Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/16968
DOI URL: http://dx.doi.org/10.21474/IJAR01/16968

RESEARCH ARTICLE

LIVE CRYPTO SENTIMENT: SOCIAL MEDIA INFLUENCE ON MULTI-SECTORAL COIN AND ITS IMPACT ON PORTFOLIO RISK MANAGEMENT, USING DATA ANALYTICS.

Aditya Kumar, Vishal Jha and Kunal Srivastav

Manuscript Info

Manuscript History
Received: 22 March 2023
Final Accepted: 25 April 2023
Published: May 2023

Abstract

The project objective is to collect historical twitter tweet data and process it further for statistical analysis. This analysis is further done to extract a general sentiment of the collective tweet data - which will imply the social media influence on coin market share, which will be used to provide a supportive role in predicting the LTP(last traded price) trend for a specific cryptocurrency and sectorial coin segments. The trends are classified as bearish, bullish and neutral. This project will only be a supportive indicator for strategies and not a sole buy/sell predictor which will support in portfolio risk management.

Introduction:-

Cryptocurrencies are an alternative kind of currency made of different kinds of decentralized crypto. These cryptocurrencies demonstrate traits of quick growth and rapid decline, indicating a high level of price volatility over time. Emotional ethics also and not just capital value, are said to impact financial system decisions, according to behavioral economists. Dolan and Edlin (2002) encouraged this idea and argued that emotions are a factor for the decisions in the market. Therefore, technologies like sentiment analysis are helpful in demonstrating how factors like emotions as well as economic fundamentals have an impact on the price of a commodity.(Panger, 2017).

Objective:-

The project objective is to collect historical twitter tweet data and process it further for statistical analysis. This analysis is further done to extract a general sentiment of the collective tweet data - which will imply the social media influence on coin market share, which will be used to provide a supportive role in predicting the LTP(last traded price) trend for a specific cryptocurrency and sectorial coin segments. The trends are classified as bearish, bullish and neutral. This project will only be a supportive indicator for strategies and not a sole buy/sell predictor which will support in portfolio risk management. **Due to their volatility in the current market, cryptocurrency price fluctuations are very tough to forecast. Currency fluctuations are significantly influenced by bank regulations, political risk, and regulatory organizations, hence it is important to have a certain degree of price variation forecast to protect investor/trader capital from worst case scenarios** - like sudden market crashes and market disruptions.

The mood expressed on social media and web search analytics tools like Google Trends and twitter activity have a significant impact on cryptocurrency price swings. Given that so many people tweet about cryptocurrencies, even when their prices fall, sentiment on Twitter about future cryptocurrency prices is often favourable.

Corresponding Author:- Aditya Kumar

Methodology:-

The "Price" of cryptocurrencies was the subject of numerous research that employed a variety of methods, including the Autoregressive Integrated Moving Average (ARIMA) time series model. Using daily, weekly, and monthly time series, they seek to predict the prices. By examining public mood on Twitter and figuring out how investor sentiments relate to one another, it is feasible to predict the volatility of cryptocurrency prices. Rahman et al. consequently presented machine learning algorithms based on the Twitter dataset. The goal of the study is to determine if user sentiment and the price of BTC are related. However, they employ a range of techniques, including Linear Regression (LR), Decision Tree Regression (DTR), and Support Vector Regression (SVR). According to the experiment, there is a discernible correlation between sentiment on Twitter and price fluctuation, with the decision tree algorithm having the highest accuracy i.e. **75%**, when compared to other algorithms. Theodore Panagiotidis, Thanasis Stengos ,Orestis Vravosinos conducted experiments to understand feature selection relative to application of lasso regression and regularization for model development.[4]

