Sentiment Analysis on Tweets to Model Cryptocurrency Volatility

MLPR Project

Team 32:-
Bhavya Pathak
Bisol Mathai
Tanu Adhikari

Introduction To Cryptocurrency

What is Cryptocurrency?

- A digital currency that operates without a central authority.
- They function on blockchain technology.
- Example: Bitcoin – the first and most popular cryptocurrency.

Bitcoin's Significance in Finance

- A cryptocurrency created in 2009.
- Limited supply (21 million coins).
- Traded globally – highly volatile

Volatility

- Volatility refers to how quickly and dramatically the price of an asset can change.
- High volatility means prices can swing wildly in short periods.
- Bitcoin is famous for its high volatility.
- Example: The price reacts to major news headlines and viral social media trends. These factor cause large price change.

Influence of Social Media on Market Dynamics

- **Social media plays a critical role in shaping public sentiment about Bitcoin.**
- **The collective 'mood' of billions of people influence investor behavior and market direction.**
- **Research indicates that positive or negative posts can significantly impact market behavior, altering price trajectories almost in real time.**

How Elon Musk's Twitter activity moves cryptocurrency markets

Twitter activity affects short-term cryptocurrency returns and volume. In other words, we investigate whether cryptocurrency markets exhibit a "Musk Effect".

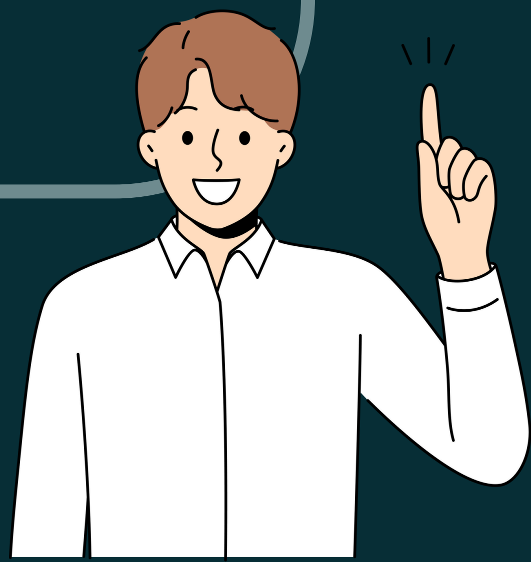
Considered in isolation, non-negative tweets from Musk lead to significantly positive abnormal Bitcoin returns. Individual tweets do raise the price of Bitcoin by 16.9% or reduce it by almost 11.8%. Our study shows the significant impact that the social media activity of influential individuals can have on cryptocurrencies. This suggests a conflict between morals, risks of market manipulation and investor protection.

Twitter, a well-known social networking site, offers users a service that allows anyone to send and receive brief text messages. Twitter allows users to view one another's posts even if they are strangers [6]. Investors commonly express their feelings on Twitter, making it a rich source of emotional intelligence and providing live updates on crypto-currency information [7]. An analysis by the American Institute of Economic Research found that news from around the world can cause significant variations in the price of BTC. Investigating how people feel about BTC via tweets is beneficial [8]. Data sentiments can be identified using deep learning (DL) technology [9].

Nair, M., Abd-Elmegid, L. A., & Marie, M. I. (2024). Sentiment analysis model for cryptocurrency tweets using different deep learning techniques. Journal of Intelligent Systems, 33(1), 20230085.https://doi.org/10.1515/jisys-2023-0085

Problem Statement

- Bitcoin is a highly volatile asset, creating a need for price prediction tools to aid investment decisions.
- Social media sentiment (e.g., on Twitter) is used as a promising indicator for predicting this volatility.
- But the validity of social media sentiment is compromised by automated bot accounts that manipulate the signal.
- Existing models often use raw, unverified tweet data, leading to noisy and unreliable predictions.
- **Our Project Goal: To achieve accurate Bitcoin volatility prediction by filtering out bot tweets and analyzing genuine human sentiment.**



Sentiment analysis

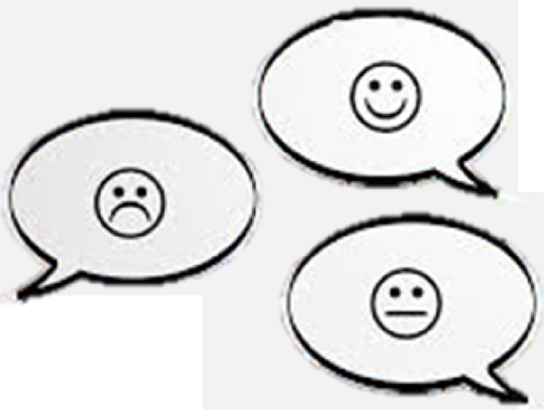
Sentiment analysis is the process of analyzing digital text to determine if the emotional tone of the message is positive, negative, or neutral.

