Live cryptocurrency price prediction and analysis

Kankaria, Romit
rkankaria@wisc.edu

Shukla, Varan
vshukla4@wisc.edu

Chunduru, Rahul
chunduru2@wisc.edu

Sinha, Sweksha
ssinha42@wisc.edu

ABSTRACT

This research paper explores the application of advanced data processing techniques for the prediction and trend analysis of cryptocurrency prices. The proposed system uses Apache Kafka for live data streaming and Apache Spark for stream processing. The processed data is then fed into the Autoregressive Integrated Moving Average (ARIMA) model for prediction. We have also evaluated the effectiveness of this approach in predicting the future trends of cryptocurrency prices against historical data. The results indicate that our system can accurately predict future trends in cryptocurrency prices and can be used as a tool for investors and traders to make informed decisions about their investments. We have also attempted to explore other factors influencing the crypto market, particularly internet sentiment, by analyzing social media content and news sentiment to gauge public opinion on various cryptocurrencies. The paper also investigates the correlation of coin prices with larger coins such as Bitcoin and Ethereum. Overall, this research highlights the potential of combining cutting-edge data processing technologies with machine learning techniques to predict the future trends of complex and volatile financial markets.

Keywords

Cryptocurrency prices, Spark Streaming, Data processing, Visualisation, Apache Kafka, ARIMA, Sentiment Analysis

1 Motivation

The motivation for building a cryptocurrency price prediction model is rooted in the significant growth and impact of the cryptocurrency market over the past few years. Cryptocurrencies have emerged as a new asset class that has the potential to revolutionize the financial industry. The scale of the cryptocurrency market has grown rapidly, with a total market capitalization of over $2 trillion as of 2021. This growth has attracted a large number of investors and traders who are interested in profiting from the market's volatility.

However, the volatility of the cryptocurrency market poses significant risks for investors and traders. Cryptocurrency prices can fluctuate rapidly and unpredictably, making it challenging to make informed investment decisions. Building a reliable price prediction model that can accurately forecast future trends in cryptocurrency prices can provide valuable insights for investors and traders, enabling them to make informed decisions and reduce risks.

Moreover, the decentralization of cryptocurrencies, which operates outside the traditional financial system, making it very challenging to apply traditional financial models and analysis to cryptocurrency markets. Therefore, building a machine learning-based model that leverages advanced data processing techniques can offer a more robust approach to understanding and predicting cryptocurrency market trends.

In conclusion, the motivation for building a cryptocurrency price prediction model lies in the potential of cryptocurrencies to revolutionize the financial industry, the rapid growth of the market, its inherent volatility, and the decentralization of cryptocurrencies. Building a reliable and accurate prediction model can offer valuable insights to investors and traders, enabling them to make informed decisions and reduce risks in this rapidly evolving market.

2 Literature Review

Numerous studies [3] - [12] have been conducted to examine and comprehend the fluctuations within the cryptocurrency market. [8] provides a comprehensive survey of 146 research papers on various aspects of cryptocurrency trading, analyzing trends, datasets, and opportunities in the field. Studying this paper can provide insights into the current state of cryptocurrency trading research, including available trading platforms, strategies, and risk management.

Several models were analysed for our prediction model. [4] compares the performance of various machine learning methods, including neural networks and support vector machines, for predicting cryptocurrency prices. The results show that some models, such as the LSTM neural network, outperform others in terms of prediction accuracy. Due to the difficulty in interpreting the internal workings of the model due to its complexity and the significant computational resources required to train, we concluded that it may not be the most suitable for real-time or high-frequency trading applications. [10] paper proposes the use of the ARIMA model for stock price prediction, achieving promising results in predicting the closing prices of Taiwan's stock market index. The ARIMA model is a widely used and well-established time series analysis technique that has been successfully applied to stock price prediction, making it a natural choice for crypto price prediction as well. Additionally, crypto coins exhibit strong autocorrelation and seasonality, which are characteristics that the ARIMA model is particularly well-suited to model.

