**MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR**
(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)
***NAAC Accredited with A++ Grade***


**Project Report**

**on**

# Prediction of Bitcoin Prices with Sentiment Analysis

**Submitted By:**

**Surabhi Verma**

**0901AM211059**

**Faculty Mentor:**

**Dr. Neelam Arya,**

**Assistant Professor**

**CENTRE FOR ARTIFICIAL INTELLIGENCE**
MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

JULY-DEC. 2023

**MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR**

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

*NAAC Accredited with A++ Grade*

# CERTIFICATE

This is certified that **Surabhi Verma** (0901AM211059) has submitted the project report titled **Prediction of Bitcoin Prices with Sentiment Analysis** under the mentorship of **Dr. Neelam Arya**, in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in **Artificial Intelligence and Machine Learning** from Madhav Institute of Technology and Science, Gwalior.

Neelam 24/11/23

Dr. Neelam Arya
Faculty Mentor
Assistant Professor
Centre for Artificial Intelligence

24/11/2023

**Dr. R. R. Singh**
Coordinator
Centre for Artificial Intelligence

**MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR**

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

***NAAC Accredited with A++ Grade***

# DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in **Artificial Intelligence and Machine Learning** at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of Dr. Neelam Arya, Assistant Professor, Centre for Artificial Intelligence

I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Surabhi Verma
0901AM211059
III$^{rd}$ Year,
Centre for Artificial Intelligence

**MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE GWALIOR**

(A Govt. Aided UGC Autonomous Institute Affiliated to RGPV, Bhopal)

***NAAC Accredited with A++ Grade***

# ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Centre for Artificial Intelligence,** for allowing me to explore this project. I humbly thank **Dr. R. R. Singh**, Coordinator, Centre for Artificial Intelligence, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr. Neelam Arya**, Assistant Professor, Centre for Artificial Intelligence, for her continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Surabhi Verma
0901AM211059
III$^{rd}$ Year,
Centre for Artificial Intelligence

# ABSTRACT

This report explores the integration of sentiment analysis and traditional financial indicators to predict Bitcoin prices. Utilizing a diverse dataset encompassing historical prices and social media data, the study employs Natural Language Processing to extract sentiment features. These features are combined with conventional financial metrics, and machine learning algorithms are applied to develop a robust predictive model. Through rigorous back testing, the research evaluates the model's performance. Additionally, temporal analysis of sentiment patterns aims to unveil nuanced relationships between changing market sentiments and Bitcoin price movements. The findings provide valuable insights for investors and contribute to the understanding of sentiment's impact on cryptocurrency markets.

**Keyword:** Bitcoin prices, Sentiment analysis, Cryptocurrency markets, Predictive modelling, Natural Language Processing (NLP), RandomForest, XGBoost Classifier, Financial indicators, Wikipedia data, Back testing, Market sentiment, Price movements, Digital assets, Rolling averages

## सार:

यह रिपोर्ट बिटकॉइन की कीमतें पूर्वानुमानित करने के लिए भावना विश्लेषण और पारंपरिक वित्तीय संकेतकों के समर्थन में अनुसंधान करती है। ऐतिहासिक मूल्यों और सोशल मीडिया डेटा को समाहित करने के लिए एक विविध डेटासेट का उपयोग करते हुए, अध्ययन में प्राकृतिक भाषा प्रसंस्करण का उपयोग संवेदना सुविधाओं को निकालने के लिए करता है। ये सुविधाएं पारंपरिक वित्तीय माप के साथ मिश्रित की जाती हैं, और मजबूत पूर्वानुमान मॉडल विकसित करने के लिए मशीन लर्निंग एल्गोरिदम का उपयोग होता है। कठिन पुनरायोजन के माध्यम से, अनुसंधान ने मॉडल के प्रदर्शन का मूल्यांकन किया है। इसके अलावा, भावना पैटर्न का कालीन विश्लेषण ने बदलते बाजार भावनाओं और बिटकॉइन कीमतों के बीच सूक्ष्म संबंधों को खोलने का प्रयास किया है। इस अनुसंधान के परिणाम निवेशकों के लिए मूल्यशील सुजान प्रदान करते हैं और क्रिप्टोकरेंसी बाजारों पर भावना के प्रभाव को समझने में योगदान करते हैं।

# LIST OF FIGURES

# TABLE OF CONTENTS

# Chapter 1: INTRODUCTION

## 1.1 Background

Cryptocurrencies, decentralized digital currencies that utilize cryptographic techniques for secure financial transactions, have witnessed a remarkable rise in popularity and recognition as a unique asset class. Among the myriad of cryptocurrencies, Bitcoin has emerged as a frontrunner, capturing the attention of investors, businesses, and the general public alike. Bitcoin's decentralized nature, borderless transactions, and potential for substantial returns have contributed to its status as a groundbreaking financial instrument.

