Title: Utilizing Social listening to identify Cryptocurrency trends
Authors: Eram Khan, Ashish Panchal, Souptik Banerjee, Nakul Kishore Dumblekar, Ashwini G, Gyanendra Kumar Patro

I. INTRODUCTION:

Traditional money market volatility is generally controlled by central authorities or intermediaries based on a monetary policy to ensure economic stability. Cryptocurrencies use decentralized ledgers without reliance on intermediaries and therefore prices are highly driven by public perception and confidence in the currency in form of crypto currency trades. This results in high volatility and price fluctuations based on perceived effect of external factors like changes in government policies, launch of new currencies with new value additions, recognition of a coin by a business or industry or changing belief of among investors and influencers. Price prediction is therefore difficult solely based on traditional techniques using historical highs/lows and trade volumes. Potential investors are dissuaded by these low accuracy models. This project attempts to reduce this uncertainty by modelling external factors along with investor's sentiment reflected on social networking platform, a primary source driving investment decision.

II. PROJECT GOALS :

The expected outcome of this effort is making investors take informed, more consistent and confident investment decisions, by providing a holistic view of current cryptomarket state, showing the broader "sentiment" about a currency in terms of market engagement and consolidated associated conversation topics with their top discussions and providing price prediction based on these identified emotions along with traditional factors such as daily price fluctuations, associated stock market indices, online search frequency etc, with a short term goal of better return on crypto asset investment and a longer term goal of increase in acceptance and utilization of crypto asset class, making transactions more secure and fast.

III. EXISTING WORK:

Bitcoin transactions reached 1 million/month in 2013, triggering a decade long effort to understand and predict crypto prices, but its disconnect with legacy financial prediction methods to understand investor sentiments have been in vain,[1], due to long curations of market maturity knowledge [2,3,4]. In the last 3 years social media has been identified as an important driving force of investor sentiments and in turn prices. Pandemic induced physical isolation has further boosted social activation [5]. Customers who seldom post have a higher impact [6], however their identification in a colossal of daily conversations is difficult. Although for specific coins, like bitcoin, dedicated forums provided better signals of the price variations [7], as this cannot be stated true for less mature and low popularity coins, these solutions effort are not scalable and cannot be generalized.

The Problem of generalization can be solved to a degree if new topics from conversation were identified from a pool, techniques like word embedding and term-frequency-inverse-document-frequency (TF-IDF) matrices [8], followed by clusters could find significant subjects in unstructured texts to identify important and similar terms, but such techniques undermine the importance of high occurrence and face difficulties in identifying trending topics. This loss of information could be compensated with topic associations to identify important sub-topics associated with important conversation topics from a large collection of social chatter, [9,10] investigate association rule mining, where the current solutions have scope of improvement in understanding the similarity and value add of the words.

As another dimension, segregating these conversations into positive and negative sentiment may give better insight [11,12,13], where most of the models fail to recognise these in long sentences, Multinomial Naive Bayes models could identify these better at an information domain level, although restricted on the trained data and lack generality. Commonly used libraries lack the understanding of modern social media jargons and knowledge about crypto and hence fail to depict the correct directions, additionally extended human emotions like fear, anger, excitement which drive investor actions are difficult to extract. Slight success has been achieved, [14] a range of eight emotional categories as per Plutchik's emotion model. But it is hampered by reliance on lexical constraints for label assignment, which may be difficult to scale with unstructured social media language. These efforts individually fall short in one or the other aspect, have proved the importance of social media sentiments in price prediction and highlighted the importance to solve the generality and scalability problems. Although Traditional financial prediction techniques demand large historical data and rely on a single parameter to estimate prices, but they do not account for external factors such as economic developments and societal sentiment that affect pricing. Recently elements such as government regulation and news, were used to model volatility [15]. Jing-Zhi Huang et al. [16] have used only technical indicators like growth rate. Models such as ARIMAX, GARCH, CART, CNN-LSTM/ResNet [17,18,19,20,21] are extensively explored, utilizing global stock indices, gold prices, and fear gauges such as the VIX and US Economic Policy Uncertainty Index, with much improvement.

IV. PROPOSED METHOD: NOVEL APPROACH DESIGN AND DESCRIPTION:

Most studies performed till date focus on price prediction accuracies. None has worked on answering the question of crypto trends illiteracy and causal sentiment analysis, in which social listening plays a crucial role. Our approach tries to fill these gaps by educating users of these ever-changing social cues and understand the impact of association between them. It will also open a window to better understand

