ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

# Master

Data analytics For Business

# Master's Final Work

Dissertation

Twitter Impact on Bitcoin price

Pedro Gaspar Jorge

October - 2023

ISEG
Lisbon School
of Economics
& Management
Universidade de Lisboa

# MASTER
DATA ANALYTICS FOR BUSINESS

## MASTER'S FINAL WORK
DISSERTATION

TWITTER IMPACT ON BITCOIN PRICE

PEDRO GASPAR JORGE

**SUPERVISION:**
PROF. CARLOS J. COSTA

Lisbon School of Economics & Management
Universidade de Lisboa

OCTOBER - 2023

## ABSTRACT

Ten years ago, the Bitcoin price was $133 on September 30, 2013. On September 30, 2023, Bitcoin registered a price of over $27,000. Bitcoin is exchanged similarly to stocks in the stock market, operating independently of traditional financial systems for portfolio diversification. Twitter played a significant role in Bitcoin's rise, using social media platforms to spread information and build communities. These communities promoted Bitcoin, creating hype. The objective is to understand if tweets impact Bitcoin's price movements and if their analysis is useful for forecasting. CryptoBERT and VADER sentiment analysis tools will be used comparing their effectiveness. Various Machine Learning models such as Random Forest, XGBoost, AdaBoost, SVM, KNN, Bayesian Regression, and GBM will be applied for a comprehensive understanding. The analysis will focus on active social media users to determine the accuracy of their information in forecasting Bitcoin fluctuations. The study's findings may contribute to ongoing research on sentiment analysis for a broader range of financial instruments, including Bitcoin.

KEYWORDS: Bitcoin; Twitter; Sentiment Analysis; Vader; CryptoBert; Machine Learning

## TABLE OF CONTENTS

## TABLE OF FIGURES

## ACKNOWLEDGMENTS

I would like to thank my advisor, Professor Carlos Costa, whose guidance, contribution, and unwavering support have been crucial throughout this journey.

I would also like to thank my family and friends for their belief in me, which has been decisive to my conquests.

Finally, I thank my outstanding girlfriend, who has been on my side during this tough couple of months while I managed to balance my life between work and the master's thesis.

# 1. INTRODUCTION

Twitter is a social media platform that allows its users to post and read short messages and tweets. The platform has over 300 million users worldwide. The company was founded in 2006 with the initial objective of communicating between individuals through short messages. However, over the years, it has evolved into a powerful platform for communicating, sharing news and information, or even promoting businesses. The fact that this social media operates in real-time provides people a free pass to indirect contact with what is happening worldwide. Twitter has also become a popular platform for leaders to communicate with their followers. Crypto influencers and CEOs can use Twitter to share their thoughts, insights, and perspectives on the market and announce new projects and partnerships. By following these accounts, investors can obtain valuable insights into the market's direction and the prospects of specific cryptocurrencies.

The platform has become more and more important when researching cryptocurrencies. Users can use Twitter's search functionality to find information about specific coins and read reviews and opinions from other investors. In fact, as a common social media user, it is interesting how active crypto-related users are, which creates a synergetic environment for crypto information. Twitter is populated by many accounts only related to crypto, providing regular updates about their development and partnerships. Moreover, there is a lot of promotion of the area in the platform, such as coin offerings and other fundraising events, which highly promotes the activity and visibility of crypto to all Twitter users. All these factors make tweets a very important source of information that can be used in many studies. Since social media is in direct contact with crypto investors, influencers, and founders, people in the industry are trying to understand and analyse how this information flows impact the market.

One way of analysing textual data like tweets is by doing a Sentiment Analysis, a field of NLP (Natural Language Processing) that deals with identifying, extracting, and analysing opinions and emotions from the text. This process determines whether a tweet, in this case, transmits a positive, neutral, or negative sentiment. Sentiment Analysis has become increasingly important in recent years due to the vast amount of information generated by users, like this example, by tweets, which can be the input to the most diversified studies. The main goal of this analysis type is to make the process of analysing and categorizing textual data more automatized, which can be achieved by NLP and ML (Machine Learning). Two different models were chosen to perform this type of Analysis: Vader and CryptoBERT.

As mentioned before, Twitter may play a significant role in sharing opinions towards Bitcoin, so there is the belief that studying the sentiment behind the tweets might be interesting since it is possible to understand how people feel about the cryptocurrency at a specific point in time. Since Bitcoin is highly volatile, the sentiment people express about it can impact whether investors are buying or selling the asset, which may directly impact the price.

Although sentiment analysis may look perfect, there are several points to consider when doing this kind of work. Since analysing sentiments imply a computerized analysis of people's expressions, a couple of factors are considered. Furthermore, the challenges and limitations of this analysis will be mentioned with its composition and results in mind.

## *1.1 Research Question and Objective*

The main objective of this study is to understand whether tweets may impact the price of Bitcoin and how they can serve as an object of study to interpret the fluctuations of the Bitcoin close price.

Since our main data source is textual data, we need to extract numerical expressions of tweets to use them as input for our models. With the use of NLP more specifically Sentiment Analysis, it is possible to calculate how positive or negative is the tweet expression in numerical data. By doing this, there is the opportunity to create models with the tweet's sentiment.

In order to capture the whole characteristics of the tweet influence, it is important to consider the information about the user who wrote that tweet, who tweeted. The characteristics that were found important to analyse jointly with the tweet sentiment were the impact and exposure of the tweet.

It is important to address how much exposure the user has and how much impact was created by the tweet to understand how the specific tweet might have influenced the movement of Bitcoin price. There is a strong belief that there is a lot of potential in the tweet analysis of Bitcoin and financial data. Although it is expected that cryptocurrencies tend to be more subjected to speculation, the conviction that there is value in analysing tweets about financial assets is strong.

> Social media content is no longer only about sharing pictures of a vacation, being entertained by the musings of kittens, or watching a tutorial on how to bake chocolate chip cookies. For many, social media has become an indispensable source of information on all aspects of life, including news, finance, and investments.
>
> In: (Matsuyama & Wood, 2022)

With the social media boom and the increase in people's life exposure, the term "influencer" has become a huge showcase for almost everything that can be purchased. Financial products are not an exception. There is a significant number of accounts in almost all social media, specifically focusing on influencing financial decisions. It is important to keep in mind that if subjected to influence, this type of content is attached to a high frequency of risk. There are many potential risks associated with providing financial information on social media, but the need to be aware of the value in extracting valuable information from there is important to focus on here.

Furthermore, there is hope that this study may provide additional insights into sentiment analyser selection and in the growing importance of analysing Twitter financial data. The importance of studying financial information in social media makes it more and more pertinent to assess the effectiveness of these two sentiment analysis models while conducting this study.

### *1.2 Thesis Structure Overview*

In terms of structure, chapter 2 starts with the literature review, where relevant literature regarding sentiment analysis of tweets will be reviewed. A lot of works were written regarding the topic, and it is relevant to share what has been done and how it relates to this work. This literature review will be transversal to all the topics that are going to be analysed throughout the work also, with the objective of identifying current gaps or limitations that the study aims to fill.

Moreover, chapter 3 will cover the methodology of the analysis. This section will present how this work was developed, from the data preparation and pre-processing to the application of the models. Since the data was collected from a dataset available online, it will be explained what the factors on the table were to either use the dataset or extract the data directly from Twitter and why the decision to rely on the dataset was made. After going through the data analysis, the models used and their relevance in the data science context are going to be approached. As mentioned before, there is the intention of capturing the whole characteristics of the tweet in the model created. So besides from the sentiment of the tweet, the features of the user who tweeted will also be analysed. One tweet may have exposure but no impact. For example, if you have a million followers and you tweet, the probability of people seeing the tweet is much higher than someone with a hundred followers. However, even with one million followers, your tweet can have 0 impact. The impact will be measured by the number of favourites. Favourites are the same as likes in other social media, which represents, in theory, that someone likes your tweet, of course, this can happen for other reasons, but by having a lot of favorites, at least you know that your tweet had an impact because it generated a reaction by another user. In this work, the level of exposure was measured by the number of followers and the level of impact by the number of favorites. These two features seem functionally relevant to improve the model since they will give weight to the polarity of the tweet, which is the main output of the sentiment analyser models.

Furthermore, chapter 4 will present the results of this study. This section will go through the outputs of the research, where the models will be compared, and the results will be evaluated. Although having mathematical results that support the original statement is fundamental, understanding the reasoning behind the results and what could have been done better if this thesis had been written today is fundamental.

Chapter 5 will cover the main findings of this work and address what can be concluded with the results interpretation.

# 2. Literature Review

## *2.1 Literature Review Overview*

Character development is a purposeful and important element in both films and novels, enabling artists to add important background information and improve the plot. A similar methodology is used in this literature review, which aims to examine pertinent studies and their significance for the subject at hand.

As mentioned before, this thesis focuses on whether Twitter as a social media can influence people to buy or sell Bitcoin (Aparicio et al.2022). Twitter is one of the most popular platforms used by millions of users worldwide. The platform provides a channel for people to express their feelings, thoughts, and opinions about pretty much everything. The ability of Twitter to spread news and information in real-time is one of the most significant factors contributing to its influence on people's activities and investments. The platform allows users to follow their favourite influencers, celebrities, and news enterprises, giving them access to real-time updates on the hot topics of the moment. When compared to some decades before, when a company gets bankrupt or is going to be financed/bought by some investor, this type of information was not published as fast as now. With social media, investors have access to information much faster, which may influence them to make more accurate decisions (Costa et al., 2021).

Nowadays, with the advances in ML and, more specifically, in NLP, it is possible to extract numerical reasoning out of text variables, which is becoming an ultimate advantage since the numerical data can serve as input to prediction models, which can help us to forecast the future. NLP has an enormous impact in this work since there was an attempt to extract value from tweets so it could be studied whether they might have an impact in Bitcoin investors' actions. Throughout this literature review, there will be a distinct analysis by topics.

