Science and Technology Publishing (SCI & TECH)
ISSN: 2632-1017
Vol. 8 Issue 6, June - 2024

Machine Learning Based Cryptocurrency Price Prediction Using Financial News

[1]**Ali Pekin**, alipekin58@gmail.com
[2]**İbrahim Enes Ulusoy**, c.enes.eng@gmail.com
[3]**Hidayet Takçı**, Sivas Cumhuriyet University, htahci@cumhuriyet.edu.tr

***Abstract*— The aim of this research is to investigate how daily financial news affects the ability to predict Bitcoin prices with the integration of various machine learning algorithms. To conduct these experiments, daily financial news on cryptocurrencies such as bitcoin and ethereum were collected for two months. Then, a dataset was created from this news data by taking the results of sensitivity analysis such as BERT and Vader. Machine learning models such as K-Nearest Neighbors, Naïve Bayes, Support Vector Machine and Gradient Boosting were created from this dataset. It is noteworthy that the best performing model was CATBOOST, which achieved an impressive accuracy score of 0.85. This study represents a significant contribution to the potential of sentiment analysis in news and the ability to use machine learning models to predict the direction of prices. This development provides stakeholders such as investors and traders with a valuable tool that enables them to make informed decisions.**

Keywords—Cryptocurrency; machine; learning; sentiment; analysis; financial news; price predictions.

I. INTRODUCTION

The potential of coins in the cryptocurrency market to disrupt traditional financial systems and offer an alternative to traditional currencies has attracted a lot of attention recently. As cryptocurrencies are becoming increasingly accepted as a form of payment, there is growing interest in predicting their price movements. Machine learning algorithms and social media data offer a promising approach to predict the market trends of coins. Previous works have shown the potential of using social media data such as Twitter posts to predict stock market trends [1] and Bitcoin market behavior [2]. However, there are few studies on how this financial news can predict the price movements of coins. The aim of this study is to develop a machine learning-based model for predicting the price of coins using market and on-chain data with a focus on daily financial news.

This research is based on sentimental analysis of news to predict market trends, with data from 74 different news sources to predict the price movements of coins. The results of our study will provide valuable information for investors, market participants and exchanges in risk management and decision-making processes. For this research, two coins have been selected for price forecasting. These are Bitcoin and Ethereum, because these coins contain more news data than others. In this research, we review the relevant literature on the use of machine learning and financial news data in predicting the market trends of Bitcoin and Ethereum. Then, machine learning models were created with the sentiment analysis results obtained from the news. Finally, the conclusions and future directions of our research will be discussed.

II. RESEARCH METHODOLOGY

The process followed for the research conducted is given in the table below.

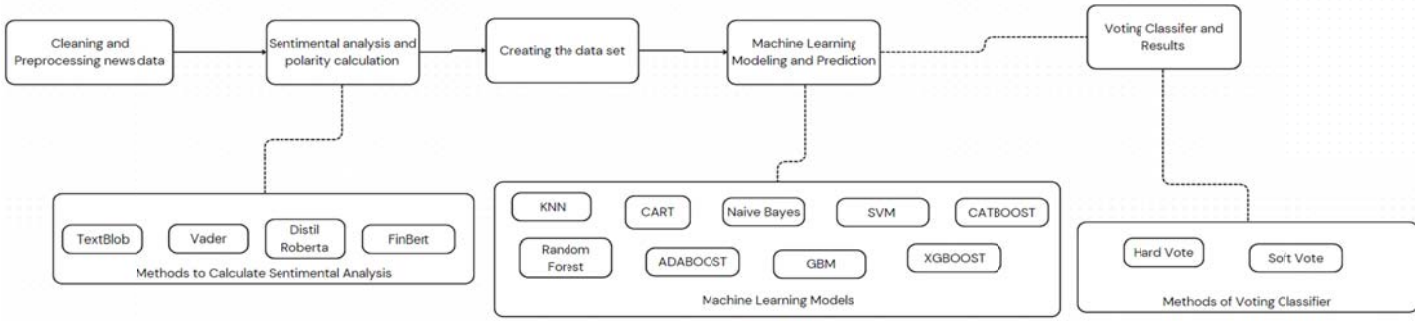

Figure-1. Research Process

In this study, firstly, the news data were subjected to a pre-processing process. Afterwards, a data set was created with the sentiment analysis results obtained from the news data. News data was collected from 74 different online news sources. Using TextBlob, VADAR, Finbert (Hugging Face Model) and DistilRoberta-financial-sentiment (Hugging Face Model), the sentiment of the news was calculated and

saved in the dataset. Nine classification models including KNN, SVM, CART, Gaussian Naive Bayes, Random Forest, ADABOOST, Gradient Boosting Classifier, XGBOOST, CATBOOST were created over the generated dataset. And finally, voting classifier models including these 9 models were created and the results of all the models were obtained and compared.

