การขุดเหมืองข้อความและการวิเคราะห์ความรู้สึกสำหรับการลงทุนในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอ

TEXT MINING AND SENTIMENT ANALYSIS FOR CRYPTOCURRENCY INVESTMENT USING COMMENTS FROM YOUTUBE

ภานุมาศ ทุกข์จาก

PHANUMART TOOKJAK

การค้นคว้าอิสระนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูลและการวิเคราะห์
ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง คณะวิทยาศาสตร์
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
2565

KMITL-2022-SC-M-017-125

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

TEXT MINING AND SENTIMENT ANALYSIS FOR CRYPTOCURRENCY INVESTMENT USING COMMENTS FROM YOUTUBE

PHANUMART TOOKJAK

AN INDEPENDENT STUDY SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE DEGREE OF MASTER OF SCIENCE IN
DATA SCIENCE AND ANALYTICS
KMITL-DIGITAL ANALYTICS AND INTELLIGENCE CENTER SCHOOL OF SCIENCE
KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG
2022
KMITL-2022-SC-M-017-125

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

COPYRIGHT 2022

SCHOOL OF SCIENCE

KING MONGKUT'S INSTITUTE OF TECHNOLOGY LADKRABANG

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

หัวข้อการค้นคว้าอิสระ การขุดเหมืองข้อความและการวิเคราะห์ความรู้สึกสำหรับการลงทุนในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอ

ชื่อนักศึกษา นางสาวภานุมาศ ทุกข์จาก

รหัสประจำตัว 63605094

ปริญญา วิทยาศาสตรมหาบัณฑิต (วิทยาการข้อมูลและการวิเคราะห์) ศูนย์วิเคราะห์ข้อมูลดิจิทัลอัจฉริยะพระจอมเกล้าลาดกระบัง

พ.ศ. 2565

อาจารย์ที่ปรึกษาการค้นคว้าอิสระ ผู้ช่วยศาสตราจารย์ ดร. กุลสวัสดิ์ จิตขจรวานิช

บทคัดย่อ

การตัดสินใจที่จะลงทุนในสกุลเงินดิจิทัลนั้นมีหลายปัจจัย ซึ่งความคิดเห็นบนโลกออนไลน์ก็เป็นปัจจัยหนึ่งที่มีผลต่อการตัดสินใจในการลงทุน ทำให้การค้นคว้าอิสระนี้นำการขุดเหมืองข้อความและการวิเคราะห์ความรู้สึกมาใช้ร่วมกับการสร้างแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network) เพื่อใช้สำหรับการวิเคราะห์ความคิดเห็นในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอ เพื่อช่วยแนะนำและใช้ประกอบการตัดสินใจในการลงทุนในสกุลเงินดิจิทัล ซึ่งประกอบไปด้วยความคิดเห็นทั้งหมด 7,795 ความคิดเห็น จาก 47 วีดีโอ โดยให้แบบจำลองทำนายความคิดเห็นออกมาว่าเป็นความคิดเห็นเชิงบวกหรือความคิดเห็นเชิงลบ พบว่าแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับแบบ BERT (Bidirectional Encoder Representations from Transformer) ซึ่งให้ค่าความถูกต้องมากกว่าแบบอื่นและมีค่าอยู่ที่ 76 – 79 เปอร์เซ็นต์ นอกจากนี้ยังมีการนำความคิดเห็นที่ทำนายมาแล้วไปประกอบเป็นคุณลักษณะหนึ่งในการนำคุณลักษณะต่าง ๆ ของวีดีโอที่มีการพูดถึงสกุลเงินดิจิทัลทั้งหมด 15 คุณลักษณะ จึงใช้เทคนิคการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) เพื่อลดจำนวนตัวแปรที่จะใช้ประกอบการตัดสินใจในการลงทุนในสกุลเงินดิจิทัล จากการทดลองพบว่าใช้ Components เท่ากับ 4 จะสามารถอธิบายความแปรปรวน และสามารถอธิบายข้อมูลทั้งหมดได้ 78.05 เปอร์เซ็นต์

คำสำคัญ : การวิเคราะห์ความรู้สึก โครงข่ายประสาทเทียมแบบวนกลับ สกุลเงินดิจิทัล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Independent Study Title	Text Mining and Sentiment Analysis For Cryptocurrency Investment using Comments from YouTube
Student Name	Miss Phanumart Tookjak
Student ID	63605094
Degree	Master of Science (Data Science and Analytics)
	KMITL-Digital Analytics and Intelligence Center
Year	2022
Idependent Study Advisor	Asst. Prof. Dr. Kulsawasd Jitkajornwanich

ABSTRACT

Decision on investing in cryptocurrency has dynamic factors. Comments on the internet is one of the factor that influences the decision on cryptocurrency's investment. Independent study leads to texting mining and sentiment analysis, which are combined to construct the model of recurrent neural network for cryptocurrency related comment on youtube's analysis. This model aims to advise and used for supporting the decision-making on cryptocurrency investment. The sample group are 7,795 comments from 47 videos on youtube. The method is to predict the comment to be positive or negative by using the model in different approaches. The result shows that the model of the recurrent neural network named the bidirectional encoder representations from transformers, which the accuracy is more precise than other model,and the result level is 76-79 percent. Moreover,the predicted comments have been assembled as one of features. The research has categorized the videos that talk about cryptocurrency into 15 features. Then,the principal component analysis is used for reducing the amount of variable that is used to make investment decision. The result reveals that the component level is 4, explain variance and describes all information as 78.05 percent.

Keywords : Sentiment Analysis, Recurrent Neural Network, Cryptocurrency

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

กิตติกรรมประกาศ

การค้นคว้าอิสระครั้งนี้จัดทำขึ้นเพื่อศึกษาวิธีการดำเนินงานเพื่อวิเคราะห์ความรู้สึกสำหรับการลงทุนในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอ ด้วยโครงข่ายประสาทเทียมแบบวนกลับ โดยการค้นคว้าอิสระครั้งนี้ สำเร็จลุล่วงไปได้ด้วยดี เนื่องจากได้รับคำแนะนำด้านวิชาความรู้ คำปรึกษา มาจากคณาจารย์หลายท่าน ซึ่งข้าพเจ้าขอขอบพระคุณทุกท่านไว้ ณ ที่นี้

ข้าพเจ้าขอขอบพระคุณ ผศ.ดร.กุลสวัสดิ์ จิตขจรวานิช ที่กรุณารับเป็นอาจารย์ที่ปรึกษาการค้นคว้าอิสระครั้งนี้โดยให้คำปรึกษา คำแนะนำด้านวิชาความรู้ กำลังใจ ตลอดจนการปรับปรุงแก้ไขข้อบกพร่องและให้ความช่วยเหลือในด้านต่าง ๆ เมื่อข้าพเจ้าพบปัญหาในระหว่างการดำเนินการค้นคว้าอิสระนี้ให้สำเร็จและมีความสมบูรณ์มากยิ่งขึ้น

ขอขอบคุณ นายกิตติศักดิ์พัฒนา ราเชาว์ดี ที่กรุณาให้คำปรึกษาในด้านโปรแกรมและสละเวลาให้ คำแนะนำรวมถึงการชี้แนะแนวทางแก้ไขจนการค้นคว้าอิสระครั้งมีความสมบูรณ์มากยิ่งขึ้น

สุดท้ายนี้ข้าพเจ้าขอกราบขอบพระคุณบิดามารดาและครอบครัวของข้าพเจ้าที่ได้ให้ชีวิตและโอกาสทางการศึกษา สนับสนุน เป็นกำลังใจและให้ความห่วงใยเสมอมา ตลอดจนคณะอาจารย์ทุกท่านที่ให้วิชาความรู้อันเป็นประโยชน์แก่ข้าพเจ้า ให้ข้าพเจ้าสามารถทำการค้นคว้าอิสระนี้สำเร็จลุล่วงไปได้ด้วยดี

ภานุมาศ ทุกข์จาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญ

	หน้า
บทคัดย่อภาษาไทย	ก
บทคัดย่อภาษาอังกฤษ	ข
กิตติกรรมประกาศ	ค
สารบัญ	ง
สารบัญตาราง	ฉ
สารบัญรูป	ช
บทที่ 1 บทนำ	1
1.1 ความเป็นมาและความสำคัญของปัญหา	1
1.2 วัตถุประสงค์ของงานวิจัย	1
1.3 ขอบเขตของงานวิจัย	2
1.4 ประโยชน์ที่คาดว่าจะได้รับ	2
1.5 ขั้นตอนการดำเนินงาน	2
บทที่ 2 ทฤษฎีและงานวิจัยที่เกี่ยวข้อง	3
2.1 สกุลเงินดิจิทัล	3
2.2 การทำเหมืองข้อความและการวิเคราะห์ความรู้สึก	3
2.3 โครงสร้างประสาทเทียมแบบวนกลับ	5
2.3.1 Long Short-Term Memory	6
2.3.2 Gated Recurrent Units	7
2.3.3 Bidirectional Encoder Representations from Transformer	7
2.4 การวิเคราะห์องค์ประกอบหลัก	8
2.5 การวิเคราะห์การถดถอยโลจิสติก	10
2.6 งานวิจัยที่เกี่ยวข้อง	11
บทที่ 3 วิธีการดำเนินงานวิจัย	12
3.1 วิธีการศึกษา	13
3.2 ขั้นตอนการรวบรวมข้อมูลและศึกษาชุดข้อมูล	13
3.3 ขั้นตอนการจัดเตรียมข้อมูล	16
3.3.1 ขั้นตอนการติดฉลากข้อมูล	16
3.3.2 ขั้นตอนการทำความสะอาดข้อมูล	17
3.3.3 ขั้นตอนการแปลงข้อมูล	19
3.3.3.1 การตัดคำ (Tokenize)	19

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.3.2 การสร้างคลังคำศัพท์ (Bag of words) 20
3.3.3.3 หาความยาวของประโยค 20
3.3.3.4 แปลงคำเป็นตัวเลข 21
3.3.3.5 ทำการ Padding 21
3.3.3.6 แปลงผลเฉลยให้เป็นตัวเลข 22
3.3.3.7 การเข้ารหัสผลเฉลยแบบวันฮอท (One-Hot Encoding) 22
3.3.4 ขั้นตอนการเพิ่มคุณลักษณะใหม่จากข้อมูล 23
3.4 ขั้นตอนการสร้างแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ 24
3.5 ขั้นตอนการติดฉลากแบบอัตโนมัติ 26
3.6 ขั้นตอนการวิเคราห์องค์ประกอบหลักและการสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก 27
3.6.1 การวิเคราห์องค์ประกอบหลัก (Principal Component Analysis) 28
3.6.2 การสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก (Logistic regression) 30
บทที่ 4 ผลการวิจัยและการอภิปรายผล 31
4.1 ผลการวิจัยของแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ 31
4.2 ผลการวิจัยของการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติก 39
4.3 อภิปรายผล 40
บทที่ 5 สรุปผลการวิจัยและข้อเสนอแนะ 41
5.1 สรุปผลการวิจัย 40
5.2 ขอบเขตและข้อจำกัด 42
5.3 ข้อเสนอแนะ 42
เอกสารอ้างอิง 43
ประวัติผู้เขียน 45

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญตาราง

ตารางที่	หน้า
3.1 ตารางอธิบายชื่อคุณลักษณะของวีดีโอที่ได้จากการดึง API	13
3.2 ตารางอธิบายชื่อคุณลักษณะข้อมูลความคิดเห็นของวีดีโอที่ได้จากการดึง API	14
3.3 ตารางอธิบายชื่อคุณลักษณะใหม่	23
3.4 แสดงการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM	25
3.5 แสดงการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ GRU	25
3.6 แสดงการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ BERT	26
3.7 ตารางแสดงคุณลักษณะตั้งต้นและผลเฉลย	27
3.8 แสดงการสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก	30
4.1 เปรียบเทียบประสิทธิภาพแบบจำลองแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ	39
4.2 เปรียบเทียบประสิทธิภาพแบบจำลองLogistic Regression	39

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

สารบัญรูป

รูปที่	หน้า
2.1 โครงสร้างของโครงข่ายประสาทเทียมแบบวนกลับ	5
2.2 ส่วนประกอบของหน่วยความจำระยะสั้นแบบยาว	6
2.3 ส่วนประกอบของ Gated Recurrent Units	7
2.4 ส่วนประกอบของBidirectional Encoder Representations from Transformer	8
2.5 รูปแบบของการลดคุณลักษณะ	9
2.6 ตัวอย่างลำดับความแปรปรวนของแต่ละองค์ประกอบ	9
2.7 Standard Logistic Function	10
3.1 ลำดับขั้นตอนการดำเนินงาน	12
3.2 แสดงข้อมูลของรายละเอียดวีดีโอที่ดึงจากจากเว็บไซต์คลังคลิปวิดีโอ	14
3.3 แสดงความคิดเห็นบางส่วนของวีดีโอที่ดึงจากจากเว็บไซต์คลังคลิปวิดีโอ	15
3.4 แสดงราคาเหรียญบิตคอยต์บางส่วน	16
3.5 ตัวอย่างความคิดเห็นที่มีผลเฉลยเป็นความรู้สึกเชิงบวก	17
3.6 ตัวอย่างความคิดเห็นที่มีผลเฉลยเป็นความรู้สึกเชิงลบ	17
3.7 เครื่องหมายวรรคตอนที่ทำการตัดออกจากประโยคความคิดเห็น	18
3.8 ตัวอย่างคำที่ไม่มีความสำคัญภาษาอังกฤษ	18
3.9 ตัวอย่างคำที่ไม่มีความสำคัญภาษาไทย	19
3.10 ตัวอย่างประโยคที่ทำความสะอาด	19
3.11 ตัวอย่างประโยคที่ทำการตัดคำ	20
3.12 ตัวอย่างคำในคลังคำศัพท์	20
3.13 ความยาวสูงสุดของคำในประโยค	20
3.14 ตัวอย่างประโยคที่แปลงคำเป็นตัวเลข	21
3.15 ตัวอย่างประโยคที่ทำการ padding	21
3.16 แปลงผลเฉลยให้เป็นตัวเลข	22
3.17 เข้ารหัสผลเฉลยแบบวันฮอท	22
3.18 แสดงตัวอย่างข้อมูลของคุณลักษณะใหม่	23
3.19 สถาปัตยกรรมโครงข่ายประสาทเทียมแบบวนกลับ	24
3.20 ติดฉลากแบบอัตโนมัติด้วยโมเดลที่ฝึกอบรมมาแล้ว	26
3.21 แสดงตัวอย่างชุดข้อมูลที่ใช้ในการประกอบการตัดสินใจลงทุนในสกุลเงินดิจิทัล	28
3.22 แสดงตัวอย่างข้อมูลที่ทำ Standardization	28

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.23 กราฟ Scree plot 29
4.1 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน 32
4.2 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน 32
4.3 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบลบอีโมติคอน 33
4.4 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบลบอีโมติคอน 33
4.5 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน 34
4.6 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน 35
4.7 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบลบอีโมติคอน 35
4.8 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบลบอีโมติคอน 36
4.9 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน 37
4.10 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน 37
4.11 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบลบอีโมติคอน 38
4.12 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบลบอีโมติคอน 38

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันสกุลเงินดิจิทัล (Cryptocurrency) (1) ถือว่าเป็นสินทรัพย์ที่กำลังมาแรงและได้รับความนิยมเป็นอย่างมาก ซึ่งมีราคากลางในการซื้อขายแปรผันตามกลไลตลาด ไม่มีศูนย์กลางหรือตัวกลาง สามารถแลกเปลี่ยนมูลค่าหรือโอนย้ายข้อมูลต่าง ๆ จากต้นทางถึงปลายทางโดยตรงผ่านเครือข่ายอินเตอร์เน็ต และมีการบันทึกการทำธุรกรรมทุกอย่างที่เกิดขึ้นไว้ในแบบที่ไม่สามารถแก้ไขข้อมูลได้ ทำให้ถือเป็นระบบที่คล่องตัว และมีความปลอดภัยสูง จึงทำให้เป็นที่นิยม เข้าถึงง่าย และได้รับความสนใจมากขึ้นเรื่อย ๆ ในขณะเดียวกันก็ถือว่าเป็นสินทรัพย์ดิจิทัลที่มีมูลค่าการลงทุนในตลาดรวมสูงมากด้วยเช่นกัน แต่เนื่องจากมีเหรียญในสกุลเงินดิจิทัลหลากหลาย อีกทั้งสกุลเงินดิจิทัลยังมีความผันผวนเป็นอย่างมากในแต่ละช่วงเวลา ปัจจัยที่ทำให้ราคาสกุลเงินดิจิทัลผันผวน เช่น ความต้องการของตลาด การนำไปใช้งานจริง กฎหมายควบคุม การแข่งขัน และ ข่าวสาร ซึ่งมีความเสี่ยงสูงในการลงทุน ทำให้ยากต่อการตัดสินใจซื้อหรือควรลงทุนในเหรียญไหนในช่วงเวลานั้น ๆ

ดังนั้นการค้นคว้าอิสระนี้จึงจัดทำการวิเคราะห์ความคิดเห็นในการลงทุนในสกุลเงินดิจิทัลโดยประยุกต์ใช้การขุดเหมืองข้อความและการวิเคราะห์ความรู้สึก โดยนำความคิดเห็นเกี่ยวกับสกุลเงินดิจิทัลจากเว็บไซต์คลังคลิปวิดีโอตั้งแต่อดีตจนถึงปัจจุบันมาสร้างแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network) เพื่อทำนายความคิดว่าเป็นเชิงบวกหรือความคิดเห็นเชิงลบ เพื่อช่วยแนะนำและช่วยสนับสนุนการตัดสินใจในการซื้อสกุลเงินดิจิทัลของผู้ลงทุน

1.2 วัตถุประสงค์ของงานวิจัย

1) ศึกษาและวิเคราะห์การลงทุนในสกุลเงินดิจิทัลโดยประยุกต์ให้การขุดเหมืองข้อความและการวิเคราะห์ความรู้สึก

2) ศึกษาและสร้างแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับด้วยข้อมูลความคิดเห็นเกี่ยวกับสกุลเงินดิจิทัล

3) เปรียบเทียบประสิทธิภาพแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับที่ถูกฝึกอบรมด้วยข้อมูลที่ถูกจัดเตรียมด้วยวิธีที่แตกต่างกัน

4) ศึกษาการวิเคราะห์องค์ประกอบหลัก

5) เพื่อทำนายความคิดเห็นและช่วยแนะนำ สนับสนุนประกอบการการตัดสินใจในการลงทุนในสกุลเงินดิจิทัล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

