การจัดกลุ่มข่าวปลอมด้วยเทคนิคการเรียนรู้เครื่อง

วิสิทธิ์ วาณิชยานนท์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิศวกรรมศาสตร์มหาบัณฑิต
สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่
วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์
มหาวิทยาลัยธุรกิจบัณฑิตย์
ปีการศึกษา 2565

FAKE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES

WISITH WANISHYANON

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering

Department of Big Data Engineering

College of Innovative Technology and Engineering,

Dhurakij Pundit University

Academic Year 2022

ใบรับรองวิทยานิพนธ์

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์

วิศวกรรมศาสตรมหาบัณฑิต สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่

หัวข้อวิทยานิพนธ์ การจัดกลุ่มข่าวปลอมด้วยเทคนิคการเรียนรู้เครื่อง

เสนอโดย วิสิทธิ์ วาณิชยานนท์

สาขาวิชา วิศวกรรมข้อมูลขนาดใหญ่

อาจารย์ที่ปรึกษาวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น

ได้พิจารณาเห็นชอบโดยคณะกรรมการสอบวิทยานิพนธ์แล้ว

(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพัชร์ อารีรัชกุลกานต์) ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตคงชื่น) กรรมการที่ปรึกษาวิทยานิพนธ์

(ดร.ธนภัทร ยังคะจิตร) กรรมการ

(ดร.เอกสิทธิ์ พัชรวงศ์ศักดา) กรรมการ

วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ รับรองแล้ว

(ดร.ชัยพร เขมะภาตะพันธ์) คณบดีวิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์

วันที่ 31 เดือน กรกฎาคม พ.ศ. 2566

หัวข้อวิทยานิพนธ์ การจัดกลุ่มข่าวปลอมด้วยเทคนิคการเรียนรู้เครื่อง
ชื่อผู้เขียน วิสิทธิ์ วาณิชยานนท์
อาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.ดวงใจ จิตต์คงชื่น
หลักสูตร วิศวกรรมข้อมูลขนาดใหญ่
ปีการศึกษา 2565

บทคัดย่อ

การตรวจจับข่าวปลอมเป็นงานที่ยาก เนื่องจากข่าวปลอมมีการพัฒนารูปแบบอย่างรวดเร็ว มีการนำเสนอข่าวเหมือนกับข่าวจริงจนแยกไม่ออก ดังนั้นเพื่อหาเทคนิคการเรียนรู้เครื่องที่ดีที่สุดในการนำมาสร้างแบบจำลองการจัดกลุ่มข่าวปลอม ผู้วิจัยนำเสนอวิธีการจัดกลุ่มข่าวปลอม โดยใช้ชุดข้อมูลจาก Fake News Copus ซึ่งเป็นชุดข้อมูลที่ประกอบด้วยหัวข้อข่าวและเนื้อหา ใช้เทคนิคการเรียนรู้เครื่อง 4 เทคนิค คือ การถดถอยโลจิสติก นาอีฟเบส์ ซัพพอร์ทเวกเตอร์แมชชีน และตัวจำแนกป่าแบบสุ่ม ประเมินผลแบบจำลองโดยวัดประสิทธิภาพการจำแนกตามแนวคิดการค้นคืนสารสนเทศโดยใช้ค่า AUC (Area Under the ROC Curve) สำหรับการตรวจสอบจากหัวข้อข่าวและเนื้อหาของข่าว

ผลการวิจัยพบว่าแบบจำลองเทคนิคการถดถอยโลจิสติก สามารถจัดกลุ่มข่าวปลอมได้ดีที่สุดด้วยค่า AUC ร้อยละ 94 ตามด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีนร้อยละ 91 เทคนิคนาอีฟเบส์ร้อยละ 89 และเทคนิคตัวจำแนกป่าแบบสุ่มร้อยละ 88 ตามลำดับ

คำสำคัญ : ข่าวปลอม,เทคนิคการเรียนรู้เครื่อง

.....

Thesis Title	FAKE NEWS CLASSIFICATION USING MACHINE LEARNING TECHNIQUES
Author	WISITH WANISHYANON
Thesis Advisor	Asst.Prof. Duangjai Jitkongchuen, Ph.D.
Program	Big Data Engineering
Academic Year	2022

ABSTRACT

News can spread like a wildfire through the network. Unfortunately, most of the fake news stories were shared before adequate contemplation on the news's integrity was conducted. Moreover, with the lighting pace of information flowing through social networking, most news once shared is out of sight out of mind. Consequently, relying on human to screen out fake news ourselves might not be very conductive and productive screening instrument. To enable automated fake news detection, this research focuses on utilizing machine learning techniques to construct fake news classification models. In this paper we created a fake news classification model using machine learning techniques with 4 techniques, Logistic regression, Naïve Bayes, Support Vector Machine and Random Forest. The experiments were conducted equally random data from Fake Corpus. As result, 32,000 records of reliable and fake news were generated. Classification performance was evaluated by using the AUC (Area Under the ROC Curve).

The result of this study showed that the Logistic Regression Model can classify fake news with 94% of the AUC was classified, following by Support Vector Machine 91%, Naïve Bayes 89% and Random Forest 88% respectively.

Keywords: Fake News, Machine Learning

.....

กิตติกรรมประกาศ

งานวิจัยชิ้นนี้ ไม่อาจสำเร็จลงได้ หากขาดการให้โอกาสและการสนับสนุนจากหลายท่าน ผู้วิจัยขอขอบพระคุณมหาวิทยาลัยธุรกิจบัณฑิตย์ ที่ให้โอกาสอบรมด้านวิศวกรรมข้อมูลตั้งแต่ยังไม่ได้เข้าเป็นนักศึกษา ขอบพระคุณอาจารย์วรพล พงษ์เพ็ชร ที่ให้คำปรึกษา ให้โอกาสทั้งทางด้านการศึกษา การทำงานและการใช้ชีวิต ขอบพระคุณอาจารย์เอกสิทธิ์ พัชรวงศ์ศักดา ที่ให้ความเมตตาและกรุณาสอนสิ่งที่ยากให้เป็นสิ่งที่ง่าย ขอบพระคุณอาจารย์ดวงใจ จิตคงชื่น ที่ให้กรุณาช่วยเหลือ ให้คำปรึกษาทั้งวิชาการและการทำงาน ขอบพระคุณอาจารย์รัฐศิลป์ รานอกภานุวัชร ที่ให้คำปรึกษาและแนะแนวทางการเรียน ขอบพระคุณอาจารย์ณัฐพัชร์ อารีรัชกุลกานต์ ที่กรุณาให้คำปรึกษาในหลาย ๆ สิ่ง และอาจารย์ธนภัทร ฆังคะจิตร ที่คำปรึกษาและคอยผลักดัน นอกจากนี้รวมถึงเจ้าหน้าที่ประจำภาคทุกท่าน โดยเฉพาะคุณกุลธิดา รอดบุญ ที่ให้กำลังใจและให้ความช่วยเหลือตลอดมา

นอกจากนี้ ผู้วิจัยขอขอบพระคุณครอบครัว และเพื่อน ๆ ที่อดทน เสียสละ ให้เวลาและให้โอกาสผู้วิจัย ได้ศึกษาหาความรู้เพิ่มเติม ผู้วิจัยหวังเป็นอย่างยิ่งว่า วิทยานิพนธ์ฉบับนี้จะเป็นแนวทางการศึกษาต่อยอด สำหรับผู้ที่สนในศึกษาในด้านนี้ต่อไป

วิสิทธิ์ วาณิชยานนท์

สารบัญ

สารบัญ (ต่อ)

หน้า

สารบัญตาราง

ตารางที่ **หน้า**

สารบัญภาพ

ภาพที่ **หน้า**

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

จากความก้าวหน้าของระบบอินเทอร์เน็ตและเทคโนโลยีการสื่อสารและการรับส่งข่าวสารที่สะดวกรวดเร็ว ทำให้เกิดการเปลี่ยนแปลงรูปแบบการใช้สื่อของมนุษย์อย่างมากมาย ทั้งการปรับปรุงให้การสื่อสารมีประสิทธิภาพเพิ่มขึ้นหรือนำเสนอข้อมูลผ่านช่องทางใหม่ ๆ จนทำให้กลายเป็นการสื่อสารไร้พรมแดน คนเราสามารถสื่อสารกันได้ทั่วทุกมุมโลก นอกจากนี้ คอมพิวเตอร์ โน้ตบุ๊คและอุปกรณ์การสื่อสารพกพา หากสามารถเข้าถึงอินเทอร์เน็ตได้ ก็สามารถเข้าถึงเครือข่ายสังคมออนไลน์ได้อย่างสะดวกทุกที่ทุกเวลา เครือข่ายสังคมออนไลน์ได้กลายเป็นพื้นที่สาธารณะที่สมาชิก ที่ไม่จำกัดเพศ อายุ เชื้อชาติ ศาสนา ระดับการศึกษา อาชีพ ทุกคนสามารถเป็นได้ทั้งผู้สื่อสารหรือเขียนบอกเล่าเนื้อหา เรื่องราว ประสบการณ์ บทความ รูปภาพและวีดิโอจากการเขียนและจัดทำขึ้นเอง หรือพบเจอจากสื่ออื่น ๆ แล้วนำมาแบ่งปันให้กับผู้อื่นซึ่งอยู่ในเครือข่ายของตนผ่านระบบอินเทอร์เน็ตและสื่อสังคมออนไลน์ (Social Media) บทบาทที่เปลี่ยนไปผู้ใช้สื่อสังคมออนไลน์จึงมีสามารถมีบทบาททั้งในฐานะของผู้ส่งสาร (Sender) และผู้รับสาร (Receiver) ได้ในเวลาเดียวกันสามารถสะท้อนความคิดเห็นและผลิตเนื้อหาได้ด้วยตนเอง หรือ กล่าวได้ว่าในปัจจุบันผู้รับสารสามารถเป็นได้ทั้ง “ผู้ชม ผู้ใช้ และผู้แชร์” [10] ซึ่งแตกต่างการรับสารผ่านสื่อดั้งเดิม จากเดิมที่ผู้รับสารจะต้องรอเวลาในการเผยแพร่เนื้อหาผ่านสื่อดั้งเดิม บทบาททางการ สื่อสารของผู้รับสาร (Receiver) จึงอยู่ในลักษณะที่เป็นผู้ตาม (Passive Receiver) เท่านั้น ส่งผลให้สื่อสังคมออนไลน์กลายเป็นที่นิยมและเป็นวัฒนธรรมการสื่อสารรูปแบบใหม่ที่มีอัตราการใช้งานที่ เติบโตอย่างรวดเร็ว ประชาชนจึงหันมานิยมรับข่าวสารผ่านทางสังคมออนไลน์ (Social Network) เพิ่มขึ้น ความนิยมนี้ทำให้มีผู้ที่ใช้ช่องทางนี้เป็นเครื่องมือสร้างผลประโยชน์ให้กับตนเองด้วยวิธีการต่าง ๆ เกิดขึ้นอย่างมากมาย วิธีที่เป็นที่นิยมมากวิธีหนึ่งคือการปล่อย “ข่าวปลอม” (Fake News) ดังปรากฏการณ์สำคัญ เช่น การลงประชามติแยกตัวออกจากสหภาพยุโรปของประเทศสหราชอาณาจักร ผลการเลือกตั้งของประเทศสหรัฐอเมริกาและประเทศฝรั่งเศส หรือกรณีสวรรคตของพระบาทสมเด็จพระปรมินทรมหาภูมิพลอดุลยเดช จุดประสงค์ของข่าวปลอมคือต้องการให้ผู้รับข่าวอยู่ในสถานะทางอารมณ์ที่สามารถโดนโน้มน้าวได้ง่าย ซึ่งจะทำให้ผู้ปล่อยข่าวมีอิทธิพลต่อการการรับรู้ การตัดสินใจและพฤติกรรมของผู้ใช้ในสื่อสังคมออนไลน์ได้ตามความต้องการ

เมื่อพฤติกรรมของผู้ใช้ในการเปิดรับข่าวสารเปลี่ยนแปลงไป ประกอบกับพฤติกรรมของมนุษย์ที่มักมีแนวโน้ม “เชื่อในสิ่งที่เราเชื่ออยู่แล้ว” ทำให้ข่าวที่อาจไม่ถูกต้องตามข้อเท็จจริง (Fact) แต่ถูกใจคนอ่านได้รับความนิยมมากขึ้นตามไปด้วย [5] ถึงแม้ว่าสื่อสังคมออนไลน์จะมีประโยชน์ในการช่วยให้เราสามารถเข้าถึงข้อมูลข่าวสารได้ง่ายดายและรวดเร็ว แต่ปัญหาในสังคมที่เกิดขึ้นจากการพึ่งพาสื่อสังคมออนไลน์ในการบริโภคข้อมูลข่าวสารมากจนเกินไปคือ การที่คนปักใจเชื่อ ข่าวสารทั่วไป รวมถึงข่าวลือ ข่าวปลอมต่าง ๆ บนสื่อสังคม

ออนไลน์นำมาเผยแพร่โดยไม่คำนึงว่าข้อมูลนั้นมีความถูกต้องมากน้อยเพียงใด ขณะที่สื่อมวลชนพากันพูดถึงทฤษฎีฟองสบู่ หรือ Filter Bubble ของ เอลี ปารีเซอร์ นักเคลื่อนไหวทางอินเทอร์เน็ต อีกครั้ง ซึ่งปารีเซอร์กล่าวว่าความหมายของอินเทอร์เน็ตแตกต่างไปจากเดิมมาก จากที่เคยเชื่อมโยง ทุกคนในโลกเข้าด้วยกัน ปัจจุบันพื้นที่ออนไลน์และเว็บไซต์ข่าวมีระบบกลไกคัดกรองข้อมูลตามความชอบและสิ่งที่ผู้ใช้งานมีปฏิสัมพันธ์โต้ตอบด้วยบ่อย ๆ กลายเป็นว่าเรากำลังอยู่ในฟองสบู่ รับรู้เฉพาะบางข้อมูลข่าวสารเท่านั้น [4]

พฤติกรรมการใช้สื่อสังคมออนไลน์ในการรับรู้และเชื่อข่าวสารโดยขาดวิจารณญาณหรือการรู้เท่าทันนั้นเป็นเรื่องสำคัญที่ไม่สมควรจะมองข้าม ข่าวสารที่ไม่มีความจริงที่เผยแพร่ทางสื่อสังคมออนไลน์นั้นอาจจะส่งผลกระทบต่อสังคมโดยรวมได้หากผู้ที่รับข่าวสารนั้นขาดความรู้เท่าทันในการ เชื่อข่าวนั้นโดยไม่ตระหนักถึงผลที่อาจจะเกิดขึ้น เช่นข่าวปลอม ข่าวลวง เกี่ยวกับการเกิดภัยพิบัติ ร้ายแรงต่าง ๆ ที่ สร้างความตื่นตระหนกแก่ผู้คนในวงกว้าง หรือข่าวปลอมเกี่ยวกับ ข้อมูลทางสุขภาพ และยา อาหารรักษาโรคในทางที่ผิด ตลอดจนข้อมูลที่บิดเบือนเกี่ยวกับการเมือง ความ ปลอดภัย และ ความมั่นคงของชาติที่อาจมีผู้ไม่หวังดีจงใจเผยแพร่เพื่อสร้างสถานการณ์ให้เกิดความสับสนวุ่นวาย [12] นอกจากนี้ยังเพิ่มความเกลียดชัง สร้างความแตกแยก และอาจกระทบถึงความมั่นคงของชาติ

ข่าวปลอม (Fake News) คือข่าวที่นำไปสู่ความเข้าใจผิดและไม่ตั้งอยู่บนพื้นฐานความเป็นจริง มีความจงใจให้สารสนเทศที่ผิดพลาดเพื่อทำให้เกิดความเข้าใจผิดและหาผลประโยชน์จากการเข้าใจผิดนั้น ผู้ที่ขาดทักษะในการจำแนกข้อเท็จจริงจึงมักมีโอกาสถูกยั่วยุให้หลงเชื่อข่าวปลอมได้โดยง่าย ทั้งนี้ ข่าวปลอมมิได้เป็นประเด็นใหม่สำหรับสังคม และได้ถูกใช้มาอย่างยาวนานในสงครามหรือเกมส์ความขัดแย้งในแต่ละยุคสมัย โดยการเผยแพร่ในรูปแบบต่าง ๆ เช่น การบอกต่อกันแบบปากต่อปาก ใบปลิว เรื่องที่เล่าต่อ ๆ กันมา ฟอเวิร์ดเมล์ จนถึงสื่อสังคมออนไลน์หรือโซเชียลมีเดีย การพิจารณาความน่าเชื่อถือของข่าวแต่ดั้งเดิมคือการพิจารณาเนื้อหาและบริบทของข่าว แล้วตีความไปตามที่ความคิด ความรู้และความเชื่อของผู้ที่รับข่าว แต่หลักการพิจารณาความน่าเชื่อถือของข่าวดังกล่าวไม่สามารถใช้ได้อีกต่อไปเนื่องจากข่าวสามารถปลอมแปลงได้ด้วยเหตุผลหลาย

ข่าวปลอมนั้นไม่ใช่เรื่องใหม่แต่เป็นเรื่องที่มีมานานตั้งแต่ยุคโบราณแต่ยังไม่แพร่กระจายมากเนื่องจากยังไม่มีสื่อออนไลน์ ที่สามารถกระจายข่าวได้อย่างรวดเร็วได้เหมือนปัจจุบัน ในอดีตข่าวปลอมถูกเผยแพร่ในรูปแบบต่าง ๆ เช่น การบอกต่อกันแบบปากต่อปาก ใบปลิว เรื่องที่เล่าต่อกันมา เมื่อมาถึงยุคที่เริ่มมีอินเทอร์เน็ต ข่าวปลอมเริ่มแพร่ระบาดทางฟอร์เวิรดเมล์ (Forward Mail) เกิดการส่งต่อกันมากขึ้น และจนมาถึงยุคสื่อสังคมออนไลน์หรือโซเชียลมีเดีย (Social Media) การตรวจสอบข่าวว่าเป็นข่าวจริงหรือข่าวปลอมในแต่ละยุคสมัยมีการเปลี่ยนแปลงไปตามปัจจัยที่เอื้ออำนวย ในยุคของการเผยแพร่ข่าวด้วยสื่อสิ่งพิมพ์นิยมใช้วิธีการตรวจที่มาและพิจารณาเนื้อหาโดยยึดจากคำสำคัญ (Keyword) โครงสร้างภาษา ความเกี่ยวข้องระหว่างหัวข้อข่าวกับเนื้อหาของข่าว จากนั้นจึงส่งให้บรรณาธิการอ่านแล้วพิจารณาว่าควรเผยแพร่หรือไม่ หลายครั้ง

เป็นการพิจารณาโดยใช้ประสบการณ์ตัดสินเพราะต้องแข่งกับเวลาเพื่อให้สามารถพิมพ์ข่าวได้ทัน ทำให้ข่าวปลอมมีโอกาสเผยแพร่ออกไป

เมื่อเทคโนโลยีจะเริ่มมีความก้าวหน้ามากขึ้น สามารถคัดกรอง ตรวจสอบความน่าเชื่อถือของข่าวได้ด้วยวิธีการและเครื่องมือที่หลากหลาย แต่หลักการตรวจสอบข่าวก็ยังคงใช้การหลักการเดิมคือพิจารณาจากเนื้อหาด้วยคำสำคัญและยังคงต้องใช้บรรณาธิการหรือ “คน” นอกจากนี้ปริมาณข่าวในแต่ละวันมีจำนวนมาก ทำให้ไม่สามารถตรวจสอบทุกข่าวได้ การตรวจสอบข้อเท็จจริงต้องอาศัยข้อมูลภายนอกซึ่งอาจเป็นเรื่องยาก เพราะข้อมูลภายนอกมีการเปลี่ยนแปลงตลอดเวลา บ่อยครั้งที่ข่าวได้รับการพิสูจน์แล้วว่าเป็นเท็จ แต่ความเสียหายได้เกิดขึ้นแล้ว และที่สำคัญข่าวปลอมมักจะหายไปอย่างรวดเร็วทำให้ยากต่อการตรวจสอบ ถึงจะมีวิธีการวิเคราะห์ทางภาษาที่ใช้แนวทางการวิเคราะห์คุณสมบัติเชิงปริมาณ เช่น โครงสร้างไวยากรณ์ การเลือกคำ เครื่องหมายวรรคตอน ฯลฯ ที่มีความซับซ้อน สามารถทำงานได้เร็วกว่ามนุษย์และใช้กับข่าวหลากหลายประเภท แต่เทคนิควิธีการเขียนข่าวก็มีความซับซ้อนขึ้นเช่นกัน เช่นข่าวเสียดสีที่มีจำนวนมากและตรวจสอบได้ง่าย แต่การใช้ถ้อยคำรุนแรง หยาบคาย ไร้สาระ ไม่เกี่ยวข้องกับเนื้อหา ทำให้การพัฒนาอัลกอริทึมในการตรวจจับข่าวปลอมมีประสิทธิภาพไม่เพียงพอ

1.2 วัตถุประสงค์ของงานวิจัย

เพื่อนำเสนอวิธีการจัดกลุ่มข่าวปลอมโดยใช้แบบจำลองการเรียนรู้เครื่องที่เพิ่มประสิทธิภาพร่วมกับการใช้ข้อมูลตัวอักษรจากหัวข้อข่าวและเนื้อหาข่าว โดยมุ่งเน้นที่การปรับปรุงรูปแบบการจำลองการเรียนรู้เครื่องให้สามารถวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ

1.3 ขอบเขตของงานวิจัย

1.3.1 ในงานวิจัยชิ้นนี้ ผู้วิจัยใช้ชุดข้อมูลสาธารณะ Fake News Curpus [15] ซึ่งมีผู้รวบรวมไว้จากเว็บไซต์ 1001 เว็บไซท์ จาก BuzzFeed และ Politifact ประกอบด้วยบทความหรือข่าวทางด้านการเมือง

1.3.2 ข้อมูลที่ใช้ เป็นข้อมูลเชิงตัวอักษรที่ คือ หัวข้อข่าวและเนื้อหาของข่าว เป็นภาษาอังกฤษ

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1 เพิ่มความน่าเชื่อถือของการตรวจจับข่าวปลอม ที่มีความสอดคล้องกับข้อเท็จจริง1.4.2 ขอบเขตของแผนงานการจัดกลุ่ม

1.4.2 สามารถนำกรอบงานวิจัยนี้ไปประยุกต์ใช้กับข้อมูลหัวข้อข่าวที่เป็นภาษาอื่นได้

1.5 นิยามศัพท์เฉพาะ

ข่าวปลอม (Fake News) หมายถึง สื่อหรือเนื้อหาของข่าวที่ถูกแต่งขึ้นทั้งหมด อาจมีความจริงเพียงเล็กน้อยหรือไม่มีเลย มีวัตถุประสงค์เพื่อหลอกลวง บิดเบือน โฆษณาชวนเชื่อ กระตุ้นการเข้าชมเว็บไซต์และเพื่อให้เกิดการแพร่กระจายข่าวผ่านทางสื่อสังคมออนไลน์โดยหวังผลให้ผู้รับข่าวปลอมเกิดการเข้าใจผิดมากกว่าความบันเทิง ทั้งนี้วัตถุประสงค์หลักของข่าวปลอมอาจมุ่งหวังผลประโยชน์ทางการเงิน ธุรกิจ การเมือง การทหาร หรือแม้กระทั่งความมั่นคงของชาติ การนำเสนอมีความเป็นมืออาชีพ ทั้งการจงใจเลียนแบบสื่อหนังสือพิมพ์ที่มีอยู่จริงไปจนถึงสื่อโฆษณาชวนเชื่อของรัฐบาลจนทำให้ผู้ใช้ไม่สามารถแยกแยะระหว่างข่าวจริงกับข่าวปลอมออกจากกันได้

การเรียนรู้เครื่อง (Machine Learning) หมายถึง การศึกษาและสร้างขั้นตอนวิธีที่ทำให้เครื่องคอมพิวเตอร์เรียนรู้ได้จากข้อมูลตัวอย่างหรือสภาพแวดล้อม เพื่อสร้างตัวแบบหรือขั้นตอนวิธีและนำไปใช้ทำนายข้อมูลใหม่ได้ โดยมีจุดมุ่งหมายคือการพัฒนาหรือปรับปรุงประสิทธิภาพการทำงานของระบบให้ดียิ่งขึ้น เมื่อเครื่องเรียนรู้แล้วความรู้ที่เรียนได้จะเก็บไว้ในฐานความรู้ด้วยรูปแบบการแทนความรู้อย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชั่น

การจำแนกประเภทข้อมูล (Classification) คือ การแก้ปัญหาพื้นฐานของการเรียนรู้แบบมีผู้สอน โดยปัญหาคือการทำนายประเภทของข้อมูลจากคุณสมบัติต่าง ๆ ของข้อมูล โดยการเรียนรู้แบบมีผู้สอนจะสร้างแบบจำลองหรือฟังก์ชั่น เชื่อมโยงระหว่างคุณสมบัติของข้อมูลกับประเภทของข้อมูลจากตัวอย่างข้อมูลสอนแล้วจึงใช้แบบจำลองหรือฟังก์ชั่นนี้ทำนายประเภทของข้อมูลที่ไม่เคยพบ เครื่องมือหรือขั้นตอนที่ใช้สำหรับการแบ่งประเภทข้อมูลเช่น โครงข่ายประสาทเทียม การวิเคราะห์การถดถอย ตัวจำแนกป่าแบบสุ่ม ซัพพอร์ทเวกเตอร์แมชชีน เนอีฟเบย์ เป็นต้น

บทที่ 2
แนวคิด ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข่าวปลอม (Fake News)

ข่าวปลอม (Fake News) หมายถึง สื่อหรือเนื้อหาของข่าวที่ถูกแต่งขึ้นทั้งหมด อาจมีความจริงเพียงเล็กน้อยหรือไม่มีเลย มีวัตถุประสงค์เพื่อหลอกลวง บิดเบือน โฆษณาชวนเชื่อ กระตุ้นการเข้าชมเว็บไซต์และเพื่อให้เกิดการแพร่กระจายข่าวผ่านทางสื่อสังคมออนไลน์โดยหวังผลให้ผู้รับข่าวปลอมเกิดการเข้าใจผิดมากกว่าความบันเทิง ทั้งนี้วัตถุประสงค์หลักของข่าวปลอมอาจมุ่งหวังผลประโยชน์ทางการเงิน ธุรกิจ การเมือง การทหาร หรือแม้กระทั่งความมั่นคงของชาติ การนำเสนอมีความเป็นมืออาชีพ ทั้งการจงใจเลียนแบบสื่อหนังสือพิมพ์ที่มีอยู่จริงไปจนถึงสื่อโฆษณาชวนเชื่อของรัฐบาลจนทำให้ผู้ใช้ไม่สามารถแยกแยะระหว่างข่าวจริงกับข่าวปลอมออกจากกันได้

2.1.1 ความหมายของข่าวปลอม

ข่าวปลอม (Fake News) เป็นคำที่ถูกใช้ในหลากบริบท หลายความหมาย และยังมีความกำกวมในเรื่องของคำนิยาม ซึ่งอาจครอบคลุมถึงเนื้อหาหลากหลายประเภท เช่น ข่าวเสียดสี ข่าวล้อเลียน เรื่องแต่งเพื่อหลอกลวง การโฆษณาชวนเชื่อ ส่งผลให้เกิดความเข้าใจผิดต่อไปอย่างไม่มีที่สิ้นสุด มีงานวิจัยจำนวนมากที่ให้ความหมาย และศึกษาถึงผลกระทบของข่าวปลอม ดังตัวอย่างต่อไปนี้

