

## บทที่ 1

### บทนำ

#### 1.1 ที่มาและความสำคัญของปัญหา

จากอดีตที่ผ่านมาสื่อดั้งเดิมที่เกิดขึ้นที่ถือเป็นยุคต้นๆ ได้แก่ หนังสือพิมพ์ วิทยุ โทรทัศน์ เป็นสื่อมวลชนที่เข้าถึงประชาชนผู้รับข่าวสารคราวละมากๆ เป็นยุคที่ผู้ผลิตสารจะเป็นผู้ให้ข้อมูล รายงานเหตุการณ์ต่างๆ ที่เกิดขึ้นไปยังผู้รับสาร ผู้รับสารได้รับเพียงทางเดียว และผู้ส่งสารไม่สามารถทราบผลกระทบที่เกิดขึ้นในเวลานั้นๆ ได้

เมื่อผู้รับสารมีความต้องการที่หลากหลายมากขึ้น ทำให้เกิดการแข่งขัน โดยนัยของการตลาดนั้นหมายถึง ความพยายามระหว่างผู้ขายที่ต้องการจะกระจายสินค้าของตนให้ได้ในราคาที่ดีที่สุด แนวทางการตลาดเป็นจุดเปลี่ยนในการทำให้ผู้ผลิตสารจะต้องปรับเปลี่ยนการนำเสนอให้มีความหลากหลายตอบสนองต่อผู้รับสารที่มีความต้องการที่แตกต่างไป ยุคที่มีหนังสือพิมพ์แบบรวมเนื้อหาหลายๆ แนวในเล่มเดียวกัน ลักษณะการนำเสนอเนื้อหาเชิงบรรณาธิการคล้ายๆ กัน (สิริทิพย์ 2543, 99) อาทิเช่น ไทยรัฐ เดลินิวส์ ซึ่งสามารถอธิบายความเป็นหนังสือพิมพ์ที่สามารถตอบสนองคนอ่านได้หลากหลาย ซึ่งเนื้อหาที่น่าสนใจมีทั้งหมดหมูที่เป็น อาชญากรรม การเมือง กีฬา บันเทิง นิยาย หนังสืออีกประเภทหนึ่งคือ หนังสือพิมพ์เฉพาะทาง อาทิเช่น มติชน ประชาชาติฐานเศรษฐกิจ สยามกีฬา ดาราเดลี เป็นต้น จะเห็นได้ว่าหนังสือเหล่านี้ จะเจาะประเด็นเฉพาะทาง อย่างเช่น มติชน จะนำเสนอข่าวการเมืองโดยตรง รวมถึงการทำงานของรัฐบาล แนวทางการจัดการบริหารบ้านเมือง ในอีกทางหนึ่งหนังสือพิมพ์สยามกีฬา ได้ให้ข้อมูลข่าวสารของแวดวงกีฬาต่าง ๆ ติดตามนักกีฬา ข่าวงในของการฝึกซ้อม เป็นต้น แต่อย่างไรก็ตามยังคงรูปแบบของผู้ผลิตสารกับผู้รับสาร ในยุคนี้ผู้รับสารสามารถเลือกได้เพียงประเภทของข่าวสารที่จะรับหรือไม่รับเท่านั้น ประชาชนไม่สามารถตอบโต้หรือแสดงความคิดเห็นลงในสื่อได้เลย (สิริทิพย์ ชันสุวรรณ, 2543, 2)

ยุคต่อมา เมื่อมีการพัฒนาการของกระบวนการสื่อสารมวลชน ทั้งขั้นตอน เทคนิคและเทคโนโลยีในการสื่อสารนั้นมีความเปลี่ยนแปลงไป โดยพัฒนาการของกระบวนการสื่อสารมวลชน มีความสัมพันธ์โดยตรงกับการค้นพบ การประดิษฐ์วัสดุอุปกรณ์เพื่อการบันทึกและเผยแพร่ ข่าวสาร และจุดเปลี่ยนที่สำคัญที่กระทบต่อวงการสื่อสารมวลชนโดยตรงก็คือการมาถึงของโลก “อินเทอร์เน็ต” (Internet) ซึ่งเป็นเครือข่ายของคอมพิวเตอร์ขนาดใหญ่ที่เชื่อมโยงเครือข่ายคอมพิวเตอร์ทั่วโลกเข้าด้วยกัน ทำให้มีการเปลี่ยนแปลงจากผู้ผลิตข่าวสารผ่านการนำเสนอบนกระดาษมาสู่โลกของแพลตฟอร์ม โดยมี “เวิลด์ไวด์เว็บ” (World Wide Web : WWW) ที่เป็นบริการค้นหาและแสดงข้อมูล

แบบมัลติมีเดียบนอินเทอร์เน็ตทุกประเภท ซึ่งข้อมูลข่าวสารจะถูกจัดเป็นหมวดหมู่ให้ง่ายต่อการค้นหา จัดอยู่ในรูปแบบของข้อความ รูปภาพ หรือ เสียง ทำให้แนวคิดการส่งจินตนาการผ่านหน้าจอคอมพิวเตอร์เข้าไปในเครือข่าย ได้รับการนำเสนอเป็นพื้นที่ซึ่งใช้ติดต่อสื่อสารกันหรือที่เรียกว่า “ไซเบอร์สเปซ” (Cyberspace) ผลกระทบของโลกาภิวัตน์ต่อการรับข้อมูลข่าวสารของประชาชนมีความสัมพันธ์กันอย่างมีนัยยะ และทำให้เกิด “สังคมเครือข่ายออนไลน์” (Social networking) สามารถเข้าถึงข้อมูลได้ง่ายดายและจับใจ ผ่านอุปกรณ์คอมพิวเตอร์เมื่ออยู่ในที่ทำงาน ผ่านแท็บเล็ตและมือถือซึ่งพกพาสะดวกได้ง่าย ทำให้ใครต่อใครก็หันไปใช้สื่อออนไลน์และโซเชียลมีเดียในการรับข้อมูลขึ้น ปรากฏการณ์เช่นนี้เกิดขึ้นทั่วโลก ผู้คนหันมาเสพเนื้อหาผ่านแพลตฟอร์มหรือสื่อโซเชียลมีเดีย เช่น Facebook, Twitter หรือ Youtube มากขึ้น

เมื่อรูปแบบการนำเสนอเนื้อหาต่าง ๆ เปลี่ยนไป พฤติกรรมของผู้เสพสื่อหรือผู้บริโภคจึงเปลี่ยนตาม การเปลี่ยนผ่านของเทคโนโลยีไปสู่โลกดิจิทัล ทำให้องค์กรสื่อสารสิ่งพิมพ์หลาย ๆ แห่งเสี่ยงจะต้องปิดตัวลง ขณะที่ภาครัฐได้ตอบสนองนโยบายไทยแลนด์ 4.0 ของรัฐบาล ที่ใช้นวัตกรรมและเทคโนโลยีในการพัฒนาประเทศได้รับการตอบรับจากหลายฝ่าย เพราะเข้ากับยุคสมัยที่ทุกอย่างมุ่งสู่โลกดิจิทัล กลายเป็นความท้าทายของสื่อที่จะต้องปรับรูปแบบจากที่เคยเป็นสื่อสิ่งพิมพ์ที่จะต้องนำเสนอข่าวสารเดิมที่ตนเคยมีให้อยู่ในรูปแบบที่ผู้อ่านชาวออนไลน์สนใจ และต้องปรับตัวให้เข้ากับความเปลี่ยนแปลงนี้ให้ได้ หรือแม้กระทั่งนักข่าว เมื่อการทำข่าวแบบเดิมไม่ตอบโจทย์อีกต่อไปและการปรับตัวเท่านั้นจึงจะอยู่รอด “นักข่าวที่สามารถปรับตัวในการรายงานให้เข้ากับหลายสื่อหลายช่องทางอย่างเช่น ทวีต วิทยู ออนไลน์ และโซเชียลมีเดีย ยังเป็นที่ต้องการของตลาดโดยเฉพาะในยุคภูมิทัศน์สื่อกำลังเปลี่ยนแปลง เช่นเดียวกันกับพฤติกรรมของผู้ชมรายการโทรทัศน์ ผู้อ่านชาวออนไลน์หนังสือพิมพ์และนิตยสาร ก็เสพสื่อในช่องทางออนไลน์มากขึ้น ในทางกลับกัน นักข่าวที่ไม่สามารถปรับตัวกับความเปลี่ยนแปลงได้ มีโอกาสถูกเลิกจ้างได้สูงมากเช่นกัน” (มานะ ตรีรยาภิวัฒน์, 2560)

อย่างไรก็ตาม ประเทศไทยมีสื่อออนไลน์หน้าใหม่เกิดขึ้นมากมายหลากหลายสำนัก อาทิ เดอะแมทเทอร์, เดอะโมเมนต์ ฯลฯ ยังรวมถึงบรรดาแอตมินเพจเฟซบุ๊กต่าง ๆ ขณะที่สื่อสิ่งพิมพ์เดิมก็หันมาจริงจังกับการทำตลาดผ่านแพลตฟอร์มและโซเชียลมีเดียเพื่อหวังแย่งชิงผู้อ่านชาวออนไลน์ข่าว แต่ก็เป็ข้อได้เปรียบของสื่อเดิมเพราะมีทีมผลิตบทความข่าว และมีคอนเทนต์ (Content) อยู่ในมือมหาศาล แต่หากมีการเก็บข้อมูลที่ไม่ดี ก็จะทำให้เกิด “สภาวะข้อมูลท่วมท้น” (Information overload) ซึ่งเกิดจากการขาดการจัดการข้อมูลที่ดี ทำให้มีข้อมูลที่ไม่เป็นสาระ ไม่สามารถนำไปใช้งานได้จริงและซ้ำซ้อน ภาวะดังกล่าวหากเกิดในแพลตฟอร์มที่เป็นสื่อก็อาจก่อให้เกิดปัญหาข่าวที่นำเสนอขึ้นไม่มีผู้อ่านชาวออนไลน์ ระบบสืบค้นและระบบแนะนำ (Recommendation System) จึงเป็นวิธีที่พัฒนาขึ้นเพื่อแก้ไขปัญหาภาวะข้อมูลท่วมท้นได้อย่างดี ทั้งนี้ระบบแนะนำเป็น

เครื่องมือ ซอฟต์แวร์ และเทคนิคการให้บริการข้อเสนอแนะสำหรับผู้ใช้งาน โดยเป็นข้อเสนอแนะที่เกี่ยวข้องกับการตัดสินใจต่างๆ (เรชา โสมพงษ์ และคณะ, 2015) เมื่อมีการเปรียบเทียบการใช้งานระหว่างระบบสืบค้นและระบบแนะนำข้อมูลแล้วพบว่า ผู้อ่านข่าวออนไลน์โดยทั่วไปชอบระบบแนะนำข้อมูลมากกว่า เพราะสามารถวิเคราะห์และเข้าถึงข้อมูลที่ตรงกับลักษณะการใช้งานจริงของผู้อ่านข่าวออนไลน์ โดยการเก็บข้อมูลจากพฤติกรรมของผู้ใช้งานอย่างอัตโนมัติและไม่ก่อให้เกิดปัญหาข้อมูลการใช้งานของผู้อ่านข่าวออนไลน์รั่วไหล ด้วยเหตุนี้เองระบบแนะนำข้อมูลจึงสามารถตอบสนองความต้องการของผู้ใช้เป็นรายบุคคลและรายครั้งตามความต้องการของผู้ใช้งานที่เปลี่ยนไป จึงเรียกได้ว่าระบบแนะนำส่วนบุคคล ที่มีผู้ใช้เป็นศูนย์กลาง (user centric) อย่างแท้จริง (Chen Li และคณะ, 2016)

ระบบแนะนำ (Recommendation System) สามารถแบ่งออกได้เป็น 3 แบบด้วยกันคือ 1. การกรองด้วยเนื้อหา (Content) คือการแนะนำที่วิเคราะห์พื้นฐานจากความชอบของผู้ใช้นั้น ๆ ในอดีตที่ผ่านมา เช่น แพลตฟอร์มแนะนำบทความที่มีเนื้อหาหรืออยู่ในกลุ่มประเภทบทความเดียวกันให้กับผู้อ่านข่าวออนไลน์ เป็นต้น 2. การกรองแบบร่วมมือ (Collaborative) จะวิเคราะห์ความชอบของผู้ใช้อื่นๆ ที่มีลักษณะ พฤติกรรม ความชอบที่คล้ายคลึงกันมาแนะนำให้กับผู้ใช้ปัจจุบัน เช่น เมื่อเข้าแพลตฟอร์มจำหน่ายสินค้าออนไลน์ หลังจากกดเลือกซื้อสินค้า A จะมีแถบโฆษณาและข้อความสินค้าตัวอื่นปรากฏขึ้น และแนะนำสินค้า B ด้วย ซึ่งเคยมีผู้ซื้อสินค้า A เลือกซื้อมาก่อน เป็นต้น 3. การกรองข้อมูลแบบผสม (Hybrid Approaches) เป็นการผสมผสานการแนะนำทั้งแบบการกรองด้วยเนื้อหาและการกรองแบบร่วมมือเข้าด้วยกัน โดยระบบที่ใช้การกรองด้วยเนื้อหานั้นได้เริ่มนำมาใช้กับระบบแนะนำข่าว เพื่อให้การแนะนำข่าวเหมาะกับผู้ใช้รายบุคคลมากขึ้น โดยระบบจะวิเคราะห์บทความข่าวว่ามีความสัมพันธ์กันหรือไม่ จากนั้นจึงเลือกแนะนำข่าวที่มีเนื้อหาใกล้เคียงกันกับข่าวก่อนหน้าให้กับผู้อ่านข่าวออนไลน์ขณะนั้น ในงานวิจัยฉบับนี้มุ่งเน้นที่จะพัฒนาโมเดลที่จะช่วยแนะนำข่าวที่เหมาะสมกับประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ โดยใช้วิธีการกรองข้อมูลแบบผสม (Hybrid Approaches) ร่วมกับวิธีการจำแนกประเภท (Classification) เพื่อหาโมเดลที่ดีที่สุดที่จะช่วยให้ระบบการแนะนำข่าวสามารถแนะนำข่าวให้ตรงกับความต้องการของผู้อ่านข่าวออนไลน์และทำให้อัตราการคลิกอ่านข่าวต่อไปสูงขึ้น และมีส่วนช่วยในการนำข้อมูลที่ได้นำไปพัฒนาการนำเสนอบทความข่าวสารบนแพลตฟอร์มที่เหมาะสมกับพฤติกรรมผู้อ่านข่าวออนไลน์

## 1.2 คำถามวิจัย

“ระบบแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าวออนไลน์ ช่วยให้  
อัตราการคลิกอ่านข่าวสูงขึ้น”

### 1.3 วัตถุประสงค์

1.3.1 เพื่อพัฒนาระบบแนะนำข่าวที่เหมาะสมกับประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์

1.3.2 เพื่อพัฒนาการนำเสนอข่าวสารบนแพลตฟอร์มเพื่อให้อัตราการคลิกอ่านข่าวที่แนะนำสูงขึ้น

### 1.4 นิยามศัพท์

- **การทำเหมืองข้อมูล (Data Mining)** หมายถึง กระบวนการวิเคราะห์ข้อมูลอย่างอัตโนมัติหรือกึ่งอัตโนมัติ เพื่อแยกประเภท ค้นหารูปแบบและแนวทางความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ หรือคลังข้อมูล ให้อยู่ในรูปแบบของกฎ (Rule) โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่องมือและหลักคณิตศาสตร์ (เรขา โสมพงษ์ และคณะ, 2015)

- **ระบบแนะนำ (Recommendation System)** เป็นวิธีที่พัฒนาขึ้นเพื่อแก้ไขปัญหาภาวะข้อมูลท่วมท้น ซึ่งเป็นเครื่องมือ ซอฟต์แวร์ และเทคนิคการให้บริการข้อเสนอแนะสำหรับผู้ใช้งาน โดยเป็นข้อเสนอแนะที่เกี่ยวข้องกับการตัดสินใจต่างๆ (เรขา โสมพงษ์ และคณะ, 2015)

- **หัวข้อข่าว** หมายถึง หมวดหมู่ของบทความข่าวที่แสดงบนในแพลตฟอร์ม

- **แพลตฟอร์ม** หมายถึง รูปแบบการแสดงผลเนื้อหาข่าวออนไลน์บนหลายอุปกรณ์ เช่น พีซี , แท็บเล็ต , หรือโทรศัพท์มือถือ และอุปกรณ์เคลื่อนที่อื่นๆ ซึ่งการนำเสนอแบบ 4 บทความ และ 8 บทความ เป็นรูปแบบการแสดงผลบนพีซี แต่บนอุปกรณ์อื่นๆจะถูกปรับแต่งตามความเหมาะสม

- **Session** หมายถึง ช่วงเวลาช่วงหนึ่งที่ผู้อ่านข่าวออนไลน์ 1 คน เข้ามาทำกิจกรรมต่างๆ บนแพลตฟอร์ม เช่น การอ่านเนื้อหา การคลิก เป็นต้น

- **หัวข้อข่าว 1** หมายถึง บทความข่าวแรกที่นำผู้อ่านข่าวออนไลน์เข้ามายังแพลตฟอร์มข่าว

- **หัวข้อความ 2** หมายถึง บทความข่าวต่อจาก หัวข้อข่าว 1 ที่ผู้อ่านข่าวออนไลน์จะคลิกเข้าไปอ่านข่าวต่อไป

- **Impressions** หมายถึง จำนวนครั้งที่มีการมองเห็นบทความข่าวที่ระบบแนะนำข่าวในแพลตฟอร์มแสดงบทความข่าวให้กับผู้อ่านข่าวออนไลน์

- **Click Through Rate** หมายถึง อัตราส่วนที่บ่งบอกว่ามีผู้อ่านบทความข่าวที่ระบบแนะนำข่าวแสดงให้เห็นบ่อยเพียงใด ซึ่งวัดจากจำนวนการคลิกหารด้วยจำนวนครั้งที่มีการมองเห็นบทความข่าวทั้งหมด

## 1.5 วิธีดำเนินการวิจัย

### 1.5.1 การเก็บรวบรวมข้อมูลและได้มาซึ่งข้อมูล (Data Extraction)

งานวิจัยนี้ผู้วิจัยได้ทำการเก็บรวบรวมข้อมูลผู้อ่านข่าวออนไลน์ข่าวที่เข้ามาผ่านแพลตฟอร์มและมีการลงทะเบียนด้วยการ Login โดยเก็บข้อมูลข้อมูลประวัติและพฤติกรรมของผู้ลงทะเบียน เป็นเวลาหนึ่งเดือน ในช่วงเดือนสิงหาคม พ.ศ. 2561 แล้วเป็นจำนวน 94,979 แถว

### 1.5.2 กระบวนการประมวลข้อมูลเบื้องต้น (Data preprocessing) ประกอบด้วย

1) การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนในการตรวจสอบและแก้ไขความถูกต้องของข้อมูลโดยผู้วิจัยตรวจสอบและแก้ไขความถูกต้องของข้อมูล เพื่อคัดข้อมูลส่วนรบกวนและส่วนที่ไม่เกี่ยวข้องออกไป

2) การรวมข้อมูล (Data Integration) หลังจากทำความสะอาดข้อมูลในเบื้องต้นแล้วผู้วิจัยทำการรวมข้อมูลประวัติผู้อ่านข่าวออนไลน์ข่าวและพฤติกรรมของผู้อ่านข่าวออนไลน์เข้าด้วยกัน

3) การคัดเลือกตัวแปรที่เหมาะสม (Data Selection) เป็นขั้นตอนการดึงข้อมูลจากแหล่งวิเคราะห์ที่บันทึกไว้ รวมทั้งตัดข้อมูลที่ไม่น่าสนใจออกไป ผู้วิจัยทำการกรองข้อมูลอีกครั้ง โดยคัดเอาเฉพาะตัวแปรที่มีประโยชน์ต่อการสร้างระบบ

4) การแปลงข้อมูลให้เหมาะสม (Data Transformation) ผู้วิจัยทำการแปลงข้อมูลที่ได้หลังจากคัดเลือกตัวแปรเพื่อให้สามารถนำไปวิเคราะห์ข้อมูลและสร้างระบบต่อไปได้

1.5.3 การพัฒนาระบบ (Data Modeling) ระบบประกอบด้วยโมเดลพยากรณ์หัวข้อข่าวและฟังก์ชันการเลือกข่าว นำการประมวลผลข้อมูล 6 แบบ ไปสร้าง 4 โมเดล และสร้างฟังก์ชันเพื่อ

เลือกข่าว จากบทความข่าวทั้งหมด 61,455 บทความ โดยนำผลลัพธ์ของโมเดลพยากรณ์หัวข้อข่าวที่ดีที่สุด มาเข้าสู่กระบวนการฟังก์ชันการเลือกข่าว

1.5.4 การวัดประสิทธิภาพของโมเดล (Model Evaluation) เป็นขั้นตอนการประเมินประสิทธิภาพการทำงานของโมเดลการจำแนกประเภท โดยผู้วิจัยใช้วิธี Cross-Validation, ตารางการจำแนกหรือเมทริกซ์ความสับสน (Confusion Matrix) และการวัดค่าความถูกต้องแม่นยำของโมเดล (Accuracy)

1.5.5 การนำระบบไปใช้จริง (Implementation) ผู้วิจัยทดสอบประสิทธิภาพของระบบด้วยการนำไปใช้งานจริงบนแพลตฟอร์มไทยรัฐ ใน 5 หัวข้อข่าว ดังนี้ การเมือง, ต่างประเทศ, อาชญากรรม, สังคม และบันเทิง ซึ่งเป็น 5 หัวข้อข่าวที่อัตราการเข้าชมสูงที่สุด โดยทำการวัดผลด้วยอัตราการคลิก

## 1.6 ขอบเขตการวิจัย

1.6.1 การทดลองนี้ใช้ข้อมูลของผู้อ่านข่าวออนไลน์ข่าวที่เข้ามาผ่านแพลตฟอร์มไทยรัฐ และมีการลงทะเบียนด้วยการ Login ซึ่งเป็นแพลตฟอร์มข่าวที่อนุญาตให้สามารถเข้าถึงข้อมูลในการทำงานวิจัย โดยเก็บข้อมูลเป็นเวลาหนึ่งเดือน ในช่วงเดือนสิงหาคม พ.ศ. 2561 ซึ่งการเก็บข้อมูลในช่วงเวลาหนึ่งๆ ที่สถานการณ์บ้านเมืองเป็นปกติ เมื่อผ่านขั้นตอนการทำความสะอาดข้อมูลแล้วเป็นจำนวน 91,243 แถว ประกอบด้วย

1.6.1.1 ข้อมูลประวัติผู้ลงทะเบียน

1.6.1.2 ข้อมูลพฤติกรรมผู้ลงทะเบียน

1.6.2 ไม่เก็บพฤติกรรมของผู้อ่านข่าวออนไลน์ย้อนหลังมาใช้เป็นข้อมูลการสร้างระบบ

1.6.3 การนำระบบมาทดสอบบนแพลตฟอร์มเพื่อเปรียบเทียบผลก่อนและหลังการใช้ระบบจะทำได้ในช่วงเวลาที่จำกัด เพียง 24 ชั่วโมงเท่านั้น

1.6.4 งานวิจัยนี้จะทำการเปรียบเทียบการทำงานระหว่างแพลตฟอร์มที่ใช้ระบบแนะนำข่าวในเชิงพาณิชย์ “C” กับระบบแนะนำข่าวของผู้วิจัย

## 1.7 ผลที่คาดว่าจะได้รับ

1.7.1 ได้ระบบการแนะนำข่าวที่เหมาะสมกับประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์

1.7.2 เพิ่มอัตราการคลิกอ่านข่าวของแพลตฟอร์มให้มากขึ้น

## บทที่ 2

### แนวคิดทฤษฎีและงานวิจัยที่เกี่ยวข้อง

การศึกษาเอกสารและทฤษฎีที่เกี่ยวข้องในบทนี้ผู้วิจัยได้ศึกษาเอกสาร แนวคิด และทฤษฎี การทำเหมืองข้อมูล โดยแยกออกเป็นหัวข้อดังนี้

#### 1. การทำเหมืองข้อมูล

- 1.1 การเก็บรวบรวมข้อมูล
- 1.2 การประมวลข้อมูลเบื้องต้น
- 1.3 การสร้างโมเดล
- 1.4 การวัดประสิทธิภาพของโมเดล
- 1.5 การนำโมเดลไปใช้จริง

#### 2. การจำแนกประเภท (Classification)

##### 2.1 เทคนิคการจำแนกประเภท

- 2.1.1 ต้นไม้ตัดสินใจ (Decision Tree)
- 2.1.2 การถดถอยเชิงโลจิสติก (Logistic Regression)
- 2.1.3 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine-SVM)
- 2.1.4 โครงข่ายประสาทเทียม (Artificial Neural Network-ANN)

##### 2.2 การทดสอบโมเดล

##### 2.3 การวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล (Classifier

Evaluation Metrics)

3. ระบบแนะนำข่าว
4. การทดสอบประสิทธิภาพของระบบด้วย A/B testing
5. งานวิจัยที่เกี่ยวข้อง

#### 1. การทำเหมืองข้อมูล (Data Mining)

การทำเหมืองข้อมูล (Data Mining) หมายถึง กระบวนการวิเคราะห์ข้อมูลอย่างอัตโนมัติหรือกึ่งอัตโนมัติ เพื่อแยกประเภท ค้นหารูปแบบและแนวทางความสัมพันธ์ของข้อมูลจากฐานข้อมูลขนาดใหญ่ หรือคลังข้อมูล ให้อยู่ในรูปแบบของกฎ (Rule) โดยอาศัยหลักสถิติ การรู้จำ การเรียนรู้ของเครื่องมือและหลักคณิตศาสตร์ (เรขา โสมพงษ์ และคณะ, 2015)

ผลลัพธ์จากการทำเหมืองข้อมูล คือ รูปแบบและแนวทางความสัมพันธ์ที่ซ่อนอยู่ในชุดข้อมูลหนึ่งๆ ซึ่งรูปแบบจะสะท้อนเหตุการณ์หรือสิ่งที่เกิดซ้ำแล้วซ้ำอีก จนสามารถทำนายได้ (ญาใจ ลิ้มปิยะกรณ, 2556, 14) ซึ่งขั้นตอนในการทำเหมืองข้อมูลมีดังนี้

### 1.1 การเก็บรวบรวมข้อมูล

เป็นขั้นตอนก่อนการทำเหมืองข้อมูล เนื่องจากข้อมูลที่ได้มายังจำเป็นต้องตรวจสอบความถูกต้อง ซึ่งอาจมีลักษณะข้อมูล ดังนี้

1.1.1 ข้อมูลไม่สมบูรณ์ (Incomplete data) คุณลักษณะบางอย่างของข้อมูลขาดหายไป (Missing value) ขาดคุณลักษณะที่น่าสนใจ หรือขาดรายละเอียดของข้อมูล เช่น พฤติกรรมการคลิกอ่านข่าวที่ไม่มีความต่อเนื่องภายในเวลาที่กำหนด คือ 1 Session (30 นาที) ทำให้ไม่สามารถนำข้อมูลนั้นมาสังกัดเพื่อทำ การทดลองได้ เป็นต้น

1.1.2 ข้อมูลรบกวน (Noisy data) ข้อมูลนั้นมีค่าผิดพลาด (Error) หรือมีค่าผิดปกติ (Outliers) เช่น ข้อมูลวัน เดือน ปีเกิด ที่คำนวณออกมาแล้วมีค่าเกิน 100 หรือมีค่าต่ำกว่า 0 เป็นต้น

1.1.3 ข้อมูลไม่สอดคล้อง (Inconsistent data) เป็นข้อมูลเดียวกัน แต่ตั้งชื่อต่างกัน หรือใช้ค่าแทนข้อมูลต่างกัน เช่น การใส่ชื่อหมวดข่าวที่มีความซ้ำซ้อนกัน Entertainment และ /Entertainment ซึ่งเป็นหมวดเดียวกันแต่กระบวนการดึงข้อมูลทำให้ข้อมูลมีความซ้ำซ้อนกัน เป็นต้น (ผศ.วิภาวรรณ บัวทอง, 2557, 2)

### 1.2 การประมวลข้อมูลเบื้องต้น

1.2.1 การทำความสะอาดข้อมูล (Data Cleaning) เป็นขั้นตอนในการตรวจสอบและแก้ไขความถูกต้องของข้อมูล เพื่อคัดข้อมูลส่วนรบกวนการทำความสะอาดข้อมูล และส่วนที่ไม่เกี่ยวข้องออกไป เช่น browser id เป็นข้อมูลที่ไม่มีความจำเป็นในการสร้างโมเดล จึงตัดออกไม่นำมาใช้ในการทดลอง เป็นต้น

