SOLENT UNIVERSITY, SOUTHAMPTON
FACULTY OF BUSINESS, LAW, AND DIGITAL TECHNOLOGIES

**MSc Applied Artificial Intelligence and Data Science**
**Academic Year 2021-2022**

Idris Babalola

# Cryptocurrency Market Sentiment Analysis

Supervisor: Dr Peyman Heydarian

September 2022

This report is submitted in partial fulfilment of the requirements of Solent University for the degree of MSc Applied Artificial Intelligence and Data Science

# Acknowledgement

First and foremost, I would like to take this opportunity to express my sincere gratitude and thanks to my supervisor Dr. Peyman Heydarian who made this work possible. His tutelage and domain expertise have been indispensable in crafting this level of work.

In addition, it would be remiss of me not to mention Dr Femi Isiaq. The exposure I gained from working on his projects in collaboration with Solent university has afforded me the luxury of being involved in research and composing research writing of the requisite standard.

I would also like to give special thanks to my wife, without whose support, patience and understanding when undertaking my research and writing project.

Finally, I would like to thank God for making this journey achievable and for giving me the ability to reach my goals.

## Abstract

Over recent years, the data generated by Social Media websites has been increasing exponentially. Opinions, rumors, news about the latest trends and technologies are generated in humongous quantity by people every day. These data are usually shared through various Social Media sites such as Twitter and Reddit.

The project aimed to use this information about the latest cryptocurrencies to predict the outcome of the cryptocurrency market. Multiple pre-processing techniques were experimented with the data before concluding data preparation. TF-IDF Vectorizer was selected for feature extraction in the model built after thorough experimentations. An imbalance among the classes was detected in the dataset which was tackled with tactful implementation of Synthetic Minority Oversampling Technique (SMOTE). C-Support Vector Classification algorithm was selected as the primary classifier after thorough testing and experimentations with several other state of the art classifiers put up against each other. The model was further optimized with hyperparameter tuning.

The results of these experimentations were finally brought together in an interactive GUI Dashboard. The dashboard allows the user to analyze the sentiment of the most recent data generated in social media sites using various visualization tools. It also allows for comparison with the live cryptocurrency market. The user may also input their own data into the dashboard to analyze the sentiment using a state-of-the-art SVC model.



**Keywords:** Cryptocurrency; NLP; Sentiment Analysis; Price; Predictive Model

# Table of Contents

# Chapter One

## 1 Introduction

With the advancement in technology and better and stable internet connections available to most individuals, there is a huge volume of data generated every day that is stored and available to all internet users. There exist a diverse range of mediums through which data is generated. Examples include online learning platforms, how-to-do articles, online gaming, freelancing platforms, blogging websites, entertainment streaming services and others. Among these, the biggest generator of data is social media which gathers it content from various popular websites such as Instagram, Tiktok, Facebook, Twitter, Reddit etc. These platforms afford people the opportunity to share their personal experiences, lifestyles and offer up their opinions and views on differing subjects ranging from the political arena, the stock market, entertainment and media reviews. People engage in discussions on these topics on social media platforms. The sentiment of the market is inferred from their collective discussions and views.

Consequently, several works have attempted different approaches to gather data and perform exploratory analysis and data mining to gain a comprehensive insight into people's behavior while posting the content. Companies can gain understanding from unstructured online text such as support tickets, emails, social media channels, blog posts, web chats, forums, and comments by using tools for sentiment analysis. Automatic, hybrid and rule-based algorithms have largely substituted manual data processing.

The effectiveness of a marketing campaign can be evaluated, the target audience or demographics can be pinpointed, customer feedback can be collected via forms online, websites or social media (such as Twitter, Instagram, Snapchat, Facebook etc.), and market research can be carried out with the help of sentiment analysis tools. These tools can be used to determine popularity, reputation and brand awareness at a certain period or over time; track consumer reception of new products or features; evaluate the success of a marketing campaign; track. People can communicate personal updates, political viewpoints, as well as film and media evaluations through the use of these websites. Social media users debate on different subjects. Their stance on these debates mirrors the general sentiment of the market.

Cryptocurrency is becoming increasingly more significant in the domain of finance. The term "cryptocurrency" refers to a form of digital or virtual currency that is protected by cryptography. Because of this protection, it is very difficult to forge or double-spend this type of currency. The term "virtual money" is frequently used when referring to an alternate method

of conducting business transactions including the exchange of products and services. It is composed of several distinct types of decentralized cryptocurrencies, each of which possesses its own unique set of characteristics and a value that is distinct from that of the others in the network. The information flow that surrounds the subject of virtual money is characterized by fluctuation, with the nature of opinion and sentiment shifting in response to a variety of different factors in a way that is dependent on those factors. This is since the notion of virtual money is still very recent. Several studies have been done to study the many different approaches that may be used to get this data and to carry out exploratory analysis to gain a better knowledge of the posting behaviors that users engage in. With the advancement in statistical modeling and machine learning, it is now possible to predict the future movements in the price of cryptocurrencies up to certain confidence with the availability of enough computing power and data.

## 1.1 Background of the Study

Social Media generates a humongous amount of data every day. People share their opinions, views, and sentiments about different topics online through various platforms such as Twitter, Facebook, YouTube, LinkedIn, Snapchat, Reddit, Instagram, etc. The collective opinion of people for some topic results in the categories such as 'trending', 'hot topic', 'viral', and many more (Ge-Stadnyk et al.,2017). Due to this interest of people in these topics, sentiment analysis has been studied over a long time to understand people's opinions and feelings about various matters. This analysis also helps in understanding the people's behavioral trait in a psychological context, for example, if a person is suffering from depression or anxiety, their recent posts on different social media platforms might hint at their neurological state (Luo et al.,2013), and if detected early, might even prevent some serious issues in the future and helps in better diagnosis and prognosis for the same (Orhan et al., 2020).

The successful initial public offering of a decentralized cryptocurrency came about in the year 2009. It was initially developed by an unknown person or group working under the guise of Satoshi Nakamoto, and the market for virtual currencies has experienced phenomenal expansion as a result. Since that time, there has been a lot of interest around cryptocurrencies in general, as well as Bitcoin specifically. On the one hand, proponents of cryptocurrencies assert that they will bring about fundamental shifts in both the global economy and the political landscape, whereas, on the other hand, there are those who assert that Bitcoin is fundamentally flawed and that it will not be around for much longer because of these flaws (Narayanan et al.

2016). According to (Liu,Tsyvinski., 2021), the returns on the prices of cryptocurrencies can primarily be forecast by momentum and the attention devoted to them by investors. This is the case since momentum tends to precede price increases. To make price forecasts using cryptocurrencies, numerous other academic organisations have successfully implemented algorithms based on deep learning. These forecasts are generated utilising historical data by the academics (Livieris. ,2020; Patel.,2020). The open and closing prices of the trading day, as well as the highest and lowest prices, the total value of bitcoin on the market, and the daily number of transactions utilising bitcoin were used as input factors in these analyses.

Due to severe market volatility, it is hard for academics to combine several approaches to ascertain the value of bitcoin (Caporale, Gil-Alana, & Plastun 2018). Bitcoin (Nakamoto, S.2008), the most well-known cryptocurrency, is and will continue to serve as the standard for cryptocurrency coin valuation. Due to its decentralised nature and immutable properties, blockchain technology is particularly well suited for applications in a wide variety of other fields. The utilisation of data obtained via social media platforms, such as Twitter, for the purpose of conducting sentiment analysis has lately acquired popularity (Dritsas, E.2018), particularly when it comes to discussion postings that investigate people' opinions and feelings regarding bitcoin.

The sentiment analysis task can be carried out in two ways. The first method involves summarizing the semantic meaning of the sentences by using the static resources, where the knowledge databases containing affective lexicons are utilized. As mentioned in (Ghiassi et al., 2013), psychological experts or automotive processes create lexicons, that select the features along with labeled text corpus. The effect of these features on a text and their capability of predicting the polarity of provided text are used to solve the required problem. Texts usually contain expressions that carry emotional valence, such as “great” (positive valence) or “terrible” (negative valence), leaving readers with a positive or negative impression. With the use of a valence dictionary, words in documents generally can be classified as either negative or positive (and sometimes as neutral) in lexicon sentiment analysis approach. Take the phrase “Good cryptocurrencies such as Bitcoin and Ethereum sometimes have bad days too.”. A valence dictionary would label the word “Good” as positive; the word “bad” as negative; and possibly the other words as neutral. Once each word in the text is labeled, an overall sentiment score is derived by counting the positive and negative words and combining these values mathematically. The sentiment score is finally used to classify the sentiment of the text. Also,

it is to be noted that in the lexicon-based approach we don't use Machine Learning models. The overall sentiment of the text is determined on-the-fly, depending only on the dictionary that is used for labeling word valence.

The second approach for carrying out sentimental analysis on text data involves treating the problem as a text classification problem. The basic pipeline for this method contains several steps starting from pre-processing the text dataset, converting pre-processed text to vector embeddings, and finally the classification using machine learning-based algorithms. Each step of this process is experimental since there are multiple substeps or choices for each step, and a different combination of those results in varying performance. The text is pre-processed in such a way that it retains the most useful information or semantic meaning of the text by removing the inherent noise as much as possible. The quality of the features generated depends a lot on the input, therefore it is a crucial step. The general pre-processing involves the following steps: -

- Uppercasing/Lowercasing: - Text is either lowercased or uppercased to remove case-sensitive scenarios and facilitate smooth text matching, for e.g., a sample tweet "The weather is BeauTIFul today." will be converted to "the weather is beautiful today." or "THE WEATHER IS BEAUTIFUL TODAY."
- Stopwords Removal: - Dimensionality is a huge problem in the training of machine learning models. Common words which don't contribute a lot to the semantics of the sentence are removed to increase the effectiveness and reduce the response time. Following are some of the keywords provided by the nltk library in python as part of stopwords:- "me", "my", "myself", "we", "am", "is", "our", "between", "into", "against", "both", "each", "few", "more", etc.
- Punctuation Removal:- Similar to stopwords, punctuations don't contribute to the sentence semantics and are therefore removed as part of preprocessing. Following punctuations are removed in general from a sentence:- ".", ",", ";", ":", "?", "!", "`", "_", "-", "()", "[]", "...", "/", "@", etc.
- Stemming/Lemmatization:- This process is used to reduce the word to a smaller form by transforming it either to a root or based on the similarity of words with the same structure in the sentence. The changes by stemming and lemmatization differ as shown in the following example:-

    - Original word - “strange”, stemmed word - “strang”, lemmatized word - “strange”.
    - Original word - “is”, stemmed word - “is”, lemmatized word - “be”.
- Tokenization:- This process splits the words from the sentences into a list of list of pieces called tokens, this process is called word tokenization and is carried out using a word tokenizer. The splitting of a sentence into a list of sentences is a process called sentence tokenization which is carried by the sent tokenizer. Tokenization helps in the creation of vector spaces which is the next crucial step in the sentiment analysis pipeline. The following examples show the difference in the working of both tokenizers:-
    - Original Sentence:- “I love cats. I also love birds.”
    - Sent Tokenizer output:- [‘I love cats.’, ‘I also love birds.’]
    - Word Tokenizer output:- [‘I’, ‘love’, ‘cats’, ‘.’, ‘I’, ‘also’, ‘love’, ‘birds’, ‘.’]

The classification models capable of handling high-dimensional data are used since the vectorization techniques for generating vectors use multiple text transformation techniques.

## 1.2 Research Question

This research seeks to answer the questions:

1. Can sentiment analysis be utilized as an effective tool to draw correlations between cryptocurrencies market and other entities?
2. Can historical and real-time data be effectively combined to predict a trend of cryptocurrency market sentiment?

