การระบุตัวตนของมนุษย์จากข้อความตัวอักษร

Human Identification from Text Messages

พลภัทร จงวัฒนศิริ

ลิขิตภูมิ ลิขิตงาม

ปริญญานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรบัณฑิต

สาขาวิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ปีการศึกษา 2565

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ปริญญานิพนธ์ปีการศึกษา 2565 ภาควิชาวิศวกรรมคอมพิวเตอร์ สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เรื่อง การระบุตัวตนของมนุษย์จากข้อความตัวอักษร

HUMAN IDENTIFICATION FROM TEXT MESSAGES

ผู้จัดทำ

1. นายพลภัทร จงวัฒนศิริ รหัสนักศึกษา 62010604
2. นายลิขิตภูมิ ลิขิตงาม รหัสนักศึกษา 62010785

_____อาจารย์ที่ปรึกษา

(รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

การระบุตัวตนของมนุษย์จากข้อความตัวอักษร

นายพลภัทร จงวัฒนศิริ 62010604
นายลิขิตภูมิ ลิขิตงาม 62010785
รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ อาจารย์ที่ปรึกษา
ปีการศึกษา 2565

บทคัดย่อ

ในปัจจุบัน เทคโนโลยีปัญญาประดิษฐ์ถือเป็นอีกศาสตร์หนึ่งที่นักวิจัยหลายๆ คนทั่วโลกให้ความสนใจ เนื่องจากเป็นศาสตร์ที่ว่าด้วยการสร้างความฉลาดให้กับเครื่องจักร โดยเฉพาะอย่างยิ่งกับระบบคอมพิวเตอร์ และมีการนำศาสตร์ทางด้านปัญญาประดิษฐ์มาประยุกต์ใช้ในหลายๆ ด้าน ซึ่งหนึ่งในนั้นก็คือ การระบุตัวตนของผู้เขียนข้อความผ่านทางตัวอักษร คณะผู้จัดทำได้พัฒนาเว็บแอปพลิเคชันที่สามารถคัดแยกตัวบุคคลผ่านทางตัวอักษร โดยใช้ปัญญาประดิษฐ์ในการคำนวณและวิเคราะห์ว่าข้อความเป็นของผู้เขียนคนไหน ซึ่งผู้ใช้สามารถกรอกข้อความ แล้วจากนั้นระบบจะแสดงตัวผู้เขียนที่ได้เขียนข้อความนั้นออกมาแสดงผลทางหน้าเว็บไซต์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Human Identification from Text Messages

Mr.	Phonlapat	Jongwatanasiri	62010604
Mr.	Likhitbhum	Likhitngam	62010785
Assoc. Prof. Dr.	Kietikul	Jearanaitanakij	Advisor

Academic Year 2022

ABSTRACT

Nowadays, artificial intelligence (AI) is one of many issues that become attractive to many researchers. It describes how to create the intelligent for the machine, especially for the computer system. Many researchers and developers have adopted the AI into many fields and which one of it is wanting to know the author of a message through letters. Our team have developed a web application that can identify people through letters. Using artificial intelligence to calculate and analyze which author's text belongs to in which the user can enter text. Then the system will display author who has written that message to be displayed on the website.

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

ปริญญานิพนธ์ฉบับนี้สามารถดำเนินการจนประสบความสำเร็จ เนื่องจากได้รับความอนุเคราะห์จาก รศ.ดร.เกียรติกูล เจียรนัยธนะกิจ อาจารย์ที่ปรึกษาปริญญานิพนธ์ที่กรุณาให้คำแนะนำ และคำปรึกษาเพื่อปรับปรุงและแก้ไขข้อบกพร่อง รวมทั้งการให้องค์ความรู้และแนวทางการศึกษาค้นคว้ามาโดยตลอด คณะผู้จัดทำขอกราบขอบพระคุณเป็นอย่างสูงไว้ ณ ที่นี้

ขอขอบพระคุณอาจารย์สรยุทธ กลมกล่อม อาจารย์ผู้ประสานงานที่ให้ข้อมูลและคำปรึกษาแก่ผู้จัดทำจนปริญญานิพนธ์สามารถสำเร็จลงด้วยดี

คณะผู้จัดทำมีความซาบซึ้งในความกรุณาของทุกท่านที่กล่าวถึงและผู้ที่ไม่ได้เอ่ยนามในที่นี้ที่มีส่วนให้ความช่วยเหลือและสนับสนุนด้วยดีตลอดมา จึงขอกราบขอบพระคุณด้วยความจริงใจ

พลภัทร จงวัฒนศิริ

ลิขิตภูมิ ลิขิตงาม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ (ต่อ)

หน้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตาราง **หน้า**

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูป **หน้า**

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาของปัญหา

ในปัจจุบัน สังคมออนไลน์การเขียนเพื่อการสื่อสารเป็นสิ่งสามัญ ในการติดต่อกับ บุคคลอื่น โดยในโลกออนไลน์นั้นการเขียนความคิดเห็นเว็บบอร์ดบล็อก หรือบทความต่างๆล้วนแล้วแต่เป็นการพิมพ์ผ่านตัวบุคคลทั้งนั้น ถึงแม้ว่าจะเขียนในรูปแบบหรือในเว็บไซต์ที่ต่างกันเช่น Facebook Pantip แต่ก็เป็นตัวบุคคลเองที่เป็นคนพิมพ์ดังนั้นทำให้สำนวนการพิมพ์ของบุคคลนั้นยังคงเหมือนเดิม

ในสังคมออนไลน์บุคคลที่กระทำผิดในโลกออนไลน์มีอยู่มากแต่ไม่สามารถระบุตัวตนได้ เนื่องจากมีการใช้อวตาร์(Avatar)ในการกระทำผิด ซึ่งถ้าหากว่าเราสามารถระบุตัวตนของบุคคลที่กระทำผิดได้นั้นก็จะสามารถทำให้สังคมออนไลน์น่าใช้มากขึ้น

ในปัจจุบันนี้มีเทคโนโลยีที่ชื่อว่า Machine Learning ซึ่งเป็นเทคโนโลยีที่สามารถทำให้ระบบคอมพิวเตอร์เรียนรู้ได้ด้วยตนเองโดยใช้ข้อมูลที่มีอยู่ซึ่งเรียกว่า Dataset โดย Machine Learning นี้สามารถทำการจัดหมวดหมู่ของข้อมูล หรือเรียกว่า Classification ได้ซึ่งจะเป็นหัวใจหลักของการแยกตัวบุคคลจากบทความ หรืองานเขียนต่างๆ และสามารถพัฒนาไปสู่การระบุความเป็นไปได้ของตัวบุคคลจากข้อความที่ไม่ระบุตัวตนในโลกออนไลน์ในอนาคตได้

1.2 วัตถุประสงค์ของโครงงาน

1) เพื่อสามารถนำไปใช้ในการระบุตัวตนบุคคลจากบทความได้
2) เพื่อสามารถนำไปพัฒนาต่อจนสามารถค้นหาบุคคลในโลกออนไลน์จากข้อความที่เขียนขึ้นได้

1.3 ขอบเขตของโครงงาน

1.3.1 ขอบเขตระบบของ Web Application

1) User สามารถพิมพ์ข้อความลงไปในช่องอินพุตเพื่อทำนายผลออกมาได้
2) User สามารถเลือกการค้นหาโดยเป็นคำหรือข้อความก็ได้
3) User พิมพ์ได้แค่ภาษาอังกฤษกับตัวเลข

1.3.2 ขอบเขตระบบฐานข้อมูล

1) จัดเก็บโมเดลที่ได้จากทำ Machine Learning

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.3.3 ขอบเขตระบบการระบุตัวตนจากข้อความ

1) รองรับเฉพาะบน Web Browser

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ได้ศึกษาการเขียนของบุคคลต่าง ๆ
2) ได้ศึกษาเกี่ยวกับการประมวลผลตามธรรมชาติ
3) ได้ศึกษาเกี่ยวกับ Machine Learning/Deep Learning
4) ได้ศึกษาโครงสร้าง Website ที่สามารถใช้งานกับโมเดลจาก Machine Learning/ Deep Learning ได้

1.5 แผนการดำเนินงาน

การพัฒนาโครงงานในปีการศึกษาที่ 1/2565 เริ่มตั้งแต่เดือนสิงหาคม พ.ศ. 2565 ถึงเดือนธันวาคม พ.ศ. 2565 ตามตารางที่ 1.1

ตาราง 1.1 แผนการดำเนินงานในปีการศึกษาที่ 1/2565

รายการ	สิงหาคม					กันยายน				ตุลาคม					พฤศจิกายน				ธันวาคม			
	1	2	3	4	5	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4
ค้นหาบทความจากผู้เขียนอย่างน้อย 50 คน																						
ตรวจสอบบทความที่ผู้เขียนได้เขียนไว้																						
เลือกบทความจากผู้เขียนที่จะใช้ และจัดระเบียบข้อความให้ถูกต้องและสมบูรณ์																						
ติดตั้ง Python																						
ติดตั้ง Scikit-learn สำหรับการสร้าง ML																						

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ติดตั้ง Pandas สำหรับใช้ในการสร้าง Data Frame																				
ติดตั้ง Numpy สำหรับใช้งานร่วมกับ Pandas																				
ติดตั้ง PythaiNLP																				
ทดสอบใช้งาน Library ที่ติดตั้งแล้วและแก้ปัญหาหากมีข้อผิดพลาด																				
นำบทความที่ได้ทั้งหมดมาทำ Text Cleaning																				
นำบทความทั้งหมดมาทำ Tokenize (ตัดคำ)																				
นำบทความทั้งหมดมาแยก Feature ด้วย TF-IDF																				
สรุปผลและจัดทำรูปเล่ม																				

