

**Department of Computer Science**

**MSc Data Science & Analytics**

**Academic Year 2022-2023**

**Transformer-‌‌‌Based Deep Learning for Finance News Sentiment 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

CHAPTER 1: Introduction 1

1.1 Background 1

1.2 Research aim and objectives 2

1.3 Research approach 2

1.4 Dissertation outline 3

CHAPTER 2: Literature Review 4

2.1 Cryptocurrency in the Literature 4

2.2 Sentiment Analysis 5

2.3 Sentiment Analysis Methods 7

2.3.1 Lexicon-Based Method 8

2.3.2 Hybrid Method 8

2.3.3 Machine Learning and Deep Learning Method 9

2.3.4 Transformer-Based Deep Learning Method 10

2.4 Time Series Analysis 14

Summary 15

CHAPTER 3: Methodology 16

3.1 Introduction to Data Mining 16

3.2 KDD Methodology 16

3.3 SEMMA Methodology 17

3.4 CRISP-DM Methodology 18

3.5 Comprehensive comparison 20

CHAPTER 4: Data Analysis 22

4.1 Business Understanding 22

4.2 Data Understanding 23

4.3 Data Preparation 23

4.3.1 Cryptocurrency price dataset preparation steps 23

4.3.2 Text (comments on online forums dataset) preparation steps 23

4.4 Modeling 23

4.4.1 Transformer-based deep learning model 24

4.4.2 Time Series Model 24

4.5 Evaluation 24

4.5.1 Performance Metrics 24

4.5.2 ROC and AUC 25

4.6 Deployment 25

CHAPTER 5: Discussion 26

Comparison with Existing Models 26

CHAPTER 6: Conclusion 26

References 26

# 

# 

# CHAPTER 1: Introduction

**Overview**

The research aims to develop a transformer-based deep learning model for sentiment analysis to investigate the impact of sentiment and emotions expressed in online forum comments on cryptocurrency portfolio performance and explore the correlation between emotions and the performance of cryptocurrency portfolios.

## 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.*, 2018a)**.**

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 model 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 in order to better predict the future movement of cryptocurrency prices. 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, 2021a).

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 research.
* **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

**Overview**

In the initial section of this chapter, we delve into the profound influence of cryptocurrencies on global markets and their correlation with sentiment analysis. We examine the advantages and disadvantages of cryptocurrencies, particularly focusing on the pioneering example of Bitcoin. Furthermore, sentiment analysis is introduced as a potent instrument with versatile applications, enabling us to grasp public opinions comprehensively. The chapter classifies sentiment analysis techniques into lexicon-based, hybrid, and machine learning approaches, showcasing their respective strengths. Moreover, we identify an intriguing research gap in comprehending the impact of sentiment and emotions on cryptocurrency portfolio performance, and we propose harnessing transformer-based deep learning models for conducting in-depth analysis in this area.

## 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, worldwide accessibility, 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 (Qaroush, Zakarneh and Dawabsheh, 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 performance (Ecer, Böyükaslan and Hashemkhani Zolfani, 2022).

In order to forecast Bitcoin market stocks 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 amount 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, 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 (Behdenna, Barigou and Belalem, 2018).

**Sentence level:** When faced with the task of extracting multiple emotions from a single document, employing a sentence-level classification approach proves to be a prudent choice. This method operates on two fundamental assumptions. Firstly, it presupposes that we can accurately identify the subject or entity to which each sentence refers. This knowledge of entities helps us grasp the context and emotions expressed within each sentence. Secondly, it assumes that every sentence carries a distinct and unique opinion or emotional perspective. By adopting this strategy of breaking down the text into individual sentences and understanding the entities involved, we can gain a more comprehensive understanding of the various emotions intertwined within the document's content (Feldman, 2013). This approach evaluates the general polarity of a document by considering both the sentiment polarity of each sentence and its importance to the overall document. The significance of a sentence within the document is determined by its position, making the initial sentence particularly crucial as it typically declares the document's main opinion (Marshan, Kansouzidou and Ioannou, 2021).

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, 2022).

**Aspect-level:** Aspect-level sentiment analysis seeks to identify sentiment-target pairs within a provided text, wherein sentiment is connected to specific aspects of the discussed entity. This approach enables a more in-depth analysis that capitalizes on the abundant information provided by the textual review. The objective of aspect-level sentiment analysis is to discover and consolidate sentiments related to entities mentioned in documents or their specific aspects, thereby generating highly detailed sentiment information with diverse practical applications (Schouten and Frasincar, 2016). Aspect-level classification finds frequent application in tasks involving customer reviews, and its workflow comprises three sequential steps: Identification, Classification, and Aggregation (Vanaja and Belwal, 2018). In the initial phase, the process involves identifying sentiment-target pairs within the text, where the target represents specific aspects of the entity being analysed. Subsequently, during the second step, these pairs are categorized into sentiment classes, such as positive or negative. Finally, in the third step, the sentiments associated with each aspect are aggregated to provide a comprehensive view of the overall sentiment (Schouten and Frasincar, 2016).

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).

Aspect Level

Phrase Level

Sentence Level

Document Level

Figure 2.1. Levels of Sentiment Analysis (Wankhade, Rao and Kulkarni, 2022b).

## 2.3 Sentiment Analysis Methods

Hamed, Ezzat and Hefny **(**2020)categorize sentiment analysis methods into three primary types: lexicon-based methods, the hybrid method, and machine learning - with a particular emphasis on deep learning. Figure 2.2 provides a graphical representation of these primary types of sentiment analysis methods.

Machine Learning

Sentiment Analysis

Lexicon-Based

Hybrid

Supervised Learning

Unsupervised Learning

Semi-Supervised

Learning

Figure 2.2.Sentiment Analysis Methods(Madhoushi, Hamdan and Zainudin, 2015).

## 2.3.1 Lexicon-Based Method

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). In relation to lexicon-based methods, the study conducted by Almatarneh and Gamallo (2018) aims 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. The outcomes of the tests demonstrate that our lexicons are more effective in detecting extreme sentiments compared to two established resources: SO-CALL and SentiWords (a variant of SentiWordNet).

