การตรวจจับข่าวปลอมบนสื่อสังคมออนไลน์: กรณีศึกษาไวรัสโคโรน่า 2019

FAKE NEWS DETECTION ON SOCIAL MEDIA: CASE STUDY OF CORONAVIRUS 2019

นางสาวรัชนีวรรณ กอวิรัตน์

RUTCHANEEWAN KOWIRAT

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
พ.ศ. 2565

KMITL-2022-SC-M-017-105

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

FAKE NEWS DETECTION ON SOCIAL MEDIA: CASE STUDY OF CORONAVIRUS 2019

RUTCHANEEWAN KOWIRAT

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS

KMITL-DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

2022

KMITL-2022-SC-M-017-105

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2022
SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อการค้นคว้าอิสระ การตรวจจับข่าวปลอมบนสื่อสังคมออนไลน์: กรณีศึกษาไวรัสโคโรน่า 2019
ชื่อนักศึกษา นางสาวรัชนีวรรณ กอวิรัตน์
รหัสประจำตัว 63605082
ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์)
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง
พ.ศ. 2565
อาจารย์ที่ปรึกษาการค้นคว้าอิสระ รองศาสตราจารย์ ดร.ละออ บุญเกษม

บทคัดย่อ

ในปัจจุบันสื่อสังคมออนไลน์ (Social Media) เป็นช่องทางการติดตามข่าวสารช่องทางหนึ่งที่ผู้ใช้นิยมใช้กันอย่างแพร่หลาย เนื่องจากสามารถเข้าถึงข้อมูลได้ง่ายและสะดวกรวดเร็ว ซึ่งข้อดีในส่วนนี้นั้น ถูกผู้ไม่ประสงค์ดีนำไปใช้ในการแพร่กระจายข่าวปลอม (Fake News) เป็นวงกว้าง โดยที่ข่าวปลอมเริ่มเป็นปัญหาที่ส่งผลกระทบต่อสังคมมากยิ่งขึ้นหลังจากเกิดการแพร่ระบาดของโควิด-19 (COVID-19) ด้วยการทำให้ผู้คนเกิดความตื่นตระหนกและมีความรู้ความเข้าใจต่อการรักษาหรือป้องกันตนเองจากโรคแบบผิดวิธี ด้วยเหตุนี้วัตถุประสงค์ของงานวิจัยฉบับนี้จึงมีจุดมุ่งเน้นไปที่การตรวจจับข่าวปลอมที่เกี่ยวข้องกับโควิด-19 บนสื่อสังคมออนไลน์เพื่อช่วยกลั่นกรองข้อมูลด้วยการจำแนกข่าวจริงและข่าวปลอมด้วยการเรียนรู้ของเครื่อง (Machine Learning) และการเรียนรู้เชิงลึก (Deep Learning) อย่างโมเดลต้นไม้ตัดสินใจ (Decision Tree) โดยมีค่าความถูกต้องอยู่ที่ 99.92% และโมเดลโครงข่ายประสาทเทียมแบบวนซ้ำ (Recurrent Neural Network: RNN) โดยมีค่าความถูกต้องอยู่ 89.41% และ 96.52% จากชุดข้อมูล 2 ชุดที่นำเข้าสู่โมเดล

คำสำคัญ: การเรียนรู้เชิงลึก, การเรียนรู้ของเครื่อง, โควิด-19, ข่าวปลอม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Independent Study Title	Fake News Detection on Social Media: Case Study of Coronavirus 2019
Student Name	Miss Rutchaneewan Kowirat
Student ID	63605082
Degree	Master of Science (Data Science and Analytics) KMITL-Digital Analytics and Intelligence Center
Year	2022
Independent Study Advisor	Assoc. Prof. Dr. Laor Boongasame

Abstract

Social media has become one of the most popular channels to keep updated with daily news because it can quickly and easily access information. This advantage is used by malicious people to spread fake news widely. Since the COVID-19 pandemic, fake news has become a huge social problem, causing people to panic and misunderstand how to cure or protect themselves from the virus. So, the goal of this research is to use machine learning and deep learning to find COVID-19 fake news on social media, such as a decision tree with an accuracy of 99.92% and a recurrent neural network (RNN) model with the highest accuracy of 89.41% and 96.52% from the two datasets imported into the model.

Keywords: Fake News, COVID-19, Decision Tree, Deep Learning, Machine Learning, Recurrent Neural Network (RNN) model, Social media

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

งานวิจัยฉบับนี้ได้รับทุนอุดหนุนจากกองทุนวิจัยสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ทำให้งานวิจัยดำเนินได้ด้วยดี ซึ่งผู้วิจัยขอกราบขอบพระคุณเป็นอย่างสูง

ขอกราบขอบพระคุณ รศ.ดร.ละออ บุญเกษม อาจารย์ที่ปรึกษาที่ให้ความรู้ คำปรึกษา และกำลังใจ ทั้งยังคอยช่วยปรับปรุงแก้ไขงานวิจัยฉบับนี้ให้มีประสิทธิภาพยิ่งขึ้น ตลอดจนคอยให้ความช่วยเหลือในเรื่องต่าง ๆ ซึ่งผู้วิจัยรู้สึกซาบซึ้งในความอนุเคราะห์นี้อย่างมาก

ขอกราบขอบพระคุณ ดร.จิรภัทร์ หยกรัตนศักดิ์ และผศ.ดร.พรพิมล ชัยวุฒิศักดิ์ คณะกรรมการสอบที่สละเวลามาให้คำแนะนำและข้อเสนอแนะต่าง ๆ ซึ่งช่วยให้งานวิจัยฉบับนี้มีประสิทธิภาพมากยิ่งขึ้น

ขอกราบขอบพระคุณคณาจารย์ทุกท่านที่ให้ความรู้ คำปรึกษา จนก่อให้เกิดประสิทธิ์ประสาทวิชาความรู้และความเชี่ยวชาญ ทั้งทางด้านทฤษฎีและการปฏิบัติ

สุดท้ายนี้ขอกราบขอบพระคุณบิดา มารดา พี่น้อง และเพื่อนทุกคนที่คอยช่วยเหลือ ให้กำลังใจและสนับสนุนอยู่เบื้องหลังเสมอ จนทำให้งานวิจัยฉบับนี้สำเร็จไปได้ด้วยดี

นางสาวรัชนีวรรณ กอวิรัตน์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของงานวิจัย

ข่าวปลอม คือ ข่าวสารที่มีการบิดเบือนข้อเท็จจริง เพื่อหลอกลวงและชักจูงผู้รับสารให้เกิดความเข้าใจผิด ซึ่งเป็นปัญหาที่เกิดขึ้นในทุกยุคสมัย จึงต้องมีการควบคุมดูแลข่าวปลอมไม่ให้มีการแพร่กระจายไปในวงกว้างโดยกฎหมายควบคุมการเผยแพร่ข้อมูลอันเป็นเท็จ [4] ซึ่งในปัจจุบันข่าวปลอมเริ่มเป็นปัญหาที่ส่งผลกระทบต่อการสังคมมากขึ้น เมื่อผู้คนสามารถเข้าถึงข้อมูลข่าวสารต่าง ๆ ผ่านทางสื่อสังคมออนไลน์ อาทิเช่น เฟซบุ๊ก (Facebook), ทวิตเตอร์ (Twitter), กูเกิ้ล (Google) และเว็บไซต์อื่น ๆ ได้อย่างง่ายดาย โดยที่ทุกคนสามารถเป็นผู้สื่อสาร หรือเขียนเล่าเนื้อหา ประสบการณ์ จัดทำรูปภาพและวิดีโอขึ้นเอง หรือพบเจอสื่ออื่น ๆ ก่อนนำมาแบ่งปันให้กับผู้อื่นที่อยู่ในเครือข่ายของตนผ่านทางสื่อสังคมออนไลน์ได้ [3] ทำให้ข่าวสารสามารถแพร่กระจายเป็นวงกว้างได้อย่างรวดเร็ว โดยที่ผู้ส่งไม่อาจทราบเลยว่า ข่าวสารที่ตนเองได้ส่งออกไปถูกรับไปโดยผู้ใด และในทางฝั่งผู้รับข่าวสารเองก็ยากที่จะทราบได้ว่าเนื้อหาข้อมูลข่าวสารที่ได้รับมาเป็นความจริงหรือเป็นข้อมูลที่ผ่านการเปลี่ยนแปลงแก้ไขที่อาจถูกบิดเบือนข้อเท็จจริงไปก่อนหน้าหรือไม่ [4] นอกจากนี้ข่าวปลอมยังสามารถเกิดขึ้นได้กับข่าวสารทุกประเภท โดยข่าวสารและข้อมูลที่ได้รับการบิดเบือนเป็นจำนวนมากในปัจจุบันคือ ข่าวที่เกี่ยวข้องกับโรคติดต่อทางเดินหายใจซึ่งเกิดจากไวรัสโคโรนา (Coronavirus) ชนิดใหม่ที่เพิ่งถูกค้นพบและเกิดการระบาดในช่วงเดือนธันวาคมปี พ.ศ. 2562 จวบจนถึงปัจจุบัน ซึ่งโรคดังกล่าวได้ถูกเรียกว่า โควิด-19 [21]

ข่าวปลอมที่เกี่ยวข้องกับโควิด-19 มีการให้ข้อมูลที่ไม่เป็นจริงหลายอย่าง อาทิเช่น การป้องกันโรค, สถานที่เสี่ยงติดเชื้อ หรือการรักษา เป็นต้น จึงเป็นสิ่งที่สร้างความตื่นตระหนกให้แก่ผู้คนและก่อให้เกิดความรู้ในการป้องกันตัวเองจากโควิด-19 แบบผิด ๆ รวมถึงเกิดผลกระทบต่อธุรกิจในบางพื้นที่ เนื่องจากมีข่าวปลอมใส่ร้ายว่าเคยมีผู้ติดเชื้อเดินทางไปยังสถานที่ดังกล่าว ดังนั้นจึงควรมีวิธีการที่สามารถตรวจสอบข่าวสารต่าง ๆ บนสื่อสังคมออนไลน์ เพื่อลดการเผยแพร่ของข่าวปลอมและเสริมสร้างความรู้ความเข้าใจในการป้องกันตัวเองจากโควิด-19 ซึ่งวิธีดังกล่าวคือ การนำการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกมาประยุกต์ใช้ในการจำแนกข่าวจริงและข่าวปลอม

สำหรับงานวิจัยฉบับนี้นั้น ทางผู้วิจัยจะทำการดึงข้อมูลจากสื่อสังคมออนไลน์อย่างทวิตเตอร์โดยตรงผ่าน Twitter API และทำการจัดเตรียมข้อมูลให้เหมาะสมก่อนนำไปประยุกต์ใช้เข้ากับโมเดลต้นไม้ตัดสินใจ เพื่อทำการวิเคราะห์คุณลักษณะของข่าวปลอมด้วยการเรียนรู้ของเครื่อง นอกจากนี้ยังทำการตรวจจับข่าวปลอมที่เกี่ยวข้องกับโควิด-19 บนสื่อสังคมออนไลน์ด้วยการเรียนรู้เชิงลึก เพื่อเป็นการช่วยกลั่นกรองข้อมูลและลดการเผยแพร่ของข่าวปลอม สำหรับโมเดลที่ใช้งานจะมุ่งเน้นไปที่โมเดล RNN ด้วย RNN Cell ประเภทหน่วยความจำระยะสั้นแบบยาว 2 ทิศทาง (Bidirectional Long Short Term Memory: BiLSTM) และหน่วยเวียนกลับแบบมีประตู 2 ทิศทาง (Bidirectional Gated Recurrent Unit: BiGRU) ทั้งยังมีการประยุกต์เข้ากับ Word2Vec และมีการปรับค่าพารามิเตอร์ต่าง ๆ เพื่อค้นหาโมเดลที่มีประสิทธิภาพดีที่สุด โดยมุ่งหวังว่างานวิจัยฉบับนี้จะสามารถทำการจำแนกข่าวจริงและข่าวปลอมที่เกี่ยวข้องกับโควิด-19 ได้อย่างมีประสิทธิภาพ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.2 วัตถุประสงค์ของงานวิจัย

1) ศึกษาคุณลักษณะของข่าวจริงและข่าวปลอม
2) ศึกษาและสร้างโมเดลเพื่อใช้ในตรวจสอบข่าวปลอมที่เกี่ยวข้องกับโควิด-19

1.3 ขอบเขตของงานวิจัย

1) สามารถวิเคราะห์คุณลักษณะของข่าวจริงและข่าวปลอม
2) สามารถจำแนกข่าวจริงและข่าวปลอมที่เกี่ยวข้องกับโควิด-19

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ทราบคุณลักษณะของข่าวจริงและข่าวปลอม ทำให้สามารถหลีกเลี่ยงการเผยแพร่ข่าวปลอมได้
2) ได้เรียนรู้และเข้าใจกระบวนการเรียนรู้ของเครื่องและการเรียนรู้เชิงลึกมากขึ้น
3) ฝึกทักษะการเขียนภาษาไพธอน (Python)
4) ลดการเผยแพร่ของข่าวปลอมที่เกี่ยวข้องกับโควิด-19 และช่วยเสริมสร้างการรับสารจากสื่อสังคมออนไลน์ได้ถูกต้องมากยิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ เป็นการวิเคราะห์คุณลักษณะและทำการจำแนกข่าวจริงและข่าวปลอมที่เกี่ยวข้องกับโควิด-19 โดยผู้จัดทำได้ทำการศึกษาและรวบรวมข้อมูลที่เกี่ยวข้อง ตามหัวข้อดังต่อไปนี้

2.1 ข่าวปลอม

ข่าวปลอม คือ ข่าวสารที่มีการบิดเบือนข้อเท็จจริง ประกอบไปด้วยข้อมูลที่ไม่ถูกต้องหรือจงใจหลอกลวงและชักจูงผู้รับสารให้เกิดความเข้าใจผิดหรือสร้างความเสียหายแก่สังคมได้ โดยข่าวปลอมสามารถเผยแพร่ได้ทุกช่องทาง ไม่ว่าจะเป็นสื่อข่าวตีพิมพ์ การแพร่สัญญาณตามปกติ หรือสื่อสังคมออนไลน์

สื่อสังคมออนไลน์เป็นช่องทางกระจายข่าวปลอมที่นิยมในปัจจุบัน เมื่อผู้คนจำนวนมากเริ่มเปลี่ยนช่องทางการรับข้อมูลข่าวสารแบบดั้งเดิมมาเป็นติดตามผ่านแพลตฟอร์มออนไลน์ เนื่องจากมีความสะดวกรวดเร็วและสามารถเข้าถึงผู้คนได้ง่ายดายกว่า โดยหัวข้อข่าวของข่าวปลอมมักจะถูกเขียนขึ้นในรูปแบบที่ฉูดฉาด เพื่อใช้ในการล่อลวงและดึงดูดความสนใจของผู้รับสาร โดยผู้เขียนข่าวปลอมมีจุดประสงค์ได้หลายรูปแบบ [6] อาทิเช่น ทำกำไรจากการ Clickbait, โจมตีคู่แข่งทางธุรกิจ, ต้องการล้วงข้อมูลส่วนตัวหรือที่เป็นความลับผ่านมัลแวร์ (Malware) ตลอดจนทำไปเพียงเพื่อความสนุกสนาน

นอกจากนี้ข่าวปลอมยังส่งผลให้เกิดความเสียหายทางเศรษฐกิจ สังคม และทำให้เกิดความหวั่นวิตกต่าง ๆ หรืออาจร้ายแรงจนถึงขั้นก่อให้เกิดความสูญเสียของบุคคลที่ได้รับรู้ข่าวปลอม [4] อาทิเช่น การสร้างข่าวปลอมว่านายกรัฐมนตรีของไทยเป็น 1 ใน 45 คนที่เป็นเศรษฐีของทวีปเอเชียเพื่อหวังผลทางการเมือง, การจัดทำภาพประกอบที่ระบุข้อความเท็จว่ามีการประกาศใช้ พ.ร.บ. ข้าวแบบใหม่ ชาวนาจะต้องขายข้าวผ่านรัฐบาลเท่านั้น หากฝ่าฝืนจะมีโทษจำคุก 5 ปี และปรับเงิน 5 แสนบาท และการเผยแพร่ความรู้เรื่องการป้องกันโควิด-19 แบบผิด ๆ อย่างการกินยาฟ้าทะลายโจรในปริมาณมากจะช่วยรักษาโควิด-19 ทั้งที่จริงนั้นการบริโภคในปริมาณที่มากเกินพอดีจะทำให้ตับถูกทำลายได้

