SENTIMENT-DRIVEN CRYPTOCURRENCY PRICE PREDICTION: A MACHINE LEARNING APPROACH UTILIZING HISTORICAL DATA AND SOCIAL MEDIA SENTIMENT ANALYSIS

Saachin Bhatt, Mustansar Ghazanfar, and Mohammad Hossein Amirhosseini

University of East London, London, E16 2RD, United Kingdom

ABSTRACT

This research explores the impact of social media sentiments on predicting Bitcoin prices using machine learning models, integrating on-chain data, and applying a Multi Modal Fusion Model. Historical crypto market, on-chain, and Twitter data from 2014 to 2022 were used to train models including K-Nearest Neighbors, Logistic Regression, Gaussian Naive Bayes, Support Vector Machine, Extreme Gradient Boosting, and Multi Modal Fusion. Performance was compared with and without Twitter sentiment data which was analysed using the Twitter-roBERTa and VADAR models. Inclusion of sentiment data enhanced model performance, with Twitter-roBERTa-based models achieving an average accuracy score of 0.81. The best performing model was an optimised Multi Modal Fusion model using Twitter-roBERTa, with an accuracy score of 0.90. This research underscores the value of integrating social media sentiment analysis and on-chain data in financial forecasting, providing a robust tool for informed decision-making in cryptocurrency trading.

KEYWORDS

Cryptocurrency, Bitcoin Price, Social Media, Sentiment Analysis, Machine Learning, Classification

1. INTRODUCTION

The potential of Bitcoin, the first and most widely used cryptocurrency, to disrupt traditional financial systems and provide an alternative to fiat currencies has attracted significant attention. With its growing acceptance as a form of payment, there is increasing interest in predicting its price movements. Machine learning algorithms and social media data, specifically Twitter, offer a promising approach to forecasting Bitcoin market trends [20, 12]. Previous studies have demonstrated the potential of using social media data, such as Twitter posts, to predict stock market trends [4] and Bitcoin market behavior [13,17]. However, few studies have explored the use of on-chain data and Multi Modal Fusion models in predicting Bitcoin price movements, which are the novel aspects of this research.

The goal of this study is to develop a machine learning-based model for predicting the price of Bitcoin using market and on-chain data, with a focus on Twitter sentiment analysis. The research builds upon previous studies that have used a combination of news articles and Twitter data to forecast Bitcoin price movements and those that have focused on the sentiment of tweets to predict market trends. The results of our study will provide valuable information for investors, market participants, and exchanges in managing risk and making decisions.

In this paper, relevant literature on the use of machine learning, on-chain data, and social media data in predicting Bitcoin market trends is reviewed. The methodology and results from this study are then presented, which includes a detailed analysis of the impact of Twitter sentiment on predicting the price of Bitcoin, the incorporation of on-chain data, and the application of a Multi Modal Fusion model. Finally, we will discuss the limitations and future directions for our research.

2. LITERATURE REVIEW

In recent years, various studies have been conducted to investigate the impact of Twitter on market price and to develop models to predict the market price based on Twitter data.Abraham et al. [1] aimed to calculate the sentiment of live tweets and its effect on the market price. They collected live Twitter data and stock market data using APIs and thenused Naïve Bayes classification to calculate the sentiment of live tweets and an XGBoost regression model to predict market price.Tandon et al. [2] also investigated the impact of Twitter on cryptocurrency. They collected Elon Musk tweets and historical Bitcoin market data and used an ARIMA model for price prediction.

Moreover, Jaquart et al. [3] aimed to predict the short-term price of Bitcoin using features from various market indexes. They gathered data from multiple sources and compared the performance of different machine learning models, including neural networks, tree-based models, and ensemble models. They found that the LSTM model performed the best on the higher time frames.Basilio and Toriola [4] also used LSTM neural networks and sentiment analysis to predict the price of Bitcoin using Twitter data. The study achieved a Root Mean Squared Error of 0.014 when using VADAR sentiment analysis to compliment the LSTM prediction. The results suggest that deep learning and sentiment analysis can be valuable tools for predicting Bitcoin prices using Twitter data. However, the data collected was over a short period of time (February to June) and was a sample from a population so may not have been representative.

Developing from this, Wang et al.[22] proposed a hybrid LSTM-CNN approach for stock market prediction. They combined LSTMs and CNNs to learn the patterns in stock prices. The hybrid LSTM-CNN model outperforms traditional machine learning models with an accuracy of 90%.As well as LSTM models, the authors explored various machine learning techniques to model the nonlinear relationship between bitcoin prices and social sentiment data and predict the price values with some lead time and found that the sentiment data model is superior in capturing the non-linear relationship compared to the conventional methods of technical indicators and decision trees, while the neural network models are robust and offer better accuracy in predicting bitcoin price. [23]

Furthermore, Joshiand Rao [5]aimed to predict stock trends using news sentiment analysis. They collected news articles from multiple sources, including Bloomberg and Reuters, and used sentiment analysis to determine the sentiment of each article. Then, they used various machine learning algorithms, including SVM, Random Forest and Naïve Bayes, to predict the trend of the stock based on the sentiment of the news articles. The study found that incorporating sentiment analysis into their models improved their accuracy, with the Random Forest algorithm achieving the highest accuracy for all test cases. The study suggests that news sentiment analysis can be a valuable predictor of stock trends when combined with machine learning algorithms. Similar findings have been claimed from another study [10].

