การจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความภาษาไทยด้วยวิธีการเรียนรู้แบบกึ่งมีผู้สอน

อมรกานต์ ใจดี[1*], ศุภร คนธภักดี[2]

บทคัดย่อ

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 และทำการทดลองโดยใช้ 3 แบบจำลอง ประกอบด้วย Support Vector Machine (SVM), Logistic Regression (LR) และ Naïve Bayes (NB) จากการทดลองพบว่าแบบจำลองที่มีประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารได้ดีที่สุด คือ SVM โดยมีความแม่นยำอยู่ที่ 0.82 โดยแบบจำลอง LR และ NB จะมีความแม่นยำอยู่ที่ 0.78 และ 0.74 ตามลำดับ

คำสำคัญ : การวิเคราะห์ข้อความ, การประมวลผลภาษาธรรมชาติ, เทคนิคการเรียนรู้แบบกึ่งมีผู้สอน

---

[1] หลักสูตรวิทยาศาสตรมหาบัณฑิต สาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

[2] คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ กรุงเทพฯ 10110

* Corresponding author:Tel.: 094-8108116 E-mail address: amornkarn.jaidee@g.swu.ac.th

Classification Of Banking Products From Thai Text Using A Semi-Supervised Learning

Amornkarn Jaidee[1*], Subhorn Khonthapagdee[2]

Abstract

The purpose of this research was to classify complaints about banking products using a semi-supervised machine learning methods. In this research, we use customer complaints obtained from Web scraping by using the 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 the 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.

Keywords : Text Analytics, Natural Language Processing, Semi-supervised Learning

---

[1] Data Science, Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

[2] Faculty of Science, Srinakharinwirot University, Bangkok, 10110, Thailand

* Corresponding author:Tel.: 094-8108116 E-mail address: amornkarn.jaidee@g.swu.ac.th

บทนำ

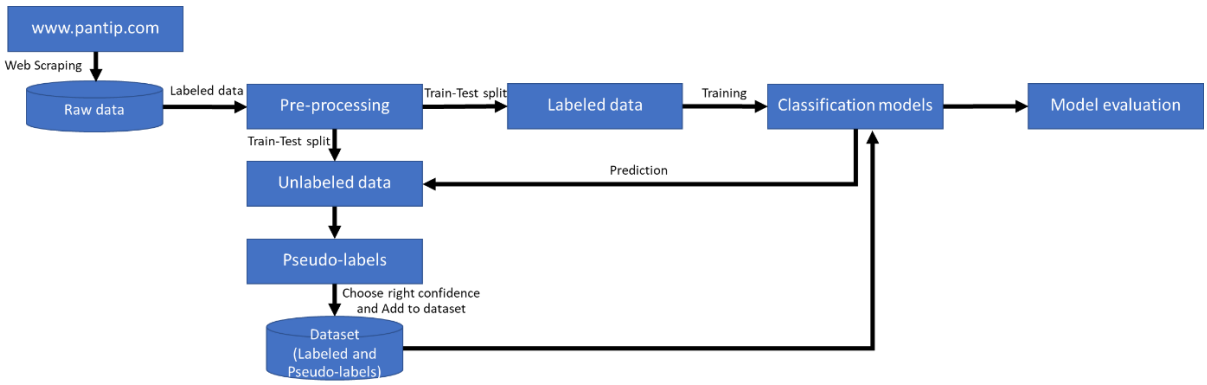
ปัจจุบัน ในทางการทำธุรกิจใดๆก็ตาม การรับฟังข้อคิดเห็น หรือข้อร้องเรียนจากลูกค้า ถือว่าเป็นเรื่องที่สำคัญอย่างยิ่ง ไม่ว่าจะเป็นในเรื่องของการแนะนำเพื่อให้องค์กรสามารถปรับตัวตามคำแนะนำได้อย่างทันท่วงที การแสดงความคิดเห็นสำหรับนำมาใช้ในการออกผลิตภัณฑ์ใหม่ๆ การขอความอนุเคราะห์ให้ดำเนินการในเรื่องต่างๆ หรือการร้องเรียนที่เกิดขึ้นจากการใช้บริการหรือผลิตภัณฑ์ของทางองค์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