Hunt Allcott จากมหาวิทยาลัยนิวยอร์ค และ Matthew Genzhow จากมหาวิทยาลัยสแตนฟอร์ด [24] ได้ทำการวิจัยเกี่ยวกับข่าวปลอม ว่ามีผลต่อการเลือกตั้งในสหรัฐอเมริกาจริงหรือไม่ โดยเตรียมข่าวสามประเภท คือ ข่าวจริง ข่าวปลอม และข่าวปลอมที่ทีมวิจัยเขียนขึ้นมาเอง โดยถามความเห็นจากผู้มีสิทธิ์และลงคะแนนเสียงเลือกตั้ง ผลการวิจัยพบว่าคนจำนวนมากเชื่อข่าวริงมากกว่าข่าวปลอม มีเพียงร้อยละ 10 เท่านั้นที่เชื่อข่าวปลอมและข่าวปลอมที่ทีมวิจัยสร้างขึ้นมา นั่นแสดงให้เห็นว่า ข่าวปลอมมีอิทธิพลเพียงเล็กน้อยต่อทัศนคติของคน แต่ทว่าสิ่งที่น่ากังวลใจระยะยาวคือ คนมีแนวโน้มที่จะโหยหาข้อมูลข่าวสารที่สอดคล้องกับทัศนคติของตนเอง และในโลกอินเทอร์เน็ตก็มีข้อมูลมากมายให้สามารถนำมาเติมเต็มความเชื่อเดิม ๆ ของตน

Wardle ได้ให้ความหมายของข่าวปลอมไว้ว่า ข่าวปลอม หมายถึง ข่าวที่มีเนื้อหาอันเป็นเท็จ (Wardle, 2017) ถูกประดิษฐ์ขึ้นโดยไม่มีข้อเท็จจริงที่สามารถตรวจสอบแหล่งที่มาหรือคำพูดได้ บางครั้งเรื่องราวเหล่านี้อาจอยู่ในรูปแบบของการโฆษณาชวนเชื่อที่มีเจตนาออกแบบมาเพื่อทำให้ผู้อ่านเข้าใจผิด โดยอาจมีการออกแบบให้พาดหัวข่าวให้ดูเกินจริงเพื่อเรียกแขก (Click bait) เพื่อหวังผลประโยชน์จากคนจำนวนมาก วิธีนี้ทำให้ข่าวปลอมเผยแพร่ผ่านสื่อสังคมออนไลน์ได้อย่างกว้างขวางเนื่องจากได้รับส่งแบ่งผลประโยชน์ได้ง่ายและรวดเร็ว [17]

ข่าวปลอมคือประเภทหนึ่งของการหลอกลวงหรือการกระจายข้อมูลที่ไม่ถูกต้องหรือเป็นข้อมูลที่เป็นเท็จ โดยใช้สื่อสิ่งพิมพ์ทั้งแบบดั้งเดิมและสื่อกระจายเสียงหรือผ่านทางสื่อสังคมออนไลน์ เพื่อให้มีคุณสมบัติเป็นข่าวปลอมเรื่องราวจะต้องมีการเขียนและตีพิมพ์โดยมีเจตนาที่จะให้ผู้อื่นเข้าใจผิดเพื่อผลประโยชน์ทางเศรษฐกิจและการเมือง [18]

ข่าวปลอมมีส่วนประกอบของข้อมูลที่เป็นเท็จ จงใจหลอกลวง และได้รับการออกแบบเพื่อประโยชน์เชิงพาณิชย์มักได้รับการเผยแพร่บนเว็บไซต์คล้ายคลึงกับข้อมูลปลอม ข่าวปลอมจะใช้ภาษาและภาพข่าวของการรายงานข่าวเพื่อสร้างความเชื่อถือกับผู้อ่านซึ่งแตกต่างจากการรายงานข่าวในหนังสือพิมพ์ รวมถึงข้อมูลที่มีการปลอมแปลงอย่างจงใจ ทำเช่นนี้เพื่อดึงดูดผู้เข้าชมเว็บไซต์เพื่อรับประโยชน์ทางการเงิน [19]

ข่าวปลอม โดยความหมายของมันคือข้อมูลชุดหนึ่งที่ทำขึ้นมาโดยมีเจตนาจะบิดเบือนและโจมตีบุคคล กลุ่มคนหรือองค์กร มีเนื้อหากระตุ้นอารมณ์เพื่อเรียกร้องความสนใจจากผู้อ่าน หรือที่เราเคยเห็นตามหน้าสื่อบ่อย ๆ ว่า “พาดหัวเรียกแขก” [20]

ข่าวปลอมที่คนเขียนตั้งใจให้เชื่อแบบผิด ๆ บ้างก็เป็นข่าวปลอมที่เกิดจากการกุเรื่องขึ้นมา แต่ทำให้กลายเป็นประเด็นข่าวได้ บ้างก็เป็นข่าวปลอมแปบที่คนเขียนไม่คิดอยู่แล้วว่าคนอ่านจะเชื่อแต่หวังยอดแชร์ บางข่าวก็ถูกมองว่าเป็นข่าวปลอมเพราะมาจากการมองต่างมุม ทำให้ตีความเป็นคนละแบบ [6]

Fake News คือข่าวที่ไม่เป็นจริง แวดวงการสื่อสารมวลชนต่างประเทศกำลังวิตกกังวลกันเป็นอย่างมาก เนื่องจากข่าวปลอมส่งผลกระทบทั้งการเมือง เศรษฐกิจ สังคม และสติปัญญาของคนในสังคม [2] จากงานวิจัยของ Reuter ในมหาวิทยาลัยออกซ์ฟอร์ด ได้เปิดเผยผลสำรวจความคิดเห็นจากผู้อ่านข่าวออนไลน์ในประเทศสหรัฐอเมริกา สหราชอาณาจักร สเปน และฟินแลนด์ พบว่ากลุ่มตัวอย่างได้นิยามความหมายของคำว่า “ข่าวปลอม” กว้างขวางขึ้นกว่าเดิมโดยไม่ได้จำกัดความหมายแค่เพียงข่าวผิด แต่หมายถึงการเขียนข่าวโฆษณาชวนเชื่อทางการเมือง ข่าวที่เลือกข้าง และข่าวที่ผู้สนับสนุนเพื่อจุดมุ่งหมายทางการเมืองด้วย [3]

2.1.2 ประเภทของข่าวปลอม

ไม่ว่าข่าวปลอมจะเกิดขึ้นโดยตั้งใจหรือไม่ก็ตาม ข่าวนั้นจะคงอยู่ภายในระบบนิเวศของข้อมูลที่ผิดพลาดและบิดเบือน รอการส่งต่อ แพร่กระจาย เพื่อสร้างอิทธิพลต่อความคิดเห็นของสารณชน หรือเพื่อปิดบังความจริง

ในการจำแนกประเภทของข่าวปลอม The European Association for Viewer Interests หรือ EAVI ซึ่งเป็นองค์กรที่ไม่แสวงผลกำไรที่สนับสนุนการรู้เท่าทันสื่อและการเป็นพลเมืองแบบสมบูรณ์ในยุโรป ได้ทำการศึกษาในหัวข้อ “Beyond Fake News – 10 Types of Misleading News” โดยแบ่งประเภทไว้ดังนี้ [18]

1. การโฆษณาชวนเชื่อ (Propaganda) มีเจตนาหลอกลวง เปลี่ยนความคิด มุมมอง ทำให้เกิดความรู้สึกคล้อยตาม เพื่อส่งเสริมบุคคล การกระทำหรือเหตุการณ์บางอย่าง บ่อยครั้งที่เนื้อหามุ่งสร้างความเกลียดชังและก่อความไม่สงบ

2. การยั่วให้คลิก (Clickbait) มีลักษณะเป็นเรื่องราวที่น่าตื่นเต้นหรือมีหัวเรื่องที่น่าสนใจ ข่าวลักษณะนี้มุ่งเน้นการ “คลิก” เพื่อเข้าชมเนื้อหา และเพื่อสร้างรายได้จากโฆษณา บ่อยครั้งที่เนื้อหาของข่าวมีลักษณะเกินจริงหรือพูดเท็จอย่างสิ้นเชิง

3. เนื้อหาที่ได้รับการสนับสนุน (Sponsored Content) มีลักษณะเป็นเรื่องราวที่สร้างขึ้นเพื่อประชาสัมพันธ์หรือการโฆษณาบนสื่อออนไลน์ ผู้ใช้สามารถเห็นข้อความได้ทันทีโดยไม่ต้องคลิกเพื่อเข้าชม จากการวิจัยของมหาวิทยาลัยสแตนฟอร์ดในปี 2016 พบว่านักเรียนมัธยมศึกษาร้อยละ 80 มีแนวโน้มที่เชื่อว่าข่าวปลอมที่อยู่ในรูปแบบนี้เป็นข่าวจริง

4. ข่าวล้อเลียนและเสียดสี (Satire and Hoax) ข่าวล้อเลียนที่เป็นเรื่องตลกที่นำเสนอในรูปแบบทั่วไป ใช้เนื้อหาที่ตลกขบขัน หรือเสียดสี ประชดประชัน เพื่อแสดงความคิดเห็นเกี่ยวกับเหตุการณ์ข่าวในโลกแห่งความเป็นจริง

5. ข่าวที่ผิดพลาด (Error) คือข่าวที่เผยแพร่จากสำนักข่าวที่เชื่อถือได้ แต่เกิดการผิดพลาดจากการสื่อสาร โดยคำพูด ภาพ หรือการกระทำ ที่ทำให้ผู้รับข่าวเข้าใจผิดไปในทิศทางอื่น หรือไม่เข้าใจในข่าวนั้น

6. ข่าวที่นำเสนอเอนเอียงเข้าข้างฝ่ายใดฝ่ายหนึ่ง (Partisan) เป็นข่าวที่บิดเบือนเนื้อหาหรือสร้างขึ้นมาเพื่อเข้าข้างหรือโจมตีฝ่ายตรงข้าม

7. ทฤษฎีสมคบคิด (Conspiracy Theory) คือข่าวที่อยู่ในรูปของเรื่องเล่า หรือบทความที่สร้างขึ้นมาจากความคิดของคนหรือกลุ่มคน โดยนำเหตุการณ์ต่าง ๆ ที่เกิดขึ้นมาปะติดปะต่อเข้าด้วยกัน มีวัตถุประสงค์ที่ซ่อนเร้นบางอย่างเพื่อให้เกิดประโยชน์หรือโทษต่อบุคคลหรือกลุ่มบุคคล หรือเพื่ออธิบายเหตุการณ์ที่เกิดขึ้น ลักษณะของทฤษฎีสมคบคิดโดยทั่วไปมีจะมีข้อเท็จจริงประกอบอยู่ด้วย ทั้งนี้ก็เพื่อเสริมให้เกิดความน่าเชื่อถือว่ามีหลักฐานสนับสนุนที่ดูเหมือนจะเกี่ยวข้องกัน แต่ข้อเท็จจริงที่นำมาประกอบมีเพียงเล็กน้อยเท่านั้น

8. วิทยาศาสตร์ลวงโลก (Pseudoscience) คือการแอบอ้าง หรือความเชื่อ หรือแนวทางการปฏิบัติที่บ่งบอกว่าเป็นวิทยาศาสตร์ แต่แท้จริงแล้วไม่ได้ผ่านกระบวนการทางวิทยาศาสตร์ที่ถูกต้อง ไม่มีหลักฐานหรือความเป็นไปได้มาสนับสนุน ข่าวประเภทนี้มีลักษณธเหมือนทฤษฎีสมคบคิด แต่มีความเป็นวิทยาศาสตร์ มักพบอยู่ในเว็บไซต์เกี่ยวกับสุขภาพหรืองานวิจัยทางวิทยาศาสตร์

9. ข่าวที่ให้ข้อมูลผิด ๆ (Misinformation) สามารถพบได้ในประเภทข่าวที่ชี้นำให้เกิดความเข้าใจผิด โดยใช้ข้อมูลที่ไม่ถูกต้องเพื่อผลประโยชน์ทางเศรษฐกิจหรือการเมือง

10. ข่าวปลอมที่ถูกสร้างขึ้นอย่างสมบูรณ์ (Bogus) คือข่าวปลอมที่เจตนาสร้างขึ้นเพื่อให้เป็นข่าวปลอมที่สมบูรณ์ อาจมีเนื้อหา ภาพ เสียง หรือข้อมูลที่เป็นเท็จมาประกอบกัน และอาจรวมถึงการแอบอ้างแหล่งข่าวที่น่าเชื่อถือหรือบุคคลที่อยู่ในเหตุการณ์

2.1.3 รูปแบบเนื้อหาของข่าวปลอม

การจำแนกรูปแบบเนื้อหาของข่าวปลอม สามารถแบ่งออได้เป็น 7 รูปแบบ คือ

1.เนื้อหาเลียนแบบ ล้อเลียน เสียดสี

2.มีเนื้อหาชี้นำ

3.มีเนื้อหาแอบอ้าง

4.เนื้อหาถูกประดิษฐ์ขึ้น

5.มีการเชื่อมโยงเนื้อหาที่ผิด

6.เนื้อหาที่ผิดบริบท

7.เนื้อหาที่หลอกลวง

ตารางที่ 2.1 การจำแนกการใช้เนื้อหาของข่าวปลอมตามวัตถุประสงค์ [21]

วัตถุประสงค์	**เนื้อหา**						
	ล้อเลียนหรือเสียดสี	ชี้นำ	แอบอ้าง	ประดิษฐ์ขึ้น	เชื่อมโยงเนื้อหาที่ผิด	ผิดบริบท	หลอกลวง
การสื่อสารที่ผิดพลาด		☐	☐	☐			
การล้อเลียนเสียดสี	☐				☐		☐
กระตุ้นหรือสร้างเรื่องเหลวไหล					☐	☐	☐
สร้างความหลงผิด				☐			
ฝักใฝ่ฝ่ายใดฝ่ายหนึ่ง			☐	☐			
ผลประโยชน์หรือผลกำไร		☐			☐		☐
อิทธิผลทางการเมือง			☐	☐		☐	☐
โฆษณาชวนเชื่อ			☐	☐	☐	☐	☐

นอกจากนี้ Thomas O' Duffy ได้นำเสนอการจำแนกข่าวปลอม โดยพิจารณาจากแหล่งที่มาของข่าว ดังนี้ [5][14]

1. บทความ หรือ เรื่องราวปลอม (Fake article or story) คือบทความปลอมที่แต่งขึ้นเองทั้งหมดนำเสนอในรูปแบบของบทความจริง

2. การอ้างอิงปลอม (Fake reference) คือบทความที่ไม่มีเจตนาจะปลอม แต่ใช้วิธีอ้างอิงปลอมเพื่อให้บทความเกิดความน่าเชื่อถือ

3. Fake Meme คือบทความที่ประกอบด้วยรูปภาพและคำพูด ได้รับความนิยมอย่างมากสำหรับการเผยแพร่ไวรัลซึ่งมักประกอบด้วยในกรณีนี้ข้อมูลที่มีข้อมูลเท็จ หรือปลอม

4. บุคลิกภาพปลอม (Fake Personality)

5. ตัวแทนปลอม (Fake representative) คือบุคคลที่แอบอ้างว่าเป็นตัวแทนขององค์กรใดองค์กรหนึ่ง เพื่อให้ได้รับความสนใจ หรือเพื่อทำให้องค์กรที่แอบอ้างนั้นเสื่อมเสียชื่อเสียง

6. เว็บเพจปลอม (Fake social page) คือหน้าเว็บเพจที่แอบอ้าง หรือแสดงตัวว่าเป็นตัวแทนของบุคคล ตราสินค้า หรือกลุ่มองค์กรที่ไม่มีอยู่จริง

7. เว็บไซต์ปลอม (Fake Website) คือเว็บไซต์ที่มีเนื้อหาปลอมทั้งเว็บไซต์

8. บทวิจารณ์ปลอม (Fake Review) คือโฆษณาที่เผยแพร่ทางสื่อสังคมออนไลน์ มีแรงจูงใจหรือมีเจตนาที่จะทำให้เกิดความรู้สึกโน้มเอียงไปในทิศทางที่ต้องการ

9. ภาพปลอม (Fake Portrayal) คือข้อมูลที่อยู่ในรูปแบบของภาพนิ่งหรือภาพเคลื่อนไหวที่แสดงถึงเหตุการณ์ หรือพฤติกรรมของบุคคลใดบุคคลหนึ่งที่ไม่ได้เกิดขึ้นจริง โดยภาพเหล่านี้ถูกอธิบายว่าเป็น "ภาพลวงตาปลอม"

10. ความจริงครึ่งเดียว (Half Truth) คือการนำเสนอ หรือรายงานข้อมูล โดยบอกความจริงเพียงครึ่งหนึ่งเพื่อหลอกลวงผู้รับสาร หรือจงใจทิ้งข้อเท็จจริงหรือบางส่วนเพื่อให้ข้อมูลดูน่าเชื่อถือ

2.1.4 สถานการณ์ข่าวปลอมที่พบในต่างประเทศ

สถานการณ์ข่าวปลอมที่พบในต่างประเทศมักเป็นเรื่องที่ส่งผลกระทบระดับประเทศ เช่น การเมืองภายในประเทศ ปรากฏการณ์สำคัญที่ข่าวปลอม หรือ Fake News ที่สร้างแรงสั่นสะเทือนไป ทั่วโลกก็คือ ชัยชนะในการเลือกตั้งประธานาธิบดีของนายโดนัลด์ ทรัมป์ เมื่อเดือนพฤศจิกายน ปี 2016 หลังนายโดนัลด์ ทรัมป์ ตัวแทนจากพรรครีพับลิกันคว้าชัยชนะเหนือนางฮิลลารี คลินตัน ตัวแทน จากพรรคเดโมแครต ในการเลือกตั้งประธานาธิบดีสหรัฐฯ

"เฟซบุ๊ก" ตกเป็นเป้าในการโจมตีอย่างหนัก หลังมีผลศึกษาออกมาว่าข่าวปลอมจากเว็บไซต์ปลอมนั้นถูกแชร์ให้ผู้คนเห็นมากกว่าข่าวจริงหรือจากเว็บไซต์ข่าวจริง ๆ บทความเว็บไซต์ BuzzFeed ชี้ว่าตัวเลขยอดการให้ความสนใจและมีสวนร่วมกับข่าวนั้นจาก 20 ข่าวปลอมในชวงสามเดือนก่อนการเลือกตั้งปี 2016 สูงถึง 8.7 ล้านครั้ง มากกว่าข่าวจริงที่มีจำนวน 7.3 ล้านครั้ง นอกจากนี้ 5 ข่าว ปลอมที่มียอดที่ผู้ใช้งาน

เฟซบุ๊กมีสวนร่วมสูงสุดในเฟซบุ๊กล้วนแล้วแต่เป็นข่าวโจมตี ฮิลลารี คลินตัน จากพรรคเดโมแครต ข่าวเหล่านี้ไม่เป็นความจริงแต่ถูกแชร์อย่างแพร่หลาย มีดังนี้

1. โป๊ปฟรานซิส ให้การรับรองโดนัลด์ ทรัมป์ (960,000 ครั้ง)
2. FBI ยืนยัน ฮิลลารี คลินตัน ขายอาวุธให้กลุ่ม ISIS (789,000 ครั้ง)
3. ข่าวอีเมลฮิลลารี คลินตัน รั่ว (754,000 ครั้ง)
4. ฮิลลารี คลินตัน ถูกตัดสิทธิ์จากสำนักงานรัฐบาลกลาง (701,000 ครั้ง)
5. พบศพเจ้าหน้าที่ FBI ที่ต้องสงสัยจากอีเมลรั่วไหลของฮิลลารี คลินตัน (567,000 ครั้ง)

นอกจากนี้การเกิดข่าวปลอมในชวงการเลือกตั้งประธานาธิบดีสหรัฐอเมริกา ยังได้สร้าง แนวคิดทางการเมืองแบบ Post-truth Politics คือการแสดงความคิดเห็นและการสนทนาเรื่อง การเมืองโดยใช้อารมณ์เป็นที่ตั้ง และไม่อยู่บนนโยบายทางการเมืองและข้อเท็จจริง จากกรณีข่าว ปลอมมีอิทธิพลต่อการเมืองสหรัฐอเมริกาทำให้เยอรมันเกิดความตื่นตัวในเรื่องข่าวปลอมบนเฟซบุ๊ก ขึ้นมา นักการเมืองเยอรมันให้ความสำคัญกับเรื่องนี้เป็นพิเศษ เนื่องจากปี 2017 เยอรมันจะมีการ เลือกตั้ง นักการเมืองหลายคนกลัวว่าข่าวปลอมจะแพร่กระจายในเฟซบุ๊กและส่งอิทธิพลต่อการเลือกตั้ง ดังที่มันสร้างอิทธิพลต่อการเลือกตั้งอเมริกันในปี 2016 มาแล้ว (สฤณี อาชวานันทกุล, 2560) แม้จำนวนข่าวปลอม และการบิดเบือนของข้อมูลในยุโรปจะยังไม่รุนแรงเท่าในสหรัฐฯ แต่ จิม วอเตอร์ สตัน (Jim Waterston) บรรณาธิการของ BuzzFeed ให้ความเห็นว่าข่าวปลอมในยุโรปนั้นแนบเนียน กว่า จึงเป็นอันตรายได้มากกว่าข่าวปลอมในสหรัฐฯ เขาได้จำแนกประเภทข่าวปลอมออกเป็น 3 ประเภท คือ

1. ข่าวที่ข้อมูลผิดทั้งหมด
2. ข่าวปลอมที่มีข้อมูลถูกบางสวน
3. ข่าวปลอมที่ตั้งใจบิดเบือนเนื้อหาจากความจริง

ซึ่งประเภทที่ 3 ข่าวปลอมที่ตั้งใจบิดเบือนเนื้อหาจากความจริง นั้นจะเป็นข่าวปลอมที่ถูกเชื่อ และแชร์ได้ง่ายมากในอังกฤษ และจะเป็นอันตรายมากกว่า เพราะข่าวพวกนี้จะจำแนกยากกว่าว่าเป็น ข่าวปลอมหรือจริง เพราะข่าวพวกนี้มีความคิดเห็นทางการเมืองแทรกอยู่ ซึ่งอาจไปถูกใจคนบางพวก บางกลุ่ม [16]

ข่าวปลอมที่เกิดขึ้นในต่างประเทศ มักจะพบว่าเป็นเรื่องการเมือง การเลือกตั้ง นอกจากนี้ยัง พบว่าข่าวปลอมยังสร้างปัญหาละเมิดสิทธิสวนบุคคลให้กับนานาประเทศ เช่น การแชร์ภาพข่าวปลอม วิกฤตโรฮิงญาทำให้เกิดสถานการณ์เลวร้ายในรัฐยะไข่ ชายแดนพม่าทำให้มีผู้เสียชีวิตมากกว่า 400 ราย จากการเผยแพร่เรื่องราวที่เป็นข่าวปลอมนี้อย่างแพร่หลายทำให้สถานการณ์เลวร้าย และเป็นส่วนหนึ่งที่ทำให้วาทกรรมความขัดแย้งในพม่าซับซ้อนมากยิ่งขึ้น

ภูมิภาคเอเชียตะวันออกเฉียงใต้ ผู้แทนสื่อมวลชนในสาขาต่าง ๆ ได้หยิบยกมาเป็นประเด็นที่พูดคุยกันอย่างกว้างขวาง ในบางประเทศให้ความสำคัญกับปัญหาข่าวลวง โดยออกกฎหมายเพื่อแก้ปัญหานี้ อย่างประเทศสิงคโปร์ได้ออกกฎหมายสำหรับป้องกันการแพร่ระบาดของข่าวปลอม โดยเน้นไปที่เว็บไซต์สื่อออนไลน์ต่าง ๆ ซึ่งสร้างความกังวลให้กับเสรีภาพของสื่อในสิงคโปร์ ทำให้การพิจารณากฎหมายดังกล่าวมีการถกกัน

อย่างกว้างขวางและยาวนานถึง 2 วัน นับเป็นการพิจารณากฎหมายที่ไม่ธรรมดาของประเทศสิงคโปร์ที่ใช้เวลามากขนาดนี้ ในที่สุดวันที่ 8 พฤษภาคม 2562 สิงคโปร์ผ่านกฎหมายกำกับข่าวลวง (The Protection from Online Falsehood and Manipulation) ให้อำนาจรัฐล้วงข้อมูลได้ โทษหนักทั้งคนแชร์ และแพลทฟอร์ม ส่วนในฟิลิปปินส์มีองค์กรที่ตรวจสอบว่าข่าวไหนเป็นข่าวจริงหรือข่าวเท็จ อย่าง Vera File อินโดนีเซีย มีองค์กร Fact Checking ที่มาใช้เพื่อตรวจสอบเรื่องการเมือง

สำหรับประเทศไต้หวัน มี “ ไต้หวันโมเดล” เป็นกลุ่มภาคประชาสังคมที่มีความสามารถทางด้านเทคโนโลยีมารวมตัวกันสร้างแชตบอตและโปรแกรมที่จะเข้าไปตรวจสอบข่าวลวงในแต่ละกลุ่ม โดยเฉพาะที่ส่งกันในกลุ่มญาติพี่น้อง หรือให้ผู้ใช้ส่งข้อความมาแล้วมีอาสาสมัครช่วยกันตรวจสอบแล้วส่งข้อเท็จจริงกลับไปให้ผู้ใช้

