



## การวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์มออนไลน์

โดย

นางสาวจันทร์ตราพร สุขโสม

โครงการพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
วิทยาศาสตรบัณฑิต  
สาขาวิชาวิทยาการคอมพิวเตอร์  
คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์  
ปีการศึกษา 2567  
ลิขสิทธิ์ของมหาวิทยาลัยธรรมศาสตร์

การวิเคราะห์ความคิดเห็นของภาพยินต์จากแพลตฟอร์มออนไลน์

โดย

นางสาวจันทร์ตราพร สุขโสม

โครงการพิเศษนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

วิทยาศาสตรบัณฑิต

สาขาวิชาวิทยาการคอมพิวเตอร์

คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

ปีการศึกษา 2567

ลิขสิทธิ์ของมหาวิทยาลัยธรรมศาสตร์

**Movie Reviews Sentiment Analysis from Online Platforms**

**BY**

**Miss Chantraphon Suksom**

**A FINAL-YEAR PROJECT REPORT SUBMITTED IN PARTIAL FULFILLMENT  
OF THE REQUIREMENTS FOR THE DEGREE OF BACHELOR OF SCIENCE**

**COMPUTER SCIENCE**

**FACULTY OF SCIENCE AND TECHNOLOGY**

**THAMMASAT UNIVERSITY**

**ACADEMIC YEAR 2024**

**COPYRIGHT OF THAMMASAT UNIVERSITY**

มหาวิทยาลัยธรรมศาสตร์  
คณะวิทยาศาสตร์และเทคโนโลยี  
รายงานโครงการนิเทศ

ของ  
นางสาวจันทร์ตราพร สุ่ม  
เรื่อง

การวิเคราะห์ความคิดเห็นของภาคผนวกจากแพลตฟอร์มออนไลน์

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์  
เมื่อ วันที่ 30 พฤษภาคม พ.ศ. 2568

อาจารย์ที่ปรึกษา

(ผศ. ดร. วิรัชัน Jarvisiri)

กรรมการสอบโครงการนิเทศ

(ผศ. ดร. วนิดา พุทธิวิจaya)

กรรมการสอบโครงการนิเทศ

(รศ. ดร. นัญชรัตน์ วงศ์วิธีชัย)

มหาวิทยาลัยธรรมศาสตร์  
คณะวิทยาศาสตร์และเทคโนโลยี  
รายงานโครงการพิเศษ

ของ

นางสาว จันทร์ตราพร สุขโสม  
เรื่อง

การวิเคราะห์ความคิดเห็นของภาคผนวกจากแพลตฟอร์มออนไลน์

ได้รับการตรวจสอบและอนุมัติ ให้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร  
หลักสูตรวิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์  
เมื่อ วันที่ 30 พฤษภาคม พ.ศ. 2568

อาจารย์ที่ปรึกษา

(ผศ. ดร. วิรัตน์ จาเรืองศักดิ์พูลย์)

กรรมการสอบโครงการพิเศษ

(ผศ. ดร. วนิดา พุทธิวิทยา)

กรรมการสอบโครงการพิเศษ

(รศ. ดร. ณัชรัตน์ วงศ์วิธีชัยร)

หัวข้อโครงการพิเศษ	การวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์มออนไลน์
ชื่อผู้เขียน	นางสาวจันทร์ตราพร สุขโสม
ชื่อปริญญา	วิทยาศาสตรบัณฑิต สาขาวิชาวิทยาการคอมพิวเตอร์
สาขาวิชา/คณะ/มหาวิทยาลัย	สาขาวิชาวิทยาการคอมพิวเตอร์ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์
อาจารย์ที่ปรึกษาโครงการพิเศษ	ผศ. ดร. วิรัตน์ จาเรืองศ์เพบูลย์
อาจารย์ที่ปรึกษาโครงการพิเศษร่วม	
ปีการศึกษา	2567

## บทคัดย่อ

งานวิจัยนี้นำเสนอการวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์มออนไลน์ถูกสร้างขึ้นมาเพื่อช่วยในการวิเคราะห์ความรู้สึกของความคิดเห็นที่มีต่อภาพยนตร์ ให้สามารถมองเห็นภาพรวมของความคิดเห็นที่มีต่อภาพยนตร์ได้ งานวิจัยนี้มุ่งเน้นในการพัฒนาระบบวิเคราะห์รู้สึก (Sentiment Analysis) โดยใช้ความคิดเห็นจาก YouTube และ X เพื่อเป็นข้อมูลในการศึกษา โดยสามารถแยกแยะความคิดเห็นออกเป็น ความคิดเก็บเชิงบวก และความคิดเห็นเชิงลบ ด้วยการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing) และเทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ในการสร้างแบบจำลอง ผลลัพธ์ที่ได้จากการวิเคราะห์ความรู้สึกของความคิดเห็นจากแพลตฟอร์มออนไลน์ที่ผ่านการทดสอบโดยโลจิสติกส์ เปรียบเทียบกับชัพพอร์ตแมชชีน พบร่วมโมเดลที่ใช้เทคนิคการถดถอยโลจิสติกส์มีค่าความถูกต้องสูงกว่าโดยมีค่าความถูกต้องเท่ากับ 89.63 เปอร์เซ็นต์

**คำสำคัญ:** การวิเคราะห์ความรู้สึก, เทคนิคการประมวลผลภาษาธรรมชาติ, เทคนิคการเรียนรู้ของเครื่อง

Thesis Title	Movie Reviews Sentiment Analysis from Online Platforms
Author	Miss Chantraphon Suksom
Degree	Bachelor of Science
Major Field/Faculty/University	Computer Science Faculty of Science and Technology Thammasat University
Project Advisor	Asst. Prof. Ph.D. Wirat Jareevongpiboon
Academic Years	2024

## ABSTRACT

This research presents a sentiment analysis system for movie reviews collected from online platforms. The system is designed to facilitate the interpretation of public opinion regarding films by providing an overview of audience sentiment. The study focuses on developing a sentiment analysis model using user comments from YouTube and X (formerly Twitter) as primary data sources. The comments are categorized into positive and negative sentiments through the application of Natural Language Processing (NLP) techniques and Machine Learning (ML) algorithms. The sentiment classification model was trained and evaluated using two machine learning methods: Logistic Regression and Support Vector Machine (SVM). The analysis results demonstrate that the Logistic Regression model achieved the highest accuracy, with a classification accuracy of 89.63%, thereby outperforming the SVM model.

**Keywords:** Sentiment Analysis, Natural Language Processing, Machine Learning Technique

### กิตติกรรมประกาศ

โครงการวิเคราะห์ความคิดเห็นของภาพนิตร์จากแพลตฟอร์มออนไลน์นี้จะไม่สามารถสำเร็จลุล่วงไปได้ด้วยดีหากไม่ได้รับความอนุเคราะห์และสนับสนุนจาก ผศ. ดร. วิรัตน์ จาเรืองศ์เพบูลย์ อาจารย์ที่ปรึกษาโครงงานที่ให้เกียรติสละเวลาอันมีค่าในการให้คำปรึกษา ให้ข้อเสนอแนะ และคำแนะนำถึงแนวทางในการแก้ไขปัญหาการจัดทำโครงงาน

ขอขอบพระคุณคณะกรรมการสอบโครงงานพิเศษทุกท่านที่สละเวลาในการให้ความรู้และให้คำแนะนำที่เป็นประโยชน์

ขอขอบคุณข้อมูลความรู้จากเว็บไซต์ซึ่งเป็นแหล่งข้อมูลในการศึกษาค้นคว้าในการจัดทำโครงงาน

ขอบคุณเพื่อน ๆ ที่ให้ความช่วยเหลือและเป็นกำลังใจในการทำโครงงาน

สุดท้ายขอขอบพระคุณคุณบิดา มารดา และน้องชาย ที่ให้การสนับสนุน และให้กำลังใจในการทำโครงงานนี้เสมอมา

นางสาวจันทร์ตราพร สุขโสม

## สารบัญ

	หน้า
บทคัดย่อ	1
ABSTRACT	2
กิตติกรรมประกาศ	3
สารบัญ	4
สารบัญตาราง	7
สารบัญภาพ	8
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของโครงการ	1
1.2 วัตถุประสงค์	2
1.3 ขอบเขตของโครงการ	2
1.4 ประโยชน์ของโครงการ	3
1.5 ข้อจำกัดของโครงการ	3
บทที่ 2 วรรณกรรมและงานวิจัยที่เกี่ยวข้อง	4
2.1 แนวคิดทฤษฎีที่เกี่ยวข้อง	4
2.1.1 การประมวลผลด้วยภาษาธรรมชาติ	4
2.1.2 การวิเคราะห์ความรู้สึก	5
2.1.3 การเตรียมข้อมูลสำหรับการประมวลผลภาษาธรรมชาติ	6
2.1.4 TF-IDF	7

2.1.5 อัลกอริทึมในการแยกประเภท	8
2.1.6 เมทริกซ์คอนฟิวชัน	13
2.1.7 การวัดประสิทธิภาพของแบบจำลอง	14
2.2 งานวิจัยที่เกี่ยวข้อง	15
2.2.1 ระบบการวิเคราะห์ความรู้สึกจากวิดีโอบันโชเชียลมีเดียด้วย ชั้พพร์ตเวกเตอร์แมชชีน	15
2.2.2 การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ ขั้นตอนวิธีชั้พพร์ตเวกเตอร์แมชชีน	16
2.2.3 การระบุรีวิวที่น่าสงสัยในผลิตภัณฑ์ดูแลผิวน้ำของลูกค้าบันดาดซื้อขาย สินค้าออนไลน์ ด้วยเทคนิคบริการเรียนรู้ของเครื่อง	16
2.2.4 ระบบวิเคราะห์ความคิดเห็นต่อละครไทยบนทวิตเตอร์	17
2.2.5 การเปรียบเทียบคุณสมบัติของงานวิจัยที่เกี่ยวข้อง	18
บทที่ 3 วิธีการวิจัย	19
3.1 ภาพรวมของโครงงาน	19
3.2 การวิเคราะห์ขอบเขตและความต้องการของระบบ	21
3.2.1 การรวบรวมข้อมูล (Data Collection)	21
3.2.2 การเตรียมข้อมูล (Data Preprocessing)	21
3.2.3 การฝึกสอนและการทดสอบแบบจำลอง (Training and Testing Model)	24
3.2.4 การประเมินผล (Evaluation)	25
3.3 ประเด็นที่นำเสนอและสิ่งที่ท้าทาย	26
3.4 ผลลัพธ์ที่คาดหวัง	26
บทที่ 4 ทรัพยากรและแผนการดำเนินงาน	28
4.1 การจัดเตรียมฮาร์ดแวร์และซอฟต์แวร์	28
4.1.1 ฮาร์ดแวร์ที่ใช้ในการพัฒนา	28

4.1.1 ซอฟต์แวร์ที่ใช้ในการพัฒนา	28
4.2 แผนการดำเนินงาน	28
4.3 ผลการดำเนินงาน	30
4.3.1 การวิเคราะห์ชุดข้อมูล	30
4.3.2 การเบรียบเทียบประสิทธิภาพ	32
4.3.1 การทดลองใช้งาน	33
บทที่ 5 สรุป	35
5.1 สรุปผลการดำเนินงาน	43
5.2 ข้อเสนอแนะและแนวทางในการพัฒนา	45
รายการอ้างอิง	44

## สารบัญตาราง

หน้า

ตารางที่ 2.2.5 ตารางเปรียบเทียบตุณสมบัติของงานวิจัยที่เกี่ยวข้อง	18
ตารางที่ 3.2.1 ตารางแสดงจำนวนข้อมูลแยกตามประเภทความรู้สึก	21
ตารางที่ 4.2.1 การดำเนินงานที่ผ่านมาจนถึงปัจจุบัน	31
ตารางที่ 4.3.1 ประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยการทดลองโดยโลจิสติกส์ หลังการปรับพารามิเตอร์	41
ตารางที่ 4.3.2 ประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยชั้พพอร์ตเวกเตอร์แมชชีน หลังการปรับพารามิเตอร์	44
ตารางที่ 4.3.3 ตารางเปรียบเทียบประสิทธิภาพของอัลกอริทึมการทดลองโดยโลจิสติกส์และชัพพอร์ต เวกเตอร์แมชชีน	47

## สารบัญภาพ

	หน้า
ภาพที่ 2.1.1 แสดงตัวอย่างการจำแนกข้อมูลด้วย Support Vector Machine	11
ภาพที่ 2.1.2 แสดงตัวอย่างการทำ Support Vector Machine ด้วย Kernel Function	12
ภาพที่ 2.1.3 แสดงตัวอย่างตาราง Confusion Matrix	13
ภาพที่ 3.1.1 ตัวอย่างสถาปัตยกรรมภาพรวมของการวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์ม ออนไลน์	19
ภาพที่ 3.2.1 ตัวอย่างการวิเคราะห์ขอบเขตและความต้องการของการวิเคราะห์	21
ภาพที่ 3.2.2 รูปแสดงตัวอย่างข้อความก่อนทำการทำความสะอาดข้อความ	22
ภาพที่ 3.2.3 รูปแสดงตัวอย่างข้อความหลังทำการทำความสะอาดข้อความ	22
ภาพที่ 3.2.4 รูปแสดงตัวอย่างข้อมูลคำที่เป็น Stop Word	23
ภาพที่ 3.2.5 รูปแสดงตัวอย่างข้อมูลหลังจากตัด Stop Word	23
ภาพที่ 3.2.6 รูปแสดงตัวอย่างข้อมูลก่อนการทำ Lemmatization	23
ภาพที่ 3.2.7 รูปแสดงตัวอย่างข้อมูลที่ผ่านการทำ Lemmatization	23
ภาพที่ 3.2.8 รูปแสดงตัวอย่างข้อมูลหลังการตัดคำ	24
ภาพที่ 3.2.9 รูปแสดงตัวอย่างโค้ดการใช้งานโมดูล TfidfVectorizer	25
ภาพที่ 3.2.10 ภาพที่ 3.2.10 รูปแสดงตัวอย่างโค้ดการใช้งานโมดูล TfidfVectorizer และค่าพารามิเตอร์ที่ใช้ในการทดลอง	25
ภาพที่ 3.2.11 รูปแสดงตัวอย่าง Unigram และ Bigram ที่พบบ่อย	26
ภาพที่ 3.2.12 รูปแสดงตัวอย่าง Confusion Matrix	28
ภาพที่ 4.3.1 กราฟเปรียบเทียบความยาวความของคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ	33
ภาพที่ 4.3.2 กราฟเปรียบเทียบความยาวความของคิดเห็นเชิงบวกและความคิดเห็นเชิงลบหลังจากการทำความสะอาดข้อมูล	34
ภาพที่ 4.3.3 ค่อนพิวชันเมทริกซ์ของการถดถอยโลจิสติกส์กับ TfidfVectorizer ที่ยังไม่ได้ทำการปรับพารามิเตอร์	35
ภาพที่ 4.3.4 ค่อนพิวชันเมทริกซ์ของการถดถอยโลจิสติกส์กับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้ว	36
ภาพที่ 4.3.5 ค่อนพิวชันเมทริกซ์ของชั้พพอร์ตเวอร์แมชีนกับ TfidfVectorizer ที่ยังไม่ได้ทำการปรับพารามิเตอร์	37

ภาพที่ 4.3.6 ค่อนพิวชันแมทริกซ์ของชั้พพร์ตเวกเตอร์เมชีนกับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้ว	38
ภาพที่ 4.3.7 ค่อนพิวชันแมทริกซ์ของ LinearSVC กับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้ว	39
ภาพที่ 4.3.8 ค่อนพิวชันแมทริกซ์ของการทดลองโดยโลจิสติกส์หลังจากการทำ GridSearchCV	40
ภาพที่ 4.3.9 ค่อนพิวชันแมทริกซ์ของชั้พพร์ตเวกเตอร์เมชีน	
การทำ RandomSearchCV	42
ภาพที่ 4.3.10 ค่อนพิวชันแมทริกซ์ของชั้พพร์ตเวกเตอร์เมชีนหลังจากการทำ GridSearchCV	43
ภาพที่ 4.3.11 ค่อนพิวชันแมทริกซ์ผลการจำแนกความรู้สึกของข้อความจาก YouTube ด้วยการทดลองโดยโลจิสติกส์หลัง	45
ภาพที่ 4.3.12 ค่อนพิวชันแมทริกซ์ผลการจำแนกความรู้สึกของข้อความจาก YouTube ด้วยชั้พพร์ตเวกเตอร์เมชีน	46
ภาพที่ 4.3.13 ภาพแสดงหน้าหลักของระบบ	48
ภาพที่ 4.3.14 ภาพแสดงหน้าหลักของระบบสำหรับการเลือกแพลตฟอร์มในการวิเคราะห์	48
ภาพที่ 4.3.15 ภาพตัวอย่าง Pie Chart จากการวิเคราะห์ความคิดเห็น	49
ภาพที่ 4.3.16 ตัวอย่างความคิดเห็นจากยูทูปที่ไม่เดลทำนายป้ายกำกับ	49
ภาพที่ 4.3.17 ตัวอย่างการทำ Word Could จากความคิดเห็นโดยแยกเป็นความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ	50
ภาพที่ 5.1.1 ตัวอย่างข้อความเชิงลบที่แบบจำลองทั้งสองทำนายผิดพลาดก่อนทำความสะอาด ข้อความ	52
ภาพที่ 5.1.2 ตัวอย่างข้อความเชิงลบที่แบบจำลองทั้งสองทำนายผิดพลาดหลังทำความสะอาด ข้อความ	52
ภาพที่ 5.1.3 ตัวอย่างข้อความเชิงบวกที่แบบจำลองทั้งสองทำนายผิดพลาดก่อนทำความสะอาด ข้อความ	52
ภาพที่ 5.1.4 ตัวอย่างข้อความเชิงบวกที่แบบจำลองทั้งสองทำนายผิดพลาดหลังทำความสะอาด ข้อความ	53
ภาพที่ 5.1.5 ตัวอย่างความคิดเห็นต่อภาพนัตร์ Sinners บนยูทูป	53
ภาพที่ 5.1.6 ตัวอย่างความคิดเห็นที่แบบจำลองการทดลองโดยโลจิสติกส์ทำนายผิดพลาด	54
ภาพที่ 5.1.7 ตัวอย่างความคิดเห็นที่แบบจำลองชั้พพร์ตเวกเตอร์เมชีนทำนายผิดพลาด	55

## บทที่ 1

### บทนำ

#### 1.1 ความเป็นมาและความสำคัญของภารกิจ

ปัจจุบันโซเชียลมีเดียกลายเป็นส่วนสำคัญต่อการใช้ชีวิตประจำวันของผู้คนในทุกมิติ ไม่ว่าจะเป็นการสื่อสาร การติดตามข่าวสาร การพักผ่อนหย่อนใจ รวมถึงการแสดงออกถึงความคิดเห็นที่มีต่อสิ่งต่างๆ ที่เกิดขึ้นรอบตัว การแสดงความคิดเห็นผ่านโลกออนไลน์จึงกลายเป็นสิ่งที่มีอิทธิพลต่อสังคมเป็นอย่างมาก การรีวิวและการแสดงความคิดเห็นกลับเป็นเสียงสะท้อนนม涓 ความคิด หรือความรู้สึกของคนในสังคม ภาพนิทรรศ์เป็นหนึ่งในสื่อบันเทิงที่ได้รับความสนใจและมักได้รับการวิพากษ์วิจารณ์ผ่านแพลตฟอร์มออนไลน์เป็นจำนวนมาก โดยเฉพาะในยุคที่สังคมออนไลน์เข้ามามีบทบาทสำคัญความคิดเห็นจากผู้ชมภาพนิทรรศ์จะเปรียบเสมือนเสียงสะท้อนที่ไม่อาจมองข้าม ความคิดเห็นเชิงบวกสามารถสร้างกระแสความนิยมและผลักดันให้ภาพนิทรรศ์ได้รับความสนใจมากขึ้น แต่ในขณะเดียวกันความคิดเห็นเชิงลบอาจส่งผลกระทบต่อความเชื่อมั่นของผู้ชมรายใหม่ที่กำลังพิจารณาเลือกชมภาพนิทรรศ์ แต่ด้วยปริมาณข้อมูลความคิดเห็นที่เพิ่มขึ้นเรื่อยๆ การวิเคราะห์ความคิดเห็นด้วยวิธีแบบเดิมๆ จึงกลายเป็นความท้าทายและสิ่นเปลี่ยนเวลา

การแสดงความคิดเห็นที่มีจำนวนมากจากผู้ชมในแพลตฟอร์มต่างๆ ส่งผลให้เกิดความยากลำบากในการทำความเข้าใจความรู้สึกโดยรวมของผู้ชมที่มีต่อภาพนิทรรศ์แต่ละเรื่อง ด้วยปริมาณข้อมูลที่เพิ่มขึ้นอย่างรวดเร็วทำให้การวิเคราะห์ความคิดเห็นแบบเดิม เช่น การอ่านหรือวิเคราะห์ด้วยมนุษย์นั้นนิยมมีข้อจำกัดทั้งในด้านของเวลาและความแม่นยำ นอกจากนี้ความคิดเห็นของผู้ชมที่ทำการโพสต์มักมีลักษณะที่ซับซ้อนหรือหลากหลายความรู้สึกแฝงอยู่ในการใช้ถ้อยคำและบริบทที่แตกต่างกัน

การวิเคราะห์ความคิดเห็นที่มีปริมาณมากด้วยเทคโนโลยีการประมวลผลภาษาธรรมชาติ (Natural Language Processing) และการเรียนรู้ของเครื่อง (Machine Learning) มีความสำคัญอย่างยิ่งในการแก้ปัญหาความยุ่งยากในการประมวลผลข้อมูลจำนวนมากและซับซ้อน ซึ่งความคิดเห็นของผู้ชมที่มีต่อภาพนิทรรศ์สามารถระบุสภาพความทึบตื้นและความรู้สึกที่มีอิทธิพลต่อความนิยมของภาพนิทรรศ์ได้ การใช้เทคโนโลยีเข้ามาช่วยในการวิเคราะห์ความคิดเห็นจะช่วยให้ผู้ผลิตสามารถรับรู้ถึงความคิดเห็นของผู้ชมได้อย่างรวดเร็วและแม่นยำ ซึ่งไม่เพียงแต่ช่วยให้เข้าใจความนิยมของลูกค้าที่แท้จริงแล้ว ยังช่วยให้ผู้ผลิตสามารถปรับปรุงและพัฒนาสินค้าและบริการให้มีประสิทธิภาพมากยิ่งขึ้น

ประสิทธิภาพหมายอีกนัยน์ จึงอีกด้วย และในส่วนของผู้บริโภคภาพนั้นการเข้าถึงความคิดเห็นที่ได้รับการวิเคราะห์ จัดระเบียบอย่างละเอียดและเข้าถึงง่ายจะช่วยให้ผู้บริโภคสามารถมองเห็นภาพรวมของมุ่งมองที่มีต่อภาพนั้นได้ เพื่อเป็นตัวช่วยในการตัดสินใจเลือกซื้อภาพนั้นที่สอดคล้องกับความชอบและความคาดหวังของตนเองได้ดียิ่งขึ้น