The volatility of cryptocurrency prices is examined by [5] by combining text mining and time-series modeling. The authors use sentiment analysis to extract information from news articles related to cryptocurrencies and then analyze the impact of this information on the volatility of cryptocur-

rency prices. Certain projects also depend on other stronger coins as concluded in [11]. [13] proposes a holistic approach to predicting cryptocurrency prices by combining machine learning and sentiment analysis. We intend to validate this hypothesis by predicting the prices of several cryptocurrencies based on factors other than historical prices.

3 System Overview

Our Real-time Crypto price analysis pipeline is designed for heavy throughput and optimal latency. At a high level, it consists of 3 stages: an ingestion pipeline, a stream data cleaning pipeline, and lastly, an analytical engine. We shall first describe the building blocks of these pipeline stages and then present our implementation details.

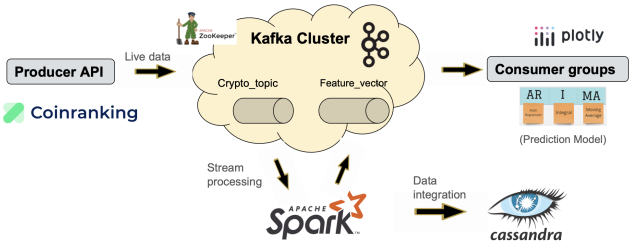


Figure 1: Real-time pipeline overview

3.1 Building Blocks

3.1.1 Docker

Docker allows the creation, deployment, and running of applications in- side container 'pods'. A container is a lightweight package that contains all the necessary software required to run an application which made it simple and easy to build our pipeline.

3.1.2 Kafka

Kafka is an open-source pub-sub platform meant for high throughput, low latency, and scalability. Kafka data can be either schematized or raw data. For our project, we use Kafka version 1.10 to store both raw json crypto price data and also augmented data. Sharding data in Kafka partitions improved publish-subscribe performance.

For our experiments, we ran Kafka with 3 shards, in Kraft mode (so that it doesn't need any configuration management a system like zookeeper) on our local machine in a docker pod.

3.1.3 Spark Streams

Apache Spark Streaming enables real-time processing of streaming data. We used to build scalable, fault-tolerant, and high-throughput streaming processing pipelines that can process data in real-time as it is generated.

Spark Streaming breaks down the continuous data stream into small batches, which are then processed by Spark's core engine in parallel. These batches are typically processed in small time windows, ranging from a few seconds to several minutes, allowing Spark to perform real-time analytics and computations on the streaming data.

3.1.4 Cassandra

Cassandra is a distributed, NoSQL database management system designed to handle large amounts of data. We use this database to store durably results from spark queries and to perform historical queries.

3.2 Ingestion Pipeline

We use the **Coin Ranking API** [1] to collect prices of different cryptocurrencies in real time. In our docker container KafkaProducer, we have a producer processes that periodically polls the API for recent prices for about **50 coins**. The polling frequency is once every 3 seconds, although it is configurable. We have observed that using a higher polling frequency leads to API throttling and is counterproductive. The collected raw data is published into Kafka's **crypto_topic**. This data is in JSON bytes format, and the following is our data schema. (The collected data is semi-structured; therefore, we don't expect it to be strict-adhering).

```
{
'coin': String,
'min_price': Double,
'max_price': Double,
'open': Double,
'close': Double,
'start_time': DateTime,
'end_time': DateTime
}
```

3.3 Stream Processing Pipeline

The raw data collected in the **'crypto_topic'** Kafka topic is processed by a Spark SQL stream pipeline. This is done for two reasons. First, we would like to run some statistical queries and visualise the data being fetched. Secondly, we want to enrich the raw data and extract **'feature vectors'** for downstream learning. Each time series data point for ML must commonly contain the price's recent history for context. The following pseudo-code is briefly the queries we implemented using Spark SQL to compute running Return on Investment on a sliding window of 5 min duration and 30s stride.