However, the unparalleled growth and adoption of Bitcoin have been accompanied by a high level of price volatility. Traditional financial modelling, which relies on historical data, fundamental analysis, and technical indicators, encounters difficulties in accurately forecasting the price movements of this decentralized digital currency. The dynamic and relatively nascent nature of the cryptocurrency market, coupled with external factors such as regulatory developments and market sentiment, has prompted researchers to seek innovative approaches to address the challenges associated with predicting Bitcoin prices.

As a result, there has been a growing interest in exploring alternative methods that go beyond conventional financial models. Sentiment analysis, a field within natural language processing (NLP), has emerged as a promising avenue. By analysing and interpreting the sentiments expressed in textual data from various sources, such as social media, news articles, and online forums, researchers aim to capture the collective mood and opinions of market participants, potentially providing valuable insights into the factors influencing Bitcoin price dynamics.

## 1.2 Objective

This report aims to:

a. **Investigate the impact of sentiment on Bitcoin prices.**

The primary objective is to explore and understand the influence of sentiment on the fluctuations of Bitcoin prices. This involves analysing sentiment data extracted from various sources, such as social media, news articles, and forums, to discern patterns and correlations with the observed price movements. By delving into the emotional aspects of market participants, this research aims to uncover how sentiment contributes to the dynamics of Bitcoin prices.

**b. Develop a predictive model that integrates sentiment analysis.**

This objective focuses on the development of a comprehensive predictive model that incorporates sentiment analysis as a crucial component. The model will utilize historical Bitcoin price data and sentiment features derived from textual data to make predictions about future price movements. The integration of sentiment analysis aims to capture the nuanced information present in qualitative data, providing a more holistic understanding of the factors driving Bitcoin prices.

The model may leverage machine learning algorithms, such as regression, neural networks, or ensemble methods, to effectively combine quantitative financial indicators with sentiment-derived features. Natural Language Processing (NLP) techniques will be employed to preprocess and extract sentiment-related information from textual data.

**c. Evaluate the effectiveness of the proposed model in comparison to traditional models.**

To assess the contribution of sentiment analysis to Bitcoin price prediction, this objective involves a comparative analysis. The developed predictive model will be benchmarked against traditional financial models, including time series analysis and machine learning models that rely solely on quantitative indicators. Evaluation metrics such as accuracy, precision, recall, and F1 score will be employed to quantify the performance of the proposed model against its traditional counterparts.

# Chapter 2: LITERATURE REVIEW

## 2.1 Cryptocurrency Price Prediction Models

As the cryptocurrency market continues to evolve, researchers have explored various models to predict price movements, aiming to provide valuable insights for investors and market participants. This literature review section presents an overview of key studies and approaches in the field of cryptocurrency price prediction.

### 2.1.1 Time Series Analysis

Time series analysis is a fundamental method for predicting cryptocurrency prices. [1] applied ARIMA models to Bitcoin prices, demonstrating the efficacy of time series methods in capturing short-term trends and cyclical patterns.

### 2.1.2 Machine Learning Models

Machine learning techniques have gained popularity for their ability to capture complex relationships in cryptocurrency data. [2] employed support vector machines and decision trees to predict Bitcoin prices, showcasing the potential of ML models in handling non-linear patterns. Deep learning models have also been explored. [3] implemented LSTM networks to predict Bitcoin prices, highlighting the effectiveness of neural networks in capturing temporal dependencies.

### 2.1.3 Sentiment Analysis Models

Sentiment analysis has emerged as a crucial factor in understanding cryptocurrency price movements. [4] conducted a pioneering study that correlated Twitter sentiment with Bitcoin prices, emphasizing the impact of social media sentiment on market dynamics. In a more recent study, [5] integrated sentiment analysis from financial news articles to predict cryptocurrency prices, underscoring the importance of incorporating qualitative information into predictive models.

### 2.1.4 Network Analysis Models

Network analysis provides insights into the structure of the cryptocurrency ecosystem. [6] utilized network analysis to study the Bitcoin transaction graph, revealing patterns that could be indicative of market behaviour.

### 2.1.5 Blockchain-Based Models

On-chain data analysis is essential for understanding cryptocurrency fundamentals. [7] explored the relationship between on-chain metrics and Bitcoin prices, highlighting the potential of blockchain-based models for predicting market trends.

