Fake Reviews Classification Using Locality-Sensitive Hasing (LSH)

Pearploy Thipprasert
(6310422030)
Department of Applied Statistics, Data Science, NIDA
pearploy.thi@stu.nida.ac.th

Rakchanok Thongkumpan
(6310422039)
Department of Applied Statistics, Data Science, NIDA
rakchanok.tho@stu.nida.ac.th

Weerasak Karoon
(6310422046)
Department of Applied Statistics, Data Science, NIDA
weerasak.karo@stu.nida.ac.th

Ruetinan Hanprasopwat
(6310422033)
Department of Applied Statistics, Data Science, NIDA
ruetinan.han@stu.nida.ac.th

Athittaya Sriaram
(6310422045)
Department of Applied Statistics, Data Science, NIDA
athittaya.sri@stu.nida.ac.th

Nachasa Khongchu
(6310422050)
Department of Applied Statistics, Data Science, NIDA
supharak.kho@stu.nida.ac.th

***Abstract*—*ในโลกที่เต็มไปด้วยข้อมูลการจะหาข้อมูลที่ถูกต้องและไม่ได้รับอิทธิพลจากการบิดเบือนเลยเป็นเรื่องที่ทำได้ยาก จุดประสงค์ของการศึกษานี้คือการแยกรีวิวบนแพล็ตฟอร์มว่าเป็นรีวิวจริงหรือรีวิวเท็จ เพื่อประโยชน์ของผู้ใช้งานแพล็ตฟอร์ม ร้านอาหารที่ถูกรีวิว และผู้ให้บริการแพล็ตฟอร์มนั้น โดยชุดข้อมูลที่ใช้ในการศึกษานี้คือรีวิวจากร้านอาหารยอดนิยมจำนวน 76 ร้านแรก ในเว็บไซต์ Wongnai จำนวน 12,859 รีวิว ซึ่งการศึกษานี้เป็นการศึกษาเพื่อจัดประเภทรีวิวนั้นว่าเป็นจริงหรือเท็จ โดยอ้างอิงจากความคล้ายกับรีวิวต้นแบบที่นำมาเป็นเกณฑ์ตัดสิน ซึ่งมีขั้นตอนในการดำเนินงานคือ ทำ Word Tokenization, Min-Hash และ LSH เพื่อหารีวิวที่คล้ายคลึงที่สุดจำนวน 10 รีวิวแรก และคำนวณหา Weighted Average Jaccard Similarity ออกมาเพื่อจำแนกรีวิวเหล่านั้น*

Keywords—fake review, LSH, MinHash, classification, word tokenization, wongnai

I. Introduction

ในยุคปัจจุบันที่มีการเติบโตทางเทคโนโลยีสูง ผู้บริโภคหันใช้งานผ่านสื่อสังคมออนไลน์มากขึ้น แม้ปัจจุบัน "ธุรกิจร้านอาหาร" ในประเทศไทยมีมูลค่าตลาดรวม 385,000 ล้านบาท แต่ก็มีแนวโน้มที่ยังคงเติบโตอย่างต่อเนื่อง ต่างพยายามปรับเปลี่ยนแนวคิดให้การตอบสนองวิถีคนเมืองที่นิยมความรวดเร็วและสะดวกสบาย ซึ่งจะเห็นได้ว่าในปัจจุบันมีผู้ให้บริการร้านอาหารบนแพล็ตฟอร์มออนไลน์มากขึ้น อาทิเช่น Wongnai, Grab, Foodpanda, Gojek เป็นต้น [1] เมื่อผู้บริโภคได้เริ่มเปลี่ยนพฤติกรรมมาใช้ช่องทางการสื่อสารออนไลน์เป็นหลัก ในการค้นหาร้านอาหารเพื่ออ่านรีวิวประกอบการตัดสินใจในการเลือกรับประทานอาหารมากขึ้น ดังนั้น “รีวิวและเรตติ้ง” ของแต่ละร้านจึงถือเป็นเครื่องมืออย่างหนึ่งที่ช่วยสร้างความเชื่อมั่นให้แก่ลูกค้าบนแพล็ตฟอร์มออนไลน์ได้เป็นอย่างมาก ซึ่งพบว่าเหตุผลอันดับหนึ่งที่ผู้คนตัดสินใจเข้าชมเว็บไซต์ขายสินค้าออนไลน์เพราะต้องการเข้าไปอ่านรีวิวจากผู้ที่เคยซื้อสินค้า และแนวโน้มที่จะเชื่อความเห็นของผู้บริโภคด้วยกันเองมากกว่าการโฆษณา [2] และ 72% ของผู้บริโภคมีแนวโน้มจะซื้อสินค้าหากรีวิวสินค้าเป็นไปในเชิงบวก และ 92% ของลูกค้าจะเลือกใช้บริการร้านค้าถ้าที่มีเรตติ้งตั้งแต่ 4 ดาวขึ้นไป [3] ดังนั้นจึงเห็นได้ว่ารีวิวของผู้บริโภคและเรตติ้งในโลกออนไลน์นอกจากจะเป็นส่วนช่วยในการสร้างความเชื่อมั่นให้แก่ผู้บริโภคแล้ว ยังเป็นตัวแปรสำคัญในกระบวนการตัดสินใจของผู้บริโภคอีกด้วย