Bassam Charif Hamdar, Tarek Saad, Mohammad Hamdar suggested an idea on how a 2 step phase can be implemented to help establish a digital currency in Lebanon, and how gold and oil prices can drastically have a change in prices of the cryptocurrency[5]. Analyzing sentiments through tweets on these can trigger the price to any side. Jethin Abraham, Daniel Higdon, John Nelson, Juan Ibarra presented a model for - Bitcoin and Ethereum price prediction on the basis of daily average tweets. Also it was suggested that tweet volumes are a better metric than sentiment value. A Linear regression algorithm was used in prediction with data from google trends and past prices which was found highly correlating. Stuart Colianni, Stephanie Rosales, Michael Signorotti researched on execution of Naive Bayes, Support logistic regression, in the Scikit Python library. Also it was suggested to formulate a dictionary where each word is highly correlated to the cryptocurrency market[6]. For the DANN model and CNN-BiLSTM model, Bhaskar Tripathi and Rakesh Kumar Sharma created a Bayesian Optimisation recommender with a reduced network size that has three and five hidden layers respectively. Their research indicates that a smaller neural network size is generally preferred because it takes less time to train it for our objective.[7]. Marzieh Rostami, Mahdi Bahaghighate Morteza Mohammadi Zanjireh provided with inexhaustive research on Grid search (GS) algorithm - in a clever strategy that uses a manually determined hyperparameter subset of the search space for a chosen algorithm to try to exhaustively explore the data space while Random search chooses values for each hyperparameter separately depending on their probability distributions.[8]. We will be using the GS algorithm for hyper parameter tuning in our proposed model. According to Arif Furkan Mendi, Twitter is an influential text-based medium for creating personal communication channels and exchanging thoughts. This website has more than 330 million active visitors, making it incredibly well-liked. Data from twitter is scraped and cleaned with useful feature selection - can be a very powerful data source for model training in sentiment analysis scope. To further delve into sentiment analysis, Sara Alqethami ,Shaimaa Alghamdi, Hosam Alhakami and Tahani Alsubait used Bayesian Theorem to solve text classification and sentiment analysis using the ML model it is used for text categorisation.[10]

Expected Outcome

A comparative examination of DL was only done in a few experiments in an effort to determine the best preprocessing method. These studies were carried out over a range of time periods, making them comparatively outdated in the context of the quickly developing crypto market. Since the literature has not been fully utilized, more study is needed to forecast bitcoin values and analyze investor sentiment utilizing DL techniques. Therefore, further research is required to both confirm that current findings are still applicable in 2022 and to identify novel

Identify, Research ,Collect Idea And Findings

""On the determinants of bitcoin returns: A LASSO approach" by Theodore Panagiotidis, Thanasis Stengos, and Orestis Vravosinos (March 2018):

This research paper investigates the determinants of bitcoin returns using a LASSO (Least Absolute Shrinkage and Selection Operator) approach. The authors consider 21 variables and utilize LASSO regression for variable selection and regularization.

"Economic and Technical Modeling of the Lebanese Crypto Currency Implication for a Digital-Lira DL" by Bassam Charif Hamdar, Tarek Saad, and Mohammad Hamdar (February 2021):

This paper proposes a two-step phase to establish a digital currency in Lebanon. It explores the impact of gold and

oil prices on the cryptocurrency market and suggests that analyzing sentiments through tweets related to these factors can influence cryptocurrency prices.

"Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis" by Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra (2018):

The authors predict the prices of Bitcoin and Ethereum based on daily average tweets. They argue that tweet volumes are a more informative metric than sentiment values. The prediction model utilizes linear regression with data from Google Trends and past prices, revealing a strong correlation.

"Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis" by Stuart Colianni, Stephanie Rosales, and Michael Signorotti (2015):

This research paper explores the use of Naive Bayes, Support Vector Machines (SVM), and logistic regression algorithms for algorithmic trading of cryptocurrencies. The authors suggest formulating a dictionary that correlates each word with the cryptocurrency market.

"Modeling Bitcoin Prices using Signal Processing Methods, Bayesian Optimization, and Deep Neural Networks" by Bhaskar Tripathi and Rakesh Kumar Sharma (October 2022):

This paper proposes a model for predicting Bitcoin prices using signal processing methods, Bayesian optimization, and deep neural networks. The authors find that the DANN model and the CNN-BiLSTM model with reduced network sizes yield favorable results in terms of training time and accuracy.

"Bitcoin daily close price prediction using optimized grid search method" by Marzieh5 Rostami, Mahdi Bahaghighate, and Morteza Mohammadi Zanjireh (February 2022):

The authors employ the grid search (GS) algorithm to predict the daily close price of Bitcoin. They compare it with the random search algorithm and find that GS, which exhaustively explores the data space, outperforms random search.