News data pulled from crypto compare platform via news api.

Cleaning and Preprocessing News Data

The diagram below shows the process of cleaning and pre-processing news data.

Figure-. News Data Pre-Processing

First, punctuation marks in the collected news data were removed from the text using Python loops. Then, the text is decomposed into tokens using the NLTK library, which is widely used in Natural Language Processing (NLP). Then, the text is decomposed into tokens using the NLTK library, which is widely used in Natural Language Processing (NLP).

Once the news was cleaned, we replaced our news data with the cleaned data.

Sentimental Analysis and Polarity Calculation

We calculated the sentiment of the news using TextBlob, VADAR, Finbert (Hugging Face Model) and DistilRoberta-financial-sentiment (Hugging Face Model) on the cleaned data.

Some of the sentimental analysis models we use are available on the hugging face platform and are bert-based.

Hugging face is a major platform in machine learning, with a growing number of users and many models and datasets. The overall aim of Hugging face is to facilitate the use of machine learning processes, to provide users with faster and more ready-to-use tools, and to increase the sharing of open source models and datasets between users[6].

Google BERT (Bidirectional Encoder Representations from Transformers) is a Machine Learning model developed by Google artificial intelligence researchers in 2018 and used in the Google algorithm to process natural languages. BERT is a transformer model that is pre-trained on a large English dataset in a self-supervised manner[7].

One of the models we use for sentimental analysis is the Finbert model and the other is the Distil Roberta Financial Sentiment model. These models are the fine-tuning of the bert model on the hugging face platform, whose success rate has been calculated over thousands of financial data and are widely used by users.

SENTIMENTAL ANALYSIS AND POLARITY CALCULATION

We calculated the sentiment of the news using TextBlob, VADAR, Finbert (Hugging Face Model) and DistilRoberta-financial-sentiment (Hugging Face Model) on the cleaned data.

Some of the sentimental analysis models we use are available on the hugging face platform and are bert-based.

Hugging face is a major platform in machine learning, with a growing number of users and many models and datasets. The overall aim of Hugging face is to facilitate the use of machine learning processes, to provide users with faster and more ready-to-use tools, and to increase the sharing of open source models and datasets between users[6].

Google BERT (Bidirectional Encoder Representations from Transformers) is a Machine Learning model developed by Google artificial intelligence researchers in 2018 and used in the Google algorithm to process natural languages. BERT is a transformer model that is pre-trained on a large English dataset in a self-supervised manner[7].

One of the models we use for sentimental analysis is the Finbert model and the other is the Distil Roberta Financial Sentiment model. These models are the fine-tuning of the bert model on the hugging face platform, whose success rate has been calculated over thousands of financial data and are widely used by users.

I. **TEXTBLOB**

TextBlob is a Python library for Natural Language Processing (NLP). It is based on NLTK (Natural Language Toolkit)[8]. When you give it a sentence, it gives back two outputs: polarity and objectivity.

The polarity score ranges from -1 to 1. A score of -1 means that the words are extremely negative, such as “disgusting” or “horrible”. A score of 1 means that the words are super positive, like “excellent” or “best”.

The subjectivity score ranges from 0 to 1. If it is close to 1, it means that the sentence contains too many personal opinions instead of facts.

In the study, we were more interested in the polarity score and the polarity values of the news data ranging between -1 and 1 were calculated and recorded in the table.

II. VADER

VADER, like TextBlob, is a dictionary-based sentiment analyzer tool with predefined rules for words or dictionaries. However, what sets VADER apart is that it can not only classify words as positive, negative or neutral, but also assess the overall sentiment of a sentence.

The output from VADER is presented in the form of a Python dictionary consisting of four keys: 'neg' for negative, 'neu' for neutral, 'pos' for positive and 'compound'. The composite score is particularly noteworthy as it represents the overall sentiment of the sentence, normalizing the other three scores (negative, neutral and positive) between -1 and +1. Similar to TextBlob, a score of -1 indicates the most negative emotion and +1 indicates the most positive emotion.

