Market News Analysis to Classify Bitcoin Pricing Trend

TEAM: SLAMMING SQUAD

TEAM MEMBERS:

HANG YANG, SHUANGXI ZHU, SHIYONG LI, PEIHONG MAN, MOHAMED SEFRI

Github Project Repo: edgeslab/Slamming-Squad

Outline

- Topic Selection
- Data collection and cleaning
- EDA and Visualization
- Feature selection and model building
- Evaluation
- Discussions

Topic Selection

Topic Selection - Why Bitcoin?

- What would celebrities talk about Bitcoin

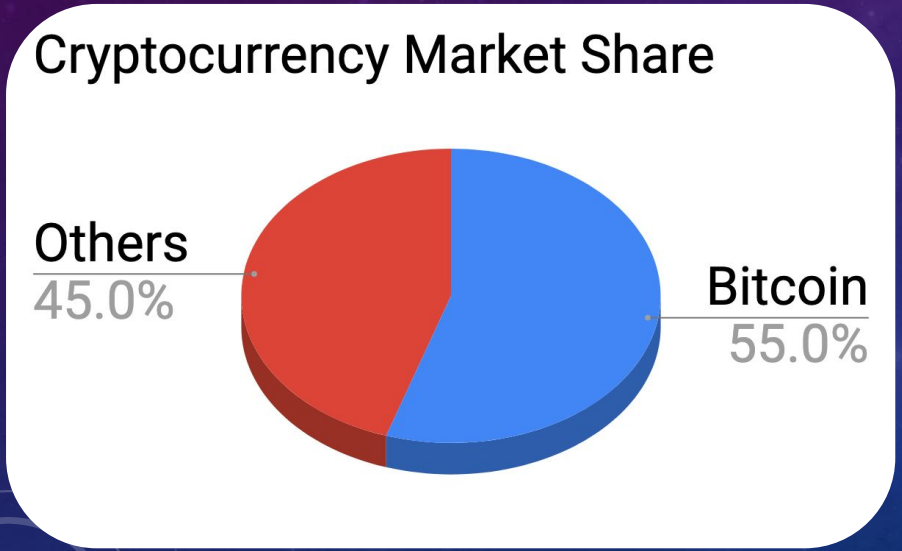
"Bitcoin is a technological tour de force."
–Bill Gates, Co-founder of Microsoft

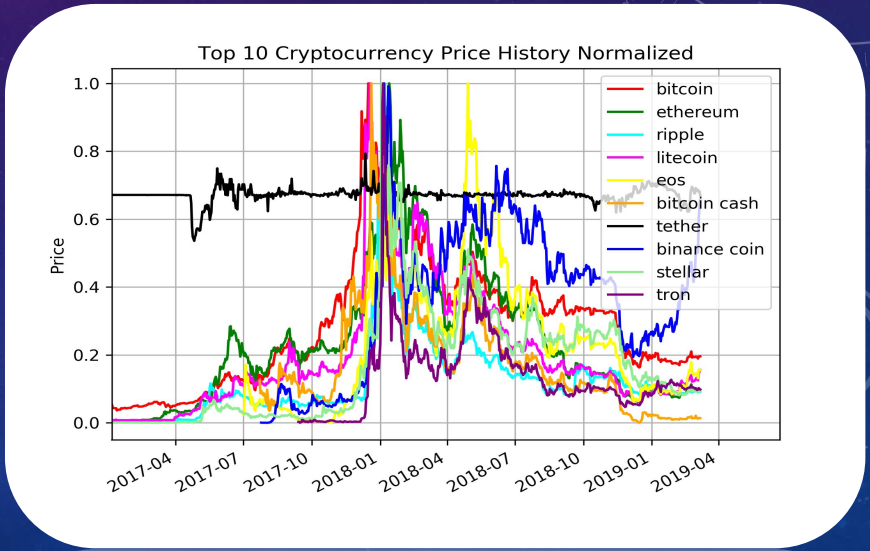
"I will eat my d–k' if I lose $500K Bitcoin bet"
–John McAfee, CEO antivirus software company McAfee Associates