ในปี 2560 เว็บไซต์ Wongnai มีคนเข้ามาค้นหาข้อมูลมากกว่า 7,500,000 ครั้ง โดยคำค้นหายอดนิยมจะเป็นชื่อเมนูอาหาร ร้านอาหาร สถานที่ เช่น ชื่อห้าง หรือย่านที่มีร้านอาหารตั้งอยู่มากมาย และหากร้านไหนเป็นที่พูดถึงและมีรีวิวที่ดีบนแพล็ตฟอร์ม ก็ยิ่งมีแนวโน้มว่าจะช่วยให้ร้านอาหารขายดีและประสบความสำเร็จมากยิ่งขึ้น ตัวเลขบ่งชี้ว่าการรีวิวมีผลต่อการตัดสินใจซื้อสินค้าทั่วไปของผู้บริโภคถึง 3 ใน 4 ในส่วนของอาหารและเครื่องดื่มก็ยังเป็นสิ่งที่ผู้บริโภคซื้อตามหลังจากเห็นรีวิวถึง 46.5%

[4] จะเห็นได้ว่าการรีวิวเป็นปัจจัยหนึ่งที่มีอิทธิพลอย่างมากต่อการตัดสินใจเลือกร้านอาหารของผู้บริโภค จึงทำให้ร้านอาหารบางร้านได้นำการเขียนรีวิวที่เป็นเท็จมาใช้ เพื่อเพิ่มยอดขายให้กับทางร้านค้า เช่นมีการจ้างบุคคลอื่นมารีวิวร้านของตนเอง เพื่อที่จะทำให้ร้านอาหารของตนเป็นที่แพร่หลายมากขึ้น ซึ่งการเขียนรีวิวเท็จนั้นมีจุดประสงค์เพื่อสร้างความเข้าใจผิดให้แก่ผู้บริโภค ไม่ว่าจะเป็นการเขียนรีวิวในเชิงบวกที่ลำเอียง เพื่อเพิ่มยอดขายให้กับทางร้านของตัวเอง หรือการเขียนรีวิวเชิงลบที่ไม่เป็นธรรม เพื่อโจมตีคู่แข่งทางธุรกิจของตน [5]

ด้วยความตระหนักถึงปัญหา และผลกระทบของรีวิวที่เป็นเท็จที่มีต่อผู้บริโภคและผู้ขาย อีกทั้งการวิจัยที่มีอยู่ในปัจจุบันมักเป็นการตรวจจับรีวิวเท็จที่เป็นภาษาอังกฤษ มีการศึกษาในภาษาไทยน้อย กลุ่มผู้วิจัยจึงมีวัตถุประสงค์ที่จะศึกษาการจำแนกรีวิวเท็จโดยใช้วิธี Locality-Sensitive Hashing (LSH) เพื่อสร้างความเชื่อมั่นให้กับทั้งผู้บริโภค และธุรกิจ โดยผู้บริโภคสามารถมั่นใจได้ว่ารีวิวที่อ่านนั้นเป็นความจริง และร้านอาหารนั้นตรงกับความต้องการของตัวเองหรือไม่ ทางด้านธุรกิจก็สามารถรับรู้ความต้องการ และความเห็นจากผู้บริโภคตัวจริง เพื่อนำไปปรับปรุงพัฒนาสินค้า และบริการอย่างต่อเนื่องให้ตอบโจทย์ผู้บริโภคมากยิ่งขึ้น อีกทั้งยังเพิ่มความน่าเชื่อถือให้กับร้านค้าและตัว platform มากขึ้น สามารถนำมาพัฒนาต่อยอดทางธุรกิจได้ในอนาคตต่อไป