ในทางธุรกิจของธนาคารก็เช่นกัน ส่วนใหญ่องค์กรธุรกิจทางธนาคารจะมีระบบที่รับฟังความคิดเห็นต่างๆจากลูกค้า หรือมีช่องทางสำหรับแจ้งเรื่องร้องเรียน ขอความอนุเคราะห์ หรือแม้กระทั่งการชมเชยก็ตาม ซึ่งช่องทางส่วนใหญ่จะเป็นในรูปแบบ สาขา เว็บไซต์ คอลเซ็นเตอร์ หรืออีเมล เป็นต้น

ซึ่งระบบของธนาคารต่างๆ ส่วนมากจะไม่มีให้ลูกค้าสามารถกรอกในส่วนของประเภทผลิตภัณฑ์หรือบริการ ที่ต้องการแจ้งข้อมูลได้ จึงทำให้การที่จะทราบได้ว่าเป็นผลิตภัณฑ์หรือบริการใดที่ลูกค้าแจ้งเข้ามานั้น พนักงานที่ดูแลระบบจะต้องอ่านทุกข้อความถึงจะระบุได้ว่าเป็นผลิตภัณฑ์ในด้านใด หรือจำแนกว่าเกี่ยวข้องกับเรื่องใด ซึ่งขั้นตอนในส่วนนี้จะทำให้ใช้เวลาเป็นอย่างมาก และทำให้การที่จะส่งเรื่องไปยังหน่วยงานต่างๆ เพื่อดำเนินการในส่วนที่เกี่ยวข้องมีความล่าช้า จนอาจทำให้การแก้ไขปัญหาไม่ทันเวลา และไม่เป็นที่พึงพอใจต่อลูกค้าได้

