# การตรวจสอบข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง

น.ส.สุปัญญา อภิวงศ์โสภณ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2561

# DETECTING FAKE NEWS WITH MACHINE LEARNING METHOD

Miss Supanya Aphiwongsophon

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy (Computer Engineering) in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2018

| | |
|---|---|
| หัวข้อวิทยานิพนธ์ | การตรวจสอบข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง |
| โดย | น.ส.สุปัญญา อภิวงศ์โสภณ |
| สาขาวิชา | วิศวกรรมคอมพิวเตอร์ |
| อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก | ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา |

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรดุษฎีบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(รองศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา)

..... กรรมการ
(รองศาสตราจารย์ ดร.ทวิตีย์ เสนีวงศ์ ณ อยุธยา)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.ชวลิต ศรีสถาพรพัฒน์)

สุปัญญา อภิวงศ์โสภณ : การตรวจสอบข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง. ( DETECTING FAKE NEWS WITH MACHINE LEARNING METHOD) อ.ที่ปรึกษาหลัก : ศ. ดร.ประภาส จงสถิตย์วัฒนา

วิทยานิพนธ์นี้นำเสนอวิธีการตรวจจับข่าวปลอมบนเครือข่ายสังคมออนไลน์ทวิตเตอร์ด้วยวิธีการเรียนรู้ด้วยเครื่อง โดยใช้การเรียนรู้ด้วยเครื่องสามวิธี ได้แก่ Naïve Bayes, Neural Network และ Support Vector Machine โดยเก็บข้อมูลจากหัวข้อข่าวที่เป็นภาษาไทย ในระหว่างเดือนตุลาคมถึงพฤศจิกายน พ.ศ. 2560 ผลการวิจัยพบว่าทั้งสามวิธีสามารถตรวจจับข่าวปลอมในชุดข้อมูลได้อย่างถูกต้อง ร้อยละความถูกต้องของวิธี Naïve Bayes คือ 96.08 เปอร์เซ็นต์ Neural Network 99.89 เปอร์เซ็นต์ และ Support Vector Machine 99.89 เปอร์เซ็นต์ นอกจากนี้ได้ทำการวิเคราะห์ข้อมูลข่าวปลอมและชี้ให้เห็นลักษณะของข่าวปลอมที่พบในชุดข้อมูล

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2561 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

# # 5771425821 : MAJOR COMPUTER ENGINEERING

KEYWORD: FAKE NEWS, MACHINE LEARNING, ONLINE SOCIAL NETWORK, TWITTER

Supanya Aphiwongsophon : DETECTING FAKE NEWS WITH MACHINE LEARNING METHOD. Advisor: Prof. Dr. Prabhas Chongstitvatana, Ph.D.

This dissertation proposes a machine learning method which can identify fake news from Twitter data. The experiment is carried out with three widely used machine learning methods: Naïve Bayes, Neural Network and Support Vector Machine using Thai's topic and collected from October to November 2017. The results show that all three methods can detect fake news in this data set accurately. The accuracy of Naïve Bayes method is 96.08 percent, Neural Network 99.89 percent and Support Vector Machine 99.89 percent. Furthermore, we analyze the data of fake news and point out some of its characteristics.

Field of Study: Computer Engineering Student's Signature .....

Academic Year: 2018 Advisor's Signature .....

# กิตติกรรมประกาศ

ขอขอบคุณศาสตราจารย์ ดร.ประภาส จงสถิตย์วัฒนา อาจารย์ที่ปรึกษาที่ให้คำแนะนำทั้งในด้านวิชาการ ให้คำชี้แนะในการแก้ไขปัญหาต่าง ๆ ให้กำลังใจในยามท้อแท้ จนทำให้วิทยานิพนธ์นี้สามารถลุล่วงไปได้ด้วยความเรียบร้อย

ขอขอบคุณคณะกรรมการสอบวิทยานิพนธ์ที่ให้ความอนุเคราะห์เป็นกรรมการสอบและให้คำแนะนำที่ทำให้วิทยานิพนธ์นี้เรียบร้อยมากยิ่งขึ้น ซึ่งคณะกรรมการสอบวิทยานิพนธ์นี้ประกอบด้วย รองศาสตราจารย์ ดร.วิวัฒน์ วัฒนาวุฒิ รองศาสตราจารย์ ดร. ทวิตีย์ เสนีวงศ์ ณ อยุธยา ผู้ช่วยศาสตราจารย์ ดร. สุกรี สินธุภิญโญ และ ผู้ช่วยศาสตราจารย์ ดร. ชวลิต ศรีสถาพรพัฒน์

ขอขอบคุณคณาจารย์ทุกท่านในภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ที่ได้ประสิทธิ์ประสาทความรู้ และให้ประสบการณ์ที่ดีตลอดระยะเวลาที่ศึกษาในหลักสูตร

ขอขอบคุณบัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย ที่สนับสนุนทุนการศึกษาหลักสูตรดุษฎีบัณฑิต "100 ปี จุฬาลงกรณ์มหาวิทยาลัย" สำหรับการดำเนินงานวิทยานิพนธ์ฉบับนี้

ขอบคุณทุกคนในครอบครัวที่รักและห่วงใย คอยให้กำลังใจกันมา ตลอดจนพี่น้องในห้องปฏิบัติการ ISL (Intelligent Systems Laboratory) รวมถึงเพื่อน ๆ ในภาควิชาวิศวกรรมคอมพิวเตอร์ทุกท่านที่เป็นกำลังใจให้กันเสมอมา

สุปัญญา อภิวงศ์โสภณ

# สารบัญ

**หน้า**

# สารบัญภาพ

# สารบัญตาราง

# บทที่ 1
# ที่มาและความสำคัญของปัญหา

ข่าวปลอม (Fake news) เป็นปัญหาสำคัญในสังคม การรับรู้ข่าวปลอมของประชาชน ส่งผลกระทบทำให้เกิดความเสียหายจากมูลเหตุอันเป็นเท็จได้ จึงมีความพยายามในการควบคุมดูแลข่าวปลอม ไม่ให้มีการแพร่กระจายไปในวงกว้างทั้งด้านกฎหมายควบคุมการเผยแพร่ข้อมูลอันเป็นเท็จ [1] ตัวอย่างการตรวจสอบข้อเท็จจริงที่มีการเผยแพร่กันในวงกว้าง อาทิเช่น รายการชัวร์ก่อนแชร์ ซึ่งเป็นการตรวจสอบข้อเท็จจริงของข่าวปลอมโดยส่งผู้สื่อข่าวออกไปสัมภาษณ์ผู้ที่มีความรู้ความเชี่ยวชาญเกี่ยวกับประเด็นที่กล่าวถึงในข่าวปลอม แล้วจึงนำมาเผยแพร่เปิดเผยข้อเท็จจริง นอกจากนั้นยังมีเว็บไซต์ที่พยายามตรวจสอบข่าวปลอมที่เกิดขึ้นในประเทศไทย

หากจะพิจารณาถึงข่าวปลอมควรเริ่มต้นจากการส่งข่าว ในอดีตเริ่มจากลักษณะของการบอกต่อกันไปเรื่อย ๆ กระจายข่าวออกไปแบบเดียวกับการกระจายข่าวลือ เนื่องจากการบอกต่อแบบคนต่อคน อาจเกิดความคลาดเคลื่อนได้จากการที่คนสื่อสารพูดผิด คนฟังเกิดความเข้าใจผิดไป การตีความผิดความหมาย หรือการขยายความโดยเพิ่มเติมความคิดเห็นส่วนตัวที่อาจเกิดจากความอคติก่อนการส่งข่าวต่อไป ทำให้ประสิทธิภาพของการส่งข่าวทำได้ไม่ดีเท่าที่ควร เนื้อหาของข่าวที่อาจเกิดเปลี่ยนแปลงไป เมื่อมีการส่งข่าวต่อกันไปในระยะเวลาหนึ่ง ต่อมาราวปี ค.ศ. 1439 เมื่อเริ่มมีการใช้กระดาษ จึงเกิดการบันทึกเนื้อหาข่าวต่าง ๆ ลงบนกระดาษเป็นสื่อสิ่งพิมพ์ก่อนส่งต่อไปยังผู้รับ [2] จนกระทั่งวิวัฒนาการต่อมาในปัจจุบันทำให้เกิดการสื่อสารข้อมูลต่าง ๆ ที่สามารถทำได้อย่างรวดเร็วผ่านระบบเครือข่ายอินเทอร์เน็ต การส่งเนื้อหาข่าวออกไป สามารถส่งได้ทั้งข้อความ ภาพหรือเสียงได้ในเวลาเดียวกัน อีกทั้งยังสามารถกระจายข่าวสารไปถึงผู้รับปริมาณได้ครั้งละมาก ๆ ในคราวเดียวกัน เกิดการเปลี่ยนแปลงข้อมูลให้อยู่ในรูปอิเล็กทรอนิกส์ทำให้สามารถส่งต่อได้อย่างง่ายดาย ทำให้เนื้อหาข่าวสารสามารถแพร่กระจายออกไปได้อย่างรวดเร็ว โดยที่ผู้ส่งไม่อาจทราบเลยว่า ข่าวสารที่ตนเองได้ส่งออกไปถูกรับไปโดยผู้ใด ในมุมมองอีกด้านหนึ่งทางฝั่งผู้รับข่าวสารเอง เมื่อได้รับข่าวสารใด ๆ จะทราบได้อย่างไรว่า เนื้อหาข้อมูลข่าวสารที่ได้รับมา เป็นความจริง หรือเป็นข้อมูลที่ผ่านการเปลี่ยนแปลงแก้ไขที่อาจถูกบิดเบือนข้อเท็จจริงไปก่อนหน้าหรือไม่

ระบบเครือข่ายอินเทอร์เน็ตที่มีผู้ใช้งานจำนวนมาก ประกอบด้วยส่วนของเครือข่ายสังคมออนไลน์ที่ปัจจุบันกำลังได้รับความนิยมใช้งานกันมาก และยังเป็นแหล่งกำเนิดข้อมูลข่าวสารต่าง ๆ ที่ผู้คนส่วนใหญ่ใช้เสพข่าวสารเป็นเรื่องปกติในชีวิตประจำวัน โดยข่าวสารที่ปรากฏตามสื่อต่าง ๆ จะพบว่ามีทั้งส่วนที่เป็นข้อเท็จจริง และส่วนที่เป็นข่าวปลอมที่ต้องการหลอกลวงให้ผู้คนหลงเชื่อ เกิดความเข้าใจผิดพลาด คลาดเคลื่อนไปจากข้อเท็จจริง อันเกิดจากความเข้าใจผิดของผู้ส่งเอง หรือ

เจตนาของผู้ส่งข่าวที่ต้องการบิดเบือนข้อเท็จจริงของข่าวนั้น อาจเป็นความพยายามสร้างความคลุมเครือระหว่างเรื่องราวเนื้อหาที่เป็นความจริงกับข้อมูลอันเป็นเท็จ เพื่อทำลายความเชื่อถือในเรื่องราวบางอย่าง [3]

การแพร่กระจายข่าวในอดีต เริ่มต้นด้วยคำพูดของคนโดยส่งต่อจากปากต่อปาก ซึ่งการแพร่กระจายข้อความผ่านทางคำพูดมีโอกาสที่จะสื่อสารกันได้ที่ไม่ถูกต้องในหลายด้าน อาทิเช่น ผู้รับตีความไม่ถูกต้อง เกิดความเข้าใจผิดจากการสื่อสารที่ผิดพลาด การมีอคติส่วนบุคคล อาจรวมถึงความคิดเห็นส่วนตัวก่อนที่จะมีการป่าวประกาศออกไป [4] ส่งผลให้ประสิทธิภาพของการสื่อสารไม่ดี โดยเนื้อหาของข่าวอาจผิดเพี้ยนไปเมื่อมีการส่งข้อความต่อ ๆ กันไปในช่วงระยะเวลาหนึ่ง [5] [6]

ข้อมูลข่าวที่อยู่ในรูปแบบอิเล็กทรอนิกส์สามารถส่งต่อได้ง่ายและรวดเร็ว เนื้อหาสามารถแพร่กระจายไปได้อย่างกว้างขวางรวดเร็ว เครือข่ายสังคมออนไลน์กลายเป็นแหล่งข่าวสำคัญ มีข้อมูลปริมาณมากมายมหาศาลสามารถส่งต่อเผยแพร่ไปบนสื่อสังคมออนไลน์ ผู้อ่านจึงได้รับรู้เนื้อหาข่าวจากหลายแหล่งในคราวเดียวกัน [7] [8] [9] แต่อย่างไรก็ตาม ข้อมูลที่ไม่ถูกต้องยังคงแพร่กระจายผ่านไปยังสื่อสังคมออนไลน์

เครือข่ายสังคมออนไลน์ทวิตเตอร์ (Twitter) เริ่มต้นใช้งานครั้งแรกตั้งแต่เดือนมีนาคม ค.ศ. 2006 ในช่วงแรกที่เปิดให้มีการใช้งาน ผู้ใช้จะสามารถส่งเฉพาะข้อความที่เป็นตัวอักษรไม่เกิน 140 ตัวอักษร ต่อมาได้ปรับเพิ่มความยาวในการส่งข้อความเป็นไม่เกิน 280 ตัวอักษร ลักษณะคล้ายกับการส่งข้อความผ่านโทรศัพท์เคลื่อนที่หรือโปรแกรมสนทนาออนไลน์ [10] ทำให้ผู้ใช้งานคุ้นชินกับการใช้งานทวิตเตอร์เพื่อส่งข่าวสั้น ๆ กระชับได้ใจความ จากรายงาน “Digital in 2019” ของ We Are Social และ Hootsuite [11] ที่มีความเกี่ยวข้องกับสถานการณ์การใช้งานดิจิทัล และอินเทอร์เน็ตประจำปีค.ศ. 2019 โดยรวบรวมข้อมูลสถิติจากทั่วโลกรวมถึงประเทศไทย พบว่าปัจจุบัน โลกมีประชากร 7,876 ล้านคน แบ่งเป็นผู้ชาย 50.5% และผู้หญิง 49.5% มีจำนวนผู้ใช้งานอินเทอร์เน็ตทั่วโลกราว 4,388 ล้านคน คิดเป็นจำนวนมากกว่า 50% ของประชากรโลก โดยมีจำนวนผู้ใช้โทรศัพท์เคลื่อนที่ทั่วโลก 5,112 ล้านคน มีจำนวนผู้ใช้เครือข่ายสังคมออนไลน์ (Online Social Network) 3,484 ล้านคน และพบว่าจำนวนผู้ใช้งานเครือข่ายสังคมออนไลน์มากถึง 3,256 ล้านคน ใช้งานผ่านโทรศัพท์เคลื่อนที่ ในเดือนเมษายน ค.ศ. 2019 พบว่ามีผู้เปิดบัญชีใช้งานทวิตเตอร์ทั่วโลกมากกว่า 330 ล้านบัญชี และรายงานในส่วนของภูมิภาคเอเชียตะวันออกเฉียงใต้ ในประเทศไทยมีการใช้งานทวิตเตอร์ในแต่ละเดือนประมาณ 89.4 ล้านครั้ง จากบัญชีผู้ใช้งานทวิตเตอร์ประมาณ 3.9 ล้านคน จากข้อมูลที่กล่าวมานี้ พบว่าในประเทศไทยเองมีการใช้งานทวิตเตอร์เพื่อส่งข่าวสารกันอย่างมากมาย จึงเกิดคำถามว่าจะทราบได้อย่างไรว่าข่าวที่กำลังอ่านอยู่เป็นเรื่องจริงหรือเป็นข่าวปลอม

จากรายงาน [11] ยังพบว่าพฤติกรรมการใช้เครือข่ายสังคมออนไลน์ทั่วโลกในการใช้งานเครือข่ายสังคมออนไลน์มีจำนวนเพิ่มมากขึ้นในประเทศกำลังพัฒนา จำนวนผู้ใช้งาน เครือข่ายสังคม

ออนไลน์ ปี ค.ศ. 2019 อยู่ที่ 3,484 ล้านคน เพิ่มมากขึ้นถึง 288 ล้านคนจากปีก่อนหน้า โดยประชากรส่วนใหญ่ที่ใช้เครือข่ายสังคมออนไลน์เป็นกลุ่มอายุ 18 – 24 ปี และ 25 – 34 ปี แต่ละคนใช้เวลาไปกับเครือข่ายสังคมออนไลน์ โดยเฉลี่ย 2 ชั่วโมง 16 นาทีต่อวัน และโดยเฉลี่ยแต่ละคนจะมีบัญชีเครือข่ายสังคมออนไลน์ 8.9 บัญชี

ปัจจุบันปี ค.ศ. 2019 ประเทศไทยมีประชากร 69.24 ล้านคน แบ่งเป็นผู้หญิง 51.3% และผู้ชาย 48.7% โดย 50% ของจำนวนประชากรทั้งหมด อาศัยอยู่ในเขตเมือง คนไทยประมาณ 57 ล้านคนสามารถเข้าถึงอินเทอร์เน็ต และนิยมใช้เครือข่ายสังคมออนไลน์ โดยคนไทย 51 ล้านคน มีการใช้งานเครือข่ายสังคมออนไลน์เป็นประจำ และมีการใช้งานอินเทอร์เน็ตเฉลี่ย 9 ชั่วโมง 11 นาทีต่อวัน โดยใช้เวลาอยู่กับเครือข่ายสังคมออนไลน์ 3 ชั่วโมง 11 นาทีต่อวัน แต่ละคนมีบัญชีเครือข่ายสังคมออนไลน์เฉลี่ย 10.5 บัญชี ประเทศไทยยังจัดอยู่ในกลุ่ม 5 อันดับสูงสุดของโลกที่มีบัญชีเครือข่ายสังคมออนไลน์เฉลี่ยสูงสุดเป็นรองจากอินเดีย (12.0 บัญชีต่อคน) อินโดนีเซีย (11.2 บัญชีต่อคน) เวียดนาม (10.8 บัญชีต่อคน) เท่ากับโคลัมเบีย (10.5 บัญชีต่อคน) และมากกว่าฟิลิปปินส์ (10.4 บัญชีต่อคน) [11]

การที่แต่ละคนมีบัญชีเครือข่ายสังคมออนไลน์จำนวนมาก แสดงให้เห็นถึงความหลากหลายของเครือข่ายสังคมออนไลน์จำนวนมากที่มีการใช้งานแล้ว ยังอาจทำให้เกิดข่าวปลอมขึ้นมาได้ เนื่องจากการใช้งานที่เกิดขึ้นในเครือข่ายสังคมออนไลน์ทำให้คนได้รับข่าวจากหลายช่องทาง อันจะก่อให้เกิดความสับสนเกิดความเข้าใจผิดพลาดส่งต่อข้อมูลโดยมิได้ระมัดระวังจนกลายเป็นภัยคุกคามทางสังคม ในสังคมปัจจุบันหลายประเทศพยายามสร้างกลไกป้องกันข่าวปลอมที่แตกต่างกัน ทั้งกลไกทางกฎหมายและมาตรการต่าง ๆ จากบริษัทผู้ให้บริการ กลุ่มคนอายุสูงวัยเป็นกลุ่มที่มีความอ่อนไหวต่อการส่งสารข้อมูลอันเป็นเท็จ เนื่องด้วยความไม่รู้เท่าทันสื่อ หรือเป็นความปรารถนาดีที่ต้องการส่งต่อไปยังผู้รับสาร โดยมิได้ตระหนักว่าการส่งต่อออกไปเป็นการช่วยแพร่กระจายข่าวปลอมให้เผยแพร่ออกไปรวดเร็วมากขึ้น ในประเทศไทยเองมีความพยายามควบคุมการเผยแพร่ข้อมูลบนเครือข่ายคอมพิวเตอร์ด้วยการออกมาตรการทางกฎหมาย เช่น การออกพระราชบัญญัติว่าด้วยการกระทำความผิดเกี่ยวกับคอมพิวเตอร์ (ฉบับที่ ๒) พ.ศ. ๒๕๖๐ [1]

การเผยแพร่ข่าวปลอมที่หวังผลทางการเมือง หรือต้องการสร้างกระแสขึ้นในสังคมในช่วงเวลาหนึ่ง ข่าวสารต่าง ๆ ที่ผู้ใช้ได้รับมีส่วนสำคัญสำหรับการตัดสินใจคือประเด็นความถูกต้องของเนื้อหาข้อเท็จจริงที่เกิดขึ้นจริง ๆ ข่าวปลอมในด้านลบก่อให้เกิดผลกระทบกับผู้ที่เกี่ยวข้องกับประเด็นเนื้อหา ตัวอย่างเช่น การปล่อยข่าวปลอมเรื่องพระอาการประชวรทรุดหนัก และการเสด็จสวรรคตของรัชกาลที่ 9 ในช่วงเดือนตุลาคม พ.ศ. 2559 ส่งผลให้นักลงทุนในตลาดหลักทรัพย์ในประเทศไทยเกิดความหวั่นวิตก ขายหุ้นออกมาจำนวนมากอย่างต่อเนื่อง ทำให้ดัชนีตลาดหลักทรัพย์ลดลงอย่างหนักโดย

ลดลงมากที่สุดถึง - 99.08 จุด ในมูลค่าการซื้อขายมากกว่า 91 ล้านบาท ภายในช่วงเวลาบ่ายของวันที่ 12 ตุลาคม พ.ศ. 2559 [12]

ประเทศไทยตั้งอยู่เหนือเส้นศูนย์สูตร อยู่ในบริเวณที่มีโอกาสเกิดมรสุมเขตร้อนได้บ่อย ดังนั้น ข่าวที่เกี่ยวข้องกับภัยพิบัติทางธรรมชาติ เช่น พายุเข้า ฝนตก น้ำท่วม เขื่อนแตกจึงเป็นข่าวที่ส่งผลกระทบต่อชีวิตความเป็นอยู่ของประชาชน การรับรู้ข่าวสารที่ถูกต้อง รวดเร็ว จะทำให้ประชาชนสามารถเตรียมพร้อม เพื่อรับสถานการณ์ภัยธรรมชาติที่จะเกิดขึ้นได้ทันเวลา ซึ่งจะช่วยลดความสูญเสียต่าง ๆ ที่อาจเกิดขึ้นได้ แต่หากประชาชนได้รับข่าวปลอมหรือข่าวที่ผิดพลาด จะเกิดความเสียหาย อาทิเช่น สถานการณ์น้ำท่วมฉับพลันที่เกิดจากการน้ำป่าที่ไหลลงมา หากประชาชนไม่ได้เตรียมความพร้อมรับสถานการณ์ล่วงหน้า ข้าวของ ทรัพย์สินต่าง ๆ ในบ้านจะเกิดความเสียหายได้

ทวิตเตอร์ (Twitter) เป็นหนึ่งเครือข่ายสังคมออนไลน์ที่กำลังได้รับความนิยมมากขึ้นเรื่อย ๆ มีผู้ใช้จำนวนมากใช้ข้อมูลที่เผยแพร่ผ่านแพลตฟอร์มเครือข่ายสังคมออนไลน์ทวิตเตอร์เป็นประจำทุกวัน การส่งข้อมูลผ่านทวิตเตอร์ยังเป็นการส่งข้อมูลที่มีประสิทธิภาพ ถึงแม้ว่าผู้ใช้จะได้รับอนุญาตให้ส่งข้อมูลในจำนวนอักขระที่จำกัด แต่ผู้ใช้มีอิสระในการเผยแพร่เนื้อหาเป็นอะไรก็ได้ ตามแต่ความต้องการ ในช่วงหลายปีมานี้ผู้ใช้งานทวิตเตอร์ที่มีความประสงค์ร้าย สามารถแพร่กระจายข้อมูลข่าวสารที่เป็นอันตราย ซึ่งส่งผลกระทบต่อความสงบสุขของสังคม หรือความปลอดภัย ดังนั้นการแพร่กระจายข่าวลือหรือข่าวปลอมต่าง ๆ ที่เกิดขึ้นในเครือข่ายสังคมออนไลน์ไม่เพียงแต่ส่งผลกระทบในทางลบ แต่ยังรวมถึงชีวิตของผู้คนที่ได้รับผลกระทบจากข่าวปลอมอีกด้วย [13]

ตัวอย่างการได้รับข้อมูลที่ผิดพลาดที่เกิดขึ้นผ่านสื่อสังคมออนไลน์อื่น ๆ เช่น WhatsApp เป็นตัวอย่างที่ชาวอินเดียห้าคนถูกทุบตีจนตาย เนื่องจากผู้ที่เข้ามาทำร้ายเชื่อว่าผู้ที่ตกเป็นเหยื่อทั้งห้าคนมีความเกี่ยวข้องกับการลักพาตัวเด็กตามข้อมูลที่การเผยแพร่กันผ่าน WhatsApp ซึ่งในประเทศอินเดียมีผู้ใช้งาน WhatsApp ประมาณ 200 ล้านคน การส่งผ่านข้อมูลใด ๆ ผ่านสื่อสังคมออนไลน์ที่คนส่วนใหญ่เชื่อถือย่อมส่งผลกระทบได้อย่างรุนแรง ดังผลของการหลอกลวงด้วยข้อความลูกโซ่หรือข้อความที่บิดเบือนความจริงที่แชร์ใน WhatsApp ที่ทำให้มีคนเสียชีวิตจริง ๆ จากการส่งต่อข้อมูลข่าวปลอม [14]

อีกตัวอย่างหนึ่งของผลกระทบที่เกิดขึ้นจากข่าวปลอมคือเรื่องสลดใจที่เกิดโดยไม่คาดฝัน เป็นเรื่องราวของสาเหตุที่ทำให้เกิดการฆ่าตัวตายของผู้อำนวยการสำนักงานเศรษฐกิจและวัฒนธรรมไทเปในนครโอซาก้า ประเทศญี่ปุ่น โดยเกิดการฆ่าตัวตายโดยการแขวนคอในบ้านตัวเอง เนื่องจากข่าวลือที่เกิดขึ้นระหว่างพายุไต้ฝุ่น Sheppard ในบริเวณเขตคันไซของญี่ปุ่น ทำให้ผู้คนเข้าใจผิดว่ามีการเลือกปฏิบัติ ในการเข้าให้ความช่วยเหลือระหว่างนักท่องเที่ยวชาวจีนที่เดินทางมาจากแผ่นดินใหญ่ กับนักท่องเที่ยวชาวไต้หวัน [15]

สองในสามของชาวอเมริกันที่ใช้งานอินเทอร์เน็ต มีการใช้งานเครือข่ายสังคมออนไลน์เพื่อรับข้อมูลข่าวสารต่าง ๆ [16] การแพร่กระจายข่าวออนไลน์ส่งผลกระทบต่อสังคมโดยรวม ตัวอย่างเช่น ในช่วงสามเดือนก่อนการเลือกตั้งประธานาธิบดีสหรัฐอเมริกาในช่วงปีพ.ศ. 2559 (ค.ศ. 2016) มีข่าวปลอมเกิดขึ้นมากมายในสื่อสังคมออนไลน์ โดยเฉพาะอย่างยิ่งมีข่าวปลอมที่เกิดขึ้นมีการแพร่กระจายไปมากกว่า 37 ล้านครั้งบนเครือข่ายสังคมออนไลน์เฟซบุ๊ก (Facebook) โดยที่จำนวนข่าวปลอมที่มีการแพร่กระจายกันมากกว่าครึ่งหนึ่งมาจากแหล่งข้อมูลที่มีความน่าเชื่อถือ [17]

ผู้อ่านจะทราบได้อย่างไรว่าข้อมูลข่าวที่กำลังอ่านเป็นเรื่องจริงหรือเท็จ เมื่อในแต่ละวันผู้อ่านได้รับข้อมูลข่าวสารปริมาณมากมาย การพิจารณาตัดสินใจเชื่อเนื้อหาของข้อมูลที่ได้รับว่าเป็นเรื่องจริงหรือเป็นเรื่องของการเปลี่ยนแปลงบิดเบือนข้อเท็จจริง เป็นเรื่องยากที่ผู้อ่านไม่อาจทราบได้ว่าเนื้อหาของข้อความหรือข่าวนั้นเป็นจริงหรือเท็จ [18]

จากงานวิจัย [19] พบว่าการกระจายข่าวปลอมเกิดขึ้นอย่างรวดเร็ว ข่าวปลอมสามารถเข้าถึงผู้อ่านได้รวดเร็วมากกว่าข่าวจริงถึง 6 เท่า ในขณะที่ปัจจุบันมีการเผยแพร่ข่าวปลอมมากกว่าข่าวจริงประมาณ 70 เปอร์เซ็นต์ นอกจากนั้นยังพบว่าการกระจายข่าวส่วนใหญ่เกิดจากคนจริง ๆ ไม่ได้เกิดจากการกระทำของหุ่นยนต์ (bot) โดยงานวิจัยนี้พบว่าเรื่องราวข่าวที่ได้รับความนิยมในการเผยแพร่มากที่สุดเป็นเรื่องที่มีความเกี่ยวข้องกับการเมือง ตามด้วยเรื่องราวความลี้ลับที่เกี่ยวข้องกับตำนานข่าวสารทางด้านเศรษฐกิจ ข่าวอาชญากรรมและการก่อการร้าย ข่าววิทยาศาสตร์และเทคโนโลยี ข่าวบันเทิง และข่าวภัยธรรมชาติ

งานวิจัยส่วนใหญ่ที่เกี่ยวข้องกับข่าวปลอม เป็นงานวิจัยในภาษาต่างประเทศ อาทิเช่น ภาษาอังกฤษ [20] [21] [22] ภาษาอาหรับ [13] [23] [24] [25] แต่มีงานวิจัยที่เกี่ยวข้องกับข่าวปลอมในภาษาไทยจำนวนไม่มากนัก โดยงานวิจัยที่มีการวิเคราะห์ข่าวเพื่อจำแนกว่าเป็นข่าวปลอมหรือไม่ด้วยการวิเคราะห์เนื้อความในข่าว [26] [27] การวิเคราะห์อารมณ์ความรู้สึกที่ได้จากการอ่านเนื้อความข่าว [28] [29] ซึ่งงานวิจัยที่มุ่งเน้นกับการวิเคราะห์เนื้อหาของข่าว เมื่อมีการเปลี่ยนแปลงเนื้อข่าวเป็นภาษาอื่น ๆ บริบททางด้านภาษา ไวยากรณ์หรือความหมายของคำในแต่ละภาษาจะเปลี่ยนแปลง ทำให้ไม่สามารถใช้ผลการวิเคราะห์ที่ได้กับเนื้อหาข่าวในภาษาที่เปลี่ยนแปลงไป นอกจากนี้งานวิจัยที่เกี่ยวข้องกับการวิเคราะห์องค์ประกอบของข่าวปลอม [2] การวิเคราะห์ลักษณะการกระจายของข่าวปลอมสู่ผู้รับสารด้วยแบบรูปต่าง ๆ [6] ยังคงเป็นเรื่องที่มีความน่าสนใจสำหรับอีกหลายงานวิจัย

การประเมินข่าวปลอมไม่เพียงแต่เกี่ยวข้องกับความน่าเชื่อถือของเนื้อหาข่าวเท่านั้น แต่ยังรวมถึงความน่าเชื่อถือของแหล่งสื่อสังคมออนไลน์อีกด้วย จากปัญหาที่กล่าวมาข้างต้น ในงานวิจัยนี้จึงเสนอวิธีการจำแนกข่าวปลอมที่เป็นภาษาไทยที่ได้จากเครือข่ายสังคมออนไลน์ทวิตเตอร์ โดยการค้นหาคุณลักษณะที่สามารถใช้จำแนกข่าวปลอมออกจากข่าวจริง โดยใช้กระบวนการวิธีการเรียนรู้

ด้วยเครื่อง (Machine learning) เพื่อจำแนกข่าวปลอมจากข้อความที่มีการส่งกันในทวิตเตอร์ โดยงานวิจัยนี้มีสมมติฐานเบื้องต้น ได้แก่

1) แหล่งข่าวที่เผยแพร่ข่าวครั้งแรก มีผลต่อความน่าเชื่อถือของผู้รับข่าวสาร

2) จำนวนการส่งต่อข่าวสาร (Retweet) ก่อนหน้า มีผลต่อการส่งต่ออีกครั้งของผู้อ่านข่าว

3) การส่งต่อของผู้อ่านข่าวโดยการระบุบุคคลที่ต้องการให้รับรู้เฉพาะ (Mention) แสดงถึงความมั่นใจในความน่าเชื่อถือของข่าวนั้น

4) ข่าวที่มีแหล่งอ้างอิง เช่น รูปภาพประกอบ คลิปวีดิทัศน์ หรือแหล่งเชื่อมโยงอื่น (URL) จะได้รับความน่าเชื่อถือมากกว่าข่าวที่มีแต่เนื้อหาเป็นข้อความเพียงอย่างเดียว

จากสมมติฐานเบื้องต้นที่ได้กล่าวมาข้างต้น จะนำไปสู่เป้าหมายสำหรับการทำวิจัยดังต่อไปนี้

1) ทำให้แยกแยะคุณลักษณะของข้อมูลที่ส่งผลต่อจำแนกข่าวปลอมในเครือข่ายสังคมออนไลน์ทวิตเตอร์ (Twitter)

2) เพื่อค้นหาปัจจัยหรือคุณลักษณะเฉพาะที่ทำให้สามารถจำแนกข่าวปลอมออกจากข่าวจริง

3) เสนอคำแนะนำเพื่อการหลีกเลี่ยงการส่งต่อ หรือขยายความข่าวปลอมออกไป

ขอบเขตการดำเนินงานวิจัย ประกอบด้วย

1) งานวิจัยนี้ใช้ข้อมูลจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ในการเก็บข้อมูล

2) เลือกเก็บข้อมูล 22 คุณลักษณะ จากเครือข่ายสังคมออนไลน์ทวิตเตอร์ ได้แก่

*Id, Name, IsVerified, ProfileImageUrl, FollowersCount, FriendsCount, FavouritesCount, StatusesCount, Description, Location, TimeZone, UserCreatedDate, Status, Url, Mentions, Number of Mentions, HashTags, Number of HashTags, RetweetCount, TweetCreatedDate, MessageText, MessageImage*

ประโยชน์ที่คาดว่าจะได้รับ

1) ทราบคุณลักษณะของข้อมูลที่ส่งผลต่อจำแนกข่าวปลอมออกจากข่าวจริงในเครือข่ายสังคมออนไลน์ทวิตเตอร์

2) สามารถตัดสินใจเชื่อข่าวได้จากคุณลักษณะที่ค้นพบ

3) ได้ข้อเสนอแนะเพื่อหลีกเลี่ยงการส่งต่อ หรือขยายความข่าวปลอม

# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

เอกสารบทนี้แสดงความเกี่ยวข้องกับหลักการ ทฤษฎีต่าง ๆ ที่เกี่ยวข้องกับข่าวสาร ทั้งในส่วนของสื่อสังคมออนไลน์ที่ใช้ในการเผยแพร่ข่าวสาร การนิยามความหมายของข่าวจริง ข่าวปลอม องค์ประกอบของข่าว ประเภทของข่าว ลักษณะเด่นของข่าวปลอม สาเหตุที่ทำให้เกิดข่าวปลอม รวมถึงเรื่องที่เกี่ยวข้องกับการจำแนกข่าวปลอม ปัญหาและผลกระทบที่เกิดจากข่าวปลอม การป้องกันการรับรู้ข่าวปลอมเบื้องต้น คุณลักษณะเด่นของทวิตเตอร์ (Twitter) ตลอดจนงานวิจัยที่เกี่ยวข้อง ดังรายละเอียดต่อไปนี้

## 2.1 สื่อสังคมออนไลน์กับการเผยแพร่ข่าวสาร

จากการใช้งานอินเทอร์เน็ตและเครือข่ายสังคมออนไลน์ในการสื่อสารระหว่างกันเพิ่มมากขึ้น โดยเฉพาะในประเทศไทยที่มีการเจริญเติบโตด้านการใช้อินเทอร์เน็ตผ่านอุปกรณ์ไร้สายมากเป็นอันดับต้น ๆ ของโลก [11] ดังนั้นจึงเกิดการรายงานข่าวที่รวดเร็ว มีการส่งต่อข่าวสารที่ได้รับทั้งเรื่องจริงหรือเรื่องหลอกลวง โดยที่ผู้ส่งเองอาจไม่ได้ตรวจสอบและไม่ทราบว่าสิ่งที่ได้ส่งต่อไปนั้นเป็นเรื่องจริงหรือไม่จริง

แต่ละวินาทีมีข้อมูลใหม่ ๆ เกิดขึ้นมากมายนับไม่ถ้วนบนเครือข่ายสังคมออนไลน์ ไม่ว่าจะเป็นข้อมูลข่าวสาร กิจกรรมต่าง ๆ เนื้อหาบทความที่มีคนสร้างขึ้นใหม่ตลอดเวลาผ่านสื่อต่าง ๆ โดยในที่นี้ *สื่อ* มีความหมายถึง ภาพ เสียง เนื้อหา ที่ประกอบด้วยข้อมูลที่ผู้รับและผู้ส่งสามารถใช้สื่อสารถึงกันได้ตามวัตถุประสงค์ที่ต้องการ

คนที่ใช้อินเทอร์เน็ตในชีวิตประจำวัน ส่วนใหญ่จะได้รับข่าวสารต่าง ๆ เป็นกิจวัตรจากแหล่งข่าวต่าง ๆ หลายแหล่ง ทั้งส่วนที่เป็นเครือข่ายสังคมออนไลน์ (Online Social network) เช่น Facebook, Twitter หรือเว็บไซต์บริการของแหล่งข่าวต่าง ๆ เช่น Thairath, Dailynews, Nations, Manager, Khaosod, Matichon, Kapook, Pantip หรือแม้แต่เว็บไซต์เลียนแบบสำนักข่าวที่ปล่อยข่าวปลอม ซึ่งปริมาณข้อมูลข่าวสารที่เกิดขึ้นมามากมายประกอบด้วยข่าวจริง ข่าวปลอม ข่าวลือ ความคิดเห็นส่วนบุคคลที่เกี่ยวกับข่าว ความหลากหลายของข่าวที่มีการผสมปนเปกันทำให้ผู้รับข่าวสารแยกแยะความจริงได้ยาก เกิดความสับสนในการรับรู้ข่าวสาร ซึ่งนำมาให้เกิดความเสียหายในชีวิตและทรัพย์สินได้

เครือข่ายสังคมออนไลน์ทวิตเตอร์ (Twitter) เป็นช่องทางการสื่อสารของข่าวทันเหตุการณ์ช่องทางหนึ่งที่ได้รับความนิยมในการกระจายข่าวสารต่าง ๆ อย่างรวดเร็ว ด้วยคุณลักษณะของ

ทวิตเตอร์ที่สามารถส่งข้อความสั้นได้อย่างกระชับ ทำให้ผู้รับสามารถรับรู้ประเด็นสำคัญของข่าวได้โดยไม่เสียเวลามากนัก และสามารถส่งต่อให้ใครก็ได้อย่างรวดเร็ว

เมื่อมีการรับส่งข่าวสารกันอย่างมากมาย ปัญหาที่ตามมาคือ การจำแนกเนื้อหาข่าวสารเมื่อผู้รับข่าวสารได้รับว่าข่าวนั้นเป็นเรื่องจริงหรือเรื่องหลอกลวง ซึ่งการส่งต่อเรื่องราวที่ไม่เป็นความจริงอาจก่อให้เกิดผลกระทบได้ทั้งในด้านดีและด้านร้ายตามมา ด้วยเนื้อหาเรื่องราวที่ปรากฏในข่าวอาจเป็นเรื่องไม่ดีสำหรับคนบางกลุ่มในขณะเดียวกันอาจเป็นเรื่องที่ดีสำหรับคนอีกกลุ่มก็เป็นได้ ตัวอย่างเช่น ข่าวเขื่อนแตกทำให้เกิดน้ำท่วม หากพิจารณาในด้านความเสียหายที่เกิดขึ้นกับกลุ่มคนที่อาศัยอยู่ในพื้นที่ที่เกิดเหตุการณ์ ย่อมได้รับผลกระทบและเกิดความสูญเสียในทรัพย์สิน ทำให้สภาพจิตใจย่ำแย่ เกิดความเครียดสะสม แต่จากเหตุการณ์เดียวกันนี้ หากพิจารณาในด้านบวกจากคนอีกกลุ่ม หลังจากเหตุการณ์น้ำท่วมจะต้องมีการฟื้นฟูสภาพบ้านเรือนที่อยู่อาศัย ย่อมเป็นโอกาสของกลุ่มผู้รับเหมาก่อสร้างที่จะมีงานทำ หรือกลุ่มสถาบันการเงินที่จะมีลูกค้าเพิ่มขึ้นจากการกู้ยืมเงินมาซ่อมแซมบ้านเรือน นักจิตวิทยามีกรณีศึกษาเพิ่มขึ้น ในการเข้าไปมีส่วนร่วมให้คำปรึกษาเพื่อฟื้นฟูสภาพจิตใจ ลดความเครียดให้ผู้ประสบภัย แต่ในด้านความรู้สึกของคนที่ไม่ได้รับผลกระทบกับเหตุการณ์โดยตรง อาจพิจารณาได้ว่าเป็นเรื่องราวที่น่าสงสาร น่าเห็นใจ และควรช่วยเหลือผู้ประสบภัยมากกว่า