## 2.2 Sentiment Analysis in Financial Markets

[4] investigated the correlation between social media sentiment and Bitcoin prices. The authors employed sentiment analysis techniques on Twitter data to extract relevant signals and examined their impact

on algorithmic trading strategies. [8] explored the relationship between sentiment and cryptocurrency prices, focusing on Bitcoin and Ethereum. The study employed sentiment analysis on social media data and news articles to gauge the overall market sentiment. The research highlighted the potential of sentiment analysis as a predictive tool and its role in capturing market dynamics. [5] delved into the impact of news sentiment on cryptocurrency prices, specifically Bitcoin. Abad et al. conducted sentiment analysis on financial news articles to extract sentiment-related features. The study proposed a predictive model that integrated sentiment analysis, demonstrating its efficacy in improving the accuracy of cryptocurrency price predictions. [9] conducted an empirical study on the relationship between sentiment and cryptocurrency prices, with a focus on Bitcoin and Ethereum. The research involved sentiment analysis on social media and news data, and the findings highlighted the potential for sentiment analysis to enhance the accuracy of cryptocurrency price predictions. [10] explored the short-term return predictability of Bitcoin by analysing daily sentiment data. The study utilized a sentiment lexicon approach to extract sentiment features from social media. The findings indicated a significant relationship between daily sentiment and short-term Bitcoin returns, emphasizing the role of sentiment in intraday price movements.

# Chapter 3: Methodology

## 3.1 Data Collection

### 3.1.1 Sentiment data from Wikipedia

By analysing edits made on the Wikipedia page of Bitcoin can indeed provide insights into the level of interest and sentiment of the public towards Bitcoin. Wikipedia edits often reflect current events, developments, and changes in perception or understanding of a topic.

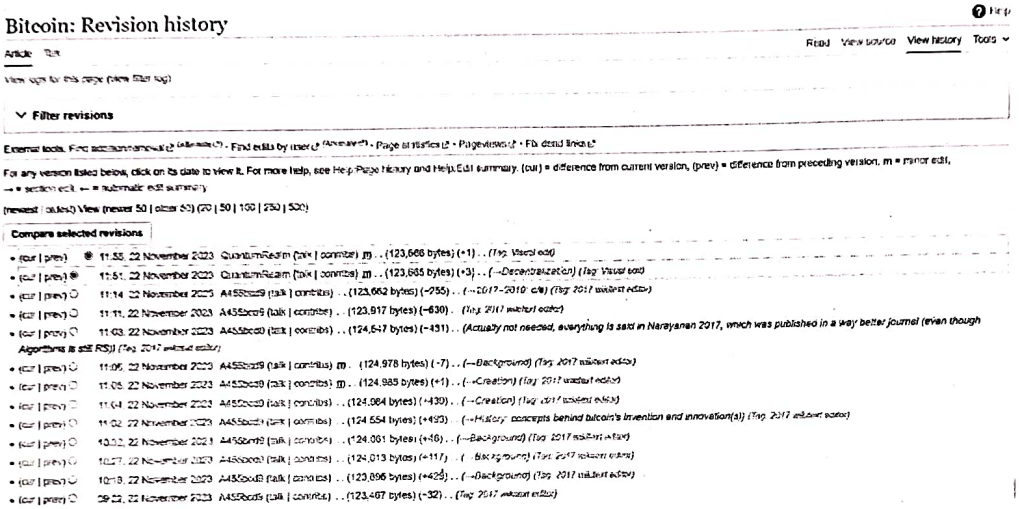


Fig 3.1. Wikipedia page of Bitcoin Revision History

These edits are made by the general public which gives us the sentiment that the public holds towards Bitcoin. These edits are downloaded in the project via mwclient library which enables us to work with specific Wikipedia page.

```
!pip install mwclient  #mediawikiclient
import mwclient,
import time

site=mwclient.Site("en.wikipedia.org")   #class that enables us to work with specific wiki site
page=site.pages["Bitcoin"]
```

Fig 3.2. Fetching the Bitcoin Page

### 3.1.2 Financial Data from Yahoo API

Yahoo Finance API provides financial data, including stock prices, historical data, market summaries, and company information. Developers can access this data through the API, facilitating the integration of real-time and historical financial information into applications, websites, or analysis tools. The API supports a range of endpoints for diverse financial data needs.