"A Sentiment Analysis Method Based on a Blockchain-Supported Long Short-Term Memory Deep Network" by Arif Furkan Mendi (June 2022):

This research paper introduces a sentiment analysis method based on a blockchain-supported Long Short-Term Memory (LSTM) deep network. The author emphasizes the influence of Twitter as a text-based medium for personal communication and exchanging thoughts.

"Cryptocurrency Price Prediction using Forecasting and Sentiment Analysis" by Shaimaa Alghamdi, Sara Alqethami, Tahani Alsubait, and Hosam Alhakami (2022):

The authors utilize Bayesian Theorem and machine learning models for text classification and sentiment analysis in the context of cryptocurrency price prediction. They highlight the effectiveness of the Bayesian approach in text categorization.

Existing Problems & Proposed Solutions

Model Building

For proposed solution, we will be creating an ensemble learning machine learning model (a combination of the least square linear regression (LSLR) and Bayesian ridge regression models)

We used a bagging method for many different models to generate the final prediction. We collected the results from different categories and either summed the averages or identified the probability of their occurrence. We found that having an ensemble method of learning was beneficial for error reduction in a particular model.[11]

We chose LSR because it minimizes any necessary error that occurs. We use an array of independent and dependent variables to determine the coefficients. And I used Bayesian ridge regression because it adds a lambda parameter to

the input values that penalizes the beta coefficients and shifts them toward zero. Bayesian ridge regression returns a probabilistic model with a Gaussian parameter.[10]

In model development, our first focus will be on the features selection process. **Feature selection is the process of selecting a subset of relevant features for use in a model. The goal of feature selection is to improve the model's performance, reduce overfitting, and reduce training time.**
Simple imputer missing values datas median value , standard scalar data , one hot encoder for converting true and false to 0 1.

There are several feature selection techniques available, including Lasso regression, Recursive Feature Elimination, and Principal Component Analysis (PCA).

Lasso regression: Lasso (Least Absolute Shrinkage and Selection Operator) is a type of regression that adds an L1 regularization penalty to the cost function. This penalty shrinks the coefficients of less important features towards zero, effectively performing feature selection. Lasso is particularly useful when dealing with high-dimensional data, where the number of features is much larger than the number of samples.[11]

Recursive Feature Elimination: Recursive Feature Elimination (RFE) is a feature selection technique that recursively removes the least important features and builds a model on the remaining features. RFE is commonly used in conjunction with linear models, such as Lasso or Ridge regression.[12]

Principal Component Analysis: Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms the data into a lower-dimensional space while preserving as much of the original information as possible. PCA works by identifying the principal components (i.e., the directions with the most variation) and projecting the data onto these components.[13]

The best feature selection technique for our specific problem will depend on the nature of your data and the goals of your model. **Lasso regression is a popular choice for high-dimensional data, while PCA is useful for reducing the dimensionality of the data. Recursive Feature Elimination can be useful when working with linear models,** but may not be appropriate for all types of models. Ultimately, the best way to determine which feature selection technique is most appropriate for our data will be to experiment with different techniques and evaluate their performance on a validation set.

After feature selection, our goal is to tune model hyperparameters. Hyperparameter tuning is the process of selecting the optimal hyperparameters for a machine learning model. Hyperparameters are the parameters that are set before training the model, such as the regularization parameter or the learning rate.

The process of hyperparameter tuning typically involves searching for the optimal values of hyperparameters using a search algorithm such as Grid Search, Randomized Search, or Bayesian Optimization.
Grid Search: Grid Search is a simple technique for hyperparameter tuning that involves exhaustively searching over a grid of hyperparameters. The user specifies the range of values for each hyperparameter, and the Grid Search algorithm evaluates the model's performance for each combination of hyperparameters.[15]

Randomized Search:
Randomized Search is similar to Grid Search, but instead of evaluating all possible combinations of hyperparameters, it randomly samples from the hyperparameter space. This can be more efficient than Grid Search when the hyperparameter space is large.

