2024 IEEE 9th International Conference for Convergence in Technology (I2CT)
Pune, India. Apr 5-7, 2024

An Empirical Study of Financial BERT Models for Sentiment Analysis and Cryptocurrency Price Correlation

Harini Anand[1], Arti Arya[2]
[1,2]PES University, Dept. of CSE, Bengaluru, India Email:
[1]harini.anand@gmail.com, [2]artiarya@pes.edu

***Abstract*—Cryptocurrency trading has become a prominent financial domain, with a market capitalization exceeding one trillion USD. As the influence of social media on crypto markets grows, leveraging sentiment analysis (SA) becomes crucial for enhancing trading models. This work showcases the comparative study of advanced sentiment analysis tools, specifically CryptoBERT, FinBERT, VADER, and SenticNet, and their effectiveness in confirming that Bitcoin is the best cryptocurrency. Also, the correlation between sentiment and prices using a pre-trained transformer DistilBERT is fine-tuned to understand how cryptocurrency prices fluctuate in the market. The study results in a correlation coefficient of 0.88.**

***Keywords*—Cryptocurrency, Bitcoin, Sentiment Analysis, DistilBERT, CryptoBERT, FinBERT, Senticnet, VADER Analysis**

I. Introduction

The cryptocurrency market is experiencing an unprecedented boom driven by a confluence of factors that have captured the attention of investors, institutions, and the broader public. The meteoric rise of Bitcoin, the pioneering cryptocurrency, has been a primary catalyst, with its market capitalization, showcasing a remarkable surge in mainstream acceptance. The surge in retail participation, facilitated by user-friendly platforms, has contributed to the democratization of access to digital assets. [1] Moreover, the rapid evolution of blockchain technology beyond cryptocurrencies into various sectors, including supply chain management and healthcare, underscores the broader potential and utility of distributed ledger technology. As these trends converge, the cryptocurrency market continues to thrive, attracting unprecedented capital and attention, reflecting a paradigm shift in how we perceive and interact with financial assets in the digital age.

Cryptocurrency is a prominent kind of virtual currency that makes use of cryptography for security and operates on decentralized networks, most commonly observed to be based on blockchain technology. [2] Here's a brief overview of cryptocurrency and some prominent kinds:

- ***Bitcoin***: Bitcoin is the first and also, the most popular kind of cryptocurrency. It exponentiated the development process for other kinds of cryptocurrencies and is also known as digital gold.
- ***Ethereum***: Ether (ETH) is the native cryptocurrency used to facilitate computation services along with transactions on the Ethereum network.
- ***Litecoin***: Litecoin is known as the silver equivalent to Bitcoin's gold. It provides a different hashing algorithm and quicker transaction confirmation times.
- ***Binance***: BNB is primarily used to pay for transaction fees on the Binance exchanges.

In this paper, we delve into the exploration of advanced sentiment analysis techniques, and through a novel approach, merge elements of natural language processing, machine learning, and financial analysis, as this research contributes to understanding the correlation between social sentiment and the prices of the most optimum cryptocurrency for strategic decision-making in the dynamic and volatile market.

II. Related Work

A. Social Media Sentiment Analysis for Cryptocurrency Market Prediction

The current research work can be categorized into two distinct parts, one being, Feature Extraction in Machine Learning techniques [3] from text sources with a method called "word counting" and the second method being a deep learning method where the corpus text is shown as a sequence of vector embeddings. The first method comes with the disadvantage of not having the ability to represent semantic information that has been derived from a structured order of words in the dataset and the second is a data-deficient approach because it needs a greater number of parameters [4] to learn from. Financial Sentiment analysis differs in both, purpose and the area of domain when compared to normal sentiment analysis techniques. The objective of the financial sentiment analysis technique is to primarily grasp the relation between information represented using texts and the overall financial market trends. One example highlighting this situation would be to make a dictionary of terms related to the finance domain with values assigned to it such as positive or negative. Then the sentiment of the documents is measured by making a summation of the words with a corresponding dictionary key value. Since there is a significant lack of large labeled financial

979-8-3503-9447-4/24/$31.00 ©2024 IEEE

datasets, [5] it is difficult to use neural networks for sentiment analysis effectively.