การพัฒนาโครงงานในปีการศึกษาที่ 2/2565 เริ่มตั้งแต่เดือนธันวาคม พ.ศ. 2565 ถึงเดือนเมษายน พ.ศ. 2565 ตามตารางที่ 1.2

ตาราง 1.2 แผนการดำเนินงานในปีการศึกษาที่ 2/2565

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รายการ	ธันวาคม				มกราคม					กุมภาพันธ์				มีนาคม				เมษายน			
	1	2	3	4	1	2	3	4	5	1	2	3	4	1	2	3	4	1	2	3	4
ทำการ Count Word คำที่กรองมาจาก TF-IDF																					
ทำข้อมูลเป็น Data Frame																					
แบ่งข้อมูลออกเป็นส่วน Test size และ Train size																					
เทรนโมเดลด้วยเทคนิค Logistic Regression																					
เทรนโมเดลด้วยเทคนิค LSTM																					
ปรับจูน Parameter เพื่อเพิ่มค่า Accuracy rate																					
หา CNN ที่เหมาะสม																					
ติดตั้ง Node.js																					
ติดตั้ง React.js																					
ติดตั้ง MongoDB																					
ติดตั้ง Axios																					
ติดตั้ง Flask																					
ติดตั้ง Bootstrap 5 สำหรับตกแต่งหน้าเว็บแอปพลิเคชัน																					

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ออกแบบ UXUI																				
สร้าง Component หน้า HOME																				
สร้าง Component หน้าค้นหาบุคคลจากบทความ(INPUT)																				
สร้าง Component หน้าค้นหาบุคคลจากบทความ(OUTPUT)																				
สร้าง Component หน้าผู้จัดทำ																				
เชื่อมต่อ Back-End กับโมเดล																				
สร้าง API สำหรับรับและส่งข้อมูลที่โมเดลต้องการ																				
สร้าง API เพื่อเชื่อมต่อ Front-End กับ Back-End																				
Deploy ระบบเว็บแอปพลิเคชันลง Server																				
ทดสอบระบบเว็บแอปพลิเคชันและแก้ไข																				
สรุปผลและจัดทำรูปเล่ม																				

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

2.1 แนวคิดและทฤษฎีที่เกี่ยวข้อง

2.1.1 การถดถอยโลจีสติก (Logistic Regression)

เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุที่มีวัตถุประสงค์เพื่อประมาณค่าหรือทำนายเหตุการณ์ที่น่าสนใจว่าจะเกิดหรือไม่เกิดเหตุการณ์นั้นภายใต้อิทธิพลของตัวปัจจัย ประกอบไปด้วยตัวแปรตาม และตัวแปรอิสระ การวิเคราะห์การถดถอยแบบโลจีสติกเกี่ยวข้องกับทฤษฎีความน่าจะเป็นทวินามถูกเรียกว่า Binomial Logistic Regression การถดถอยโลจีสติกจัดเป็นเครื่องมือวิเคราะห์ข้อมูลในการศึกษาวิจัยเพื่อทำนายเหตุการณ์ หรือประเมินความเสี่ยง จึงมีการประยุกต์ในงานวิจัยหลากหลายสาขา ทั้งสาขาทางการแพทย์ วิศวกรรมศาสตร์ นิเวศวิทยา เศรษฐศาสตร์ และสังคมศาสตร์

การสร้างแบบจำลองหรือสมการถดถอยโลจีสติก กระทำโดยการประมาณค่าสัมประสิทธิ์จากชุดข้อมูลสังเกตที่เก็บวัดมาได้ โดยใช้วิธีประมาณความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation; MLE) ซึ่งต่างจากการวิเคราะห์การถดถอยทั่วไปที่ใช้วิธีกำลังสองน้อยที่สุด (Least Square Method) ที่มีแนวคิดว่าต้องให้ได้ค่าส่วนเหลือ (Residuals) มีการกระจายแบบปกติ ดังนั้นการวิเคราะห์แบบ MLE เพื่อให้ได้โมเดลที่เหมาะสมที่สุดจึงใช้กระบวนการวิเคราะห์แบบเวียนซ้ำ (Iteral Process) โดยเริ่มจากการประมาณค่าสัมประสิทธิ์ของตัวแปรเพื่อให้ได้สมการตั้งต้น หลังจากนั้นก็ใช้สมการเพื่อทำนายค่าแล้วนำมาคำนวณซ้ำเพื่อหาค่าสัมประสิทธิ์ใหม่ที่ทำให้ได้ความน่าจะเป็นสูงสุดเพื่อให้สามารถทำนายค่าของตัวแปรตามได้ใกล้เคียงค่าของข้อมูลจริงมากที่สุด

2.1.2 การประมวลผลภาษาธรรมชาติ (Natural Language Processing)

เป็นหนึ่งในเทคโนโลยีปัญญาประดิษฐ์เพื่อให้เกิดการสื่อสารข้อมูล รวมทั้งการวิเคราะห์ข้อมูลต่างๆออกมาได้อย่างที่มนุษย์สื่อสารกัน และมีวัตถุประสงค์ในการปิดช่องว่างทางการสื่อสารระหว่างมนุษย์และระบบคอมพิวเตอร์

แม้ว่าเทคนิคการทำงานทั้งแบบ Supervised learning และ Unsupervised Learning โดยเฉพาะอย่างยิ่งกระบวนการทำงานแบบ Deep Learning จะได้ถูกนำมาใช้งานอย่างแพร่หลายในการสร้างแบบจำลองวิเคราะห์ภาษาของมนุษย์แล้วก็ตาม ก็ยังคงมีความจำเป็นในการสร้างความเข้าใจทางภาษาศาสตร์ที่ลึกและซับซ้อนยิ่งขึ้น รวมถึงความรู้ความเข้าใจเฉพาะด้าน ซึ่งแตกแขนงความชำนาญย่อยออกไปจากเทคนิค Machine Learning ตามปกติอีกด้วย ด้วยเหตุนี้ NLP จึงมีความสำคัญในการลดความสับสนทางการวิเคราะห์ภาษาลง และเพิ่มมิติให้แก่ข้อมูลในรูปของ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

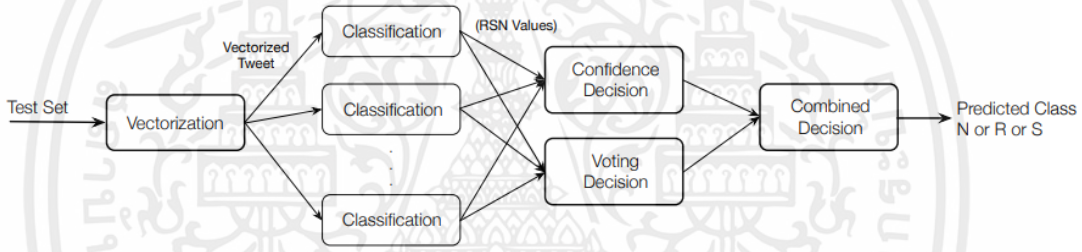
ตัวเลข เพื่อการนำไปใช้งานต่าง ๆ ต่อไป เช่น ในการทำ Speech Recognition หรือการใช้งาน Text Analytics

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 การตรวจจับข้อความที่ไม่ดีใน Twitter โดยใช้ Deep Learning

2.2.1.1 คำอธิบายของ Recurrent Neural Network (RNN)

การที่ Neural Network มีความสามารถในการค้นหาข้อมูล ซึ่งเป็นประโยชน์ในการจำแนกประเภท (Classification) การที่ Recurrent Neural Network เป็น Neural Network ชนิดพิเศษ ซึ่งจะเกิดการเพิ่มลูปในกระบวนการทำ โดยที่ RNN นั้นใช้ Back Propagation ในกระบวนการฝึกอบรมเพื่ออัพเดท Weight ในทุกๆเลเยอร์ โดยในการทดลองนี้จะใช้ RNN ที่มีชื่อว่า Long Short-Term Memory (LSTM) ซึ่งมีลำดับการทำงานดังรูปที่ 2.1



รูป 2.1 ภาพรวมของระบบที่มี Multiple Classification

การทำขั้นตอนแรกจะต้องใช้ความละเอียดที่มากเพื่อแยกคลาสของผู้ใช้แต่ละคนออกมาเป็นข้อความที่เป็นกลาง ข้อความที่เหยียดเชื้อชาติ และข้อความที่กีดกันทางเพศ โดยกำหนด Feature ทั้ง 3 อย่างคือ T_{Na}, T_{Ra} และ T_{Sa} ซึ่งแสดงถึงแนวโน้มข้อความของผู้ใช้ที่มีเนื้อหาที่เป็นกลาง แบ่งแยกเชื้อชาติ และเพศ ตามลำดับ ให้ m_a แทนเซตของผู้ใช้และให้ m_{Na}, m_{Ra} และ m_{Sa}แทนสับเซตของผู้ใช้ที่โพสต์เนื้อหาที่เป็นกลาง แบ่งแยกเชื้อชาติ และเพศ ตามลำดับ และ Feature ที่ได้ออกมาจะเป็นไปตามสมการที่ 2.1 ถึงสมการที่ 2.3

$$T_{Na} = \frac{|m_{Na}|}{|m_a|} \qquad (2.1)$$

$$T_{Ra} = \frac{|m_{Ra}|}{|m_a|} \qquad (2.2)$$

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

$$T_{Sa} = \frac{|m_{Sa}|}{|m_a|} \qquad (2.3)$$

เพื่อปรับปรุงความสามารถในการแยกประเภท เราจะใช้ Long Short-Term Memory (LSTM)