Bayesian Optimization:
Bayesian Optimization is a more sophisticated technique for hyperparameter tuning that uses a probabilistic model to guide the search for optimal hyperparameters. Bayesian Optimization uses the information gained from evaluating the model to update the probabilistic model and select the next set of hyperparameters to evaluate.[15]

To determine which technique is the best fit for our purpose, we will experiment with all three techniques and see which one works best for your particular problem.. **Some common hyperparameters to tune include the learning rate, regularization parameter, and number of hidden layers in a neural network.**

We will use a validation set to evaluate the performance of different hyperparameters, and then test the final model on a separate test set to ensure that the model has not overfit to the hyperparameters.

Our final step to model building would be selecting a regularization technique. **Regularization is a technique used to prevent overfitting in machine learning models.** The basic idea behind regularization is to **add a penalty term to the loss function during training, which discourages the model from learning overly complex relationships between the input and output.**

There are two commonly used regularization methods in machine learning: L1 regularization (Lasso) and L2 regularization (Ridge).

L1 regularization (Lasso):

L1 regularization adds a penalty term to the loss function that is proportional to the absolute value of the model's weights. This has the effect of shrinking some weights to zero, effectively performing feature selection. L1 regularization can be useful when dealing with high-dimensional data, where many of the input features may be irrelevant.

L2 regularization (Ridge):

L2 regularization adds a penalty term to the loss function that is proportional to the square of the model's weights. This has the effect of shrinking all the weights towards zero, but not necessarily to zero. L2 regularization can be useful when the model has many input features, but all of them may be relevant to the output.

L1 regularization may be useful if we have high-dimensional data and want to perform feature selection, while L2 regularization may be useful if we have a large number of input features and want to avoid overfitting.

We will decide on regularization technique selection only after finalizing the input data feature set completion and selection model run.

Data Building

Before moving on model building, we will be web scraping our twitter data.

Our selected features for Twitter data include:

Text content: The most basic feature of a tweet is the text content itself. This can include the words used, the hashtags, and any URLs or mentions included in the tweet.

User information: Twitter users have a profile that includes information about themselves, such as their username, location, bio, and number of followers and following.

Metadata:

Each tweet also includes metadata that provides information about the tweet, such as the time it was posted, the number of retweets and favorites it has received, and the language in which it was written.

Sentiment:

Sentiment analysis can be used to extract the emotional content of a tweet, which can be useful for understanding the attitudes and opinions of Twitter users.

Network information:

Twitter is a social network, so information about the relationships between users can be useful for understanding the structure of the network and how information spreads through it.

Here high dimensionality can lead to overfitting and increased computational complexity.

In general, we will start with a small number of features and gradually increase them if needed. It's also important to consider the potential interactions between features, as these can add additional complexity to the model.

Dimensionality reduction techniques such as principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) will be used to reduce the number of features while retaining the most important information.

We ran our PCA modified dataset on multiple models from different model families. We compared models using various performance metrics listed below,
Kappa, measures the agreement between predicted and actual values, taking into account the possibility of agreement occurring by chance. The higher the Kappa value, the better the model's performance.

F1, is the harmonic mean of precision and recall, providing a balanced measure of the model's accuracy. A higher F1 score indicates better overall performance.

MSE metrics measure the average squared error between predicted and actual values, with lower values indicating better performance.

RMSE metrics measure the average squared error between predicted and actual values, with lower values indicating better performance.

Results:

	Model	MSE	RMSE	F1
0	Linear Regression	0.580095	0.761640	0.379559
1	Ensemble 1 (LR + BR)	0.578938	0.760880	0.376155
2	Ensemble 2 (LR + DT + RF)	0.064708	0.254378	0.996056
3	Ensemble 3 (LR + DT + RF + BR)	0.145010	0.380802	0.978487
4	Ensemble 4 (LR + BR + Support Vector)	0.060249	0.245456	0.994735

We chose model 4, i.e., Ensemble 4 (LR + BR + SupportVectorMachine) because, It demonstrates high performance across all evaluation metrics, including a high F1 score, high Kappa value, and relatively low MSE and RMSE. It also provided more accurate results in case of predicting buy/sell indicator.

The Least Squared Linear Regression + Bayesian Regressor Model gave the following result:
Model Details
We have used linear and bayesian_ridge models to get this much accuracy for our model.Accuracy achieved for predicting the directional movement of bitcoin is **77.22%**.

