Sensing Social Media Signals for Cryptocurrency News

Johannes Beck[*]
ETH Zurich
Switzerland
beckjoh@ethz.ch

Roberta Huang[*]
ETH Zurich
Switzerland
huangr@ethz.ch

David Lindner[*]
ETH Zurich
Switzerland
lindnerd@ethz.ch

Tian Guo
Computational Social Science,
ETH Zurich
Switzerland
tian.guo@gess.ethz.ch

Ce Zhang
Department of Computer Science,
ETH Zurich
Switzerland
ce.zhang@inf.ethz.ch

Dirk Helbing
Computational Social Science,
ETH Zurich
Switzerland
dirk.helbing@gess.ethz.ch

Nino Antulov-Fantulin
Computational Social Science,
ETH Zurich
Switzerland
anino@ethz.ch

ABSTRACT

The ability to track and monitor relevant and important news in real-time is of crucial interest in multiple industrial sectors. In this work, we focus on cryptocurrency news, which recently became of emerging interest to the general and financial audience. In order to track popular news in real-time, we (i) match news from the web with tweets from social media, (ii) track their intraday tweet activity and (iii) explore different machine learning models for predicting the number of article mentions on Twitter after its publication. We compare several machine learning models, such as linear extrapolation, linear and random forest autoregressive models, and a sequence-to-sequence neural network.

CCS CONCEPTS

• **Networks** → *Social media networks*; • **Computing methodologies** → *Machine learning algorithms.*

KEYWORDS

Social media, mining and learning, cryptocurrency

ACM Reference Format:
Johannes Beck, Roberta Huang, David Lindner, Tian Guo, Ce Zhang, Dirk Helbing, and Nino Antulov-Fantulin. 2019. Sensing Social Media Signals for Cryptocurrency News. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW'19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3308560.3316706

1 INTRODUCTION

Social media and news play an important role in driving the fluctuation of economic indicators and financial markets [5, 10, 13, 18] in a nontrivial fashion. Recently, novel financial markets have emerged, that are exchanging between fiat money and cryptocurrencies [9]. As of December 2018, cryptocurrencies have a total market capitalization of $120 billion, with more than 250000 transactions per day. Therefore, it is not surprising that the rapid development of cryptocurrency has attracted increasing attention from news and social media.

A large volume of news articles about cryptocurrencies, published daily can make it hard for individuals or traders to filter out relevant information and make informed decisions in this domain. Fortunately, people share and discuss news every day in large quantities on social media platforms, e.g. on Twitter, which is the focus of this paper. Therefore, social media can be a good proxy to monitor and track ”important” news about cryptocurrencies. Our work is motivated by the hypothesis that high engagement with a news article on Twitter is related to the ”importance” of an article.

In this paper, we introduce an online data mining system which connects news and tweets discussing it. We also perform preliminary data exploratory and predictive analytics using machine learning and deep learning. Overall, the contribution of this paper is as follows: (i) We build an online data mining pipeline to extract news articles from a discussion on Twitter and collect tweets associated with the articles. This paired news and tweet data is continuously updated in a cloud database. This data is a rich source for studying public interest and attention on cryptocurrency and the potential effect of social media on the market. (ii) Based on the news and associated tweets collected by the pipeline, we perform exploratory data analysis to characterize news discussion on Twitter. (iii) We apply machine learning and deep learning models to predict the popularity of news articles on Twitter. We aim to predict the number of tweets mentioning a news article related to cryptocurrencies, which we consider as a measure of its ”importance”.

[*] Authors contributed equally to this research.

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316706

2 RELATED WORK

Many studies have focused on the relationship between social media, news, and other information from the www onto financial markets [1, 6, 13]. However, the main focus of our work is modeling and prediction of news popularity via social media. In [17], the authors link a given news article to social media utterances that implicitly reference it through a dedicated query model. Tracking and automatically connecting news articles to Twitter conversations by Twitter hashtags was studied in [14]. In [3], the authors constructed a multi-dimensional feature space derived from an article and use a conventional SVM to predict its popularity. The authors in [11] show how the class of temporal point processes (Hawkes) can be used for predicting Retweet dynamics. The authors in [8] propose how to leverage knowledge base information for improving popularity prediction. Starting from the idea that only a small amount of news articles become popular, [12] focused on the subset of the most popular news to rank articles. In [2] it formulates article importance prediction as a classification task using SVM.

