การรู้จำอาหารไทยจากรูปภาพโดยใช้เทคนิคการเรียนรู้แบบโครงข่ายประสาทแบบคอนโวลูชัน

คณุตม์ เมฆอัคฆกรณ์[1] และ ศิวะวงศ์ วงศ์เลิศชัย[2]

[1]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

[2]คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

Emails: not41298@gmail.com, swonglerthchai@gmail.com

บทคัดย่อ

ในปัจจุบันการรักษาสุขภาพได้รับความนิยมมากขึ้น ไม่ว่าจะเป็นการออกกำลังกาย การพักผ่อนให้เพียงพอ และการเลือกรับประทานอาหารที่ดีต่อสุขภาพ ในงานวิจัยนี้สนใจในการทำนายรูปภาพอาหาร โดยเฉพาะอาหารไทย เราจึงนำการรู้จำภาพ ที่เป็นเทคนิคหนึ่งของการเรียนรู้ด้วยเครื่องมาใช้เพื่อสามารถนำไปใช้กับแอปพลิเคชันในอนาคต โดยจะนำภาพถ่ายของอาหารไทยจากอินเทอร์เน็ตรวบรวมเป็นชุดข้อมูลเพื่อมาใช้กับการรู้จำภาพ และเราได้ใช้ โครงข่ายประสาทเทียมแบบคอนโวลูชัน ซึ่งเป็นหนึ่งในเทคนิคของการเรียนรู้เชิงลึก และได้รับความนิยมมากสำหรับการรู้จำภาพอาหาร โดยเลือกแบบจำลองและงานวิจัยที่อยู่ในหัวข้อการเรียนรู้ของเครื่อง โดยพิจารณาจากประสิทธิภาพสูงและมีอัตราความผิดพลาดต่ำจากเว็บไซต์ PapersWithCode ที่เป็นเว็บไซต์ศูนย์รวมงานวิจัยในหัวข้อการเรียนรู้ของเครื่องเอาไว้ อย่างเช่น InceptionV3 แบบจำลองถูกสร้างขึ้นในปี 2015 มีจุดประสงค์เพิ่มความแม่นยำของการรู้จำรูปภาพโดยเฉพาะรูปภาพที่ซับซ้อนและหลากหลาย และทางเราได้เลือกแบบจำลองที่เกี่ยวข้องกับการรู้จำภาพอาหารไทย เช่น Nu-InNet ที่เป็นแบบจำลองที่มีขนาดเล็กแต่มีความแม่นยำที่สูงในทำนายรูปภาพอาหารไทย จากนั้นจึงนำแบบจำลองที่ได้เลือกเอาไว้มาปรับให้เหมาะสมและได้ผลลัพธ์ที่สูงอย่างเช่น การนำ Image Augmentation มาใช้ร่วมด้วย จากนั้นจึงเลือกแบบจำลองที่ได้ผลลัพธ์มาใช้ และก่อนที่เราจะนำชุดข้อมูลมาฝึกสอนกับแบบจำลอง เราจะนำชุดข้อมูลไปทำ Texture Segmentation หลังจากนั้นนำไปฝึกสอนกับแบบจำลอง U-Net หลังจากนั้นจึงนำไปฝึกแบบจำลองที่มีเป้าหมายสำหรับระบุชื่ออาหารไทย

1. บทนำ

ในปัจจุบันผู้คนให้ความสนใจกับการควบคุมน้ำหนักและการดูแลรักษาสุขภาพของตนมากขึ้น โดยมีสาเหตุมาจากผู้คนต้องการลดการเจ็บป่วยและไม่ต้องการเสียค่ารักษาพยาบาลที่สูง ผู้คนจึงหาข้อมูลและวิธีการดูแลสุขภาพ เช่น วิธีการออกกำลังกายที่ถูกต้อง สารอาหารที่ร่างกายต้องการ อาหารที่รับประทานแล้วส่งผลดีต่อร่างกายและอาหารที่ส่งผลเสียต่อร่างกาย ซึ่งวิธีดูแลสุขภาพที่ง่ายที่สุดคือ การรับประทานอาหารที่สมดุลและมีคุณค่าทางโภชนาการที่ดี เพราะอาหารถือเป็นปัจจัยหลักของการดำเนินชีวิต โดยเราสามารถเลือกรับประทานอาหารที่มีประโยชน์ต่อร่างกาย ที่อ้างอิงมาจากข้อมูลคุณค่าโภชนาการอาหาร โดยใช้อินเทอร์เน็ตที่มีการใช้งานอย่างแพร่หลาย ผู้คนจึงสามารถเข้าถึงข้อมูลเกี่ยวกับอาหารได้ง่ายมากขึ้น รวมไปถึงการใช้แอปพลิเคชันเพื่อค้นหาข้อมูลเกี่ยวกับอาหารผ่านการใช้รูปภาพอีกด้วย

การรู้จำรูปภาพอาหาร เป็นเทคโนโลยีที่ได้รับการให้ความสนใจเป็นอย่างมาก ส่งผลการทำงานวิจัยในด้านการรู้จำรูปภาพ ประเภทรูปภาพอาหารมากยิ่งขึ้น โดยในปี 2015 และปี 2017 มีการจัดการแข่งขันที่ชื่อว่า ILSVRC โดยมีหัวข้อย่อยที่เกี่ยวข้องกับการรู้จำรูปภาพ

อาหารอยู่ด้วย โดยผู้แข่งขันในงานนี้มีการใช้เทคนิคสำหรับการรู้จำภาพอย่างหลากหลายโดยหนึ่งในนั้นคือโครงข่ายประสาทเทียมแบบคอนโวลูชัน ด้วยงานแข่งขันนี้ทำให้เกิดเทคนิคอื่นๆ เพื่อประสิทธิภาพให้กับการรู้จำรูปภาพ เช่น การสกัดคุณลักษณะ การพัฒนาสถาปัตยกรรมของโครงข่าย แต่จากงานวิจัยส่วนใหญ่ล้วนเป็นงานวิจัยที่เกี่ยวข้องกับการรู้จำรูปภาพอาหารทางฝั่งชาวตะวันตก เมื่อเปรียบเทียบกับงานวิจัยหัวข้อการเรียนรู้จำรูปภาพอาหารชาวเอเชีย จะพบว่ามีจำนวนน้อย โดยเฉพาะอย่างยิ่งในหัวข้ออาหารไทยที่มีจำนวนไม่มาก เพราะอาหารไทยยังไม่เป็นที่นิยมมากเมื่อเปรียบอาหารตะวันตก และอาหารตะวันตกมีจุดเด่นที่ในหลายๆ เมนูเราสามารถแยกวัตถุในเมนูได้ด้วยตาเปล่า ซึ่งแตกต่างจากอาหารชาวเอเชียที่หลายๆ เมนูใส่วัตถุหลายอย่าง และใช้ตาเปล่าแยกวัตถุดิบได้ยาก ด้วยเหตุผลนี้เราจึงเล็งเห็นปัญหานี้จึงได้เริ่มศึกษางานวิจัยเกี่ยวกับ การรู้จำรูปภาพอาหารไทย ที่มีอยู่ อย่างงานวิจัยของคุณ Natta และคณะที่มีจุดประสงค์ในการหาปริมาณแคลอรี่ของอาหารจากรูปภาพ มุ่งเน้นไปที่อาหารไทย โดยใช้เครื่องมือหลายอย่างที่เกี่ยวกับการจัดกลุ่มภาพ และการสกัดคุณลักษณะซึ่งเป็นการจัดเตรียมชุดข้อมูล และใช้ฝึกฝนกับแบบจำลองแบบ SVM ซึ่งแบบจำลองสามารถทำนายอาหารไทยได้แม่นยำ แต่ยังทำนายจำนวนแคลอรี่ได้ไม่แม่นยำนัก[1] งานวิจัยต่อไปของคุณ Punnarumol และคณะ ได้นำ transfer learning มาใช้เพื่อลดระยะเวลาในการฝึกสอนแบบจำลอง และจะได้ความแม่นยำเพิ่มขึ้นด้วย แต่เมื่อนำแบบจำลองที่ได้มาใช้ในโทรศัพท์ในทำนายอาหารถึงแม้จะมีความแม่นยำที่สูงแต่ใช้เวลาในการทำนายที่นาน[2] และในหลายๆ งานวิจัยในหัวข้อการรู้ภาพอาหารมักเจอปัญหาอย่าง แบบจำลองมีประสิทธิภาพและความแม่นยำสูง แต่มีของแบบจำลองมีขนาดใหญ่ ใช้เวลาฝึกสอนนาน ใช้ทรัพยากรที่สูงสำหรับการฝึกสอนแบบจำลอง ชุดข้อมูล มีข้อมูลอื่นๆ ที่ไม่เกี่ยวข้องอยู่ด้วยทำให้เวลาฝึกสอนแบบจำลองแล้วได้ประสิทธิภาพที่ต่ำ ดังนั้นเราจึงได้ทำ Image segmentation ก่อนจะนำไปฝึกสอนแบบจำลอง และทำ fine-tuned กับแบบจำลอง งานวิจัยนี้มีจุดประสงค์เพื่อหาแบบจำลองที่ทำนายรูปภาพอาหารไทยได้แม่นยำที่สุด และมีขนาดเล็กพอเพื่อนำไปใช้กับแอพลิเคชันทำนายรูปภาพอาหารในภายหลัง และงานวิจัยในหัวข้อการรู้จำรูปภาพไทย ก็มักจะมีข้อเสียที่เหมือนๆกันคือขนาดของชุดข้อมูลที่ใช้ฝึกสอนนั้นมีจำนวนที่น้อยเกินไป และเครื่องมือที่ใช้ฝึกสอนแบบจำลองที่มีประสิทธิภาพที่ไม่สูงทำให้ฝึกสอนแบบจำลองได้น้อย

2. การทบทวนวรรณกรรมที่เกี่ยวข้อง

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1การรู้จำรูปภาพ (Image Recognition)

การรู้จำรูปภาพ คือกระบวนการระบุหรือแยก สิ่งของหรือรูปแบบจากรูปภาพหรือวิดีโอ โดยใช้อัลกอริทึมและเทคนิคของการเรียนรู้ด้วยเครื่อง(Machine Learning) โดยมีจุดประสงค์ให้คอมพิวเตอร์ สามารถเข้าใจหรือตีความข้อมูลจากภาพ ช่วยให้คอมพิวเตอร์สามารถจำแนกหรือจัดหมวดหมู่จาก รูปภาพ เทคนิคที่นิยมใช้ในการรู้จำรูปภาพคือโครงข่ายประสาทเทียมแบบคอนโวลูชัน(Convolution neural networks:CNNs) ในปัจจุบันการรู้จำรูปภาพได้ถูกนำไปใช้หลากหลายได้แก่ รถยนต์ไร้คนขับ ระบบรักษาความปลอดภัย ภาพทางการแพทย์ ระบบจดจำใบหน้า

2.1.2 โครงข่ายประสาทเทียมแบบคอนโวลูชัน (Convolution Neural Network: CNN)

โครงข่ายประสาทเทียมแบบคอนโวลูชัน ใช้หลักการจากพีชคณิตเชิงเส้น โดยเฉพาะ การคูณเม ทริกซ์เพื่อระบุรูปแบบภายในภาพ โครงข่ายประสาทเทียมแบบคอนโวลูชันแตกต่างจากโครงข่ายประสาทอื่นๆ ด้วยประสิทธิภาพที่ดีกว่าในการป้อนข้อมูลภาพ เสียงพูด หรือเสียง โครงข่ายประสาทเทียม แบบคอนโวลูชัน ประกอบไปด้วยเลเยอร์หลัก 3 ประเภทได้แก่

คอนโวลูชันเลเยอร์(Convolution layer) เป็นส่วนสำคัญของโครงข่ายประสาทเทียมแบบคอน โวลูชัน และเป็นที่ที่มีการคำนวณเกิดขึ้นมากที่สุด โดยคอนโวลูชันเลเยอร์นั้นต้องการข้อมูล อินพุต ตัวกรอง(Filter) และFeature Mapโดยในเลเยอร์นี้จะมีการสกัด Feature จากข้อมูลอินพุต ผ่านตัวกรอกโดยกระบวนการนี้เราเรียกว่า คอนโวลูชัน โดยหลังจากเกิดกระบวนการคอนโวลู ชันแต่ละ

ครั้ง จะใช้ ReLU เพื่อเปลี่ยนผลลัพธ์ให้เป็น Feature Map เพื่อเพิ่มความไม่เป็นเส้นตรง ให้กับแบบจำลอง

พูลลิงเลเยอร์ (Pooling Layer) หรือที่รู้จักกันว่า Downsampling เป็นเลเยอร์ที่จะลดขนาดมิติ ของข้อมูลอินพุตและลดจำนวน Parameter ของข้อมูลอินพุต โดยจะส่งผลให้ขนาดของ Feature map ลดลงด้วย

เลเยอร์เชื่อมต่ออย่างสมบูรณ์(Fully Connected Layer) ในเลเยอร์นี้ทุกๆ โหนดจะเชื่อมต่อกัน หมดทั้งโหนดข้อมูลอินพุตหรือโหนดจากเลเยอร์ก่อนหน้า ไปยังโหนดข้อมูลเอาท์พุต

2.1.3 การแปลงฟูรีเย (Fourier transform)

การแปลงฟูรีเย คือเทคนิคการแปลงรูปภาพโดยจะทำการแปลงโดเมนตำแหน่ง (Spatial Domain) ไปเป็นโดเมนความถี่ (Frequency domain) โดยมีวัตถุประสงค์ของการแปลงรูปภาพคือเพื่อระบุหรือ สกัดคุณลักษณะที่ไม่สามารถทำได้ในโดเมนตำแหน่ง การแปลงฟูริเยจะทำการแยกองค์ประกอบ รูปภาพขาเข้า ให้อยู่ในองค์ประกอบ sin กับ cos บนโดเมนความถี่ ซึ่งจะสอดคล้องกับโดเมนตำแหน่ง โดยการแปลงฟูรีเยมีอยู่ 2 แบบหลักๆ คือ