The first topic is tweet analysis and social media's influence nowadays, focusing on the financial world. Secondly, the different ways of analysing this type of data are by going through the typically used models. Furthermore, it will also be discussed which machine learning models perform better regarding both financial data and sentiment analysis outputs. Since there is a large group of sentiment analysis algorithms and machine learning models, this analysis pretends to address which are the most popular since it can support the model selection (Costa et al, 2021, Aparicio et al, 2021).

### *2.2 Literature Review by Thematic*

### *2.2.1 Tweet Analysis*

Insights, trends, and feelings come from the succinct, casual, and contextually rich communications broadcast on social media platforms through the process of tweet analysis. The power of Twitter is essentially the production of news in real-time (Trigka et al., 2022). Understanding public opinion and the voice of the people on diverse issues requires the use of this practice. The escalating study of Twitter with the understanding of its own impacts and multifaceted dynamics has been a focal point for researchers and organizations that are interested in leveraging this information. With the ability to see when a tweet was posted, it is also possible to tell how those feelings change over time. This makes Twitter an excellent resource for collecting text data on a topic like cryptocurrencies to explore the possible relationships between them and prices (Abraham et al., 2018). User-generated material, particularly tweets, has increased at an unprecedented rate because of social media platforms' exponential expansion. Because of the enormous amount of data it contains and the instantaneous window it provides into public opinion, tweet analysis is a useful tool for academics, companies, and governments (Aparicio, et al. 2021, Costa et al, 2021).

The use of tweet analysis in the world of cryptocurrencies has been investigated by numerous studies and frameworks. In this setting, tweets are essential for capturing current investor perspectives, market mood, and new patterns in the dynamic world of cryptocurrency, and one of the most followed ways of capturing these variables is by doing a sentiment analysis (Trigka et al., 2022; Abraham et al., 2018; Stenqvist and Lönnö, 2017). Discussions on cryptocurrencies frequently use industry-specific jargon, technical words, and references. It is possible to find a lot of hashtags, cashtags, crypto wallet addresses, URLs, and usernames, so accurate emotion interpretation requires a mixed knowledge of both the language and the complex operations of the crypto environment (Kulakowski and Frasincar, 2023). Furthermore, quick research is necessary to ensure relevance, given the quick pace of advancements and market moves in the Bitcoin industry.

Despite the insights that could be developed, there are certain drawbacks to using tweet analysis in cryptocurrency. The perception of cryptocurrencies can be incredibly unstable, influenced by news, legislative developments, and even endorsements from famous people. It might be difficult to distinguish between sincere sentiment and planned attempts to manipulate the market. Cross-lingual sentiment research is also considered a challenge given the fact that the vast range of languages and cultures that result from cryptocurrencies' decentralized and international character (Kulakowski and Frasincar, 2023).

Another topic that cannot be forgotten is the ethics behind working with Twitter (Abraham et al., 2018), it is crucial that the academics who work with this data don't disseminate false information to further increase the advantage provided by such a model. The product of this type of study should serve as a tool to help understand if Twitter can be seen as a factor that influences people to buy or sell Bitcoin.

By looking into creative strategies that incorporate sentiment analysis methods with cryptocurrency domain expertise, this research aims to address these limitations. This study contributes to improving the accuracy and efficacy of sentiment analysis for examining Bitcoin price fluctuations and investor sentiment in the constantly changing world of crypto finance by leveraging the power of ML and NLP tailored to the cryptocurrency context.

### *2.2.2 Bitcoin*

Bitcoin first introduction was made in 2008 by Nakamoto (2008). Since then, a lot of unexpected events happened. One of them was the boom of cryptocurrencies and the main event, the boom of Bitcoin. Bitcoin stands as a pioneering example of a decentralized digital currency, which caused a profound shift in the way it is conceived, stored, and exchanged value. The fact that this cryptocurrency can actually generate financial profit is a clear way to explain why new and inventive approaches are seen when studying or predicting Bitcoin prices (Aparicio et al. 2022). Blockchain is also an important factor to consider, keeping in mind that the applications of blockchain go beyond peer-to-peer payments, bringing an even more sustained value to the table since it brings security, privacy, and a distributed ledger, bringing value to an innumerous number of applications (Abreu et la, 2018, Bernadino et al, 2022, Cesario et al, 2023, ). Since cryptocurrencies are connected to blockchain because they provide an incentive for machines, electricity consumption (Aparicio, Romao, and Costa, 2022), and the run and validations of the blockchain, there is a direct relation between both, which brings us to the conclusion that the increased usage of blockchain is also an increased usage of cryptocurrencies (Abraham et al., 2018)

The relationship between Bitcoin and social media, particularly Twitter, has become crucial in influencing discussions and estimates of the value of Bitcoin. Within the context of these forums, public opinion has a significant impact on the price fluctuation of Bitcoin. Social media's real-time capabilities allow for the quick broadcast of news, opinions, and analyses, quickly influencing market perceptions and causing price changes. As a result, the relationship between the opinions shared on social media sites like Twitter and the following changes in Bitcoin's value highlights the dynamic changes in contemporary financial markets (Bernardino et al., 2022).

### *2.2.3 Sentiment Analysis and NLP*

Data is the primary power source for modern decision-making in the digital age. With the increasing amount of information on internet platforms, it is crucial to be able to access and understand this enormous volume of textual material. The crucial link, NLP, enables machines to comprehend, examine, and gain knowledge from human speech. Since a considerable amount of data that is generated is unstructured text like tweets, NLP becomes fundamental when doing this type of analysis (Abraham et al., 2018).

Sentiment analysis, commonly called opinion mining, is a computational technique that involves analysing text data to determine the emotional tone and attitudes being communicated. It is essential for comprehending trends in social media, client feedback, and public opinion. Businesses and researchers can learn how people feel about particular issues, goods, or events by extracting feelings from text, which can help them make wise judgments, which goes well with the source of information being tweets.

There are several different sentiment analysis models, each with distinct strengths and limitations. In this study, it was applied two different algorithms: VADER (Hutto, C. J. and Gilbert, 2014) and CryptoBERT (Kulakowski and Frasincar, 2023). VADER is a lexicon-based algorithm which means that it relies on sentiment lexicons, which are lists of words with assigned sentiment scores. Words are assigned positive, negative, or neutral scores, and the overall sentiment of a text is calculated based on the cumulative scores of its constituent words. While simple, these methods might struggle with context and sarcasm. It is commonly used in most sentiment analysis studies since it provides great results when analysing textual data (Costa et al, 2021).

On the other hand, CryptoBERT is a machine learning-based model that was created by a post-train and fine-tuning of a Twitter-oriented model based on the BERT architecture, BERTweet (Nguyen et al., 2020) on the cryptocurrency domain.

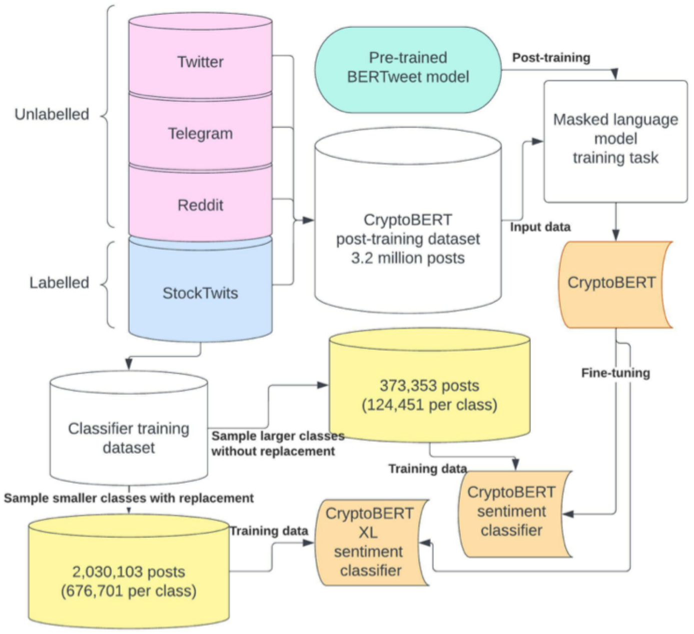

Unlabelled
Labelled
Twitter
Telegram
Reddit
StockTwits
Pre-trained BERTweet model
Post-training
Masked language model training task
CryptoBERT post-training dataset 3.2 million posts
Input data
CryptoBERT
Fine-tuning
373,353 posts (124,451 per class)
Classifier training dataset
Sample larger classes without replacement
Training data
CryptoBERT sentiment classifier
Sample smaller classes with replacement
Training data
CryptoBERT XL sentiment classifier
2,030,103 posts (676,701 per class)

Figure 1. CryptoBERT Post-Training and Fine-Tuning Flowchart

As mentioned before, BERT is an NLP model developed by Google researchers (Devlin et al., 2019). It was designed for a wide range of NLP tasks and has significantly advanced the state-of-the-art in various language understanding practices.

Since this model is state-of-the-art BERT-based, there is value in comparing his performance with VADER, which is seen as an established model among studies and researchers. BERT is based on transformer architecture and can be trained on a large, unsupervised corpus of text to learn the characteristics of a given language domain, which happens to be our source of data. In its paper (Kulakowski and Frasincar, 2023) CryptoBERT has already shown great results in comparison to VADER when handling crypto data. However, there is the intention to compare both since VADER was the sentiment analysis tool chosen in the articles referred to previously for the analysis (Trigka et al., 2022; Abraham et al., 2018; Stenqvist and Lönnö, 2017).

### *2.2.4 Machine Learning Models*

When training a model that could deliver great results in the current days, it is important to lean on machine learning. Ensemble techniques have gained prominence for enhancing the predictive performance of machine learning models by combining the outputs of multiple base models (Aparicio, Romao, and Costa, 2022).

By utilizing the combined knowledge of several base models, ensemble approaches have become an effective way to overcome the constraints of individual models. A well-known advanced boosting technique is XGBoost (Chen and Guestrin, 2016), which uses adaptive learning to iteratively improve model predictions. By integrating the results of weak learners, such as decision trees, XGBoost improves accuracy by concentrating on situations when sentiment classification is difficult. The model that emerges catches subtle sentiment hints, enabling a more thorough comprehension of textual emotion.