2.2 ทวิตเตอร์

ทวิตเตอร์เป็นสื่อสังคมออนไลน์รูปแบบไมโครบล็อกกิ้ง (Microblogging) [4] [3] ที่ใช้ในการรับ-ส่งข้อมูลแบบสั้น ๆ ที่เรียกว่า ทวิต (Tweet) โดยการทวิต 1 ครั้งสามารถพิมพ์ข้อความได้ไม่เกิน 280 ตัวอักษร [10] และในส่วนของการแชร์ข้อมูลข่าวสารจะถูกเรียกว่า รีทวิต (Retweet) ทำให้ผู้ใช้งานสามารถแชร์ทวิตของผู้ใช้งานคนอื่นที่ตนเองสนใจได้ นอกจากนี้ยังสามารถติดตามผู้คนที่เราสนใจได้ผ่านการ Follow โดยที่คนอื่น ๆ สามารถติดตามเรากลับได้เช่นกัน โดยเราจะเรียกผู้คนที่ติดตามเราว่า Follower ทั้งนี้ยังสามารถค้นหาข่าวสารหรือสร้างเทรนด์ใหม่ ๆ ได้อย่างง่ายดายผ่านระบบ Hashtag อีกด้วย

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในปัจจุบันทวิตเตอร์ได้รับความนิยมอย่างมากทั่วโลก แม้ผู้ใช้งานจะถูกจำกัดจำนวนตัวอักษรในการส่งข้อมูล ทว่ากลับมีอิสระในการเผยแพร่เนื้อหา ทวิตเตอร์จึงนับว่าเป็นแพลตฟอร์มที่สะดวก ทั้งยังสามารถส่งข้อมูลได้อย่างมีประสิทธิภาพ [4] ผู้คนจึงนิยมใช้ในการติดตามข่าวสารและเผยแพร่เรื่องราวหรือประสบการณ์ของตนเองไปพร้อมกัน โดยข้อดีของทวิตเตอร์คือความรวดเร็วในการส่งข่าว หากมีเหตุการณ์ฉุกเฉินที่ต้องการความช่วยเหลือแบบเร่งด่วน หรือต้องการติดตามข่าวใดข่าวหนึ่งอย่างกระชั้นชิด ทว่าความรวดเร็วนี้เองก็นับว่าเป็นข้อเสียเนื่องจากเป็นสิ่งที่ทำให้ข่าวปลอมสามารถแพร่กระจายได้เป็นอย่างดี ไม่ว่าผู้เขียนจะมีเจตนาสร้างข่าวปลอมหรือไม่ เพราะกว่าที่จะสามารถแก้ไข้ความใจผิดข่าวปลอมอาจจะแพร่กระจายไปไกลแล้วก็เป็นได้

2.3 ต้นไม้ตัดสินใจ [5]

ต้นไม้ตัดสินใจ เป็นกระบวนการเรียนรู้ของเครื่อง กระบวนการหนึ่งที่ช่วยในการตัดสินใจและวิเคราะห์แนวโน้มของข้อมูลผ่านรูปแบบของกราฟต้นไม้ โดยที่จะมีโหนดราก (Root Node) เป็นตัวตั้งต้น ก่อนที่จะแตกกระจายการตัดสินใจอื่น ๆ ลงมาจนสิ้นสุดที่ใบ (Leaf Node) หรือโหนดที่ไม่มีโหนดอื่นมาต่อท้ายนั่นเอง ซึ่งวิธีการนี้ช่วยทำให้สามารถแยกแยะลักษณะและรูปแบบของข้อมูลได้ง่ายขึ้น โดยขั้นตอนการสร้างต้นไม้ตัดสินใจ มีดังนี้

2.3.1 คำนวณหา Entropy โดยรวม

คำนวณหา Entropy โดยรวมของต้นไม้ตามสูตรที่ (1)

$$Entropy_{(Tree)} = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t) \quad (1)$$

2.3.2 คำนวณหา Entropy ของแต่ละคุณลักษณะ (Attribute)

คำนวณหา Entropy ของแต่ละคุณลักษณะ ตามสูตรที่ (1) จากนั้นทำการหาค่า Gain โดยการนำค่าที่ได้มาลบกับ Entropy โดยรวมของต้นไม้ ตามสูตรที่ (2)

$$Gain = Entropy_{(Tree)} - Entropy_{(\text{Attribute})} \quad (2)$$

2.3.3 ทำการเปรียบเทียบค่า Gain

เมื่อได้ค่า Gain ของแต่ละคุณลักษณะแล้ว ให้ทำการเปรียบเทียบและเลือกคุณลักษณะที่มีค่า Gain สูงที่สุดเป็นเกณฑ์ในการแบ่งกิ่ง

2.3.4 ทำการตัดคุณลักษณะและคำนวณซ้ำ

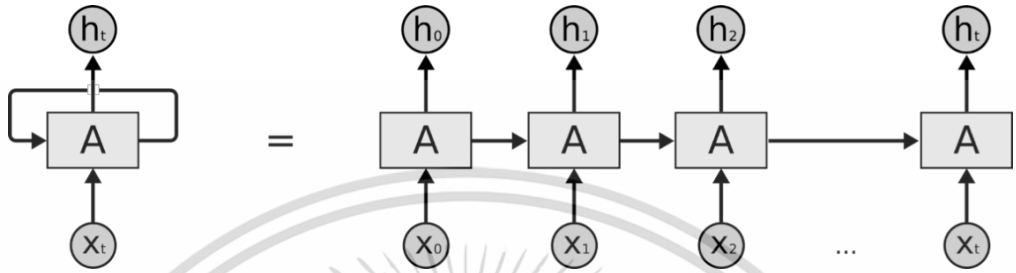
ทำการตัดคุณลักษณะที่เคยใช้เป็นเกณฑ์การแบ่งออกก่อนจะทำการคำนวณซ้ำอีกครั้งตั้งแต่ขั้นตอนแรกจนกว่าจะได้ต้นไม้ที่สมบูรณ์

2.4 โครงข่ายประสาทเทียมแบบวนซ้ำ [17, 20]

โครงข่ายประสาทเทียบแบบวนซ้ำ หรือ RNN คือ โครงข่ายประสาทเทียม (Artificial Neural Network) รูปแบบหนึ่งที่รับชุดข้อมูล (Data Set) ผ่านโหนดขาเข้า (Input Node) แบบ Time

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Series หรือข้อมูลที่มีลักษณะเป็นลำดับ (Sequence) เช่น วิดีโอ (Video) ที่เป็นลำดับของภาพ (Sequence of images) หรือข้อความ (Text) ที่เป็นลำดับของคำ (Sequence of words) ดังนั้น RNN จึงเป็นโมเดลที่เหมาะสำหรับการจำแนกกลุ่ม (Classification), การตัดคำ, Named Entity Recognition (NER) และ Sequence Tagging ตลอดจนข้อมูลที่จำเป็นต้องสนใจข้อมูลในอดีตเพื่อใช้ในการวิเคราะห์ เป็นต้น



รูปที่ 2.1 การทำงานของ RNN

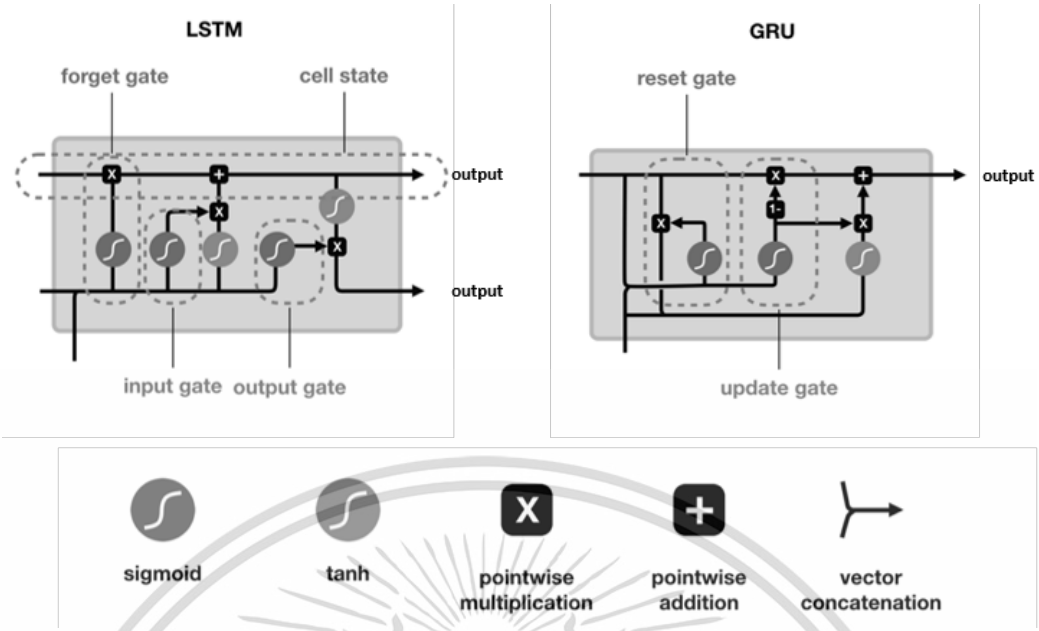
(ที่มา: https://colah.github.io/posts/2015-08-Understanding-LSTMs/)

สำหรับการทำงานของ RNN จากรูปที่ 2.1 เมื่อ A คือ RNN Cell, X(t) คือ ข้อมูลขาเข้า (Input data) ณ เวลา t และ h(t) คือ ข้อมูลขาออก (Output data) ณ เวลา t จะเห็นได้ว่าเมื่อข้อมูลขาเข้า เข้าสู่ RNN Cell นอกจากจะส่งข้อมูลขาออกมาในรูปแบบเวกเตอร์ (Vector) แล้ว ยังมีการวนข้อมูล Output state ของ RNN Cell รอบนั้น ๆ ให้เป็นข้อมูล Input state ในรอบถัดไปด้วย ซึ่งการทำงานนี้จะช่วยให้จดจำบริบทของคำในขณะการฝึกอบรม อาทิเช่น การระบุชื่อและนามสกุล หากคำก่อนหน้ามีแนวโน้มเป็นชื่อแล้ว คำถัดไปจะถูกระบุให้กลายเป็นนามสกุลแทน เป็นต้น

อย่างไรก็ตาม RNN นั้นมีจุดอ่อนอยู่ที่ขั้นตอนของการปรับค่าน้ำหนัก (Weight) โดยสามารถแบ่งปัญหาได้เป็น 2 ประเภท คือ Exploding gradient ที่เป็นกรณีที่ค่าน้ำหนัก ขยายใหญ่มากขึ้นเรื่อย ๆ เมื่อมีการปรับค่าน้ำหนักในแต่ละครั้ง และ Vanishing gradient ที่เป็นกรณีที่ค่าน้ำหนักมีขนาดเล็กลงเรื่อย ๆ จนเข้าใกล้ศูนย์และทำให้การปรับค่าน้ำหนักไม่มีผลอีกต่อไป สำหรับการแก้ปัญหาทั้งสองนั้น สามารถแก้ปัญหาได้โดยหลากหลายวิธี อาทิเช่น LSTM, GRU, Batch Normalization หรือการใช้ ReLU เป็น Activation function นั่นเอง

นอกจากนี้ภายใน RNN Cell ยังมีวงจรสำหรับนำข้อมูลขาเข้ามาถ่วงค่าน้ำหนักในรูปแบบต่างๆ ซึ่งวงจรที่ได้รับความนิยมในปัจจุบัน คือ LSTM และ GRU

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.2 วงจรแบบ LSTM และ GRU

(ที่มา: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21)

จากรูปที่ 2.2 จะเห็นได้ว่าวงจร LSTM จะประกอบไปด้วย Input Gate, Output Gate, Forget Gate, Cell State และมีขาส่งข้อมูลเวกเตอร์ออก 2 ขา ในขณะที่วงจร GRU จะประกอบไปด้วย Reset Gate, Update Gate และมีขาส่งข้อมูลเวกเตอร์ ออกเพียงขาเดียวเท่านั้น โดยที่ LSTM นั้นเหมาะที่จะใช้งานในรูปแบบที่ต้องการดูข้อมูลเป็นลำดับ ส่วน GRU นั้นเหมาะกับการใช้งานในรูปแบบการแยกองค์ประกอบของคำ นอกจากนี้ LSTM และ GRU ยังมีการทำงานแบบ BiLSTM และ BiGRU คือสามารถวิเคราะห์บริบทของข้อมูลได้ 2 ทิศทาง ไม่เพียงแต่นำข้อมูลในอดีตมามีส่วนช่วยในการวิเคราะห์เท่านั้น แต่ยังสามารถนำข้อมูลในอนาคตมามีส่วนช่วยในการตัดสินใจอีกด้วย ซึ่งเป็นการแก้ปัญหาในกรณีที่อาจจะมีการสูญเสียข้อมูลเมื่อมีการย้อนกลับไปปรับค่าน้ำหนัก หากมีการรับข้อมูลขาเข้าที่มีความยาวมาก ๆ ได้

2.5 หน่วยความจำระยะสั้นแบบยาว [2, 16]

หน่วยความจำระยะสั้นแบบยาว หรือ LSTM ถูกออกแบบมาเพื่อแก้ปัญหาของ RNN ในกรณีที่มีการปรับค่าน้ำหนักของข้อมูลที่มีความยาวมาก ๆ ซึ่งจากรูปที่ 2.2 จะเห็นได้ว่าภายในวงจรจะมี Sigmoid และ Tanh ในการคูณค่าเพื่อทำการปรับค่าน้ำหนัก เนื่องจาก Sigmoid จะทำการบีบให้ค่าน้ำหนักอยู่ในช่วง 0 ถึง 1 เท่านั้น ในขณะที่ Tanh จะทำการบีบให้ค่าอยู่ระหว่าง -1 ถึง 1 ซึ่งการที่ค่าน้ำหนักถูกบีบบังคับให้อยู่ในช่วงค่าใดค่าหนึ่งนี้เอง จึงทำให้ลดการเกิดปัญหา Exploding gradient และ Vanishing gradient

นอกจากนี้ Sigmoid ไม่เพียงแต่ช่วยบีบค่าน้ำหนักเท่านั้น แต่ยังช่วย Gate ต่าง ๆ ตัดสินใจว่าข้อมูลใดบ้างที่มีความสำคัญ ด้วยการพิจารณาจากค่าน้ำหนักของข้อมูล หากมีค่าเข้าใกล้ 1 มาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

แสดงว่าข้อมูลนั้นมีความสำคัญ ควรค่าที่ส่งข้อมูลเพื่อทำการอัปเดตต่อไป ในทางกลับกันหากมีค่าเข้าใกล้ 0 มาก ก็แสดงว่าข้อมูลนั้นไม่มีความสำคัญมากพอและควรจะทำการลบข้อมูลนั้นไป