1.3 ขอบเขตของงานวิจัย

1) วิเคราะห์ความรู้สึกของความคิดเห็นในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอ ในช่วง 1 มกราคม 2564 ถึง 30 เมษายน 2565 โดยทำการดึงวีดีโอที่อัปโหลดเดือนละ 4-5 วีดีโอ โดยดึงวีดีโอที่อัปโหลดในช่วงต้นเดือน กลางเดือน ปลายเดือนของในแต่ละเดือน

2) วิเคราะห์องค์ประกอบหลักที่ช่วยในการตัดสินใจจากคุณลักษณะวีดีโอสกุลเงินดิจิทัลและราคาของสกุลเงินดิจทัล

3) ใช้ภาษาไพธอนในการวิเคราะห์เหมืองข้อความ การวิเคราะห์ความรู้สึก และการวิเคราะห์องค์ประกอบหลัก

1.4 ประโยชน์ที่คาดว่าจะได้รับ

1) ผู้ลงทุนในสกุลเงินดิจิทัลสามารถตัดสินใจลงทุนซื้อเหรียญต่าง ๆ ได้

2) ได้เรียนรู้และเข้าใจวิธีการเรียนรู้ด้วยเครื่องของอัลกอริทึมโครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network) และการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) มากขึ้น

1.5 ขั้นตอนการดำเนินงาน

1) หาหัวข้อค้นคว้า

2) เก็บรวบรวมข้อมูลความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอและราคาสกุลเงินดิจิทัล

3) พัฒนาแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network) ด้วยข้อมูลที่จัดเตรียมไว้และทำการเลือกองค์ประกอบหลัก

4) สรุปผลการค้นคว้า

5) จัดทำรูปเล่มการค้นคว้า

6) ตรวจสอบรูปเล่มการค้นคว้า

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีที่เกี่ยวข้อง ซึ่งจะประกอบไปด้วย สกุลเงินดิจิทัล การทำเหมืองข้อความ และวิเคราะห์ความรู้สึก โครงข่ายประสาทเทียมแบบวนกลับแบบ Long Short-Term Memory โครงข่ายประสาทเทียมแบบวนกลับแบบ Gated Recurrent Units โครงข่ายประสาทเทียมแบบวนกลับแบบ Bidirectional Encoder Representations from Transformer และการวิเคราะห์องค์ประกอบหลัก โดยที่จะแสดงรายละเอียดเพิ่มเติมในหัวข้อ 2.1 ถึง 2.6

2.1 สกุลเงินดิจิทัล

สกุลเงินดิจิทัล หรือ คริปโตเคอเรนซี คือ สกุลเงินเข้ารหัส โดยสกุลเงินดิจิทัล หรือ สินทรัพท์ดิจิทัลนั้นได้รับการออกแบบมาเพื่อเป็นสื่อกลางของการแลกเปลี่ยนโดยใช้วิทยาการเข้ารหัสหรือรหัสลับ (Cryptography) ในการป้องกันและยืนยันธุรกรรม โดยพื้นฐานแล้วสกุลเงินดิจิทัลจะมีการจำกัดการเข้าถึงฐานข้อมูล ซึ่งนั่นทำให้การเปลี่ยนแปลงข้อมูลเป็นไปได้ยาก เราสามารถซื้อ ขาย แลกเปลี่ยน และสะสมสกุลเงินดิจิทัลได้ นอกจากนี้สกุลเงินดิจิทัลยังสามารถใช้แทนมูลค่าของบริการที่ถูกสร้างไว้บนเทคโนโลยีบล็อกเชนอีกด้วย โดยบล็อกเชน หรือ บางทีถูกเรียกว่าระบบบันทึกรายการธุรกรรมดิจิทัลบนเครือข่ายคอมพิวเตอร์ (Distributed Ledger Technology หรือ DLT) นั้น (2) ทำให้ธุรกรรม ของสินทรัพย์ดิจิทัลสามารถเกิดขึ้นได้อย่างปลอดภัย โปร่งใส และปลอมแปลงได้ยาก ด้วยการใช้เครือข่ายแบบกระจายศูนย์ (Decentralization) และการใช้รหัสผ่านแบบสร้างข้อมูลแทนตัว (Cryptographic Hashing) สกุลเงินดิจิทัลส่วนใหญ่ทำงานอยู่บนบล็อกเชนสาธารณะหรือบล็อกเชนเปิด (Public Blockchain) อย่างไรก็ตามบล็อกเชนยังมีอีกหนึ่งประเภทนั่นก็คือบล็อกเชนส่วนบุคคลหรือบล็อกเชนปิด (Private Blockchain)

2.2 การทำเหมืองข้อความและการวิเคราะห์ความรู้สึก

การขุดข้อความ (Text Mining) หรือที่เรียกว่าการทำเหมืองข้อมูลข้อความ (Text Data Mining) เป็นกระบวนการในการแปลงข้อความที่ไม่มีโครงสร้างให้อยู่ในรูปแบบที่มีโครงสร้างเพื่อระบุรูปแบบที่มีความหมายและข้อมูลเชิงลึกใหม่ ด้วยการใช้เทคนิคการเรียนรู้ของเครื่อง สถิติ และภาษาศาสตร์ และอัลกอริทึมการเรียนรู้เชิงลึกอื่น ๆ ด้วยการแปลงข้อมูลให้อยู่ในรูปแบบที่มีโครงสร้างมากขึ้นผ่านการทำเหมืองข้อความและการวิเคราะห์ข้อความ การขุดข้อความสามารถสำรวจและค้นพบ ความสัมพันธ์ที่ซ่อนอยู่ภายในข้อมูลที่ไม่มีโครงสร้างได้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ข้อความเป็นหนึ่งในประเภทข้อมูลที่พบบ่อยที่สุดในฐานข้อมูล ข้อมูลนี้สามารถจัดเป็น ข้อมูลที่ มีโครงสร้าง ข้อมูลนี้ได้รับการกำหนดมาตรฐานให้อยู่ในรูปแบบตารางที่มีแถวและคอลัมน์จำนวนมาก ทำให้ง่ายต่อการจัดเก็บและประมวลผลสำหรับอัลกอริทึมสำหรับการวิเคราะห์และแมชชีนเลิร์นนิง ข้อมูลที่มีโครงสร้างอาจรวมถึงการป้อนข้อมูล เช่น ชื่อ ที่อยู่ และหมายเลขโทรศัพท์ ข้อมูลที่ไม่มีโครงสร้าง ข้อมูลนี้ไม่มีรูปแบบข้อมูลที่กำหนดไว้ล่วงหน้า ซึ่งอาจรวมถึงข้อความจากแหล่งที่มา เช่น โซเชียลมีเดียหรือบทวิจารณ์ผลิตภัณฑ์ หรือรูปแบบสื่อสมบูรณ์ เช่น ไฟล์วิดีโอและไฟล์เสียง ข้อมูลกึ่งโครงสร้าง ข้อมูลนี้เป็นการผสมผสานระหว่างรูปแบบข้อมูลที่มีโครงสร้างและไม่มีโครงสร้าง แต่ก็มีโครงสร้างไม่เพียงพอที่จะตอบสนองความต้องการของฐานข้อมูลเชิงสัมพันธ์ ตัวอย่างของข้อมูลกึ่งโครงสร้าง ได้แก่ ไฟล์ XML, JSON และ HTML เนื่องจาก 80 เปอร์เซ็นต์ ของข้อมูลในโลกอยู่ในรูปแบบที่ไม่มีโครงสร้าง การทำเหมืองข้อความจึงเป็นแนวทางปฏิบัติที่มีคุณค่าอย่างยิ่งภายในองค์กร เครื่องมือขุดข้อความและเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing) ทำให้เราสามารถแปลงเอกสารที่ไม่มีโครงสร้างให้อยู่ในรูปแบบที่มีโครงสร้างเพื่อให้สามารถวิเคราะห์และสร้างข้อมูลเชิงลึกได้ ซึ่งจะช่วยปรับปรุงการตัดสินใจขององค์กรนำไปสู่ผลลัพธ์ทางธุรกิจที่ดีขึ้น

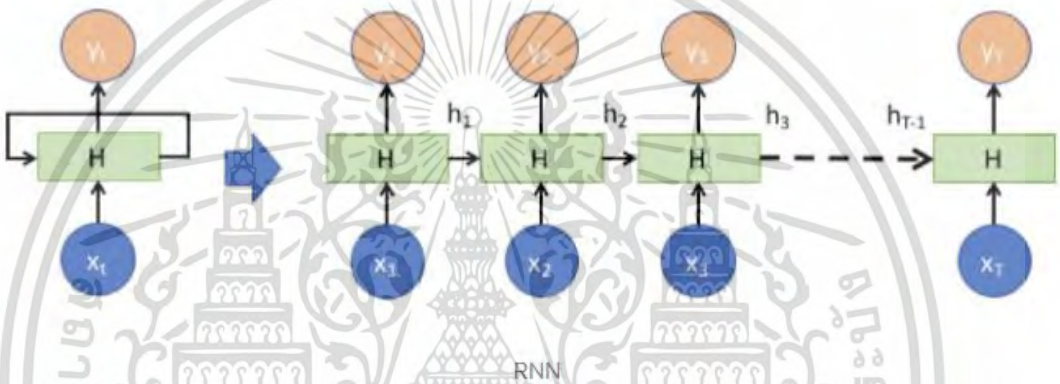
การวิเคราะห์ความคิดเห็นเป็นกระบวนการในการตรวจจับความรู้สึกเชิงบวกหรือเชิงลบในข้อความ ธุรกิจมักใช้เพื่อตรวจจับความรู้สึกในข้อมูลโซเชียล วัดชื่อเสียงของแบรนด์ และทำความเข้าใจลูกค้า เนื่องจากลูกค้าแสดงความคิดและความรู้สึกของตนอย่างเปิดเผยมากกว่าที่เคย การวิเคราะห์ ความรู้สึกจึงกลายเป็นเครื่องมือสำคัญในการติดตามและทำความเข้าใจความรู้สึกนั้น การวิเคราะห์ ความคิดเห็นของลูกค้าโดยอัตโนมัติ เช่น ความคิดเห็นในการตอบแบบสำรวจและการสนทนาทางโซเชียลมีเดีย ช่วยให้แบรนด์ต่าง ๆ เรียนรู้สิ่งที่ทำให้ลูกค้ามีความสุขหรือหงุดหงิด เพื่อให้พวกเขาสามารถปรับแต่งผลิตภัณฑ์และบริการให้ตรงกับความต้องการของลูกค้าได้

การวิเคราะห์ความรู้สึก หรือที่เรียกว่าการขุดความคิดเห็น ทำงานด้วยการประมวลผลภาษาธรรมชาติ และอัลกอริทึมการเรียนรู้ของเครื่องที่อยู่เบื้องหลังการสนทนาออนไลน์โดยอัตโนมัติมีอัลกอริทึมต่าง ๆ ที่สามารถนำไปใช้ในแบบจำลองการวิเคราะห์ความคิดเห็นขึ้นอยู่กับจำนวน ข้อมูลที่ต้องการวิเคราะห์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3 โครงสร้างประสาทเทียมแบบวนกลับ

โครงข่ายประสาทเทียมแบบวนกลับ (Recurrent Neural Network) เป็นโครงข่ายประสาทเทียมชนิดที่ทำงานสำหรับข้อมูลอนุกรมเวลาหรือข้อมูลที่เกี่ยวข้องกับลำดับ อย่างไรก็ตามหากเรามีข้อมูลตามลำดับโดยที่จุดข้อมูลหนึ่งจุดขึ้นอยู่กับจุดข้อมูลก่อนหน้า เราจำเป็นต้องแก้ไขโครงข่ายประสาทเทียมเพื่อรวมการพึ่งพาระหว่างจุดข้อมูลเหล่านี้ โครงสร้างประสาทเทียมแบบวนกลับมีแนวคิดเรื่องหน่วยความจำที่ช่วยให้จัดเก็บสถานะหรือข้อมูลของอินพุตก่อนหน้าเพื่อสร้างเอาต์พุตถัดไปของลำดับในการคำนวณหาค่า โดยสามารถอธิบายโครงสร้างของโครงข่ายประสาทเทียมแบบวนกลับ(3) ตามรูปที่ 2.1



รูปที่ 2.1 โครงสร้างของโครงข่ายประสาทเทียมแบบวนกลับ

หมายเหตุ. จาก https://medium.com/@sinart.t

โดยที่

H = hidden layer

yt = output จาก RNN ที่เวลา *t*

xt = input data ที่เวลา *t*

ht = hidden state ที่เวลา *t*

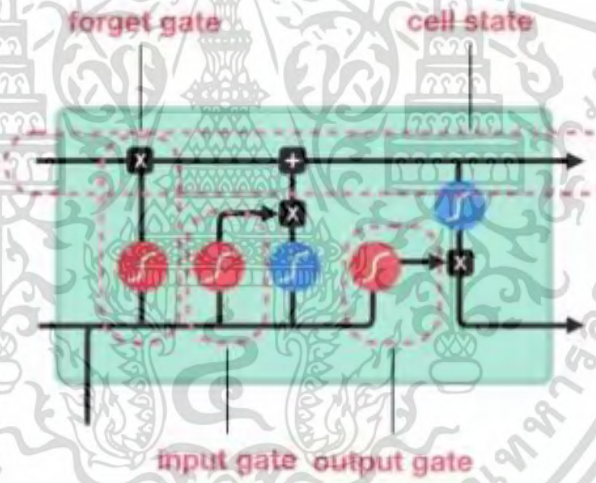
เนื่องจากโครงสร้างประสาทเทียมแบบวนกลับ ใช้ข้อมูลจากเครือข่ายก่อนหน่าทำให้สามารถทำงานได้ดีในข้อมูลอนุกรมเวลา นำข้อมูลเวลาก่อนหน้ามาหาต่อกับเวลาปัจจุบัน ข้อมูลอนุกรมเวลานั้นรวมถึงข้อมูลแบบข้อความและข้อมูลเสียง แต่ข้อเสียของโครงสร้างประสาทเทียมแบบวนกลับ คือ มันสามารถดูย้อนกลับได้แค่เพียงในช่วงระยะเวลาสั้น ๆ เท่านั้น ซึ่งปัญหาหลัก ๆ ของโครงสร้างประสาทเทียมแบบวนกลับ เกิดมาจากค่าเกรเดียนที่เริ่มน้อยลงในข้อมูลที่มีความยาวมากขึ้น จนแทบจะไม่สามารถเห็นความเปลี่ยนแปลงของเกรเดียนได้เลย ซึ่งปัญหานี้ถูกเรียกว่า ปัญหาการสูญหายของเกรเดียน (Vanishing Gradient Problem: VGP) ซึ่งปัญหานี้ถูกแก้ไขโดยใช้เกทแบบวนกลับ (Gated

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

Recurrent Unit: GRU) และหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) โดยในบทนี้จะกล่าวถึงโครงข่ายประสาทเทียมแบบวนกลับทั้งหมด 3 อัลกอริทึม ได้แก่ 1. หน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory) 2.หน่วยที่เกิดซ้ำแบบ gated (Gated Recurrent Units) เนื่องจากเป็นโครงสร้างประสาทเทียมแบบวนกลับที่เทรนง่ายขึ้นเพราะแก้ปัญหาการสูญหายของเกรเดียน (Vanishing Gradient) ได้ดี 3. BERT (Bidirectional Encoder Representations from Transformer) มีฐานข้อมูลขนาดใหญ่ในการวิเคราะห์คำศัพท์ ทำให้ได้ค่าความถูกต้องค่อนข้างสูง BERT ได้รับการฝึกฝนเรียนรู้เข้าใจภาษาจากเนื้อหาข้อความภาษาจำนวนมาก โดยใช้คลังคำศัพท์ Corpus ของ Wikipedia, Toronto Book Corpus

2.3.1 หน่วยความจำระยะสั้นแบบยาว

หน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) การทำงานภายในของเครือข่ายหน่วยความจำระยะสั้นแบบยาว ประกอบด้วยสามส่วนและแต่ละส่วนทำหน้าที่แยกกัน ดังที่แสดงในรูปที่ 2.2



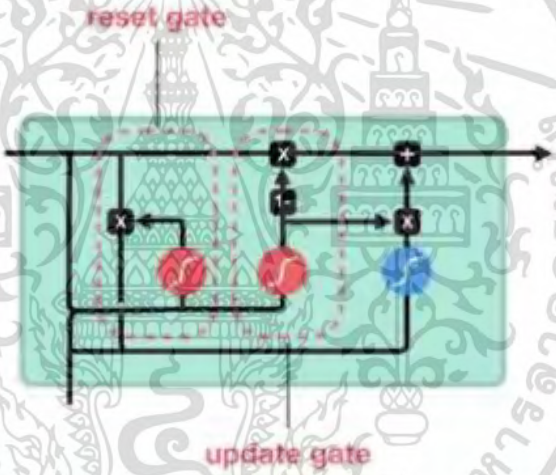
รูปที่ 2.2 ส่วนประกอบของหน่วยความจำระยะสั้นแบบยาว
หมายเหตุ. จาก https://towardsdatascience.com

ในเซลล์ของเครือข่ายหน่วยความจำระยะสั้นแบบยาว ขั้นตอนแรกคือ Forget Gate ทำหน้าที่ตัดสินใจว่าควรเก็บข้อมูลจากการประทับเวลาก่อนหน้าหรือไม่เก็บข้อมูลก่อนหน้า ส่วนที่สองเรียกว่า Input Gate ใช้เพื่อวัดความสำคัญของข้อมูลใหม่ที่ดำเนินการโดยอินพุต และส่วนสุดท้ายคือ Output Gate ทำหน้าที่ตัดสินใจอย่างมีเงื่อนไขว่าจะส่งออกอะไรตามอินพุตและหน่วยความจำ (4)

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.3.2 Gated Recurrent Units (GRU)

Gated Recurrent Units (GRU) ใช้เกตเพื่อควบคุมการไหลของข้อมูล ที่คล้ายกับ Long Short-Term Memory (LSTM) แต่ Gated Recurrent Units มี Reset Gate จะช่วยตัดสินใจว่าจะเก็บข้อมูลใดจากขั้นตอนเวลาก่อนหน้าพร้อมกับอินพุตใหม่ โดยใช้ทั้งสถานะที่ซ่อนอยู่จากขั้นตอนเวลาก่อนหน้าและข้อมูลอินพุต ที่ขั้นตอนเวลาปัจจุบัน และ Update Gate ใช้สถานะที่ซ่อนอยู่ก่อนหน้าและข้อมูลอินพุตปัจจุบันช่วย ให้โมเดลกำหนดว่าข้อมูลในอดีตที่เก็บไว้ในสถานะที่ซ่อนอยู่ก่อนหน้านั้นจะต้องถูกเก็บรักษาไว้ในอนาคตมากน้อยเพียงใด Gated Recurrent Units มีประสิทธิภาพใกล้เคียงกับ Long Short-Term Memory ในหลาย ๆ งาน แต่เนื่องจาก Parameter น้อยกว่า Long Short-Term Memory เนื่องจากไม่มี Output Gate ทำให้เทรนได้ง่ายกว่า เร็วกว่า และในบางงานที่ชุดข้อมูลมีขนาดเล็กพบว่า Gated Recurrent Units ประสิทธิภาพดีกว่า (5) ดังที่แสดงในรูปที่ 2.3



