โครงการ

การเรียนการสอนเพื่อเสริมประสบการณ์

ชื่อโครงการ การจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง
Classifying Thai News Dialogues into Topic Types Using Machine Learning Technique

ชื่อนิสิต 1. นางสาวศลิษา ชูชื่นพฤกษาพันธ์ 6033661023
2. นางสาวไอศวรรย์ ธโนศวรรย์ 6033673523

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์
สาขาวิชาวิทยาการคอมพิวเตอร์

ปีการศึกษา 2563

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง

นางสาวศลิษา ชูชื่นพฤกษาพันธ์

นางสาวไอศวรรย์ ธโนศวรรย์

โครงการนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรวิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ปีการศึกษา 2563

Classifying Thai News Dialogues into Topic Types Using Machine Learning Technique

Salisa Chuchuenprueksaphan

Isawan Thanosawan

A Project Submitted in Partial Fulfillment of the Requirements

for the Degree of Bachelor of Science Program in Computer Science

Department of Mathematics and Computer Science

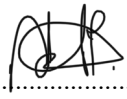
Faculty of Science

Chulalongkorn University

Academic Year 2020

หัวข้อโครงงาน	การจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง
โดย	นางสาวศลิษา ชูชื่นพฤกษาพันธ์
	นางสาวไอศวรรย์ ธโนศวรรย์
สาขาวิชา	วิทยาการคอมพิวเตอร์
อาจารย์ที่ปรึกษาโครงงานหลัก	รองศาสตราจารย์ ดร.ศุภกานต์ พิมลธเรศ
อาจารย์ที่ปรึกษาโครงงานร่วม	ผู้ช่วยศาสตราจารย์ ศศิภา พันธุวดีธร

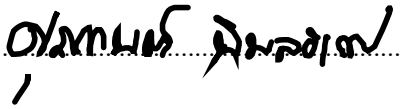
ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับโครงงานฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาบัณฑิตในรายวิชา 2301499 โครงงานวิทยาศาสตร์ (Senior Project)



..... หัวหน้าภาควิชาคณิตศาสตร์

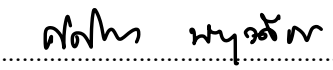
(ศาสตราจารย์ ดร.กฤษณะ เนียมมณี) และวิทยาการคอมพิวเตอร์

คณะกรรมการสอบโครงงาน



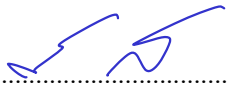
..... อาจารย์ที่ปรึกษาโครงงานหลัก

(รองศาสตราจารย์ ดร.ศุภกานต์ พิมลธเรศ)



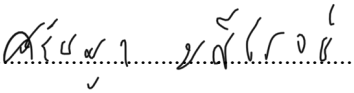
..... อาจารย์ที่ปรึกษาโครงงานร่วม

(ผู้ช่วยศาสตราจารย์ ศศิภา พันธุวดีธร)



..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.กรุง สินอภิรมย์สราญ)



..... กรรมการ

(รองศาสตราจารย์ ดร.ศรันญา มณีโรจน์)

บทคัดย่อภาษาไทย

นางสาวศลิษา ชูชื่นพฤกษาพันธ์, นางสาวไอศวรรย์ ธโนศวรรย์: การจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง. (Classifying Thai News Dialogues into Topic Types Using Machine Learning Technique)

อ.ที่ปรึกษาโครงงานหลัก: รองศาสตราจารย์ ดร.ศุภกานต์ พิมลธเรศ, อ.ที่ปรึกษาโครงงานร่วม: ผู้ช่วยศาสตราจารย์ ศศิภา พันธุวดีธร, 77 หน้า

รายการข่าวเป็นสื่อที่มีความสำคัญต่อการติดตามเหตุการณ์ใหม่และความเปลี่ยนแปลงของสังคมที่เกิดขึ้นตลอดเวลา ซึ่งรายการข่าวมักนำเสนอข่าวหัวข้อข่าวที่หลากหลายรวมอยู่ในรายการเดียวกัน โครงงานนี้มีจุดประสงค์เพื่อสร้างตัวจำแนกและเปรียบเทียบประสิทธิภาพการจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อ ในการศึกษานี้ ตัวจำแนกชุดคำโต้ตอบภาษาไทยหกตัวที่ใช้ขั้นตอนวิธีที่แตกต่างกันได้นำมาใช้เพื่อจำแนกประเภทคำโต้ตอบข่าวไทยออกเป็นประเภทของข่าวหกประเภท ได้แก่ ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม ข่าวบันเทิง ข่าวกีฬา และข่าวสิ่งแวดล้อม ตัวจำแนกประเภทห้าตัวได้แก่ นาอีฟเบย์แบบอเนกนาม เพื่อนบ้านใกล้ที่สุดเคตัว ป่าสุ่ม ซัพพอร์ตเวกเตอร์แมชชีน และเพอร์เซปตรอนหลายชั้นใช้เวกเตอร์คุณลักษณะที่ได้จากความถี่ของคำและความถี่ของเอกสารที่ผกผัน ทว่าตัวจำแนกอีกตัวคือเพอร์เซปตรอนหลายชั้นใช้เวกเตอร์ความน่าจะเป็นของหัวข้อที่ได้จากการจัดสรรของดีรีเคลท์แฝง ผลการทดลองพบว่าตัวจำแนกที่สามารถจำแนกคำโต้ตอบข่าวไทยได้ดีที่สุดคือ เพอร์เซปตรอนหลายชั้นที่ใช้เวกเตอร์คุณลักษณะบนพื้นฐานของความถี่ของคำและความถี่ของเอกสารที่ผกผัน และให้ค่าความเที่ยงเฉลี่ยเป็น 0.9622 ค่าความครบถ้วนเฉลี่ยเป็น 0.9609 และคะแนน F1 เฉลี่ยเป็น 0.9609

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต .....

ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาโครงงานหลัก .....

ปีการศึกษา 2563 ลายมือชื่อ อ.ที่ปรึกษาโครงงานร่วม .....

บทคัดย่อภาษาอังกฤษ

6033661023, 6033673523: MAJOR COMPUTER SCIENCE

KEYWORDS: MACHINE LEARNING/ CLASSIFICATION ALGORITHM/ MULTINOMIAL NAIVE BAYES/ K-NEAREST NEIGHBORS/ RANDOM FOREST/ SUPPORT VECTOR MACHINES/ MULTI-LAYER PERCEPTRON/ LATENT DIRICHLET ALLOCATION

SALISA CHUCHUENPRUEKSAPHAN, ISAWAN THANOSAWAN: CLASSIFYING THAI NEWS DIALOGUES INTO TOPIC TYPES USING MACHINE LEARNING TECHNIQUE.

ADVISOR: ASSOC. PROF. SUPHAKANT PHIMOLTARES, Ph.D., CO-ADVISOR: ASST. PROF. SASIPA PANTHUWADEETHORN, 77 pp.

News programs are an important media to keep up with new events and social changes which happen all the time and news programs mostly present various news topics in the same program. The purpose of this project is to create classifiers and compare performance of classifying Thai news dialogues as topic types. In this study, six Thai news dialogues classifiers using different algorithms were used to classify Thai news dialogues into six types of news, which are political news, economic news, crime news, entertainment news, sports news, and environmental news. Five classifiers, which are Multinomial Naive Bayes, K-Nearest Neighbors, Random Forest, Support Vector Machines, and Multi-Layer Perceptron used feature vectors obtained from Term Frequency-Inverse Document Frequency whereas the other classifier is Multi-Layer Perceptron using the topic probability vectors obtained from Latent Dirichlet Allocation. The experimental results showed that the best Thai news dialogues classifier was Multi-Layer Perceptron using feature vectors based on Term Frequency-Inverse Document Frequency and yielded an average precision of 0.9622, average recall of 0.9609, and average F1-score of 0.9609.

Department : Mathematics and Computer Science Student's Signature Salisa

Student's Signature Isawan

Field of Study : Computer Science Advisor's Signature

Academic Year : 2020 Co-advisor's Signature

กิตติกรรมประกาศ

ในงานวิจัย “การจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง” นี้ ได้รับการสนับสนุนและความช่วยเหลืออย่างเต็มที่จาก รองศาสตราจารย์ ดร.ศุภกานต์ พิมลธเรศ อาจารย์ที่ปรึกษาโครงงานหลัก และผู้ช่วยศาสตราจารย์ ศศิภา พันธุวดีธร อาจารย์ที่ปรึกษาโครงงานร่วมในการให้คำปรึกษาและคำชี้แนะต่าง ๆ อันเป็นประโยชน์ต่อการทำงานวิจัย ตรวจทานแก้ไขข้อผิดพลาด รวมถึงคอยให้กำลังใจและติดตามให้โครงงานนี้มีความคืบหน้าอยู่เสมอจนกระทั่งสำเร็จไปได้ด้วยดี

ขอขอบพระคุณผู้ช่วยศาสตราจารย์ ดร.กรุง สินอภิรมย์สราญ และรองศาสตราจารย์ ดร.ศรันญา มณีโรจน์ กรรมการสอบโครงงาน ซึ่งได้ช่วยชี้แนะให้โครงงานมีความสมบูรณ์มากขึ้น

ขอขอบพระคุณอาจารย์ท่านอื่นที่มิได้กล่าวนามไว้ ณ ที่นี้ ที่ได้ถ่ายทอดความรู้ให้ผู้วิจัยได้มีความรู้ และเข้าใจในทฤษฎีต่าง ๆ อันเป็นประโยชน์ในการดำเนินงานวิจัยครั้งนี้

ขอขอบพระคุณภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ที่ได้จัดสถานที่ให้นิสิตในการดำเนินงานวิจัย รวมถึงงบประมาณค่าใช้จ่ายในการดำเนินงานวิจัยนี้

ขอขอบพระคุณบิดา มารดา ญาติมิตร และเพื่อนทุกท่านที่ได้สนับสนุน ให้คำปรึกษา ส่งผลให้งานวิจัยนี้สำเร็จลุล่วงไปได้ด้วยดี

สารบัญ

สารบัญตาราง

สารบัญรูปภาพ

บทที่ 1
บทนำ

1.1 ความเป็นมาและเหตุผลการวิจัย

ในทุกวินาทีของโลกปัจจุบันมีเหตุการณ์ใหม่ ๆ ที่น่าสนใจเกิดขึ้นอยู่ตลอดเวลา ซึ่งการติดตามเหตุการณ์ความเปลี่ยนแปลงที่เกิดขึ้นที่รวดเร็วและมีประสิทธิภาพนั้นต้องอาศัยการติดตามข่าวสารผ่านสื่อต่าง ๆ เช่น หนังสือพิมพ์รายวัน บทความข่าวบนเว็บไซต์ โดยเฉพาะอย่างยิ่ง “รายการข่าว” ไม่ว่าจะเป็นรายการข่าว โทรทัศน์ วิทยุ รวมไปถึงบนแพลตฟอร์มออนไลน์ จากการสำรวจของ กสทช. [1] พบว่าความนิยมรายการข่าวของคนไทยในปี 2563 ยังคงเป็นไปอย่างต่อเนื่อง โดยจุดเด่นของการรับข่าวสารผ่านรายการข่าวคือผู้ชมจะได้ฟังการรายงานข่าวจากผู้รายงานข่าวที่ผ่านการคัดกรอง เรียบเรียง วิเคราะห์ และสรุปมาจากแหล่งข่าวที่เชื่อถือได้ รายการข่าวจึงเป็นสื่อที่มีความสำคัญอย่างมากในสังคมปัจจุบัน

นอกจากนี้แล้วรายการข่าวยังสามารถแบ่งได้ตามลักษณะการรายงานข่าว คือ การรายงานข่าวแบบบรรยายซึ่งมักใช้ภาษาระดับทางการในการอ่านข่าว และไม่มีการแสดงความคิดเห็นจากผู้บรรยายข่าว และอีกลักษณะคือการรายงานข่าวแบบเล่าข่าวที่มีจุดเด่นเป็นการใช้ภาษาในระดับที่ไม่เป็นทางการ และมีการแสดงความคิดเห็นหรือโต้ตอบกันระหว่างผู้รายงานข่าว ทำให้ชุดคำโต้ตอบข่าวมีความเป็นธรรมชาติของภาษาและมีลักษณะคล้ายกับภาษาพูดที่ใช้สื่อสารในชีวิตจริงมากกว่าการรายงานข่าวแบบบรรยาย อย่างไรก็ตามในรายการข่าวหนึ่งรายการมักจะมีการนำเสนอข่าวในหลากหลายหัวข้อ เช่น ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวบันเทิง ข่าวอาชญากรรม ข่าวต่างประเทศ และอื่น ๆ ทั้งหมดนี้ถูกรวมอยู่ในรายการข่าวเดียวกัน อีกทั้งบางรายการข่าวมีความยาวมากถึง 2 ชั่วโมง จึงยากสำหรับคนที่ต้องการเลือกฟังข่าวแบบเจาะจงหัวข้อ เนื่องจากต้องเสียเวลาในการค้นหาข่าวที่ตนสนใจ

จากที่กล่าวมาข้างต้น ทำให้ผู้พัฒนามีความสนใจที่จะศึกษาเกี่ยวกับการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าว เพื่อบอกประเภทหัวข้อข่าวว่าเป็นข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม ข่าวบันเทิง ข่าวกีฬา หรือ ข่าวสิ่งแวดล้อม ซึ่งในการพัฒนา ผู้พัฒนาจะเก็บชุดข้อมูลเสียงโต้ตอบระหว่างผู้รายงานข่าวจากวิดีโอย้อนหลังของรายการข่าวและแปลงเสียงเป็นข้อความด้วยวิธี Speech-to-Text (STT) แล้วจึงนำข้อความมาสกัดเป็นเวกเตอร์คุณลักษณะ (feature vector) จากนั้นจึงสร้างแบบจำลองสำหรับจำแนกหัวข้อข่าวโดยใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อนำเสนอวิธีที่สามารถจำแนกประเภทข่าวจากคำโต้ตอบที่มีประสิทธิภาพ

1.2 วัตถุประสงค์การวิจัย

เพื่อพัฒนาตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง

1.3 ขอบเขตการวิจัย

1. ข้อมูลนำเข้าคือชุดคำโต้ตอบของข่าวภาษาไทย
2. ชุดคำโต้ตอบรวบรวมจากวิดีโอรายการข่าวไทย โดยมีอายุข่าวย้อนหลังไม่เกิน 2 ปี คือตั้งแต่เดือนตุลาคม 2561 ถึง ตุลาคม 2563 และมีผู้รายงานข่าว 2 คนขึ้นไป
3. ชุดคำโต้ตอบแบ่งตามประเภทหัวข้อข่าวได้ 6 ประเภท ได้แก่ ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม ข่าวบันเทิง ข่าวกีฬา และข่าวสิ่งแวดล้อม โดยมีนิยามของแต่ละประเภทข่าว ดังนี้
    ก. ข่าวการเมือง คือข่าวการเคลื่อนไหวของพรรคการเมือง นักการเมือง และกระบวนการต่าง ๆ ทางการเมือง
    ข. ข่าวเศรษฐกิจ คือข่าวความเคลื่อนไหวทางเศรษฐกิจรวมถึงโครงการกระตุ้นเศรษฐกิจของรัฐบาล
    ค. ข่าวอาชญากรรม คือข่าวเกี่ยวข้องกับคดีอาชญากรรมต่าง ๆ การเข้าจับกุมคนร้าย
    ง. ข่าวบันเทิง คือข่าวที่นำเสนอเรื่องราวในวงการบันเทิงของดารา นักร้อง และศิลปิน
    จ. ข่าวกีฬา คือข่าวที่รายงานเกี่ยวกับเรื่องกีฬาและนักกีฬาต่าง ๆ รวมถึงการวิเคราะห์และรายงานผลการแข่งขันกีฬา
    ฉ. ข่าวสิ่งแวดล้อม คือข่าวเกี่ยวกับมลพิษทางสิ่งแวดล้อม และภัยพิบัติต่อสิ่งแวดล้อม
4. จำนวนชุดคำโต้ตอบประเภทละ 100 ชุด
5. ชุดคำโต้ตอบแต่ละชุดมีระยะเวลาไม่เกิน 15 วินาที

1.4 ขั้นตอนการวิจัย

ก. แผนการศึกษา

1. ศึกษาค้นคว้าทฤษฎีที่เกี่ยวข้อง
2. ศึกษางานวิจัยที่เกี่ยวข้อง
3. เตรียมชุดข้อมูลคำโต้ตอบข่าวไทย
4. วิเคราะห์ ออกแบบ และสร้างตัวแบบในการจำแนกประเภทหัวข้อข่าว
5. ทดสอบประสิทธิภาพของตัวแบบ
6. วิเคราะห์และอภิปรายผล
7. จัดทำเอกสาร

ข. ระยะเวลาที่ศึกษา

ขั้นตอนการดำเนินงาน	ปี 2563					ปี 2564			
	ส.ค.	ก.ย.	ต.ค	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1. ศึกษาค้นคว้าทฤษฎีที่เกี่ยวข้อง									
2. ศึกษางานวิจัยที่เกี่ยวข้อง									
3. เตรียมชุดข้อมูลคำโต้ตอบข่าวไทย									
4. วิเคราะห์ ออกแบบ และสร้างตัวแบบในการจำแนกประเภทหัวข้อข่าว									
5. ทดสอบประสิทธิภาพของตัวแบบ									
6. วิเคราะห์และอภิปรายผล									
7. จัดทำเอกสาร									

1.5 ประโยชน์ที่ได้รับ

ก. ประโยชน์ที่ได้ต่อผู้พัฒนา

1. ได้รับความรู้จากการศึกษาค้นคว้าทฤษฎีและงานวิจัยที่เกี่ยวข้อง
2. ได้ฝึกฝนและพัฒนาทักษะในการสร้างตัวจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง
3. ได้ฝึกฝนและพัฒนาทักษะการวางแผนและทำงานเป็นขั้นตอน
4. ได้ฝึกการทำงานเป็นกลุ่ม การยอมรับความคิดเห็นผู้อื่น และความรับผิดชอบในหน้าที่
5. ได้พัฒนาศักยภาพในการเรียนรู้ด้วยตัวเอง

ข. ประโยชน์ที่ได้ต่อผู้ใช้

1. ได้แนวทางในการจำแนกประเภทหัวข้อข่าวจากชุดคำโต้ตอบ
2. สามารถนำชุดข้อมูลข่าวไทย และแนวทางที่เสนอไปพัฒนาต่อยอดได้

1.6 โครงสร้างของรายงาน

บทที่ 2 กล่าวถึงหลักการและทฤษฎีที่เกี่ยวข้องกับการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าว รวมทั้งงานวิจัยที่เกี่ยวข้อง

บทที่ 3 กล่าวถึงวิธีการวิจัยในการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าว

บทที่ 4 กล่าวถึงกระบวนการทดลองและผลของการดำเนินการวิจัยของการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าวโดยแสดงและวิเคราะห์ประสิทธิภาพของวิธีการรวมถึงการอภิปรายผลการทดลอง

บทที่ 5 กล่าวถึงการสรุปผลการวิจัยการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าวและข้อเสนอแนะ

บทที่ 2
งานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงหลักการและทฤษฎีที่เกี่ยวข้องกับการสร้างแบบจำลองสำหรับจำแนกประเภทหัวข้อจากคุณลักษณะที่ได้จากข้อมูลคำโต้ตอบข่าวภาษาไทย รวมถึงงานวิจัยที่เกี่ยวข้อง

2.1 หลักการและทฤษฎีที่เกี่ยวข้อง

2.1.1 Term Frequency–Inverse Document Frequency (TF-IDF)

ในการพิจารณาความสำคัญของคำในเอกสารใช้เทคนิค TF-IDF [2] เข้ามาช่วยในการแยกคำตามความสำคัญจากน้ำหนักของคำแต่ละคำ TF-IDF นั้นประกอบด้วย 2 ส่วนคือ TF (Term Frequency) และ IDF (Inverse Document Frequency) โดยมีสมการที่ใช้ในการคำนวณ ดังสมการที่ 1

$$TF\text{-}IDF(t,d) = TF(t,d) * IDF(t) \quad (1)$$

ส่วนแรกคือ TF เป็นการดูความถี่ของคำที่เกิดขึ้นในเอกสาร โดยหาสัดส่วนระหว่างจำนวนครั้งที่คำ t ปรากฏในเอกสาร d กับจำนวนรวมคำทั้งหมดในเอกสาร d ดังสมการที่ 2

$$TF(t,d) = \frac{f(t,d)}{\sum_{t' \in d} f(t',d)} \quad (2)$$

ส่วนต่อมาคือ IDF หรือความถี่ของเอกสารที่ผกผัน ใช้บอกถึงความสำคัญของคำในเอกสารทั้งหมด กล่าวคือเมื่อคำใดปรากฏอยู่ทั่วไปในหลายเอกสาร คำนั้นจะมีค่า IDF ต่ำ หมายความว่าเป็นคำที่มีความสำคัญน้อย โดยคำนวณได้ตามสมการที่ 3

$$IDF(t) = log\left(\frac{1+n}{1+df(t)}\right) + 1 \quad (3)$$

เมื่อ n แทน จำนวนเอกสารทั้งหมดในชุดข้อมูล

$df(t)$ แทน จำนวนเอกสารในชุดข้อมูลที่มีคำ t

2.1.2 Multinomial Naive Bayes (MNB)

Naive Bayes [3] คือขั้นตอนวิธีของการเรียนรู้แบบมีผู้สอนแบบหนึ่งที่เหมาะกับการจำแนกประเภทแบบหลายคลาสและสามารถทำงานได้ดีกับข้อมูลที่มีขนาดมิติสูง จึงเป็นขั้นตอนวิธีที่เป็นที่นิยมสำหรับงานจำแนกประเภทข้อความ Naive Bayes อาศัยทฤษฎีความน่าจะเป็นตามทฤษฎีบทของเบย์ (Bayes' Theorem) ที่ใช้การคำนวณความน่าจะเป็นของเหตุการณ์ที่เป็นอิสระต่อกันแบบมี

เงื่อนไขให้ชุดสอน $T = \{(\boldsymbol{x}_1, y_1),\ (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)\}$ และป้ายกำกับ y โดยที่ $y_i \in \{c_1, c_2, \dots, c_k\}$ และ $i \in \{1,2, \dots, N\}$

สมมติให้จำนวนค่าที่เป็นไปได้ของ $x^{(l)}$ เท่ากับ S_l โดยที่ $l = 1,2, \dots, n$ และจำนวนค่าที่เป็นไปได้ของ Y เท่ากับ k Naive Bayes จะเรียนรู้การแจกแจงความน่าจะเป็นร่วม $P(X|Y)$ ของข้อมูลนำเข้าและข้อมูลส่งออกโดยใช้การแจกแจงความน่าจะเป็นแบบมีเงื่อนไข ตามสมมติฐานของความอิสระต่อกันแบบมีเงื่อนไข ดังสมการที่ 4

$$P\left(X = \boldsymbol{x}\middle|Y = c_j\right) = P\left(X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)}\middle|Y = c_j\right) = \prod_{l=1}^{n} P\left(X^{(l)} = x^{(l)}\middle|Y = c_j\right); j = 1,2, \dots, k \quad (4)$$

ตัวจำแนก Multinomial Naive Bayes (MNB) [4] เป็นการใช้ขั้นตอนวิธี Naive Bayes เพื่อใช้กับข้อมูลที่มีลักษณะเป็นอเนกนาม (multinomial) ในการสร้างตัวจำแนก MNB จากข้อมูลนำเข้า $\boldsymbol{x}$ จะคำนวณหาป้ายกำกับ y ที่มีความน่าจะเป็นภายหลัง (posterior probability) ที่มากที่สุด เพื่อหาป้ายกำกับ y ที่เป็นไปได้มากที่สุด ซึ่งคำนวณด้วยทฤษฎีบทของเบย์ได้ดังสมการที่ 5 และ 6

$$P(Y = c_j|X = \boldsymbol{x}) = \frac{P(X = x|Y = c_j)P(Y = c_j)}{\sum_j P(X = x|Y = c_j)P(Y = c_j)} \quad (5)$$

$$y = \underset{c_j}{argmax}\, P(Y = c_j) \prod_l P\left(X^{(l)} = x^{(l)}\middle|Y = c_j\right) \quad (6)$$

ทั้งนี้ในการคำนวณความน่าจะเป็นของขั้นตอนวิธี Naive Bayes อาจเกิดปัญหาความน่าจะเป็นที่มีค่าเป็นศูนย์ (zero probability) ดังนั้นจึงต้องทำการปรับปรุงพารามิเตอร์เพื่อให้ค่าประมาณที่คำนวณได้มีค่าไม่เท่ากับศูนย์ เรียกว่าการทำ smoothing โดยอาศัยเทคนิค Additive smoothing ดังสมการที่ 7 ซึ่งเป็นการกำหนดพารามิเตอร์ alpha (α) ให้เป็นพารามิเตอร์การปรับให้เรียบ โดยที่ alpha ต้องมีค่ามากกว่า 0 เมื่อ alpha มีค่าเท่ากับ 1 เรียกว่า Laplace smoothing และเมื่อ alpha มีค่าน้อยกว่า 1 จะเรียกว่า Lidstone smoothing

$$P\left(X^{(l)} = x^{(l)}\middle|Y = c_j\right) = \frac{N_{x^{(l)},c_j} + \alpha}{N_{c_j} + \alpha n} \quad (7)$$

เมื่อ $N_{x^{(l)},c_j}$ แทน จำนวนข้อมูลในคลาส c_j ที่คุณลักษณะ $X^{(l)}$ มีค่าเท่ากับ $x^{(l)}$

N_{c_j} แทน จำนวนข้อมูลในคลาส c_j

α แทน พารามิเตอร์การปรับให้เรียบ

n แทน จำนวนคุณลักษณะทั้งหมด

ตัวจำแนก MNB เหมาะสำหรับการจำแนกคุณลักษณะแบบไม่ต่อเนื่อง เช่น การนับจำนวนคำ TF-IDF เป็นต้น ตัวจำแนกแบบ MNB จึงเป็นที่นิยมในงานการจำแนกข้อความ [5]

2.1.3 K-Nearest Neighbors (KNN)

K-nearest neighbors (KNN) [6] คือขั้นตอนวิธีของการเรียนรู้แบบมีผู้สอนแบบหนึ่ง ใช้สำหรับงานในการจำแนกประเภทและวิเคราะห์การถดถอย เป็นขั้นตอนวิธีที่ไม่ซับซ้อน ใช้เวลาในการฝึกฝนชุดสอนน้อยกว่าขั้นตอนวิธีอื่น ๆ โดยอาศัยแนวคิดที่ว่าจุดใด ๆ มักใช้คุณสมบัติร่วมกันกับจุดใกล้เคียง การเลือกวิธีคำนวณระยะห่างระหว่างจุดนั้นขึ้นอยู่กับการใช้งานวิธีที่นิยมใช้ เช่น การคำนวณระยะห่างแบบ Euclidean และการคำนวณระยะห่างระหว่างข้อความแบบ Manhattan