In addition, created Gitari *et al.*(2015) a detection system, utilizing a lexicon built from subjectivity and hate-related attributes, to discern and evaluate hate speech online, emphasizing aspects of race, nationality, and faith. Data was gathered from 100 blog posts on different websites with a focus on race, nationality, and religion. They assessed their hate speech classifier using precision, recall, and F-score metrics against a reference corpus. Precision measured the accuracy of categorization, while recall gauged how many sentences were correctly identified. The F-score combined these values. Their results, detailed in tables, showed the impact of different features on two data sets. When compared to earlier methods for identifying racist content, their lexicon-driven approach showed enhanced precision and recall. The research tested a classifier for hate speech detection using various feature sets. For the FIRST corpus, the best results (73.42% precision, 68.42% recall, and 70.83% F-score) were achieved using a combination of semantic, hate verb, and theme-based features. Similar enhancements were observed for the SECOND corpus with precision at 71.55%, recall at 68.24%, and F-score at 69.85% using the same feature set. When not differentiating between subjective and objective sentences, performance dropped notably across both corpora. Compared to prior methods, their approach displayed improved precision and recall in classifying racist/non-racist content

Furthermore,Mehmood and Balakrishnan (2020) introduced an improved method for lexicon-based sentiment analysis in social issues. The research utilized tweets related to illegal immigration, amassing 694,141 entries over a quarter of a year. Post initial refinement and organization, they set aside 2,500 tweets into two separate collections for assessment. 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.

## 2.3.2 Hybrid Method

In sentiment analysis, a hybrid technique combines the power of machine learning, which discerns patterns from data, with lexicon-based strategies, which utilize established word lists and their corresponding sentiments. The term "hybrid" signifies the blending of these two unique approaches (Hassonah *et al.*, 2020). Hybrid methods are employed as well to address the limitations of the separate techniques (Bonta, Kumaresh and Janardhan, 2019). 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. 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 to pinpoint central communicators and how information was spread. Such insights can assist governmental bodies in refining their rescue and relief strategies. For companies, this methodology aids in promptly addressing feedback, allocating resources effectively, identifying key opinion leaders, and keeping pace with market shifts. In essence, the sentiment analysis model improves decision-making and maximizes resource utilization for different parties involved. 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.

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, two publicly available Twitter datasets referred to as Twitter A and Twitter B, containing tweets with positive and negative sentiments, and the Movie Review Dataset provided by Pang and Lee, used in various Sentiment Analysis experiments and studies. The examined model demonstrated remarkable efficiency, registering an accuracy of 88.02% and a precision of 84.24% when tested on datasets similar to Twitter's. This approach notably surpassed the results of Naïve Bayes (NB) and Maximum Entropy (ME) techniques. Additionally, the hybrid system could discern various intensities in sentiment polarity and, through the application of fuzzy sets, was able to interpret sentences with 'poor' polarity as potentially neutral.

Additionally, using hybrid approach Zainuddin, Selamat and Ibrahim (2017) introduced a method 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. Their data consists of the Hate Crime Twitter Sentiment (HCTS) dataset with 1078 tweets across various hate crime categories, the Stanford Twitter Sentiment (STS) dataset with 353 tweets based on different categories, and the Sanders Twitter Corpus (STC) dataset with 1091 manually classified tweets focusing on four tech companies. The study's findings emphasized the effectiveness of the novel hybrid technique for Twitter aspect-based sentiment analysis. By utilizing association rule mining (ARM) and integrating heuristic combinations in parts-of-speech (POS) patterns, along with tools like the Stanford dependency parser (SDP) for recognizing both overt and concealed sentiment aspects, the approach achieved substantial enhancements. It increased the precision of traditional sentiment classification methods by 76.55%, 71.62%, and 74.24%.

## 2.3.3 Machine Learning and Deep Learning Method

In sentiment analysis, machine learning methodologies employ elements like Parts of Speech and n-grams for textual evaluation. These techniques can be categorized into three primary groups: supervised learning, where the system learns from annotated data; semi-supervised learning, which leverages both annotated and non-annotated data; and unsupervised learning, that detects inherent patterns in datasets without predefined labels (Hamed, Ezzat and Hefny, 2020). Using machine learning for sentiment analysis, the research conducted by Kilimci (2020) 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. As well, Siripurapu et al. (2021) addressed 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.

## 2.3.4 Transformer-Based Deep Learning Method

Transformer-based deep learning is an innovative architecture designed for sequential inputs, utilizing an attention mechanism. It follows an encoder-decoder structure, with the encoder converting input sequences into continuous representations through auto-regressive steps, and the decoder generating output sequences. The encoder and decoder stacks consist of pointwise and fully connected layers, along with a self-attention mechanism. This transformative approach has revolutionized natural language processing (Yadav, Kumar and Chauhan, 2020). In comparison to traditional deep learning models, transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) have gained significant popularity in the field of natural language processing. BERT, being a pre-trained transformer-based language model, surpasses many NLP tasks, including text classification, entity recognition, and question answering, delivering improved accuracy, particularly in sentiment analysis. The appeal of transformer-based models over traditional deep learning models lies in their ability to handle long-range dependencies in sequential data more effectively. Traditional models like RNNs (Recurrent Neural Networks) encounter the vanishing gradient problem when dealing with long sequences, limiting their capacity to capture long-term dependencies. In contrast, transformer-based models exploit self-attention mechanisms to capture dependencies between all positions in a sequence, regardless of their distance, without requiring recurrent connections. This makes them more efficient and successful in processing long sequences, which is crucial for tasks like sentiment analysis, text classification, and language translation. Furthermore, transformer-based models like BERT are pre-trained on vast amounts of data, enabling them to learn general language representations that can be fine-tuned for specific tasks with smaller data, resulting in enhanced performance (Aurpa, Sadik and Ahmed, 2022).

