An analysis of social media sentiment using the transformer language model to predict cryptocurrency market trends

Done by: Maedeh Maghsoudi

Supervised by: Mr. Anotida Madzvamuse

Candidate Number: 245414

**Abstract**

The purpose of this thesis is to determine whether the cryptocurrency market is influenced by social media (in our case, Reddit); second, to forecast cryptocurrency trends using sentiment analysis of social media postings; Third explore a few models to solve social media Sentiment Analysis problems and get the best accuracy.

Text sentiment analysis, also known as opinion mining, is research on the calculation of people's views, evaluations, attitudes, and emotions expressed by entities.

Research on the pre-trained transformer models has significantly improved the performance of many natural language processing tasks. In recent years, pre-trained models (PTM) have been applied in sentiment analysis. Therefore, a question has arisen, which is whether PTMs contain sufficient syntactic information for Sentiment Analysis.

In this research, various models were explored to solve the social media Sentiment Analysis problem: the known model, VADER, which is a popular combined lexicon and rule-based sentiment analytic software. the recent DeBERTa which is a transformer based model developed by Microsoft. Lastly, a sentence transformer model named “all-MiniLM-L6-v2” that maps sentences and paragraphs to a 384-dimensional dense vector space and can be used for tasks such as clustering or semantic search.

Keywords: Cryptocurrency, Bitcoin, Sentimental Analysis, social media, Transformers, DeBERTa, VADER, Sentence-Transformers.

**Table of Contents**

[Chapter 1: Introduction 1](#_Toc112951555)

[Chapter 2: Background 3](#_Toc112951556)

[2.1 cryptocurrency and Bitcoin 3](#_Toc112951557)

[2.2 NLP and sentiment analysis 4](#_Toc112951558)

[2.2.1 Lexicon-based approach 4](#_Toc112951559)

[2.2.2 transformer based 5](#_Toc112951560)

[Chapter 3: Literature Review 6](#_Toc112951561)

[3.1 sentiment analysis 6](#_Toc112951562)

[3.2 social media sentiment analysis 7](#_Toc112951563)

[3.3 crypto prediction using sentiment analysis in social media 9](#_Toc112951564)

[Chapter 4: data 15](#_Toc112951565)

[4.1 Collecting Data 15](#_Toc112951566)

[4.2 Clean Up Data Reddit Data 15](#_Toc112951567)

[4.3 Roadblocks 16](#_Toc112951568)

[4.4 Labels 17](#_Toc112951569)

[4.5 Stats 18](#_Toc112951570)

[Chapter 5: Method 20](#_Toc112951571)

[5.1 Lexicon-based approach 20](#_Toc112951572)

[5.1.1 VADER 20](#_Toc112951573)

[5.2 Transformer 23](#_Toc112951574)

[5.2.1 BERT and RoBERTa 23](#_Toc112951575)

[5.2.2 DeBERTa 24](#_Toc112951576)

[5.3 Sentence Similarity 26](#_Toc112951577)

[5.3.1 all-MiniLM-L6-v2 26](#_Toc112951578)

[Chapter 6: Result and Discussion 28](#_Toc112951579)

[Chapter 7: conclusion and Future work 30](#_Toc112951580)

**Figures**

[Figure 1 Deberta Architecture 25](#_heading=h.2bn6wsx)

[Figure 2 Fine Tuning DeBERTa 29](#_heading=h.49x2ik5)

**Tables**

[Table 1 Data details 19](#_heading=h.2jxsxqh)

[Table 2 Post Segmentation Stats 19](#_heading=h.z337ya)

[Table 3 Models' Accuracy Comparison Results 30](#_heading=h.2p2csry)

# Introduction

Market growth for cryptocurrencies has been phenomenal in recent years and as a result of this market's lucrative potential, it is attracting a large amount of new investors. Cryptocurrencies have grown in importance in the national markets ever since the birth of Bitcoin in 2009. The fact that no central body governs Bitcoin makes it a decentralized currency and with lower barriers of entry to transact bitcoin due to the fact that they are not governed by any one nation, unlike more traditional forms of currency exchange like the banking system, which force users to comply with rules and authorities that are rife with fraud and severe corruption, Bitcoin has stamped its importance on the world economy. As a result, many nation states are now focusing on Bitcoin, and there is a possibility that in the future nations worldwide will adopt their own cryptocurrency. In comparison to other conventional forms of transactions, again like banking, it is clear from prior research studies that bitcoin has several special qualities such as the previously mentioned decentralization, its cheap transaction costs, the fact that nobody can freeze your Bitcoins, you can transact your bitcoin 24 hours a day, 365 days a year with no amount limit, and no restrictions of country borders or sanctions, you’re free to transact it however you wish. Bitcoin does not adhere to the typical institutional rules and models and therefore its price fluctuations are determined by public perception and opinion.[1]

We are all acutely aware of the interconnectedness of social media and the influence it has on people's lives. With readily available technology, it is now very simple to share your opinions on social media in the form of blog entries, online forums, reviews, and feeds (including those available on Twitter or Reddit) [2].

This brings us to the astounding quantity of messages sent every second. According to recent research, the typical person creates 1.7 MB of data per second, 102 MB in a minute, and 18.7 billion texts are sent daily [3]. The decentralized digital currency Bitcoin (BTC) is comparable to the majority of other well-known currencies in that it is influenced by socially created ideas, regardless of whether such opinions are supported by facts or not. The price of Bitcoin increased by as much as 5.2 percent on 24 March 2021 when Elon Musk tweeted Tesla would accept Bitcoin as payment [4], and it also fell by as much as 9.5 percent on 13 May 2021 when Elon Musk tweeted to question the energy consumption from Bitcoin mining. These examples show how tweets and news can change the dynamic pricing of cryptocurrencies. [5]

For our text data sources, we mainly used Reddit. Reddit is a social media network similar to Twitter, in that it compiles information from several sources and posts it in one place. A key distinction is that you follow topics rather than people on Reddit.

The three main objectives of this paper are to: first, determine whether or not the cryptocurrency market is influenced by social media (in our case, Reddit); second, forecast cryptocurrency trends using sentiment analysis of social media postings; and third exploring various models to solve the social media Sentiment Analysis problem: VADER, DeBERTa, all-MiniLM-L6-v2. The known model VADER (Valence Aware Dictionary for Sentiment Reasoning), proposed by Hutto and Gilbert, which is a popular a combined lexicon and rule-based sentiment analytic software. The recent DeBERTa model (Decoding enhanced BERT with disentangled attention) which is a kind of neural language model based on transformer, which uses self-supervised learning to pre-train on a large number of original text corpora. Model named all-MiniLM-L6-v2, which is a sentence-transformers model and it maps sentences & paragraphs to a 384-dimensional dense vector space and can be used for tasks such as clustering or semantic search.

Apart from the introduction, this thesis is divided into 6 chapters. The background first discusses the fundamentals of cryptocurrencies and sentiment analysis. Second, a literature review takes into consideration earlier related work produced by other scholars. The third and fourth chapters describe the data gathering procedure and methodology. The results and experiments are then produced in chapter five. The last chapter includes the thesis' conclusion and explores future study opportunities.

# Background

## cryptocurrency and Bitcoin

The term "cryptocurrency" itself was derived from the field of cryptography. Researchers investigating information confidentiality are known as cryptographers, and their research has deeper roots because these techniques have been used for a very long time. Cryptocurrency is now seen as a revolutionary notion in the development of payment systems and technology due to the fact that it is an asset that can be used widely. [6]

The most well-known and established cryptocurrency is Bitcoin. In the initial Bitcoin whitepaper, published in 2008 [7] , Satoshi Nakamoto proposed a trust-free electronic cash system that is supported by a blockchain that can be cryptographically verified and serves as an immutable public ledger for all transactions.

The blockchain may be used for any system in which one would trade value since it prevents asset duplications, even if the Bitcoin currency is its most well-known application. [8]. Bitcoin is a decentralized currency. Because there are no rules and regulations that can control the supply of bitcoin, it is impossible to artificially inflate or deflate the maximum supply of the currency. A fixed number of new bitcoins are produced every 10 minutes when a new block is added to the blockchain, making the supply of new bitcoins predictable and steady. Due to its limited supply and increased demand for bitcoin historically, the price has increased on a macro level. This demand is frequently impacted by public perceptions, attitudes, or news on bitcoin.  [9]

For example, the price of bitcoin dropped about 30% in a single day on 19/05/2021, following the disclosure of the Chinese government's plans to control cryptocurrencies, from $43,546.12 to $30,681.50. Elon Musk, the CEO of Tesla and SpaceX, has also made a number of tweets regarding Bitcoin that have significantly affected its price.

