# Bitcoin price trend prediction using tweet sentiment analysis

Krishna Sameer Kummetha
210470013
Eranjan Padumadasa
MSc Big Data Science

*Abstract*— **A lot of crypto currency traders seek Twitter tweets for guidance for trading cryptocurrency on a daily basis. In this project I tried to analyze the ability of Twitter data to predict price trends of Bitcoin cryptocurrency. Sequential CNN model was used to analyze sentiments across the tweet. The sentiment analyzed is being used to predict the upcoming trends of Bitcoin price using various supervised machine learning classifiers. This will be very useful to short term crypto currency traders for whom it will work as a tool which will assist them in their trades.**

*Keywords*— **Bitcoin, Sentiment Analysis, Supervised ML Classifiers**

## I. Introduction

As Payments and banking are getting digitalised along with the increase in access of internet to every individual, we are getting closer to being a digital payment society. This has paved way for the emergence of cryptocurrencies.

A cryptocurrency is a decentralised unit of currency which is monitored by peer to peer network called blockchain[1] which allows payment verifications and various other transactions without any central authority[2]. Bitcoin was the first decentralised digital currency introduced in 2008 by Satoshi Nakamoto in white paper “Bitcoin: A Peer-to-Peer Electronic Cash System”. Bitcoin is a digital currency independent of governments and banks where sender and receiver of the transactions cannot be tracked[3]. In the beginning, Bitcoin was primarily used by black markets like Silk Road[4] that accepted payments only through bitcoin but over time various online merchants started accepting payments through Bitcoin[5]. Since 2012 venture capitalists have started investing in Bitcoin based startups[6][7]. All these contributed to the popularity of Bitcoin as an investment option which has resulted in drastic increase in it’s price. In 2017 it reached it’s all time high of USD 19,783 in December 2017[8]. This has further increased the popularity of alternative currency idea among general public. Not only did the bitcoin revolutionised digital currencies but also led to the blockchain based tech development such as smart contracts [9] and block chain-based supply chain management [10]. Crypto currency price depends a lot on people’s perception about it that is sentiments play huge role on it’s price. People share opinions mostly on micro blogging websites such as Twitter, Tumblr and Pinterest etc.

Micro-blogging acts as broadcast medium for content such as short sentences, individual images. Most prominent of the micro-blogging platforms is twitter with 229 million as of May,2022[11]. Users can post tweets of length up to 280 characters long and attach videos, images, GIFs and links to websites with tweets. Twitter is a convenient way to broadcast views globally for millions around the world. Also, tweets can be marked with hashtags which can used to search a trending topic. Each day there are more than 800 millions tweets being tweeted on various topics [12].

## II. Related work

Primary source of information about crypto has been social media, which can be classified as Twitter, cryptocurrency forums and various new sources pertaining to cryptocurrency. Along with the various news sources [13], researchers have also used forums such as Bitccointalk.org and Reddit[14,15,16,17] to perform sentiment analysis and forecast Bitcoin price. The predictive capability of news and social media sentiment for BTC prices and volume is recognised by many researchers to be spanning over a short span of one week and long period of 1-3 months. Even the number of messages correlates with trade volume of BTC[14]. In addition to this, Karalevicius et al[13] also confirmed what was mentioned earlier; crypto investors often exxagerate to information leading to trend in the market where initially shift in markets with the sentiment and later slowly gets back to normal. Twitter sentiment analysis was used in numerous studies to predict price fluctuations for BTC.

In a study done by A.Derakshan and H. Beigy’s [18], they used text which is in English and also Parsian posted on Yahoo Finance, focussing on a huge number of stocks. They achieved an accuracy of 56% in predicting upward or downward trends, setting a standard to compare out results. They primarily focussed on stock prices prediction rather than single cryptocurrency and also we focussed on just one single cryptocurrency whereas they focussed on 18 US equities. I chose twitter because of the ease with which we can collect twitter data unlike Yahoo Finance.