II. DATASET

A. *Data Collection*

เนื่องจากการศึกษานี้เป็นการจำแนกรีวิวที่เป็นเท็จ ผู้วิจัยจึงเลือกใช้ชุดข้อมูลจากเว็บไซต์วงใน (Wongnai) ซึ่งเป็นเว็บไซต์ที่นำเสนอเนื้อหาและข้อมูลรีวิวจากผู้ใช้งานจริงแบบครบวงจร ทั้งร้านอาหาร สูตรอาหาร ความสวยความงาม และการท่องเที่ยว [6] โดยผู้ใช้งานสามารถค้นหาร้านอาหาร ข้อมูล รูป และคำวิจารณ์จากสมาชิกคนอื่นได้

การศึกษานี้จึงเลือกใช้รีวิวจากร้านอาหาร โดยทำการดึงข้อมูลมาจากเว็บไซต์วงใน (ตัวอย่างข้อมูลหน้าเว็บไซต์ดังรูปที่ 1) ซึ่งเลือกจากร้านอาหารยอดนิยมจำนวน 76 ร้านแรก ในวันที่ 19 เมษายน 2564 โดยใช้วิธีการดึงข้อมูลด้วยการทำ Web Scraping ด้วยโปรแกรมภาษา Python ที่ใช้ Selenium ในการควบคุม Chrome Webdriver ให้ทำการคลิกลิงค์ต่างๆแบบอัตโนมัติ จากนั้นจึง scrape ข้อมูล ที่ต้องการมาเก็บเป็น dataset ในรูปแบบ dataframe ของ pandas ซึ่งได้ข้อมูลรีวิวมาทั้งหมด 12,859 รีวิว

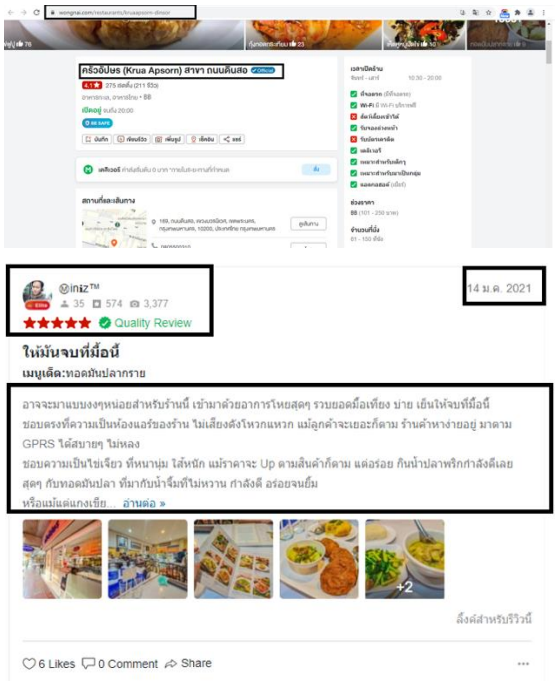

Fig. 1. ตัวอย่างข้อมูลที่ทำการ Scrape มาจากเว็บไซต์วงใน

B. *Data Preparation and Labeling*

ชุดข้อมูลรีวิวที่ดึงมาจากเว็บไซต์วงในสามารถแบ่งได้เป็น 2 แบบ ได้แก่

1. ข้อมูลเกี่ยวกับตัวรีวิว ประกอบไปด้วย restaurant_name, comment, score, date และ length ซึ่งบอกให้รู้ว่าเป็นรีวิวจากร้านค้าใด รีวิวว่าอย่างไร ให้เรตติ้งเท่าไร และมีความยาวเท่าไร
2. ข้อมูลเกี่ยวกับผู้รีวิว ประกอบไปด้วย reviewer, follower, review, photo ซึ่งบอกให้รู้ว่าใครเป็นผู้เขียนรีวิว และผู้เขียนรีวิวนี้มีคนติดตามอยู่เท่าไร เคยเขียนรีวิวบนเว็บไซต์นอกจากรีวิวนี้กี่ครั้ง และมีการอัพโหลดรูปลงมาในเว็บไซต์กี่รูป

จากนั้นนำข้อมูลที่ได้มาทำ Data Cleansing คือลบข้อมูลที่ scrape มาไม่ครบถ้วน รวมทั้งลบรีวิวที่เป็นภาษาต่างประเทศ และภาษาถิ่นออก เนื่องจากการศึกษาครั้งนี้เรามุ่งเน้นไปที่ภาษาไทยเท่านั้น และเนื่องจากรีวิวที่ scrape มาไม่สามารถระบุได้ว่า รีวิวนั้นเป็นรีวิวจริง หรือรีวิวเท็จ ผู้วิจัยจึงต้องทำ Data Labeling ด้วยตนเอง