รูปที่ 2.3 ส่วนประกอบของ Gated Recurrent Units

หมายเหตุ. จาก https://towardsdatascience.com

2.3.3 Bidirectional Encoder Representations from Transformer

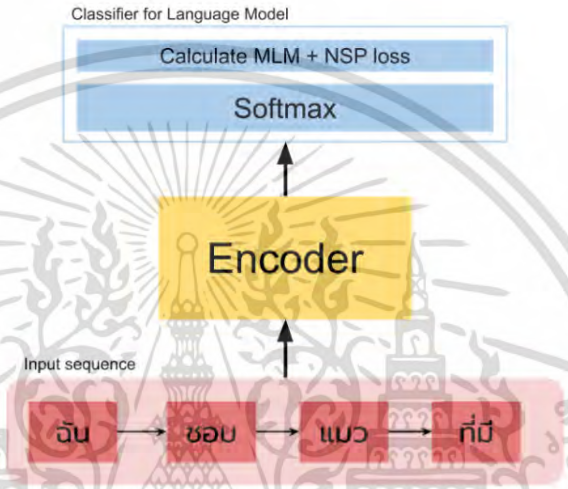
‘BERT’ นั้นย่อมาจาก Bidirectional Encoder Representations from Transformer ถูกพัฒนาโดย Google ก็คือ Artificial Intelligent (AI) ของ Google Algorithm เวอร์ชั่นใหม่ล่าสุด เหมือนเครื่องมือที่ทำให้สามารถเรียนรู้ภาษาของมนุษย์ (Natural Language Processing: NLP) ได้ดีขึ้น

BERT จึงเป็นความพยายามในการออกแบบ Language Model ที่สามารถเรียนรู้จากทั้งบริบทที่อยู่ทางซ้าย และขวา พร้อมๆกัน โดยใช้เทคนิค Self-Attention สถาปัตยกรรม

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ของ BERT ถูกออกแบบให้เลือกใช้เฉพาะส่วนที่เป็น encoder ซึ่งทำหน้าที่แปลงคำในประโยคให้เปลี่ยนไปเป็นเวกเตอร์

เพื่อให้ encoder สามารถทำหน้าที่เป็น Language Model ได้ Bidirectional Encoder Representations from Transformer จึงเพิ่มโมเดลอีก 1 ตัว ต่อจาก encoder ที่มีอยู่เดิม เพื่อทำหน้าที่เป็น Classifier นำเวกเตอร์ที่ได้จาก encoder ไปคำนวนต่อให้ได้คำตอบในรูปแบบคล้ายๆกับ Language Model ทั่วๆไป (6) ดังที่แสดงในรูปที่ 2.4

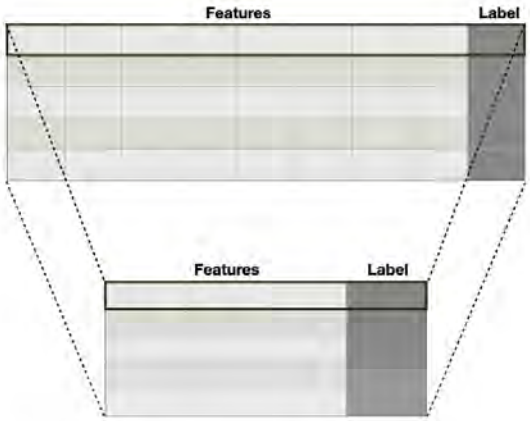


รูปที่ 2.4 ส่วนประกอบของ Bidirectional Encoder Representations from Transformer หมายเหตุ. จาก https://medium.com/@chameleontk

2.4 การวิเคราะห์องค์ประกอบหลัก

การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis: PCA) เป็นวิธีที่ใช้วิเคราะห์ข้อมูลหลายตัวแปร เพื่อหาความสัมพันธ์ของตัวแปรเหล่านั้นส่งผลทำให้เกิดการลดมิติของข้อมูล (High dimensional) ให้มีมิติลดลง เห็นได้บ่อยครั้งในการใช้การวิเคราะห์องค์ประกอบหลักสำหรับงาน Image ที่ นำมาประยุกต์เพื่อลดจำนวนของคุณลักษณะ (Features) ให้มีจำนวนลดลง ซึ่งจะส่งผลทำให้การสร้างโมเดลนั้นใช้เวลาน้อยลงซึ่งมีประโยชน์มากในงาน Regression เป็นต้น การทำการวิเคราะห์องค์ประกอบหลักถือเป็นการแก้ไขปัญหาความสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) ไปโดยอัตโนมัติ เพราะกระบวนการในการสร้างองค์ประกอบ (Components) ของ การวิเคราะห์องค์ประกอบหลักจะไม่มีความสัมพันธ์กันเกิดขึ้น โดยแสดงรูปแบบการลดจำนวนคุณลักษณะในรูปที่ 2.5

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



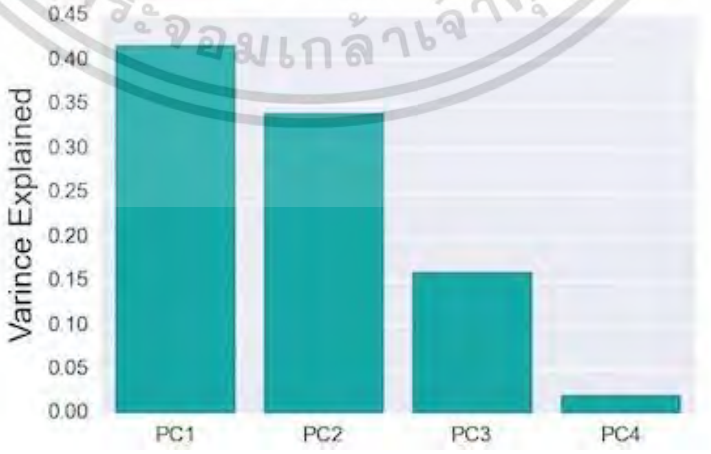
รูปที่ 2.5 รูปแบบของการลดคุณลักษณะ

หมายเหตุ. จาก https://towardsdatascience.com/using-principal-component-analysis

ยกตัวอย่าง สมมุติว่ามีตัวแปรหรือคุณลักษณะประมาณ 50 ตัว แต่อาจจะไม่ได้ใช้งานตัวแปรเหล่านี้ทั้งหมด เราต้องอาศัยการเรียงลำดับความสำคัญของตัวแปร เช่น ตัวแปรใดบ้างส่งผลกระทบก็นำมาใช้งาน หรือ ตัวแปรใดแทบไม่มีเกี่ยวข้องเลย

โดยการนำข้อมูลเข้าไปวิเคราะห์องค์ประกอบหลักนั้นจะต้องทำการการแปลงข้อมูลก่อน สาเหตุที่ต้องทำนั้นเพราะว่ามีข้อมูลที่ลักษณะที่ต่างกันมากเกินไป ตัวอย่างเช่น น้ำหนัก กับ ส่วนสูง เป็นข้อมูลที่อยู่คนละหน่วยวัด ดังนั้นควรมีการปรับช่วงข้อมูลให้อยู่ในช่วงเดียวกันก่อน ซึ่งสามารถใช้การปรับมาตรฐาน (Standardization หรือ Z-Score Normalization) มาจัดการได้ (7)

Principal Component Analysis จะทำการสร้างตัวแปรที่เรียกว่า องค์ประกอบ (component) หรือ n_components ซึ่งหมายถึงจำนวนมิติของข้อมูลที่ต้องการลดทอนลงไป โดยโดย component ตัวแรกจะมีค่าความแปรปรวน (variance) สูงที่สุด ซึ่งจะอธิบาย ข้อมูลได้มากที่สุด และตัวถัดๆ ไปก็จะค่าความแปรปรวนลดลงตามลำดับ แสดงดังรูปที่ 2.6 ซึ่งจำนวนองค์ประกอบที่เหมาะสมที่ถูกเลือกมาใช้จะครอบคลุมความแปรปรวนประมาณ 80–90 เปอร์เซ็นต์ (8)



รูปที่ 2.6 ตัวอย่างลำดับความแปรปรวนของแต่ละองค์ประกอบ

หมายเหตุ. จาก https://kongruksiam.medium.com

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.5 การวิเคราะห์การถดถอยโลจิสติก

การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression) คือเป็นโมเดลทางสถิติที่จำลองความน่าจะเป็นของเหตุการณ์ที่เกิดขึ้นเพื่อการประมาณค่าหรือทำนายเหตุการณ์ที่สนใจว่าจะเกิดหรือไม่เกิดเหตุการณ์ภายใต้อิทธิของตัวปัจจัย แบบจำลอง การถดถอยโลจิสติกประกอบด้วยตัวแปรตามแบบไบนารี (Binary Dependent Variable) จะถูกแปลงเป็น 0 และ 1 ในขณะที่ตัวแปรอิสระแต่ละตัวสามารถเป็นตัวแปรไบนารี หรือตัวแปรต่อเนื่อง (Continuous Variable) ก็ได้ การวิเคราะห์การถดถอยโลจิสติกเกี่ยวข้องกับทฤษฎีความน่าจะเป็นทวินามถูกเรียกว่า Binomial Logistic Regression ถ้าตัวแปรตามเป็นพหุนามจะเรียกว่า Multinomial Logistic Regression การวิเคราะห์การถดถอยโลจิสติกจัดเป็นเครื่องมือวิเคราะห์ข้อมูลในการศึกษาวิจัยที่มีวัตถุประสงค์เพื่อทำนายเหตุการณ์ หรือประเมินความเสี่ยง จึงมีการประยุกต์ใช้ในงานวิจัยหลากหลายสาขา ทั้งสาขาทางการแพทย์วิศวกรรมศาสตร์นิเวศวิทยา เศรษฐศาสตร์และสังคมศาสตร์ (9)

ในการคำนวณการวิเคราะห์การถดถอยโลจิสติกที่ใช้กับปัญหาตัวแปรตามแบบไบนารี แบบ 0 และ 1 สามารถเขียนฟังชั่นเป็น y = f(x) โดย y = {0,1} และได้โมเดลเป็น z = b0 + b1x1 + b2x2 + ... bnxn เพื่อคำนวณหาค่า z และนำฟังก์ชันซิกมอยด์ (Sigmoid Function) เพื่อแปลงค่า z ให้มีค่าอยู่ระหว่าง 0 ถึง 1 ถ้า prob >= 0.5 predict y=1และถ้า prob < 0.5 predict y=0 การวัดความแม่นยำของโมเดลคือการนำ predicted y ไปเทียบกับค่า y ของจากเหตุการณ์จริง

รูปที่ 2.7 Standard Logistic Function

หมายเหตุ. จาก https://forest-admin.forest.ku.ac.th

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

2.6 งานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้มีจุดประสงค์เพื่อสร้างระบบแนะนำและช่วยสนับสนุนการตัดสินใจในการลงทุนในสกุลเงินดิจิทัล โดยได้รับแรงบันดาลใจจากงานวิจัยดังต่อไปนี้

2.6.1 Emre Sasmaz และ F. Boray Tek (2021) ได้ทำวิจัยเกี่ยวกับการวิเคราะห์ความความรู้สึกของข้อความทวีตบนทวิตเตอร์สำหรับสกุลเงินดิจิทัล (Cryptocurrency) เพื่อแนะนำการซื้อขายสกุลเงินดิจิทัลรายวัน ในการวิจัยนี้ได้ทำการตรวจสอบความเป็นไปได้ของการวิเคราะห์ความรู้สึกอัตโนมัติสำหรับสกุลเงินดิจิทัลโดยใช้แบบจำลองโครงข่ายประสาทเทียมแบบวนกลับแบบ BERT (Bidirectional Encoder Representations from Transformer) โดยได้กำหนดเป้าหมายเป็นหนึ่งสกุลเงินดิจิทัล คือ NEO โดยจำนวนทวีตที่ใช้ในการทดลอง 846,790 ทวีต และราคาของสกุลเงินดิจิทัลครอบคลุม 5 ปี และทำการติดป้ายกำกับทวีตเป็นความรู้สึกเชิงบวก เชิงลบ และเป็นกลาง และนำไปผ่านการฝึกอบรมและทดสอบตัวจำแนกประเภทโดยการสร้างต้นไม้ตัดสินใจหลายๆต้น (Random Forest) บนข้อมูลที่มีป้ายกำกับ โดยมีความแม่นยำอยู่ที่ 77 เปอร์เซ็นต์ จากการวิจัยพบว่าความรู้สึกรายวันของทวีตมีความสัมพันธ์กับราคารายวันของสกุลเงินดิจิทัล

2.6.2 Mohammad J. Hamayel และ Amani Yousef Owda (2021) ได้ทำการวิจัยเกี่ยวกับการสร้างแบบจำลองใช้ในการทำนายราคาสกุลเงินดิจิตัล (Cryptocurrency) สามประเภท ได้แก่ บิตคอยน์ (Bitcoin : BTC) ไลท์คอยน์ (Litecoin : LTC) และ อีเธอร์เลียม (Ethereum : ETH) โดยใช้แบบจำลองโครงข่ายประสาทเทียมแบบวนกลับทั้งหมดสามอัลกอรึทึม ได้แก่ หน่วยที่เกิดซ้ำแบบ gated (Gated Recurrent Units : GRU) นำเสนอการคาดการณ์ด้วย ค่าเฉลี่ยของร้อยละความผิดพลาดสัมบูรณ์ (Mean Absolute Percentage Error : MAPE) ดีกว่าแบบจำลองหน่วยความจำระยะสั้นระยะยาว (Long Short-Term Memory : LSTM) และ หน่วยความจำระยะสั้นระยะยาวแบบสองทิศทาง (Bidirectional Long Short-Term Memory : bi - LSTM) โดยรวมแล้วแบบจำลองการทำนายในบทวิจัยนี้แสดงผลลัพธ์ที่แม่นยำใกล้เคียงกับราคาจริงของสกุลเงินดิจิทัลโดยสามารถช่วยนักลงทุนและผู้ค้าในการระบุการขายและการซื้อสกุลเงินดิจิทัล

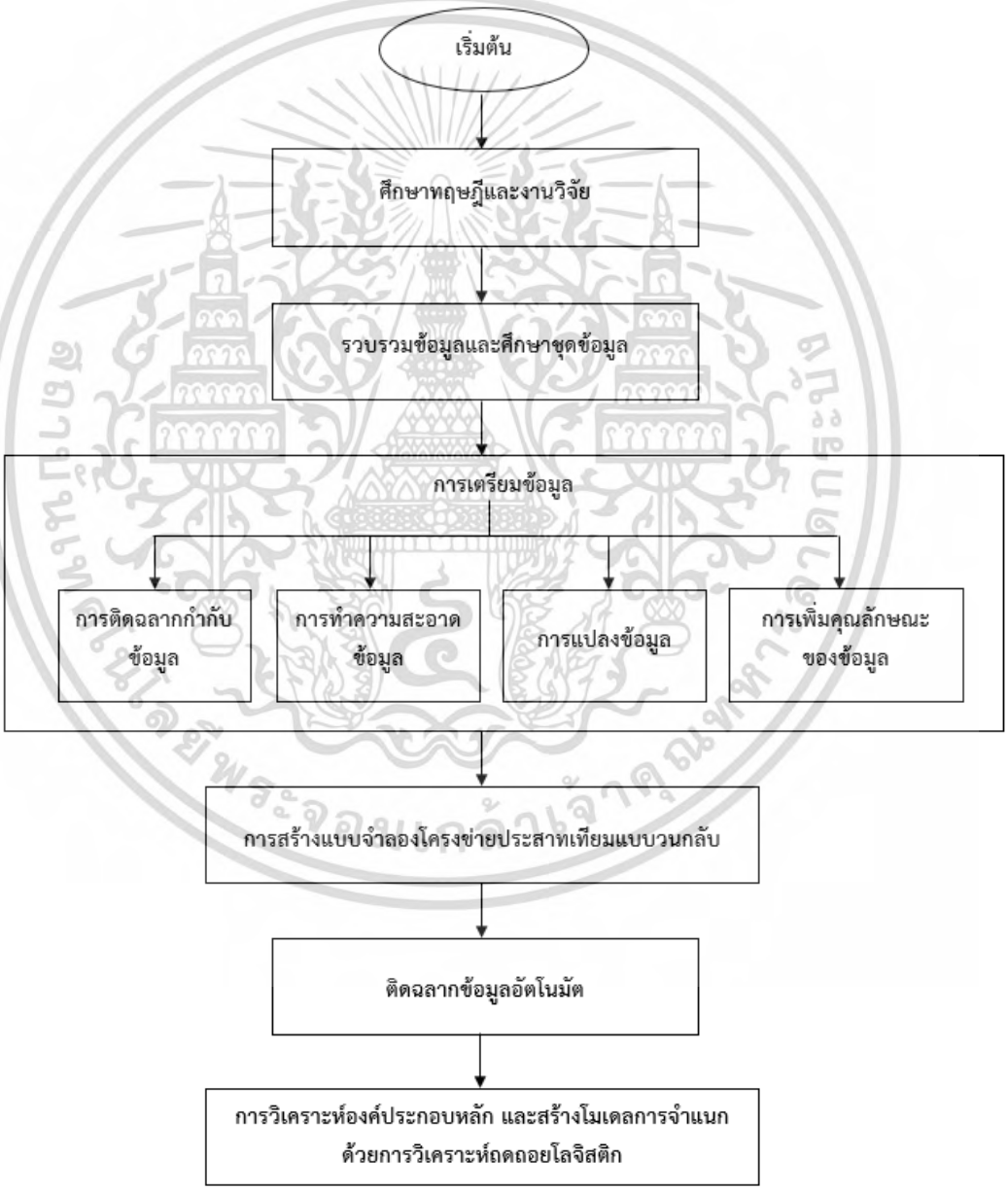
2.6.3 María Guinda และ Ritabrata Bhattacharyya (2021) ได้ทำการวิจัยเกี่ยวกับการวิเคราะห์องค์ประกอบหลักเกี่ยวกับความสัมพันธ์ของสกุลเงินดิจิทัล (Cryptocurrency) เพื่อสร้างพอร์ตโฟลิโอที่หลากหลาย การสร้างพอร์ตโฟลิโอสกุลเงินดิจิทัลที่หลากหลายซึ่งเป็นผลมาจากการวิเคราะห์ Principal Component Analysis กับชุดข้อมูลที่ประกอบด้วยผลความสัมพันธ์ของแต่ละสินทรัพย์ จากนั้นแสดงค่าลักษณะเฉพาะเพื่อตรวจสอบว่าเปอร์เซ็นต์ของการเปลี่ยนแปลงที่อธิบายโดยตัวแปรที่คำนวณใหม่แต่ละตัวคือเท่าใด มิติข้อมูลแรกแสดงถึง 71.7% ของการเปลี่ยนแปลงและ 5.2% สำหรับมิติที่สอง ซึ่งหมายความว่าด้วยตัวแปรใหม่เพียงสองตัว (Dimension1 และ Dimension2) สามารถแสดงข้อมูล 76.9% ที่มีอยู่ในตัวแปรสหสัมพันธ์ 30 ตัว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 3