Such events have demonstrated that the demand for bitcoin can be unpredictable. However, the price of bitcoin has increased exponentially over time [10]. The deflationary nature of Bitcoin has now made its use as a payment method less desirable. Instead, the profitable potential of bitcoin as an opportunity for investment and a store of value similar to gold and other precious metals has come into greater emphasis.[11]

## NLP and sentiment analysis

Natural language processing (NLP) has expanded over the past ten years with the goal of improving the ability of computational systems to reason about natural language. The term "sentiment analysis" refers to the process of analyzing and extracting sentiment, opinion, subjectivity, and polarity from the text. [12]

Technically speaking, sentiment analysis as a component of NLP entails training a classifier on a labelled corpus to find and establish attributes and features that can then be applied to processing new input data and determining its nature.

Finding the underlying opinions, sentiment, and subjectivity in texts—all significant elements in influencing behaviour—is the crux of sentiment analysis.[13]In the same paper, they point out that there are several subtasks that makeup sentiment analysis, including polarity classification, opinion summarization, subjectivity classification, sarcasm identification, and the detection of misleading opinions.

Sentiment Analysis (SA), commonly referred to as opinion analysis or emotion AI, is the process of scoring emotions, opinions, and attitudes. Furthermore, analysis may be done using this score, which often falls into one of three categories: positive, negative, or neutral.

For a variety of fields, such as stock price prediction, polarity classification has been proven to achieve excellent accuracy in forecasting change or trends in public opinion. [14]

### Lexicon-based approach

Lexicon-based techniques compare the sentiment scores of a lexicon (a list of words) with the text being classified which results in a final polarity score. A lexicon is a collection of features (e.g., words and their sentiment classification). A typical technique in sentiment analysis is the lexicon-based approach, which compares a text passage to a vocabulary and assigns sentiment categories. Lexicons can be difficult to build, but once built, they are usually quite easy to use. Well-designed lexicons can achieve high accuracy and are held in high regard in the scientific community. [15]

#### VADER

Software called Valence Aware Dictionary and Sentiment Reasoner (VADER), created by Hutto and Gilbert combines lexicon- and rule-based sentiment analysis [16] . Text sentiment polarity (positive, neutral, or negative) and intensity may both be detected using VADER. The difficulties in assessing the language, symbols, and style employed in writings, particularly those in the social media realm, led to the development of VADER. [17]

VADER scores each tweet with a negative, positive, neutral and compound polarity score. The compound score is a sum of the individual sentiment scores, adjusted according to a set of rules and normalized to fall within the [−1→+1] range. VADER was widely used in related work and provides advantages including the following: it is open source and free; it is human validated and tuned for social media content[18]; and it has also been shown to perform competitively with human annotators and has outperformed several benchmarks, especially when based on social media content[16].

### transformer based

A Transformer-based language model is composed of stacked Transformer blocks [19]. Each block contains a multi-head self-attention layer followed by a fully connected positional feed-forward network. The standard self-attention mechanism lacks a natural way to encode word position information. Thus, existing approaches add a positional bias to each input word embedding so that each input word is represented by a vector whose value depends on its content and position. The positional bias can be implemented using absolute position embedding or relative position embedding [20]. It has been shown that relative position representations are more effective for natural language understanding and generation tasks. Large-scale Transformer-based PLMs are typically pre-trained on large amounts of text to learn contextual word representations using a self-supervision objective, known as Masked Language Model (MLM) [21]. The disentangled attention mechanism differs from all existing approaches in that it represents each input word using two separate vectors that encode a word’s content and position, respectively, and attention weights among words are computed using disentangled matrices on their contents and relative positions, respectively.

# Literature Review

Natural language text is nothing more to a computer than encoded bytes in its raw form. To help computational systems reason more effectively in regard to natural language, progress has been achieved in the field of natural language processing (NLP) over the past ten years. As the name suggests, sentiment analysis analyses text and draws out its sentiment, opinion, subjectivity, and polarity. The body of existing research includes a sizable number of studies on the use of social network data to implement various language processing algorithms.

## sentiment analysis

There are several applications for sentiment analysis, such as automatic positive, negative, and possibly harmful comment flagging on websites and social media platforms. This research report is based on a variety of relevant research topics and concepts.

Sentiment analysis, or the process of examining how the public feels about a particular entity. Opinion mining can be used in sentiment analysis's place, according to [22] , because the two terms are comparable. Opinions describe the holder's preferences and are defined as an individual's perception of something that is different from that of another person.

Sentimental analysis, according to [23], is a component of artificial intelligence. commonly referred to as the computation, evaluation, processing, sentiment, coding of opinions, and subjectivity within the word of text.

Provided in [24] are the technological models based on annotations that constitute the language processing system, the pipeline of processing procedures that Stanford CoreNLP contains, and information on how to utilize it.

A paper by [25] in-depth describes the techniques and algorithms the Stanford CoreNLP system employs to determine the sentiment of a text. The idea of Treebank Sentiment, a corpus of parsed texts organized as a tree with the structure of a syntactically and semantically labeled and annotated phrase, is introduced in this article.

The capabilities of the IBM Watson system are shown in [26], whose conjunction in the working process creates effective solutions. These abilities include dynamic learning, which enhances the result-based learning process to become more intelligent with each iteration and interaction. Natural language processing, which recognizes the complexities of unstructured data, [20] and proposes answer sets based only on relevant evidence.

The authors of the study[27] attempt an analysis that varies in that the selected texts were manually sorted into four categories: positive, negative, neutral, and irrelevant after being translated into English from foreign languages.

The Attention-based LSTM with Aspect Embedding (ATAE-LSTM) model was developed by [28] In order to determine the attention weight, the paper connected aspect embeddedness to each word vector of the input phrase, then used this information to calculate aspect-specific sentence embeddedness in order to categories sentiments. The development of increasingly intricate attentional processes followed, with the aim of improving the learning of aspect-specific representations.

Through the transformation of the work into a pair classification problem, a study makes use of BERT's capabilities [20]. Using a model of the sentence's syntactic structure, the ASC research determines the polarity of the sentiment aspects. Word dependencies and syntactical information were used by authors to model the dependency tree. They make use of a graph neural network (GNN).

## social media sentiment analysis

Social media posts were an instrumental data source that we used to measure and analyse sentiment. More specifically, posts from websites where they are open to the public. Social media sentiment analysis postings may be used in a variety of commercial applications, such as marketing, customer satisfaction analysis, or even trading stocks/ cryptocurrency for profit. Additionally, it has been determined that social media usage is increasing rapidly, with an estimated 2.77 billion active users worldwide at present and an expected increase in the coming years [29].

The internet is a vast virtual wasteland where one might discuss and express their thoughts with little limitations, which impacts different aspects of life and has an impact on both communication and marketing[30].The judgment of services and goods by potential clients is greatly influenced by ratings and reviews posted online. Reviews are regarded as a crucial factor in the choice making of purchasers in many different industries.

Customers trust other customers' opinions and reviews because, as reported in [31], more than 30% of internet users assess the services or goods online before purchasing. Particularly, those who have used the service or product in the past, as compared to the business marketing. Social media is thereby impacting consumer preferences through modifying their actions and attitudes.

Retailers are creating sales in conventional stores, according to [32] they discovered that their exposure was much increased by their social media presence. Additionally, a friendly and engaging presence in a chat room or social media network might significantly improve the brand's reputation. Additionally, helping the company get very beneficial and unstructured data on the demand patterns in a "nonintrusive manner." Monitoring social media activity is also a fantastic way to track metrics such as customer loyalty, consumer attitudes about a product or brand, the success and engagement of marketing efforts, and other data. Additionally, tracking social media activity enables the identification of key influencers who are pertinent to a campaign, business, or product.

The relevance of a microblogging platform as a source for sentiment analysis is discussed in [33]. The importance and necessity of Twitter as a data source for processing opinions increases with the volume of postings providing thoughts on various goods, services, affinities, and themes.

According to an article by[34], the analysis of tweets to determine opinions on subjects included in their text serves as a reliable measure of popularity. The author suggests a traditional categorization of emotions into three groups: positive, negative, and neutral. Using the previously categorized tweets as learning data, the real-time Twitter data is collected and categorized into one of the categories. The Naive Bayes (NB) algorithm serves as the foundation for the categorization process.