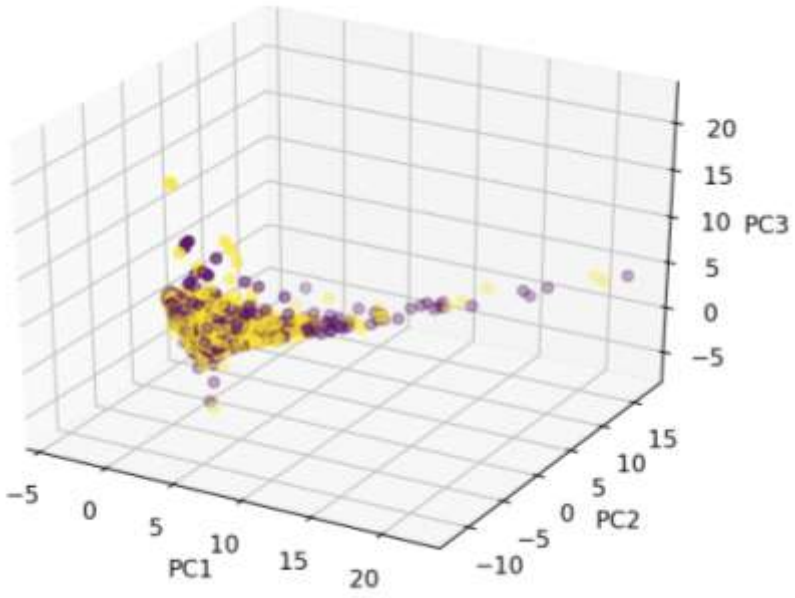
Mean Squared Error (**MSE**): **0.578938**
Root Mean Squared Error (**RMSE**): **0.760880**

The Least Squared Linear Regression + Bayesian Regressor + Support Vector Machine gave the following result:
Model Details
Accuracy achieved in this model is **100%** in which we predicted **764 predictions**. This model was trained on a larger dataset in which there were **15-minute intervals** over a period of **699 days**.

Mean Squared Error (**MSE**): **0.060249**
Root Mean Squared Error (**RMSE**): **0.245456**
Coefficient of Determination (**R-Squared**): **0.916**

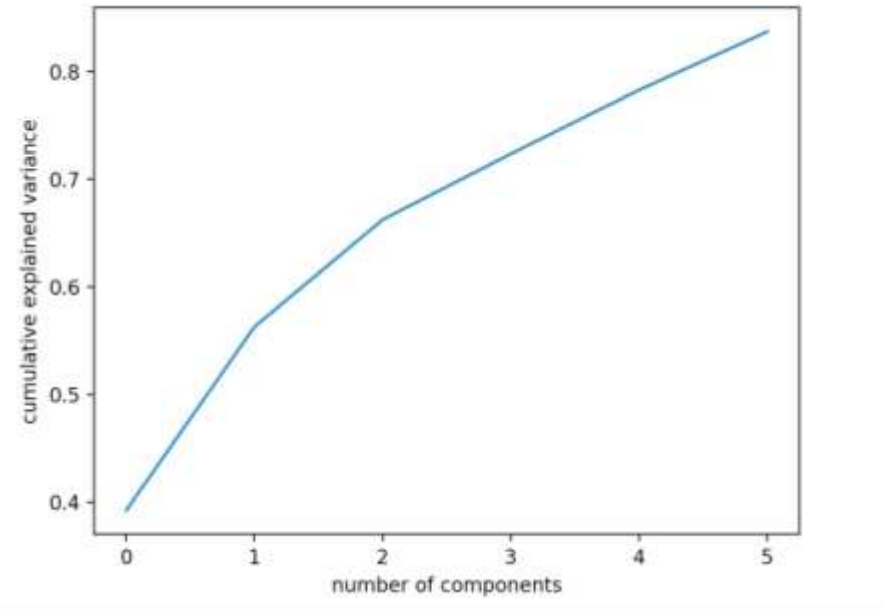
All methods and concepts we have used here are,
Cleaning data Removing all the null values and making the data accurate so proper feature selection can be done we did the cleaning of data.
Sklearn We used sklearn for data analysis and Machine learning in our model.
Text Blob It is used for sentiment analysis of the tweets given by user
Tweepy- It is a framework that we use to access the twitter API easily from which we extracted the twitter data easily.
Finance- It is a framework that we use to access the yahoo finance API easily from which we extracted the Bitcoin data set.
Resampling- Re sampling was done by us so we get a more precise data set.
Least Squared Linear regression- Used to determine the best fit line of the data set.
Bayesian Ridge- As we have a big data set and did our best to remove the anomalies from the data sets we have used a bayesian ridge mechanism to deal with insufficient data.
Mean Squared Error- It is a risk function that helps measure the square of errors

