Cryptocurrencies Price Prediction Using News and Social Networks Data

Arnaud Autef
Catherine Gu
Olivier Pham
Charbel Saad
Benoît Zhou

Overview

1. Problem description
2. Data Collection
3. Feature Engineering
4. Model & Evaluation
5. Conclusion & Takeaways

Overview

Top 100 Cryptocurrencies by Market Capitalization

Cryptocurrencies | Exchanges | Watchlist | USD | Next 100 → | View All

#	Name	Market Cap	Price	Volume (24h)	Circulating Supply	Change (24h)	Price Graph (7d)
1	Bitcoin	$141,491,248,675	$7,976.12	$23,153,242,863	17,739,362 BTC	-7.40%	
2	Ethereum	$26,268,025,515	$247.00	$9,872,839,695	106,346,402 ETH	-7.68%	
3	XRP	$17,376,386,250	$0.411938	$2,553,674,556	42,181,995,112 XRP *	-8.36%	
4	Bitcoin Cash	$7,046,271,276	$395.44	$2,202,815,324	17,818,775 BCH	-9.79%	
5	Litecoin	$6,438,460,335	$103.73	$4,019,675,150	62,068,951 LTC	-9.20%	
6	EOS	$6,138,276,865	$6.69	$4,014,765,857	917,678,573 EOS *	-11.02%	
7	Binance Coin	$4,302,042,406	$30.47	$419,528,781	141,175,490 BNB *	-5.81%	
8	Bitcoin SV	$3,839,016,605	$215.47	$1,108,711,999	17,816,861 BSV	-1.95%	
9	Tether	$3,129,805,000	$0.997987	$24,228,629,558	3,136,118,221 USDT *	-0.55%	

Problem description

- Low barrier to entry for trading cryptocurrencies

- Large tradable market by dollar value
  - >2200 cryptocurrencies
  - 19000 markets trading crypto
  - $250bn market cap

- Is there a relationship between social network/news data and cryptocurrencies price?

Research Roadmap

Data Collection

- Market Data from Coinbase
- Google Trends search volume
- Twitter textual data

Exploratory Analysis

- Google Trend volume // BTC volume
- BTC keyword lagged // BTC volume
- Keyword volume // volatility

Directional prediction

- Sentiment Analysis
- Directional imbalances
- Condition on volatility

Model & Evaluation

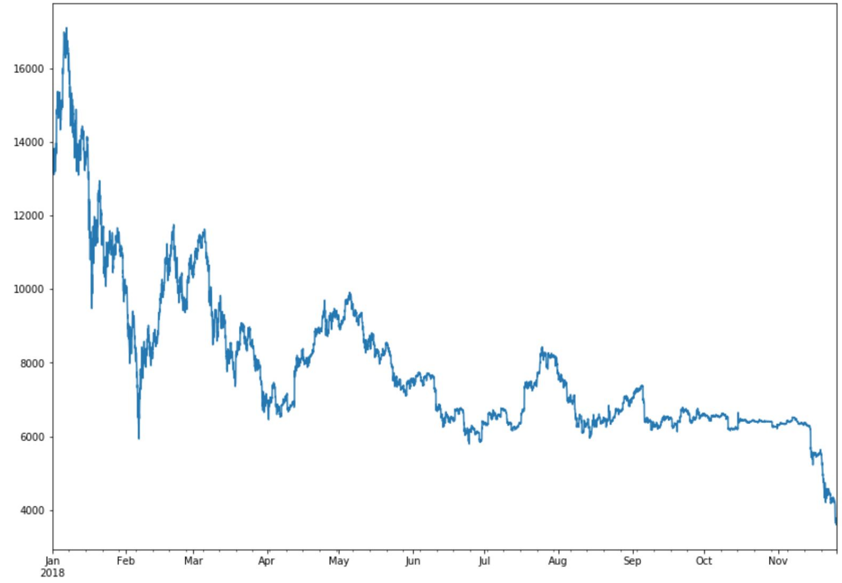
- Build a linear model
- Roll-forward Backtest

Data Collection

Market Data Collection

- L0 data from **Coinbase API** of currency pairs for 2018
- Different market behavior