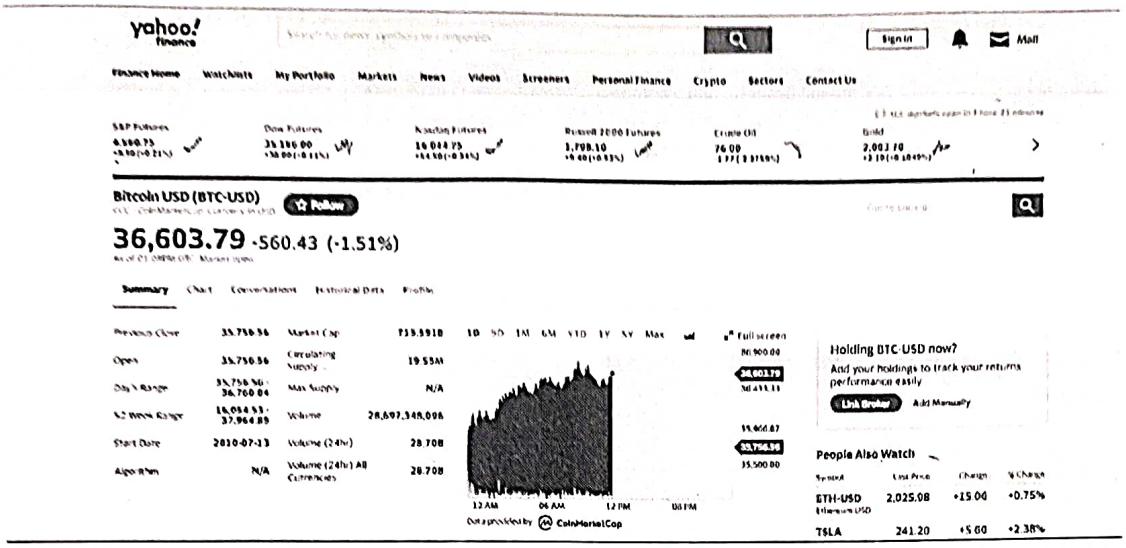


Fig 3.3. Yahoo Finance showing current Bitcoin Price

The price history of bitcoin against Indian Rupee from its starting date to current date (on which the project was completed) is fetched by Yahoo API.

| Date | Open | High | Low | Close | Volume | Dividends | Stock Splits |
|---|---|---|---|---|---|---|---|
| 2014-09-17 00:00:00+00:00 | 2.844333e+04 | 2.854223e+04 | 2.755250e+04 | 2.785164e+04 | 1282359120 | 0.0 | 0.0 |
| 2014-09-18 00:00:00+00:00 | 2.782277e+04 | 2.782277e+04 | 2.508574e+04 | 2.577412e+04 | 2093892320 | 0.0 | 0.0 |
| 2014-09-19 00:00:00+00:00 | 2.575365e+04 | 2.598884e+04 | 2.336609e+04 | 2.402334e+04 | 2307413745 | 0.0 | 0.0 |
| 2014-09-20 00:00:00+00:00 | 2.401585e+04 | 2.575756e+04 | 2.372438e+04 | 2.488181e+04 | 2243150060 | 0.0 | 0.0 |
| 2014-09-21 00:00:00+00:00 | 2.483197e+04 | 2.509612e+04 | 2.392506e+04 | 2.426826e+04 | 1617399085 | 0.0 | 0.0 |
| ... | ... | ... | ... | ... | ... | ... | ... |
| 2023-11-18 00:00:00+00:00 | 3.006040e+06 | 3.057181e+06 | 2.988633e+06 | 3.048202e+06 | 1869485764064 | 0.0 | 0.0 |
| 2023-11-19 00:00:00+00:00 | 3.050592e+06 | 3.068408e+06 | 3.017936e+06 | 3.047287e+06 | 990007668332 | 0.0 | 0.0 |
| 2023-11-20 00:00:00+00:00 | 3.047292e+06 | 3.124238e+06 | 3.034126e+06 | 3.114010e+06 | 1075801516348 | 0.0 | 0.0 |
| 2023-11-21 00:00:00+00:00 | 3.112970e+06 | 3.147153e+06 | 3.074437e+06 | 3.122672e+06 | 1740456922727 | 0.0 | 0.0 |
| 2023-11-22 00:00:00+00:00 | 2.980164e+06 | 3.062876e+06 | 2.980164e+06 | 3.052020e+06 | 2380659425280 | 0.0 | 0.0 |

