Journal of Finance – Marketing Research; Vol. 15, Issue 08; 2024
p-ISSN: 3030-4296; e-ISSN: 3030-430X
DOI: https://doi.org/10.52932/jfm.v15i8

Journal of Finance – Marketing Research

http://jfm.ufm.edu.vn

BITCOIN PRICE MOVEMENT PREDICTION BY NEWS SENTIMENT USING MACHINE LEARNING APPROACH

Phan Huy Tam[1*], Chu Quang Thuy[1]

[1]University of Economics and Law, Vietnam

ARTICLE INFO

DOI:
10.52932/jfm.v15i8.515

Received:
March 31, 2024
Accepted:
August 03, 2024
Published:
November 25, 2024

Keywords:
Bitcoin movement;
Google news;
Machine learning;
Sentiment analysis.

JEL codes:
G17, C53

ABSTRACT

This study investigates the potential of news sentiment, derived from Google News, in predicting Bitcoin price movements. It explores the correlation between sentiment in news headlines and Bitcoin's market behavior. Employing a data set of news headlines related to Bitcoin from Google News, this research applies sentiment analysis and various machine learning algorithms, including Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Naïve Bayes, and KNN. The approach involves extracting sentiment scores and correlating these with historical Bitcoin price data. The analysis revealed that Decision Tree and Random Forest algorithms provide a balanced prediction for Bitcoin price movements. Logistic Regression and Support Vector Machine showed high AUC scores but with unbalanced class predictions. Naïve Bayes and KNN were less effective. Overall, sentiment analysis of news headlines can predict short-term Bitcoin price movements to a reasonable degree. This research offers a novel tool for investors and market analysts, enhancing understanding of the impact of news sentiment on cryptocurrency prices. It provides a predictive model to assist in investment decision-making and market analysis. This study contributes to the literature on financial forecasting by integrating sentiment analysis with machine learning for cryptocurrency price prediction. It's one of the few studies to analyze the impact of news sentiment on Bitcoin price, offering a new perspective in the realm of financial technology and digital currencies.

*Corresponding author:
Email: tamphan.ntc@gmail.com

1. Introduction

Since the launch of Bitcoin, this type of cryptocurrency has become a superstar and gained the trust of an increasing community. Bitcoin has the highest market capitalization in terms of cryptocurrency and is considered to be called "digital gold". This achievement is based on investors' faith in the future of Bitcoin and the underlying technology (Prajapati, 2020). In recent days, the huge demand for Bitcoin trading shows the interest of the market in cryptocurrencies and blockchain technology in general. Bitcoin is a decentralized cryptocurrency that does not require any central authority such as banks and could be transferred and used for many purposes through a peer-to-peer network. The circulation of Bitcoin uses a special algorithm which was introduced under the name of Satoshi Nakamoto, all transactions are public on a system of distributed ledge to ensure transparency and anonymity at the same time (Nakamoto, 2008). Figure 1 shows the historical price of Bitcoin since its inception. The graph shows a sharp increase in Bitcoin value since the first released in 2009 to a peak of around $20,000 in 2017. This indicates that Bitcoin is a good investment term and attracts a vast amount of capital around the world. The number of Bitcoin users was around 4 million when it reached the highest value of all time in 2017 (Hileman & Rauchs, 2017).

Even if the Bitcoin owner used it as a means of payment or treated it as an investment, the fluctuation in the value of Bitcoin is uncertain. This new store of value is considered a unique asset and behaves in ways similar to both standard financial and speculative assets (Kristoufek, 2015). This makes Bitcoin prices extremely hard to predict (Abraham et al., 2018). Since people's trust is involved in the rise of the cryptocurrency market, the sentiment of the general population does make a huge impact on the future of cryptocurrency market capitalization (Kristoufek, 2015). The application of text-based data with sentiment analysis to predict the future trend of cryptocurrencies is widely used across academic fields. The data source varies from journal news headlines, content, or social media posts like Facebook, Twitter, or even anonymous communities such as Reddit (Abraham et al., 2018; Gerritsen et al., 2022; Huang et al., 2021; Kristoufek, 2015; Mittal et al., 2019; Vo et al., 2019).