general investor sentiments, dependencies of price trend on specific sentiments and find the most probable explanation of market movements, therefore result in more informed decisions. Most research, as mentioned in the previous section focuses on micro and short term prediction, i.e few hours to 1 day, only and of use to daily traders not long term investors. Understanding social cues demystify signals on medium to long term and explain non-cyclic but possibly recurring events along with transfer of these learnings from old currencies to new. To make this possible our solution contains 2 parts. **(1) Extracting conversation topics by processing social media data**, finding association between them and quantifying popularity/hotness. Along with determining investor sentiments with models trained on twitter conversations of past 3 years, understanding crypto jargons/associated slangs and emotions, compared with historically known approaches. Due to higher quality and content relevance with reduced noise to information ratio Reddit is used as social listening data source, identified as more informative than twitter [22,23]. To create a generalized approach, and capture information about every possible new event, our approach is to target common pool forums and subreddits, specific information can be segregated with topic identification. Enabling us to process Crypto currencies, along with shared applications, usage platforms, industry news etc. The framework is scalable and can combine information from multiple information sources like multiple subreddits, social media platforms or forums. **(2) Quantifying the impact of external, derived financial factors along with quantified social media data** on crypto price and to gauge the prediction improvement. For generalization our approach will predict price for 20 crypto currencies and tokens and understand the relationship and impact of social media and external factors on different type of coins and tokens, including matured tokens, new coins, application driven currencies, exchange swap token and meme/alt tokens. Our initial hypothesis was that prices of different type of currencies relate differently with social media, some may be actively driven and some follow passive conversation on the forums, which might have an active impact in medium to long term, extending to our 2nd and 3rd hypothesis that price prediction accuracy for some currencies are better at medium to longer term and long term learning for some currencies may result in inductive biases. Learning period and prediction accuracy depends on investor profiles which could be gauged by social profiling from online chatters. Our aim is to create a dynamic prediction framework, which could profile the price trends using external and social media factors and forecast these trend segments per currency. We have segmented currencies based on their price variations, determined correlation with various factors and performed Grid Search to tune our hyper parameters. We have thus developed a generalized prediction framework which also provides higher accuracy when compared to existing solutions. By the process of distillation, social flavour is added, along with fundamental and technical analysis enabling investors to weight such factors in their investment decisions. (3) **User Interface**: Two views on tableau with flat file structure *(fig 3.)*, The first view helps user understand more about various cryptocurrencies, find out more inter-related currencies, identify top conversation topics (including coins, applications, trending terms) and associated popularity ranking and top associated terms in the conversations. It also helps in building connections between different topics, context from top associated posts and comments and description. The latter provides forecast for the currency of interest, along with trend and weighted impact of identified factors, with comparative analysis on these factors, for making logical insights.

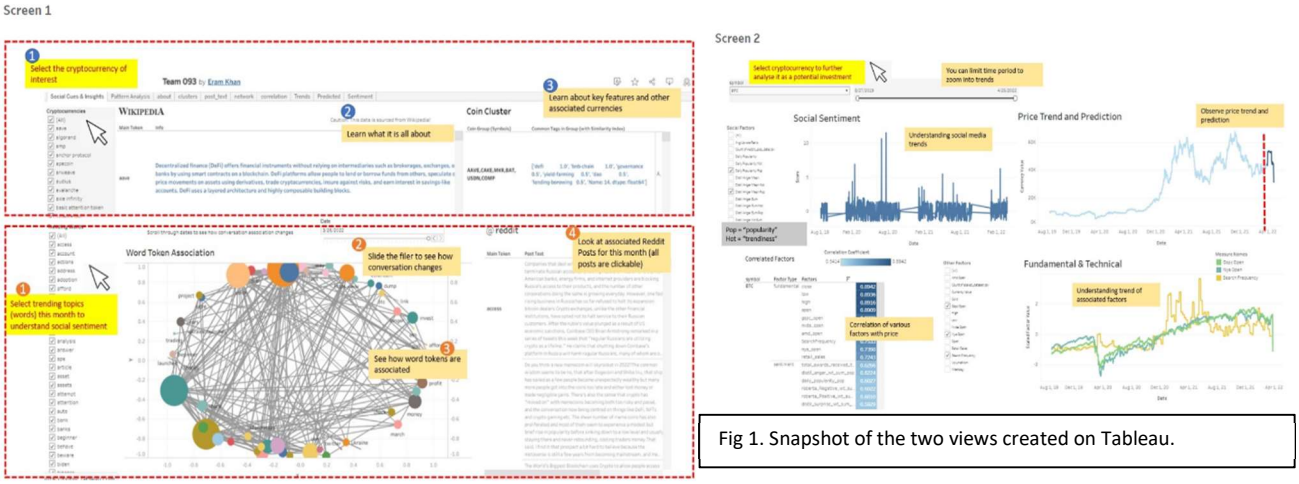


Fig 1. Snapshot of the two views created on Tableau.

Data Description and Extraction Process:

Data	Extraction Process
Social Media Data	3 years of post and comment data was extracted from 2 most active sub-reddits, r/Cryptocurrencies and r/Cryptomarket with **5.5M** total and 4/2K hourly active memebers. The conversations are collected in 2 steps, 1st the collections of **446k Posts** with a rate of 12,000 top posts per month, extracted at day level with **~260k words per day**, performed with **Pushshift API**, taking **21 hours** of crawling. 2nd collection additional post parameters and **172.8 M Comments** associated with these posts, extracted over **54 hours** of crawling using Reddit's API with a dedicated account.
Crypto Price Data	Market price values for 100 currencies are sourced from an existing dataset in data.world which is refreshed with an active connection to the Coinmarketcap.com service, with about **150K records** for the last 3 years.
Online Search Frequency	Data was gathered using **Pytrends** library. This data was fetched for all currencies. A spearman correlation of 0.78 was observed with the 60 days in future price data.
Driving firm stock and commodity price	AMD and NVDIA stock price: **AMD** and **NVIDIA** provide key components for bitcoin mining and therefore the confidence in these companies which is reflected in their stock price is an important data point to consider, with a correlation of ~0.8. Along with Day wise information of global stock indices **DJI, GSPC, IXIC** and **NYA** and **Gold** were fetched from **Yahoo Financial**.
Macroeconomic data	An important factor to estimate future demand for a cryptocurrency. Data on **retail sales, consumer sentiment** and **unemployment** was fetched from **Alphavantage.co**