In this paper, we exploit ensemble machine learning and sequence to sequence (seq2seq) deep learning to study the predictability of crypto news popularity on Twitter in real-time mode. In contrast to others, our analysis is focused on the intraday importance prediction.

3 DATA PIPELINE

The data pipeline consists of a real-time online system, with the following components: (i) Twitter collection, (ii) news article collection, and (iii) tweet-article matching.

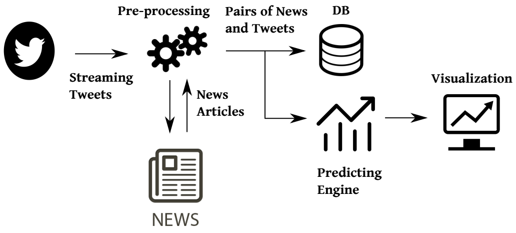


Figure 1: Architecture of the data pipeline.

The Twitter data collection was implemented by using the publicly accessible Twitter streaming API with real-time filtering by a list of cryptocurrency related keywords.

The article data collection is obtained by scraping news from the dynamic set of gazetteer source URLs. The set of gazetteer source URLs is automatically updated by extracting the URLs from the content of downloaded tweets.

The tweet-article matching data is the document-oriented database, that contains matchings between news and tweets. The matching exists if the tweet explicitly contains the URL of an article.

First post-processing step is **merging** of articles. We merge the matchings of two articles if they fulfill the following 3 conditions: (1) the URLs of both articles share the same host as well as the same path; (2) both articles have the same title; and (3) both articles were published at the same time. These conditions allow for merging of same articles of which the URLs have different query strings. While merging the articles we also remove duplicate entries for the same tweets which are sometimes present in the database. For instance, Fig. 2 and 3 demonstrate the top publisher, keywords and news ranking snapshot from the pipeline.

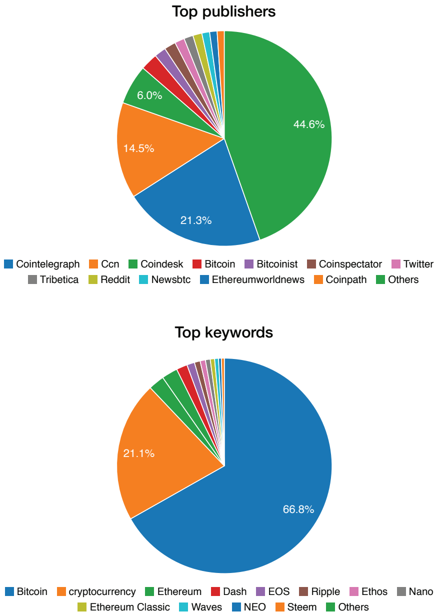
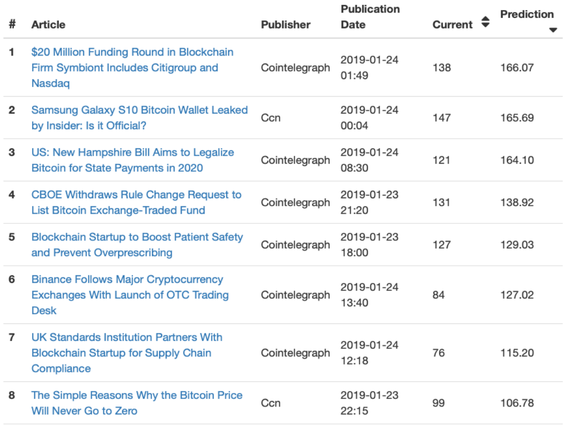


Figure 2: Top publishers and top keywords on 24th Jan 2019, visualized by our system.



