

**Department of Computer Science**

**MSc Data Science & Analytics**

**Academic Year 2022-2023**

Transformer-based Deep Learning for Finance News Sentiment and Emotion Analysis for Cryptocurrency Portfolios

Azadeh Almasi (2207479)

A report submitted in partial fulfilment of the requirement for the degree of Master of Science

Brunel University

Department of Computer Science

Uxbridge, Middlesex UB8 3PH

United Kingdom

Tel: +44 (0) 1895 203397

Fax: +44 (0) 1895 251686

**ABSTRACT**

*Give a summary of your dissertation. Try to include the following:*

* *A high-level description of the topic area*
* *An overview of the problem studied and why this is interesting/relevant/ important*
* *A high-level description of the approach taken*
* *A summary of the contribution (what did you find/what was the outcome? What is the key implication of your work?)*

*If possible, try not to exceed 500 words. You must not exceed this one page.*

**ACKNOWLEDGEMENTS**

*Who are you grateful to? Your supervisor?*

Sign in the box below to certify that the work carried out is your own. By signing this box you are certifying that your dissertation is free from plagiarism. Make sure that you are fully aware of the Department guidelines on plagiarism (see the student handbook). The penalties if you are caught are severe. All material from other sources must be properly referenced and direct quotes must appear in quotation marks.

I certify that the work presented in the dissertation is my own unless referenced

Signature..........................................

Date.....................................

1

Insert a word count. This is the sum of the words in all the chapters only. The sum should exclude the words in the title page, abstract, acknowledgements, table of contents, references and any appendices.

**TOTAL NUMBER OF WORDS: 12,000**

[*All the above should not exceed this one page]*

# Table of Contents

Table of Contents v

List of Tables vi

List of Figures vii

CHAPTER 1: Introduction 1

*1.1 Introduction* 1

*1.2 Research aim and objectives* 1

*1.3 Research approach* 1

*1.4 Dissertation outline* 1

CHAPTER 2: Literature Review 2

*2.1 Heading* 2

*2.2 Heading* 2

*2.3 Heading* 2

*2.N Summary* 2

CHAPTER 3: Research Approach 3

*3.1 Heading* 3

CHAPTER 4: Data Analysis 4

*4.1 Heading* 4

CHAPTER 5: Discussion 5

*5.1 Heading* 5

CHAPTER 6: CONCLUsion 6

*6.1 Summary of the dissertation* 6

*6.2 Research contributions* 6

*6.3 Future research and development* 6

REFERENCES 7

*APPENDIX A:* ETHICAL *APPROVAL* 8

*APPENDIX B:* OTHER *APPENDICES* 9

**CHAPTER 1: Introduction**

**1.1 Background**

The market for cryptocurrencies has experienced substantial expansion ever since the first Bitcoin was released in 2009 (Farell, 2015), and has seen a remarkable value increase that has surpassed the most significant historical bubbles over the last three centuries (Naeem, Mbarki and Shahzad, 2021). The number of cryptocurrencies has increased from just one in July 2010 to 2419 by February 2020 (Anamika, Chakraborty and Subramaniam, 2023). As a result, both professional and academic researchers have shown a strong desire to comprehend the behaviour of these newly emerging assets (Naeem, Mbarki and Shahzad, 2021). In the present day, Both investors and speculators are increasingly interested in Bitcoin, which is emerging as the prevailing digital currency(Serafini *et al.*, 2020)**.**

It could be difficult to figure out the worth of cryptocurrencies because there is much debate about their nature, such as whether they are a currency, a financial bubble, or simply a digital asset. As a result, there is no agreement on which factors drive cryptocurrency prices. Consequently, cryptocurrency pricing is heavily reliant on widely disseminated opinions, sentiments, and emotions in finance related topics (Naeem, Mbarki and Shahzad, 2021). Furthermore, cryptocurrency prices behave differently than traditional currencies, making it extremely difficult to forecast their prices(Abraham *et al.*, 2018)**.**

