Bitcoin Value and Sentiment Expressed in Tweets

Bernhard Preisler* **Margot Mieskes†** **Christoph Becker†**
University of Applied Sciences Darmstadt
Germany
†firstname.lastname@h-da.de

Abstract

In recent years, traditional economic models failed to forsee several developments resulting in a considerable economic crisis. Other phenomena, such as the increase in Bitcoin value cannot be completely modeled by these traditional means either. As Bitcoin and other cryptocurrencies are a playground for technically interested people, it might be worthwhile to look into other communication channels, such as Social Media to find clues for the development we observe. We hypothesize that sentiment expressed in, for example, might model the development of Bitcoin value better than traditional models. In this work, we present a data set of Tweets covering almost one year, which we annotated for Sentiment. Additionally, we show results from preliminary experiments which support our hypothesis that sentiment information is highly predictive of the value development.

1 Introduction

Financial markets sometimes exhibit tendencies, Keynes (1936) describes as *Animal Spirits* and in the past, traditional models failed to support all that was observable from an economic point of view. Kindleberger (1978) was able to show already in 1978 in the context of a financial crisis that opinions and beliefs of investors are related to news and journal articles. This is especially true for cryptocurrencies such as Bitcoin, which showed a rather erratic behaviour in the past two years. To get new insights into market behaviour, we decide to use Twitter and evaluate whether Tweets can give us more information on the currency's behaviour than traditional models. The hypothesis behind this is, that people investing in Bitcoin might also voice their opinions and/or beliefs through Social Media channels, such as Twitter and therefore influence the market on a subjective level. To that end, we collect Tweets and perform a sentiment analysis on them. Our main question is whether sentiments expressed in Tweets correlate with the value of the crypotcurrency. Our preliminary results indicate that the degree of sentiment does strongly correlate with the development of the currency and that information found in Tweets could improve traditional economic models.

Our major contributions are[1]:

- A dataset of Tweets related to Bitcoin.
- A subset of the main data set that was manually annotated for sentiment.
- An evaluation of various off-the shelf machine learning methods to automatically classify sentiment in Tweets.
- A preliminary analysis of the development of sentiment in Tweets in correlation to the development of the value of the cryptocurrency Bitcoin.

The paper is structured as follows: Section 2 gives an overview on the relevant related work. In Section 3 we describe the data collection and manual annotation. In Section 4 we describe the machine learning and baseline methods used and the features extracted from the data. Section 5 presents the results and their discussion and we finalize the paper with our conclusions and some pointers for future work in Section 6.

*The author was doing his final thesis at the University of Applied Sciences Darmstadt.

[1]The data set and its annotations are available at https://github.com/mieskes/BitcoinTweets

2 Related Work

Work on Sentiment analysis is available in abundance and reviewing the whole field is beyond the scope of this paper. Mäntyläki et al. (2018) present a survey on the topic of sentiment analysis by looking at over 6000 publications, of which 99% were published after 2004. Therefore, we focus on work that was most influential for us.

Gonçalves et al. (2013) look into methods for assigning sentiment to five data sets. They test various methods, including lexicon-based approaches. Their results indicate, that machine learning works best for Twitter.

With respect to sentiment analysis of Twitter the SemEval tasks are of specific interest. Results from the 2016 installment (Nakov et al., 2016), especially subtask A "Message Polarity Classification" and subtask B "Classification to a two-point scale" show that accuracy ranges from 0.646 for the best team to 0.342 for the baseline on Task A. For Task B the accuracy is at 0.862 for the best system and 0.778 for the baseline.

In 2017 the subtask A aimed at a three-point classification (positive, negative and neutral), while subtask B was the same as in 2016 (Rosenthal et al., 2017). Results are again in the range of 0.651 (accuracy) for the best system. The baseline is annotating all Tweets into either positive, negative or neutral and results range from 0.193 for the case, where everything was labeled as positive to 0.483 for labeling everything as neutral.

For 2018 the tasks changed slightly to look at emotions and valence. The annotation for the valence task was done on a 7-point scale, ranging from *very positive mental state* to -3 *very negative mental state*.[2]

With respect to Bitcoin, Kim (2014) analysed comments in a Bitcoin Forum in order to predict the value development of the currency. The author uses data from three years and analyses the comments for sentiment. Using machine learning, the author models the comments and the currency development based on 90% of their data and test the resulting model on 10% of the data. The accuracy is at 80% correct for the prediction of currency value based on comments.