The use of multi-modal fusion in stock market prediction has attracted significant attention from researchers in recent years, as it holds the potential to enhance the accuracy and reliability of forecasting models [11]. Multi-modal fusion combines various data sources, such as textual, nu-

merical, and visual data, to capture a broader range of market factors and trends [12]. Furthermore, multi-modal fusion leverages advanced machine learning techniques, such as deep learning and natural language processing, to analyze and integrate heterogeneous data sources effectively[13]. Several studies have reported improved stock market prediction performance when using multi-modal fusion compared to single-modal approaches [14]. Critien et al. [24] employed a multi-modal fusion strategy for forecasting Bitcoin's short-term price. In their investigation, the fusion of Twitter data, market data, and news data amplified the accuracy of their prediction model, further bolstering the claim for Twitter's relevance in market prediction. However, further research is needed to optimize the fusion process and address the challenges associated with handling large-scale, high-dimensional, and noisy data inherent in financial markets [15].

Twitter data has demonstrated considerable potential for predicting market prices. However, the efficacy of such predictions hinges on the choice of an appropriate predictive model, often leading to the use of time series models and neural networks. The model's effectiveness depends on the specificity of the task and dataset, with the inclusion of sentiment analysis and multi-modal fusion significantly improving prediction precision. Neural networks are particularly promising when dealing with noisy data. Yet, there are challenges to overcome. Variability in Twitter data quality, the complex nature of the relationship between tweets and market prices, and the prevalence of noisy data that complicates data cleaning procedures all present obstacles. Despite these hurdles, the studies reviewed suggest a promising future for utilizing Twitter data in market price predictions. Therefore, further research is essential to tackle these challenges and improve the methods for harnessing Twitter data for effective market price predictions.

Sentiment analysis is the process of using natural language processing and computational linguistics to identify and extract subjective information from text. It is often used to analyse the attitudes, opinions, and emotions expressed in text data to better understand how people feel about a particular topic or subject [6]. To conduct Sentiment Analysis on the Tweets, multiple techniques have been explored in this research.

2.1. TextBlob

TextBlob is a library developed from the Natural Language Tool Kit (NLTK) for Natural Language Processing. This includes Sentiment Analysis. There are two main outputs from TextBlob: polarity and subjectivity. The polarity values indicate the extent to which a text is positive or negative. The subjectivity describes how much the text is subjective or objective. TextBlob uses a rule-based method. There is a pre-defined Lexicon and conditions that constitute to what's classified as positive, negative, and neutral. Tweets, like most social media posts, can be very informal and hence the words and phrases used can stem from a form of informal writing e.g., sarcasm and irony [7]. Being a rule-based model, TextBlob won't be able to pick up on these and hence can incorrectly classify a lot of Tweets, making it non-ideal for our dataset.

2.2. VADAR

VADAR (Valence Aware Dictionary for Sentiment Reasoning) utilises both a lexicon and rule-based approach. It uses a combination of pre-labelled lexical features (key words which are labelled as having a positive or negative sentiment) to classify new words into having either a positive or negative sentiment. The compound VADAR score is a normalized, weighted composite score that is calculated based on the sum of all the lexicon ratings which have been standardized to range between -1 (most extreme negative) and +1 (most extreme positive) [8]. Despite also being rule-based, it is a lexicon that is used to express social media sentiment which supports emoticons in the generation of the Sentiment, making it a strong model for the Twitter data [8].

2.3. BERT, RoBERTa & Twitter-RoBERTa

RoBERTa is an extension of BERT (Bidirectional Encoder Representations from Transformers) which is a transformer based, pre-training model which make use of embedding vector space which allows for a deeper understand of context rather than just analysis on a word-by-word basis. BERT was originally designed to create pre-training bi-directional representations to extract context-specific information from the input. The bi-directional nature of BERT allows it to read and understand the text from left-to-right and right-to-left, ensuring minimal information loss. RoBERTa is a state-of-the-art BERT model which has been trained on 160GB of additional data, hence increasing the robustness of BERT and performing with much better results. Twitter-RoBERTa (a Hugging Face model) is a model that has been pre-training of a huge number of Tweets from twitter [9]. This makes it ideal to use for our Twitter dataset. The BERT model is typically composed of multiple layers of self-attention and feedforward neural networks, which allow it to process input sequences of any length. The self-attention layers of the BERT model use a multi-headed attention mechanism, which allows the model to simultaneously attend to different parts of the input sequence and to weight those parts differently when generating output. This makes the BERT model particularly effective at natural language processing tasks, such as language translation and sentiment analysis. In addition to its self-attention and feedforward layers, the BERT model also uses a technique called masked language modelling, which involves randomly masking a portion of the input sequence and training the model to predict the masked tokens. This allows the model to learn the relationships between different words in the input sequence and to better understand the meaning and context of the words in a sentence [5].

3. METHODOLOGY

The diagram below explains the methodology used in this research. We calculated the sentiment of Tweets using VADAR and Twitter-RoBERTa, and trained five classification models including Naïve Bayes, K-Nearest Neighbours, Support Vector Machine, Logistic Regression, and XGBoost, as well as a Multi Modal Fusion Model.