Social media is a key medium for engagement and communication in the modern day. For many people, social networking sites act as information hubs. As a result, social media has emerged as a major source of information. Another study [35] asserts that the recommender system and decision-making heavily rely on sentiment analysis inside social media assessments. Making decisions enables investing and purchasing goods. Users frequently use their peers' experiences when deciding whether to buy a product or make an investment. In the modern day, there are several ratings accessible online that are challenging for consumers or investors to understand.

A study in [36] notes that sentiment analysis simplified and made this process simpler. According to [37], who are explaining review polarity, it is possible to tell if a review is positive or negative without even reading the entire line. enhancing the buyer's convenience in the purchasing decision.

The study hypothesis offered in[38] is different and original. The method of operation tries to discover popular sentiments of the platform's most prominent and influential users.

Sentiment analysis, also known as opinion mining or contextual mining, is used in computational linguistics, text analysis, and Natural Language Processing (NLP). This also helps in locating, quantitatively organizing, and extracting the tailored data. Sentimental analysis mostly uses client feedback in the form of comments or reviews on products or materials [39]. As an illustration, the user may buy any product online. Therefore, the buyer reads product reviews before making a purchase. Hence, supporting the buyer in decision making regarding the merchandise.

Three terms for expressing the sentiments are derived from sentiment analysis. First, objectivity with relation to the topic of the opinion. The object's feature comes next. Last but not least, the person expressing the view on the thing. Sentimental analysis addresses a number of issues, including object recognition, feature extraction, and opinion orientation discovery.

## crypto prediction using sentiment analysis in social media

Over the previous several years, numerous academics tried establishing various approaches for predicting the cryptocurrency/ Bitcoin market. A common trend in this research is the utilization of two key information sources: historical pricing and Internet-based data about Bitcoin. This section will mostly cover some recent noteworthy efforts that use online social signals to predict cryptocurrency due to the scope of our article.

Recent studies have made it clearly evident that people's purchasing decisions are impacted by the data provided on websites and on social media so sentiment analysis has been used in a number of efforts to forecast early cryptocurrency market movement using social media sentiment.

In another paper [40], the authors evaluated the price returns and daily trading volumes of cryptocurrencies with the causality of tweet sentiments, tweet volume, and the buyers to sellers' ratio on Twitter. It has been claimed that opinions posted on Twitter may be useful for forecasting changes in Bitcoin prices.

A strategy for forecasting changes in Bitcoin prices based on data from Google Trends and Twitter was put out by [41] The analysis of the input data revealed a strong correlation between price and both tweet volume and Google Trends. With these input features, a multiple linear regression model was employed to forecast the price of cryptocurrency.

In their study, [42] demonstrated how information from social media affected the movements of the Bitcoin price. For the purpose of analysis, tweets and forum comments on Bitcointalk were gathered. The findings indicated that information shared on social media, particularly on the Bitcointalk forum, had a significant influence on the price of bitcoin.

A technique for forecasting changes in cryptocurrency price and transaction volume was described by [43] The Granger causality test was performed by the authors to examine the relationship between user feedback from forums and social media and pricing. Variations were predicted using the AODE (Average One-Dependent Estimators) model.

Using previous price data,[44]. forecasted the price of bitcoin and its moves. For the prediction task, recurrent neural network (RNN) and long short-term memory (LSTM) models with Bayesian optimization were applied. In order to compare the performance of the time series model with the neural network model, an ARIMA (Autoregressive Integrated Moving Average) time series model is also created. RNN was outperformed by the LSTM, while ARIMA was outperformed by deep learning models in terms of prediction.

Time series analysis was suggested by[45] as a way to determine how the price of Bitcoin relates to the economy, technological factors, and sentiment on Twitter. Support Vector Machines (SVM) were employed in the sentiment analysis task to calculate the daily sentiment ratio for Twitter. The author estimated two regression models for the time series analysis task: a short-run Ordinary Least Square (OLS) estimate and a long-run Vector Error-Correction Model (VECM). The findings demonstrated that Twitter sentiment ratio has a favorable immediate effect on Bitcoin price. Moreover, it was discovered that the volume of Wikipedia search requests had a favorable impact on Bitcoin's price.

A technique for examining the link between Bitcoin price and tweet sentiment and views for Bitcoin on Google Trends was provided by[46] The results of the investigation revealed a strong link between price and Google Trends SVI.

By using Twitter sentiment to train an Extreme Gradient Boosting Regression tree model (XGBoost) to forecast ZClassic price movements, [47] aimed to illustrate this idea. Using recent and previous data from Twitter emotions and Bitcoin closing prices, this study allowed the KryptoOracle to forecast the Bitcoin price for the next minute. Regression tree model XGBoost was chosen due to its effectiveness, speed, and ease of retraining.

Researchers in [48] tried to forecast the Bitcoin and Litecoin values two hours in advance using the sentiments indicated in the most recent tweets. They were looking at whether social characteristics might be used to forecast cryptocurrency values. Thus, they employed a Multiple Linear Regression (MLR) model to forecast a bi-hourly average price from the quantity of positive, neutral, and negative tweets gathered every two hours.

Another paper [49] compared the importance of several preprocessing methods for sentiment analysis of tweets. They experimented with 16 different preprocessing techniques and four different machine learning algorithms to categorize tweets. According to their findings, lemmatization, replacing repetitive punctuation, substituting contractions, or eliminating numerals were all advised.

Authors in [50] made an effort to categorize Twitter users who mention COVID-19 on the social media platform and trained many machine learning algorithms to identify these people. Logistic Regression, Random Forest, Support Vector Machine, Stochastic Gradient Descent, Multi-Layer Perceptron, and XGBoost were some of the machine learning algorithms trained on these properties. When trained on the baseline, demographic, and geolocation data, Random Forest got the best AUC-ROC score of any system.

Similar research on the use of sentiment in tweets to influence investment choices, especially in Bitcoin, were conducted by[51] Hour by hour and day by day, the authors' supervised machine learning methods that finally produced accuracy results that were above 90%. The author noted that the robust error analysis on the input data, which on average produced a 25% greater accuracy, was what enabled the 90% accuracy.

In the paper Trading on Twitter: Using social media sentiment to predict stock returns by[52] , 2.5 million tweets concerning S&P 500 companies were applied to the authors' unique sentiment classifier and stock returns were compared. The findings demonstrated that quickly spreading emotion is expected to be reflected in a stock price on the same trading day, but sentiment that spreads more slowly is more likely to be reflected on subsequent trading days.

According to [53], who used a variety of machine learning algorithms to forecast the values of twelve cryptocurrencies and obtained impressive outcomes, the market is not efficient since previous price movements influence future ones.

This paper [54] states that techniques based on machine learning tend to perform worse than modern investing strategies like buy and hold. The challenge of predicting cryptocurrency prices is still very new, and therefore there exist conflicting literature studies on the subject. Yet, a lot of them did not take social media impact into account when determining influences. Several publications have previously made an effort to examine how social media and the cryptocurrency market interact.

Authors in this paper [55] used an ELMo embedding model and an SVM (Support Vector Machines) classifier to attempt to forecast the Bitcoin price using the sentiment from the social media platform Twitter. One key finding from this study is emphasized in the current research, and it is that Bitcoin prices are not influenced by investor sentiment, at least not on Twitter. The findings of [56], which demonstrate a substantial correlation between Twitter sentiment and Bitcoin returns, and of [15], who employed causality analysis to find that Twitter sentiment may be used to forecast the values of Bitcoin, Bitcoin Cash, and Litecoin, respectively, are in direct opposition to this conclusion.

The purpose of the article[43] is to investigate the impact of online community debates on the development of the cryptocurrency market. Users have access to the Stanford CoreNLP tool's capabilities and techniques through a library that is defined in Stanford CoreNLP.