Positive tweet

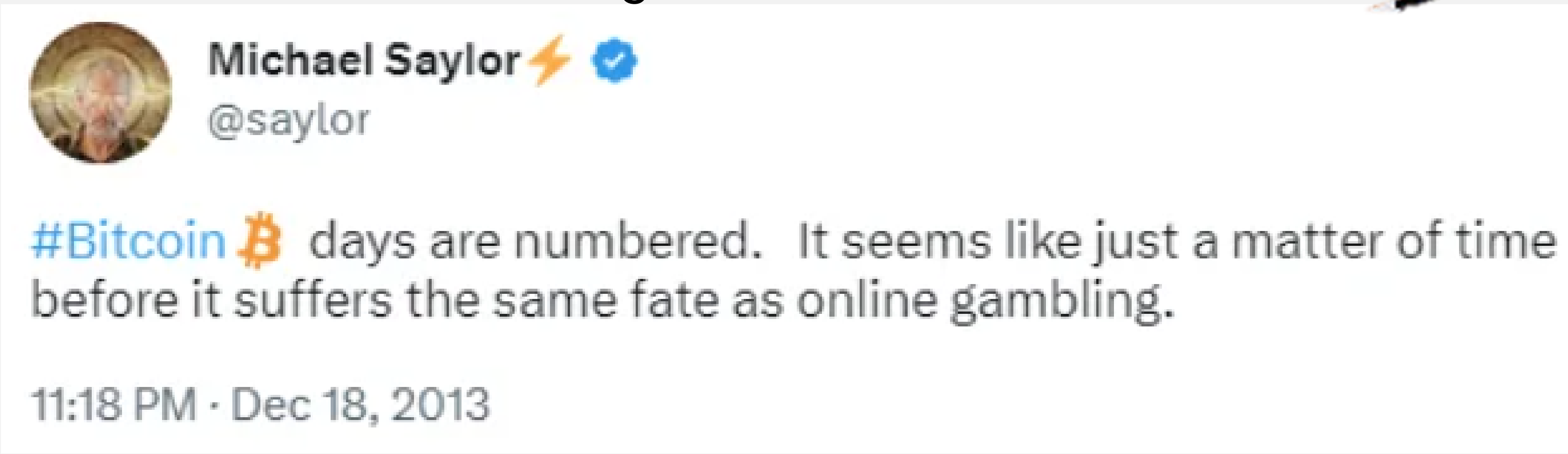

darknes23
@darknes23

Just sold 100 #bitcoins for $220 cash. I tripled my investment. Buyer was happy: 100 BTC are worth $260 right now. #BTC ₿ economy is real.

5:42 pm · 29 Apr 11 · Twitter Web Client



Negative tweet


Michael Saylor⚡
@saylor

#Bitcoin ₿ days are numbered. It seems like just a matter of time before it suffers the same fate as online gambling.

11:18 PM · Dec 18, 2013

Neutral tweet


christian walker
@javashaman

Woot! Finally generated a bitcoin! Actually... 100 bitcoins! Am I rich yet??

12:41 AM · Aug 12, 2010 · Twitter Web Client

- Social media sentiment is commonly used to predict crypto price movements.
- Automated bot accounts distort sentiment by inflating engagement.
- High bot activity reduces the predictive power of tweet-based models.
- Raw, unfiltered tweet data leads to misleading forecasts.
- Filtering out artificial activity is crucial for accurate sentiment analysis.

Qureshi, K., & Zaman, T. (2023). Social media engagement and cryptocurrency performance. Plos one, 18(5), e0284501.

Literature review

1 Bot detection Models

This study uses Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes, K-Nearest Neighbors, Kernel SVM, Decision Trees, and Random Forest—for Twitter bot detection.

These models rely on a combination of account metadata features (such as follower count, verified status, description keywords) to distinguish between bot and human accounts.

- Random Forest achieved the highest accuracy of 85.25%. However, Logistic Regression demonstrated the highest true positive rate (94.14%), showing strong capability in correctly identifying bot accounts.

.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 4, April 2021
DOI 10.17148/IJARCCE.2021.10417

Twitter Bot Detection

Jison M Johnson[1], Prince John Thekkadayil[2], Mansi D Madne[3], Atharv Nilesh Shinde[4], N Padmashri[5]
Student, Computer Department, Fr. Conceicao Rodrigues Institute of Technology, Navi Mumbai, India[1-4]
Asst. Professor, Computer Department, Fr. Conceicao Rodrigues Institute of Technology, Navi Mumbai, India[5]

Abstract: Today, social media platforms are being utilized by a gazillion of people which covers a vast variety of media. Among this, around 192 million active accounts are stated as Twitter users. This discovered an increasing number of bot accounts are problematic that spread misinformation and humor, and also promote unverified information which can adversely affect various issues. So in this paper, we will detect bots on Twitter using machine learning techniques. A web application where we can verify if the account is a bot or a genuine account. We analyze the dataset extracted from the Twitter API which consists of both human and bot accounts. We analyze the important features like tweets, likes, retweets, etc., which are required to provide us with good results. We use this data to train our model using machine learning methods Decision Trees and Random Forest. For linking our model with the web content, we used the flask server. Our result on our framework indicates that the user belongs to a human account or a bot with reasonable accuracy.

V. RESULTS

The model required for this project aims to predict whether a given Twitter account is a genuine user or a bot and also to have the best accuracy. For this reason, we implemented various types of classification algorithms and selected the one with the maximum accuracy score. The below table shows the algorithms and their accuracy scores.

Algorithms	Accuracy Rate	Misclassification Rate	True Positive Rate
Logistic Regression	71.85%	28.15%	94.14%
KNN Classifier	83.71%	16.29%	85.67%
SVM Classifier	67.28%	32.72%	93.50%
Kernel SVM	69.28%	30.72%	91.11%
Naive Bayes	62.14%	37.86%	96.22%
Random Forest	85.25%	14.75%	83.50%

But the model did not incorporate tweet content and it was limited to static metadata features.
We plan to Include text-based features such as tweet content, hashtags, and sentiment scores and integrating bot detection as a preprocessing step before sentiment classification, ensuring more accurate analysis by filtering out the tweets generated by bots bots.

Aslam, N., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2022). Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model. https://doi.org/10.1109/ACCESS.2022.3165621

2 Sentiment-Based Models

This study uses Support Vector Machines (SVM), Logistic Regression (LR), Naive Bayes, and Random Forest—for sentiment classification
. These models rely on feature extracting techniques like Bag-of-Words (BoW) , TF-IDF and word2vec to convert tweets into numerical representations

- BoW consistently outperformed TF-IDF and Word2Vec in accuracy across all tested models.
- SVM and LR emerged as the most accurate classifiers, achieving 98% accuracy for sentiment classification using BoW features.

Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model

NAILA ASLAM[1,2], FURQAN RUSTAM[3], ERNESTO LEE[4], PATRICK BERNARD WASHINGTON[5], AND IMRAN ASHRAF[6]
[1]School of Electronics and Information Engineering, Hebei University of Technology, Tianjin 300401, China
[2]Department of Computer Science, School of Systems and Technology, University of Management and Technology, Lahore 54000, Pakistan
[3]Department of Software Engineering, School of Systems and Technology, University of Management and Technology, Lahore 54770, Pakistan
[4]Department of Computer Science, Broward College, Broward County, Fort Lauderdale, FL 33301, USA
[5]Division of Business Administration and Economics, Morehouse College, Atlanta, GA 30314, USA
[6]Department of Information and Communication Engineering, Yeungnam University, Gyeongsan-si 38544, South Korea
Corresponding authors: Imran Ashraf (ashrafimran@live.com) and Ernesto Lee (elee@broward.edu)

This work was supported by the Florida Center for Advanced Analytics and Data Science funded by Ernesto.Net (under the Algorithms for Good Grant).

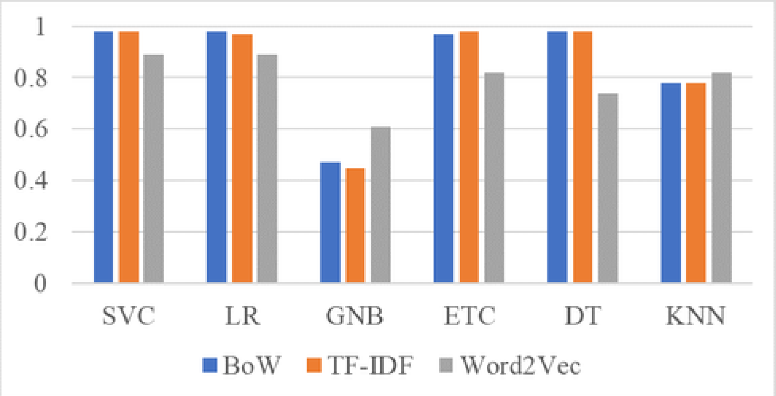


TABLE 11. Sentiment analysis results using BoW features.

Model	Accuracy	Precision	Recall	F1 score	G mean
SVM	0.98	0.98	0.97	0.98	0.97
LR	0.98	0.98	0.96	0.97	0.97
GNB	0.47	0.55	0.57	0.45	0.56
ETC	0.97	0.97	0.93	0.95	0.95
DT	0.98	0.97	0.97	0.97	0.97
KNN	0.78	0.86	0.67	0.72	0.76

The referenced study did not account for the presence of bot-generated tweets, which can distort sentiment analysis results by artificially inflating sentiment polarity

We plan to integrate bot detection techniques to filter out bot generated tweets to ensure more reliable and human-centric sentiments.

Aslam, N., Rustam, F., Lee, E., Washington, P. B., & Ashraf, I. (2022). Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model. https://doi.org/10.1109/ACCESS.2022.3165621

3 Market value prediction model

This study employs Long Short-Term Memory (LSTM) and a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model for Bitcoin price volatility prediction.

- LSTM and CNN-LSTM models consistently outperformed traditional models, with CNN-LSTM improving short-term (7-day) forecast performance by up to 9.77% compared to HAR.

Contents lists available at ScienceDirect
Journal of International Financial Markets, Institutions & Money
journal homepage: www.elsevier.com/locate/intfin

Forecasting Bitcoin volatility using machine learning techniques

Zih-Chun Huang[a], Ivan Sangiorgi[a], Andrew Urquhart[b,*]

5.2.1. Comparisons with benchmark models

Table 6 presents the relative forecasting power and RMSE of the hybrid CNN-LSTM model compared to HAR model.[23] There are three main results: first of all, it is noticeable that by employing 10-min frequency data instead of using standard inputs in the HAR model, 1-day, 7-day average, and 30-day average data, or feeding log returns as in GARCH-type models, the CNN-LSTM models rank at the top and outperform LSTM in 7-day to 60-day ahead prediction. Therefore, the CNN-LSTM model that uses high-frequency data outperforms models that use daily data. Furthermore, our model is a better choice for short-term volatility forecasting, especially in 7-day ahead forecasting. The best performance for the 7-day ahead prediction generated by the hybrid CNN-LSTM model is 9.77% higher than the HAR model. However, the HAR model with daily data outperforms the neural network model with high-frequency data for 1-day ahead prediction.

Second, compared to LSTM, the hybrid CNN-LSTM exploits the benefits of image classification on MTF images encoded by the transition probability information. The result shows that the prediction accuracy of the hybrid CNN-LSTM model along with image

that while HAR generally performs better in 1-day forecasts, the CNN-LSTM model outperforms in 7-day forecasts. We can observe that in 1-day ahead forecasting, the HAR model is more effective at handling Bitcoin's volatility spikes, likely originating from its design, which incorporates 7-day and 30-day average volatility to mitigate the impact of sudden changes. However, for 7-day ahead forecasts, the CNN-LSTM model can reduce the influence of the volatility spikes compared to 1-day ahead predictions. Moreover, the CNN-LSTM model surpasses the HAR model, particularly when dealing with volatility clustering, which the HAR model struggles to capture. This advantage is due to the memory unit of the LSTM, which allows the CNN-LSTM model to outperform in both short-term and long-term forecasts.

However, the referenced study does not incorporate external market signals such as sentiment scores or trading behavior, and it primarily forecasts volatility magnitude, not directional movement (up/down).

We plan to extend this approach by integrating sentiment analysis (based on filtered human-generated tweets) to forecast the direction of price movement.