[2] https://competitions.codalab.org/competitions/17751

3 Data Collection and Annotation

As Bitcoin values evolve rapidly, we assume that a medium that allows for rapid communication, such as Twitter more closely reflects the development of the currency.

From Twitter we extract Tweets with relation to Bitcoin, by identifying them through their respective hashtags, such as #bitcoin, #btc, #cryptocurrency etc. We collected data from January 2018 until August 2018 and restricted our collection to English Tweets only, to reduce the chance to have a mixed-language data set. The total data set contains over 50 million Tweets[3].

Figure 2 shows how often Hashtags related to Bitcoin and cryptocurrencies occur in our data set. We observe that only approximately 17% of the Tweets are actually marked with *bitcoin*, while a lot of Tweets refer to other cryptocurrencies or deal with general topics related to them, such as *mining*. To reduce the data set we removed duplicate Tweets, as identified by their ID and also retweeted Tweets.

3.1 Preprocessing

We perform a range of preprocessing steps inspired by Martínez-Cámara et al. (2013) in order to extract features and feed the data to the machine learning algorithms. These preprocessing steps included filtering for stop words, removal of hashtags, UserIDs and URLs within the Tweets. The remaining data only contains plain text.

3.2 Annotation

To be able to train a machine learning model, we need training data. We handed slightly less than 2000 Tweets to human annotators via Amazon Mechanical Turk to annotate them for sentiment.

Figure 1 shows the task description and the annotation interface as displayed on Amazon Mechanical Turk. We coloured the various levels of sentiment for ease of use. In the description, we refer to positive sentiment as indication of rising value and negative sentiment as indication for dropping value of the currency. Apart from the plain text, Turkers did not get any meta data on the Tweets.

Each Tweet is annotated by 7 Turkers and results were averaged. Average values ≥ 0.15 are considered positive Tweets, ≤ -0.15 are considered

[3] The set of Tweet IDs are available at https://github.com/mieskes/BitcoinTweets

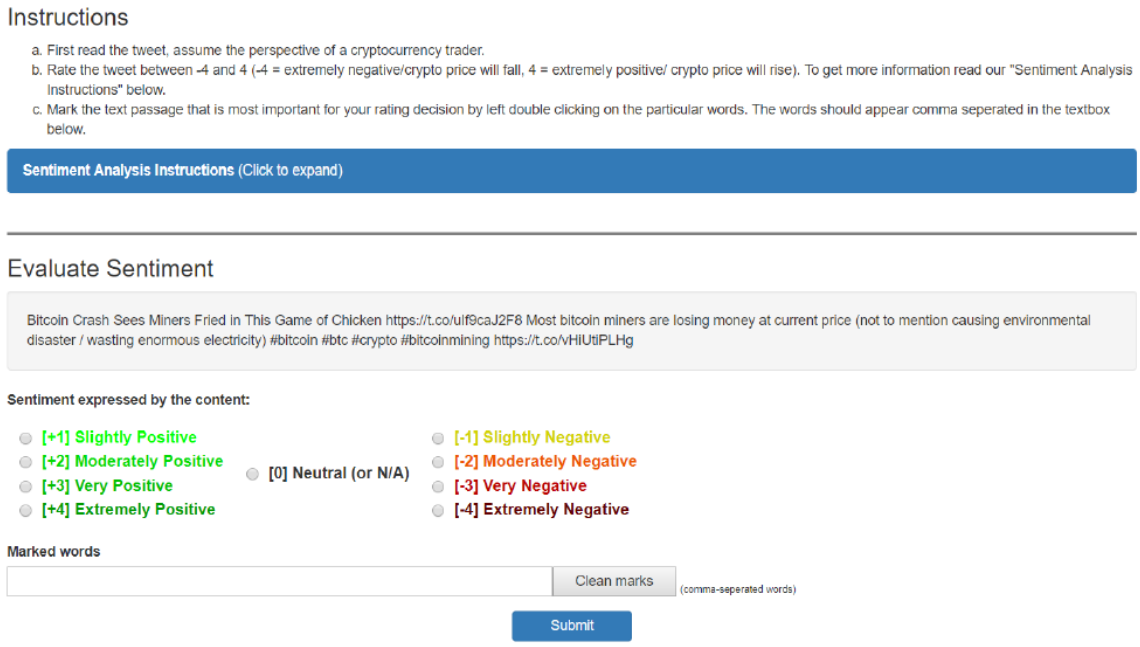

Figure 1: Task description and Annotation Interface for Amazon Mechanical Turk.

