Syracuse University
College of Engineering
& Computer Science

Cryptocurrency Projections from Tweet Sentiment Analysis

Nicholas Mohan, Nicholas Landry,
Isaiah Plummer, Nathan Fenske

Introduction

- Through tweets, we are predicting the next hours price change for three different cryptocurrencies; Bitcoin, Dogecoin, and Ethereum
- Utilizing the Twitter search api, as well as the Coinbase price api, we are able to create a training set for a Naive Bayes Classifier
- The hope is to give the user an idea based on the current circumstances if the next hour these crypto prices will increase or decrease

Collecting Tweets

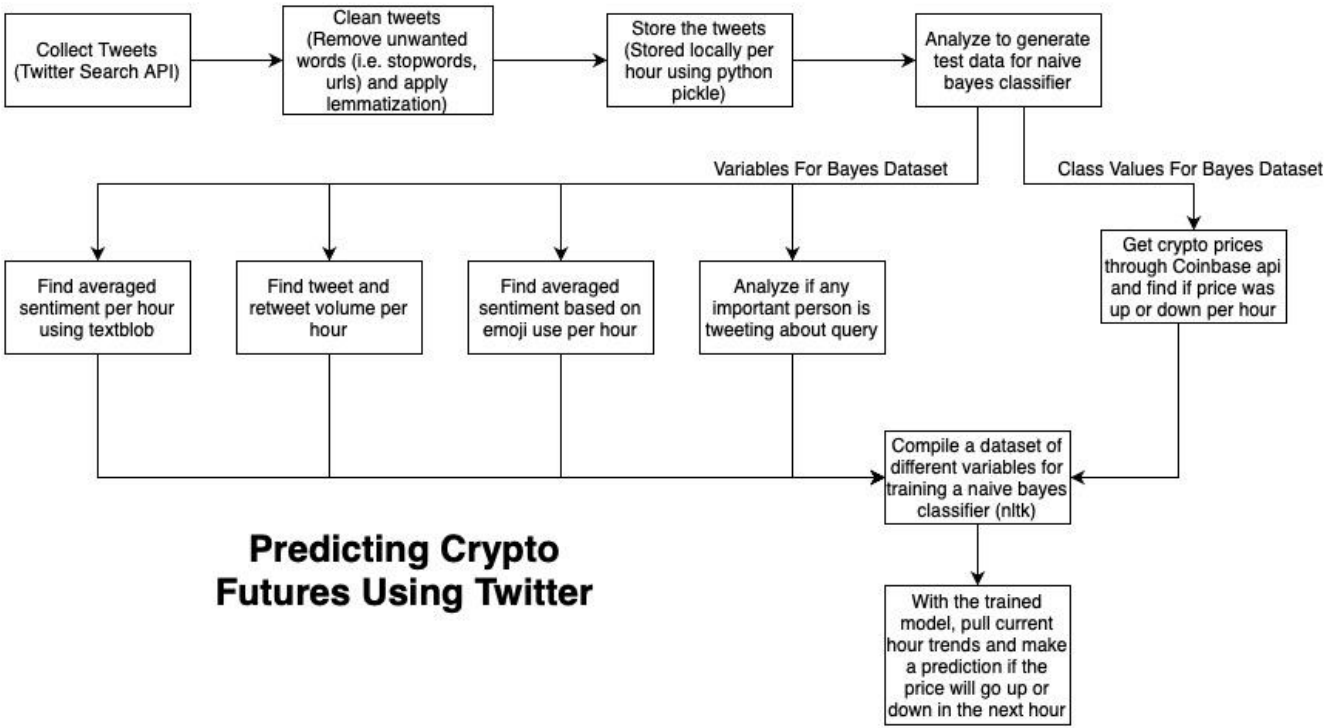
- Utilizing the twitter search API to receive all tweet data for the previous seven days
- This module is run through a shell script that receives in a list of dates, as well as a query, and automatically collects all tweets for those days that match the query
- This data is separated by each hour of each day, and stored as python pickle objects to external files
- Other data, such as tweet volume per hour as well as retweets are counted as they are used later for the classifier.