V. METHODOLOGY

1. Topic Identification: Daily social chatter topics are identified for post and comments separately Steps:***(i)*** Removing Stop words and count vectorization of posts and calculating post frequencies of top 3000 words per day. TFIDF cannot be directly used a) as it represents relative importance of topics by uniqueness within data, deprioritizes possible popular topics, b) Also as the score is a relative ratio or term frequency and inverse document frequency, it loses ability to identify hot or rising topics as it is unable to gauge relative change from the previous days. Using frequency can help us identify popular words but giving higher scores to common words which can be removed based on low *TFIDF score identified over an adjacent 1 year of post data.* ***(ii)***Calculation of daily popularity as a function of number of posts x 10, number of upvotes, number of comments, number of rewards x10. This is further used to calculate popularity score, as a 1 month rolling sum of daily popularity, hotness as 3 days rolling sum of daily popularity. **(iii)**Initial Hypothesis and observations: Social media chatter around a currency is an early identifier for its price change. (Fig. 2) In the start of Dip Cycle investors interests start to

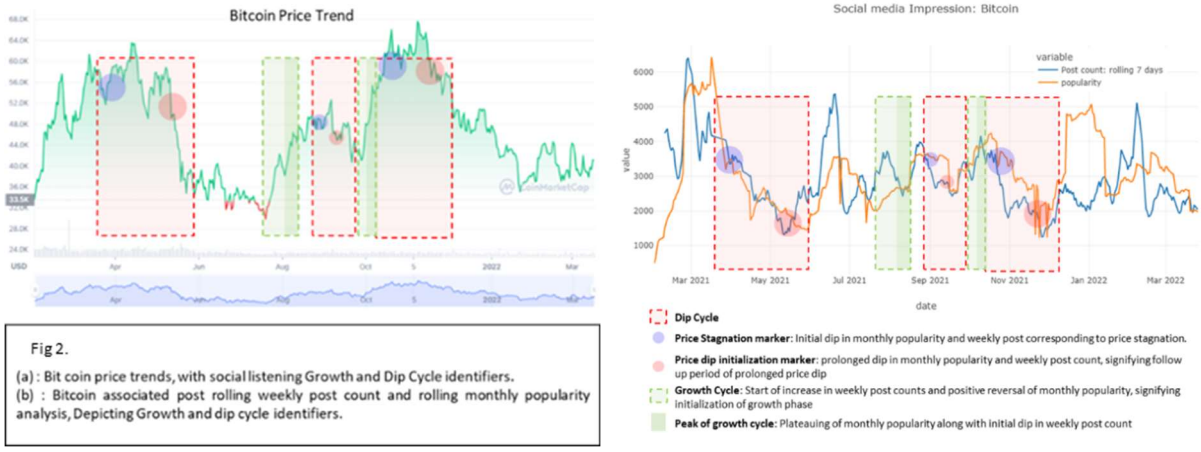


Fig 2.
(a) : Bit coin price trends, with social listening Growth and Dip Cycle identifiers.
(b) : Bitcoin associated post rolling weekly post count and rolling monthly popularity analysis, Depicting Growth and dip cycle identifiers.

dwindle, however the general monthly popularity is still high, which marks the period of price stagnation at high values, followed by a period of decrease in popularity and conversation. A negative change in price is visible with a reversal in conversations while the popularity is still low where investors react to external news and sources. In the start of the growth cycle, it is generally observed when the investors' weekly chatter increases following, additionally, the popularity marking the longer term interest also observes a positive reversal, and continues to grow, even after weekly interactions decrease, which marks the end of growth. Popularity helps in segregating the sentiment trends, as conversations alone spikes during both growth and dip cycles, which may not be useful. **2. Topic Association: (i)** sub-topics associated with a specific currency are identified by performing the same procedure as post topics, however only on comments associated with the posts conversing about the currency. **(ii)** Additionally, the strength of connection of a topic and subtopic is calculated by the ratio of posts sharing the topic and subtopic, to the total posts conversing about the topic. The importance score of each subtopic for a topic is created by subtopic popularity, hotness and rising score based on connections strength, number of comments on the subtopic, votes and awards gained. **(iii)** Initial Hypothesis and observations: Investor conversations