## 1.3 Aims and Objectives

The research is aimed at studying Cryptocurrency Market Sentiment Analysis using data obtained from different social media platforms, for the identification of most desirable machine learning model that will aid in accurate prediction of cryptocurrency market. Some of the objectives of the research include:

1. Using Natural Language Processing to predict the sentiment of the cryptocurrency market.
2. Developing a working code for counting positive and negative sentiments with an intuitive GUI.

3. Carrying out sentiment analysis using a combination of different machine learning-based approaches and text embeddings methods and comparing them using an evaluation metric.
4. Developing an ensemble model that employs different state-of-the-art classifiers used for sentiment analysis.

The scope of this research includes creating a labeled dataset using multiple methods for tweets from Twitter and posts from the different Reddit subreddits related to Bitcoin, Ethereum, and Litecoin cryptocurrencies. It also includes performing sentiment analysis classification by experimenting with different preprocessing pipelines, text embedding methods, and classification models along with the creation of an ensemble model of pre-existing SOTA (State of the art) models and baseline classifiers. The analysis will also include the change in the correlation of sentiments with the change in pricing of the different cryptocurrencies. All the visualizations will be presented in an intuitive graphical user interface such as dashboards. Dashboards will be interactive with the ability to switch the analysis for different cryptocurrencies.

# Chapter 2

## 2 Literature Review

The economics of cryptocurrencies have been the subject of numerous studies, which have focused on topics such as market efficiency, price volatility, and price discovery. The herding of cryptocurrencies is analysed in other publications. Herding in the cryptocurrency market is something that is worth investigating because it has the potential to create an inefficient market that is unable to implement economically sound pricing models. In this study, we will investigate how sentiment analysis and users of social media platforms impact the popularity of cryptocurrencies.

Additionally, this study investigates the mood of cryptocurrency market sentiment by analysing different tweets and posts available on Twitter and Reddit and subsequently classifying them into neutral, positive and negative. The data on these platforms are unstructured but for the most part in text format. Text format is easier to analyse and manipulate than other types of media such as videos, images and audio. The analysis will be carried out using various machine learning algorithms combined with text embedding methods and evaluation metrics.

(Maria Trigka et al, 2022) A study that examines Twitter accounts to forecast the popularity of a cryptocurrency to find its impact. Twitter engagement is largely associated with the number of retweets, followers, likes and the ratings of tweet sentiment to measure popularity. Kendall Correlation Coefficients, Spearman and Pearson were Post hoc techniques employed to provide support for hypotheses concerning the influence and characteristics of users. Valence Aware Dictionary and Sentiment Reasoner, popularly referred to as VADER, was employed to assign a score to the general tone of tweets sent during a specified window of time. To determine which variable is the most important for accurately predicting the popularity of cryptocurrencies in the future, a Granger causality test was carried out. This test examined the statistical importance of a number of characteristics in predicting time series popularity.

The purpose of the study that was published in (Stenqvist, 2017) was to investigate whether or not the sentiment analysis of tweets that were connected to bitcoin could be used as a basis for predicting whether or not the price of bitcoin would rise. The authors of the research paper (Colianni et al,2022) discuss a variety of machine learning end-to-end workflows with the aim of conducting sentiment analysis on Twitter data and identifying market activity in the cryptocurrency Bitcoin. They produce hourly and daily predictions with an accuracy of more

than 90% by making use of a number of supervised learning algorithms. A study that looks at how the prices of cryptocurrencies interact with one another and are regularly correlated with the sentiment values gleaned from Twitter and StockTwits posts is described in (Aste T,2019). The study was conducted to investigate these questions. The authors investigate the possibilities of a specific characteristic structure being considered within a market and inquire about the most valuable cryptocurrencies.

A study conducted by (Abraham et al, 2018) used data from Twitter and Google Trends to forecast price changes for Bitcoin and Ethereum. He found that the volume of tweets, rather than the sentiment of tweets (perpetually positive), is the optimum prediction of price direction. They managed to reliably predict changes in prices by using data from tweets and Google Trends as inputs. People can make more informed decisions regarding the purchasing and selling of Ethereum and cryptocurrencies using this methodology.

((Şaşmaz, E. & Tek, F. B., 2021) finds that twitter tweets are trusted by and used by many investors to guide their everyday bitcoin trading. In this study, the chances of successfully automating cryptocurrency sentiment research were investigated with focus on one alternative cryptocurrency (NEO) and gathered data for this study. The study relied heavily on the data collecting and cleaning processes. To begin, we pulled daily tweets over the past five years that had the hashtags #NEO from Twitter. After collecting tweets, they were screened to exclude any that did not specifically reference NEO. As an example, we took a small sample of tweets and manually assigned positive, negative, and neutral labels to them. We used the labelled data to train and evaluate a Random Forest classifier, and we were able to achieve an accuracy of 77% on the test set. The second part of the research looked at whether the NEO price fluctuated in tandem with the mood of the tweets posted that day. We discovered a favorable relationship between the volume of daily tweets and the pricing of several cryptocurrencies. The information is made public by us.

The goal of this paper (Dulau., 2019) is to examine, identify, and develop a Java solution for the purpose of retrieving the sentiment connected to the cryptocurrencies phenomenon, from the posted content available on particular social platforms. This will be done by analysing the content of the posts, looking for patterns, and developing the solution. The paper proposes exploring, identifying, and developing the solution. Finding out if the sentiment is good, negative, or neutral is a relevant approach for demonstrating human nature perspective on the

much-debated subject of cryptocurrency. This method is selected as a suitable way of presenting the perception of human nature.

Since their inception in 2009, cryptocurrencies have been progressively taking on an increasingly important part in the alternative economies of the world. Nevertheless, having a clear picture of their real effect and performance can be a challenging endeavor. In this work (Nebojsa Horvat et al,2020), an architecture was proposed for processing real-time data and conducting analysis relying upon the Lambda architecture. The design was developed for bitcoin data. The architecture that has been described has the ability to process both continuous and batch data of cryptocurrency blocks and transactions. Additionally, it can perform analysis of a variety of patterns that may be found in an exchange market and blockchain system. The architecture proposed is transposable, which makes it easier to implement components that are only loosely coupled. It also offers a number of advantages for cryptocurrency exploration, including real-time observation of events concerning blockchain and statistical exploration, trends in trading cryptocurrency, and social media events that are related to cryptocurrency reputation.

(Huang et al,2021) forecasts the highly unpredictable price of cryptocurrency by analysing the sentiment of social media. Previous research has analysed the sentiment of social media posts written in English; here, an approach was proposed for the recognition of posts sentiment on Sina Weibo documented in Chinese. They created an end-to-end workflow to gather posts on Weibo, elaborating on the development of a cryptocurrency sentiment dictionary, and suggested an LSTM-based recurrent neural network as a means of predicting future price trend at intervals. Experimentation has shown that the proposed method is superior to the current auto regressive model par excellence by a difference of 18.5% in precision and 15.4% in recall. (Chen, R., & Lazer, M., 2011) investigates the relationship between the content of Twitter feeds and stock market movement. They were interested in determining whether or not the sentiment information contained in these feeds is capable of accurately predicting future price movements and, if so, to what extent. To find a solution to this problem, they created a model, determined how accurate it was, and then test it on real market data by using a fictitious portfolio. According to the results of the analysis, the model is profitable.

(Bo Pang & Lee Lillian, 2008) They were the pioneers in the development of sentiment analysis. The primary objective was to classify written material based not only on its subject matter but

also on its overall attitude; for instance, they wanted to categorize film reviews as either positive or negative. A machine learning algorithm was applied on a database of film reviews, and the results demonstrate that this type of algorithm performs better than those developed by humans. Support vector machines, maximum entropy and Naive Bayes were the machine learning algorithms chosen. When they look at several different factors, they come to the realisation that it is extremely challenging to classify people's feelings. They provide evidence that the bedrock upon which sentiment analysis is built are supervised machine learning algorithms.

(Colianni et al. 2015) set out to demonstrate whether Bitcoin data gathered from Twitter could possibly be employed to create cryptocurrency trading strategies that yield gains. Support vector machines, logistic regression and Naive Bayes were the machine learning algorithms chosen. The classifiers were put through their paces utilising two different data sets. Naive Bayes allowed them to increase their accuracy to 95% per day and 76.23% per hour. They tried again with the textprocessing.com API to assign each tweet a score between 0 and 1 for positivity, 0 for neutrality, and 1 for negativity. Logistic Regression was able to classify data achieving a daily accuracy of 86% & an hourly accuracy of 98.58% using this method.

Similar to (Dickinson & Hu, 2015), (Raheman et al,. 2022) focuses on the correlation of the movements in the sentiments metrics with the changes in the prices of Bitcoin, a popular cryptocurrency based on Blockchain Technology. They utilized Natural Language Processing and interpretable artificial intelligence methods over non-explainable and non-interpretable ones. For data collection, they fetched 100000 news items from the Twitter and Reddit posts for the period of July to December 2021 using the official Twitter and Reddit APIs. Since the data was unlabeled, manual labeling was done by two independent researchers, and the ground truth was selected as the average of both. The sentimental analysis was modeled and experimented with 21 different models famous in the NLP domain such as Afinn, Vader, TextBlob, GoogleNLP, AWS, Aigents, and BERT-based models. Major challenges that they observed were the use of sarcasm, idioms, negations, and non-text data. The model aigents performed best out of all models.

### Results obtained in previous works related to cryptocurrencies

- **Paper:-** (Raheman et. al., 2022)

- **Dataset Used:-** 0.1 million tweets and Reddit posts across 77 public Twitter timelines and Reddit subreddits over a six-month period labeled by two experts in the range of -1.0 to 1.0 for sentiment classification of Bitcoin cryptocurrency.
- **Results Obtained:-** The average Pearson correlation between sentiment metrics "predicted" by respective models and "ground truth" provided by humans. As illustrated the "out of the box" Aigents model "aigents" possess a correlation of -0.33, and after fine tuning. "agents" possess a correlation of -0.57. "ensemble (all)" equate to average metrics across all models, and "ensemble (top 3)" equate to the average of the best three models (aigents, aigents and finBERT). The following figure shows the comparison of different models evaluated in the given research paper:-

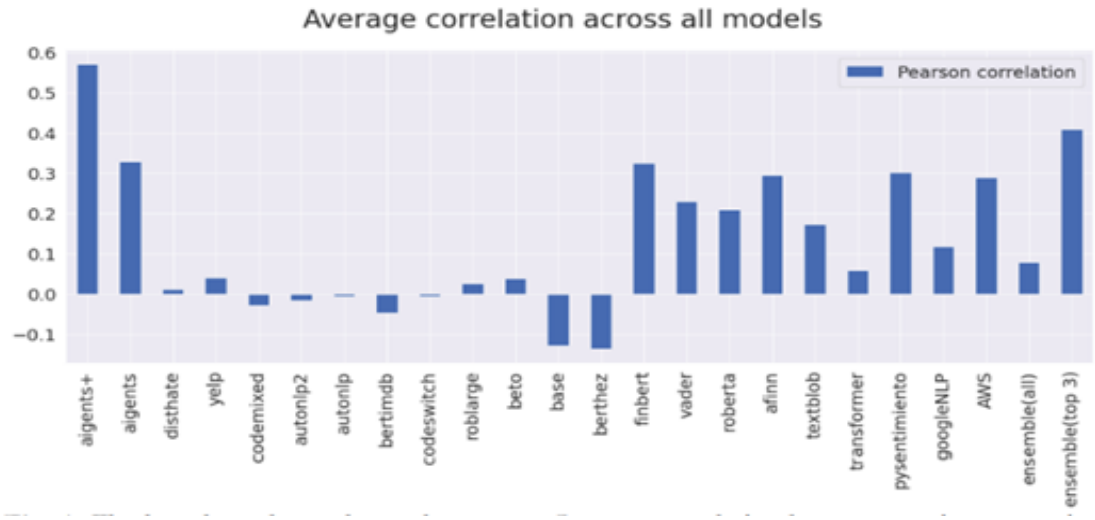


Figure 2.1 Pearson correlation between sentiment metrics

- **Paper:-** (Ider et. al., 2022)

