Influence of Social Media on Crypto Market based on Sentiment Analysis

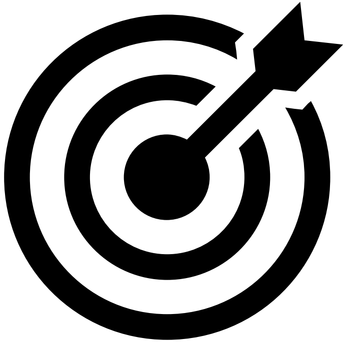
Text Mining Project

Alexandra Baur, Joris Bertens, Marius Löwe, Luis Riveros

M20180037, M20180423, M20180410, M20180753

What influence does social media have on the Bitcoin price or does a change in price trigger social media behaviour?

Objectives



1. Insight on interaction of social media behaviour and cryptocurrency market prices
2. Process real-time data in order to predict changes of stock prices or development of social media behaviour

We used the scrapy library to obtain 800.000 BTC related tweets via the twitter API

NOVA IMS Information Management School

Methodology -Data retrieval

Utilize Scrapy package

Scrapy

Twitter API

twitter

Obtain Dataset

- ~ 800.000 tweets
- January 2019 only
- Filtered for: bitcoin, Bitcoin, #bitcoin and #Bitcoin

Later we found that lemmatizing or stemmatizing did not improve the performance of the classifier

NOVA IMS Information Management School

Methodology - Preprocessing

1. about 95% in English language (filtered)
2. Removal of smileys, @usermentions, #-signs, links, etc.
3. NLP specific Preprocessing
   - Tokenization
   - Stop word removal
   - Normalization
   - Lemmatization (not included in final model)
   - Stemmatization (not included in final model)

Model 1 & 2: Build bottom line model with existing libraries to build upon in further analysis

Methodology - Classification

1 **Apply pre-trained models**

NLTK

TextBlob

2 **Training data**

IMDb (*)

Vote classifier

LogR

Gauss NB

Nom. NB

SGD

LinSVC

Bernoulli NB

NuSVC

No measure of accuracy on our own data set, except individual assessment

* Source: https://pythonprogramming.net/static/downloads/short_reviews/

Model 3: Training a self constructed classifier

Methodology - Preprocessing

3

Tagging tweets manually

4000 tweets

+ 615

± 2970

− 415

Balancing data

+ 415

− 415

Vote classifier

LogR

Gauss NB

Nom. NB

SGD

LinSVC

Bernoulli NB

NuSVC

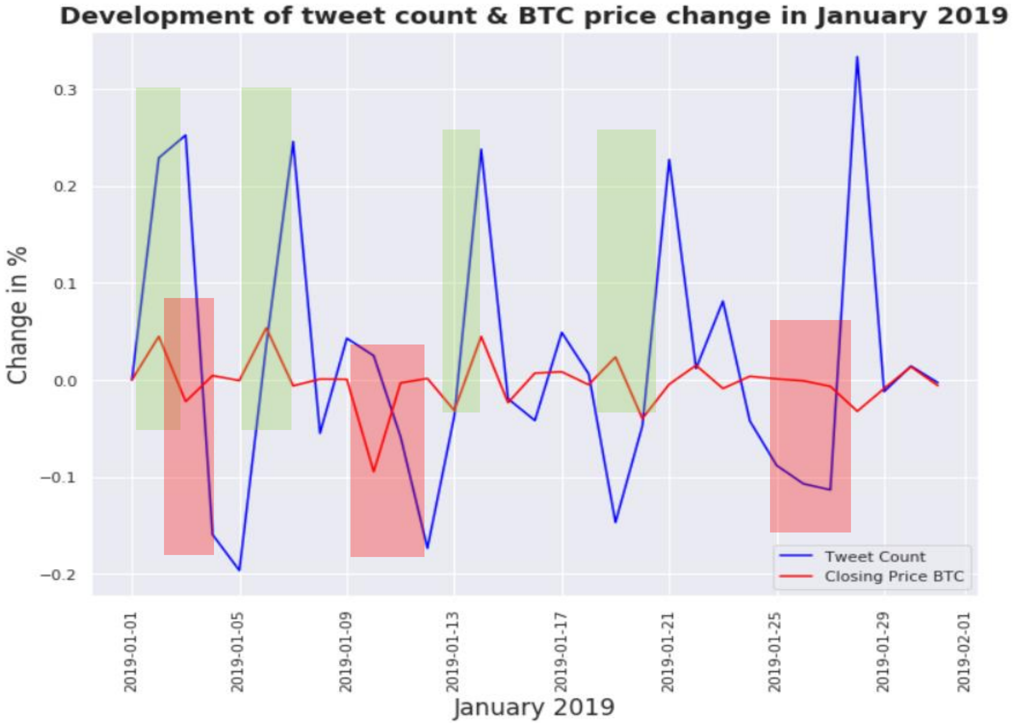
We found 4000 tweets as training data to be few, and difficult “Crypto-Slang”