In order to predict market movements, several methods have been applied, such as statistical analysis, pattern recognition, machine learning, sentiment analysis, and hybrid approaches. Statistics, as the oldest approach, is employed for data analysis. After that, pattern recognition is a visual strategy that has gained widespread adoption among traders. Recognizing trends and patterns in the stock market's data is required for this. Machine Learning, particularly with the advancements in deep learning tools, has gained significant popularity for predicting time-series data. The introduction of computer-based recognition through machine learning has further amplified the importance of pattern recognition theory. Sentiment Analysis takes a different route by analysing crowd-sourced data. It relies on the principle of "wisdom of crowds," considering the collective opinion of individuals as reliable as that of a single expert. This approach leverages news, current events, public releases, and social media to make market forecasts. In the Hybrid method, a combination of the mentioned approaches is utilized. This comprehensive approach amalgamates statistical analysis, visual pattern recognition, machine learning techniques, and sentiment analysis to enhance market analysis and prediction.(Serafini *et al.*, 2020)

**1.2 Research aim and objectives**

The aim of this study is to develop a hybrid method by integrating a Transformer-based deep learning model with sentiment analysis methodologies to investigates the impact of sentiment and emotions expressed in finance-related news articles on cryptocurrency portfolio performance. Through the application of deep learning techniques and sentiment analysis methodologies, this research seeks to explore the correlation between sentiment and emotions found in online forum comments and the performance of cryptocurrency portfolios. The following objectives will be pursued as part of the research in order to accomplish the stated goals:

1. 1) To study sentiment analysis, deep learning, and cryptocurrency portfolio performance literature. Gain significant insights, develop a basis for the study, and choose the best data science methodology based on the literature.
2. To search for and acquire an open-source dataset that includes cryptocurrency prices over time, as well as sentiment and emotional comments from online forums. This dataset will be used for training and testing the developed model, requiring thorough exploration of diverse sources and repositories.
3. To explore the relationship between sentiments, emotions in online forum comments, and cryptocurrency portfolio performance using a Transformer-based deep learning model. Analyse sentiments and emotions for correlations with portfolio performance.
4. To carry out exhaustive testing, evaluation, and discussion of the research findings. This will expand the field's understanding and provide new research avenues.

**1.3 Research approach**

Data science projects can profit from project management and process methodologies. Such methodologies work as a success factor (Schröer, Kruse and Gómez, 2021).

As CRISP-DM (CRoss Industry Standard Process for Data Mining) is one of the project management and process methods that offers a framework for executing big data projects which is independent of both the industry sector and the technology employed, it is one of the project management and process methodologies. This model is intended to make large data mining initiatives less expensive, more trustworthy, repeatable, easier to organize, and quicker(Wirth and Hipp, 2000). Accordingly, we will use CRISP-DM as the process methodology that includes the following steps: 1) Business Understanding, 2) Data Understanding, 3) Data 4) Preparation, 5) Modelling, 6) Evaluation, and 7) Deployment.

**1.4 Dissertation outline**

* The evaluation of the literature will look at the various models and sentiment analysis methodologies employed in these investigations, as well as their accuracy and potential for boosting the profitability of cryptocurrency portfolios. By combining these insights, we hope to identify crucial insights and knowledge gaps that will help us design our own deep-learning model.
* Afterward through the proposed methodology for this research, we will follow the CRISP-DM process model, which provides a framework for carrying out big data projects. The five steps of CRISP-DM methodology will be clarified.
* In the part on data analysis, our primary focus is on the creation of a deep learning model that is based on the Transformer. This article investigates the methods that was applied in order to analyse the data and construct the model. The process of data analysis initiates with the collection of relevant datasets, continues with the pre-processing of the data, then moves on to the training of the deep learning model, and finally concludes with the evaluation and testing of the model.
* In the discussion section, we will thoroughly analyse and interpret the results gained from the data analysis. This section offers a comprehensive examination and interpretation of the findings, highlighting their relevance within the framework of the research objectives and aims. Moreover, it addresses any limitations or challenges faced during the research and presents an assessment of the implications arising from the findings.
* The conclusion acts as a comprehensive summary of the entire dissertation and serves as the last component of the dissertation. This is accomplished by presenting a summary of the most important findings from the data analysis and the debate, as well as by highlighting the most important contributions made by the research. This part offers a comment not just on the research method but also on its implications for the overarching topic of study. In addition to this, it indicates prospective topics for future research and development, providing paths for further exploration and enhancement.