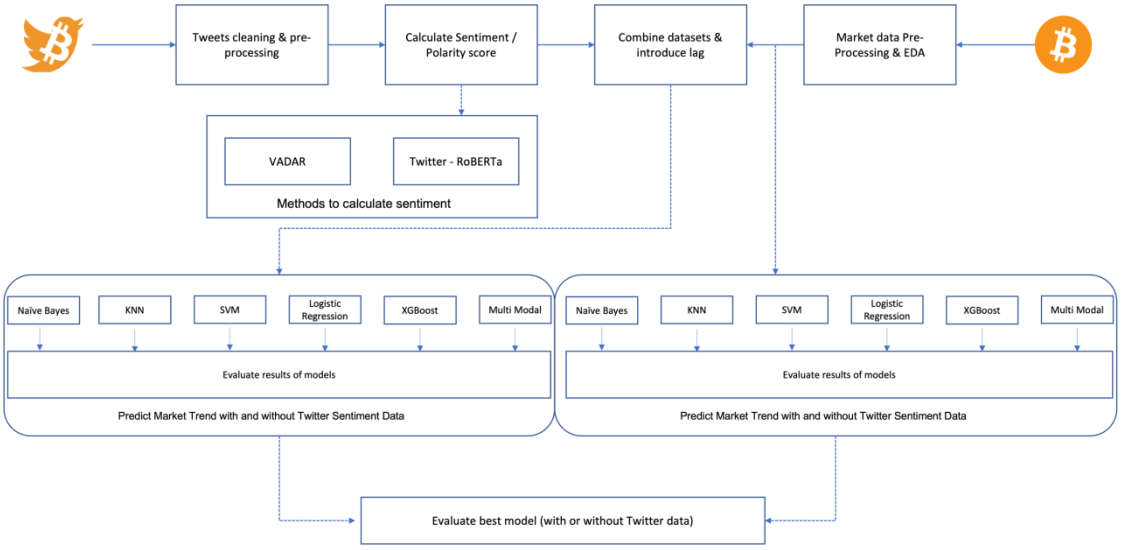


Figure 1. Experimental Pipeline

The experimental process used in this project follows the pipeline demonstrated in Figure 1. We are looking to investigate whether the inclusion of Twitter sentiment data improves the perfor-

mance models when predicting market trend. Thus, as shown in Figure 1, we keep the models and evaluation technique the same when testing with and without Twitter sentiment.

3.1. Dataset

There are three main datasets used in this research including historic Bitcoin tweets , historical Bitcoin market data and historical Bitcoin specific on-chain Blockchain data, all of which are obtained via Kaggle repository. The historic Bitcoin market data provides the daily attributes including (1) open, (2) low, (3) close, (4) high, and (5) volume traded between 2014 and 2022. The Twitter data has been filtered to only contain those tweets relating to Bitcoin for the same period.

3.2. Exploratory Data Analysis

EDA helps to identify potential issues with the data and to ensure that the data is suitable for the intended analysis. It also helps to identify trends and patterns in the data that may not be immediately obvious, and to develop a better understanding of the data.

As we start to analyse the stock market data, we can utilize Plotly to create a detailed candle stick chart that incorporates all the relevant attributes in our dataset. This chart, shown in Figure 2, allows us to easily visualise and interpret the trends and patterns in the data.

Figure 2. Market price over time

A visual inspection of the stock market data reveals an exponential increase in market price starting in late 2020, with a clear bull-bear pattern emerging between 2021 and 2022. This suggests that the market experienced significant growth during this time, likely driven by various factors such as economic conditions, investor sentiment, and company performance. After calculating the sentiment scores for the tweets, we can align them with the corresponding market prices to see if there is a visible relationship between the two. This may provide insights into how sentiment influences market behaviour and could potentially be used to improve the accuracy of market trend predictions. We then moved on to exploring the relationships between the attributes within our dataset.

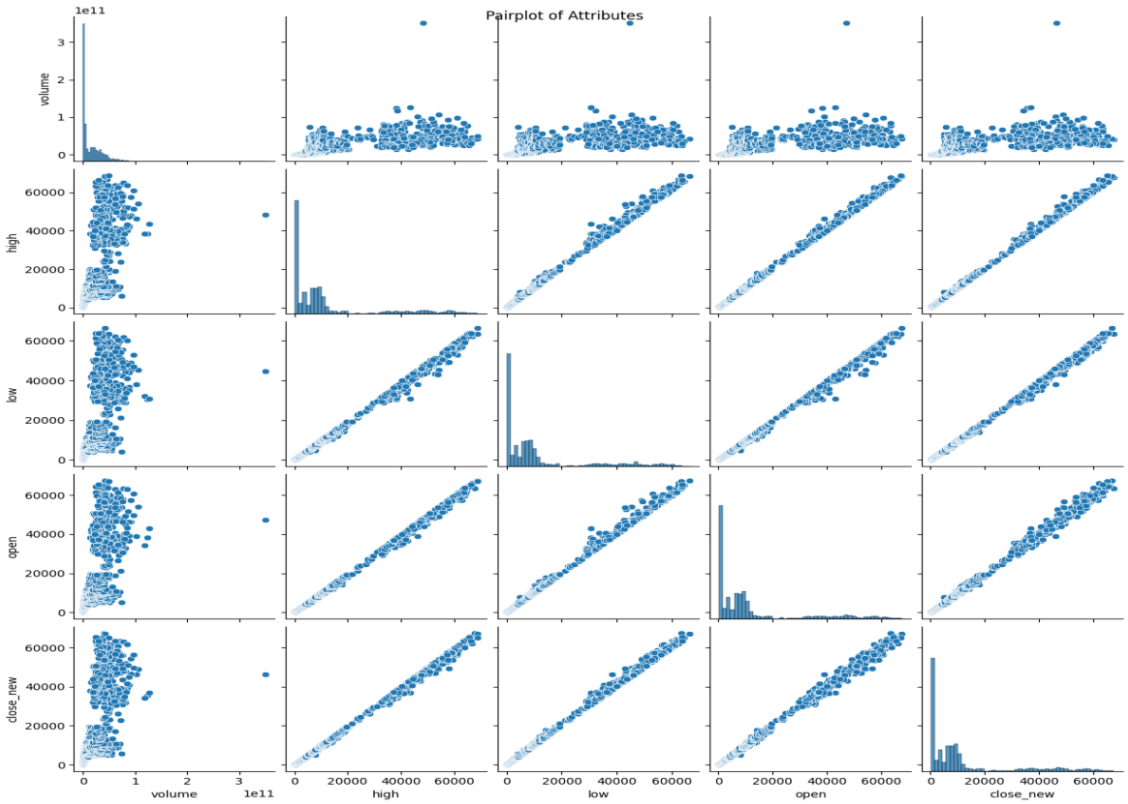


Figure 3. Relationship between attributes

Figure 3 demonstrates the strong positive correlations between attributes, as evidenced by the diagonal scatter plots. Additionally, the attribute distributions are heavily skewed to the right.

