การวิเคราะห์ความรู้สึกสื่อสังคมเพื่อใช้ในการพยากรณ์ราคาบิทคอยน์

(Twitter Sentimental Analysis For Bitcoin's Return Prediction)

นาย	เริงฤทธิ์	แจ้งศรี	รหัสนักศึกษา	6102686455
นางสาว	มณฑกาญจน์	นาเพาะผล	รหัสนักศึกษา	6202685217

รายงานนี้เป็นส่วนหนึ่งของการเรียนรายวิชา EC200 วิทยาศาสตร์ข้อมูลสำหรับการวิเคราะห์เศรษฐกิจ

มหาวิทยาลัยธรรมศาสตร์ รังสิต

ภาคเรียนที่ 1 ปีการศึกษา 2564

คำนำ

รายงานนี้ฉบับนี้จัดทำขึ้นเพื่อเป็นส่วนหนึ่งของวิชา EC200 วิทยาศาสตร์ข้อมูลสำหรับการวิเคราะห์เศรษฐกิจ โดยมีจุดประสงค์เพื่อศึกษาและเรียนรู้เกี่ยวกับการวิเคราะห์ข้อมูลโดยใช้เครื่องมือที่ได้ศึกษามา

คณะผู้จัดทำได้เลือก การวิเคราะห์ความรู้สึกสื่อสังคมเพื่อใช้ในการพยากรณ์ราคาบิทคอยน์ (Twitter Sentimental Analysis For Bitcoin's Return Prediction) มาวิเคราะห์และทำแบบจำลองในรายงาน เนื่องจากคณะผู้จัดทำมองว่า การวิเคราะห์ความรู้สึกสื่อสังคมเพื่อใช้ในการพยากรณ์ราคาบิทคอยน์นั้นเป็นการวิเคราะห์ที่น่าสนใจและในขณะนี้เป็นเรื่องที่กำลังนิยม ทำให้คณะผู้จัดทำสามารถเข้าใจในรายละเอียดและสามารถหาข้อมูลเพิ่มเติมได้อย่างครบถ้วน

คณะผู้จัดทำจะต้องขอขอบคุณ อ.ดร.วศิน ศิวสฤษดิ์ ซึ่งเป็นผู้ให้ความรู้ในหลักการการทำแบบจำลองสำหรับการวิเคราะห์ต่างๆ ทางคณะผู้จัดทำหวังว่ารายงานฉบับนี้จะให้ความรู้ และเป็นประโยชน์แก่ผู้อ่านไม่มากก็น้อย

คณะผู้จัดทำ

1.บทนำ (Introduction)

Bitcoin คือ สกุลเงินดิจิทัลที่ทุกคนบนโลกสามารถใช้แลกเปลี่ยนกันได้อย่างอิสระ ถูกสร้างขึ้นมาด้วยภาษาคอมพิวเตอร์ ไม่มีใครเป็นเจ้าของ ไม่มีรูปร่าง และไม่สามารถจับต้องได้เหมือนธนบัตรหรือเหรียญเงินบาท Bitcoin ถูกสร้างขึ้นมาด้วยกลุ่มนักพัฒนาเล็กๆกลุ่มหนึ่งตลอดจนบริษัทใหญ่ๆทั่วโลก โดยระบบของ Bitcoin ถูกรันโดยคอมพิวเตอร์ของผู้ใช้งานทั่วโลก โดยใช้ระบบซอฟต์แวร์ในการถอดสมการคณิตศาสตร์ โดยความพิเศษของสกุลเงินนี้เป็นที่นิยม เนื่องจากมันถูกควบคุมแบบกระจาย (decentralize) กล่าวคือไม่มีสถาบันการเงินไหนสามารถควบคุม Bitcoin ได้ ซึ่งนั่นทำให้ผู้คนที่เลือกใช้ Bitcoin ส่วนใหญ่สบายใจ เนื่องจากแม้แต่ธนาคารก็ไม่สามารถควบคุม Bitcoin ได้ โดยผู้ที่สร้างและพัฒนา Bitcoin ใช้นามแฝงมีชื่อว่า “ซาโตชิ นาคาโมโตะ” มีจุดประสงค์ คือมีความต้องการสร้างสกุลเงินที่เป็นอิสระจากรัฐบาลและธนาคาร โดยสามารถส่งหากันผ่านระบบอินเทอร์เนตและมีค่าธรรมเนียมราคาถูก