1.2.2 การรวมข้อมูล (Data Integration) เป็นขั้นตอนการรวมแหล่งข้อมูล ซึ่งมีข้อมูลหลายแห่งมารวมไว้ที่เดียวกัน เช่น การนำข้อมูลประวัติผู้อ่านข่าวออนไลน์มารวมกับข้อมูลของพฤติกรรมผู้อ่านข่าวออนไลน์ เป็นต้น

1.2.3 การคัดเลือกข้อมูล (Data Selection) เป็นขั้นตอนการคัดเลือกตัวแปรที่เหมาะสม รวมทั้งตัดข้อมูลที่จำเป็นออกไป เช่น การตัดข้อมูลประวัติผู้อ่านข่าวออนไลน์ที่มีอายุต่ำกว่า 15 ปี และสูงกว่า 80 ปีออกไป เพราะมีความเป็นไปได้ว่ากลุ่มของอายุที่ตัดออกนั้นจะไม่ได้เป็นผู้ลงทะเบียนด้วยการ Login ด้วยตัวเอง เป็นต้น



1.2.4 การแปลงข้อมูล (Data Transformation) เป็นขั้นตอนการแปลงข้อมูลให้เหมาะสม สำหรับนำไปวิเคราะห์ข้อมูลและสร้างโมเดลต่อไป เช่น การแปลงวัน เวลา การเข้ามาอ่านข่าว ให้เป็นข้อมูลรายสัปดาห์ และรายชั่วโมง เป็นต้น

### 1.3 การสร้างโมเดล

เป็นการสร้างโมเดลจากข้อมูลที่มี โดยการสร้างโมเดล คือการให้อัลกอริทึมทำการค้นหารูปแบบที่เป็นประโยชน์จากชุดข้อมูลที่ใส่เข้าไป เช่น การสร้างโมเดลจากอัลกอริทึมซัพพอร์ทเวกเตอร์แมชชีน (SVM) เพื่อพยากรณ์พฤติกรรมผู้อ่านข่าวออนไลน์ว่าจะมีการคลิกอ่านข่าวที่ระบบแนะนำให้หรือไม่ เป็นต้น

### 1.4 การวัดประสิทธิภาพของโมเดล

เป็นขั้นตอนการประเมินประสิทธิภาพการ ทำงาน ของ โมเดล โดยค่าที่ใช้วัดประสิทธิภาพของโมเดล เช่น อัตราความถูกต้อง (Accuracy) ซึ่งผลลัพธ์ที่ได้จะออกมาเป็นเปอร์เซ็นต์

### 1.5 การนำโมเดลไปใช้จริง

เป็นขั้นตอนการนำเสนอความรู้ที่ค้นพบ โดยใช้เทคนิคในการนำเสนอเพื่อให้ เข้าใจ และนำโมเดลไปใช้จริง เช่น การนำโมเดลที่ให้ผลลัพธ์ที่ดีที่ได้ไปทำนายพฤติกรรมการเข้าอ่านข่าวของผู้อ่านข่าวออนไลน์ใหม่ที่ไม่เคยเข้าแพลตฟอร์มมาก่อน และวัดผลเปรียบเทียบก่อน-หลังการนำโมเดลไปใช้งาน เป็นต้น (Han, J. & Kamber, M., 2006 : 7)

## 2. การจำแนกประเภท (Classification)

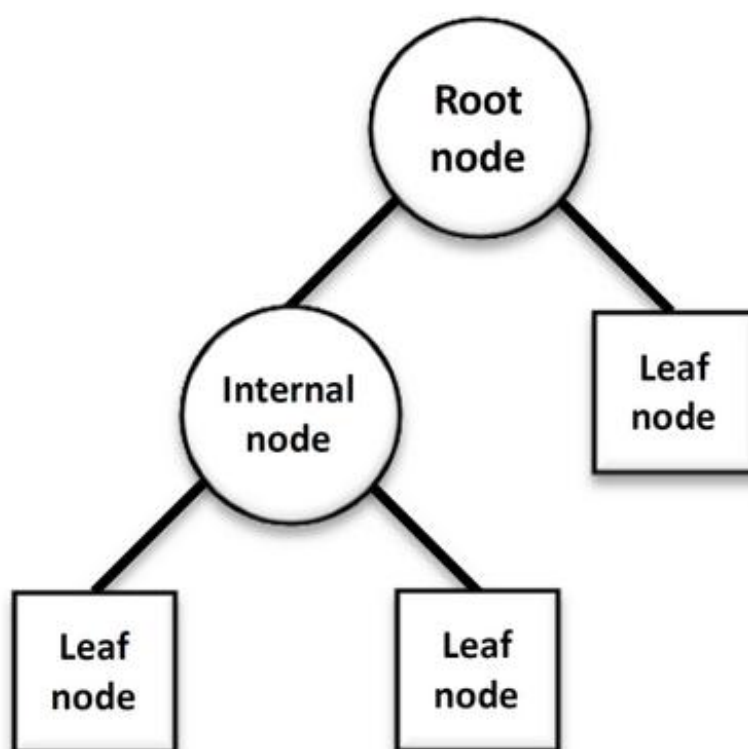
การจำแนกประเภท เป็นกระบวนการจำแนกข้อมูลให้เป็นหมวดหมู่เพื่อให้ได้ผลลัพธ์ที่เหมาะสมที่สุดต่อผู้ใช้และสามารถใช้ได้อย่างมีประสิทธิภาพ ซึ่งการจำแนกข้อมูลที่ดีจะทำให้ข้อมูลสำคัญสามารถค้นหาและเรียกค้นข้อมูลได้ง่าย (Margaret Rouse , 2014, Paragraph 1) การจำแนกประเภทข้อมูลประกอบด้วยการทำนายผลบางอย่างขึ้นอยู่กับข้อมูลที่กำหนดหรือที่เรียกว่า Predictive Modeling เป็นเทคนิคที่นิยมใช้กันมากในการวิเคราะห์ข้อมูลและการทำงานวิจัยเชิงประยุกต์ ซึ่งกระบวนการจำแนกประเภทข้อมูลจะแบ่งเป็นสองส่วนคือ (1) การนำข้อมูลสอน (training

data) มาสร้างโมเดลและวัดประสิทธิภาพของโมเดล และ (2) การนำโมเดลที่ได้ไปใช้ทำนาย (predict) เพื่อหาคำตอบให้กับข้อมูลใหม่ (เอกสิทธิ์ พัทธวงศ์ศักดิ์, 2557:50)

## 2.2 เทคนิคการจำแนกประเภท

### 2.1.1 ต้นไม้ตัดสินใจ (Decision Tree)

ต้นไม้ตัดสินใจ เป็นการจำแนกประเภท ข้อมูลออกเป็นกลุ่ม (class) ต่างๆ โดยใช้คุณลักษณะ (attribute) ข้อมูลในการจำแนกประเภท ต้นไม้ตัดสินใจที่ได้จากการเรียนรู้ ทำให้ทราบว่าคุณลักษณะใดของข้อมูลที่เป็นตัวกำหนดการจำแนกประเภท และคุณลักษณะแต่ละตัวมีความสำคัญมากน้อยต่างกันอย่างไรต่อการจำแนกประเภท ช่วยให้สามารถวิเคราะห์ข้อมูลและตัดสินใจได้ถูกต้องมากยิ่งขึ้น (ญาใจ ลิ้มปิยะกรณ, 2556:120) ดังภาพประกอบที่ 2.1



ภาพประกอบที่ 2.1 ส่วนประกอบของต้นไม้ตัดสินใจ

ส่วนประกอบของต้นไม้ตัดสินใจ

1) โหนดภายใน (Internal node) คือ คุณลักษณะต่างๆ ของข้อมูล ซึ่งเมื่อข้อมูลใดๆ ตกลงมาถึงโหนด จะใช้คุณลักษณะนี้เป็นตัวตัดสินใจว่าข้อมูลจะไปทิศทางใด (เงื่อนไข) โดยโหนดภายในที่เป็นจุดเริ่มต้นของต้นไม้ เรียกว่า โหนดราก (Root node)

2) กิ่ง (branch, link) เป็นค่าของคุณลักษณะในโหนดภายในที่แตกกิ่งนี้ออกมา ซึ่งโหนดภายในจะแตกกิ่งเป็นจำนวนเท่ากับจำนวนของคุณลักษณะในโหนดภายในนั้น

3) โหนดใบ (Leaf node) คือกลุ่มต่างๆ ซึ่งเป็นผลลัพธ์ในการจำแนกประเภทข้อมูล (การกระทำ)

### ข้อดีของต้นไม้ตัดสินใจ

- 1) เหมาะสมกับข้อมูลที่มีตัวแปรเชิงคุณภาพ หรือ ตัวแปรไม่ต่อเนื่อง
- 2) มีความทนทานต่อข้อมูลรบกวน เช่น คุณลักษณะที่ไม่เกี่ยวข้อง และค่าคุณลักษณะที่ขาดหาย
- 3) การเรียนรู้มีความรวดเร็วเมื่อเทียบกับอัลกอริทึมสำหรับการจำแนกประเภทชนิดอื่น
- 4) แต่ละเส้นทาง จากโหนดรากถึงโหนดใบสามารถแสดงความรู้ในรูปกฎ IF-THEN ได้
- 5) ผลการเรียนรู้แสดงอยู่ในรูปที่เข้าใจได้ง่ายทำให้ง่ายต่อการวิเคราะห์คุณลักษณะที่มีผลต่อการจำแนกประเภทกลุ่มต่างๆ (ญาใจ ลิมปิยะกรณ, 2556:121)

### การคัดเลือกตัวแปร (Feature selection)

การสร้างโมเดลด้วยวิธีต้นไม้ตัดสินใจ จะทำการคัดเลือกคุณลักษณะ (Feature) ที่มีความสัมพันธ์กับกลุ่มข้อมูล (class) มากที่สุดขึ้นมาเป็นโหนดบนสุด (root node) ของต้นไม้ หลังจากนั้นจะหาคุณลักษณะของข้อมูลต่อไปเรื่อยๆ (เอกสิทธิ์ พัทธวงศ์ศักดิ์, 2557: 59)

ในการจำแนกประเภทข้อมูล จะพบว่าคุณลักษณะของข้อมูลมีจำนวนมาก ซึ่งบางข้อมูลก็ไม่ได้มีความสำคัญในการแยกกลุ่มข้อมูล (class) ตามที่ต้องการ จึงจำเป็นต้องการคัดเลือกคุณลักษณะข้อมูลมาใช้งาน ซึ่งขั้นตอนการคัดเลือกจะแบ่งออกเป็น 2 กลุ่มใหญ่คือ

**1. Filter Approach** เป็นการคัดเลือกตัวแปรโดยการคำนวณหาค่าน้ำหนักซึ่งอาจเป็นการหาความสัมพันธ์ในแต่ละตัวแปรและกลุ่มข้อมูล และจะเลือกตัวแปรโดยเรียงลำดับตามค่าน้ำหนักที่คำนวณได้แล้วเลือกตัวแปรที่มีค่าน้ำหนักมากกว่าตามที่ต้องการมาใช้งาน

**2. Wrapper Approach** เป็นการคัดเลือกตัวแปรด้วยการสร้างโมเดล จำแนกประเภทขึ้นมาจากชุดของตัวแปรที่กำหนดไว้และวัดประสิทธิภาพการทำงานของโมเดล และเลือกชุดของตัวแปร

แปรที่ทำให้โมเดลมีประสิทธิภาพมากที่สุดมาใช้งาน เช่น โมเดลที่ให้ค่าความถูกต้องมากที่สุด การคัดเลือกตัวแปรด้วยวิธีการนี้แบ่งย่อยได้เป็น 2 แบบ คือ

- Forward Selection เป็นการสร้างโมเดลโดยการเพิ่มตัวแปรข้อมูลที่ละหนึ่งตัวแปร ถ้าคุณลักษณะข้อมูลที่ใส่เพิ่มให้ประสิทธิภาพที่ดีก็จะเก็บไว้และเลือกตัวแปรอื่นๆ มาเพิ่มต่อไปจนประสิทธิภาพของโมเดลไม่ได้ดีขึ้นก็จะหยุดทำงาน

- Backward Elimination เป็นการสร้างโมเดลที่เริ่มจากการใช้ตัวแปรทั้งหมดก่อนและตัด (Eliminate) ตัวแปรที่ไม่สำคัญทิ้งไปทีละหนึ่งตัวแปร ถ้าประสิทธิภาพดีขึ้นก็ตัดตัวแปรอื่นๆ ต่อไป

### การสร้างต้นไม้ตัดสินใจ

ในช่วงปลายของยุค 1970 มีนักวิจัยด้านการเรียนรู้ของเครื่อง (Machine Learning) คือ J. Ross Quinlan ได้คิดค้นอัลกอริทึมสำหรับสร้างต้นไม้ตัดสินใจที่มีชื่อว่า ID3 (Iterative Dichotomiser) ต่อมาได้พัฒนาต่อยอด ID3 ไปเป็น C4.5 ซึ่งได้กลายมาเป็นอัลกอริทึมพื้นฐานที่ใช้สำหรับเปรียบเทียบประสิทธิภาพการทำงานของอัลกอริทึมต่างๆ ทางด้านการเรียนรู้แบบมีผู้สอน (Supervised Learning)

ID3 และ C4.5 ได้ทำการประยุกต์ใช้วิธีการเชิงละโมภ (Greedy approach) ในการสร้างต้นไม้ภายใต้วิธีการแบบ “Top-down recursive divide-and-conquer” โดยทำการพิจารณาชุดข้อมูลสอน (Training data) ด้วยการแบ่งข้อมูลออกเป็นส่วนย่อยๆ ในระหว่างกระบวนการสร้างต้นไม้ (โกเมศ: 6)

- เริ่มต้นด้วยนำตัวอย่างชุดข้อมูลสอน (Training data) มาสร้างเป็นราก (Root node)

- คุณลักษณะของข้อมูลควรอยู่ในรูปของข้อมูลเชิงคุณภาพ คือข้อมูลชนิดกลุ่ม หากเป็นข้อมูลในเชิงปริมาณ ควรทำการแบ่งข้อมูลให้เป็นกลุ่มก่อนเสียก่อน

- การสร้างต้นไม้ตัดสินใจนั้นมีพื้นฐานมาจากวิธีการเลือกคุณลักษณะของข้อมูล

- จะหยุดสร้างต้นไม้ตัดสินใจเมื่อชุดข้อมูลที่ตกอยู่ในโนหนดใบ (Leaf node) มีคลาสเดียวกันทั้งหมดหรือเกือบทั้งหมด (หทัยรัตน์, 2014: 18)

### 2.1.2 การถดถอยเชิงโลจิสติก (Logistic Regression Analysis)

การถดถอยเชิงโลจิสติก นำมาใช้วิเคราะห์ข้อมูลเพื่อทำนายว่าเหตุการณ์หนึ่งจะเกิดขึ้นได้หรือไม่ หรือมีโอกาสเกิดขึ้นมากน้อยเพียงใด โดยกำหนดว่ามีตัวแปรหนึ่งหรือหลายตัวที่ส่งผลต่อเหตุการณ์นั้นๆ ทำให้ทราบถึงเหตุผลของการเกิดเหตุการณ์นั้นหรือไม่เกิดเหตุการณ์นั้น การถดถอยเชิงโลจิสติก จะทำให้ทราบว่ามีสาเหตุใดที่บ่งชี้ความสำเร็จของเหตุการณ์ที่เกิดขึ้นเหล่านี้ ค่าของตัวแปรตามที่ปรากฏเป็นข้อมูลสำหรับวิเคราะห์จะมีเพียง 2 ค่าเท่านั้น กล่าวคือ ใช่ (Yes) หรือ ไม่ใช่ (No) ต่อไปจะแทนด้วยค่า 1 และ 0 ตามลำดับ ดังนั้นในการอนุมานทางสถิติจึงไม่ต้องมีข้อสมมติฐานการแจกแจงของค่าความคลาดเคลื่อนมีการแจกแจงแบบปกติ เนื่องจากค่าของตัวแปรตามที่ถูกวัดเป็นค่า 1 และ 0 ไม่ใช่ค่าต่อเนื่อง แต่ค่าในที่สุดค่าที่ถูกทำนายจากสมการ Logistic Regression จะให้ค่าที่อยู่ระหว่าง 0 และ 1 หมายถึงค่าของความน่าจะเป็นของการเกิดเหตุการณ์ที่น่าสนใจ (อุไรวรรณ, 2546: 26) ประเภทของการถดถอยเชิงโลจิสติกมี 2 ประเภทดังนี้

1) การถดถอยเชิงโลจิสติกทวิ (Binary Logistic) ความสัมพันธ์จะอยู่ในรูปแบบของสมการเส้นถดถอย (Regression Equation) มีตัวแปรที่มีค่าเพียง 2 ค่า (Dichotomous Variable) คือ เป็น 0 และ 1 ส่วนคุณลักษณะข้อมูลอาจมีค่าเพียงค่าเดียวหรือหลายตัวก็ได้ จากการที่ Logistic Regression Analysis เป็นการทำนายค่าความน่าจะเป็นของการเกิดหรือไม่เกิดเหตุการณ์ที่น่าสนใจ ซึ่งมีข้อมูลตัวแปรตาม ดังนั้นเทคนิคนี้อาจถูกเรียกว่า Binary Logistic Regression สมการแสดงความสัมพันธ์ระหว่าง  $x$  และ  $y$  จะอยู่ในรูปเชิงเส้น ดังนี้

$$y = \beta_0 + \beta_1 x + e \quad (1)$$

แต่สำหรับการวิเคราะห์ความถดถอยเชิงโลจิสติก ตัวแปรตามหรือ  $y$  มีค่าได้เพียง 2 ค่า คือ ไม่เกิดเหตุการณ์ ( $y = 0$ ) และเกิดเหตุการณ์ ( $y = 1$ ) ซึ่งความสัมพันธ์ไม่ได้อยู่ในรูปเชิงเส้น เนื่องจากตัวแปรตามมีค่า 0 กับ 1 ความสัมพันธ์จะอยู่ในรูปดังสมการนี้

$$P(Y) = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}} \quad (2)$$

เรียกสมการที่ (2) ว่า Logistic Response Function โดยที่

$$0 \leq P(Y) \leq 1$$

เมื่อ (Y) คือความน่าจะเป็นของการเกิดเหตุการณ์ Y และ

$e$  คือ exponential function

และความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจ ( $y = 0$ ) มีค่าเท่ากับ  $1 - P(Y)$

2) โมเดลการวิเคราะห์การถดถอยเชิงโลจิสติกแบบพหุกลุ่ม

วิธีนี้จะใช้เมื่อตัวแปรตามเป็นตัวแปรเชิงกลุ่มที่มีค่ามากกว่า 2 ตัวขึ้นไปซึ่งกรณีที่มีตัวแปรอิสระมากกว่า 1 ตัว หรือมีตัวแปรอิสระ  $n$  ตัว จะได้สมการความสัมพันธ์  $x$  และ  $y$  ดังนี้

$$P(Y) = \frac{e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_n x_n)}}{1 + e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_n x_n)}} \quad (3)$$

$$P(\text{ไม่เกิดเหตุการณ์}) = 1 - P(Y)$$

จากสมการความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามของการวิเคราะห์การถดถอยเชิงโลจิสติกไม่ได้อยู่ในรูปเชิงเส้น โดยปรับให้อยู่ในรูปของ Odds หรือ Odd Ratio ซึ่ง Odd Ratio หมายถึงอัตราส่วนระหว่างโอกาสที่จะเกิดเหตุการณ์ที่สนใจ ( $y = 1$ ) กับโอกาสที่ไม่เกิดเหตุการณ์ ดังนั้น Odds ของการเกิดเหตุการณ์ และแปลงให้เป็นเส้นตรงได้ดังสมการ (4) (5) และ (6) ดังนี้

$$\frac{P}{1-P} = e^{(\beta_0 + \beta_1 X_1 + \dots + \beta_n x_n)} \quad (4)$$

$$\log \log \left( \frac{P}{1-P} \right) = \beta_0 + \beta_1 x \quad (5)$$

$$\text{Log(odds of } P) = \text{logit}(P) = \beta_0 + \beta_1 x + \dots + \beta_n x_n + E \quad (6)$$

$P$  หมายถึง ความน่าจะเป็นที่เกิดเหตุการณ์

$\beta_0$  หมายถึง ค่าของ  $y$  เมื่อ  $x = 0$

$\beta_n$  หมายถึง อัตราการเปลี่ยนแปลงของ  $y$  เมื่อ  $x_n$  เปลี่ยนไป 1 หน่วย โดยที่ตัวแปรอิสระอื่นๆ คงที่

$\mathcal{E}$  หมายถึง ความคลาดเคลื่อนของการพยากรณ์

### 2.1.3 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine - SVM)

เป็นอัลกอริทึมในการคัดแยกกลุ่มเพื่อจัดประเภทหรือจำแนกประเภทแบบการเรียนรู้โดยอาศัยตัวอย่างประเภทหนึ่ง ซึ่งมีความสามารถในการจัดหมวดหมู่และการทำนาย โดยพื้นฐานจะมีการคำนวณแบบเชิงเส้น จัดอยู่ในประเภทมุ่งหาผลลัพธ์ที่ดีที่สุดในการเรียนรู้ (Discriminative Training) บนการเรียนรู้เชิงสถิติของข้อมูล ซึ่งในงานวิจัยนี้การนำข้อมูลประวัติและพฤติกรรมผู้อ่านข่าวออนไลน์มาใช้ในการหาขอบเขตการตัดสินใจในการแบ่งข้อมูลออกเป็นสองส่วน โดยใช้สมการเส้นตรงเพื่อแบ่งเขตข้อมูล 2 กลุ่มออกจากกัน โดยมีวัตถุประสงค์เพื่อลดความผิดพลาดจากการทำนายผลลัพธ์ (Minimize error) พร้อมกับเพิ่มระยะห่างระหว่างขอบเขตข้อมูลทั้ง 2 กลุ่ม ให้มากที่สุด (Maximize Margin) ของระนาบตัดสินใจ (Decision Hyperplane) ในการแยกกลุ่มของข้อมูลสอนออกจากกัน หรือเรียกว่าการจัดหมวดหมู่โดยค่าระยะขอบที่มากที่สุด (Maximize Margin Classifier)

กำหนดให้  $(x_i, y_i), \dots, (x_n, y_n)$  เมื่อ  $x \in R^m, y \in \{-1, 1\}$  โดย

$n$  คือ จำนวนข้อมูลตัวอย่าง

$m$  คือ จำนวนมิติของข้อมูลเข้า

$x$  คือ ข้อมูลนำเข้า

$y$  คือ ประเภทหรือกลุ่มของข้อมูล ประกอบด้วย 2 กลุ่ม มีค่า +1 หรือ -1

สำหรับปัญหาเชิงเส้น ข้อมูลมิติขนาดสูง ถูกแบ่งเป็น 2 กลุ่ม โดยใช้ระนาบตัดสินใจพิจารณาชุดของกลุ่มข้อมูล  $x$  โดยกำหนดให้กลุ่มข้อมูล  $x_1$  เป็นเวกเตอร์ข้อมูลที่มีค่า  $y$  เป็นบวก และ  $x_2$  เป็นเวกเตอร์ข้อมูลที่มีค่า  $y$  เป็นลบ การสร้างระนาบตัดสินใจเพื่อแบ่งแยกกลุ่มข้อมูลสามารถคำนวณได้ดังสมการที่ 2

$$(w * x_1) + b > 0 \text{ ถ้า } y_i = +1 \text{ และ } (w * x_2) + b < 0 \text{ ถ้า } y_i = -1 \quad (2)$$

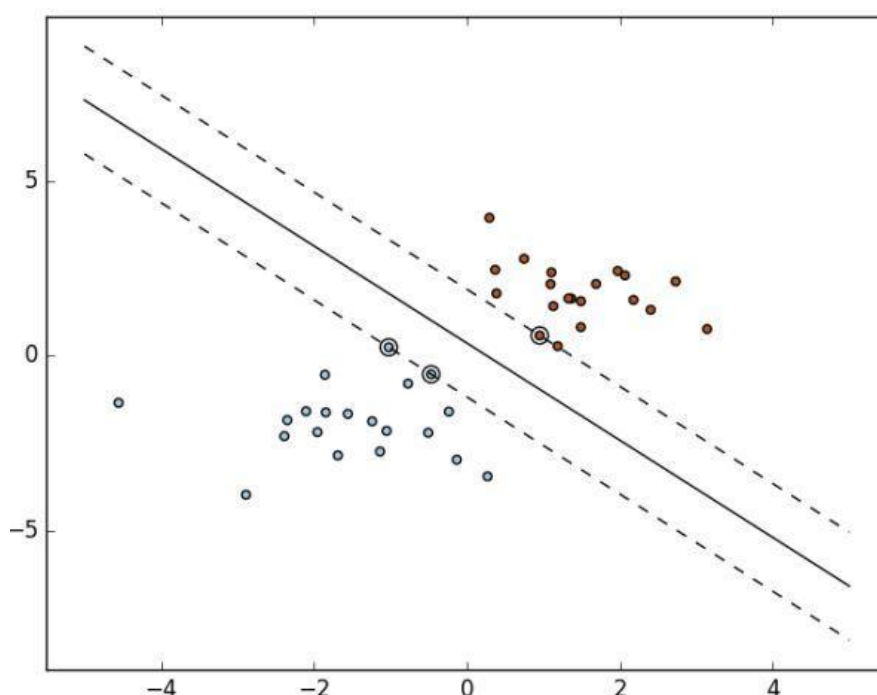
$w$  คือ เวกเตอร์น้ำหนัก

$x_1$  คือ เวกเตอร์ข้อมูลที่มีค่าเป็นบวก

$x_2$  คือ เวกเตอร์ข้อมูลที่มีค่าเป็นลบ

$b$  คือ ค่า bias

ในการหาระนาบเกินที่เหมาะสมที่สุด จะทำการหาตำแหน่งของซัพพอร์ตเวกเตอร์แมชชีน เพื่อเป็นตัวแทนของกลุ่มข้อมูลทั้งคู่ ในการพิจารณาเกณฑ์การแบ่งกลุ่มโดยอาศัยหลักการคือ จะใช้ ระนาบเกินที่เป็นระยะห่างที่สุทธระหว่างข้อมูล 2 กลุ่ม ที่อยู่ใกล้กันมากที่สุดเพียงระนาบเดียวเท่านั้น จากนั้นจึงหาระนาบที่รักษาระยะห่างจากขอบมากที่สุด (Maximum Margin) และถือว่าระนาบดังกล่าวคือระนาบสำหรับการแบ่งกลุ่มที่เหมาะสมที่สุด ดังภาพประกอบ 2.2



ภาพประกอบ 2.2 แสดงระนาบสำหรับการแบ่งกลุ่มที่เหมาะสมที่สุด ของ SVM (Lasse Schultebrucks, 2017)

#### 2.1.4 โครงข่ายประสาทเทียม (Artificial Neural Network - ANN)



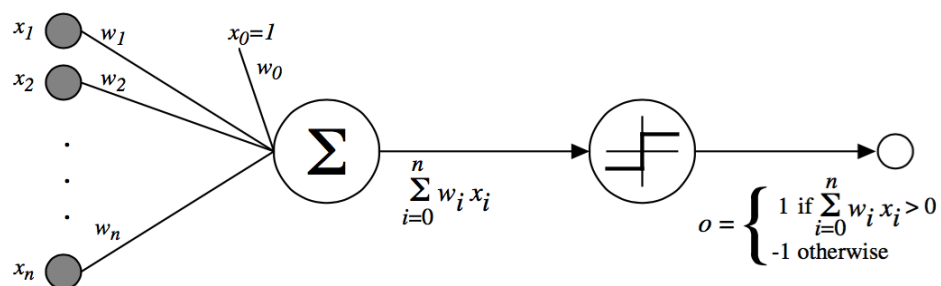
โครงข่ายประสาทเทียม หรือในชื่ออื่น เช่น ข่ายงานประสาทเทียม และนิวรัลเน็ตเวิร์ค (Neural Network) เป็นเทคนิคการเรียนรู้หนึ่งในศาสตร์การเรียนรู้ของเครื่อง (Machine Learning) ซึ่งมีที่มาจากการทำงานของเซลล์ประสาทในสมองของมนุษย์ ทั้งนี้ โครงข่ายประสาทเทียมใช้วิธีการเรียนรู้เชิงแนะนำ (Supervised Learning) เนื่องจากชุดข้อมูลสอนโครงข่ายประสาทเทียมจะเป็นชุดข้อมูลที่ทราบคำตอบ (Class Label) ล่วงหน้าแล้วนั่นเอง

#### 2.1.4.1 เพอร์เซปตรอน (Perceptron)

เพอร์เซปตรอน เป็นหน่วยย่อยที่สุดของโครงข่ายประสาทเทียม โดยจะรับข้อมูลขาเข้าเป็นตัวเลขจำนวน  $n$  ค่า ตั้งแต่  $x_1, \dots, x_n$  จากนั้นจะผ่านขั้นตอน ดังนี้