2.1.5 สถานการณ์ข่าวปลอมในประเทศไทย

ในประเทศไทยมีข่าวปลอมอยู่มากมาย มีทั้งข่าวปลอมที่สร้างขึ้นมาใหม่ทั้งหมด ข่าวปลอมที่บิดเบือนข้อเท็จจริงหรือให้ข้อมูลผิด ๆ ข่าวปลอมมักใช้เทคนิคให้คนสนใจ ไม่ว่าจะเป็นบ่อนทำลายหน่วยงาน บุคคล หวังผลทางการเมือง หรือผลประโยชน์ทางธุรกิจต่าง ๆ ข่าวปลอมบน เฟซบุ๊กที่พบมากในประเทศไทยช่วงปี 2558-2560 นอกจากข่าวปลอมที่ถูกสร้างขึ้นมาโดยสมบูรณ์ คือเว็บไซต์ประเภทคลิกเบท (Click bait) โดยใช้เทคนิคพาดหัวข่าวที่นาสนใจ โดยเฉพาะข่าวแปลก ๆ ที่ทำให้ตกใจ หรือประหลาดใจ โดยเว็บไซต์เหล่านี้จะมีหัวข้อ หรือพาดหัวข่าวให้นาสนใจ และจงใจ เขียนเพื่อดึงดูดให้เข้าไปอ่าน และอยากกดไลค์ กดแชร์ อยากคอมเมนท์ ไม่ว่าจะเป็น ‘อึ้ง! ทึ่ง! ตะลึง!’ และทิ้งท้ายข้อความไว้ว่า ‘เป็นเพราะสิ่งนี้...’ หรือ ‘เพราะทำแบบนี้...’ [5] ซึ่งข่าวปลอมประเภทนี้ก็จะมีรายได้จากคาโฆษณา ยิ่งมีผู้หลงเชื่อคลิกเข้าไปดูมากเท่าใดยิ่งมีรายได้เพิ่มมากขึ้นเท่านั้น เว็บไซต์ข่าวปลอมประเภทคลิก เบท มักจะไม่มีตัวตนจริง หรือ เลือกปกปิดผู้จดทะเบียนเว็บไซต์เพื่อไม่ให้ทางการจับได้เวลามีปัญหา การนำเสนอข่าวเชิงคลิกเบทมีการปรับตัวอยู่เสมอ มักปรับเปลี่ยนรูปแบบการนำเสนอข่าว เพื่อเรียกร้องความสนใจของผู้อ่านเป็นหลัก โดยการนำเสนอ ในรูปแบบของเทคนิคหรือรูปแบบการดำเนินชีวิต เพื่อให้ผู้อ่านแชร์ต่อในโซเชียลมีเดีย ในด้านจริยธรรมสื่อมวลชนถือว่าการนำเสนอข่าวเชิงคลิกเบทเป็นข่าวที่ไม่ค่อยมีความน่าเชื่อถือ นำเสนอข้อเท็จจริงให้กับผู้อ่านเพียงบางส่วนหรืออาจมีการบิดเบือนข่าวให้กับผู้อ่านข่าว ซึ่งผลให้ผู้อ่านจะได้รับรู้ข่าวที่ไม่ครบถ้วนรอบด้าน [9] และสิ่งที่คนทำเว็บคลิกเบท พัฒนาขึ้น มาอีกขั้น คือการปลอมชื่อเว็บไซต์ ให้ใกล้เคียงกับเว็บไซต์สำนักข่าวต่าง ๆ โดยใช้กลลวงในการเปลี่ยน ชื่อลิงก์เว็บไซต์ URL สร้างเว็บไซต์ใหม่ในชื่อใกล้เคียงกัน จากนั้นก็ปลอมหน้าเว็บไซต์ให้เหมือนกับเว็บไซต์สำนักข่าวจริงจนถึงเว็บไซต์หน่วยงานต่าง ๆ เพื่อให้ผู้อ่านที่อ่านข่าวไม่ละเอียด หรือไม่ได้ ตรวจสอบให้ชัดเจนหลงเชื่อ [5] และเนื้อหาข่าวปลอมพวกนี้จะใช้วิธีการจับคำค้นหา (Keyword) ที่คนสนใจชวงนั้น ๆ มารวมกันในข่าวเดียว

จากการตรวจสอบข่าวปลอมและข้อเท็จจริงบนโลกออนไลน์ของศูนย์ชัวร์ก่อนแชร์ สำนักข่าว ไทย อสมท. [7] ตั้งแต่เดือนมีนาคมปี 2558 เพื่อแก้ปัญหาการแพร่กระจายของข้อมูลไม่ถูกต้อง และได้ ช่วยทำ

ความเข้าใจเกี่ยวกับข้อมูลเท็จ พบว่าข้อมูลเท็จบนโลกออนไลน์ที่ประชาชนสอบถามมา สวนมากมักเป็นเรื่องสุขภาพโดยเฉพาะกลุ่มโรคไม่ติดต่อเรื้อรัง เพราะผู้ผลิตเรื่องเท็จเหล่านี้ มองเห็น พฤติกรรมของผู้อ่านที่ขาดทักษะการตรวจสอบข้อมูล โดยเฉพาะกลุ่มผู้อ่านที่ค่อนข้างสูงอายุที่เข้าสู่โลกอินเทอร์เน็ตตอนอายุมากอาจจะขาดเรื่องความเท่าทันสื่อดิจิทัล สิ่งที่อันตรายกว่าการส่งต่อข้อมูล เท็จโดยรู้เท่าไม่ถึงการณ์ คือข้อมูลเท็จในลักษณะข่าวปลอม ที่ปลอมหน้าเว็บข่าวด้วยเจตนาหลอกลวง ตลอดปี 2559 สามารถเก็บข้อมูลได้ว่ามีข่าวปลอมรวมกว่า 300 หัวข้อ และพบว่าแต่ละหัวข้อมีการ ไลค์และแชร์บนเฟซบุ๊กรวมกันอยู่ในหลักแสน จากการสืบค้นของพบว่าเว็บไซต์ข่าวปลอมเหล่านี้มีอยู่ ไม่น้อย เมื่อตามรอยเว็บไซต์ข่าวปลอมขนาดเล็กแห่งหนึ่งได้พบว่ามีคนหลงคลิกเข้าไปชมเป็นจำนวน หลักหมื่น เว็บไซต์เหล่านี้มีความพยายามควบคุมบัญชีเฟซบุ๊กของผู้หลงเข้าไปเพื่อไปใช้ประโยชน์ทางการค้า

หากดูสถิติเกี่ยวกับปัญหาข่าวปลอมของไทย ศูนย์สำรวจความคิดเห็นบ้านสมเด็จโพลล์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏบ้านสมเด็จเจ้าพระยา [11] ได้เก็บจากกลุ่มตัวอย่างจากประชาชนที่อาศัยอยู่ในกรุงเทพมหานคร จำนวนทั้งสิ้น 1,211 กลุ่มตัวอย่าง เมื่อวันที่ 20-23 กุมภาพันธ์ 2562 พบว่ากลุ่มตัวอย่างส่วนใหญ่เคยพบเห็นข่าวปลอม ร้อยละ 85.1 และเคยตรวจสอบข้อมูลข่าวสารที่คิดว่าเป็นข่าวปลอม ร้อยละ 65.8 และเคยพบเห็นข่าวปลอมในลักษณะข่าวที่ตั้งใจให้เกิดความเข้าใจผิด มากที่สุด ร้อยละ 31.3 อันดับที่สองคือข่าวที่นำภาพปลอมหรือภาพที่ไม่เกี่ยวข้องนำมาเป็นภาพประกอบ ร้อยละ 17.8 อันดับที่สามคือข่าวที่มีการตัดต่อภาพ และข้อมูลในข่าวที่ไม่มีความจริงใด ๆ ร้อยละ 16.8 อันดับที่สี่คือข่าวปลอมแบบล้อเลียนขำขัน ร้อยละ 15.2 และอันดับที่ห้าคือข่าวที่นำคำพูดของบุคคลที่ไม่ได้พูดจริงมาอ้างถึง ร้อยละ 7.6

กลุ่มตัวอย่างส่วนใหญ่ใช้วิธีการในการตรวจสอบข้อมูลข่าวสารที่คิดว่าเป็นข่าวปลอมโดยดูแหล่งที่มา / ผู้เขียน ร้อยละ 33.1 อันดับที่สองคือไม่คิดจะตรวจสอบ ร้อยละ 26.8 อันดับสามคือค้นหาจากแหล่งข่าวที่น่าเชื่อถือ ร้อยละ 17.0 อันดับที่สี่คือตรวจสอบวันเวลาที่เผยแพร่ ร้อยละ 9.6 และอันดับที่ห้าคือตรวจสอบว่าเป็นการตัดต่อ ร้อยละ 8.2

การพบเห็นข่าวปลอมกลุ่มตัวอย่างส่วนใหญ่พบผ่านสื่อเฟซบุ๊ก (Facebook) เป็นอันดับหนึ่ง ร้อยละ 72.7 อันดับสอง คือ มีคนเล่าให้ฟัง ร้อยละ 10.3 และอันดับสาม คือ ผ่านสื่อไลน์ (Line) ร้อยละ 8.8 และพบเห็นข่าวปลอมที่มีเนื้อหาเกี่ยวกับประเด็นเรื่องการเมืองเป็นอันดับหนึ่ง ร้อยละ 28.2 อันดับสองคือประเด็นเรื่องดารา ร้อยละ 26.9 อันดับสามคือประเด็นเรื่องหลอกขายสินค้า ร้อยละ 17.3 อันดับที่สี่คือประเด็นเรื่องสุขภาพ ร้อยละ 15.6 อันดับที่ห้าคือประเด็นเรื่องภัยพิบัติ ร้อยละ 7.3 และอันดับสุดท้ายคือประเด็นเรื่องศาสนา ร้อยละ 4.6

กลุ่มตัวอย่างส่วนใหญ่คิดว่าปัญหาข่าวปลอมเป็นปัญหาที่ควรมีการแก้ไขอย่างเร่งด่วน ร้อยละ 84.5 และอยากให้ภาครัฐมีมาตรการในการป้องกันและปราบปราบ ข่าวปลอม (Fake News) ร้อยละ 85.0

2.1.6 องค์ประกอบที่ทำให้หลงเชื่อข่าวปลอมบนสื่อสังคมออนไลน์

แนวคิดเกี่ยวกับการการเผยแพร่ข่าวปลอมบนสื่อสังคมออนไลน์อาจส่งผลให้ผู้ใช้มีแนวโน้มที่จะรับข่าว ปลอมและสับสนระหว่างสิ่งที่เป็นจริงกับสิ่งที่ไม่เป็นจริง ความสับสนนี้อาจทวีความรุนแรงขึ้นจาก หลายองค์ประกอบดังนี้ [10]

1. ความอคติหรือเอนเอียงเพื่อยืนยันความคิดของตนเอง (Confirmation Bias) คือ การมีแนวโน้มที่จะหลีกเลี่ยงข้อมูลที่ไม่เห็นด้วยและหาเนื้อหาที่ยืนยันความเชื่อของเราที่มีมาแต่เดิมเลือกที่จะเชื่อข้อมูลที่ตรงกับความคิดหรือหาเหตุผลมาสนับสนุนความคิดของตนเองที่จะเชื่อ

2. อัลกอริทึมสื่อสังคมออนไลน์ (Social Media Algorithms) การจัดเรียงเนื้อหาที่จะนำมา เสนอแก่ผู้ใช้ เช่น อัลกอริทึมเฟซบุ๊กจะมีการแสดงผลของนิวส์ฟีดอิงจากความสนใจของผู้ใช้เป็นหลัก หากเนื้อหาได้รับความความสนใจหรือมีปฏิสัมพันธ์ก็จะนำเสนอข้อมูลนั้นหรือข้อมูลที่เกี่ยวข้อง ใกล้เคียงเสมอ

3. ห้องเสียงสะท้อนของสื่อ (Echo chamber) เป็นคำเปรียบเทียบถึงห้องที่ออกแบบให้มี การสะท้อนเสียงกลับไปมา หมายถึงสถานการณ์ที่ข้อมูล ความคิด ความเชื่อหนึ่ง ๆ ถูกขยายหรือถูก สนับสนุนผ่านการสื่อสารและการทำซ้ำภายในระบบหนึ่ง ๆ ใน “ห้อง” นี้ แหล่งที่มาของข้อมูลนั้น ๆ มักไม่ถูกตั้งคำถาม มุมมองที่แตกต่างหรือท้าทายต่อข้อมูลเดิมมักถูกปิดกั้น หรือถูกนำเสนอน้อยกว่า ข้อมูลที่สอดคล้องกัน ในสื่อสังคมออนไลน์มักจะถูกป้อนข่าวสารในเฉพาะสิ่งที่บุคคลสนใจ หรือมี ความคิดเห็นไปในทิศทางเดียวกัน และคนมักจะเปิดรับข้อมูลที่ถูกจริตของตนและปิดกั้นชุดข้อมูลที่ ขัดความรู้สึก ซึ่งเชื่อมโยงกับหลักทฤษฎีการเลือกรับ (Selective Exposure) และความไม่ลงรอยกัน ของความคิด (Cognitive Dissonance) หากไม่มีการถกเถียงแลกเปลี่ยนข้อมูลมาก ๆ อาจทำให้ผู้รับ สารคิดว่าคนสวนใหญ่ในสังคมคิดเหมือนกันกับตน จึงเป็นเสียงสะท้อนเฉพาะสิ่งที่ผู้รับสารอยากได้ยิน

4. ผลของสมัยนิยมหรือการเห็นตามคนหมู่มาก (Bandwagon Effect) เป็นปรากฏการณ์คนคิดเชื่อ หรือกระทำตามคนหมู่มากไม่ว่าสิ่งที่กำลังตามนั้นถูกหรือผิด แต่ขอให้อยู่ในกระแสนิยม เช่น เมื่อคนจำนวนมากในเฟซบุ๊กเชื่อข่าวใดก็จะเชื่อข่าวนั้นตามไปด้วย หรือ มีความคิดเห็นไปใน ทิศทางใดก็พร้อมที่จะมีความคิดเห็นไปในทิศทางนั้นเช่นกันโดยขาดการพิจารณาอย่างรอบคอบ

2.2 ทฤษฎีการเรียนรู้ของเครื่อง (Machine Learning)

การศึกษาด้านการเรียนรู้ของเครื่อง มีวัตถุประสงค์เพื่อให้โปรแกรมคอมพิวเตอร์สามารถพัฒนาความสามารถของตัวเองโดยอัตโนมัติ จากตัวอย่างที่มีผู้สอน (Supervised Learning) ป้อนให้ ตัวอย่างเช่น โปรแกรมคอมพิวเตอร์ที่สามารถเรียนรู้เพื่อเล่นเกมส์หมากรุก สามารถพัฒนาความสามารถในการเล่นซึ่งวัดจากเปอร์เซ็นต์ในการแข่งขันชนะคู่แข่งขันซึ่งเป็นมนุษย์โดยใช้ตัวอย่างจากการเล่นกับตัวเอง หรือการสร้างโปรแกรมที่สามารถรู้จำภาพตัวเขียนตัวอักษรได้โดยวัดความสามารถในการรู้จำจากเปอร์เซ็นต์ของภาพตัวอย่างอักษรที่รู้จำได้อย่างถูกต้องโดยใช้ตัวอย่างเป็นภาพตัวอักษรซึ่งผู้สอนป้อนให้

ได้มีการนำวิธีการเรียนรู้ของเครื่องไปใช้เป็นส่วนประกอบสำคัญในการสร้างโปรแกรมเพื่อใช้งานทั่วไป รวมถึงการประยุกต์ใช้ในงานด้านต่าง ๆ เช่น การสร้างโปรแกรมประยุกต์เพื่อประโยชน์ในด้านธุรกิจการอนุมัติบัตรเครดิต การค้นพบความรู้จากฐานข้อมูล การรู้จำเสียงพูด (Speech Recognize) การรู้จำภาพ (Image Recognition) การทำนายอัตราการฟื้นตัวของผู้ป่วยโรคปอดอักเสบ การสืบหากรณีที่เจ้าของบัตรเครดิตชำระเงินไม่ตรงเวลา การบังคับรถโดยอัตโนมัติบนถนนหลวง หรือการสร้างกลยุทธ์ต่าง ๆ ในการเล่นเกมส์ เป็นต้น มีงานวิจัยจำนวนมากได้พัฒนาวิธีการเรียนรู้ของเครื่องให้มีประสิทธิภาพมากยิ่งขึ้น เช่น การหาความสัมพันธ์พื้นฐานของตัวอย่างที่ป้อนให้เพื่อทำการเรียนรู้ การหาจำนวนสมมติฐานที่เหมาะสมในการนำมาพิจารณา การทำนายความผิดพลาดของสมมติฐานต่าง ๆ การสร้างแบบจำลองการเรียนรู้ของมนุษย์และสัตว์ รวมถึงการทำความเข้าใจความสัมพันธ์ของการเรียนรู้ของมนุษย์กับสัตว์ เป็นต้น

2.2.1 เทคนิคการเรียนรู้เครื่อง (Machine Learning)

การเรียนรู้เครื่องคือการทำให้เครื่องเรียนรู้จากข้อมูลตัวอย่างหรือสภาพแวดล้อม จุดมุ่งหมายคือการปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น เมื่อเรียนรู้แล้วจึงจัดเก็บข้อมูลไว้ในฐานความรู้ด้วยรูปแบบอย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชั่น แบ่งออกเป็นแบบมีผู้สอน (Supervised Leaning) และแบบไม่มีผู้สอน (Unsupervised Leaning) โดยในงานวิจัยชิ้นนี้ใช้แบบมีผู้สอน ประกอบด้วยเทคนิคการถดถอยโลจิสติก นาอีฟเบส์ ซัพพอร์ทเวกเตอร์แมชชีน และตัวจำแนกป่าแบบสุ่ม

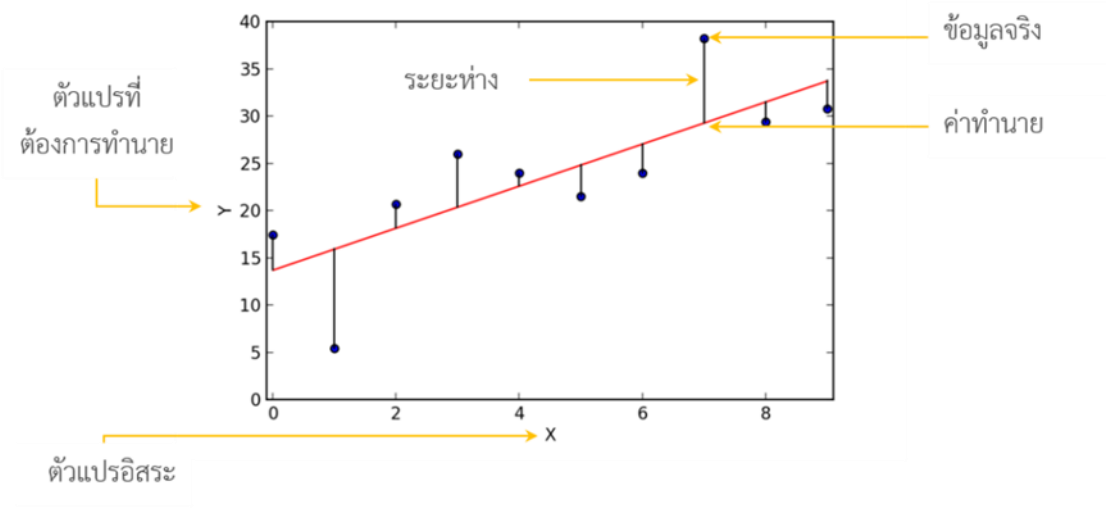
2.2.1.1 การถดถอยโลจิสติก (Logistic Regression) [8] เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุประเภทหนึ่ง มีวัตถุประสงค์ในการหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วนเพื่อประมาณหรือทำนายความน่าจะเป็นของการเกิดหรือไม่เกิดเหตุการณ์ที่กำลังสนใจ มีงานวิจัยจำนวนมากที่ใช้เทคนิคนี้ในการวิเคราะห์ข้อมูลและแสดงให้เห็นถึงความมั่นใจและความน่าเชื่อถือ

การถดถอยโลจิสติก (Logistic Regression) เป็นการพยายามฟิตเส้นโค้งซิกมอยด์ (Sigmoid Curve) ให้เข้ากับข้อมูลจริงมากที่สุดเท่าที่ทำได้ในหลักการที่คล้ายกับการวิเคราะห์การถดถอยเชิงเส้น (Linear Regression) ซึ่งเป็นการพยายามฟิตสมการเส้นตรงเข้ากับข้อมูลจริง สิ่งที่แตกต่างคือข้อมูลที่ทำนายจากการทำ Logistic Regression นั้นเป็นการทำนายระหว่างค่า 1 (ในเกณฑ์) และ ค่า 0 (นอกเกณฑ์) ในขณะที่การทำนายจากการทำ Linear Regression นั้นค่าที่ได้จากการทำนายอาจเป็นค่าตัวเลขใด ๆ ก็ได้ เพื่อให้ผลจากการทำนายโดย Logistic Regression มีค่าเป็น 0 หรือ 1 เท่านั้น ตัวแปรที่ส่งผลต่อการทำนายจะถูกคำนวณให้กลายเป็นค่าความน่าจะเป็น (Probability) ว่าค่านั้น ๆ มีความเป็นไปได้ที่จะเป็น 1 เท่าใด โดยสมการดังกล่าวสามารถเขียนได้ในรูปของสมการ

$$P(1) = \frac{e^{f(x,\beta)}}{1 + e^{f(x,\beta)}}$$

เมื่อ x คือเซ็ตของตัวแปรที่ส่งผลต่อการทำนาย, \beta คือเซ็ตของพารามิเตอร์ของโมเดลทำนาย, f(x) เป็นฟังก์ชันเชิงเส้นในรูปของ f(x)=\beta_0+\beta_1 x_1 + \beta_2 x_2 + ... และ P(1) คือค่าความน่าจะเป็นที่ค่าตัวแปรนั้นจะถูกทำนายให้เป็น 1

ในขณะที่ผลจากการทำนายด้วย Linear Regression สามารถเขียนแทนได้ด้วยเส้นตรง (สำหรับกรณีทำนายผลจากหนึ่งตัวแปร) ผลรวมระยะห่างระหว่างค่าทำนายกับค่าของข้อมูลจริงบ่งชี้ความแม่นยำของการทำนายดังกล่าว เพื่อหาเส้นตรงที่เหมาะสม เราอาจสร้างเส้นตรงหลาย ๆ เส้นก่อนจะเฟ้นหาเส้นตรงที่ให้ผลรวมระยะห่างที่ต่ำที่สุด ซึ่งจะถือเป็นโมเดลที่ดีที่สุดที่ใช้อธิบายข้อมูลชุดนั้น ๆ

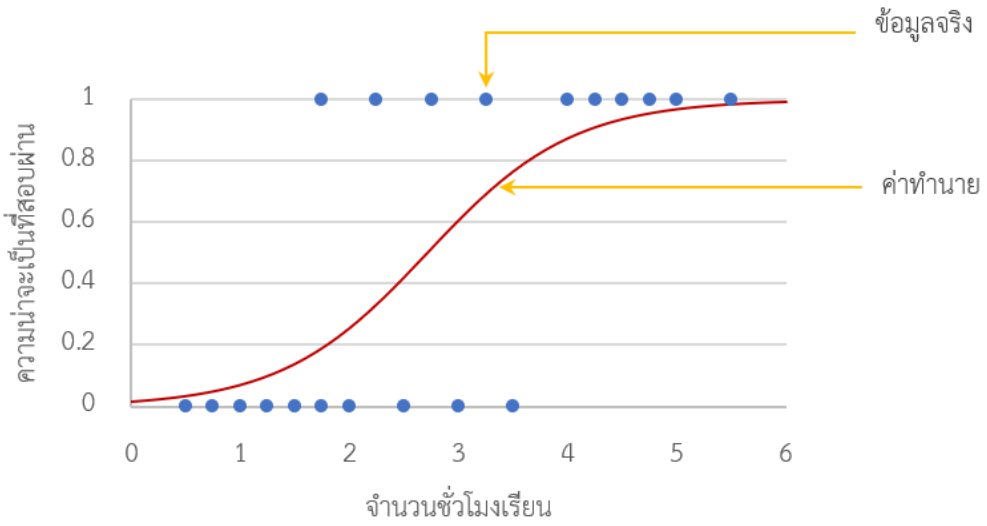


ภาพที่ 2.1 Linear Regression วัดค่าความแม่นยำจากผลรวมของระยะห่างระหว่างข้อมูลจริงกับค่าที่ทำนายจากโมเดล

อย่างไรก็ดีเมื่อใช้หลักการดังกล่าวกับการทำนายด้วย Logistic Regression พบว่าค่าผลรวมที่ได้จากการเทียบระยะห่างระหว่างเส้นโค้งซิกมอยด์กับข้อมูลจริงสามารถหาจุดที่ต่ำที่สุดได้ยาก ในทางปฏิบัติเราจะคำนวณหาค่าผลคูณความน่าจะเป็น ซึ่งสะท้อนความศักยภาพของโมเดลในการฟิตเข้ากับตัวข้อมูล (Likelihood) แทนผลรวมระยะห่าง โมเดลเส้นโค้งซิกมอยด์ที่ให้ค่าความเป็นไปได้สูงสุด (Maximum Likelihood) จะถูกเลือกให้เป็นโมเดลที่ดีที่สุดที่ใช้อธิบายข้อมูลชุดนั้น ในทางปฏิบัติเรานิยมคำนวณค่าล็อกการิธึมของผลคูณความน่าจะเป็น (Log Likelihood) โดยสมการผลคูณความน่าจะเป็นจะถูกเปลี่ยนให้อยู่ในรูปของผลรวม ซึ่งสามารถคำนวณหาค่าสูงสุดผ่านการหาอนุพันธ์ (Differentiation) ได้ง่ายกว่าการหาอนุพันธ์ของผลคูณ ผลบวกดังกล่าวอยู่ในรูปของสมการ

$$LL = \log(Likelihood) = \sum_i \begin{cases} \log\left(\frac{e^{(x_i^T \beta)}}{1 - e^{(x_i^T \beta)}}\right), y_i = 1 \\ \log\left(1 - \frac{e^{(x_i^T \beta)}}{1 - e^{(x_i^T \beta)}}\right), y_i = 0 \end{cases}$$

โดยพจน์แรกว่าด้วยความเป็นไปได้ที่จุด *i* ใด ๆ ที่ทราบมาก่อน (labelled) ว่าเป็น 1 นั้น จะถูกโมเดลทำนายว่าเป็น 1 ในขณะที่พจน์ที่สองว่าด้วยความเป็นไปได้ที่จุด *i* ใด ๆ ที่ทราบมาก่อนว่าเป็น 0 นั้นจะถูกโมเดลทำนายว่าเป็น 0 ด้วยเหตุนี้ผลรวมของสองพจน์จึงเป็นผลรวมความเป็นไปได้สำหรับกรณีที่โมเดลทายถูก (1 หรือ 0) ทั้งหมด โมเดลที่มีค่า \beta ที่สร้างความเป็นไปได้ที่สูงที่สุด (Maximum Likelihood) จึงถูกเลือกให้เป็นโมเดลที่เป็นตัวแทนของข้อมูลชุดนั้น



ภาพที่ 2.2 ตัวอย่างการใช้ Sigmoid Curve เพื่อใช้ทำนายความน่าจะเป็นที่จะสอบผ่านเนื่องจากจำนวนชั่วโมงเรียน

2.2.1.2 นาอีฟเบส์ (Naïve Bayes : NB) [18] เป็นขั้นตอนวิธีที่ได้รับความนิยมและถูกนำมาใช้อย่างแพร่หลายในงานจำแนกหมวดหมู่เอกสาร เนื่องจากความเรียบง่ายของขั้นตอนวิธีและให้ประสิทธิภาพการจำแนกที่ดี นาอีฟเบส์เป็นขั้นตอนวิธีที่มีพื้นฐานมาจากทฤษฎีเบส์ (Bayes' Theorem) ซึ่งอาศัยหลักความ

น่าจะเป็นในการทำนายผลลัพธ์ โดยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์ สำหรับรูปแบบการคำนวณความน่าจะเป็นของนาอีฟเบย์สสามารถคำนวณได้จากสมการ

$$p(Class_j|Topics_i) = \frac{p(Class_j)\times p(Topics_i|Class_j)}{p(Topics_i)}$$

โดยที่ $p(Class_j|Topics_i)$ ความน่าจะเป็นที่หัวข้อ *(Place)* ที่ *i* จะอยู่ในหมวดหมู่ *(Class)* ที่ *j* เมื่อ $1 \le I \le n$, $1 \le j \le 5$ และ *n* คือ จำนวนหัวข้อทั้งหมด

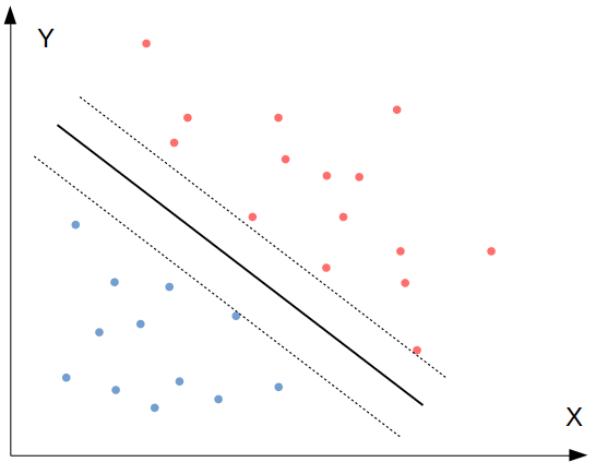
$p(Class_j)$ ความน่าจะเป็นของหมวดหมู่ *(Class)* ที่ *j*