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

วิธีดำเนินการ



ภาพประกอบ 1 แสดงกระบวนการดำเนินงานวิจัย

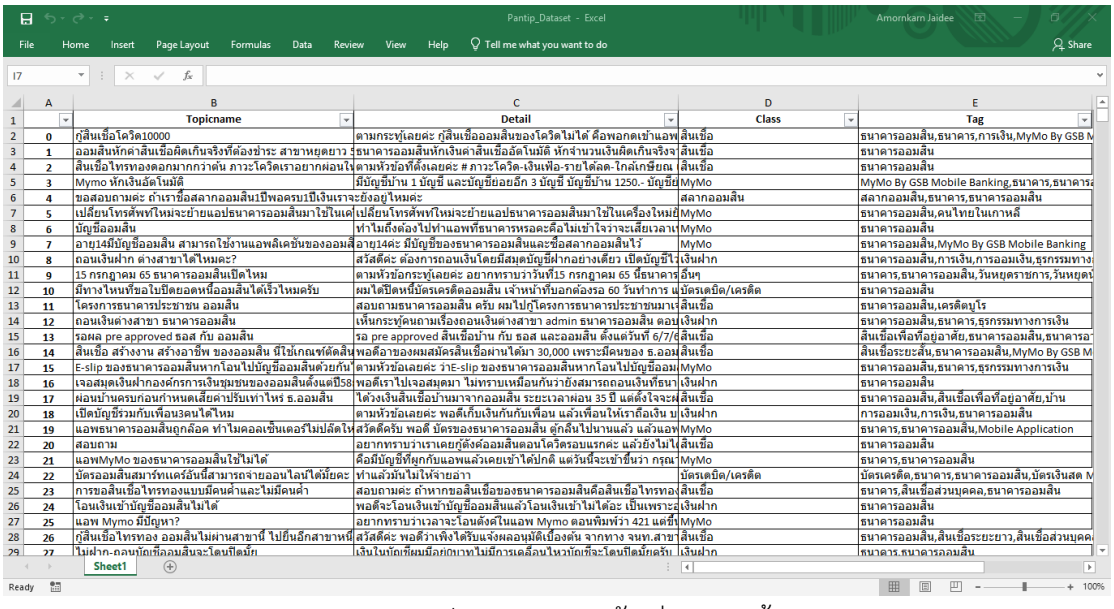
ขั้นตอนที่ 1 : การเก็บรวบรวมข้อมูล

ในงานวิจัยนี้เก็บรวบรวมข้อมูลที่เป็นข้อความจากความคิดเห็น ข้อเสนอแนะ ขอความอนุเคราะห์ และข้อร้องเรียน จากการทำ Web scraping โดยใช้ชุดคำสั่ง Selenium ในการดึงข้อมูลจากเว็บไซต์ www.pantip.com และเลือกแท็กเป็นธนาคารออมสิน ซึ่งผู้วิจัยได้ทำการดึงข้อมูล ณ วันที่ 13 กรกฎาคม 2565 มีจำนวนกระทู้ทั้งหมด 600 กระทู้ โดยข้อมูลจะประกอบไปด้วย 3 คอลัมน์ คือ หัวข้อกระทู้ (Topic name), รายละเอียดในกระทู้ (Detail) และแท็ก (Tag)

เนื่องจากข้อมูลดังกล่าว เป็นข้อมูลที่ยังไม่ได้มีการระบุประเภทของผลิตภัณฑ์ไว้ ทางผู้วิจัยจึงได้ทำการเพิ่มคอลัมน์ “Class” ที่เป็นคอลัมน์ target เพื่อระบุประเภทของข้อมูลกระทู้ (Labeled data)