3354 rows × 7 columns

Fig 3.4. Fetched Price data from Yahoo API

## 3.2 Sentiment Analysis

### 3.2.1 Downloading pre-trained Sentiment Analysis model

The pipeline module from the transformers library simplifies the usage of pre-trained models for various natural language processing tasks. It allows you to perform tasks like text generation, sentiment analysis, and question answering with minimal code. For example, pipeline('sentiment-analysis') initializes a sentiment analysis pipeline that you can use to analyse the sentiment of a given text. This library is built on top of Hugging Face's Transformer models.

```
!pip install transformers  #pre-trained deep learning models
from transformers import pipeline
sentiment_pipeline=pipeline("sentiment-analysis")  #this downloads the sentiment analysis model
```

Fig 3.5. Pre-Trained Sentiment Analysis Model using transformers library

### 3.2.2 Analysing sentiment of Revision edits

Each comment edit that was made on the Wikipedia page is evaluated whether it has positive, neutral or negative sentiment. A score is assigned to the sentiment that is calculated by the model. Positive having score equal to 1 and negative as -1.

```
edits={}

for rev in revs:
  date=time.strftime("%Y-%m-%d",rev ["timestamp"])  #convert timestamp datatype (named tuple) into string

  if date not in edits:
    edits[date] = dict(sentiments=list(), edit_count=0)

  edits[date]["edit_count"] +=1

  comment = rev.get("comment", "")
  edits[date]["sentiments"].append(find_sentiment(comment))
#edits=dictionary with date as key and no. of times the page was edited along with the sentiment of the comment on that particular date
```

Fig 3.6. This code calculates all edits in a single day and find average sentiment of each day

All the missing dates are filled using pandas.date_range and in reindexed so that no missing values is there. Rolling Average is used to determine the average sentiment of last 30 days for each day. This 'edits' is then converted into pandas Dataframe and merged with market price data to be used for prediction model

| | edit_count | sentiment | neg_sentiment |
|---|---|---|---|
| 2009-04-06 | 0.133333 | 0.005820 | 0.025000 |
| 2009-04-07 | 0.000000 | 0.000000 | 0.000000 |
| 2009-04-08 | 0.000000 | 0.000000 | 0.000000 |
| 2009-04-09 | 0.000000 | 0.000000 | 0.000000 |
| 2009-04-10 | 0.000000 | 0.000000 | 0.000000 |
| ... | ... | ... | ... |
| 2023-10-29 | 0.766667 | 0.109872 | 0.141026 |
| 2023-10-30 | 0.766667 | 0.109872 | 0.141026 |
| 2023-10-31 | 0.766667 | 0.109872 | 0.141026 |
| 2023-11-01 | 0.766667 | 0.109872 | 0.141026 |
| 2023-11-02 | 0.766667 | 0.109872 | 0.141026 |

5324 rows × 3 columns

Fig 7. Sentiment Data

## 3.3 Predictive Model

### 3.3.1 Exploratory Data analysis

Features like 'stock splits' and 'dividends' were removed as they were not useful to the model. The date time index for price data was converted from aware to naïve to match the sentiment date time index. All the column names are converted into lowercase for the ease of programming.