Algorithm 1 Ensemble classifier

1: **for** $tw \in$ {tweets} **do**
2: **for** $cl \in$ { classifiers } **do**
3: $(N_{cl}, R_{cl}, S_{cl}) \leftarrow \text{classifier}_{cl}(tw)$
4: $v_{cl} \leftarrow \max(N_{cl}, R_{cl}, S_{cl})$
5: $id_{cl} \leftarrow \arg\max(N_{cl}, R_{cl}, S_{cl})$
6: **end for**
7: $m \leftarrow \text{mode}(id_1, id_2, id_3)$
8: **if** $m \in$ {Neutral, Racist, Sexism} **then**
9: decision $\leftarrow m$
10: **else**
11: decision $\leftarrow id_{\arg\max(v_1, v_2, v_3)}$
12: **end if**
13: **print** decision for tw
14: **end for**

รูป 2.2 อัลกอริทึม 1 การนำเอา classifier หลายตัวมาใช้

โดยการทำการ Classification ก็จะใช้อัลกอริทึมดังรูปที่ 2.2 โดยจะมีการทำงานคือ ดูข้อความจากทวิตเตอร์แล้วทำการเพิ่มอินพุตคลาสเป็น id_{cl} โดยดูค่าฐานนิยม (Mode) ที่ได้จาก id_1, id_2, id_3 จากนั้นส่งข้อมูลออกเป็นเอาต์พุตในรูปแบบของผลลัพธ์ที่เป็นกลาง เหยียดเชื้อชาติ หรือกีดกันทางเพศ

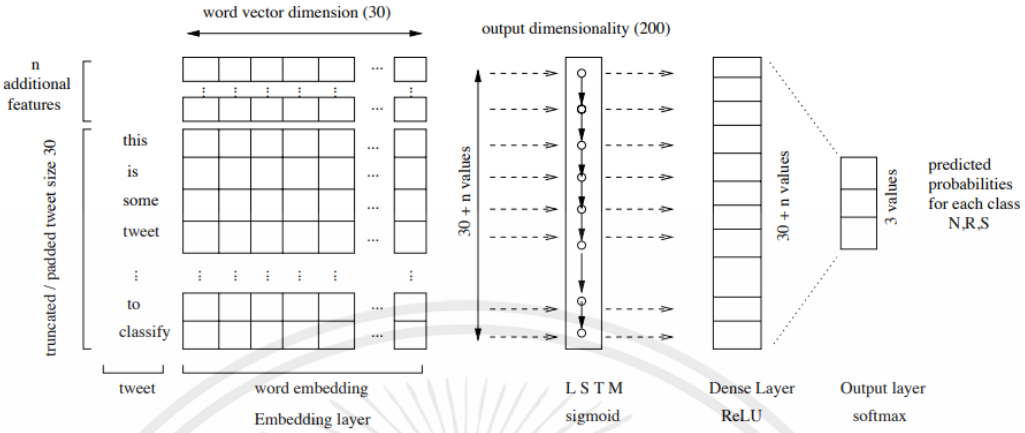
ต่อมาเป็นการทำขั้นตอน Data Processing ก่อนทำการเทรน Neural Network จะต้องทำการตัดคำ หรือกระบวนการ Tokenization ในทุกๆ โพสต์ของทวิตเตอร์ ในการทดลองนี้ได้กำหนดความยาวของข้อความไว้ที่ 30 คำต่อโพสต์ และโพสต์ที่น้อยกว่า 30 คำ จะทำการ Padding ให้ค่าเป็น 0 จากนั้นโพสต์จะถูกแปลงไปเป็นเวกเตอร์

ตาราง 2.1 ตารางแสดงขนาดของ Input Dimension ของเวกเตอร์

Combination	Additional features	Features	Input Dimension
O	No additional features	-	30
NS	Neutral & Sexism	$t_{N,a}$, $t_{S,a}$	32
NR	Neutral & Racism	$t_{N,a}$, $t_{R,a}$	32
RS	Racism & Sexism	$t_{R,a}$, $t_{S,a}$	32
NRS	Neutral, Racism & Sexism	$t_{N,a}$, $t_{R,a}$, $t_{S,a}$	33

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ในการ Evaluate ของงานวิจัยนี้จะต้องใช้โมเดลที่มี 4 เลเยอร์คือ Input Layer, Hidden Layer, Dense Layer และ Output Layer โดยมีการทำงานดังรูปที่ 2.3



รูป 2.3 Deep Learning Model ของงานวิจัย

การทดสอบจะใช้ Dataset ที่ประกอบไปด้วยโพสต์ในทวิตเตอร์ที่เป็นโพสต์ที่แบ่งแยกเชื้อชาติจำนวน 1943 โพสต์ โพสต์ที่กีดกันทางเพศ 3166 โพสต์ และโพสต์ทั่วไปอีก 10889 โพสต์ ทดลองโดยใช้วิธี 10-Fold Cross Validation โดยแบ่งเป็น Training Set 85% และโมเดลสร้างจาก Keras และการที่จะทำให้ไม่เกิดการ Over-Fitting โมเดลจะรันได้มากสุดที่ 100 รอบ (Epochs) วิธีดังกล่าวสามารถทำให้เกิด Accuracy ที่มากที่สุด แต่ก็ยังมี Error ที่เกิดขึ้น โดยแต่ละรอบจะต่างกันประมาณ ±1 % และจำนวนรอบ (Epochs) ที่มีประสิทธิภาพที่สุดจะอยู่ที่ 30-40 รอบ (Epochs) และได้ผลลัพธ์ออกมาดังตารางที่ 2.5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

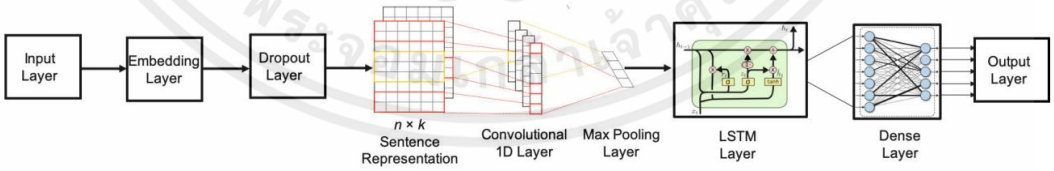
ตาราง 2.2 ตารางผลลัพธ์การทดลอง

Approach	Characteristics	Precision	Recall	F-Score
single classifier *(i)*	O	0.9175	0.9218	0.9196
single classifier *(ii)*	NS	0.9246	0.9273	0.9260
single classifier *(iii)*	NR	0.9232	0.9259	0.9245
single classifier *(iv)*	RS	0.9232	0.9264	0.9248
single classifier *(v)*	NRS	0.9252	0.9278	0.9265
ensemble *(i)*	O + NRS + NR	0.9283	0.9315	0.9298
ensemble *(ii)*	O + NRS + NS	0.9288	0.9319	0.9303
ensemble *(iii)*	O + NRS + RS	0.9283	0.9315	0.9299
ensemble *(iv)*	O + NS + RS	0.9277	0.9310	0.9293
ensemble *(v)*	O + NS + NR	0.9276	0.9308	0.9292
ensemble *(vi)*	O + RS + NR	0.9273	0.9306	0.9290
ensemble *(vii)*	NRS + NR + RS	0.9292	0.9319	0.9306
ensemble *(viii)*	NRS + NR + NS	0.9295	0.9321	0.9308
ensemble *(ix)*	NRS + NS + RS	0.9294	0.9321	0.9308
ensemble *(x)*	NS + RS + NR	0.9286	0.9314	0.9300
ensemble *(xi)*	O + NS + RS + NR + NRS	0.9305	0.9334	**0.9320**
Badjatiya et al. (2017)	LSTM + Random Embedding + GBDT	0.9300	0.9300	0.9300
Waseem and Hovy (2016)	Unsupervised List of Criteria	0.7290	0.7774	0.7391
Waseem (2016)	Unsupervised Expert annotators only	0.9159	0.9292	0.9153
Park and Fung (2017)	2 step HybridCNN (Word Vec. / Char Vec.)	0.8270	0.8270	0.8270

2.2.2 การตรวจจับผู้ใช้ที่ไม่น่าเชื่อถือ โดยใช้ Deep Learning

2.2.2.1 การขัดแยกข่าวสาร (News Classification)

การที่เราจะสามารถทำการแยกข่าวออกจากกันได้นั้น เราจะต้องใช้ Neural Network โดยเป็นการใช้ Deep Learning โดยการนำ Long Short-Term Memory (LSTM) และ Convolutional Neural Network (CNN) มาทำงานร่วมกัน ดังรูป 2.6



รูป 2.4 โครงสร้างของ News Classification

โดยก่อนที่เราจะเริ่มนั้นจะต้องมีการทำ Tokenization กับ Input ก่อน โดย Tokenization นั้นคือการแบ่งประโยคหรือบทความที่เรานำมากใช้ออกเป็นคำแยกออกจากกัน โดยได้แบ่งออกเป็น 30,000 Token ต่อชุดข้อมูล

