Prediction of the Cryptocurrency Behavior Using Sentiment Analysis of Twitter and News data

Team 137 CSE 6242: Aram Saponjyan, Charles E Fryer, Farhad Sedaghati, Mandeep Mundy

1 INTRODUCTION

Cryptocurrency prices are very volatile and can be subject to significant changes in a very short period of time. Expert traders can use the patterns that exist in the real-world market behavior to correctly predict if a stock price will go up to buy it before the price increases or it will go down to sell it before the price decreases. Various parameters can affect the price of a given stock. One of the factors shown to have a strong correlation with the crypto price movement is business news outlets or data or financial reports. Cryptocurrency is a hot topic these days, and many investors invest their money on buying or selling. With increasing interest in the crypto field; however, there is not much research on the effects of news and Tweets on the crypto market. Instead, the majority of research studies focus on the stock price change.

2 PROBLEM DEFINITION

Sentiment analysis, also known as opinion mining, is the analysis of the feelings such as attitudes, emotions, or opinions, expressed in the news reports/blog posts/Tweets, using natural language processing (NLP) tools to infer whether a section of text is negative, positive, or neutral. The goal in sentiment analysis is to extract useful information from semi-structured or unstructured data from different sources such as articles/blogs/tweets. In this project, we study the impacts of business news items as well as Tweets data on crypto price by performing a sentiment analysis. Results from this project can be used to predict the direction of the market and whether it is a right time to buy or sell crypto to maximize the return and profit.

3 LITERATURE SURVEY

This paper [5] employs two types of machine learning models: generative and discriminative. The models were used to analyze tweets and compute sentiment scores. The accuracy was low but the LSTM model showed predictive power in forecasting the BTC price. However, both models fail to capture the correct sentiment when synthetic data was used to evaluate them. This is useful because we have to be careful with oversampling techniques to balance our data. Similarly, the objective of this paper [11] is to determine the price direction of BTC. They applied sentiment analysis and machine learning to tweets and Reddit posts. They analyzed the time series model prediction of BTC prices using LSTM techniques and found it performed best. This is useful because it confirms the LSTM is a proven technique so we will explore the practice. We will improve upon this paper by exploring more than one data source for sentiment and modeling. Using tweets and google trends, this paper [7] analyzes the influence of social media on price movements for BTC and ADA. The researchers utilized VADER and found that it consistently outperformed other techniques. This is useful because we can explore and apply the VADER model to analyze sentiment. We will improve upon this paper by exploring more than two data sources for sentiment and modeling.

This paper [8] presents three types of RNN algorithms to predict the prices of BTC, LTC, and ETH. Results from the experiments show (GRU) performed better in prediction than (LSTM) and (bi-LSTM) models. This is useful because other papers used LSTM and found it to have some predictive power but this study found the GRU to outperform LSTM which we can study and explore when modeling. This paper [9] proposes a hybrid model of GRU and LSTM to predict the price. Researchers used a combination of time series and neural networks to predict the price and found adding sentimental factors improve performance. This isn't necessarily useful for our project but supports our efforts in that there is a lot of power in text to predict crypto.

The following papers combine data from multiple sources to analyze market sentiment and stock price prediction. [1] used SVM and a daily stock price return response variable looking across indices, sectors, and stocks. The researchers in [2] used an SVM model from yahoo to predict if stocks will either go up or down. This is useful because it concludes that aspect-based sentiment outperformed general message sentiment. The researchers in [3] considered 2 different methods to predict stock price based on financial news articles. Since the model performance wasn't that great, we hope to improve upon the performance by testing and applying these methodologies to cryptocurrencies.

4 PROPOSED METHOD

Current research technique uses either a single social media or news outlet to perform sentiment analysis. Our approach built upon the techniques used in these studies by utilizing multiple social media and financial news sentiment sources into an overall predictive model.