- **Dataset Used:-** 4780 news articles, tweets, and Reddit posts labeled by three experts for sentiment classification of Bitcoin cryptocurrency. Weak labels were generated for a large corpus using BERT models and zero-shot classification as positive, negative, or neutral.
- **Results Obtained:-** The following figure shows the comparison of different models evaluated in the given research paper:-

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| BART 0Shot-TC | 0.790 | 0.775 | 0.771 | 0.773 |
| FinBERT | 0.822 | 0.787 | 0.823 | 0.805 |
| BERT-Frozen | 0.688 | 0.622 | 0.583 | 0.602 |
| BERT-Unfrozen | 0.789 | 0.769 | 0.754 | 0.761 |
| Ensemble 0B-NB | 0.795 | 0.816 | 0.717 | 0.763 |
| Ensemble 0BB | 0.787 | 0.784 | 0.737 | 0.760 |

Figure 2.2 Sentiment classification metrics

# Chapter 3

## 3 Research Methodology

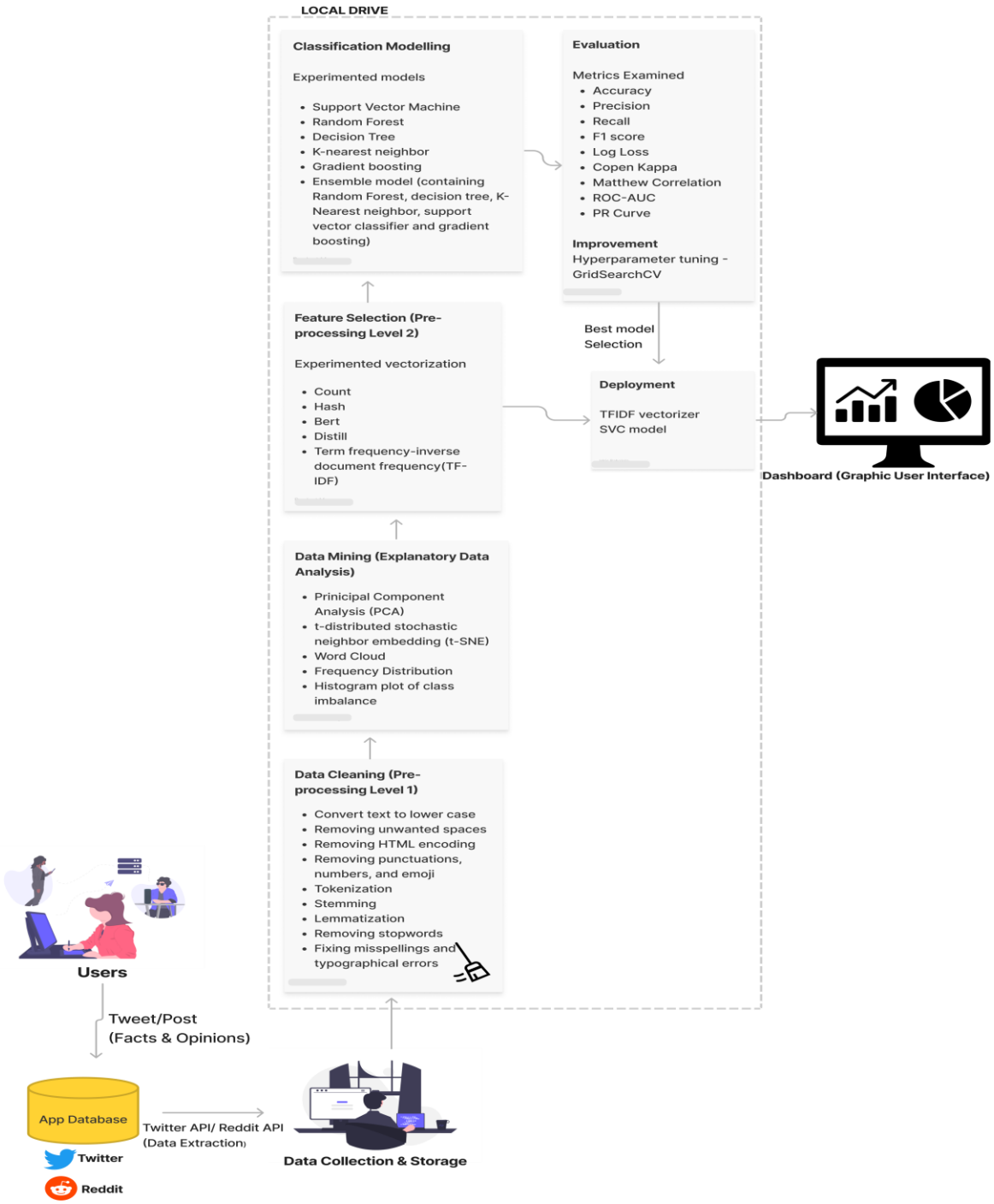


Figure 3.1 System Architecture

In this section, the architecture of the cryptocurrency market sentiment analysis developed in Figure 3.1 is explained along with the description of each component. Roughly, the platform fetches the data through multiple data sources and stores it in a repository, performs NLP on the data from the stored repository, and provides sentiment analysis for desired parameters. The detailed components of the system are as follows: -

## 3.1 Data Collection

The process of gathering data related to the problem definition and research use case is known as data collecting. Data collection and storing must be done in a way that is relevant to the specific area of interest in order to produce effective and efficient machine learning solutions.

For this research, cryptocurrency sentiments and opinions has been collected from two social media platforms namely, Twitter and Reddit. The twitter data was collected using the Twitter API key and tweepy python library. The retrieved tweet targeted cryptocurrency topics by specifying hashtags (i.e #Bitcoin, #Ethereum and #Litecoin), time frame of required data, column name and record limit. A total of 3000 tweets was collected setting the timeframe between 28$^{th}$ July 2022 and 3$^{rd}$ August 2022 with column name as *Tweets*. The unstructured text data collected was stored in a pandas data frame and exported to a CSV file. In addition, the Reddit data was collected using the Twitter API key and praw python library. The retrieved tweet focused on hot news around cryptocurrency topics by specifying hashtags (i.e #Bitcoin, #Ethereum, #Litecoin) and record limit. The praw library does not support data retrieval for a specific period at the time of data collection, it only supports fixed time such as daily, monthly, quarterly. A total of 3000 daily post was collected on 28$^{th}$ July 2022 with column name as *Title*, stored in a dataframe and exported into a CSV file.

## 3.2 Data Annotation

Data annotation is the technique through which we label data to make objects, text, input recognizable by machines. There are two approaches commonly used to annotate, one entails the use of human involvement and the other uses programming script. (Sasmaz, Emre & Tek, F.. 2021) in a similar work investigated the two approaches and presented that the sentiment analyzer is prone to misclassifying the labels and overfitting. For this work, the manual labelling

was adopted and given precedence over the sentiment analyzer. The twitter and reddit dataset were accessed in Microsoft Excel, the entire records were evaluated intuitively and assigned a label as positive, negative and neutral on a new column depending on the emotions present in the record. The labelled datasets were observed to contain duplicate rows, 240 and 12 records were removed from the twitter and reddit data respectively using the drop function in pandas. The remaining labelled records were counted using the pandas count method - 2323 positive sentiments, 381 negative sentiments and 56 neutral sentiments for twitter data with 2155 positive sentiments, 691 negative sentiments and 94 neutral sentiments for the reddit data.

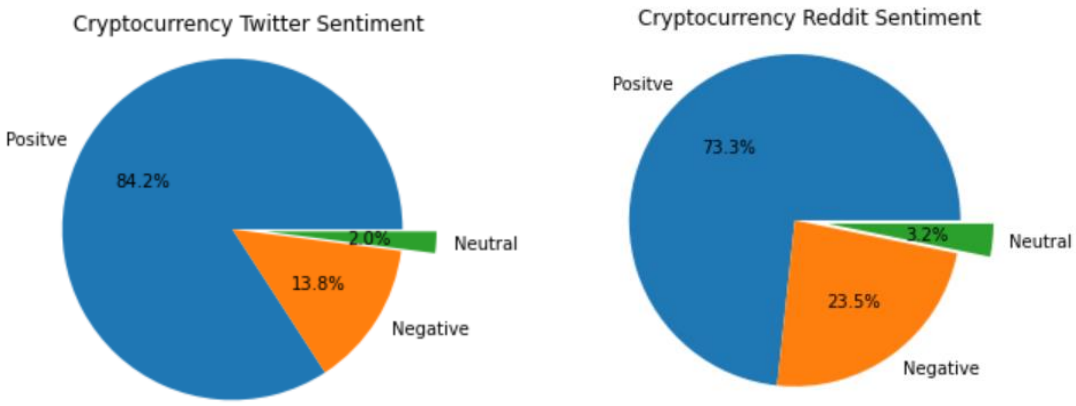


Figure 3.2 Labelled data sentiment distribution

Based on the approach adopted, neutrality is addressed in different ways in sentiment classification. The words neutrality score is used in lexicon-based approaches to either detect objective opinions (Ding and Liu, 2008) or screen them out, allowing algorithms to target words with positive and negative and positive valence (Taboada et al, 2010). When statistical methods are utilised, however, the way neutrals are handled vary substantially. Some studies indicate that the objective sentences in the text are less significant, therefore they take them out and concentrate solely on the subjectivity in order to optimize the binary categorization (Bo Pang and Lillian Lee, 2002).

In other instances, they employ hierarchical classification, which determines neutrality first and then sentiment polarity (Wilson et al, 2005). It is a technique which aims at classifying text documents into classes that are organized into a hierarchy. Unlike flat text classification, hierarchical text classification considers the interrelationships among classes and allows for organizing documents into a natural hierarchical structure. This technique's filtering approach decreases gradually the original dataset in relation to the contextual polarity and common

terms in a document. As a result, textual evidence inferring neutral feelings becomes less meaningful than data inferring stronger sentiments on opposite ends of the polarities (Li, 2015).

Most research studies on sentiment analysis that employ statistical methods exclude the neutral class, assuming that the neutral texts are close to the boundary of the binary classifier. Furthermore, it is considered that neutral texts have less to convey than those with evident negative or positive emotions. For this work, the positive and negative sentiment were used, hence dropping the neutral sentiments.

## 3.3 Data Cleaning

Data annotation step is followed by the data cleaning. The data cleaning phase involves the recognition of variables and rectifying anomalies in the dataset which could have a negative effect on the predictive model. For this work, a custom function was created to match the required pattern inside the text using regex library and removes it. Another custom function was created to split the text into words and replace the word with the value mapped in the dictionary if present.