โดยในแต่ละ Input layer และ Output Layer จะมีหน้าที่ดังนี้

ดั้งนั้นจะเริ่มจาก Embedding layer โดย Layer นี้นั้นจะขยายขนาด Vector

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ของแต่ละ input ให้ใหญ่ขึ้นโดยในส่วนแรกจะมีค่าเป็น 30,000 เท่ากับ token ที่มี
ในส่วนที่สองจะเป็น 128 หมายความว่าแต่ละ Token จะมี Vector ได้ 128 มิติ
ต่อไปก็จะเป็น Dropout Layer เป็นส่วนที่ทำให้โมเดลง่าย และธรรมดายิ่งขึ้น และยังลดการเกิด overfitting อีกด้วย

Conv1D layer เป็นส่วนที่จะทำการจัด และสลับข้อมูลเพื่อลดการซ้ำกัน หรือเหมือนกันของข้อมูล

Maxpooling1D layer จะทำการลดขนาดของ Input ลง เพื่อลดจำนวน Parameter ของโมเดลลง

LSTM layer จะนำข้อมูลที่เคย Train เก็บไว้ดึงออกมาใช้เพื่อให้เร็วและมีผลลัพธ์ที่ใกล้เคียงกันถ้าข้อมูลมีความคล้ายกันเพื่อลงการแปรผันของ Output

Dense Layer เป็นส่วนที่จะรวมข้อมูลทั้งหมดและทำการบีบข้อมูล แล้วทำการตัดสิ้นใจเลือก output ที่เหมาะสม

หลังจากที่เราสร้างโมเดลแล้วก็จะเริ่มทำการ Train Model โดยจะดูค่าเหล่านี้ประกอบการตัดสิ้นใจในการปรับเปลี่ยน

Loss, Optimizer และ Metric โดยที่

Loss จะใช้ Binary Cross-Entropy เพราะ Output มีแค่ 0 กับ 1

Optimizer จะใช้ Adam Optimizer Algorithm

Metric ดูที่ Accuracy

ส่วนสุดท้ายเป็นการแบ่ง Training Set และ Test Set โดยจะเป็น 80% และ 20% ตามลำดับ

2.2.2.2 การขัดแยกผู้ใช้ (User Profile Classification)

การขัดแยกผู้ใช้ (User Profile Classification) จะเป็นการตรวจหาสิ่งที่เกี่ยวข้องกับ user นั้น โดยจะใช้ Deep Neural Network

Neural Network Classification ในส่วนนี้จะเป็นการจัดการข้อมูลต่าง ๆ ก่อนทำโมเดลโดยที่ข้อมูลจาก Social ต่าง ๆนั้นยากต่อการนำมาใช้โดนจะต้องมี Layer ที่สร้างขึ้นเพื่อลงปัญหานี้จะเกิดขึ้นโดนสร้างตามสมการที่ 2.4

$$N_h = \frac{N_s}{\left(\alpha(N_i + N_0)\right)} \quad (2.4)$$

N_h คือ Hyperparameter ใน Dense Layer

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

N_i คือ Input N_0 คือ Output

N_s คือ จำนวน Dataset

Baseline Classifiers จะเอา Classification 3 แบบมาใช้

1) LINEAR SUPPORT VECTOR MACHINE CLASSIFIER (SVC)
2) SUPPORT VECTOR MACHINE CLASSIFIER OPTIMIZED BY STOCHASTIC GRADIENT DESCENT (SVM-SGD)
3) K-NEAREST-NEIGHBOR CLASSIFIER (KNN)

Dataset จะมี profile และ ข่าวสารต่าง ๆ โดยจะมี

1) 563,315 ข่าวสารจาก Politi-Fact.com
2) 62,367 ข่าวสารทั่วไป แบ่งเป็น
    1) 34,426 Fake News
    2) 29,938 Verified News
3) 4,022 profile แบ่งเป็น
    1) 2,013 Fake News Profile
    2) 2,008 Real News Profile

2.2.2.3 ผลลัพธ์การทดลอง

ตาราง 2.3 ตารางผลลัพธ์จากการทดลอง

	Training/Test		Cross-validation
	Accuracy	**Loss**	**Accuracy**
NN	91.47%	21.32%	92.89%
SVC	90.02%	17.79%	90.47%
SVM-SGD	89.26%	18.77%	88.51%
KNN	81.54%	28.41%	81.86%

ตาราง 2.4 ตารางผลลัพธ์จากการทดลอง

	Training/Test						Cross-validation
	Precision	**Recall**	**F1-Score**	**Accuracy**	**Average Precision**	**Loss**	**Accuracy**
SVC	90%	90%	90%	90.13%	87.34%	9.48%	91%
SVM-SGD	90%	90%	90%	90.94%	88.07%	10.91%	91%
KNN	84%	83%	84%	83.13%	81.19%	15.86%	84%

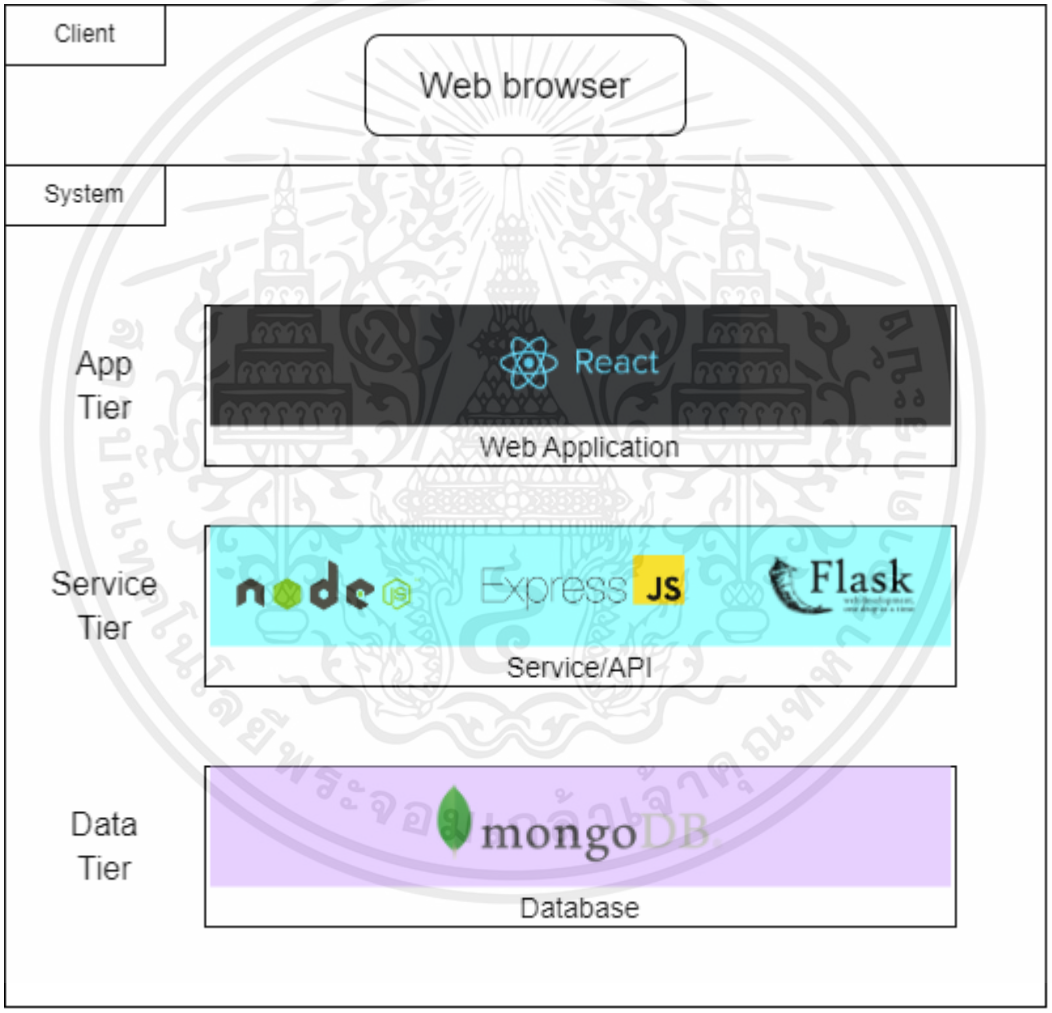
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3

การออกแบบ

ในการออกแบบและพัฒนาเว็บไซต์การระบุตัวตนของมนุษย์จากข้อความจะเป็นการพัฒนาเว็บไซต์ขึ้นมาใหม่ ให้มีรูปแบบที่ทันสมัย ใช้งานง่าย และตอบโจทย์ผู้ใช้งาน เพื่อใช้ในการค้นหาบุคคลจากข้อความที่ได้กรอกเข้าไป