1. การแปลงฟูรีเยแบบไม่ต่อเนื่อง (Discrete Fourier Transform) ใช้สำหรับการวิเคราะห์สัญญาณ โดยแปลงโดเมนเวลาจากสัญญาณให้เป็นโดเมนความถี่ การแปลงแบบฟูรีเยแบบไม่ต่อเนื่อง จะมีข้อจำกัดในการใช้ไม่เหมือนกับการแปลงฟูรีเยแบบเร็ว ตัวอย่างที่นำการแปลงฟูรีเยแบบ ไม่ต่อเนื่องไปใช้งานคือ แอปพลิเคชันประมวลผลสัญญาณดิจิตอล (Digital Signal Processing Application)

2. การแปลงฟูรีเยแบบเร็ว (Fast Fourier Transform) เป็นการนำ การแปลงฟูรีเยแบบไม่ต่อเนื่องมา ปรับปรุงโดยจะเปรียบเสมือนการแปลงฟูรีเยแบบไม่ต่อเนื่องที่คำนวณได้เร็วขึ้น โดยจะแปลง 14 โดเมนตำแหน่ง หรือพื้นที่ให้เป็นโดเมนความถี่ ตัวอย่างที่นำการแปลงฟูรีเยแบบเร็วไปใช้งาน คือ การประมวลผลกับรูปภาพ การประมวลผลกับเสียง

2.1.4 การรวมข้อมูล (Data Fusion)

การรวมข้อมูลคือกระบวนการ การนำข้อมูลจากหลาย ๆ แหล่ง หรือข้อมูลที่อยู่คนละประเภท กัน มาผสมผสานกันโดยมีจุดประสงค์เพื่อเพิ่มประสิทธิภาพและความแม่นยำให้กับแบบจำลอง เช่น ข้อมูลประเภทข้อความกับข้อมูลประเภทรูปภาพ การรวมข้อมูลเป็นเทคนิคที่มีมานานแต่จะมีส่วน หลักๆ ดังนี้

1. การรวมข้อมูลขาเข้า (Input Level Fusion) เป็นการรวมชุดข้อมูลจากหลายแหล่งและเป็นชุด ข้อมูลที่ยังไม่ผ่านการประมวลผล จากนั้นจึงนำมาฝึกสอนแบบจำลอง

2. การรวมข้อมูลคุณลักษณะ (Feature Level Fusion) เป็นการสกัดคุณลักษณะจากชุดข้อมูลจาก หลายแหล่ง โดยจะสกัดคุณลักษณะแย่งกันจากนั้นจึงนำคุณลักษณะที่ได้มาต่อกัน จากนั้นจึงนำ ไปฝึกสอนแบบจำลอง

3. การข้อมูลการตัดสินใจ (Decision Level Fusion) เป็นรวมผลลัพธ์ของแบบจำลองที่การฝึกฝนด้วยชุดข้อมูลที่แตกต่างกัน

2.1.5 แบบจำลอง CentralNet

CentralNet เป็นแบบจำลองที่ถูกเสนอโดย คุณ Valentin และคณะ โดยแบบจำลองจะใช้ชุดข้อมูลแตกต่าง กัน 2 ประเภทสำหรับการฝึกฝน โดยทุกๆ เลเยอร์ของแบบจำลองจะทำรวม โดยสิ่งที่นำมารวมคือคุณลักษณะที่สกัดมาได้ จากนั้นนำคุณลักษณะที่สกัดได้มาฝึกฝน จากนั้นจึงนำผลลัพธ์ที่ได้ทั้งหมด มาทำนายโดยแบบจำลอง CentralNet สามารถใช้หาความสัมพันธ์ระหว่าง 2 ข้อมูลได้ เช่นการจำแนก วัตถุที่มีการพูดคุยในวิดีโอ หรือการระบุอารมณ์จากภาพและเสียง

2.2 งานวิจัยที่เกี่ยวข้อง

การรู้จำรูปภาพอาหารเป็นหัวข้องานวิจัยที่ได้รับความนิยมมาก โดยส่วนใหญ่จะเป็นอาหารในประเทศตะวันตก หรืออาหารญี่ปุ่น รวมไปถึงอาหารไทยด้วย ถึงแม้

จะ จำนวนงานวิจัยในหัวข้อนี้จะมีไม่เยอะเท่าอาหารในประเทศที่กล่าวมา แต่เราก็ได้เลือกที่จะศึกษางาน วิจัยที่มีหัวข้อหลักเป็น การรู้จำรูปภาพอาหารไทย อย่างในปี 2014 คุณ Natta และคณะ ได้เลือกใช้เทคนิค Texture Segmentation จากการตั้งสมมุติฐานที่ว่า ในส่วนอื่นๆ ที่ไม่เกี่ยวข้องกับ อาหารจะมีรายละเอียดของ Texture ที่น้อยกว่าส่วนที่เกี่ยวข้องกับอาหารที่อยู่ในรูปภาพ จึง แบ่งรูปภาพออกเป็นส่วนที่เกี่ยวข้องกับอาหารและส่วนที่ไม่เกี่ยวข้อง จากนั้นทำการ masking และ crop รูปภาพเพื่อตัดส่วนที่เป็นสีดำออกก่อนจะนำไปฝึกสอน แบบจำลอง [1]

ในปี 2017 คุณ Chakkrit Termritthikun และคณะ ได้นำเสนอแบบจำลองที่ชื่อว่า nu-innet ซึ่งดัดแปลงมาจาก inception module ของ googLeNet เพื่อให้เหมาะสมกับการนำไปใช้ในโทรศัพท์ในเวอร์ชั่น เพื่อลดจำนวนของพารามิเตอร์ลง แต่ยังคงประสิทธิภาพใน การทำนายรูปภาพเอาไว้ ทำให้สามารถฝึกแบบจำลองได้ในเวลาไม่นานและแบบจำลองมีขนาดเล็ก[3] และในปีเดียวกันคุณ Chakkrit และคณะก็ได้นำ nu-innet ทั้ง 2 เวอร์ชันที่กล่าวมา มาปรับปรุง โดย เพิ่มความลึกให้กับตัว Inception module ที่ผ่านการดัดแปลง ซึ่งในงานวิจัยนี้ใช้ค่าความลึกเมื่อความเพิ่มความลึกจะส่งผลให้ฝึกสอนใช้เวลานานขึ้นและมีจำนวน พารามิเตอร์ที่สูงขึ้น เมื่อเปรียบเทียบผลลัพธ์ทั้ง 4 ความลึกแล้ว ค่าความลึกที่ 4 นั้นจะมีความแม่นยำ มากที่สุดและเมื่อเปรียบเทียบกับแบบจำลอง GoogLeNetและแบบจำลอง BN-Inception ที่ผ่านการ ฝึกสอนด้วยชุดข้อมูลเดี่ยวกัน Nu-InNet ที่มีการกำหนดค่าความลึกเป็น 4 ก็มีค่าความแม่นยำที่สูงกว่า มี parameter จำนวนที่น้อยและแบบจำลองมีขนาดเล็กตามจุดประสงค์ในการสร้างแบบจำลองที่ว่า ต้องการนำแบบจำลองไปใช้กับแอพลิเคชันในโทรศัพท์มือถือ[4]

3. วิธีการดำเนินการวิจัย

3.1 ชุดข้อมูลที่ใช้

1) THFOOD-50 คือชุดข้อมูลรูปภาพอาหารไทยที่เป็นที่นิยม 50 ชนิดถูกเก็บรวบรวม ซึ่งรูปภาพ เหล่านี้ถูกนำมาจากเครื่องมือค้นหาต่างๆ ได้แก่ Google, Bing และ Flickr โดยแต่ละชนิดของ อาหารไทยนั้นมีราว ๆ 200 ถึง 700 ภาพ และถูกแบ่งออกเป็น 2 ชุดคือชุดฝึกสอน จำนวน 80% และชุดทดสอบ จำนวน 20% ชุดข้อมูลนี้มีจำนวน รายการอาหารไทยที่เป็นของคาวทั้งหมด 50 รายการ มีจำนวนรูปภาพทั้งหมด 15,772 ภาพ แบ่ง

2) FoodyDudy คือชุดข้อมูลรูปภาพอาหารไทยที่เป็นที่นิยม 48 เมนู ผู้เก็บรวบรวมชุดข้อมูลนี้คือคุณ Phacharaphol (Gemmy) Somboontham มีจุดประสงค์เพื่อให้ชาวต่างชาติสามารถรู้ได้ว่าเมนูนี้มีชื่อว่าอะไร และนำแบบ จำลองที่ฝึกด้วยข้อมูลนี้ไปใส่ในเว็บแอพลิเคชัน โดยมีรูปภาพทั้งหมด 14,400 รูปภาพ

3) ชุดข้อมูลรูปภาพอาหารไทยที่เก็บด้วยตนเอง เก็บโดยใช้สคริปต์ดาวโหลดรูปภาพจากเว็บไซต์ Google โดยนำรายการอาหารมาจากตารางคุณโภชนาการปีพศ.2561 ของสำนักโภชนาการ มีรูปภาพทั้งหมด 18,923 รูปภาพ

3.2 การตั้งค่าแบบจำลองและการประเมินผล

ในการทดลองเราได้แบ่งชุดข้อมูลเป็นชุดข้อมูลฝึกฝนแบบจำลองและชุดข้อมูลทดสอบแบบจำลอง โดยแบ่งเป็นอัตราส่วน 80% เป็นชุดข้อมูลที่ใช้ฝึกฝนแบบจำลอง และ 20% ทำการตั้งค่าพารามิเตอร์กับแบบจำลองดังนี้

1) ค่าอัตราการเรียนรู้ (learning rate) ถูกกำหนดเป็น 0.001 เนื่องจากค่านี้เหมาะสมกับงานการจัด แยกอาหาร และช่วยลดความเสี่ยงในการเกิด overfitting

2) การใช้ Dropout เป็นเทคนิคที่ช่วยลดการเกิด overfitting ในโมเดล

3) จำนวนรอบการฝึก กำหนดรอบการฝึกเป็น 200 เนื่องจากโมเดล Food classification จำเป็นต้อง ฝึกฝนให้มากพอที่จะได้รับการเรียนรู้อย่างเต็มที่

4) การใช้งาน Early stopping เป็นเทคนิคในการหยุดการฝึกโมเดลก่อนจนกว่าจะเต็มจำนวนรอบ การฝึก โดยใช้เกณฑ์เป็นค่าความสูญเสียของชุดข้อมูลทดสอบ (validation loss) เพื่อลดเวลาใน การฝึกและลดโอกาสในการเกิด overfitting

5) การใช้งาน Model checkpointing เป็นเทคนิคในการบันทึกโมเดลที่ดีที่สุดหลังจากแต่ละรอบ การฝึก เพื่อใช้ในการทดสอบหรือใช้กับชุดข้อมูลใหม่ในอนาคต

6) การใช้ Optimizer Optimizer ถูกกำหนดเป็น Adam optimizer เพื่อช่วยให้โมเดลฝึกฝนได้อย่างมีประสิทธิภาพและความเร็วในการเรียนรู้ที่มาก

7) การใช้งาน Cross Validation ใช้เทคนิค K-fold cross-validation โดยกำหนด K=5 เพื่อทดสอบ และวัดประสิทธิภาพของโมเดลโดยการแบ่งชุดข้อมูลการเรียนรู้เป็น 5 ส่วนเท่าๆ กัน โดยจะเลือกใช้ 4 ส่วนในการแบ่งชุดข้อมูลการเรียนรู้และใช้ส่วนที่เหลือ 1 ส่วนเป็นชุดข้อมูลทดสอบซึ่งนำทดสอบแต่ละรอบจะใช้ส่วนที่ไม่เหมือนกัน เพื่อให้ทุกส่วนของชุดข้อมูลถูกนำมาใช้ทั้งในการสร้างแบบจำลองและการทดสอบแบบจำลอง จะทำซ้ำกัน 5 รอบ เพื่อให้ทุกส่วนของชุดข้อมูลถูกนำมาใช้ฝึกฝน

ในส่วนของการวัดประสิทธิผลในเราได้นำชุดข้อมูลทดสอบซึ่งประกอบไป ด้วยภาพที่คำตอบที่ถูกต้องและคำตอบที่ถูกทำนายโดยโมเดล โดยใช้การเปรียบเทียบ คำตอบที่ถูกต้องกับคำตอบที่ถูกทำนายด้วยเกณฑ์ โดย Top-1 Accuracy คือ ความแม่นยำของโมเดล ที่ทำนายคำตอบถูกต้องในหมวดหมู่ที่เป็นที่สุดหรือหมวดหมู่ที่มีความน่าจะเป็นสูงที่สุดและ Top-5 Accuracy คือ ความแม่นยำของโมเดลที่ทำนายคำตอบถูกต้องในหมวดหมู่ที่ถูกต้อง หรือหมวดหมู่ที่มี ความน่าจะเป็นสูงถึง 5 อันดับแรก

3.3 การฝึกฝนแบบจำลอง

ในการฝึกฝนแบบจำลอง เราได้ใช้แบบจำลองที่เกี่ยวข้องกับการรู้จำภาพอาหารและแบบจำลองที่ เป็น State of the Art ทั้งหมด 10 แบบจำลองได้แก่ Nu-inNet 1.0 , Nu-inNet 1.1, AlexNet, SqueezeNet, Resnet50V2, InceptionV3, DenseNet201, MobileNetV2, GoogLeNet, EfficientNetV2 นั้นเราได้แบ่งวิธีการฝึกแบบจำลองเป็นดังนี้

1) การฝึกฝนแบบจำลองด้วยชุดข้อมูลที่ไม่ผ่านการทำ Image Segmentation

2) การฝึกฝนแบบจำลองด้วยชุดข้อมูลที่ผ่านการทำ Image Segmentation ด้วยโปรแกรม LabelMe ในการทำ Masking เพื่อครอปรูปภาพจากนั้นจึงนำผลลัพธ์ที่ได้ไปใช้ฝึกฝนแบบจำลอง

3) การฝึกฝนแบบจำลองวิธี end-to-end โดยการทำ Image Segmentation กับชุดข้อมูลด้วยโปรแกรม LabelMe ในการทำ Masking เพื่อครอปรูปภาพซึ่งจะคล้ายๆ กับวิธีที่ 2 แต่เราจะนำผลลัพธ์จากการทำ Masking ไปฝึกกับแบบจำลอง U-Net และนำผลลัพธ์ที่ได้ไปทดสอบกับแบบจำลองที่ได้จากข้อที่ 1

