การสังเคราะห์ข้อความเพื่อเพิ่มตัวอย่างการตรวจจับข้อความประทุษวาจาในข้อความภาษาไทย

นายธโนภาส วรรณวโรทร

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2564

Text synthesis to add an example for detecting hate speech in Thai massages.

Mr. Thanopath Wanwarotorn

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Science in Computer Science

Department of Computer Engineering

FACULTY OF ENGINEERING

Chulalongkorn University

Academic Year 2021

หัวข้อวิทยานิพนธ์	การสังเคราะห์ข้อความเพื่อเพิ่มตัวอย่างการตรวจจับข้อความประทุษวาจาในข้อความภาษาไทย
โดย	นายธโนภาส วรรณวโรทร
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(รองศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ)

ธโนภาส วรรณวโรทร : การสังเคราะห์ข้อความเพื่อเพิ่มตัวอย่างการตรวจจับข้อความประทุษวาจาในข้อความภาษาไทย. (Text synthesis to add an example for detecting hate speech in Thai massages.) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุกรี สินธุภิญโญ

ในงานวิจัยนี้เป็นการศึกษาวิธีการแก้ไขปัญหาในการจำแนกข้อความประทุษวาจา ด้วยวิธีการสังเคราะห์ข้อความขึ้นเพื่อแก้ไขปัญหาของการเกิดชุดข้อมูลไม่สมดุลที่ปรากฏในข้อมูลที่เก็บรวบรวมมาจากทวิตเตอร์ ซึ่งหลังจากเก็บรวบรวม ทำความสะอาดข้อมูลและติดฉลากข้อมูลแล้ว ผู้วิจัยได้สร้างตัวอย่างเพิ่มเติม 3 วิธีคือ คือ 1. การสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์ (Synthetic Minority Over-sampling Technique: SMOTE) 2. เทคนิคการสร้างข้อความเพิ่ม (Text generation) 3.เทคนิคคำฝังตัว (Word Embedding) เป็นวิธีการในการใช้สังเคราะห์ตัวอย่างเพิ่มเติม ให้เกิดความสมดุลก่อนที่จะนำข้อมูลชุดใหม่ที่สร้างขึ้นใหม่แบ่ง ตัวอย่างเป็น 3 รูปแบบในการจำแนกข้อความประทุษวาจา คือ 1. อัลกอริทึมนาอีฟเบย์ (Navie bays) 2. หน่วยความจำระยะสั้นแบบยาว (LSTM) 3. หน่วยความจำระยะสั้นแบบยาว ร่วมกับ โครงข่ายประสาทแบบคอนโวลูชัน (LSTM + CNN) เพื่อเป็นการจำแนกข้อความประทุษวาจา ในชุดข้อความที่เป็นข้อความธรรมดา โดยผลการทดลองการจำแนกข้อความมีความหมายเชิงประทุษวาจา ซึ่งในการทดลองแรกได้ลองใช้ข้อมูลที่ไม่สมดุล จากผลการทดลองทั้ง 3 รูปแบบที่ใช้ในการจำแนกซึ่งให้ความถูกต้องไม่สูงเท่าที่ควร จากนั้นจึงทำการแก้ไขปัญหาในชุดของข้อมูลทำให้ได้ความถูกต้องสูงขึ้นในทุกชุดของทุกโมเดล

สาขาวิชา วิทยาศาสตร์คอมพิวเตอร์ ลายมือชื่อนิสิต .....

ปีการศึกษา 2564 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6270117721 : MAJOR COMPUTER SCIENCE

KEYWORD: Natural Language Processing Text Classification Long Short-Term Memory Convolutional Neural Network Word2Vec

Thanopath Wanwarotorn : Text synthesis to add an example for detecting hate speech in Thai massages.. Advisor: Asst. Prof. SUKREE SINTHUPINYO, Ph.D

In this paper, we present a method for solving a problem in classifying text messages containing Hate Speech by synthesizing messages to solve the problem of the imbalance in text corpuses that were collected from Twitter. After collecting, cleansing, and labeling the data, we augmented samples using three methods, namely 1) Synthetic Minority Over-sampling Technique (SMOTE), 2) Text generation technique, and 3) Word Embedding. In this research, we used three text classification techniques: Naive Bayes, Long Short-Term Memory (LSTM), and a combination of Long Short-Term Memory and Convolutional Neural Network (CNN). The accuracy of the text classification on imbalanced text data was not high. However, after we added the text from minority class to the training set, the accuracy become higher in all classification models.

Field of Study: Computer Science Student's Signature .....

Academic Year: 2021 Advisor's Signature .....

กิตติกรรมประกาศ

ผู้วิจัยขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ อาจารย์ที่ปรึกษาวิทยานิพนธ์ที่กรุณาคอยให้คำปรึกษาจนวิทยานิพนธ์นี้สามารถสาเร็จได้ด้วยดี

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์ ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ และ ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ ผู้ให้เกียรติเป็นกรรมการสอบวิทยานิพนธ์ และชี้แนะแนวทางในการปรับปรุงวิทยานิพนธ์ให้ดียิ่งขึ้น

ขอขอบคุณคุณแม่ที่คอยอยู่ข้างเคียงและสนับสนุนมาตลอด ทำให้ผู้วิจัยได้มาถึงที่ฝันไว้ และขอขอบคุณเพื่อนๆ พี่ น้องๆทุก ๆ ท่านที่เคยให้ความช่วยเหลือผู้วิจัยตลอดมา ที่ไม่ได้กล่าวถึงในที่นี้

ธโนภาส วรรณวโรทร

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่1 บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในปัจจุบันเกิดกลุ่มข้อความที่แสดงออกถึงการแสดงการความเกลียดชังเกิดขึ้นมากมายและส่งผลต่อบุคคลหรือกลุ่มสังคมเป็นอย่างมาก [3] ในปัจจุบันงานวิจัยที่สามารถทำงานในโดเมนของข้อความประทุษวาจาได้ยังไม่แพร่หลาย ส่วนหนึ่งอาจเพราะปัญหาจากการเก็บรวบรวมข้อมูลที่รวบรวมมาแล้วเกิดปัญหาเป็นกลุ่มข้อมูลส่วนน้อยในข้อความที่เป็นประทุษวาจา ทางผู้วิจัยจึงได้ทำการศึกษาวิธีการสังเคราะห์ข้อความที่มีความหมายในเชิงการแสดงความเกลียดชัง หรือ ประทุษวาจา เพื่อใช้ในการแก้ไขปัญหาของชุดข้อมูลกลุ่มน้อยให้เกิดความสมดุลกัน เพื่อพัฒนาประสิทธิภาพการตรวจจับข้อความประทุษวาจาในข้อความภาษาไทย ที่มีความหมายในเชิงการแสดงความเกลียดชังที่เก็บรวบรวมมาจากแพลตฟอร์มทวิตเตอร์[7] อีกปัญหาของการที่เก็บรวบรวมข้อความตัวอย่างที่ถูกรวบรวมมานั้นเกิดการเลื่อมล้ำของข้อความส่งผลให้เกิดปัญหาต่อทำให้ผู้วิจัยเห็นถึงปัญหาของปริมานข้อมูลที่มีไม่เท่ากัน เพราะธรรมชาติของข้อความที่เก็บรวบรวมมาจากแพลตฟอร์มทวิตเตอร์นั้นเป็นข้อความที่ถูกพูดถึงซ้ำกันในหลาย ๆ ครั้งและเป็นข้อความสั้น ๆ ที่เกิดในช่วงเวลาช่วงนึง จึงทำให้ข้อมูลที่เป็นประทุษวาจามีปริมาณที่น้อยกว่าจำนวนข้อความธรรมดา เป็นเหตุที่ทำให้ข้อมูลที่ใช้ในการเรียนรู้ของเครื่องมีความแตกต่างกันมาก และส่งผลให้ความแม่นยำในการจำแนกข้อความมีผลลัพธ์ที่ไม่ดีเท่าที่ควร [1, 5]

ในงานวิจัยนี้ผู้วิจัยทำการศึกษาวิธีการแก้ปัญหาข้อมูลที่ไม่สมดุลในเทคนิคต่างๆ ด้วยวิธีทั้งหมด 3 คือ 1.การสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์ (Synthetic minority over-sampling technique: SMOTE) [9] 2. เทคนิคการสร้างข้อความเพิ่มจากการใช้เทคนิคความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร (Term Frequency Inverse Document Frequency หรือ TF-IDF) [8] 3.เทคนิคการเปลี่ยนคำเป็นเวกเตอร์ (Word2VEC) [11] วิธีการแก้ปัญหาชุดข้อมูลไม่สมดุลนั้น การสังเคราะห์กลุ่มข้อมูลเพิ่มเพื่อให้จำนวนตัวอย่างในชุดตัวอย่างที่เป็นข้อความประเภทประทุษวาจา จากนั้นนำข้อมูลที่ผ่านการสังเคราะห์ข้อมูลที่สมดุลด้วยวิธีการข้างต้น นำเข้าสู่ตัวแบบเพื่อทำการจำแนกข้อความประทุษวาจาโดยทั้งหมด 3 รูปแบบ คือ 1. อัลกอริทึมนาอีฟเบย์ (Naive Bayes) [9] ซึ่งเป็นวิธีการจำแนกข้อความที่เป็นเทคนิคพื้นฐานที่ใช้ทั่วไปในการใช้จำแนกข้อความ 2. หน่วยความจำระยะสั้นแบบยาว (LSTM) [12] เป็นอีกเทคนิคที่นิยมใช้ในการจำแนกข้อความ รวมถึงการสร้างข้อความเพิ่มเติม จึงเป็นอีกวิธีที่นำมาใช้ในการทดลองนี้ [8] และ 3. หน่วยความจำระยะสั้นแบบยาว ร่วมกับ โครงข่ายประสาทแบบคอนโวลูชัน (LSTM + CNN) ซึ่งเป็นตัวแบบที่นิยมใช้อย่างแพร่หลายในการจำแนกข้อความที่พัฒนาเพิ่มเติมมาจากวิธีที่ 2 ในสุดท้ายจะนำผลลัพธ์ของโมเดลทั้ง 3 ชุด เป็นตัวเปรียบเทียบของวิธีการสังเคราะห์ตัวอย่างเพิ่มเติมในกลุ่มน้อย ที่กล่าวมาในข้างต้นว่าวิธีการที่ผู้วิจัยนำเสนอ

1.2 วัตถุประสงค์ของงานวิจัย

ในงานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการสังเคราะห์ข้อมูลตัวอย่างในกลุ่มน้อย เพื่อเพิ่มตัวอย่างข้อมูลที่เป็นข้อความประทุษวาจาความเป็นความหมายเชิงประทุษวาจาในแฟลตฟอร์มทวิตเตอร์ เพื่อแก้ปัญหาชุดข้อมูลไม่สมดุลในการจำแนกข้อความประทุษวาจา ที่จะมีผลกับตัวแบบในการจำแนกให้ดีขึ้น

1.3 ขอบเขตของการวิจัย

การเก็บรวบรวมข้อมูลในงานวิจัยจะเก็บข้อมูลจากทวิตเตอร์ โดยพิจารณาจากทวีตที่เป็นภาษาไทยในระหว่างวันที่ 1 พฤศจิกายน 2563 ถึง 31 มกราคม 2564

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1.4.1. สามารถเสนอวิธีการที่เหมาะสมสำหรับช่วยให้การจำแนกประเภทข้อความบนชุดข้อมูลที่ไม่สมดุลมีประสิทธิภาพมากขึ้น

1.4.2 สามารถสร้างตัวแบบที่มีประสิทธิภาพสำหรับจำแนกประเภทของข้อความประทุษวาจาภาษาไทย เพื่อเป็นประโยชน์ต่อผู้ต้องการนำไปใช้ในการจำแนกข้อความประทุษวาจาต่อไป

1.5 ขั้นตอนการดำเนินงานวิจัย

1.5.1. ศึกษาทฤษฎีที่เกี่ยวข้อง

1.5.2. ศึกษางานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ข้อความที่เป็นข้อความเชิงประทุษวาจา

1.5.3. ศึกษาวิธีการเก็บรวบรวมข้อมูลทวิตเตอร์

1.5.4. เก็บรวบรวมข้อมูล ทำความสะอาดข้อมูล และทำติดฉลากให้ข้อมูล

1.5.6. ทำการสังเคราะห์ข้อมูลเพิ่มเติม

1.5.7. เขียนบทความเพื่อเผยแพร่ผลงานวิชาการ

1.5.8. สอบโครงร่างวิทยานิพนธ์

1.5.9. ปรับปรุงวิธีการสังเคราะห์ข้อมูลเพิ่มเติม

1.5.10. สรุปและเรียบเรียงวิทยานิพนธ์

1.5.11. สอบวิทยานิพนธ์

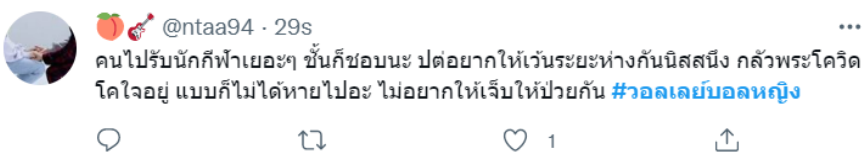
บทที่2 ทฤษฎีที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่นำมาประยุกต์ใช้กับการเพิ่มกลุ่มตัวอย่างที่ไม่สมดุลและจำแนกข้อความที่เป็นข้อความประทุษวาจา บนสื่อสังคมออนไลน์ โดยได้ทำการค้นคว้า หลักการ ทฤษฎีและงานวิจัยที่เกี่ยวข้อง เพื่อนำไปสู่แนวคิดและวิธีการดำเนินงานซึ่งทฤษฎีที่เกี่ยวข้องได้แก่ ทวิตเตอร์ ทวิตเตอร์เอพีไอ การสังเคราะห์ข้อมูลเพิ่ม การตัดคำภาษาไทย การจำแนกแบบนาอีฟเบย์ โครงข่ายประสาทแบบคอนโวลูชัน หน่วยความจำชนิดสั้นแบบยาว การวัดประสิทธิภาพการจำแนกของแบบจำลอง งานวิจัยที่เกี่ยวข้องกับการจำแนกข้อความประทุษวาจาบนสื่อสังคมออนไลน์ รวมถึงวิธีการวัดประสิทธิภาพของโมเดล

