# Machine Learning For Sentiment Analysis

Mandeep Ahuja
BrainStation

# Can machine learning be applied to public discourse on social media to extract actionable insights?

Assuming public sentiment correlates with market trends, can I use machine learning to decide if I should buy or sell Bitcoin?

# DATA

- Collection
  - Kaggle dataset of over 22 million tweets with #Bitcoin

- Description
  - DateTime, Username, Tweet

- Pre-processing
  - Missing values
  - Tweets that sound like advertisements by defining ‘noise words’
  - Tweets without ‘#bitcoin’
  - Eliminating unwanted characters such as #s, URLs, whitespace & special characters
  - Duplicate tweets
  - Tweets with new year wishes & birthday wishes
  - Datetime column into year and month
  - StopWords library without the word ‘not’
  - Lemmatization; example ‘doing’, ‘does’, ‘do’  become ‘do’
  - Tokenizing into individual words
  - Random sample of 1 million tweets was retrieved as our dataset

# DATA

| | year | month | day | time | datetime | username | tweet |
|---|---|---|---|---|---|---|---|
| 0 | 2022 | 6 | 22 | 13:20:08 | 2022-06-22 13:20:08-04:00 | dasrecord | @allen_drewe @financebrah2 @newzealandhodl @florida_btc @ln_prints @awaitedsavior @ropeium @theguyswann @downwarddave @davidgshort @rangershodl @based_fyodor @federalistfiles @notephesians @mr_vril_maxxed @bitcoin_fuckboi @andhans_jail @janetystackx @joshmanmode_ you mean your fiat is slowly dying. |
| 1 | 2021 | 7 | 4 | 13:35:49 | 2021-07-04 13:35:49-04:00 | CryptoEugene101 | @themoonboyz \n\n$qnt “the image below explains \n what quant is all about!”\n\n“expect a parabolic breakout!”\n\n“$qnt is all around us & the majority don’t even realise/ know it yet, but soon they will!”\n\n@quant_network @quantoverledger \n@ripple @bitcoin https://t.co/nqqfe8fw6c |
| 2 | 2021 | 9 | 8 | 13:05:41 | 2021-09-08 13:05:41-04:00 | BotcoinUSD | on the 09/08/2021 at 05:05 1btc was worth $46377.00 #bitcoin #crypto #botcoinusd #bitfinex |
| 3 | 2021 | 4 | 23 | 05:49:48 | 2021-04-23 05:49:48-04:00 | emylacapra | @shanstory the best position is to love both when it goes up or down😉\nit's never enough of bitcoin bought, don't you think? |
| 4 | 2021 | 5 | 3 | 00:14:42 | 2021-05-03 00:14:42-04:00 | x_rhodium | @xrpology you’ll have to forgive me because it’s a fight to the bitter end. the evolution and the revolution continue on side-by-side. \nquestion on your preference:\nphysical gold\nbitcoin \nfiat\nor xrp\nwhich would give you the most comfort sleeping at night |

| | year | month | day | time | datetime | username | tweet | cleaned_tweet |
|---|---|---|---|---|---|---|---|---|
| 0 | 2022 | 6 | 22 | 13:20:08 | 2022-06-22 13:20:08-04:00 | dasrecord | @allen_drewe @financebrah2 @newzealandhodl @florida_btc @ln_prints @awaitedsavior @ropeium @theguyswann @downwarddave @davidgshort @rangershodl @based_fyodor @federalistfiles @notephesians @mr_vril_maxxed @bitcoin_fuckboi @andhans_jail @janetystackx @joshmanmode_ you mean your fiat is slowly dying. | drewe btc print fyodor vril maxxed fuckboi jail mean fiat slowly die |
| 1 | 2021 | 7 | 4 | 13:35:49 | 2021-07-04 13:35:49-04:00 | CryptoEugene101 | @themoonboyz \n\n$qnt “the image below explains \n what quant is all about!”\n\n“expect a parabolic breakout!”\n\n“$qnt is all around us & the majority don’t even realise/ know it yet, but soon they will!”\n\n@quant_network @quantoverledger \n@ripple @bitcoin https://t.co/nqqfe8fw6c | image explain quant expect parabolic breakout around us amp majority not even realise know yet soon network |
| 2 | 2021 | 9 | 8 | 13:05:41 | 2021-09-08 13:05:41-04:00 | BotcoinUSD | on the 09/08/2021 at 05:05 1btc was worth $46377.00 #bitcoin #crypto #botcoinusd #bitfinex | btc worth bitcoin crypto botcoinusd bitfinex |
| 3 | 2021 | 4 | 23 | 05:49:48 | 2021-04-23 05:49:48-04:00 | emylacapra | @shanstory the best position is to love both when it goes up or down😉\nit's never enough of bitcoin bought, don't you think? | best position love go never enough bitcoin buy not think |
| 4 | 2021 | 5 | 3 | 00:14:42 | 2021-05-03 00:14:42-04:00 | x_rhodium | @xrpology you’ll have to forgive me because it’s a fight to the bitter end. the evolution and the revolution continue on side-by-side. \nquestion on your preference:\nphysical gold\nbitcoin \nfiat\nor xrp\nwhich would give you the most comfort sleeping at night | forgive fight bitter end evolution revolution continue side side question preference physical gold bitcoin fiat xrp would give comfort sleep night |

# Word2Vec Model

- Trained model size - over 96 million words with over 300 million word-word pairs
- Similarity Scores
- Topic Modeling

```
Similarity score between 'bitcoin' and 'ethereum': 0.3686
Similarity score between 'bitcoin' and 'crypto': 0.4681
Similarity score between 'bitcoin' and 'blockchain': 0.2206
Similarity score between 'ethereum' and 'crypto': 0.6041
Similarity score between 'ethereum' and 'blockchain': 0.4814
Similarity score between 'crypto' and 'blockchain': 0.4952
```

# Word2Vec - Topic Modelling

Topic 1: General Bitcoin Perception and Usage: This topic seems to revolve around general attitudes and perceptions towards Bitcoin. Words like "not," "people," "like," "money," "buy," "would," "use," and "know" suggest discussions about the practicality, understanding, and public sentiment regarding Bitcoin. It indicates conversations about whether or not to use Bitcoin, its value as money, and public knowledge or skepticism about it.

Topic 2: Bitcoin and Related Cryptocurrencies: This topic focuses on Bitcoin in the context of the broader cryptocurrency market. It includes terms like "btc" (Bitcoin), "crypto," "ethereum," "cryptocurrency," "eth" (Ethereum), "blockchain," "nft" (non-fungible tokens), "usd" (U.S. Dollar), and "binance" (a cryptocurrency exchange). This suggests discussions about Bitcoin's position in the crypto market, its comparison with other cryptocurrencies, and its relation to broader market platforms and technologies.

Topic 3: Market Dynamics and Investment Strategy: This topic appears to concentrate on market trends and investment strategies involving Bitcoin. Words like "go," "buy," "btc," "time," "market," "see," "get," "year," and "next" imply conversations about buying Bitcoin, market timing, future predictions, and overall investment strategies.