3.3. Modelling

When it comes to Natural Language Processing(NLP) and sentiment analysis, pre-processing steps are required to produce the best results. Pre-processing is an essential phase in NLP that involves preparing text data for further analysis. Gupta et al. [21] explained that the pre-processing steps used in NLP may vary depending on the specific application and the goals of the analysis, but some common steps include:

- Tokenization: This involves dividing the text into smaller units called tokens, which can be individual words, phrases, or other elements of the text.
- Stop word removal: This involves removing common words such as ”the” and ”a” that are not relevant to the analysis and do not add meaningful information to the text.
- Stemming and lemmatization: These techniques are used to reduce words to their base form, which can help reduce the dimensionality of the data and make it easier to analyse.
- Part-of-speech tagging: This involves identifying the part of speech (e.g. noun, verb, adjective) of each word in the text, which can be useful for understanding the structure and meaning of the text.
- Normalization: This involves converting text to a standard format, such as lowercase, to reduce variability and make it easier to analyse.

In this research, data transformation and cleaning process were applied to the tweets. As a results, duplicated entries were removed in the first step. Then, user mentions (words that begin with @ + username) were removed. URL mentioned (e.g., 'https') were also removed as this information is not important for calculating the sentiment of the tweet. Furthermore, stop words ('a', 'it', 'the' etc.) were removed as they are classified as low-level information, so removing this allows the models to focus on the important information. Removing the stop words also reduces the size of the dataset so can allow for faster run time. In addition, punctuations were removed. These steps were followed by word stemming and lemmatisation. Table 1 shows the cleaning process for a tweet.

Table 1. Cleaning process for a tweet.

Original Tweet	Remove Punctuation	Tokenise	Remove Stop Words	Stemming	Lemmatising
$BTC a big chance in a billion!	BTC a big chance in a billion	[btc, a, big, chance, in, a, billion]	[btc, big, chance, billion]	[btc, big, chanc, billion]	[btc, big, chance, billion]

Once the tweets were cleaned, we then merged the dataset with the cleaned Bitcoin market and on-chain data. Table 2 shows the attributes and samples in the Bitcoin market data. Descriptions for the attributes are provided in table 3.

Table 2. Raw Bitcoin Market Data

.

Date	Volume	High	Low	Open	Close
2022-09-17	21056800	468.174011	452.421997	465.864014	457.334015
2022-09-18	34483200	456.859985	413.104004	456.859985	424.440002
2022-09-19	37919700	427.834991	384.532013	424.102997	394.795990

Table 3. Bitcoin On-Chain Data.

Date	TxVolume	TxCount	avgDifficulty	generatedCoins	paymentCount
2022-09-17	240000000	77185	2.982973e+10	4653	142642
2022-09-18	34483200	69266	6.653303e+12	4475	120597
2022-09-19	37919700	59636	2.98297331e+10	3425	423180

Table 4. Descriptions for the attributes in Bitcoin Market and On-Chain data.

Attribute	Description
Volume	Total number of shares that were traded
High	Highest price at which the stock was traded during that period (day)
Low	Lowest price at which the stock was traded during the period (day)
Open	Price at which the stock began trading at the start of the period (day)
Close	Price at which the stock ended trading at the end of the period (day)
TxVolume	Total transaction volume in a day
TxCount	Total transactions executed in a given period (day)
avgDifficulty	Average computational complexity required to mine a new block in the Blockchain
generatedCoins	Total number of coins mined during the period (day)
paymentCount	Total number of unique payments executed within the blockchain network during a given period (day)

The values for 'Volume' are much greater than the other attributes. As a result, we normalised the dataset using the MinMaxScalar() library from Sklearn. Table 5 shows the samples after normalisation.

Table 5. Bitcoin Market and On-Chain Data after Normalisation

	Date		
Attribute	2022-09-17	2022-09-18	2022-9-19
Volume	0.000043	0.000081	0.000091
High	0.003739	0.003574	0.003151
Low	0.004243	0.003649	0.003217
Open	0.004289	0.004155	0.003669
Close	0.004144	0.003655	0.003216
TxVolume	0.000042	0.000086	0.000093
TxCount	0.003721	0.003649	0.003188
avgDifficulty	0.004312	0.003673	0.003244
generatedCoins	0.004198	0.004126	0.003653
paymentCount	0.004123	0.003612	0.003229
avgDifficulty	0.005829	0.004312	0.003673
generatedCoins	0.004254	0.004198	0.004126
paymentCount	0.004323	0.004123	0.003612

We went onto predicting the sentiment using both VADAR and Twitter-RoBERTa. Table 6 demonstrates some samples inthe new dataset which includes the clean tweets and the predicted sentiments.

Table 6. Sentiment Dataset

Date	Lemmatised Tweets	Sentiment
2022-09-17	[btc, big, chance, billion]	0.022796
2022-09-18	[btc, best, nft, stock]	0.055593
2022-09-19	[bitcoin, expect, rise]	0.059745

The purpose is to predict market trend whether tomorrow's closing price is greater or less than today's closing price. For this research, we assumed that the features used to predict market trend took affect after one day (including sentiment), hence we lagged each feature by 1 day. Table 7 shows the samples in the final dataset with 1 day lag.

