# CLASSIFICATION OF BANKING PRODUCTS FROM THAI TEXT USING A SEMI-SUPERVISED LEARNING

อมรกานต์ ใจดี

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2565

# การจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความภาษาไทย ด้วยวิธีการเรียนรู้แบบกึ่งมีผู้สอน

อมรกานต์ ใจดี

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2565

# CLASSIFICATION OF BANKING PRODUCTS FROM THAI TEXT USING A SEMI-SUPERVISED LEARNING

AMORNKARN JAIDEE

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2022

สารนิพนธ์

เรื่อง

การจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความภาษาไทย

ด้วยวิธีการเรียนรู้แบบกึ่งมีผู้สอน

ของ

อมรกานต์ ใจดี

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

---

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

---

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.ศุภร คนธภักดี)

..... ประธาน

(ผู้ช่วยศาสตราจารย์ ดร.รัตน์ชัยนันท์ ธรรมสุจริต)

..... กรรมการ

(อาจารย์ ดร.โสภณ มงคลลักษมี)

| | |
|---|---|
| ชื่อเรื่อง | การจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความภาษาไทย ด้วยวิธีการเรียนรู้แบบกึ่งมีผู้สอน |
| ผู้วิจัย | อมรกานต์ ใจดี |
| ปริญญา | วิทยาศาสตรมหาบัณฑิต |
| ปีการศึกษา | 2565 |
| อาจารย์ที่ปรึกษา | อาจารย์ ดร. ศุภร คนธภักดี |

งานวิจัยนี้มีวัตถุประสงค์เพื่อให้สามารถจำแนกประเภทของผลิตภัณฑ์ธนาคารด้วยวิธีการเรียนรู้แบบกึ่งมีผู้สอน จากข้อความของการแสดงความคิดเห็น การขอความอนุเคราะห์ หรือการร้องเรียนจากลูกค้า ที่ได้จากการทำ Web scraping โดยใช้ชุดคำสั่ง Selenium ในการดึงข้อมูลจากเว็บไซต์ www.pantip.com ที่ได้ทำการดึงข้อมูล ณ วันที่ 13 กรกฎาคม 2565 ซึ่งมีจำนวนกระทู้ทั้งหมด 600 กระทู้ และใช้เทคนิคการประมวลผลภาษาธรรมชาติในการเตรียมความพร้อมของข้อมูล ในความเป็นจริงการที่จะระบุประเภทของผลิตภัณฑ์ธนาคารเป็นเรื่องที่ต้องใช้ทรัพยากรคน และเวลาจำนวนมาก ดังนั้นผู้วิจัยจึงได้แบ่งข้อมูลออกเป็นข้อมูลที่มีการระบุและไม่ระบุประเภทของผลิตภัณฑ์ธนาคารไว้ และใช้เทคนิคการเรียนรู้แบบกึ่งมีผู้สอนในการฝึกฝนแบบจำลอง และทำการทดลองโดยใช้เทคนิคการประมวลผลธรรมชาติ 3 แบบจำลอง ประกอบด้วย Support Vector Machine (SVM), Logistic Regression (LR) และ Naïve Bayes (NB) จากการทดลองพบว่าแบบจำลองที่มีประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารได้ดีที่สุดคือ SVM โดยมีความแม่นยำอยู่ที่ 0.82 โดยแบบจำลอง LR และ NB จะมีความแม่นยำอยู่ที่ 0.78 และ 0.74 ตามลำดับ

คำสำคัญ : การวิเคราะห์ข้อความ, การประมวลผลภาษาธรรมชาติ, เทคนิคการเรียนรู้แบบกึ่งมีผู้สอน

| | |
|---|---|
| Title | CLASSIFICATION OF BANKING PRODUCTS FROM THAI TEXT USING A SEMI-SUPERVISED LEARNING |
| Author | AMORNKARN JAIDEE |
| Degree | MASTER OF SCIENCE |
| Academic Year | 2022 |
| Thesis Advisor | Subhorn Khonthapagdee , Ph.D. |

The purpose of this research is to classify complaints about banking products using semi-supervised machine learning methods. In this research, customer complaints were obtained from Web scraping by using Selenium to retrieve data from www.pantip.com. The data was extracted as of July 13, 2022, with a total of 600 posts. Natural language processing techniques were used to prepare the data. Currently, all comments and complaints are primarily screened by humans. Consequently, there is a significant delay in this process. The experiments were performed using three models: Support Vector Machine (SVM), Logistic Regression (LR), and Naïve Bayes (NB). The best accuracy was achieved by SVM, with an accuracy of 0.82. The LR and NB models was having an accuracy of 0.78 and 0.74, respectively.

Keyword : Text Analytics; Natural Language Processing; Semi-supervised Learning

# กิตติกรรมประกาศ

สารนิพนธ์ฉบับนี้สำเร็จเรียบร้อยได้ด้วยดี เนื่องจากได้รับความอนุเคราะห์ช่วยเหลือจาก อาจารย์ ดร. ศุภร คนธภักดี ซึ่งเป็นอาจารย์ที่ปรึกษาสารนิพนธ์ ที่ให้คำปรึกษา คำแนะนำ และดูแลเอาใจใส่อย่างดีมาตลอด ทางผู้วิจัยรู้สึกซาบซึ้ง และขอขอบพระคุณเป็นอย่างสูงไว้ ณ โอกาสนี้

ขอขอบพระคุณคณาจารย์ทุกท่านในสาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ได้ให้ความรู้ คำแนะนำ ตลอดจนความช่วยเหลือให้การทำวิจัยนี้สำเร็จได้ด้วยดี

ขอขอบพระคุณบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ที่ให้ความช่วยเหลือ และให้คำแนะนำที่เป็นประโยชน์ในการทำวิจัยครั้งนี้

ขอขอบพระคุณเว็บไซต์ Pantip.com ที่ให้ความอนุเคราะห์ใช้ข้อมูลภายในเว็บไซต์ เพื่อใช้ในการดำเนินงานวิจัยครั้งนี้

ขอขอบพระคุณพี่ๆทุกท่านในฝ่ายระบบงานบริการลูกค้าหลัก และฝ่ายลูกค้าสัมพันธ์ ธนาคารออมสิน ที่ได้ให้ความช่วยเหลือ และคำแนะนำที่มีประโยชน์ในการทำวิจัยนี้ให้สำเร็จลุล่วงไปได้ด้วยดี และขอขอบพระคุณธนาคารออมสิน ที่ให้การสนับสนุนในการศึกษา ตลอดจนการทำงานเป็นอย่างดี

ขอขอบคุณ นายพรภวิษย์ สารบูรณ์ และ นางสาวณัชชา ชินนาพันธ์ ที่ให้คำปรึกษา และให้ความช่วยเหลือ ในการศึกษาและการทำวิจัยนี้ตลอดมา

ขอขอบคุณเพื่อนๆ สาขาวิชาวิทยาการข้อมูล คณะวิทยาศาสตร์ ที่ได้ให้ความช่วยเหลือและให้คำแนะนำที่เป็นเป็นประโยชน์ต่อผู้วิจัยตลอดมา

สุดท้ายนี้ ผู้วิจัยขอขอบพระคุณครอบครัว สำหรับแรงผลักดัน การให้กำลังใจ ให้ความช่วยเหลือและให้การสนับสนุนในด้านการศึกษาอย่างดีตลอดมา ขอขอบพระคุณเป็นอย่างสูง

อมรกานต์ ใจดี

# สารบัญ

หน้า

# สารบัญตาราง

**หน้า**

# สารบัญรูปภาพ

**หน้า**

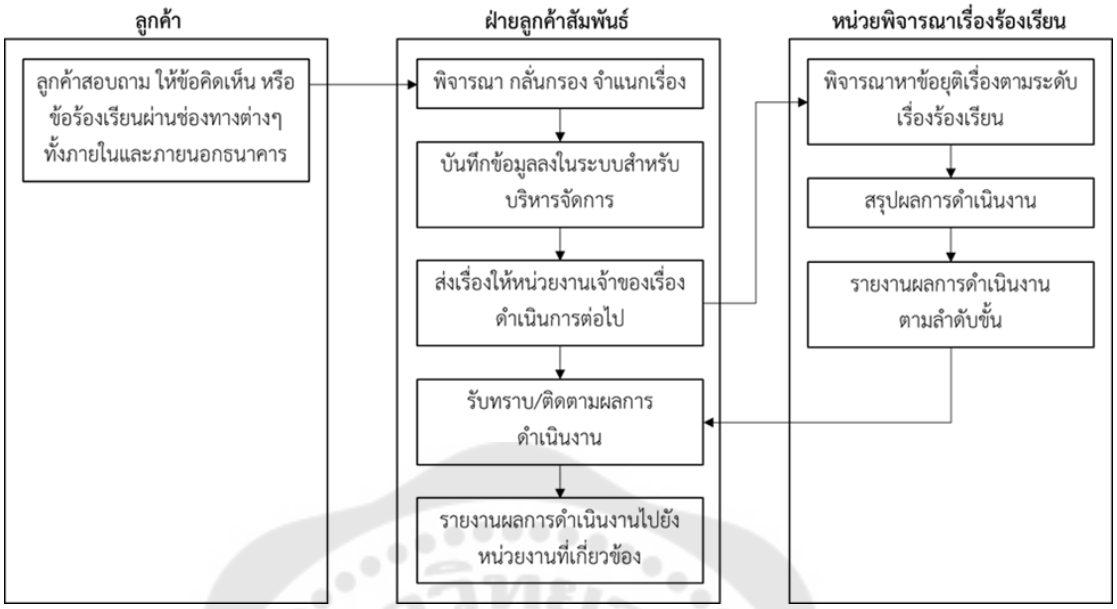
# บทที่ 1
# บทนำ

## 1. ภูมิหลัง

ปัจจุบัน ในทางการทำธุรกิจใดๆก็ตาม การรับฟังข้อคิดเห็น หรือข้อร้องเรียนจากลูกค้า ถือว่าเป็นเรื่องที่สำคัญอย่างยิ่ง ไม่ว่าจะเป็นในเรื่องของการแนะนำเพื่อให้องค์กรสามารถปรับตัวตามคำแนะนำได้อย่างทันท่วงที การแสดงความคิดเห็นสำหรับนำมาใช้ในการออกผลิตภัณฑ์ใหม่ๆ การขอความอนุเคราะห์ให้ดำเนินการในเรื่องต่างๆ หรือการร้องเรียนที่เกิดขึ้นจากการใช้บริการหรือผลิตภัณฑ์ของทางองค์กรธุรกิจ จากตัวอย่างดังกล่าว จะเห็นว่าข้อมูลทั้งหมดมีประโยชน์อย่างมากในการทำให้องค์กรประสบความสำเร็จด้านความพึงพอใจของลูกค้า และรักษาลูกค้าไว้ให้มีความสัมพันธ์กับองค์กรต่อไป

ในทางธุรกิจของธนาคารก็เช่นกัน ส่วนใหญ่องค์กรธุรกิจทางธนาคารจะมีระบบที่รับฟังความคิดเห็นต่างๆจากลูกค้า หรือมีช่องทางสำหรับแจ้งเรื่องร้องเรียน ขอความอนุเคราะห์ หรือแม้กระทั่งการชมเชยก็ตาม ซึ่งช่องทางส่วนใหญ่จะเป็นในรูปแบบ สาขา เว็บไซต์ คอลเซ็นเตอร์ หรืออีเมล เป็นต้น

ซึ่งระบบของธนาคารต่างๆ ส่วนมากจะไม่มีให้ลูกค้าสามารถกรอกในส่วนของประเภทผลิตภัณฑ์หรือบริการ ที่ต้องการแจ้งข้อมูลได้ จึงทำให้การที่จะทราบได้ว่าเป็นผลิตภัณฑ์หรือบริการใดที่ลูกค้าแจ้งเข้ามานั้น พนักงานที่ดูแลระบบจะต้องอ่านทุกข้อความ ถึงจะระบุได้ว่าเป็นผลิตภัณฑ์ในด้านใด หรือจำแนกว่าเกี่ยวข้องกับเรื่องใด ซึ่งขั้นตอนในส่วนนี้จะทำให้ใช้เวลาเป็นอย่างมาก และทำให้การที่จะส่งเรื่องไปยังหน่วยงานต่างๆ เพื่อดำเนินการในส่วนที่เกี่ยวข้องมีความล่าช้า จนอาจทำให้การแก้ไขปัญหาไม่ทันเวลา และไม่เป็นที่พึงพอใจต่อลูกค้าได้ ตามกระบวนการทำงานในภาพประกอบ 1




ภาพประกอบ 1 กระบวนการจัดการเรื่องร้องเรียนต่างๆ ของธนาคารออมสิน

ดังนั้น ในงานวิจัยนี้จึงมุ่งเน้นในการที่จะสามารถระบุผลิตภัณฑ์จากข้อความที่ลูกค้าให้ความเห็น หรือร้องเรียนเข้ามา เพื่อให้สามารถจัดการเรื่องเหล่านี้ได้อย่างรวดเร็วยิ่งขึ้น โดยใช้การเรียนรู้ของเครื่อง (Machine Learning Model) ในการวิเคราะห์ข้อความ เนื่องจากข้อมูลบางส่วนที่ไม่ได้มีการระบุประเภทของผลิตภัณฑ์ไว้ ทางผู้วิจัยจึงใช้การเรียนรู้ของเครื่องแบบ “การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning)” ในการจำแนกประเภทของผลิตภัณฑ์ธนาคาร รวมถึงการใช้เทคนิคที่จะหาคำสำคัญที่จะพบบ่อยในผลิตภัณฑ์นั้นๆ ซึ่งโมเดลที่ได้จะมีการเรียนรู้และทดสอบกับข้อมูลจริง รวมถึงการประเมินผลความถูกต้องแม่นยำ โดยใช้เทคนิคการวัดผลที่เป็นมาตรฐาน

โดยในงานวิจัยนี้จะใช้ข้อมูลจากการทำ Web scraping และใช้ชุดคำสั่ง Selenium ในการดึงข้อมูลจากเว็บไซต์ www.pantip.com ที่ถือเป็นอีกหนึ่งช่องทางให้ลูกค้าสามารถตั้งกระทู้สอบถาม ให้ข้อคิดเห็น หรือข้อร้องเรียนได้

ซึ่งทางผู้วิจัยได้เลือกตัวอย่างข้อมูลฯ เป็นของธนาคารออมสิน ที่ทำการดึงข้อมูล ณ วันที่ 13 กรกฎาคม 2565 จำนวน 600 กระทู้ โดยแบ่งประเภทของผลิตภัณฑ์ธนาคาร จำนวน 6 ประเภท ดังนี้

1. สินเชื่อ เป็นธุรกรรมทางการเงิน ที่ธนาคารอนุมัติให้แก่บุคคลธรรมดาหรือนิติบุคคลกู้ยืมเพื่อวัตถุประสงค์ต่างๆ โดยมีเงื่อนไขให้ชำระหนี้สินครบตามกำหนดในระยะเวลาของสัญญา

2. เงินฝาก เป็นบัญชีธนาคารที่ดูแลโดยสถาบันการเงิน ซึ่งลูกค้าสามารถฝากและถอนเงินได้ บัญชีเงินฝากอาจเป็นบัญชีออมทรัพย์ บัญชีกระแสรายวัน หรือบัญชีประเภทอื่น ๆ

3. สลากออมสิน เป็นประเภทการออมทรัพย์ที่ผู้ฝากจะได้รับดอกเบี้ยควบคู่กับสิทธิได้รับรางวัลในแต่ละงวดและจะต้องถือครองจนครบกำหนด จึงจะได้รับดอกเบี้ยตามที่ระบุไว้

4. บัตรเดบิต/เครดิต

1) บัตรเดบิต เป็นผลิตภัณฑ์ธนาคารประเภทบัตรที่ผูกไว้กับบัญชีเงินฝากของผู้ถือบัตรเพื่อใช้ทำรายการที่เครื่องถอนเงินอัตโนมัติ (ATM) และชำระค่าสินค้าและบริการ รวมถึงการซื้อสินค้าออนไลน์ โดยจะเป็นการหักเงินจากบัญชีเงินฝากทันที
2) บัตรเครดิต เป็นบริการบัตรที่ออกให้แก่ลูกค้าเพื่อใช้จ่ายแทนเงินสด และสามารถใช้ได้ตามจำนวนวงเงินบัตรที่อนุมัติ

5. แอปพลิเคชัน MyMo เป็นระบบ Mobile Banking สำหรับธนาคารออมสิน ที่ให้บริการทำธุรกรรมทางการเงินผ่านโทรศัพท์มือถือ

6. อื่นๆ เป็นเรื่องที่นอกเหนือจากผลิตภัณฑ์ธนาคารดังกล่าว เช่น การสมัครงาน การค้นหาที่ตั้งของธนาคาร เป็นต้น