สำหรับการทำงานของวงจร LSTM จากรูปที่ 2.2 จะเห็นได้ว่า Gate ที่ใช้ในการควบคุมการไหลของข้อมูลนั้น ประกอบไปด้วย Input Gate, Output Gate และ Forget Gate ซึ่งสำหรับขั้นตอนการทำงานจะเริ่มจากนำข้อมูลขาเข้าที่ประกอบไปด้วยข้อมูลจากปัจจุบันและข้อมูลที่ได้รับมาจาก Hidden State ก่อนหน้าเข้าสู่วงจร จากนั้นจะนำข้อมูลเข้าสู่ Forget Gate เพื่อทำการพิจารณาความสำคัญของข้อมูล ข้อมูลที่ไม่มีความสำคัญจะถูกลบ ส่วนข้อมูลที่มีความสำคัญจะถูกส่งไปยัง Input Gate ที่ทำหน้าที่ในการอัปเดตข้อมูลใน Cell State ซึ่งจะนำข้อมูลขาเข้ามาคูณด้วย Sigmoid และ Tanh ก่อนจะนำข้อมูลทั้ง 2 มาคูณเข้าด้วยกันอีกครั้ง เพื่อให้ Sigmoid ช่วยในการตัดสินใจว่าข้อมูลใดใน Tanh มีความสำคัญควรค่าแก่การนำไปอัปเดตใน Cell State ส่วน Gate สุดท้ายอย่าง Output Gate จะนำข้อมูลขาเข้ามาคูณด้วย Sigmoid และนำข้อมูลที่ได้จาก Cell State มาคูณด้วย Tanh แล้วจึงนำค่าที่ได้มาคูณเข้าด้วยกันอีกครั้ง เพื่อหาข้อมูลที่มีความสำคัญและทำการส่งข้อมูลนั้นออกไปเป็นข้อมูลขาออก ควบคู่กับการส่งข้อมูล Hidden State เพื่อใช้เป็นข้อมูลขาเข้าในรอบถัดไป ด้วยเหตุนี้ LSTM จึงเหมาะสมกับการที่ต้องการดูข้อมูลแบบลำดับ

2.6 หน่วยเวียนกลับแบบมีประตู [2, 16]

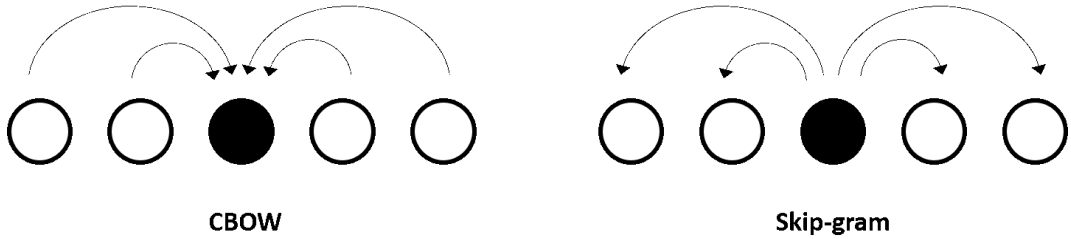
หน่วยเวียนกลับแบบมีประตู หรือ GRU มีการทำงานที่คล้ายคลึงกันกับ LSTM ทว่าจากการพิจารณารูปที่ 2.2 จะเห็นได้ว่าแตกต่างกันตรงที่ Gate จะประกอบไปด้วย Reset Gate และ Update Gate โดยการทำงานจะเริ่มจาก Reset Gate จะทำหน้าที่พิจารณาความสำคัญของข้อมูลว่าควรลบหรือเก็บเอาไว้ แล้วจึงไปยัง Update Gate เพื่อทำการพิจารณาว่าข้อมูลใดควรที่จะถูกส่งออกไปเป็นข้อมูลขาออก ด้วยจำนวน Gate ที่น้อยกว่า LSTM ทำให้การทำงานของ GRU มีความรวดเร็วกว่านั่นเอง

2.7 Word2Vec [19, 13]

Word2Vec เป็นโมเดลที่ใช้สร้าง Word Embedding ซึ่งจะทำการแปลงคำให้เป็นตัวเลขในรูปแบบของ Word Vector โดยที่ยังมีความสัมพันธ์กับคำรอบข้างที่อยู่ในประโยค สำหรับการสร้าง Word Embedding ของ Word2Vec จะมีวิธีการอยู่ 2 แบบ คือ Continuous Bag of Words (CBOW) และ Continuous Skip-gram (Skip-gram)

ดังรูปที่ 2.3 การทำงานของ CBOW จะทำการอ้างอิง Word Vector ของคำที่อยู่โดยรอบคำที่กำหนดมาใช้ในการทำนาย (Predict) หา Word Vector ในขณะที่ Skip-gram จะทำกำหนดคำขึ้นมาคำหนึ่งก่อน ให้เป็นคำแกนกลาง (Center word) เพื่อใช้คำนั้นในการทำนายคำที่อยู่รอบข้าง ทั้งนี้สามารถกำหนดความยาวของบริบทที่ใช้ได้ ด้วยการกำหนด n-Gram โดยจะทำการพิจารณาเฉพาะ n คำที่อยู่ด้านหน้าและด้านหลังคำแกนกลางเท่านั้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 2.3 การทำงานของ CBOW และ Skip-gram

ยกตัวอย่างดังรูปที่ 2.4 จะเห็นได้ว่าในส่วนของ CBOW เมื่อต้องการทำนายหา Word vector ของคำว่า color จะมีการอ้างอิงคำรอบข้าง ในขณะที่ Skip-gram จะทำการกำหนดให้ color เป็นคำแกนกลางก่อน แล้วจึงค่อยใช้ทำนายคำรอบข้าง ซึ่งจากตัวอย่างนั้นเป็นแบบ 2-Gram ซึ่งครอบคลุม 2 คำหน้าและหลังของคำแกนกลาง

รูปที่ 2.4 ตัวอย่างการทำงานของ CBOW และ Skip-gram

2.8 เมทริกซ์วัดประสิทธิภาพ (Confusion Matrix) [1, 4]

เมทริกซ์วัดประสิทธิภาพ เป็นการประเมินประสิทธิภาพของผลลัพธ์การทำนายเปรียบเทียบกับผลลัพธ์จริง โดยจะทำการแสดงผลเป็นตารางในรูปเมทริกซ์จัตุรัส ดังตารางที่ 2.1

ตารางที่ 2.1 ตัวอย่าง Confusion Matrix ขนาด 2x2

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

จากตารางที่ 2.1 มีรายละเอียดดังต่อไปนี้

1) True Positive (TP) คือ จำนวนข้อมูลที่ได้ทำการทำนายว่าเป็นจริงและผลลัพธ์จริงมีค่าเป็นจริง
2) False Negative (FN) คือ จำนวนข้อมูลที่ได้ทำการทำนายว่าเป็นเท็จแต่ผลลัพธ์จริงมีค่าเป็นจริง
3) True Negative (TN) คือ จำนวนข้อมูลที่ได้ทำการทำนายว่าเป็นเท็จและผลลัพธ์จริงมีค่าเป็นเท็จ
4) False Positive (FP) คือ จำนวนข้อมูลที่ได้ทำการทำนายว่าเป็นจริงแต่ผลลัพธ์จริงมีค่าเป็นเท็จ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

นอกจากนี้ยังสามารถนำค่าต่าง ๆ ที่ได้จากตารางที่ 2.1 มาคำนวณเพื่อใช้วัดประสิทธิภาพต่าง ๆ ของโมเดลได้ ดังนี้

ค่าความถูกต้อง (Accuracy) เป็นการวัดค่าความถูกต้องของแบบจำลองที่สามารถทำนายได้ถูกต้องตรงตามผลลัพธ์จริงทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (3)

$$Accuracy = \frac{True\ Positive + True\ Negative}{(True\ Positive + True\ Negative + False\ Positive + False\ Negative)} \quad (3)$$

Precision เป็นการวัดค่าความเที่ยงของข้อมูลที่ได้ทำนายว่าเป็นจริงมีความถูกต้องเพียงใด โดยสามารถคำนวณได้จากสมการที่ (4)

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} \quad (4)$$

Recall, True Positive Rate (TPR) หรือ Sensitivity เป็นการวัดความถูกต้องของแบบจำลองที่สามารถทำนายได้ถูกต้องเป็นอัตราส่วนของผลลัพธ์จริงที่มีค่าเป็นจริง โดยสามารถคำนวณได้จากสมการที่ (5)

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} \quad (5)$$

F-measure เป็นการวัดค่าความแม่นยำเฉลี่ยของแบบจำลอง โดยคำนวณเฉลี่ยจากค่า Precision และ Recall โดยสามารถคำนวณได้จากสมการที่ (6)

$$F - measure = 2 * \frac{(Precision * Recall)}{(Precision + Recall)} \quad (6)$$

True Negative Rate (TNR) หรือ Specificity เป็นการวัดค่าความแม่นของข้อมูลที่ได้ทำนายว่าเป็นเท็จจากคำตอบที่เป็นเท็จทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (7)

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Negative}{(True\ Negative + True\ Positive)} \quad (7)$$

False Positive Rate (FPR) เป็นการวัดค่าความแม่นของข้อมูลที่ได้ทำนายว่าเป็นจริงจากคำตอบที่เป็นเท็จทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (8)

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{(True\ Negative + False\ Positive)} \quad (8)$$

False Negative Rate (FNR) เป็นการวัดค่าความแม่นของข้อมูลที่ได้ทำนายว่าเป็นเท็จจากคำตอบที่เป็นจริงทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (9)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$False\ Negative\ Rate\ (FNR)\ =\ \frac{False\ Negative}{(True\ Positive\ +\ False\ Nagative)} \quad (9)$$

2.9 งานวิจัยที่เกี่ยวข้อง

งานวิจัยที่เกี่ยวข้องสามารถแบ่งได้เป็น 2 ส่วนคือ งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข่าวปลอมด้วยการเรียนรู้ของเครื่อง และงานวิจัยที่เกี่ยวข้องกับการตรวจจับข่าวปลอมด้วยการเรียนรู้เชิงลึก ดังนี้

2.9.1 งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข่าวปลอมด้วยการเรียนรู้ของเครื่อง

แบ่งเป็น 2 ประเภทตามลักษณะการดึงข้อมูลเพื่อใช้ในการวิเคราะห์ ดังนี้

1) ดึงข้อมูลโดยตรงจากสื่อสังคมออนไลน์: งานวิจัย [4] และ [6] เป็นงานวิจัยที่ทำการดึงข้อมูลที่ใช้ในการวิเคราะห์โดยตรงจากสื่อสังคมออนไลน์ โดยที่งานวิจัย [4] ทำการดึงข้อมูลจากทวิตเตอร์ ส่วนงานวิจัย [6] ดึงข้อมูลจากเฟซบุ๊ก, Forex และ Reddit

สำหรับขั้นตอนของการตรวจสอบข่าวปลอมนั้น ในส่วนของงานวิจัย [4] จะทำการแปลงข้อมูลที่อยู่ในรูปแบบไม่เป็นโครงสร้าง (Unstructured data) ให้เป็นข้อมูลที่อยู่ในรูปแบบเป็นโครงสร้าง (Structure data) พร้อมทำการปรับค่าข้อมูลแบบเข้ารหัส โดยใช้วิธีการ Ordinal และทำการจัดเก็บเป็นไฟล์ .csv ก่อนที่จะนำข้อมูลเข้าสู่การเรียนรู้ของเครื่องต่อไป ซึ่งหลังจากทำการเก็บข้อมูล งานวิจัยนี้ได้มีการวิเคราะห์ความสำคัญของคุณลักษณะ หากมีการพิจารณาตัดคุณลักษณะใดออกแล้วจะส่งผลต่อประสิทธิภาพของโมเดลในการวิเคราะห์ข่าวปลอมหรือไม่อีกด้วย และในส่วนงานวิจัย [6] จะทำการจัดอันดับให้คุณลักษณะ โดยใช้ Pearson's ในการวัดค่าความสัมพันธ์ เพื่อเพิ่มความแม่นยำและลดเวลาการ Training ข้อมูล โดยงานวิจัยทั้งสองมีการกำหนดให้ใช้ 10-fold cross validation สำหรับการทดสอบประสิทธิภาพของโมเดล

2) ดึงข้อมูลจากชุดข้อมูลสำเร็จ (Data Set): งานวิจัย [7], [8] และ [11] เป็นงานวิจัยที่ทำการดึงข้อมูลที่ใช้ในการวิเคราะห์โดยตรงจากชุดข้อมูลสำเร็จ โดยที่งานวิจัย [8] และ [11] มีการใช้ชุดข้อมูลสำเร็จที่ได้จากเว็บไซต์ Kaggle และจาก GitHub เพิ่มเติมด้วยในงานวิจัย [11] ในขณะที่งานวิจัย [7] ไม่ได้ทำการระบุแหล่งที่มา

สำหรับขั้นตอนของการตรวจสอบข่าวปลอมนั้น ในส่วนของงานวิจัย [8] จะทำการใช้ข้อมูล 2 ชุด ในการ Train โมเดลต่าง ๆ โดยแบ่งสัดส่วนเป็น 60:20:20 สำหรับการ Train, Validation และ Test ส่วนงานวิจัย [11] ทำการเตรียมข้อมูลด้วยวิธีการ Linguistic Inquiry and Word Count (LIWC) และประยุกต์ใช้อัลกอรึทึม (Algorithm) Self-Adaptive Harmony Search (SAHS) กับโมเดลต่าง ๆ และในส่วนของงานวิจัย [7] มีการกำหนดให้ใช้ 5-fold cross validation สำหรับการทดสอบประสิทธิภาพของโมเดล

2.9.2 งานวิจัยที่เกี่ยวข้องกับการตรวจจับข่าวปลอมด้วยการเรียนรู้เชิงลึก

งานวิจัย [12] มีการดึงข้อมูลจากทวิตเตอร์โดยตรง ผ่าน Twitter API โดยข้อมูลที่เป็นข่าวจริงจะทำการดึงมาจากแหล่งข่าวที่น่าเชื่อถือเท่านั้น ในขณะที่ข้อมูลข่าวปลอมจะทำการดึงข้อมูลจากคีย์เวิร์ด 'Fake news covid' โดยมีการใช้โมเดลต้นไม้ตัดสินใจ ในการทำนายแนวโน้มของข่าวปลอมว่ามีลักษณะแบบใด ดังนั้นข้อมูลที่ทำการดึงมาจะถูกกรองให้เหลือเพียงข้อมูลเชิงปริมาณ (Quantitative data) เท่านั้น นอกจากนี้ยังมีการ Normalization ข้อมูล ด้วย Min-Max

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Normalization เพื่อปรับค่าของข้อมูลให้มีช่วงไม่ต่างกันมากนักก่อนที่จะนำข้อมูลเข้าสู่โมเดลต่อไป ซึ่งจากผลการทดลองสามารถวัดค่าความถูกต้องได้ที่ 99.92%

งานวิจัย [18] มีการนำเข้าข้อมูลจากชุดข้อมูลสำเร็จจากหลายแหล่ง และทำการตรวจจับข่าวปลอมด้วยวิธีการเพอร์เซปตรอนชั้นเดียว (Single-Layer Perceptron: SLP), เพอร์เซปตรอนหลายชั้น (Multi-Layer Perceptron: MLP) และโครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolutional Neural Network: CNN) สำหรับการทดสอบโมเดลจะมีการแปลงข้อมูลให้เป็น Pooled Output ด้วย BERT และ RoBERTs โดยที่โมเดล CNN จะมีการแปลงข้อมูลให้เป็น Sequence Outputs เพิ่มเติมด้วย GPT2 และ Funnel Transformer นอกจากนี้ยังมีการประยุกต์ใช้ Gaussian Noise layer กับ CNN ป้องกันการเกิด Overfitting และใช้ Keras ในการตัดคำอีกด้วย

งานวิจัย [15] มีการดึงข้อมูลข่าวปลอมจากสื่อสังคมออนไลน์ต่าง ๆ อาทิเช่น ทวิตเตอร์ และเฟซบุ๊ก เป็นต้น ในขณะที่ชุดข้อมูลข่าวจริงจะทำการดึงข้อมูลจากบัญชีทวิตเตอร์ที่ได้รับการยืนยัน (Verified) จากทางทวิตเตอร์ เช่น WHO, CDC และ ICMR และในส่วนของการเตรียมข้อมูลนั้นมีการลบข้อมูลบางจำพวก เช่น ตัวเลข, ลิงก์ (Links) และคำที่ไม่ค่อยสื่อความหมายอย่าง Stop Words เป็นต้น นอกจากนี้ยังมีการใช้ TF-IDF ในการแสดงความถี่ของคำและทำการให้ค่าความสำคัญอีกด้วย

งานวิจัย [14] มีการนำเข้าข้อมูลจากชุดข้อมูลสำเร็จจากหลายแหล่ง มีการใช้ Keras ในการตัดคำ นอกจากนี้ยังมีการทำ Word Embedding ด้วย Word2Vec และ GloVe อีกด้วย