$$next\ day\ close_n = close_{n+1} \quad (1)$$

$$trend_n = next\ day\ close_n - close_n \quad (2)$$

Table 7. Complete Modelling Dataset with 1 Day Lag

	Date		
Attribute	2022-09-17	2022-09-18	2022-9-19
Volume	0.000062	0.000043	0.000081
High	0.003219	0.003739	0.003574
Low	0.003918	0.004243	0.003649
Open	0.004263	0.0042289	0.004155
Close	0.003971	0.004144	0.003655
Next Day Close	0.003655	0.003216	0.003425
Lagged Sentiment	0.0013893	0.022796	0.055593
Trend	-0.000488	-0.000440	0.000209
TxVolume	0.000050	0.000042	0.000086
TxCount	0.002975	0.003721	0.003649
avgDifficulty	0.005829	0.004312	0.003673
generatedCoins	0.004254	0.004198	0.004126
paymentCount	0.004323	0.004123	0.003612

3.4. Multi Modal Fusion

We used a deep learning approach that combined sentiment analysis of Twitter sentiment with on-chain attributes data such as transaction volume, difficulty, and payment count. The novelty of our approach lies in its ability to extract and combine multiple types of data, each with its own unique temporal characteristics and properties, using an LSTM model. The LSTM model effectively captures both the short-term dynamics of the sentiment analysis and the longer-term trends of the on-chain attributes data.

Specifically, our model included two separate branches: one for the sentiment analysis modality and another for the on-chain attributes modality. Each branch processed its respective modality independently, with the sentiment analysis branch using a dense layer and the on-chain attributes branch using an LSTM layer. The output of each branch was then concatenated, followed by another dense layer and a final sigmoid activation layer to produce the prediction.

We also included a feature selection step in our model to determine the relative importance of each modality. This involved training the model with and without specific modalities and comparing the resulting performance. Additionally, we included hyperparameter tuning to optimize the model's performance and ensure appropriate complexity.

3.5. Classification

We split the dataset into training set and test set using train_test_split() method from Sklearn library. 80% of the data was used for training, 20% was used for testing and 'random_state' parameter was set to 0. Moreover, hyperparameter tuning is an important step in the process of building a classification model. In this research, we explored hyperparameter tuning for five different classifiers including Support Vector Machine (SVM), Logistic Regression (LR), K-Nearest Neighbours (KNN), Gaussian Naïve Bayes, and XGBoost. We used various techniques for hyperparameter tuning including GridSearchCV, RandomizedSearchCV, and BayesSearchCV.

Table 8 shows a summary of the parameters chosen for each model. For all five models, all possible combinations of the hyperparameters were investigated during the hyperparameter tuning process and the combinations presented in table 8 produced the best results.

Table 8. Hyperparameter tuning for the proposed models

Model	Parameters	Value
SVM	C	0.1
	Gamma	10
	Kernel	poly
LR	C	3.364
	Penalty	L2
	Solver	Saga
KNN	N_neighbours	7
	P	2
	Weights	Uniform
XGBoost	Colsample_bytree	0.3
	Gamma	0
	Max_depth	2
	N_estimators	100
Gaussian Naïve Bayes	priors	None
	var_smoothing	1e-9
Multi Modal Fusion	lstm_units	53
	dense_units	116
	dropout_rate	0.393
	optimizer	Adam

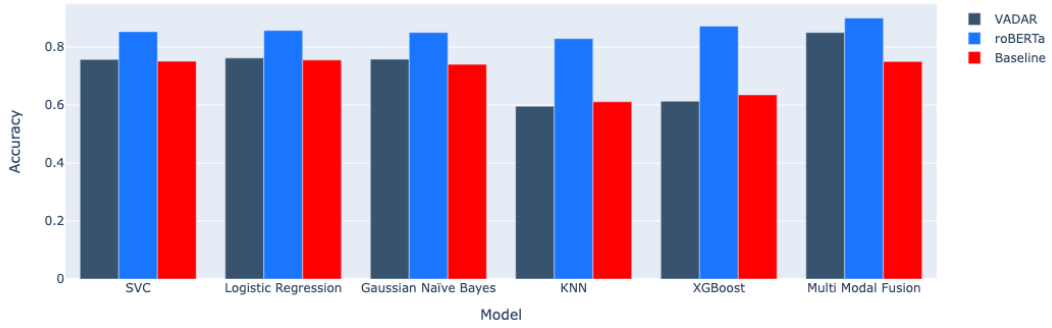
4. RESULTS

The aim of this paper was to determine whether the inclusion of Twitter data provides greater performance in our models when trying to predict Bitcoin market trend. As a results, we compared the performance of five different classification models and a Multi Modal Fusion model for predicting the market trend of Bitcoin using the VADER and Twitter-roBERTa sentiment analysis models. The models were trained and tested using two different datasets, one with sentiment data included and the other without. The performance of the implemented models have been evaluated and compared. Table 9 shows the accuracy, F1 score, precision, and recall values for each model with sentiment data and without sentiment data.

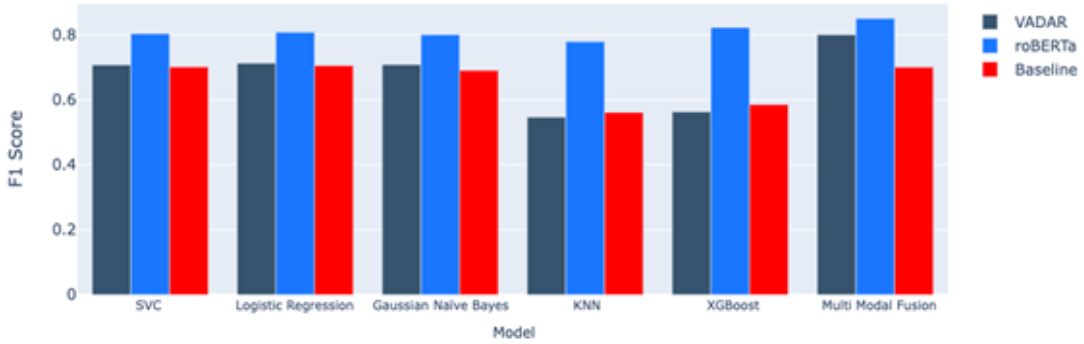
Table 9. Modelling Results – Accuracy, F1 score, Precision, and Recall