Huang, Z.-C., Sangiorgi, I., & Urquhart, A. (2024). Forecasting Bitcoin volatility using machine learning techniques. C, 97, 102064. https://doi.org/10.1016/j.intfin.2024.102064

Data Preprocessing

About the Dataset

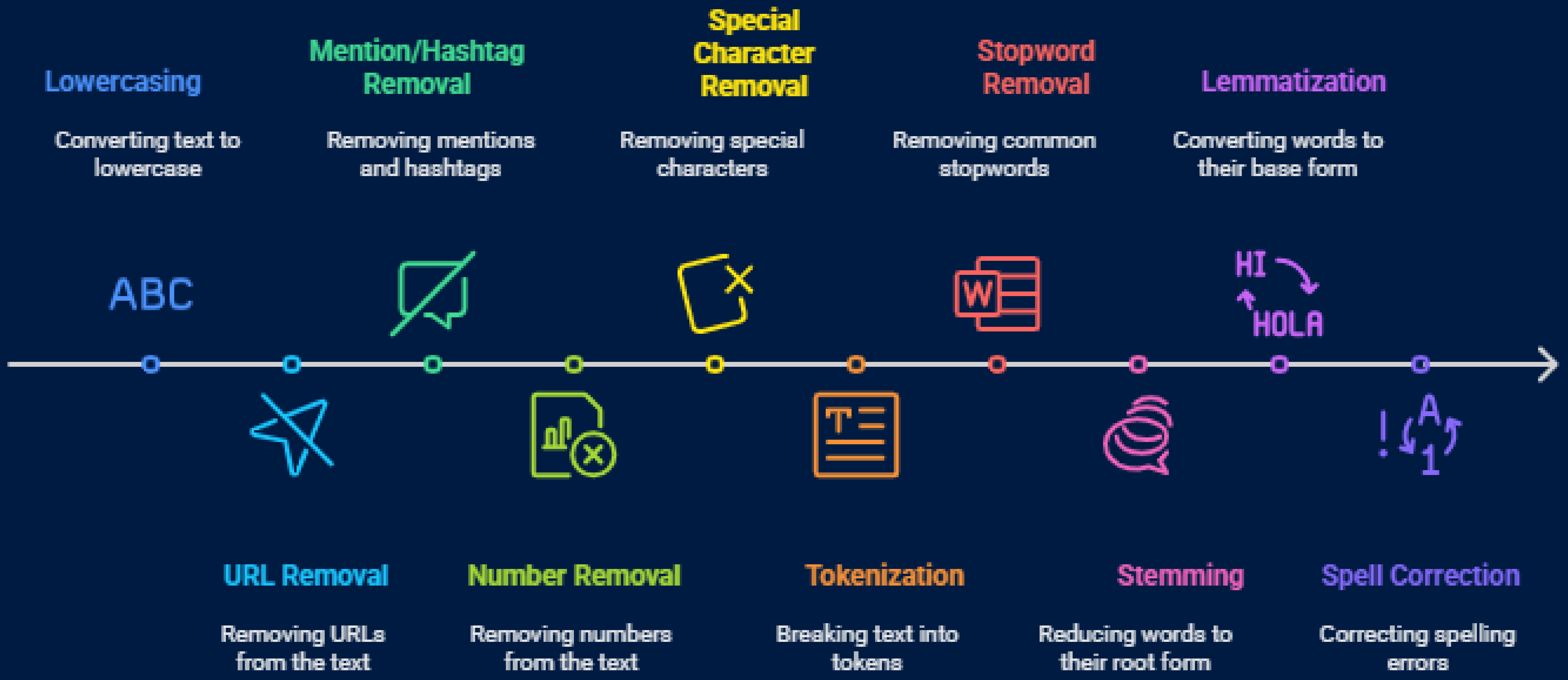
- **Dataset Name**: Bitcoin Tweets
- **Source:** Publicly available tweets collected using Twitter API.
- **Collection Method:** Tweets containing hashtags #Bitcoin and #BTC were collected using Twitter's Streaming/Search API.
- Collection began on 6th February 2021.
- **Initial volume:** Over 100,000 tweets.
- **Collected Fields:**

text, date, user followers, user friends, hashtags etc.