จากปัญหาข้างต้นผู้จัดทำจึงมีแนวคิดที่จะพัฒนาการวิเคราะห์ความคิดเห็นของภาพนั้นผ่านแพลตฟอร์มออนไลน์ที่สามารถประมวลผลความคิดเห็นของผู้ชมที่ได้ โดยระบบนี้จะนำเทคโนโลยีการประมวลผลภาษาธรรมชาติ (Natural Language Processing) และการเรียนรู้ของเครื่อง (Machine Learning) มาช่วยในการจำแนกความคิดเห็นเป็นเชิงบางหรือเชิงลบ โดยจะแสดงประเด็นความคิดเห็นของภาพนั้นให้อยู่ในรูป Word Cloud เพื่อให้ผู้ใช้งานสามารถเห็นภาพรวมของความรู้สึกและมุ่งมองในแต่ละด้านที่มีต่อภาพนั้น เพื่อช่วยลดระยะเวลาในการประมวลผลความคิดเห็นจำนวนมากด้วยมุ่งเน้นและช่วยให้ผู้ผลิตภาพนั้นสามารถใช้ข้อมูลที่ได้จากการวิเคราะห์ไปปรับปรุงคุณภาพให้ตอบโจทย์ความต้องการของผู้ชมได้อย่างตรงจุด

## 1.2 วัตถุประสงค์

โครงการนี้มีเป้าหมายเพื่อวิเคราะห์ความรู้สึกและจำแนกประเภทของความคิดเห็นบนแพลตฟอร์มออนไลน์ เพื่อให้บรรลุ เป้าหมายดังกล่าว จึงกำหนดวัตถุประสงค์ของโครงการดังต่อไปนี้

- เพื่อให้สามารถวิเคราะห์ความรู้สึกของความคิดเห็นที่มีต่อภาพนั้นบนแพลตฟอร์มออนไลน์ได้ทันต่อเวลา
- เพื่อให้ผู้สร้างสรรค์ผลงานได้ข้อมูลที่เป็นประโยชน์ในการนำข้อมูลที่ได้จากการวิเคราะห์ไปปรับปรุงผลงาน
- เพื่อให้ผู้บริโภคภาพนั้นสามารถเข้าถึงความคิดเห็นและมุ่งมองของผู้ชมท่า�อื่น ๆ ประกอบการตัดสินใจเลือกซื้อภาพนั้นได้

## 1.3 ขอบเขตของโครงการ

เพื่อให้เป็นไปตามวัตถุประสงค์ของโครงการ จึงได้กำหนดขอบเขตของการพัฒนาระบบทั้งนี้

- จัดเก็บและประมวลผลข้อมูลเฉพาะความคิดเห็นที่เป็นภาษาอังกฤษ

2. ข้อมูลสำหรับการฝึกสอนเป็นข้อความแสดงความคิดเห็นต่อภาพณัตร์บน [www.imdb.com](http://www.imdb.com) จำนวน 50,000 ข้อความ
3. การรวบรวมข้อมูลที่ใช้ในการวิเคราะห์จะเป็นการดึงความคิดเห็นจากไฟล์วิดีโອน YouTube และความคิดเห็นจาก X ผ่าน API
4. การวิเคราะห์ความคิดเห็นจะให้ผลลัพธ์เป็น ความรู้สึกเชิงบวก(Positive) และความรู้สึกเชิงลบ (Negative) เท่านั้น
5. อัลกอริทึมทางด้านการเรียนรู้ของเครื่องที่ใช้ในการเปรียบเทียบความสามารถในการจำแนกหมวดหมู่ ได้แก่ Logistics Regression และ Support Vector Machine

#### 1.4 ประโยชน์ของโครงการ

1. ผู้ชุมภาพญตร์ได้ข้อมูลในการตัดสินใจเลือกชมภาพณัตร์จากข้อมูลสรุปความชื่นชอบ หรือคำติชมจากผู้ชมท่านอื่น
2. ผู้สร้างภาพณัตร์ได้รับคำติชมในประเด็นต่างๆ ได้จาก Word Cloud และสามารถนำไปประยุกต์ในการพัฒนาสร้างภาพณัตร์ให้ตรงกับความต้องการของผู้ชมมากขึ้น

#### 1.5 ข้อจำกัดของโครงการ

1. แบบจำลองวิเคราะห์ความรู้สึกสามารถจำแนกความรู้สึกออกเป็นความรู้สึกเชิงบวกและความรู้สึกเชิงลบเท่านั้น
2. การดึงข้อมูลความคิดเห็นเป็นแบบ Realtime แต่ด้วยข้อจำกัดการดึงข้อมูลจาก platform ทำให้ระบบถูกจำกัดจำนวนครั้งของการทดสอบต่อวัน โดย
  - 2.1 การดึงข้อมูลจากยูทูบ (YouTube) สามารถดึงได้เพียง 500 โพสต์ต่อเดือน เท่านั้น และ
  - 2.2 การดึงข้อมูลจาก X สามารถดึงข้อมูลได้เพียง 100 ครั้งต่อเดือนเท่านั้น

## บทที่ 2

### วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

#### 2.1 แนวคิดทฤษฎีที่เกี่ยวข้อง

##### 2.1.1 การประมวลผลด้วยภาษาธรรมชาติ

การประมวลผลด้วยภาษาธรรมชาติ หรือ Natural Language Processing (NLP) คือวิทยาการแขนงหนึ่งในหมวดหมู่ของเทคโนโลยีปัญญาประดิษฐ์ หรือ Artificial Intelligence (AI) ซึ่งถูกพัฒนามาเพื่อการปฏิสัมพันธ์ระหว่างคอมพิวเตอร์และมนุษย์ผ่านภาษาธรรมชาติ โดยเกี่ยวข้อง กับการพัฒนาอัลกอริทึมและเทคนิคที่ช่วยให้คอมพิวเตอร์สามารถเข้าใจ ตีความ ตลอดจนสามารถใช้งานภาษาธรรมชาติได้ เทคโนโลยี Natural Language Processing (NLP) ครอบคลุมการทำงานที่หลากหลายทั้งการแปลภาษา การวิเคราะห์ความรู้สึก การสรุปข้อความ และอื่นๆ โดยมีเป้าหมายเพื่อ ช่วยลดช่องว่างการสื่อสารระหว่างคอมพิวเตอร์กับมนุษย์ ให้คอมพิวเตอร์สามารถประมวลผลและ วิเคราะห์ข้อมูลภาษาธรรมชาติจำนวนมหาศาลได้อย่างมีประสิทธิภาพ และแม่นยำ

เทคโนโลยี Natural Language Processing (NLP) มีรากฐานมาจากหลายสาขาวิชา ด้วยกัน โดยมีพื้นฐานเกิดจากการทำงานระหว่างภาษาศาสตร์ทางคอมพิวเตอร์ หรือ Computational Linguistics (การใช้คอมพิวเตอร์วิเคราะห์ข้อมูลทางภาษา) และในส่วนของอัลกอริทึมการเรียนรู้ของ เครื่อง (Machine learning) กับการเรียนรู้เชิงลึก (Deep learning) เป็นต้น โดยหลัก Computational Linguistics ที่ NLP ใช้ในการวิเคราะห์แบ่งออกเป็น 2 ประเภท ได้แก่ การวิเคราะห์ไวยากรณ์ (Syntactical Analysis) คือ กระบวนการที่ใช้ในการกำหนดความหมายของคำ วลี หรือ ประโยค และอีกประเภทคือ การวิเคราะห์เชิงความหมาย (Semantical Analysis) เป็นวิธีการที่ใช้ ผลลัพธ์ที่ได้มาจากการ Syntactical Analysis มาแทนนิยาม และตีความหมายของคำตามรูปแบบ โครงสร้างประโยค

โดยเทคโนโลยี NLP นั้นสามารถนำไปใช้ประโยชน์ได้หลากหลายด้าน ตัวอย่างเช่น ใช้ประโยชน์ให้คอมพิวเตอร์หรือเครื่องจักรสามารถเข้าใจภาษาของมนุษย์ได้ทั้งภาษาพูดและข้อความ ช่วยสกัดข้อมูลที่มีความหมายอ้อมจากข้อความ ช่วยในการจัดการข้อมูลที่เป็นข้อความหรือเอกสาร ให้สามารถค้นหาข้อมูลและสกัดข้อมูลได้อย่างมีประสิทธิภาพ และสามารถช่วยในการวิเคราะห์ ความรู้สึกและความคิดเห็นได้อีกด้วย ซึ่งในปัจจุบันเทคโนโลยี NLP ถูกนำมาประยุกต์ใช้และมีบทบาท

สำคัญในชีวิตประจำวันมากขึ้นเรื่อย ๆ ตัวอย่างเช่น Siri หรือ Google Assistant ที่ถูกนำมาใช้เป็นผู้ช่วยส่วนตัวในการค้นหาข้อมูล ตั้งเตือน และใช้ควบคุมอุปกรณ์ไฮเทคผ่านเสียงพูดหรือข้อความ Grammarly หรือ Microsoft Word ที่ใช้เทคโนโลยีในการตรวจสอบคำผิดหรือปรับปรุงประโยคให้ถูกต้อง ด้วยเหตุนี้ NLP จึงกลายเป็นเครื่องมือที่จำเป็นสำหรับผู้ที่ต้องการทำงานหรือจัดการกับข้อมูลด้านภาษา ไม่ว่าจะเป็นการแปลภาษา สร้างคำอธิบาย สร้างโมเดลภาษา หรือช่วยในการตัดสินใจโดยอ้างอิงข้อมูลจากข้อความ

### 2.1.2 การวิเคราะห์ความรู้สึก

การวิเคราะห์ความรู้สึก หรือ Sentiment Analysis คือ กระบวนการที่ใช้เทคโนโลยีคอมพิวเตอร์ในการวิเคราะห์และประมวลผล ตรวจสอบและระบุความรู้สึกหรืออารมณ์ที่แสดงออกมาจากข้อความหรือเสียง โดยจะใช้เทคนิคสำหรับการจำแนกข้อความออกเป็นกลุ่มด้วยวิธีการทางปัญญาประดิษฐ์ (AI) และ กระบวนการ NLP โดยการจำแนกความรู้สึกของเนื้อหาว่ามีแนวโน้มเป็นไปในเชิงบวก (Positive), เชิงลบ (Negative) หรือเป็นกลาง (Neutral) ซึ่งการจำแนกขึ้นอยู่กับข้อความหรือคำพูดที่ถูกเลือกใช้ในการอธิบายหัวข้อ สินค้าหรือบริการนั้นๆ

ในปัจจุบันเป็นยุคที่โซเชียลมีเดียเป็นหนึ่งในปัจจัยหลักของชีวิตประจำวันของมนุษย์ การทำ Sentiment Analysis จึงสำคัญอย่างยิ่งในการนำมารวิเคราะห์ลูกค้าในตลาดออนไลน์ เนื่องจากปัจจุบันโซเชียลมีเดียมีการเปิดกว้างให้คนสามารถแสดงความคิดเห็นได้อย่างเสรีมากขึ้น คนส่วนใหญ่จึงมักรีวิวหรือแสดงความคิดเห็นผ่านโซเชียลมีเดีย ดังนั้นโซเชียลมีเดียจึงกลายเป็นแหล่งข้อมูลขนาดใหญ่ที่สามารถใช้ในการวิเคราะห์และรับฟังความคิดเห็นของผู้บริโภค การใช้เทคโนโลยี Sentiment Analysis ใน การวิเคราะห์ความคิดเห็น จึงเป็นตัวช่วยที่มีประโยชน์อย่างมากในการทำธุรกิจ โดยในทางธุรกิjm ก าม น า Sentiment Analysis มาใช้ในการทำธุรกิจ ตัวอย่างเช่น ใช้ในการพัฒนาสินค้าให้ตรงกับความต้องการในตลาด การทำ Market Research หรือ ใช้เพื่อดูความคิดเห็นหรือรีวิวที่ลูกค้าพูดถึงสินค้าและบริการทางโซเชียลมีเดียว่าลูกค้ามีความคิดเห็นต่อแบรนด์ไปทิศทางไหน

### 2.1.3 การเตรียมข้อมูลสำหรับการประมวลผลภาษาธรรมชาติ

ในการสร้างแบบจำลองของข้อมูลนั้นจำเป็นต้องมีการเตรียมข้อมูลให้พร้อมก่อนที่จะนำไปใช้งาน ซึ่งวิธีการในการเตรียมข้อมูลนั้นมีความแตกต่างกันไปขึ้นอยู่กับโครงสร้างภาษาที่นำมาใช้ การวิเคราะห์ โดยมีขั้นตอนในการจัดเตรียมข้อมูล ดังนี้

#### 2.1.3.1 การทำความสะอาดข้อความ

การทำความสะอาดข้อความ หรือ Text Cleaning เช่น การปรับตัวอักษรให้เป็นตัวอักษรพิมพ์เล็ก การลบเครื่องหมายวรรคตอน (Punctuation) การลบเครื่องหมายหรืออักษรพิเศษ หรือ การลบข้อความที่มี HTML Tag เพื่อให้มีเนื้อนำข้อมูลไปทำการวิเคราะห์สามารถวิเคราะห์ข้อมูลได้ง่ายและมีประสิทธิภาพมากขึ้น

#### 2.1.3.2 การตัดคำ

การตัดคำ หรือ Word Tokenization เป็นกระบวนการที่มักใช้ในการตัดข้อความตัวหนังสือที่เป็นรูปประโยคให้อยู่ในหน่วยที่เล็กลง ซึ่งอาจอยู่ในรูปของคำ หน่วยย่อของคำ หรือตัวอักษรก็ได้ โดยการตัดคำนั้นมีความสำคัญอย่างยิ่งในเทคโนโลยี NLP คือสามารถใช้เพื่อค้นหาข้อความหรือคีย์เวิร์ดจากประโยคได้ในภาษาอังกฤษการตัดคำสามารถทำได้ง่ายเมื่อเทียบกับภาษาไทยโดยใช้ช่องว่างเป็นตัวช่วยในการตัดแบ่งคำ

#### 2.1.3.3 การลบคำที่เป็นคำพื้นฐาน

การลบคำที่เป็นคำพื้นฐาน หรือ Stop Word Removal คือ การลบคำที่มักพบบ่อยในประโยคหรือข้อความ แต่ไม่ได้ใช้ในการสื่อสารความหมาย เช่น a, an, the, also, just, quite, unless

#### 2.1.3.4 การStemming และ Lemmatization

การ Stemming และ Lemmatization เนื่องจากในภาษาอังกฤษมีการผันไวยกรณ์ที่ทำให้คำที่มีความหมายเดียวกันอยู่ในรูปที่ไม่เหมือนกัน ตัวอย่างเช่น run, runs และ running แต่ละคำมีรากศัพท์มาจากคำว่า run เมื่ອอกัน แต่ถูกผันไปตามรูปแบบประโยค ดังนั้นจึงต้องทำการลดรูปคำเดียวกันให้อยู่ในรูปแบบพื้นฐานของคำนั้นๆ โดย Stemming จะเป็นการตัดส่วนท้ายของคำออกไปเพื่อให้เหลือรากศัพท์ของคำนั้นๆ เช่น run, runs, running ตัดให้อยู่ในรูปของ run ส่วน Lemmatization จะเป็นการแปลงคำค้ำค้ำต่างๆ ให้อยู่ในรูปแบบพื้นฐานของคำนั้นๆ เช่น is, am, are อยู่ในรูปของ be

#### 2.1.4 TF-IDF

TF-IDF หรือ Term Frequency – Inverse Document Frequency เป็นวิธีการถ่วงน้ำหนักที่นิยมใช้กันมากในการทำงานด้านการสืบค้นสารสนเทศ และ Text Mining เป็นวิธีการทางสถิติที่ใช้ในการประเมินความสำคัญต่อเอกสาร ความสำคัญของคำเป็นสัดส่วนโดยตรงกับจำนวนครั้งที่คำคำนั้นถูกพบในเอกสารแต่ถูกลดความสำคัญโดยความถี่ของคำนั้นในกลุ่มเอกสารทั้งหมด

##### 2.1.4.1 Term Frequency

TF ย่อมาจาก Term Frequency หรือ ความถี่ของคำศัพท์เป็นค่าสำหรับดูความถี่ว่าคำนั้นๆ ปรากฏบ่อยมากน้อยแค่ไหนในเอกสาร ซึ่งการหาค่า TF สามารถหาได้จากสูตร

$$TF_{(t,d)} = \frac{f_{(t,d)}}{\sum_{t' \in d} f_{(t',d)}}$$

ซึ่ง  $t$  คือ term หรือ คำ

$d$  คือ เอกสารนั้นๆ

กล่าวคือค่า TF สามารถหาได้จากการคำนวณคำที่ปรากฏในเอกสาร/จำนวนคำทั้งหมดในเอกสาร

##### 2.1.4.2 Inverse Document Frequency

IDF ย่อมาจาก Inverse Document Frequency หรือการผกผันในความถี่ของเอกสาร ซึ่งมีที่มาจากการที่มีค่า TF มาเกี้ยงไม่สามารถตอบได้ว่าเป็นคำสำคัญเนื่องจาก TF จะทำการคำนวณจากแต่ละเอกสาร แต่ IDF จะเป็นค่าที่คำนวณจากเอกสารทั้งหมด ซึ่งค่า IDF สามารถคำนวณได้จากสูตร

$$IDF_{(t)} = \log [ (1 + n) / (1 + DF_{(t)}) ] + 1$$

ซึ่ง  $t$  คือ term หรือ คำ 1 คำ

$n$  คือ จำนวนเอกสารทั้งหมดที่มีอยู่

$DF(t)$  คือ จำนวนเอกสารที่พบของคำ  $t$

ดังนั้นค่า TF-IDF สามารถหาได้จากการนำค่า TF และค่า IDF มาคูณกัน เพื่อให้น้ำหนักของ TF และน้ำหนักของ IDF ที่ซึ่งมีค่าต่างข้ามกันนั้นเป็นน้ำหนักที่สามารถแยกคำที่มีความสำคัญของมากได้ด้วยสูตร

$$TFIDF_{(t,d)} = TF_{(t,d)} * IDF_{(t)}$$

## 2.1.5 อัลกอริทึมในการแยกประเภท

### 2.1.5.1 การวิเคราะห์การถดถอยโลจิสติก

การวิเคราะห์การถดถอยโลจิสติก หรือ Logistic Regression Analysis เป็นเทคนิควิเคราะห์ตัวแปรเชิงพุทธิ์มีวัตถุประสงค์เพื่อประมาณค่าหรือทำนายเหตุการณ์ที่สนใจว่าเกิดหรือไม่เกิดเหตุการณ์นั้น โดยแบบจำลองโลจิสติกจะทำนายตัวแปรออกเป็นทวินาม (Dichotomous Variable) กล่าวคือมีได้สองค่า เช่น “เกิด” กับ “ไม่เกิด” หรือ “เสี่ยง” กับ “ไม่เสี่ยง” ส่วนค่าที่ใช้ทำนายอาจมีตัวเดียวหรือหลายตัวที่เป็นได้ทั้งตัวแปรเชิงกลุ่ม (Categorical Variable) หรือตัวแปรแบบต่อเนื่อง (Continuous Variable) การวิเคราะห์การถดถอยโลจิสติกที่เกี่ยวข้องกับทฤษฎีความน่าจะเป็นทวินามถูกเรียกว่า Binomial Logistic Regression ถ้าตัวแปรตามเป็นพหุนามจะเรียกว่า Multinomial Logistic Regression

โมเดลในการวิเคราะห์การถดถอยโลจิสติกแบ่งออกเป็น 3 ประเภท ดังนี้

- (1) การวิเคราะห์การถดถอยโลจิสติกแบบทวินาม หรือ Binary Logistic Regression Analysis เป็นการวิเคราะห์ความสัมพันธ์ของตัวแปรอิสระที่ได้ค่าตัวแปรตามออกมาเป็นสองค่า คือ 0 และ 1 ซึ่งจะเป็นการทำนายว่าเกิดหรือไม่เกิดเหตุการณ์ที่สนใจ เช่น ตัวแปรตามเป็น 0 คือผู้ป่วยไม่เป็นโรคเบาหวาน หรือ 1 คือผู้ป่วยเป็นโรคเบาหวาน เป็นต้น
- (2) การวิเคราะห์การถดถอยโลจิสติกแบบพหุนาม หรือ Multinomial Logistic Regression Analysis เป็นการวิเคราะห์ความสัมพันธ์ของตัวแปรอิสระที่ได้ค่าตัวแปรอกรามากกว่า 2 ค่าขึ้นไป ตัวอย่างเช่น การจำแนกกลุ่มออกเป็น สีขาว, สีดำ, สีแดง

(3) การวิเคราะห์การถดถอยโลจิสติกแบบเรียงลำดับ หรือ Multinomial Logistic Regression Analysis เป็นการวิเคราะห์ความสัมพันธ์ของตัวแปรอิสระที่ได้ค่าตัวแปรอีกมากกว่า 2 ค่าขึ้นไป ตัวอย่างเช่น การจำแนกกลุ่มออกเป็น สีขาว, สีดำ, สีแดง

การวิเคราะห์การถดถอยโลจิสติกเป็นการประมาณค่าของการเกิดเหตุการณ์ หากมีตัวแปรเพียงแค่ตัวเดียวฟังก์ชันโลจิสติกจะแสดงความน่าจะเป็นของเหตุการณ์ที่มีดังต่อไปนี้

$$Prob(event) = \frac{e^{(\beta_0 + \beta_1 x)}}{1 + e^{(\beta_0 + \beta_1 x)}}$$

หรือ

$$Prob(event) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

เมื่อ

$\beta_0$  คือ ค่าคงที่

$\beta_i$  คือ ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระตัวที่  $i$

$x$  คือ ตัวแปรอิสระ

$e$  คือ ค่าลอการิทึมธรรมชาติ (มีค่าประมาณ 2.71828...)

