตั้งแต่เริ่มสร้างบัญชีผู้ใช้

ด้านเนื้อหา (Content-Based) ได้แก่ ช่วงเวลาการโพสต์ และคะแนนความรู้สึกของเนื้อหาข่าว

ด้านการมีส่วนร่วม (Social Based) ได้แก่ จำนวนครั้งในการกดชื่นชอบ และจำนวนครั้งในการแบ่งปัน

6. การอภิปรายผล

จากปัญหาที่เกิดขึ้นจากข่าวปลอมที่แพร่กระจายได้อย่างรวดเร็ว และทำให้ผู้รับข่าวสารเข้าใจผิด ทำให้ส่งผลเสียมหาศาลที่ตามมาได้ในภายหลัง ผู้วิจัยจึงมีแนวคิดที่จะทำการค้นหาคุณลักษณะที่สำคัญต่อการตรวจสอบข่าวปลอม และค้นหาตัวแบบการเรียนรู้ของเครื่องที่เหมาะสมสำหรับการตรวจสอบข่าวปลอมที่เป็นภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์ โดยการเปรียบเทียบความถูกต้องของตัวแบบ ประกอบด้วยตัวแบบที่ใช้เทคนิคต้นไม้ตัดสินใจ ซัปพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม

ซึ่งผลการเปรียบเทียบตัวแบบการเรียนรู้ของเครื่องที่เหมาะสมสำหรับการตรวจสอบข่าวปลอมที่เป็นภาษาไทยบนสื่อสังคมทวิตเตอร์ คือตัวแบบที่ใช้เทคนิคโครงข่ายประสาทเทียมโดยให้ค่าความถูกต้องมากที่สุดถึง 97% โดยใช้ข้อมูลจากชุดข้อมูล 70 เปอร์เซ็นต์ในการสอนและ 30 เปอร์เซ็นต์ในการทดสอบ และคุณลักษณะที่สำคัญของการตรวจสอบข่าวปลอม ได้แก่ จำนวนการโพสต์ตั้งแต่เริ่มสร้างบัญชีผู้ใช้ ช่วงเวลาการโพสต์ของข่าว คะแนนความรู้สึกของเนื้อหาข่าว จำนวนครั้งการกดชื่นชอบ และจำนวนครั้งการแบ่งปันต่อ

งานวิจัยฉบับนี้สามารถต่อยอดได้โดยการนำไปใช้ในการประยุกต์เพื่อใช้สร้างระบบตรวจสอบข่าวปลอมในแบบทันเวลาจริง (Real-Time Verification) และสามารถเก็บข้อมูลข่าวจำนวนมาก เพื่อใช้สอน และทดสอบตัวแบบเพื่อให้ได้ความถูกต้องที่มากยิ่งขึ้น

7. เอกสารอ้างอิง

[1] Soriush Vosoughi and Deb Roy and Sinan Aral. (2018). The spread of true and false news online, *Science*. vol. 359, pp. 1146-1151.

[2] David Natingga. (2017). *Data Science Algorithms in a Week: Top 7 algorithms for computing*, Packt Publishing Ltd., Birmingham.

[3] Joel Grus. (2017). *Data Science from Scratch*, O' Reilly Media, Inc., United State of America.

[4] Nick McClure. (2017). *TensorFlow Machine Learning Cookbook Second Edition*, Packt Publishing Ltd., Birmingham.

[5] Aurelien Geron. (2017). *Hands-On Machine Learning with Scikit- Learn & TensorFlow*, O'Reilly Media, Inc., United State of America.

[6] Kai Shu, Amy Sliva, Jiliang Tang and Huan Liu. (2016). Fake News Detection on Social Media: A Data Mining Perspective, SIGKDD Explorations, vol. 19, pp. 22–36.

[7] Atefeh Farzindar and Diana Inkpen. (2015). *Natural Language Processing for Social Media: Second Edition (Synthesis Lectures on Human Language Technologies)*, Morgan & Claypool Publishers, United State of America.

[8] Saranya Krishnan and Min Chen. (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.

[9] Oluwaseun Ajao, Deepayan Bhowmik and Shahrzed Zargari. (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.

[10] Natali Ruchansky Sungyong Seo and Yan Liu. (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.

[11] Wannaphong Phatthiyaphaibun (2016). PyThaiNLP, URL: https://pythainlp.readthedocs.io, access on 27/11/2018.

ประวัติผู้เขียน

ชื่อ สกุล ชวัล วัฒนากิจจากุล

รหัสประจำตัวนักศึกษา 6110121001

วุฒิการศึกษา

วุฒิ	ชื่อสถาบัน	ปีที่สำเร็จการศึกษา
วิศวกรรมศาสตรบัณฑิต (วิศวกรรมคอมพิวเตอร์)	มหาวิทยาลัยสงขลานครินทร์	2553

ตำแหน่งและสถานที่ทำงาน

กรรมการผู้จัดการบริษัท ฟูบา ซิสเต็ม จำกัด อำเภอหาดใหญ่ จังหวัดสงขลา
กรรมการผู้จัดการบริษัท ซคูลเมท เอ็ดดูเคชัน จำกัด อำเภอหาดใหญ่ จังหวัดสงขลา
กรรมการผู้จัดการบริษัท บีแดร์ จำกัด อำเภอหาดใหญ่ จังหวัดสงขลา
กรรมการผู้จัดการบริษัท โลเร็มบอร์ด จำกัด อำเภอถลาง จังหวัดภูเก็ต

การตีพิมพ์เผยแพร่ผลงาน

ชวัล วัฒนากิจจากุล, อนันท์ ชกสุริวงค์. 2563. ตัวแบบการเรียนรู้ของเครื่องเพื่อการตรวจสอบข่าวปลอมภาษาไทยบนสื่อสังคมออนไลน์ทวิตเตอร์. การประชุมวิชาการ วิศวกรรมศาสตร์ เทคโนโลยีและสถาปัตยกรรมศาสตร์ ครั้งที่ 11, 1289-1295, นครราชสีมา, ประเทศไทย, 21 สิงหาคม 2563