Bitcoin holy high
-Slamming squad, uic

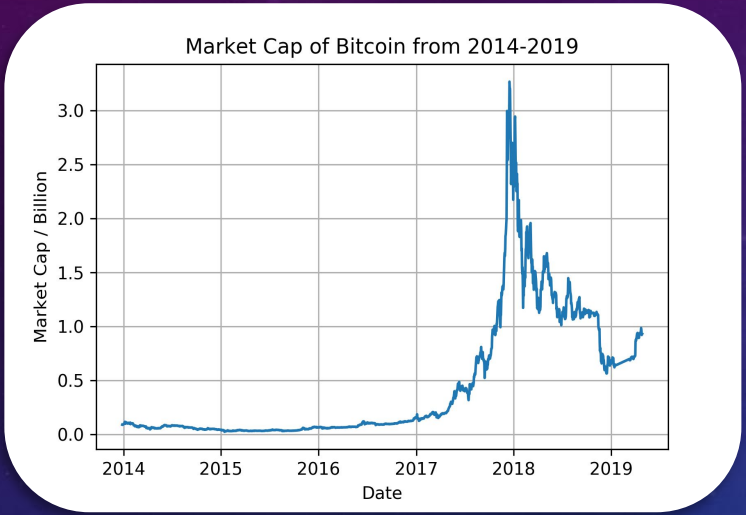
Topic Selection - Why Bitcoin?

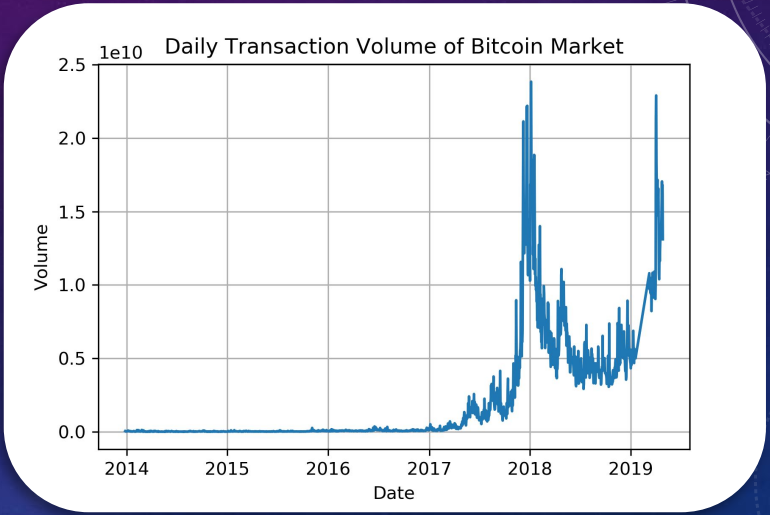
Bitcoin market dominance is around 55%;
We could find that top 10 cryptocurrency nearly have similar variation trend.