ในกรณีที่มีตัวแปรทำนายหลายตัวฟังก์ชันโลจิสติกจะเป็นดังนี้

$$Prob(event) = \frac{e^z}{1 + e^z}$$

หรือ

$$Prob(event) = \frac{1}{1 + e^{-z}}$$

เมื่อ  $z$  คือ Linear Combination ที่อยู่ในรูป

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

โดยค่าความน่าจะเป็นของเหตุการณ์จะมีค่าอยู่ระหว่าง 0 ถึง 1 เช่น ความน่าจะเป็นที่ผู้ป่วยป่วยเป็นโรคเบาหวานเท่ากับ 0.25 หมายถึง ผู้ป่วยมีโอกาสป่วยเป็นโรคเบาหวาน 25 เปอร์เซ็นต์

#### ข้อมูลที่เหมาะสมกับการวิเคราะห์การถดถอยโลจิสติก

- ข้อมูลที่ใช้สำหรับการแบ่งแยกหรือจำแนกประเภท โดยส่วนใหญ่จะเป็นข้อมูลที่แบ่งแบบทวินาม (Binomial) คือข้อมูลที่สามารถแบ่งออกเป็น 2 ประเภท เช่น ใช่/ไม่ใช่ และข้อมูลที่มีการแบ่งมากกว่า 2 ประเภท (Multinomial)

#### จุดเด่นของการวิเคราะห์การถดถอยโลจิสติก

- มีแนวโน้มที่จะเกิดการ Overfit น้อย แต่อาจเกิดขึ้นได้ในกรณีที่ High Dimensional Dataset
- สามารถใช้งานได้กับ Dataset ที่มีความสัมพันธ์แบบ Linear
- ทำงานได้กับข้อมูลที่มี Feature data เยอะแต่ตัวอย่างน้อย
- สามารถทำการเข้าใจการวิเคราะห์ได้ง่าย ไม่เปลืองทรัพยากรในการประมวลผล

#### จุดอ่อนของการวิเคราะห์การถดถอยโลจิสติก

- ไม่เหมาะสมกับข้อมูลที่มี Feature Data น้อย
- สามารถใช้ได้กับ Discrete Function เท่านั้น

#### 2.1.5.2 ชั้พพอร์ตเวกเตอร์แมชชีน

ชั้พพอร์ตเวกเตอร์แมชชีน หรือ Support Vector Machine เป็นหนึ่งในวิธีการเรียนรู้ของเครื่อง (Machine Learning) ด้วยขั้นตอนวิธีที่ใช้ในการวิเคราะห์ข้อมูลเพื่อจำแนกประเภทข้อมูล (Classification) และการวิเคราะห์การถดถอย (Regression) ชั้พพอร์ตเวกเตอร์แมชชีนได้ถูกพัฒนาในกรอบแนวคิดของทฤษฎีการเรียนรู้ทางสถิติและถูก

นำไปประยุกต์ใช้ในหลาย ๆ ด้าน เช่น เพื่อสนับสนุนการตัดสินใจของแพทย์, การระบุใบหน้า เป็นต้น

ชั้พพอร์ตเวกเตอร์เมชชีนเป็นกระบวนการเรียนรู้ที่ต้องมีผู้สอน (Supervised Learning) นั่นคือจะต้องมีผู้สอนก่อนจึงจะสามารถแบ่งกลุ่ม (Classify) ข้อมูลได้ โดยผู้สอน ที่ว่านี้คือเซตของข้อมูลที่เรารู้อยู่แล้วว่าข้อมูลนั้นอยู่กลุ่ม (Class) ใดในการสอนเพื่อสามารถ แบ่งกลุ่มแบบเชิงเส้นได้ โดยการแบ่งกลุ่มจะทำการแบ่งด้วยการใช้เส้นระนาบเกิน หรือ Hyperplane ใน การแบ่งข้อมูลในแต่ละกลุ่มออกจากกันด้วยระยะห่างที่มากที่สุด (Maximum Margin) หลักการหา Hyperplane จะใช้หลักการของฟังก์ชันเชิงเส้น (Linear Function) ที่อยู่ตรงกลางระหว่างกลุ่มใดๆ ในกรณีของการแบ่งกลุ่มออกจากกันจะแทนค่า เป้าหมาย (Target) กลุ่มนึงด้วยค่า “+1” และ กลุ่มที่สองด้วยค่า “-1” ซึ่งจะได้สมการ ดังนี้

$$f(x) = (w_0x) + b$$

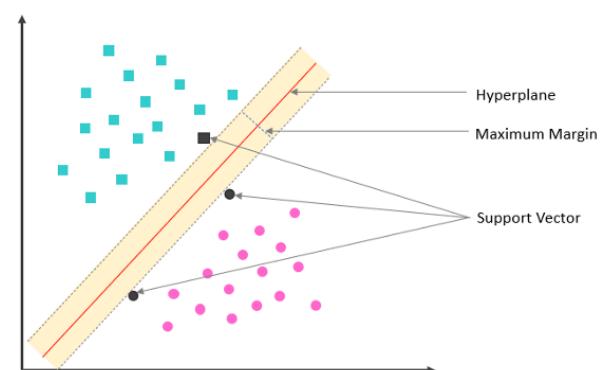
เมื่อ

$f(x)$  คือ ค่าที่คำนวณโดย model

$w$  คือ ค่าน้ำหนักในการคำนวณ

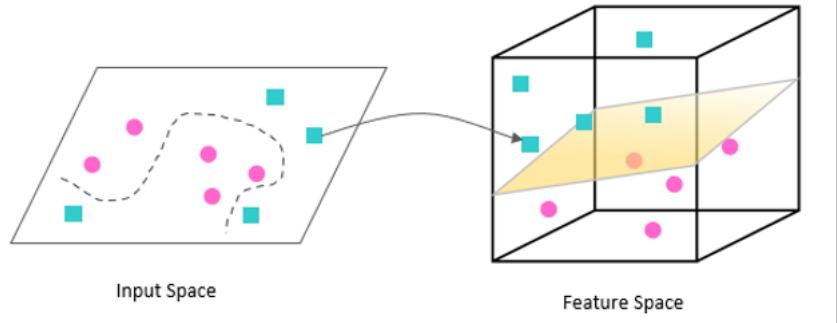
$b$  คือ ค่าไบอัส (Bias)

เมื่อทำการสอนเสร็จสิ้นแล้วการนำเวกเตอร์  $x$  ได้มาจัดเข้ากลุ่มสามารถ ทำได้โดยการแทนค่า  $x$  ลงในสมการ และพิจารณาค่า  $f(x)$  ที่เกิดขึ้น หากค่าที่ได้เป็นบวก แสดงว่าค่านั้นอยู่ในกลุ่มที่หนึ่ง หากค่าที่ได้เป็นลบก็แสดงว่าค่านั้นอยู่ในกลุ่มที่สองและหาก เท่ากับศูนย์จะสามารถเลือกให้ข้อมูลอยู่ในกลุ่มไหนก็ได้



ภาพที่ 2.1.1 แสดงตัวอย่างการจำแนกข้อมูลด้วย Support Vector Machine

ในกรณีที่ชุดข้อมูลไม่สามารถแบ่งออกด้วย Linear Hyperplane ได้ ฟังก์ชัน Kernal จึงถูกนำมาใช้ในการหารูปแบบและความสัมพันธ์ของข้อมูลเพื่อช่วยให้การแบ่งที่เป็น Non Linear Function โดยวิธีการคือการสร้างมิติจากเดิม 2 มิติ ให้เป็น 3 มิติ



ภาพที่ 2.1.2 แสดงตัวอย่างการทำ Support Vector Machine ด้วย Kernel Function

ที่มา : <https://www.nectec.or.th/news/news-public-document/machine-learning-manufact-1.html>

#### ข้อมูลที่เหมาะสมกับการทำซัพพอร์ตเวกเตอร์แมชชีน

- ชุดข้อมูลที่มี Feature จำนวนมาก แต่มีปริมาณข้อมูลน้อยถึงปานกลาง เหมาะกับข้อมูลที่มีการแยกประเภทหรือจำแนกประเภท

#### จุดเด่นของซัพพอร์ตเวกเตอร์แมชชีน

- สามารถจัดการกับข้อมูลที่เป็น Non Linear ได้ด้วยการใช้เทคนิค Kernel
- มีความยืดหยุ่นและทำงานได้ดี โดยเฉพาะข้อมูลที่มีความซับซ้อนหลาย Feature
- ในการใช้ Kernel Function มีให้เลือกใช้งานได้หลายชนิด เช่น Linear , Polynomial, RBF (Radial Basis Function)

#### จุดอ่อนของซัพพอร์ตเวกเตอร์แมชชีน

- อาจทำงานได้ไม่ดีกับข้อมูลที่เป็น Dataset ขนาดใหญ่ เนื่องจากใช้เวลาในการฝึกสอนนาน
- เนื่องจากมี Kernel Function ให้เลือกใช้งานได้หลายชนิดจึงต้องระมัดระวังในการเลือกใช้งานให้เหมาะสมสมด้วย

### 2.1.6 เมทริกซ์คอนพิวชัน

## Confusion Matrix

		Actually Positive (1)	Actually Negative (0)	
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)		
	False Negatives (FNs)	True Negatives (TNs)		

ภาพที่ 2.1.3 แสดงตัวอย่างตาราง Confusion Matrix

ที่มา : <https://everydaymarketing.co/business-and-marketing-case-study/ai/confusion-matrix-ml-evaluation/>

เมทริกซ์คอนพิวชัน หรือ Confusion Matrix คือตารางที่ใช้ประเมินประสิทธิภาพของ Classification Machine Learning ในรูปแบบต่างๆ ร่วมกับชุดข้อมูลทดสอบที่เราทราบค่าหรือผลลัพธ์ที่แท้จริงแล้ว ด้วยการประเมินค่าจริง (Actual value) เทียบกับค่าที่ได้จากผลลัพธ์ของ Machine Learning model หรือค่าที่คาดการณ์ไว้ (Predicted Value) แล้วสรุปค่าอุอกมาในรูปแบบของตารางเมทริกซ์ที่ประกอบด้วยค่า True negative (TN), False positive (FP), False negative (FN) และ True positive (TP) ซึ่งแต่ละค่ามีความหมาย ดังนี้

- (1) True Positive (TP) คือ จำนวนข้อมูลที่ถูกจำแนกได้อย่างถูกต้องอยู่ในกลุ่มประเภทที่กำลังต้องการหรือสนใจ
- (2) True Negative (TN) คือ จำนวนข้อมูลที่ถูกจำแนกถูกต้องในกลุ่มที่ไม่ใช่เป้าหมายหรือกลุ่มที่ไม่ได้สนใจ
- (3) False Positive (FP) คือ จำนวนข้อมูลที่ถูกทำนายผิดในกลุ่มข้อมูลที่เป็นกลุ่มเป้าหมายหรือกลุ่มที่สนใจ
- (4) False Negative (FN) คือ จำนวนข้อมูลที่ถูกจำแนกผิดในกลุ่มข้อมูลที่ไม่ใช่เป้าหมายหรือกลุ่มข้อมูลที่ไม่ได้สนใจ

### 2.1.7 การวัดประสิทธิภาพของแบบจำลอง

#### 2.1.7.1 ค่าความถูกต้อง

ค่าความถูกต้อง หรือ Accuracy เป็นค่าที่นิยมนำมาใช้วัดประสิทธิภาพของโมเดลอย่างแพร่หลาย ใช้สำหรับบอกว่าโมเดลทำนายได้ถูกต้องกับสิ่งที่เกิดขึ้นจริงเท่าไหร่ เทียบกับส่วนที่ทำนายทั้งหมด จากสูตร

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

กล่าวคือค่าความถูกต้องจะใช้บอกความสามารถของแบบจำลองโดยรวมในการทำนายทั้งกรณี Positive และ Negative

#### 2.1.7.2 ค่าความแม่นยำ

ค่าความแม่นยำ หรือ Precision คือ ค่าความถูกต้องของประเด็นที่โมเดลทำนาย ซึ่งสามารถหาได้จาก

$$\text{ค่าความแม่นยำ} = \frac{TP}{(TP+FP)}$$

กล่าวคือเป็นค่าที่จะใช้พิจารณาถึงความสามารถของระบบในการทำนาย positive โดยเฉพาะ

#### 2.1.7.3 ค่าความอ่อนไหว

ค่าความเริ่ม ค่าไว หรือ Recall ในทางสถิติจะใช้คำว่า Sensitivity คือค่าที่บ่งบอกว่าแบบจำลองที่สร้างนั้นเจอ Actual Positive Class ที่เราสนใจมากน้อยเพียงใดด้วยสูตร

$$\text{ค่าความอ่อนไหว} = \frac{TP}{(TP+FN)}$$

กล่าวคือเป็นค่าที่จะใช้ประเมินความสามารถของแบบจำลองในการประเมินผล Positive ในกรณีที่มี Positive จริงๆ

#### 2.1.7.4 F1 Score

เป็นการเอาค่า Recall กับ Precision มาหาเฉลี่ยแบบ Harmonic (Harmonic Mean) ซึ่งทำให้ค่าที่ออกมาเป็นการวัดประสิทธิภาพของโมเดลด้วย 2 เมทริก รวมกัน ซึ่งสามารถหาได้จากสูตร

$$F1\ score = 2 \frac{(Precision \times Recall)}{(Precision + Recall)}$$

โดยที่ค่า F1-Score มีค่าสูงเมื่อทั้งความแม่นยำในการ Predict กลุ่มเป้าหมาย และความสามารถในการประเมิน Positive มีค่าสูงพร้อมกัน

## 2.2 งานวิจัยที่เกี่ยวข้อง

### 2.2.1 ระบบการวิเคราะห์ความรู้สึกจากวิดีโอบนโซเชียลมีเดียด้วยชั้ฟพอร์ตเวกเตอร์แมชชีน

ปรีดาวรรณ เกษเมธีการุณ, พยุง มีส์จ และ สุมิตรา นวลมีศรี (2559) ได้จัดทำวิจัย ระบบการวิเคราะห์ความรู้สึกจากวิดีโอบนโซเชียลมีเดียด้วยชั้ฟพอร์ตเวกเตอร์แมชชีน เป็นการพัฒนา ระบบการวิเคราะห์ความรู้สึกจากวิดีโอด้วยเทคนิคชั้ฟพอร์ตเวกเตอร์แมชชีนเพื่อเป็นแนวทางในการ ตัดสินใจในการซื้อของผู้บริโภคและการจัดการข้อมูลของภาคธุรกิจต่อวิดีโอบนสื่อสังคมออนไลน์ทั้ง เชิงบวกและเชิงลบด้วยเทคนิคการเรียนรู้ของเครื่อง ซึ่งงานวิจัยเป็นการรวบรวมข้อมูลวิดีโอบทวิจารณ์ จาก YouTube จำนวน 500 วิดีโอ ซึ่งแต่ละวิดีโอมีความยาวอยู่ระหว่าง 2-5 นาที และนำมาสกัดเป็น ข้อความและตัดคำเพื่อนำไปสกัดคุณลักษณะของคำโดยอาศัยน้ำหนักของคำนำมานำมาคำนวณค่า Term Frequency (TF) และ Inverse Documents Frequency (IDF) และนำไปวิเคราะห์ความรู้สึกว่า เป็นเชิงบวกหรือเชิงลบโดยทำการเปรียบเทียบกับคลังข้อมูลด้วยเทคนิคชั้ฟพอร์ตเวกเตอร์แมชชีน จากการทดลองพบว่าระบบมีประสิทธิภาพดีทั้งด้านความถูกต้องอยู่ที่ร้อยละ 98 เมื่อเทียบกับ ผู้เชี่ยวชาญ 15 คน

ข้อจำกัดของงานวิจัยระบบการวิเคราะห์ความรู้สึกจากวิดีโอบนโซเชียลมีเดียด้วยชั้ฟ พอร์ตเวกเตอร์แมชชีน คือ ระบบใช้ข้อมูลในการสอนเป็นวิดีโocom ความคิดเห็นเกี่ยวกับผลิตภัณฑ์ เครื่องสำอางจำนวน 500 วิดีโอเท่านั้น โดยสามารถจำแนกได้เป็นความรู้สึกเชิงบวกและความรู้สึกเชิง ลบ ภาษาไทยเท่านั้น ซึ่งข้อมูลที่นำมาทดสอบมาจากผู้ทำแบบทดสอบจำนวน 60 คน ซึ่งเป็นนักศึกษา ระดับปริญญาตรี

### 2.2.2 การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีชัฟ พอร์ตเวกเตอร์แมชชีน

ริสุดา เทศเมือง และ นิเวศ จิระวิชิตชัย (2560) ได้จัดทำงานวิจัยเรื่องการวิเคราะห์ ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีชัฟพอร์ตเวกเตอร์แมชชีน ซึ่ง

เป็นการจำแนกความคิดเห็นโดยการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ ด้านการบริการห้องพัก โรงแรม รีสอร์ท จาก Agoda Thailand และ Twitter Thailand ที่จะทะเบียนหลักทรัพย์ โดยจัดเก็บข้อมูลจำนวน 2,890 ข้อมูล โดยใช้เทคนิคเหมืองข้อมูลวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ และสร้างแบบจำลองด้วยอัลกอริทึมทั้งหมด 4 วิธี ได้แก่ ชัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ นาอีฟเบย์ และเคนเนยสเรสเนเบอร์ เพื่อเปรียบเทียบประสิทธิภาพในการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ และจากการทดลองพบว่าอัลกอริทึมที่มีคุณลักษณะดีที่สุด คือ ชัพพอร์ตเวกเตอร์แมชชีน ซึ่งมีค่าความถูกต้องสูงที่สุดเท่ากับ 83.38%

ข้อจำกัดของวิจัยการวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์ โดยใช้ขั้นตอนวิธีชัพพอร์ตเวกเตอร์แมชชีน คือ สามารถวิเคราะห์ได้เฉพาะข้อความที่เป็นภาษาไทยเท่านั้น และเป็นข้อมูลที่ได้จาก Agoda Thailand และ Twitter Thailand เท่านั้น โดยระบบจะทำการคำนวณรูปแบบของภาษาไทยและภาษาอื่นๆ ไม่สามารถวิเคราะห์ได้

### **2.2.3 การระบุรีวิวที่น่าสงสัยในผลิตภัณฑ์ดูแลผิวน้ำของลูกค้าบนตลาดชื้อขายสินค้าออนไลน์ ด้วยเทคนิควิธีการเรียนรู้ของเครื่อง**

เพญุนภา จุมพลพงษ์ (2564) ได้จัดทำงานวิจัยการระบุรีวิวที่น่าสงสัยในผลิตภัณฑ์ดูแลผิวน้ำของลูกค้าบนตลาดชื้อขายสินค้าออนไลน์ด้วยเทคนิควิธีการเรียนรู้ของเครื่องซึ่งเป็นงานวิจัยที่ต้องการระบุรีวิวที่มีความน่าสงสัยในผลิตภัณฑ์ดูแลผิวน้ำที่สามารถช่วยลูกค้าที่ต้องการอ่านรีวิวอันดับสิบในชื่อผลิตภัณฑ์ดูแลผิวน้ำบนตลาดสินค้าออนไลน์ โดยทำการเก็บรวบรวมข้อมูลจากรีวิวผลิตภัณฑ์ดูแลผิวน้ำบนตลาดชื้อขายสินค้าออนไลน์แห่งหนึ่งที่เป็นภาษาไทยระหว่างเดือนมกราคมถึงเดือนพฤษจิกายน 2564 จากผลิตภัณฑ์ดูแลผิวน้ำ 10 ยี่ห้อ ที่มียอดขายสูงสุดตามลำดับจำนวน 60,081 รีวิว และนำมาพัฒนาแบบจำลองโดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning) ประกอบด้วย 3 เทคนิค ได้แก่ เทคนิคนาอีฟเบย์ (Naives Bays) เทคนิคชัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) เทคนิคต้นไม้ตัดสินใจ (Decision Tree) จำนวน 5 แบบ ในการประเมินค่าความถูกต้อง (Accuracy) มากที่สุดเท่ากับ 85%

ข้อจำกัดของวิจัยการระบุรีวิวที่น่าสงสัยในผลิตภัณฑ์และผิวน้ำของลูกค้าบันดาดซื้อขายสินค้าออนไลน์ด้วยเทคนิคบริการเรียนรู้ของเครื่องคือประชาร์ทที่ใช้ในการทำวิจัย ได้แก่ รีวิวผลิตภัณฑ์ดูแลผิวน้ำบันดาดซื้อขายสินค้าออนไลน์ โดยเลือกเก็บข้อมูล 8 คุณลักษณะ ได้แก่ ยี่ห้อสินค้า ราคา ความเร็วในการตอบกลับ ชื่อลูกค้า ข้อมูลรีวิว วันเวลาที่รีวิว และจำนวนดาว ซึ่งทำการเก็บรวบรวมข้อมูลระหว่างเดือนพฤษภาคม - ธันวาคม พ.ศ.2564 โดยสามารถจำแนกประเภทของรีวิวเป็นรีวิวจริง (Real) รีวิวน่าสงสัย (Suspicious) รีวิวที่ไม่มีข้อความ (Missing Value) และค่าที่ไม่เกี่ยวข้อง (Error) และความรู้สึกเชิงบวก ความรู้สึกเชิงลบ และความรู้สึกเป็นกลางเท่านั้น

#### 2.2.4 ระบบวิเคราะห์ความคิดเห็นต่อ栎ครไทยบันทวิตเตอร์

ธนสิทธิ์ เร่งสมบูรณ์สุข และ กวินธิดา สายยศ (2562) ได้จัดทำงานวิจัยระบบวิเคราะห์ความคิดเห็นต่อ栎คนไทยบันทวิตเตอร์ขึ้นมาเนื่องจากปัจจุบันผู้บริโภคส่วนใหญ่ที่สนใจหรือซึ่งชอบจะเปลี่ยนความคิดเห็นผ่านสื่อออนไลน์ เช่น ทวิตเตอร์ (Twitter) ในการแลกเปลี่ยนความคิดเห็นทำให้มีข้อความคิดเห็นจำนวนมากและต้องใช้เวลาในการอ่านและทำความเข้าใจข้อความทั้งหมด จึงได้พัฒนาระบบนี้ขึ้นมาเพื่อวิเคราะห์ความคิดเห็นเกี่ยวกับ栎คนไทยบันทวิตเตอร์ เพื่อเป็นเครื่องมือที่ช่วยวิเคราะห์และสรุปผลความชอบที่มีต่อ栎ประเทศไทย เป็นตัวช่วยในการตัดสินใจเลือกซื้อผลิตภัณฑ์ง่ายและรวดเร็ว โดยวิเคราะห์ข้อความแสดงความคิดเห็นออกเป็น ขอบ ไม่ชอบ และความคิดเห็นระดับกลาง ในการพัฒนาระบบใช้ไลบรารี PythaiNLP ช่วยในการตัดคำ และสร้างโมเดลสำหรับการเรียนรู้ด้วยเครื่องโดยใช้เทคนิคนาอีฟเบย์ จากการทดสอบการจำแนกข้อความแสดงความคิดเห็นพบว่าค่าความถูกต้องและความแม่นยำมีค่าประมาณ 70% มีประสิทธิภาพยังไม่เพียงพอในการทำนายความคิดเห็นเชิงลบ

ข้อจำกัดของงานวิจัยระบบวิเคราะห์ความคิดเห็นต่อ栎คนไทยบันทวิตเตอร์ คือ สามารถวิเคราะห์ได้เฉพาะข้อมูลภาษาไทยเท่านั้น และไม่สามารถวิเคราะห์คำที่สะกดผิดและคำแสลงได้ สามารถจำแนกความคิดเห็นออกเป็น ขอบ ไม่ชอบ และความรู้สึกเป็นกลาง ข้อมูลที่ใช้ในการสอนมีจำนวนจำกัดทำให้ไม่สามารถจำแนกข้อความด้านลบได้ดีพอ