In a study done by Pangolu et al[19], they used blogging in social media for forecasting stock price, as it represents public sentiments perfectly about current events. They used sentiment analysis and supervised ML methods to tweets and studied the correlation between twitter sentiment and stock price movement of company. They concluded that there is clear correlation between rise and fall of stock prices with public tweet sentiments. I have done something similar but my analysis was on Bitcoin and not stocks.

In an another research done by D.R. Pant[20], they did something close to the analysis that we did. They also used twitter as their input and tried to predict price of Bitcoin. But they achieved a relatively high accuracy than us as they labelled the data manually which resulted on a very small dataset of 7455 tweets, almost 1770 of which were declared neutral or redundant. To have a huge dataset and generalise our model more, I decided to use VADER sentiment analysis tool for labelling the tweets so I can easily label huge amount of data in a short interval of time.

In an another study done by Sara Abdali and Ben Hoskins[21], they used a pre trained Bert model[22] to make the prediction using Twitter sentiment analysis whereas I used a sequential model to make the prediction they achieved an accuracy of 53.2% which is little less than what I achieved.

## III. METHODOLOGY

### A. Dataset

There are two datasets being used here. I initially planned on using Twitter API for downloading tweets which mentioned Bitcoin for creating the dataset but then twitter placed a limit on the number of tweets that I could download. So I switched to Kaggle[23] to obtain daily tweets data for bitcoin which contains data from 01/2/2021 to 01/03/2022 and the second dataset of daily prices of Bitcoin from using Python's yahoo finance API[24] for the same time period. I chose this time period as Bitcoin faced heavy fluctuations during this period as it reached it's peak and also one of the lowest price during this period. "Fig.1. shows the trend

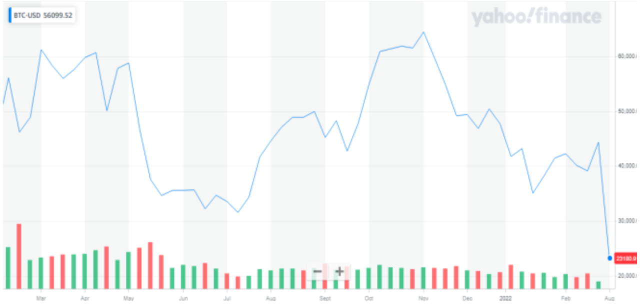


Fig. 1. Bitcoin price chart from 01/2/2021 to 01/03/2022

The dataset has 2259788 tweets that were posted frequently within the time period mentioned above. The dataset is a combination of expressions, hashtags, URL's and user's mentions. As a result of the casual usage by people in the social media these raw tweets usually contains noise because of which a lot of useful features gets lost. This unstructured data needs to be structured to create a dataset that can be learned easily by various classifiers.

### B. Data Pre-processing

A number of pre-processing steps have been implemented to eliminate inconsistencies and standardise the dataset. Pre-processing of data is done using NLTK as it is an open-source which provides APIs which are easy to use for numerous text pre-processing methods. Below are the pre-processing steps that were performed.

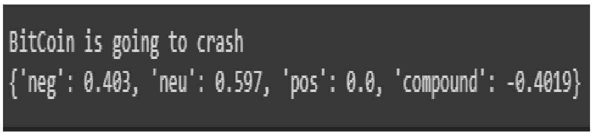
- Removed all Nan values
- Removed URLs
- Removed hashtags
- Removed mentions and characters which are not in English
- Removed punctuations
- Tokenization using tweet tokenizer:- Splitting string into a list of tokens
- Removed stop word
- Performed stemming and lemmatization

"Table 1" shows the steps being performed for pre processing tweets

TABLE I. STEP BY STEP PREPROCESSING RESULTS

| | *Pre Processing Technique* | *Result* |
|---|---|---|
| 0. | Actual Tweet | SS@bitcoin https:/twitter.com/FT Bitcoin ETF crashed but buy, think about it #btc #buy |
| 1. | Removing URL | SS@bitcoin Bitcoin ETF crashed but buy, think about it #btc #buy |
| 2. | Removing Hashtag | SS@bitcoin Bitcoin ETF crashed but buy, think about it |
| 3. | Removing Mention | bitcoin ETF crashed but buy, think about it |
| 4. | Tokenize | [' bitcoin ', 'ETF', 'crashed', 'but', ',' 'buy', 'think, 'about', 'it'] |
| 5. | Remove Punctuations | [' bitcoin ', 'ETF', 'crashed', 'but', 'buy', 'think, 'about', 'it'] |
| 6. | Removing Stopwords | [' bitcoin ', 'ETF', 'crashed', 'but', 'buy', 'think, 'about', 'it'] |
| 7. | Lemmetisation | [' bitcoin ', 'ETF', 'crashed', 'but', 'buy', 'think, 'about', 'it'] |