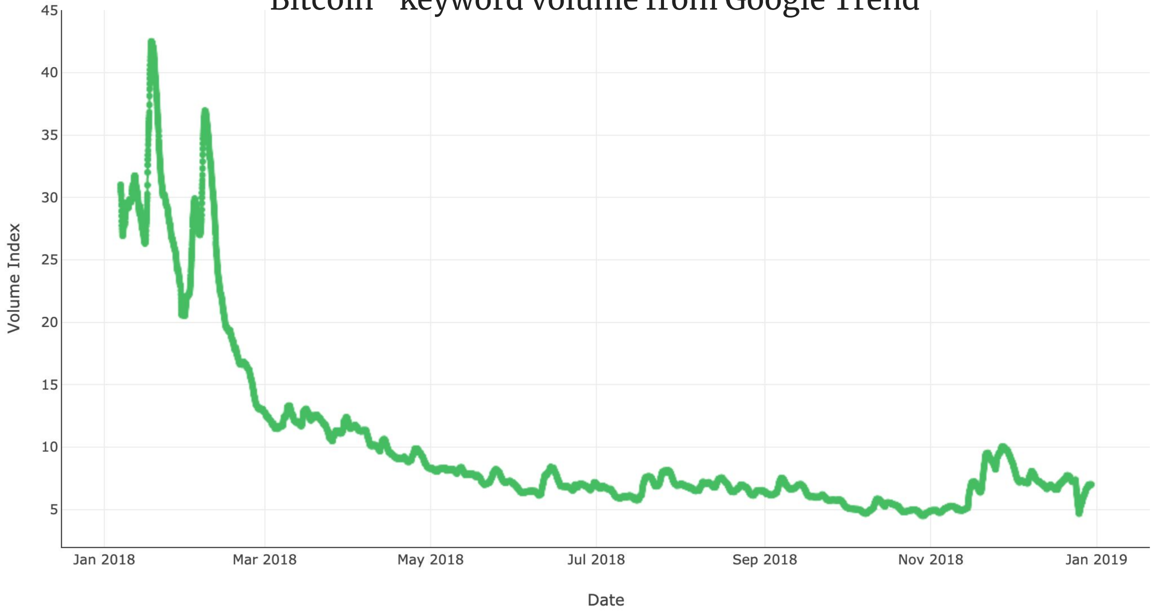
Price



Google Trends Data

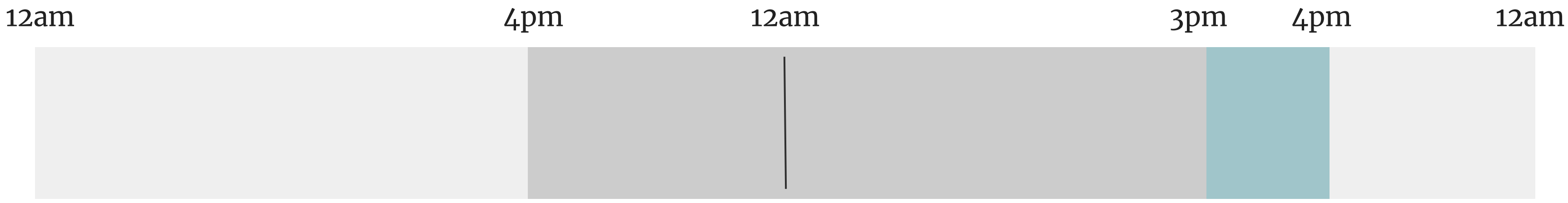
Example: Google Trend on “*bitcoin*”

“Bitcoin” keyword volume from Google Trend



Twitter Textual Data

- Used the “GetOldTweet” repository
- Returns all tweets in **chronological order** between 4pm to 4pm of the next day: **too much data**
- Selected only 1000 tweets between **3pm** and **4pm**