4) การฝึกฝนแบบจำลองโดยนำ Fast Fourier transform เข้ามาใช้กับชุดข้อมูลขาเข้า โดยจะเปลี่ยนรูปภาพให้เป็นโดเมนความถี่และฝึกฝนแบบจำลองจากนั้นจะฝึกแบบจำลองด้วยวิธีที่ 1 ถึง 3 จากนั้นทำการรวมด้วย Fully Connected จากนั้นจึงทำการทดสอบแบบจำลอง

5) การฝึกฝนแบบจำลองด้วยวิธีการแบบ CentralNet โดยจะฝึกแบบจำลอง 2 อันไปพร้อมกัน อันแรกเป็นแบบจำลองที่ใช้ Fast Fourier Transform เข้ามาใช้กับชุดข้อมูลขาเข้า และอันที่ 2 จะฝึกฝนด้วยชุดข้อมูลแบบข้อที่ 1 โดยทั้ง 2 แบบจำลองจะมีการสกัดคุณลักษณะของทุกเลเยอร์ จากนั้นเราจะทำนำคุณลักษณะที่สกัดได้นั้น มารวม จากนั้นจะนำคุณลักษณะที่ได้สกัดจากเลเยอร์ออกมาแล้วนำไปฝึกต่อจากนั้นจึงนำแบบจำลองมาทดสอบ

4. ผลการทดลอง

4.1 ผลการรวมชุดข้อมูลด้วยตนเอง

นำเมนูอาหารไทยทั้งหมด 234 รายการมาค้นหารูปภาพใน search engine อย่างเช่น google โดยจะดูผลลัพธ์ที่ได้หลังจากค้นหารูปภาพแล้วว่า ตรงกับที่เราต้องการหรือไม่ โดยสิ่งที่เราต้องการคือ รูปภาพเมนูที่เรา

ต้องการ จากนั้น จึงเขียน script เพื่อดาวน์โหลดรูปภาพจากผลลัพธ์ที่ได้ และนำรูปภาพที่ดาวน์โหลดสำเร็จแล้วมาเปลี่ยนชื่อไฟล์ให้เมนูอาหารไทยแบบภาษาอังกฤษ จากนั้นจะทำการคัดเลือกรูปภาพที่ตรงกับเมนูที่เราค้นหาเห็นเมนูอาหารอย่างชัดเจนที่สุดและมีขนาดของรูปภาพขั้นต่ำเป็น 300x300 โดยจะรวมรูปภาพที่ผ่านการคัดเลือกแล้ว ให้อยู่ในโฟลเดอร์เพื่อจัดการได้ง่ายและได้แบบจำลองที่ทำนายผลลัพธ์ได้ดีที่สุด จากรายการอาหารทั้งหมดจำนวน 234 รายการ มีบางรายการที่ซ้ำกันได้แก่ ซุปหน่อไม้, แกงหอยขม, ข้าวหมกไก่, ขนมจีนน้ำยาป่า, ขนมจีนน้ำพริก และมีบางรายการที่ไม่สามารถหารูปภาพได้ ได้แก่ ปลาดุกอุยทอดรสเผ็ด, ปลาหลังขาว 3 รส, ปลาปากคม 3 รส, นาชิดาแฆ, นาชิดาแฆกุ้ง, นาชิมีเยาะ, นาชิตีเน๊ะ, หอยแครงปรุงรส, ปลาทู 3 รส, ต้มข่าทูน่า, ปลาทูน่าผสมกระเทียม, หมี่กรอบอาหารเจ, ปลาเส้นผสมสาหร่ายอาหารเจ, แกงอ่อมมะระใส่เห็ดอาหารเจ, หมู - ไตอาหารเจ, ยำก๋วยเตี๋ยวเซียงไฮ้อาหารเจ, ขนมจีนน้ำละซอ, ปลาซาร์ดีนในซอสมะเขือเทศ เป็น 24 รายการจึงเหลือรานการทั้งหมด 210 รายการ ชุดข้อมูลนี้มีจำนวนรูปภาพทั้งหมด 18923 รูปภาพ

4.2 ผลลัพธ์ของแบบจำลองของชุดข้อมูลที่รวบรวมด้วยตนเอง

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	64.8 ± 5.4	84.8 ± 7.1
Nu-inNet1.0	62.1 ± 6.9	82.7 ± 7.0
AlexNet	53.9 ± 2.9	76.8 ± 8.5
SqueezeNet	63.2 ± 8.6	91.3 ± 4.1
Resnet50V2	59.4 ± 15.9	84.5 ± 7.5
InceptionV3	63.4 ± 12.2	88.0 ± 10.0
DenseNet201	55.9 ± 9.3	83.4 ± 6.1
Mobilenet_v2	57.0 ± 4.2	77.5 ± 5.4
GoogLeNet	63.9 ± 2.8	87.7 ± 3.6
EfficientNetV2	64.3 ± 4.3	86.5 ± 3.3

รูปที่ 1 ผลลัพธ์ของแบบจำลองที่ไม่ผ่านการทำ Image Segmentation

จากรูปที่ 1 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 64.8% และ Top-5 สูงสุดคือ SqueezeNet ที่ 91.3%

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	67.4 ± 5.8	87.0 ± 5.3
Nu-inNet1.0	65.0 ± 5.3	85.2 ± 7.3
AlexNet	56.4 ± 3.6	79.0 ± 10.6
SqueezeNet	65.6 ± 8.5	92.5 ± 3.4
Resnet50V2	62.4 ± 14.9	86.0 ± 7.5
InceptionV3	66.1 ± 14.2	89.5 ± 9.9
DenseNet201	58.8 ± 8.5	85.6 ± 2.2
Mobilenet_v2	59.2 ± 4.2	78.8 ± 6.1
GoogLeNet	65.4 ± 2.5	89.3 ± 3.4
EfficientNetV2	66.5 ± 3.9	88.2 ± 3.9

รูปที่ 2 ผลลัพธ์ของแบบจำลองที่ฝึกด้วยชุดข้อมูลที่ผ่านทำ Image Segmentation

จากรูปที่ 2 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 67.4% และ Top-5 สูงสุดคือ SqueezeNet ที่ 92.5% และเปรียบเทียบผลลัพธ์ระหว่างรูปที่ 1 กับ รูปที่ 2 การทำ Image Segmentation นั้นสามารถเพิ่มค่าความแม่นยำเฉลี่ย Top-1 ได้ถึง 1.5-3% และ ค่าความแม่นยำเฉลี่ย Top-5 เพิ่มขึ้น 1.1-2.8%

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	65.9 ± 5.5	87.1 ± 3.2
Nu-inNet1.0	65.6 ± 3.4	83.8 ± 7.3
AlexNet	55.1 ± 3.3	77.3 ± 11.7
SqueezeNet	64.9 ± 8.5	86.8 ± 3.2
Resnet50V2	61.4 ± 15.1	84.7 ± 7.9
InceptionV3	64.9 ± 13.8	88.5 ± 9.9
DenseNet201	57.6 ± 8.5	84.4 ± 2.8
Mobilenet_v2	58.3 ± 3.9	77.9 ± 6.0
GoogLeNet	64.0 ± 2.5	86.0 ± 1.8
EfficientNetV2	65.0 ± 3.9	86.9 ± 3.9

รูปที่ 3 ผลลัพธ์ของแบบจำลองที่ฝึกด้วยเทคนิค end-to-end

จากรูปที่ 3 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 65.9% และ Top-5 สูงสุดคือ InceptionV3 ที่ 88.5% และเปรียบเทียบผลลัพธ์ระหว่างรูปที่ 1 กับ รูปที่ 3 การฝึกฝนแบบจำลองโดยใช้เทคนิค End-to-End ค่าความ แม่นยำเฉลี่ย Top-1 มีค่าเพิ่มขึ้น 0.1-2% และค่าความแม่นยำเฉลี่ย Top-5 มีค่าเพิ่มขึ้น 0.2-1.6%

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	66.7 ± 5.0	88.6 ± 2.8
EfficientNetV2	66.4 ± 3.5	88.3 ± 2.8
InceptionV3	65.5 ± 10.9	89.4 ± 8.9
Resnet50	61.3 ± 14.2	86.2 ± 6.4

รูปที่ 4 ผลลัพธ์ของแบบจำลองที่ใช้ Fast Fourier Transform

จากรูปที่ 4 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 66.7% และ Top-5 สูงสุดคือ InceptionV3 ที่ 89.4% และเมื่อเปรียบเทียบระหว่างรูปที่ 1 กับรูปที่ 4 แบบจำลองทั้ง 4 นั้นมีค่าความแม่นยำเฉลี่ย Top-1 เพิ่มขึ้นถึง 1.9-2.1 % และค่าความแม่นยำเฉลี่ย Top-5 เพิ่มขึ้น 1.5-1.9 %

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	69.3 ± 5.3	90.0 ± 2.1
EfficientNetV2	68.5 ± 3.4	89.9 ± 3.1
InceptionV3	68.2 ± 12.0	91.0 ± 8.5
Resnet50	64.4 ± 13.2	87.9 ± 7.0

รูปที่ 5 ผลลัพธ์ของแบบจำลองที่ใช้ Fast Fourier Transform และฝึกฝนด้วยชุดข้อมูลที่ผ่านการทำ Image Segmentation

จากรูปที่ 5 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 69.3% และ Top-5 สูงสุดคือ InceptionV3 ที่ 91.0% และเมื่อเปรียบเทียบระหว่างรูปที่ 2 กับรูปที่ 5 แบบจำลองทั้ง 4 นั้นมีค่าความแม่นยำเฉลี่ย Top-1 เพิ่มขึ้นถึง 1.9-2.1 % และค่าความแม่นยำเฉลี่ย Top-5 เพิ่มขึ้น 1.5-1.9 %

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	67.8 ± 5.0	89.1 ± 3.1
EfficientNetV2	67.1 ± 3.6	88.5 ± 3.5
InceptionV3	67.2 ± 11.7	89.9 ± 8.7
Resnet50	63.9 ± 13.1	86.2 ± 7.2

รูปที่ 6 ผลลัพธ์ของแบบจำลองที่ใช้ Fast Fourier Transform และเทคนิค End-to-End

จากรูปที่ 6 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 67.8% และ Top-5 สูงสุดคือ InceptionV3 ที่ 89.9% และเมื่อเปรียบเทียบระหว่างรูปที่ 3 กับรูปที่ 6 แบบจำลองทั้ง 4 นั้นมีค่าความแม่นยำเฉลี่ย Top-1 เพิ่มขึ้นถึง 1.8-2.6 % และค่าความแม่นยำเฉลี่ย Top-5 เพิ่มขึ้น 1.4-1.9 %

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	70.1 ± 5.0	90.8 ± 2.8
EfficientNetV2	69.9 ± 3.2	90.7 ± 2.7
InceptionV3	69.1 ± 9.7	91.4 ± 8.2
Resnet50	68.2 ± 7.6	89.3 ± 6.0

รูปที่ 7 ผลลัพธ์ของแบบจำลองฝึกด้วยวิธีแบบ CentralNet

จากรูปที่ 7 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 70.1% และ Top-5 สูงสุดคือ InceptionV3 ที่ 91.4% และเมื่อเปรียบเทียบระหว่างรูปที่ 1 กับรูปที่ 7 มีค่าความแม่นยำเฉลี่ย Top-1 เพิ่มขึ้นถึง 5.3-8.8 % และมีค่าความ แม่นยำเฉลี่ย Top-5 เพิ่มขึ้นถึง 3.4-4.8 %

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	72.7 ± 4.9	93.6 ± 2.4
EfficientNetV2	72.2 ± 3.8	92.7 ± 2.3
InceptionV3	72.1 ± 9.2	93.4 ± 7.8
Resnet50	71.1 ± 7.9	91.7 ± 5.8

รูปที่ 8 ผลลัพธ์ของแบบจำลองฝึกด้วยวิธีแบบ CentralNet ด้วยชุดข้อมูลที่ผ่านการทำ Image Segmentation

จากรูปที่ 8 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 72.7% และ Top-5 สูงสุดคือ InceptionV3 ที่ 93.4% และเมื่อเปรียบเทียบระหว่างรูปที่ 2 กับรูปที่ 8 มีค่าความแม่นยำเฉลี่ย Top-1 เพิ่มขึ้นถึง 5.3-8.7 % และมีค่าความ แม่นยำเฉลี่ย Top-5 เพิ่มขึ้นถึง 3.9-5.7 %

Model	Top-1 Avg Accuracy(%)	Top-5 Avg Accuracy(%)
Nu-innet1.1	71.0 ± 4.9	92.2 ± 2.5
EfficientNetV2	70.8 ± 3.3	91.2 ± 2.7
InceptionV3	70.1 ± 9.4	92.0 ± 8.2
Resnet50	69.3 ± 7.5	90.1 ± 5.8

รูปที่ 9 ผลลัพธ์ของแบบจำลองฝึกด้วยวิธีแบบ CentralNet ด้วยเทคนิค End-to-End

จากรูปที่ 9 แบบจำลองที่มีค่าความแม่นยำเฉลี่ย Top-1 สูงสุดคือ Nu-innet1.1 ที่ 71.0% และ Top-5 สูงสุดคือ Nu-innet1.1 ที่ 92.2% และเมื่อเปรียบเทียบระหว่างรูปที่ 3 กับรูปที่ 9 ค่าความแม่นยำเฉลี่ย Top-1 เพิ่มขึ้นถึง 5.1-7.9 % และมีค่าความ แม่นยำเฉลี่ย Top-5 เพิ่มขึ้นถึง 3.5-5.3 %

5. บทสรุป

5.1 การรวบรวมชุดข้อมูลรูปภาพอาหารไทย

การรวบรวมชุดข้อมูลรูปภาพอาหารไทย จะอาหารบางรายการที่บางรายการที่มีจำนวนรูปภาพที่น้อย เมื่อไปฝึกแบบจำลองจะส่งผลให้มีความแม่นยำลดลง โดยมีแนวทางการแก้ไขปัญหานี้คือ ควรค้นหารูปภาพจากหลายๆ