Next instead of manually labelling the tweets as positive negative or neutral I have used I have used VADER(Valence Aware Dictionary and Sentiment Reasoner)[25] for this purpose. I have explored other options of Pattern, BERT, Textblob [26] for labelling the dataset. But decided to go with VADER as it is specifically attuned to social media text. It's trained to decode even emoji's sentiment. This makes it effective in deciphering and labelling content properly from social media platforms such as Facebook and Twitter. It gives us result in the form of positive, negative , neutral and compound which is computed by normalising the scores of positive, negative and neutral. The result looks as shown in "Fig.2".



```
BitCoin is going to crash
{'neg': 0.403, 'neu': 0.597, 'pos': 0.0, 'compound': -0.4019}
```

Fig. 2. Vader Scoring

For my research purpose I am only taking the compound scores as it gives me the overall sentiment of tweet. I then categorised it as positive, negative or neutral based on the compound value i.e

TABLE II. COMPOUND SCORING

| | |
|---|---|
| Compound > 0 | positive |
| Compound < 0 | negative |
| Compound = 0 | neutral |

Then I split the dataset into training and testing dataset and then integer encoded the input data so that each word is a unique integer which is done using Tokeniser API. The I used to ensure that all sequences have the same length.

### C. *Tweet Sentiment Analysis*

I have used a sequential model for the sentiment analysis of the tweets. For a plain stack of layers where each layer has precisely one input tensor and one output tensor, a sequential model is appropriate [27]. Added Embedding layer. The need of embedding layer is that it enables us to convert each word into a fixed length vector of definite size. The resulting vector has real values and is dense. Word vectors' fixed length and decreased dimensions enable us to represent words more effectively. The embedding layer functions this way like a lookup table. In this table, the dense word vectors serve as the values and the words serve as the keys. The output of this layer is passed through 1D convolutional layer. In order to build a tensor of outputs, this layer creates a convolution kernel that is convolved with the layer input over a single spatial dimension[28].

The output is then passed through a MaxPooling1D[29] which down samples the input representation by taking the maximum value over a spatial window of size pool_size.

Then LSTM(Long Short Term Memory) layer is used. LSTM is a kind of recurrent neural network but works better than a normal recurrent neural networks in terms of memory. Short-term memory is a problem for traditional neural networks. By retaining the crucial information and identifying patterns, LSTMs effectively increase performance. A dense layer with softmax activation function is used which converts a vector of values to a probability distribution. Categorical cross entropy function is used. When there are two or more output labels in a multi-class classification model, Categorical cross entropy is used as a loss function. One-hot category encoding value in the form of 0s and 1 are allocated to the output label. The "Fig.3" depicts the sequential model that I have built.