There seems to be a delay in the twitter reactions to bitcoin price-changes

Results – Tweet Count

- Amount of tweets correlates with the changes in price

	Day Tweet = Day Sent.	Day Sent. -1	Day Sent. -2
Mean Error	0.09	0.102	0.102
Correlation	0.022	0.157	0.111



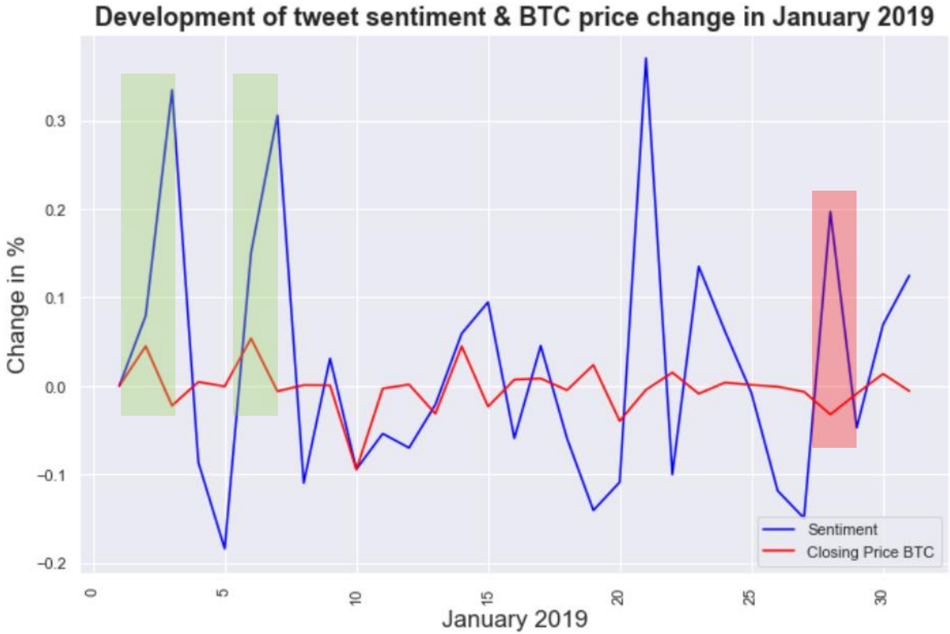
One day time shift in reaction of tweet sentiment to the BTC-price shows slight correlation

NOVA IMS Information Management School

Results – Tweet sentiment NLTK

- Positivity of tweets correlates with the amount of tweets
- Overall the tweets about Bitcoin are more positive.

	Day Tweet = Day Sent.	Day Sent. -1	Day Sent. -2
Mean Error	0.10	0.11	0.12
Correlation	0.052	0.259	0.038