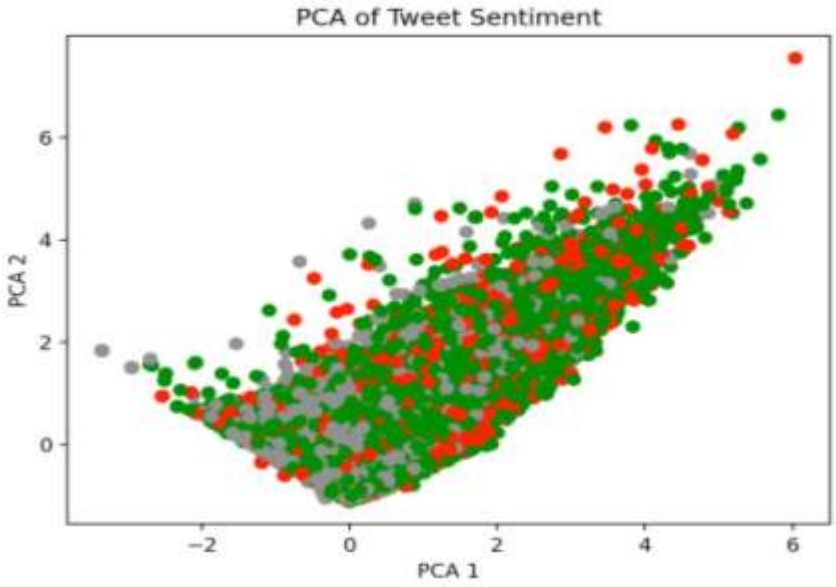
Root Mean Squared Error- It is the root of MSE the lower the RMSE value the more accurate our result is.
Normalisation - It was done by us to transform the features and bring them to a similar set of values.
Principal Component Analysis - It was done by us to reduce the redundant features and reduce the dimensionality of our data so we can receive accurate results.
L1 regression- It helped us reduce the coefficient values to almost 0 so we can receive accurate results
One hot encoding - It converts statements from true or false or in general statements to 0 and 1 and as we have a matrix formed with binary digits it is easier to process the data for training.
SVM -Although SVM works better on small data sets it has helped us receive optimal results for the data given to it.



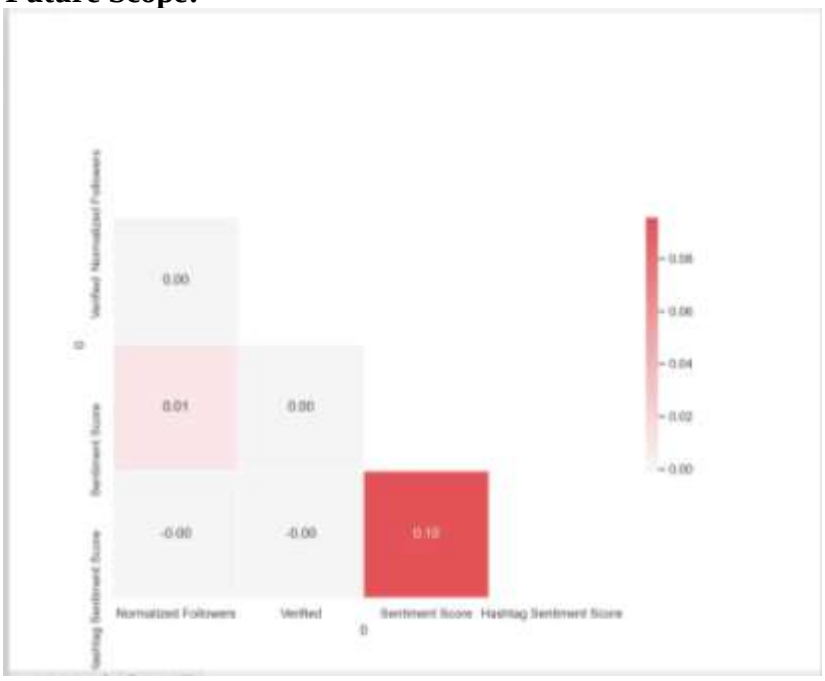
PCA Observed Graphs
Principal Component Analysis - It was done by us to reduce the redundant features and reduce the dimensionality of our data so we can receive accurate results. We have selected 4/5 PCAs as they fell in the preferred 80-85% cumulative explained variance range.