ด้วยเหตุนี้เองทำให้ในปัจจุบันต่างมีผู้คนให้ความสนใจกับสกุลเงินดิจิทัลมากขึ้น ไม่เพียงแค่ Bitcoin เท่านั้น แต่ยังมีสกุลเงินดิจิทัลอื่นๆ ที่ได้รับความสนใจ แต่ทางคณะผู้จัดทำให้ความสนใจกับ Bitcoin เป็นพิเศษ เนื่องจากว่า เป็นสกุลเงินดิจิทัลที่เป็นที่รู้จักมากที่สุด และมีผู้คนให้ความสนใจมากกว่าสกุลเงินอื่นๆ ทำให้เป็นที่น่าสนใจว่าจะสามารถสร้างแบบจำลองที่ทำนายราคาของ Bitcoin โดยใช้ข้อมูลจากสื่อสังคมได้หรือไม่ และหากสามารถทำนายได้จริงในระดับความแม่นยำที่เหมาะสม จะสามารถเป็นประโยชน์แก่นักลงทุน, ธนาคาร, องค์กร และอื่นๆไม่มากก็น้อย

ในทศวรรษที่ผ่านมาจะเห็นได้ว่าอินเทอร์เนตมีการพัฒนาอย่างแพร่หลาย ทั้งในเรื่องของการแลกเปลี่ยนข้อมูลและการแลกเปลี่ยนประสบการณ์ กลายเป็นเรื่องง่ายในการพูดคุยผ่านสื่อสังคม ไม่ว่าจะเป็น Twitter, Instagram, Facebook, Blogs และอื่นๆ ตัวอย่างเช่น จะเห็นได้ว่าใน Twitter มีผู้ใช้จำนวนหลายแสนรายในการแลกเปลี่ยนข้อมูลหรือพูดคุยเรื่องที่เกี่ยวข้องกับ Bitcoin ผ่านตามแฮชแทคทุกวัน ซึ่งข้อมูลจำนวนมากนี้เป็นประโยชน์กับการหารูปแบบการขึ้นลงของราคา Bitcoin ผ่านการใช้เทคโนโลยี เช่น Natural Language Processing และอื่นๆ โดยในรายงานฉบับนี้ ทางคณะผู้จัดทำได้เลือกใช้ Decision Tree และ Naïve Bayes มาใช้ในการสร้างแบบจำลอง

2.แนวทางการวิเคราะห์ข้อมูล (Approach)

2.1 ข้อมูล

ทางคณะผู้จัดทำได้นำข้อมูลมาจากเว็บไซต์ [1]ที่ใช้สำหรับการวิเคราะห์ข้อมูลผ่านสื่อสังคมโดยใช้ Tweets ที่เกี่ยวข้องกับ Bitcoin ในช่วงวันที่ 5 กุมภาพันธ์ ค.ศ. 2021 จนถึง วันที่ 12 ธันวาคม ค.ศ. 2021 เป็นรายชั่วโมง รวมทั้งหมด 1,950,228 ข้อมูล โดยข้อมูลเป็นรูปแบบข้อมูลภาคตัดขวางหลายช่วงเวลา (Pooled Cross Sectional Data) ข้อมูลประกอบไปด้วย วันที่ผู้ใช้งาน Tweet ชื่อผู้ใช้งาน ตำแหน่งของผู้ใช้งาน คำอธิบายผู้ใช้งาน วันที่ผู้ใช้สมัคร จำนวนผู้ติดตาม จำนวนเพื่อนของผู้ใช้งาน จำนวนคนถูกใจ บัญชีผู้ใช้ที่ผ่านการตรวจสอบ และ เนื้อหาของ Tweet