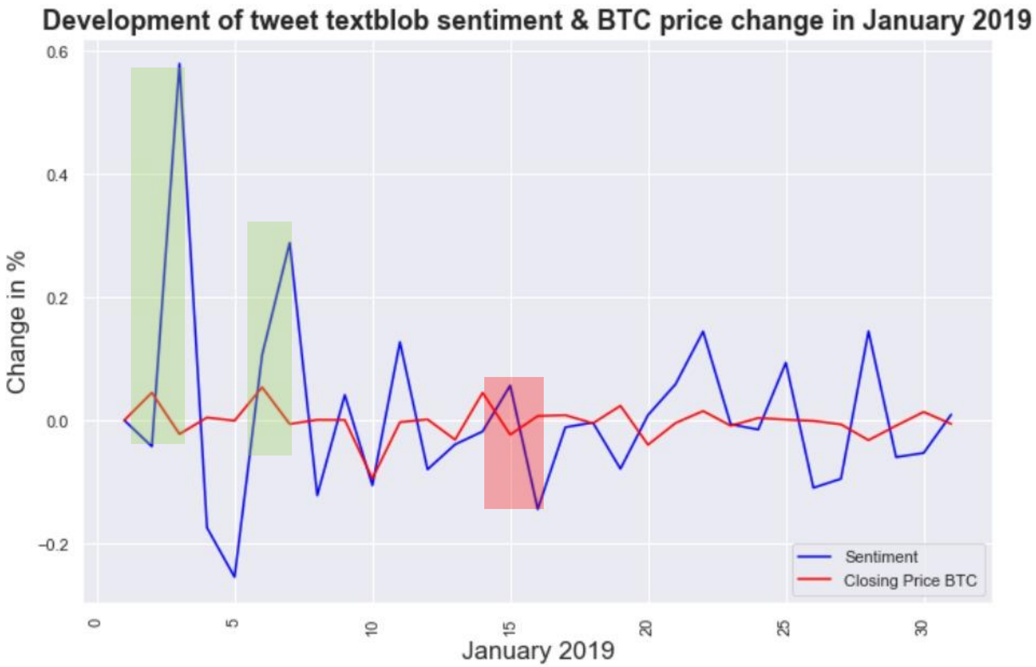


Reaction in Social Media is in general stronger (changes) than the price development

Results – Tweet sentiment TextBlob

- Best results regarding correlation of the curves

	Day Tweet = Day Sent.	Day Sent. -1	Day Sent. -2
Mean Error	0.10	0.09	0.10
Correlation	-0.064	0.322	-0.106

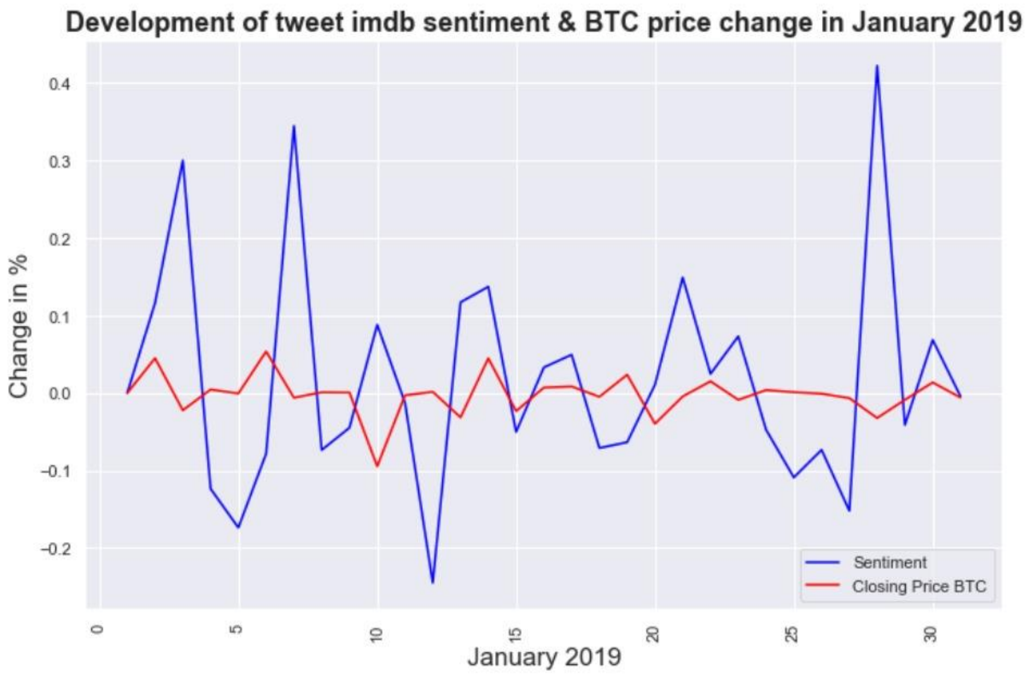


Classifier trained on tweets, but not about the same topic do not outperform the library classifiers

Results – Tweet sentiment IMDB

- No outstanding performance as expected due to movie related training

	Day Tweet = Day Sent.	Day Sent. -1	Day Sent. -2
Mean Error	0.11	0.11	0.11
Correlation	-0.215	0.201	0.163