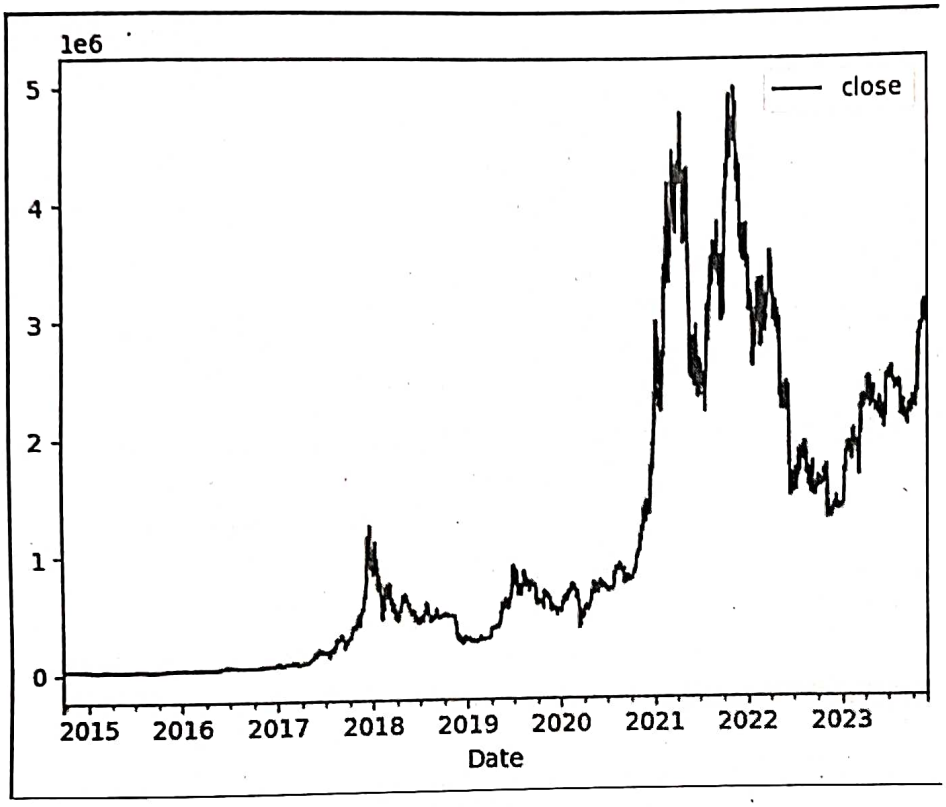


Fig 3.8. Plot between closing price and date

Fig 8. shows that price of bitcoin has spiked suddenly in the year 2021 and 2022.

### 3.3.2 Combining the sentiment data with Price data

The previously made sentiment data is imported and combined with the price data and only those rows are considered which have same dates. We then observe whether the price goes up or down the next day by comparing the closing price of next day and closing price of current day. The target variable would be 1 if the prices go up and 0 if it goes down. We observed that there are as much increase in price as there are decrease in price.

```
btc["target"].value_counts()

1    1824
0    1510
Name: target, dtype: int64
```

Fig 3.9. Target variable signifies increase or decrease in price per day

| | open | high | low | close | volume | edit_count | sentiment | neg_sentiment | tomorrow | target |
|---|---|---|---|---|---|---|---|---|---|---|
| 2014-09-17 | 2.844333e+04 | 2.854223e+04 | 2.785250e+04 | 2.788104e+04 | 1282359120 | 5.033333 | 0.244480 | 0.532716 | 24803.121094 | 0 |
| 2014-09-18 | 2.782277e+04 | 2.782277e+04 | 2.508574e+04 | 2.577412e+04 | 2093992320 | 5.066667 | 0.245040 | 0.532716 | 26600.681641 | 1 |
| 2014-09-19 | 2.575365e+04 | 2.598884e+04 | 2.336609e+04 | 2.402334e+04 | 2307413745 | 5.200000 | 0.285074 | 0.540308 | 25705.070000 | 1 |
| 2014-09-20 | 2.401595e+04 | 2.575756e+04 | 2.372438e+04 | 2.488181e+04 | 2243150060 | 5.200000 | 0.259807 | 0.549385 | 25241.833984 | 1 |
| 2014-09-21 | 2.483197e+04 | 2.509612e+04 | 2.392506e+04 | 2.426826e+04 | 1617399085 | 5.233333 | 0.272244 | 0.532716 | 24726.542969 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2023-10-29 | 2.828465e+06 | 2.869478e+06 | 2.825719e+06 | 2.843635e+06 | 847539744277 | 0.766667 | 0.109872 | 0.141026 | NaN | 0 |
| 2023-10-30 | 2.843618e+06 | 2.898163e+06 | 2.831789e+06 | 2.880921e+06 | 930904073848 | 0.766667 | 0.109872 | 0.141026 | NaN | 0 |
| 2023-10-31 | 2.880359e+06 | 2.902052e+06 | 2.840071e+06 | 2.872238e+06 | 1430597552698 | 0.766667 | 0.109872 | 0.141026 | NaN | 0 |
| 2023-11-01 | 2.872047e+06 | 2.892442e+06 | 2.837775e+06 | 2.887054e+06 | 1312313305463 | 0.766667 | 0.109872 | 0.141026 | NaN | 0 |
| 2023-11-02 | 2.886180e+06 | 2.958294e+06 | 2.844646e+06 | 2.950284e+06 | 1868736385252 | 0.766667 | 0.109872 | 0.141026 | NaN | 0 |

3334 rows × 10 columns

Fig 3.10. Combined Data of sentiment analysis and Price data

### 3.3.3 Training Baseline Model

Implement a baseline Random Forest model using the pre-processed data. Random Forest is chosen for its simplicity, interpretability, and ability to handle both numerical and categorical features. Utilize libraries such as scikit-learn for model implementation.

```
#random forest- avoids overfitting and fast
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=100, min_samples_split=50, random_state=1)
# n_estimators=100 individual decision tree
#min_sample_split- each descion tree should not split unless it has 50 samples, reduces overfitting

#splitting data for training and testing, last 200 days are testing set, rest are training
train = btc.iloc[:-200]
test = btc.iloc[-200:]

#no cross validation technique is used as the data is time series data, order is important
predictors = ["close", "volume", "open", "high", "low", "edit_count", "sentiment", "neg_sentiment"]
model.fit(train[predictors], train["target"])
```

```
RandomForestClassifier
RandomForestClassifier(min_samples_split=50, random_state=1)
```