ขั้นตอนวิธี KNN [7] เป็นการหาป้ายกำกับของจุดข้อมูล x_i จากการลงคะแนนเสียงข้างมากของป้ายกำกับของจุดที่ใกล้ที่สุดจำนวน K จุด ดังนั้นจึงจำเป็นที่จะต้องวัดความคล้ายคลึงกันระหว่างจุดข้อมูล x_i และ θ_j โดยใช้เมทริกซ์มินคอฟสกี (p-norm) ในปริภูมิ n มิติของจำนวนจริง ดังสมการที่ 8 สำหรับค่า p เป็น 2 คือการคำนวณระยะห่างแบบ Euclidean และสำหรับค่า p เป็น 1 คือการคำนวณระยะห่างแบบ Manhattan

$$d(x_i, \theta_j) = \left(\sum_{m=1}^{n} \left|x_{im} - \theta_{jm}\right|^p\right)^{\frac{1}{p}} \quad (8)$$

เมื่อ x_i, θ_j แทน จุดใด ๆ ใน n มิติ

และในปริภูมิข้อมูลอื่น ๆ ต้องเลือกการคำนวณระยะห่างที่เหมาะสม เช่น การคำนวณระยะห่างแบบ Hamming ในปริภูมิ q มิติของจำนวนเต็ม

2.1.4 Random Forest (RF)

Random Forest (RF) [8] คือขั้นตอนวิธีของการเรียนรู้แบบมีผู้สอนแบบหนึ่งที่ใช้งานง่ายและได้รับการพิสูจน์แล้วว่าสามารถทำงานได้ดีในหลากหลายปัญหา Random Forest นั้นเป็นขั้นตอนวิธีที่ใช้การรวมแบบจำลองการเรียนรู้หลากหลายแบบเข้าด้วยกัน (ensemble learning) ทำให้ได้จุดที่สมดุลกันระหว่าง ค่า bias และความแปรปรวนของข้อมูลที่ทำให้เกิดค่าความผิดพลาดในการจำแนกที่น้อยที่สุด หลักการสำคัญของวิธี Random Forest คือการสร้างต้นไม้ตัดสินใจอย่างง่ายจำนวนมากในขั้นตอนการฝึกฝน และใช้การลงคะแนนเสียงข้างมากในขั้นตอนการจำแนก ซึ่งการลงคะแนนเสียงนี้จะช่วยแก้ไขคุณสมบัติที่ไม่พึงประสงค์ของโครงสร้างการตัดสินใจให้เหมาะกับชุดสอน จะได้ค่าตัดสินใจสุดท้าย $\hat{f}$ ดังสมการที่ 9 [9]

ในขั้นตอนฝึกฝนตัวจำแนก Random Forest จะใช้เทคนิคที่เรียกว่า Bagging (Bootstrap Aggregation) คือการสร้างข้อมูลย่อยหลายชุดจากชุดสอนด้วยวิธีสุ่มแบบใส่คืน (Sampling with

Replacement) และสร้างต้นไม้ตัดสินใจหลาย ๆ ต้นจากชุดข้อมูลดังกล่าว ซึ่งต้นไม้แต่ละต้นจะไม่มีการตัดแต่ง (pruning) โดยการสุ่มข้อมูลแต่ละครั้งจะมีข้อมูลส่วนหนึ่งที่ไม่ถูกเลือก เรียกว่า ข้อมูลนอกถุง (out-of-bag) ซึ่งเป็นข้อมูลที่สามารถนำไปใช้ตรวจสอบความแม่นของต้นไม้ตัดสินใจแต่ละต้นได้จากการคำนวณค่าความผิดพลาดของข้อมูลนอกถุง (out-of-bag error)

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b(\boldsymbol{x}') \tag{9}$$

เมื่อ B แทน จำนวนต้นไม้

$\hat{f}_b$ แทน ฟังก์ชันตัดสินใจของต้นไม้ต้นที่ b

$\boldsymbol{x}'$ แทน ชุดสอนตัวอย่าง

ในการสร้างต้นไม้จะต้องอาศัยเกณฑ์ในการแบ่งต้นไม้ที่มีอยู่ 2 เกณฑ์หลัก ได้แก่ ความไม่บริสุทธิ์ของจีนี และเอนโทรปี โดยมีสูตรคำนวณตามสมการที่ 10 และ 11 ตามลำดับ

$$Gini: I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) \tag{10}$$

$$Entropy: I_E(f) = -\sum_{i=1}^{m} f_i \, log_2 f_i \tag{11}$$

เมื่อ f_i แทน ความน่าจะเป็นที่จะได้ข้อมูลที่อยู่ในคลาส i

m แทน จำนวนคลาสทั้งหมด

2.1.5 Support Vector Machines (SVM)

SVM เป็นขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนที่ได้รับความนิยมในงานการจำแนกประเภทข้อความเป็นอย่างมาก เนื่องจาก SVM สามารถทำงานได้ดีกับข้อมูลที่มีขนาดมิติสูงหรือข้อมูลที่มีขนาดมิติมากกว่าจำนวนตัวอย่าง อีกทั้งทำงานได้ดีกับข้อมูลที่ไม่มีโครงสร้าง (unstructured data) และข้อมูลกึ่งโครงสร้าง (semi-structured data) เช่น ข้อมูลประเภทข้อความ รูปภาพ เป็นต้น [10] SVM แบบดั้งเดิมนั้นมีไว้สำหรับแก้ปัญหาการจำแนกแบบไบนารี แต่ไม่ใช่กับการจำแนกแบบหลายคลาส สำหรับการจำแนกแบบหลายคลาสจะใช้หลักการแยกปัญหาการจำแนกแบบหลายคลาสออกเป็นปัญหาการจำแนกแบบไบนารีหลายปัญหา สามารถทำได้ 2 วิธี คือ กลยุทธ์หนึ่งต่อส่วนที่เหลือ (One-vs-Rest: OvR) และกลยุทธ์แบบหนึ่งต่อหนึ่ง (One-vs-One: OvO) สำหรับกลยุทธ์ OvR เริ่มจากแบ่งปัญหาหลายคลาสออกเป็นปัญหาไบนารีหลาย ๆ ปัญหา จากนั้นจึงใช้การจำแนกแบบไบนารีและเปรียบเทียบแต่ละคลาสกับส่วนที่เหลือ ขณะที่กลยุทธ์ OvO เป็นการเปรียบเทียบแต่ละคู่คลาส [11]

SVM อาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นหรือระนาบ (Hyperplane) สำหรับแบ่งกลุ่มข้อมูลออกจากกัน ซึ่งสามารถใช้การแบ่งด้วยสมการเชิงเส้น ดังแสดง

ในรูปที่ 2.1 ระนาบแบ่งที่เหมาะสมที่สุดคือระนาบที่ทำให้เกิดระยะขอบ (margin) ระหว่างสองคลาสมากที่สุด โดยจะเลือกระนาบจากการดูระยะห่างที่มีค่าสูงที่สุดระหว่างระนาบแบ่งไปยังจุดข้อมูลที่ใกล้ที่สุดของคลาสทั้งสอง และเรียกจุดข้อมูลที่ใกล้ที่สุดนี้ว่าเวกเตอร์สนับสนุน (support vector) ในปริภูมิ m มิติระนาบแบ่ง [12] ดังสมการที่ 12

$$\boldsymbol{w}^T\boldsymbol{x} + b = 0 \tag{12}$$

เมื่อ $\boldsymbol{w}$ แทน เวกเตอร์ถ่วงน้ำหนัก

$\boldsymbol{x}$ แทน เวกเตอร์คุณลักษณะนำเข้า

b แทน ค่า bias

โดยคำนวณหาระยะขอบได้ดังสมการที่ 13

$$Maximal\ margin = \frac{2}{\|\boldsymbol{w}\|} \tag{13}$$

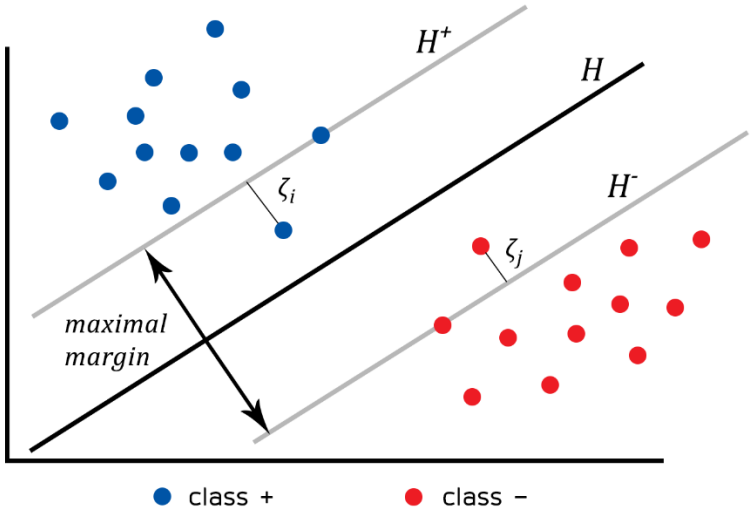
ดังนั้น ในการหาระยะขอบที่มากที่สุดสามารถหาได้จากขนาดของเวกเตอร์ถ่วงน้ำหนัก ($\|\boldsymbol{w}\|$) ที่น้อยที่สุด ดังสมการที่ 14

$$minimize\ \Phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{14}$$

โดยที่ $y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1;\ i = 1,2,\dots,n$

เมื่อ y_i แทน คลาสข้อมูล; โดยมีเงื่อนไขดังนี้

1. $y_i = +1$ สำหรับข้อมูล; คลาสบวก
2. $y_i = -1$ สำหรับข้อมูล; คลาสลบ



รูปที่ 2.1 การแยกข้อมูลสองคลาสโดยใช้ระนาบแบ่งใน SVM

จากรูปที่ 2.1 แสดงให้เห็นว่าแบบจำลองประกอบด้วยระนาบแบ่ง 3 ระนาบ ได้แก่ ระนาบ H คือระนาบที่อยู่ตรงกลาง มีสมการคือ $\boldsymbol{w}^T\boldsymbol{x} - b = 0$ และระนาบ H^+, H^- คือระนาบที่อยู่เหนือและใต้ระนาบ H เรียกว่า supporting plane ซึ่งได้จากการเพิ่มระยะขอบของระนาบ H มีสมการเป็น $\boldsymbol{w}^T\boldsymbol{x} + b = +1$ และ $\boldsymbol{w}^T\boldsymbol{x} + b = -1$ ตามลำดับ

ซึ่งโดยทั่วไปแล้วไม่สามารถใช้วิธีจำแนกแบบเชิงเส้นเพื่อแบ่งข้อมูลให้ถูกต้องได้ทั้งหมด ดังนั้นนอกจากการหาระยะขอบที่มากที่สุดแล้ว ยังจำเป็นต้องหาค่าความผิดพลาดที่น้อยที่สุดด้วย ซึ่งสามารถทำได้โดยอาศัยตัวแปรส่วนขาด หรือ slack variable (ζ)

Slack variable (ζ) คือ ระยะห่างระหว่างจุดข้อมูลกับ supporting plane (ระนาบ H^+ หรือ H^-) ดังนั้นเมื่อต้องการระนาบแบ่งที่ให้ค่าความผิดพลาดที่น้อยที่สุด จะต้องทำให้ ζ มีค่าน้อย ๆ หรือใกล้เคียงศูนย์ เพื่อลดผลรวมของระยะห่างระหว่าง supporting plane กับจุดข้อมูลที่ถูกจำแนกผิด ดังนั้นสามารถคำนวณหาขนาดของเวกเตอร์ถ่วงน้ำหนักที่น้อยที่สุด ดังสมการที่ 15

$$minimize\ \Phi(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_i \zeta_i \tag{15}$$

โดยที่ $\quad y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1 - \zeta_i;\ i = 1,2,\dots,n$

เมื่อ C คือ ค่าที่กำหนดเพื่อทำให้เป็นมาตรฐาน (Regularization) หาก C มีค่าน้อย ๆ ระยะขอบของระนาบแบ่งจะลดลง ทำให้ตัวจำแนกมีความซับซ้อนมากขึ้น และถ้าหาก C มีค่ามาก ๆ ระยะขอบของระนาบแบ่งจะเพิ่มขึ้น ทำให้ตัวจำแนกมีความเรียบง่ายมากขึ้น

สำหรับการแปลงข้อมูลนำเข้าให้อยู่ในรูปแบบของข้อมูลการประมวลผลที่ต้องการ จะใช้ฟังก์ชันที่เรียกว่า เคอร์เนล [13] ซึ่งทำหน้าที่แปลงชุดสอนเพื่อให้เส้นแบ่งไม่เป็นฟังก์ชันเชิงเส้น และสามารถเปลี่ยนเป็นสมการเชิงเส้นในปริภูมิมิติที่สูงขึ้นได้ โดยประเภทของเคอร์เนลที่พบมากที่สุด ได้แก่ เคอร์เนลพหุนาม เคอร์เนลซิกมอยด์ และเคอร์เนลฐานแนวรัศมี

เคอร์เนลพหุนาม (Polynomial Kernel) แสดงถึงความคล้ายคลึงกันของเวกเตอร์ในชุดสอนในปริภูมิคุณลักษณะเหนือพหุนามของตัวแปรดั้งเดิม โดยจะดูจากการรวมกันของคุณลักษณะนำเข้า เรียกว่า คุณลักษณะการโต้ตอบ สามารถคำนวณได้จากผลคูณเชิงสเกลาร์ยกกำลังด้วยดีกรีของเคอร์เนล d ดังสมการที่ 16

$$K(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{y} + c)^d \tag{16}$$

เมื่อ $\quad \boldsymbol{x}, \boldsymbol{y}$ แทน เวกเตอร์ในปริภูมิข้อมูล

$\quad\quad d$ แทน ดีกรีของเคอร์เนล

เคอร์เนลฟังก์ชันฐานแนวรัศมี (Radial Basis Function Kernel: RBF) เป็นฟังก์ชันค่าจริงซึ่งคำนวณจากการหาระยะห่างแบบ Euclidean ระหว่างจุดสังเกตสองจุด ใช้กับข้อมูลที่ไม่เป็นเชิงเส้น เป็นเคอร์เนลได้รับความนิยมมากที่สุดที่ใช้ในการจำแนกประเภทใน SVM สามารถคำนวณหาขอบเขตข้อมูลได้จากสมการที่ 17, 18

$$K(\boldsymbol{x}, \boldsymbol{y}) = exp\ (-\Upsilon \|\boldsymbol{x} - \boldsymbol{y}\|^2) \tag{17}$$

$$\Upsilon = \frac{1}{2\sigma^2} \tag{18}$$

เมื่อ σ แทน ส่วนเบี่ยงเบนมาตรฐาน

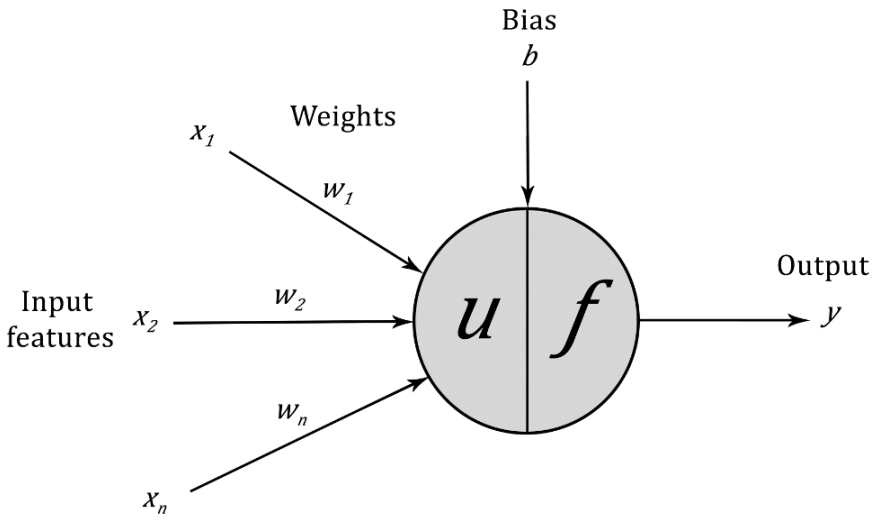
คอร์เนลซิกมอยด์ (Sigmoid Kernel) มีต้นกำเนิดมาจากฟังก์ชันกระตุ้นของโครงข่ายประสาทเทียม ซึ่งเทียบเท่ากับแบบจำลองเพอร์เซปตรอนสองชั้น แสดงดังสมการที่ 19

$$K(\boldsymbol{x}, \boldsymbol{y}) = tanh\ (\alpha \boldsymbol{x}^T \boldsymbol{y} + c) \tag{19}$$

เมื่อ α แทน ค่าสัมประสิทธิ์

2.1.6 Multi-Layer Perceptron (MLP)

โครงข่ายประสาทเทียมเพอร์เซปตรอนแบบหลายชั้น (Multi-Layer Perceptron: MLP) [14] เป็นส่วนเสริมของโครงข่ายประสาทแบบป้อนไปข้างหน้า และเป็นโครงสร้างของข่ายงานระบบประสาทอย่างง่ายที่ได้รับความนิยม [15] โดย MLP ประกอบด้วยเซลล์ประสาทที่เรียกว่าเพอร์เซปตรอน ซึ่งโครงสร้างทั่วไปของเพอร์เซปตรอน แสดงดังรูปที่ 2.2



รูปที่ 2.2 โครงสร้างของเพอร์เซปตรอนที่มี n คุณลักษณะนำเข้า

จากรูปที่ 2.2 เพอร์เซปตรอนรับ n คุณลักษณะเป็นข้อมูลนำเข้า ซึ่งแต่ละคุณลักษณะจะมีค่าน้ำหนักที่กำหนดไว้ โดยคุณลักษณะนำเข้าต้องเป็นข้อมูลประเภทตัวเลข ซึ่งหากไม่ใช่ข้อมูลประเภทตัวเลข จะต้องถูกแปลงให้เป็นตัวเลขก่อนเสมอเพื่อให้สามารถใช้เพอร์เซปตรอนได้ ตัวอย่างเช่น คุณลักษณะที่เป็นหมวดหมู่ที่มีจำนวนค่าที่เป็นไปได้เท่ากับ p สามารถแปลงเป็นคุณลักษณะนำเข้าจำนวน p คุณลักษณะ เพื่อใช้แทนการ ‘มี’ หรือ ‘ไม่มี’ คุณลักษณะดังกล่าว โดยเรียกตัวแปรนี้ว่าตัวแปรหุ่น (dummy variables)

คุณลักษณะนำเข้าจะถูกส่งต่อไปยังฟังก์ชันนำเข้า u แทนด้วย $u(\boldsymbol{x})$ เพื่อคำนวณผลรวมคุณลักษณะนำเข้าที่ถ่วงน้ำหนัก ดังสมการที่ 20

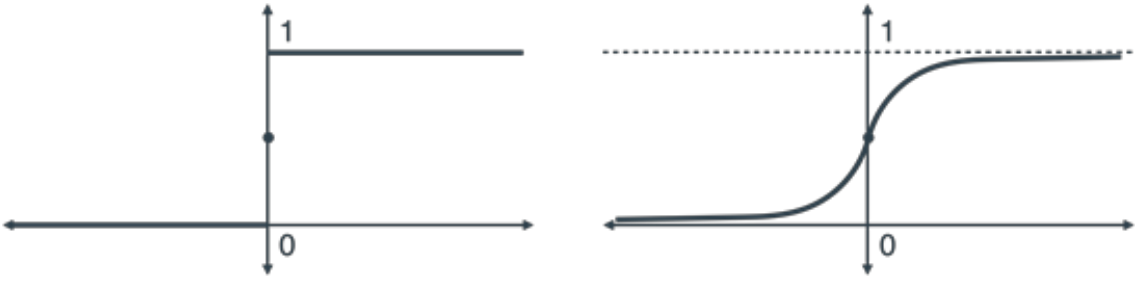
$$u(\boldsymbol{x}) = \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{x}_i + b \tag{20}$$

เมื่อ b แทน ค่า bias

หลังจากนั้นผลลัพธ์ของการคำนวณจะถูกส่งต่อไปยังฟังก์ชันกระตุ้น f เพื่อสร้างข้อมูลนำออกของเพอร์เซปตรอน และสำหรับในเพอร์เซปตรอนแบบดั้งเดิม ฟังก์ชันกระตุ้นนี้จะอยู่ในรูปของฟังก์ชันขั้นบันได โดยมี θ เป็นเกณฑ์ขั้นต่ำ ดังสมการที่ 21

$$y = f\big(u(\boldsymbol{x})\big) = \begin{cases} 1, & u(\boldsymbol{x}) > \theta \\ 0, & u(\boldsymbol{x}) \leq \theta \end{cases} \tag{21}$$

ตัวอย่างเช่น ฟังก์ชันขั้นบันไดที่มี $\theta = 0$ แสดงดังรูปที่ 2.3 (a) ดังนั้นจะได้ว่าเพอร์เซปตรอนจะมีค่าเป็นจริงหรือเท็จ สามารถดูได้จาก $u(\boldsymbol{x}) - \theta > 0$ ทำให้ได้สมการเส้นแบ่ง (hyperplane) คือ $u(\boldsymbol{x}) - \theta = 0$ โดยที่เพอร์เซปตรอนจะส่งออกค่า 1 สำหรับจุดรับเข้าใด ๆ ที่อยู่เหนือเส้นแบ่ง และส่งออกค่า 0 สำหรับจุดใด ๆ ที่อยู่บนหรือใต้เส้นแบ่ง ดังนั้นเพอร์เซปตรอนจึงเป็นตัวจำแนกแบบเชิงเส้น กล่าวคือสามารถทำงานได้ดีกับข้อมูลที่แยกกันในรูปแบบเชิงเส้น จะเห็นได้ว่าการเรียนรู้ของเพอร์เซปตรอนสามารถทำได้โดยปรับค่าน้ำหนักเพื่อหาเส้นแบ่งที่สามารถแยกชุดสอนได้ดี



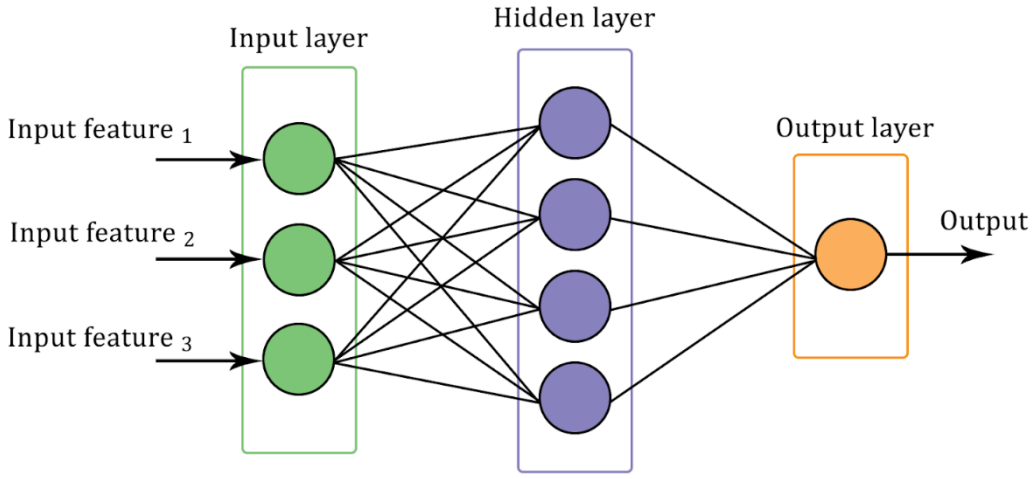
(a) ตัวอย่างฟังก์ชันขั้นบันได (b) ตัวอย่างฟังก์ชันซิกมอยด์

รูปที่ 2.3 ตัวอย่างฟังก์ชันกระตุ้น

MLP เป็นการรวมเซลล์ประสาทหลาย ๆ เซลล์เข้าด้วยกัน จึงทำให้ MLP มีความสามารถในการประมาณค่าฟังก์ชันต่อเนื่องได้ โดย MLP มีโครงสร้างที่ประกอบด้วยชั้นอย่างน้อย 3 ชั้น ได้แก่ ชั้นนำเข้า (Input Layer) ชั้นซ่อน (Hidden Layer) และชั้นส่งออก (Output Layer) โดยในแต่ละชั้นมีรายละเอียดดังนี้

- ชั้นนำเข้า จำนวน 1 ชั้น ทำหน้าที่ส่งคุณลักษณะนำเข้าไปยังชั้นซ่อน ชั้นแรก
- ชั้นซ่อน จำนวนอย่างน้อย 1 ชั้น เป็นชั้นที่อยู่ตรงกลาง โดยชั้นซ่อนชั้นแรกทำหน้าที่รับคุณลักษณะนำเข้าจากชั้นนำเข้า และชั้นซ่อนชั้นอื่น ๆ ทำหน้าที่รับข้อมูลส่งออกจากแต่ละเพอร์เซปตรอนของชั้นก่อนหน้า
- ชั้นส่งออก จำนวน 1 ชั้น ทำหน้าที่รับข้อมูลส่งออกจากแต่ละเพอร์เซปตรอนของชั้นซ่อนชั้นสุดท้าย

ตัวอย่างโครงสร้างของ MLP สามชั้น ที่มีชั้นนำเข้ามีคุณลักษณะนำเข้า 3 คุณลักษณะ ชั้นซ่อนมีเซลล์ประสาท 4 โหนด และชั้นส่งออกมีข้อมูลส่งออก 1 โหนด แสดงดังรูปที่ 2.4



รูปที่ 2.4 โครงสร้างของ MLP สามชั้น

เพอร์เซปตรอนของ MLP ส่วนใหญ่มักใช้ฟังก์ชันกระตุ้นอื่น ๆ ที่ไม่ใช่ฟังก์ชันขั้นบันได ซึ่งฟังก์ชันซิกมอยด์ (sigmoid) เป็นฟังก์ชันที่นิยมใช้ในเซลล์ประสาทชั้นซ่อน ตัวอย่างของฟังก์ชันซิกมอยด์ แสดงในรูปที่ 2.3 (b)

โดยมีการคำนวณเซลล์ประสาทส่งออกและเซลล์ประสาทซ่อน ตามสมการที่ 22 และ 23 ตามลำดับ ดังนี้

$$o(\boldsymbol{x}) = G(b_2 + \boldsymbol{w}_2 \cdot h(\boldsymbol{x})) \tag{22}$$

$$h(\boldsymbol{x}) = s(b_1 + \boldsymbol{w}_1 \cdot \boldsymbol{x}) \tag{23}$$

เมื่อ $\boldsymbol{w}$ แทน เวกเตอร์น้ำหนัก

$G(\cdot), s(\cdot)$ แทน ฟังก์ชันกระตุ้น

โดยที่ $G(\cdot), s(\cdot)$ สามารถเป็นได้หนึ่งในฟังก์ชัน ต่อไปนี้

ฟังก์ชัน hyperbolic tangent (tanh) โดยที่ $f(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})}$ หรือ