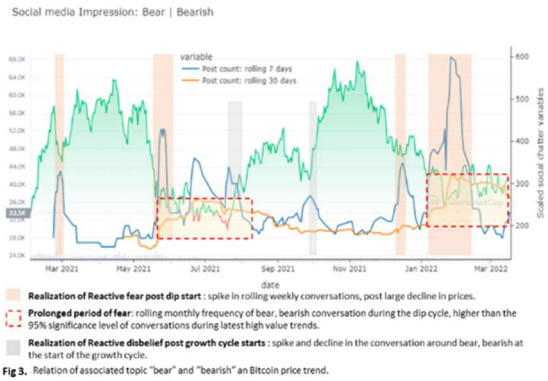


Fig 3. Relation of associated topic "bear" and "bearish" an Bitcoin price trend.

associated with specific subtopics, containing information about their views, news or applications. Known/reoccurring keywords may depict the reactive or active state of general investor sentiment. As depicted in fig. 3., user conversation involving "bear" an "bearish" effectively shows the realization of fear in investors, where the conversations containing these keywords spike after the start of dip cycle, at its largest initial drop. The size and width of the spike is proportional to impact of the drop. While the conversation of these keyword remains higher than the 1.96σ level of the conversations before the dip, the price value was observed to stay down or worsened, gauging the width of possible down period, and is valid until the growth cycle starts. With the start of the growth cycle, investors become sceptic, resulting in spike in keyword conversation, and fades away as the price remains high or grows further, marking the start of acceptance of growth.

3. Sentiment Analysis: Sentiments classification of the conversations adds another dimension, segregating the increase in posts with negative vs positive investor sentiments, which may enable directional relationship analysis with price movement. Comparative analysis was performed on various frameworks, including text2emotion, NLTK's VADER [24], DistilBERT [25], and Twitter-roberta-base model [26,27], where latter is trained on ~124M tweets from Jan-2018 to Dec-2021, with efficient evaluation of recent slangs along with blockchain conversation/domain knowledge and fine-tuned for sentiment analysis with the TweetEval benchmark. text2emotion results were not included in the model as it was not able to understand jargon based sentiments and slangs. Roberta, DistilBERT and Vader resulted in more accurate results, hence were included as parameters for prediction model, including total, average and weighted average popularity and hotness of sentiments. **4. Price forecasting using fundamental and technical factors**: 3 models were tested including ARIMAX, Random Forest and XGBoost Regression on crypto price, search frequency, stock price and macroeconomic data. *(*1) ARIMAX: No trend or seasonality was observed on out-of-sample prediction over multiple currencies for longer horizons. Existing research only show results for a few hours post the training period and is not useful for investment decisions in crypto. *(*2) Random Forest: produced acceptable results, implying non-linear relationship with factors *(*3) XGBoost regression was further explored to optimize the prediction model, giving superior performance *4.1* Xgboost Parameter Tuning and performance: 5 parameters were tuned for prediction optimization. (1)Learning Rate α: to control speed of optimality convergence and avoid overfitting, was tuned with grid-search over a range of 0 to 1, and delta of 0.003. (2) L1 (Lasso) and L2 (Ridge) Regularization parameters were tuned to avoid overfitting. L1 adding sum of absolute error and L2 adds sum of squared error. (3) Horizon: 60, 30 and 10 day future prediction were experimented to evaluate change in prediction accuracy over time. Better prediction result were obtained for longer horizons. Based on these results, we decided to proceed with XGBoost based prediction considering a horizon of 30 days and included social factors as well, as input parameters. (4) Tree Depth: to control number of branch splits and reducing overfitting, searched over 4 – 30, best at 6 (5) Minimum leaves to split: controlling learning and under-fitting, searched over 1 – 50, best at 2. PFB initial tuning results, trained only on technical and price factors, with further improvements as discussed Section VI.

#	Sym	Learning Rate	L1	L2	Horizon	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error
1	BTC	0.3	0.14	0.4	60	0.1296	0.0243	0.1559
2	BTC	0.45	0.135	0.37	30	0.1225	0.0218	0.1478
3	BTC	0.3	0.13	0.45	10	0.2722	0.0888	0.2981
4	ETH	0.35	0.137	0.35	60	0.148	0.0285	0.1687
5	ETH	0.3	0.13	0.35	30	0.1693	0.0473	0.2174
6	ETH	0.37	0.135	0.45	10	0.2969	0.1148	0.3388
7	XRP	0.45	0.15	0.43	60	0.8252	0.793	0.793
8	XRP	0.37	0.14	0.35	30	0.6782	0.4924	0.7017
9	XRP	0.45	0.14	0.45	10	0.4776	0.37	0.6083

Clustering: To develop a generalized framework for prediction and understand crypto universe better we performed clustering on currency descriptive parameters and cross correlation coefficients, finding similarity between cryptocurrencies. However, after analysing results we determined that price based cross correlation is more important for prediction than other descriptive parameters. Therefore, we have used two separate techniques,(1) K-Modes clustering is used to segregate cryptocurrencies on shared descriptive attributes like "underlying technology", "key usage", "unique proposition" to understand a currency qualitatively. K-Means clustering is used for identifying currencies following similar price trends. Pearson correlation across all cryptocurrencies was determined and used as an attribute for clustering.