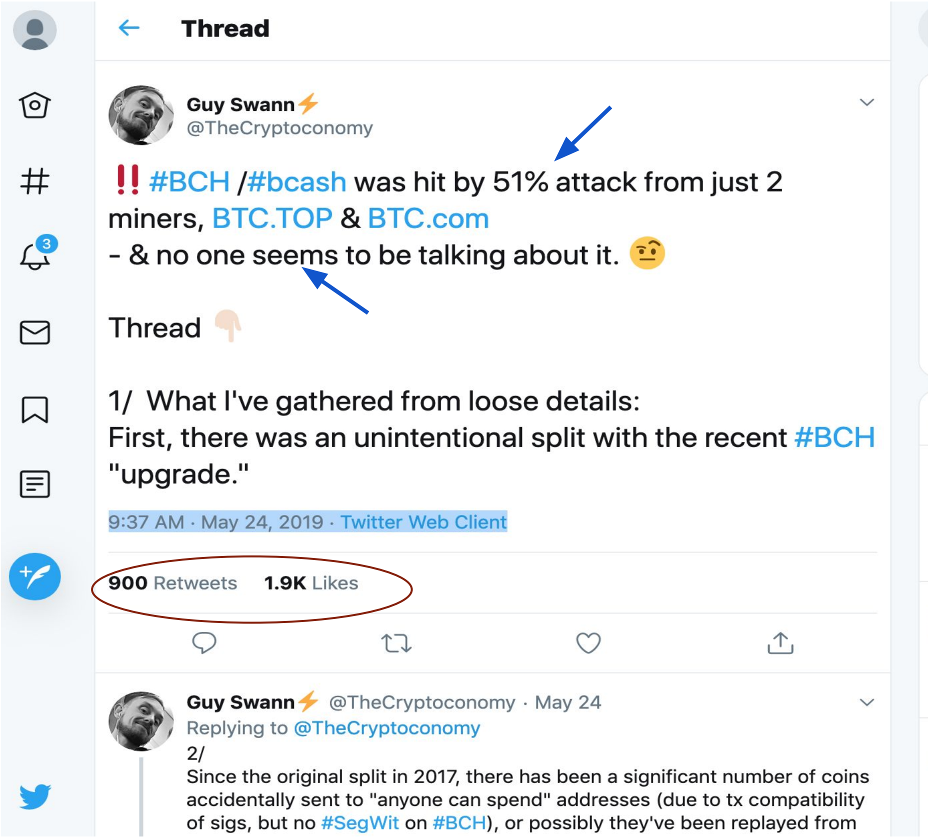


Feature Engineering

Tweets:

- Some very informative tweets but...