Although the primary word embedding layers are first initialized with pre-trained values, the remaining model possesses an impending requirement to grasp difficult and abstract relations with labeled data that is smaller in size. [6] A possible solution is to initialize the entire model to the maximum extent with pre-trained values and further fine-tune those values for performing any kind of classification task. SenticNet is a semantic tool that is publically available and is used for concept-level sentiment analysis. [7] Instead of making use of graph-mining as well as dimensionality-reduction techniques, SenticNet 3 uses a concept called "energy flows" in order to join multiple parts of extended knowledge representations. The SenticNet 3 models are contextualized sentics that consist of information related to natural linguistic expressions, that are not single word in nature. [8] and as a result, represent the information of an intermediate nature between neural networks and that of typical symbolic systems. [9] Past studies show that the price prediction of cryptocurrency has been constantly experimented with and explored, and as a result, cryptocurrencies have gained considerable fame. The latest work shows that there is an active effort put in to improve the performance of price prediction methods by utilizing deep learning models. [10]However, most studies have thrown light on predicting cryptocurrency prices for the following day and therefore, the investors face a severe disadvantage in adjusting to the necessity of making complex decisions on actions that support maximizing their profit, such as "Sell", "Buy", and "Wait" at a rapid and constant cycle. Not a lot of studies explore the application and utility of deep learning models to provide recommendations for these activities, and the overall performance metrics of these models are low.

B. Research Gaps

There are lesser findings on comparative comprehensive studies on leveraging social media data and using real-time cryptocurrency data extracted from market values to understand the relationship between social perception and the optimum currency to invest in. There is a gap in exploring more sophisticated approaches for initializing and fine-tuning pre-trained models to enhance the effectiveness of sentiment analysis models in financial contexts.

III. Proposed Approach

As shown in Figure 1, we pre-process the Bitcoin prices data scraped from reliable sources and forums from the year 2018 to 2023 and, perform comparative analysis to maximize the accuracy of the model architectures used and for refining trading models. It aims to solve the growing need for precise and informed decision-making in cryptocurrency trading. In this work, advanced sentiment analysis techniques, specifically employing CryptoBERT [11], FinBERT [12], VADER, and SenticNet [13] are compared to understand the general market perception on social media forums and are correlated to the prices of the most preferred cryptocurrency. The research extends beyond sentiment analysis by introducing a predictive model using a pre-trained Transformer neural network that can determine investor sentiment on the most traded cryptocurrency, Bitcoin, and determine whether the Bitcoin price is likely to move up or down as a result of that sentiment.

In this study, Vader sentiment analysis is used to modify the current collated dataset by adding sentiment labels based on the compound score for performing binary classification and eliminating neutral sentiments that don't have any effect on cryptocurrency prices. It was employed to gauge the sentiment expressed in online discussions over the years and predict Bitcoin's favorability as compared to other cryptocurrencies. The time series data was carefully cropped to align with specific time frames as seen in Figure 2, ensuring a consistent and relevant analysis. To enhance the comparability of sentiment scores across different periods, the data was normalized as seen in Figure 3. We introduced a novel approach by calculating the derivative of both the cryptocurrency values and sentiment scores as seen in Figure 4.

This technique aids in observing correlations between the fluctuations in sentiment scores and changes in cryptocurrency values, providing deeper insights into the dynamic nature of the sentiments expressed and their potential impact on the perceived favorability of Bitcoin in comparison to other cryptocurrencies. After performing VADER Analysis, the sentiment score label column was added to the dataset.

The VADER sentiment score is calculated using the following equation:

$$S_i = P_i - N_i \tag{1}$$

where S_i is the sentiment score for sentence i, P_i is the positive probability , and N_i is the negative probability.

The sentiment analysis model, CryptoBERT, was trained with a max sequence length of 128, and demonstrated its proficiency in classifying sentiments into two categories: "Bearish," and "Bullish."