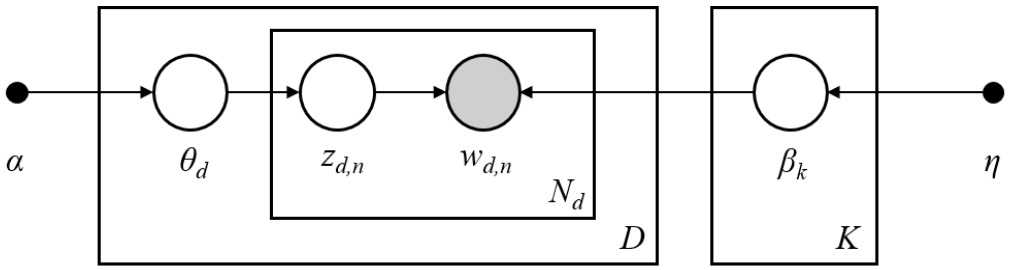
ฟังก์ชัน logistic sigmoid (sigmoid) โดยที่ $f(x) = \frac{1}{(1+e^{-x})}$ หรือ

ฟังก์ชัน rectified linear unit (relu) โดยที่ $f(x) = max\ (0, x)$

ในการเรียนรู้ของ MLP จะมีการปรับค่าน้ำหนักของ เพอร์เซปตรอนเพื่อลดความผิดพลาดของผลลัพธ์จากชุดสอน โดยปกติแล้วจะมีการใช้ขั้นตอนวิธีแบบแพร่ย้อนกลับ (backpropagation) ที่พยายามลดค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error: MSE) ให้น้อยที่สุด

2.1.7 Latent Dirichlet Allocation (LDA)

การจัดสรรของดีรีเคลท์แฝง (Latent Dirichlet Allocation: LDA) [16] เป็นขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอนที่ใช้ในการสร้างแบบจำลองหัวข้อหรือแบบจำลองความน่าจะเป็นโดยกำเนิด (generative probabilistic model) ที่ใช้ในการทำเหมืองข้อความสำหรับประมวลผลภาษาธรรมชาติ นอกจากนี้ยังสามารถใช้ในงานแยกคุณลักษณะ เช่น การจัดกลุ่มเอกสาร ได้อีกด้วย [17] โดยใช้ความน่าจะเป็นในการอธิบายหัวข้อของการกระจายหัวข้อในแต่ละเอกสาร และการกระจายของคำในแต่ละหัวข้อ อธิบายกระบวนการต่าง ๆ ใน LDA ด้วยแบบจำลองเชิงรูปภาพ ได้ดังรูปที่ 2.5



รูปที่ 2.5 แบบจำลองเชิงรูปภาพของ LDA

จากรูปที่ 2.5 แสดงโหนดต่าง ๆ โดยแต่ละโหนดเป็นตัวแปรสุ่มและมีบทบาทในกระบวนการกำเนิด โหนดสีเทาแสดงถึงตัวแปรที่สังเกตได้ (observed variable) และโหนดสีขาวแสดงถึงตัวแปรที่ซ่อนอยู่หรือตัวแปรแฝง (latent variable) ดังนั้นมีตัวแปรคำ (w) เพียงตัวแปรเดียวที่สังเกตได้ ส่วนตัวแปรแฝงจะเป็นตัวกำหนดการรวมแบบสุ่มของหัวข้อในชุดข้อมูลและการกระจายของคำในเอกสาร เป้าหมายของ LDA คือการใช้คำที่สังเกตได้เพื่อสรุปโครงสร้างหัวข้อที่ซ่อนอยู่ [18]

รายละเอียดตัวแปรมีดังนี้

ให้ α แทน ตัวควบคุมสัดส่วนการกระจายของหัวข้อในแต่ละเอกสาร

θ_d แทน ความน่าจะเป็นของหัวข้อในเอกสาร d

$z_{d,n}$ แทน หัวข้อของคำในเอกสาร d

$w_{d,n}$ แทน คำที่ปรากฏในเอกสาร d มีทั้งหมด n คำ

β_k แทน ความน่าจะเป็นในการปรากฏของคำในหัวข้อ k

η แทน ตัวควบคุมสัดส่วนการกระจายของคำในแต่ละหัวข้อ

อธิบายกระบวนการเกิดได้ดังนี้

1. ในแต่ละหัวข้อ K หัวข้อ สุ่มหยิบ β_k ด้วยวิธีการกระจายแบบดีรีเคลท์ (dirichlet distribution) โดยมี η เป็นตัวควบคุม จะได้เมทริกซ์ความน่าจะเป็นในการปรากฏของคำในแต่ละหัวข้อ
2. ในแต่ละเอกสาร D เอกสาร สุ่มหยิบ θ_d ด้วยวิธีการกระจายแบบดีรีเคลท์โดยมี α เป็นตัวควบคุม จะได้เมทริกซ์ความน่าจะเป็นของหัวข้อในแต่ละเอกสาร
3. สำหรับในแต่ละคำ คำที่ w_d ในเอกสาร d
    3.1 สุ่มหยิบหัวข้อ z_d ด้วยวิธีการแจกแจงแบบอเนกนาม (multinomial distribution) ของ θ_d
    3.2 สุ่มหยิบคำที่ w_d ด้วยวิธีการแจกแจงแบบอเนกนาม (multinomial distribution) ในหัวข้อที่สุ่มได้จากข้อ 3.1

ทำการอนุมานจากสมมติฐานในการสร้างเอกสาร โดยใช้ข้อมูลที่มีอยู่เพื่อเรียนรู้ตัวแปรที่ต้องการ จากการคำนวณค่าการแจกแจงภายหลัง (posterior distribution) โดยหาความน่าจะเป็นร่วมของหัวข้อ (z) การกระจายหัวข้อในเอกสาร (θ) และการกระจายคำในหัวข้อ (β)

เมื่อนำคำใด ๆ มาพิจารณาจะมีค่าเท่ากับความน่าจะเป็นร่วมของหัวข้อ (z) คำ (w) คำนั้นที่จะปรากฏ (w_d) การกระจายหัวข้อในเอกสาร (θ) และการกระจายคำในหัวข้อ (β) หารด้วยความน่าจะเป็นของคำนั้น ๆ ดังสมการที่ 24

$$p(z,\theta,\beta|w,\alpha,\eta) = \frac{p(z,\theta,\beta|\alpha,\eta)}{p(w|\alpha,\eta)} \tag{24}$$

แต่ในความเป็นจริงจะไม่สามารถคำนวณได้โดยตรง จึงต้องใช้ขั้นตอนวิธีอื่น ๆ ในการประมาณค่า เช่น การประมาณแบบลาปลาซ (Laplace approximation) การประมาณแบบแปรผัน (variational approximation) เป็นต้น

2.1.8 เมทริกซ์ความสับสน (Confusion Matrix)

เมทริกซ์ความสับสน [19] เป็นเครื่องมือที่ใช้สำหรับประเมินประสิทธิภาพการจำแนกซึ่งเป็นที่นิยมในการรายงานผลลัพธ์ของปัญหาการจำแนกประเภทไบนารี และการจำแนกประเภทหลายคลาส [20] สำหรับเมทริกซ์ความสับสนสำหรับการจำแนกประเภทแบบไบนารีในตารางที่ 2.1 บอกถึงค่าต่าง ๆ ดังนี้

1. ค่าผลบวกจริง (True Positive: TP) คือจำนวนตัวอย่างที่ถูกทำนายว่าเป็นคลาสบวก และมีป้ายกำกับเป็นคลาสบวก
2. ค่าผลลบจริง (True Negative: TN) คือจำนวนตัวอย่างที่ถูกทำนายว่าเป็นคลาสลบ และมีป้ายกำกับเป็นคลาสลบ
3. ค่าผลบวกเท็จ (False Positive: FP) คือจำนวนตัวอย่างที่ถูกทำนายว่าเป็นคลาสบวก แต่มีป้ายกำกับเป็นคลาสลบ
4. ค่าผลลบเท็จ (False Negative: FN) คือจำนวนตัวอย่างที่ถูกทำนายว่าเป็นคลาสลบ แต่มีป้ายกำกับเป็นคลาสบวก

ตารางที่ 2.1 เมทริกซ์ความสับสนสำหรับการจำแนกประเภทแบบไบนารี

		Predicted	
		Positive	Negative
Actual	Positive	TP	FN
	Negative	FP	TN

จากตารางที่ 2.1 แสดงเมทริกซ์ความสับสนสำหรับการจำแนกประเภทแบบไบนารี ค่าบนเส้นทแยงมุม หมายถึงจำนวนตัวอย่างที่แบบจำลองทำนายได้ถูกต้อง (TP+TN) ได้แก่ จำนวนตัวอย่างที่ถูกทำนายว่าเป็นคลาสบวกและมีป้ายกำกับเป็นคลาสบวก (TP) และจำนวนตัวอย่างที่ถูกทำนายว่าเป็นคลาสลบและมีป้ายกำกับเป็นคลาสลบ (TN) ส่วนค่าที่อยู่นอกเหนือจากเส้นทแยงมุม หมายถึงจำนวนตัวอย่างที่ตัวจำแนกทำนายผิด ซึ่งมีได้สองกรณี คือผลการทำนายเป็นคลาสลบแต่มีป้ายกำกับเป็นคลาสบวก (FN) และผลการทำนายเป็นคลาสบวกแต่มีป้ายกำกับเป็นคลาสลบ (FP)

ในการประเมินประสิทธิภาพจะพิจารณาจากหลาย ๆ ค่าประกอบกัน [20] [21] หากพิจารณาค่าความแม่นเพียงอย่างเดียว อาจทำให้การประเมินบางส่วนผิดพลาดไป สำหรับชุดข้อมูลที่ไม่สมดุลกัน ดังนั้นจึงได้ใช้เครื่องมือประเมินประสิทธิภาพอื่น ๆ อย่างเมทริกซ์ความสับสน ค่าความเที่ยง

ค่าความครบถ้วน และคะแนน F1 มาช่วยประเมินประสิทธิภาพ เพื่อช่วยลดความผิดพลาดในการประเมิน โดยมีรายละเอียดดังนี้

1. ค่าความแม่น (Accuracy)

ค่าความแม่นเป็นการประเมินแบบหนึ่งที่นิยมใช้มากที่สุดสำหรับการประเมินประสิทธิภาพการจำแนกประเภท โดยคำนวณจากผลรวมของตัวเลขบนเส้นทแยงมุมในเมทริกซ์ความสับสนหารด้วยจำนวนตัวอย่างทั้งหมด กล่าวคืออัตราส่วนระหว่างจำนวนตัวอย่างที่จำแนกได้ถูกต้องต่อจำนวนตัวอย่างทั้งหมด ดังสมการที่ 25

$$Accuracy = \frac{TN+TP}{TN+FP+FN+TP} \tag{25}$$

2. ค่าความเที่ยง (Precision)

ค่าความเที่ยงหรือค่าทำนายผลบวก (Positive Predictive Value: PPV) คืออัตราส่วนระหว่างผลทำนายเป็นจริงและป้ายกำกับเป็นจริง (TP) ต่อจำนวนตัวอย่างที่ผลทำนายเป็นจริงทั้งหมด ดังสมการที่ 26

$$PPV = Precision = \frac{TP}{FP+TP} \tag{26}$$

3. ค่าความครบถ้วน (Recall)

ค่าความครบถ้วนหรืออัตราผลบวกจริง (True Positive Rate: TPR) คือจำนวนตัวอย่างที่เป็นบวกที่ทำนายได้ถูกต้องเทียบกับจำนวนตัวอย่างที่เป็นบวกทั้งหมด แสดงดังสมการที่ 27

$$TPR = Recall = \frac{TP}{TP+FN} = \frac{TP}{P} \tag{27}$$

4. คะแนน F1 (F1-score)

คะแนน F1 คือค่าเฉลี่ยฮาร์โมนิก (Harmonic mean) ของ Precision และ Recall เมื่อคะแนน F1 มีค่าสูง หมายความว่ามีประสิทธิภาพในการจำแนกประเภทสูง แสดงดังสมการที่ 28

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \tag{28}$$

2.2 งานวิจัยที่เกี่ยวข้อง

จากการศึกษาเรื่องของการจำแนกประเภทหัวข้อ โดยใช้คุณลักษณะที่สกัดได้จากข้อความ พบว่ามีงานวิจัยที่เกี่ยวข้องดังต่อไปนี้

งานวิจัยของ Arisara Noppakaow และคณะ [22] ได้ศึกษาเกี่ยวกับแบบจำลองอัตโนมัติเพื่อจำแนกบทความข่าวภาษาไทยออกเป็น 4 ประเภท ได้แก่ ข่าวอาชญากรรม ข่าวการเมือง ข่าวกีฬา และข่าวบันเทิง ด้วยขั้นตอนวิธี Decision Tree, Support Vector Machine (SVM) และ Multilayer Perceptron (MLP) มีความแม่นอยู่ที่ 86% 94% และ 95% ตามลำดับ

งานวิจัยของ Wiphada Jirasirilerd และ Pikulkaew Tangtisanon [23] ได้ศึกษาเกี่ยวกับวิธีการติดป้ายกำกับอัตโนมัติสำหรับบทความข่าวบนเว็บไซต์ภาษาไทยโดยใช้การกระจายข้อความในเอกสาร โดยสกัดคำที่มีความหมายคล้ายกันจากเวกเตอร์ย่อหน้าของข่าวแต่ละประเภท และนำไปกำหนดเป็นป้ายกำกับ เนื่องจากโปรแกรมแบ่งกลุ่มคำภาษาไทยที่มีอยู่นั้นให้อัตราความแม่นที่ต่ำ จึงได้ใช้โครงข่ายประสาทแบบคอนโวลูชันร่วมกับวิธีการจำแนกแบบไบนารีในการแยกคำออกจากประโยคเพื่อประสิทธิภาพที่ดีขึ้น และพบว่าแบบจำลองเวกเตอร์ที่นำเสนอให้ความแม่นที่ดีกว่าแบบจำลองเวกเตอร์อื่น ๆ

งานวิจัยของ Seonggyu Lee และคณะ [24] ได้เพิ่มประสิทธิภาพของการจำแนกประเภทเอกสาร โดยนำเสนอวิธีการที่พัฒนามาจากพื้นฐานของ Latent Dirichlet Allocation (LDA) พร้อมกับพิจารณาน้ำหนักของคำในการสุ่มตัวอย่างและเพิ่มความสมดุลในการกระจายหัวข้อ โดยทดลองกับชุดข้อมูล 20 Newsgroups ซึ่งเป็นชุดข้อมูลที่รวบรวมเอกสารกลุ่มข่าวไว้จำนวน 20 ประเภท ผลการทดลองแสดงให้เห็นว่าการสร้างแบบจำลองหัวข้อแบบถ่วงน้ำหนักสมดุล (Balance Weighted Topic Modeling) ทำให้ได้คุณลักษณะที่ช่วยให้การจำแนกประเภทเอกสารมีประสิทธิภาพดีขึ้น

งานวิจัยของ D. E. Cahyani และ K. A. P. Nuzry [25] ได้ศึกษาเกี่ยวกับการจำแนกหัวข้อที่เป็นที่นิยมบนทวิตเตอร์ทั้งหมด 210 หัวข้อ แบ่งเป็น 6 ประเภท ได้แก่ การเมือง กีฬา ความบันเทิง การท่องเที่ยว ธุรกิจ และข่าวอื่น ๆ โดยทดลองกับจำนวนทวีตที่แตกต่างกัน ได้แก่ 30, 100, 200 และ 500 ทวีตต่อหัวข้อ และสร้างตัวจำแนกโดยใช้ขั้นตอนวิธี Multinomial Naive Bayes (MNB) สำหรับงานที่มีป้ายกำกับเดียว (single-label) ซึ่งได้ความแม่นสูงสุดอยู่ที่ 82.53% กับข้อมูลจำนวน 500 ทวีตต่อหัวข้อ และใช้ขั้นตอนวิธี K-Nearest Neighbor (KNN) สำหรับงานที่มีป้ายกำกับหลายประเภท (multi-label) ซึ่งได้ความแม่นสูงสุดอยู่ที่ 88.05% กับข้อมูลจำนวน 500 ทวีตต่อหัวข้อ

งานวิจัยของ Awet Fesseha และคณะ [26] ได้ศึกษาเกี่ยวกับการจำแนกประเภทบทความข่าวภาษาทิกริญญา (Tigrigna) และผู้วิจัยได้สร้างชุดข้อมูลใหม่จากแหล่งข่าวต่าง ๆ ของทิกริญญา โดยแบ่งออกเป็น 6 ประเภท ได้แก่ เกษตรกรรม กีฬา สุขภาพ การศึกษา ศาสนา และการเมือง แล้วใช้ตัวจำแนกการเรียนรู้ของเครื่องที่แตกต่างกันเพื่อตรวจสอบประสิทธิภาพในชุดข้อมูลใหม่ รวมทั้งหมด 7 ตัวจำแนกที่เป็น

ที่นิยม ได้แก่ Logistic Regression, Nearest Centroid, Decision Tree (DT), Support Vector Machines (SVM), K-nearest neighbors (KNN), Random Forest และ Multi-Layer Perceptron (MLP) นอกจากนี้ยังมีการรวมแบบจำลองเพื่อให้ได้ความแม่นสูงสุดโดยการรวมตัวจำแนกที่ดีที่สุดตามตัวจำแนกเสียงส่วนใหญ่ (majority-voting) ผลการทดลองแสดงให้เห็นว่าตัวจำแนกประเภท SVM ได้ค่าความแม่นสูงที่สุดคือ 96% และตัวจำแนกประเภท Nearest-centroid ได้ค่าความแม่นต่ำที่สุดคือ 89%

งานวิจัยของ Nicolas และ Zach CHASE [27] ได้ศึกษาเกี่ยวกับการจำแนกประเภทหัวข้อของบทความข่าวที่มีป้ายกำกับหลายประเภท และวิเคราะห์ข้อบกพร่องของขั้นตอนวิธีต่าง ๆ รวมถึง Naive Bayes แล้วผู้วิจัยได้นำเสนอตัวจำแนก Naive Bayes แบบหนึ่งต่อทั้งหมด โดยใช้ตัวจำแนกหนึ่งตัวต่อคลาส ซึ่งได้ประสิทธิภาพมากกว่าวิธีการที่ได้จาก Term Frequency–Inverse Document Frequency (TF-IDF)

จากงานวิจัยที่กล่าวมาข้างต้นสังเกตได้ว่ามีการนำข้อความมาสกัดเป็นเวกเตอร์คุณลักษณะ (feature vector) จากนั้นจึงนำไปสร้างแบบจำลองสำหรับจำแนกหัวข้อข่าวโดยใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งจะกล่าวในบทที่ 3 ต่อไป

บทที่ 3
วิธีการวิจัย

ในบทนี้จะกล่าวถึงวิธีการวิจัยในการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าว ซึ่งในงานวิจัยนี้จะแบ่งกระบวนการออกเป็น 4 ขั้นตอน ได้แก่

1. การรวบรวมข้อมูล (Data Collection)
2. การเตรียมข้อมูล (Data Pre-processing)
3. การสกัดคุณลักษณะ (Features Extraction)
4. การหาพารามิเตอร์ที่เหมาะสม (Parameters Optimization)

3.1 การรวบรวมข้อมูล (Data Collection)

ข้อมูลชุดคำโต้ข่าวภาษาไทยที่ใช้ในงานวิจัยนี้ เก็บรวบรวมจากรายการข่าวย้อนหลังบนเว็บไซต์ ซึ่งข้อมูลข่าวแต่ละประเภทจะนำมาจากรายการข่าวที่แตกต่างกันไป โดยแหล่งที่มาของข่าวทั้ง 6 ประเภทมีดังนี้

1. **ข่าวการเมือง** นำมาจากรายการคุยข่าวเช้า ช่อง 8 และรายการคับข่าวครบประเด็น ช่อง MCOT HD
2. **ข่าวเศรษฐกิจ** นำมาจากรายการเรื่องเล่าเช้านี้ และเรื่องเล่าเสาร์อาทิตย์ ช่อง 3 HD
3. **ข่าวอาชญากรรม** นำมาจากรายการเที่ยงวันทันเหตุการณ์ ช่อง 3 HD
4. **ข่าวบันเทิง** นำมาจากรายการวันบันเทิง ช่อง one31 และรายการไนน์เอ็นเตอร์เทน ช่อง MCOT HD
5. **ข่าวกีฬา** นำมาจากรายการ Siamsport Halftime และรายการฟุตบอลไทยวาไรตี้ ช่อง Siamsport
6. **ข่าวสิ่งแวดล้อม** นำมาจากรายการเรื่องเล่าเช้านี้ และเรื่องเล่าเสาร์อาทิตย์ ช่อง 3 HD

3.2 การเตรียมข้อมูล (Data Pre-processing)

หลังจากการรวบรวมชุดข้อมูลคำโต้ตอบข่าวภาษาไทยจากวิดีโอรายการข่าวต่าง ๆ เป็นที่เรียบร้อย ในขั้นตอนถัดมาจะเป็นการเตรียมข้อมูลให้พร้อมสำหรับการนำไปสร้างตัวจำแนก แบ่งออกเป็น 4 ขั้นตอน เริ่มจากแปลงเสียงเป็นข้อความ ติดป้ายกำกับ ทำความสะอาดข้อความ และสุดท้ายคือการตัดคำ ซึ่งแต่ละขั้นตอนมีรายละเอียดดังนี้

3.2.1 การแปลงเสียงเป็นข้อความ (Speech-to-Text: STT)

เริ่มจากการนำข้อมูลเสียงจากวิดีโอรายการข่าวย้อนหลังของแต่ละรายการมาแปลงเสียงเป็นข้อความด้วยวิธี Speech-to-Text (STT) ที่ใช้เทคนิค Speech Recognition ในงานวิจัยนี้อาศัยเครื่องมือ Google Voice Typing ซึ่งเป็นเครื่องมือสำเร็จรูปเป็นเครื่องมือหลักในการทำ STT โดยกำหนดให้วิดีโอมีความยาว 15 วินาทีต่อหนึ่งชุดคำโต้ตอบ คิดเป็น 50 คำต่อหนึ่งชุดคำโต้ตอบ

โดยเฉลี่ย และทำการตรวจสอบความถูกต้องของคำที่ได้หลังจากแปลงเสียงเป็นข้อความ จนได้เป็นชุดข้อมูลประเภทข้อความ

3.2.2 การติดป้ายกำกับ (Labeling)

การติดป้ายกำกับเป็นการแบ่งชุดข้อมูลที่รวบรวมได้ออกเป็นประเภทต่าง ๆ เพื่อนำข้อมูลไปใช้กับการเรียนรู้แบบมีผู้สอน ซึ่งในขั้นตอนนี้จะเป็นการติดป้ายกำกับแบบเดี่ยว โดยจะใช้ผู้ติดป้ายกำกับจำนวน 2 คน เพื่อติดป้ายกำกับให้กับข้อมูลทั้งหมด และหากมีชุดข้อมูลใด ๆ ที่ถูกติดป้ายกำกับต่างกัน จะทำการคัดชุดข้อมูลนั้นออกเพื่อให้ข้อมูลมีความกำกวมน้อยที่สุด เช่น

“ทางศบศ.เคาะออกมาที่ประชุมศบศ. นายกรัฐมนตรี พลเอกประยุทธ์ จันทร์โอชา ท่านมาเป็นประธานการประชุมนะคะ มีการเคาะมาตรการเงินช่วยโควิด จะช่วยผ่านบัตรคนจนค่ะ”

จะเห็นว่าชุดโต้ตอบข้างต้นสามารถติดป้ายกำกับได้ 2 ประเภท คือ เศรษฐกิจและการเมือง จึงต้องตัดชุดข้อมูลนี้ออก

สำหรับตัวอย่างชุดคำโต้ตอบข่าวทั้ง 6 ประเภท ติดป้ายกำกับแบบเดี่ยวได้ดังนี้

- ตัวอย่างชุดคำโต้ตอบข่าวประเภทการเมือง

  “จริงๆต้องระมัดระวังทั้งเรื่องของการใช้เวลา ใช่ และเนื้อหา ค่ะ สาระของการหาเสียง ซึ่งจริงๆไม่ใช่เฉพาะพลเอกประยุทธ์เท่านั้นครับ ทุกพรรคการเมืองก็ต้องระมัดระวังสิ่งเหล่านี้เช่นเดียวกัน ไม่ได้แตกต่างอะไรกันนะครับ รองนายกรัฐมนตรี”

- ตัวอย่างชุดคำโต้ตอบข่าวประเภทเศรษฐกิจ

  “อ่อ มันออโต้ให้เรา มันออโต้ให้เรา ไม่งั้นงงตาย เอ้อ ใช่มะ พี่นิป ใช่ไง ต้องมานั่งคิดอีก เอ๊ะมันเท่าไหร่ คำนวณไม่ถูก ไม่ต้องงงตรงนี้ เพราะมันคำนวณให้เลย และภาระจะตกไปอยู่ที่เจ้าของโรงแรม เจ้าของโรงแรมจะต้องไปเอาเงินจากรัฐบาลเอง คือ 40%”

- ตัวอย่างชุดคำโต้ตอบข่าวประเภทอาชญากรรม

  “ผู้เสียชีวิตด้วยนะครับ ไม่น่าเกิดเรื่องแบบนี้เลย แล้วแทงกันกลางวันแสกๆ อ่า คนอยู่เยอะแยะนะฮะ คือ ค่ะ อุกอาจมาก ค่ะคือแม่ค้าขายลูกชิ้นเนี่ยบอกว่า คือพอผู้ตายกลับมา พอลงจากรถที่เค้า”

- ตัวอย่างชุดคำโต้ตอบข่าวประเภทบันเทิง

  “ไรงี้ครับผม ก็เซ็นต์ไว้ 3 ปี ทางค่าย 0316 Entertainment ตอนนี้ก็ผ่านมา 2 ปีแล้ว ก็มีเพลงอะไรออกมาเรื่อยๆนะครับ แต่ว่าทางเพลงที่เวียดนาม หรือว่าทางผลงานที่เวียดนามเนี่ยยังไม่ ยังไม่มี”

- ตัวอย่างชุดคำโต้ตอบข่าวประเภทกีฬา

  “คราวนี้เรามาอัพเดทผลการแข่งขันของฟุตบอลไทยลีกกันบ้างดีกว่านะครับ ช่วงสุดสัปดาห์ที่ผ่านมา โห สนุกเลยนะมันค่อนข้างที่จะมีสีสันมากๆสำหรับสัปดาห์ที่ผ่านมามาดูครับ สิงห์เชียงรายยูไนเต็ดนี้ก็เฉือนเอาชนะ”

- ตัวอย่างชุดคำโต้ตอบข่าวประเภทสิ่งแวดล้อม

  “แต่ว่าสิ่งหนึ่งที่มาพร้อมกับความหนาวนั่นคือฝุ่น PM 2.5 ค่ะ เมื่อวานนี้พื้นที่กรุงเทพมหานครเนี่ยเห็นได้ชัดเจนนะครับ และใครที่ไวต่ออากาศแบบนี้ภูมิแพ้เนี่ยรู้สึกได้ทันทีนะฮะ มันจะคันในจมูก ผมเนี่ย”