Social networking sites like Twitter, Facebook, and several news sites like Reddit all have very high levels of activity amongst their users. This essay demonstrates how to accurately examine and categorize someone's emotions or written work. The study also recommends Twitter as a social media medium due to the frequent tweets related to cryptocurrency. However, despite the relatively low frequency of news broadcasts, they still have an influence on cryptocurrency prices. [57]

The paper “Predicting Fluctuations in prices based on User Comments and Replies on News” [43] talks about how the volume of transactions, user feedback, and user responses affect Bitcoin prices. Three cryptocurrencies—Bitcoin, Ethereum, and Ripple—are the main emphasis of the article. Additionally, it shows that the price of Bitcoin is positively impacted by user comments (but is unaffected by negative remarks), but the prices of Ripple and Ethereum are greatly impacted by negative user comments and answers. The report also argued that while Ripple's community is young and immature, since it was just created in 2015, Bitcoin's projections were more accurate owing to its vibrant community.Gallen Thomas demonstrated a correlation between Twitter data analysis and public perceptions of the value of cryptocurrencies. In addition, he made the argument that social media, namely Twitter emotions, have a greater influence on bitcoin users than consumers' emotional states do. [58]

Research on Twitter attitude against the stock market was done by [59]. He employed data analysis and neural networks in his research project to predict the future direction of the bitcoin price. From his findings, it was possible to forecast changes in the stock market for the upcoming days, based on the previous week's Tweets.

The Author conducted another study in which they developed a learning model using tensor networks. He focused on Twitter data in order to analyze the sentiment and see whether the findings could be formulated, and see how it related to other stochastic occurrences. [60]

Additionally, authors in [61] created a hybrid model using a recurrent neural network and multiple linear regressions that attempted to address the shortcomings of each model utilized.

The results of a study by[62] demonstrated the importance of using online search tools like Google Trends and social media such as Twitter comments to estimate the value of cryptocurrencies like Bitcoin. For search phrases relating to mining and blockchains, two crucial components of cryptocurrencies, Nie discovered that these elements yielded a very high significance with a very low P-value.

Pre-trained language models like BERT and RoBERTa have significantly improved NLP tasks in recent years. Simply utilizing these contextualized embeddings has resulted in significant performance improvements due to the quantity of knowledge gained during the pre-training period.

Researchers in this paper [63] tried using several stacked standard prediction layers on PTM for the E2E-ABSA tasks. On the basis of the original Bert model, RoBERTa has proved the following points through experiments: further increasing the number of pre-training data can improve the model effect; extending the pre-training time or increasing the number of pre-training steps can also improve the model effect.

This Paper [64] demonstrated that it was possible to use a combined strategy to predict the price of bitcoin using both social media signals and conventional financial modeling techniques. Target words, mood, and other characteristics that define the changing social media environment, such as post frequency and comments, were included in these signals. These researchers used a method that resulted in a daily increase of 32.29 percent. Valence metrics by themselves produced a daily increase of 0.1183. Trading Bitcoin could be profitable at these prices if you have adequate funds. The back testing that the researchers did on their findings gives their prediction model some further assurance.

In a work [65] gathered tweets relevant to the price of Bitcoin and created a model that was helpful to forecast the price of Bitcoin. They assessed each tweet's impact as positive, negative, or neutral using the Valence Aware Dictionary and Sentiment Reasoner (VADER) analysis tool. Only tweets that were either negative or positive were saved and utilized for study.

Our project is differentiated from the projects previously outlined due to several key outliers. Firstly, we attempt to market sentiment predictions using new data on Reddit, as you can find most of the papers use Twitter and Google Trends for their source of data. Secondly, we forecast Bitcoin and cryptocurrency trends rather than market and coin price fluctuations. Finally, and most importantly, we concentrated on several approaches to label our data based on the most current transformer-based model, which has attracted less attention in recent works.

# data

## Collecting Data

The work starts by taking two different sources of data acquired in the research phase. These sources are Social Media and Crypto Compare. Social Media Source: Within the study, social media sources for sentiment analysis are Reddit posts. We chose this platform because it is easier to obtain information from users securely without affecting the quality of the extract. The Reddit API does have search limitations, nevertheless, due to the lower number of posts produced per day, the API is more than enough to fulfill what is required to give accurate predictions. Also, the Reddit platform’s API offers developers a very wide range of endpoints that can be used, each having different functions: from posting, creating, or editing posts, comments, messages, personal information, or account information, to searching and obtaining different items such as subreddits, posts, users, accounts, messages, conversations, names, etc. After a multitude of criteria specified in the request. On Reddit, there were extracted cryptocurrency subreddits, which facilitated the subsequent selection of posts and comments. During the collection of data, different approaches and sources were utilized to obtain data necessary for analysis.

To obtain Reddit post/comment data, the thesis used "PSAW" which is a minimalist wrapper for searching public Reddit comments/submissions via the pushshift.io API. The PushShift API helps provide the improved searching capability to obtain Reddit comments and submissions. Pushshift divides the data into two sections, i.e., submissions representing the posts on the Reddit platform and the comments on the Reddit platform. These Pushshift API features allowed fetching the data from every cryptocurrency media source. The study acquired hourly data from the subreddit and the comments on BTC and crypto in cryptocurrency subreddits.

## Clean Up Data Reddit Data

Posts contain a large amount of noise, such as hashtags, URLs, and emotions. These characters make Reddit sentiment analysis a challenging assignment. Pre-processing of the data is a very important step as it dictates the efficiency of the other steps down the line for sentiment analysis. So firstly we need to do the pre-processing to clean the data and make it less noisy. This pre-processing step is a common practice in NLP models to ensure that the remaining word tokens are meaningful. Each post has gone through the following process in sequence:

* Converted all English alphabet characters to lowercase.
* Removed all the URLs.
* Removed all the characters that are not in the English alphabet, to filter out numbers and non-English posts using the library spaCy
* Remove too short and too long posts.
* Remove specific words and phrases.

Many research papers have approached the problem that social media bots impose on sentiment analysis methodologies on the cryptocurrency subject like,[66] or [67].

Some heuristics are proposed for bot identification on Reddit:

* The post contains “give away” or “giving away”.
* The post contains “deleted” or “removed”.
* The post contains “I am a bot, and this action was performed automatically”.
* The post contains “pump”, “register”, or “join”.
* The post contains more than 14 hashtags.
* The post contains more than 14 ticker symbols.

Also, we can see after cleaning, our fear data decreased from 3 million to 0.8 million likewise greed data decreased from 3.08 million to 1.9 million

## Roadblocks

During these experiments, we encountered numerous challenges.

* Sarcasm. People use sarcasm in their posts or conversation, it is a way of expressing a negative sentiment using a backhanded compliment. This situation can make it difficult for the sentiment model to understand the true context of the texts. If most of the texts contain sarcasm, it results in a higher number of positive sentiments even though in reality, it was a negative sentiment.
* Idioms. Sentiment analysis methods are still not mature enough to understand the idioms used in the texts.
* Negations. High use of negation leads to misclassification. For example, “not bad” is positive but for most of the lexicon-based models, it will be negative because we are using a negative word with negation. This word order makes it positive, but most lexicon-based models consider it negative.
* Non-text data. Reddit is not limited to texts only. Users can upload audio, images, and videos. If the images contain a strong indication of price change, the sentiment model will miss that.

## Labels

Fear and Greed are two emotions that can drive prices in opposite directions. When investors are fearful, they tend to sell assets to avoid losses and if they are greedy, they may be more likely to buy assets to make back the money they previously lost. This can result in prices rising even higher. It uses data from various exchanges and news sources to calculate the market’s “fear and greed” score. The Fear & Greed Index can be a helpful tool for crypto investors, as it can give them a quick snapshot of market sentiment. However, it’s important to note that the index is not a perfect indicator and should be used as one of many factors when making investment decisions.

The index looks at different factors to come up with its reading. Some of these indicators are relatively simple, such as the overall market capitalization of Bitcoin. Other indicators are more complex, such as the ratio of put options to call options or the amount of news coverage about Bitcoin. These indicators are:

1. Social media volume. The social media volume is calculated by taking the total number of tweets about Bitcoin and Ethereum over a while and dividing it by the total number of tweets about all cryptocurrencies. This gives us a measure of how much attention the two largest cryptocurrencies are getting on social media. The amount of news coverage can also indicate investor sentiment. A lot of news coverage generally means a lot of interest in that specific asset, which can be either good or bad.
2. Market volatility. Market volatility is measured by taking the daily returns’ standard deviation over time. This gives us a measure of how volatile the market is.
3. BTC dominance. BTC dominance is simply the percentage of the total market cap that Bitcoin makes up. This gives us an idea of Bitcoin’s importance to the overall market.
4. Transactions per day. Transactions per day are a measure of how much activity is happening on the blockchain in one given day. This gives us an idea of how much actual use the cryptocurrency is getting.
5. The popularity of search terms. The more times an item is looked up, the more interest there is in that asset. The popularity of search terms is measured by taking the number of Google searches for Bitcoin and Ethereum over a period of time and dividing it by the total number of Google searches for all cryptocurrencies. This gives us an idea of the interest in the two largest cryptocurrencies.
6. Price changes over time. Price changes over time measure how much the price has changed over a while. This gives us an idea of how volatile the market is.
7. Exchange inflow/outflow ratio. Put/call ratio: This ratio measures the number of put options (options to sell a security at a certain price) relative to call options (options to buy a security at a certain price). A high put/call ratio indicates that investors are more worried about a possible drop in the underlying asset price than they are about a possible rise.
8. Technical indicators. Technical indicators are mathematical calculations that can help traders identify trends in the market. Some popular technical indicators include moving averages and support and resistance levels.