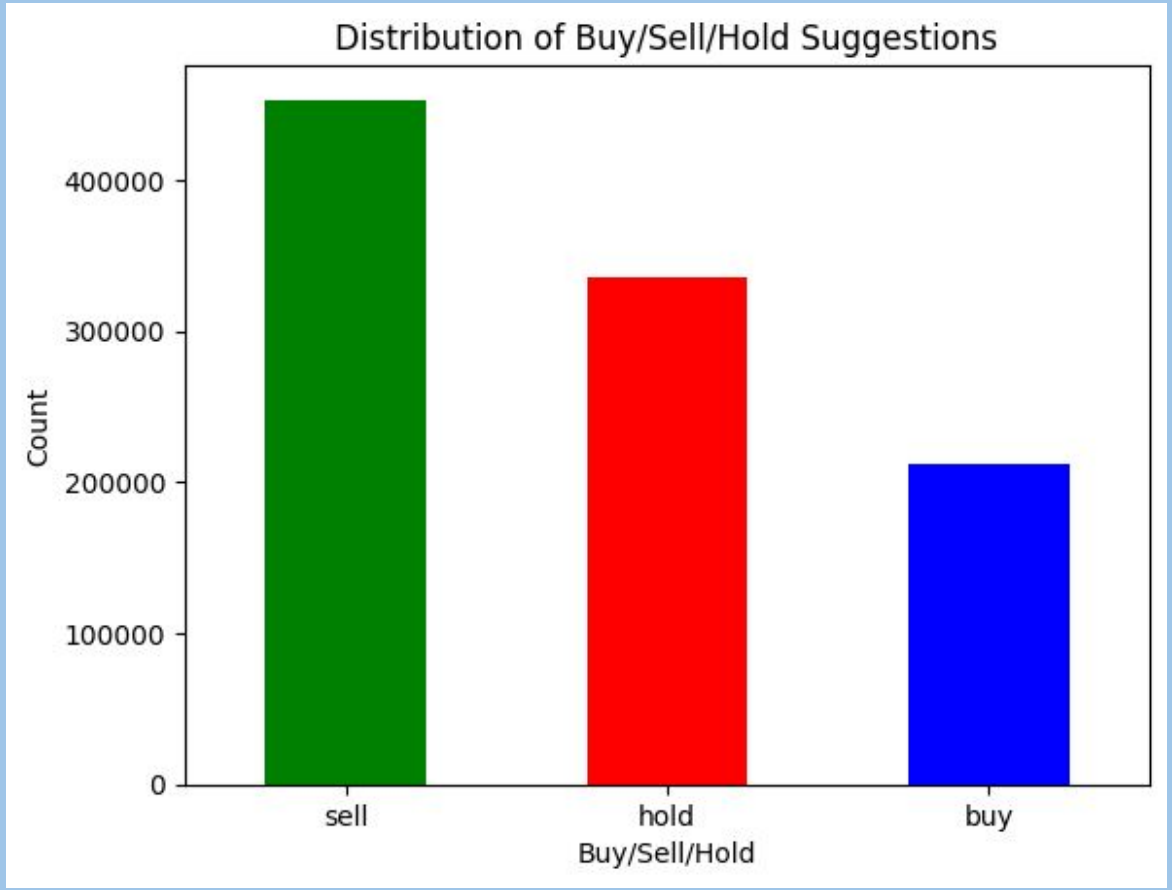
Topic 4: Cryptocurrency Projects and Community Involvement: This topic is likely related to specific cryptocurrency projects and community activities. Terms such as "project," "crypto," "cryptocurrency," "block," "airdrop," "great," "join," "bsc" (Binance Smart Chain), and "good" suggest discussions about various crypto-related projects, community initiatives like airdrops (free distribution of new tokens), and involvement in blockchain-based platforms.

Topic 5: News and Updates in the Crypto World: This topic seems to focus on news, updates, and mining in the cryptocurrency world. Words like "crypto," "news," "via," "mine," "market," "new," "cryptocurrency," "amp" (possibly short for 'and'), and "elon" (potentially referring to Elon Musk, a notable figure in tech and crypto) suggest discussions centered around the latest news, mining activities, market updates, and influential figures in the crypto space.

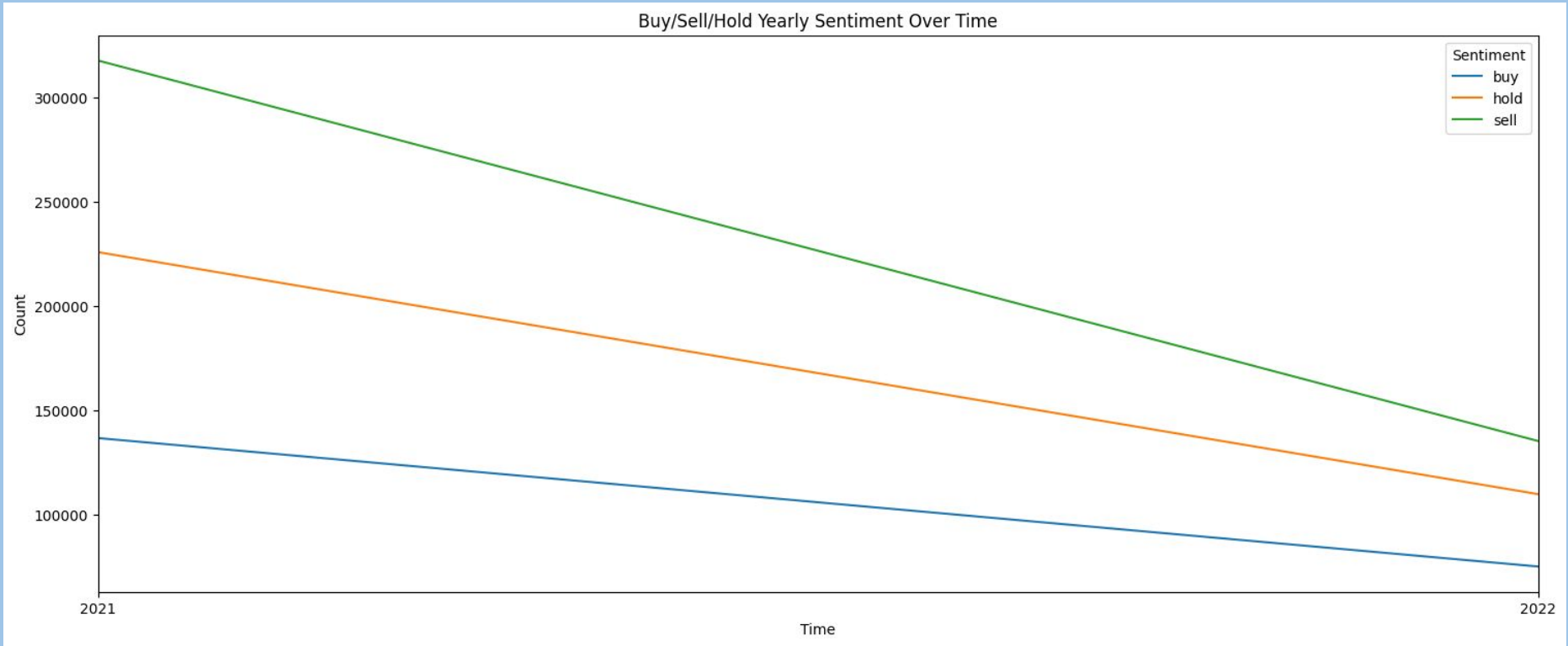
# VADER + Keywords

| | year | month | day | time | datetime | username | tweet | cleaned_tweet | sentiment | category |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2021 | 6 | 30 | 12:53:23 | 2021-06-30 12:53:23-04:00 | StckPro | $ms new article : morgan stanley just purchased a huge amount of grayscale bitcoin trust https://t.co/p8vzmp6sqb get all the latest $ms related news here : https://t.co/hu36emjo1v https://t.co/qyir5l2ujs | new article morgan stanley purchase huge amount grayscale bitcoin trust get latest relate news | 0.6808 | buy |
| 1 | 2021 | 8 | 12 | 03:00:47 | 2021-08-12 03:00:47-04:00 | bitcoinagile | primexbt makes cov staking easier with direct cov token purchase press release #bitcoin news #btc #crypto https://t.co/mw2igpz2fs https://t.co/aqpx7asnat | primexbt make cov stake easier direct cov token purchase press release bitcoin news btc crypto | 0.4215 | buy |
| 2 | 2021 | 2 | 25 | 15:50:11 | 2021-02-25 15:50:11-05:00 | eurushaga | grow your asset more than your cost capital. that's why companies are flipping their balance sheet to bitcoin and the shareholders are leveraging this strategy. #bitcoin #moneyhasmoved | grow asset cost capital company flip balance sheet bitcoin shareholders leverage strategy bitcoin moneyhasmoved | 0.3612 | neutral |
| 3 | 2021 | 4 | 17 | 23:51:59 | 2021-04-17 23:51:59-04:00 | PicoDeGallo1122 | #doge #hodl #dontsell #dogecoin #dogecoinrise #dogecointothemoon dont sell guys this isnt bitcoin this is doge | doge hodl dontsell dogecoin dogecoinrise dogecointothemoon not sell guy not bitcoin doge | 0.0000 | sell |
| 4 | 2021 | 10 | 1 | 18:37:38 | 2021-10-01 18:37:38-04:00 | rizalrvc_ | italian luxury fashion house dolce & gabbana sells nft collection for $5.7 million – blockchain bitcoin news https://t.co/unyvso2e6x | italian luxury fashion house dolce amp gabbana sell nft collection million blockchain bitcoin news | 0.0000 | sell |

