Sentiment Analysis on Cryptocurrency Tweets Using Machine Learning

Murugesapandian

Computer Science and Engineering & SMK Fomra Institute Of Technology

---***---

Abstract - Cryptocurrency works similar to standard currency, however, virtual payments are made for goods and services without the intervention of any central authority. Many investors believe in and use Twitter tweets to guide their daily cryptocurrency trading. In this project, we investigated the feasibility of sentiment analysis and emotion for cryptocurrencies. For the study, we targeted (BTC) Bitcoin and collected related data. The data collection, cleaning were essential components of the study. Analysis of the sentiments about cryptocurrency is highly desirable to provide a holistic view of peoples' perceptions. In this regard, this study performs both sentiment analysis and emotion detection using the tweets related to the Bitcoin which are widely used for predicting the market prices of cryptocurrency. For increasing the efficacy of the analysis, a deep learning ensemble model LSTM-GRU is proposed that combines two recurrent neural networks applications. Comparatively, a larger number of people feel happy with the use of cryptocurrency, followed by fear and surprise emotions. The model achieves the highest performance for sentiment analysis with a 0.91 accuracy score and the highest emotion 0.83. Similarly, LSTM-GRU outperforms all other models in terms of correct and wrong predictions for both sentiment analysis 0.99 and emotion detection 0.98.

Key Words: bitcoin, sentiment analysis, machine learning, cryptocurrencies, tweets

1.INTRODUCTION

Cryptocurrency market has been developed at an exceptional pace since its emergence. Cryptocurrency is a digital currency however it is not controlled by any central authority to make online payments. It uses system ledger entries called ‘tokens’ to make online payments for goods and services. Elliptical curve encryption and public-private key pairs are used as cryptographic algorithms. Similarly, hashing functions are utilized to protect online payments and ensure legitimate and unique transactions.

Cryptocurrency comes under many names. You have probably read about some of the most popular types of cryptocurrencies such as Bitcoin, Litecoin, and Ethereum. Cryptocurrencies are increasingly popular alternatives for online payments. Before converting real dollars, euros, pounds, or other traditional currencies into ₿ (the symbol for Bitcoin, the most popular cryptocurrency), you should understand what cryptocurrencies are, what the risks are in using cryptocurrencies, and how to protect your investment.

What is cryptocurrency? A cryptocurrency is a digital currency, which is an alternative form of payment created using encryption algorithms. The use of encryption technologies means that cryptocurrencies function both as a currency and as a virtual accounting system. To use cryptocurrencies, you need a cryptocurrency wallet. These wallets can be software that is a cloud-based service or is stored on your computer or on your mobile device. The wallets are the tool through which you store your encryption keys that confirm your identity and link to your cryptocurrency.

What are the risks to using cryptocurrency? Cryptocurrencies are still relatively new, and the market for these digital currencies is very volatile. Since cryptocurrencies don't need banks or any other third party to regulate them; they tend to be uninsured and are hard to convert into a form of tangible currency (such as US dollars or euros.) In addition, since cryptocurrencies are technology-based intangible assets, they can be hacked like any other intangible technology asset. Finally, since you store your cryptocurrencies in a digital wallet, if you lose your wallet (or access to it or to wallet backups), you have lost your entire cryptocurrency investment.

Follow these tips to Protect your cryptocurrencies:

- Look before you leap! Before investing in a cryptocurrency, be sure you understand how it works, where it can be used, and how to exchange it. Read the webpages for the currency itself (such as Ethereum, Bitcoin or Litecoin) so that you fully understand how it works, and read independent articles on the cryptocurrencies you are considering as well.
- Use a trustworthy wallet. It is going to take some research on your part to choose the right wallet for your needs. If you choose to manage your cryptocurrency wallet with a local application on your computer or mobile device, then you will need to protect this wallet at a level consistent with your investment. Just like you wouldn't carry a million dollars around in a paper bag, don't choose an unknown or lesser- known wallet to protect your cryptocurrency. You want to make sure that you use a trustworthy wallet.
- Have a backup strategy. Think about what happens if your computer or mobile device (or wherever you store your wallet) is lost or stolen or if you don't otherwise have access to it. Without a backup strategy, you will have no way of getting your cryptocurrency back, and you could lose your investment.