Removing Noise

- Start off by tokenizing the tweets, which separates individual words into strings which are stored in a list
- Filter out stop words and punctuation
  - Unique characters, emojis, words, @’s and #’s remained
- Filter out tokens under 2 characters
- Checked if its a token that we want to keep
  - If token is an @
  - If token is a #
  - If token is an emoji using emoji.UNICODE_EMOJI_ENGLISH
  - If a word using nltk.corpus.words
- Autocorrected leftover tokens and checked if they were actual words
- Filter named entities

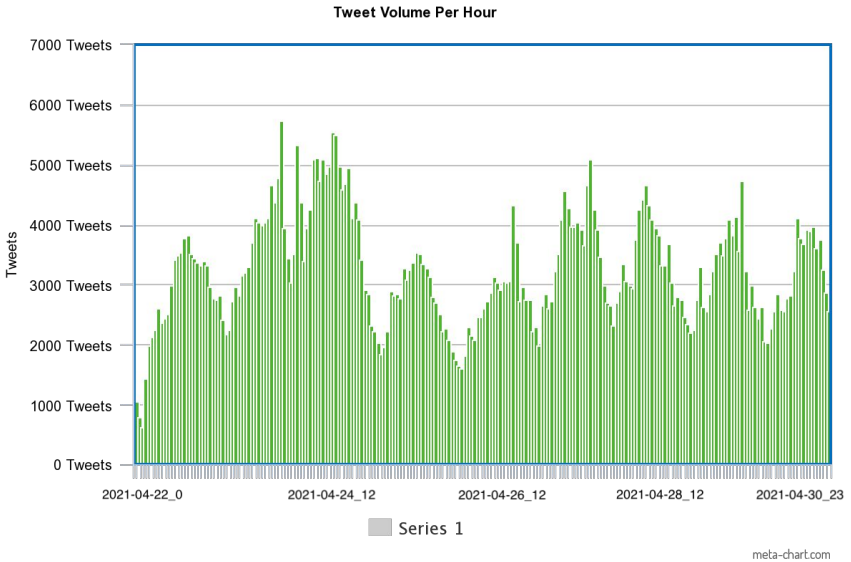
#hashtag

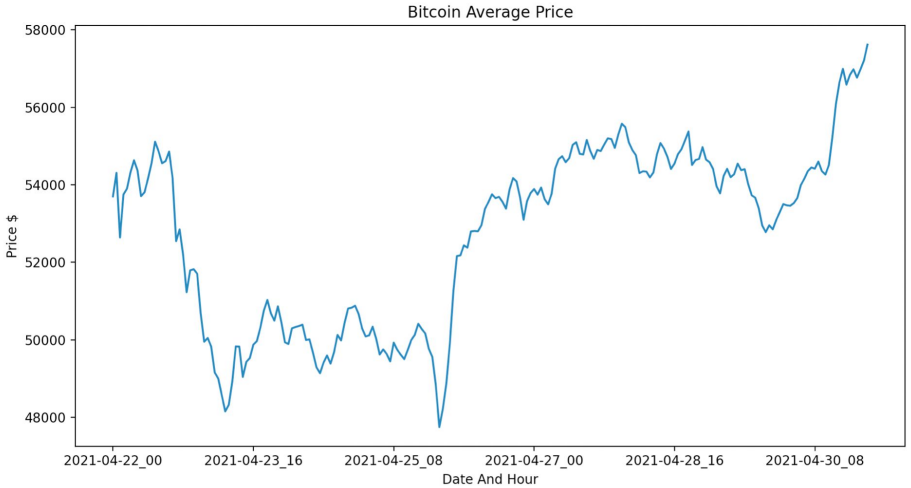
Flow Chart


Collect Tweets (Twitter Search API)
Clean tweets (Remove unwanted words (i.e. stopwords, urls) and apply lemmatization)
Store the tweets (Stored locally per hour using python pickle)
Analyze to generate test data for naive bayes classifier
Variables For Bayes Dataset
Class Values For Bayes Dataset
Find averaged sentiment per hour using textblob
Find tweet and retweet volume per hour
Find averaged sentiment based on emoji use per hour
Analyze if any important person is tweeting about query
Get crypto prices through Coinbase api and find if price was up or down per hour
Compile a dataset of different variables for training a naive bayes classifier (nltk)
With the trained model, pull current hour trends and make a prediction if the price will go up or down in the next hour
Predicting Crypto Futures Using Twitter

Creating A Prediction

- Why not just use twitter sentiment? What other factors should we consider?





Bayes V1

```
Added to Training Set. Ready For Next Hour.

Sampling training set...

Training...

Accuracy:
0.5562
Most Informative Features
         contains(Black) = True              rise : fall   =     10.8 : 1.0
          contains(Hunt) = True              fall : rise   =      9.0 : 1.0
 contains(coinhuntworld) = True              fall : rise   =      9.0 : 1.0
         contains(vault) = True              fall : rise   =      9.0 : 1.0
          contains(Coin) = True              fall : rise   =      8.5 : 1.0
            contains(BN) = True              fall : rise   =      7.5 : 1.0
   contains(environment) = True              rise : fall   =      7.3 : 1.0
           contains(sir) = True              rise : fall   =      6.5 : 1.0
       contains(success) = True              rise : fall   =      6.5 : 1.0
       contains(BitNorm) = True              fall : rise   =      6.5 : 1.0
           contains(may) = True              fall : rise   =      6.5 : 1.0