In the study, negative 'neg', neutral 'neu' and positive 'pos' values were calculated separately for each news data and recorded in the table.

III. FINBERT

FinBERT is a pre-trained NLP model to analyze sentiment of financial text. It is built by further training the BERT language model in the finance domain, using a large financial corpus and thereby fine-tuning it for financial sentiment classification[9].

The data used to train FinBERT are texts from financial news services and a wide range of financial data from many other sources.

Based on this model, sentimental analysis was performed on the news data and recorded in the table.

DISTILROBERTA FINANCIAL

This model is an NLP model pre-trained to analyze the sentiment of financial texts.

DistilRoBerta was created by further training the language model in finance using a financial corpus, thus fine-tuning it for financial sentiment classification. Financial PhraseBank by Malo and in-house JSL documents and annotations were used for fine-tuning.The Financial PhraseBank dataset consists of 4840 sentences from English financial news categorized by sentiment. These sentences were then annotated by 16 people with a background in finance and business[10].

Based on this model, sentimental analysis was performed on the news data and recorded in the table.

MACHINE LEARNING MODELS

Machine learning models were created over the created data set. 9 different machine learning models were used. Supervised learning algorithms were used since the feature of our dataset to be predicted is labeled.These are KNN, SVM, CART, Gaussian Naive Bayes, Random Forest, ADABOOST, Gradient Boosting Classifier, XGBOOST, CATBOOST algorithms.

The KNN algorithm, also known as the K-Nearest Neighbor algorithm, is one of the most widely known and used machine learning algorithms. Classification is done by using the closeness between a selected feature and the feature closest to it. The K value here is expressed by a number such as 3 or 5 for example[11].

It was decided to use the KNN algorithm in this research because of its advantages such as being a simple but effective algorithm, giving good results when the correct k value is found, and not requiring much cost since the data set is not very large.

Support Vector Machines is a machine learning technique that can be applied independently of the distribution and can provide solutions to classification and regression problems in a supervised or semi-supervised manner[12].

The Support Vector Machine algorithm was decided to be used in this study due to its advantages such as its ability to clearly distinguish between classes, its ability to give successful results with a small number of data in a large number of dimensions, and its robustness against noise.

Decision tree algorithm is one of the tree-based learning algorithms in supervised learning. It can be used for both classification and regression[13].

The decision tree algorithm was decided to be used in this study due to its advantages such as being easy to interpret and being successful in unbalanced data sets.

The naive bayes classification model is a machine learning algorithm that is frequently used in text classification problems. This model is based on Bayes' Theorem and the reason why it is called “naive” (pure, simple) is due to the assumption of independence between features during the classification process. This means that for each feature, the other features that affect the classification process are independent of each other[14].

The Gaussian Naive Bayes algorithm was decided to be used in this study due to its advantages such as being very fast when training the model and giving good results in previous studies.

Random Forests is a supervised learning algorithm. This algorithm is frequently used for both classification and regression. The basis of this algorithm is to evaluate the predictions produced by multiple decision trees by aggregating them[15].

The Random Forest algorithm was decided to be used in this study due to its advantages such as being less overlearning (overfit) and being robust against outliers compared to other algorithms.

Gradient Boosting is a machine learning algorithm used to create a strong learner by combining weak learners together. These weak learners can be, for example, decision trees. GBM builds the next tree by trying to minimize the errors of the previous tree[16].

The Gradient Boositng algorithm was decided to be used in this research because of its advantages over the random forest algorithm, such as being more accurate and capturing complexities well as it tries to improve its continuous error.

XGBoost (eXtreme Gradient Boosting) is a high performance version of the Gradient Boosting algorithm optimized with various adjustments. It was introduced to our lives with the article “XGBoost: A Scalable Tree Boosting System” published by Tianqi Chen and Carlos Guestrin in 2016. The most important features of the algorithm are that it can achieve high predictive power, prevent overlearning, manage empty data and do it fast. According to Tianqi, XGBoost works 10 times faster than other popular algorithms.[17]

The XGBOOST algorithm was decided to be used in this research due to its advantages such as its success in natural language processing projects and its fast speed.

The AdaBoost algorithm is called an ensemble classifier, which represents a strong classifier resulting from the combination of weak classifiers. The general working logic of the model starts with re-running the classifier at each stage by increasing the weight of the incorrect predictions made as a result of the previous stage. With these operations, it is aimed to focus on the incorrect predictions and increase the accuracy rate of the model in classification[18].

The ADABOOST algorithm was decided to be used in this research because of its advantages such as its long-time use in classification problems and its good results.