**Literature Review**

**Literature Review**

**Background and Needs**

**Methods**

In order to forecast market movements, various techniques have been employed, including statistical analysis, pattern recognition, machine learning, sentiment analysis, and hybrid approaches. Statistical analysis, being the oldest method, is used for analysing data. Pattern recognition, which relies on visual strategies, has become widely adopted by traders to identify trends and patterns in stock market data. Machine learning, especially with advancements in deep learning tools, has gained significant popularity for predicting time-series data. The introduction of machine learning-based recognition has further emphasized the significance of pattern recognition theory. Sentiment analysis takes a different approach by analysing crowd-sourced data. It relies on the concept of "wisdom of crowds," considering the collective opinion of individuals as reliable as that of a single expert. This method utilizes news, current events, public releases, and social media to make market forecasts. In the hybrid approach, a combination of the aforementioned methods is utilized. This comprehensive approach integrates statistical analysis, pattern recognition, machine learning techniques, and sentiment analysis to enhance market analysis and prediction (Serafini *et al.*, 2020)**.**

This literature review intends to study the current research on the application of several techniques, such as statistical analysis, pattern recognition, machine learning, sentiment analysis, and hybrid approaches, in two separate parts to investigate the effect of sentiment and emotions in finance-related contexts on the performance of cryptocurrency portfolios.

1. **Sentiment** **Analysis**

Currently, in the present era, characterized by the proliferation of Web 2.0 applications, users are generating vast quantities of data in an expansive and ever-changing manner. In light of this, sentiment analysis has emerged as a significant instrument, enabling the automated extraction of valuable insights from the data generated by users (Habimana *et al.*, 2019). Sentiment analysis is a valuable tool that holds significant potential for businesses, governments, and researchers. It enables the extraction and analysis of public sentiment and opinions, providing valuable insights for informed decision-making and enhancing business strategies (Birjali, Kasri and Beni-Hssane, 2021).

**1.1 Machine Learning Methods**

In a study three different machine learning models have been used: neural networks (NN), support vector machines (SVM), and random forests (RF). They analysed the performance of these models in predicting the price fluctuations of Bitcoin, Ethereum, Ripple, and Litecoin utilizing data from Twitter and market data as input features. Using machine learning and sentiment analysis, they discovered that neural networks outperformed other models in predicting prices (Valencia, Gómez-Espinosa and Valdés-Aguirre, 2019).

In a research Haritha and Sahana (2023) discussed using Twitter as a news source and space for discussing Bitcoin and other cryptocurrencies. They developed an end-to-end model which includes metrics such as "tweet volume", "user following", and "user verification" to predict the price of Bitcoin using historical prices and sentiment of tweets. They applied a neural network model founded on "Bidirectional Encoder Representations from Transformers (BERT)" and a "Gated Recurrent Unit (GRU)" to predict sentiment and prices, respectively. The average MAPE for the sentiment prediction was 9.45%, while the MAPE for the price forecast was 3.6%. “FinBERT”, a language model trained on financial texts, is used for sentiment analysis in this study, enabling contextual embedding and greater precision. Nevertheless, they focused just on Bitcoin, which might limit the generalizability of the recommended model to other cryptocurrencies.