Muhammad et al. (2023) applied a transformer-based deep learning model to predict stock prices in the Bangladesh stock market. The primary objective of the study is to propose a precise model for forecasting the closing price returns of specific stocks. The transformer architecture, renowned for its effectiveness in natural language processing tasks, is chosen due to its capability to handle sequential data and capture long-term relationships. The model's design involves a transformer encoder-decoder structure, which takes historical stock price sequences as input and generates predictions for the next day or week's closing price return. During training, a regression loss function is employed to minimize the disparity between predicted and actual closing price returns. The model is trained using historical stock price datasets from the selected companies. To enhance the model's performance, the study explores the incorporation of supplementary features such as technical indicators and news sentiment scores. While these additions result in some improvements, the core transformer-based architecture remains the key factor responsible for accurate predictions. The study's scope encompasses eight companies listed in the Bangladesh stock market, including AAMRANET and AGRANINS. Both daily and weekly chart data are analysed. In the case of daily chart-based models, predictions rely on the preceding eight days' data, while weekly chart-based models use information from the past eight weeks. The results demonstrate that the proposed transformer-based model achieves remarkable accuracy with low error rates, surpassing alternative methods for stock price prediction. As a result, the study provides valuable insights for investors seeking well-informed decisions in the stock market.

In addition, an improved BERT-based CBRNN model was developed by Tabinda Kokab, Asghar, and Naz (2022) for sentence-level sentiment analysis of social media data. The study's evaluation was based on the SemEval-2017 dataset. To enhance the model's performance, the authors employed a zero-shot algorithm to annotate data and generate semantic and contextual embeddings using the BERT model. Additionally, they utilized a dilated CNN model to extract both local and global sentimental features from embedded features. To capture long-term dependencies within lengthy texts, a Bi-LSTM model was integrated to process word sequences in both forward and backward directions. The proposed model's parameters were optimized using a grid search CV algorithm, and a comparative analysis was conducted against other sentiment analysis models. Notably, the results demonstrated the superior performance of the proposed BERT-based CBRNN model compared to state-of-the-art models on the SemEval-2017 dataset. The evaluation of the model was conducted through an array of key metrics, including Recall, Precision, F-score, Accuracy, the Receiver Operating Characteristic (ROC), and the Area Under the Curve (AUC). Notably, the model's performance was marked by an outstanding F1-score of 0.87, exceeding the top result in the SemEval-2017 competition.

