

**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]*

# 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. 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

* **Chapter 1:** In the introduction chapter, we will provide an overview of the research objective, which is to develop a transformer-based deep learning model for sentiment analysis in cryptocurrency portfolio performance. We will present background information on the growth of cryptocurrencies and the importance of sentiment analysis in pricing decisions. Additionally, we will outline the specific research aims and objectives, highlighting the exploration of sentiments and emotions in online forums' impact on portfolio outcomes. The chapter will also discuss the adoption of the CRISP-DM process model to ensure a systematic approach throughout the research.
* **Chapter 2:** 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.
* **Chapter 3:** After conducting an extensive review of different methodologies, we have decided to implement the CRISP-DM process model. This well-regarded framework is specifically designed to guide and structure big data projects effectively, ensuring a successful and systematic approach to our endeavour.
* **Chapter 4:** In the part on data analysis, our primary focus is on the creation of a deep learning model for sentiment analysis that is based on the Transformer. The process of data analysis initiates 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.
* **Chapter 5:** 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.
* **Chapter 6:** 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.

# CHAPTER 2: Literature Review

## 2.1 Cryptocurrency in the Literature

Eight years ago, a peer-to-peer network for conducting encrypted digital trade called cryptocurrency was created. While cryptocurrencies are not expected to replace traditional fiat currency, they could transform the way Internet-connected global markets engage with each other, clearing away barriers surrounding normal national currencies and exchange rates. Long-established and unaltered financial payment systems that have been in place for many years are being disrupted by Bitcoin, the first and most well-known cryptocurrency (DeVries, 2016). This has resulted in cryptocurrencies growing quickly and are now common assets on the international financial markets, garnering interest from the media, retail investors, institutional investors, and regulators as well as emerging as a significant and current study topic across a number of academic sectors (Almeida and Gonçalves, 2023).

Cryptocurrencies are made possible by blockchain technology, which supports their decentralised character. To improve security and effectiveness, it integrates cutting-edge cryptography, distributed consensus techniques, and incentives. In comparison to traditional currencies, cryptocurrencies have a number of benefits, including decentralisation, privacy, quick transactions, cheap costs, accessibility worldwide, transparency, and protection from inflation. However, they also have drawbacks like limited acceptance, investment concerns, and limited understanding. These issues can be addressed, encouraging the responsible usage and acceptance of cryptocurrencies while reducing possible risks, through increased knowledge, education, and regulatory measures (Palestine Technical University *et al.*, 2022).

There exists a growing market in the cryptocurrency world, a great deal of potential for cryptocurrency in the financial world, and the fact that cryptocurrency is legal tender in some countries. Furthermore, cryptocurrency has the potential to reach a number of achievable goals in the future, all of which make it crucial to predict cryptocurrency's importance and performance (Ecer, Böyükaslan and Hashemkhani Zolfani, 2022).

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)**.**

## 2.2 Sentiment Analysis

The increasing use of the Internet and social media has had a profound impact on shaping people's opinions in various areas, including social, political, religious, and economic domains. Consequently, there has been a notable surge in research dedicated to opinion mining and sentiment analysis (Almatarneh and Gamallo, 2018). Sentiment analysis has many uses across many different industries. These programs analyze client comments and feelings in order to better understand it, and they also use social media posts from patients to gauge their mental health. The range of applications for sentiment analysis has substantially increased thanks to the development of technologies like Blockchain, IoT, Cloud Computing, and Big Data. Because of these technological developments, sentiment analysis may now be used in practically any industry or discipline to extract insightful information from textual data (Wankhade, Rao and Kulkarni, 2022a).

In research, sentiment analysis refers to the exploration and evaluation of sentiments present in textual data. It includes the processing and analysis of emotions and thoughts expressed in the text (Liaqat *et al.*, 2022). 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). Sentiment analysis has been examined at various levels as explained below.