Approach

We built an interactive python dash application that is used to inform cryptocurrency day trading decisions. The trading recommendations and corresponding visualizations are powered by a machine learning model that considers multiple social media and news outlet sentiment signals.

Data Retrieval: API: We pull the training and live data in with the Twitter and News API. We're pulling tweets associated with the cryptocurrency coin with the Twitter developer API. We were able to get the developer account since the project is for academia. There are no historical restrictions with the developer account which is opportune since there is no historical limits. Additionally, we have an improved API GET method rate since the developer account has less restrictions which is key to the training and prediction modules.

We are also utilizing the News API which is a REST API for searching and retrieving live articles from all over the web. With the News API, we can search on the keyword, date published, source domain name, and language. With the News API, we search on the

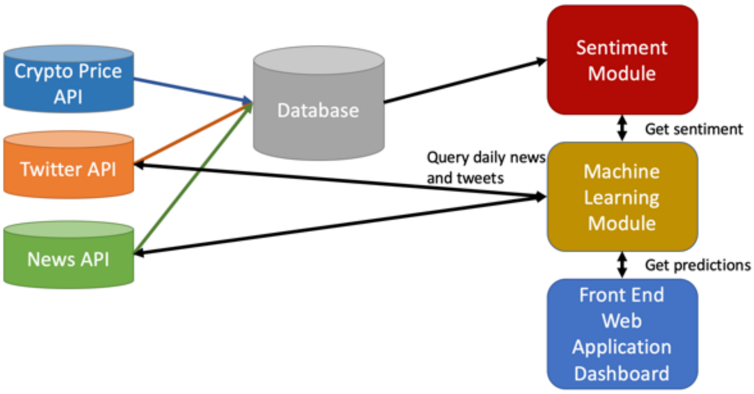


Figure 1—The trading recommendations and visualizations are powered by a machine learning model that considers multiple social media and news outlet sentiment signals.

keyword or phrase of the specific coin to pull in all associated news for that day. We were also able to sort by date published, relevancy to search keyword, and popularity of source with the News API to extract the most relevant crypto news.

To pull in the crypto coin price, we leverage the CoinGecko API. The CoinGecko API allows us to retrieve cryptocurrency data such as price, volume, market cap, and exchange data.

Training Data: Twitter + News

The training data is from March 15, 2022 to April 15, 2022. It consists of tweets and news articles associated with our selected cryptocurrency coin. The model was trained on Bitcoin, Dogecoin, Ethereum, Litecoin, and XRP. For each coin, we pulled in the news and Twitter data for the past 30 days and calculated the sentiment.

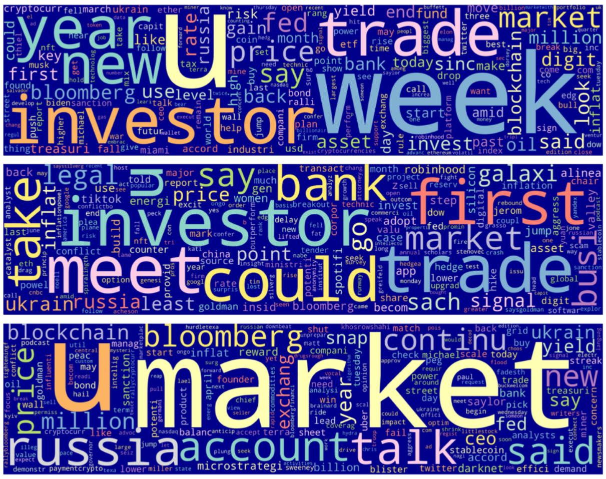


Figure 2—Trained data word clouds by sentiment: Neutral, Positive, and Negative

Choosing an NLP Library: To find an NLP library to implement for sentiment analysis, we looked into Flair, SpaCy, VADER, and TextBlob. These libraries use rule-based sentiment analysis which is fundamental. We tried Flair at first but we noticed conflicting sentiment.