Google News is a powerful search engine to collects posted news from various sources worldwide. Furthermore, this nice source of text-based data is capable of crawling news based on selected keywords as well as other information related to the text-based data such as posted date, author… In this research, the author uses text-based data to feed in the sentiment analysis and then combines it with the historical price and volume of Bitcoin. The processed data is used to train a machine-learning model to investigate the correlation between these variables.

This research examines the prediction power of newspaper headlines collected from the Google News platform, using sentiment analysis and various machine learning algorithms. This type of non-quantifiable data showed a potential effect in predicting cryptocurrency trend direction by capturing the overall market point of view about the volatility of the cryptocurrencies. The test results provide evidence of the relationship between text-based data from news headlines and the short-term movement of Bitcoin price in both academics and practice.

The paper includes five parts: Section 1 introduces research issues; Section 2 presents a theoretical overview; Section 3 presents data collection and methods; Section 4 presents the results of empirical research and discussion; the final Section 5 presents the conclusion and recommendation.

2. Literature review

Sentiment analysis

According to Zhao et al. (2016), sentiment analysis, sometimes called "opinion mining" is a method that attempts to quantify people's opinions, sentiments, emotions, appraisals, and attitudes toward entities such as products, services, organizations, individuals, issues, events, topics, and their attributes. Thanks to the rapid development of online community platforms like Blogs, Facebook, Twitter, and search engines such as Google News, Yahoo… A huge volume of data that contains opinions is available to research in various fields including sciences, business, finance… From the point of view of psychology, human behaviors are heavily affected by opinion or in other words, the way we see the world. All choices, activities, or even just thoughts are formed by the beliefs and perceptions of reality and the conditions that how the individual or the whole community evaluates it (Zhang et al., 2018).

On the other hand, the diversification of topics and the nature of text-based information from online sources raise difficulties in finding and processing the sentiment. Also, large and complex data from websites like blogs and forums contain huge amounts of opinion which is almost impossible to identify or extract relevant information by human nature. This is the opportunity for automated text-based data processors such as sentiment analysis models. Many institutions or companies have paid great attention and invested heavily in building sentiment analytics systems. (Zhao et al., 2016).

The combination of sentiment analysis and machine learning algorithms has been applied in much previous research. The machine learning models used to process the sentiment data vary from supervised machine learning methods including Naïve Bayes, Support Vector Machine, Decision Tree… and other unsupervised algorithms. The emergence of this technique enlarges the applications domains ranging from computer vision and speech recognition to NLP. In short, sentiment analysis is a method that can transform are extract opinions from text-based data. The result of this process not only spreads from positive, negative, or neutral but also gives researchers the numeric evaluation of the sentiment such as weights, subjectivity, and polarity scores. These data could be used as input for machine learning algorithms to analyze the effect of the text-based content on the historical price of cryptocurrencies.

Previous studies

Text-based data is used widely in machine learning applications and natural language processing techniques in both practical and academic fields. Many studies use text-based data combined with machine learning algorithms to predict financial asset price or price movement. For example, Gerritsen's (2022) research uses a hand-collected dataset containing bullish, neutral, and bearish predictions for Bitcoin published by crypto experts and concluded that neutral and bearish predictions are followed by negative abnormal returns whereas bullish predictions are not associated with nonzero abnormal returns.

Besides, recent studies in big data analytics and natural language processing have developed automatic techniques for analyzing sentiment in social media information. In addition, the growing user base of social media and the high volume of posts also provide valuable sentiment information to predict the price fluctuation of the cryptocurrency. The research of Vo et al. (2019) is directed at predicting the volatile price movement of cryptocurrency by analyzing the sentiment in social media and finding the correlation between them using a method to identify the sentiment of the Chinese social media posts from the most popular Chinese social media platform Sina-Weibo. The study proposes a long short-term memory (LSTM) based recurrent neural network alongside the historical cryptocurrency price movement to predict the price trend for future time frames. Furthermore, Prajapati (2020) researches cryptocurrencies and Reddit posts using the LSTM model. This research states that social sentiment gives a good estimate of how future Bitcoin values may move.