#	Article	Publisher	Publication Date	Current ⬍	Prediction ▾
1	$20 Million Funding Round in Blockchain Firm Symbiont Includes Citigroup and Nasdaq	Cointelegraph	2019-01-24 01:49	138	166.07
2	Samsung Galaxy S10 Bitcoin Wallet Leaked by Insider: Is it Official?	Ccn	2019-01-24 00:04	147	165.69
3	US: New Hampshire Bill Aims to Legalize Bitcoin for State Payments in 2020	Cointelegraph	2019-01-24 08:30	121	164.10
4	CBOE Withdraws Rule Change Request to List Bitcoin Exchange-Traded Fund	Cointelegraph	2019-01-23 21:20	131	138.92
5	Blockchain Startup to Boost Patient Safety and Prevent Overprescribing	Cointelegraph	2019-01-23 18:00	127	129.03
6	Binance Follows Major Cryptocurrency Exchanges With Launch of OTC Trading Desk	Cointelegraph	2019-01-24 13:40	84	127.02
7	UK Standards Institution Partners With Blockchain Startup for Supply Chain Compliance	Cointelegraph	2019-01-24 12:18	76	115.20
8	The Simple Reasons Why the Bitcoin Price Will Never Go to Zero	Ccn	2019-01-23 22:15	99	106.78

Figure 3: Top article ranked by our system on 24th Jan 2019.

We work with three entities in our data: news articles, tweets, and matchings. Article entity has the following properties: `URL`, `title`, `publication time` and `text`. Tweet entity has the following properties: `user-id`, `text`, `publication time` and `links`. Each matching entity has `article-id` and `list of matched tweets`.

4 PREDICTIVE ANALYTICS

Let t_0, t_1 be two timestamps with $t_0 < t_1 < t_0 + \delta$, where we set δ to 24 hours in this paper. Given an article published at time t_0 and all tweets published between t_0 and t_1 that mention the article, the task is to predict the cumulative number of tweets mentioning the article between time t_0 and $t_0 + \delta$. Here t_1 is the prediction starting time, i.e. how much historical data can be used to predict.

4.1 Feature Extraction

The **time series feature** f_k is given by the number of mentions of the article between t_0 and $t_0 + kd$ where $k \in \{j \in \{1, 2, ...\} | t_0 + jd \leq t_1\}$. As an example suppose that t_1 is 3 hours after t_0 and the article is mentioned twice, once and three times in hours 1, 2 and 3 since publication respectively. Then there are 3 time series features: $f_1 = 2, f_2 = 3, f_3 = 6$. Note that the number of these features is not constant, but depends on t_1.

We extract a vector of **content features** from each article, by using a keyword list to allow the models to learn individual dynamics for articles related to different cryptocurrencies. Each cryptocurrency is represented by a binary feature, that is set to 1 if one of the keywords related to the concept is present in the title of the article.

The amount of Twitter mentions might further be related to the publisher of an article. We then extract the 10 publishers with the highest numbers of mentions. For each of these publishers, we introduce a **context binary feature** set to 1 if the article was published by the respective publisher.

4.2 Predictive Models

As a **baseline model**, we use a linear extrapolation of the last k time series features by fitting a linear function of the time step to the dataset given by $\{(K - k + 1, f_{K-k+1}), ..., (K, f_K)\}$. The model ignores content and context features. In our experiments, we will choose $k = 3$.

An **autoregressive model** (AR) [15] of order k predicts the value at the next timestep $K + 1$ based on the values observed at the previous k timesteps $K - k + 1, ..., K$. In our experiments, we provide K as an additional input to the model. The idea is, that the dynamics can be very different a few hours after the publication and shortly before the end of the prediction window. In our case, we have to predict multiple steps in the future. This is achieved by recursively predicting next timestep and then using it as an input feature for multistep prediction. We use two autoregressive models: (i) linear AR model and (ii) random forest.

A **random forest** [4] is an ensemble of decision trees. The total response of the random forest model is the average prediction of all decision trees. In order to increase the variety of the individual decision trees, each tree is trained on a bootstrapped sample from the original dataset and uses only random subsets of the features for each decision.

Sequence-to-sequence model [16] consists of two recurrent neural networks (RNN): encoder and decoder. The encoder receives as inputs all available time series features. The initial hidden state is given by the final hidden state of the encoder. In our architecture, the output of the decoder at each timestep serves as input for a fully connected network with one hidden layer that outputs the predicted value for the next timestep. If context or content features are used, those features serve as additional inputs to the fully connected network. The predicted value is then used as the input at the next timestep. As a loss function, we use the sum of squared errors between the predicted values and the real values. Our sequence-to-sequence model implementation is based on the gated recurrent unit (GRU) [7], a variant of RNN.

5 EVALUATION AND RESULTS

5.1 Dataset

In this paper, we have done the evaluation of our real-time model with 23535 articles published between 2018-12-02 00:00:00 and 2018-12-09 00:00:00, and all tweets mentioning those articles. We want to estimate confidence intervals of the performance of the different models instead of just obtaining point estimates. Therefore, we use bootstrapping to generate 100 new datasets consisting of 2000 samples from the validation set.