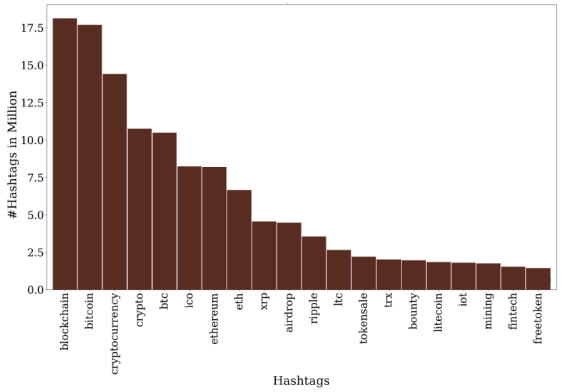

Figure 2: Distribution of Hashtags in the Data set

negative Tweets and results in between are considered neutral. Our final training data set contains 1042 positive Tweets, 727 negative Tweets and 88 neutral Tweets.

We evaluate the annotation quality using Krippendorffs α. As expected, the inter-annotator agreement for the full distinction is fairly low ($\alpha = 0.13$). As we are primarily interested in positive, negative and neutral sentiment, we collapsed the annotations to represent only the three main classes (Details are described above). Nevertheless, the result ($\alpha = 0.43$) was considered improvable. A more detailed look at the annotation revealed, that in some cases individual annotators annotated the complete or near opposite of what the

mean	Text	Sent
-2.3	@SilverBulletBTC Damn, and I can not buy . . .	-1
0.4	Gauthier-Mohammed: I will be a father of . . .	1
-3.4	Oh my! So many #scam these days . . .	-1
1.7	New #Blockchain marketplace Repayment . . .	1

Table 1: Exemplary Tweets including average and mapped sentiment classifications.

majority had done. We identified these instances and removed them from consideration. This left us with enough annotations to create a gold standard on it and raised the inter-annotator agreement to $\alpha = 0.53$, which, considering the complexity of the taks, is a good result.

Table 1 shows example Tweets from the training data. The first column shows the average sentiment value based on all annotations and the last column shows the mapped sentiment classification.

Figure 4 shows the distribution of tweets annotated with a specific sentiment class. We observe, that more tweets receive a positive classification, while fewer receive a negative classification. Most tweets are annotated as *Moderately* or *Very Positive*, while on the negative side, the various subclasses are more evenly distributed. It is interesting to note, that very few tweets are marked as *Extremely Negative*, while on the positive side, a considerable amount of tweets are marked as *Extremely Positive*. This indicates, that most tweets are positive, up to the degree of being enthusiastic.

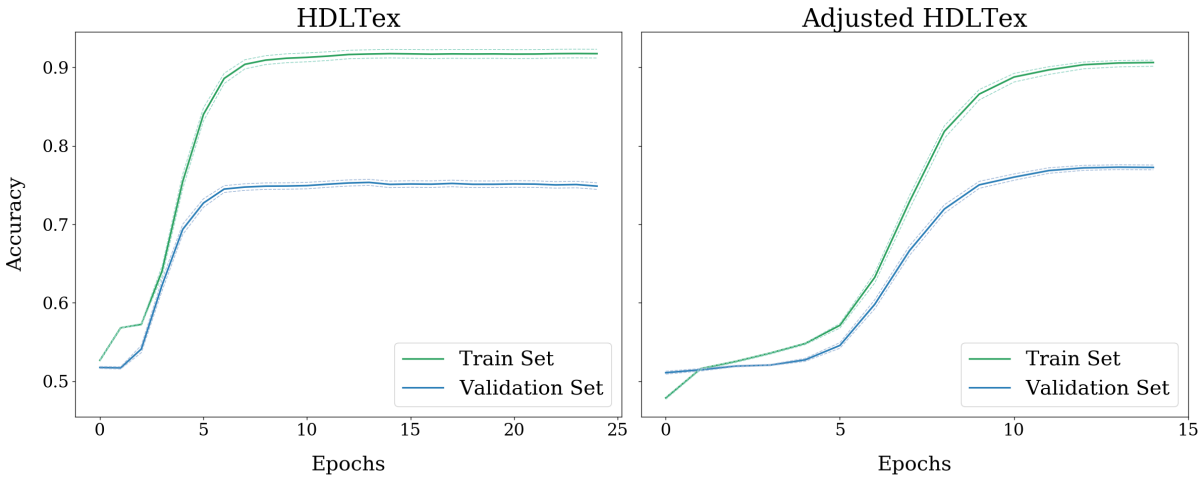

Figure 3: Development of the model using traning and test data both for the original HDLTex and the modified HDLTex architecture.