3.1 ภาพรวมของระบบ



รูป 3.1 3-Tier Architecture ของระบบเว็บไซต์การระบุตัวตนของมนุษย์จากข้อความ

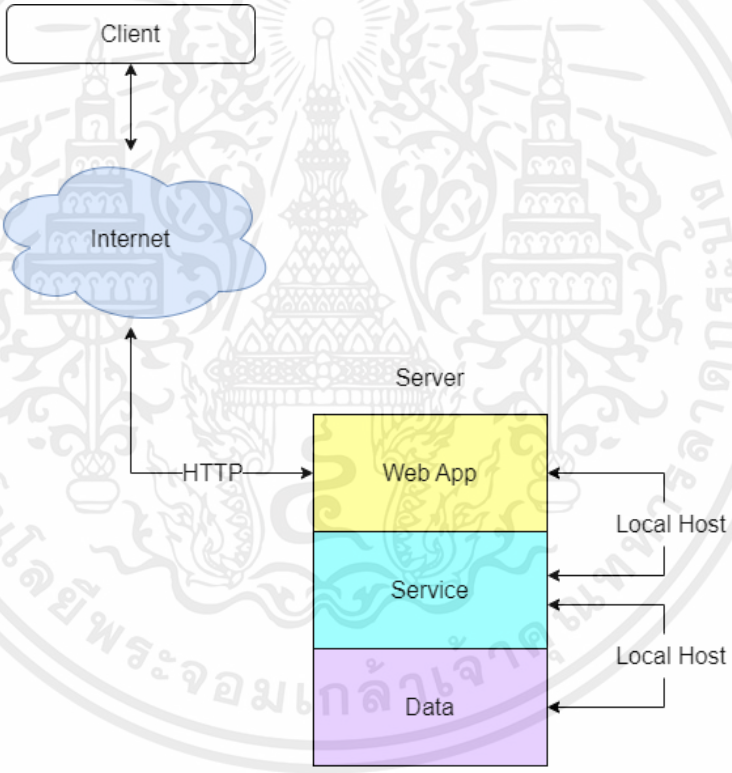
โดย 3-Tier Architecture ของระบบเว็บไซต์การระบุตัวตนของมนุษย์จากข้อความ แบ่งออกเป็น 3 ชั้นดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ชั้นที่ 1 App Tier หรือส่วนที่ติดต่อกับผู้ใช้งาน ซึ่งในส่วนนี้ จะใช้ React Framework สำหรับสร้างเว็บแอปพลิเคชันในฝั่งไคลเอ็นในรูปแบบของ HTML, CSS และ JavaScript ซึ่งง่ายต่อผู้จัดทำเพราะคุ้นเคยกับ React มาระดับนึงแล้ว

ชั้นที่ 2 Service Tier ซึ่งเป็นส่วนที่ทำหน้าที่ติดต่อ Function ประมวลผลการทำงาน และประสานงานระหว่าง Data Tier และ App Tier ซึ่ง API นี้จะใช้ Flask ซึ่งทำได้ง่าย และใช้ภาษา Python ในการทำงาน

ชั้นที่ 3 Data Tier เป็นส่วนที่ติดต่อกับฐานข้อมูล จัดเก็บข้อมูล และเข้าถึงข้อมูล โดยจะใช้ MongoDB ในการจัดการฐานข้อมูล ซึ่งเป็น NoSQL Database ทำให้จัดการได้ง่าย รวมทั้งยังสามารถจัดเก็บได้โดยการนำ Dataset ที่เป็นไฟล์นามสกุล .CSV เข้าไปใน MongoDB Compass สามารถสร้าง Collection ได้อย่างสะดวกรวดเร็ว จึงทำให้พวกเราเลือกใช้ MongoDB

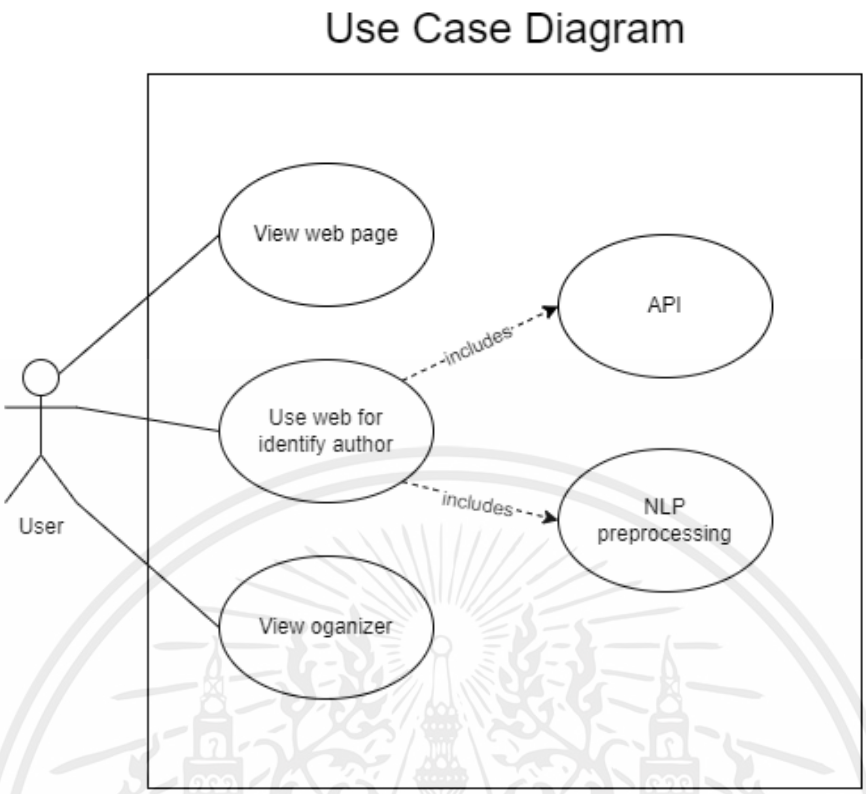


รูป 3.2 ระบบเครือข่ายแบบ Client/Server

3.2 Use Case Diagram (Overview)

เว็บไซต์การระบุตัวตนของมนุษย์จากข้อความโดยผู้ใช้งานระบบจะสามารถใช้งานเว็บไซต์ได้ดังรูปที่ 3.3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูป 3.3 แผนภาพ Use Case เว็บไซต์ระบุตัวตนของมนุษย์จากข้อความ

ตาราง 3.1 ประเภทของผู้ววใช้งานระบบ

ประเภทผู้ใช้งาน	รายละเอียด
1. User	ผู้ใช้งานทั่วไป ไม่สามารถดำเนินการภายในระบบได้

3.3 System Requirement Specification

ความต้องการของระบบเว็บไซต์การระบุตัวตนของมนุษย์จากข้อความ

ตาราง 3.2 รายการความสามารถของระบบ

ID	Detail	Type	Priority
R1	ระบบสามารถแสดงหน้าค้นหาผู้ที่เขียนบทความ	Functional	Must have
R2	ระบบสามารถแสดงรูปของเจ้าของบทความที่ทำนายได้ออกมา	Functional	Must have

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

R3	ระบบสามารถทำนายเจ้าของบทความได้อย่างถูกต้อง	Functional	Must have
R4	ระบบสามารถแสดงหน้าหลักที่มีข้อความรายละเอียดของเว็บไซต์	Non-Functional	Must have
R5	ระบบสามารถแสดงหน้าของผู้จัดทำ	Non-Functional	Should have
R6	ผู้ใช้สามารถพิมพ์ข้อความในช่องค้นหาได้	Functional	Should have
R7	ผู้ใช้สามารถกดปุ่มค้นหาเจ้าของบทความนั้นได้	Functional	Must have
R8	ผู้ใช้สามารถกดปุ่มค้นหาอีกครั้งได้หลังจากค้นหาบทความไปแล้ว	Functional	Must have
R9	ผู้ใช้สามารถกด Navigation bar ด้านบนได้	Functional	Must have

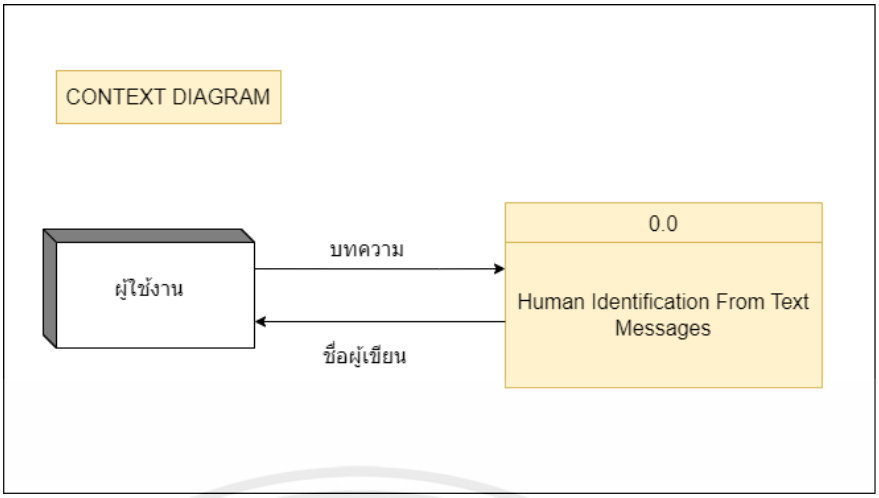
3.4 Database Design

ส่วนของฐานข้อมูลจะใช้เป็น NoSQL Database ในการเก็บข้อมูลต่าง ๆ ของโมเดลโดยใช้บริการ MongoDB เก็บข้อมูลต่าง ๆ ของโมเดลที่เราทำเอาไว้ และโมเดลของเราจะเก็บในรูปแบบของไฟล์ CSV และนำไปเก็บเป็น Collection ซึ่ง MongoDB เป็น NoSQL Database จึงทำให้เราสามารถเพิ่ม Collection หรือ Table ได้ง่ายและรวดเร็ว