Another well-known ensemble method is Random Forest (Breinman, 2001), which makes use of the power of many decision trees that have all been trained on various data subsets.

Although knowing these two algorithms generally are the ones who perform better, it is also great to provide more studies and research so there will be compared not only the two previously talked about but also AdaBoost (Freund and Schapire, 1999), KNN (Fix and Hodges, 1951), SVM (Cortes and Vapnik, 1995), Bayesian Regression (MacKay, 2009) and GBM (Natekin and Knoll, 2013).

Moreover, machine learning techniques offer insightful information about the relative relevance of features, assisting in understanding the results of sentiment analysis. In terms of accuracy and generalisation, researchers have found that ensemble-based sentiment analysis models that include the previously mentioned routinely beat standalone models, so it will be great to incorporate these models with the inputs chosen to not only compare them but also to provide better forecasting materials.

Most of the studies related to this subject focus on classification models by analysing whether the sentiment of the tweet is positive or negative and the relation with the increase or decrease of the Bitcoin close price as Stenqvist and Lönnö (2017) and Abraham et al. (2018). Also, a type of analysis that can be seen in various research papers regarding the prediction of bitcoin price is the time series analysis that is performed by Shakri(2021) using neural networks (Livieris et al.,2021).

This analysis pretended to capture not only the sentiment behind the tweet but also the characteristics of each user to understand whether they can be merged into a model capable of not only predicting if the Bitcoin price is going to increase or decrease but also the numerical expression of that increase or decrease forecasting a potential close price value.

# 3. Methodology

## *3.1 Methodological Approach*

The generic methodological approach followed was supported and inspired by CRISP-DM (Costa and Aparicio, 2020, 2021). This approach starts with the domain understanding and then will go to data understanding and data preparation. Then, it is performed the modelling, data evaluation, and deployment. This approach may also be considered in the context of a design science approach (Aparicio et al., 2023).

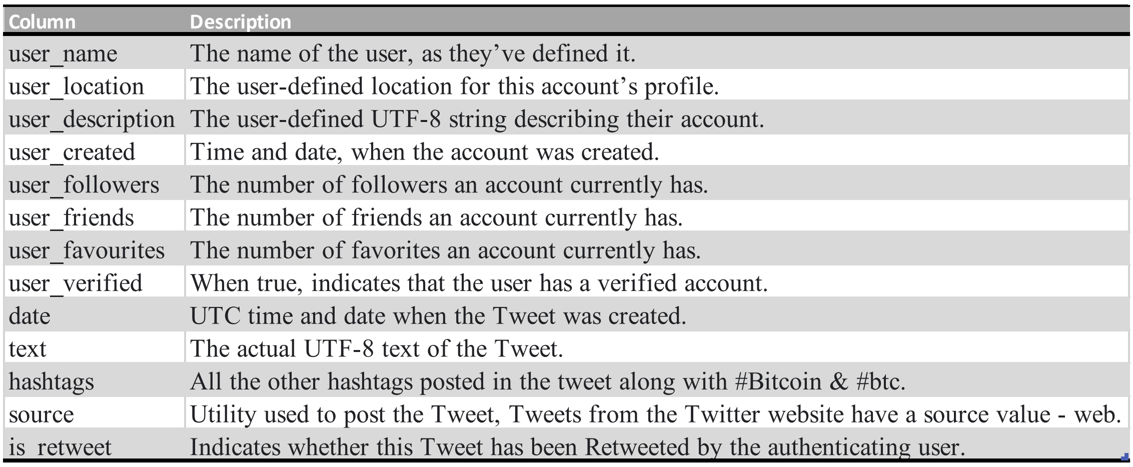
### *3.2 Dataset*

When Twitter was selected as the main data source for this study, the idea of extracting massively tweets specifically about Bitcoin was not clear. After some research, it was found that the most common way of doing that was by using the Twitter API. Twitter API is Twitter Application Programming Interface, and it is available for everyone who creates a Twitter developer account. With the application, any user can theoretically perform diverse tasks related to manipulating tweet data. However, there were some technical challenges while using this tool. The first barrier was the authentication and rate limit, which was the main challenge. However, it is quite easy to get a developer account. Accessing Twitter API requires authentication using tokens and imposes a limit on the API collection of tweets. As the problem was collecting a lot of data, the bottleneck was already created since it was not possible to get sample representativeness from that point onwards. Therefore, a consistent extraction of tweets could not be guaranteed since there was not a high volume of data that is needed for a good sample.

The solution chosen to go forward with the study was searching for a dataset online that was convenient and could address the research questions. The dataset chosen is a collection of tweets that have either #bitcoin or #btc in their text. This extraction started on 6/2/2021 with an initial 100000 tweets and keeps continuing daily. With this method of extraction, a sample that is large enough for this kind of study could be guaranteed. The relevance of the data to the study could also be guaranteed since the dataset would only comprehend tweets about Bitcoin. These factors ensured the necessity of analysing the data and understanding what needed to be done to guarantee data quality.

Knowing that the main data source for the study is Twitter and the focus is the tweets, this dataset gathers all the data needed as the basis of the work since it contains relevant information regarding each tweet and the user that tweeted.

For each tweet, there is access to the following columns:



| Column | Description |
|---|---|
| user_name | The name of the user, as they've defined it. |
| user_location | The user-defined location for this account's profile. |
| user_description | The user-defined UTF-8 string describing their account. |
| user_created | Time and date, when the account was created. |
| user_followers | The number of followers an account currently has. |
| user_friends | The number of friends an account currently has. |
| user_favourites | The number of favorites an account currently has. |
| user_verified | When true, indicates that the user has a verified account. |
| date | UTC time and date when the Tweet was created. |
| text | The actual UTF-8 text of the Tweet. |
| hashtags | All the other hashtags posted in the tweet along with #Bitcoin & #btc. |
| source | Utility used to post the Tweet, Tweets from the Twitter website have a source value - web. |
| is_retweet | Indicates whether this Tweet has been Retweeted by the authenticating user. |

Figure 2. Columns of the dataset

Although most of these fields may seem interesting to analyse in this study, the focus will be only on some of them since this analysis will be more analytical than exploratory.

### *3.3 Data Quality and Data Pre-processing*

This study focuses on the fluctuations of Bitcoin over time, so the need to choose an interval of time is crucial. Analysing two different intervals of time when studying a cryptocurrency like Bitcoin can be very different since cryptocurrencies are more volatile than a more traditional form of asset. The interval of time chosen was from 2021-02-19 to 2023-01-09, corresponding to over 95 million tweets in this dataset. However, this number might seem great for the analysis of the assumption that everyone's influences cannot be made.

An influencer in this context is someone with significant relevance that creates either impact or interaction with a significant number of people in the cryptocurrency world. However, there are people who have such a huge influence in one area that they project that influence on another area, such as Elon Musk, who is known for being the owner of Tesla but has already projected his influence in the cryptocurrency world. There is something possible to assume: there is no Twitter account that influences without having followers. Influencers, in general, have a substantial number of followers and produce content consistently, creating a sense of constant interaction with their followers.

The accounts associated with these personalities are also the ones that generate money in different ways, either by sponsorships, affiliate marketing, or other means. Recognizing that it was aggregated only users in the dataset that can potentially generate some impact in other users to buy or sell bitcoin and can be considered influencers. It was established a value of 30000 followers. This filter not only makes sure that the accounts comprehended in the analysed data have a high number of followers but also lowers the number of tweets to 187664, making the data easier to deal with with just a simple computer power.

As mentioned before, the data quality of Twitter requires a pre-processing of data, guaranteeing several quality needs. Twitter data can be notoriously messy due to its unstructured and informal nature. Character limits and users' overuse of abbreviations and slang make text normalization essential. Furthermore, hashtags and mentions also add further complexity to it, requiring a tokenization process. By acknowledging these characteristics of Twitter data, there was the need to ensure these factors in the pre-processing of data:

- Text normalisation: To make it simpler to compare and analyse tweets accurately, the text is normalised by changing it to lowercase, deleting unusual characters, and handling contractions and abbreviations.
- Tokenization: Text can be broken down into tokens (words or phrases) to allow for more detailed analysis, such as sentiment analysis or keyword extraction, which can help us understand user sentiments and themes better.
- Stopword Removal: Reducing the dataset's size and concentrating on key terms improves the effectiveness of subsequent analysis. Common words like "and," "the," or "is" are examples of words that can be eliminated.

By following these practices, it was possible to enhance the quality of Twitter's collected data, making it more suitable for sentiment analysis. Thus facilitating the extraction of valuable insights from these information-rich data while reducing the impact of irrelevant or noisy information. This process was only applied to the tweets that served as a source for VADER sentiment analysis since CryptoBERT accepted entire tweets and was created to process the whole corpus of the tweet, capturing all the essence behind the text.

### *3.4 Sentiment Analysis*

Sentiment analysis is an NLP technique used to determine the sentiment or emotional tone expressed in textual data. Nowadays, multiple sentiment analysis tools with a different genesis could be used in this study for the specific context of analysing the sentiment behind the tweets. However, there was not only the objective of using one of the most used sentiment analysis tools to understand whether they could perform with crypto tweets, but there was also the objective of using a state-of-the-art sentiment analysis tool so it would be possible to study something new. There was the opportunity to bring some innovation to the work. VADER and CryptoBERT were the two tools chosen.

The main objective of the use of sentiment analysis tools in a practical sense is to get the numerical expression of the tweets that can serve as input when modelling. Having the tweets pre-processed, VADER (Hutto, C. J., and Gilbert, 2014) searches the tokenized words of each tweet in the sentiment lexicon dictionary and attribute scores. The tool gives us four different scores: positive, neutral, negative, and compound. The first three provide us with the weights of each polarity of the tweet. The positive score quantifies the weight of positive sentiment in the text, while the negative and neutral scores provide the weight of the negative and neutral sentiment in the text. The last one gives us one numerical value between -1 and 1, representing the tweet's overall sentiment polarity. However, this score is not calculated by average. It is calculated by an algorithm that combines the words and evaluates the intensification and direction of it, giving a final score. In this work, the focus will be on the three individual sentiment scores, which provide more specific information about the sentiment expressed in the text.

