MINISTERUL EDUCAŢIEI NAŢIONALE
ROMÂNIA
UNIVERSITATEA DE MEDICINĂ, FARMACIE, ŞTIINŢE ŞI TEHNOLOGIE „GEORGE EMIL PALADE" DIN TÂRGU MUREŞ

Acta Marisiensis. Seria Technologica
Vol. 16 (XXXIII) no. 2, 2019
ISSN 2668-4217, ISSN-L 2668-4217

sciendo

10.2478/amset-2019-0009

CRYPTOCURRENCY – SENTIMENT ANALYSIS IN SOCIAL MEDIA

Tudor-Mircea DULĂU[1], **Mircea DULĂU**[2]
[1]*Robert Bosch SRL Romania*
14 Somesului st., 400145, Cluj-Napoca, Romania
[1]dulau.tudor@yahoo.com
[2]*George Emil Palade University of Medicine, Pharmacy, Science, and Technology of Targu Mures*
38 Gheorghe Marinescu st., no. 38, 540139, Targu Mures, Romania
[2]mircea.dulau@umfst.ro

Abstract

The paper proposes the exploration, identification and development of a Java solution for extracting the sentiment related to the cryptocurrencies phenomenon, from the content of the posts of certain popular social networks.

Detecting the positive, neutral or negative character of the sentiment is adopted as a relevant method of establishing the nature of the human perception on the topical issue defined by cryptocurrencies.

Keywords: cryptocurrency, sentiment analysis, Stanford CoreNLP, IBM Watson

1. Introduction

Social platforms have gained a great deal of popularity in recent years, primarily due to the natural and continuous human desire for interaction and the ease with which these virtual connections can be formed and maintained.

The social networks communities currently offer a vast source of information and opinions on any topic, exploiting this data becoming a priority in many research and language processing activities.

The virtual currency represents an alternative means of exchange, consisting of various types of decentralized cryptocurrencies, each one having its particularities and value. The information flow on the topic of virtual currency is characterized by diversity, the nature of the opinion and sentiment fluctuating based on many variables.

From a technical standpoint, the sentiment analysis is part of the natural language processing and involves training a classifier on a labeled corpus, in order to identify and establish attributes / features which can afterwards be used in processing new input data and evaluating its nature.

Because the language is interpretable, dominated by ambiguities and assumptions, and subjected to many specific individual reactions, perceptions, insinuations and idioms, it is extremely imprecise. However, due to the transmission of such a large and important meaning, capturing it is essential. In general, identifying this signification is performed much better by an individual than by a computer.

However, contrary to the predisposition to inconsistencies, unclarities, gaps and contradictions, the comprehension and transmission of language can be characterized to a great extent by accuracy when using appropriate processing systems.

Based on these, the study proposes a solution for analyzing the sentiment in texts extracted from two social networks: Twitter and Reddit, that are related to cryptocurrencies. For this, two important processing tools are used.

One of these, Stanford CoreNLP, is a concentration of tools dedicated to natural language processing and analysis. It was developed by Stanford Natural Language Processing Group using Java programming language and is used both in research areas focused on natural language processing, and in commercial and government applications that need such functionalities.

Another system, IBM Watson, is a highly evolved cognitive one, based on applying innovative concepts on the existing foundations in artificial intelligence. With a great experience in the development of processing techniques of any kind, this tool can be considered among the most important performance benchmarks.

In Section 2 of the paper it is presented a short review of studies from the literature, on natural language analysis and processing, including the architectures.

In Section 3, the pipeline, the architecture and the components of the system are explained.

The results obtained after analyzing the content of the texts with Stanford CoreNLP system and using IBM Watson, are detailed in Section 4. The conclusions are provided in the final section.

2. State of the art

The existing literature gathers a relatively great number of studies on manipulating social networks data in order to perform different language processing algorithms. However, the number of high performance systems that can be utilized in such purposes is reduced.

Within [1], there is presented the pipeline of processing procedures that Stanford CoreNLP has, the method of use, the utility, the constitutive elements and the annotations based technological models composing the language processing system.

Article [2] extensively presents the methods and algorithms used by the Stanford CoreNLP system in order to analyze the sentiment in a text. Here, it is introduced the concept of Treebank Sentiment, a corpus of parsed texts, in the form of a tree, with the structure of a syntactically and semantically labeled and annotated sentence.

In [3] are presented the capabilities of the IBM Watson system, whose combination in the working process produces efficient solutions. These capabilities are natural language processing, by understanding the complexities of unstructured data, generating and evaluating hypotheses, by applying analytical techniques to evaluate and propose answer sets based only on relevant evidence, and dynamic learning, by improving the result based learning process, to become more intelligent with each iteration and interaction.