groups	Cluster Attributes
BTC,ETH,DOGE,LTC,BCH,XMR,ZEC,MIOTA,BSV	mineable 0.888889 pow 0.888889 medium-of-exchange 0.777778 bnb-chain 0.666667
USDT,USDC,BUSD,DAI,TUSD	stablecoin 1.0 bnb-chain 1.0 moonriver-ecosystem 0.8 arbitrum-ecosytem 0.8 asset-backed-stablecoin 0.8 avalanche-ecosystem 0.6
BNB,FTT,HBAR,OKB	marketplace 1.00 centralized-exchange 0.75
XRP,SHIB,TRX,HNT,FLOW,XEC,KSM,GMT,NEXO,AMP,USDP,AUDIO	0.166667 payments 0.166667 enterprise-solutions 0.166667 bnb-chain 0.166667 web3 0.166667 solana-ecosystem 0.166667
SOL,CELO,MINA	pos 1.000000 zero-knowledge-proofs 0.666667
ADA,ETC,ALGO,VET,ICP,WAVES,STX,ZIL,QNT,KDA,XEM,QTUM	smart-contracts 0.833333 platform 0.833333
LUNA,WBTC,CRO,RUNE,CVX,LRC,CRV,SCRT,ROSE,KNC,KAVA,ANC	defi 0.833333 cosmos-ecosystem 0.583333

VI. EXPERIMENTAL DESIGN & OBSERVATIONS

There were three primary objectives in our experimental design 1) To determine if social media parameters impact crypto currency price movement. 2) To determine if there is any similarity in

different cryptocurrency price trends 3) Identify key features and tune hyper parameters for prediction model to minimize prediction error.

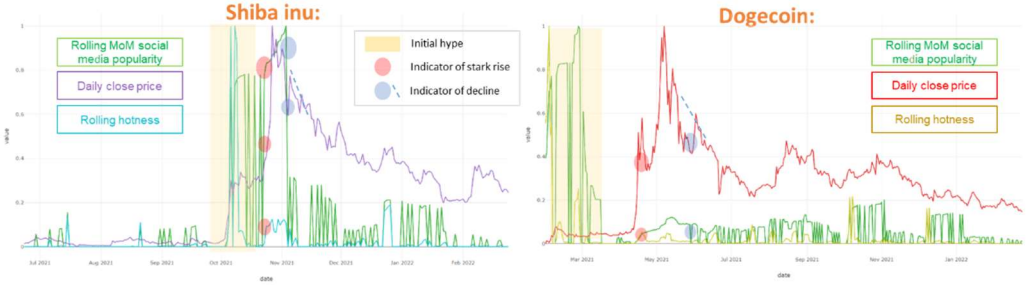


Fig 4. Day on Day price trend with social factors for (a) Shiba INU and (b) DOGECOIN, Depicting initial association of social media popularity and price Hype, Climb and Dive pattern

Impact of Social Media parameters: Meme/alt coins *(Fig. 4)* are highly social media driven, where the price and chatter move very closely. Initial price movements of such new coins could be identified from social listening. There are distinct phases observed in price variations. *Phase1: HYPE* – there is initial hype, indicated by abrupt spike in conversation hotness, monthly popularity and increase in price. *Phase 2: CLIMB* – following a fad, semi-gradual growth in price, monthly popularity and hotness. Followed a continuous price increase over a short interval of time. *Phase 3: DIVE* – sudden drop in social popularity and hotness, quickly followed by a price decline wedge, with subsequent peaks following a constant down trend.

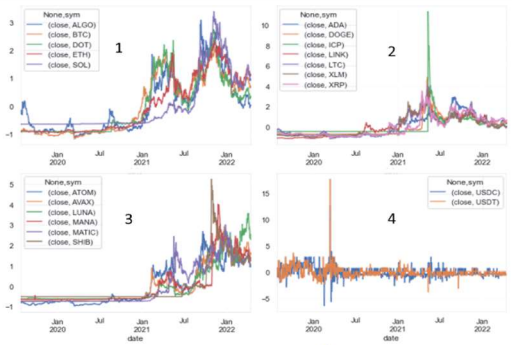
The effect observed in more stable coins like Ethereum (Fig. 5.) is *very different*. *Phase 1: HOPE* – avg weighted positive sentiment increases during a slow price grow or deterioration period, and following a trend decline. *Phase 2 : BEGINNING* however when investors' positive sentiment stops to decline or starts to increase along with price, it marks the initialization of short term high growth. *Phase 3: CLIMB* – the increase in price is generally steep, giving investors very short window to take decision, and is followed by a decline, possibly initialized by profit realization.

Fig. 5. Day on day price trend Ethereum with social factors

Based on these observations we established that there is a dependency of currency price on social factors however the extent and effect varies. We further tested this hypothesis by observing GMIC (Generalized Mean Information Coefficient) value of price with social factors and observed varying results. Coins like AVAX observed GMIC of 0.79 with Popularity but for BTC it was only 0.65, for XRP even lesser at 0.25. We therefore decided to make factors considered for the model dynamic and filtered based on GMIC value.