Words that we would associate with positive sentiment were labeled as more negative due to the 'style' of crypto news and social media. Generic sentiment models seem to incorrectly characterize articles describing a "surge" in Bitcoin price as negative. The word "surge" can be characterized as negative in a general context if it relates to surging gas prices or other consumer goods, but in a financial context, the word "surge" is positive.

We opted to use VADER and TextBlob capitalizing on the strengths of each. Textblob returns polarity and subjectivity properties for a given input sentence. Subjective refers to emotion or judgment which is the type of text data we are analyzing. Valence aware dictionary for sentiment reasoning (VADER) is another sentiment analyzer implemented. It uses a list of lexical features which are labeled as positive or negative according to their semantic orientation to calculate the text sentiment. Vader is also optimized for social media data and can yield good results when used with social media data.

Ensemble the Sentiment: Ensemble models are a machine learning technique which combines numerous modeling techniques. Ensemble reduces variance, noise/bias, and improves accuracy. We approach the model features with an ensemble point of view. We use the VADER compound score sentiment for twitter data and TextBlob for news sentiment as 2 sets of independent variables which are input into the LDA model. This final data set is what the model trains on.

Linear Discriminant Analysis: The selected model is Linear Discriminant Analysis (LDA) which is a dimensionality reduction technique. The goal of LDA is to project the features in higher dimensional space onto a lower-dimensional space in order to avoid dimensionality and also reduce resources and dimensional costs. In simple terms, they reduce the features in a particular data set while retaining most of the data. The LDA model verified with k fold validation performed best so it was the selected model. While training the machine learning model, buy / sell labels were restricted to +/- 3 percent daily price movements.

Hosting the Project: The main goal was to make a user-friendly web interface that makes our dashboard interactive and fulfills the user needs. We designed a day trading web interface and ran python through Flask. The web application is hosted in AWS and leverages the CoinGecko API to retrieve cryptocurrency data, the Twitter API for social media data, and the News API to retrieve news data. The application URL can be found here.

We store the pretrained models in AWS and it runs live minute interval predictions on BTC, LTC, ETH, DGC, and XRP. Since the web application is hosted on AWS, the load balancing is handled automatically. We use AWS Elastic Load Balancing and EC2 Auto Scaling which automatically distribute our incoming traffic across multiple EC2 instances.

5 EXPERIMENTS/ EVALUATION

Overview: Questions we sought to answer

We tested if a model could accurately predict the direction of a cryptocurrency using sentiment data. There is a lot of research utilizing different NLP and sentiment techniques to predict price but we sought to implement the ensemble sentiment technique combined with multiple data sets and coins to create the most accurate model. The initial hypothesis was combining the sentiment techniques, multiple data sources and coins would improve overall performance.

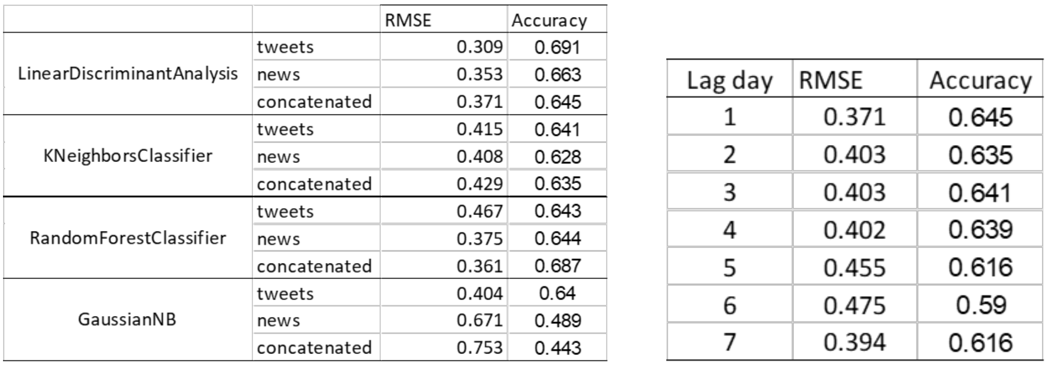
We also investigated the effect of n-day lag day in the model. N-day lag day indicates that today's news headlines and twitter data are correlated with returns on n-day later. In this analysis, we used 1 day lag up to 1-week lag and obtained the RMSE and accuracy using concatenated news and twitter data for all 5 coins with the k-fold cross validation technique.