## Stats

A total of 6 million posts/comments were collected for BTC and Cryptocurrency related news. We gathered data from the date range 2021-08-01 to 2022-08-01. In this period, we had 57 Greed days that were the bottleneck. So, we chose 57 Greed days plus 57 randomly selected days from 308 Fear days.

Stopwords, ‘bitcoin’, ‘btc’ and ‘cryptocurrency’ have been excluded from the counting process, as unsurprisingly, they are the most frequent words given our search criteria when constructing the dataset.

The gathered data for Fear and Greed had some anomalies that needed to be cleaned. Statistics about the data before and after cleaning described in the Table 1.

**Table 1 Data details**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Fear (Original) | Fear (Cleaned) | Greed | Greed (Cleaned) |
| Posts count | 1401208 | 1322036 | 3084992 | 2894292 |
| Mean word count | 22.964472 | 21.547232 | 19.847671 | 18.214493 |
| std word count | 43.877997 | 44.216340 | 28.707581 | 28.157411 |
| min word count | 1 | 1 | 1 | 1 |
| 25%-word count | 6 | 6 | 6 | 5 |
| 50%-word count | 13 | 12 | 12 | 11 |
| 75%-word count | 26 | 23 | 24 | 21 |
| Max word count | 1759 | 1759 | 1720 | 1720 |

By a general cleaning of obvious anomalies, data statistics weren't changed significantly. Greed data only reduced by 6% and the Fear data 5.9%. The other problem with the data was the huge difference between min and max number of words. Problems that we could introduce to our method is dimension inconsistency and usage of meaningless single words (or a couple of word phrases) in the context. So, we limited the dataset to posts between 25% and 75%, for example, for the cleaned Fear dataset, we only used posts that have a word count in the inclusive range of 6 and 23.

**Table 2 Post Segmentation Stats**

|  |  |  |
| --- | --- | --- |
|  | Fear 25-75% | Greed 25-75% |
| Posts count | 709843 | 1612975 |
| Mean word count | 12.476560 | 11.145330 |
| std word count | 4.851446 | 4.603789 |
| min word count | 6 | 5 |
| 25%-word count | 8 | 7 |
| 50%-word count | 12 | 10 |
| 75%-word count | 16 | 15 |
| Max word count | 23 | 21 |

It is shown in Table 2, that we ended up having 70 thousand Fear data and 1.6million Greed data. To have a balanced dataset, we used all the remaining posts with Fear label, and sample equivalent number of Fear labeled posts from the posts with greed label. The total data for the mixed data set would have 1419686 posts, with the same number of posts for each label to make sure the dataset is unbiased.

# Method

## Lexicon-based approach

Lexicon-based techniques compare the sentiment scores of a lexicon (a list of words) with the text being classified to result in a final polarity score. A lexicon is a collection of features (e.g., words and their sentiment classification). A typical technique in sentiment analysis is the lexicon-based approach, which compares a text passage to a vocabulary and assigns sentiment categories. Lexicons can be difficult to build, but when they are, usually are quite easy to use. Well-designed lexicons can achieve high accuracy and are held in high regard in the scientific community. [15]

### VADER

VADER (Valence Aware Dictionary for Sentiment Reasoning), proposed by [16], is a popular combined lexicon and rule-based sentiment analytic software. The difficulties in analyzing the language, symbols, and style employed in writings, particularly those in the social media realm, led to the development of VADER. The authors created a vocabulary of 9,000 lexicons with emotion scores ranging from -4 to 4, with 4 denoting extremely negative sentiment and 4 denoting extremely positive sentiment. The writers compile a sentiment score of 10 persons for each vocabulary. The average of 10 people's scores is used to evaluate each lexicon's score in the dictionary. This task is done through crowdsourcing on Amazon Mechanical Turk (AMT). In short, VADER takes a single text document as input, assesses its sentiment level, and then outputs a score called the sentiment score. VADER initially counts the total of sentiment scores of all terms in the document that occur in the above sentiment lexicon dictionary on each input document. Because each document differs in length, the authors proposed the following formula to normalize sentiment score to a range of -1 to 1:

s =

where:

* s is normalized sentiment score,
* x is the sum of sentiment score of a text document,
* α is a normalizing parameter, default value is 15.

VADER was created by looking through and choosing features from three previously created and verified lexicons: Linguistic Inquiry and Word Count (LIWC), Affective Norms for English Words (ANEW), and General Inquirer (GI). The writers also included emojis, slang, and widely used acronyms from social media. A valence value was assigned to each characteristic, and using this new data, 7500 features were chosen to be part of the VADER vocabulary. Hutto and Gilbert examined 800 tweets' syntax, grammar, and perceived valence value in addition to the word bank. Five key behaviours that are utilized to affect a tweet's intensity were identified through the investigation indicated above and turned into rules. These rules and this lexicon makeup VADER altogether:

* Punctuation: This can be done to make a phrase sound more intense. Compared to "Bitcoin price is growing," the phrase "Bitcoin price is increasing!!!" is stronger.
* Capitalization: Using ALL-CAPS for words that are particularly indicative of a feeling can enhance that sentiment and engage readers' attention. It is more expressive to say: "It is GREAT NEWS for all traders around the world." than "It is great news for all traders around the world."
* Degree adverbs: These words have the power to either intensify or lessen feelings. For instance, the phrase "This law is exceedingly awful news for Blockchain fans" intensifies the negative connotation, but the phrase "The Number of transactions is slightly lower than yesterday" lessens its impact.
* “But”: The sentence after "but" may come to dominate the overall opinion as a result of this contrastive conjunction. Many individuals could consider the second phrase in the example below to be more significant: "More and more scholars are working on cryptocurrency-related research topics, but relatively few of them really have novel insights."
* Trigram: Nearly 90% of instances where negation changes the polarity of the text were detected by the authors by looking at the tri-gram before a sentiment-heavy lexical characteristic. "The food here is not actually that fantastic," would be a negated statement.

A compound sentiment score, calculated by adding the adjusted valence (intensity) ratings of all words in the lexicon, is what VADER will provide for a given input of text. The compound score is a decimal value between [−1, 1] (−1 being the most negative, 1 being the most positive). Using the threshold values recommended by the authors, the compound score can alternatively be interpreted as a categorical sentiment value if necessary:

|  |  |  |
| --- | --- | --- |
|  | Positive | If compound > 0.05 |
| Sentiment(compound) | Neutral | If -0.05 <= compound <= 0.05 |
|  | Negative | If compound < -0.05 |

VADER produces the proportions of positive, neutral, and negative terms in the text in addition to the compound score. According to Hutto et al., VADER is a rule-based model with its own gold-standard lexicon that assigns each word a valence score—a decimal number between [-4, 4]—that represents the word's level of intensity. The valence of each word in the phrase is taken into account by VADER when calculating the compound sentiment. For instance, the intensity ratings for the terms "okay" and "best" are 0.9 and 3.2, respectively, while the scores for the words "uncomfortable" and "horrific" are 1.6 and 3.4, respectively. Because Twitter's random tweets were used to construct the gold-standard vocabulary, VADER is especially useful in social media settings. As a result of this, the lexicon also includes present-day Internet colloquialisms, such as initialisms (LOL, WTF), emoticons ( :(, ;-) ), as well as emojis ( page14image566832).

*Hutto, C. and E. Gilbert[16] express the goals on which they based the creation of VADER as the following:*

* Works well on social media style text, yet readily generalizes to multiple domains
* Requires no training data, but is constructed from a generalizable, valence-based human-curated gold standard sentiment lexicon
* Is fast enough to be used online with streaming data,
* Does not severely suffer from a speed-performance tradeoff.