แหล่ง เช่น ค้นหาชื่ออาหารในเว็บไซต์ที่เป็น search engine อื่นๆ อย่าง Bing และควรค้นหาชื่ออาหารในเว็บไซต์ที่มีการเก็บรวบรวมรูปอย่าง Flickr หรือเว็บไซต์ที่มีความ เกี่ยวข้องกับอาหารอย่างเว็บไซต์ Wongnai ที่มีคนรีวิวอาหารพร้อมกับรูปภาพอาหาร จากแนวทางที่ได้เสนออาจทำหารูปภาพได้มากขึ้น

5.2 การรวบรวมชุดข้อมูลรูปภาพอาหารไทย

การฝึกแบบจำลองกับชุดข้อมูลที่รวบรวมด้วยตนเองนั้น การใช้การทำ Image Segmentation นั้นได้ประสิทธิสูงกว่าเมื่อเทียบกับเทคนิค End-to-End โดยค่าความแม่นยำเฉลี่ย Top-1 นั้นมีค่ามากกว่าถึง 0.5-1.5 % และค่าความแม่นยำเฉลี่ย Top-5 มี ค่ามากกว่าถึง 0.8-1.7 % ดังในรูปที่ 2 กับรูปที่ 3

การใช้ Fast Fourier Transform ร่วมกับการทำ Image Segmentation นั้นมีประสิทธิภาพที่ดีกว่าการใช้เทคนิค End-to-End โดย Image Segmentation มีค่าความแม่นยำเฉลี่ย Top-1 มากกว่าการใช้เทคนิค End-to-End ถึง 0.5-1.6 % และมี ค่าความแม่นยำเฉลี่ย Top-5 มากกว่าการใช้เทคนิค End-to-End ถึง 0.9-1.7 % ดังรูปที่ 5 และ 6

การรวมข้อมูลแบบ CentralNet ร่วมกับ การทำ Image Segmentation นั้นมีประสิทธิภาพที่ดีกว่าการรวมข้อมูลแบบ CentralNet ร่วมกับการ ใช้เทคนิค End-to-End โดย Image Segmentation มีค่าความแม่นยำเฉลี่ย Top-1 มากกว่าการใช้เทคนิค End-to-End ถึง 1.5-2 % และมีค่าความแม่นยำเฉลี่ย Top-5 มากกว่าการใช้เทคนิค End-to-End ถึง 1.4- 1.7 % ดังรูปที่ 8 กับรูปที่ 9

เอกสารอ้างอิง

[1] N. Tammachat and N. Pantuwong, “Calories analysis of food intake using image recognition,” 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), 2014.

[2] P. Temdeeand S. Uttama,“Food recognition on smartphone using transfer learning of convolution neural network,” 2017 Global Wireless Summit (GWS), 2017.

[3] C. Termritthikun, P. Muneesawang, and S. Kanprachar, “Nu-innet: Thai food image recognition using convolutional neural networks on smartphone,” Journal of Telecommunication, Electronic and Computer Engineering (JTEC), vol. 9, no. 2-6, pp. 63–67, 2017.

[4] C. Termritthikun and S. Kanprachar, “Accuracy improvement of Thai food image recognition using deep convolutional neural networks,” in 2017 international electrical engineering congress (IEECON). IEEE, 2017, pp. 1–4

วิเคราะห์และทำนายความนิยมของคอนเทนต์ในสื่อสังคมออนไลน์โดยการเรียนรู้ด้วยเครื่องคอมพิวเตอร์

Analysis and prediction of social media content popularity based on machine learning approach

บทคัดย่อ

โครงงานมีวัตถุประสงค์เพื่อสร้างโมเดลที่สามารถทำนายความนิยมของคอนเทนต์ในสื่อสังคมออนไลน์ด้วยการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ โดยมุ่งหวังให้ผู้สร้างคอนเทนต์นำผลการทำนายที่ได้ไปประกอบการตัดสินใจต่อไป ข้อมูลคอนเทนต์ที่ใช้ในการศึกษาทั้งหมดได้รวบรวมมาจากโพสบน Facebook fanpage สาธารณะ เฉพาะที่เขียนด้วยภาษาไทยเป็นหลักและเกี่ยวกับการขายสินค้า ข่าวสาร หรือการให้ความรู้ “คะแนนความนิยม” จะคำนวณจากจำนวนการกดถูกใจคอนเทนต์ คณะผู้วิจัยได้นำอัลกอริทึม XGBoost มา พัฒนาโมเดลทั้งหมด 2 ประเภท ได้แก่ Classification Model และ Regression Model ผลการทดลองพบว่า Classification Model มีความแม่นยำในการทำนายอยู่ที่ 38.45% โดยที่ความแม่นยำจะเพิ่มขึ้นเป็น 61.97% เมื่อนับรวมคำทำนายของคลาสข้างเคียงว่าเป็นค่าที่ถูกต้องด้วย (ความแม่นยำข้างเคียง) ส่วน Regression Model มีค่าเฉลี่ยความผิดพลาดสัมบูรณ์อยู่ที่ 150.34%

KEYWORDS

social media content , popularity analysis and prediction, machine learning

1. INTRODUCTION ปัจจุบันคนส่วนใหญ่หันมาใช้โซเชียลมิเดียในการดำเนินชีวิตและการประกอบอาชีพกันมากยิ่งขึ้น โดย Facebook เป็นแพลตฟอร์มที่มีความนิยมเป็นอย่างมากของนักเขียนคอนเทนต์ในไทย ด้วยความนิยมที่มากและความหลากหลายของรูปแบบการใช้งานจึงทำให้แพลตฟอร์ม Facebook เป็นหนึ่งในช่องทางที่เหมาะสมที่จะนำมาใช้ในการสร้างคอนเทนต์เพื่อโปรโมทสินค้าและบริการ แต่อย่างไรก็ตามการเขียนคอนเทนต์ให้ได้รับความนิยมนั้นไม่ใช่สิ่งที่ไม่ว่าใครก็สามารถจะทำได้ และยังเป็นเรื่องยากที่ผู้เขียนจะทราบได้ว่าคอนเทนต์ที่ตนกำลังเขียนนั้นจะได้รับความนิยม หรือได้รับความสนใจจากผู้อ่านหรือไม่ เพราะปัจจัยที่ส่งผลต่อความนิยมของบทความมีความหลากหลายและเปลี่ยนแปลงไปตามเวลาและสถานการณ์ ดังนั้นในงานวิจัยฉบับนี้คณะผู้จัดทำจึงได้จัดทำโครงงานวิเคราะห์และทำนายความนิยมของคอนเทนต์ในสื่อสังคมออนไลน์โดยการเรียนรู้ด้วยเครื่องคอมพิวเตอร์ขึ้น ด้วยการสร้างโมเดลในรูปแบบต่าง ๆ จาก XGboots เพื่อทำนายผลว่า คอนเทนต์ที่เขียนนั้นจะได้รับความนิยมมากน้อยเพียงใด โดยการนำคอนเทนต์เก่าจาก แพลทฟอร์มของ Facebook มาวิเคราะห์และนำไปให้โมเดลเรียนรู้ โดยมีวัตถุประสงค์เพื่อวิเคราะห์ปัจจัยที่ส่งผลต่อความนิยมของคอนเทนต์ เพื่อสร้างโมเดลที่สามารถทำนายความนิยมของคอนเทนต์ และเพื่อช่วยให้ผู้เขียนคอนเทนต์ได้นำไปใช้ประโยชน์ในการปรับปรุงคอนเทนต์ ของตนเองต่อไป

2. RELATED WORK

2.1 งานวิจัย Sequential Prediction of Social Media Popularity with

จากงานวิจัยนี้ได้เสนอวิธีการจัดการเตรียมข้อมูลของจำนวนการกดไลค์โดยวิธีการ Log normalization เพื่อจัดการกับปัญหาเรื่องการต่างกันของเวลาในแต่ละโพสเนื่องจากการเก็บข้อมูลนั้นเวลาในการโพสที่ผ่านมาของแต่ละโพสไม่เท่ากันจึงจำเป็นต้องทำการ Log

normalization เพื่อเป็นการลดBias ของข้อมูลโดยสมการของ Log normalization มีดังนี้

$$s = \log_2 \frac{r}{d} + 1$$

2.2 งานวิจัย Intrinsic Image Popularity Assessment จากงานวิจัยนี้ได้ทำการสร้างโมเดลการทำนายค่าความนิยมของรูปภาพโดยใช้ Deep learing ในการทำการทดลองและทำนาย

2.3 งานวิจัย NIMA: Neural Image Assessment จากงานวิจัยนี้ได้ทำการทดลองสร้าง convolutional neural network เพื่อใช้ในการทำนาย คุณภาพของรูปภาพ

2.4 Deep Age Estimation From Classification to Ranking จากงานวิจัยนี้ได้เสนอวิธีการทำนาย MultiClass แบบใหม่โดยใช้ Convolutional Neural Network (CNN)-based framework, ranking-CNN โดยวิธีการ แบบเก่าคือทำการทำนาย เลขกลุ่มต่างๆออกมาเป็นจำนวนนับ

3. METHOD

3.1 data gathering

วิธีการที่ 1 คณะวิจัยได้ทำการเก็บรวบรวมข้อมูลจากแฟนเพจสาธารณะจาก เฟซบุ๊คด้วย Facebook Graph api โดยวิธีการเก็บข้อมูลชุดแรกทำการเก็บข้อมูลโดยทำการเก็บข้อมูล 15 โพสล่าสุดของแต่ละแฟนเพจ ทั้งหมด 96 แฟนเพจ จะได้โพสทั้งหมด 1440 โพส

วิธีการที่ 2 เนื่องจากคณะวิจัยได้ทำการทดลองแล้วพบปัญหาเรื่องโพสที่มีเวลาในการโพสที่แตกต่างกันอาจจะทำให้โมเดลเกิด bias ขึ้นได้ คณะวิจัยจึงตัดสินใจทำการเก็บชุดข้อมูลชุดที่ 2 ขึ้นโดยทางคณะวิจัยจะทำการเก็บข้อมูลของ โพสที่พึ่งถูกโพสและเก็บต่อไป ทุกๆ 2ชั่วโมงเป็นเวลา 2อาทิตย์ได้โพสทั้งหมด 506 โพส

3.2 data understanding

ข้อมูลที่ทำการรวบรวมและนำมาใช้ในการเรียนรู้และทำนายผลของโมเดลประกอบด้วย

message post คือ ตัวอักษรในโพส

tag post คือ แท็กที่อยู่ในโพส

create time คือ เวลาที่โพส

like คือ จำนวนการกด like

Page follow count คือ จำนวนยอดติดตามของแฟนเพจ

Page like count คือจำนวนการกดถูกใจของแฟนเพจ

3.3 data prepare

โดยการทำความสะอาดข้อมูล จะทำกับข้อมูลในส่วนของ message post และ tag post โดยจะทำการ ลบอีโมจิ ลบตัวเลข และ ลบอักษรพิเศษ

การเตรียมฟีเจอร์สำหรับการทำนายผล

โดยฟีเจอร์ที่เป็นส่วนประกอบในการทำนายยอดไลค์ได้แก่ 1. message post และ message_tag ทำการแยกคำทำโดยใช้ word_tokenize 2. create time นำ เดือน วัน และชั่วโมง ที่โพสออกมาจาก create time และแบ่งเป็นช่วงของเวลา 3. Image ใช้ข้อมูลที่มาจากโมเดลภายนอก nima , iipa page_followers_count, page_likes_count โดยข้อมูลทั้งหมดที่เหลือรวมถึงข้อมูลที่ทำการแบ่งกลุ่มแล้วนำ normalize ให้อยู่ในค่า 0 – 1 เพื่อให้ข้อมูลที่จำนวนตัวเลขใกล้เคียงกัน

การเตรียมฟีเจอร์สำหรับเป็นเป้าหมายที่จะทำนาย

โดยการทำนายผลจะใช้ 2 วิธี คือ classification และ regression ซึ่งทั้ง 2 วิธีนี้จะมีวิธีการเตรียมข้อมูลที่เอาไว้สำหรับเป็นเป้าหมายที่จะทำนาย เพื่อที่จะนำไปให้โมเดลได้เรียนรู้ไม่เหมือนกัน โดยจะมีวิธีการเตรียมข้อมูล ดังนี้

Classification จะทำการแบ่งค่าไลค์ออกเป็นกลุ่ม โดยจะใช้วิธีแบบ log scale โดยจะได้ข้อมูลในแต่ละช่วงต่างกันเพราะ ในช่วงของจำนวนตัวเลขน้อยๆ เช่น ตัวเลข 20 กับ 30 นั้นจะเห็นได้ว่ามีความต่างกันมาก แต่ตัวเลข 1000 กับ 1010 นั้นต่างกันไม่มากจึงให้ความสำคัญกับตัวเลขที่มีจำนวนน้อยๆมากกว่า

Regression จะทำการเปลี่ยนค่าไลค์ที่มีจำนวนต่างกันโดยการใช้ฟังกชัน log normalization กับค่า ไลค์ ดังสมการนี้

3.4 model

โดยการทดลองนี้จะใช้โมเดล XGBoost ในการทดลอง ซึ่งพัฒนาการต่อยอดมาจาก gradient boosting จึงได้ชื่อว่า extreme gradient boosting ซึ่ง XGBoost พัฒนาขึ้นในหลายๆด้านไม่ว่าจะเป็นลดทรัพยากรในการใช้งาน, ได้ผลทำนายที่ดียิ่งขึ้น, ใช้เวลาน้อยลง