### 2.2.5 การเปรียบเทียบคุณสมบัติของงานวิจัยที่เกี่ยวข้อง

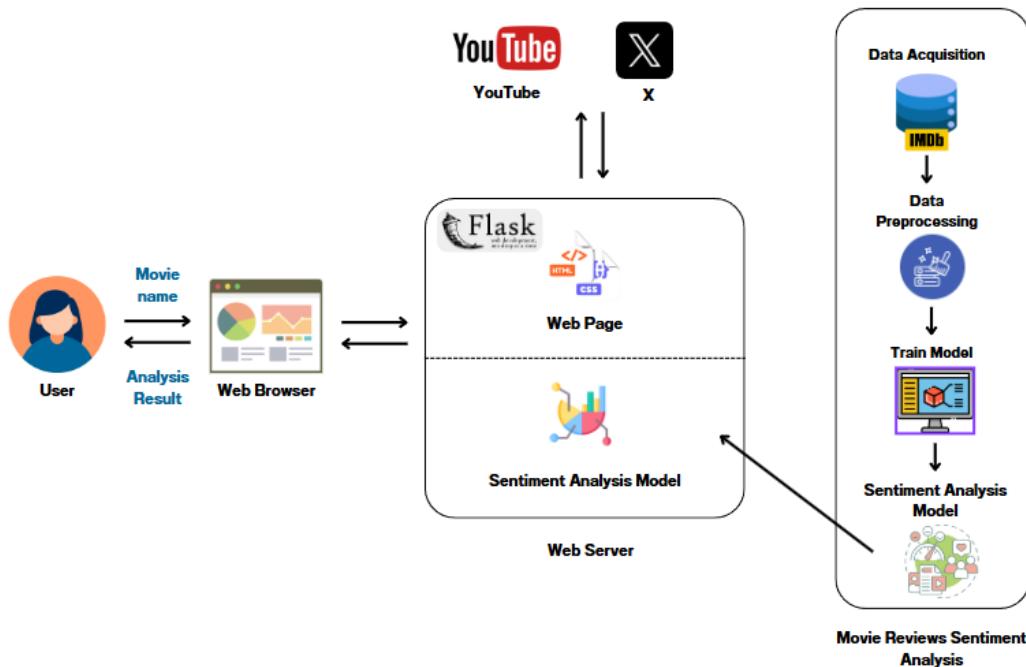
ตารางที่ 2.2.5 ตารางเปรียบเทียบคุณสมบัติของงานวิจัยที่เกี่ยวข้อง

	ระบบการวิเคราะห์ความรู้สึกจากวิดีโอบนโซเชียลมีเดียด้วยซัพพอร์ตเวกเตอร์แมชชีน	การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวสินค้าออนไลน์โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน	การระบุรีวิวที่นำเสนอในผู้คนที่ดูแลผู้ห้ามทดลองลูกค้าบนตลาดอื่นขายสินค้าออนไลน์ ด้วยเทคนิควิธีการเรียนรู้ของเครื่อง	ระบบวิเคราะห์ความคิดเห็นต่อผลกระทบไทยบนทวิตเตอร์	การวิเคราะห์ความคิดเห็นของภาคยนตร์ผ่านแพลตฟอร์มออนไลน์
อัลกอริทึมที่ใช้ในการทดสอบ	ซัพพอร์ตเวกเตอร์แมชชีน	ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ นาอีฟเบย์ และเคเนียสเรสนेबอร์	ซัพพอร์ตเวกเตอร์แมชชีน ต้นไม้ตัดสินใจ และนาอีฟเบย์	นาอีฟเบย์	ซัพพอร์ตเวกเตอร์แมชชีน และการถดถอยโลจิสติก
การจำแนกประเภทของข้อความ	จำแนกเป็นความรู้สึกเชิงบวกและความรู้สึกเชิงลบ	จำแนกเป็นความรู้สึกเชิงบวกและความรู้สึกเชิงลบ	จำแนกความคิดเห็นเป็นรีวิวจริง รีวิวน่าสนใจ และรีวิวที่ไม่มีข้อความ	จำแนกความคิดเห็นเป็นช่อง ไม่ชอบ และเป็นกลาง	จำแนกเป็นความรู้สึกเชิงบวกและความรู้สึกเชิงลบ
ภาษาที่ใช้ในการวิเคราะห์	ภาษาไทย	ภาษาไทย	ภาษาไทย	ภาษาไทย	ภาษาอังกฤษ
แหล่งข้อมูลใน การวิเคราะห์	คลิปวิดีโอบน YouTube	Agoda Thailand และ Twitter Thailand (X)	Twitter (X)	Twitter (X)	YouTube และ X

## บทที่ 3

### วิธีการวิจัย

#### 3.1 ภาพรวมของโครงการ



ภาพที่ 3.1.1 ตัวอย่างสถาปัตยกรรมภาพรวมของการวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์มออนไลน์

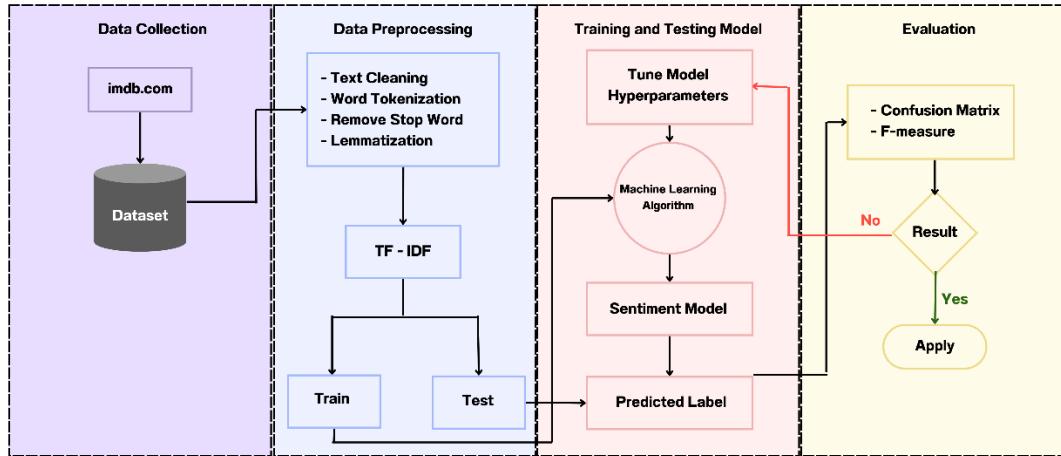
คำอธิบายสถาปัตยกรรมภาพรวมของการวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์มออนไลน์มีขั้นตอนดังนี้

- User หรือ ผู้ใช้งาน คือ บุคคลที่ร้องขอใช้งานการวิเคราะห์ความคิดเห็นของภาพยนตร์จากแพลตฟอร์มออนไลน์ผ่านทางเว็บเบราว์เซอร์
- Web Browser หรือ เบราว์เซอร์ คือ ตัวกลางในการแลกเปลี่ยนข้อมูลระหว่างผู้ใช้งานและเครื่องแม่ข่าย (Web Server) ของระบบ ในการทดลองใช้ Microsoft Edge
- Web Server หรือ เครื่องแม่ข่าย เก็บโนเดลการวิเคราะห์และประมวลผล เป็นส่วนบริการผู้ใช้งานเมื่อได้รับคำร้องขอจากเว็บเบราว์เซอร์โดยให้บริการการวิเคราะห์ความคิดเห็นของภาพยนตร์

- YouTube เป็นแพลตฟอร์มที่เก็บความคิดเห็นของภาพยนตร์ที่ระบบจะดึงความคิดเห็น (Comments) ได้คลิปวิดีโอที่เกี่ยวข้องกับภาพยนตร์ผ่านการใช้ YouTube API ซึ่งทำการค้นหาด้วยชื่อของภาพยนตร์ที่ต้องการนำข้อมูลมาวิเคราะห์
- X เป็นแพลตฟอร์มเก็บข้อมูลข้อความแสดงความคิดเห็นที่เกี่ยวกับภาพยนตร์ที่ระบบจะดึงข้อมูลผ่านการใช้ X API ซึ่งข้อมูลความคิดเห็นที่นำมาวิเคราะห์จะทำค้นหาผ่านแฮชแท็ก (Hashtag) ของชื่อเรื่องภาพยนตร์
- การวิเคราะห์ความคิดเห็นของภาพยนตร์ (Movie Reviews Sentiment Analysis)
  - IMDB คือ แหล่งข้อมูลเกี่ยวกับความคิดเห็นต่อภาพยนตร์ที่ถูกรวบรวมเอาไว้ทั้งหมด 50,000 ข้อความ จากเว็บไซต์ [www.imdb.com](http://www.imdb.com) เป็นข้อมูลสำหรับการฝึกสอนโมเดล
  - การเตรียมข้อมูล (Data Preprocessing) เป็นขั้นตอนสำหรับการเตรียมข้อมูลให้พร้อมนำไปใช้งานใน Machine Learning เพื่อให้แบบจำลองสามารถทำงานค่าได้อย่างมีประสิทธิภาพมากที่สุด การทำความสะอาดข้อความ (Data Cleaning) มีหลายรูปแบบ ได้แก่ การลบเครื่องหมายหรืออักษรพิเศษ ลบข้อความที่มี HTML Tag การตัดคำ (word Tokenization) เพื่อให้ตัวหนังสือที่อยู่ในรูปประโยคข้อความมีหน่วยเล็กลงเป็นคำ การลบคำที่เป็นคำพิธีนฐาน (Stop Word Removal) เพื่อลบคำที่ไม่ได้ใช้ในการสื่อความหมาย การ Stemming และ การ Lemmatization เพื่อแปลงไวยากรณ์ให้อยู่ในรูปเดียวกัน
  - การสร้างและฝึกสอนระบบด้วยแบบจำลองสองแบบ ได้แก่ แบบจำลองการวิเคราะห์ด้วยซัพพอร์ตเวกเตอร์เมชัน และแบบจำลองการวิเคราะห์การถดถอยโลจิสติก เพื่อหาแบบจำลองที่มีประสิทธิภาพมากที่สุดมาใช้สำหรับการวิเคราะห์ความคิดเห็น

หลังจากผ่านกระบวนการในการเรียนรู้ข้อมูลแล้วแบบจำลองจะสามารถวิเคราะห์และจำแนกความคิดเห็นที่มีต่อภาพยนตร์ได้ว่าเป็นความคิดเห็นเชิงบวกหรือความคิดเห็นเชิงลบ เพื่อนำไปใช้และติดตั้งบน Web Server ต่อไป

### 3.2 การวิเคราะห์ข้อบอกร่องและความต้องการของระบบ



ภาพที่ 3.2.1 ตัวอย่างการวิเคราะห์ข้อบอกร่องและความต้องการของการวิเคราะห์

#### 3.2.1 การรวบรวมข้อมูล (Data Collection)

โครงการนี้ใช้เป็นชุดข้อมูลที่สามารถเข้าถึงได้สาธารณะจากเว็บไซต์ [www.kaggle.com](http://www.kaggle.com) ซึ่งเป็นชุดข้อมูลความคิดเห็นต่อภาพยนตร์จากเว็บไซต์ [www.imdb.com](http://www.imdb.com) จำนวน 50,000 ข้อความ ชุดข้อมูลมีป้ายกำกับ (Label) ความคิดเห็นออกเป็น 2 ประเภท ได้แก่ ความคิดเห็นเชิงบวก (Positive) และความคิดเห็นเชิงลบ (Negative) โดยจำนวนข้อมูลแยกตามประเภทความรู้สึกได้ ดังนี้

ตารางที่ 3.2.1 ตารางแสดงจำนวนข้อมูลแยกตามประเภทความรู้สึก

ความรู้สึก	จำนวนข้อมูล (Rows)
เชิงบวก	25,000
เชิงลบ	25,000
รวม	50,000

#### 3.2.2 การเตรียมข้อมูล (Data Preprocessing)

##### 3.2.2.1 การจัดการกับข้อความ

- (1) การทำความสะอาดข้อความ (Text Cleaning)

เนื่องจากข้อมูลที่นำมาทำการทดลองมาจากแพลตฟอร์มออนไลน์ ผู้ใช้งานส่วนใหญ่มักพิมพ์ข้อความด้วยอักษรย่อ หรือสัญลักษณ์อิโมจิ ซึ่งเป็นตัวอักษรเครื่องหมาย หรืออักษรที่ทำให้เมื่อนำข้อมูลมาใช้สร้างแบบจำลองแล้วส่งผลให้แบบจำลองไม่สามารถทำงานได้อย่างมีประสิทธิภาพเท่าที่ควรจึงควรทำความสะอาดข้อมูลก่อนเพื่อให้ได้ข้อมูลที่เหมาะสมและพร้อมนำไปใช้งานอย่างมีประสิทธิภาพมากที่สุด โดยวิธีทำความสะอาดข้อมูลที่ใช้ได้แก่ การปรับตัวอักษรให้ถูกต้องเป็นตัวอักษรพิมพ์เล็กทั้งหมด การลบเครื่องหมายวรรคตอน (Punctuation) การลบเครื่องหมายหรืออักษรพิเศษ การลบข้อความที่มี HTML Tag และการขยายคำย่อภาษาอังกฤษ (Contractions)

A wonderful little production. <br /><br />The filming technique is very unassuming - very old-time-BBC fashion and gives a comforting, and sometimes discomforting, sense of realism to the entire piece. <br /><br />The actors are extremely well chosen - Michael Sheen not only "has got all the polaris" but he has all the voices down pat too! You can truly see the seamless editing guided by the references to Williams' diary entries, not only is it well worth the watching but it is a terrifically written and performed piece. A masterful production about one of the great masters of comedy and his life. <br /><br />The realism really comes home with the little things: the fantasy of the guard which, rather than use the traditional 'dream' techniques remains solid then disappears. It plays on our knowledge and our senses, particularly with the scenes concerning Orton and Halliwell and the sets (particularly of their flat with Halliwell's murals decorating every surface) are terribly well done

ภาพที่ 3.2.2 รูปแสดงตัวอย่างข้อความก่อนทำการทำความสะอาด

a wonderful little production the filming technique is very unassuming very old time bbc fashion and gives a comforting and sometimes discomforting sense of realism to the entire piece the actors are extremely well chosen michael sheen not only has got all the polaris but he has all the voices down pat too you can truly see the seamless editing guided by the references to williams diary entries not only is it well worth the watching but it is a terrifically written and performed piece a masterful production about one of the great masters of comedy and his life the realism really comes home with the little things the fantasy of the guard which rather than use the traditional dream techniques remains solid then disappears it plays on our knowledge and our senses particularly with the scenes concerning orton and halliwell and the sets particularly of their flat with halliwell s murals decorating every surface are terribly well done

ภาพที่ 3.2.3 รูปแสดงตัวอย่างข้อความหลังทำการทำความสะอาด

## (2) การลบคำหยุด (Removal Stop Word)

เป็นขั้นตอนสำหรับการลบคำที่เป็นคำพื้นฐานที่มักพบบ่อยในประโยคหรือข้อความแต่ไม่ได้ใช้ในการสื่อความหมาย เช่น is, was, and ,this วิธีทำงานคือหากพบคำใดที่ตรงกับคำที่อยู่ในลิสต์ของ stopwords จะทำการลบคำนั้นออกไป

wonderful little production filming technique is very unassuming very old time fashion and gives a comforting and sometimes discomforting sense of realism to the entire piece actors are extremely well chosen not only has got all the polar but he has all the voices down pat too can truly see the seamless editing guided by the references to diary entries not only is it well worth the watching but it is a terribly written and performed piece masterful production about one of the great master of comedy and his life the realism really comes home with the little things the fantasy of the guard which rather than use the traditional dream techniques remains solid then disappears It plays on our knowledge and our senses particularly with the scenes concerning and and the sets particularly of their flat with halliwell murals decorating every surface are terribly well done

ภาพที่ 3.2.4 รูปแสดงตัวอย่างคำที่เป็น Stop Word

wonderful little production filming technique unassuming old time bbc fashion gives comforting sometimes discomforting sense realism entire piece actors extremely well chosen michael sheen got polari voices pat truly see seamless editing guided references williams diary entries well worth watching terrificly written performed piece masterful production one great master comedy life realism really comes home little things fantasy guard rather use traditional dream techniques remains solid disappears plays knowledge senses particularly scenes concerning orton halliwell sets particularly flat halliwell murals decorating every surface terribly well done

ภาพที่ 3.2.5 รูปแสดงตัวอย่างข้อมูลหลังจากตัด Stop Word

## (3) Lemmatization

เป็นขั้นตอนในการผันไวยากรณ์ในภาษาอังกฤษให้คำที่มีความหมายเดียวกันอยู่ในรูปเดียวกัน เช่น “Run”, “Running”, “Runs” จะถูกแปลงอยู่ในรูปเดียวกัน คือ Run

wonderful little production filming technique unassuming old time bbc fashion gives comforting sometimes discomforting sense realism entire piece actors extremely well chosen michael sheen got polari voices pat truly see seamless editing guided eferences williams diary entries well worth watching terrificly written performed piece masterful production one great master comedy life realism really comes home little things fantasy guard rather use traditional dream techniques remains solid disappears plays knowledge senses particularly scenes concerning orton halliwell sets particularly flat halliwell murals decorating surface terribly well done

ภาพที่ 3.2.6 รูปแสดงตัวอย่างข้อมูลก่อนการทำ Lemmatization

wonderful little production filming technique unassuming old time bbc fashion give comfort sometimes discomforte sense realism entire piece actor extremely well choose michael sheen get polari voice pat truly see seamless editing guide reference williams diary entry well worth watch terrificly write perform piece masterful production one great master comedy life realism really come home little thing fantasy guard rather use traditional dream technique remain solid disappear play knowledge sense particularly scene concern orton halliwell set particularly flat halliwell mural decorate every surface terribly well do

ภาพที่ 3.2.7 รูปแสดงตัวอย่างข้อมูลที่ผ่านการทำ Lemmatization

#### (4) การตัดคำ (Word Tokenization)

เป็นขั้นตอนที่ใช้ในการตัดข้อความตัวหนังสือที่อยู่ในรูปประโยคให้อยู่ในหน่วยที่เล็กลง ซึ่งอาจอยู่ในรูปของคำ (Token) เช่น ข้อความ “This movie was fantastic and amazing” จะสามารถตัดคำได้เป็นลิสต์ของคำ คือ [‘This’, ‘movie’, ‘was’, ‘fantastic’, ‘and’, ‘amazing’] เป็นต้น

[‘wonderful’, ‘little’, ‘production’, ‘filming’, ‘technique’, ‘unassuming’, ‘old’, ‘time’, ‘bbc’, ‘fashion’, ‘give’, ‘comfort’, ‘sometimes’, ‘discomforte’, ‘sense’, ‘realism’, ‘entire’, ‘piece’, ‘actor’, ‘extremely’, ‘well’, ‘choose’, ‘michael’, ‘sheen’, ‘get’, ‘polari’, ‘voice’, ‘pat’, ‘truly’, ‘see’, ‘seamless’, ‘editing’, ‘guide’, ‘reference’, ‘williams’, ‘diary’, ‘entry’, ‘well’, ‘worth’, ‘watch’, ‘terrificly’, ‘write’, ‘perform’, ‘piece’, ‘masterful’, ‘production’, ‘one’, ‘great’, ‘master’, ‘comedy’, ‘life’, ‘realism’, ‘really’, ‘come’, ‘home’, ‘little’, ‘thing’, ‘fantasy’, ‘guard’, ‘rather’, ‘use’, ‘traditional’, ‘dream’, ‘technique’, ‘remain’, ‘solid’, ‘disappear’, ‘play’, ‘knowledge’, ‘sense’, ‘particularly’, ‘scene’, ‘concern’, ‘orton’, ‘halliwell’, ‘set’, ‘particularly’, ‘flat’, ‘halliwell’, ‘mural’, ‘decorate’, ‘every’, ‘surface’, ‘terribly’, ‘well’, ‘do’]

ภาพที่ 3.2.8 รูปแสดงตัวอย่างข้อมูลหลังการตัดคำ

### 3.2.2.2 TF-IDF

หลังจากการทำความสะอาดและจัดการกับข้อความแล้วจะนำข้อความที่ได้มาทำการคัดแยกคำตามความสำคัญโดยการให้ค่าน้ำหนักของคำในแต่ละคำโดยคำนึงถึงสองปัจจัยคือ ค่า TF หรือค่าความถี่ของคำศัพท์ที่พบในข้อความ และค่า IDF คือค่าความถี่ผกผันของคำที่พบในข้อความซึ่งเป็นค่าที่คำนวณจากคำที่พบในข้อความทั้งหมด ในการเขียนโปรแกรมสามารถใช้ฟังก์ชันในไลบรารีไซคิท-ลิรัน ซึ่งเป็นโมดูลในภาษาไพทอนมาสำหรับใช้คำนวณค่า TF-IDF ของข้อมูลได้ดังต่อไปนี้

```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf = TfidfVectorizer()
x = tfidf.fit_transform(x)

x.shape
✓ 3.8s
```

Python

ภาพที่ 3.2.9 รูปแสดงตัวอย่างโค้ดการใช้งานโมดูล TfidfVectorizer

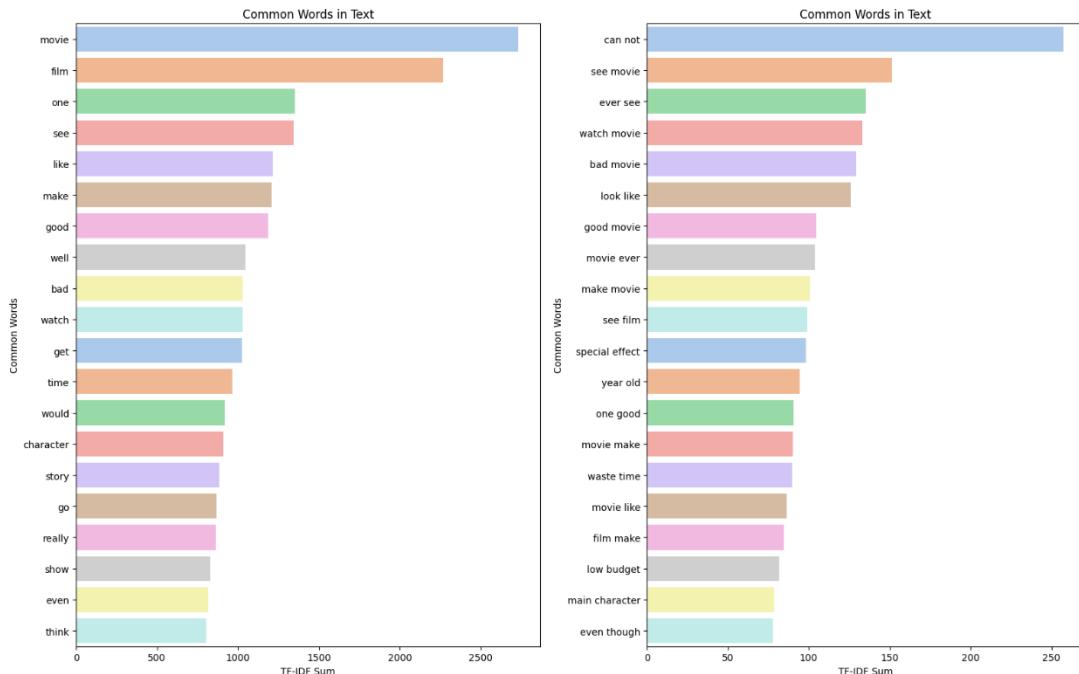
ซึ่งจะทดลองใช้ TF-IDF แบบที่ใช้พารามิเตอร์ตั้งต้น (Default) และ TF-IDF แบบที่ปรับพารามิเตอร์ เพื่อเปรียบเทียบประสิทธิภาพ โดยพารามิเตอร์คือ

```
tfidf_vectorizer = TfidfVectorizer(
    ngram_range=(1, 2),
    min_df=5,
    max_df=0.7,
    max_features=5000,
    stop_words='english'
)
tfidf_vectorizer.fit(x_train)
```