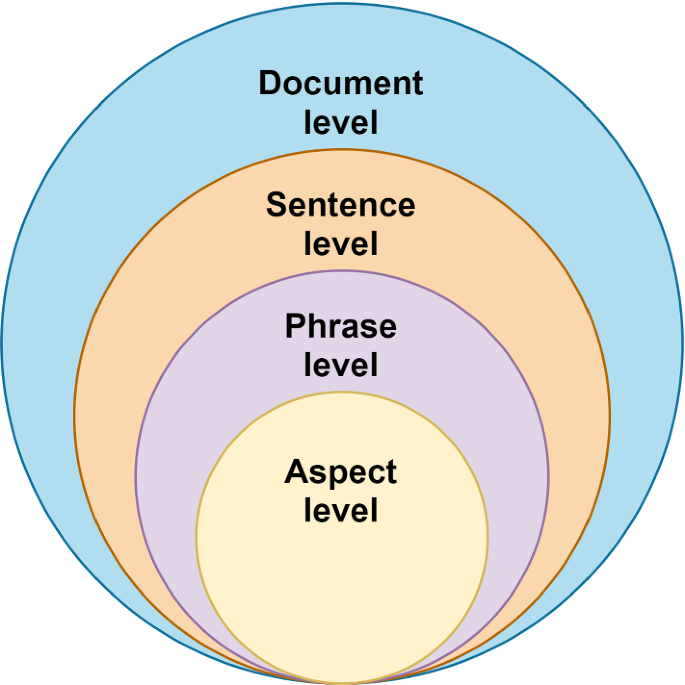
**Document-level**: The analysis of sentiment at the document level involves examining the overall sentiment expressed in a complete document, such as a review or a news article. The objective is to determine whether the sentiment conveyed in the document is positive, negative, or neutral. This form of analysis proves valuable in grasping the general sentiment within a large collection of documents, such as customer reviews for a product or service. Various techniques, including machine learning algorithms, lexicon-based approaches, and rule-based methods, can be employed to perform document level sentiment analysis. However, it is important to acknowledge that this approach has limitations as it may not capture the subtle nuances of sentiment expressed within individual sentences or specific aspects of a product or service (Behdenna, Barigou and Belalem, 2018).

**Sentence level**: At the sentence level, each sentence is analysed individually to determine its corresponding polarity. This level of analysis is particularly valuable when a document contains a diverse range of sentiments. Subjective classification is closely associated with this level of analysis. Similar methodologies used at the document level can be applied, but sentence-level analysis requires more training data and computational resources. The polarity of each sentence can be aggregated to find the sentiment of the overall document or considered individually. In certain cases, document-level sentiment analysis may be insufficient. Previous research on sentence-level analysis has primarily focused on identifying subjective sentences. However, more challenging tasks, such as handling conditional sentences or ambiguous statements, require the importance of sentence-level sentiment analysis (Wankhade, Rao and Kulkarni, 2022b).

**Phrase-level**: Phrase level sentiment analysis involves examining sentiment at a granular level by mining opinion words within individual phrases and classifying them based on their sentiment. Each phrase may express sentiment towards multiple aspects or focus on a single aspect. This approach proves beneficial when analysing product reviews that span multiple lines, as it allows for sentiment evaluation specific to particular aspects expressed within a phrase (Wankhade, Rao and Kulkarni, 2022b).

**Aspect-level**: Aspect-level sentiment analysis involves analysing sentiment at the level of specific aspects or entities mentioned in a sentence. Each sentence can contain multiple aspects, and aspect-level sentiment analysis focuses on assigning polarity to each aspect. The analysis considers all the aspects mentioned in the sentence and calculates an aggregate sentiment for the entire sentence based on the individual polarities assigned to each aspect. Researchers have dedicated significant attention to aspect-level sentiment analysis, with various studies exploring different approaches and techniques for this task (Wankhade, Rao and Kulkarni, 2022b). The availability of detailed sentiment information offers valuable insights for a wide range of applications across different domains. In sentiment analysis, the overall sentiment pertains to the entity under discussion, whereas aspect-level sentiment analysis focuses on the sentiment associated with specific aspects of the entity (Schouten and Frasincar, 2016).

Figure 1, Levels of Sentiment Analysis (Wankhade, Rao and Kulkarni, 2022b).



Sentiment analysis has a wide range of applications that span various domains. These applications include analysing customer opinions to understand their feedback and sentiments, as well as assessing the mental health status of patients based on their social media posts. The advancement of technologies like Blockchain, IoT, Cloud Computing, and Big Data has significantly expanded the scope of applications for sentiment analysis. As a result, sentiment analysis can now be applied in almost any field or discipline, leveraging these technological advancements to gain valuable insights from textual data (Wankhade, Rao and Kulkarni, 2022b).

There are three types of sentiment analysis method: machine learning (ML), lexicon-based methods, and hybrid method (Hamed, Ezzat and Hefny, 2020).