Fig 3.11. RandomForest Classifier

### 3.3.4 Using XgBoost Classifier along with Back testing

An XGBoost classifier is employed for predicting Bitcoin prices by integrating sentiment analysis. Historical data, encompassing cryptocurrency prices and sentiment features, is used for training and

testing the model. The trading strategy is back tested, simulating its performance using historical data. The XGBoost classifier is assessed based on accuracy metrics, and the back testing results, such as cumulative returns and other performance indicators, guide the refinement of the model and strategy. The combination of XGBoost classification and back testing provides a comprehensive framework for predicting Bitcoin prices and evaluating the effectiveness of sentiment-based trading strategies.

```
#this function would combine all the predicted values with actual values
def predict(train, test, predictors, model):
    model.fit(train[predictors], train["target"])
    preds = model.predict(test[predictors])
    preds = pd.Series(preds, index=test.index, name="predictions")
    combined = pd.concat([test["target"], preds], axis=1)
    return combined

#start defines how much historical data to be used (1095=3yrs)
#step defines period the amount we want to make predictions for after the start (150=6 months)
def backtest(data, model, predictors, start=1095, step=150):
    all_predictions = []

    for i in range(start, data.shape[0], step):
        train = data.iloc[0:i].copy()
        test = data.iloc[i:(i+step)].copy()
        predictions = predict(train, test, predictors, model)
        all_predictions.append(predictions)

    return pd.concat(all_predictions)
```

Fig 3.12. Predict function combines the predicted values with actual values

Back test function keeps making new predictions after 1095 days for every next 150 days

```
#learning rate=lower the number less overfitting
from xgboost import XGBClassifier

model = XGBClassifier(random_state=1, learning_rate=.1, n_estimators=200)
predictions = backtest(btc, model, predictors)
```

Fig 3.13. XgBoost Classifier

### 3.3.5 Improving the Model using different time horizons (Trends)

Trends in the data are computed using a function compute_rolling to compute rolling averages and ratio for various horizons and add them as new columns to the input DataFrame (btc). It also returns a modified DataFrame along with a list of new predictor column names.

```
#find trends in various column in the last period
def compute_rolling(btc):
    horizons = [2,7,60,365]
    new_predictors = ["close", "sentiment", "neg_sentiment"]

    for horizon in horizons:
        #compute rolling averages
        rolling_averages = btc.rolling(horizon, min_periods=1).mean()

        #Compute close ratio
        ratio_column = f"close_ratio_{horizon}"
        btc[ratio_column] = btc["close"] / rolling_averages["close"]

        # Compute sentiment-related features
        edit_column = f"edit_{horizon}"
        btc[edit_column] = rolling_averages["edit_count"]

        # Compute rolling averages for the target variable
        rolling = btc.rolling(horizon, closed='left', min_periods=1).mean()
        trend_column = f"trend_{horizon}"
        btc[trend_column] = rolling["target"]

        # Update the list of new predictor columns
        new_predictors+= [ratio_column, trend_column, edit_column]
    return btc, new_predictors
```

Fig 3.14. Computing rolling averages to find trends in data

# Chapter 4: Evaluation of the Predictive Model

## 4.1 Evaluation of RandomForest Model

The RandomForest model achieved a precision of 55%, indicating its ability to correctly identify relevant instances among the total predicted positives. Precision, a crucial metric in classification evaluation, signifies the proportion of true positives among all instances predicted as positive.

```
#evaluating the predictions
from sklearn.metrics import precision_score

preds = model.predict(test[predictors])
preds = pd.Series(preds, index=test.index)
precision_score(test["target"], preds)
```

```
0.5504587155963303
```

Fig 4.1. Precision score of RandomForest model.