          contains(Blue) = True              fall : rise   =      5.9 : 1.0
         contains(banks) = True              fall : rise   =      5.9 : 1.0
      contains(@spinte8) = True              rise : fall   =      5.6 : 1.0
         contains(Click) = True              rise : fall   =      5.6 : 1.0
            contains(MB) = True              rise : fall   =      5.6 : 1.0
          contains(Pool) = True              rise : fall   =      5.6 : 1.0
           contains(car) = True              rise : fall   =      5.6 : 1.0
         contains(subiu) = True              rise : fall   =      5.6 : 1.0
           contains(van) = True              rise : fall   =      5.6 : 1.0
            contains(□) = True               rise : fall   =      5.6 : 1.0
        contains(milyon) = True              rise : fall   =      5.4 : 1.0
             contains(¡) = True              rise : fall   =      5.4 : 1.0
       contains(nothing) = True              fall : rise   =      5.4 : 1.0
 contains(@WClementeIII) = True              fall : rise   =      4.9 : 1.0
```

Process:

- Creates a variable representing the presence of each of the 2000 most popular tokens from a set date
  - Tokenization removes punctuation, stopwords, and named entities [named entities got added in later don't worry (:]
- Compiles training set from a different set date, collecting values for each of the 2000 token features
- Predicts rise/fall with a defined threshold of > +/- .5%

Accuracy: 57% w/ training, test size 5k

Problems:

- No lemmatization, incomplete tokenization (non-english words, urls, etc.)
- Token features are not enough to encapsulate the environment which would cause a rise/fall
- Team decided to try out more categories

Bayes V2

```
Training...

Accuracy:
0.6196
Most Informative Features
      contains(cashback) = True              C1 : C3     =     44.3 : 1.0
         contains(Bulls) = True              C1 : C3     =     31.7 : 1.0
         contains(cover) = True              C2 : C3     =     19.4 : 1.0
       contains(indices) = True              C2 : C3     =     19.4 : 1.0
contains(@ICOAnnouncement) = True              C1 : C3     =     19.0 : 1.0
contains(@bitcoin_paris) = True              C1 : C3     =     19.0 : 1.0
          contains(Hash) = True              C1 : C3     =     19.0 : 1.0
         contains(Hours) = True              C1 : C3     =     19.0 : 1.0
          contains(Look) = True              C1 : C3     =     19.0 : 1.0
       contains(Looking) = True              C1 : C3     =     19.0 : 1.0
    contains(Prediction) = True              C1 : C3     =     19.0 : 1.0
```