**2. ความมุ่งหมายของงานวิจัย**

ในการวิจัยครั้งนี้ผู้วิจัยได้ตั้งความมุ่งหมายไว้ดังนี้

1. เพื่อให้สามารถจำแนกประเภทของผลิตภัณฑ์ จำนวน 6 ประเภท จากข้อความของการแสดงความคิดเห็น การขอความอนุเคราะห์ หรือการร้องเรียนจากลูกค้า โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบ Semi-Supervised Learning

2. เพื่อให้สามารถเปรียบเทียบประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารระหว่างแบบจำลอง Support Vector Machine (SVM), Naïve Bayes และ Logistic Regression

**3. ขอบเขตของการวิจัย**

**ประชากรที่ใช้ในการวิจัย**

ข้อความจากความคิดเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ และข้อร้องเรียน จากเว็บไซต์ Pantip ที่ทำการเก็บข้อมูล ณ วันที่ 13 กรกฎาคม 2565 จำนวน 600 ข้อความ

**กลุ่มตัวอย่างที่ใช้ในการวิจัย**

ข้อความจากความคิดเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ และข้อร้องเรียน ที่ตัวอย่างข้อมูลเป็นของธนาคารออมสิน จำนวน 600 ข้อความ

**ตัวแปรที่ศึกษา**

ตัวแปรอิสระ ที่เป็นข้อความจากความคิดเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ และข้อร้องเรียน

**4. กรอบแนวคิดในงานวิจัย**

งานวิจัยนี้มุ่งเน้นในการที่จะสามารถจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความที่ลูกค้าให้ความเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ หรือร้องเรียนเข้ามา ซึ่งเป็นข้อความภาษาไทย โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning Model) ในการวิเคราะห์ข้อความ เนื่องจากข้อมูลบางส่วนที่ไม่ได้มีการระบุประเภทของผลิตภัณฑ์ไว้ ทางผู้วิจัยจึงใช้การเรียนรู้ของเครื่องแบบ “การเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning)” ในการจำแนกประเภทของผลิตภัณฑ์ธนาคาร โดยใช้อัลกอริทึม ได้แก่ Support Vector Machine (SVM) Naïve Bayes และ Logistic Regression โดยมีการเตรียมความพร้อมของข้อมูลที่เป็นข้อความที่ใช้การแบ่งประโยค (Sentence tokenize) การตัดคำ (Word tokenize) การลบคำที่ไม่สื่อความหมายของประโยค (Remove Stop word) รวมถึงการใช้เทคนิคที่จะหาคำสำคัญที่จะพบบ่อยในผลิตภัณฑ์นั้นๆ และนำมาเป็นคุณสมบัติสำหรับใช้ในการจำแนกฯ ได้แก่ การคัดแยกคำ

ตามความสำคัญ (TF-IDF) การสร้างกลุ่มคำที่จะแสดงขนาดตามจำนวนความถี่ของคำ (Word Cloud) และมีการประเมินผลของแบบจำลองด้วยเทคนิคการวัดผลที่เป็นมาตรฐาน เช่น Accuracy, Precision, Recall และ F1 Score

**5. สมมติฐานงานวิจัย**

การสร้างแบบจำลองที่ใช้เทคนิคการเรียนรู้แบบกึ่งมีผู้สอน (Semi-Supervised Learning) สามารถจำแนกประเภทของผลิตภัณฑ์ธนาคาร จากข้อความภาษาไทยได้ โดยผู้วิจัยได้แบ่งประเภทของผลิตภัณฑ์เป็น 6 กลุ่มใหญ่ๆ ดังนี้

1. สินเชื่อ
2. เงินฝาก
3. สลากออมสิน
4. บัตรเดบิต/เครดิต
5. แอปพลิเคชัน MyMo
6. อื่นๆ

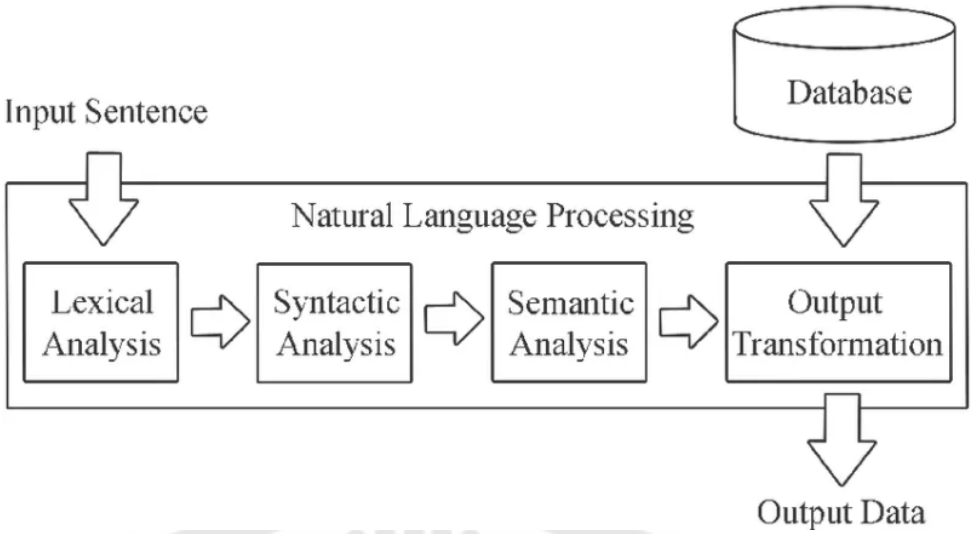
# บทที่ 2
# ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้องกับการจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความ และได้นำเสนอตามหัวข้อต่อไปนี้

1. การประมวลผลภาษาธรรมชาติ
2. การเตรียมความพร้อมของข้อมูลภาษาไทย
3. เทคนิคการเรียนรู้ของเครื่อง
4. แบบจำลองที่ใช้ในงานวิจัย
5. งานวิจัยที่เกี่ยวข้อง

## 1. การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ หรือ Natural language processing (NLP) เป็นการผสมผสานเทคโนโลยีปัญญาประดิษฐ์ (Artificial Intelligence : AI) วิทยาการคอมพิวเตอร์ และภาษาศาสตร์เชิงคำนวณเข้าด้วยกัน เพื่อช่วยให้คอมพิวเตอร์สามารถตีความ จัดการ และทำความเข้าใจภาษามนุษย์ได้ ซึ่งในโลกปัจจุบันมีข้อมูลเสียงและข้อความจำนวนมาก ที่มาจากช่องทางการสื่อสารต่างๆ เช่น อีเมล ข้อความ ฟีดข่าว โซเชียลมีเดีย วิดีโอ เสียง และอื่นๆ โดยสามารถใช้ NLP เพื่อประมวลผลข้อมูลนี้ได้แบบอัตโนมัติ วิเคราะห์เจตนาหรือความรู้สึกในข้อความ และตอบสนองต่อการสื่อสารของมนุษย์ได้อย่างทันที NLP จะแบ่งข้อความหรือคำพูดของมนุษย์ออกเป็นส่วนย่อยๆ ที่โปรแกรมคอมพิวเตอร์สามารถเข้าใจได้ง่าย ตัวอย่างเช่น การใช้ NLP เข้ามาเป็นส่วนหนึ่งของการบริการลูกค้า ในรูปแบบของระบบการสนับสนุนลูกค้าแบบอัตโนมัติโดยใช้แชทบอท (Chatbot) เป็นการรู้จำคำพูดของมนุษย์ผ่าน NLP ที่ช่วยให้สามารถตอบคำถามของผู้ใช้งานแบบทันท่วงที และจัดการปัญหาได้อย่างเหมาะสม รวมถึงสามารถลดต้นทุนและเวลาในการบริการลูกค้าลงได้กระบวนการทำงานของ NLP มี 4 ขั้นตอนหลักตามภาพประกอบ 2




ภาพประกอบ 2 กระบวนการทำงานของ NLP

ที่มา : https://research.aimultiple.com/nlp/

1) การวิเคราะห์คำศัพท์ (Lexical Analysis) เป็นกระบวนการแบ่งประโยคออกเป็นคำ เพื่อระบุความหมายของประโยคและความสัมพันธ์กับประโยคทั้งหมด

2) การวิเคราะห์ในเชิงโครงสร้าง (Syntactic Analysis) เป็นกระบวนการตรวจสอบการวางตำแหน่งของกลุ่มคำ และระบุความสัมพันธ์ระหว่างคำและกลุ่มคำต่างๆ ภายในประโยค

3) การวิเคราะห์ในเชิงความหมาย (Semantic Analysis) เป็นกระบวนการตรวจสอบความถูกต้องในเชิงความหมายของประโยค โดยประโยคที่วางกลุ่มคำชนิดต่างๆ ตามโครงสร้างไวยากรณ์ จะมีความหมาย

4) การสร้างผลลัพธ์ (Output Transformation) เป็นกระบวนการที่สร้างผลลัพธ์จากการวิเคราะห์ความหมายของข้อความหรือคำพูดที่เหมาะสมกับเป้าหมายที่ต้องการ

**2. การเตรียมความพร้อมของข้อมูลภาษาไทย**

การทำความสะอาดข้อมูลเป็นขั้นตอนที่สำคัญ เนื่องจากเป็นการทำให้ข้อมูลมีความพร้อมที่จะใช้กับโมเดล ซึ่งข้อมูลที่เป็นข้อความส่วนใหญ่จะมีสัญญาณรบกวนในรูปแบบต่างๆ เช่น อีโมติคอน (Emoticon) สัญลักษณ์พิเศษต่างๆ หรือคำที่ไม่สื่อ

ความหมาย เป็นต้น โดยทางผู้วิจัยได้ใช้ไลบรารี่ PyThaiNLP ในการทำความสะอาดข้อมูลที่เป็นภาษาไทย ซึ่งใช้เทคนิคดังต่อไปนี้

2.1 การแบ่งประโยค (Sentence tokenize) เป็นการแบ่งประโยคออกจากข้อความที่มีความยาว โดยใช้อัลกอริทึม crfcut ในการตัดคำและกำหนด label สำหรับแต่ละคำ ว่ามีคำใดเป็นคำที่อยู่ในประโยค และคำที่สิ้นสุดประโยค

2.2 การตัดคำ (Word tokenize) เป็นการตัดคำออกจากประโยค โดยใช้อัลกอริทึมหลักของ PyThaiNLP คือ newmm จะทำการจับคู่กับคำในพจนานุกรมภาษาไทย และแยกเป็นรายการคำ

2.3 การลบคำที่ไม่สื่อความหมายของประโยค (Remove Stop word) เป็นการลบคำที่เกิดขึ้นบ่อยครั้ง และไม่ได้มีส่วนสำคัญต่อความหมายโดยรวมของประโยค ตัวอย่างเช่น มี, การ, ความ เป็นต้น

2.4 การคัดแยกคำตามความสำคัญ (TF-IDF) เป็นเทคนิคการคัดแยกคำตามความสำคัญโดยการให้น้ำหนักคำในแต่ละคำ ซึ่งใช้ 2 ปัจจัยคือค่า TF และ IDF นำโดยมาคูณกันและได้ Output ออกมาเป็น Vector ที่เป็นค่าต่อเนื่อง ซึ่งมีรายละเอียดดังนี้

2.4.1 Term Frequency (TF) เป็นการนับแต่ละคำ และหารด้วยจำนวนคำ เพื่อให้ได้เปอร์เซ็นต์น้ำหนักของแต่ละคำที่มีในประโยค (Document) นั้นๆ

2.4.2 Inverse Document Frequency (IDF) เป็นการนับว่าแต่ละคำปรากฏทั้งหมดใน Document จำนวนเท่าไร หากปรากฏใน Document จำนวนมาก จะให้ค่าน้ำหนักที่น้อย เนื่องจาก IDF จะใช้ log เพื่อกำหนดช่วงค่าของน้ำหนักให้ต่ำลงมา (Scale down)

2.5 การสร้างกลุ่มคำ (Word Cloud) เป็นการ Visualize ข้อมูลแบบหนึ่ง โดยการจับกลุ่มของคำที่ปรากฏในจำนวนที่แตกต่างกัน ยิ่งคำปรากฏใหญ่ขึ้นและโดดเด่นขึ้นแสดงว่ามีการกล่าวถึงบ่อยและมีความสำคัญในข้อมูลนั้นๆ

**3. เทคนิคการเรียนรู้ของเครื่อง**

การเรียนรู้ของเครื่อง (Machine Learning) คือการนำคอมพิวเตอร์ มาเรียนรู้จากข้อมูลกลุ่มตัวอย่างจำนวนมาก เป็นการทำงานในรูปแบบของการจดจำข้อมูลและการทำนายเมื่อมีข้อมูลใหม่เข้ามา โดยใช้คุณลักษณะ (Feature) ในการแบ่งแยกข้อมูลที่สนใจกับข้อมูลอื่นๆ ออกจากกัน การเรียนรู้ของเครื่องแบ่งออกเป็น 4 ประเภทหลัก ดังต่อไปนี้

3.1 การเรียนรู้แบบมีผู้สอน (Supervised learning) เป็นการให้คอมพิวเตอร์เรียนรู้จากข้อมูลที่มีการระบุประเภทข้อมูล (Labeled data) เรียบร้อยแล้ว เพื่อให้สามารถทำนายผลลัพธ์ (Output) ที่จะเกิดขึ้นได้ การเรียนรู้แบบมีผู้สอนแบ่งออกเป็น 2 ประเภท ดังนี้

3.1.1 การจำแนกประเภทข้อมูล (Classification) เป็นอัลกอริทึมที่ใช้เพื่อแก้ปัญหาการจำแนกประเภทที่ Output เป็นข้อมูลที่จัดเป็นกลุ่ม (Categorical data) ตัวอย่างเช่น การทำนายการหมุนเปลี่ยนของพนักงาน (Employee turnover), การทำนาย Spam mail เป็นต้น

3.1.2 การวิเคราะห์การถดถอย (Regression) เป็นอัลกอริทึมที่ใช้เพื่อแก้ปัญหาการถดถอยซึ่งมีความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปร Input และ Output ที่เป็นค่าต่อเนื่อง (Continuous value) ตัวอย่างเช่น การทำนายปริมาณน้ำฝน, การทำนายค่าฝุ่น PM 2.5, การทำนายราคาบ้าน เป็นต้น

3.2 การเรียนรู้แบบไม่มีผู้สอน (Unsupervised learning) จะตรงกันข้ามกับ Supervised learning คือการให้คอมพิวเตอร์เรียนรู้จากข้อมูลที่ไม่มีการระบุประเภทข้อมูลไว้ และทำนายผลลัพธ์ออกมา เป้าหมายหลักคือการจัดกลุ่มหรือจัดหมวดหมู่ชุดข้อมูล การเรียนรู้แบบไม่มีผู้สอนแบ่งออกเป็น 2 ประเภท ดังนี้

3.2.1 การแบ่งกลุ่มข้อมูล (Clustering) คือ วิธีการจัดกลุ่มข้อมูลให้เป็นกลุ่มเพื่อให้ข้อมูลที่มีความคล้ายคลึงกันมากที่สุดยังคงอยู่ในกลุ่มเดียวกัน หากมีความคล้ายคลึงกันน้อยลงหรือไม่มีเลยก็จะไปอยู่ภายใต้กลุ่มอื่นๆ ตัวอย่างเช่น การจัด

กลุ่มของลูกค้าตามพฤติกรรมในการซื้อสินค้า หรือการแบ่งกลุ่มผลิตภัณฑ์จากการขายสินค้า เป็นต้น

3.2.2 การหาความสัมพันธ์ของข้อมูล (Association) คือ การค้นหาความสัมพันธ์ระหว่างตัวแปรภายในชุดข้อมูล (Dataset) จุดมุ่งหมายหลักคือการค้นหาความเกี่ยวข้องกันของระหว่างข้อมูลหนึ่งกับข้อมูลอื่นๆ เพื่อทำการจับกลุ่มข้อมูลเหล่านั้น เพื่อนำไปใช้ให้เกิดประโยชน์มากที่สุด ตัวอย่างเช่น การหาสินค้าที่ลูกค้ามักจะซื้อพร้อมกัน หรือการหาสินค้าที่ลูกค้าซื้อแล้ว จะต้องซื้อสิ่งอื่นๆตามมาด้วย เพื่อใช้ในการจัดโปรโมชั่น หรือการจัดเรียงสินค้าให้เหมาะสมกับความต้องการ

3.3 การเรียนรู้แบบกึ่งมีผู้สอน (Semi-supervised learning) เป็นการเรียนรู้ของเครื่องที่ใช้ทั้งเทคนิค Supervised learning (Labeled data) และ Unsupervised learning (Unlabeled data) โดยจะใช้ในกรณีที่ Labeled data มีไม่เพียงพอ จึงทำให้ข้อมูลส่วนใหญ่จะเป็นข้อมูลแบบ Unlabeled data ข้อดีก็คือสามารถลดระยะเวลาที่จะต้อง labeled ข้อมูลทั้งหมดซึ่งมีจำนวนมาก โดยมีวิธีการทำงานดังต่อไปนี้