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



	A	B	C	D	E
1		Topicname	Detail	Class	Tag
2	0	กู้สินเชื่อโควิด10000	ตามกระทู้เลยค่ะ กู้สินเชื่อออมสินของโควิดไม่ได้ คือพอกดเข้าแอพ	สินเชื่อ	ธนาคารออมสิน,ธนาคาร,การเงิน,MyMo By GSB M
3	1	ออมสินหักค่าสินเชื่อผิดเกินจริงที่ต้องชำระ สาขาหยุดยาว 5	ธนาคารออมสินหักเงินค่าสินเชื่ออัตโนมัติ หักจำนวนเงินผิดเกินจริงจ	สินเชื่อ	ธนาคารออมสิน
4	2	สินเชื่อไทรทองดอกมากกว่าต้น ภาวะโควิดเราอยากผ่อนไ	ตามหัวข้อที่ตั้งเลยค่ะ # ภาวะโควิด-เงินเฟ้อ-รายได้ลด-ใกล้เกษียณ	สินเชื่อ	ธนาคารออมสิน
5	3	Mymo หักเงินอัตโนมัติ	มีบัญชีบ้าน 1 บัญชี และบัญชีย่อยอีก 3 บัญชี บัญชีบ้าน 1250.- บัญชีย	MyMo	MyMo By GSB Mobile Banking,ธนาคาร,ธนาคาร
6	4	ขอสอบถามค่ะ ถ้าเราซื้อสลากออมสินเป็นพอครบ1ปีเงินเราจะยังอยู่ไหมค่ะ		สลากออมสิน	สลากออมสิน,ธนาคาร,ธนาคารออมสิน
7	5	เปลี่ยนโทรศัพท์ใหม่จะย้ายแอปธนาคารออมสินมาใช้ในเค	เปลี่ยนโทรศัพท์ใหม่จะย้ายแอปธนาคารออมสินมาใช้ในเครื่องใหม่	MyMo	ธนาคารออมสิน,คนไทยในเกาหลี
8	6	บัญชีออมสิน	ทำไมถึงต้องไปทำแอพที่ธนาคารหรอคะคือไม่เข้าใจว่าจะเสียเวลาเ	MyMo	ธนาคารออมสิน
9	7	อายุ14มีบัญชีออมสิน สามารถใช้งานแอพลิเคชั่นของออมสิ	อายุ14ค่ะ มีบัญชีของธนาคารออมสินและซื้อสลากออมสินไว้	MyMo	ธนาคารออมสิน,MyMo By GSB Mobile Banking
10	8	ถอนเงินฝาก ต่างสาขาได้ไหมคะ?	สวัสดีค่ะ ต้องการถอนเงินโดยมีสมุดบัญชีฝากอย่างเดียว เปิดบัญชีไ	เงินฝาก	ธนาคารออมสิน,การเงิน,การออมเงิน,ธุรกรรมทาง
11	9	15 กรกฎาคม 65 ธนาคารออมสินเปิดใหม	ตามหัวข้อกระทู้เลยค่ะ อยากทราบว่าวันที่15 กรกฎาคม 65 นี้ธนาคา	อื่นๆ	ธนาคาร,ธนาคารออมสิน,วันหยุดราชการ,วันหยุดป
12	10	มีทางไหนที่ขอใบปิดยอดหนี้ออมสินได้เร็วไหมครับ	ผมได้ปิดหนี้บัตรเครดิตออมสิน เจ้าหน้าที่บอกต้องรอ 60 วันทำการ แ	บัตรเดบิต/เครดิต	ธนาคารออมสิน
13	11	โครงการธนาคารประชาชน ออมสิน	สอบถามธนาคารออมสิน ครับ ผมไปกู้โครงการธนาคารประชาชนมาเ	สินเชื่อ	ธนาคารออมสิน,เครดิตบูโร
14	12	ถอนเงินต่างสาขา ธนาคารออมสิน	เห็นกระทู้คนถามเรื่องถอนเงินต่างสาขา admin ธนาคารออมสิน ตอบ	เงินฝาก	ธนาคารออมสิน,ธนาคาร,ธุรกรรมทางการเงิน
15	13	รอผล pre approved ธอส กับ ออมสิน	รอ pre approved สินเชื่อบ้าน กับ ธอส และออมสิน ตั้งแต่วันที่ 6/7/6	สินเชื่อ	สินเชื่อเพื่อที่อยู่อาศัย,ธนาคารออมสิน,ธนาคารอ
16	14	สินเชื่อ สร้างงาน สร้างอาชีพ ของออมสิน นี่ใช้เกณฑ์ตัดสิน	พอดีอาของผมสมัครสินเชื่อผ่านได้มา 30,000 เพราะมีคนของ ธ.ออม	สินเชื่อ	สินเชื่อระยะสั้น,ธนาคารออมสิน,MyMo By GSB M
17	15	E-slip ของธนาคารออมสินหากโอนไปบัญชีออมสินด้วยกัน	ตามหัวข้อเลยค่ะ ว่าE-slip ของธนาคารออมสินหากโอนไปบัญชีออม	MyMo	ธนาคารออมสิน,ธนาคาร,ธุรกรรมทางการเงิน
18	16	เจอสมุดเงินฝากองค์กรการเงินชุมชนของออมสินตั้งแต่ปี58	พอดีเราไปเจอสมุดมา ไม่ทราบเหมือนกันว่ายังสามารถถอนเงินที่ธนา	เงินฝาก	ธนาคารออมสิน
19	17	ผ่อนบ้านครบกำหนดเสียค่าปรับเท่าไหร่ ธ.ออมสิน	ได้วงเงินสินเชื่อบ้านมาจากออมสิน ระยะเวลาผ่อน 35 ปี แต่ตั้งใจจะผ	สินเชื่อ	ธนาคารออมสิน,สินเชื่อเพื่อที่อยู่อาศัย,บ้าน
20	18	เปิดบัญชีร่วมกับเพื่อน3คนได้ไหม	ตามหัวข้อเลยค่ะ พอดีเก็บเงินกันกับเพื่อน แล้วเพื่อนให้เราถือเงิน บ	เงินฝาก	การออมเงิน,การเงิน,ธนาคารออมสิน
21	19	แอพธนาคารออมสินถูกล็อค ทำไมคอลเซ็นเตอร์ไม่ปลดให	สวัสดีครับ พอดี บัตรของธนาคารออมสิน ถูกฉีนไปนานแล้ว แล้วแอพ	MyMo	ธนาคาร,ธนาคารออมสิน,Mobile Application
22	20	สอบถาม	อยากทราบว่าเราเคยกู้ดิ้งค์ออมสินตอนโควิดรอบแรกค่ะ แล้วยังไม่ไ	สินเชื่อ	ธนาคารออมสิน
23	21	แอพMyMo ของธนาคารออมสินใช้ไม่ได้	คือมีบัญชีที่ผูกกับแอพแล้วเคยเข้าได้ปกติ แต่วันนี้จะเข้าขึ้นว่า กรุณ	MyMo	ธนาคาร,ธนาคารออมสิน
24	22	บัตรออมสินสมาร์ทแคร์อันนี้สามารถจ่ายออนไลน์ได้มั้ยคะ	ทำแล้วมันไม่ให้จ่ายอ่า	บัตรเดบิต/เครดิต	บัตรเครดิต,ธนาคาร,ธนาคารออมสิน,บัตรเงินสด M
25	23	การขอสินเชื่อไทรทองแบบมีคนค้ำและไม่มีคนค้ำ	สอบถามค่ะ ถ้าหากขอสินเชื่อของธนาคารออมสินคือสินเชื่อไทรทอง	สินเชื่อ	ธนาคาร,สินเชื่อส่วนบุคคล,ธนาคารออมสิน
26	24	โอนเงินเข้าบัญชีออมสินไม่ได้	พอดีจะโอนเงินเข้าบัญชีออมสินแล้วโอนเงินเข้าไม่ได้อะ เป็นเพราะอ	เงินฝาก	ธนาคารออมสิน
27	25	แอพ Mymo มีปัญหา?	อยากทราบว่าเวลาจะโอนดังค์ในแอพ Mymo ตอนพิมพ์ว่า 421 แต่ขึ้น	MyMo	ธนาคารออมสิน
28	26	กู้สินเชื่อไทรทอง ออมสินไม่ผ่านสาขาหนึ่	สวัสดีค่ะ พอดีว่าเพิ่งได้รับแจ้งผลอนุมัติเบื้องต้น จากทาง จนท.สาขา	สินเชื่อ	ธนาคารออมสิน,สินเชื่อระยะยาว,สินเชื่อส่วนบุคค
29	27	ไม่ฝาก-ถอนบัตรเพื่อออมสินจะโดนปิดบัญ	เงินในบัตรที่ผมมีอยู่ปัจจุบันไม่มีการเคลื่อนไหวบัตรจะโดนปิดบัญครับ	เงินฝาก	ธนาคาร,ธนาคารออมสิน