A diagram of a diagram

Description automatically generated

Figure 2, Sentiment Analysis Techniques (Hamed, Ezzat and Hefny, 2020).

Machine learning-based techniques involve the extraction and analysis of sentences and phrases utilising features such as Parts of Speech Marks (POS), n-gram, bi-gram, monogram, and bag-of-words. This method combines three essential methods for categorization of sentences and phrases: Naive Bayes, Support Vector Machines (SVM), and Maximum Entropy. There are three types of machine learning approaches: supervised, semi-supervised, and unsupervised, all of which offer automation and the ability to handle massive amounts of data, making them ideal for sentiment analysis (Hamed, Ezzat and Hefny, 2020). In order to Lexicon-Based methods a predetermined lexicon is necessary. A lexicon is a collection of terms that are specific to a field or language (Bonta, Kumaresh and Janardhan, 2019). Lexicon-based algorithms, classify emotions using decision trees such as k-Nearest Neighbourhood (k-NN), Conditional Random Field (CRF), Hidden Markov Model (HMM), Class Dimensional Classification (SDC), and Chain Optimisation (SMO) (Hamed, Ezzat and Hefny, 2020). Hybrid methods are employed as well to address the limitations of the separate techniques (Bonta, Kumaresh and Janardhan, 2019).

## Lexicon-Based Method

In a study, Almatarneh and Gamallo (2018) aim to highlight the importance of extreme opinions across a variety of fields. The researchers devised an automated method to create a lexicon containing highly negative and positive words from labelled corpora. This lexicon was then integrated into a classifier to detect extreme reviews. The classifier employed a two-step process: identifying documents with extremely negative sentiments and classifying documents with extremely positive sentiments. The classification algorithm relied on a simple word-matching technique for unsupervised sentiment analysis. To determine their effectiveness, they compared automatically generated lexicons with manually crafted ones. The manual lexicons were divided into partitions based on the polarity weight of each word, and separate experiments were conducted for each partition. According to the results, the researchers' vocabulary outperformed SO-CALL and SentiWords (a version of SentiWordNet).

To recognize hate speech in blogs and forums, a classifier was created. Data was gathered from 100 blog posts on different websites with a focus on nationality, religion, and ethnicity. They created a vocabulary based on the subjective and semantic traits connected to hate speech. They developed a hate speech detection classifier with a 70% accuracy rate using this lexicon (Gitari *et al.*, 2015).

Mehmood and Balakrishnan (2020) introduce an improved method for lexicon-based sentiment analysis in social issues. They enhance the technique by incorporating verbs with multi-level grammatical dependencies and improving the General Inquirer sentiment lexicon. To assess the effectiveness of their approach, the researchers compare it to ten online sentiment analysis tools. The results show that their proposed solution not only achieves higher overall accuracy than the online tools but also outperforms them in classifying positive, negative, and neutral sentiments.

## 3.2 Hybrid Method

Mendon *et al.* (2021), have devised a framework for analysing sentiments expressed by Twitter users regarding natural disasters. They have employed a hybrid approach, combining machine learning, statistical modelling, and a lexicon-based approach. Data pre-processing techniques were applied, sentiment classification was performed, and themes within the Twitter discussions were identified. The study examined 243,746 tweets related to the 2018 natural disasters in Kerala, India, and analysed the fluctuation of positive and negative sentiments over time and across locations. The researchers also identified key users who played a critical role in disseminating information during the disaster and demonstrated how polarity indices varied over time. These findings can be valuable for government agencies and other organizations involved in rescue and relief operations.

Based on research conducted by Appel *et al.(*2016), a novel hybrid approach to sentence-level sentiment analysis has been introduced. They utilized natural language processing techniques, an improved sentiment lexicon incorporating SentiWordNet, and fuzzy sets to estimate the polarity and intensity of semantic orientation in sentences. The proposed method was applied to three diverse datasets, and the results were compared with those obtained from Naïve Bayes and Maximum Entropy techniques. The findings demonstrate that the hybrid approach outperforms both Naïve Bayes and Maximum Entropy techniques when the latter are utilized in isolation.

The researchers introduced a novel hybrid approach for aspect-based sentiment analysis of Twitter data. This approach combined rule-based methods with feature selection techniques such as principal component analysis (PCA), latent semantic analysis (LSA), and random projection (RP) in their experiments. They also made available a dataset called Hate Crime Twitter Sentiment (HCTS) for evaluation by the research community. The experimental results demonstrated that the proposed hybrid approach led to improved classification accuracies. Additionally, the Support Vector Machine (SVM) performed well even when different features were used in the classification process. Among the various features, the part-of-speech (POS) tags feature was identified as the most effective for representing the tweets during classification (Zainuddin, Selamat and Ibrahim, 2017).