3.5 evaluation

ทางคณะวิจัยได้เลือกใช้ตัวแปรในการวัดค่าของโมเดล regression ได้แก่ 1. R square 2. Mean Squre Error (MSE) 3. Mean Absolute Error(MAE) และ เสนอการวัดแบบ Mean Absolute Error percentage (MAEP) เพื่อที่จะได้เห็นความผิดพลาดของโมเดลได้ง่ายยิ่งขึ้น จากสมการดังนี้

$$(abs(\hat{Y} - Y)/ Y) * 100$$

Y แฮด คือ ค่าที่ทำนายออกมาได้

Y คือ ไลค์จริงๆของโพสนั้น

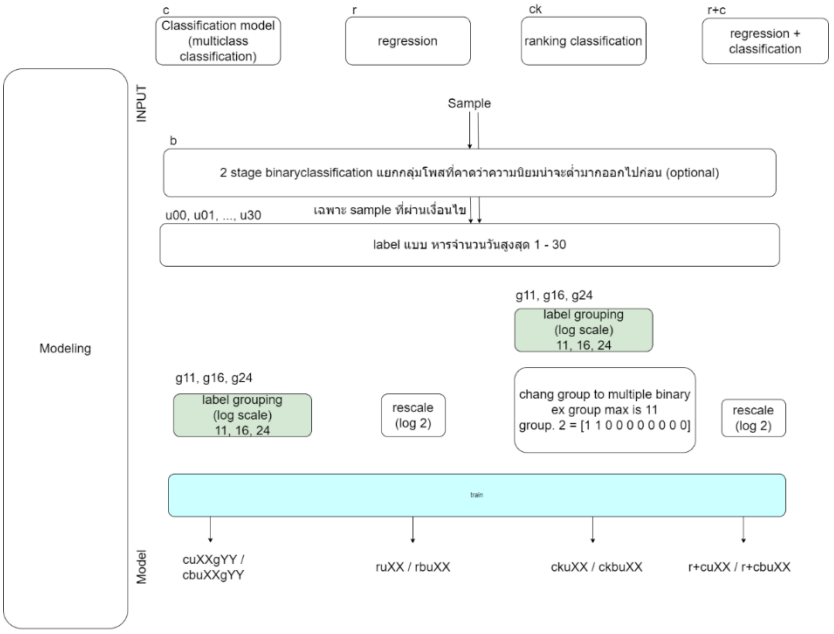
classification ได้แก่ 1. Accuracy 2. Recall 3. F1 4. Precision และเสนอการวัดประสิทธิภาพของโมเดล อีกแบบซึ่งสามารถใช้ได้กับงานนี้คือการวัดประสิทธิภาพแบบ adjacent group (กลุ่มใกล้เคียง) สาเหตุที่ทางคณะวิจัยเสนอการวัดแบบนี้ขึ้นมาเพราะว่าค่าการ prediction นั้นเป็นค่าที่เป็น categories โดยค่าจริงๆที่ได้อาจจะเป็นค่าที่ใกล้กับกลุ่มอื่นๆจึงได้เสนอวิธีการวัดผลแบบนี้ขึ้นมา

3.6 Feature Selection

[9] การเลือกบาง feature ที่มีความสำคัญต่อโมเดลมากๆออกมาใช้สำหรับการสร้างโมเดล เพื่อลดเวลาในการเรียนรู้ของโมเดลได้อีกด้วย จากการศึกษาข้อมูลของ [10]how to choose feature selection ที่ได้มีการทำวิธีการเลือก feature selection จึงทำให้ทางคณะผู้จัดทำมีความสนใจที่จะใช้ feature selection แบบ f_classif โดยการใช้ฟังก์ชั่นทางสถิติ ANOVA มาใช้กับการทำนายกับ regression และในส่วนของ classification จะใช้ chi2 โดยใช้ฟังก์ชั่นทางสถิติ chi square มาช่วยคำนวณด้วย

4. EXPERIMENTS

โดยทางคณะได้ตั้งชื่อโมเดลแต่ละประเภทไว้เป็นตัวอักษรย่อเพื่อให้สามารถเข้าใจได้โดยตั้งไว้เป็นรหัสดังต่อไปนี้



4.1 Classification

พารามิเตอร์

พารามิเตอร์	Default	Normal	Selection
base_score	0.5	0.5	0.5
booster	gbtree	btree	btree
colsample_bylevel	1	1	1
colsample_bynode	1	1	1
colsample_bytree	0.3	0.4	0.3
gamma	0.1	0.3	0.0
learning_rate	0.15	0.25	0.2
max_delta_step	0	0	0
max_depth	12	15	5
min_child_weight	1	7	3
missing	None	None	None
n_estimators	100	100	100
nthread	1	1	1
objective	multi:softprob	multi:softprob	multi:softprob
reg_alpha	0	0	0
reg_lambda	1	1	1
scale_pos_weight	1	1	1
seed	0	0	0
subsample	1	1	1
verbosity	1	1	1

ผลการทดลอง

model-code	Accuracy	recall	f1	precision	Accuracy adjacent-1
cu00g11	0.2478	0.2070	0.2080	0.2175	0.5126
cu01g11	0.2478	0.2070	0.2080	0.2175	0.5126
cu02g11	0.2836	0.2385	0.2314	0.2342	0.5441
cu03g11	0.2962	0.2201	0.2186	0.2215	0.5273
cu04g11	0.3172	0.2182	0.2079	0.2093	0.5735
cu05g11	0.3151	0.2030	0.1994	0.2015	0.5357
cu06g11	0.3172	0.2145	0.2009	0.2067	0.5525
cu07g11	0.3109	0.1907	0.1853	0.2013	0.5441
cu08g11	0.3382	0.2184	0.2127	0.2259	0.5588
cu09g11	0.3424	0.2093	0.2074	0.2214	0.5672
cu10g11	0.3424	0.2038	0.1958	0.1956	0.5546
cu11g11	0.3403	0.2025	0.1957	0.1954	0.5756
cu12g11	0.3550	0.2503	0.2257	0.2177	0.5903
cu13g11	0.344	0.2403	0.2120	0.2050	0.5882
cu14g11	0.3697	0.2574	0.2342	0.2295	0.5882
cu15g11	0.3424	0.2350	0.2122	0.2056	0.5819
cu16g11	0.3487	0.2540	0.2238	0.2153	0.5924
cu17g11	0.3529	0.2450	0.2246	0.2173	0.5945
cu18g11	0.3508	0.2557	0.2309	0.2200	0.6029
cu19g11	0.3697	0.2471	0.2280	0.2233	0.6092
cu20g11	0.3655	0.2517	0.2244	0.2175	0.6113
cu21g11	0.3781	0.2448	0.2283	0.2238	0.6071
cu22g11	0.3676	0.2488	0.2331	0.2288	0.6008
cu23g11	0.3697	0.2682	0.2386	0.2292	0.6050
cu24g11	0.3760	0.2611	0.2369	0.2292	0.6113
cu25g11	0.3571	0.2298	0.2167	0.2146	0.6029
cu26g11	0.3676	0.2378	0.2198	0.2145	0.6092
cu27g11	0.3466	0.2304	0.2077	0.2017	0.6050
cu28g11	0.3571	0.2264	0.2101	0.2056	0.6050
cu29g11	0.3697	0.2316	0.2086	0.2038	0.5945
cu30g11	0.3760	0.2440	0.224	0.2181	0.6071

ผลการทดลองที่ดีที่สุด

model-code	Accuracy	recall	f1	precision	Accuracy adjacent-1
cu29g16	0.3844	0.2170	0.2084	0.2056	0.6176
cbu03g11	0.2155	0.2864	0.2151	0.2060	0.5868
cu23g11	0.3697	0.2682	0.2386	0.2292	0.6050
cu04g16	0.3172	0.2220	0.2192	0.2343	0.5735
cu25g16	0.3634	0.2129	0.2012	0.1988	0.6197

adjacence accuracy คือการคิด accuracy แบบกลุ่มค้างเคียงด้วย

adjacence accuracy – 1 หมายถึง ถ้าค่า predict ได้แล้วไปตกอยู่ในกลุ่มข้างเคียง 1 กลุ่มจากกลุ่มที่ predict ได้จริงๆจะถือว่าทายถูก

จะเห็นได้ว่า accuracy นั้นเพิ่มขึ้นอย่างเห็นได้ชัดเนื่องจากมีการทำนายส่วนมากนั้นตกไปอยู่กลุ่มข้างเคียง

Classification model (binary classification)

ทำการทดลองทำนายจำแนกโพสว่าโพสไหนที่จะมีค่าความนิยมโดยเฉลี่ยต่อวันน้อยกว่า 50 โดยจะจะเป็นโมเดลเสริมสำหรับนำมาใช้ในส่วนของ regression

4.2 Regression

พารามิเตอร์

พารามิเตอร์	Default	Normal	Selection
base_score	0.5	0.5	0.5
booster	gbtree	btree	btree
colsample_bylevel	1	1	1
colsample_bynode	1	1	1
colsample_bytree	1	0.7	1
gamma	0	0	0
learning_rate	0.1	0.1	0.04
max_delta_step	0	0	0
max_depth	3	5	6
min_child_weight	1	1	1
missing	None	None	None
n_estimators	100	100	200
nthread	1	1	1
objective	reg:linear	reg:linear	reg:linear
reg_alpha	0	0	0
reg_lambda	1	1	1
scale_pos_weight	1	1	1
seed	0	0	0
subsample	1	0.8	0.44
verbosity	1	1	1
importance_type	gain	gain	gain

ผลการทดลองไม่ผ่านการทำ binary class

Model code	mae	mse	rmse	r2	mean%
ru01	1630.66	20447825.89	4521.92	0.2692	2409.23
ru02	1588.80	17899638.06	4230.79	0.3603	1872.98
ru03	1628.04	21941593.72	4684.18	0.2158	1605.92
ru04	1576.30	18975614.26	4356.10	0.3218	1439.32
ru05	1582.73	18460453.11	4296.56	0.3402	2217.77
ru06	1555.75	17920223.6	4233.22	0.3595	2172.38
ru07	1582.26	18445069.55	4294.77	0.3408	1762.08
ru08	1617.23	19504647.32	4416.40	0.3029	2388.63
ru09	1573.65	18128751.7	4257.78	0.3521	2253.68
ru10	1635.21	20090695.13	4482.26	0.2820	2561.26
ru11	1630.17	17511790.72	4184.70	0.3741	3673.45
ru12	1546.36	18180186.46	4263.82	0.3502	2659.40
ru13	1546.50	16797609.18	4098.48	0.3996	2088.57
ru14	1605.07	18637678.65	4317.13	0.3339	2330.14
ru15	1646.76	19549354.19	4421.46	0.3013	3451.59
ru16	1613.81	21042640.57	4587.22	0.2479	2146.81
ru17	1708.93	22675403.2	4761.86	0.1896	3557.98
ru18	1699.29	22431716	4736.21	0.1983	3700.23
ru19	1674.70	22041147.58	4694.80	0.2122	2917.94
ru20	1657.83	20101288.91	4483.44	0.2816	2764.77
ru21	1629.25	20787278.13	4559.30	0.2571	3396.28
ru22	1631.19	19934186.18	4464.77	0.2875	3330.17
ru23	1726.40	22686421.62	4763.02	0.1892	3331.16
ru24	1673.84	21459598.29	4632.45	0.2330	2675.84
ru25	1623.59	21222260.31	4606.76	0.2415	2232.04
ru26	1617.43	21153269.34	4599.26	0.2440	2912.64
ru27	1669.59	20234384.98	4498.26	0.2768	3693.84
ru28	1723.08	22197651.42	4711.43	0.2067	5262.24
ru29	1643.83	20362673.61	4512.50	0.2722	3395.31
ru30	1644.59	20566430.47	4535.02	0.2650	3400.78

ผลการทดลองผ่านการทำ binary class

Model code	mae	mse	rmse	r2	mean%
rbu01	2489.31	29833617.71	5462.01	0.3862	160.08
rbu02	2584.61	35416841.15	5951.20	0.2713	158.34
rbu03	2542.00	34709435.85	5891.47	0.2859	161.70
rbu04	2635.45	33358314.34	5775.66	0.3137	175.69
rbu05	2455.49	31956258.51	5652.98	0.3425	151.81
rbu06	2570.51	35949668.45	5995.80	0.2604	152.94
rbu07	2580.57	33450103.53	5783.60	0.3118	150.33
rbu08	2653.48	35481856.57	5956.66	0.2700	155.42
rbu09	2578.66	34717914.23	5892.19	0.2857	153.78
rbu10	2687.31	39123771.78	6254.89	0.1951	186.21
rbu11	2590.84	35976885.26	5998.07	0.2598	156.44
rbu12	2603.71	36348542.67	6028.97	0.2522	169.06
rbu13	2575.90	36364814.47	6030.32	0.2518	176.21
rbu14	2587.45	35956446.34	5996.36	0.2602	164.97
rbu15	2590.72	38932406.22	6239.58	0.1990	165.49
rbu16	2556.59	36901878.03	6074.69	0.2408	166.15
rbu17	2612.91	38184421.69	6179.35	0.2144	179.11
rbu18	2546.75	36511153.99	6042.44	0.2488	189.53
rbu19	2688.40	36698076.29	6057.89	0.2450	175.22
rbu20	2655.80	37403536.22	6115.84	0.2305	163.29
rbu21	2682.54	38770372.91	6226.58	0.2023	167.18
rbu22	2603.83	35142733.14	5928.13	0.2770	184.20
rbu23	2516.56	35418132.22	5951.31	0.2713	162.57
rbu24	2488.78	35283180.57	5939.96	0.2741	155.12
rbu25	2588.46	34315317.96	5857.92	0.2940	165.56
rbu26	2474.60	31274254.5	5592.33	0.3566	163.10
rbu27	2522.70	30853462.02	5554.58	0.3652	164.80
rbu28	2443.73	31192100.08	5584.98	0.3583	170.20
rbu29	2584.89	37687303.95	6138.99	0.2246	181.90
rbu30	2480.58	32905208.15	5736.30	0.3230	181.51

ผลการทดลองที่ดีที่สุด

model-code	mae	mse	rmse	r2	mean%
ru12	1546.36	18180186.46	4263.82	0.3502	2659.40
ru13	1546.50	16797609.18	4098.48	0.3996	2088.57
rbu07	2580.57	33450103.53	5783.60	0.3118	150.33

จะเห็นได้ว่าเมื่อนำข้อมูลที่มีค่าความนิยมต่ำกว่า 50 ออกไปแล้วทำการทดลอง mean% ที่เราทำการทดลองคิดออกมานั้นลดลงไปอย่างเห็นได้ชัด เมื่อเทียบกับโมเดลที่ไม่ได้นำค่าความนิยมต่ำออก เปอร์เซ็น error นั้นจะมากกว่า 1000 เพราะ regression นั้นทำนายค่าความนิยมต่ำๆได้ไม่ดี