5.2 Set-up

For the baseline model, we choose a linear interpolation of the most recent 3 time series features. The linear autoregressive model is evaluated for orders 1, 3 and 5. The random forest autoregressive models are all of order 3 with 50 and 500 estimators. The sequence-to-sequence model has a hidden state size of 200 for the encoder and the decoder. The hidden dense layer consists of 200 units. The network is then trained in an end-to-end fashion, using back-propagation with training batches of size 64. The baseline model is only provided the time series features. The autoregressive models and the sequence-to-sequence model are trained using the time series, content and context features.

5.3 Evaluation metrics

The goal of our prediction is to extract the most popular cryptocurrency related articles. Therefore, our evaluation focuses on the top k articles with most mentions on Twitter. We use standard *mean absolute percentage error* (MAPE) to measure the accuracy of the predicted number of mentions and *normalized discounted cumulative gain* (NDCG) to measure the quality of the induced ranking of articles. MAPE computes by how many percents the predicted value deviates from the actual value on average, while NDCG value of 1 indicates the correct ranking.

We vary the prediction start time to be 5, 10, 15 or 20 hours after publication time while keeping the target prediction time fixed at $\delta = 24$ hours after the publication.

5.4 Overall performance

In Fig. 4 we see that all models make better predictions, the closer the prediction start time is to the target time. After 15 and 20 hours from the publication time, even the baseline model already achieves very good performance with MAPE of 20%. At prediction start times of 15 and 20 hours, the random forest (RF) and sequence-to-sequence (S2S) model achieve a lower MAPE than the baseline.

At prediction start time 5 hours after publication advanced models achieve a significantly lower MAPE than the baseline. RF and S2S model achieve a MAPE around $30 - 40\%$, while the linear model is at about 45% and the baseline at 70%. For predictions starting 10

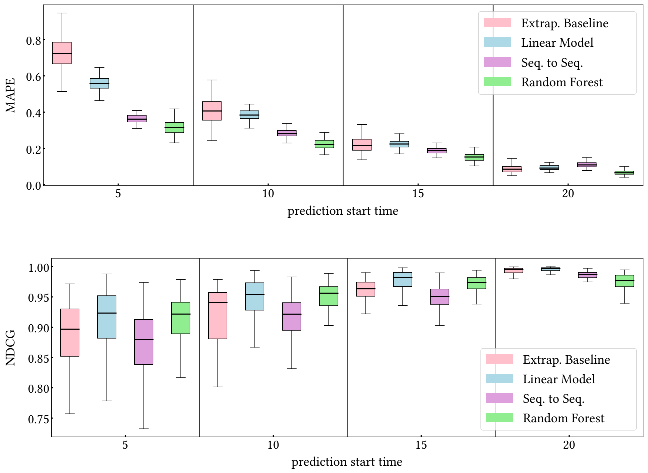


Figure 4: MAPE and NDCG of the different models evaluated on the test data set. The quantiles are determined using 100 bootstrap samples. Predictions are evaluated at 4 different prediction times, 5, 10, 15, and 20 hours after publication of an article.

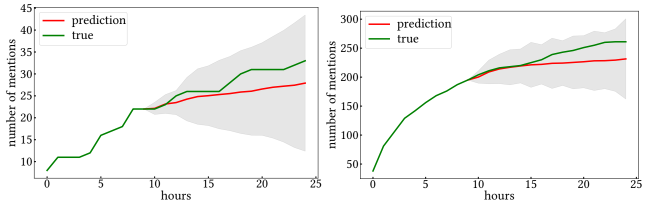


Figure 5: Predictions of the RFAR model for two articles from the test set. The shaded area shows 95% prediction intervals that are determined from the distribution of the predictions by the ensemble estimators.

hours after publication, the baseline and the linear model improve significantly over their performance 5 hours after publication. However, RF and S2S model still achieve a significantly lower MAPE. It is instructive to look at the NDCG as well. Here, the baseline model achieves an NDCG of around 0.9 only 5 hours after publication.

In addition to achieving the best model performance in our experiments, the RF model also gives us a natural way to quantify the **prediction uncertainty**. Instead of just calculating the mean of the ensemble predictions, we can calculate percentiles of the predictions to get prediction intervals with 95% coverage. This is shown in two example time series in Fig. 5.