จากการศึกษาลักษณะของรีวิวเท็จ [7] มักไม่ค่อยมีข้อมูลเกี่ยวกับผู้รีวิว เป็นบัญชีที่ไม่ค่อยมีการเคลื่อนไหว หรือมีจำนวนรีวิวที่เขียนไว้น้อย มีข้อความสั้นจนเกินไป ซึ่งในข้อความจะระบุชื่อร้าน ชื่อเมนูเยอะจนเกินพอดี ใช้คำพูดที่เกินจริงอย่างเห็นได้ชัด

รวมทั้งอาจมีข้อมูลที่ไม่เกี่ยวกับสินค้า และกล่าวถึงข้อมูลส่วนตัวเยอะ ในขณะที่รีวิวจริงมักจะเขียนอย่างไม่ยืดเยื้อ ซึ่งในการ Label ข้อมูลเหล่านี้ สามารถแบ่งออกมาได้เป็น 2 วิธีคือ

1. Conditional Labeling
   - รีวิวที่สั้นจนเกินไปกำหนดให้ มีอักขระไม่ถึง 35 ตัวในรีวิว หรือ length น้อยกว่า 35 โดยอ้างอิงตัวเลขจากการแปลซับไตเติลที่ต้องใช้ความกระชับในการแปล [8]
   - ไม่ค่อยมีข้อมูลเกี่ยวกับผู้รีวิวกำหนดให้ จำนวน follower, review, และ photo ทั้งหมดน้อยกว่า Quartile ที่ 1 และ review น้อยกว่าหรือเท่ากับ 1 รีวิวทั้งเว็บไซต์
   - รีวิวที่สุ่มเสี่ยงว่าอาจจะเป็นการว่าจ้างโฆษณาโดย Influencer หรือ Blogger กำหนดให้ จำนวน follower, review, และ photo ทั้งหมดมากกว่า Quartile ที่ 3 ดังตาราง

TABLE I. ตำแหน่ง QUARTILE ที่ 1 และ 3

Attributes	Q1	Q3 - maximum
follower	5	265-75841
review	28	512-8040
photo	98	3115-996544

Fig. 2. ตำแหน่ง Quartile ที่ 1 และ 3

2. Manual Labeling

นอกเหนือจากเงื่อนไขที่กำหนดไว้ข้างต้นแล้ว เราจำเป็นที่จะต้องทำการ Label ด้วยคน ซึ่งจะมีเกณฑ์ในการตัดสินใจดังนี้

- ข้อความที่ไม่เกี่ยวข้องกับร้านนั้นๆ
- มีข้อความที่ระบุชื่อร้านเต็ม มีชื่อเมนูอาหาร ที่ตั้ง เวลาเปิดปิด ให้ข้อมูลที่ครบถ้วนจนเกินไป
- ใช้คำที่เกินจริงอย่างเห็นได้ชัด ทั้งในแง่บวกและลบ

เมื่อทำการเตรียมข้อมูลและ Label แล้ว เหลือข้อมูลที่นำไปใช้ได้จริง 12,113 รีวิว ซึ่งแบ่งออกเป็น รีวิวจริง 8,923 รีวิว ซึ่งคิดเป็น 78.96% และรีวิวเท็จ 3,190 รีวิว ซึ่งคิดเป็น 21.04%

III. CLASSIFICATION MODEL

เมื่อทำการเตรียมข้อมูลเรียบร้อยแล้ว นำข้อมูลที่ได้มาทำ Classification Model ด้วยโปรแกรมภาษา Python โดยใช้ Library ต่างๆ ได้แก่ pythainlp datasketch โดยมีขั้นตอนตาม Fig. 3 ด้านล่าง