ภาพประกอบ 2 แสดงตัวอย่างของชุดข้อมูล

ขั้นตอนที่ 2 : การนำเข้าข้อมูล ตรวจสอบข้อมูล และเตรียมความพร้อมของข้อมูล

ใช้ภาษาไพทอนในการวิเคราะห์ข้อมูลและการเรียนรู้ของเครื่อง โดยเริ่มจากการนำเข้าไฟล์ชุดข้อมูล เพื่อเข้าสู่การเตรียมความพร้อมของข้อมูลที่เป็นข้อความที่ใช้ไลบรารี่ PyThaiNLP ได้แก่ การแบ่งประโยค ออกจากข้อความที่มีความยาว โดยใช้อัลกอริทึม crfcut การตัดคำออกจากประโยค โดยใช้อัลกอริทึม newmm การลบคำที่ไม่สื่อความหมายของประโยค การสร้างกลุ่มคำ (Word Cloud) เพื่อดูคำที่ปรากฏบ่อยในชุดข้อมูล รวมถึงการคัดแยกคำตามความสำคัญ (TF-IDF) ในการคำนวณน้ำหนักของคำในชุดข้อมูล สำหรับสร้างคุณสมบัติ (feature) เพื่อใช้กับแบบจำลองต่างๆ โดยใช้ TfidfVectorizer กำหนด max_features=200

ขั้นตอนที่ 3 : การแบ่งชุดข้อมูล

ผู้วิจัยได้นำข้อมูลที่มีการระบุประเภทไว้มาทำการแบ่งชุดข้อมูล ออกเป็น ชุดข้อมูลฝึกฝนและชุดข้อมูลทดสอบที่อัตราส่วน 70:30 จากนั้นได้ทำการแบ่งชุดข้อมูลฝึกฝนออกเป็นครึ่งหนึ่งอีกครั้งเป็นส่วนที่มีการระบุประเภทและอีกส่วนที่เราจะทำเหมือนว่าชุดข้อมูลนั้นไม่มีการระบุประเภทไว้ ในอัตราส่วนที่ 70:30 เช่นกัน จึงได้จำนวนข้อมูลของแต่ละชุดข้อมูลเป็น ชุดข้อมูลฝึกฝน จำนวน 294 ตัวอย่าง, ชุดข้อมูลทดสอบ จำนวน 180 ตัวอย่าง และชุดข้อมูลที่ไม่ได้มีการระบุประเภท จำนวน 126 ตัวอย่าง