$$running\_RoI = inp.groupBy(\text{"}time\text{"},\text{"}5min\text{"},\text{"}30s\text{"}),\text{"}coin\text{"})$$
$$.(\text{"}RoI\text{"}, \frac{(\text{"}max\_price\text{"} - \text{"}min\_price\text{"}) * 100.0}{\text{"}min\_price\text{"}})$$

The next query selects only prominent crypto currencies from input data.

$$popularCoins = List(\text{"}Bitcoin\text{"},\text{"}Ethereum\text{"},\text{"}Dogecoin\text{"})$$
$$prominentPrices = input.filter(\text{"}coin\text{"}.in(popularCoins))$$

Finally, the following is the query to collect different timestamp prices into a single datapoint -

```
SELECT window, coin, collect_list(price)
FROM (
  SELECT price, coin, window(time, '5 min')
  AS window FROM inputDF
)
GROUP BY window, coin
```

3.4 Visualization

The raw data collected into the Kafka queue from the API is ingested by a daemon process that pushes the streaming data into **Plotly Dash**, which can be visualised at *localhost:8050*. Plotly dash picks up data from the Kafka consumer every 1 sec, and the graph is refreshed at the web endpoint.

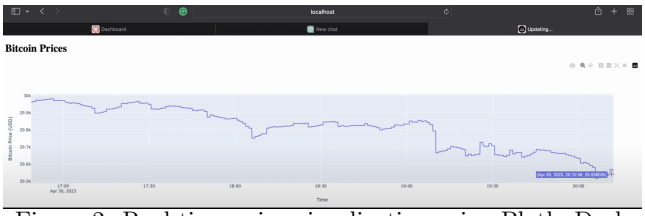


Figure 2: Real-time price visualization using Plotly Dash

Users can also choose to observe the price in a sliding window of 10 minutes instead of observing all the data collected from the API since start time.

3.5 Analytical Engine

We used the Auto-regressive Integral Moving Average (aka ARIMA) model inspired by Dhinakaran *et al.* [9] for our price prediction task. ARIMA is a frequently used model to forecast the price of equities and variables, among other data, that often vary, even in the order of a per-second change. Our analysis combines the historical dataset from kaggle[2] with the real-time data being streamed using the API services to help predict the price of Bitcoin (aka BTC) for approximately three months. In addition, our analytical engine also considers sentiments about these coins over the years and tries to gauge if social media does cause a substantial difference in the price of these cryptocurrencies.

4 Methodology

4.1 The ARIMA Model

Deciding on a model to predict crypto price with good accuracy given the volatility required surveying existing research. Which model, at what frequency and for what window of time into the future will the prediction be practical? These are some critical questions we intend to look at going further ahead.

The Autoregressive Integrated Moving Average (ARIMA) is a method developed by George Box and Gwilyn Jenkins in 1970 and is commonly referred to as the **Box-Jenkins method** [6, 7]. ARIMA is a time series forecasting model that combines autoregressive (AR) and moving average (MA) models with differencing. The general form is $ARIMA(p, d, q)$, where: p: the number of autoregressive terms (lags)
d: the degree of differencing (the number of times the data is differenced)
q: the number of moving average terms (lags of the forecast errors)

$$ARIMA(p, d, q) : y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

Where the important terms are:
- y_t is the time series at time t
- ε_t is the error term at time t
- d is the degree of differencing applied to the time series