- 1) นำข้อมูลขาเข้าแต่ละค่าไปคูณถ่วงน้ำหนักด้วยค่าน้ำหนักของตัวเองซึ่งแทนด้วย  $w_1, \dots, w_n$
- 2) นำค่าจากข้อ 1) แต่ละค่าไปบวกรวมกันพร้อมกับบวกด้วยค่าน้ำหนักพิเศษ  $w_0$  หรือในตำราบางเล่มอาจเรียกว่าค่า  $b$  ซึ่งมาจากคำว่า ไบแอส (Bias) ทั้งนี้ ค่าน้ำหนักพิเศษอาจมองเป็นข้อมูลขาเข้าตัวที่ 0 ซึ่งกำหนดตายตัวว่ามีค่าเป็น 1 เสมอก็ได้ กล่าวคือ  $x_0 = 1$
- 3) ค่าผลรวมจากข้อ 2) ไปเข้าฟังก์ชันกระตุ้น (Activation function) ซึ่งมีลักษณะเป็นฟังก์ชันขั้นบันได (Step function) ให้ผลลัพธ์เป็นค่าใดค่าหนึ่งระหว่าง 1 กับ -1

ซึ่งผลลัพธ์จากข้อ 3) จะเป็นผลลัพธ์ของเพอร์เซปตรอนด้วย สามารถสรุปได้ในรูปแบบสมการ ดังนี้



ภาพประกอบ 2.3 โครงสร้างเพอร์เซปตรอน

ซึ่งฟังก์ชันขั้นบันไดที่ให้ผลลัพธ์เป็น 1 หรือ -1 นิยมเรียกว่าฟังก์ชันสองขั้ว (Bipolar function) แต่ฟังก์ชันขั้นบันไดที่ให้ผลลัพธ์ระหว่าง 1 กับ 0 จะนิยมเรียกว่า ฟังก์ชันสองคำตอบ (Binary function)

การเรียนรู้ของเพอร์เซปตรอน จะมีกฎการสอนตามสมการ ดังนี้

$$w_i \leftarrow w_i + w_i \quad (1)$$

$$w_i \leftarrow (t - )x_i \quad (2)$$

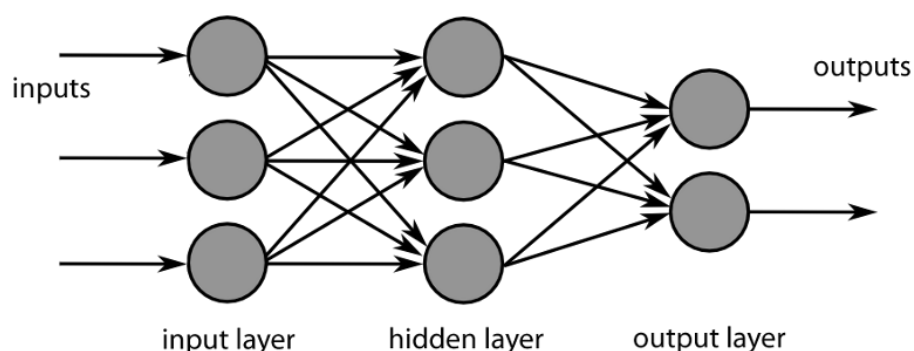
โดยที่  $w_i$  คือ น้ำหนักของเพอร์เซปตรอนตัวที่  $i$  ซึ่งเราจะปรับในสมการ (2) เรียกว่า อัตราการเรียนรู้ (Learning rate) เป็นค่าที่กำหนดว่าในการเรียนรู้แต่ละรอบจะเปลี่ยนแปลงน้ำหนักของข้อมูลขาเข้าทั้งหมดด้วยอัตราส่วนเท่าใด เมื่อเทียบผลต่างระหว่างคำตอบที่ถูกต้อง  $t$  กับผลลัพธ์

#### 2.1.4.2 โครงข่ายประสาทเทียมแบบไปข้างหน้า (Feedforward neural network)

ในบางปัญหา เช่น การจำลองฟังก์ชันเอ็กกอร์ (XOR หรือ Exclusive or) เพอร์เซปตรอนเพียงตัวเดียวไม่สามารถตอบคำถามได้ จึงเป็นที่มาของโครงข่ายประสาทเทียมแบบไปข้างหน้า ซึ่งเป็นรูปแบบโครงข่ายประสาทเทียมที่ง่ายที่สุด (D. Jurafsky and J. H. Martin, 2017) ชื่อของโครงข่ายประสาทเทียมชนิดนี้เกิดจากผลลัพธ์ของเพอร์เซปตรอนของแต่ละชั้นจะไม่ย้อนกลับ

มาเป็นข้อมูลขาเข้าของชั้นก่อนหน้าอีก เป็นการส่งต่อข้อมูลแบบเดินทางเดียว สำหรับโครงสร้างของโครงข่ายประสาทเทียมชนิดนี้ แบ่งออกเป็น 3 ชั้น ประกอบด้วย

- 1) ชั้นข้อมูลขาเข้า (Input layer) ยังคงเป็นค่าตัวเลขจำนวน  $n$  ค่า ตั้งแต่  $x_1, \dots, x_n$  เช่นเดิม
- 2) ชั้นซ่อน (Hidden layer) ประกอบด้วย เพอร์เซปตรอนหลายตัว ซึ่งข้อมูลขาเข้าทุกตัวจากชั้นข้อมูลขาเข้าจะผ่านเข้าเพอร์เซปตรอนเหล่านี้ และแต่ละตัวจะคำนวณผลลัพธ์ด้วยค่าน้ำหนักที่แตกต่างกันจนได้ผลลัพธ์ออกมา ผลลัพธ์เพอร์เซปตรอนทุกตัวในชั้นนี้จะนำเข้าสู่ชั้นผลลัพธ์ต่อไป
- 3) ชั้นผลลัพธ์ (Output layer) ประกอบด้วยเพอร์เซปตรอนอีกเช่นกัน และการทำงานคล้ายชั้นซ่อน แต่ข้อมูลขาเข้าเพอร์เซปตรอนในชั้นนี้ คือผลลัพธ์จากเพอร์เซปตรอนแต่ละตัวในชั้นซ่อนนั่นเอง แต่เนื่องจากผลลัพธ์นี้จะให้ค่าเป็นผลลัพธ์ของโครงข่ายประสาทเทียมแบบไปข้างหน้า ดังนั้นจำนวนของเพอร์เซปตรอนในชั้นผลลัพธ์นี้จะขึ้นอยู่กับลักษณะของปัญหาที่ใช้โครงข่ายรูปแบบนี้ในการแก้ปัญหาด้วย



ภาพประกอบ 2.4 โครงข่ายประสาทเทียมแบบไปข้างหน้า (Jeff Hu, 2018)

โดยสมการที่เกิดขึ้นในโครงข่ายประสาทเทียมแบบไปข้างหน้าโดยใช้เพอร์เซปตรอน คือ

$$h_i(x) = \begin{cases} 1 & \text{if } w_0 + \sum_{i=1}^n w_i x_i > 0 \\ -1 & \text{otherwise} \end{cases}$$

$$o_i(x) = \begin{cases} 1 & \text{if } w_0 + \sum_{i=1}^p w_i h_i > 0 \\ -1 & \text{otherwise} \end{cases}$$

กำหนดให้

$h_i$  เป็น สมการของเพอร์เซปตรอนตัวที่  $i$  ในชั้นซ่อน

$p$  เป็น จำนวนเพอร์เซปตรอนในชั้นซ่อน

$O_i$  เป็น สมการของเพอร์เซปตรอนตัวที่  $i$  ในชั้นผลลัพธ์

#### 2.1.4.3 กฎการสอนโครงข่ายประสาทเทียมแบบไปข้างหน้า

ในการสอนโครงข่ายประสาทเทียมแบบไปข้างหน้าที่มีองค์ประกอบเป็นหน่วยย่อย (ไม่ใช่เพอร์เซปตรอน) มักมีค่าความผิดพลาด (Error) เกิดขึ้นเสมอ ดังนั้นกฎการสอนจะเป็นการลดค่าความผิดพลาดให้เหลือน้อยที่สุด กล่าวคือ เข้าใกล้ 0 ให้มากที่สุดนั่นเอง ทั้งนี้ เราจะมองค่าความผิดพลาดเป็นฟังก์ชันต้นทุน (Cost function) หรือฟังก์ชันสูญเสีย (Loss function) เพื่อใช้สอนโครงข่าย

กำหนดให้

$E$  คือ ค่าความผิดพลาด หรือฟังก์ชันการสูญเสีย

$N$  คือ จำนวนข้อมูลสอนทั้งหมด

$t_i$  คือ คำตอบที่ถูกต้องของข้อมูลสอนลำดับที่  $i$

$O_i$  คือ ผลลัพธ์ปัจจุบันของข้อมูลสอนลำดับที่  $i$  (ซึ่งต้องปรับให้ตรงกับคำตอบที่ถูกต้องหากผลลัพธ์ยังผิดอยู่)

ค่าความผิดพลาดที่นิยมเลือกใช้ มีดังนี้

- 1) ค่าเฉลี่ยความผิดพลาดยกกำลังสอง (Mean Squared Error : MSE)

$$E = \frac{1}{2} \sum_{i=1}^N (t_i - o_i)^2$$

2) ค่าเฉลี่ยครอส-เอนโทรปีแบบทวิภาค (Binary Cross-Entropy : BCE)

$$E = -\frac{1}{N} \sum_{i=1}^N t_i \cdot \log(o_i) + (1 - t_i) \cdot \log(1 - o_i)$$

3) ค่าติดลบอัลกอริทึมของความเป็นไปได้ (Negative Log Likelihood : NLL) หากกำหนดให้  $C$  เป็นจำนวนประเภทของคำตอบทั้งหมดที่เป็นไปได้และ  $d_i$  คือผลต่างของความน่าจะเป็นระหว่างคำตอบที่ถูกต้องกับคำตอบทั้งหมดที่ทำนาย สามารถเขียนสมการฟังก์ชันการสูญเสียได้ว่า

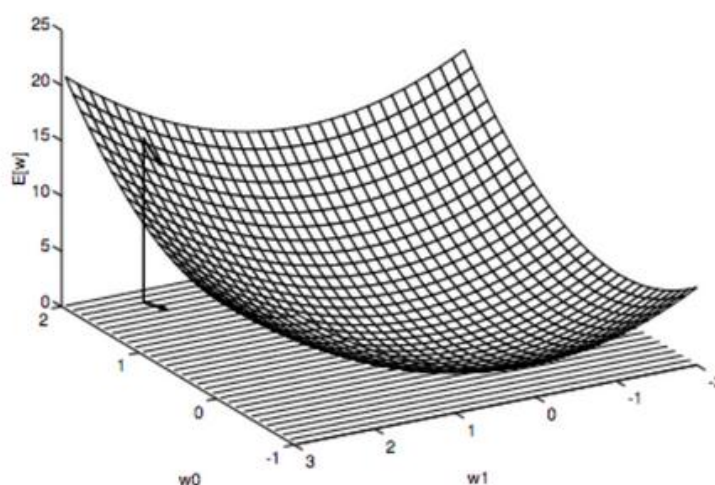
$$E = -\frac{1}{C} \sum_{i=1}^C \log d_i$$

ในการสอนโครงข่ายประสาทเทียมแบบไปข้างหน้า จะมีการหาค่าที่เหมาะสมที่สุด (Optimization) เพื่อมุ่งลดความผิดพลาดให้ได้มากที่สุดในการสอนข้อมูลแต่ละรอบด้วย วิธีการที่นิยมใช้กันมากวิธีหนึ่งคือ เอสจีดี (SGD ย่อมาจาก Stochastic Gradient Descent) โดยจะนำค่าฟังก์ชันการสูญเสียมาหาค่าเกรเดียนเทียบน้ำหนัก ทั้งนี้ ปรับน้ำหนักสำหรับข้อมูลขาเข้าลำดับที่  $i$  ได้ดังนี้

$$w_i \leftarrow w_i + \Delta w_i$$

$$\Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

ถ้าเรามีเพอร์เซปตรอนหรือหน่วยย่อยที่มีข้อมูลขาเข้า 1 ตัว และเรานำค่าน้ำหนักของขาเข้านั้นแทนด้วย  $w_1$  ค่าน้ำหนักพิเศษ  $w_0$  และค่าความผิดพลาด  $E[w]$  มาสร้างสเปซความผิดพลาด จะพบว่าผิวของค่าความผิดพลาดมีลักษณะเป็นผิวโค้ง โดยค่าความผิดพลาดจะลดลงในลักษณะการลงเนินเขาเพื่อไปหาจุดต่ำสุด



ภาพประกอบ 2.5 สเปซแสดงค่าความผิดพลาดในการสอนเพอร์เซปตรอน  
(Tom M. Mitchell, 1997)

## 2.2 การทดสอบโมเดล

ในการสร้างโมเดลนั้น จำเป็นต้องมีการทดสอบโมเดลเพื่อให้ทราบว่าโมเดลที่สร้างขึ้นมามีประสิทธิภาพเพียงใด ซึ่งโดยทั่วไปมี 3 วิธีคือ

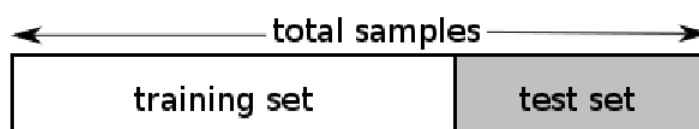
1. Self Consistency Test เป็นการเอาชุดข้อมูลสอนมาเป็นตัวทดสอบเลย เริ่มจากการสร้างโมเดลด้วยข้อมูลสอน หลังจากนั้นนำโมเดลที่สร้างได้มาพยากรณ์ข้อมูลสอนชุดเดิม ซึ่งการทดสอบโมเดลด้วยวิธีนี้จะให้ผลการทดสอบที่มีค่าสูง อาจใกล้เคียง 100% เนื่องจากเป็นข้อมูลชุดเดิมที่ระบบได้ทำการเรียนรู้มาแล้ว แต่ผลการทดสอบที่ได้ไม่เหมาะที่นำไปใช้ในงานวิจัยต่างๆ แต่เหมาะสำหรับการใช้ทดสอบเพื่อดูแนวโน้มของโมเดลที่สร้างขึ้น ซึ่งถ้าผลที่ออกมามีเปอร์เซ็นต์ที่น้อย อาจหมายถึงชุดข้อมูลกับโมเดลไม่เหมาะสมกัน เป็นต้น

2. Split Test เป็นการแบ่งข้อมูลด้วยการสุ่มออกเป็น 2 ส่วน เช่น 70 : 30 หรือ 80 : 20 โดยข้อมูลส่วนแรกใช้ในการสร้างโมเดล และข้อมูลส่วนที่สองใช้ในการทดสอบโมเดล ซึ่งวิธีการนี้

หากใช้การสุ่มเพียงครั้งเดียว ผลการทดสอบออกมาในลักษณะดีหรือแย่ขึ้นอาจขึ้นอยู่กับวิธีการเลือกสุ่มข้อมูล ดังนั้นหากต้องการใช้วิธีนี้ในการทดสอบโมเดลให้ได้ผลดี ควรทำการสุ่มข้อมูลหลายๆ ครั้ง

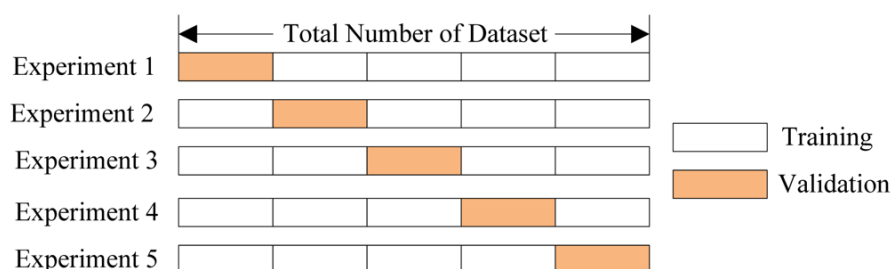
3. Cross-Validation เป็นวิธีการทดสอบโมเดล โดยการนำข้อมูลเข้านั้นจะต้องแยกข้อมูลบางส่วนออกก่อนที่จะเริ่มทำการสอนและใช้ข้อมูลที่แยกออกมานั้นใช้ในการทดสอบ โดย Cross-Validation มีหลายวิธี เช่น

- Holdout Method ชุดข้อมูลจะถูกแบ่งออกเป็นสองชุด คือชุดการสอน (Training Set) และชุดข้อมูลทดสอบ (Testing Set) ข้อดีของวิธีนี้คือสามารถประเมินข้อมูลชุดใหญ่ในเวลาไม่นาน



ภาพประกอบ 2.6 ตัวอย่างการแบ่งข้อมูลชุดการสอนและข้อมูลชุดทดสอบตามหลักการของ Cross-Validation แบบ Holdout Method (Nikolaos Kosmas Chlis, 2015)

- K-fold Cross Validation วิธีการนี้จะทำการแบ่งข้อมูลออกเป็นหลายส่วน (มักแสดงด้วยค่า K) เช่น 5-fold Cross Validation คือการแบ่งข้อมูลออกเป็น 5 ส่วน โดยที่แต่ละส่วนมีจำนวนข้อมูลเท่ากัน หลังจากนั้นข้อมูลส่วนหนึ่งจะใช้เป็นตัวทดสอบประสิทธิภาพของโมเดล และทำวนไปเช่นนี้จนครบจำนวนที่แบ่งไว้ (เอกสิทธิ์ พัทธวงศ์ศักดิ์, 2557) ดังตัวอย่างภาพประกอบที่ 2.7 ที่มีการแบ่งชุดข้อมูลเป็น 5 ส่วน



ภาพประกอบ 2.7 ตัวอย่างการแบ่งข้อมูลชุดข้อมูล K-fold (K=5) ตามหลักการของ Cross-Validation แบบ K-fold Cross Validation

### 2.3 เกณฑ์การวัดประสิทธิภาพของโมเดลการจำแนกประเภทข้อมูล (Classifier Evaluation Metrics)

การวัดประสิทธิภาพของโมเดลมีด้วยกันหลายวิธี ในงานวิจัยนำเสนอการวัดประสิทธิภาพของโมเดลด้วยกัน 2 วิธี ดังนี้

1) ตารางการจำแนกหรือเมทริกซ์ความสับสน (Confusion Matrix)

ตารางการจำแนกหรือเมทริกซ์ความสับสน เป็นตารางความถี่สองทางที่นับจำนวนความถี่ของเหตุการณ์ที่สนใจกับเหตุการณ์ที่ไม่เกิดขึ้นจริง กับเหตุการณ์ที่ใช้สมการทำนายในการจำแนกเหตุการณ์โดยใช้ความน่าจะเป็นที่กำหนดตามตารางการจำแนกหรือเมทริกซ์ความสับสนในตารางที่ 2.1

	สมการทำนาย		รวม
	เหตุการณ์ไม่สนใจ	เหตุการณ์สนใจ	
เหตุการณ์ไม่สนใจ	TN	FP	TN+FP
ข้อมูลจริง เหตุการณ์ สนใจ	FN	TP	FN+TP
รวม	TN+FN	FP+TP	TN+FN+FP+TP P=n

#### ตารางที่ 2.1 ตารางการจำแนกหรือเมทริกซ์ความสับสน

เกณฑ์ในการตรวจสอบตัวแบบที่ได้จากตารางจำแนกหรือเมทริกซ์ความสับสนมีดังนี้

- ลบจริง (True Negative) หมายถึง เหตุการณ์ที่ไม่สนใจที่สมการทำนายจำแนกเป็นเหตุการณ์ไม่สนใจ และให้ TN แทนจำนวนลบจริง



- บวกเท็จ (False Positive) หมายถึง เหตุการณ์ไม่สนใจที่สมการทำนายจำแนกเป็น เหตุการณ์สนใจ และให้ FP แทนจำนวนบวกเท็จ
- ลบเท็จ (False Negative) หมายถึง เหตุการณ์สนใจที่สมการทำนายจำแนกเป็น เหตุการณ์ไม่สนใจ และให้ FN แทนจำนวนลบเท็จ
- บวกจริง (True Positive) หมายถึง จำนวนเหตุการณ์ที่สนใจที่สมการทำนายจำแนก เป็นเหตุการณ์สนใจ และให้ TP แทนจำนวนบวกจริง
- TN+FN หมายถึง จำนวนเหตุการณ์ทั้งหมดที่สมการทำนายจำแนกเป็นเหตุการณ์ ไม่สนใจ
- FP+TP หมายถึง จำนวนเหตุการณ์ทั้งหมดที่สมการทำนายจำแนกเป็นเหตุการณ์ สนใจ
- TN+FP หมายถึง จำนวนเหตุการณ์ไม่สนใจทั้งหมด
- FN+TP หมายถึง จำนวนเหตุการณ์สนใจทั้งหมด
- n เป็น จำนวนเหตุการณ์ทั้งหมด (จิราวัลย์, 2558 : 338)

2) การวัดค่าความถูกต้องแม่นยำของโมเดล (Accuracy หรือ Correct Percentage) เป็น เกณฑ์วัดค่าความถูกต้องแม่นยำของโมเดล ในการจำแนกประเภท เพื่อบ่งบอกระดับความถูกต้องใน การจำแนกประเภทข้อมูล โมเดลที่ได้จากการประมวล ได้แก่ สัดส่วนระหว่างจำนวนข้อมูลทั้งหมดที่ จำแนกประเภทถูกต้องทั้งประเภท Positive และ Negative กับจำนวนข้อมูลทั้งหมดที่มีการถูก จำแนกประเภท ดังสมการที่ (1)

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \quad (1)$$

### 3. ระบบแนะนำข่าว (News Recommendation System)

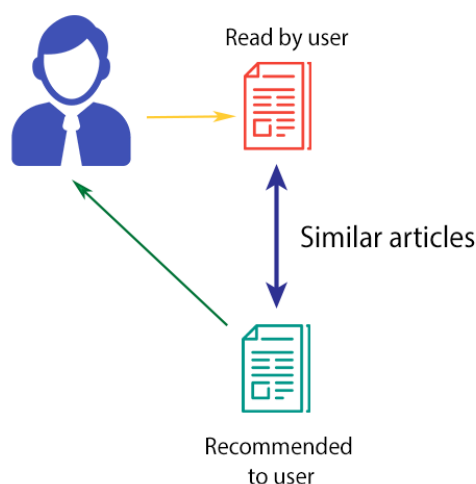
ระบบแนะนำข่าว (Lei Li และคณะ, 2011) คือ การให้บริการการแนะนำข่าวที่มีจำนวนมหาศาลในโลกออนไลน์ออกมาจำนวนหนึ่งให้กับผู้อ่านข่าวออนไลน์ เมื่อพิจารณาตามวิธีการวิจัยจะสามารถแบ่งวิธีการแนะนำข่าวออกได้เป็น 3 กลุ่ม คือ การกรองด้วยเนื้อหา (Content Filtering), การกรองแบบร่วมมือ (Collaborative Filtering) และการกรองข้อมูลแบบผสม (Hybrid Approaches)

### 3.1 การกรองด้วยเนื้อหา (Content Filtering)

ในระบบแนะนำข่าวที่ใช้การกรองด้วยเนื้อหา (Content filtering) เป็นวิธีการค้นคืนบทความข่าวโดยการจับคู่จากความคล้ายคลึงกันระหว่างบทความข่าวใหม่กับประวัติการอ่านข่าวที่ผ่านมาของผู้อ่านข่าวออนไลน์ว่าตรงกันหรือไม่ ถ้าใช่ก็จะนำเสนอข้อมูลนั้นทันที (นลินี โสพัทธิต, 2555 : 7)

ดังนั้น วิธีนี้เป็นการคำนวณหาค่าความคล้ายคลึงระหว่างเอกสารกับประวัติการอ่านข่าวที่ผ่านมาของผู้อ่านข่าวออนไลน์ โดยนำเนื้อหาในข้อมูลข่าว เช่น คำสำคัญ (Keywords), วลี (Phrases) หรือคุณลักษณะ (Feature) มาสร้างเป็นประวัติผู้อ่านข่าวออนไลน์แต่ละคน เพื่อค้นหาข้อมูลที่ผู้อ่านข่าวออนไลน์คนนั้นสนใจ

ดังภาพประกอบ 2.8



ภาพประกอบที่ 2.8 การกรองด้วยเนื้อหา (Content Filtering)

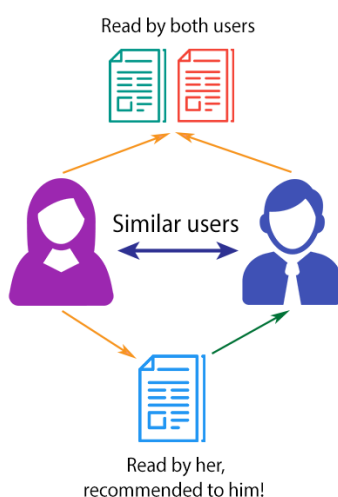
วิธีการของการกรองด้วยเนื้อหานั้น จะแตกต่างจากวิธีการกรองแบบร่วมมือคือ แทนที่จะแนะนำข่าวโดยใช้ข้อมูลของผู้อ่านข่าวออนไลน์ หรือแนะนำข่าวที่มีความใกล้เคียงกัน ระบบจะทำการเลือกข่าวโดยอิงจากความพึงพอใจของผู้อ่านข่าวออนไลน์ว่าชอบในข่าวที่มีคุณลักษณะใดคุณสมบัติอย่างไร ซึ่งยังแสดงให้เห็นอีกด้วยว่าผู้อ่านข่าวออนไลน์แต่ละคนมีเกณฑ์การเลือกข่าวอย่างไรหรือยึดในคุณลักษณะใดของข่าวเป็นหลัก โดยการกรองด้วยเนื้อหานั้นมีขั้นตอนการทำงานดังนี้

- 1) สร้างประวัติผู้อ่านข่าวออนไลน์โดยอิงจากคุณลักษณะหรือคุณสมบัติของบทความข่าว ซึ่งคำนวณจากพฤติกรรมการอ่านข่าวที่ผ่านๆ มาของผู้อ่านข่าวออนไลน์
- 2) คำนวณความสัมพันธ์ระหว่างประวัติผู้อ่านข่าวออนไลน์และบทความข่าว
- 3) นำผลลัพธ์ที่ได้จากข้อ 2) มาปรับค่าน้ำหนักแต่ละคุณลักษณะของบทความข่าว
- 4) เลือกบทความข่าวที่มีความสัมพันธ์กับผู้อ่านข่าวออนไลน์ใหม่มากที่สุด โดยดูจากความสำคัญที่ผู้อ่านข่าวออนไลน์ใหม่ให้ไว้ในแต่ละคุณลักษณะและค่าน้ำหนักของแต่ละคุณลักษณะของบทความข่าวที่มีอยู่ในระบบ

ซึ่งวิธีการของการกรองด้วยเนื้อหานั้นจะแก้ปัญหาที่เกิดขึ้นในระบบการกรองแบบร่วมมือได้ เพราะการคำนวณแต่ละครั้งจะไม่ใช่การใช้ข้อมูลจากผู้ใช้ที่เคยเลือกบทความข่าวที่มีความใกล้เคียงกันกับผู้อ่านข่าวออนไลน์ใหม่ ระบบจึงสามารถแนะนำข่าวได้แม้ว่ามีข้อมูลอยู่ในระบบเป็นจำนวนน้อย (Cold Start)

### 3.2 การกรองแบบร่วมมือ (Collaborative Filtering)

ระบบจะวิเคราะห์ข้อมูลจากประวัติการของผู้อ่านข่าวออนไลน์ที่ผ่านมา โดยอาจจะใช้ข้อมูลของกลุ่มผู้อ่านข่าวออนไลน์ที่เหมือนกันเพื่อคาดการณ์การเหตุการณ์ในอนาคต หรือสร้างตัว



แบบความน่าจะเป็น (Lei Li และคณะ, 2011) ดังภาพประกอบ 2.9

### ภาพประกอบ 2.9 การกรองแบบร่วมมือ (Collaborative Filtering)

วิธีการทำงานของการกรองแบบร่วมมือคือ จะพิจารณาผู้อ่านข่าวออนไลน์ที่มีพฤติกรรมการอ่านข่าวลักษณะคล้ายกันในระบบ เช่นอาจจะเคยอ่านข่าวเดียวกันมาก่อน เป็นต้น มาทำนายข่าวใหม่และแนะนำให้กับผู้อ่านข่าวออนไลน์ใหม่ โดยระบบแนะนำประเภทนี้จะมีขั้นตอนการประมวลผล 3 ขั้นตอนใหญ่ๆ คือ

- 1) สร้างประวัติผู้อ่านข่าวออนไลน์ตามข้อมูลที่จะใช้เป็นพื้นฐานของระบบ
- 2) คัดเลือกผู้อ่านข่าวออนไลน์ที่เคยเลือกวัตถุนั้น (Co-rated item) ซึ่งในที่นี้เราหมายถึง บทความข่าว ที่มีความใกล้เคียงหรือคล้ายคลึงกันขึ้นมาตามจำนวนที่กำหนดไว้ โดยการเปรียบเทียบประวัติผู้อ่านข่าวออนไลน์ตามข้อมูลที่ใช้เป็นพื้นฐานของระบบ
- 3) คำนวณว่าผู้อ่านข่าวออนไลน์ข่าวและข่าวนั้นมีความเหมาะสมกันมากน้อยเพียงใด โดยอาศัยข้อมูลที่หาได้ในขั้นตอนที่ 2 แล้วจึงเลือกบทความข่าวที่มีความสัมพันธ์กับพฤติกรรมมากที่สุดให้กับผู้อ่านข่าวออนไลน์