date	user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	text
2021-02-05 11:00:00	Iconic Holding	Frankfurt am Main, Germany	Professional Crypto Asset Ventures \nhttps://t...	2021-01-05 13:22:24	301.0	1075	361	False	2️⃣ Debunking 9 #Bitcoin Myths by @Patrick_Lo...
2021-02-05 11:00:00	Iconic Holding	Frankfurt am Main, Germany	Professional Crypto Asset Ventures \nhttps://t...	2021-01-05 13:22:24	301.0	1075	361	False	📖 Weekend Read 📖 \n\nKeen to learn about #cryp...
2021-02-05 11:00:00	Iconic Holding	Frankfurt am Main, Germany	Professional Crypto Asset Ventures \nhttps://t...	2021-01-05 13:22:24	301.0	1075	361	False	4️⃣📡 Bloomberg LP #CryptoOutlook 2021 with @...
2021-02-05 11:00:00	Iconic Holding	Frankfurt am Main, Germany	Professional Crypto Asset Ventures \nhttps://t...	2021-01-05 13:22:24	301.0	1075	361	False	5️⃣#Blockchain 50 2021 by @DelRayMan, @Forbe...
2021-02-05 11:00:00	Nick Doevendans	Edam-Volendam, Nederland	Amateur historicus m.n. WW2, schrijver, muziek...	2020-06-12 16:50:07	37.0	123	410	False	#reddcoin #rdd @reddcoin to the moon #altcoin ...
...	...	...	...	...	...	...	...	...	...
2021-12-12 00:00:00	Saylor Bot	NaN	This bot will tweet @MicroStrategy and @michae...	2021-08-18 05:11:37	44.0	3.0	5.0	False	Michael Saylor's Bitcoin Average: ~$29534\n\nP...
2021-12-12 00:00:00	STUDIO192.NL	Apeldoorn (GLD) Holland	On-line Radio Station. In Apeldoorn,Holland	2010-01-15 15:57:04	1114.0	622.0	233.0	False	[950] #Glasgow If You Join In At #CryptoTab no...
2021-12-12 00:00:00	Crypto Pricing	NaN	Given cryptocurrency's current price in USD. P...	2021-03-10 13:17:40	15.0	1.0	421.0	False	Crypto Prices (USD/B)\n\nBitcoin ➡ $49361.3809...
2021-12-12 00:00:00	Mister Crypto	Netherlands	My goal is to educate the world about #crypto ...	2021-10-04 11:37:09	1986.0	16.0	14526.0	False	#Bitcoin retested the previous support as res...

รูปภาพที่ 1 : ตัวอย่างข้อมูลภาคตัดขวางหลายช่วงเวลา (Pooled Cross Sectional Data)

[1] https://www.kaggle.com/kaushiksuresh147/bitcoin-tweets

อีกทั้งได้นำเข้าข้อมูลราคาของ Bitcoin ผ่านไลบรารี่ (**Library**) ชื่อ **“ccxt”** โดยเป็นไลบรารี่ที่ใช้สำหรับเชื่อมต่อและซื้อขายกับตลาดสกุลเงินดิจิทัล **ccxt** ให้การเข้าถึงข้อมูลตลาดเพื่อ การจัดเก็บ วิเคราะห์ และแสดงผลข้อมูล

2.2 การเตรียมข้อมูล

2.2.1 ทางคณะผู้จัดทำได้เห็นว่าวันที่ในชุดข้อมูลไม่ได้เป็นประเภท datetime64 ซึ่งเป็นประเภทข้อมูลของไลบรารี่ NumPy ที่รองรับการทำงานรูปของวันที่ จึงทำการเปลี่ยนวันที่ให้เป็นประเภท datetime64

หลังจากนั้นทำการปัดข้อมูลวันที่ให้เป็นรายชั่วโมง ตัวอย่างเช่น 02/05/2021 12:15:45 เป็น 02/05/2021 12:00:00 เพื่อลดความซับซ้อนและขนาดของข้อมูล

2.2.2 เนื่องจากชุดข้อมูลมีขนาดใหญ่จึงต้องทำการลดขนาดข้อมูล โดยการนำ Tweet ที่เกิดขึ้นในวันเวลาเดียวกันรวมเป็นข้อมูลเดียว

date	text
2021-02-05 11:00:00	Prices update in $USDT(1hour)$: \n\nBTC - 3...
2021-02-05 11:00:00	The believe is still strong, anything can happ...
2021-02-05 11:00:00	WARNING! POWEREARN - OUTSIDE PROJECT - FAST SC...
2021-02-05 11:00:00	PENN ASIA IS DEVELOPING A DIGITAL ASSET CALLED...
2021-02-05 11:00:00	$BTC SELLING PRESSURE ALERT Price tradi...
...	...
2021-12-12 00:00:00	IOG Has Already Started Process of Making Card...
2021-12-12 00:00:00	Buy #Bitcoin and take profit at 100k$
2021-12-12 00:00:00	@airdropinspect Awesome project, this project ...
2021-12-12 00:00:00	B1 = 49, 470(23 : 33UTC)\nBTC prices conti...
2021-12-12 00:00:00	$BTC #Bitcoin Bull case, 3 peaks and Domed hou...