Symbol	Popularity	Distill Weighted Anger	Distil Joy Sum	Distil Love Mean	Distil Sadness Weighted	Distill Surprise Sum	Roberta Negative Weighted
AVAX	0.7916	0.8352	0.9391	0.8809	0.8348	0.9193	0.8442
ALGO	0.7188	0.7343	0.7099	0.7149	0.723	0.692	0.7244
SOL	0.7134	0.7136	0.7651	0.714	0.712	0.7148	0.7644
XLM	0.6694	0.6566	0.6888	0.6601	0.6607	0.6782	0.6661
BTC	0.6534	0.6507	0.4056	0.5789	0.5911	0.4316	0.6305
MANA	0.6121	0.5914	0.5876	0.5687	0.588	0.6063	0.589
SHIB	0.6096	0.7007	0.6597	0.6415	0.7	0.6599	0.7007
DOT	0.5664	0.5948	0.6244	0.6314	0.5963	0.6072	0.5937
ICP	0.5443	0.4989	0.5844	0.6704	0.5188	0.7526	0.5103
ATOM	0.5131	0.4356	0.3965	0.396	0.4495	0.3788	0.4326
LUNA	0.4451	0.4695	0.4874	0.4625	0.4576	0.4793	0.4733
DOGE	0.4033	0.3453	0.325	0.2405	0.4226	0.3349	0.3565
ADA	0.3021	0.3111	0.3665	0.3279	0.3133	0.3529	0.3104
MATIC	0.2793	0.274	0.2171	0.2146	0.2594	0.2372	0.2509
LINK	0.2669	0.2316	0.201	0.2137	0.2066	0.2094	0.2078
XRP	0.2506	0.2087	0.2264	0.2138	0.2009	0.2136	0.2162
ETH	0.2502	0.2923	0.2849	0.3233	0.2726	0.2918	0.2305
LTC	0.2374	0.2097	0.2714	0.2396	0.2027	0.2444	0.2158
USDC	0.16	0.1451	0.1509	0.1531	0.1405	0.1383	0.1489
USDT	0.1154	0.0982	0.0898	0.0957	0.1119	0.0971	0.1036

GMIC value of social factors determined by Distill and Roberta for 20 crypto currencies



Price Trend similarities: On performing K-means clustering using Pearson correlation, to validate for linear dependencies, for top 20 cryptocurrencies, following groups were identified (i) relatively older/popular coins like ALGO, BTC, DOGE, ETH and SOL. (ii) ADA, ICP, LNK, LTC, XLM, XRP (iii) ATOM, AVAX, LUNA, MANA, MATIC, SHIB (iv) stable coins like USDC, USDT. We inferred that the price peak and drop significantly varies for different currencies, resulting in change in viable period for trained and tested, significantly effecting prediction performance for different currencies. Therefore, a dynamic approach was designed to determine the viable Time period of training and testing, starting with 1st encounter of price mode than the "Mean of Overall Price / 2.3" for each coin. **Identifying key features and tuning hyper parameters for prediction model to minimize prediction error:** Initial results showed extreme predictions using social media metrics generated by Roberta, Distill-Bert and Vader directly as parameter, due to fluctuations in conversations and

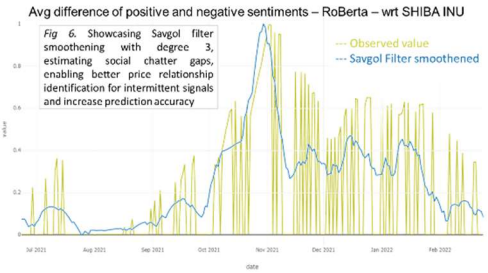


Fig 6. Showcasing Savgol filter smoothening with degree 3, estimating social chatter gaps, enabling better price relationship identification for intermittent signals and increase prediction accuracy

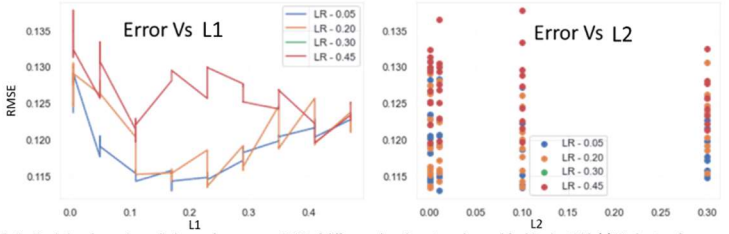


Fig 7. : Depicting change in prediction performance as RMSE of difference learning rates, observed for bitcoin. With (a) L1 : best performance achieved at L1 between 0.1~0.2, with lower Learning rates . (b) L2 best with lower LRs, with slightly better results at lower L2.