## 3.3 Machine Learning and Deep Learning Method

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).

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).

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). As an Example 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).

## Transformer-Based Deep Learning Method

In this article conducted by Zhou, Ji and Zhang (2022), a deep-learning model named TransPCNet is introduced, which utilizes transformers to classify sewer defects using 3D point clouds. The researchers conducted a series of experiments on a dataset comprising 827 real and 16,200 synthetic point clouds of sewers. Through their study, they compared the performance of TransPCNet with other advanced methods and observed that TransPCNet exhibited superior accuracy, precision, recall, and F1-score. Furthermore, the researchers assessed the robustness of TransPCNet by evaluating its performance in various scenarios derived from the dataset, consistently achieving excellent results. These findings suggest that TransPCNet holds significant promise for detecting sewer defects and has potential applications in civil and environmental engineering.

In this study, researchers created and assessed MassGenie, a sophisticated deep-learning approach based on transformers, aimed at recognizing small molecules based on their mass spectra. To train the MassGenie model, they utilized a dataset containing more than one million mass spectra, and its performance was subsequently evaluated on multiple benchmark datasets. The findings demonstrated that MassGenie surpassed alternative methods for small molecule identification, showcasing exceptional accuracy and precision. The results indicate that MassGenie could become a valuable asset for identifying small molecules in diverse domains such as drug discovery and metabolomics (Shrivastava *et al.*, 2021).

Using the Transformer architecture, Wu *et al.* (2021) present a novel deep learning model for identifying communication-oriented entities (CEs) from information and communication technology (ICT) patents in the construction industry. They have addressed the key technical hurdles and categories of CEs for recognition, outlined the architecture of their proposed TBNN model, validated the model using training and testing instances, and compared its performance with a baseline model. The findings demonstrate that the performance of the proposed TBNN model surpasses expectations when compared to similar natural language processing (NLP) tasks in previous studies. This model offers an efficient alternative to labour-intensive and time-consuming manual searching. However, it does not capture the relationships between recognized CEs, and its effectiveness decreases when dealing with raw texts containing numerous ambiguous entities.

The study conducted by Hu *et al.* (2023) proposes a deep learning model named VGG-TSwinformer for the early prediction of Alzheimer's disease using brain structural magnetic resonance imaging (sMRI) data. They conduct experiments to assess the model's performance on two tasks: (1) predicting the progression of mild cognitive impairment (MCI) from baseline to 24 months (pMCI vs sMCI), and (2) classifying patients into Alzheimer's disease (AD), MCI, or cognitively normal (CN) categories. The experimental results demonstrate that the VGG-TSwinformer model surpasses several state-of-the-art deep learning models on both tasks, achieving high levels of accuracy, sensitivity, specificity, and area under the curve (AUC) values. The researchers also provide an extensive analysis of the model's architecture and its capacity to capture temporal patterns of brain structural changes associated with disease progression. Overall, the study suggests that the VGG-TSwinformer model exhibits significant potential for early prediction of Alzheimer's disease, offering the potential to enhance the accuracy and efficiency of clinical diagnosis and treatment.

A study carried out by Muhammad *et al.* (2023) investigates the feasibility of employing deep learning models to predict stock market trends and reviews related studies in the field. The scholars created a transformer-based deep learning model designed specifically for forecasting stock prices in the DSE. They integrated time2vec embedding into their model. The outcomes indicated that their model achieved favourable results when applied to DSE data.

## 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).

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).

**Summary**

Although previous research has demonstrated the potential of various techniques, particularly machine learning and deep learning models, in diverse applications such as sentiment analysis for price prediction within the realm of cryptocurrencies, a significant gap remains in the literature regarding the examination of sentiment and emotions in finance-related comments on online forums and their subsequent impact on cryptocurrency portfolio performance, particularly utilizing transformer-based deep learning models. Therefore, the objective of this research is to bridge this gap by developing a transformer-based deep learning model that can effectively analyse sentiment and emotions, subsequently investigating their influence on cryptocurrency portfolio performance.