Bitcoin was the first blockchain-based cryptocurrency introduced in 2009 and it remains important and leading the market today. In addition to Bitcoin, a large number of cryptocurrencies have been introduced over time, each with its opportunities and functions to provide different features and specifications. Such cryptocurrencies include Bitcoin clones, as well as, entirely new currencies with additional features.

Cryptocurrency investors expect both profit and loss due to ups and downs in the crypto market. For this purpose, many tools are available which can forecast the crypto market and occasionally investors invest based on such forecasts. The rise and fall in the demand for cryptocurrencies are also affected by general public opinion or Governmental policies. In this regard, peoples' sentiments and emotions can help in determining the up and down of cryptocurrency market value, especially, sentiment analysis is trendy nowadays for investment in cryptocurrency [1], [2]. Investors first perform an analysis of peoples' sentiment for a specific currency and then make investments according to the sentiments [3]. Because of that, sentiment analysis on cryptocurrency markets has become a task of great importance [4]. Studies show that tweets containing positive sentiments have a substantial impact on the demand for cryptocurrencies and vice versa [5], [6].

1.2. NEED FOR THE STUDY

Despite the proposal of several sentiment analysis approaches, several challenges require further research efforts. For example, sentiment annotation is challenging when the sentence structure is complex. Often, simple sentences are needed to produce high- accuracy annotations. Similarly, a single approach cannot be generalized and applicable to all the corpus. An approach designed for sentiment analysis in one domain does not necessarily produce good results in another domain. In addition, the role of a specific feature extraction technique cannot be ignored fully. From this perspective, this study is specially designed for predicting people's sentiments and emotions on the cryptocurrency market using supervised machine learning models.

Owing to the wide use of Twitter for expressing opinions and thoughts on specific topics, this study leverages a tweets dataset for this purpose. This study makes the following contributions

- An ensemble model is proposed to perform sentiment analysis with high accuracy. For this purpose, the advantages of long short-term memory (LSTM) and gated recurrent unit (GRU) are combined.
- Sentiment analysis and emotion analysis are performed. TextBlob is used for annotating the sentiments data while emotions are annotated using the LeXmo model. Positive, negative, and neutral sentiments are used while emotions are classified into happy, sad, surprise, angry, and fear.
- The suitability and performance of three feature engineering approaches are studied including term frequency-inverse document frequency (TF-IDF), bag of words (BoW), and Word2Vec. Experiments are performed using several well-known machine learning models such as support vector machine (SVM), logistic regression (LR), Gaussian Naive Bayes (GNB), extra tree classifier (ETC), decision tree (DT), and k nearest neighbor (KNN). Additionally, the performance of LSTM and GRU models is also analyzed.

1.3. LITERATURE SURVEY

M. Hasan et al., (2019) in their work proposes a machine learning approach for the automatic detection of emotions from the text posted on social networks. Emotions are detected by performing the text classification. The study investigates several problems including semantic complexity of text messages, casual style of micro-blogs, multiple sentiments in text, and different states of emotions. Binary classifiers are used to distinguish tweets with emotions and tweets without emotions. Two main tasks of the approach include offline training and online classification task. The developed emotion classification system Emotex can obtain a classification accuracy of 90% for text messages.

S. Sharifirad et al., (2019) predicted emotion detection and emotion intensity degree. For this purpose, natural language processing (NLP) tools are used on sexist tweets which are categorized into indirect harassment, physical harassment, and sexual harassment. Additionally, emotions of anger, joy, sadness and fear are investigated containing low, medium, and high intensity. For multilabel classification SVM, Naïve Bayes (NB), KNN, Multi-layer perceptron (MLP), LSTM, and convolutional neural network (CNN) are used with Word2Vec, global vector (Glove), and FastText for achieving high classification accuracy. In the same manner, the study investigates 3 categories of speech containing sexist remarks to find the intensity of each emotion. Results show that joy feeling and indirect harassment have direct relation and anger is associated with sexual harassment when the intensity is considered. Similarly, anger, joy, and sadness feelings are associated with physical harassment.

The study of F. M. Shah, et al., (2019) conducts experiments to detect tweets' emotions using the AIT-2018 dataset. The authors propose a new model that leverages EmoSenticNet andWordNetAffect for detecting emotions. Results show that the performance is affected by the small dataset and language ambiguity problems. Accuracy is reduced by the text containing multiple emotions.