3.2.3 การทำความสะอาดข้อความ (Text Cleaning)

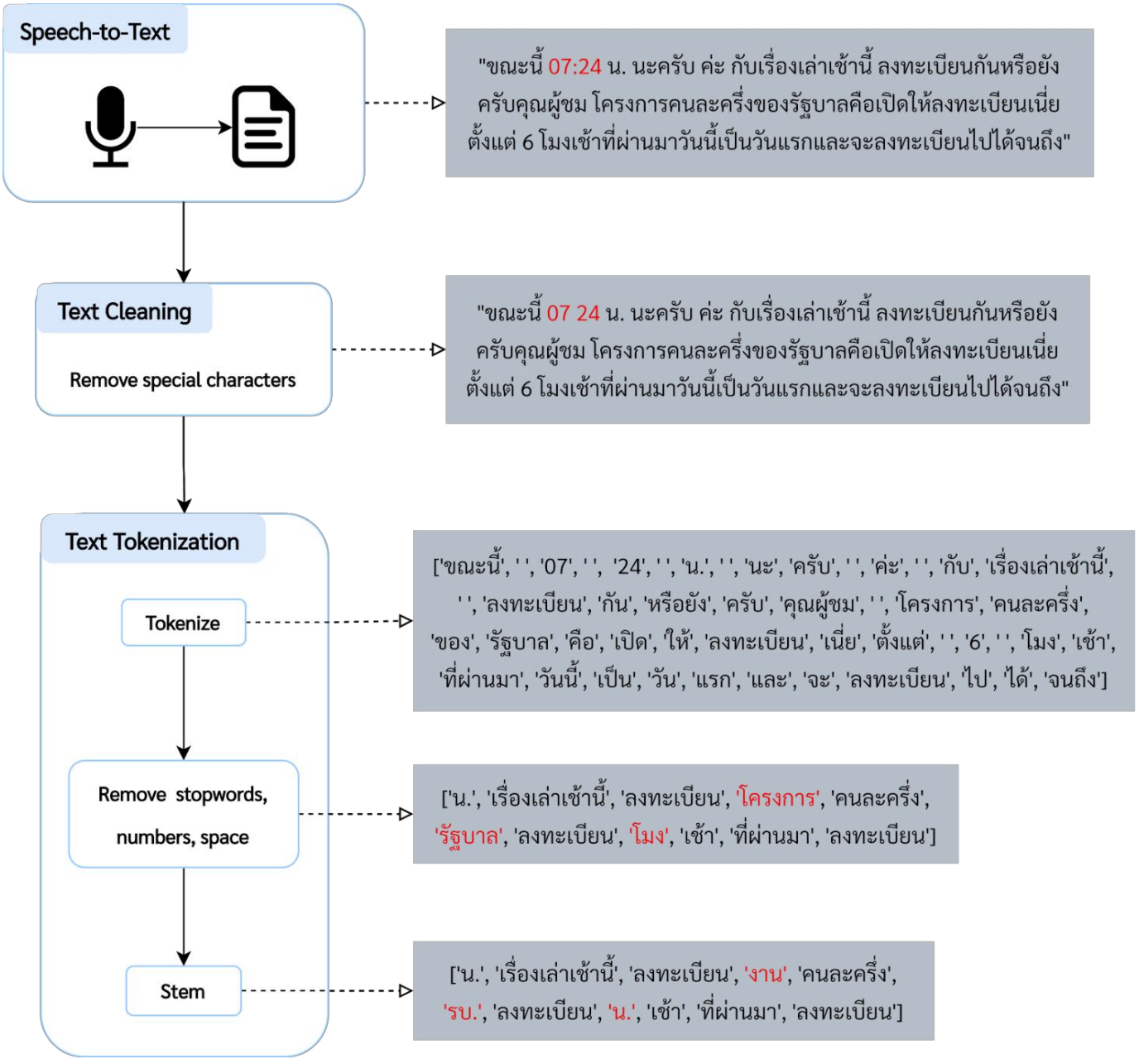
นำชุดข้อมูลประเภทข้อความที่ได้มาทำความสะอาด โดยการลบอักขระพิเศษ (special characters) ยกเว้นเครื่องหมายมหัพภาค (.) ซึ่งเป็นอักขระที่ไม่เกี่ยวข้องกับความหมายของประโยค ได้แก่ !"#$%&'()*+,-/:;<=>?@[\]^_`{|}~ เนื่องจากอักขระเหล่านี้ไม่มีผลต่อการจำแนกประเภทของข่าว โดยที่อักขระอื่น ๆ นอกเหนืออักขระพิเศษ ได้แก่ ตัวอักษรภาษาไทยและภาษาอังกฤษ ตัวเลข และเครื่องหมายมหัพภาค จะเก็บใช้ไว้สำหรับขั้นตอนการตัดคำต่อไป

3.2.4 การตัดคำ (Text Tokenization)

ในขั้นตอนนี้มีจุดประสงค์เพื่อแบ่งข้อความหรือประโยคออกเป็นคำสั้น ๆ โดยแต่ละคำจะต้องเป็นคำที่มีความหมาย โดยในขั้นตอนนี้จะอาศัยไลบารี PyThaiNLP ซึ่งภายในไลบารีจะประกอบด้วยพจนานุกรมที่ใช้สำหรับตัดคำและหารากคำ ทั้งภาษาไทยและภาษาอังกฤษ และเพื่อให้คำมีความหลากหลายและครอบคลุมกับชุดข้อมูลที่ใช้ในงานนี้มากที่สุด ผู้วิจัยจึงได้เพิ่มคำลงในพจนานุกรมที่ใช้ตัดคำ โดยคำที่เพิ่มจะเป็นคำเฉพาะหรือคำที่ไม่มีอยู่ในพจนานุกรมทั่วไป เช่น ชื่อนักการเมือง ดารา นักแสดง และคนดัง เป็นต้น และยังได้ลบคำที่ไม่ส่งผลต่อความหมายของประโยค เช่น คำสันธาน คำบุพบท คำลงท้าย หรือคำวิเศษณ์บางคำ เช่น ขณะนี้ วันนี้ เป็นต้น ซึ่งเป็นคำฟุ่มเฟือย (stopwords) โดยเลือกลบตามรายการคำฟุ่มเฟือยที่อยู่ในไลบารี PyThaiNLP รวมไปถึงคำฟุ่มเฟือยที่ผู้วิจัยได้เพิ่มเข้ามา และสุดท้ายคือการตัดส่วนขยายคำ (stemming) ซึ่งเป็นกระบวนการลดการผันคำให้อยู่ในรูป

ของรากคำเพื่อให้คำที่มาจากรากเดียวกันแต่มีส่วนขยายต่างกันอยู่ในรูปแบบเดียวกัน ซึ่งจะได้ผลลัพธ์เป็นหน่วยย่อยของคำที่ต้องการ เช่น

- รัฐบาล ลดให้อยู่ในรูปของ รบ.
- โรงเรียน ลดให้อยู่ในรูปของ ร.ร.
- ภาพยนตร์ ลดให้อยู่ในรูปของ หนัง
- คุณแม่ ลดให้อยู่ในรูปของ แม่



รูปที่ 3.1 ขั้นตอนการเตรียมชุดคำโต้ตอบข่าวภาษาไทย

จากรูปที่ 3.1 แสดงให้เห็นถึงขั้นตอนการเตรียมชุดคำโต้ตอบข่าวภาษาไทยและตัวอย่างชุดคำโต้ตอบ ในขั้นตอนนี้เริ่มต้นจากการนำชุดคำโต้ตอบที่รวบรวมได้จากรายการข่าวต่าง ๆ มาแปลงเสียงเป็นข้อความด้วยวิธี Speech-to-Text ยกตัวอย่างข้อความของชุดคำโต้ตอบที่ได้ ดังนี้

“ขณะนี้ 07:24 น. นะครับ ค่ะ กับเรื่องเล่าเช้านี้ ลงทะเบียนกันหรือยังครับคุณผู้ชม โครงการคนละครึ่งของรัฐบาลคือเปิดให้ลงทะเบียนเนี่ยตั้งแต่ 6 โมงเช้าที่ผ่านมาวันนี้เป็นวันแรกและจะลงทะเบียนไปได้จนถึง”

จากนั้นนำข้อความที่ได้มาทำความสะอาด (Text Cleaning) โดยการลบอักขระพิเศษต่าง ๆ ออกไปข้อความ จากข้อความตัวอย่างจะเห็นว่าอักขระที่ถูกลบออกคือ ‘:’ และได้ผลลัพธ์ดังนี้

“ขณะนี้ 07 24 น. นะครับ ค่ะ กับเรื่องเล่าเช้านี้ ลงทะเบียนกันหรือยังครับคุณผู้ชม โครงการคนละครึ่งของรัฐบาลคือเปิดให้ลงทะเบียนเนี่ยตั้งแต่ 6 โมงเช้าที่ผ่านมาวันนี้เป็นวันแรกและจะลงทะเบียนไปได้จนถึง”

ขั้นตอนถัดมาคือการตัดคำ (Text Tokenization) ซึ่งเป็นการแบ่งข้อความหรือประโยคออกเป็นคำสั้น ๆ โดยแต่ละคำจะต้องเป็นคำที่มีความหมาย โดยอาศัยพจนานุกรมภาษาไทยและภาษาอังกฤษจากไลบารี PyThaiNLP รวมถึงพจนานุกรมที่ผู้วิจัยเพิ่มเข้ามา

['ขณะนี้', ' ', '07', ' ', '24', ' ', 'น.', ' ', 'นะ', 'ครับ', ' ', 'ค่ะ', ' ', 'กับ', 'เรื่องเล่าเช้านี้', ' ', 'ลงทะเบียน', 'กัน', 'หรือยัง', 'ครับ', 'คุณผู้ชม', ' ', 'โครงการ', 'คนละครึ่ง', 'ของ', 'รัฐบาล', 'คือ', 'เปิด', 'ให้', 'ลงทะเบียน', 'เนี่ย', 'ตั้งแต่', ' ', '6', ' ', 'โมง', 'เช้า', 'ที่ผ่านมา', 'วันนี้', 'เป็น', 'วัน', 'แรก', 'และ', 'จะ', 'ลงทะเบียน', 'ไป', 'ได้', 'จนถึง']

จากนั้นนำแถวของคำที่ได้มาทำความสะอาดอีกครั้งโดยการลบคำฟุ่มเฟือย (stopwords) ตัวเลขและช่องว่าง ในขั้นตอนนี้มีจุดประสงค์เพื่อกำจัดคำที่ไม่มีผลต่อความหมายของประโยค ซึ่งเป็นการทำให้ชุดข้อมูลอยู่ในรูปอย่างง่าย ได้ผลลัพธ์ดังนี้

['น.', 'เรื่องเล่าเช้านี้', 'ลงทะเบียน', 'โครงการ', 'คนละครึ่ง', 'รัฐบาล', 'ลงทะเบียน', 'โมง', 'เช้า', 'ที่ผ่านมา', 'ลงทะเบียน']

ขั้นตอนสุดท้ายของการเตรียมข้อมูลคือการตัดส่วนขยายคำ (stemming) เพื่อให้คำที่มาจากรากเดียวกันแต่มีส่วนขยายต่างกันอยู่ในรูปแบบเดียวกัน เช่น ‘โครงการ’ เป็น ‘งาน’, ‘รัฐบาล’ เป็น ‘รบ.’, ‘โมง’ เป็น ‘น.’ จากข้อความตัวอย่างจะได้ผลลัพธ์จากการตัดส่วนขยายคำดังนี้

['น.', 'เรื่องเล่าเช้านี้', 'ลงทะเบียน', 'งาน', 'คนละครึ่ง', 'รบ.', 'ลงทะเบียน', 'น.', 'เช้า', 'ที่ผ่านมา', 'ลงทะเบียน']

เมื่อเสร็จขั้นตอนการทำความสะอาดข้อความและการตัดคำ จะได้ชุดโต้ตอบข่าวภาษาไทยที่มีลักษณะเป็นคำสั้น ๆ จัดเก็บอยู่ในรูปของรายการ จำนวน 600 รายการ หรือ 600 ชุดคำโต้ตอบ โดยมีจำนวนคำทั้งสิ้น 10,140 คำ เฉลี่ย 16 คำต่อหนึ่งชุดคำโต้ตอบ

3.2.5 การแบ่งชุดข้อมูล (Data Splitting)

ชุดข้อมูลทั้งหมดมีจำนวน 600 ชุด แบ่งออกเป็น 6 ประเภท กระจายอย่างเท่ากัน จำนวนประเภทละ 100 ชุด ในการจำแนกประเภทหัวข้อข่าวแบ่งชุดข้อมูลทั้งหมดออกเป็น 2 ชุดย่อย ดังนี้

1. ชุดสอน (Training set) ประกอบด้วยชุดคำตอบโต้ 420 ชุด แต่ละคลาสมีจำนวน 70 ชุด
2. ชุดทดสอบ (Test set) ประกอบด้วยชุดคำตอบโต้ 180 ชุด แต่ละคลาสมีจำนวน 30 ชุด

โดยสัดส่วนชุดสอนต่อชุดทดสอบคิดเป็น 70:30

3.3 การสกัดคุณลักษณะ (Features Extraction)

ในขั้นตอนนี้จะอาศัยการคำนวณความถี่ของคำ (Term Frequency: TF) และคำนวณความถี่ของเอกสารที่ผกผัน (Inverse Document Frequency: IDF) ซึ่งเป็นขั้นตอนวิธีสำหรับการสกัดคุณลักษณะอย่างง่ายโดยแปลงหน่วยย่อยหรือคำที่ได้มาจากขั้นตอนก่อนหน้ามาอยู่ในรูปของเวกเตอร์

สำหรับชุดสอนจะได้เวกเตอร์คุณลักษณะที่มีขนาดเท่ากับ 2,255 จำนวน 420 เอกสาร และสำหรับชุดทดสอบจะได้เวกเตอร์คุณลักษณะที่มีขนาดเท่ากับ 2,255 จำนวน 180 เอกสาร อธิบายได้ดังตัวอย่างต่อไปนี้

กำหนดให้เอกสาร ก มีชุดคำโต้ตอบดังนี้

“ขณะนี้ 07:24 น. นะครับ ค่ะ กับเรื่องเล่าเช้านี้ ลงทะเบียนกันหรือยังครับคุณผู้ชม โครงการคนละครึ่งของรัฐบาลคือเปิดให้ลงทะเบียนเนี่ยตั้งแต่ 6 โมงเช้าที่ผ่านมาวันนี้เป็นวันแรกและจะลงทะเบียนไปได้จนถึง”

รูปที่ 3.2 แสดงตัวอย่างคำจำนวน 10 คำ จากทั้งหมด 2,255 คำ พร้อมผลลัพธ์จากการคำนวณความถี่ของคำและความถี่ของเอกสารที่ผกผัน ก ของแต่ละคำ โดยเรียงจากคะแนน TF-IDF จากมากไปน้อย

	tf	idf	tfidf
ลงทะเบียน	0.159508	3.951590	0.63031
น.	0.106339	4.740048	0.50405
เรื่องเล่าเช้านี้	0.053169	6.349486	0.337598
คนละครึ่ง	0.053169	4.270044	0.227035
เช้า	0.053169	4.209419	0.223812
รบ.	0.053169	4.152261	0.220773
ที่ผ่านมา	0.053169	4.098194	0.217898
งาน	0.053169	3.608646	0.191869
สืบสวนสอบสวน	0.000000	6.349486	0.000000
สิ่งแวดล้อม	0.000000	5.656338	0.000000

รูปที่ 3.2 ความถี่ของคำและความถี่ของเอกสารที่ผกผัน ก ของตัวอย่างคำจำนวน 10 คำ

จะเห็นว่าคะแนน TF-IDF แสดงถึงความสำคัญของคำแต่ละคำในเอกสาร ก ถ้าคะแนน TF-IDF มีค่ามากหมายถึงคำ ๆ นั้นมีความสำคัญต่อเอกสารนั้นมาก เช่น คำว่า ‘ลงทะเบียน’ มีการปรากฏมากที่สุดในเอกสาร ก และมีการปรากฏอยู่ในเอกสารอื่น ๆ เป็นจำนวนน้อย ในทางกลับกันเมื่อพิจารณาคำว่า ‘สิ่งแวดล้อม’ จะพบว่าเป็นคำที่ไม่มีการปรากฏอยู่ในเอกสาร ก แต่ปรากฏอยู่ทั่วไปในหลาย ๆ เอกสาร ทำให้คะแนน TF-IDF เป็น 0 ซึ่งหมายความว่าคำ ๆ นี้ไม่สำคัญกับเอกสาร ก ดังนั้นสามารถตีความได้ว่า คำว่า ‘ลงทะเบียน’ มีความสำคัญกับเอกสาร ก มากที่สุด ตามด้วย ‘น.’ และ ‘เรื่องเล่าเช้านี้’

ดังนั้นจะได้เวกเตอร์คุณลักษณะที่สร้างจากการคำนวณคะแนน TF-IDF ของแต่ละคำในแต่ละเอกสารที่มีขนาดเท่ากับ 2,255 จากคำในตัวอย่างข้างต้น แสดงเวกเตอร์คุณลักษณะได้ดังนี้

```
(0, 1903)  0.3375977133305172
(0, 1790)  0.22381189151090813
(0, 1276)  0.6303098940432628
(0, 1190)  0.22077281860577022
(0, 752)   0.5040500493286206
(0, 717)   0.2178981022485478
(0, 364)   0.19186916534539342
(0, 252)   0.22703526027097307
```

เมื่อ ตัวเลขแรก คือ ดัชนีของเอกสาร โดยมีค่าตั้งแต่ 0-599
ตัวเลขที่สอง คือ ดัชนีของคำ โดยมีค่าตั้งแต่ 0-2254
ตัวเลขที่สาม คือ คะแนน TF-IDF ที่คำนวณได้

จากเวกเตอร์คุณลักษณะที่คำนวณได้จะถูกนำไปใช้เป็นข้อมูลนำเข้าในขั้นตอนถัดไป

3.4 การหาพารามิเตอร์ที่เหมาะสม (Parameters Optimization)

ในขั้นตอนนี้เป็นการสร้างตัวจำแนกประเภทโดยกำหนดพารามิเตอร์ที่เหมาะที่สุดกับชุดข้อมูลที่ใช้ในงานวิจัยนี้ โดยแบ่งการทดลองออกเป็น 2 ส่วน ได้แก่ การหาพารามิเตอร์สำหรับขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน และการหาพารามิเตอร์สำหรับขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอน ในแต่ละส่วนมีรายดังละเอียดดังนี้

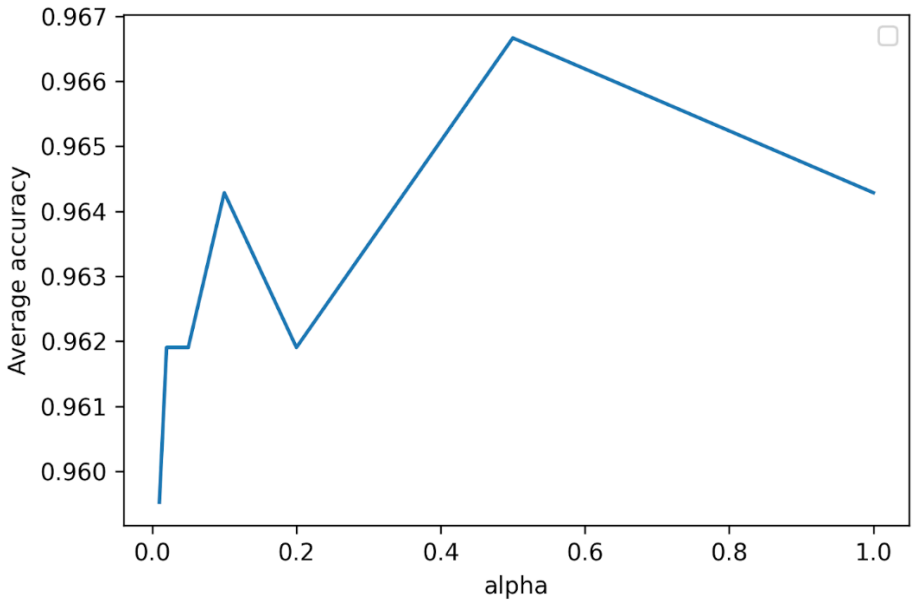
3.4.1 ขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน (Supervised Learning Algorithm)

สำหรับการหาพารามิเตอร์สำหรับขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน จะนำชุดสอนที่อยู่ในรูปของเวกเตอร์คุณลักษณะมาทำการค้นหาแบบกริด (Grid Search) และกำหนดให้แบ่งกลุ่มชุดสอนออกเป็น 5 กลุ่ม สำหรับทำการตรวจสอบแบบไขว้ (Cross Validation) โดยทดลองหาพารามิเตอร์กับทุก ๆ ขั้นตอนวิธีที่แตกต่างกันจำนวน 5 ขั้นตอนวิธี ได้แก่ Multinomial Naive Bayes (MNB), K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machines (SVM) และ Multi-Layer Perceptron (MLP) โดยแต่ละขั้นตอนวิธีมีรายละเอียดดังนี้

1. Multinomial Naive Bayes (MNB)

MNB เป็นขั้นตอนวิธีที่อาศัยทฤษฎีความน่าจะเป็นตามทฤษฎีบทของเบย์ ซึ่งการที่จะคำนวณความน่าจะเป็นได้นั้นต้องอาศัยพารามิเตอร์ alpha (α) ซึ่งเป็นพารามิเตอร์ที่ช่วยปรับปรุงค่าประมาณของพารามิเตอร์ที่คำนวณได้ให้มีค่าไม่เท่ากับศูนย์ โดยหาพารามิเตอร์ alpha ที่เหมาะสมด้วยวิธีการค้นหาแบบกริดโดยปรับพารามิเตอร์ alpha ที่ค่าต่าง ๆ ได้แก่ 0.01, 0.02, 0.05, 0.1, 0.2, 0.5 และ 1.0

กราฟเส้นสำหรับค่าพารามิเตอร์ alpha ที่ปรับให้เหมาะสมของ MNB จะแสดงในรูปที่ 3.3 จะเห็นว่าค่าความแม่นเฉลี่ยมีค่าสูงเมื่อกำหนดให้ alpha มีค่าน้อย ๆ ดังนั้นจะได้พารามิเตอร์ที่เหมาะสม คือ alpha = 0.5



รูปที่ 3.3 ค่าความแม่นเฉลี่ย เมื่อกำหนด alpha ให้มีค่าตั้งแต่ 0.01 ถึง 1.0

2. K-Nearest Neighbors (KNN)

KNN เป็นขั้นตอนวิธีหาป้ายกำกับที่ได้จากการรวมผลโหวตเสียงข้างมากของจุดเพื่อนบ้านที่อยู่รอบ ๆ จุดเป้าหมาย ซึ่งจำเป็นต้องอาศัยพารามิเตอร์ต่าง ๆ ในการกำหนดเงื่อนไขของจุดเพื่อนบ้านที่จะนำมาพิจารณา ได้แก่

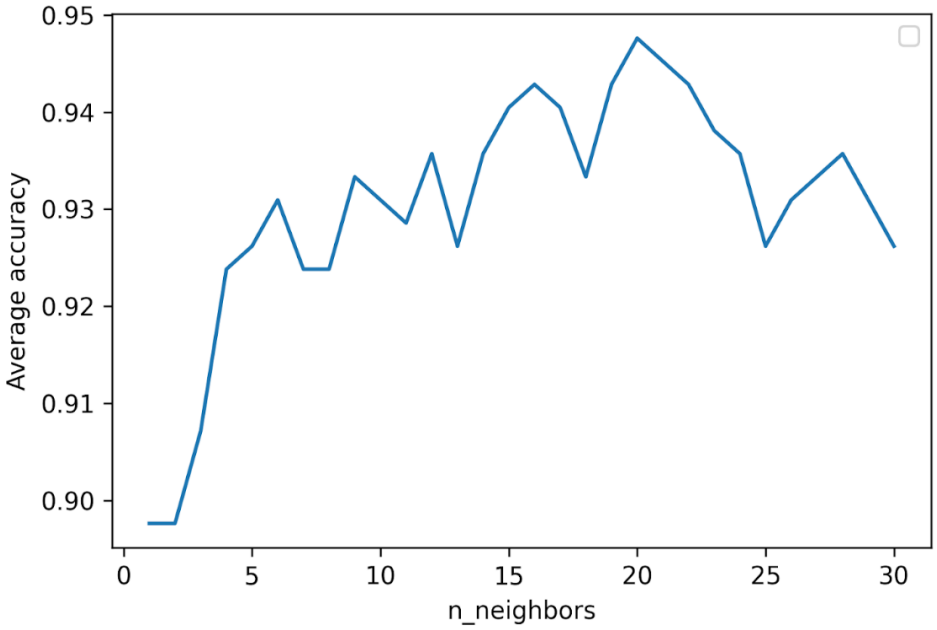
- **n_neighbors** คือ จำนวนจุดเพื่อนบ้านที่ใช้ (K)
- **p** คือ พารามิเตอร์ยกกำลังของเมทริกซ์มินคอฟสกี โดยใช้การหาระยะทางแบบแมนฮัตตัน (l1) เมื่อ $p = 1$ และใช้การหาระยะทางแบบยุคลิด (l2) เมื่อ $p = 2$
- **weights** คือ ฟังก์ชันถ่วงน้ำหนักที่ใช้ในการทำนาย โดยทุกจุดเพื่อนบ้านจะถูกถ่วงน้ำหนักให้เท่ากันเมื่อใช้ฟังก์ชันถ่วงน้ำหนักแบบสม่ำเสมอ (uniform) และจะคำนวณน้ำหนักตามส่วนกลับของระยะทางของจุดเมื่อใช้ฟังก์ชันถ่วงน้ำหนักตามระยะ (distance)

การหาค่าที่เหมาะสมสำหรับพารามิเตอร์ดังกล่าวด้วยวิธีการค้นหาแบบกริด ทำโดยกำหนดค่าพารามิเตอร์ต่าง ๆ ดังตารางที่ 3.1

ตารางที่ 3.1 พารามิเตอร์ KNN สำหรับการค้นหาแบบกริด

p	weights
1, 2	uniform, distance

จะได้ค่าพารามิเตอร์ที่เหมาะสมคือ p = 2 และ weights = distance จากนั้นจะนำชุดพารามิเตอร์ที่ได้ไปใช้เพื่อหาค่าที่เหมาะสมสำหรับ n_neighbors โดยกำหนดให้ n_neighbors มีค่าตั้งแต่ 1 ถึง 30 ดังรูปที่ 3.4



รูปที่ 3.4 ค่าความแม่นเฉลี่ย เมื่อกำหนด n_neighbors ให้มีค่าตั้งแต่ 1 ถึง 30

จากกราฟในรูปที่ 3.4 แสดงค่าความแม่นเฉลี่ย เมื่อกำหนด n_neighbors ให้มีค่าตั้งแต่ 1 ถึง 30 จะเห็นได้ว่าค่าพารามิเตอร์ที่เหมาะสมที่ทำให้ค่าความแม่นเฉลี่ยมีค่าสูงที่สุด คือ n_neighbors = 20