Market Cap and Daily transaction volume of Bitcoin / Billions

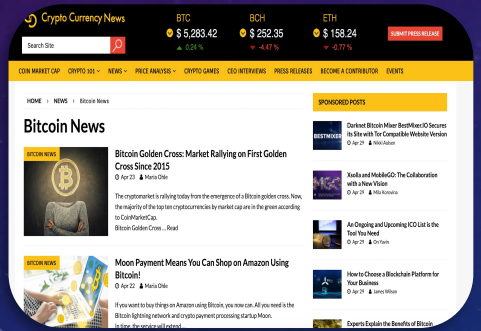


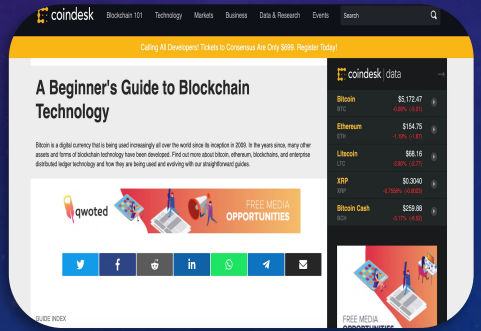


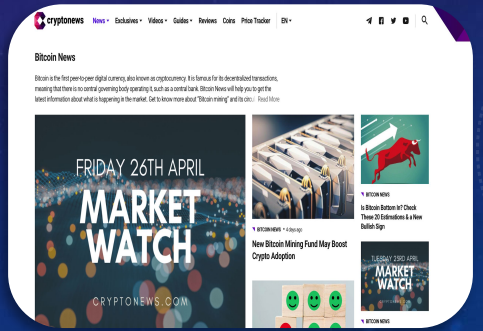
- The 2018 Bitcoin Crash is a nightmare (nearly 80% collapsed) to many investors, but the daily transaction volume and the market cap in 2019 seems to release a signal of Bitcoin revival.

Our goal is simple – to study or to make a prediction on the Bitcoin's price trend based on the news

- Stories tells. We try to build a classification model based on the most popular browsed Bitcoin news medias, as more and more people rely on information from internet feeds.







Data collection and cleaning

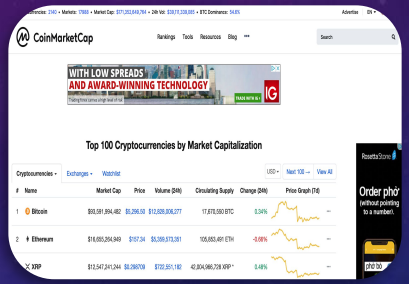
Data collection and cleaning

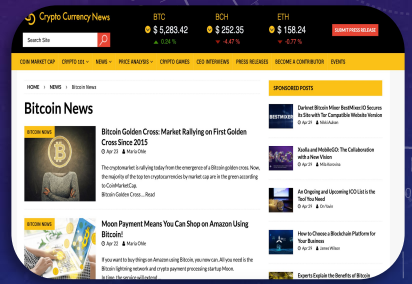
Web crawler by python

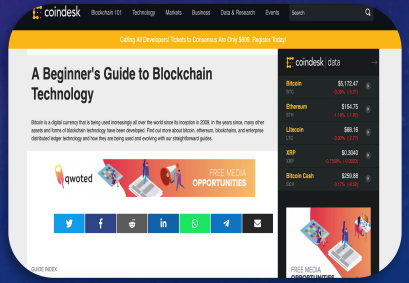
- Restful API
- requests
- Beautifulsoup4
- Selenium/webdriver

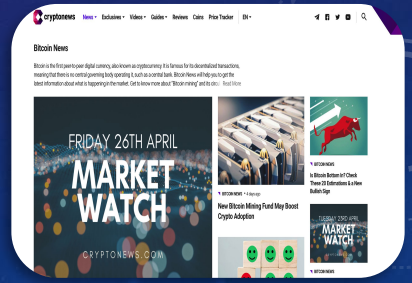
Websites:

- ccn.com
- cryptonews.com
- cryptocurrencynews.com









Data cleaning

- Data transformation: object-> datetime, object -> string, object-> float
- Data standardization/normalization
- Data aggregation: combine sets of media data together
- Data Label: If Bitcoin each day' (close-open) is positive, we label it as 1, otherwise we label it as 0.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2142 entries, 0 to 2141
Data columns (total 7 columns):
Date           2142 non-null object
Open           2142 non-null object
High           2142 non-null object
Low            2142 non-null object
Close          2142 non-null object
Volume         2142 non-null object
Market Cap     2142 non-null object
dtypes: object(7)
memory usage: 117.2+ KB
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2142 entries, 0 to 2141
Data columns (total 3 columns):
date      2142 non-null datetime64[ns]
open      2142 non-null float64
close     2142 non-null float64
dtypes: datetime64[ns](1), float64(2)
memory usage: 50.3 KB
```

EDA And Visualization

Raw dataset matters

- Challenges & Solutions:
  - Bitcoin price changes every second
    - Our product mainly aims at long transaction sellers.
  - Bitcoin news websites have many restrictions on crawling (CCN has its firewall from avoiding hacking their website)
    - Our team members crawl CCN weekly(slowly) based on their limits and guidelines
  - Different data format collected from different websites
    - EDA and Pre-process

EDA – Pricing dataset

	Date	Open	High	Low	Close	Volume	Market Cap