⇩

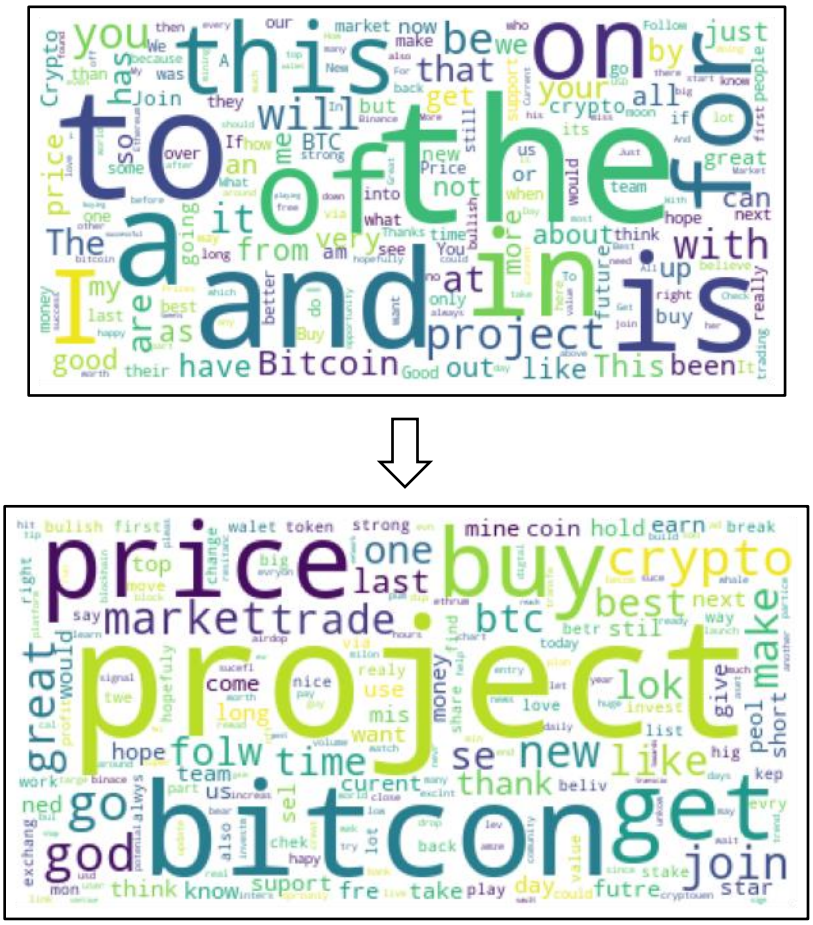
	text
2021-02-05 11:00:00	Prices update in $USDT(1hour)$: \n\nBTC - ...
2021-02-05 12:00:00	Stablecoin flow into exchanges #BTC #onchain...
2021-02-05 13:00:00	@CoinsioCom Got enough Coins? Earn, Trade, Se...
2021-02-05 14:00:00	@_Chuks1 @dmitriidavs Right here w/ @binance!...
2021-02-05 15:00:00	1750$ \n\n#Ethereum \n\n https://t.co/briBX...
...	...
2021-12-11 20:00:00	@binance Buy feg now\n\nMaybe tomorrow will b...
2021-12-11 21:00:00	#Saitama is on fire!!!\nLFG !!! \n@SaitaM...
2021-12-11 22:00:00	Dapper Labs Is Bringing Its NFTs to the NFL: ...
2021-12-11 23:00:00	MACD 8H fast line crossed above the slow line...
2021-12-12 00:00:00	#bitcoin Structure and Anxiety - More Extreme...

รูปภาพที่ 2 : ตัวอย่างการลดขนาดของข้อมูล

2.2.3 ต่อไปจะเป็นกระบวนการ Text Preprocessing เนื่องจากข้อมูลประกอบไปด้วย ตัวอักษร อิโมติคอน (Emoticon) URL และสัญลักษณ์ต่างๆ จึงต้องมีการทำความสะอาดก่อนนำไปใช้วิเคราะห์ โดยเริ่มจากการถอนคำ (Tokenization) ตัวอย่างเช่น “I love Python” เป็น [‘I’, ‘love’, ‘Python’]

จากนั้นทำการลบคำที่ไม่เกี่ยวข้องกับการวิเคราะห์ โดยจะทำการตรวจสอบว่าเป็นตัวอักษร A-Z หรือไม่ และทำการลบคำที่มีรูปแบบผิดปกติ โดยทำการตรวจสอบ Ascii Code หากไม่อยู่ระหว่าง 65-122 จะทำการลบคำนั้นออก