2.1 ทวิตเตอร์

ทวิตเตอร์ (Twitter) เปิดตัวในปี 2549 และมีต้นกำเนิดในชื่อ บริษัท Podcasting Odeo นับตั้งแต่นั้นมา แพลตฟอร์มดังกล่าวได้เติบโตขึ้นเป็นเครื่องมือสื่อสารระดับโลกและพื้นที่ทางสังคมดิจิทัลซึ่งโดยทั่วไปเรียกว่า "Twitter verse" ซึ่งส่งเสริมการสนทนาและการรวบรวมข้อมูลผ่านทวีตที่ส่งและค้นหาได้ นับตั้งแต่ก่อตั้งขึ้นมา แพลตฟอร์มโซเชียลได้พัฒนาเป็นหนึ่งในเครือข่ายโซเชียลที่ได้รับความนิยมมากที่สุดในโลกที่ผู้คนเชื่อมต่อถึงกันผ่านการสนทนาการค้นหาและการค้นพบ

ทวิตเตอร์ เป็นเครือข่ายสังคมออนไลน์และบริการไมโครบล็อกสำหรับการสื่อสารแบบเรียลไทม์ที่ใช้โดยผู้คนและองค์กรหลายล้านคน ผู้ใช้ ทวิตเตอร์ ติดต่อกันโดยการโพสต์อัปเดตที่เรียกว่า "ทวีต" ไปยังไซต์เพื่อแบ่งปันแลกเปลี่ยนและค้นหาข้อมูล ทวีตประกอบด้วยอักขระ 280 ตัวหรือน้อยกว่า และสามารถมีแนวคิดและข้อมูลประเภทต่าง ๆ เช่นภาพถ่ายวิดีโอและลิงก์ไปยังข้อความ ผู้ใช้สามารถเข้าถึงข้อความที่โพสต์เหล่านี้บน ทวิตเตอร์ผ่านทางเว็บหรือบนอุปกรณ์มือถือที่มีการเชื่อมต่ออินเทอร์เน็ต เมื่อผู้ใช้โพสต์ทวีต ข้อความจะถูกโพสต์ไปยังโปรไฟล์ของพวกเขาและส่งไปยังหน้าแรก หรือฟีด ของผู้ใช้และผู้ติดตามของผู้ใช้สามารถรับการอัปเดต ทวีตเตอร์ ของผู้ใช้นั้นๆ ผู้ใช้สามารถเลื่อนดูฟีดเพื่อดูการอัปเดต ทวีตเตอร์ จากผู้ติดตามซึ่งอาจเป็นอะไรก็ได้ตั้งแต่การสังเกตส่วนบุคคลไปจนถึงข่าวด่วน ผู้ใช้สามารถติดตามหรือเป็นผู้ติดตามของเพื่อนครอบครัวเพื่อนร่วมงานองค์กรธุรกิจและบุคคลสาธารณะเพื่อดูข้อมความที่ทวีตเตอร์ในหน้าแรกได้ คล้ายกระดานข่าว



รูปภาพ 1 รูปตัวอย่างหน้าทวิตเตอร์ของผู้ใช้รายหนึ่ง

ผู้ติดตามสามารถสื่อสารกับผู้ใช้โดยการตอบกลับทวีตหรือ "รีทวีต (Retweet)" โพสต์เพื่อแชร์การอัปเดตกับผู้ติดตามของตนเองอีกครั้ง ทวีตและการตอบกลับเป็นสาธารณะและทุกคนสามารถมองเห็นได้เว้นแต่ทวีตจะได้รับการปกป้องด้วยการตอบสนองและแบ่งปันทวีตผู้ใช้มีส่วนร่วมในแบบเรียลไทม์ผ่านเนื้อหาที่น่าสนใจหัวข้อยอดนิยมและเหตุการณ์ปัจจุบัน

ผู้ใช้จะถูกระบุบน ทวิตเตอร์ โดยหมายเลขอ้างอิงซึ่งเป็นชื่อผู้ใช้ที่กำหนด ผู้ใช้กล่าวถึงผู้ติดตามหรือผู้ใช้รายอื่นโดยใช้สัญลักษณ์ @ และ ทวีตยังสามารถค้นหาได้โดยใช้ แฮทแท็ก (hashtags) และติดแท็กสัญลักษณ์ # ตามด้วยคำสำคัญ

ส่วนประกอบของ ทวิตเตอร์มีเครื่องมือที่ใช้ในทวิตเตอร์มีดังนี้

- ตอบกลับ (Reply): คุณสามารถ ตอบกลับ ทวีตของคนอื่นได้
- ข้อความส่งตรง (Direct Message) : สามารถส่งข้อความส่วนตัวถึงคนอื่น โดยไม่มีใครเห็น
- ชื่นชอบ (Favorites): สามารถเก็บ ทวีต ที่สนใจใน ชื่นชอบ ได้
- รีทวิต (Retweet) : ถ้าเห็น ทวิต ของใครน่าสนใจ อยากจะ รีทวิต ในกลุ่มของคุณก็ทำได้คล้ายการส่งต่อ
- ลิ้งก์ (Links) : ใส่ ลิ้งค์ ในข้อความที่ต้องการ ค้นหาคำ หรือเรื่องที่คน ทวีต
- เรื่องที่ได้รับความนิยม (Trending Topics): เรื่องที่กำลังอยู่ในความสนใจของคนใช้ทวิตเตอร์

โดยทั่วไป มีการใช้ ทวิตเตอร์กันมาก และกำลังเป็นเทรนด์ที่กำลังมาแรง ซึ่ง ณ ตอนนี้ ทวิตเตอร์ ถูกนำมาใช้งานในในลักษณะที่แตกต่างกัน บางคนใช้เพื่อตอบคำถามและติดตามเพื่อนๆ คนรู้จักว่าเขากำลังทำอะไรอยู่ที่ไหน แต่มีหลายคน รวมถึงนักการตลาดที่เห็นประโยชน์ของทวิตเตอร์ มากกว่านั้น จึงเริ่มมีการใช้งาน ทวิตเตอร์ ในด้านอื่นๆ เช่น คนทั่วไปใช้ ทวิตเตอร์ ให้เพื่อนๆ และคนสนิท ติดตามซึ่งกันและกัน หรือ ใช้ ทวิตเตอร์ เป็นเสมือนอีกหนึ่งสังคมออนไลน์ เพื่อสนทนาและพูดคุยสร้างสัมพันธ์กับคนอื่นมากขึ้น กูรู หรือคนที่เชียวชาญและชำนาญในเฉพาะเรื่อง นิยมใช้ ทวิตเตอร์ ในการสร้างชื่อเสียงให้กับตัวเอง เจ้าของสินค้า หรือ แบรนด์ เริ่มใช้ ทวิตเตอร์ ในประชาสัมพันธ์สินค้า และบริการของแบรนด์ หรือร้านค้าต่างๆ รวมไปถึงการใช้เป็นเครื่องมือติดต่อและสร้างสัมพันธ์กับลูกค้า แทนการใช้คอลเซ็นเตอร์ หรือ ดารา นักแสดง นักร้อง นิยมใช้ทวิตเตอร์ให้แฟนคลับได้ติดตามแบบส่วนตัว

2.2 ทวิตเตอร์เอพีไอ

ทวิตเตอร์เอพีไอ (Twitter API) [7] เป็นเครื่องมือที่ทางทวิตเตอร์ อนุญาตให้เข้าถึงส่วนต่างๆ ของแพลตฟอร์มของทวิตเตอร์ เพื่อให้ผู้คนสามารถสร้างซอฟต์แวร์ที่ทำงานร่วมกับทวิตเตอร์ ซึ่งโดยธรรมชาติของข้อมูลบนทวิตเตอร์นั้นรูปแบบไม่ซ้ำกับข้อมูลที่แชร์โดยแพลตฟอร์มโซเชียลอื่น ๆ เพราะส่วนใหญ่ข้อมูลของทวิตเตอร์มีผู้ใช้เลือกที่จะแบ่งปันแบบสาธารณะมากกว่าแพลตฟอร์มโซเชียล ดังนั้นทวิตเตอร์เอพีไอของทวิตเตอร์ อนุญาตให้การเข้าถึงข้อมูลสาธารณะได้ในวงกว้างซึ่งผู้ใช้เลือกที่จะแบ่งปันกับคนทั้งโลก ทวิตเตอร์เอพีไอ ช่วยให้คุณเข้าถึงข้อมูลและส่งข้อมูล เช่น ข้อความและสื่อโดยไม่ต้องเปิดแอปพลิเคชันทวิตเตอร์ คุณส่งข้อความหาพวกเพื่อนโดยใช้คำสั่งของภาษาโปรแกรมต่างๆ ซึ่งทวิตเตอร์เอพีไอเป็นเครื่องมือที่มอบสิทธิ์เข้าถึงข้อมูลตามสิทธิที่ทางทวิตเตอร์กำหนดไว้ให้

การเข้าถึงข้อมูล ทวิตเตอร์ เมื่อมีคนต้องการเข้าถึง เอพีไอของทวิตเตอร์ พวกเขาจะต้องลงทะเบียน แอปพลิเคชันตามค่าเริ่มต้น โดยปกติแล้วแอปพลิเคชันสามารถเข้าถึงข้อมูลสาธารณะบน ทวิตเตอร์ เท่านั้น ข้อมูลบาง

ประเภทเช่น การส่งหรือรับข้อความโดยตรง จำเป็นต้องได้รับการอนุญาตเพิ่มเติมจากผู้ใช้งานก่อนจึงจะสามารถเข้าถึงข้อมูลของผู้ใช้งานได้ สิทธิ์เหล่านี้ไม่ได้รับตามค่าเริ่มต้น ผู้ใช้จะต้องเลือกตามแต่ละแอปพลิเคชันว่าจะให้การเข้าถึงนี้หรือไม่ และสามารถควบคุมแอปพลิเคชันทั้งหมดที่ได้รับอนุญาตในบัญชีของคุณ ทวิตเตอร์เอพีไอ มีการเข้าถึงข้อมูลที่หลากหลาย ซึ่งแบ่งออกเป็นห้ากลุ่มหลัก

บัญชีและผู้ใช้เราอนุญาตให้นักพัฒนาจัดการโปรไฟล์และการตั้งค่าของบัญชีโดยทางโปรแกรม ปิดเสียงหรือบล็อกผู้ใช้ จัดการผู้ใช้และผู้ติดตาม ขอข้อมูลเกี่ยวกับกิจกรรมของบัญชีที่ได้รับอนุญาต และอื่นๆ จุดเหล่านี้สามารถช่วยบริการในเรื่องเหตุต่าง เช่น แผนกการจัดการเหตุฉุกเฉินในเครือจักรภพแห่งเวอร์จิเนีย ซึ่งให้ข้อมูลแก่ผู้อยู่อาศัยเกี่ยวกับการรับมือเหตุฉุกเฉินและการแจ้งเตือนเหตุฉุกเฉิน

2.3 การสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ในงานจำแนกประเภทของข้อมูล (Classification) ผลลัพธ์ของประสิทธิภาพของแบบจำลอง ขึ้นอยู่กับปริมาณของกลุ่มตัวอย่างข้อมูล (Dataset) เป็นส่วนสำคัญหากชุดข้อมูล ถ้าเกิดเหตุการณ์ชุดมีปริมาณที่แตกต่างกัน (Imbalanced Class) การประมวลผลข้อมูลย่อมทำให้ประสิทธิภาพของตัวแบบลดลง เพราะในการเรียนรู้จะมีแต่ข้อมูลในกลุ่มมากทำให้มีการเรียนรู้แต่ข้อมูลกลุ่มมาก (Majority Class) ผลที่ได้ก็จะจำแนกไปในข้อมูลกลุ่มมากที่เก่งมาก ส่วนข้อมูลกลุ่มน้อย (Minority Class) ถูกจำแนกได้น้อยกว่าตัวอย่างหลักเป็นเหตุให้ ผลลัพธ์ของประสิทธิภาพของแบบจำลองได้ผลไม่ดีเท่าที่ควรซึ่งในงานวิจัยนี้ ชุดข้อมูลกลุ่มน้อยคือข้อความประทุษวาจาที่ทางวิจัยสนใจ

การแก้ปัญหาการแบ่งกลุ่มข้อมูลที่ไม่สมดุลแบ่งออกเป็น 2 วิธี ได้แก่ วิธีการแก้ปัญหาด้วยอัลกอริทึม การเรียนรู้ (Algorithmic Level Approach) โดยทำการปรับปรุงวิธีการหรือรูปแบบและเทคนิคสำหรับการเรียนรู้เพื่อให้ตัวจำแนกมีประสิทธิภาพมากที่สุด เช่น การใช้แบบจำลองหลายตัวเพื่อช่วยในการหาคำตอบ หรือ การสุ่มข้อมูลฝึกสอนเป็นหลายชุด แต่สร้างด้วยตัวแบบเดียวกันเทคนิคเดียวกันทั้งหมดเป็นต้น และอีกวิธีงานวิจัยนี้เลือกใช้คือ วิธีแก้ปัญหาระดับข้อมูล (Data Level Approach)

แก้ปัญหาระดับข้อมูลสำหรับการจัดการปัญหาในระดับข้อมูลนั้นสามารถแก้ปัญหาได้โดยใช้วิธีการสุ่มตัวอย่างซ้ำ(Resampling Techniques) ซึ่งมีวิธีการเลือกใช้ที่หลากหลายเช่น โอเวอร์แซมปลิ้ง เป็นวิธีการที่ใช้ในการสุ่มเพิ่มจำานวนข้อมูลในกลุ่มข้างน้อยให้มีจำนวนใกล้เคียงกับข้อมูลในกลุ่มข้างมากและ อันเดอร์แซมปลิ้ง (Under-sampling) เป็นวิธีในการสุ่มเลือกข้อมูลจากกลุ่มข้างมากให้ได้จำนวนที่ใกล้เคียงกับกลุ่มข้างน้อย ในงานวิจัยนี้ได้แก้ปัญหาความไม่สมดุลของคลาสในระดับข้อมูลด้วยเทคนิคการสุ่มเพิ่มตัวอย่างข้างน้อยสังเคราะห์ (Synthetic Minority Over-sampling Technique : SMOTE) [8]

เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย เป็นวิธีในการสร้างตัวอย่างข้อมูลสังเคราะห์ที่มีความคล้ายคลึงกับตัวอย่างดั้งเดิมโดยการสุ่มตัวอย่างแบบโอเวอร์แซมปลิ้งกับข้อมูล กลุ่มที่มีจำนวนน้อยให้มีจำนวนมากขึ้นจนใกล้เคียง

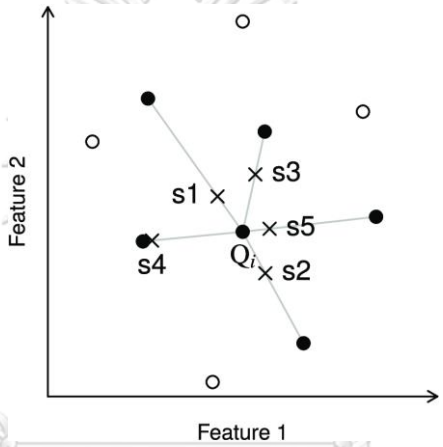
หรือเท่ากับข้อมูลกลุ่มที่มีจำนวนมาก เพื่อให้ตัวแบบจำแนกประเภทสามารถทำนายตัวอย่างใหม่เป็นกลุ่มข้อมูลที่มีจำนวนน้อยมากขึ้นโดย การทำงานของเทคนิควิธีการสุ่มเพิ่มตัวอย่างกลุ่มน้อยร่วม สามารถอธิบายได้ดังต่อไปนี้

(1) พิจารณาแต่ละตัวอย่างเฉพาะที่เป็นตัวอย่างกลุ่มน้อย มองหาตัวอย่างที่เป็นเพื่อนบ้านใกล้สุดกับตัวอย่างดังกล่าวจำนวน k ตัว

(2) สุ่มเลือกตัวอย่างที่เป็นเพื่อนบ้านใกล้สุดมาหนึ่งตัวอย่าง

(3) ลากเส้นเชื่อมตามระยะทางแบบยูคลิเดียนจากตัวอย่างที่กำลังพิจารณาไปยังตัวอย่าง เพื่อนบ้านใกล้สุดที่สุ่มมาได้

(4) สุ่มจุดที่อยู่บนเส้นเชื่อมดังกล่าวขึ้นมาเป็นจำนวนเท่ากับจำนวนตัวอย่างใหม่ที่ต้องการให้ จุดเหล่านั้นเป็นตัวอย่างสังเคราะห์ตัวอย่างใหม่



รูปภาพ 2 การสุ่มข้อมูลตัวอย่างด้วยเทคนิค SMOTE

2.4 เทคนิคความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร

เป็นเทคนิคการคัดแยกคำตามความสำคัญที่ถูกใช้ในข้อความทั้งหมดที่เก็บรวบรวมมา โดยเทคนิคนี้ใช้ในการประเมินความสำคัญของคำในข้อความทั้งหมด โดยที่ความสำคัญจะมีสัดส่วนเพิ่มตามจำนวนครั้งของคำที่เกิดขึ้นของคำนั้นๆ ในข้อความทั้งหมด เพื่อเปรียบเทียบกับสัดส่วนผกผันของ คำนั้น ๆ ในข้อความทั้งหมด หรือพูดอีกนัยหนึ่งว่า ความถี่ของค่า-ส่วนกลับความถี่ของเอกสารจะช่วยกรองคำที่มีความถี่สูง [2] ซึ่งในงานวิจัยนี้จะทำการนำคำที่ความสำคัญไปใช้ต่อในการสังเคราะห์ตัวอย่างเพิ่มเติมเพื่อแก้ปัญหาข้อมูลไม่สมดุล โดย TF-IDF สามารถคำนวณได้ดังสมการที่ (1)

$$TF_{ij} = \frac{f_{ij}}{n_j} \quad (1)$$

เมื่อ f_{ij} แทนจำนวนความถี่ของคำ i ในข้อความ j

n_j แทนจำนวนคำทั้งหมดในข้อความ

$$IDF_i = 1 + log\frac{N}{c_i} \quad (2)$$

เมื่อ N แทนจำนวนข้อความทั้งหมด

c_i แทนจำนวนข้อความที่มีคำ i ปรากฏอยู่

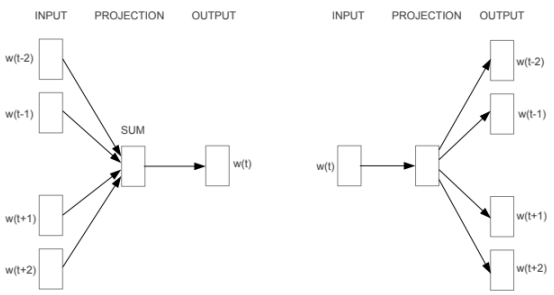
$$w_{ij} = TF_{ij} \times IDF_i \quad (3)$$

เมื่อ w_{ij} แทนจำนวนคะแนนของการคัดแยกคำตามความสำคัญ

2.5 การแปลงคำเป็นเวกเตอร์

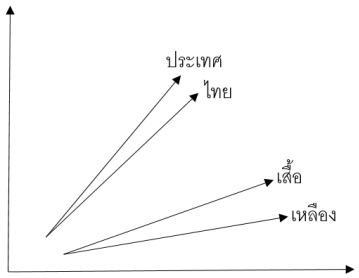
การแปลงคำเป็นเวกเตอร์ (Word2Vec) [20] เป็นแบบจำลองที่ใช้สร้างเวกเตอร์ที่ใช้ในการแทนค่าของคำให้อยู่ในรูปแบบของเวกเตอร์ที่มีความยาวจำกัด โดยอดีตรูปแบบของการเปลี่ยนแปลงคำเวกเตอร์คือการเข้ารหัสแบบวันฮ็อท (one-hot) ซึ่งจะเป็นการแทนคำด้วย เวกเตอร์ที่มีค่า 1 แค่ 1 ตำแหน่งตามความยาวของประโยค โดยไม่สนลำดับของคำที่เกิดขึ้นในประโยค ซึ่งจะส่งผลให้จำนวนคำทั้งหมดในคลังข้อความจะไม่เกิดความสัมพันธ์ใดๆ ระหว่างเวกเตอร์ของคำแต่ละคำเนื่องจากการแปลงเวกเตอร์จะเป็นการสุ่มขึ้นมา

เทคนิคในการแปลงคำเป็นเวกเตอร์รูปแบบใหม่ที่เกิดขึ้นเพื่อแก้ปัญหาในการหาความสัมพันธ์จะใช้เทคนิคโครงข่ายประสาทเทียม (Neural Network) แบบ การเข้ารหัส (Encoder-Decoder) มีเลเยอร์ (Layer) จำนวน 2 เลเยอร์ซึ่ง มีหลักการในการเปรียบเทียบเวกเตอร์ทางความหมายของคำทั้ง 2 คำ แล้วคืนค่าออกมาเป็นตัวเลขตั้งแต่ -1 ถึง 1 บ่งชี้ถึงความใกล้เคียงทางความหมายให้ค่าน้อยไปมาก หรือพูดอีกนัยหนึ่งว่าคำที่มีบริบทการปรากฏคล้ายคลึงกันควรเป็นคำที่มีความหมายคล้ายกันด้วย [11] ซึ่งวิธีการแปลงคำเป็นเวกเตอร์



รูปภาพ 3 สถาปตยากรรมของเทคนิคการแปลงคำเป็นเวกเตอร์ [20]

ซึ่งเวกเตอร์ของคำต่าง ๆ จะถูกคำนวณจากบริบทรอบข้างซึ่งการสร้างเวกเตอร์จะอาศัยคำสมมติฐานว่าคำอยู่ในบริบทเดียวกันในคำที่มีความหายใกล้เคียง โดยเมื่อมีการแปลงคำเป็นเวกเตอร์แล้วนั้น เวกเตอร์จะถูกมาคำนวณหาความคล้ายคลึงกัน[16]โดยใช้เทคนิคของการวัดค่าความคล้ายเชิงมุมซึ่งคำที่มีความหมายใกล้เคียงกัน จะมีค่าของความคล้ายเชิงมุมที่สูง



รูปภาพ 4 เวกเตอร์ที่มีความคล้ายเชิงมุม

2.7 การจำแนกแบบนาอีฟเบย์

นาอีฟเบย์ [9] เป็นวิธีการจำแนกข้อมูลประเภทหนึ่งซึ่งมีแนวคิดมาจาก ทฤษฎีของ ชองเบย์ โดยอาศัยหลักการทางสถิติมาใช้จำแนกข้อมูล วิธีการนาอีฟเบย์จะพิจารณาความน่าจะเป็นของข้อมูลที่จะเป็นคลาสใด ๆ โดยคำนวณจากความน่าจะเป็นของการเกิดแต่ละคลาส ร่วมกับความน่าจะเป็นของการเกิดแต่ละคุณลักษณะของแต่ละคลาส อย่างไรก็ตามวิธีการนาอีฟเบย์ มีสมมติฐานว่าคุณลักษณะแต่ละตัวของข้อมูลไม่ขึ้นต่อกัน ทำให้สามารถลดความซับซ้อนในการคำนวณลงไปมากแต่ยังคงมีประสิทธิภาพในการจำแนกประเภทข้อมูลจากที่ได้เห็น ความสำเร็จจากการประยุกต์ใช้วิธีการนาอีฟเบย์ในหลากหลายสาขา รวมไปถึงการจำแนกประเภทข้อความอีกด้วย การเรียนรู้แบบนาอีฟเบย์มีขั้นตอนดังต่อไปนี้

(1) รับชุดข้อมูลสอนเข้ามา คำนวณความน่าจะเป็นในการเกิดคลาสแต่ละคลาสแล้วบันทึก ค่าที่ได้เอาไว้ดังนี้ สำหรับแต่ละคลาส C_i
บันทึกค่า $P'(C_i)$ ซึ่งใช้ประมาณ $P(C_i)$

(2) คำนวณความน่าจะเป็นในการเกิดแต่ละค่าคุณลักษณะของแต่ละคลาส แล้วบันทึกค่าที่ได้ เอาไว้ ดังนี้
สำหรับแต่ละค่าคุณลักษณะ $A_j = a_j$ ของแต่ละคลาส C_i
บันทึกค่า $P'(A_j = a_j \mid C_i)$ ซึ่งใช้ประมาณ $P(A_j = a_j \mid C_i)$
เมื่อมีตัวอย่างใหม่เข้ามา สามารถใช้ความรู้ที่อยู่ในรูปแบบความน่าจะเป็นที่ได้ถูกบันทึกเอาไว้มาคำนวณเพื่อหาว่าตัวอย่างดังกล่าวมีความน่าจะเป็นสูงสุดที่จะเป็นคลาสใด แล้วจึงจำแนกข้อมูล เป็นคลาสนั้น ดังต่อไปนี้

$$C = \quad =1\ m\ P'(C_i) \prod P'(A_j = a_j \mid C_i)\ n\ j=1 \qquad (4)$$

โดยกำหนดให้

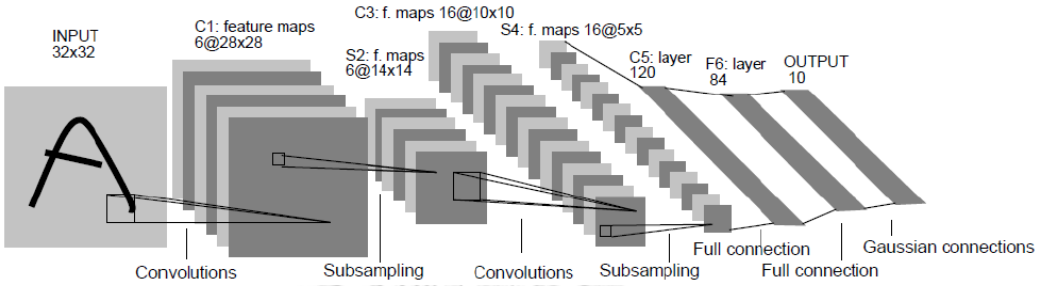
Ci; = 1, 2, ..., m หมายถึง คลาสของข้อมูล ซึ่งมีได้ตั้งแต่คลาสที่ 1 จนถึง m

A_i; = 1, 2, ..., n หมายถึง ค่าคุณลักษณะของข้อมูล ซึ่งมีได้ตั้งแต่ค่าคุณลักษณะที่ 1 จนถึง n

C หมายถึง คลาสซึ่งเป็นคำตอบของตัวอย่างใหม่

2.8 โครงข่ายประสาทแบบคอนโวลูชัน

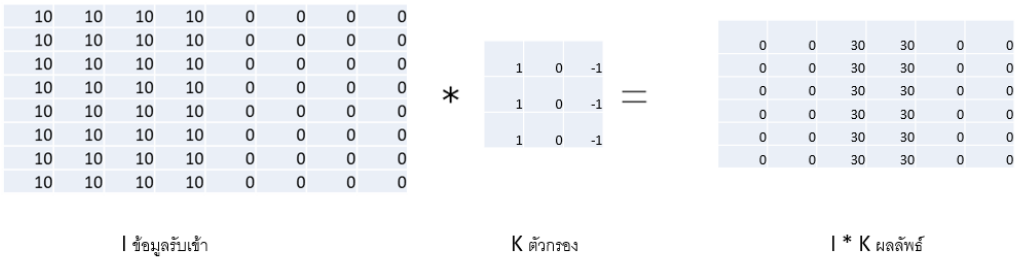
โครงข่ายประสาทแบบคอนโวลูชัน[15] เป็นนิวรอลเน็ตเวิร์กเชิงลึกที่มีจุดเริ่มต้นจากงานวิจัยด้านการจำแนกภาพตัวอักษร [22] โดยใช้ตัวกรอง (Filter) เพื่อสร้างเป็นฟีเจอร์ใหม่ (Feature Map) นำไปใช้เป็นข้อมูลรับเข้าของชั้นถัดไป โครงสร้างของนิวรอลเน็ตเวิร์กคอนโวลูชันแสดงได้ดังรูปที่ 5 ซึ่งเกิดจากนำชั้นหลาย ๆ ประเภทดังต่อไปนี้มาประกอบเข้าด้วยกัน ดังนี้