The study of Huang et al. (2021) tries to analyze the ability of news data to predict price fluctuations for the second largest cryptocurrency in terms of market capitalization: Ethereum. The model can directly predict price direction by indicating whether to buy, sell, or hold. The final version of the model could correctly predict cryptocurrency prices using historical data and sentimental information gained from news data. This paper showed that sentiment analysis is an important perspective for cryptocurrency price prediction due to the interactive nature of financial activities.

Other sources of text-based data on socio-media such as Twitter, and Facebook… are very potential. Abraham et al. (2018) present a method for predicting changes in Bitcoin and Ethereum prices utilizing Twitter data and Google Trends data and conclude that tweet volume, rather than tweet sentiment (which is invariably overall positive regardless of price direction), is a price direction predictor. Mittal et al. (2019) prove the correlation between Bitcoin price and Twitter and Google search patterns. The test result based on Linear regression, polynomial regression, Recurrent Neural Network, and Long Short-Term Memory analysis concludes that there is a relevant degree of correlation between Google Trends and Tweet volume data with the price of Bitcoin, and with no significant relation with the sentiments of tweets.

3. Methodology

This study uses headlines on news from the search query for the keyword "Bitcoin" on the Google News search platform and the historical price of Bitcoin. The text data is scraped using the pygooglenews library and the Bitcoin daily historical price is collected using the Yahoofinance library on Python. The time length of the data is 5 years from April 01, 2017, to April 01, 2022, which results in 34,734 new articles. The collected article headlines are limited in this time frame because of the available limitation of historical posts extracted from online newspapers newspaper.

The data is processed by the following steps: (i) scrap and collect raw data; (ii) pre-processing data; (iii) performing sentiment analysis on text data; (iv) gathering sentiment data and historical data; and (v) applying different machine learning models on data. In that:

- Scrap and collect raw data: use pygooglenews and Yahoofiance library on Python
- Pre-processing data: clean data, deal with unstructured data, missing data…
- Perform sentiment analysis: use the Textblob library on Python to identify polarity, subjectivity, and sentiment on text data.
- Gather data: both sentiment data from the previous step and historical data of Bitcoin price are gathered daily. In that, if there is more than 1 new article in 1 day, the polarity, and subjectivity of that day is the average of polarity and subjectivity of all new articles published on that day. Besides, the Bitcoin price historical data is transferred into category data (1 if the daily price increased, and 0 if otherwise).
- Apply machine learning models: the final data is used as input for different classifier

machine learning models including Logistic Regression, Support Vector Machine, Naïve Bayes, KNN, Decision Tree, and Random Forest.

TextBlob is a straightforward yet powerful library for processing textual data, offering a consistent API for various natural language processing tasks, including sentiment analysis. In the context of predicting Bitcoin price movements based on news sentiment, TextBlob's sentiment analysis utilizes two main attributes: polarity and subjectivity. Polarity measures positivity or negativity degree in a text, ranging from -1.0 (most negative) to +1.0 (most positive). A negative polarity score indicates negative sentiment, a positive polarity score indicates positive sentiment and a score close to 0 indicates neutral sentiment. This is particularly useful in analyzing news headlines to gauge market sentiment toward Bitcoin. Subjectivity quantifies the degree to which a text is subjective or objective, ranging from 0.0 (very objective) to 1.0 (very subjective). Objective texts are based on facts and evidence, while subjective texts reflect personal opinions, feelings, and beliefs. Understanding the subjectivity of news articles helps assess the reliability and bias of the information influencing Bitcoin prices.

TextBlob's sentiment analysis is based on a pre-trained model that uses a large corpus of labeled data, incorporating a lexicon with associated polarity and subjectivity scores. The lexicon-based approach involves checking each word in the text against the lexicon. If a word is found, its polarity and subjectivity scores are used in the analysis. The sentiment analysis mechanism involves several steps. First, the text is tokenized into individual words or tokens. The polarity and subjectivity scores for each word are then retrieved from the lexicon. The overall polarity of the text is computed by averaging these scores, and a similar process is applied to calculate subjectivity. The final sentiment score for the entire text is obtained by aggregating the individual word scores. In the context of predicting Bitcoin price movements, the sentiment scores derived from news headlines are used as inputs for various machine learning models. This process allows for the assessment of how market sentiment, as captured through news headlines, correlates with historical Bitcoin prices. TextBlob's lexicon-based approach, relying on a pre-trained model, provides an efficient method for quantifying sentiment, making it a valuable tool in financial forecasting research (Zhao et al., 2016; Zhang et al., 2018).