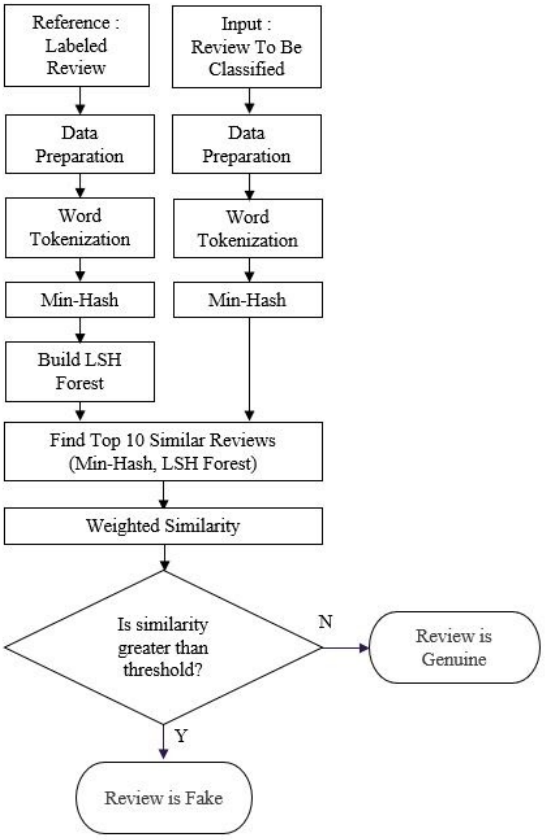

Fig. 3. ขั้นตอนการดำเนินงาน

A. *Word Tokenization*

เป็นขั้นตอนในการทำให้ข้อมูลรีวิวพร้อมใช้งาน [9] เช่นการทำความสะอาดข้อมูลโดยการลบช่องว่าง (White space) การลบเครื่องหมาย และการลบ icon ต่างๆ ออกจากข้อมูลรีวิว จากนั้นนำข้อมูลรีวิวมาแบ่งออกเป็นคำ ซึ่งงานวิจัยนี้เกี่ยวกับรีวิวที่เป็นภาษาไทยดังนั้นจึงเลือกใช้ไลบรารี PyThainlp ที่สามารถตัดแบ่งรีวิวออกเป็นคำได้อย่างถูกต้องในระดับที่ยอมรับได้ นั่นคือผลลัพธ์ที่ออกมาอ่านแล้วสามารถเข้าใจได้ โดยมี 2 ขั้นตอนคือ

จัดการกับข้อความที่เรียงคำไม่ถูกต้องหรือการใช้อักษรซ้ำซ้อน รวมไปถึงการใช้วรรณยุกต์ซ้ำซ้อน เช่น "แ" พิมพ์เป็น "เ เ" หรือ "ไส้" เป็นต้น โดยการทำ Normalize ซึ่งใช้คำสั่งดังนี้

```
from pythainlp.util import normalize

print(normalize('ไส้อั่วอร่อยมาก'))
```

ผลลัพธ์

'ไส้อั่วอร่อยมาก'

การแยกคำออกจากประโยค โดยมีกระบวนการตัดประโยค (string) ให้เป็นคำ (token) ซึ่งใช้คำสั่งดังนี้

```
from pythainlp.util import normalize
from pythainlp.tokenize import word_tokenize

print( word_tokenize(normalize('ไส้อั่วอร่อยมาก')))
```

ผลลัพธ์

['ไส้อั่ว', 'อร่อย', 'มาก']

B. Local-Sensitive Hashing (LSH)

การที่เรามีข้อมูลจำนวนมากที่มีลักษณะเป็นเซ็ตคอลเล็กชั่นขนาดใหญ่ เมื่อเราต้องการคัดเลือกหรือค้นหาเฉพาะข้อมูลที่ต้องการ (query) โดยที่ข้อมูลที่ต้องการค้นหามีลักษณะเป็นเซ็ต และต้องใช้ cost มหาศาลในการประมวลผล จึงมีการนำเสนอแนวคิดที่เรียกว่า Locality Sensitive Hashing (LSH) เพื่อจัดการกับปัญหาดังกล่าว และทำให้การค้นหาข้อมูลมีประสิทธิภาพ ขั้นตอนของอัลกอลิธึมนี้ เริ่มที่การทำให้เซ็ตของข้อมูลทั้งหมดอยู่ในรูปแบบ Shingle สร้าง MinHash สำหรับทุกเซ็ต โดยที่ยังสามารถใช้เป็นตัวแทนของข้อมูลดั้งเดิมได้ (Signature Matrix) และเมื่อต้องการจะค้นหาข้อมูลที่มีความคล้ายคลึงกันมากที่สุด จะคำนวนหาค่า Jaccard similarities ระหว่าง query MinHash และทุก MinHash ของ collection [10]

เนื่องจากข้อมูลรีวิวที่ทำการศึกษามีปริมาณมาก อีกทั้งมีขนาดที่ไม่แน่นอน เราจึงใช้หลักการของ LSH มาประยุกต์ใช้เพื่อจัดการกับปัญหาที่กล่าวข้างต้น ซึ่งสามารถจัดการกับข้อมูลที่เป็น Documents จำนวนมากได้ โดยที่สามารถลดขนาดของข้อมูล และยังคงเอกลักษณ์ของข้อมูลได้ในเวลาเดียวกัน อีกทั้งยังสามารถคำนวนหารีวิวที่มีลักษณะคล้ายคลึงกันออกมาได้ โดยเริ่มจากการนำคำ (Token) ที่แยกออกมาจากรีวิวให้อยู่ในรูปแบบ Set of Shingles ที่ประกอบไปด้วยคำในแต่ละรีวิวนั้น ๆ ซึ่งรีวิวที่คล้ายกันจะมี Shingles ตัวเดียวกันอยู่ เพื่อให้สามารถเทียบเคียงความเหมือนของรีวิวจากแต่ละ Token ได้