Three dictionaries retrieved from the NLTK module and pre-loaded to memory using the pickle library. One of the dictionaries was containing commonly used words with apostrophe such as *"wasn't, couldn't"* which in turn maps it to *"was not, could not"*. The second was used to map commonly used short-words and abbreviation such as *"asl"* with its full-text *"age, sex, location"*. The third was used to match certain special character combinations with commonly used emotion icons. Table 3.1 outlines the different text cleaning steps carried out on the data.

| | Process | Tool |
|---|---|---|
| **Basic Cleaning** | Converting text to lower case | Custom function/ Regex library |
| | Removing unwanted spaces | Custom function/ Regex library |
| | Removing HTML encoding | HTML Parser |
| | Removing punctuations, numbers | Custom function/ Regex library |
| **Intermediate Cleaning** | Tokenization | WordTokenize |
| | Stemming | Porter Stemmer in NLTK |
| | Lemmatization | WordNetLemmatizer in NLTK |
| | Removing stopwords | |
| **Advanced Cleaning** | Fixing misspellings and typographical errors | Textblob library |

Table 3.1 Text cleaning steps

After the cleaning steps were completed, the cleaned text was utilized in the next phase.

## 3.4 Feature Engineering and Selection

When training a model, it is critical to identify unique characteristics and features for optimizing the desire output and performance. This step comprises creating new features as needed and selecting appropriate features based on the problem requirement.

As part of this work, the feature engineering approach facilitated creation of five (5) new columns namely, '*TweetsClean*', '*TweetsTokens*', '*TweetsTokenClean*', '*TweetStemmed*' *and* '*TweetLammatized*'. This is illustrated in Figure 3.3 below, hence increasing the dimensionality of the existing features. '*TweetsTokens*' feature was generated by tokenizing the cleaned text data. '*TweetStemmed*' *and* '*TweetLammatized*' were stemmed and lemmatized features respectively generated from the '*TweetsToken*' feature after eliminating the stopwords.

| | Title | Sentiment | TweetsClean | TweetsTokens | TweetsTokensClean | TweetsStemmed | TweetsLemmatized |
|---|---|---|---|---|---|---|---|
| 0 | Brazil Are Looking To Ethereum To Improve Thei... | 1 | brazil are looking to ethereum to improve thei... | [brazil, are, looking, to, ethereum, to, impro... | [brazil, looking, ethereum, improve, voting, s... | brazil look ethereum improv vote system | brazil looking ethereum improve voting system |
| 1 | BTC and LTC - Friends forever. | 1 | btc and ltc friends forever | [btc, and, ltc, friends, forever] | [btc, ltc, friends, forever] | btc ltc friend forev | btc ltc friend forever |
| 2 | Google you need to fix this :( | 1 | google you need to fix this sad | [google, you, need, to, fix, this, sad] | [google, need, fix, sad] | googl need fix sad | google need fix sad |
| 3 | 'India today is making the move to outlaw Bitc... | 1 | india today is making the move to outlaw bitco... | [india, today, is, making, the, move, to, outl... | [india, today, making, move, outlaw, bitcoin, ... | india today make move outlaw bitcoin surveil c... | india today making move outlaw bitcoin surveil... |
| 4 | Salaries in Argentina to be paid in Bitcoin, b... | 1 | salaries in argentina to be paid in bitcoin bi... | [salaries, in, argentina, to, be, paid, in, bi... | [salaries, argentina, paid, bitcoin, bill, pro... | salari argentina paid bitcoin bill propos approv | salary argentina paid bitcoin bill proposed ap... |

Fig 3.3 Dataframe of data after feature engineering

The stemmed and lemmatized features of the cleaned data were considered for the feature selection and importance which was done as part of data pre-processing to understand the independent features with significant parameters and weight associated with the prediction or target variable(s). Stemming and lemmatization are vital in improving an information extraction system's recall ability (Kanis and Sko-rkovsk a, 2010; Kettunen et al., 2005). It was observed in the investigation of (Balakrishnan & Ethel, 2014) that lemmatization gave better precision than stemming, however the variations are negligible. Generally, the results suggested language modelling techniques increase text extraction, with lemmatization yielding the best outcomes.

For this study, the lemmatized feature was adopted for the modelling. Lemmatization enables an algorithm to determine meaning of each word properly; the process optimises the data accessible to it more precisely and, unlike stemming, avoids eliminating a large number of words due to shallow, imperfect filtering. This substantially increases a system's comprehension of contexts and capacity to connect the meaning of one text to another. Furthermore, lemmatization requires more processing resources and time since it must derive the meaning of the word rather than simply cutting it off. However, because the dataset in this work is not particularly large, we chose to proceed with lemmatization (Christopher D et al, 2008).

## 3.5 Data analysis and Visualization

Data analysis is a vital step of performing exploratory studies on data to discover correlations, detect anomalies, testable theories, and cross-check assumptions using summaries of statistical results and graphical illustration. Data visualisation is a broad phrase that refers to any endeavour to assist people grasp the importance of data by presenting it in a visual format. Text visualization helps to recognize undetected patterns, trends and correlations.

Presentation of results from applied text analytics is significantly less straightforward compared to numeric data. We have to find the most informative features such as words, phrases, fragments. It is easy to create visualisations, but it is far more challenging to make good ones. For this work, high-dimensional embeddings was projected onto the 2-d graphs utilizing T-SNE from the Yellow Brick library, a dimensionality-reduction methods. The clusters are mapped on the projected graph using the labels. A Principal Component Analysis (PCA) was carried out to reduce the dimensionality of the features using the scikit learn library and matplotlib was employed for the visualization. The word cloud of the corpus was generated with Matplotlib library in python. Word clouds provide a visualization of words. It shows the most popular phrases and words based on their frequency and relevance. The frequency distribution was visualized with the NLTK frequency distribution function. Frequency distribution is useful in counting how many times a specific word in repeated with a text. The histogram of classes was also visualized to check for class imbalance. A class imbalance was observed between the proportion of positive sentiment and negative sentiment, oversampling techniques was applied to the corpus to remediate the problem using the Synthetic Minority Over-sampling Technique (SMOTE) function in Imbalanced-learn library. The SMOTE over-sampling method involves generating "synthesised" instances of the minority class instead of over-sampling using substitution. This approach has been shown to be effective in recognizing handwritten characters. It was employed by (Ha & Bunke, 1997) to generate additional training data by carrying out certain actions on real data.

## 3.6 Feature Encoding/Data Encoding

Encoding comprises rendering categorical data into numerical for the purpose of improving model performance. The data transformation implemented on the dataset were five text embedding technique (namely: count, tfidf, hash, bert, distil) to compare results, present findings and choose the best. The text embedding technique were tested with an ensemble model (containing decision tree, support vector classifier, gradient boosting, K-Nearest neighbor and Random Forest). A custom function was created to display the metric as illustrated in Table 3.2 below. The F1 metric was used to determine the best where distil and bert outperformed all. But the performance difference with tfidf was less and since tfidf is extremely more efficient on computational resources and time (Senbel, S, 2021).

| S/N | Vectorizer | ROC-AUC | PR AUC | F1 Score | Copen Kappa | Matthew CC | Log Loss |
|---|---|---|---|---|---|---|---|
| 1 | Count | 0.712551 | 0.886150 | 0.806394 | 0.313672 | 0.345226 | 0.565700 |
| 2 | TFIDF | 0.690992 | 0.876965 | 0.868435 | 0.355933 | 0.358053 | 0.451015 |
| 3 | Hash | 0.695759 | 0.878623 | 0.873922 | 0.370361 | 0.371741 | 0.443332 |
| 4 | Bert | 0.750844 | 0.899274 | 0.878237 | 0.443847 | 0.451531 | 0.420087 |
| 5 | Distill | 0.737818 | 0.894244 | 0.878556 | 0.429619 | 0.434817 | 0.418620 |

Table 3.2

For this work the tfidf technique was adopted. Also in the experimentations, tfidf was found to be more efficient.

## 3.7 Classification Modelling

The data modelling approaches varies depending on the dataset and the objective of the research. Since the dependent variable is present, it is a supervised machine learning task. The task is considered as a regression problem if the prediction or dependent variable is of numeric data type, and as a classification problem if the prediction or dependent variable is of categorical data type. Since the target variable Sentiment is present in this dataset, modelling may be conducted both supervised (using the dependent variable) and unsupervised (by not considering the dependent variable).

For this supervised classification problem, decision tree, support vector classifier, gradient boosting, K-Nearest neighbor and Random Forest classification technique were employed. The Twitter dataset appeared to be small as the model was not generalizing properly. The twitter and reddit dataset were combined using the concatenation function in pandas to enable the models learn as many features as possible to improve the model performance since the data were collected from the same time for the same cryptocurrencies (BTC, ETH, LTC). A total of 5799 features were used in this classification modelling with an 80:20 data split for training data and test data respectively. TFIDF vectorizer was tested on the individual model. A custom function was used to generate an evaluation metric as illustrated in Table 3.3.

| S/N | Model | ROC-AUC | PR AUC | F1 Score | Copen Kappa | Matthew CC | Log Loss |
|---|---|---|---|---|---|---|---|
| 1 | RF | 0.694248 | 0.877941 | 0.882979 | 0.383026 | 0.383123 | 0.450812 |
| 2 | DT | 0.689128 | 0.876446 | 0.857300 | 0.338388 | 0.343263 | 6.308887 |
| 3 | KNN | 0.615957 | 0.855268 | 0.436989 | 0.105949 | 0.211486 | 15.134003 |
| 4 | SVC | 0.652895 | 0.863013 | 0.915254 | 0.387395 | 0.428018 | 0.433369 |
| 5 | GBC | 0.737935 | 0.896388 | 0.798304 | 0.334395 | 0.379702 | 0.495134 |
| 6 | Ensemble | 0.698466 | 0.879665 | 0.870461 | 0.368635 | 0.371036 | 0.450099 |

Table 3.3 Evaluation metric to select best model

The Receiver Operator Characteristic (ROC) curve is a famous measure for evaluating binary classification tasks. It is a probability curve that shows the True Positive Rate versus the False Positive Rate at varying threshold values, separating between the 'signal' & 'noise.' The Area Under the Curve (AUC) is an indicator of a classifier's capacity to differentiate between classes and is utilized to summarise the ROC curve. However, in the event of unbalanced datasets, ROC curves may present an overly optimistic assessment of performance.

Precision is a measure that counts the number of positive forecasts that are correct. Recall is a measure that counts the number of correct positive predictions produced out of all possible positive predictions. Both precision and recall are fixated on the positive class (the minority class) and are unfixated with the true negatives (majority class), thereby making it ideal for evaluating a classifier's performance on the minority class. The PR curve's emphasis on the minority class makes it a useful test for imbalanced binary classifier models (Davis, J. & Goadrich, M.,2006).

The Precision-Recall AUC works similarly to the ROC-AUC in the way it summarises the curve with a spectrum of threshold values into a single score. The score can be utilised to compare multiple models on a binary classification use case, with 1.00 representing a model with perfect skill.

Cohen's kappa coefficient is a measure of inter-rater consensus for qualitative variables. Since k considers the agreement occuring by chance, it is typically perceived to be a more robust measure than simple percent agreement computation. Cohen's kappa calculates the degree of agreement between two raters that classify N things into C mutually exclusive groups. The Matthews Correlation Coefficient, just like the F1 score, is a highly dependable and well-rounded metric (Chicco, D., & Jurman, G,2020).

In this study, the F1 score was used to evaluate the performance of classification algorithms. This is primarily due to its capacity to deliver accurate results on both balanced and imbalanced datasets. However, because it takes into account both the precision and recall abilities of the model, it serves as a comprehensive evaluator of model performance.