### 3.3 การกรองข้อมูลแบบผสม (Hybrid Approaches)

การกรองข้อมูลแบบผสม เป็นการรวม การกรองด้วยเนื้อหา (Content Filtering) และการกรองแบบร่วมมือ (Collaborative Filtering) เข้าด้วยกัน เป็นวิธีผสมผสานเพื่อให้ผลการแนะนำที่ดีขึ้น มีความแม่นยำมากขึ้น (นลินี โสพัตสถิต, 2555: 15) ซึ่งเป็นการหลีกเลี่ยงข้อจำกัดของการกรองด้วยเนื้อหาและการกรองแบบร่วมมือ แต่อาจทำให้มีความซับซ้อนในการใช้ทรัพยากรในการแนะนำสูง

## 4. การทดสอบประสิทธิภาพของระบบด้วย A/B testing

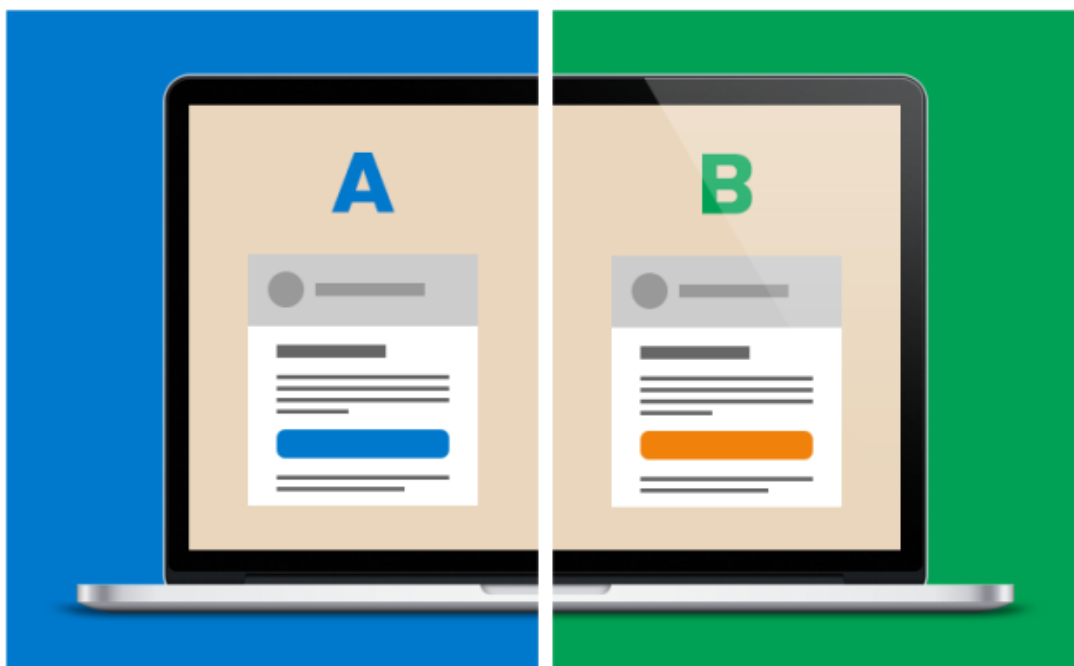
A/B testing หรือ Split test เป็นการทดสอบรูปแบบส่วนประกอบต่างๆ บนหน้าแพลตฟอร์มเพื่อหารูปแบบที่ทำให้ได้ผลลัพธ์ที่ดีที่สุด โดยการแบ่งกลุ่มเป้าหมายที่จะทำการทดสอบออกเป็น 2 กลุ่มเท่าๆ กัน กลุ่มแรกให้เห็นหน้าแพลตฟอร์มแบบ A กลุ่มที่ 2 ให้เห็นหน้าแพลตฟอร์ม

แบบ B แล้ววัดผลว่าแบบใดให้ผลลัพธ์ตามเป้าหมายที่ตั้งไว้ได้ดีที่สุด ในงานวิจัยฉบับนี้จะหมายถึงการวัดอัตราการคลิกอ่านข่าวในหน้าแพลตฟอร์ม

ในกรณีที่มีการทดสอบมากกว่า 2 แบบ จะเรียกรูปแบบการทดสอบนั้นว่า A/B/n testing ซึ่งอาจจะเป็น A/B/C/D testing ก็ได้ แต่การที่จะสามารถทดสอบได้หลายๆ รูปแบบนั้น ต้องแน่ใจว่ากราฟฟิคมีในหน้านั้นมากพอสำหรับการทำทดสอบเพื่อให้ได้ผลลัพธ์ที่มีความถูกต้อง

A/B Testing จะพิจารณาจากอัตราการเข้าชม ซึ่งแพลตฟอร์มแต่ละประเภทจะมี อัตราการเข้าชมที่แตกต่างกัน เช่น

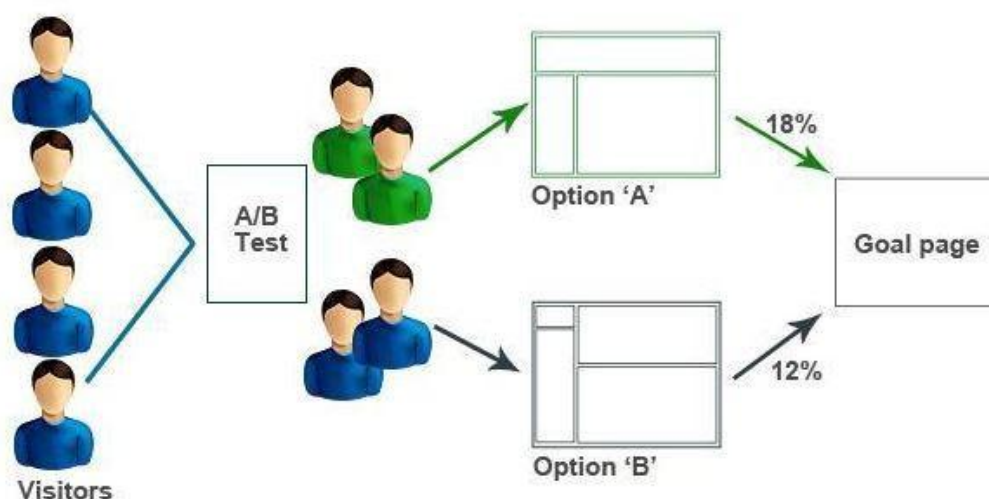
1. E-Commerce Website อาจจะวัดจาก ผู้เข้าชมแพลตฟอร์มที่ซื้อสินค้ากับผู้เข้าชมแพลตฟอร์มเฉยๆ
2. Software As A Service Web App อาจจะวัดจากผู้เข้าใช้ Application ที่ลงทะเบียนทดลองใช้ (Trail version) และเปลี่ยนมาเป็นจ่ายเงิน (paid version)
3. ข่าว หรือ Media แพลตฟอร์ม อาจจะวัดจากผู้เข้าชมแพลตฟอร์มที่คลิกโฆษณาหรือกดติดตาม (Subscriptions) กับผู้เข้าชมแพลตฟอร์ม ซึ่ง Conversion rate นี้จะเป็นตัววัดประสิทธิภาพของตัวแปร A, B ว่าแบบไหนที่ออกแบบแล้วมีประสิทธิภาพมากกว่ากัน



ภาพประกอบที่ 2.10 การทำ A/B testing

### ขั้นตอนการทำ A/B Testing มีดังนี้

1. ศึกษาข้อมูลแพลตฟอร์ม โดยใช้ Google Analytics เพื่อหาปัญหา เช่น ผู้อ่านข่าวออนไลน์แพลตฟอร์มเข้ามาอ่านข่าวแล้วออกจากเว็บไปโดยไม่ดูหน้าอื่นต่อ
2. การตั้งสมมติฐาน โดยสมมติฐานของผู้วิจัยคือ บทความข่าวที่แนะนำโดยโมเดลการแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าว จะทำให้อัตราการคลิกอ่านข่าวสูงขึ้น
3. ทดสอบสมมติฐาน โดยการสร้างตัวแปรที่สอดคล้องกับสมมติฐานจากข้อ 2 โดยเทียบกับแพลตฟอร์มปัจจุบัน เช่น ตัวแปร A เป็นแพลตฟอร์มปัจจุบันที่ไม่มีการเปลี่ยนแปลง และตัวแปร B เป็นแพลตฟอร์มที่มีรูปโปรโมชันที่มีขนาดใหญ่ขึ้น มีเนื้อหาดึงดูดให้เข้ามาคลิก ในที่นี้ผู้วิจัยเลือกให้ตัวแปร B เป็นแพลตฟอร์มที่มีระบบแนะนำข่าวรายบุคคล
4. ดูผลลัพธ์จากการสร้างตัวแปร A/B test วิเคราะห์ผลลัพธ์จากข้อ 3 ว่าตัวแปร B ที่เราสร้างรูปโปรโมชันใหม่ มีผู้เข้าชมเข้ามาคลิก เพิ่มขึ้นขนาดไหน ถ้ามีคนคลิกเยอะ ซึ่งอาจมีการทดสอบสมมติฐานว่ามีคนคลิกเพิ่มมากขึ้นอย่างมีนัยสำคัญหรือไม่ ซึ่งแสดงว่าควรที่จะเพิ่มระบบแนะนำข่าวรายบุคคลให้กับผู้อ่านข่าวออนไลน์ เพื่อเพิ่มอัตราการเข้าชมให้มากขึ้น แต่ถ้าไม่มีการเปลี่ยนแปลง ระบบแนะนำข่าวรายบุคคลไม่มีผลกับแพลตฟอร์ม ควรจะกลับไปข้อ 2 เพื่อตั้งสมมติฐานใหม่และลองใหม่เพื่อหาสาเหตุปัญหาบนแพลตฟอร์ม



ภาพประกอบที่ 2.11 ตัวอย่างการวัดผลที่ได้จาก A/B testing

## 5. งานวิจัยที่เกี่ยวข้อง

**Jiahui Liu และคณะ** (Jiahui Liu, Peter Dolan, Elin Ronby Pedersen. 2010.) ได้ทำ งานวิจัยเรื่อง Personalized News Recommendation Based on Click Behavior เป็นการพัฒนาระบบแนะนำข่าวสารส่วนบุคคลใน Google News สำหรับผู้ใช้ที่ล็อกอินและเปิดใช้ประวัติเว็บอย่าง ชัดเจน โดยได้ใช้เทคนิคการพัฒนาระบบแนะนำข้อมูล (Recommender System) 2 รูปแบบ ผสมผสานกันคือ การกรองด้วยเนื้อหา (Content filtering) และการกรองแบบร่วมมือ (Collaborative filtering) หลังจากนั้นจะทำการวัดผลโดยผู้ใช้งานจะถูกแบ่งเป็น 2 กลุ่มในปริมาณเท่าๆ กันด้วย วิธีการสุ่ม คือกลุ่มทดลอง และกลุ่มควบคุม เมื่อผู้ใช้เข้าใช้งาน Google News Session ข่าวแนะนำ ที่ได้คัดสรรข่าวตามความสนใจของผู้ใช้คนนั้นโดยเฉพาะจะถูกสร้างขึ้น จากนั้นระบบจะเริ่มคำนวณ อัตราการคลิก (click rate) จากประวัติการคลิกของผู้ใช้ โดยผลการทดลองสรุปว่า ระบบแนะนำข่าว พัฒนาขึ้นมาเป็นการกรองแบบผสมผสาน เพิ่มประสิทธิภาพการแนะนำข่าวได้ดีกว่าการใช้การกรอง แบบร่วมมือเพียงอย่างเดียว และเข้าใช้งานแพลตฟอร์มถี่ขึ้นโดยใช้ระยะเวลาเท่าเดิม

งานวิจัยนี้แตกต่างจากงานวิจัยของผู้วิจัยในแง่ที่ว่า ผู้วิจัยใช้โมเดลแนะนำข่าวรายบุคคล โดยใช้เทคนิคการจำแนกข้อมูล (Classification) ร่วมกับการกรองด้วยเนื้อหา (Content filtering)

นอกจากนั้นยังมีฟังก์ชันในการเลือกข่าว รวมถึงการวัดประสิทธิภาพของของระบบด้วยอัตราการคลิก (click rate) โดยเปรียบเทียบผลด้วย A/B Testing อีกด้วย

**Chen Li และคณะ** (Chen Li, Zhengtao Jiang. 2016.) ได้ทำการวิจัยเรื่อง A Hybrid News Recommendation Algorithm based on User's Browsing Path เป็นการออกแบบอัลกอริทึมของระบบแนะนำข่าวด้วยข้อมูล browsing path ของผู้ใช้ ซึ่งเป็นการนำอัลกอริทึมมาวิเคราะห์และเลียนแบบพฤติกรรมผู้ใช้เป็นรายบุคคล จากนั้นนำหลักการเดียวกันไปทำนายพฤติกรรมผู้ใช้งานใหม่ และประเมินผลว่าระบบแนะนำนั้นสอดคล้องกับลักษณะของผู้ใช้งานคนนั้น ๆ หรือไม่ โดยเก็บข้อมูล ID ผู้ใช้ , ID ข่าว และระยะเวลาที่ผู้ใช้คลิก จากการทดลองพบว่า การแนะนำโดยใช้เนื้อหาเป็นเกณฑ์ (recommendation based on content) ไม่ใช่อัลกอริทึมที่ดีที่สุด เนื่องจากใช้เวลาในการประมวลผลนานเกินไป ส่วนผลการทดลองที่ออกมาดีคือการใช้อัลกอริทึมแบบผสม (hybrid recommend algorithm) คือการรวมกันระหว่างการกรองด้วยเนื้อหา (Content filtering) และการกรองแบบร่วมมือ (collaborative filtering)

งานวิจัยนี้แตกต่างจากงานวิจัยของผู้วิจัยในแง่ที่ว่า ผู้วิจัยใช้การกรองด้วยเนื้อหา (Content filtering) ร่วมกับการจำแนกประเภท (Classification) โดยการกรองด้วยเนื้อหานั้นจะแนะนำข่าวที่มีคุณลักษณะคล้ายกัน อยู่ในหมวดหมู่เดียวกันกับข่าวที่ผู้อ่านข่าวออนไลน์อ่านก่อนหน้า เพื่อแนะนำข่าวถัดไปให้กับผู้อ่านข่าวออนไลน์ ซึ่งมีข้อดีคือถ้าผู้อ่านข่าวออนไลน์เป็นผู้อ่านข่าวออนไลน์ที่เข้ามาใช้งานแพลตฟอร์มเป็นครั้งแรกก็สามารถแนะนำข่าวถัดไปให้ผู้อ่านข่าวออนไลน์ได้โดยพิจารณาคุณลักษณะของเนื้อหาที่ผ่านมา โดยไม่ต้องสนใจรูปแบบหรือคุณลักษณะของผู้อ่านข่าวออนไลน์ ประกอบการแนะนำข่าว ส่วนการจำแนกประเภท (Classification) จะหาโมเดลที่ดีที่สุดในการคัดเลือกข่าวที่เหมาะสมกับผู้อ่านข่าวออนไลน์ที่มีพฤติกรรมสอดคล้องกับการคลิกอ่านบทความข่าว

**Wei Chu และคณะ** (Wei Chu, Seung-Taek Park. 2009.) ได้ทำงานวิจัยเรื่อง Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models โดยนำเสนอรูปแบบการวิเคราะห์การถดถอยเชิงเส้นคู่ สำหรับระบบแนะนำข่าวรายบุคคลที่มีการเปลี่ยนแปลงเนื้อหาอยู่เสมอ โดยใช้คุณสมบัติเป็นเกณฑ์ในการวิเคราะห์ค่าความสัมพันธ์คุณลักษณะของผู้อ่านข่าวออนไลน์และคุณลักษณะของเนื้อหา คำนวณด้วยฟังก์ชัน Bilinear Regression กับผลตอบรับของผู้อ่านข่าวออนไลน์ ซึ่งในการทดลองนี้เพื่อทำการค้นหารอบการทำงาน (framework) ที่มีความยืดหยุ่นกับการแนะนำข่าวรายบุคคล ซึ่งได้ผลดีกับการทดลองในระบบปิด (offline) โดยบทความข่าวที่มีคุณสมบัติแปรผัน (dynamic feature) สามารถแนะนำข่าวใหม่ได้ทันทีและแม่นยำ งานวิจัยนี้แตกต่างจากงานวิจัยของผู้วิจัยในแง่ที่ว่า ผู้วิจัยได้ทำการทดลองกับผู้อ่านข่าวออนไลน์ที่ใช้



งานบนระบบเปิด (online) และผลวัดได้ทันทีโดยใช้เครื่องมือ A/B Testing เพื่อเปรียบเทียบอัตราการคลิกเข้ามาอ่านข่าว

**นิเวศ จิระวิจิตชัย และคณะ** (นิเวศ จิระวิจิตชัย, ปริญญา สงวนสัตย์, พยุง มีสัจ .2553.) ได้ทำการวิจัยวิจัยเรื่อง การพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ โดยในการทดลองใช้โมเดลการพัฒนาประสิทธิภาพการจัดหมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ โดยทำการทดสอบด้วยวิธี 10-fold cross validation โดยจะตัดค่าแบบวิธีการตัดค่า แล้วจึงส่งข้อมูลเข้าทดสอบกับอัลกอริทึมการจำแนกข้อมูล เปรียบเทียบค่าความถูกต้อง (Precision) ความแม่นยำ (Recall) และนำค่าที่ได้เข้าโมเดล classifier โดยผลการทดลองในการลดทรัพยากรระบบและระยะเวลาในการประมวลผล อัลกอริทึม SVM ลดได้สูงสุด 94.3% งานวิจัยฉบับนี้ทำการทดลองกับแพลตฟอร์มไทยรัฐมาก่อน โดยทำการทดลองการจัดระบบหมวดหมู่ข่าวเพื่อให้ระบบการแนะนำข่าวบนแพลตฟอร์มเป็นอัตโนมัติมากขึ้น และช่วยลดอัตราการผิดพลาดของผู้ดูแลแพลตฟอร์ม

งานวิจัยนี้แตกต่างจากงานวิจัยของผู้วิจัยในแง่ที่ว่า งานของผู้วิจัยนั้น มีการวัดประสิทธิภาพของโมเดล (F-Measure) ในขั้นตอนการสร้างโมเดล และใช้เครื่องมือในการวัดผลสุดท้ายของงานวิจัยคืออัตราการคลิก (click rate) งานวิจัยนี้มีวัตถุประสงค์แตกต่างจากงานของผู้วิจัย ซึ่งงานนี้สามารถเอามาต่อยอดจากงานผู้วิจัยได้

### บทที่ 3

#### ระเบียบวิธีวิจัย

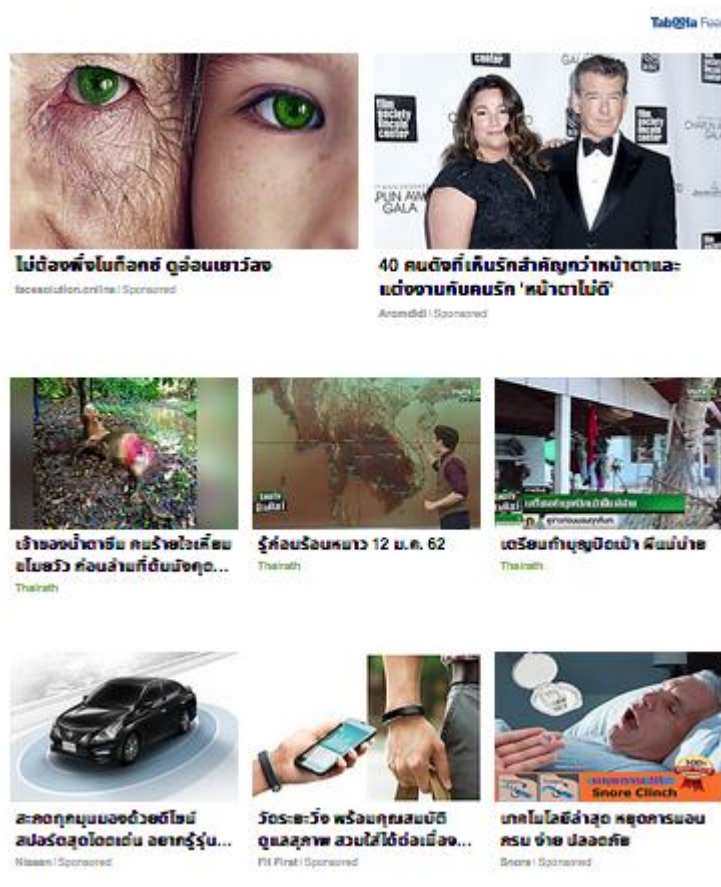
การวิเคราะห์และพัฒนาระบบแนะนำข่าวรายบุคคลโดยใช้ประวัติและพฤติกรรมผู้อ่านข่าวออนไลน์ โดยผู้วิจัยใช้เทคนิคการกรองด้วยเนื้อหาผสมกับเทคนิคการจำแนกข้อมูล (Classification) แบ่งวิธีการดำเนินงานออกเป็น 4 ขั้นตอน ดังนี้

- 3.1 การศึกษาข้อมูลและปัญหาของระบบเดิม
- 3.2 การออกแบบสถาปัตยกรรมระบบ
- 3.3 การพัฒนาระบบ
- 3.4 การวัดประสิทธิภาพของระบบ

### 3.1 การศึกษาข้อมูลและปัญหาของระบบเดิม

จากการศึกษาการทำงานของระบบเดิมของแพลตฟอร์มไทยรัฐในส่วนของข่าวนั้นพบว่า ผู้อ่านข่าวออนไลน์แพลตฟอร์มไทยรัฐส่วนใหญ่หลังจากที่เข้ามาอ่านเนื้อหาบนแพลตฟอร์ม แล้วมีผู้อ่านข่าวออนไลน์จำนวนมากออกจากแพลตฟอร์มไปโดยที่ไม่คลิกอ่านเนื้อหาอื่นต่อ โดยจากข้อมูล Truehits ในปี 2560 เวลาเฉลี่ยของผู้อ่านข่าวออนไลน์ที่เข้ามาดูแพลตฟอร์มซึ่งใช้เวลาเฉลี่ยที่  $0 < = 1$  นาที คิดเป็น 54.69% เมื่อผู้วิจัยสำรวจและวิเคราะห์องค์ประกอบของแพลตฟอร์มพบว่า แพลตฟอร์มใช้ระบบแนะนำข่าวเชิงพาณิชย์ที่ชื่อ "C" แต่บทความข่าวที่แนะนำละกันไปมีบทความข่าวจากหลายหัวข้อข่าว และมีผู้อ่านข่าวออนไลน์จำนวนมากไม่คลิกอ่านข่าวในหน้าอื่นต่อ ผู้วิจัยจึงอยากเปรียบเทียบระบบแนะนำข่าวที่ใช้ระบบแนะนำข่าวรายบุคคลของผู้ใช้เปรียบเทียบกับระบบแนะนำข่าวเชิงพาณิชย์

คุณอาจสนใจข่าวนี้



The screenshot displays a grid of recommended news items. At the top, there are two large featured articles. Below them, there are three rows of smaller article thumbnails, each with a title and a source attribution.

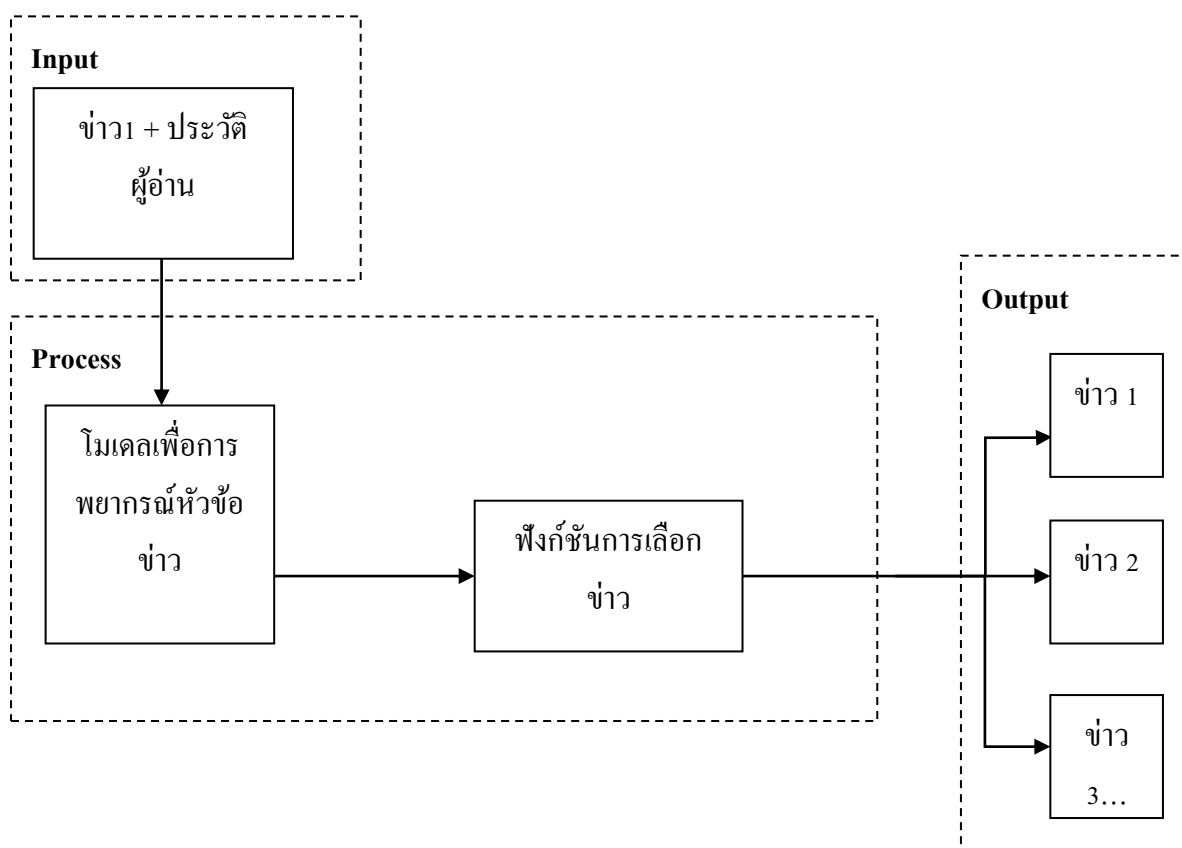
- Featured Article 1:** Image of a person's face with green eyes. Title: "ไม่ต้องพึ่งใบกอล์ฟ ดูอันเขี้ยวจ" (Don't rely on golf leaves, look at the fangs). Source: faceofthionline | Sponsored.
- Featured Article 2:** Image of a couple on a red carpet. Title: "40 คนดังก็เห็นรักสำคัญกว่าหน้าตาและ... แต่งงานกับคนรัก 'หน้าตาไม่ดี'" (40 celebrities also see love as more important than looks and... marry their love 'not good looking'). Source: Aromdidi | Sponsored.
- Row 1:**
  - Image of a person in a red shirt. Title: "เจ้าของน้ำเต้าหู้ คบรายใจเคียด... โยนขวว ก่อนล้มก็ยื่นมือจับ..." (The owner of the soybean pudding, a heartless person... threw a brick before falling and then reaching out to catch...). Source: Thairath.
  - Image of a person in a red shirt. Title: "รู้ก่อนร้อนหนาว 12 ม.ค. 62" (Know before hot or cold, Jan 12, 62). Source: Thairath.
  - Image of a person in a red shirt. Title: "เตรียมทำบุญบิณฑบาต ผัดบิณฑบาต" (Prepare to do good deeds, alms-giving, and cooking alms). Source: Thairath.
- Row 2:**
  - Image of a car. Title: "สละทุกบุญบิณฑบาตให้แม่ สบร้อนสุดโหดเย็น อากาศรู้ร..." (Sacrifice all good deeds and alms-giving to mother, experience the hottest and coldest weather, the weather knows...). Source: Nissan | Sponsored.
  - Image of a person's hand. Title: "วัดระฆัง: ฝรั่งคนสวยคนดี ดูสุขภาพ สบใสได้ต่อเนื่อง..." (Wat Rong: Beautiful and good foreigner, look at health, stay healthy continuously...). Source: Ptt First | Sponsored.
  - Image of a person's face. Title: "ปากไมโลส่าสุด หลอดการฉลน ครบ จำ ปอดคห" (The most beautiful mouth, the most beautiful person, complete memory, lungs). Source: Siam | Sponsored.