Furthermore, a comprehensive research project was carried out with the aim of forecasting the upward or downward trends in Bitcoin's value by analysing textual data from online sources. This data, collected between February 20, 2017, and April 6, 2019, came from platforms such as Reddit, Bitcointalk, and CryptoCompare and included approximately 2 million comments, headlines, and forum discussions. Along with this, essential daily financial information about Bitcoin was also examined, including its starting and ending prices, highest and lowest values, trading volume, and overall market worth. The research utilized specific sentiment analysis techniques, known as aspect-based sentiment analysis (ABSA), employing models like the Joint Sentiment Topic model (JST) and Topic-Sentiment Latent Dirichlet Allocation (TS-LDA). These models were used to pinpoint both the underlying sentiments and topics within the collected texts. By doing this, the researchers were able to accurately mirror the market's behaviour and focus on categorizing Bitcoin's returns as either positive or negative. Several machine learning methods were explored, all in various settings, and the collected data was divided into two parts: 55.4% reflecting positive directions and 44.6% reflecting negative directions. The research team assessed the outcomes using a standard measurement called Receiver Operating Characteristic Area Under the Curve (ROC AUC), and the results were benchmarked against basic models that only considered financial aspects. The detailed evaluation revealed that this innovative approach enhanced the existing methods for predicting Bitcoin's returns by a minimum of three percentage points. The integration of JST and TS-LDA features contributed to this improvement or, at the very least, maintained the same level of performance as conventional methods. The research findings made it clear that the newly devised method offered a substantial improvement in forecasting the directional shifts in Bitcoin's value. Bitcointalk emerged as the most effective source, delivering a ROC AUC rating of 0.58, likely owing to the detailed and frequent comments. While textual attributes were shown to be beneficial, their effect differed depending on factors like the number of comments and their length. Furthermore, the inclusion of ABSA features in the analysis offered a more subtle appreciation of the sentiments and topics presented in the text. This contributed to a clearer understanding of the data. The study ultimately concluded that employing aspect-based sentiment analysis is a promising advancement in the field of predicting trends in cryptocurrency. It presents an opportunity for not only improving predictive accuracy but also for gaining a more in-depth understanding of the market's complex dynamics (Loginova *et al.*, 2021).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Refs. | Aim | Method | Data | Evaluation Metrics | | Results |
| Almatarneh and Gamallo (2018) | To address the oversight of extreme opinions in sentiment analysis by using an unsupervised approach with a new lexicon to capture the most intense positive and negative expressions. | A lexicon-based approach augmented with machine learning and statistical modelling. | Publicly available website data, product and movie reviews, and lexicons from text corpora including a Stanford University Movie Review Dataset. | Accuracy, Precision, Recall, and F1-Score. | | According to the results, the researchers' vocabulary outperformed SO-CALL and SentiWords (a version of SentiWordNet). |
| Gitari et al. (2015) | To use a specialized lexicon to discern and assess online hate speech, particularly regarding race, nationality, and faith. | An innovative technique combining lexicon-based methods, detection of subjective phrases, and evaluation through a dedicated corpus. | An annotated corpus consisting of 500 labelled paragraphs related to hate speech. | Accuracy, Precision, Recall, and F1-Score. | | The researchers developed a classifier for hate speech detection, with the best results achieved using semantic, hate verb, and theme-based features: 73.42% precision and 68.42% recall for the FIRST corpus, and 71.55% precision and 68.24% recall for the SECOND corpus. |
| Mehmood and  Balakrishnan (2020) | To create an advanced lexicon-driven sentiment analysis approach for societal concerns, with an emphasis on illegal immigration. | An advanced lexicon-driven sentiment analysis technique on Twitter content related to illegal immigration, merging the General Inquirer with layered grammatical structures and the significance of verbs. | 694,141 tweets on illegal immigration from Twitter. | Accuracy, Precision, Recall, and F1-Score. | | The results indicate that their method surpasses online tools in overall accuracy and in categorizing positive, negative, and neutral sentiments. |
| Mendon et al. (2021) | To create a structure for examining emotions on Twitter related to natural calamities, employing a blend of techniques that include machine learning, statistical analysis, and lexicon-based methods. | A hybrid approach, mixing machine learning, statistical modelling, and lexicon-based methods. | 43,746 tweets collected from 8 August 2018 to 23 February 2019 about Kerala's natural disasters. | ---------- | | The research crafted a sentiment analysis model that supports rescue activities and empowers businesses to grasp customer feelings and take timely, informed actions. |
| Appel et al. (2016) | To present a hybrid approach to Sentiment Analysis at the sentence level, demonstrating its enhanced accuracy and precision compared to traditional techniques like Naïve Bayes and Maximum Entropy. | A hybrid approach to Sentiment Analysis, combining natural language processing, sentiment lexicons, and fuzzy sets. | Twitter datasets, and the Movie Review Dataset by Pang and Lee. | Accuracy, Precision, Recall, and F1-Score | | With Twitter-like datasets, the study achieved an accuracy of 88.02% and a precision of 84.24%, an improvement over Naive Bayes and Maximum Entropy. |
| Zainuddin, Selamat and Ibrahim (2017) | To propose and evaluate a new hybrid sentiment classification method for Twitter, using feature selection techniques like PCA, LSA, and RP, to perform finer-grained, aspect-based analysis, and demonstrate improvements in accuracy over existing methods. | A hybrid sentiment classification for Twitter, using principal component analysis (PCA), latent semantic analysis (LSA), and random projection (RP). | Their data includes the Hate Crime Twitter Sentiment (HCTS) dataset with 1078 tweets, the Stanford Twitter Sentiment (STS) dataset with 353 tweets, and the Sanders Twitter Corpus (STC) dataset with 1091 tweets. | | Accuracy, Precision, Recall, and F1-Score | The hybrid sentiment analysis approach showcased significant enhancements in performance, elevating the accuracy of standard sentiment analysis techniques by 76.55%, 71.62%, and 74.24% respectively. |
| Kilimci (2020) | To explore the applicability of sentiment analysis using deep learning and word embedding models in estimating the direction of Bitcoin price. | CNNs, RNNs, and LSTMs, and word embedding models like Word2Vec, GloVe, and FastText. | English Twitter dataset, 17,629 Bitcoin-related tweets. | | Accuracy, Precision, Recall, and F1-Score. | The study found that the FastText word embedding model, combined with deep learning algorithms, achieved the highest accuracy of 89.13% in forecasting Bitcoin price direction. This model outperformed other methods, including LSTM and VADER. |
| Siripurapu *et al.* (2021) | To provide investors and traders with accurate forecasts and predictions for Bitcoin's price. | Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN) | The data was extracted in real time from Quandl.com, covering the period from 2011 to the present | | Mean Absolute Error (MAE) and Loss Values | The study's results showed that the CNN algorithm was more effective in predicting Bitcoin prices, outperforming the LSTM model. |
| Aurpa, Sadik and Ahmed (2022) | To identify and classify abusive comments in the Bangla language on Facebook using transformer-based models like BERT and ELECTRA. | Transformer-based deep neural network models, specifically BERT (Bidirectional Encoder Representations from Transformers) and ELECTRA (Efficiency Learning an Encoder that Classifies Token Replacements Accurately). | A novel dataset comprising 44,001 comments collected from various Facebook posts. | | Accuracy, Precision, Recall, and F1-Score. | Accuracies of 85.00% using the BERT architecture and 84.92% using the ELECTRA architecture. |
| Muhammad et al. (2023) | To propose a precise model for forecasting the closing price returns of specific stocks in the Bangladesh stock market. | A transformer-based deep learning model with an encoder-decoder structure. | Historical stock price datasets from eight companies listed in the Bangladesh stock market, including both daily and weekly chart data. | | RMSE and, MAE | The study's transformer-based models successfully predicted stock prices for selected companies in the Dhaka Stock Exchange (DSE) with satisfactory low error rates, demonstrating promising performance in both daily and weekly predictions. |
| Tabinda Kokab, Asghar, and Naz (2022) | To develop a model that improves sentiment analysis by addressing challenges like noisy data and vocabulary limitations. They propose using a BERT-based Convolution Bi-directional Recurrent Neural Network to analyse social media reviews without losing information. | An improved BERT-based CBRNN model for analysing the sentiment of social media data. | The SemEval-2017 dataset | | Recall, Precision, F1-Score, Accuracy, ROC AUC. | The result of the study was that the proposed BERT-based CBRNN model demonstrated superior performance compared to other state-of-the-art models on the SemEval-2017 dataset. Specifically, the model achieved an impressive F1-score of 0.87, surpassing the best-performing model in the SemEval-2017 competition. |
| Loginova *et al.* (2021) | To improve cryptocurrency price return predictions in addition by extracting joint topical-sentiment features and assessing text subjectivity, they seek to demonstrate that the proposed approach not only enhances interpretability but also boosts predictive performance, offering a new method in cryptocurrency market forecasting. | Aspect-based sentiment analysis models, specifically JST and TS-LDA. | Reddit, Bitcointalk and CryptoCompare. | | ROC, AUC, and Accuracy | The researchers' use of aspect-based sentiment analysis improved Bitcoin returns predictions by at least 3%, advancing cryptocurrency trend forecasting. |

## Time Series Analysis

In the context of statistical analysis the article conducted by Roy, Nanjiba and Chakrabarty (2018) focuses on predicting the market price of Bitcoin using time series analysis, specifically the Autoregressive Integrated Moving Average (ARIMA) model. They utilize 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%.