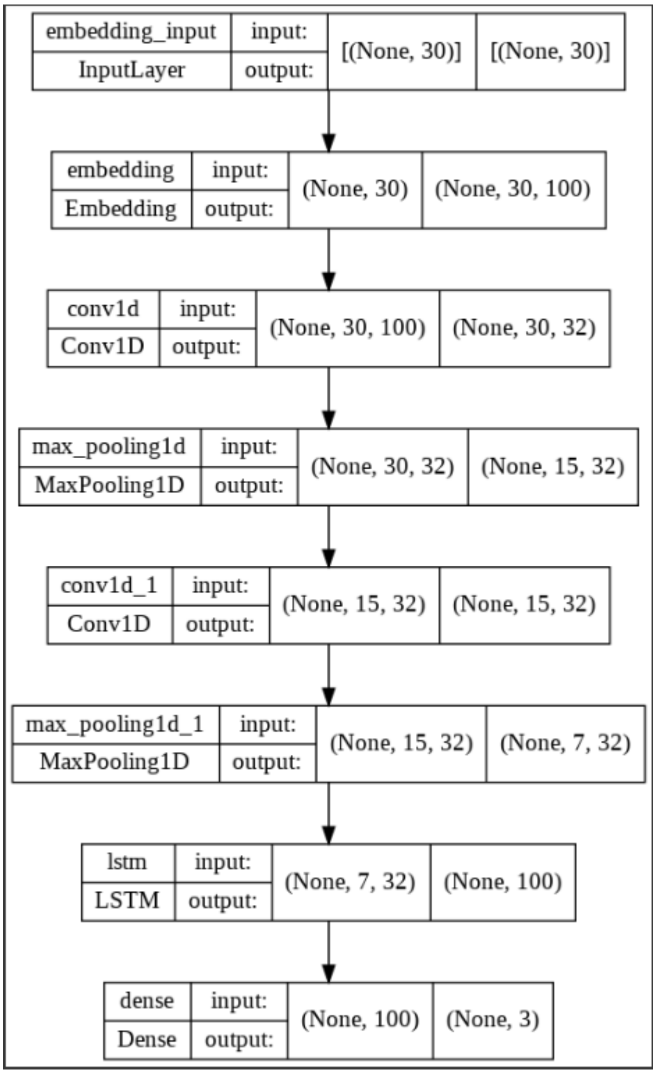


Fig. 3. Sequential Model

As we can see from the confusion matrix our model is able to strongly predict the twitter sentiments.

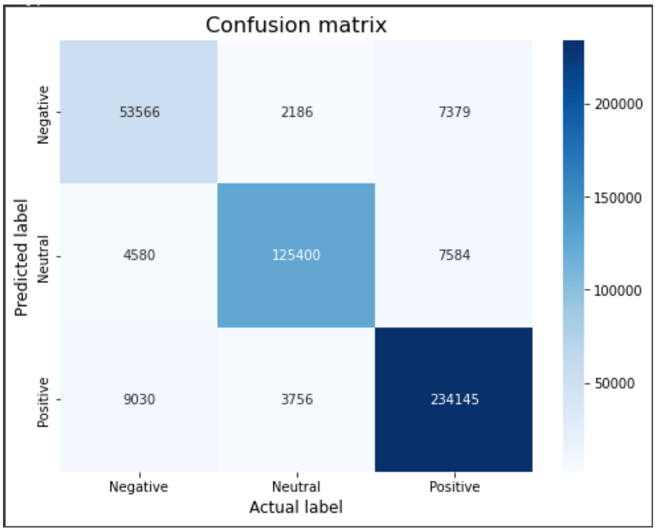


Fig. 4. Sequential Model Confusion Matrix

F1 score is the harmonic mean of precision and recall. A model with high precision score means it is performing well. We achieved high very F1 score of 0.94 for positive and 0.93 for neutral tweets indicating that our model is able to predict these correctly . Whereas the F1 score of Negative tweets is relatively low as compared to positive and neutral tweets. “Table III” shows precision, recall and F1-scores obtained by our model.

TABLE III CLASSIFICATION REPORT

| ***Label*** | ***precision*** | ***recall*** | ***F1-score*** |
|---|---|---|---|
| Negative | 0.85 | 0.80 | 0.82 |
| Neutral | 0.91 | 0.95 | 0.93 |
| positve | 0.95 | 0.94 | 0.94 |

The below graph “Fig.5” shows the training and validation accuracy of my model. It achieved a training accuracy of 96% and a validation accuracy of 92% indicating that there is only slight difference between my training and validation accuracy. After a few epochs model starts to have validation accuracy slightly dropping indicating that our model is slightly overfitting