X. Zhang, et al., (2020) perform sentiment analysis on online social networks using machine learning and lexicon based methods. A multi-label learning algorithm is introduced for this purpose. The proposed approach aims at multiple level emotion detection concerning user view and incorporates machine learning approaches to achieve a multi-label emotion detection system. The authors discover social correlation and temporal correlation, as well as, emotion label correlation. Despite its capability, the proposed approach is limited by the use of a small dataset and low accuracy.

NAILA ASLAM et al., (2022) Predicted that, besides using the machine learning models, two deep learning models are also used for experiments including LTSM and GRU, in addition to the proposed ensemble model. The performance of the proposed model achieves the highest performance for sentiment analysis. The ensemble recurrent structure model LSTM-GRU outperforms with significant performance in comparison with other models including both machine and deep learning models.

2. PROPOSED METHODOLOGY

This research conducts tests on cryptocurrency-related tweets to analyze sentiment and detect emotions. We introduce an ensemble model aimed at enhancing the accuracy of sentiment and emotion classification. The proposed method's architecture is depicted in Figure 1. We executed all tests on an

Intel Core i7 11th generation computer running the Windows OS. We used Python for implementing machine and deep learning models, utilizing the TensorFlow, and scikit-learn frameworks.

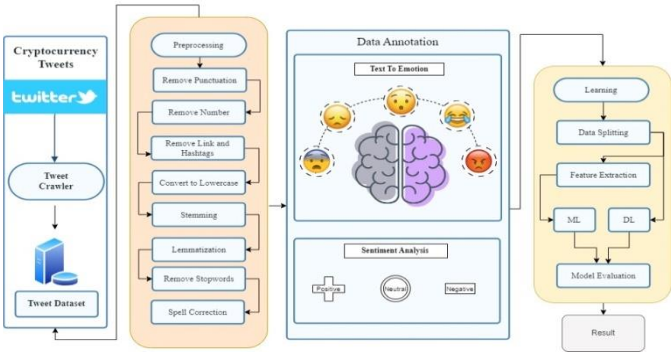

Fig -1: Architecture

Collect the dataset from Kaggle.com . The collected dataset shape is 1048576 rows, 10 columns. Tweets contain links, tags, usernames, numbers, and other characters which are not useful for the machine learning models' training so we can pre-process the data. We will skip and going to take 3576 rows and 3 columns. Sample dataset is shown below

To increase the learning efficiency of machine learning models, preprocessing techniques are used to clean the data. The following steps are carried out for data preprocessing. We remove these meaningless data from tweets using preprocessing techniques such as stemming, lemmatization, spell correction, stop words removal 1. Tokenization 2. Remove punctuation 3. Remove numbers 4. Stemming 5. Lemmatization 6. Stop word removal 7. Spelling correction

After preprocessing dataset, that data will be given to the machine learning algorithm. Machine learning algorithm analyzes the data. Experiments are performed using the machine learning models with BoW, TF-IDF, and Word2Vec features. Experimental results indicate that SVM and LR show more significant results as compared to other models with BoW features, each with a 0.90 accuracy score. This significant performance with both linear models is based on the large feature set. The tweets generate a large feature set that is suitable for the SVM and LR. While RF and DT achieve 0.97 and 0.96 accuracy scores, respectively, KNN and GNB could not perform well. Overall, SVM and LR show a superb performance with a large feature set and multiclass prediction. 1. SVM support vector machine 2. KNN K nearest neighbor 3. GNB Gaussian naive base 4. DT Decision tree 5. LR Logistic regression 6. RF Random forest

SMOTETomek is a hybrid sampling method that combines the oversampling technique called Synthetic Minority Oversampling Technique (SMOTE) and the undersampling technique called Tomek Links. The purpose of SMOTETomek is to address the problem of imbalanced class distribution in machine learning. SMOTE generates synthetic samples for the minority class by interpolating new examples between existing minority class instances. The goal of this technique is to create new synthetic data that can help balance the class distribution. However, SMOTE may produce noisy samples if the minority class is highly overlapping with the majority class. On the other hand, Tomek Links is an undersampling technique that identifies pairs of instances that belong to different classes and are nearest neighbors. These pairs are called Tomek Links, and they can be used to remove the majority class instances that are close to the minority class instances. By doing so, Tomek Links can help improve the decision boundary and reduce the overlap between the two classes. The SMOTETomek algorithm combines these two techniques by first applying SMOTE to oversample the minority class and then applying Tomek Links to undersample the majority class. This can help create a better balance between the classes and reduce the noise that may have been introduced by SMOTE. The result is a dataset with a more balanced class distribution that can be used for machine learning algorithms. One of the advantages of SMOTETomek is that it can be used with any classification algorithm and does not require any modification to the algorithm. However, it is important to note that SMOTETomek may not always improve the performance of the classifier and should be used with caution. Additionally, SMOTETomek can be computationally expensive, especially for large datasets, and may require careful tuning of the hyperparameters.