### ภาพประกอบ 3.1 ตัวอย่างบทความข่าวในหน้าเว็บที่แนะนำโดยระบบแนะนำข่าวเชิงพาณิชย์

จากภาพประกอบ 3.1 จะเห็นว่าข่าวที่แสดงในหัวข้อ “เรื่องที่คุณอาจสนใจ” ซึ่งการแนะนำบทความข่าวดังกล่าวเป็นการแนะนำโดยใช้ระบบแนะนำข่าวในเชิงพาณิชย์

### 3.2 สถาปัตยกรรมระบบ

จากการศึกษาปัญหาของระบบเดิม พบว่าแพลตฟอร์มเดิมเลือกที่จะนำเสนอ “ข่าวอื่นที่เกี่ยวข้อง” และ “เรื่องที่คุณอาจสนใจ” ด้วยการใช้คำสำคัญเป็นแรงจูงใจผู้อ่านข่าวออนไลน์อาจยังไม่มากพอ จึงทำให้ผู้อ่านข่าวออนไลน์ไม่อ่านข่าวอื่นต่อ ดังนั้น ผู้วิจัยจึงได้ทำการออกแบบสถาปัตยกรรมระบบใหม่ ซึ่งจะช่วยแก้ปัญหาการที่ผู้อ่านข่าวออนไลน์เข้ามาในแพลตฟอร์มเพียงข่าวเดียวแล้วออกไปทันทีให้เปลี่ยนเป็นอ่านข่าวอื่นที่แนะนำเพิ่มมากขึ้น ดังภาพประกอบที่ 3.3



ภาพประกอบที่ 3.2 แสดงสถาปัตยกรรมระบบ

จากภาพประกอบ 3.2 สถาปัตยกรรมระบบข้างต้นสามารถอธิบายรายละเอียดได้ดังนี้

#### **Input**

ข่าว 1 + ประวัติผู้อ่านข่าวออนไลน์ หมายถึง ข้อมูลที่ผู้อ่านข่าวออนไลน์ที่เข้ามาอ่านข่าวแรกในแพลตฟอร์มในช่วงเวลาหนึ่งๆ และประวัติผู้อ่านข่าวออนไลน์ที่เข้ามามีการคลิกอ่านข่าวต่อกันเป็นจำนวน n ข่าว ภายใน 1 session (30 นาที)

#### **Process**

โมเดลเพื่อการพยากรณ์หัวข้อข่าว หมายถึง โมเดลที่รับข้อมูลเข้ามาคือข่าว 1 และประวัติผู้อ่านข่าวออนไลน์ และนำไปพยากรณ์หัวข้อข่าว (หมวดข่าว) เพื่อแนะนำข่าวถัดไปให้กับผู้อ่านข่าวออนไลน์

ฟังก์ชันการเลือกข่าว หมายถึง ฟังก์ชันทำการคัดเลือกข่าวตามเกณฑ์ที่กำหนด เพื่อส่งข่าวออกไป ซึ่งเป็นข่าวที่แนะนำโดยระบบแนะนำข่าว

#### **Output**

ข่าว 1, ข่าว 2, ข่าว3... หมายถึง ข่าวที่ถูกคัดเลือกมาแล้วจากฟังก์ชันการเลือกข่าว เพื่อแสดงผลออกมาให้ผู้อ่านข่าวออนไลน์

### **3.3 การพัฒนาระบบ**

กระบวนการพัฒนาระบบ เป็นดังนี้

#### **1) การเก็บข้อมูลจริงและลักษณะข้อมูล**

ข้อมูลที่นำมาศึกษาเป็นข้อมูลจากผู้อ่านข่าวออนไลน์ที่ลงทะเบียนเข้าใช้งานระบบแพลตฟอร์มไทยรัฐ ซึ่งข้อมูลที่เก็บนั้นประกอบด้วยประวัติของผู้อ่านข่าวออนไลน์ซึ่งจะเป็นตัวบอกคุณลักษณะของผู้อ่านข่าวออนไลน์ และพฤติกรรมการคลิกอ่านข่าวของผู้อ่านข่าวออนไลน์คนนั้นๆ โดยกลุ่มข้อมูลที่ศึกษาจะอยู่ในช่วงเดือนสิงหาคม 2561

#### **2) การเตรียมข้อมูล**

##### **2.1) การทำความสะอาดข้อมูล**

จากฐานข้อมูลที่เก็บได้จริงทั้งหมด จำนวน 50,000,000 แถว (ห้าสิบล้านแถว) เมื่อผ่านขั้นตอนการทำความสะอาดข้อมูล และกำจัดข้อมูลขาดหายแล้ว เหลือข้อมูลที่

น่าสนใจที่ผู้วิจัยจะสามารถนำไปทดสอบทดลองต่อได้ เป็นจำนวน 91,234 แถว หลังจากนั้นจึงได้ทำการวิเคราะห์ความเกี่ยวข้องของข้อมูล พบว่ามีข้อมูลจำนวนมากที่มาจากพฤติกรรมของผู้อ่านข่าวออนไลน์นี้ไม่คลิกอ่านข่าวอย่างต่อเนื่องในช่วงเวลา 1 session ทำให้ความต่อเนื่องของข้อมูลขาดหายไปไม่สามารถนำมาทำการทดลองได้ ทำให้เหลือมีข้อมูลที่สามารถใช้งานได้เพียง 91,234 แถว จากนั้นจะถูกนำไปแปลงข้อมูลจากข้อมูลดิบ ดังตารางที่ 3.1 เพื่อทำข้อมูลให้อยู่ในรูปแบบโครงสร้าง ซึ่งจะอธิบายต่อในหัวข้อ 2.2

Attribute	ข้อมูลดิบ
audienceld	58ec4866d67ab41b600072a7
trafficId	5b7cab2e1829e032294346ac
sessionId	d5nuk79dqj82fs0hbnmr6vc9t6
trafficType	hit
entityId	908066
Previous Referer	https://www.thairath.co.th/tags/%E0%B8%95%E0%B9%88%E0%B8%B2%E0%B8%87%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8
Referer	android-app://com.google.android.googlequicksearchbox/https/www.google.com
memberId	5b44a81c6cfab32c746a56d9
Birthday	1 มกราคม 2530
Age	No data
Province	48
RegionId	3
postCode	32000
Gender	male
ContentId	912101
Topic	/entertain/news

title	ใจร้าย! ย้อนรอยทัวร์แสบจัดทริปลุยนอกलयแพนิกเที่ยวปาดหน้าตา คาสนามบิน
OS Type	Mozilla/5.0 (iPhone; CPU iPhone OS 11_4_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/11.0 Mobile/15E148 Safari/604.1

### ตารางที่ 3.1 ตัวอย่างข้อมูลดิบของผู้อ่านข่าวออนไลน์และบทความข่าว

#### 2.2) การแปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน (Data transformation)

เมื่อได้ข้อมูลที่ได้จากการทำความสะอาดแล้วและเหลือเฉพาะข้อมูลที่น่ามาใช้งานได้ ขั้นตอนต่อไปคือการแปลงข้อมูลให้อยู่ในรูปแบบที่เหมาะสมที่จะนำไปเข้า Process โมเดลพยากรณ์ ตัวอย่างเช่น มีข้อมูล Previous Referer ซึ่งจะเป็นการบอกว่า เพจก่อนหน้าที่ผู้อ่านข่าวออนไลน์จะเข้ามาอ่านข่าวนั้นอยู่ที่ไหนซึ่งข้อมูลที่ได้คือ

<https://www.thairath.co.th/tags/%E0%B8%95%E0%B9%88%E0%B8%B2%E0%B8>

[/%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8](https://www.thairath.co.th/tags/%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8)

ซึ่งลิงก์ที่ได้นั้นเป็นลิงก์ที่ยาวเกินไปและมีข้อมูลที่ไม่สามารถนำไปวิเคราะห์ต่อได้ จึง

ต้องแปลงข้อมูลนั้นให้เป็น [www.thairath.co.th](http://www.thairath.co.th) หรือ ข้อมูล Birthday ซึ่งเป็นข้อมูลวันเดือนปีเกิด

จะต้องแปลงข้อมูลให้เป็นตัวเลข 2 หลักเพื่อให้นำไปสร้างโมเดลต่อได้ เป็นต้น ตัวอย่างการแปลงข้อมูล ดังตารางที่ 3.2 และเมื่อแปลงข้อมูลแล้วจะได้ข้อมูลที่มีประโยชน์ต่อการนำไปสร้างโมเดล ดังภาพประกอบที่ 3.3

Attribute	ข้อมูลดิบ	ข้อมูลที่แปลงแล้ว
audienceld	58ec4866d67ab41b600072a7	
trafficld	5b7cab2e1829e032294346ac	
sessionId	d5nuk79dqj82fs0hbnmr6vc9t6	

trafficType	hit	
entityId	908066	
Previous-Referer	https://www.thairath.co.th/tags/%E0%B8%95%E0%B9%88%E0%B8%B2%E0%B8%87%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8	www.thairath.co.th
Referer	android-app://com.google.android.googlequicksearchbox/https/www.google.com	www.google.com
memberId	5b44a81c6cfab32c746a56d9	
Birthday	1 มกราคม 2530	31
Age	No data	31
Province	48	จังหวัดลำปาง
RegionId	3	ภาคเหนือ
postCode	52100	รหัสไปรษณีย์จังหวัดลำปาง
Gender	male	ชาย
ContentId	912101	
Topic	/entertain/news	entertainnews
title	ใจร้าย! ——— ย้อนรอยทัวร์แสวง จัดทริปลุยนอกलयแพนก์เที่ยวภาค หน้าตากาสนามบิน	

OS Type	Mozilla/5.0 (iPhone; CPU iPhone OS 11_4_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/11.0 Mobile/15E148 Safari/604.1	iOS
---------	---	-----

**ตารางที่ 3.2 ตัวอย่างการแปลงข้อมูลให้เหมาะสมกับการใช้งาน  
และการตัดข้อมูลที่ไม่จำเป็นออก**

memberId	gender	age	provinceld	regionId	OS Type	weekday	hour	referrer	section	topic	ContentId
5b469cd47aa83678885e4290	male	32	0	0	Android	2	19	http://m.facebook.com/	entertain	entertainnews	/content/1363855
5b469cd47aa83678885e4290	male	32	0	0	Android	2	20	http://m.facebook.com/	news	society	/content/1363933
5b457f997aa8367e1a680a	male	32	0	0	Android	3	19	https://www.thairath.co.th/	news	crime	/content/1359949
5b457f997aa8367e1a680a	male	32	0	0	Android	7	20		news	society	/content/1362521
5b457f997aa8367e1a680a	male	32	0	0	Android	1	12		entertain	entertainnews	/content/1362033
5b46ed557aa836722e13167f	male	34	0	0	Android	7	16	http://m.facebook.com	news	society	/content/1362426
5b46ed557aa836722e13167f	male	34	0	0	Android	7	16	http://m.facebook.com	entertain	entertainnews	/content/1360715
5b38a8d12f776852100d554e	male	35	0	0	Android	1	21	https://www.thairath.co.th/	entertain	entertainnews	/content/1358286
5b38a8d12f776852100d554e	male	35	0	0	Android	1	21	https://www.thairath.co.th/	news	politic	/content/1358657

ภาพประกอบที่ 3.3 ตัวอย่างข้อมูลที่ผ่านการแปลงแล้ว

ตารางที่ 3.3 จะเป็นตารางที่แสดงข้อมูลที่แปลงแล้วและสามารถนำไปใช้ในการสร้างโมเดล ซึ่งมีทั้งหมด 12 คุณลักษณะ คือ memberId, Sex, Age, Provinceld, Region, OS Type, Weekday, hour, referrer, Section, Topic และ ContentId

ชื่อ	ความหมาย	ค่าของตัวแปร
memberId	Primary Key	เป็น ID ยืนยันตัวตนของผู้ใช้งานว่าเป็นคนเดิมหรือไม่
Sex	เพศ	แบ่งออกเป็น 2 กลุ่ม คือ - เพศชาย - เพศหญิง
Age	อายุ	อายุ
Provinceld	จังหวัด	แบ่งออกเป็น 78 กลุ่ม คือ 0 = ไม่ทราบจังหวัด 1 = กระบี่ 2 = กรุงเทพฯ



		3 = กาญจนบุรี ... ฯลฯ
Region	ภูมิภาค	แบ่งออกเป็น 6 กลุ่ม คือ 0 = ไม่ทราบภูมิภาค 1 = ภาคเหนือ 2 = ภาคใต้ 3 = ภาคกลาง 4 = ภาคตะวันออกถึงเหนือ 5 = ภาคตะวันออก

ชื่อ	ความหมาย	ค่าของตัวแปร
OS Type	ระบบปฏิบัติการที่ใช้งาน	แบ่งออกเป็น 8 กลุ่ม คือ - Android - Chrome OS - iOS - Linux - Mac OS - Ubuntu - Windows
Weekday	7 วันใน 1 สัปดาห์	แบ่งออกเป็น 7 กลุ่ม คือ - อาทิตย์ - จันทร์ - อังคาร - พุธ - พฤหัสบดี - ศุกร์ - เสาร์

hour	ช่วงเวลาที่เข้าแพลตฟอร์ม	แบ่งออกเป็น 24 กลุ่ม คือ แยกรายชั่วโมง ตั้งแต่ 0.00 – 23.59 น.
referrer	แหล่งที่มา	แพลตฟอร์มที่ใช้งานอยู่ก่อนเข้ามา อ่านข่าวในแพลตฟอร์มไทยรัฐ

ชื่อ	ความหมาย	ค่าของตัวแปร
Section	หมวดข่าว	หมวดข่าวใหญ่ที่เลือกใช้ แบ่ง ออกเป็น 2 กลุ่ม คือ 1. บันเทิง 2. ข่าว
Topic	หัวข้อข่าว	แบ่งออกเป็น 5 กลุ่ม คือ 1. ต่างประเทศ 2. การเมือง 3. สังคม 4. อาชญากรรม 5. บันเทิง
ContentId	ID ของข่าว	แยกตาม กลุ่มของหัวข้อข่าว คือ 1. ต่างประเทศ 2. การเมือง 3. สังคม 4. อาชญากรรม 5. บันเทิง

ตารางที่ 3.3 แสดงรายละเอียดการแปลงข้อมูล

### 2.3) การจัดการกับข้อมูลสูญหาย (Handling Missing Data)

ข้อมูลสูญหายเป็นปัญหาในการวิเคราะห์ข้อมูล โดยเฉพาะอย่างยิ่งในฐานข้อมูลที่มีข้อมูลจำนวนมาก ในการทดลองนี้ผู้วิจัยเลือกวิธีการจัดการกับข้อมูลสูญหาย (Handling Missing Data) ด้วยวิธีการตัดระเบียน ที่มีการสูญหายออกจากการวิเคราะห์ เนื่องจากรูปแบบของข้อมูลสูญหายอาจจะนำข้อมูลไปสู่ความเอนเอียง (bias) ได้ นอกจากนี้จะสูญเสียข้อมูลอันเป็นประโยชน์อื่นในขอบเขตข้อมูลอื่นๆ ทั้งหมดอีกด้วย (สายชล สันสมบัติทอง, 2558 : 37)

ตัวอย่างดังตารางที่ 3.6 แสดงตัวอย่างข้อมูลที่มีการสูญหาย ไม่สามารถนำไปใช้งานได้

Attribute	ข้อมูลดิบ
audienceld	58ec4866d67ab41b600072a7
trafficld	5b7cab2e1829e032294346ac
sessionId	d5nuk79dqj82fs0hbnmr6vc9t6
trafficType	hit
entityId	908066
Previous Referer	<a href="https://www.thairath.co.th/tags/%E0%B8%95%E0%B9%88%E0%B8%B2%E0%B8%87%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8">https://www.thairath.co.th/tags/%E0%B8%95%E0%B9%88%E0%B8%B2%E0%B8%87%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%97%E0%B8%A8</a>
Referer	android-app://com.google.android.googlequicksearchbox/https/www.google.com
memberId	5b44a81c6cfab32c746a56d9
Birthday	No data
Age	No data
Province	48
RegionId	3
postCode	52100
Education	มหาวิทยาลัยเชียงใหม่

Gender	male
ContentId	912101
Topic	No data
title	ใจร้าย! ย้อนรอยทัวร์แสวงจัดทริปลุยนอกलयแพนก์เที่ยวปาดหน้าดา คาสนามบิน
OS Type	Mozilla/5.0 (iPhone; CPU iPhone OS 11_4_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/11.0 Mobile/15E148 Safari/604.1

ตารางที่ 3.6 แสดงตัวอย่างของข้อมูลสูญหาย ถูกตัดทิ้งด้วยวิธีการตัดระเบียบ

### 3.3 การพัฒนาระบบ

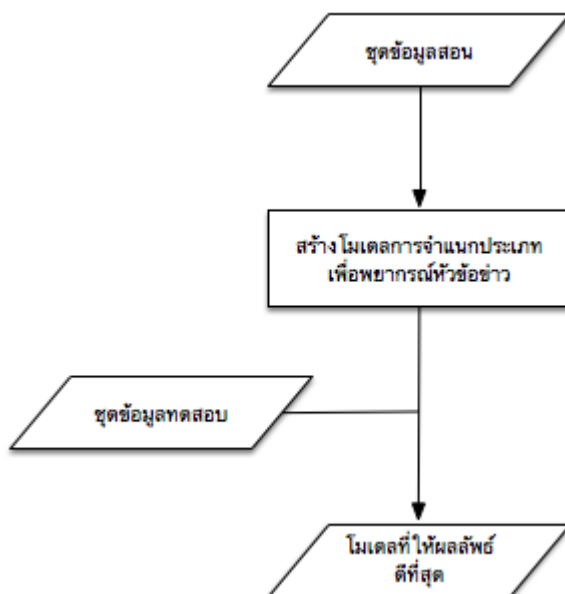
จากภาพประกอบที่ 3.3 แสดงสถาปัตยกรรมระบบ แสดงให้เห็นว่าการพัฒนาระบบจะประกอบด้วย 2 ขั้นตอน มีรายละเอียดของขั้นตอน ดังนี้

- สร้างโมเดล เพื่อพยากรณ์หัวข้อข่าว
- สร้างฟังก์ชันการเลือกบทความข่าวที่โมเดลพยากรณ์มาให้

#### 3.3.1 สร้างโมเดลเพื่อพยากรณ์หัวข้อข่าว

ข้อมูลที่ลงทะเบียนเข้ามาในระบบคือข่าวที่ผู้อ่านข่าวออนไลน์ข่าวในขณะนั้น (หัวข้อข่าว 1) และประวัติของผู้อ่านข่าวออนไลน์ ข้อมูลจะถูกแยกออกเป็น 2 ชุด คือชุดข้อมูลสอน (Training Data) และชุดข้อมูลทดสอบ (Testing Data)

นำชุดข้อมูลสอน (Training Data) มาเข้าสู่กระบวนการทำเหมืองข้อมูล เพื่อสร้างโมเดลในการพยากรณ์หาข่าวที่อยู่ในหัวข้อข่าว โดยจะทำการสร้างหลาย ๆ โมเดล เพื่อให้ทราบว่ามีโมเดลที่ถูกสร้างขึ้นนั้น โมเดลใดมีประสิทธิภาพและให้ผลลัพธ์ออกมาดีที่สุด และวัดความถูกต้องของแต่ละโมเดลด้วยการป้อนข้อมูลชุดข้อมูลทดสอบ (Testing Data) เข้าไปทดสอบแต่ละโมเดล ดังภาพประกอบ 3.4



ภาพประกอบที่ 3.4 แสดงขั้นตอนการสร้างโมเดลเพื่อพยากรณ์หัวข้อข่าว

### 3.3.2 สร้างฟังก์ชันการเลือกข่าว (News Selection Function) ตามหัวข้อข่าวที่โมเดลพยากรณ์ให้

นำผลลัพธ์ของโมเดลจากการพยากรณ์หัวข้อข่าวที่ดีที่สุด จากภาพประกอบ 3.4 มาป้อนเข้าสู่กระบวนการฟังก์ชันการเลือกข่าว โดยจะเลือกใช้ 2 วิธีคือ ข่าวที่มีคนอ่านมากที่สุด และข่าวล่าสุด ซึ่งมีข้อดี และข้อเสีย ต่างกันในแต่ละวิธี ดังนี้

#### ข่าวที่มีผู้อ่านมากที่สุด

##### ข้อดี

- เป็นประเด็นที่มีคนกลุ่มใหญ่ให้ความสนใจ
- เนื้อหามีการปรับปรุงและเพิ่มเติมเนื้อหา โดยกองบรรณาธิการข่าวให้มีความ

ครบถ้วนสมบูรณ์

##### ข้อเสีย

- ข่าวใหม่จะไม่มีโอกาสถูกแนะนำเพราะยังมีผู้อ่านข่าวออนไลน์ไม่มากนัก ทำให้เกิดปัญหา Cold Start นั่นคือภาวะของข้อมูลที่น้อยเกินไปหรือไม่เพียงพอต่อการทำการทดลอง เพราะไม่มีผู้คลิกอ่าน

- บทความข่าวอาจไม่ทันต่อเหตุการณ์ หรือเป็นข่าวเก่า

### ข่าวล่าสุด

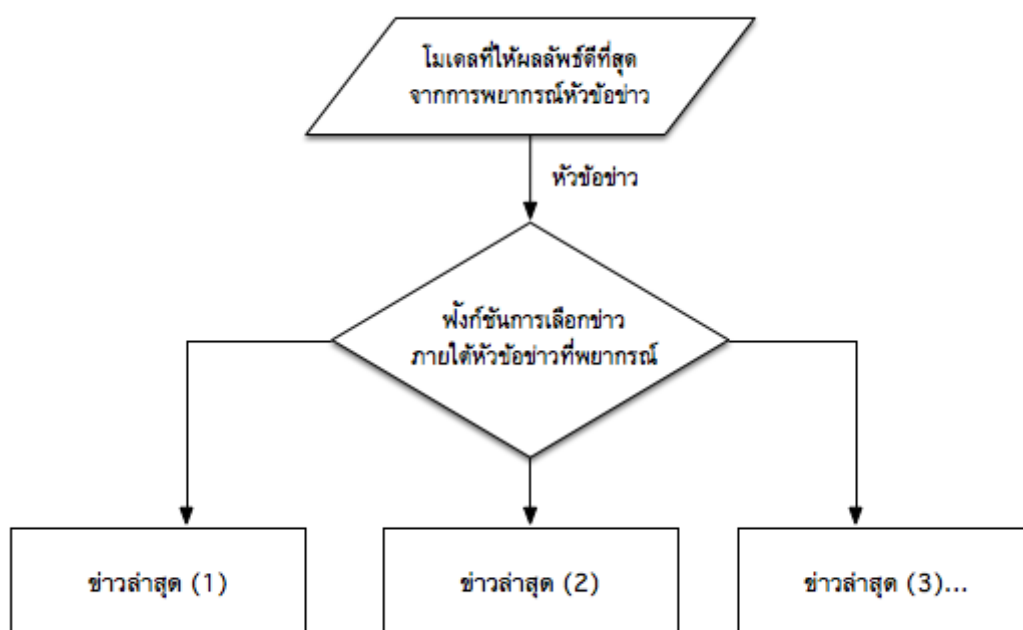
#### ข้อดี

- ข่าวสดใหม่ ทันต่อเหตุการณ์
- ผู้คนอาจให้ความสนใจ
- แก้ปัญหา Cold Start ที่เกิดกับการเลือกข่าวแบบ “ข่าวที่มีผู้อ่านข่าวออนไลน์มากที่สุด” เนื่องจากข่าวล่าสุดเป็นข่าวที่สดใหม่และยังมิมีผู้คลิกอ่านข่าวน้อย การแนะนำข่าวล่าสุดให้ผู้อ่านข่าวออนไลน์ได้เห็น เพื่อให้ผู้อ่านข่าวออนไลน์ได้คลิกอ่านต่อไปจึงสามารถช่วยแก้ปัญหานี้ได้

#### ข้อเสีย

- เนื่องจากการนำเสนอข่าวที่รวดเร็ว ความถูกต้องครบถ้วนของข่าวอาจยังไม่เพียงพอ

เมื่อผ่านกระบวนการฟังก์ชันการเลือกข่าวแล้วจึงจะแสดงผลออกให้แก่ผู้อ่านข่าวออนไลน์ ดังภาพประกอบที่ 3.5



### ภาพประกอบที่ 3.5 ฟังก์ชันเพื่อเลือกข่าว

#### 3.4 การวัดประสิทธิภาพของระบบ

ในการพัฒนาระบบ ตัวชี้วัดว่าระบบที่พัฒนานั้นมีประสิทธิภาพหรือไม่ ผู้วิจัยเลือกใช้วิธีการวัดอัตราการคลิก (Measurement Click Rate) โดยใช้เครื่องมือวัดผลคือ A/B Testing เพื่อทดสอบว่าแพลตฟอร์มเดิมที่แนะนำบทความข่าวด้วยการใช้ระบบแนะนำข่าวเชิงพาณิชย์ “C” เมื่อเทียบกับแพลตฟอร์มที่แนะนำบทความข่าวด้วยโมเดลการแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าวนั้น แบบใดให้ผลลัพธ์ออกมาดีกว่ากัน

##### ขั้นตอนการทำ A/B Testing

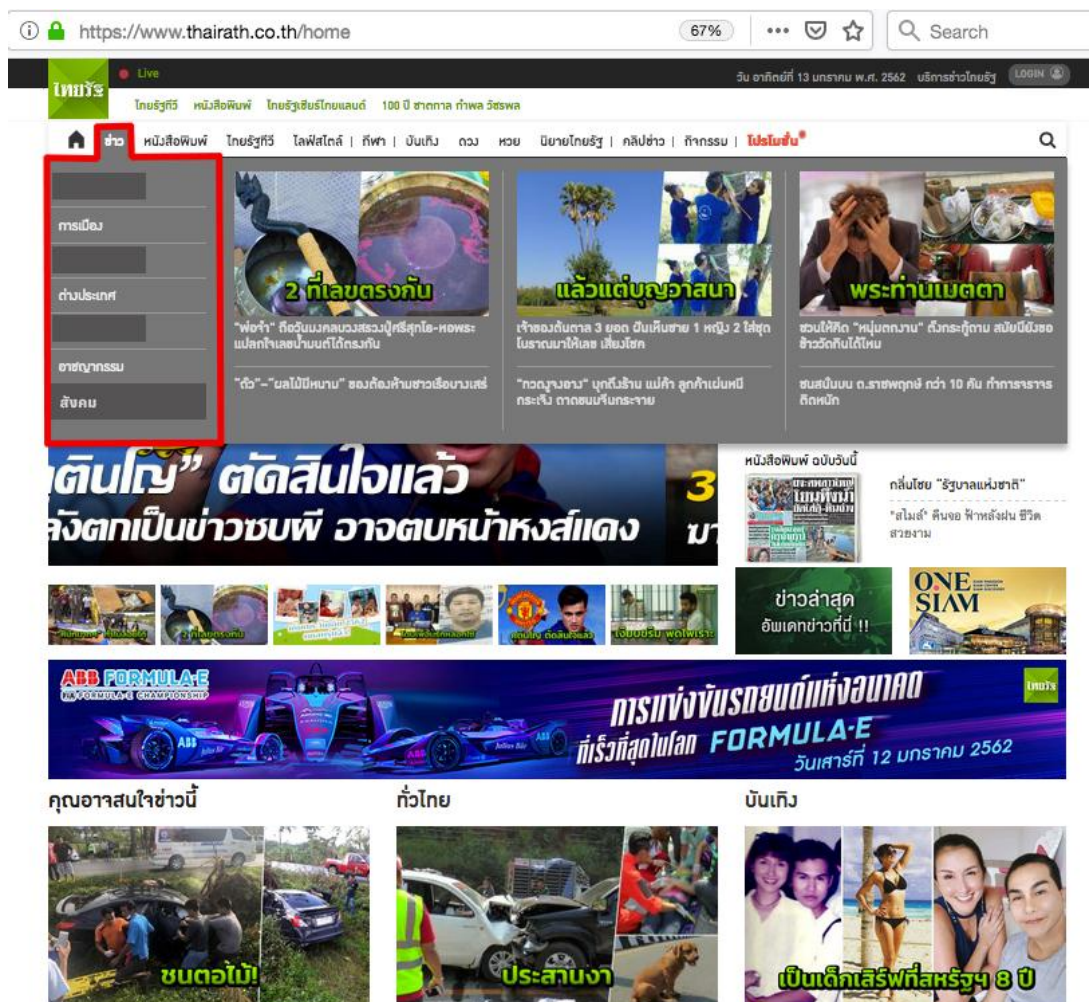
1. ศึกษาข้อมูลแพลตฟอร์ม โดยใช้ Google Analytics เพื่อหาปัญหา เช่น ผู้อ่านข่าวออนไลน์แพลตฟอร์มเข้ามาอ่านข่าวแล้วออกจากแพลตฟอร์มไปโดยไม่ดูหน้าอื่นต่อ
2. การตั้งสมมติฐาน โดยสมมติฐานของผู้วิจัยคือ บทความข่าวที่แนะนำโดยโมเดลการแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าว จะทำให้อัตราการคลิกอ่านข่าวสูงขึ้น
3. ทดสอบสมมติฐาน โดยการสร้างตัวแปรที่สอดคล้องกับสมมติฐานจากข้อ 2 โดยเทียบกับแพลตฟอร์มปัจจุบัน เช่น ตัวแปร A เป็นแพลตฟอร์มปัจจุบันที่ไม่มีการเปลี่ยนแปลงและมีการใช้ระบบแนะนำข่าวเชิงพาณิชย์ และตัวแปร B ในที่นี้ผู้วิจัยเลือกให้ตัวแปร B เป็นแพลตฟอร์มที่มีระบบแนะนำข่าวรายบุคคล โดยเลือกใช้โมเดลที่ให้ผลลัพธ์ที่ดีที่สุดนำมาทดสอบ
4. ดูผลลัพธ์จากการสร้างตัวแปร A/B test โดยส่งกลุ่มทดสอบยังหน้าแพลตฟอร์มที่เป็นตัวแปร A และ B จำนวนหนึ่ง แล้วนับอัตราการคลิกอ่านต่อของ A และ B และเอามาเปรียบเทียบกัน และทำการวัดอัตราการคลิก