งานวิจัย [9] มีการดึงข้อมูลจากสื่อสังคมออนไลน์ต่าง ๆ เช่น ทวิตเตอร์, เฟซบุ๊ก และ Reddit ผ่านวิธีการ BERT และ GLoVe ก่อนจะทำการประยุกต์ใช้ News Embedding Block (NEB) และ Multi-Scale Feature Block (MSFB) เข้ากับโมเดลต่าง ๆ นอกจากนี้ยังมีการใช้ TF-IDF ในการแสดงความถี่ของคำและทำการให้ค่าความสำคัญอีกด้วย

งานวิจัย [19] มีการนำเข้าข้อมูลจากชุดข้อมูลสำเร็จจากหลายแหล่ง ประยุกต์ใช้เข้ากับโมเดล CNN, BiLSTM และ Residual Network (ResNet) และใช้ Word2Vec, GloVe, และ fastText ในการทำ Word Embedding ทั้งยังมีการใช้ Optimizer 7 วิธี คือ SGD, RMSprop, Adam, Adadelta, Adagrad, Adamax และ Nadam

งายวิจัย [8] มีการนำเข้าข้อมูลจาก 2 ชุดข้อมูลสำเร็จจาก Kaggle และประยุกต์ใช้เข้ากับโมเดล CNN, Vanilla RNN, LSTM และ BiLSTM และมีการใช้ Softmax ในการคาดคะเนคำตอบว่าเป็นข่าวปลอมหรือไม่ ทั้งนี้ยังมีการใช้ Optimizer 3 วิธี คือ RMSprop, Adam และ Adagrad

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3
วิธีการดำเนินงานวิจัย

งานวิจัยนี้ มีวิธีการดำเนินงานเป็น 2 ส่วน คือ การวิเคราะห์ข่าวปลอมด้วยการเรียนรู้ของเครื่อง และการวิเคราะห์ข่าวปลอมด้วยการเรียนรู้เชิงลึก

3.1 การวิเคราะห์ข่าวปลอมด้วยการเรียนรู้ของเครื่อง

ในส่วนนี้จะเป็นการวิเคราะห์คุณลักษณะและทำนายแนวโน้มของข่าวปลอมว่ามีรูปแบบหรือลักษณะอย่างไรด้วยการเรียนรู้ของเครื่อง ซึ่งจะมีวิธีการดำเนินงานดังนี้

3.1.1 การเก็บรวบรวมข้อมูล

งานวิจัยนี้ได้ทำการดึงข้อมูลมาจากสื่อสังคมออนไลน์อย่างทวิตเตอร์โดยตรง ผ่าน Twitter API ในช่วงวันที่ 9 กุมภาพันธ์ 2564 จนถึงวันที่ 16 กุมภาพันธ์ 2564 โดยในส่วนของข้อมูลที่เป็นข่าวจริงจำนวน 2,684 แถว จะทำการดึงมาจากแหล่งข่าวที่น่าเชื่อถือได้ และในส่วนของข่าวปลอมจำนวน 1,500 แถว จะทำการดึงจากคีย์เวิร์ด ‘Fake news covid’ โดยคุณลักษณะของข้อมูลที่ทำการเลือกได้แนวคิดจากงานวิจัย [4] ที่มีการดึงข้อมูลผ่านทางทวิตเตอร์เช่นเดียวกัน โดยข้อมูลสามารถแบ่งออกเป็น 2 ส่วนคือ ข้อมูลของผู้ใช้งาน และข้อมูลของข้อความที่ทำการทวิต ดังตารางที่ 3.1 และ 3.2 ตามลำดับ

ตารางที่ 3.1 ข้อมูลของผู้ใช้งาน

คุณลักษณะ	คำอธิบาย
1) User_name	ชื่อของผู้ใช้งาน
2) User_screen_name	ชื่อบัญชีของผู้ใช้งาน
3) User_description	รายละเอียดข้อมูลที่ผู้ใช้งานได้ทำการเขียนเอาไว้
4) User_followers_count	จำนวนผู้ติดตามของผู้ใช้งาน
5) User_friends_count	จำนวนที่ผู้ใช้งานทำการติดตาม
6) User_statuses_count	จำนวนครั้งที่ผู้ใช้งานเคยทำการทวิต หรือ รีทวิตทั้งหมด
7) User_favorites_count	จำนวนครั้งที่ผู้ใช้งานเคยกดถูกใจทั้งหมด
8) User_date	วันที่สร้างบัญชี

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 3.2 ข้อมูลของข้อความที่ทำการทวิต

คุณลักษณะ	คำอธิบาย
1) Tweet_text	ข้อความที่ทำการทวิต
2) Tweet_hashtag	Hashtag ที่ทำการติด
3) Tweet_hashtag_count	จำนวน Hashtag ที่ทำการติด
4) Tweet_language	ภาษาที่ใช้ทวิต
5) Tweet_retweet_count	จำนวนครั้งที่มีการรีทวิต
6) Tweet_favorite_count	จำนวนครั้งที่มีการกดถูกใจ
7) Tweet_date	วันที่ทำการทวิต

หลังจากทำการดึงข้อมูลทั้ง 2 ส่วนดังกล่าว ขั้นตอนต่อไปจะทำการตรวจสอบว่ามีข้อมูลที่ซ้ำซ้อนกันหรือมีข้อมูลที่เกิดการสูญหาย (Missing value) หรือไม่ หากมีให้ทำการลบข้อมูลส่วนที่เกิดการซ้ำหรือสูญหายออก เมื่อได้ข้อมูลที่สมบูรณ์แล้วให้ทำการรวบรวมข้อมูลทั้ง 2 ส่วนเข้าเป็นชุดเดียวกัน จากนั้นจึงทำการระบุว่าเป็นข่าวจริงหรือข่าวปลอมตามวิธีการที่ดึงมาจากทวิตเตอร์ เมื่อได้ชุดข้อมูลทั้งข่าวจริงและข่าวปลอมแล้ว ให้ทำการรวมชุดข้อมูลทั้งสองเข้าด้วยกัน และตรวจสอบความซ้ำซ้อนของข้อมูลอีกครั้ง แล้วจึงสามารถนำข้อมูลชุดนี้ไปใช้ในการเรียนรู้ของเครื่อง

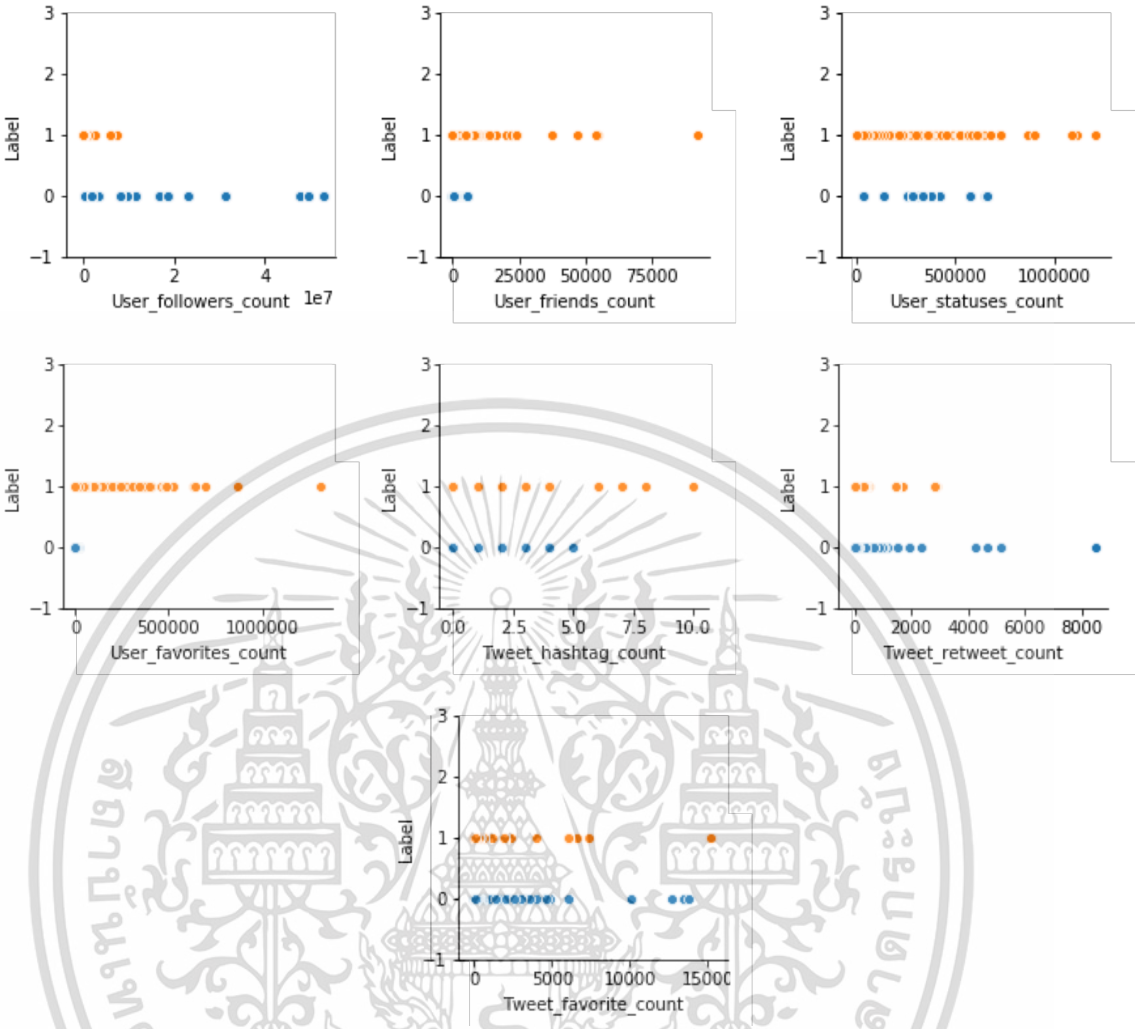
3.1.2 โมเดลการเรียนรู้ของเครื่องที่เลือกใช้

ในส่วนของโมเดลที่ใช้ในการตรวจสอบและวิเคราะห์ข่าวดังกล่าวเป็นข่าวปลอมหรือไม่ งานวิจัยนี้จะใช้โมเดลต้นไม้ตัดสินใจ ในการวัดประสิทธิภาพการวิเคราะห์ เนื่องจากตัวโมเดลต้นไม้ตัดสินใจ เป็นโมเดลที่รับแต่ข้อมูลเชิงปริมาณเท่านั้น ดังนั้นเราจึงจำเป็นต้องทำการกรองคุณลักษณะของข้อมูลให้เหลือเพียงข้อมูลเชิงปริมาณ ซึ่งคุณลักษณะของข้อมูลที่ได้ทำการเลือกมี 7 คุณลักษณะ คือ

1) User_followers_count
2) User_friends_count
3) User_statuses_count
4) User_favorites_count
5) Tweet_hashtag_count
6) Tweet_retweet_count
7) Tweet_favorite_count

หลังจากนั้นเราจะทำการนำคุณลักษณะของข้อมูลดังกล่าวไปใช้ในการวิเคราะห์แนวโน้มเบื้องต้นของข้อมูล โดยการสร้างเป็นแผนภาพเพื่อให้ง่ายต่อการวิเคราะห์และทำความเข้าใจ ดังรูปที่ 3.1 โดยที่เราสร้างและกำหนด Label ขึ้นมาใหม่เพื่อใช้ในการระบุว่าข่าวใดเป็นข่าวจริงหรือข่าวปลอม โดยที่ Label เท่ากับ 0 คือ ข่าวจริง ส่วน Label เท่ากับ 1 คือ ข่าวปลอม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 3.1 แผนภาพการวิเคราะห์แนวโน้มเบื้องต้นของข้อมูล

จากรูปที่ 3.1 จะเห็นได้ว่าแนวโน้มและการกระจายของข่าวจริงและข่าวปลอมแตกต่างกันอย่างเห็นได้ชัด เนื่องจากในส่วนของข้อมูลข่าวจริงเราได้ทำการรวบรวมมาจากแหล่งข่าวที่น่าเชื่อถือ ดังนั้นในการวิเคราะห์เราจะใช้คำว่า ‘แหล่งข่าวที่น่าเชื่อถือ’ แทนส่วนของผู้ใช้งานที่เผยแพร่ข่าวจริง ซึ่งสามารถทำการวิเคราะห์แนวโน้มเบื้องต้นของข้อมูลได้ดังต่อไปนี้

- User_followers_count: ยอดผู้ติดตามของแหล่งข่าวที่น่าเชื่อถือมีมากกว่าผู้ใช้งานที่เผยแพร่ข่าวปลอมอย่างเห็นได้ชัด เพราะโดยปกติแล้วแหล่งข่าวดังกล่าวมักจะมียอดผู้ติดตามสูงมากกว่าบุคคลทั่วไปนั่นเอง ยกตัวอย่างเช่น แอ็กเคานต์ (Account) ของ BBC Breaking News ที่ได้ทำการดึงข้อมูลมานั้นมียอดผู้ติดตามสูงถึง 47.5 ล้าน นับว่าเป็นยอดติดตามที่หาได้ยากแม้กระทั่งบุคคลหรือองค์กรที่มีชื่อเสียง
- User_friends_count: แหล่งข่าวที่น่าเชื่อถือจะมีการติดตามผู้อื่นน้อย เนื่องจากแหล่งข่าวเหล่านี้จะไม่ทำการติดตามบุคคลทั่วไปแต่มักจะติดตามแหล่งข่าวที่มาจากเครือเดียวกัน หรือแหล่งข่าวที่น่าเชื่อถือด้วยกันเองเท่านั้น แตกต่างจากส่วนของผู้ใช้งานที่เผยแพร่ข่าวปลอมที่มักเป็นบุคคลทั่วไป และมีการปฏิสัมพันธ์กับผู้ใช้งานคนอื่น ๆ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

มากกว่า ยกตัวอย่างเช่น แอ็กเคานต์ของ BBC Breaking News ได้ทำการติดตามเพียง 3 แอ็กเคานต์ ซึ่งเป็นแหล่งข่าวในเครือเดียวกันทั้งหมด คือ BBC News (World), BBC News (World) และ BBC Sport

- User_statuses_count: แหล่งข่าวที่น่าเชื่อถือจะมีความถี่ของการกระจายข้อมูลปานกลาง กล่าวคือไม่ถี่เกินไปหรือน้อยเกินไป เพราะแม้มีการทวิตเป็นประจำทุกวัน ทว่าจะมีการทวิตเป็นเป็นข่วงเวลาหรือตามแต่ข่าวด่วนในบางครั้งเท่านั้น ทั้งยังไม่นิยมการรีทวิตแหล่งข่าวหรือบุคคลอื่นใด ส่วนทางฝั่งผู้ใช้งานที่เผยแพร่ข่าวปลอมอาจเป็นเพราะมีจุดประสงค์ในการต้องการกระจายข่าวปลอมจึงมีการทวิตบ่อยครั้ง ดังนั้นความถี่ของการกระจายข้อมูลจึงพุ่งสูง ยกตัวอย่างเช่น แอ็กเคานต์ของ BBC Breaking News จากข้อมูลที่ทำการดึงมานั้น แม้จะทำการสร้างแอ็กเคานต์ตั้งแต่ปี 2007 ทว่ากลับมี User_statuses_count เพียงแค่ประมาณ 37,000 เท่านั้น
- User_favorites_count: คล้ายคลึงกันกับการวิเคราะห์ของ User_friends_count แหล่งข่าวที่น่าเชื่อถือไม่นิยมการกดถูกใจแหล่งข่าวหรือบุคคลอื่นใด ในขณะที่ผู้ใช้งานที่เผยแพร่ข่าวปลอมนั้นมีการปฏิสัมพันธ์กับผู้ใช้งานคนอื่น ๆ มากกว่าการกระจายของข้อมูลจึงค่อนข้างสูง ยกตัวอย่างเช่น แอ็กเคานต์ของ BBC Breaking News มี User_favorites_count เป็น 0 หรือก็คือไม่มีการกดถูกใจทวิตอื่นใดทั้งสิ้น
- Tweet_hashtag_count: โดยทั่วไปแหล่งข่าวที่น่าเชื่อถือจะทำการติดเฉพาะ Hashtag ที่เกี่ยวข้องกับเนื้อหาข่าวเท่านั้น ในขณะที่ฝั่งข่าวปลอมอาจจะมีการติด Hashtag อื่นที่ไม่เกี่ยวข้องเพื่อเป็นการกระจายข่าวปลอมอีกทางหนึ่ง ยกตัวอย่างเช่น แอ็กเคานต์ของ BBC Breaking News จะไม่นิยมติด Hashtag แต่จะทำการทวิตแค่หัวข้อและเนื้อหาข่าวเพียงเท่านั้น
- Tweet_retweet_count และ Tweet_favorite_count: เนื่องจากแหล่งข่าวที่น่าเชื่อถือมีความน่าเชื่อถือมากกว่าผู้ใช้งานที่เผยแพร่ข่าวปลอมที่มักเป็นบุคคลทั่วไป ดังนั้นยอดการรีทวิตและกดถูกใจจึงมีมากกว่า ทั้งนี้โดยพื้นฐานแล้วแหล่งข่าวที่น่าเชื่อถือมียอดผู้ติดตามที่สูง จึงมีโอกาสที่ผู้ใช้งานทวิตเตอร์จะเห็นมากกว่าอีกด้วย ยกตัวอย่างจากรูปที่ 3.1 จะพบว่าในส่วนของชุดข้อมูลข่าวจริง Tweet_retweet_count และ Tweet_favorite_count มีการกระจายตัวโดยรวมสูงและคงที่กว่าข่าวปลอม