In order to evaluate VADER's performance, it was put up against eleven additional semantic analytic tools and approaches for polarity categorization (positive, neutral, and negative sentiment), which was done across four diverse domains. In every test instance, VADER regularly ranks in the top three results and outperforms the competition in the social media text area. Since a lexicon is already there, VADER doesn't need any training data or training compared to approaches to sentiment analysis based on machine learning. VADER can handle vast amounts of text quickly since it is totally rule-based. It is possible to modify the vocabulary while using VADER using the Natural Language Toolkit (NLTK) module in Python by adding extra tokens and valence values. The lexicon and Python-specific modules were provided by the authors under an MIT License, making them open source and free to use.

## Transformer

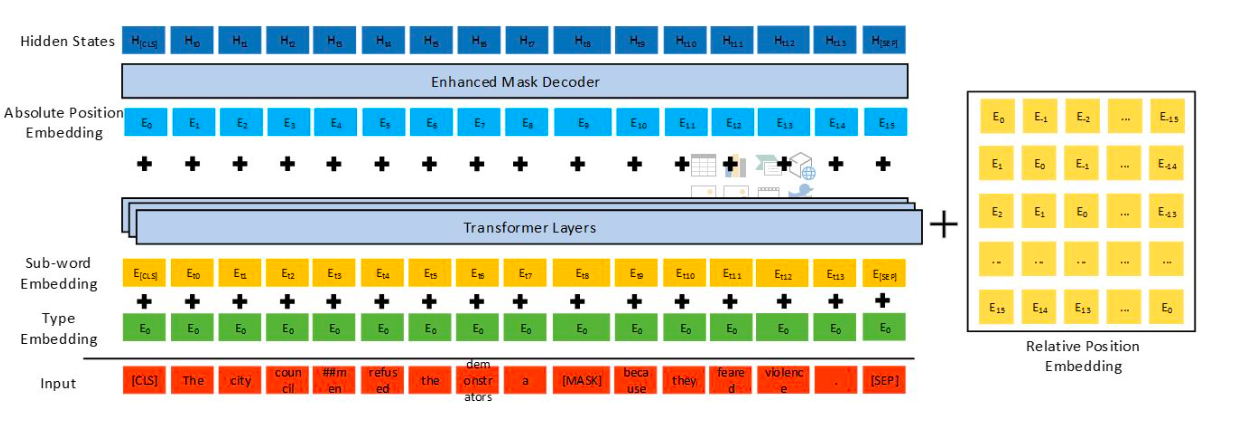
For neural language modeling, The Transformer has been established as the most successful neural network model. Transformers apply self-attention to compute in parallel an attention weight for each word from the input text that measures the influence each word has on another, in contrast to recurrent neural networks (RNNs) that process text in sequence. This allows for much more parallelization than RNNs for large-scale model training [19]. Large-scale Transformer-based Pre-trained Language Models (PLMs) have become more popular since 2018; examples include GPT, BERT RoBERTa, XLNet, UniLM, ELE. Due to the use of task-specific labels, these PLMs have been improved, setting new standards for numerous downstream natural language processing (NLP) activities.

### BERT and RoBERTa

Aspect sentiment classification (ASC) activities from the past were often reliant on manually created functions, including term frequency. Deep learning-based ASC problems have received a lot of attention recently. For aspect-based sentiment analysis (ABSA), deep neural language models based on BERT are frequently used. Bert evaluates the most recent deep bidirectional encoder representation from transformers, which reads text while taking the left and right sides of the word's context into account at all levels. Bert is able to provide additional text semantic representations, where each word is translated into an embedded vector based on its location inside the phrase. BERT is one of the pre-training techniques (PTMs) that has significantly improved ASC task performance [68]. Next, RoBERTa (A Robustly Optimized BERT Pre-training Approach) can outperform the performance of the traditional BERT technique by training the model for a longer period of time using larger batches of data.[69]This study will make use of the DeBERTa model. This is an improved version of the well-known RoBERTa and BERT models, which perform better than the majority of the current language models at this time.

### DeBERTa

DeBERTa (Decoding-enhanced BERT with disentangled attention) is a Transformer-based neural language model that is pre-trained using self-supervised learning on large raw text corpora. DeBERTa is designed to learn universal language representations that may be adapted to a variety of downstream NLU tasks, much like other PLMs. DeBERTa uses three innovative approaches shown in Figure 1 to improve prior state-of-the-art PLMs (such as BERT, RoBERTa, and UniLM): a disentangled attention mechanism, an enhanced mask decoder, and a virtual adversarial training method for fine-tuning.[70]



**Figure 1 Deberta Architecture**

DeBERTa is represented using two vectors that encode its position and content, respectively, and the attention weights among words are calculated using disentangled matrices based on their respective relative positions and contents. Given the small number of parameters (1,5B), this model will be accurate and fast so it can be used in a fast-moving market like cryptocurrency also, it will be at an affordable cost. DeBERTa model with pre-trained weights will be downloaded from HuggingFace. We will examine the results and evaluate them against the prior state-of-the-art studies by fine-tuning this pre-trained model with our dataset.

DeBERTa enhances prior state-of-the-art PLMs with three unique methods: a disentangled attention mechanism, an enhanced mask decoder, and a virtual adversarial training approach for fine-tuning.

**Disentangled attention**: DeBERTa represents each word using two vectors that encode its content and position, in contrast to BERT, which represents each word in the input layer using a vector that is the sum of its word (content) embedding and position embedding. The attention weights among words are calculated using disentangled matrices based on their contents and relative positions, respectively. The finding that the attention weight of a word pair depends not only on its contents but also on its relative placements serves as the basis for this. For instance, the relationship between the terms "deep" and "learning" is considerably more dependent when they are used together than when they are used separately.[71]

**Enhanced mask decoder**: DeBERTa is pre-trained utilizing masked language modeling (MLM), the same as BERT. MLM is a fill-in-the-blank challenge where a model is trained to predict what a masked word should be based on the words around the token. DeBERTa uses the content and position information of the context words for MLM. The disentangled attention mechanism already considers the contents and relative positions of the context words, but not the absolute positions of these words, which in many cases are crucial for the prediction. Consider the sentence “a new store opened beside the new mall” with the italicized words “store” and “mall” masked for prediction. Despite having comparable local contexts, the two terms have different syntactic functions in the phrase. (In this case, the word "store" rather than "mall" is the subject of the phrase.) It is crucial to take a word's absolute location into consideration in the language modeling process since these syntactical intricacies heavily depend on the absolute positions of the words in the sentence. Right before the SoftMax layer, where the model decodes the masked words based on the aggregated contextual embeddings of word contents and positions, DeBERTa incorporates absolute word position embeddings.[72]

**Scale Invariant Fine-Tuning improves training stability**: A regularization technique for raising the generalization of models is virtual adversarial training. It accomplishes this by enhancing a model's robustness to adversarial instances, which are produced by introducing little input perturbations. The model is regularized to yield the same output distribution on a task-specific example as it does on an adversarial perturbation of that example. For NLU challenges, the word embedding is perturbed rather than the initial word sequence. The embedding vectors' value ranges (norms) change between various models and words. The variance gets larger for bigger models with billions of parameters, leading to some instability of adversarial training. A Scale-Invariant-Fine-Tuning (SiFT) approach is created, inspired by layer normalization, to enhance training stability by applying perturbations to the normalized word embeddings.[70] The disentangled mechanism already took into account the content and relative positions of the context words. So, we do not do much fine-tuning on the DeBERTa model. DeBERTa only needs half the data and is better than BERT and RoBERTa.

## Sentence Similarity

Sentence Similarity is the task of determining how similar two texts are in other words, Sentence similarity or semantic textual similarity is a measure of how similar two pieces of text are, or to what degree they express the same meaning [73]. Sentence similarity models convert input texts into vectors (embeddings) that capture semantic information and calculate how close (similar) they are between them. The common methods used for text similarity range from simple word-vector dot products to pairwise classification, and more recently, deep neural networks. This task is particularly useful for information retrieval and clustering/grouping.The sentence similarity scores can be used in a wide variety of applications, such as search/retrieval, nearest-neighbor or kernel-based classification methods, recommendations, and ranking tasks. Related tasks include paraphrase or duplicate identification, searching, and matching applications.

Sentence similarity is normally calculated by the following two steps:

* Obtaining the embeddings of the sentences.
* Taking the cosine similarity between them.