ขั้นตอนที่ 1 ฝึกฝนโมเดลด้วย training data ที่มีจำนวนน้อย ซึ่งจะ train จนกว่าโมเดลจะมีผลลัพธ์ที่แม่นยำ

ขั้นตอนที่ 2 อัลกอริทึมนี้จะถูกใช้ทำนาย Unlabeled data ซึ่งเรียกว่าการทำ Pseudo label data โดยจะกำหนดค่า Confidence level ที่ยอมรับได้ หากข้อมูลมีค่าเกินกว่าที่กำหนด จึงจะเพิ่มลงในชุดข้อมูล Labeled data

ขั้นตอนที่ 3 Labeled data และ Pseudo label data จะเชื่อมโยงเข้าด้วยกันเป็น Dataset ใหม่ และถูกนำมา train โมเดลอีกครั้ง โดยเป็นการทำซ้ำหลายๆ รอบ หากข้อมูลมีความเหมาะสมกับกระบวนการ ประสิทธิภาพของโมเดลจะเพิ่มขึ้นเรื่อยๆ ในการทำซ้ำแต่ละครั้ง

3.4 การเรียนรู้แบบเสริมแรง (Reinforcement Learning) เป็นการเรียนรู้ของเครื่องที่ทำการ train โมเดล โดยจะใช้การตัดสินใจแบบเรียงลำดับเพื่อให้ได้ผลลัพธ์ที่เหมาะสมที่สุดสำหรับการแก้ปัญหานั้นๆ ในการเรียนรู้แบบเสริมแรง จะไม่มีข้อมูลที่มี

การระบุประเภทข้อมูล (Labeled data) เหมือนการเรียนรู้แบบมีผู้สอน และตัวแทน (Agent) จะเรียนรู้จากประสบการณ์เท่านั้น ตัวอย่างเช่น การสร้างเกมส์ หรือการสร้างหุ่นยนต์ เป็นต้น

**4. แบบจำลองที่ใช้ในงานวิจัย**

4.1 Support Vector Machine (SVM) เป็นแบบจำลองที่มุ่งเน้นที่จะทำให้พื้นที่เส้นแบ่งระหว่างประเภทของข้อมูล (Maximal margin classification) มีขนาดใหญ่มากที่สุด เพื่อที่จะลดผลเสียที่จะเกิดขึ้นจากการทายข้อมูลผิดประเภท (Variance) ตัวอย่างเช่น การที่จุดข้อมูลสองจุดอยู่ใกล้เคียงกัน ควรจะถูกจัดกลุ่มเป็นประเภทเดียวกัน ถ้าหากมี Decision boundary แบ่งระหว่างจุดข้อมูลนี้จะทำให้เกิด variance เพิ่มมากขึ้น โดย SVM จะใช้ Loss Function เป็น Hinge loss โดยมีจุดมุ่งหมายคือ ลดข้อผิดพลาดของแบบจำลอง และเพิ่มขนาดของพื้นที่เส้นแบ่งระหว่างประเภทของข้อมูล รายละเอียดตามสมการ ดังนี้

$$arg\min_{w,b} C\sum_{i=1}^{n}\max\{0,1-y_i(w^T x_i+b)\}+||w||_2^2$$

4.2 Naïve Bayes เป็นแบบจำลองที่จำแนกประเภทของข้อมูลโดยใช้หลักความน่าจะเป็น โดยจะใช้เมื่อแต่ละ Features เป็นอิสระต่อกัน ตามทฤษฎีบทของ Bayes ดังสมการต่อไปนี้

$$P(y|x)=\frac{P(x|y).P(y)}{P(x)}$$

$P(y|x)$ : ความน่าจะเป็นที่จะเกิดเหตุการณ์ y เมื่อเกิดเหตุการณ์ x ขึ้นก่อน

$P(x|y)$ : ความน่าจะเป็นที่จะเกิดเหตุการณ์ x เมื่อเกิดเหตุการณ์ y ขึ้นก่อน

$P(x)$ : ความน่าจะเป็นที่จะเกิดเหตุการณ์ x

$P(y)$ : ความน่าจะเป็นที่จะเกิดเหตุการณ์ y

4.3 Logistic Regression เป็นแบบจำลองที่จำแนกประเภทของข้อมูล ให้ผลลัพธ์ออกมาเป็นค่าระหว่าง 0-1 (ความน่าจะเป็น) ได้ โดยการใช้ Sigmoid function (Logistic function) ตามสมการดังต่อไปนี้

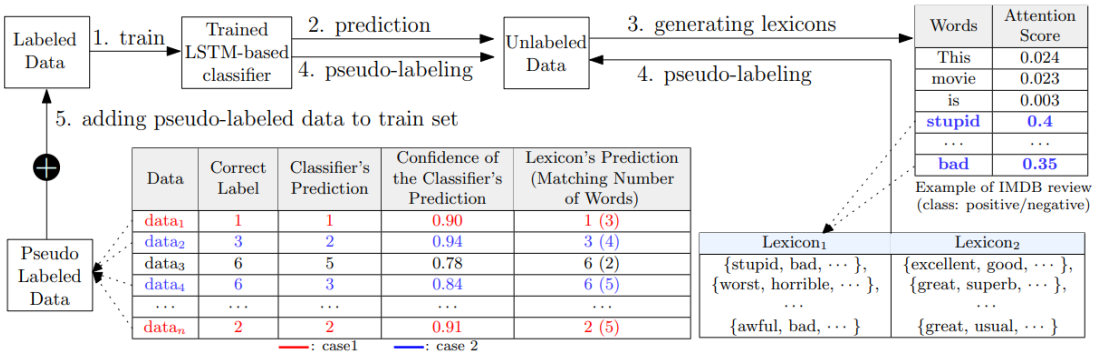
$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

**5. งานวิจัยที่เกี่ยวข้อง**

ในงานวิจัยนี้ได้ทบทวนวรรณกรรมและศึกษางานวิจัยที่มุ่งเน้นการอธิบายการจำแนกประเภทของข้อความ ด้วยเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน และเทคนิคในการเตรียมความพร้อมของข้อมูลที่เป็นข้อความ โดยมีงานวิจัยที่เกี่ยวข้อง ดังต่อไปนี้

5.1 บทความวิจัยเรื่อง SALNet: Semi-supervised Few-Shot Text Classification with Attention-based Lexicon Construction (Lee et al., 2021)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของข้อความ ด้วยเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน และการสร้างคลังคำศัพท์จาก Attention mechanism (SALNet) โดยใช้แบบจำลอง Long Short-Term Memory (LSTM) ซึ่งเป็นโครงข่ายประสาทเทียมประเภท Recurrent Neural Network (RNN) ที่ใช้สำหรับประมวลผล ทำนาย และจำแนกข้อมูลแบบ Time-series ซึ่งมีกระบวนการดำเนินงานดังภาพประกอบที่ 3




ภาพประกอบ 3 แสดงกระบวนการดำเนินงานวิจัยของบทความ

ที่มา : (Lee et al., 2021)

1. สร้างแบบจำลอง LSTM จากข้อมูลที่มีการระบุประเภท (Labeled Data) ซึ่งมีจำนวนน้อย โดยแบบจำลองจะต้องประกอบด้วย Attention Mechanism เพื่อเก็บคำสำคัญสำหรับข้อมูลแต่ละประเภท

2. Train แบบจำลองที่มี Attention Mechanism อีกครั้ง กับข้อมูลที่ไม่ได้ระบุประเภท (Unlabeled Data)

3. จะได้ set ของคำสำคัญออกมา ซึ่งจะเรียกว่า Lexicon หรือก็คือการที่เราใช้ชุดข้อมูลและ Attention score สร้าง lexicon ที่ประกอบด้วยคำสำคัญๆไว้

4. ทำนายประเภทของข้อมูลที่อยู่ในชุดข้อมูล โดยใช้แบบจำลองที่ผ่านการ train มาแล้ว และ lexicon

5. เพิ่มข้อมูลใหม่ที่มีการระบุประเภทไว้ ในชุดข้อมูล training และ train แบบจำลองอีกครั้ง โดยเริ่มจากขั้นตอนแรกใหม่

6. ทำขั้นตอนข้างต้นซ้ำจนกว่าจะไม่มีการเพิ่มข้อมูลที่ถูกระบุประเภทปลอม (Pseudo label)ในชุดข้อมูล training set อีกต่อไป โดยในการ train แบบจำลองแต่ละครั้งจะใช้ชุดข้อมูลที่ทำการปรับปรุงแล้ว

จากผลการศึกษาพบว่าการใช้อัลกอริทึม SALNet (attention-based LSTM + BERT) ให้ประสิทธิภาพดีกว่าการใช้ BERT แบบดั้งเดิม

5.2 บทความวิจัยเรื่อง Semi supervised Text Classification Using Unsupervised Topic Information (Dorado & Ratte, 2016)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของข้อความ ด้วยเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน โดยใช้ข้อมูลหัวข้อที่ไม่ได้มีการระบุประเภทข้อมูล โดยมีการใช้อัลกอริทึม การจัดสรรดีรีเคลแฝง (Latent Dirichlet Allocation : LDA) ในการจัดหมวดหมู่ข้อความ และใช้แบบจำลอง Naïve Bayes และ Support Vector Machine (SVM) ในการเพิ่มจำนวนข้อมูล (Augment) โดยใช้คีย์เวิร์ดที่ได้จากการทำ LDA เป็น feature และศึกษาขนาดของข้อมูลที่มีการระบุประเภทไว้ ซึ่งจำเป็นในการฝึกฝนโมเดลในการจำแนกประเภทต่างๆ โดยใช้ข้อมูลทั้งที่มีการระบุและไม่ได้ระบุประเภทของข้อมูล และข้อมูลที่เพิ่มเข้ามาใหม่ จากผลการศึกษาพบว่าการใช้มูลเพียง 3% (20 ตัวอย่างต่อหมวดหมู่) ของชุด

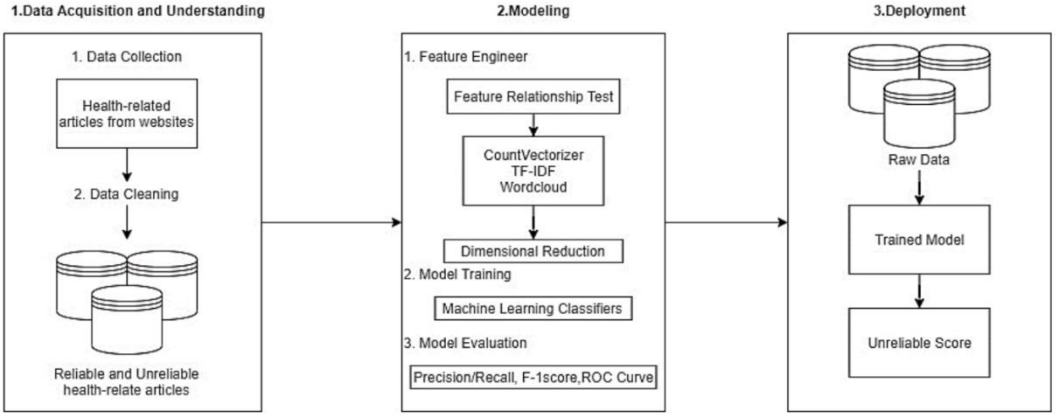
ข้อมูล 600 ตัวอย่าง สามารถให้ accuracy ได้ถึง 80% โดยใช้อัลกอริทึม SVM และ AUG

5.3 บทความวิจัยเรื่อง Text Classification Based On Semi-Supervised Learning (Thanh et al., 2013)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของข้อความ ด้วยเทคนิคการเรียนรู้แบบกึ่งมีผู้สอน โดยใช้แบบจำลอง SVM และสร้างแบบจำลอง Feature ที่ใช้ข้อมูลที่มีการระบุประเภท (Labeled data) และปรับปรุงแบบจำลองให้ดีขึ้น ด้วยข้อมูลที่ไม่ได้มีการระบุประเภท (Unlabeled data) ซึ่งทดลองโดยใช้ training data จำนวน 600 ข้อความ และทดสอบจำนวน 10 ครั้ง ในแต่ละครั้งจะสุ่มเลือกข้อมูลที่ไม่ได้มีการระบุประเภท จำนวน 10 ข้อความ และจะเพิ่มขนาดของ training data ด้วยข้อมูลที่ไม่ได้มีการระบุประเภท จำนวนตั้งแต่ 100 – 600 ข้อความ จากผลการศึกษาพบว่าคุณภาพของระบบจะดีขึ้นเมื่อเพิ่มจำนวนของข้อมูลที่ไม่ได้มีการระบุประเภท

5.4 บทความวิจัยเรื่อง Detection of Unreliable Medical Articles on Thai Websites (Saengkhunthod et al., 2021)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของข้อความในบทความบนเว็บไซต์ เพื่อตรวจจับบทความทางการแพทย์ที่หลอกลวง ด้วยเทคนิคการเรียนรู้ของเครื่อง มีการใช้แบบจำลอง XGBoost, Decision tree, SVM, Logistic Regression และ k-NN และมีการเปรียบเทียบสิทธิภาพระหว่าง Model ต่างๆ โดยงานวิจัยมีขั้นตอนการดำเนินงานดังภาพประกอบ 4




ภาพประกอบ 4 แสดงวิธีการดำเนินการวิจัย

ที่มา : (Saengkhunthod et al., 2021)

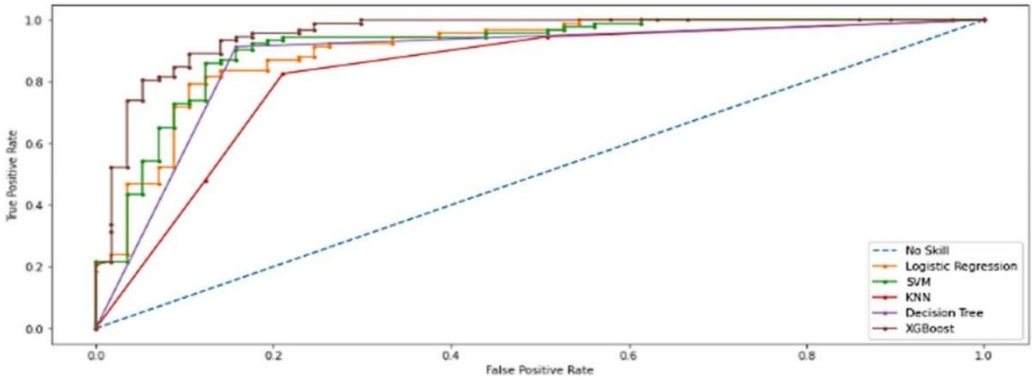
โดยการดำเนินงานจะแบ่งออกเป็นสามส่วนหลักๆ คือ Data Acquisition, Modeling, Deployment

1. การเก็บข้อมูล (Data Acquisition) จะประกอบไปด้วยการเก็บข้อมูลบทความทางการแพทย์จากเว็บไซต์ต่าง ซึ่งการเก็บตัวอย่างของข้อมูลจะใช้เทคนิคการดึงข้อมูลจากเว็บไซต์ (Web Scraping) และใช้โมดูล Beautiful Soup ในการเก็บข้อมูลจากเว็บไซต์ต่างๆ ข้อมูลที่ได้รับมาจึงถูกแบ่งออกเป็น 2 ประเภท คือ ข้อมูลจริง และข้อมูลเท็จ และนำมา ทำความสะอาดข้อมูล

2. การสร้างแบบจำลอง (Modeling) จะประกอบไปด้วยการทำ features relationship test โดยใช้เทคนิค Chisquare test เพื่อดูความสัมพันธ์ระหว่าง Features ที่เลือกมาและความน่าเชื่อถือของบทความ (Target), มีการใช้เทคนิค TF-IDF ,การสร้าง Word cloud ที่สามารถตรวจจับคำที่แสดงให้เห็นถึงคำที่พบบ่อยที่สุดในบทความที่น่าเชื่อถือ , การทำ Dimensional reduction เพื่อลดขนาดของข้อมูลเป็น 2 มิติ โดยใช้ t-SNE ที่แสดงให้เห็นการจัดกลุ่มของข้อมูล ,การฝึกฝนแบบจำลอง และการประสิทธิภาพของแบบจำลองโดยใช้ Precision, Recall, F1 Score และ ROC curve

3. การปรับใช้งาน (Deployment) เป็นการนำแบบจำลองที่ได้รับการฝึกฝนมาจำแนกบทความความที่หลอกลวง