Wooley et al (2019) investigated the relationship between public sentiment and cryptocurrency prices, utilizing data gathered from a network of 24 Reddit communities associated with Bitcoin, Ethereum, and other cryptocurrencies. Building on prior research that utilized sentiment analysis and graph theory to model public sentiment and identify influencers in online communities, the study authors construct a comprehensive graph containing 599,489 nodes representing users and 7,278,390 edges representing user interactions. By employing sentiment analysis on Reddit data with the VADER tool, the authors create distinct time series to track the frequency of each sentiment label in submission titles, submission bodies, and comments. They engineer a total of 112 different time series to capture changes in the graph structure, content-sentiment over time, and author significance. Using Granger causality tests on these engineered time series, they examine their correlation with cryptocurrency price movements. Subsequently, they employ classification models to forecast cryptocurrency price movements based on the generated time series data. Remarkably, the findings reveal that by solely considering lagged price values and lagged values from a single Reddit data-derived feature, the direction of Bitcoin and Ethereum price movements can be predicted with an accuracy of 74.2% and 73.1%, respectively.

## 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 little research is done to examine the effect 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

**Overview**

In this chapter, we delve into the world of data mining methodologies, exploring three prominent approaches: KDD, SEMMA, and CRISP-DM. Each methodology's fundamental principles, stages, and advantages are thoroughly examined, enabling us to make an informed decision for our research project. Through a comprehensive comparison, we identify CRISP-DM as the ideal choice, given its well-structured approach and wide applicability. We will adopt CRISP-DM to analyse cryptocurrency prices and sentiment using transformer-based deep learning models.

## Introduction to Data Mining

In recent years, data mining has emerged as a crucial tool for various industries, corporations, and businesses. It has revolutionized how we utilize vast amounts of previously untapped data, enabling us to analyse and predict trends and patterns. This has been particularly valuable, as it helps us avoid the risk of overlooking the wealth of valuable information hidden within extensive databases. To make the most of these opportunities, the adoption of appropriate data mining techniques is essential for gaining valuable insights and knowledge (Chen, Yu and Han, 1996). The rise of data mining has led to the introduction of various process models that serve as guiding frameworks for conducting data mining tasks and applications. These models are designed to effectively handle the challenges of working with massive datasets (Shafique and Qaiser, 2014).Therefore, there has been a significant increase in the use of end-to-end data mining approaches, including methodologies like KDD process, SEMMA and, CRISP-DM (Plotnikova, Dumas and Milani, 2020).

In the forthcoming sections, we will thoroughly explore each methodology, clarifying its core principles, stages, and the advantages it offers. Furthermore, we will conduct a comprehensive comparison, assessing the appropriateness and adaptability of each methodology concerning diverse factors.

## KDD Methodology

The Knowledge Discovery Databases (KDD) approach is a method characterized by its iterative and interactive nature (Wright, 1998). Knowledge Discovery in Databases (KDD) involves the systematic extraction of hidden knowledge from databases. To effectively conduct KDD, it is essential to possess relevant prior knowledge and a clear understanding of the application domain and goals (Shafique and Qaiser, 2014). The KDD structured process is illustrated in Figure 3 by Plotnikova, Dumas and Milani (2020). It consists of five stages (Plotnikova, Dumas and Milani, 2020):

**Selection -** This initial stage involves creating a target dataset or focusing on a specific subset of variables or data samples for the discovery.

**Pre-processing -** The target data undergoes a rigorous cleaning and pre-processing phase to ensure consistency and reliability.

**Transformation -** In this stage, data is transformed using various techniques such as dimensionality reduction to enhance its representation.

**Data Mining -** The focus here is on searching for relevant patterns in the data, usually with a predictive objective.

**Interpretation/Evaluation -** The final stage entails interpreting and evaluating the mined patterns to gain meaningful insights and assess their significance and quality (Azevedo and Santos, 2008)

Target Data



Pre-processed

Data

Transformed Data

Pattern

v

v

v

Data

Knowledge

Pre-processing

Transformation

Data Mining

Evaluation

Selection

Figure 3.3. An overview of KDD's five stages (Plotnikova, Dumas and Milani, 2020).

## SEMMA Methodology

SEMMA, an acronym for Sample, Explore, Modify, Model, and Access, is a powerful data mining method pioneered by SAS Institute. It provides a systematic approach for understanding, organizing, and building data mining projects, effectively addressing various business problems and goals. SEMMA's integration with SAS Enterprise Miner makes it a structured framework, housing essential functional tools for efficient data analysis and decision-making (Shafique and Qaiser, 2014). SEMMA represents a comprehensive five-stage process model that encompasses the following stages: sampling, exploring, modifying, modelling, and assessing (Omari Firas, 2023).

**Sampling -** At this stage, a portion of the large dataset is extracted, striking a balance between containing significant information and enabling quick manipulation. It is worth noting that this step is considered optional.

**Exploration -** During this phase, the data is thoroughly explored to identify unexpected trends and anomalies, thereby fostering a deeper understanding, and generating new insights.

**Data Modification -** In this step, variables are created, selected, and transformed to streamline the model selection process, focusing on pertinent aspects of the data.

**Modelling -** At this stage, the data is subjected to modelling, wherein the software automatically searches for a combination of data that can reliably predict the desired outcome.

**Assessment -** The final stage involves evaluating the data mining process's findings, and determining their usefulness, reliability, and performance (Azevedo and Santos, 2008).

## CRISP-DM Methodology

Recent years have seen significant growth and development in the field of Data Mining. Various initiatives are underway to establish standardized practices within this domain. In the pursuit of establishing standards for the field, CRISP-DM (CRoss Industry Standard Process for Data Mining) emerges as a prominent example. This framework has evolved into an industry-standard approach, outlining a series of sequential steps aimed at providing guidance for the implementation of data mining applications (Azevedo and Santos, 2009). CRISP-DM is widely recognized as the prevailing standard and an industry-neutral process model utilized for implementing data mining initiatives (Schröer, Kruse and Gómez, 2021b).