วิธีการดำเนินงานวิจัย

วิธีการดำเนินงานวิจัยการขุดเหมืองข้อความและการวิเคราะห์ความรู้สึกสำหรับการลงทุนในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอด้วยโครงข่ายประสาทเทียมแบบวนกลับ เพื่อให้การวิจัยในครั้งนี้เป็นไปตามวัตถุประสงค์ จึงมีขั้นตอนการดำเนินงานวิจัยและแสดงรายละเอียดดังรูปที่ 3.1



รูปที่ 3.1 ลำดับขั้นตอนการดำเนินงาน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.1 วิธีการศึกษา

วิธีการศึกษาเพื่อดำเนินงานวิจัยการขุดเหมืองข้อความและการวิเคราะห์ความรู้สึกสำหรับการลงทุนในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอด้วยโครงข่ายประสาทเทียมแบบวนกลับ ผู้ศึกษาได้ศึกษาและวิเคราะห์กระบวนการดำเนินวานจากงานวิจัยที่เกี่ยวข้องพร้อมทั้งทฤษฎีต่าง ๆ จากเอกสาร โดยมีรายละเอียดดังนี้

1. ศึกษาข้อมูลเกี่ยวกับการขุดเหมืองข้อความและการวิเคราะห์ความรู้สึก
2. ศึกษาความรู้พื้นฐานของสกุลเงินดิจิทัล
3. ศึกษาและวิเคราะห์จากงานวิจัยที่เกี่ยวข้อง
4. ศึกษาแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ
5. ศึกษาการวิเคราะห์องค์ประกอบหลักและการวิเคราห์ถดถอยโลจิสติก

3.2 ขั้นตอนการรวบรวมข้อมูลและศึกษาชุดข้อมูล

ในการวิจัยครั้งนี้ ผู้พัฒนาได้ทำการดึงข้อมูลความคิดเห็น (comments) และ รายละเอียดของวีดีโอจากเว็บไซต์คลังคลิปวิดีโอ (YouTube) เนื่องจากในสถานการณ์ปัจจุบันความคิดเห็นบนโลกออนไลน์มีผลต่อราคาและการผันผวกของสกุลเงินดิจิทัล และมีผลต่อการตัดสินใจลงทุนในสกุลเงินดิจิทัล โดยทำการดึงข้อมูลโดยใช้ API (Application Programming Interface) ของ Google ซึ่ง YouTube Data API v3 เป็น API ที่ให้การเข้าถึงข้อมูลของเว็บไซต์คลังคลิปวิดีโอเช่น วิดีโอ เพลย์ลิสต์ และช่อง โดยทำการดึงข้อมูลของวีดีโอที่พูดถึงสกุลเงินดิจิทัลในช่วงวันที่ 1 มกราคม 2564 ถึงวันที่ 30 เมษายน 2565 โดยมีจำนวนทั้งสิ้น 72 วีดีโอโดยทำการดึงวีดีโอที่อัปโหลดเดือนละ 4-5 วีดีโอ โดยดึงวีดีโอที่อัปโหลดในช่วงต้นเดือน กลางเดือน ปลายเดือนของในแต่ละเดือน และประกอบด้วยความคิดเห็นจำนวน 8,509 ความคิดเห็น โดยข้อมูลรายละเอียดของวีดีโอประกอบด้วยคุณลักษณะดังตารางที่ 3.1 และตัวอย่างข้อมูลรายละเอียดของวีดีโอในรูปที่ 3.2

ตารางที่ 3.1 ตารางอธิบายชื่อคุณลักษณะของวีดีโอที่ได้จากการดึง API

ลำดับที่	ชื่อคอลัมน์	คำอธิบายชื่อคอลัมน์
1	Channel_name	ชื่อช่อง
2	Channel_id	Id ช่อง
3	publishedAt	วันที่อัพโหลดวีดีโอบน YouTube
4	video_name	ชื่อวีดีโอ
5	Description	คำอธิบายวีดีโอ
6	Video_id	Id วีดีโอ
7	viewCount	จำนวนคนเข้าชมวีดีโอ ณ วันที่ดึงข้อมูล

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

8	likeCount	จำนวนคนกดถูกใจวีดีโอ ณ วันที่ดึงข้อมูล
9	commentCount	จำนวนความคิดเห็นในวีดีโอ ณ วันที่ดึงข้อมูล

จากรูปที่ 3.2 รูปภาพแสดงตัวอย่างข้อมูลรายละเอียดของวีดีโอที่เกี่ยงกับสกุลเงินดิจิทัล ซึ่งเป็นตัวอย่างข้อมูลที่ถูกนำมาใช้ในการสร้างโมเดลการวิเคราะห์ถดถอยแบบโลจิสติก

	C	D	E	F	G	H	I	J	K
1	Channel_name	Channel_id	publishedAt	video_name	Description	Video_id	viewCount	likeCount	commentCount
2	Paul Pattarapon พอล ‡	UCunVACpk7PRWB1tz	2021-02-01T10:00:28Z	ห้ามพลาด !! ลงทุนใน	EP. นี้ต้องบอกเลยว่าห้า	oBhgk5Atla0	1590518	48578	936
3	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-04-14T10:09:41Z	Dogecoin จะไปถึง $1 ไ	Dogecoin จะไปถึง $1 ไ	Rw4GXtm0JKM	7129	152	9
4	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-04-18T10:36:06Z	เหรียญ Cardano (ADA)	เหรียญ Cardano (ADA)	YZsdNxSDIrE	95569	1910	56
5	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-04-22T01:41:25Z	เหรียญ ALTCOIN ที่น่าส	พูดคุยกันได้ที่ Discord	I9uDK_Y7NQc	59731	1558	36
6	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-04-27T13:46:43Z	เหรียญ IOST คืออะไร ใ	พูดคุยกันได้ที่ Discord	QdU1Ms7Jydg	184314	3727	120
7	CHAITA StudioTM	UCREmkzfFnO8yJm2o1	2021-05-04T15:39:35Z	วิเคราะห์ BITCOIN และ	วิเคราะห์ BITCOIN เ	L7uZJ4HPKGw	1410	38	2
8	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-05-05T02:24:50Z	(ข่าว) ใช่งานจริง DOGE	พูดคุยกันได้ที่ Discord	5hd-rVcBD6U	28274	777	44
9	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-05-10T13:29:55Z	เหรียญ SHIBA INU (SHI	ซื้อกาแฟให้ 1 แก้วhttp	eWMB3FJYm5I	116878	2088	159
10	เหมียวกวักทรัพย์ - Cat	UCoZH6WYW71Lct8pm	2021-05-12T07:45:30Z	พูดถึงและวิเคราะห์เหรีย	วันนี้ผมจะมาพูดถึง เหรี	MAfK7yodBHY	59853	1332	227
11	Crypto Dog	UC2JQKOnnNbw5Qnbi	2021-05-13T01:54:46Z	(ข่าว) แดงทั้งกระดาน /	ซื้อกาแฟให้ 1 แก้วhttp	ZVOwfXun9Kk	44498	1175	102

รูปที่ 3.2 แสดงข้อมูลบางส่วนของรายละเอียดวีดีโอที่ดึงจากเว็บไซต์คลังคลิปวิดีโอ

ในส่วนต่อไปจะแสดงข้อมูลตัวอย่างของความคิดเห็นของวีดีโอที่มีการพูดถึงสกุลเงินดิจิทัล ซึ่งความคิดเห็นในเว็บไซต์คลังคลิปวิดีโอ (YouTube) นั้นแตกต่างจากข้อความรูปแบบอื่น ทั้งมีไวยากรณ์ที่ผิดปกติอย่างมาก มีการใช้อิโมติคอนสูง การการแสดงความคิดเห็นชิงเสียดสี คิดเห็น โดยข้อมูลที่ทำการดึงความคิดเห็นของวีดีโอจะประกอบด้วยคุณลักษณะดังตารางที่ 3.2 และตัวอย่างข้อมูลความคิดเห็นของวีดีโอ ในรูปที่ 3.3

ตารางที่ 3.2 ตารางอธิบายชื่อคุณลักษณะข้อมูลความคิดเห็นของวีดีโอที่ได้จากการดึง API

ลำดับที่	ชื่อคอลัมน์	คำอธิบายชื่อคอลัมน์
1	Video_id	Id วีดีโอ
2	comment	ความคิดเห็น
3	likeCount	จำนวนที่กดถูกใจความคิดเห็น
4	publishedAt	วันที่แสดงความคิดเห็น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

	A	B	C	D
1	Video_id	comment	likeCount	publishedAt
2	uDYJhiCV2U8	ขอข้อมูล แนวทาง Dopple and Twin ด้วยครับ	0	2021-11-14T04:13:12Z
3	uDYJhiCV2U8	ชอบรายการนี้ สนุกดีมีความรู้ 😄	0	2021-11-14T04:08:37Z
4	uDYJhiCV2U8	พวกใจปลาซิวขายไป	0	2021-11-14T01:04:59Z
5	uDYJhiCV2U8	bitkub ท็อป เค้าเก่งนะ เรื่ององค์ความรู้ทางการเงิน และการตลาด (หลังปั้นบิตคับ) แต่ตอ	0	2021-11-13T23:19:43Z
6	uDYJhiCV2U8	คลิปก่อนพูดเรื่องSANDกำลังมาแรง จบคลิปไปซื้อแป๊ปเดียวพุ่งเลย แม่นจริง	0	2021-11-13T20:37:05Z
7	uDYJhiCV2U8	มันมีรีเสิร์ชอยู่นะครับว่าคนคิดน้อยส่วนใหญ่ใช้สัญชาตญาณเทรดหุ้นได้กำไรมากกว่าคนรู้	0	2021-11-13T19:36:12Z
8	uDYJhiCV2U8	จบรายการปุ๊บราคาพุ่งปั๊บ 5555+	2	2021-11-13T16:21:46Z
9	uDYJhiCV2U8	อาไร คือดิ่งหนัก 55	0	2021-11-13T16:15:07Z
10	uDYJhiCV2U8	ช่องนี้ให้ความรู้ที่ดีมากๆ ครับ .. ผมแชร์ไปให้เพื่อนๆ มากมาย บางเรื่องเราก็ไม่เข้าใจ ก็เล	3	2021-11-13T15:44:05Z
11	uDYJhiCV2U8	เขียวทั้งกระดาน 😱	1	2021-11-13T15:41:26Z
12	uDYJhiCV2U8	thumbnail อะ ไม่ต้องทำให้แพนิคขนาดนั้นก็ได้มั้ง	0	2021-11-13T15:41:05Z
13	uDYJhiCV2U8	ชอบๆ ในการสนทนานานาสาระ/การแสดงความคิดเห็นครับ ..	1	2021-11-13T15:41:01Z
14	uDYJhiCV2U8	สามพี่น้อง กลับมาแล้ว	1	2021-11-13T15:39:15Z

รูปที่ 3.3 แสดงความคิดเห็นบางส่วนของวีดีโอที่ดึงจากเว็บไซต์คลังคลิปวิดีโอ (YouTube)

ในส่วนสุดท้ายของการเก็บรวมรวมข้อมูลจะเป็นการดึงข้อมูลราคาของสกุลเงินดิจิทัลในช่วงวันที่ 1 มกราคม 2564 ถึง วันที่ 30 เมษายน 2565 เพื่อนำมาใช้ในการวิจัยครั้งนี้ด้วย เพราะราคานั้นก็มีความสำคัญและเป็นส่วนหนึ่งในการประกอบการตัดสินใจว่าจะลงทุนในสกุลเงินดิจิทัลหรือไม่ พร้อมทั้งทำให้เห็นถึงความผกผันของราคาและอุปสงค์ อุปธานของความต้องการลงทุนในสกุลเงินดิจิทัล

ในการวิจัยครั้งนี้ผู้พัฒนาได้ทำการดึงราคาของสกุลเงินดิจิทัลจากเว็บไซต์ CoinMarketCap ซึ่งเป็นข้อมูลจาก Binance เป็นกระดานเทรดสกุลเงินดิจทัลชั้นนำของโลก โดยข้อมูลที่ได้รับจากผู้บริการเว็บไซต์มีดังนี้

1. ข้อมูลวันเดือนปี (dt_price)
2. ราคาเปิดของวันนั้น (open)
3. ราคาที่สูงที่สุดในวันนั้น (high)
4. ราคาที่ต่ำที่สุดในวันนั้น (low)
5. ราคาปิดของวันนั้น (close)
6. ปริมาณการซื้อขายในวันนั้น (volume)

โดยความถี่ของราคาที่นำไปวิเคราะห์ความถี่เป็นวัน โดยทำการเลือกราคาของเหรียญที่ชื่อว่าบิตคอยต์ (Bitcoin) ดังแสดงข้อมูลราคาของสกุลเงินดิจิทัลรูปที่ 3.4

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

	A	B	C	D	E	F
1	dt_price	open	high	low	close	volume
2	2021-01-01 00:00:00	28,923.63	29,600.00	28,624.57	29,331.69	54,182.93
3	2021-01-02 00:00:00	29,331.70	33,300.00	28,946.53	32,178.33	129,993.87
4	2021-01-03 00:00:00	32,176.45	34,778.11	31,962.99	33,000.05	120,957.57
5	2021-01-04 00:00:00	33,000.05	33,600.00	28,130.00	31,988.71	140,899.89
6	2021-01-05 00:00:00	31,989.75	34,360.00	29,900.00	33,949.53	116,050.00
7	2021-01-06 00:00:00	33,949.53	36,939.21	33,288.00	36,769.36	127,139.20
8	2021-01-07 00:00:00	36,769.36	40,365.00	36,300.00	39,432.28	132,825.70
9	2021-01-08 00:00:00	39,432.48	41,950.00	36,500.00	40,582.81	139,789.96
10	2021-01-09 00:00:00	40,586.96	41,380.00	38,720.00	40,088.22	75,785.98
11	2021-01-10 00:00:00	40,088.22	41,350.00	35,111.11	38,150.02	118,209.54
12	2021-01-11 00:00:00	38,150.02	38,264.74	30,420.00	35,404.47	251,156.14
13	2021-01-12 00:00:00	35,410.37	36,628.00	32,531.00	34,051.24	133,948.15
14	2021-01-13 00:00:00	34,049.15	37,850.00	32,380.00	37,371.38	124,477.91
15	2021-01-14 00:00:00	37,371.38	40,100.00	36,701.23	39,144.50	102,950.39
16	2021-01-15 00:00:00	39,145.21	39,747.76	34,408.00	36,742.22	118,300.92
17	2021-01-16 00:00:00	36,737.43	37,950.00	35,357.80	35,994.98	86,348.43
18	2021-01-17 00:00:00	35,994.98	36,852.50	33,850.00	35,828.61	80,157.73

รูปที่ 3.4 แสดงราคาเหรียญบิตคอยต์บางส่วน

3.3 ขั้นตอนการจัดเตรียมข้อมูล

3.3.1 ขั้นตอนการติดฉลากข้อมูล

หลังจากได้เก็บรวบรวมข้อมูลความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอ (YouTube) มาแล้วนั้นต้องทำการติดฉลากให้ความคิดเห็น เนื่องจากต้องนำความคิดเห็นพร้อมผลเฉลยไปเป็นข้อมูลชุดการเรียนรู้ (Tanning Set) สำหรับใช้เทรนแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับเพื่อสร้างแบบจำลองวิเคราะห์ความรู้สึกแบบอัตโนมัต

โดยการทำติดฉลากความคิดเห็นทั้งหมดจำนวน 703 รายการ จาก 25 วีดีโอ โดยแบ่งออกเป็น 2 ผลเฉลย ประกอบด้วย POS (Positive) คือ ความคิดเห็นนั้นเป็นความรู้สึกเชิงบวก จำวน 366 รายการ และ NEG (Negative) คือ ความคิดเห็นนั้นเป็นความรู้สึกเชิงลบ จำนวน 337 รายการ โดยจะแสดงตัวอย่างการติดฉลากให้กับชุดข้อมูลเรียนรู้ดังรูปที่ 3.5 และ 3.6

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

	A	B
1	comment	label
2	BTC วิ่งขึ้นเพราะ London Hard fork ของ ETH กลัวจะน้อยหน้าหรือเปล่า	pos
3	ขอฟุ่งแรงๆ ไป 50,000+ เลย อิอิ	pos
4	วันนี้ที่ราคาบิดดอย เอ้ยย!! บิทคอยพุ่งขึ้นไปนั้นเหตุเกิดจากบริษัทใหญ่ๆเข้ามาลงทุนเพิ่มในปริมาณที่มากขึ้น	pos
5	คือหลายวันมานี้ราคา Bitcoin มันขึ้นเรื่อยๆ พอมาถึงวันสิ้นเดือน เงินเดือนออกแบบนี้ รายย่อยรายจิ๋ว ก็ซื้อเ	pos
6	ราคาขึ้นหรือไม่ขึ้นสำหรับหนูไม่สำคัญค่ะ สำคัญที่ถือนานแค่ไหนมากกว่าค่ะ ส่วนตัวหนูถือแต่ ETH นะค่ะ	pos
7	น่าจะมาจากที่เยอรมันออกมาประกาศให้กองทุนและสถาบันเข้ามาลงทุนใน Bitcoin	pos
8	ราคา Bitcoin พุ่งไปแตะ 4 หมื่นเหรียญต้นๆ.. มาจาก short squeeze แล้วต่อด้วย panic buy ครับ :)	pos
9	ขอบคุณสำหรับข่าวสารดีๆแบบนี้	pos
10	น่าจะเด้บโตอีกเยอะครับ ขอบคุณสำหรับข่าวสารครับ	pos
11	มีสถาบันใหญ่ๆ ให้ความสนใจมากขึ้นราคาจึงพุ่งขนาดนี้ครับ	pos
12	BTC is hope ครับ	pos
13	ขึ้นตาม ดีมาน ที่แท้จริงครับ ภาพเริ่มชัดเจนเจ้าเก็บของได้เยอะแล้ว ถึงเวลาปล่อยข่าวเชิงบวกละครับต่อจาก	pos
14	คนเริ่มสนใจมากขึ้นครับ	pos