# CHAPTER 3: Methodology

# CHAPTER 4: Data analysis

**References**

1. Abraham, J. *et al.* (2018) ‘Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis’, 1(3).
2. Almatarneh, S. and Gamallo, P. (2018) ‘A lexicon based method to search for extreme opinions’, *PLOS ONE*. Edited by E. Cambria, 13(5), p. e0197816. Available at: https://doi.org/10.1371/journal.pone.0197816.
3. Almeida, J. and Gonçalves, T.C. (2023) ‘A systematic literature review of investor behavior in the cryptocurrency markets’, *Journal of Behavioral and Experimental Finance*, 37, p. 100785. Available at: https://doi.org/10.1016/j.jbef.2022.100785.
4. Ammer, M.A. and Aldhyani, T.H.H. (2022) ‘Deep Learning Algorithm to Predict Cryptocurrency Fluctuation Prices: Increasing Investment Awareness’, *Electronics*, 11(15), p. 2349. Available at: https://doi.org/10.3390/electronics11152349.
5. 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.
6. Appel, O. *et al.* (2016) ‘A hybrid approach to the sentiment analysis problem at the sentence level’, *Knowledge-Based Systems*, 108, pp. 110–124. Available at: https://doi.org/10.1016/j.knosys.2016.05.040.
7. Behdenna, S., Barigou, F. and Belalem, G. (2018) ‘Document Level Sentiment Analysis: A survey’, *EAI Endorsed Transactions on Context-aware Systems and Applications*, 4(13), p. 154339. Available at: https://doi.org/10.4108/eai.14-3-2018.154339.
8. Birjali, M., Kasri, M. and Beni-Hssane, A. (2021) ‘A comprehensive survey on sentiment analysis: Approaches, challenges and trends’, *Knowledge-Based Systems*, 226, p. 107134. Available at: https://doi.org/10.1016/j.knosys.2021.107134.
9. Bonta, V., Kumaresh, N. and Janardhan, N. (2019) ‘A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis’, *Asian Journal of Computer Science and Technology*, 8(S2), pp. 1–6. Available at: https://doi.org/10.51983/ajcst-2019.8.S2.2037.
10. DeVries, P.D. (2016) ‘An Analysis of Cryptocurrency, Bitcoin, and the Future’, 1(2).
11. Ecer, F., Böyükaslan, A. and Hashemkhani Zolfani, S. (2022) ‘Evaluation of Cryptocurrencies for Investment Decisions in the Era of Industry 4.0: A Borda Count-Based Intuitionistic Fuzzy Set Extensions EDAS-MAIRCA-MARCOS Multi-Criteria Methodology’, *Axioms*, 11(8), p. 404. Available at: https://doi.org/10.3390/axioms11080404.
12. Farell, R. (2015) ‘An Analysis of the Cryptocurrency Industry’.
13. Gitari, N.D. *et al.* (2015) ‘A Lexicon-based Approach for Hate Speech Detection’, *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), pp. 215–230. Available at: https://doi.org/10.14257/ijmue.2015.10.4.21.
14. Habimana, O. *et al.* (2019) ‘Sentiment analysis using deep learning approaches: an overview’, *Science China Information Sciences*, 63(1), p. 111102. Available at: https://doi.org/10.1007/s11432-018-9941-6.
15. Hamed, S., Ezzat, M. and Hefny, H. (2020) ‘A Review of Sentiment Analysis Techniques’, *International Journal of Computer Applications*, 176(37), pp. 20–24. Available at: https://doi.org/10.5120/ijca2020920480.
16. Haritha G, B. and Sahana N, B. (2023) ‘Cryptocurrency Price Prediction using Twitter Sentiment Analysis’, in *Natural Language Processing, Information Retrieval and AI*. *4th International Conference on Natural Language Processing, Information Retrieval and AI*, Academy and Industry Research Collaboration Center (AIRCC), pp. 13–22. Available at: https://doi.org/10.5121/csit.2023.130302.
17. Hu, Z. *et al.* (2023) ‘VGG-TSwinformer: Transformer-based deep learning model for early Alzheimer’s disease prediction’, *Computer Methods and Programs in Biomedicine*, 229, p. 107291. Available at: https://doi.org/10.1016/j.cmpb.2022.107291.
18. Liaqat, M.I. *et al.* (2022) ‘Sentiment analysis techniques, challenges, and opportunities: Urdu language-based analytical study’, *PeerJ Computer Science*, 8, p. e1032. Available at: https://doi.org/10.7717/peerj-cs.1032.
19. Mehmood, Y. and Balakrishnan, V. (2020) ‘An enhanced lexicon-based approach for sentiment analysis: a case study on illegal immigration’, *Online Information Review*, 44(5), pp. 1097–1117. Available at: https://doi.org/10.1108/OIR-10-2018-0295.