The SVC model outperformed all models in every evaluation metric in this experimentation. The experimental findings reveal that SVM consistently performs well on text categorization tasks, exceeding previous approaches appreciably. SVMs, due to their capability to generalize effectively in large dimensional feature spaces, it eliminates the necessity for feature selection, making text categorization more easier to apply. The robustness of SVMs over traditional techniques is another benefit. SVMs perform well across all experiments, minimizing catastrophic failure as was seen with some tasks using conventional approaches. Furthermore, since SVMs can dynamically determine appropriate configurations, no parameter tuning is necessary. SVMs are a very attractive approach and user-friendly method for learning text classifiers from examples as a result of all of this (Joachims, 1998).

## 3.8 Hyperparameter Optimization

Grid Search is the ideal hyperparameter optimizer (Zoller M-A, 2019). Iterating endlessly over all potential combinations of hyperparameter adjustment is the grid search modus operandi (Hutter F et al, 2019). The method of grid search is frequently used to find the ideal model setup parameters. It takes a simple, use-case-related approach to choosing the appropriate hyperparameter of interest (Mesafint et al, 2021). The Support Vector classifier model was enhanced for this work utilising Scikit-GridSearchCV Learn's for hyperparameter tuning.

## 3.9 Graphic User Interface

Rather than utilizing conventional command-line text-based navigation software, the graphical user interface, or GUI, enables users to interact with the programme using graphical indicators and audio indicators. GUI varies depending on the type of interaction. On desktops and laptops, websites are typically used for interface, mobile software application are used for phones, and dashboards are deployed for solutions particular to data science.

The technology utilised to create these interfaces change depending on the type of interface. Websites are created using tools such as CSS, Javascript, HTML whereas mobile apps are created using tools such as Dart, kotlin and flutter. Both of these technology frameworks can be

employed to create dashboards, however python programmers can create dashboards using streamlit, plotly, flask, django etc. The dashboard for this project is created using streamlit. With Streamlit, an open-source application framework, anyone can quickly create online applications with no prior front-end development knowledge. It offers a free Python library that has comprehensive documentation and boilerplate code. The produced streamlit dashboard can easily be connected with the machine learning model made in the preceding steps. There are two methods for integrating; either by pre-training the model on the data, or by dumping it using the pickle library, then loading it when the dashboard is launched and using it for prediction. The alternative is to just train it when a prediction is made. When training a model takes a lot of time, like when training RNNs or LSTMs over a large number of epochs, the first strategy is helpful. The second approach performs best when many models are required, but not all at once, and minimal training time. The first method is implemented to develop and integrate the machine learning model with the GUI since the SVC model is chosen as the study's final model. The dashboard design was made intuitive and characterized by a number of features:

- Using a custom function for real-time collection of cryptocurrency data via Yahoo finance API in python and displaying a trend of opening price of previous last one week to the user.

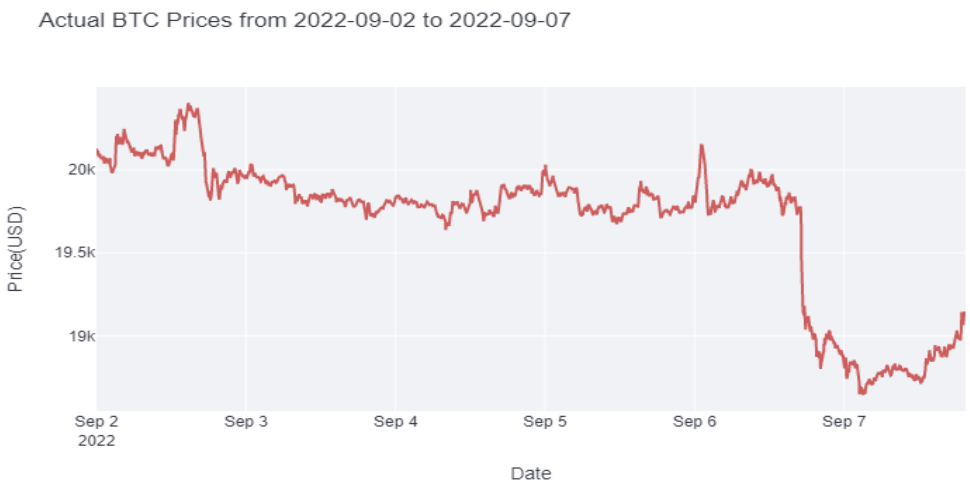


Figure 3.4 Cryptocurrency price plot

- the dashboard allows the user to scrap latest tweet data of the last one week for analysis

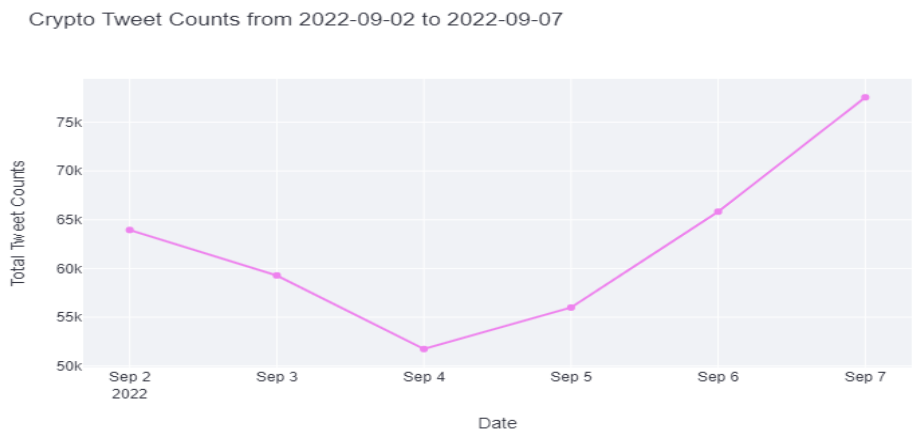


Fig 3.5 Cryptocurrency tweet count plot

- a custom function is used to compute positive sentiment ratio and a trend is displayed to the user.

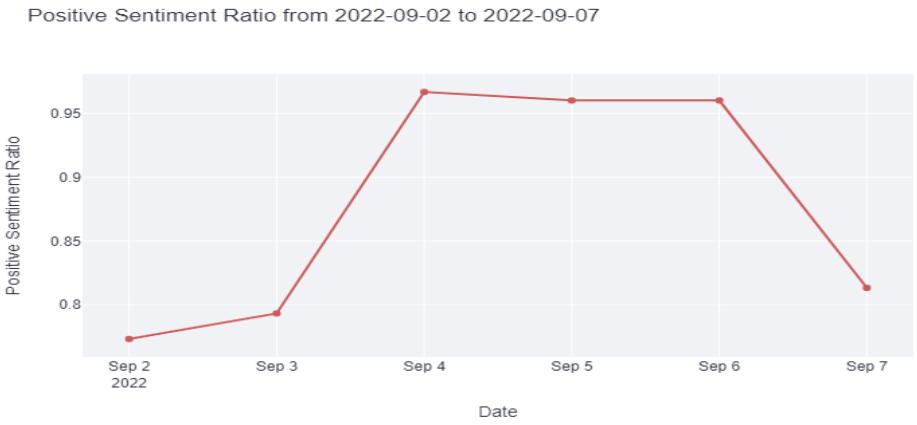


Fig 3.6 Positive sentiment plot over specific time

- a wordcloud plot trend is generated from the scrapped twitter data for a six-day interval. The data is preprocessed and cleaned before displaying to users real-time.


Figure 3.7 Word cloud plot

- a pie donut trend of the mood of the market is created. This is achieved by classifying sentiments into positive and negative and displayed to users real-time in a plot.

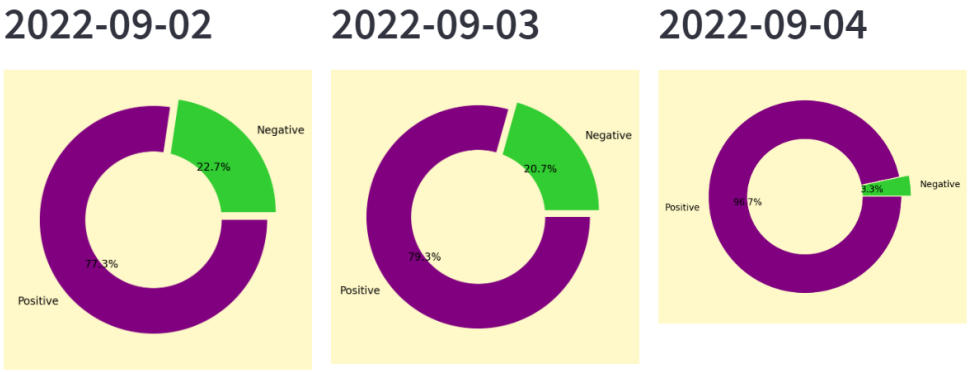


Figure 3.8 Pie count plot

- a text field was designed to collect real-time text input from the user. The collected input is classified with the trained model and output a sentiment result.

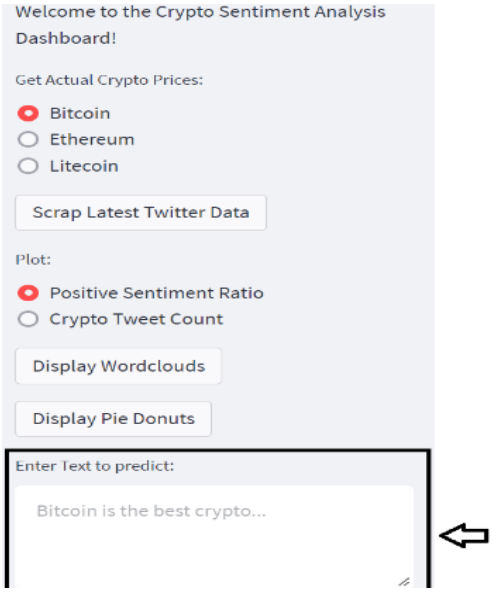


Fig 3.9 Textbox for reading collecting input for the user

Finally, the user interactive GUI Dashboard was deployed to cloud using streamlit. The cloud link to the dashboard is https://eidreiz01-deployment-app-103zqn.streamlitapp.com/ . It allows anyone to access the functionalities of the GUI dashboard on the cloud.

# Chapter 4

## 4 Results

### Evaluation

Prior to development, machine learning models must be tested by a sufficient number of parameters (Xie Y et al, 2018). A high accuracy score might give the impression that the model is performing very well. Accuracy, by itself, cannot reveal which class was incorrectly predicted. The scikit learn package was used to assess the performance of the classification model using different metrics which includes accuracy, F1 score, precision, recall, Cohen Kappa, log loss, ROC-AUC and Matthew Correlation Coefficient (MCC).

For this work, the ensemble model was used to test 5 vectorizations techniques and the best one chosen. A heatmap confusion matrix was plotted using seaborn library as shown in Figure 4.1, Figure 4.2, Figure 4.3, Figure 4.4 and Figure 4.5.