$p(Topics_i)$ ความน่าจะเป็นของหัวข้อ *(Topics)* ที่ *i*

$p(Topics_i|Class_j)$ ความน่าจะเป็นที่คุณลักษณะ (f_{1-n}) ของหัวข้อ *(Topics)* ที่ *i* ปรากฏในหมวดหมู่ *(Class)* ที่ *j* สามารถคำนวณได้จากสมการ

$$p(Topics_i|Class_j) = p(f_1, f_2, \dots, f_n|Class_j) = \prod_{k=1}^{n}(f_k|Class_j)$$

2.2.1.3 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) เป็นวิธีการจำแนกกลุ่มข้อมูลที่อาศัยระนาบการตัดสินใจมาใช้ในการแบ่งข้อมูลออกเป็น 2 ส่วน โดยพยายามสร้างเส้นแบ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตทั้งสองกลุ่มมากที่สุด ซึ่งซัพพอร์ทเวกเตอร์แมชชีนจะใช้ฟังก์ชั่นแม็พพิ้ง (Mapping Function) เพื่อแปลงข้อมูลจากโดเมนเดิมไปยังโดเมนที่เรียกว่า ฟีเจอร์ สเปซ (Feature Space) และใช้ฟังก์ชั่นเคอร์เนล (Kernel Function)ในการวัดความคล้ายกันของข้อมูลในฟีเจอร์ สเปซ ในงานวิจัยนี้ใช้เคอร์เนลฟังก์ชั่น คือ โพลิโนเมียล เคอร์เนล (Polynomial Kernel) เนื่องจากการเป็นวิธีที่ดีที่สุด



ภาพที่ 2.3 ตัวอย่างระนาบการตัดสินใจของซัพพอร์ทเวกเตอร์แมชชีน

ที่มา: (Ali, Shamsuddin and Ismail, 2011)

จากภาพเป็นปัญหา Binary classification ที่จำแนกข้อมูลออกเป็นสองพวก คือสีน้ำเงินและสีแดง สิ่งที่ SVM ทำ คือการหาเส้นแบ่งการตัดสินใจที่เป็นเส้นทึบ ซึ่งเส้นนี้จะเกิดขึ้นระหว่างกลางของเส้นประด้านซ้ายและขวา โดยมีเงื่อนไขว่าจะต้องหาคู่ของเส้นประที่กว้างที่สุดเท่าที่จะเป็นไปได้ โดยคู่ของเส้นประที่กว้างที่สุดเท่าที่จะเป็นไปได้นี้ จะมีสองแบบ คือ 1) Hard margin classification คือคู่เส้นประที่ห้ามไม่ให้มีจุดข้อมูลอยู่ในพื้นที่ระหว่างเส้นประ และ 2) Soft margin classification คืออนุญาตให้มีข้อมูลอยู่ในพื้นที่ระหว่างเส้นประได้บ้าง

SVM ใช้ Hypothesis function แบบเส้นตรง เหมือนกับ Linear regression นั่นคือ

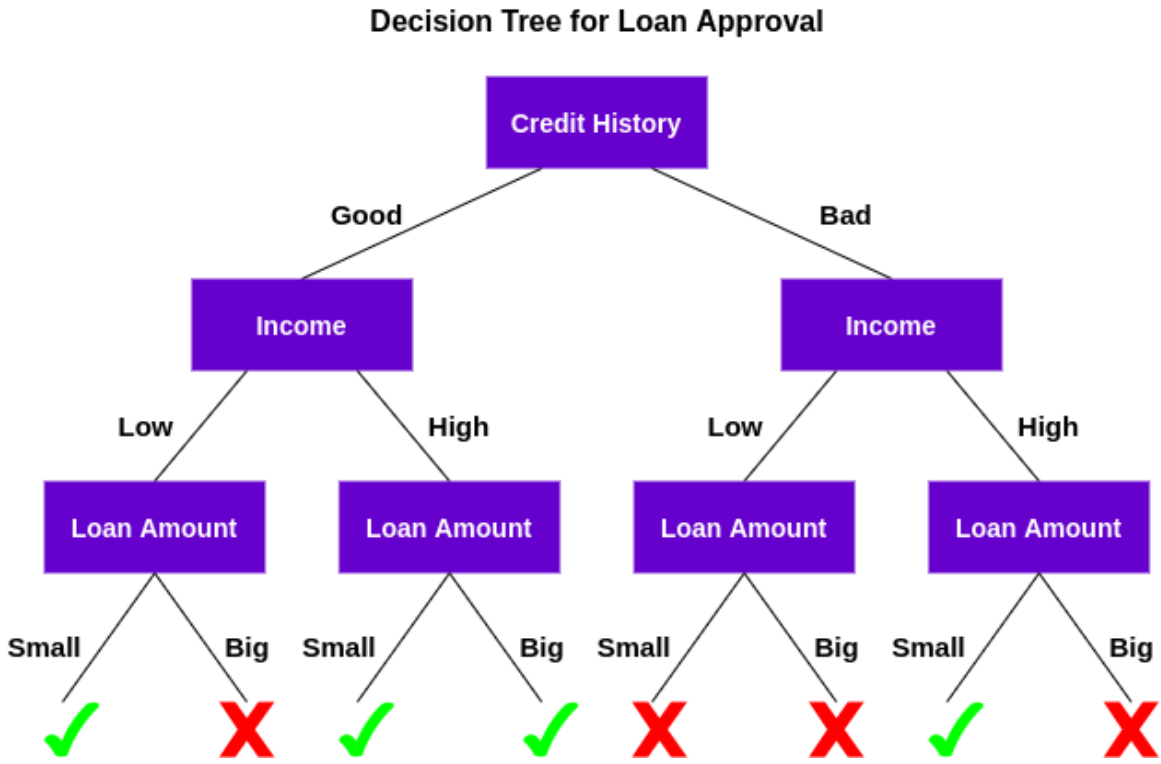
$$h_\theta(x) = w_1x_1 + w_2x_2 + \cdots w_nx_n + b$$
$$= w^Tx + b$$

โดยถ้าผลลัพธ์เป็นบวก จะทำนาย Class $\hat{y}$ ว่าเป็น 1 ส่วนถ้าเป็นลบ ทำนายว่าเป็น 0 จะสามารถเขียนวิธีการตัดสินใจตามเงื่อนไขดังกล่าวได้ดังนี้

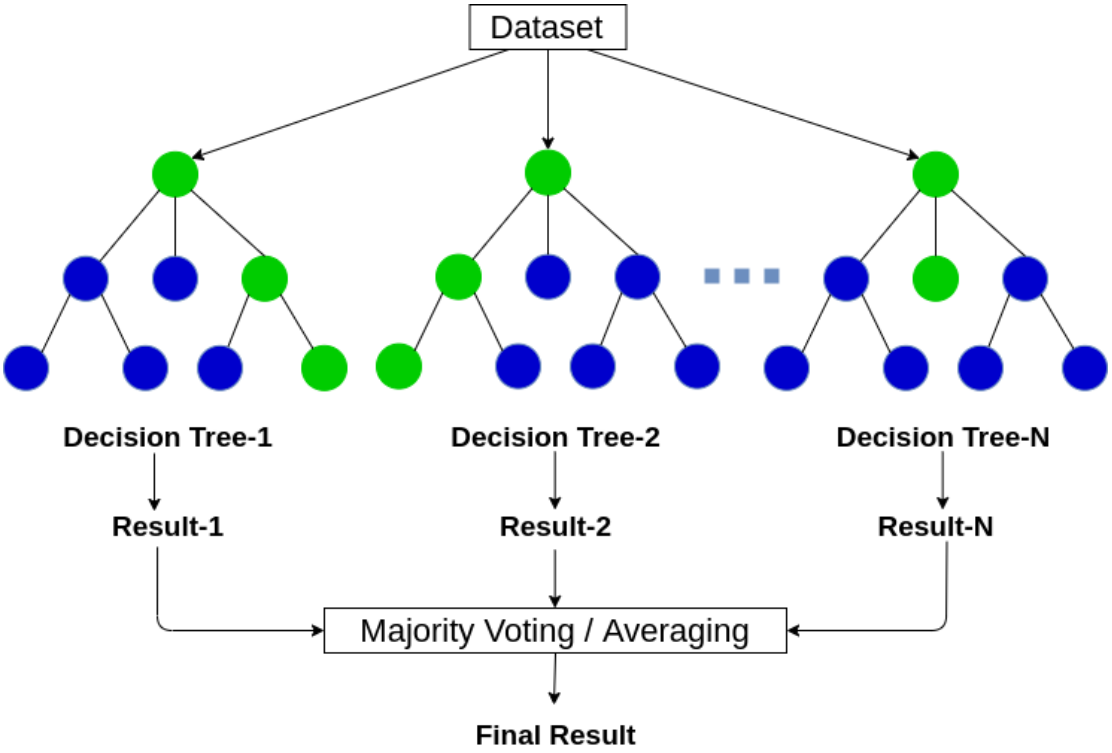
$$\hat{y} = \begin{cases} 0 \; if \; w^Tx + b < 0, \\ 1 \; if \; w^Tx + b \leq 0 \end{cases}$$

เมื่อนิยามเส้นแบ่งการตัดสินใจแล้ว (เส้นทึบ) จะสามารถกำหนดเส้นประทั้งสองด้านของเส้นทึบ โดยเส้นประแต่ละด้านคือตำแหน่งที่ $h_\theta(x)$ เท่ากับ *-1* และ *1* และเมื่อพิจารณาว่า ความชันของฟังก์ชันการตัดสินใจ เท่ากับ *Norm* ของ *Vector* ค่าน้ำหนัก *w*

2.2.1.4 ตัวจำแนกป่าแบบสุ่ม (Random Forest) คือการสร้างแบบจำลองด้วยการนำแนวคิดโมเดลต้นไม้ตัดสินใจ (Decision Tree) เข้ามาใช้ โดยจะทำการสร้างโมเดล Decision Tree ขึ้นมาหลาย ๆ โมเดลโดยการสุ่มตัวแปร แล้วนำผลที่ได้แต่ละโมเดลมารวมกันพร้อมกับคำนวณผลที่มีจำนวนซ้ำกันมากที่สุดสกัดออกมาเป็นผลลัพธ์สุดท้าย วิธีการของโมเดลต้นไม้ประกอบไปด้วยโหนด (Node) และกิ่ง (Branch) แต่ละโหนดจะถูกแทนด้วยคุณลักษณะ (Feature) ของชุดข้อมูลที่นำมาเรียนรู้และทดสอบ แต่ละกิ่งของต้นไม้แสดงผลในการในการทดสอบ และลีฟโหนด (Leaf Node) แสดงหมวดหมู่ที่ผู้ใช้กำหนด ส่วนเกณฑ์การเลือกคุณลักษณะเพื่อนำมาเป็นโหนดของต้นไม้นั้นมาจากการคำนวณค่าเกนสารสนเทศ (Information Gain) โดยพิจารณาคุณลักษณะที่มีค่าเกนสารสนเทศหรือมีค่าเอ็นโทรพี (Entropy) ต่ำ หมายความว่าคุณลักษณะนั้นมีความสามารถในการจำแนกหมวดหมู่สูง



ภาพที่ 2.4 กระบวนการทำงานของ Decision Tree Algorithm



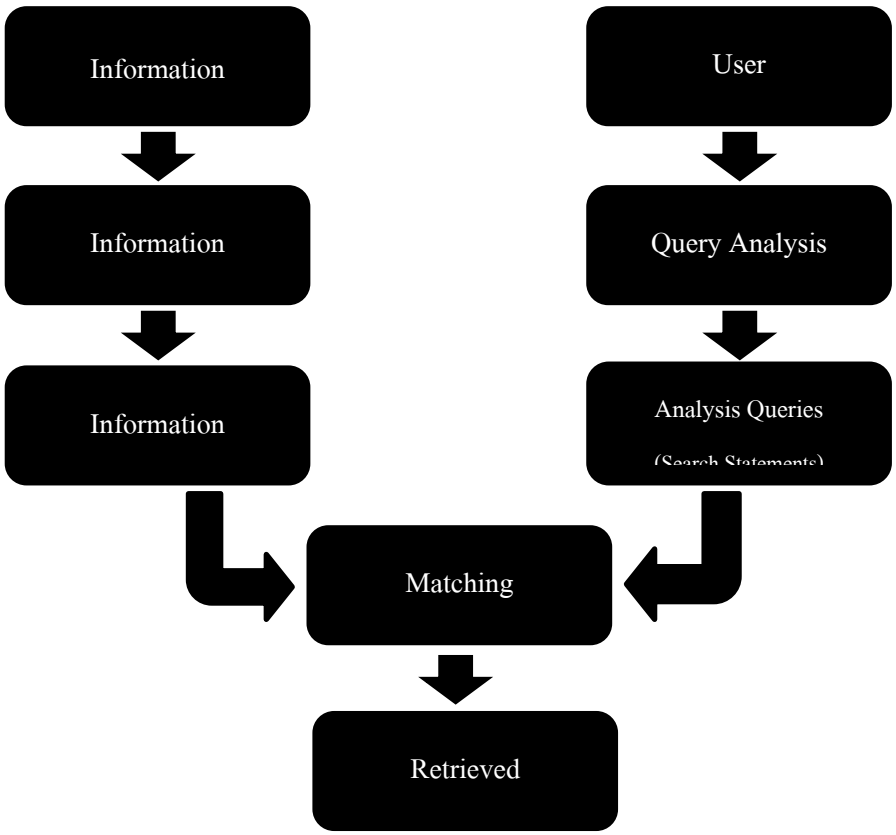
ภาพที่ 2.5 กระบวนการทำงานของ Random Forest Algorithm

ในขั้นตอนการทำงานของ Random Forest จะทำการจำแนกต้นไม้หลาย ๆ ต้น ซึ่งในแต่ละต้นมีการแบ่งเป็นคลาส โดยที่ต้นไม้แต่ละต้นจะถูกสร้างขึ้นจากกลุ่มตัวอย่างที่แตกต่างกันจากกระบวนการของ Decision Tree ถูกสร้างขึ้นหลายกระบวนการจนกลายเป็นป่า (forest) จนกระทั่งวิเคราะห์การตัดสินใจจากต้นไม้แต่ละต้นที่อยู่ในป่า ดังนั้นจึงสามารถสรุปได้ว่า Random Forest เป็น Algorithm ประเภทหนึ่งของ Decision Tree ที่มีลักษณะแบบไม่ตัดแต่งกิ่ง (Unpruned) หรือต้นไม้ถดถอย (Regression Trees) ซึ่งถูกสร้างจากกการนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูล และคุณลักษณะข้อมูลแล้วนำมาสร้างเป็น Decision Tree ซึ่งมีตัวอย่างส่วนหนึ่งซึ่งไม่ถูกเลือกจะถูกนำมาใช้ในการทดสอบ Decision Tree เรียกข้อมูลส่วนนี้ว่า Out-of-bag (OOB) โดยวิธีการนี้เรียกว่า Bagging ผลลัพธ์ที่ได้อย่างอิสระจาก Decision Tree แต่ละต้นจะถูกนำมาคิดเป็นผลการโหวตที่มากที่สุด ทั้งนี้ Random Forest ไม่จำเป็นต้องมีข้อมูลทดสอบเพื่อประมาณความผิดพลาด เพราะข้อมูล OOB นั้นถูกนำมาใช้ทดสอบ Decision Tree อยู่แล้ว

2.3 การจัดเก็บและค้นคืนข้อมูลสารสนเทศ

เทคโนโลยีการค้นคืนข้อมูลสารสนเทศ คือ เทคโนโลยีที่ว่าด้วยการจัดเก็บ ประมวลผล ค้นคืน และนำเสนอข้อมูลและเอกสารที่มีความสัมพันธ์กับสิ่งที่ผู้ใช้ต้องการค้นคืน (Relevant Documents) ซึ่งเทคโนโลยีการค้นคืนข้อมูลสารสนเทศนี้มีต้นกำเนิดในช่วงทศวรรษที่ 1960 โดยในช่วงแรกการค้นคืนข้อมูลยังจำกัดอยู่เพียงการค้นคืนเอกสารจำนวนไม่มาก แต่การเกิดขึ้นของอินเทอร์เน็ตช่วงทศวรรษที่ 1990 ส่งผลให้ข้อมูลมีปริมาณเพิ่มขึ้นอย่างรวดเร็ว ความจำเป็นที่ตามมาคือความต้องการในการค้นหาเอกสาร ด้วยเหตุนี้เองจึงมีการศึกษาวิจัยเทคนิคและวิธีการที่ทำให้การค้นคืนข้อมูลสารสนเทศมีประสิทธิผลมากขึ้น

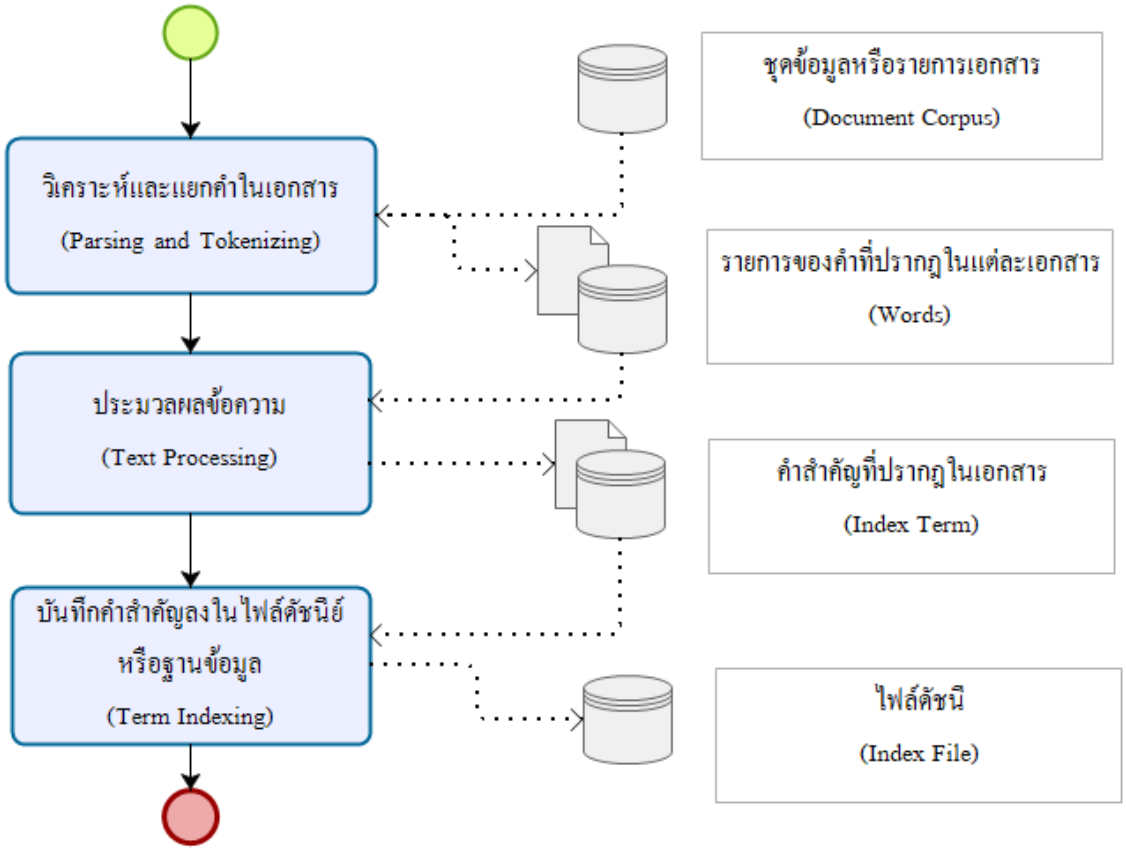
โดยทั่วไปเทคนิคของการค้นคืนสารสนเทศนั้นระบบจะทำการประมวลผลเอกสารที่อยู่ถูกเก็บอยู่ในคลังข้อมูล (Documents Corpus) เพื่อสกัดคำสำคัญที่ปรากฏในเอกสาร มาสร้างเป็นดัชนีคำสำคัญ (Indexes) เก็บไว้ในฐานข้อมูล จากนั้นเมื่อผู้ใช้ทำการค้นคืนข้อมูลด้วยคำค้นคืน (Query) ระบบจะทำการเปรียบเทียบคำค้นคืนกับรายการดัชนีที่บันทึกอยู่ในฐานข้อมูล จากนั้นจึงแสดงผลรายการเอกสารที่มีความสัมพันธ์กับคำค้นคืน (Relevant Documents) นำเสนอแก่ผู้ใช้ ดังที่แสดงให้เห็นในภาพด้านล่าง



ภาพที่ 2.6 การทำงานของระบบค้นคืนทั่วไป

กลุ่มนักวิจัยและนักพัฒนาระบบสารสนเทศเฉพาะทางในสาขาการสืบค้นข้อมูลให้การรับรองว่าทฤษฎี หลักการและวิธีปฏิบัติในศาสตร์การค้นคืนข้อมูลนั้น สามารถนำไปประยุกต์ใช้เพื่อการออกแบบระบบที่มีความสามารถในการระบุองค์ประกอบสำคัญ เนื้อหา ทั้งยังสามารถแสดงขั้นตอนการประมวลผลเอกสารได้ เป็นต้น ในขณะเดียวกัน ได้มีการศึกษาค้นคว้าเพื่อพัฒนาเทคนิคการค้นหาให้มีความสามารถในการเรียนรู้ให้มีประสิทธิภาพมากยิ่งขึ้นโดยนำเสนอเทคนิควิธีที่หลากหลายสำหรับการให้ผลลัพธ์ที่เท่าเทียมกันทั้งผู้ใช้งานทั่วไปและผู้ใช้งานผ่านระบบเครือข่าย

2.3.1 กระบวนการจัดเก็บข้อมูล กระบวนการนี้มีวัตถุประสงค์ในการสร้างดัชนีคำสำคัญที่ใช้บันทึกข้อมูลดัชนีคำสำคัญ (Index Terms) ซึ่งเป็นเสมือนตัวแทนที่ใช้ในการค้นคืนเอกสารในคลังเพื่อใช้ในการเปรียบเทียบกับคำค้นคืน โดยกระบวนการจะมีขั้นตอนเป็นไปตามที่แสดงในแผนภาพ



ภาพที่ 2.7 แผนภาพกิจกรรมแสดงขั้นตอนการจัดเก็บข้อมูลและสร้างดัชนีคำสำคัญ

วิเคราะห์และแยกคำในเอกสาร (Parsing and Tokenizing) เป็นขั้นตอนที่เริ่มจากการนำชุดข้อมูลหรือเอกสารที่ถูกจัดเก็บอยู่ในคลังมาทำการวิเคราะห์และแบ่งเป็นรายการของคำที่ปรากฏในแต่ละเอกสาร (Words)

ประมวลผลข้อความ (Text Processing) หลังจากที่ได้รายการคำที่ปรากฏในเอกสารมาแล้วนั้น เราจะนำคำที่ได้เหล่านั้นมาทำการประมวลผลเพื่อให้ได้ออกมาเป็นดัชนีคำสำคัญ โดยขั้นตอนในการประมวลผลข้อความหลัก ๆ จะเป็นการนำคำที่ได้มาทำการกรองคำที่ไม่มีความหมาย (Stop word Removal) โดยเทียบจากรายการคำศัพท์ที่มักปรากฏในเอกสาร (Stop list) เช่น คำว่า 'the', 'an', 'of' เป็นต้น เนื่องจากคำเหล่านี้มีคุณสมบัติในการระบุลักษณะเนื้อหาของเอกสารน้อย นอกจากนี้ยังมีการนำรายการคำที่ได้มาทำการแปลงให้อยู่ในรูปรากศัพท์ (Stemming and Plural Removal) เช่นคำว่า 'computing' หรือ 'computed' จะถูกแปลงให้อยู่ในรูป 'compute' เพื่อทำให้สามารถค้นคืนเจอได้ง่ายขึ้น

บันทึกคำสำคัญลงในไฟล์ดัชนีหรือฐานข้อมูล (Term Indexing)

รายการคำสำคัญที่ได้หลังจากการประมวลผลข้อความนั้นจะถูกนำมาสร้างเป็นไฟล์ดัชนีคำสำคัญเพื่อนำไปใช้ในการค้นคืน โดยในไฟล์ดัชนีนั้นจะมีการให้น้ำหนักของคำสำคัญแต่ละคำ (Term Weighting) ซึ่งวิธีในการให้น้ำหนักความสำคัญนั้นมีหลายวิธี สำหรับในงานวิจัยนี้เราได้ใช้วิธีคำนวณน้ำหนักโดยพิจารณาความถี่ของคำที่ปรากฏในเอกสาร (TF-IDF, Term Frequency - Inverse Document Frequency) โดยวิธีนี้เป็นวิธีทางสถิติที่ใช้ประเมินความสำคัญของคำที่มีต่อเอกสารแต่ละรายการ กล่าวคือความสำคัญของคำจะเป็นสัดส่วนโดยตรงกับจำนวนครั้งที่คำสำคัญปรากฏในเอกสารแต่จะถูกลดความสำคัญลงหากคำนั้นปรากฏในเอกสารอื่น ๆ ในคลังด้วย ค่าน้ำหนักของคำสำคัญแต่ละคำสามารถคำนวณได้จากสมการ

$$w(i,j) = tf(i,j) \cdot idf(i)$$

โดยที่ $w(i,j)$ คือค่าน้ำหนักของคำสำคัญ i ที่ปรากฏในเอกสาร j ส่วน $idf(i)$ คือค่า Inverse Document Frequency ที่ใช้วัดความทั่วไปของคำสำคัญ i ที่ปรากฏในเอกสารทั้งหมดในคลัง สามารถคำนวณได้จากสูตรสมการที่ xx ส่วน $tf(i,j)$ คือจำนวนครั้งที่คำสำคัญ i ปรากฏในเอกสาร j ที่ถูกนอร์มัลไลซ์เพื่อลดความคาดเคลื่อนแล้ว สามารถคำนวณได้จากสมการ

$$idf(i) = \log_2(n) - \log_2(docfreq(i)) + 1$$

โดยที่ n คือจำนวนเอกสารที่มีทั้งหมดในคลังและ $docfreq(i)$ คือจำนวนเอกสารที่มีคำสำคัญ i ปรากฏอยู่

$$tf(i,j) = \frac{freq(i,j)}{\max_l freq(l,j)}$$

โดยที่ $freq(i,j)$ คือจำนวนครั้งที่คำสำคัญ i ปรากฏในเอกสาร j และ $\max_l freq(l,j)$ คือจำนวนคำสำคัญทั้งหมดที่ปรากฏในเอกสาร j

2.3.2 การสกัดคำคุณลักษณะ

การสกัดคุณลักษณะของบทความหรือข่าวคือการสร้างตัวแทนคุณลักษณะของเอกสารซึ่งอาจจะใช้คำเดี่ยว วลี หรือประโยค คุณลักษณะที่สกัดได้จะถูกจัดรูปแบบให้อยู่ในลักษณะของเวกเตอร์และถูกแทนด้วยลักษณะของค่าความจริงหรือแทนด้วยค่าความถี่ของคำ [18] สำหรับงานวิจัยชิ้นนี้ใช้คำเดี่ยวที่ได้จากกระบวนการตัดคำเป็นคุณลักษณะและการหาค่าน้ำหนักของคำในการสกัดคุณลักษณะ จากนั้นจึงลดคุณลักษณะของเอกสารด้วยการละเว้นคำที่เป็นคำบุพบทและคำสันธาน [19]