Although CryptoBERT (Kulakowski and Frasincar, 2023) is also a sentiment analysis tool, it is quite different from VADER. In the creation of CryptoBERT, Mikolaj Kulakowski and Flavius Frasincar post-trained and fine-tuned a Twitter-oriented model based on the Bidirectional Encoder Representations from Transformers (BERT) architecture, BERTweet, on the cryptocurrency domain, resulting in the final result CryptoBERT. Although there was already a BERT-based model trained with a large corpus of financial news and reported to analyse the sentiment of financial texts, finBERT, the new model came up from the objective of creating a tool that could specifically integrate cryptocurrency knowledge.

When operating, CryptoBERT also works differently from VADER. While VADER gives us different sentiment scores, the BERT-based model gives us a label that can be "bearish", "neutral" or "bullish". These labels represent a categorization of a financial price movement. When an asset's price increases, it is popular to say that the specific asset is bullish. On the other hand, when an asset's price decreases, it is popular to say that the asset is bearish. It is possible to conclude that bearish is the equivalent to a negative sentiment score and bullish to a positive sentiment score. However, our objective is not to work with categories but with numbers and analytical reasoning. In order to address this, there was a need to understand how these labels are given. The tool also computes scores like VADER does. However, the output is not a score. CryptoBERT calculates the negative, neutral, and positive scores and the negative, neutral, and positive scores and then assigns a label according to the higher score of the text data. For example, if a tweet has a higher positive score, the algorithm keeps that score and assigns bullish labels. To analyse the outputs of this tool just like VADER, it was collected the higher score for each tweet and stored in the negative, positive, or neutral score if the label was bearish, bullish, or neutral. In this way, the categorized output represented by a negative, positive, or neutral could be weighted and more related to the output given by VADER.

By the end of this process for each tweet, there was access to 6 columns, where 3 of them were the individual sentiment scores calculated for VADER, vader_pos, vader_neg, and vader_neu, which correspond to the positive, negative, and neutral sentiment scores and three columns of the individual scores of CryptoBERT, bert_pos, bert_neg and bert_neu which correspond to the positive sentiment score, negative sentiment score and neutral sentiment score calculated by CryptoBERT.

## *3.5 Bitcoin Close Price and Volatility*

### *3.5.1 Bitcoin Price Analysis*

Significant discrepancies in analysis are revealed when analysing an asset over time, and this difference is especially obvious when examining Bitcoin. Depending on the timeframe chosen, Bitcoin's price volatility leads to different analysis, ranging from potential long-term store of value to short-term speculative assets. Different storylines are highlighted by historical price trends for Bitcoin, giving weight to its function as an investment or a speculative vehicle. Its association with traditional markets and other macro factors changes with time, changing how risk and reward are perceived. The market's attitude is heavily influenced by news and social media, which causes abrupt price changes.

In the following graph, it is possible to see the close price (USD) movement over time:

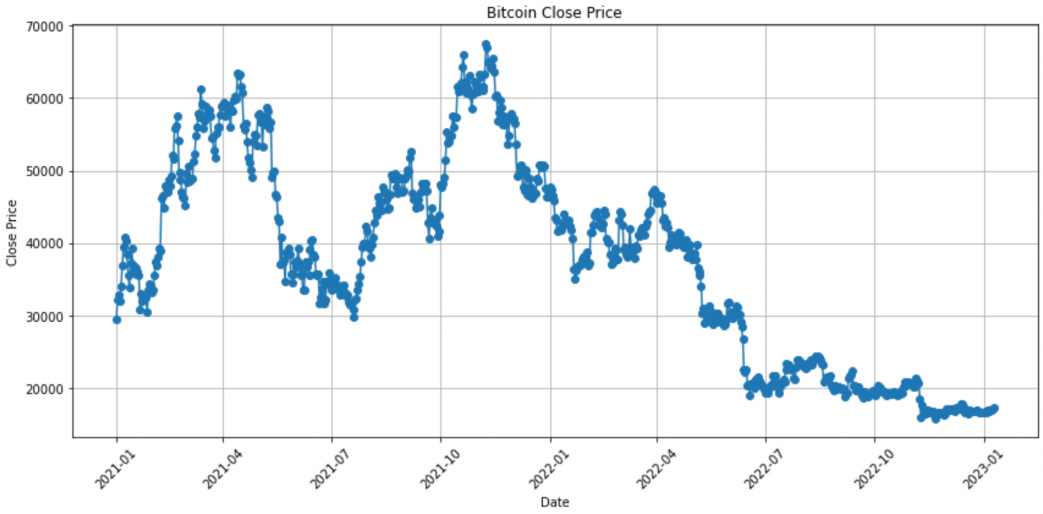


Figure 3. Bitcoin Daily Close Price

It is important to acknowledge that at the beginning of this time plot, the world was under a pandemic, and it is interesting the fact that there were some significant high points during such an unpredictable period. However, there is also a clear downward trend at the beginning of 2022, which is related to the invasion of Ukraine by Russia.

Compared to SP500, which comprises 500 of the largest publicly traded companies and is considered a key indicator of the overall health of the U.S. stock market, Bitcoin presents itself as a much more volatile asset.

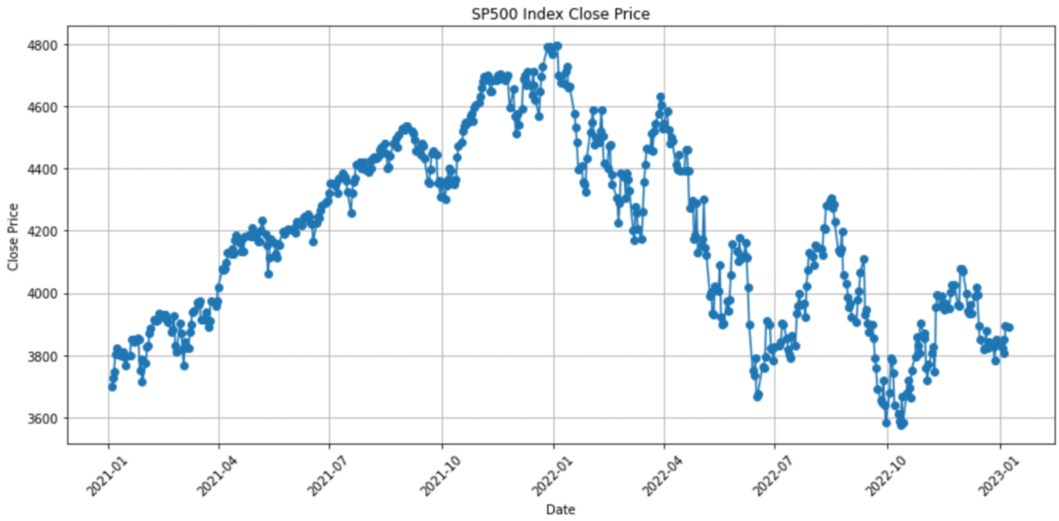


Figure 4. SP500 Index Close Price

While Bitcoin doubled and halved its price more than once in a short period of time, finishing the time interval with close to half the price compared to the beginning of the time interval, the SP500 followed a clear upwards and then downwards trend over this period, finishing the time period with an increase comparing to the beginning of the period which can relate with the global economy, since the upward trend relates to the pandemic ending period and the downward trend relates to the war impact in the economy.

### *3.5.2 Integration of Bitcoin in the analysis*

In order to integrate Bitcoin in this work it, Yahoo Finance was used to extract Bitcoin's daily close prices. With Yahoo Finance, the cryptocurrency data was merged with the tweet text and the sentiment scores, having a complete data frame with all the information necessary for this analysis.

However, it doesn't make sense to analyse a tweet with the close price of Bitcoin in the day that the tweet was written because there was no time for the tweet to make enough impact on influencing the price to move. Knowing this, each row containing the tweet's information was connected to the Bitcoin close price on the day after the tweet was written instead of the close price of that day. By doing this, the analysis of how the information written about Bitcoin influenced its price to move in the day after can be made, capturing the impact and range that the information written might had on the cryptocurrency.

### *3.6 Model Selection and Feature Selection*

The application of sophisticated ML algorithms to estimate Bitcoin values and improve sentiment analysis is one of the most encouraging breakthroughs. While offering deeper insights into market emotion and behaviour, these technical advancements have the potential to change how investors and traders approach the cryptocurrency market completely. Cryptocurrency price forecasting has long been a challenge due to the highly speculative characteristics of these assets. However, recent technological advances in ML have brought us quite closer to more precise predictions.

Though designed to uncover emotional undertones from textual data, sentiment analysis algorithms may have trouble dealing with human language's complexities and intrinsic complexities. Utilizing recent ML models is extremely important due to the maximization of potentialities to innovate while introducing incremental advances and upcoming insights that, although minimal, can benefit new studies and research.

In this work there is the expectation to bring and evaluate models that are known to be the ones that better perform and operate in this kind of study. To better understand how each model can improve and operate with each one of the features, they will be presented in the data science and sentiment analysis context.

#### *3.6.1 Model Selection*

The early 2000s saw the development of the potent ML algorithm known as Random Forest, principally by Leo Breiman and Adele Cutler. Its importance comes from its capacity to increase prediction accuracy and manage challenging jobs by utilizing ensemble learning. During training, Random Forest builds a great number of decision trees, each employing a random subset of the data and features. The final forecast is then voted on by all these trees, which aids in strengthening the model's robustness and minimising overfitting. Random Forest is a useful method for sentiment categorization when used in the context of sentiment analysis because it effectively captures subtle patterns and sentiments in text data by combining the predictions of multiple decision trees, especially when dealing with noisy or unstructured textual data from social media, as Twitter. Random Forest's adaptability and dependability make it a valuable tool in the field of natural language processing, considerably enhancing the precision of tasks like sentiment analysis and text classification, making the model a great tool for this study.