SenticNet uses subsymbolic models, which contain language models that are auto-regressive as well as kernel methods, to make symbolic representations which facilitate the transformation of natural language into precursor reconstructed form of languages. This conversion process further improves the system's capability to precisely deduce polarity from the given data. The framework comprises a systematic knowledge datahouse with 361,654 words and concepts linked to emotional information.

FinBERT is used primarily for financial sentiment analysis, as it leverages BERT-based architecture, enabling it to capture contextual nuances and semantic meanings within large financial texts. It is employed in our work to assess the sentiment surrounding Bitcoin and seeks to decipher sentiments expressed in textual data, shedding light on public perceptions and market sentiment regarding Bitcoin.

We further use DistilBERT, a pre-trained Transformer model preferred for its 60% faster processing compared to the original pre-trained Transformer, BERT, while retaining 97% of BERT's language understanding capabilities.

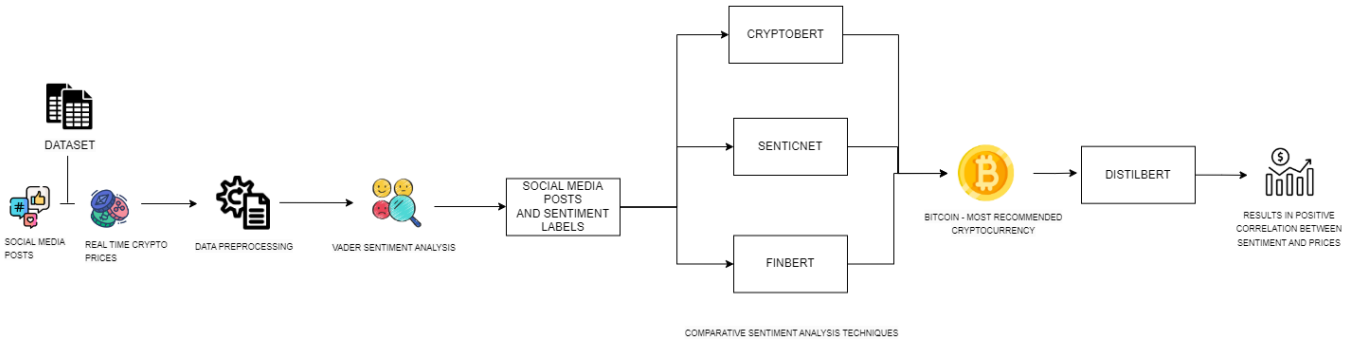

Fig. 1: Architectural Diagram for the Proposed Approach

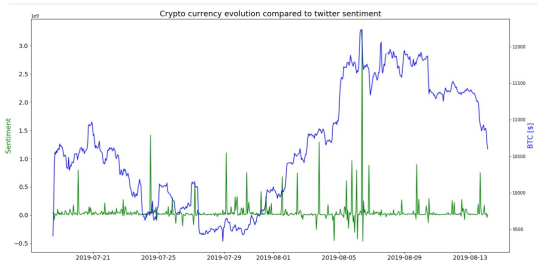

Fig. 2: Time series data cropped to match the time frames

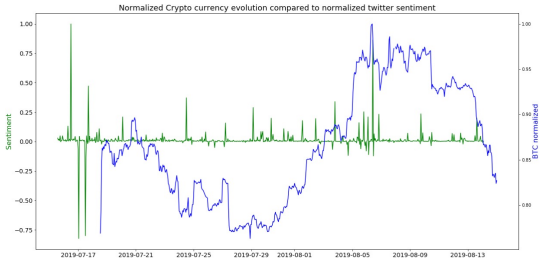

Fig. 3: Normalized Crypto currency evolution compared to normalized Twitter sentiment

The sentiment-labeled social media dataset is utilized to fine-tune the DistilBERT model [16], which subsequently predicts positive or negative sentiment in an additional corpus of unlabeled tweets. The analysis explores the relationship between the sentiment expressed in social media comments and tweets and its potential impact on Bitcoin prices.