หลังจากนั้นนำ Set of Shingling มาสร้าง Signature Matrix ด้วยการทำ Minhash เริ่มจากการ random permutation จำนวน 128 แถว และ hash value ของข้อมูลแบบ 32-bit hash function ด้วยวิธี SHA1 [11]

และในขั้นตอนสุดท้าย เป็นการนำข้อมูลที่ถูกลดขนาดด้วย Minhash แล้วมาสร้างเป็น LSH Forest เพื่อใช้ในการคำนวณหาความคล้ายคลึงกันของข้อมูล ซึ่งข้อมูลที่จะนำมาสร้างจะเป็นข้อมูลต้นแบบ (Train Data) เท่านั้น การสร้าง LSH Forest มีข้อดีคือสามารถประมวลผล (Query) ได้เร็วกว่าการทำ LSH ที่มี BigO(N) เนื่องจาก LSH Forest สามารถ Query แบบ Binary search tree ได้ BigO(log(N)) [12]

LSH Forest เป็นการทำ hash table ของ LSH ในรูปแบบ prefix tree ที่มี depth มากที่สุดที่เป็นไปได้เท่ากับ จำนวน hash values (เลข permutation) / จำนวน tree ซึ่งในที่นี้กำหนดไว้ที่ 128/8 พร้อมทั้งมี key สำหรับอ้างอิงรีวิวตั้งต้น อีกทั้งยังมีอัลกอริทึมในการค้นหาที่ทำการหา key ของรีวิวที่คล้ายกันมากที่สุดจาก forest โดยวัดจากค่าประมาณ Jaccard Similarity ที่มากที่สุด เพื่อที่จะระบุรีวิวจาก train forest ว่าเป็นรีวิวใด รวมถึง LSH Forest ยังสามารถกำหนดจำนวนผลลัพธ์ที่ต้องการได้ [13] ซึ่งเหมาะกับการศึกษานี้ เพราะเราต้องการรีวิวที่เหมือนที่สุดจำนวน 10 อันดับ ในขณะที่ LSH จะต้องกำหนดค่า minimum Jaccard Similarity ที่เหมาะสมเพื่อให้ได้รีวิวที่มีความคล้ายกันมากกว่าค่านั้น ซึ่งจะไม่สามารถทราบได้ว่าผลลัพธ์ที่คืนค่ากลับมาจะมีจำนวนเท่าใด หรือค่าที่กำหนดนั้นอาจจะไม่คืนค่าใดกลับมาเลย เนื่องจากไม่มีรีวิวที่มีค่ามากกว่าค่า Jaccard Similarity ที่กำหนด

IV. RESULT

A. Review Classification

ผลลัพธ์จากการทำ LSH Forest คือ ได้ชุดข้อมูลรีวิว ที่มีค่า Jaccard Similarity เรียงลำดับจากมากสุดไปจำนวน 10 อันดับ จากนั้นนำผลลัพธ์นี้มาคำนวณแบบถ่วงน้ำหนักตามสมการ ดังนี้

$$Y_{pred} = \sum_{i=1}^{10} \frac{X_i Y_i}{X_i}$$

โดยที่ X_i คือค่า Jaccard Similarity

Y_i คือ label ของรีวิวมีค่าเป็น 0 และ 1

i คือ Row index ของชุดข้อมูลซึ่งมีจำนวน 10 อันดับ

จากนั้นนำผลลัพธ์ที่ได้ Y_{pred} มาเปรียบเทียบกับค่า Optimal Threshold จะได้ว่า

$$Y_{pred} = \begin{cases} 1; Y_{pred} \geq Optimal\ Threshold \\ 0; Y_{pred} < Optimal\ Threshold \end{cases}$$

ผลลัพธ์ที่เราได้จากแบบจำลองการจำแนกประเภทของรีวิวนั้น จะได้ผลลัพธ์เป็นค่าทำนายความน่าจะเป็นของ label ซึ่ง label ที่ศึกษานี้ มีรูปแบบเป็น binary class คือ 0 และ 1 ซึ่งต้องมีการเลือกค่าที่เหมาะสมที่สุดที่ใช้ในการจำแนกประเภทรีวิว (Optimal Threshold) [14]