Extreme Gradient Boosting, often known as XGBoost, is a state-of-the-art machine learning technique that has become very popular in recent years. It is well known for its remarkable predicted accuracy and adaptability and is a member of the gradient-boosting family of algorithms. A group of decision trees is successively trained by XGBoost, with each one aiming to fix the flaws of the previous one. It is very skilled at managing complicated information and catching subtle patterns since it uses gradient descent optimisation techniques to minimise a defined loss function. The main advantages of XGBoost include handling missing data, preventing overfitting using regularisation, and offering insight into feature relevance. It has applications in a variety of fields, such as financial modelling and natural language processing, and as a result, it is being used more and more when working in data science projects when the best results possible are needed for their predictive modelling assignments.

AdaBoost, or Adaptive Boosting, is a popular ensemble learning algorithm that has gained popularity for its ability to enhance the performance of poor learners, often decision trees. It works by repeatedly training weak classifiers on the same dataset, giving extra weight to cases that the prior classifiers incorrectly classified. AdaBoost can concentrate on the difficult-to-classify samples thanks to this adaptive weighting method, significantly improving the ensemble's overall classification accuracy. AdaBoost ultimately creates a strong ensemble model—often outperforming individual base classifiers—by combining the results of these weak classifiers. AdaBoost is a useful machine learning method, especially for problems like binary classification, due to its simplicity, versatility, and resistance to overfitting.

A straightforward yet effective supervised machine learning method used for both classification and regression applications is the k-Nearest Neighbours (KNN) algorithm. KNN is based on the idea that items tend to belong to the same class or have similar numerical values if they have similar features. It operates by determining the separation between the data point that needs to be predicted or categorised and its k closest neighbours in the training dataset, where k is a user-defined number. The target data point is given the class or value that appears the most frequently among these neighbours. KNN is a non-parametric algorithm, which means that it doesn't make any firm assumptions about the distribution of the underlying data. However, the selection of the distance measure and the amount of k can have an impact.

Bayesian Regression is a statistical modelling approach that combines linear regression methods with the concepts of Bayesian inference. Bayesian Regression treats model parameters as probability distributions as opposed to classical linear regression, which estimates model parameters as fixed values. It starts with a prior distribution that represents our initial assumptions about parameter values and then computes a posterior distribution by updating this distribution with observed data. This posterior distribution offers a richer and more insightful perspective on the model by quantifying the uncertainty linked to the parameters as well as providing point estimates of those parameters. Thus, it is an important tool for enhancing the accuracy and robustness of an ensemble learning system. A more dependable and accurate overall prediction is created by combining the predictions of various models, or an ensemble. As one of the ensemble's constituent models, Bayesian Regression can be incorporated because of its capacity to model uncertainty and offer probabilistic predictions.

Gradient Boosting Machine (GBM) is renowned for its extraordinary versatility and predictive power. GBM is a member of the ensemble learning family and works by repeatedly fusing various weak learners, frequently decision trees, to produce a powerful prediction model, similar to the other boosting models. GBM develops each new tree by concentrating on the errors or residuals of the preceding trees, unlike some other ensemble approaches, gradually increasing the model's accuracy. GBM can balance between overfitting and underfitting by varying the learning rate and the number of trees in the ensemble, making it resilient and able to handle complicated datasets. It has applications in a variety of industries, including banking, healthcare, and natural language processing, and excels at a wide range of tasks, including regression.

The improvement of research and studies is fundamentally based on the methodical application of multiple machine learning models and the rigorous comparison of their outcomes.

### *3.6.2 Features Selection*

Having all the tools to the modelling set up, it was time to put together all the features that were going to be part of our model. As mentioned before our main inputs of our model were the output of the sentiment analysers. Since our objective was to understand whether Twitter impacts the price of Bitcoin, our main features that constitute the model are the numerical expressions of the feeling behind each tweet. These outputs were stored in variables, either the VADER or the CryptoBERT ones.

However, it is important to consider that a tweet written by someone who has 30 thousand followers, or 30 million followers is not the same, making it impossible to give the same weight to someone with 30 thousand or 30 million followers. This applies in the same way to the tweet's impact. For instance, if a tweet had 30 favourites or 30 thousand favourites, it is not the same since the second option might generate more impact since it caused a reaction by a greater number of users.

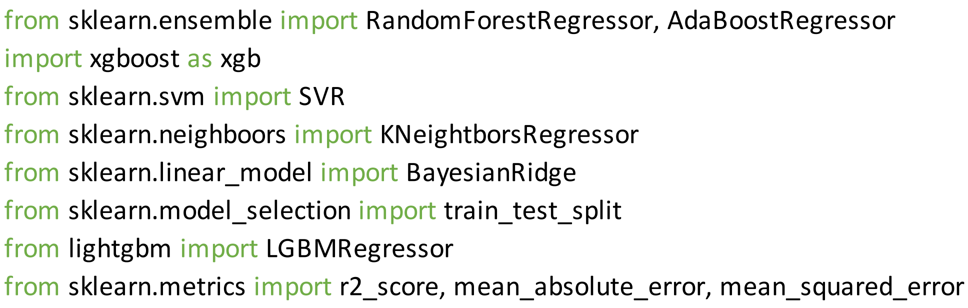
In order to address this problem, it was considered two extra features besides the sentiment analysers outputs: the impact and the exposure. The impact is related with the number of favourites that the tweet had. A favourite has mentioned earlier it's a reaction applied to a tweet by a different user. Although this reaction nowadays is designed as a heart, the assumption that every user that puts a favourite in a tweet loves what was written cannot be made, however it is possible to assume that a reaction was made so the tweet had impact in other user, which makes a tweet with more favourites a tweet that generated more impact. The exposure is related to the number of followers that the user has. Nowadays, influencers are paid more and more because of their number of followers since they reach a higher audience when they publish. However, some followers are more loyal or active than others. Someone with a higher number of followers has more exposure since what they publish is seen by a higher number of people.

### *3.6.3 Implementation*

To answer the main research question, there was the need for a model that has as input the numerical expression of a tweet, the sentiment analyser output, and understand whether it could predict or could be considered accurate with the Bitcoin price of the day after it was written. After collecting the tweets, transforming them into their own numerical expression, and choosing the features and models the data was prepared for implementation.

These were the libraries and functions imported to perform the models and to measure their results.

It was defined a train percentage split of 20% and the parameters of the machine learning models performed were the default from the functions imported.



```
from sklearn.ensemble import RandomForestRegressor, AdaBoostRegressor
import xgboost as xgb
from sklearn.svm import SVR
from sklearn.neighboors import KNeightborsRegressor
from sklearn.linear_model import BayesianRidge
from sklearn.model_selection import train_test_split
from lightgbm import LGBMRegressor
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
```

Figure 5. Figure Libraries and Functions Imported

When implementing the main predictive model to answer the research question, an idea surged. The data aggregated a large number of users since there were users who tweeted only a few regarding Bitcoin. However, there were users with a lot of activity about Bitcoin, generating a large number of tweets. Considering that these accounts with many followers create a vast quantity of content about Bitcoin, could they predict the price of the cryptocurrency well? To answer this question, the same models, and features to each one of the users with over a thousand tweets in the considered interval of time.

# 4. RESULTS

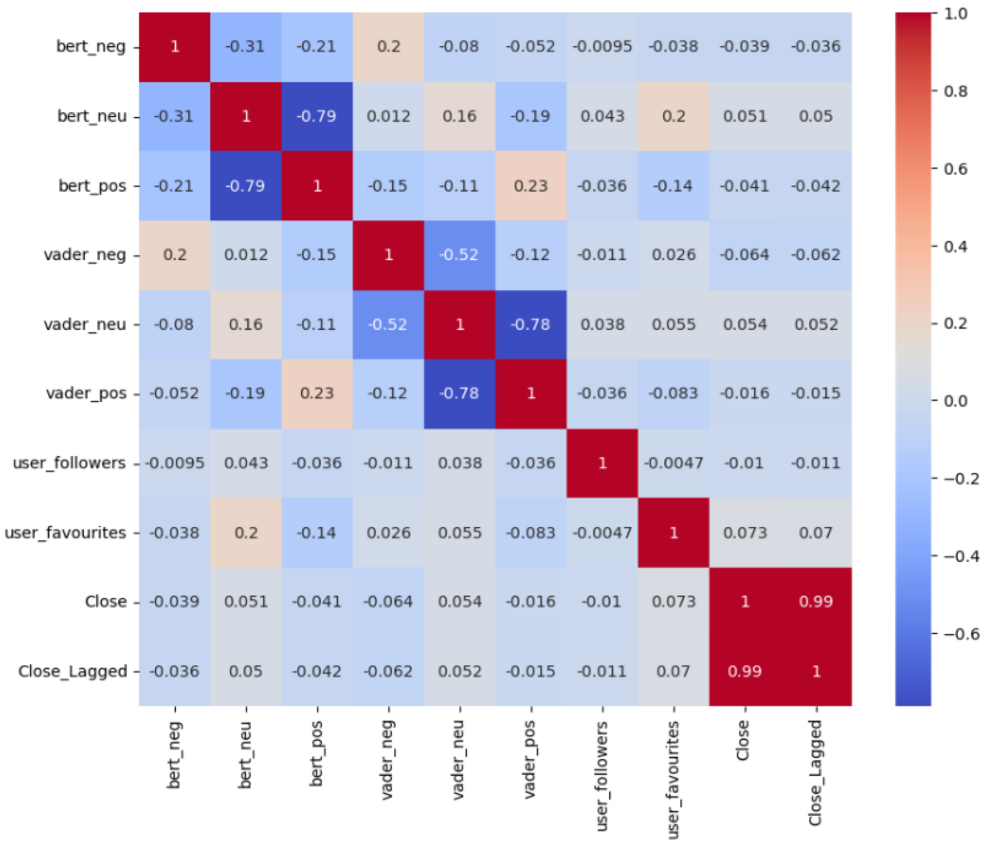
This chapter of the thesis goes into the fundamental essence of this research, which involves comprehensively presenting its findings. These results serve as the study's nucleus, emanating from applying the research methodology and the inherent logical reasoning underpinning this work. While the overarching objective remains the attainment of optimal outcomes, this chapter primarily centres on extracting meaningful insights from the numerical manifestations of the collected data, thereby facilitating the formulation of data-driven conclusions.