20. Mendon, S. *et al.* (2021) ‘A Hybrid Approach of Machine Learning and Lexicons to Sentiment Analysis: Enhanced Insights from Twitter Data of Natural Disasters’, *Information Systems Frontiers*, 23(5), pp. 1145–1168. Available at: https://doi.org/10.1007/s10796-021-10107-x.
21. Mounika, S. (2021) ‘Crypto-Currency Price Prediction using CNN and LSTM Models’, *International Journal for Research in Applied Science and Engineering Technology*, 9(3), pp. 107–114. Available at: https://doi.org/10.22214/ijraset.2021.33191.
22. Muhammad, T. *et al.* (2023) ‘Transformer-Based Deep Learning Model for Stock Price Prediction: A Case Study on Bangladesh Stock Market’, *International Journal of Computational Intelligence and Applications*, p. 2350013. Available at: https://doi.org/10.1142/S146902682350013X.
23. 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.
24. Palestine Technical University *et al.* (2022) ‘Cryptocurrencies Advantages and Disadvantages: A Review’, *International Journal of Applied Sciences and Smart Technologies*, 4(1), pp. 1–20. Available at: https://doi.org/10.24071/ijasst.v4i1.4610.
25. Roy, S., Nanjiba, S. and Chakrabarty, A. (2018) ‘Bitcoin Price Forecasting Using Time Series Analysis’.
26. Schouten, K. and Frasincar, F. (2016) ‘Survey on Aspect-Level Sentiment Analysis’, *IEEE Transactions on Knowledge and Data Engineering*, 28(3), pp. 813–830. Available at: https://doi.org/10.1109/TKDE.2015.2485209.
27. 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.
28. Seabe, P.L., Moutsinga, C.R.B. and Pindza, E. (2023) ‘Forecasting Cryptocurrency Prices Using LSTM, GRU, and Bi-Directional LSTM: A Deep Learning Approach’, *Fractal and Fractional*, 7(2), p. 203. Available at: https://doi.org/10.3390/fractalfract7020203.
29. 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.
30. Shrivastava, A.D. *et al.* (2021) ‘MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra’, *Biomolecules*, 11(12), p. 1793. Available at: https://doi.org/10.3390/biom11121793.
31. 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. Available at: https://doi.org/10.3390/e21060589.
32. Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022a) ‘A survey on sentiment analysis methods, applications, and challenges’, *Artificial Intelligence Review*, 55(7), pp. 5731–5780. Available at: https://doi.org/10.1007/s10462-022-10144-1.
33. Wankhade, M., Rao, A.C.S. and Kulkarni, C. (2022b) ‘A survey on sentiment analysis methods, applications, and challenges’, *Artificial Intelligence Review*, 55(7), pp. 5731–5780. Available at: https://doi.org/10.1007/s10462-022-10144-1.
34. Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: Towards a Standard Process Model for Data Mining’.
35. Wołk, K. (2020) ‘Advanced social media sentiment analysis for short‐term cryptocurrency price prediction’, *Expert Systems*, 37(2). Available at: https://doi.org/10.1111/exsy.12493.
36. Wooley, S. *et al.* (2019) ‘Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment’, in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, Boca Raton, FL, USA: IEEE, pp. 500–505. Available at: https://doi.org/10.1109/ICMLA.2019.00093.
37. Wu, H. *et al.* (2021) ‘A transformer-based deep learning model for recognizing communication-oriented entities from patents of ICT in construction’, *Automation in Construction*, 125, p. 103608. Available at: https://doi.org/10.1016/j.autcon.2021.103608.
38. Zainuddin, N., Selamat, A. and Ibrahim, R. (2017) ‘Hybrid sentiment classification on twitter aspect-based sentiment analysis’, *Applied Intelligence* [Preprint]. Available at: https://doi.org/10.1007/s10489-017-1098-6.
39. Zhou, Y., Ji, A. and Zhang, L. (2022) ‘Sewer defect detection from 3D point clouds using a transformer-based deep learning model’, *Automation in Construction*, 136, p. 104163. Available at: https://doi.org/10.1016/j.autcon.2022.104163.