จากผลการศึกษาพบว่า แบบจำลองที่แม่นยำที่สุด คือ XGBoost โดยมีความแม่นยำ (Accuracy) อยู่ที่ 90.60% และค่าอื่นๆก็ยังคงสูงที่สุดจากโมเดลทั้งหมด รองลงมาก็จะเป็น Decision Tree และใน ROC Curve จะสังเกตเห็นว่ายิ่งเข้าใกล้ 1 curve ของ XGBoost และ Decision Tree มีประสิทธิภาพที่ใกล้เคียงกัน แต่ยังเป็น XGBoost ที่มีประสิทธิภาพดีที่สุดจากโมเดลทั้งหมด รายละเอียดตามภาพประกอบที่ 5




ภาพประกอบ 5 แสดง Roc curve ที่วัดประสิทธิภาพของแบบจำลอง

ที่มา : (Saengkhunthod et al., 2021)

5.5 บทความวิจัยเรื่อง Article Classification Using Natural Language Processing and Machine Learning (Dien et al., 2019)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของบทความ ด้วยเทคนิคการประมวลผลภาษาธรรมชาติ และเทคนิคการเรียนรู้ของเครื่อง มีการใช้แบบจำลอง SVM, Naïve Bayes และ k-NN โดยมีการเปรียบเทียบสิทธิภาพระหว่าง Model ต่างๆ และมีการเตรียมความพร้อมของข้อมูลที่เป็นข้อความ ได้แก่ การแยกคำ (Word segmentation), การลบคำที่ไม่สื่อความหมายของประโยค (Remove Stop word) รวมถึงมีการใช้เทคนิคที่จะหาคำสำคัญที่จะพบบ่อยบทความนั้นๆ และนำมาเป็นคุณสมบัติสำหรับใช้ในการจำแนกฯ ได้แก่ การคัดแยกคำตามความสำคัญ (TF-IDF) และมีการประเมินประสิทธิภาพของแบบจำลองโดยใช้ Precision, Recall, F1 Score จากผลการศึกษาพบว่า แบบจำลอง SVM มีประสิทธิภาพในการจำแนกประเภทของบทความที่ดีที่สุด

โดยมีความแม่นยำเฉลี่ย มากกว่า 91% จึงมีความเป็นไปได้ในการแบบจำลองมาพัฒนาระบบการจำแนกประเภทของบทความแบบอัตโนมัติ ดังภาพประกอบ 6

| | SVM | | | Naïve Bayes | | | kNN | | |
|---|---|---|---|---|---|---|---|---|---|
| Topics | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ | Precision | Recall | $F_1$ |
| Technology | 0.857 | 0.857 | 0.857 | 0.857 | 0.857 | 0.857 | 0.400 | 0.571 | 0.471 |
| Environment | 1.000 | 0.333 | 0.500 | 0.400 | 0.333 | 0.364 | 0.667 | 0.333 | 0.444 |
| Natural Sciences | 0.750 | 1.000 | 0.857 | 0.667 | 0.667 | 0.667 | 0.600 | 1.000 | 0.750 |
| Animal husbandry | 1.000 | 1.000 | 1.000 | 1.000 | 0.500 | 0.667 | 1.000 | 0.500 | 0.667 |
| Biotechnology | 1.000 | 0.500 | 0.667 | 1.000 | 0.500 | 0.667 | 1.000 | 0.500 | 0.667 |
| Agriculture | 0.786 | 1.000 | 0.880 | 0.846 | 1.000 | 0.917 | 0.733 | 1.000 | 0.846 |
| Fisheries | 0.947 | 1.000 | 0.973 | 0.857 | 1.000 | 0.923 | 0.947 | 1.000 | 0.973 |
| Education | 1.000 | 1.000 | 1.000 | 1.000 | 0.500 | 0.667 | 1.000 | 1.000 | 1.000 |
| Social sciences and Humanities | 1.000 | 1.000 | 1.000 | 0.600 | 0.750 | 0.667 | 0.600 | 0.750 | 0.667 |
| Economics | 1.000 | 1.000 | 1.000 | 0.900 | 0.818 | 0.857 | 1.000 | 0.545 | 0.706 |
| Average accuracy rate | | | **91.2%** | | | **80.9%** | | | **76.5%** |

ภาพประกอบ 6 แสดงการเปรียบเทียบประสิทธิภาพระหว่างแบบจำลอง

ที่มา : (Dien et al., 2019)

5.6 บทความวิจัยเรื่อง Thai Clickbait Headline News Classification and its Characteristic (Wongsap et al., 2018)

งานวิจัยนี้ได้เสนอการจำแนกประเภทและลักษณะของการใช้คำพาดหัวข่าวที่หลอกลวง ของข้อมูลที่เป็นภาษาไทย ด้วยเทคนิคการเรียนรู้ของเครื่อง เพื่อศึกษาลักษณะเฉพาะของการพาดหัวข่าวที่หลอกลวงของไทย และเปรียบเทียบผลกระทบของอักขระพิเศษว่าเกี่ยวข้องกับการพาดหัวข่าวที่หลอกลวงหรือไม่ ซึ่งแต่ละชุดข้อมูลจะมีการคัดเลือก Features ด้วยเทคนิคการประมวลผลข้อความ ได้แก่ n-gram และ TF-IDF

และใช้ features เหล่านั้นเป็น input ในการสร้างแบบจำลอง Decision Tree, SVM และ Naïve Bayes จากผลการศึกษาพบว่าสัญลักษณ์พิเศษในหัวข้อข่าวมี

บทบาทสำคัญในการจำแนกประเภทฯ โดยแบบจำลอง Decision Tree ให้ประสิทธิภาพดีที่สุด มีความแม่นยำได้สูงที่สุด 99.9%

5.7 บทความวิจัยเรื่อง Automated Classification of Criminal and Violent Activities in Thailand from Online News Articles (Thaipisutikul et al., 2021)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของอาชญากรและกิจกรรมที่รุนแรงในประเทศไทยจากบทความข่าวออนไลน์ ด้วยเทคนิคการเรียนรู้ของเครื่อง เพื่อศึกษาและวิเคราะห์รูปแบบอาชญากรรมที่จะเกิดขึ้น มีการแบ่งประเภทของอาชญากรรมออกเป็น 5 ประเภท ได้แก่ การลักทรัพย์ ยาเสพติด ฆาตกรรม อุบัติเหตุ และการทุจริต และมีการเตรียมความพร้อมของข้อมูลที่เป็นข้อความ ได้แก่ การตัดคำ (Word tokenization), การลบคำที่ไม่สื่อความหมายของประโยค (Remove Stop word), การลดรูปคำให้เป็นแบบพื้นฐาน (Word stemming) เป็นต้น รวมถึงมีการสกัดคุณสมบัติที่สำคัญ (Feature extraction) โดยใช้ TF-IDF ในงานวิจัยนี้จะใช้แบบจำลอง Multinomial Naive Bayes (MNB), Gradient Boosting Machine (GBM), Random Forest (RF), K-Nearest Neighbors (KNN), Multinomial Logistic Regression (MLG) และ Support Vector Machine (SVM)

จากผลการศึกษาพบว่าแบบจำลองที่มีประสิทธิภาพดีที่สุดคือ SVM และ MLG โดยมีความแม่นยำอยู่ที่ 79.41% และ 78.53% ตามลำดับ

5.8 บทความวิจัยเรื่อง Examinations on the Performance of Classification Models for Thai News Articles (Noppakaow & Uchida, 2019)

งานวิจัยนี้ได้เสนอการวัดประสิทธิภาพของแบบจำลองที่ใช้ในการจัดประเภทบทความข่าวในประเทศไทย ซึ่งชุดข้อมูลจะประกอบด้วยบทความข่าวจำนวน 6,000 บทความจากเว็บไซต์ข่าวหลัก 3 แห่ง บทความข่าวแบ่งออกเป็น 4 ประเภท ได้แก่ ข่าวอาชญากรรม ข่าวการเมือง ข่าวกีฬา และข่าวบันเทิง โดยใช้แบบจำลอง Decision Tree, Support Vector Machine (SVM) และ Multilayer Perceptron (MLP) และมีการประเมินประสิทธิภาพของแบบจำลองโดยใช้ Accuracy, Precision, Recall และ F1 Score

จากผลการศึกษาพบว่าแบบจำลองที่มีประสิทธิภาพดีที่สุดคือ MLP, SVM และ Decision Tree คือ 95%, 94% และ 86% ตามลำดับ

5.9 บทความวิจัยเรื่อง Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach (Gaydhani et al., 2018)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของคำพูดแสดงความเกลียดชังและภาษาที่ไม่เหมาะสมบน Twitter ด้วยเทคนิคการเรียนรู้ของเครื่อง โดยใช้แนวทาง N-gram และ TF-IDF ข้อมูลจะถูกแบบออกเป็น 3 ประเภท คือ คำพูดแสดงความเกลียดชัง (hateful), คำพูดก้าวร้าว (offensive) และคำพูดปกติ (clean) ซึ่งทำการทดสอบโดยการพิจารณาคุณสมบัติ (features) จาก N-gram และการคัดแยกคำตามความสำคัญ (TF-IDF) จะถูกนำไปใช้ในแบบจำลองต่างๆ มีการวิเคราะห์เปรียบเทียบแบบจำลองโดยพิจารณาจากค่าต่างๆ ของ n ใน N-gram และ TF-IDF normalization

หลังจากการปรับแต่งแบบจำลอง พบว่าแบบจำลองที่มีประสิทธิภาพดีที่สุด คือ Naïve Bayes โดยใช้ช่วงของ n-gram เท่ากับ 1-3 และใช้ L2 normalization

5.10 บทความวิจัยเรื่อง Fake News Detection Using Neural Network (U et al., 2023)

งานวิจัยนี้ได้เสนอการจำแนกประเภทของข่าวปลอม โดยใช้เทคนิคการเรียนรู้ของเครื่อง และโครงข่ายประสาทเทียม เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองโดยใช้ชุดข้อมูลจากประเภทข่าวที่หลากหลาย เช่น ข่าวการเมือง กีฬา และไวรัสโคโรนา เป็นต้น โดยใช้แบบจำลอง Logistic Regression(LR), Decision Trees (DT), Convolution Neural Networks (CNN) และ Long Short Term Memory (LSTM) และมีการประเมินประสิทธิภาพของแบบจำลอง โดยใช้ Accuracy, F1-score, Recall และ Precision

จากผลการศึกษาพบว่าแบบจำลองที่มีประสิทธิภาพดีที่สุดคือ LSTM, LR, CNN และ DT คือ 99.07%, 98.34%, 97.46% และ 96.16% ตามลำดับ

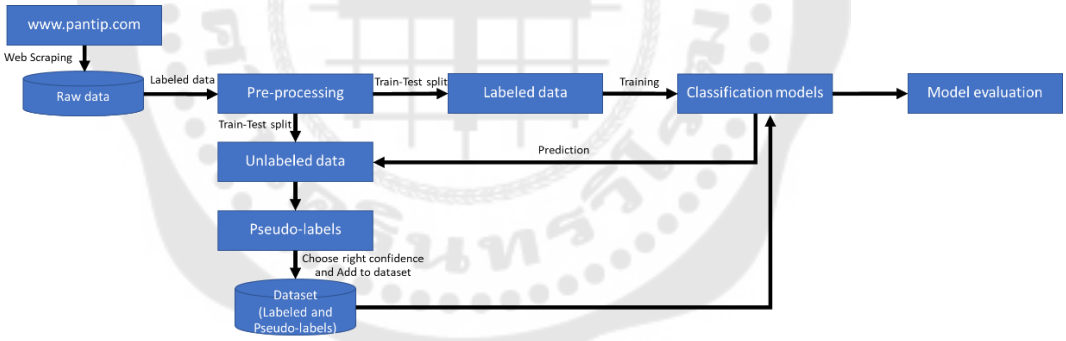
# บทที่ 3
# วิธีดำเนินการวิจัย

ในงานวิจัยครั้งนี้ ผู้วิจัยได้ดำเนินการตามขั้นตอนดังนี้

1. การออกแบบกระบวนการดำเนินการวิจัย
2. การเก็บรวบรวมข้อมูล
3. การเตรียมความพร้อมของข้อมูล
4. การสร้างแบบจำลอง
5. การประเมินผลแบบจำลอง

## 1. การออกแบบกระบวนการดำเนินงานวิจัย

การออกแบบกระบวนการดำเนินงานวิจัย ประกอบไปด้วยกระบวนการหลัก ตามภาพประกอบ 7

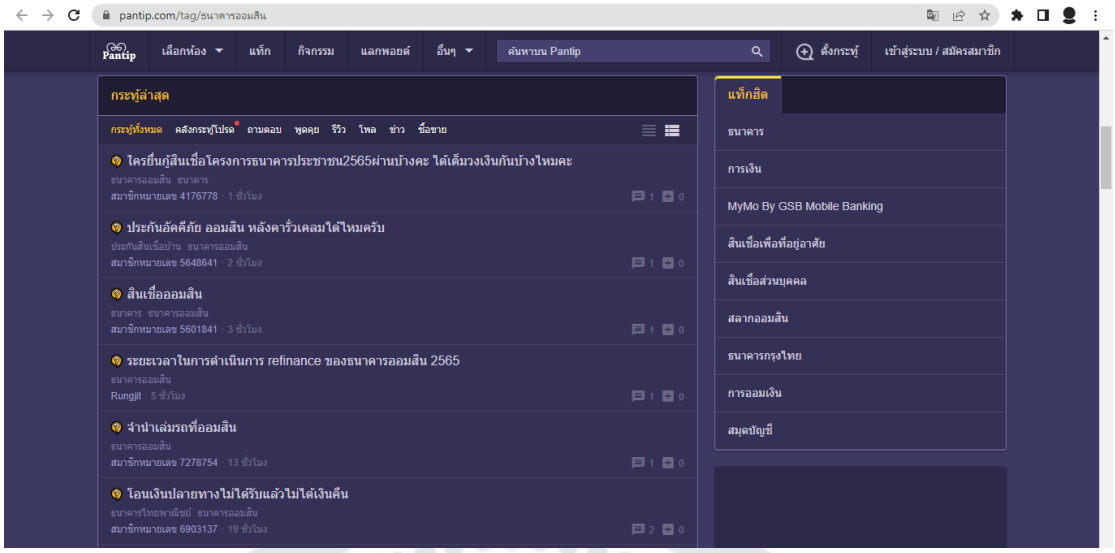



ภาพประกอบ 7 แสดงกระบวนการดำเนินงานวิจัย

## 2. การเก็บรวบรวมข้อมูล

2.1 การดึงข้อมูล

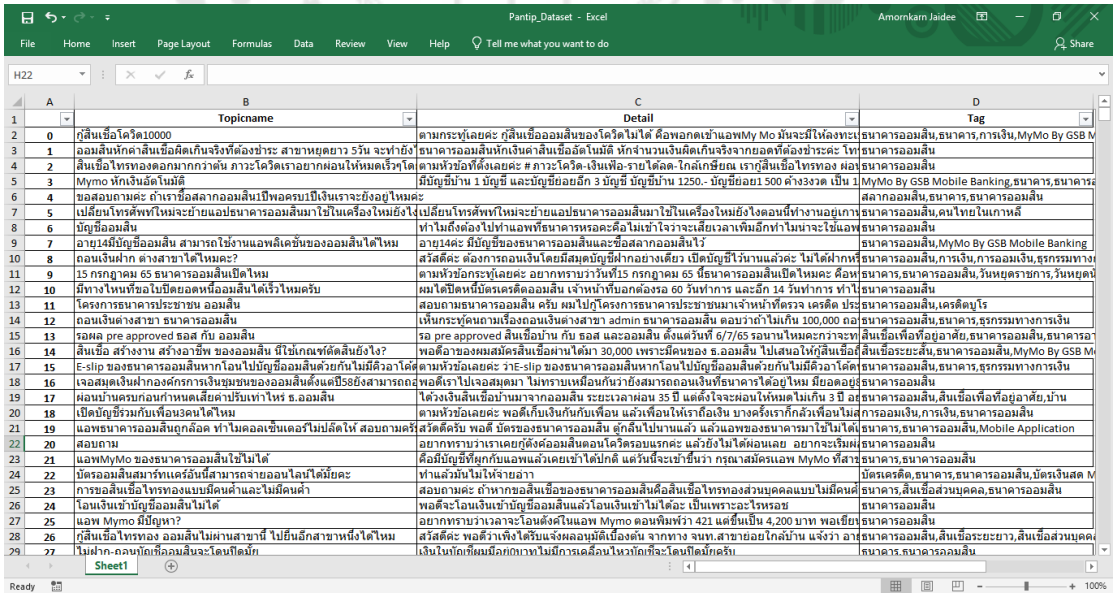
ในงานวิจัยนี้เก็บรวบรวมข้อมูลที่เป็นข้อความจากความคิดเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ และข้อร้องเรียน จากการทำ Web scraping โดยใช้ชุดคำสั่ง Selenium ในการดึงข้อมูลจากเว็บไซต์ www.pantip.com และเลือกแท็กเป็น ธนาคารออมสิน รายละเอียดตามภาพประกอบ 8