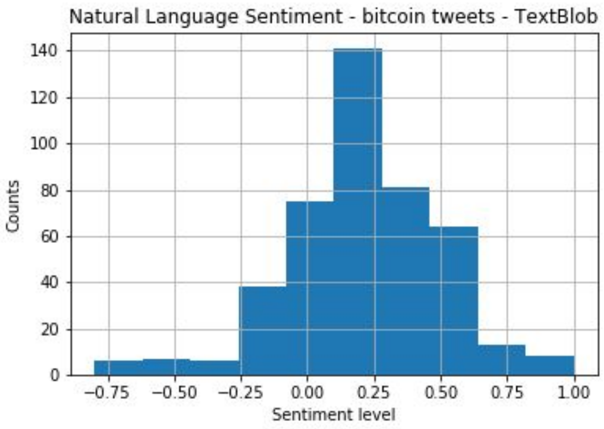
- Not structured
- Noisy

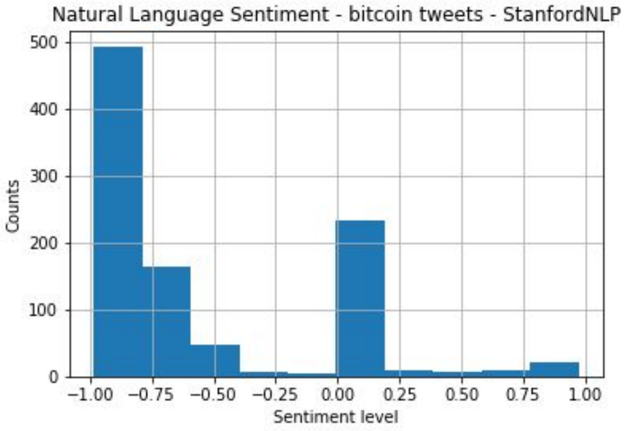


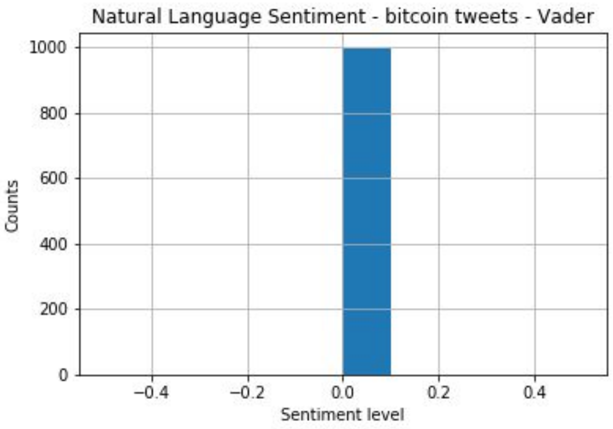
Sentiment Analysis for tweets

Compared 3 libraries on 1k tweets:

- *TextBlob* (naive Bayes classifier)
- *StanfordNLP*
- *Vader*







Keyword volume

- Counted positive and negative keywords volume in Tweets and Google Trends volume
- *Imbalance =*

  # positive words – # negative words

Positive Keywords:	**Negative Keywords:**
'conviction','bold','up', 'buy', 'bullish', 'bull', 'free money', 'long', 'rise', 'boom', 'bid', 'support', 'grow', 'get', 'make', 'earn', 'investment', 'invest', 'investing', 'invested', 'buying', 'bought', 'pump', 'like', 'skyrocket',	*'scam', 'capitulation', 'down', 'fork', 'sell', 'short', 'bear', 'bearish', 'bubble', 'stop', 'crash', 'clamp', 'shut', 'freeze', 'fall', 'bust', 'trash', 'forbid', 'oppose','dash', 'sold', 'selling', 'collapse', 'plummet', 'plunge'*

Predictor Construction for Twitter

	Positive Tone	Negative Tone
Positive Imbalance	Buy	Sell
Negative Imbalance	Sell	Buy

Keyword Score = *Imbalance* x *sign(sentiment score)*

Twitter Predictors

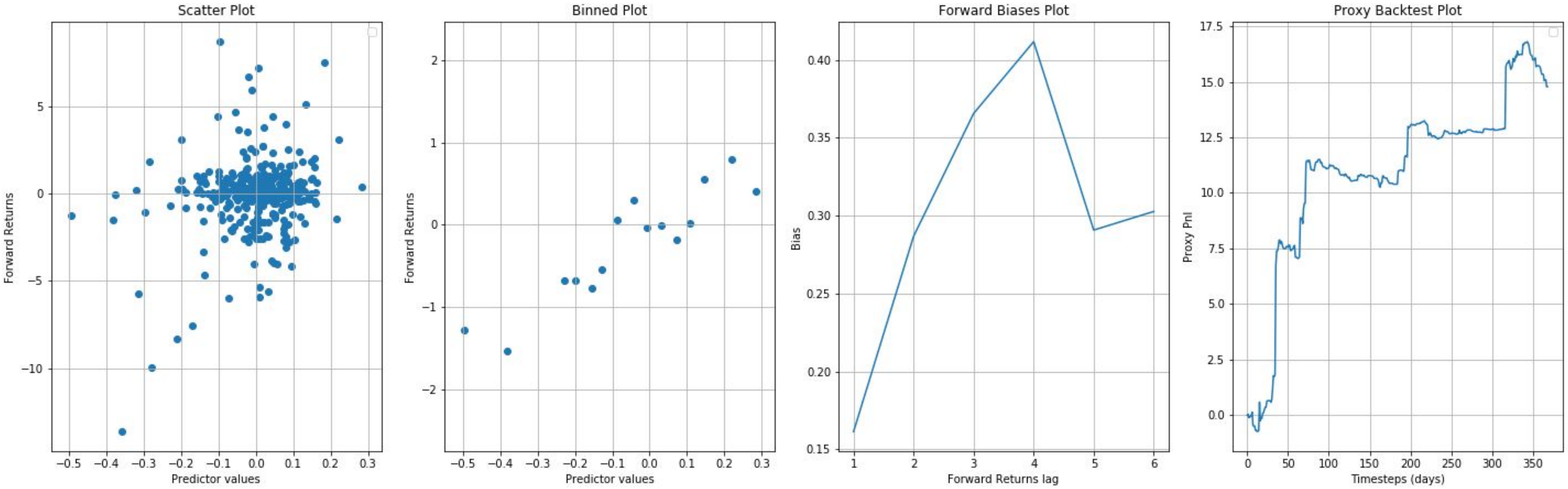
- Buy signal: Tweets with positive Sentiment Analysis score over the last hour.
- Sell signal: Tweets with negative Sentiment Analysis score over the last hour.
- Normalized signal: $$\frac{\text{Buy} - \text{Sell}}{\text{Buy} + \text{Sell}}$$