ขั้นตอนที่ 4 : การสร้างแบบจำลอง

ในงานวิจัยนี้ได้ใช้แบบจำลองที่เหมาะสมกับงาน Classification ทางผู้วิจัยได้ดำเนินการทดลองเปรียบเทียบประสิทธิภาพของแบบจำลองจำนวน 3 แบบจำลอง โดยใช้ค่ามาตรฐาน (Default) ของแต่ละแบบจำลอง และทำการฝึกฝนแบบจำลองร่วมกับการทำ 5-Folds Cross Validation ได้แก่

- แบบจำลองที่ 1 Support Vector Machine (SVM)
- แบบจำลองที่ 2 Naïve Bayes (NB)
- แบบจำลองที่ 3 Logistic Regression (LR)

จากนั้นได้นำแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝน มาทำนายชุดข้อมูลที่ไม่ได้มีการระบุประเภท เพื่อให้สามารถจำแนกประเภทของผลิตภัณฑ์ธนาคารจากข้อความเหล่านั้นได้ หรือที่เรียกว่า Pseudo Label ซึ่งจะเลือกข้อมูล Pseudo Label Data ที่มีค่าความน่าจะเป็น (Probability) ผ่านเกณฑ์ของค่าความเชื่อมั่น (Confidence Level) ซึ่งในงานวิจัยนี้จะใช้เป็นค่ามัธยฐาน (Median) เพื่อเพิ่มลงในชุดข้อมูลฝึกฝน โดยจะทำซ้ำเรื่อย ๆ จนกว่าจะไม่มีข้อมูลที่ผ่านเกณฑ์ และจากนั้นจึงนำแบบจำลองมาฝึกฝนกับชุดข้อมูลฝึกฝนที่มีทั้ง Labeled data และ Pseudo label data

ขั้นตอนที่ 5 : การประเมินผลแบบจำลอง

ในงานวิจัยนี้ใช้เทคนิคการประเมินประสิทธิภาพของแบบจำลองที่เป็นมาตรฐาน และเหมาะสำหรับการจำแนกประเภท (Classification) 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 ที่ใช้ในแอปพลิเคชัน ที่ข้อมูล 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

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

ผลการวิจัยและอภิปรายผลการวิจัย

จากการทดลอง ประเมิน และเปรียบเทียบประสิทธิภาพ ของแบบจำลอง SVM, Logistic Regression และ Naïve Bayes ได้ผลลัพธ์ ดังต่อไปนี้

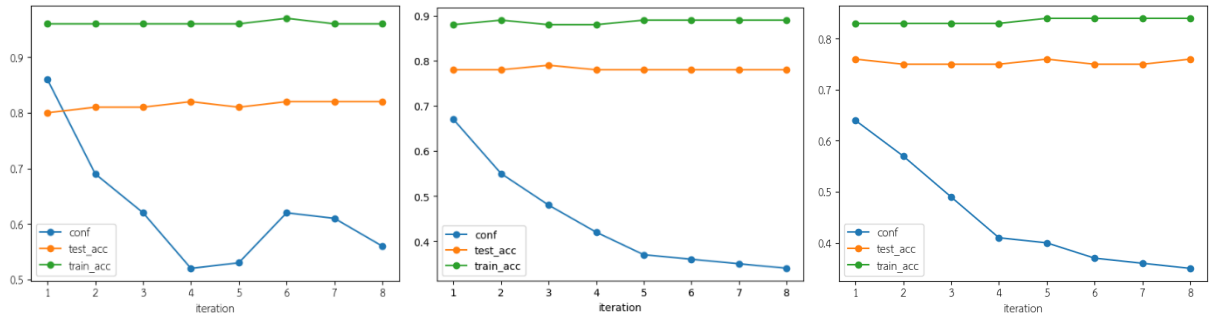
1. ผลลัพธ์ของการจำแนกประเภทของผลิตภัณฑ์ธนาคาร ด้วยเทคนิคการเรียนรู้ของเครื่องแบบกึ่งมีผู้สอน (Semi-Supervised Learning)