0	2019-04-27	5,279.47	5,310.75	5,233.64	5,268.29	1.311127	0.930862
1	2019-04-26	5,210.30	5,383.63	5,177.37	5,279.35	1.681211	0.932723
2	2019-04-25	5,466.52	5,542.24	5,181.34	5,210.52	1.533028	0.920465
3	2019-04-24	5,571.51	5,642.04	5,418.26	5,464.87	1.704803	0.965300
4	2019-04-23	5,399.37	5,633.80	5,389.41	5,572.36	1.586731	0.984174

- Structure: CSV file, column types are object
- Granularity: each row/record represents daily statistics on Bitcoin
- Scope: the earliest data we got is from 2013 to now
- Temporality: we collected the data weekly to get the latest
- Faithfulness: the data is collected from reliable authorized websites

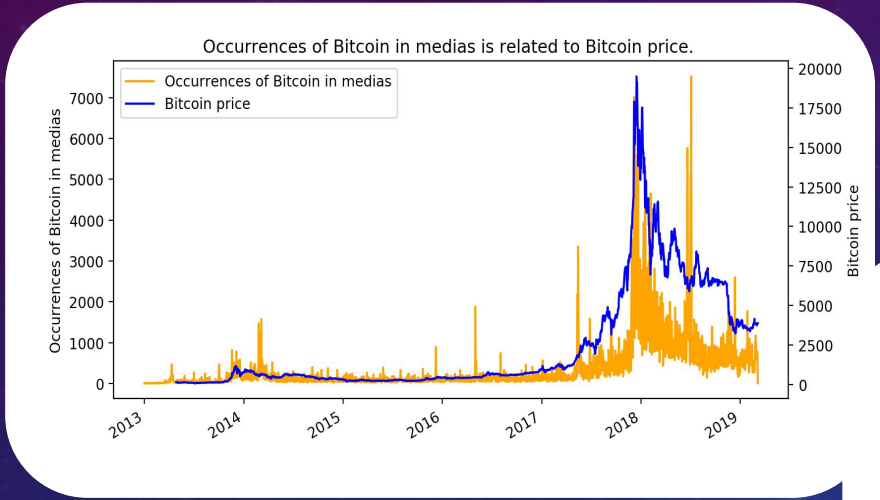
EDA – Media dataset

	text	date
0	The past two sessions saw a lot of volatility...	2019-04-26
1	Recently, bitcoin price started a downside co...	2019-04-25
2	Entrepreneur and Bitcoin bull John McAfee say...	2019-04-24
3	Specialist blockchain and artificial intellig...	2019-04-24
4	Two-thirds of 10,000 surveyed Europeans belie...	2019-04-2

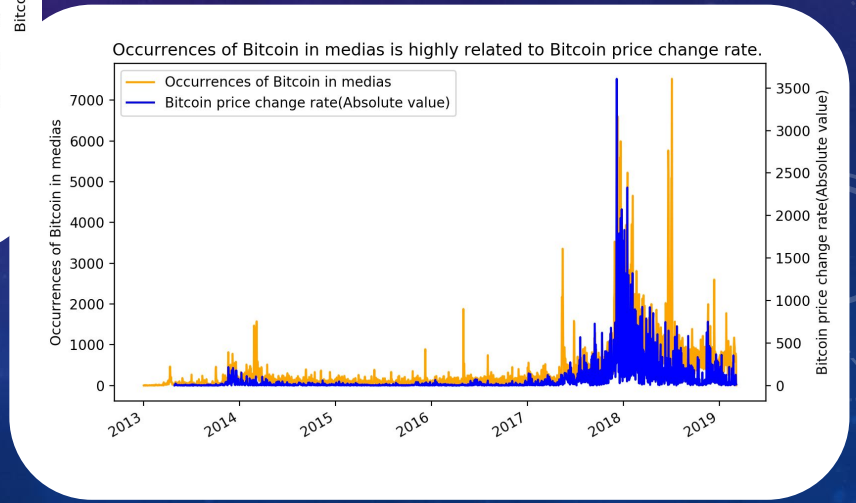
	text	date	year
0	The cryptomarket is rallying today from the e...	23-Apr	2019
1	If you want to buy things on Amazon using Bit...	22-Apr	2019
2	The cryptocurrency top ten are in the green! ...	10-Apr	2019
3	Cryptocurrency investment is on the rise righ...	9-Apr	2019
4	It's a bullish day in the cryptosphere. Paypa...	2-Apr	2019

- Structure: CSV file, column types are object
- Granularity: each row/record represents an article about Bitcoin
- Scope: collected sample size is small, we combined several media together.
- Temporality: we collected the data weekly to get the latest data