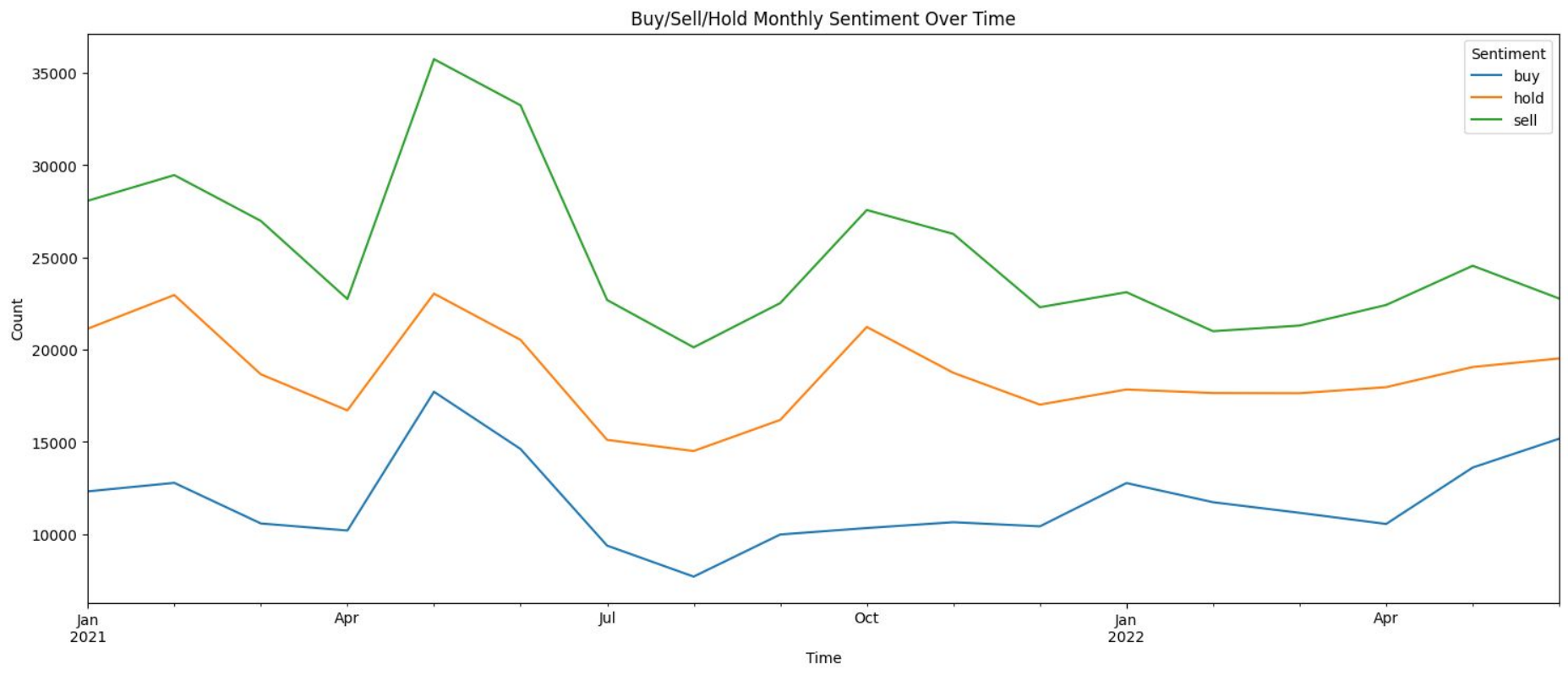
# Sentiment Distribution - Tweets



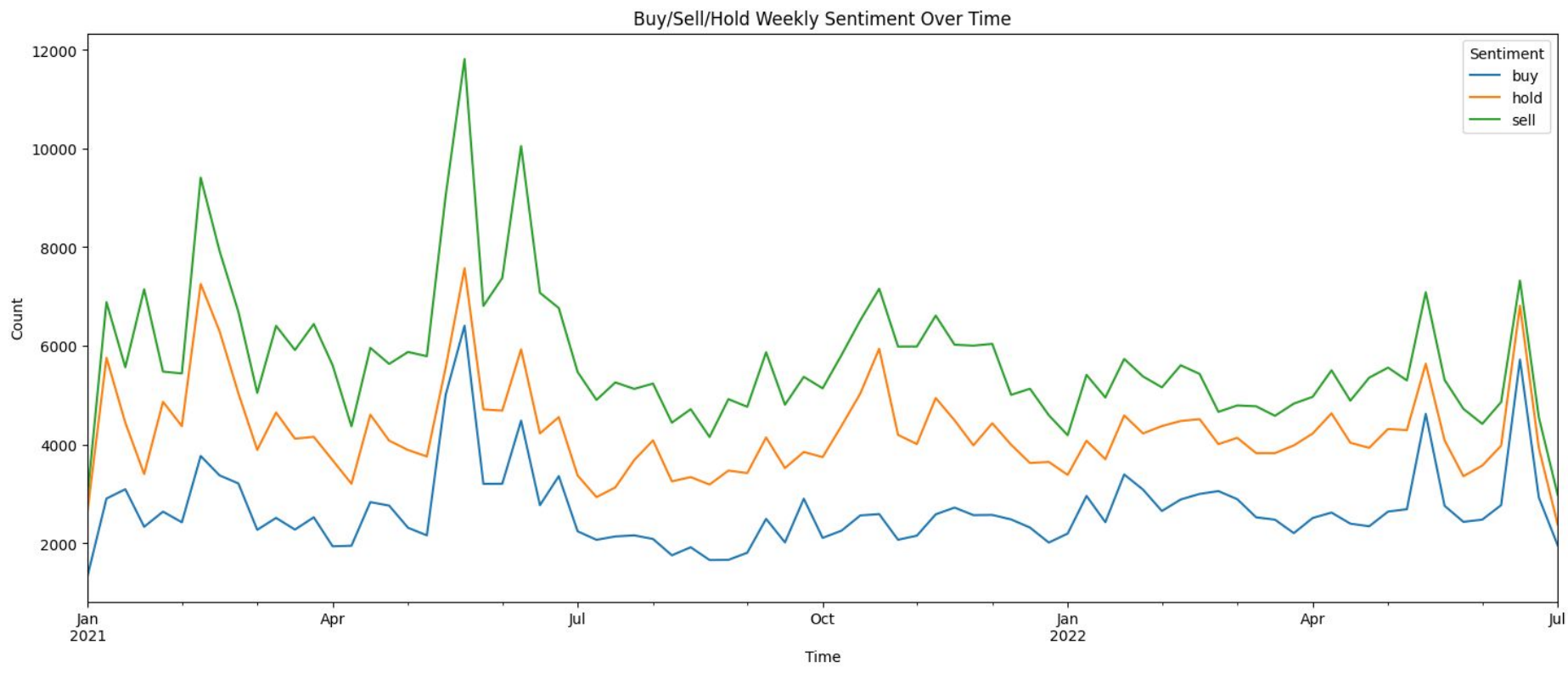
# Buy/Sell/Hold Yearly Sentiment



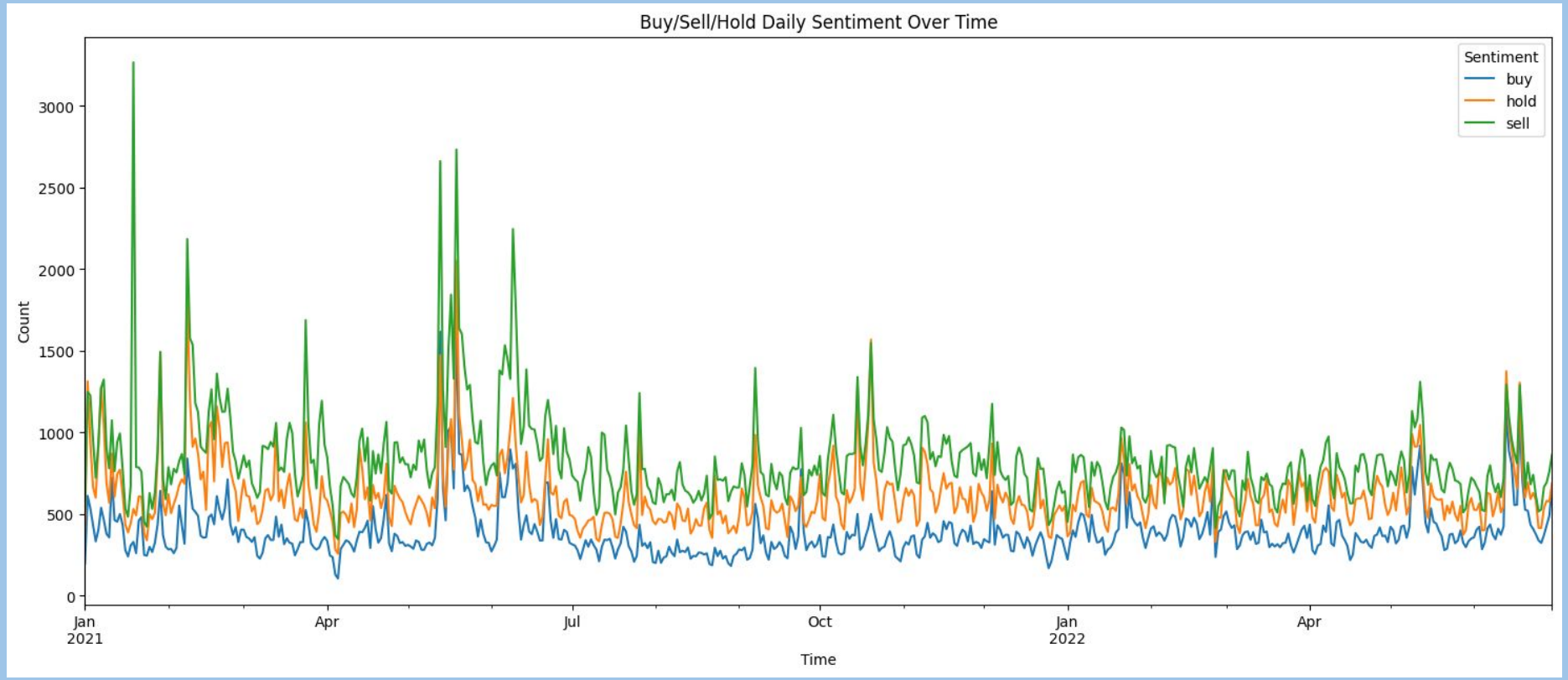
# Buy/Sell/Hold Monthly Sentiment



# Buy/Sell/Hold Weekly Sentiment



# Buy/Sell/Hold Daily Sentiment



# Bitcoin Chart - U.S.Dollar

# Word2VEC & VADER**

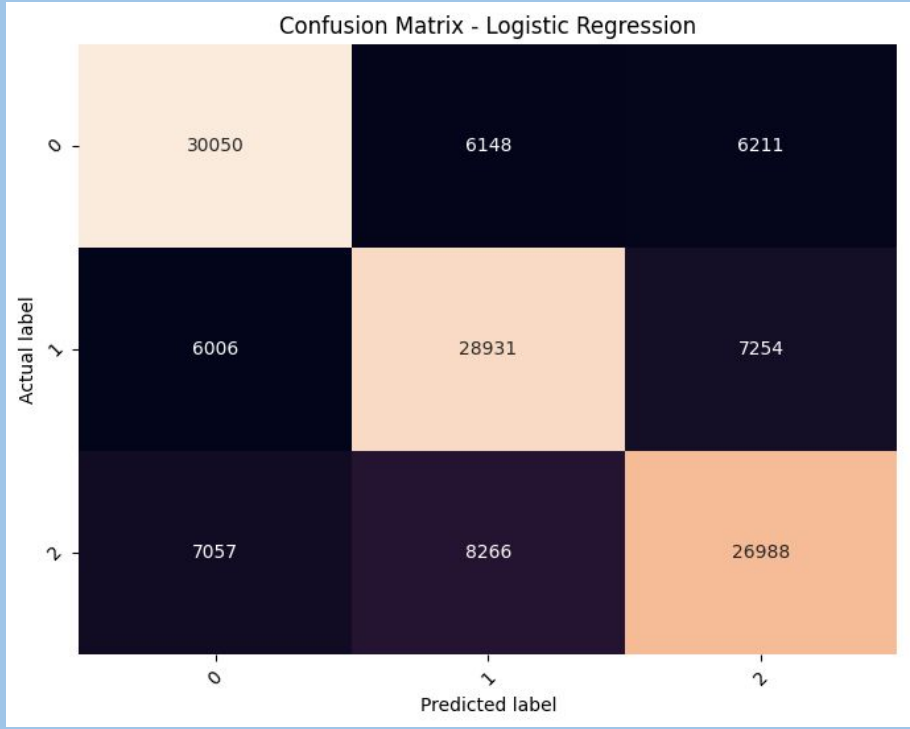
| | year | month | day | time | datetime | username | tweet | cleaned_tweet | sentiment | category | category_mapped | tweet_vectors |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2021 | 8 | 3 | 03:11:20 | 2021-08-03 | BrieflyBitcoin | bank of america sees benefit in adopting #bitc... | bank america see benefit adopt bitcoin legal t... | 0.9186 | sell | 2 | [-0.31757963, 0.10027568, -0.18486142, 0.27022... |
| 1 | 2022 | 5 | 2 | 14:54:32 | 2022-05-02 | cryptoredline | @mikealfred since 2016 i have bitcoin and ethe... | since bitcoin ethereum wait day bitcoin go evo... | 0.3612 | sell | 2 | [-0.10168886, 0.4265832, 0.15867133, -0.229852... |