ภาพประกอบ 8 แสดงตัวอย่างหน้าจอเว็บไซต์ www.pantip.com

2.2 ชุดข้อมูล

ผู้วิจัยได้ทำการดึงข้อมูล ณ วันที่ 13 กรกฎาคม 2565 มีจำนวนกระทู้ทั้งหมด 600 กระทู้ โดยข้อมูลจะประกอบไปด้วย 3 คอลัมน์ คือ หัวข้อกระทู้ (Topic name), รายละเอียดในกระทู้ (Detail) และแท็ก (Tag) รายละเอียดตามภาพประกอบ 9




| | A | B | C | D |
|---|---|---|---|---|
| 1 | | Topicname | Detail | Tag |
| 2 | 0 | กู้สินเชื่อโควิด10000 | ตามกระทู้เลยค่ะ กู้สินเชื่อออมสินของโควิดไม่ได้ คือพอกดเข้าแอพMy Mo มันจะมีให้ลงทะเ | ธนาคารออมสิน,ธนาคาร,การเงิน,MyMo By GSB M |
| 3 | 1 | ออมสินหักค่าสินเชื่อผิดเกินจริงที่ต้องชำระ สาขาหยุดยาว 5วัน จะทำยังไ | ธนาคารออมสินหักเงินค่าสินเชื่อผิดเกินจริงจากยอดที่ต้องชำระค่ะ โท | ธนาคารออมสิน |
| 4 | 2 | สินเชื่อไทรทองดอกมากกว่าต้น ภาวะโควิดเราอยากผ่อนให้หมดเร็วๆโด | ตามหัวข้อที่ตั้งเลยค่ะ # ภาวะโควิด-เงินเฟ้อ-รายได้ลด-ใกล้เกษียณ เรากู้สินเชื่อไทรทอง ผอ | ธนาคารออมสิน |
| 5 | 3 | Mymo หักเงินอัตโนมัติ | มีบัญชีบ้าน 1 บัญชี และบัญชีย่อยอีก 3 บัญชี บัญชีบ้าน 1250.- บัญชีย่อย1 500 ค้างงวด เป็น 1 | MyMo By GSB Mobile Banking,ธนาคาร,ธนาคาร |
| 6 | 4 | ขอสอบถามค่ะ ถ้าเราชื่อสลากออมสินปีพอครบปีเงินเราจะยังอยู่ไหมคะ | | สลากออมสิน,ธนาคาร,ธนาคารออมสิน |
| 7 | 5 | เปลี่ยนโทรศัพท์ใหม่จะย้ายแอปธนาคารออมสินมาใช้ในเครื่องใหม่ยังไง | เปลี่ยนโทรศัพท์ใหม่จะย้ายแอปธนาคารออมสินมาใช้ในเครื่องใหม่ยังไงตอนนี้ทำงานอยู่เกา | ธนาคารออมสิน,คนไทยในเกาหลี |
| 8 | 6 | บัญชีออมสิน | ทำไมถึงต้องไปทำแอพที่ธนาคารหรอคะคือไม่เข้าใจว่าจะเสียเวลาเพิ่มอีกทำไมน่าจะใช้แอพ | ธนาคารออมสิน |
| 9 | 7 | อายุ14มีบัญชีออมสิน สามารถใช้งานแอพพลิเคชั่นของออมสินได้ไหม | อายุ14ค่ะ มีบัญชีของธนาคารออมสินและซื้อสลากออมสินไว้ | ธนาคารออมสิน,MyMo By GSB Mobile Banking |
| 10 | 8 | ถอนเงินฝาก ต่างสาขาได้ไหมคะ? | สวัสดีค่ะ ต้องการถอนเงินโดยมีสมุดบัญชีฝากอย่างเดียว เปิดบัญชีไว้นานแล้วค่ะ ไม่ได้ฝากหรื | ธนาคารออมสิน,การเงิน,การออมเงิน,ธุรกรรมทาง |
| 11 | 9 | 15 กรกฎาคม 65 ธนาคารออมสินเปิดไหม | ตามหัวข้อกระทู้เลยค่ะ อยากทราบว่าวันที่15 กรกฎาคม 65 นี้ธนาคารออมสินเปิดไหมคะ คือห | ธนาคาร,ธนาคารออมสิน,วันหยุดราชการ,วันหยุดน |
| 12 | 10 | มีทางไหนที่ขอให้ปิดยอดหนี้ออมสินได้เร็วไหมครับ | ผมได้ปิดหนี้บัตรเครดิตออมสิน เจ้าหน้าที่บอกต้องรอ 60 วันทำการ และอีก 14 วันทำการ ทำไ | ธนาคารออมสิน |
| 13 | 11 | โครงการธนาคารประชาชน ออมสิน | สอบถามธนาคารออมสิน ครับ ผมไปกู้โครงการธนาคารประชาชนมาเจ้าหน้าที่ตรวจ เครดิต ประ | ธนาคารออมสิน,เครดิตบูโร |
| 14 | 12 | ถอนเงินต่างสาขา ธนาคารออมสิน | เห็นกระทู้คนถามเรื่องถอนเงินต่างสาขา admin ธนาคารออมสิน ตอบว่าถ้าไม่เกิน 100,000 ถอ | ธนาคารออมสิน,ธนาคาร,ธุรกรรมทางการเงิน |
| 15 | 13 | รอผล pre approved ธอส กับ ออมสิน | รอ pre approved สินเชื่อบ้าน กับ ธอส และออมสิน ตั้งแต่วันที่ 6/7/65 รอนานไหมคะกว่าจะท | สินเชื่อเพื่อที่อยู่อาศัย,ธนาคารออมสิน,ธนาคารอ |
| 16 | 14 | สินเชื่อ สร้างงาน สร้างอาชีพ ของออมสิน มีใช้เกณฑ์ตัดสินยังไง? | พอดีอาของผมสมัครสินเชื่อผ่านไต้มา 30,000 เพราะมีคนของ ธ.ออมสิน ไปเสนอให้กู้สินเชื่อธ | สินเชื่อระยะสั้น,ธนาคารออมสิน,MyMo By GSB M |
| 17 | 15 | E-slip ของธนาคารออมสินหากโอนไปบัญชีออมสินด้วยกันไม่มีค่าโอนได้ | ตามหัวข้อเลยค่ะ วิธี-slip ของธนาคารออมสินหากโอนไปบัญชีออมสินด้วยกันไม่มีค่าโอนได้ | ธนาคารออมสิน,ธนาคาร,ธุรกรรมทางการเงิน |
| 18 | 16 | เจอสมุดเงินฝากองค์กรการเงินชุมชนของออมสินตั้งแต่ปี58ยังสามารถถ | พอดีเราไปเจอสมุดมา ไม่ทราบเหมือนกันว่ายังสมารถถอนเงินที่ธนาคารได้อยู่ไหม มียอดอยู่ | ธนาคารออมสิน |
| 19 | 17 | ผ่อนบ้านครบก่อนกำหนดเสียค่าปรับเท่าไหร่ ธ.ออมสิน | ได้วงเงินสินเชื่อบ้านมาจากออมสิน ระยะเวลาผ่อน 35 ปี แต่ตั้งใจจะผ่อนให้หมดไม่เกิน 3 ปี อย | ธนาคารออมสิน,สินเชื่อเพื่อที่อยู่อาศัย,บ้าน |
| 20 | 18 | เปิดบัญชีร่วมกับเพื่อน3คนได้ไหม | ตามหัวข้อเลยค่ะ พอดีเก็บเงินกันกับเพื่อน แล้วเพื่อนให้เราถือเงิน บางครั้งเราก็กลัวเพื่อนไม่ส | การออมเงิน,การเงิน,ธนาคารออมสิน |
| 21 | 19 | แอพธนาคารออมสินถูกล็อค ทำไมคอลเซ็นเตอร์ไม่ปลดให้ สอบถามครั | สวัสดีครับ พอดี บัตรของธนาคารออมสิน ตู้กลืนไปนานแล้ว แล้วแอพของธนาคารมาใช้ไม่ได้ | ธนาคาร,ธนาคารออมสิน,Mobile Application |
| 22 | 20 | สอบถาม | อยากทราบว่าเราเคยกู้ดังต่อออมสินตอนโควิดรอบแรกค่ะ แล้วยังไม่ได้ผ่อนเลย อยากจะเริ่มผ่ | ธนาคารออมสิน |
| 23 | 21 | แอพMyMo ของธนาคารออมสินใช้ไม่ได้ | คือมีบัญชีที่ผูกกับแอพแล้วเคยเข้าได้ปกติ แต่วันนี้จะเข้าขึ้นว่า กรุณาสมัครแอพ MyMo ที่สา | ธนาคาร,ธนาคารออมสิน |
| 24 | 22 | บัตรออมสินสมาร์ทแคร์อันนี้สามารถจ่ายออนไลน์ได้มั้ยคะ | ทำแล้วมันไม่ให้จ่ายอ่า | บัตรเครดิต,ธนาคาร,ธนาคารออมสิน,บัตรเงินสด M |
| 25 | 23 | การขอสินเชื่อไทรทองแบบมีคนค้ำและไม่มีคนค้ำ | สอบถามค่ะ ถ้าหากขอสินเชื่อของธนาคารออมสินคือสินเชื่อไทรทองส่วนบุคคลแบบไม่มีคนค้ำ | ธนาคาร,สินเชื่อส่วนบุคคล,ธนาคารออมสิน |
| 26 | 24 | โอนเงินเข้าบัญชีออมสินไม่ได้ | พอดีจะโอนเงินเข้าบัญชีออมสินแล้วโอนเงินเข้าไม่ได้อะ เป็นเพราะอะไรหรอช | ธนาคารออมสิน |
| 27 | 25 | แอพ Mymo มีปัญหา? | อยากทราบว่าเวลาจะโอนดังต์ในแอพ Mymo ตอนพิมพ์ว่า 421 แต่ขึ้นเป็น 4,200 บาท พอเขียน | ธนาคารออมสิน |
| 28 | 26 | กู้สินเชื่อไทรทอง ออมสินไม่ผ่านสาขาขึ้ ไปยื่นอีกสาขาหนึ่งได้ไหม | สวัสดีค่ะ พอดีว่าเพิ่งได้รับแจ้งผลอนุมัติเบื้องต้น จากทาง จนท.สาขายอยใกล้บ้าน แจ้งว่า อาร | ธนาคารออมสิน,สินเชื่อระยะยาว,สินเชื่อส่วนบุคค |
| 29 | 27 | ไม่ฝาก-ถอนบัตรเดบิตออมสินจะโดนปิดมั้ย | เงินในบัตรที่ผมมีอยู่เกือบไม่มีการเคลื่อนไหวบัตรใช้จะโดนปิดบัตรไหม | ธนาคาร,ธนาคารออมสิน |

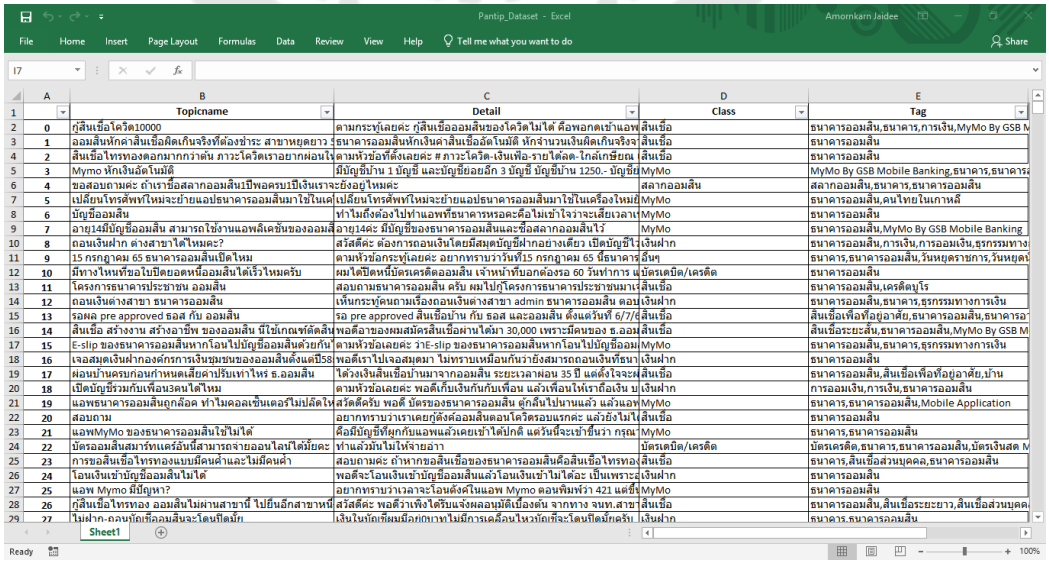
ภาพประกอบ 9 แสดงตัวอย่างไฟล์ชุดข้อมูล

2.3 การระบุประเภทของข้อมูล

ผู้วิจัยได้ทำการเพิ่มคอลัมน์ “Class” ที่เป็นคอลัมน์ target เพื่อระบุประเภทของข้อมูลกระทู้ (Labeled data) จำนวน 600 กระทู้ โดยมีจำนวนของข้อมูลแต่ละประเภทผลิตภัณฑ์ ตามตารางที่ 1 ตัวอย่างของชุดข้อมูลมีรายละเอียดตามภาพประกอบ 10

ตาราง 1 แสดงจำนวนข้อมูลของแต่ละประเภทผลิตภัณฑ์

| ลำดับ | ประเภทผลิตภัณฑ์ | จำนวน |
|---|---|---|
| 1 | สินเชื่อ | 220 |
| 2 | เงินฝาก | 180 |
| 3 | แอปพลิเคชัน MyMo | 90 |
| 4 | อื่นๆ | 68 |
| 5 | บัตรเดบิต/เครดิต | 25 |
| 6 | สลากออมสิน | 17 |




ภาพประกอบ 10 แสดงตัวอย่างของชุดข้อมูล

**3. การเตรียมความพร้อมของข้อมูล**

ในการสร้างแบบจำลองเพื่อจำแนกประเภทของผลิตภัณฑ์จากธนาคาร มีการเตรียมความพร้อมของข้อมูลที่เป็นข้อความ เพื่อเพิ่มประสิทธิภาพของแบบจำลอง โดยมีรายละเอียด ดังต่อไปนี้

3.1 การแบ่งประโยค (Sentence tokenize) ในงานวิจัยนี้ใช้อัลกอริทึม crfcut ในการแบ่งประโยคจากข้อความในเนื้อหาของกระทู้ ซึ่งมีความยาวออกเป็นประโยค ดังภาพประกอบ 11

| | all | Class |
|---|---|---|
| 0 | กู้สินเชื่อโควิด10000 ตามกระทู้เลยค่ะ กู้สินเช... | สินเชื่อ |
| 1 | ออมสินหักค่าสินเชื่อผิดเกินจริงที่ต้องชำระ สาข... | สินเชื่อ |
| 2 | สินเชื่อไทรทองดอกมากกว่าต้น ภาวะโควิดเราอยากผ่... | สินเชื่อ |
| 3 | Mymo หักเงินอัตโนมัติ มีบัญชีบ้าน 1 บัญชี และบ... | MyMo |
| 4 | ขอสอบถามค่ะ ถ้าเราชื้อสลากออมสิน1ปีพอครบ1ปีเงิ... | สลากออมสิน |

| | all | Class |
|---|---|---|
| 0 | [กู้สินเชื่อโควิด10000 ตามกระทู้เลยค่ะ , กู้สิ... | สินเชื่อ |
| 1 | [ออมสินหักค่าสินเชื่อผิดเกินจริงที่ต้องชำระ สา... | สินเชื่อ |
| 2 | [สินเชื่อไทรทองดอกมากกว่าต้น , ภาวะโควิดเราอยา... | สินเชื่อ |
| 3 | [Mymo หักเงินอัตโนมัติ , มีบัญชีบ้าน 1 บัญชี แ... | MyMo |
| 4 | [ขอสอบถามค่ะ ถ้าเราชื้อสลากออมสิน1ปีพอครบ1ปีเง... | สลากออมสิน |

ภาพประกอบ 11 แสดงตัวอย่างการแบ่งประโยค โดยใช้อัลกอริทึม crfcut

3.2 การตัดคำ (Word tokenize) ในงานวิจัยนี้ใช้อัลกอริทึม newmm ในการตัดคำออกจากประโยค และแยกเป็นรายการคำ ดังภาพประกอบ 12

| | all_newmm | Class |
|---|---|---|
| 0 | [กู้, สินเชื่อ, โควิด, ตาม, กระทู้, เลย, ค่ะ, ... | สินเชื่อ |
| 1 | [ออมสิน, หัก, ค่า, สินเชื่อ, ผิด, เกิน, จริง, ... | สินเชื่อ |
| 2 | [สินเชื่อ, ไทรทอง, ดอก, มากกว่า, ตัน, ภาวะ, โค... | สินเชื่อ |
| 3 | [หัก, เงิน, อัตโนมัติ, มี, บัญชี, บ้าน, บัญชี,... | MyMo |
| 4 | [ขอ, สอบถาม, ค่ะ, ถ้า, เรา, ซื้อ, สลาก, ออมสิน... | สลากออมสิน |