Diving deep into our analysis, we now talk about the key decisions we took with respect to using the ARIMA model for our analysis. Firstly, we note that forecasting the prices of assets that vary quite a bit over small intervals is complex and needs a high level of accuracy. This is the case for Bitcoin, as we can see, its price fluctuates significantly even over smaller periods, say one day. In 2021, readers may remember that there was a net 28% change in the price of BTC in a single day. Our dataset comprises per-minute data having feature vectors Open, Close, High, Low, Volume of the currency traded in that period. ARIMA with **exogenous regressors** is used with exogenous variables: the feature vector of open, close, high, low, and per cent change can be included as exogenous variables to help improve the model's accuracy. The specific values for p and q are determined through model se- lection techniques based on the characteristics of the data. d: the degree of differencing is 1

We use historical data set ranging from the start of 2015 up until now. We have shown a visualization of trends shown in the prices of BTC for the time-period in Figure 3. Here we see the volatility of BTC even on a daily basis, leave alone quarterly or yearly basis.

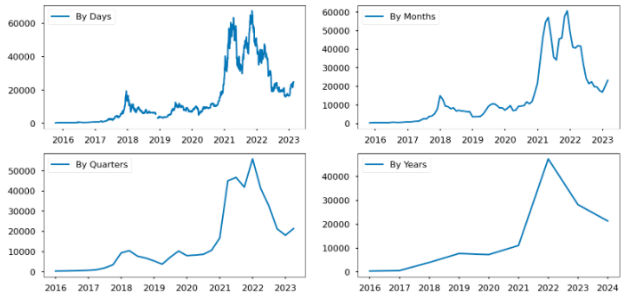


Figure 3: Price Trends of BTC.

We decided to go ahead with a per-minute frequency dataset looking at the above visualizations as the data points for quarterly or even monthly changes would be less for performing regression. We make use of a lot of inbuilt Python libraries for data preprocessing. Initially, the feature vectors are combined to form a weighted price by taking averages for the feature vectors mentioned above and scaling them with the standard of the volumes. This would be our crucial feature-vector going ahead for our analysis.

Now, we describe how the ARIMA model is actually used for our task. First, we perform the **Dickey-Fuller** test to determine whether our data is stationary. We find out that it is not fixed, with the test giving results tending to 0. We transform the data using the Box-Cox transformation to a normal distribution to confirm our findings. In achieving this, we are guaranteed that our dataset is well suited to have regression performed on it. Next, we pass this transformed dataset through several parameters to determine which would be well served here. In doing so, we figured out the autocorrelation and the extent of variation of prices of BTC over the years that helped us to determine the input parameters to our ARIMA model. Figure 4 plots the autocorrelation described above.

We passed the weighted data feature vector generated to

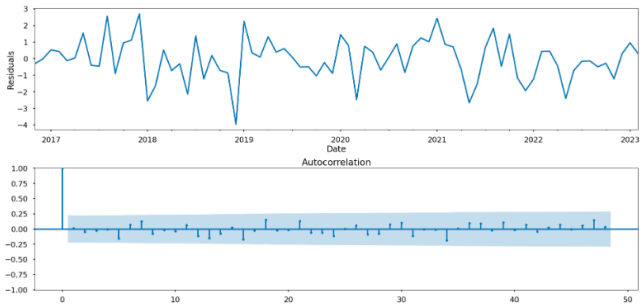


Figure 4: Autocorrelation of Prices for Tuning.

the model along with some timestamps converted into the datetime datatype to validate the performance of our model. A detailed evaluation analysis is taken care of in section 5.

4.2 Sentiment Analysis

The volatility in cryptocurrency may be due to more than just how well a project is doing over some time. Many factors, like market demand, adoption rate, network effects, supply and demand dynamics, regulatory changes, security concerns, and technological advancements, can influence the price of a cryptocurrency. In addition, social media posts by influential personalities cause chaos in the crypto market, causing sentiment changes on social media which might be tightly coupled with the price changes seen in the cryptocurrency market.

There is always a lot of chatter on Social Media about cryptocurrencies. Some posts are directed at the progress being made by the projects, while most result from influential personalities or government organizations announcing something concerning these coins. For example, in 2021-2022, Elon Musk tweeted giving Dogecoin much support. This led to a massive uproar on social media and eventually led to a considerable spike in the currency's price of about 1000%. Many people made money, while most lost much of it as this change was too volatile to persist.