จากนั้นทำการลดรูปคำศัพท์ แปลงคำให้อยู่ในรูปพื้นฐาน โดยใช้วิธี Lemmatization ตัวอย่างเช่น [“running”, “ran”] เป็น [“run”, “run”] โดยในที่นี้ทางคณะผู้จัดทำได้ใช้ฟังก์ชัน WordNetLemmatizer จากไลบรารี่ nltk หลังจากทำการลดรูปคำเรียบร้อย จะทำการลบคำที่พบบ่อยแต่ไม่ได้สื่อความหมาย (Stop Words) โดยใช้ฟังก์ชัน stopwords จากไลบรารี่ nltk



รูปภาพที่ 3 : ตัวอย่างคำศัพท์จากการถูกลดรูปคำ

2.2.4 ต่อมาเป็นขั้นตอน Text Vectorizer คือขั้นตอนการถอนคำแต่ละคำ (Token) ซึ่งเป็นข้อมูลแบบข้อความแปลงไปเป็นข้อมูลเชิงตัวเลข (Numerical) โดยการนับจำนวนคำที่เกิดขึ้น (Counter) การคำนวณสัดส่วนความถี่ที่พบในเอกสาร (TF) หรือ การคำนวณการผกผันของสัดส่วนความถี่ที่พบในเอกสาร (IDF) ซึ่งทางคณะผู้จัดทำได้ใช้ TF-IDF ซึ่งคือ TF คูณด้วย IDF จากนั้นจัดเรียงข้อมูลให้อยู่ในรูปเมทริกซ์

2.2.5 จากนั้นทำการแบ่ง Class ของชุดข้อมูลราคา Bitcoin โดยเริ่มจากการคำนวณอัตราผลตอบแทนของ Bitcoin รายชั่วโมง จากนั้นจำแนกข้อมูลออกมาเป็น 2 ประเภท 1. อัตราผลตอบแทนเป็นบวก 2. อัตราผลตอบแทนเป็นลบ โดยหากเป็นประเภทแรกจะแทนด้วยค่า 1 นอกจากนั้นแทนด้วยค่า 0

2.2.6 จากนั้นนำชุดข้อมูลที่ได้ไปประกอบกับชุดข้อมูลประเภทของอัตราผลตอบแทนของ Bitcoin และชุดข้อมูลอื่นๆ เช่น จำนวนผู้ติดตาม จำนวนคนถูกใจ เพื่อนำไปใช้ในการพัฒนาแบบจำลอง

3.วิธีการวิจัย (Experimental Methodology)

3.1 แบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ (Decision Tree)

หลังจากที่จัดเรียงชุดข้อมูลเรียบร้อยดังรูปที่ 4 ทำการแบ่งข้อมูลออกเป็น feature และ class โดยจะกำหนดให้ TF-IDF ของคำแต่ละคำ จำนวนเฉลี่ยของผู้ติดตามในวันที่ Tweet และ จำนวนเฉลี่ยของจำนวนคนถูกใจในวันที่ Tweet เป็น feature และ ข้อมูลที่ได้จาก 2.2.5 เป็น class โดยจะแบ่งเป็น 2 แบบจำลองคือ แบบจำลองแรกเป็นแบบจำลองที่รวม feature จำนวนเฉลี่ยของผู้ติดตามในวันที่ Tweet จำนวนเฉลี่ยของจำนวนคนถูกใจในวันที่ Tweet และ Class ของอัตราผลตอบแทนวัน ณ ปัจจุบันแบบจำลองที่สอง จะไม่รวม feature ดังกล่าว

	ab	abc	abcd	abcde	abcish	abck	abcot	abcs	abcus	abcxyz	...	zymergn	zymurgenc	zynga	zyskind	zytar	zytewq
2021-02-05 12:00:00	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-02-05 13:00:00	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-02-05 14:00:00	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-02-05 15:00:00	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-02-05 16:00:00	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2021-12-11 20:00:00	0.000000	0.021306	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-12-11 21:00:00	0.000000	0.004167	0.007594	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-12-11 22:00:00	0.004863	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-12-11 23:00:00	0.000000	0.007204	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0
2021-12-12 00:00:00	0.000000	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0

รูปภาพที่ 4 : ตัวอย่างชุดข้อมูลที่ถูกจัดเรียง

จากนั้นทำการแบ่งข้อมูลออกเป็น 80% เพื่อใช้ในการสอนแบบจำลองให้มีความสามารถในการจำแนก และ 20% ใช้สำหรับทดสอบแบบจำลองว่ามีประสิทธิภาพ โดยการประเมินแบบจำลองจะทำการวัดผลจาก Confusion Matrix และค่า Precision Recall f-1 score