3.5 Data Flow Diagram (Overview)

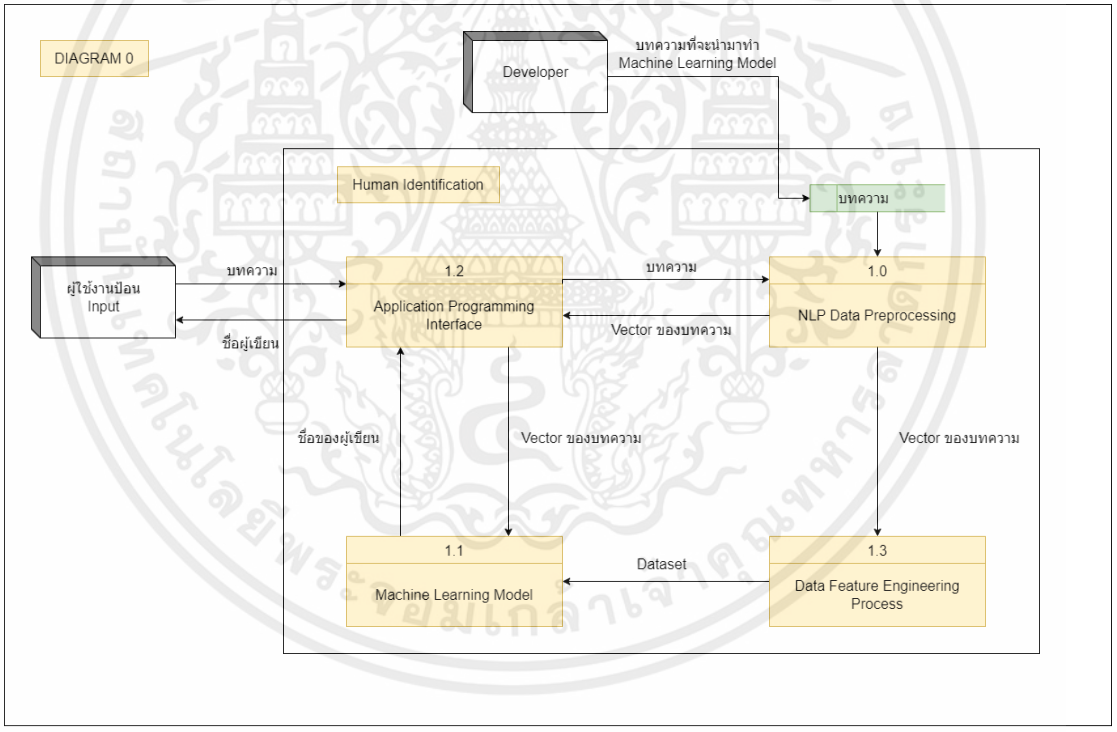
ส่วนของ Data Flow Diagram จะแสดงเป็นการทำงานของเว็บไซต์โดยมีผู้ใช้งานเข้ามาใช้งาน โดยจะแบ่งเป็น Context Diagram, Diagram 0 และ Diagram 1 ตามรูปที่ 3.4, 3.5 และ 3.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูป 3.4 รูปภาพแสดง Context Diagram

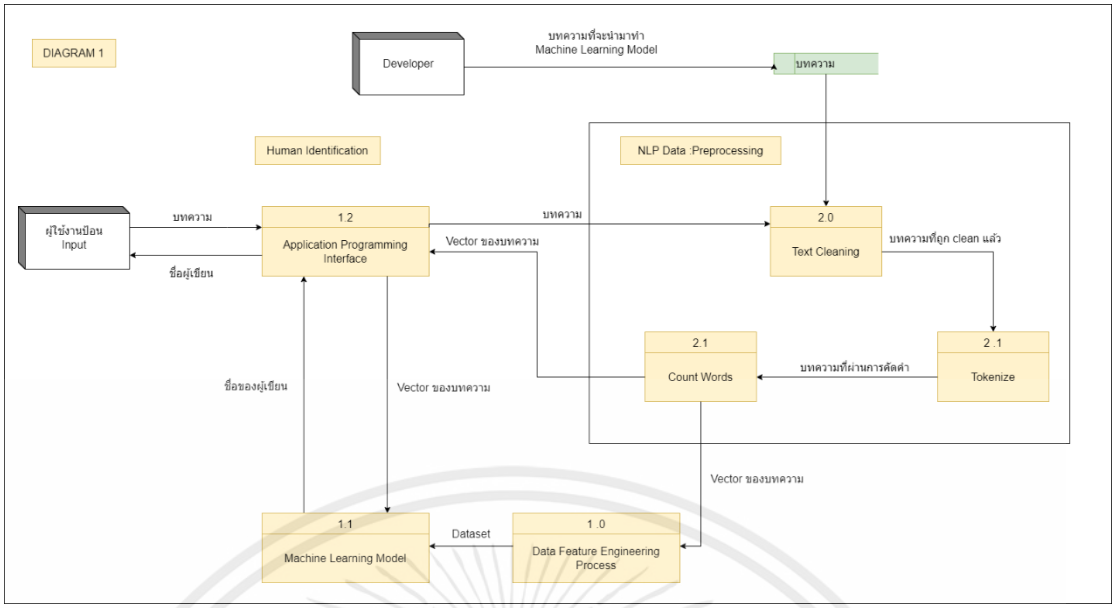
การส่งข้อความที่ใช้ตอนการตรวจสอบเข้าใน Human Identification From Text Messages แล้วส่งชื่อผู้เขียน 5 อันดับแรกสุดที่มีความคล้ายที่สุดกับ Input กลับมาให้ผู้ใช้



รูป 3.5 รูปภาพแสดง Diagram 0

Human Identification From Text Messages หลังจากรับบทความมาจากผู้ใช้ก็จะแปลงเป็น Vector ใน NLP แล้วก็ส่งกลับไปที่ Machine Learning แล้วก็ส่งชื่อผู้เขียนกลับไป ส่วนของผู้พัฒนาก็เป็นการทำ Model Machine Learning

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



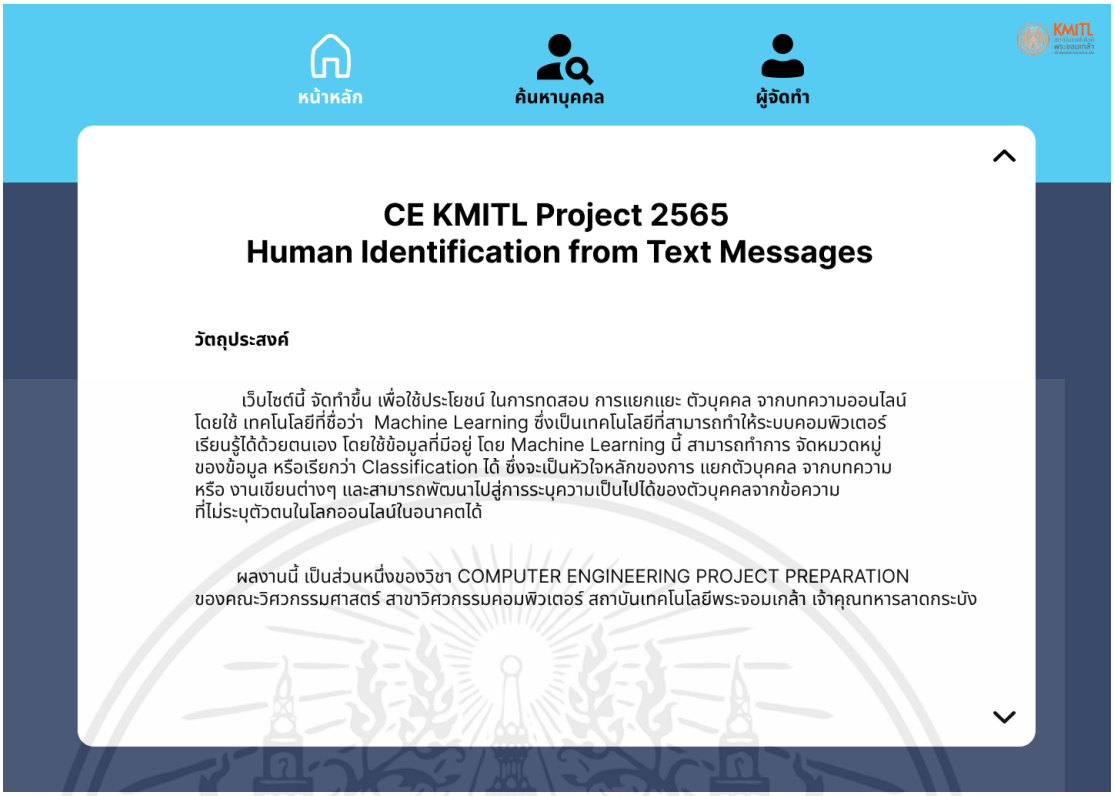
รูป 3.6 รูปภาพแสดง Diagram 1

NLP จะมีการทำงานหลักๆสามอย่างหลังรับของความมาจะมีการทำ Text Cleaning ทำให้นำไปใช้งานได้ง่ายขึ้น หลักจากนั้นจะเป็นทำ Tokenize คือการทำให้แยกประโยคออกเป็นคำ แล้วส่วนสุดท้ายจะเป็นการเปลี่ยนข้อมูลเป็น Vector แล้วส่งกลับไป