รูปที่ 3.5 ตัวอย่างความคิดเห็นที่มีผลเฉลยเป็นความรู้สึกเชิงบวก

	A	B
1	comment	label
2	ไม่มีหลักประกัน อะไรเลยใน Bitcoin ต้องยอมรับความเสี่ยงนะ ปลาใหญ่กินปลาเล็ก	neg
3	หลอกทุบ แน่นอนครับ	neg
4	ถ้าขึ้นจริงก็ตกรถล่ะ ขายทิ้งทั้งพอร์ตไปลง Revo กับ Mist หมดแล้ว 555	neg
5	คนทำราคา โหดนะเนี่ย สลัดสบัดๆ รายเล็กๆ Hodl ไว้ครับทุกคน	neg
6	เหรียญที่ซื้อไม่เขียว เหรียญที่เขียวไม่ซื้อ	neg
7	ตอนนี้ตกละครับพี่น้องแต่ไม่รู้ว่าจะพุ่งอีกไหม(Zil)	neg
8	ข่าวมาช้าไปครับ เข้าตอนนี้ดอยครับ	neg
9	ขายหมู ดีกว่าติดดอยใช่มั้ยครับพี่	neg
10	ตกรถมารวมกันตรงนี้	neg
11	ตกรถครับ555	neg
12	ตกรถครับ	neg
13	Doge ข่าวดีมาเรื่อยๆ แต่ราคาไม่ขึ้นเลยครับ	neg
14	ตอนนี้สภาพเศษฐกิจโลก ไม่เอื้อให้กับการลงทุนในตอนนี้ btcจะลงมาอีกทีครับ 37000โดยประมาน	neg

รูปที่ 3.6 ตัวอย่างความคิดเห็นที่มีผลเฉลยเป็นความรู้สึกเชิงลบ

3.3.2 ขั้นตอนการทำความสะอาดข้อมูล

โดยขั้นตอนทำความสะอาดข้อมูล (Text Preprocessing) ผู้วิจัยจะดำเนินการตามหลักการที่อ้างอิงมาจากบทความการวิเคราะห์ความรู้สึก (10)

ก่อนนำข้อมูลเข้าไปเรียนรู้เพื่อสร้างแบบจำลองโครงข่ายประสามเทียมแบบวนกลับนั้น จะต้องมีการทำความสะอาดข้อมูลก่อนนำทำการวิเคราะห์ความรู้สึก โดยทำการ Normalization เป็นขั้นตอนการปรับรูปแบบของคำให้เหมือนกัน เช่น การแปลงให้เป็นตัวอักษรภาษาอังกฤษเป็นตัวเล็กทั้งหมด เช่น คำว่า Bitcoin จะแปลงเป็น bitcoin และ ETH จะแปลงเป็น eth ซึ่งเป็นตัวอักษรตัวเล็กทั้งหมดและทำการตัดเครื่องหมายวรรคตอน (punctuation) ดังแสดงในรูปที่ 3.7

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
# pun = string.punctuation
pun = '"#\'()*,-.;<=>[\\]^_`{|}~'
pun
```

```
'"#\'()*,-.;<=>[\\]^_`{|}~'
```

รูปที่ 3.7 เครื่องหมายวรรคตอนที่ทำการตัดออกจากประโยคความคิดเห็น

นอกจากนี้ยังทำการจำกัดคำที่ไม่สำคัญ (Stop words) ออกจากประโยคความคิดเห็น ทั้งคำที่ไม่มีความสำคัญภาษาไทยหรือ คำที่เจอบ่อยๆแต่ไม่สื่อความหมายอะไรมาก เช่น การ, ความ, คือ, ที่, ซึ่ง เป็นต้น และ คำที่ไม่มีความสำคัญภาษาอังกฤษ เช่น a, and, the ทิ้งไป ดังแสดงตามรูปที่ 3.8 คำที่ไม่มีความสำคัญภาษาอังกฤษ และ รูปที่ 3.9 คำที่ไม่มีความสำคัญภาษาไทย การตัด stop words จะมีข้อดี คือ ทำให้คลังคำศัพท์ (bag of words) ไม่ใหญ่เกินไปโดยไม่จำเป็น

```
list(stop_words)
```

```
['i',
 'me',
 'my',
 'myself',
 'we',
 'our',
 'ours',
 'ourselves',
 'you',
 "you're",
 "you've",
 "you'll",
 "you'd",
 'your',
 'yours',
 'yourself',
 'yourselves',
 'he',
 'him',
 'his',
```

รูปที่ 3.8 ตัวอย่างคำที่ไม่มีความสำคัญภาษาอังกฤษ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
thai_stopwords = list(thai_stopwords())
thai_stopwords
```

```
['ขณะ',
 'ไป',
 'เฉพาะ',
 'คราวใด',
 'ดังเก่า',
 'พร้อมทั้ง',
 'หรือเปล่า',
 'อย่างน้อย',
 'เสร็จสิ้น',
 'ต่อ',
 'แค่เพียง',
 'เรียก',
 'กำลังจะ',
 'เสียนี่',
 'คงจะ',
 'บางๆ',
 'ซึ่งได้แก่',
```

รูปที่ 3.9 ตัวอย่างคำที่ไม่มีความสำคัญภาษาไทย

หลังจากทำความสะอาดข้อมูลเสร็จแล้วจะได้ประโยค ดังแสดงรูปที่ 3.10

	comment	label	clean_comments
0	ตกรถครับ555	neg	ตกรถ555
1	ขายหมู ดีกว่าติดดอยใช่มั้ยครับพี่ 😭	neg	ขายหมู ดีกว่าติดดอยพี่ 😭
2	เลือก ETH แต่เอาไปต่ำกู้ BTC ครับ 55555	pos	เลือก eth ต่ำกู้ btc 55555
3	รอ eth $3150	pos	รอ eth $3150
4	LUNA ร่วงจังเลยครับ	neg	luna ร่วง

รูปที่ 3.10 ตัวอย่างประโยคที่ทำความสะอาด

3.3.3 ขั้นตอนการแปลงข้อมูล

หลังจากที่ได้ทำการทำความสะอาดข้อมูลแล้ว ต่อไปจะทำการแปลงข้อมูล (Text Preprocessing) เพื่อการวิเคราะห์ข้อความ

3.3.3.1 การตัดคำ (Tokenize)

นำข้อความหรือประโยคที่ทำความสะอาดแล้วมาดำเนินการตัดออกเป็นคำ โดยไลบรารี่สำหรับตัดคำภาษาไทยที่ใช้ตัดคำของการวิจัยนี้ คือ PyThaiNLP โดยใช้ฟังก์ชั่น word_tokenize ตัวอย่างประโยคที่ทำการตัดคำแล้ว ดังแสดงในรูปที่ 3.11

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

```
['ลงเรื่อยๆยังหาจุดยืนไม่เจอ']
[['ลง', 'เรื่อยๆ', 'ยัง', 'หา', 'จุดยืน', 'ไม่', 'เจอ']]
```

รูปที่ 3.11 ตัวอย่างประโยคที่ทำการตัดคำ

3.3.3.2 การสร้างคลังคำศัพท์ (Bag of words)

การสร้างคลังคำศัพท์ที่ไม่ซ้ำกันจากข้อความทั้งหมดของชุดข้อมูลเรียนรู้ และทำการแปลงคำแต่ละคำเป็นตัวเลข (unique id) จากชุดข้อมูลเรียนรู้ (training set) มีคำที่ไม่ซ้ำกันอยู่จำนวน 1,932 คำ ดังแสดงในรูปที่ 3.12

```
train_word_tokenizer = create_tokenizer(df['clean_comments'])
vocab_size = len(train_word_tokenizer.word_index) + 1

train_word_tokenizer.word_index

{'ผม': 1,
 'btc': 2,
 'ซื้อ': 3,
 'eth': 4,
 'ราคา': 5,
 'ขาย': 6,
 'เหรียญ': 7,
```

รูปที่ 3.12 ตัวอย่างคำในคลังคำศัพท์

3.3.3.3 หาความยาวของประโยค

ทำการสร้างฟังก์ชันเพื่อหาความยาวสูงสุดของของคำในประโยคในชุดข้อมูลเรียนรู้ ซึ่งเราจะค้นหาประโยคที่มีความยาวสูงสูดโดยใช้ Parameter key = len กำหนดความยาวสูงสุดของคำในประโยคให้กับ max_length เพื่อเตรียมทำ Padding ดังแสดงในรูปที่ 3.13

```
def max_length(words):
    return(len(max(words, key = len)))

max_length = max_length(cleaned_words)
max_length

349
```

รูปที่ 3.13 ความยาวสูงสุดของคำในประโยค

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.3.4 แปลงคำเป็นตัวเลข

แปลงคำในประโยคที่ได้ทำ Cleaning เป็นตัวเลข ด้วย Keras Tokenizer Object ยกตัวอย่างประโยคในชุดข้อมูลเรียนรู้ที่ทำการแปลงคำเป็นตัวเลข แสดงตามรูปที่ 3.14

```
encoded_doc = encoding_doc(train_word_tokenizer, cleaned_words)

print(cleaned_words[4])
print(encoded_doc[4])
```

```
เหรียญ ที่ ซื้อ ไม่ เขียว   เหรียญ ที่ เขียว ไม่ ซื้อ
[7, 1117, 3, 97, 7, 1117, 97, 3]
```

รูปที่ 3.14 ตัวอย่างประโยคที่แปลงคำเป็นตัวเลข

3.3.3.5 ทำการ Padding

การทำ Padding คือ ตัวเลขที่แทนแต่ละคำในประโยค โดยส่วนใหญ่จะกำหนดให้มีการเติม 0 เพื่อให้แต่ละประโยคมีความยาวเท่ากัน เนื่องจากการนำข้อมูลอินพุต (input) เข้าโครงข่ายประสาทเทียมแบบวนกลับนั้น ต้องทำให้ความยาวประโยคเท่ากันก่อน โดยจะแสดงตัวอย่างประโยคที่ทำการ padding แล้วในรูปที่ 3.15

```
padded_doc = padding_doc(encoded_doc, max_length)
print("Shape of padded docs = ",padded_doc.shape)

print(cleaned_words[4])
print(encoded_doc[4])
print(padded_doc[4])
```

```
Shape of padded docs =  (703, 349)
เหรียญ ที่ ซื้อ ไม่ เขียว   เหรียญ ที่ เขียว ไม่ ซื้อ
[7, 1117, 3, 97, 7, 1117, 97, 3]
[   7 1117    3   97    7 1117   97    3    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0    0
    0    0    0    0    0    0    0    0    0    0    0    0    0]
```

รูปที่ 3.15 ตัวอย่างประโยคที่ทำการ padding

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า
ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.3.6 แปลงผลเฉลยให้เป็นตัวเลข

โดยการนำข้อมูลเข้าโครงข่ายประสามเทียมแบบวนกลับจะต้องทำการแปลงผลเฉลยให้เป็นตัวเลขก่อนโดยใช้ output_tokenizer ดังแสดงผลลัพธ์ในรูปที่ 3.16

```
encoded_output = encoding_doc(output_tokenizer, label)
print(label)
print(encoded_output)
```

```
['pos', 'neg']
[[1], [2]]
```

รูปที่ 3.16 แปลงผลเฉลยให้เป็นตัวเลข

3.3.3.7 การเข้ารหัสผลเฉลยแบบวันฮอท (One-Hot Encoding)

การเข้ารหัสแบบวันฮอท (One-Hot Encoding) คือการทำข้อมูลที่ถูกเก็บในลักษณะตัวแปรเชิงกลุ่ม (Categorical Variable) ทั้งในลักษณะที่มีลำดับ (Ordinal number) และไม่มีลำดับ (Nominal number) เปลี่ยนให้อยู่ในรูปแบบของ Binary values ที่มีค่า 0 หรือ 1 เท่านั้น ข้อดีการเข้ารหัสแบบวันฮอท คือการแปลงข้อมูลให้เครื่องคอมพิวเตอร์สามารถเรียนรู้และเข้าใจตรรกะที่เป็นตัวเลขได้ ส่วนข้อเสียคือการใช้ทรัพยากรเครื่องที่เพิ่มขึ้นตามจำนวนข้อมูล เพราะเข้ารหัสแบบวันฮอท ข้อมูลถูกแปลงให้อยู่ในรูป Matrix ที่ประกอบไปด้วยเลข 0 เป็นจำนวนมาก ดังแสดงผลลัพธ์ในรูปที่ 3.17

```
output_one_hot = one_hot(encoded_output)
print(encoded_output[0])
print(output_one_hot[0])
```

```
[1]
[1. 0.]
```

รูปที่ 3.17 เข้ารหัสผลเฉลยแบบวันฮอท

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.3.4 ขั้นตอนการเพิ่มคุณลักษณะใหม่จากข้อมูล

การเพิ่มคุณลักษณะใหม่ (Features) จากข้อมูลในขั้นตอนนี้เป็นการเตรียมชุดข้อมูลสำหรับนำไปใช้ในการวิเคราห์องค์ประกอบหลักและการสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก เพื่อสร้างโมเดลในการทำนายว่าจะซื้อหรือไม่ซื้อสกุลเงินดิจิทัล และช่วยสร้างโมเดลที่ใช้สำหรับการสนับสนุนการตัดสินใจในการลงทุนดิจิทัล

การเพิ่มคุณลักษณะใหม่ (Features) โดยมีคุณลักษณะใหม่ ดังต่อไปนี้ โดยคุณลักษณะใหม่แสดงดังตารางที่ 3.3 และตัวอย่างข้อมูลคุณลักษณะใหม่ในรูปที่ 3.18

ตารางที่ 3.3 ตารางอธิบายชื่อคุณลักษณะใหม่

ลำดับที่	ชื่อคอลัมน์	คำอธิบายชื่อคอลัมน์
1	len_descvideo	ความยาวของคำอธิบายวีดีโอ
2	commentcount_pos	จำนวนความคิดเห็นที่ทำนายว่าเป็นเชิงบวกของวีดีโอ
3	commentcount_neg	จำนวนความคิดเห็นที่ทำนายว่าเป็นเชิงลบของวีดีโอ
4	len_max_comment	ความยาวสูงสุดของความคิดเห็นในวีดีโอ
5	len_min_comment	ความยาวต่ำสุดของความคิดเห็นในวีดีโอ
6	max_token	จำนวนความคิดเห็นที่มีคำสูงสุดในวิดีโอ
7	min_token	จำนวนความคิดเห็นที่มีคำต่ำสุดในวิดีโอ

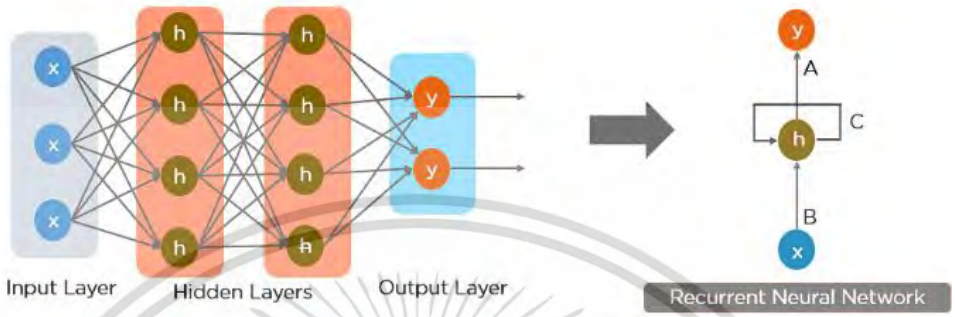
	A	B	C	D	E	F	G	H
1	len_descvideo	commentcount_pos	commentcount_neg	len_max_comment	len_min_comment	max_token	min_token	label
2	53	29	7	529	21	132	7	buy
3	959	16	10	253	11	68	4	not buy
4	1633	112	51	98	7	35	2	buy
5	735	193	66	169	6	43	2	not buy
6	964	4	3	114	10	42	2	buy
7	330	30	17	718	8	258	2	not buy
8	443	49	12	283	9	86	2	not buy
9	870	103	48	334	4	86	1	buy
10	644	28	6	402	6	114	1	buy
11	2110	13	5	428	1	152	1	not buy
12	712	99	66	313	6	85	1	buy
13	692	282	77	122	3	66	1	not buy
14	995	725	210	236	3	69	1	not buy
15	1503	74	27	403	5	109	1	not buy

รูปที่ 3.18 แสดงตัวอย่างข้อมูลของคุณลักษณะใหม่

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.4 ขั้นตอนการสร้างแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ

ขั้นตอนการฝึกอบรมและประเมินประสิทธิภาพ (Training and Evaluating Model) ของแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ ซึ่งในขั้นตอนแรกจะทำการกำหนดโครงสร้างของแบบจำลองโครงข่ายประสาทเทียมอธิบายได้ดังรูปที่ 3.19



รูปที่ 3.19 สถาปัตยกรรมโครงข่ายประสาทเทียมแบบวนกลับ

หมายเหตุ. จาก https://www.simplilearn.com

1. ชั้นข้อมูลเข้า (Input Layer) เป็นชั้นที่นำข้อมูลแบบจำลอง โดยในการวิจัยนี้ใช้ฟังก์ชันการเปิดใช้งาน (Activation Function) เป็น Relu
2. ชั้นซ่อน (Hidden Layer) เป็นชั้นที่อยู่ตรงกลาง ซึ่งจะมีผลอย่างมากต่อประสิทธิภาพในการเรียนรู้ของโมเดล ซึ่ง hidden layer นั้นจะมีกี่ชั้นก็ได้ โดยในชั้นนี้จะมี ชั้น Dense Layer 1 และชั้น Dense Layer 2 โดยที่มีการเปิดฟังก์ชันการใช้งาน (Activation Function) เป็น Relu
3. ชั้นข้อมูลออก (Output Layer) จำนวนโหนดของชั้นนำออกจะเท่ากับคลาสคำตอบ โดยในการวิจัยนี้จะใช้ฟังก์ชันการใช้งาน (Activation Function) เป็น Softmax

โดยการวิจัยนี้ทางผู้พัฒนาได้ออกแบบการทดลองโดยใช้ข้อมูลชุดการเรียนรู้ (Training Set) ชุดเดียวกันทั้งหมด แต่นำไปเข้าทำการฝึกอบรมและประเมินประสิทธิภาพของแบบบจำลองโครงข่ายประสาทเทียมแบบวนกลับที่แตกต่างกันทั้งหมด 3 อัลกอรึทึม และแต่ละอัลกอรึทึมทำการเปรียบเทียบประสิทธิภาพระหว่างแบบจำลองที่นำข้อมูลแบบมีอีโมติคอน และ แบบไม่มีอีโมติคอน โดยแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับที่แตกต่างกันทั้งหมด 3 อัลกอรึทึมดังนี้

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.4.1 แบบจำลองโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมแบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) ทำการทดลองดังตารางที่ 3.4