The data fed to machine learning models is divided into 2 sets: training set and test set which account for 80% and 20% of the data respectively. The data of the training set is used to train and fit the machine learning algorithms. After that, the model is evaluated using the test data set. The results of classifier prediction are evaluated by the confusion matrix, precision, recall, and f1-score. In that, the confusion matrix: consists of 4 numbers divided into actual class and predicted class, in the research case of this paper.

4. Results and discussion

This paper aims to predict the movement of Bitcoin by using sentiment text-based data from news article headlines with different machine learning algorithms. The data includes sentiment derived from the new article headlines on the the Google News search platform and Bitcoin's daily historical price. The summary of sentiment analysis is as follows:

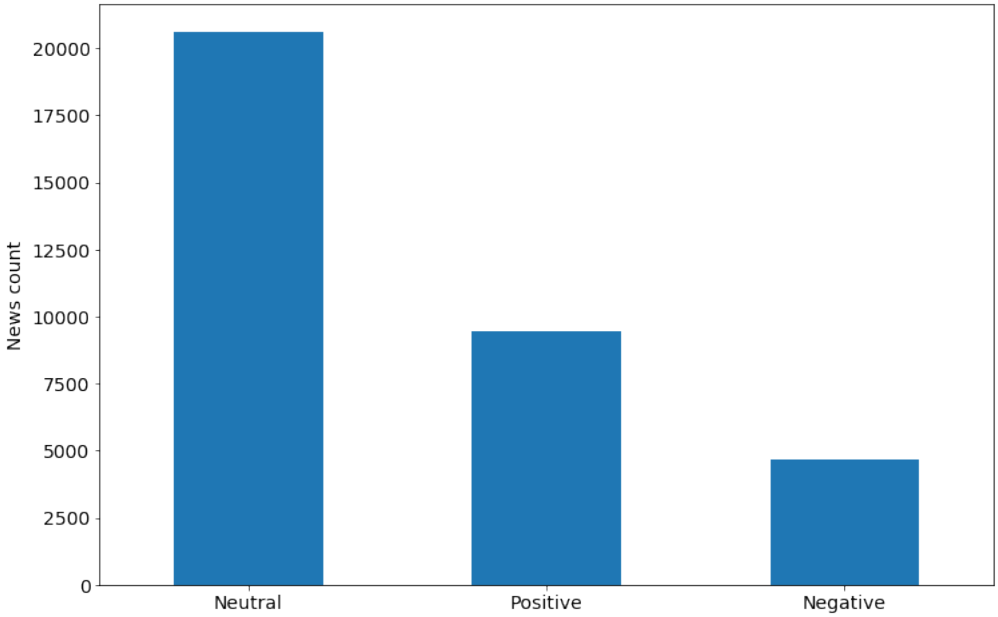


Figure 1: Sentiment analysis data count

Figure 1 shows the number of new articles according to the sentiment analysis result. Most new articles are classified as neutral (more than 20,000 articles). The number of positive articles is nearly 10,000 and almost double the number of negative articles.