3.2 Naïve Bayes

ในการทดลองครั้งนี้จะใช้แบบจำลอง Naïve Bayes อยู่ 2 ประเภท คือ Multinomial Naïve Bayes และ Gaussian Naïve Bayes โดยเหตุผลที่ไม่ใช่ Bernoulli เนื่องจากตัวแบบจำลอง Bernoulli Naïve Bayes นั้นจะสนใจเพียงแค่การมีอยู่ของข้อมูล โดยไม่สนความถี่ของข้อมูล และเนื่องจากการจัดการข้อมูลของทางคณะผู้จัดทำได้ใช้การนับคำแบบ TF-IDF ดังนั้นการใช้ Bernoulli หรือ Multinomial จึงน่าจะไม่ให้ผลลัพธ์ต่างกัน ในส่วนของแบบจำลอง Gaussian Naïve Bayes นั้น จะต่างกับ Multinomial ตรงที่ Probabilistic Distribution ของ Likelihood ของ Features โดย Multinomial นั้นจะเหมาะกับ Features ที่มีลักษณะไม่ต่อเนื่อง (Discrete) ในขณะที่ Gaussian นั้นจะเหมาะกับ Features ที่มีลักษณะต่อเนื่อง (Continuous) โดยการทดสอบจะใช้เกณฑ์ในการประเมินเดียวกับแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้

4. ผลการศึกษา (Results)

4.1 แบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ (Decision Tree)

ในส่วนของแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่รวม Quantitative Features ซึ่งคือ จำนวนเฉลี่ยของผู้ติดตามในวันที่ Tweet จำนวนเฉลี่ยของจำนวนคนถูกใจในวันที่ Tweet และ Class ของอัตราผลตอบแทนวัน ณ ปัจจุบัน ผลจากรูปที่ 5 ทำให้เห็นได้ว่า Precision ของ Class 0 นั้นเท่ากับ 51% และในส่วนของ precision ของ Class 1 นั้นเท่ากับ 54% เช่นกัน Recall ของ Class 0 เท่ากับ 48% และ Recall ของ Class 1 เท่ากับ 56% และตัวแบบจำลองมี f1-score เท่ากับ 50% สำหรับ Class 0 และ 55% สำหรับ Class 1 อีกทั้งยังมี Accuracy rate เท่ากับ 52% ดังรูปที่ 5

	precision	recall	f1-score	support
0.0	0.51	0.48	0.50	198
1.0	0.54	0.56	0.55	210
accuracy			0.52	408
macro avg	0.52	0.52	0.52	408
weighted avg	0.52	0.52	0.52	408

รูปภาพที่ 5 : ตาราง Classification Report ของแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่รวม Quantitative Features

ในส่วนของแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่ไม่รวม Quantitative Features ผลจากรูปที่ 6 ทำให้เห็นได้ว่า Precision ของ Class 0 นั้นเท่ากับ 54% และในส่วนของ precision ของ Class 1 นั้นเท่ากับ 56% เช่นกัน Recall ของ Class 0 เท่ากับ 52% และ Recall ของ Class 1 เท่ากับ 58% และตัวแบบจำลองมี f1-score เท่ากับ 53% สำหรับ Class 0 และ 57% สำหรับ Class 1 อีกทั้งยังมี Accuracy rate เท่ากับ 55% ดังรูปที่ 6

	precision	recall	f1-score	support
0.0	0.54	0.52	0.53	198
1.0	0.56	0.58	0.57	210
accuracy			0.55	408
macro avg	0.55	0.55	0.55	408
weighted avg	0.55	0.55	0.55	408

รูปภาพที่ 6 : ตาราง Classification Report ของแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่ไม่รวม Quantitative Features

4.2 Multinomial Naïve Bayes

ในส่วนของแบบจำลอง Multinomial Naïve Bayes ที่รวม Quantitative Features ผลจากรูปที่ 7 ทำให้เห็นได้ว่า Precision ของ Class 0 นั้นเท่ากับ 48% และในส่วนของ precision ของ Class 1 นั้นเท่ากับ 50% เช่นกัน Recall ของ Class 0 เท่ากับ 60% และ Recall ของ Class 1 เท่ากับ 38% และตัวแบบจำลองมี f1-score เท่ากับ 53% สำหรับ Class 0 และ 43% สำหรับ Class 1 อีกทั้งยังมี Accuracy rate เท่ากับ 49% ดังรูปที่ 7

	precision	recall	f1-score	support
0.0	0.48	0.60	0.53	198
1.0	0.50	0.38	0.43	210
accuracy			0.49	408
macro avg	0.49	0.49	0.48	408
weighted avg	0.49	0.49	0.48	408