รูปภาพ 5 โครงข่ายประสาทแบบคอนโวลูชัน [21]

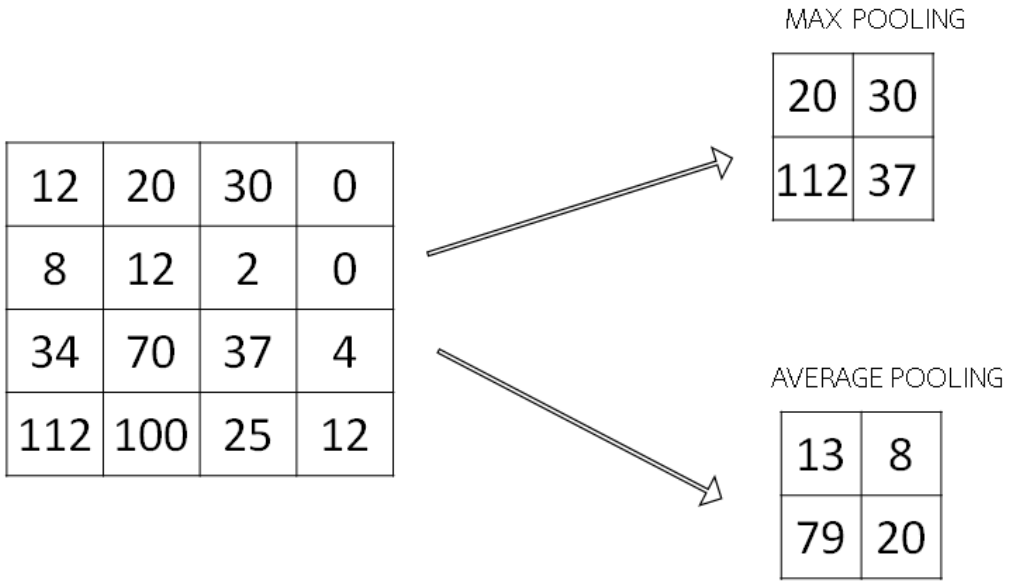
1. ชั้นคอนโวลูชัน (Convolutional Layer) เป็นชั้นที่ทำการหาฟีเจอร์จากลุ่มของข้อมูลรับเข้าที่อยู่ใกล้ ๆ กัน โดยใช้วิธีการดอทเมทริกซ์ กับตัวกรองโดยที่น้ำหนักของตัวกรองจะใช้ร่วมกันในทุก ๆ การทำคอนโวลูชันของข้อมูลรับเข้า กำหนดให้ข้อมูลรับเข้าแทนด้วยเมทริกซ์ I และตัวกรองแทนด้วยเมทริกซ์ K ซึ่งมีขนาด h × W ผลลัพธ์ของการทำคอนโวลูชัน สามารถคำนวณได้จากสมการที่ 5

$$(I * K) = \sum\sum Kij\, w\, j = 1\, .Ix + i - 1, y + j - 1\, h\, i = 1 \qquad (5)$$



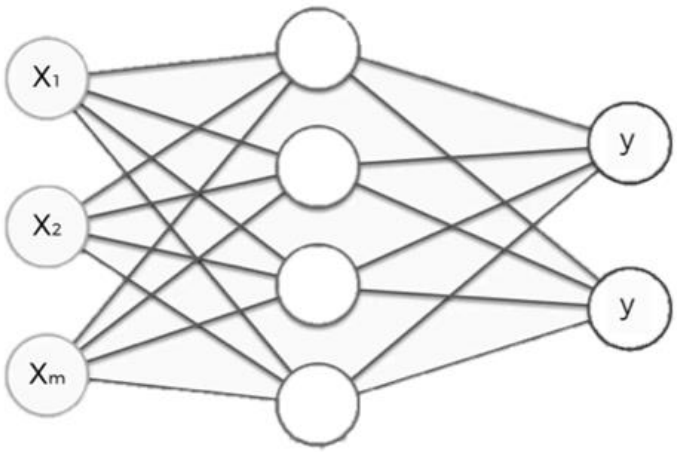
รูปภาพ 6 ตัวอย่างการทำคอนโวลูชัน

2. ชั้นการรวม (Pooling Layer) ทำหน้าที่ลดขนาดของข้อมูล เพื่อให้เหลือเฉพาะข้อมูลที่สำคัญ ๆ เท่านั้น ซึ่งมักจะนิยมนำมาต่อกับชั้นคอนโวลูชันโดยทั่วไปนิยมใช้การเลือกข้อมูลที่มีค่ามากที่สุด (Max Pooling) หรือค่าเฉลี่ย (Average Pooling) มาจากแต่ละช่วงของเมตริกซ์เพื่อสร้างเป็นเมทริกซ์ที่มีขนาดเล็ก ดังรูปที่ 7



รูปภาพ 7 ตัวอย่างการรวมชั้นของชั้นการรวม

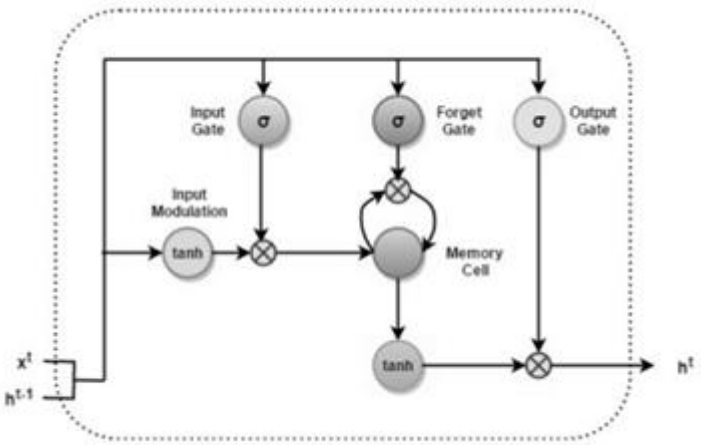
3 ชั้นการเชื่อมโยงเต็มรูปแบบ (Fully Connected Layer) หลังจากการประกอบกันของชั้นคอนโวลูชันและชั้นการรวมจำนวนหนึ่งแล้ว ในชั้นสุดท้ายของ นิวรอลเน็ตเวิร์กคอนโวลูชันจะเป็นการเชื่อมโยงเต็มรูปแบบ คือ ในชั้นนี้จะประกอบด้วยชั้นย่อย ๆ ที่มีเพอร์เซ็ปตรอนอยู่จำนวนหนึ่ง โดยที่เพอร์เซ็ปตรอนแต่ละตัว จะมีเส้นเชื่อมกันเพอร์เซ็ปตรอนทุก ตัวในชั้นก่อนหน้าและเพอร์เซ็ฟตรอน ทุกตัวในชั้นถัดไป ทำให้สามารถทำการคำนวณค่าที่การป้อนไป ข้างหน้าและการแพร่กระจายย้อนกลับได้ด้วยวิธีการปกติได้ชั้นการเชื่อมโยงเต็มรูปแบบแสดงดังรูปที่ 8



รูปภาพ 8 ชั้นการเชื่อมโยงเต็มรูปแบบ

2.9 หน่วยความจำระยะสั้นแบบยาว

หน่วยความจำระยะสั้นแบบยาว เป็นโครงข่ายประสาทเทียมแบบหนึ่งที่ถูกออกแบบมาสำหรับการประมวลผลลำดับ (sequence) โครงข่ายประสาทเทียมแบบวนกลับชนิดพิเศษ ถูกนำเสนอโดย Hochreiter และ Schmid Huber ในปี 1997 [12] ซึ่งเป็นวิธีการที่เกิดขึ้นมาเพื่อแก้ไขปัญหาของเครือข่ายประสาทเทียมแบบอาร์เอ็นเอ็น (Recurrent neural network :RNN) ที่มีปัญหาในกรณีที่ปัญหาในกรณีที่ข้อมูลมีความยาว ซึ่งในหน่วยความจำระยะสั้นแบบยาวพัฒนาต่อมาจากประสาทเทียมแบบอาร์เอ็นเอ็น ซึ่งทำงานได้ดีในการเรียนรู้แบบระยะยาว



รูปภาพ 9 หน่วยความจำระยะสั้นแบบยาว

โดยที่หน่วยความจำระยะสั้นแบบยาวจะมีองค์ประกอบ ดังนี้อินพุตเกต (Input gate) ซึ่งมีหน้าที่ควบคุมการอัปเดตข้อมูล, เอาท์พุตเกต (Output gate) ซึ่งมีหน้าที่ควบคุมการส่งออกข้อมูล และฟอร์เก็ตเกต (Forget gate) ซึ่งมีหน้าที่ควบคุมการกำจัดข้อมูลจากหน่วยก่อนหน้า

หลักการทำงานของหน่วยความจำระยะสั้นแบบยาว คือในส่วนแรกจะมีชั้นซิกมอยด์ (sigmoid Layer) ซึ่งให้ค่าออกมาระหว่าง 0 กับ 1 จะได้ค่าออกมาซึ่งจะนำไปใช้ในการคูณกับสถานะ(State) ของหน่วยก่อนหน้าเพื่อเป็นการบอกกับหน่วยความจำว่าจะใช้ค่าสถานะของหน่วยก่อนหน้าหรือลบทิ้ง

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right) \tag{6}$$

ในส่วนถัดไปเราก็จะมีการคำนวณส่วนของ tanh แล้วนำค่านั้นไปคูณกับค่าที่ได้จาก ชั้นซิกมอยด์ เพื่อตั้งค่า น้ำหนัก (Weight) ในข้อมูลใหม่

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right) \tag{7}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C) \tag{8}$$

จากนั้นเราก็นำค่าของสองส่วนจากสมาการด้านบนมารวมกัน ให้ลืมค่าเก่าบางส่วน และรับบางส่วนจากของใหม่มา จะได้ค่า Cell State

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{9}$$

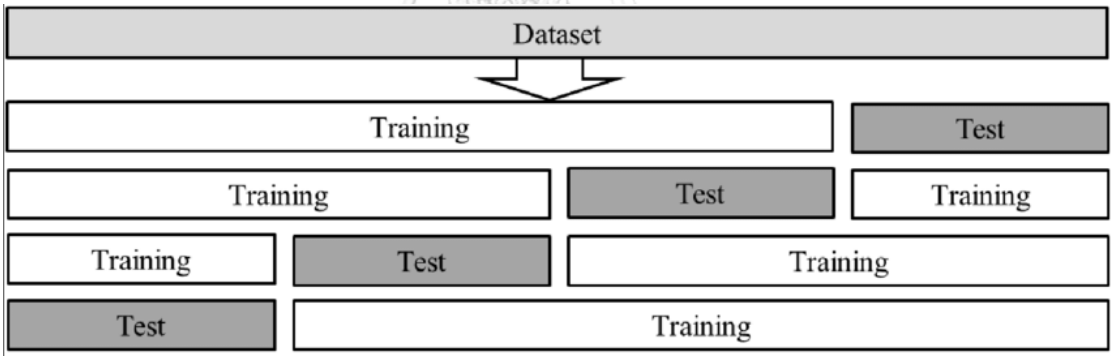
จากนั้นเราก็นำค่า Cell State มาคำนวณหาค่า tanh และนำค่าที่ได้มาคูณกับค่าจาก Sigmoid Layer เพื่อตั้ง Weight ให้อีกครั้ง และจะได้ออกมาเป็นค่า h(t) [12]

$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right) \tag{10}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{11}$$

2.10 เทคนิคไขว้ข้ามกลุ่ม (K-Fold Cross Validation)

เป็นหนึ่งในหัวข้อสำคัญเกี่ยวกับการทดสอบโมเดลการเรียนรู้ ซึ่งถูกคิดค้นโดย Dunlap K. และ Popper K. R การวัดประสิทธิภาพนี้เป็นการแบ่งข้อมูลเป็นส่วนๆเท่าๆกัน เพื่อสร้างข้อมูลเรียนรู้และข้อมูลทดสอบเพื่อทดสอบในโมเดลเพื่อคำนวณหาค่าเฉลี่ย ค่าความถูกต้อง หรือ ความผิดพลาด ก่อนที่จะนำโมเดลไปใช้ทำนายข้อมูลทดสอบ มีสองขั้นตอนหลักการแยกข้อมูลออกเป็นส่วนย่อยทำการ ชุดข้อมูลเรียนนรู้ และ ชุดข้อมูลทดสอบ จากนั้นจะทำการเรียนรู้จากชุดที่เป็นเรียนรู้และชุดทดสอบ วนไปจนครบทุกข้อมูลที่แบ่งส่วนตามรูปที่ 10



รูปภาพ 10 เทคนิคไขว้ข้ามกลุ่ม

2.11 คอนฟิวเมทริกซ์ (Confusion Matrix)

คอนฟิวเมทริกซ์ [18] เป็นการประเมินผลลัพธ์ของการทำนายของตัวแบบ เทียบกับผลลัพธ์ที่เกิดจากคน โดยรูปแบบจะเป็นตาราง 2 มิติ ระหว่างประเภทจริง (actual class) กับ ประเภทที่ถูกทำนาย (predicted class) ที่แสดงให้เห็นถึงประสิทธิภาพการทำงานของตัวแบบ ดังตารางที่ 1

ตาราง 1 ตารางคอนฟิวเมทริกซ์ Confusion Matrix

		Predicted Class	
		Yes	No
Actual Class	Yes	TP (true positive)	FN (false negative)
	No	FP (false positive)	TN (true negative)