ภาพที่ 3.2.10 รูปแสดงตัวอย่างโค้ดการใช้งานโมดูล TfidfVectorizer และค่าพารามิเตอร์ที่ใช้ในการทดลอง

- Ngram\_range = (1,2) คือ การพิจารณากลุ่มคำทั้งจาก Unigram หรือคำเดียว ๆ (เช่น “good”, “movie”) และ Bigram พิจารณาลำดับของสองคำที่ติดกัน (เช่น “good movie”, “very bad”) เพื่อช่วยให้โมเดลสามารถหาความสัมพันธ์ของคำที่อยู่ติดกัน 2 คำได้

- Min\_df = 5 คือ พารามิเตอร์ที่ใช้กรองคำซึ่ง คำ หรือ n-gram จะถูกเก็บไว้เป็นฟีเจอร์ก็ต่อเมื่อปรากฏในข้อความอย่างน้อย 5 ข้อความที่แตกต่างกัน
- Max\_df = 0.7 คือ พารามิเตอร์ที่ใช้ในการกรองคำโดยจะถูกพิจารณาเป็นฟีเจอร์ก็ต่อเมื่อ คำ หรือ n-gram ปรากฏในข้อความไม่เกิน 70 เปอร์เซ็นต์ของข้อความทั้งหมด
- Max\_feature = 5000 คือ TfidfVectorizer จะเลือกเพียง 5,000 ฟีเจอร์ที่มีค่า TF-IDF สูงสุด เพื่อช่วยลดมิติของข้อมูล
- Stop\_word = ‘English’ คือ การกรองคำ Stop words ที่พบบ่อยในภาษาอังกฤษออก



ภาพที่ 3.2.11 รูปแสดงตัวอย่าง Unigram และ Bigram ที่พบบ่อย

### 3.2.2.3 การแบ่งข้อมูลเพื่อการฝึกสอนและทดสอบ (Train Test Split)

หลังจากการทำ TF-IDF เพื่อเตรียมความพร้อมของข้อมูลให้อยู่ในรูปที่คอมพิวเตอร์สามารถคำนวณผลได้แล้ว ขั้นตอนต่อไปคือการแบ่งข้อมูลออกเป็น ข้อมูลฝึกสอน (Training set) คือข้อมูลที่ถูกใช้เพื่อฝึกสอนแบบจำลอง และข้อมูลทดสอบ (Test set) คือข้อมูลที่ใช้เพื่อทดสอบประสิทธิภาพของแบบจำลอง โดยแบ่งสัดส่วนเป็นชุดข้อมูล

ฝึกสอน 80 เปอร์เซ็นต์ หรือ 40,000 ข้อความ และ ชุดข้อมูลทดสอบ 20 เปอร์เซ็นต์ หรือ 10,000 ข้อความ โดยมีสัดส่วนของความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบที่เท่ากัน

### 3.2.3 การฝึกสอนและการทดสอบแบบจำลอง (Training and Testing Model)

#### 3.2.3.1 อัลกอริทึมการเรียนรู้ของเครื่อง

เป็นการนำชุดข้อมูลสำหรับการฝึกสอนที่ผ่านการเตรียมข้อมูล และการแทนค่าด้วยเวกเตอร์ TF-IDF แล้ว มาใช้ฝึกสอนอัลกอริทึมการเรียนรู้ของเครื่องแบบมิผู้สอน เพื่อสร้างแบบจำลอง ซึ่งจะเป็นการทดลองเพื่อเปรียบเทียบประสิทธิภาพของ 2 แบบจำลอง ได้แก่ แบบจำลองการถดถอยโลจิสติกส์ และ ชัพพอร์ตเวกเตอร์แมชชีน

#### 3.2.3.2 แบบจำลองความรู้สึก (Sentiment Model)

เป็นการสร้างแบบจำลองจากอัลกอริทึม 2 ตัว ได้แก่ การวิเคราะห์การถดถอยโลจิสติก และชัพพอร์ตเวกเตอร์แมชชีน ซึ่งจะนำทั้ง 2 แบบจำลองมาใช้กับ TF-IDF 2 แบบ คือ TF-IDF ที่ยังไม่ได้ทำการปรับพารามิเตอร์ และ TF-IDF ที่ปรับค่าพารามิเตอร์แล้ว เพื่อเปรียบเทียบว่าแบบใดมีประสิทธิภาพในการจำแนกที่ดีกว่ากัน

#### 3.2.3.3 การปรับไsex เบอร์พารามิเตอร์ของแบบจำลอง (Tune Model Hyperparameters)

เป็นการปรับพารามิเตอร์ให้เหมาะสมกับแบบจำลองเพื่อให้แบบจำลองแต่ละตัวสามารถทำนายผลได้อย่างมีประสิทธิภาพมากที่สุดผ่านการทำ GridSearchCV โดยได้ค่าพารามิเตอร์สำหรับแต่ละแบบจำลอง ดังนี้

- พารามิเตอร์ที่ใช้ในการปรับประสิทธิภาพของ LogisticRegression

```
param_grid_logreg = {
    'C': [0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear'],
    'class_weight': [None, 'balanced'],
    'max_iter': [100, 200, 300],
```

```
'tol': [1e-4, 1e-3]}
```

- พารามิเตอร์ที่ใช้สำหรับแบบจำลอง LinearSVC

```
param_grid_svm = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'loss': ['hinge', 'squared_hinge'],
    'dual': [True, False],
    'class_weight': [None, 'balanced'],
    'max_iter': [1000, 2000, 3000],
    'tol': [1e-4, 1e-3, 1e-2],
    'random_state': [42]}
```

### 3.2.3.4 การทำนายป้ายกำกับ (Predict Label)

เป็นการนำชุดข้อมูลสำหรับการทดสอบที่เลือกมาเฉพาะข้อมูลความคิดเห็น (ปิดป้ายกำกับไว้) มาทำการทดสอบกับแบบจำลองทั้งสองเพื่อสังเกตการทำนายผลว่าการแยกประเภทของหมวดหมู่ตรงกับที่ป้ายกำกับที่ปิดไว้หรือไม่

## 3.2.4 การประเมินผล (Evaluation)

### 3.2.4.1 เมทริกซ์/confusion matrix และค่าประสิทธิภาพของแบบจำลอง

เป็นขั้นตอนที่จะนำผลจากการทำนายป้ายกำกับเข้าสู่กระบวนการประเมินประสิทธิภาพของแบบจำลองโดยใช้มาตรฐานต่าง ๆ เช่น ค่าความถูกต้อง (Accuracy) ดังสมการที่ 1 ค่าความแม่นยำ (Precision) ดังสมการที่ 2 ค่าความระลึก (Recall) ดังสมการที่ 3 และค่าเฉลี่ยแบบต่อหน้าหลังของ Precision และ Recall (F-1 score) ดังสมการที่ 4

		Prediction	
		yes	no
Actual	yes	TP	FN
	no	FP	TN

ภาพที่ 3.2.12 รูปแสดงตัวอย่าง Confusion Matrix

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$Precision = \frac{TP}{(TP+FP)} \quad (2)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (3)$$

$$F1 score = 2 \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (4)$$

TP คือ โมเดลทำนายเป็นความคิดเห็นเชิงบวกและข้อความนั้นเป็นความคิดเห็นเชิงบวก

TN คือ โมเดลทำนายเป็นความคิดเห็นเชิงลบและข้อความนั้นเป็นความคิดเห็นเชิงลบ

FP คือ โมเดลทำนายเป็นความคิดเห็นเชิงบวกแต่ข้อความนั้นเป็นความคิดเห็นเชิงลบ

FN คือ โมเดลทำนายเป็นความคิดเห็นเชิงลบแต่ข้อความนั้นเป็นความคิดเห็นเชิงบวก

### 3.2.4.2 ผลลัพธ์ (Result)

ตรวจสอบความถูกต้องของข้อมูลและค่าที่ได้จากการประเมินประสิทธิภาพของแบบจำลองว่าแบบจำลองทำงานได้อย่างมีประสิทธิภาพหรือไม่ หรือค่าจากประเมินอยู่ในเกณฑ์ที่สามารถรับได้หรือไม่ ด้วยค่าที่ได้จากการวัดต่างๆ เช่น ตารางคอนฟิวชัน ค่าความถูกต้อง ค่าความแม่นยำ และอื่นๆ

### 3.2.4.3 การประยุกต์ (Apply)

การประยุกต์ใช้แบบจำลองที่ได้โดยนำแบบจำลองไปใช้กับแอปพลิเคชันบนเว็บเพื่อให้ผู้ใช้งานสามารถวิเคราะห์ความรู้สึกจากภาพนั้นตัวการคิร์ช้อปภาพนั้นที่ต้องการ

## 3.3 ประเด็นที่น่าสนใจและท้าทาย

1. การทำความสะอาดข้อความก่อนจะนำมาทำการสอนแบบจำลองเนื่องจากในภาษาอังกฤษ มีการใช้ไวยากรณ์ที่ไม่เหมือนกันในคำศัพท์คำเดียวกันทำให้หากไม่ทำการทำความสะอาด หรือทำการทำความสะอาดที่ไม่มีประสิทธิภาพ แบบจำลองอาจทำนายคำที่มีความหมาย เหมือนกันเป็นต่างกัน
2. ภาษาที่นำมาใช้เป็นภาษาที่ไม่เป็นทางการมีการใช้คำแสง คำย่อ อิโมจิ คำไม่สุภาพ คำประชดประชัน

3. การดึงข้อมูลจากแพลตฟอร์มออนไลน์มาทำการวิเคราะห์เบื้องจากต้องทำการเชื่อมต่อ API มีแพลตฟอร์มที่นำเสนอจำนวนมากหลายแพลตฟอร์ม แต่มักมีข้อจำกัดในการเชื่อมต่อผ่าน API หรือ บางตัวไม่มี API ที่ให้บริการในการดึงความคิดเห็นมาวิเคราะห์

#### 3.4 ผลลัพธ์ที่คาดหวัง

สามารถค้นหาแบบจำลองที่มีประสิทธิภาพดีที่สุดมาใช้ในการทำนายความรู้สึกของความคิดเห็นต่อภาพพยนตร์ได้ และสามารถแยกคำที่พบบ่อยในรูปแบบ Word Cloud โดยแยกออกเป็นคำที่พบบ่อยในความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบได้

บทที่ 4

#### 4.1 การจัดเตรียมhaar์ดแวร์และซอฟต์แวร์

#### 4.1.1 ฮาร์ดแวร์ที่ใช้ในการพัฒนา

- CPU : AMD Ryzen 7 5800H with Radeon Graphics
  - GPU : NVIDIA GeForce RTX 3050
  - Memory : 16 GB
  - Storage : ssd 512 GB
  - OS : window 11

#### 4.1.2 ซอฟต์แวร์ที่ใช้ในการพัฒนา

- Complier: Python 3.12.5
  - Editor: Jupyter notebook, Visual Studio Code
  - Language: Python, Java, HTML,CSS

#### 4.2 ระยะเวลาในการดำเนินงาน

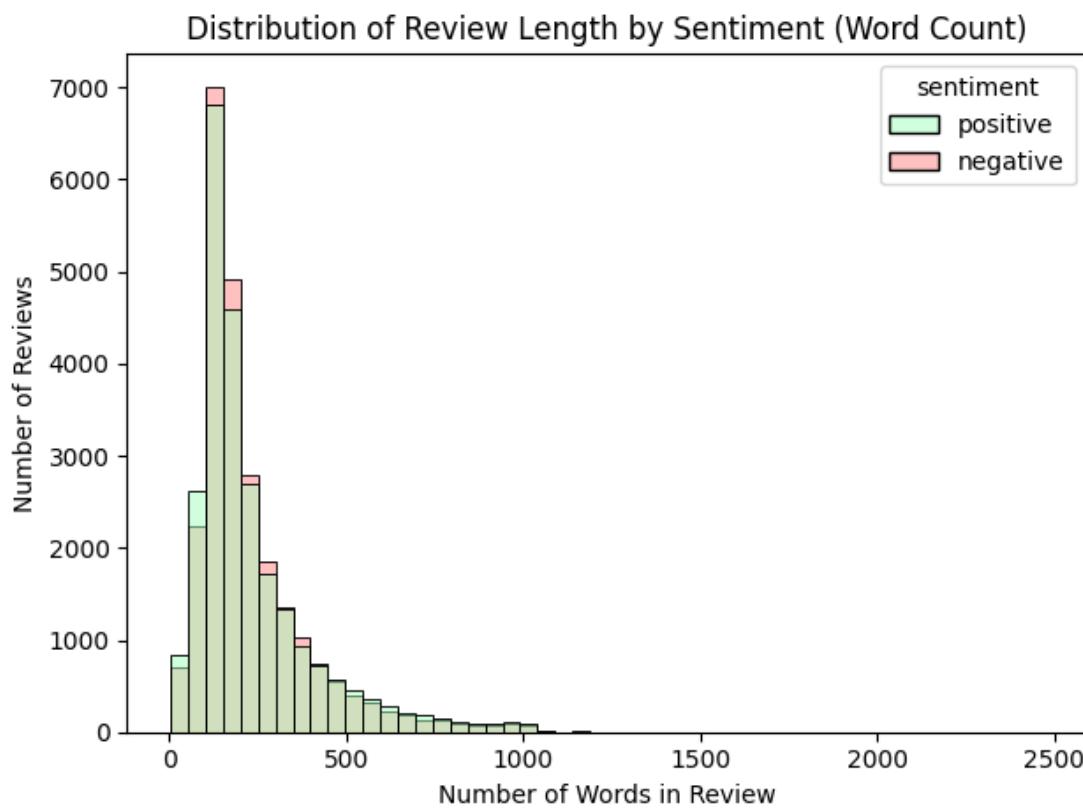
ตารางที่ 4.2.1 การดำเนินงานที่ผ่านมาจนถึงปัจจุบัน



### 4.3 ผลการดำเนินงาน

#### 4.3.1 การวิเคราะห์ชุดข้อมูล

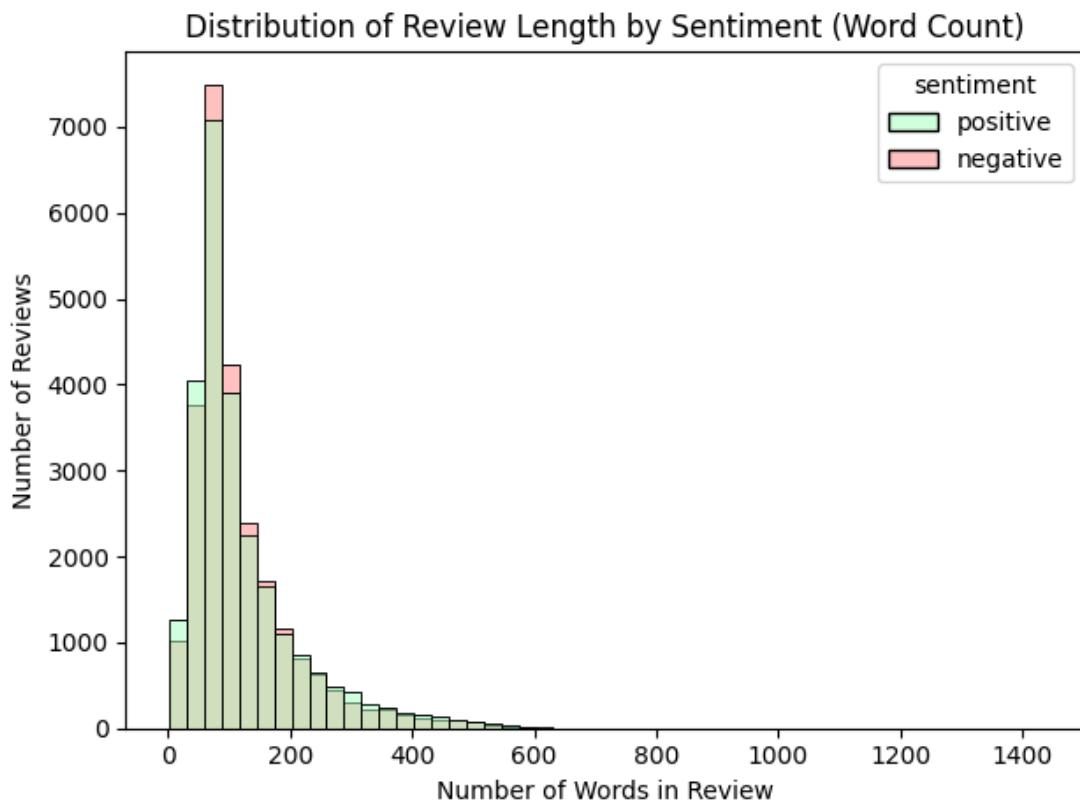
ชุดข้อมูลจากการรวบรวมข้อความคิดเห็นต่อภาพนัตรจำนวน 50,000 ความคิดเห็น ชุดข้อมูลถูกจำแนกออกเป็น 2 ประเภท คือ ความรู้สึกเชิงบวกและความรู้สึกเชิงลบ ซึ่งถูกแบ่งออกเป็นความรู้สึกเชิงบวกจำนวน 25,000 ความคิดเห็น และความรู้สึกเชิงลบจำนวน 25,000 ความคิดเห็น ชุดข้อมูลความคิดเห็นเป็นข้อความภาษาอังกฤษที่มีความยาวเฉลี่ยประมาณ 1,309 ตัวอักษร หรือ 231 คำ



ภาพที่ 4.3.1 กราฟเปรียบเทียบความยาวความของความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ

จากราฟแสดงการกระจายตัวของจำนวนคำในแต่ละข้อความคิดเห็นจำแนกตามประเภทของความรู้สึก พบร่วมกันของทั้งสองประเภทมีลักษณะที่คล้ายคลึงกันโดยมีแนวโน้มกระจุกตัวอยู่ในช่วงความยาว 100 - 200 คำ และส่วนใหญ่จะอยู่ในช่วงความยาว 0 - 500 คำ ซึ่งหมายความว่าความคิดเห็นส่วนใหญ่นั้นมีความยาวไม่มากนัก จำนวนความคิดเห็นเชิงบวกที่มีความยาวอยู่ในช่วง 0 - 150 คำ และมากกว่า 400 คำ มีจำนวนมากกว่าความคิดเห็นที่เป็นความรู้สึกเชิงลบ ซึ่งอาจต้องสมมติฐานได้ว่าคนมักจะแสดงความคิดเห็นด้วยข้อความสั้นๆ หรือไม่ก็แสดงความ

คิดเห็นด้วยข้อความยาวๆไปเลย ค่าความยาวเฉลี่ยของข้อความเชิงบวกมีความยาวประมาณ 233 คำ และความคิดเห็นเชิงลบมีความยาวประมาณ 229 คำ โดยหลังจากทำความสะอาดข้อความแล้วขนาดความยาวเฉลี่ยของข้อความทั้งหมดมีขนาดสั้นลงเหลือเพียงประมาณ 120 คำ ซึ่งขนาดความยาวเฉลี่ยของความคิดเห็นเชิงบวกมีความยาวประมาณ 122 คำ ความคิดเห็นเชิงลบมีความยาวเฉลี่ยประมาณ 119 คำ ถึงแม้จะทำการทำความสะอาดข้อมูลแต่สัดส่วนความยาวของความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบแบบจำเพาะมีความแตกต่างจากเดิมสังเกตได้จากภาพที่ 4.3.2 ความคิดเห็นเชิงบวกยังมีจำนวนมากกว่าความคิดเห็นเชิงลบในช่วงความคิดเห็นที่มีความยาวในช่วง 0 – 150 คำ และมากกว่า 400 คำ อุญจี



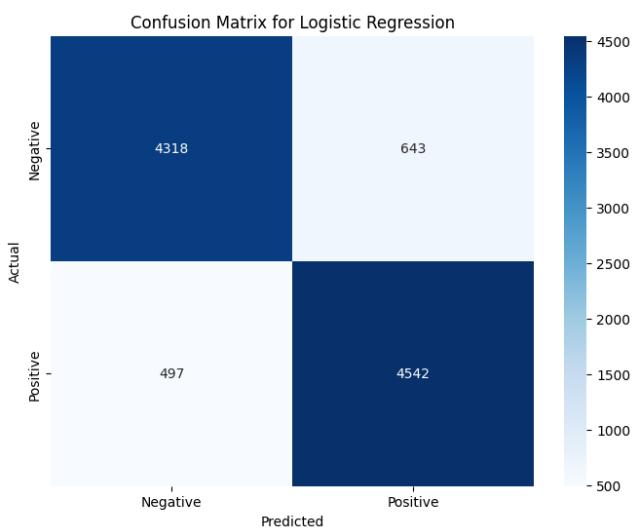
ภาพที่ 4.3.2 กราฟเปรียบเทียบความยาวความของคิดเห็นเชิงบวกและความคิดเห็นเชิงลบหลังจากทำความสะอาดข้อมูล

### 4.3.2 การเปรียบเทียบประสิทธิภาพ

เนื่องจากผู้จัดทำโครงการต้องการให้แบบจำลองสามารถทำนายความรู้สึกของข้อความคิดเห็นต่อภาพนั้นได้อย่างแม่นยำมากที่สุด ผู้จัดทำจึงได้สร้างโมเดลมาเพื่อเปรียบเทียบประสิทธิภาพความแม่นยำของแบบจำลอง โดยแต่ละโมเดลมีการใช้อัลกอริทึมและค่าพารามิเตอร์ที่แตกต่างกัน ซึ่งอัลกอริทึมที่ใช้ในการเปรียบเทียบประสิทธิภาพได้แก่ อัลกอริทึมการถดถอยโลจิสติกส์ และ ชัพพอร์ตเวกเตอร์แมชีน ซึ่งทั้งสองอัลกอริทึมจะมีการปรับพารามิเตอร์เพื่อหาค่าพารามิเตอร์ที่ดีที่สุด เพื่อให้ได้ค่าความถูกต้อง, ค่าความแม่นยำ, ค่าความอ่อนไหว และค่า F1-Score ที่ดีที่สุด จากชุดข้อมูลฝึกสอนทั้งหมด 40,000 ข้อความ และชุดข้อมูลทดสอบทั้งหมด 10,000 ข้อความ ซึ่งมีอัตราส่วนของความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบที่เท่ากันทั้งในชุดข้อมูลฝึกสอนและชุดข้อมูลทดสอบ