popularity. *(fig 6.)* We Applied Savgol filter smoothening to reduce high fluctuations in social data and determine a trend, yielding better performance. The Significance of factors was observed to be different for different cryptocurrencies, hence we added factor filtration based on GMIC score, removing low importance factors for model predictions, with GMIC lower than 0.75 with 30 days ahead coin price. However, some currencies like LINK, SHIB and DOGE associated with very few variables above the threshold. To answer this, in case of less than 5 parameters at 0.75 threshold, a dynamic reduction of 0.02 in the threshold was introduced until atleast 5 most significant factors were identified. After building the initial module for factor selection we employed Grid Search for hyper parameter tuning, determining the Optimal values for Learning rate, Ridge and Lasso regularization coefficients, Max Depth and Minimum split leaves by using Root Mean Squared Error as the Objective *(fig 7.)*. Initially tuning with cross validation of 3 but later moving it to 5 and reducing mean error by ~10%. Post optimization, for 30 day prediction, we achieved an MSE of 0.005 for relatively less socially driven BTC at a Learning rate of 0.05, L1 - 0.17, L2 - 0.01 and for more socially driven coins like AVAX an MSE of 0.08 was achieved by setting Learning rate to 0.03, L1 to 0.21 and L2 to 0.001 (*fig. 8.*).

Fig 8. : Test prediction for AVAX and BTC during latest 5% of the significant coin lifetime before 26th March 2022. Model was observed to identify trend significantly well, while faced minor challenges with possible small future fluctuations. Deeming better for long term investment.

MODEL PERFORMANCE TABLE

sym	learning_rate	reg_alpha	reg_lambda	mse	mae	rmse
BTC	0.05	0.17	0.01	0.00531	0.0512	0.0729
XRP	0.45	0.05	0.1	0.0159	0.114	0.1261
SOL	0.45	0.29	0.1	0.06954	0.2508	0.2637
ADA	0.45	0.005	0.01	0.12702	0.283	0.3564
LTC	0.2	0.21	0.01	0.0289	0.1622	0.17
XLM	0.45	0.45	0.1	0.04562	0.2062	0.2136
USDT	0.2	0.05	0.01	0.00049	0.0176	0.0221
USDC	0.03	0.29	0.001	0.00051	0.0194	0.0225
ETH	0.2	0.41	0.3	0.00224	0.0365	0.0474
ALGO	0.45	0.05	0.1	0.14701	0.3557	0.3834
ICP	0.03	0.005	0.0001	2.25E-05	0.0041	0.0047
LINK	0.45	0.45	0.01	0.00937	0.0696	0.0968
MATIC	0.2	0.37	0.01	0.00776	0.0745	0.0881
SHIB	0.2	0.005	0.1	0.00015	0.011	0.0122
LUNA	0.03	0.45	0.01	0.00417	0.0546	0.0646
MANA	0.2	0.45	0.001	0.00317	0.0485	0.0563
ATOM	0.2	0.13	0.0001	0.02323	0.1036	0.1524
DOGE	0.45	0.37	0.01	0.00893	0.0447	0.0945
DOT	0.2	0.05	0.1	0.00466	0.0576	0.0682
AVAX	0.03	0.21	0.0001	0.00686	0.0751	0.0828

Model achieved less than 0.01 RMSE for 13 out of 20 cryptocurrencies.

Assessing Scalability & User Acceptance Testing: Scalability was assessed by varying A) Number of Cryptocurrencies: Parameter tuning and prediction for top 20 currencies currently covered in our visualization took ~5 hours to run. For executing this script for ~100 currencies on coinmarketcap data it took a little over 28 hours. B) Reddit data timeframe: We were able to extract 3 years of post data from Reddit. Overall it took ~10 hours to download data and ~10 hours to extract sentiment factors from this data. Including these factors in our XGBoost prediction framework reduced error by ~30% where effects varied from currency to currency. **Acceptance Testing**: 8/10 people felt that they can now make a more informed investment decision. 7/10 people felt that they were able to discover more insights about the currency of their interest.

VII. Conclusion

Our experiments show that social listening has significant importance in predicting future values for some currencies at different maturities and multiple currencies share similarities in term of popularity and sentiment association and in term of feature importance. Additionally, our observations from topic associations, show that significant amount of knowledge can be gained succinctly, making connections on topics like blockchain, NFT, scams etc. with a high potential to add value awareness about ongoing/changing driving forces, including their current importance to the world. This effort brings in a solution to better understand the global investor profiles and their changing behaviour. We believe this approach can also be extrapolated to other diversified fields to understand consumer importance and their needs. Our next step is to identify and quantify the impact of not just technology associated topics but also human association including social, economic and political factors to create a more robust and generalized approach to help investors better understand the crypto world. We also aim to scale the framework to utilize critic and influencer information from journals, newsletters, articles and active media advertisements, showcasing and quantifying the impact of multiple dimensions.

VIZ URL : https://team93-group-project.herokuapp.com/

All team members have contributed a similar amount of effort.

VIII. References

[1] Yukun Liu, Aleh Tsyvinski, Risks and Returns of Cryptocurrency, The Review of Financial Studies, Volume 34, Issue 6, June 2021, Pages 2689–2727, https://doi.org/10.1093/rfs/hhaa113

[2] Giudici, G., Milne, A. & Vinogradov, D. Cryptocurrencies: market analysis and perspectives. J. Ind. Bus. Econ. 47, 1–18 (2020). https://doi.org/10.1007/s40812-019-00138-6

[3] Khan, Ruby & Hakami, Tahani. (2021). Cryptocurrency: usability perspective versus volatility threat. Journal of Money and Business. ahead-of-print. 10.1108/JMB-11-2021-0051.