ความสัมพันธ์ระหว่างผู้ใช้ในเครือข่ายสังคมออนไลน์ เป็นความสัมพันธ์รูปแบบเฉพาะที่ไม่สามารถพบได้ในการสำรวจผู้อ่านข่าวจากหนังสือพิมพ์หรือสื่ออื่น ๆ ทั่วไป การที่ผู้ใช้ในเครือข่ายสังคมออนไลน์สามารถส่งต่อข่าวสารให้คนอื่นที่รู้จักกันเป็นการส่วนตัวหรือใครก็ได้ ซึ่งเป็นการบ่งชี้ประการหนึ่งได้ว่า คนรับข่าวมีความเชื่อมั่นในเนื้อหาข่าว มั่นใจว่าการเผยแพร่กระจายข่าวนั้นเป็นเรื่องจริง สามารถส่งต่อไปยังผู้อื่นได้ ดังนั้นแหล่งข่าวตั้งต้นที่เริ่มต้นเผยแพร่ข่าวครั้งแรกไม่ว่าจะมีความน่าเชื่อถือมากหรือน้อยเพียงใด ย่อมส่งผลทำให้ผู้ใช้ที่ได้รับข่าวยินยอมที่จะเผยแพร่ข่าวต่อไปหรือไม่

การเผยแพร่ข่าวสารในแนวรูปแบบของ Clickbait โดยการใช้ประเด็นข่าวปลอมในการพาดหัวข่าว หลอกลวงให้ผู้อ่านเกิดความสนใจแล้วคลิกเพื่อหารายละเอียดมากขึ้น เกิดขึ้นกับหัวข้อพาดข่าวทั้งข่าวจริงและข่าวปลอม โดยการหลอกล่อให้คลิกลิงค์ด้วยการใช้ข้อความพาดหัวที่เกินจริง หรือเป็นเรื่องที่คนส่วนใหญ่กำลังอยากรู้ในช่วงเวลานั้น ทำให้สามารถดึงดูดความสนใจคนได้จำนวนมาก ๆ แต่เมื่อเข้าไปดูรายละเอียดข่าวแล้วจะพบว่าไม่มีสาระตามหัวข้อหรือข้อความพาดหัวข่าว บางครั้งอาจสร้างข่าวปลอมขึ้นมาเพื่อหลอกผู้อ่านหลงเชื่อ ทำให้คนสนใจแต่พาดหัวข่าวส่งต่อข้อมูลออกไป ดังนั้นข่าวปลอมเหล่านี้จึงสามารถกระจายออกไปเป็นวงกว้างได้เรื่อย ๆ โดยเป้าหมายของการทำ Clickbait คือต้องการให้เกิดการคลิกเข้าไปดูจำนวนมาก ๆ หรือเข้าไปมีปฏิสัมพันธ์กับหน้าเว็บเพจเพื่อแสดงโฆษณา ในสื่อสังคมออนไลน์มีความพยายามในการเผยแพร่ข่าวสาร ประเด็นสำคัญที่ทำให้ผู้คนเกิด

ความสนใจและติดตาม ทำให้ผู้สร้างข้อมูลได้รับประโยชน์จากเงินค่าโฆษณาเมื่อมีคนสนใจจำนวนมาก ซึ่งปัญหาที่เกิดขึ้นนี้ไม่เพียงหลอกให้ผู้สนใจข่าวเสียเวลา ในบางกรณีรายละเอียดของเนื้อข่าวยังไม่เป็นความจริงอีกด้วย ปัญหานี้ไม่เพียงแต่พบมากในประเทศไทย ยังเป็นปัญหาที่สามารถพบได้ทั่วโลก และยังคงเป็นที่สนใจในผู้คนมากมาย โดยเฉพาะหลังจากเกิดเหตุการณ์การปล่อยข่าวลือ ข่าวปลอมต่าง ๆ ช่วงก่อนการเลือกตั้งประธานาธิบดีประเทศสหรัฐอเมริกาเมื่อปี ค.ศ. 2016 [17]

ข่าวปลอมส่วนมากจะมีการอ้างอิงถึงข่าวลือในประเด็นที่กำลังเป็นที่ถกเถียงกันในสังคม อีกทั้งข่าวปลอมที่ส่งเป็นข้อความยังมีความคล้ายคลึงกับข่าวจริงหรือข่าวลือมาก ๆ ที่เป็นประเด็นที่กำลังมีการพูดถึงกันอย่างมากในสังคม ข่าวปลอมสามารถแพร่กระจายได้อย่างง่ายดายทางสื่อสังคมออนไลน์ จดหมายอิเล็กทรอนิกส์ การส่งข้อความคุยกันโดยตรง และการส่งข้อความไปยังทุกคนโดยไม่ระบุชื่อผู้รับปลายทาง ซึ่งเป็นการเผยแพร่แบบสาธารณะได้ คุณลักษณะของข่าวปลอมส่วนมากใช้ข้อมูลพื้นฐานจากความจริงและเป็นสิ่งที่คนกำลังให้ความสนใจ โดยมีการเติมแต่งความคิดเห็นส่วนบุคคลเพิ่มเติมลงไป [30]

ข่าวปลอมสามารถก่อให้เกิดความผิดพลาดจากความหวาดระแวงและความกลัวในภัยธรรมชาติ การจัดการลงทุนทางธุรกิจผิดพลาดมาจากการตัดสินใจจากการได้รับข่าวปลอมเช่นกัน หรือแม้แต่ข้อมูลประกอบการตัดสินใจเลือกตั้ง สามารถเกิดความผิดพลาดจากการได้รับข่าวปลอมแง่ลบในฝ่ายตรงข้ามได้ [19]

ข่าวปลอมที่เกิดขึ้นในทวิตเตอร์ เริ่มต้นจากผู้ใช้ที่อ้างสิทธิ์ถึงข้อความ รูปภาพ หรือส่วนเชื่อมโยงไปยังบทความออนไลน์หรือหลักฐานที่มีความเกี่ยวข้องกับเนื้อหาที่ต้องการสื่อสาร แล้วคนอื่นที่เห็นเผยแพร่ต่อด้วยการส่งต่อ ข้อความนั้นอีกครั้ง มีหลายหน่วยงานให้ความสำคัญในการวิเคราะห์ข่าว จากตัวอย่างที่เกิดเว็บไซต์บริการตรวจสอบค้นหาความจริง โดยการทำงานของเว็บไซต์เหล่านี้ต้องการให้มีการระบุหัวข้อข่าว เนื้อหาข่าว เพื่อจะนำข้อมูลไปประมวลผลออกมาเป็นความน่าเชื่อถือของหัวข้อข่าวที่ระบุ [6] [21] [24] โดยมีทั้งส่วนที่ใช้คนในการตรวจสอบความจริงและใช้ขั้นตอนวิธีหรืออัลกอริทึม (Algorithm) ในการประมวลผล [31] ความน่าเชื่อถือของข่าว ควรพิจารณาถึงที่มาของข่าวจากแหล่งข่าวที่มีความน่าเชื่อถือ อาจเป็นองค์กรที่นำเสนอข่าว หรือเป็นบุคคลที่มีความรู้ความเชี่ยวชาญเฉพาะด้าน หรือเป็นผู้ที่สามารถให้ข้อเท็จจริงได้ โดยไม่มีความผิดพลาดในการเสนอเนื้อหา ตลอดจนไม่มีการบิดเบือน หรือแสดงความคิดเห็นที่มีความลำเอียงกับข้อเท็จจริงของเนื้อหา หรือแสดงอคติกับประเด็นเนื้อหาที่ต้องการนำเสนอ [31]

## 2.2 นิยามของข่าวและข่าวปลอม

จากพจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๕๔ [32] ให้นิยามความหมายของคำที่เกี่ยวข้องกับข่าวที่เกี่ยวข้องกับงานวิจัยนี้ ดังต่อไปนี้

"ข่าว" หมายถึง *คำบอกเล่าเรื่องราวซึ่งโดยปรกติมักเป็นเรื่องเกิดใหม่หรือเป็นที่สนใจ คำบอกกล่าว คำเล่าลือ*

"ข่าวกรอง" หมายถึง *ข่าวที่ได้ตรวจสอบหลักฐานแล้วว่าเป็นข่าวที่เชื่อถือได้*

"ข่าวลือ" หมายถึง *ข่าวที่พูดกันทั่วไป แต่ยังไม่มีอะไรยืนยันได้แน่นอน*

ข่าว หมายถึง การรายงานสถานการณ์หรือเหตุการณ์ที่เป็นข้อเท็จจริงที่เกิดขึ้นหรือข้อคิดเห็นจากบุคคลสำคัญที่คนให้ความสนใจ อาจจะเป็นเรื่องที่มีผลกระทบต่อคนในสังคม หรือเป็นเรื่องราวใหม่ ๆ ข้อมูลใหม่ที่มีเนื้อหาที่คนส่วนใหญ่ยังไม่ทราบ หรืออาจเป็นเรื่องราวเสียดสี ในบางครั้งอาจมีการนำเรื่องจริงมาเขียนล้อเลียนให้กลายเป็นเรื่องตลกขบขัน

จากความหมายของคำว่าข่าวในพจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๕๔ ไม่ปรากฏการนิยามความหมายของคำว่า "ข่าวปลอม" หรือ "ข่าวลวง" ในที่นี้จึงใช้ความหมายจากคำที่เกี่ยวข้องในส่วนของ "ข่าวกรอง" เป็นความหมายในทางตรงกันข้าม และ "ข่าวลือ" ในความหมายที่ข่าวลืออาจจะเป็นข้อมูลจริงก็ได้หรืออาจจะเป็นส่วนหนึ่งของข่าวปลอม

ดังนั้น "ข่าวปลอม" หรือ "ข่าวลวง" จะหมายถึง คำบอกเล่าเรื่องราวที่เป็นเรื่องเกิดใหม่หรือเป็นเรื่องที่ผู้คนให้ความสนใจ อาจเป็นคำบอกกล่าว คำเล่าลือที่ยังไม่มีหลักฐานยืนยันว่าเป็นเรื่องจริงที่น่าเชื่อถือ หรืออาจเป็นเหตุการณ์ที่มีผู้เกี่ยวข้องออกมาแถลงการณ์ยืนยันแล้วว่าเนื้อหาข่าวได้ถูกบิดเบือนไป หรือไม่ได้เป็นความจริงตามที่เนื้อข่าวนั้นได้กล่าวอ้าง

การตรวจสอบข้อเท็จจริงของข่าวว่าเป็นความจริงหรือเป็นข่าวปลอม จึงมีความจำเป็นต้องศึกษาวิจัยค้นคว้า เพื่อค้นหาปัจจัยหรือคุณลักษณะเฉพาะที่ทำให้สามารถจำแนกข่าวปลอมออกจากข่าวจริงได้ ซึ่งจะเป็นการช่วยให้ผู้รับข่าวสารสามารถตัดสินใจเลือกว่าจะส่งต่อข่าวสารนั้นต่อไปอีกหรือไม่ เพื่อให้สังคมเกิดการส่งต่อข่าวสารที่เป็นจริง ส่งผลให้เกิดความช่วยเหลือกันในเหตุการณ์จำเป็นอย่างทันเวลา เกิดการรับรู้เรื่องราวดี ๆ ที่ทำให้ทุกคนมีความเบิกบานใจ ไม่เครียด เมื่อผู้คนสบายใจจะทำให้จำนวนผู้ป่วยต่าง ๆ ที่เกิดความอาการผิดปกติที่มีสาเหตุเนื่องมาจากความไม่สบายใจมีจำนวนลดลง [3]

หากเนื้อหาข่าวเป็นความจริง จะต้องมีหลักฐานจากข้อเท็จจริงที่ปรากฏอย่างชัดเจน สามารถพิสูจน์ได้ แต่หากเป็นข่าวปลอม ที่อาจเกิดจากคำเล่าลือ การบอกต่อ รวมถึงการส่งต่อที่เกิดขึ้นในสื่อสังคมออนไลน์ต่าง ๆ จะไม่สามารถหาหลักฐานมายืนยันความถูกต้อง อีกทั้งข่าวปลอมที่เกิดขึ้น ยังอาจส่งผลเสียหายต่อบุคคล สังคม เศรษฐกิจ ประเทศชาติ

ตัวอย่างข่าวปลอมที่เป็นปรากฏการณ์ครั้งแรกที่มีการบันทึกไว้ ถูกค้นพบเมื่อปีค.ศ. 1835 ในหนังสือพิมพ์ The Sun ของนิวยอร์ก เป็นบทความที่มีเนื้อหาทางด้านดาราศาสตร์ ของ Sir John Herschel [33] ที่กล่าวถึงสิ่งมีชีวิตมหัศจรรย์ที่อาศัยอยู่บนดวงจันทร์ โดยอ้างอิงจากการสำรวจดวงจันทร์ด้วยกล้องโทรทรรศน์แบบพิเศษ ซึ่งเป็นบทความ 6 เรื่องที่ในเวลาต่อมาเป็นรู้จักกัน

ในชื่อ The Great Moon Hoax ถึงแม้ว่าเนื้อหาในบทความจะเป็นเรื่องไม่จริง แต่ก็กลายเป็นประเด็นในวงการดาราศาสตร์ยุคนั้น เนื่องจากมีผู้คนมากมายให้ความสนใจ และต่อมาภายหลังบทความนี้ได้กลายเป็นเรื่องอ่านสนุกราวกับนวนิยาย เมื่อทุกคนทราบความจริงว่าเนื้อหาในบทความไม่ใช่เรื่องจริง

ข่าวปลอมเกิดจากข้อมูลที่ผิด [29] [33] และข้อมูลที่ถูกบิดเบือน หลายหน่วยงานมีความพยายามในการให้คำนิยามเกี่ยวกับข่าวปลอม ดังตัวอย่างต่อไปนี้

– firstdraftnews.com กล่าวว่า ข่าวปลอม เป็นเพียงส่วนหนึ่งของข่าวที่เกิดจากการสื่อสารข้อมูลที่ผิดพลาดและถูกบิดเบือน

– ห้องปฏิบัติการวิจัย Nieman Lab ในมหาวิทยาลัย Harvard พิจารณาจากบทสรุปในการศึกษาวิจัยที่เกี่ยวข้องกับการเพิ่มขึ้นของการรายงานข่าวจากข้อมูลที่เป็นข่าวปลอม เกิดจากข้อมูลที่ผิดพลาด เนื้อหาที่บิดเบือน การรับรู้ข่าวสารต่าง ๆ ของผู้คน และยังมีการพิจารณาการเผยแพร่ข่าวปลอมรวมถึงการแพร่กระจายของข้อมูลข่าวที่มีเนื้อหาไม่ถูกต้อง

– แนวคิดจากศาสตราจารย์ Melissa Zimdars ผู้เชี่ยวชาญทางด้านนิเทศศาสตร์ของวิทยาลัย Merrimack ได้กล่าวถึงข่าวปลอมว่าเป็นข้อมูลที่ไม่เป็นความจริง เป็นข้อมูลที่ทำให้คนเกิดความเข้าใจผิด หรือข่าวเสียดสีต่าง ๆ ที่จำต้องพิจารณาการวิเคราะห์แหล่งข่าว ตลอดจนวิเคราะห์จำนวนรายการต่าง ๆ ที่มีการสื่อสารผิดพลาด รวมถึงแหล่งข่าวที่เผยแพร่ข่าวเสียดสีออกสู่สาธารณชน

ข่าวปลอมส่วนใหญ่เกิดจากบัญชีที่ไม่ได้เป็นที่รู้จักกันมากนัก บางครั้งไม่ได้เป็นบุคคลที่มีความโดดเด่นหรือมีชื่อเสียง ไม่ได้เป็นกลุ่มคนที่มีอิทธิพลทางออนไลน์ ต้องการสร้างกระแสในโลกออนไลน์ ข่าวปลอม การแพร่กระจายออกไปได้ไกลมากเพียงใดขึ้นอยู่กับการส่งต่อของกลุ่มคนที่มีอิทธิพลในโลกออนไลน์ (Online Influencer) เนื่องจากกลุ่มคนเหล่านี้มีผู้ติดตามจำนวนมาก ดังนั้นการแสดงหรือส่งต่อข่าวปลอม หรือเพียงการกล่าวว่ามีคนลือว่า ก็อาจเป็นประเด็นที่ก่อให้เกิดความสงสัยขึ้นมาได้ ในทางการเมืองข่าวปลอมที่เกิดขึ้นส่งผลทางจิตวิทยากับประชาชน แม้ว่าผู้ส่งข่าวต่ออาจจะทราบว่าเป็นข่าวปลอม แต่ยังมีการส่งข่าวต่อไปอย่างต่อเนื่องที่จะทำให้เกิดผลบวกกับฝ่ายการเมืองฝั่งที่ตนเองชื่นชอบนั่นเอง [17]

ข่าวปลอมที่เกิดจากความเข้าใจคลาดเคลื่อน (Mis-Information) เป็นการส่งต่อข่าวปลอมโดยไม่ได้ตั้งใจ ไม่มีเจตนาปั่นป่วน ไม่เจตนาทำร้ายใคร แต่มีการส่งต่อข่าวไปจากความไม่รู้ โดยไม่ทราบว่าสิ่งที่ส่งต่อไปนั้นเป็นเรื่องไม่จริง เป็นเรื่องที่ถูกหลอกลวง หรือในบางครั้งข่าวที่ส่งต่ออาจเป็นเรื่องที่ข้อเท็จจริงยังไม่คลี่คลาย แต่ในท้ายที่สุดพบว่าเป็นเรื่องไม่จริง

## 2.3 องค์ประกอบของข่าว

ข่าว ที่ทำให้ผู้อ่านให้ความสนใจ ประกอบด้วยโครงสร้างสำคัญ [2] [3] ดังต่อไปนี้

– ที่มาของข่าว อาจเป็นผู้เขียนข่าวหรือสำนักพิมพ์หรือหน่วยงานที่เผยแพร่ข่าว

– พาดหัวข่าวเป็นข้อความสั้น ๆ และกระชับ ใช้แสดงใจความหลักและประเด็นสำคัญที่เป็นความน่าสนใจของข่าว เพื่อดึงดูดความสนใจของผู้อ่าน อีกทั้งยังเป็นส่วนสำคัญที่สามารถตอบสิ่งที่ผู้อ่านต้องการทราบจากข่าวได้ โดยสามารถตอบคำถามในรูปแบบ 5W1H ได้แก่ ใคร (Who) ทำอะไร (What) เมื่อไหร่ (When) ที่ไหน (Where) ทำไม (Why) อย่างไร (How)

– เนื้อหาข่าวเป็นข้อความที่แสดงรายละเอียดที่เป็นใจความสำคัญของเรื่องหรือเหตุการณ์ เป็นลำดับเหตุการณ์หรือรายละเอียดเฉพาะที่ต้องการสื่อในข่าว

– รูปภาพ หรือ วีดิทัศน์ แสดงรายละเอียดประกอบเนื้อหาข่าวที่ทำให้เห็นอ่านเห็นภาพของเรื่องราวข่าวได้ชัดเจนกว่าการอ่านข้อความเพียงอย่างเดียว

เงื่อนไขของเหตุการณ์ที่นำมาเสนอเป็นข่าว เพื่อให้ได้รับความสนใจจากผู้อ่าน ประกอบด้วย

– มีการรายงานด้วยความรวดเร็ว ทันต่อเหตุการณ์ที่กำลังเกิดขึ้นในขณะนั้น หรือเพิ่งจะเกิดขึ้นในเวลาไม่นาน หรือสามารถนำเสนอข่าวสารได้อย่างรวดเร็วทันทีทันใด

– เป็นข้อมูลของเหตุการณ์ที่เกิดขึ้นแล้วจะมีความเกี่ยวข้องกับผู้อ่าน เป็นเรื่องใกล้ตัวที่อาจพบได้ในชีวิตประจำวัน หรือเป็นเรื่องส่งผลกระทบโดยตรงต่อผู้อ่าน เป็นเรื่องที่ส่งผลต่อความเป็นอยู่ของผู้คนแต่ละคนมากน้อยเพียงใด หรือส่งผลกระทบต่อคนส่วนใหญ่ในสังคม หากเป็นเรื่องที่มีผลกระทบมาก ข่าวนั้นย่อมจะได้รับความสนใจจากผู้อ่านมากขึ้น

– เป็นเหตุการณ์ที่เกิดกับบุคคลสำคัญ บุคคลที่มีชื่อเสียง หรือสถานที่สำคัญที่ทุกคนรู้จักกันดี

– ความผิดปกติหรือเหตุการณ์แปลกประหลาดที่เกิดขึ้น ที่ไม่เคยเกิดขึ้นมาก่อน เหตุการณ์ที่ทำให้ผู้คนเกิดความสงสัยและสนใจ

– ความมีเงื่อนงำ ความน่าสงสัยของเหตุการณ์ที่สาเหตุทำให้เกิดนั้นยังไม่ถูกเปิดเผย จึงยังทำให้เหตุการณ์ยังเป็นที่น่าสนใจของคนที่มีความอยากรู้อยากเห็นกับเหตุการณ์นั้น ๆ

– ความขัดแย้งกันที่เกิดขึ้นในเหตุการณ์มีการให้ข้อมูลที่ไม่ตรงกัน โดยปกติคนส่วนใหญ่จะให้ความสนใจกับประเด็นความขัดแย้ง ไม่ว่าจะเป็นความขัดแย้งด้านการแสดงความคิดเห็น หรือความสัมพันธ์ต่าง ๆ และก่อให้เกิดปัญหาอื่น ๆ ตามมา

– เป็นเรื่องราวที่ทำให้เกิดอารมณ์ร่วมไปกับเหตุการณ์หรือบุคคลในข่าว อาจจะเป็นเรื่องน่าตื่นเต้น ทำให้เกิดความปลื้มปิติ ดีใจ ให้ความชื่นชมยินดี หรือก่อให้ความเศร้าโศกสะเทือนใจ

– ความเจริญก้าวหน้าในการพัฒนาด้านต่าง ๆ ทั้งด้านเทคโนโลยี การประดิษฐ์คิดค้นนวัตกรรมใหม่ ๆ

## 2.4 ประเภทของข่าว

การนำเสนอข่าวส่วนใหญ่แบ่งตามประเภทของหัวข้อการพาดข่าว ดังนี้

- เหตุการณ์ปรากฏการณ์ทางธรรมชาติ และภัยพิบัติจากภัยธรรมชาติ
- เศรษฐกิจ การค้าและการลงทุน
- อาชญากรรม
- สงครามและการก่อการร้าย
- การเมือง
- วิทยาศาสตร์เทคโนโลยี
- สิ่งแวดล้อม
- กีฬา
- ข่าวบันเทิง
- ข่าวทั่วไป

จากงานวิจัย [27] พบว่าข่าวปลอมที่พบมากในปัจจุบัน สามารถแบ่งประเภทตามระดับความรุนแรงของผลกระทบได้ดังต่อไปนี้

- ข่าวหลอกลวงที่ส่งผลรุนแรง เป็นรายงานความรุนแรงที่ไม่เคยมีการรายงานหรือกล่าวถึงมาก่อนทั้งในสื่อสิ่งพิมพ์และสื่อออนไลน์ต่าง ๆ การเปิดเผยประเด็นข่าวลักษณะนี้มีความเหมาะสมที่จะนำมาจัดเก็บเป็นฐานข้อมูลข่าวปลอม แต่การเก็บข้อมูลในลักษณะแบบนี้ต้องใช้เวลานานและจะต้องมีการเก็บสำเนาต้นฉบับเพื่อใช้อ้างอิงข่าวที่เกิดขึ้น
- ข่าวหลอกลวงที่ส่งผลกระทบในวงกว้าง เป็นการหลอกลวงเป็นอีกรูปแบบหนึ่งที่มีเจตนาปลอมแปลง ในสื่อหลักต่าง ๆ บนเครือข่ายสังคมออนไลน์ เจตนาและมีความพยายามที่จะบิดเบือนข่าวโดยการหลอกลวงผู้อ่านด้วยเนื้อหาข่าวอันเป็นเท็จ อาจทำให้สื่ออื่นหลงเชื่อ ร่วมรายงานข่าวหลอกลวงนี้ด้วยความไม่รู้จริง และนำไปรายงานซ้ำอีกด้วย
- ข่าวปลอมที่เป็นเรื่องตลก อาจเป็นเรื่องน่าขำขัน ที่ผู้อ่านจะรู้สึกถึงอารมณ์ร่วมไปกับการรายงานข่าว เช่น ข่าวล้อเลียน ข่าวหยอกล้อทางการเมือง การวิจัยในข่าวกลุ่มนี้มีความพยายามใช้เทคโนโลยีที่หลากหลายเพื่อการวิเคราะห์อารมณ์ความรู้สึกด้านบวกหรือด้านลบหลังจากที่ผู้อ่านได้รับรู้เนื้อหาข่าว

## 2.5 ลักษณะของข่าวปลอม

ปัญหาหนึ่งในการวิเคราะห์ข่าวคือการแยกข่าวปลอมออกจากข่าวจริง ซึ่งจุดเด่นของข่าวปลอมส่วนใหญ่ที่พบมีข้อสังเกตดังนี้ [22] [31]

– Clickbait มีการพาดหัวข้อข่าวที่ทำให้ตกใจ คนอ่านเกิดความสนใจอยากรู้จึงคลิกเข้ามาดูรายละเอียดเนื้อหามากขึ้น แต่เนื้อหารายละเอียดภายในข่าวไม่ปรากฏสาระสำคัญ หรือในบางครั้งอาจเป็นเรื่องไม่จริงทั้งหมด เพียงต้องการหลอกให้คนตื่นตกใจ อยากรู้อยากเห็นและคลิกเข้ามาดูรายละเอียดเพิ่มขึ้นเท่านั้น [34]

– Propaganda เป็นข่าวในลักษณะของการโฆษณาชวนเชื่อที่มีเจตนาหลอกให้คนอ่านเข้าใจผิด แสดงการต่อต้าน หรือมีการยุยงปลุกปั่น ก่อให้เกิดความเกลียดชังกัน และเกิดความไม่สงบ

– Commentary/Opinion เป็นการแสดงความคิดเห็นหรือแสดงปฏิกิริยาตอบโต้ด้วยความลำเอียงด้านใดด้านหนึ่งกับเหตุการณ์ที่เกิดขึ้นในขณะนั้น ในบางครั้งมีความพยายามชักจูงให้ผู้อ่านเกิดความเห็นคล้อยตามกันไปด้วย

– Humor/Satire เป็นบทความเน้นด้านความบันเทิง อาจเป็นเรื่องตลก เรื่องราวการล้อกันเล่น เพื่อให้เกิดความสนุกสนานเพลิดเพลินในการรับชมและไม่ให้เกิดความเครียด

หากพิจารณาจากมุมมองตามแบบสังคมศาสตร์ (Social Science) ลักษณะของข่าวปลอมแบ่งเป็นประเภทต่าง ๆ ดังนี้ [35]

- Visual-based: ใช้กราฟิกหรือสื่อประสมต่าง ๆ ในการแสดงเนื้อหาข่าว อาจอยู่ในรูปแบบรูป ภาพ วีดิทัศน์ หรือทั้งสองแบบ
- User-based: ผู้ใช้งานใช้บัญชีปลอมในการเผยแพร่ข่าว เพื่อให้ส่งต่อข่าวได้ตรงกับกลุ่มเป้าหมายที่ต้องการ โดยอาจกำหนดประวัติผู้ใช้งานตามกลุ่มอายุ เพศ เรื่องราวที่สนใจร่วมกัน
- Post-based: ตามลักษณะการโพสต์บนกลุ่มที่ข่าวนั้นไปปรากฏบนสื่อเครือข่ายสังคมออนไลน์ ในรูปแบบข้อความ รูปภาพ ภาพเคลื่อนไหว หรือวีดิทัศน์
- Network-based: เครือข่ายที่เชื่อมต่อกัน หมายถึงกลุ่มคนที่มีความสนใจบางอย่างร่วมกันหรือชอบเข้าร่วมกิจกรรมบางอย่างด้วยกัน
- Knowledge-based: กลุ่มที่แบ่งปันข้อมูลผิด ๆ บางอย่างร่วมกัน บนความเชื่อร่วมกันว่าสิ่งนั้นเป็นความจริง
- Style-based: เป็นรูปแบบที่ข่าวปลอมถูกนำเสนอโดยบุคคลธรรมดาที่ไม่ใช่นักข่าวอาชีพ เพียงแต่ต้องการร่วมรายงานข่าวคราวต่าง ๆ
- Stance-based: เน้นไปที่วิธีการแสดงข้อความข่าวทั้งในส่วนของการพาดหัวข่าวแสดงความเห็นด้วยหรือไม่เห็นด้วยกับแต่ละข่าวเฉพาะ

## 2.6 สาเหตุที่ทำให้เกิดข่าวปลอม

สาเหตุที่ผู้อ่านให้ความสนใจข่าวปลอม ข่าวลวง เกิดจาก

– ผู้คนสนใจอยากรู้อยากเห็นและสงสัยในข่าวที่เป็นประเด็น อาจเป็นเรื่องที่อยู่บนพื้นฐานข้อเท็จจริงบางอย่าง หรืออาจเป็นเรื่องหลอกที่แต่งขึ้นก็ได้ ซึ่งความไม่แน่นอนที่ปรากฏในเนื้อหาข่าว จะกระตุ้นให้ผู้คนสนใจติดตามอย่างไม่รู้ตัว

– ผู้คนส่วนใหญ่มีความอยากรู้อยากเห็นในเรื่องคนอื่น เนื่องจากมนุษย์เป็นสัตว์สังคม จึงมีความอยากรู้อยากเห็นในเรื่องราวต่าง ๆ ที่เกิดขึ้นในสังคม ที่แม้จะไม่ใช่เรื่องของตนเองก็ตาม

– ผู้คนชอบเรื่องที่เห็นคนอื่นทำอะไรผิดไปจากทำนองครองธรรมที่ควรจะเป็น เนื่องจากจะทำให้รู้สึกสะใจแบบลึก ๆ เหมือนกำลังได้ติดตามนิยายเรื่องหนึ่ง

– ผู้อ่านชอบอ่านเรื่องราวที่ตัวเองจะเดาตอนจบของเรื่องได้ ไม่ว่าเรื่องราวนั้นจะเป็นจริงหรือไม่ก็ตาม หรือจะมีความเป็นไปอย่างไร ซึ่งเรื่องราวที่ได้อ่าน ทำให้เกิดความตื่นเต้น สนใจติดตามตอนจบจริง ๆ ที่จะเกิดขึ้น

– ผู้อ่านชอบการวิพากษ์วิจารณ์ ซึ่งในข่าวฉาวส่วนใหญ่จะเป็นเรื่องราวการกระทำความผิดของคน ซึ่งเมื่อเรื่องราวถูกเปิดเผยสู่สาธารณชนแล้ว จะมีการวิเคราะห์โจษจันในลักษณะว่ากล่าวติติง นับว่าเป็นการลงโทษจากสังคมออนไลน์ โดยผู้คนส่วนใหญ่จะรู้สึกพอใจกับผลการกระทำจากสังคมออนไลน์ในลักษณะที่พึงพอใจกับความยุติธรรมที่เกิดขึ้น และมีความพึงพอใจกับการลงโทษคนผิดในประเด็นข่าว

“ความอยากรู้อยากเห็น” คือ ความต้องการที่จะเรียนรู้หรืออยากรู้เรื่องราวต่าง ๆ โดยนักจิตวิทยา George Loewenstein ได้กล่าวว่า ความอยากรู้อยากเห็นเป็นแรงจูงใจสำคัญที่ส่งผลต่อพฤติกรรมของมนุษย์ [29]

วัตถุประสงค์ของผู้สร้างข่าวปลอม [36]

1) ต้องการสร้างกระแสสังคม เพื่อให้ผู้คนหันมาให้ความสนใจผู้สร้างข่าวปลอม

2) อยากเป็นบุคคลที่มีชื่อเสียง และเป็นที่น่าจดจำของคนทั่วไป

3) การหวังผลทางการเมือง โดยการปล่อยข่าวที่ทำให้เกิดกระแสความนิยมกับนักการเมืองบางคน หรือข่าวที่ทำให้เกิดความเข้าใจผิดในฝ่ายตรงข้าม ซึ่งอาจก่อให้เกิดกระแสต่อต้านอันส่งผลต่อสถานภาพทางการเมือง เช่น ข่าวลือ ข่าวปลอมต่าง ๆ ที่เกิดขึ้นในช่วงระหว่างการหาเสียงเลือกตั้งประธานาธิบดีในประเทศสหรัฐอเมริกาเมื่อปี ค.ศ. 2016 [17]

## 2.7 วิธีการจำแนกข่าวปลอม

เมื่อผู้อ่านได้รับข่าวสารใด ๆ จะมีความเชื่อมั่นความจริงที่ปรากฏในเนื้อหาข่าวนั้นได้มากน้อยเพียงใด มีงานวิจัย [20] ได้กล่าวถึงวิธีการพิจารณาข่าวปลอมเบื้องต้นไว้ดังนี้

1) แหล่งที่มาของข่าว

– แหล่งที่มาที่นำเสนอข่าว เป็นแหล่งข่าวที่เป็นองค์กรนำเสนอข่าวอาชีพ หรือแหล่งข่าวที่เป็นบุคคลที่มีชื่อเสียง หรือเป็นเพียงบุคคลทั่วไป [35]

– ความน่าเชื่อถือของแหล่งข่าว เป็นแหล่งข่าวที่ไม่เคยเสนอข่าวผิดพลาด ข่าวปลอม หรือข่าวลือต่าง ๆ

– แหล่งข่าวนั้นมาจากผู้ที่มีความรู้จริง หรือเป็นผู้ที่มีความเชี่ยวชาญเกี่ยวข้องกับเนื้อหาที่ปรากฏ หรือเป็นนักวิชาการด้านต่าง ๆ ที่เกี่ยวข้อง

– แหล่งข่าวที่ให้ข้อมูลเป็นบุคคลที่ได้รับผลกระทบโดยตรงจากการนำเสนอข่าวปลอม อาจจะเป็นบุคคลในข่าว หรือผู้ที่อยู่ร่วมในเหตุการณ์

– พิจารณาเรื่องราวที่ปรากฏในเว็บไซต์เพื่อสำรวจเว็บไซต์ ภารกิจของหน่วยงาน และมีข้อมูลสำหรับการติดต่อ

– ข่าวจากแหล่งเดียวกัน สามารถเผยแพร่กระจายออกไปอย่างรวดเร็วโดยการส่งต่อในหลายช่องทางเครือข่ายสังคมออนไลน์ทั้งทางเว็บไซต์ หรือเครือข่ายสังคมออนไลน์ [37]

2) ผู้เสนอข่าว

– ประวัติของผู้นำเสนอข่าว ความเชี่ยวชาญ ชำนาญในงานเฉพาะด้าน ความน่าเชื่อถือที่เกี่ยวข้องกับเนื้อหาที่นำเสนอนั้น ๆ ซึ่งอาจเป็นบุคคลหรือองค์กรสื่อสารมวลชน

– ผู้นำเสนอข่าว แสดงข้อมูลที่เกี่ยวข้องที่มีความน่าเชื่อถือ และความมีตัวตนจริง ๆ โดยที่อาจเป็นสำนักข่าว ผู้สื่อข่าว รวมถึงนักเขียนบทความผ่าน blogger

– ความถี่ในการโพสต์ ความถี่ในการนำเสนอข่าว หรือความถี่ในการกระจายข่าวส่งต่อข่าว

– การตอบสนองต่อเนื้อหาที่นำเสนอ หรือการตอบโต้ด้วยการแสดงความคิดเห็นต่าง ๆ

– พิจารณาความคิดเห็น ความเชื่อส่วนบุคคล และผลกระทบกับการตัดสินใจ มีความน่าเชื่อถือระดับใด

3) เนื้อหา

– การหัวข้อพาดข่าวมีความสัมพันธ์กับเนื้อหาทั้งหมดหรือไม่ หรือเพียงต้องการชักชวนให้ผู้อ่านสนใจคลิกเข้าไปดูรายละเอียดของข่าวเท่านั้น

– รายละเอียดของเนื้อหาข่าว อยู่ในหลายรูปแบบ ดังต่อไปนี้ [35] [38] [39] [40] [41] [42] [43] [44] [45] [46]

– ตัวอักษรที่เป็นข้อความในการบรรยายเนื้อหาข่าว มีการวิเคราะห์ความหมายของข้อความ การวิเคราะห์ความหมายแฝง การวิเคราะห์หัวข้อข่าว

– รูปภาพประกอบที่เกี่ยวข้องกับเนื้อหา หรือเหตุการณ์ที่ต้องการนำเสนอ อาจพิจารณาความเหมือนความคล้ายกันของรูปภาพ ความหลากหลายของรูปภาพที่แสดง การค้นคืนรูปภาพที่เกี่ยวข้อง การพิจารณาคุณลักษณะต่าง ๆ ของรูปภาพในเชิงสถิติ

– วีดิทัศน์ที่เกี่ยวข้องกับเหตุการณ์ข่าวนั้น ๆ

– จุดเชื่อมโยง อ้างอิงไปยังแหล่งข่าวอื่นที่เกี่ยวข้องกับเหตุการณ์ มีแหล่งข้อมูลที่สนับสนุนเนื้อหาข่าวมีความน่าเชื่อถือมากน้อยเพียงใด

4) เหตุการณ์

– พิจารณาจากประเภทของเหตุการณ์ที่เกิดขึ้น

– ผลกระทบที่เกิดจากเหตุการณ์ เกิดขึ้นกับบุคคล องค์กร หรือสังคม

– สามารถแสดงหลักฐานหรือข้อพิสูจน์ทางวิทยาศาสตร์ที่สามารถยืนยันความถูกต้องของข้อมูลที่นำเสนอในเหตุการณ์ได้

– เนื้อหาข่าวที่เกี่ยวข้องกับเหตุการณ์ การประมวลผลภาพ การพิจารณาความสัมพันธ์ระหว่างตำแหน่งที่เกิดกับภาษาที่ใช้ในการนำเสนอเหตุการณ์ ประเภทของเหตุการณ์ อาทิเช่น อาชญากรรม รัฐบาลและการเมือง เศรษฐกิจและสังคม สภาพภูมิอากาศ เรื่องราวที่เกี่ยวข้องกับความบันเทิง ข่าวสารที่เกี่ยวกับราชวงศ์ บุคคลที่มีชื่อเสียงและทรงอิทธิพลในแต่ละวงการ

5) เวลา

– เวลาที่เกิดเหตุการณ์

– เวลาที่มีการนำเสนอข่าวที่เกี่ยวข้องกับเหตุการณ์ครั้งแรก

– ระยะเวลาที่ยังคงมีการกระจายข่าวที่เกี่ยวข้องกับเหตุการณ์นั้น

– ความทันสมัยของวันเวลาที่นำเสนอเหตุการณ์ การนำเสนอเรื่องราวเก่า ๆ ที่มีความเกี่ยวข้องกับเรื่องราวในปัจจุบัน

– การเสนอข่าวสารในเวลาที่ล่วงเลยมาแล้ว จะต้องสามารถแสดงหลักฐานเพื่อประกอบการพิจารณาตัดสินใจความน่าเชื่อได้หรือไม่ได้ แต่หากเป็นเวลาที่เหตุการณ์นั้นยังไม่เกิดขึ้น ยังไม่สามารถพิจารณาว่าเกิดเหตุการณ์จริง เช่นเดียวกับที่กล่าวหรือไม่ หรือเกิดเหตุการณ์ขัดแย้งกับประเด็นข่าวที่มีการกล่าวอ้างถึงหรือไม่ ทำได้เพียงรอให้เวลามาถึงแล้วจึงจะทราบข้อเท็จจริง

การวิเคราะห์ข้อความข่าวเพียงอย่างเดียว ไม่สามารถนำมาใช้วิเคราะห์แยกข่าวปลอมได้ถูกต้อง การประเมินข้อมูลจากสื่อเครือข่ายสังคมออนไลน์ การพิจารณาอ้างอิงแหล่งข้อมูลที่มีความ