CatBoostClassifier is a classification algorithm based on the gradient boosting method. CatBoost is a machine learning algorithm known for its direct processing of categorical variables, fast training times, and high performance.[19]

The CatBoost algorithm was decided to be used in this research due to its advantages such as analyzing the categorical variable without the need to convert it to numerical and achieving a high success rate with a decision tree structure using gradient boosting.

Models were created with the 9 machine learning algorithms and datasets mentioned above.

CONCLUSIONS

In this study, accuracy, precision, recall, sensitivity, recall, roc auc score values were run one by one for KNN, CART, NB, RF, XGBOOST, GBM, ADABOOST, CATBOOST and SVM algorithms and a comparison was madeThe values of the algorithms before using K-fold cross validation and grid search are shown in table 1.

Table-1

	KNN	CART	NB	RF	XGBOOST	GBM	ADA	CAT	SVM
Accuracy	0.83	0.78	0.76	0.79	0.79	0.79	0.76	0.79	0.78
Precision	0.81	0.88	0.78	0.78	0.8	0.74	0.78	0.78	0.82
Recall	0.88	0.7	0.78	0.88	0.82	0.95	0.78	0.88	0.75
F1 Score	0.84	0.76	0.75	0.82	0.81	0.83	0.77	0.82	0.78
Roc Auc Score	0.88	0.86	0.85	0.83	0.82	0.84	0.81	0.83	0.87

Looking at the best accuracy value for the best parameters after using cross validation and grid search, the accuracy values of the algorithms shown in table 1 have changed and the new accuracy values are shown in table 2.

Table-2

	KNN	CART	NB	RF	XGBOOST	GBM	ADA	CAT	SVM
Accuracy	0.68	0.72	0.73	0.76	0.73	0.73	0.63	0.78	0.68
Precision	0.68	0.74	0.77	0.74	0.76	0.79	0.69	0.78	0.69
Recall	0.75	0.75	0.75	0.82	0.72	0.7	0.62	0.85	0.75
F1 Score	0.7	0.73	0.72	0.78	0.74	0.73	0.62	0.8	0.71
Roc Auc Score	0.78	0.71	0.83	0.82	0.81	0.8	0.7	0.8	0.81

Performance values after feature engineering are shown in table 3.

Table-3

Table 2 and Table 3 show that there is an increase in all performance values for all algorithms after using cross validation and grid search.

	KNN	CART	NB	RF	XGBOOST	GBM	ADA	CAT	SVM
Accuracy	0.79	0.85	0.79	0.85	0.85	0.85	0.83	0.86	0.82
Precision	0.85	0.87	0.76	0.84	0.86	0.88	0.85	0.85	0.81
Recall	0.78	0.85	0.9	0.88	0.85	0.82	0.82	0.9	0.85
F1 Score	0.79	0.85	0.82	0.86	0.85	0.85	0.84	0.87	0.83
Roc Auc Score	0.89	0.87	0.88	0.91	0.88	0.91	0.89	0.89	0.89

Accordingly, the KNN algorithm gave the highest accuracy value before feature selection. After feature selection, the highest accuracy value was CATBOOST algorithm with 0.86. Some results were obtained by using a voting classifier on top of these models.

The Voting Classifier is a machine learning model that trains on a group of multiple models and predicts an output (class) based on the highest probability of the class selected as the output[20].

In soft voting, the probability values from each classifier are taken, summed and divided by the total number of classifiers. If the average probability value is greater than 0.5, it becomes 1 and if it is less than 0, it becomes 0. In hard voting, labels are taken from each classifier. The most repeated label is the result of hard voting.

Since there are no probability values in hard voting, the roc curve cannot be drawn and no roc auc score is generated.

The voting classifier results can be found in table-4.

Table-4

	Soft Vote	Hard Vote
Accuracy	0.85	0.87
Precision	0.85	0.86
Recall	0.9	0.9
F1 Score	0.87	0.88
Roc Auc Score	0.92	

This research reveals how news affects the cryptocurrency market. After analyzing the news data obtained over a two-month period on a daily basis, models were created with 9 different machine learning algorithms. The most successful of these models was CATBOOST with an accuracy rate of 86%. The hard vote value from the voting classifier was the model with the highest accuracy rate of 87%.

In light of the insights from this research, investors can use the models created by sentiment analysis on news data to learn about the direction of the cryptocurrency market.