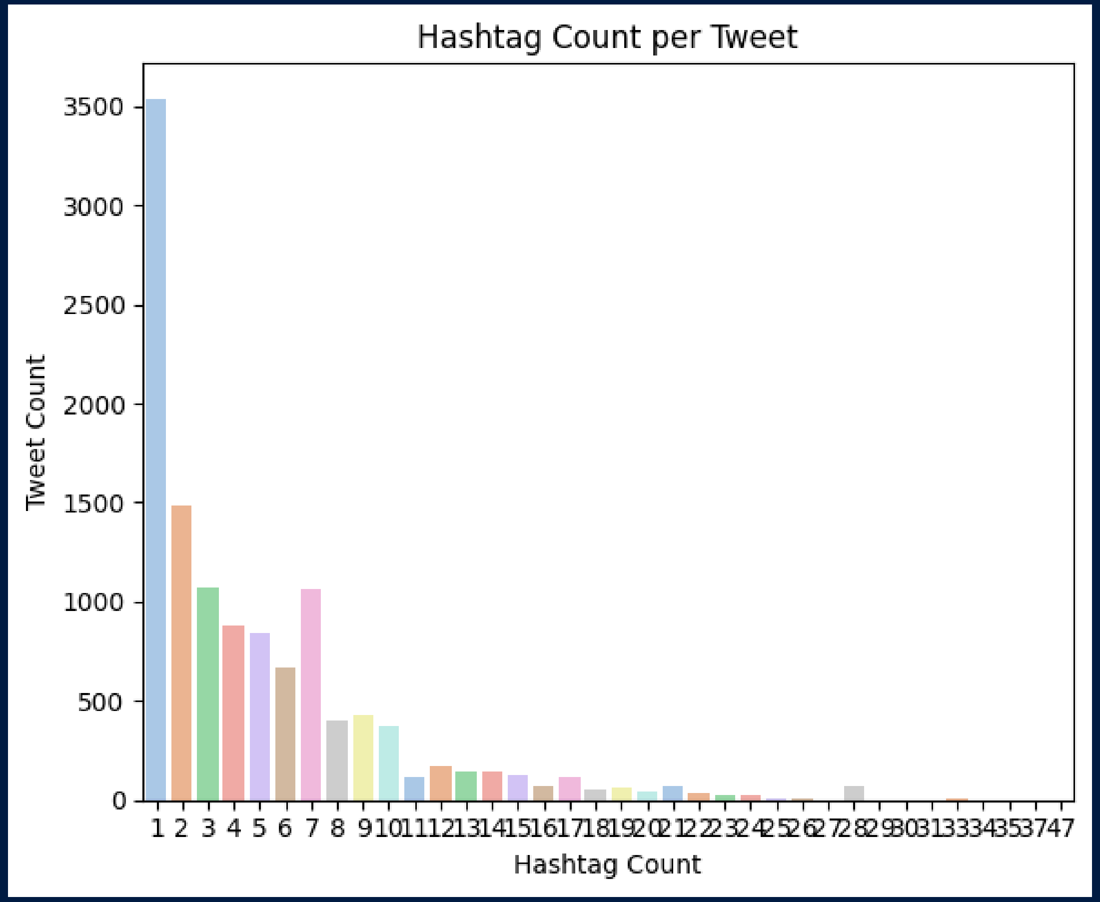
user_name	user_location	user_description	user_created	user_followers	user_friends	user_favourites	user_verified	date	text	hashtags	source	is_retweet
ChefSam	Sunshine Stat	Culinarian | Hot Sa	#############	4680	2643	6232	FALSE	############	Which #bitcoin books shou	['bitcoin']	Twitter for iPhon	FALSE
Royâšįï¸		Truth-seeking pleb	#############	770	1145	9166	FALSE	############	@ThankGodForBTC I appre	['Bitcoin']	Twitter for iPhon	FALSE
Ethereum Yoda		UP or DOWN...	#############	576	1	0	FALSE	############	#Ethereum price update:	['Ethereur	Twitter Web App	FALSE
Viction	Paris, France	https://t.co/8M3r	#############	236	1829	2195	FALSE	############	CoinDashboard v3.0 is	['Bitcoin']	Twitter for Andro	FALSE
Rosie	London	The flower langua	#############	12731	46	134	FALSE	############	#Bitcoin Short Term	['Bitcoin',	Twitter Web App	FALSE
AkinHack	United States	Professional Reco	#############	197	48	13	FALSE	############	Yâ€™all Message me for an	['CYBER', '	Twitter for Andro	FALSE
CAIR (Pump/Dump)		CAIR is a smart me	#############	5976	1	107	FALSE	############	PUMP : 4-Hour Chart (1x!)	['FILUSDT'	BinanceTW	FALSE
NFTevening	ðŸ‘‡ Join 25k+	The best newslett	#############	26940	2050	12198	FALSE	############	ðŸ“°TwelveFold by	['Bitcoin',	Hypefury	FALSE
AbdeL		#sol #NFT	#############	792	75	409	FALSE	############	@BitcoinBullsNFT The first	['NFT', 'Bit	Twitter Web App	FALSE
PUBLORD	probably dow	TOXIC HAPPY HOU	#############	22414	3225	63945	FALSE	############	Your first #Bitcoin Halving w	['Bitcoin']	Twitter for Andro	FALSE
Bitcoin Canc	Brazil	Robot that posts th	1/6/2021 1:36	40	4	1	FALSE	############	Candle of day 01/03/2023	['Bitcoin',	Bitcoin Candle Bc	FALSE
FYC crypto |	Sharing Information And Learnin		#############	6	23	1	FALSE	############	UBS Strategists Predict	['BTC', 'Bit	Twitter Web App	FALSE
Ethereum Yoda		UP or DOWN...	#############	576	1	0	FALSE	############	#Ethereum price update:	['Ethereur	Twitter Web App	FALSE
BNB Price Tracker		#BNB #BNBTracker	#############	492	5	0	FALSE	############	#BinanceCoin price	['Binance(	Twitter Web App	FALSE
â˜£ï¸ Hugo S	In your head	#Bitcoin miner/inv	#############	3307	476	9782	FALSE	############	@stacyherbert My tweets a	['Bitcoin',	Twitter Web App	FALSE