น่าเชื่อถือ การระบุบุคคลที่เขียนบทความ การพิจารณาลำดับความสำคัญของข้อมูลบางประเภท การทำความเข้าใจกับคุณสมบัติพื้นฐานของผู้นำเสนอข่าว อาจเป็นส่วนหนึ่งในการพิจารณาความน่าเชื่อถือของข้อมูลที่ได้รับจากสื่อสังคมออนไลน์

## 2.8 ปัญหาและผลกระทบที่เกิดจากข่าวปลอม

การที่คนได้รับข่าวสารมากมายในปัจจุบัน ทั้งเรื่องจริงเรื่องปลอม โดยที่ไม่สามารถแยกแยะได้ว่าเรื่องราวของข่าวไม่ใช่เรื่องจริง อาจส่งผลให้เกิดผลเสียตามมา เกิดความหวั่นวิตกในเหตุการณ์ร้ายแรงต่าง ๆ อาจส่งผลให้เกิดความเสียหายทางเศรษฐกิจและสังคม [17] อาทิเช่น เมื่อคราวที่รัชกาลที่ ๙ ใกล้จะเสด็จสวรรคต มีข่าวลือ ข่าวปลอมสร้างกระแสออกมาตลอดเวลา ทำให้ประชาชนเกิดความหวั่นไหว ดัชนีตลาดหลักทรัพย์แห่งประเทศไทยได้รับผลกระทบ มีการเทขายหุ้นออกมาจนกระทั่งมีค่าติดลบมากถึงเกือบหนึ่งร้อยจุด [12] ทำให้นักลงทุนบางคนอาจได้รับความเสียหายจากเหตุการณ์นี้ [47]

ตัวอย่างข่าวปลอมทำให้สังคมเกิดความสับสน และมีการตรวจสอบยืนยันข้อเท็จจริงจากผู้ที่เกี่ยวข้องแล้ว [48] เช่น

ประเทศไทยอนุญาตให้บุคคลสัญชาติไทยที่มีสมาชิกในครอบครัวอย่างน้อย 3 คน สามารถปลูกกัญชาได้ไม่เกินครอบครัวละ 50 ต้น อีกทั้งยังสามารถใช้กัญชาในทางสันทนาการได้ จากการตรวจสอบข้อมูลพบว่าข้อความมีเนื้อหาบิดเบือนไปจากข้อเท็จจริง โดยพระราชบัญญัติยาเสพติดให้โทษ ฉบับที่ 7 พ.ศ.2562 ระบุว่า อนุญาตให้สามารถนำกัญชาไปใช้ประโยชน์เฉพาะทางการแพทย์เท่านั้น ส่วนการนำไปใช้เพื่อหวังผลด้านอื่น ๆ เช่น ใช้ในด้านสันทนาการ หรือการเปิดเสรีการปลูกกัญชาได้ครอบครัวละไม่เกิน 50 ต้น ยังไม่มีการประกาศอนุญาตอย่างเป็นทางการ ดังนั้นผู้ใดที่ลักลอบครอบครองหรือจำหน่าย ยังคงมีความผิดตามกฎหมาย

การกินน้ำที่เย็นจัดบ่อย ๆ จะทำให้เป็นนิ่ว จึงต้องระวังและหลีกเลี่ยงการดื่มน้ำเย็น สำนักงานคณะกรรมการอาหารและยา (อย.) ชี้แจงว่าข้อความนี้ไม่เป็นความจริงเนื่องจากการกินน้ำเย็นจัดบ่อย ๆ ไม่มีความเกี่ยวข้องกับการเป็นโรคนิ่วในไต สาเหตุของการเกิดโรคนิ่วเกิดได้จากหลายปัจจัยร่วมกัน ตั้งแต่ปัสสาวะมีความเข้มข้นสูง ดื่มน้ำน้อยและสูญเสียเหงื่อมาก หรือกินอาหารที่มีแคลเซียมสูง โปรตีนสูง โซเดียมสูง เป็นประจำ หรือมีภาวะยูริกในเลือดสูงจากโรคเกาต์ หรือคนที่ได้รับเคมีบำบัดระหว่างเป็นโรคมะเร็ง และกินยาในกลุ่มที่มีความเสี่ยงสูง เช่น ยากันชัก (ไดแลนติน) หรือมีประวัติคนในครอบครัวเป็นโรคนิ่วมาก่อน

ภาพประกอบข้อความที่ระบุว่า มีการประกาศใช้ พ.ร.บ. ข้าว ที่ระบุว่า ชาวนาต้องขายข้าวผ่านรัฐบาลเท่านั้น หากฝ่าฝืนจะได้รับโทษหนักจำคุก 5 ปี ปรับเป็นเงิน 5 แสนบาท ข้อเท็จจริงคือสภานิติ

บัญญัติแห่งชาติ (สนช.) ได้ถอดถอนวาระการพิจารณาร่าง พ.ร.บ. ข้าว ออกไปอย่างไม่มีกำหนด เพื่อไม่ให้เกิดความขัดแย้งทั้งฝ่ายที่เห็นด้วยและฝ่ายที่ไม่เห็นด้วย

เว็บไซต์อินเวสติ้ง เว็บไซต์ข่าวด้านการเงิน เศรษฐกิจ และการลงทุน ได้รายงานชื่อ รองนายกรัฐมนตรีของไทยเป็นหนึ่งในบุคคลที่มีฐานะร่ำรวยในระดับเศรษฐีของทวีปเอเชียทั้งหมด 45 คน ทางโฆษกกระทรวงกลาโหมออกมาแถลงข้อเท็จจริงว่าเป็นข่าวเท็จถูกจัดทำเพื่อหวังผลทางการเมือง

จากปัญหาที่เกิดจากข่าวปลอมที่ส่งผลกระทบต่อบุคคล ซึ่งอาจก่อให้เกิดการสูญเสียของบุคคลจากการรับรู้ข่าวปลอมอีกด้วย [14] [15]

การตรวจสอบความน่าเชื่อถือของสื่อสังคมออนไลน์นั้น มีความซับซ้อนมากขึ้น เนื่องจากการเติบโตของเครือข่ายอินเทอร์เน็ต ความน่าเชื่อถือของข้อมูลที่พบในข่าวส่วนใหญ่นำเสนอผ่านความสัมพันธ์ระหว่างการใช้คำและเนื้อหา นอกจากนี้ความน่าเชื่อถือของข่าวอาจต้องพิจารณารายละเอียดในองค์ประกอบที่เกี่ยวข้องของความน่าเชื่อถือของแหล่งข่าว และปัจจัยที่ใช้ในการพิจารณาความน่าเชื่อถือของเนื้อหาข่าว สามารถพิจารณาได้ในหลายมุมมอง [49] [50] [51] [52] [53]

**วัตถุประสงค์การสร้างข่าวปลอม** [54] ได้แก่

- หวังผลสนับสนุนส่งเสริมประโยชน์ของฝ่ายตนเองและทำลายฝ่ายตรงข้าม ทั้งในเรื่องส่วนตัว การฉ้อโกง ธุรกิจ การลงทุน หรือแม้แต่เรื่องการหวังผลทางการเมือง
- การทำให้คนอื่นหลงเชื่อสนองเพื่อความสนุกส่วนตัว โดยไม่ได้หวังผลชัดเจน อาจเป็นการหมิ่นประมาท

เมื่อผู้อ่านข่าวที่ขาดทักษะการรู้เท่าทันสื่อที่มีความซับซ้อน เพราะเคยชินกับการได้รับข้อมูลจากสื่อหลักที่มีการกลั่นกรอง จึงไม่สามารถทำความเข้าใจเนื้อหาข่าวปลอม หรือเข้าใจผิดไปหลังจากที่อ่านจบ จึงไม่สามารถตัดสินใจเชื่อข่าวที่อ่านได้หรือไม่ ซึ่งอีกปัญหาหนึ่งที่เกิดจากข่าวปลอมและทำให้เกิดผลเสียหายทางด้านการเมืองอย่างชัดเจน ที่มีการกล่าวถึงกันอย่างมากคือในช่วงระหว่างการหาเสียงก่อนการเลือกตั้งประธานาธิบดีสหรัฐอเมริกาเมื่อปี ค.ศ. 2016 ข่าวลือต่าง ๆ ทั้งข่าวจริงข่าวปลอมถูกปล่อยออกมาอย่างมากมาย ความขัดแย้งทางการเมืองส่งผลกระทบต่อความรู้สึก ความเชื่อมั่นของประชาชนต่อผู้ลงสมัครรับเลือกตั้งเพื่อดำรงตำแหน่งประธานาธิบดี ทำให้คะแนนเสียงมีการเปลี่ยนแปลงไป [17]

ผลเสียที่เกิดจากการส่งต่อข้อมูลการสร้างข่าวเท็จ อาจก่อให้เกิดปัญหาระหว่างบุคคลจนกลายเป็นคดีความ ทำให้บุคคลอื่นและสังคมเกิดความตื่นตระหนก ในบางครั้งอาจร้ายแรงจนส่งผลกระทบในวงกว้างต่อสังคมและเศรษฐกิจของประเทศ การเผยแพร่ออกไปเป็นวงกว้างอาจสร้างความเสียหายส่งผลกระทบต่อประเทศได้

## 2.9 การป้องกันข่าวปลอมเบื้องต้น

วิธีการตรวจสอบข่าวปลอม เมื่อได้รับข้อมูลข่าวสารบนสื่อเครือข่ายสังคมออนไลน์ไม่ควรรีบหลงเชื่อในทันที ควรพิจารณาความน่าเชื่อถือของที่มาแหล่งข่าวว่ามีความน่าเชื่อถือเพียงใด และพยายามหาข้อมูลประกอบจากแหล่งอื่น ๆ เพื่อสนับสนุนเนื้อหาข่าวดังกล่าว หากไม่พบข้อมูลอื่นที่เกี่ยวข้องให้ตั้งข้อสันนิษฐานว่าเป็นข่าวปลอม และเมื่อพบเห็นประเด็นที่น่าสงสัย ไม่ควรหลงเชื่อและส่งต่อในทันที ควรรอจนกว่าจะได้รับการยืนยันข้อเท็จจริงจากบุคคลหรือหน่วยงานที่เกี่ยวข้อง และหากมั่นใจว่าเป็นข่าวปลอม ควรรีบแจ้งเตือนผู้ใช้อื่น เพื่อช่วยลดการแพร่กระจายข่าวปลอม

ตัวอย่างเว็บไซต์ข่าวภาษาไทยที่มีการตรวจสอบข้อเท็จจริง ด้วยการไปสืบหาข้อเท็จจริงมายืนยัน หรือโดยการไปสัมภาษณ์สอบถามข้อเท็จจริงจากผู้เชี่ยวชาญแต่ละด้าน แล้วจึงนำมาเฉลยว่าข่าวเหล่านี้เป็นข่าวปลอม สามารถพิจารณาได้โดยการเข้าถึงจากหัวข้อ “ข่าวปลอม” ดังตัวอย่างเว็บไซต์ต่อไปนี้

https://twitter.com/hashtag/ข่าวปลอม

https://www.thairath.co.th/tags/ข่าวปลอม

https://mgronline.com/tags/ข่าวปลอม

https://news.kapook.com/topics/ข่าวปลอม

https://news.mthai.com/tag/ข่าวปลอม

https://www.sanook.com/news/tag/ข่าวปลอม

เครือข่ายสื่อสังคมออนไลน์เฟซบุ๊ก (Facebook) เสนอวิธีการสังเกตข่าวปลอมเบื้องต้นสำหรับผู้ใช้ทั่วไปดังนี้ [55]

– ไม่ควรหลงเชื่อหัวข้อข่าว ให้สร้างความสงสัยกับข้อความพาดหัวข่าวก่อน เนื่องจากข่าวปลอมส่วนใหญ่จะใช้ข้อความพาดหัวที่สะดุดตาเพื่อดึงดูดความสนใจผู้อ่าน โดยหากเป็นภาษาอังกฤษอาจใช้ตัวอักษรเป็นตัวพิมพ์ใหญ่ทั้งหมด หรือการเน้นข้อความด้วยตัวหนา และการใช้เครื่องหมายสัญลักษณ์ต่าง ๆ เพื่อทำให้เกิดความสะดุดตา ทำให้คนสนใจประเด็นข่าวนั้น ๆ หากพิจารณาข้อความหัวข้อข่าวพาดหัวแล้วพบว่ามีความหวือหวา น่าตื่นตระหนก ดูเกินจริงกว่าปกติที่เคยเกิดขึ้น จะทำให้เป็นเรื่องเหลือเชื่อ และไม่น่าจะเป็นไปได้ ให้สันนิษฐานข้อความข่าวนั้นก่อนว่าน่าจะเป็นข่าวปลอม

– พิจารณาสังเกตที่จุดเชื่อมต่อหรือลิงค์ (URL) อย่างละเอียด โดยส่วนใหญ่จุดเชื่อมต่อปลอมหรือหลอกลวงจะมีลักษณะคล้ายคลึงกับจุดเชื่อมต่อจริง ๆ แต่มีการปรับเปลี่ยนหรือดัดแปลงเพียงเล็กน้อยเพื่อเลียนแบบแหล่งข่าวจริง หากผู้ใช้ไม่ทันสังเกตจะหลงเชื่อคลิกตามไป ดังนั้นหากพบจุดเชื่อมต่อแปลก ๆ อาจเป็นสัญญาณเตือนของข่าวปลอมได้ โดยผู้ใช้สามารถไปที่เว็บไซต์ข่าวปลอมและเปรียบเทียบกับแหล่งข่าวที่ได้รับการยอมรับเพื่อเปรียบเทียบข้อมูลได้

– ตรวจสอบแหล่งข่าว โดยการสังเกตแหล่งที่มา และตรวจสอบเรื่องราวให้เกิดความมั่นใจว่าเป็นแหล่งข่าวที่น่าเชื่อถือและมีชื่อเสียงด้านการให้ข้อมูลที่ถูกต้อง หากเรื่องราวนั้นมาจากแหล่งข่าวที่ไม่รู้จักหรือจากองค์กรที่ชื่อไม่คุ้นเคย ควรตรวจสอบที่ส่วน “เกี่ยวกับ” ของแหล่งข่าวนั้น ๆ

– สังเกตสิ่งที่ผิดปกติ ในเว็บไซต์ข่าวปลอมส่วนใหญ่สะกดคำผิดหรือมีการจัดวางรูปแบบที่ดูไม่เป็นมืออาชีพ มีการสะกดคำผิดพลาดหรือวางรูปแบบโครงร่างที่ไม่ปกติ หากพบลักษณะเหล่านี้ควรระมัดระวังในการอ่านข่าวนั้น

– พิจารณารูปภาพประกอบ เรื่องราวของข่าวปลอมส่วนใหญ่ จะใช้รูปภาพประกอบหรือวีดิทัศน์ที่มีการบิดเบือนไปจากความจริง บางครั้งรูปภาพอาจเป็นรูปจริง แต่ไม่เกี่ยวข้องกับบริบทของเรื่องราว ผู้ใช้สามารถตรวจสอบหาแหล่งที่มา โดยการค้นหารูปภาพนั้น

– ตรวจสอบวันเวลาที่ข่าวนั้นปรากฏ เนื่องจากเรื่องราวของข่าวปลอมส่วนใหญ่อาจมีการเปลี่ยนแปลงวันเวลาที่เกิดเหตุการณ์ ทำให้ลำดับเหตุการณ์สับสนและไม่สมเหตุสมผล

– ตรวจสอบหลักฐานและแหล่งที่มาของข้อมูลเพื่อยืนยันความถูกต้อง หากไม่มีหลักฐานประกอบข่าว หรือมีการอ้างอิงโดยที่ไม่ได้ระบุชื่อผู้เชี่ยวชาญ ข่าวดังกล่าวอาจเป็นข่าวปลอม

– เปรียบเทียบข่าวกับรายงานข่าวจากแหล่งอื่น ๆ หากไม่มีแหล่งข่าวอื่นที่รายงานข่าวในเรื่องเดียวกัน อาจเป็นสัญญาณบ่งชี้ว่าเป็นข่าวปลอม ข่าวจริงควรมีการรายงานจากหลากหลายแหล่งข่าวที่น่าเชื่อถือ

– พิจารณาเนื้อหาเรื่องราวของข่าว เป็นเรื่องตลกขำขัน หรือเป็นเรื่องล้อกันเล่นหรือไม่ ในบางครั้งผู้อ่านแยกข่าวปลอมออกจากมุกตลกหรือข่าวเสียดสีหรือการล้อเลียนได้ยาก การตรวจสอบเพื่อพิจารณารายละเอียด สนใจแหล่งที่มาของเรื่องที่ล้อเลียน ตลอดจนการใช้น้ำเสียงในการเล่าเรื่องมีลักษณะในเชิงล้อกันเล่นเพื่อความสนุกสนานหรือเป็นเรื่องจริงจัง

– บางเรื่องมีเจตนาและมีความจงใจสร้างขึ้นเพื่อให้เป็นข่าวปลอม ดังนั้นผู้อ่านจึงต้องใช้วิจารณญาณในการอ่าน เพื่อคิดวิเคราะห์เรื่องราว และส่งต่อเฉพาะข่าวที่มีความมั่นใจว่าเชื่อถือได้เท่านั้น

## 2.10 คุณลักษณะเด่นของทวิตเตอร์

– ทวิตเตอร์เน้นการแสดงข้อความและมีรูปภาพแนบขยายความ

– การใช้งานของทวิตเตอร์ ผู้ใช้บางคนมีการใช้งานเสมือนเป็นการประโยคบอกเล่า ไม่มีการระบุผู้รับ เป็นลักษณะของการบอกกล่าวอย่างไม่ระบุเหตุผล การคำใช้พูดพร่ำบ่น การแสดงความคิดเห็น หรือการเขียนข้อความใด ๆ แบบไม่เป็นทางการ

– ไม่มีระบบเพื่อน กลุ่ม หรือหน้าเว็บเพจ จึงไม่มีระบบคัดกรองคนดู ข้อความที่มีการแสดงออกมาทั้งหมด จะถูกตั้งค่าการมองเห็นเป็นแบบสาธารณะทั้งหมด ดังนั้นใคร ๆ สามารถดู

ข้อความได้ทั้งหมดทุกคน ไม่ว่าจะเป็นผู้ใช้งานที่มีบัญชีผู้ใช้งานหรือไม่ และจะเป็นผู้ติดตามบัญชีนั้นหรือไม่ก็ตาม

– ระบบการติดตาม เมื่อผู้ใช้ติดตาม (follow) บัญชีที่ตนเองสนใจแล้ว บัญชีผู้ใช้ที่ต้องการติดตามได้ตั้งค่าความเป็นส่วนตัวเป็นแบบสาธารณะ (public) เมื่อผู้ใช้ได้ติดตามบัญชีนั้นแล้วจะเห็นข้อความของบัญชีนั้นในส่วนการรับข่าวสารได้ก่อนข้อความอื่น ๆ

– กรณีที่สร้างบัญชีแบบปิด ซึ่งเป็นบัญชีเฉพาะที่สามารถกำหนดให้บุคคลที่อนุญาตให้ติดตามเท่านั้น จึงจะสามารถดูข้อความที่แสดงได้ทั้งหมด

– ทุกคนมีสิทธิ์สมัครบัญชีผู้ใช้งานทวิตเตอร์ได้มากกว่า 1 บัญชี ดังนั้นแต่ละผู้ใช้จึงสามารถเปิดหลายบัญชีแยกกันสำหรับใช้งานตามวัตถุประสงค์ส่วนตัวได้

– ผู้ใช้ทวิตเตอร์เน้นการอ่านข้อความ ซึ่งอาจเป็นข้อความข่าวหรือข้อความที่มีคนกล่าวถึง หากหัวข้อข่าว (Topic) เป็นที่น่าสนใจ จะมีการใช้แฮชแท็ก (# Hashtag) ปริมาณมากจนขึ้นเป็นเทรนด์ ตัวอย่างเช่น ละคร การเชียร์กีฬา การเมือง

– การค้นหาเรื่องที่สนใจในทวิตเตอร์สามารถค้นหาผ่านแฮชแท็ก หากต้องการให้ข้อความไปอยู่ในหมวดใด สามารถเพิ่มแฮชแท็กได้ เพื่อให้คนที่สนใจในเรื่องเดียวกันสามารถค้นหาข้อความที่มีการแสดงออกไป

## 2.11 การเรียนรู้ด้วยเครื่อง

กระบวนการเรียนรู้ด้วยเครื่อง เป็นที่ยอมรับในด้านการประยุกต์ใช้งานกับแก้ไขปัญหาหลายด้าน หรือใช้สำหรับโปรแกรมประยุกต์มากมาย โดยแต่ละกระบวนการเรียนรู้มีความเหมาะสมกับข้อมูลในรูปแบบเฉพาะ กระบวนการเรียนรู้ด้วยเครื่องที่แก้ปัญหาด้วยการสอนโดยใช้ข้อมูลที่มีคำตอบที่แท้จริงของปัญหาเรียกว่าการเรียนรู้แบบได้รับการแนะนำหรือมีผู้สอน (Supervised Learning) กระบวนการเรียนรู้ด้วยเครื่องโดยมีข้อมูลที่ไม่มีคำตอบเรียกว่าเป็นวิธีการเรียนรู้แบบไม่ได้รับการแนะนำ หรือการเรียนรู้โดยไม่มีผู้สอน (Unsupervised Learning) [56]

การเรียนรู้ของเครื่องด้วยวิธีการจำแนกแบบ Naïve Bayes เป็นวิธีการหนึ่งในการจำแนกข้อมูลด้วยค่าความน่าจะเป็นอย่างง่าย โดยเป็นวิธีการที่งานวิจัยส่วนใหญ่นิยมใช้สำหรับการจำแนกข้อมูล เนื่องจากเป็นวิธีการที่เข้าใจได้ง่ายและได้ผลการจำแนกข้อมูลที่เป็นที่ยอมรับได้ โดยอาศัยหลักการพื้นฐานของการใช้กฎของ Bayes ภายใต้สมมติฐานที่ว่าข้อมูลทุกคุณลักษณะต้องเป็นอิสระจากกัน ไม่มีคุณลักษณะใดขึ้นอยู่กับคุณลักษณะอื่นที่ใช้ [26] [57] [58]

เริ่มมีการศึกษาใช้งาน Naïve Bayes กันอย่างแพร่หลายตั้งแต่ทศวรรษ 1960 ในกลุ่มงานวิจัยที่เกี่ยวข้องกับการสืบค้นข้อความในเอกสารต่าง ๆ โดยเป็นวิธีที่นิยมสำหรับการจัดหมวดหมู่ข้อความ การพิจารณาการแยกประเภทเอกสารออกจากเอกสารประเภทอื่น มีการนำไปใช้ประยุกต์กับงานด้าน

การจัดกลุ่มเอกสารตามจำนวนความถี่ของคำสำคัญที่ค้นพบในเอกสารต่าง ๆ ซึ่งจำนวนความถี่ของคำเป็นคุณสมบัติที่เหมาะสม สำหรับการนำไปใช้ในการประมวลผลข้อมูลเอกสารเบื้องต้น

ตัวจำแนก Naïve Bayes มีความยืดหยุ่นสูง สามารถนำไปใช้จำแนกข้อมูลจำนวนมาก ๆ ได้ โดยการใช้จำนวนพารามิเตอร์หลายตัวกับตัวแปรที่เป็นคุณลักษณะหรือตัวพยากรณ์ ในปัญหาการเรียนรู้ การฝึกสอนที่ให้ค่าความน่าจะเป็นสูงสุดสามารถทำได้โดยการประเมินแบบ Closed-form expression ที่ใช้เวลามากกว่าการประเมินด้วย Iterative approximation ที่ใช้สำหรับการจำแนกประเภทอื่น ๆ

Naïve Bayes เป็นขั้นตอนวิธีการจำแนกประเภทตามคุณลักษณะ โดยสมมติว่ามีการแจกแจงของข้อมูลที่เป็นอิสระต่อกัน ความน่าจะเป็นก่อนหน้าสำหรับแต่ละคลาสคำตอบจะถูกคำนวณจากข้อมูลการฝึกสอน และกำหนดสมมติฐานว่าเป็นอิสระจากกัน เพื่อหาค่าสมมติฐานที่น่าจะถูกต้องที่สุด

Naïve Bayes เป็นเทคนิคสำหรับการสร้างตัวแยกประเภท แบบจำลองที่กำหนดคำตอบของคลาสให้ปัญหา แสดงด้วยค่าคุณลักษณะซึ่งมีการดึงคำตอบของคลาสจากชุด จำกัด บางชุด ค่าของคุณสมบัติเฉพาะนั้นไม่ขึ้นอยู่กับค่าของคุณสมบัติอื่นใด ๆ ยกตัวอย่างเช่น ผลไม้ผลหนึ่งจะถูกพิจารณาว่าเป็นส้ม ก็ต่อเมื่อเป็นผลที่มีสีส้ม สีเหลือง หรือสีเขียว มีลักษณะทรงค่อนข้างกลม และมีเส้นผ่านศูนย์กลางประมาณ 10 เซนติเมตร เมื่อพิจารณาคุณสมบัติแต่ละตัว เพื่อสนับสนุนความน่าจะเป็นที่จะจำแนกให้ได้ว่าผลไม้ผลนี้จะเป็นส้มหรือไม่ Naïve Bayes จะไม่คำนึงถึงความสัมพันธ์ที่เป็นไปได้ระหว่างคุณลักษณะของสี ลักษณะทรงวัตถุ หรือขนาดเส้นผ่านศูนย์กลางวัตถุ

สำหรับแบบจำลองที่สร้างจาก Naïve Bayes สามารถได้รับฝึกสอนให้เกิดประสิทธิภาพสูงมากในสภาพแวดล้อมการเรียนรู้แบบมีผู้สอน การใช้งานจริงหลายอย่างใช้การประมาณค่าพารามิเตอร์สำหรับแบบจำลอง Naïve Bayes โดยใช้วิธีการที่เลือกค่าความน่าจะเป็นสูงสุด

ข้อได้เปรียบของ Naïve Bayes คือต้องการข้อมูลการฝึกสอนจำนวนไม่มากเพื่อการประมาณค่าพารามิเตอร์ที่จำเป็นสำหรับการจำแนกประเภท

Naïve Bayes เป็นแบบจำลองความน่าจะเป็นแบบมีเงื่อนไข สำหรับการแก้ปัญหาการจำแนกกลุ่มข้อมูล โดยอาศัยความรู้ก่อนหน้า (Prior knowledge) ซึ่งคือความน่าจะเป็นก่อนหน้าของสมมติฐานหนึ่งร่วมกับข้อมูลที่เกิดขึ้น ปัญหาการจำแนกกลุ่มข้อมูลที่ใช้กฎของ Bayes ในการอธิบายสามารถคำนวณค่าความน่าจะเป็นของการจำแนกคลาสเมื่อทราบแต่ละข้อมูล ได้จากสมการที่ (1)

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (1)$$

โดยที่

$k$ แทนด้วยค่าที่เป็นไปได้ในคลาสคำตอบ $C_k$

$x = (x_1, \ldots, x_n)$ แทนค่าเวกเตอร์ข้อมูลที่ใช้คำนวณแจกแจงความน่าจะเป็น ซึ่ง $n$ เป็นจำนวนคุณลักษณะของข้อมูลทั้งหมดที่แต่ละตัวมีความเป็นอิสระไม่ขึ้นต่อกัน

$p(C_k|x)$ แทนค่าความน่าจะเป็นที่จะเกิดคลาสคำตอบ $C_k$ เมื่อทราบค่า $x$

$p(C_k)$ แทนค่าความน่าจะเป็นก่อนหน้าที่จำแนกได้คลาสคำตอบ $C_k$

$p(x|C_k)$ แทนค่าความน่าจะเป็นที่จะเกิด $x$ เมื่อทราบค่าคลาสคำตอบ $C_k$

$p(x)$ แทนค่าความน่าจะเป็นก่อนหน้าที่จะเกิด $x$

การสร้างตัวจำแนกข้อมูลจากแบบจำลองความน่าจะเป็น ซึ่งรูปแบบคุณลักษณะที่เป็นอิสระต่อกัน ทำให้แบบจำลองความน่าจะเป็นของ Naïve Bayes เป็นการรวมกฎการตัดสินใจ โดยการเลือกค่าสมมติฐานที่เป็นไปได้มากที่สุด และฟังก์ชันที่กำหนดคลาสคือ $\hat{y} = C_k$ สำหรับค่า $k$ บางตัวตามสมการที่ (2)

$$\hat{y} = argmax_{k \in \{1, \ldots, k\}} p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \quad (2)$$

Neural Network เป็นรูปแบบการเรียนรู้ของเครื่องที่เลียนแบบจากลักษณะทั่วไปทางชีววิทยาของโครงสร้างสมอง เพื่อจำลองการทำงานของสมองมนุษย์ ประกอบด้วยหน่วยประมวลผลย่อย (Neuron) จำนวนมากที่ทำงานร่วมกันโดยส่งสัญญาณผ่านตัวเชื่อมที่เรียกว่า (Connection link) ในแต่ละเส้นเชื่อมจะมีค่าความสำคัญต่างกันจากค่าน้ำหนักประจำเส้น (Weight) และแต่ละหน่วยประมวลผลย่อยหรือโหนด (Node) จะส่งค่าสัญญาณเมื่อได้รับค่าจากฟังก์ชันการกระตุ้น (Activation function) ไปยังหน่วยประมวลผลย่อยอื่น โดยโครงสร้างระบบเป็นการทำงานแบบขนาน ในการทำงานเครือข่ายประสาทเทียมจะแบ่งการทำงานเป็นสามชั้นหลัก โดยรับข้อมูลเข้าเป็นชั้นแรก (Input layer) และประมวลผลเป็น หลายชั้นเพื่อทดสอบหาค่าลักษณะสำคัญจากเส้นทางภายในโครงข่ายในเรียกว่าชั้นซ่อน (Hidden layer) จนกระทั่งได้เป็นคำตอบที่ดีที่สุดอยู่ในชั้นผลลัพธ์ (Output layer) ซึ่งขั้นตอนการเรียนรู้ มีการใช้วิธีการแพร่ย้อนกลับ (Back propagation) เพื่อเลือกค่าลักษณะที่เหมาะสมที่สุดในเส้นทางต่าง ๆ ในเครือข่าย ทำให้สามารถใช้แยกข้อมูลได้หลายประเภท มีความแม่นยำสูงในการแยกแยะข้อมูล เหมาะกับข้อมูลหลายประเภททั้งแบบต่อเนื่องและไม่ต่อเนื่อง และมีความทนทานต่อสัญญาณรบกวนได้ดี แต่มีข้อจำกัดในการอธิบายความหมายของค่าน้ำหนักที่เกิดขึ้นได้ยาก เสียเวลาในการเรียนรู้ข้อมูลค่อนข้างนานเพื่อการสร้างแบบจำลองที่ดี [59]

หลักสำคัญในการกระตุ้นความสนใจในเครือข่ายประสาทเทียมและการเรียนรู้คือขั้นตอนวิธีการแพร่กระจายย้อนกลับของ Werbos (1975) ที่ทำให้เกิดการฝึกสอนเครือข่ายที่เป็นไปได้แบบหลายชั้นและมีประสิทธิภาพ การกระจายแบบ Backpropagation ปรับแก้ค่าน้ำหนักของแต่ละโหนดเพื่อลดค่าความผิดพลาดของแต่ละชั้นในการฝึกสอนแบบจำลอง ต่อมาในช่วงกลางทศวรรษ 1980 การ

ประมวลผลแบบกระจายขนานเป็นที่ได้รับความนิยมจากแนวคิดของ Rumelhart และ McClelland (1986) จึงได้อธิบายวิธีการใช้หลักการนี้ เพื่อเชื่อมต่อการจำลองกระบวนการทำงานของระบบประสาท

Vanishing gradient ถูกใช้เป็นปัญหาในการอธิบายการใช้งานเครือข่ายแบบหลายชั้น โดยการทำ Feedforward ที่ใช้ Backpropagation ซึ่งเป็นที่รู้จักกันในชื่อ Recurrent Neural Network (RNNs) เมื่อเกิดค่าความผิดพลาดแพร่กระจายจากชั้นหนึ่งไปอีกชั้นหนึ่ง จะมีส่งผลให้มีการปรับค่าน้ำหนักของแต่ละโหนดที่จำนวนการปรับค่าขึ้นอยู่กับค่าผิดพลาดที่เกิดขึ้น และทำให้เกิดการใช้งานโครงข่ายหลายชั้นที่มีความลึกมากขึ้น สำหรับการแก้ปัญหานี้ Schmidhuber ได้นำวิธีการใช้ลำดับชั้นของเครือข่ายหลายระดับ (1992) ที่ได้รับการฝึกสอนมาก่อนระดับหนึ่งทีละครั้ง สำหรับการฝึกสอนที่ไม่มีคำตอบล่วงหน้า Backpropagation ถูกปรับแต่งโดย Behnke (2003) อาศัยเครื่องหมายของ gradient (Rprop) ในการใช้แก้ปัญหา

แนวความคิดของ Support Vector Machine (SVM) เกิดจากการที่นำค่าของกลุ่มข้อมูลมาวางลงในฟีเจอร์สเปซ (Feature Space) จากนั้นจึงหาเส้นที่ใช้แบ่งข้อมูลทั้งสองออกจากกัน โดยจะสร้างเส้นตรงที่ใช้แบ่ง (Hyperplane) ขึ้นมา และเพื่อให้ทราบว่าเส้นตรงที่แบ่งสองกลุ่มออกจากกันโดยเลือกเส้นตรงที่ดีที่สุดในการแบ่งกลุ่ม การจำแนกข้อมูลบนระนาบหลายมิติ จะใช้ส่วนการเลือกที่มีความเหมาะสมที่สุดเรียกว่า โครงสร้างในการคัดเลือก (Feature selection) ซึ่งโครงสร้างในการคัดเลือกมาจากข้อมูลที่สอนให้ระบบเรียนรู้ จำนวนกลุ่มของโครงสร้างที่ใช้อธิบาย เรียกว่า เวกเตอร์ (Vector) โดยตัวแบบ SVM ต้องการแบ่งแยกกลุ่มของเวกเตอร์ด้วยหนึ่งกลุ่มของตัวแปรเป้าหมายที่อยู่ข้างหนึ่งของระนาบ และกรณีอื่นที่อยู่ทางระนาบต่างกัน ซึ่งเวกเตอร์ที่อยู่ข้างระนาบหลายมิติทั้งหมดเรียกว่า Support Vectors

ขั้นตอนวิธีในการจำแนกโดยหลักการทำงานคือ การให้ข้อมูลนำเข้าที่ใช้ฝึกสอนเป็นเวกเตอร์ใน N มิติ เช่นถ้าในกรณีของ 2 มิติ และ 3 มิติ จะเป็นจุดที่อยู่ในพิกัด xy และ xyz ตามลำดับ จากนั้นจึงสร้าง Hyperplane ที่ใช้แยกกลุ่มของเวกเตอร์ข้อมูลนำเข้าออกเป็นประเภทต่าง ๆ ในกรณีที่ข้อมูลเป็น 2 มิติ Hyperplane เป็นเส้นตรง และหากข้อมูลเป็น 3 มิติ Hyperplane จะเป็นลักษณะของระนาบ อาศัยหลักการของการหาสัมประสิทธิ์ของสมการ เพื่อสร้างเส้นแบ่งแยกกลุ่มข้อมูลที่ถูกป้อนเข้าสู่กระบวนการสอนให้ระบบเรียนรู้ โดยเน้นไปยังเส้นแบ่งแยกกลุ่มข้อมูล (Margin) ได้ดีที่สุด เป็นการเรียนรู้ของเครื่องเพื่อแบ่งตัวอย่างสอนในสองกลุ่ม โดยการสร้างระนาบหลายมิติแบ่งแยกแบบดีที่สุด (Optimal Separating Hyperplane) แบ่งแต่ละกลุ่มให้อยู่บนแต่ละด้านของระนาบหลายมิติ จุดเด่นของ SVM คือสามารถจับคู่ (Map) เวกเตอร์ในมิติข้อมูลนำเข้า ให้กลายเป็น Feature Space โดยใช้ฟังก์ชันที่เรียกว่า kernel กรณีที่ข้อมูลนำเข้าเป็นแบบ 2 มิติ สามารถจำแนกข้อมูลนำเข้าเป็น 2 กลุ่ม โดยใช้ Hyperplane เส้นตรง โดยทั่วไปมีเส้นตรงจำนวนมากที่สามารถแบ่งกลุ่มข้อมูลออกเป็น

สองส่วนได้ แต่เส้นตรงเส้นใดสามารถใช้แบ่งกลุ่มได้ดีที่สุด (Optimal Line) จะพิจารณาจาก ที่เป็นผลรวมระยะห่างของเส้นตรงที่เป็น Hyperplane ถึงเส้นตรงที่ผ่านข้อมูลนำเข้าที่อยู่ใกล้เส้นมากสุดและขนานกับ Hyperplane ของทั้งสองกลุ่ม ระยะที่กล่าวถึงนี้สามารถมองเป็นเวกเตอร์ได้ และเรียกว่า Support Vector Machine โดยขั้นตอนวิธี SVM จะเลือก Hyperplane ที่ให้ค่าเส้นแบ่งแยกกลุ่มข้อมูล และมีระยะห่างจากข้อมูลสอนมายังระนาบหลายมิติที่กว้างที่สุดหรือมีค่าสูงสุด

SVM เป็นวิธีการเรียนรู้ของเครื่องที่ได้รับความนิยม ให้ความแม่นยำสูงสำหรับการจำแนกข้อมูล เป็นวิธีการเรียนรู้แบบมีผู้สอนที่ใช้สำหรับการวิเคราะห์การจำแนกประเภทข้อมูล [60] สามารถใช้ช่วยแก้ปัญหาการจำแนกข้อมูล ใช้ในการวิเคราะห์ข้อมูลและจำแนกข้อมูล [61]

การนำแบบจำลอง (Model) ไปใช้งานจำเป็นต้องมีการวัดประสิทธิภาพของแบบจำลองก่อน โดยทั่วไปนิยมใช้ตาราง Confusion Matrix ในการวัดประสิทธิภาพแบบจำลองสามารถ ตารางนี้อยู่ในรูปเมตริกซ์จัตุรัส โดยจะมีจำนวนแถว เท่ากับ จำนวนคอลัมน์มีค่าเท่ากับจำนวนคลาสคำตอบ ยกตัวอย่างเช่น หากคำตอบมีเพียง 2 ค่า คือ จริง และ เท็จ จะได้ตารางเมตริกซ์ขนาด 2x2 โดยข้อมูลในด้านคอลัมน์คือ ค่าคำตอบที่ปรากฏอยู่ในชุดข้อมูล และข้อมูลในแนวแถวเป็นค่าที่แบบจำลองทำนายได้ อธิบายได้ดังรูปที่ 2.1

| ค่าที่ทำนาย/ค่าคำตอบ | จริง | เท็จ |
|---|---|---|
| จริง | True Positive (TP) | False Positive (FP) |
| เท็จ | False Negative (FN) | True Negative (TN) |

**รูปที่ 2.1** ตัวอย่าง Confusion Matrix ขนาด 2×2

โดยที่

True Positive (TP) คือจำนวนข้อมูลที่ทำนายว่าเป็นจริงและคำตอบเป็นจริง

True Negative (TN) คือจำนวนข้อมูลที่ทำนายถูกต้องว่าเป็นเท็จจากคำตอบที่เป็นเท็จ

False Positive (FP) คือจำนวนข้อมูลที่ทำนายผิดว่าเป็นจริงแต่คำตอบจริงเป็นเท็จ

False Negative (FN) คือจำนวนข้อมูลที่ทำนายผิดว่าเป็นเท็จแต่คำตอบที่ถูกต้องเป็นจริง

Accuracy เป็นการวัดความถูกต้องของแบบจำลองที่ตอบได้ถูกต้องจากข้อมูลทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (3)

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

Precision เป็นการวัดความแม่นยำของข้อมูลที่ทำนายว่ามีความถูกต้องตามจริงหรือไม่ เป็นค่าที่บอกว่าการทำนายว่าจริงมีความถูกต้องเพียงใด โดยสามารถคำนวณได้จากสมการที่ (4)

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

Recall หรือ True Positive Rate (TPR) เป็นการวัดความถูกต้องของแบบจำลองที่ตอบได้ถูกต้องเป็นอัตราส่วนจากข้อมูลที่จริงทั้งหมดเท่าใด โดยสามารถคำนวณได้จากสมการที่ (5)