Since the precision of this model is not very high, therefore refinement of this model is done by using XGBoost Classifier

## 4.2 Evaluation of XGBoost Classifier Model

The XGBoost model demonstrated a precision of 54%, indicating its ability to accurately identify relevant instances among the total predicted positives. Precision, a key classification metric, highlights the model's capability to minimize false positives. We did not gain much different result from Xgboost even after using back testing along with it.

```
predictions["predictions"].value_counts()
```

```
1    1178
0    1061
Name: predictions, dtype: int64
```

```
precision_score(predictions["target"], predictions["predictions"])
```

```
0.5449915110356537
```

Fig 4.2. Precision score of XGBoost Model

## 4.3 Evaluation of Model considering the trends

The model with rigorous back testing and rolling averages along different time horizons (2, 7, 60, 365 days) yielded a precision score of 71 % indicating the highest precision recorded among other models. This would lead to highly accurate predictions which could be further used for real-life prediction upon fine tuning.

```
predictions = backtest(btc, model, new_predictors)
```

```
precision_score(predictions["target"], predictions["predictions"])
```

```
0.7141744548286605
```

Fig 4.3. Precision score of model after back testing the trends

# Chapter 5: Conclusion

## 5.1 Future Scope

The research on predicting Bitcoin prices with sentiment analysis opens avenues for future exploration and advancements in understanding the intricate dynamics of cryptocurrency markets. Several promising directions emerge, shaping the future scope of this research:

### 5.1.1 Enhanced Sentiment Analysis Techniques

Further refinement of sentiment analysis techniques, including advanced Natural Language Processing (NLP) approaches, sentiment context analysis, and sentiment disambiguation, holds potential for capturing more nuanced and accurate sentiment signals. Advancements in these techniques will contribute to the precision of predictive models, enhancing their effectiveness in forecasting Bitcoin prices.

### 5.1.2 Integration of External Data Sources

Expanding the scope of data integration to include macroeconomic indicators, regulatory developments, and blockchain transaction data can provide a more holistic view of the cryptocurrency ecosystem. Future research should explore comprehensive frameworks for seamlessly incorporating these external factors into sentiment analysis models, aiming for a more comprehensive understanding of market influences.

### 5.1.3 Real-Time Sentiment Analysis

The cryptocurrency market's rapid pace requires a shift toward real-time sentiment analysis. Developing methodologies for capturing and analysing sentiment in near real-time will empower market participants to make agile and well-informed decisions amidst swiftly changing market conditions. Investigating technologies and strategies for dynamic sentiment analysis will be crucial for staying ahead in this dynamic environment.

### 5.1.4 Explainable AI in Cryptocurrency Predictions

The demand for transparent and interpretable AI models is growing. Future research should prioritize the development of explainable AI models specific to cryptocurrency predictions. Models that provide clear insights into the features influencing predictions will enhance trust among users, enabling better-informed decision-making in the cryptocurrency market.

## 5.2 Limitations of this study

While this study provides valuable insights, certain limitations should be acknowledged:

### 5.2.1 Data Limitations

The study relies on the availability and quality of historical Bitcoin prices and sentiment data. Data limitations, including gaps or inaccuracies, may impact the model's performance and generalizability.

### 5.2.2 Model Generalization

The predictive models developed in this research may have limitations in generalizing to different market conditions or time periods. Future studies should explore model robustness across diverse scenarios.

### 5.2.3 External Factors

The research considers sentiment and historical data, but external factors such as regulatory changes, macroeconomic shifts, or unforeseen events were not extensively incorporated. Further research could explore ways to integrate a broader range of external factors.

### 5.2.4 Dynamic Nature of Cryptocurrency Markets

Cryptocurrency markets are highly dynamic, subject to rapid changes and market sentiment shifts. The study's static analysis may not fully capture the dynamic nature of these markets.

### 5.2.5 Model Interpretability

While predictive models show promise, their interpretability remains a challenge. Enhancements in model interpretability are essential for fostering user trust and understanding.

# References

[1] Tsantekidis et al., "Forecasting Cryptocurrency Prices with Deep Learning under Data scarcity," 2018

[2] Singh and Srivastava, "Bitcoin Price Prediction Using Machine Learning Algorithms: An Approach to Find the Optimal Model," 2018

[3] Zhang et al., "Time series prediction using a deep learning model with LSTM network architecture," 2019

[4] Garcia and Schweitzer, "Social signals and algorithmic trading of Bitcoin," 2015

[5] Abad et al., "Predicting Cryptocurrency Prices Using News Sentiment Analysis," 2020

[6] Ron and Shamir, "Quantitative Analysis of the Full Bitcoin Transaction Graph," 2013

[7] Hayes and Danezis, "A survey of cryptocurrency crime and blockchain forensics," 2018

[8] Bouras et al., "Cryptocurrency Price Prediction Using Sentiment Analysis," 2019

[9] Mittal and Goel, "Cryptocurrency Price Prediction using Sentiment Analysis: An empirical study," 2021

[10] Li et al., "Daily Bitcoin Sentiment and Short-Term Return Predictability," 2018