Using a structured approach, this model guarantees project quality and reduces the need for specialized expertise, resulting in cost and time savings, while also making large data mining initiatives more reliable, repeatable, easier to organize, and quicker (Wirth and Hipp, 2000). The methodology's adaptability to various applications and environments and its independence from specific tools and techniques enhance its appeal. Furthermore, it supports knowledge transfer and training, making it an ideal framework for developing transformer-based deep learning models to analyse sentiment (Chapman et al., 1999). CRISP-DM is highly regarded as an effective methodology for data science projects due to its clear and intuitive steps, fostering a shared understanding within the team. As data scientists naturally align with a CRISP-DM-like process, its adoption requires minimal training, role adjustments, or conflict within the organization. The iterative nature of data science is well-supported by CRISP-DM, allowing for a flexible implementation that encompasses many of the advantages of agile principles and practices (Saltz, 2021). According to Figure 3, CRISP-DM comprises six major phases of iteration (Palacios *et al.*, 2017). For our research, we will adopt CRISP-DM as the process methodology, incorporating the following essential steps:

**Diagram

Description automatically generated**

Figure 3.4. CRISP-DP Process (Palacios *et al.*, 2017).

**Business Understanding -** The Business Understanding phase within CRISP-DM involves establishing the project's objectives, comprehending the customer's requirements, and formulating a detailed data mining plan. This phase lays a robust foundation for the project and ensures that it aligns with the business objectives (Saltz, 2021).

**Data Understanding -** In the Data Understanding phase of CRISP-DM, the process commences with the collection of the initial dataset. Subsequently, various activities are undertaken to develop a comprehensive understanding of the data, pinpoint potential data quality issues, and extract valuable insights or intriguing subsets, which can serve as the basis for further exploration and the formulation of hypotheses (Azevedo and Santos, 2008).

**Data Preparation -** In CRISP-DM's Data Preparation phase, raw data is converted for analysis through cleaning, integration, and selecting relevant elements, ensuring it's ready for modelling and evaluation (Shafique and Qaiser, 2014).

**Modelling -** The fourth phase of the CRISP-DM process involves selecting and applying diverse modelling techniques. Multiple models are constructed, and various parameters are configured to address the same data mining problem.

**Evaluation -** In this phase, a comprehensive assessment of the model(s) is conducted, and the steps involved in its construction are reviewed to ensure alignment with the business objectives.

**Deployment -** The project doesn't conclude merely with the creation of the model. Even if the model's purpose is to enhance data knowledge, the acquired insights must be organized and presented in a manner that is usable for the customer (Saltz, 2021).

## Comprehensive comparison

When comparing the KDD, CRISP-DM, and SEMMA methodologies, we find interesting connections. At first glance, KDD and SEMMA stages match up: Sample is like Selection, Explore is like Preprocessing, Modify is like Transformation, Model is like Data Mining, and Assess is like Interpretation/Evaluation. Digging deeper, we see that SEMMA's five stages practically implement the KDD process, with a direct link to SAS Enterprise Miner. Contrasting KDD with CRISP-DM is a bit more complex. CRISP-DM includes steps before and after KDD: the Business Understanding phase, which aligns with understanding the application domain and user goals, and the Deployment phase, which consolidates knowledge into the system. To sum it up, Data Understanding in CRISP-DM combines Selection and Preprocessing, Data Preparation aligns with Transformation, Modelling with Data Mining, and Evaluation with Interpretation/Evaluation. For a quick summary, check out Table 1 (Azevedo and Santos, 2008).

After thoroughly comparing the KDD, CRISP-DM, and SEMMA methodologies, we have decided to use CRISP-DM for our research. CRISP-DM offers a comprehensive and well-structured approach that covers all the essential phases, from understanding the business objectives to modelling and evaluation. Unlike SEMMA, which is more tailored to SAS Enterprise Miner, CRISP-DM provides greater flexibility and broader applicability. It ensures a holistic understanding of the application domain through its Data Understanding and Data Preparation phases, which are crucial for our research.

CRISP-DM's alignment with industry best practices and its well-defined stages make it easier for us to manage and execute the data mining project effectively. By leveraging the Data Understanding, Data Preparation, Modelling, and Evaluation phases, we can efficiently handle large datasets and extract valuable insights to address our research objectives successfully. Overall, CRISP-DM's systematic approach, adaptability, and guidance through each essential step in the data mining process have made it the optimal choice for our research project.

|  |  |  |
| --- | --- | --- |
| KDD | SEMMA | CRISP-DM |
| Pre KDD | ------------- | Business Understanding |
| Selection | Sample | Data Understanding |
| Pre-Processing | Explore |
| Transformation | Modify | Data Preparation |
| Data Mining | Model | Modelling |
| Interpretation/Evaluation | Assessment | Evaluation |
| Post KDD | ------------- | Deployment |

Table 3.1. Comparison among KDD, SEMMA, and CRISP-DM (Azevedo and Santos, 2008).

# 

# CHAPTER 4: Data Analysis

**Overview**

## Business Understanding

Several hedge funds and asset managers have integrated cryptocurrency-related assets into their investment portfolios and trading strategies. The academic community has also devoted substantial efforts to researching cryptocurrency trading (Fang *et al.*, 2022). As a burgeoning market and research area, cryptocurrencies and cryptocurrency trading have witnessed significant advancements and a noteworthy surge in attention and engagement (Farell, 2015). According to Figure 4, it is evident that over 85% of the papers have been published since 2018, indicating the emergence of cryptocurrency trading as a fresh research domain in the financial trading field. The survey covers the time span from 2013 to June 2021 (Fang *et al.*, 2022). Cryptocurrencies are relatively young, extremely volatile, and prone to frequent shocks. It is challenging to value cryptocurrencies as there are no underlying fundamental factors and their value can be greatly impacted by investors' sentiments (Anamika, Chakraborty and Subramaniam, 2023). By understanding the relationship between sentiment and cryptocurrency portfolio performance, stakeholders can make more informed investment decisions and improve the overall portfolio management. The stakeholders include investors, financial analysts, and cryptocurrency traders who will benefit from the insights gained from the model (Anamika, Chakraborty and Subramaniam, 2023).