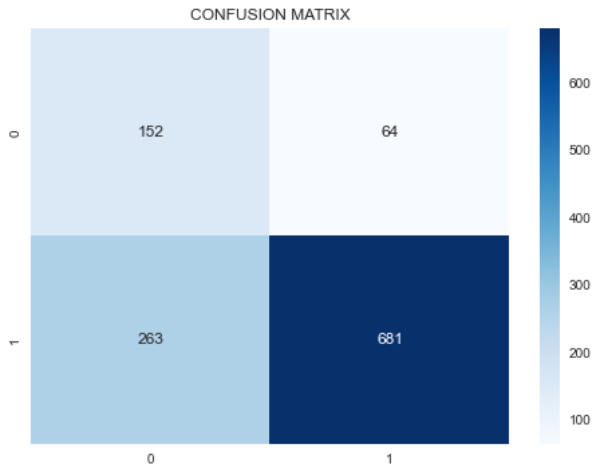


Fig 4.1- Count

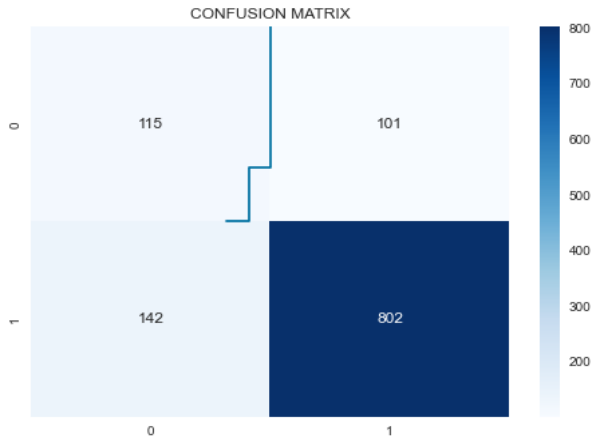


Fig 4.2- TFIDF

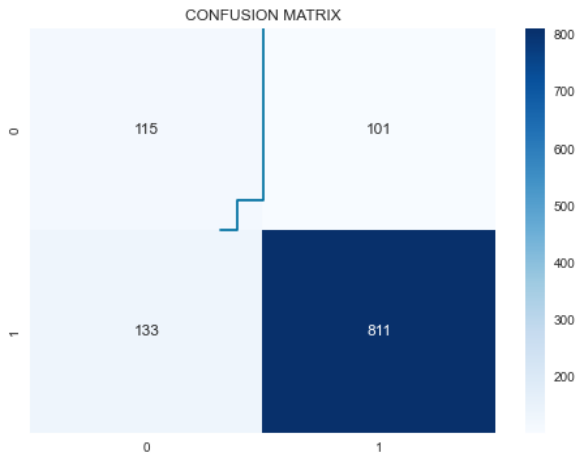


Figure 4.3 - Hash

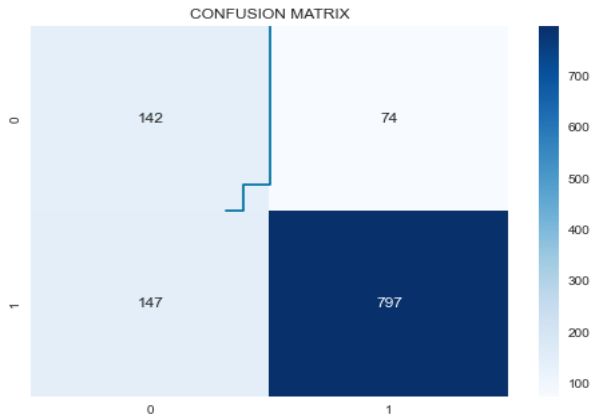


Figure 4.4 -Bert

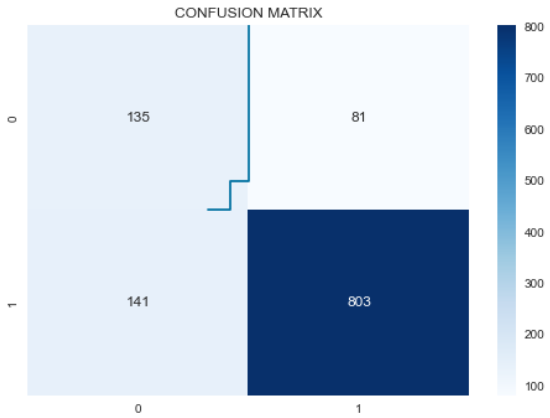


Fig 4.5- Distill

The individual models (RF, DT, KNN, SVC and GBC) and ensemble model were tested with TFIDF the selected vectorization approach. The evaluation metrics are illustrated in Table 4.1 and 4.2 depicting the results. The KNN model had a poor accuracy and SVC had a good accuracy which appeared to have generalized properly.

| Model | Target | Precision | Recall | F1 Score | Accuracy (%) |
|---|---|---|---|---|---|
| SVC | Positive Sentiment | 0.86 | 0.97 | 0.92 | 85 |
| | Negative Sentiment | 0.73 | 0.33 | 0.46 | |
| Random Forest | Positive Sentiment | 0.89 | 0.88 | 0.88 | 81 |
| | Negative Sentiment | 0.49 | 0.51 | 0.50 | |
| Decision Tree | Positive Sentiment | 0.89 | 0.83 | 0.86 | 78 |
| | Negative Sentiment | 0.42 | 0.55 | 0.48 | |
| KNN | Positive Sentiment | 0.96 | 0.28 | 0.44 | 41 |
| | Negative Sentiment | 0.23 | 0.95 | 0.37 | |
| GBC | Positive Sentiment | 0.93 | 0.70 | 0.80 | 71 |
| | Negative Sentiment | 0.37 | 0.78 | 0.50 | |
| Ensemble | Positive Sentiment | 0.89 | 0.85 | 0.87 | 79 |
| | Negative Sentiment | 0.46 | 0.55 | 0.50 | |

Table 4.1: Classification report of all models tested

| S/N | Model | ROC-AUC | PR AUC | F1 Score | Copen Kappa | Matthew CC | Log Loss |
|---|---|---|---|---|---|---|---|
| 1 | RF | 0.694248 | 0.877941 | 0.882979 | 0.383026 | 0.383123 | 0.450812 |
| 2 | DT | 0.689128 | 0.876446 | 0.857300 | 0.338388 | 0.343263 | 6.308887 |
| 3 | KNN | 0.615957 | 0.855268 | 0.436989 | 0.105949 | 0.211486 | 15.134003 |
| 4 | SVC | 0.652895 | 0.863013 | 0.915254 | 0.387395 | 0.428018 | 0.433369 |
| 5 | GBC | 0.737935 | 0.896388 | 0.798304 | 0.334395 | 0.379702 | 0.495134 |
| 6 | Ensemble | 0.698466 | 0.879665 | 0.870461 | 0.368635 | 0.371036 | 0.450099 |

Table 4.2: Evaluation Metrics of all models tested

The best model was selected to be SVC. The overall accuracy of the Support Vector model was computed to be 85% with a poor recall and F1 score as shown in table 4.3. To improve these metrics, hyperparameter optimization was employed using GridSearchCV in scikit learn. GridSearchCV is particularly useful when tuning multiple hyperparameters, the best parameters for this work are captured in table 4.4 below. The result had slight improvement in both metrics and the accuracy score did not improve as illustrated in table 4.3. By hyperparameter tuning, five (5) folds cross validation on a stratified K-fold was implemented during grid search to ensure the model does not overfit and maintains its capability to generalise well on data and the detection rate (recall) for our minority class was optimized significantly. Recall is a more important metric in imbalanced datasets because it tells us how well our model can detect the minority class from the dataset (Weng, Cheng & Poon, Josiah; 2008).

| **Evaluation without Hyperparamater tuning** | | | | **Evaluation with Hyperparamater tuning** | | | |
|---|---|---|---|---|---|---|---|
| **Overall Accuracy: 85%** | | | | **Overall Accuracy: 85%** | | | |
| Target | Precision | Recall | F1 Score | Target | Precision | Recall | F1 Score |
| Positive Sentiment | 0.86 | 0.97 | 0.92 | Positive Sentiment | 0.87 | 0.96 | 0.91 |
| Negative Sentiment | 0.73 | 0.33 | 0.46 | Negative Sentiment | 0.69 | 0.37 | 0.48 |

Table 4.3: Evaluation Metrics Table for Support Vector Classifier

| Parameters | Best Parameter |
|---|---|
| Kernels | 'rbf' |
| Regularization parameter, C | 25 |

Table 4.4: Support Vector Classifier- Best parameters for tuning

To visualize clustering of the data we use vector embedding method on the "Twitter Token Clean" feature and get the result as shown in Figure 4.6 for Twitter data and Figure 4.7 for Reddit data respectively.

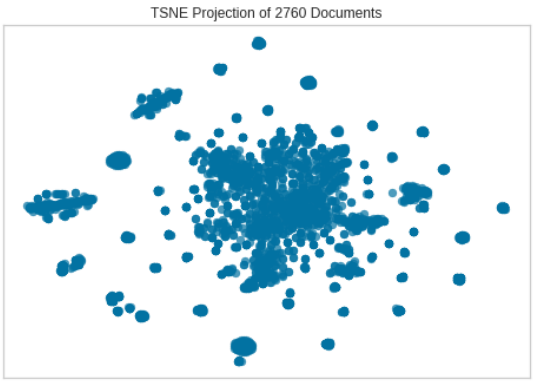


Figure 4.6 T-SNE Twitter data

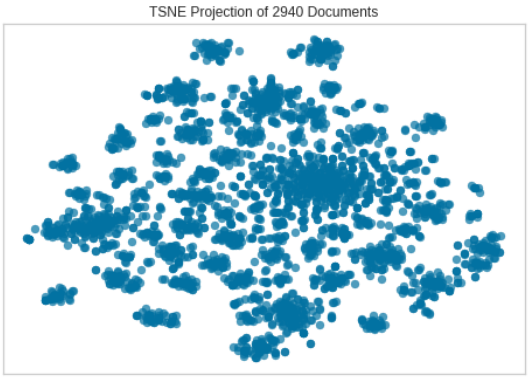


Figure 4.7 T-SNE Reddit Data

A visualization of the clustering of the combined dataset represented in Figure 4.8.

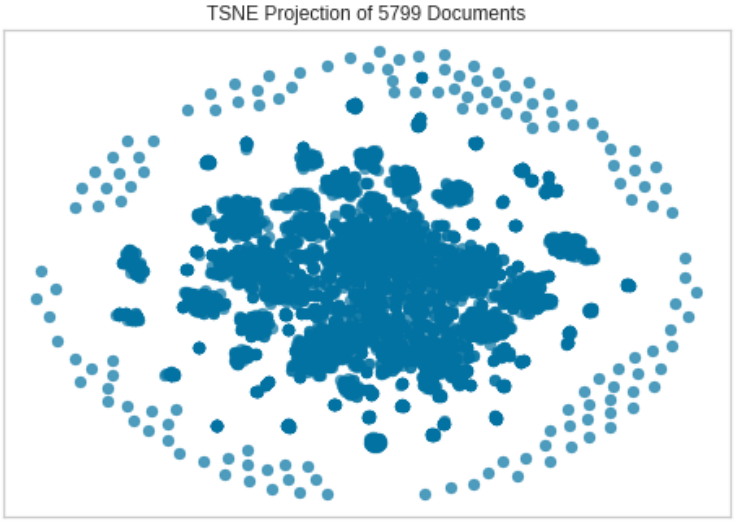


Fig 4.8 T-SNE Twitter & Reddit Data combined

Plotting class imbalance using histogram we get Figure 4.9 for Twitter data and Figure 4.10 for Reddit data.