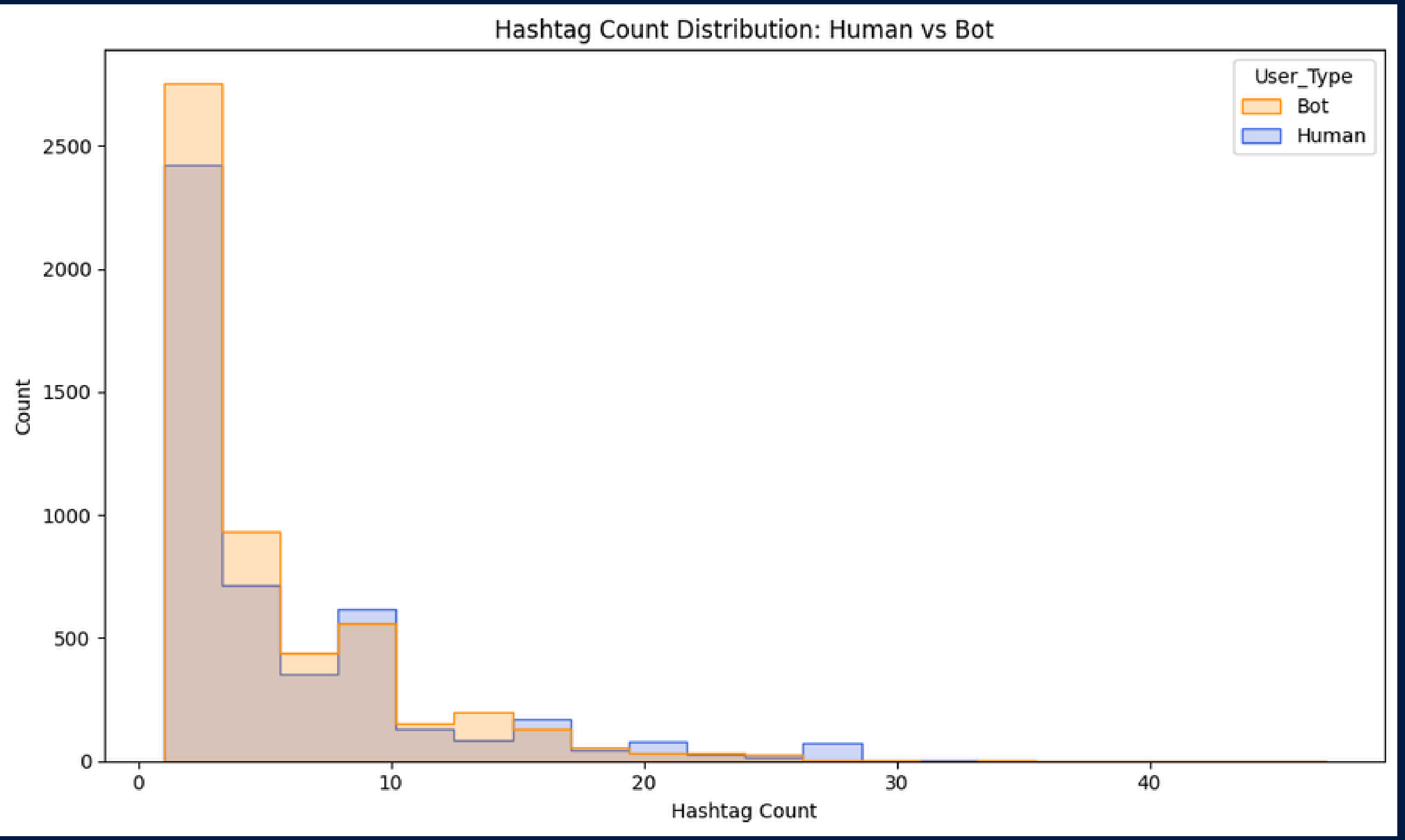
https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets?select=Bitcoin_tweets.csv