5. CONCLUSION

ในการทดลองทั้งหมดที่ผ่านมา โมเดล Classification , Regression, Regression + Classification และ Ranking Classification นั้นทั้งหมดล้วนมีประสิทธิภาพยังไม่ดีมากนัก โดยสามารถสรุปผลของแต่ละโมเดลได้ดังนี้

1.Classificartion

เป็นวิธีที่ได้ค่าความแม่นยำออกมาดีที่สุดในกลุ่มประเภท ที่ทำนายเป็น Multiclassfication เหมือน ๆ กันและหากใช้ผลการทำนายแบบ กลุ่มใกล้เคียงก็จะยิ่งแม่นยำขึ้นไปอีกแต่ช่วงของค่าความนิยมที่ทำนายออกมาก็จะยิ่งกว้างมากขึ้นไปอีก วิธีการคัดกรองโพสก่อนนำมาทำนายผลไม่ได้ทำให้โมเดลประเภทนี้ดีขึ้น

2.Regression

ค่าความผิดพลาดยังสูงมากอยู่แต่หากทำการกรองโพสที่มีค่าความนิยมต่ำ ๆ ออกไปก่อนทำนายจะช่วยให้ ค่าความผิดพลาดดีขึ้นได้ เนื่องจากโมเดลนี้ทำนายโพสที่มีค่าความนิยมต่อวัน ต่ำๆได้ไม่ดี

REFERENCES

[1] สสปท-TOSH. “New Normal” [Online]

Available: https://www.tosh.or.th/covid-19/index.php/new-normal

[2] Narongyod Mahittivanicha “สถิติและพฤติกรรมการใช้ social media ทั่วโลก Q1 ปี 2020” [Online]

Available: https://www.twfdigital.com/blog/2020/02/global-social-media-usage-stats-q1-2020/

[3] Aksarapak C. “อาชีพ Content Creator คืออะไร” [Online]

Available: https://contentshifu.com/blog/what-is-content-creator

[4] Kingcopywriting. “คอนเทนต์ (Content) คืออะไร” [Online]

Available: https://kingcopywriting.com/what-is-content/

[5] โต๊ะข่าวไอที ดิจิทัล“ข้อมูลล่าสุดสถิติใช้ดิจิทัลในไทยจากรายงาน We are social” [Online]

Available: https://www.bangkokbiznews.com/columnist/989552

[6] James Thorn “Decision Trees Explained” [Online]

Available: https://towardsdatascience.com/decision-trees-explained-3ec41632ceb6

[7] Jason Brownlee “A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning”[Online]

Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/

[8] Joel Jorly “XGBOOST — IN A NUTSHELL” [Online]

Available: https://ai.plainenglish.io/xgboost-in-a-nutshell-211e170e8b48

[9] Arnon Puitrakul. “Feature Selection ใน Machine Learning”[Online]

Available:https://arnondora.in.th/feature-selection-machine-learning/

[10] Jason Brownlee. “How to Choose a Feature Selection Method For Machine Learning” [Online]

Available:https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

[11] Chew Jian Chieh “USING ANOVA TO FIND DIFFERENCES IN POPULATION MEANS”

[Online]

Available: https://www.isixsigma.com/tools-templates/analysis-of-variance-anova/using-anova-find-differences-population-means

การวิเคราะห์และจำแนกข้อความที่เป็นพิษ

พิพัฒน์พงศ์ พิทักษ์ก่อผล[1]

[1]*คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง กรุงเทพฯ*

[2]*คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบังฯ*

Emails: 62070140@kmitl.ac.th

บทคัดย่อ

รายงานฉบับนี้มีวัตถุประสงค์ในการกล่าวถึงการรายงานของวิชาโครงงาน คณะเทคโนโลยีสารสนเทศ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง ในเรื่องการวิเคราะห์และจำแนกข้อความที่เป็นพิษ

ผู้ที่พิมพ์ข้อความลงไปบนโซเชียลมีเดียอาจจะเจตนาหรือไม่เจตนาที่จะพิมพ์ข้อความที่เป็นพิษลงไป ข้อความเหล่านั้นอาจจะทำให้ผู้ที่อ่านเกิดปมหรือไปกระทบจิตใจ หรืออาจทำให้ผู้ที่ถูกกล่าวถึงในข้อความเสียหายได้

โดยมีการสกัดคุณลักษณะของข้อความ 2 แบบได้แก่ TF-IDF และการใช้ Universal Sentence Encoder Multilingua สร้างแบบจำลอง Traditional Machine Learning ได้แก่ Random Forest, Decision Tree, Extra Tree, K-Nearest Neighbors และสร้างแบบจำลอง Deep learning โดยใช้ Keras

คำสำคัญ – ข้อความ; เป็นพิษ; โซเชียลมีเดีย; แบบจำลอง

1. บทนำ

ในปัจจุบันสื่อโซเชียลมีเดียเป็นสื่อกลางที่สามารถสนทนาและแลกเปลี่ยนความคิดเห็นต่างๆผ่านอินเทอร์เน็ตแต่บางประโยคก็อาจจะเป็นพิษทำให้ผู้ที่อ่านหรือโดนกล่าวถึงเกิดปมในใจหรือเสียหายได้ ถ้าเป็นเด็กที่ยังไม่บรรลุนิติภาวะเข้ามาอ่านอาจจะทำให้จำข้อความที่เป็นพิษไปใช้ในชีวิตประจำวันได้ความคิดเห็นที่เป็นพิษจะทำให้ส่งผลต่อความสัมพันธ์และก่อให้เกิดการทำร้ายอีกฝ่ายไม่ว่าจะเป็นการทำร้ายร่างกายหรือจิตใจ มักทำให้ผู้ถูกกระทำเป็นคนที่ต้องรู้สึกหวาดกลัว จึงทำให้ผู้ใช้มากมายสามารถเข้ามาแสดงความคิดเห็นได้บางโพสต์หรือบางการแสดงความคิดเห็นอาจจะเป็นข้อความที่ไม่เหมาะสมหรือเป็นพิษอาจจะทำให้ผู้อ่านที่เป็นเด็กหรือยังขาดวุฒิภาวะอาจจะลอกเลียนแบบการใช้คำพูดหรือพฤติกรรมที่ไม่เหมาะสม หรือถ้าเป็นโพสต์ที่มีการขายของการมีข้อความที่ไม่เหมาะสมหรือเป็นพิษอาจจะทำให้สินค้าขาดความน่าเชื่อถือและส่งผลกระทบต่อสินค้าหรือร้านค้า ดั้งนั้นการวิเคราะห์และจำแนก

1.1 วัตถุประสงค์

เพื่อศึกษาการสร้างแบบจำลองที่เหมาะสมกับการวิเคราะห์และจำแนกข้อความที่เป็นพิษ และสร้างกระบวนการการคัดกรองข้อความที่เป็นพิษและเพื่อลดข้อความที่เป็นพิษบนสื่อออนไลน์

2. ทฤษฎีที่เกี่ยวข้อง

2.1 ความรุนแรงของข้อความที่เป็นพิษบนสื่อสังคม

2.1.1 ข้อความที่เป็นพิษ

ข้อความที่เป็นพิษ[1] หมายถึง ข้อความที่ถูกพูดไปในเชิงลบเนื่องจากคำว่าเป็นพิษมักจะนำมาเปรียบเทียบกับการกระทำหรือความสัมพันธ์ที่เป็นการทำร้ายอีกฝ่ายไม่ว่าจะเป็นด้านร่างกายหรือจิตใจ ความสัมพันธ์ที่เป็นพิษมักจะทำให้ผู้ถูกกระทำเป็นคนที่ต้องรู้สึกหวาดกลัว การที่โพสต์ข้อความที่

เป็นพิษลงบนโซเชียลมีเดียจะทำให้เกิดความเสียหายที่ตามมากับผู้ที่โพสต์หรือผู้ที่ถูกกล่าวถึงคือ อาจมีการทะเลาะเบาะแว้งทำร้ายร่างกายผู้อื่นหรือไม่ก็ทำร้ายตัวเองเนื่องจากไปพูดทำร้ายจิตใจ หรือบางครั้งก็สามารถทำให้ธุรกิจเกิดความเสียหายได้เนื่องจากการโพสต์รีวิวด้วยข้อความที่เป็นพิษทำให้

2.1.2 ความรุนแรง

ความรุนแรง [2] หมายถึง การเจตนาใช้กำลัง หรือใช้อำนาจในการข่มขู่ คุกคาม ซึ่งจะส่งผลกระทบต่อร่างกายหรือจิตใจของผู้ที่ถูกกระทำ อาจเป็นสาเหตุทำให้บาดเจ็บเสียชีวิตหรือสะเทือนใจ ซึ่งการเกิดภาวะความรุนแรงต่อตัวบุคคลสามารถเกิดได้จากการใช้ความรุนแรงทางกายโดยเจตนาทำให้ผู้อื่นได้รับความเดือดร้อนบาดเจ็บหรือถึงแก่ชีวิต หรือบังคับผู้อื่นให้ทำตามที่ตนเองต้องการ หรือการใช้ความรุนแรงที่เกิดขึ้นจากอาการทางจิต ซึ่งในปัจจุบันความรุนแรงได้มีผลอย่างแพร่หลายในบริบทสังคมนั้นล้วนสามารถมีความรุนแรงเกิดขึ้นได้ทั้งนั้นไม่ว่าจะเป็นสังคมในกลุ่มเพื่อน หรือแม้แต่ครอบครัวสามารถเกิดความรุนแรงได้หลากหลายรูปแบบ เช่น การทุบตี ด่าทอ หรือทิ้งทอด ในปัจจุบันสื่อสังคมเป็นสื่อกลางที่ให้บุคคลทั่วไปมีส่วนร่วมสร้างและแลกเปลี่ยนความคิดเห็นต่าง ๆ ผ่านอินเทอร์เน็ต เนื่องจากบนสื่อสังคมสามารถทำให้ผู้คนเข้าถึงข้อมูลต่างๆได้อย่างรวดเร็ว อาจทำให้บางเนื้อหาแพร่กระจายอย่างรวดเร็ว ทำให้เกิดการแสดงความคิดเห็นที่อย่างหลากหลายโดยยังไม่ได้ทราบข้อเท็จจริงของข้อมูลทำให้เกิดการขัดแย้งกันบนสื่อสังคม จึงเป็นสาเหตุทำให้มีความรุนแรงต่าง ๆ มากมาย และจะส่งผลต่อสุขภาพจิต และสภาพจิตใจ

2.1.3 ประเภทของความรุนแรง

1. ความรุนแรงทางร่างกาย (Physical Violence) หมายถึง การได้รับบาดเจ็บโดยผู้กระทำความรุนแรงต่อผู้ถูกกระทำ เช่น การทะเลาะวิวาท เตะ ต่อย และเป็นเหตุการณ์ที่ไม่ใช่อุบัติเหตุ

2. ความรุนแรงทางเพศ (Sexual Violence) หมายถึง การกระทำต่าง ๆ ที่มีเป้าหมายเพื่อใช้ผู้ถูกกระทำเป็นเครื่องตอบสนองความต้องการทางเพศของผู้กระทำ โดยอาจใช้กำลังบังคับหลอกลวง ข่มขู่ ชักชวนให้สิ่งตอบแทน เป็นต้น

3. ความรุนแรงทางจิตใจ (Psychological Violence) หมายถึง การถูกทำร้ายจิตใจและควบคุมบังคับจิตใจทำให้รู้สึกอับอาย รู้สึกด้อยค่า หรือลดคุณค่าความเป็นมนุษย์ของผู้อื่นลง

2.1.4 สื่อสังคม (Social Media)

สื่อสังคมหรือ Social media หมายถึงสื่อกลางอิเล็กโทรนิกส์ที่ให้บุคคลทั่วไปได้มีส่วนร่วมสร้างและแลกเปลี่ยนความคิดเห็นต่างๆ ผ่านอินเทอร์เน็ตได้ในปัจจุบันสื่อสังคมเป็นมากกว่าที่ที่ใช้ในการสนทนาแลกเปลี่ยนความคิดแต่ยังเป็นที่สามารถดำเนินธุรกิจได้อีกด้วยโซเชียลมีเดียจึงควรมีการคัดกรองข้อความที่เป็นพิษออกไปเพื่อประโยชน์ต่อผู้ที่เข้ามาพูดคุยหรือผู้ที่ใช้โซเชียลมีเดีย

2.2 ความหมายของป้ายกำกับข้อความ (Label)

2.2.1 ข้อความที่เป็นพิษ (Toxic)

ข้อความที่เป็นพิษ เป็นข้อความที่มีคำพูดหยาบคาย คำพูดเสียดสี หรือเป็นข้อความที่อาจจะทำให้ 2 ฝ่ายเกิดการแตกแยกกัน

2.2.2 ข้อความที่เป็นพิษมาก (Severe toxic)

ข้อความที่เป็นพิษ เป็นข้อความที่มีคำพูดหยาบคาย คำพูดเสียดสี หรือเป็นข้อความที่อาจจะทำให้ 2 ฝ่ายเกิดการแตกแยกกัน

2.2.3 ข้อความที่เป็นการหมิ่นประมาท (Insult)

ข้อความที่เป็นการหมิ่นประมาท เป็นข้อความที่มีการพูดหมิ่นประมาทจากการกระทำของผู้อื่น พูดถึงลักษณะร่างกาย เพศสภาพ ของผู้อื่น

2.2.4 ข้อความที่เป็นลามกอนาจาร (Obscene)

ข้อความที่เป็นลามกอนาจาร เป็นข้อความที่มีการพูดถึงสิ่งลามกอนาจาร การกระทำที่ลามกอนาจาร หรือเป็นข้อความที่นัดชักชวนผู้อื่นไปทำเรื่องลามกอนาจาร

2.2.5 ข้อความที่เป็นการข่มขู่ (Threat)

ข้อความที่เป็นการข่มขู่ เป็นข้อความที่มีการพูดข่มขู่ทำร้ายร่างกายหรือจิตใจ เช่นขู่ว่าจะไปทำร้ายร่างกาย หรือจะให้ผู้อื่นไปข่มขืน เป็นต้น

2.2.6ข้อความที่เป็นการเกลียดชัง (Identity hate)