The architectural idea described by [3] and identified in [4] states that after deciphering the text of the question and the potential answers in the corpus, the system examines the question in relation to the semantic context in hundreds of ways. Then, it uses the results to get a degree of confidence in interpreting the question and the potential answers.

According to article [5], the analysis of the tweets performed to extract the opinion about the topics referred to in their text, represents a good indicator of popularity. Among the eloquent examples mentioned are the tracking of trends by companies to adapt products and services to current needs. The authors propose a classic classification of feelings into positive, negative and neutral categories. The data from Twitter is taken in real time and classified into one of the categories, using the previously classified tweets as learning data. The classification mechanism is based on the Naive Bayes (NB) algorithm.

In [6], it is shared the vision on the importance of the microblogging platform as a source for sentiment analysis. The higher the number of posts containing views on different products, services, affinities and topics, the greater the value and necessity of Twitter, as a data source for processing opinions.

The system's architectural typology found in most of the works described is the one identified in [7].

In article [8], the authors try an analysis that differs in the fact that the selected texts have been translated into English, being initially in other languages, and manually categorized into four categories: positive, negative, neutral and irrelevant. The approach proposed in [9] uses a mix between three different datasets to establish the corpus, each one having the Twitter platform as the source.

In [10], a distinct and unique study idea is presented. The way of working aims to identify the opinion about the most famous and influential users of the platform. The authors of paper [11] intend to identify if the cryptocurrency related data from the online environment can be used to develop suitable and advantageous trading strategies. Using Twitter as the source for the data corpus and Bitcoin as the study currency, the authors try to apply different learning options to obtain results regarding the evolution of the currency. Article [12] is dedicated to studying the influence of the online communities discussions on the evolution of the digital money market of cryptocurrencies. The possibilities and methods of the Stanford CoreNLP tool are made available to users in the form of a library that is described in [13] and supplemented with the examples in [14].

The capabilities of the IBM cognitive system are available to developers through various external services that include sets of operations. The appropriate API in this situation is the one described in [15] and dedicated to natural language processing operations.

3. Execution pipeline. System architecture

To achieve the objectives, the approach consists in extracting data related to cryptocurrencies from several sources, filtering and storing it, respectively applying suitable text processing algorithms on it.

Initially, in order to fulfill the extraction of the most important currencies, an endpoint of a specialized platform called Coin Market Cap is interrogated. The external services provided by Twitter and Reddit platforms are then repeatedly used to retrieve texts and information about the previously gathered currencies.

For the sentiment analysis there are two methods applied, corresponding to the two tools used, Stanford CoreNLP and IBM Watson.

Fig. 1 presents the abstract architecture of the execution pipeline used to obtain predictions related to the nature of the sentiment in the extracted texts. An abstract general architecture of the components and the interactions between them is illustrated in Fig. 2, in which the dependencies of the software system's elements can be identified.

The first module (Fig. 1) is the one for composing the data corpus, exposing the constituent for extracting the coins and the one for extracting the data from social platforms.

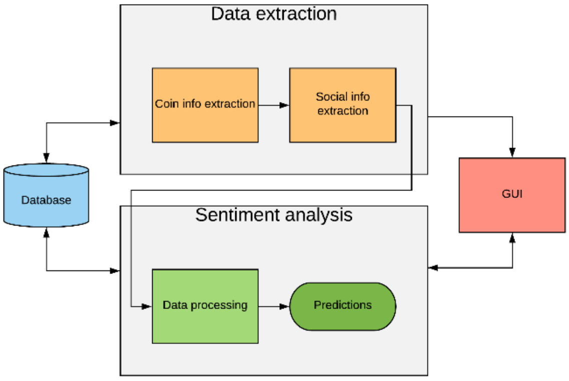


Fig. 1: Simplified system architecture

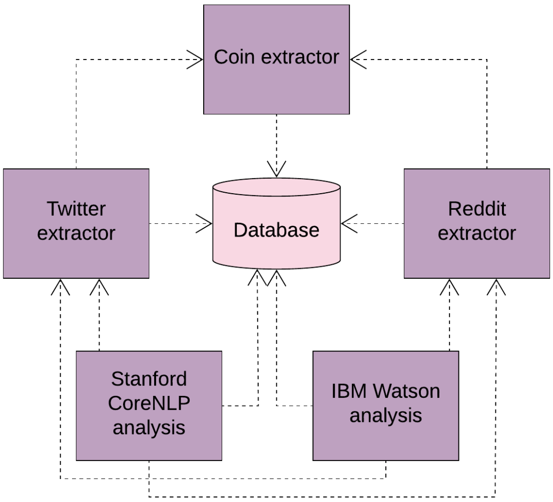


Fig. 2: System components

The module for sentiment analysis contains the data processing component for classifying the text, which provides the results.