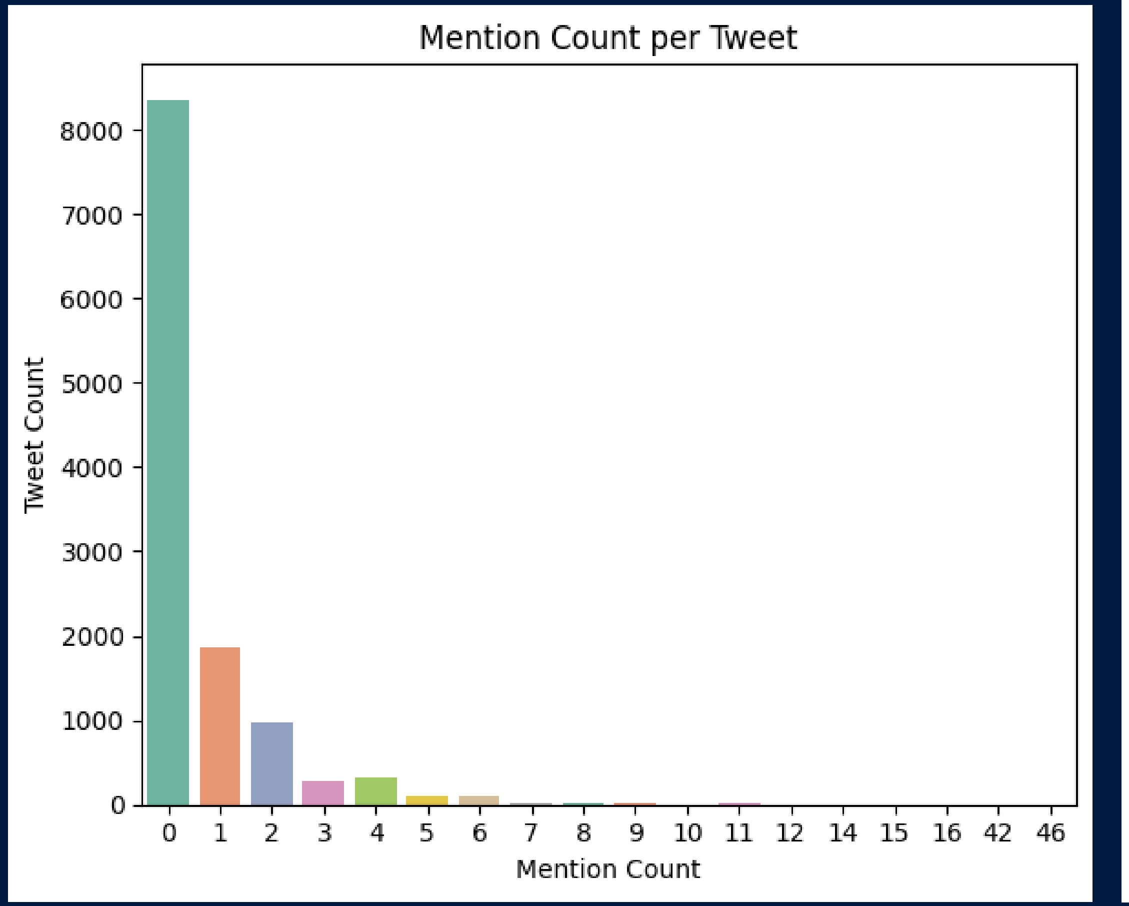
Data Preprocessing Sequence

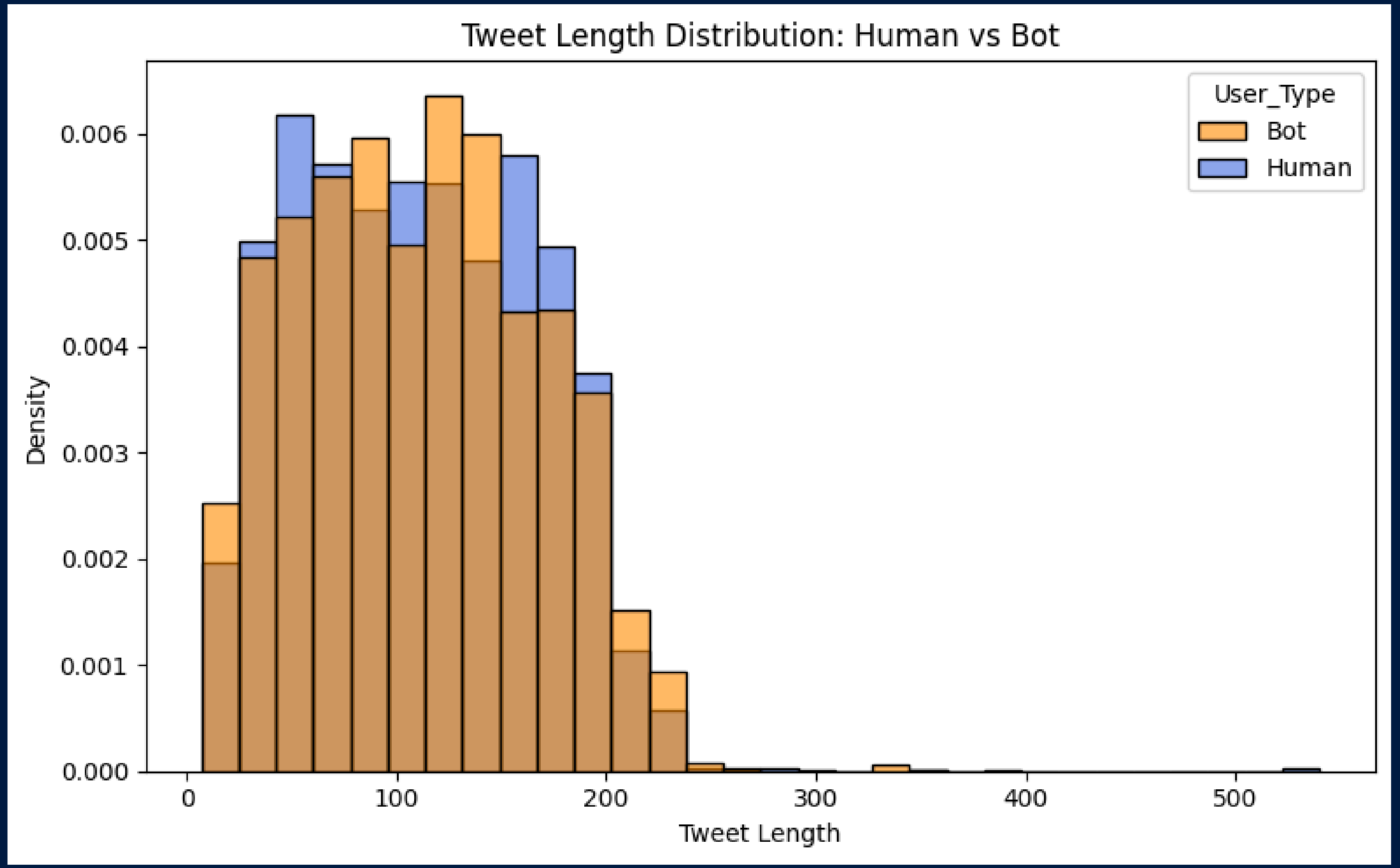


Data Visualisation









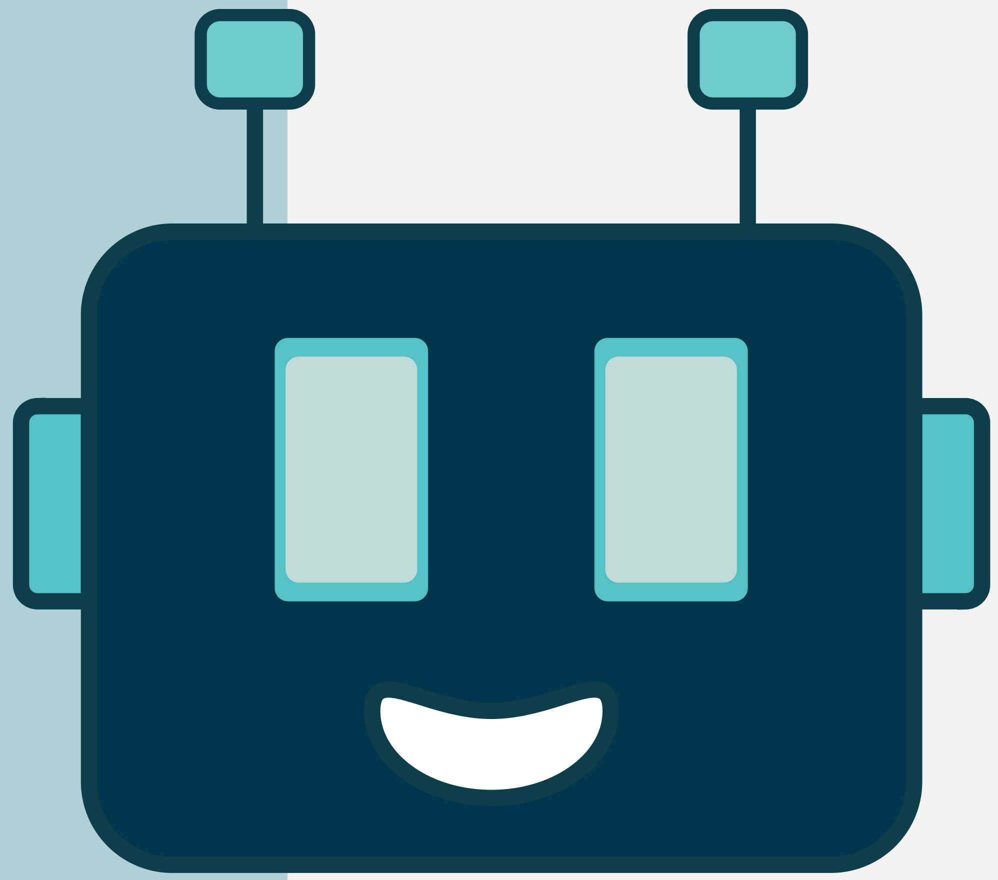
Methodology

Bot-Detection

1. **Manual Labeling of Bots**
   - We manually labeled accounts as bots or humans.
   - Looked into account metadata like Followers-to-friends ratio, Account activity , Profile details and behavior patterns