We expect signals to work best when using their discrete derivative, so we substract them from their moving average.

Plots & Metrics

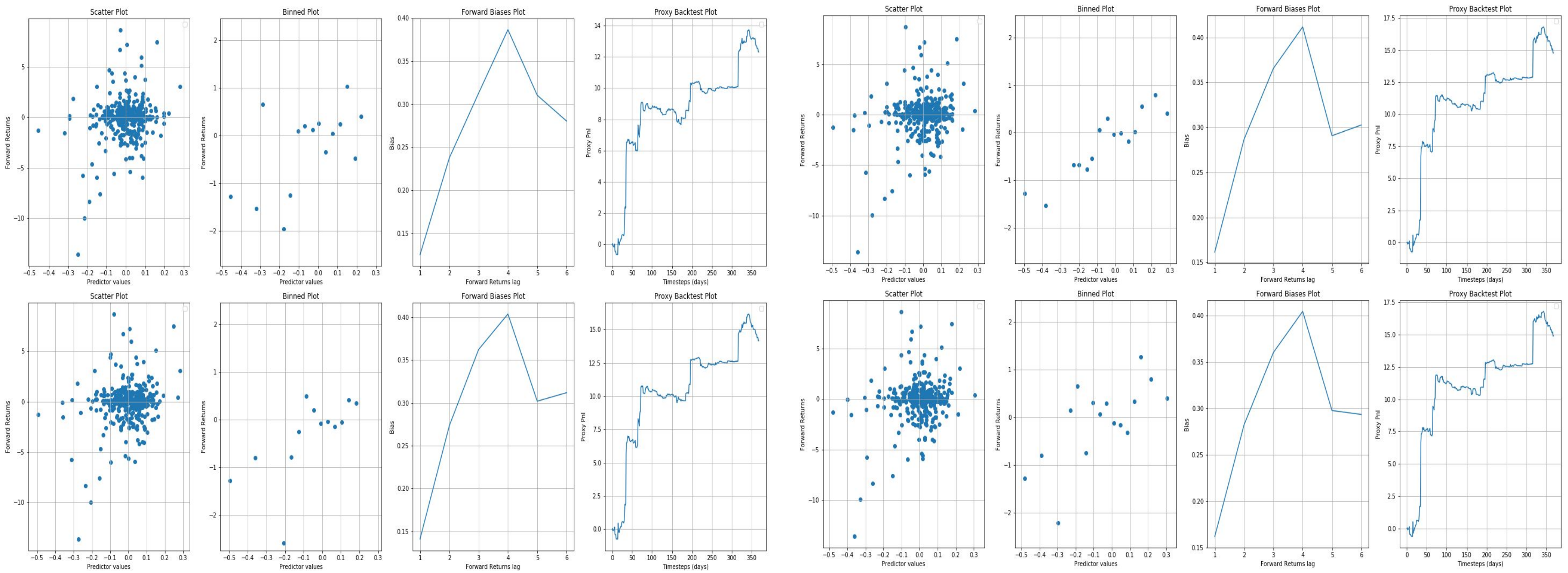
$$bias(lag) = \mathrm{corr}\left[return(t+lag),\, predictor(t)\right] \cdot \sigma(t+lag)$$