3. Random Forest (RF)

RF ใช้หลักการสร้างต้นไม้ตัดสินใจอย่างง่ายหลาย ๆ ต้น มารวมเข้าด้วยกัน ในการหาต้นไม้ที่เหมาะสมนั้น สามารถหาได้จากการปรับค่าพารามิเตอร์ต่าง ๆ ให้เหมาะสมด้วยวิธีการค้นหาแบบกริด พารามิเตอร์ที่นำมาพิจารณา มีดังนี้

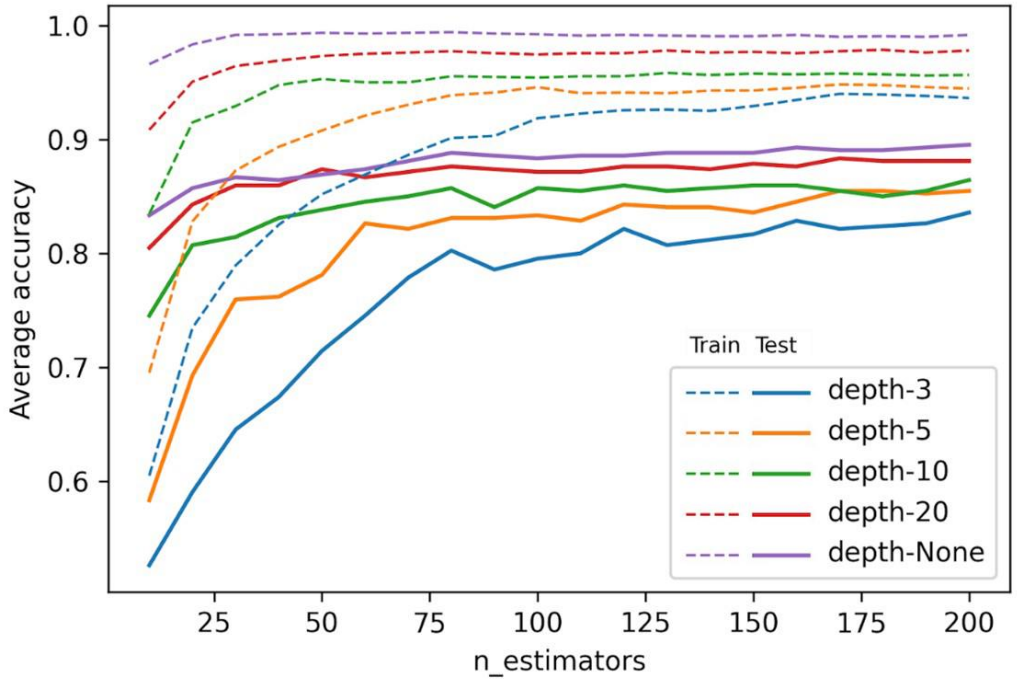
- **n_estimators** คือ จำนวนต้นไม้ทั้งหมดที่ต้องการสร้าง
- **criterion** คือ ฟังก์ชันในการวัดคุณภาพของการแบ่งโหนด ซึ่งใช้เกณฑ์ในการพิจารณาอยู่ 2 เกณฑ์ คือ ความไม่บริสุทธิ์ของจีนี และเอนโทรปี
- **max_depth** คือ ระดับความลึกที่มากที่สุดของต้นไม้
- **min_samples_leaf** คือ จำนวนชุดสอนขั้นต่ำในโหนดใบ (leaf node)
- **min_samples_split** คือ จำนวนชุดสอนขั้นต่ำในการแบ่งโหนดภายใน (internal node)

กำหนดค่าพารามิเตอร์ต่าง ๆ ได้ดังตารางที่ 3.2

ตารางที่ 3.2 พารามิเตอร์ RF สำหรับการค้นหาแบบกริด

criterion	min_samples_leaf	min_samples_split
gini, entropy	1, 3, 10	1, 3, 10, 20, 50

จากการตรวจสอบแบบไขว้จะได้ค่าพารามิเตอร์ที่เหมาะสมคือ criterion = gini, min_samples_leaf = 1 และ min_samples_split = 20 จากนั้นนำค่าพารามิเตอร์ที่ได้ไปใช้เพื่อหาค่าที่เหมาะสมสำหรับ n_estimators และ max_depth โดยกำหนดให้ n_estimators มีค่าตั้งแต่ 10 ถึง 200 และ max_depth มีค่าเท่ากับ 3, 5, 10, 20 และ None ดังรูปที่ 3.5



รูปที่ 3.5 ค่าความแม่นเฉลี่ยชุดสอนและชุดทดสอบในระดับความลึกของต้นไม้ที่แตกต่างกัน เมื่อกำหนด n_estimators ให้มีค่าตั้งแต่ 10 ถึง 200

จากกราฟในรูปที่ 3.5 จะได้ว่าที่ระดับความลึกของต้นไม้เป็น None ซึ่งหมายความว่าไม่กำหนดระดับความลึกของต้นไม้ ทำให้ได้ค่าความแม่นเฉลี่ยสูงที่สุดที่จำนวนต้นไม้เท่ากับ 160 ดังนั้นจะได้ค่าพารามิเตอร์ที่เหมาะสมคือ n_estimators = 160 และ max_depth = None

4. Support Vector Machines (SVM)

SVM คือขั้นตอนวิธีที่อาศัยหลักการของการหาสัมประสิทธิ์ของสมการเพื่อสร้างเส้นสำหรับแบ่งกลุ่มข้อมูลออกจากกัน ซึ่งอาศัยพารามิเตอร์ต่าง ๆ ในการกำหนดเงื่อนไขสำหรับสร้างเส้นแบ่ง โดยมีพารามิเตอร์ที่สำคัญ ได้แก่

- **C** คือ พารามิเตอร์การทำให้เป็นมาตรฐาน (Regularization parameter) เมื่อ C มีค่าน้อย ๆ ระยะขอบของเส้นแบ่งจะลดลง ทำให้ตัวจำแนกมีความซับซ้อนมากขึ้น และเมื่อ C มีค่ามาก ๆ ระยะขอบของเส้นแบ่งจะเพิ่มขึ้น ทำให้ตัวจำแนกมีความเรียบง่ายมากขึ้น
- **kernel** คือ ประเภทของเคอร์เนลที่ใช้ในขั้นตอนวิธี ได้แก่ เคอร์เนลเชิงเส้น เคอร์เนลพหุนาม เคอร์เนลฟังก์ชันฐานแนวรัศมี และคอร์เนลซิกมอยด์
- **gamma** คือ ค่าสัมประสิทธิ์เคอร์เนลของเคอร์เนลพหุนาม เคอร์เนลฟังก์ชันฐานแนวรัศมี และคอร์เนลซิกมอยด์

หาค่าที่เหมาะสมสำหรับพารามิเตอร์ดังกล่าวด้วยวิธีการค้นหาแบบกริด โดยกำหนดค่าพารามิเตอร์ต่าง ๆ ดังตารางที่ 3.3

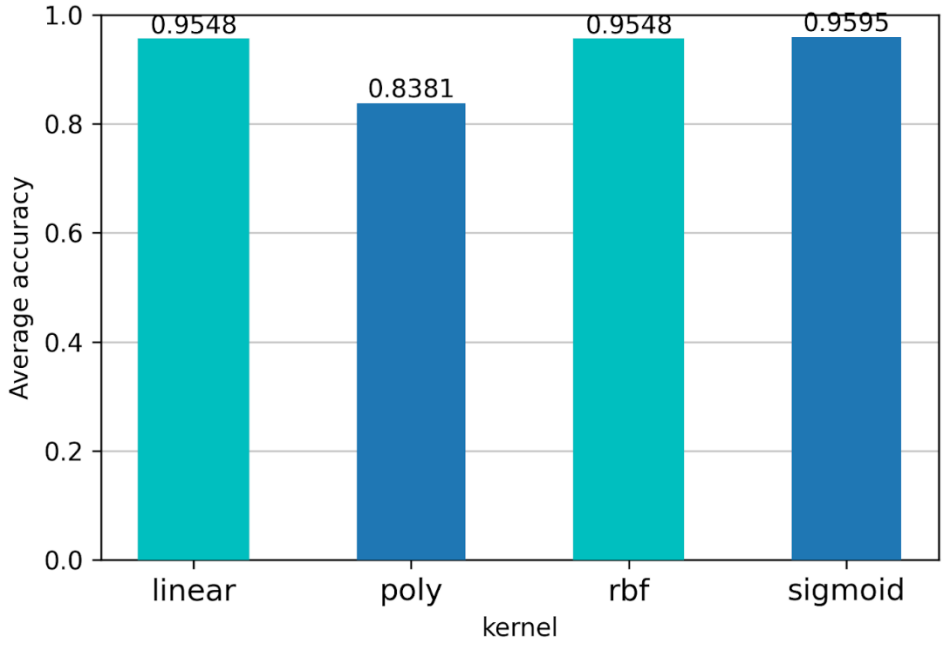
ตารางที่ 3.3 พารามิเตอร์ SVM สำหรับการค้นหาแบบกริด

kernel	C	gamma
linear, poly, rbf, sigmoid	1, 10, 100	0.01, 0.05, 0.1, 0.2, 0.5, 1

จากการค้นหาแบบกริดจะได้ว่าสำหรับ kernel ที่แตกต่างกัน จะมี C และ gamma ที่เหมาะที่สุดแตกต่างกัน ได้ผลลัพธ์ดังนี้

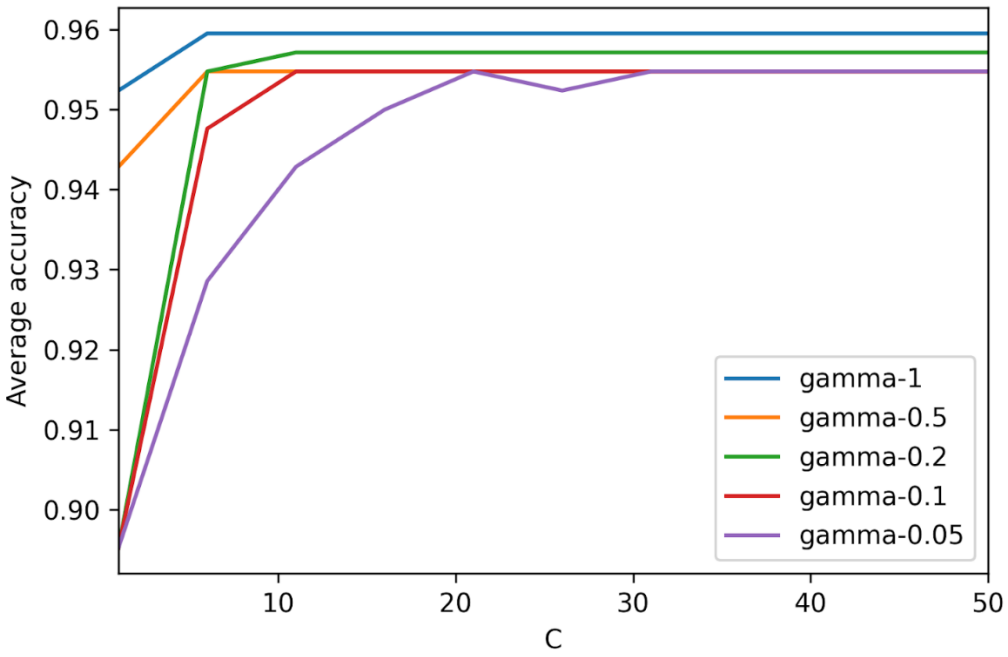
- {'kernel': 'linear', 'C': 1, 'gamma': 1}
- {'kernel': 'poly', 'C': 10, 'gamma': 1}
- {'kernel': 'rbf', 'C': 100, 'gamma': 0.01}
- {'kernel': 'sigmoid', 'C': 10, 'gamma': 1}

เมื่อพิจารณาค่าความแม่นเฉลี่ยจากการตรวจสอบแบบไขว้ จะได้พารามิเตอร์ที่เหมาะสมที่ให้ค่าความแม่นเฉลี่ยสูงสุด คือ เคอร์เนลซิกมอยด์ ดังรูปที่ 3.6



รูปที่ 3.6 ค่าความแม่นเฉลี่ยของพารามิเตอร์เคอร์เนลแบบต่าง ๆ

จากกราฟในรูปที่ 3.6 แสดงค่าความแม่นเฉลี่ยของพารามิเตอร์เคอร์เนลทั้ง 4 ประเภท โดยใช้ C และ gamma ที่แตกต่างกัน จะเห็นได้ว่าพารามิเตอร์ที่เหมาะสมที่ให้ค่าความแม่นเฉลี่ยสูงสุด คือ เคอร์เนลประเภทซิกมอยด์ ดังนั้นจึงกำหนดพารามิเตอร์ kernel = sigmoid



รูปที่ 3.7 ค่าความแม่นเฉลี่ยของเคอร์เนลซิกมอยด์ของแต่ละค่า gamma ที่แตกต่างกัน เมื่อกำหนด C ให้มีค่าตั้งแต่ 1 ถึง 51

จากกราฟในรูปที่ 3.7 แสดงค่าความแม่นเฉลี่ยของเคอร์เนลซิกมอยด์ของค่า gamma ที่แตกต่างกันจำนวน 5 ค่า ได้แก่ 0.05, 0.1, 0.2, 0.5 และ 1 เมื่อกำหนดให้ C มีค่าตั้งแต่ 1 ถึง 51 เนื่องจากค่าความแม่นเฉลี่ยสูงสุดคือเส้นที่ gamma = 1 และ C ที่มีค่าตั้งแต่ 6 ขึ้นไป ดังนั้นเลือก C ที่มีค่าไม่สูงมาก เพราะอาจให้ตัวจำแนกมีความเรียบง่ายเกินไป จะได้พารามิเตอร์ที่เหมาะสมคือ C = 6 และ gamma = 1

5. Multi-Layer Perceptron (MLP)

MLP ใช้หลักการรวมกันของเซลล์ประสาทหลายเซลล์เข้าด้วยกัน ประกอบกันหลาย ๆ ชั้น ในการเรียนรู้ของ MLP นั้น จะต้องปรับค่าน้ำหนักของเพอร์เซปตรอน ในการปรับค่าน้ำหนักนั้นขึ้นกับพารามิเตอร์ต่าง ๆ ดังนี้

- **hidden_layer_sizes** คือ จำนวนโหนดในชั้นซ่อนแต่ละชั้น
- **activation function** คือ ฟังก์ชันกระตุ้นของชั้นซ่อน ได้แก่ ฟังก์ชัน identity ฟังก์ชัน hyperbolic tangent (tanh) ฟังก์ชัน logistic sigmoid (sigmoid) และฟังก์ชัน rectified linear unit (relu)
- **solver** คือ พารามิเตอร์สำหรับการปรับน้ำหนักให้เหมาะสม โดยมี lbfgs เป็นตัวเพิ่มประสิทธิภาพในตระกูลของวิธีการกึ่งนิวตัน โดยที่ adam และ sgd ใช้ความชัน (gradient) ในการบอกขนาดและทิศทางในการปรับพารามิเตอร์
- **alpha** คือ พารามิเตอร์การทำให้เป็นมาตรฐาน (Regularization parameter) เมื่อ alpha มีค่าน้อย ตัวจำแนกจะมีความซับซ้อนมากขึ้น และเมื่อ alpha มีค่ามาก ตัวจำแนกจะเรียบง่ายมากขึ้น
- **learning_rate_init** คือ อัตราการเรียนรู้เริ่มต้นสำหรับการปรับค่าน้ำหนัก

ใช้วิธีการค้นหาแบบกริดเพื่อหาพารามิเตอร์ที่เหมาะสม โดยกำหนดพารามิเตอร์ได้ดังตารางที่ 3.4

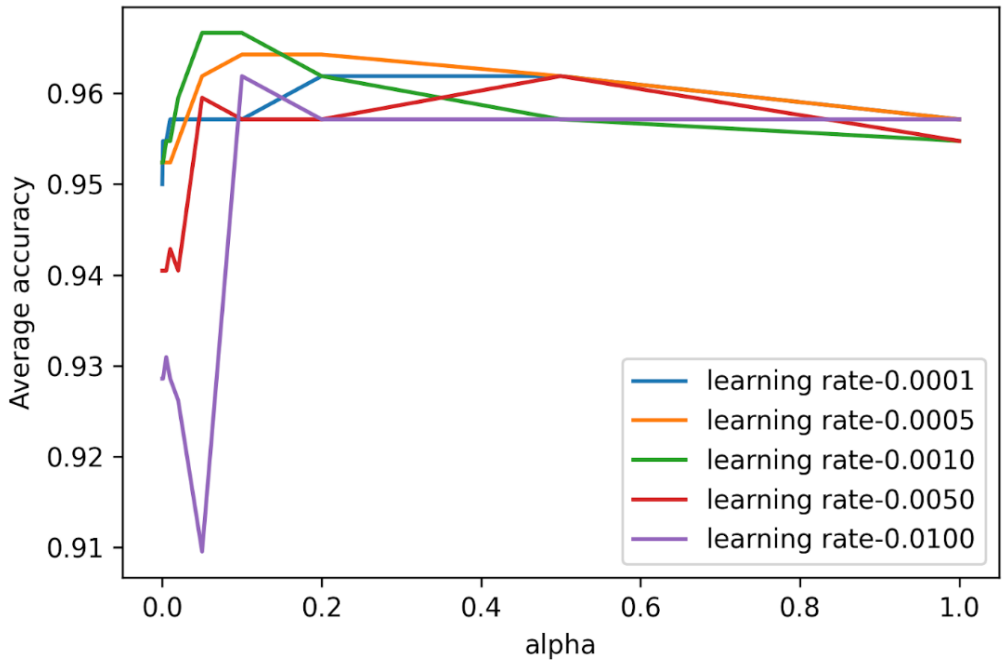
ตารางที่ 3.4 พารามิเตอร์ MLP สำหรับการค้นหาแบบกริด ครั้งที่ 1

hidden_layer_sizes	activation	solver
(64), (128), (256), (32, 32), (64, 64), (128, 128), (256, 256), (16, 16, 16), (32, 32, 32), (64, 64, 64), (128, 128, 128),	tanh, relu, logistic, identity	sgd, adam, lbfgs

จะได้พารามิเตอร์ที่เหมาะสมที่ให้ค่าความแม่นเฉลี่ยสูงสุด คือ alpha = 0.05, hidden_layer_sizes = (128, 128), activation = tanh, และ solver = adam จากนั้นนำค่าพารามิเตอร์ที่ได้ไปใช้เพื่อหาค่าที่เหมาะสมสำหรับ alpha และ learning_rate_init โดยกำหนดค่า alpha และ learning_rate_init ดังตารางที่ 3.5

ตารางที่ 3.5 พารามิเตอร์ MLP สำหรับการค้นหาแบบกริด ครั้งที่ 2

alpha	learning_rate_init
0.0001, 0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1	0.0001, 0.0005, 0.001, 0.005, 0.01



รูปที่ 3.8 ค่าความแม่นเฉลี่ยของ MLP ของแต่ละค่า learning_rate_init ที่แตกต่างกัน เมื่อกำหนด alpha ให้มีค่าตั้งแต่ 0.0001 ถึง 1

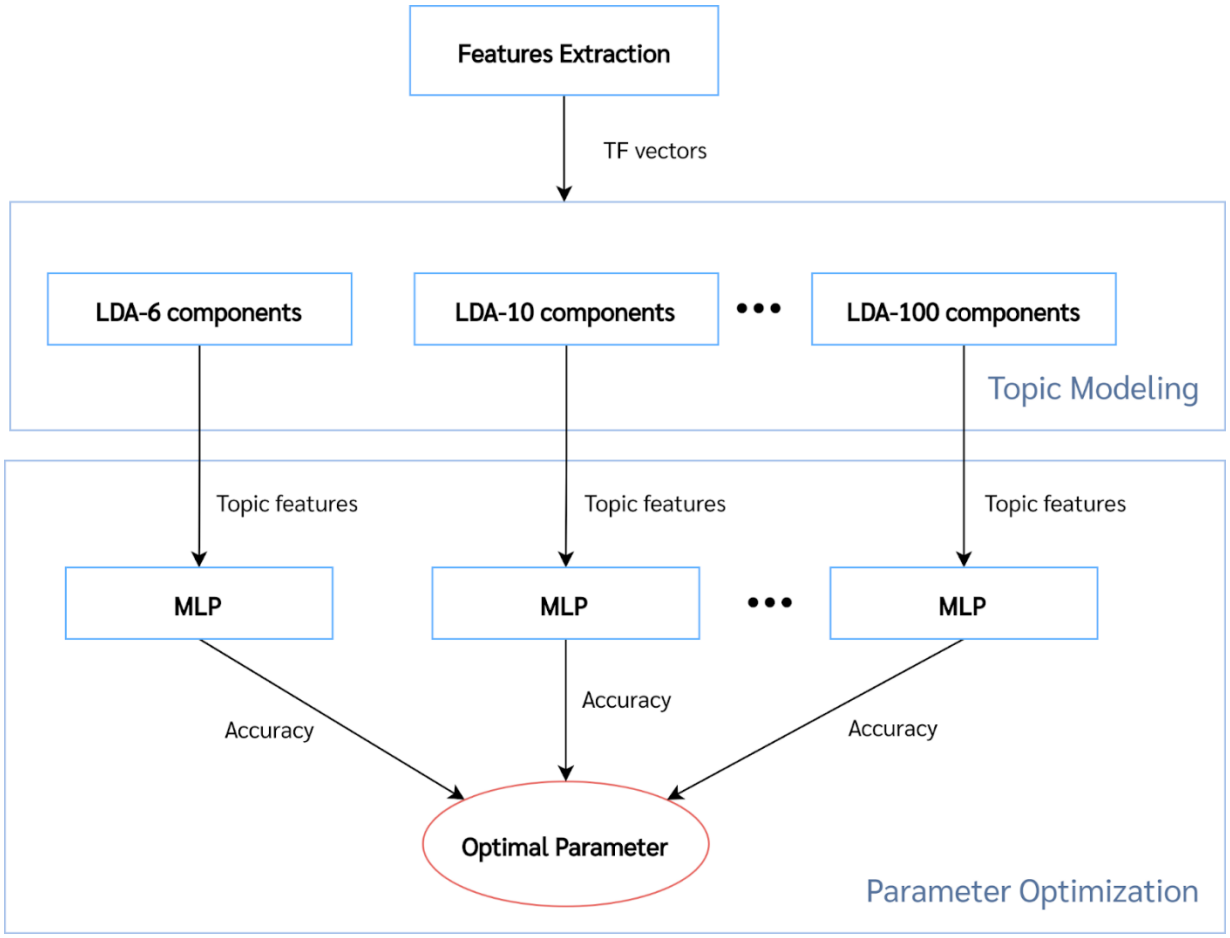
จากกราฟในรูปที่ 3.8 แสดงค่าความแม่นเฉลี่ย MLP ของค่า learning_rate_init ที่ต่างกัน จำนวน 5 ค่า ได้แก่ 0.0001, 0.0005, 0.001, 0.005 และ 0.01 เมื่อกำหนดให้ alpha มีค่าตั้งแต่ 0.0001 ถึง 1 จะได้ว่าค่าความแม่นเฉลี่ยสูงที่สุดคือเส้นที่ learning_rate_init = 0.001 และ alpha = 0.05 ดังนั้นจะได้ค่าพารามิเตอร์ที่เหมาะสมคือ learning_rate_init = 0.001 และ alpha = 0.05

3.4.2 ขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning Algorithm)

สำหรับขั้นตอนวิธีการจัดสรรของดีรีเคลท์แฝง (Latent Dirichlet Allocation: LDA) เป็นขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอนจะใช้เวกเตอร์คุณลักษณะที่ได้จากการคำนวณคะแนน TF ของชุดสอนและพารามิเตอร์ที่เหมาะสมมาสร้างแบบจำลองหัวข้อ กำหนดพารามิเตอร์จำนวนหัวข้อ (n_components) ที่แตกต่างกัน ได้แก่ 6, 10, 20, 30, 40, 50, 60, 70, 80, 90 และ 100 หัวข้อ เพื่อหาจำนวนหัวข้อที่เหมาะสมกับชุดสอน และกำหนดค่าพารามิเตอร์อื่น ๆ ที่เกี่ยวข้อง ดังนี้

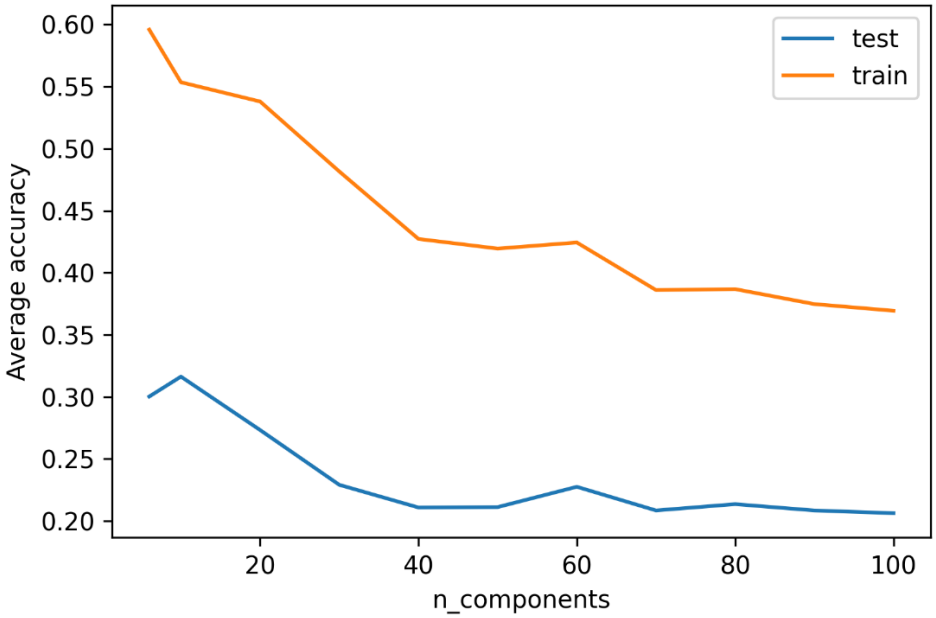
- **doc_topic_prior** คือ ตัวควบคุมสัดส่วนการกระจายของหัวข้อในแต่ละเอกสาร (α) กำหนดให้มีค่าเท่ากับ 1/n_components
- **topic_word_prior** คือ ตัวควบคุมสัดส่วนการกระจายของคำในแต่ละหัวข้อ (η) กำหนดให้มีค่าเท่ากับ 1/n_components