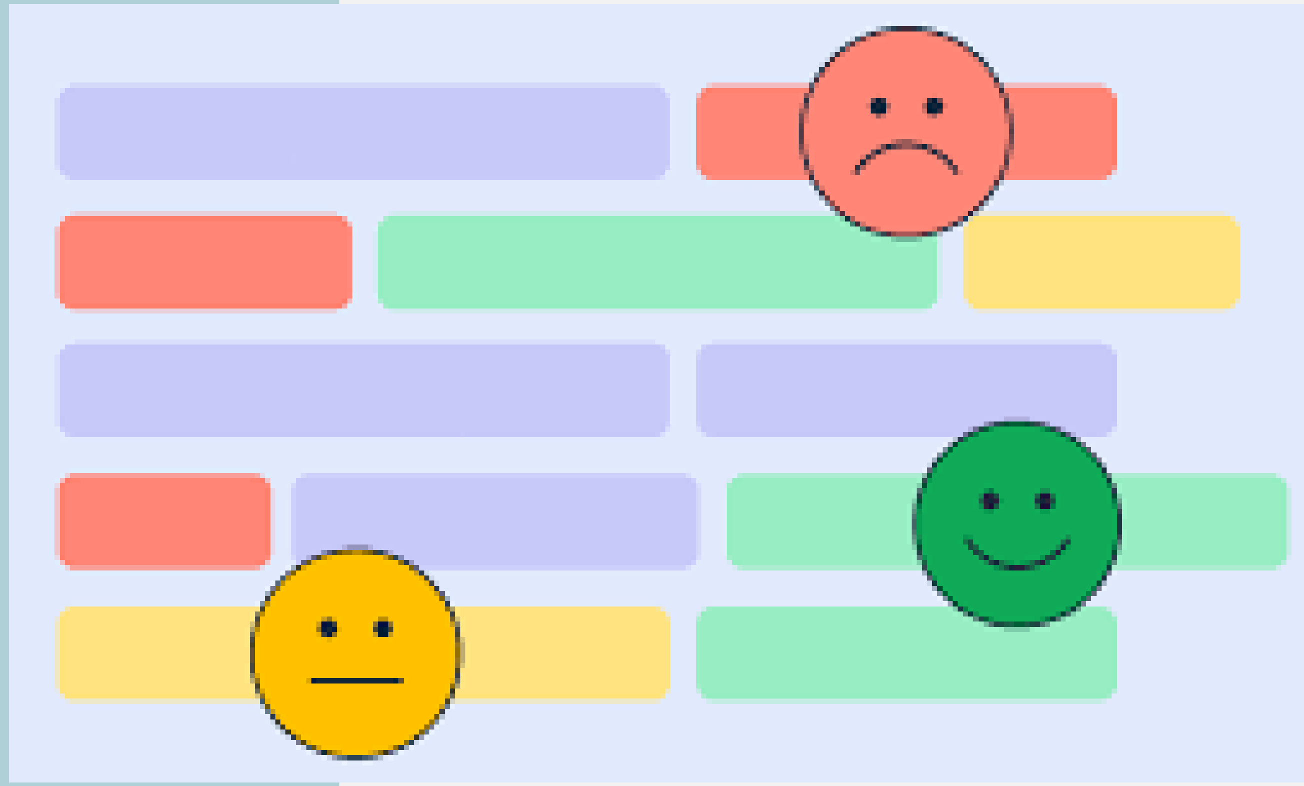
2. Feature Engineering
   - Split users into train and test groups
   - Used TF-IDF to convert cleaned tweet text into numeric vectors (unigrams and bigrams)..
   - Combined the text and metadata features into a single feature set.

3. Model Training
   - Chose Logistic Regression for classification.
   - Used GridSearchCV to find the best hyperparameters (C, penalty, solver).



Sentiment Analysis

1. **Sentiment Labeling with VADER**
   - We used VADER to assign sentiment labels (positive, negative, neutral) to tweets
   - VADER calculates a polarity score and classifies each tweet based on that score.
2. **Feature Extraction using Bag of Words**
   - Converted tweets into numeric vectors using BoW
   - This captures how often key words appear in each tweet
   - Split the dataset into training (80%) and testing (20%) sets.
3. **Sentiment Classification using SVM**
   - Trained an SVM (Support Vector Machine) to predict sentiment labels using linear “kernal”
   - SVM was chosen because it’s fast, efficient, and works well for text classification



Volatility Prediction

1. **Feature Engineering**
- Created features from Bitcoin price data and tweet sentiment scores, including volatility, price changes, and lagged values.

2. **Data Preparation**
- Normalized features and converted data into sequences of 10 time steps for time series modeling.
- Model: LSTM + CNN
- Used an LSTM layer to capture temporal patterns, followed by a 1D CNN to extract local features, ending with a fully connected layer for binary volatility classification.

3. **Training and Evaluation**
- Trained with an 80/20 train-test split. Evaluated model accuracy and classification report on test data.

Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	Benchmark (Approx.)
Volatility Prediction	82%	0.83	0.82	0.82	85-91% (typical in finance forecasti
Bot Detection	97%	0.97	0.97	0.97	95-98% (standard in bot detection
Sentiment Analysis	92%	0.92	0.92	0.92	90-95% (common in social media

Thank you!