ตาราง 2 แสดงจำนวนข้อมูลที่ผ่านเกณฑ์หรือไม่ผ่านเกณฑ์ของค่าความเชื่อมั่น

แบบจำลอง	จำนวนข้อมูลที่ผ่านเกณฑ์ค่าความเชื่อมั่น	จำนวนข้อมูลที่ไม่ผ่านเกณฑ์ค่าความเชื่อมั่น
SVM	97	29
Logistic Regression	92	34
Naïve Bayes	90	36

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 ตามผลลัพธ์ที่แสดงในตาราง 2

2. ผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น



ภาพประกอบ 3 แสดงผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น ของแบบจำลอง SVM, LR, NB

จากผลลัพธ์ของความสัมพันธ์ระหว่าง ค่าความแม่นยำของชุดข้อมูลฝึกฝน ชุดข้อมูลทดสอบ และค่าความเชื่อมั่น จะพบว่าค่าความเชื่อมั่นมีค่าลดลงเรื่อย ๆ เนื่องจากข้อมูลที่ไม่ได้มีการระบุประเภท มีจำนวนลดลง ส่งผลให้ข้อมูลที่ผ่านเกณฑ์ของค่าความเชื่อมั่นมีจำนวนลดลงไปด้วย และเมื่อมีการเพิ่มข้อมูล Pseudo Label Data ลงในชุดข้อมูลฝึกฝนจะพบว่า หากมีข้อมูลที่ถูกเพิ่มเป็นจำนวนมาก จะส่งผลให้ความแม่นยำของชุดข้อมูลฝึกฝน และชุดข้อมูลทดสอบเพิ่มมากขึ้น ซึ่งจะเห็นได้ชัดจากแบบจำลอง Support Vector Machine (SVM) ตามภาพประกอบ 3

3. เปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝนทั้งหมด

ตาราง 3 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

จากแบบจำลองที่ได้จากการฝึกฝนกับชุดข้อมูลฝึกฝนที่ประกอบไปด้วย Labeled data และ Pseudo label data และนำมาทำนายกับชุดข้อมูลทดสอบ จะได้ผลลัพธ์ของการเปรียบเทียบประสิทธิภาพของแบบจำลองที่ผ่านการฝึกฝนกับชุดข้อมูลฝึกฝนทั้งหมด จะเห็นได้ว่าทั้ง 3 แบบจำลอง จะมีค่าความแม่นยำที่ห่างกันไม่มากนัก แต่แบบจำลอง SVM เป็นแบบจำลองที่มี

ประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารได้ดีกว่าแบบจำลองอื่นๆ 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 ดังตาราง 3

สรุปผลการวิจัย

งานวิจัยนี้ได้ทำการจำแนกประเภทของผลิตภัณฑ์ธนาคาร 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 ได้แก่ Support Vector Machine (SVM), Logistic Regression (LR) และ Naïve Bayes (NB) จากผลการทดลองพบว่า แบบจำลอง SVM มีประสิทธิภาพในการจำแนกประเภทของผลิตภัณฑ์ธนาคารดีที่สุด

งานวิจัยในอนาคต ผู้วิจัยจะมีการเพิ่มจำนวนข้อมูลที่ใช้ในการฝึกฝน และจัดการกับ Imbalanced data มีการทดสอบกับค่าทางสถิติอื่นๆ รวมถึงการใช้เทคนิคการปรับจูนพารามิเตอร์ (Hyperparameter Tuning) เพื่อหาพารามิเตอร์ที่เหมาะสมกับแบบจำลอง และทำให้ประสิทธิภาพของแบบจำลองเพิ่มมากขึ้น

กิตติกรรมประกาศ

การจัดทำวิจัยได้รับการสนับสนุนจากบัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ ในการนำเสนอผลงานวิจัย ผู้วิจัยจึงขอขอบคุณมา ณ ที่นี้

เอกสารอ้างอิง

[1] 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),

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

[3] 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.

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

[5] 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),

[6] 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),

[7] 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),

[8] 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),

[9] 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),