IV. Dataset Description

The study utilizes a diverse and extensive dataset for sentiment analysis and cryptocurrency price prediction. The dataset is a combination of labeled social media posts and cryptocurrency price data, collected from various reliable online platforms. [14] Here's a breakdown of the key components:

Social Media Posts

- **Size and Sources:** The dataset consists of 2 million labeled StockTwits posts and 3.2 million social media posts from platforms such as StockTwits, Telegram, Reddit, and Twitter whose sentiment distribution.
- **Date Range:** Posts from StockTwits were collected from November 1, 2018, to June 16, 2023. Telegram data spans from November 16, 2018, to January 30, 2023. Reddit comments were collected from May 2018 to May 2023, and Twitter posts with specific hashtags were collected for May 2018.
- **Preprocessing:** The data has undergone cleaning and preprocessing, involving the removal of cashtags, hashtags, usernames, URLs, crypto wallets, non-English characters, and posts shorter than 4 words. Additionally, the study converted all text to lowercase and addressed spacing and punctuation issues and is represented using a heatmap as seen in Figure 5, where rows represent time intervals, and columns represent different sentiment categories.

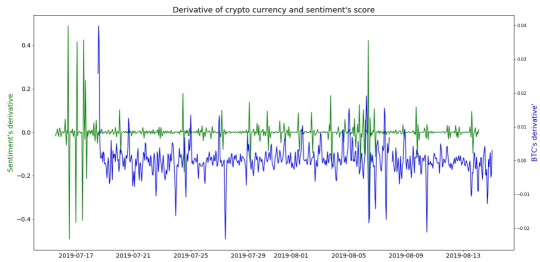

Fig. 4: Derivative of cryptocurrency and sentiment's score

Cryptocurrency Price Data

- **Bitcoin and Data:** The study [15] includes six years of Bitcoin price data (2018/01/01 to 2023/08/27) and five years of Ethereum data (2018/05/09 to 2023/04/27) obtained from CoinAPI with trends depicted using Rolling Statistics as seen in Figure 6.
- **Attributes:** The bitcoin price dataset comprises columns such as Open, High, Low, Close, etc..
- **Daily Updates:** The dataset is updated daily, reflecting real-time changes in cryptocurrency prices as seen in Figure 7.

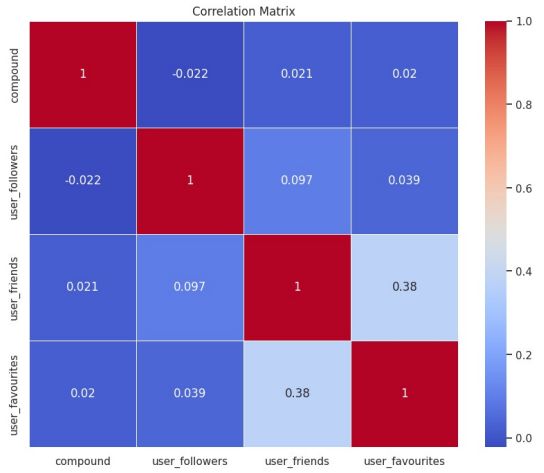

Fig. 5: Correlation matrix to show the intensity of sentiments over time.

Twitter Sentiment Data

- **Data Collection:** Daily sentiment against Bitcoin(BTC) and Ethereum(ETH) [17] is gauged by scraping 10,000 tweets each day from Twitter, covering the date range from 2018/11/01 to 2023/06/27 for BTC(Bitcoin) and 2018/05/09 to 2021/04/27 for ETH(Ethereum).
- **Attributes:** Each scraped tweet includes 'username', 'userlocation', 'userdescription', 'usercreated', 'user-followers', 'userfriends', 'userfavourites', userverified", 'date', 'text', 'hashtags', 'source', 'isretweet

User-Centric Features

- The approach incorporates unique features such as the number of followers per user and the volume of daily trades, providing a broader perspective beyond conventional methods that rely solely on historical price and sentiment data.

In summary, the dataset is comprehensive, spanning various social media platforms, financial domains, and cryptocurrency price data, aiming to enhance sentiment analysis [18] and predictive modeling in the context of cryptocurrency markets.

V. Algorithm for Proposed Approach

The section gives detailed insight into the proposed approach.