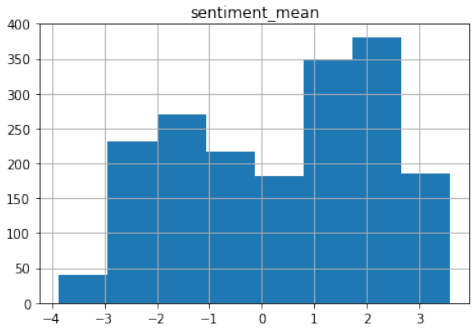

Figure 4: Distribution of manual annotations into the various sentiment classes.

4 Sentiment Classification

We experiment with a range of machine learning methods – both classical and deep learning-based. We use SVMs and Random Forest in addition to two deep learning based methods, which we describe in the following.

4.1 Baselines

We employ two baseline systems in our experiments. Hutto and Gilbert (2014) describe vadersentiment[4] as a lexicon and rule-based sentiment analysis tool, which is specifically targeted towards Social Media. On Social Media the authors achieve an overall F_1 score for the classification of positive, negative and neutral sentiment of 0.96. The tool is implemented in Python.

Sentimentr[5] is implemented in R and is also lexicon-based. The implementation is tested on three different review datat sets (Amazon, Yelp and IMDB) and achieve accuracy rates between 76.5% for the Amazon Review data set and 71.5% for the Yelp data set.

4.2 Machine Learning Approaches

We also experiment with various machine learning approaches. Two serve as baselines and are traditional machine learning systems, while two are deep-learning based.

4.2.1 Baselines

We use Random Forest and Support Vector Machines (SVM) in their implementation in R using standard features. Using a GridSearch and 10-fold cross-validation, we experimentally determine the best parameters for both SVM and Random Forest and use them to classify the data.

4.2.2 HDLTex

The Hierarchical Deep Learning for Text Classification has been developed specifically for text classification (Kowsari et al., 2017). In its original implementation it contains an Artifical Neural Network (ANN), a Convolutional Neural Network (CNN) and a Recurrent Neural Network (RNN).

We experimentally adapt the model with respect to the various parameters. Most importantly, we increase the drop out to 65% and use only 15 epochs.

Figure 3 shows how the accuracy of the models using the original (left side) and modified (right

[4] https://github.com/cjhutto/vaderSentiment

[5] https://github.com/trinker/sentimentr;https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf

Method	Class 1 F_1	Class -1 F_1	Accuracy
CNNSC	0.79	0.86	0.80
ad. CNNSC	0.86	0.89	0.85
HDLTex	0.69	0.81	0.75
ad. HDLTex	0.75	0.83	0.77
RandomForest	0.73	0.86	0.73
SVM	0.79	0.83	0.79
vaderSentiment	0.85	0.90	0.85
setimentr	0.80	0.87	0.79

Table 2: Results for the various automatic sentiment annotation methods examined.

side) HDLTex architecture develop. We see that both methods reach the plateau measured in accuracy between 5 to 10 epochs.[6] But the modified HLDTex architecture achieves a higher accuracy on the test data than the original HDLTex architecture.

4.2.3 CNNSC

We use the Convolutional Neural Network for Sentence Classification (CNNSC) by Kim (2014) with pretrained Word2Vec-based Vectors from the Twitter domain.

Similar to the HDLTex we experimentally create a modified architecture, which uses fewer epochs (20), more filters (128) and a higher drop out rate (75%).

Figure 5 shows how the models using the original (left side) and modified (right side) CNNSC architecture develop. While the original architecture shows a somewhat "bumpy" start in the first 5 epochs, the learning curve for the modified architecture is considerably smoother. Furthermore, the modified CNNSC achieves a higher accuracy both in the training and the test data.