จากค่าความน่าจะเป็น ขึ้นอยู่กับความสำคัญของค่าผลบวกจริงและผลบวกลวง (True Positive Rate and False Positive Rate) ในที่นี้จะให้ความสำคัญเท่าๆกัน เพื่อให้ได้ค่าผลบวกจริงสูงและผลบวกลวงต่ำ โดยใช้วิธีการวิเคราะห์ Receiver operating characteristic (ROC) curve และ Area under the curve (AUC) เพื่อหาจุดที่มีความแม่นยำเกิดขึ้นสูงที่สุด น่าเชื่อถือที่สุดและผิดพลาดน้อยที่สุด ROC curve นั้นจะแสดงให้เห็นว่าเมื่อค่า Threshold เปลี่ยนไปจะทำให้ค่าผลบวกจริงและผลบวกลวงเปลี่ยนตามไปด้วย

วิธีการเลือกค่า Optimal Threshold นั้นใช้หลักการของ Youden's J statistic [15] หรือค่า Youden's index ที่ให้ค่า index ที่มากที่สุด จากสมการ

$$J = TruePositiveRate - FalsePositiveRate$$

B. Result Evaluation

จากขั้นตอนดังกล่าว เมื่อนำผลลัพธ์ที่ได้จากแบบจำลองมาทำการวิเคราะห์ ROC curve และ Area AUC ผ่านโปรแกรมภาษา Python โดยใช้ Library pandas, numpy, sklearn และ matplotlib ได้ผลลัพธ์ ดัง Fig. 4

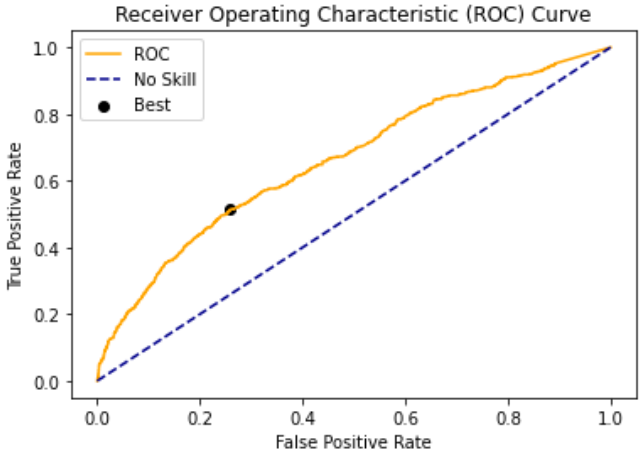

Fig. 4. ผลลัพธ์จากการวิเคราะห์ ROC curve

เมื่อเทียบกับการสุ่มเลือก ที่จะมีค่า AUC = 0.500 การทำแบบจำลองด้วย LSH ได้ค่า AUC = 0.663 แสดงให้เห็นว่าการทำแบบจำลองได้ผลลัพธ์ที่ดีกว่าการเดาสุ่ม และเมื่อคำนวนหาค่า Maximum Youden's index แล้ว ได้ค่า Optimal Threshold อยู่ที่ 0.319 จากค่า Threshold ดังกล่าว จะทำให้ได้ค่าชี้วัดผลลัพธ์การทำนายต่างๆ ดังนี้

Recall = 0.514	Precision = 0.409
Accuracy = 0.682	F1 = 0.456

V. Conclusion

จากผลการศึกษาพบว่า วิธีการหาชุดข้อมูลที่คล้ายกันด้วย Local-Sensitive Hasing สามารถนำมาประยุกต์ใช้ในการจำแนกรีวิวจริงและรีวิวเท็จได้ ในแง่ของความคล้ายคลึงกันของข้อความรีวิว แต่อย่างไรก็ตามในการศึกษานี้ใช้เพียงแค่ข้อมูลจากข้อความรีวิวในช่วงเวลาที่กำหนดจากแพล็ตฟอร์มเดียวเท่านั้น อีกทั้งยังเป็นแค่การประยุกต์ใช้แนวคิดเพียงแนวคิดเดียว ซึ่งยังสามารถพัฒนาการศึกษาให้มีความถูกต้องแม่นยำยิ่งขึ้นด้วยการเพิ่มตัวแปรอื่นๆที่เกี่ยวข้อง รวมถึงการนำหลักทฤษฎีอื่นๆมาใช้ เพื่อนำไปต่อยอดได้ในอนาคต