Because in the proposed application the tweets are required, the Search API described by [16] is necessary.

Also, Reddit platform's API, illustrated in [17], offers developers a very wide range of endpoints that can be used, each having different functions: from posting, creating or editing posts, comments, messages, personal information or account information, to searching and obtaining different items such as subreddits, posts, users, accounts, messages, conversations, names, etc. after a multitude of criteria specified in the request.

The predictions consist of different information resulting from the analysis, namely the type of the general sentiment of a text, the related class, the score distribution between classes, the type of the sentiment in relation to a certain currency, the type of the general emotion in the sentence, its class, the type of the emotion that targets a certain currency etc.

The programming language used to implement the system is Java. All the analysis results are saved in a database.

To manipulate the data collection stored in the MySql system, the tools provided by Hibernate are utilized. For the actual implementation, the Java libraries provided by Stanford CoreNLP and IBM Watson are used.

4. Tests and results

Text searching cannot be accomplished without a set of representative cryptocurrencies names. Thus, all the social network selections will have in the content the name of a currency.

The specialized platform called Coin Market Cap offers the ranking of the cryptocurrencies on the capital market.

The Twitter platform represents a very rich source for the cryptocurrencies topic. Therefore, its posts will be subjected to the analysis of the project's text classifiers.

In Fig. 3 there are presented, for guidance purposes, the top 5 coins on the platform [18], respectively the market share that represents the ordering criterion. In Fig. 4, there is illustrated the large number of daily tweets regarding each of the first three cryptocurrencies in the ranking indicated in Fig. 3, for a period of one year [19].

The developed algorithm involves adapting the processing results obtained using the Stanford CoreNLP software tool, in situations where certain terms in the jargon of the cryptocurrency domain influence the nature of the sentiment in an undetected manner.

The following data was collected through the extraction components:

- number of cryptocurrencies: 300;
- number of texts having Twitter as source: 6323;
- number of texts having Reddit as source: 1008 (838 posts and 170 comments).

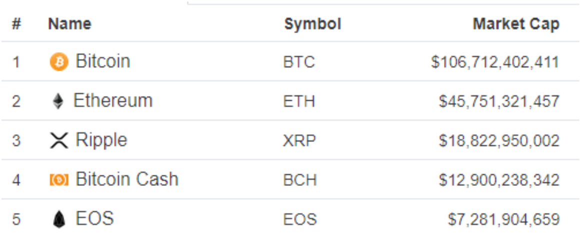
The difference in the amount of texts extracted from the two platforms is due to the increased popularity Twitter has and the fact that Reddit applies filters on its content and has moderators.

The results obtained after analyzing the contents of 7331 texts using the Stanford CoreNLP system, respectively 7327 texts using IBM Watson, are presented in Table 1.

The texts from Twitter were chosen to indicate the number of negative sentiments found in each day for a period of approximately one and a half months.

The statistics identify the results of both types of processing and are shown in Fig. 5.

Similarly, the positive ones from the same period of time are highlighted in Fig. 6.



#	Name	Symbol	Market Cap
1	Bitcoin	BTC	$106,712,402,411
2	Ethereum	ETH	$45,751,321,457
3	Ripple	XRP	$18,822,950,002
4	Bitcoin Cash	BCH	$12,900,238,342
5	EOS	EOS	$7,281,904,659

Fig. 3: Ranking of cryptocurrencies based on market share

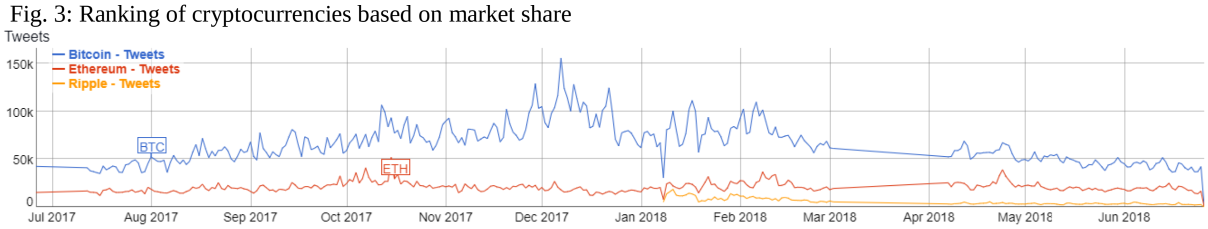


Fig. 4: Number of daily posts about top cryptocurrencies

Table 1: Number of texts in each sentiment class

Sentiment / Analysis	Very negative	Negative	Neutral	Positive	Very positive
Stanford CoreNLP	4	2905	3295	1105	22
IBM Watson	-	1954	3650	1723	-