รูปภาพที่ 7 : ตาราง Classification Report ของแบบจำลอง Multinomial Naïve Bayes ที่รวม Quantitative Features

ในส่วนของแบบจำลอง Multinomial Naïve Bayes ที่ไม่รวม Quantitative Features ผลจากรูปที่ 8 ทำให้เห็นได้ว่า Precision ของ Class 0 นั้นเท่ากับ 48% และในส่วนของ precision ของ Class 1 นั้นเท่ากับ 44% เช่นกัน Recall ของ Class 0 เท่ากับ 88% และ Recall ของ Class 1 เท่ากับ 9% และตัวแบบจำลองมี f1-score เท่ากับ 62% สำหรับ Class 0 และ 14% สำหรับ Class 1 อีกทั้งยังมี Accuracy rate เท่ากับ 47% ดังรูปที่ 8

	precision	recall	f1-score	support
0.0	0.48	0.88	0.62	198
1.0	0.44	0.09	0.14	210
accuracy			0.47	408
macro avg	0.46	0.48	0.38	408
weighted avg	0.46	0.47	0.37	408

รูปภาพที่ 8 : ตาราง Classification Report ของแบบจำลอง Multinomial Naïve Bayes ที่ไม่รวม Quantitative Features

4.3 Gaussian Naïve Bayes

ในส่วนของแบบจำลอง Gaussian Naïve Bayes ที่รวม Quantitative Feature ผลจากรูปที่ 9 ทำให้เห็นได้ว่า Precision ของ Class 0 นั้นเท่ากับ 48% และในส่วนของ precision ของ Class 1 นั้นเท่ากับ

38% เช่นกัน Recall ของ Class 0 เท่ากับ 97% และ Recall ของ Class 1 เท่ากับ 1% และตัวแบบจำลองมี f1-score เท่ากับ 65% สำหรับ Class 0 และ 3% สำหรับ Class 1 อีกทั้งยังมี Accuracy rate เท่ากับ 48% ดังรูปที่ 9

	precision	recall	f1-score	support
0.0	0.48	0.97	0.65	198
1.0	0.38	0.01	0.03	210
accuracy			0.48	408
macro avg	0.43	0.49	0.34	408
weighted avg	0.43	0.48	0.33	408

รูปภาพที่ 9 : ตาราง Classification Report ของแบบจำลอง Gaussian Naïve Bayes ที่รวม Quantitative Features

ในส่วนของแบบจำลอง Gaussian Naïve Bayes ที่ไม่รวม Quantitative Feature ผลจากรูปที่ 10 ทำให้เห็นได้ว่า Precision ของ Class 0 นั้นเท่ากับ 46% และในส่วนของ precision ของ Class 1 นั้นเท่ากับ 48% เช่นกัน Recall ของ Class 0 เท่ากับ 57% และ Recall ของ Class 1 เท่ากับ 38% และตัวแบบจำลองมี f1-score เท่ากับ 51% สำหรับ Class 0 และ 42% สำหรับ Class 1 อีกทั้งยังมี Accuracy rate เท่ากับ 47% ดังรูปที่ 10

	precision	recall	f1-score	support
0.0	0.46	0.57	0.51	198
1.0	0.48	0.38	0.42	210
accuracy			0.47	408
macro avg	0.47	0.47	0.47	408
weighted avg	0.47	0.47	0.47	408

รูปภาพที่ 10 : ตาราง Classification Report ของแบบจำลอง Gaussian Naïve Bayes ที่ไม่รวม Quantitative Features

5. การอภิปรายผล (Discussion)

ในการวิเคราะห์ผลการทดสอบ ทางคณะผู้จัดทำจะยึดหลัก “ตกรถดีกว่าติดดอย” ดังนั้น ควรเลือกแบบจำลองที่มี High Precision, Low Recall ของ Class 1 เนื่องจากว่า ในการตัดสินใจที่จะซื้อ Bitcoin หรือไม่ การที่จะใช้ตัวแบบจำลองในการประกอบการตัดสินใจนั้นย่อมกังวลว่า ตัวแบบจำลองจะพยากรณ์ว่าอัตราผลตอบแทนจะเป็นบวกผิด เพราะแท้จริงแล้วอัตราผลตอบแทนจะเป็นลบ ดังนั้น High Precision หมายความว่า ตัวแบบจำลองพยากรณ์ว่า อัตราผลตอบแทนจะเป็นบวกและข้อมูลที่พยากรณ์เป็นบวกจริงๆ หรืออีกนัยหนึ่ง แบบจำลองมีการพยากรณ์ False Positive น้อย ส่วน Low Recall หมายความว่า จากข้อมูลอัตราผลตอบแทนบวกทั้งหมด แบบจำลองพยากรณ์ว่าเป็นบวกได้ถูกต้องน้อย หรืออีกนัยหนึ่งหมายความว่า มี False Negative สูง หรือ คือการตกรถนั่นเอง