Machine Learning Model	Sentiment Model	Accuracy		F1 Score		Precision		Recall	
		With Sentiment	Without Sentiment	With Sentiment	Without Sentiment	With Sentiment	Without Sentiment	With Sentiment	Without Sentiment
SVC	VADAR	0.757	0.751	0.707	0.701	0.700	0.701	0.715	0.701
	roBERTa	0.853	0.751	0.803	0.701	0.800	0.701	0.807	0.701
Logistic Regression	VADAR	0.762	0.755	0.712	0.705	0.705	0.705	0.720	0.705
	roBERTa	0.857	0.755	0.807	0.705	0.805	0.705	0.812	0.705
Gaussian Naïve Bayes	VADAR	0.758	0.740	0.708	0.690	0.702	0.690	0.717	0.690
	roBERTa	0.850	0.740	0.800	0.690	0.802	0.690	0.809	0.690
KNN	VADAR	0.596	0.611	0.546	0.561	0.535	0.561	0.555	0.561
	roBERTa	0.829	0.611	0.779	0.561	0.777	0.561	0.785	0.561
XGBoost	VADAR	0.613	0.635	0.563	0.585	0.552	0.585	0.570	0.585
	roBERTa	0.872	0.635	0.822	0.585	0.820	0.585	0.827	0.585
Multi Modal Fusion	VADAR	0.850	0.750	0.800	0.700	0.795	0.700	0.805	0.700
	roBERTa	0.900	0.750	0.850	0.700	0.850	0.700	0.855	0.700

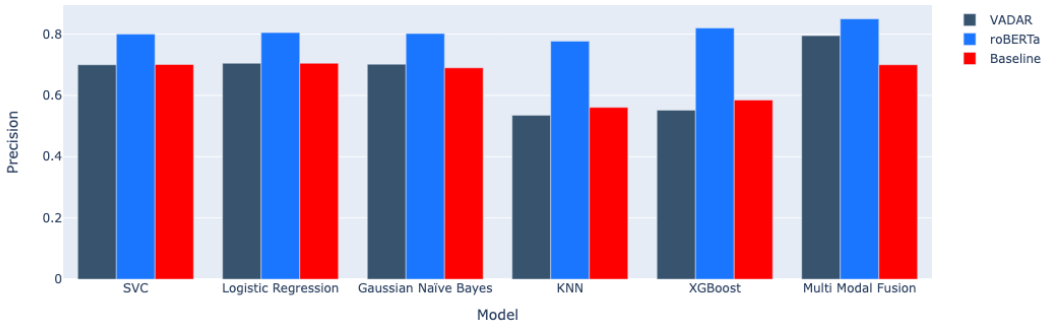
It's evident that using Twitter-RoBERTa as a sentiment model results in consistently better performance. Figure 5 demonstrates the results of comparing the performance of the models.



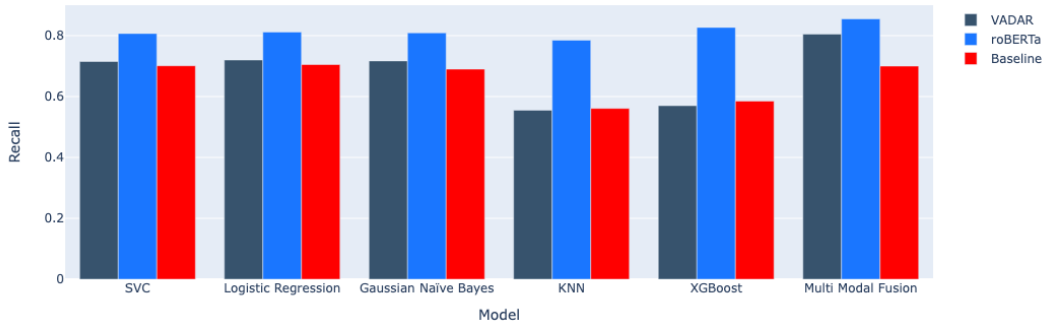
(A) Comparing the accuracy score for the models



(B) Comparing the F1 score for the models



(C)Comparing the precision value for the models



(D) Comparing the recall valuefor the models

Figure 5. Comparing the performance of the models with and without sentiment data

The results show that including sentiment data improves the performance of most models. Multi Modal Fusion with Twitter-roBERTa achieved the highest accuracy score of 0.90, F1 score of 0.85, Precision score of 0.85, and recall score of 0.85, indicating its effectiveness in predicting market trends based on the given data. XGBoost with Twitter-roBERTa sentiment analysis also produced impressive results, achieving an accuracy score of 0.87, and F1, Precision, and Recall scores of 0.82. However, models such as KNN and Gaussian Naïve Bayes had lower scores with sentiment data. When sentiment data was not included, the performance of all models decreased, which suggests that sentiment data is useful in predicting market trends. These results suggest that XGBoost with roBERTa sentiment analysis is the best-performing model for the given task, while Logistic Regression with roBERTa sentiment analysis also produced competitive results.

5. DISCUSSION

The results presented in this paper point towards the superior performance of roBERTa-based models compared to their VADAR counterparts. This finding is consistent across all machine learning models tested, as reflected by the accuracy, F1 score, precision, and recall.

Comparing F1 scores between the sentiment models, it is apparent that models using roBERTa consistently outperform those using VADAR. For instance, roBERTa-based models with SVC, Logistic Regression, Gaussian Naïve Bayes, KNN, XGBoost, and Multi Modal Fusion achieved F1 scores of 0.803, 0.807, 0.800, 0.779, 0.822, and 0.850 respectively, whereas their VADAR counterparts achieved lower scores. This demonstrates the higher harmonic mean of precision and recall achieved by the roBERTa models, thus leading to better overall performance.