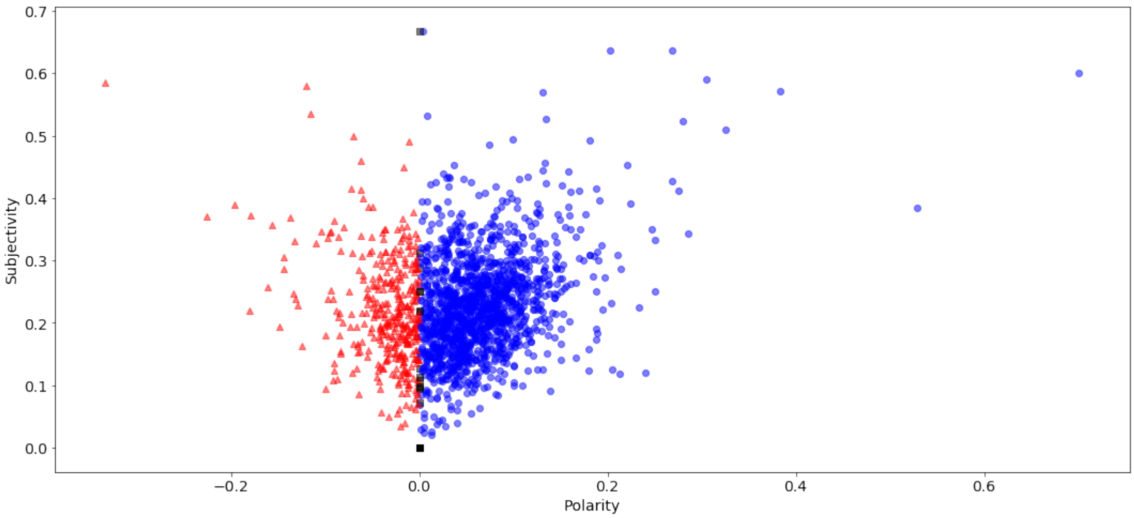


Figure 2: Polarity vs subjectivity

Figure 2 shows the scatter plot between polarity and the subjectivity. The number of articles with positive polarity is much higher than the number of articles with negative polarity while the subjectivity of articles is spread almost evenly from the base point of 0.

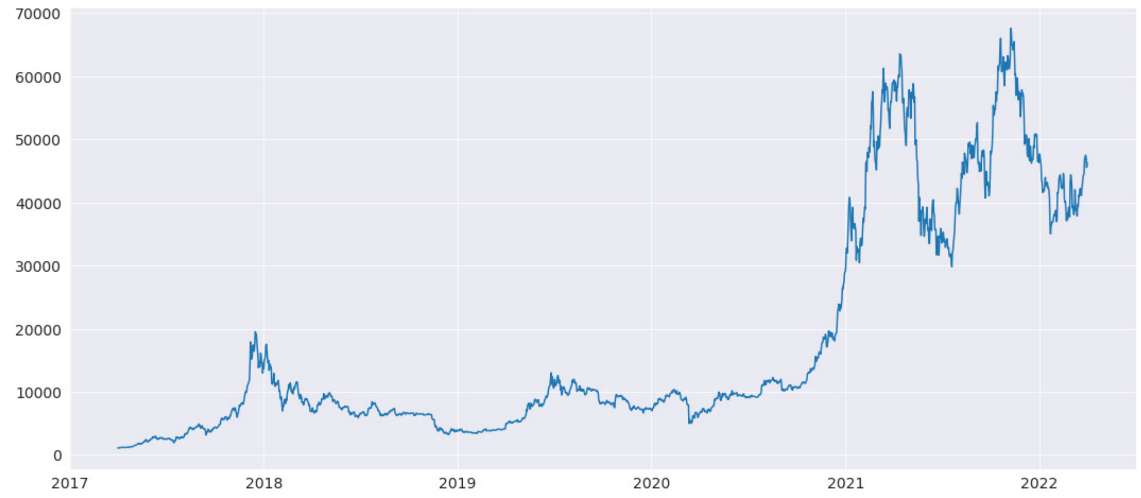
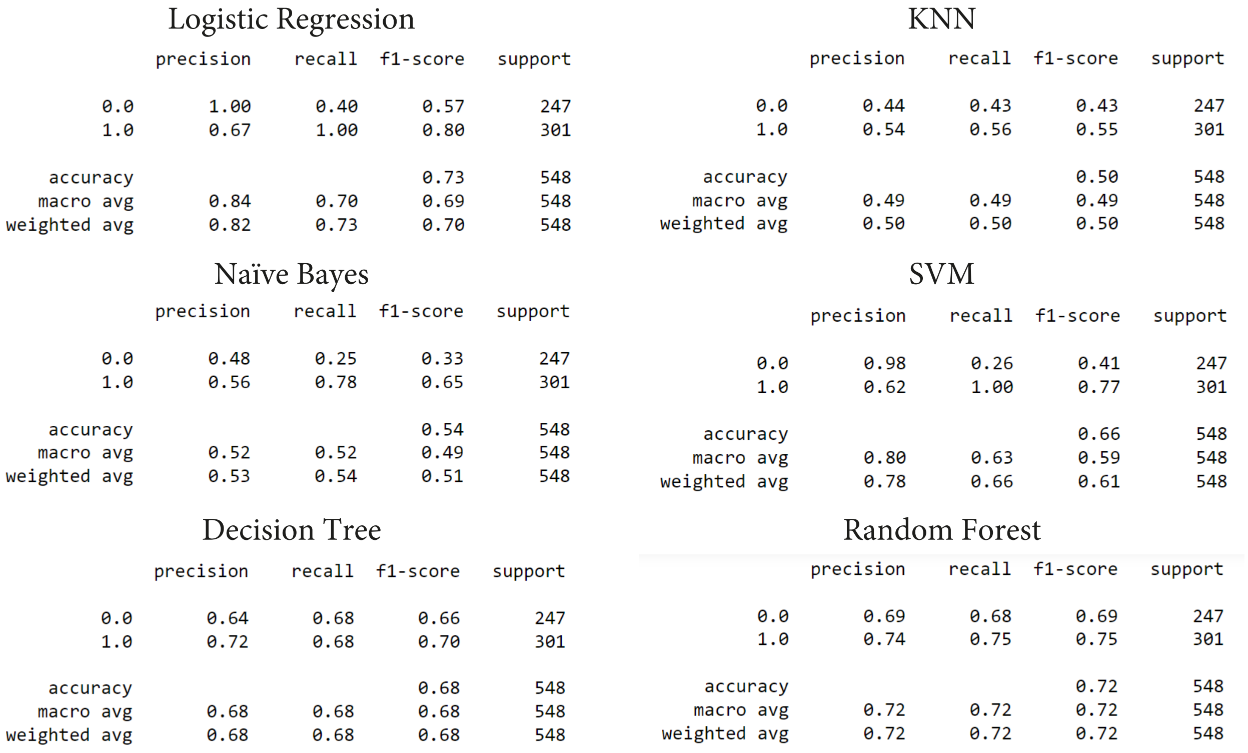