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

F-measure เป็นการวัดค่าประสิทธิภาพโดยรวมของแบบจำลองโดยคำนวณจากเฉลี่ยระหว่าง Precision และ Recall สามารถคำนวณได้จากสมการที่ (6)

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

True Negative Rate (TNR) คือ ค่าที่บอกว่าทำนายว่าเท็จ เป็นอัตราส่วนเท่าใดจากคำตอบที่เป็นเท็จทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (7)

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

False Positive Rate (FPR) คือ ค่าที่บอกว่าทำนายว่าจริง เป็นอัตราส่วนเท่าใดจากคำตอบที่เป็นเท็จทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (8)

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

False Negative Rate (FNR) คือ ค่าที่บอกว่าโปรแกรมทำนายว่าไม่จริง เป็นอัตราส่วนเท่าใดจากคำตอบจริงทั้งหมด โดยสามารถคำนวณได้จากสมการที่ (9)

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

ก่อนที่จะนำผลลัพธ์ที่ได้ไปใช้งานการสร้างแบบจำลองด้วยเทคนิค Classification มีการทดสอบประสิทธิภาพของแบบจำลอง 3 แบบ ได้แก่

(1) วิธี Self-Consistency Test หรือ Use Training Set เป็นวิธีใช้ชุดข้อมูลเดียวกันในการสร้างแบบจำลองและทดสอบแบบจำลอง โดยจากสร้างแบบจำลองด้วยชุดข้อมูลที่มีการฝึกสอน (Training data) หลังจากนั้นนำแบบจำลองที่สร้างได้มาทำนายผลด้วยชุดข้อมูลเดิม การวัดประสิทธิภาพแบบจำลองวิธีนี้ ให้ผลการวัดประสิทธิภาพที่มีค่าสูง เนื่องจากเป็นชุดข้อมูลที่นำมาทดสอบเป็นข้อมูลเดียวกับแบบจำลองที่ได้มีการเรียนรู้แล้ว ดังนั้นวิธีการนี้เหมาะสำหรับการใช้ดูความเหมาะสมของแบบจำลองที่สร้างขึ้นกับชุดข้อมูล

(2) วิธี Split Test เป็นการแบ่งชุดข้อมูลออกเป็น 2 ส่วน โดยการสุ่มข้อมูล ตัวอย่างเช่นอาจแบ่งชุดข้อมูลออกเป็นร้อยละ 50 ต่อร้อยละ 50 หรือร้อยละ 60 ต่อร้อยละ 40 หรืออื่น ๆ โดยการแบ่งชุดข้อมูลส่วนแรกสำหรับใช้ในการสร้างแบบจำลอง และชุดข้อมูลในส่วนที่เหลือเก็บไว้ใช้ในการทดสอบประสิทธิภาพการทำงานของแบบจำลอง วิธีการนี้ควรทำซ้ำหลาย ๆ ครั้งเพื่อให้เกิดความมั่นใจในการสุ่มข้อมูลที่แบ่งส่วน เหมาะกับการสร้างแบบจำลองที่มีชุดข้อมูลขนาดใหญ่มาก ๆ

(3) วิธี Cross-validation Test แบ่งข้อมูลออกเป็นหลายส่วนโดยกำหนดเป็นค่า k ส่วน ในการแบ่งชุดข้อมูลออกเป็น k ส่วนที่มีขนาดเท่า ๆ กัน แล้วแบ่งหนึ่งส่วนไว้ใช้ในการทดสอบประสิทธิภาพของแบบจำลอง และชุดข้อมูลส่วนที่เหลือทั้งหมดจะถูกนำไปใช้สร้างแบบจำลองแต่ละรอบ โดยจะทำซ้ำวนไปจนกระทั่งครบตามจำนวน k ที่ระบุ

## 2.12 งานวิจัยที่เกี่ยวข้อง

วิวัฒนาการของสื่อในช่วงศตวรรษที่ 19 สิ่งพิมพ์ในรูปแบบกระดาษยังมีราคาถูก จึงทำให้ผู้คนสามารถเข้าถึงหนังสือพิมพ์ได้ง่าย [17] การเผยแพร่ข่าวสารปริมาณมาก ทำให้เกิดความกังวลว่าสื่อมวลชนจะมีส่วนสำคัญในการชักจูงแนวความคิดของประชาชนให้โน้มเอียงไปตามที่ต้องการได้มากกว่าผู้นำประเทศ ต่อมามีการเปลี่ยนรูปแบบการสื่อสารไปสู่สื่อสังคมออนไลน์มีโครงสร้างแตกต่างกันอย่างมากกับเทคโนโลยีสื่อก่อนหน้า การเจริญเติบโตของข่าวออนไลน์ทำให้เกิดความหวั่นวิตกกังวล เกี่ยวกับความคิดเห็นที่เกิดขึ้นมีความหลากหลายมากเกินไป จะทำให้ประชาชนสามารถสร้างเสียงสะท้อน ข้อเรียกร้องต่าง ๆ มากขึ้น อีกทั้งยังเกิดความกังวลเกี่ยวกับแพลตฟอร์มใหม่ที่เกิดขึ้นอาจทำให้การทำงานของสื่อมวลชนถูกตรวจสอบมากยิ่งขึ้น เนื้อหาข่าวสารต่าง ๆ สามารถถ่ายทอดให้กับผู้ใช้ได้โดยตรง มีการเผยแพร่กระจายข่าวกันได้โดยไม่ต้องมีการกลั่นกรองหรือการตรวจสอบความจริงหรือการตีความของสื่อหรือบุคคลใด ๆ และผู้ใช้ยังมีทางเลือกในการเข้าถึงสื่อมากขึ้นสามารถเข้าถึงสื่อได้หลายราย และสามารถเปรียบเทียบกันได้อีกด้วย

ข่าวสารที่มีการเผยแพร่กันบนสื่อสังคมออนไลน์มีทั้งข่าวจริงที่ได้ผ่านการตรวจสอบก่อนการนำเสนอ และข่าวปลอมที่ถูกนำเสนอด้วยความรวดเร็ว ในงาน [62] กล่าวว่าการประชาสัมพันธ์ที่เกี่ยวกับข่าวจริงสามารถแพร่กระจายไปทั่วโลกอย่างรวดเร็ว แต่ข่าวที่เกี่ยวกับข่าวปลอมมีการปรับเปลี่ยนเนื้อหาสามารถแพร่กระจายออกไปได้มากกว่า ในการศึกษาการแพร่กระจายข่าวที่เป็นจริงและข่าวปลอมบนเครือข่ายสังคมออนไลน์ใน [6] แสดงให้เห็นความขัดแย้งกันที่สามารถนำมาจำลองการเผยแพร่ข่าวสารของผู้เข้าร่วมงานวิจัย ทำให้เห็นว่าประชาชนสามารถมีส่วนร่วมกับงานวิจัยโดยสมัครรับข่าวสาร นอกจากนั้น [29] ได้วิเคราะห์หาปัจจัยที่เกี่ยวข้องกับข้อความที่ส่งผลต่อทัศนคติด้านลบและความเชื่อที่ไม่ถูกต้องตามข้อมูลที่ได้รับ [3] ได้เสนอคำแนะนำเกี่ยวกับวิธีจัดการกับความเสี่ยงที่เกิดขึ้นได้อย่างมีประสิทธิภาพ ตลอดจนการปรับปรุงวิธีการรับรู้ข่าวสาร ความเข้าใจ การ

แก้ปัญหาในระยะยาวของปัญหาของข่าวปลอม ในงาน [2] ได้สำรวจบทความวิชาการต่าง ๆ เพื่อช่วยในการทำงานวิจัยที่เกี่ยวกับปัญหาข่าวปลอม มีการนำเสนอการตรวจสอบข่าวปลอมที่เกี่ยวกับสื่อสังคมออนไลน์ รวมทั้งการปลอมตัวของข่าวปลอมในเชิงจิตวิทยาและทฤษฎีทางสังคม อัลกอริทึมทางด้านการทำเหมืองข้อมูลที่มีการใช้งาน ชุดข้อมูลตัวแทน โดย [2] [63] ได้เสนอมาตรวัดเพื่อจัดระเบียบประเภทข่าวปลอม และแสดงให้เห็นถึงความสัมพันธ์ที่แตกต่างกันของแต่ละวิธีการ คุณสมบัติของข่าวปลอมที่แตกต่างกับข่าวจริงคือผู้เขียนตั้งใจจะหลอกลวงผู้อ่าน ปัจจัยแรงจูงใจด้านการเงินทำให้เกิดการสร้างข่าวปลอมขึ้นมาเพื่อหลอกให้ผู้อ่านติดตามรายละเอียด

งานวิจัย [62] วิเคราะห์โครงสร้างวิวัฒนาการของการเผยแพร่ข่าวปลอมบนสื่อสังคมออนไลน์ โดยงาน [19] พบว่าข่าวเท็จเป็นข่าวแปลกใหม่กว่าข่าวจริง แสดงให้เห็นว่าผู้คนมีแนวโน้มที่จะแบ่งปันข้อมูลใหม่ ๆ แม้ว่าเรื่องราวนั้นจะเป็นเรื่องไม่จริงก็ตาม โดยงาน [62] พบว่าข้อความที่เกี่ยวกับข่าวจริงเกิดการเผยแพร่ในแนวกว้างและมีความลึกที่สั้นกว่าข้อความเกี่ยวกับข่าวปลอม และในงาน [64] แสดงให้เห็นว่าข่าวลือที่ได้รับการพิสูจน์แล้วว่าจริง มีแนวโน้มที่จะได้รับการแก้ไขอย่างรวดเร็วกว่าที่เป็นจริง นอกจากนี้ [19] [29] [62] พบว่าผู้คนส่วนใหญ่จะให้ความสนใจในการตอบกลับเรื่องราวใหม่มากกว่าเรื่องจริงที่ถูกเปิดเผยในภายหลัง ซึ่งงาน [64] ได้เสนอวิธีการที่ทำให้สามารถรวบรวม จำแนก และใส่คำอธิบายประกอบชุดข้อมูลข่าว เพื่อทำความเข้าใจว่าผู้ใช้แพร่กระจายสนับสนุนหรือปฏิเสธข่าวลือที่ได้รับการพิสูจน์แล้ว และวิเคราะห์บทบาทของผู้ใช้ที่แตกต่างกัน

การแบ่งประเภทข่าวปลอมในงาน [27] ได้ใช้การพิจารณาจากน้ำหนักข้อดีข้อเสียจากการใช้คลังข้อมูลคำ สำหรับการวิเคราะห์ข้อความและการทำนายแบบคาดเดา การกรองข้อมูล การตรวจสอบและการยืนยันข้อมูลออนไลน์ยังเป็นสิ่งสำคัญ ในงาน [24] มีวิเคราะห์ข้อมูลข่าวปลอมที่ขึ้นอยู่กับความหมายของข้อความที่ส่งออกไป เช่น การวิเคราะห์ความหมายของคำ การวิเคราะห์ความเชื่อมั่นทางอารมณ์ในการทำเหมืองข้อมูล [21] ได้ทดสอบประสิทธิภาพในการตรวจสอบความน่าเชื่อถือของการมีอารมณ์ร่วมจากการตรวจสอบวีดิทัศน์ และสำรวจรูปแบบข้อมูลที่มีการพิมพ์ข้อความเปรียบเทียบกับข้อมูลรูปแบบวิดีโอ และเปรียบเทียบอารมณ์ที่ได้จากการรับรู้ข้อมูลแบบมีอารมณ์ขันและไม่มีอารมณ์ขัน และ [24] ให้ความสำคัญของคุณลักษณะที่ใช้ การลดคุณลักษณะที่เน้นถึงคุณลักษณะที่แสดงถึงความน่าเชื่อถือที่ดีที่สุด ส่วน [65] เสนอระบบการวิเคราะห์ความน่าเชื่อถือใหม่สำหรับการประเมินความน่าเชื่อถือของข้อมูลบนทวิตเตอร์ องค์ประกอบที่อ้างอิงจากความมีชื่อเสียง เครื่องมือแบ่งประเภทความน่าเชื่อถือ องค์ประกอบด้านประสบการณ์การใช้งานของผู้ใช้ และอัลกอริทึมสำหรับการจัดอันดับคุณลักษณะ นอกจากนี้มีการพัฒนาวิธีการตรวจหาข่าวปลอมบนทวิตเตอร์แบบอัตโนมัติ โดยการเรียนรู้จากการคาดการณ์การประเมินความถูกต้องในชุดข้อมูลทวิตเตอร์ที่น่าเชื่อถือสองชุด การวิเคราะห์คุณลักษณะจะระบุถึงคุณลักษณะที่คาดการณ์ได้มากที่สุดสำหรับการประเมินความถูกต้องที่เข้มงวดของนักข่าว [45] [66]

การวิเคราะห์ความน่าเชื่อถือของข้อมูลข่าวสารที่แพร่กระจายผ่านทางทวิตเตอร์ที่เป็นบริการที่ได้รับความนิยมมาก [6] [39] [65] [67] ในงานวิจัย [10] พบว่าข้อความส่วนใหญ่ที่โพสต์ไว้ที่ทวิตเตอร์เป็นจริง แต่ยังมีการแพร่กระจายข้อมูลที่ผิดพลาดและข่าวลือที่เป็นเท็จ โดยส่วนใหญ่มาจากความไม่ได้ตั้งใจ งานวิจัยนี้ให้ความสำคัญกับวิธีการการประเมินความน่าเชื่อถือของข้อมูลที่มีการส่งในทวิตเตอร์แบบอัตโนมัติ โดยเฉพาะการวิเคราะห์โพสต์ที่เกี่ยวข้องกับหัวข้อที่อยู่ในกลุ่ม "Trending" และแยกกลุ่มข้อความที่น่าเชื่อถือหรือไม่น่าเชื่อถือ โดยคุณลักษณะจากการสกัดข้อมูล ในงาน [20] ใช้แพลตฟอร์มในสื่อสังคมออนไลน์เพื่อให้เกิดโอกาสในการศึกษารูปแบบการแบ่งปันข้อมูลร่วมกันของผู้ใช้ การปรึกษาหารือเรื่องเกี่ยวกับข่าวลือ และวิธีการประเมินความถูกต้องโดยอัตโนมัติโดยใช้การประมวลผลภาษาธรรมชาติและเทคนิคการทำเหมืองข้อมูล เป้าหมายในการพัฒนาระบบ และประเภทข่าวลือที่ประกอบด้วย 4 ส่วน ได้แก่ การตรวจจับข่าวลือ การติดตามข่าวลือ การแบ่งประเภท การจัดประเภทความถูกต้องของข่าวลือ

งานวิจัย [27] [66] [68] [69] ต่างมุ่งเน้นวิธีการประมวลผลกับข้อความโดยใช้กระบวนการวิธีการเรียนรู้ด้วยเครื่องในการวิเคราะห์ข่าวปลอมแบบอัตโนมัติ

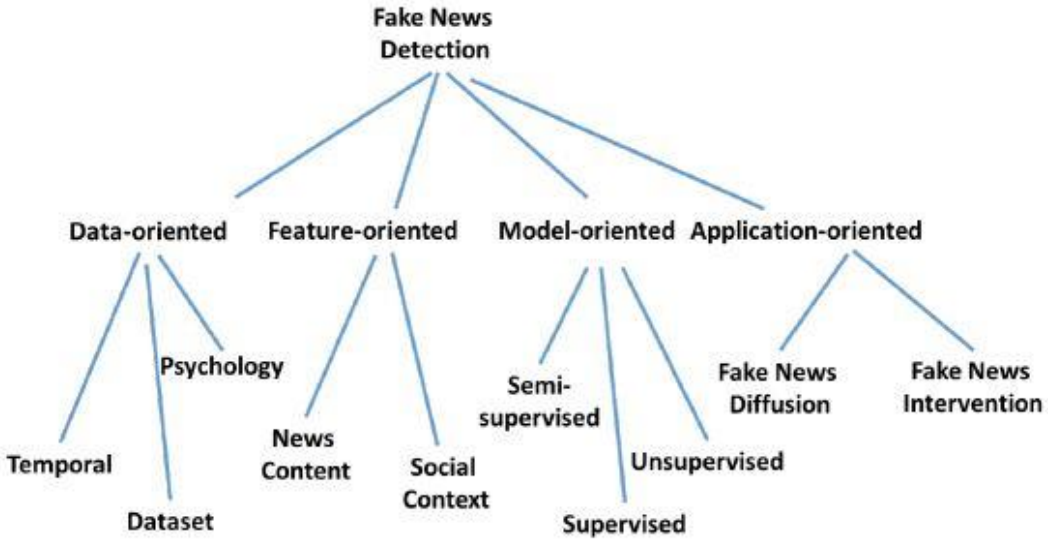
งานวิจัย [26] แสดงวิธีการตรวจสอบข่าวปลอมอย่างง่ายด้วยการใช้ตัวแบ่งประเภท Naïve Bayes โดยใช้เป็นระบบทดสอบกับชุดข้อมูลจากเฟซบุ๊กมีความแม่นยำในการจัดหมวดหมู่ประมาณ 74% งาน [23] ใช้การจำแนกด้วย Naïve Bayes ในการวิเคราะห์ความน่าเชื่อถือในกรอบงานใหม่แบบหลายขั้นตอน เพื่อระบุเนื้อหาที่ไม่น่าเชื่อถือในทวิตเตอร์ และมีการปรับปรุงความสำคัญสัมพัทธ์ของคุณลักษณะที่ใช้เพื่อปรับปรุงความถูกต้องในการจัดหมวดหมู่ ผลการวิจัย [23] พบว่าสามารถจำแนกความน่าเชื่อถือได้ด้วยความถูกต้อง 90.3% ความแม่นยำ 98.24% และการเรียกคืน 98.8%

SAMAR [28] เป็นระบบการเรียนรู้ด้วยเครื่องแบบ Supervised learning ในการวิเคราะห์เนื้อหาและความรู้สึก สำหรับสื่อสังคมออนไลน์ภาษาอาหรับ ในบริบทนี้จะกล่าวถึงปัญหาที่เกี่ยวข้องดังต่อไปนี้ การแสดงข้อมูลเกี่ยวกับคำศัพท์ที่ดีที่สุด คุณสมบัติมาตรฐานที่ใช้สำหรับภาษาอังกฤษมีประโยชน์สอดคล้องกับภาษาอาหรับหรือไม่ วิธีจัดการภาษาอาหรับสำหรับคุณลักษณะเฉพาะประเภทที่ส่งผลกระทบต่อประสิทธิภาพการทำงาน [25] งาน [27] ได้เสนอความต้องการของระบบคลังข้อมูลคำศัพท์สำหรับการตรวจสอบข่าวปลอม และข้อเสนอแนะในการจัดการข่าวปลอมแต่ละประเภท และยังมีงานวิจัยที่วิเคราะห์ข่าวปลอมในเนื้อหาข่าวอีกหลายภาษา เช่น ภาษาจีน ภาษาอินโดนีเซีย [6] [39] [45] [57] [65] [67] [65]

ตัวอย่างการวิเคราะห์สื่อสังคมออนไลน์ที่มีผลต่อระบอบประชาธิปไตยในประเทศฝั่งตะวันตก งาน [3] และข่าวการเมืองในประเทศสหรัฐอเมริกา [17] ได้รับการตีแผ่ซ้ำ ๆ เนื่องจากการเปลี่ยนแปลงเทคโนโลยีของสื่อทำให้คนรับข่าวสารได้ง่ายขึ้น จึงส่งผลให้ประชาชนเกิดความสนใจใน

เรื่องการเมืองมากขึ้น งาน [62] ได้ตรวจสอบปัญหาข่าวปลอมด้านบริบทเนื้อความข่าว ในการเลือกตั้งประธานาธิบดีของประเทศสหรัฐอเมริกาในปี ค.ศ. 2016

จากการศึกษาตัวอย่างงานวิจัยที่มีความเกี่ยวข้องกับการตรวจสอบข่าวปลอมบนสื่อสังคมออนไลน์ แบ่งออกเป็น 4 ด้าน ได้แก่ Data-oriented, Feature-oriented, Model-oriented และ Application-oriented ดังรายละเอียดในรูปที่ 2.2 [2]




**รูปที่ 2.2** การแบ่งประเภทงานวิจัยที่เกี่ยวข้องกับการตรวจจับข่าวปลอม

จากรูปที่ 2.2 แสดงการแบ่งประเภทงานวิจัยที่เกี่ยวข้องกับการตรวจจับข่าวปลอมมีรายละเอียดของงานวิจัยที่เกี่ยวข้องแต่ละด้านดังนี้

– Data-oriented ให้ความสนใจกับความแตกต่างของลักษณะของข้อมูล ทั้งในส่วนของชุดข้อมูล (Dataset) ที่ปัจจุบันยังไม่มีชุดข้อมูลที่เป็นมาตรฐานจริง ๆ สำหรับใช้ในการวิเคราะห์ข่าวปลอม ส่วนในด้านของมุมมองด้านเวลา พิจารณาที่ช่วงระยะเวลาของเนื้อหาข่าวปลอมสามารถปรากฏอยู่ ก่อนที่ความจริงจะถูกเปิดเผย ซึ่งจะมีรูปแบบที่พบแตกต่างกันมากมาย การวิเคราะห์ในส่วนนี้จะมุ่งเน้นการตรวจหาข่าวปลอมเพื่อแจ้งเตือนผู้ใช้ให้ระมัดระวังตัว และส่วนสุดท้ายเป็นการวิเคราะห์ข่าวสารในเชิงจิตวิทยาที่มีความพยายามในการวิเคราะห์ด้านพฤติกรรม ความรู้สึก ปัจจัยที่มีความเกี่ยวข้องกับจิตวิทยาออกมาในรูปแบบตัวเลขเชิงปริมาณ ที่มีความน่าสนใจทั้งด้านการวัดและการประเมินตัวเลขทางสถิติต่าง ๆ [13] นอกจากนั้นการวิเคราะห์เนื้อหาข่าวยังประกอบด้วยการวิเคราะห์ความหมายของคำที่ปรากฏในเนื้อหาข่าวรูปแบบไวยากรณ์ที่ใช้ในเนื้อหาข่าว การวิเคราะห์โครงสร้างประโยคที่ใช้ในข่าวความพยายามในการสร้างฐานความรู้ที่เป็นชุดข้อมูล (Corpus) ที่

ประกอบด้วยคำศัพท์เทคนิคคำศัพท์เฉพาะ และคำที่ใช้งานทั่วไป ตลอดจนการวิจัยที่เกี่ยวข้องกับอารมณ์ที่เกิดขึ้นและการรับรู้ความรู้สึกเมื่ออ่านคำที่ปรากฏในเนื้อหาข่าว [68] [70] การประมวลผลภาษาธรรมชาติถูกนำมาใช้ในการวิเคราะห์คำที่มีการใช้งาน ตลอดจนความหมายที่ต้องการสื่อของเนื้อหาข่าวปลอม เพื่อจำแนกรายละเอียดต่าง ๆ [18] [71]

– Feature-oriented สนใจในส่วนของการตรวจหาคุณลักษณะข้อมูลที่สามารถใช้จำแนกข่าวปลอมได้ แบ่งเป็นสองส่วนคือ การวิเคราะห์เนื้อหาข่าว และการวิเคราะห์บริบทที่เกี่ยวข้องกับข่าว โดยการวิเคราะห์เนื้อหาข่าวจะมุ่งเน้นไปในส่วนของการวิเคราะห์ทางภาษาศาสตร์ หรือการวิเคราะห์ภาพ เพื่อสกัดเอาคุณลักษณะเฉพาะออกจากข้อความ แต่การวิเคราะห์บริบทที่เกี่ยวข้องกับข่าวมุ่งเน้นไปในอีกสามด้าน ได้แก่ ด้านคุณลักษณะที่เกี่ยวข้องกับผู้ใช้งาน จะให้ความสนใจกับผู้ใช้ที่มีประวัติผู้ใช้ทั่วไปมากกว่าผู้ใช้ที่มีคุณลักษณะเฉพาะบุคคล ด้านคุณลักษณะของการโพสต์ที่มีการใช้เทคนิค Convolution Neural Networks (CNNs) เพื่อช่วยให้เข้าใจความคิดเห็นของผู้ใช้และปฏิกิริยาการตอบสนองต่อข่าวปลอม และด้านคุณลักษณะที่เกี่ยวข้องเครือข่ายความสัมพันธ์ระหว่างผู้ใช้งาน โดยจะมุ่งเน้นไปในส่วนของรูปแบบในการเชื่อมต่อกันระหว่างการติดต่อสื่อสารของผู้คนที่ใช้งานสื่อสังคมออนไลน์เดียวกันหรือต่างกัน และการเชื่อมโยงข้อมูลระหว่างเครือข่ายสังคมออนไลน์ต่างกันเพื่อค้นหาความจริง

– Model-oriented เป็นการปรับปรุงประสิทธิภาพในการตรวจจับข่าวปลอมให้สามารถทำงานได้ดีมากยิ่งขึ้น โดยการใช้เทคนิคการจำแนกข้อมูลหลายแบบ เช่น Supervised classification model, Unsupervised classification model และ Semi-supervised classification model โดยตัวอย่างของเทคนิคที่ใช้ในการสร้างแบบจำลองการจำแนกข้อมูลที่ได้รับความนิยมใช้กัน ได้แก่ Naïve Bayes, [26] [57] Decision Tree, Logistic Regression, K-Nearest Neighbor (KNN), Neural Network, [72] Random Forest [23] และ Support Vector Machine (SVM) [73]

– Application-oriented เป็นงานวิจัยที่เน้นการประยุกต์การใช้งาน เช่น การกระจายข่าวปลอม ในส่วนของการหารูปแบบการกระจาย [64] หรือการแทรกแซงข่าวปลอมด้วยการปรับปรุงเปลี่ยนแปลงข้อเท็จจริงให้บิดเบือนไป

หลายหน่วยงานมีความพยายามในการหยุดการเผยแพร่ข่าวปลอม [74] โดยการตรวจสอบความน่าเชื่อถือของข่าว เพื่อหยุดการกระจายของข่าวปลอม ด้วยวิธีการดังนี้

– การใช้คนตรวจสอบความถูกต้องของข้อมูล โดยมีการอ้างอิงกับเครือข่ายตรวจสอบข้อเท็จจริงระดับนานาชาติ (International Fact Checking Network : IFCN) ที่อนุญาตให้ผู้ใช้งานเครือข่ายสังคมออนไลน์ สามารถกำหนดค่าสถานะของบทความที่ผิดพลาดแบบจงใจได้ นอกจากนี้ข่าวปลอมยังถูกตรวจสอบจากองค์กรสื่อสารมวลชนต่าง ๆ เช่น หนังสือพิมพ์ Washington Post และเว็บไซต์ Snopes.com

– การใช้อัลกอริทึมในการตรวจสอบ [22] [75] [76] [77] เนื่องจากการแพร่กระจายของข่าวปลอมได้ใช้อัลกอริทึมในการเผยแพร่ ดังนั้นจึงเกิดแนวคิดในการใช้อัลกอริทึมมาจัดการแพร่กระจายของข่าวปลอมด้วยการระบุเนื้อหาและแหล่งที่มาของข่าว โดยแบ่งประเภทเป็นดังนี้

– อัลกอริทึมที่ขึ้นอยู่กับเนื้อหาของข่าว

– อัลกอริทึมที่ขึ้นอยู่กับรูปแบบทิศทางในการกระจายของข้อความข่าวออกไป

– อัลกอริทึมแบบผสมขึ้นอยู่กับการให้ค่าน้ำหนักความสำคัญของค่าที่สนใจ หรือกลุ่มของคุณลักษณะที่ส่งผลต่ออัลกอริทึมที่ใช้เรียนรู้

การใช้คนตรวจสอบข้อเท็จจริงของเนื้อหาข่าวว่าเป็นข่าวจริงหรือข่าวปลอมเป็นสิ่งที่น่าเชื่อถือมากกว่าการใช้เทคโนโลยีใด ๆ เพียงแต่การใช้แรงงานคนในการตรวจสอบข้อเท็จจริงข่าวนั้น ต้องใช้ทรัพยากรและค่าใช้จ่ายในการดำเนินการสูงมาก หน่วยงานที่ให้บริการตรวจสอบความจริงข่าวจึงมีความพยายามในการใช้วิธีการในกระบวนการเรียนรู้ด้วยเครื่อง โดยอาศัยหลักการให้เหตุผลทางตรรกศาสตร์ที่เกี่ยวข้องมาช่วยแก้ไขปัญหา [21] [22] ตัวอย่างเว็บไซต์บริการตรวจสอบความถูกต้องความน่าเชื่อถือของข่าว ได้แก่

– FactCheck.Org ตรวจสอบความถูกต้องของแถลงการณ์ โฆษณา นโยบายทางการเมืองจากนักการเมือง ผู้เชี่ยวชาญ รวมถึงกลุ่มบุคคลที่มีส่วนได้ส่วนเสียที่เกี่ยวข้อง

– Michigan Truth Squad ประเมินความถูกต้องและความจริงของโฆษณาทางการเมืองใน Michigan โดยการแสดงผลการประเมินในหัวข้อเฉพาะ

– Politifact.com ดำเนินการโดยหนังสือพิมพ์ St. Petersburg Times เป็นเว็บไซต์ที่แสดงค่าความจริงด้วย "Truth-o-meter" ในการจำแนกเรื่องจริงจากเรื่องราวในแถลงการณ์ทางการเมือง รวมถึงโฆษณาต่าง ๆ ที่เกิดขึ้น

– Project Vote Smart ตรวจสอบประวัติการลงคะแนนเสียง ประวัติที่มาของบุคคลและข้อความสาธารณะที่ปรากฏจากผู้สมัครต่าง ๆ ทั่วประเทศ

– ProPublica เป็นองค์กรอิสระที่ไม่หวังผลกำไร โดยหน่วยงานจะสร้างงานวารสารในเชิงสืบสวนสอบสวนข้อเท็จจริงที่ประชาชนให้ความสนใจ

– Fact Checker ดำเนินการโดย Glenn Kessler ซึ่งเป็นผู้สื่อข่าวของหนังสือพิมพ์ Washington Post เพื่อตรวจสอบความถูกต้องของแถลงการณ์ทางการเมืองที่เกี่ยวข้องกับตัวเลขที่สำคัญ ปัญหาที่มีความสำคัญในระดับท้องถิ่น ระดับชาติในประเทศสหรัฐอเมริกา และระดับนานาชาติ

– Snopes.com เป็นเว็บไซต์ตรวจสอบเรื่องราวที่เกี่ยวข้องกับการเมือง เรื่องชาวบ้าน เรื่องเล่าต่าง ๆ ข่าวลือที่เกิดขึ้น

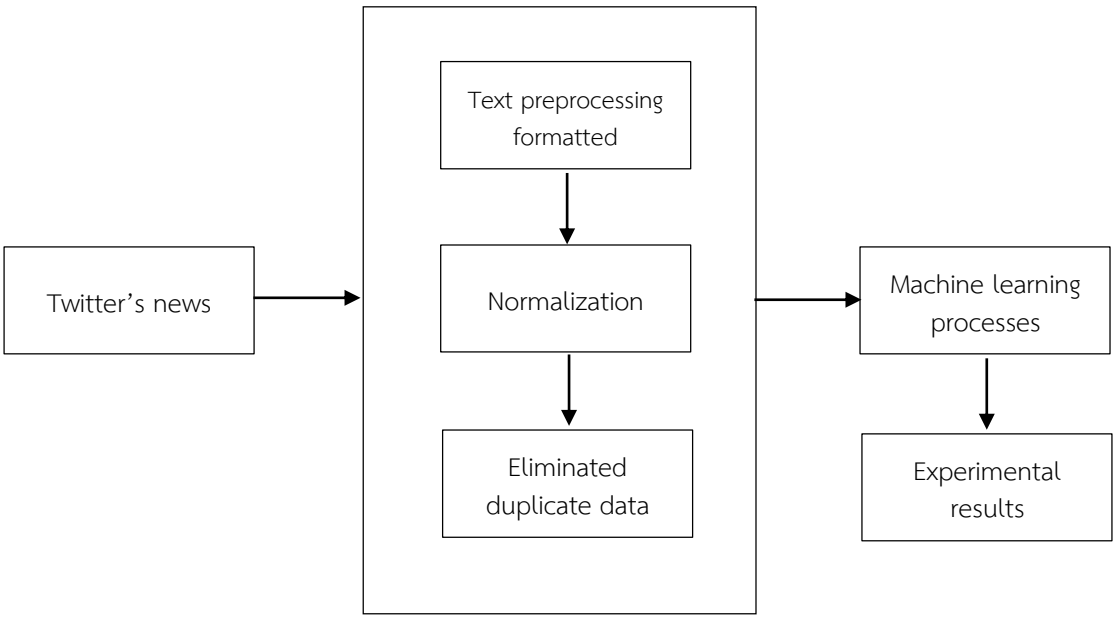
นอกจากประเด็นที่กล่าวมาข้างต้น [2] ได้อภิปรายเกี่ยวกับพื้นที่ที่เกี่ยวข้องกับงานวิจัย ปัญหาและแนวทางในการวิจัยในอนาคตสำหรับการตรวจสอบข่าวปลอมบนสื่อสังคมออนไลน์

# บทที่ 3
# วิธีการดำเนินงานวิจัย

งานวิจัยนี้เสนอ วิธีการจำแนกข่าวปลอมบนสื่อสังคมออนไลน์ทวิตเตอร์ด้วยวิธีการเรียนรู้ด้วยเครื่อง เนื้อหาในบทนี้ประกอบด้วย ภาพรวมของระบบ การเก็บข้อมูลข่าวจากสื่อสังคมออนไลน์ทวิตเตอร์ ก่อนจะนำไปประมวลผลด้วยวิธีการเรียนรู้ด้วยเครื่องการวัดและประเมินผล ดังรายละเอียดต่อไปนี้

## 3.1 ภาพรวมของระบบ

การเก็บข้อมูลเพื่อใช้ในงานวิจัยที่พัฒนาแบ่งออกเป็นสามส่วนหลัก ได้แก่ ส่วนแรกเป็นการเก็บข้อมูลหัวข้อข่าวที่สนใจจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ ส่วนต่อมาเป็นส่วนของกระบวนการปรับปรุงเปลี่ยนแปลงรูปแบบข้อมูลให้เหมาะสมกับการนำไปประมวลผล และส่วนสุดท้ายคือการนำข้อมูลไปผ่านกระบวนการเรียนรู้ด้วยเครื่อง โดยรายละเอียดของภาพรวมกระบวนการดำเนินงานในงานวิจัยนี้แสดงดังรูปที่ 3.1




**รูปที่ 3.1** ภาพรวมกระบวนการดำเนินงานในงานวิจัย

จากรูปที่ 3.1 แสดงภาพรวมกระบวนการดำเนินงานในงานวิจัยประกอบด้วยสามส่วน โดยเริ่มต้นจากการเก็บข้อมูลหัวข้อข่าวที่สนใจจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ จากนั้นจึงทำการประมวลผลจากข้อมูลดิบของหัวข้อข่าวที่กำหนดให้ที่อยู่ในรูปแบบไม่เป็นโครงสร้าง (Unstructured data) ให้เป็นข้อมูลที่อยู่ในรูปแบบเป็นโครงสร้าง (Structure data) ที่เหมาะสมโดยจัดเก็บเป็นไฟล์ .csv เนื่องจากข้อมูลที่มีลักษณะไม่เป็นโครงสร้างนั้นไม่เหมาะสมกับการนำไปใช้ในกระบวนการเรียนรู้ด้วยเครื่อง [56] จากนั้นจึงนำไฟล์ .csv ที่ได้มาปรับเปลี่ยนรูปแบบข้อมูลที่อยู่ในโครงสร้างที่กำหนดไว้จากข้อมูลที่เป็นข้อความต่าง ๆ ให้เปลี่ยนค่าเป็นค่าตัวเลขโดยใช้เงื่อนไขที่กำหนดไว้ในตารางที่ 3.1 แล้วจึงทำความสะอาดข้อมูลโดยการขจัดข้อมูลส่วนที่ซ้ำกันออกไปให้เหลือเฉพาะข้อมูลที่ไม่ซ้ำกันเลย และเข้าสู่กระบวนการสุดท้ายคือการนำข้อมูลไปผ่านกระบวนการเรียนรู้ด้วยเครื่องได้เป็นผลลัพธ์ออกมาในที่สุด

### 3.2 การเก็บข้อมูลข่าวจากสื่อสังคมออนไลน์ทวิตเตอร์

การเลือกหัวข้อข่าวสำหรับงานวิจัยนี้ ให้ความสำคัญกับหัวข้อข่าวที่มีความเกี่ยวข้องกับเหตุการณ์จากปรากฏการณ์ทางธรรมชาติ และเหตุการณ์สำคัญเฉพาะกาล ซึ่งภัยคุกคามทางธรรมชาติเป็นเรื่องที่คาดเดาได้ยากว่าจะเกิดเหตุการณ์ร้ายแรงขึ้นมาเมื่อใด บางครั้งเมื่อมีข่าวลวงเกิดขึ้นมาในระหว่างที่เกิดเหตุการณ์ คนทั่วไปจะทราบได้อย่างไรว่าข่าวที่มีการบอกต่อกันนั้นเกิดจากข้อเท็จจริง หรือเกิดจากการเข้าใจคลาดเคลื่อนในการแจ้งข่าวสารก่อให้เกิดความเท็จแล้วส่งต่อกันไปจึงทำให้ผู้คนเกิดความเข้าใจผิดพลาดไปได้

การดำเนินงานวิจัยในเบื้องต้น ได้จัดเก็บข้อมูลจากสื่อสังคมออนไลน์ทวิตเตอร์ ในช่วงที่มีการเปลี่ยนแปลงฤดูกาล ระหว่างเดือนตุลาคม ถึงพฤศจิกายน พ.ศ. 2560 โดยเลือกเก็บข้อมูลในหัวข้อเรื่องที่เกี่ยวข้องกับภัยที่เกิดจากผลกระทบสืบเนื่องจากปรากฏการณ์ทางธรรมชาติ เช่น น้ำท่วม, น้ำท่วมกรุงเทพ, ฝนตก, น้ำป่า, เขื่อนแตก, พายุ, พายุขนุน, พายุไต้ฝุ่นล้ง, พายุดีเพรสชัน, ปล่อยน้ำท่วมรังสิต, ปล่อยน้ำท่วม, แผ่นดินไหว, เข้าสู่ฤดูหนาว, ภัยหนาว, อุณหภูมิลดลง, สภาพอากาศแปรปรวน, หนาว เป็นต้น นอกจากนั้นในช่วงเวลาดังกล่าวยังมีข่าวที่เกี่ยวข้องกับงานที่มีความสำคัญต่อจิตใจของประชาชนชาวไทยอีกประเด็นหนึ่งคือข่าวที่เกี่ยวข้องกับงานพระราชพิธีถวายพระเพลิงพระบรมศพพระบาทสมเด็จพระเจ้าอยู่หัว รัชกาลที่ 9 ซึ่งเป็นเหตุการณ์เฉพาะกาลในช่วงเวลานั้น เนื่องจากในรัชสมัยของรัชกาลที่ 9 มีช่วงระยะเวลายาวนานถึง 70 ปี มีโครงการอันเนื่องมาจากพระราชดำริ และจากพระราชกรณียกิจที่เกิดขึ้นเพื่อประโยชน์สุขของประชาชนจำนวนมากมาย ดังนั้นประชาชนไทยส่วนมากจึงให้ความรักความเทิดทูนต่อสถาบันพระมหากษัตริย์เป็นอย่างยิ่ง ด้วยความรักความผูกพันที่เกิดขึ้นนี้ทำให้ประเด็นข่าวใดก็ตามที่มีความเกี่ยวข้องกับพระมหากษัตริย์จะเป็นที่ได้รับความสนใจ

จากประชาชนอย่างมากเป็นพิเศษ ในงานวิจัยนี้จึงได้เลือกหัวข้อข่าวที่มีคำสำคัญที่เกี่ยวข้องกับรัชกาลที่ 9 ในช่วงเวลาดังกล่าว เช่น รัชกาลที่ 9, พระราชพิธีถวายพระเพลิง, ดอกไม้เพื่อพ่อ, ปากคลองตลาด, นิทรรศการงานพระราชพิธี, สู่ฟ้าเสวยสวรรค์, สถิตในดวงใจไทยนิรันดร์, ส่งเสด็จสู่สวรรคาลัย, เติมน้ำมันฟรี, บางจากเติมฟรี, ดอกไม้จันทน์, พันก้านดอกไม้จันทน์จากสีดำเป็นขาว, ย้ายผู้ว่าชลบุรี นนทบุรี เป็นต้น และยังมีหัวข้อข่าวอื่นที่เป็นเหตุการณ์ที่คนในสังคมให้ความสนใจ เช่น การลดหย่อนภาษี การเก็บเงินสมทบกองทุนประกันสังคม เป็นต้น