ตารางที่ 3.4 แสดงการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM

ลำดับ	รายละเอียดการทดลอง
1	ทดลองนำข้อมูลแบบไม่ลบอิโมติคอน โดยกำหนด Epoch = 300 และ Batch Size = 128 Input Layer และ Hidden Layer ใช้ Activation Function เป็น Relu Output Layer ใช้ Activation Function เป็น Softmax Optimizer ใช้ Adam และ Learning Rate = 0.001
2	ทดลองนำข้อมูลแบบลบอิโมติคอน โดยกำหนด Epoch = 300 และ Batch Size = 128 Input Layer และ Hidden Layer ใช้ Activation Function เป็น Relu Output Layer ใช้ Activation Function เป็น Softmax Optimizer ใช้ Adam และ Learning Rate = 0.001

3.4.2 แบบจำลองโครงข่ายประสาทเทียมโดยใช้อัลกอรึมแบบเกทแบบวนกลับ (Gated Recurrent Unit: GRU) ทำการทดลองดังตารางที่ 3.5

ตารางที่ 3.5 แสดงการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ GRU

ลำดับ	รายละเอียดการทดลอง
1	ทดลองนำข้อมูลแบบไม่ลบอิโมติคอน โดยกำหนด Epoch = 300 และ Batch Size = 128 Input Layer และ Hidden Layer ใช้ Activation Function เป็น Relu Output Layer ใช้ Activation Function เป็น Softmax Optimizer ใช้ Adam และ Learning Rate = 0.001
2	ทดลองนำข้อมูลแบบลบอิโมติคอน โดยกำหนด Epoch = 300 และ Batch Size = 128 Input Layer และ Hidden Layer ใช้ Activation Function เป็น Relu Output Layer ใช้ Activation Function เป็น Softmax Optimizer ใช้ Adam และ Learning Rate = 0.001

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.4.3 แบบจำลองโครงข่ายประสาทเทียมโดยใช้อัลกอริมแบบ Bidirectional Encoder Representations from Transformer: BERT) ทำการทดลองดังตารางที่ 3.6

ตารางที่ 3.6 แสดงการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ BERT

ลำดับ	รายละเอียดการทดลอง
1	ทดลองนำข้อมูลแบบไม่ลบอีโมติคอน โดยกำหนด Epoch = 5 และ Batch Size = 24 Learning Rate = 0.0002
2	ทดลองนำข้อมูลแบบลบอีโมติคอน โดยกำหนด Epoch = 5 และ Batch Size = 24 Learning Rate = 0. 0002

3.5 ขั้นตอนการติดฉลากแบบอัตโนมัต

หลังจากการฝึกอบรมและประเมินประสิทธิภาพแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับด้วยการทดลองที่แตกต่างกันออกไปแล้วนั้น พบว่าโครงข่ายประสาทเทียมแบบวนกลับแบบ Bidirectional Encoder Representations from Transformer และข้อมูลนำเข้าทดลองไม่ลบอิโมติคอนนั้นให้ความถูกต้องที่ดีสุด คือ 79 เปอร์เซนต์ ดังนั้นจึงนำโมเดลที่ผ่านการฝึกอบรมมาแล้วนั้นไปทำนายข้อมูลที่ไม่มีผลเฉลยจำนวน 7,795 ความคิดเห็น จากวีดีโอทั้งหมด 47 วีดีโอ ทำการติดฉลากแบบอัตโนมัต ดังแสดงตามรูปที่ 3.20

```
def predictions(text):
    predict = reloaded_predictor.predict(text)

    return predict

df_un['label'] = df_un['clean_comments'].apply(predictions)
df_un
```

```
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 30ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 37ms/step
1/1 [==============================] - 0s 31ms/step
1/1 [==============================] - 0s 31ms/step
```

รูปที่ 3.20 ติดฉลากแบบอัตโนมัตด้วยโมเดลที่ฝึกอบรมมาแล้ว

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6 ขั้นตอนการวิเคราห์องค์ประกอบหลักและการสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก

ในการวิจัยครั้งนี้จะคุณลักษณะต่าง ๆ ของวีดีโอที่เกี่ยวกับสกุลเงินดิจิทัลและราคามาทำนายว่าจะตัดสินใจซื้อหรือไม่ซื้อสกุลเงินดิจิทัล โดยมีคุณลักษณะทั้งหมด 15 คุณลักษณะ (Features)

เนื่องจากมีจำนวนคุณลักษณะที่เยอะ และอยากทราบว่าในการตัดสินใจว่าจะลงทุนในสกุลเงินดิจิทัลนั้นจะต้องดูคุณลักษณะไหนเป็นหลักในการประกอบการตัดสินใจว่าจะซื้อหรือไม่ซื้อหรือเป็นคุณลักษณะที่มีผลต่อการตัดสินใจในการลงทุนในสกุลเงินดิจิทัล โดยรายละเอียดของคุณลักษณะทั้ง 15 คุณลักษณะและผลเฉลยแสดงดังตารางที่ 3.7

ตารางที่ 3.7 ตารางแสดงคุณลักษณะตั้งต้นและผลเฉลย

ลำดับ	ชื่อคุณลักษณะ	คำอธิบายคุณลักษณะ
1	len_descvideo	ความยาวของคำอธิบายวีดีโอ
2	viewcount	จำนวนคนเข้าชมวีดีโอ ณ วันที่ดึงข้อมูล
3	likecount	จำนวนคนกดถูกใจวีดีโอ ณ วันที่ดึงข้อมูล
4	commentcount	จำนวนความคิดเห็นในวีดีโอ ณ วันที่ดึงข้อมูล
5	commentcount_pos	จำนวนความคิดเห็นที่ทำนายว่าเป็นเชิงบวกของวีดีโอ
6	commentcount_neg	จำนวนความคิดเห็นที่ทำนายว่าเป็นเชิงลบของวีดีโอ
7	len_max_comment	ความยาวสูงสุดของความคิดเห็นในวีดีโอ
8	len_min_comment	ความยาวต่ำสุดของความคิดเห็นในวีดีโอ
9	max_token	จำนวนความคิดเห็นที่มีคำสูงสุดในวิดีโอ
10	min_token	จำนวนความคิดเห็นที่มีคำต่ำสุดในวิดีโอ
11	open	ราคาเปิดของวันนั้น
12	high	ราคาที่สูงที่สุดในวันนั้น
13	low	ราคาที่ต่ำที่สุดในวันนั้น
14	close	ราคาปิดของวันนั้น
15	volumn	ปริมาณการซื้อขายในวันนั้น
16	label	การตัดสินใจลุงทุนในสกุลเงินดิจิทัล โดยแบ่งออกเป็น 2 หมวดหมู่ ดังนี้ 0 หมายถึง ตัดสินใจไม่ซื้อ 1 หมายถึง ตัดสินใจซื้อ

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6.1 การวิเคราห์องค์ประกอบหลัก (Principal Component Analysis: PCA)

ชุดข้อมูลวีดีโอที่นำเข้าไปวิเคราะห์องค์ประกอบหลักมีทั้งหมด 47 รายการ ประกอบด้วยคุณลักษณะตั้งต้น 15 ลักษณะตามที่ได้กล่าวในตารางที่ 3.7 โดยเป้าหมาย (Target) มีค่าเป็น 0 หมายถึง ตัดสินใจไม่ซื้อ และ 1 หมายถึง ตัดสินใจซื้อ โดยแสดงตัวอย่างชุดข้อมูลที่ใช้ในการประกอบการตัดสินใจลงทุนในสกุลเงินดิจิทัลดังรูปที่ 3.21

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	len_descv	viewcou	likecour	commentco	commentcount_pc	commentcount_neg	len_max	len_min	max_toke	min_to	open	high	low	close	volumn	label
2	964	5981	257	7	4	3	114	10	42	2	54716.5	57445.1	53256.64	57274.9	36163.7	0
3	1414	6544	411	23	17	6	359	1	110	0	41679.7	42786.7	41200.02	41864.6	22724.4	0
4	35	7129	152	9	7	2	331	3	97	1	41571.7	41913.7	40843.01	41382.6	32532.3	0
5	2378	10556	738	46	36	9	426	4	115	1	62896.5	63086.3	60677.01	61395	35930.9	0
6	2110	13443	566	18	13	5	428	1	152	1	34259	34678.4	32658.34	33086.6	48181.4	0
7	330	14097	810	47	30	17	718	8	258	2	56670	58000	48600	49631.3	99842.8	0
8	1004	15672	565	12	9	3	367	1	108	0	50053.9	50190	45672.75	46702.8	50869.5	0
9	494	17257	971	59	46	13	664	4	197	1	64122.2	65000	63360.22	64380	22505	0
10	0	19455	683	32	24	8	338	3	112	1	57138.3	59398.9	57000	58960.4	42153.5	0
11	116	28274	777	44	34	10	1330	4	431	1	43160	44949	42809.98	44421.2	61743.1	0
12	443	28378	759	61	49	12	283	9	86	2	60006.7	60499	50931.3	56150	124882	0

รูปที่ 3.21 แสดงตัวอย่างชุดข้อมูลที่ใช้ในการประกอบการตัดสินใจลงทุนในสกุลเงินดิจิทัล

จากชุดข้อมูลจะเห็นได้ว่าแต่ละคุณลักษณะมีช่วงของข้อมูลที่แตกต่างกันมากเกินไป ตัวอย่างเช่น คุณลักษณะจำนวนความคิดเห็นในวีดีโอ ณ วันที่ดึงข้อมูล (commentcount) กับ ปริมาณการซื้อขายในวันนั้น (volumn) เราจึงจำเป็นต้องทำการ Standardization หรือ Z-Score Normalization คุณลักษณะตั้งต้น

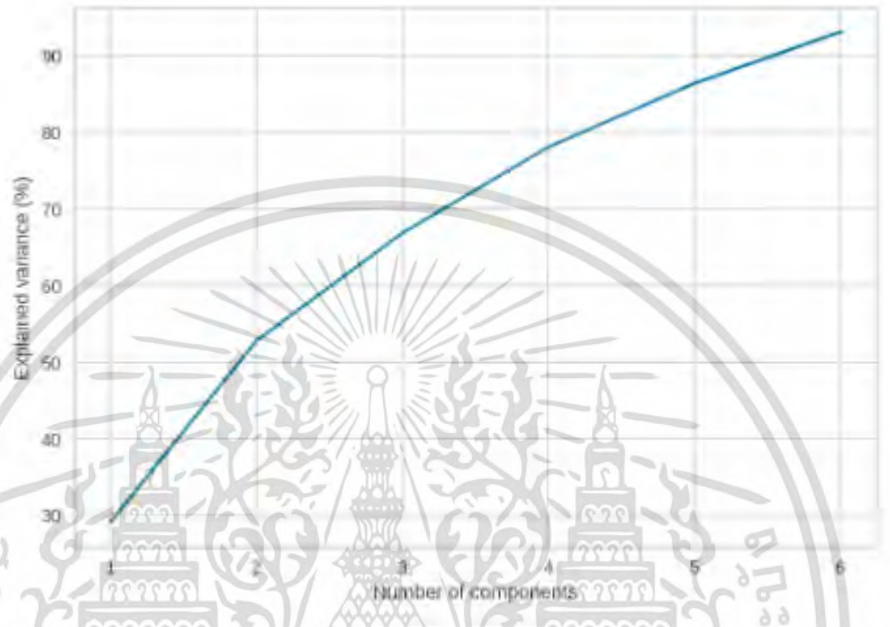
Standardization (z-score) คือ การนำข้อมูลคุณลักษณะ (Feature) มาปรับให้ค่าเฉลี่ย (Mean) มีค่าเท่ากับ 0 และ ค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) คือ การวัดการกระจายของกลุ่มข้อมูล มีค่าเท่ากับ 1 โดยคุณลักษณะตั้งต้นที่ทำการ Standardization แล้วแสดงดังรูปที่ 3.22

	z_len_descvideo	z_viewcount	z_likecount	z_commentcount	z_commentcount_pos	z_commentcount_neg
0	0.3763	-0.4809	-0.4931	-1.0026	-0.8316	-0.8432
1	1.2114	-0.4785	-0.4728	-0.8706	-0.7302	-0.7682
2	-1.3478	-0.4760	-0.5070	-0.9861	-0.8082	-0.8683
3	3.0004	-0.4614	-0.4296	-0.6809	-0.5821	-0.6932
4	2.5030	-0.4490	-0.4523	-0.9118	-0.7614	-0.7932

รูปที่ 3.22 แสดงตัวอย่างข้อมูลที่ทำ Standardization

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากนั้นนำค่าความแปรปรวนสะสม (Cumulative Variances) มาพล็อตรวมกันเป็นกราฟ เรียกว่า Scree plot เพื่อดูว่าการอธิบายของแต่ละ Components มีค่าเป็นเท่าใดและควรใช้ Components เท่าไหร่ในการลดมิติของคุณลักษณะ (Features) ดังแสดงกราฟ Scree plot ในรูปที่ 3.23



รูปที่ 3.23 กราฟ Scree plot

จะเห็นได้ว่าจากกราฟ Components โดยให้ n = 1 จะสามารถอธิบาย Variance ได้ 29.03 เปอร์เซ็นต์ แต่ถ้ามีการเพิ่ม Components ให้เป็น n = 4 จะสามารถอธิบายข้อมูลทั้งหมดได้ 78.05 เปอร์เซ็นต์ ค่าที่ออกมาจากทำการวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) คือ ค่าไอเก้น (Eigenvalues) ซึ่งเป็นค่าที่ใช้บอกว่าในการ แตกจำนวนคุณลักษณะออกมาเป็นส่วนย่อย (Decompose) ด้วยการวิเคราะห์องค์ประกอบหลักแล้ว Components ต่าง ๆ ที่เกิดขึ้นมาสามารถอธิบายค่าความแปรปรวน (Explain variance) ได้มากน้อยแค่ไหน ซึ่งค่าไอเก้นที่มีค่ามากสามารถอธิบายค่าความแปรปรวานได้มาก

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

3.6.2 การสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก (Logistic regression)

ทำการสร้างโมเดลเพื่อทำนายว่าจะตัดสินใจลงทุนในสกุลเงินดิจิทัลหรือไม่ จากชุดข้อมูลที่ได้กล่าวไปในตารางที่ 3.7 ผู้พัฒนาได้ทำการออกแบบการทดลองเพื่อทำการเปรียบเทียบประสิทธิภาพของโมเดลว่าการจัดเตรียมข้อมูลแบบใดให้ผลลัพธ์ที่ดีที่สุด จึงสามารถออกแบบการทดลองได้ดังตารางที่ 3.8

ตารางที่ 3.8 แสดงการสร้างโมเดลการจำแนกด้วยการวิเคราะห์ถดถอยโลจิสติก

ลำดับ	รายละเอียดการทดลอง
1	ใช้การวิเคราะห์ถดถอยโลจิสติกเพื่ออบรมโมเดลโดยใช้คุณลักษณะทั้งหมด 15 รายการในชุดข้อมูล โดยแบ่งเป็นชุดข้อมูลฝึกอบรม 70 เปอร์เซ็นต์และชุดข้อมูลทดลอง 30 เปอร์เซ็นต์
2	ใช้การวิเคราะห์ถดถอยโลจิสติกโดยใช้การวิเคราะห์องค์ประกอบหลัก (PCA) เพื่ออบรมโมเดลโดยใช้ Component เท่ากับ 4 โดยแบ่งเป็นชุดข้อมูลฝึกอบรม 70 เปอร์เซ็นต์และชุดข้อมูลทดลอง 30 เปอร์เซ็นต์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 4
ผลการวิจัยและการอภิปรายผล

ในบทนี้จะกล่าวถึงผลลัพธ์จากการดำเนินงานโดยแบ่งออกเป็น ผลการดำเนินงานของการสร้างแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับของการทดลองประสิทธิภาพแบบจำลองทั้งหมด 6 รายการ ผลการดำเนินงานของของการสร้างโมเดลการวิเคราะห์การถดถอยโลจิสติกทั้งหมด 2 รายการ ซึ่งสามารถเปรียบเทียบประสิทธิภาพของแบบจำลองด้วยการวัดผลลัพธ์จากความถูกต้อง (Accuracy) และการอภิปรายผลการดำเนินงาน

4.1 ผลการวิจัยของแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ

จากการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับทั้ง 6 รายการนั้น ใช้ชุดข้อมูลในการเรียนรู้ชุดเดียวกัน โดยสิ่งที่แตกต่างกันนั้นคือ อัลกอรึทึมในการเรียนรู้ของแบบแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับทั้งหมด 3 อัลกอรึทึม โดยแต่ละอัลกอรึทึมจะทำการเรียนรู้ทั้งแบบลบอีโมติคอนในประโยคและไม่ลบอิโมติคอนในประโยค

ผลลัพธ์จากแบบจำลองโครงข่ายประสาทเทียมแบบ Long Short-Term Memory การทดลองที่ 1 โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอนและ กำหนด Epoch = 300 และ Batch Size = 128 ซึ่งสามารถอธิบายได้จากรูปที่ 4.1 และรูปที่ 4.2 พบว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) มีค่าประมาณ 0.2 และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 3 จะเห็นได้ว่าทั้งสองค่าไม่ได้มีการลู่เข้าหากันจากกราฟค่าสูญเสีย (Loss) รูปที่ 4.1 พบว่ากราฟมีการฝึกอบรมมากขึ้น ค่าการฝึกอบรม (Train Loss) จะลงอย่างต่อเนื่องในขณะที่ค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) จะลดลงถึงจุดหนึ่งแล้วหลังจากนั้นค่าจะกลับมาเพิ่มขึ้นเรื่อย ๆ จึงสรุปได้ว่าแบบจำลองเกิดปัญหา Overfitting

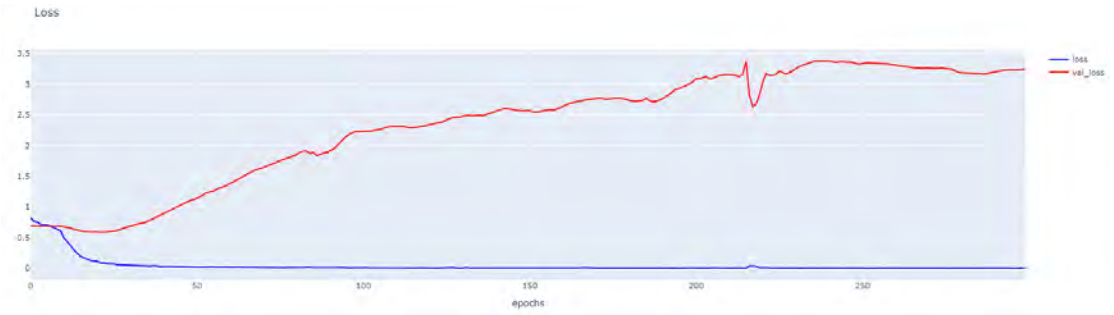
อีกทั้งยังมีค่าความถูกต้องจากข้อมูลฝึกอบรมและค่าความถูกต้องจากข้อมูลตรวจสอบอยู่ประมาณ 0.9 และ 0.67 ตามลำดับ และเมื่อนำแบบจำลองนี้ไปใช้กับข้อมูลทดสอบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ 0.6738 หรือ 67.38 เปอร์เซ็นต์ และมีค่า คะแนนเอฟวัน (F1-Score) อยู่ที่ประมาณ 0.6736