Figure 3: Bitcoin historical price from April 01, 2017, to April 01, 2022

Figure 3 shows the historical price of Bitcoin from April 01, 2017, to April 01, 2022. Overall, Bitcoin price experienced an uptrend with a huge jump during late 2020 and early 2021.

Table 1. Classification report for machine learning algorithms



Logistic Regression

```
              precision    recall  f1-score   support

         0.0       1.00      0.40      0.57       247
         1.0       0.67      1.00      0.80       301

    accuracy                           0.73       548
   macro avg       0.84      0.70      0.69       548
weighted avg       0.82      0.73      0.70       548
```

KNN

```
              precision    recall  f1-score   support

         0.0       0.44      0.43      0.43       247
         1.0       0.54      0.56      0.55       301

    accuracy                           0.50       548
   macro avg       0.49      0.49      0.49       548
weighted avg       0.50      0.50      0.50       548
```

Naïve Bayes

```
              precision    recall  f1-score   support

         0.0       0.48      0.25      0.33       247
         1.0       0.56      0.78      0.65       301

    accuracy                           0.54       548
   macro avg       0.52      0.52      0.49       548
weighted avg       0.53      0.54      0.51       548
```

SVM

```
              precision    recall  f1-score   support

         0.0       0.98      0.26      0.41       247
         1.0       0.62      1.00      0.77       301

    accuracy                           0.66       548
   macro avg       0.80      0.63      0.59       548
weighted avg       0.78      0.66      0.61       548
```

Decision Tree

```
              precision    recall  f1-score   support

         0.0       0.64      0.68      0.66       247
         1.0       0.72      0.68      0.70       301

    accuracy                           0.68       548
   macro avg       0.68      0.68      0.68       548
weighted avg       0.68      0.68      0.68       548
```

Random Forest

```
              precision    recall  f1-score   support

         0.0       0.69      0.68      0.69       247
         1.0       0.74      0.75      0.75       301

    accuracy                           0.72       548
   macro avg       0.72      0.72      0.72       548
weighted avg       0.72      0.72      0.72       548
```

Table 1 shows the classification report for 6 different machine-learning algorithms. In that, KNN and Naïve Bayes have the lowest performance for both precision and recall aspects. Decision Tree and Random Forest have significantly better results but are still lower than Logistic Regression and SVM models.