เมื่อทำการวิเคราะห์แนวโน้มเบื้องต้นของข้อมูลเรียบร้อยแล้ว ขั้นตอนต่อไปจะทำการ Normalization ข้อมูล เพื่อปรับค่าของข้อมูลให้มีช่วงไม่ต่างกันมากนัก โดยวิธีการที่ใช้คือ Min-Max Normalization ที่มีค่าช่วงอยู่ 0 ถึง 1 หลังจากนั้นจะนำค่าที่รับแล้วมาทำการแบ่งเป็น Train 70% และ Test 30% ก่อนจะนำเข้าสู่โมเดลต่อไป โดยมีอัลกอริทึมในการตรวจสอบข่าวปลอมดังรูปที่ 3.2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Algorithm 1 Detect fake news

1: Input : Read data from CSV
2: Output : A Decision Tree of Detect fake news
3: Extract Variable data and Label (true or fake) from CSV file to create new table
4: Identify min and max in each column
5: For each (column) do

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

6: Set X = Xnew, Y = Label
7: Split Train 70% and Test 30%
8: Compute Decision tree
9: Return Decision Tree of Detect fake news

รูปที่ 3.2 อัลกอริทึมการตรวจสอบข่าวปลอม

3.2 การวิเคราะห์ข่าวปลอมด้วยการเรียนรู้เชิงลึก

ในส่วนนี้จะมุ่งเน้นไปที่การวิเคราะห์ข้อความด้วยการเรียนรู้เชิงลึก ซึ่งจะมีวิธีการดำเนินงานดังต่อไปนี้

3.2.1 ข้อมูลที่ใช้ในงานวิจัย

งานวิจัยฉบับนี้ใช้ชุดข้อมูลจากงานวิจัย [15] ที่มีขนาด 2,140 แถว โดยจะทำการเรียกชุดข้อมูลนี้ว่า Dataset-1 และใช้ชุดข้อมูลจากงานวิจัย [12] หรือข้อมูลที่ได้ทำการดึงจากทวิตเตอร์โดยตรงผ่าน Twitter API ซึ่งมีขนาด 4,184 แถว โดยจะทำการเรียกชุดข้อมูลนี้ว่า Dataset-2 ทว่าชุดข้อมูลนี้นั้นจะถูกนำมากรองข้อมูลด้วยโมเดลต้นไม้ตัดสินใจ ให้เหลือเพียงข้อมูลที่มีการทำนายถูกต้องเท่านั้น และตัดคุณลักษณะอื่น ๆ ออก ให้เหลือเพียง Tweet_Text และ Label โดยสามารถเขียนเป็นอัลกอริทึมได้ดังรูปที่ 3.3 โดยก่อนที่จะนำข้อมูลไปใช้ต่อไปนั้น จะทำการตัดข้อมูลที่มีการซ้ำซ้อนกันออกก่อน ซึ่งเมื่อผ่านการกรองข้อมูลด้วยวิธีดังกล่าวแล้ว ชุดข้อมูลจะมีขนาดเล็กลงเป็น 3,015 แถว

Algorithm 2 Filtering of data

1: Input : Read data from Decision Tree of Detect fake news
2: Output : A filtered data.
3: **For** each (column) **do**
    if (label == [prediction(label)]) **then**
        predict = “true”
    else
        predict = “false”
    end
end
4: **Return** Tweet Text and label columns with predict column equal true.

รูปที่ 3.3 อัลกอริทึมการกรองข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หลังจากทำการกรองข้อมูลแล้ว ขั้นตอนต่อไปจะทำการความสะอาดข้อความด้วยการแปลงข้อความให้เป็นตัวเล็กทั้งหมด พร้อมทั้งลบอักขระที่ไม่ใช้ตัวอักษร และข้อความที่ไม่มีผลต่อประสิทธิภาพโมเดล อาทิเช่น URL และชื่อบัญชี เป็นต้น ทั้งยังทำการกำจัดคำไม่สำคัญอย่าง Stop Words หรือคำที่แม้จะนำออกไปจากประโยคแล้วก็ยังสามารถเข้าใจความหมายของประโยคได้ รวมถึงทำการเปลี่ยนรูปคำให้อยู่ในรูปแบบดั้งเดิม (Lemmatization) เพื่อช่วยให้ข้อความมีขนาดและความซับซ้อนน้อยลง นอกจากนี้ในส่วนของชุดข้อมูล Dataset-2 ที่มี Label เป็นข่าวปลอมจะถูกทำการกรองว่า ‘Fake news’ ออก เนื่องจากในขั้นตอนการดึงข้อมูลได้ทำการดึงคำที่มีคีย์เวิร์ดดังกล่าวเพื่อให้ไม่เกิดความเป็นนิเสธของข้อมูล จึงทำการกรองให้เหลือแต่เนื้อหาที่เป็นข่าวปลอมเท่านั้น

ขั้นตอนต่อไปจะทำการสุ่มข้อมูลของ Label ให้มีขนาดเท่ากัน เพื่อเป็นการลดโอกาสที่จะเกิดการ Bias ของข้อมูลเพื่อนำข้อมูลเข้าสู่โมเดล โดยชุดข้อมูลจาก Dataset-1 สุ่มข้อมูลได้เป็น Fake news และ Real news อย่างละ 1,014 แถว ส่วนชุดข้อมูลที่ Dataset-2 นั้น สุ่มข้อมูลได้เป็น Fake news และ Real news อย่างละ 789 แถว ทั้งนี้จำนวนขนาดข้อมูลที่ทำการสุ่มจะอิงจากจำนวนขนาดของ Label ที่มีค่าน้อยกว่า

เมื่อ Label มีจำนวนเท่ากันแล้ว จะนำข้อความมาทำการสร้าง Keras Tokenizer Object เพื่อให้ได้ Bag of word ของจำนวนคำทั้งหมด แล้วจึงหาความยาวสูงสุดของคำในประโยคเพื่อนำมาทำ Padding พร้อมทำการแปลงแต่ละคำในประโยคเป็นตัวเลข และเติมเลขศูนย์เพิ่มเติมเพื่อให้ทุกประโยคมีความยาวเท่ากัน นอกจากนี้ในส่วนของ Label จะมีการเข้ารหัสข้อมูลแบบ One Hot เพื่อให้อยู่ในรูปของตัวเลขนั่นเอง

3.2.2 โมเดลการเรียนรู้เชิงลึกที่เลือกใช้

งานวิจัยฉบับนี้จะใช้โมเดล RNN ในการตรวจจับข่าวปลอม ทั้งนี้เพื่อลดการสูญเสียข้อมูลเมื่อมีการย้อนกลับไปปรับค่าน้ำหนัก จึงจะมีการใช้ RNN Cell ที่เป็นการเรียนรู้ข้อมูลแบบสองทิศทาง ดังนั้นงานวิจัยฉบับนี้จะใช้โมเดล RNN แบบ BiLSTM และ BiGRU เป็นโมเดลตั้งต้น พร้อมประยุกต์ทั้ง 2 โมเดลเข้ากับ Word2Vec เพื่อใช้ในการเปรียบเทียบและวัดประสิทธิภาพของโมเดลที่เหมาะสมในแต่ละชุดข้อมูล นอกจากนี้ก่อนนำข้อมูลเข้าสู่โมเดลจะทำการแบ่งข้อมูลออกเป็น Train 80% และ Test 20%

สำหรับการนิยามโมเดลนั้น จะมีการกำหนดโมเดลให้เป็นแบบ Sequential และมีการใช้ Optimizer แบบ Adam ส่วนการกำหนดชั้น Layer จะเรียงตามลำดับ ดังนี้

1) Input Layer: ชั้นสำหรับป้อนข้อมูลขาเข้า
2) Embedding Layer: ใช้สำหรับการฝังคำ
3) Bidirectional Layer: ในกรณีที่เป็นแบบ BiGRU นั้นจะใช้ Activation เป็น ReLU ในขณะที่แบบ BiLSTM จะใช้ค่าพารามิเตอร์ (Parameter) ที่ชื่อว่า merge_mode ที่เป็นการกำหนดวิธีการนำ Vector มาต่อกัน โดยแบบที่เลือกคือแบบ concat
4) Dense Layer: ทำการกำหนดมิติเป็นขนาด 128 และ Activation เป็น ReLU
5) Dropout Layer: ทำการสุ่มปิดโหนดบางตัวในระบบ เพื่อลดความซับซ้อนของข้อมูล
6) Dense Layer: ทำการลดขนาดมิติลง โดยกำหนดมิติเป็นขนาด 64
7) Dropout Layer: ทำการสุ่มปิดโหนดอีกครั้ง
8) BatchNormalization Layer: ทำการ Normalize ข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

9) Dense Layer (Output Layer): กำหนด Activation เป็น Softmax เพื่อให้ผลที่ได้ออกมาเป็นค่าความน่าจะเป็น

จากการนิยามโมเดลจะเห็นได้ว่านอกจากจะมีการใช้ Bidirectional Layer เพื่อหลีกเลี่ยงปัญหาการสูญเสียข้อมูลในระหว่างการปรับค่าน้ำหนัก ยังมีการใช้ BatchNormalization Layer และ Activation Function เป็น ReLU เพื่อแก้ปัญหา Exploding gradient และ Vanishing gradient อีกด้วย

นอกจากนี้ในกรณีที่มีการประยุกต์ร่วมกับ Word2Vec ซึ่งจะมีการกำหนดค่าพารามิเตอร์ในการสร้างโมเดล Word2Vec เป็น min_count = 1, dimension = 128, workers = 6, sg = 1 (Skip-gram) และ iter = 1000 ก่อนที่จะทำการสร้าง Embedding matrix แล้วจึงนำไปใส่ใน Embedding Layer ด้วยการกำหนดน้ำหนัก ให้เป็น Embedding matrix ที่ได้นั่นเอง

ทั้งนี้ในส่วนของการเทรนโมเดล (Train Model) จะมีการกำหนดโมเดลตั้งต้นดังตารางที่ 3.3 โดยเริ่มเทรนโมเดลที่ค่า Dropout = 0.10 หลังจากนั้นจะมีการปรับจูน Hyperparameters ต่าง ๆ เพื่อใช้ในการเปรียบเทียบและวิเคราะห์ประสิทธิภาพโมเดล ในกรณีที่มีการปรับเปลี่ยนค่าพารามิเตอร์ การเพิ่มจำนวนรอบของการเทรนโมเดล ตลอดจนการเปลี่ยนแปลงค่า Dropout จะส่งผลต่อประสิทธิภาพของโมเดลหรือไม่ หากส่งผลจะผลอย่างไร เพื่อค้นหาโมเดลที่ดีที่สุด

ตารางที่ 3.3 รายละเอียดโมเดลตั้งต้น

Hyperparameters	ค่าที่ใช้
Loss Function	Cross-Entropy
Optimizer	Adam
Learning Rate	0.001
Epochs	10
Batch Sizes	32

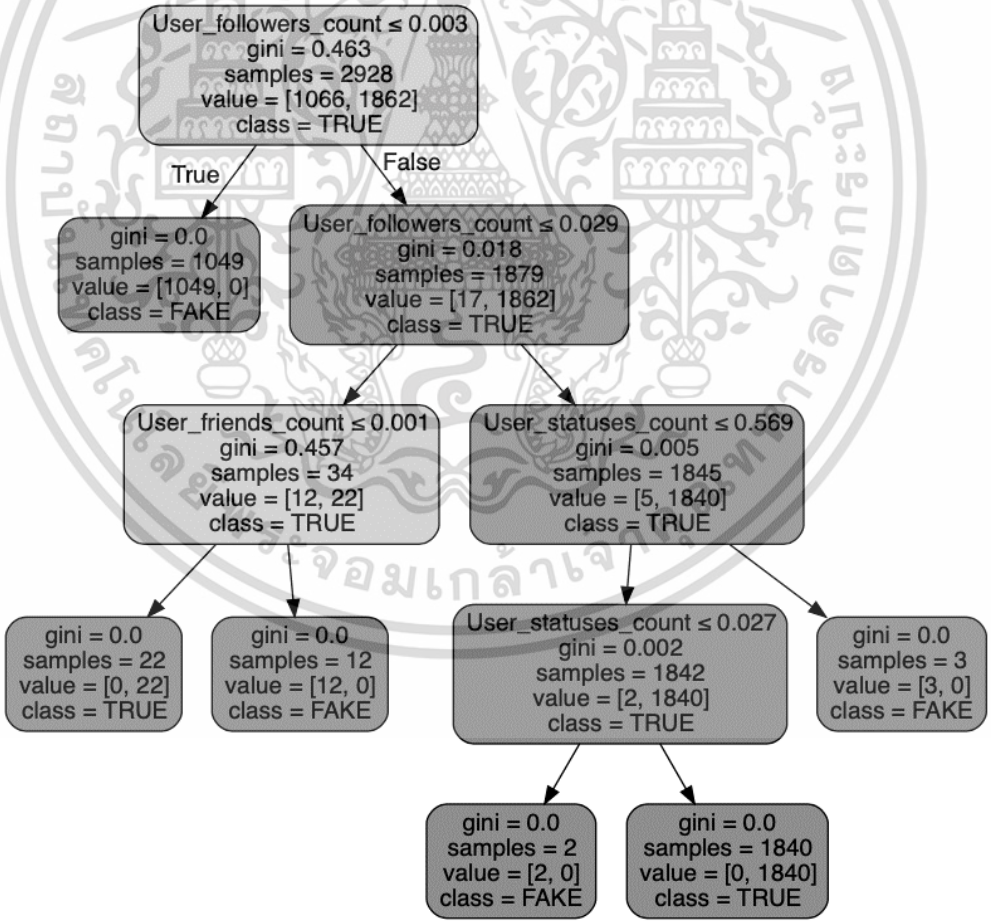
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิจัยและการอภิปรายผล

งานวิจัยนี้ ประกอบไปด้วยผลวิจัยและการอภิปรายผล 2 ส่วนด้วยกัน คือ ผลวิจัยจากการเรียนรู้ของเครื่อง และผลวิจัยจากการเรียนรู้เชิงลึก

4.1 ผลวิจัยจากการเรียนรู้ของเครื่อง

หลังจากนำข้อมูลเข้าสู่โมเดลต้นไม้ตัดสินใจ เราจะทำการสร้างแผนภาพต้นไม้ ดังรูปที่ 4.1 ขึ้นมาเพื่อให้ง่ายต่อการวิเคราะห์แนวโน้มของข่าวจริงและข่าวปลอม ซึ่งจะเห็นได้ว่าแนวโน้มของการตัดสินใจว่าข่าวใดเป็นข่าวจริงและข่าวใดเป็นข่าวปลอมจะอยู่ที่คุณลักษณะข้อมูลที่อยู่ในส่วนของผู้ใช้งานเป็นหลัก นอกจากนี้โมเดลต้นไม้ตัดสินใจ สามารถวัดค่าความถูกต้องได้ที่ 99.92%