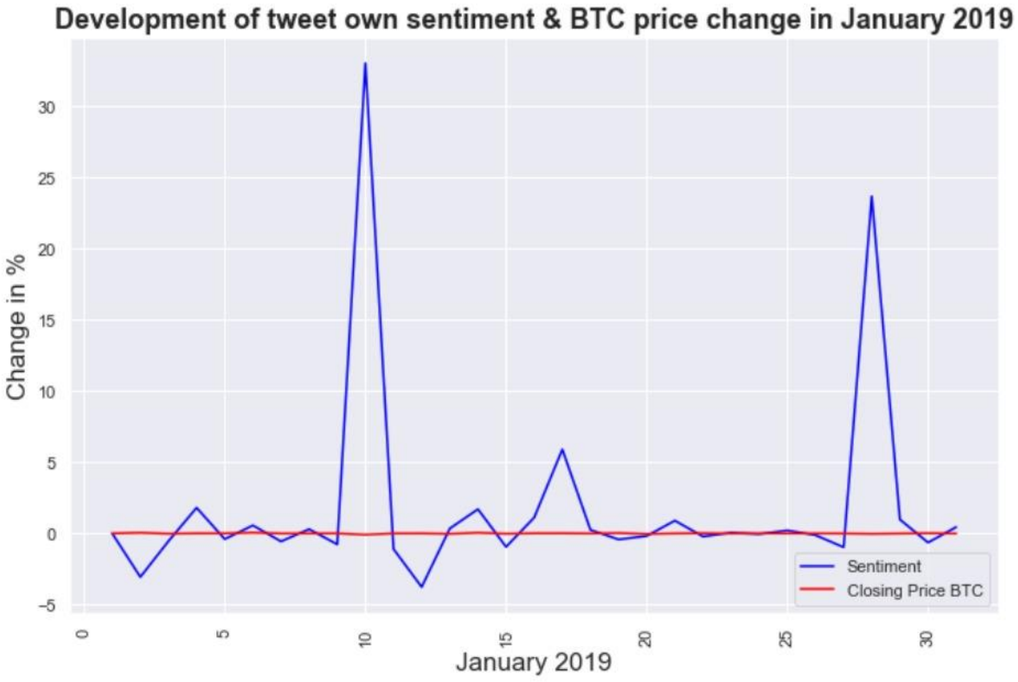


Own model needs to be trained with more data to improve

Results – Tweet sentiment Vote Classifier

- Few extreme values sekwed the heavily
- Therefore results hard to interpret

	Day Tweet = Day Sent.	Day Sent. -1	Day Sent. -2
Mean Error	2.74	2.84	2.81
Correlation	-0.626	-0.022	0.106



Data retrieval | Data Preprocessing | NLP Model Selection | **Results & Visualization**

The best results were achieved with the self trained model

Results – Table

	Count comparison	Baseline NLTK	Baseline TextBlob	Self trained model	IMBD trained model
Test accuracy**	not applicable	44%	50%	63%	55%
Mean Error*	0.10	0.11	0.09	2.81	0.11
Correlation*	0.157	0.259	0.322	0.106	0.201

* Calculation based on Day Sent. -1 **May vary due to shuffling of testing data

Equivocal results to test accuracy and correlation

Demonstration

Information Management School

Methodology – Live Streaming

Data retrieval | Data Preprocessing | NLP Model Selection | **Results & Visualization**

Bitcoin is thoroughly discussed on social media but does this mean they affect each others behaviour?

Challenges

Data retrieval	Data Preprocessing	NLP Model Selection	Results & Visualization
1. **Web scraping (API restrictions)** 2. Various languages 3. Retrieving live tweet data stream	1. Cleaning tweets effectively 2. Numbers treatment 3. Method selection for different models	1. Find BTC / Stock related classifier 2. **Find / produce suitable training data**	1. **Choosing metrics for results measurement** 2. Visualize effectively

Social media behaviour seems to be influenced by the bitcoin price

NOVA IMS Information Management School

Conclusion

- Social media is more likely to be influenced by the crypto market development than vice versa
- Live sentiment may give sentiment indication when lager price changes occurr
- Better training data & classifier performance would allow for deeper analysis

Not useful for use as Bitcoin investment tool, but as an supportive tool

Future work

1. Continue training to further improve the model
2. Considering retweets and followers reached
3. Include other currencies or hastags (ETH, #btc, etc.)
4. Check for other social media channels, e.g. Facebook
5. Check more historical data

References

- Bitcoin data: https://www.coindesk.com/price/bitcoin
- Twitter data: https://twitter.com/
- IMDB data: https://pythonprogramming.net/static/downloads/short_reviews/

NOVA IMS Information Management School

Code

Link: https://github.com/jorisbertens/Text_Mining_MASTER