ภาพประกอบ 12 แสดงตัวอย่างการตัดคำ โดยใช้อัลกอริทึม newmm

3.3 การลบคำที่ไม่สื่อความหมายของประโยค (Remove Stop word) ในงานวิจัยนี้มีการลบคำที่เกิดขึ้นบ่อยครั้ง และไม่ได้มีส่วนสำคัญต่อความหมายโดยรวมของประโยค โดยจะลบคำที่มีอยู่ในชุดข้อมูลคำที่ไม่สื่อความหมายของประโยคภาษาไทยของไลบรารี PyThaiNLP ดังภาพประกอบ 13

| | all_newmm | Class |
|---|---|---|
| 0 | [กู้, สินเชื่อ, โควิด, กระทู้, กู้, สินเชื่อ, ... | สินเชื่อ |
| 1 | [ออมสิน, หัก, ค่า, สินเชื่อ, ชำระ, สาขา, หยุด,... | สินเชื่อ |
| 2 | [สินเชื่อ, ไทรทอง, ดอก, ตัน, ภาวะ, โควิด, ผ่อน... | สินเชื่อ |
| 3 | [หัก, เงิน, อัตโนมัติ, บัญชี, บ้าน, บัญชี, บัญ... | MyMo |
| 4 | [สอบถาม, ซื้อ, สลาก, ออมสิน, ปี, ปี, เงิน, ยัง... | สลากออมสิน |

ภาพประกอบ 13 แสดงตัวอย่างการลบคำที่ไม่สื่อความหมายของประโยค

3.4 การสร้างกลุ่มคำ (Word Cloud) ในงานวิจัยนี้จะใช้ Word Cloud เพื่อแสดงความถี่ของคำที่ปรากฏอยู่ในแต่ละประเภทของผลิตภัณฑ์ธนาคาร หากคำใดที่ปรากฏบ่อยจะมีขนาดใหญ่ ซึ่งจะโดดเด่นกว่าคำอื่นๆ ดังภาพประกอบ 14 ถึง ภาพประกอบ 19


ภาพประกอบ 14 แสดงตัวอย่างการสร้างกลุ่มคำที่ปรากฏในผลิตภัณฑ์ธนาคารประเภทสินเชื่อ

จากภาพประกอบ 14 จะเห็นว่าคำที่มีความถี่มากที่สุดและเกี่ยวข้องกับข้อความที่เป็นประเภทผลิตภัณฑ์สินเชื่อ คือ “สินเชื่อ, กู้, หนี้” เป็นต้น


ภาพประกอบ 15 แสดงตัวอย่างการสร้างกลุ่มคำที่ปรากฏในผลิตภัณฑ์ธนาคารประเภทเงินฝาก

จากภาพประกอบ 15 จะเห็นว่าคำที่มีความถี่มากที่สุดและเกี่ยวข้องกับข้อความที่เป็นประเภทผลิตภัณฑ์เงินฝาก คือ “บัญชี, โอน, ถอนเงิน” เป็นต้น


ภาพประกอบ 16 แสดงตัวอย่างการสร้างกลุ่มคำที่ปรากฏในผลิตภัณฑ์ธนาคารประเภทสลากออมสิน

จากภาพประกอบ 16 จะเห็นว่าคำที่มีความถี่มากที่สุดและเกี่ยวข้องกับข้อความที่เป็นประเภทผลิตภัณฑ์สลากออมสิน คือ “สลาก, ออมสิน, รางวัล” เป็นต้น


ภาพประกอบ 17 แสดงตัวอย่างการสร้างกลุ่มคำที่ปรากฏในผลิตภัณฑ์ธนาคารประเภทบัตรเดบิต/เครดิต

จากภาพประกอบ 17 จะเห็นว่าคำที่มีความถี่มากที่สุดและเกี่ยวข้องกับข้อความที่เป็นประเภทผลิตภัณฑ์บัตรเดบิต/เครดิต คือ “บัตร, เอทีเอ็ม, บัตรเครดิต” เป็นต้น


ภาพประกอบ 18 แสดงตัวอย่างการสร้างกลุ่มคำที่ปรากฏในผลิตภัณฑ์ธนาคารประเภทแอปพลิเคชัน MyMo

จากภาพประกอบ 18 จะเห็นว่าคำที่มีความถี่มากที่สุดและเกี่ยวข้องกับข้อความที่เป็นประเภทผลิตภัณฑ์แอปพลิเคชัน MyMo คือ “แอพ, ระบบ, แอพธนาคารออมสิน” เป็นต้น


ภาพประกอบ 19 แสดงตัวอย่างการสร้างกลุ่มคำที่ปรากฏในผลิตภัณฑ์ธนาคารประเภทอื่นๆ

จากภาพประกอบ 19 จะเห็นว่าคำที่มีความถี่มากที่สุดและเกี่ยวข้องกับข้อความที่เป็นประเภทผลิตภัณฑ์อื่นๆ คือ “สอบ, สอบถาม, เข้าทำงาน” เป็นต้น

3.5 การคัดแยกคำตามความสำคัญ (TF-IDF) ในงานวิจัยนี้จะใช้ TF-IDF ในการคำนวณน้ำหนักของคำในชุดข้อมูล สำหรับสร้างคุณสมบัติ (feature) เพื่อใช้กับแบบจำลอง ดังภาพประกอบ 20

| | กด | กระทู้ | กรุงไทย | กลับมา | กลัว | กสิกร | กี่ | กู้ | กู้เงิน | ขอบคุณ |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.151666 | 0.199442 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.335919 | 0.0 | 0.0 |
| **1** | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 |
| **2** | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.085567 | 0.0 | 0.0 |
| **3** | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 |
| **4** | 0.000000 | 0.000000 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.000000 | 0.0 | 0.0 |

ภาพประกอบ 20 แสดงตัวอย่างการคุณสมบัติจากการคำนวณ TF-IDF

**4. การสร้างแบบจำลอง**

4.1 การแบ่งชุดข้อมูล

ผู้วิจัยได้นำข้อมูลทั้งหมดมาทำการแบ่งชุดข้อมูลออกเป็น ชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบที่อัตราส่วน 70:30 จากนั้นได้ทำการแบ่งชุดข้อมูลฝึกฝนออกเป็นครึ่งหนึ่งอีกครั้งเป็นส่วนที่มีการระบุประเภทและอีกส่วนที่เราจะทำเหมือนว่าชุดข้อมูลนั้นไม่มีการระบุประเภทไว้ ในอัตราส่วนที่ 70:30 เช่นกัน จึงได้จำนวนข้อมูลของแต่ละชุดข้อมูลเป็น ชุดข้อมูลฝึกฝน จำนวน 294 ตัวอย่าง, ชุดข้อมูลทดสอบ จำนวน 180 ตัวอย่าง และชุดข้อมูลที่ไม่ได้มีการระบุประเภท จำนวน 126 ตัวอย่าง ตามภาพประกอบ 21

```
# train-test split
# training data split data to unlabel
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.3,random_state=42)
X_train, X_unl, y_train, y_unl = train_test_split(X_train, y_train, test_size=0.3,random_state=42)
```

ภาพประกอบ 21 แสดงตัวอย่างชุดคำสั่งสำหรับการแบ่งชุดข้อมูล

4.2 การฝึกฝนแบบจำลอง

ในงานวิจัยนี้จะใช้แบบจำลองที่เหมาะสำหรับการจำแนกประเภท จำนวน 3 แบบจำลอง ดังนี้

1. Support Vector Machine (SVM)
2. Naïve Bayes
3. Logistic Regression

โดยงานวิจัยนี้จะใช้ค่ามาตรฐาน (Default) ของแต่ละแบบจำลอง และทำการฝึกฝนแบบจำลองร่วมกับการทำ 5-Folds Cross Validation สำหรับจำแนกประเภทของผลิตภัณฑ์ธนาคาร โดยนำแบบจำลองมาฝึกฝนกับชุดข้อมูลฝึกฝน และใช้ในการทำนายชุดข้อมูลทดสอบและชุดข้อมูลที่ไม่ได้มีการระบุประเภท เพื่อทำการหาแบบจำลองที่มีประสิทธิภาพสูงที่สุดจากทั้งหมด ตามภาพประกอบ 22

```
from sklearn import svm
from sklearn.model_selection import cross_val_score
import numpy as np

clf = svm.SVC(probability=True)
scores = cross_val_score(clf, X=X_train, y=y_train , cv=5)
scores.mean()
```

ภาพประกอบ 22 แสดงตัวอย่างชุดคำสั่งสำหรับสร้างแบบจำลอง SVM

4.3 การใช้แบบจำลองทำนายชุดข้อมูลที่ไม่ได้มีการระบุประเภท

ผู้วิจัยได้นำแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝน มาทำนายชุดข้อมูลที่ไม่ได้มีการระบุประเภท เพื่อให้สามารถจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความเหล่านั้นได้ หรือที่เรียกว่า Pseudo Label

| | C1Prob | C2Prob | C3Prob | C4Prob | C5Prob | C6Prob | lab | actual | max |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.067430 | 0.035929 | 0.011705 | 0.059816 | 0.719243 | 0.105878 | อื่นๆ | อื่นๆ | 0.719243 |
| 1 | 0.542941 | 0.021820 | 0.305054 | 0.024147 | 0.064923 | 0.041115 | MyMo | สลากออมสิน | 0.542941 |
| 2 | 0.003124 | 0.007788 | 0.001714 | 0.975910 | 0.008510 | 0.002954 | สินเชื่อ | สินเชื่อ | 0.975910 |
| 3 | 0.012841 | 0.011967 | 0.002816 | 0.808840 | 0.128584 | 0.034952 | สินเชื่อ | สินเชื่อ | 0.808840 |
| 4 | 0.083451 | 0.019431 | 0.006628 | 0.037213 | 0.136320 | 0.716957 | เงินฝาก | เงินฝาก | 0.716957 |

ภาพประกอบ 23 แสดงตัวอย่างข้อมูล Pseudo Label

4.4 การเพิ่มข้อมูล Pseudo Label Data ลงในชุดข้อมูลฝึกฝน

ผู้วิจัยจะเลือกข้อมูล Pseudo Label Data ที่มีค่าความน่าจะเป็น (Probability) ผ่านเกณฑ์ของค่าความเชื่อมั่น (Confidence Level) ซึ่งในงานวิจัยนี้จะใช้เป็นค่ามัธยฐาน (Median) เพื่อเพิ่มลงในชุดข้อมูลฝึกฝน โดยจะทำซ้ำเรื่อย ๆ จนกว่าจะไม่มีข้อมูลที่ผ่านเกณฑ์ และจากนั้นจึงนำแบบจำลองมาฝึกฝนกับชุดข้อมูลฝึกฝนที่มีทั้ง Labeled data และ Pseudo label data

**5. การประเมินผลแบบจำลอง**

ในงานวิจัยนี้ใช้เทคนิคการประเมินประสิทธิภาพของแบบจำลองที่เป็นมาตรฐานและเหมาะสำหรับการจำแนกประเภท โดยใช้ตาราง Confusion Matrix ได้แก่ Accuracy, Precision, Recall และ F1 Score

5.1 Accuracy เป็นความแม่นยำสำหรับการจำแนกประเภท ซึ่งวัดจากจำนวนการทำนายประเภทข้อมูลตัวอย่างที่ถูกต้องหารด้วยจำนวนข้อมูลตัวอย่างทั้งหมด ตามสมการดังต่อไปนี้

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positive (TP) หมายถึง การที่แบบจำลองทำนายว่าเป็น positive และความจริงก็เป็น positive

True Negative (TN) หมายถึง การที่แบบจำลองทำนายว่าเป็น negative และความจริงก็เป็น negative

False Positive (FP) หมายถึง การที่แบบจำลองทำนายว่าเป็น positive แต่ความจริงเป็น negative

False Negative (FN) หมายถึง การที่แบบจำลองทำนายว่าเป็น negative แต่ความจริงเป็น positive

5.2 Precision - Recall เป็นการวัดความแม่นยำประเภทหนึ่ง ที่ใช้ในแอปพลิเคชันที่ข้อมูล positive มีจำนวนน้อย แต่มีข้อมูล negative จำนวนมาก ทำให้ accuracy ไม่ถูกต้อง ตามสมการดังต่อไปนี้

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

5.3 F1 Score เป็นการวัดความแม่นยำ โดยการคำนวณมาจาก Precision และ Recall เพื่อใช้เป็นต้นแบบในการเปรียบเทียบประสิทธิภาพของแบบจำลอง ตามสมการดังต่อไปนี้

$$F1\ Score = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$$

# บทที่ 4
# ผลการดำเนินงานวิจัย

ในงานวิจัยนี้ได้ทำการจำแนกประเภทของผลิตภัณฑ์ธนาคาร จำนวน 6 ประเภท จากข้อความของการแสดงความคิดเห็น การขอความอนุเคราะห์ หรือการร้องเรียนจากลูกค้า จำนวน 600 ข้อความ โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติ (Natural language processing) ในการเตรียมความพร้อมของข้อมูลสำหรับแบบจำลอง โดยมีการแบ่งข้อมูลออกเป็น ชุดข้อมูลฝึกฝน จำนวน 294 ตัวอย่าง, ชุดข้อมูลทดสอบ จำนวน 180 ตัวอย่าง และชุดข้อมูลที่ไม่ได้มีการระบุประเภท จำนวน 126 ตัวอย่าง และมีการประเมินประสิทธิภาพของแบบจำลองที่ใช้ในงานวิจัย โดยให้เป็นไปตามวัตถุประสงค์ที่ได้กำหนดไว้ ดังนี้

1. ผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning)

2. ผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น

3. เปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝนทั้งหมด

## 1. **ผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน** (Semi-Supervised Learning)

การจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ของแต่ละแบบจำลอง ได้ผลลัพธ์ดังต่อไปนี้

ตาราง 2 แสดงผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ของแบบจำลอง SVM

| ครั้งที่ | จำนวนข้อมูลฝึกฝน | ค่าความเชื่อมั่น | จำนวนข้อมูลที่ผ่านค่าความเชื่อมั่น | ค่าความแม่นยำของชุดข้อมูลฝึกฝน | ค่าความแม่นยำของชุดข้อมูลทดสอบ |
|---|---|---|---|---|---|
| 1 | 294 | 0.86 | 46 | 0.96 | 0.80 |
| 2 | 340 | 0.69 | 24 | 0.96 | 0.81 |
| 3 | 364 | 0.62 | 12 | 0.96 | 0.81 |
| 4 | 376 | 0.52 | 8 | 0.96 | 0.82 |
| 5 | 384 | 0.53 | 4 | 0.96 | 0.81 |
| 6 | 388 | 0.62 | 2 | 0.97 | 0.82 |
| 7 | 390 | 0.61 | 1 | 0.96 | 0.82 |
| 8 | 391 | 0.56 | 0 | 0.96 | 0.82 |

ตาราง 3 แสดงผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ของแบบจำลอง Logistic Regression

| ครั้งที่ | จำนวนข้อมูลฝึกฝน | ค่าความเชื่อมั่น | จำนวนข้อมูลที่ผ่านค่าความเชื่อมั่น | ค่าความแม่นยำของชุดข้อมูลฝึกฝน | ค่าความแม่นยำของชุดข้อมูลทดสอบ |
|---|---|---|---|---|---|
| 1 | 294 | 0.67 | 47 | 0.88 | 0.78 |
| 2 | 341 | 0.55 | 24 | 0.89 | 0.78 |
| 3 | 365 | 0.48 | 10 | 0.88 | 0.79 |
| 4 | 375 | 0.42 | 5 | 0.88 | 0.78 |
| 5 | 380 | 0.37 | 3 | 0.89 | 0.78 |
| 6 | 383 | 0.36 | 2 | 0.89 | 0.78 |
| 7 | 385 | 0.35 | 1 | 0.89 | 0.78 |
| 8 | 386 | 0.34 | 0 | 0.89 | 0.78 |