3.6 User Interface

หน้าเว็บไซต์การระบุตัวตนของมนุษย์จากข้อความมีการออกแบบให้ใช้งานง่าย และมีสีที่ตัดกับตัวหนังสือได้ชัดเจนเพื่อความอ่านง่าย ทำให้มีฟังก์ชันหลากหลายในการทำโดยมีหลาย Model ในการทำนาย เพื่อให้มีความหลากหลายต่อผู้ใช้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูป 3.7 หน้าหลัก

รูป 3.8 หน้าค้นหาผู้ที่เขียนบทความ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

รูป 3.9 หน้าแสดงรูปเจ้าของบทความที่ทำนายออกมาได้

รูป 3.10 หน้าแสดงผู้จัดทำ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.7 การออกแบบโมเดล

3.7.1 การ Train Model โดย Machine Learning

เลือก Machine Learning มา 3 เทคนิคได้แก่ Support Vector Machine, Logistic Regression, Decision Tree และนำตารางข้อมูลที่มีไปให้ Machine Learning นำไปเรียนรู้ เพื่อเปรียบเทียบโมเดลที่ดีที่สุด

ตาราง 3.3 ค่า Accuracy ของ Machine Learning Model

Machine learning model	Accuracy
Support Vector Machine	89%
Logistic Regression	90%
Decision Tree	78%

3.7.2 การ train model โดย deep learning

ในส่วนของ Deep learning นั้นหลังจากที่เราได้ทำ Machine Learning แล้วเราจึงได้ตัดสินใจในการทำ Model Deep Learning เพื่อนำมาเปรียบเทียบและเราคาดว่า Deep Learning น่าจะเหมาะกับงานนี้มากกว่า Machine Learning

3.7.3 Data Preparation

3.7.3.1 Text Tokenizer

นำ Dataset ที่มีอยู่มาทำการตัดคำออกจากกันและแบ่งแต่ละผู้เขียนออกจากกัน

3.7.3.2 NLP: Doc2Vec

เป็นการเปลี่ยนจากคำแต่ละคำที่เราตัดออกจากกันมาไปแปลงให้เป็น Vector เพื่อให้เหมาะแก่การนำไปใช้ Train Model

3.7.4 การประเมินค่าและวัดผลการทำงานของ Model

ทำการประเมินผลโดยการนาการทดลองที่ 3 มาวัดประสิทธิภาพโมเดล นอกจากจะประเมินด้วยค่า Loss และ Accuracy ทางคณะผู้จัดทำจะทำการประเมินบนพื้นฐาน Precision, Recall, F-Score และ Confusion Matrix

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4

ผลการทดลอง

4.1 ผลการทดลองโมเดลสำหรับทำนายผู้เขียน

4.1.1 ผลการทดลอง Machine Learning Model

จากการศึกษาและทดลองโมเดลต่างๆดูแล้ว ได้ข้อสรุปว่า Logistic Regression นั้นมีผลลัพธ์ที่ดีที่สุดเทียบกับ Decision Tree และ Support Vector Machine แต่ในการทดลองของเรานั้นเป็นโมเดลขนาดเล็กๆเวลาใช้ Dataset ที่มากขึ้นอาจจะมีผลลัพธ์ที่แย่ลง เราเลยจะทำ Logistic Regression หลายโมเดลให้เท่ากับ Class ทั้งหมด 109 Class แล้วเอาผลจากทั้งหมดนั้นมาเทียบกัน Dataset มีทั้งหมด 109 Class แต่เราจะทำการแบ่งเป็นจากทั้งหมดไม่ได้ เพราะแต่ละ Class มีข้อมูลไม่เท่ากัน เราจึงจะแบ่งแต่ละ Class เป็น Train Set 80% และ Test Set 20%

ตาราง 4.1 ตารางแสดงค่าผลลัพธ์ต่างๆของโมเดล Logistic Regression

	Precision	Recall	F1 score	Test set
Author1	0.83	0.89	0.86	500
Author2	0.94	0.91	0.93	1000
Macro Avg	0.89	0.90	0.89	1500
Weighted Avg	0.91	0.90	0.90	1500
Accuracy	0.90			1500

ตาราง 4.2 ตาราง Confusion Metrix ของโมเดล Logistic Regression

	Actual Positive	Actual Negative
Predicted Positive	445	55
Predicted Negative	90	910

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตาราง 4.3 ตารางแสดงค่าผลลัพธ์ต่างๆของโมเดล Decision Tree

	Precision	Recall	F1 Score	Test Set
Author1	0.66	0.73	0.69	500
Author2	0.86	0.81	0.83	1000
Macro Avg	0.76	0.77	0.76	1500
Weighted Avg	0.79	0.78	0.79	1500
Accuracy	0.78			1500

ตาราง 4.4 ตาราง Confusion Metrix ของโมเดล Decision Tree

	Actual Positive	Actual Negative
Predicted Positive	367	133
Predicted Negative	191	809

ตาราง 4.5 ตารางแสดงค่าผลลัพธ์ต่างๆของโมเดล SVM

	Precision	Recall	F1 Score	Test Set
Author1	0.81	0.87	0.84	500
Author2	0.93	0.90	0.92	1000
Macro Avg	0.87	0.89	0.88	1500
Weighted Avg	0.89	0.89	0.89	1500
Accuracy	0.89			1500

ตาราง 4.6 ตาราง Confusion Metrix ของโมเดล SVM

	Actual Positive	Actual Negative
Predicted Positive	436	64
Predicted Negative	100	900

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หลังจากการที่ได้ลองนำ Machine Learning Model มาเทียบกัน เพื่อที่จะเลือกโมเดลที่ดีที่สุด ซึ่งในตอนแรกพวกเราตัดสินใจจะใช้ Logistic Regression เพราะได้ผลลัพธ์ที่ดีที่สุด แต่โมเดลของเรานั้นมี 109 Class ถ้าจะให้ทำเป็นโมเดลเดียว อาจจะไม่มีประสิทธิภาพมากพอเราเลยจะทำโมเดลให้สำหรับแต่ละ Class ดังนั้นพอเราได้ทดลองทำจริงแล้ว พวกเราเลยตัดสินใจว่า เราจะใช้ SVM แทนเพราะเหมาะสำหรับโมเดลที่มี Class เดียว

ตาราง 4.7 ตารางแสดงค่าผลลัพธ์ต่างๆของโมเดล SVM

F1 Score	0.689
Recall	0.548
Precision	0.927

ตาราง 4.8 ตาราง Confusion Metrix ของโมเดล SVM

	Actual Positive	**Actual Negative**
Predicted Positive	597640	492360
Predicted Negative	47089	46148

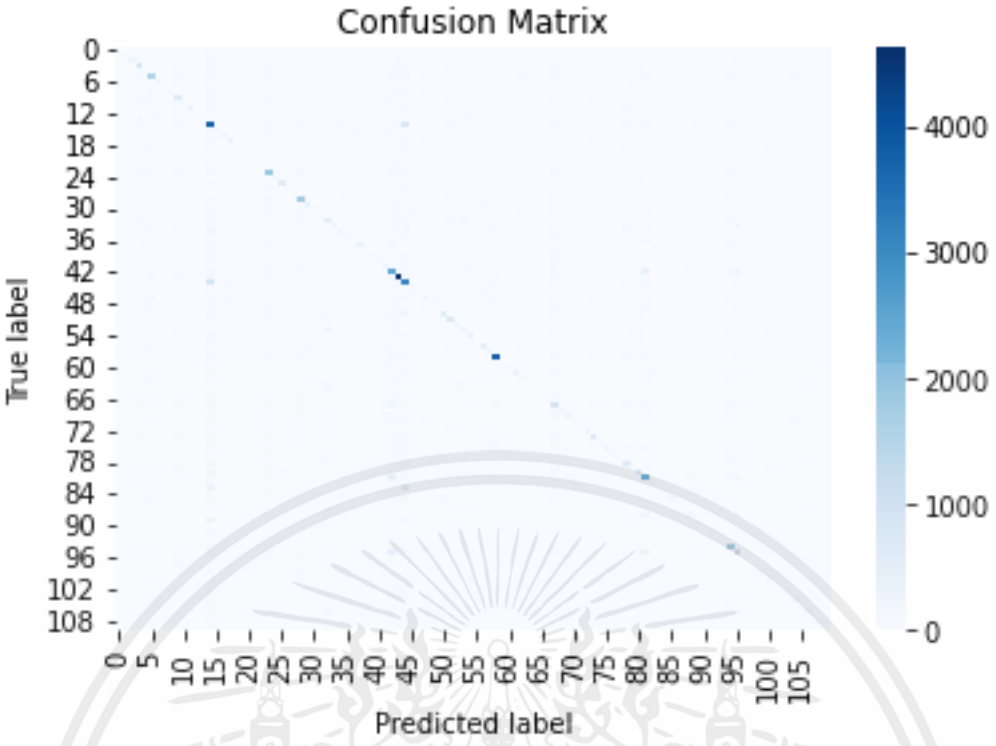
4.1.1 ผลการทดลอง Deep Learning Model

หลังจากที่เราได้ทำโมเดล SVM ได้สำเร็จแล้วเราได้ตัดสินใจที่จะทำ deep learning model แต่ในรอบนี้เราไม่ได้ทำการเปรียบเทียบเหมือน Machine learning model ดังนั้นเราจึงเลือกใช้โมเดล LSTMโดยที่แบ่ง เป็น 80% Train Set กับ Validation ในส่วน 20% ที่เหลือเป็นของ Test Set

ตาราง 4.9 ตารางแสดงค่าผลลัพธ์ต่างๆของโมเดล LSTM

F1 Score	0.418
Recall	0.456
Precision	0.467

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูป 4.1 กราฟ Confusion Metrix ของโมเดล LSTM