The results will be divided in two subsections. Firstly, the focus will be on the main analysis of this study that corresponds to answering his work's research question. The formulation of this analysis comprehends a model which will be analysed by the relation between its features and by performance and relevance. Secondly, the focus will be on the most active user's analysis where the same methodology was applied, although done to each of the Twitter users with more than a thousand tweets during the chosen time interval.

## *4.1 Main Results*

### *4.1.1 Linear Correlation Analysis*

In statistical analysis and data science, a correlation matrix is a fundamental tool used to measure the correlations between variables or features within a dataset. It offers a methodical methodology to assess the potency and direction of linear relationships between two sets of variables. The Pearson correlation coefficients must be calculated for each pair of variables to create a Pearson correlation matrix. These correlation coefficients vary from -1 to 1, where -1 denotes a perfect negative correlation, 1 denotes a perfect positive correlation, and 0 denotes no correlation.



|  | bert_neg | bert_neu | bert_pos | vader_neg | vader_neu | vader_pos | user_followers | user_favourites | Close | Close_Lagged |
|---|---|---|---|---|---|---|---|---|---|---|
| bert_neg | 1 | -0.31 | -0.21 | 0.2 | -0.08 | -0.052 | -0.0095 | -0.038 | -0.039 | -0.036 |
| bert_neu | -0.31 | 1 | -0.79 | 0.012 | 0.16 | -0.19 | 0.043 | 0.2 | 0.051 | 0.05 |
| bert_pos | -0.21 | -0.79 | 1 | -0.15 | -0.11 | 0.23 | -0.036 | -0.14 | -0.041 | -0.042 |
| vader_neg | 0.2 | 0.012 | -0.15 | 1 | -0.52 | -0.12 | -0.011 | 0.026 | -0.064 | -0.062 |
| vader_neu | -0.08 | 0.16 | -0.11 | -0.52 | 1 | -0.78 | 0.038 | 0.055 | 0.054 | 0.052 |
| vader_pos | -0.052 | -0.19 | 0.23 | -0.12 | -0.78 | 1 | -0.036 | -0.083 | -0.016 | -0.015 |
| user_followers | -0.0095 | 0.043 | -0.036 | -0.011 | 0.038 | -0.036 | 1 | -0.0047 | -0.01 | -0.011 |
| user_favourites | -0.038 | 0.2 | -0.14 | 0.026 | 0.055 | -0.083 | -0.0047 | 1 | 0.073 | 0.07 |
| Close | -0.039 | 0.051 | -0.041 | -0.064 | 0.054 | -0.016 | -0.01 | 0.073 | 1 | 0.99 |
| Close_Lagged | -0.036 | 0.05 | -0.042 | -0.062 | 0.052 | -0.015 | -0.011 | 0.07 | 0.99 | 1 |

Figure 6. Correlation Matrix

Since this figure aims to present the coefficients of correlation between variables, there are several conclusions possible to take from the linear relationships between them. Starting by looking at the chart's last column, it is possible to conclude that there is no clear linear relationship between the features and the independent variables. In order for the features to have a linear relationship with the close price of Bitcoin, their coefficient of correlation should be either close to 1 or -1. Given that all values closely approximate 0, it is reasonable to infer the absence of a discernible linear relationship between them. While this outcome may initially appear disheartening, it leads us to the inference that the model in question is not of a linear nature.

The values closer to the limits 1 and -1 are the ones from the relations between the sentiment analysis outputs. The relation between the positive and negative individual sentiment scores is the relation with more linear strength, which makes sense since a phrase to be negative cannot be positive. Also, the relationships between themselves always present a negative coefficient of correlation, although sometimes weak, meaning that for an individual sentiment score to go up, the others go down.

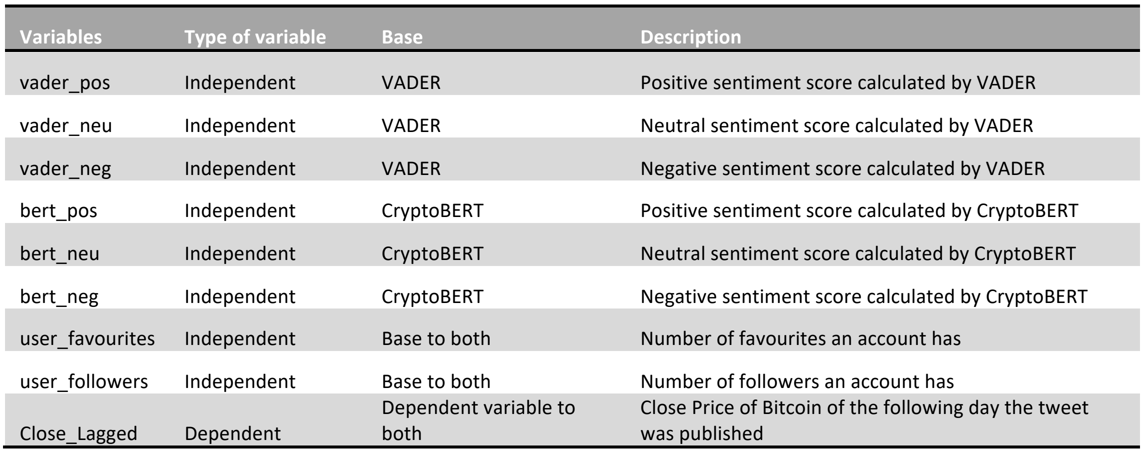
Although it is impossible to state that a relationship is a strong linear one, some values are still important to analyse. By looking at the polarity scores, for example, positive VADER and positive CryptoBERT scores, it is possible to see a weak positive relationship. Also this weak positive relation can also be seen between their negative scores and their neutral scores, which means that there is a weak positive linear relationship in how both tools scored the tweets.

Moreover, there is a weak positive linear relationship between the variable user_favourites and the individual neutral sentiment score of CryptoBERT, which shows that tweets with more favourites had a higher neutral sentiment score calculated by CryptoBERT.

In Trigka et al.(2022), Pearson's coefficient calculated between user followers and user favourites was 0.163, which differs from ours. Although both values are close to zero, there is no conclusion to take regarding that information.

### *4.1.2 Modelling*

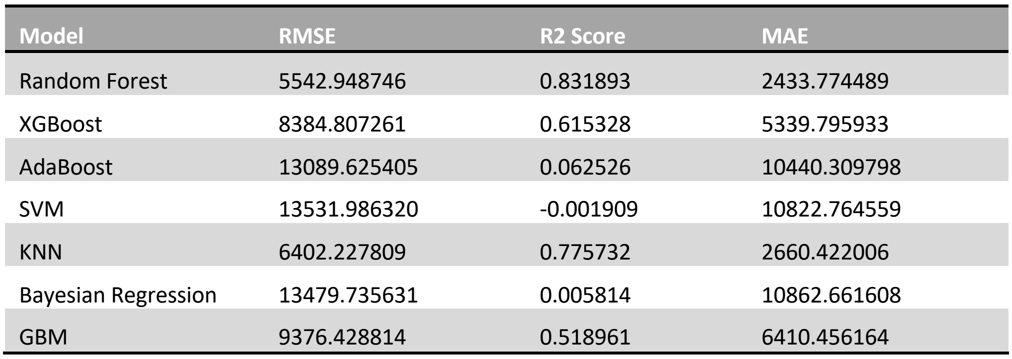
Since it is impossible to identify a linear regression between the dependent and independent variables, this study performed machine learning models to the referred features.



| Variables | Type of variable | Base | Description |
|---|---|---|---|
| vader_pos | Independent | VADER | Positive sentiment score calculated by VADER |
| vader_neu | Independent | VADER | Neutral sentiment score calculated by VADER |
| vader_neg | Independent | VADER | Negative sentiment score calculated by VADER |
| bert_pos | Independent | CryptoBERT | Positive sentiment score calculated by CryptoBERT |
| bert_neu | Independent | CryptoBERT | Neutral sentiment score calculated by CryptoBERT |
| bert_neg | Independent | CryptoBERT | Negative sentiment score calculated by CryptoBERT |
| user_favourites | Independent | Base to both | Number of favourites an account has |
| user_followers | Independent | Base to both | Number of followers an account has |
| Close_Lagged | Dependent | Dependent variable to both | Close Price of Bitcoin of the following day the tweet was published |

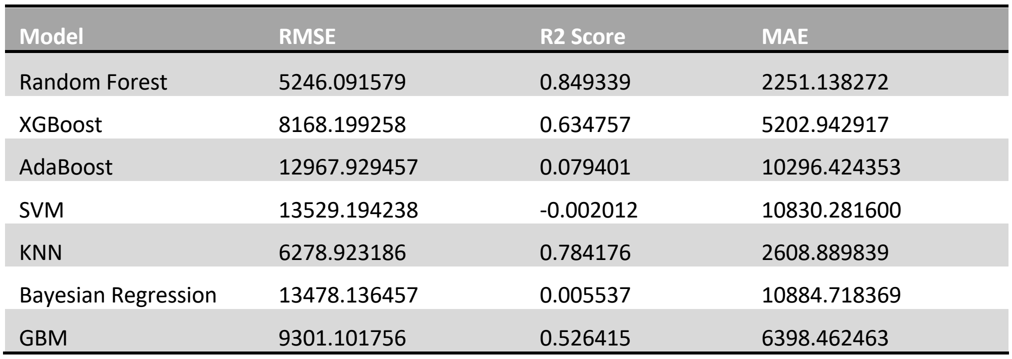
Figure 7. Variables of the models

In the figures 8 and 9 are presented the results of their performance. The common features between the models are the number of favourites of the tweet and the number of followers of the account who posted the tweet, or as mentioned before, the impact and exposure. The features that differ between pictures are the outputs of the sentiment analysis. While Figure 8 uses the sentiment analysis outputs calculated by VADER, figure 9 uses the sentiment analysis outputs calculated by CryptoBERT.



