Influence of Social Media on Crypto Market based on Sentiment Analysis

Group: Alexandra Baur - M20180037, Joris Bertens - M20180423, Marius Löwe - M20180410, Luis Riveros - M20180753

The objective of the project was to get an insight on the interaction of social media behavior and crypto currency market prices and to be able to process real-time data in order to predict changes of stock prices or development of social media behavior.

For that purpose it was decided to analyze data of January 2019. The data for the bitcoin price development was available on coindesk.com and could be exported as csv-file. The twitter data was web-scraped using the scrapy library and the twitter API. It was scraped for tweets that contain the following key works: Bitcoin, bitcoin, #Bitcoin, #bitcoin and resulted in around 800.000 tweets.

It was filtered for English tweets only, which were roughly 95% of the dataset (lang_detect library). @mentions, hashtags, numbers, links and break lines were removed. Additionally, the tweets were formatted and briefly cleaned. Further preprocessing of the tweets was done by tokenization, normalization, stop word removal. Lemmatization and Stemmatization was performed but later excluded from the model due to worse results when applying.

In order to determine the sentiment of the tweets, multiple classifiers were applied to find the best one. In the first place, the nltk and textblob libraries were used to classify the tweets. After receiving unsatisfactory results, it was decided to train an own developed vote-classifier that was trained on IMdB Twitter data. For this, seven different algorithms containing variations on Naïve Baise, Support-Vector-Classification and Gradient Decent were applied and voted on for a compound score. Since the classifiers used were trained with movie-related tweets, there was still room for improvement to classify crypto-related tweets.

For this reason, 4000 random tweets were labeled manually. From this 415 positive and 415 negative (already balanced) were included for training the data. Based on a validation threshold (70%) for the relative number of classifiers voting for a certain sentiment, the vote classifier was tested.

Finally all classifiers were evaluated by an accuracy score on own labeled test data of crypto-related tweets, indicating that the own created model has the highest accuracy.

To compare the development of crypto currency price development and twitter sentiment graphs were created for every classifier, after capturing the sentiment for all the January-tweets (800.000). The graphs represent the % change rate per day over the month of January 2019, were some correlation can be seen, but with some time-shift.

To evaluate the curves created by the different classifiers, the correlation and the mean absolute error between the curves was computed for the actual curve and the shifted curve by one and two days.

The results show that correlation increases in general if the bitcoin-curve is shifted by one day towards the sentiment curve, which indicates that bitcoin prices influence social media behavior / sentiment around one day later.

For the live sentiment analysis data is constantly retrieved from the Twitter API filtered for the same tweets and from the Lomond (WebSocket) API the current Bitcoin price is obtained.

TextBlob was used to capture the sentiment of each incoming tweet because it performs the fastest. Based on this assessment the tweet with the according sentiment is displayed (if in English). Furthermore a graph shows the current price and overall sentiment development in real time.

The findings of the real time data are can´t be used to predict bitcoin prices, but serve as a complimentary tool to capture live sentiment especially when there are lager changes in price.