โดยที่ TP คือ จำนวนครั้งที่ตัวแบบทำนายว่าจริงและคนบอกว่าจริง

TN คือ จำนวนครั้งที่ตัวแบบตอบว่าไม่จริงและคนบอกว่าไม่จริง

FP คือ จำนวนที่ตัวแบบบอกว่าจริงแต่คนบอกไม่จริง

FN คือ จำนวนที่ตัวแบบบอกว่าไม่จริงแต่คนบอกว่าจริง

โดยค่าของ TP, TN, FP และ FN จะถูกนำไปคำนวณเพื่อสร้างเป็นเกณฑ์ต่างๆ ในการวัดค่าความถูกต้องในการจำแนกประเภท ได้แก่

1. ค่าความถูกต้อง (Accuracy) เป็นเกณฑ์วัดความถูกต้องโดยรวมในการจำแนกประเภท เพื่อพิจราณาการทำนายของตัวแบบ โดยนับจำนวนที่ตัวแบบทายถูกทั้งหมดเทียบกำจำนวนข้อมูลทั้งหมด ดังสมาการที่ 12

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (12)$$

2. ค่าความแม่นยำ(precision) เป็นเกณฑ์การวัดความที่ถูกต้องที่จะสนใจในผลลัพธ์ที่ทำนายถูกว่ามีกี่เปอร์เซนต์ที่ทำนายถูก

$$\text{Precision} = \frac{TP}{TP+FP} \quad (13)$$

3. ค่าความถูกต้องประเภทที่สนใจ (recall) เป็นเกณฑ์การวัดความถูกต้องในการจำแนกประเภท โดยจะพิจารณาว่าหากความจริงเป็นประเภทที่สนใจแล้วตัวแบบจะทำนายถูกกี่เปอร์เซ็นต์

$$\text{Recall} = \frac{TP}{TP+FN} \quad (14)$$

4. ค่าเอฟวัน (F-1 measure) เป็นการวัดความถูกต้องในการจำแนกประเภทที่ใช้ค่าเฉลี่ยฮาร์โมนิค ระหว่างค่าความแม่นยำ กับ ค่าความถูกต้อง ตามสมาการ

$$F_1 \text{ Measure} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (15)$$

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

บทที่ 3 แนวคิดและวิธีการดำเนินงาน

ในส่วนนี้ของงานวิจัยจะกล่าวถึงแนวทางการดำเนินงานโดยนำทฤษฎีที่เกี่ยวข้องจากบทที่ 2 มาประยุกต์ใช้กับการสังเคราะห์ตัวอย่างเพิ่มเพื่อการจำแนกข้อความภาษาไทยที่มีความหมายเป็นประทุษวาจาให้มีประสิทธิภาพมากขึ้น โดยจะแบ่งเป็นหัวข้อ ดังนี้ 1. เก็บรวบรวมข้อมูลจากทวิตเตอร์ 2.การทำความสะอาดข้อมูล3.การติดฉลากข้อมูล 3.การตัดคำ 4.การเปลี่ยนคำเป็นเวกเตอร์ 5.การจำแนกข้อความประทุษวาจาในข้อความภาษาไทย 3.6การจำแนกข้อความประทุษวาจา

3.1 เก็บรวบรวมข้อมูล

ในส่วนนี้ทางผู้วิจัยจะทำการเก็บรวบรวมข้อมูลจากทวิตเตอร์ด้วยวิธีการค้นหาทวิตเตอร์ด้วยคำที่มีความหมายเชิงวาจาสร้างความเกลียดชังในรูปแบบต่างๆ คือ สลิ่ม ,ส้มเน่า ,ขี้ข้าเผด็จการ, มารศาสนา, เกลียดอิสลาม, ตลาดล่าง, กระเทย, ลัทธิจานบิน เป็นคำตั้งต้นในการค้นหาข้อความจากทวิตเตอร์ จากนั้นทำการนำข้อความที่ได้จากการค้นหาของแต่ละคำมาทำตัดคำเพื่อหาคำที่มีความถี่เยอะที่สุดในชุดการค้นหานั้นๆ มาจัดอันดับตั้งแต่ 1 ถึง 3 จากนั้น จะนำคำจากอันดับ 1 ถึง 3 ไปใช้ในครั้งถัดๆไป ซ้ำไปเรื่อย ๆ จนกระทั่งได้อันดับของค่าความถี่คำที่คงที่ ทางผู้วิจัยจะนำคำเหล่านั้นมาใช้ค้นหาคำในทวิตเตอร์โดยจะทำการเก็บข้อความจากคำที่นำมาใช้เท่านั้น ตามขั้นตอนวิธีที่1 (algorithm) ดังต่อไปนี้

ALGORITHM1 : Collect Tweets

***INPUT**: word(เช่น สลิ่ม ,ส้มเน่า ,ขี้ข้าเผด็จการ,มารศาสนา,เกลียดอิสลาม,ตลาดล่าง,กระเทย ,ลัทธิจานบิน)*

***OUTPUT:** Final_sentence*

```
Assign list of sentences // ประโยคที่ค้นหา
Assign list of token // คำที่ถูกตัด
Assign list of word_rank // คำตามลำดับ
Assign list of final_word // คำที่ใช้ค้นหาสุดท้าย
Assign list of final_sentence // ข้อมูลที่จะเก็บรวบรวมไปใช้
function collect_tweets(input_word):
for word in input_word:
        Sentence = Collect_Tweets_API(word) // ค้นหาทวิตด้วย api
        token = Tokenization(Sentence) // ตัดคำจากประโยค
return  token
end
collect_tweets(input_word)
word_rank = ranking_term_frequency(token,sort=descending)// จัดอันดับค่าความถี่ของคำ เรียงจากมากไปหาน้อย
word_rank = word_rank[1:3] // pick top 3 of word
while:
```

```
16          if  word_rank != word_rank:
17              collect_tweets(word_rank)
18          else :  break
19          end if
20      for word in final_word:
21          final_sentence = collect_Tweets_API(word)
22      end
```

หลังจากที่ได้คำที่ใช้ในการค้นหาแล้วจะนำคำเหล่านั้นไปทำการค้นหาข้อมูลด้วย ทวิตเตอร์เอพีไอ ได้ดังตัวอย่างตารางด้านล่าง

ตาราง 2 แสดงตัวอย่างข้อความที่เก็บรวบรวมมา

ถ้าไม่ช่วยกันตอนนี้ ก็ไม่รู้จะมีโอกาศแบบนี้อีกมั้ย นับวันมันยิ่งโหดร้ายเหลือเกิน
บ้าน Rapperline9899 ขออนุญาตใช้พื้นที่ในการเป็นกระบอกเสียง และขอเป็นส่วนหนึ่งในการเรียกร้องประชาธิปไตย การแสดงความ
เวลาตี5ครึ่ง กลุ่มธรรมศาสตร์เพื่อเสรีประชาธิปไตยขอเชิญชวนนักศึกษามหาวิทยาลัยธรรมศาสตร์ที่สะดวกมารวมตัวกันที่โดมบริหารเพื่อ
อะไรมันจะเกิดมันต้องเกิดการเปลี่ยนแปลงธรรมดาของโลก..ช่วยหยุดสงครามกลางเมือง
ผู้มาก่อนกาล กุเคยด่าเค้าไปได้ไง แต่เค้ารู้ได้ไงวะ มาก่อนกาลสัด

3.2 การทำความสะอาด

จากข้อมูลที่เก็บรวบรวมมาข้อความจากทวิตเตอร์เป็นข้อมูลที่ประกอบด้วยอักขระ ภาษาอังกฤษ สัญลักษณ์พิเศษต่างๆเช่น : , ‘’, ; เป็นต้น ซึ่งไม่สามารถนำไปใช้ได้ในงานวิจัยนี้เนื่องจากส่งผลกับตัวแบบในการเรียนรู้ จึงต้องทำความสะอาดก่อน และข้อความบางส่วนอาจจะมีการซ้ำกัน หรือ อิโมติคอนติดและตัวอักขระซ้ำเช่น คำว่า “แงงงง” ที่มีการใส่ ง เพิ่มขึ้นมาส่งผลต่อขั้นตอนการตัดคำ และ ความสามารถให้ความจำแนกข้อความของโมเดลลดลง ดังนั้นข้อมูลจะถูกทำการลบชุดที่ข้อมูลที่ซ้ำออกและลบสัญลักษณ์พิเศษออกเนื่องจากไม่ได้ให้ข้อมูลใดๆ เกี่ยวกับลักษณะของข้อความคือ เครื่องหมายวรรคตอน ตัวเลข อักขระพิเศษ และอิโมจิ จะถูกลบออก ท้ายสุดจะนำเอาอักษรภาษาอังกฤษออกด้วยเพราะในงานวิจัยนี้จะทำการสังเคราะห์และตรวจจับข้อความที่เป็นภาษาไทยเท่านั้น

ตาราง 3 ตัวอย่างข้อความที่ทำความสะอาดแล้ว

ข้อความ	ข้อความที่ผ่านการทำความสะอาดแล้ว
 ถ้าไม่ช่วยกันตอนนี้ ก็ไม่รู้จะมีโอกาศแบบนี้อีกมั้ย นับวันมันยิ่งโหดร้ายเหลือเกิน	ถ้าไม่ช่วยกันตอนนี้ ก็ไม่รู้จะมีโอกาศแบบนี้อีกมั้ยนับวันมันยิ่งโหดร้ายเหลือเกิน
 บ้าน Rapperline9899 ขออนุญาตใช้พื้นที่ในการเป็นกระบอกเสียง และขอเป็นส่วนหนึ่งในการเรียกร้องประชาธิปไตย การแสดงความ	บ้านขออนุญาตใช้พื้นที่ในการเป็นกระบอกเสียง และขอเป็นส่วนหนึ่งในการเรียกร้องประชาธิปไตย การแสดงความ

3.3 การติดฉลากข้อมูล

ขั้นตอนนี้จะเป็นการนำข้อมูลที่ทำความสะอาดเรียบร้อยแล้วจะถูกนำส่งข้อมูลทั้งหมดไปยังนักข่าวสามคนที่ใช้ทวิตเตอร์เป็นประจำในการทำงาน เพื่อติดป้ายกำกับให้กับข้อมูลที่ซึ่งคลาส ‘No’ มีความหมายแสดงความเกลียดชัง และ คลาส ‘Hate’ แสดงถึงว่าข้อความนี้เป็นข้อความที่แสดงความเกลียดชังในวิจัยนี้จะให้นักข่าวทั้ง 3 คนทำการเลือกข้อมูลที่ส่งให้แยกคนละ 1 ชุด จากนั้นจะส่งข้อมูลให้แต่ละคน คนละ1ชุดเพื่อติดฉลากพร้อมทั้งบอกเหตุผลว่าข้อความนี้เป็นประทุษวาจาหรือไม่ ในขั้นสุดท้ายจะทำการเก็บรวบรวมข้อมูลทั้งหมดมาแล้วดูคะแนนเสียงในการเลือกของผู้เชี่ยวชาญแต่ละคน คือถ้ามีคนโหวตให้ข้อความนี้เป็นประทุษวาจามากกว่า 2 คนขึ้นไป ข้อความนั้นจะถูกใส่คลาส ‘Hate’ ถ้าไม่จะเป็นคลาส ‘No’ โดยตัวอย่างข้อความที่จะใช้ในการทดสอบตามตารางที่ 4

ตาราง 4 ตัวอย่างข้อความที่ติดฉลากแล้ว

ข้อความ	คลาส	เหตุผล
เกลียดพวกที่อ้างว่ารักสถาบันเพื่อเอาไว้เป็นฉากบังหน้าใช้ทำร้ายฆ่าหรือโบ้ยความผิดให้ผู้อื่นที่เห็นต่าง	Hate	กล่าวหาผู้รักสถาบัน ว่าใช้สถาบันเป็นฉากบังหน้าให้ทำร้ายผู้เห็นต่าง
คือแบบ เห็นแบบนี้แล้วไม่แปลกใจเลย ทำไมประเทศชาติถึงไม่เจริญสักที เพราะมีผู้ใหญ่แบบนี้ไง แก่แล้วแก่เลย ขนาดรูปตัดต่อลวกๆ	Hate	เพราะมีผู้ใหญ่แบบนี้ไง แก่แล้วแก่เลย" ประเทศชาติเจริญ/ไม่เจริญ อาจไม่ใช่เพราะผู้ใหญ่อย่างเดียว
มาตามนัดมายด์ ภัสราวลี หนึ่งในแกนนำกลุ่มคณะราษฎร 2563 เดินทางมาที่ สน.ทุ่งมหาเมฆ เพื่อมารับทราบข้อกล่าวหา ม.112	No	เป็นข้อความเล่าถึงสถานการณ์ไม่ได้การโจมตีใคร
ขณะนี้ แกนนำผู้ถูกหมายจับพร้อมอานนท์และประสิทธิ์ อยู่ระหว่างการสอบปากคำของเจ้าหน้าที่ตำรวจ มีอาจารย์และทนายติดตามอย่างใกล้ชิด	No	เป็นข้อความประโยคบอกเล่าถึงสถานการณ์

3.3 การตัดคำ

ข้อมูลที่ผ่านการติดฉลาดเรียบร้อยแล้ว จะถูกนำมาตัดให้เหลือเป็นคำโดยการใช้ไลบารี่ ‘Newmm’ เป็นสิ่งที่ใช้ตัดและจากการที่ลองตัดคำมาแล้วนั้น จะมีศัพท์บางอย่างที่ไม่สามารถตัดได้ในไลบารี่นี้ สามารถทำการเพิ่มคำศัพท์เข้าไปได้ ตัวอย่างเล่นคำว่าตลาดล่าง ถ้าปกติจะ ตัดออกมาเป็น ‘ตลาด’,‘ล่าง’ แต่ในงานวิจัยนี้ต้องผลลัพธ์ของการตัดคำเป็นคำว่า ‘ตลาดล่าง’ จึงได้มีการเพิ่มคำศัพท์ใหม่เข้าไป จากนั้นจะทำการเปลี่ยนข้อความให้เหลือเพียงคำๆ ซึ่งผลลัพธ์ของการตัดคำแล้วจะได้ ลักษณะตามตารางที่ 5 นี้