จากหัวข้อข่าวที่กล่าวมาข้างต้น การเลือกเก็บข้อมูลจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ผ่าน Twitter API ด้วย 22 คุณลักษณะ ดังต่อไปนี้

Id, Name, IsVerified, ProfileImageUrl, FollowersCount, FriendsCount, FavouritesCount, StatusesCount, Description, Location, TimeZone, UserCreatedDate, Status, Url, Mentions, Number of Mentions, HashTags, Number of HashTags, RetweetCount, TweetCreatedDate, MessageText, MessageImage

โดยการเลือกใช้ 22 คุณลักษณะในงานวิจัยนี้ ได้แนวความคิดมาจากงานวิจัย [23] ที่มีการใช้ 20 คุณลักษณะในการประเมินความน่าเชื่อถือข้อความภาษาอาหรับในเครือข่ายสังคมออนไลน์ทวิตเตอร์ ซึ่งในงานวิจัยนี้หลังจากเก็บข้อมูลได้มีการวิเคราะห์ความสำคัญของคุณลักษณะหากมีการพิจารณาตัดคุณลักษณะใดออกแล้วจะส่งผลต่อค่าร้อยละของความถูกต้องในการวิเคราะห์ข่าวปลอม รายละเอียดจะกล่าวถึงในหัวข้อถัดไป

เมื่อจัดเก็บข้อมูลจากทวิตเตอร์ตามเงื่อนไขที่ได้กล่าวมาข้างต้นแล้ว ได้ข้อความที่เกี่ยวข้องกับหัวข้อข่าวที่กำหนด นำข้อความที่ได้มาผ่านกระบวนการทำความสะอาดข้อมูลก่อนนำไปประมวลผล จำนวนทั้งหมด 948,373 ข้อความ

เมื่อจัดเก็บข้อมูลจากทวิตเตอร์และปรับรูปแบบข้อมูลในอยู่ในรูปโครงสร้างด้วยไฟล์ .csv แล้ว ลำดับต่อไปในการจัดเตรียมข้อมูลสำหรับกระบวนการเรียนรู้ด้วยเครื่องคือการนำข้อมูลมาปรับค่าให้เป็นตัวเลข (Encoder to data normalization) ซึ่งเป็นการจัดหมวดหมู่ข้อมูลเพื่อทำให้ง่ายและเกิดประสิทธิภาพในกระบวนการเรียนรู้มากขึ้น การนำข้อมูลมาปรับค่าให้เป็นตัวเลข ก่อนนำไปเข้าสู่กระบวนการเรียนรู้ด้วยเครื่อง

รูปแบบการปรับค่าข้อมูลให้เป็นตัวเลขมีหลายวิธี [78] ยกตัวอย่างเช่น

1) การปรับค่าข้อมูลโดยการเข้ารหัสแบบคลาสสิค (Classic Encoders)

ข้อมูลที่อยู่ในรูปแบบตัวอักษร หรือข้อความต่าง ๆ หากใช้การเข้ารหัสแบบคลาสสิคจะใช้วิธีการแปลงค่าข้อมูลที่อยู่ในรูปแบบตัวอักษรหรือตัวเลขให้เปลี่ยนแปลงไปเป็นตัวเลขหนึ่งเพื่อสื่อความหมายตามที่ต้องการ เช่น

– Ordinal—เป็นวิธีการแปลงข้อมูลอย่างง่ายที่กำหนดให้ข้อความต่าง ๆ เปลี่ยนไปเป็นตัวเลขจำนวนเต็มตามที่ต้องการ โดยตัวเลขผลการแปลงค่า จะมีค่าอยู่ในช่วงระหว่าง 1 ถึง k โดยที่ค่า k เป็นจำนวนข้อมูลที่ไม่ซ้ำกันเลยในจำนวนมิติข้อมูลที่ต้องการแปลง

– OneHot—เป็นการแปลงข้อมูลจากค่าหนึ่งไปเป็นตัวเลขที่มีค่าเพียงค่าใดค่าหนึ่งจากค่าในหลายมิติที่กำหนด โดยที่ค่าที่อยู่ในมิติอื่น ๆ จะมีค่าเป็นศูนย์หรือไม่มีค่า เพื่อทำให้สามารถเปรียบเทียบกับค่าที่ได้กับค่าในมิติอื่น ๆ

– Binary—เป็นวิธีการแปลงค่าแต่ละตัวเลขให้อยู่ในรูปแบบตัวเลขฐานสอง ซึ่งแต่ละหลักของเลขฐานสองจะถูกเก็บในแต่ละคอลัมน์แยกกัน ซึ่งรูปแบบนี้อาจทำให้ข้อมูลบางส่วนหายไปได้ในกรณีที่มีจำนวนคอลัมน์สำหรับจัดเก็บข้อมูลน้อยกว่าจำนวนหลักจริงของค่าตัวเลขฐานสองที่แปลงได้

– BaseN—เป็นการแปลงค่าลักษณะเดียวกันกับการแปลงค่าให้อยู่ในรูปแบบเลขฐานสอง แต่ใช้ฐานเลขเป็นจำนวน N ใด ๆ แทน ซึ่งข้อจำกัดจะน้อยกว่าแบบ Binary เนื่องจากช่วงค่าข้อมูลมีค่ามากกว่า

– Hashing—เป็นการแปลงข้อมูลจากค่าหนึ่งไปเป็นตัวเลขที่มีค่าในคอลัมน์ใดเพียงคอลัมน์หนึ่งลักษณะเช่นเดียวกับแบบ OneHot แต่วิธีการนี้จะใช้จำนวนคอลัมน์ที่น้อยกว่าและมีการใช้ฟังก์ชันการคำนวณผลลัพธ์ ซึ่งอาจทำให้ผลที่ได้เกิดปัญหาการชนกันของข้อมูล (Collision)

2) วิธีการเข้ารหัสแบบเปรียบเทียบ (Contrast Encoders)

การเข้ารหัสในลักษณะนี้ใช้สำหรับกรณีที่มีหลายปัญหาที่ขัดแย้งกัน ซึ่งผลลัพธ์ที่ได้จากการแปลงค่าจะได้ข้อมูลหนึ่งคอลัมน์สำหรับข้อมูลหนึ่งค่า

– Helmert (reverse)—ใช้กับตัวแปรที่มีความอิสระต่อกัน เพื่อใช้เปรียบเทียบค่าเฉลี่ยกับความสัมพันธ์ที่ขึ้นอยู่กับตัวแปรก่อนหน้า

– Sum—เปรียบเทียบค่าเฉลี่ยของตัวแปรที่ขึ้นกับค่าตัวแปรก่อนหน้ากับค่าเฉลี่ยของตัวแปรทั้งหมดในระดับเดียวกัน

– Backward Difference—เปรียบเทียบค่าเฉลี่ยตัวแปรตามในระดับข้อมูลกับค่าเฉลี่ยตัวแปรตามของข้อมูลระดับก่อนหน้า

– Polynomial—การเปรียบเทียบความแตกต่างระหว่างค่าเฉลี่ยข้อมูลด้วย Orthogonal polynomial โดยแสดงออกมาในรูปกราฟความสัมพันธ์ เพื่อวิเคราะห์แนวโน้มโดยพิจารณาจากค่าสัมประสิทธิ์ของกราฟที่ได้ หากเป็นรูปแบบกราฟเส้นตรงแสดงว่าได้ค่าสัมประสิทธิ์ยก

กำลังหนึ่ง (Linear) กราฟเส้นโค้งแสดงว่าได้ค่าสัมประสิทธิ์ยกกำลังสอง (Quadratic) หรือค่าสัมประสิทธิ์ยกกำลังสาม (Cubic)

งานวิจัยนี้เลือกใช้วิธีการแปลงเป็นตัวเลขแบบ Ordinal ซึ่งเข้าใจได้ง่ายเป็นเงื่อนไขการแปลงข้อมูลให้อยู่ในช่วงตัวเลขจำนวนเต็มที่เหมาะสมกับแต่ละคุณลักษณะข้อมูล โดยมีรายละเอียดที่ใช้ในการแปลงข้อมูลเป็นตัวเลขดังต่อไปนี้

1) Id หรือรหัสผู้ใช้เป็นตัวเลขที่ทวิตเตอร์ใช้ในการจำแนกแต่ละบัญชีผู้ใช้งาน ประกอบด้วยตัวเลขสองลักษณะคือลักษณะแรกประกอบด้วยตัวเลขจำนวน 9-10 หลักใช้สำหรับบัญชีผู้ใช้ที่ได้เริ่มต้นใช้งานในช่วงแรกที่ทวิตเตอร์เริ่มเปิดให้บริการ และอีกลักษณะหนึ่งประกอบด้วยตัวเลขจำนวน 18 หลักเป็นบัญชีของผู้ใช้ที่ขอใช้งานทวิตเตอร์มาไม่นาน

2) Name ซึ่งคือชื่อผู้ใช้งานทวิตเตอร์ โดยรูปแบบของชื่อผู้ใช้งานที่ปรากฏแบ่งได้เป็น 4 รูปแบบ ได้แก่ ชื่อที่ประกอบด้วยตัวอักขระภาษาไทยเท่านั้น ชื่อที่ประกอบด้วยตัวอักขระภาษาอังกฤษเท่านั้น ชื่อที่ประกอบด้วยตัวอักขระภาษาไทยผสมกับภาษาอังกฤษกับตัวเลข และชื่อที่ประกอบด้วยตัวอักขระภาษาอื่น ๆ

3) Isverified เป็นคุณลักษณะของผู้ใช้ที่ผู้ใช้ได้ยืนยันตัวบุคคลกับทางทวิตเตอร์ ซึ่งเงื่อนไขนี้มีค่าอยู่สองแบบคือ true สำหรับผู้ใช้งานที่ผ่านการยืนยันตัวตนผู้ใช้แล้ว และ false สำหรับผู้ใช้ที่ไม่มีการยืนยันตัวตน

4) ProfileImageUrl คือตำแหน่งที่จัดเก็บ URL ที่เชื่อมโยงไปยังรูปภาพประวัติส่วนตัวของผู้ใช้ ซึ่งผู้ใช้บางคนอาจไม่มีรูปภาพประวัติส่วนตัว หรือมีรูปภาพที่มีนามสกุลเป็น .jpg หรือรูปภาพที่มีนามสกุลเป็น .png หรือรูปภาพที่มีนามสกุลอื่น ๆ

5) FollowersCount ซึ่งเป็นตัวเลขการนับจำนวนผู้ติดตามของบัญชีผู้ใช้ Twitter แต่ละคน ตัวเลขจำนวนผู้ติดตามนี้ประกอบด้วยตัวเลขหลายหลัก จึงใช้วิธีการคำนวณโดยการหารด้วยเลข 10 ได้ผลออกมาเป็นตัวเลขจำนวนหลักของผู้ติดตามแทนจำนวนนับจริง ๆ จากข้อมูลเบื้องต้นในที่นี้มีค่าอยู่ในช่วงระหว่าง 1-7 โดยเริ่มต้นจาก 1 หมายถึงบัญชีนั้นมีจำนวนผู้ติดตาม 0-9 คน และมีค่าไม่เกิน 7 ที่หมายถึงมีผู้ติดตามบัญชีนั้น 1,000,000-9,999,999 คน

6) Friendscount, Favoritescount, และ StatusesCount หมายถึง จำนวนเพื่อนที่มีของบัญชี จำนวนข้อความที่ผู้ใช้เคยกดถูกใจ และจำนวนสถานะของบัญชี ตามลำดับ ซึ่งค่าที่กล่าวมานี้เป็นค่าตัวเลขลักษณะแบบเดียวกับ FollowersCount จึงใช้เงื่อนไขในการแปลงค่าตัวเลขเช่นเดียวกัน

7) Description คือรายละเอียดของผู้ใช้ที่ต้องการอธิบายเกี่ยวกับตัวเอง เงื่อนไขที่ใช้ในการแปลงเป็นตัวเลขของเงื่อนไขนี้เป็นเช่นเดียวกับ Name เนื่องจากมีคุณลักษณะข้อมูลแบบเดียวกัน

8) Location เป็นสถานที่ที่ผู้ใช้โพสต์ข้อความ TimeZone เป็นเขตเวลาที่บัญชีผู้ใช้นั้นถูกสร้างขึ้น ทั้งสองคุณลักษณะนี้เป็นชื่อสถานที่เช่นเดียวกันจึงใช้เงื่อนไขเดียวกัน โดยแบ่งเขตเวลาและสถานที่ตามตำแหน่งบริเวณที่ตั้งของแต่ละภูมิภาค ในงานวิจัยนี้สนใจข้อความทวิตเตอร์ภาษาไทยจึงต้องการเน้นเฉพาะส่วนประเทศไทยจึงแยกออกจากมาเป็นอีกส่วน โดยรายละเอียดทั้ง 7 เขต ได้แก่ ในประเทศไทย, เอเชียตะวันออกเฉียงใต้, เอเชีย, ออสเตรเลีย/นิวซีแลนด์, ยุโรป/รัสเซีย, สหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย และ แอฟริกา นอกจากนี้อาจมีผู้ใช้บางคนไม่ได้ระบุสถานที่และเขตเวลาใน 7 บริเวณที่กล่าวมาข้างต้น

9) UserCreateDate เป็นวันที่ผู้ใช้สร้างบัญชีผู้ใช้งานขึ้นมา ทวิตเตอร์ถูกสร้างขึ้นมาใช้งานครั้งแรกในเดือนมีนาคม ค.ศ. 2006 และมีการเริ่มต้นใช้งานในเดือนกรกฎาคมในปีเดียวกัน เพื่อให้สามารถแยกกลุ่มอายุการใช้งานของผู้ใช้ได้ ในงานวิจัยนี้จึงได้ประมวลผลค่าวันที่บัญชีผู้ใช้งานถูกสร้างขึ้นเป็นอายุการใช้งานของบัญชีแล้วเลือกแบ่งกลุ่มย่อยตามระยะเวลาที่บัญชีได้เปิดใช้งานมาแล้ว โดยกำหนดการแบ่งช่วงห่างของระยะเวลาการเปิดใช้งานแต่ละช่วงเท่ากับ 6 เดือน หรือมีค่าเท่ากับ 0.5 ปี เช่น 0.5, 1, 1.5, ...

10) Status เป็นคุณลักษณะที่ผู้ใช้บ่งบอกสถานะของบัญชีผู้ใช้งาน ในเงื่อนไขนี้สนใจเฉพาะการมีค่าสถานะปรากฏอยู่ หรือไม่มีค่าสถานะ

11) Url เป็นข้อความของการเชื่อมโยงไปยังสื่อปลายทางที่ข้อความนั้นต้องการระบุถึง ในที่นี้แต่ละข้อความอาจไม่มี URL ปรากฏเลย หรืออาจมีเพียงหนึ่ง URL หรืออาจมีมากกว่าก็ได้

12) Mentions เป็นรายละเอียดของชื่อผู้ใช้ที่ถูกกล่าวถึงในข้อความ ซึ่งจะปรากฏในข้อความตามหลังสัญลักษณ์ @ และในแต่ละข้อความอาจจะไม่มี Mentions เลย หรือมี Mentions เพียงหนึ่งสัญลักษณ์ หรือมากกว่าก็ได้

13) Number of Mentions เป็นจำนวนสัญลักษณ์ @ ที่ปรากฏในข้อความ ซึ่งค่าดังกล่าวนี้ได้ตัวเลขมาจาก Twitter API

14) HashTags เป็นรายละเอียดที่ผู้ใช้ต้องการอธิบายหรือระบุเฉพาะเจาะจงในหัวข้อเรื่องของข้อความโดยปกติข้อความที่เป็น HashTag จะปรากฏสัญลักษณ์ # หน้าข้อความ และอาจไม่มีในข้อความเลยหรือมีมากกว่าหนึ่งสัญลักษณ์ในข้อความก็ได้

15) Number of HashTags เป็นจำนวนสัญลักษณ์ # ที่ปรากฏในข้อความ ซึ่งได้ค่าตัวเลขนี้ได้มาจาก Twitter API เช่นเดียวกันกับจำนวนตัวเลขของ Mentions

16) RetweetCount เป็นจำนวนตัวเลขของข้อความที่ถูกนำไปเผยแพร่ซ้ำอีกครั้งโดยบุคคลอื่นที่ไม่ใช่เจ้าของข้อความ เนื่องจากจำนวนค่าตัวเลขที่ได้มีความคล้ายกับจำนวนผู้ติดตาม จึงใช้เงื่อนไขในการแปลงเป็นตัวเลขเช่นเดียวกัน ซึ่งจากข้อมูลที่ใช้ในงานวิจัยจะได้ค่าผลลัพธ์ที่ได้อยู่ในช่วงระหว่าง 1-7

17) TweetCreatedDate เป็นวันที่เวลาที่ข้อความนั้นถูกสร้างขึ้น ซึ่งค่าตัวเลขที่นำมาใช้ในการแปลงได้แบ่งตามช่วงเวลาที่สร้างข้อความขึ้นมา เพื่อให้สามารถแยกแยะได้ว่าข้อความนั้นถูกสร้างในเวลากลางวันหรือกลางคืน จึงแบ่งช่วงเวลาของข้อความที่ถูกสร้างเป็น 4 ช่วงดังนี้ 06.01-12.00 น., 12.01-18.00 น., 18.01-24.00 น., 00.01-06.00 น.

18) MessageText เป็นรายละเอียดของข้อความที่มีการส่งถึงกัน โดยข้อความที่ส่งนี้อาจเป็นข้อความที่เขียนขึ้นมาเอง หรือเป็นข้อความที่ส่งต่อจากบัญชีผู้ใช้อื่นก็ได้ ดังนั้นเงื่อนไขในการแปลงนี้จึงตรวจสอบความเป็นเจ้าของหรือไม่ใช่แต่เป็นข้อความที่ถูกส่งต่ออีกครั้ง

19) MessageImage เป็นที่อยู่การเชื่อมโยง URL ไปยังรูปภาพที่เกี่ยวข้องกับข้อความ โดยในแต่ละข้อความอาจจะไม่ปรากฏหรืออาจมีเพียงหนึ่งหรืออาจมีมากกว่าหนึ่งการเชื่อมโยงไปยังรูปภาพที่เกี่ยวข้องก็ได้

จากเงื่อนไขที่กล่าวมาข้างต้น สรุปข้อมูลแสดงในรูปแบบตารางได้ดังตารางที่ 3.1

**ตารางที่ 3.1** เงื่อนไขที่ใช้ในการแปลงข้อมูลแต่ละคุณลักษณะเป็นตัวเลข

| **คุณลักษณะ** | **เงื่อนไขการแปลงค่า** | |
|---|---|---|
| | **รายละเอียดที่ตรวจสอบ** | **ผลการแปลงค่า** |
| Id | ไม่ปรากฏข้อมูล | 0 |
| | 9-10 หลัก | 1 |
| | 18 หลัก | 2 |
| Name, Description | ไม่ปรากฏข้อมูล | 0 |
| | ตัวอักขระภาษาไทยทั้งหมด | 1 |
| | ตัวอักขระภาษาอังกฤษทั้งหมด | 2 |
| | ตัวอักขระภาษาไทยหรือภาษาอังกฤษหรือตัวเลข | 3 |
| | ตัวอักขระภาษาอื่น ๆ | 4 |
| | สัญลักษณ์อื่น ๆ | 5 |

**ตารางที่ 3.1 (ต่อ)** เงื่อนไขที่ใช้ในการแปลงข้อมูลแต่ละคุณลักษณะเป็นตัวเลข

| คุณลักษณะ | เงื่อนไขการแปลงค่า | |
|---|---|---|
| | **รายละเอียดที่ตรวจสอบ** | **ผลการแปลงค่า** |
| IsVerified | ไม่ปรากฏข้อมูล | 0 |
| | True | 1 |
| | False | 2 |
| ProfileImageUrl | ไม่ปรากฏข้อมูล | 0 |
| | .jpg | 1 |
| | .png | 2 |
| | อื่น ๆ | 3 |
| FollowersCount, FriendsCount, FavouritesCount, StatusesCount, RetweetCount, | ไม่ปรากฏข้อมูล | 0 |
| | 1-9 | 1 |
| | 10-99 | 2 |
| | 100-999 | 3 |
| | 1,000-9,999 | 4 |
| | 10,000-99,999 | 5 |
| | 100,000-999,999 | 6 |
| | 1,000,000-9,999,999 | 7 |
| Location, TimeZone | ไม่ปรากฏข้อมูล | 0 |
| | ประเทศไทย | 1 |
| | เอเชียตะวันออกเฉียงใต้ยกเว้นประเทศไทย | 2 |
| | เอเชียยกเว้นเอเชียตะวันออกเฉียงใต้ | 3 |
| | ออสเตรเลีย/นิวซีแลนด์ | 4 |
| | ยุโรป/รัสเซีย | 5 |
| | สหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย | 6 |
| | แอฟริกา | 7 |
| | อื่น ๆ | 8 |
| CreatedDate | ไม่ปรากฏข้อมูล | 0 |
| | น้อยกว่า 0.5 ปี | 1 |
| | ระหว่าง 0.5 ปี ถึง 1 ปี | 2 |
| | ... (เพิ่มขึ้นช่วงละ 0.5 ปี) | ... |

**ตารางที่ 3.1 (ต่อ)** เงื่อนไขที่ใช้ในการแปลงข้อมูลแต่ละคุณลักษณะเป็นตัวเลข

| คุณลักษณะ | เงื่อนไขการแปลงค่า | |
|---|---|---|
| | **รายละเอียดที่ตรวจสอบ** | **ผลการแปลงค่า** |
| Url, MessageImage, Mentions, HashTags, Number of Mentions, Number of HashTags | ไม่ปรากฏข้อมูล | 0 |
| | 1 ลิงค์ หรือ 1 @ หรือ 1 # | 1 |
| | 2 ลิงค์ หรือ 2 @ หรือ 2 # | 2 |
| | 3 ลิงค์ หรือ 3 @ หรือ 3 # | 3 |
| | มากกว่า 3 ลิงค์ หรือ 3 @ หรือ 3 # | 4 |
| Status | ไม่ปรากฏข้อมูล | 0 |
| | มีค่าสถานะ | 1 |
| TweetCreatedDate | ไม่ปรากฏข้อมูล | 0 |
| | 06.01-12.00 น. | 1 |
| | 12.01-18.00 น. | 2 |
| | 18.01-24.00 น. | 3 |
| | 00.01-06.00 น. | 4 |
| MessageText | ข้อความที่เขียนเอง | 0 |
| | ข้อความที่ Retweet | 1 |

ขั้นตอนนี้มีการระบุประเภทของข่าว (Class) แต่ละรายการโดยค่าของรายการที่เป็นเนื้อหาข่าวจริงมีค่าเป็น 1 และข่าวปลอมมีค่าเป็น 0 หลังจากกระบวนการปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบตัวเลขทั้งหมดแล้ว จะได้ไฟล์ .csv ที่มีชุดข้อมูลตัวเลขแทนข้อความที่เกี่ยวข้องกับหัวข้อข่าวทั้งหมด 948,373 รายการ ต่อจากนั้นยังได้ทำความสะอาดชุดข้อมูลโดยการกำจัดข้อมูลส่วนที่ซ้ำกันออกไป เพื่อให้กระบวนการเรียนรู้ด้วยเครื่องทำงานได้อย่างมีประสิทธิภาพมากขึ้น หลังจากการเอารายการที่ซ้ำกันออก จึงเหลือชุดข้อมูลที่มี่ค่าไม่ซ้ำกันเลยจำนวน 327,784 รายการ สำหรับนำไปใช้ในกระบวนการเรียนรู้ด้วยเครื่องเพื่อจำแนกข่าวปลอมต่อไป ตัวอย่างรายการบางส่วนของชุดข้อมูลที่ไม่ซ้ำกันแสดงดังตารางที่ 3.2

**ตารางที่ 3.2** ตัวอย่างรายการข้อมูลหลังจากกระบวนการปรับเปลี่ยนข้อมูลให้อยู่ในรูปแบบตัวเลข

| Id | Name | IsVerified | ProfileImageUrl | FollowersCount | FriendsCount | FavouritesCount | StatusesCount | Description | Location | TimeZone | CreatedDate | Status | Url | Mentions | Number of Mentions | HashTags | Number of HashTags | RetweetCount | TweetCreatedDate | MessageText | MessageImage | Class |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3 | 2 | 1 | 3 | 3 | 5 | 5 | 3 | 1 | 0 | 4 | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 1 | 1 | 0 | 1 |
| 2 | 3 | 2 | 1 | 2 | 3 | 3 | 5 | 3 | 1 | 0 | 1 | 1 | 3 | 2 | 2 | 2 | 2 | 2 | 4 | 1 | 0 | 0 |

จากตัวอย่างรายการข้อมูลในบรรทัดแรกของตารางที่ 3.2 สามารถอธิบายหมายความของข้อมูลรายการนี้ได้ดังรายละเอียดต่อไปนี้

Id มีค่าเท่ากับ 1 แสดงว่ารหัสของบัญชีผู้ใช้เป็นข้อมูลที่เป็นตัวเลขขนาด 9-10 หลักที่เป็นบัญชีผู้ใช้ที่ได้เริ่มต้นใช้งานในช่วงแรกที่ทวิตเตอร์เริ่มเปิดให้บริการ

Name มีค่าเท่ากับ 3 แสดงว่าเป็นชื่อผู้ใช้งานทวิตเตอร์ที่ประกอบด้วยตัวอักขระภาษาไทยผสมกับภาษาอังกฤษกับตัวเลข

IsVerified มีค่าเท่ากับ 2 แสดงถึงคุณลักษณะของผู้ใช้ที่ผู้ใช้ได้ยืนยันตัวบุคคลกับทางทวิตเตอร์ มีค่าเป็น false นั่นคือผู้ใช้นี้ไม่มีการยืนยันตัวตน

ProfileImageUrl มีค่าเท่ากับ 1 แสดงว่าตำแหน่งที่จัดเก็บ URL ที่เชื่อมโยงไปยังรูปภาพประวัติส่วนตัวของผู้ใช้เป็นรูปภาพที่มีนามสกุลเป็น .jpg

FollowersCount มีค่าเท่ากับ 3 แสดงว่าจำนวนผู้ติดตามของบัญชีผู้ใช้งานนี้ อยู่ในช่วงค่าระหว่าง 100-999

FriendsCount มีค่าเท่ากับ 3 แสดงว่าจำนวนเพื่อนที่มีของบัญชีผู้ใช้งานนี้ อยู่ในช่วงค่าระหว่าง 100-999

FavouritesCount มีค่าเท่ากับ 5 แสดงว่าจำนวนข้อความที่ผู้ใช้นี้เคยกดถูกใจ อยู่ในช่วงค่าระหว่าง 10,000-99,999

StatusesCount มีค่าเท่ากับ 5 แสดงว่าจำนวนสถานะของบัญชี อยู่ในช่วงค่าระหว่าง 10,000-99,999

Description มีค่าเท่ากับ 3 แสดงว่าเป็นรายละเอียดของผู้ใช้ที่ต้องการอธิบายเกี่ยวกับตัวเองที่ประกอบด้วยตัวอักขระภาษาไทยผสมกับภาษาอังกฤษกับตัวเลข

Location มีค่าเท่ากับ 1 แสดงว่าสถานที่ที่ผู้ใช้โพสต์ข้อความอยู่ในประเทศไทย

TimeZone มีค่าเท่ากับ 0 แสดงว่าไม่มีการระบุค่าเขตเวลาที่บัญชีผู้ใช้นั้นถูกสร้างขึ้น

CreatedDate มีค่าเท่ากับ 4 แสดงว่าบัญชีผู้ใช้นี้ได้เปิดใช้งานมาแล้วเป็นระยะเวลานานระหว่าง 1.5 ปี ถึง 2 ปี

Status มีค่าเท่ากับ 1 แสดงว่าบัญชีผู้ใช้งานนี้มีค่าสถานะปรากฏอยู่

Url มีค่าเท่ากับ 0 แสดงว่าข้อความนี้ไม่มีจุดเชื่อมโยงไปยังสื่อปลายทางอื่น ๆ

Mentions มีค่าเท่ากับ 2 แสดงว่ามีรายละเอียดของชื่อผู้ใช้ที่ถูกกล่าวถึงในข้อความตามหลังสัญลักษณ์ @ เพียง 2 ชื่อ

Number of Mentions มีค่าเท่ากับ 2 แสดงว่า Twitter API คืนค่าจำนวนสัญลักษณ์ @ ที่ปรากฏในข้อความมาเป็น 2 ซึ่งเป็นที่สังเกตว่ามีค่าเท่ากับจำนวน Mentions ที่พบในข้อมูลรายการเดียวกัน

HashTags มีค่าเท่ากับ 3 แสดงว่ามีรายละเอียดที่ผู้ใช้ต้องการอธิบายหรือระบุเฉพาะเจาะจงในข้อความตามหลังสัญลักษณ์ # จำนวน 3 ประเด็น

Number of HashTags มีค่าเท่ากับ 3 แสดงว่า Twitter API คืนค่าจำนวนสัญลักษณ์ # ที่ปรากฏในข้อความมาเป็น 3 ซึ่งเป็นที่สังเกตว่ามีค่าเท่ากับจำนวน HashTags ที่พบในข้อมูลรายการเดียวกัน

RetweetCount มีค่าเท่ากับ 4 แสดงว่าข้อความนี้มีการนำไปเผยแพร่ซ้ำอีกจำนวนอยู่ในช่วงค่าระหว่าง 1,000-9,999 ครั้งโดยบุคคลอื่นที่ไม่ใช่เจ้าของข้อความ

TweetCreatedDate มีค่าเท่ากับ 1 แสดงว่าวันที่เวลาที่ข้อความนั้นถูกสร้างขึ้นอยู่ในช่วงเวลาระหว่าง 06.01-12.00 น.

MessageText มีค่าเท่ากับ 1 แสดงว่ารายละเอียดของข้อความที่มีการส่งถึงกันเป็นข้อความที่ถูกส่งต่ออีกครั้ง (Retweet)

MessageImage มีค่าเท่ากับ 0 แสดงว่าไม่มีการเชื่อมโยง URL ไปยังรูปภาพที่เกี่ยวข้องกับข้อความ

Class มีค่าเท่ากับ 1 แสดงว่า ข้อความรายการนี้เป็นข่าวจริง

การเก็บข้อมูลข่าวในงานวิจัยนี้เลือกเก็บข้อมูลจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ระหว่างเดือนตุลาคม ถึงเดือนพฤศจิกายน 2560 โดยมีรายละเอียดจำนวนหัวข้อข่าวที่ได้จัดเก็บรวมมาจำนวน 948,373 ข้อความ แบ่งตามประเภทข่าวจริงจำนวน 827,073 ข้อความ (คิดเป็นร้อยละ 87.21 จากหัวข้อข่าวที่เก็บมาทั้งหมด) ข่าวปลอมจำนวน 121,300 ข้อความ (คิดเป็นร้อยละ 12.79 จากหัวข้อข่าวที่เก็บมาทั้งหมด) ซึ่งรายละเอียดทั้งหมดแสดงดังตารางที่ 3.3

**ตารางที่ 3.3** รายละเอียดจำนวนข่าวที่มีการจัดเก็บข้อมูล

| หัวข้อข่าว | จำนวนข่าว | | | คิดเป็นร้อยละ | |
|---|---|---|---|---|---|
| | รวม | ข่าวจริง | ข่าวปลอม | ข่าวจริง | ข่าวปลอม |
| ดอกไม้จันทน์, ก้านดอกไม้จันทน์ดำเป็นขาว, ดอกไม้เพื่อพ่อ, ปากคลองตลาด, สนามหลวง, ถวายอาลัย, น้อมส่งเสด็จสู่สวรรคาลัย, ส่งเสด็จสู่สวรรคาลัย, สู่ฟ้าเสวยสวรรค์, รัชกาลที่ 9, พระราชพิธีถวายพระเพลิง, สถิตในดวงใจนิรันดร์, นิทรรศการงานพระราชพิธี, เครื่องราชอิสริยาภรณ์, น้ำมันฟรี, สลับผู้ว่า ชลบุรี นนทบุรี | 363,639 | 357,485 | 6,154 | 98.31 | 1.69 |
| น้ำท่วม, เขื่อนแตก, น้ำเหนือมา, ปล่อยน้ำท่วมรังสิต, ฝนตก, พายุขนุน, ไต้ฝุ่นล้ง, หนาว, อากาศหนาว, แผ่นดินไหว | 361,751 | 254,777 | 106,974 | 70.43 | 29.57 |
| ก้าวคนละก้าว, กินเจ, เลือกตั้งญี่ปุ่น, อุบัติเหตุ, เจ้าชายซาอุฯ, มาร์คเยือนไทย, ร้านสะดวกซื้อ ขายเบียร์สด, เพิ่มเงินสมทบประกันสังคม, ช้อปช่วยชาติ ลดหย่อนภาษี, อาบอบนวด ลดหย่อนภาษี | 222,983 | 214,811 | 8,172 | 96.34 | 3.66 |
| **จำนวนรวม** | **948,373** | **827,073** | **121,300** | **87.21** | **12.79** |

จากตารางที่ 3.3 หัวข้อข่าวที่จัดเก็บมาในกลุ่มแรก ได้แก่ ข่าวที่เกี่ยวข้องกับพระมหากษัตริย์ เช่น ดอกไม้จันทน์, ก้านดอกไม้จันทน์ดำเป็นขาว, ดอกไม้เพื่อพ่อ, ปากคลองตลาด, สนามหลวง, ถวายอาลัย, น้อมส่งเสด็จสู่สวรรคาลัย, ส่งเสด็จสู่สวรรคาลัย, สู่ฟ้าเสวยสวรรค์ , รัชกาลที่ 9, พระราชพิธีถวายพระเพลิง, สถิตในดวงใจนิรันดร์, นิทรรศการงานพระราชพิธี, เครื่องราชอิสริยาภรณ์, น้ำมันฟรี, สลับผู้ว่า ชลบุรี นนทบุรี จำนวนข่าวรวม 363,639 ข้อความ แบ่งเป็นข่าวจริง 357,485 ข้อความ (คิดเป็นร้อยละ 98.31 จากหัวข้อข่าวที่เก็บมา) ข่าวปลอม 6,154 ข้อความ (คิดเป็นร้อยละ 1.69 จากหัวข้อข่าวที่เก็บมา) หัวข้อข่าวกลุ่มที่สอง ได้แก่ น้ำท่วม, เขื่อนแตก, น้ำเหนือมา, ปล่อยน้ำท่วมรังสิต, ฝนตก, พายุขนุน, ไต้ฝุ่นล้ง, หนาว, อากาศหนาว, แผ่นดินไหว จำนวนข่าวรวม 361,751 ข้อความ

แบ่งเป็นข่าวจริง 254,777 ข้อความ (คิดเป็นร้อยละ 70.43 จากหัวข้อข่าวที่เก็บมา) ข่าวปลอม 106,974 ข้อความ (คิดเป็นร้อยละ 29.57 จากหัวข้อข่าวที่เก็บมา) และหัวข้อข่าวกลุ่มสุดท้าย ได้แก่ ก้าวคนละก้าว, กินเจ, เลือกตั้งญี่ปุ่น, อุบัติเหตุ, เจ้าชายซาอุฯ, มาร์ค เยือนไทย, ร้านสะดวกซื้อ ขายเบียร์สด, เพิ่มเงินสมทบประกันสังคม, ช้อปช่วยชาติ ลดหย่อนภาษี, อาบอบนวด ลดหย่อนภาษี จำนวนข่าวรวม 222,983 ข้อความ แบ่งเป็นข่าวจริง 214,811 ข้อความ (คิดเป็นร้อยละ 96.34 จากหัวข้อข่าวที่เก็บมา) ข่าวปลอม 8,172 ข้อความ (คิดเป็นร้อยละ 3.66 จากหัวข้อข่าวที่เก็บมา)



การเก็บรวบรวมข้อมูลจากเครือข่ายสังคมออนไลน์ทวิตเตอร์

เริ่มจากการกำหนดค่าคุณลักษณะต่าง ๆ ที่ต้องการเรียกเก็บจาก Twitter API และ ระบุคำสำคัญที่เป็นหัวข้อข่าวที่ต้องการเก็บข้อมูล โดยการเก็บข้อมูลจากทวิตเตอร์มีข้อจำกัดในสิทธิการเข้าถึงข้อมูลคือสำหรับแต่ละค่าโทเคน (token) จะสามารถใช้เรียกเก็บข้อมูลได้ทุก ๆ 15 นาทีต่อครั้ง และเรียกข้อความได้มากที่สุดครั้งละไม่เกิน 20,000 ข้อความ การเรียกข้อมูลย้อนหลังทำได้ไม่เกิน 7 วัน ดังนั้นหากต้องการข้อมูลที่มีคำสำคัญเดียวกันมาก ๆ จำเป็นต้องมีการเรียกเก็บข้อมูลด้วยคำสำคัญเดิมซ้ำกันหลายครั้ง แล้วจึงนำไปทำความสะอาดข้อมูลโดยการตัดรายการที่ซ้ำออกในภายหลัง

วิธีการพิจารณาข่าวจริงข่าวปลอมของข้อความที่ได้จากเครือข่ายสังคมออนไลน์ทวิตเตอร์

เมื่อพิจารณาข้อความที่ได้จากเครือข่ายสังคมออนไลน์ทวิตเตอร์ และตรวจพบสิ่งที่ไม่ใช่ข่าว ซึ่งไม่ใช่ข้อความที่ต้องการใช้ในงานวิจัย จะลบข้อความเหล่านี้ทิ้งไป ซึ่งสิ่งที่ไม่ใช่ข่าวมีลักษณะดังต่อไปนี้

– คำบ่น คำพูดกล่าวขึ้นมาลอย ๆ ไม่มีที่มาที่ไป
– ประโยคคำถาม
– คำโฆษณาชวนเชื่อ
– การแสดงความคิดเห็นที่เกี่ยวข้องกับประเด็นข่าว
– คำชักชวน ปลุกระดม ไปร่วมทำกิจกรรมต่าง ๆ
– การโจมตี กล่าวหาให้ร้ายบุคคลอื่น/องค์กรต่าง ๆ
– คำหยาบ คำไม่สุภาพ คำสบถ คำด่า คำที่มีความหมายเชิงหมิ่นเหม่ในเรื่องเพศ คำพูดที่ต้องตีความในลักษณะต้องคิดลึกจึงจะเข้าใจความหมาย
– การใช้คำที่เป็นเทรนด์ แต่ข้อความในเนื้อหาไม่มีความเกี่ยวข้องกับ hashtag ที่อ้างถึง
– คำบัญญัติเองเป็นคำที่ไม่มีความหมายตามพจนานุกรมฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๕๔

การตรวจสอบข้อความว่าเป็นข่าวปลอมพิจารณาจาก

– ตรวจสอบข้อมูลจากแหล่งข่าวในประเด็นที่เนื้อหาข่าวอ้างถึงจากหลาย ๆ แหล่ง
– หน่วยงานที่รับผิดชอบที่เกี่ยวข้องกับประเด็นเนื้อหา
– ผู้ที่มีความเกี่ยวข้อง/ถูกพาดพิงในประเด็นเนื้อหาข่าว
– แหล่งข่าวอื่น ๆ เพื่อเปรียบเทียบเนื้อหาตรงกันหรือไม่

– ใช้เครื่องมือค้นหา (search engine) ค้นหาประเด็นที่เกี่ยวข้อง พิจารณาความสัมพันธ์กับข่าวที่สนใจ

– เปรียบเทียบรูปภาพ แหล่งอ้างอิงที่ปรากฏในเนื้อหา เป็นเรื่องเก่าแล้วนำมาเล่าใหม่ หรือมีความเกี่ยวข้องกับประเด็นที่ข่าวอ้าง/กล่าวถึง

– เว็บไซต์แหล่งข่าวที่มีการตรวจสอบข้อมูลโดยผู้เชี่ยวชาญยืนยันหรือหน่วยงานผู้รับผิดชอบ

– ความสัมพันธ์ระหว่างเวลากับเหตุการณ์ที่กล่าวถึงในเนื้อหามีความสมเหตุสมผลเข้ากันหรือไม่

<u>ข้อระวังในการแยกข่าวปลอม</u>

– ข้อความที่เลือกมาจากหัวข้อข่าวที่มีคำสำคัญเหมือนกัน แต่อาจมีรายละเอียดเนื้อหาไม่ตรงกับหัวข้อข่าวนั้น