| Model | RMSE | R2 Score | MAE |
|---|---|---|---|
| Random Forest | 5542.948746 | 0.831893 | 2433.774489 |
| XGBoost | 8384.807261 | 0.615328 | 5339.795933 |
| AdaBoost | 13089.625405 | 0.062526 | 10440.309798 |
| SVM | 13531.986320 | -0.001909 | 10822.764559 |
| KNN | 6402.227809 | 0.775732 | 2660.422006 |
| Bayesian Regression | 13479.735631 | 0.005814 | 10862.661608 |
| GBM | 9376.428814 | 0.518961 | 6410.456164 |

Figure 8. VADER Models Results



| Model | RMSE | R2 Score | MAE |
|---|---|---|---|
| Random Forest | 5246.091579 | 0.849339 | 2251.138272 |
| XGBoost | 8168.199258 | 0.634757 | 5202.942917 |
| AdaBoost | 12967.929457 | 0.079401 | 10296.424353 |
| SVM | 13529.194238 | -0.002012 | 10830.281600 |
| KNN | 6278.923186 | 0.784176 | 2608.889839 |
| Bayesian Regression | 13478.136457 | 0.005537 | 10884.718369 |
| GBM | 9301.101756 | 0.526415 | 6398.462463 |

Figure 9. CryptoBERT Models Results

In order to evaluate the performance of each model, it was chosen three indicators that are usually popular for this practice. A key factor in determining how well a machine learning model captures the underlying variance in the data is the R2 Score, often known as the coefficient of determination. This statistic provides a measurable assessment of the model's capacity to account for and explain the variations or fluctuations seen in the target variable. When calculating the R2 Score, it is possible to measure how much the model explains the Bitcoin price variation. On the other hand, RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) provide information on the model's accuracy in making predictions and the size of its errors. While MAE is more robust in the presence of outliers and offers a clear interpretation of the average prediction error, RMSE is particularly effective for penalising greater errors and making it sensitive to outliers.

In this comparative analysis the model composed by the CryptoBERT outputs presents a higher R2 Score than the model composed by VADER outputs when using all the machine learning models referred to before. The results provide a data-driven conclusion that CryptoBERT based models performed better in explaining the variations of Bitcoin price than the VADER based ones. The model that presents the higher R2 Score was the Random Forest applied to the CryptoBERT outputs, which was around 0.85, which means that the model explains 85% of the variance. Moreover, the models that use CryptoBERT-produced outputs presented predicted values with lower errors from the true values than VADER-based models.

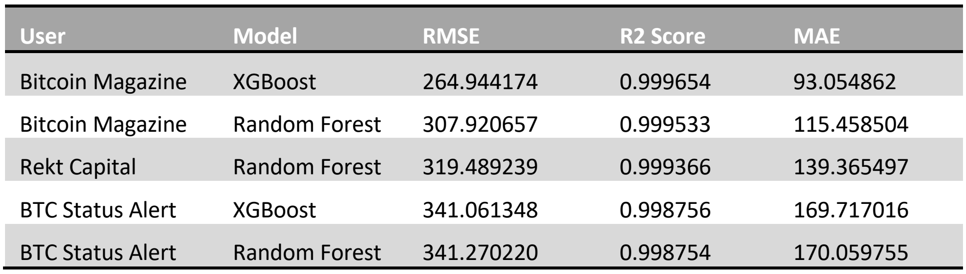
Regarding the performance of each machine learning model the three models that performed better were Random Forest, KNN, and XGBoost. These results are transversal to both sentiment analysis tools since, by the order mentioned, Random Forest, KNN, and XGBoost present higher values of R2 Score and the lowest values of RMSE and MAE, meaning that these models explain better the variance of the bitcoin price and predicted prices that are closer to the true price values.

### *4.2 Users Analysis*

In this subsection, the analysis was done regarding the most active users of the dataset. It made a filter selecting only the users that tweeted more than a thousand times in the chosen period of time. It was performed the same batch of models referred before to this subset of the dataset to each one of the users. By doing this analysis there is an individual overview of each user and understand whether some tweet more accurately towards bitcoin price than others.

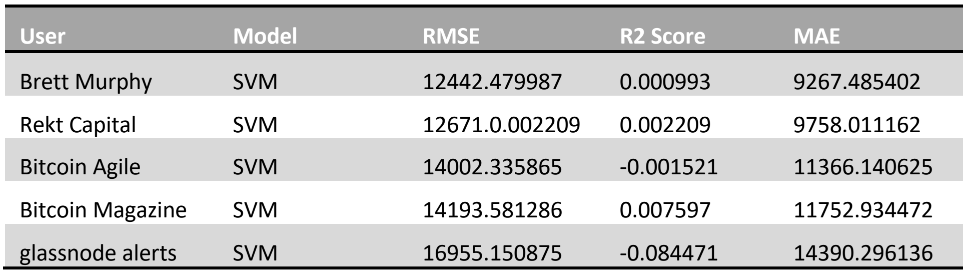
Figures 10 and 11 present the top and bottom five VADER-based models performance while the figures 12 and 13 present the top and bottom five CryptoBERT-based models performance. In Figure 10 and 11, the features combined with the VADER outputs were applied, while in Figure 12 and 13, the features combined with the CryptoBERT outputs. By comparing these results with the main results presented before, it is possible to conclude that there are top users who create content on social media with consistent and accurate data. The top 5 models composed of 5 five different users between the two figures present an R2 Score close to 1. This statistically means that the model almost fully explains the variance of bitcoin price. Likewise, the models present error values that are low, keeping in mind the Bitcoin close price at this paper's published date.

When comparing the results between the models performed with VADER features and CryptoBERT features, the VADER model appears to present better models. The top 3 VADER-based models present a higher R2 Score with scores of more than 0.999 and lower RMSE and MAE than the top 1 CryptoBERT-based model.



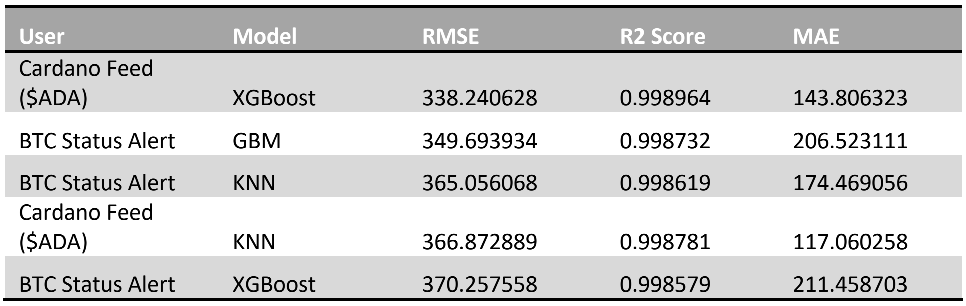
| User | Model | RMSE | R2 Score | MAE |
|---|---|---|---|---|
| Bitcoin Magazine | XGBoost | 264.944174 | 0.999654 | 93.054862 |
| Bitcoin Magazine | Random Forest | 307.920657 | 0.999533 | 115.458504 |
| Rekt Capital | Random Forest | 319.489239 | 0.999366 | 139.365497 |
| BTC Status Alert | XGBoost | 341.061348 | 0.998756 | 169.717016 |
| BTC Status Alert | Random Forest | 341.270220 | 0.998754 | 170.059755 |

Figure 10. VADER Top5 User Models



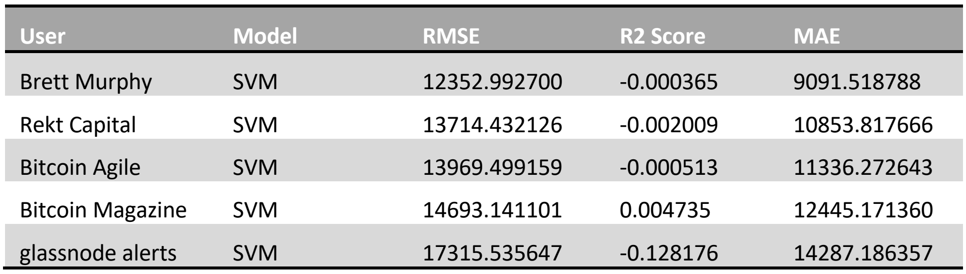
| User | Model | RMSE | R2 Score | MAE |
|---|---|---|---|---|
| Brett Murphy | SVM | 12442.479987 | 0.000993 | 9267.485402 |
| Rekt Capital | SVM | 12671.0.002209 | 0.002209 | 9758.011162 |
| Bitcoin Agile | SVM | 14002.335865 | -0.001521 | 11366.140625 |
| Bitcoin Magazine | SVM | 14193.581286 | 0.007597 | 11752.934472 |
| glassnode alerts | SVM | 16955.150875 | -0.084471 | 14390.296136 |

Figure 11. VADER Bottom5 User Models



| User | Model | RMSE | R2 Score | MAE |
|---|---|---|---|---|
| Cardano Feed ($ADA) | XGBoost | 338.240628 | 0.998964 | 143.806323 |
| BTC Status Alert | GBM | 349.693934 | 0.998732 | 206.523111 |
| BTC Status Alert | KNN | 365.056068 | 0.998619 | 174.469056 |
| Cardano Feed ($ADA) | KNN | 366.872889 | 0.998781 | 117.060258 |
| BTC Status Alert | XGBoost | 370.257558 | 0.998579 | 211.458703 |