ตาราง 5 แสดงตัวอย่างข้อความภาษาไทยที่ถูกตัดคำ

ประโยคเดิม	ตัดคำแล้ว
พวกขยะสังคมวันสร้างแต่ความเดือดร้อนรำคาญ	['พวก', 'ขยะสังคม', 'วัน', 'สร้าง', 'แต่', 'ความเดือดร้อน', 'รำคาญ'],
ม็อบก่อกวนสร้างความเดือดร้อนให้สังคมไทยรายวัน	['ม็อบ','ก่อกวน', 'สร้าง', 'ความเดือดร้อน', 'ให้', 'สังคม', 'ไทย', 'รายวัน']

3.4 การเปลี่ยนคำเป็นเวกเตอร์

ในขั้นตอนนี้จะนำข้อมูลที่ถูกตัดเป็นคำๆ แล้วนำมาแปลงเป็นเวกเตอร์ซึ่งมีขนาด 300 มิติ ซึ่งในขั้นตอนนี้จะใช้ไลบารี่ ‘GENSIM’ [19] ในโปรแกรมภาษาไพธอนเป็นเครื่องมือในการแปลงคำทั้งหมดจะถูกกำหนดด้วยเวกเตอร์ โดยจากการตัดคำแล้วนำมาสร้างเวกเตอร์นั้น ประกอบด้วยคำทั้งหมด 5729 คำซึ่งแต่ละคำประกอบด้วยทั้งหมด 300 มิติ ได้ผลลัพธ์เป็นตารางที่ 6

ตาราง 6 ตัวอย่างคำที่แปลงเป็นเวกเตอร์

คำ	0	1	2	3	...	300
นับวัน	-0.001500436	0.00151545	0.000346311	-0.000874488	...	-0.00099
พาดหัว	0.001150172	-0.000589385	-0.000500596	-0.001136635	...	0.000133
ข่าวสด	0.000229587	0.000423785	0.001442365	0.001493414	...	0.040018
เปรี้ยว	-0.000802141	-0.000114384	-0.001124568	0.001153794	...	0.001123

3.5 การสังเคราะห์ข้อความเพิ่มเติมในกลุ่มน้อย

จากข้อมูลทั้งหมดที่เก็บรวมมาแล้ว ทำให้สังเกตเห็นว่าข้อความมีความไม่สมดุลของข้อมูล โดยที่ข้อมูลที่มีคลาสเป็น Hate มีจำนวน 87 ข้อความจากข้อมูลทั้งหมด ซึ่งในขั้นตอนนี้จะทำการใช้ เทคนิคทั้งหมด 3 วิธีคือ (1) เทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อย (2) การใช้ความถี่ของคำ เพื่อเพิ่มจำนวนให้มีความสมดุลกัน (3) การใช้ค่าความคล้ายเชิงมุม

3.5.1 ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคการสร้างข้อความจากความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร

ในขั้นตอนของการสร้างข้อมูลเพิ่มเติมนี้ ทางผู้วิจัยจะทำการนำคำทั้งหมดมาทำการเรียงลำดับตามค่าความถี่ของค่า-ส่วนกลับความถี่ของเอกสารจากมากไปหาน้อย แล้วทำการสุ่มคำในช่วงความถี่สูงที่สุดในช่วงตั้งแต่ 10 เปอร์เซ็นต์แรกของอันดับทั้งหมดเป็นช่วงแรกในการสุ่มคำที่จะนำมาใช้ในการสร้างตัวอย่างเพิ่มโดยการค้นหาข้อมูลที่มีคำที่สุ่มขึ้นมาแล้วเพิ่มคำนั้นเข้าไปเพื่อสร้างเป็นชุดข้อมูลใหม่ จากนั้นจะทำซ้ำโดยเพิ่มขึ้นทีละ 10 เปอร์เซ็นต์ ช่วงจาก 10 เป็น 20 30 40 ไปจนกระทั่งถึงการใช้คำทั้งหมดโดยวิธีการสร้างข้อมูลใหม่ตามอัลกอริทึมที่ 2

ALGORITHM 2 : Generate new data by TF-IDF

INPUT: sentence
OUTPUT: output_data_Tf-IDF
1 *Assign list of new_data*
2 *Assign list of word_ranking*
3 *word_ranking = gen_top_tf-idf(sentence)*
4 *new_word = word_ranking [1:10]*
5 *random_word = random(new_word)*
6 *output_data ← Add random_word to sentence*
7 *end*

ตาราง 7 อันดับของค่าความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร ในชุดข้อมูล

คำ	ค่า TF-IDF
คน	0.064182
เสื้อ	0.024845
ทหาร	0.022257
แดง	0.021739
สื่อ	0.020704
เมือง	0.020186
ไพร่	0.019669
ไทย	0.013975
สลิม	0.010352

โดยวิธีการสร้างข้อมูลเพิ่มคือการเอาคำที่อยู่สุ่มขึ้นมาจากในช่วงนั้น ๆ มาทำการสร้างข้อมูลใหม่ด้วยกัน การเพิ่มคำเข้าไปในประโยคด้วยการต่อท้ายของคำนั้น ๆ จะได้เป็นชุดข้อมูลใหม่ตามตัวอย่าง ในตารางที่ 8

ตาราง 8 ตัวอย่างของข้อมูลที่เกิดจาการสร้างของความถี่คำด้วยการเพิ่มคำเข้าไปในประโยค

คำที่สุ่มขึนมา	ประโยคเดิม	ประโยคสร้างใหม่
ไพร่	ประเทศพังเพราะอีไพร่และพวกไพร่	ประเทศพังเพราะอีไพร่ไพร่และพวกไพร่ไพร่

ตลาดล่าง	พวกนรเตรียมทหารก็ยังตลาดล่างอยู่ดี	พวกนรเตรียมทหารก็ยังตลาดล่างตลาดล่างอยู่ดี
ทหาร	การค้าทาสยุคใหม่ก็คืออยู่ในรูปแบบการเกณฑ์ทหาร	การค้าทาสยุคใหม่ก็คืออยู่ในรูปแบบการเกณฑ์ทหารทหาร
ลัทธิจานบิน	เบื่อลัทธิจานบินใช้การสวดมนต์มาเป็นกำบังปากสวดแต่ใจไม่สงบก็ไม่มีประโยชน์ป่าวว	เบื่อลัทธิจานบินใช้การสวดมนต์มาเป็นกำบังปากสวดแต่ใจไม่สงบก็ไม่มีประโยชน์ป่าวว

ในการเพิ่มคำเข้าไปในคำที่มีความถี่สูงเพิ่มเข้าไปในชุดตัวอย่างที่เป็นกลุ่มน้อยนั้นทำให้ค่าของความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร ข้อความมีค่าเพิ่มขึ้นจากเดิมส่งผลให้การจำแนกข้อความของตัวแบบจะมีประสิทธิภาพเพิ่มขึ้นเนื่องจากคำที่เพิ่มขึ้นจากการสร้างใหม่ทำให้เกิดแพทเทิร์นในการตรวจจับข้อความได้ดีขึ้น

3.5.2 การสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

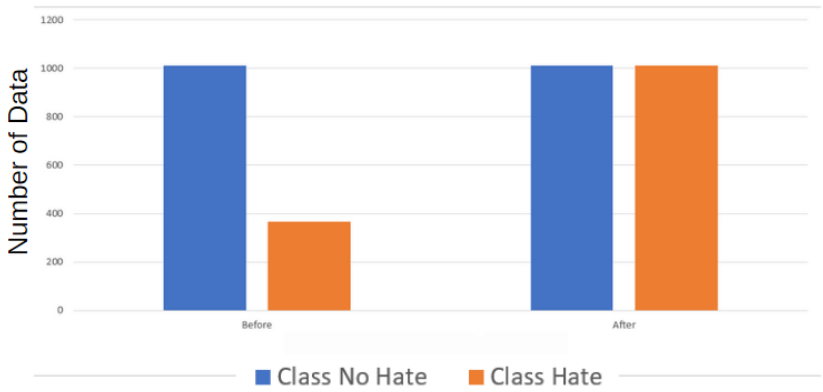
หลังจากในขั้นตอนวิธีการแทนค่าคำด้วยเวกเตอร์ของแต่ละคำแล้ว จากนั้นจะนำข้อมูลที่เป็นกลุ่มน้อยไปสร้างตัวอย่างเพิ่มด้วย เทคนิคการสุ่มตัวอย่างส่วนน้อยด้วยการสังเคราะห์ เพื่อสร้างตัวอย่างเพิ่มเติมขึ้นมาซึ่งขั้นตอนในการทำจะมีขั้นตอนดังนี้

(1) สุ่มจุดตัวอย่างขึ้นมา 5 จุดจากนั้นมองหาจุดของข้อมูลกลุ่มน้อยที่เป็นเพื่อนบ้านใกล้สุดกับจุดตัวอย่างที่สุ่มขึ้นมา

(2) สุ่มเลือกตัวอย่างที่เป็นเพื่อนบ้านใกล้สุดมาหนึ่งตัวอย่าง

(3) ลากเส้นเชื่อมตามระยะทางแบบยูคลิเดียนจากจุดตัวอย่างที่กำลังพิจารณาไปยังตัวอย่างเพื่อนบ้านใกล้สุดที่สุ่มมาได้

(4) สุ่มจุดที่อยู่บนเส้นเชื่อมดังกล่าวขึ้นมาเป็นจำนวนเท่ากับจำนวนตัวอย่างใหม่ที่ต้องการให้ จุดเหล่านั้นเป็นตัวอย่างสังเคราะห์ตัวอย่างใหม่



รูปภาพ 11 ผลลัพธ์ของจำนวนของมูลที่ผ่านการ smote

3.5.3 การสร้างตัวอย่างส่วนน้อยเพิ่มด้วยเทคนิคความคล้ายเชิงมุม

ในการสร้างตัวอย่างเพิ่มด้วยวิธีนี้ จะเป็นการนำคำที่อยู่ช่วงของค่าความถี่ของคำที่ถูกเรียงลำดับความถี่นำมาเป็นคำตั้งต้น จากนั้นจะทำการหาคำที่คล้ายกันในเชิงมุมโดยจะจัดอันดับจากคำที่มีค่าความคล้ายเชิงมุมมากไปหาคำที่น้อยกว่า ซึ่งจะใช้ผลลัพธ์ในช่วง 10 คำแรกที่มีค่ามาก จะเป็นชุดข้อมูลในการสร้างข้อมูลเพิ่ม จากนั้นทำการนำคำมาต่อกันโดยวิธีการสุ่มจนครบ 10 คำเพื่อสร้างเป็นข้อมูลชุดใหม่ โดยขั้นตอนวิธีการในการสร้างคำคล้ายคลึงที่เกิดซ้ำและนำคำดังกล่าวมาเพิ่มเป็นคุณลักษณะใหม่เป็นไปตามขั้นตอนวิธีดังต่อไปนี้

ALGORITHM3 : Generate new data by similar words

***INPUT**: word // คำที่ต้องการจะหาคำคล้ายเชิงมุม*

***OUTPUT:** output_data //ข้อมูลใหม่ที่เกิดจากความคล้ายเชิงมุม*

```
Assign list of new_data
Assign list of sim_word
sim_word = gen_top_sim_word(word)// หาค่าความคล้ายเชิงมุม
new_data = sim_words[1:10] // เลือกคำที่มีค่าสูงสุด 10 คำแรก
for word in range(10):
    random_word = random(new_data) // สุ่มตัวอย่างคำจากข้อมูล new data
    if random_word not in output_data:
        output_data ← random_word //ต่อคำที่สุ่มขึ้นมาเพิ่มเข้าไป
    end if
end
```

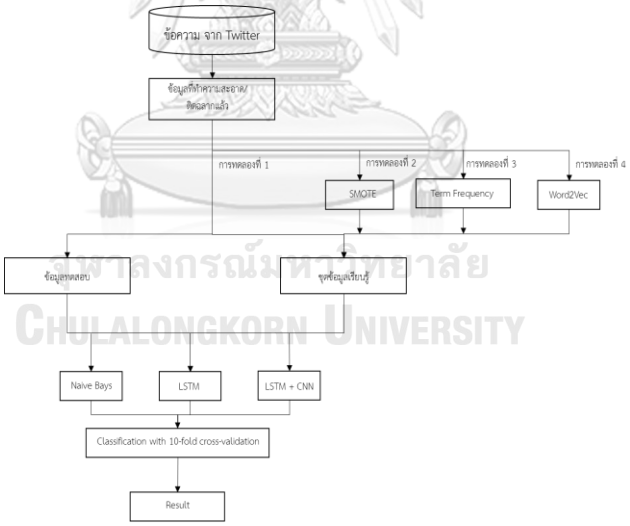
ตาราง 9 ตัวอย่างข้อมูลที่มาจากความคล้ายเชิงมุม

คำ	คำที่คล้ายคลึง	ประโยคสร้างใหม่
สถาบัน	ประเทศ,ไทย,ทหาร,ชุมนุม,บิ๊ก,กษัตริย์,เหลือง,ตำรวจ,ตุลา,เสื้อ	ตุลากษัตริย์เหลืองประเทศตำรวจเสื้อบิ๊กชุมนุม
ทหาร	คน,ประชาชน,ไทย,สื่อ,ตำรวจ,ชุมนุม,ประเทศ,บิ๊ก,ข่าว,ใหม่	บิ๊กประชาชนใหม่คนไทยสื่อตำรวจชุมนุมข่าว
ประเทศ	ไทย,ตำรวจ,ทหาร,ประชาชน,ชุมนุม,ชนะ,สถาบัน,กษัตริย์,แดง,ประกาศ	ประเทศทหารตำรวจประชาชนไทยชนะสถาบันประกาศชุมนุมกษัตริย์แดง