We created a Google survey and collected detailed information on the application usability for day traders and users. The survey URL can be found here.

Selecting the Best Algorithm

We tested 4 different popular classification algorithms, including: linear discriminant analysis, K-nearest neighbors, Random Forest, and Gaussian Naive Bayes. To perform the comparison test, we combined the data for all 5 coins (Bitcoin, Litecoin, Ethereum, Dogecoin, and XRP). They were trained and tested on the model using the k-fold cross validation technique. Gaussian Naive Bayes is the worst method with the lowest accuracy and highest RMSE. The top accuracy are related to LDA and Random Forest classifiers; while LDA has the lowest average RMSE. Therefore, we pick the LDA classifier as the final classifier to train the model and use it for future predictions. Figure 3a shows the corresponding root mean square errors and accuracy for different models we tested.

Additionally, experimenting with the lag day we found that the 1 day lag results in the highest accuracy and lowest RMSE for the model indicating the price change in the next day are more correlated to the news and Twitter's data sentiment scores, see figure 3b.



		RMSE	Accuracy
LinearDiscriminantAnalysis	tweets	0.309	0.691
	news	0.353	0.663
	concatenated	0.371	0.645
KNeighborsClassifier	tweets	0.415	0.641
	news	0.408	0.628
	concatenated	0.429	0.635
RandomForestClassifier	tweets	0.467	0.643
	news	0.375	0.644
	concatenated	0.361	0.687
GaussianNB	tweets	0.404	0.64
	news	0.671	0.489
	concatenated	0.753	0.443

(a) Model test results

Lag day	RMSE	Accuracy
1	0.371	0.645
2	0.403	0.635
3	0.403	0.641
4	0.402	0.639
5	0.455	0.616
6	0.475	0.59
7	0.394	0.616

(b) Lag test

Figure 3—RMSE and Accuracy

The google survey results in figure 4 found the application to have an easy to navigate and intuitive user interface.