$$lag^* = \mathrm{argmax}_{lag}(bias) = 4\mathrm{hours}$$



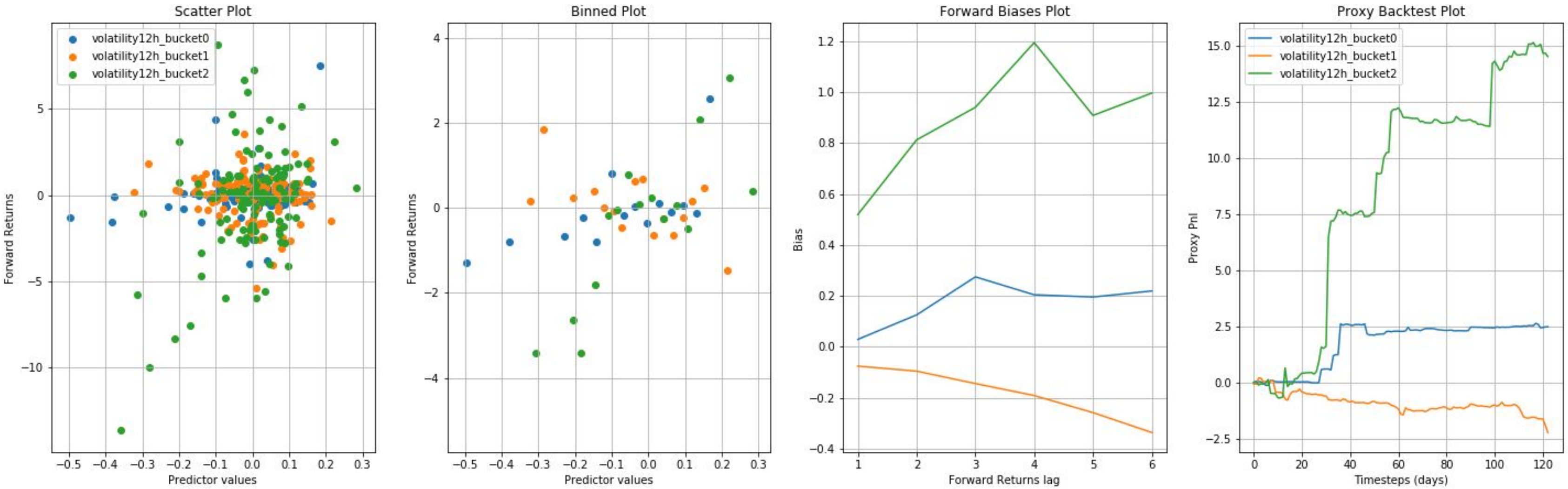
Parameters selection

- Moving average: 3, 5, 7 & 9 days
- Returns horizon



Final predictor

Normalized predictor + Conditioning on volatility regime



High

Medium

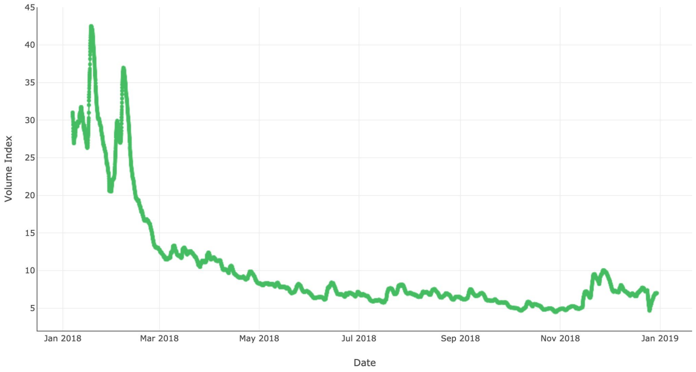
Low

Predictor Construction for Google Trends

Signal for Google Trends:

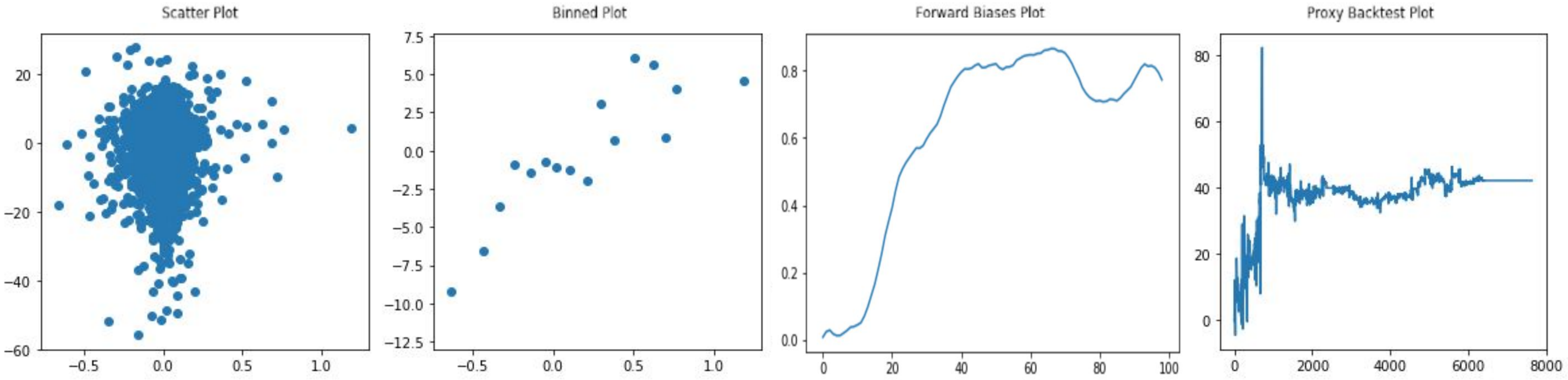
$$\frac{\text{Positive Volume} - \text{Negative Volume}}{\text{Total Volume}}$$

Centered with the ewm



Google Trends Predictors

- Positive correlation
- Longer horizon (2 days)
- Limited prediction power after the beginning of the year



Model & Evaluation

Model

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- No intercept
- Y = BTC-USD return 4 hours ahead of 4pm
- X_i = sentiment feature conditioned on past 12 hour volatility regime: low/median/high
- Moving average of 7 days
- In-sample $R^2 = 0.041$

Evaluation

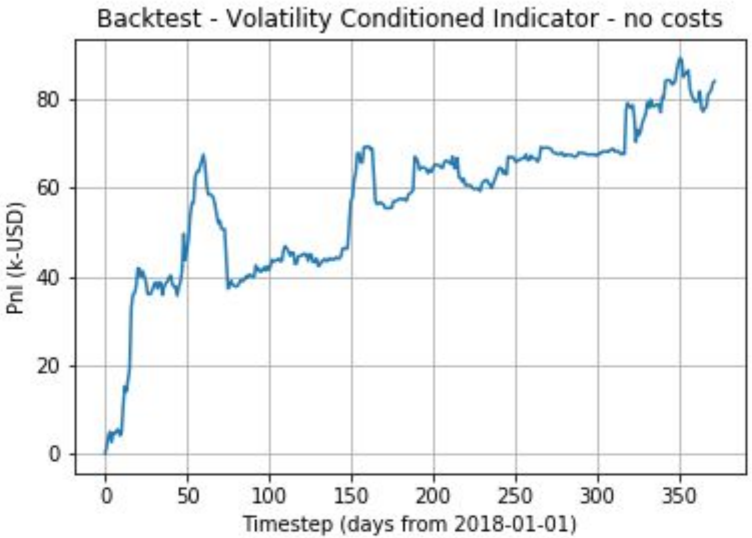
Roll-forward backtest of two weeks span

Week 1	Week 2	Week 3	Week 4				
Week 1	Week 2	Week 3	Week 4	Week 5	Week 6		
Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8

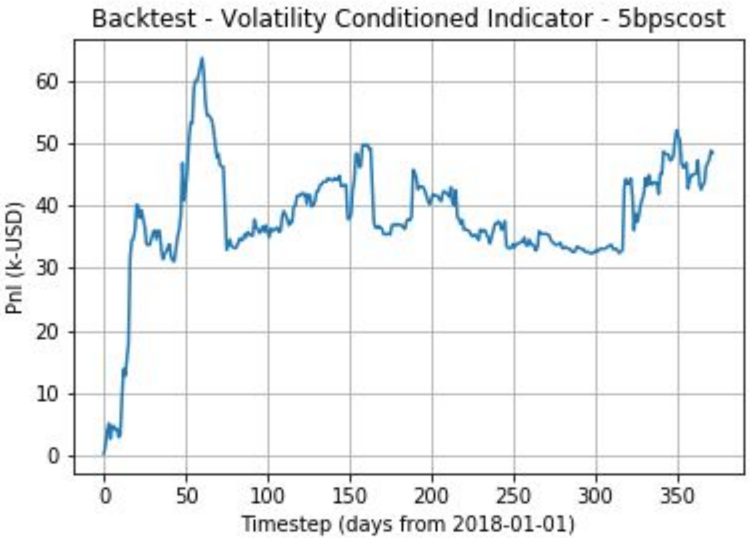
- Threshold based trading strategy
- Threshold Γ= trading cost
- Linear trading costs