- Faithfulness: the data is collected from reliable authorized media websites

Visualization - The word “bitcoin” occurrence frequency in the media with the bitcoin pricing trend



Occurrences of Bitcoin in medias is related to Bitcoin price.

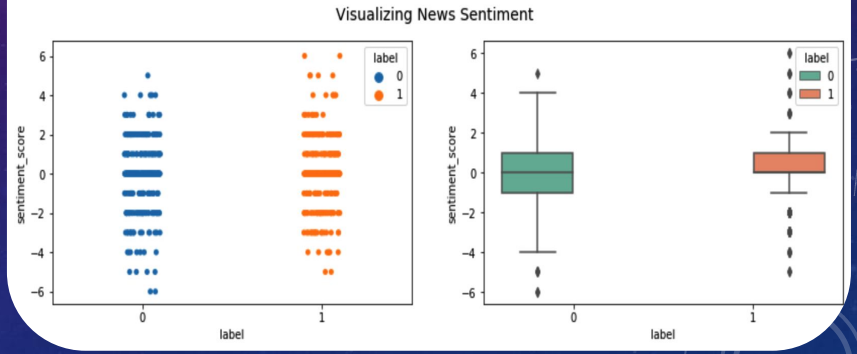


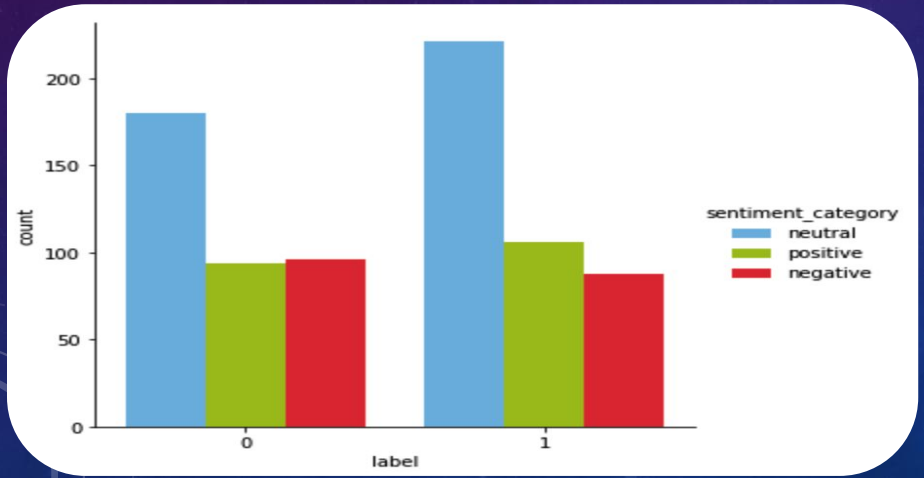
Occurrences of Bitcoin in medias is highly related to Bitcoin price change rate.

Visualization - media sentiment analysis with afinn

It tells that there is more polarity in the label-1 ‘rising’ trend on the Bitcoin price. Positive words in “rising” trend makes sense.

	sentiment_score							
	count	mean	std	min	25%	50%	75%	max
label								
0	370.0	-0.108108	1.654161	-6.0	-1.0	0.0	1.0	5.0
1	415.0	0.050602	1.600073	-5.0	0.0	0.0	1.0	6.0





Model selection and building

Text preprocess and feature extraction

- Lowercase
- Punctuation handling
- Stopword
- Remove rare words
- Tokenization
- Lemmatization
- tfidfvectorizer

Classifier and Model training

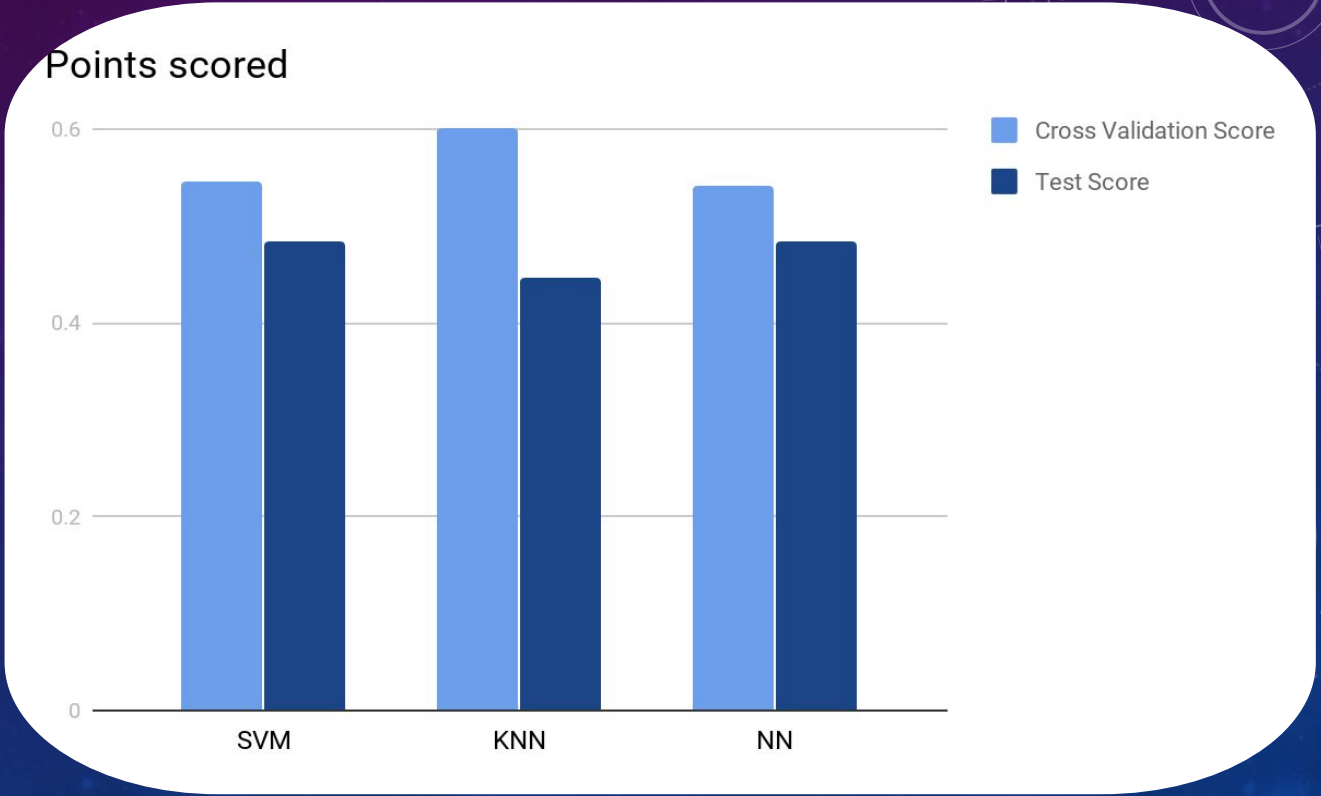
- Support Vector Machine(SVM)
- K Nearest Neighbor(KNN)
- Neutral Network(NN)

Evaluation

Better models are SVM and NN

Nearly 56% training accuracy

Nearly 50% test accuracy.



Discussions/ Takeaways

1. Data collection: causal or effect on news. How to distinguish such kind of news?
2. Try more ML models to see if there is a better fit.
3. Limit the number of features in feature matrix.
4. Sample Biased processing (SMOTE Oversampling, Random Oversampling)
5. Maybe more coarse grained data label is needed??

Thank you!