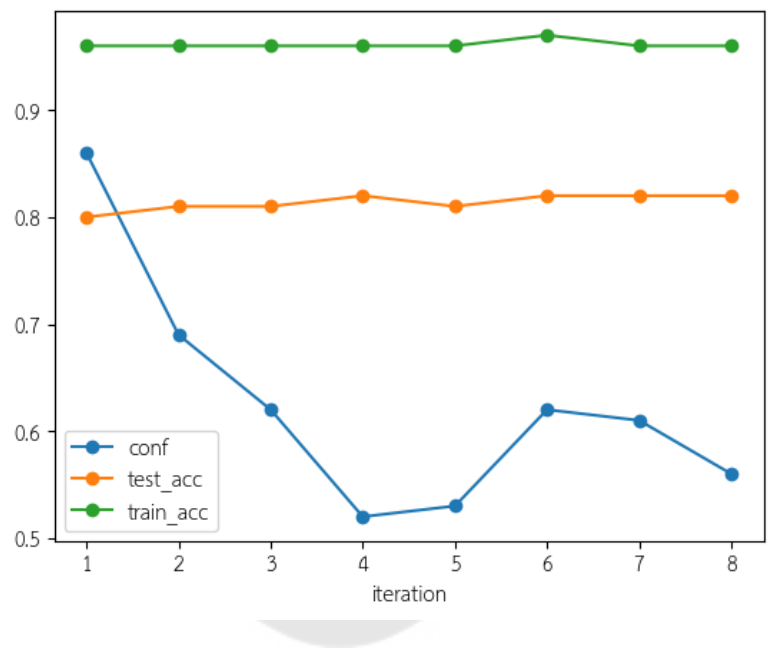
ตาราง 4 แสดงผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ของแบบจำลอง Naïve Bayes

| ครั้งที่ | จำนวนข้อมูลฝึกฝน | ค่าความเชื่อมั่น | จำนวนข้อมูลที่ผ่านค่าความเชื่อมั่น | ค่าความแม่นยำของชุดข้อมูลฝึกฝน | ค่าความแม่นยำของชุดข้อมูลทดสอบ |
|---|---|---|---|---|---|
| 1 | 294 | 0.64 | 46 | 0.83 | 0.76 |
| 2 | 340 | 0.57 | 22 | 0.83 | 0.75 |
| 3 | 362 | 0.49 | 11 | 0.83 | 0.75 |
| 4 | 373 | 0.41 | 6 | 0.83 | 0.75 |
| 5 | 379 | 0.40 | 3 | 0.84 | 0.76 |
| 6 | 382 | 0.37 | 1 | 0.84 | 0.75 |
| 7 | 383 | 0.36 | 1 | 0.84 | 0.75 |
| 8 | 384 | 0.35 | 0 | 0.84 | 0.76 |

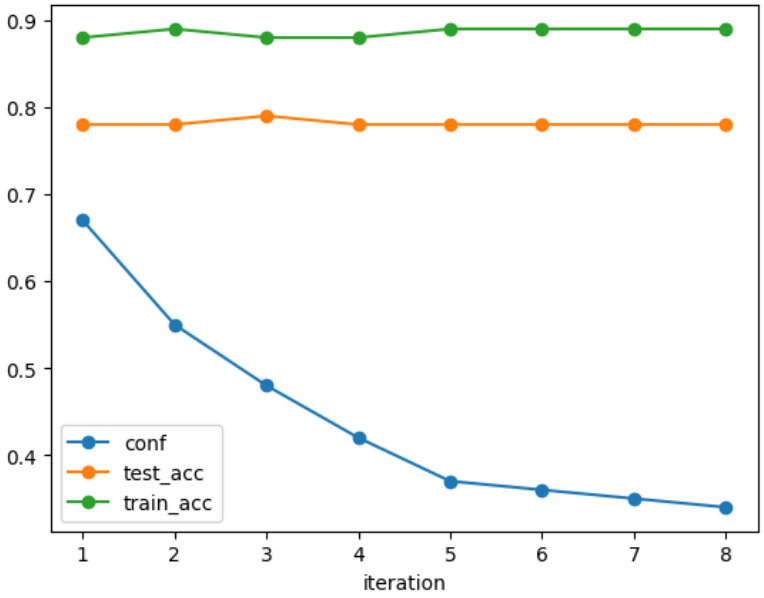
จากผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ของทั้ง 3 แบบจำลอง พบว่า แบบจำลอง Support Vector Machine (SVM) มีข้อมูล Pseudo label data ที่ผ่านเกณฑ์ค่ามัธยฐาน (Median) และเพิ่มลงในชุดข้อมูลฝึกฝน มีจำนวนมากที่สุด ซึ่งมีจำนวนข้อมูลที่ผ่านเกณฑ์เท่ากับ 97 ตัวอย่าง จากข้อมูลที่ไม่ได้มีการระบุประเภทไว้ทั้งหมด จำนวน 126 ตัวอย่าง คงเหลือข้อมูลจำนวน 29 ตัวอย่าง ที่ไม่ผ่านเกณฑ์ ซึ่งน้อยที่สุดในแบบจำลองทั้งหมด

**2. ผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น**

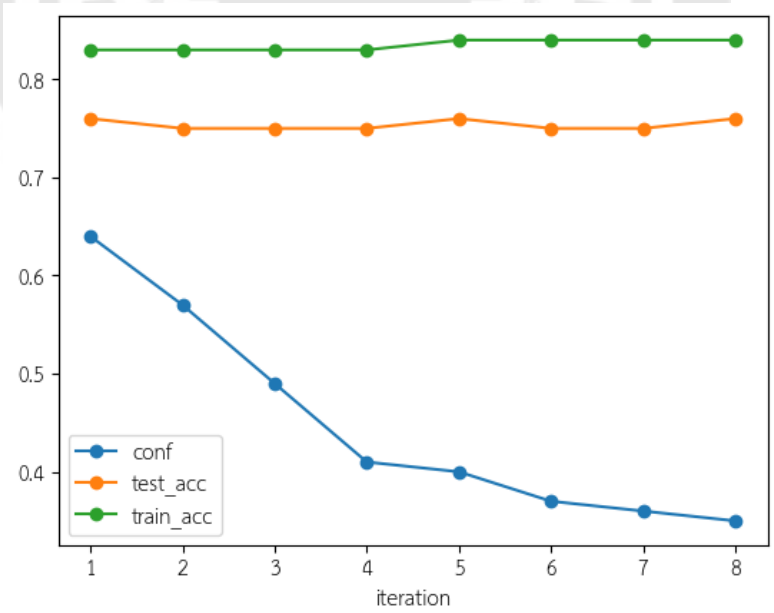
จากการทดลองสร้างแบบจำลองเพื่อฝึกฝนกับชุดข้อมูลฝึกฝน และการทำนายชุดข้อมูลทดสอบ และชุดข้อมูลที่ไม่ได้มีการระบุประเภท เพื่อให้ได้ข้อมูล Pseudo Label Data และเพิ่มข้อมูลที่ผ่านเกณฑ์ของค่าความเชื่อมั่นลงในชุดข้อมูลฝึกฝน ซึ่งจะมีการทำซ้ำจนกว่าจะไม่มีข้อมูลที่ผ่านเกณฑ์ จึงทำให้ได้ผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่นของทั้ง 3 แบบจำลอง ดังภาพประกอบ 24 ถึง ภาพประกอบ 26




ภาพประกอบ 24 แสดงผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น ของแบบจำลอง SVM




ภาพประกอบ 25 แสดงผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น ของแบบจำลอง Logistic Regression

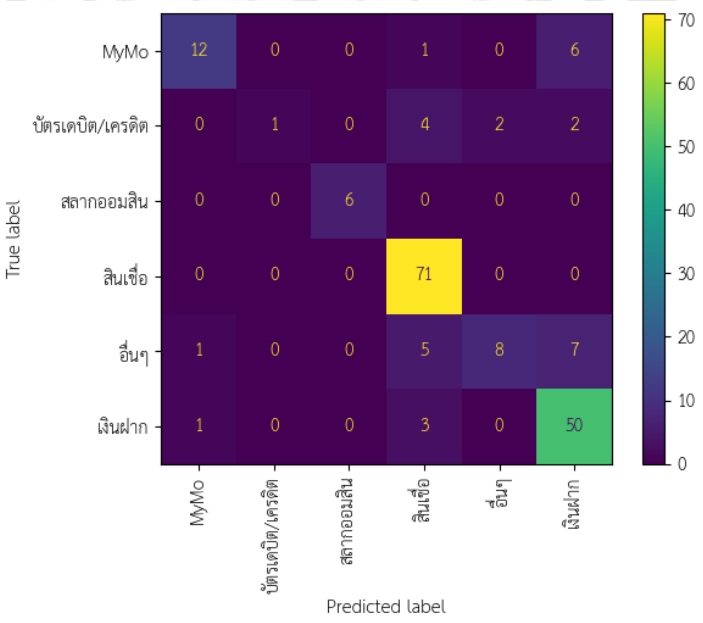



ภาพประกอบ 26 แสดงผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น ของแบบจำลอง Naïve Bayes

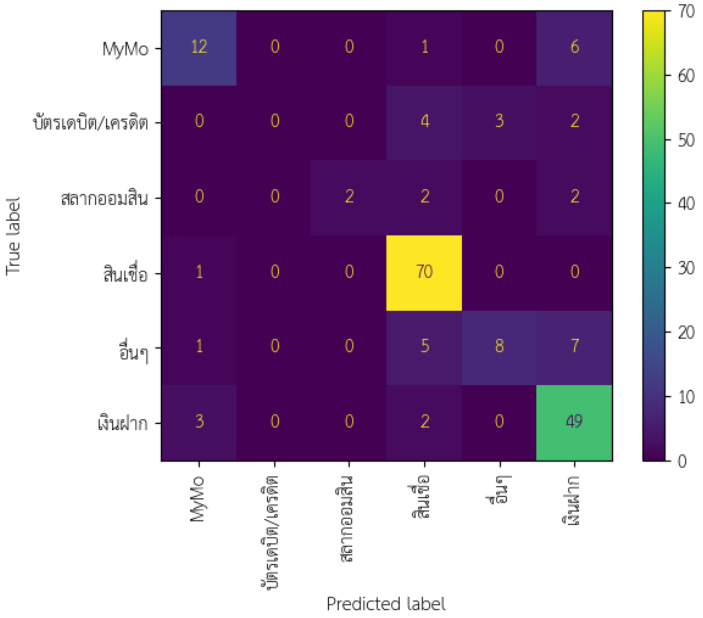
จากผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น จะพบว่าค่าความเชื่อมั่นมีค่าลดลงเรื่อย ๆ เนื่องจากข้อมูลที่ไม่ได้มีการระบุประเภท มีจำนวนลดลง ส่งผลให้ข้อมูลที่ผ่านเกณฑ์ของค่าความเชื่อมั่นมีจำนวนลดลงไปด้วย และเมื่อมีการเพิ่มข้อมูล Pseudo Label Data ลงในชุดข้อมูลฝึกฝนจะพบว่า หากมีข้อมูลที่ถูกเพิ่มเป็นจำนวนมาก จะส่งผลให้ความแม่นยำของชุดข้อมูลฝึกฝน และชุดข้อมูลทดสอบเพิ่มมากขึ้น ซึ่งจะเห็นได้ชัดจากแบบจำลอง Support Vector Machine (SVM) ตามภาพประกอบ 18

**3. เปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝนทั้งหมด**

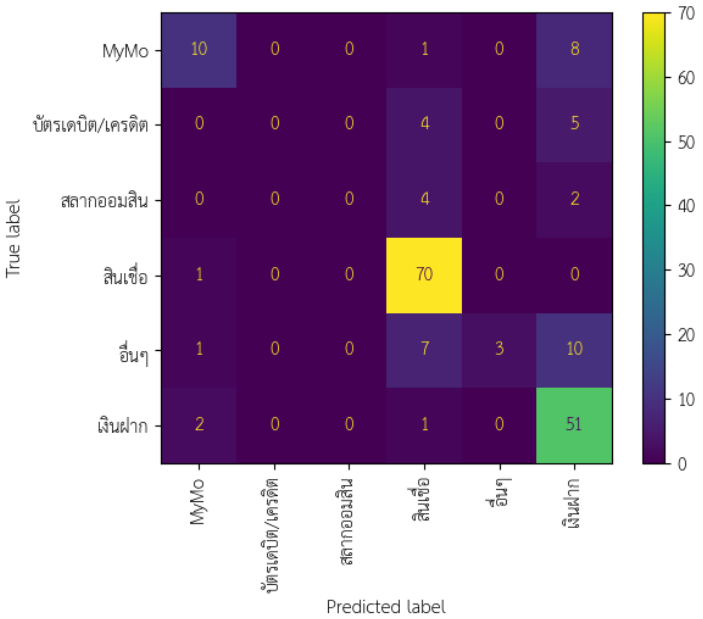
จากแบบจำลองที่ได้จากการฝึกฝนกับชุดข้อมูลฝึกฝนที่ประกอบไปด้วย Labeled data และ Pseudo label data และนำมาทำนายกับชุดข้อมูลทดสอบ จะได้ผลลัพธ์ของการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝนทั้งหมด ดังภาพประกอบ 27 ถึง ภาพประกอบ 29 และ ตาราง 5




ภาพประกอบ 27 แสดงตาราง Confusion Matrix ของแบบจำลอง SVM




ภาพประกอบ 28 แสดงตาราง Confusion Matrix ของแบบจำลอง Logistic Regression




ภาพประกอบ 29 แสดงตาราง Confusion Matrix ของแบบจำลอง Naïve Bayes

ตาราง 5 แสดงตารางเปรียบเทียบผลการประเมินประสิทธิภาพของแบบจำลอง SVM, Logistic Regression และ Naïve Bayes

| Models | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| SVM | 0.82 | 0.88 | 0.68 | 0.70 |
| Logistic Regression | 0.78 | 0.67 | 0.54 | 0.56 |
| Naïve Bayes | 0.74 | 0.53 | 0.43 | 0.42 |

จากตาราง 5 จะเห็นได้ว่าทั้ง 3 แบบจำลอง จะมีค่าความแม่นยำที่ห่างกันไม่มากนัก แต่แบบจำลอง SVM เป็นแบบจำลองที่มีประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารได้ดีกว่าแบบจำลองอื่นๆ โดยมีค่า Accuracy เท่ากับ 0.82 , ค่า Precision เท่ากับ 0.88 , ค่า Recall เท่ากับ 0.68 และ ค่า F1-Score เท่ากับ 0.70 รองลงมาจะเป็นแบบจำลอง Logistic Regression และ Naïve Bayes ตามลำดับ

# บทที่ 5
# สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการในการวิจัยนี้ เพื่อศึกษาการจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความที่ลูกค้าให้ความเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ หรือร้องเรียนเข้ามา ซึ่งเป็นข้อความภาษาไทย โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ผู้วิจัยได้ประเมินประสิทธิภาพของแบบจำลอง โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อเสนอแนะ

## 1. สรุปผลการวิจัย

การรับฟังข้อคิดเห็น หรือข้อร้องเรียนจากลูกค้า ถือว่าเป็นเรื่องที่สำคัญ และประโยชน์อย่างมากในการทำให้องค์กรประสบความสำเร็จด้านความพึงพอใจของลูกค้า และรักษาลูกค้าไว้ให้มีความสัมพันธ์กับองค์กรต่อไป โดยการที่จะทราบได้ว่าเรื่องที่ลูกค้าแจ้งเข้ามาเป็นผลิตภัณฑ์หรือบริการใด พนักงานที่ดูแลระบบจะต้องอ่านทุกข้อความ ถึงจะระบุได้ว่าเป็นผลิตภัณฑ์ในด้านใด หรือจำแนกว่าเกี่ยวข้องกับเรื่องใด ซึ่งขั้นตอนในส่วนนี้จะทำให้ใช้เวลาเป็นอย่างมาก และทำให้การที่จะส่งเรื่องไปยังหน่วยงานต่างๆ เพื่อดำเนินการในส่วนที่เกี่ยวข้องมีความล่าช้า จนอาจทำให้การแก้ไขปัญหาไม่ทันเวลา และไม่เป็นที่พึงพอใจต่อลูกค้าได้

ในงานวิจัยนี้ได้ทำการจำแนกประเภทของผลิตภัณฑ์ธนาคาร จากข้อความของการแสดงความคิดเห็น การขอความอนุเคราะห์ หรือการร้องเรียนจากลูกค้า จำนวน 600 ข้อความ ที่เก็บรวบรวมจากการทำ Web scraping จากเว็บไซต์ www.pantip.com และเลือกแท็กเป็นธนาคารออมสิน โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning) ร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติ (Natural language processing) ในการเตรียมความพร้อมของข้อมูลสำหรับแบบจำลอง โดยมีการแบ่งชุดข้อมูลออกเป็น ชุดข้อมูลฝึกฝนและชุดข้อมูล