รูปที่ 4.1 แผนภาพต้นไม้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 4.1 จะพบว่าข่าวจริงและข่าวปลอมมีความแตกต่างกันอย่างเห็นได้ชัด โดยที่ข่าวจริงจะมีแนวโน้มไปทางบัญชีทวิตเตอร์ที่มียอดผู้ติดตามเยอะ แต่มีการติดตามผู้อื่นน้อยและมีจำนวนครั้งในการทวิตหรือรีทวิตปานกลาง ในขณะที่บัญชีทวิตเตอร์ข่าวปลอมจะมียอดผู้ติดตามน้อย แต่ทำการติดตามผู้อื่นเยอะและมีจำนวนครั้งในการทวิตหรือรีทวิตน้อยหรือมากจนเกินไป นอกจากนี้จะเห็นได้ว่าแนวโน้มการตัดสินใจจะเอนเอียงไปที่ข้อมูลในส่วนของผู้ใช้งานมากกว่ารายละเอียดของข้อมูลที่เผยแพร่ สาเหตุหลักเกิดจากเราได้ทำการดึงข้อมูลข่าวจริงมาจากแหล่งข่าวที่น่าเชื่อถือ ซึ่งแหล่งข่าวดังกล่าวมักจะมีลักษณะที่เหมือนกัน ดังนี้

1) มียอดผู้ติดตามเยอะ เนื่องจากเป็นแหล่งข่าวที่น่าเชื่อถือทำให้มีผู้คนจำนวนมากทำการติดตามเพื่อรับข้อมูลข่าวสารที่เชื่อถือได้ ยกตัวอย่างจากข้อมูลที่ได้ทำการดึงมานั้น แหล่งข่าวที่น่าเชื่อถือที่มีผู้ติดตามน้อยที่สุด ยังมีจำนวนผู้ติดตามไม่ต่ำกว่า 170,000 ซึ่งเป็นจำนวนที่หาได้ยากในบุคคลทั่วไป
2) มีการติดตามผู้อื่นน้อย มักจะติดตามแหล่งข่าวที่มาจากเครือเดียวกัน หรือแหล่งข่าวที่น่าเชื่อถือด้วยกันเอง ยกตัวอย่างจากข้อมูลที่ได้ทำการดึงมานั้น แหล่งข่าวที่น่าเชื่อถือที่จะมียอดการติดตามเฉลี่ยเพียง 1,400 ในขณะที่แอ็กเคานต์ในส่วนของข่าวปลอมมียอดการติดตามเฉลี่ยสูงถึง 6,000 ถือว่ามีสัดส่วนค่าเฉลี่ยที่แตกต่างกันหลายเท่าตัว
3) จำนวนการทวิตปานกลาง เนื่องจากเป็นแหล่งข่าวจึงต้องมีการอัพเดทข่าวสารทุกวัน ทว่าจะมีการทวิตเป็นเวลาและไม่ถี่จนเกินไป ซึ่งจากการสุ่มสำรวจแอ็กเคานต์ที่ได้ทำการดึงข้อมูลมาจากทวิตเตอร์นั้น ค้นพบว่าแหล่งข่าวที่น่าเชื่อถือมักจะทำการทวิตข่าวประจำวัน หรือข่าวที่มีประเด็นน่าติดตามเท่านั้น และไม่นิยมที่จะรีทวิตของผู้อื่นมากนัก ดังนั้นยอดรวมจำนวนการทวิตจึงไม่สูงมากนักเมื่อเทียบกับวันที่เริ่มใช้งาน ในขณะที่แอ็กเคานต์ในส่วนของข่าวปลอมนั้น มักจะทำการทวิตเรื่องอื่น ๆ ด้วย และนิยมการรีทวิตผู้อื่น ทำให้มียอดรวมจำนวนการทวิตสูงกว่าเล็กน้อย

จะเห็นได้ว่าลักษณะของข่าวจริงที่ได้ทำการวิเคราะห์จากรูปที่ 4.1 นั้นตรงกับที่ได้ทำการวิเคราะห์แนวโน้มเบื้องต้นของข้อมูลที่ได้กล่าวไปในบทที่ 3 ถึง 3 ตัวแปร คือ User_followers_count, User_friends_count และ User_statuses_count ในขณะที่ตัวแปรอื่น ๆ ยังไม่มีข้อสรุปว่าตรงกับการวิเคราะห์หรือไม่ เนื่องจากตัวโมเดลต้นไม้ตัดสินใจ ไม่ได้นำตัวแปรอื่น ๆ ว่าเป็นเกณฑ์ในการแบ่งแผนภาพต้นไม้ ทั้งนี้อาจเป็นเพราะตัวแปรอื่น ๆ ยังไม่มีผลต่อการตัดสินใจที่มากพอ หรือไม่ก็เพียงใช้ตัวแปรทั้ง 3 เป็นเกณฑ์ก็อาจเพียงพอต่อการจำแนก เนื่องจากในชุดข้อมูลของข่าวจริงเราทำการดึงจากแหล่งข่าวที่น่าเชื่อถือทั้งหมด ทำให้สัดส่วนบางอย่างแตกต่างจาก ชุดข้อมูลของข่าวปลอมอย่างชัดเจน อาทิเช่น ยอดผู้ติดตาม เป็นต้น

4.2 ผลวิจัยจากการเรียนรู้เชิงลึก

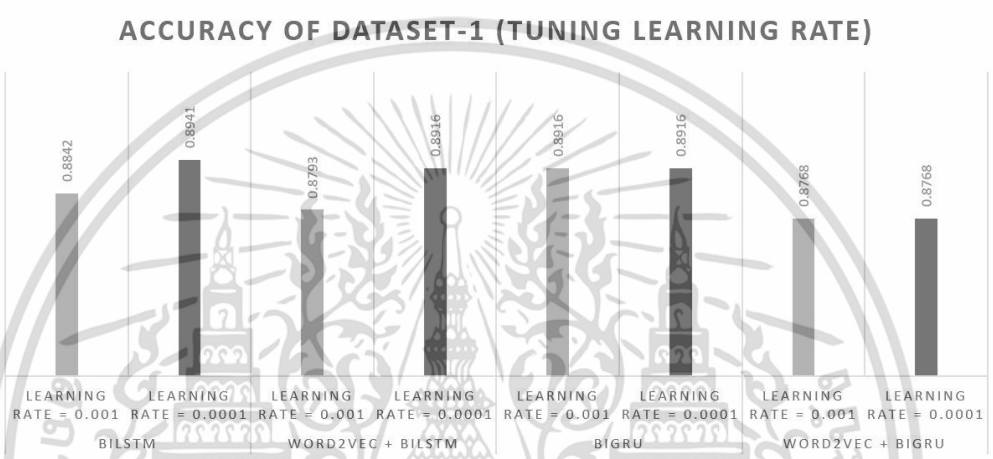
สำหรับการเทรนโมเดลแรกเริ่มทั้ง 2 ชุดข้อมูลจะทำการเทรนโมเดลที่มีรายละเอียดโมเดลตั้งต้นที่เหมือนกัน ดังที่แสดงในตารางที่ 3.3 ของบทที่ 3 โดยที่จะเริ่มต้นจากค่า Dropout = 0.1 djvo0t ทำการปรับจูนค่า Learning Rate เป็น 0.0001 เพิ่มขึ้นอีกหนึ่งโมเดล เพื่อค้นหาค่า Learning Rate ที่เหมาะสม ซึ่งหลังจากนำข้อมูลเข้าสู่โมเดลต่าง ๆ แล้ว จะทำการนำผลลัพธ์ค่าความสูญเสีย (Loss) และค่าความถูกต้องจากทั้ง 4 โมเดล คือ BiLSTM, BiGRU, Word2Vec + BiLSTM และ Word2Vec

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

+ BiGRU มาทำการวิเคราะห์และเปรียบเทียบเพื่อค้นโมเดลที่ดีที่สุดในแต่ละชุดข้อมูล โดยจะแบ่งผลวิจัยออกเป็น 2 ชุดตามแต่ชุดข้อมูล คือ ผลวิจัยของชุดข้อมูล Dataset-1 ที่เป็นชุดข้อมูลที่ได้จากงานวิจัย [15] และผลวิจัยของชุดข้อมูล Dataset-2 จากงานวิจัย [12] หรือชุดข้อมูลที่ได้ทำการดึงจากทวิตเตอร์โดยตรงผ่าน Twitter API ดังนี้

4.2.1 ผลวิจัยของชุดข้อมูล Dataset-1

สำหรับผลวิจัยของชุดข้อมูล Dataset-1 หลังจากนำข้อมูลเข้าโมเดลตั้งต้นพร้อมทั้งทำการปรับค่า Learning Rate เป็น 0.0001 ประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.2 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.1



รูปที่ 4.2 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Learning Rate ของ Dataset-1

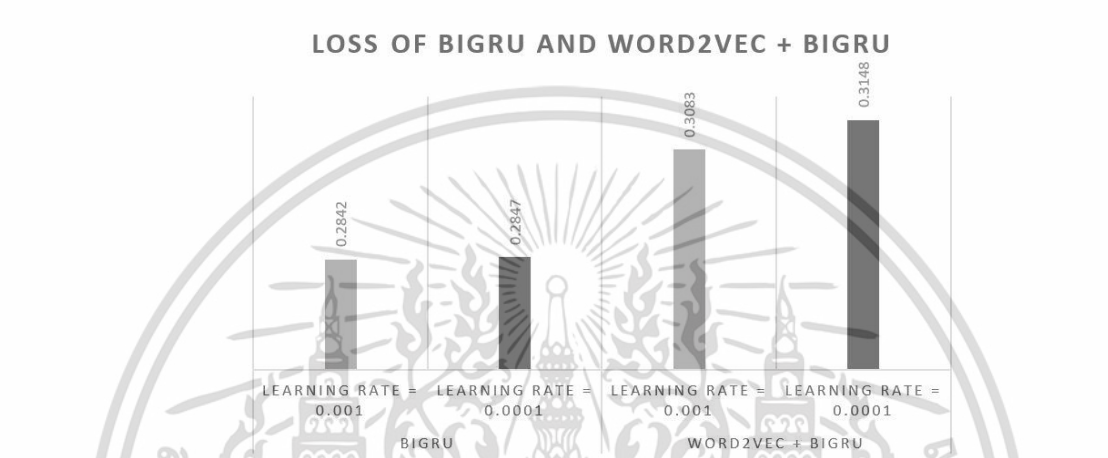
ตารางที่ 4.1 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Learning Rate ของ Dataset-1

Model	Learning Rate	Dataset-1	
		Loss	Accuracy
BiLSTM	0.001	0.3073	0.8842
	0.0001	**0.2579**	**0.8941**
Word2Vec + BiLSTM	0.001	0.2625	0.8793
	0.0001	**0.2681**	**0.8916**
BiGRU	0.001	**0.2842**	**0.8916**
	0.0001	0.2847	0.8916
Word2Vec + BiGRU	0.001	**0.3083**	**0.8768**
	0.0001	0.3148	0.8768

จากรูปที่ 4.2 ค้นพบว่า เมื่อมีการปรับค่า Learning Rate เป็น 0.0001 จะทำงานได้ดีบนโมเดล BiLSTM และ BiGRU เนื่องจากมีค่าความสูงต้องสูงกว่า ในขณะที่โมเดล BiGRU และ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

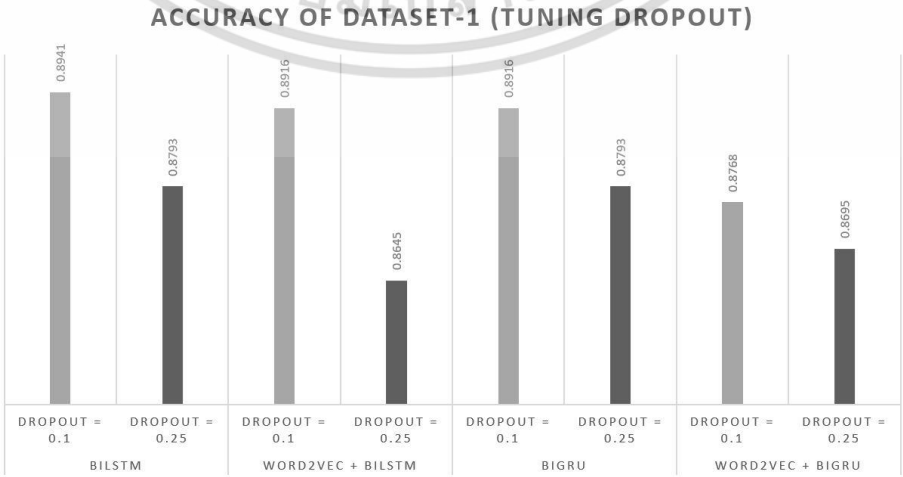
Word2Vec + BiGRU แม้ค่าความถูกต้องจะไม่เปลี่ยนแปลง เมื่อมีการเปลี่ยนค่า Learning Rate ทว่าเมื่อมีการพิจารณาที่ค่าความสูญเสีย ค้นพบว่าเมื่อค่า Learning Rate = 0.001 จะมีค่าความสูญเสียที่น้อยกว่า ดังนั้นจึงสรุปได้ว่าโมเดล BiGRU และ Word2Vec + BiGRU ทำงานได้ดีบนค่า Learning Rate = 0.001 เนื่องจากค่าความสูญเสียคือ ค่าที่ใช้ในการเปรียบเทียบผลลัพธ์ที่โมเดลทำนายมาได้กับค่าจริงว่ามีค่าความคลาดเคลื่อน (Error) เท่าใด ดังนั้นหากยิ่งมีค่าความสูญเสียน้อยยิ่งมีประสิทธิภาพ ทั้งนี้เพื่อให้ง่ายต่อการพิจารณาจึงนำค่าความสูญเสียจากตารางที่ 4.1 มาแสดงเป็นกราฟแท่งเพิ่มเติมในรูปที่ 4.3



รูปที่ 4.3 เปรียบเทียบค่าความสูญเสียของโมเดล BiLSTM และ BiGRU ของ Dataset-1

ทั้งนี้เมื่อทำการพิจารณาจากค่าความถูกต้องสูงสุดจากตารางที่ 4.1 ค้นพบว่าโมเดล BiLSTM ให้ค่าความถูกต้องสูงที่สุดเมื่อมีการใช้ค่า Learning Rate = 0.0001 ดังนั้นจึงถือว่าค่า Learning Rate นี้เหมาะสมกับชุดชุดข้อมูล Dataset-1 และใช้ค่านี้ในการปรับจูนพารามิเตอร์ครั้งต่อไป

หลังจากได้ค่า Learning Rate ที่เหมาะสมแล้ว ลำดับต่อไปจะทดลองปรับจูนค่า Dropout จากเดิมที่มีค่าเป็น 0.10 เป็น 0.25 เพื่อค้นหาว่าหากทำการลดความซับซ้อนของโมเดลลงด้วยการเพิ่มการสุ่มปิดโหนดจาก 10% เป็น 25% จะส่งผลต่อโมเดลอย่างไร โดยที่ประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.4 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.2



รูปที่ 4.4 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Dropout ของ Dataset-1

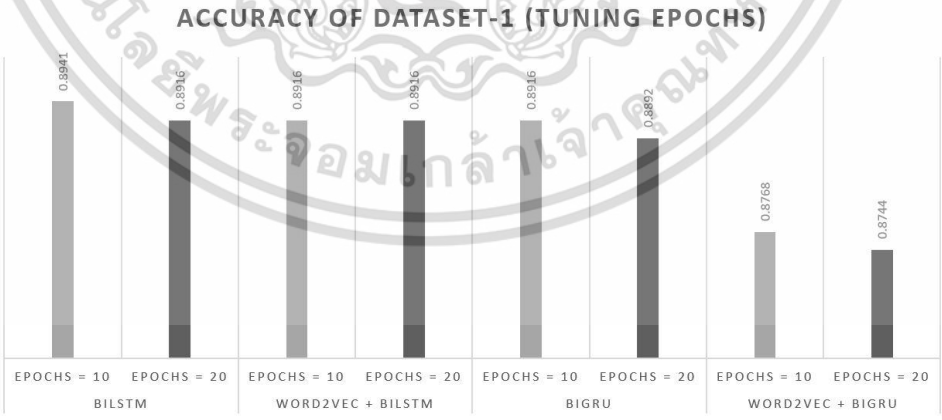
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.2 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Dropout ของ Dataset-1