#### 4.3.2.1 การจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์

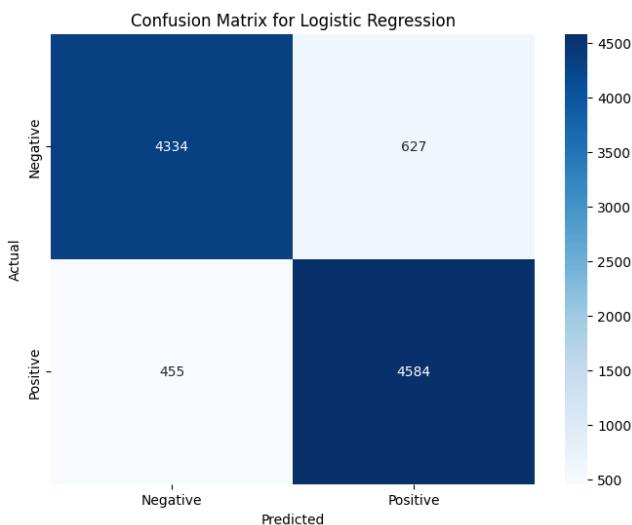
การจำแนกความรู้สึกของความคิดเห็นด้วยอัลกอริทึมการถดถอยโลจิสติกส์ที่ใช้พารามิเตอร์ตั้งต้นทดลองกับการทำ TfIdfVectorizer ที่ยังไม่ได้ทำการปรับค่ารามิเตอร์ และ TfIdfVectorizer ที่ปรับพารามิเตอร์แล้วก่อน เพื่อวัดหลังจากปรับพารามิเตอร์ TfIdfVectorizer และแบบจำลองมีประสิทธิภาพที่ดีมากขึ้นหรือไม่



ภาพที่ 4.3.3 คุณพิวชันเมทริกซ์ของการถดถอยโลจิสติกส์กับ TfIdfVectorizer ที่ยังไม่ได้ทำการปรับพารามิเตอร์

จากค่อนพิวชันเมทริกซ์ของอัลกอริทึมการทดสอบโดยโลจิสติกส์ที่ใช้พารามิเตอร์  
ตั้งต้นกับ TfidfVectorizer ที่ยังไม่ทำการปรับพารามิเตอร์

- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถ  
ทำนายได้ถูก 4,318 ข้อความ และทำนายผิด 643 ข้อความ
- **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถ  
ทำนายได้ถูก 4,542 ข้อความ และทำนายผิด 497 ข้อความ



ภาพที่ 4.3.4 ค่อนพิวชันเมทริกซ์ของการทดสอบโดยโลจิสติกส์กับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้ว

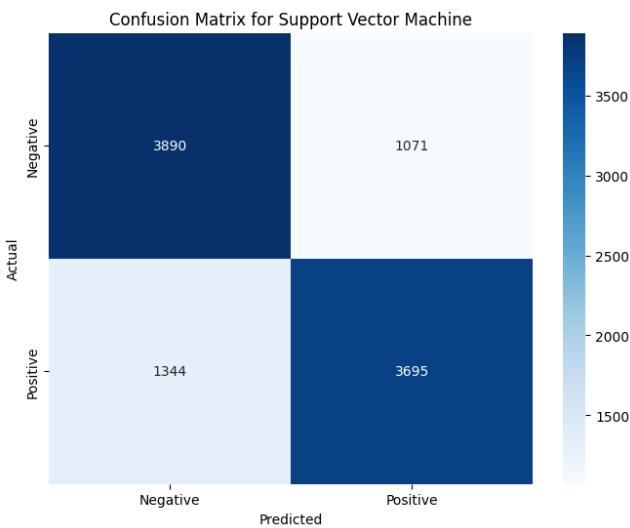
จากค่อนพิวชันเมทริกซ์ของอัลกอริทึมการทดสอบโดยโลจิสติกส์ที่ใช้พารามิเตอร์  
ตั้งต้นกับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้ว

- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถ  
ทำนายได้ถูก 4,334 ข้อความ และทำนายผิด 627 ข้อความ
- **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถ  
ทำนายได้ถูก 4,584 ข้อความ และทำนายผิด 455 ข้อความ

จากค่อนพิวชันเมทริกซ์ในภาพที่ 4.3.3 และ 4.3.4 จะเห็นได้ว่าเมื่อทำการ  
ปรับพารามิเตอร์ของ TfidfVectorizer แล้วแบบจำลองมีประสิทธิภาพในการทำนายถูกต้อง<sup>เพิ่มขึ้น</sup>จากก่อนปรับพารามิเตอร์ทำนายถูกทั้งหมด 8,860 ข้อความ เป็น 8,918 ซึ่งทำนาย  
ถูกเพิ่มขึ้น 58 ข้อความ หรือทำนายถูกเพิ่มขึ้น 0.655 เปอร์เซ็นต์

#### 4.3.2.2 การจำแนกความรู้สึกของข้อความด้วยชั้พพอร์ตเวกเตอร์แมชชีน

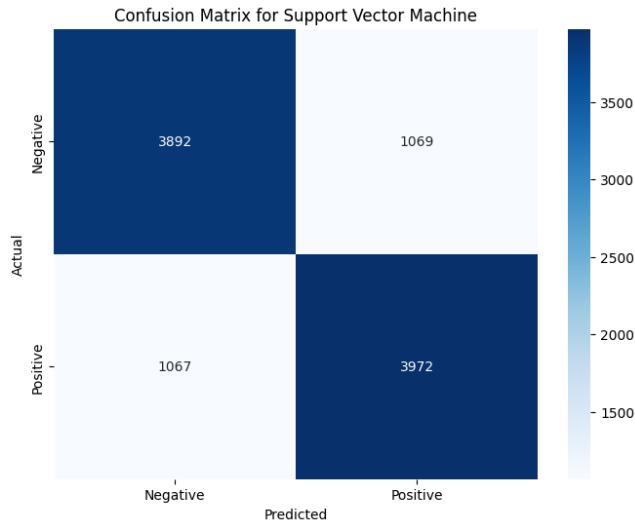
การจำแนกความรู้สึกของความคิดเห็นด้วยอัลกอริทึมชั้พพอร์ตเวกเตอร์แมชชีนที่ใช้พารามิเตอร์ตั้งต้นทดลองกับการทำ TF-IDF ที่ยังไม่ได้ทำการปรับค่ารามิเตอร์ และ TF-IDF ที่ปรับพารามิเตอร์แล้วก่อน เพื่อดูว่าหลังจากปรับพารามิเตอร์ TF-IDF แล้ว แบบจำลองมีประสิทธิภาพที่ดีมากขึ้นหรือไม่



ภาพที่ 4.3.5 คอนฟิวชันเมทริกซ์ของชัพพอร์ตเวกเตอร์แมชชีนกับ TfidfVectorizer ที่ยังไม่ได้ทำการปรับพารามิเตอร์

จากคอนฟิวชันเมทริกซ์ของอัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีนที่ใช้พารามิเตอร์ตั้งต้นกับ TF-IDF ที่ยังไม่ทำการปรับพารามิเตอร์

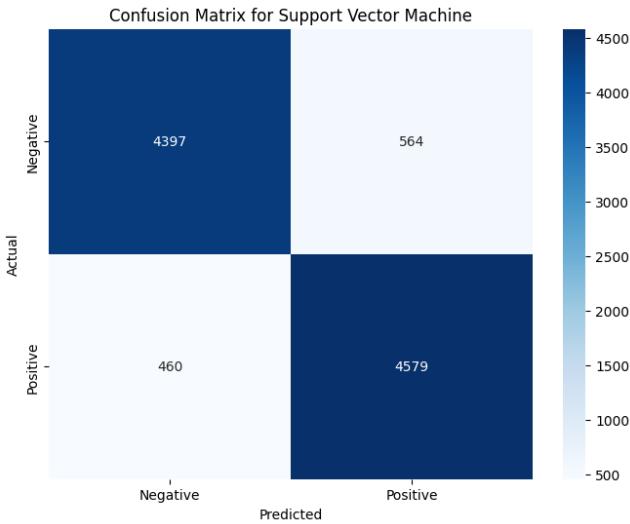
- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 3,890 ข้อความ และทำนายผิด 1,071 ข้อความ
- **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 3,695 ข้อความ และทำนายผิด 1,344 ข้อความ



ภาพที่ 4.3.6 /confusion matrix ของชั้นพารามิเตอร์ที่ปรับค่าพารามิเตอร์แล้ว

- จาก/confusion matrix ของอัลกอริทึมชั้นพารามิเตอร์ที่ใช้พารามิเตอร์ตั้งต้นกับ TfidfVectorizer ที่ทำการปรับค่าพารามิเตอร์แล้ว
- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 3,892 ข้อความ และทำนายผิด 1,069 ข้อความ
  - **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 3,972 ข้อความ และทำนายผิด 1,067 ข้อความ

จาก/confusion matrix ในภาพที่ 4.3.5 และ 4.3.6 พบว่าเมื่อทำการปรับพารามิเตอร์ของ TfidfVectorizer และ พบร่วมแบบจำลองมีความสามารถในการทำนายข้อมูลได้ถูกต้องมากขึ้นจากการปรับพารามิเตอร์ทำนายถูกต้อง 7,585 ข้อความ หลังจากปรับพารามิเตอร์ TfidfVectorizer ทำนายถูกต้อง 7,864 ข้อความ แบบจำลองสามารถทำนายได้ถูกต้องมากขึ้นถึง 279 ข้อความ หรือทำนายถูกเพิ่มขึ้น 3.68 เปอร์เซ็นต์



ภาพที่ 4.3.7 /confusion matrix ของ LinearSVC กับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้ว

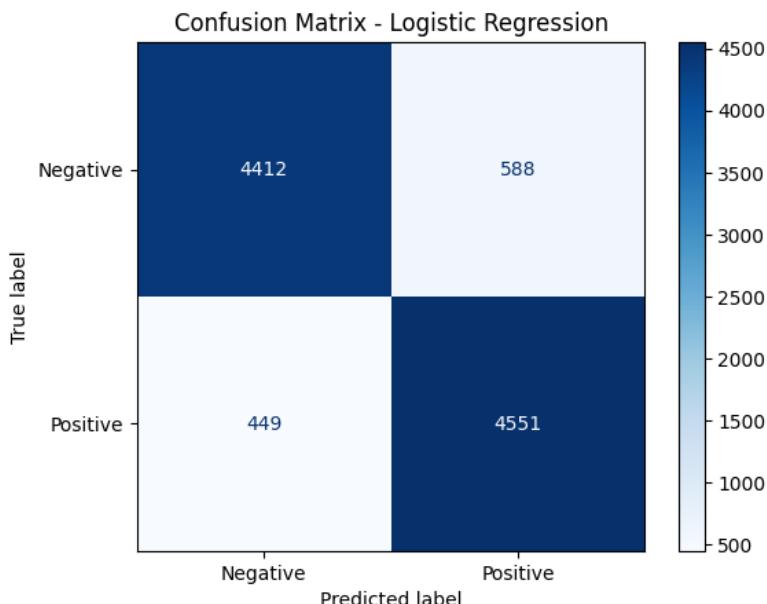
- จาก/confusion matrix ของอัลกอริทึมชั้นพ้อร์ตเวกเตอร์แมชชีนที่ใช้ LinearSVC กับ TfidfVectorizer ที่ทำการปรับค่าพารามิเตอร์แล้ว
- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,397 ข้อความ และทำนายผิด 564 ข้อความ
  - **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,579 ข้อความ และทำนายผิด 460 ข้อความ

จาก/confusion matrix ในภาพที่ 4.3.6 และ 4.3.7 พบร่วมกับเครื่องเรียนแบบจำลองชั้นพ้อร์ตเวกเตอร์ที่ใช้ SVC กับ LinearSVC แบบจำลองที่ใช้ LinearSVC มีประสิทธิภาพในการทำนายดีกว่าแบบจำลองที่ใช้ SVC ธรรมด้า ซึ่งทำได้ข้อความได้ถูกต้องเพียง 7,864 ข้อความ เมื่อเทียบกับแบบจำลองที่ใช้ LinearSVC ที่ทำนายได้ถูกต้อง 8,976 ข้อความ แบบจำลอง LinearSVC สามารถทำนายได้ถูกต้องมากขึ้นถึง 1,112 ข้อความ หรือเพิ่มขึ้น 14.14 เปอร์เซ็นต์ จากแบบจำลองที่ใช้ SVC ธรรมด้า

#### 4.3.2.3 ประสิทธิภาพของแบบจำลองหลังจากการปรับพารามิเตอร์

##### (1) การหาพารามิเตอร์ที่ดีที่สุดของการถดถอยโลจิสติกส์

นำแบบจำลองการถดถอยโลจิสติกส์ที่ใช้ TfidfVectorizer ที่มีการปรับค่าพารามิเตอร์แล้วมาทำการหาค่าพารามิเตอร์ที่ดีที่สุดด้วยการทำ GridSearchCV ทั้งหมด 5 Fold ซึ่งได้ค่าพารามิเตอร์ที่ดีที่สุด คือ {'C': 10, 'class\_weight': None, 'max\_iter': 100, 'penalty': 'l2', 'solver': 'liblinear', 'tol': 0.0001} ซึ่งได้ค่าอนพิวชันเมทริกซ์ ดังนี้



ภาพที่ 4.3.8 ค่าอนพิวชันเมทริกซ์ของการถดถอยโลจิสติกส์หลังจากการทำ GridSearchCV

จากค่าอนพิวชันเมทริกซ์ของการถดถอยโลจิสติกส์หลังจากการปรับค่าพารามิเตอร์ จะได้ข้อสรุปว่า

- **ความคิดเห็นเชิงลบ** : ใช้ข้อมูลทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,412 ข้อความ และทำนายผิด 588 ข้อความ
- **ความคิดเห็นเชิงบวก** : ใช้ข้อมูลทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,551 ข้อความ และทำนายผิด 449 ข้อความ

รวมทั้งสองประเภทของความคิดเห็นที่ใช้ข้อมูลในการทดสอบทั้งหมดจำนวน 10,000 ข้อความ ทำนายถูก 8,963 ข้อความ ดังนั้นคิดเป็นเปอร์เซ็นต์ความถูกต้องคือ 89.63 เปอร์เซ็นต์ ซึ่งจากการสังเกตค่าอนพิวชันเมทริกซ์ไม่เคลื่อนไหวสามารถในการทำนายข้อความที่เป็นเชิงบวกได้ดีกว่าข้อความที่เป็นเชิงลบ

ตารางที่ 4.3.1 ประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยการถดถอยโลจิสติกส์หลังการปรับพารามิเตอร์

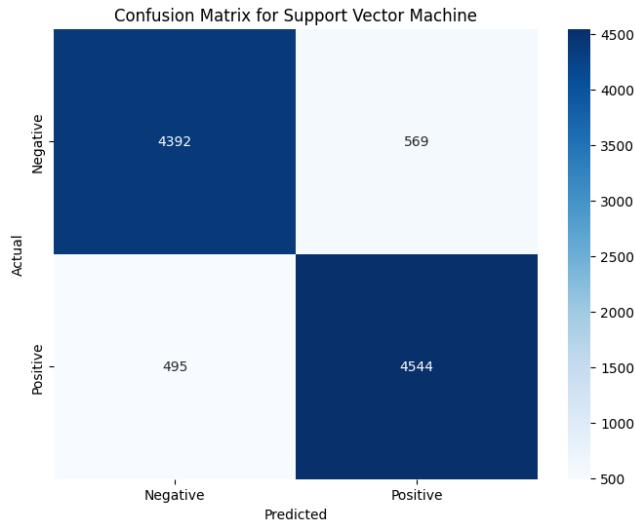
กลุ่มของข้อความ	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าอ่อนไหว	F1-Score
ลบ	0.90	0.91	0.88	0.89
บวก		0.89	0.91	0.90

จากตารางที่ 4.3.2 พบร่วมกันว่าการจำแนกความคิดเห็นด้วยการถดถอยโลจิสติกส์ให้ค่าความถูกต้องถึง 0.90 มีค่าความแม่นยำในกลุ่มข้อความที่เป็นเชิงลบ คือ 0.91 และความแม่นยำในกลุ่มเชิงบวก คือ 0.89 ซึ่งมีค่าความแม่นยำในการจำแนกไม่ต่างกันมาก ในส่วนของค่าอ่อนไหวในกลุ่มข้อความที่เป็นเชิงบวกมีค่าอ่อนไหวสูงกว่าอยู่ที่ 0.91 ซึ่งไม่ได้ต่างกับค่าอ่อนไหวในกลุ่มความคิดเห็นเชิงลบมากนัก โดยค่า F1-score ของกลุ่มความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบมีค่าที่ใกล้กันคือ 0.90 และ 0.89 ตามลำดับ ซึ่งหมายความว่าไม่เดลิมีความสามารถทำนายข้อมูลของหัวข้อส่องกลุ่มข้อมูลใกล้เคียงกัน

## (2) การหาพารามิเตอร์ที่ดีที่สุดของชั้นพัฒนาแบบตัวอย่างแมชชีน

นำแบบจำลองชั้นพัฒนาแบบตัวอย่างแมชชีนที่ใช้กับ TfidfVectorizer ที่ปรับค่าพารามิเตอร์แล้วมาทำการหาค่าพารามิเตอร์ที่ดีที่สุดด้วยการทำ RandomSearchCV สำหรับ SVC และ GridSearchCV สำหรับ LinearSVC

แบบจำลองชั้นพัฒนาแบบตัวอย่างแมชชีนที่ทำการหาค่าพารามิเตอร์ด้วย RandomSearchCV โดยการสุ่มข้อความคิดเห็นออกมา 20 เบอร์เซ็นต์ หรือ 7,000 ข้อความ จากชุดข้อมูลทดสอบจำนวน 17,500 ข้อความ สำหรับโมเดลที่ใช้ SVC ได้พารามิเตอร์ที่ดีที่สุด คือ {'C': 2, 'kernel': 'linear', 'shrinking': True, 'gamma': 1.0, 'degree': 2, 'class\_weight': 'balanced} และเมื่อเอาค่าพารามิเตอร์ที่ดีที่สุดมาใช้กับชุดข้อมูลทดสอบได้ค่อนพิวนั้นเมทริกซ์ดังนี้

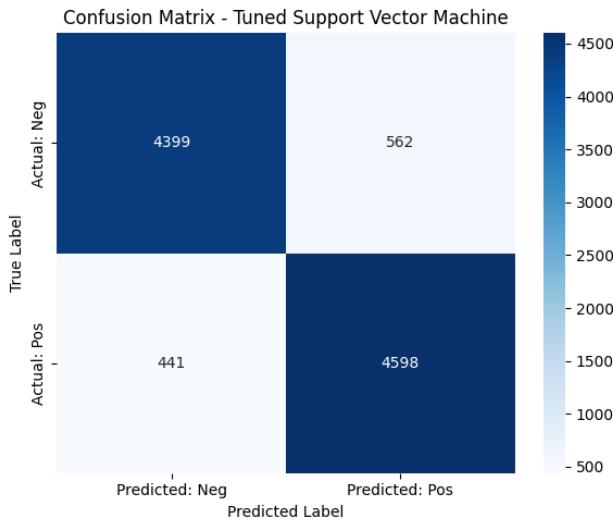


ภาพที่ 4.3.9 คอนฟิวชันเมทริกซ์ของชัพพอร์ตเวกเตอร์แมชีนหลังจากการทำ RandomSearchCV

- จากคอนฟิวชันเมทริกซ์ของชัพพอร์ตเวกเตอร์แมชีนหลังจากการปรับค่าพารามิเตอร์ด้วย RamdomSearchCV จะได้ข้อสรุปว่า
- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,392 ข้อความ และทำนายผิด 569 ข้อความ
  - **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,544 ข้อความ และทำนายผิด 495 ข้อความ

รวมทั้งสองประเภทของความคิดเห็นที่ใช้ข้อมูลในการทดสอบทั้งหมดจำนวน 10,000 ข้อความ ทำนายถูก 8,936 ข้อความ ดังนั้นคิดเป็นเปอร์เซ็นต์ความถูกต้อง คือ 89.36 เปอร์เซ็นต์ ซึ่งจากการสังเกตคอนฟิวชันเมทริกซ์ไม่เดล米ความสามารถในการทำนายข้อความที่เป็นเชิงบวกได้ดีกว่าข้อความที่เป็นเชิงลบ

แบบจำลองชัพพอร์ตเวกเตอร์แมชีนที่ทำการหาค่าพารามิเตอร์ด้วย GridSearchCV สำหรับโมเดลที่ใช้ LinearSVC จำนวน 5 Fold ซึ่งได้พารามิเตอร์ที่ดีที่สุดคือ `{'C': 1, 'class_weight': None, 'dual': True, 'loss': 'squared_hinge', 'max_iter': 1000, 'penalty': 'l2', 'tol': 0.0001}` ซึ่งได้คอนฟิวชันเมทริกซ์ดังนี้



ภาพที่ 4.3.10 คอนฟิวชันเมทริกซ์ของชั้พพร็อกเตอร์แมชชีนหลังจากการทำ GridSearchCV

จากคอนฟิวชันเมทริกซ์ของชั้พพร็อกเตอร์แมชชีนหลังจากการทำการปรับค่าพารามิเตอร์ จะได้ข้อสรุปว่า

- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,399 ข้อความ และทำนายผิด 562 ข้อความ
- **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 5,000 ข้อความ ซึ่งสามารถทำนายได้ถูก 4,598 ข้อความ และทำนายผิด 441 ข้อความ

รวมทั้งสองประเภทของความคิดเห็นที่ใช้ข้อมูลในการทดสอบทั้งหมดจำนวน 10,000 ข้อความ ทำนายถูก 8,997 ข้อความ ดังนั้นคิดเป็นเปอร์เซ็นต์ความถูกต้อง คือ 89.97 เปอร์เซ็นต์ ซึ่งจากการสังเกตคอนฟิวชันเมทริกซ์ไม่เดล米ความสามารถในการทำนายข้อความที่เป็นเชิงบวกได้ดีกว่าข้อความที่เป็นเชิงลบ

ตารางที่ 4.3.2 ประสิทธิภาพการจำแนกความรู้สึกของข้อความด้วยชั้พพอร์ตเวกเตอร์แมชีนหลังการปรับ  
พารามิเตอร์