Prediction:

- Category of price change C1 … C5 in the current hour
  - C1: >1%, C2: >.5%, C3: [.5%, -.5%], C4: <-.5%, C5: <-1%

New Features:

- Category price change C1 … C5 for previous hour
- Tweet volume/current hour (0+,5k+,10k+)
- Retweets/tweet (0+,10+,50+)
- Trading volume/previous hour (0+,500+,1k+)

Accuracy: 61% w/ training, test size 5k

Bayes V3

```
Training...

Accuracy:
0.6483
Most Informative Features
contains(@Bitcoin_K_S_A) = True              C4 : C3     =     20.4 : 1.0
         contains(token) = True              C1 : C4     =     20.2 : 1.0
           contains(due) = True              C1 : C3     =     19.7 : 1.0
           contains(mum) = True              C1 : C3     =     19.7 : 1.0
contains(@dogecoin_rise) = True              C1 : C3     =     19.7 : 1.0
       contains(Capital) = True              C1 : C3     =     19.7 : 1.0
     contains(Dogecoins) = True              C1 : C3     =     19.7 : 1.0
           contains(GET) = True              C1 : C3     =     19.7 : 1.0
           contains(Got) = True              C1 : C3     =     19.7 : 1.0
       contains(However) = True              C1 : C3     =     19.7 : 1.0
           contains(Tax) = True              C1 : C3     =     19.7 : 1.0
         contains(alone) = True              C1 : C3     =     19.7 : 1.0
        contains(closer) = True              C1 : C3     =     19.7 : 1.0
        contains(effect) = True              C1 : C3     =     19.7 : 1.0
         contains(event) = True              C1 : C3     =     19.7 : 1.0
          contains(grip) = True              C1 : C3     =     19.7 : 1.0
         contains(guess) = True              C1 : C3     =     19.7 : 1.0
          contains(mass) = True              C1 : C3     =     19.7 : 1.0
         contains(pizza) = True              C1 : C3     =     19.7 : 1.0
          contains(pour) = True              C1 : C3     =     19.7 : 1.0
       contains(respect) = True              C1 : C3     =     19.7 : 1.0
```

Modifications:

- Better tokenization, lemmatization included

New Features:

- Contains emoji, contains repeated emoji (T/F)
- Influencer tagged, special influencer tagged (T/F)

Problems:

- New features not weighted with respect to token variables
- Could stand to benefit from an actual SA metric
- New tokenization runs quite slow

Accuracy: 64% w/ training, test size 5k

Results

Classifying strictly negative/positive price changes by hour, we've been able to achieve a maximum accuracy of 71% using our current build.

Classifying 5 separate price changes, we've been able to achieve a maximum accuracy of 64% with our current build. Better than a coin toss, but this is still not satisfactory to act as a predictor for hopeful investors.

Ultimately, we still increased our accuracy build by build, indicating that tweets can act as a reliable source of prediction for crypto price changes. This would suggest that our original assumption is valid, that the valuation of cryptocurrencies is highly subjective.

What We Are Still Working On

- Adding in explicit SA features (TextBlob)
- Modifying weights of features to not so heavily favor token features (ratio of about 500:1 token to non-token features)
- Expanding the training dates
  - One day may not be representative of a standard day of trading
  - May be worth sampling from a week rather than grabbing a full day
- Identify ideal # and thresholds for Bayes classifier
  - 5 may be too many categories for our training data to work well with
  - Similarly, our thresholds may limit the amount of training data per category (i.e. the ratio of C1 to C2 tweets may be great enough to skew the outcome of the prediction)

Syracuse University
College of Engineering
& Computer Science

Questions?