จากรูปที่ 4.1 แสดงค่าสูญเสีย (Loss) จากผลการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น loss หมายถึงค่าสูญเสียจากชุดข้อมูลฝึกอบรม (Train Loss) มีค่าประมาณ 0.2

2) เส้น val_loss หมายถึงค่าสูญเสียจากชุดข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 3

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

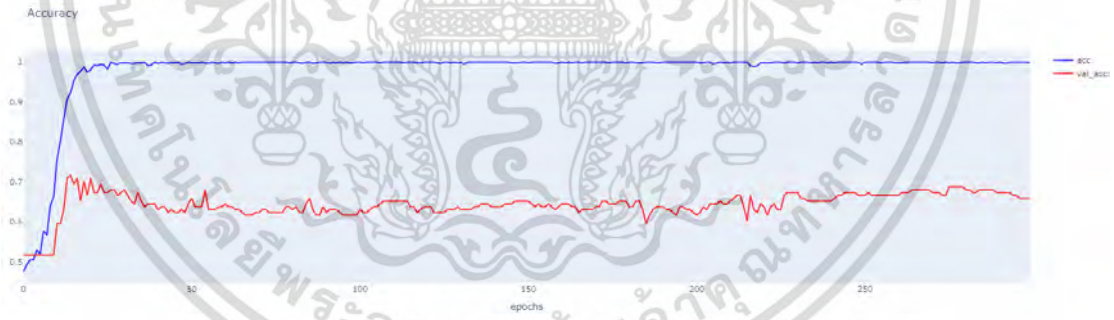


รูปที่ 4.1 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน

จากรูปที่ 4.2 แสดงค่าความถูกต้องของข้อมูลฝึกอบรมและชุดข้อมูลตรวจสอบของการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น acc หมายถึงค่าความถูกต้องจากชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 90 เปอร์เซ็นต์

2) เส้น acc_loss หมายถึงค่าความถูกต้องชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 67 เปอร์เซ็นต์



รูปที่ 4.2 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน

ผลลัพธ์จากแบบจำลองโครงข่ายประสาทเทียมแบบ Long Short-Term Memory การทดลองที่ 2 โดยทดลองกับข้อมูลแบบลบอีโมติคอนและ กำหนด Epoch = 300 และ Batch Size = 128 ซึ่งสามารถอธิบายได้จากรูปที่ 4.3 และรูปที่ 4.4 พบว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) มีค่าประมาณ 0.1 และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 3 จะเห็นได้ว่าทั้งสองค่าไม่ได้มีการลู่เข้าหากันจากกราฟค่าสูญเสีย (Loss) รูปที่ 4.3 พบว่ากราฟมีการฝึกอบรมมากขึ้น ค่าการฝึกอบรม (Train Loss) จะลงอย่างต่อเนื่องในขณะที่ค่าสูญเสียจากข้อมูล

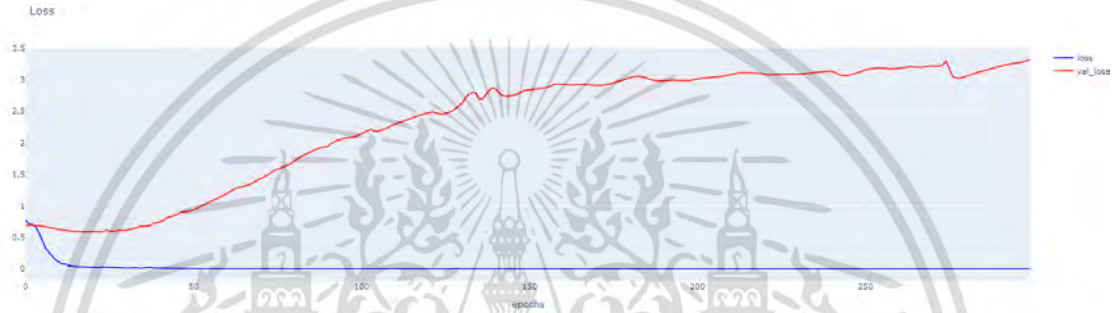
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตรวจสอบ (Validation Loss) จะลดลงถึงจุดหนึ่งแล้วหลังจากนั้นค่าจะกลับมาเพิ่มขึ้นเรื่อย ๆ จึงสรุปได้ว่าแบบจำลองเกิดปัญหา Overfitting อีกทั้งยังมีค่าความถูกต้องจากข้อมูลฝึกอบรมและค่าความถูกต้องจากข้อมูลตรวจสอบอยู่ประมาณ 0.9 และ 0.73 ตามลำดับ และเมื่อนำแบบจำลองนี้ไปใช้กับข้อมูลทดสอบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ 0.7305 หรือ 73.05 เปอร์เซ็นต์ และมีค่าคะแนนเอฟวัน (F1-Score) อยู่ที่ประมาณ 0.7301

จากรูปที่ 4.3 แสดงค่าสูญเสีย (Loss) จากผลการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น loss หมายถึงค่าสูญเสียจากชุดข้อมูลฝึกอบรม (Train Loss) มีค่าประมาณ 0.1

2) เส้น val_loss หมายถึงค่าสูญเสียจากชุดข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 3

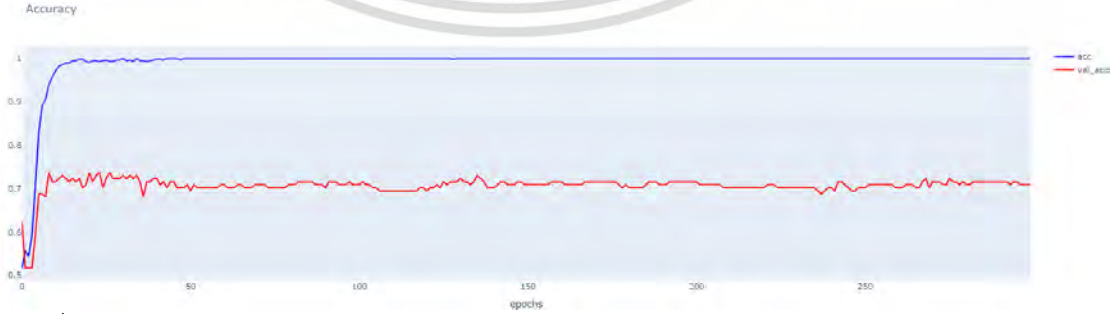


รูปที่ 4.3 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบลบอีโมติคอน

จากรูปที่ 4.4 แสดงค่าความถูกต้องของข้อมูลลฝึกอบรมและชุดข้อมูลตรวจสอบของผการทดลองลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น acc หมายถึงค่าความถูกต้องจากชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 90 เปอร์เซ็นต์

2) เส้น acc_loss หมายถึงค่าความถูกต้องชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 73 เปอร์เซ็นต์



รูปที่ 4.4 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ LSTM โดยทดลองกับข้อมูลแบบลบอีโมติคอน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ผลลัพธ์จากแบบจำลองโครงข่ายประสาทเทียมแบบ Gated Recurrent Units การทดลองที่ 1 โดยทดลองข้อมูลแบบไม่ลบอีโมติคอนและ กำหนด Epoch = 300 และ Batch Size = 128 ซึ่งสามารถอธิบายได้จากรูปที่ 4.5 และรูปที่ 4.6 พบว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) มีค่าประมาณ 0.1 และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 2.5 รูปที่ 4.5 พบว่ากราฟมีการฝึกอบรมมากขึ้น ค่าการฝึกอบรม (Train Loss) จะลงอย่างต่อเนื่องในขณะที่ค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) จะลดลงถึงจุดหนึ่งแล้วหลังจากนั้นค่าจะกลับมาเพิ่มขึ้นเรื่อย ๆ จึงสรุปได้ว่าแบบจำลองเกิดปัญหา Overfitting อีกทั้งยังมีค่าความถูกต้องจากข้อมูลฝึกอบรมและค่าความถูกต้องจากข้อมูลตรวจสอบอยู่ประมาณ 0.9 และ 0.69 ตามลำดับ และเมื่อนำแบบจำลองนี้ไปใช้กับข้อมูลทดสอบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ 0.6950 หรือ 69.50 เปอร์เซ็นต์ และมีค่าคะแนนเอฟวัน (F1-Score) อยู่ที่ประมาณ 0.6947

จากรูปที่ 4.5 แสดงค่าสูญเสีย (Loss) จากผลการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ GRUโดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น loss หมายถึงค่าสูญเสียจากชุดข้อมูลฝึกอบรม (Train Loss) มีค่าประมาณ 0.1

2) เส้น val_loss หมายถึงค่าสูญเสียจากชุดข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 2.5

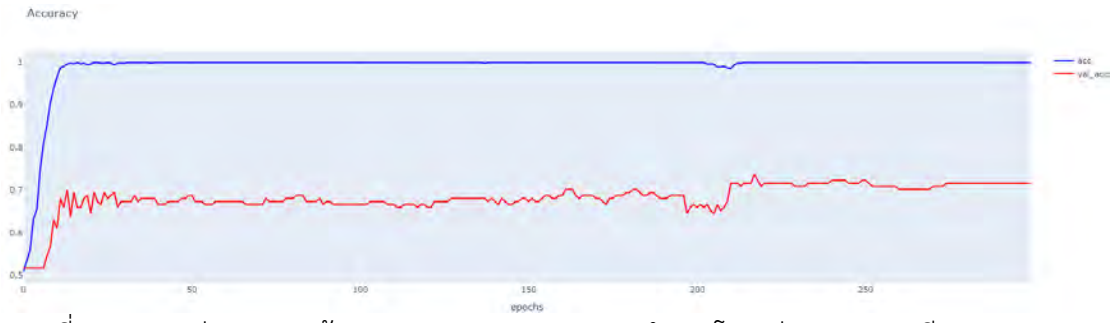
รูปที่ 4.5 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน

จากรูปที่ 4.6 แสดงค่าความถูกต้องของข้อมูลลฝึกอบรมและชุดข้อมูลตรวจสอบของผการทดลองลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น acc หมายถึงค่าความถูกต้องจากชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 90 เปอร์เซ็นต์

2) เส้น acc_loss หมายถึงค่าความถูกต้องชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 69 เปอร์เซ็นต์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.6 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน

ผลลัพธ์จากแบบจำลองโครงข่ายประสาทเทียมแบบ Gated Recurrent Units การทดลองที่ 2 โดยทดลองข้อมูลแบบลบอีโมติคอนและ กำหนด Epoch = 300 และ Batch Size = 128 ซึ่งสามารถอธิบายได้จากรูปที่ 4.7 และรูปที่ 4.8 พบว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) มีค่าประมาณ 0.1 และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 2.5 รูปที่ 4.7 พบว่ากราฟมีการฝึกอบรมมากขึ้น ค่าการฝึกอบรม (Train Loss) จะลงอย่างต่อเนื่องในขณะที่ค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) จะลดลงถึงจุดหนึ่งแล้วหลังจากนั้นค่าจะกลับมาเพิ่มขึ้นเรื่อย ๆ จึงสรุปได้ว่าแบบจำลองเกิดปัญหา Overfitting อีกทั้งยังมีค่าความถูกต้องจากข้อมูลฝึกอบรมและค่าความถูกต้องจากข้อมูลตรวจสอบอยู่ประมาณ 0.9 และ 0.7 ตามลำดับ และเมื่อนำแบบจำลองนี้ไปใช้กับข้อมูลทดสอบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ 0.7234 หรือ 72.34 เปอร์เซ็นต์ และมีค่าคะแนนเอฟวัน (F1-Score) อยู่ที่ประมาณ 0.7188

จากรูปที่ 4.7 แสดงค่าสูญเสีย (Loss) จากผลการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ GRUโดยทดลองกับข้อมูลแบบลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น loss หมายถึงค่าสูญเสียจากชุดข้อมูลฝึกอบรม (Train Loss) มีค่าประมาณ 0.1

2) เส้น val_loss หมายถึงค่าสูญเสียจากชุดข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 2.5

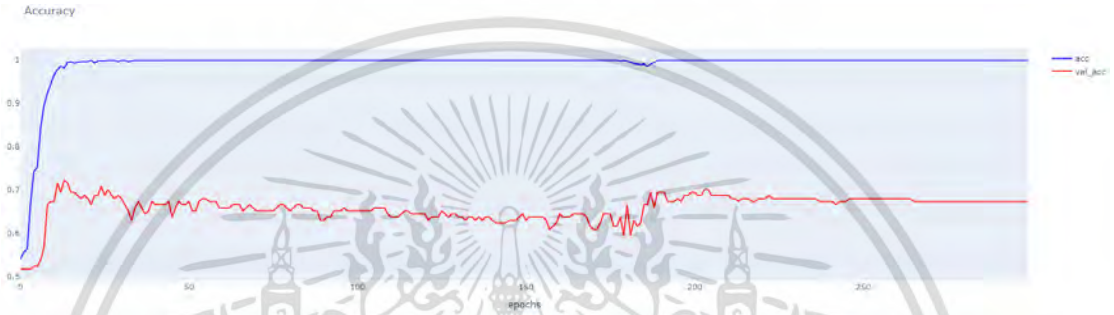
รูปที่ 4.7 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบลบอีโมติคอน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

จากรูปที่ 4.8 แสดงค่าความถูกต้องของข้อมูลฝึกอบรมและชุดข้อมูลตรวจสอบของผการทดลองลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น acc หมายถึงค่าความถูกต้องจากชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 90 เปอร์เซ็นต์

2) เส้น acc_loss หมายถึงค่าความถูกต้องชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 70 เปอร์เซ็นต์



รูปที่ 4.8 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ GRU โดยทดลองกับข้อมูลแบบลบอีโมติคอน

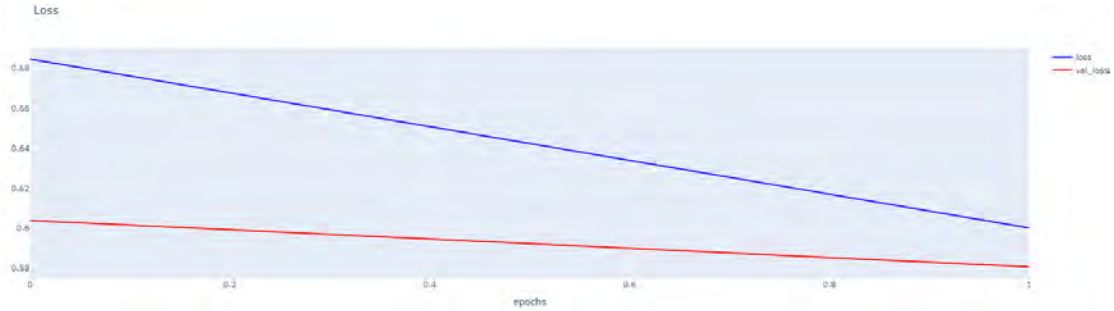
ผลลัพธ์จากแบบจำลองโครงข่ายประสาทเทียมแบบ Bidirectional Encoder Representations from Transformer การทดลองที่ 1 โดยทดลองข้อมูลแบบไม่ลบอีโมติคอนและกำหนด Epoch = 2 และ Batch Size = 24 ซึ่งสามารถอธิบายได้จากรูปที่ 4.9 และรูปที่ 4.10 พบว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) มีค่าประมาณ 0.6 และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 0.58 รูปที่ 4.9 พบว่ากราฟค่าสูญเสีย (Loss) จะเห็นว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าลดลงอย่างต่อเนื่อง ซึ่งกราฟทั้งสองเส้นจะมีช่องว่างระหว่างกันน้อยแสดงว่าเป็นแบบจำลองที่มีการเรียนรู้ที่ดี สามารถนำไปใช้ทำนายข้อมูลที่ไม่เคยพบเห็นมาก่อยได้อย่างแม่นยำ อีกทั้งยังมีค่าความถูกต้องจากข้อมูลฝึกอบรมและค่าความถูกต้องจากข้อมูลตรวจสอบอยู่ประมาณ 0.69 และ 0.79 ตามลำดับ และเมื่อนำแบบจำลองนี้ไปใช้กับข้อมูลทดสอบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ 0.79 หรือ 79 เปอร์เซ็นต์ และมีค่าคะแนนเอฟวัน (F1-Score) อยู่ที่ประมาณ 0.79

จากรูปที่ 4.9 แสดงค่าสูญเสีย (Loss) จากผลการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น loss หมายถึงค่าสูญเสียจากชุดข้อมูลฝึกอบรม (Train Loss) มีค่าประมาณ 0.6

2) เส้น val_loss หมายถึงค่าสูญเสียจากชุดข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 0.58

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้



รูปที่ 4.9 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน

จากรูปที่ 4.10 แสดงค่าความถูกต้องของข้อมูลลฝึกอบรมและชุดข้อมูลตรวจสอบของผการทดลองลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น acc หมายถึงค่าความถูกต้องจากชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 69 เปอร์เซ็นต์

2) เส้น acc_loss หมายถึงค่าความถูกต้องชุดข้อมูลฝึกอบรม Train Accuracy) มีค่าประมาณ 79 เปอร์เซ็นต์

รูปที่ 4.10 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน

ผลลัพธ์จากแบบจำลองโครงข่ายประสาทเทียมแบบ Bidirectional Encoder Representations from Transformer การทดลองที่ 2 โดยทดลองข้อมูลแบบไม่ลบอีโมติคอนและกำหนด Epoch = 5 และ Batch Size = 24 ซึ่งสามารถอธิบายได้จากรูปที่ 4.11 และรูปที่ 4.12 พบว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) มีค่าประมาณ 0.64 และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 0.58 รูปที่ 4.11 พบว่ากราฟค่าสูญเสีย (Loss) จะเห็นว่าค่าสูญเสียจากข้อมูลการฝึกอบรม (Train Loss) และค่าสูญเสียจากข้อมูลตรวจสอบ (Validation Loss) มีค่าลดลงอย่างต่อเนื่อง ซึ่งกราฟทั้งสองเส้นจะมีช่องว่างระหว่างกันน้อยแต่ยังน้อยกว่าแบบจำลองโครงข่ายประสาทเทียมแบบ BERTโดยทดลองกับข้อมูลแบบไม่ลบอีโมติคอน อีกทั้งยังมีค่า