### all-MiniLM-L6-v2

This is a sentence transformers model, it maps sentences and paragraphs to a 384-dimensional dense vector space and can be used for tasks like clustering or semantic search.Sentence-Transformers is a Python framework for state-of-the-art sentence, text, and image embeddings. We can use this framework to compute sentence/text embeddings for more than 100 languages. These embeddings can then be compared e.g., with cosine-similarity to find sentences with similar meanings. This can be useful for semantic textual similarity, semantic search, or paraphrase mining.[74] The project aims to train sentence embedding models on very large sentence-level datasets using a self-supervised contrastive learning objective. This model used the pre-trained “nreimers/MiniLM-L6-H384-uncased” model and fine-tuned it on a 1B sentence pairs dataset. This model uses a contrastive learning objective. Given a sentence from the pair, the model should predict which out of a set of randomly sampled other sentences, was actually paired with it in our dataset.

This model is developed during the Community week using JAX/Flax for NLP & CV, organized by Hugging Face. The model was developed as part of the project: Train the Best Sentence Embedding Model Ever.

# Result and Discussion

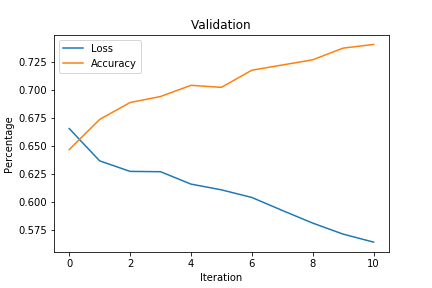
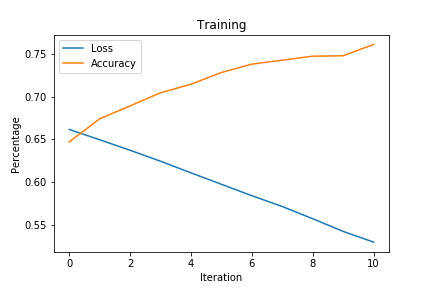
We used 3 different approaches to understand the sentiment of a Reddit post. The same dataset is used for different methods; therefore, we can safely compare accuracy metrics between methods. VADER method used as is, from NLTK library, that achieved around 50% accuracy. Given that we have 2 labels, this accuracy is as good as random.

We expected transformer models to outperform VADER, because of the complexity and potentials of these models, yet we tried VADER to make sure transformers are not overkill for this task.

DeBERTa model performed better in classifying posts based on their sentiments in comparison to VADER, by achieving 65% accuracy. This result raises the question whether the training context for the model was much more different than our context, so fine tuning it will be worth a shot.

We were able to increase the accuracy, by fine-tuning the model using our dataset. The fine-tuned model achieved 75% in the training phase and 72% in the validation stage. These results proved our hypothesis that different contexts can affect the results, even with a model trained with a billion record datasets.

The fine-tuning process results are shown in the Figure 2.



**Figure 2 Fine Tuning DeBERTa**

Another approach to do the same task, was to compare semantic similarity between each post and each label, Fear and Greed. Using all-MiniLM-L6-V2, we compared similarities of each post, by the keyword, Fear and Greed. The predicted label is determined by comparing the similarity score between the two keywords.

‘Fear’ if similarity(post,fear).score > similarity(post,greed).score else ‘Greed’

Although it is a rare occasion, we ignore the fact that these two scores can be equivalent to each other, but in our method, that case will be considered as Greed.

This way, we were able to achieve 90% accuracy without any training or fine-tuning as shown in Table 3.

**Table 3 Models' Accuracy Comparison Results**

|  |  |  |  |
| --- | --- | --- | --- |
| VADER | DeBERTa | DeBERTa (fine-tuned) | all-MiniLM-L6-v2 |
| 0.502 | 0.650 | 0.720 | 0.908 |

The results show that we can reliably predict accurate market sentiment by looking at social media data and expect changes in the crypto market before they happen. Any sign of panic or fear will be obvious in social media posts because people tend to post on social media about their next moves before doing it.

# conclusion and Future work

We were able to achieve an accurate method to detect market sentiment based on Reddit social media posts by incorporating an existing transformer model and mixing this with a new approach. We also demonstrate that methods like VADER that are supposed to work on social media textual contexts are not reliable anymore due to the high level of complexity and rapid changes in language.

Classification using state-of-the-art models may not be a good practice to solve such problems, as we showed using DeBERTa this was not the best way of predicting market sentiment based on textual data.

Future works on this subject could be done by finding different data streams, such as Twitter and creating a merged dataset from multiple sources to reduce the bias and be more inclusive demographically, due to the fact that not every crypto trader uses Reddit. Also, using different sentence similarity transformer models, and fine-tuning them with crypto-related context, could be a way in which to increase the accuracy, thus, reliability of this approach.

REFERENCES

1. Berentsen, A. and F. Schär, *A short introduction to the world of cryptocurrencies.* FRB of St. Louis Working Review, 2018.

2. Phillips, R.C. and D. Gorse. *Predicting cryptocurrency price bubbles using social media data and epidemic modelling*. in *2017 IEEE symposium series on computational intelligence (SSCI)*. 2017. IEEE.

3. Alonso-Monsalve, S., et al., *Convolution on neural networks for high-frequency trend prediction of cryptocurrency exchange rates using technical indicators.* Expert Systems with Applications, 2020. **149**: p. 113250.

4. Ante, L., *How Elon Musk's twitter activity moves cryptocurrency markets.* Available at SSRN 3778844, 2021.

5. Huang, X., et al. *Lstm based sentiment analysis for cryptocurrency prediction*. in *International Conference on Database Systems for Advanced Applications*. 2021. Springer.

6. Conti, M., et al., *A survey on security and privacy issues of bitcoin.* IEEE Communications Surveys & Tutorials, 2018. **20**(4): p. 3416-3452.

7. Nakamoto, S., *Bitcoin: A peer-to-peer electronic cash system.* Decentralized Business Review, 2008: p. 21260.

8. Mao, H., S. Counts, and J. Bollen, *Predicting financial markets: Comparing survey, news, twitter and search engine data.* arXiv preprint arXiv:1112.1051, 2011.

9. Keningi, E., *Predicting the price of Bitcoin using the sentiment of popular Bitcoin-related Tweets.* 2022.

10. Marshall, R., *Bitcoin: Where two worlds collide.* Bond L. Rev., 2015. **27**: p. 89.

11. Baur, D.G., K. Hong, and A.D. Lee, *Bitcoin: Medium of exchange or speculative assets?* Journal of International Financial Markets, Institutions and Money, 2018. **54**: p. 177-189.

12. Pang, B. and L. Lee, *Opinion mining and sentiment analysis.* Foundations and Trends® in information retrieval, 2008. **2**(1–2): p. 1-135.

13. Taboada, M., et al., *Lexicon-based methods for sentiment analysis.* Computational linguistics, 2011. **37**(2): p. 267-307.

14. Khuwaja, P., S.A. Khowaja, and K. Dev, *Adversarial learning networks for FinTech applications using heterogeneous data sources.* IEEE Internet of Things Journal, 2021.

15. Kraaijeveld, O. and J. De Smedt, *The predictive power of public Twitter sentiment for forecasting cryptocurrency prices.* Journal of International Financial Markets, Institutions and Money, 2020. **65**: p. 101188.

16. Hutto, C. and E. Gilbert. *Vader: A parsimonious rule-based model for sentiment analysis of social media text*. in *Proceedings of the international AAAI conference on web and social media*. 2014.

17. Pano, T. and R. Kashef, *A complete VADER-based sentiment analysis of bitcoin (BTC) tweets during the era of COVID-19.* Big Data and Cognitive Computing, 2020. **4**(4): p. 33.

18. Tan, X. and R. Kashef. *Predicting the closing price of cryptocurrencies: a comparative study*. in *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems*. 2019.

19. Vaswani, A., et al., *Attention is all you need.* Advances in neural information processing systems, 2017. **30**.

20. Sun, C., L. Huang, and X. Qiu, *Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence.* arXiv preprint arXiv:1903.09588, 2019.

21. Devin, J., et al. *BERT: Pre-training of deep bidirectional transformers for language onderstanding*. in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019.

22. Paltoglou, G. and M. Thelwall, *Twitter, MySpace, Digg: Unsupervised sentiment analysis in social media.* ACM Transactions on Intelligent Systems and Technology (TIST), 2012. **3**(4): p. 1-19.

23. Thakor, P. and S. Sasi, *Ontology-based sentiment analysis process for social media content.* Procedia Computer Science, 2015. **53**: p. 199-207.

24. Manning, C.D., et al. *The Stanford CoreNLP natural language processing toolkit*. in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014.