Additionally, researchers examine the difficulties in creating reliable price predictions for cryptocurrencies due to the market's nonlinearity. It suggests employing three types of Recurrent Neural Networks (RNNs) to anticipate Bitcoin, Ethereum, and Litecoin exchange rates. According to the study, the "Bi-Directional LSTM (Bi-LSTM)" compared to LSTM and GRU had better accuracy. The essay also includes an outline of the current monetary system and the advent of blockchain technology and cryptocurrencies as a new asset class in the international financial landscape (Seabe, Moutsinga and Pindza, 2023).

Related to machine learning approach, other research has explored a long short-term memory (LSTM) algorithm for forecasting the values of four different types of cryptocurrencies: "AMP", "Ethereum", "Electro-Optical System", and "XRP". They gathered “CoinMarketCap” data on the daily closing prices of the selected cryptocurrencies from January 1, 2019, to August 14, 2020, and divided it into training and testing sets. The LSTM method was tested using Normalize Root Mean Square Error (NRMSE), Root Mean Square Error (RMSE), and Mean Square Error (MSE) examinations, and the results showed that the LSTM algorithm had superior performance in forecasting all types of cryptocurrencies. The importance of using these models is that they may have significant economic ramifications by assisting investors and traders in recognizing trends in the sales and purchases of various types of cryptocurrencies. The findings of the LSTM model were compared to those of current systems, and the study proved that the proposed model provided greater accuracy based on the proposed system's reduced prediction errors (Ammer and Aldhyani, 2022).

**1.2 Deep Learning Methods**

In recent times, there has been a surge in the proposition of deep learning methodologies for various sentiment analysis undertakings, and these approaches have consistently attained cutting-edge outcomes (Habimana *et al.*, 2019).

Mounika (2021) addresses the growing desire for the digital marketplace and Bitcoin's significance as the most important cryptocurrency for investors. The significant volatility and moves in Bitcoin's price make accurate price forecasting and prediction difficult. For this goal, models of machine learning, particularly deep learning algorithms like Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) have been proposed. The study's goal is to give investors and traders with accurate forecasts and predictions. The results reveal that the suggested system employing the CNN model outperforms other price-prediction models. Overall, the article emphasizes the significance of machine learning in comprehending the complexities and volatility of cryptocurrency markets.

In addition to prior researches, this research covers sentiment analysis and its applicability in analysing user opinions on diverse topics. The researchers used deep learning and word embedding models to estimate the direction of the Bitcoin price by analysing user opinions on social media, notably Twitter. Deep learning architectures including CNNs, RNNs, and LSTMs are employed, as well as word embedding models like Word2Vec, GloVe, and FastText. The evaluations are carried out on an English Twitter dataset, and the findings demonstrate that the “FastText” model, as a word embedding model, outperforms the others in estimating the direction of Bitcoin price with 89.13% accuracy. This research was the first effort to use deep learning and word embedding models to predict Bitcoin price variations (Kilimci, 2020).

1. **Time Series Analysis**

An article in the context of statistical analysis, focuses on predicting the market price of Bitcoin using time series analysis, specifically the Autoregressive Integrated Moving Average (ARIMA) model. The study utilizes four years of Bitcoin data from 2013 to 2017 and aims to achieve a 90% accuracy in predicting the volatility of Bitcoin prices in the short run. The research is motivated by the increasing popularity of Bitcoin as a decentralized cryptocurrency and the interest of both investors and researchers in understanding and predicting its value. They discuss the testing of the models against previously unused data and the calculation of accuracy using the normalized "root mean squared error (RMSE)" method. The study emphasizes two essential characteristics of financial data: price movements in trends and the tendency of history to repeat itself. The research predicts Bitcoin prices for the next ten days based on the chosen model and compares the results with actual prices.

In conclusion, the study proposes a suitable model, ARIMA, for predicting the market price of Bitcoin using time series analysis. The research demonstrates the potential of this approach in forecasting Bitcoin price volatility in the short run, with a claimed accuracy of 90% (Roy, Nanjiba and Chakrabarty, 2018).