3.6 การจำแนกข้อความในการทดลอง

ในงานวิจัยนี้ในส่วนของการเลือกแบบจำลองสำหรับการจำแนกข้อความภาษาธรรมชาติเพื่อใช้ในงานวิจัยที่เกี่ยวกับการประมวลผลข้อความที่เป็นประทุษวาจา โดยผู้วิจัยเปรียบเทียบเทคนิคของการสร้างตัวอย่างเพิ่มเติมระหว่างการสร้างข้อมูลด้วยข้อความจากความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร การสุ่มตัวอย่างเพิ่มส่วนน้อยด้วยการสังเคราะห์ และ การใช้ความคล้ายคลึงของคำในการสร้างข้อมูลเพิ่มเติม

โดยในงานวิจัยนี้เลือกใช้แบบจำลองทั้งหมด 3 รูปแบบคือ 1.นาอีฟเบย์ซึ่งเป็นแบบจำลองที่ได้รับความนิยมในการจำแนกข้อความ เพราะใช้งานง่ายอาศัยหลักการคือการคำนวณความน่าจะเป็นที่จะเกิดเหตุการณ์บางอย่าง โดยพิจารณาเหตุการณ์ก่อนหน้า ข้อดีคือความรวดเร็วในการสอนการเรียนรู้ 2. LSTM [15] เพราะลำดับของแต่ละคำมีความสำคัญ ส่งผลต่อความหมายของประโยคซึ่งเป็นแบบจำลองที่เหมาะกับข้อมูลที่เป็นลำดับ (Sequence Data) โดยการใช้ข้อมูลส่งออก (Output) ของสถานะก่อนหน้ากลับมาเป็นข้อมูลนำเข้า (Input) ของสถานะปัจจุบันในชั้นข้อมูลซ่อน (Hidden Layer) ร่วมกับข้อมูลขาเข้าในสถานะปัจจุบัน ทำให้เข้าใจความหมายของประโยคได้ดีกว่าเพราะได้อ่านข้อมูลจากคำก่อนหน้ามาแล้ว และ รูปแบบสุดท้ายที่นำมาใช้จะเป็นในส่วนการร่วมกันระหว่าง หน่วยความจำสั้นแบบยาว และ โครงข่ายประสาทคอนโวลูชัน ที่ทำงานร่วมกันเป็นรูปแบบสุดท้ายที่ใช้ในการทดลอง ซึ่งรูปแบบของการทดลองจะเป็นไปตามภาพที่ 12



รูปภาพ 12แผนภาพขั้นตอนการทดลอง

บทที่ 4 ผลการทดลอง

ในบทนี้ทางผู้วิจัยจะแสดงให้วิธีการทดลองและผลลัพธ์ของการทดลองการจำแนกข้อความภาษาไทยที่มีความหมายในเชิงประทุษวาจาในกรณีที่ ข้อมูลที่นำมามีความไม่สมดุลกัน เนื่องจากในงานวิจัยนี้สนใจในวิธีการเพิ่มข้อมูลเพื่อแก้ปัญหาข้อมูลไม่สมดุล ในการทำวิจัยต้องการเพิ่มประสิทธิภาพในการจำแนกข้อมูลกลุ่มน้อยให้ดีขึ้น จึงได้ทำการทดลองเปรียบเทียบจะเทียบเฉพาะค่าความแม่นยำ ค่าความเที่ยง และ ค่าเอฟเอฟ ของค่าข้อมูลที่เป็นประทุษวาจาเท่านั้น ซึ่งจะได้ผลลัพธ์แบ่งตามรูปแบบการทดลองของรูปแบบทั้ง 3 ดังนี้ 1.รูปแบบ นาอีฟเบย์ 2. รูปแบบหน่วยความจำระยะสั้นแบบยาว 3.รูปแบบหน่วยความจำสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน

4.1 รูปแบบนาอีฟเบย์

ในการทดลองนี้จะทำการทดลองเพื่อแสดงให้เห็นถึงปัญหาของความแม่นยำในตัวแบบ จากการเรียนรู้ที่เกิดจากข้อมูลที่ไม่สมดุล ข้อมูลที่มาจากเทคนิคการสังเคราะห์กลุ่มน้อย เทคนิคการสร้างข้อความจากความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร และเทคนิคการสร้างคำจากค่าความคล้ายเชิงมุม แล้วนำไปป้อนข้อมูลเข้าสู่รูปแบบนาอีฟเบย์ ซึ่งได้ผลลัพธ์ประเมินตัวแบบได้ผลลัพธ์ของแต่ละตัวแบบตามตารางต่อไปนี้

ตาราง 10 แสดงค่าความถูกต้องค่าความแม่น ค่าระลึกและค่าเอฟเมเซอร์ ของข้อมูลไม่สมดุล

นาอีฟเบย์				
	ความถูกต้อง	ความแม่นยำ	ค่าระลึก	เอฟวัน
ไม่สมดุล	0.559	0.683	0.382	0.49
SMOTE	0.58	0.5	0.13	0.2
TF-IDF	0.67	0.626	0.89	0.89
Word2vec	0.606	0.552	0.64	0.484

4.2 รูปแบบหน่วยความจำระยะสั้นแบบยาว

ในการทดลองนี้จะทำการทดลองเพื่อแสดงให้เห็นถึงปัญหาของความแม่นยำในตัวแบบ จากการเรียนรู้ที่เกิดจากข้อมูลที่ไม่สมดุล ข้อมูลที่มาจากเทคนิคการสังเคราะห์กลุ่มน้อย เทคนิคการสร้างข้อความจากความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร และเทคนิคการสร้างคำจากค่าความคล้ายเชิงมุม แล้วนำไปป้อนข้อมูลเข้าสู่รูปแบบหน่วยความจำสั้นแบบยาว ซึ่งได้ผลลัพธ์ประเมินตัวแบบได้ผลลัพธ์ของแต่ละตัวแบบตามตารางต่อไปนี้

ตาราง 11 แสดงค่าความถูกต้องค่าความแม่น ค่าระลึกและค่าเอฟเมเซอร์*ของข้อมูลไม่สมดุล*

หน่วยความจำระยะสั้นแบบยาว				
	ความถูกต้อง	ความแม่นยำ	ค่าระลึก	เอฟวัน
ไม่สมดุล	0.61	0.687	0.343	0.458
SMOTE	0.55	0.541	0.406	0.464
TF	0.7	0.427	0.676	0.523
Word2vec	0.78	0.7	0.98	0.82

4.3 รูปแบบหน่วยความจำสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน

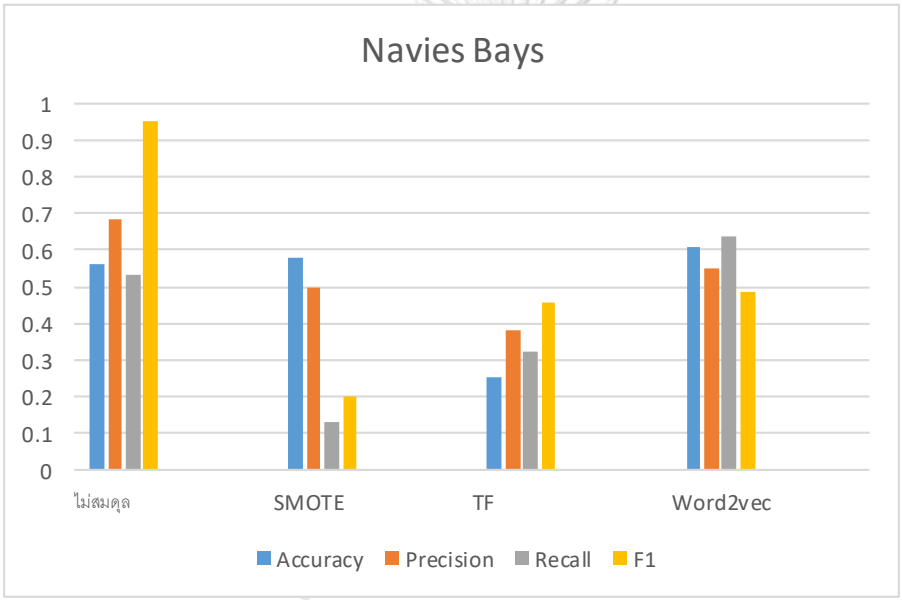
ในการทดลองนี้จะทำการทดลองเพื่อแสดงให้เห็นถึงปัญหาของความแม่นยำในตัวแบบ จากการเรียนรู้ที่เกิดจากข้อมูลที่ไม่สมดุล ข้อมูลที่มาจากเทคนิคการสังเคราะห์กลุ่มน้อย เทคนิคการสร้างข้อความจากความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร และเทคนิคการสร้างคำจากค่าความคล้ายเชิงมุม แล้วนำไปป้อนข้อมูลเข้ารูปแบบหน่วยความจำสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน ได้ผลลัพธ์ของแต่ละตัวแบบตามตารางต่อไปนี้

ตาราง 12 แสดงค่าความถูกต้องค่าความแม่น ค่าระลึก และค่าเอฟเมเซอร์ ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน

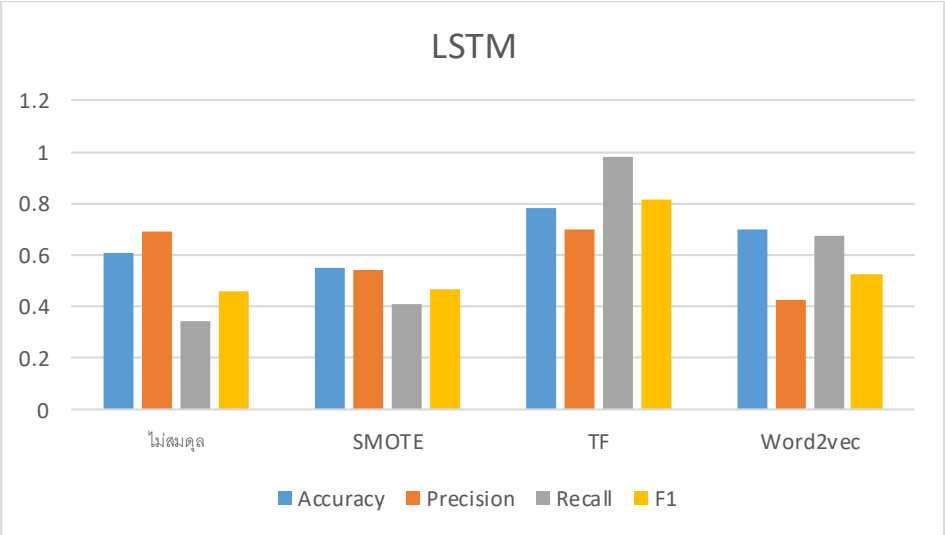
ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน				
	ความถูกต้อง	ความแม่นยำ	ค่าระลึก	เอฟวัน
ไม่สมดุล	0.61	0.667	0.375	0.48
SMOTE	0.64	0.682	0.468	0.556
TF	0.71	0.635	1	0.777
Word2vec	0.82	0.742	0.959	0.837

4.4 เปรียบเทียบการทดลอง

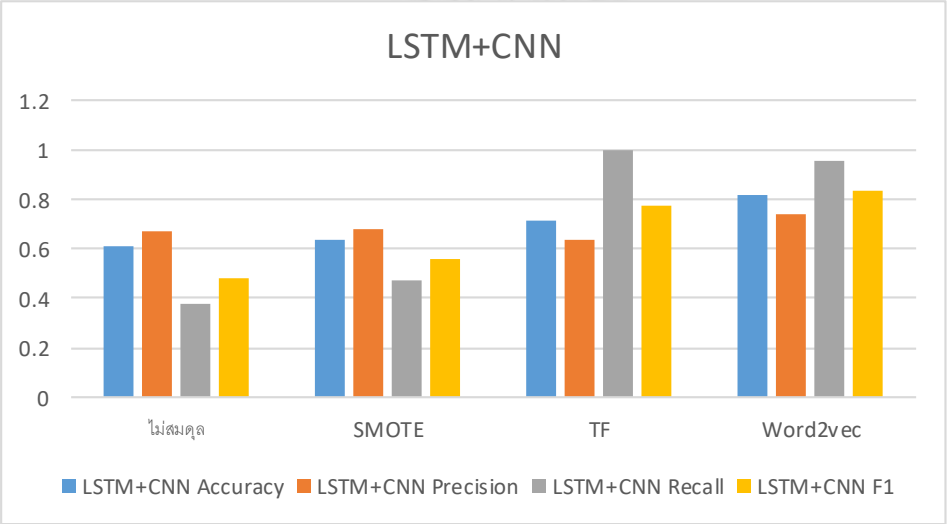
ในหัวข้อนี้จะเป็นการเปรียบเทียบผลการทดลองทั้งสามการทดลองที่ใช้ข้อมูลที่ แต่ต่างกัน ได้แก่ 1 ชุดข้อมูลที่ไม่สมดุล 2 ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคสุ่มตัวอย่างกลุ่มร่วมน้อย 3 ชุดข้อมูลที่สมดุลด้วยเทคนิคการสร้างข้อความจากความถี่ของคำ และ 4.ชุดข้อมูลที่สมดุลด้วยการใช้เทคนิคความคล้ายเชิงมุม โดยทุกชุดข้อมูลจะถูกนำไปทดลองรูปแบบ ผู้วิจัยได้เลือกผลลัพธ์ค่าความถูกต้อง ค่าความแม่นยำ ค่าระลึก และ ค่าเอฟวัน เมื่อจำนวนคุณลักษณะที่มีความเหมาะสมที่ทำให้ตัวแบบมีประสิทธิภาพในการทำนายตัวอย่างเป็นคลาสที่เป็นมากที่สุด กล่าวคือเป็นจำนวนคุณลักษณะที่ทำให้ตัวแบบได้ค่าระบึกสูงที่สุด ซึ่งอาจแตกต่างกันไปในแต่ละชุดข้อมูลแต่ละตัวมแบบการเรียนรู้เครื่องทั้งสามวิธี



รูปภาพ 13กราฟแท่งของการทดลองรูปแบบนาอีฟเบย์



รูปภาพ 14กราฟแท่งของการทดลองรูปแบบหน่วยความจำสั้นแบบยาว



รูปภาพ 15 กราฟแท่งของการทดลองรูปแบบหน่วยความจำสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน

4.5 ตัวอย่างผลลัพธ์ของการทำนายของข้อมูล

ในหัวข้อจะแสดงให้ถึงตัวอย่างของข้อมูลที่ผู้ทำวิจัยใช้ในการทดสอบข้อมูลจากที่ทำการทดลองจากรูปแบบจำแนกประเภทนาอีฟเบย์ ที่ทำการจำแนกข้อความ โดยจะแบ่งเป็นทั้งหมด 3 รูปแบบตามขั้นตอนของการเพิ่มตัวอย่าง เป็นดังตัวอย่างตามนี้

ตาราง 13 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบนาอีฟเบย์ของข้อมูลไม่สมดุล

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ยำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	Hate
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไห	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	Hate
ตัดไปตอนที่มึงด่าคนติงจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	Hate
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	Hate
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	Hate
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปเป็นผลลัพธ์การทำนายในรูปแบบนาอีฟเบย์ของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ตาราง 14 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบนาอีฟเบย์ของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ยำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No

ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไท	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสริฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติงจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	Hate
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	Hate
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นผลลัพธ์การทำนาย ในรูปแบบนาอีฟเบย์ของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการที่มาจากความถี่ของคำ

ตาราง 15 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบนาอีฟเบย์ของข้อมูลที่มาจากความถี่ของคำ

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย่ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไท	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสริฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติงจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	Hate

ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	Hate
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นผลลัพธ์การทำนาย ในรูปแบบนาอีฟเบย์ของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการที่มาจากเทคนิคความคล้ายเชิงมุม

ตาราง 16 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบนาอีฟเบย์ของข้อมูลที่มาจากเทคนิคความคล้ายเชิงมุม

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไท	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติ่งจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No

เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวของข้อมูลไม่สมดุล

ตาราง 17 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวของข้อมูลไม่สมดุล

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไท	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติงจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทำพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	N
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวของ จากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ตาราง 18 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาว ของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ข้อความ	คลาสติดฉลาก	คลาสทำนาย

ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไห	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติ่งจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวของข้อมูลที่มาจากความถี่ของคำ

ตาราง 19 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาว ของข้อมูลที่มาจากความถี่ของคำ

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไห	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No

ตัดไปตอนที่มึงด่าคนติ่งจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวของ ข้อมูลที่มาจากของข้อมูลที่มาจากเทคนิคความคล้ายเชิงมุม

ตาราง 20 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาว ของข้อมูลที่มาจากเทคนิคความคล้ายเชิงมุม

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไห	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น‘ไม่ผิด’ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนทีมึงด่าคนติงจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนีทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No

ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นตัวอย่างผลลัพธ์การทำนายในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลไม่สมดุล

ตาราง 21 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลไม่สมดุล

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไห	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	Hate
ตัดไปตอนที่มึงด่าคนติงจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทำพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	Hate
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ต่อไปจะเป็นตัวอย่างผลลัพธ์การทำนายในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ตาราง 22 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลที่มาจากเทคนิคการสุ่มตัวอย่างส่วนน้อยเพิ่มด้วยการสังเคราะห์

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไห	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสรีฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติ่งจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	Hate
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลที่มาจากความถี่ของคำ

ตาราง 23 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลที่มาจากความถี่ของคำ

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	Hate

ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไท	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสริฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติ่งจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No
ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	No
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	No

ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลที่มาจากความถี่ของคำ

ตาราง 24 ตัวอย่างผลลัพธ์การทำนาย ในรูปแบบหน่วยความจำระยะสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชันของข้อมูลที่มาจากเทคนิคความคล้ายเชิงมุม

ข้อความ	คลาสติดฉลาก	คลาสทำนาย
ดรอาทิตย์ย้ำคนไทยไม่เกลียดทหารแต่รังเกียจทหารที่มายึดอำนาจและแสวงหาผลประโยชน์	Hate	Hate
ด้วยความหวังดีคุณจตุพรสิ่งดีดีที่เคยทำมาสำหรับคนเสื้อแดงเหนื่อยก็พักเถอะครับให้คนเสื้อแดง	Hate	No
ดีที่คอภิสิทธิ์ไม่ถอยตามที่พวกเขาพยายามจะกดดันเพราะถ้าถอยก็จะกลายเป็นว่าใครรวบรวมมวลชนได้ก็สามารถที่จะล้มรัฐบาลไท	Hate	Hate
ต้องเข้าใจกันก่อนว่าระบบทุนนิยมและนายทุนนั้น'ไม่ผิด'ผิดที่ระบบประเทศเรามันเสริฟนายทุนจนเกินไปเพราะจากความเคยชินที่เคยเป็นท	Hate	No
ตัดไปตอนที่มึงด่าคนติ่งจีนวันไชน่าเหยียดสารพัดจะเหยียดทีแบบนี้ทาพูดว่าไม่มีสิทธิ์หาความสุขใส่ตัวบ้างเหรอ	Hate	Hate
ชีวิตตามปกติของน้องม็อบมุ้งมิ้ง	No	Hate
ตอนนี้ม็อบมุ้งมิ้งตาบอดหูหนวกค่ะ	No	No

ถ้านายประยุทธ์จันทร์โอชาไม่ได้เป็นทหารทุกท่านคิดว่านายประยุทธ์จะประกอบอาชีพอะไรได้1รับเหมาก่อสร้าง2เปิดร้านขายของ	No	No
ทรัมป์เหมือนสนธิสุเทพนี้หละ	No	Hate
ทักษิณพูดในคลิปว่าถ้าย้อนเวลาไปได้อยากแก้ไขการจัดการปัญหาภาคใต้ด้วยการเจรจามากกว่าที่เคยทำไว้ในสมัยยิ่งลักษณ์จึงมีการสร้างเว	No	No
เหี้ยมากใครก็ได้ฝากเปิดในหน่อย ข่าวในพระราชสำนัก รีมิกซ์	No	Hate

บทที่ 5 สรุปผลการทดลอง

5.1 สรุปผลการทดลอง

หลังจากการเตรียมข้อมูลซึ่งเป็นข้อความที่เก็บรวมรวมจากทวิตเตอร์ โดยใช้การตัดคำการแทนใช้เทคนิคความถี่ของค่า-ส่วนกลับความถี่ของเอกสาร การนำเทคนิคการสุ่มเพิ่มตัวอย่างกลุ่มน้อยมาใช้ประยุกต์ใช้กับชุดข้อมูล ดังกล่าวสามารถทำให้ตัวแบบจำแนก ประเภทนาอีฟเบย์ หน่วยความจำระยะสั้นแบบยาว และโครงข่ายประสาทแบบคอนโวลูชัน ร่วมกับ หน่วยความจำระยะสั้นแบบยาว มีประสิทธิภาพเพิ่มขึ้นในแง่ ของค่าความแม่นและค่าระลึก โดยสังเกตจากการเพิ่มขึ้นของค่าดังกล่าวในเกือบทุกชุดข้อมูล ของแต่ละชุดข้อมูล จะพบว่าเมื่อจำนวนของข้อมูลมีความสมดุลกันขึ้น สามารถทำให้ค่าของ ตัววัดต่างๆมีค่าสูงขึ้น

ต่อมาผู้วิจัยได้เสนอเทคนิคการเพิ่มตัวอย่างจากการใช้เทคนิคของการใช้ค่าความคล้ายคลึงในเชิงมุมในการสร้างตัวอย่างเพิ่ม โดยตั้งสมมติฐานว่าจากการทดลองที่ทำการเพิ่มกลุ่มข้อมูลโดยเทคนิคความถี่ของค่า-ส่วนกลับความถี่ของเอกสารที่คำที่มีค่าสูงนั้นจะมีผลต่อการเรียนรู้ของตัวแบบคือแพทเทิร์นของคำที่เป็นคำที่ใช้บ่อยในชุดของข้อมูลจะส่งผลให้ การสร้างข้อมูลใหม่ขึ้นมานั้นมีประสิทธิภาพที่สูงขึ้น หลังจากประยุกต์ใช้เทคนิคดังกล่าวแล้วพบว่าเทคนิคนี้สามารถเพิ่มประสิทธิภาพในแง่ ของ ค่าระลึกและ ค่าความถูกต้องในตัวแบบที่เป็นตัวแบบของรูปแบบหน่วยความจำสั้นแบบยาวร่วมกับโครงข่ายประสาทแบบคอนโวลูชัน แต่ในขณะที่ค่าความถูกต้องลดต่ำลงในตัวแบบ รูปแบบหน่วยความจำสั้นแบบยาว ซึ่งมีค่าไม่ต่างกันมาก

โดยสรุปการใช้เทคนิคสังเคราะห์ข้อมูลแต่ละรูปแบบที่เสนอมาสามารถทำให้ตัวแบบทำนายข้อมูลเป็นกลุ่มที่เป็นคลาสบวกซึ่งเป็นคลาสกลุ่มน้อยมากขึ้นได้ จากการที่ใช้เทคนิคในการเพิ่มกลุ่มตัวอย่างทั้ง 2 วิธีที่เสนอไปสามารถเพิ่มประสิทธิภาพได้ทั้งหมดทุกวิธี ซึ่งในการทดลองนี้เนื่องจากคำมีผลกับชุดข้อมูลจึงเป็นผลให้ ค่าต่างๆของตัวแบบมีประสิทธิภาพเพิ่มขึ้นได้

5.2 ผลงานตีพิมพ์จากงานวิจัย

ส่วนหนึ่งของวิทยานิพนธ์นี้ได้ตีพิมพ์และนำเสนอในที่ประชุมวิชาการดังนี้

- บทความวิชาการเรื่อง “การสังเคราะห์ข้อความเพื่อเพิ่มตัวอย่างในการตรวจจับข้อความประทุษวาจาในภาษาไทย” โดย ธโนภาส วรรณวโรทร และ สุกรี สินธุภิญโญ ใน การจัดการประชุมวิชาการเสนอผลงานวิจัยระดับชาติด้านวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยราชภัฏจันทรเกษม ครั้งที่ 5 (The 5th CRU-National Conference in Science and Technology : NCST 5th 2022)

บรรณานุกรม

[1] Alon Jacovi, Oren Sar Shalom, Yoav Goldberg. (2020).Understanding Convolutional Neural Networks for Text Classification 2020 Computation and Language (cs.CL)

[2] B. Go´ralewicz. (2018). The TFIDF Algorithm Explained.

[3] Chikashi N., Joel T., Achint T., Yashar M., and Yi Chang. (2016). Abusive language detection in online user content. In Proceedings of the 25th International Conference on World Wide Web, pages 145–153.

[4] D. Devi., S. K. Biswas and B. Purkayastha.(2020),A Review on Solution to Class Imbalance Problem: Undersampling Approaches, 2020 International Conference on Computational Performance Evaluation (ComPE), 2020, pp. 626-631, doi: 10.1109/ComPE49325.2020.9200087.

[5] H.Rathpisey., T.B.Adji.(2019), Handling Imbalance Issue in Hate Speech Classification using Sampling-based Methods, 2019 5th International Conference on Science in Information Technology (ICSITech).

[6] H. Zhang., D. Li. (2007), "Naïve Bayes Text Classifier," (2007) IEEE International Conference on Granular Computing (GRC 2007), 2007, pp. 708-708, doi: 10.1109/GrC. 2007.40.

[7] I. Twitter. The Search API. Available: https://dev.twitter.com/rest/public/search

[8] M. Ibrahim, M. Torki and N. El-Makky.V. (2018),"Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning," In the 17th IEEE International Conference on Machine Learning and Applications, 2018.

[9] N.V. Chawla., K.W.B., L.O. Hall., W.P. Kegelmeyer. (2002), SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002.

[10] P. Sarakit, T. Theeramunkong and C. Haruechaiyasak. (2015), Improving emotion classification in imbalanced YouTube dataset using SMOTE algorithm, In the 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA),Chonburi, 2015.

[11] Piyaphakdeesakun C., Facundes, N., Polvichai, J. (2019). Thai comments sentiment analysis on social networks with deep learning approach. In 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) (pp. 1-4). IEEE.

[12] S. Hochreiter., J. Schmidhuber. (1997). Long short-term memory. Neural computation, 15 Nov 1997;9 (8):1735-80.

[13] T.H. Wen, M. Gasic, N. Mrksic, P.H. Su, D. Vandyke and S. Young. (2015), "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," arXiv preprint arXiv:1508.01745, 7 Aug 2015.

[14] Winda K. S., Dian P.i R., Reza F. M., (2020), Multilabel Classification for News Article Using Long Short-Term Memory, Sriwijaya Journal of Informatic and Applications Vol. 01, No. 01, August 2020, pp. 34-44

[15] Yuandong L., Shaofu L. (2019), Research on Text Classification Based on CNN and LSTM, 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)

[16] Y. Liu, Q. Xu and Z. Tang, "Research on Text Classification Method Based on PTF-IDF and Cosine Similarity," 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2019, pp. 205-208, doi: 10.1109/ICIIBMS46890.2019.8991542.

[17] P., Tang L., Liu H. (2009) Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-39940-9_565

[18] Visa, S., Ramsay, B., Ralescu, A.L. and Van Der Knaap, E., 2011. Confusion matrix-based feature selection. MAICS, 710(1), pp.120-127.

[19] Rehurek, R., & Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2), 2.

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient E stimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

[21] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324

[22] Cherian, T., Badola, A., & Padmanabhan, V. (2018). Multi-cell LSTM based neural language model. arXiv preprint arXiv:1811.06477.

ประวัติผู้เขียน

ชื่อ-สกุล	ธโนภาส วรรณวโรทร
วัน เดือน ปี เกิด	16 พฤศจิกายน 2536
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	วศ.บ. วิศวกรรมศาสตร์ มหาวิทยาลัยพระจอมเกล้าพระนครเหนือ
ที่อยู่ปัจจุบัน	159/66 รามอินทรา 8 ถนน รามอินทรา แขวงอนุเสาวีย์ เขตบางเขน กรุงเทพ