We are using SMOTETomek Algorithm to balance the dataset. SMOTETomek is a data resampling technique used in machine learning to address the issue of imbalanced datasets. It is a combination of two other techniques, namely, Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links. SMOTE works by creating synthetic samples of the minority class, whereas Tomek Links is used to remove noisy and borderline instances from the majority class. SMOTETomek combines the strengths of both techniques to balance the class distribution by oversampling the minority class and undersampling the majority class. The SMOTE algorithm generates synthetic samples of the minority class by interpolating between existing samples, while Tomek Links identify pairs of instances that are closest to each other from different classes and remove the majority class instance. By applying both techniques together, SMOTETomek helps to remove noise from the majority class and generates synthetic samples of the minority class, resulting in a more balanced dataset. Overall, SMOTETomek is a useful technique for improving the performance of machine learning models when working with imbalanced datasets. Based on this dataset we will use Neural network LSTM and GRU to predict the emotion and sentiment.

Following figure shows a set of results which we got for each ml algorithm

Table: 1 Algorithm Results

MODEL NAME	TYPE	ACCURACY	PRECESSION	RECALL
Random Forest	EMOTION	0.801	0.801	0.801
	SENTIMENT	0.424	0.424	0.424
Decision Tree	EMOTION	0.746	0.746	0.746
	SENTIMENT	**0.912**	**0.912**	**0.912**
SVM	EMOTION	0.827	0.827	0.827
	SENTIMENT	0.885	0.885	0.885
KNN	EMOTION	0.587	0.587	0.587
	SENTIMENT	0.681	0.681	0.681
LR	EMOTION	**0.836**	**0.836**	**0.836**
	SENTIMENT	0.875	0.875	0.875
GNB	EMOTION	0.664	0.664	0.664
	SENTIMENT	0.633	0.633	0.633
ETC	EMOTION	0.798	0.798	0.798
	SENTIMENT	0.907	0.907	0.907

3. CONCLUSIONS

This study performs sentiment analysis and emotion detection on tweets related to cryptocurrency in specific to Bitcoin. Sentiment analysis of cryptocurrency holds potential significance as it is widely used for predicting the market price of the cryptocurrency which necessitates sentiments classification with high accuracy. For experiments, tweets are extracted from Twitter, and the dataset is annotated using TextBlob and LeXmo for sentiments and emotions, respectively. Besides the use of several machine learning and deep learning models for classification, this study leverages recurrent neural networks LSTM and GRU to form an ensemble model to enhance classification performance. In addition, TFIDF, and Word2Vec features are used as feature extraction techniques for the machine learning models. The model achieves the highest performance for sentiment analysis with a 0.91 accuracy score and the highest emotion 0.83. Similarly, LSTM-GRU outperforms all other models in terms of correct and wrong predictions for both sentiment analysis 0.99 and emotion detection 0.98. Dataset balancing implement by the under sampling and oversampling dataset with LSTM-GRU.

ACKNOWLEDGEMENT

I extend heartfelt gratitude to my supervisor, Mr. VIJAI ANAND S K, whose invaluable guidance and unwavering motivation were instrumental in shaping this work. I'm also thankful to Mr. BOOPATHI U for his constructive feedback and evaluations. A collective thanks goes to the entire Computer Science and Engineering faculty for their consistent support and collaboration in ensuring the project's success.