A. Input

- Cryptocurrency prices data from reliable sources and forums (2018-2023).
- Social media discussions and comments related to Bitcoin.

B. Output

- Most preferred cryptocurrency
- Predictive model outcomes exploring sentiment and preferred cryptocurrency price movements relation.

C. Steps

1) Pre-process prices data and the social media posts.
2) Utilize Vader sentiment analysis to add sentiment label to the dataset.
3) Apply advanced sentiment analysis techniques (CryptoBERT, FinBERT, VADER, SenticNet).
4) Train CryptoBERT for sentiment analysis.
5) Apply SenticNet for symbolic representations.
6) Utilize FinBERT for financial sentiment analysis.
7) Perform comparative analysis on model architectures and deduce the most preferred cryptocurrency.
8) Use DistilBERT for faster processing and finding the correlation between prices and investor perception of the most preferred cryptocurrency.

This algorithm outlines a comprehensive process for refining trading models and predicting sentiment and price movements in the Bitcoin market.

VI. Overview on BERT Architectures

1) CryptoBERT [11]: CryptoBERT is a model specialized for financial sentiment analysis. Unlike FINBERT, CryptoBERT is pre-trained in the finance domain, utilizing the Financial Phrase Bank dataset.

The model architecture of CryptoBERT extends the BERT language model, specifically tailored for financial sentiment analysis. It undergoes additional training on a large financial corpus, ensuring its proficiency in sentiment classification within the financial domain.

2) FINBERT [12]: It is a sentiment analysis model designed for the cryptocurrency domain. In its pre-training phase, the model is exposed to a diverse dataset comprising 3.2 million social media posts related to various cryptocurrencies, collected from platforms such as StockTwits, Telegram, Reddit, and Twitter, spanning from 2018 to 2023. The cleaning process involves the removal of cashtags, hashtags, usernames, URLs, and crypto-wallets. Non-duplicate posts with a length above 4 words are considered for further analysis.

The model architecture of FINBERT is built upon the BERT language model. It undergoes fine-tuning on a balanced dataset of 2 million labeled StockTwits posts, enabling it to discern sentiments effectively. It supports a maximum sequence length of 128, showcasing its capability to handle sequences of up to 514 tokens.

VII. RESULTS

A. CryptoBERT

The predictions, including labels and scores, as seen in Table 1 are as follows:

TABLE I: Post Sentiment Analysis Results

Post	Sentiment	Accuracy
Unlabelled Post 1	Bullish	86.33%
Unlabelled Post 2	Bearish	92.85%
Unlabelled Post 3	Bullish	68.46%

Neutral sentiment is not considered as it doesn't have any effect on cryptocurrency prices. These results underscore