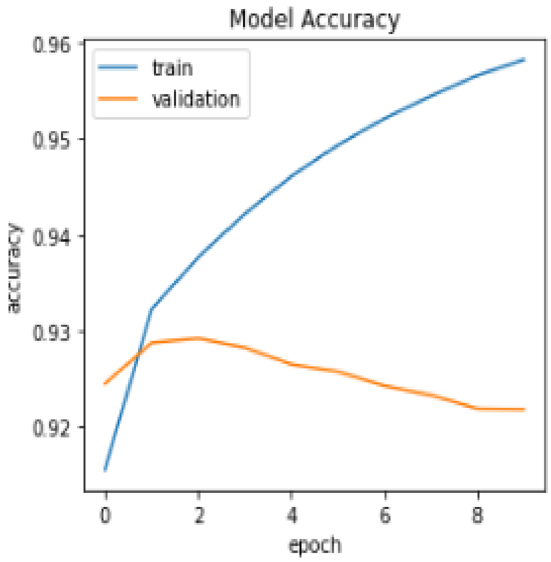


Fig. 5. Sequential Model Accuracy

As you can see, our model training loss decreases but the validation losses increases again indicating slight overfitting. The graph “Fig.6” shows the loss curve of our model.

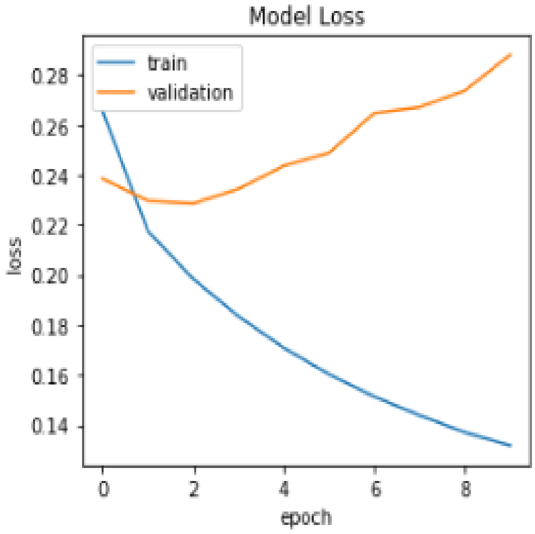


Fig.6. Sequential Model Loss

*C.Bitcoin Trend Prediction*

The output of the Tweet sentiment analysis i.e the tweets with their scores is used to predict the trend of the bitcoin for each day that is to check if the price is increasing or decreasing based on the compound score of tweets. First a new column called price sentiment is created which is what we are trying to predict which is labelled as ‘0’ or ‘1’ based on if the price decreased or increased respectively for that specific day. Next from our tweet sentiment table we extract each date along with the sentiments of all the tweets for that day. Then we split the entire data into training and testing data. The way I did the splitting is that I have divided the tweets for each day into 80% training data and 20% testing data. Then I have stacked all the training data of each day together and have done the same for the testing data. Then I used the following models to make the prediction.

*Classification*

*A.Naïve Bayes classification*:

In Bayesian classification we're interested in determining the likelihood of a label given some features [30] which can be written as

$$P(L \mid \text{features}) = \frac{P(\text{features} \mid L)P(L)}{P(\text{features})} \qquad (1)$$

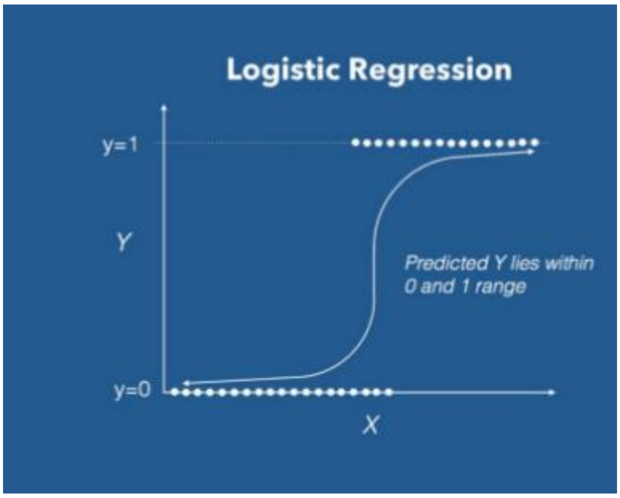
Where L is the label whose probability that we are trying to predict. Over here I am using Gaussian Naïve Bayes where the assumption is that each label’s data is drawn from a simple Gaussian distribution [30].