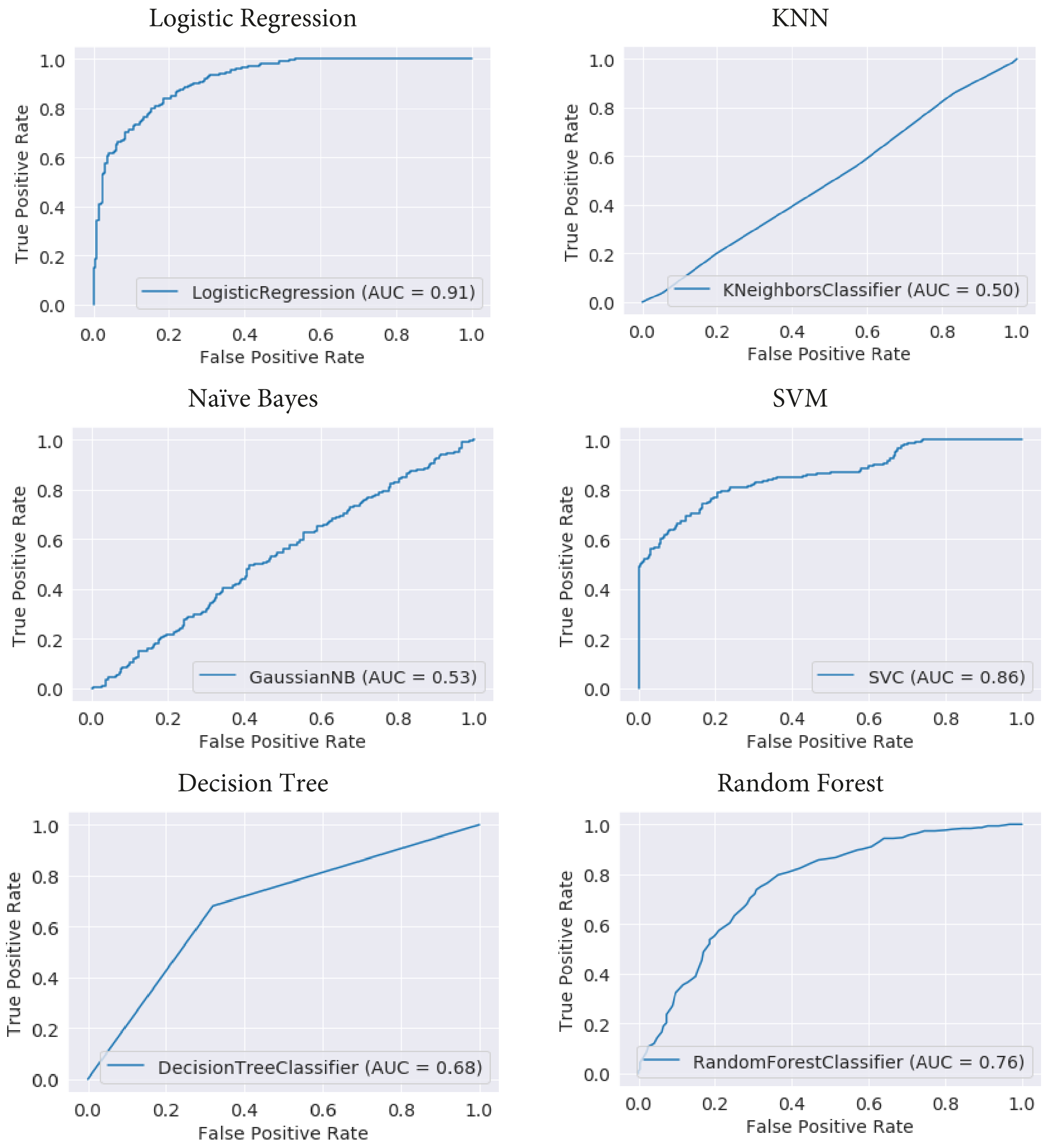


Figure 4. ROC for different machine-learning algorithms

Figure 4 shows the ROC curve for 6 machine-learning algorithms. Overall, the AUC of KNN and Naïve Bayes are very close to 0.5, which means the result of model KNN and Naïve Bayes are almost useless. The AUC of the Decision Tree, Random Forest, and SVM are relatively high at 0.68, 0.76, and 0.86 respectively. The highest AUC is from Logistic Regression at 0.91. So, based on the perspective of ROC, Logistic Regression is the best model compared to the other 5 models.

5. Conclusion

This research aims to analyze the impact of newspaper headlines by the keyword "Bitcoin" scrapped from the Google news search platform and the Bitcoin historical price movement. Applying sentiment analysis using supported packages in Python, and 6 different machine learning algorithms, the test results show that the performance in predicting movement of Bitcoin price varies between those algorithms. In more detail, the Logistic Regression model and Random Forest have the highest performance compared to the other models. SVM and Decision Tree models also show potential while Naïve Bayes and KNN models are considered useless in predicting Bitcoin price movement by sentiment data on article news.

The author acknowledges the limitations of this paper: *(i)* only focus on new articles scraped from the Google news platform. Further research could extend the source of new scrapped to increase the sample size; *(ii)* this research only uses the article headlines but ignores the content and other information in the articles. Further studies might include some other content of the article such as full written content, source of the news… *(iii)* the study only applies a few common machine learning algorithms, further research could use more complex algorithms and have a wider range of comparison between machine learning algorithms.

References