[4] Chua, Eunice & Rustico, Ed Mark. (2018). ACCEPTABILITY OF INVESTING IN CRYPTOCURRENCIES.

[5] Tsao, Shu-Feng & Chen, Helen & Tisseverasinghe, Therese & Yang, Rena & Li, Lianghua & Butt, Zahid. (2021). What social media told us in the time of COVID-19: a scoping review. The Lancet Digital Health. 3. 10.1016/S2589-7500(20)30315-0.

[6] FENG MAI , ZHE SHAN, QING BAI, XIN (SHANE) WANG, ROGER H.L. CHIANG: How Does Social Media Impact Bitcoin Value? A Test of the Silent Majority Hypothesis,2018, Journal of Management Information Systems 35(1):19-52

[7] Poongodi M. , Tu N. Nguyen , Mounir Hamdi , Korhan Cengiz: Global cryptocurrency trend prediction using social media, 2021, Information Processing and Management 58(6):102708

[8] Stephan A.Curiskis, BarryDrakeThomas, R.OsbornPaul, J.Kennedy: An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit,2020, Information Processing & Management

[9] Anagha R Kulkarni, Vrinda Tokekar, Parag Kulkarni: Identifying Context of Text Documents using Naïve Bayes Classification and Apriori Association Rule Mining, 2012 CSI Sixth International Conference on Software Engineering (CONSEG)

[10] Hany Mahgoub, Dietmar Rösner, Nabil Ismail and Fawzy Torkey: A Text Mining Technique Using Association Rules Extraction, World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:2, No:6, 2008

[11] Van Minh Hao, Nguyen Huynh Huy, Bob Dao, Thanh-Tan Mai, Khuong Nguyen-An: Predicting Cryptocurrency Price Movements Based on Social Media, 2019 International Conference on Advanced Computing and Applications (ACOMP)

[12] Lars Steinert, Christian Herff: Predicting altcoin returns using social media,2018, PLoS ONE 13(12): e0208119

[13] Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, Iqbal H. Sarker: SentiLSTM: A Deep Learning Approach for Sentiment Analysis of Restaurant Reviews, 20th International Conference on Hybrid Intelligent Systems (HIS 2020)

[14] Erik Tromp, Mykola Pechenizkiy: Rule-based Emotion Detection on Social Media: Putting Tweets on Plutchik's Wheel, arXiv, 1412.4682 [cs.CL], 2014

[15] AHMED ZAHIM DAHHAM DAHHAM, Abdullahi Abdu Ibrahim: EFFECTS OF VOLATILITY AND TREND INDICATOR FOR IMPROVING PRICE PREDICTION OF CRYPTOCURRENCY, 2020, IOP Conference Series Materials Science and Engineering 928(3):032043

[16] Jing-Zhi Huang, William Huang, Jun Ni: Predicting bitcoin returns using high dimensional technical indicators, 2018, The Journal of Finance and Data Science 5(3)

[17] G. Vidyulatha , M. Mounika , N. Arpitha: Crypto Currency Prediction Model using ARIMA, Turkish Journal of Computer and Mathematics Education Vol.11 No.03 (2020), 1654-1660

[18] Rama K. Malladi, Prakash L. Dheeriya: Time Series Analysis of Cryptocurrency returns and volatilities, Journal of Economics and Finance (2021) 45:75-94

[19] Vasily Derbentsev, Natalia Datsenko, Olg Stepanenko, Vitaly Bezkorovainyi: Forecasting cryptocurrency prices time series using Machine Learning,2019, SHS Web of Conferences 65(1):02001

[20] Ioannis E. Livieris , Niki Kiriakidou , Stavros Stavroyiannis and Panagiotis Pintelas: An Advanced CNN-LSTM Model for Cryptocurrency Forecasting, Electronics 2021, 10, 287.

[21] Ahmed.F. Ibrahim , Liam Corrigan, Rasha Kashef: Predicting the Demand in Bitcoin Using Data Charts: A Convolutional Neural Networks Prediction Model, 2020 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)

[22] Kris Shaffer, With Twitter's poor signal-to-noise ratio, should social academia look to less corporate and more localised networks?, The London School of economics and political sciences, 2014, 11, 13

[23] Priya, Shalini & Sequeira, Ryan. (2018). Where should one get news updates: Twitter or Reddit. Online Social Networks and Media. 9. 17-29. 10.1016/j.osnem.2018.11.001.

[24] Hutto, C.J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[25] Sanh, Victor & Debut, Lysandre & Chaumond, Julien & Wolf, Thomas. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing - NeurIPS 2019

[26] Barbieri, Francesco & Camacho-Collados, José & Espinosa-Anke, Luis & Neves, Leonardo. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. 1644-1650. 10.18653/v1/2020.findings-emnlp.148.

[27] Loureiro, Daniel & Barbieri, Francesco & Neves, Leonardo & Espinosa-Anke, Luis & Camacho-Collados, José. (2022). TimeLMs: Diachronic Language Models from Twitter.