Finally, the trained AR model is **deployed** to do online predictions on real-time data. The data extraction server is deployed on the Google Cloud, which constantly retrieves new tweets and articles and finds the tweet-article matchings. The matchings are saved as a new batch into a document database, deployed on Amazon Web Services. The online news popularity prediction is visualized in an interactive webpage[1], which also provides the open sourced datasets collected by the pipeline.

[1] Link to the webpage: http://cryptodatathon.com/ranknews

6 CONCLUSION

In this paper, we introduce an online data mining system relating cryptocurrency news to the tweets discussing them. This data pipeline paves the way for monitoring cryptocurrency news of public's interest, identifying and predicting popular news, and tracking public opinions towards cryptocurrencies.

Data exploration on the collected paired news articles and tweets characterized top publishers, and top cryptocurrencies discussed on Twitter. We also perform preliminary predictive analytics using machine learning and deep learning models. This work is a first step towards providing a prediction system, that detects articles that are going to become popular shortly after they are published.

Acknowledgments

N.A.-F. and T.G. are grateful for financial support from the EU Horizon 2020 projects: SoBigData under grant agreement No. 654024. J.B., R.H. and D.L are grateful for the support of Professor A. Krause on the Data Science Lab 2018 course at ETH Zurich.

REFERENCES

[1] Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Clara Vega. 2007. Real-time price discovery in global stock, bond and foreign exchange markets. *Journal of International Economics* 73, 2 (2007), 251–277.
[2] Ioannis Arapakis, B Barla Cambazoglu, and Mounia Lalmas. 2014. On the feasibility of predicting news popularity at cold start. In *International Conference on Social Informatics*.
[3] Roja Bandari, Sitaram Asur, and Bernardo A Huberman. 2012. The pulse of news in social media: Forecasting popularity. *ICWSM* 12 (2012), 26–33.
[4] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[5] Sunandan Chakraborty, Ashwin Venkataraman, Srikanth Jagabathula, and Lakshminarayanan Subramanian. 2016. Predicting socio-economic indicators using news events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1455–1464.
[6] Hailiang Chen, Prabuddha De, Yu Jeffrey Hu, and Byoung-Hyoun Hwang. 2012. Customers as Advisors: The Role of Social Media in Financial Markets. *SSRN Electronic Journal* (2012).
[7] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv e-prints* abs/1412.3555 (2014).
[8] Hongjian Dou, Wayne Xin Zhao, Yuanpei Zhao, Daxiang Dong, Ji-Rong Wen, and Edward Y Chang. 2018. Predicting the Popularity of Online Content with Knowledge-enhanced Neural Networks. In *ACM KDD*.
[9] Tian Guo, Albert Bifet, and Nino Antulov-Fantulin. 2018. Bitcoin Volatility Forecasting with a Glimpse into Buy and Sell Orders. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE.
[10] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 261–269.
[11] Ryota Kobayashi and Renaud Lambiotte. 2016. Tideh: Time-dependent hawkes process for predicting retweet dynamics. In *Tenth International AAAI Conference on Web and Social Media*.
[12] Nuno Moniz, Luís Torgo, and Fátima Rodrigues. 2014. Resampling approaches to improve news importance prediction. In *International Symposium on Intelligent Data Analysis*.
[13] Matija Piškorec, Nino Antulov-Fantulin, Petra Kralj Novak, Igor Mozetič, Miha Grčar, Irena Vodenska, and Tomislav Šmuc. 2014. Cohesiveness in Financial News and its Relation to Market Volatility. *Scientific Reports* 4, 1 (2014).
[14] Bichen Shi, Georgiana Ifrim, and Neil Hurley. 2014. Insight4news: Connecting news to relevant social conversations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 473–476.
[15] Robert H. Shumway and David S. Stoffer. 2017. *Time Series Analysis and Its Applications*. Springer International Publishing.
[16] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*.
[17] Manos Tsagkias, Maarten De Rijke, and Wouter Weerkamp. 2011. Linking online news and social media. In *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 565–574.
[18] Wenbin Zhang and Steven Skiena. 2010. Trading strategies to exploit blog and news sentiment. In *In Fourth Int. Conf. on Weblogs and Social Media (ICWSM)*.