| 2 | 2022 | 6 | 26 | 20:50:11 | 2022-06-26 | WTFutures | #etc is creating new resistance zones after d... | etc create new resistance zone destroy hour ho... | -0.3400 | buy | 1 | [0.29176408, 0.23502524, 0.27072465, -0.278395... |
| 3 | 2022 | 1 | 10 | 00:02:55 | 2022-01-10 | ConnNFT | who's buying #bitcoin ? is this the dip to be ... | buy bitcoin dip buy go lower eth btc cryptocur... | -0.2960 | buy | 1 | [-0.00104879, 0.30820906, 0.2549957, -0.111654... |
| 4 | 2021 | 3 | 10 | 01:14:06 | 2021-03-10 | JimBob30238733 | @ryanfeepoker ryan how do u buy bitcoin ? i tr... | ryan buy bitcoin try coinbase customer support... | 0.4215 | sell | 2 | [-0.18221024, 0.17337537, -0.13183504, -0.2366... |

**Disclaimer: Not an industry practice to rely on machine generated labels as ground truth. For learning purposes only.

# Word2VEC & VADER**
# Logistic Regression

Test Accuracy: 0.6773959704044566
Test Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.70 | 0.71 | 0.70 | 42409 |
| 1 | 0.67 | 0.69 | 0.68 | 42191 |
| 2 | 0.67 | 0.64 | 0.65 | 42311 |
| accuracy | | | 0.68 | 126911 |
| macro avg | 0.68 | 0.68 | 0.68 | 126911 |
| weighted avg | 0.68 | 0.68 | 0.68 | 126911 |