ภาพประกอบ 3.6 แสดงพื้นที่การลงทะเบียนด้วยการ Login เข้าแพลตฟอร์ม

ภาพประกอบ 3.6 ผู้วิจัยจะใช้ข้อมูลที่นำมาทำการทดลองจากข้อมูลของผู้อ่านข่าวออนไลน์  
ข่าวที่เข้ามาผ่านแพลตฟอร์มและมีการลงทะเบียนด้วยการ Login





ภาพประกอบ 3.7 แสดงพื้นที่ หัวข้อข่าว ข่าวบนแพลตฟอร์ม คือ การเมือง, ต่างประเทศ, อาชญากรรม และสังคม



ภาพประกอบ 3.8 แสดงพื้นที่ หัวข้อข่าว ข่าวบนแพลตฟอร์ม คือ บันเทิง

ภาพประกอบที่ 3.7 และ 3.8 แสดงพื้นที่หน้าแพลตฟอร์มที่ผู้วิจัยเลือกทำการวัดอัตราการคลิกด้วยวิธี A/B Testing โดยการเลือกหัวข้อข่าวบนแพลตฟอร์มที่มาทำการทดลอง 5 หัวข้อข่าวซึ่งเป็น 5 หัวข้อข่าวที่อัตราการเข้าชมสูงที่สุด คือ การเมือง, ต่างประเทศ, อาชญากรรม, สังคม และบันเทิง



ภาพประกอบ 3.9 เปรียบเทียบหน้าข่าว A กับ หน้าข่าว B บนแพลตฟอร์ม

ภาพประกอบ 3.9 เป็นการเปรียบเทียบข่าวที่แนะนำจากระบบแนะนำข่าว 2 ระบบ โดย A คือข่าวที่แนะนำโดยระบบแนะนำข่าวเชิงพาณิชย์ซึ่งใช้กับแพลตฟอร์มเดิม และ B คือข่าวที่แนะนำโดยระบบแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าว หลังจากนั้นจึงทำการวัดผลด้วยอัตราการคลิกและนำไปทดสอบสมมติฐานต่อไป

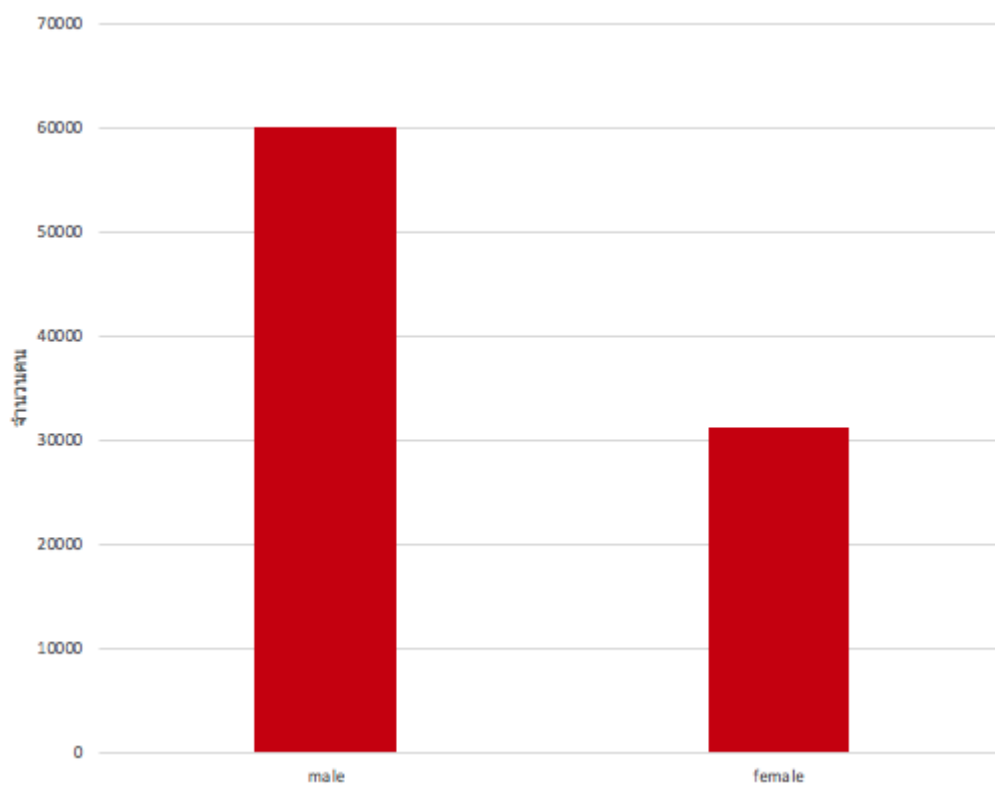
## บทที่ 4

### ผลการวิจัย

การวิจัยเรื่องระบบเพื่อแนะนำหัวข้อข่าวโดยใช้ประวัติและพฤติกรรมผู้อ่านข่าวออนไลน์ ซึ่งข้อมูลที่เก็บนั้นประกอบด้วย 1) ประวัติของผู้อ่านข่าวออนไลน์ซึ่งจะเป็นตัวบอกคุณลักษณะของผู้อ่านข่าวออนไลน์ และ 2) พฤติกรรมการคลิกอ่านข่าวของผู้อ่านข่าวออนไลน์คนนั้นๆ โดยข้อมูลที่เก็บได้จริงทั้งหมด จำนวน 50,000,000 แถว (ห้าสิบล้านแถว) เมื่อผ่านขั้นตอนการทำความสะอาดข้อมูล และกำจัดข้อมูลสูญหายแล้ว เหลือข้อมูลที่น่าสนใจที่ผู้วิจัยจะสามารถนำไปทำการทดลองต่อได้ เป็นจำนวน 91,234 แถว โดยผู้วิจัยสร้างระบบแนะนำข่าวโดยใช้เทคนิคการกรองแบบผสม (Hybrid) กับเทคนิคการจำแนกข้อมูล (Classification) โดยวิเคราะห์และเปรียบเทียบผลการทดลองโมเดลด้วยอัลกอริทึม ต้นไม้ตัดสินใจ (Decision Tree) , การถดถอยเชิงโลจิสติก (Logistic Regression) , ซัพพอร์ตเวกเตอร์แมชชีน (SVM) และโครงข่ายประสาทเทียม (Artificial Neural Network - ANN) เพื่อนำอัลกอริทึมที่ให้ผลลัพธ์ออกมาดีที่สุดนำมาใช้ในการพัฒนาฟังก์ชันการเลือกข่าวและวัดประสิทธิภาพของระบบ ในบทนี้ผู้วิจัยขอเสนอผลการวิเคราะห์ 3 หัวข้อ ดังนี้

- 4.1 การสำรวจข้อมูลเบื้องต้น
- 4.2 การพัฒนาระบบ
  - 4.2.1 การพยากรณ์หัวข้อข่าว
  - 4.2.2 ฟังก์ชันการเลือกข่าว
- 4.3 การนำโมเดลไปใช้งานจริง

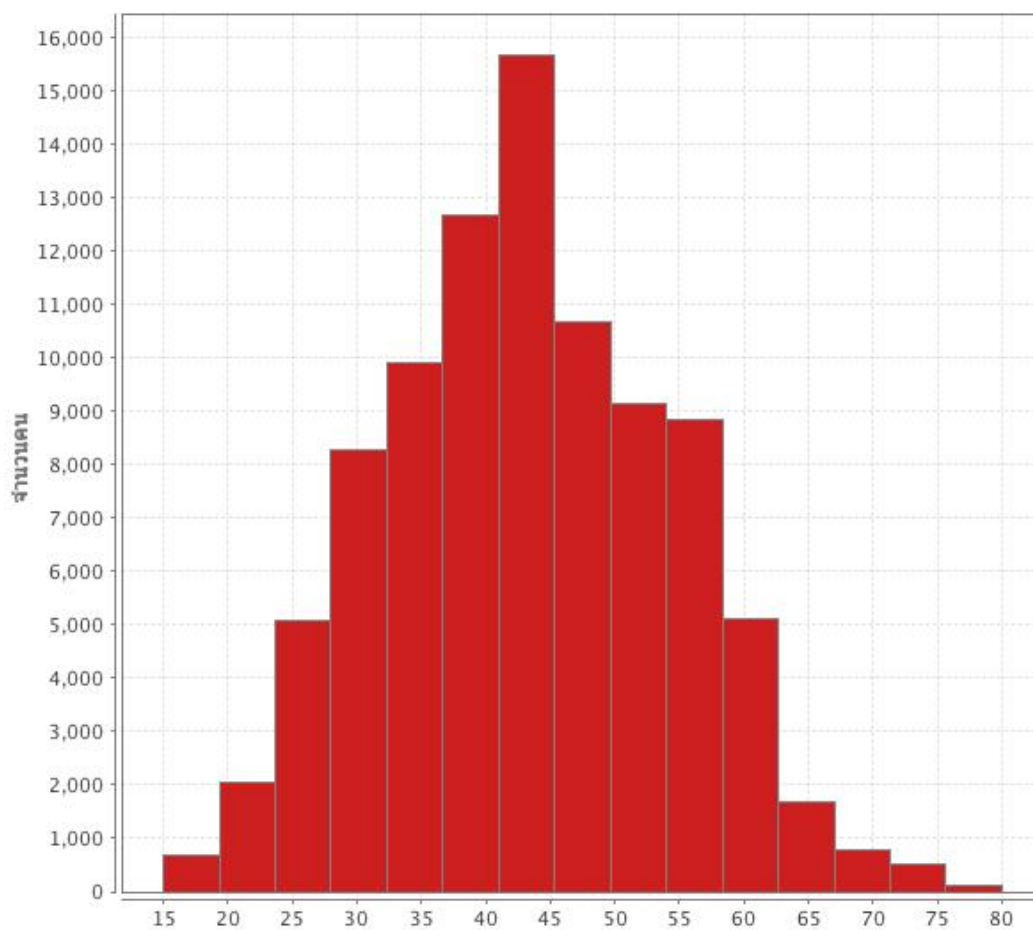
#### 4.1 การสำรวจข้อมูลเบื้องต้น



ภาพประกอบ 4.1 แสดงข้อมูลเพศ

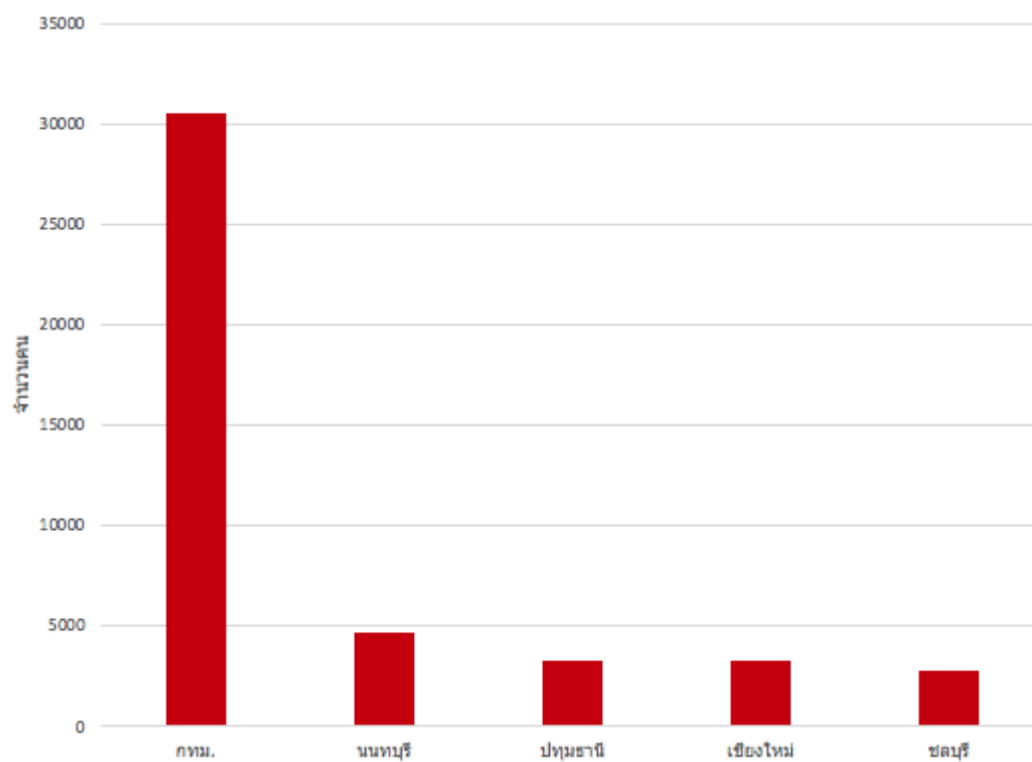
จากภาพประกอบ 4.1 จำนวนผู้ลงทะเบียนเข้าแพลตฟอร์ม จำนวน 91,234 แถว โดยแบ่งเป็นเพศหญิง จำนวน 31,169 แถว และเพศชาย จำนวน 60,074 แถว





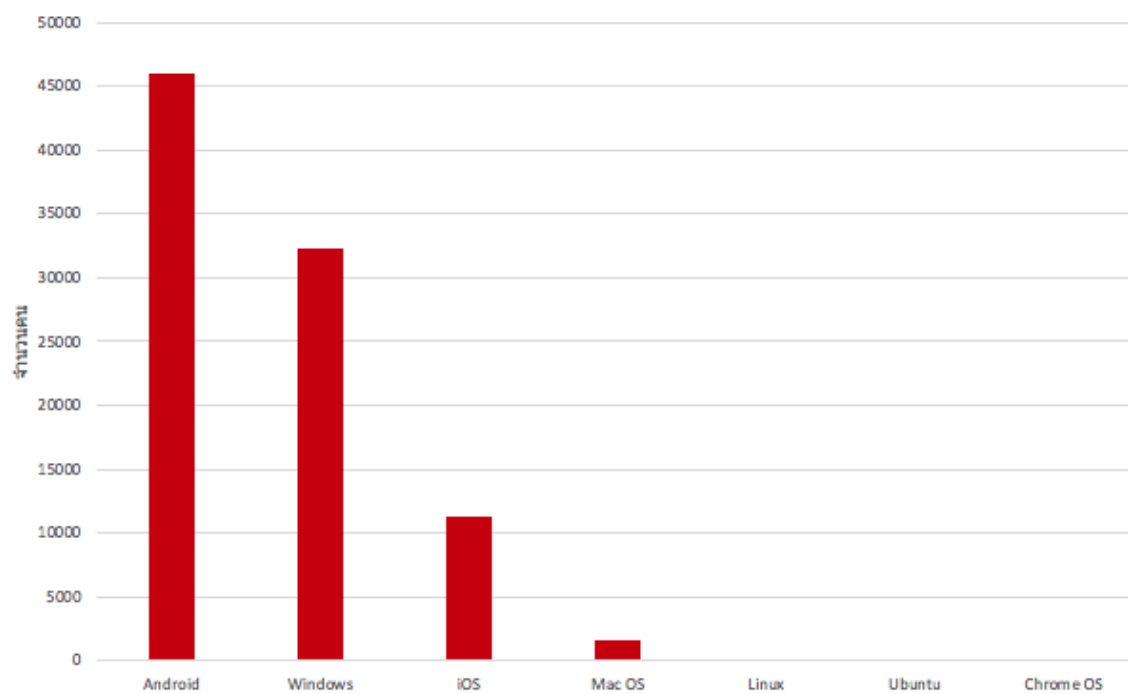
ภาพประกอบ 4.2 แสดงข้อมูลอายุ

จากภาพประกอบ 4.2 เป็นข้อมูลอายุของผู้ลงทะเบียนเข้าแพลตฟอร์ม ซึ่งผู้วิจัยทำการตัดข้อมูลขาดหายออกและคัดเลือกข้อมูลที่เหมาะสม โดยมีอายุตั้งแต่ 15 – 80 ปี โดยมีค่าเฉลี่ยอายุอยู่ที่ 42.97 ปี



ภาพประกอบ 4.3 แสดงข้อมูลจังหวัด

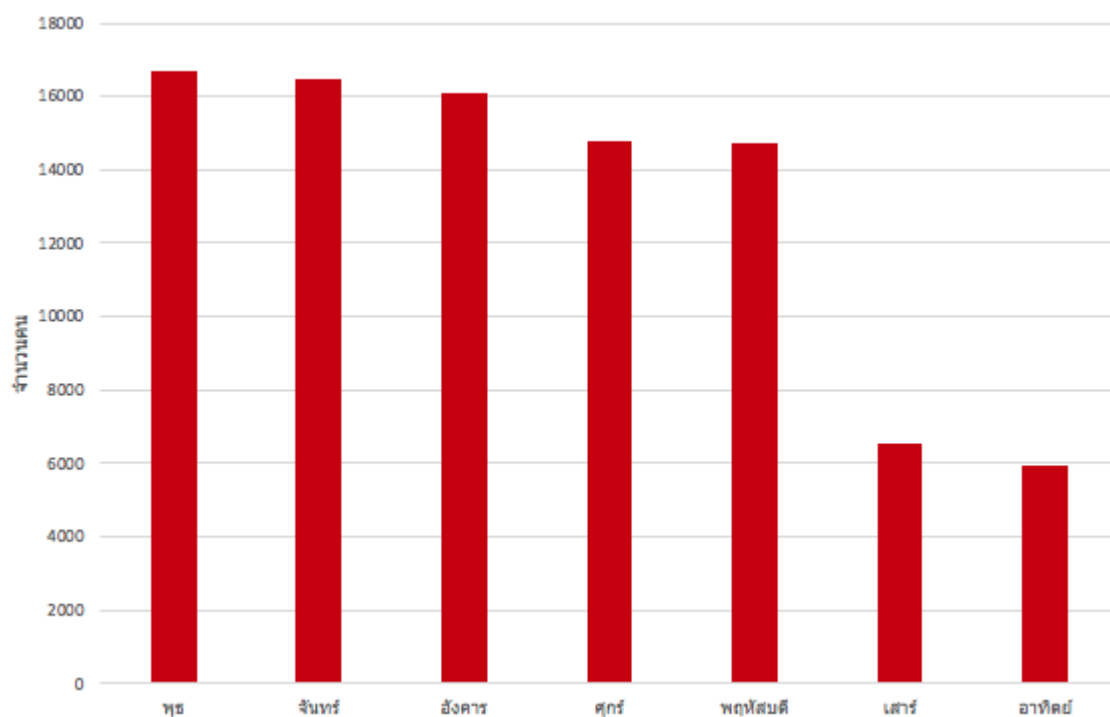
จากภาพประกอบ 4.3 เป็นข้อมูลจังหวัดของผู้มีที่ลงทะเบียนเข้าแพลตฟอร์ม 5 อันดับแรก คือ กรุงเทพมหานคร นนทบุรี ปทุมธานี เชียงใหม่ และชลบุรี



ภาพประกอบ 4.4 แสดงข้อมูลระบบปฏิบัติการ

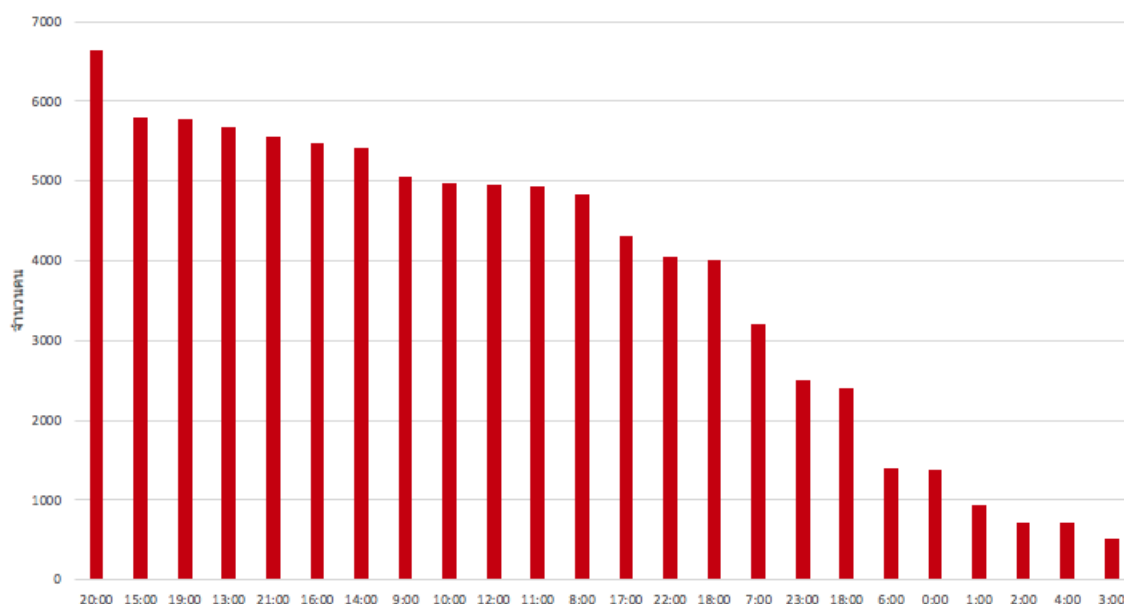
จากภาพประกอบ 4.4 ข้อมูลของระบบปฏิบัติการที่ใช้งานของผู้ลงทะเบียนเข้าแพลตฟอร์มสูงที่สุดคือ Android รองลงมาคือ Windows, iOS , Mac OS





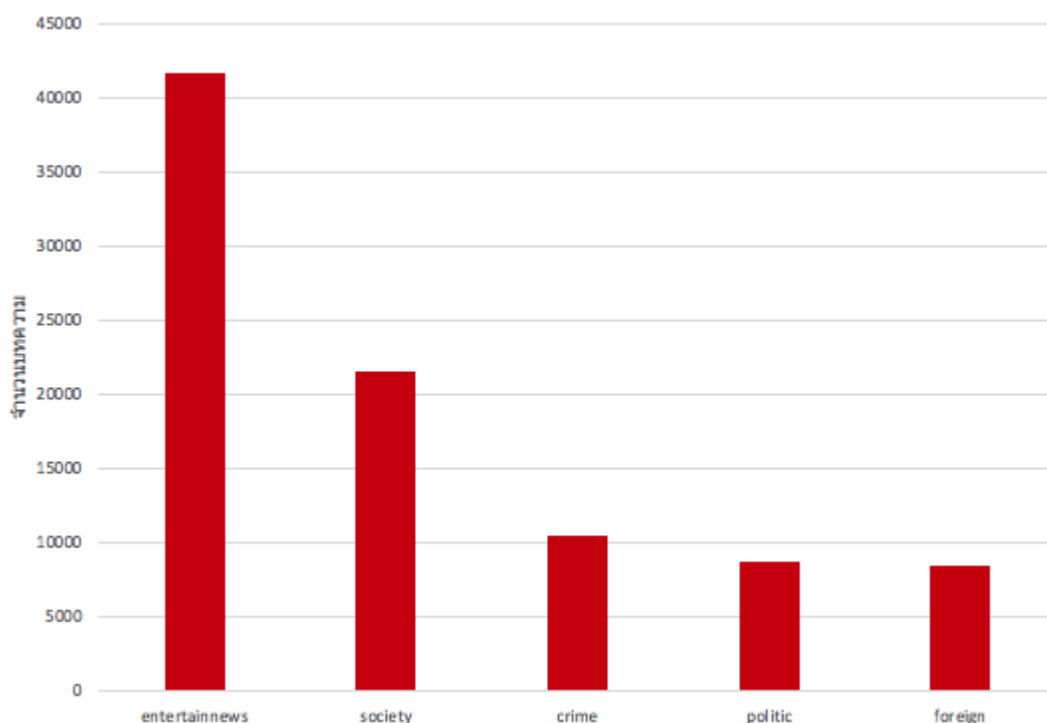
ภาพประกอบ 4.5 แสดงข้อมูลการผู้อ่านข่าวในหนึ่งสัปดาห์

จากภาพประกอบ 4.5 เป็นข้อมูลของผู้ลงทะเบียนเข้าแพลตฟอร์มในหนึ่งสัปดาห์ โดยเรียงลำดับตามวันที่มีผู้อ่านข่าวมากที่สุด ดังนี้ วันพุธ, วันจันทร์, วันอังคาร, วันศุกร์, วันพฤหัสบดี, วันเสาร์ และวันอาทิตย์



ภาพประกอบ 4.6 แสดงข้อมูลผู้อ่านข่าวรายชั่วโมง

จากภาพประกอบ 4.6 แสดงข้อมูลผู้อ่านข่าวนับเป็นรายชั่วโมง โดยช่วงเวลาที่ผู้อ่านข่าวมากที่สุดคือ 20.00 น., ช่วงเวลา 15.00 น., และ 19.00 น. มีจำนวนผู้อ่านข่าวออนไลน์ข่าวใกล้เคียงกัน และช่วงเวลาที่ผู้อ่านข่าวออนไลน์ข่าวน้อยที่สุดคือ 03.00 น.