– บางข้อความมีคำสำคัญเหมือนกัน แต่มีคำแสดงความขัดแย้งกันในเนื้อหา หรือคำในเชิงปฏิเสธ จำต้องพิจารณาความหมายของคำเชิงปฏิเสธในเนื้อหาประกอบด้วย เช่น ข่าว...เป็นข่าวปลอม เรื่อง....เป็นเรื่องลวง ไม่มีเรื่อง...เกิดขึ้นจริง เนื้อหา...ไม่เป็นความจริง

ส่วนความหมายของแต่ละคุณลักษณะข้อมูลที่เลือกใช้ในชุดข้อมูล สามารถอธิบายรายละเอียดได้ดังนี้

| *ชื่อคุณลักษณะที่ใช้* | *ความหมาย* |
|---|---|
| Id | รหัสผู้ใช้เป็นตัวเลขที่ทวิตเตอร์ใช้ในการจำแนกแต่ละบัญชีผู้ใช้งาน |
| Name | ชื่อผู้ใช้งานทวิตเตอร์ |
| IsVerified | คุณลักษณะการยืนยันตัวบุคคลกับทางทวิตเตอร์ของผู้ใช้ |
| ProfileImageUrl | ตำแหน่งที่จัดเก็บ URL ที่เชื่อมโยงไปยังรูปภาพประวัติส่วนตัวของผู้ใช้ |
| FollowersCount | จำนวนผู้ที่ติดตามบัญชีผู้ใช้งานนี้ |
| FriendsCount | จำนวนที่เจ้าของบัญชีนี้กำลังติดตาม |
| FavouritesCount | จำนวนเจ้าของบัญชีนี้เคยกดถูกใจ |
| StatusesCount | จำนวนสถานะที่เคยโพสต์ของบัญชีนี้ |
| Description | รายละเอียดของผู้ใช้ที่ต้องการอธิบายเกี่ยวกับตัวเอง |
| Location | สถานที่ที่ผู้ใช้โพสต์ข้อความ |
| TimeZone | เขตเวลาที่บัญชีผู้ใช้นั้นถูกสร้างขึ้น |
| CreatedDate | วันที่ผู้ใช้สร้างบัญชีผู้ใช้งานขึ้นมา |
| Status | คุณลักษณะที่ผู้ใช้บ่งบอกสถานะปัจจุบันของบัญชีผู้ใช้งาน |

| | |
|---|---|
| Url | ตำแหน่งการเชื่อมโยงไปยังสื่อปลายทางที่ข้อความนั้นต้องการระบุถึง |
| Mentions | รายละเอียดของชื่อผู้ใช้ที่ถูกกล่าวถึงในข้อความ ปรากฏหลังสัญลักษณ์ @ |
| NumberofMentions | จำนวนสัญลักษณ์ @ ที่ปรากฏในข้อความ ได้ค่าจาก Twitter API |
| HashTags | รายละเอียดพิเศษที่ผู้ใช้ต้องการอธิบายหรือระบุเฉพาะเจาะจงในหัวข้อเรื่องของข้อความ ปรากฏสัญลักษณ์ #หน้าข้อความพิเศษ |
| NumberofHashTags | จำนวนสัญลักษณ์ # ที่ปรากฏในข้อความ ได้ค่าจาก Twitter API |
| RetweetCount | จำนวนครั้งที่ข้อความถูกนำไปเผยแพร่ซ้ำอีกครั้งโดยบุคคลอื่นที่ไม่ใช่คนโพสต์ข้อความครั้งแรก |
| TwittCreatedDate | วันที่เวลาที่ข้อความนั้นถูกสร้างขึ้น |
| MessageText | รายละเอียดของข้อความ |
| MessageImage | ตำแหน่งที่อยู่การเชื่อมโยงไปยังรูปภาพที่เกี่ยวข้อง |

ส่วนต่อไปจะเป็นการอธิบายรายละเอียดของชุดข้อมูลที่ใช้กับคุณลักษณะที่เลือกใช้ เมื่อนำชุดข้อมูลข่าวที่จัดเก็บมาได้ตามรายละเอียดในตารางที่ 3.3 มาผ่านกระบวนการปรับค่าข้อมูลให้เป็นตัวเลขทั้งหมดตามเงื่อนไขที่กล่าวมาในบทก่อนหน้า แล้วจึงนำข้อมูลที่ได้มาขจัดความซ้ำซ้อนของข้อมูล โดยลบข้อมูลส่วนที่ซ้ำกันออก เหลือเฉพาะข้อมูลส่วนที่มีค่าไม่ซ้ำกันเลย จำนวน 327,784 ข้อความ และเมื่อนำชุดข้อมูลนี้มาพิจารณาแต่ละคุณลักษณะที่ได้เลือกใช้ จะปรากฏรายละเอียดดังต่อไปนี้

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Id พบว่าลักษณะของข่าวที่นำมาใช้ส่วนใหญ่เป็นบัญชีผู้ใช้ที่มีรหัสผู้ใช้งาน (Id) ที่มีการใช้งานมานานแล้วซึ่งใช้รหัสผู้ใช้ขนาด 9-10 หลัก มากกว่า 18 หลัก โดยในข่าวจริงเป็นข่าวของบัญชีผู้ใช้ที่มีรหัสผู้ใช้ขนาด 9-10 หลัก จำนวนร้อยละ 71.75 และในส่วนของข่าวปลอมเป็นข่าวของบัญชีผู้ใช้ที่มีรหัสผู้ใช้ขนาด 9-10 หลัก จำนวนร้อยละ 89.35 ดังรายละเอียดในตารางที่ 3.4

**ตารางที่ 3.4** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Id

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Id | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | 9-10 หลัก | 71.75 | 89.35 |
| | 18 หลัก | 28.25 | 10.65 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Name พบว่าลักษณะของข่าวที่นำมาใช้ส่วนใหญ่เป็นบัญชีผู้ใช้ที่มีการใช้ชื่อผู้ใช้งาน (Name) เป็นตัวอักขระผสมที่เป็นภาษาไทยหรือภาษาอังกฤษหรือตัวเลขผสมกันมากกว่ารูปแบบอื่น ๆ กล่าวคือ ในบัญชีที่มีการแสดงข่าวจริงเป็นบัญชีชื่อตัวอักขระผสมภาษาไทยหรือภาษาอังกฤษหรือตัวเลข จำนวนร้อยละ 74.14 บัญชีชื่อตัวอักขระภาษาไทย จำนวนร้อยละ 23.21 บัญชีชื่อตัวอักขระภาษาอังกฤษ จำนวนร้อยละ 2.66 ไม่พบชื่อบัญชีผู้ใช้ที่เป็นตัวอักขระภาษาอื่น ๆ และสัญลักษณ์อื่น ๆ และในส่วนของข่าวปลอมมีลักษณะการใช้ชื่อบัญชีเช่นเดียวกับบัญชีข่าวจริง กล่าวคือ ในบัญชีที่มีการแสดงข่าวปลอมเป็นบัญชีชื่อตัวอักขระผสมภาษาไทยหรือภาษาอังกฤษหรือตัวเลข จำนวนร้อยละ 76.04 บัญชีชื่อตัวอักขระภาษาไทย จำนวนร้อยละ 20.41 บัญชีชื่อตัวอักขระภาษาอังกฤษ จำนวนร้อยละ 3.55 และไม่พบชื่อบัญชีผู้ใช้ที่เป็นตัวอักขระภาษาอื่น ๆ และสัญลักษณ์อื่น ๆ เช่นเดียวกับบัญชีผู้ใช้ที่แสดงข่าวจริง ดังรายละเอียดในตารางที่ 3.5

**ตารางที่ 3.5** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Name

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Name | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | ตัวอักขระภาษาไทยทั้งหมด | 23.21 | 20.41 |
| | ตัวอักขระภาษาอังกฤษทั้งหมด | 2.66 | 3.55 |
| | ตัวอักขระภาษาไทยหรือภาษาอังกฤษหรือตัวเลข | 74.14 | 76.04 |
| | ตัวอักขระภาษาอื่น ๆ | 0.00 | 0.00 |
| | สัญลักษณ์อื่น ๆ | 0.00 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ IsVerified พบว่าลักษณะของข่าวที่นำมาใช้ส่วนใหญ่เป็นบัญชีผู้ใช้ที่มีคุณลักษณะ IsVerified เป็น FALSE แสดงว่าไม่มีการยืนยันตัวตนทั้งในส่วนของการแสดงข่าวจริงและข่าวปลอม โดยบัญชีผู้ใช้ที่แสดงข่าวจริง จำนวนร้อยละ 99.22 ไม่มีการยืนยันตัวตน และบัญชีผู้ใช้ที่แสดงข่าวปลอมทั้งหมด (ร้อยละ 100) ไม่มีการยืนยันตัวตน ดังรายละเอียดในตารางที่ 3.6

**ตารางที่ 3.6** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ IsVerified

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| IsVerified | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | TRUE | 0.78 | 0.00 |
| | FALSE | 99.22 | 100.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ ProfileImageUrl ซึ่งเป็นรายละเอียดลิงค์ของภาพประวัติของบัญชีผู้ใช้งานทวิตเตอร์ พบว่าโดยส่วนใหญ่เป็นบัญชีผู้ใช้มีการใช้ภาพประวัติเป็นไฟล์ .jpg ทั้งในส่วนของบัญชีผู้ใช้ที่แสดงข่าวจริงจำนวนร้อยละ 94.92 และบัญชีผู้ใช้ที่แสดงข่าวปลอมจำนวนร้อยละ 92.01 โดยมีบางส่วนใช้ภาพประวัติผู้ใช้เป็นไฟล์ .png จำนวนร้อยละ 2.00 ในบัญชีผู้ใช้ที่แสดงข่าวจริง และจำนวนร้อยละ 1.48 ในบัญชีผู้ใช้ที่แสดงข่าวปลอม และที่เหลือเป็นรูปแบบไฟล์รูปภาพอื่น ๆ โดยที่บัญชีผู้ใช้ทั้งหมดไม่มีบัญชีใดที่ไม่ปรากฏข้อมูลรูปภาพประวัติผู้ใช้ ดังรายละเอียดในตารางที่ 3.7

**ตารางที่ 3.7** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ ProfileImageUrl

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| ProfileImageUrl | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | .jpg | 94.92 | 92.01 |
| | .png | 2.00 | 1.48 |
| | อื่น ๆ | 3.08 | 6.51 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ FollowersCount คือ จำนวนผู้ที่ติดตามบัญชี ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงพบว่าร้อยละ 41.49 มีจำนวนผู้ติดตามระหว่าง 100-999 บัญชี ร้อยละ 35.04 มีจำนวนผู้ติดตามระหว่าง 10-99 บัญชี ร้อยละ 10.14 มีจำนวนผู้ติดตามระหว่าง 1,000-9,999 บัญชี ร้อยละ 9.48 มีจำนวนผู้ติดตามระหว่าง 1-9 บัญชี ร้อยละ 1.82 ไม่มีผู้ติดตาม ร้อยละ 1.40 มีจำนวนผู้ติดตามระหว่าง 10,000-99,999 บัญชี ร้อยละ 0.42 มีจำนวนผู้ติดตามระหว่าง 100,000-999,999 บัญชี และร้อยละ 0.21 มีจำนวนผู้ติดตามระหว่าง 1,000-9,999 บัญชี ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าร้อยละ 48.82 มีจำนวนผู้ติดตามระหว่าง 100-999 บัญชี ร้อยละ 17.75 มีจำนวนผู้ติดตามระหว่าง 10-99 บัญชี ร้อยละ 15.98 มีจำนวนผู้ติดตามระหว่าง 1,000-9,999 บัญชี ร้อยละ 7.10 มีจำนวนผู้ติดตามระหว่าง 10,000-99,999 บัญชี ร้อยละ 3.55 มีจำนวนผู้ติดตามระหว่าง 1,000,000-9,999,999 บัญชี จำนวนผู้ติดตามระหว่าง 100,000-999,999 บัญชีเท่ากับจำนวนผู้ติดตามระหว่าง 1-9 บัญชีคือ ร้อยละ 3.25 และร้อยละ 0.30 ไม่มีผู้ติดตาม ดังรายละเอียดในตารางที่ 3.8

**ตารางที่ 3.8** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ FollowersCount

| **คุณลักษณะ** | **รายละเอียด** | **ร้อยละของข้อมูล** | |
|---|---|---|---|
| | | **ข่าวจริง** | **ข่าวปลอม** |
| FollowersCount | ไม่ปรากฏข้อมูล | 1.82 | 0.30 |
| | 1-9 | 9.48 | 3.25 |
| | 10-99 | 35.04 | 17.75 |
| | 100-999 | 41.49 | 48.82 |
| | 1,000-9,999 | 10.14 | 15.98 |
| | 10,000-99,999 | 1.40 | 7.10 |
| | 100,000-999,999 | 0.42 | 3.25 |
| | 1,000,000-9,999,999 | 0.21 | 3.55 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ FriendsCount คือจำนวนที่ผู้ใช้บัญชีกำลังติดตามซึ่งเท่ากับจำนวน Follower ที่ปรากฏในหน้าบัญชีผู้ใช้ แต่การเรียกค่าผ่าน Twitter API จะใช้คำว่า Friend แทน Follower ซึ่งในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงพบว่า

ร้อยละ 63.33 มีจำนวนผู้กำลังติดตามระหว่าง 100-999 บัญชี ร้อยละ 25.14 มีจำนวนผู้กำลังติดตามระหว่าง 10-99 บัญชี ร้อยละ 9.98 มีจำนวนผู้กำลังติดตามระหว่าง 1,000-9,999 บัญชี ร้อยละ 1.30 มีจำนวนผู้กำลังติดตามระหว่าง 1-9 บัญชี ร้อยละ 0.23 ไม่มีผู้กำลังติดตาม และร้อยละ 0.03 มีจำนวนผู้กำลังติดตามระหว่าง 10,000-99,999 บัญชี ไม่พบบัญชีผู้ใช้ที่มีการแสดงข่าวมีจำนวนผู้กำลังติดตามมากกว่า 100,000 บัญชีขึ้นไป ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าร้อยละ 65.38 มีจำนวนผู้กำลังติดตามระหว่าง 100-999 บัญชี ร้อยละ 18.64 มีจำนวนผู้กำลังติดตามระหว่าง 10-99 บัญชี ร้อยละ 13.61 มีจำนวนผู้กำลังติดตามระหว่าง 1,000-9,999 บัญชี ร้อยละ 1.48 มีจำนวนผู้กำลังติดตามระหว่าง 1-9 บัญชี ร้อยละ 0.89 ไม่มีผู้กำลังติดตาม และไม่พบบัญชีผู้ใช้ที่มีการแสดงข่าวปลอมมีจำนวนผู้กำลังติดตามตั้งแต่ 10,000 บัญชีขึ้นไป ดังรายละเอียดในตารางที่ 3.9

**ตารางที่ 3.9** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ FriendsCount

| **คุณลักษณะ** | **รายละเอียด** | **ร้อยละของข้อมูล** | |
|---|---|---|---|
| | | **ข่าวจริง** | **ข่าวปลอม** |
| FriendsCount | ไม่ปรากฏข้อมูล | 0.23 | 0.89 |
| | 1-9 | 1.30 | 1.48 |
| | 10-99 | 25.14 | 18.64 |
| | 100-999 | 63.33 | 65.38 |
| | 1,000-9,999 | 9.98 | 13.61 |
| | 10,000-99,999 | 0.03 | 0.00 |
| | 100,000-999,999 | 0.00 | 0.00 |
| | 1,000,000-9,999,999 | 0.00 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ FavouritesCount คือจำนวนสิ่งที่สนใจของเจ้าของบัญชีหรือจำนวนที่เจ้าของบัญชีเคยไปกดถูกใจ ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงพบว่าร้อยละ 37.55 มีจำนวนสิ่งที่สนใจ 1,000-9,999 ครั้ง ร้อยละ 28.75 มีจำนวนสิ่งที่สนใจระหว่าง 100-999 ครั้ง ร้อยละ 17.91 มีจำนวนสิ่งที่สนใจระหว่าง 10,000-99,999 ครั้ง ร้อยละ 10.45 มีจำนวนสิ่งที่สนใจระหว่าง 10-99 ครั้ง ร้อยละ 3.00 มีจำนวนสิ่งที่สนใจระหว่าง 1-9

ครั้ง ร้อยละ 1.24 มีจำนวนสิ่งที่สนใจระหว่าง 100,000-999,999 ครั้ง ร้อยละ 1.10 ไม่มีมีสิ่งที่สนใจ และไม่มีบัญชีใดที่มีจำนวนสิ่งที่สนใจมากกว่า 1,000,000 ครั้งขึ้นไป ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าร้อยละ 34.62 มีจำนวนสิ่งที่สนใจระหว่าง 1,000-9,999 ครั้ง ร้อยละ 24.56 มีจำนวนสิ่งที่สนใจระหว่าง 100-999 ครั้ง ร้อยละ 15.38 มีจำนวนสิ่งที่สนใจระหว่าง 10,000-99,999 ครั้ง ร้อยละ 10.95 มีจำนวนสิ่งที่สนใจระหว่าง 10-99 ครั้ง ร้อยละ 7.69 มีจำนวนสิ่งที่สนใจระหว่าง 1-9 ครั้ง ร้อยละ 1.78 มีจำนวนสิ่งที่สนใจระหว่าง 100,000-999,999 ครั้ง และไม่มีบัญชีใดที่มีจำนวนสิ่งที่สนใจมากกว่า 1,000,000 ครั้งขึ้นไป ดังรายละเอียดในตารางที่ 3.10

**ตารางที่ 3.10** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ FavouritesCount

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| FavouritesCount | ไม่ปรากฏข้อมูล | 1.10 | 5.03 |
| | 1-9 | 3.00 | 7.69 |
| | 10-99 | 10.45 | 10.95 |
| | 100-999 | 28.75 | 24.56 |
| | 1,000-9,999 | 37.55 | 34.62 |
| | 10,000-99,999 | 17.91 | 15.38 |
| | 100,000-999,999 | 1.24 | 1.78 |
| | 1,000,000-9,999,999 | 0.00 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ StatusesCount คือจำนวนสถานะของเจ้าของบัญชีได้ทวีตออกไปแล้วนับแต่เปิดบัญชีผู้ใช้งาน ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงพบว่าร้อยละ 44.40 มีจำนวนสถานะ 10,000-99,999 สถานะ ร้อยละ 25.97 มีจำนวนสถานะระหว่าง 1,000-9,999 สถานะ ร้อยละ 18.76 มีจำนวนสถานะระหว่าง 100,000-999,999 สถานะ ร้อยละ 8.37 มีจำนวนสถานะระหว่าง 100-999 สถานะ ร้อยละ 2.05 มีจำนวนสถานะระหว่าง 10-99 สถานะ ร้อยละ 0.39 มีจำนวนสถานะระหว่าง 1-9 สถานะ ร้อยละ 0.04 มีจำนวนสถานะระหว่าง 1,000,000-9,999,999 สถานะ และไม่มีบัญชีใดเลยที่ไม่มีมีสถานะ ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าร้อยละ 37.87 มีจำนวนสถานะระหว่าง 10,000-99,999 สถานะ

ร้อยละ 34.02 มีจำนวนสถานะระหว่าง 100,000-999,999 สถานะ ร้อยละ 20.71 มีจำนวนสถานะระหว่าง 1,000-9,999 สถานะ ร้อยละ 4.14 มีจำนวนสถานะระหว่าง 100-999 สถานะ ร้อยละ 1.48 มีจำนวนสถานะระหว่าง 10-99 สถานะเท่ากับร้อยละ 1.48 มีจำนวนสถานะระหว่าง 1,000,000-9,999,999 สถานะ ร้อยละ 0.30 มีจำนวนสถานะระหว่าง 1-9 สถานะ และไม่มีบัญชีใดเลยที่ไม่มีมีสถานะ ดังรายละเอียดในตารางที่ 3.11



**ตารางที่ 3.11** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ StatusesCount

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| StatusesCount | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | 1-9 | 0.39 | 0.30 |
| | 10-99 | 2.05 | 1.48 |
| | 100-999 | 8.37 | 4.14 |
| | 1,000-9,999 | 25.97 | 20.71 |
| | 10,000-99,999 | 44.40 | 37.87 |
| | 100,000-999,999 | 18.76 | 34.02 |
| | 1,000,000-9,999,999 | 0.04 | 1.48 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Description พบว่าลักษณะของบัญชีผู้ใช้ที่มีการใช้รายละเอียดของผู้ใช้งาน (Description) ในบัญชีที่มีการแสดงข่าวจริงจำนวนร้อยละ 99.73 เป็นตัวอักขระผสมภาษาไทยหรือภาษาอังกฤษหรือตัวเลขผสมกันมากกว่ารูปแบบอื่นที่มีเพียงร้อยละ 0.27 ที่เป็นบัญชีรายละเอียดที่เป็นตัวอักขระภาษาอังกฤษทั้งหมด และไม่พบรายละเอียดของบัญชีผู้ใช้ที่เป็นตัวอักขระภาษาไทยทั้งหมด ภาษาอื่น ๆ และสัญลักษณ์อื่น ๆ ส่วนของข่าวปลอมมีลักษณะการใช้รายละเอียดของบัญชี โดยมีลักษณะคล้ายกับบัญชีข่าวจริงโดยใช้ตัวอักขระผสมภาษาไทยหรือภาษาอังกฤษหรือตัวเลขทั้งหมดร้อยละ 100 โดยไม่พบบัญชีผู้ใช้ที่มีรายละเอียดในแบบอื่น ๆ ดังรายละเอียดในตารางที่ 3.12

**ตารางที่ 3.12** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Description

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Description | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | ตัวอักขระภาษาไทยทั้งหมด | 0.00 | 0.00 |
| | ตัวอักขระภาษาอังกฤษทั้งหมด | 0.27 | 0.00 |
| | ตัวอักขระภาษาไทยหรือภาษาอังกฤษหรือตัวเลข | 99.73 | 100.00 |
| | ตัวอักขระภาษาอื่น ๆ | 0.00 | 0.00 |
| | สัญลักษณ์อื่น ๆ | 0.00 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Location พบว่าลักษณะของตำแหน่ง (Location) ของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงจำนวนร้อยละ 43.80 เป็นสถานที่อื่น ๆ ที่ไม่ตรงกับข้อมูลในตารางที่ 4.11 ร้อยละ 33.06 แสดงตำแหน่งบัญชีผู้ใช้อยู่ในประเทศไทย ร้อยละ 17.83 แสดงตำแหน่งบัญชีผู้ใช้ในเอเชียตะวันออกเฉียงใต้ยกเว้นประเทศไทย ร้อยละ 2.32 แสดงตำแหน่งบัญชีผู้ใช้ในเอเชียยกเว้นเอเชียตะวันออกเฉียงใต้ ร้อยละ 2.08 แสดงตำแหน่งบัญชีผู้ใช้ที่อยู่ในสหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย ร้อยละ 0.63 แสดงตำแหน่งบัญชีผู้ใช้ในยุโรป/รัสเซีย ร้อยละ 0.16 แสดงตำแหน่งบัญชีผู้ใช้ใน ออสเตรเลีย/นิวซีแลนด์ ร้อยละ 0.12 ไม่ปรากฏตำแหน่งของบัญชีผู้ใช้ และร้อยละ 0.02 แสดงตำแหน่งบัญชีผู้ใช้ในแอฟริกา ส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวปลอมจำนวนร้อยละ 37.87 เป็นสถานที่อื่น ๆ ที่ไม่ตรงกับข้อมูลในตารางที่ 4.11 ร้อยละ 29.29 แสดงตำแหน่งบัญชีผู้ใช้ที่อยู่ในประเทศไทย ร้อยละ 27.81 แสดงตำแหน่งในเอเชียตะวันออกเฉียงใต้ยกเว้นประเทศไทยร้อยละ 2.66 แสดงตำแหน่งบัญชีผู้ใช้ในเอเชียยกเว้นเอเชียตะวันออกเฉียงใต้ ร้อยละ 1.78 แสดงตำแหน่งบัญชีผู้ใช้ในสหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย ร้อยละ 0.59 แสดงตำแหน่งบัญชีผู้ใช้ในยุโรป/รัสเซีย ไม่พบบัญชีผู้ใช้ที่อยู่ในออสเตรเลีย/นิวซีแลนด์ หรือแอฟริกา และบัญชีผู้ใช้ไม่ปรากฏข้อมูลตำแหน่งของบัญชีผู้ใช้ ดังรายละเอียดในตารางที่ 3.13

**ตารางที่ 3.13** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Location

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Location | ไม่ปรากฏข้อมูล | 0.12 | 0.00 |
| | ประเทศไทย | 33.06 | 29.29 |
| | เอเชียตะวันออกเฉียงใต้ยกเว้นประเทศไทย | 17.83 | 27.81 |
| | เอเชียยกเว้นเอเชียตะวันออกเฉียงใต้ | 2.32 | 2.66 |
| | ออสเตรเลีย/นิวซีแลนด์ | 0.16 | 0.00 |
| | ยุโรป/รัสเซีย | 0.63 | 0.59 |
| | สหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย | 2.08 | 1.78 |
| | แอฟริกา | 0.02 | 0.00 |
| | อื่น ๆ | 43.80 | 37.87 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ TimeZone ที่แสดงตำแหน่งตามการแบ่งเส้นเขตเวลาโลก พบว่าทั้งหมดของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงและบัญชีผู้ใช้ที่มีการแสดงข่าวปลอมไม่ปรากฏข้อมูลแสดงตำแหน่งตามการแบ่งเส้นเขตเวลาโลก แม้ว่าจะมีเงื่อนไขการพิจารณาเวลาเช่นเดียวกับตำแหน่งสถานที่โดยแบ่งเป็นส่วนย่อย ๆ ตามภูมิภาคเช่น ประเทศไทย เอเชียตะวันออกเฉียงใต้ยกเว้นประเทศไทย เอเชียยกเว้นเอเชียตะวันออกเฉียงใต้ ออสเตรเลีย/นิวซีแลนด์ ยุโรป/รัสเซีย สหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย แอฟริกา และอื่น ๆ รายละเอียดดังตารางที่ 3.14

**ตารางที่ 3.14** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ TimeZone

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| TimeZone | ไม่ปรากฏข้อมูล | 100.00 | 100.00 |
| | ประเทศไทย | 0.00 | 0.00 |
| | เอเชียตะวันออกเฉียงใต้ยกเว้นประเทศไทย | 0.00 | 0.00 |
| | เอเชียยกเว้นเอเชียตะวันออกเฉียงใต้ | 0.00 | 0.00 |
| | ออสเตรเลีย/นิวซีแลนด์ | 0.00 | 0.00 |
| | ยุโรป/รัสเซีย | 0.00 | 0.00 |
| | สหรัฐอเมริกา/แคนาดา/อลาสก้า/ฮาวาย | 0.00 | 0.00 |
| | แอฟริกา | 0.00 | 0.00 |
| | อื่น ๆ | 0.00 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ CreatedDate ที่แสดงวันที่บัญชีผู้ใช้ทวิตเตอร์ถูกสร้างขึ้น สำหรับในงานวิจัยนี้ได้นำข้อมูลวันที่บัญชีผู้ใช้ทวิตเตอร์ถูกสร้างขึ้นมาประมวลผลข้อมูลโดยการคำนวณเป็นอายุการใช้งานของบัญชีที่สร้างขึ้น และได้แบ่งช่วงการแสดงผลช่วงละ 0.5 ปี พบว่าบัญชีผู้ใช้ที่มีการแสดงข่าวจริงร้อยละ 18.68 มีอายุการใช้งานน้อยกว่า 0.5 ปี ร้อยละ 12.62 มีอายุการใช้งานระหว่าง 0.5 ปี ถึง 1 ปี ร้อยละ 11.19 มีอายุการใช้งานระหว่าง 3 ปี ถึง 3.5 ปี ร้อยละ 10.24 มีอายุการใช้งานระหว่าง 1.5 ปี ถึง 2 ปี ร้อยละ 10.06 มีอายุการใช้งานระหว่าง 3.5 ปี ถึง 4 ปี ร้อยละ 9.78 มีอายุการใช้งานระหว่าง 1 ปี ถึง 1.5 ปี ร้อยละ 9.31 มีอายุการใช้งานระหว่าง 2 ปี ถึง 2.5 ปี ร้อยละ 8.71 มีอายุการใช้งานระหว่าง 4 ปี ถึง 4.5 ปี ร้อยละ 0.35 มีอายุการใช้งานระหว่าง 4.5 ปี ถึง 5 ปี ร้อยละ 0.13 มีอายุการใช้งานระหว่าง 5 ปี ถึง 5.5 ปี และไม่พบบัญชีผู้ใช้ที่มีอายุการใช้งานมากกว่า 5.5 ปีขึ้นไป ส่วนบัญชีผู้ใช้ที่มีการแสดงข่าวปลอมร้อยละ 29.68 มีอายุการใช้งานระหว่าง 4 ปี ถึง 4.5 ปี ร้อยละ 16.77 มีอายุการใช้งานระหว่าง 3.5 ปี ถึง 4 ปี ร้อยละ 14.84 มีอายุการใช้งานระหว่าง 3 ปี ถึง 3.5 ปี ร้อยละ 12.26 มีอายุการใช้งานระหว่าง 2.5 ปี ถึง 3 ปี ร้อยละ 9.03 มีอายุการใช้งานน้อยกว่า 0.5 ปี ร้อยละ 5.81 มีอายุการใช้งานระหว่าง 2 ปี ถึง 2.5 ปี ร้อยละ 5.16 มีอายุการใช้งานระหว่าง 0.5 ปี ถึง 1 ปี ร้อยละ 3.87 มีอายุการใช้งานระหว่าง 1.5 ปี ถึง 2 ปี ร้อยละ 2.58 มีอายุการใช้งานระหว่าง 1 ปี ถึง 1.5 ปี และไม่พบบัญชีผู้ใช้ที่มีอายุการใช้งานมากกว่า 4.5 ปีขึ้นไป ดังรายละเอียดในตารางที่ 3.15

**ตารางที่ 3.15** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ CreatedDate

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| CreatedDate | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | น้อยกว่า 0.5 ปี | 18.68 | 9.03 |
| | ระหว่าง 0.5 ปี ถึง 1 ปี | 12.62 | 5.16 |
| | ระหว่าง 1 ปี ถึง 1.5 ปี | 9.78 | 2.58 |
| | ระหว่าง 1.5 ปี ถึง 2 ปี | 10.24 | 3.87 |
| | ระหว่าง 2 ปี ถึง 2.5 ปี | 9.31 | 5.81 |
| | ระหว่าง 2.5 ปี ถึง 3 ปี | 8.94 | 12.26 |
| | ระหว่าง 3 ปี ถึง 3.5 ปี | 11.19 | 14.84 |
| | ระหว่าง 3.5 ปี ถึง 4 ปี | 10.06 | 16.77 |
| | ระหว่าง 4 ปี ถึง 4.5 ปี | 8.71 | 29.68 |
| | ระหว่าง 4.5 ปี ถึง 5 ปี | 0.35 | 0.00 |
| | ระหว่าง 5 ปี ถึง 5.5 ปี | 0.13 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Status ซึ่งเป็นรายละเอียดของสถานะของเจ้าของบัญชี ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงและบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าไม่มีบัญชีใดเลยที่ไม่มีมีสถานะ ทุกบัญชีมีสถานะทั้งหมด ดังรายละเอียดในตารางที่ 3.16

**ตารางที่ 3.16** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Status

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Status | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | มีค่าสถานะ | 100.00 | 100.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Url ซึ่งเป็นรายละเอียดของลิงค์ (Link) หรือจุดเชื่อมโยงข้อมูลของข้อความที่ทวีต โดยในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริง ร้อยละ 50.41 มีจำนวน 3 ลิงค์ที่ปรากฏในข้อความ ร้อยละ 49.46 ไม่ปรากฏจำนวนลิงค์ในข้อความ ร้อยละ 0.12 มีจำนวน 1 ลิงค์ที่ปรากฏในข้อความ และร้อยละ 0.01 มีจำนวนมากกว่า 3 ลิงค์ที่ปรากฏในข้อความ ส่วนของบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่า ร้อยละ 89.94 มีจำนวน 3 ลิงค์ที่ปรากฏในข้อความ ร้อยละ 10.06 ไม่ปรากฏจำนวนลิงค์ในข้อความ และไม่มีบัญชีใดเลยที่มีจำนวน 1 ลิงค์หรือ 2 ลิงค์ หรือมากกว่า 3 ลิงค์ที่ปรากฏในข้อความ ดังรายละเอียดในตารางที่ 3.17

**ตารางที่ 3.17** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Url

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Url | ไม่ปรากฏข้อมูล | 49.46 | 10.06 |
| | 1 ลิงค์ | 0.12 | 0.00 |
| | 2 ลิงค์ | 0.00 | 0.00 |
| | 3 ลิงค์ | 50.41 | 89.94 |
| | มากกว่า 3 ลิงค์ | 0.01 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ Mentions ที่เป็นการระบุถึงบัญชีผู้ใช้อื่นที่พบในข้อความที่ทวีต จะมีค่าเท่ากับ Number of Mentions ที่เป็นค่าจาก Twitter API โดยแสดงจำนวนการระบุถึงบัญชีผู้ใช้อื่นที่พบในข้อความที่ทวีต ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริง ร้อยละ 89.68 มีจำนวน 1 @ ปรากฏในข้อความ ร้อยละ 6.72 ไม่ปรากฏจำนวน @ ใน

ข้อความ ร้อยละ 3.11 มีจำนวน 2 @ ปรากฏในข้อความ ร้อยละ 0.40 มีจำนวน 3 @ ปรากฏในข้อความ และร้อยละ 0.09 มีจำนวนมากกว่า 3 @ ปรากฏในข้อความ ส่วนของบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่า ร้อยละ 68.93 มีจำนวน 1 @ ปรากฏในข้อความ ร้อยละ 28.99 ไม่ปรากฏจำนวน @ ในข้อความ ร้อยละ 1.78 มีจำนวน 2 @ ปรากฏในข้อความ ร้อยละ 0.30 มีจำนวน 3 @ ปรากฏในข้อความ และไม่มีบัญชีใดเลยที่มีจำนวนมากกว่า 3 @ ปรากฏในข้อความ ดังรายละเอียดในตารางที่ 3.18



**ตารางที่ 3.18** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ Mentions และ Number of Mentions

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| Mentions และ Number of Mentions | ไม่ปรากฏข้อมูล | 6.72 | 28.99 |
| | 1 @ | 89.68 | 68.93 |
| | 2 @ | 3.11 | 1.78 |
| | 3 @ | 0.40 | 0.30 |
| | มากกว่า 3 @ | 0.09 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ HashTags ที่เป็นการระบุถึงข้อความเฉพาะที่ต้องการเน้นโดยใช้สัญลักษณ์ # ประกอบในข้อความ โดยมีค่าเท่ากับ Number of HashTags ที่เป็นค่าจาก Twitter API จากข้อมูลพบว่าส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริง ร้อยละ 49.635 ไม่ปรากฏจำนวน # ในข้อความ ร้อยละ 26.71 มีจำนวน 1 # ปรากฏในข้อความ ร้อยละ 13.70 มีจำนวน 2 # ปรากฏในข้อความ ร้อยละ 5.50 มีจำนวน 3 # ปรากฏในข้อความ และร้อยละ 4.44 มีจำนวนมากกว่า 3 # ปรากฏในข้อความ ส่วนของบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่า ร้อยละ 53.25 ไม่ปรากฏข้อมูลจำนวน # ในข้อความ ร้อยละ 22.78 มีจำนวน 1 # ปรากฏในข้อความ ร้อยละ 19.23 มีจำนวน 3 # ปรากฏในข้อความ ร้อยละ 1.18 มีจำนวน 2 # ปรากฏในข้อความ และร้อยละ 3.55 มีจำนวนมากกว่า 3 # ปรากฏในข้อความ ดังรายละเอียดในตารางที่ 3.19

**ตารางที่ 3.19** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ HashTags และ Number of HashTags

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| HashTags และ Number of HashTags | ไม่ปรากฏข้อมูล | 49.65 | 53.25 |
| | 1 # | 26.71 | 22.78 |
| | 2 # | 13.70 | 1.18 |
| | 3 # | 5.50 | 19.23 |
| | มากกว่า 3 # | 4.44 | 3.55 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ RetweetCount เป็นจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำ ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริง พบว่าร้อยละ 28.91 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำ 1,000-9,999 ครั้ง ร้อยละ 23.02 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำระหว่าง 100-999 ครั้ง ร้อยละ 20.87 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำระหว่าง 10,000-99,999 ครั้ง ร้อยละ 14.60 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำระหว่าง 10-99 ครั้ง ร้อยละ 6.84 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำระหว่าง 1-9 ครั้ง ร้อยละ 5.77 ไม่มีการส่งต่อข้อความทวีตซ้ำ และไม่มีบัญชีใดที่มีจำนวนครั้ง ที่มีการส่งต่อข้อความทวีตซ้ำมากกว่า 100,000 ครั้งขึ้นไป ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าร้อยละ 52.66 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำระหว่าง 10-99 ครั้ง ร้อยละ 24.85 มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำระหว่าง 1-9 ครั้ง ร้อยละ 21.89 ไม่มีการส่งต่อข้อความทวีตซ้ำ และไม่มีบัญชีใดที่มีจำนวนครั้งที่มีการส่งต่อข้อความทวีตซ้ำมากกว่า 1,000 ครั้งขึ้นไป ดังรายละเอียดในตารางที่ 3.20

**ตารางที่ 3.20** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ RetweetCount

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| RetweetCount | ไม่ปรากฏข้อมูล | 5.77 | 21.89 |
| | 1-9 | 6.84 | 24.85 |
| | 10-99 | 14.60 | 52.66 |
| | 100-999 | 23.02 | 0.59 |
| | 1,000-9,999 | 28.91 | 0.00 |
| | 10,000-99,999 | 20.87 | 0.00 |
| | 100,000-999,999 | 0.00 | 0.00 |
| | 1,000,000-9,999,999 | 0.00 | 0.00 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ TweetCreatedDate เป็นวันเวลาที่ข้อความทวีตถูกสร้างขึ้น ในงานวิจัยนี้สนใจช่วงเวลาที่แต่ละข้อความถูกสร้างขึ้น จึงแบ่งช่วงเวลาในวันเป็น 4 ช่วงเวลาได้แก่ 06.01-12.00 น. 12.01-18.00 น. 18.01-24.00 น. และ 00.01-06.00 น. ซึ่งพบว่าข้อมูลส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงร้อยละ 31.59 ข้อความถูกสร้างขึ้นในช่วงเวลา 00.01-06.00 น. ร้อยละ 30.29 ข้อความถูกสร้างขึ้นในช่วงเวลา 12.01-18.00น. ร้อยละ 28.18 ข้อความถูกสร้างขึ้นในช่วงเวลา 06.01-12.00 น. และร้อยละ 9.94 ข้อความถูกสร้างขึ้นในช่วงเวลา 18.01-24.00 น. ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่าร้อยละ 48.22 ข้อความถูกสร้างขึ้นในช่วงเวลา 06.01-12.00 น. ร้อยละ 29.88 ข้อความถูกสร้างขึ้นในช่วงเวลา 12.01-18.00 น. ร้อยละ 20.12 ข้อความถูกสร้างขึ้นในช่วงเวลา 00.01-06.00 น. และร้อยละ 1.78 ข้อความถูกสร้างขึ้นในช่วงเวลา 18.01-24.00 น. ดังรายละเอียดในตารางที่ 3.21

**ตารางที่ 3.21** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ TweetCreatedDate

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| TweetCreatedDate | ไม่ปรากฏข้อมูล | 0.00 | 0.00 |
| | 06.01-12.00 น. | 28.18 | 48.22 |
| | 12.01-18.00 น. | 30.29 | 29.88 |
| | 18.01-24.00 น. | 9.94 | 1.78 |
| | 00.01-06.00 น. | 31.59 | 20.12 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ MessageText พิจารณาข้อความที่ทวีตเป็นข้อความที่สร้างจากบัญชีผู้ใช้เองหรือส่งต่อข้อความทวีตของบัญชีอื่นซ้ำ ในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริงพบว่าร้อยละ 5.77 เป็นข้อความที่เจ้าของบัญชีผู้ใช้เขียนขึ้นมาเอง และร้อยละ 94.23 เป็นข้อความที่ส่งต่อข้อความทวีตของบัญชีอื่น ส่วนในบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมร้อยละ 21.89 เป็นข้อความที่เจ้าของบัญชีผู้ใช้เขียนขึ้นมาเอง และร้อยละ 78.11 เป็นข้อความที่ส่งต่อข้อความของบัญชีอื่น ดังรายละเอียดในตารางที่ 3.22