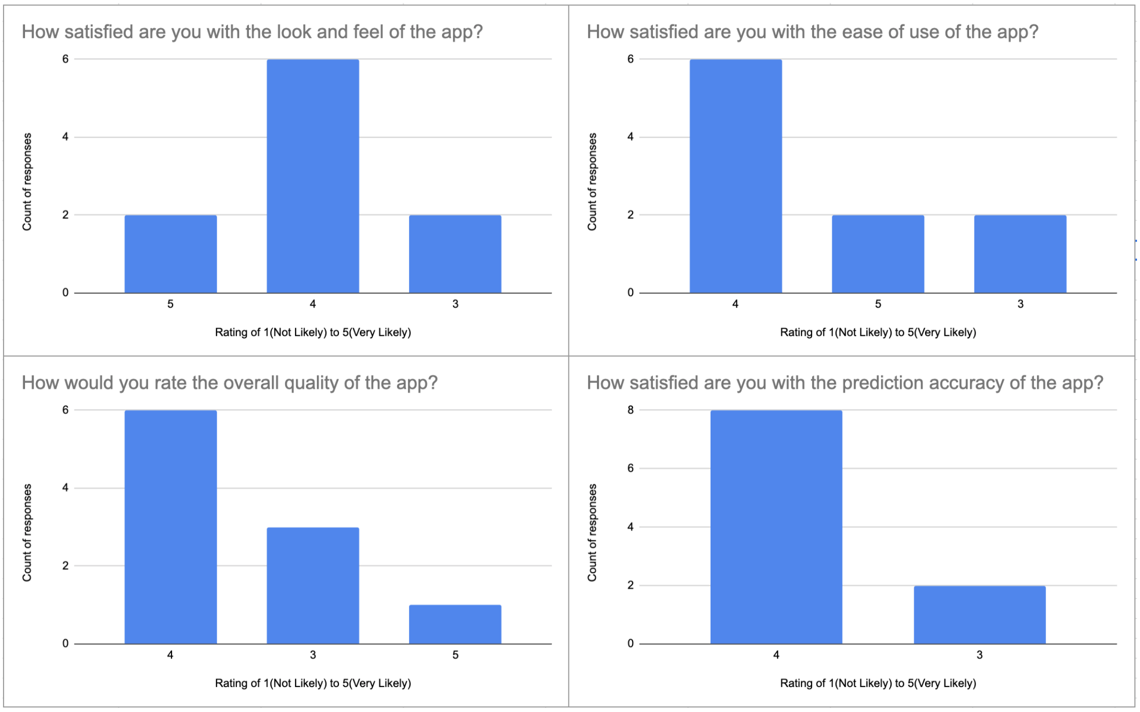


Figure 4—User survey results

6 CONCLUSIONS AND DISCUSSION

Experimentation has highlighted the need to combine news and social media data sources alongside different sentiment analyzers to predict price direction. Current research accuracy varied between 55 and 80 percent. With our approach, we were able to implement a solution with accuracy in this range.

The work has been equally distributed across all team members. We originally started the project with 6 members however one person had to drop the class. Another member was unresponsive during the proposal process so we had to drop them also. So our final group consists of 4 members.

REFERENCES

[1] Li, X., Xie, H., Chen, L. (2014). News impact on stock price return via sentiment analysis. Knowledge-Based Systems, 69, 14-23.

[2] Nguyen, T. H., Shirai, K., Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42, 24, 9603-9611.

[3] Mohan, S., Mullapudi, S., Sammeta, S. (2019). Stock Price Prediction Using News Sentiment Analysis. 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications, 205-208.

[4] Derakhshan, A., Beigy, H. (2019). Sentiment analysis on stock social media for stock price movement prediction. Engineering Applications of Artificial Intelligence, 85, 569-578.

[5] Wong, E. W. (2021). Prediction of Bitcoin prices using Twitter Data and Natural Language Processing. Prediction of Bitcoin Prices Using Twitter Data and Natural Language Processing.

[6] Makrehchi, M., Shah, S., Liao, W. (2013, November). Stock prediction using event-based sentiment analysis. In 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT) (Vol. 1, pp. 337-342). IEEE.

[7] Khurshid, A. R. K. (2017). Cryptocurrency Price Prediction using Sentiment Analysis. Cryptocurrency Price Prediction Using Sentiment Analysis.

[8] Hamayel, M. J., Owda, A. Y. (2021). A Novel Cryptocurrency Price Prediction Model Using GRU, LSTM and bi-LSTM Machine Learning Algorithms. AI, 2(4), 477–496.

[9] Tanwar, S., Patel, N. P., Patel, S. N., Patel, J. R., Sharma, G., Davidson, I. E. (2021). Deep Learning-Based Cryptocurrency Price Prediction Scheme With Inter-Dependent Relations. IEEE Access, 9, 138633–138646.

[10] Ko, C. R., Chang, H. T. (2021). LSTM-based sentiment analysis for stock price forecast. PeerJ Computer Science, 7, e408.

[11] Tarif, A. M. T., Raju, S. M. R. (2020). Real-Time Prediction of BITCOIN Price using Machine Learning Techniques and Public Sentiment Analysis. Real-Time Prediction of BITCOIN Price Using Machine Learning Techniques and Public Sentiment Analysis.

[12] Gupta, R., Chen, M. (2020). Sentiment analysis for stock price prediction. In 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 213-218). IEEE.