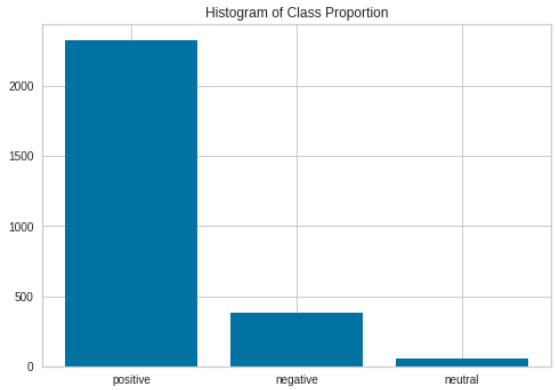


Figure 4.9 Twitter data class imbalance

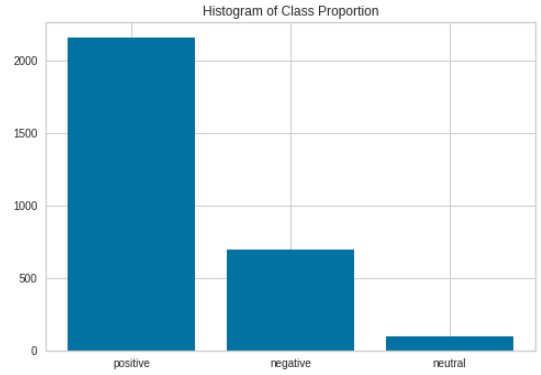


Figure 4.10 Reddit data class imbalance

A plot of class imbalance was carried out using histogram after combining the twitter and reddit dataset as shown in Figure 4.11. A count of 4721 positive sentiment was obtained and 1078 negative after dropping the neutral class.

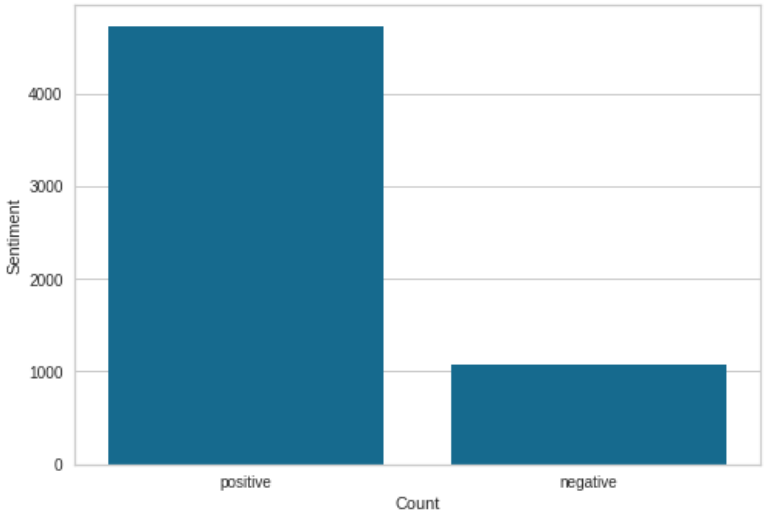


Figure 4.11 Imbalanced class. Combined dataset

After using the SMOTE object to eliminate the class imbalance we get Figure 4.12 for Twitter data and Figure 4.13 for Reddit data.

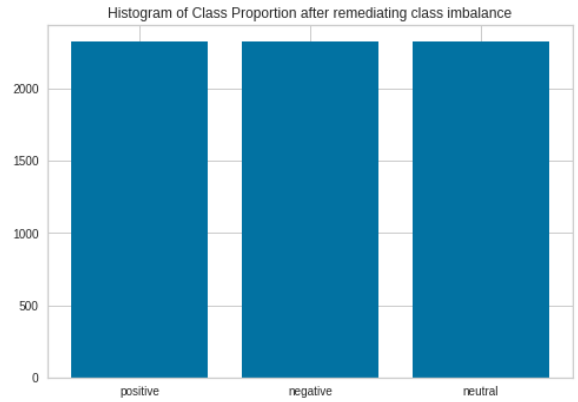


Figure 4.12 Balanced Class Twitter data

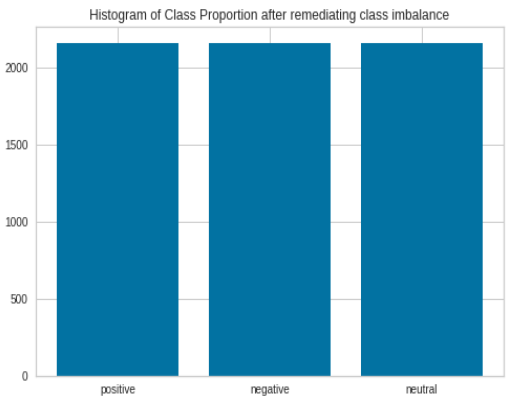


Figure 4.13 Balanced class Reddit data

The word cloud visualization of the Twitter and Reddit dataset combined illustrated in Figure 4.14.

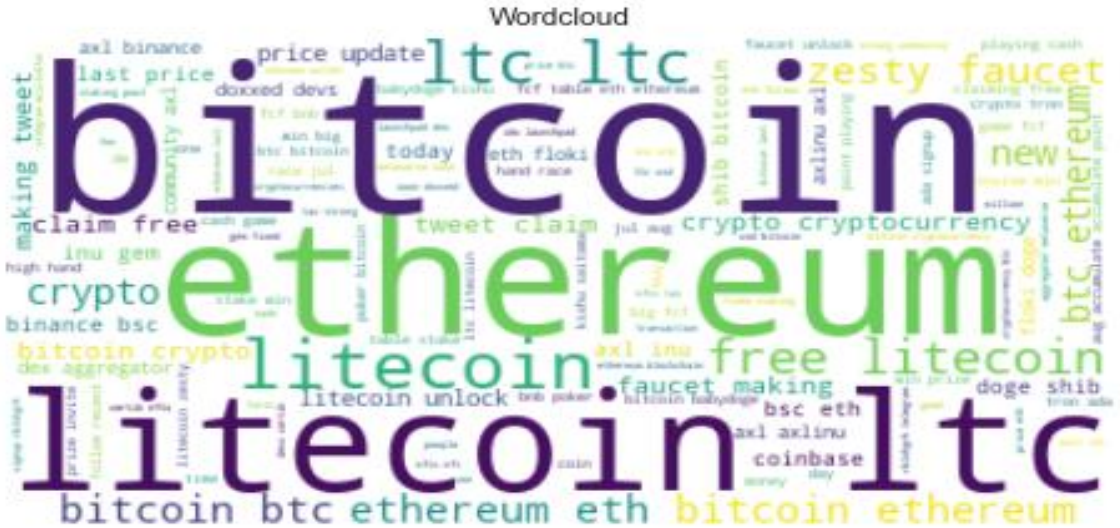


Figure 4.14 Word Cloud Plot

A frequency distribution of fifty (50) words was plotted in Figure 4.15, the most occurrent words in ascending order were observed to be ‘nft’, ‘binance,’, ‘price’, ‘cryptocurrency’, ‘Litecoin’, ‘Ethereum’, ‘Bitcoin’, ranging from about 500-2500 word count.

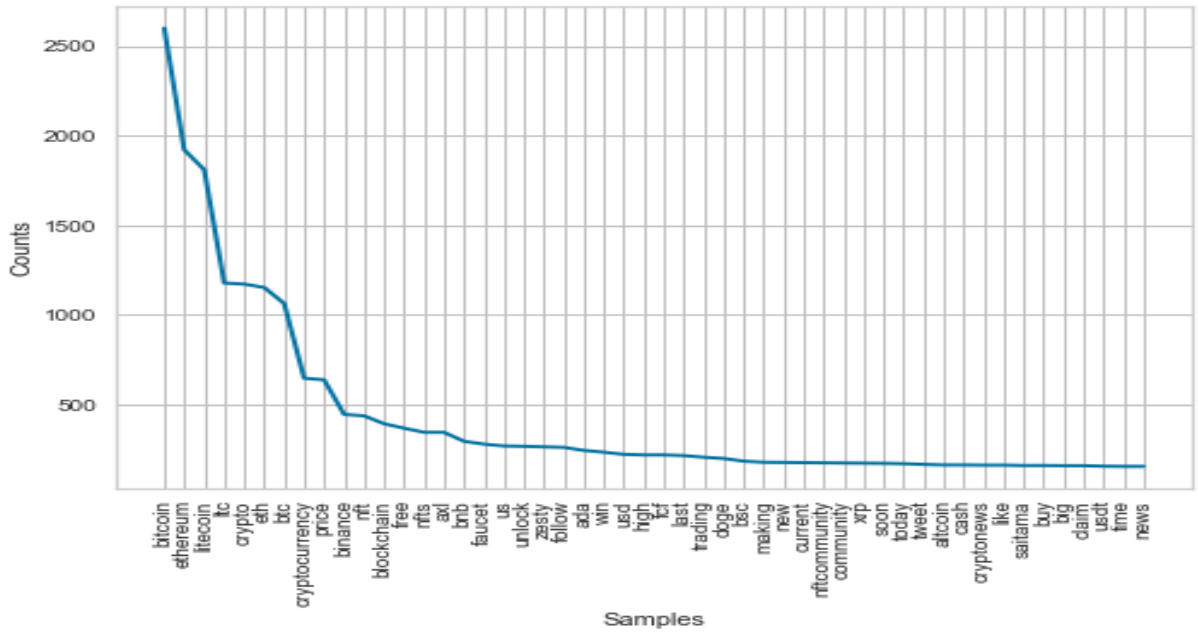


Figure 4.15 Frequency distribution plot

A principal component analysis (PCA) plot was carried out. It was discovered that total features dimension from 5799 was reduced to 380 of the total features to explain 50% of the variance in the data as shown in Figure 4.16. It demonstrates how effective PCA can be in certain situations to reduce dimensions.

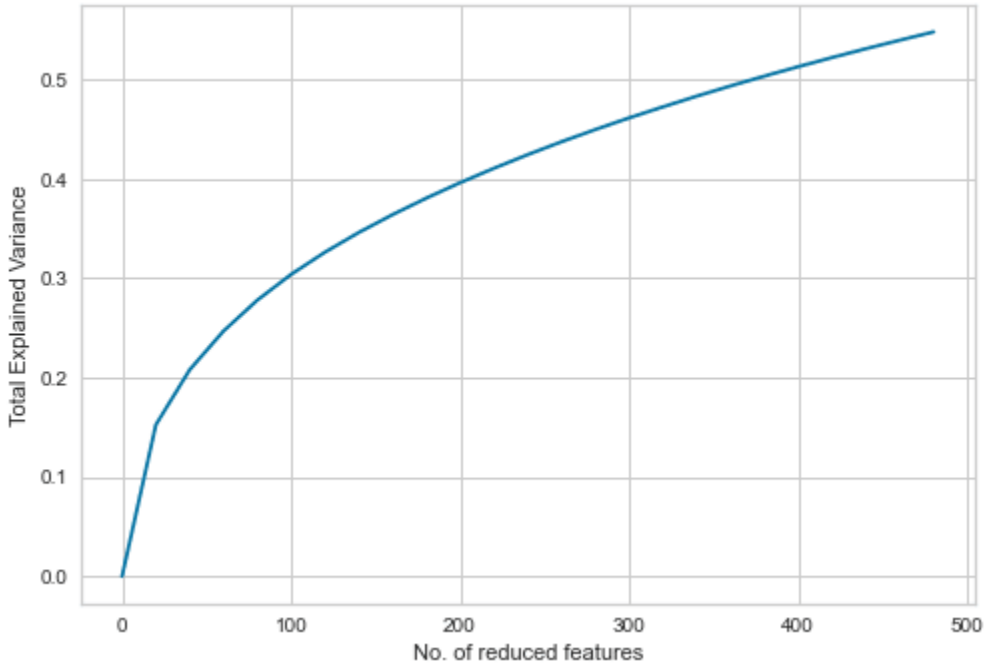


Figure 4.16 Principal Component Analysis Plot

## 4.1Discussions

This frequency distribution plot further supports the investigation on non-fungible token (NFT) and its strong dependency on the cryptocurrency market. A strong correlation was established. It was concluded that a trigger in bitcoin price increases the sales of NFT and a trigger in Ethereum price affects the NFT wallet (Ante, 2021).

Furthermore, the Principal Component Analysis results in this work corroborate with finding in (Gewers et al, 2021) where 50% of the variance in the datasets behaved in the manner of a standardized data.