Model	Dropout	Dataset-1	
		Loss	Accuracy
BiLSTM	0.10	**0.2579**	**0.8941**
	0.25	0.3026	0.8793
Word2Vec + BiLSTM	0.10	**0.2681**	**0.8916**
	0.25	0.2953	0.8645
BiGRU	0.10	**0.2847**	**0.8916**
	0.25	0.3236	0.8793
Word2Vec + BiGRU	0.10	**0.3148**	**0.8768**
	0.25	0.3280	0.8695

จากรูปที่ 4.4 ค้นพบว่าที่เมื่อทำการเพิ่มการปิดโหนด ผลลัพธ์ค่าความถูกต้องของโมเดลกลับลดลง ในกรณีนี้อาจเป็นเพราะชุดข้อมูล Dataset-1 มีความซับซ้อนน้อยอยู่แล้ว การสุ่มปิดโหนดเพิ่มจึงเป็นการลดประสิทธิภาพของโมเดลแทน ดังนั้นจึงสรุปได้ว่าในเบื้องต้นค่า Dropout = 0.10 เป็นค่าที่เหมาะสมชุดข้อมูล Dataset-1 และควรใช้ค่านี้ในการปรับจูนพารามิเตอร์ต่อไป

ขั้นตอนต่อไปจะเป็นการเพิ่มจำนวนรอบของการเทรนโมเดล โดยปรับจูนค่า Epochs เป็น 20 เพื่อวิเคราะห์ว่าจำรอบที่เหมาะสมในการเทรนโมเดลควรมีค่าเท่าใด โดยประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.5 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.3



รูปที่ 4.5 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Epochs ของ Dataset-1

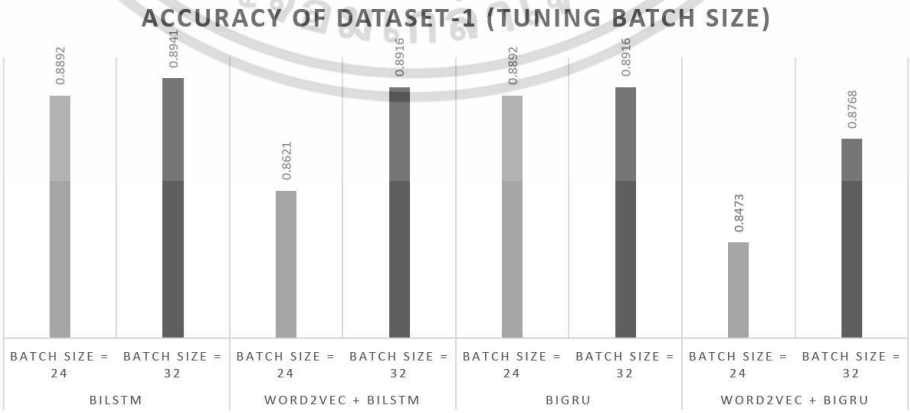
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.3 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Epochs ของ Dataset-1

Model	Epochs	Dataset-1	
		Loss	Accuracy
BiLSTM	10	**0.2579**	**0.8941**
	20	0.2526	0.8916
Word2Vec + BiLSTM	10	**0.2681**	**0.8916**
	20	0.2747	0.8916
BiGRU	10	**0.2847**	**0.8916**
	20	0.2823	0.8892
Word2Vec + BiGRU	10	**0.3148**	**0.8768**
	20	0.2992	0.8744

จากรูปที่ 4.5 ค้นพบว่าจำนวนที่เหมาะสมในการเทรนโมเดลอยู่ที่จำนวน 10 รอบ เนื่องจากมีค่าความถูกต้องที่สูงกว่า เมื่อเปรียบกับการเพิ่มจำนวนรอบเป็น 20 รอบ แม้ว่าโมเดล Word2Vec + BiLSTM จะมีค่าความถูกต้องเท่ากันแม้จะมีการเปลี่ยนแปลงจำนวนรอบการเทรนโมเดล ทว่าเมื่อพิจารณาจากค่าความสูญเสียที่มีค่าน้อยกว่า เมื่อทำการเทรนโมเดลที่ 10 รอบ ดังนั้นจึงทำการสรุปว่าค่า Epochs = 10 เป็นค่าที่เหมาะสมชุดข้อมูล Dataset-1 และควรใช้ค่านี้ในการปรับจูนพารามิเตอร์ต่อไป

สำหรับขั้นตอนสุดท้ายจะทดลองปรับจูนค่า Batch Size จาก 32 เป็น 24 ซึ่งเป็นการค้นหาว่าควรให้โมเดลมีการเรียนรู้ข้อมูลกี่ครั้งก่อนการปรับค่าน้ำหนัก จึงจะทำให้เกิดประสิทธิภาพที่ดีที่สุด โดยประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.6 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.4



รูปที่ 4.6 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Batch Size ของ Dataset-1

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

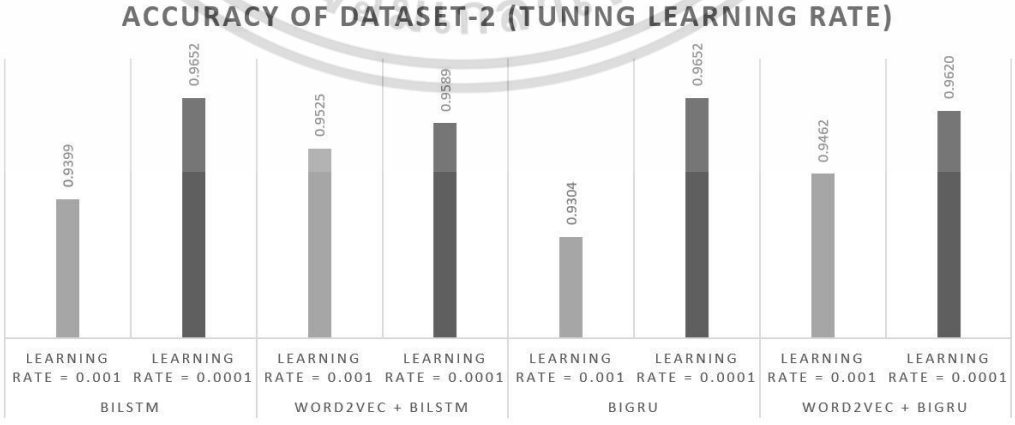
ตารางที่ 4.4 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Batch Size ของ Dataset-1

Model	Batch Size	Dataset-1	
		Loss	Accuracy
BiLSTM	24	0.2956	0.8892
	32	**0.2579**	**0.8941**
Word2Vec + BiLSTM	24	0.2778	0.8621
	32	**0.2681**	**0.8916**
BiGRU	24	0.2667	0.8892
	32	**0.2847**	**0.8916**
Word2Vec + BiGRU	24	0.3335	0.8473
	32	**0.3148**	**0.8768**

จากรูปที่ 4.6 จะเห็นได้ว่าค่า Batch Size ที่ 32 เป็นค่าที่เหมาะสมกับทุกโมเดล ทั้งนี้หากจะทำการเลือกโมเดลที่เหมาะสมกับชุดข้อมูล Dataset-1 จะเห็นได้ว่าโมเดล BiLSTM ให้ค่าความถูกต้องที่สูงที่สุดเมื่อเทียบกับโมเดลอื่น ๆ ดังนั้นจึงสรุปได้ว่าโมเดลที่เหมาะสมสำหรับชุดข้อมูล Dataset-1 คือ โมเดล BiLSTM ที่ค่า Learning Rate = 0.0001, Dropout = 0.10, Epochs = 10 และ Batch Size = 32 โดยมีค่าความถูกต้องอยู่ที่ 89.41%

4.2.2 ผลวิจัยของชุดข้อมูล Dataset-2

สำหรับผลวิจัยของชุดข้อมูล Dataset-2 หลังจากนำข้อมูลเข้าโมเดลตั้งต้นพร้อมทั้งทำการปรับค่า Learning Rate เป็น 0.0001 ประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.7 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.5



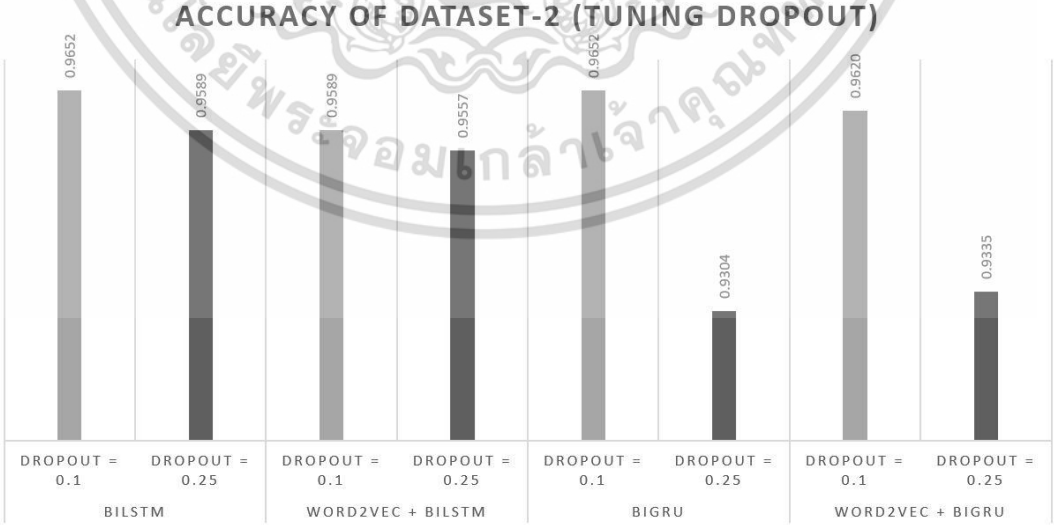
รูปที่ 4.7 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Learning Rate ของ Dataset-2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.5 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Learning Rate ของ Dataset-2

Model	Learning Rate	Dataset-2	
		Loss	Accuracy
BiLSTM	0.001	0.2140	0.9399
	0.0001	**0.1552**	**0.9652**
Word2Vec + BiLSTM	0.001	0.1689	0.9525
	0.0001	**0.1153**	**0.9589**
BiGRU	0.001	0.1788	0.9304
	0.0001	**0.1291**	**0.9652**
Word2Vec + BiGRU	0.001	0.1749	0.9462
	0.0001	**0.1519**	**0.9620**

จากรูปที่ 4.7 หลังพิจารณาจากค่าความถูกต้อง ค้นพบว่าทุกโมเดลทำงานได้ดีกว่า เมื่อใช้ค่า Learning Rate ที่ 0.0001 ดังนั้นจึงถือว่าค่า Learning Rate นี้เหมาะสมกับชุดชุดข้อมูล Dataset-2 และใช้ค่านี้ในการปรับจูนพารามิเตอร์ครั้งต่อไป ซึ่งหลังจากได้ค่า Learning Rate ที่เหมาะสมแล้ว ลำดับต่อไปจะเป็นเช่นเดียวกับกับผลวิจัยชุดข้อมูล Dataset-1 คือการทดลองปรับจูนค่า Dropout จาก 0.10 เป็น 0.25 เพื่อพิจารณาประสิทธิภาพที่ได้หลังจากการเพิ่มการสุ่มปิดโหนดจาก 10% เป็น 25% จะส่งผลต่อโมเดลอย่างไร โดยที่ประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.8 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.6



รูปที่ 4.8 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Dropout ของ Dataset-2

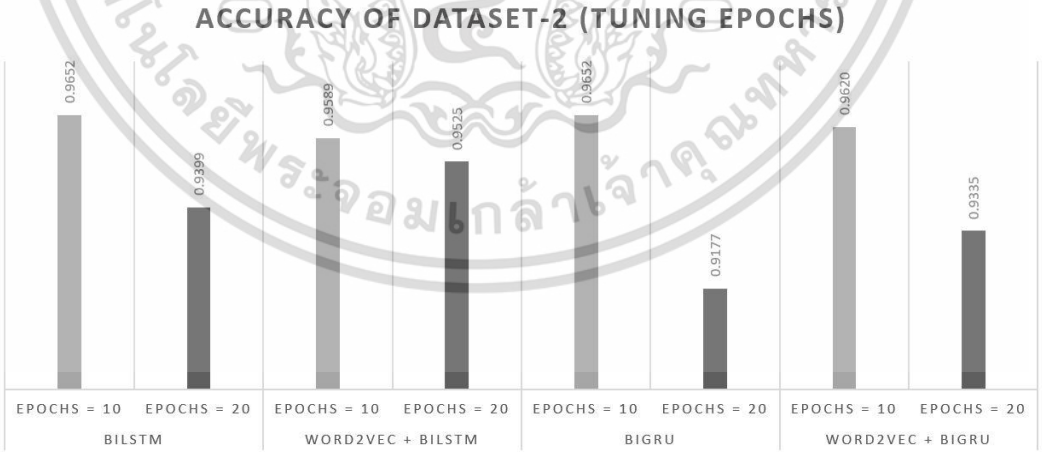
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.6 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Dropout ของ Dataset-2

Model	Dropout	Dataset-2	
		Loss	Accuracy
BiLSTM	0.10	**0.1552**	**0.9652**
	0.25	0.1662	0.9589
Word2Vec + BiLSTM	0.10	**0.1153**	**0.9589**
	0.25	0.1649	0.9557
BiGRU	0.10	**0.1291**	**0.9652**
	0.25	0.2015	0.9304
Word2Vec + BiGRU	0.10	**0.1519**	**0.9620**
	0.25	0.1845	0.9335

จากรูปที่ 4.8 ค้นพบว่าที่เมื่อทำการเพิ่มการปิดโหนดจาก 10% เป็น 25% ค่าความถูกต้องของทุกโมเดลลดลง ดังนั้นจึงสรุปได้ว่าค่า Dropout = 0.10 เป็นค่าที่เหมาะสมชุดข้อมูล Dataset-2 และควรใช้ค่านี้ในการปรับจูนพารามิเตอร์ต่อไป

ขั้นตอนต่อไปจะเป็นการปรับจูนค่า Epochs ให้เพิ่มขึ้นจาก 10 เป็น 20 เพื่อวิเคราะห์ว่าจำรอบที่เหมาะสมในการเทรนโมเดลควรมีค่าเท่าใด โดยประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.9 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.7



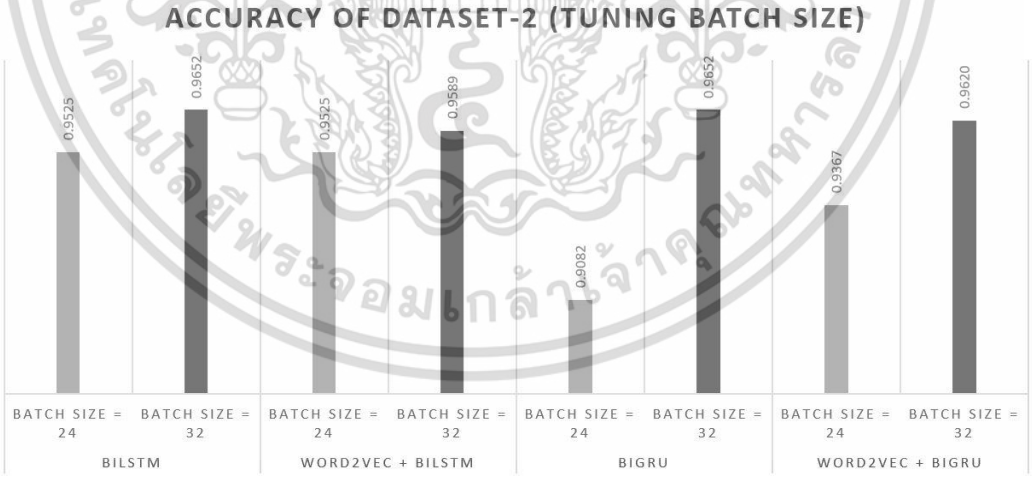
รูปที่ 4.9 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Epochs ของ Dataset-2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.7 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Epochs ของ Dataset-2