2.3.3 การตัดคำ (Word Segmentation)

การตัดคำเป็นการแบ่งข้อความที่เรียงต่อเนื่องกันออกมาเป็นหน่วยคำเพื่อหาขอบเขตของแต่ละหน่วยคำ ซึ่งวิธีการตัดคำจะขึ้นอยู่กับลักษณะของภาษาที่นำมาวิเคราะห์ เช่นข้อความภาษาไทยจะมีลักษณะการเขียนที่ต่อเนื่องกัน การหาจุดจบของขอบเขตของคำทำได้ยาก ต่างจากภาษาอังกฤษที่มีการเว้นวรรคคำอย่างชัดเจนและสามารถหาจุดจบของข้อความได้

2.3.4 การสร้างตัวแทนเอกสาร

เพื่อให้คอมพิวเตอร์เข้าใจภาษามนุษย์ การเรียนรู้ของเครื่องจึงนิยมใช้ลักษณะตัวแทนของคำหรือข้อความมากกว่าความหมายของคำ ซึ่งตัวแทนของคำหรือข้อความมักอยู่ในรูปของเวกเตอร์ของน้ำหนักคำ โดยมากใช้ค่าน้ำหนักของคำเป็นไบนารีหรือไม่เป็นไบนารีก็ได้ขึ้นอยู่กับวิธีการคำนวณ สำหรับงานวิจัยชิ้นนี้จะใช้วิธีการคำนวณค่าน้ำหนักด้วยวิธีไบนารี

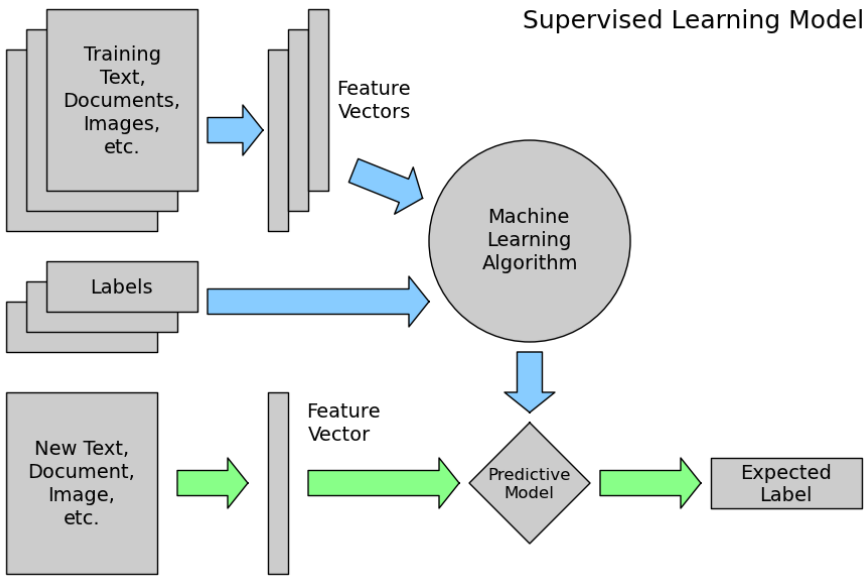
2.4 การจำแนกประเภท

การจำแนกประเภท (Classification) เป็นการวิเคราะห์คุณลักษณะของข้อมูลตัวอย่างเพื่อใช้ในการจัดกลุ่ม ซึ่งในที่นี่หมายถึงตัวแปรที่ใช้ในการทำนาย เพื่อที่จะสามารถจำแนกกลุ่มได้จะต้องทราบล่วงหน้าว่ามีกลุ่มอะไรบ้าง (Predefined Categories) และมีข้อมูลของกลุ่มตัวอย่างในแต่ละกลุ่มก่อน การจำแนกประเภทเป็นการทำเหมืองข้อมูลแบบที่นิยมใช้แพร่หลายซึ่งครอบคลุมการใช้งานที่หลากหลาย (Tan, Steinbach & Kumar, 2006 หน้า 145) เช่น การคัดแยกอีเมล์ขยะจากอีเมล์ปกติโดยพิจารณาหัวเรื่องหรือเนื้อหาของอีเมล์ การจำแนกเซลที่เป็นหรือไม่เป็นอันตรายจากผลการตรวจด้วย MRI การจำแนกกาแลกซี่โดยอาศัยลักษณะรูปร่างของมัน หรือการทำนายการเป็นมะเร็งโดยอาศัยข้อมูลระดับการแสดงออกของยีน เป็นต้น

เทคนิคการจำแนกประเภทเป็นวิธีการสร้างตัวแบบการจำแรกประเภท (Classification Model) อย่างมีระบบ จากข้อมูลนำเข้า (Input Data) ตัวอย่างเทคนิคการจำแรกประเภทได้แก่ ตัวจำแนกต้นไม้การตัดสินใจ (Decision Tree Classifier) ตัวจำแนกด้วยกฎ (Rule-based Classifier) ข่ายงานประสาท (Neural Network) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine หรือ SVM) ตัวจำแนกนาอีฟเบย์ (Naïve Bayes Classifier) เป็นต้น โดยที่แต่ละเทคนิคใช้อัลกอริทึมการเรียนรู้เพื่อสร้างตัวแบบที่สอดคล้องกับข้อมูลที่สุด โดยแสดงความสัมพันธ์ระหว่างตัวแปรทำนาย และกลุ่มของข้อมูลนำเข้า ตัวแบบที่สร้างขึ้นนี้นอกจากจะ

มีความสอดคล้องกับข้อมูลนำเข้าแล้ว ยังควรมีความสามารถในการจำแนกประเภทข้อมูลใหม่ที่ไม่ได้ใช้ในการสร้างตัวแบบได้อย่างถูกต้องอีกด้วย

แนวทางการแก้ปัญหาการจำแนกประเภทเริ่มจากการแบ่งข้อมูลออกเป็นสองส่วน ได้แก่ ข้อมูลฝึกฝน (Training Data) และข้อมูลทดสอบ (Testing Data) แต่ละชุดข้อมูลมีข้อมูลของตัวแปรทำนาย และกลุ่มที่แต่ละหน่วยตัวอย่างเป็นสมาชิก ข้อมูลฝึกฝนจะถูกนนำไปใช้เพื่อจำแนกประเภทข้อมูลของหน่วยตัวอย่างในข้อมูลทดสอบ ตัวแปรที่ได้รับการทดสอบว่ามีความถูกต้องสูงก็จะถูกนำไปใช้ในการทำนายข้อมูลใหม่ แนวทางการสร้างตัวแปรแบบจำแนกประเภท แสดงดังภาพที่ 2.8



ภาพที่ 2.8 แนวทางการสร้างตัวแบบการจำแนกประเภท

2.5 การประเมินประสิทธิภาพของตัวแบบการจำแนกประเภท

การประเมินประสิทธิภาพของตัวแบบการจำแนกประเภท อาศัลผลการทำนายการเป็นสมาชิกกลุ่มข้อมูล เมื่อใช้ตัวแบบกับข้อมูลทดสอบโดยจับจำนวนตัวอย่างที่ได้รับการทำนายที่ถูกต้อง และที่ไม่ถู กต้อง ค่าจำนวนตัวอย่างเหล่านี้ถูกนำมาสร้างเป็นตารางแสดงผลลัพธ์การจำแรกประเภทที่เรียกว่า Confusion Matrix ซึ่งใช้สำหรับกรณีปัญหาการจำแนกประเภทที่แบ่งกลุ่มข้อมูลเป็นสองกลุ่ม ตัวเลขที่แสดงในราคางเป็นจำนวนตัวอย่างในแต่ละสถานการณ์ ดังนี้

a_1 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มที่ 1 และได้รับการทำนายว่าอยู่ในกลุ่มที่ 1

a_2 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มที่ 1 และได้รับการทำนายว่าอยู่ในกลุ่มที่ 2

a_3 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มที่ 2 และได้รับการทำนายว่าอยู่ในกลุ่มที่ 1

a_4 หมายถึง จำนวนหน่วยตัวอย่างที่อยู่ในกลุ่มที่ 2 และได้รับการทำนายว่าอยู่ในกลุ่มที่ 2

จำนวนหน่วยตัวอย่างที่ได้รับการทำนายได้อย่างถูกต้องมีจำนวน $a_1 + a_4$ ตัวอย่างและจำนวนหน่วยตัวอย่างที่ได้รับการทำนายไม่ถูกต้อง มีจำนวน $a_2 + a_3$ ตัวอย่าง ดังแสดงในตารางที่ 2.2

ตารางที่ 2.2 ผลการจำแนกประเภทของหน่วยตัวอย่าง

ประเภทของหน่วยตัวอย่างที่เป็นจริง	**ประเภทของหน่วยตัวอย่างที่ถูกทำนาย**	
	กลุ่มที่ 1	กลุ่มที่ 2
กลุ่มที่ 1	a_1	a_2
กลุ่มที่ 2	a_3	a_4

เนื่องจากจำนวนข้อมูลข่าวปลอมและข่าวจริงในชุดข้อมูล Fake Corpus ไม่สมดุลกันทำให้คุณสมบัติของข้อมูลส่วนใหญ่บดบังคุณสมบัติของข้อมูลส่วนน้อย และทำให้ประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยนั้นไม่ดีเท่าที่ควร ดังนั้นเพื่อแก้ปัญหาดังกล่าว ผู้วิจัยจึงได้นำ Confusion Matrix มาใช้ในการประเมินผลลัพธ์ของการจัดกลุ่ม โดยที่หากข่าวปลอมใดที่ระบุไว้ว่าเป็นข่าวปลอมและสามารถจำแนกได้อย่างถูกต้องว่าเป็นข่าวปลอม ถือเป็น True Positive (TP) หากข่าวปลอมใดที่ระบุไว้ว่าเป็นข่าวจริงแต่ไม่สามารถจำแนกได้ว่าเป็นข่าวปลอม ถือเป็น False Negative (FN) หากข่าวจริงใดที่ระบุไว้ว่าเป็นข่าวจริงและสามารถจำแนกได้ว่าเป็นข่าวจริง ถือเป็น True Negative (TN) และสุดท้าย หากข่าวจริงใดถูกระบุไว้ว่าเป็นข่าวปลอมแต่ไม่สามารถจำแนกได้ว่าข่าวจริง ถือเป็น False Positive (FP)

ในการสร้างแบบจำลองให้มีประสิทธิภาพ ผู้วิจัยได้พยายามลดจำนวนการพยากรณ์ที่ผิดพลาด ทั้ง False Negative และ False Positive โดยการนำวิธีการวัดประสิทธิภาพของแบบจำลอง F-measure (F-1) มาใช้ เพื่อให้เกิดความสมดุลระหว่างความแม่นยำ (Precision) กับการเรียกคืน (Recall)

ค่าความแม่นยำ (Precision) เป็นการวัดความแม่นยำของโมเดล โดยพิจารณาแยกทีละคลาสแล้วนำมาหาค่าเฉลี่ย ดังสมการ

$$Precision_i = \frac{TP_i}{TP_i + FP_i}$$

ค่าเรียกคืน (Recall) เป็นการวัดการคืนค้นของโมเดลโดยพิจารณาแยกทีละคลาสแล้วนำมาหาค่าเฉลี่ย ดังสมการ

$$Recall_i = \frac{TP_i}{TP_i + FN_i}$$

ค่าความถ่วงดุล (F-Measure) หรือค่าประสิทธิภาพโดยรวม เป็นการวัดประสิทธิภาพโดยรวมของทั้งสองค่า ระหว่างค่าของความแม่นยำ และค่าเรียกคืน ดังสมการ

$$F - Measure_i = \frac{2 \text{ x } (Recall_i \text{ x } Precision_i)}{Recall_i \text{ x } Precision_i}$$

นอกจากนี้ เพื่อให้เห็นภาพผลการทดลอง ผู้วิจัยได้นำ AUC (Area Under the ROC Curve) มาใช้เพื่อเป็นตัวบอกประสิทธิภาพในการทดสอบ (test performance) ว่าแบบจำลองสามารถระบุข่าวปลอมและข่าวจริงได้ดีมากน้อยเพียงใด และระบุจุดตัด (cut-off) ของการทดสอบที่มีความแม่นยำและน่าเชื่อถือมากที่สุดเพื่อนำค่าดังกล่าวมาใช้ในการจำแนกข่าวจริงและข่าวปลอมให้ถูกต้องมากที่สุดและผิดพลาดน้อยที่สุด

2.6 งานวิจัยที่เกี่ยวข้อง

การตรวจจับข่าวปลอมเป็นประเด็นที่น่าสนใจและมีผู้วิจัยเพื่อหาเทคนิควิธี การตรวจจับหลายวิธีด้วยกัน จากการศึกษางานวิจัยในด้านการตรวจจับข่าวปลอม นักวิจัยหลายท่านได้นำเสนอวิธีการแก้ปัญหาและการตรวจจับข่าวปลอมไว้สองวิธีใหญ่ ๆ ด้วยกัน คือ: 1) วิธีการศึกษาทางภาษาศาสตร์ 2) วิธีการศึกษาบนเครือข่าย

2.6.1 วิธีการศึกษาทางภาษาศาสตร์ (Linguistic Approaches)

Mihalcea และ Strapparvva 2009 นำเสนอเทคนิคการประมวลผลภาษาธรรมชาติ (Neural Language Processing) เพื่อแก้ปัญหานี้ Bing Liu และคณะ [21] ได้ศึกษาวิเคราะห์คำวิจารณ์ปลอม (Fake Reviews) ในเว็บไซต์ Amazon โดยจากการวิเคราะห์ความเชื่อมั่น คำศัพท์ ความคล้ายคลึงกันของเนื้อหาและความไม่สอดคล้องกันทางความหมาย เพื่อระบุทวิจารณ์ปลอม Hai และคณะ [22] เสนอวิธีการเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised) เพื่อตรวจจับข้อความขยะ (Spam Detection) ซึ่งจากการวิจัยทางด้านภาษาศาสตร์ แม้จะได้ผลเป็นที่น่าพอใจ แต่วิธีการวิเคราะห์เพียงเนื้อหา หรือคำ ไม่เพียงพอต่อการตรวจจับข่าวปลอม เพราะข่าวปลอมมักมีรูปแบบการเขียนที่เปลี่ยนแปลงไปและมีการพัฒนาให้มีความคล้ายคลึงกับข่าวจริงมากยิ่งขึ้น ดังนั้น นักวิจัยหลายท่านจึงมุ่งความสนใจไปที่โครงสร้างภาษาที่ลึกกว่าเดิม เช่นแผนผังไวยากรณ์ ซึ่งในกรณีนี้ประโยคจะถูกแสดงเป็นลักษณะโครงสร้างต้นไม้ เพื่อธิบายโครงสร้างไวยากรณ์ เช่นคำนาม คำกริยา วลี

2.6.2 วิธีการศึกษาบนเครือข่าย (Network-based Approaches)

อีกวิธีหนึ่งในการตรวจจับข่าวปลอมคือการวิเคราะห์โครงสร้างเครือข่าย ความเชื่อมโยงข้อมูลและพฤติกรรมการเผยแพร่ข้อมูล ซึ่งเป็นคุณลักษณะที่สำคัญในขณะที่การพัฒนากราฟความรู้ (Knowledge Graph) ซึ่งเป็นประโยชน์มากในการตรวจสอบข้อเท็จจริงโดยพิจารณาตามความสัมพันธ์ระหว่างผู้เผยแพร่และผู้รับข่าวสาร Ciampaglia และคณะ [23] ได้นำเสนอแนวคิดใหม่ โดยใช้ตัวแปรในเครือข่ายที่มีผลกระทบต่อการแพร่กระจายของข่าว และเพื่อวิเคราะห์ความน่าจะเป็นของประเภทข่าว วิธีการตามการวิเคราะห์กราฟความรู้สามารถให้ผลความแม่นยำ (Precision) ร้อยละ 61 ถึงร้อยละ 95 โดยทิศทางการวิจัยที่มีแนวโน้มอีกประการหนึ่งคือการใช้ประโยชน์จากพฤติกรรมของเครือข่ายสังคมออนไลน์เพื่อระบุข่าวปลอม รวมถึงการหลอกลวงประเภทอื่น

สุปัญญา อภิวงศ์โสภณ [13] ได้ทำการศึกษาและตรวจจับข่าวปลอมโดยใช้เทคนิคการเรียนรู้ของเครื่องยอดนิยมสามประเภท คือนาอีฟเบส์ (Naive Bayes: NB) โครงข่ายประสาทเทียม (Neural Network : NN) และซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) ผลที่ได้จากการศึกษาแสดงให้เห็นว่าเทคนิคนาอีฟเบส์สามารถตรวจสอบข่าวปลอมมีความถูกต้อง (Accuracy) ร้อยละ 96.08 ส่วนโครงข่ายประสาทเทียมและซัพพอร์ทเวกเตอร์แมชชีน มีความถูกต้องที่ร้อยละ 99.90

วิริยาภรณ์ ทองสุข [9] ศึกษาเรื่อง “การวิเคราะห์การนำเสนอข่าวเชิงคลิกเบทของเว็บไซต์”เป็นการวิจัยแบบผสมผสาน โดยใช้ระเบียบวิธีวิจัยแบ่งออกเป็น 3 สวน คือ สวนที่ 1 การ วิเคราะห์คุณลักษณะการพาดหัวข่าวและความสอดคล้องการพาดและเนื้อหาข่าวเชิงคลิกเบทของ เว็บไซต์ สวนที่ 2 การสัมภาษณ์เชิงลึก (in-depth Interview) ตัวแทนนักวิชาการและนักวารสาร ศาสตร์ด้านสื่อสารมวลชน และสวนที่ 3 การสนทนากลุ่ม (Focus Group) จำนวน 2 กลุ่ม คือตัวแทน กลุ่มวัยทำงานและตัวแทนกลุ่มนักศึกษา ผลการศึกษาพบว่ามีคุณลักษณะการพาดหัวข่าวที่ใช้รูปแบบ ของวลีที่นำมาเรียง ๆ กัน เพื่อใช้ในการพาดหัวข่าวมากที่สุด ในสวนของภาษาที่ใช้การพาดหัวข่าวใช้ คำเรียกชื่อจริง/ชื่อเล่น ประกอบกับการใช้เครื่องหมายอัศเจรีย์ (!) คำสแลง และภาษาต่างประเทศ เพื่อเป็นการกระตุ้นความอยากรู้ให้กับผู้อ่านคลิกเข้าไปอ่านข่าว ส่วนเนื้อหาข่าวของเว็บไซต์ที่นำเสนอ เป็นข้อความสั้น ๆ ประกอบกับภาพนิ่ง และการวิเคราะห์ความสอดคล้องเนื้อหาข่าวและพาดหัวข่าว ของเว็บไซต์ พบว่าสวนใหญ่ยังมีความสอดคล้องต้องกันอยู่ การศึกษาแนวโน้มของการนำเสนอข่าวเชิง คลิกเบทของเว็บไซต์ จากตัวแทนนักวิชาการและนักวารสารศาสตร์ ด้านสื่อสารมวลชน ระบุว่า การ นำเสนอข่าวเชิงคลิกเบทมีการปรับตัวอยู่เสมอมักปรับเปลี่ยนรูปแบบการนำเสนอข่าว เพื่อเรียกร้อง ความสนใจของผู้อ่านเป็นหลัก โดยการนำเสนอในรูปแบบของกลเม็ดชีวิต รูปแบบการดำเนินชีวิต เพื่อให้ผู้อ่านแชร์ต่อในโซเชียลมีเดีย ในด้านจริยธรรมสื่อมวลชนถือว่าการนำเสนอข่าวเชิงคลิกเบทเป็นข่าวที่ไม่ค่อยมีความนาเชื่อถือ เพราะได้นำเสนอข้อเท็จจริงให้กับผู้อ่านเพียงบางสวนหรืออาจมีการบิดเบือนข่าวให้กับผู้อ่านข่าว ซึ่งผลให้ผู้อ่านจะได้รับรู้ข่าวที่ไม่ครบถ้วนรอบด้าน ด้านการศึกษาความคิดเห็นผู้อ่านข่าวเชิงคลิกเบทของเว็บไซต์ ตัวแทนกลุ่มวัยทำงานและตัวแทนกลุ่มนักศึกษา ให้ ความคิดที่สอดคล้องกันเห็นว่า การนำเสนอข่าวเชิงคลิกเบทของเว็บไซต์มีลักษณะเป็นด้านลบ เพราะ นำเสนอข่าวที่เกินความจริง

และขาดความน่าเชื่อถือหลายสวน ถึงแม้ว่าผู้อ่านจะรู้เท่าทันสื่อแต่แนวโน้มของการนำเสนอข่าวเชิงคลิกเบทจะยังอยู่ในโลกออนไลน์เพราะผู้อ่านยังคงให้ความสนใจกับ ข่าวสังคม ข่าวบันเทิง ผู้นำเสนอข่าวเชิงคลิกเบทจึงมีการปรับเปลี่ยนรูปแบบการนำเสนอเพื่อให้ผู้อ่าน เกิดความรู้สึกแปลกใหม่ และไม่ให้รู้เท่าทันอยู่เสมอ

Victoria L. Rubin และคณะ [13] ได้ทำการวิจัยโดยใช้ซัพพอร์ทเวกเตอร์แมชชีน เพื่อทำนายลักษณะและจัดกลุ่มข่าวออกเป็น 5 ประเภท จากข่าวทั้งหมด 360 ข่าว ผลการวิจัยสามารถจัดกลุ่มได้อย่างแม่นยำถึงร้อยละ 90 ซึ่งผลการวิจัยชิ้นนี้สามารถช่วยลดผลกระทบที่อาจเกิดขึ้นจากข่าวปลอม

Alcott & Gentzkow [24] ศึกษาเรื่อง “Social Media and Fake News in the 2016 Election” โดยมีวัตถุประสงค์ในการศึกษาว่าข่าวปลอมมีอิทธิพลหลักต่อผลการเลือกตั้งสหรัฐอเมริกา ในปี 2016 หรือไม่ โดยเตรียมข่าวสามประเภทไปถามความเห็นจากผู้ลงคะแนนเสียงเลือกตั้ง ข่าวสาม ประเภทนั้นได้แก่ ข่าวจริง ข่าวปลอมจริง และข่าวปลอมปลอมที่ทีมวิจัยเขียนขึ้นมาเอง ผ่านการเก็บ ข้อมูลโดยใช้แบบสอบถามออนไลน์สำรวจความคิดเห็นกลุ่มตัวอย่างชาวอเมริกันจำนวน 1,208 คน ที่มีอายุ 18 ปีขึ้นไป ผลการศึกษาพบว่า กลุ่มตัวอย่างจำนวนร้อยละ 14 ระบุว่าสื่อสังคมออนไลน์ว่าเป็น แหล่งข่าวที่สำคัญที่สุดของข่าวการเลือกตั้ง กลุ่มตัวอย่างสามารถจดจำ แยกแยะข่าวปลอมในช่วงการ เลือกตั้งได้และเชื่อในข่าวจริงมากกว่าข่าวปลอม แต่สำหรับข่าวปลอมที่มีการเปลี่ยนแปลงผลการ เลือกตั้งบทความเรื่องหนึ่งที่ปลอมแปลงจะต้องมีผลเหมือนกันเช่นเดียวกับการโฆษณาทางโทรทัศน์ 36 รายการ ข้อสรุปที่น่าสนใจคือข่าวปลอมมีอิทธิพลเพียงเล็กน้อยต่อทัศนคติของคน แต่สิ่งที่น่ากังวล ในระยะยาวคือ คนมีแนวโน้มจะโหยหาข้อมูลข่าวสารที่สอดคล้องกับอคติของตัวเอง ซึ่งโลกดิจิทัลมี ข้อมูลมากมายให้ใช้เติมเต็มความเชื่อเดิม ๆ

บทที่ 3
ระเบียบวิธีวิจัย

ขั้นตอนการจัดกลุ่มข้อมูลข่าวปลอมแบ่งออกเป็น 4 ขั้นตอน คือ 1) การเตรียมข้อมูลและนำมาจัดกลุ่ม 2) ขั้นตอนก่อนกระบวนการประมวลผล (Pre-processing) การนำข้อมูลที่ถูกจัดกลุ่มมาตัดคำและกำกับหน้าที่ของคำ จากนั้นทำกระบวนการสกัดคำคุณลักษณะเพื่อสร้างเวกเตอร์ข้อมูล พร้อมใส่น้ำหนักของคุณลักษณะ 3) การสร้างแบบจำลองโดยนำข้อมูลประมวลผลด้วยเทคนิคการจัดกลุ่ม 4 เทคนิค ได้แก่ การถดถอยโลจิสติก นาอีฟเบส์ ตัวจำแนกป่าแบบสุ่ม และซัพพอร์ทเวกเตอร์แมชชีน 4) การประเมินผลแบบจำลองโดยวัดประสิทธิภาพการจำแนกตามแนวคิดการค้นคืนสารสนเทศ โดยขั้นตอนทั้งหมดสามารถอธิบายได้ดังภาพ

ภาพที่ 3.1 แสดงขั้นตอนการจัดกลุ่มข่าวปลอม

3.1 การเตรียมข้อมูล

3.1.1 การวิเคราะห์ข้อมูล

เพื่อตรวจสอบการค้นพบจากข้อมูลดิบได้ทำการตรวจสอบอย่างละเอียดเพื่อศึกษาข้อมูลข้อความและรูปภาพในบทความข่าว มีความแตกต่างบางประการระหว่างข่าวจริงและของปลอม

3.1.2 ชุดข้อมูล

ในงานวิจัยชิ้นนี้ การวิเคราะห์ข้อมูลทางภาษาศาสตร์ ผู้วิจัยได้ใช้ชุดข้อมูลสาธารณะ Fake Corpus [16] ซึ่งมีผู้รวบรวมไว้ จำนวน 9,408,908 ข่าว ประกอบด้วยบทความหรือข่าวที่คัดลอกมาจากรายชื่อเว็บไซต์ที่เป็นแหล่งข่าวปลอม กับรายชื่อเว็บไซต์ที่มีเนื้อหาน่าเชื่อถือ จำนวน 1,001 เว็บไซต์โดยใช้ชุดไลบรารี่ของ python ที่ชื่อ Scrapy แล้วจึงนำมาบันทึกข้อมูลตามรูปแบบที่ต้องการด้วยชุดไลบรารี่ของ python ที่ชื่อ newspaper โดยแต่ละบทความจะถูกติดป้ายกำกับไว้ตามความน่าเชื่อถือหรือประเภทของเว็บไซต์ มีจำนวนดังแสดงดังตารางที่ 3.1

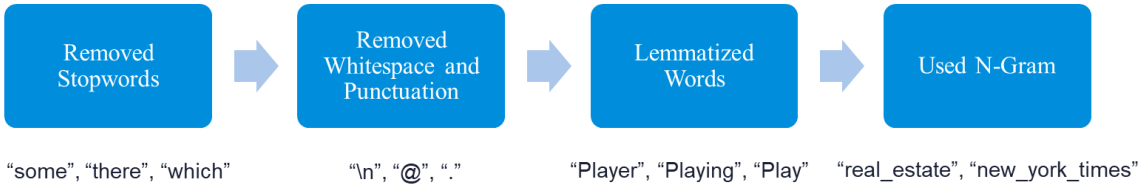
ตารางที่ 3.1 จำนวนบทความหรือข่าว แยกตามประเภท [5]