**ตารางที่ 3.22** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ MessageText

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| MessageText | ข้อความที่เขียนเอง | 5.77 | 21.89 |
| | ข้อความที่มีการส่งต่อซ้ำอีกครั้ง (Retweet) | 94.23 | 78.11 |

พิจารณาชุดข้อมูลที่นำมาใช้ในงานวิจัย เมื่อวิเคราะห์คุณลักษณะ MessageImage ซึ่งเป็นรายละเอียดของลิงค์ของตำแหน่งที่จัดเก็บรูปภาพประกอบข้อความข่าว พบว่าในส่วนของบัญชีผู้ใช้ที่มีการแสดงข่าวจริง ร้อยละ 72.33 ไม่ปรากฏจำนวนลิงค์ของรูปภาพในข้อความ ร้อยละ 27.03 มีจำนวน 1 ลิงค์ของรูปภาพที่ปรากฏในข้อความ ร้อยละ 0.62 มีจำนวน 3 ลิงค์ของรูปภาพที่ปรากฏในข้อความ ร้อยละ 0.03 มีจำนวนมากกว่า 3 ลิงค์ของรูปภาพที่ปรากฏในข้อความ และไม่พบบัญชีที่มี

จำนวน 2 ลิงค์ของรูปภาพที่ปรากฏในข้อความ ส่วนของบัญชีของผู้ใช้ที่มีการแสดงข่าวปลอมพบว่า ร้อยละ 73.96 ไม่ปรากฏจำนวนลิงค์ของรูปภาพในข้อความ ร้อยละ 25.44 มีจำนวน 1 ลิงค์ของรูปภาพที่ปรากฏในข้อความ ร้อยละ 0.59 มีจำนวน 3 ลิงค์ของรูปภาพที่ปรากฏในข้อความ มีจำนวนมากกว่า 3 ลิงค์ของรูปภาพที่ปรากฏในข้อความ และไม่พบบัญชีที่มีจำนวน 2 ลิงค์ และมากกว่า 3 ลิงค์ของรูปภาพที่ปรากฏในข้อความ ดังรายละเอียดในตารางที่ 3.23

**ตารางที่ 3.23** ผลการวิเคราะห์ข้อมูลข่าวตามคุณลักษณะ MessageImage

| คุณลักษณะ | รายละเอียด | ร้อยละของข้อมูล | |
|---|---|---|---|
| | | ข่าวจริง | ข่าวปลอม |
| MessageImage | ไม่ปรากฏข้อมูล | 72.33 | 73.96 |
| | 1 ลิงค์ | 27.03 | 25.44 |
| | 2 ลิงค์ | 0.00 | 0.00 |
| | 3 ลิงค์ | 0.62 | 0.59 |
| | มากกว่า 3 ลิงค์ | 0.03 | 0.00 |

บทนี้ได้กล่าวถึงรายละเอียดของชุดข้อมูลที่จัดเก็บจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ ส่วนในเอกสารบทต่อไปเป็นการนำเสนอรายละเอียดของการทดลอง และการพิจารณาคุณลักษณะสำคัญของชุดข้อมูลที่เหมาะสมในการใช้จำแนกข่าวปลอม

# บทที่ 4
# ผลงานวิจัย

บทนี้นำเสนอรายละเอียดของผลที่ได้จากการทดลองได้จากการประมวลผลด้วยวิธีการเรียนรู้ด้วยเครื่องทั้ง 3 วิธีการ ได้แก่ Naïve Bayes, Neural Network และ Support Vector Machine และการทดลองเพื่อปรับปรุงความถูกต้องในการจำแนกข่าวปลอม โดยการเลือกคุณลักษณะที่เหมาะสมสำหรับการทดลองในชุดข้อมูลที่จัดเก็บ โดยพิจารณาตัดบางคุณลักษณะออกก่อนการประมวลผลด้วยวิธีการเรียนรู้ด้วยเครื่องอีกครั้งเพื่อตรวจสอบค่าความถูกต้องในการจำแนกข่าวปลอม โดยมีรายละเอียดดังต่อไปนี้

## 4.1 การเรียนรู้ด้วยเครื่อง

เมื่อข้อมูลได้ผ่านกระบวนการปรับค่าข้อมูลให้เป็นตัวเลขทั้งหมด และนำไปขจัดความซ้ำซ้อนของข้อมูลออกไปเหลือข้อมูลที่มีค่าไม่ซ้ำกันเลย จำนวน 327,784 ข้อความ จากนั้นนำข้อมูลนี้ไปผ่านกระบวนการเรียนรู้ด้วยเครื่อง เพื่อจำแนกข้อมูลที่ข่าวปลอม โดยเลือกใช้วิธีการเรียนรู้ของเครื่อง 3 วิธีการ ได้แก่ Naïve Bayes, Neural Network และ Support Vector Machine ด้วยการใช้โปรแกรม Weka ได้ผลการจำแนกข้อมูลในการทดลอง กำหนดให้ใช้ 10-fold cross validation สำหรับการทดสอบประสิทธิภาพของโมเดล โดยการแบ่งข้อมูลเป็น 10 ส่วน เพื่อนำไปเทรน 9 ส่วน และใช้ทดสอบ 1 ส่วน โดยผลลัพธ์ที่ได้จากการทดลองแสดงดังตารางที่ 4.1

ส่วนของโปรแกรม Weka ที่เลือกใช้ มีการกำหนดค่าตัวแปรต่าง ๆ ของแต่ละวิธีการเรียนรู้ของเครื่อง ดังรายละเอียดต่อไปนี้

Naïve Bayes:

```
function NaiveBayes,
batchSize = 100,
numDecimalPlaces = 2,
useKernelEstimator = False,
useSupervisedDiscretization = False
```

Neural Network:

```
function MultilayerPerceptron,
batchSize = 100,
hiddenLayers = a, 1
earningRate = 0.3,
momentum = 0.2,
nominalToBinaryFilter = True,
nomalizeAttribute = True,
nomalizeNumericalClass = True,
numDecimalPlaces = 2,
seed = 0,
trainingTime = 500,
validationSetSize = 0,
validationThreshold = 20
```

Support Vector Machine:

```
function SGD,
batchSize = 100,
epochs = 500,
epsilon = 0.001,
lambda 1.0E-4,
learningRate = 0.01,
lossFunction = Hinge loss,
numDecimalPlaces = 2,
seed = 1
```

จากผลการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องทั้ง 3 วิธี ได้ตารางประเมินการทำงานของแบบจำลอง (Confusion matrix) ของแต่ละวิธีการ โดยค่าตัวเลขเป็นจำนวนข้อความข่าว โดยค่าที่อยู่ในแนวตั้งเป็นคำตอบจริงของข้อมูลข่าวจริงและข่าวปลอมตามลำดับ ส่วนค่าที่อยู่ในแนวนอนเป็นค่าที่แบบจำลองทำนายผลเป็นข่าวจริงหรือข่าวปลอมตามลำดับ ดังรายละเอียดในตารางที่ 4.1

**ตารางที่ 4.1** ผลการทำนายโดยแบบจำลองที่ได้จากแต่ละวิธีการเรียนรู้ด้วยเครื่อง

| **ค่าที่ทำนาย/ค่าคำตอบ** | ข่าวจริง | ข่าวปลอม |
|---|---|---|
| ข่าวจริง | True Positive (TP) | False Positive (FP) |
| ข่าวปลอม | False Negative (FN) | True Negative (TN) |
| **วิธีการเรียนรู้ด้วยเครื่อง** | **จำนวนข่าว** | |
| Naïve Bayes | 283,745 | 2,595 |
| | 10,239 | 31,205 |
| Neural Network | 293,728 | 115 |
| | 256 | 33,685 |
| Support Vector Machine | 293,738 | 99 |
| | 246 | 33,701 |

จากข้อมูลในตารางที่ 4.1 ในส่วนแรกเป็นการนิยามโครงสร้างของตารางประเมินการทำงานของแบบจำลอง ที่ใช้แสดงผลการทำนายจากแบบจำลองของแต่ละวิธีการ ซึ่งมีความหมายดังต่อไปนี้

– True Positive (TP) หมายถึง จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง

– False Positive (FP) หมายถึง จำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง

– False Negative (FN) หมายถึง จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม

– True Negative (TN) หมายถึง จำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม

ความสัมพันธ์ของจำนวนข่าวที่ปรากฏในตารางที่ 4.1 มีดังนี้

– จำนวนข่าวจริงทั้งหมดที่ใช้ในชุดข้อมูล มีค่าเท่ากับ ผลรวมของจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง กับ จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม

– จำนวนข่าวปลอมทั้งหมดที่ใช้ในชุดข้อมูล มีค่าเท่ากับ ผลรวมของจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง กับ จำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม

– จำนวนข่าวที่แบบจำลองทายว่าเป็นข่าวจริง มีค่าเท่ากับ จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง กับ จำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง

– จำนวนข่าวที่แบบจำลองทายว่าเป็นข่าวปลอม มีค่าเท่ากับ จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม กับ จำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม

จากข้อมูลในตารางที่ 4.1 เมื่อพิจารณาในส่วนของวิธีการเรียนรู้ด้วยเครื่องแบบ Naïve Bayes จะได้

True Positive (TP) = 283,745 False Positive (FP) = 2,595

False Negative (FN) = 10,239 True Negative (TN) = 31,205

เมื่อนำไปคำนวณค่า Precision ซึ่งในที่นี้ใช้เพื่อวัดความแม่นยำในการทำนายของแบบจำลอง โดยแสดงสัดส่วนระหว่าง จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง กับ ผลรวมระหว่างจำนวนข่าวจริงที่แบบจำลองทำนายว่าจริงและจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง ตามการคำนวณในสมการที่ (4) จะได้

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} = \frac{283{,}745}{283{,}745+2{,}595} = 0.9909$$

คำนวณค่า Recall หรือ True Positive Rate (TPR) เป็นการวัดค่าความถูกต้องที่แบบจำลองทำนายข้อมูลข่าวจริงได้ถูกต้องเมื่อเทียบกับ ผลรวมระหว่างจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริงและจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม ตามการคำนวณในสมการที่ (5) จะได้

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} = \frac{283{,}745}{283{,}745+10{,}239} = 0.9652$$

คำนวณค่า F-measure เพื่อการวัดค่าประสิทธิภาพโดยรวมของแบบจำลองโดยคำนวณจากเฉลี่ยระหว่าง Precision และ Recall สามารถคำนวณได้จากสมการที่ (6) จะได้

$$F\text{-}measure=2* \frac{(Precision*Recall)}{(Precision+Recall)} = 2* \frac{(0.9909*0.9652)}{(0.9909+0.9652)} = 0.9779$$

คำนวณค่า True Negative Rate (TNR) คืออัตราส่วนจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม กับ ผลรวมระหว่างจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอมและจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง โดยสามารถคำนวณได้จากสมการที่ (7)

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Negative}{(True\ Negative + False\ Positive)} = \frac{31{,}205}{31{,}205+2{,}595} = 0.9232$$

คำนวณค่า False Positive Rate (FPR) คืออัตราส่วนจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง กับ ผลรวมระหว่างจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอมและจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริง โดยสามารถคำนวณได้จากสมการที่ (8)

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{(True\ Negative + False\ Positive)} = \frac{2{,}595}{31{,}205+2{,}595} = 0.0768$$

คำนวณค่า False Negative Rate (FNR) คืออัตราส่วนจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม กับ ผลรวมระหว่าง จำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริงและจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม โดยสามารถคำนวณได้จากสมการที่ (9)

$$\textit{False Negative Rate (FNR)} = \frac{\textit{False Negative}}{(\textit{True Positive} + \textit{False Negative})} = \frac{10{,}239}{283{,}745+10{,}239} = 0.0348$$

คำนวณค่า Accuracy ที่ใช้เพื่อการวัดค่าความถูกต้องในการทำนายของแบบจำลองที่สามารถทำนายได้ผลถูกต้องคืออัตราส่วนของผลรวมระหว่างจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวจริงและจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอม กับ จำนวนข้อมูลทั้งหมดในชุดข้อมูล โดยสามารถคำนวณได้จากสมการที่ (3) และคิดเป็นค่าร้อยละได้โดยการคูณด้วย 100

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

$$= \frac{283{,}745+31{,}205}{283{,}745+31{,}205+2{,}595+10{,}239} = 0.9608$$

จากข้อมูลในตารางที่ 4.1 เมื่อพิจารณาในส่วนของวิธีการเรียนรู้ด้วยเครื่องแบบ Neural Network จะได้

True Positive (TP) = 293,728 False Positive (FP) = 115

False Negative (FN) = 256 True Negative (TN) = 33,685

เมื่อนำไปคำนวณค่า Precision ตามการคำนวณในสมการที่ (4) จะได้

$$\textit{Precision} = \frac{\textit{True Positive}}{(\textit{True Positive} + \textit{False Positive})} = \frac{293{,}728}{293{,}728+115} = 0.9996$$

คำนวณค่า Recall หรือ True Positive Rate (TPR) ตามการคำนวณในสมการที่ (5) จะได้

$$\textit{Recall} = \frac{\textit{True Positive}}{(\textit{True Positive} + \textit{False Negative})} = \frac{293{,}728}{293{,}728+256} = 0.9991$$

คำนวณค่า F-measure ได้จากสมการที่ (6) จะได้

$$\textit{F-measure} = 2 * \frac{(\textit{Precision} * \textit{Recall})}{(\textit{Precision}+\textit{Recall})} = 2 * \frac{(0.9996*0.9991)}{(0.9996+0.9991)} = 0.9994$$

คำนวณค่า True Negative Rate (TNR) ได้จากสมการที่ (7) จะได้

$$\textit{True Negative Rate (TNR)} = \frac{\textit{True Negative}}{(\textit{True Negative} + \textit{False Positive})} = \frac{33{,}685}{33{,}685+115} = 0.9966$$

คำนวณค่า False Positive Rate (FPR) ได้จากสมการที่ (8) จะได้

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{(True\ Negative + False\ Positive)} = \frac{115}{33{,}685+115} = 0.0034$$

คำนวณค่า False Negative Rate (FNR) คำนวณได้จากสมการที่ (9) จะได้

$$False\ Negative\ Rate\ (FNR) = \frac{False\ Negative}{(True\ Positive + False\ Negative)} = \frac{256}{293{,}728+256} = 0.0009$$



คำนวณค่า Accuracy ได้จากสมการที่ (3) จะได้

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

$$= \frac{293{,}728+33{,}685}{293{,}728+33{,}685+115+256} = 0.9989$$

จากข้อมูลในตารางที่ 4.1 เมื่อพิจารณาในส่วนของวิธีการเรียนรู้ด้วยเครื่องแบบ Support Vector Machine จะได้

True Positive (TP) = 293,738 False Positive (FP) = 99

False Negative (FN) = 246 True Negative (TN) = 33,701

เมื่อนำไปคำนวณค่า Precision ตามการคำนวณในสมการที่ (4) จะได้

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)} = \frac{293{,}738}{293{,}738+99} = 0.9997$$

คำนวณค่า Recall หรือ True Positive Rate (TPR) ตามการคำนวณในสมการที่ (5) จะได้

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)} = \frac{293{,}738}{293{,}738+246} = 0.9992$$

คำนวณค่า F-measure ได้จากสมการที่ (6) จะได้

$$F\text{-}measure = 2* \frac{(Precision*Recall)}{(Precision+Recall)} = 2* \frac{(0.9997*0.9992)}{(0.9997+0.9992)} = 0.9994$$

คำนวณค่า True Negative Rate (TNR) ได้จากสมการที่ (7) จะได้

$$True\ Negative\ Rate\ (TNR) = \frac{True\ Negative}{(True\ Negative + False\ Positive)} = \frac{33{,}701}{33{,}701+99} = 0.9971$$

คำนวณค่า False Positive Rate (FPR) ได้จากสมการที่ (8) จะได้

$$False\ Positive\ Rate\ (FPR) = \frac{False\ Positive}{(True\ Negative + False\ Positive)} = \frac{99}{33{,}701+99} = 0.0029$$

คำนวณค่า False Negative Rate (FNR) คำนวณได้จากสมการที่ (9) จะได้

$$False\ Negative\ Rate\ (FNR) = \frac{False\ Negative}{(True\ Positive + False\ Negative)} = \frac{246}{293{,}738+246} = 0.0008$$

คำนวณค่า Accuracy ได้จากสมการที่ (3) จะได้

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

$$= \frac{293{,}738+33{,}701}{293{,}738+33{,}701+99+246} = 0.9989$$

จากการคำนวณข้างต้นของแต่ละแบบจำลองที่สร้างจากวิธีการเรียนรู้ด้วยเครื่องทั้ง 3 ได้แก่ Naïve Bayes, Neural Network และ Support Vector Machine แสดงรายละเอียดดังตารางที่ 4.2

**ตารางที่ 4.2** ผลลัพธ์จากการทดลองการจำแนกข้อมูลด้วยวิธีการเรียนรู้ด้วยเครื่อง

| | Precision | Recall (True Positive Rate) | F-Measure | True Negative Rate | False Positive Rate | False Negative Rate | Accuracy |
|---|---|---|---|---|---|---|---|
| **Naïve Bayes** | 0.9909 | 0.9652 | 0.9779 | 0.9232 | 0.0768 | 0.0348 | 96.08% |
| **Neural Network** | 0.9996 | 0.9991 | 0.9994 | 0.9966 | 0.0034 | 0.0009 | 99.89% |
| **Support Vector Machine** | 0.9997 | 0.9992 | 0.9994 | 0.9971 | 0.0029 | 0.0008 | 99.89% |

จากผลที่ได้ในตารางที่ 4.2 พบว่าค่า Precision ที่ใช้เพื่อวัดความแม่นยำในการทำนายของแบบจำลองที่สร้างด้วย Support Vector Machine ให้ค่าดีที่สุดมากถึง 0.9997 ตามมาด้วย Neural Network ที่ให้ค่า 0.9996 ซึ่งมีความแตกต่างกันน้อยมาก ส่วน Naïve Bayes ให้ค่าความแม่นยำมากถึง 0.9909 แม้ว่าจะมีค่าน้อยกว่าอีกสองวิธี แต่ก็เป็นตัวเลขที่มีค่าค่อนข้างสูง ส่วนค่า Recall หรือ True Positive Rate (TPR) เป็นการวัดค่าความถูกต้องที่แบบจำลองทำนายข้อมูลข่าวจริงได้ถูกต้องเมื่อเปรียบเทียบกับจำนวนข่าวจริงทั้งหมดที่ปรากฏอยู่ในชุดข้อมูล ซึ่งผลการคำนวณที่ได้เป็นไปในแนวทางเดียวกับค่า Precision กล่าวคือ Support Vector Machine ให้ค่าดีที่สุดมากถึง 0.9992 ตามมาด้วย Neural Network ที่ให้ค่า 0.9991 และ Naïve Bayes ให้ค่า 0.9652 ดังนั้นการคำนวณค่า F-measure ที่ใช้วัดค่าประสิทธิภาพโดยรวมของแบบจำลองที่คำนวณจาก ค่าเฉลี่ยระหว่าง

Precision และ Recall ของแบบจำลองที่ได้จาก Support Vector Machine และ Neural Network จึงมีค่าเท่ากันคือ 0.9994 ส่วน Naïve Bayes ให้ค่าต่ำกว่าเพียงเล็กน้อยคือ 0.9779

ค่า True Negative Rate (TNR) เป็นอัตราส่วนของจำนวนข่าวปลอมที่แบบจำลองสามารถทำนายได้ถูกต้องว่าเป็นข่าวปลอมกับจำนวนข่าวปลอมทั้งหมดที่ปรากฏในชุดข้อมูล พบว่าแบบจำลองที่สร้างจากวิธี Support Vector Machine ให้ค่า 0.9971 ส่วน Neural Network ให้ค่า 0.9966 และ Naïve Bayes ให้ค่า 0.9232 แสดงว่าทุกแบบจำลองที่สร้างขึ้นจากสามวิธีการเรียนรู้ด้วยเครื่องสามารถใช้จำแนกข่าวปลอมได้อย่างถูกต้องมากกว่าร้อยละ 90 ขึ้นไป

ค่า False Positive Rate (FPR) เป็นอัตราส่วนของจำนวนข่าวปลอมที่แบบจำลองทำนายได้ว่าเป็นข่าวจริงซึ่งเป็นผลทำนายที่ผิดพลาดกับจำนวนข่าวปลอมทั้งหมดที่ปรากฏในชุดข้อมูล ดังนั้นค่าที่ได้นี้ควรจะเป็นค่าตัวเลขน้อย ๆ ยิ่งวิธีการใดให้ค่ายิ่งน้อยแสดงว่าแบบจำลองนั้นทำนายข้อมูลข่าวปลอมผิดน้อยมากเท่านั้น จากผลการวิจัยพบว่าแบบจำลองที่สร้างจากวิธี Support Vector Machine ให้ค่าน้อยที่สุดคือ 0.0029 ตามด้วยค่าที่ได้จาก Neural Network คือ 0.0034 และค่าที่ได้จาก Naïve Bayes มีค่าเป็น 0.0768

ค่า False Negative Rate (FNR) เป็นอัตราส่วนของจำนวนข่าวจริงที่แบบจำลองทำนายได้ว่าเป็นข่าวปลอมซึ่งเป็นผลทำนายที่ผิดพลาดกับจำนวนข่าวจริงทั้งหมดที่ปรากฏในชุดข้อมูล ซึ่งค่าที่ได้นี้ควรจะเป็นค่าตัวเลขน้อย ๆ ยิ่งวิธีการใดให้ค่ายิ่งน้อยแสดงว่าแบบจำลองนั้นทำนายข้อมูลข่าวจริงผิดไปน้อยมากเท่านั้น จากผลการวิจัยพบว่าแบบจำลองที่สร้างจากวิธี Support Vector Machine ให้ค่าน้อยที่สุดคือ 0.0008 ตามด้วยค่าที่ได้จาก Neural Network คือ 0.0009 และค่าที่ได้จาก Naïve Bayes มีค่าเป็น 0.0348

คำนวณค่า Accuracy เป็นวัดค่าความถูกต้องในการทำนายของแบบจำลองที่สามารถทำนายได้ผลถูกต้องทั้งในส่วนของการทำนายข่าวจริงได้ถูกต้องและทำนายข่าวปลอมได้อย่างถูกต้องจากจำนวนข้อมูลทั้งหมดในชุดข้อมูล ดังนั้นหากค่าที่ได้มีค่ามาก ๆ แสดงว่าแบบจำลองที่ได้จากวิธีการนั้นสามารถจำแนกข่าวได้อย่างถูกต้องมาก เพื่อความสะดวกในการเปรียบเทียบค่าจึงได้คิดคำนวณเป็นค่าร้อยละด้วยการคูณด้วย 100 จะได้ผลเป็น

ค่าร้อยละของความถูกต้องที่ได้จาก Naïve Bayes มีค่าเท่ากับ 96.08% แต่วิธีการ Neural Network และ Support Vector Machine ให้ผลลัพธ์ค่าร้อยละความถูกต้อง (Accuracy) สูงกว่าวิธีการของ Naïve Bayes ซึ่งค่าร้อยละความถูกต้องที่ได้จาก Neural Network มีค่าเท่ากันกับค่าที่ได้จาก Support Vector Machine นั่นคือมีค่าเท่ากับ 99.89% ซึ่งเป็นข้อสังเกตได้ว่าแบบจำลองที่สร้างจากวิธีการเรียนรู้ด้วยเครื่องทั้ง 3 วิธี ให้ผลการจำแนกข่าวปลอมได้อย่างถูกต้อง

**4.2 การเลือกคุณลักษณะที่เหมาะสม**

การเลือกพิจารณาตัดคุณลักษณะที่มีความสำคัญน้อยออกเพื่อให้ประสิทธิภาพการทำนายได้ค่าความถูกต้องสูงมากขึ้น โดยการคัดสรรคุณลักษณะ Attribute Selection โดยการเลือกลดจำนวนคุณลักษณะที่ไม่เกี่ยวข้องออก ให้เหลือเฉพาะคุณลักษณะที่มีความสัมพันธ์กันเท่านั้น ซึ่งทำให้ผลการจำแนกได้ค่าความถูกต้องมากขึ้นและลดเวลาที่ใช้ในการประมวลผลลง ยกตัวอย่างเช่น

– การลดคุณลักษณะด้วย InfoGain Attribute Evaluation เป็นการลดคุณลักษณะโดยการวัดค่า Information gain ที่คุณลักษณะมีความสัมพันธ์กับคลาส

– GainRatio Attribute Evaluation เป็นการประเมินค่าคุณลักษณะด้วยการวัดค่า Gain Ratio ที่วัดความสัมพันธ์ของคุณลักษณะโดยมีการปรับตามค่าขอบเขตของข้อมูลในคุณลักษณะที่สนใจกับคลาส

– OneR Attribute Evaluation เป็นการลดจำนวนคุณลักษณะโดยใช้กฎ (Rule) ด้วยการสร้างกฎจากต้นไม้ตัดสินใจหนึ่งระดับ โดยกฎที่สร้างได้จากแต่ละคุณลักษณะจะมีความแตกต่างกัน โดยจะเลือกใช้กฎที่มีค่าความผิดพลาดน้อยสุดเพียงกฎเดียว โดยค่าของคุณลักษณะที่ให้ค่าความผิดพลาดน้อยสุดจะเป็นค่าที่ดีที่สุด

– ChiSquare Attribute Evaluation เป็นการลดจำนวนคุณลักษณะที่ใช้ โดยการคำนวณค่า Chi-Square ทางสถิติ

การปรับปรุงความถูกต้องในการจำแนกข่าวปลอม เมื่อพิจารณาคุณลักษณะที่คาดว่าจะส่งผลต่อความแม่นยำในการจำแนกข่าวปลอม ในงานวิจัยนี้ได้เลือกวิธีการทดลองโดยการลดคุณลักษณะออกทีละหนึ่ง แล้วใช้วิธีการเรียนรู้ด้วยเครื่องแต่ละวิธีการจำแนกข่าวปลอม ผลของร้อยละความถูกต้องในการจำแนกข้อมูลเมื่อมีการพิจารณาลดแต่ละคุณลักษณะสำหรับการทดลองด้วยแต่ละวิธีการเรียนรู้ด้วยเครื่อง ได้แก่ Naïve Bayes, Neural Network และ Support Vector Machine แสดงค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมแต่ละวิธีการเรียนรู้ด้วยเครื่องโดยการลดทีละหนึ่งคุณลักษณะและค่าเปรียบเทียบที่เปลี่ยนแปลงไปจากเดิมที่จำแนกด้วยคุณลักษณะทั้ง 22 คุณลักษณะ

หากผลที่ได้ที่เป็นค่าตัวเลขบวกหมายถึงได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมเพิ่มมากขึ้น แสดงว่าการลดคุณลักษณะนั้น ทำให้ได้ความถูกต้องในการจำแนกมากยิ่งขึ้น จึงควรไปพิจารณาคุณลักษณะนั้น แต่หากผลที่ได้ที่เป็นค่าตัวเลขลบจะหมายความว่าได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมลดลงจากเดิม แสดงว่าการลดคุณลักษณะนั้น ทำให้ได้ความถูกต้องในการจำแนกลดลง จึงไม่ควรพิจารณาการลดคุณลักษณะนั้น ดังตารางที่ 4.3

อธิบายความหมายของตัวอย่างของผลการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องเมื่อมีการลดหนึ่งคุณลักษณะ (Id) ที่ได้แสดงในตารางที่ 4.3 ดังต่อไปนี้

เมื่อลดคุณลักษณะ Id และนำข้อมูล 21 คุณลักษณะที่เหลือไปจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Naïve Bayes ได้ค่าร้อยละความถูกต้องในการจำแนกเป็น 96.0828 เมื่อเปรียบเทียบกับการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Naïve Bayes โดยใช้ 22 คุณลักษณะที่ได้ค่าร้อยละความถูกต้องในการจำแนกเป็น 96.0825 จะได้ความแตกต่างเมื่อเปรียบเทียบกันเป็น 0.0003 แสดงว่าการลดคุณลักษณะ Id และนำข้อมูลที่เหลือไปจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Naïve Bayes แล้วจะได้ค่าร้อยละความถูกต้องในการจำแนกดีขึ้นเพียงร้อยละ 0.0003 ซึ่งเป็นค่าที่ได้เป็นค่าน้อยมาก มีความแตกต่างกันไม่มาก

เมื่อลดคุณลักษณะ Id ออกไป และนำข้อมูล 21 คุณลักษณะที่เหลือไปจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Neural Network ได้ค่าร้อยละความถูกต้องในการจำแนกเป็น 99.8957 เมื่อเปรียบเทียบกับการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Neural Network โดยใช้ 22 คุณลักษณะที่ได้ค่าร้อยละความถูกต้องในการจำแนกเป็น 99.8969 จะได้ความแตกต่างเมื่อนำค่ามาเปรียบเทียบกันเป็น -0.0012 แสดงว่าการลดคุณลักษณะ Id และนำข้อมูลที่เหลือไปจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Neural Network แล้วจะได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมลดลงร้อยละ 0.0012

เมื่อลดคุณลักษณะ Id ออกไป และนำข้อมูล 21 คุณลักษณะที่เหลือไปจำแนกข่าวปลอมปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Support Vector Machine ได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมเป็น 99.8969 ซึ่งมีค่าเท่ากับผลของร้อยละความถูกต้องในการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Support Vector Machine โดยใช้ 22 คุณลักษณะ ค่าเปรียบเทียบที่เปลี่ยนแปลงไปจึงมีค่าเท่ากับ 0 แสดงว่าการลดคุณลักษณะ Id เพียงค่าเดียวนี้ไม่ส่งผลต่อการเปลี่ยนแปลงของค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Support Vector Machine

**ตารางที่ 4.3** ผลการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องเมื่อมีการลดหนึ่งคุณลักษณะ

| คุณลักษณะข้อมูลที่ลดไป | ร้อยละความถูกต้องในการจำแนกข่าวปลอม | | | ค่าเปรียบเทียบที่เปลี่ยนแปลงไป | | |
|---|---|---|---|---|---|---|
| | Naïve Bayes | Neural Network | Support Vector Machine | Naïve Bayes | Neural Network | Support Vector Machine |
| *ไม่ลดคุณลักษณะใด ๆ* | *96.0825* | *99.8969* | *99.8969* | | | |
| Id | 96.0828 | 99.8957 | 99.8969 | 0.0003 | -0.0012 | 0 |
| Name | 96.2060 | 99.8963 | 99.8969 | 0.1235 | -0.0006 | 0 |
| IsVerified | 96.1935 | 99.8954 | 99.8969 | 0.1110 | -0.0015 | 0 |
| ProfileImageUrl | 96.2286 | 99.8969 | 99.8969 | 0.1461 | 0 | 0 |
| FollowersCount | 96.3693 | 99.8957 | 99.8969 | 0.2868 | -0.0012 | 0 |
| FriendsCount | 96.1835 | 99.8963 | 99.8969 | 0.1010 | -0.0006 | 0 |
| FavouritesCount | 96.2512 | 99.8969 | 99.8969 | 0.1687 | 0 | 0 |
| StatusesCount | 96.1584 | 99.8966 | 99.8969 | 0.0759 | -0.0003 | 0 |
| Description | 96.1935 | 99.8963 | 99.8969 | 0.1110 | -0.0006 | 0 |
| Location | 96.2045 | 99.896 | 99.8969 | 0.1220 | -0.0009 | 0 |
| TimeZone | 96.1935 | 99.8966 | 99.8969 | 0.1110 | -0.0003 | 0 |
| CreatedDate | 96.0489 | 99.896 | 99.8969 | -0.0336 | -0.0009 | 0 |
| Status | 96.1935 | 99.8963 | 99.8969 | 0.1110 | -0.0006 | 0 |
| Url | 94.9012 | 99.8969 | 99.8969 | -1.1813 | 0 | 0 |
| Mentions | 98.4169 | 99.8957 | 99.8969 | 2.3344 | -0.0012 | 0 |
| Number of Mentions | 98.4169 | 99.8957 | 99.8969 | 2.3344 | -0.0012 | 0 |
| HashTags | 96.1819 | 99.8963 | 99.8969 | 0.0994 | -0.0006 | 0 |
| Number of HashTags | 96.1819 | 99.8963 | 99.8969 | 0.0994 | -0.0006 | 0 |
| RetweetCount | 97.1326 | 99.8969 | 99.8969 | 1.0501 | 0 | 0 |
| TweetCreatedDate | 96.1871 | 99.8969 | 99.8969 | 0.1046 | 0 | 0 |
| MessageText | 97.096 | 99.8954 | 99.8969 | 1.0135 | -0.0015 | 0 |
| MessageImage | 96.2088 | 99.8969 | 99.8969 | 0.1263 | 0 | 0 |

จากข้อมูลทั้งหมดที่ปรากฏในตารางที่ 4.3 พบว่าแต่ละหนึ่งคุณลักษณะที่ลดหายไปไม่ส่งผลที่ดีขึ้นกับค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Neural Network เนื่องจากค่าผลต่างของการเปรียบเทียบที่ได้ทั้งหมดเป็นค่าตัวเลขติดลบหรือศูนย์ ส่วน

ผลต่างของค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Support Vector Machine เมื่อลดแต่ละหนึ่งคุณลักษณะแล้วมีค่าเป็นศูนย์ทั้งหมด จึงกล่าวได้ว่าการลดแต่ละหนึ่งคุณลักษณะของชุดข้อมูลแล้วนำไปจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Neural Network และ วิธี Support Vector Machine ไม่ส่งผลให้การจำแนกข่าวปลอมทำได้ถูกต้องมากขึ้น แต่มีคุณลักษณะที่ลดไปแล้วส่งผลที่ดีขึ้นสำหรับค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องวิธี Naïve Bayes จำนวน 4 คุณลักษณะ ได้แก่ Mentions, Number of Mentions, RetweetCount และ MessageText แต่เนื่องด้วยค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมเมื่อลดคุณลักษณะ Mentions มีค่าเท่ากับ Number of Mentions จึงเลือกพิจารณาเฉพาะ Mentions แทนคุณลักษณะ Number of Mentions ในการทดลองลดคุณลักษณะมากกว่าหนึ่งคุณลักษณะซึ่งให้ผลดังตารางที่ 4.4

**ตารางที่ 4.4** ผลการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่องเมื่อลดมากกว่าหนึ่งคุณลักษณะ

| คุณลักษณะข้อมูลที่ลดไป | ร้อยละความถูกต้องในการจำแนกข่าว | | | ค่าเปรียบเทียบที่เปลี่ยนแปลงไป | | |
|---|---|---|---|---|---|---|
| | Naïve Bayes | Neural Network | Support Vector Machine | Naïve Bayes | Neural Network | Support Vector Machine |
| *ไม่ลดคุณลักษณะใด ๆ* | *96.0825* | *99.8969* | *99.8969* | | | |
| Mentions, RetweetCount | 99.0390 | 99.8969 | 99.8969 | 2.9565 | 0.0000 | 0.0000 |
| Mentions, MessageText | 98.9896 | 99.8960 | 99.8969 | 2.9071 | -0.0009 | 0.0000 |
| RetweetCount, MessageText | 98.2998 | 99.8969 | 99.8969 | 2.2173 | 0.0000 | 0.0000 |
| Mentions, RetweetCount, MessageText | 99.3898 | 99.8969 | 99.8969 | 3.3073 | 0.0000 | 0.0000 |

จากตารางที่ 4.4 แสดงค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมโดยการลดมากกว่าหนึ่งคุณลักษณะแล้วนำไปผ่านกระบวนการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง Naïve Bayes, Neural Network และ Support Vector Machine โดยเลือกคุณลักษณะที่สนใจจากคุณลักษณะที่ส่งผลต่อค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมมีค่าเพิ่มมากขึ้นจากการลดทีละหนึ่งคุณลักษณะ ได้แก่ Mentions, RetweetCount และ MessageText ซึ่งได้ผลปรากฏว่า

การเปลี่ยนแปลงค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมที่เกิดจากการลดมากกว่าหนึ่งคุณลักษณะของชุดข้อมูล แล้วนำไปผ่านกระบวนการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง Naïve Bayes ได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมที่ถูกต้องมากยิ่งขึ้น โดยการลดสองคุณลักษณะของ Mentions กับ RetweetCount พร้อมกันทำให้ได้ค่าร้อยละความถูกต้องใน

การจำแนกข่าวปลอมเป็นร้อยละ 99.0390 โดยมีค่าเพิ่มมากขึ้นร้อยละ 2.9565 เมื่อเปรียบเทียบกับค่าเดิมของค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมที่ไม่มีการลดคุณลักษณะใด ๆ จาก 22 คุณลักษณะคือร้อยละ 96.0825 เมื่อพิจารณาการลดสองคุณลักษณะของ Mentions กับ MessageText พร้อมกัน ทำให้ได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมสูงขึ้นเป็นร้อยละ 98.9896 โดยมีค่าเพิ่มมากขึ้นร้อยละ 2.9071 และการลดสองคุณลักษณะของ RetweetCount กับ MessageText พร้อมกัน ทำให้ได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมสูงขึ้นเป็นร้อยละ 98.2998 โดยมีค่าเพิ่มมากขึ้นร้อยละ 2.2173 เมื่อเปรียบเทียบกับค่าเดิม และพิจารณาการลดคุณลักษณะพร้อมกันทั้งสามคุณลักษณะที่คาดว่าจะส่งผลต่อค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมคือ Mentions, RetweetCount และ MessageText ได้ค่าร้อยละความถูกต้องในการจำแนกข่าวปลอมสูงขึ้นเป็นร้อยละ 99.3898 โดยมีค่าเพิ่มมากขึ้นร้อยละ 3.3073 ซึ่งเป็นค่าร้อยละความถูกต้องที่มีค่ามากที่สุดในการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง Naïve Bayes สำหรับชุดข้อมูลนี้

เมื่อพิจารณาการลดมากกว่าหนึ่งคุณลักษณะของชุดข้อมูลแล้วนำไปผ่านกระบวนการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง Neural Network พบว่าค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมที่มีการลดคุณลักษณะพร้อมกันสองคุณลักษณะคือ Mentions กับ MessageText ส่งผลต่อค่าเปรียบเทียบที่เปลี่ยนแปลงไปโดยมีค่าแย่ลงร้อยละ -0.0009 นอกจากนั้นผลจากการลดมากกว่าหนึ่งคุณลักษณะแบบอื่น ได้แก่ Mentions พร้อมกับ RetweetCount หรือ RetweetCount พร้อมกับ MessageText หรือการลดสามคุณลักษณะในคราวเดียวโดยลดคุณลักษณะ Mentions, RetweetCount และ MessageText พร้อมกัน ผลปรากฏว่าไม่มีความแตกต่างกันในค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมได้ค่าเปรียบเทียบออกมาเป็นศูนย์ทั้งหมด

พบว่าค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง Support Vector Machine ไม่มีการเปลี่ยนแปลงใด ๆ ได้ค่าเปรียบเทียบที่เปลี่ยนแปลงไปมีความแตกต่างกันเป็นศูนย์ทั้งหมด

# บทที่ 5
# วิเคราะห์ผลการวิจัย

เอกสารในบทนี้เป็นบทสุดท้าย ประกอบด้วยสองส่วน ได้แก่ ส่วนแรกเป็นการสรุปผลการดำเนินงานของงานวิจัยนี้และอภิปรายผลการดำเนินงานที่ได้ ในส่วนสุดท้ายเป็นข้อจำกัดต่าง ๆ รวมถึงแนวทางวิจัยต่อในอนาคต ดังรายละเอียดต่อไปนี้