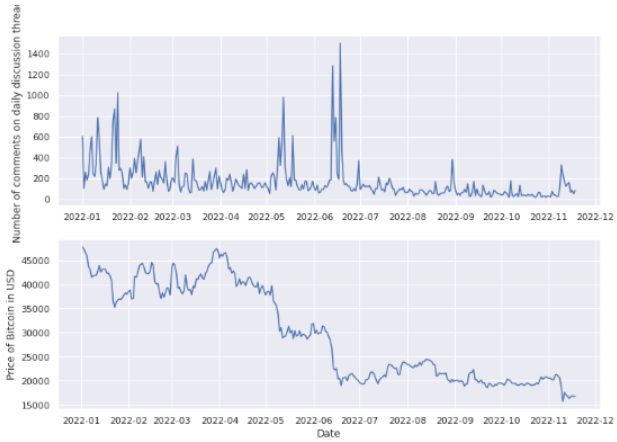


Figure 5: Volatility observed in 2022.

For this scenario, we sourced Reddit and Twitter posts from Kaggle for a period ranging from 2014 to 2022.

4.2.1 *Twitter Trends*

For Twitter, we collected data from Kaggle which looks like a screen capture given below. We cleaned up the tweets, stripping off the text that included hashtags as most of them were found to be redundant and promotional. We used regular expressions for the same. In addition, we removed any hyperlinks that were present. Post cleaning, we made use of the subjectivity and polarity methods of the **TextBlob** library in order to predict if the tweets led to a positive outlook, negative outlook or the audience as a whole remained neutral to it.

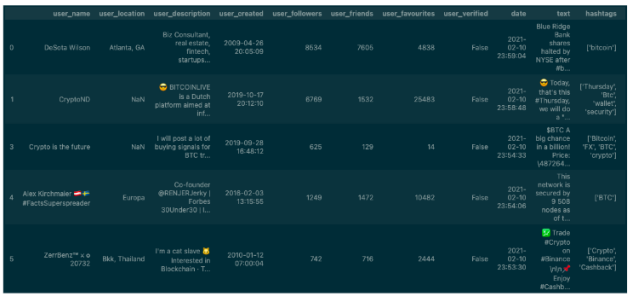
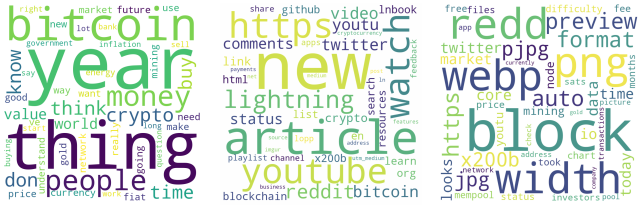


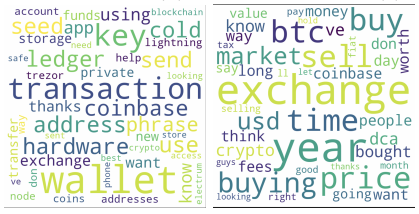
Figure 6: Sample Tweets for BTC.

4.2.2 *Reddit Trends*

Reddit space consists of many different quorums, or channels, that discuss finance, other currencies, projects and equities making it a huge source for text analysis. We used data from Reddit from 2014 to 2022, the same duration as our historical price dataset from Kaggle. We faced several data cleaning challenges before processing this dataset, such as missing rows, columns etc. The feature we were most interested in was the post's content that was present in the "self-test" feature vector. With it, the features relevant to our analysis are upvote ratio, score, total awards received, and unlimited comments. We have also filtered out posts with less than 20 characters in the self-test feature.