ในการหาพารามิเตอร์ n_compoents ที่เหมาะสมหาได้จากการเปรียบเทียบค่าความแม่นของตัวจำแนกที่สร้างจากแบบจำลองหัวข้อที่แตกต่างกัน ซึ่งทำได้โดยการอาศัยขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนที่ใช้คุณลักษณะหัวข้อที่ได้จากขั้นตอนวิธี LDA เป็นข้อมูลนำเข้า และจากขั้นตอนที่ 3.4.1 เมื่อพิจารณาค่าความแม่นที่ได้จากการตรวจสอบแบบไขว้ของขั้นตอนวิธี MLP เทียบกับขั้นตอนวิธีอื่น ๆ จะเห็นว่า MLP มีแนวโน้มที่จะให้ผลการจำแนกที่มีค่าความแม่นสูงกว่าขั้นตอนวิธีอื่น ๆ ดังนั้นจึงเลือกใช้ MLP ในการหาพารามิเตอร์ที่เหมาะสมสำหรับ LDA ดังแสดงในรูปที่ 3.9



รูปที่ 3.9 การหาพารามิเตอร์ที่เหมาะสมของขั้นตอนวิธีการจัดสรรของดีรีเคลท์แฝง

ดังนั้น ในการหาพารามิเตอร์จำนวนหัวข้อที่เหมาะสม จะสามารถหาได้จากการสร้างตัวจำแนก LDA-MLP ที่กำหนดพารามิเตอร์ n_components ที่แตกต่างกัน แล้วนำมาหาค่าความแม่นเฉลี่ยของแต่ละตัวจำแนกโดยใช้ตัวตรวจสอบแบบไขว้ด้วยวิธีการสุ่มแบ่งแบบชั้น (Stratified Shuffle Split cross-validator) โดยกำหนดจำนวนรอบการสุ่มใหม่เท่ากับ 200 รอบ และแบ่งชุดสอนออกเป็น 2 ส่วน ในอัตราส่วน 70:30 สำหรับสอนและทดสอบตามลำดับ แล้วจึงนำค่าความแม่นเฉลี่ยของแต่ละตัวจำแนก LDA-MLP มาเปรียบเทียบกันเพื่อหา n_components ที่ทำให้ค่าความแม่นเฉลี่ยมีค่าสูงสุด ซึ่งได้ผลลัพธ์ดังรูปที่ 3.10



รูปที่ 3.10 ค่าความแม่นเฉลี่ยชุดสอนและชุดทดสอบของ LDA-MLP
เมื่อกำหนดจำนวนหัวข้อตั้งแต่ 6 ถึง 100 หัวข้อ

จากรูปที่ 3.10 แสดงกราฟความสัมพันธ์ระหว่างจำนวนหัวข้อของ LDA กับค่าความแม่นเฉลี่ยที่ได้จากการประเมินประสิทธิภาพของตัวจำแนก MLP จะได้ว่าจำนวนหัวข้อที่ให้ค่าความแม่นเฉลี่ยสูงที่สุดคือ 6 หัวข้อ ดังนั้นจึงกำหนดให้พารามิเตอร์ n_components = 6 เป็นพารามิเตอร์ที่เหมาะสมสำหรับสร้างแบบจำลองหัวข้อ LDA

หลังจากที่ได้พารามิเตอร์ที่เหมาะสม นำพารามิเตอร์ที่ได้จากขั้นตอนนี้และเวกเตอร์คุณลักษณะ TF ที่ได้จากขั้นตอน 3.3 มาสร้างแบบจำลองหัวข้อ LDA ซึ่งจะกล่าวในบทถัดไป

บทที่ 4
ผลการทดลอง

ในบทนี้จะกล่าวถึงการสร้างตัวจำแนก ผลการทดลอง และการอภิปรายผลการทดลองของการสร้างตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อ

4.1 การสร้างตัวจำแนก

สร้างตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่องด้วยวิธีการเรียนรู้แบบมีผู้สอนและแบบไม่มีผู้สอนด้วยขั้นตอนวิธีที่แตกต่างกัน โดยอาศัยไลบรารี Scikit-learn เพื่อที่จะเปรียบเทียบประสิทธิภาพของตัวจำแนก

สำหรับการสร้างตัวจำแนกประเภทด้วยขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน ใช้ข้อมูลนำเข้าเป็นเวกเตอร์คุณลักษณะที่ได้จากการคำนวณคะแนน TF-IDF ของชุดสอน และตั้งค่าพารามิเตอร์ตามชุดพารามิเตอร์ที่เหมาะสมที่ได้จากขั้นตอน 3.4.1 โดยสร้างทั้งหมด 5 ตัวจำแนกจากขั้นตอนวิธีที่แตกต่างกัน ได้แก่ MNB, KNN, RF, SVM และ MLP

สำหรับการสร้างตัวจำแนกประเภทด้วยขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอน ใช้ข้อมูลนำเข้าเป็นเวกเตอร์คุณลักษณะความน่าจะเป็นของหัวข้อที่ได้จากการคำนวณความน่าจะเป็นของหัวข้อในแต่ละชุดคำโต้ตอบของชุดสอนที่คำนวณได้จากขั้นตอนวิธี LDA โดยตั้งค่าพารามิเตอร์ตามจำนวนหัวข้อที่เหมาะสมตามที่ได้จากขั้นตอนที่ 3.4.2 ในการสร้างตัวจำแนกเพื่อประเมินประสิทธิภาพ ได้เลือกขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนเพียงหนึ่งขั้นตอนวิธีจากทั้งหมด 5 ขั้นตอนวิธีที่จะนำมาใช้กับ LDA ขั้นตอนวิธีทั้ง 5 ขั้นตอนได้แก่ MNB, KNN, RF, SVM และ MLP สำหรับการตั้งค่าพารามิเตอร์ในขั้นตอนวิธีต่าง ๆ มีรายละเอียดตามขั้นตอนที่ 3.4.1 และจากการประเมินประสิทธิภาพเพื่อเลือกตัวจำแนกสำหรับ LDA พบว่าขั้นตอนวิธี MLP ให้ค่าความแม่นสูงสุดจากทั้งหมด จึงได้ตัวจำแนกสุดท้ายที่เลือกมาเป็น LDA-MLP

4.2 ผลการทดลอง

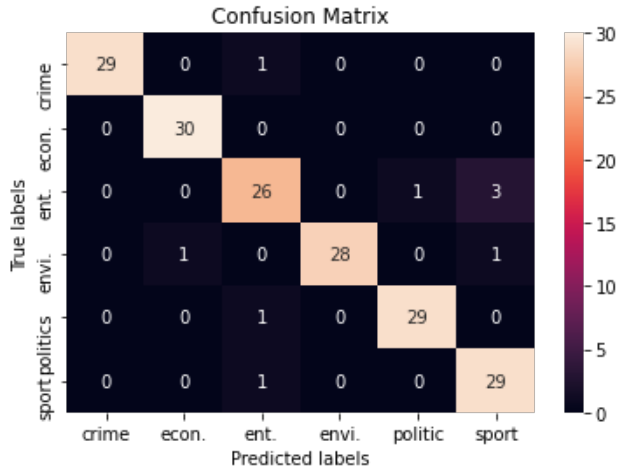
จากการสร้างตัวจำแนกประเภททั้งหมด 6 ตัวจำแนก ได้แก่ MNB, KNN, RF, SVM, MLP และ LDA-MLP จะนำตัวจำแนกทั้งหมดมาประเมินประสิทธิภาพของการจำแนกประเภทคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อ โดยทดสอบกับชุดทดสอบที่ได้เตรียมไว้ในขั้นตอนที่ 3.2.5 เพื่อพิจารณาค่าความแม่น (Accuracy) ค่าความเที่ยง (Precision) ค่าความครบถ้วน (Recall) และคะแนน F1 (F1-score)[1] พร้อมทั้งแสดงเมทริกซ์ความสับสนที่ได้จากการจำแนกประเภทหัวข้อข่าวทั้ง 6 ประเภท

[1] คะแนน F1 ของตัวจำแนกอาจไม่ได้อยู่ในช่วงระหว่างค่าความเที่ยงกับค่าความครบถ้วน เนื่องจากใช้วิธีเฉลี่ยแบบมหภาค (Macro F1-score)

จากการจำแนกประเภทชุดทดสอบของตัวจำแนก MNB พบว่าความแม่นโดยรวมมีค่าเท่ากับ 95.00% โดยมีค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 เท่ากับ 95.16%, 95.00% และ 95.01% ตามลำดับ ดังผลลัพธ์ในตารางที่ 4.1 และได้เมทริกซ์ความสับสน ดังแสดงในรูปที่ 4.1 จะเห็นได้ว่าชนิดหัวข้อข่าวที่ตัวจำแนกจำแนกได้ดีที่สุดคือหัวข้อข่าวประเภทเศรษฐกิจ โดยมีค่าความแม่นเท่ากับ 100.00% ตามด้วยข่าวอาชญากรรม ข่าวการเมือง และข่าวกีฬา ซึ่งทั้ง 3 ประเภทข่าวนี้มีค่าความแม่นเท่ากัน คือ 96.67%

ตารางที่ 4.1 การประเมินประสิทธิภาพของตัวจำแนก MNB ของชุดทดสอบ

ประเภทข่าว	**การประเมินประสิทธิภาพ**		
	Precision	Recall	F1-score
อาชญากรรม (Crime)	1.0000	0.9667	0.9831
เศรษฐกิจ (Economic)	0.9677	1.0000	0.9836
บันเทิง (Entertainment)	0.8966	0.8667	0.8814
สิ่งแวดล้อม (Environment)	1.0000	0.9333	0.9655
การเมือง (Politic)	0.9667	0.9667	0.9667
กีฬา (Sport)	0.8788	0.9667	0.9206
รวม	**0.9516**	**0.9500**	**0.9501**

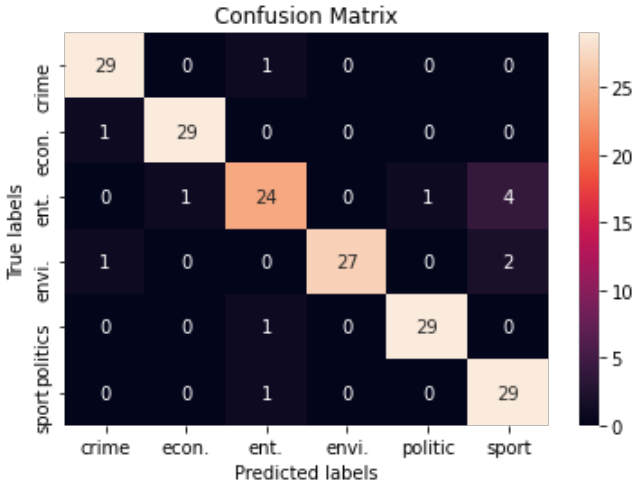


รูปที่ 4.1 เมทริกซ์ความสับสนของตัวจำแนก MNB

จากการจำแนกประเภทชุดทดสอบของตัวจำแนก KNN พบว่าความแม่นโดยรวมมีค่าเท่ากับ 92.78% โดยมีค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 เท่ากับ 93.10%, 92.78% และ 92.77% ตามลำดับ ดังผลลัพธ์ในตารางที่ 4.2 และได้เมทริกซ์ความสับสน ดังแสดงในรูปที่ 4.2 จะเห็นได้ว่าชนิดหัวข้อข่าวที่ตัวจำแนกจำแนกได้ดีที่สุดคือหัวข้อข่าวประเภทอาชญากรรม เศรษฐกิจ การเมือง และกีฬา ได้ค่าความแม่นเท่ากัน คือ 96.67%

ตารางที่ 4.2 การประเมินประสิทธิภาพของตัวจำแนก KNN ของชุดทดสอบ

ประเภทข่าว	**การประเมินประสิทธิภาพ**		
	Precision	Recall	F1-score
อาชญากรรม (Crime)	0.9355	0.9667	0.9508
เศรษฐกิจ (Economic)	0.9667	0.9667	0.9667
บันเทิง (Entertainment)	0.8889	0.8000	0.8421
สิ่งแวดล้อม (Environment)	1.0000	0.9000	0.9474
การเมือง (Politic)	0.9667	0.9667	0.9667
กีฬา (Sport)	0.8286	0.9667	0.8923
รวม	**0.9310**	**0.9278**	**0.9277**

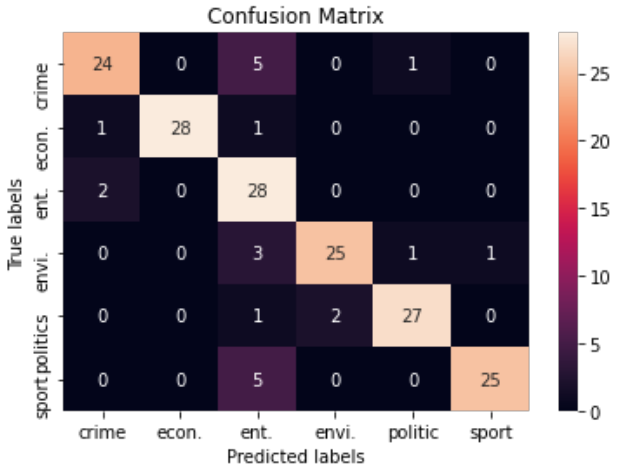


รูปที่ 4.2 เมทริกซ์ความสับสนของตัวจำแนก KNN

จากการจำแนกประเภทชุดทดสอบของตัวจำแนก RF พบว่าความแม่นโดยรวมมีค่าเท่ากับ 87.22% โดยมีค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 เท่ากับ 89.31%, 87.22% และ 87.67% ตามลำดับ ดังผลลัพธ์ในตารางที่ 4.3 และได้เมทริกซ์ความสับสน ดังแสดงในรูปที่ 4.3 จะเห็นได้ว่าชนิดหัวข้อข่าวที่ตัวจำแนกจำแนกได้ดีที่สุดคือหัวข้อข่าวประเภทเศรษฐกิจ และบันเทิง ซึ่งได้ค่าความแม่นเท่ากัน คือ 93.33% ตามด้วย 83.33% ของข่าวสิ่งแวดล้อม และข่าวกีฬา

ตารางที่ 4.3 การประเมินประสิทธิภาพของตัวจำแนก RF ของชุดทดสอบ

ประเภทข่าว	**การประเมินประสิทธิภาพ**		
	Precision	Recall	F1-score
อาชญากรรม (Crime)	0.8889	0.8000	0.8421
เศรษฐกิจ (Economic)	1.0000	0.9333	0.9655
บันเทิง (Entertainment)	0.6512	0.9333	0.7671
สิ่งแวดล้อม (Environment)	0.9259	0.8333	0.8772
การเมือง (Politic)	0.9310	0.9000	0.9153
กีฬา (Sport)	0.9615	0.8333	0.8929
รวม	**0.8931**	**0.8722**	**0.8767**

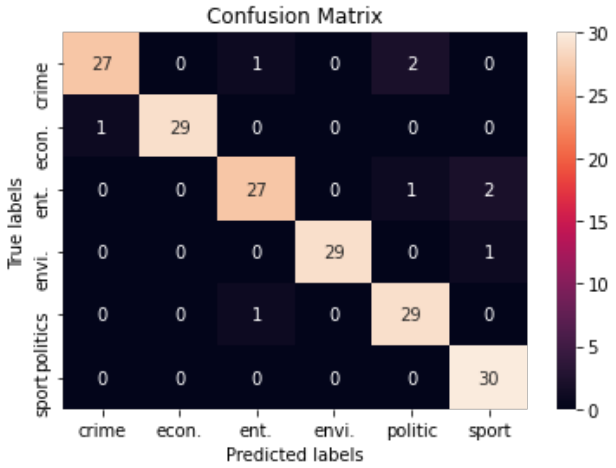


รูปที่ 4.3 เมทริกซ์ความสับสนของตัวจำแนก RF

จากการจำแนกประเภทชุดทดสอบของตัวจำแนก SVM พบว่าความแม่นโดยรวมมีค่าเท่ากับ 95.00% โดยมีค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 เท่ากับ 95.18%, 95.00% และ 95.00% ตามลำดับ ดังผลลัพธ์ในตารางที่ 4.4 และได้เมทริกซ์ความสับสน ดังแสดงในรูปที่ 4.4 จะเห็นได้ว่าชนิดหัวข้อข่าวที่ตัวจำแนกจำแนกได้ดีที่สุดคือหัวข้อข่าวประเภทกีฬา ได้ค่าความแม่นเท่ากับ 100.00% ตามด้วย 96.67% ของข่าวเศรษฐกิจ ข่าวสิ่งแวดล้อม และข่าวการเมือง

ตารางที่ 4.4 การประเมินประสิทธิภาพของตัวจำแนก SVM ของชุดทดสอบ

ประเภทข่าว	**การประเมินประสิทธิภาพ**		
	Precision	Recall	F1-score
อาชญากรรม (Crime)	0.9643	0.9000	0.9310
เศรษฐกิจ (Economic)	1.0000	0.9667	0.9831
บันเทิง (Entertainment)	0.9310	0.9000	0.9153
สิ่งแวดล้อม (Environment)	1.0000	0.9667	0.9831
การเมือง (Politic)	0.9062	0.9667	0.9355
กีฬา (Sport)	0.9091	1.0000	0.9524
รวม	**0.9518**	**0.9500**	**0.9500**

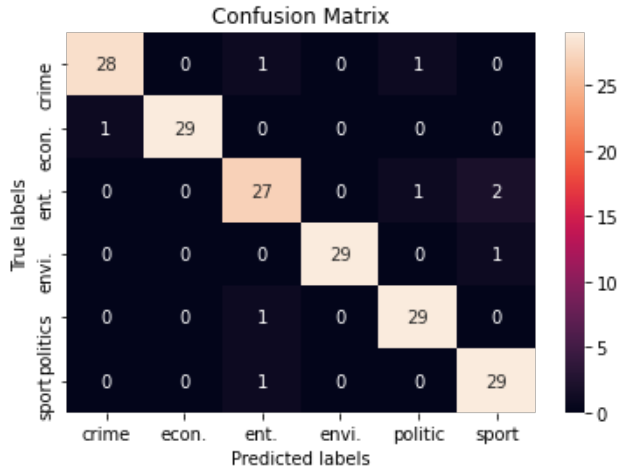


รูปที่ 4.4 เมทริกซ์ความสับสนของตัวจำแนก SVM

จากการจำแนกประเภทชุดทดสอบของตัวจำแนก MLP พบว่าความแม่นโดยรวมมีค่าเท่ากับ 95.00% โดยมีค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 เท่ากับ 95.12%, 95.00% และ 95.03% ตามลำดับ ดังผลลัพธ์ในตารางที่ 4.5 และได้เมทริกซ์ความสับสน ดังแสดงในรูปที่ 4.5 จะเห็นได้ว่าชนิดหัวข้อข่าวที่ตัวจำแนกจำแนกได้ดีที่สุดคือหัวข้อข่าวประเภทเศรษฐกิจ สิ่งแวดล้อม การเมือง และกีฬา ได้ค่าความแม่นเท่ากันที่ 96.67%

ตารางที่ 4.5 การประเมินประสิทธิภาพของตัวจำแนก MLP ของชุดทดสอบ

ประเภทข่าว	**การประเมินประสิทธิภาพ**		
	Precision	Recall	F1-score
อาชญากรรม (Crime)	0.9655	0.9333	0.9492
เศรษฐกิจ (Economic)	1.0000	0.9667	0.9831
บันเทิง (Entertainment)	0.9000	0.9000	0.9000
สิ่งแวดล้อม (Environment)	1.0000	0.9667	0.9831
การเมือง (Politic)	0.9355	0.9667	0.9508
กีฬา (Sport)	0.9062	0.9667	0.9355
รวม	**0.9512**	**0.9500**	**0.9503**

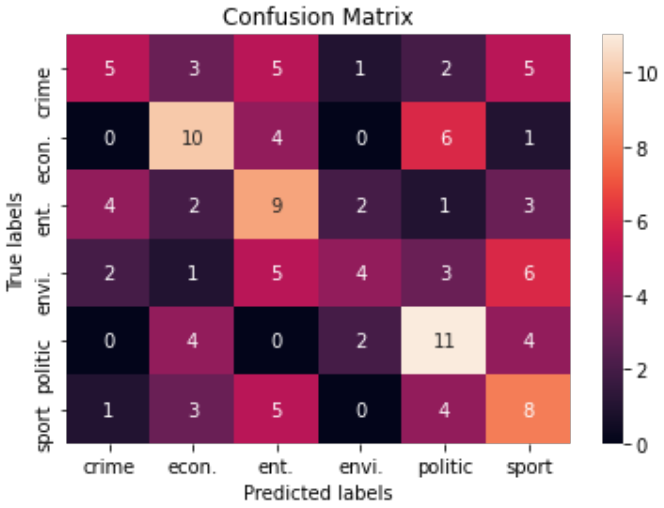


รูปที่ 4.5 เมทริกซ์ความสับสนของตัวจำแนก MLP

จากการจำแนกประเภทชุดทดสอบของตัวจำแนก LDA-MLP พบว่าความแม่นโดยรวมมีค่าเท่ากับ 37.30% โดยมีค่าความเที่ยง ค่าความครบถ้วน และคะแนน F1 เท่ากับ 38.68% 37.30% และ 36.39% ตามลำดับ ดังผลลัพธ์ในตารางที่ 4.6 และได้เมทริกซ์ความสับสน ดังแสดงในรูปที่ 4.6 จะเห็นได้ว่าชนิดหัวข้อข่าวที่ตัวจำแนกจำแนกได้ดีที่สุดคือหัวข้อข่าวประเภทการเมือง ได้ค่าความแม่นเท่ากับ 52.38% ตามด้วยข่าวเศรษฐกิจ และข่าวบันเทิง ซึ่งมีค่าความแม่นเท่ากับ 47.62% และ 42.86% ตามลำดับ แม้ว่าความแม่นในการจำแนกข่าวประเภทการเมือง จะสามารถทำได้สูงถึง 52.38% แต่ในทางกลับกัน สำหรับการจำแนกข่าวประเภทสิ่งแวดล้อม กลับทำได้เพียง 19.05% ด้วยเหตุนี้จึงทำให้ตัวจำแนก LDA-MLP มีประสิทธิภาพในการจำแนกต่ำ

ตารางที่ 4.6 การประเมินประสิทธิภาพของตัวจำแนก LDA-MLP ของชุดทดสอบ

ประเภทข่าว	**การประเมินประสิทธิภาพ**		
	Precision	**Recall**	**F1-score**
อาชญากรรม (Crime)	0.4167	0.2381	0.3030
เศรษฐกิจ (Economic)	0.4348	0.4762	0.4545
บันเทิง (Entertainment)	0.3214	0.4286	0.3673
สิ่งแวดล้อม (Environment)	0.4444	0.1905	0.2667
การเมือง (Politic)	0.4074	0.5238	0.4583
กีฬา (Sport)	0.2963	0.3810	0.3333
รวม	**0.3868**	**0.3730**	**0.3639**



รูปที่ 4.6 เมทริกซ์ความสับสนของตัวจำแนก LDA-MLP

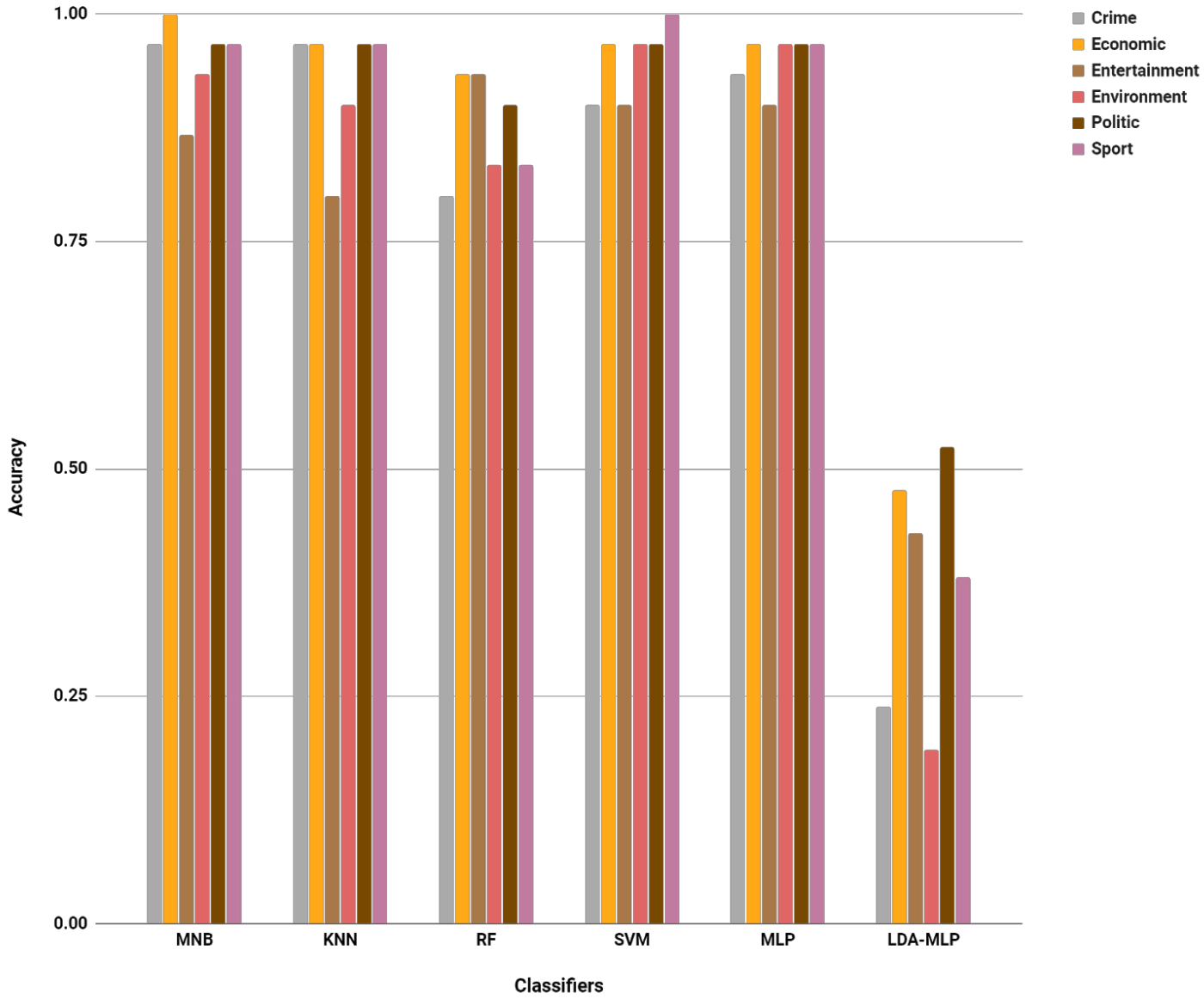
จากผลการประเมินประสิทธิภาพด้วยชุดทดสอบในตารางที่ 4.1-4.6 นำมาสรุปเปรียบเทียบประสิทธิภาพของตัวจำแนกประเภททั้งหมด 6 ตัวจำแนก ได้ดังตารางที่ 4.7 และรูปที่ 4.7 และสามารถสรุปได้ดังนี้