*B. Logistic Regression:*

Despite the name, logistic regression is more of a classification model than a regression model. Logistic regression. For situations involving linear and binary classification, logistic regression is a straightforward and more effective approach. It's a classification model that's incredibly simple to implement and performs admirably with linearly separable classes. It is extensively used for classification in the industry. It is a statistical method for binary classification that be generalised to multiclass classification [31].

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{2}$$

Fig.7. Logistic Regression



*C.Decision Tree*

A supervised machine learning algorithm called Decision Tree uses a set of guidelines to make judgments, much like how people do. The idea behind Decision Trees is to repeatedly partition the dataset until all the data points that belong to each class are isolated by using the dataset features to produce yes/no questions [30]. The advantage of decision tree is that it can easily interpreted, and also data robustness that is the algorithm can handle any type of data. It can handle a mix of categorical and numerical data.

*D.k-NN classifier:*

k-NN (K-Nearest Neighbour) algorithm is one of the simplest supervised machine learning algorithm. The closest training data point to the point we wish to generate a prediction for is the sole point that the k-NN algorithm takes into account in its most basic form. The known output for this training point is then the prediction [32].

After the predictions were made for all the tweets . The aggregate of the probabilities was taken by taking the mean of all the predictions for a day and if the probability was greater than 0.5 it was labelled as 1 indicating an increase in price and if the probability was less than 0.5 it was labelled as 0 indicating a decrease in price.

## IV. Results

*A. Naïve Bayes*

Naïve Bayes produced a **training accuracy** of **55.49%** and **testing accuracy** of **55.13%.**

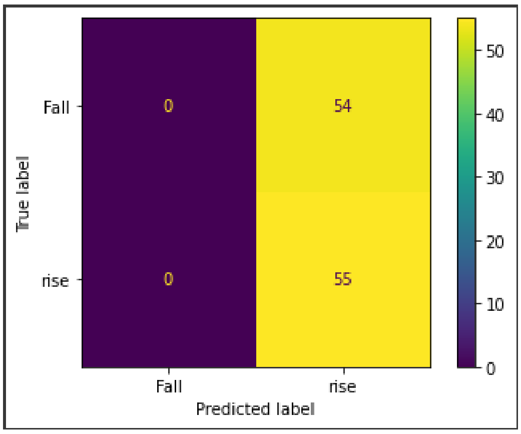


Fig .8. Naïve Bayes Confusion matrix

From the above confusion matrix we can see that True positives and false postives are high that is it is able to predict all the upward trend of increase in price but is very poor at predicting decrease in price.

*B.Logistic Regression*

Logistic regression produced a **training accuracy** of **55%** and **testing accuracy** of **55%**. The graph below “Fig.8” depicts the confusion matrix obtained after classification.

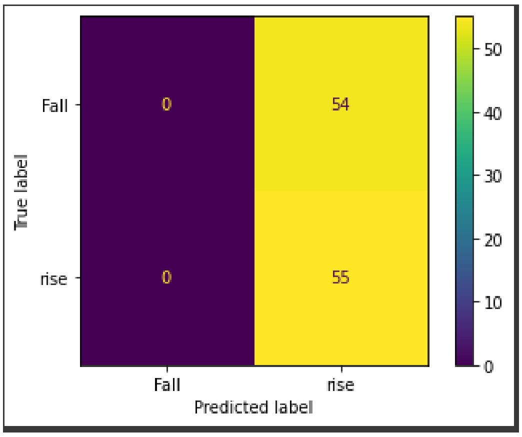


Fig .8. Logistic Regression Confusion matrix

Logistic regression produced a similar result as that of Naïve Bayes from the above confusion matrix we can see that True positives and false negatives are high that is it is able to predict all the upward trend of increase in price but is very poor at predicting decrease in price.

### *C.Decision Tree Classifier*

Decision Tree classifier produced a **training accuracy** of **87%** and **testing accuracy** of **56%**. The graph below “Fig.9” depicts the confusion matrix obtained after classification.