5 Results

In the following we present results for the sentment classification and the relation of the sentiment index to the development of the Bitcoin value.

5.1 Sentiment Classification

Table 2 shows the results for the various machine learning methods and the two baselines we used (see Section 4) for details. We observe that all methods are fairly close together in terms of F_1 and overall accuracy. For the negative class, the modified CNNSC achieves the best results, while for the positive class vaderSentiment achieves the best results. Both methods perform similarly with respect to overall accuracy. This lack of difference between the two methods might be due to the comparably small data set used for training and that a larger data set might boost the performance of the deep learning-based system. Overall, our results are comparable to what has been reported in the literature.

Figure 6 shows the unigram features ranked by their importance. We observe that the most predictive unigrams are actually easily associated with positive or negative sentiment. Words like *join* are less clear, but nevertheless rank comparably high for the sentiment classification.

An initial error analysis shows that, as expected, the neutral class, which makes up about 5% of our data set, causes misclassifications. Either because neutral tweets are classified as having positive or negative sentiment or the other way around. Therefore, improving the classification of the neutral class might also improve the overall classification.

5.2 Sentiment Index and Bitcoin Value

In the next step, we apply the adapted CNNSC to the whole data set in order to classify the data from the complete observed time frame. Figure 7 shows the results for the sentiment development in comparison to the Bitcoin value. The index is normalized to range between 0 and 1. The negative value for the sentiment index at the starting point is an artefact due to lack in previous data. We observe that the Bitcoin value constantly dropped during the observed time-frame, with some bumps in between. The sentiment index closely follows this development and reflects it.

In addition, we perform initial experiments using time-series analysis. For this, we look at the development of the sentiment index and the Bitcoin value on a daily basis. These preliminary results indicate that the sentiment index is a highly significant predictor for the Bitcoin value. But as both Twitter and Bitcoin are rapidly developing and changing, it would be interesting to also investigate shorter time-frames, such as half-day or hourly predictions.

6 Conclusion

We presented a data set of Tweets related to Cryptocurrencies. We manually analysed a subset of the Tweets in order to re-train and evaluate various machine learning and off-the-shelf sentiment classification methods. The main question though was to analyse the development of the sentiment

[6] The graph on the right is based on fewer epochs.

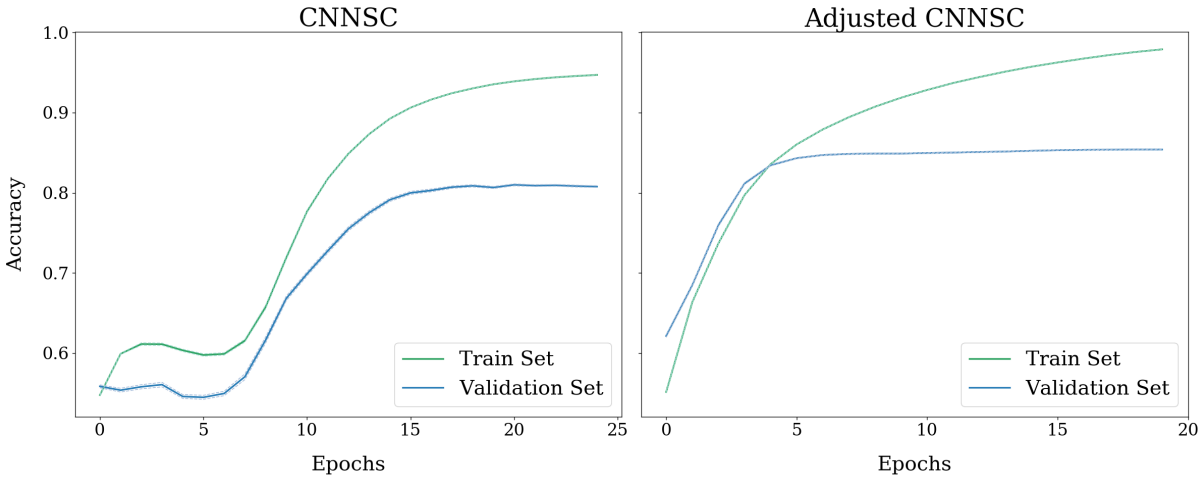

Figure 5: Development of the model using training and test data both for the original CNNSC and the modified CNNSC architecture.