A research was conducted on cryptocurrencies like Bitcoin and Ethereum to determine whether public sentiment influences their price. Through an analysis of 24 Reddit communities related to cryptocurrencies, the authors built a set of 112 time-series features from submissions and comments on these subreddits. A Granger causality test is then run on the engineered time series against cryptocurrency price movements, and then the engineered time series are used to estimate cryptocurrency price movements. With only lagged price data and lagged values from a single Reddit data-derived feature, the direction of Bitcoin and Ethereum price movements could be predicted with 74.2% and 73.1% accuracy, respectively (Wooley *et al.*, 2019). Is this a statistical method? Can it be a part of time series analysis part?

As another approach Krysztof Wolk (2020) analysed data from cryptocurrency prices, Twitter sentiments, and Google Trends using predictive and descriptive models. He discusses how social media platforms such as Twitter and Google Trends can be used to monitor public sentiment toward cryptocurrencies and predict price changes. He also examines the correlation across the total number of tweets and data collected from web searches with crypto market prices, as well as the utility of sentiment analysis for forecasting price changes. Least Square Linear Regression (LSLR) and Bayesian Ridge Regression were utilized by the authors. They discovered that using an ensemble learning method was effective for error reduction in a specific model, and they compared the linear regression and ensemble learning methods, discovering that the latter worked better. Because there is a correlation between Twitter data and crypto price movements, the article suggests that sentiment analysis of Twitter and Google Trends can be effective in anticipating crypto fluctuations in the prices (Wołk, 2020).

Examine the correlation between the sentiment analysis results derived from Twitter data, Google Trends data, and the movements in cryptocurrency prices.

While the research has displayed the potential of deep learning models and machine learning models in analysing sentiment and predicting cryptocurrency prices in the context of cryptocurrencies, there is still a gap in the literature when it comes to investigating the effect of sentiment and emotions in finance-related comments on online forums on cryptocurrency portfolio performance using transformer-based deep learning model. The goal of this research is to fill up this gap by developing a transformer-based deep learning model to analyse sentiment and emotions and investigate their impact on cryptocurrency portfolio performance.

**References**

1. Abraham, J. *et al.* (2018) ‘Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis’, 1(3).
2. Anamika, Chakraborty, M. and Subramaniam, S. (2023) ‘Does Sentiment Impact Cryptocurrency?’, *Journal of Behavioral Finance*, 24(2), pp. 202–218. Available at: https://doi.org/10.1080/15427560.2021.1950723.
3. Farell, R. (2015) ‘An Analysis of the Cryptocurrency Industry’.
4. Mariappan, L.T. *et al.* (2023) ‘A Forecasting Approach to Cryptocurrency Price Index Using Reinforcement Learning’, *Applied Sciences*, 13(4), p. 2692. Available at: https://doi.org/10.3390/app13042692.
5. Naeem, M.A., Mbarki, I. and Shahzad, S.J.H. (2021) ‘Predictive role of online investor sentiment for cryptocurrency market: Evidence from happiness and fears’, *International Review of Economics & Finance*, 73, pp. 496–514. Available at: https://doi.org/10.1016/j.iref.2021.01.008.
6. Schröer, C., Kruse, F. and Gómez, J.M. (2021) ‘A Systematic Literature Review on Applying CRISP-DM Process Model’, *Procedia Computer Science*, 181, pp. 526–534. Available at: https://doi.org/10.1016/j.procs.2021.01.199.
7. Serafini, G. *et al.* (2020) ‘Sentiment-Driven Price Prediction of the Bitcoin based on Statistical and Deep Learning Approaches’, in *2020 International Joint Conference on Neural Networks (IJCNN)*. *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom: IEEE, pp. 1–8. Available at: https://doi.org/10.1109/IJCNN48605.2020.9206704.
8. Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: Towards a Standard Process Model for Data Mining’.