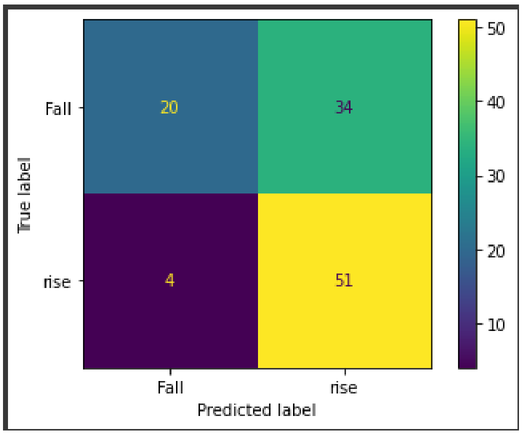


Fig .9. Decision Tree Confusion matrix

Decision Tree classifier produced a better result when compared to Naïve Bayes and Logistic Regression. As we can see almost all the positive trends predicted properly and majority of the negative trends for each day were predicted correctly.

### *D.k-NN Classifier*

k-NN classifier produced a **training accuracy** of **68%** and **testing accuracy** of **54%**. The k-NN classifier could predict all the negative trends properly but it fails to predict the positive trends in price. The graph below “Fig.10” depicts the confusion matrix obtained after classification.

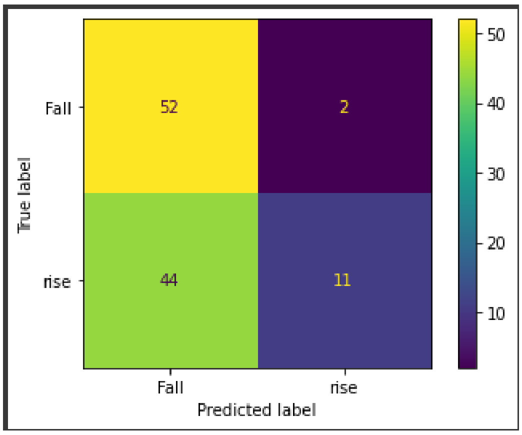


Fig .10. k-NN Confusion matrix

## V. Conclusion/Fututre work

Bitcoin is highly volatile market. The price movements depend a lot on a lot of factors. Considering this I achieved a good result of 56% in predicting the price trend using Decision Tree classifier on tweets. This work can be extended by considering various other features such as number of reshares of the tweets, considering if the tweet is verified or not , number of followers the user has among other features. We could also combine sentiments from various other sources such as reddit and social media platform to generalize the sentiment more. More advanced algorithms such as LSTMs and GRU can be used for predicting the price movements more accurately. Nonetheless, this project shows us that it is possible to predict the price movements with very little data sourced only from social media.

## References

[1] M. D. Pierro, ”What is the blockchain?”, in Computing in Science & Engineering, pp. 92 - 95, Sep 2017

[2] I. Makarov, and A. Schoar, ”Trading and arbitrage in cryptocurrency markets” in Journal of Financial Economics, vol. 135, pp. 293-319, 2020.

[3] S. Bj¨orn, ”What is bitcoin”, in Sveriges riksbank economic review, pp. 71-87, 2014.

[4] R. Bohme, N. Christin, B. Edelman, and T. Moore, “Bitcoin: Eco- ¨ nomics, technology, and governance,” Journal of economic Perspectives, vol. 29, no. 2, pp. 213–38, 2015.e

[5] A. Cuthbertson, “Bitcoin is now accepted at starbucks, whole foods and 100s of other shops,” May 2019. [Online]. Available: https://www.independent.co.uk/life-style/gadgets-andtech/news/bitcoin-stores-spend-where-starbucks-whole-foods-cryptoa8913366.html

[6] T. Simonite, “Bitcoin millionaires become investing angels,” Apr 2020. [Online]. Available: https://www.technologyreview.com/2013/06/12/15919/bitcoinmillionaires-become-investing-angels/

[7] “Ten-hut! bitcoin recruits snap to,” Dec 2014. [Online]. Available: https://www.wsj.com/articles/DJFVW00120141201eac1ag1qi

[8] D. Z. Morris, “Bitcoin hits a new record high, but stops short of 20,000 usd,” Dec 2017. [Online]. Available: http://fortune.com/2017/12/17/bitcoin-record-high-short-of-20000

[9] A. Kosba, A. Miller, E. Shi, Z. Wen, and C. Papamanthou, “Hawk: The blockchain model of cryptography and privacy-preserving smart contracts,” in 2016 IEEE symposium on security and privacy (SP). IEEE, 2016, pp. 839–858.