## 5.1 สรุปผล

ข้อมูลจากเครือข่ายสังคมออนไลน์ทวิตเตอร์ ในช่วงระหว่างเดือนตุลาคมถึงพฤศจิกายน พ.ศ. 2560 ประเทศไทยมีการปกครองในระบอบประชาธิปไตยอันมีพระมหากษัตริย์เป็นพระประมุข ดังนั้นสถาบันพระมหากษัตริย์จึงมีความสัมพันธ์กับประชาชนอย่างแนบแน่น โดยเฉพาะอย่างยิ่งในหลวงรัชกาลที่ 9 ทรงครองราชย์เป็นระยะเวลายาวนานกว่า 70 ปี พระราชกรณียกิจต่าง ๆ ที่เคยทรงปฏิบัติมาตลอดรัชกาล ได้ส่งผลต่อวิถีชีวิตความเป็นอยู่ของผู้คนส่วนใหญ่ในประเทศ ดังนั้นประชาชนโดยส่วนใหญ่จึงมีความรักความผูกพันเทิดทูนพระมหากษัตริย์เหนือกว่าสิ่งอื่นใด เมื่อมีเหตุการณ์ที่เกี่ยวข้องกับพระองค์ อาทิเช่น ข่าวการประชวร การเสด็จสวรรคต ตลอดจนพระราชพิธีที่เกี่ยวข้องกับพระมหากษัตริย์ จึงมีผลกับประชาชนอย่างยิ่ง ทั้งทางตรงและทางอ้อม ส่งผลต่อสภาพจิตใจ สังคม และเศรษฐกิจภายในประเทศ ดังนั้นการส่งข่าวถึงกันในเรื่องข่าวคราวต่าง ๆ ที่เกี่ยวข้องกับสถาบันพระมหากษัตริย์จึงได้รับความสนใจมากกว่าประเด็นข่าวอื่น ๆ ซึ่งข่าวที่เกิดขึ้นไม่ว่าจะเป็นข่าวจริงหรือข่าวปลอมย่อมส่งผลต่อความรู้สึกของประชาชนอย่างมากมาย

ข้อมูลที่เกี่ยวข้องกับเหตุการณ์ปรากฏการณ์ทางธรรมชาติเป็นเรื่องที่ส่งผลกระทบต่อชีวิตความเป็นอยู่ของชาวบ้าน หากเกิดภัยพิบัติจากปรากฏการณ์ทางธรรมชาติโดยมิได้มีการเตรียมตัวระวังป้องกันภัยไว้ล่วงหน้า จะก่อให้เกิดความเดือดร้อน ส่งผลให้เกิดความเสียหายต่อทรัพย์สินได้ ดังนั้นข้อมูลข่าวสารด้านการแจ้งเตือนภัยธรรมชาติจึงเป็นเรื่องที่ควรให้ความสำคัญเป็นอย่างยิ่ง ข่าวต่าง ๆ ที่เกิดในช่วงเหตุการณ์ภัยธรรมชาติมีทั้งข่าวจริงข่าวปลอมจำนวนมาก ดังนั้นประชาชนจึงอาจเกิดความสับสนในการเตรียมการระวังภัยได้ หัวข้อข่าวอื่น ๆ ที่มีผลต่อผู้คนสังคมโดยตรง คนส่วนใหญ่จะให้ความสำคัญรับฟังข่าวสารและแสดงความคิดเห็นกันมากกว่าประเด็นข่าวที่เป็นเรื่องไกลตัว อาทิเช่น ข่าวเรื่องการเสียภาษี การลดหย่อนภาษี การจ่ายเงินสมทบเข้ากองทุนประกันสังคม ซึ่งในงานวิจัยนี้ได้รวบรวมจัดเก็บข้อมูลข่าวทั่วไปไว้ด้วย

จากหัวข้อเรื่องที่เกี่ยวกับเหตุการณ์ข่าวปรากฏการณ์ทางธรรมชาติ ข่าวทั่วไป และข่าวที่เกี่ยวข้องกับเหตุการณ์พระราชพิธีพระบรมศพรัชกาลที่ 9 ปรากฏว่า เมื่อได้วิเคราะห์คุณลักษณะของ

ข้อมูลข่าวที่จัดเก็บมา ส่วนใหญ่ไม่พบบัญชีทวิตเตอร์ที่มีสถานะ IsVerified เป็น TRUE ในบัญชีทวิตเตอร์ที่โพสต์ข้อความ ส่วนค่าของจำนวนผู้ที่ติดตามบัญชีผู้ใช้ (FollowersCount) ที่มีค่ามากที่สุด มีค่ามากกว่าร้อยละ 40 และจำนวนบัญชีที่ผู้ใช้กำลังติดตาม (FriendsCount) ของบัญชีทวิตเตอร์มีค่ามากที่สุด มีค่ามากกว่าร้อยละ 60 ทั้งที่โพสต์ข่าวจริงและข่าวปลอมมีค่ามากที่สุดอยู่ในช่วงระหว่าง 100-999 คน เช่นเดียวกัน ค่าของจำนวนสิ่งที่เคยกดถูกใจ (FavouritesCount) ของบัญชีทวิตเตอร์ทั้งที่โพสต์ข่าวจริงและข่าวปลอมมีค่ามากที่สุดอยู่ในช่วงระหว่าง 1,000-9,999 โดยมีค่ามากที่สุดมากกว่าร้อยละ 30 ค่าของจำนวนสถานะที่เจ้าของบัญชีเคยแสดงมา (StatusesCount) ทั้งบัญชีทวิตเตอร์ที่โพสต์ข่าวจริงและข่าวปลอมมีค่าเป็นไปในแนวทางเดียวกัน คือมีค่าอยู่ในช่วงระหว่าง 10,000-99,999 ครั้ง มีค่ามากที่สุดร้อยละ 44.40 และร้อยละ 37.87 ตามลำดับ เมื่อพิจารณาวันที่สร้างบัญชีทวิตเตอร์ (CreatedDate) ของบัญชีทวิตเตอร์ที่โพสต์ข่าวจริงส่วนใหญ่ (ร้อยละ 18.68) เป็นบัญชีที่มีอายุการใช้งานน้อยกว่า 0.5 ปี แต่สำหรับบัญชีทวิตเตอร์ที่โพสต์ข่าวปลอมส่วนใหญ่ (ร้อยละ 29.68) เป็นบัญชีที่มีอายุการใช้งานระหว่าง 4-4.5 ปี ข้อความที่พบในข่าว (MessageText) ของบัญชีทวิตเตอร์ที่แสดงข่าวจริง มีสัดส่วนของข้อความที่เขียนเองต่อข้อความรีทวิตที่ส่งต่อเป็นร้อยละ 8.17 ต่อ 91.83 น้อยกว่าบัญชีทวิตเตอร์ที่โพสต์ข่าวปลอม ซึ่งมีสัดส่วนระหว่างข้อความที่เขียนเองกับข้อความรีทวิตซ้ำจากคนอื่น เป็นร้อยละ 21.89 ต่อ 78.11 จำนวนครั้งของการส่งต่อข้อความ (RetweetCount) ของบัญชีทวิตเตอร์ที่โพสต์ข่าวจริงมีค่ามากที่สุดอยู่ในช่วงระหว่าง 1,000-9,999 ครั้ง แต่บัญชีทวิตเตอร์ที่โพสต์ข่าวปลอมมีการรีทวิตซ้ำมากที่สุดเพียง 10-99 ครั้งเท่านั้น จากชุดข้อมูลที่จัดเก็บมาได้พบว่าวันเวลาที่โพสต์ข้อความข่าว (TweetCreatedDate) ของบัญชีทวิตเตอร์ที่แสดงข่าวจริงส่วนน้อยที่สุด (ร้อยละ 9.94) แสดงในช่วงเวลา 18.01น.-24.00 น. เนื่องจากในช่วงเวลาดังกล่าวเป็นเวลาที่คนส่วนใหญ่พักผ่อน นั่งเล่น ดูหนัง ฟังเพลง ใช้เวลาไปกับการท่องอินเทอร์เน็ต หลังจากเสร็จสิ้นจากภารกิจประจำวัน แต่จากชุดข้อมูลที่จัดเก็บมาพบว่าบัญชีทวิตเตอร์ที่โพสต์ข่าวปลอมส่วนใหญ่จะโพสต์ข้อความในช่วงกลางวันระหว่างเวลา 06.00-18.00น. ซึ่งน่าจะสันนิษฐานได้ว่า ผู้ส่งข่าวปลอมมีความตั้งใจที่จะให้คนอื่นได้เห็นข่าวนั้นในช่วงเวลาหลังเลิกงานเวลาเย็น อีกข้อสังเกตหนึ่งที่ได้จากการจัดเก็บข้อมูลจากบัญชีทวิตเตอร์ที่โพสต์ข่าวจริงและข่าวปลอมพบว่า จากข้อความที่โพสต์ทั้งหมดโดยประมาณครึ่งหนึ่งในข้อความข่าวจริงคือร้อยละ 49.65 และร้อยละ 53.25 ของข้อความข่าวปลอม ไม่ปรากฏสัญลักษณ์เครื่องหมาย HashTags (#) ในข้อความข่าว ซึ่งเป็นที่น่าแปลกใจเป็นอย่างยิ่ง จึงสันนิษฐานว่าอาจเกิดจากหัวข้อข่าวเหล่านั้นต้องการแสดงรายละเอียดเนื้อหาข่าวมากกว่าการเน้นหรือกล่าวอ้างถึงประเด็นอื่นด้วยการใช้เครื่องหมาย # จำนวนจุดเชื่อมโยงข้อมูล (URL) ที่พบทั้งบัญชีทวิตเตอร์ข่าวจริงและข่าวปลอม ส่วนใหญ่พบว่ามีจำนวนจุดเชื่อมโยงที่สามารถเข้าถึงแหล่งข้อมูลมากกว่า 3 ลิงค์ ส่วนของการระบุถึงบัญชีผู้ใช้อื่น (Mentions)

และจำนวนการระบุถึงบัญชีผู้ใช้อื่น (Number of Mentions) ของบัญชีผู้ใช้ทวิตเตอร์ที่แสดงข่าวจริงและข่าวปลอม ส่วนใหญ่มีหนึ่งการอ้างอิงถึง

จากผลการวิเคราะห์ข้อมูลข่าวปลอมที่ใช้ในงานวิจัยนี้ สามารถสรุปได้ว่าคุณลักษณะของข่าวปลอมที่ปรากฏในชุดข้อมูลมีลักษณะดังต่อไปนี้

ข่าวปลอมที่พบส่วนมาก 89.35% จากข้อมูลข่าวปลอมที่พบในชุดข้อมูลทั้งหมดมาจากบัญชีผู้ใช้งานที่เปิดมานาน (ขนาด Id ยาว 9-10 ตัว) โดยอายุการใช้งานเฉลี่ยของบัญชีผู้ใช้งานทวิตเตอร์อยู่ในช่วงระหว่าง 4-4.5 ปี ชื่อบัญชีที่ใช้ส่วนใหญ่เป็นตัวอักขระผสมภาษาไทยภาษาอังกฤษ ตัวเลขมากกว่าแบบอื่น ๆ และเกือบครึ่งหนึ่งของบัญชีที่โพสต์ข่าวปลอมร้อยละ 48.82 มีผู้ติดตามและมีจำนวนบัญชีที่มีการติดตามบัญชีอื่น ๆ อยู่ในช่วงเดียวกันคืออยู่ในช่วงระหว่าง 100-999 บัญชี และกดถูกใจสิ่งที่ชอบมากระหว่าง 1,000-9,999 บัญชีผู้ใช้ บัญชีที่แสดงการส่งข้อความที่เป็นข่าวปลอมส่วนใหญ่มีจำนวนสถานะของการส่งข้อความทวีตออกไปแล้วของเจ้าของบัญชีนับแต่เริ่มเปิดบัญชีผู้ใช้งานอยู่ในช่วงระหว่าง 10,000-99,999 ครั้ง เป็นจำนวนมากถึงร้อยละ 37.87 และมีจำนวนผู้ใช้งานที่ส่งข่าวปลอมทำการส่งข้อความไปแล้ว 100,000-999,999 ข้อความมากถึงร้อยละ 34.02 รายละเอียดของชื่อบัญชีผู้ใช้ส่วนใหญ่เลือกใช้ตัวอักขระผสมมากกว่า การใช้ตัวอักขระแบบภาษาไทยทั้งหมดหรือภาษาอังกฤษทั้งหมด นอกเหนือจากส่วนของข้อมูลที่แสดงตำแหน่งเป็นตำแหน่งอื่น ๆ ที่ไม่อยู่ในตารางที่ 3.13 แล้วพบว่าบัญชีผู้ใช้งานส่วนใหญ่เป็นบัญชีที่พบในประเทศไทยและเอเชียตะวันออกเฉียงใต้มากกว่าบริเวณอื่น ๆ ซึ่งมีความสอดคล้องกับชุดข้อมูลที่จัดเก็บจากหัวข้อของข้อความข่าวที่เป็นภาษาไทย แต่จากข้อมูลทั้งหมดพบว่าไม่มีการระบุบริเวณตำแหน่งเวลาในข้อความที่เป็นข่าวปลอมถูกนำไปเผยแพร่ใหม่ (retweeted) เพียง 10-99 ครั้ง เนื่องจากความจริงปรากฏแล้ว ข้อความเหล่านั้นจึงไม่มีการเผยแพร่ซ้ำออกไปอีก อีกข้อสังเกตหนึ่งที่พบคือข้อความข่าวปลอมที่จะโพสต์ส่วนใหญ่ไม่ได้เป็นข้อความที่เจ้าของบัญชีผู้ใช้สร้างขึ้นเอง แต่เป็นการส่งต่อข้อความจากที่มีเคยมีการส่งไปแล้วมากกว่า และการโพสต์ครั้งแรกส่วนใหญ่ทำในช่วงเวลากลางวันมากกว่ากลางคืน

การเรียนรู้ของเครื่องในการตรวจจับข่าวปลอมไม่ใช่เรื่องยาก หากข้อมูลที่ใช้ในการจัดหมวดหมู่เป็นข้อมูลที่มีคำตอบที่ชัดเจน [38] ผลการวิจัยนี้ใช้วิธีการเรียนรู้ด้วยเครื่องเพื่อการจำแนกข่าวปลอมสามวิธีการ ได้แก่ Naïve Bayes, Neural Network และ Support Vector Machine ซึ่งความถูกต้องของการจำแนกข่าวปลอม สำหรับการจำแนกข่าวปลอมของชุดข้อมูลตัวอย่างนี้ได้ดี จึงเป็นไปในแนวทางเช่นเดียวกับงานวิจัยของ [79] ที่เลือกใช้วิธีการ Support Vector Machine ในการวิเคราะห์ความน่าเชื่อถือของเหตุการณ์ที่เป็นข่าว นอกจากนั้นยังมีงานวิจัยของ [26] [65, 69, 72] ที่ใช้วิธีการเรียนรู้ด้วยเครื่อง Naïve Bayes, Neural Network ในการจำแนกข่าวปลอมเช่นกัน รวมถึงเริ่มมีการใช้ deep Neural Network ในการปรับปรุงประสิทธิภาพการทำงาน

ข้อสังเกตที่ได้จากการพิจารณาข่าวปลอมที่อยู่ในทวิตเตอร์พบว่า ข่าวปลอมที่เกิดขึ้นส่วนใหญ่เกิดขึ้นในระยะเวลาสั้น ๆ แล้วจะหายไปเมื่อความจริงปรากฏ จากผลการเก็บข้อมูลของงานวิจัยนี้พบว่าระยะเวลาที่ใช้ในการเผยแพร่ข่าวจริงและข่าวปลอม มีความแตกต่างกันอย่างมีนัยยะสำคัญ โดยระยะเวลาเฉลี่ยที่มีการเผยแพร่ข่าวปลอมเป็น 5 วัน 1 ชั่วโมง 19 นาที ซึ่งสั้นกว่าระยะเวลาเฉลี่ยที่มีการเผยแพร่ข่าวจริง 7 วัน 7 ชั่วโมง 13 นาที ค่าความแปรปรวนของระยะเวลาเฉลี่ยที่ใช้ในการเผยแพร่ข่าวจริงเป็น 7.34 และข่าวปลอมเป็น 4.78 ดังนั้นจึงกล่าวได้ว่าระยะเวลาวงจรชีวิตของข่าวปลอมสั้นกว่าข่าวจริง ช่วงระยะเวลาที่ข่าวปลอมมีการเผยแพร่กันในช่วงข่าวที่เผยแพร่อย่างรวดเร็วบนทวิตเตอร์ บางข่าวปลอมมีการแพร่กระจายกันนานจนกระทั่งความจริงปรากฏขึ้นแล้วข่าวปลอมนั้นจะหายไปอย่างเงียบ ๆ แต่ปัญหาคือตราบใดที่ยังไม่มีการเปิดเผยความจริง ข่าวนั้นจะยังคงเผยแพร่ต่อไปเรื่อย ๆ ซึ่งจะทำให้เกิดความเข้าใจผิด และในบางประเด็นหากที่เป็นหัวข้อข่าวที่เป็นเรื่องที่สังคมมีความอ่อนไหว อาจสร้างความเสียหายต่อผู้คนและสังคมได้

จากข้อมูลในงานวิจัยนี้พบว่า ช่วงระยะเวลาที่ข่าวปลอมมีการเผยแพร่ในระยะเวลาอันจำกัดและเป็นเพียงช่วงเวลาสั้น ๆ มากกว่าการเผยแพร่ข่าวจริงที่มีการส่งต่อเผยแพร่กันอย่างยาวนานกว่า ซึ่งสอดคล้องกับผลงานวิจัยของ [80]

คุณลักษณะที่ส่งผลต่อจำแนกข่าวปลอมออกจากข่าวจริงในเครือข่ายสังคมออนไลน์ทวิตเตอร์มีผลต่อค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมของชุดข้อมูลนี้ ได้แก่ Mentions, RetweetCount และ MessageText จากการทดลองลดค่ามากกว่าหนึ่งคุณลักษณะแล้วนำไปผ่านกระบวนการจำแนกข่าวปลอมด้วยวิธีการเรียนรู้ด้วยเครื่อง Naïve Bayes, Neural Network และ Support Vector Machine แล้วได้ผลค่าร้อยละความถูกต้องของการจำแนกข่าวปลอมมีค่ามากขึ้น

## 5.2 ข้อจำกัดและแนวทางวิจัยในอนาคต

ข้อจำกัดของวิธีการที่ใช้ในงานวิจัยนี้คือสามารถใช้งานได้ดีกับข่าว "จริง" ที่ไม่ได้ใช้เนื้อหาของข่าวในการวิเคราะห์ วิธีการเรียนรู้ด้วยเครื่องสำหรับการจำแนกข่าวปลอมใช้การแปลงข้อมูลจากกฎการแปลงข้อมูลแปลงให้เป็นตัวเลข โดยไม่ได้พิจารณาเนื้อหาของข่าว ในอนาคตเพื่อขยายขอบเขตของข่าวที่สามารถจำแนกได้ จะใช้วิธีการวิเคราะห์คำความหมายของเนื้อข่าวมาร่วมพิจารณา

การเรียนรู้ด้วยเครื่องยังมีอีกข้อจำกัดหนึ่งคือ กรณีที่มีข้อมูลที่ไม่รู้จักจำนวนมากแบบจำลองจะไม่สามารถจำแนกข้อมูลได้อย่างถูกต้อง หากไม่มีข้อมูลการเทรนนิ่งในจำนวนมากเพียงพอ ดังนั้นจึงจำเป็นต้องมีการรวบรวมข้อมูลจำนวนมากเพื่อรองรับข่าวที่หลากหลายมากขึ้น

เนื่องจากระยะเวลาอันจำกัดที่สามารถดึงข้อมูลย้อนหลังจากเครือข่ายสังคมออนไลน์ทวิตเตอร์เพียงไม่เกิน 7 วันย้อนหลัง ดังนั้นการเก็บข้อมูลข่าวหัวข้อใด ๆ ก็ตาม หากเป็นเรื่องที่ยังอยู่ในความสนใจที่ยังมีการพูดถึงกันในสังคม จะต้องทำการเก็บข้อมูลในหัวข้อเดิมซ้ำ ๆ กันอย่างต่อเนื่อง เพื่อให้

ได้ข้อมูลที่เกี่ยวข้องมากที่สุด นอกจากนั้นข่าวปลอมที่เกิดมานั้นยังมีระยะเวลาจำกัด เนื่องจากเมื่อความจริงปรากฏขึ้นข่าวปลอมเหล่านั้นก็จะหายไป ไม่มีใครกล่าวถึงอีก ดังนั้นการจะเก็บข้อมูลข่าวปลอมจะต้องเฝ้าคอยเก็บหลาย ๆ หัวข้อข่าว โดยที่ในระหว่างการเก็บข้อมูลจะยังไม่ทราบล่วงหน้าว่าหัวข่าวนั้นจะเป็นข่าวจริงหรือข่าวปลอม อีกทั้งขึ้นอยู่กับประเด็นที่สนใจอีกด้วย

ข้อจำกัดในงานวิจัยนี้คือแบบจำลองที่ใช้ในการระบุข่าวปลอมที่สร้างขึ้น เกิดจากการรวบรวมข้อมูลในช่วงระยะเวลาไม่นานมาก ซึ่งชุดข้อมูลที่นำมาใช้เป็นข่าวในหัวข้อจำกัด ดังนั้นเมื่อเวลาเปลี่ยนแปลงไป ปัญหาข่าวต่าง ๆ มีการเปลี่ยนแปลงไป แบบจำลองที่สร้างขึ้นนี้ อาจไม่สามารถใช้ในการจำแนกข่าวปลอมได้อย่างมีประสิทธิภาพ

ในอนาคตอาจมีข้อมูลเพิ่มเติมที่ใช้ในการสร้างแบบจำลองที่ดียิ่งขึ้น เพื่อปรับปรุงประสิทธิภาพการจำแนกข่าวปลอมให้สามารถรองรับประเด็นข่าวที่มากขึ้น และทันสมัยเข้ากับข่าวปัจจุบัน

# บรรณานุกรม


[1] ส. อ. ห. สพธอ., "รู้จัก พ.ร.บ. คอมพิวเตอร์ฯ ฉบับที่ 2 พ.ศ. 2560," 2016. [Online]. Available: https://ictlawcenter.etda.or.th/news/detail/computer-2559.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," *SIGKDD Explor. Newsl.,* vol. 19, no. 1, pp. 22-36, 2017, doi: 10.1145/3137597.3137600.

[3] K. Niklewicz, "Weeding out fake news: an approach to social media regulation," *European View,* vol. 16, no. 2, pp. 335-335, 2017/12/01 2017, doi: 10.1007/s12290-017-0468-0.

[4] A. Campan, A. Cuzzocrea, and T. M. Truta, "Fighting fake news spread in online social networks: Actual trends and future research directions," in *2017 IEEE International Conference on Big Data (Big Data)*, 11-14 Dec. 2017 2017, pp. 4453-4457, doi: 10.1109/BigData.2017.8258484.

[5] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards Detecting Compromised Accounts on Social Networks," *IEEE Transactions on Dependable and Secure Computing,* vol. 14, no. 4, pp. 447-460, 2017, doi: 10.1109/TDSC.2015.2479616.

[6] A. Ehsanfar and M. Mansouri, "Incentivizing the dissemination of truth versus fake news in social networks," in *2017 12th System of Systems Engineering Conference (SoSE)*, 18-21 June 2017 2017, pp. 1-6, doi: 10.1109/SYSOSE.2017.7994981.

[7] H. Bansal and M. Misra, "Sybil Detection in Online Social Networks (OSNs)," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 27-28 Feb. 2016 2016, pp. 569-576, doi: 10.1109/IACC.2016.111.

[8] P. Jain and V. Singh, "CredRank: Evaluating tweet credibility during high impact events," in *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 14-17 Dec. 2016 2016, pp. 553-557, doi: 10.1109/IC3I.2016.7918025.

[9] H. Berghel, "Alt-News and Post-Truths in the "Fake News" Era," *Computer,* vol. 50, no. 4, pp. 110-114, 2017, doi: 10.1109/MC.2017.104.

[10] C. Castillo, M. Mendoza, and B. Poblete, *Information credibility on Twitter*. 2011, pp. 675-684.

[11] W. A. S. a. Hootsuite, "Digital in 2019 Global Digital Overview," 2019. [Online]. Available: https://wearesocial.com/global-digital-report-2019.

[12] WorkpointTV, "หุ้นไทยวิกฤต! ภาคบ่ายร่วงติดลบเกือบ 100 จุด คาดสถานการณ์ความมั่นคงในประเทศ," 2016. [Online]. Available: https://www.facebook.com/WorkpointNews/posts/321351604900791/.

[13] M. Alrubaian, M. Al-Qurishi, M. M. Hassan, and A. Alamri, "A Credibility Analysis System for Assessing Information on Twitter," *IEEE Transactions on Dependable and Secure Computing,* vol. 15, no. 4, pp. 661-674, 2018, doi: 10.1109/TDSC.2016.2602338.

[14] S. Phartiyal. "India asks WhatsApp to curb spread of false messages." The Thomson Reuters https://www.reuters.com/article/us-whatsapp-india-fakenews/india-asks-whatsapp-to-curb-spread-of-false-messages-idUSKBN1JT1Z2 (accessed 2018).

[15] Y. L. Mari Saito. "Taiwan representative in Japan's Osaka commits suicide." The Thomson Reuters. https://www.reuters.com/article/us-japan-taiwan/taiwan-representative-in-japans-osaka-commits-suicide-idUSKCN1LV067 (accessed 2018).

[16] E. S. a. J. Gottfried. "News Use Across Social Media Platforms 2017." Pew Research Center. http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/ (accessed 2018).

[17] M. G. Hunt Allcott, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives,* vol. 31, no. 2, pp. 211-236, 2017.

[18] S. Y. Bhat and M. Abulaish, "Community-based features for identifying spammers in Online Social Networks," in *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)*, 25-28 Aug. 2013 2013, pp. 100-107, doi: 10.1145/2492517.2492567.

[19] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science,* vol. 359, no. 6380, p. 1146, 2018, doi: 10.1126/science.aap9559.

[20] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and Resolution of Rumours in Social Media: A Survey " presented at the 2017 12th System of Systems Engineering Conference (SoSE), Waikoloa, HI, 2018.

[21] D. G. Young, K. Hall Jamieson, S. Poulsen, and A. Goldring, *Fact-Checking Effectiveness as a Function of Format and Tone: Evaluating FactCheck.org and FlackCheck.org*. 2017, p. 107769901771045.

[22] W. Vorhies, "Using Algorithms to Detect Fake News – The State of the Art," *Data Science Central* May 1, 2017 at 3:30pm 2017. [Online]. Available: https://www.datasciencecentral.com/profiles/blogs/using-algorithms-to-detect-fake-news-the-state-of-the-art.

[23] M. AlRubaian, M. Al-Qurishi, M. Al-Rakhami, S. M. M. Rahman, and A. Alamri, "A Multistage Credibility Analysis Model for Microblogs," presented at the Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, Paris, France, 2015.

[24] R. El Ballouli, W. El-Hajj, A. Ghandour, S. Elbassuoni, H. Hajj, and K. Shaban, "CAT: Credibility Analysis of Arabic Content on Twitter," Valencia, Spain, apr 2017: Association for Computational Linguistics, in Proceedings of the Third Arabic Natural Language Processing Workshop, pp. 62-71, doi: 10.18653/v1/W17-1308. [Online]. Available: https://www.aclweb.org/anthology/W17-1308

[25] R. Mouty and A. Gazdar, "Survey on Steps of Truth Detection on Arabic Tweets," in *2018 21st Saudi Computer Society National Computer Conference (NCC)*, 25-26 April 2018 2018, pp. 1-6, doi: 10.1109/NCG.2018.8593060.

[26] M. Granik and V. Mesyura, "Fake news detection using naive Bayes classifier," in *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, 29 May-2 June 2017 2017, pp. 900-903, doi: 10.1109/UKRCON.2017.8100379.

[27] V. Rubin, Y. Chen, and N. Conroy, *Deception Detection for News: Three Types of Fakes*. 2015.

[28] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Computer Speech & Language,* vol. 28, no. 1, pp. 20-37, 2014/01/01/ 2014, doi: https://doi.org/10.1016/j.csl.2013.03.001.

[29] M.-P. S. Chan, jones, hall, and D. Albarracin, *Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation*. 2017.

[30] Y. U. Chandra, Surjandy, and Ernawaty, "Higher education student behaviors in spreading fake news on social media: A case of LINE group," in *2017 International Conference on Information Management and Technology (ICIMTech)*, 15-17 Nov. 2017 2017, pp. 54-59, doi: 10.1109/ICIMTech.2017.8273511.

[31] Á. Figueira and L. Oliveira, *The current state of fake news: challenges and opportunities*. 2017, pp. 817-825.

[32] สำนักงานราชบัณฑิตยสภา, "ระบบค้นหาคำศัพท์," พจนานุกรม ฉบับราชบัณฑิตยสถาน พ.ศ. ๒๕๕๔, 2011. [Online]. Available: http://www.royin.go.th/dictionary/index.php.

[33] S. J. Herschel, "The Great Moon Hoax," *New York Sun,* 1835. [Online]. Available: http://hoaxes.org/archive/permalink/the_great_moon_hoax.

[34] D. Bufnea and D. Şotropa, "A Community Driven Approach for click Bait Reporting," in *2018 26th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 13-15 Sept. 2018 2018, pp. 1-6, doi: 10.23919/SOFTCOM.2018.8555759.

[35] S. B. Parikh and P. K. Atrey, "Media-Rich Fake News Detection: A Survey," presented at the 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 2018.

[36] B. Nyhan and J. Reifler, *Displacing Misinformation about Events: An Experimental Test of Causal Corrections*. 2015, pp. 1-13.

[37] A. Algarni, Y. Xu, and T. Chan, "Measuring Source Credibility of Social Engineering Attackers on Facebook," presented at the 2016 49th Hawaii International Conference on System Sciences (HICSS), 2016.

[38] J. Hyman, "Addressing fake news: Open standards & easy identification," in *2017 IEEE 8th Annual Ubiquitous Computing, Electronics and Mobile Communication*

*Conference (UEMCON)*, 19-21 Oct. 2017 2017, pp. 63-69, doi: 10.1109/UEMCON.2017.8248986.

[39] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel Visual and Statistical Image Features for Microblogs News Verification," *IEEE Transactions on Multimedia,* vol. 19, no. 3, pp. 598-608, 2017, doi: 10.1109/TMM.2016.2617078.

[40] H. S. Al-Ash and W. C. Wibowo, "Fake News Identification Characteristics Using Named Entity Recognition and Phrase Detection," in *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*, 24-26 July 2018 2018, pp. 12-17, doi: 10.1109/ICITEED.2018.8534898.

[41] A. Dey, R. Z. Rafi, S. H. Parash, S. K. Arko, and A. Chakrabarty, "Fake News Pattern Recognition using Linguistic Analysis," in *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, 25-29 June 2018 2018, pp. 305-309, doi: 10.1109/ICIEV.2018.8641018.

[42] S. Gheewala and R. Patel, "Machine Learning Based Twitter Spam Account Detection: A Review," in *2018 Second International Conference on Computing Methodologies and Communication (ICCMC)*, 15-16 Feb. 2018 2018, pp. 79-84, doi: 10.1109/ICCMC.2018.8487992.

[43] G. B. Guacho, S. Abdali, N. Shah, and E. E. Papalexakis, "Semi-supervised Content-Based Detection of Misinformation via Tensor Embeddings," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 28-31 Aug. 2018 2018, pp. 322-325, doi: 10.1109/ASONAM.2018.8508241.

[44] N. Kim, D. Seo, and C. Jeong, "FAMOUS: Fake News Detection Model Based on Unified Key Sentence Information," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, 23-25 Nov. 2018 2018, pp. 617-620, doi: 10.1109/ICSESS.2018.8663864.

[45] H. Nurrahmi and D. Nurjanah, "Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 6-7 March 2018 2018, pp. 543-548, doi: 10.1109/ICOIACT.2018.8350758.

[46] M. L. D. Vedova, E. Tacchini, S. Moret, G. Ballarin, M. DiPierro, and L. d. Alfaro, "Automatic Online Fake News Detection Combining Content and Social Signals," in *2018 22nd Conference of Open Innovations Association (FRUCT)*, 15-18 May 2018 2018, pp. 272-279, doi: 10.23919/FRUCT.2018.8468301.

[47] V. Bakir and A. McStay, *Fake News and The Economy of Emotions: Problems, causes, solutions*. 2017, pp. 1-22.

[48] SpringNews, "ดิสเครดิตรัฐบาล!! 3 ข่าวปลอมทำสังคมสับสน," 2019. [Online]. Available: https://www.springnews.co.th/truth/truth-election-62/485023

[49] F. Benamara, C. Bosco, E. Fersini, G. Pasi, V. Patti, and M. Viviani, "SeCredISData 2018: Special Session on Sentiment, Emotion, and Credibility of Information in Social Data," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 1-3 Oct. 2018 2018, pp. 638-640, doi: 10.1109/DSAA.2018.00082.

[50] S. B. Fazili and M. Ahmad, "Guassian Gradient Descent Model for Trust Inference in Imbalanced Data," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, 14-15 June 2018 2018, pp. 929-934, doi: 10.1109/ICCONS.2018.8663243.

[51] J. Hodson and B. Traynor, "Design Exploration of Fake News: A Transdisciplinary Methodological Approach to Understanding Content Sharing and Trust on Social Media," in *2018 IEEE International Professional Communication Conference (ProComm)*, 22-25 July 2018 2018, pp. 1-5, doi: 10.1109/ProComm.2018.00008.

[52] C. M. M. Kotteti, X. Dong, N. Li, and L. Qian, "Fake News Detection Enhancement with Data Imputation," in *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, 12-15 Aug. 2018 2018, pp. 187-192, doi: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00042.

[53] K. Ootani and H. Yamana, "External Content-dependent Features for Web Credibility Evaluation," in *2018 IEEE International Conference on Big Data (Big Data)*, 10-13 Dec. 2018 2018, pp. 5414-5416, doi: 10.1109/BigData.2018.8622398.

[54] ว. ศรีโหมด, ""หยุดคิด เช็คต้นตอ" หนทางยุติ "ข่าวปลอม" ที่กำลังเป็นปัญหาโลก," *PostToday.*, 2019. [Online]. Available: https://www.posttoday.com/social/general/543708.

[55] Facebook. "เคล็ดลับในการสังเกตข่าวปลอม." https://www.facebook.com/help/spotfalsenews (accessed 2019).

[56] S. R. Guruvayur and R. Suchithra, "A detailed study on machine learning techniques for data mining," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, 11-12 May 2017 2017, pp. 1187-1192, doi: 10.1109/ICOEI.2017.8300900.

[57] I. Y. R. Pratiwi, R. A. Asmara, and F. Rahutomo, "Study of hoax news detection using naïve bayes classifier in Indonesian language," in *2017 11th International Conference on Information & Communication Technology and System (ICTS)*, 31-31 Oct. 2017 2017, pp. 73-78, doi: 10.1109/ICTS.2017.8265649.

[58] Wikipedia. "Naive Bayes classifier." https://en.wikipedia.org/wiki/Naive_Bayes_classifier (accessed 2019).

[59] Wikipedia. "Artificial neural network." https://en.wikipedia.org/wiki/Artificial_neural_network (accessed 2019).

[60] A. Pal and A. Y. K. Chua, "Classification of rumors and counter-rumors," in *2018 4th International Conference on Information Management (ICIM)*, 25-27 May 2018 2018, pp. 81-85, doi: 10.1109/INFOMAN.2018.8392814.

[61] Wikipedia. "Support-vector machine." https://en.wikipedia.org/wiki/Support-vector_machine (accessed 2019).

[62] M. Jones - Jang *et al.*, *A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis*. 2018, pp. 103-113.

[63] M. Verstraete, D. E. Bambauer, and J. R. Bambauer, *Identifying and Countering Fake News*. 2017.

[64] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, *Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads*. 2016.

[65] S. K. a. M. Chen, "Identifying Tweets with Fake News," presented at the 2018 IEEE International Conference on Information Reuse and Integration (IRI), Salt Lake City, UT, USA, 2018. [Online]. Available: doi.ieeecomputersociety.org/10.1109/IRI.2018.00073.

[66] C. Buntain and J. Golbeck, "Automatically Identifying Fake News in Popular Twitter Threads," in *2017 IEEE International Conference on Smart Cloud (SmartCloud)*, 3-5 Nov. 2017 2017, pp. 208-215, doi: 10.1109/SmartCloud.2017.40.

[67] Y. Qin, W. Dominik, and C. Tang, "Predicting Future Rumours," *Chinese Journal of Electronics,* vol. 27, no. 3, pp. 514-520, 2018, doi: 10.1049/cje.2018.03.008.

[68] N. Conroy, V. Rubin, and Y. Chen, "Automatic Deception Detection: Methods for Finding Fake News," in *ASIS&T2015*, St. Louis, MO, USA, 2015, 2015. [Online]. Available: https://www.researchgate.net/publication/281818865_Automatic_Deception_Detection_Methods_for_Finding_Fake_News. [Online]. Available: https://www.researchgate.net/publication/281818865_Automatic_Deception_Detection_Methods_for_Finding_Fake_News

[69] L. Krzysztof, S.-W. Jacek, J.-L. Michal, and G. Amit, "Automated Credibility Assessment on Twitter," *Computer Science,* vol. 16, no. 2, 2015, doi: 10.7494/csci.2015.16.2.157.

[70] M. Thandar and S. Usanavasin, "Measuring Opinion Credibility in Twitter," in *Recent Advances in Information and Communication Technology 2015*, Cham, H. Unger, P. Meesad, and S. Boonkrong, Eds., 2015// 2015: Springer International Publishing, pp. 205-214.

[71] D. Wang, L. Kaplan, T. Abdelzaher, and C. C. Aggarwal, "On Credibility Estimation Tradeoffs in Assured Social Sensing," *IEEE Journal on Selected Areas in Communications,* vol. 31, no. 6, pp. 1026-1037, 2013, doi: 10.1109/JSAC.2013.130605.

[72] S. T. Allaparthi, G. Yaparla, and V. Pudi, "Sentiment and Semantic Deep Hierarchical Attention Neural Network for Fine Grained News Classification," in *2018 IEEE International Conference on Big Knowledge (ICBK)*, 17-18 Nov. 2018 2018, pp. 65-72, doi: 10.1109/ICBK.2018.00017.

[73] O. Ajao, D. Bhowmik, and S. Zargari, *Fake News Identification on Twitter with Hybrid CNN and RNN Models*. 2018.

[74] D. Berkowitz and D. A. Schwartz, "Miley, CNN and The Onion," *Journalism Practice,* vol. 10, no. 1, pp. 1-17, 2016/01/02 2016, doi: 10.1080/17512786.2015.1006933.

[75] S. Gilda, "Evaluating machine learning algorithms for fake news detection," in *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, 13-14 Dec. 2017 2017, pp. 110-115, doi: 10.1109/SCORED.2017.8305411.

[76] I. P. Benitez, A. M. Sison, and R. P. Medina, "An improved genetic algorithm for feature selection in the classification of Disaster-related Twitter messages," in *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, 28-29 April 2018 2018, pp. 238-243, doi: 10.1109/ISCAIE.2018.8405477.

[77] S. Girgis, E. Amer, and M. Gadallah, "Deep Learning Algorithms for Detecting Fake News in Online Text," in *2018 13th International Conference on Computer Engineering and Systems (ICCES)*, 18-19 Dec. 2018 2018, pp. 93-97, doi: 10.1109/ICCES.2018.8639198.

[78] J. Hale, "Smarter Ways to Encode Categorical Data for Machine Learning Exploring Category Encoders," *TowardsDataScience,* 2018. [Online]. Available: https://towardsdatascience.com/smarter-ways-to-encode-categorical-data-for-machine-learning-part-1-of-3-6dca2f71b159.

[79] S. Krishnan and M. Chen, "Identifying Tweets with Fake News," presented at the 2018 IEEE International Conference on Information Reuse and Integration (IRI), 2018.

[80] S. H. a. H. Paulheim, "Weakly Supervised Learning for Fake News Detection on Twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Barcelona, Spain, 2018, vol. 00, pp. 274-277, doi: 10.1109/ASONAM.2018.8508520.

# บรรณานุกรม

# ประวัติผู้เขียน

| | |
|---|---|
| **ชื่อ-สกุล** | สุปัญญา อภิวงศ์โสภณ |
| **วุฒิการศึกษา** | วท.บ. (วิทยาการคอมพิวเตอร์) มหาวิทยาลัยหัวเฉียวเฉลิมพระเกียรติ<br>วท.ม. (วิทยาการสารสนเทศ) สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง<br>วศ.ด. (วิศวกรรรมคอมพิวเตอร์) จุฬาลงกรณ์มหาวิทยาลัย (กำลังศึกษา) |
| **ที่อยู่ปัจจุบัน** | 37/3 หมู่ที่ 17 ตำบลคลองนครเนื่องเขต อำเภอเมือง จังหวัดฉะเชิงเทรา |
| **ผลงานตีพิมพ์** | S. Aphiwongsophon and P. Chongstitvatana, "Detecting Fake News with Machine Learning Method," 2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), Chiang Rai, Thailand, 2018, pp. 528-531. doi: 10.1109/ECTICon.2018.8620051 |