**Disclaimer: Not an industry practice to rely on machine generated labels as ground truth. For learning purposes only.

- Pretraining on TweetEval by Cardiff University, UK

- TweetEval
  - Train
  - Test
  - Validation

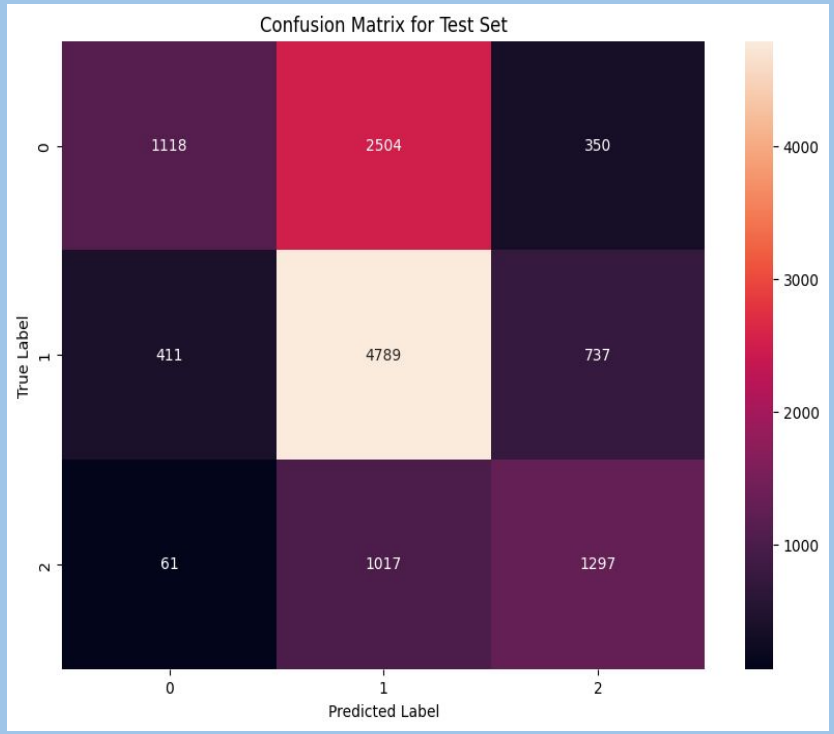
- Models trained
  - Logistic Regression
  - VADER
  - Random Forest
  - Perceptron
  - RoBERTa

# Another method: Model Training

# Logistic Regression

Logistic Regression Classification Report for Test Set:

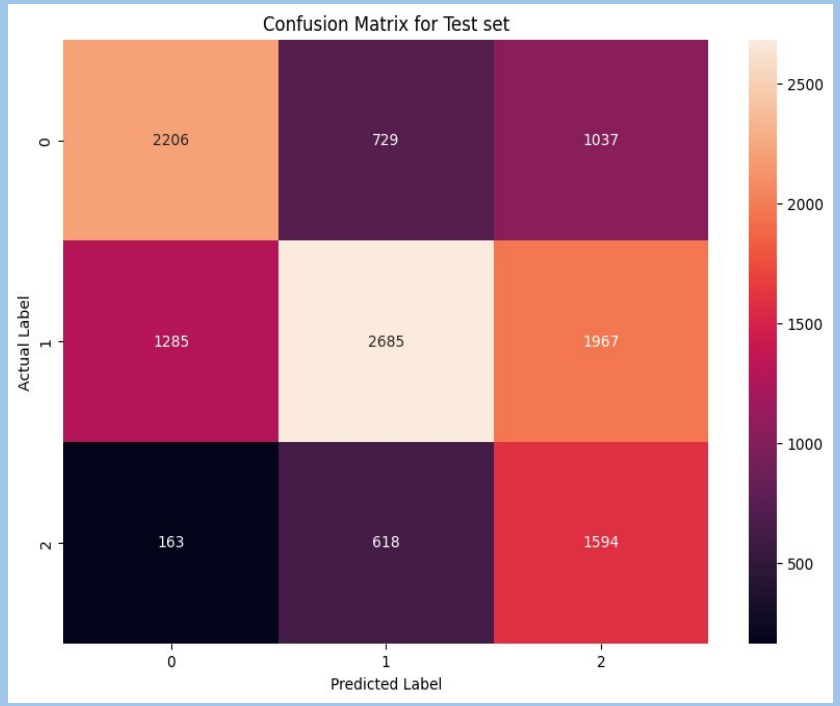
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.70 | 0.28 | 0.40 | 3972 |
| 1 | 0.58 | 0.81 | 0.67 | 5937 |
| 2 | 0.54 | 0.55 | 0.55 | 2375 |
| accuracy | | | 0.59 | 12284 |
| macro avg | 0.61 | 0.54 | 0.54 | 12284 |
| weighted avg | 0.61 | 0.59 | 0.56 | 12284 |



# VADER Lexicon

VADER Classification Report Test set

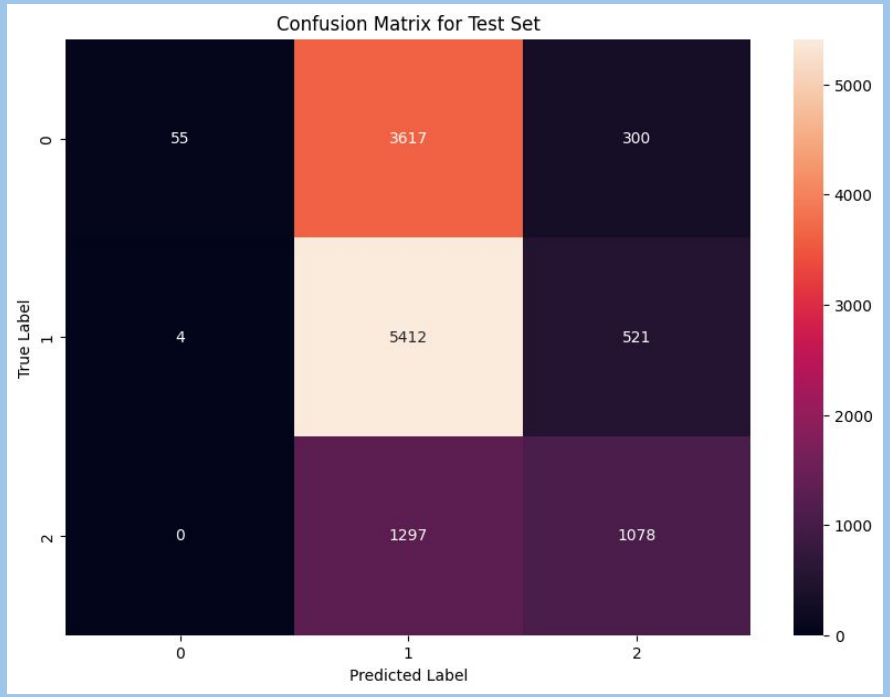
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.60 | 0.56 | 0.58 | 3972 |
| 1 | 0.67 | 0.45 | 0.54 | 5937 |
| 2 | 0.35 | 0.67 | 0.46 | 2375 |
| accuracy | | | 0.53 | 12284 |
| macro avg | 0.54 | 0.56 | 0.52 | 12284 |
| weighted avg | 0.58 | 0.53 | 0.54 | 12284 |



# Random Forest

Random Forest Classification Report for Test Set:

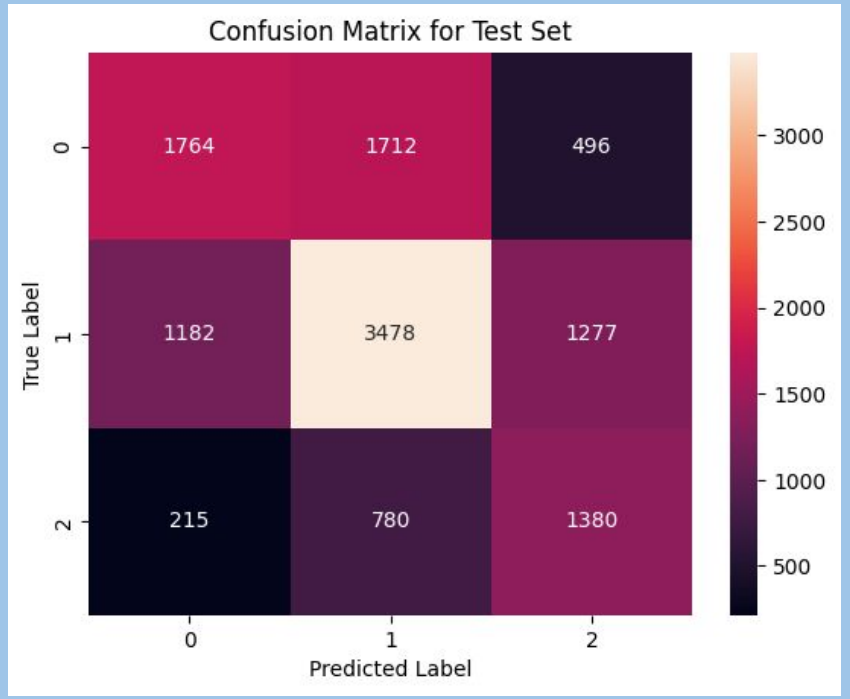
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.93 | 0.01 | 0.03 | 3972 |
| 1 | 0.52 | 0.91 | 0.67 | 5937 |
| 2 | 0.57 | 0.45 | 0.50 | 2375 |
| accuracy | | | 0.53 | 12284 |
| macro avg | 0.67 | 0.46 | 0.40 | 12284 |
| weighted avg | 0.66 | 0.53 | 0.43 | 12284 |