Figure 12. CryptoBERT Top5 User Models

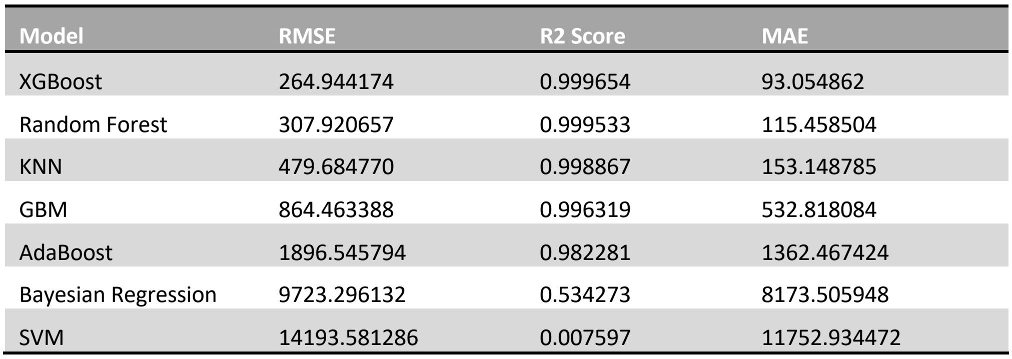


| User | Model | RMSE | R2 Score | MAE |
|---|---|---|---|---|
| Brett Murphy | SVM | 12352.992700 | -0.000365 | 9091.518788 |
| Rekt Capital | SVM | 13714.432126 | -0.002009 | 10853.817666 |
| Bitcoin Agile | SVM | 13969.499159 | -0.000513 | 11336.272643 |
| Bitcoin Magazine | SVM | 14693.141101 | 0.004735 | 12445.171360 |
| glassnode alerts | SVM | 17315.535647 | -0.128176 | 14287.186357 |

Figure 13. CryptoBERT Bottom5 User Models

Having in mind the users that appear to be more accurate in their tweets regarding the variance of Bitcoin price, the only user that appears in both figures is BTC Status Alert. This account happens to be a bot automatized by another account. In fact, all the accounts that are present in figures 10 and 11 create a lot of statistical content regarding the financial analysis of cryptocurrency data. Bitcoin Magazine is the oldest and most established source of news, information, and expert commentary on Bitcoin, its underlying blockchain technology, and the industry built around it. Since 2012, Bitcoin Magazine has provided analysis, research, education, and thought leadership at the intersection of finance and technology.

Regarding Bitcoin Magazine, the username attached to the account that provided the data leading to a better model presents an R2 Score of 0.999 and an average absolute error of 93, representing 0.3% of the Bitcoin close price at the data this paper published.



| Model | RMSE | R2 Score | MAE |
|---|---|---|---|
| XGBoost | 264.944174 | 0.999654 | 93.054862 |
| Random Forest | 307.920657 | 0.999533 | 115.458504 |
| KNN | 479.684770 | 0.998867 | 153.148785 |
| GBM | 864.463388 | 0.996319 | 532.818084 |
| AdaBoost | 1896.545794 | 0.982281 | 1362.467424 |
| Bayesian Regression | 9723.296132 | 0.534273 | 8173.505948 |
| SVM | 14193.581286 | 0.007597 | 11752.934472 |

Figure 14. Bitcoin Magazine User Models

Figure 14 shows that XGBoost, Random Forest, and KNN were the models that performed better again, showing consistency in performance compared to other models. It is also possible to state that SVM appears consistently at the bottom, showing poor performance in relation to the others.

## 5. Conclusion

This thesis approaches the sentiment analysis of tweets and how to capture the characteristics of a tweet about Bitcoin to predict the close price of the day after the tweet was written. The methodology followed produced three main outputs: the correlation matrix, the results of the model focusing on the whole dataset, and the results of the users' models. Regarding the correlation matrix, although it is possible to identify some weak linear relationships, there is nothing concrete to take conclusions about clear linear relationships since the only strong linear relationships were between the sentiment analysis tools, which are evident for the tool to function properly. The only conclusion that can be made is that there were no clear linear relationships between the features and the Bitcoin price the day after.

The results of the models were positive regarding the methods of evaluation chosen. For the whole dataset, the models that presented the best results were both Random Forest based, while the ones that presented the worst results were SVM-based.

Random Forest presented a higher R2 Score followed by a lower value of both RMSE and MAE. Regarding the sentiment analysis tool, CryptoBERT showed better results than VADER for the whole dataset.

The results were even better in the user analysis, with users presenting R2 Score values close to 1. Although having good results is the objective of every research, there is the belief that results that appear to be too good always generate questions. In fact, the user that populated the top 5 models performing for both methodologies is a bot that tweets statistical information about bitcoin price and volume status. However, it might be interesting to realize that accounts that only tweet about present statistical information can produce content that highly explains the movement of the coin. Furthermore, XGBoost, Random Forest, and KNN were the Machine Learning models that performed better, while CryptoBERT and VADER produced very good results.

Regarding these results, it is important to acknowledge that one of the key factors for this analysis to be significant is the amount of data that was worked with, and although there were analysed thousands of tweets, the amount is still low to be considered representative. Moreover, our main data source, the tweets, is a type of unstructured data containing a lot of noise, and it is hard to work with, making capturing value harder.

In our data-driven world, it is crucial to keep relentlessly pursuing the extraction of value from textual data regarding Bitcoin and other financial assets. Identifying market trends, hazards, and any kind of shift that may affect the price movement of assets can be critical to the economy's evolution and the companies' valorisation.

# References

Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review*, 1(3), Article 1.

Abreu, P. W., Aparicio, M., & Costa, C. J. (2018). Blockchain technology in the auditing environment. In *2018 13th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/10.23919/CISTI.2018.8399460

Aparicio, J. T., de Sequeira, J. S., & Costa, C. J. (2021). Emotion analysis of portuguese political parties communication over the covid-19 pandemic. In 2021 16th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI52073.2021.9476557

Aparicio, J. T., Romao, M., & Costa, C. J. (2022). Predicting Bitcoin prices: The effect of interest rate, search on the internet, and energy prices. In *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-5). Madrid, Spain. https://doi.org/10.23919/CISTI54924.2022.9820085

Aparicio, J. T., Aparicio, M., & Costa, C. J. (2023). Design Science in Information Systems and Computing. In Proceedings of International Conference on Information Technology and Applications: ICITA 2022 (pp. 409-419). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-9331-2_35

Bernardino, C., Costa, C. J., & Aparicio, M. (2022). Digital Evolution: blockchain field research. In 2022 17th Iberian Conference on Information Systems and Technologies (CISTI) (pp. 1-6). IEEE https://doi.org/10.23919/CISTI54924.2022.9820035

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5–32. https://doi.org/10.1023/A:1010933404324

Cesario, F., J. Costa, C., Aparicio, M., & Aparicio, J. (2023). Blockchain Technology Adoption: Factors Influencing Intention and Usage. In A. R. da Silva, M. M. da Silva, J. Estima, C. Barry, M. Lang, H. Linger, & C. Schneider (Eds.), Information Systems Development, Organizational Aspects and Societal Trends (ISD2023 Proceedings). Lisbon, Portugal: Instituto Superior Técnico. https://doi.org/10.62036/ISD.2023.9

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A methodology to boost data science. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. IEEE https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C. J., & Aparicio, J. T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In *Advances in Parallel & Distributed Processing, and Applications: Proceedings from PDPTA'20, CSC'20, MSV'20, and GCC'20* (pp. 65-75). Springer International Publishing. https://doi.org/10.1007/978-3-030-69984-0_7

Costa, C. J., Aparicio, M., & Aparicio, J. (2021). Sentiment Analysis of Portuguese Political Parties Communication. In Proceedings of the 39th ACM International Conference on the Design of Communication (SIGDOC'21) (pp. 63-69). Association for Computing Machinery (ACM). https://doi.org/10.1145/3472714.3473624

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186). Minneapolis, Minnesota: Association for Computational Linguistics.

González-Mendes, S, González-Sánchez, R., Costa, C. & García-Muiña, F (2023) "Analysing the state of the art of Blockchain application in Smart Cities: A bibliometric study," 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), Aveiro, Portugal, pp. 1-6, doi: 10.23919/CISTI58278.2023.10211371.

Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225.

Kulakowski, M., & Frasincar, F. (2023). Sentiment Classification of Cryptocurrency-Related Social Media Posts. *IEEE Intelligent Systems*, 38(4), 5-9.

Livieris, I. E., Kiriakidou, N., Stavroyiannis, S., & Pintelas, P. (2021). An Advanced CNN-LSTM Model for Cryptocurrency Forecasting. Electronics, 10(3), 287. MDPI AG.

MacKay, D. J. C. (2009). Bayesian interpolation. Neural Computation, 4(3), 415–447.

Matsuyama, A., & Wood, T. (2022). The rising role of social media 'finfluencers' June 29 2022. Site: https://www2.deloitte.com/cn/en/blog/financial-advisory-financial-services-blog/2022/the-rising-role-of-social-media-influencers-in-finance.html

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized business review*.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21.

Nguyen, D. Q., Vu, T., & Nguyen, A. T. (2020). BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 9-14). Association for Computational Linguistics.

Pedregosa, F., Varoquaux, Ga"el, Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825–2830.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37–52).

Shakri, I. (2021). Time series prediction using machine learning: a case of Bitcoin returns. Studies in Economics and Finance. ahead-of-print.

Silverman, B. W., & Jones, M. C. (1989). E. Fix and J.L. Hodges (1951): An Important Contribution to Nonparametric Discriminant Analysis and Density Estimation: Commentary on Fix and Hodges (1951). *International Statistical Review / Revue Internationale de Statistique*, 57(3), 233–238.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis. Master Computer Science, KTH Sweeden.

Sul, H., Dennis, A. R., & Yuan, L. I. (2014). Trading on Twitter: The Financial Information Content of Emotion in Social Media. *2014 47th Hawaii International Conference on System Sciences*, Waikoloa, HI, USA, pp. 806-815

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1 (9/1/2001), 211–244.

Trigka, M., Kanavos, A., Dritsas, E., Vonitsanos, G., & Mylonas, P. (2022). The Predictive Power of a Twitter User's Profile on Cryptocurrency Popularity. *Big Data and Cognitive Computing*, 6(2), 59.

# APPENDICES