1. ความแม่นของตัวจำแนกแต่ละประเภท เมื่อพิจารณาพบว่า ตัวจำแนกที่ให้ค่าความแม่นสูงที่สุดมีอยู่ 3 ตัวด้วยกัน คือ MLP MNB และ SVM มีค่าเท่ากับ 95.00% ตามด้วยตัวจำแนก KNN, ตัวจำแนก RF และสุดท้ายคือตัวจำแนก LDA-MLP มีค่าความแม่นเท่ากับ 92.78%, 87.22% และ 37.30% ตามลำดับ

2. ความแม่นในตามประเภทหัวข้อข่าวที่ตัวจำแนกจำแนกได้ เมื่อพิจารณาพบว่า ค่าความแม่นตามประเภทหัวข้อข่าวที่ให้ค่าสูงสุดคือ ข่าวเศรษฐกิจโดยส่วนใหญ่ และตัวจำแนก MNB มีความสามารถจำแนกข่าวเศรษฐกิจสูงที่สุด สามารถจำแนกได้ถูกต้อง 100.00% ตามด้วยตัวจำแนก MLP และ KNN มีค่าความแม่นเท่ากันคือ 96.67% ในขณะที่ประเภทหัวข้อข่าวที่ถูกจำแนกได้ถูกต้องน้อยที่สุดเมื่อเทียบกับประเภทหัวข้อข่าวอื่น ๆ โดยส่วนใหญ่เป็นข่าวบันเทิง และตัวจำแนกที่จำแนกข่าวประเภทนี้ที่ให้ค่าความแม่นสูงสุด คือ ตัวจำแนก MLP และ SVM ที่ให้ค่าความแม่นอยู่ที่ 90.00% ตามด้วยตัวจำแนก MNB ที่ค่าความแม่น 86.67%

ตารางที่ 4.7 การเปรียบเทียบของตัวจำแนกประเภททั้งหมด 6 ตัวจำแนก ของชุดทดสอบ

	ค่าความแม่น					
ประเภทข่าว	MNB	KNN	RF	SVM	MLP	LDA-MLP
อาชญากรรม (Crime)	0.9667	0.9667	0.8000	0.9000	0.9333	0.2381
เศรษฐกิจ (Economic)	1.0000	0.9667	0.9333	0.9667	0.9667	0.4762
บันเทิง (Entertainment)	0.8667	0.8000	0.9333	0.9000	0.9000	0.4286
สิ่งแวดล้อม (Environment)	0.9333	0.9000	0.8333	0.9667	0.9667	0.1905
การเมือง (Politic)	0.9667	0.9667	0.9000	0.9667	0.9667	0.5238
กีฬา (Sport)	0.9667	0.9667	0.8333	1.0000	0.9667	0.3810
รวม	**0.9500**	**0.9278**	**0.8722**	**0.9500**	**0.9500**	**0.3730**

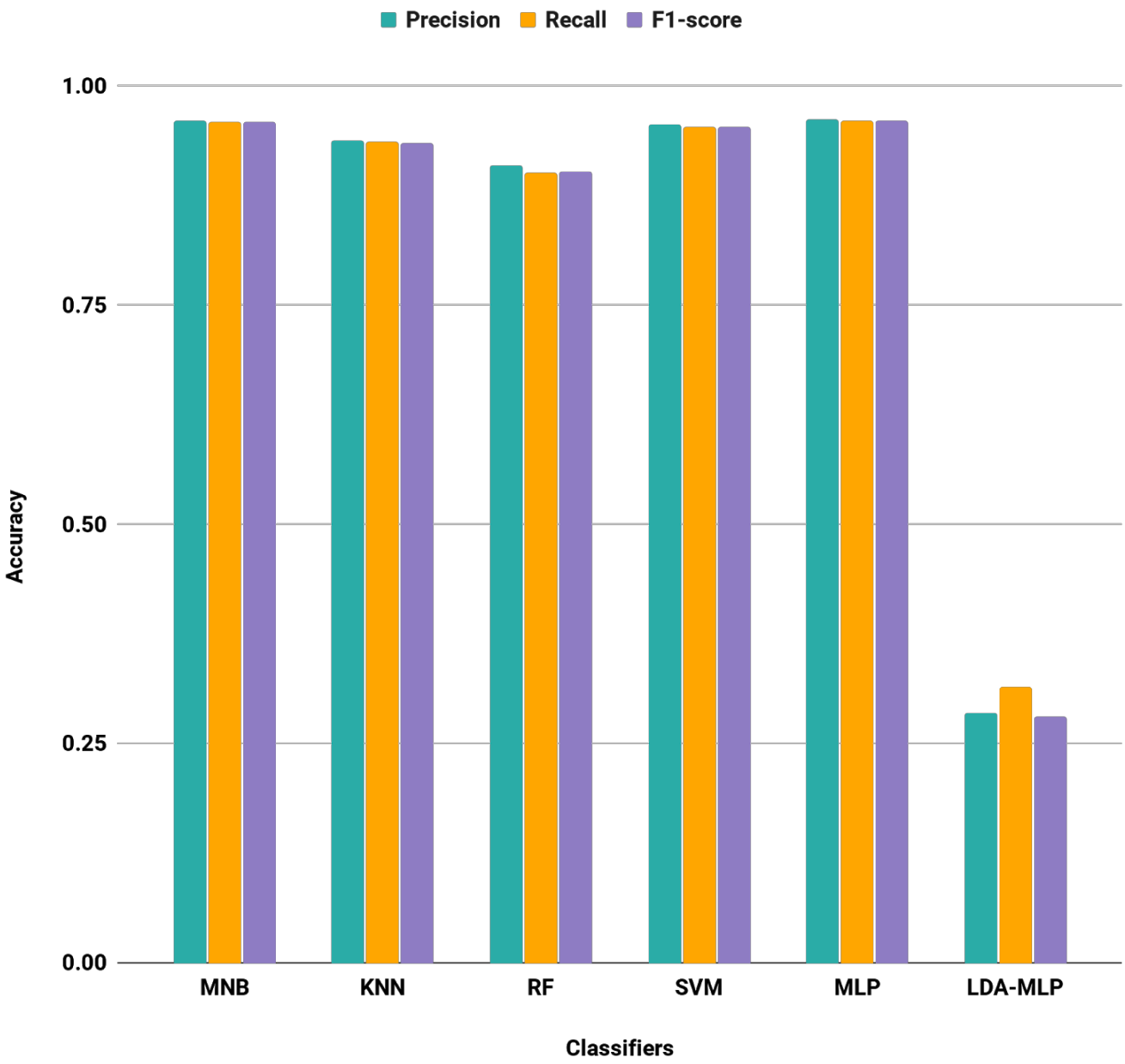


รูปที่ 4.7 กราฟค่าความแม่นของชุดทดสอบตามประเภทหัวข้อข่าวของแต่ละตัวจำแนก

ในการสรุปผลจากการทดลอง เพื่อที่จะทำให้ค่าความแม่นที่ได้จากการประเมินประสิทธิภาพมีความเสถียรมากที่สุด จึงนำตัวจำแนกที่สร้างจากชุดพารามิเตอร์ที่เหมาะสมที่หาได้จากขั้นตอนที่ 3.4 มาทำการตรวจสอบแบบไขว้ด้วยวิธีการสุ่มแบ่งแบบชั้นจำนวน 200 รอบ และใช้ผลการทำนายที่ได้จากทั้ง 200 รอบ มาคิดค่าความเที่ยงเฉลี่ย ค่าความครบถ้วนเฉลี่ย และคะแนน F1 เฉลี่ย ได้ผลลัพธ์ดังตารางที่ 4.8 และรูปที่ 4.7

ตารางที่ 4.8 ค่าความเที่ยงเฉลี่ย ค่าความครบถ้วนเฉลี่ย และคะแนน F1 เฉลี่ยของแต่ละตัวจำแนกประเภท

	การประเมินประสิทธิภาพ		
ตัวจำแนกประเภท	**Precision**	**Recall**	**F1-score**
MNB	0.9609	0.9595	0.9594
KNN	0.9381	0.9358	0.9354
RF	0.9090	0.9006	0.9017
SVM	0.9561	0.9534	0.9537
MLP	0.9622	0.9609	0.9609
LDA-MLP	0.2851	0.3150	0.2812



รูปที่ 4.8 กราฟค่าความเที่ยงเฉลี่ย ค่าความครบถ้วนเฉลี่ย และคะแนน F1 เฉลี่ยของแต่ละตัวจำแนก

จากตารางที่ 4.8 และกราฟในรูปที่ 4.8 สามารถสรุปได้ดังนี้

1. ตัวจำแนกประเภทที่มีค่าความเที่ยงเฉลี่ยสูงสุดคือตัวจำแนก MLP มีค่าเท่ากับ 96.22% ตามด้วย MNB และ SVM มีค่าเท่ากับ 96.09% และ 95.61% ตามลำดับ
2. ตัวจำแนกประเภทที่มีค่าความครบถ้วนเฉลี่ยสูงสุดคือตัวจำแนก MLP มีค่าเท่ากับ 96.09% ตามด้วย MNB และ SVM มีค่าเท่ากับ 95.95% และ 95.34% ตามลำดับ
3. ตัวจำแนกประเภทที่มีคะแนน F1 เฉลี่ยสูงสุดคือตัวจำแนก MLP มีค่าเท่ากับ 96.09% ตามด้วย MNB และ SVM มีค่าเท่ากับ 95.94% และ 95.37% ตามลำดับ

จากการตรวจสอบแบบไขว้ด้วยวิธีการสุ่มแบ่งแบบชั้นจะสามารถสรุปค่าความแม่นเฉลี่ย พร้อมค่าเบี่ยงเบนมาตรฐานของค่าความแม่นของตัวจำแนกประเภทแต่ละแบบ ได้ดังตารางที่ 4.9

ตารางที่ 4.9 ค่าความแม่นเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของตัวจำแนกประเภทแต่ละแบบ

ตัวจำแนกประเภท	ค่าความแม่นเฉลี่ย	ส่วนเบี่ยงเบนมาตรฐาน
MNB	0.9595	0.0119
KNN	0.9358	0.0154
RF	0.9006	0.0227
SVM	0.9534	0.0136
MLP	0.9609	0.0126
LDA-MLP	0.3150	0.0287

จากตารางที่ 4.9 จะได้ว่าประสิทธิภาพของการจำแนกประเภทคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อด้วยตัวจำแนกแบบต่าง ๆ MLP ให้ค่าความแม่นเฉลี่ยสูงสุดเท่ากับ 96.09% และมีส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 0.0126 ตามด้วยตัวจำแนก MNB และ SVM ซึ่งมีค่าความแม่นเฉลี่ยเท่ากับ 95.95% และ 95.34% และมีส่วนเบี่ยงเบนมาตรฐานอยู่ที่ 0.0119 และ 0.0136 ตามลำดับ

4.3 การอภิปรายผลการทดลอง

จากการสร้างตัวจำแนกประเภทเพื่อเปรียบเทียบประสิทธิภาพการจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง สามารถสรุปได้ดังนี้

จากผลการทดลองของตัวจำแนกประเภทที่สร้างจากขั้นตอนวิธีการเรียนรู้แบบมีผู้สอน เมื่อเปรียบเทียบค่าความแม่นเฉลี่ยของแต่ละตัวจำแนก จะเห็นได้ว่าตัวจำแนกประเภทที่ให้ผลดีที่สุดคือตัวจำแนกประเภท MLP ที่มีความแม่นเฉลี่ยเท่ากับ 96.09% โดยที่ตัวจำแนกอื่น ๆ ที่มีประสิทธิภาพรองลงมาคือตัวจำแนก MNB ตามด้วยตัวจำแนก SVM โดยมีความสามารถต่างกันเพียงเล็กน้อย จะเห็นว่ามีค่าความแม่นเฉลี่ยต่างกันเพียงไม่ถึง 1% ในทางกลับกัน จะเห็นได้ว่าค่าความแม่นเฉลี่ยของตัวจำแนก RF ได้ประสิทธิภาพไม่ดีเท่าตัวจำแนกอื่น ๆ อย่างเห็นได้ชัด ซึ่งอาจเป็นเพราะขั้นตอนวิธี RF ไม่เหมาะกับข้อมูลที่มีคุณลักษณะจำนวนมาก เนื่องจากยิ่งจำนวนของคุณลักษณะมากขึ้น ยิ่งทำให้เกิดการแบ่งโหนดมากขึ้น อีกทั้งทำให้ต้นไม้มีความลึกมากขึ้น จากรูปที่ 3.5 แสดงค่าความแม่นเฉลี่ยของตัวจำแนก RF ที่ได้จากการตรวจสอบแบบไขว้โดยเปรียบเทียบระหว่างค่าความแม่นของชุดสอนและชุดทดสอบ จะเห็นได้ว่าค่าความแม่นเฉลี่ยของชุดทดสอบมี

ค่าความแม่นที่ต่ำกว่าชุดสอนอย่างชัดเจน ซึ่งเป็นปัญหา Overfitting ที่ส่งผลให้ประสิทธิภาพในการจำแนกลดลง และอีกเหตุผลหนึ่งคือ เวกเตอร์คุณลักษณะ TF-IDF ที่ใช้เป็นข้อมูลนำเข้าของขั้นตอนวิธี RF อยู่ในรูปของเมทริกซ์มากเลขศูนย์ (Sparse Matrix) ดังนั้นในการเลือกคุณลักษณะที่จะใช้สำหรับการแบ่งโหนดของต้นไม้แต่ละครั้ง จึงมีโอกาสสูงที่จะได้เป็นคุณลักษณะที่มีค่าเป็น 0 จึงทำให้การแบ่งโหนดของต้นไม้ทำได้ไม่ดีนัก ด้วยเหตุนี้ตัวจำแนกประเภท RF จึงอาจไม่เหมาะกับงานในการจำแนกข้อความที่มีจำนวนคุณลักษณะสูง

สำหรับผลการวิจัยของตัวจำแนกประเภท LDA-MLP ที่สร้างจากขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอนและอาศัยขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนมาช่วยในการประเมินประสิทธิภาพ พบว่ามีค่าความแม่นเฉลี่ยเพียง 31.50% ซึ่งต่างกับตัวจำแนกประเภท MLP ที่เป็นตัวจำแนกที่ดีที่สุดมากถึง 60% จากผลการประเมินประสิทธิภาพของชุดทดสอบของตัวจำแนก LDA-MLP จะพบว่า LDA-MLP ให้ประสิทธิภาพในการจำแนกต่ำที่สุดในทุก ๆ ประเภทข่าว ซึ่งอาจมีสาเหตุมาจากการที่ขั้นตอนวิธี LDA เป็นขั้นตอนวิธีที่ใช้หลักการความน่าจะเป็นในการอธิบายการกระจายหัวข้อในแต่ละเอกสาร และเป็นขั้นตอนวิธีการเรียนรู้แบบไม่มีผู้สอน จึงไม่มีการใช้ประโยชน์จากป้ายกำกับหัวข้อข่าว ซึ่งต่างจากตัวจำแนกประเภทแบบอื่น ๆ จึงทำได้เพียงจัดกลุ่มให้กับข้อมูลอีกทั้งด้วยชุดคำโต้ตอบที่ใช้ในงานวิจัยนี้มีความยาวเฉลี่ยเพียง 50 คำต่อหนึ่งชุดคำโต้ตอบ ในขณะที่งานในการจัดกลุ่มหัวข้อทั่วไปที่ใช้ LDA มักใช้บทความข่าวที่โดยปกติแล้วมีความยาวอยู่ที่ประมาณ 400 คำต่อหนึ่งบทความ ดังนั้นด้วยจำนวนคำที่น้อยเกินไป อาจเป็นอุปสรรคต่อการหาการกระจายหัวข้อ เพราะจำนวนคำที่สังเกตได้มีอยู่น้อย ทำให้ไม่เพียงพอที่จะสรุปออกมาเป็นหัวข้อที่แม่นยำได้

จากการประเมินประสิทธิภาพการจำแนกประเภทคำโต้ตอบข่าวไทยออกเป็นประเภทของข่าว ได้แก่ ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม ข่าวบันเทิง ข่าวกีฬา และข่าวสิ่งแวดล้อม พบว่าการจำแนกชุดคำโต้ตอบของข่าวบางประเภทยังจำแนกได้ไม่ดีนัก ซึ่งอาจมีสาเหตุมาจากการที่เวกเตอร์คุณลักษณะที่ได้จากการคำนวณคะแนน TF-IDF สามารถบอกได้เพียงความสำคัญในการปรากฏของคำ ๆ หนึ่งที่มีต่อเอกสารเท่านั้น แต่ไม่สามารถบ่งบอกได้ถึงบริบทของคำที่อาจเปลี่ยนไปตามคำอื่น ๆ ในประโยค ดังนั้นการใช้เวกเตอร์คุณลักษณะ TF-IDF เพียงอย่างเดียว จึงไม่สามารถบอกได้ถึงประเภทหัวข้อข่าวที่แท้จริงที่แฝงอยู่ในสิ่งที่ผู้รายงานต้องการจะสื่อ ตัวอย่างชุดคำโต้ตอบ เช่น

ตัวอย่างที่ 1

“เฮ้ยๆเบาๆ คือพอขับแรงเนี่ย เดี๋ยวเค้าด่า คลื่นมันก็ซัดเข้าไปอะค่ะ ครับ
รถมอเตอร์ไซค์ที่จอดอยู่ล้มระเนระนาดเลยนะคะ นี่อย่างกับทะเลเลยเนี่ยค่ะ”

จากตัวอย่างชุดคำโต้ตอบข้างต้น จะเห็นว่ามีคำเด่น ๆ ที่ปรากฏในชุดคำโต้ตอบนี้คือ “รถมอเตอร์ไซค์” “ล้มระเนระนาด” ทำให้ตัวจำแนกทำนายได้ว่าเป็นข่าวอาชญากรรม ซึ่งผิดพลาดจากประเภทหัวข้อข่าวที่แท้จริงที่แฝงอยู่คือข่าวสิ่งแวดล้อม

ตัวอย่างที่ 2

“คือภาคอีสาน ภาคเหนือ ภาคตะวันออก ภาคกลางเนี่ยทุกพื้นที่ต้องควรเฝ้าระวังคาดการณ์ที่ท่านอธิบดีบอกว่ามีสองเส้นทางคุณผู้ชมลองดูนะคะ เดี๋ยวดูจากภาพกราฟฟิคนะคะ แบบจำลองที่หนึ่งคือ”

จากตัวอย่างชุดคำโต้ตอบข้างต้น จะเห็นว่ามีคำเด่น ๆ ที่ปรากฏในชุดคำโต้ตอบนี้คือ “ท่าน” “เฝ้า” “ระวัง” ทำให้ตัวจำแนกทำนายได้ว่าเป็นข่าวการเมือง ซึ่งผิดพลาดจากประเภทหัวข้อข่าวที่แท้จริงที่แฝงอยู่คือข่าวสิ่งแวดล้อม

ตัวอย่างที่ 3

“มันก็มีแต่ไม่ชัดเท่าเดิม ใช่ครับ จริงๆน้ำหนักขึ้นนะครับ ขึ้นมาสองโล แต่ว่าเหมือนพอเราเต้นแล้วก็ออกกำลังกาย มันเหมือนการคาดิโอตลอดเวลา ก็เลยทำให้ยังมีอยู่ ใช่”

จากตัวอย่างชุดคำโต้ตอบข้างต้น จะเห็นว่ามีคำเด่น ๆ ที่ปรากฏในชุดคำโต้ตอบนี้คือ “น้ำหนัก” “ออกกำลังกาย” ทำให้ตัวจำแนกทำนายได้ว่าเป็นข่าวกีฬา ซึ่งผิดพลาดจากประเภทหัวข้อข่าวที่แท้จริงที่แฝงอยู่คือข่าวบันเทิง

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

ในบทนี้จะกล่าวถึงการสรุปผลการวิจัยการสร้างตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง และข้อเสนอแนะ โดยมีรายละเอียดดังนี้

5.1 สรุปผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อสร้างตัวจำแนกและเปรียบเทียบประสิทธิภาพตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อ ตัวจำแนกถูกสร้างขึ้นจากเวกเตอร์คุณลักษณะที่ได้จากการคำนวณคะแนน TF-IDF และเวกเตอร์ความน่าจะเป็นของหัวข้อสำหรับแต่ละชุดคำโต้ตอบที่หาได้จากขั้นตอนวิธี LDA สำหรับเทคนิคในการจำแนกประเภทหัวข้อข่าวนั้น ใช้เทคนิคการเรียนรู้ของเครื่องที่หลากหลาย ทั้งแบบขั้นตอนวิธีการเรียนรู้แบบมีผู้สอนและแบบไม่มีผู้สอน เพื่อจำแนกประเภทคำโต้ตอบข่าวไทยออกเป็นประเภทของข่าวทั้งหมด 6 ประเภทด้วยกัน ได้แก่ ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม ข่าวบันเทิง ข่าวกีฬา และข่าวสิ่งแวดล้อม โดยขั้นตอนในสร้างตัวจำแนกมีดังนี้ การรวมรวบรวมข้อมูล การเตรียมข้อมูล การสกัดคุณลักษณะ และการหาพารามิเตอร์ที่เหมาะสม แล้วจึงนำไปสร้างตัวจำแนก รวม 6 ตัวจำแนกจากขั้นตอนวิธีที่แตกต่างกัน ได้แก่ MNB, KNN, RF, SVM, MLP และ LDA-MLP

จากการเปรียบเทียบประสิทธิภาพการจำแนกคำตอบข่าวไทยของแต่ละตัวจำแนก พบว่าตัวจำแนกที่สามารถจำแนกคำตอบข่าวไทยได้ดีที่สุดคือตัวจำแนก MLP ซึ่งมีค่าความแม่นเท่ากับ 96.09% ตามด้วยตัวจำแนก MNB, SVM, KNN, RF และ LDA-MLP ซึ่งมีค่าความแม่นเท่ากับ 95.95%, 95.34%, 93.58%, 90.06% และ 31.50% ตามลำดับ

5.2 ข้อเสนอแนะ

จากการทดลองสร้างตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยมีข้อเสนอแนะดังต่อไปนี้

1. ในการทดลองสร้างตัวจำแนกประเภทของชุดข้อมูลประเภทข้อความ ผู้ทดลองควรให้ความสำคัญกับขั้นตอนการเตรียมข้อมูลเป็นอย่างมาก เนื่องจากขั้นตอนนี้เป็นขั้นตอนที่ช่วยกำจัดส่วนที่ไม่มีประโยชน์หรือไม่เกี่ยวข้องกับเนื้อหาออกไป ดังนั้นการดำเนินงานในขั้นตอนนี้จึงส่งผลต่อลักษณะของเวกเตอร์คุณลักษณะที่สกัดได้เป็นอย่างมาก จึงต้องให้ความสำคัญเป็นอย่างยิ่ง ตัวอย่างเช่น ในขั้นตอนการลบคำฟุ่มเฟือย และการตัดคำ โดยปกติจะอาศัยพจนานุกรมที่อยู่ในไลบารี PyThaiNLP มาใช้ในการตัดคำและลบคำฟุ่มเฟือยออกจากชุดข้อมูล แต่คำศัพท์ที่อยู่พจนานุกรมอาจไม่เพียงพอหรือไม่คร อบคลุม

กับชุดข้อมูลที่มีอยู่ ดังนั้นจึงควรเพิ่มคำลงในพจนานุกรม โดยคำที่เพิ่มจะเป็นคำเฉพาะหรือคำที่ไม่มีอยู่ในพจนานุกรมทั่วไป เช่น “คุณผู้ชม” “ฮะ” “เงี้ย” “แล้วก็”

2. เนื่องจากลักษณะเนื้อหาของข่าวบางประเภทจะขึ้นอยู่กับช่วงเวลาหรือยุคสมัยที่ออกอากาศ ดังนั้นการกำหนดเงื่อนไขอายุข่าวจึงช่วยให้ชุดคำโต้ตอบที่รวบรวมได้มีลักษณะของเนื้อหาไปในทิศทางเดียวกัน ซึ่งจะส่งผลต่อความสามารถในการจำแนก ตัวอย่างเช่น ข่าวการเมืองในช่วงปี 2561 ที่นับเป็นปีแห่งการเลือกตั้ง ข่าวที่รวบรวมได้ส่วนใหญ่จะมีเนื้อหาเกี่ยวกับการเลือกตั้ง กฎเกณฑ์หรือข้อปฏิบัติในการเลือกตั้ง เป็นต้น ในขณะที่ข่าวการเมืองในปี 2563 ส่วนใหญ่จะมีเนื้อหาเกี่ยวกับการชุมนุม การประกาศข้อเรียกร้อง การดำเนินคดีกับผู้ชุมนุม เป็นต้น ดังนั้นสำหรับการทดลองที่มีเวลาจำกัด จึงควรกำหนดขอบเขตของอายุข่าวให้ไม่มากเกินไป เพื่อช่วยลดความหลากหลายของเนื้อข่าวได้ในระดับหนึ่ง โดยในงานวิจัยนี้ได้กำหนดอายุข่าวไว้ที่ไม่เกิน 2 ปี นับจากเวลาที่ดำเนินการเก็บรวบรวมข้อมูล (เดือนตุลาคม 2561 ถึง ตุลาคม 2563)

3. การสกัดคุณลักษณะเพื่อสร้างเวกเตอร์คุณลักษณะสามารถทำได้หลายเทคนิค เช่น Wordcount TF-IDF หรือ n-grams เป็นต้น หากต้องการเพิ่มประสิทธิภาพของการจำแนกโดยพิจารณาถึงบริบทของคำใกล้เคียง สามารถสร้างเวกเตอร์คุณลักษณะด้วยเทคนิค n-grams ได้ ซึ่งเป็นเทคนิคที่ช่วยควบคุมขนาดของเมทริกซ์คุณลักษณะได้ โดยขึ้นอยู่กับค่า n ที่เลือกใช้ เช่น bigrams (n = 2) หรือ trigrams (n = 3) เป็นต้น โดยที่ค่า n สูง ๆ มีผลทำให้เมทริกซ์คุณลักษณะมีขนาดใหญ่ขึ้น

5.3 ปัญหาและอุปสรรค

เนื่องจากชุดข้อมูลคำโต้ตอบข่าวไทยไม่มีแหล่งข้อมูลแบบเปิดที่นำมาใช้งานได้ ดังนั้นในงานวิจัยนี้จึงได้ทำการเก็บรวบรวมข้อมูลขึ้นใหม่ด้วยตนเอง โดยสำหรับชุดข้อมูลประเภทคำโต้ตอบข่าวจำเป็นต้องรวบรวมจากวิดีโอย้อนหลังของรายการข่าว ทำให้ขั้นตอนการคัดเลือกวิดีโอรายการข่าวและการเก็บรวบรวมข้อมูลนั้นใช้เวลานานพอสมควร อีกทั้งหลังจากที่คัดเลือกวิดีโอข่าวและดำเนินการแปลงเสียงเป็นข้อความ ยังต้องทำการตรวจสอบความถูกต้องของคำที่ได้หลังจากแปลงเสียงเป็นข้อความอีกด้วย ดังนั้นด้วยเวลาดำเนินการวิจัยที่มีอยู่อย่างจำกัด จึงสามารถรวบรวมปริมาณชุดข้อมูลและความยาวต่อชุดคำโต้ตอบได้เพียงจำนวนหนึ่งเท่านั้น ซึ่งอาจเป็นสาเหตุให้ผลการจำแนกของตัวจำแนก LDA-MLP มีค่าไม่ดีเท่าที่ควรจะเป็น

เอกสารอ้างอิง

[1] กสทช., “พัฒนาการความนิยม รายการข่าว เทียบกับ รายการบันเทิง,” 2563. [ออนไลน์]. เข้าถึงได้จาก: http://broadcast.nbtc.go.th/academic/?type=NjIwNzAwMDAwMDAy. [สืบค้นเมื่อวันที่ 20 ตุลาคม 2563].