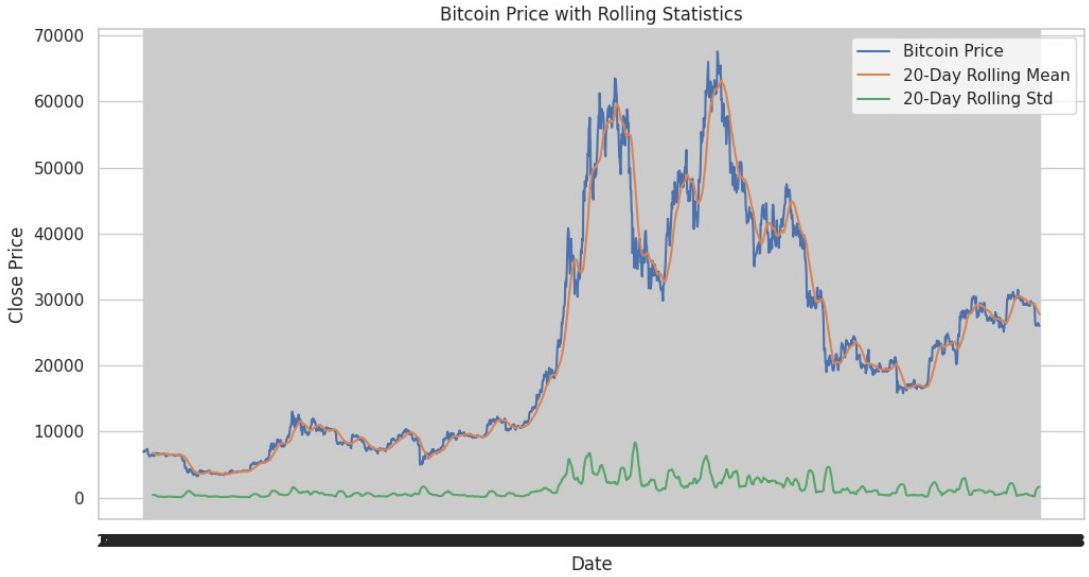

Fig. 6: Rolling mean and standard deviation to visualize trends and volatility from 2018-2023

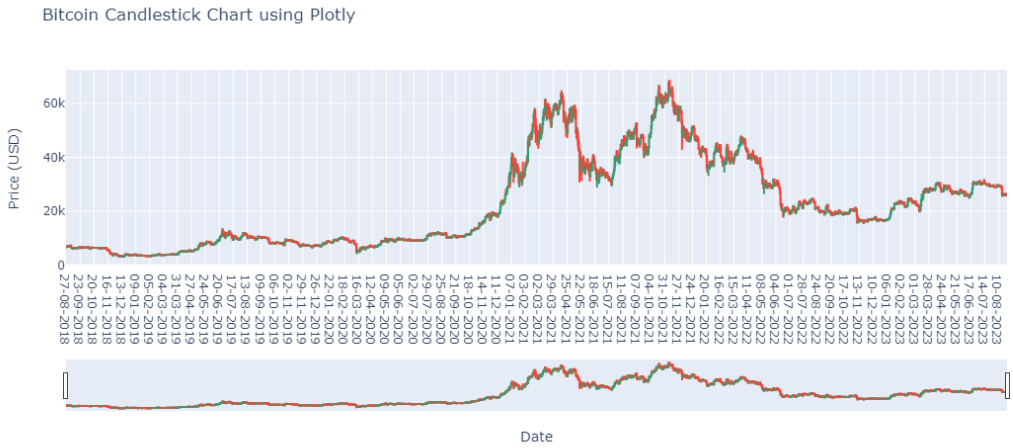

Fig. 7: Candlestick for Price Action Patterns and Volatility Management

CryptoBERT's effectiveness in discerning sentiments, with high confidence scores. The model's ability to provide nuanced predictions contributes to its reliability in capturing the dynamic nature of cryptocurrency-related discussions.

B. SenticNet

In evaluating sentiment analysis performance using accuracy and macro-averaged F1-Score metrics, SenticNet demonstrates a noteworthy performance with an accuracy of 0.76 and an F1 score of 0.59. These results underscore SenticNet's effectiveness in determining sentiment polarity, indicating Bitcoin as the best-performing cryptocurrency in this context.

C. FinBERT

In the analysis of sentiment labels within the dataset,the distribution of sentiment labels among these tweets is as follows: 28.1% were classified as positive, 12.4% as negative, and the majority, 59.4%, were labeled as neutral. This breakdown provides insights into the prevalence of different sentiments expressed in financial-related tweets within the dataset

The model achieved an impressive accuracy of 91% and an F1 Score of 0.89. These performance metrics highlight the effectiveness of FinBERT in accurately classifying financial text data.

Hence, from the three state-of-the-art architecture models, it is clear that Bitcoin is the most preferred cryptocurrency to invest in.

D. DistilBERT

The hyperparameters employed during the training process were carefully chosen to strike a balance between model

convergence and computational efficiency. A learning rate of 2e-05 was selected to guide the optimization process, while both the training and evaluation batch sizes were set to 16, ensuring efficient use of computational resources. The seed, initialized to 42, allowed for reproducibility in the training procedure. The Adam optimizer, configured with betas=(0.9, 0.999) and epsilon=1e-08, facilitated the weight updates during backpropagation. The linear learning rate scheduler, `lr_scheduler_type: linear`: linear, helped modulate the learning rate across epochs. The training spanned two epochs, providing sufficient iterations for the model to learn from the labeled dataset. These hyperparameters collectively contributed to the successful fine-tuning of the DistilBERT model for sentiment analysis on the cryptocurrency dataset. The model observed an accuracy of 0.86 and a loss of 0.460 as seen in Table II. The model found a strong correlation between daily investor sentiment as reflected by the social media posts and the Bitcoin price the following day, with a correlation coefficient of 0.88. In other words, if sentiment was positive, the price would likely rise the next day, conversely, if sentiment was negative, the price would likely fall.