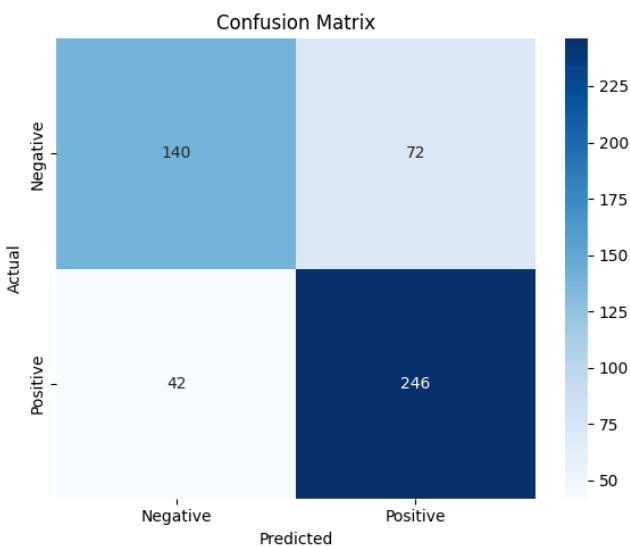
กลุ่มของข้อความ	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าอ่อนไหว	F1-Score
ลบ	0.90	0.91	0.89	0.90
บวก		0.89	0.91	0.90

จากตารางที่ 4.3.2 พบว่าการจำแนกความคิดเห็นด้วยชั้พพอร์ตเวกเตอร์แมชีนให้ค่าความถูกต้องถึง 0.90 มีค่าความแม่นยำในกลุ่มข้อความที่เป็นเชิงลบ คือ 0.91 และความแม่นยำในกลุ่มเชิงบวก คือ 0.89 ซึ่งมีค่าความแม่นยำในการจำแนกไม่ต่างกันมาก ในส่วนของค่าอ่อนไหวในกลุ่มข้อความที่เป็นเชิงบวกมีค่าอ่อนไหวสูงกว่าอยู่ที่ 0.91 ซึ่งไม่ได้ต่างกับค่าอ่อนไหวในกลุ่มความคิดเห็นเชิงลบมากนัก โดยค่า F1-score ของกลุ่มความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบมีค่าที่ใกล้กันคือ 0.9 และ 0.90 ตามลำดับ ซึ่งหมายความว่าโมเดลมีความสามารถทำนายข้อมูลของห้องสองกลุ่มข้อมูลได้ใกล้เคียงกัน

#### 4.3.2.4 การทดสอบประสิทธิภาพของอัลกอริทึม

จากนั้นนำผลลัพธ์ที่ดีที่สุดที่ได้จากการปรับพารามิเตอร์ของแต่ละอัลกอริทึมมาเปรียบเทียบกันเพื่อหาโมเดลที่มีประสิทธิภาพมากที่สุด ซึ่งทั้งสองโมเดลมีค่าความถูกต้อง ค่าความแม่นยำ ค่าระลึกถึง และค่า F1-Score ที่ใกล้เคียงกันมาก ผู้จัดทำจึงนำโมเดลทั้ง 2 ไปทดลองใช้ทำงานยังกับความคิดเห็นของภาพนัตรเรื่อง “Sinners” ที่ดึงมาจากยูทูป ซึ่งเป็นชุดข้อมูลความคิดเห็นทั้งหมด 500 ข้อความ เพื่อเปรียบเทียบประสิทธิภาพความสามารถในการทำงานของข้อมูลที่มาจากการแหล่งข้อมูลคนละที่

##### (1) ประสิทธิภาพของการถดถอยโลจิสติกส์กับความคิดเห็นบนยูทูป



ภาพที่ 4.3.11 ค่อนพิวชันเมทริกซ์ผลการจำแนกความรู้สึกของข้อความจาก YouTube ด้วยการถดถอยโลจิสติกส์ หลัง

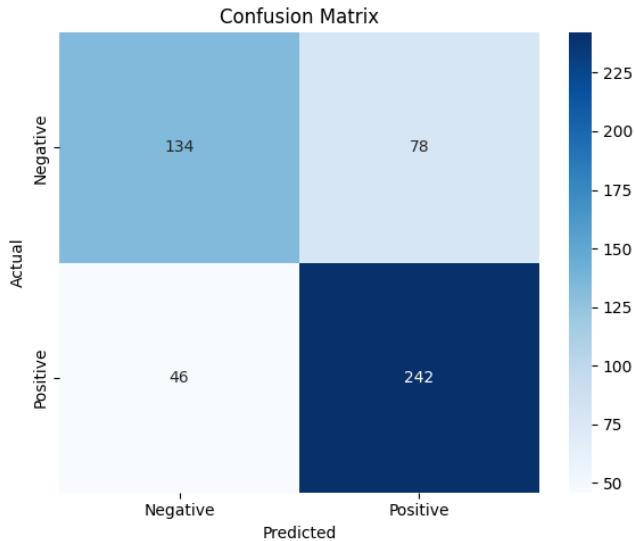
จากค่อนพิวชันเมทริกซ์ของการถดถอยโลจิสติกส์ที่วิเคราะห์ความคิดเห็นบนยูทูป จำนวน 500 ความคิดเห็น

- **ความคิดเห็นเชิงลบ** : ใช้ข้อความทดสอบจำนวน 212 ข้อความ ซึ่งสามารถทำงานได้ถูก 140 ข้อความ และทำงานผิด 72 ข้อความ
- **ความคิดเห็นเชิงบวก** : ใช้ข้อความทดสอบจำนวน 288 ข้อความ ซึ่งสามารถทำงานได้ถูก 246 ข้อความ และทำงานผิด 42 ข้อความ

รวมทั้งสองประเภทของความคิดเห็นที่ใช้ข้อมูลในการทดสอบทั้งหมดจำนวน 500 ข้อความ ทำงานถูก 8,943 ข้อความ ดังนั้นคิดเป็นเปอร์เซ็นต์ความถูกต้อง คือ 75.2 เปอร์เซ็นต์ ซึ่ง

จากการสังเกตค่อนพิวชันเมทริกซ์มีความสามารถในการทำนายข้อความที่เป็นเชิงบวกได้ดีกว่าข้อความที่เป็นเชิงลบ

(2) ประสิทธิภาพของชั้พพร์ตเวกเตอร์แมชชีนกับความคิดเห็นบนบัญชี



ภาพที่ 4.3.12 ค่อนพิวชันเมทริกซ์ผลการจำแนกความรู้สึกของข้อความจาก YouTube ด้วยชัพพร์ตเวกเตอร์แมชชีน  
จากค่อนพิวชันเมทริกซ์ของชัพพร์ตเวกเตอร์แมชชีนที่วิเคราะห์ความคิดเห็นบนบัญชี  
ทุปจำนวน 500 ความคิดเห็น

- ความคิดเห็นเชิงลบ : ใช้ข้อความทดสอบจำนวน 212 ข้อความ ซึ่งสามารถทำนายได้ถูก 134 ข้อความ และทำนายผิด 78 ข้อความ
- ความคิดเห็นเชิงบวก : ใช้ข้อความทดสอบจำนวน 288 ข้อความ ซึ่งสามารถทำนายได้ถูก 242 ข้อความ และทำนายผิด 46 ข้อความ

รวมทั้งสองประเภทของความคิดเห็นที่ใช้ข้อมูลในการทดสอบทั้งหมดจำนวน 500 ข้อความ ทำนายถูก 8,943 ข้อความ ดังนั้นคิดเป็นเปอร์เซ็นต์ความถูกต้อง คือ 77.2 เปอร์เซ็นต์ ซึ่งจากการสังเกตค่อนพิวชันเมทริกซ์มีความสามารถในการทำนายข้อความที่เป็นเชิงบวกได้ดีกว่าข้อความที่เป็นเชิงลบ

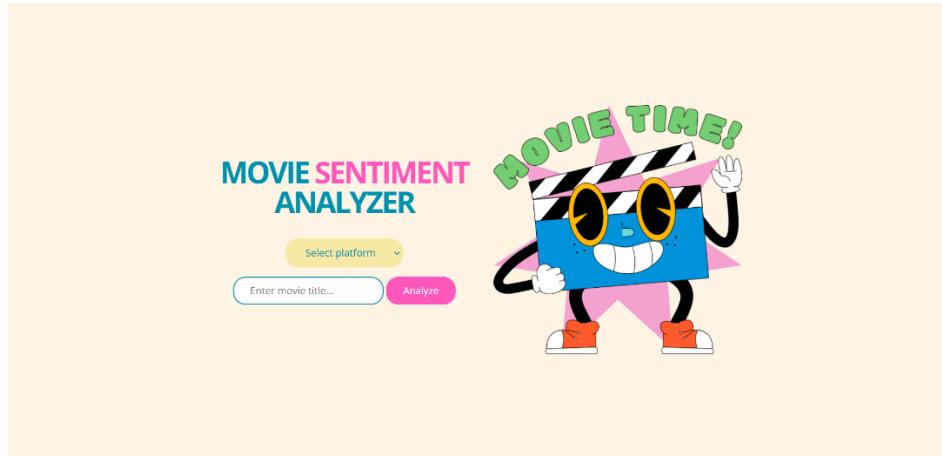
ตารางที่ 4.3.3 ตารางเปรียบเทียบประสิทธิภาพของอัลกอริทึมการตัดถ้อยโลจิสติกส์และชั้พพร์เวกเตอร์แมชีน

	ประเภทข้อความ	Logistic Regression	Support Vector Machine
ค่าความถูกต้อง		<b>0.77</b>	0.75
ค่าความแม่นยำ	ลบ	<b>0.77</b>	0.74
	บวก	<b>0.77</b>	0.76
ค่าอ่อนไหว	ลบ	<b>0.66</b>	0.63
	บวก	<b>0.85</b>	0.84
F1-Score	ลบ	<b>0.71</b>	0.68
	บวก	<b>0.81</b>	0.80

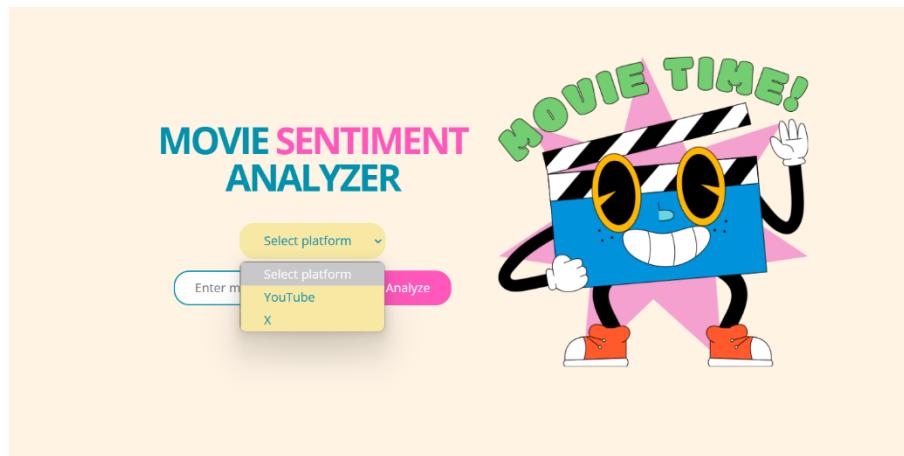
จากการเปรียบเทียบประสิทธิภาพของโมเดลทั้งสองชุดโดยใช้ค่าความถูกต้อง, ค่าความแม่นยำ, ค่าอ่อนไหว และค่า F1-Score พบร่วมกัน โดยแบบจำลองมีค่าเท่า ๆ กัน โดยแบบจำลองการตัดถ้อยโลจิสติกสมีประสิทธิภาพที่สูงกว่าเพียงเล็กน้อย

### 4.3.3 การทดลองใช้งาน

ทดลองใช้งานโมเดลด้วยการดึงข้อมูลความคิดเห็นเกี่ยวกับภาพยนตร์บนแพลตฟอร์มออนไลน์มาวิเคราะห์ โดยสามารถเลือกแพลตฟอร์มที่ต้องการวิเคราะห์ความคิดเห็น (ยูทูป หรือ X) และจากนั้นกรอกชื่อภาพยนตร์ที่ต้องการวิเคราะห์



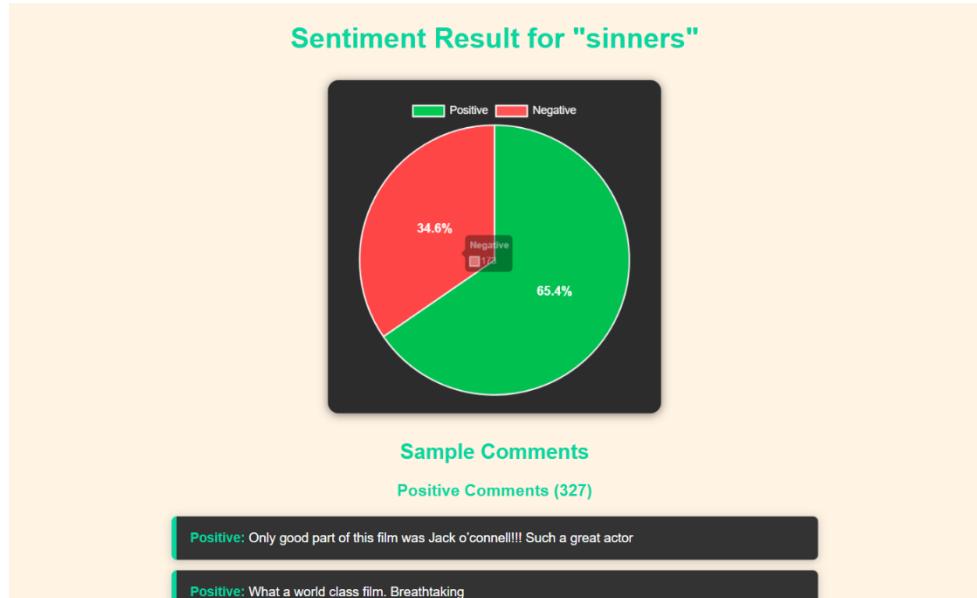
ภาพที่ 4.3.13 ภาพแสดงหน้าหลักของระบบ



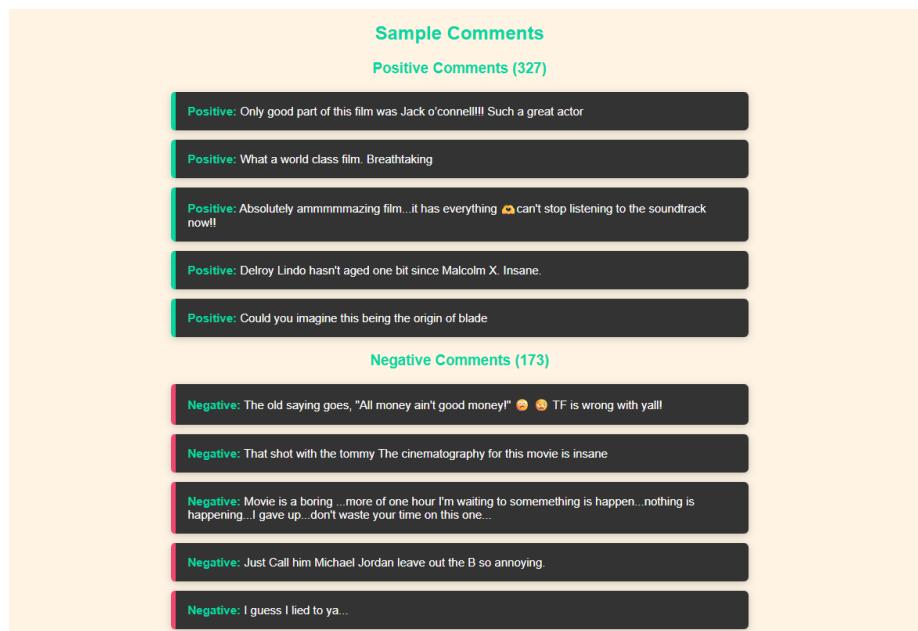
ภาพที่ 4.3.14 ภาพแสดงหน้าหลักของระบบสำหรับการเลือกแพลตฟอร์มในการวิเคราะห์

เนื่องจากโมเดลที่พัฒนาเป็นโมเดลวิเคราะห์ความคิดเห็นภาษาอังกฤษเท่านั้น ดังนั้น คำค้นหาจึงควรเป็นภาษาอังกฤษ โดยจะทำการทดสอบดึงความคิดเห็นของภาพยนตร์ Sinners จากความคิดเห็นบนยูทูป และ X มาใช้ในการทดลองโมเดล ด้วยการนำข้อมูลความคิดเห็นไปวิเคราะห์ด้วยโมเดล ผลปรากฏว่าสามารถวิเคราะห์ความรู้สึกของข้อความได้ โดยสามารถแยกเป็นความคิดเห็น

เชิงบวกและความคิดเห็นเชิงลบ พร้อมทั้งแสดงแผนภูมิวงกลม และ Word Cloud ซึ่งมีผลลัพธ์ดังต่อไปนี้



ภาพที่ 4.3.15 ภาพตัวอย่าง Pie Chart จากการวิเคราะห์ความคิดเห็น



ภาพที่ 4.3.16 ตัวอย่างความคิดเห็นจากยุทปท์โมเดลทำนายป้ายกำกับ



ภาพที่ 4.3.17 ตัวอย่างการทำ Word Could จากความคิดเห็นโดยแยกเป็นความคิดเห็นเชิงบวกและความคิดเห็นเชิงลบ

## บทที่ 5

### สรุป

#### 5.1 สรุปผลดำเนินงาน

##### 5.1.1 การสร้างแบบจำลอง

ในการสร้างแบบจำลองด้วยอัลกอริทึมการถดถอยโลจิสติกส์ และชัพพอร์ตเวกเตอร์แมชชีนในตอนต้นที่ยังไม่ได้ทำการปรับค่าพารามิเตอร์ของแบบจำลองจะสังเกตได้ว่าแบบจำลองที่ใช้อัลกอริทึมการถดถอยโลจิสติกส์สามารถทำนายข้อความที่ถูกปิดป้ายกำกับไว้ได้ถูกต้องมากกว่าแบบจำลองที่ใช้อัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีนอย่างเห็นได้ชัดแบบจำลองการถดถอยโลจิสติกส์ทำนายความคิดเห็นผิดพลาด 1,082 ข้อความ เมื่อเทียบกับชัพพอร์ตเวกเตอร์แมชชีนที่ทำนายผิดพลาดถึง 2,136 ข้อความ ซึ่งทำนายผิดพลาดมากกว่าถึงเกือบสองเท่า จากผลการทดลองนี้อาจหมายความได้ว่าชุดข้อมูลนี้เหมาะสมกับแบบจำลองประเภทเชิงเส้น มากกว่า เนื่องจากพารามิเตอร์ตั้งต้นของ LogisticRegression ใช้ Kenel เป็น ‘linear’ ส่วน SVC ใช้ kernel เป็น ‘rbf’ และเมื่อเปลี่ยนอัลกอริทึมชัพพอร์ตเวกเตอร์แมชชีนที่มา kernel มาเป็น ‘linear’ แบบจำลองมีประสิทธิภาพสูงขึ้นอย่างเห็นได้ชัดโดยแบบจำลองสามารถจำแนกประเภทความรู้สึกของข้อความได้ถูกต้องมากขึ้นถึง 1,112 ข้อความ ดังนั้นจึงเลือกแบบจำลองชัพพอร์ตเวกเตอร์แมชชีนที่ใช้ LinearSVC มาใช้ในการปรับค่าพารามิเตอร์

ซึ่งประสิทธิภาพของทั้งสองแบบจำลองหลังจากการปรับค่าพารามิเตอร์พบว่าทั้งสองแบบจำลองมีค่าประสิทธิภาพต่างกันเพียงเล็กน้อย แบบจำลองการถดถอยโลจิสติกสมีค่าความถูกต้องอยู่ที่ 89.63 เปอร์เซ็นต์ แบบจำลองชัพพอร์ตเวกเตอร์แมชชีนมีความถูกต้อง เท่ากับ 89.43 เปอร์เซ็นต์ ดังนั้นจึงสามารถเลือกแบบจำลองได้ไปใช้งานก็ได้เนื่องจากแบบจำลองมีประสิทธิภาพใกล้เคียงกัน โดยในการสร้างระบบวิเคราะห์ภาษาญี่ปุ่นตัวจากแพลตฟอร์มออนไลน์เลือกแบบจำลองการถดถอยโลจิสติกส์ไปใช้ในการสร้างระบบ

### 5.1.2 ตัวอย่างข้อความที่ทำนายผิดพลาดของแบบจำลองบนชุดข้อมูลทดสอบ

ตัวอย่างข้อความที่แบบจำลองทั้ง 2 แบบจำลองทำนายผิดพลาดบนชุดข้อมูลทดสอบ

The film quickly gets to a major chase scene with ever increasing destruction. The first really bad thing is the guy hijacking Steven Seagal would have been beaten to pulp by Seagal's driving, but that probably would have ended the whole premise for the movie.  
It seems like they decided to make all kinds of changes in the movie plot, so just plan to enjoy the action, and do not expect a coherent plot. Turn any sense of logic you may have, it will reduce your chance of getting a headache.  
I does give me some hope that Steven Seagal is trying to move back towards the type of characters he portrayed in his more popular movies.

ภาพที่ 5.1.1 ตัวอย่างข้อความเชิงลบที่แบบจำลองทั้งสองทำนายผิดพลาดก่อนทำการทดสอบ ภาพที่ 5.1.1 ตัวอย่างข้อความเชิงลบที่แบบจำลองทั้งสองทำนายผิดพลาดหลังทำการทดสอบ

จากภาพที่ 5.1.1 เป็นความคิดเห็นที่ถูกป้ายกำกับไว้ว่าเป็นความคิดเห็นเชิงลบ ซึ่งเป็นข้อความที่ยังไม่ได้ทำการทดสอบข้อความ ภาพที่ 5.1.2 คือข้อความคิดเห็นหลังจากการทำการทดสอบข้อความแล้ว พบร้าข้อความที่ถูกเน้นด้วยสีมีคำที่ถูกลบไปในขั้นตอนการทำความสะอาดข้อมูลด้วยการลบ Stopword ซึ่งคำที่ถูกลบออกไปนั้นเป็นคำที่สื่อไปในทิศทางลบ เห็นได้ชัดจากประโยค “do not expect a coherent plot” เมื่อถูกลบ Stopword ออกไปจะเหลือเพียง “expect coherent plot” ซึ่งคำที่ซื้อความหมายในเชิงลบอย่างคำว่า “not” ถูกลบออกไปทำให้แบบจำลองทำนายข้อความนี้ผิดเป็นความคิดเห็นเชิงบวก

This film was more effective in persuading me of a Zionist conspiracy than a Muslim one. And I'm Jewish.  
Anbody go to journalism school? Read an editorial? Freshman year rhetoric? These alarmist assertions, presented in a palatable way, might prove persuasive. But by offering no acknowledgment of possible opposing arguments, nor viable (or any at all) solutions, few sources and each of dubious origin, makes the argument an ineffectual diatribe.  
And thank goodness for that -- I wouldn't want anyone to leave the theatre BELIEVING any of this racist claptrap.  
A good lesson for me -- and hopefully a cautionary tale for you -- to actually read about a film before seeing it.