[10] M. Andoni, V. Robu, D. Flynn, S. Abram, D. Geach, D. Jenkins, P. McCallum, and A. Peacock, “Blockchain technology in the energy sector: A systematic review of challenges and opportunities,” Renewable and Sustainable Energy Reviews, vol. 100, pp. 143–174, 2019.

[11] “Twitter: number of monetizable daily active users worldwide 2017-2022,” May 2022.[Online] Available : https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/#:~:text=Twitter%3A%20number%20of%20monetizable%20daily%20active%20users%20worldwide%202017%2D2022&text=In%20the%20last%20reported%20quarter,mDAU%20in%20the%20previous%20quarter.

[12] “How Many Tweets per Day 2022” June 2022. [Online] Availablle: https://www.renolon.com/number-of-tweets-per-day/

[13] Karalevicius, Vytautas, N. Degrande, J. De Weerdt. "Using sentiment analysis to predict interday Bitcoin price movements." The Journal of Risk Finance (2018)

[14] F. Mai, Q. Bai, J. Shan, “From Bitcoin to Big Coin: The Impacts of Social Media on Bitcoin Performance”, SSRN Electron. J 1-16, January 2015.

[15] Y. Kim, J. Kim, W. Kim, “Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies”, PloS ONE 11(8), August 2016.

[16] Y. Kim, J. Lee, N. Park, "When Bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation", PloS ONE 12(5), May 2017.

[17] Xie, Peng. "Predicting digital currency market with social data: Implications of network structure and incentive hierarchy". Diss. Georgia Institute of Technology, 2017.

[18] Ali Derakhshan and H. Beigy,"Sentiment analysis on stock social media for stock price movement prediction". Available at SSRN 2607167 (2015).

[19] V. Pagolu, K. Challa, G. Panda, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements", International conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016.

[20] D. Pant and Prasanga Neupane,"Recurrent Neural Network Based Bitcoin Price Prediction by Twitter Sentiment Analysis"

[21] Sara Abdali and Ben Hoskins,"Twitter Sentiment Analysis for Bitcoin Price Prediction"(2021)

[22] Jacob Devlin, Ming-Wei Chang,"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding"(2018)

[23]https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets. [Online]. Available: https://www.kaggle.com/datasets/kaushiksuresh147/bitcoin-tweets

[24] https://finance.yahoo.com/quote/BTC-USD/history. [Online]. Available: https://finance.yahoo.com/quote/BTC-USD/history/

[25] https://github.com/cjhutto/vaderSentiment. [Online]. Available: https://github.com/cjhutto/vaderSentiment

[26] https://medium.com/geekculture/what-nlp-library-you-should-use-for-your-sentimental-analysis-project-bef6b357a6db. [Online]. Available: https://medium.com/geekculture/what-nlp-library-you-should-use-for-your-sentimental-analysis-project-bef6b357a6db

[27] https://keras.io/guides/sequential_model/. [Online]. Available: https://keras.io/guides/sequential_model

[28] https://keras.io/api/layers/convolution_layers/convolution1d/. [Online]. Available: https://keras.io/api/layers/convolution_layers/convolution1d/

[29] https://keras.io/api/layers/pooling_layers/max_pooling1d/. [Online]. Available: https://keras.io/api/layers/pooling_layers/max_pooling1d/

[30] Jake VanderPlas "Python Data Science Handbook: Essential Tools for Working with Data"

[31] Abdulhamit Subasi, in Practical Machine Learning for Data Analysis Using Python, 2020

[32] Andreas C. Müller & Sarah Guido "Introduction to Machine Learning with Python"