TABLE II: Summary of key metrics obtained during the training of DistilBERT

Training Loss	Epoch	Step	Validation Loss	Accuracy
0.2823	1.0	4109	0.2658	0.8540
0.1905	2.0	10218	0.4060	0.8615

In terms of computational efficiency, SenticNet and FinBERT may introduce additional processing overhead due to semantic analysis and domain-specific training. CryptoBERT, tailored for cryptocurrency discussions, stands out in capturing context-relevant sentiments. The choice among these methods depends on the desired balance between computational efficiency and domain-specific accuracy in sentiment analysis for cryptocurrency-related content. SenticNet excels in handling complex concepts and subtle sentiment variations, providing a deeper understanding of context. On the other hand, VADER proves effective in identifying informal language and slang in tweets, showcasing its proficiency in capturing sentiment polarity efficiently. To make the best use of the inductive biases learned by larger models while getting pretrained, DistilBERT introduces a triple loss which is a summation of language modeling, distillation, as well as cosine-distance losses for exploring the proportional relationship between price values and the investor perception observed on social media. This analysis is instrumental for investors, traders, and analysts seeking timely insights into the workings of the cryptocurrency market, particularly with Bitcoin's pivotal role as a front-runner in the crypto space.

The heatmap as seen in Figure 8, prominently highlights Bitcoin as the standout performer among its counterparts. The chosen metrics, meticulously evaluated across multiple dimensions, consistently position Bitcoin as the leader in terms of market performance, transactional efficiency, and overall market sentiment. This comprehensive visualization

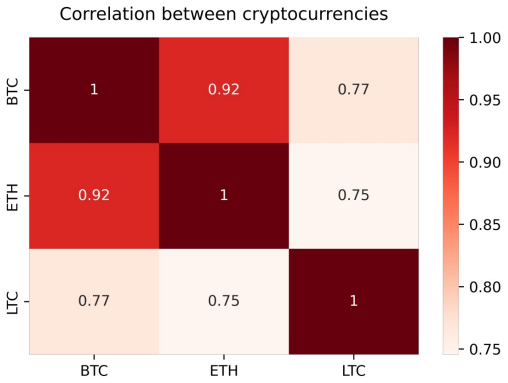

Fig. 8: Heatmap depicting the comparative strength of various cryptocurrencies

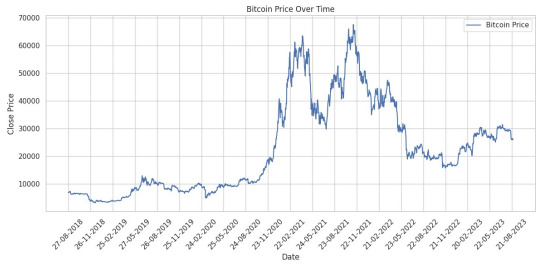

Fig. 9: Price Trend of Bitcoin over the years 2018-2023

provides invaluable insights for investors, signaling Bitcoin's dominance and sustained appeal within the dynamic landscape of the cryptocurrency market.

As seen in Figure 9, the graph shows a stable price trend over the past 5 years helping us conclude that Bitcoin is the most recommended and popular cryptocurrency to be invested in.

VIII. Conclusion

This work presents a comprehensive investigation into the intricate dynamics of cryptocurrency prices, employing a state-of-the-art Cryptocurrency Recommender System. The integration of advanced sentiment analysis techniques, including VADER, SenticNet, FinBERT, and CryptoBERT, enabled a meticulous evaluation of sentiments associated with various cryptocurrencies. Through rigorous comparative analysis, we demonstrated the efficacy of our proposed system, with a particular emphasis on predicting Bitcoin prices. Our findings underscore the significance of sentiment analysis in enhancing predictive models for cryptocurrency markets. As the cryptocurrency landscape continues to evolve, our research adds to the ever-evolving body of information aimed at harnessing the power of sentiment analysis for precise and effective cryptocurrency price predictions.