ข้อความที่เป็นการเกลียดชัง เป็นข้อความที่แสดงความเกลียดชังในตัวตนของคนคนหนึ่งเช่น ศาสนา เพศ รูปร่างหน้าตา เป็นต้น

2.3 เครื่องมือ (Tools)

2.3.1 ภาษาไพทอน (Python)

Python[3] เป็นภาษาการเขียนโปรแกรมระดับสูงนิยมใช้ในการเขียนโปรแกรมอย่างมากในปัจจุบัน ถูกออกแบบมาให้ง่ายต่อการเรียนรู้และมีไวยากรณ์ที่ช่วยให้เขียนโค้ดสั้นกว่าภาษาอื่นๆ สนับสนุนการเขียนโปรแกรมเชิงวัตถุ (Object Oriented Programming)

2.3.2.การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP)

การประมวลผลภาษาธรรมชาติหรือ NLP [4] เป็นวิทยาการแขนงหนึ่งในหมวดหมู่ของเทคโนโลยีปัญญาประดิษฐ์ เป็นกระบวนการที่ทำให้คอมพิวเตอร์สามารถเข้าใจภาษาของมนุษย์ในรูปแบบข้อความหรือเสียง ทำให้สามารถคอมพิวเตอร์สามารถตีความภาษาของมนุษย์ ไปจนถึงสามารถวัดอารมณ์ความรู้สึกที่แฝงอยู่ในข้อความหรือสามารถกลั่นกรองใจความสำคัญของข้อความออกมาได้การทำงานขั้นพื้นฐานเบื้องต้นของ NLP มีดังนี้

1. Tokenization คือการแยกส่วนของข้อความออกเป็นหน่วยเล็กๆที่เรียกว่าโทเคน หรือในภาษาไทยชอบเรียกขั้นตอนนี้ว่าเป็นการตัดคำ โดยโทเคนสามารถเป็นได้ทั้ง คำ, อักษร และคำย่อย
2. Parsing คือการวิเคราะห์คำหลังจากวิธีการ Tokenization ว่าคำเหล่านั้นมีโครงสร้างประโยคและไวยากรณ์ถูกต้องหรือไม่
3. Lemmatization คือการเปลี่ยนคำให้อยู่ในรูปแบบพื้นฐานโดยอ้างอิงจากพจนานุกรมทั้งนี้ขึ้นอยู่กับว่าคำนั้นมีโครงสร้างประโยคเป็นอะไรด้วย
4. Part of Speech Tagging คือการแท็กคำในประโยคว่าประเภทของคำนั้นคืออะไร เช่น คำนาม คำกริยา คำบุพบท เป็นต้น

2.3.3. นิพจน์ทั่วไป (Regular Expression)

นิพจน์ทั่วไป [5] คือ การกำหนดรูปแบบอักขระ เพื่อใช้ในการค้นหาหรือตรวจสอบค่าในรูปแบบของสตริง (String) โดยนิพจน์ทั่วไปมีพื้นฐานมาจากทฤษฎีทางคณิตศาสตร์ที่ใช้ในการเปรียบเทียบข้อความที่ต้องการค้นหากับรูปแบบที่กำหนดว่ามีความสอดคล้องกัน

2.3.4.ฟังก์ชันการตัดคำของ PyThaiNLP (Tokenize)

PyThaiNLP Tokenize [6] คือ การแยกส่วนของข้อความออกเป็นหน่วยเล็กๆที่เรียกว่าโทเคน โดยฟังก์ชันการตัดคำของ PyThaiNLP สามารถรองรับการตัดคำ, ประโยค ซึ่งเอนจินหรืออัลกอริทึมที่นำมาใช้ในการทดลองมี 5ตัว ได้แก่

1. newmm เป็นการใช้วิธี Maximum Matching Algorithm (MMA) ซึ่งเป็นวิธีการตัดคำแบบ rule-based
2. longest เป็นการใช้วิธี Longest Matching Algorithm ในการตัดคำ วิธีนี้จะเลือกคำที่มีความยาวมากที่สุด
3. multi_cut เป็นการผสมผสานของการตัดคำแบบ rule-based และการตัดคำแบบ statistical โดยจะทำการตัดคำด้วย MMA แล้วใช้วิธีการตัดคำแบบ statistical เพื่อคัดเลือกคำที่เหมาะสม
4. deepcut เป็นการใช้วิธีตัดคำแบบ neural network โดยจะใช้โมเดล Deep Neural Network (DNN) เพื่อเรียนรู้รูปแบบการตัดคำในภาษาไทยจากข้อมูลที่มีอยู่แล้ว
5. attacut เป็นการใช้วิธีการตัดคำแบบ neural network โดยจะใช้โมเดล Deep Neural Network (DNN) ที่ใช้เทคนิค attention mechanism เพื่อจัดการกับปัญหาการตัดคำที่มีความยาวของคำแตกต่างกัน

2.3.5. Tweepy

Tweepy [7] เป็นไลบารีภาษา Python ที่สามารถใช้สำหรับการเก็บข้อมูลแบบเรียลไทม์ หรือเก็บข้อมูลย้อนหลังในอดีตของแอปพลิเคชันทวิตเตอร์ และสามารถกรองข้อมูลตามความสนใจ เช่น เข้าถึงข้อมูลโดยระบุวันที่ เวลาการใช้งาน, เข้าถึงข้อมูลตามความสนใจของผู้ใช้งาน (Hashtag), เข้าถึงข้อมูลตามคำสำคัญ หรือคำที่สนใจ (keyword)

2.3.6. RandomizedSearchCV

RandomizedSearchCV เป็นเทคนิคในการทำ hyperparameter tuning ใน traditional machine learning โดยจะทำการสุ่มค่าพารามิเตอร์ที่กำหนดให้เป็นไปตามการแจกแจงที่ระบุไว้

2.3.7. Tensorflow

TensorFlow เป็นซอฟต์แวร์โอเพนซอร์สสำหรับการเรียนรู้เชิงลึก (deep learning) ที่ถูกพัฒนาโดย Google Brain Team โดย TensorFlow จะใช้กราฟในการแทนและคำนวณรูปแบบและน้ำหนักของโมเดลการเรียนรู้เชิงลึก

2.3.8. Keras

Keras เป็นชุดเครื่องมือสำหรับการสร้างและฝึกโมเดลปัญญาประดิษฐ์ซึ่งช่วยให้ผู้ใช้สามารถสร้างโมเดลได้อย่างรวดเร็วและง่ายมากยิ่งขึ้น โดย Keras ได้รวมแนวคิดจากโมดูล deep learning libraries อย่าง Theano และ TensorFlow เข้าด้วยกันเพื่อทำให้การสร้างและฝึกโมเดลได้สะดวกและมีประสิทธิภาพมากขึ้น

2.3.9. Streamlit

Streamlit เป็นเครื่องมือสำหรับพัฒนาแอปพลิเคชันเว็บที่ใช้ภาษา Python เพื่อแสดงผลและสร้างตัวควบคุม โดย Streamlit สามารถสร้างเว็บแอปพลิเคชันที่มี Interface ที่สวยงามได้อย่างรวดเร็วและง่าย ซึ่งจะช่วยให้นักพัฒนาสามารถเน้นการพัฒนาฟังก์ชันโดยไม่ต้องคำนึงถึงการตกแต่งเว็บได้มากขึ้น Streamlit มีฟังก์ชันการแสดงผลข้อมูลแบบ interactiveได้อย่างรวดเร็วและง่ายดายซึ่งเหมาะสำหรับการสร้างเว็บแอปพลิเคชันที่ต้องการการแสดงผลลัพธ์แบบ Real time

2.4 การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

2.4.1. TF-IDF

TF-IDF เป็นวิธีหาคำสำคัญในเอกสารโดยดูจากเนื้อหาของเอกสารทั้งหมด โดยมีสองส่วนในการหาคำสำคัญในเอกสาร

2.4.2. Word2Vec

เป็นการเรียนรู้ความหมายของคำและสร้างเวกเตอร์ที่บ่งบอกความหมายของคำโดยใช้โมเดล neural network ที่เรียนรู้จากข้อความ โดยคำแต่ละคำจะถูกแทนด้วยเวกเตอร์ที่มีความยาวเท่ากัน ซึ่งคำที่ถูกเปลี่ยนเป็นเวกเตอร์จะสามารถใช้งานในการหาความสัมพันธ์ระหว่างคำ หรือการแทนคำเป็นเวกเตอร์ เพื่อนำไปใช้ในงานต่อไป เช่น การจัดกลุ่มคำที่มีความหมายใกล้เคียงกัน หรือการหาคำที่เกี่ยวข้องกันในประโยค

2.4.3. Universal Sentence Encoder Multilingual (USE-M)

Universal Sentence Encoder Multilingual (USE-M) เป็น Sentence Encoder ที่ถูกพัฒนาโดยใช้โมเดล Transformer ของ Google

2.5 การเรียนรู้ด้วย Zero-Shot Learning และ One-Shot Learning

2.5.1 Zero-Shot Learning

Zero-Shot Learning[8] คือ การใช้โมเดลภาษาที่ถูกฝึกฝนมาก่อนแล้วในการทำนายหรือจัดหมวดหมู่ของคลาสโดยที่ไม่เห็นหรือไม่ได้ถูกฝึกในประเภทนั้นระหว่างการฝึก Zero-Shot Learning ทำให้สามารถใช้โมเดลทั่วไปในการจำแนกข้อมูลโดยไม่ต้องฝึกสอนประเภทข้อมูลที่เราต้องการจำแนกใหม่

2.5.2 One-Shot Learning

One-Shot Learning[9] คือ การจัดหมวดหมู่ของคลาสโดยใช้ข้อมูลตัวอย่างเพียงหนึ่งตัวเท่านั้นโดยโมเดลจะถูกฝึกอบรมด้วยชุดข้อมูลที่มีเพียงหนึ่งตัวอย่างสำหรับแต่ละคลาส One-Shot Learning สามารถแก้ปัญหาการจัดหมวดหมู่ของคลาสที่มีข้อมูลมากแก้ปัญหา Imbalance ของข้อมูลได้

2.6 การสร้างแบบจำลอง (Modeling)

2.6.1 Decision Tree

Decision Tree[10] คือเทคนิคที่ให้ผลลัพธ์ในลักษณะเป็นโครงสร้างของต้นไม้ภายในต้นไม้จะประกอบไปด้วยโหนด (node) ซึ่งแต่ละโหนดจะมีเงื่อนไขของคุณลักษณะเป็นตัวทดสอบ กิ่งของต้นไม้ (branch) แสดงถึงค่าที่เป็นไปได้ของคุณลักษณะที่ถูกเลือกทดสอบ และใบ (leaf) เป็นสิ่งที่อยู่ล่างสุดของต้นไม้แสดงถึงกลุ่มของข้อมูล (class) ก็คือผลลัพธ์ที่ได้จากการพยากรณ์ซึ่งข้อดีของวิธีการนี้คือให้ผลการพยากรณ์ที่แม่นยำและเกิดปัญหา overfitting น้อย

2.6.2 Randomforest

Randomforest[11] เป็นโมเดลที่ถูกพัฒนามาด้วยวิธีการ Decision

Tree ขึ้นมาหลายๆ โมเดลโดยวิธีการสุ่มตัวแปรแล้วนำผลลัพธ์ที่ได้แต่ละโมเดลมารวมกันโดยผลลัพธ์สุดท้ายจะมาจากการโหวตจากผลลัพธ์ที่ได้ในแต่ละโมเดลวิธีการนี้เรียกว่า Bagging หรือ Boostraping

2.6.3 Extratree

Extratree[12] เป็นโมเดลที่มีการทำงานคล้ายคลึงกับ Randomforest แต่การสร้างต้นไม้จะมีความแตกต่างออกไปและจะมีข้อแตกต่างอยู่ 2 ประการคือ
1.จะไม่มีการใช้ boostrap มาแทนที่กลุ่มตัวอย่างทำให้สามารถใช้ชุดข้อมูลที่มีอยู่ทั้งหมดและสามารถลดการ bias ได้
2.ในการแยกโหนดจะใช้วิธีแบบ random split แทนที่วิธี best split ซึ่งสามารถลดการแปรปรวนลงได้เนื่องจากฟีเจอร์บางอย่างที่มีการทำงานซับซ้อนจะไม่ถูกเลือกมา

2.6.4 K-Nearest Neighbor (KNN)

KNN[13] เป็นโมเดลที่สามารถเรียนรู้ได้ทั้ง 2 แบบคือ การถดถอย (regression) และ การจำแนกประเภท (classification) KNN จะทำนายคลาสที่ถูกต้องโดยคำนวณระยะห่างระหว่างข้อมูล

2.7 การประเมินประสิทธิภาพ (Evaluation)

1. Accuracy คือตัววัดประสิทธิของการทำงานของอัลกอริทึมโดยจะดูว่าค่าที่ทำนายถูกต้อง และมีความแม่นยำเท่าไหร่ในการทำนาย

2. Recall คือการวัดค่าความไว เป็นสัดส่วนของจำนวนการจำแนกที่เป็น True positive ต่อข้อมูลที่เกิดขึ้นจริง เพื่อดูว่าเมื่อเทียบกับการจำแนกที่เราสนใจแล้วมีความแม่นยำมากน้อยแค่ไหนเมื่อเทียบกับข้อมูลจริง

3. F1 - Score เป็นการผสมระหว่าง Precision และ Recall เข้าด้วยกันเพื่อใช้วัดความสามารถของโมเดล

4. Precision เป็นสัดส่วนของจำนวนการจำแนกที่เป็น True positive ต่อจำนวนการจำแนกที่เราสนใจหรือเป็นความแม่นยำที่สนใจแค่ในส่วนของการทำนาย

3. วิธีการดำเนินงานวิจัย

3.1 ขอสิทธิ์การเข้าถึงข้อมูล Twitter

ในขั้นตอนนี้จะทำเพื่อขอสิทธิ์การเข้าถึง Twitter API หรือการเข้าถึงข้อมูลของ Twitter ซึ่งถ้าไม่ขอสิทธิ์การเข้าถึงนี้ก็ไม่สามารถดึงข้อมูลของ Twitter ออกมาใช้ได้