ทดสอบที่อัตราส่วน 70:30 จากนั้นได้ทำการแบ่งชุดข้อมูลฝึกฝนออกเป็นครึ่งหนึ่งอีกครั้งเป็นส่วนที่มีการระบุประเภทและอีกส่วนที่เราจะทำเหมือนว่าชุดข้อมูลนั้นไม่มีการระบุประเภทไว้ ในอัตราส่วนที่ 70:30 เช่นกัน จึงได้จำนวนข้อมูลของแต่ละชุดข้อมูลเป็นชุดข้อมูลฝึกฝน จำนวน 294 ตัวอย่าง, ชุดข้อมูลทดสอบ จำนวน 180 ตัวอย่าง และชุดข้อมูลที่ไม่ได้มีการระบุประเภท จำนวน 126 ตัวอย่าง

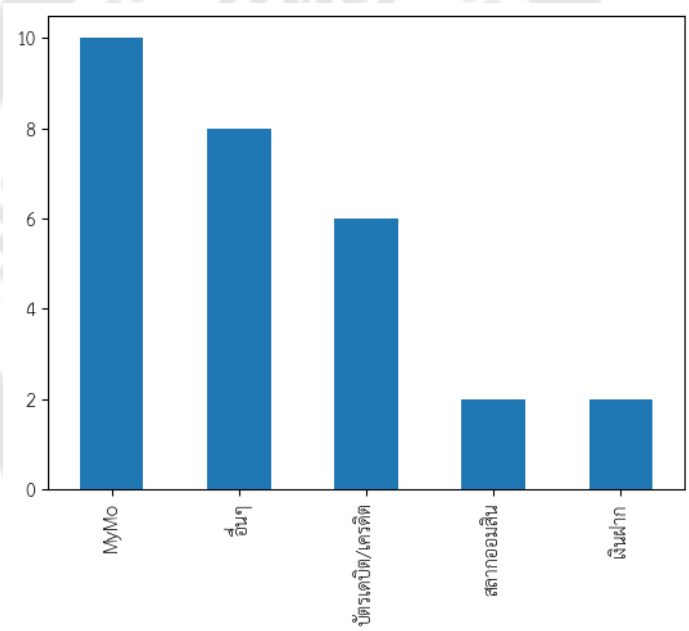
จากผลลัพธ์ของการทดลองและการวัดประสิทธิภาพของแบบจำลองทั้งหมด พบว่า แบบจำลอง SVM สามารถทำให้มีการเพิ่มข้อมูล Pseudo Label Data ลงในข้อมูลฝึกฝนได้มากที่สุด จำนวน 97 ตัวอย่าง จากข้อมูลที่ไม่ได้มีการระบุประเภทไว้ทั้งหมด จำนวน 126 ตัวอย่าง คงเหลือข้อมูลจำนวน 29 ตัวอย่าง ที่ไม่ผ่านเกณฑ์และไม่สามารถเพิ่มลงในชุดข้อมูลฝึกฝน ซึ่งน้อยกว่าแบบจำลอง Logistic Regression และ Naïve Bayes และ SVM ยังเป็นแบบจำลองที่มีประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารได้ถูกต้องและดีกว่าแบบจำลองอื่นๆ โดยมีค่า Accuracy เท่ากับ 0.82 , ค่า Precision เท่ากับ 0.88 , ค่า Recall เท่ากับ 0.68 และ ค่า F1-Score เท่ากับ 0.70 รองลงมาจะเป็นแบบจำลอง Logistic Regression และ Naïve Bayes ตามลำดับ

**2. อภิปรายผลการวิจัย**

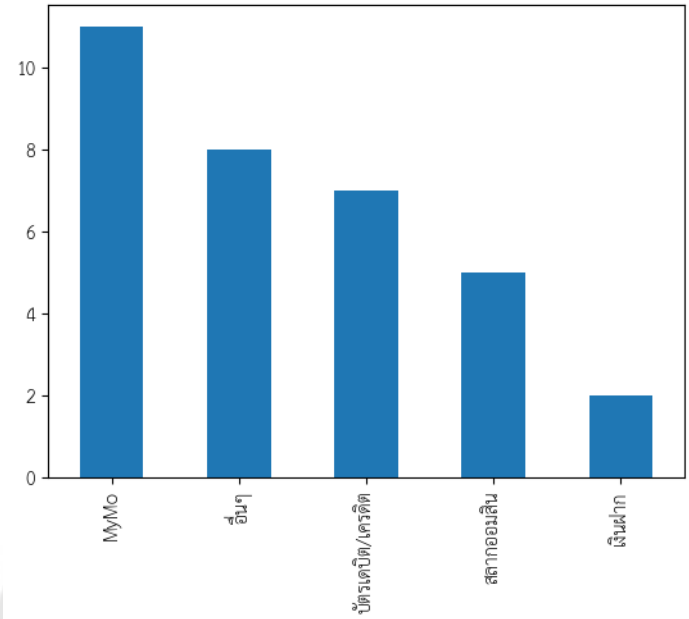
ในงานวิจัยนี้ได้สร้างแบบจำลองเพื่อการจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความภาษาไทย โดยใช้เทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน ร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติ เมื่อทำการทดลองสร้างแบบจำลองเพื่อฝึกฝนกับชุดข้อมูลฝึกฝน และการทำนายชุดข้อมูลทดสอบ และชุดข้อมูลที่ไม่ได้มีการระบุประเภท เพื่อให้ได้ข้อมูล Pseudo Label Data และเพิ่มข้อมูลที่ผ่านเกณฑ์ของค่าความเชื่อมั่นลงในชุดข้อมูลฝึกฝน ซึ่งจะมีการทำซ้ำจนกว่าจะไม่มีข้อมูลที่ผ่านเกณฑ์ จะพบว่าค่าความเชื่อมั่นมีค่าลดลงเรื่อย ๆ เนื่องจากข้อมูลที่ไม่ได้มีการระบุประเภท มีจำนวนลดลง ส่งผลให้ข้อมูลที่ผ่านเกณฑ์ของค่าความเชื่อมั่นมีจำนวนลดลงไปด้วย และเมื่อมีการเพิ่มข้อมูล Pseudo Label Data ลงในชุดข้อมูลฝึกฝนจะพบว่า หากมี

ข้อมูลที่ถูกเพิ่มเป็นจำนวนมาก จะส่งผลให้ความแม่นยำของชุดข้อมูลฝึกฝน และชุดข้อมูลทดสอบเพิ่มมากขึ้น

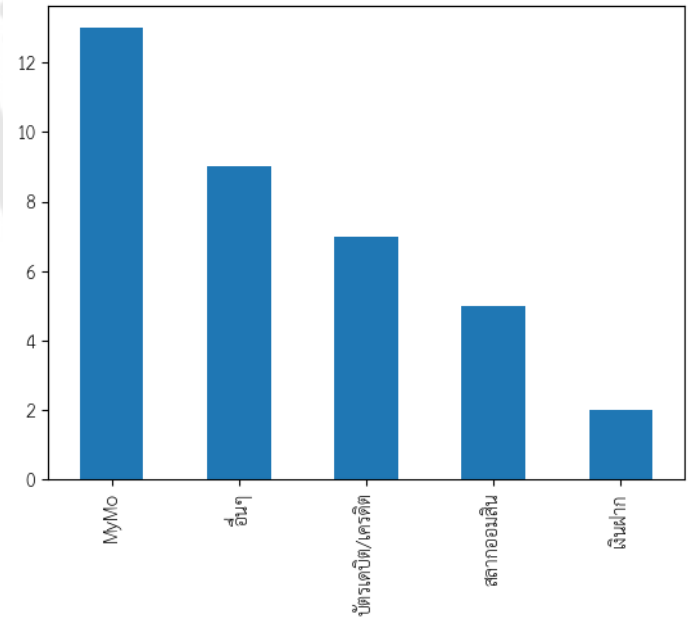
ในส่วนข้อมูลที่ไม่ผ่านเกณฑ์ของค่าความเชื่อมั่น จะพบว่าเป็นข้อมูลของประเภทผลิตภัณฑ์ ที่มีจำนวนข้อมูลในการฝึกฝนน้อย แม้แต่แบบจำลอง SVM ที่มีประสิทธิภาพดีที่สุด ยังมีข้อจำกัดในการฝึกฝนกับข้อมูลที่มีจำนวนน้อยและมีลักษณะที่ไม่สมดุลเช่นเดียวกับแบบจำลองอื่นๆ ซึ่งประเภทของผลิตภัณฑ์ที่ไม่สามารถผ่านเกณฑ์ค่าความเชื่อมั่นมากที่สุด ได้แก่ MyMo, อื่นๆ, บัตรเดบิต/เครดิต, สลากออมสิน และเงินฝาก จะเห็นได้ว่าประเภทผลิตภัณฑ์สินเชื่อ ไม่มีข้อมูลที่ไม่ผ่านเกณฑ์เลย ซึ่งเป็นไปตามภาพประกอบ 30 ถึง ภาพประกอบ 33




ภาพประกอบ 30 แสดงประเภทผลิตภัณฑ์ฯ ที่ไม่ผ่านเกณฑ์ค่าความเชื่อมั่นของแบบจำลอง SVM




ภาพประกอบ 31 แสดงประเภทผลิตภัณฑ์ฯ ที่ไม่ผ่านเกณฑ์ค่าความเชื่อมั่นของแบบจำลอง Logistic Regression




ภาพประกอบ 32 แสดงประเภทผลิตภัณฑ์ฯ ที่ไม่ผ่านเกณฑ์ค่าความเชื่อมั่นของแบบจำลอง Naïve Bayes

| | prep_sentence | Class |
|---|---|---|
| 498 | ปิดบัญชี,สลาก,ดิจิทัล,สมัคร,ออนไลน์,ออมสิน,สมั... | สลากออมสิน |
| 398 | สอบสัมภาษณ์,ธนาคารออมสิน,สอบสัมภาษณ์,เข้าทำงาน... | อื่นๆ |
| 328 | รหัส,บัตร,เอทีเอ็ม,ออมสิน,รหัส,รหัส | บัตรเดบิต/เครดิต |
| 466 | ลงทะเบียน,บัตร,คนจน,ทราบ,ตอนนี้,ลงทะเบียน,เตรี... | อื่นๆ |
| 224 | ออมสิน,บันทึก,สลิป,แชร์,สลิป,ขึ้นชื่อ,ผู้รับ,ผ... | MyMo |
| 35 | ติดต่อ,ธนาคารออมสิน,เบอร์,ติดต่อ,ต้องการ,ติดต่... | บัตรเดบิต/เครดิต |
| 71 | ออมสิน,ช่อง,ทางการ,จ่าย,เงิน,ส่วนตัว,ออมสิน,เว... | MyMo |
| 34 | สอบถาม,ธนาคารออมสิน,ผม,ขำ,ระ,สินเชื่อ,ตามกำหนด... | เงินฝาก |
| 47 | ตอนนี้,ออมสิน,เข้าทำงาน,หัวข้อ,คน,สอบสัมภาษณ์,... | อื่นๆ |
| 108 | กระปุกออมสิน,ลงทะเบียน,เม,ต้าเวิร์ส,บ้าน,หรือว... | อื่นๆ |

ภาพประกอบ 33 แสดงตัวอย่างของข้อมูลที่ไม่ผ่านเกณฑ์ค่าความเชื่อมั่น

อีกหนึ่งปัญหาที่ผู้วิจัยได้พบคือ เมื่อทดลองนำแต่ละแบบจำลองมาทำนายกับชุดข้อมูลทดสอบ พบว่าข้อความของผลิตภัณฑ์ธนาคารบางประเภทมีผลลัพธ์ของการทำนายถูกต้องได้น้อย หรือไม่ถูกต้องเลย เนื่องจากชุดข้อมูลดังกล่าวมีความไม่สมดุลกันอยู่ (Imbalanced Data) ทำให้แบบจำลองฝึกฝนกับข้อมูลประเภทผลิตภัณฑ์นั้นๆ ได้น้อย ได้แก่ ประเภทบัตรเดบิต/เครดิตที่มีข้อมูลเพียง 25 ข้อความ และ ประเภทสลากออมสินที่มีข้อมูลเพียง 17 ข้อความ ซึ่งจะสังเกตได้ว่าแบบจำลองจะทำนายเป็นประเภทผลิตภัณฑ์ที่มีข้อความจำนวนมากๆ

ทั้งนี้ แบบจำลองทั้งหมดสามารถจำแนกประเภทของผลิตภัณฑ์ธนาคาร ที่มีประสิทธิภาพของแบบจำลองอยู่ในเกณฑ์ที่ดี ซึ่งเป็นไปตามสมมติฐานที่ผู้วิจัยได้ตั้งไว้ข้างต้น

**3. ข้อเสนอแนะ**

3.1 ในงานวิจัยนี้ใช้ชุดข้อมูลที่มีจำนวนข้อความเพียง 600 ข้อความ ซึ่งมีจำนวนน้อยเกินไปในการฝึกฝนแบบจำลอง หากสามารถเพิ่มจำนวนข้อมูลที่ใช้ในการฝึกฝนได้อาจทำให้ประสิทธิภาพของแบบจำลองเพิ่มมากขึ้น

3.2 ในชุดข้อมูลยังมีความไม่สมดุลกันอยู่ (Imbalanced Data) ควรจะมีการทำ Over Sampling, Under Sampling หรือ SMOTE เป็นต้น เพื่อทำให้ข้อมูลทุกประเภทมีความสมดุลกัน ซึ่งจะส่งผลต่อประสิทธิภาพของแบบจำลอง

3.3 ในงานวิจัยนี้ใช้ค่าทางสถิติ คือ ค่ามัธยฐาน (Median) เป็นเกณฑ์ของค่าความเชื่อมั่น (Confidence Level) เพียงค่าเดียว อาจมีการทดสอบกับค่าทางสถิติอื่นๆร่วมด้วย เช่น ค่าเฉลี่ย (Mean) หรือค่าฐานนิยม (Mode) เป็นต้น ว่าส่งผลในการเพิ่มประสิทธิภาพของแบบจำลองหรือไม่

3.4 ในงานวิจัยนี้สร้างแบบจำลองโดยใช้ค่ามาตรฐาน (Default) เพื่อฝึกฝนในแต่ละแบบจำลอง ซึ่งในการเพิ่มประสิทธิภาพของแบบจำลอง อาจใช้เทคนิคการปรับจูนพารามิเตอร์ (Hyperparameter Tuning) เพื่อหาพารามิเตอร์ที่เหมาะสมกับแบบจำลองและทำให้ประสิทธิภาพของแบบจำลองเพิ่มมากขึ้น

3.5 แบบจำลองในงานวิจัยนี้สามารถประยุกต์ใช้กับโปรแกรมการจำแนกประเภทของผลิตภัณฑ์ธนาคาร เพื่อช่วยลดเวลา และลดทรัพยากรแรงงานในการที่จะต้องระบุประเภทฯ ด้วยตนเอง

# บรรณานุกรม


Dien, T. T., Loc, B. H., & Thai-Nghe, N. (2019, 26-28 Nov. 2019). Article Classification using Natural Language Processing and Machine Learning. 2019 International Conference on Advanced Computing and Applications (ACOMP),

Dorado, R., & Ratte, S. (2016). Semi-Supervised Text Classification Using Unsupervised Topic Information. 2016 29th International Florida Artificial Intelligence Research Society Conference,

Gaydhani, A., Doma, V., Kendre, S., & B B, L. (2018). *Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach*.

Lee, J. H., Ko, S.-K., & Han, Y.-S. (2021). SALNet: Semi-supervised Few-Shot Text Classification with Attention-based Lexicon Construction. AAAI,

Noppakaow, A., & Uchida, O. (2019, 10-11 Oct. 2019). Examinations on the Performance of Classification Models for Thai News Articles. 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE),

Saengkhunthod, C., Kerdnoonwong, P., & Atchariyachanvanich, K. (2021, 21-24 Jan. 2021). Detection of Unreliable Medical Articles on Thai Websites. 2021 13th International Conference on Knowledge and Smart Technology (KST),

Thaipisutikul, T., Tuarob, S., Pongpaichet, S., Amornvatcharapong, A., & Shih, T. K. (2021, 21-24 Jan. 2021). Automated Classification of Criminal and Violent Activities in Thailand from Online News Articles. 2021 13th International Conference on Knowledge and Smart Technology (KST),

Thanh, V. D., Tuan, P. M., Hung, V. T., & Ban, D. V. (2013, 15-18 Dec. 2013). Text classification based on semi-supervised learning. 2013 International Conference on Soft Computing and Pattern Recognition (SoCPaR),

U, P., Naik, A., Gurav, S., Kumar, A., C, S. R., & M, B. S. (2023, 24-25 Feb. 2023). Fake News Detection Using Neural Network. 2023 IEEE International Conference on Integrated Circuits and Communication Systems (ICICACS),

Wongsap, N., Prapphan, T., Lou, L., Kongyoung, S., Jumun, S., & Kaothanthong, N. (2018, 7-9 May 2018). Thai Clickbait Headline News Classification and its Characteristic. 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems (ICESIT-ICICTES),

# ประวัติผู้เขียน

**ผลงานตีพิมพ์** -

**รางวัลที่ได้รับ** -