Evaluation

Roll-forward backtest of two weeks span



no trading cost

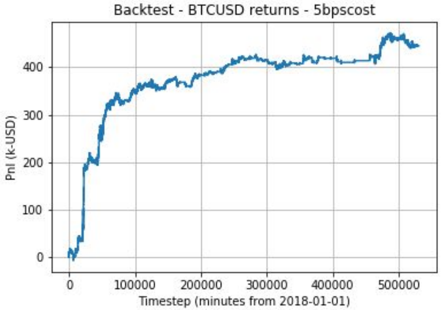


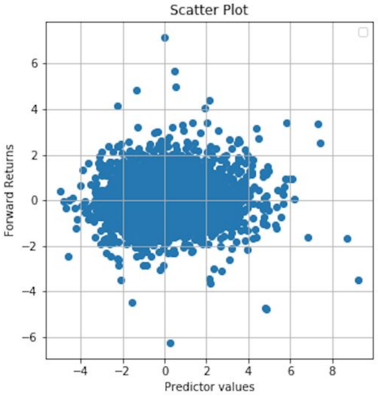
trading cost = 5 bps

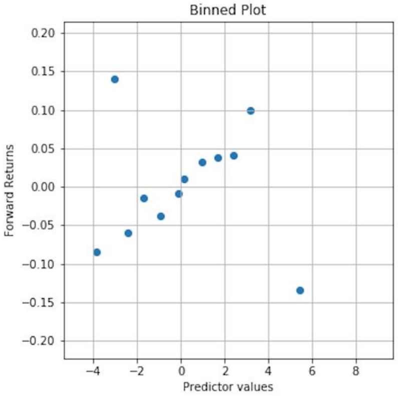
Sharpe ratio with no trading cost: $S = 2.44$

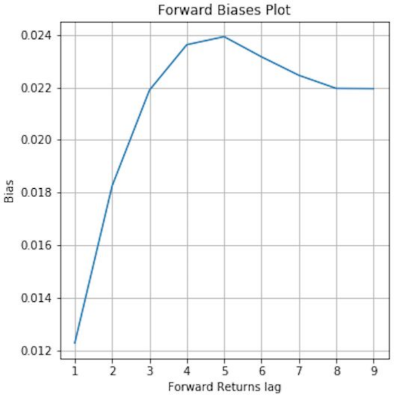
Market Data based model

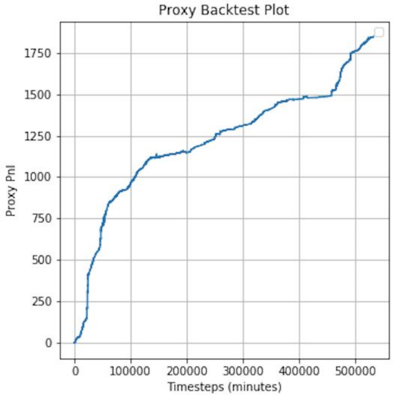
- <u>Extension</u>: Market data for higher frequency strategies
- Lead lag Bitcoin ➡ LiteCoin 5 min
- Sharpe ratio of 38.5











Conclusion & Takeaways

- Choice of sentiment analysis package is important
- Engineering challenges: quality of sentiment data, feature construction
- Filtering methods (e.g. volatility conditioning) produces more reliable results and align with the fundamentals observed in finance
- Portfolio construction: accounting for trading cost, introducing alpha threshold to determine position turnover, imitating real trading environment
- Extension: combining our signals to produce lower risk reliable strategies

Questions