# Perceptron

Perceptron Classification Report for Test Set:

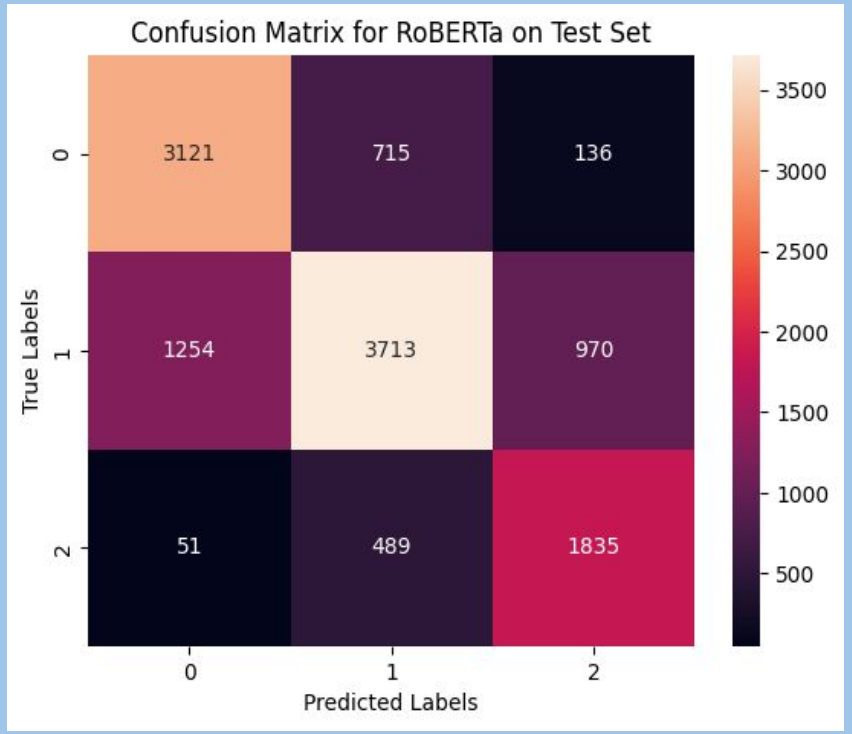
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.56 | 0.44 | 0.49 | 3972 |
| 1 | 0.58 | 0.59 | 0.58 | 5937 |
| 2 | 0.44 | 0.58 | 0.50 | 2375 |
| accuracy | | | 0.54 | 12284 |
| macro avg | 0.53 | 0.54 | 0.53 | 12284 |
| weighted avg | 0.55 | 0.54 | 0.54 | 12284 |



# RoBERTa

RoBERTa Classification Report for Test Set:

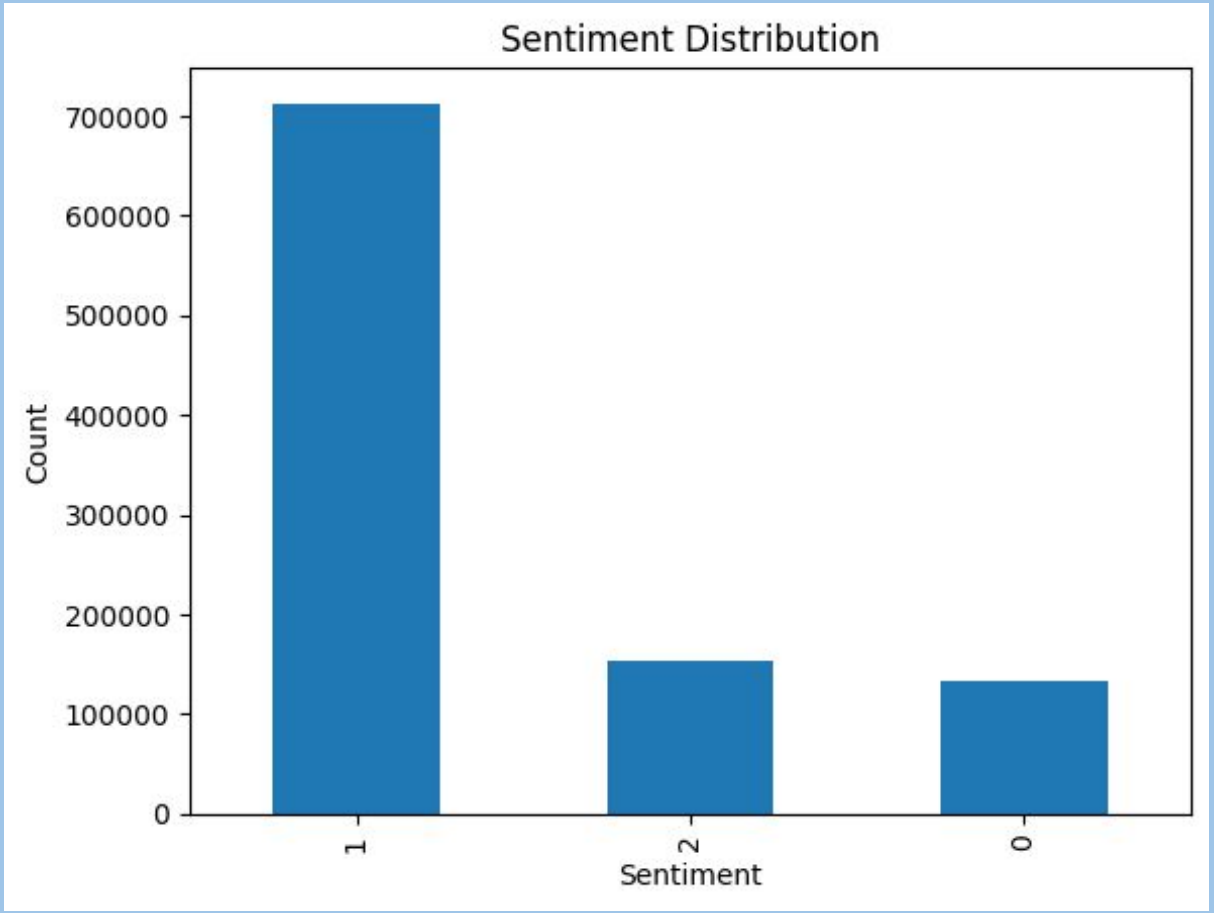
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.71 | 0.79 | 0.74 | 3972 |
| 1 | 0.76 | 0.63 | 0.68 | 5937 |
| 2 | 0.62 | 0.77 | 0.69 | 2375 |
| accuracy | | | 0.71 | 12284 |
| macro avg | 0.69 | 0.73 | 0.71 | 12284 |
| weighted avg | 0.71 | 0.71 | 0.70 | 12284 |