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review, 1*(3). https://scholar.smu.edu/datasciencereview/vol1/iss3/1

Gerritsen, D. F., Lugtigheid, R. A. C., & Walther, T. (2022). Can bitcoin investors profit from predictions by crypto experts? *Finance Research Letters, 46*. https://doi.org/10.1016/j.frl.2021.102266

Hileman, G., & Rauchs, M. (2017). 2017 Global cryptocurrency benchmarking study. *SSRN*. https://dx.doi.org/10.2139/ssrn.2965436

Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Liu, Z., & Zhang, J. (2021). LSTM based sentiment analysis for cryptocurrency prediction. In C. S. Jensen, E.-P. Lim, D.-N. Yang, W.-C. Lee, V. S. Tseng, V. Kalogeraki, J.-W. Huang, C.-Y. Shen (Eds.). *Procededings of the Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021,* 617-621. Springer International Publishing. https://doi.org/10.1007/978-3-030-73200-4_47

Kristoufek, L. (2015). What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0123923

Mittal, A., Dhiman, V., Singh, A., & Prakash, C. (2019). Short-term bitcoin price fluctuation prediction using social media and web search data. *Paper presented at the 2019 Twelfth International Conference on Contemporary Computing (IC3),* 1-6. Noida, India. DOI: 10.1109/IC3.2019.8844899

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Bitcoin*. https://bitcoin.org/bitcoin.pdf

Prajapati, P. (2020). *Predictive analysis of Bitcoin price considering social sentiments*. arXiv. https://doi.org/10.48550/arXiv.2001.10343

Vo, A.-D., Nguyen, Q.-P., & Ock, C.-Y. (2019). Sentiment analysis of news for effective cryptocurrency price prediction. *International Journal of Knowledge Engineering*, *5*(2), 47-52. DOI: 10.18178/ijke.2019.5.2.116

Zhang, L., Wang, S., Liu, B. (2018). Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4). https://doi.org/10.1002/widm.1253

Zhao, J., Liu, K., & Xu, L. (2016). Sentiment analysis: mining opinions, sentiments, and emotions. *Computational Linguistics, 42*(3), 595-598. https://doi.org/10.1162/COLI_r_00259