ดังนั้น เกณฑ์ที่จะใช้ตัดสินใจในการเลือกแบบจำลอง จะตัดสินใจจาก Precision ที่สูงของ Class 1 และตามด้วย Recall ที่สูงของ Class 0 เพราะฉะนั้นหากใช้เกณฑ์นี้แบบจำลองที่เหมาะสมจะเรียงได้ ดังตารางที่ 1

อันดับ	แบบจำลอง
1.	แบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่ไม่รวม Quantitative Features
2.	แบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่รวม Quantitative Features
3.	แบบจำลอง Multinomial Naïve Bayes ที่รวม Quantitative Features
4.	แบบจำลอง Gaussian Naïve Bayes ที่ไม่รวม Quantitative Features
5.	แบบจำลอง Multinomial Naïve Bayes ที่ไม่รวม Quantitative Features
6.	แบบจำลอง Gaussian Naïve Bayes ที่รวม Quantitative Features

ตารางที่ 1 : แสดงอันดับที่มี Precision สูงของ lass 1 และตามด้วย Recall ที่สูงของ Class 0

แต่หากพิจารณาจากอัตราผลตอบแทน (Geometric mean) จากการซื้อขาย Bitcoin ตามแบบจำลอง โดยหากแบบจำลองพยากรณ์ว่าเป็น Class 1 จะทำการซื้อในวันนั้น แต่ถ้าเป็น Class 0 จะไม่ทำอะไร โดยจะได้ผลลัพธ์ตามตารางที่ 2 จะเห็นได้ว่าแบบจำลองที่พยากรณ์ได้ถูกต้องมาก อย่างอันดับที่ 1 ในตารางที่ 1 กลับอยู่ในอันดับที่ 2 ในตารางที่ 2 อาจเป็นเพราะว่าแบบจำลองนั้นสามารถพยากรณ์ถูกแต่วันที่มีอัตราผลตอบแทนน้อย

อันดับ	แบบจำลอง	อัตราผลตอบแทน	อัตราผลตอบแทนสูงกว่าปกติ
1.	แบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่รวม Quantitative Features	87.68%	ใช่
2.	แบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ที่ไม่รวม Quantitative Features	70.25%	ใช่
3.	แบบจำลอง Gaussian Naïve Bayes ที่ไม่รวม Quantitative Features	36.50%	ใช่
4.	แบบจำลอง Multinomial Naïve Bayes ที่รวม Quantitative Features	2.45%	ไม่
5.	แบบจำลอง Multinomial Naïve Bayes ที่ไม่รวม Quantitative Features	-1 .61%	ไม่
6.	แบบจำลอง Gaussian Naïve Bayes ที่รวม Quantitative Features	-2.56%	ไม่

ตารางที่ 2 : แสดงอันดับที่พิจารณาจากอัตราผลตอบแทน (Geometric mean) จากการซื้อขาย Bitcoin

6. สรุปผลการวิจัย (Conclusion)

จากการทำรายงานฉบับนี้ ทางคณะผู้จัดทำได้มีการศึกษาหาความรู้เพิ่มเติมไม่ว่าจะเป็นในเรื่องของการทำกระบวนการ Text Preprocessing ที่นำมาใช้ในเรื่องของการถอนคำ (Tokenization) หรือ การประมวลผลข้อมูลที่มีขนาดใหญ่ เพื่อที่จะนำเอาข้อมูลเหล่านั้นไปวิเคราะห์ โดยหลังจากที่ทำการวิเคราะห์ผลลัพธ์พบว่า กลุ่มแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ (Decision Tree) สามารถสร้างอัตราผลตอบแทนได้มากกว่า ทั้งนี้ในการเปรียบเทียบแบบจำลองการตัดสินใจแบบโครงสร้างต้นไม้ (Decision Tree) นั้น ทางคณะผู้จัดทำเห็นว่าควรที่จะใช้วิธี Random Forest เข้ามาช่วยในการตัดสินใจ