ภาพประกอบ 4.8 แสดงจำนวนหัวข้อข่าว

ภาพประกอบ 4.8 แสดงจำนวนของ หัวข้อข่าว ข่าวที่เก็บข้อมูลเพื่อนำมาทดสอบโมเดล 5 หัวข้อข่าว และเรียงลำดับผู้อ่านข่าวออนไลน์จากมากไปน้อย คือ บันเทิง, สังคม, อาชญากรรม, การเมือง และต่างประเทศ

#### การประมวลผลข้อมูล 6 แบบ การแบ่งชุดข้อมูลเพื่อนำไปทดสอบกับโมเดล

ผู้วิจัยได้ทำการประมวลผลข้อมูล เพื่อหารูปแบบการประมวลผลที่เหมาะสมกับการนำไปสร้างโมเดล โดยได้แบ่งย่อยออกไปอีกกว่าชุดของลักษณะข้อมูลใดที่สามารถที่สามารถแบ่งย่อยได้ในหลายระดับ ผู้วิจัยพบว่า เวลา และ จังหวัด สามารถแบ่งข้อมูลย่อยได้ โดย เวลา แบ่งรูปแบบข้อมูลออกเป็น ทุก 6 ชั่วโมง ทุกชั่วโมง และจังหวัด แบ่งรูปแบบข้อมูลออกโดยจัดกลุ่มแบ่งออกเป็นภูมิภาค และย่อยออกเป็นรายจังหวัด ทั้งหมด 77 จังหวัด เพื่อดูว่าการแบ่งย่อยข้อมูลออกเป็นกลุ่มใหญ่หรือแบบแบ่งย่อย ลักษณะจะให้ผลการทดลองออกมาดีกว่า ซึ่งสรุปการประมวลผลข้อมูลออกมาได้ 6 แบบ โดยการประมวลผลแต่ละแบบแยกรายละเอียดได้ดังนี้

##### แบบที่ 1

- ผู้ลงทะเบียนเข้าแพลตฟอร์ม เก็บข้อมูลเป็นรหัสแยกตามรายบุคคล (memberId)

- เพศ เก็บข้อมูลแยกเป็น เพศชาย (male) และเพศหญิง (female)
- การจัดข้อมูลผู้อ่านชาวออนไลน์ที่มีอายุ (age) ที่ต่ำกว่า 15 ปี และสูงกว่า 80 ปี ออกไป
- กลุ่มผู้อ่านชาวออนไลน์แยกรายจังหวัด โดยข้อมูลดิบเก็บเป็นรหัสจังหวัด (ProvinceId) และเรียงรหัสจังหวัดตามตัวอักษร โดยให้รหัสจังหวัด 0 คือ ข้อมูลของผู้อ่านชาวออนไลน์ที่ไม่ระบุจังหวัด, 1 คือ รหัสจังหวัดกระบี่, 2 คือรหัสจังหวัดกรุงเทพมหานคร เป็นต้น
- ระบบปฏิบัติการสำหรับการลงทะเบียนเข้าแพลตฟอร์ม เก็บข้อมูลแยกเป็นประเภท (OS Type) มี 7 ประเภทคือ Android, Windows, iOS , Mac OS, Linux, Ubuntu และ Chrome OS
- ผู้ลงทะเบียนเข้าแพลตฟอร์มในหนึ่งสัปดาห์ โดยเก็บเป็นรหัสแยกตามวัน (weekday) และเรียงรหัสตามวัน คือ 1 = วันจันทร์, 2 = วันอังคาร, 3 = วันพุธ, 4 = วันพฤหัสบดี, 5 = วันศุกร์, 6 = วันเสาร์ และ 7 = วันอาทิตย์
- เก็บข้อมูลผู้ลงทะเบียนเข้าแพลตฟอร์มแยกออกเป็นช่วงเวลา (timeperiod) โดยแบ่งออกเป็น 4 ช่วงเวลา คือ P1 คือ 00:00 – 05:59 น., P2 คือ 06:00 – 11:59 น., P3 คือ 12:00 – 17:59 น. และ P4 คือ 18:00 – 23:59 น.
- เก็บข้อมูลของผู้อ่านชาวออนไลน์ข่าวจากที่มาก่อนจะเข้าแพลตฟอร์ม (Referer) โดยที่มาที่เก็บได้ เช่น [www.google.com](http://www.google.com), [www.facebook.com](http://www.facebook.com) หรือผู้อ่านข่าวอยู่ในแพลตฟอร์มไทยรัฐอยู่แล้ว [www.thaiarath.co.th](http://www.thaiarath.co.th) เป็นต้น
- แยกข้อมูลข่าวออกเป็น 2 กลุ่มใหญ่ (section) คือ ข่าว (news) และบันเทิง (entertainment)
- กลุ่มหมวดข่าวย่อย (หัวข้อข่าว) ออกเป็น 5 กลุ่ม คือ ต่างประเทศ, การเมือง, สังคม, อาชญากรรม และบันเทิง

รายละเอียดดังตารางที่ 4.1

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
memberId	5b469cd47aa83678885e4290
Sex	- เพศชาย (male) - เพศหญิง (female)
Age	อายุ 15 – 80 ปี
ProvinceId	เรียงรหัสจังหวัดตามตัวอักษร โดย 0 = ไม่ระบุจังหวัด 1 = รหัสจังหวัดกระบี่ 2 = รหัสกรุงเทพมหานคร 3 = รหัสจังหวัดกาญจนบุรี ... ฯลฯ
OS Type	- Android - Chrome OS - iOS - Linux - Mac OS - Ubuntu - Windows
Weekday	- อาทิตย์ - จันทร์ - อังคาร - พุธ - พฤหัสบดี - ศุกร์ - เสาร์

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
Timeperiod	<ul style="list-style-type: none"> <li>- P1 คือ 00:00 – 05:59 น.</li> <li>- P2 คือ 06:00 – 11:59 น.</li> <li>- P3 คือ 12:00 – 17:59 น.</li> <li>- P4 คือ 18:00 – 23:59 น.</li> </ul>
Referer	<ul style="list-style-type: none"> <li>- www.google.com</li> <li>- www.facebook.com</li> <li>- www.thaiarath.co.th</li> </ul>
Section	<ul style="list-style-type: none"> <li>1. บันเทิง (entertainment)</li> <li>2. ข่าว (news)</li> </ul>
Topic	<ul style="list-style-type: none"> <li>1. ต่างประเทศ</li> <li>2. การเมือง</li> <li>3. สังคม</li> <li>4. อาชญากรรม</li> <li>5. บันเทิง</li> </ul>

ตาราง  
ที่

#### 4.1 การประมวลผลข้อมูลแบบที่ 1

### แบบที่ 2 แตกต่างจากแบบที่ 1 ดังนี้

- กลุ่มผู้อ่านข่าวออนไลน์แยกตามภูมิภาค โดยข้อมูลดิบเก็บเป็นรหัสตามภูมิภาค (RegionId) และเรียงรหัสภูมิภาค คือ 1 = ภาคเหนือ, 2 = ภาคใต้, 3 = ภาคกลาง, 4 = ภาคตะวันออกเฉียงเหนือ และ 5 = ภาคตะวันออก

### รายละเอียดดังตารางที่ 4.2

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
RegionId	ภูมิภาค แบ่งออกเป็น 6 กลุ่ม คือ 0 = ไม่ทราบภูมิภาค 1 = ภาคเหนือ 2 = ภาคใต้ 3 = ภาคกลาง 4 = ภาคตะวันออกเฉียงเหนือ 5 = ภาคตะวันออก

### ตารางที่ 4.2 การประมวลผลข้อมูลแบบที่ 2

### แบบที่ 3 แตกต่างจากแบบที่ 1 ดังนี้

- กลุ่มผู้อ่านข่าวออนไลน์แยกรายจังหวัด โดยข้อมูลดิบเก็บเป็นรหัสจังหวัด (ProvinceId) และเรียงรหัสจังหวัดตามตัวอักษร โดยให้รหัสจังหวัด 0 คือข้อมูลของผู้อ่านข่าวออนไลน์ที่ไม่ระบุจังหวัด, 1 คือรหัสจังหวัดกระบี่, 2 คือรหัสจังหวัดกรุงเทพมหานคร เป็นต้น
- กลุ่มผู้อ่านข่าวออนไลน์แยกตามภูมิภาค โดยข้อมูลดิบเก็บเป็นรหัสตามภูมิภาค (RegionId) และเรียงรหัสภูมิภาค คือ 1 = ภาคเหนือ, 2 = ภาคใต้, 3 = ภาคกลาง, 4 = ภาคตะวันออกเฉียงเหนือ และ 5 = ภาคตะวันออก

รายละเอียดดังตารางที่ 4.3

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
ProvinceId	เรียงรหัสจังหวัดตามตัวอักษร โดย 0 = ไม่ระบุจังหวัด 1 = รหัสจังหวัดกระบี่ 2 = รหัสกรุงเทพมหานคร 3 = รหัสจังหวัดกาญจนบุรี ...ฯลฯ
RegionId	แบ่งออกเป็น 6 กลุ่ม คือ 0 = ไม่ทราบภูมิภาค 1 = ภาคเหนือ 2 = ภาคใต้ 3 = ภาคกลาง 4 = ภาคตะวันออกเฉียงเหนือ 5 = ภาคตะวันออก

ตารางที่ 4.3 การประมวลผลข้อมูลแบบที่ 3

## แบบที่ 4 แตกต่างจากแบบที่ 1 ดังนี้

- กลุ่มผู้อ่านข่าวออนไลน์แยกรายจังหวัด โดยข้อมูลดิบเก็บเป็นรหัสจังหวัด (ProvinceId) และเรียงรหัสจังหวัดตามตัวอักษร โดยให้รหัสจังหวัด 0 คือข้อมูลของผู้อ่านข่าวออนไลน์ที่ไม่ระบุจังหวัด, 1 คือ รหัสจังหวัดกระบี่, 2 คือ รหัสจังหวัดกรุงเทพมหานคร เป็นต้น

- เก็บข้อมูลผู้ลงทะเบียนเข้าแพลตฟอร์ม 24 ชั่วโมง โดยเก็บแยกเป็นรายชั่วโมง (hour) โดยเริ่มตั้งแต่เวลา 01.00 น. ไปจนถึงเวลา 23:59 น.



รายละเอียดดังตารางที่ 4.4

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
ProvinceId	เรียงรหัสจังหวัดตามตัวอักษร โดย 0 = ไม่ระบุจังหวัด 1 = รหัสจังหวัดกระบี่ 2 = รหัสกรุงเทพมหานคร 3 = รหัสจังหวัดกาญจนบุรี ... ฯลฯ
Hour	แบ่งออกเป็น 24 กลุ่ม แยกรายชั่วโมง คือ 00:00 – 00:59 01:00 – 01:59 02:00 – 02:59 03:00 – 03:59 04:00 – 04:59 05:00 – 05:59 06:00 – 06:59 07:00 – 07:59 08:00 – 08:59 09:00 – 09:59 10:00 – 10:59 11:00 – 11:59 12:00 – 12:59 13:00 – 13:59 14:00 – 14:59 15:00 – 15:59 16:00 – 16:59 17:00 – 17:59

	18:00 – 18:59	ตาราง ที่
	19:00 – 19:59	
	20:00 – 20:59	
	21:00 – 21:59	
	22:00 – 22:59	
	23:00 – 23:59	

#### 4.4 การประมวลผลข้อมูลแบบที่ 4

แบบที่ 5 แตกต่างจากแบบที่ 1 ดังนี้

- กลุ่มผู้อ่านข่าวออนไลน์แยกตามภูมิภาค โดยข้อมูลดิบเก็บเป็นรหัสตามภูมิภาค (RegionId) และเรียงรหัสภูมิภาค คือ 1 = ภาคเหนือ, 2 = ภาคใต้, 3 = ภาคกลาง, 4 = ภาคตะวันออกเฉียงเหนือ และ 5 = ภาคตะวันออก
- เก็บข้อมูลผู้ลงทะเบียนเข้าแพลตฟอร์ม 24 ชั่วโมง โดยเก็บแยกเป็นรายชั่วโมง (hour) โดยเริ่มตั้งแต่เวลา 01.00 น. ไปจนถึงเวลา 23:59 น.

รายละเอียดดังตารางที่ 4.5

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
RegionId	แบ่งออกเป็น 6 กลุ่ม คือ 0 = ไม่ทราบภูมิภาค 1 = ภาคเหนือ 2 = ภาคใต้ 3 = ภาคกลาง 4 = ภาคตะวันออกเฉียงเหนือ 5 = ภาคตะวันออก
Hour	แบ่งออกเป็น 24 กลุ่ม แยกรายชั่วโมง คือ 00:00 – 00:59 01:00 – 01:59 02:00 – 02:59 03:00 – 03:59

04:00 – 04:59
05:00 – 05:59
06:00 – 06:59
07:00 – 07:59
08:00 – 08:59
09:00 – 09:59
10:00 – 10:59
11:00 – 11:59
12:00 – 12:59
13:00 – 13:59
14:00 – 14:59
15:00 – 15:59
16:00 – 16:59
17:00 – 17:59
18:00 – 18:59
19:00 – 19:59
20:00 – 20:59
21:00 – 21:59
22:00 – 22:59
23:00 – 23:59

ตาราง  
ที่

#### 4.5 การประมวลผลข้อมูลแบบที่ 5

แบบที่ 6 แตกต่างจากแบบที่ 1 ดังนี้

- กลุ่มผู้อ่านข่าวออนไลน์แยกรายจังหวัด โดยข้อมูลดิบเก็บเป็นรหัสจังหวัด (ProvinceId) และเรียงรหัสจังหวัดตามตัวอักษร โดยให้รหัสจังหวัด 0 คือข้อมูลของผู้อ่านข่าวออนไลน์ที่ไม่ระบุจังหวัด , 1 คือรหัสจังหวัดกระบี่ , 2 คือรหัสจังหวัดกรุงเทพมหานคร เป็นต้น
- กลุ่มผู้อ่านข่าวออนไลน์แยกตามภูมิภาค โดยข้อมูลดิบเก็บเป็นรหัสตามภูมิภาค (RegionId) และเรียงรหัสภูมิภาค คือ 1 = ภาคเหนือ , 2 = ภาคใต้ , 3 = ภาคกลาง , 4 = ภาคตะวันออกเฉียงเหนือ และ 5 = ภาคตะวันออก

- เก็บข้อมูลผู้ลงทะเบียนเข้าแพลตฟอร์ม 24 ชั่วโมง โดยเก็บแยกเป็นรายชั่วโมง (hour) โดยเริ่มตั้งแต่เวลา 01.00 น. ไปจนถึงเวลา 23:59 น.

รายละเอียดดังตารางที่ 4.6

คุณลักษณะข้อมูล	ค่าที่เป็นไปได้
ProvinceId	เรียงรหัสจังหวัดตามตัวอักษร โดย 0 = ไม่ระบุจังหวัด 1 = รหัสจังหวัดกระบี่ 2 = รหัสกรุงเทพมหานคร 3 = รหัสจังหวัดกาญจนบุรี ... ฯลฯ
RegionId	แบ่งออกเป็น 6 กลุ่ม คือ 0 = ไม่ทราบภูมิภาค 1 = ภาคเหนือ 2 = ภาคใต้ 3 = ภาคกลาง 4 = ภาคตะวันออกเฉียงเหนือ 5 = ภาคตะวันออก
hour	แบ่งออกเป็น 24 กลุ่ม แยกรายชั่วโมง คือ 00:00 – 00:59 01:00 – 01:59 02:00 – 02:59 03:00 – 03:59 04:00 – 04:59 05:00 – 05:59 06:00 – 06:59 07:00 – 07:59 08:00 – 08:59 09:00 – 09:59 10:00 – 10:59

	11:00 – 11:59
	12:00 – 12:59
	13:00 – 13:59
	14:00 – 14:59
	15:00 – 15:59
	16:00 – 16:59
	17:00 – 17:59
	18:00 – 18:59
	19:00 – 19:59
	20:00 – 20:59
	21:00 – 21:59
	22:00 – 22:59
	23:00 – 23:59

**ตารางที่ 4.6 การประมวลผลข้อมูลแบบที่ 6**

## 4.2 การพัฒนาระบบ

ระบบประกอบด้วยโมเดลพยากรณ์หัวข้อข่าว และฟังก์ชันการเลือกข่าว ซึ่งมีขั้นตอนการพัฒนา ดังนี้

### 4.2.1 การพยากรณ์หัวข้อข่าว

นำการประมวลผลข้อมูลทั้ง 6 แบบ ไปสร้าง 4 โมเดล ได้แก่ ต้นไม้ตัดสินใจ, การถดถอยเชิงโลจิสติก, ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม ได้ผลลัพธ์ค่าความถูกต้องของโมเดลพยากรณ์หัวข้อข่าวออกมาเป็นเปอร์เซ็นต์ ดังตารางที่ 4.7

โมเดล	ข้อมูล แบบ ที่ 1	ข้อมูล แบบ ที่ 2	ข้อมูล แบบ ที่ 3	ข้อมูล แบบ ที่ 4	ข้อมูล แบบ ที่ 5	ข้อมูล แบบ ที่ 6
ต้นไม้ ตัดสินใจ	44.20%	44.48%	44.41%	43.60%	43.18%	43.12%
การถดถอย เชิงโลจิสติก	44.48%	44.47%	44.52%	44.75%	44.61%	44.56%
ซัพพอร์ต เวกเตอร์แม ชชีน	49.04%	49.06%	48.91%	47.91%	48.39%	48.11%
โครงข่าย ประสาท เทียม	49.40%	<b>50.59%</b>	49.97%	49.09%	49.22%	49.67%

ตารางที่ 4.7 ตารางแสดงผลลัพธ์เปอร์เซ็นต์ค่าความถูกต้องของโมเดลจาก 4 โมเดล

จากตารางที่ 4.7 จะเห็นได้ว่าโมเดลที่ให้ผลลัพธ์เปอร์เซ็นต์ค่าความถูกต้องของโมเดลพยากรณ์หัวข้อข่าวออกมาดีที่สุด คือ โครงข่ายประสาทเทียม กับการประมวลผลข้อมูลแบบที่ 2 โดยได้เปอร์เซ็นต์ค่าความถูกต้องของโมเดลพยากรณ์หัวข้อข่าว 50.59%

#### 4.2.2 ฟังก์ชันการเลือกข่าว

ผู้วิจัยได้สร้างฟังก์ชันการเลือกข่าว จากบทความข่าวทั้งหมด 61,455 บทความ โดยนำผลลัพธ์ของโมเดลพยากรณ์หัวข้อข่าวที่ดีที่สุด มาเข้าสู่กระบวนการฟังก์ชันการเลือกข่าว โดยจะเลือกใช้ข่าวล่าสุด โดยการแสดงผลของบทความข่าวนั้นขึ้นอยู่กับที่เข้าถึงบทความด้วย เช่น หากเป็นโทรศัพท์มือถือ จะแสดงผลบทความข่าวทีละ 1 บทความ โดยใช้วิธีเลื่อนข่าวจากด้านบนสุดลงสู่บทความข่าวถัดไปที่อยู่ด้านล่าง เป็นต้น ทั้งนี้การแสดงผลบทความจะถูกปรับเปลี่ยนไปตามแพลตฟอร์มที่ผู้อ่านข่าวเข้าถึงบทความข่าว ซึ่งผู้วิจัยเลือกใช้การแสดงผลบทความข่าวบนพีซี และด้วยข้อจำกัดของหน้าแพลตฟอร์มบนพีซี ที่ผู้วิจัยใช้ทำการทดลองนั้นถูกออกแบบมาให้แสดงผลบทความได้ 2 แบบคือ 4 หรือ 8 บทความ ดังภาพประกอบ 4.9



1) แสดงผล 4 บทความ

2) แสดงผล 8 บทความ

ภาพประกอบ 4.9 แสดงภาพตัวอย่างพื้นที่หน้าแพลตฟอร์มที่เลือกมาทดสอบฟังก์ชันการเลือกข่าว

จากผลการคำนวณโดยใช้ชุดข้อมูลฝึกสอนพบว่า มีค่าความถูกต้องในการแสดงผลแบบ 4 บทความ และแบบ 8 บทความ เป็นดังนี้

จำนวน บทความ ข่าว	ข้อมูลแบบ ที่1	ข้อมูลแบบ ที่2	ข้อมูลแบบ ที่3	ข้อมูลแบบ ที่4	ข้อมูลแบบ ที่5	ข้อมูลแบบ ที่6
ค่าความถูกต้องของการ แสดงผล 4 บทความ	39.62%	39.64%	39.64%	39.64%	39.63%	39.64%
ค่าความถูกต้องของการ แสดงผล 8 บทความ	50.00%	50.04%	50.01%	50.03%	50.00%	50.03%

**ตาราง 4.8 แสดงผลลัพธ์การหาค่าความถูกต้องของการคลิกอ่านบทความที่ 1  
ไปยังบทความที่ 2**

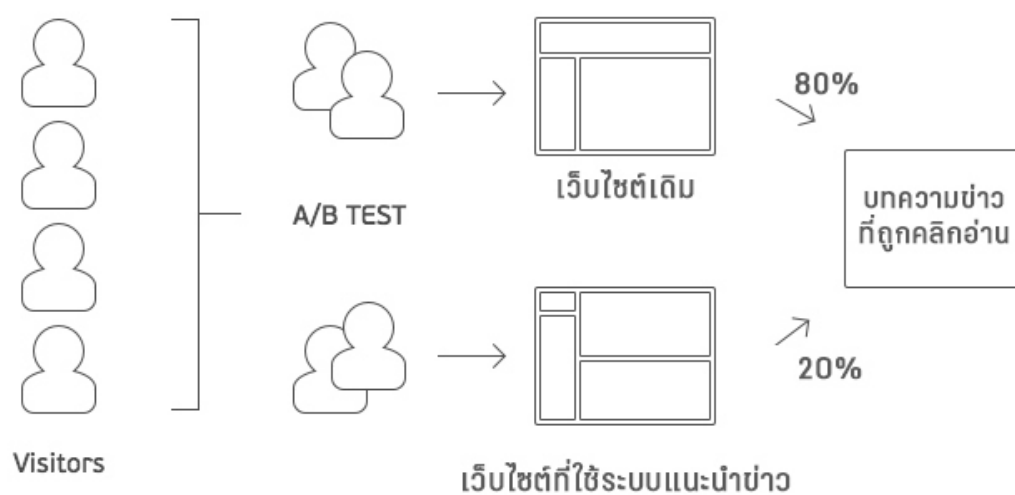
จากข้อมูลตาราง 4.8 พบว่าผลลัพธ์ของการหาโอกาสที่ข่าวล่าสุดจะถูกคลิกอ่านจากบทความที่ 1 ไปยังบทความที่ 2 ได้ผลดีที่สุดคือ 50.04% กับการแสดงผล 8 บทความ และชุดข้อมูลการประมวลผลข้อมูลแบบที่ 2

#### 4.3 การนำระบบไปใช้งานจริง

เมื่อผู้วิจัยนำระบบพยากรณ์หัวข้อข่าวที่สร้างด้วยโมเดลโครงข่ายประสาทเทียม โดยใช้ชุดข้อมูลการประมวลผลแบบที่ 2 และแสดงผล 8 บทความ ไปทดสอบบนแพลตฟอร์ม ด้วยวิธีการเปรียบเทียบแพลตฟอร์มข่าวเดิมซึ่งแนะนำบทความข่าวโดยระบบเชิงพาณิชย์ กับหน้าแพลตฟอร์มที่แสดงผลแบบเดียวกันแต่ใช้ระบบพยากรณ์หัวข้อข่าว โดยกำหนดให้สัดส่วนการมองเห็น



(Impressions) บทความข่าวทั้งหมด 5,265,988 ครั้ง สัดส่วนเป็น 80% : 20% โดยกระบวนการทำงานอธิบายได้ดังภาพประกอบ 4.10



ภาพประกอบ 4.10 กระบวนการทำงานของ A/B Test

ผู้วิจัยเริ่มทดสอบในวันที่ 8 มกราคม 2562 ช่วงเวลาตั้งแต่ 18:00 น. และสิ้นสุดที่เวลา 17.59 น. ของวันถัดไป เป็นเวลา 24 ชั่วโมง และเปรียบเทียบผลลัพธ์อัตราการคลิก ดังสมการ (1)

$$CTR(\text{อัตราการคลิก}) = \frac{\text{Number of Click Through Rate}}{\text{Number of Impressions}} \times 100 \quad (1)$$

### เปรียบเทียบผลลัพธ์อัตราการคลิกได้ดังตารางที่ 4.9

	ระบบแนะนำข่าวรายบุคคล (20%)	ระบบเชิงพาณิชย์ (80%)
การมองเห็น (Impressions)	877,678	4,388,310
อัตราการคลิก (Click Through Rate)	55,542	240,656
คิดเป็นเปอร์เซ็นต์ (CTR)	<b>6.32%</b>	<b>5.48%</b>

ตาราง 4.9 แสดงผลการเปรียบเทียบอัตราการคลิกระหว่างระบบแนะนำข่าวรายบุคคล  
กับระบบเชิงพาณิชย์

จากตาราง 4.9 เปรียบเทียบผลลัพธ์อัตราการคลิกระหว่างระบบแนะนำข่าวรายบุคคล กับระบบเชิงพาณิชย์พบว่า ระบบแนะนำข่าวรายบุคคลที่กำหนดให้มีสัดส่วนการมองเห็น 20% นั้นให้ผลลัพธ์อัตราการคลิก 6.32% ส่วนระบบเชิงพาณิชย์ที่กำหนดให้มีสัดส่วนการมองเห็น 80% ให้ผลลัพธ์อัตราการคลิก 5.48%

นอกจากนี้ ผู้วิจัยยังสรุปข้อมูลกลุ่มของผู้คลิกอ่านบทความข่าวที่แนะนำโดยระบบแนะนำข่าวรายบุคคล ได้ดังนี้

#### 1. ข้อมูลเพศ ดังแสดงในตาราง 4.10

เพศ	จำนวนคลิก (ครั้ง)	เปอร์เซ็นต์
เพศชาย	27,804	50.06
เพศหญิง	27,738	49.94

ตาราง 4.10 แสดงข้อมูลสัดส่วนเพศของผู้คลิกอ่านบทความจากระบบแนะนำข่าวรายบุคคล

จากตาราง 4.10 กลุ่มของผู้คลิกล่ามบทยความข่าวจากจำนวนการคลิกล่ามบทยความข่าว 55,542 ครั้ง มีผู้คลิกล่ามบทยความข่าวที่เป็นผู้ชาย 27,804 ครั้ง คิดเป็น 50.06% และผู้คลิกล่ามบทยความข่าวที่เป็นผู้หญิง 27,738 ครั้ง คิดเป็น 49.94%

## 2. ข้อมูลอายุ ดังแสดงในตาราง 4.11

กลุ่มอายุ	อายุ (ปี)
ช่วงอายุ	30 - 60
เฉลี่ย	45
พบมากที่สุด	37
พบมากเป็นอันดับ 2	45
พบมากเป็นอันดับ 3	36

ตาราง 4.11 แสดงข้อมูลสัดส่วนอายุของผู้คลิกล่ามบทยความข่าวจากระบบแนะนำข่าวรายบุคคล

จากตาราง 4.11 จากจำนวนการคลิกล่ามบทยความข่าว 55,542 ครั้ง พบว่า กลุ่มของผู้คลิกล่ามบทยความข่าวอยู่ในช่วงอายุ 30 – 60 ปี โดยอายุเฉลี่ยที่คลิกล่ามบทยความข่าว คือ 45 ปี และช่วงอายุที่คลิกล่ามบทยความข่าวมากที่สุด คือ 37 ปี รองลงมาคืออายุ 45 ปี และ 36 ปี ตามลำดับ

## 3. ข้อมูลจังหวัด ดังแสดงในตาราง 4.12

กลุ่มจังหวัด	จังหวัด
จังหวัดที่พบมากที่สุด	ประจวบคีรีขันธ์
พบมากเป็นอันดับ 2	ปทุมธานี
พบมากเป็นอันดับ 3	สมุทรสาคร

ตาราง 4.12 แสดงข้อมูลจังหวัดของผู้คลิกล่ามบทยความข่าวจากระบบแนะนำข่าวรายบุคคล

จากตาราง 4.12 จากจำนวนการคลิกอ่านบทความข่าวทั้งหมด 55,542 ครั้ง พบว่า กลุ่มจังหวัดที่มีการคลิกอ่านบทความข่าวมากที่สุด คือ จังหวัดประจวบคีรีขันธ์ รองลงมาคือ จังหวัดปทุมธานี และจังหวัดสมุทรสาคร

#### 4. ข้อมูลภูมิภาค ดังแสดงในตาราง 4.13

ภูมิภาค	จำนวนครั้ง (คลิก)	เปอร์เซ็นต์
ใต้	14,440	26
กลาง	13,885	25
ตะวันออกเฉียงเหนือ	12,219	22
เหนือ	9,997	18
ตะวันออก	5,001	9

ตาราง 4.13 แสดงข้อมูลตามภาคของผู้คลิกอ่านบทความจากระบบ  
แนะนำข่าวรายบุคคล

จากตาราง 4.13 จากจำนวนการคลิกอ่านบทความข่าวทั้งหมด 55,542 ครั้ง พบว่า ภูมิภาคที่มีการคลิกอ่านบทความข่าวมากที่สุด ได้แก่ ภาคใต้ มีจำนวนการคลิก 14,440 ครั้ง คิดเป็น 26% รองลงมาคือ ภาคกลาง มีจำนวนการคลิก 13,885 ครั้ง คิดเป็น 25% และภาคตะวันออกเฉียงเหนือ มีจำนวนการคลิก 12,219 ครั้ง คิดเป็น 22%

#### 5. ระบบปฏิบัติการ ดังแสดงในตาราง 4.14

ระบบปฏิบัติการ	จำนวนครั้ง (คลิก)	เปอร์เซ็นต์
Android	36,657	66
iOS	12,219	22
Windows	4,998	9
อื่นๆ	1,668	3

ตาราง 4.14 แสดงข้อมูลระบบปฏิบัติการของผู้คลิกอ่านบทความจากระบบ

### แนะนำข่าวรายบุคคล

จากตาราง 4.14 จากจำนวนการคลิกอ่านบทความข่าวทั้งหมด 55,542 ครั้ง พบว่า มีการเข้าถึงบทความข่าวด้วยระบบปฏิบัติการ Android จำนวนการคลิก 36,657 ครั้ง คิดเป็น 66% เข้าถึงบทความข่าวด้วยอุปกรณ์ iOS จำนวนการคลิก 12,219 ครั้ง คิดเป็น 22% เข้าถึงบทความข่าวด้วยระบบปฏิบัติการ Windows จำนวนการคลิก 4,998 ครั้ง คิดเป็น 9% และเข้าถึงด้วยระบบปฏิบัติการอื่นๆ จำนวนการคลิก 1,668 ครั้ง คิดเป็น 3%

### 6. ช่วงเวลา ดังแสดงในตาราง 4.15

ช่วงเวลา	จำนวนครั้ง (คลิก)	เปอร์เซ็นต์
18:00 - 23:59	21,661	39
12:00 - 17:59	15,551	28
06:00 - 11:59	14,440	26
00:00 - 05:59	3,890	7

ตาราง 4.15 แสดงข้อมูลช่วงเวลาของผู้คลิกอ่านบทความจากระบบแนะนำข่าวรายบุคคล

จากตาราง 4.15 จากจำนวนการคลิกอ่านบทความข่าวทั้งหมด 55,542 ครั้ง พบว่า ช่วงเวลาที่มีผู้คลิกอ่านบทความข่าว เรียงลำดับ ดังนี้

- ช่วงเวลา 18:00 - 23:59 น. จำนวนการคลิก 21,661 ครั้ง คิดเป็น 39%
- ช่วงเวลา 12:00 - 17:59 น. จำนวนการคลิก 15,551 ครั้ง คิดเป็น 28%
- ช่วงเวลา 06:00 - 11:59 น. จำนวนการคลิก 14,440 ครั้ง คิดเป็น 26%
- ช่วงเวลา 00:00 - 05:59 น. จำนวนการคลิก 3,890 ครั้ง คิดเป็น 7%