Correlation Matrix We used this matrix to establish useful relations between features we generated in our dataset.





Conclusion and Future Scope:-



As the results show we were able to predict the market trend for BTC using Sentiment Analysis. We used twitter data to compare the two data sets of the last 11 months, we used the twitter users sentiment and gave polarity to the sentiments of tweets. The data then went under Principal component analysis and further ahead feature selection was done. Moving ahead then we used two model LSLR(least squared linear regression),SVM and Bayesian Ridge model to train the data. Which then helps us compare the backset data and the predicted data. At first the data set that we have taken is a little bit smaller than the second one and we have not applied SVM to it so the accuracy we have received is **77.22%, the MSE is 0.155 , the RMSE is 0.394**. The second data set we have taken is of 11 months and gives us a more accurate prediction and we have put SVM on it. The accuracy received on this data set is 100% the **MSE is 0.083 , the RMSE is 0.288 and the R squared value is 0.916.**

Due to lack of resources there are some limitations that we have faced during the process.The only tweets we have

considered here are of the english language although we have come to a good prediction but when predicting the price it will be more helpful if we consider tweets from all languages.

Also as we have just predicted the market trend further ahead we can take into considerations various factors like that of -
MACD Curve Buy and sell signal
RSI overbought/oversold
Bollinger Bands - Volatility levels
9 EMA short term trend
21 EMA Entry and exit points
50 EMA place top loss
200 EMA Long Term Trend
VWAP Intraday Breakouts
ADX Strength of the Trend

Comparing these with the data sets can help us gather a more accurate prediction about the market and although we have predicted the market trend to avoid immediate loss or small gains which might be negated due to the taxations these things have to be kept in mind for a more accurate study.

References:-

1. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.
2. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.
3. Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(3), 211-244.
4. Theodore Panagiotidis, Thanasis Stengos ,Orestis Vravosinos "On the determinants of bitcoin returns: A LASSO approach", Finance research letters Volume 27, December 2018, Pages 235-240
5. Bassam Charif Hamdar, Tarek Saad, Mohammad Hamdar, "Economic and Technical Modeling of the Lebanese Crypto Currency : Implication for a Digital-Lira (DL)", February 2021 International Journal of Business Administration
6. Stuart Colianni, Stephanie Rosales, Michael Signorotti, "Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis"
7. Bhaskar Tripathi, Rakesh Kumar Sharma "Modeling Bitcoin Prices using Signal Processing Methods, Bayesian Optimization, and Deep Neural Networks"
8. Marzieh Rostami, Mahdi Bahaghighate, Morteza Mohammadi Zanjireh "Bitcoin daily close price prediction using optimized grid search method"
9. Arif Furkan Mendi "A Sentiment Analysis Method Based on a Blockchain-Supported Long Short-Term Memory Deep Network"
10. Shaimaa Alghamdi1 , Sara Alqethami , Tahani Alsubait , Hosam Alhakami "Cryptocurrency Price Prediction using Forecasting and Sentiment Analysis"
11. Thomas G Dietterich, Ensemble Methods in Machine Learning, "Oregon state university, Corvallis" (2001)
12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer.
13. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. Machine learning, 46(1-3), 389-422.
14. Jolliffe, I. T. (2011). Principal component analysis. Springer.
15. Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of Machine Learning Research, 13(Feb), 281-305.