Figure 4.6. The cumulative number of publications on cryptocurrency trading (Fang *et al.*, 2022).



## Data Understanding

In the data understanding phase of the CRISP-DM methodology, we started by exploring and gaining insights into two public datasets that we worked with:

**Comments from online forums -** This dataset covers the period from 7 August 2015 to 7 April 2019 and is intended for the development of a deep learning model based on a transformer to examine the sentiments and emotions expressed in articles related to cryptocurrency.

**Cryptocurrency price dataset -** This dataset will be used for time series analysis to understand the correlations between the prices of cryptocurrencies over the same period as the forum comments and the sentiments and emotions extracted from textual data.

## Data Preparation

In order to improve the effectiveness of deep learning models, data pre-processing techniques that are a very important step will be employed to clean the data (Aslam *et al.*, 2022, Abraham *et al.*, 2018). According to the purpose of this research, we used two different datasets therefore the preparation has been done in two general parts as outlined below:

## Cryptocurrency price dataset preparation steps

The first step of data preparation is to make sure of the quality and accuracy of the variables. Next, according to those results, we did the data preparation steps including Principal Component Analysis (PCA), Missing and Irrelevant Values Removing, Data Transformation, Outlier Handling, and Data Normalization.

## Text (comments on online forums dataset) preparation steps

We conducted the textual data pre-processing as follows:

1) Tokenization,

2) Punctuation Removal,

3) Removing Numbers, Characters, and Special Characters,

4) Stemming and lemmatization,

5) Removing stop words and short words,

6) spell correction,

Tokenization divides the text into smaller parts called tokens, while punctuation removal eliminates symbols that limit the learning capability of algorithms. Removing numbers, characters, and special characters reduces complexity and enhances model efficiency. Stemming and lemmatization clarify affixes from sentences and convert words to their root form, respectively. Removing stop words and short words improves model performance, while spell correction corrects misspelled words. All these steps improve the quality of data and ultimately enhance the performance of models (Aslam *et al.*, 2022).

## Modeling

To achieve the purpose of this research, we will train a model consisting of two sub-models on the prepared data.

## Transformer-based deep learning model

The aim of the first model is to detect the emotions from the textual data and another to study the relationship between emotions and crypto prices. The first transformer-based deep learning model will recognize emotions from cryptocurrency-related comments. The transformer design will let the model understand the semantic meaning of the text and determine if it is positive, negative, or neutral. This model improves sentiment analysis by collecting word dependencies and text context. This model outputs comments sentiment and emotion scores.

## Time Series Model

The time series model will investigate the correlation of historical cryptocurrency prices and sentiment and emotion extracted from the first model. The model will capture the link between emotions and cryptocurrency prices to be used to predict cryptocurrency portfolio performance.

## Evaluation

The evaluation of our novel transformer-based deep learning model for sentiment analysis involves a comprehensive array of performance metrics and methodologies to gauge its efficacy and draw comparisons with existing models. In this section, we provide a detailed overview of the evaluation process.

## Performance Metrics

In evaluating our transformer-based sentiment analysis model, we employed performance metrics, including:

* **Accuracy:** Assessing the overall correctness of sentiment label predictions on the given text data (Tabinda Kokab, Asghar and Naz, 2022).
* **Precision:** Measuring the proportion of true positive sentiment predictions among all positive predictions, showcasing the model's ability to avoid false positives (Tabinda Kokab, Asghar and Naz, 2022).
* **Recall:** Reflecting the model's capability to identify positive sentiment instances, recall calculates the proportion of true positive predictions among all actual positive cases (Tabinda Kokab, Asghar and Naz, 2022).
* **F1-Score:** Balancing precision and recall, the F1-score provides a comprehensive measure of the model's performance (Tabinda Kokab, Asghar and Naz, 2022).
* **Confusion Matrix:** Visually summarizing classification results, the confusion matrix presents the number of accurate and incorrect sentiment predictions for each sentiment class, encompassing true positives, true negatives, false positives, and false negatives (Tabinda Kokab, Asghar and Naz, 2022).

## ROC and AUC

To further gauge our transformer-based sentiment analysis model's performance, we utilize Receiver Operating Characteristic (ROC) Curves. These curves illustrate the trade-off between the true positive rate and false positive rate at different sentiment classification thresholds. Additionally, the Area Under the Curve (AUC) metric quantifies the overall performance, with higher AUC values signifying superior sentiment analysis capabilities (Tabinda Kokab, Asghar and Naz, 2022).

## Deployment

As deployment is not a part of my research objectives so we will not discuss how we will deploy the model.

# CHAPTER 5: Discussion

## Comparison with Existing Models

In order to showcase the effectiveness of our model, we will conduct a thorough comparative analysis against other leading sentiment analysis models in the field. By employing various embedding techniques and deep learning architectures, we ensure a comprehensive evaluation and highlight our model's potential superiority. This is covered with extensive details in Chapter 5.