ผลการวิเคราะห์เปอร์เซ็นต์การคลิกบทความข่าวจาก หัวข้อข่าว 1 ไป หัวข้อข่าว 2 แสดงใน ตาราง 4.16

หัวข้อข่าว2 หัวข้อข่าว1	บันเทิง	สังคม	อาชญากรรม	การเมือง	ต่างประเทศ
บันเทิง	62.73	0	0	1.01	36.26
สังคม	26.05	0	0	9.76	64.20
อาชญากรรม	0.11	0	49.66	0	50.22
การเมือง	0	0	0	30.22	69.78
ต่างประเทศ	1.29	0	0	1.7	97.01

ตาราง 4.16 เปอร์เซนต์การคลิกบทความข่าวจาก หัวข้อข่าว 1 ไป หัวข้อข่าว 2

จากตาราง 4.16 พบว่าวิเคราะห์การคลิกอ่านบทความข่าวจากบทความที่ 1 ไปยังบทความถัดไปได้ดังนี้

- หากผู้อ่านข่าวออนไลน์อ่านบทความแรกใน หัวข้อข่าว บันเทิง บทความถัดไปที่คลิกอ่านยังคงเป็น หัวข้อข่าว บันเทิงถึง 62.73% รองลงมาเป็น หัวข้อข่าว ต่างประเทศ 36.26%
- หากผู้อ่านข่าวออนไลน์อ่านบทความแรกใน หัวข้อข่าว สังคม บทความถัดไปที่คลิกอ่านเป็น หัวข้อข่าว ต่างประเทศถึง 64.20% รองลงมาเป็น หัวข้อข่าว บันเทิง 26.05%
- หากผู้อ่านข่าวออนไลน์อ่านบทความแรกใน หัวข้อข่าว อาชญากรรม บทความถัดไปที่คลิกอ่านเป็น หัวข้อข่าว ต่างประเทศถึง 50.22% รองลงมาเป็น หัวข้อข่าว อาชญากรรม 49.66%
- หากผู้อ่านข่าวออนไลน์อ่านบทความแรกใน หัวข้อข่าว การเมือง บทความถัดไปที่คลิกอ่านเป็น หัวข้อข่าว ต่างประเทศถึง 69.78% รองลงมาเป็น หัวข้อข่าว การเมือง 30.22%

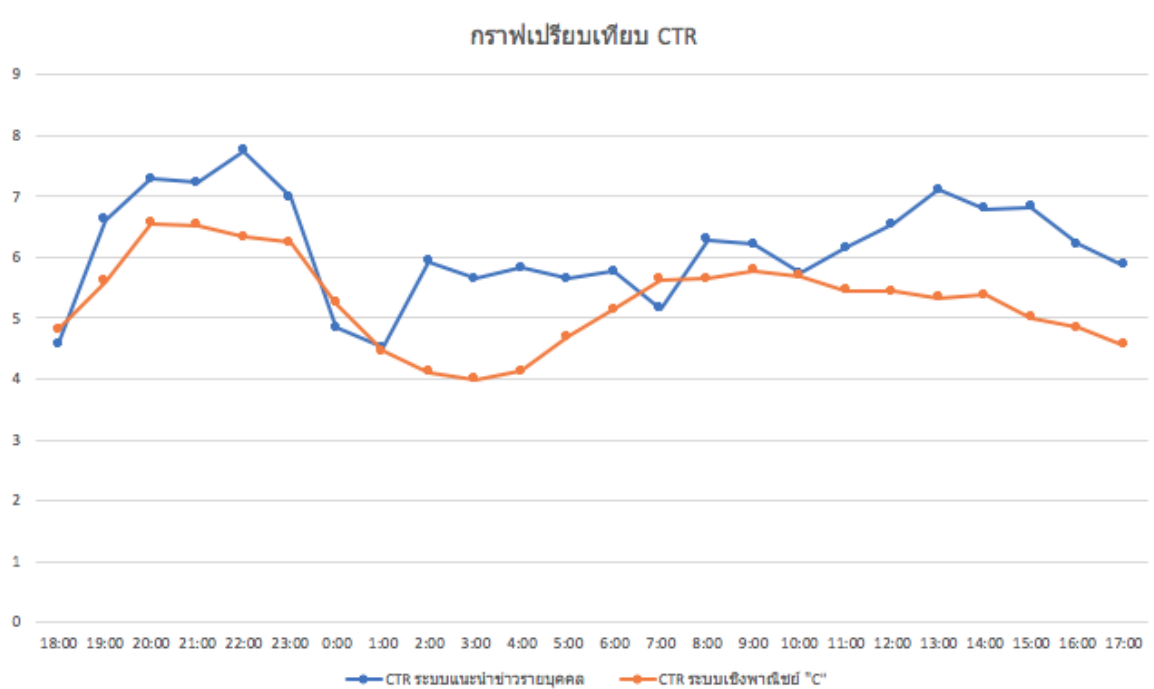
- หากผู้อ่านข่าวออนไลน์อ่านบทความแรกใน หัวข้อข่าว ต่างประเทศ บทความถัดไปที่คลิกอ่านยังคงเป็น หัวข้อข่าว ต่างประเทศถึง 97.01% รองลงมาเป็น หัวข้อข่าว บันเทิง 1.29%

ผู้วิจัยยังเปรียบเทียบการคลิกเป็นรายชั่วโมง เพื่อวิเคราะห์แนวโน้มระบบแนะนำข่าวรายบุคคลที่จะมีอัตราการคลิกดีกว่าระบบแนะนำเชิงพาณิชย์หรือไม่ ดังนี้

เวลา	CTR ระบบแนะนำ ข่าวรายบุคคล (20%)	CTR ระบบเชิง พาณิชย์ (80%)
18:00	4.57	4.81
19:00	6.62	5.61
20:00	7.28	6.55
21:00	7.22	6.53
22:00	7.75	6.33
23:00	6.99	6.25
00:00	4.85	5.24
01:00	4.52	4.46
02:00	5.93	4.11
03:00	5.64	4.00
04:00	5.83	4.12
05:00	5.64	4.70
06:00	5.77	5.14
07:00	5.16	5.63
08:00	6.29	5.64
09:00	6.22	5.78
10:00	5.74	5.69
11:00	6.15	5.45
12:00	6.54	5.44
13:00	7.11	5.34
14:00	6.79	5.38
15:00	6.82	5.00

16:00	6.21	4.85
17:00	5.87	4.56

ตารางที่ 4.17 เปรียบเทียบการคลิกเป็นรายชั่วโมงระหว่างระบบแนะนำข่าวรายบุคคล กับระบบเชิงพาณิชย์



ภาพประกอบ 4.9 กราฟวิเคราะห์แนวโน้ม CTR ระหว่างระบบแนะนำข่าวรายบุคคล กับระบบเชิงพาณิชย์

จากตาราง 4.17 เป็นการเปรียบเทียบอัตราการคลิกอ่านบทความข่าวโดยเทียบระหว่างระบบแนะนำข่าวรายบุคคลกับระบบเชิงพาณิชย์ เป็นรายชั่วโมง พบว่า มีช่วงเวลาที่ระบบพยากรณ์หัวข้อข่าว มีอัตราการคลิกอ่านบทความข่าวที่แนะนำสูงกว่าระบบเชิงพาณิชย์ มีด้วยกัน 4 ช่วงเวลา คือ 19:00 – 23:00 น., 02:00 – 06:00 น., 08:00 – 09:00 น., 11:00 – 17:00 น. ซึ่งจากกราฟในภาพประกอบ 4.9 แสดงให้เห็นว่าอัตราการคลิกอ่านบทความข่าวตลอด 24 ชั่วโมงนั้น ส่วนใหญ่มาจากระบบพยากรณ์หัวข้อข่าว

เพื่อเป็นการทดสอบว่าอัตราการคลิกรายชั่วโมงของระบบแนะนำข่าวรายบุคคลมากกว่าระบบเชิงพาณิชย์ "C" อย่างมีนัยสำคัญทางสถิติ ผู้วิจัยจึงได้ทำการทดสอบสมมติฐานทางสถิติเพื่อเปรียบเทียบค่าเฉลี่ยของอัตราการคลิกระหว่างสองระบบ โดย ได้ตั้งสมมติฐานดังนี้



ให้  $\mu$  แทนค่าเฉลี่ยของประชากร

$\bar{x}$  แทนค่าเฉลี่ยของอัตราการคลิกرایชั่วโมง

$$H_0: \mu_1 \leq \mu_2 \quad H_1: \mu_1 > \mu_2$$

สูตรคำนวณ

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)_0}{\sqrt{s_p^2/n_1 + s_p^2/n_2}}$$

t-Test: Two-Sample Assuming Unequal Variances		
	4.57	4.81
Mean	6.214782609	5.295652174
Variance	0.643680632	0.555680237
Observations	23	23
Hypothesized Mean Difference	0	
df	44	
t Stat	4.025002248	
P(T<=t) one-tail	0.000110594	
t Critical one-tail	1.680229977	
P(T<=t) two-tail	0.000221187	
t Critical two-tail	2.015367574	

ภาพประกอบ 4.10 ผลการทดสอบสมมติฐานทางสถิติ

จากภาพประกอบ 4.10 ผลการทดสอบ ได้ค่า  $P - Value = 0.0001 < 0.05$  จึงปฏิเสธ  $H_0$  และยอมรับ  $H_1$

สรุปได้ว่า ค่าเฉลี่ยของอัตราการคลิกของระบบเรา มากกว่าระบบเชิงพาณิชย์ อย่างมีนัยสำคัญที่ความเชื่อมั่น 95%

### ผลการพัฒนาระบบ สรุปได้ดังนี้

1. ระบบแนะนำข่าวรายบุคคล ที่สร้างด้วยโมเดลโครงข่ายประสาทเทียม โดยใช้ชุดข้อมูลการประมวลผลแบบที่ 2 และแสดงผล 8 บทความ ซึ่งชุดข้อมูลที่ใช้สอน ระบบกับเวลาในการทดสอบ 24 ชั่วโมงนั้น มีอัตราการคลิก 6.32% ซึ่งสูงกว่าระบบเชิงพาณิชย์ “C” ที่มีอัตราการคลิก 5.48% ผลการเปรียบเทียบอัตราการคลิกระบบแนะนำข่าวรายบุคคลสูงกว่าระบบเชิงพาณิชย์ “C” จึงสามารถตอบสมมติฐานที่ว่าระบบการแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าว สามารถช่วยให้มีอัตราการคลิกอ่านบทความข่าวที่สูงขึ้น

2. พฤติกรรมของผู้อ่านข่าวออนไลน์ข่าวบนแพลตฟอร์มมีแนวโน้มเลือกอ่านข่าวล่าสุดตามหมวดที่ตนเองสนใจ ดังนั้น โมเดลเพื่อแนะนำหัวข้อข่าว โดยแนะนำข่าวล่าสุดตามหมวดที่ผู้อ่านข่าวออนไลน์แต่ละคนสนใจ จะสามารถตอบโต้กับพฤติกรรมกรรมการอ่านได้ ทั้งนี้ จากข้อมูลข่าวสารที่มีมากมายในปัจจุบัน ทำให้ผู้เสพสื่อหรือผู้รับสารอยู่ในมุมของตัวเอง หรือที่เรียกว่า Echo Chamber (ประชากร, 2560) มีผลทำให้ผู้คนเลือกที่จะอ่านบทความข่าวนั้นแล้วข่าวต่อไปที่เลือกอ่านยังคงอยู่ในหัวข้อข่าว เดิม โดยจากผลการทดลองดังแสดงในตาราง 4.16 พบว่า ผู้อ่านข่าวออนไลน์อ่านบทความแรกใน หัวข้อข่าว บันเทิง บทความถัดไปที่คลิกอ่านยังคงเป็น หัวข้อข่าว บันเทิงถึง 62.73% หรือผู้อ่านข่าวออนไลน์ที่อ่านบทความแรกใน หัวข้อข่าว ต่างประเทศ บทความถัดไปที่คลิกอ่านยังคงเป็น หัวข้อข่าว ต่างประเทศถึง 97.01% นอกจากนั้นผู้อ่านข่าวออนไลน์ยังเลือกอ่านบทความใน หัวข้อข่าว ที่มีเนื้อหาใกล้เคียงกับ หัวข้อข่าว เดิมที่เคยอ่านมา ซึ่งจากผลการทดลองพบว่าผู้อ่านข่าวออนไลน์ที่อ่านบทความแรกใน หัวข้อข่าว การเมือง หรือสังคม พบว่าบทความถัดไปที่คลิกอ่านซึ่งมีเนื้อหาใกล้เคียงกันคือ หัวข้อข่าว ต่างประเทศมีผู้คลิกอ่านถึง 69.78% และ 64.20% ตามลำดับ

## บทที่ 5

### สรุปผลการวิจัย และข้อเสนอแนะ

ในงานวิจัยนี้ผู้วิจัยได้เสนอขั้นตอนวิธีการศึกษาและพัฒนาระบบ เพื่อนำระบบที่ได้ไปพยากรณ์บทความข่าวที่จะแนะนำให้กับผู้อ่านข่าวออนไลน์ เพื่อทดสอบสมมติฐานที่ว่า ระบบการแนะนำข่าวจากประวัติและพฤติกรรมของผู้อ่านข่าวออนไลน์ข่าว ช่วยให้อัตราการคลิกอ่านข่าวสูงขึ้น โดยขั้นตอนวิธีที่นำเสนอสามารถแบ่งได้ดังนี้

**ส่วนแรก** เป็นการเตรียมข้อมูล โดยเลือกช่วงเวลาในการเก็บข้อมูลผู้ลงทะเบียนเข้าแพลตฟอร์ม เตรียมความพร้อมของข้อมูล ศึกษาคุณลักษณะของข้อมูล โดยวิเคราะห์ข้อมูลของผู้ลงทะเบียนเข้าแพลตฟอร์มและรูปแบบของพฤติกรรม ศึกษาปัญหาที่ผู้อ่านข่าวออนไลน์ออกจากแพลตฟอร์มไปโดยที่ไม่คลิกอ่านบทความอื่นต่อ

**ส่วนที่สอง** เป็นการพัฒนาระบบที่จะนำมาใช้ในการพยากรณ์และสร้าง 4 โมเดล ได้แก่ ต้นไม้ตัดสินใจ , การถดถอยเชิงโลจิสติก , ซัพพอร์ตเวกเตอร์แมชชีน และโครงข่ายประสาทเทียม พบว่าโมเดลที่ได้ผลลัพธ์ดีที่สุดในการพยากรณ์หัวข้อข่าว คือ โครงข่ายประสาทเทียม มีผลการทดลอง มีด้วยกัน 2 ส่วนคือ

1. ผลของการสร้างและทดสอบโมเดลแนะนำหัวข้อข่าว โดยทั้ง 4 โมเดลกับการประมวลผลข้อมูล 6 แบบ ที่สร้างนั้นให้ผลลัพธ์ค่าความถูกต้องของโมเดล ดังนี้ ต้นไม้ตัดสินใจ 44.48% ในการประมวลผลข้อมูลแบบที่ 2 , การถดถอยเชิงโลจิสติก 44.75% ในการประมวลผลข้อมูลแบบที่ 4 , ซัพพอร์ตเวกเตอร์แมชชีน 49.06% ในการประมวลผลข้อมูลแบบที่ 2 และ 50.59% ในการประมวลผลข้อมูลแบบที่ 2 ดังนั้น โมเดลโครงข่ายประสาทเทียม ในการประมวลผลข้อมูลชุดที่ 2 สามารถแนะนำหัวข้อข่าวดีที่สุด

2. ผลการทดสอบโมเดลแนะนำหัวข้อข่าวกับชุดข้อมูล ซึ่งมีการประมวลผลข้อมูลออกมา 6 แบบ พบว่าการหาค่าความถูกต้องของการคลิกอ่านบทความที่ 1 ไปยังบทความที่ 2 ได้ผลลัพธ์ ดังนี้ การแสดงผล 4 บทความ ให้ค่าความถูกต้องดีที่สุด 39.64% ในการประมวลผลข้อมูลแบบที่ 2 , 3 , 4 และ 6 เมื่อเปรียบเทียบกับผลการแสดงผล 8 บทความ ให้ค่าความถูกต้องดีที่สุด 50.04% ในชุดข้อมูลการประมวลผลข้อมูลแบบที่ 2

**ส่วนที่สาม** นำระบบแนะนำหัวข้อข่าวมาทำการทดสอบกับแพลตฟอร์มข่าวจริง และวัดประสิทธิภาพของระบบแนะนำข่าวรายบุคคล ด้วยอัตราการคลิก จากบทความข่าวทั้งหมด 61,455

บทความ จากนั้นนำผลลัพธ์ของโมเดลพยากรณ์หัวข้อข่าวที่ดีที่สุด มาเข้าสู่กระบวนการฟังก์ชันการเลือกข่าว โดยจะเลือกใช้ข่าวล่าสุด ซึ่งผลการทดสอบเมื่อนำระบบไปใช้งานจริงบนแพลตฟอร์มเป็นเวลา 24 ชั่วโมง พบว่า อัตราการคลิกระหว่างระบบแนะนำข่าวรายบุคคล กับระบบเชิงพาณิชย์ มีความใกล้เคียงกัน โดยระบบแนะนำข่าวรายบุคคล ให้ผลอัตราการคลิกอ่านบทความสูงกว่าเล็กน้อย อยู่ที่ 6.32% ขณะที่ระบบเชิงพาณิชย์ให้ผลอัตราการคลิกอ่านข่าวอยู่ที่ 5.48%

### ข้อเสนอแนะสำหรับงานในอนาคต

ในการสร้างระบบแนะนำข่าวรายบุคคลโดยใช้ประวัติและพฤติกรรมผู้อ่านข่าวออนไลน์นั้น สามารถทำได้เฉพาะเงื่อนไขที่ตรงกับรูปแบบข้อมูลที่กำหนดไว้ ทั้งนี้ยังมีข้อแนะนำหากต้องการพัฒนาระบบให้มีประสิทธิภาพดียิ่งขึ้น เช่น

1. การวิจัยนี้เป็นการหาโมเดลในการแนะนำบทความข่าวบทของแพลตฟอร์มข่าวเพียงแพลตฟอร์มเดียวเท่านั้น หากต้องการนำเทคนิคนี้ไปทดสอบกับแพลตฟอร์มข่าวอื่นค่าความถูกต้องอาจเปลี่ยนแปลง ขึ้นอยู่กับข้อมูลที่ใช้ในการสร้างระบบ อาจต้องปรับแต่งระบบใหม่เพื่อให้เหมาะสมกับพฤติกรรมผู้ใช้งานแพลตฟอร์มข่าวดังกล่าว

2. เนื่องจากการวิจัยนี้ถูกทำในเวลาจำกัด จึงทำให้ผู้วิจัยต้องเลือกโมเดลที่ได้รับความนิยมในการทำการวิจัยมาทดสอบ ซึ่งผู้วิจัยเลือกมาเพียง 4 โมเดล หากต้องการนำไปทดสอบเพิ่มควรทดสอบระบบกับโมเดลอื่นๆ เพื่อความหลากหลายเพิ่มขึ้น ซึ่งอาจทำให้ค่าความถูกต้องของโมเดลมีค่าสูงขึ้น

3. ในการประมวลผลข้อมูล ควรลองประมวลผลข้อมูลแบบแบ่งช่วงเวลา แบบ ทุก 4 ชั่วโมง ซึ่งอาจได้ค่าความแตกต่างของข้อมูลเพิ่ม

4. ระบบแนะนำข่าวรายบุคคลของผู้วิจัยสามารถนำไปประยุกต์ใช้กับการกรองด้วยเนื้อหาได้ ซึ่งเป็นการกรองที่ที่ไม่มีข้อมูลประวัติ (Anonymous) ของผู้อ่านข่าวออนไลน์ ได้ เช่น ไม่มีข้อมูล ชื่อ เพศ อายุ เป็นต้น แต่จะสามารถแนะนำบทความข่าวโดยใช้พฤติกรรมผู้อ่านข่าวออนไลน์เพียงอย่าง

เดียว โดยนำข้อมูลส่วนอื่นๆ มาพิจารณาเพิ่มเติม เช่น คำสำคัญในข่าว (keyword , tag) , การนำข้อความแสดงความคิดเห็นต่อบทความข่าวในโซเชียลมีเดียมาวิเคราะห์ เป็นต้น

## บรรณานุกรม

- กัลยา วาณิชย์บัญชา. 2542. การวิเคราะห์สถิติ : สถิติเพื่อการตัดสินใจ. กรุงเทพฯ : โรงพิมพ์  
แห่งจุฬาลงกรณ์วิทยาลัย
- จันทิมา พลพิณ. 2548. **Content-based Probabilistic Text Classifier for Pornographic Web  
Filtering**. มหาสารคาม : มหาวิทยาลัยมหาสารคาม
- ญาใจ ลิมปิยะภรณ์. 2556. ตำราวิศวะ 100 ปี 2456-2556 การทำเหมืองข้อมูล. กรุงเทพฯ :  
ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย
- นลินี โสพัตสถิต. 2555. การใช้ระบบแนะนำสนับสนุนการตัดสินใจ. กรุงเทพมหานคร :  
มหาวิทยาลัยราชภัฏสวนสุนันทา
- นิเวศ จิระวิชิตชัย, ปริญญญา สงวนสัตย์, พยุง มีสัจ. 2553. การพัฒนาประสิทธิภาพการจัด  
หมวดหมู่เอกสารภาษาไทยแบบอัตโนมัติ. กรุงเทพฯ : NIDA Development Journal  
ประชากรธรรม. 2560 .ความท้าทายกลายเป็น “สื่อใหม่” ที่มากกว่า “New Platform” [Online] 20  
ธันวาคม 2560, เข้าถึงได้จาก : [www.prachatham.com/article\\_detail.php?id=485](http://www.prachatham.com/article_detail.php?id=485)
- พรเทพ เขตรัมย์. 2559. **A/B testing** คืออะไร สรุปรขั้นตอนการดำเนินงานตั้งแต่เริ่มจนจบเพื่อเพิ่ม  
**conversion** ให้กับแพลตฟอร์ม [ออนไลน์]. 30 พฤศจิกายน 2560, เข้าถึงได้จาก:  
<https://gooleanalyticsthailand.wordpress.com/เรียน-สอน-google-analytics/>
- พรพล ธรรมรงค์รัตน์. 2552. การจำแนกประเภทเว็บเพจโดยวิธีลดขนาดลักษณะเฉพาะและซัพ  
พอร์ต เวกเตอร์แมชชีน. สงขลา : มหาวิทยาลัยสงขลานครินทร์
- มานะ ตีรียาภรณ์. 2560. สิ่งพิมพ์ไทยยุค 4.0 ยังซบเซา นักข่าวเสี่ยงตกงานเพิ่ม [ออนไลน์].  
15 ธันวาคม 2560, เข้าถึงได้จาก : <http://www.bbc.com/thai/thailand-39176803>
- เรชา โสภพพงษ์ และธงชัย แก้วกิริยา. ระบบแนะนำสถานที่ท่องเที่ยวสำหรับนักท่องเที่ยวโดยใช้  
เทคนิคดาต้าไมนิ่ง [Online] 12 มกราคม 2561, เข้าถึงได้จาก :  
<https://www.tcithaijo.org/index.php/sjss/aticle/download/26576/22557/>

- วิภาวรรณ บัวทอง. 2557. Data Preprocessing [ออนไลน์] . 1 มิถุนายน 2557, เข้าถึงได้จาก : <https://wipawanblog.files.wordpress.com/2014/06/chapter-3-data-preprocessing.pdf>
- สายชล สันสมบุรณ์ทอง. 2558. การทำเหมืองข้อมูล. กรุงเทพฯ : บริษัท จามจุรีโปรดักส์ จำกัด
- สิทธิชัย คำคง. 2554. การทำเหมืองข้อมูล (Data Mining) [ออนไลน์]. 30 พฤศจิกายน 2560, เข้าถึงได้จาก:<https://mahara.org/artefact/file/download.php?file=162194&view=45421>
- สิริทิพย์ ชันสุวรรณ . 2543. งานหนังสือพิมพ์ JOURNALISM. กรุงเทพฯ : มหาวิทยาลัยกรุงเทพ
- สุภา จันทา และนลินภัทร์ ปรวัฒน์ปรีกร . 2556. ระบบจำแนกและค้นคืนข้อมูลเว็บกระทำด้วยโครงข่ายประสาทเทียมเปอร์เซ็ปตรอนแบบหลายชั้น (A Web News Information Classification and Retrieval System using Multilayer Perception Neural Network). กรุงเทพมหานคร : มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าพระนครเหนือ
- อดุลย์ ยิ้มงาม. การทำเหมืองข้อมูล (Data Mining) [ออนไลน์]. 30 พฤศจิกายน 2560, เข้าถึงได้จาก: <http://compcenter.bu.ac.th/news-information/data-mining>
- เอกสิทธิ์ พัทธวงศ์ศักดิ์. 2557. การวิเคราะห์ข้อมูลด้วยเทคนิคดาต้า ไมน์นิ่ง เบื้องต้น. กรุงเทพมหานคร : บริษัท เอเชีย ดิจิตอลการพิมพ์ จำกัด
- เอกสิทธิ์ พัทธวงศ์ศักดิ์. 2557. การแบ่งข้อมูลเพื่อนำทดสอบประสิทธิภาพของโมเดล [ออนไลน์]. 30 พฤศจิกายน 2560, เข้าถึงได้จาก: <http://dataminingtrend.com/2014/data-mining-techniques/cross-validation/>
- เอกสิทธิ์ พัทธวงศ์ศักดิ์. dataminingtrend [ออนไลน์]. 30 พฤศจิกายน 2560, เข้าถึงได้จาก: <http://dataminingtrend.com/2014/data-mining-techniques/feature-selection-information-gain/>
- อุไรวรรณ, อมรนิมิตร. 2546. การวิเคราะห์ข้อมูลโดยใช้ Logistic Regression : ทางเลือกของการวิเคราะห์ความเสี่ยง. กรุงเทพมหานคร : มหาวิทยาลัยหอการค้าไทย
- Chen Li, Zhengtao Jiang. 2016. A Hybrid News Recommendation Algorithm Based on User's Browsing Path. Okayama, Japan.
- D. Jurafsky and J. H. Martin. 2017. Neural Nets and Neural Language Models [Online] 18 ธันวาคม 2560, เข้าถึงได้จาก : <https://web.stanford.edu/~jurafsky/slp3/8.pdf>
- Han, J. & Kamber, M. (2006). Data mining concepts and techniques (2nd ed.). United States of America: Morgan Kaufman Publishers.

- Jeff Hu . 2018. **A Simple Starter Guide to Build a Neural Network** [Online] 15 ธันวาคม 2560, เข้าถึงได้จาก : <https://www.kdnuggets.com/2018/02/simple-starter-guide-build-neural-network.html>
- Jiahui Liu, Peter Dolan, Elin Rønby Pedersen . 2010. **Personalized News Recommendation Based on Click Behavior**. USA : Google Inc.
- Kan Ouivirach. 2009. การวัดประสิทธิภาพของโมเดล (**Two-class prediction**) [ออนไลน์]. 15 ธันวาคม 2560, เข้าถึงได้จาก : <https://www.kanouivirach.com/2009/11/ประสิทธิภาพของโมเดล> **What is A/B Testing?** [Online]. 17 ธันวาคม 2560, เข้าถึงได้จาก : <https://www.mockingfish.com/resources/ab-testing/>
- Kittiphat Dumrongprat. **ทำความรู้จัก A/B Testing** [ออนไลน์]. 9 ธันวาคม 2560, เข้าถึงได้จาก : <http://www.stream.co.th/2017/05/ทำความรู้จัก-ab-testing/>
- Lasse Schultebrucks. 2017. **Introduction to Support Vector Machines** [Online] 15 ธันวาคม 2560, เข้าถึงได้จาก : <https://medium.com/@LSchultebrucks/introduction-to-support-vector-machines-9f8161ae2fcb>
- Lei Li, Ding-Ding Wang, Shun-Zhi Zhu and Tao Li. 2011. **Personalized News Recommendation: A Review and an Experimental Investigation**. U.S.A
- Leo Breiman. 2001. **Random Forests**. California, USA.
- Margaret Rouse. 2014. **data classification** [Online] . 15 ธันวาคม 2560, เข้าถึงได้จาก : <http://searchdatamanagement.techtarget.com/definition/data-classification>
- Nikolaos Kosmas Chlis. 2015. **Comparison of Statistical Methods for Genomic Signature Extraction** [Online] 18 ธันวาคม 2560, เข้าถึงได้จาก : [https://www.researchgate.net/figure/Holdout-validation-method\\_fig9\\_266617511](https://www.researchgate.net/figure/Holdout-validation-method_fig9_266617511)
- Tom M. Mitchell. 1997. **Machine Learning**. McGraw-Hill Science, USA.
- Wei Chu, Seung-Taek Park. 2009. **Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models**. Madrid, Spain.