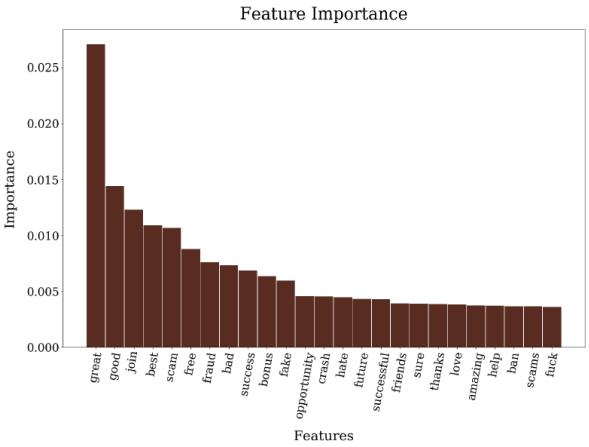

Figure 6: Feature Importance in Sentiment Classification.

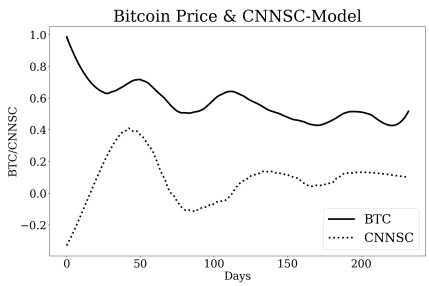

Figure 7: Sentiment and Bitcoin value development in the observed time frame.

expressed in Tweets in relation to the development of the currency's value. We found that off-the-shelf tools perform well enough to automatically analyse this type of data. Moreover, the sentiment index closely reflected the Bitcoin value, which indicates that the analysis of social media data could support current economical models in predicting future developments. Initial results using time-series analysis indicate that the sentiment index is highly predictive of the currency development.

Future Work The first next step is to extend the time-series analysis and evaluate if the predictions also hold on a shorter time-frame (i.e., half-day or hourly predictions). Additionally, looking not only at sentiment, but also at emotions and especially extreme emotions might provide additional information.

We currently only looked at positive, negative and neutral sentiment. Extending this to cover the whole annotated range could give additional improvement on the prediction and the currency value development. Finally, it would be interesting to evaluate whether these findings also hold in other areas of economics. Work by (Soo, 2018) on the american housing market indicates that analysing textual data with respect to economical data could improve current models.

Acknowledgments

This work was supported by the research center for Applied Computer Science (FZAI) and the Faculty for Mathematics and Natural Sciences, University of Applied Sciences Darmstadt.

References

Pollyanna Gonçalves, Matheus Araújo, Fabrício Benevenuto, and Meeyoung Cha. 2013. Comparing and combining sentiment analysis methods. In *Proceedings of the First ACM Conference on Online Social Networks*. ACM, New York, NY, USA, COSN '13, pages 27–38. https://doi.org/10.1145/2512938.2512951.

C.J. Hutto and Eric Gilbert. 2014. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. pages 216–225.

John Maynard Keynes. 1936. *The General Theory of Employment, Interest and Money*. Springer.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 1746–1751. https://doi.org/10.3115/v1/D14-1181.

Charles Kindleberger. 1978. Manias, Panics, and Charts: A History of Financial Crises. *Oxford University Press* .

K. Kowsari, D. E. Brown, M. Heidarysafa, K. Jafari Meimandi, M. S. Gerber, and L. E. Barnes. 2017. HDLTex: Hierarchical Deep Learning for Text Classification. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. pages 364–371. https://doi.org/10.1109/ICMLA.2017.0-134.

Mika V. Mäntyläki, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysis – A review of research topics, venues, and top cited papers. *Computer Science Review* 27:16 – 32.

Eugenio Martínez-Cámara, Arturo Montejo-Ráez, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López. 2013. SINAI: Machine Learning and Emotion of the Crowd for Sentiment Analysis in Microblogs. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Association for Computational Linguistics, pages 402–407. http://aclweb.org/anthology/S13-2066.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. Association for Computational Linguistics, San Diego, California, pages 1–18. http://www.aclweb.org/anthology/S16-1001.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*. Vancouver, Canada, pages 502–518.

Cindy K. Soo. 2018. Quantifying Sentiment with News Media across Local Housing Markets. *The Review of Financial Studies* 31(10):3689–3719. https://doi.org/10.1093/rfs/hhy036.