Furthermore, the models using roBERTa sentiment also consistently displayed higher accuracy compared to VADAR. The most significant increase in accuracy can be observed in the XGBoost model, with an increase from 0.613 to 0.872. These results indicate the capability of roBERTa sentiment models to correctly classify a larger portion of the total predictions, as compared to those using VADAR.

From a precision perspective, the roBERTa sentiment models surpassed the VADAR models in every machine learning algorithm tested. Precision measures the proportion of true positives out of the total predicted positives, and a higher precision score indicates a lower false-positive rate. The Multi Modal Fusion roBERTa model displayed the highest precision score at 0.85, while the same model using VADAR scored only 0.795.

Recall, the metric that measures the proportion of actual positives that are correctly identified, also showcases a similar trend. roBERTa models consistently achieved higher recall values. For instance, in Logistic Regression and Gaussian Naïve Bayes models, roBERTa's recall scores were 0.812 and 0.809, respectively, compared to VADAR's 0.720 and 0.717.

Interestingly, while roBERTa sentiment models performed better on every count, the inclusion or exclusion of sentiment in the models had a less impact. Differences between results 'With Sentiment' and 'Without Sentiment' were relatively small across all models and metrics, with the notable exception of Multi Modal Fusion, which experienced significant score drops without sentiment. This may indicate that sentiment analysis plays a less critical role than initially anticipated in the models' success, warranting further investigation.

While the results demonstrate the superiority of roBERTa over VADAR in terms of accuracy, precision, recall, and F1 score, it is important to consider the inherent limitations of these metrics. Accuracy measures the proportion of correct predictions (both positive and negative) made by the model. However, it does not consider the distribution of those correct predictions, potentially leading to a misleading performance assessment if the results are skewed towards one class. Given our dataset, while accuracy scores were relatively high for all models, it does not provide insight into how the models performed on individual classes of sentiment (positive, negative, neutral).

Precision assesses the model's ability to avoid false positives. While the precision scores were higher for roBERTa models, this metric does not account for false negatives, i.e., the instances where the model incorrectly classifies a true positive case as negative. Thus, while the high precision suggests that the models are reliable when they predict a particular sentiment, it does not reveal whether all instances of that sentiment have been accurately identified.

Recall, on the other hand, takes into account false negatives but ignores false positives. While the recall rates were relatively high, indicating a lower rate of false negatives, they do not provide a complete picture of the model's performance. For instance, in our XGBoost model with roBERTa, a recall of 0.827 suggests the model is capturing a significant portion of the positive instances. However, this does not reflect how many negative instances may have been incorrectly classified as positive.

The F1 score tries to address the limitations of precision and recall by considering both in its calculation, providing a balanced measure of the model's performance. However, it assumes equal importance of precision and recall, which may not always be the case. Depending on the application, one might be more important than the other. In our **case**, we might be more interested in high recall if we want to capture as many instances of a particular sentiment as possible, even if it means incorrectly classifying some negative instances.

In conclusion, the study's findings strongly suggest that roBERTa sentiment models consistently outperform VADAR sentiment models across a range of machine learning models and performance metrics. This reinforces roBERTa's reputation as a robust language model that is capable of superior performance in sentiment analysis tasks. Nonetheless, further studies should be undertaken to validate these results and to better understand the role of sentiment in these models.

6. CONCLUSION

In this paper, we focussed on examining the effect of Twitter sentiment data on the performance of machine learning models for predicting market trend, using historical Bitcoin tweets and market data. The research followed a clearly defined pipeline and produced compelling results that demonstrated the efficacy of incorporating Twitter sentiment data into machine learning models.

Most notably, the study found that using a Multi Modal Fusion with Twitter-roBERTa outperformed all other models. The superiority of this model can be attributed to the sophisticated linguistic analysis of Twitter data conducted by the Twitter-RoBERTa algorithm, which makes it particularly suited for sentiment analysis on social media data, as well as using an LSTM to capture temporal changes.

However, it is crucial to acknowledge the study's limitations and the need for further research to validate these findings. The dataset used only covers tweets, market and on-chain data from 2014 to 2022 and generalised these results to other datasets or contexts may not be appropriate. Further developments (1) include further market features, (2) incorporate other social media data, (3) evaluate longer time frame, and (4) Exploring optimal lag.

In conclusion, this research provides valuable insights into the use of Twitter sentiment data to improve machine learning models' performance in predicting market trends, with a Multi Modal Fusion model emerging as the most effective. These findings are significant, given the increasing importance of social media data in financial markets and the potential benefits of integrating this data into predictive models.

REFERENCES

[1] Abraham, J., Higdon, D., Nelson, J., Ibarra, J., Nelson, J. (2018) Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. SMU Data Science Review, 1(3).[online] Available: https://scholar.smu.edu/cgi/viewcontent.cgi?article=1039&context=datasciencereview.

[2] Barretto, M.C.(2017) Sentiment Analysis Tools and Techniques: A Comprehensive Survey. International Journal for Research in Applied Science and Engineering Technology, V(XI), pp.3030–3035. [online] Available: doi:https://doi.org/10.22214/ijraset.2017.11419.

[3] Basilio, J., Toriola, A.(2021) Prediction of Bitcoin Prices Using Deep learning and Sentiment Analysis Based on Bitcoin Tweets, MSc Research Project, School of Computing National College of Ireland. [online] Available: https://norma.ncirl.ie/5230/1/adebayojosephtoriola.pdf.

[4] Bollen, J., Mao, H. and Zeng, X. (2011) Twitter mood predicts the stock market. Journal of Computational Science, 2(1), pp.1-8.

[5] Devlin, J., Chang, M. W., Lee, K., Google, K., Language, A.(2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, pp.4171–4186. [online] Available: https://aclanthology.org/N19-1423.pdf.