References

[1] Mukhopadhyay, U., Skjellum, A., Hambolu, O., Oakley, J., Yu, L. and Brooks, R., 2016, December. A brief survey of cryptocurrency systems. In 2016 14th annual conference on privacy, security and trust (PST) (pp. 745-752). IEEE.

[2] Wu, J., Liu, J., Zhao, Y. and Zheng, Z., 2021. Analysis of cryptocurrency transactions from a network perspective: An overview. Journal of Network and Computer Applications, 190, p.103139.

[3] Raheman, A., Kolonin, A., Fridkins, I., Ansari, I. and Vishwas, M., 2022. Social media sentiment analysis for cryptocurrency market prediction. arXiv preprint arXiv:2204.10185.

[4] Wołk, K., 2020. Advanced social media sentiment analysis for short-term cryptocurrency price prediction. Expert Systems, 37(2), p.e12493.

[5] Valencia, F., Gómez-Espinosa, A. and Valdés-Aguirre, B., 2019. Price movement prediction of cryptocurrencies using sentiment analysis and machine learning. Entropy, 21(6), p.589.

[6] Huang, X., Zhang, W., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Liu, Z. and Zhang, J., 2021. Lstm based sentiment analysis for cryptocurrency prediction. In Database Systems for Advanced Applications: 26th International Conference, DASFAA 2021, Taipei, Taiwan, April 11–14, 2021, Proceedings, Part III 26 (pp. 617-621). Springer International Publishing.

[7] Abraham, J., Higdon, D., Nelson, J. and Ibarra, J., 2018. Cryptocurrency price prediction using tweet volumes and sentiment analysis. SMU Data Science Review, 1(3), p.1.

[8] Cambria, E., Speer, R., Havasi, C. and Hussain, A., 2010, November. Senticnet: A publicly available semantic resource for opinion mining. In 2010 AAAI fall symposium series.

[9] Cambria, E., Olsher, D. and Rajagopal, D., 2014, June. SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In Proceedings of the AAAI conference on artificial intelligence (Vol. 28, No. 1).

[10] Parekh, R., Patel, N.P., Thakkar, N., Gupta, R., Tanwar, S., Sharma, G., Davidson, I.E. and Sharma, R., 2022. DL-GuesS: Deep learning and sentiment analysis-based cryptocurrency price prediction. IEEE Access, 10, pp.35398-35409.

[11] Kulakowski, M. and Frasincar, F., 2023. Sentiment Classification of Cryptocurrency-Related Social Media Posts. IEEE Intelligent Systems, 38(4), pp.5-9.

[12] Yang, Y., Uy, M.C.S. and Huang, A., 2020. Finbert: A pre-trained language model for financial communications. arXiv preprint arXiv:2006.08097.

[13] Biagioni, R. and Biagioni, R., 2016. Senticnet. The SenticNet Sentiment Lexicon: Exploring Semantic Richness in Multi-Word Concepts, pp.17-31.

[14] Batra, R. and Daudpota, S.M., 2018, March. Integrating StockTwits with sentiment analysis for better prediction of stock price movement. In 2018 international conference on computing, mathematics and engineering technologies (ICoMET) (pp. 1-5). IEEE.

[15] Vidal-Tomás, D., 2022. Which cryptocurrency data sources should scholars use?. International Review of Financial Analysis, 81, p.102061.

[16] Sanh, V., Debut, L., Chaumond, J. and Wolf, T., 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[17] Mohapatra, S., Ahmed, N. and Alencar, P., 2019, December. KryptoOracle: a real-time cryptocurrency price prediction platform using twitter sentiments. In 2019 IEEE international conference on big data (Big Data) (pp. 5544-5551). IEEE.

[18] https://www.kaggle.com/code/erdeq1024/bitcoin-price-analysis-by-tweets