ประเภท	ป้ายกำกับ	จำนวน
Fake News	fake	928,083
Satire	satire	146,080
Extreme Bias	bias	1,300,444
Conspiracy Theory	conspiracy	905,981
Junk Science	junksci	144,939
Hate News	hate	117,374
Clickbait	clickbait	292,201
Proceed With Caution	unreliable	319,830
Political	political	2,435,471
Credible	reliable	1,920,139
State News	state	0

3.2 การแปลงข้อความเป็นคุณลักษณะ (Convert Text to Features)

เนื่องจากจำนวนข้อมูลข่าวปลอมและข่าวจริงในชุดข้อมูล Fake Corpus ไม่สมดุลกัน (Imbalance Data) ทำให้คุณสมบัติของข้อมูลส่วนใหญ่บดบังคุณสมบัติของข้อมูลส่วนน้อย และทำให้ประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยนั้นไม่ดีเท่าที่ควร ดังนั้นเพื่อแก้ปัญหาดังกล่าว ผู้วิจัยจึงสุ่มเลือกใช้ข้อมูลประเภท Fake News ที่มีป้ายกำกับว่า fake และ Credible ที่มีป้ายกำกับว่า reliable จำนวนประเภทละ 34,000 รายการ จากนั้น เพื่อนำข้อมูลไปสร้างแบบจำลองผู้วิจัยจึงติดป้ายกำกับไว้ สำหรับชุดข้อมูลที่เป็น fake จะติดป้ายกำกับไว้ว่า 1 และชุดข้อมูลที่เป็น reliable จะติดป้ายกำกับไว้ว่า 0

คุณภาพของการจัดรูปแบบการจำแนกข้อมูลขึ้นอยู่กับคำในคลังข้อมูล (Corpus) และคุณลักษณะของคำเหล่านั้น ในงานวิจัยชิ้นนี้จะใช้ชุดไลบรารี่ของ python ที่ชื่อ Spacy และ Gensim เพื่อจัดเตรียมข้อมูลมาใช้ในการวิเคราะห์ โดยขั้นตอนการทำงานเป็นดังภาพที่ 3.2



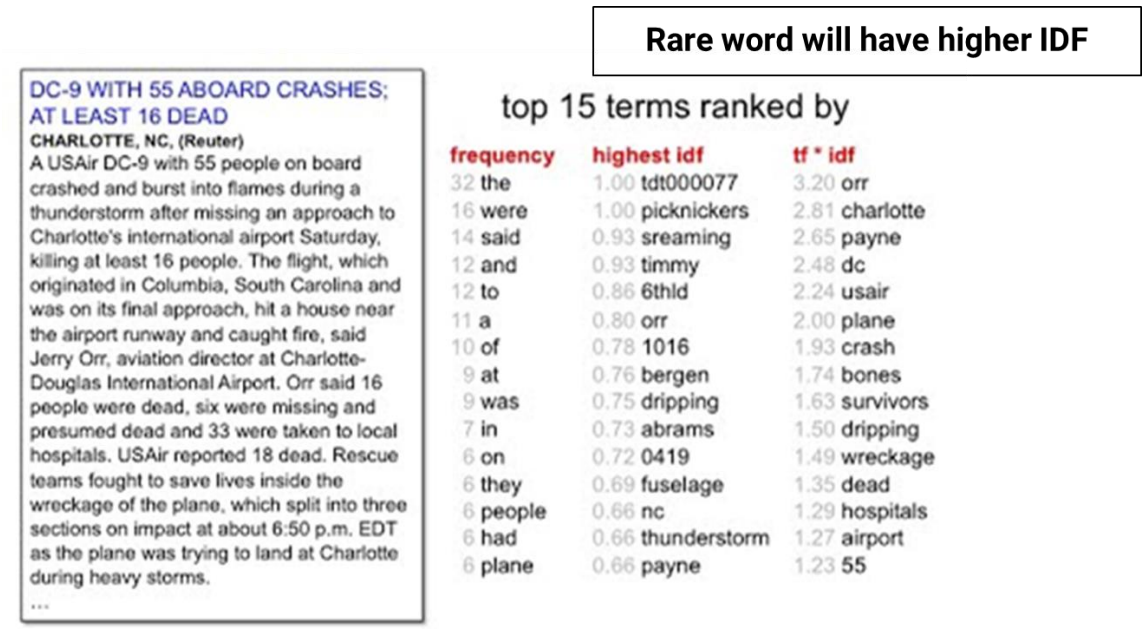
ภาพที่ 3.2 ขั้นตอนการแปลงข้อความเป็นคุณลักษณะ

ขั้นตอนข้างต้นจะช่วยลดขนาดและเพิ่มบริบทของข้อความก่อนที่จะแปลงให้อยู่ในรูปของคุณลักษณะโดยเฉพาะอย่างยิ่งการจัดคำให้อยู่ในรูปแบบปกติ (Lemmatization) ที่จะช่วยแปลงแต่ละคำที่อยู่ในรูปแบบต่าง ๆ ให้อยู่ในรูปแบบเดียว ส่วน n-Gram จะช่วยให้คำใกล้เคียงหรือมีบริบทเดียวกัน ให้เป็นคำเดียวกัน ซึ่งในการวิจัยครั้งนี้จะใช้ทั้งแบบ bi-Grams และ tri-Grams

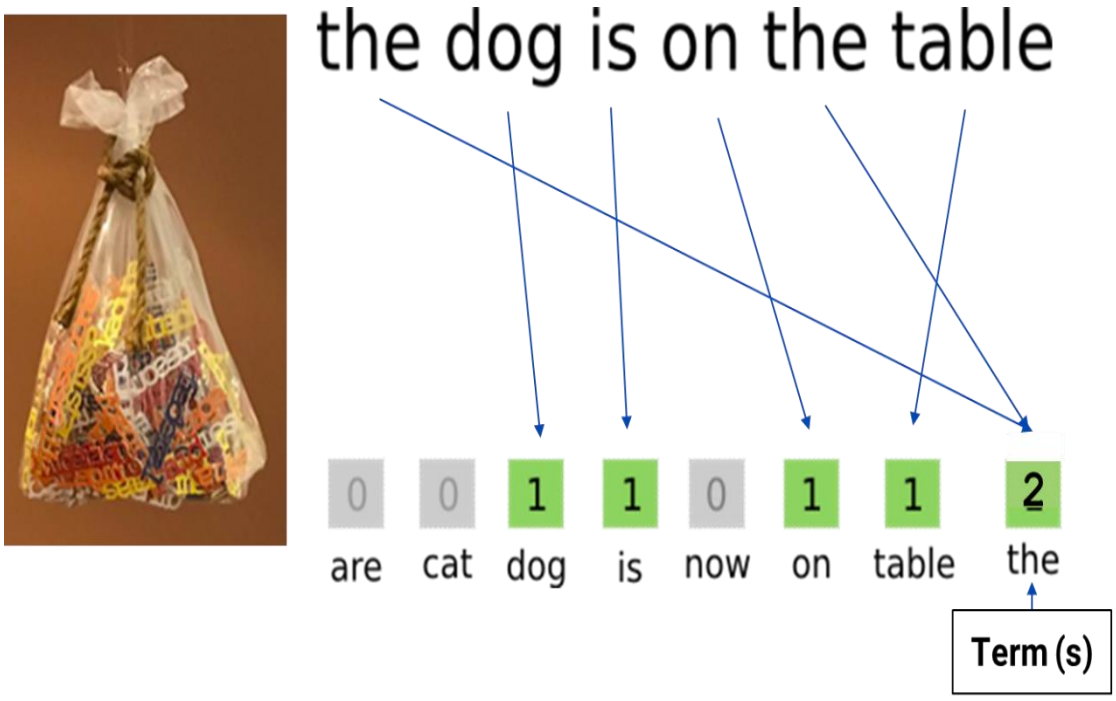
ในการวิเคราะห์และสร้างโมเดล หลังจากที่ได้มีการจัดเตรียมข้อมูลไว้แล้ว จะต้องมีการแปลงข้อมูลให้มีลักษณะเป็นคุณลักษณะ โดยเทคนิคที่จะนำมาใช้คือ TF-IDF

TF-IDF (Term Frequency Inverse Document Frequency) เป็นสถิติที่มีวัตถุประสงค์เพื่อคำนวณหาน้ำหนักจากความถี่ของการปรากฏคำในเอกสาร และพิจารณาความถี่ของคำนั้น ๆ ที่ปรากฏในเอกสารอื่นร่วมด้วย โดยมีแนวคิดที่ว่า คำที่ปรากฏในเอกสารน้อยฉบับจะมีค่าน้ำหนักสูง ส่วนคำที่ปรากฏในเอกสารหลายฉบับจะมีค่าน้ำหนักต่ำ เนื่องจากเป็นคำที่ไม่แสดงถึงคุณลักษณะเฉพาะของเอกสาร และแสดงว่าคำดังกล่าวไม่สามารถเป็นตัวแทนของเอกสารใด ๆ ได้ ซึ่งคำเหล่านั้นเรียกว่าคำหยุด (Stop Word) เช่น a and the เป็นต้น สามารถดูได้จากสมการ

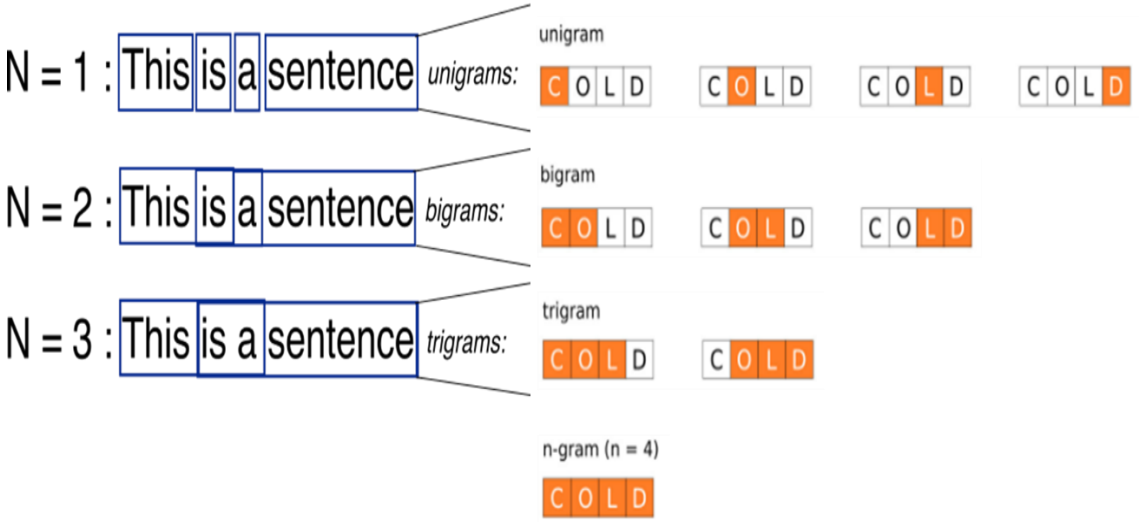
$$W_{(i,j)} = tf_{(i,j)} \times \log\left(\frac{N}{df_{(i)}}\right)$$



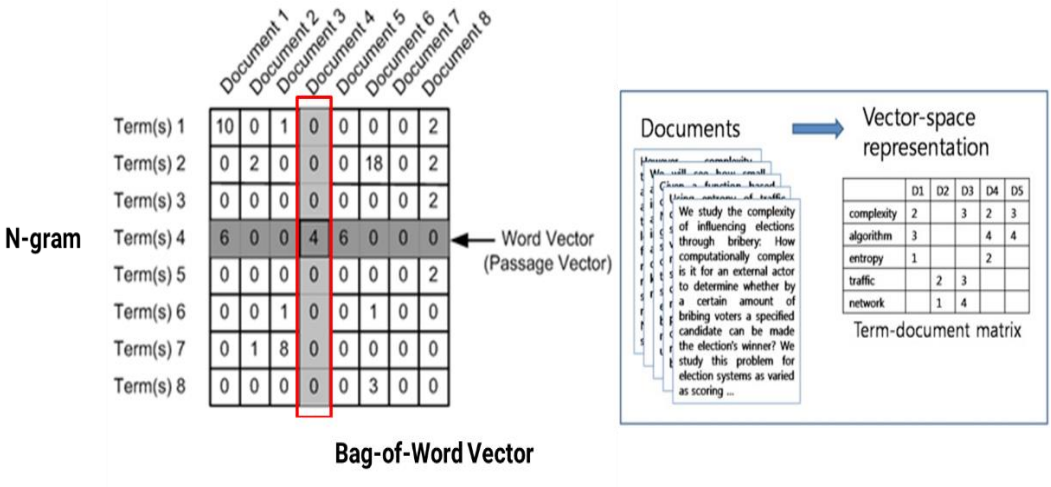
ภาพที่ 3.3 แสดงสถิติเพื่อคำนวณหาน้ำหนักจากความถี่ของการปรากฏคำในเอกสาร



ภาพที่ 3.4 แสดงลักษณะการทำงานของ Bag of Word



ภาพที่ 3.5 ขั้นตอนการแปลงข้อความเป็นคุณลักษณะ n-gram



ภาพที่ 3.6 แสดงการทำงานร่วมกันของ n-gram และ bag of word

บทที่ 4
ผลการวิจัย

รายละเอียดของผลลัพธ์ที่ได้จากการทดลองการจัดกลุ่มข่าวปลอมด้วยการเรียนรู้เครื่อง จำนวน 4 วิธี คือ การถดถอยโลจิสติก (Logistic Regression) นาอีฟเบส์ (Naïve Bayes) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) และ ตัวจำแนกป่าแบบสุ่ม (Random Forest) มีดังนี้

4.1 การประเมินผลแบบทดลอง

วิจัยนำ Confusion Matrix มาใช้ในการประเมินผลลัพธ์ของการจัดกลุ่ม โดยที่หากข่าวปลอมใดที่ระบุไว้ว่าเป็นข่าวปลอมและสามารถจำแนกได้อย่างถูกต้องว่าเป็นข่าวปลอม ถือเป็น True Positive (TP) หากข่าวปลอมใดที่ระบุไว้ว่าเป็นข่าวจริงแต่ไม่สามารถจำแนกได้ว่าเป็นข่าวปลอม ถือเป็น False Negative (FN) หากข่าวจริงใดที่ระบุไว้ว่าเป็นข่าวจริงและสามารถจำแนกได้ว่าเป็นข่าวจริง ถือเป็น True Negative (TN) และสุดท้าย หากข่าวจริงใดถูกระบุไว้ว่าเป็นข่าวปลอมแต่ไม่สามารถจำแนกได้ว่าข่าวจริง ถือเป็น False Positive (FP) ดังตารางที่ 4.1

ตารางที่ 4.1 ตัวอย่างนิยามของตารางผลจำนวนข่าวจากการทำนายได้จากการเรียนรู้เครื่อง

ทำนาย	**คำตอบ**	
	ข่าวจริง	ข่าวปลอม
ข่าวจริง	True Positive	False Positive
ข่าวปลอม	False Negative	True Negative

ในการสร้างแบบจำลองให้มีประสิทธิภาพ ผู้วิจัยได้พยายามลดจำนวนการพยากรณ์ที่ผิดพลาด ทั้ง False Negative และ False Positive โดยการนำวิธีการวัดประสิทธิภาพของแบบจำลอง F-measure (F-1) มาใช้ เพื่อให้เกิดความสมดุลระหว่างความแม่นยำ (Precision) กับการเรียกคืน (Recall)

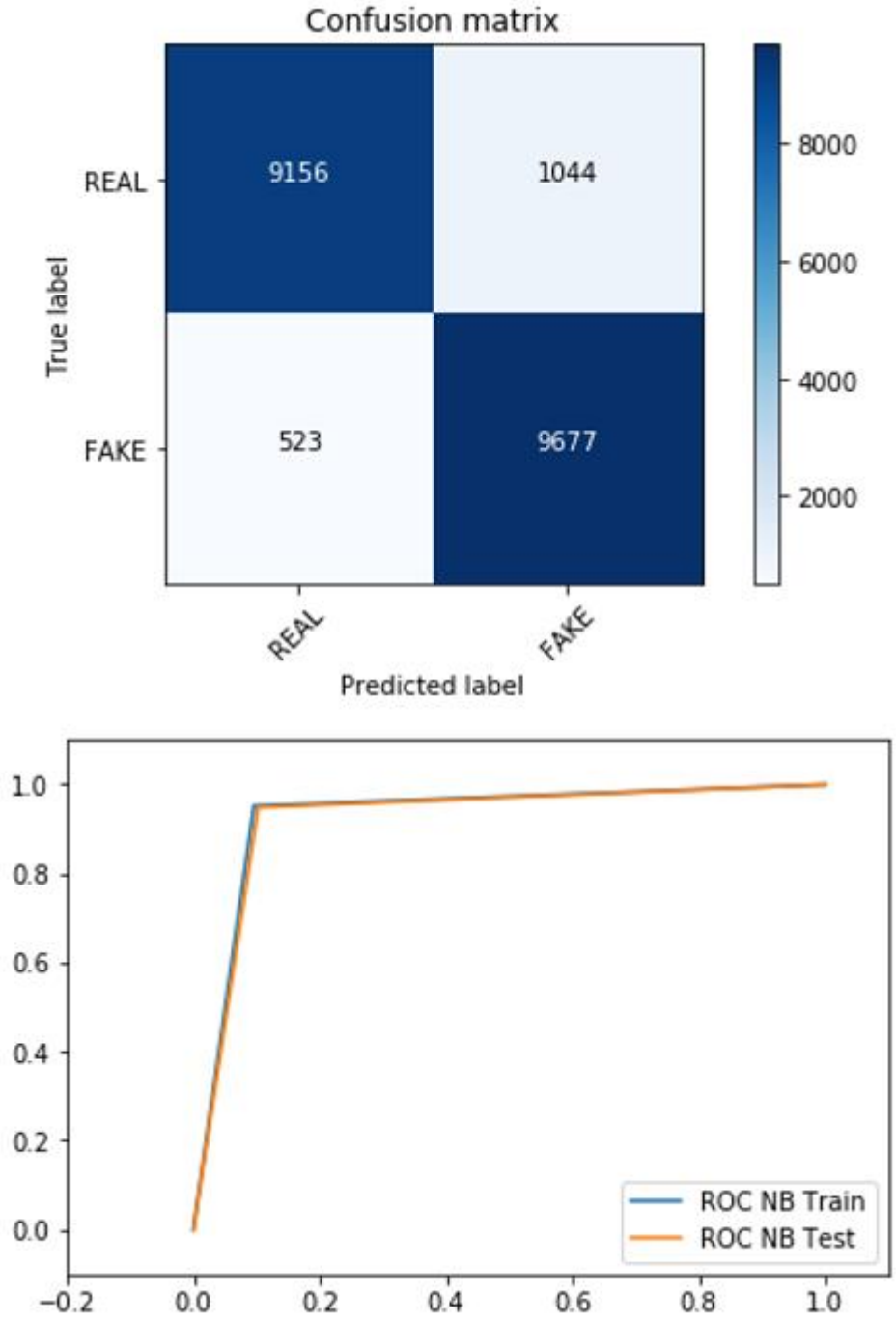
เพื่อให้เห็นภาพผลการทดลอง ผู้วิจัยได้นำ AUC (Area Under the ROC Curve) มาใช้เพื่อเป็นตัวบอกประสิทธิภาพในการทดสอบ (test performance) ว่าแบบจำลองสามารถระบุข่าวปลอมและข่าวจริงได้ดีมากน้อยเพียงใด และระบุจุดตัด (cut-off) ของการทดสอบที่มีความแม่นยำและน่าเชื่อถือมากที่สุดเพื่อนำค่าดังกล่าวมาใช้ในการจำแนกข่าวจริงและข่าวปลอมให้ถูกต้องมากที่สุดและผิดพลาดน้อยที่สุด

4.2 ผลการทดลอง

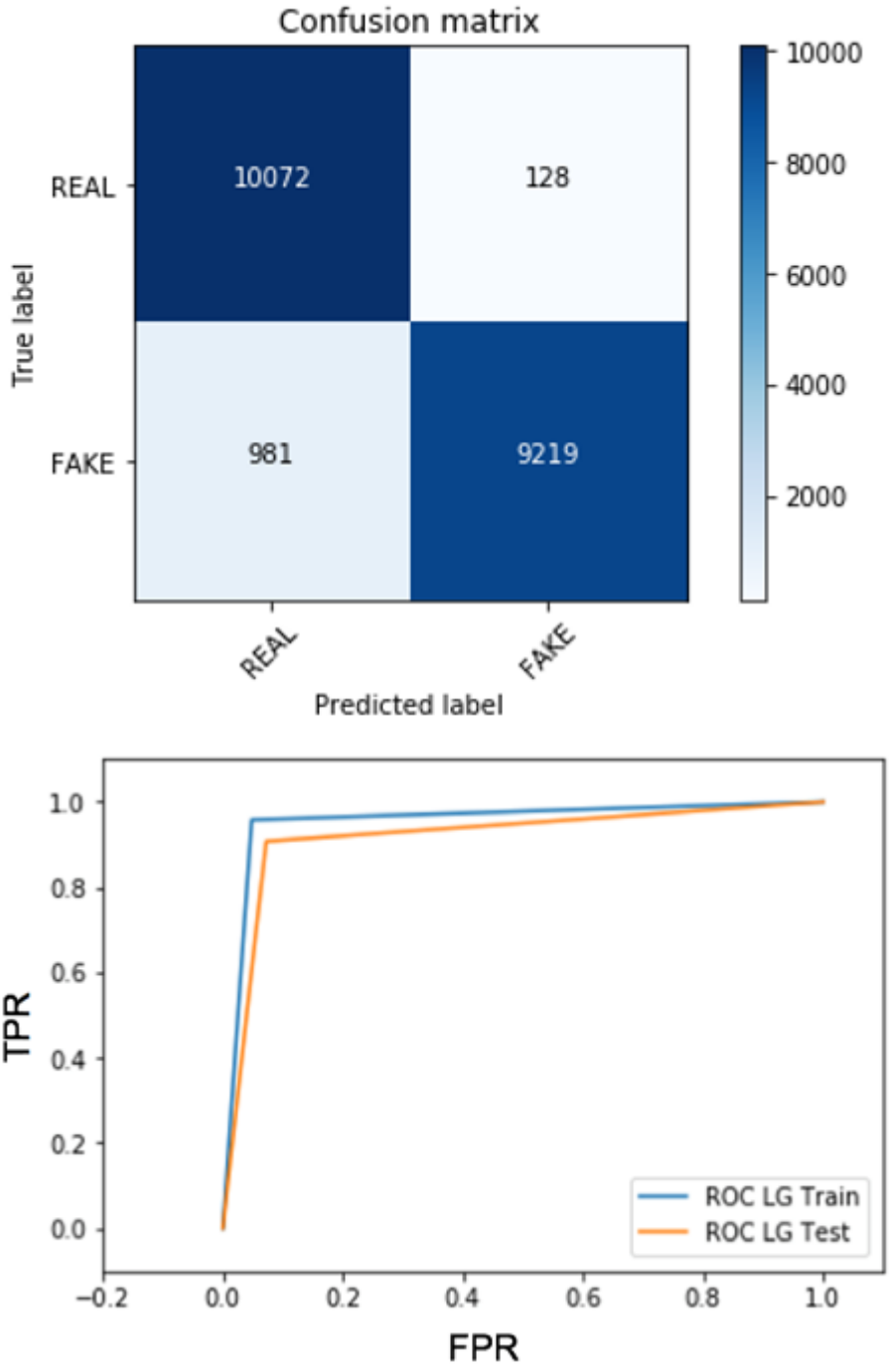
ในการหาค่าพารามิเตอร์ที่ดีที่สุดมาใช้กับการหาค่าความถี่ของคำ หรือ TF-IDF และเพื่อใช้กับแบบจำลองแบบต่าง ๆ ผู้วิจัยได้นำเทคนิค Cross Validation มาใช้ร่วมกับ Grid Search [11] ผลการวิจัยพบว่าแบบจำลองเทคนิคการถดถอยโลจิสติก สามารถจัดกลุ่มข่าวปลอมได้ดีที่สุดด้วยค่า AUC ร้อยละ 94 ตามมาด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีนร้อยละ 91 เทคนิคนาอีฟเบส์ร้อยละ 89 และเทคนิคตัวจำแนกป่าแบบสุ่มร้อยละ 88 ตามลำดับ ดังตารางที่ 4.2

ตารางที่ 4.2 แสดงผลการทดสอบแต่ละแบบจำลองการเรียนรู้เครื่อง

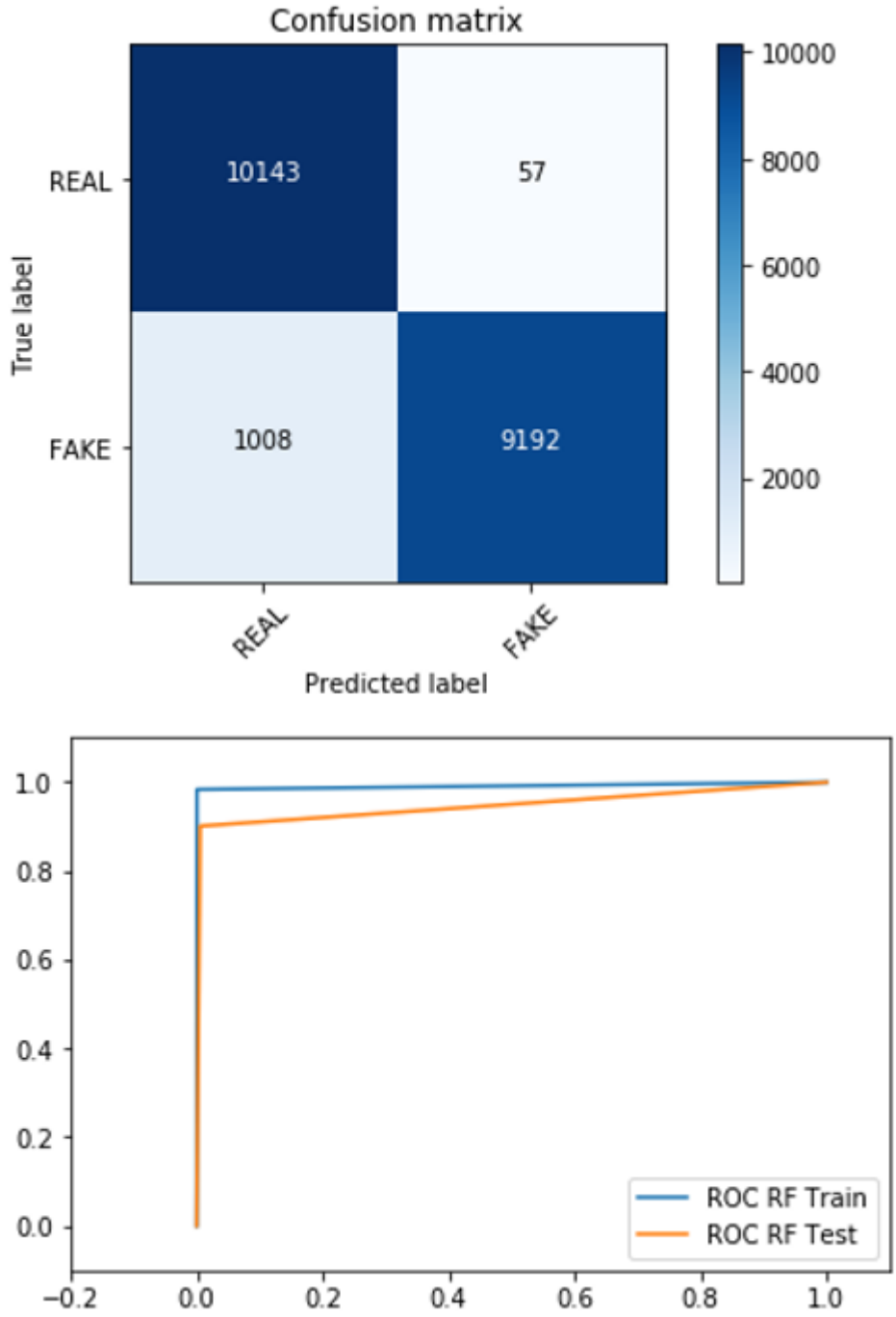
Model	Train (%)	Test (%)			
	ROC AUC	ROC AUC	Precision	Recall	F Measure
Logistic Regress	0.95	0.94	0.85	0.91	0.88
Random Forest	0.99	0.88	0.90	0.82	0.86
SVM	0.93	0.91	0.83	0.90	0.87
Naïve Bayes	0.92	0.89	0.88	0.82	0.85