(a) topic_0 (b) topic_1 (c) topic_2



(d) topic_3 (e) topic_4

Figure 7: Topics and their corresponding words for Reddit trend analysis

We use **Tokenization** and **Nonnegative Matrix Factorization** to begin sentiment analysis of text posts. The top 2000 words are being used. The 5 topics as shown in Fig 7, are pitted based on the post's sentiment. We plot the case that contributed most to the upvotes (i.e. positive) feelings. We later use this tokenization methodology to form clusters

for our price regression. We use the K-Nearest Neighbours model for the same. We also analyzed the BTC price trends and the Reddit posts to find if any particular day was entirely in the discussion and if there was any correlation. The vice-versa for this case would also hold. We also intended to see if there was an overlap between the topics we generated from the above-mentioned words. The correlation between topics is shown in Fig 8. The regression would not be arbitrary and random for the same reason.

Figure 8: Correlation between the 5 topics.

4.3 Correlation between Cryptocurrency Prices

We have also analysed the dataset from Kaggle to see the variation of different cryptocurrency prices with the more prominent ones like BTC (Bitcoin) and ETH (Ethereum). Certain meme coins like DOGE vary a lot with BTC and ETC. These projects depend on posts by influential personalities like Elon Musk and need a proper infrastructure. However, coins like LTC and XRP hold their fort regarding fluctuations because of the fantastic things they have been doing in the blockchain space. The Fig 9 depicts the correlation between the prices of various cryptocurrencies.

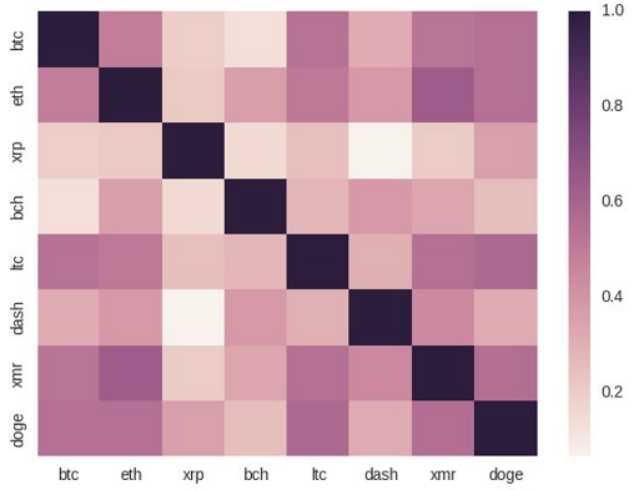


Figure 9: Cryptocurrency Correlation.

Based on the above methodology, we can determine if the general sentiment for a particular time-period is positive or negative. We can also correlate this to the BTC price trends, and determine if the price would go up or down based on the sentiments for the next day. The results for all would be discussed in the evaluation section.

5 Pipeline Evaluation

With the methodology mentioned above, we are able to achieve reasonably efficient and satisfactory results. We ran various datasets to compute the accuracy of the ARIMA model for BTC forecasting and generated visualizations for the sentiment analysis, in addition to using K-NN for price regression accuracy for a particular day.

5.1 Real-time performance metrics

Our pipeline is able to ingest upto 80 price datapoints from coinranking API, with about 150ms delay for an ingested raw datapoint to be reflected in our output. We believe that the current bottleneck in our pipeline are the web-scrappers for price which are using free-version of API while polling. Using the paid version that provides many more crypto coin's price data without throttling will put our pipeline to better utilization.

5.2 Prediction accuracy

Plotting the trends helped us identify ARIMA to be good for our use case. We tried to take various date ranges into consideration to see where the pricing trends might go out of order. Looking at the volatility in BTC prices, the forecast was good for a period of 3 months. The probability of predicting the right price for the future period came out at **81.26%** for the same time period which is quite a good result if we consider how BTC price has ranged in the years before. We have plotted the prediction using ARIMA model below which forecasts the price for the months of May, June and July of 2023. As we can see from figure 10, the prediction is quite in line with the historical prices and as we stand today, BTC price has been on the price ranging from 29,000 to 30,500 USD.