3.1.1 สมัครบัญชี Twitter จากนั้นไปยังเว็บไซต์ Twitter Developers เพื่อสมัครการเข้าถึง Developer Portal

3.1.2 จะมีคำถามให้เขียนคำตอบอธิบายเป็นภาษาอังกฤษ

1. คำถามแรกจะถามว่า คุณมีแผนที่จะใช้ Twitter API หรือ Twitter Data อย่างไร
2. คำถามที่สองจะถามก่อนว่า คุณมีแผนที่จะวิเคราะห์ข้อมูล Twitter หรือไม่โดยสามารถเลือกได้ว่าจะวิเคราะห์หรือไม่ ถ้าตอบใช่ก็ให้อธิบายว่าจะวิเคราะห์ยังไงและเป็นการวิเคราะห์โพสต์ (Tweet) หรือวิเคราะห์ผู้ใช้ Twitter
3. คำถามที่สามจะถามก่อนว่า ตัวแอพพลิเคชันของคุณ จะใช้ฟังก์ชัน โพสต์ (Tweet), แชร์โพสต์ (Retweet), ชอบ (Like), ติดตาม (Follow) หรือข้อความโดยตรง (Direct Message) หรือไม่ถ้าตอบว่าใช่ก็ต้องอธิบายแผนที่จะใช้ฟีเจอร์เหล่านี้
4. คำถามที่สี่จะถามก่อนว่า คุณวางแผนที่จะแสดงทวีตหรือข้อมูลรวมเกี่ยวกับเนื้อหา Twitter ภายนอก Twitter หรือไม่ ถ้าตอบว่าใช่ให้อธิบายว่าจะนำไปแสดงนอก Twitter อย่างไรและที่ใด
5. คำถามที่ห้าจะถามก่อนว่า ผลิตภัณฑ์ บริการ หรือการวิเคราะห์ของคุณจะทำให้เนื้อหา Twitter หรือข้อมูลที่ได้มาพร้อมใช้งานสำหรับหน่วยงานของรัฐหรือไม่

3.1.3 หลังจากตอบคำถามครบแล้วให้รออีเมลตอบรับจากทาง Twitter ตอบคำถามเพิ่มเติมกลับไปยังอีเมลที่ส่งมา

3.2 การรวบรวมและทำความเข้าใจข้อมูล

3.2.1 ชุดข้อมูลถูกรวบรวมข้อความที่มีการโพสบนแอปพลิเคชัน Twitter โดยใช้ไลบารีของ python ที่ชื่อว่า Tweepy ในการรวบรวบข้อความโดยข้อความที่ถูกเลือกมาจากการกำหนด

keyword ภาษาไทยที่ได้เตรียมไว้ หากในข้อความมีคำที่ตรงกับ keyword ที่กำหนดไว้ข้อความนั้นก็จะถูกเก็บไว้ในชุดข้อมูล

3.2.2 หลังจากที่ได้ข้อมูลมาแล้วก็มาดูว่าใน 1 ข้อความมีส่วนประกอบอะไรบ้างโดยใน 1 โพสต์ประกอบด้วย ชื่อผู้ใช้ที่ถูกกล่าวถึง, ข้อความ, แฮชแท็ก (เรื่องหัวข้อที่เกี่ยวข้องกับโพสต์), อีโมจิ, ตัวเลข, ภาษาอังกฤษ, ลิงก์ต่างๆเช่น เว็บไซต์ รูปภาพ GIF จะเห็นได้ว่าใน 1 Tweet หรือโพสมีส่วนที่ไม่จำเป็นในการนำมาวิเคราะห์ประเภทของข้อความจึงต้องในข้อความเหล่านี้ไปทำความสะอาดด้วยในปัจจุบันงานวิจัยนี้มีชุดข้อมูลที่ได้ติดแท็กป้ายกำกับข้อความแล้วทั้งหมด 10176 แถว 8 คอลัมน์โดยชุดข้อมูลจะมีลักษณะดังตาราง

3.3 การติดแท็กข้อมูล (Label)

ในการติดแท็กข้อมูลภาษาไทยที่รวบรวมมาจากTwitter ได้ทำการติดแท็กข้อมูลเองโดยใช้มนุษย์อ่านและติดแท็ก โดยชื่อของป้ายกำกับข้อความได้แรงบันดาลมาจาก Dataset ใช้ในการแข่งขัน Toxic Comment Classification Challenge บนเว็บไซต์ Kaggle โดยจะลอง plot word cloud ของแต่ละคลาสออกมาดูว่าในแต่ละคลาสส่วนใหญ่จะเจอคำศัพท์ประมาณไหนโดยคลาสทั้งหมดจะมีทั้งหมด 6 class

3.4 การเตรียมข้อมูล (Data Preparation)

3.4.1 การทำความสะอาดข้อมูล (Data Cleaning)
ขั้นตอนการทำความสะอาดข้อมูลเป็นกระบวนการตรวจสอบและแก้ไขข้อมูลให้อยู่ในรูปแบบที่พร้อมใช้งานรวมไปถึงเป็นการจัดการข้อมูลที่ไม่จำเป็นออกไปจากข้อมูลเพราะหากมีข้อมูลที่ไม่จำเป็นจะส่งผลต่อการวิเคราะห์หรือประมวลผลได้ ขั้นตอนการทำความสะอาดข้อมูลเป็นขั้นตอนที่สำคัญจึงต้องเตรียมข้อมูลให้มีความสมบูรณ์และมีคุณภาพมากที่สุด โดยจะใช้เครื่องมือ Regular Expression ในการกำหนดรูปแบบอักขระเพื่อกำจัดส่วนที่ไม่จำเป็นของข้อความทิ้งไป

3.4.2 การแยกส่วนของข้อความหรือการตัดคำ (Word Tokenization)

โดยจะใช้ PythaiNLP ซึ่งเป็นไลบารีของไพทอนใช้สำหรับประมวลผลข้อความภาษาไทย โดยการตัดคำจะใช้เอนจิน (engine) ทั้งหมด 5 ตัวในการทำการทดลองได้แก่
1. Newmm 2. Attacut 3. Longest 4. Multicut 5. Deepcut

3.4.3 การแบ่งชุดข้อมูล (Train test split)

เป็นขั้นตอนการแบ่งสัดส่วนข้อมูลเบื้องต้นจะแบ่งข้อมูลโดยมีสัดส่วน (70:30) Training set มีจำนวนประมาณ 70% ของข้อมูลทั้งหมดจะนำไปใช้สำหรับให้โมเดลเรียนรู้ข้อมูล ส่วน Test set มีจำนวนประมาณ 30% ของข้อมูลใช้สำหรับทดสอบหลังจากเรียนรู้จากชุดข้อมูล Training set ว่าโมเดลจะทำงานได้ดีแค่ไหนเมื่อเจอข้อมูลที่ไม่เคยเห็นมาก่อน

3.4.4 การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

เป็นกระบวนการที่เตรียมข้อมูลให้พร้อมสำหรับนำไปวิเคราะห์ข้อมูลหรือนำไปใช้ในการสร้างแบบจำลองโดยขั้นตอนนี้จะทำการแปลงรูปแบบของข้อมูลจากข้อความให้กลายเป็นเวกเตอร์ ซึ่งวิธีที่นำมาใช้คือ TF-IDF, Word2vec และ Sentence Embedding โดยใช้ Universal sentence encoderโดยการนำ Sentence Embedding ไปใช้งานไม่จำเป็นต้องทำขั้นตอน Word Tokenizationจึงเกิดเป็นการทดลองให้เปรียบเทียบระหว่างการใช้ Sentence Embedding คือใช้งานโดยทำและไม่ทำWord Tokenization

3.5 การสร้างแบบจำลอง (Modeling)

ซึ่งโครงงานนี้จะทดลองการสกัดคุณลักษณะของข้อมูลโดยใช้ TF-IDF และ Universal Sentence Encoder Multilingua และแบบจำลองที่นำมาใช้ในทำนายได้แก่

1.Random forest

2.Decision tree

3.Extra tree

4.K-Nearest Neighbor

3.6 การทำ Fine-tuned

เป็นการนำข้อมูลของเราไปทำ Sentence Encoder โดยใช้ Pre-trained model ที่ชื่อว่า Universal Sentence Encoder – Multilingua แล้วใช้ชุดข้อมูลใหม่หรือข้อมูลดั้งเดิมของเราที่ต้องการจะใช้ในแบบจำลองเพื่อปรับแต่งและเพิ่มประสิทธิภาพให้เหมาะสมกับการใช้งานโดยจะสร้างแบบจำลองโดยใช้ Keras

4. ผลลัพธ์และการอภิปรายผล

จากการสำรวจชุดข้อมูลที่นำมาจาก Kaggle ชุดข้อมูล train มีขนาด 159571 แถว ชุดข้อมูลที่นำไปฝึกฝนถูกแบ่งมาเป็น 70% จากชุดข้อมูลทั้งหมดโดยใช้ library จาก scikit learn ที่ชื่อว่า train-test split ในงานวิจัยนี้ชุดข้อมูลที่นำมาฝึกสอนมีทั้งหมด 10167 แถว ซึ่งงานวิจัยครั้งนี้มีการเปรียบเทียบประสิทธิภาพในแต่ละแบบจำลองได้แก่ Random forest, Decision tree, Extra tree, K-Nearest Neighbor โดยจะใช้ตัวตัดคำที่เหมาะสมที่สุดคือ attacut การใช้ Sentence Encoder ที่ชื่อว่า Universal Sentence Encoder Multilingua(USE-M)

5. สรุปผลการดำเนินงาน

งานวิจัยในครั้งนี้ได้ทำการสร้างแบบจำลองที่สามารถวิเคราะห์ข้อความที่เป็นพิษแบบ Multi-label โดยแบบจำลองที่ใช้ในการวิจัยนี้ได้แก่ Random forest, Decision tree, Extra tree, K-Nearest Neighbor, Deep Learning การวิจัยในครั้งนี้เริ่มจากการ รวบรวมข้อมูล ทำความเข้าใจข้อมูล ติดป้ายกำกับข้อความ การเตรียมข้อมูลให้พร้อมใช้งาน การสร้างแบบจำลอง และการปรับแต่งพารามิเตอร์สำหรับแบบจำลองที่ให้ประสิทธิภาพดีที่สุด และการประเมินประสิทธิภาพโดยวัดจากค่า Accuracy, F1-score, Precision, Recall

5.1 ปัญหาที่พบ

ข้อมูลมีการ Imbalance ข้อความที่เป็น Positive และ Negative มีขนาดที่แตกต่างกัน ในบาง Class มีข้อความที่เป็น Positive น้อยเกินไปอาจจะทำให้แบบจำลองไม่สามารถจำแนกข้อความ Class นั้นได้หลังจากทำความสะอาดข้อมูลแล้วมีคำศัพท์บางคำเปลี่ยนบริบท

5.2 แนวทางการพัฒนาต่อ

1. ในการติดแท็ก Label ข้อความควรมีมนุษย์มาแท็กข้อความเดิมซ้ำเพื่อให้แน่ใจว่าข้อความที่ทำการติดแท็กมานั้นติดแท็กได้ถูกต้องและควรมี guideline ในการติดแท็กข้อความ
2. ในการแบ่งคลาสของข้อความควรมีหลักการหรืออ้างอิงจาก Literature เพื่อทำให้มีน่าเชื่อถือเพิ่มมากขึ้น
2. ควรหาข้อความเพิ่มเพื่อให้ข้อความบางประเภทที่มีจำนวนน้อยสามารถเรียนรู้และตรวจจับข้อความประเภทนั้นๆได้มีประสิทธิภาพมากขึ้น
3. ในการทำความสะอาดข้อมูลควรจะกำหนด Stop words เองเพื่อไม่ให้บริบทของคำเปลี่ยนไป
4. ควรลองเลือกใช้ Pre-trained ตัวอื่นนอกเหนือจาก Universal Sentence Encoder Multilingua

เอกสารอ้างอิง

[1]incwaran. (2 เมษายน 2564). Toxic Words คำพูดง่าย ๆ แต่ความหมายสุดทิ่มแทง
https://www.mangozero.com/toxic-words/

[2]สุภาพร กมลฉ่ำ. (19 กันยายน 2564). การให้ความหมายความรุนแรงในโลกออนไลน์: กรณีศึกษานิสิตจุฬาลงกรณ์มหาวิทยาลัย

[3]meck. (19 ธันวาคม 2563). ทำความรู้จักกับ Python
https://www.borntodev.com/c/xakhrcirthnchotichwalwithy/

[4]SAS. การประมวลผลภาษาธรรมชาติ (Natural language processing)
https://www.sas.com/th_th/insights/analytics/what-is-natural-language-processing-nlp.html

[5]ณัฐธิดา หมวดเพ็ชร. (30 มิถุนายน 2558). Regular Expression
https://swiftlet.co.th/regular-expression/

[6] PyThaiNLP. Tokenize
https://pythainlp.github.io/dev-docs/api/tokenize.html

[7]Jason Rigden. (23 มกราคม 2561). Tweepy: a Python Library for the Twitter API

[8] Huggine Face. Zero-Shot Classification
https://huggingface.co/tasks/zero-shot-classification

[9] Nat Khaokum. One-Shot Learning ตอนที่ 1
https://medium.com/super-ai-engineer/one-shot-learning-%E0%B8%95%E0%B8%AD%E0%B8%99%E0%B8%97%E0%B8%B5%E0%B9%88-1-98e2bbffe427
https://medium.com/@jasonrigden/tweept-a-python-library-for-the-twitter-api-9d0537dcebd4

[10]manav. (25 ตุลาคม 2563). Decision Trees: Explained in Simple Steps
https://medium.com/analytics-vidhya/decision-trees-explained-in-simple-steps-39ee1a6b00a2

[11]Harshdeep Singh. (4 มีนาคม 2562). Understanding Random Forests
https://medium.com/@harshdeepsingh_35448/understanding-random-forests-aa0ccecdbbbb

[12]Karun Thankachan. (7 สิงหาคม). What? When? How?: ExtraTrees Classifier
https://towardsdatascience.com/what-when-how-extratrees-classifier-c939f905851c

[13]Antony Christopher. (2 กุมภาพันธ์ 2564). K-Nearest Neighbor
https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4