Model	Epochs	Dataset-2	
		Loss	Accuracy
BiLSTM	10	**0.1552**	**0.9652**
	20	0.1561	0.9399
Word2Vec + BiLSTM	10	**0.1153**	**0.9589**
	20	0.1715	0.9525
BiGRU	10	**0.1291**	**0.9652**
	20	0.2435	0.9177
Word2Vec + BiGRU	10	**0.1519**	**0.9620**
	20	0.1845	0.9335

จากรูปที่ 4.9 ค้นพบว่าจำนวนที่เหมาะสมในการเทรนโมเดลของชุดข้อมูล Dataset-2 คือ 10 รอบ และควรใช้ค่านี้ในการปรับจูนพารามิเตอร์ต่อไปในขั้นตอนสุดท้ายที่เป็นการทดลองปรับจูนค่า Batch Size จาก 32 เป็น 24 โดยประสิทธิภาพที่ได้ของแต่ละโมเดลจะแสดงในรูปที่ 4.10 และแสดงรายละเอียดเพิ่มเติมในตารางที่ 4.8



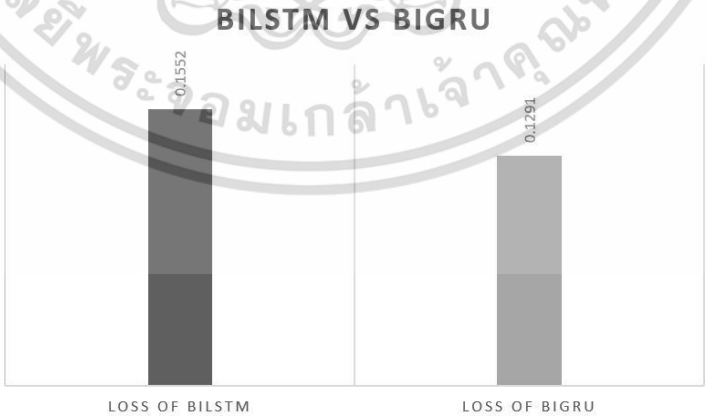
รูปที่ 4.10 ประสิทธิภาพของโมเดล เมื่อมีค่าปรับจูนค่า Batch Size ของ Dataset-2

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.8 ประสิทธิภาพของโมเดลที่ได้ เมื่อมีค่าปรับจูนค่า Batch Size ของ Dataset-2

Model	Batch Size	Dataset-2	
		Loss	Accuracy
BiLSTM	24	0.1806	0.9525
	32	**0.1552**	**0.9652**
Word2Vec + BiLSTM	24	0.1715	0.9525
	32	**0.1153**	**0.9589**
BiGRU	24	0.2133	0.9082
	32	**0.1291**	**0.9652**
Word2Vec + BiGRU	24	0.1809	0.9367
	32	**0.1519**	**0.9620**

จากรูปที่ 4.10 จะเห็นได้ว่า เมื่อทำการเปลี่ยนค่า Batch Size จาก 32 เป็น 24 ประสิทธิภาพของทุกโมเดลลดลง ดังนั้นค่า Batch Size ที่เหมาะสมกับชุดข้อมูล Dataset-2 มีค่าอยู่ที่ 32 นอกจากนี้ จะเห็นได้ว่าโมเดล BiLSTM และ BiGRU ให้ค่าความถูกต้องที่สูงที่สุดเมื่อเทียบกับโมเดลอื่น ๆ ทว่าเมื่อนำค่าความสูญเสียของทั้ง 2 โมเดลจากตารางที่ 4.8 มาทำการแสดงเป็นกราฟแท่งดังรูปที่ 4.11 เพื่อทำการเปรียบเทียบว่าโมเดลใดมีค่าความสูญเสียน้อยกว่ากัน พบว่าโมเดล BiGRU มีค่าความสูญเสียน้อยกว่า ดังนั้นจึงสรุปได้ว่าโมเดลที่เหมาะสมสำหรับชุดข้อมูล Dataset-2 คือ โมเดล BiGRU ที่ค่า Learning Rate = 0.0001, Dropout = 0.10, Epochs = 10 และ Batch Size = 32 โดยมีค่าความถูกต้องอยู่ที่ 96.52%



รูปที่ 4.11 เปรียบเทียบค่าความสูญเสียของโมเดล BiLSTM และ BiGRU ของ Dataset-2

จากผลวิจัยของชุดข้อมูล Dataset-1 และ Dataset-2 จะเห็นได้ว่ามีค่าพารามิเตอร์ที่เหมาะสมเหมือนกัน ทว่าโมเดลที่เหมาะสมกลับแตกต่างกัน ดังแสดงในตารางที่ 4.9 โดยที่ชุดข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Dataset-1 จะเหมาะสมกับโมเดล BiLSTM ในขณะที่ชุดข้อมูล Dataset-2 เหมาะสมกับโมเดล BiGRU มากกว่า ซึ่งอาจเกิดจากเนื้อหาของข้อความในแต่ละชุดข้อมูล เนื่องจากชุดข้อมูล Dataset-2 มีการเก็บชุดข้อมูลข่าวปลอมจากทวิตเตอร์โดยตรงผ่าน Twitter API ด้วยคีย์เวิร์ด ‘Fake news covid’ แม้ว่าจะทำการกรองคีย์เวิร์ด ‘Fake news’ เพื่อลดความเป็นนิเสธของข้อมูลแล้ว แต่ในชุดข้อมูลข่าวปลอมอาจจะยังคงหลงเหลือคีย์เวิร์ดบางส่วนอยู่ อาทิเช่น ‘Fake’, ‘news’ และ ‘covid’ จึงอาจเป็นสาเหตุให้ชุดข้อมูล DataSet-2 เหมาะสมกับโมเดลที่มีแยกองค์ประกอบคำมากกว่าการพิจารณาข้อมูลเป็นลำดับ ทั้งนี้สำหรับทั้ง 2 ชุดข้อมูล แม้ว่าการพิจารณาความสัมพันธ์ของประโยคจะให้ประสิทธิภาพที่ดี ทว่ายังไม่ใช่โมเดลที่ดีที่สุดเมื่อเทียบกับโมเดลที่ไม่มีการประยุกต์ใช้ Word2Vec

ตารางที่ 4.9 รายละเอียดโมเดลที่เหมาะสมและค่าความถูกต้องของแต่ละชุดข้อมูล

	Dataset-1	Dataset-2
Model	BiLSTM	BiGRU
Loss Function	Cross-Entropy	Cross-Entropy
Optimizer	Adam	Adam
Learning Rate	0.0001	0.0001
Dropout	0.10	0.10
Epochs	10	10
Batch Size	32	32
Accuracy	89.41%	96.52%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

งานวิจัยฉบับนี้สามารถวิเคราะห์คุณลักษณะและทำการตรวจจับข่าวปลอมในสื่อสังคมออนไลน์ได้ในเบื้องต้น ทว่าในสื่อสังคมออนไลน์นั้นทุกคนสามารถเป็นผู้เผยแพร่ข้อมูล ดังนั้นจึงไม่จำเป็นที่ข่าวจริงจะถูกเผยแพร่โดยแหล่งข่าวที่น่าเชื่อถือเท่านั้น ผู้ใช้งานทั่วไปก็สามารถเผยแพร่ข่าวจริงได้เช่นกัน ในส่วนนี้ถือเป็นข้อจำกัดของการค้นคว้าอิสระอย่างหนึ่งที่สามารถระบุข่าวจริงได้เพียงจากแหล่งข่าวที่น่าเชื่อถือ ยังไม่สามารถระบุข่าวจริงจากผู้ใช้งานทั่วไปได้ และในส่วนของข่าวปลอมสามารถดึงข้อมูลได้จากคีย์เวิร์ด ‘Fake news covid’ เท่านั้นเช่นเดียวกัน ส่งผลให้เกิดการเกิดการเอนเอียงของข้อมูล ทำให้ในขั้นตอนการวิเคราะห์แนวโน้มและลักษณะของข่าวปลอมอาจเกิดความคลาดเคลื่อนได้ เมื่อต้องนำมาประยุกต์ใช้กับการตรวจสอบชุดข้อมูลอื่น ๆ ทั้งนี้ในส่วนของขั้นตอนการดึงข้อมูลจากทวิตเตอร์ผ่าน Twitter API นั้น ยังไม่สามารถดึงข้อมูลได้มากนัก เนื่องด้วยทางผู้วิจัยยังมีทรัพยากรของเครื่องที่ยังไม่สูงมากพอ ในส่วนของข่าวปลอมจึงสามารถดึงข้อมูลได้สูงสุด 1,500 แถวต่อการดึง 1 ครั้ง

นอกจากนี้งานวิจัยนี้สามารถจำแนกประเภทของข่าวปลอมได้ด้วยการพิจารณาจากบริบทของคำในประโยค ทว่าจากการทดลองจะเห็นได้ว่าโมเดลที่เหมาะสมที่จะใช้ในการจำแนกนั้นแตกต่างออกไปในแต่ละชุดข้อมูล ด้วยเหตุนี้จึงควรจะประยุกต์ชุดข้อมูลเข้ากับโมเดลต่าง ๆ แล้วนำผลลัพธ์ที่ได้มาทำการเปรียบเทียบ เพื่อทำการค้นหาโมเดลที่ให้ประสิทธิภาพดีมากที่สุด

5.2 ข้อเสนอแนะ

ในอนาคตทางผู้วิจัยจะทำการเพิ่มแหล่งข้อมูล ตรวจจับคำเพื่อค้นหาคีย์เวิร์ดที่นิยมใช้ในทั้งข่าวจริงและข่าวปลอม ตลอดจนการค้นหาโมเดลอื่น ๆ ที่เหมาะสมจะนำมาทดลองเพิ่มเติม รวมถึงการค้นหาอัลกอริทึมที่สามารถใช้ค้นหาค่าที่เหมาะสมในการปรับจูนโมเดล และทดลองปรับเปลี่ยนค่าพารามิเตอร์ เพื่อลดความเอนเอียงของข้อมูลและค้นหาโมเดลที่เหมาะสมกับชุดข้อมูลที่สุดต่อไป

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

[1] กิตติพงศ์ ชมบุญ. (2558). **“เทคนิคการจำแนกประเภทข้อมูลส่วนน้อยบนข้อมูลไม่สมดุลด้วยวิธีการแบ่งข้อมูล.”** วิทยานิพนธ์วิศวกรรมศาสตรดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์, มหาวิทยาลัยเทคโนโลยีสุรนารี.

[2] ธนดล สิงขรอาสน์. 2564. **“การเรียนรู้เชิงลึกสำหรับการตรวจจับและรู้จำคำบรรยายในวีดิทัศน์.”** วิทยานิพนธ์วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ, มหาวิทยาลัยมหาสารคาม.

[3] นันทิกา หนูสม. 2561. **“ลักษณะของข่าวปลอมในประเทศไทยและระดับความรู้เท่าทันข่าวปลอมบนเฟซบุ๊กของผู้รับสารในเขตกรุงเทพมหานคร.”** วิทยานิพนธ์นิเทศศาสตรมหาบัณฑิต, มหาวิทยาลัยกรุงเทพ.

[4] สุปัญญา อภิวงศ์โสภณ. 2561. **“การตรวจสอบข่าวปลอมด้วยวิธีการเรียนรู้ของเครื่อง.”** วิทยานิพนธ์วิศวกรรมศาสตรดุษฎีบัณฑิต สาขาวิชาวิศวกรรมคอมพิวเตอร์, จุฬาลงกรณ์มหาวิทยาลัย.

[5] สายชล สินสมบูรณ์ทอง. 2560. **การทำเหมืองข้อมูล เล่ม 1 : การค้นหาความรู้จากข้อมูล.** พิมพ์ครั้งที่ 2. กรุงเทพฯ : จามจุรีโปรดักส์.

[6] Aldwairi, M. and Alwahedi, A. 2018. “Detecting Fake News in Social Media Networks.” in: The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2018). pp. 215-222.

[7] Alkhodaira, S.A. Ding, S.H.H. Fung, B.C.M. Fung. and Liu. J. 2020. “Detecting breaking news rumors of emerging topics in social media.” *Information Processing and Management*. 57(2020), Article 102018.

[8] Bahad, P. Saxena, P. and Kamal, R. 2019. “Fake News Detection using Bi-directional LSTM-Recurrent Neural Network.” in: International Conference on Recent Trends in Advanced Computing 2019 (ICRTAC 2019). 165(2019), pp. 74-82.

[9] Choudhary, M., Chouhan, S.S., Pilli, E.S., & Vipparthi, S.K. (2021). BerConvoNet: A deep learning framework for fake news classification. *Applied Soft Computing*. 110, Article 107614.

[10] Developer Platform. 2564. **Counting character**. [Online]. Available: https://developer.twitter.com/en/docs/counting-characters.

[11] Huang, Y.-F. and Chen, P.-H. 2020. “Fake news detection using an ensemble learning model based on Self-Adaptive Harmony Search algorithms.” *Expert Systems with Applications.* 159(2020), Article 113584.

[12] Kowirat, R. and Boongasame, L. 2021. “Fake News Detection on Social Media: Case Study of 2019 Novel Coronavirus” in: 2021 3rd International Conference on E-Business and E-commerce Engineering (EBEE). pp. 289-302.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

[13] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). “Efficient estimation of word representations in vector space.” ArXiv Prepr. arXiv:1301.3781, http://arxiv.org/abs/1301.3781.

[14] Nasir, J.A. Khan, O.S. and Varlamis, I. 2021. “Fake news detection: A hybrid CNN-RNN based deep learning approach.” *International Journal of Information Management Data Insights.* 1(1), Article 100007.

[15] Patwa, P. Sharma, S. PYKL, S. Guptha, V. Kumari, G. Akhtar, A.S. Ekbai, A. Das, A. and Chakraborty, T. 2021. “Fighting an Infodemic : COVID-19 Fake News Dataset.” In Combating Online Hostile Posts in Regional Languages during Emergency Situation (pp. 21–29). Springer International Publishing.

[16] Phi, M. (2018). Illustrated Guide to LSTM’s and GRU’s: A step by step explanation. [Online]. Available : https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21.

[17] Promrit, N. (2020). Sentiment Analysis 101. [Online]. Available: https://blog.pjjop.org/sentiment-analysis-101/?fbclid=IwAR1YLY8egVdu04K1B3wM7WRdka5358qOU6H-Ev0mPJDeOA_5AiCE_qR1l9A.

[18] Samadi, M. Mousavian, M. and Momtazi, S. 2021. “Deep contextualized text representation and learning for fake news detection.” *Information Processing and Management.* 58(6), Article 102723.

[19] Sastrawan, I.K. Bayupati, I.P.A., and Arsa, D.M.S.. 2021. “Detection of fake news using deep learning CNN-RNN based methods.” The Korean Institute of Communications and Information Sciences. pp. 1-13.

[20] Tangruamsub, S., (2017). Long Short-Term Memory (LSTM). [Online]. Available: https://medium.com/@sinart.t/long-short-term-memory-lstm-e6cb23b494c6.

[21] World Health Organization. 2564. **Coronavirus disease (COVID-19)**. [Online]. Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/q-a-coronaviruses.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นางสาวรัชนีวรรณ กอวิรัตน์
วัน เดือน ปีเกิด	18 มิถุนายน 2538
ที่อยู่ปัจจุบัน	87/20 หมู่ที่1 ซอยตั้งพัฒนา ถนนเลียบคลองสี่วาพาสวัสดิ์ ตำบลนาดี อำเภอเมืองสมุทรสาคร จังหวัดสมุทรสาคร 74000
ประวัติการศึกษา	(2561) วิศวกรรมศาสตรบัณฑิต สาขาวิศวกรรมคอมพิวเตอร์ เกรดเฉลี่ย 3.73 (มหาวิทยาลัยกรุงเทพ) (กำลังศึกษา) วิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูลและการวิเคราะห์ (สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง)
ทุนการศึกษาที่ได้รับ	ทุนอุดหนุนจากกองทุนวิจัยสถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง [KREF016310]
ผลงานทางวิชาการ	1. Kowirat, R. and Boongasame, L. 2021. “Fake News Detection on Social Media: Case Study of 2019 Novel Coronavirus” in: 2021 3rd International Conference on E-Business and E-commerce Engineering (EBEE). pp. 289-302.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้