ภาพที่ 5.1.3 ตัวอย่างข้อความเชิงบวกที่แบบจำลองทั้งสองทำนายผิดพลาดก่อนทำการทดสอบ

jane austen would definitely approve one gwyneth paltrow awesome job capture attitude emma funny without  
excessively silly yet elegant put convince british accent british maybe good judge fool also excellent slide door  
sometimes forget american also brilliant jeremy northam sophie thompson phyllida law emma thompson sister mother  
bate woman nearly steal show ms law even line highly recommend

ภาพที่ 5.1.4 ตัวอย่างข้อความเชิงบวกที่แบบจำลองทั้งสองทำนายผิดพลาดหลังทำการทดสอบ

จากภาพที่ 5.1.3 เป็นความคิดเห็นที่ถูกป้ายกำกับไว้ว่าเป็นความคิดเห็นเชิงบวก ซึ่งเป็นข้อความที่ยังไม่ได้ทำการทดสอบ ภาพที่ 5.1.4 คือข้อความคิดเห็นหลังจากการทำการทดสอบ ข้อความเหล่านี้ พบร่วมกับข้อความคิดเห็นนั้นพบคำที่อาจจะทำให้แบบจำลองทำนายข้อความผิด เช่น “fool” หรือ “silly” โดยปกติแล้วคำดังกล่าวมักถูกพบในข้อความคิดเห็นเชิงลบมากกว่า ความคิดเห็นเชิงบวก การที่พบคำดังกล่าวในข้อความอาจทำให้แบบจำลองทำนายข้อความผิดเป็นข้อความเชิงลบได้ และ TF-IDF อาจไม่เข้าใจบริบทของข้อความได้เนื่องจากได้กำหนดค่าพารามิเตอร์ของ TfidfVectorizer เป็น Unigram และ Bigram ซึ่งทำให้ TF-IDF พิจารณาเป็นคำเดียวๆ หรือสองคำที่อยู่ติดกันเท่านั้น ตัวอย่างเช่น “without excessively silly” TF-IDF จะพิจารณาเป็น “without”, “excessively”, “silly”, “without excessively” หรือ “excessively silly” ทำให้จากข้อความที่เป็นเชิงบวกแบบจำลองอาจจะทำนายเป็นข้อความเชิงลบได้

### 5.1.2 การทดสอบประสิทธิภาพของแบบจำลองกับความคิดเห็นบนยูทูป

การทดสอบประสิทธิภาพของแบบจำลองทั้งสองกับชุดข้อความคิดเห็นต่อภาพนัตร์เรื่อง Sinners ซึ่งเป็นภาพนัตร์เกี่ยวกับแรมไพร์และมีนักแสดงส่วนใหญ่เป็นชาวผิวดำ โดยทำการจัดเก็บความคิดเห็นจากยูทูปทั้งหมด 500 ข้อความ ซึ่งได้นำข้อความคิดเห็นทั้งหมดไปติดป้ายกำกับด้วย ChatGPT เพื่อจำแนกความคิดเห็นว่าเป็นข้อความเชิงบวกหรือลบ

review  
You keep dancing with the Devil one day he's gon follow you home? waaaah waaaah ☺☺☺☺☺ that doesn't make sense.  
This movie was one of the biggest hype . It was slow and boring. Everything felt rushed closer to the end not a lot explanation. Confusing.  
Save yourselves the cost of a ticket. Go watch From Dusk Till Dawn instead. If you were engrossed by this borefest, you have no knowledge of movie history  
For those vampires...who seek in others...to obtain...what they could not achieve with their own gift!!!  
Rebuke every evil in the name Jesus christand protect all that watch dis ,they silly and running out of time 😂  
Is that a DJ I see on the left side at 1:55 ? Time traveling vampires  
Big success In The US, but a failure overseas, see Box Office Mojo  
We're are the gingers there black now to you know  
30 days of night with a few twists.  
Buddy, the Vampire Slayer ! great ambiance, wonderfull music, such a pleasure to have a movie dedicated to the Blues Music ! please, no sequel!!!  
Just saw this movie tonight - Amazing 😊  
Grateful to be alive for the golden age of blockbuster black horror  
This gives me lovecraft country vibes.  
This film is one of the greatest Vampire films next to Dracula (1931). Fright Night (1985) and the Fright Night remake (2011). and The Lost Boys (1987).

ภาพที่ 5.1.5 ตัวอย่างความคิดเห็นต่อภาพนัตร์ Sinners บนยูทูป

จะเห็นได้ว่าความคิดเห็นส่วนใหญ่มีขنادของความยำที่สั้นมากเป็นข้อความที่ไม่เกิน 2 บรรทัด และเป็นข้อความที่มีการใช้คำแสดงรวมถึงอิโมจิในข้อความเพื่อแสดงความรู้สึก

เมื่อนำแบบจำลองทั้ง 2 ไปทำการจำแนกประเภทของแบบจำลองพบว่าแบบจำลองทั้งสองมีความสามารถในการทำนายประสิทธิภาพของแบบจำลองที่เท่ากัน คือ แบบจำลองการถดถอยโลจิสติกสม์ค่าความถูกต้องอยู่ที่ 0.77 และ แบบจำลองชัฟพอร์ตเวกเตอร์แมชชีนมีค่าความถูกต้องเท่ากับ 0.75 ซึ่งทั้งสองแบบจำลองมีความสามารถในการทำนายความคิดเห็นเชิงบวกได้มากกว่าความคิดเห็นเชิงลบทั้งค่าสั้งเกตได้จากค่าความอ่อนไหวในกลุ่มความคิดเห็นเชิงลบของแบบจำลองการถดถอยโลจิสติกส์เท่ากับ 0.66 และแบบจำลองชัฟพอร์ตเวกเตอร์แมชชีนเท่ากับ 0.63 เมื่อเทียบกับค่าความอ่อนไหวในกลุ่มความคิดเห็นเชิงบวกของแบบจำลองการถดถอยโลจิสติกส์เท่ากับ 0.85 และแบบจำลองชัฟพอร์ตเวกเตอร์แมชชีนเท่ากับ 0.84 เห็นได้ชัดว่าทั้งสองแบบจำลองสามารถทำนายข้อความเชิงบวกได้ถูกต้องมากกว่าอย่างเห็นได้ชัด ซึ่งอาจเพราะการลบคำที่เป็น Stopword ที่สื่อความหมายในเชิงลบออกไปทำให้มีอัตราความจริงที่มีขนادสั้นอย่างเช่น “That looks not great” แบบจำลองอาจทำนายผิดพลาดได้

review	Logistic Regression	SVM	Label
We're are the gingas there black now to you know	Positive	Negative	Negative
Just got home after walking out on this abomination. Blaxxploitation is back, baby! Was it written by students?	Positive	Negative	Negative
Where is the go woke go broke crew now 😂	Positive	Negative	Negative
Came here from the viral clip of Hayley Steinfeld. Not at all the tone I expected, but my interest is piqued. Honestly,	Negative	Positive	Positive
Not a good movie	Positive	Negative	Negative
It's was the best could watch it again	Negative	Positive	Positive
Is this from dusk till dawn part 2 ?	Positive	Negative	Negative
Since it's the same directors from creed and black panther and there's a common actor I see favoritism	Positive	Negative	Negative
i'm here after watching the movie and oh my god it was incredible. i walked into the movie basically blind coz i didn	Negative	Positive	Positive
i'm here after watching the movie and oh my god it was incredible. i walked into the movie basically blind coz i didn	Negative	Positive	Positive
Is that movie shot in 3:1 or am I hallucinating? What an interesting choice!	Negative	Positive	Positive
This movie was well maid !!! Soo good ! Loved that fact that they mixed history with science fiction	Negative	Positive	Positive
why its giving the plot of season "FROM"	Positive	Negative	Negative
Ive literally heard nothing but GOOD things about this movie. Going to see it tomorrow!!!	Negative	Positive	Positive
This trailer is 🤣	Positive	Negative	Negative
That looks not great	Positive	Negative	Negative
Ehhh 3/10 movie give me my time back	Positive	Negative	Negative
One thing for certain and two things for sure you can always count on people to hate!	Positive	Negative	Negative
The country accents 😂	Positive	Negative	Negative

ภาพที่ 5.1.6 ตัวอย่างความคิดเห็นที่แบบจำลองการถดถอยโลจิสติกส์ทำนายผิดพลาด

review		Logistic Regression	SVM	Label
For those vampires...who seek in others...to obtain...what they could not achieve with their own gift!!!		Negative	Positive	Negative
This gives me lovecraft country vibes.		Positive	Negative	Positive
This film is one of the greatest Vampire films next to Dracula (1931), Fright Night (1985) and the Fright Night remake (2011).		Negative	Positive	Negative
After a long time came across a movie made for the cinema thank you Ryan Coogler and Ludwig Goransson what a ride!		Positive	Negative	Positive
This movie is better than any trailer could ever do justice! This doesn't even capture half of the themes. Go see it now!		Negative	Positive	Negative
Is it horror?		Positive	Negative	Positive
This trailer deserves an award		Positive	Negative	Positive
The Irish guy depicted the dancing as if it was researched. Loved the scene.		Negative	Positive	Negative
I thought it would be a good film for the music. I made it about 5 minutes in until they started that evil whitey BS.	😊	Positive	Negative	Positive
So it's like that series called From? Where creatures comes out at night disguising as humans, and they can be repelled.		Positive	Negative	Positive
It seems like Abraham Lincoln didn't get rid of all the vampires during the Civil War. 😊		Positive	Negative	Positive
1:30 Thats some scary 🧛‍♂️ to say to someone!! "If you keep dancing with the devil, One day he's going to follow you."		Positive	Negative	Positive
0:14 That's how I be feeling rolling the blunt on a cold windy day!!!		Positive	Negative	Positive
Vampires here remind me a bit of a trailer of a new video game - The Blood of Dawnwalker .		Positive	Negative	Positive
Just seen this movie . The concept is sooo good but they unfortunately do too much story telling .. I want drama !! M		Negative	Positive	Negative
Don't tell me there's a curtain on the light in a room nd someone goes under the bed someone sleeping on nd they		Negative	Positive	Negative

ภาพที่ 5.1.7 ตัวอย่างความคิดเห็นที่แบบจำลองทัพพอร์ตເຕ່ອງແມ່ນໜຶ່ນທຳນາຍຜິດພາດ

#### การวิเคราะห์ข้อความที่แบบจำลองທຳນາຍຜິດພາດ

- จากข้อความที่ທຳນາຍຜິດພາດอาจเป็นเพราะข้อความมีการใช้อิโมจิໃນข้อความ ซึ่งข้อมูลที่ใช้ฝึกสอนแบบจำลองไม่มีการใช้อิโมจิในการฝึกสอนจึงทำให้แบบจำลองທຳນາຍຜິດພາດ ตัวอย่างเช่น “This trailer is 🎉” ให้ความหมายเป็นข้อความเชิงลบแต่แบบจำลองທຳນາຍເປັນเชิงบวก
- ข้อความคิดเห็นที่ภาพรวมเป็นความคิดเห็นเชิงลบแต่เมื่อคำที่มีความหมายเชิงบวกประกอบอยู่ด้วยจึงทำให้แบบจำลองທຳນາຍข้อความนั้นເປັນข้อความเชิงบวก เช่น “That looks not great” ซึ่ง “great” มักถูกพับໃນข้อความเชิงบวก เป็นส่วนใหญ่จึงทำให้แบบจำลองอาจທຳນາຍຜິດພາດ และອີກປັບຈັກຄືວິທີ ในการฝึกสอนแบบจำลองขั้นตอนการทำความสะอาดข้อมูลมีการลบ Stopword ที่มี “not” ประกอบอยู่ด้วยจึงทำให้ TF-IDF อาจไม่ได้ให้น้ำหนักคำนี้และมองข้อความนี้เป็นข้อความเชิงบวก
- ข้อความมีความหมายเชิงประชดประชัน หรือ ใช้คำแสลง เช่น “Where is the go woke go broke crew now 😂” ซึ่งแปลได้ว่า “พวกที่ตื่นรู้ไปไหนแล้วล่ะ” มีการใช้คำแสลง “woke” ที่แปลว่า “ตื่นรู้” และมีความหมายประชดประชัน ซึ่งสื่อความหมายไปในทางลบเนื่องจากภาษาญี่ปุ่นเรื่องนี้ເປັນการพูดถึง

แรมไพร์ที่มีสีผิวดำและประโภคนี้สือไปในทางเหยียดสีผิว ซึ่งแบบจำลองยังไม่มีความสามารถในการจำแนกข้อความที่มีความหมายประชดประชันได้

- ข้อความที่มีคำศัพท์ที่ไม่ค่อยพบเจอบ่อยๆ เช่น “Is that movie shot in 3:1 or am I hallucinating? What an interesting choice!” ให้ความหมายในเชิงบวก แต่แบบจำลองทำนายเป็นเชิงลบ และเมื่อลองหาสาเหตุแล้วพบว่าในชุดข้อมูลทดสอบนั้นพบคำศัพท์ “hallucinating” เพียงแค่ 9 ข้อความจากข้อความคิดเห็นทั้งหมด 50,000 ข้อความเท่านั้น จึงอาจทำให้ข้อความนี้ไม่ถูกให้น้ำหนักโดย TF-IDF
- บางข้อความมีความคลุมเครือว่าเป็นความคิดเห็นเชิงบวกหรือความคิดเห็นเชิงลบ ตัวอย่างเช่น “Scariest movie I've ever been in as a stunt actor.” ซึ่งเป็นข้อความที่มีความหมายกลางๆ ไม่ได้สื่อไปในทิศทางบวกหรือลบอย่างชัดเจน

## 5.2 ข้อเสนอแนะและแนวทางในการพัฒนา

การวิเคราะห์ความคิดเห็นของภาพนิทรรศจากแพลตฟอร์มออนไลน์ (Movie Reviews Sentiment Analysis from Online Platforms) สามารถนำไปพัฒนาต่ออยอดได้โดยทำให้มีความสมบูรณ์มากยิ่งขึ้น โดยทางผู้จัดทำโครงงานได้มีข้อเสนอแนะซึ่งเป็นแนวทางในการพัฒนาต่ออยอดดังต่อไปนี้

1. เพิ่มชุดข้อมูลฝึกสอนให้มีความหลากหลายมากขึ้น เช่น ข้อความที่มีอิโมจิประกอบข้อความที่มีความประชดประชัน หรือ คำแสดง
2. สามารถแยกประเภทของข้อความเป็น ความคิดเห็นเชิงบวก ความคิดเห็นเชิงลบ และความคิดเห็นกลาง ได้เนื่องจากบางข้อความอาจจะไม่ได้สื่อไปในทางทิศทางบวกหรือทางลบอย่างชัดเจน
3. สามารถดึงข้อมูลได้จากหลากหลายแพลตฟอร์มมากกว่านี้ และมีแพลตฟอร์มที่นำเสนอเจาะลึกอย่าง Facebook, เว็บไซต์ IMDB หรือ Reddit
4. ในข้อความที่มีอิโมจิสามารถเปลี่ยนอิโมจิเป็นคำกริยาของอิโมจินั้นๆ แล้วนำมายัดได้
5. สามารถพัฒนาโมเดลการวิเคราะห์ข้อความสำหรับภาษาไทยเพิ่มขึ้นได้
6. พัฒนาโมเดลการวิเคราะห์ให้มีประสิทธิภาพเพิ่มขึ้นด้วยโมเดลที่สามารถเข้าใจภาษาธรรมชาติ (NLP) ได้ เช่น BERT

## รายการอ้างอิง

- Chakrit. (28 พฤษภาคม 2562). *tf-idf-ทำงานยังไง*. เข้าถึงได้จาก softnix.co.th:  
<https://www.softnix.co.th/2019/05/28/tf-idf-%E0%B8%97%E0%B8%B3%E0%B8%87%E0%B8%B2%E0%B8%99%E0%B8%A2%E0%B8%B1%E0%B8%87%E0%B9%84%E0%B8%87/>
- Datawow. (17 มิถุนายน 2563). *ขอให้โชคดีมีชัยในโลก NLP — Part 1*. เข้าถึงได้จาก .datawow.io: <https://www.datawow.io/blogs/wish-you-luck-in-nlp-world>
- Dynamic Intelligence Asia. (ม.ป.ป.). *NATURAL LANGUAGE PROCESSING ตัวเข็มการลือสารมนุษย์-คอมพิวเตอร์*. เข้าถึงได้จาก dia.co.th: <https://www.dia.co.th/articles/what-is-nlp/>
- Jirapat Klaokliang. (9 ตุลาคม 2564). *ประเมินประสิทธิภาพไม้เดลต์ด้วย Confusion Matrix*. เข้าถึงได้จาก medium.com:  
<https://nattrio.medium.com/%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B9%80%E0%B8%A1%E0%B8%B4%E0%B8%99%E0%B8%9B%E0%B8%A3%E0%B8%B0%E0%B8%AA%E0%B8%B4%E0%B8%97%E0%B8%98%E0%B8%B4%E0%B8%A0%E0%B8%B2%E0%B8%9E%E0%B9%82%E0%B8%A1%E0%B9%80%E0%B8%94%E0%B8%A5%E0%B8%94%E0%B9%89%E0>
- Mandala. (18 กันยายน 2567). *Sentiment Analysis คือ? ช่วยให้การทำธุรกิจอย่างไร [พร้อมตัวอย่าง]*. เข้าถึงได้จาก blog.mandalasystem.com:  
<https://blog.mandalasystem.com/th/sentiment-analysis>
- Narut Soontranon. (15 ตุลาคม 2566). *SVM คือ อะไร*. เข้าถึงได้จาก nerd-data.com:  
<https://www.nerd-data.com/svm/>
- Nectec. (19 สิงหาคม 2565). *การประยุกต์ใช้ Machine Learning กับงานในภาคอุตสาหกรรม (ตอนที่ 1)*. เข้าถึงได้จาก nectec.or.th: <https://www.nectec.or.th/news/news-public-document/machine-learning-manufact-1.html>
- Ninenox. (24 กันยายน 2563). *ทำความเข้าใจ accuracy,precision,recall,f1-score*. เข้าถึงได้จาก ninenox.com:  
<https://www.ninenox.com/2020/09/24/%E0%B8%97%E0%B8%B3%E0%B8%84>

%E0%B8%A7%E0%B8%B2%E0%B8%A1%E0%B9%80%E0%B8%82%E0%B9%89  
 %E0%B8%B2%E0%B9%83%E0%B8%88-accuracyprecisionrecallf1-score/

Panaya Sudta. (19 กันยายน 2566). นักการตลาดกับ AI: Confusion Matrix และหลักการประเมินประสิทธิภาพ ML Model. เข้าถึงได้จาก everydaymarketing.co:  
<https://everydaymarketing.co/business-and-marketing-case-study/ai/confusion-matrix-ml-evaluation/>

Pisinee T. (12 พฤษภาคม 2567). *Sentiment Analysis* เมื่อความรู้สึกของลูกค้าสามารถวิเคราะห์ได้ผ่านชื่อความ. เข้าถึงได้จาก contentshifu.com:  
[https://contentshifu.com/blog/what-is-sentiment-analysis#Sentiment\\_Analysis](https://contentshifu.com/blog/what-is-sentiment-analysis#Sentiment_Analysis)

Surapong Kanoktipsatharporn. (17 พฤษภาคม 2562). *Stop Words* คืออะไร ใน Natural Language Processing – NLP ep.2. เข้าถึงได้จาก <https://www.bualabs.com/>:  
<https://www.bualabs.com/archives/2937/what-is-stop-words-in-natural-language-processing-nlp-ep-2/>

Thaiware. (4 สิงหาคม 2567). *tips.thaiware.com*. เข้าถึงได้จาก Natural Language Processing คืออะไร ? รู้จักการประมวลผลภาษาธรรมชาติกัน:  
<https://tips.thaiware.com/2558.html#history-of-natural-language-processing>  
 วิสัย เอไอ. (ม.ป.ป). เอไอตัดคำภาษาไทยคืออะไร? เข้าถึงได้จาก <https://acp.visai.ai/>:  
<https://acp.visai.ai/th/product/word-tokenization>

ธนาสิทธิ์ เร่งสมบูรณ์สุข และ กวินริดา สายยศ. (2562). ระบบวิเคราะห์ความคิดเห็นต่อองค์กรไทย บนหัวตัวเตอร์. มหาวิทยาลัยจุฬาลงกรณ์, กรุงเทพฯ

เพ็ญนา จุ่มพลพงษ์. (2564). การระบุรีวิวที่น่าสงสัยในผลิตภัณฑ์และผิวน้ำของลูกค้าบนตลาดชื่อขายลินค์ออนไลน์ ด้วยเทคนิคปริมาณการเรียนรู้ของเครื่อง. มหาวิทยาลัยแม่ฟ้าหาร, พิษณุโลก

รวิสุดา เทศเมือง และ นิเวศ จิระวิชิตชัย. (2560). การวิเคราะห์ความคิดเห็นภาษาไทยเกี่ยวกับการรีวิวลินค์ออนไลน์ โดยใช้ขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน. มหาวิทยาลัยศรีปทุม, กรุงเทพฯ

## ภาคผนวก

### ภาคผนวก ก วิธีการดึงข้อมูลมาไว้เคราะห์

#### 1. การดึงข้อมูลจากยูทูปมาไว้เคราะห์ด้วย API

```

from googleapiclient.discovery import build
from langdetect import detect, DetectorFactory
from langdetect.lang_detect_exception import LangDetectException

DetectorFactory.seed = 0

API_KEY = "REDACTED"

def fetch_youtube_comments(query, max_comments=500):
    try:
        youtube = build("youtube", "v3", developerKey=API_KEY)
        search = youtube.search().list(q=query, part="id", maxResults=1, type="video").execute()
        video_id = search['items'][0]['id']['videoId']
        comments = []
        next_page_token = None

        while len(comments) < max_comments:
            response = youtube.commentThreads().list(
                part="snippet",
                videoId=video_id,
                maxResults=100,
                textFormat="plainText",
                pageToken=next_page_token
            ).execute()

            for item in response.get("items", []):
                comment_text = item["snippet"]["topLevelComment"]["snippet"]["textDisplay"]
                try:
                    if detect(comment_text) == "en":
                        comments.append(comment_text)
                        if len(comments) >= max_comments:
                            break
                except LangDetectException:
                    continue

            next_page_token = response.get("nextPageToken")
            if not next_page_token:
                return comments

    except Exception as e:
        print("Error fetching comments:", e)
        return []

```

## 2. การดึงข้อมูลจากยูปมารวิเคราะห์ด้วย API

```
import requests

BEARER_TOKEN = 'REDACTED'

def fetch_twitter_comments(hashtag, max_results=50):
    if not BEARER_TOKEN:
        raise ValueError("TWITTER_BEARER_TOKEN is missing.")

    query = f"#{hashtag} lang:en -is:retweet"
    url = "https://api.twitter.com/2/tweets/search/recent"
    headers = {"Authorization": f"Bearer {BEARER_TOKEN}"}
    params = {
        "query": query,
        "tweet.fields": "lang,text",
        "max_results": min(max_results, 100)
    }

    response = requests.get(url, headers=headers, params=params)
    if response.status_code != 200:
        return []

    tweets = response.json().get("data", [])
    return [tweet["text"] for tweet in tweets if tweet["lang"] == "en"]
```