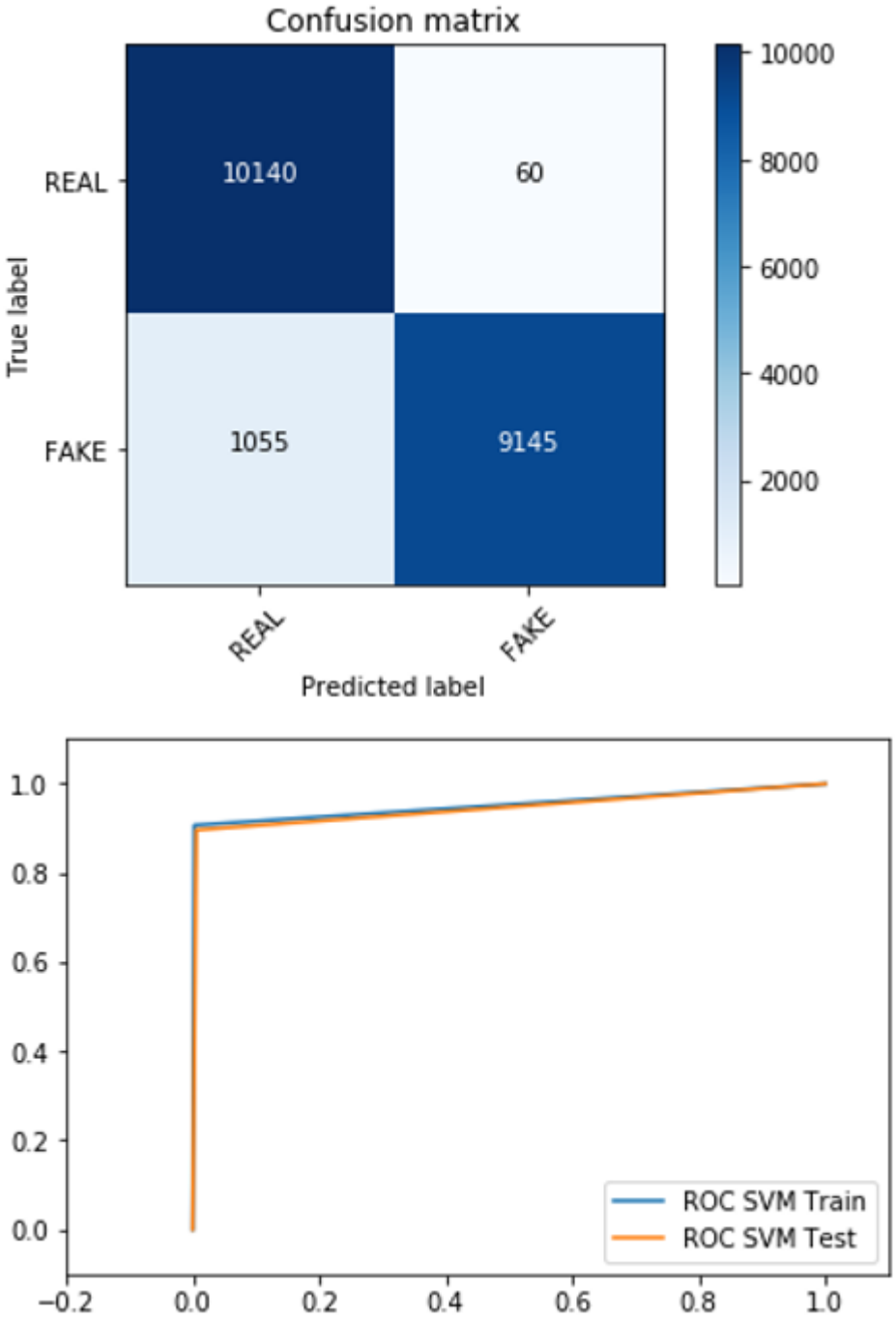
ภาพที่ 4.1 แสดงผลการทดลอง Confusion Matrix และ AUC/ROC ของแบบจำลองนาอีฟเบย์



ภาพที่ 4.2 แสดงผลการทดลอง Confusion Matrix และ AUC/ROC ของแบบจำลอง การถดถอยโลจิสติกส์



ภาพที่ 4.3 แสดงผลการทดลอง Confusion Matrix และ AUC/ROC ของแบบจำลอง ตัวจำแนกป่าแบบสุ่ม



ภาพที่ 4.4 แสดงผลการทดลอง Confusion Matrix และ AUC/ROC ของแบบจำลองซัพพอร์ทเวกเตอร์แมชชีน

บทที่ 5
บทสรุปและข้อเสนอแนะ

ในบทนี้เป็นบทสุดท้าย ประกอบด้วยสองส่วน คือ ส่วนแรกเป็นการสรุปผลการดำเนินงานของงานวิจัยนี้ และอภิปรายผลการดำเนินงานที่ได้ ในส่วนสุดท้ายเป็นข้อจำกัดต่าง ๆ รวมถึงแนวทางวิจัยต่อไปในอนาคต ดังรายละเอียดต่อไปนี้

5.1 สรุปผลงานวิจัย

งานวิจัยนี้ศึกษาการจัดกลุ่มข้อมูลเชิงพื้นที่แบบไม่มีลาเป้าหมายกำกับ เพื่อนำเป็นตัวแทนสำหรับการทำความเป็นกลุ่มก้อนของโครงการที่อยู่อาศัย โดยใช้เทคนิคอัลกอริทึมการจัดกลุ่ม ประกอบด้วย DBSCAN และ HDBSCAN จำนวนทั้งหมด 4 แบบจำลอง แล้ววัดผลด้วยดัชนีวัดคุณภาพ Silhouette coefficient, CH index, CDbw และ DBCV พบว่า การเรียนรู้ของเครื่องในการจัดกลุ่มข่าวปลอมเป็นเรื่องง่ายหากข้อมูลที่ใช้ในการจัดกลุ่มเป็นข้อมูลที่มีคำตอบ (label) ชัดเจน งานวิจัยชิ้นนี้มีวัตถุประสงค์เพื่อนำเสนอการจัดกลุ่มข่าวปลอมด้วยวิธีการเรียนรู้ของเครื่อง ประกอบด้วย 4 วิธี คือ การถดถอยโลจิสติก (Logistic Regression) นาอีฟเบส์ (Naïve Bayes) ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine) และ ตัวจำแนกป่าแบบสุ่ม (Random Forest) โดยผลวิธีการจัดกลุ่มข่าวปลอม การถดถอยโลจิสติกได้ค่าความถูกต้องมากที่สุด คือร้อยละ 94 ส่วนซัพพอร์ทเวกเตอร์แมชชีน นาอีฟเบส์ และตัวจำแนกป่าสุ่มมีผลต่างกันเพียงเล็กน้อย ความถูกต้องอยู่ที่ร้อยละ 91 ร้อยละ 89 และ ร้อยละ 88 ตามลำดับ ค่าประสิทธิภาพความถูกต้องของแบบจำลองเมื่อเปรียบเทียบด้วยค่า Precision Recall และ F-Measure ของแบบจำลองการถดถอยโลจิสติก คือ ร้อยละ 85 ร้อยละ 91 และร้อยละ 88 ตามลำดับ

5.2 ข้อจำกัดและแนวทางวิจัยในอนาคต

ในงานวิจัยนี้ ใช้วิธีการจัดกลุ่มข่าวปลอมโดยวิเคราะห์จากหัวข้อข่าวและเนื้อหาของข่าวโดยสนใจการวิเคราะห์หัวข้อข่าวและเนื้อหาของข่าวในด้านการวิเคราะห์ทางภาษาศาสตร์ ที่สามารถเป็นตัวแทนของข่าวปลอม และสามารถแยกข่าวปลอมออกจากข่าวจริง โดยไม่ได้สนใจการวิเคราะห์คุณลักษณะอื่นของข่าว ในอนาคตควรเพิ่มการวิเคราะห์จากคุณลักษณะอื่น ๆ ของข่าวร่วมด้วย เพื่อปรับปรุงประสิทธิภาพในการจำแนกประเภทของข่าวปลอมให้มีความแม่นยำมากขึ้น

บรรณานุกรม

บรรณานุกรม

[1] กานท์กลอน รักธรรม. *“คุณคือสำนักข่าวปลอม!”* ทรัมป์จวก CNN และ BuzzFeed คือสื่อขยะ!. เข้าถึงเมื่อ: 5 มกราคม 2563. เข้าถึงได้: https://themomentum.co/momentum-feature-fake-news-cnnbuzzfeed/

[2] กองบรรณาธิการจุลสารราชดำเนิน. *“เฟค นิวส์ (Fake News) วัชพืชแห่งวารสารศาสตร์”* [จุลสาร]. กรุงเทพฯ: สมาคมนักข่าวนักหนังสือพิมพ์แห่งประเทศไทย. 2560

[3] ณภัชชา นิลแก้ว และคณะ. *“Fake News วิกฤตศรัทธาต่อองค์กรสื่อ”*. ในการสัมมนาทางวิชาการเรื่อง Fake News วิกฤตการสื่อสารในยุคดิจิทัล (หน้า 1-7). นนทบุรี: มหาวิทยาลัยสุโขทัย ธรรมาธิราช.

[4] ปิยพร อรุณเกรียงไกร. *“คุณกำลังถูกเฟซบุ๊กต้มตุ๋น!? วิกฤตข่าวปลอมทะลักโลกออนไลน์ เมื่อวิจารณญาณก็อาจไม่เพียงพอ”*. เข้าถึงเมื่อ: 15 มกราคม 2560. เข้าถึงได้: https://themomentum.co/momentum-feature-social-media-s-algorithm/

[5] นันทิกา หนูสม. *“ลักษณะของข่าวปลอมในประเทศไทยและระดับความรู้เท่าทันข่าวปลอมบนเฟซบุ๊กของผู้รับสารในเขตกรุงเทพมหานคร”*. นิเทศศาสตรมหาบัณฑิต สาขาวิชาการสื่อสารเชิงกลยุทธ, มหาวิทยาลัยกรุงเทพ. 2560

[6] พิณ พัฒนา. *“รู้เขาหลอกแต่เต็มใจให้หลอก รู้จัก ‘Fake News’ ข่าวปลอมออนไลน์ที่เราชักเจอจริงบ่อยขึ้นทุกวัน”*. เข้าถึงเมื่อ: 15 มกราคม 2560. เข้าถึงได้: http://www.kmutt.ac.th/organization/ssc334/asset5.html

[7] พีรพล อนุตรโสตถิ์. *“ทำไมสังคมไทยต้อง #ชัวร์ก่อนแชร์”*. ศูนย์ชัวร์ก่อนแชร์ สำนักข่าวไทย อสมท. (มมป.)

[8] ยุทธ ไกยวรรณ์ *“หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย”*, วารสารวิจัย มหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย. 2555

[9] วิริยาภรณ์ ทองสุข. *“การวิเคราะห์การนำเสนอข่าวเชิงคลิกเบทของเว็บไซต์”*. สาขาวิชาการบริหารสื่อสารมวลชน, คณะวารสารศาสตร์และการสื่อสารมวลชน, มหาวิทยาลัยธรรมศาสตร์. 2559

[10] ศุปภศิลป์ กุลจิตต์เจือวงศ์ . *“การวิเคาะห์ผู้รับสารในยุคดิจิทัล”*. สาขาวิชาการจัดการการสื่อสาร, มหาวิทยาลัยราชภัฏรำไพพรรณี. 2560

[11] ศูนย์สำรวจความคิดเห็นบ้านสมเด็จโพลล์ สถาบันวิจัยและพัฒนา มหาวิทยาลัยราชภัฏบ้านสมเด็จเจ้าพระยา *“คน กทม 85.1 % เคยพบเห็นข่าวปลอม (Fake News) เจอข่าวปลอม (Fake News) จากเฟสบุ๊ค (Facebook) และเรื่องการเมืองมากสุด”* เข้าถึงเมื่อ: 5 มกราคม 2563. เข้าถึงได้: http://brms.bsru.ac.th/download/8-BSRU-poll/62-2-28.pdf

บรรณานุกรม (ต่อ)

[12] สุกัญญา บูรณเดชาชัย. *“ไม่ชัวร์แชร์ไป...สังคมวุ่นวาย”*. เข้าถึงเมื่อ: 5 มกราคม 2563. เข้าถึงได้: http://imgs.mcot.net/images/2018/05/ 1525684457247.pdf

[13] สุปัญญา อภิวงศ์โสภณ (2561) *“การตรวจสอบข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง”*. ภาควิชาวิศวกรรมคอมพิวเตอร์, คณะวิศวกรรมศาสตร์, จุฬาลงกรณ์มหาวิทยาลัย. 2561

[14] อิสริยะ ไพรีพ่ายฤทธิ์. (2560). *“Fake News ข่าวปลอม ปัญหาใหญ่ของโลกอินเทอร์เน็ต”*. เข้าถึงเมื่อ: 12 กุมภาพันธ์ 2561. เข้าถึงได้: http://www.okmd.tv/ blogs/ all-things-digital/fake-news-ข่าวปลอม-ปัญหาใหญ่ของโลก

[15] Maciej Szpakowski. *“Fake News Corpus”*. 2018. Accessed On: Feb 12, 2018. Available URL: https://github.com/several27/FakeNewsCorpus

[16] Hunt Allcott and Matthew Gentzkow. *“Social Media and Fake News in the 2016 Election”*. Journal of Economic Perspectives—Volume 31, Number 2-p. 211–236, 2017

[17] Desai, S., Mooney, H., & Oehrli, J. A. “Fake News, Lies and Propaganda: How to Sort Fact from Fiction”. 2017. Accessed On: Jan 5, 2020 Available URL: http://guides.lib.umich.edu/fakenews .

[18] The European Association for Viewers Interests. Infographic: Beyond FakeNews – 10 Types of Misleading News. 2016. Accessed On: Jan 5, 2020. Available URL: https://eavi.eu/beyondfake-news-10-types-misleading-info/

[19] Nurse, M. Fake news and other types of misinformation defined. 2016. Accessed On: Jan 6 2020. Available URL: http://communicationscience.org.au/fake-news-and-other-forms-ofmisinformation-defined/ .

[20] Sunnywalker. *“สรุปปัญหาข่าวปลอมที่รุนแรง จนบริษัทโซเชียลต้องกลับไปรื้อนโยบาย ทบทวนตัวเองใหม่”*. Accessed On: Jan 6 2020. Available URL: https://www.blognone.com/node/96867

[21] Bing Liu. *"The Science of Detecting Fake Reviews"*. Accessed On: Jan 4, 2021 Available URL: https://content26.com/blog/bing-liu-the-science-of-detecting-fake-reviews/

[22] Zhen Hai, Peilin Zhao, Peng Cheng, Peng Yang, XiaoLi Li, Guangxia Li, and Ant Financial. 2016. “*Deceptive review spam detection via exploiting task relatedness and unlabeled data. In EMNLP”*., 2016

บรรณานุกรม (ต่อ)

[23] Giovanni Luca Ciampaglia. 2020. *“Finding Streams in Knowledge Graphs to Support Fact Checking”*

[24] Hunt Allcott Matthew Gentzkow. *“Social Media and Fake News in the 2016 Election”* JOURNAL OF ECONOMIC PERSPECTIVES VOL. 31, NO. 2, SPRING 2017 (p. 211-36)

ภาคผนวก

ภาคผนวก ก

ผลงานตีพิมพ์

Friday 6th July 2018		
	Room I: DATA SCIENCE AND MACHINE LEARNING	**Page**
09:00 – 09:20 NCCIT2018-72	**Detection and Classification of Vehicles using Deep Learning Algorithm for Video Surveillance Systems** *Phonratichi Wairotchanaphuttha, Narong Boonsirisumpun, Wichai Puarungroj*	402
09:20 – 09:40 NCCIT2018-13	**Opinion Analysis System to Business by Text Mining On Twitter** *Chanita Lisirikul, Jeerasak Numpradit*	408
09:40 – 10:00 NCCIT2018-158	**Effective Comparison of Algorithm for Analysis Defect of Software Testing Process using Data Mining Technique** *Chanyanut Kaewtip, Sakchai Tangwannawit*	414
10:00 – 10:20	*Coffee Break*	
10:20 – 10:40 NCCIT2018-117	**Deep Learning Method for Recommender System** *Suphawit Wongadoonwit, Saharat Haritrak, Assadavud Thongthammachad, Akadej Udomchaiporn*	420
10:40 – 11:00 NCCIT2018-147	**Factor Extraction Method to Fault Diagnosis in Accident of Railway System Using Text Mining** *Thanyapawn Kransuk, Tanapon Jensuttiwetchakul*	426
11:00 – 11:20 NCCIT2018-194	**Comparisons of Predictive Models of High Vocational Certificate Students entering the University Using Data Mining Techniques** *Patcharanikarn Pongthanoo, Jim Yuennan*	432
11:20 – 11:40 NCCIT2018-160	**Fake News Classification using Machine Learning Technique** *Wisith Wanishyanon, Ratthaslip Ranokphanuwat, Worapol Pongpech*	438
11:40 – 12:00 NCCIT2018-150	**Customer Segmentation Model for e-Banking Transaction using Data Mining** *Chutimon Chaiyo, Sakchai Tangwannawit*	444
12:00 – 13:00	*Lunch*	
13:00 – 13:20 NCCIT2018-134	**Applied an Artificial Intelligence (AI) with Chatbot's Facebook Messenger to Help Sales Management for SME 4.0** *Chatchitsanu Pothisakha, Jeerasak Numpradit*	450
13:20 – 13:40 NCCIT2018-12	**Automatic Text Clustering System using Machine Learning Techniques** *Peeraphat Komolruchinonth, Wisarut Yawut, Wanthanee Prachuabsupakij*	456
13:40 – 14:00 NCCIT2018-250	**Applying Clustering Algorithm for Primary Screening of Patients** *Sutat Gammanee, Sunantha Sodsee*	462
14:00 – 14:20 NCCIT2018-219	**A Factor Analysis of Not Renew Policy Using Data Mining Technique** *Matee Iemprapai, Tanapon Jensuttiwetchakul*	468
14:20 – 14:40 NCCIT2018-174	**A Comparison of Data Mining Algorithms for Vehicle Type Classification Model** *Ngamjit Phueknarin, Mahasak Ketcham*	474
14:40 – 15:00 NCCIT2018-204	**Development of a Sustainable Water Management System for Agriculture by Using Data Mining Techniques** *Anon Eagtasi, Jeerasak Numpradit*	480
15:00 – 15:20	*Coffee Break*	

การจัดกลุ่มข่าวปลอมด้วยเทคนิคการเรียนรู้เครื่อง
Fake News Classification using Machine Learning Technique

วิสิทธิ์ วาณิชยานนท์ (Wisith Wanishyanon)[1] รัฐศิลป์ รานอกภานุวัชร์ (Ratthaslip Ranokphanuwat)[2]
และ วรพล พงษ์เพ็ชร (Worapol Pongpech)[3]
[1,2] สาขาวิชาวิศวกรรมข้อมูลขนาดใหญ่ วิทยาลัยนวัตกรรมด้านเทคโนโลยีและวิศวกรรมศาสตร์ มหาวิทยาลัยธุรกิจบัณฑิตย์
[3]คณะสถิติประยุกต์ (GSAS) สถาบันบัณฑิตพัฒนบริหารศาสตร์
[1]595162020005@dpu.ac.th, [2]udom.ran@dpu.ac.th, [3]worapol@as.nida.ac.th

บทคัดย่อ

ในสถานการณ์ที่เป็นที่สนใจของคนทั่วไป ผู้ใช้ในโลกสังคมออนไลน์มักเชื่อเนื้อหาข่าวที่เกี่ยวข้องกับเหตุการณ์นั้น ๆ ได้ง่ายและส่งต่อไปในวงกว้าง แต่เป็นที่น่าเสียดายที่ผู้ใช้โดยมากมักมีความเข้าใจเกี่ยวกับเหตุการณ์น้อยเกินไปหรือไม่ได้ตรวจสอบความถูกต้องของเนื้อหาทำให้กลายเป็นการกระจายข่าวปลอม ดังนั้นเพื่อหาเทคนิคการเรียนรู้เครื่องที่ดีที่สุดในการนำมาสร้างแบบจำลองการจัดกลุ่มข่าวปลอม ผู้วิจัยใช้เทคนิคการเรียนรู้เครื่อง 4 เทคนิค คือ การถดถอยโลจิสติก นาอีฟเบส์ ซัพพอร์ทเวกเตอร์แมชชีน และต้นไม้แบบสุ่ม โดยใช้ชุดข้อมูลจาก Fake Corpus สุ่มเลือกข้อมูลที่เป็น reliable และ fake ประเภทละ 32,000 รายการ ประเมินผลแบบจำลองโดยวัดประสิทธิภาพการจำแนกตามแนวคิดการค้นคืนสารสนเทศโดยใช้ค่า AUC (Area Under the ROC Curve) ซึ่งสร้างจากการจำแนกประเภทข้อมูล จนสามารถทำนายข้อมูลใหม่ได้ ซึ่งผลการวิจัยพบว่าแบบจำลองเทคนิคการถดถอยโลจิสติกสามารถจัดกลุ่มข่าวปลอมได้ดีที่สุดด้วยค่า AUC ร้อยละ 94 ตามด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีนร้อยละ 91 เทคนิคนาอีฟเบส์ร้อยละ 89 และเทคนิคต้นไม้แบบสุ่มร้อยละ 88 ตามลำดับ

คำสำคัญ: ข่าวปลอม เทคนิคการเรียนรู้เครื่อง นาอีฟเบส์ การถดถอยโลจิสติก ซัพพอร์ทเวกเตอร์แมชชีน ต้นไม้แบบสุ่ม การจัดกลุ่ม

Abstract

Fake news tend to gain a lot of traction in stressful or emotional situations. This is even more so in the present faster pace social networking environment where fake news can spread like a wildfire through the network. Unfortunately, most of the fake news stories were shared before adequate contemplation on the news's integrity was conducted. Moreover, with the lighting pace of information flowing through social networking, most news once shared is out of sight out of mind. Consequently, relying on human to screen out fake news ourselves might not be very conductive and productive screening instrument. To enable automated fake news detection, this research focuses on utilizing machine learning techniques to construct fake news classification models. In this paper we created a fake news classification model using machine learning techniques with 4 techniques, Logistic regression, Naïve Bayes, Support Vector Machine and Random Forest. The experiments were conducted equally random data from Fake Corpus. As result, 32,000 records of reliable and fake news were generated. Classification performance was evaluated by using the AUC (Area Under the ROC Curve). The result of this study showed that the Logistic Regression Model can classify fake news with 94% of the AUC was classified, following by Support Vector Machine 91%, Naïve Bayes 89% and Random Forest 88% respectively.