25. Socher, R., et al. *Recursive deep models for semantic compositionality over a sentiment treebank*. in *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013.

26. High, R., *The era of cognitive systems: An inside look at IBM Watson and how it works.* IBM Corporation, Redbooks, 2012. **1**: p. 16.

27. Agarwal, A., et al. *Sentiment analysis of twitter data*. in *Proceedings of the workshop on language in social media (LSM 2011)*. 2011.

28. Wang, Y., et al. *Attention-based LSTM for aspect-level sentiment classification*. in *Proceedings of the 2016 conference on empirical methods in natural language processing*. 2016.

29. Cao, X. and J. Sun, *Exploring the effect of overload on the discontinuous intention of social media users: An SOR perspective.* Computers in human behavior, 2018. **81**: p. 10-18.

30. Neri, F., et al. *Sentiment analysis on social media*. in *2012 IEEE/ACM international conference on advances in social networks analysis and mining*. 2012. IEEE.

31. Dhaoui, C., C.M. Webster, and L.P. Tan, *Social media sentiment analysis: lexicon versus machine learning.* Journal of Consumer Marketing, 2017.

32. Yue, L., et al., *A survey of sentiment analysis in social media.* Knowledge and Information Systems, 2019. **60**(2): p. 617-663.

33. Pak, A., et al. *Normalization of term weighting scheme for sentiment analysis*. in *Language and Technology Conference*. 2011. Springer.

34. Dulău, T.-M. and M. Dulău, *Cryptocurrency–sentiment analysis in social media.* Acta Marisiensis. Seria Technologica, 2019. **16**(2): p. 1-6.

35. Nemes, L. and A. Kiss, *Social media sentiment analysis based on COVID-19.* Journal of Information and Telecommunication, 2021. **5**(1): p. 1-15.

36. Poecze, F., C. Ebster, and C. Strauss, *Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts.* Procedia computer science, 2018. **130**: p. 660-666.

37. Dridi, A. and D. Reforgiato Recupero, *Leveraging semantics for sentiment polarity detection in social media.* International Journal of Machine Learning and Cybernetics, 2019. **10**(8): p. 2045-2055.

38. Bae, Y. and H. Lee, *Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers.* Journal of the American Society for Information Science and technology, 2012. **63**(12): p. 2521-2535.

39. Singh, N.K., D.S. Tomar, and A.K. Sangaiah, *Sentiment analysis: a review and comparative analysis over social media.* Journal of Ambient Intelligence and Humanized Computing, 2020. **11**(1): p. 97-117.

40. Naeem, M.A., I. Mbarki, and S.J.H. Shahzad, *Predictive role of online investor sentiment for cryptocurrency market: evidence from happiness and fears.* International Review of Economics & Finance, 2021. **73**: p. 496-514.

41. Abraham, J., et al., *Cryptocurrency price prediction using tweet volumes and sentiment analysis.* SMU Data Science Review, 2018. **1**(3): p. 1.

42. Mai, F., et al., *How does social media impact Bitcoin value? A test of the silent majority hypothesis.* Journal of management information systems, 2018. **35**(1): p. 19-52.

43. Kim, Y.B., et al., *Predicting fluctuations in cryptocurrency transactions based on user comments and replies.* PloS one, 2016. **11**(8): p. e0161197.

44. McNally, S., J. Roche, and S. Caton. *Predicting the price of bitcoin using machine learning*. in *2018 26th euromicro international conference on parallel, distributed and network-based processing (PDP)*. 2018. IEEE.

45. Georgoula, I., et al., *Using time-series and sentiment analysis to detect the determinants of bitcoin prices.* Available at SSRN 2607167, 2015.

46. Matta, M., I. Lunesu, and M. Marchesi. *Bitcoin Spread Prediction Using Social and Web Search Media*. in *UMAP workshops*. 2015.

47. Zhang, Z., et al., *XGBoost-based on-line prediction of seam tensile strength for Al-Li alloy in laser welding: Experiment study and modelling.* Journal of Manufacturing Processes, 2021. **64**: p. 30-44.

48. Jain, A., et al. *Forecasting price of cryptocurrencies using tweets sentiment analysis*. in *2018 eleventh international conference on contemporary computing (IC3)*. 2018. IEEE.

49. Symeonidis, S., D. Effrosynidis, and A. Arampatzis, *A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis.* Expert Systems with Applications, 2018. **110**: p. 298-310.

50. Lyu, H., et al., *Sense and sensibility: Characterizing social media users regarding the use of controversial terms for covid-19.* IEEE Transactions on Big Data, 2020. **7**(6): p. 952-960.

51. Stenqvist, E. and J. Lönnö, *Predicting Bitcoin price fluctuation with Twitter sentiment analysis*. 2017.

52. Sul, H.K., A.R. Dennis, and L. Yuan, *Trading on twitter: Using social media sentiment to predict stock returns.* Decision Sciences, 2017. **48**(3): p. 454-488.

53. Akyildirim, E., A. Goncu, and A. Sensoy, *Prediction of cryptocurrency returns using machine learning.* Annals of Operations Research, 2021. **297**(1): p. 3-36.

54. Chevallier, J., D. Guégan, and S. Goutte, *Is It Possible to Forecast the Price of Bitcoin?* Forecasting, 2021. **3**(2): p. 377-420.

55. Verma, M. and P. Sharma, *Money Often Costs Too Much: A Study to Investigate The Effect Of Twitter Sentiment On Bitcoin Price Fluctuation.* 2020.

56. Jahjah, F.H. and M. Rajab, *Impact of Twitter Sentiment Related to Bitcoin on Stock Price Returns.* Journal of Engineering, 2020. **26**(6): p. 60-71.

57. Inamdar, A., et al. *Predicting cryptocurrency value using sentiment analysis*. in *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. 2019. IEEE.

58. Wołk, K., *Advanced social media sentiment analysis for short‐term cryptocurrency price prediction.* Expert Systems, 2020. **37**(2): p. e12493.

59. Bollen, J., H. Mao, and X. Zeng, *Twitter mood predicts the stock market.* Journal of computational science, 2011. **2**(1): p. 1-8.

60. Prosky, J., et al., *Sentiment predictability for stocks.* arXiv preprint arXiv:1712.05785, 2017.

61. Rather, A.M., A. Agarwal, and V. Sastry, *Recurrent neural network and a hybrid model for prediction of stock returns.* Expert Systems with Applications, 2015. **42**(6): p. 3234-3241.

62. Nie, S. and Q. Ji. *Feature learning using Bayesian linear regression model*. in *2014 22nd International Conference on Pattern Recognition*. 2014. IEEE.

63. Zhang, W., et al., *A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges.* arXiv preprint arXiv:2203.01054, 2022.

64. Garcia, D. and F. Schweitzer, *Social signals and algorithmic trading of Bitcoin.* Royal Society open science, 2015. **2**(9): p. 150288.

65. Stenqvist, E. and J. Lönnö. *Predicting Bitcoin price fluctuation with Twitter sentiment analysis*. 2017.

66. Wright, J. and O. Anise, *Don’t@ me: Hunting twitter bots at scale.* Blackhat USA, 2018.

67. Anwar, A. and U. Yaqub. *Bot detection in twitter landscape using unsupervised learning*. in *The 21st Annual International Conference on Digital Government Research*. 2020.

68. Devlin, J., et al., *Bert: Pre-training of deep bidirectional transformers for language understanding.* arXiv preprint arXiv:1810.04805, 2018.

69. Rothman, D., *Transformers for Natural Language Processing: Build innovative deep neural network architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and more*. 2021: Packt Publishing Ltd.

70. He, P., et al., *Deberta: Decoding-enhanced bert with disentangled attention.* arXiv preprint arXiv:2006.03654, 2020.

71. Silva, E.H. and R.M. Marcacini. *Aspect-based sentiment analysis using BERT with Disentangled Attention*. in *Proceedings of the 8th ICML Workshop on Automated Machine Learning (AutoML 2021), Virual Event*. 2021.

72. He, P., J. Gao, and W. Chen, *Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.* arXiv preprint arXiv:2111.09543, 2021.

73. Farouk, M., *Measuring sentences similarity: a survey.* arXiv preprint arXiv:1910.03940, 2019.

74. Hertling, S., J. Portisch, and H. Paulheim, *KERMIT--A Transformer-Based Approach for Knowledge Graph Matching.* arXiv preprint arXiv:2204.13931, 2022.