REFERENCES

1. J. Abraham, D. Higdon, J. Nelson, and J. Ibarra, “Cryptocurrency price prediction using tweet volumes and sentiment analysis” SMU Data Sci. Rev., vol. 1, no. 3, p. 1, 2018.
2. S. Colianni, S. Rosales, and M. Signorotti, “Algorithmic trading of cryptocurrency based on Twitter sentiment analysis,” CS229 Project, Stanford Univ., Stanford, CA, USA, Tech. Rep., 2015, pp. 1_5.
3. A.Inamdar, A. Bhagtani, S. Bhatt, and P. M. Shetty, “Predicting cryptocurrency value using sentiment analysis,” in Proc. Int. Conf. Intell. Comput. Control Syst. (ICCS), May 2019, pp. 932_934.
4. D. L. K. Chuen, L. Guo, and Y.Wang, ``Cryptocurrency: A new investment opportunity?'' J. Alternative Investments, vol. 20, no. 3, pp. 16_40, 2017.
5. K. Woak, ``Advanced social media sentiment analysis for short-term cryptocurrency price prediction,'' Expert Syst., vol. 37, no. 2, p. e12493, Apr. 2020.
6. C. Lamon, E. Nielsen, and E. Redondo, ``Cryptocurrency price prediction using news and social media sentiment,'' SMU Data Sci. Rev, vol. 1, no. 3, pp. 1_22, 2017.
7. M. Hasan, E. Rundensteiner, and E. Agu, ``Automatic emotion detection in text streams by analyzing Twitter data,'' Int. J. Data Sci. Anal., vol. 7, no. 1, pp. 35_51, Feb. 2019.
8. S. Shari_rad, B. Jafarpour, and S. Matwin, ``How is your mood when writing sexist tweets? Detecting the emotion type and intensity of emotion using natural language processing techniques,'' 2019, arXiv:1902.03089.
9. F. Calefato, F. Lanubile, and N. Novielli, ``EmoTxt: A toolkit for emotion recognition from text,'' in Proc. 7th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW), Oct. 2017, pp. 79_80.
10. S. A. Salam and R. Gupta, ``Emotion detection and recognition from text using machine learning,'' Int. J. Comput. Sci. Eng., vol. 6, no. 6, pp. 341_345, Jun. 2018.
11. F. M. Shah, A. S. Reyadh, A. I. Shaafi, S. Ahmed, and F. T. Sithil, ``Emotion detection from tweets using AIT-2018 dataset,'' in Proc. 5th Int. Conf. Adv. Electr. Eng. (ICAEE), Sep. 2019, pp. 575580.
12. D. Haryadi and G. Putra, ``Emotion detection in text using nested long short-term memory,'' Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 6, pp. 17, 2019.
13. F. Ghanbari-Adivi and M. Mosleh, ``Text emotion detection in social networks using a novel ensemble classifier based on Parzen tree estimator (TPE),'' Neural Comput. Appl., vol. 31, no. 12, pp. 8971-8983, Dec. 2019.
14. E. Junianto and R. Rachman, ``Implementation of text mining model to emotions detection on social media comments using particle swarm optimization and naive Bayes classifier,'' in Proc. 7th Int. Conf. Cyber IT Service Manage. (CITSM), Nov. 2019, pp. 1- 6.
15. X. Zhang, W. Li, H. Ying, F. Li, S. Tang, and S. Lu, ``Emotion detection in online social networks: A multilabel learning approach,'' IEEE Internet Things J., vol. 7, no. 9, pp. 8133-8143, Sep. 2020.
16. D. Seal, U. K. Roy, and R. Basak, ``Sentence-level emotion detection from text based on semantic rules,'' in Information and Communication Technol- ogy for Sustainable Development. Singapore: Springer, 2020, pp. 423-430.
17. S. M Raju and A. M. Tarif, ``Real-time prediction of BITCOIN price using machine learning techniques and public sentiment analysis,'' 2020, arXiv:2006.14473.
18. X. Huang, W. Zhang, X. Tang, M. Zhang, J. Surbiryala, V. Iosifidis, Z. Liu, and J. Zhang, ``LSTM based sentiment analysis for cryptocurrency prediction,'' 2021, arXiv:2103.14804.
19. E. Sa3maz and F. B. Tek, ``Tweet sentiment analysis for cryptocurrencies,'' in Proc. 6th Int. Conf. Comput. Sci. Eng. (UBMK), Sep. 2021, pp. 613618.
20. T. Mehta, G. Kolase, V. Tekade, R. Sathe, and A. Dhawale, ``Price prediction and analysis of financial markets based on news, social feed, and sentiment index using machine learning and market data,'' Int. Res. J. Eng. Technol., vol. 7, no. 6, 2020. [Online]. Available: https://www.irjet.net/archives/V7/i6/IRJET-V7I688.pdf
21. NAILA ASLAM , FURQAN RUSTAM , ERNESTO LEE , PATRICK BERNARD WASHINGTON, AND IMRAN ASHRAF “Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model “ April 7, 2022.