In the model classifier experimentation, KNN performed poorly in all evaluation metrics. Generally, ensemble models are widely considered to perform better than individual models. However, the SVC model outperformed the ensemble in all forms of evaluation. This is similar to predictive modelling where deep learning models can sometimes produce results with less interpretability than baseline model such as logistics regression.

The count vectorization technique experimented on the ensemble model produced a poor recall. The Bert and Hash technique took considerable computational resource and processing time. The TFIDF method was discovered to outperform the other vectorization techniques employed in this work.

Finally, the GUI interaction was tested. The output of the cryptocurrency price, pie donut plot and the other visualization was consistent with current happening except for the text field which reads in data from the users and predict a sentiment. It was observed to generalize in the classification and misclassifies the sentiment occasionally.

# Chapter 5

## Conclusion

In this research, it was possible to successfully understand the latest trends on social media sites about the most popular cryptocurrencies through our various experiments and we were also successful at modeling it using the C-Support Vector Classification algorithm. Finally, it was possible to successfully have brought together all our findings and results into an interactive GUI Dashboard which is accessible by all users.

From these results, the researcher was able to conclude that there is indeed a correlation between the user sentiments and the movements of the cryptocurrency market. This showed us that the experimentation in this work were indeed pointing towards the right direction as we were able to extract out meaningful representations. However, there are several limitations in this research due to which we cannot come to a concrete conclusion about the topic yet. These limitations include:

- only a small sub sample of data extract was possible from a short timeline of 7 days (28th July 2022 to 3rd August 2022) due to the limitations of the API used for data scraping purposes.
- The use of only limited amount of data due to the constraints of time and effort it requires for manual labelling.
- the data used to train our classifier model is static and acquired from a specific period only. Without regular updates to the model, the pre-trained model may become irrelevant over time with the real-time changes in trends of the cryptocurrency market.
- Three (3) cryptocurrencies were used, namely Bitcoin, Ethereum and Litecoin for the complete representation of the market. Although these are among the most popular cryptocurrencies, but in some situations, they might not be able to entirely represent the trends of the cryptocurrency market.
- The researcher was unable to perform a long-term correlation analysis and hypothesis testing on whether the sentiments indeed have a statistically significant effect in the cryptocurrency market due to limitations of time.
- data was extracted from two social media sites, namely Reddit and Twitter out of the numerous social media platforms out there due to limitations of API availability.

As our next steps, this work can be furthered with research on how the data generated by the Social Media sites about the latest trends can be used to understand the movements of the

market. For the next approach, I plan to collect the social media data and the cryptocurrency market data on a regular basis for a long period of time to perform correlation analysis and hypothesis tests on whether such information can be used to predict and anticipate the movements of the actual market.

# References


- Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," *SMU Data Science Review*: Vol. 1: No. 3, Article 1.
- Ante, Lennart. (2021). The non-fungible token (NFT) market and its relationship with Bitcoin and Ethereum. SSRN Electronic Journal. 10.2139/ssrn.3861106.
- Aste, T (2019). "Cryptocurrency Market Structure: Connecting Emotions and Economics". Digit. Finance, 1, 5–21.
- Balakrishnan, Vimala & Ethel, Lloyd-Yemoh. (2014). Stemming and Lemmatization: A Comparison of Retrieval Performances. Lecture Notes on Software Engineering. 2. 262-267. 10.7763/LNSE.2014.V2.134.
- Bo Pang and Lillian Lee (2008). Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. 2, 1–2 (January 2008), 1–135. https://doi.org/10.1561/1500000011
- Caporale, G.M., Gil-Alana, A., Plastun, A.(2018). Persistence in the cryptocurrency market. Res. Int. Bus. Finance46, 141–148
- Chen, R., & Lazer, M. (2011). Analysis of Twitter Feeds for the Prediction of Stock Market Movement.
- Chicco, D., Jurman, G (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics 21, 6. https://doi.org/10.1186/s12864-019-6413-7
- Christopher D. Manning, Prabhakar Raghavan and HinrichSchütze, 2008. "Introduction to Information Retrieval", Cambridge University Press, New York, NY, USA.
- Colianni, S.; Rosales, S.; Signorotti, M (2015). Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis; CS229 Project, pp. 1–5. Available online: https://www.semanticscholar.org/paper/Algorithmic-Trading-of-Cryptocurrency-Based-onColianni-Rosales/9b838a3177523b8179511283b9489caa0f69910d
- Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In ICML'06: Proc. of the 23rd Int. Conf. on Machine Learning, ACM ICPS, pages 233–240. ACM.
- Dickinson, B. and Hu, W. (2015) Sentiment Analysis of Investor Opinions on Twitter. Social Networking, 4, 62-71
- Dritsas, E.; Livieris, I.E.; Giotopoulos, K.; Theodorakopoulos, L(2018). An apache spark implementation for graph-based hashtag sentiment classification on twitter. In

Proceedings of the 22nd Pan-Hellenic Conference on Informatics, Athens, Greece; pp. 255–260.

- Dulau, Tudor-Mircea, and Mircea Dulau (2019). ”Cryptocurrency Sentiment Analysis in Social Media.” Acta Marisiensis. Seria Technologica 16.2: 1-6.
- Ge-Stadnyk, Jing & Alonso-Vazquez, Marisol & Gretzel, Ulrike. (2017). Sentiment analysis. 10.4324/9781315565736-21.
- Gewers, Felipe & Rodrigues Ferreira, Gustavo & Arruda, Henrique & Silva, Filipi & Comin, Cesar & Amancio, Diego & da F. Costa, Luciano. (2021). Principal Component Analysis: A Natural Approach to Data Exploration. ACM Computing Surveys. 54. 1-34. 10.1145/3447755.
- Ghiassi, M., Skinner, J., and Zimbra, D. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. Expert Systems with Applications, 40(16):6266 – 6282.
- Ha, T. M., & Bunke, H. (1997). Off-line, Handwritten Numeral Recognition by Perturbation Method. Pattern Analysis and Machine Intelligence, 19/5, 535–539.
- Huang, X., Zhang, W., Huang, Y., Tang, X., Zhang, M., Surbiryala, J., Iosifidis, V., Liu, Z., & Zhang, J. (2021). LSTM Based Sentiment Analysis for Cryptocurrency Prediction. DASFAA.
- Jakub Kanis and Lucie Skorkovsk´a. 2010. Comparison of different lemmatization approaches through the means of information retrieval performance. In Petr Sojka, Ale˘s Horak, Ivan Kopecek, and Karel Pala, editors, Text, Speech and Dialogue, pages 93–100, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Joachims, Thorsten. (1998). Text Categorization with Support Vector Machines. Proc. European Conf. Machine Learning (ECML'98). 10.17877/DE290R-5097.
- Li, Jinyan(Leo). (2015). Hierarchical classification in text mining for sentiment analysis of online news. Soft Computing.
- Liu, Yukun and Aleh Tsyvinski (2021). Risks and returns of cryptocurrency. The Review of Financial Studies, 34(6):2689–2727.
- Livieris, Ioannis E, Emmanuel Pintelas, Stavros Stavroyiannis, and Panagiotis Pintelas (2020). Ensemble deep learning models for forecasting cryptocurrency time-series. Algorithms, 13(5):121.
- Luo, Tiejian & Chen, Su & Xu, Guandong & Zhou, Jia. (2013). Sentiment Analysis. 10.1007/978-1-4614-7202-5_4.

- Kimmo Kettunen, Tuomas Kunttu, and KalervoJ¨arvelin. 2005. To stem or lemmatize a highly in-flectional language in a probabilistic ir environment? Journal of Documentation, 61(4):476–496.
- Maria, Trigka & Dritsas, Elias & & Kanavos, Andreas & Vonitsanos, Gerasimos & Mylonas, Phivos. (2022). The Predictive Power of a Twitter User's Profile on Cryptocurrency Popularity. Big Data and Cognitive Computing. 6. 59. 10.3390/bdcc6020059.
- Nakamoto, S. (2008): Bitcoin: a peer-to-peer electronic cash system. Bitcoin.org
- Narayanan, Arvind, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder (2016). Bitcoin and cryptocurrency technologies: a comprehensive introduction. Princeton University Press.
- Nebojša, Horvat & Ivković, Vladimir & Todorović, Nikola & Ivančević, Vladimir & Gajic, Dusan & Luković, Ivan. (2020). Big Data Architecture for Cryptocurrency Real-time Data Processing.
- Orhan, A., Vogelsang, R. P., Andersen, M. B., Madsen, M. T., Hölmich, E. R., Raskov, H., & Gögenur, I. (2020). The prognostic value of tumour-infiltrating lymphocytes in pancreatic cancer: a systematic review and meta-analysis. *European journal of cancer (Oxford, England : 1990)*, *132*, 71–84. https://doi.org/10.1016/j.ejca.2020.03.013
- Patel, Mohil Maheshkumar, Sudeep Tanwar, Rajesh Gupta, and Neeraj Kumar (2020). A deep learning-based cryptocurrency price prediction scheme for financial institutions. Journal of Information Security and Applications, 55:102583.
- Raheman, A., Kolonin, A., Fridkins, I., Ansari, I. and Vishwas, M., (2022). Social Media Sentiment Analysis for Cryptocurrency Market Prediction. arXiv preprint arXiv:2204.10185
- Şaşmaz, E. & Tek, F. B. (2021). Tweet sentiment analysis for cryptocurrencies. 2021 6th International Conference on Computer Science and Engineering (UBMK), 613-618. doi:10.1109/UBMK52708.2021.9558914
- Senbel, S. (2021). Fast and memory-efficient TFIDF calculation for text analysis of large datasets. In H. Fujita, A. Selamat, J. CW. Lin, & M. Ali (Eds.), Advances and trends in artificial intelligence: Artificial intelligence practices (pp. 557-563). Springer. Doi: 10.1007/978-3-030-79457-6_47
- Stenqvist, E.; Lönnö, J (2017). Predicting Bitcoin Price Fluctuation with Twitter Sentiment Analysis; KTH Royal Institute of Technology, School of Computer Science and Communication: Stockholm, Sweden.

- Weng, Cheng & Poon, Josiah. (2008). A New Evaluation Measure for Imbalanced Datasets. Seventh Australasian Data Mining Conference (AusDM 2008). 87. 27-32.

## Figure References

- Figure 1 - Raheman, A., Kolonin, A., Fridkins, I., Ansari, I. and Vishwas, M., (2022). Social Media Sentiment Analysis for Cryptocurrency Market Prediction. arXiv preprint arXiv:2204.10185
- Figure 2- Ider, D., (2022). Cryptocurrency Return Prediction Using Investor Sentiment Extracted by BERT-Based Classifiers from News Articles, Reddit Posts and Tweets. arXiv preprint arXiv:2204.05781.

# 7 Appendices

## 7.1 Appendix A Ethic Approval


## 7.2 Appendix B Code to Artefact

Click on the link below to access the coding script consisting of:

- data collection code for scrapping Twitter and Reddit data (https://github.com/eidreiz01/Thesis-Scripts-2022/blob/main/data_collection_code.ipynb)
- end-to-end pipeline modelling the cryptocurrency market sentiment analysis (https://github.com/eidreiz01/Thesis-Scripts-2022/blob/main/sentiment_analysis_modelling.ipynb)
- interactive dashboard (https://github.com/eidreiz01/Thesis-Scripts-2022/blob/main/app.py)
- for all file, saved model and dataset (https://github.com/eidreiz01/Thesis-Scripts-2022)

Recommended IDE for running the script is Google Colab. Free access to GPU and pre-installed python library.