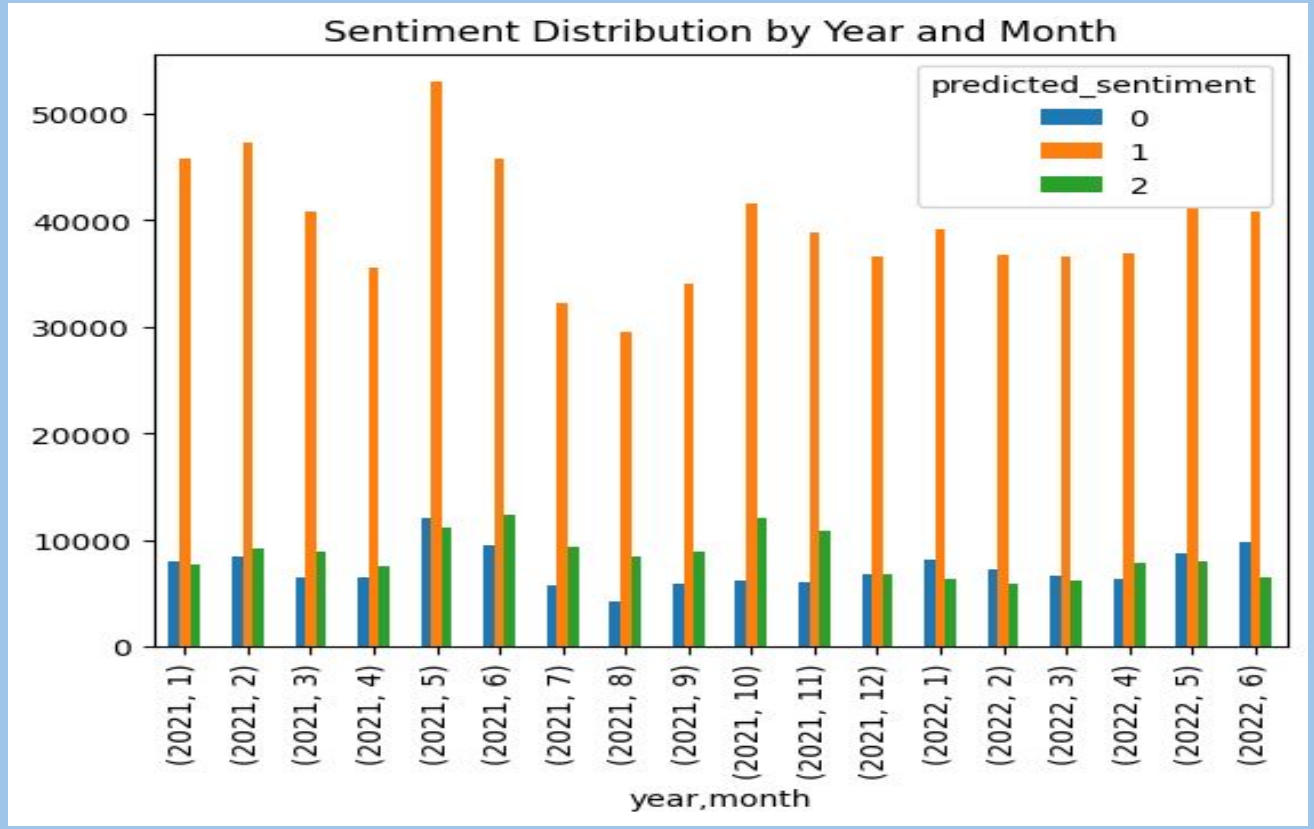
# RoBERTa On #Bitcoin

| | year | month | username | tweet | cleaned_tweet | predicted_sentiment |
|---|---|---|---|---|---|---|
| **50** | 2022 | 2 | norriegche07 | tim dillon goes off on the bitcoin conference ... | tim dillon go bitcoin conference handle heckle... | 1 |
| **51** | 2021 | 10 | aexsegg | @bitcoin_monk @olya_borderless @cryptonawaz at... | monk borderless like tiny fraction compare unv... | 0 |
| **52** | 2021 | 2 | MarioMartReq | @edugaresp el bitcoin ahora mismo a 53k @remin... | el bitcoin ahora mismo ofthis three years | 1 |
| **53** | 2021 | 1 | BlockWatcher | sun jan 31 05:28:35 2021 (3:43)\nusd : 33,991.... | sun jan usd wght blk size txs pool mb bitcoin | 1 |
| **54** | 2022 | 6 | InfinityTokenIO | rewards are available for claim!\n\nmined #bit... | reward available claim mine bitcoin gt ethereu... | 1 |
| **55** | 2021 | 8 | DefencePirc | @empty_banks i honestly want bitcoin to be in ... | bank honestly want bitcoin weeks accumulate | 1 |
| **56** | 2021 | 11 | UAPVee | @ruralindia @sonaliranade very poor article. s... | poor article idea talk seem think bitcoin cryp... | 0 |
| **57** | 2021 | 5 | sunrise_guide | https://t.co/kcfjtdqubo #industryevents #bitco... | industryevents bitcoin alexgladstein interview... | 1 |
| **58** | 2022 | 3 | johny_0722 | #bitcoin at 9:05am and then 9:07am. https://t.... | bitcoin | 1 |
| **59** | 2021 | 1 | hrizek | 30k before 2021? #bitcoin @novogratz ? | bitcoin | 1 |
| **60** | 2021 | 3 | CryptoSquawk | 🥿 increased liquidity 🦕 41 $btc traded so far ... | increase liquidity trade far bitcoin crypto btc | 1 |
| **61** | 2021 | 5 | SasiPallempati | @deitaone probably to buy more #bitcoin | probably buy bitcoin | 1 |
| **62** | 2021 | 5 | jpd_1 | i like the new look. now i just want my #bitco... | like new look want bitcoin reward card | 2 |
| **63** | 2021 | 5 | InTheLightsGlow | each bitcoin has an address which is a randoml... | bitcoin address randomly generate public key p... | 1 |
| **64** | 2021 | 5 | ymingc918 | @rick_bitcoin me too! | bitcoin | 1 |
| **65** | 2022 | 3 | btcjerk | @paultang chainanalysis is the weapon against ... | chainanalysis weapon criminal money effective ... | 0 |