This study is based on two months of news data and the forecasts are based on bitcoin and ethereum cryptocurrencies. More successful results can be obtained by increasing the number of data and currencies forecasted.

REFERENCES

[1] Bollen, J., Mao, H. and Zeng, X., 2011. Twitter mood predicts the stock market. Journal of Computational Science, 2(1), pp.1-8.

[2] Kristoufek, L., 2013. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific reports, 3, p.3415.

[3] Abraham, J., Higdon, D., Nelson, J., Ibarra, J., Nelson, J.: Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. SMU Data Science Review, 1(3), (2018). [online] Available: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview.

[4]Tandon,C.(2021,November) How can we predict the impact of the social media messages on the value of cryptocurrency? Insights from big data analytics, Available: https://www.sciencedirect.com/science/article/pii/S2667096821000288

[5] Basilio, J., Toriola, A.: Prediction of Bitcoin Prices Using Deep learning and Sentiment Analysis Based on Bitcoin Tweets, MSc Research Project, School of Computing National College of Ireland, (2021). [online] Available: https://norma.ncirl.ie/5230/1/adebayojosephtoriola.pdf.

[6] Kemal,M.(2023,21 Kasım). Makine öğrenmesinde giderek gelişen Hugging Face nedir? ı. Erişim adresi (27 Mayıs 2024): https://weblo.com.tr/blog/makine-ogrenmesinde-giderek-gelisen-hugging-face-nedir/37

[7] Karataş,E.(2023).Google Bert. Erişim adresi (27 Mayıs 2024): https://www.seolog.com.tr/google-bert/

[8] Deliloglu,T.(2024,Nisan). Python ile NLP : Duygu Analizi. Erişim adresi (27 Mayıs 2024):

https://medium.com/@taner.dll/python-nlp-duygu-analizi-sentiment-analysis-d626cbc5a8d2

[9]Genc,Z.(2020,Haziran)Finbert. Erişim adresi (27 Mayıs 2024):https://huggingface.co/ProsusAI/finbert

[10]Romero,M.(2022) DistilRoberta-financial-sentiment .Erişim adresi (27 Mayıs 2024): https://huggingface.co/mrm8488/distilroberta-finetuned-financial-news-sentiment-analysis

[11] Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., vd. (2016). KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi. Marmara Fen Bilimleri Dergisi, 28(3), 89-94. https://doi.org/10.7240/mufbed.69674

[12] Çomak, E. (2008). Destek vektör makinelerinin etkin eğitimi için yeni yaklaşımlar. Yayınlanmamış Doktora Tezi, Selçuk Üniversitesi, Fen Bilimleri Enstitüsü, Elektrik Elektronik Mühendisliği Anabilim Dalı, Konya.

[13] Atcılı,A.(2022,4 Haziran). Karar Agaclari Algoritmasi. Erişim adresi (27 Mayıs 2024): https://medium.com/machine-learning-t%C3%BCrkiye/karar-agaclari-algoritmasi-b823c23997d0

[14] Bilgili,B.(2023,19 Kasım). Naive Bayes Sınıflandırma. Erişim adresi (27 Mayıs 2024): https://medium.com/@batubilgili1907.bb/naive-bayes-s%C4%B1n%C4%B1fland%C4%B1rma-6dad0795f825

[15] Şenol,O.(2021,7 Mart). Random Forests. Erişim adresi (27 Mayıs 2024):

https://medium.com/yazılım-ve-bilişim-kulübü/random-forests-92fd17d9aa4f

[16] Güler,E.(2023,14 Mart). Gradient Boosting Nedir. Erişim adresi (27 Mayıs 2024):

https://medium.com/gradient-boosting-nedir-2ba518700777

[17] Muratlar,E.(2020,12 Mart). XGBoost Nasıl Çalışır. Erişim adresi (27 Mayıs 2024): https://www.veribilimiokulu.com/xgboost-nasil-calisir/

[18] Akel,U.(2020,12 Nisan). ADABOOST Algoritması. Erişim adresi (27 Mayıs 2024):

https://www.bilisimkitabi.com/adaboost-algoritmasi

[19] Aydinl,M.(2023,21 Haziran). CATBOOST Classifier. Erişim adresi (27 Mayıs 2024):

https://medium.com/@meltem.aydin1875/catboostclassifier-ne-demek-d5656d5f9fd9

[20] Shahane,S.(2021). Voting Classifier. Erişim adresi (27 Mayıs 2024): https://www.kaggle.com/code/saurabhshahane/voting-classifier