# CHAPTER 6: Conclusion

# References

1. Abraham, J. *et al.* (2018a) ‘Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis’, 1(3).
2. Abraham, J. *et al.* (2018b) ‘Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis’, 1(3).
3. 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.
4. 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.
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. Aslam, N. *et al.* (2022) ‘Sentiment Analysis and Emotion Detection on Cryptocurrency Related Tweets Using Ensemble LSTM-GRU Model’, *IEEE Access*, 10, pp. 39313–39324. Available at: https://doi.org/10.1109/ACCESS.2022.3165621.
8. Aurpa, T.T., Sadik, R. and Ahmed, M.S. (2022) ‘Abusive Bangla comments detection on Facebook using transformer-based deep learning models’, *Social Network Analysis and Mining*, 12(1), p. 24. Available at: https://doi.org/10.1007/s13278-021-00852-x.
9. Azevedo, A. and Santos, M.F. (2008) ‘KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW’.
10. Azevedo, A. and Santos, M.F. (no date) ‘KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW’.
11. 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.
12. 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.
13. 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.
14. Chapman, P. *et al.* (1999) ‘The CRISP-DM user guide’, in *4th CRISP-DM SIG Workshop in Brussels in March*. sn.
15. Chen, M., Yu, P.S. and Han, J. (1996) ‘Data Mining: An Overview from a Database Perspective’, *IEEE Transactions on Knowledge &amp; Data Engineering*, 8(06), pp. 866–883. Available at: https://doi.org/10.1109/69.553155.
16. DeVries, P.D. (2016) ‘An Analysis of Cryptocurrency, Bitcoin, and the Future’, 1(2).
17. 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.
18. Fang, F. *et al.* (2022) ‘Cryptocurrency trading: a comprehensive survey’, *Financial Innovation*, 8(1), p. 13. Available at: https://doi.org/10.1186/s40854-021-00321-6.
19. Farell, R. (2015) ‘An Analysis of the Cryptocurrency Industry’.
20. Farell, R. (no date) ‘An Analysis of the Cryptocurrency Industry’.
21. Feldman, R. (2013) ‘Techniques and applications for sentiment analysis’, *Communications of the ACM*, 56(4), pp. 82–89. Available at: https://doi.org/10.1145/2436256.2436274.
22. 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.
23. 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.
24. 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.
25. Hassonah, M.A. *et al.* (2020) ‘An efficient hybrid filter and evolutionary wrapper approach for sentiment analysis of various topics on Twitter’, *Knowledge-Based Systems*, 192, p. 105353. Available at: https://doi.org/10.1016/j.knosys.2019.105353.
26. Kilimci, Z.H. (2020) ‘Sentiment Analysis Based Direction Prediction in Bitcoin using Deep Learning Algorithms and Word Embedding Models’, *International Journal of Intelligent Systems and Applications in Engineering*, 8(2), pp. 60–65. Available at: https://doi.org/10.18201/ijisae.2020261585.
27. 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.
28. Loginova, E. *et al.* (2021) ‘Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data’, *Machine Learning* [Preprint]. Available at: https://doi.org/10.1007/s10994-021-06095-3.
29. Marshan, A., Kansouzidou, G. and Ioannou, A. (2021) ‘Sentiment Analysis to Support Marketing Decision Making Process: A Hybrid Model’, in K. Arai, S. Kapoor, and R. Bhatia (eds) *Proceedings of the Future Technologies Conference (FTC) 2020, Volume 2*. Cham: Springer International Publishing (Advances in Intelligent Systems and Computing), pp. 614–626. Available at: https://doi.org/10.1007/978-3-030-63089-8\_40.
30. 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.
31. 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.
32. 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.
33. 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.
34. Omari Firas (2023) ‘A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach’, *World Journal of Advanced Engineering Technology and Sciences*, 8(1), pp. 009–014. Available at: https://doi.org/10.30574/wjaets.2023.8.1.0147.
35. Palacios, H.J.G. *et al.* (2017) ‘A comparative between CRISP-DM and SEMMA through the construction of a MODIS repository for studies of land use and cover change’, *Advances in Science, Technology and Engineering Systems Journal*, 2(3), pp. 598–604. Available at: https://doi.org/10.25046/aj020376.
36. 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.
37. Plotnikova, V., Dumas, M. and Milani, F. (2020) ‘Adaptations of data mining methodologies: a systematic literature review’, *PeerJ Computer Science*, 6, p. e267. Available at: https://doi.org/10.7717/peerj-cs.267.
38. Qaroush, Z., Zakarneh, S. and Dawabsheh, A. (2022) ‘Cryptocurrencies Advantages and Disadvantages: A Review’, *International Journal of Applied Sciences and Smart Technologies*, 4(1), pp. 1–20.
39. Roy, S., Nanjiba, S. and Chakrabarty, A. (2018) ‘Bitcoin Price Forecasting Using Time Series Analysis’.
40. Saltz, J.S. (2021) ‘CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps’, in *2021 IEEE International Conference on Big Data (Big Data)*. *2021 IEEE International Conference on Big Data (Big Data)*, Orlando, FL, USA: IEEE, pp. 2337–2344. Available at: https://doi.org/10.1109/BigData52589.2021.9671634.
41. Schouten, K. and Frasincar, F. (2016a) ‘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.
42. Schouten, K. and Frasincar, F. (2016b) ‘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.
43. Schröer, C., Kruse, F. and Gómez, J.M. (2021a) ‘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.
44. Schröer, C., Kruse, F. and Gómez, J.M. (2021b) ‘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.
45. 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.
46. Shafique, U. and Qaiser, H. (2014) ‘A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA)’, 12(1).
47. Siripurapu, M. *et al.* (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.
48. Tabinda Kokab, S., Asghar, S. and Naz, S. (2022) ‘Transformer-based deep learning models for the sentiment analysis of social media data’, *Array*, 14, p. 100157. Available at: https://doi.org/10.1016/j.array.2022.100157.
49. Vanaja, S. and Belwal, M. (2018) ‘Aspect-Level Sentiment Analysis on E-Commerce Data’, in *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*. *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)*, Coimbatore: IEEE, pp. 1275–1279. Available at: https://doi.org/10.1109/ICIRCA.2018.8597286.
50. 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.
51. 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.
52. Wirth, R. and Hipp, J. (2000) ‘CRISP-DM: Towards a Standard Process Model for Data Mining’.
53. Wright, P. (1998) ‘Knowledge discovery in databases’, 5(2), pp. 23–26.
54. Yadav, J., Kumar, D. and Chauhan, D. (2020) ‘Cyberbullying Detection using Pre-Trained BERT Model’, in *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*. *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India: IEEE, pp. 1096–1100. Available at: https://doi.org/10.1109/ICESC48915.2020.9155700.
55. 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.