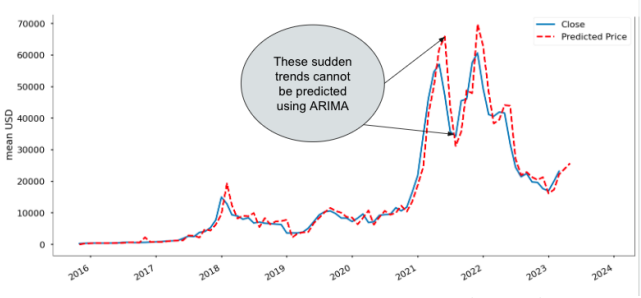


Figure 10: Price Forecasting using ARIMA.

5.3 Twitter Sentiment Analysis

For this, we visualized the polarity and subjectivity calculated and slotted them in positive, negative or neutral sections. This was more of an analysis as to how many people have been positive about BTC over a period of 8 years. Turns out very few were negative, but that negativity also had a lot to do with the crash seen in the year 2018 & 2021. We have plotted the polarity for the same in figure 11.

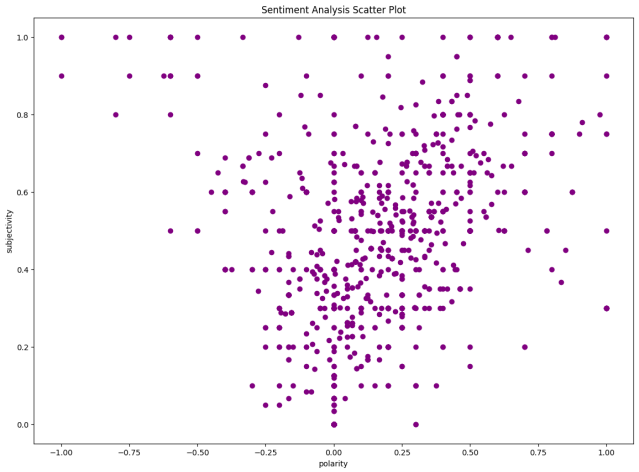


Figure 11: Scatter Plot for Twitter Sentiment Analysis.

5.4 Reddit Sentiment Analysis

We used tokenization for this. The best single day topic was found out to be for **topic 3** as depicted in the figure 12. These were the words that led to the **maximum positive price change**. Using the conclusions, we used the K-nearest neighbours model and determine if the price would go up or down based on the sentiments for the next day. For example, if we consider N = 6 (where N is the number of clusters), the accuracy when predicting if the next day markets are up or down is **0.544**, which is a bit towards the lower side, but again the volatility creeps in this uncertainty but is a good metric to find out the impact of these sentiments.

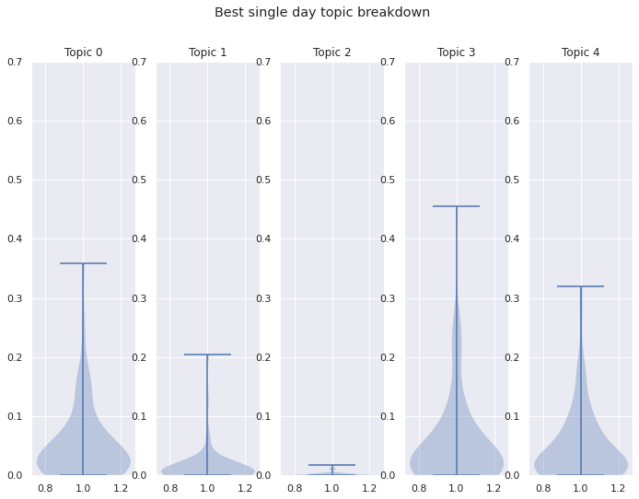


Figure 12: Best Single Day Topic.