Keyword: Fake News, Machine Learning, Naïve Bayes, Logistics Regression, Support Vector Machine, Random Forest, Classification

1. บทนำ

จากการเกิดขึ้นของเครือข่ายอินเตอร์เน็ตทำให้การติดต่อสื่อสารและการรับส่งข่าวสารสะดวกรวดเร็ว ประชาชน

จึงทันมานิยมรับข่าวสารผ่านทางสังคมออนไลน์ หรือ Social Network เพิ่มขึ้น ในทางกลับกันมีผู้ที่ใช้ช่องทางนี้เป็นเครื่องมือสร้างผลประโยชน์ให้กับตนเองด้วยวิธีการต่าง ๆ มากมาย โดยเฉพาะข่าวปลอม ดังปรากฏการณ์สำคัญ เช่น การลงประชามติแยกตัวออกจากสหภาพยุโรปของประเทศสหราชอาณาจักร [1] ผลการเลือกตั้งของประเทศสหรัฐอเมริกา [2] และประเทศฝรั่งเศส [3] หรือแม้กระทั่งกรณีสาวรรคดของพระบาทสมเด็จพระปรมินทรมหาภูมิพลอดุลยเดช [4] จะเห็นได้ว่าข่าวปลอมมีอิทธิพลต่อการการรับรู้ การตัดสินใจและพฤติกรรมของผู้ใช้ในสื่อสังคมออนไลน์เป็นอย่างยิ่ง

ข่าวปลอม (Fake News) คือข่าวที่ถูกสร้างขึ้นมาโดยที่ไม่ได้มีเหตุการณ์ของข่าวเกิดขึ้นจริง หรือเป็นข่าวที่มีเนื้อหาไม่ถูกต้อง บิดเบือน ขาดความเป็นกลางแม้ว่าจะดูว่ามีลักษณะเป็นมืออาชีพ ดังนั้นหากสามารถระบุได้ว่าข่าวใดเป็นข่าวจริงหรือข่าวใดเป็นข่าวปลอมได้อย่างทันท่วงที ก็อาจสามารถหยุดยั้งการแพร่กระจายของข่าวปลอมและลดผลกระทบทางสังคมได้

ในส่วนที่ 2 ผู้วิจัยนำเสนองานของผู้วิจัยท่านอื่นที่ผ่านมาในด้านการจำแนกกลุ่มข้อมูลโดยใช้เทคนิคการเรียนรู้เครื่อง ส่วนที่ 3 อธิบายชุดข้อมูลที่ใช้สำหรับการทดสอบ วิธีการสร้างคุณลักษณะ ขั้นตอนการประมวลผล การสร้างแบบจำลองที่เกิดขึ้นจริงโดยในงานวิจัยชิ้นนี้ ผู้วิจัยได้ทำการศึกษาเปรียบเทียบประสิทธิภาพของแบบจำลองการจัดกลุ่มด้วยเทคนิคการเรียนรู้เครื่อง 4 แบบจำลองด้วยกัน คือ การถดถอยโลจิสติก นาอีฟเบส์ ซัพพอร์ทเวกเตอร์แมชชีน และตัวจำแนกป่าแบบสุ่ม และส่วนที่ 4 สรุปข้อมูลเพื่อให้เกิดการอภิปรายและเพื่อใช้ในการวิจัยครั้งต่อไป

2. ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 ข่าวปลอม (Fake News)

ข่าวปลอม หมายถึง สื่อหรือเนื้อหาของข่าวที่ถูกแต่งขึ้นทั้งหมด อาจมีความจริงเพียงเล็กน้อยหรือไม่มีเลยมีวัตถุประสงค์เพื่อหลอกลวง บิดเบือน โฆษณาชวนเชื่อ กระตุ้นการเข้าชมเว็บไซต์และเพื่อให้เกิดการแพร่กระจายข่าวผ่านทางสื่อสังคมออนไลน์โดยหวังผลให้ผู้รับข่าวปลอมเกิดการเข้าใจผิดมากกว่าความบันเทิง ทั้งนี้วัตถุประสงค์หลักของข่าวปลอมอาจมุ่งหวังผลประโยชน์ทางการเงิน ธุรกิจ การเมือง การทหาร หรือแม้กระทั่งความมั่นคงของชาติ การนำเสนอมีความเป็นมืออาชีพ ทั้งการจงใจเลียนแบบสื่อหนังสือพิมพ์ที่มีอยู่จริงไปจนถึงสื่อโฆษณาชวนเชื่อของรัฐบาลจนทำให้ผู้ใช้ไม่สามารถแยกแยะระหว่างข่าวจริงกับข่าวปลอมออกจากกันได้ [5][6][7]

2.2 การสกัดคำคุณลักษณะ

การสกัดคุณลักษณะของบทความหรือข่าวคือการสร้างตัวแทนคุณลักษณะของเอกสารซึ่งอาจจะใช้คำเดี่ยว วลี หรือประโยค คุณลักษณะที่สกัดได้จะถูกจัดรูปแบบให้อยู่ในลักษณะของเวกเตอร์และถูกแทนด้วยลักษณะของค่าความจริงหรือแทนด้วยค่าความถี่ของคำ [8] สำหรับงานวิจัยชิ้นนี้ใช้คำเดี่ยวที่ได้จากกระบวนการตัดคำเป็นคุณลักษณะและการหาค่าน้ำหนักของคำในการสกัดคุณลักษณะ จากนั้นจึงลดคุณลักษณะของเอกสารด้วยการละเว้นคำที่เป็นคำบุพบทและคำสันธาน [9]

2.3 การสร้างตัวแทนเอกสาร

เพื่อให้คอมพิวเตอร์เข้าใจภาษามนุษย์ การเรียนรู้ของเครื่องจึงนิยมใช้ลักษณะตัวแทนของคำหรือข้อความมากกว่าความหมายของคำ ซึ่งตัวแทนของคำหรือข้อความมักอยู่ในรูปของเวกเตอร์ของน้ำหนักคำ โดยมากใช้ค่าน้ำหนักของคำเป็นไบนารีหรือไม่เป็นไบนารีก็ได้ขึ้นอยู่กับวิธีการคำนวณ สำหรับงานวิจัยชิ้นนี้จะใช้วิธีการคำนวณค่าน้ำหนักด้วยวิธีไบนารี

2.4 เทคนิคการเรียนรู้เครื่อง (Machine Learning)

การเรียนรู้เครื่องคือการทำให้เครื่องเรียนรู้จากข้อมูลตัวอย่างหรือสภาพแวดล้อม จุดมุ่งหมายคือการปรับปรุงประสิทธิภาพการทำงานของระบบให้ดีขึ้น เมื่อเรียนรู้แล้วจึงจัดเก็บข้อมูลไว้ในฐานความรู้ด้วยรูปแบบอย่างใดอย่างหนึ่ง เช่น กฎ ฟังก์ชั่น แบ่งออกเป็นแบบมีผู้สอน (Supervised Leaning) และแบบไม่มีผู้สอน (Unsupervised Leaning) โดยในงานวิจัยชิ้นนี้ใช้แบบมีผู้สอน ประกอบด้วยเทคนิคการถดถอยโลจิสติก นาอีฟเบส์ ซัพพอร์ทเวกเตอร์แมชชีน และตัวจำแนกป่าแบบสุ่ม

2.4.1 การถดถอยโลจิสติก (Logistic Regression) [8] เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุประเภทหนึ่ง มีวัตถุประสงค์ในการหาระนาบการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วนเพื่อประมาณหรือทำนายความน่าจะเป็นของการเกิดหรือไม่เกิดเหตุการณ์ที่กำลังสนใจ มีงานวิจัยจำนวนมากที่ใช้เทคนิคนี้ในการวิเคราะห์ข้อมูลและแสดงให้เห็นถึงความมั่นใจและความน่าเชื่อถือ

2.4.2 นาอีฟเบส์ (Naïve Bayes) [8] เป็นขั้นตอนวิธีที่ได้รับความนิยมและถูกนำมาใช้อย่างแพร่หลายในงานจำแนกหมวดหมู่เอกสาร เนื่องจากความเรียบง่ายของขั้นตอนวิธีและให้ประสิทธิภาพการจำแนกที่ดี นาอีฟเบส์เป็นขั้นตอนวิธีที่มีพื้นฐานมาจากทฤษฎีเบส์ (Bayes' Theorem) ซึ่งอาศัยหลักความน่าจะเป็นในการทำนายผลลัพธ์ โดยการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเพื่อใช้ในการสร้างเงื่อนไขความน่าจะเป็นสำหรับแต่ละความสัมพันธ์

2.4.3 ซัพพอร์ทเวกเตอร์แมชชีน (Support Vector Machine: SVM) [8][10][11][12] เป็นวิธีการจำแนกกลุ่มข้อมูลที่อาศัยระนาบการตัดสินใจมาใช้ในการแบ่งข้อมูลออกเป็น 2 ส่วน โดยพยายามสร้างเส้นแบ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตทั้งสองกลุ่มมากที่สุด ซึ่งซัพพอร์ทเวกเตอร์แมชชีนจะใช้ฟังก์ชั่นแม็พพิ้ง (Mapping Function) เพื่อแปลงข้อมูลจากโดเมนเดิมไปยังโดเมนที่เรียกว่า ฟีเจอร์ สเปซ (Feature Space) และใช้ฟังก์ชั่นเคอร์เนล (Kernel Function) ในการวัดความคล้ายกันของข้อมูลในฟีเจอร์ สเปซ ในงานวิจัยนี้ใช้เคอร์เนลฟังก์ชั่น คือ โพลิโนเมียล เคอร์เนล (Polynomial Kernel) เนื่องจากการเป็นวิธีที่ดีที่สุด

2.4.4 ตัวจำแนกป่าแบบสุ่ม (Random Forest) [12] คือการสร้างแบบจำลองด้วยการนำแนวคิดโมเดลต้นไม้ตัดสินใจ (Decision Tree) เข้ามาใช้ โดยจะทำการสร้างโมเดลต้นไม้ตัดสินใจขึ้นมาหลาย ๆ โมเดลโดยการสุ่มตัวแปร แล้วนำผลที่ได้แต่ละโมเดลมารวมกันพร้อมกับคำนวณผลที่มีจำนวนซ้ำกันมากที่สุด สกัดออกมาเป็นผลลัพธ์สุดท้าย วิธีการของโมเดลต้นไม้ประกอบไปด้วยโหนด (Node) และกิ่ง (Branch) แต่ละโหนดจะถูกแทนด้วยคุณลักษณะ (Feature) ของชุดข้อมูลที่นำมาเรียนรู้และทดสอบ แต่ละกิ่งของต้นไม้แสดงผลในการทดสอบ และลีฟโหนด (Leaf Node) แสดงหมวดหมู่ที่ผู้ใช้กำหนด ส่วนเกณฑ์การเลือกคุณลักษณะเพื่อนำมาเป็นโหนดของต้นไม้นั้นมาจากการคำนวณค่าเกนสารสนเทศ (Information Gain) โดยพิจารณาคุณลักษณะที่มีค่าเกนสารสนเทศหรือมีค่าเอ็นโทรพี (Entropy) ต่ำ หมายความว่าคุณลักษณะนั้นมีความสามารถในการจำแนกหมวดหมู่สูง

2.5 งานวิจัยที่เกี่ยวข้อง

Victoria L. Rubin และคณะ [13] ได้ทำการวิจัยโดยใช้ซัพพอร์ทเวกเตอร์แมชชีน เพื่อทำนายลักษณะและจัดกลุ่มข่าวออกเป็น 5 ประเภท จากข่าวทั้งหมด 360 ข่าว ผลการวิจัยสามารถจัดกลุ่มได้อย่างแม่นยำถึงร้อยละ 90 ซึ่งผลการวิจัยชิ้นนี้สามารถช่วยลดผลกระทบที่อาจเกิดขึ้นจากข่าวปลอม

กานต์ อัจฉราภรณ์ [14] ได้ศึกษาเปรียบเทียบแบบจำลองการจำแนกกลุ่มโรงงานอุตสาหกรรมโดยใช้เทคนิคการทำเหมืองข้อความ กรณีศึกษาโรงงานอุตสาหกรรมจังหวัดปทุมธานี พบว่าแบบจำลองนาอีฟเบส์มีความเหมาะสมในการจำแนกกลุ่มโรงงานอุตสาหกรรมโดยมีค่าความถูกต้องร้อยละ 62.82 ซึ่งดีกว่าแบบจำลองต้นไม้ตัดสินใจและซัพพอร์ทเวกเตอร์แมชชีน นอกจากนี้ แบบจำลองนาอีฟเบส์ยังมีการทำงานที่ง่ายและให้ผลการวิเคราะห์ได้อย่างรวดเร็ว

สุพัตรา วิริยะวิสุทธิกุลและคณะ [15]ได้ทำการศึกษาและนำเสนอระบบแจ้งเตือนธุรกิจบนสื่อสังคมออนไลน์โดยเก็บความคิดเห็นภาษาไทยที่เกี่ยวข้องกับธุรกิจจากเว็บบอร์ดพันธ์ทิพย์นำมาจำแนกเป็น 2 กลุ่มคือเชิงลบและเชิงไม่ลบ ใช้ TF-IDF ในการสกัดคุณลักษณะและใช้ซัพพอร์ทเวกเตอร์แมชชีนเพื่อจำแนกความคิดเห็น ผลการวิจัยพบว่าหากความคิดเห็นเป็นเชิงลบระบบจะทำการแจ้งเตือนไปยังธุรกิจ แต่หากเป็นความคิดเห็นเชิงไม่ลบระบบจะทำการระบุและเก็บข้อมูลไว้ให้สามารถนำมาวิเคราะห์ได้ในภายหลัง ค่า ROC คิดเป็นร้อยละ 85

3. วิธีดำเนินการวิจัย

ขั้นตอนการจัดกลุ่มข้อมูลข่าวปลอมแบ่งออกเป็น 4 ขั้นตอน คือ 1) การเตรียมข้อมูลและนำมาจัดกลุ่ม 2) ขั้นตอนก่อนกระบวนการประมวลผล (Pre-processing) การนำข้อมูลที่ถูกจัดกลุ่มมาตัดคำและกำกับหน้าที่ของคำ จากนั้นทำกระบวนการสกัดคำคุณลักษณะเพื่อสร้างเวกเตอร์ข้อมูล พร้อมใส่น้ำหนักของคุณลักษณะ 3) การสร้างแบบจำลองโดยนำข้อมูลประมวลผลด้วยเทคนิคการจัดกลุ่ม 4 เทคนิค ได้แก่ การถดถอยโลจิสติก นาอีฟเบส์ ตัวจำแนกป่าแบบสุ่ม และซัพพอร์ทเวกเตอร์แมชชีน 4) การประเมินผลแบบจำลองโดยวัดประสิทธิภาพการจำแนกตามแนวคิดการค้นคืนสารสนเทศ

3.1 การเตรียมข้อมูล

ในงานวิจัยชิ้นนี้ ผู้วิจัยได้ใช้ชุดข้อมูลสาธารณะ Fake Corpus [16] ซึ่งมีผู้รวบรวมไว้ จำนวน 9,408,908 ข่าว ประกอบด้วยบทความหรือข่าวที่คัดลอกมาจากรายชื่อเว็บไซต์ที่เป็นแหล่งข่าวปลอม กับรายชื่อเว็บไซต์ที่มีเนื้อหาน่าเชื่อถือ จำนวน 1,001 เว็บไซต์ โดยใช้ชุดไลบรารี่ของ python ที่ชื่อ Scrapy แล้วจึงนำมาบันทึกข้อมูลตามรูปแบบที่ต้องการด้วยชุดไลบรารี่ของ python ที่ชื่อ newspaper โดยแต่ละบทความจะถูกติดป้ายกำกับไว้ตามความน่าเชื่อถือหรือประเภทของเว็บไซต์ มีจำนวนดังแสดงดังตารางที่ 1

ตารางที่ 1 จำนวนบทความหรือข่าว แยกตามประเภท

ประเภท	ป้ายกำกับ	จำนวน
Fake News	fake	928,083
Satire	satire	146,080
Extreme Bias	bias	1,300,444
Conspiracy Theory	conspiracy	905,981
Junk Science	junksci	144,939
Hate News	hate	117,374
Clickbait	clickbait	292,201
Proceed With Caution	unreliable	319,830
Political	political	2,435,471
Credible	reliable	1,920,139
State News	state	0

3.2 การแปลงข้อความเป็นคุณลักษณะ (Convert Text to Features)

เนื่องจากจำนวนข้อมูลข่าวปลอมและข่าวจริงในชุดข้อมูล Fake Corpus ไม่สมดุลกัน (Imbalance Data) ทำให้คุณสมบัติของข้อมูลส่วนใหญ่บดบังคุณสมบัติของข้อมูลส่วนน้อย และทำให้ประสิทธิภาพในการจำแนกข้อมูลส่วนน้อยนั้นไม่ดีเท่าที่ควร ดังนั้นเพื่อแก้ปัญหาดังกล่าว ผู้วิจัยจึงสุ่มเลือกใช้ข้อมูลประเภท Fake News ที่มีป้ายกำกับว่า fake และ Credible ที่มีป้ายกำกับว่า reliable จำนวนประเภทละ 32,000 รายการ จากนั้น เพื่อนำข้อมูลไปสร้างแบบจำลองผู้วิจัยจึงติดป้ายกำกับไว้ สำหรับชุดข้อมูลที่เป็น fake จะติดป้ายกำกับไว้ว่า 1 และชุดข้อมูลที่เป็น reliable จะติดป้ายกำกับไว้ว่า 0

คุณภาพของการจัดรูปแบบการจำแนกข้อมูลขึ้นอยู่กับคำในคลังข้อมูล (Corpus) และคุณลักษณะของคำเหล่านั้น ในงานวิจัยชิ้นนี้จะใช้ชุดไลบรารี่ของ python ที่ชื่อ Spacy และ Gensim เพื่อจัดเตรียมข้อมูลมาใช้ในการวิเคราะห์ โดยขั้นตอนการทำงานเป็นดังภาพที่ 1

ภาพที่ 1 ขั้นตอนการแปลงข้อความเป็นคุณลักษณะ

ขั้นตอนข้างต้นจะช่วยลดขนาดและเพิ่มบริบทของข้อความก่อนที่จะแปลงให้อยู่ในรูปของคุณลักษณะ โดยเฉพาะอย่างยิ่งการจัดคำให้อยู่ในรูปแบบปกติ (Lemmatization) ที่จะช่วยแปลงแต่ละคำที่อยู่ในรูปแบบต่าง ๆ ให้อยู่ในรูปแบบเดียว ส่วน n-Gram จะช่วยให้คำใกล้เคียงหรือมีบริบทเดียวกัน ให้เป็นคำเดียวกัน ซึ่งในการวิจัยครั้งนี้จะใช้ทั้งแบบ bi-Grams และ tri-Grams

ในการวิเคราะห์และสร้างโมเดล หลังจากที่ได้มีการจัดเตรียมข้อมูลไว้แล้ว จะต้องมีการแปลงข้อมูลให้มีลักษณะเป็นคุณลักษณะ โดยเทคนิคที่จะนำมาใช้คือ TF-IDF

TF-IDF (Term FrequencyInverse Document Frequency) [17] เป็นสถิติที่มีวัตถุประสงค์เพื่อคำนวณหาน้ำหนักจากความถี่ของการปรากฏคำในเอกสาร และพิจารณาความถี่ของคำนั้น ๆ ที่ปรากฏในเอกสารอื่นร่วมด้วย โดยมีแนวคิดที่ว่า คำที่ปรากฏในเอกสารน้อยฉบับจะมีค่าน้ำหนักสูง ส่วนคำที่ปรากฏในเอกสารหลายฉบับจะมีค่าน้ำหนักต่ำ เนื่องจากเป็นคำที่ไม่แสดงถึงคุณลักษณะเฉพาะของเอกสาร และแสดงว่าคำดังกล่าวไม่สามารถเป็นตัวแทนของเอกสารใด ๆ ได้ ซึ่งคำเหล่านั้นเรียกว่าคำหยุด (Stop Word) เช่น a and the เป็นต้น สามารถดูได้จากสมการดังนี้

$$W_{(i,\ j)} = tf_{(i,\ j)} \times \log(\frac{N}{df_{(i)}}) \quad (1)$$

3.3 การประเมินผลแบบทดลอง

ผู้วิจัยนำ Confusion Matrix มาใช้ในการประเมินผลลัพธ์ของการจัดกลุ่ม โดยที่หากข่าวปลอมใดที่ระบุไว้ว่าเป็นข่าวปลอมและสามารถจำแนกได้อย่างถูกต้องว่าเป็นข่าวปลอม ถือเป็น True Positive (TP) หากข่าวปลอมใดที่ระบุไว้ว่าเป็นข่าว

จริงแต่ไม่สามารถจำแนกได้ว่าเป็นข่าวปลอม ถือเป็น False Negative (FN) หากข่าวจริงใดที่ระบุไว้ว่าเป็นข่าวจริงและสามารถจำแนกได้ว่าเป็นข่าวจริง ถือเป็น True Negative (TN) และสุดท้าย หากข่าวจริงใดถูกระบุไว้ว่าเป็นข่าวปลอมแต่ไม่สามารถจำแนกได้ว่าข่าวจริง ถือเป็น False Positive (FP)

ในการสร้างแบบจำลองให้มีประสิทธิภาพ ผู้วิจัยได้พยายามลดจำนวนการพยากรณ์ที่ผิดพลาด ทั้ง False Negative และ False Positive โดยการนำวิธีการวัดประสิทธิภาพของแบบจำลอง F-measure (F-1) มาใช้ เพื่อให้เกิดความสมดุลระหว่างความแม่นยำ (Precision) กับการเรียกคืน (Recall) [18]

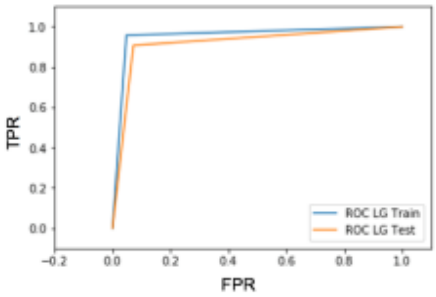
เพื่อให้เห็นภาพผลการทดลอง ผู้วิจัยได้นำ AUC (Area Under the ROC Curve) มาใช้เพื่อเป็นตัวบอกประสิทธิภาพในการทดสอบ (test performance) ว่าแบบจำลองสามารถระบุข่าวปลอมและข่าวจริงได้ดีมากน้อยเพียงใด และระบุจุดตัด (cut-off) ของการทดสอบที่มีความแม่นยำและน่าเชื่อถือมากที่สุดเพื่อนำค่าดังกล่าวมาใช้ในการจำแนกข่าวจริงและข่าวปลอมให้ถูกต้องมากที่สุดและผิดพลาดน้อยที่สุด

3.4 ผลการทดลอง

ในการหาค่าพารามิเตอร์ที่ดีที่สุดมาใช้กับการหาค่าความถี่ของคำ หรือ TF-IDF และเพื่อใช้กับแบบจำลองแบบต่าง ๆ ผู้วิจัยได้นำเทคนิค Cross Validation มาใช้ร่วมกับ Grid Search [19] ผลการวิจัยพบว่าแบบจำลองเทคนิคการถดถอยโลจิสติกสามารถจัดกลุ่มข่าวปลอมได้ดีที่สุดด้วยค่า AUC ร้อยละ 94 ตามมาด้วยเทคนิคซัพพอร์ทเวกเตอร์แมชชีนร้อยละ 91 เทคนิคนาอีฟเบส์ร้อยละ 89 และเทคนิคต้วจำแนกป่าแบบสุ่มร้อยละ 88 ตามลำดับ ดังตารางที่ 2

ตารางที่ 2 แสดงผลการทดสอบแต่ละแบบจำลอง

Model	Train (%)	Test (%)			
	ROC AUC	ROC AUC	Precision	Recall	F1 Score
Logistic Regress	0.95	0.94	0.85	0.91	0.88
Random Forest	0.99	0.88	0.90	0.82	0.86
SVM	0.93	0.91	0.83	0.90	0.87
Naïve Bayes	0.91	0.89	0.88	0.82	0.85



ภาพที่ 2 กราฟ ROC Curve จากผลการทดลองของการถดถอยโลจิสติก

4. บทสรุป

ปัจจุบันข้อมูลถูกสร้างขึ้นอย่างมากมายมหาศาลจากการใช้สื่อสังคมออนไลน์ เป็นการยากที่จะเชื่อถือได้ว่าข้อมูลเหล่านั้นมีความถูกต้องน่าเชื่อถือ ข่าวปลอมถือเป็นปรากฏการณ์ที่ท้าทายการบริโภคสื่อสังคมออนไลน์ซึ่งผู้ใช้ส่วนใหญ่มักจะถูกหลอกให้หลงเชื่อได้ง่าย เพื่อแก้ปัญหาดังกล่าว ผู้วิจัยได้เปรียบเทียบแบบจำลองการตรวจจับข่าวปลอมที่ปรากฏในสื่อสังคมออนไลน์จากพื้นฐานของข้อมูลที่มีคุณภาพ แสดงให้เห็นว่า แบบจำลองประเภทการถดถอยโลจิสติกให้ผลได้ดีที่สุด

แม้ว่าแบบจำลองประเภทการถดถอยโลจิสติกจะมีประสิทธิภาพในการจัดกลุ่มข่าวปลอมได้ดี แต่เมื่อพิจารณาตัวแปรอื่น ๆ ที่เกี่ยวข้องเช่น แหล่งที่มา ผู้แต่ง รูปภาพประกอบ เป็นต้น อาจทำให้ความน่าเชื่อถือและความเชื่อมั่นไม่ดีเท่าที่ควร ดังนั้น เพื่อปรับปรุงประสิทธิภาพของแบบจำลองให้ดียิ่งขึ้นทั้งในด้านของความถูกต้อง ความแม่นยำ ผู้วิจัยมีแนวคิดที่จะนำเสนอการจัดกลุ่มประเภทของข่าวปลอมด้วยเทคนิคการเรียนรู้เครื่องเชิงลึก (Deep Learning) และพิจารณารูปภาพที่ใช้ประกอบเนื้อหาซึ่งเป็นวิธีการที่สามารถทำควบคู่กันได้ นำมาวิเคราะห์เพิ่มเติมเพื่อประโยชน์ต่อการวิจัยในครั้งต่อไป

เอกสารอ้างอิง

[1] Martin Moore. "Inquiry into Fake News". Centre for the Study of Media, Communication and Power. King's College London .

[2] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election", Journal of Economic

Perspectives. Volume 31, Number 2-Spring 2017. Pages 211–236

[3] Emalio Ferrara. "Disinformation and Social Bot Operations in the run up to the 2017 French Presidential Election". University of Southern California., Information Sciences Institute.

[4] พิจิตรา สึคาโมโต และ นที ธรรมพัฒนพงศ์, "ข่าวลือทำมกลางวิกฤตการณ์ทางการเมืองในสื่อใหม่: กรณีศึกษาทวิตเตอร์" https://www.scribd.com/document/334294719/Rumor-Analysis-Cutweet-Final-Sukosol-Final สืบค้นเมื่อวันที่ 16 สิงหาคม 2560

[5] Beating the hell out of fake news. Ethical Record: The Proceedings of the Conway Hall Ethical Society 122 (6).

[6] Sander van der Linden (November 14, 2017). "How to Spot Fake News". Psychology Today. Retrieved November 16, 2017.

[7] Wardle, Claire . "Fake news. It's complicated". https://firstdraftnews.org/fake-news-complicated/. สืบค้นเมื่อวันที่ 12 กุมภาพันธ์ 2561

[8] Piyatida Inrak and Sukree Sinthupinyo, "Applying latent semantic analysis to classify emotions in Thai text", The 2nd International Conference on Computer Engineering and Technology (ICCET 2010), Chengdu, 2010, pp. V6-450-V6-454. doi: 10.1109/ICCET.2010.5486137

[9] นิเวศ จิระวิชิตชัย และ นรินทร์ พนาวาส, "การจำแนกความคิดเห็นโดยใช้เทคนิคการเรียนรู้ของเครื่อง", 2011 Eighth International.

[10] อุทธ ไกยวรรณ์. "หลักการและการใช้การวิเคราะห์การถดถอยโลจิสติคสำหรับการวิจัย", วารสารวิจัยมหาวิทยาลัยเทคโนโลยีราชมงคลศรีวิชัย 4(1): 1-12 (2555).

[11] วาทินี บุ้ยเพียร และคณะ, 2553, "การเปรียบเทียบประสิทธิภาพและวิเคราะห์การจำแนกข้อมูลด้วยวิธีการทางเครือข่ายประสาทเทียม", ประชุมวิชาการระดับชาติด้านคอมพิวเตอร์และเทคโนโลยีสารสนเทศ 5: 131-138.

[12] P. Sanguansat, Machine Learning. Thailand: Panyapiwat Instutute of Management, 2015

[13] Victoria L. Rubin, Niall J. Conroy, Yimin Chen, and Sarah Cornwell. "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

[14] กานต์ เจริญจิตร. "การเปรียบเทียบแบบจำลองการจำแนกกลุ่มโรงงานอุตสาหกรรมโดยใช้เทคนิคการทำเหมืองข้อความกรณีศึกษาโรงงานอุตสาหกรรมจังหวัดปทุมธานี". Joint Conference on ACTIS & NCOBA 2015, Jan 30-31, Nakorn Phanom, Thailand. ISSN: 1606-9006.

[15] สุพัตรา วิริยะวิสุทธิสกุล และคณะ. "ระบบแจ้งเตือนไซเบียมีเดียสำหรับธุรกิจด้วยซัพพอร์ทเวกเตอร์แมชชีน". คณะวิศวกรรมศาสตร์และเทคโนโลยี สถาบันการจัดการปัญญาภิวัฒน์. Panyapiwat Journal Vol.8 Special Issue August 2016 หน้า 223

[16] Fake News Corpus : https://github.com/several27/FakeNewsCorpus สืบค้นเมื่อวันที่ 12 กุมภาพันธ์ 2561

[17] ณิชาพร สุระ, "การจำแนกหมวดหมู่เอกสารภาษาไทยอัตโนมัติโดยใช้ขั้นตอนวิธี FPTC" วิทยานิพนธ์สาขาวิทยาการคอมพิวเตอร์ บัณฑิตวิทยาลัยสถาบันเทคโนโลยีพระจอมเกล้าพระนครเหนือ, 2549.

[18] ดร.เอกสิทธิ์ พัชรวงศ์ศักดา. "An Introduction to Data Mining Techniques: Thai Version", พิมพ์ครั้งที่ 1 ปี 2557

[19] C.-L. Huang, M.-C. Chen, and C.-J. Wang, "Credit scoring with a data mining approach based on support vector machines", Expert Systems with Applications, vol.33, pp. 847-856, 2007

ประวัติผู้เขียน

ชื่อ – นามสกุล วิสิทธิ์ วาณิชยานนท์

ประวัติการศึกษา

พ.ศ. 2548 สาขาสังคมวิทยาและมานุษยวิทยา
คณะมานุษยศาสตร์ มหาวิทยาลัยรามคำแหง

ประสบการณ์ทำงาน

พ.ศ. 2566 นักวิเคราะห์ข้อมูล (อิสระ)