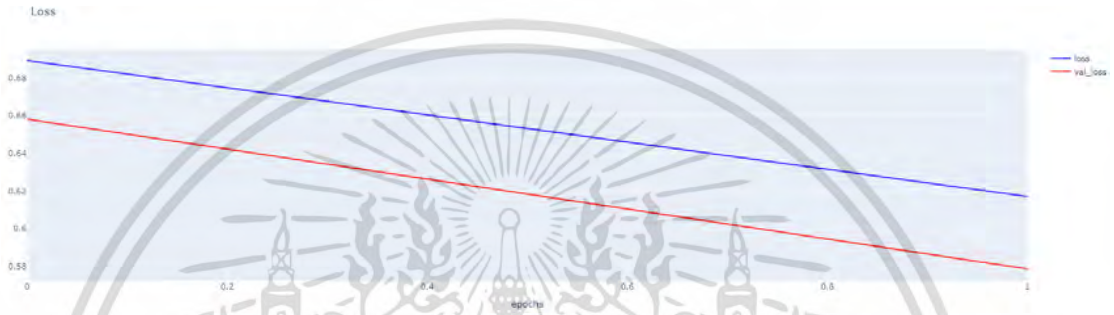
เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ความถูกต้องจากข้อมูลฝึกอบรมและค่าความถูกต้องจากข้อมูลตรวจสอบอยู่ประมาณ 0.65 และ 0.63 ตามลำดับ และเมื่อนำแบบจำลองนี้ไปใช้กับข้อมูลทดสอบให้ค่าความถูกต้อง (Accuracy) อยู่ที่ 0.63 หรือ 63 เปอร์เซ็นต์ และมีค่าคะแนนเอฟวัน (F1-Score) อยู่ที่ประมาณ 0.61

จากรูปที่ 4.11 แสดงค่าสูญเสีย (Loss) จากผลการทดลองแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น loss หมายถึงค่าสูญเสียจากชุดข้อมูลฝึกอบรม (Train Loss) มีค่าประมาณ 0.64

2) เส้น val_loss หมายถึงค่าสูญเสียจากชุดข้อมูลตรวจสอบ (Validation Loss) มีค่าประมาณ 0.58

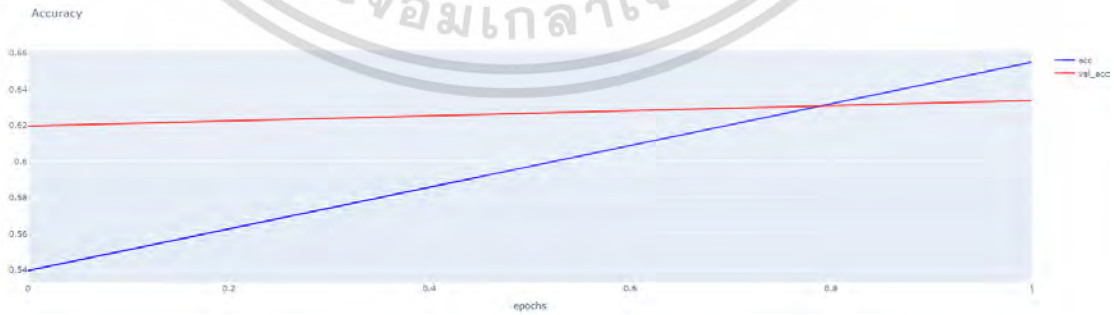


รูปที่ 4.11 แสดงค่าสูญเสีย (Loss) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบลบอีโมติคอน

จากรูปที่ 4.12 แสดงค่าความถูกต้องของข้อมูลลฝึกอบรมและชุดข้อมูลตรวจสอบของผการทดลองลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบลบอีโมติคอน ซึ่งมีรายละเอียดดังนี้

1) เส้น acc หมายถึงค่าความถูกต้องจากชุดข้อมูลฝึกอบรม (Train Accuracy) มีค่าประมาณ 65 เปอร์เซ็นต์

2) เส้น acc_loss หมายถึงค่าความถูกต้องชุดข้อมูลฝึกอบรม (Train Accuracy) มีค่าประมาณ 63 เปอร์เซ็นต์



รูปที่ 4.12 แสดงค่าความถูกต้อง (Accuracy) จากผลแบบจำลองโครงข่ายประสาทเทียมแบบ BERT โดยทดลองกับข้อมูลแบบลบอีโมติคอน

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ตารางที่ 4.1 เปรียบเทียบประสิทธิภาพของแบบจำลองแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ

ลำดับที่	Model	Epoch	Batch Size	Accuracy (%)	F1-Score
1	LSTM	300	128	67.38	0.67
2	LSTM (Remove Emoji)	300	128	73.05	0.73
3	GRU	300	128	69.50	0.69
4	GRU (Remove Emoji)	300	128	72.34	0.71
5	BERT	2	24	79	0.79
6	BERT (Remove Emoji)	2	24	63	0.61

4.2 ผลการวิจัยของการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติก

จากการทดลองการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติกเพื่อทำนายว่าจะตัดสินใจลงทุนในสกุลเงินดิจิทัลหรือไม่ทั้ง 2 รายการนั้น ได้ผลลัพธ์ดังนี้

ผลลัพธ์จากการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติก การทดลองที่ 1 โดยใช้คุณลักษณะทั้งหมด 15 รายการ ไม่มีการปรับช่วงของข้อมูลโดยมีจำนวนข้อมูล 47 รายการ โดยแบ่งเป็นชุดข้อมูลฝึกอบรม 70 เปอร์เซ็นต์และชุดข้อมูลทดลอง 30 เปอร์เซ็นต์ โดยให้ค่าความถูกต้องอยู่ที่ 66 เปอร์เซ็นต์

ผลลัพธ์จากการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติก การทดลองที่ 2 โดยใช้การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) ที่มีการปรับช่วงของข้อมูลโดยมีจำนวนข้อมูล 47 รายการ โดยแบ่งเป็นชุดข้อมูลฝึกอบรม 70 เปอร์เซ็นต์และชุดข้อมูลทดลอง 30 เปอร์เซ็นต์ โดยให้ค่าความถูกต้องอยู่ที่ 73 เปอร์เซ็นต์ จากการใช้ component เท่ากับ 4 ซึ่งจะสามารถอธิบายข้อมูลได้ทั้งหมด 78.05 เปอร์เซ็นต์

ตารางที่ 4.2 เปรียบเทียบประสิทธิภาพของแบบจำลอง Logistic Regression

ลำดับที่	รายละเอียดการทดลอง	ค่าความถูกต้อง
1	ใช้การวิเคราะห์ถดถอยโลจิสติกเพื่ออบรมโมเดลโดยใช้คุณลักษณะทั้งหมด 15 รายการในชุดข้อมูล โดยแบ่งเป็นชุดข้อมูลฝึกอบรม 70 เปอร์เซ็นต์และชุดข้อมูลทดลอง 30 เปอร์เซ็นต์	66.66
2	ใช้การวิเคราะห์ถดถอยโลจิสติกโดยใช้การวิเคราะห์องค์ประกอบหลัก (PCA) เพื่ออบรมโมเดลโดยใช้ Component เท่ากับ 4 โดยแบ่งเป็นชุดข้อมูลฝึกอบรม 70 เปอร์เซ็นต์และชุดข้อมูลทดลอง 30 เปอร์เซ็นต์	73.33

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

4.3 อภิปรายผล

จากการทดลองการวิเคราห์ความรู้สึกความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอเกี่ยวกับสกุลเงินดิจิทัลทั้ง 6 แบบ พบว่าการทดลองด้วยโครงข่ายประสาทเทียมโดยใช้อัลกอรึมแบบ Bidirectional Encoder Representations from Transformer: BERT) นั้นให้ผลลัพธ์ดีที่สุด โดยแบบจำลองนั้นไม่ท่องจำข้อมูลในการเรียนรู้มากเกินไป โดยมีค่าคะแนนเอฟวัน (F1-Score) ของผลเฉลยที่ทำนายเป็นความรู้สึกเชิงบวก (Positive) อยู่ที่ 0.80 ส่วนค่าคะแนนเอฟวัน (F1-Score) ของผลเฉลยที่ทำนายเป็นความรู้สึกเชิงลบ (Negative) อยู่ที่ 0.78 และมีค่าความถูกต้องอยู่ที่ 79 เปอร์เซนต์ ถือว่าเป็นแบบจำลองที่มีประสิทธิภาพค่อนข้างดี

ในส่วนของการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติกเพื่อทำนายว่าจะตัดสินใจลงทุนในสกุลเงินดิจิทัลทั้ง 2 แบบ จากผลลัพธ์สรุปได้ว่าสร้างโมเดลการวิเคราะห์ถดถอยโลจิสติกโดยใช้การวิเคราะห์องค์ประกอบหลัก (Principal Component Analysis) นั้นให้ผลลัพธ์ที่แม่นยำกว่าผลลัพธ์จากการสร้างโมเดลการวิเคราะห์ถดถอยโลจิสโดยใช้คุณลักษณะทั้งหมด 15 รายการ จากการใช้ component เท่ากับ 4 ซึ่งจะสามารถอธิบายข้อมูลได้ทั้งหมด 78.05 เปอร์เซ็นต์โดยโมเดลให้ค่าความถูกต้องอยู่ที่ 73 เปอร์เซ็นต์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

การค้นคว้าอิสระนี้จัดทำขึ้นเพื่อทำการวิเคราะห์ความรู้สึกสำหรับการลงทุน ในสกุลเงินดิจิทัล โดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอที่ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับ และเพื่อทำนายการตัดสินใจลงทุนในสกุลเงินดิจิทัล โดยในการวิจัยครั้งนี้ได้ทำการทดลองโมเดลการวิเคราะห์ถดถอยโลจิสโดยใช้คุณลักษณะทั้งหมดกับการทดลองโมเดลการวิเคราะห์ถดถอยโลจิสโดยใช้การวิเคราะห์องค์ประกอบหลัก ซึ่งในบทนี้จะกล่าวถึงข้อสรุปของการดำเนินงานในครั้งนี้

5.1 สรุปผลการวิจัย

การค้นคว้าอิสระนี้จัดทำขึ้นเพื่อทำการวิเคราะห์ความรู้สึกสำหรับการลงทุน ในสกุลเงินดิจิทัล โดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอที่ได้ทำการเปรียบเทียบประสิทธิภาพของแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับทั้งหมด 3 อัลกอริทึม ได้แก่ 1. แบบจำลองโครงข่ายประสาทเทียมโดยใช้อัลกอริทึมแบบหน่วยความจำระยะสั้นแบบยาว (Long Short-Term Memory: LSTM) 2. แบบจำลองโครงข่ายประสาทเทียมโดยใช้อัลกอริมแบบเกทแบบวนกลับ (Gated Recurrent Unit: GRU) 3. แบบจำลองโครงข่ายประสาทเทียมโดยใช้อัลกอริมแบบ Bidirectional Encoder Representations from Transformer: BERT) และทั้ง 3 อัลกอริทึมจะทำการเปรียบเทียบประสิทธิภาพการนำข้อมูลไปเรียนรู้แบบไม่ลบอิโมติคอนและแบบลบอิโมติคอน

จากผลการทดลองสร้างแบบจำลองด้วยชุดข้อมูลความคิดเห็นเกี่ยวกับสกุลเงินดิจิทัล พบว่าแบบจำลองที่ฝึกฝนโครงข่ายประสาทเทียมโดยใช้อัลกอริมแบบ Bidirectional Encoder Representations from Transformer: BERT) ให้ค่าความถูกต้องสูงสุดอยู่ที่ 79 เปอร์เซ็นต์ โดยอ้างอิงจากตารางที่ 4.1 จึงถูกเลือกเป็นแบบจำลองที่ใช้ในการวิเคราะห์ความรู้สึกแบบอัตโนมัติด้วยข้อมูลที่แบบจำลองไม่เคยเรียนรู้ทั้งหมด 7,795 รายการ จากจำนวน 47 วีดีโอ โดยทำนายออกมาเป็นความรู้สึกเชิงบวก (Positive) เป็นจำนวน 5,825 รายการ และ ความรู้สึกเชิงลบ (Negative) เป็นจำนวน 1,970 รายการ

จากนั้นนำชุดข้อมูลคุณลักษณะของวีดีโอที่เกี่ยวกับสกุลเงินดิจิทัล ทั้งหมดจำนวน 47 วีดีโอไปทำการทำนายเพื่อใช้ในการตัดสินใจลงทุนในสกุลเงินดิจิทัล โดยในการวิจัยครั้งนี้ได้ทำการทดลองโมเดลการวิเคราะห์ถดถอยโลจิสโดยใช้คุณลักษณะทั้งหมด 15 รายการ และไม่ได้ทำการปรับช่วงของข้อมูล พบว่าให้ค่าความถูกต้องต่ำกว่าการทดลองโมเดลการวิเคราะห์ถดถอยโลจิสโดยใช้การวิเคราะห์

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

องค์ประกอบหลัก (Principal Component Analysis) โดยการแตกย่อยจำนวนคุณลักษณะออกมาเป็นส่วนย่อย (Decompose) หรือจำนวน component เท่ากับ 4 นั้น พบว่าให้ค่าความถูกต้องของโมเดล 73 เปอร์เซ็นต์ โดยอ้างอิงจากตารางที่ 4.2 และสามารถอธิบายข้อมูลได้ทั้งหมด 78.05 เปอร์เซ็นต์

5.2 ขอบเขตและข้อจำกัด

จากการดำเนินการค้นคว้าอิสระครั้งนี้เพื่อการวิเคราะห์ความรู้สึกสำหรับการลงทุน ในสกุลเงินดิจิทัลโดยใช้ความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอและทำนายการตัดสินใจลงทุนในสกุลเงินดิจิทัลนั้นมีขอบเขตและข้อจำกัดในการดำเนินการดังนี้

1) มีปริมาณข้อมูลความคิดเห็นที่มีผลเฉลยที่ใช้ในการฝึกอบรมแบบลำลองที่น้อย เนื่องจากข้อมูลความคิดเห็นมีความหลากหลาย เนื่องจากทำการเก็บข้อมูลความคิดเห็นในช่วง 1 มกราคม 2564 ถึง วันที่ 30 เมษายน 2565 โดยมีข้อมูลจำนวน 703 รายการ

5.3 ข้อเสนอแนะ

จากผลการดำเนินการค้นคว้าอิสระครั้งนี้ สามารถนำการไปเพิ่มประสิทธิภาพแบบจำลองโครงข่ายประสาทเทียมแบบวนกลับเพื่อทำนายความมรู้สึกความคิดเห็นจากเว็บไซต์คลังคลิปวิดีโอเพื่อใช้ในการตัดสินใจในการลงทุนสกุลเงินดิจิทัลให้มีประสิทธิภาพมากขึ้น ดังนี้

1) เพิ่มจำนวนรายการข้อมูลความคิดเห็นเกี่ยวกับสกุลเงินดิจิทัลจากเว็บไซต์คลังคลิปวิดีโอที่มีผลเฉลให้มีปริมาณมากขึ้นในการฝึกอบรมแบบจำลอง

2) เพิ่มจำนวนผลเฉลยให้ครอบคุมมากขึ้น

3) ปรับจูนไฮเปอร์พารามิเตอร์ (Hyperparameter) แบบจำลองโครงข่ายประสาทเทียมแบบวนกลับเพื่อให้แบบจำลองมีประสิทธิภาพดียิ่งขึ้น

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

เอกสารอ้างอิง

(1) sciplanet, **“Cryptocurrency คืออะไร?”** [Online]. Available : https://sciplanet.org/content/8355

(2) bangkokbiznews, **“‘เงินดิจิทัล’ ควรลงทุนหรือไม่?”** [Online]. Available : https://www.bangkokbiznews.com/business/934903

(3) Sinart, T. 2017 **“Long Short-Term Memory (LSTM)”** [Online]. Available **:** https://medium.com/@sinart.t/long-short-term-memory-lstm-e6cb23b494c6

(4) tongkornkitt. 2017 **“[ML] LSTMs แบบ (เกือบ) ละเอียดยิบ ๆ”** [Online]. Available : https://medium.com/@tongkornkitt/ml-lstms

(5) Surapong, K. 2019 **“Recurrent Neural Network (RNN) คืออะไร Gated Recurrent Unit (GRU) คืออะไร สอนสร้าง RNN ถึง GRU ด้วยภาษา Python – NLP ep.9”** [Online]. Available : https://www.bualabs.com/archives/3103/what-is-rnn-recurrent-neural-network-what-is-gru-gated-recurrent-unit-teach-how-to-build-rnn-gru-with-python-nlp-ep-9/

(6) Pakawat, N. 2022 **“ทำความเข้าใจ BERT”** [Online]. Available : https://medium.com/@chameleontk

(7) Sasiwut, C. 2020 **“รู้จัก Principal Components Analysis (PCA) ในเชิง Practical”** [Online]. Available : https://lengyi.medium.com/principal-components-analysis-pca-e97c976ff130

(8) Kong Ruksiam. 20020 **“สรุป Machine Learning(EP.6)- การวิเคราะห์องค์ประกอบหลัก (PCA)”** [Online]. Available : **https://kongruksiam.medium.com/-machine-learning-ep-5pca**

(9) กาญจน์เขจร ชูชีพ. 2018 **“การถดถอยโลจิสติก”** [Online]. Available : https://forest-admin.forest.ku.ac.th%20Logistic%20Regression

(10) Nuttachot, P. 2020 “Sentiment Analysis 101” [Online]. Available : https://blog.pjjop.org/sentiment-analysis-101

(11) Wei-Meng, L. **“Using Principal Component Analysis (PCA) for Machine Learning”** [Online]. Available :

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

https://towardsdatascience.com/using-principal-component-analysis-pca-for-machine-learning-b6e803f5bf1e

(12) Emre Sasmaz and F. Boray Tek. 2021. **"Tweet Sentiment Analysis for Cryptocurrencies"** Department of Computer Science and Engineering, Isik University

(13) Mohammad J. Hamayel and Amani Yousef Owda. 2021. **"A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms"** Department of Natural, Engineering and Technology Sciences, Arab American University

(14) María Guinda and Ritabrata Bhattacharyya. 2021. **"Using Principal Component Analysis on Crypto Correlations to Build a Diversified Portfolio"** WorldQuant University

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้

ประวัติผู้เขียน

ชื่อ	นางสาวภานุมาศ ทุกข์จาก
วัน เดือน ปีเกิด	02 พฤษภาคม 2540
ที่อยู่ปัจจุบัน	131/2 ต.หัวไทร อ.บางคล้า จ.ฉะเชิงเทรา 24110
	โทร. 088-751-8592
ประวัติการศึกษา	(2561) วิทยาศาสตรบัณฑิต สาขาวิทยาการคอมพิวเตอร์ เกรดเฉลี่ย 3.20
	สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

เอกสารนี้เป็นเอกสารที่สงวนไว้สำหรับการใช้งานเพื่อการศึกษาเท่านั้น ไม่อนุญาตให้นำไปใช้ประโยชน์ด้านการค้า ไม่ว่ากรณีใดๆ ทั้งสิ้น อีกทั้งห้ามมิให้ดัดแปลงเนื้อหา และต้องอ้างอิงถึงเจ้าของเอกสารทุกครั้งที่มีการนำไปใช้