[6] Gao, J., Li, P., Chen, Z. and Zhang, J. (2020). A Survey on Deep Learning for Multimodal Data Fusion. Neural Computation, 32(5), pp.829–864. doi:https://doi.org/10.1162/neco_a_01273.

[7] He, S. and Gu, S. (2021). Multi-modal Attention Network for Stock Movements Prediction. [online] arXiv.org. Available at: https://arxiv.org/abs/2112.13593.

[8] Hutto, C., Gilbert, E.(2014) VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), pp.216–225. [online] Available: doi:https://doi.org/10.1609/icwsm.v8i1.14550.

[9] Jaquart, P., Dann, D., Weinhardt, C. (2021)Short-term bitcoin market prediction via machine learning. The Journal of Finance and Data Science, 7, pp.45–66. [online] Available: doi:https://doi.org/10.1016/j.jfds.2021.03.001.

[10] Jiang, W. (2021). Applications of deep learning in stock market prediction: Recent progress. Expert Systems with Applications, 184, p.115537. doi:https://doi.org/10.1016/j.eswa.2021.115537.

[11] Joshi, K., Rao, J.(2016) Stock Trend Prediction Using News Sentiment Analysis. International Journal of Computer Science and Information Technology, 8(3), pp.67–76. [online] Available: doi:https://doi.org/10.5121/ijcsit.2016.8306.

[12] Kim, Y.B., Kim, J.G., Kim, W., Im, J., Kim, T.H., Kang, S.J. and Kim, C.H. (2016) Predicting fluctuations in cryptocurrency transactions based on user comments and replies. PloS one, 11(8), p.161197.

[13] Kristoufek, L. (2013)BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. Scientific reports, 3, p.3415.

[14] Li, Q., Tan, J., Wang, J. and Chen, H. (n.d.). A Multimodal Event-driven LSTM Model for Stock Prediction Using Online News A Multimodal Event-driven LSTM Model for Stock Prediction Using Online News. [online] Available at: https://ailab-ua.github.io/courses/resources/Qing_TKDE_2020.pdf.

[15] Li, Y. and Pan, Y. (2021). A novel ensemble deep learning model for stock prediction based on stock prices and news. International Journal of Data Science and Analytics. doi:https://doi.org/10.1007/s41060-021-00279-9.

[16] Loria, S.(2018)TextBlob: Simplified Text Processing. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 54-59.

[17] Li, X., Wang, C.A., Dong, H. and Wang, M. (2018) A sentiment-analysis-based approach for predicting the cryptocurrency market. 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 830-834.

[18] Tandon, C., Revankar, S., Palivela, H., Parihar, S.S.(2021) How can we predict the impact of the social media messages on the value of cryptocurrency? Insights from big data analytics. International Journal of Information Management Data Insights, 1(2), p.100035. [online] Available: doi:https://doi.org/10.1016/j.jjimei.2021.100035.

[19] Wasiat, K., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., Alfakeeh, A. S.: Stock market prediction using machine learning classifiers and social media, news, Journal of Ambient Intelligence and Humanized Computing, 1-24, (2020).

[20] Zhang, Y., Welker, C. and Eksioglu, B., 2017. A study of the impact of social media sentiment on cryptocurrency market value. 2017 IEEE International Conference on Big Data (Big Data), pp.2030-2037.
[21] Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2021). Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter. IEEE Transactions on Computational Social Systems. doi:10.1109/tcss.2020.3042446
[22] Lu, W., Li, J., Li, Y., Sun, A. and Wang, J. (2020). A CNN-LSTM-Based Model to Forecast Stock Prices. [online] Complexity. Available at: https://www.hindawi.com/journals/complexity/2020/6622927/.
[23] Yan, Pang., Ganeshkumar, Sundararaj., Jiewen, Ren. (2020). Cryptocurrency Price Prediction using Time Series and Social Sentiment Data. Elements,
[24] Critien, J.V., Gatt, A. and Ellul, J. (2022). Bitcoin price change and trend prediction through twitter sentiment and data volume. Financial Innovation, 8(1). doi:https://doi.org/10.1186/s40854-022-00352-7.

AUTHORS

Saachin Bhatt is a Data Scientist and a final year BSc (Hons) Digital and Technology Solutions Degree Apprentice, sponsored by Accenture. With over two and a half years of experience in the field, he has a proven track record of designing and implementing effective analytics solutions for a range of large FTSE 100 and Global Fortune 500 organisations in the Financial Services, Aerospace and Defence, Insurance, and Chemicals industries. As a subject matter expert, Saachin has published articles in well-renowned industry journals on InsurTech and financial markets. He is also a recipient of several academic achievement awards for his expertise in mathematics, information systems design, and web development.

Mustansar Ali Ghazanfar (MAG) is a renowned computer scientist and AI expert, holding a PhD from the University of Southampton and ongoing MBA from Heriot-Watt University. He is a senior lecturer and AI course leader at University of East London. He has a strong record of accomplishment of managing innovative projects, as evidenced by his successful completion as project director/investigator of the $3.8M "JACC-upgrade" project focussing on data analytics, which contributed to Pakistan's socio-economic development. His research excellence is demonstrated by over 60 published articles and an impressive citation rate of 1.1 per day according to Google Scholar. He has also gained valuable experience as a machine learning developer at Brandwatch UK and held visiting fellow positions at organizations such as IOTA and ELM London.

Dr Mohammad Hossein Amirhosseini is a lecturer in Computer Science and Digital Technologies at University of East London. He is the course leader of BSc and MSc Digital and Technology Solutions Apprenticeship courses. He has contributed to national and international granted research projects, and has published papers in different international refereed journals and conferences. He has also been the main organiser and chair of different special sessions in renowned international conferences, guest editor of special issues in international refereed journals, and the reviewer for different journals, conferences.