6 Conclusion

Cryptocurrencies are a booming technology with a growing market and immense potential for substantial investment returns. Getting the future predictions right can be a game changer. Still, it involves crunching a lot of historical and live data while carefully testing and selecting models to predict prices in near future with those models while keeping in mind the volatility of the market sentiment. Our pipeline is built around a containerised system using docker for ease of orchestration and scalability. It scrapes such real-time data from a reliable API and publishes them into Kafka topics for high throughput and low latency processing. It has a simple yet powerful spark SQL pipeline to enrich the raw data and extract feature vectors for downstream learning. Our ARIMA model is a good predictor for prices in the near future, and integrating it into the live pipeline could help investors make a sound decision based on the history and real-time trends. The volatile sentiments of the market can easily be observed from the social media trends, as shown by us in the analysis of Twitter and Reddit posts and their correlation with the pricing of these currencies. In future, a streamlined pipeline to fetch and process live social media and news feeds for consumption by a more complex model would be ideal.

7 Future Works

The following are some possible future directions for research based on the findings of our cryptocurrency price evaluation paper:

- Examining the impact of regulatory changes and geopolitical events on cryptocurrency prices.
- Build a holistic model incorporating the factors studied in this paper along with other financial and economic indicators to further improve the prediction accuracy of the model.
- Extending the time horizon of the model to capture long-term price trends and cycles. A long-running setup will enable us to collect live data for long and make meaningful predictions based on it.
- Conducting a comparative study of various prediction models and evaluating their performance. This would involve exploring the potential of deep learning and other advanced machine learning techniques for cryptocurrency price prediction.

8 References

[1] CoinRanking API. https://developers.coinranking.com/api.

[2] Kaggle Crypto Dataset. https://www.kaggle.com/competitions/g-research-crypto-forecasting/data.

[3] Jg Aravindan and Rama Krishnan V Sankara. Parent coin based cryptocurrency price prediction using regression techniques. In *2022 IEEE Region 10 Symposium (TENSYMP)*, pages 1–6, 2022.

[4] Atieh Armin, Ali Shiri, and Behnam Bahrak. Comparison of machine learning methods for cryptocurrency price prediction. In *2022 8th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–6, 2022.

[5] Francesco Benvenuto, Massimiliano Drago, and Gabriella Pasi. Cryptocurrency price fluctuations: An analysis through text mining and time-series modeling. *IEEE Access*, 6:41161–41169, 2018.

[6] George EP Box and Gwilym M Jenkins. A distribution-free method for non-parametric analysis of covariance. *Technometrics*, 12(4):677–685, 1970.

[7] Estelle Bee Dagum. The x-ii-arima seasonal adjustment method. *Journal of Official Statistics*, 21(4):517–526, 2005.

[8] Fan Fang, Carmine Ventre, Michail Basios, Leslie Kanthan, Lingbo Li, David Martinez-Regoband, and Fan Wu. Cryptocurrency trading: A comprehensive survey, 2022.

[9] Dhinakaran K, Baby Shamini P, Divya J, Indhumathi C, and Asha R. Cryptocurrency exchange rate prediction using arima model on real time data. In *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, pages 914–917, 2022.

[10] Shu-Ping Lo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*, pages 170–174. IEEE, 2014.

[11] Nicolau Reinhard and Lionel Martellini. Cross-correlations between cryptocurrency markets: A multi-scale analysis. *Journal of Risk and Financial Management*, 12(2):67, 2019.

[12] Danial Saef, Odett Nagy, Sergej Sizov, and Wolfgang Karl Härdle. Understanding jumps in high frequency digital asset markets, 2021.

[13] Minul Wimalagunaratne and Guhanathan Poravi. A predictive model for the global cryptocurrency market: A holistic approach to predicting cryptocurrency prices. In *2018 8th International Conference on Intelligent Systems, Modelling and Simulation (ISMS)*, pages 78–83, 2018.