A. Future Work

ในการศึกษาการจำแนกรีวิวเท็จยังสามารถทำแบบจำลองเพิ่มเติมได้ โดย

- ใช้หลักการทาง Natural Language Processing (NLP) อื่นๆมาประยุกต์ใช้เพิ่มเติม เช่น TF-IDF หรือ Term Frequency-Inverse Document Frequency ซึ่งเป็นหนึ่งในวิธีหา คำ (term) ที่สำคัญ ใน เอกสาร (document) โดยดูจากเนื้อหาของเอกสารทั้งหมด มักจะใช้ในงานพวก Information-retrieval หรือ Text mining เพื่อศึกษาว่าคำใด ที่มักจะปรากฏอยู่ในรีวิวเท็จ
- สร้างแบบจำลองเพื่อการทำ Sentimental Analysis วิเคราะห์ความรู้สึก อารมณ์ต่างๆที่ผู้เขียนต้องการสื่อจากข้อความรีวิว เช่น การรีวิวเชิงบวก เชิงลบ
- สร้างแบบจำลองจากแบบจำลอง หรือ การทำ Model over model เช่น การใช้ผลลัพธ์ หรือค่าต่างๆที่ได้จากแบบจำลองหนึ่ง เป็นตัวแปรประกอบในอีกแบบจำลองหนึ่ง ตัวอย่างเช่น จากการศึกษานี้ สามารถนำค่า Jaccard Similarity มาเป็นหนึ่งในตัวแปร ร่วมกับตัวแปรอื่นๆที่

ทำการ scrape มาได้ เช่น คะแนนรีวิว วันที่รีวิว จำนวนรีวิวของผู้เขียน มาทำแบบจำลอง Classification เพื่อจำแนกรีวิวเท็จได้

- สร้างแบบจำลองอื่น ๆ ในการทำนายผลลัพธ์ และนำมาเปรียบเทียบกันเพื่อ comparing model เพื่อหาว่าแบบจำลองใดเหมาะสมที่สุดที่จะใช้ในการจำแนกรีวิวเท็จ

REFERENCES

[1] กรมพัฒนาธุรกิจการค้า, กรมพัฒน์ฯ ดึงพลัง Startup ดันธุรกิจร้านอาหารให้เติบโตแบบก้าวกระโดด, 2018, Accessed on: April 22, 2021, Available: https://www.dbd.go.th/news_view.php?nid=469407297

[2] สำนักงานพัฒนาธุรกรรมทางอิเล็กทรอนิกส์, เจาะพฤติกรรมซื้อสินค้าออนไลน์, 2015, Accessed on: April 25, 2021, Available: https://www.digitalagemag.com/เจาะพฤติกรรม-ซื้อสินค้า

[3] S. Rudolph, The Impact of Online Reviews on Customers' Buying Decisions [Infographic], 2015, Accessed on: May 2, 2021. Available: https://www.business2community.com/infographics/impact-online-reviews-customers-buying-decisions-infographic-01280945.

[4] SME Thailand, ปรับกลยุทธ์ร้านอาหารรับเทรนด์ ลูกค้าหารีวิวก่อนกิน, 2019, Accessed on: April 20, 2021, Available: https://www.smethailandclub.com/marketing-4178-id.html

[5] K. McCabe, 9 Ways to Spot a Fake Review (+How Amazon is Fighting Back), 2019, Accessed on: May 2, 2021, Available: https://learn.g2.com/fake-reviews

[6] "ข้อมูลเกี่ยวกับวงใน" Accessed on: April 20, 2021, Available: https://www.wongnai.com/about

[7] M.C.Ashwini,M.C. Padma, Efficiently analyzing and detecting fake reviews through opinion mining, International Journal of Computer Science and Mobile Computing, 2020, Vol.9, Issue.7, p. 97-108.

[8] "Netflix: Timed Text Style Guides" Accessed on: May 1, 2021. Available: https://partnerhelp.netflixstudios.com/hc/en-us/articles/220448308-Thai-Timed-Text-Style-Guide.

[9] C. Tapsai, P. Meesad, H. Unger, An Overview on the Development of Thai Natural Language Processing, Information Technology Journal, 2019, Vol.15.

[10] J.Leskovec, Stanford University, California, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, pp 68-122.

[11] S.Aradhana, S. M. Ghosh, Review Paper on Secure Hash Algorithm With Its Variants, International Journal of Technical Innovation in Modern Engineering & Science, 2017, Vol.3, Issue. 5.

[12] B. Mayank, T. Condie, P. Ganesan, LSH Forest: Self-Tuning Indexes for Similarity Search, International World Wide Web Conference Committee (IW3C2), 2005.

[13] M. Cochez, V. Terziyan, V. Ermolayev, Large Scale Knowledge Matching with Balanced Efficiency-Effectiveness Using LSH Forest, Transactions on Computational Collective Intelligencev XXVI, 2017, pp. 46-66.

[14] J. Brownlee, A Gentle Introduction to Threshold-Moving for Imbalanced Classification, 2020, Accessed on: May 1, 2021, Available: https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/

[15] M.D. Ruopp, J.N. Perkins, B.W. Whitcomb, E. F. Schisterman, Youden Index and Optimal Cut-Point Estimated from Observations Affected by a Lower Limit of Detection, Biometrical Journal, 2008, Vol.50, p.419-430.