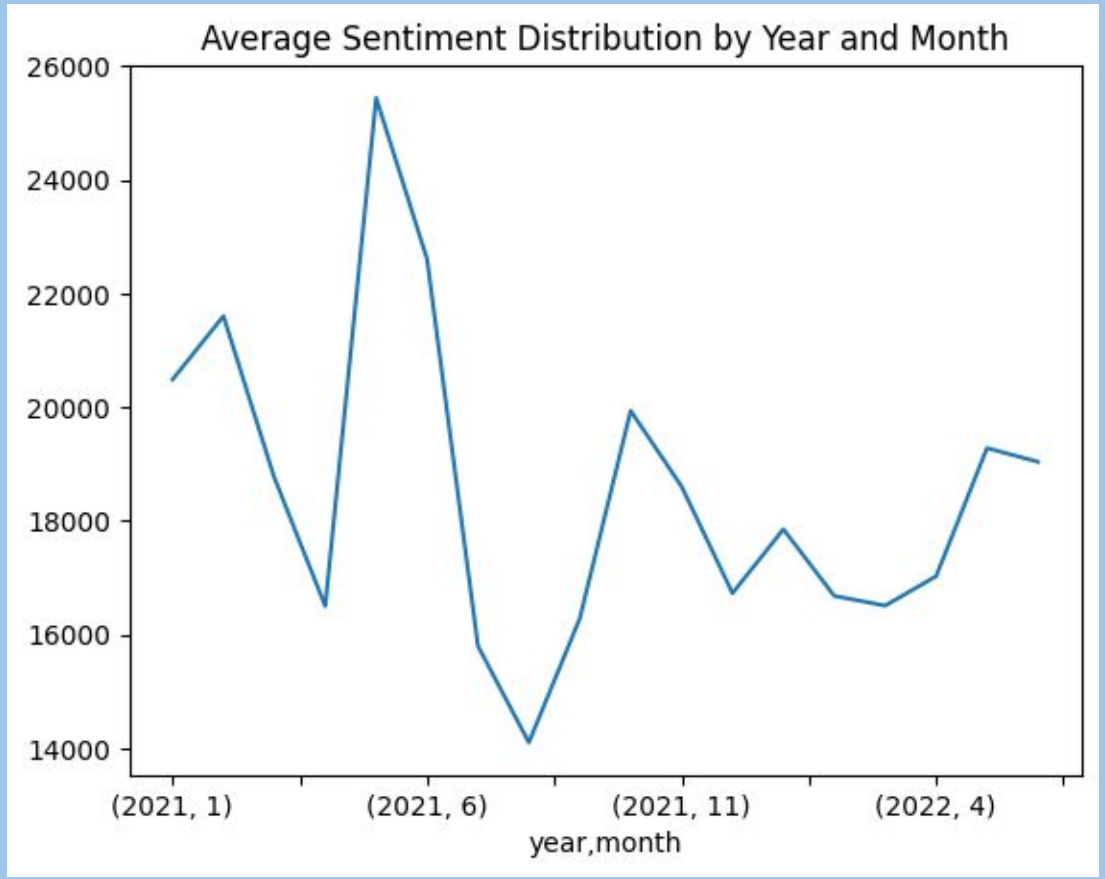
# #Bitcoin Sentiment Distribution



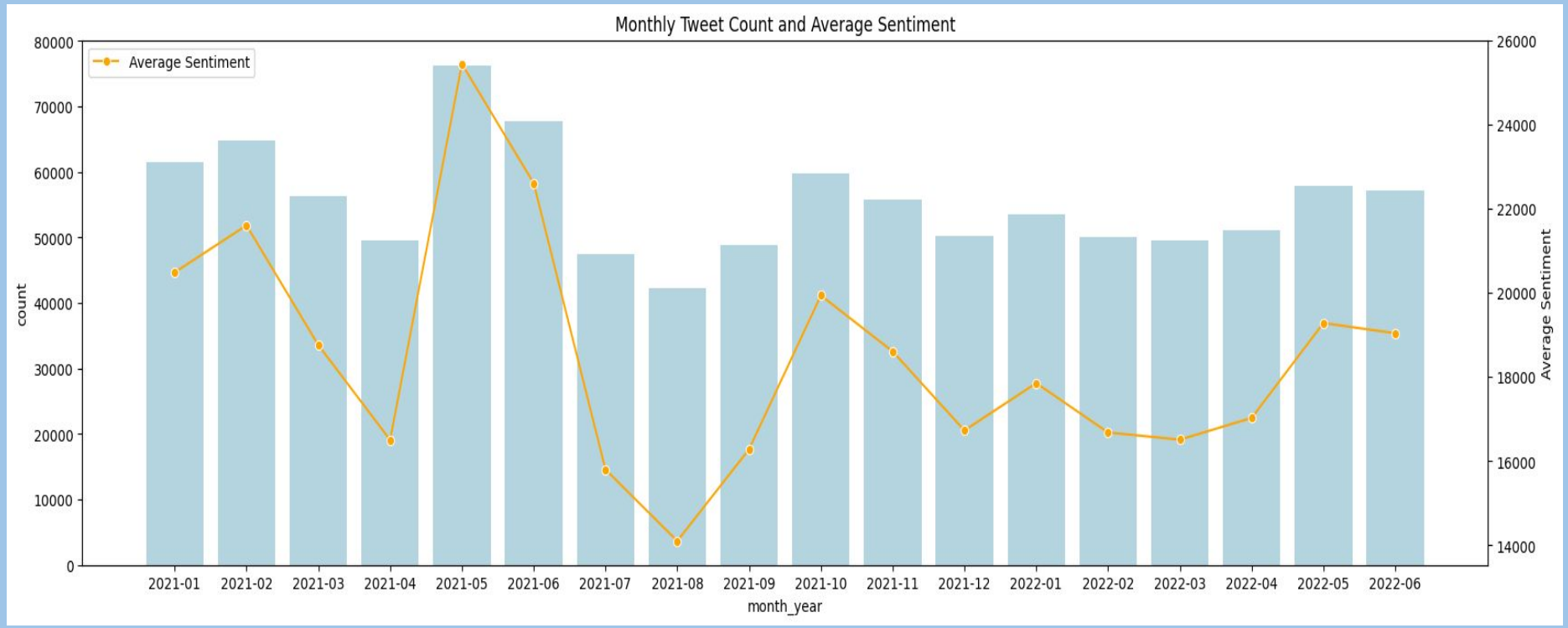
# #Bitcoin Sentiment Distribution By Year-Month



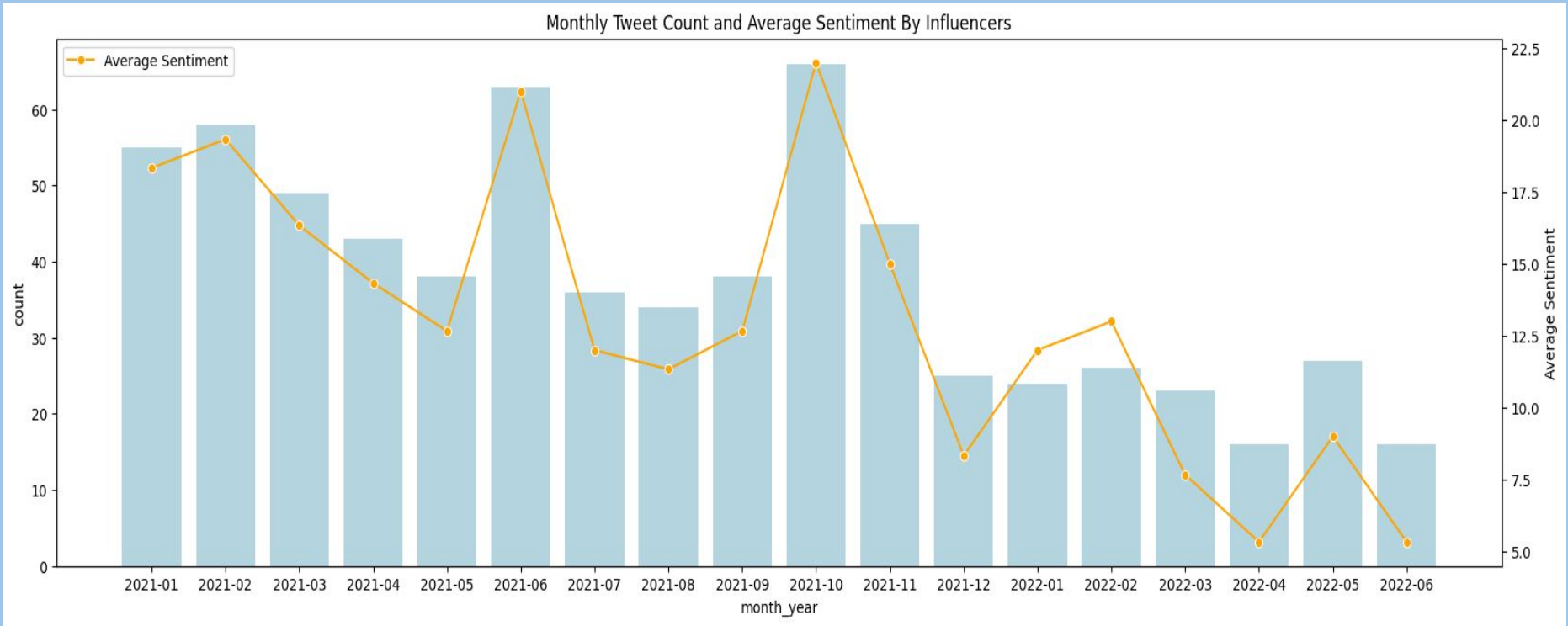
# #Bitcoin Average Sentiment Distribution By Year-Month



# #Bitcoin Monthly Tweet Count & Average Sentiment



# #Bitcoin Monthly Tweet Count & Average Sentiment - Influencers



# US Market Bitcoin Chart 2021-2022 - Tradingview

# #Bitcoin Data Quality

```
relevant_terms = ['market', 'buy', 'sell', 'trend', 'bullish', 'liquidation',
'sell-off', 'pump and dump', 'consolidate', 'capital gain', 'buy back','buyback',
'sell off','trade', 'exchange', 'risk', 'margin', 'analysis', 'bear market', 'bull,
'market', 'bubble', 'correction', 'death cross']
```

```
Corpus Size: ~1 million
```

```
Count of tweets with relevant terms: 252660
```

Only ~25% of this dataset has important market related terminology.

# Future Work

- Combine the power of Word2Vec vectors with TweetEval pre-training
- Improve model accuracy

- Data collection can be done with more market related terms

- Can this be extended to other areas?
  - Brand image
  - Public opinion of products / services
  - Customer service
  - Policy making

- SaaS product
  - Chose social media platform
  - Extract content : keywords, #hastags, @usernames, timeline
  - Preprocess
  - Analyse

# Thank You!