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการทดลอง

5.1 สิ่งที่ดำเนินการไปแล้ว

5.1.1 ส่วนการออกแบบระบบ

ทำการออกแบบโมเดล และออกแบบระบบของเว็บไซต์ในส่วนของ Front-End, Back-End และ Database ซึ่งสามารถเขียนเป็นข้อได้ ดังนี้

1) ศึกษาการออกแบบ User Interface สำหรับแอปพลิเคชันบนเว็บไซต์
2) ออกแบบ User Interface ทั้งหมดของระบบ รวมถึงจัดทำ Prototype ของเว็บไซต์
3) ออกแบบโมเดลต่างๆได้แก่ Logistic Regression, Decision Tree และ SVM
4) ออกแบบ Use Case Diagram ทั้งในส่วนของ Overview ซึ่งแสดงฟังก์ชันหลักของแอปพลิเคชันและส่วนของ Detail ที่แสดงฟังก์ชันทั้งหมดของแอปพลิเคชัน
5) ออกแบบ Data Flow Diagram ในส่วนของ Overview ซึ่งแสดงการทำงานของเว็บไซต์
6) ออกแบบโมเดลต่างๆได้แก่ Logistic regression, Decision Tree และ SVM
7) ออกแบบ 3-Tier Application Architecture

5.1.2 ส่วนการพัฒนาเว็บแอปพลิเคชัน

เริ่มจากการศึกษาเครื่องมือต่างๆที่ใช้ในการพัฒนา, ศึกษาการใช้ React, ศึกษาการใช้งาน Flask, ศึกษาการใช้งาน MongoDB, ศึกษาการใช้งานฐานข้อมูล NoSQL Database ในการพัฒนาฐานข้อมูล

จากนั้นทำการพัฒนาส่วนต่างๆของแอปพลิเคชัน ทั้งในส่วนของ Front-End, Back-End และ Database โดยพัฒนาในส่วนของฟังก์ชันหลักเสร็จทั้งหมด และทำการแก้ไขจุดต่างๆ เมื่อเจอข้อผิดพลาดของระบบ

5.1.3 ส่วนการพัฒนาโมเดล

เริ่มจากการจัดการ dataset ที่หามาได้โดยตัดผู้เขียนที่มีจำนวนข้อมูลน้อยๆออกไป, ศึกษาวิธีการทำโมเดลแต่ละประเภท, ศึกษาเทคนิคและวิธีที่สามารถนำมาทำโมเดลที่มี Dataset จำนวนเยอะ, คำนวณหาโมเดลที่มีประสิทธิภาพหรือค่า Accuracy เยอะที่สุด ซึ่งจากการทดลองของพวกเรานั้นพวกเราได้นำ Machine Learning หลากหลายตัวมาเปรียบเทียบกันแล้วได้ข้อสรุปว่า Logistic Regression จะเป็นอันที่ดีที่สุดในอันที่พวกเรานำมาทดลอง แต่หลังจากที่เราได้ลองนำ Logistic Regression มาลองใช้ดูแล้ว Logistic Regression Model นี้นั้นไม่เหมาะการวิธีการที่เราจะนำมาใช้ ดังนั้นแล้วจึงนำ SVM ที่ดีรองลงมามาใช้แทน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

พวกเราได้ตัดสินใจที่เราจะทำ Deep Learning Model ด้วยโดยเราได้ตัดสินใจที่จะใช้ Long Short-Term Memory(LSTM) เพื่อจะได้นำมาเปรียบเทียบกับกันระหว่างโมเดล และเราคาดว่า LSTM นั้นน่าจะประสิทธิภาพมากกว่า SVM

5.1.4 ส่วนการรวมระบบ

นำ Model ที่ Train ไว้ทั้งหมดมาใช้ใน Back-End(Flask) จากนั้นใช้ Method POST และ GET เพื่อนำข้อมูลจาก Database(MongoDB) มาแสดงผลในส่วน Front-End(React) ของหน้าแรก และตัวอย่างบทความของแต่ละคนโดยใช้การสุ่มมา ซึ่งตอนโหลดข้อมูลเราได้ใช้ตัว Loading เพื่อเป็นการทำให้ User รู้ว่ากำลังรอข้อมูลที่จะมาแสดงผลอยู่

5.1.5 ส่วนการ Deployment

ในส่วนของ Front-End(React) Deploy ผ่าน Netify ในส่วนของ Back-End(Flask) Deploy ผ่าน Amazon Web Service(AWS) และ Database ใช้ MongoDB Atlas แล้วทำการทดลองผ่านหน้าเว็บไซต์

5.2 ปัญหาและแนวทางการแก้ไข

1) คณะผู้จัดทำใช้รูปแบบการทำงานแบบ Agile ที่มีการวางแผนชั่วโมงในการทำงานในแต่ละ Sprint ซึ่งในบาง Sprint มี Task ที่ไม่สามารถปิดได้ เนื่องจากเวลาไม่เพียงพอ รวมไปถึงการที่ผู้จัดทำมีธุระ หรือไม่พร้อมในการทำงานเนื่องจากไม่สบายเป็นต้น แนวทางในการแก้ไขปัญหาคือ คณะผู้จัดทำต้องพยายามจัดสรรชั่วโมงการทำงานให้มีความเหมาะสมของงานให้มากขึ้น
2) MongoDB และ Flask เป็น NoSQL Database และ Framework ที่ไม่มีความคุ้นเคยมาก่อน ทำให้การพัฒนาเว็บแอปพลิเคชันในช่วงแรกมีความล่าช้า
3) เครื่องคอมพิวเตอร์ที่ใช้ในการพัฒนาแอปพลิเคชันมีทรัพยากร ที่ไม่เพียงพอต่อการทำงาน ทำให้เกิดความล่าช้า, เกิดอาการเครื่องค้าง และไม่มีความต่อเนื่องในการทำงาน
4) ในการพัฒนาเว็บแอปพลิเคชันอาจจะมีส่วนที่ผิด เราเลยทำการ Testing หลายๆรอบ แต่เมื่อ code มีการเปลี่ยนแปลงจึงทำให้ส่งผลต่อการ Testing เก่าๆด้วย เลยทำให้ส่งผลทำให้งานเพิ่มขึ้นเป็นทวีคูณ และทำให้งานล่าช้า ทางคณะผู้จัดทำจึงแก้ไขปัญหาด้วยการทำ Testing เมื่อ Code เริ่มคงที่แล้ว
5) การเทรนโมเดลใช้เวลา และทรัพยากรค่อนข้างมาก จนทำให้เกิดความล่าช้า

5.3 ข้อเสนอแนะในการพัฒนาต่อ

5.3.1 ส่วนการพัฒนาแอปพลิเคชัน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1) เพิ่ม Model ใหม่ในการทำนายผู้เขียน เพื่อที่จะทำให้ผู้ใช้รู้ว่า Model ที่ดีที่สุดควรจะใช้ Model ไหน

2) เพิ่ม/ลด และปรับจูนพารามิเตอร์เพื่อให้ผลลัพธ์ของแต่ละ Model ออกมาดีขึ้น

3) เพิ่มฟังก์ชันสแกน QR Code เพื่อใช้ในการสแกนแล้วเปิดหน้าหาข้อมูลได้เลย

4) เพิ่ม Platform ในมือถือ เพื่อจะได้ใช้ได้ง่ายขึ้น และสามารถเข้าถึงได้ง่ายขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บรรณานุกรม

Georgios K. Pitsilis, Heri Ramampiaro, Helge Langseth. 2018. “Detecting Offensive Language in Tweets Using Deep Learning.” Department of Computer Science, Norwegian University of Science and Technology

Giuseppe Sansonetti, Fabio Gasparetti, Giuseppe D’Aniello, Alessandro Micarelli. 2020. “Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection.” Department of Computer Science, Norwegian University of Science and Technology

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ภาคผนวก ก
การดำเนินงานตามปริมาณขั้นต่ำ

ตาราง ก.1 ตารางปริมาณงานที่ได้ดำเนินการไปในช่วงโครงงาน 1

รายการ	ปริมาณงาน	หมายเหตุ
นำบทความทั้งหมดมาทำ Text Cleaning	100%	
นำบทความทั้งหมดมา Tokenize (ตัดคำ)	100%	
นำบทความทั้งหมดมาแยก Feature	100%	
ทำการ Count Word คำที่กรองมา	100%	
ทำการเทรนโมเดล Logistic Regression	100%	
ทำการเทรนโมเดล Decision Tree	100%	
ทำการเทรนโมเดล SVM	100%	
ทำการเทรนโมเดลอื่นๆ	100%	
คำนวณค่าความแม่นยำของแต่ละโมเดล	100%	
หา CNN ที่เหมาะสม	100%	
ปรับจูน Parameter	100%	
ทำ Front-End ของเว็บแอปพลิเคชัน	100%	
ทำ Back-End ของเว็บแอปพลิเคชัน	100%	
เชื่อมต่อ Front-End และ Back-End	100%	
Deploy ระบบเว็บแอปพลิเคชันลง Server	100%	
รวม	100%	

จากตารางแสดงรายละเอียดงานและปริมาณงานเฉพาะในส่วนของการพัฒนาโมเดลและเว็บแอปพลิเคชัน จะเห็นได้ว่ามีความก้าวหน้าอยู่ที่ 100%

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหาและต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้