[2] scikit-learn. (n.d.). “sklearn.feature_extraction.text.TfidfTransformer”. [ONLINE] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfTransformer.html#sklearn.feature_extraction.text.TfidfTransformer. [Accessed on 20 February 2021].

[3] Liu, R., Yang, B., Zio, E., & Chen, X. (2018). Artificial intelligence for fault diagnosis of rotating machinery: A review. Mechanical Systems and Signal Processing, 108, 33–47. doi: 10.1016/j.ymssp.2018.02.016.

[4] Huwaidah, A., Adiwijaya, & Faraby, S. A. (2021). Argument Identification in Indonesian Tweets on the Issue of Moving the Indonesian Capital. Procedia Computer Science, 179, 407–415. doi: 10.1016/j.procs.2021.01.023.

[5] scikit-learn. (n.d.). “Naive Bayes”. [online] Available at: https://scikit-learn.org/stable/modules/naive_bayes.html#naive-bayes. [Accessed 20 February 2021].

[6] Jeffers, J., Reinders, J., & Sodani, A. (2016). Chapter 24 - Machine learning. In Intel Xeon Phi Processor High Performance Programming: Knights Landing Edition 2nd Edition (2nd ed., pp. 527–548). Morgan Kaufmann. doi: 10.1016/B978-0-12-809194-4.00024-7.

[7] Kramer O. (2013) K-Nearest Neighbors. In: Dimensionality Reduction with Unsupervised Nearest Neighbors. Intelligent Systems Reference Library, vol 51. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-38652-7_2.

[8] Md, C. S. (2020). Chapter 8 - Precision medicine in digital pathology via image analysis and machine learning. In Artificial Intelligence and Deep Learning in Pathology (1st ed., pp. 149–173). Elsevier. doi: 10.1016/B978-0-323-67538-3.00008-7.

[9] Buyya, R., Calheiros, R. N., & Dastjerdi, V. A. (2016). Chapter 15 - A Case Study in Big Data Analytics: Exploring Twitter Sentiment Analysis and the Weather. In Big Data: Principles and Paradigms (1st ed., pp. 357–388). Morgan Kaufmann. doi: 10.1016/B978-0-12-805394-2.00015-5.

[10] Malek, Sorayya & Hui, Cham & Na, Aziida & Cheen, Song & Toh, Sooh & Milow, Pozi. (2018). Ecosystem Monitoring Through Predictive Modeling. doi: 10.1016/B978-0-12-809633-8.20060-5.

[11] Huwaidah, A., Adiwijaya, & Faraby, S. A. (2021). Argument Identification in Indonesian Tweets on the Issue of Moving the Indonesian Capital. Procedia Computer Science, 179, 407–415. doi: 10.1016/j.procs.2021.01.023.

[12] Dukart, Juergen & Roche, F.. (2015). Basic Concepts of Image Classification Algorithms Applied to Study Neurodegenerative Diseases. Brain Mapping: An Encyclopedic Reference. 3. doi: 641-646. 10.1016/B978-0-12-397025-1.00072-5.

[13] S. Ghosh, A. Dasgupta & A. Swetapadma, "A Study on Support Vector Machine based Linear and Non-Linear Pattern Classification," 2019 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2019, pp. 24-28, doi: 10.1109/ISS1.2019.8908018.

[14] Menzies, T., Kocaguneli, E., Turhan, B., Minku, L., & Peters, F. (2014). Chapter 24 - Using Goals in Model-Based Reasoning. In: Sharing Data and Models in Software Engineering (1st ed., pp. 321–353). Morgan Kaufmann. doi: 10.1016/B978-0-12-417295-1.00024-2.

[15] Abirami S, Chitra P. (2020). Chapter Fourteen - Energy-efficient edge based real-time healthcare support system. In: 1206714775 899858081 P. (Ed.), Advances in Computers (Vol. 117, pp. 336-368). doi: 10.1016/bs.adcom.2019.09.007.

[16] DM Blei, AY Ng, MI Jordan. (2003). “Latent dirichlet allocation” Journal of Machine Learning Research 3: 993-1022.

[17] Q. Wang, R. Peng, J. Wang, Y. Xie & Y. Zhou, "Research on Text Classification Method of LDA- SVM Based on PSO optimization," 2019 Chinese Automation Congress (CAC), Hangzhou, China, 2019, pp. 1974-1978, doi: 10.1109/CAC48633.2019.8996952.

[18] scikit-learn. (n.d.). 2.5. Decomposing signals in components (matrix factorization problems). [online] Available at: https://scikit-learn.org/stable/modules/decomposition.html#latent-dirichlet-allocation-lda. [Accessed on 20 February 2021].

[19] Tharwat, Alaa. (2018). Classification Assessment Methods: a detailed tutorial. doi: 10.1016/j.aci.2018.08.003.

[20] Batarseh, F. A., & Yang, R. (2020). 5 - Foundations of data imbalance and solutions for a data democracy. In Data Democracy: At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering (1st ed., pp. 83–163). Academic Press. doi: 10.1016/B978-0-12-818366-3.00005-8.

[21] Powers, David & Ailab. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J. Mach. Learn. Technol. 2. 2229-3981. doi: 10.9735/2229-3981.

[22] A. Noppakaow & O. Uchida, “Examinations on the Performance of Classification Models for Thai News Articles,” 2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE), Pattaya, Thailand, 2019, pp. 1-4, doi: 10.1109/ICITEED.2019.8929959.

[23] W. Jirasirilerd & P. Tangtisanon, “Automatic Labeling for Thai News Articles Based on Vector Representation of Documents,” 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST), Phuket, 2018, pp. 1-4, doi: 10.1109/ICEAST.2018.8434457.

[24] S. Lee, J. Kim & S. Myaeng, “An extension of topic models for text classification: A term weighting approach,” 2015 International Conference on Big Data and Smart Computing (BIGCOMP), Jeju, 2015, pp. 217-224, doi: 10.1109/35021BIGCOMP.2015.7072834.

[25] D. E. Cahyani & K. A. P. Nuzry, “Trending Topic Classification for Single-Label Using Multinomial Naive Bayes (MNB) and Multi-Label Using K-Nearest Neighbors (KNN),” 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 2019, pp. 547-552, doi: 10.1109/ICITISEE48480.2019.9003944.

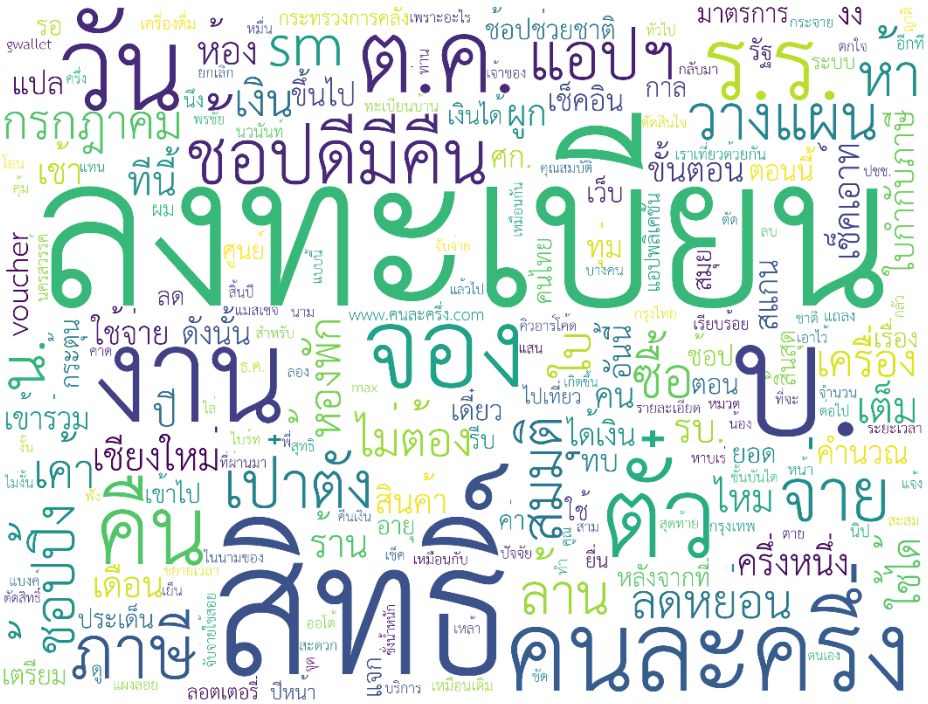
[26] A. Fesseha, S. Xiong, E. D. Emiru & A. Dahou, “Text Classification of News Articles Using Machine Learning on Low-resourced Language: Tigrigna,” 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2020, pp. 34-38, doi: 10.1109/ICAIBD49809.2020.9137443.

[27] Nicolas, Z.C., 2013. “Learning Multi-Label Topic Classification Of News Articles”. [online] Semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Learning-Multi-Label-Topic-Classification-of-News-Nicolas/e5490a559c56df10715542f2960767a2abdb8459. [Accessed on 18 October 2020].

ภาคผนวก

ภาคผนวก ก
รูปแสดงกลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของแต่ละประเภทหัวข้อข่าว

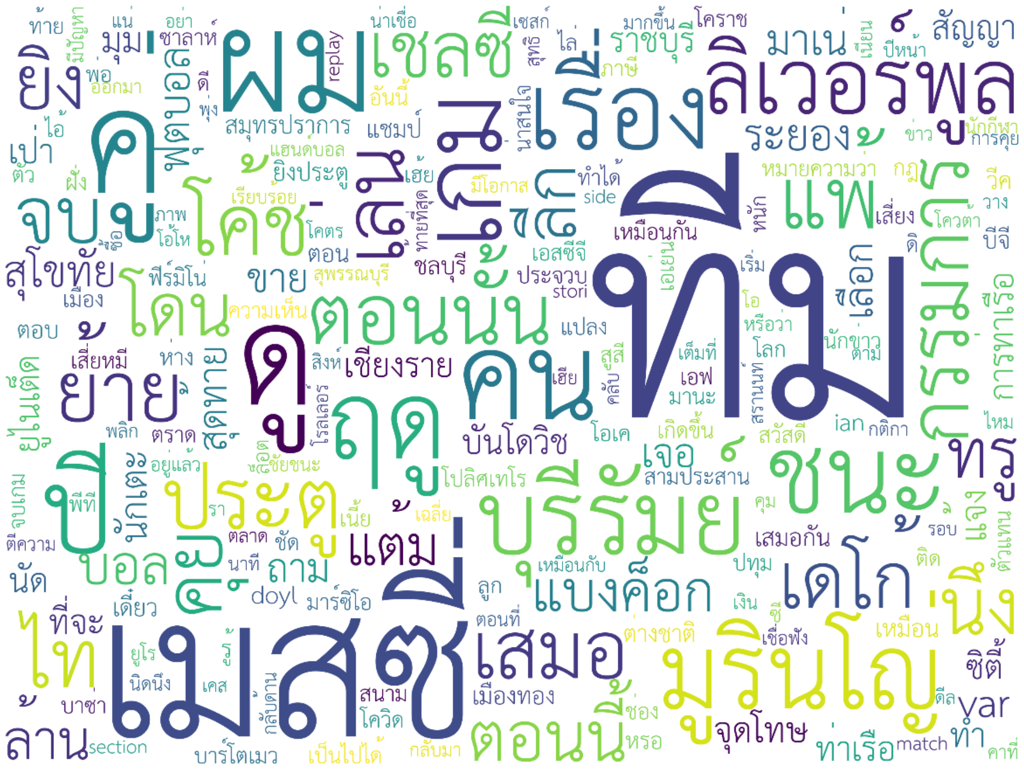
รูปที่ 1 กลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของข่าวอาชญากรรม



รูปที่ 2 กลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของข่าวเศรษฐกิจ

รูปที่ 3 กลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของข่าวบันเทิง

รูปที่ 4 กลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของข่าวสิ่งแวดล้อม



รูปที่ 5 กลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของข่าวกีฬา

รูปที่ 6 กลุ่มคำของชุดสอนที่แยกลำดับตามคะแนน TF-IDF ของข่าวการเมือง

ภาคผนวก ข
แบบเสนอหัวข้อโครงงาน รายวิชา 2301499 Project Proposal
ปีการศึกษา 2563

ชื่อโครงงาน (ภาษาไทย)	การจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง
ชื่อโครงงาน (ภาษาอังกฤษ)	Classifying Thai News Dialogues into Topic Types Using Machine Learning Technique
อาจารย์ที่ปรึกษา	1. รองศาสตราจารย์ ดร.ศุภกานต์ พิมลธเรศ 2. ผู้ช่วยศาสตราจารย์ ศศิภา พันธุวดีธร
ผู้ดำเนินการ	1. นางสาวศลิษา ชูชื่นพฤกษาพันธ์ เลขประจำตัวนิสิต 6033661023 2. นางสาวไอศวรรย์ ธโนศวรรย์ เลขประจำตัวนิสิต 6033673523 สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

หลักการและเหตุผล

ในทุกวินาทีของโลกปัจจุบันมีเหตุการณ์ใหม่ ๆ ที่น่าสนใจเกิดขึ้นอยู่ตลอดเวลา ซึ่งการติดตามเหตุการณ์ความเปลี่ยนแปลงที่เกิดขึ้นที่รวดเร็วและมีประสิทธิภาพนั้นต้องอาศัยการติดตามข่าวสารผ่านสื่อต่าง ๆ เช่น หนังสือพิมพ์รายวัน บทความข่าวบนเว็บไซต์ โดยเฉพาะอย่างยิ่ง “รายการข่าว” ไม่ว่าจะเป็นรายการข่าวโทรทัศน์ วิทยุ รวมไปถึงบนแพลตฟอร์มออนไลน์ จากการสำรวจของ กสทช. [1] พบว่าความนิยมรายการข่าวของคนไทยในปี 2563 ยังคงเป็นไปอย่างต่อเนื่อง โดยจุดเด่นของการรับข่าวสารผ่านรายการข่าวคือผู้ชมจะได้ฟังการรายงานข่าวจากผู้รายงานข่าว ที่ผ่านการคัดกรอง เรียบเรียง วิเคราะห์ และสรุปมาจากแหล่งข่าวที่เชื่อถือได้ รายการข่าวจึงเป็นสื่อที่มีความสำคัญอย่างมากในสังคมปัจจุบัน อย่างไรก็ตามในรายการข่าวหนึ่งรายการมักจะมีการนำเสนอข่าวในหลากหลายหัวข้อ เช่น ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวบันเทิง ข่าวอาชญากรรม ข่าวต่างประเทศ และอื่น ๆ ทั้งหมดนี้ถูกรวมอยู่ในรายการข่าวเดียวกัน อีกทั้งบางรายการข่าวมีความยาวมากถึง 2 ชั่วโมง จึงยากสำหรับคนที่ต้องการเลือกฟังข่าวแบบเจาะจงหัวข้อ เนื่องจากต้องเสียเวลาในการค้นหาข่าวที่ตนสนใจ

ในการศึกษาเกี่ยวกับการจำแนกประเภทบทความข่าวออนไลน์ของไทย จากงานวิจัยของ Arisara Noppakaow และคณะ [2] ได้ศึกษาเกี่ยวกับแบบจำลองอัตโนมัติเพื่อจำแนกบทความข่าวออกเป็น

4 ประเภท ได้แก่ ข่าวอาชญากรรม ข่าวการเมือง ข่าวกีฬา และข่าวบันเทิง ด้วยขั้นตอนวิธี Decision Tree, Support Vector Machine (SVM) และ Deep Learning models มีค่าความแม่นอยู่ที่ 86% 94% และ 95% ตามลำดับ และจากงานวิจัยของ Wiphada Jirasirilerd และ Pikulkaew Tangtisanon [3] ได้ศึกษาเกี่ยวกับวิธีการติดป้ายกำกับอัตโนมัติสำหรับบทความข่าวบนเว็บไซต์ภาษาไทยโดยใช้การกระจายข้อความในเอกสาร โดยสกัดคำที่มีความหมายคล้ายกันจากเวกเตอร์ย่อหน้าของข่าวแต่ละประเภท และนำไปกำหนดเป็นป้ายกำกับ เนื่องจากโปรแกรมแบ่งกลุ่มคำภาษาไทยที่มีอยู่นั้นให้อัตราของค่าความแม่นที่ต่ำ จึงได้ใช้โครงข่ายประสาทแบบคอนโวลูชันร่วมกับวิธีการจำแนกแบบไบนารีในการแยกคำออกจากประโยคเพื่อประสิทธิภาพที่ดีขึ้น และพบว่าแบบจำลองเวกเตอร์ที่นำเสนอให้ค่าความแม่นที่ดีกว่าแบบจำลองเวกเตอร์อื่น ๆ

จากงานวิจัยของ Nicolas และ Zach CHASE [4] ได้ศึกษาเกี่ยวกับการจำแนกประเภทหัวข้อของบทความข่าวที่มีป้ายกำกับหัวข้อที่เกี่ยวข้องหลายรายการ และวิเคราะห์ข้อบกพร่องของขั้นตอนวิธีต่าง ๆ รวมถึง Naive Bayes แล้วผู้วิจัยได้นำเสนอตัวจำแนก Naive Bayes แบบหนึ่งต่อทั้งหมด โดยใช้ตัวจำแนกหนึ่งตัวต่อคลาส ซึ่งได้ประสิทธิภาพมากกว่าวิธีการที่ได้จาก Term Frequency–Inverse Document Frequency (TF-IDF)

งานวิจัยของ Seonggyu Lee และคณะ [5] ได้เพิ่มประสิทธิภาพของการจำแนกประเภทเอกสาร โดยนำเสนอวิธีการที่พัฒนามาจากพื้นฐานของ Latent Dirichlet Allocation (LDA) พร้อมกับพิจารณาน้ำหนักของคำในการสุ่มตัวอย่างและเพิ่มความสมดุลในการกระจายหัวข้อ โดยทดลองกับชุดข้อมูล 20 Newsgroups ซึ่งเป็นชุดข้อมูลที่รวบรวมเอกสารกลุ่มข่าวไว้จำนวน 20 ประเภท ผลการทดลองแสดงให้เห็นว่าการสร้างแบบจำลองหัวข้อแบบถ่วงน้ำหนักสมดุล (Balance Weighted Topic Modeling) ทำให้ได้คุณลักษณะที่ช่วยให้การจำแนกประเภทเอกสารมีประสิทธิภาพดีขึ้น

จากที่กล่าวมาข้างต้น ทำให้ผู้พัฒนามีความสนใจที่จะศึกษาเกี่ยวกับการจำแนกคำโต้ตอบข่าวภาษาไทยตามชนิดหัวข้อข่าว เพื่อบอกประเภทหัวข้อข่าวว่าเป็นข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม ข่าวบันเทิง ข่าวกีฬา หรือ ข่าวสิ่งแวดล้อม ซึ่งในการพัฒนา ผู้พัฒนาจะเก็บชุดข้อมูลเสียงโต้ตอบระหว่างผู้รายงานข่าวจากวิดีโอย้อนหลังของรายการข่าวและแปลงเสียง
เป็นข้อความด้วยวิธี Speech-to-Text (STT) แล้วจึงนำข้อความมาสกัดเป็นเวกเตอร์คุณลักษณะ (feature vector) จากนั้นจึงสร้างแบบจำลองสำหรับจำแนกหัวข้อข่าวโดยใช้เทคนิคการเรียนรู้ของเครื่อง เพื่อนำเสนอวิธีที่สามารถจำแนกประเภทข่าวจากคำโต้ตอบที่มีประสิทธิภาพ

วัตถุประสงค์

เพื่อพัฒนาตัวจำแนกคำโต้ตอบข่าวไทยเป็นแบบชนิดหัวข้อโดยใช้เทคนิคการเรียนรู้ของเครื่อง

ขอบเขตของโครงงาน

1. ข้อมูลนำเข้าคือชุดคำโต้ตอบของข่าวภาษาไทย
2. ชุดคำโต้ตอบรวบรวมจากวิดีโอรายการข่าวไทย โดยมีอายุข่าวย้อนหลังไม่เกิน 2 ปี คือตั้งแต่เดือนตุลาคม 2561 ถึง ตุลาคม 2563 และมีผู้รายงานข่าว 2 คนขึ้นไป
3. ชุดคำโต้ตอบแบ่งตามประเภทหัวข้อข่าวได้ 4 ประเภท ได้แก่ ข่าวการเมือง ข่าวเศรษฐกิจ ข่าวอาชญากรรม และข่าวบันเทิง
4. จำนวนชุดคำโต้ตอบประเภทละ 100 ชุด
5. ชุดคำโต้ตอบแต่ละชุดมีระยะเวลาไม่เกิน 15 วินาที

วิธีการดำเนินงาน

1. ศึกษาค้นคว้าทฤษฎีที่เกี่ยวข้อง
2. ศึกษางานวิจัยที่เกี่ยวข้อง
3. เตรียมชุดข้อมูลคำโต้ตอบข่าวไทย
4. วิเคราะห์ ออกแบบ และสร้างตัวแบบในการจำแนกประเภทหัวข้อข่าว
5. ทดสอบประสิทธิภาพของตัวแบบ
6. วิเคราะห์และอภิปรายผล
7. จัดทำเอกสาร

ตารางเวลาการดำเนินงาน

ขั้นตอนการดำเนินงาน	ปี 2563					ปี 2564			
	ส.ค.	ก.ย.	ต.ค	พ.ย.	ธ.ค.	ม.ค.	ก.พ.	มี.ค.	เม.ย.
1. ศึกษาค้นคว้าทฤษฎีที่เกี่ยวข้อง									
2. ศึกษางานวิจัยที่เกี่ยวข้อง									
3. เตรียมชุดข้อมูลคำโต้ตอบข่าวไทย									
4. วิเคราะห์ ออกแบบ และสร้างตัวแบบในการจำแนกประเภทหัวข้อข่าว									
5. ทดสอบประสิทธิภาพของตัวแบบ									
6. วิเคราะห์และอภิปรายผล									
7. จัดทำเอกสาร									

ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่ได้ต่อผู้พัฒนา

1. ได้ฝึกฝนและพัฒนาทักษะในการสร้างตัวจำแนกด้วยเทคนิคการเรียนรู้ของเครื่อง
2. ได้ฝึกฝนและพัฒนาทักษะการวางแผนและทำงานเป็นขั้นตอน
3. ได้ฝึกการทำงานเป็นกลุ่ม การยอมรับความคิดเห็นผู้อื่น และความรับผิดชอบในหน้าที่
4. ได้พัฒนาศักยภาพในการเรียนรู้ด้วยตัวเอง

ประโยชน์ที่ได้ต่อผู้ใช้

1. ได้แนวทางในการจำแนกประเภทหัวข้อข่าวจากชุดคำโต้ตอบ
2. สามารถนำชุดข้อมูลข่าวไทย และแนวทางที่เสนอไปพัฒนาต่อยอดได้

อุปกรณ์และเครื่องมือที่ใช้

1. ฮาร์ดแวร์

   คอมพิวเตอร์พกพาที่มีหน่วยประมวลผลกลางความเร็วไม่ต่ำกว่า 2.5 GHz ความจุของหน่วยความจำไม่ต่ำกว่า 8 GB และหน่วยความจำสำรองความจุไม่ต่ำกว่า 256 GB

2. ซอฟต์แวร์

   2.1 ระบบปฏิบัติการ Windows 10

   2.2 Visual Studio Code เวอร์ชัน 1.50.1

   2.3 Google Colab Python notebook

   2.4 ไลบรารี Python สำหรับการทำงานและคำนวณข้อมูล เช่น Scikit-learn, PyThaiNLP

งบประมาณ

1.	ฮาร์ดดิสก์ชนิด SSD แบบพกพา ความจุ 500 GB	2 ชิ้น	7,000 บาท
2.	เมาส์ไร้สาย	2 ชิ้น	2,000 บาท
3.	อุปกรณ์ต่อพ่วง USB	2 ชิ้น	1,000 บาท
		รวม	**10,000 บาท**

สามารถถัวเฉลี่ยได้ทุกรายการ

เอกสารอ้างอิง

[1] กสทช., “พัฒนาการความนิยม รายการข่าว เทียบกับ รายการบันเทิง,” 2563. [ออนไลน์]. เข้าถึงได้จาก: http://broadcast.nbtc.go.th/academic/?type=NjIwNzAwMDAwMDAy. [สืบค้นเมื่อวันที่ 20 ตุลาคม 2563].

[2] A. Noppakaow and O. Uchida, "Examinations on the Performance of Classification Models for Thai News Articles," *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Pattaya, Thailand, 2019, pp. 1-4, doi: 10.1109/ICITEED.2019.8929959.

[3] W. Jirasirilerd and P. Tangtisanon, "Automatic Labeling for Thai News Articles Based on Vector Representation of Documents," *2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST)*, Phuket, 2018, pp. 1-4, doi: 10.1109/ICEAST.2018.8434457.

[4] Nicolas, Z.C., 2013. *Learning Multi-Label Topic Classification Of News Articles*. [online] Semanticscholar.org. Available at: https://www.semanticscholar.org/paper/Learning-Multi-Label-Topic-Classification-of-News-Nicolas/e5490a559c56df10715542f2960767a2abdb8459. [Accessed 18 October 2020].

[5] S. Lee, J. Kim and S. Myaeng, "An extension of topic models for text classification: A term weighting approach," *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, Jeju, 2015, pp. 217-224, doi: 10.1109/35021BIGCOMP.2015.7072834.

ประวัติผู้เขียน

ชื่อ-สกุล ศลิษา ชูชื่นพฤกษาพันธ์

รหัสประจำตัวนิสิต 6033661023

วัน เดือน ปี เกิด 24 มิถุนายน 2542

สถานที่เกิด กำแพงเพชร

วุฒิการศึกษา กำลังศึกษาในระดับปริญญาตรี สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ที่อยู่ปัจจุบัน 90/113 ซ.โกสุมรวมใจ38 แขวงดอนเมือง เขตดอนเมือง จ.กรุงเทพมหานคร 10210

ชื่อ-สกุล ไอศวรรย์ ธโนศวรรย์

รหัสประจำตัวนิสิต 6033673523

วัน เดือน ปี เกิด 22 เมษายน 2542

สถานที่เกิด กรุงเทพมหานคร

วุฒิการศึกษา กำลังศึกษาในระดับปริญญาตรี สาขาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ที่อยู่ปัจจุบัน 59/793 หมู่ 5 ต.ลาดสวาย อ.ลำลูกกา จ.ปทุมธานี 12150