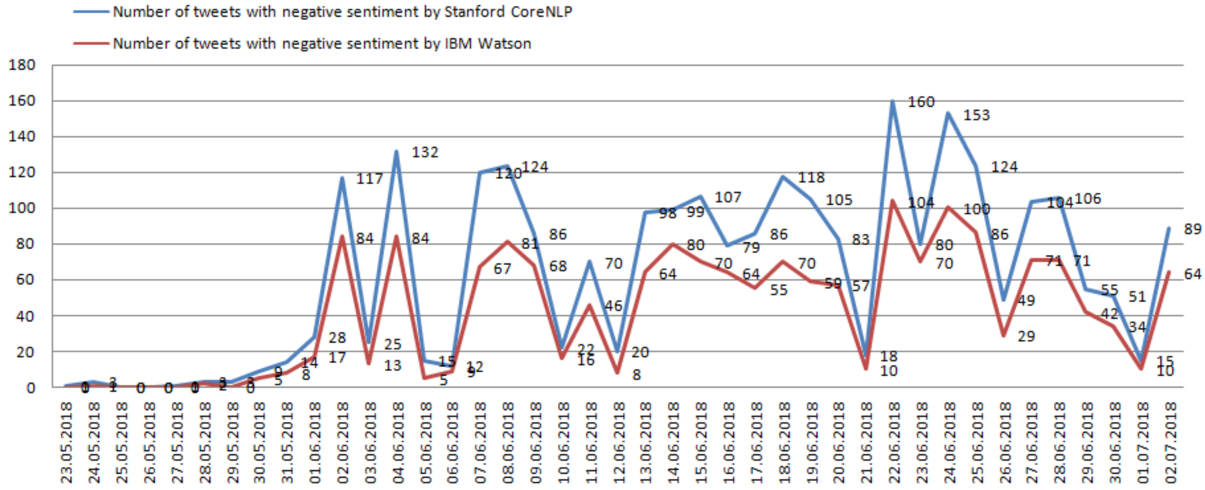


Fig. 5: Number of tweets with negative sentiment

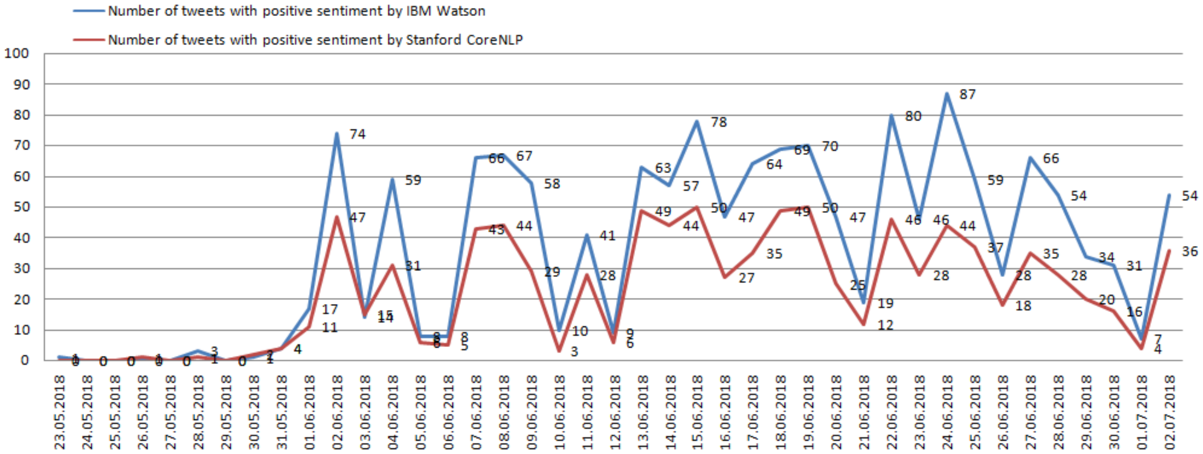


Fig. 6: Number of tweets with positive sentiment

5. Conclusions

The paper aims to analyze the sentiment that characterizes the texts posted by the users of social networks regarding the phenomenon of cryptocurrencies. By interrogating the services provided by the platforms, it was possible to select the coins of interest for the study and the actual data based on which the analysis is done, and by integrating and deriving the different language processing techniques, the nature of the sentiment was identified.

The platform utilized for collecting cryptocurrencies is Coin Market Cap, and those whose services were called for the information destined for processing are Twitter and Reddit. Tweets and details about tweets were collected from Twitter and from Reddit there were extracted subreddits, which facilitated subsequent selection of posts and comments.

After a complex data filtering, it is transmitted to the analyzers in the required format and using the specific methods. To identify the sentiment in each text, the Stanford CoreNLP and IBM Watson systems were used, intervening on the first one with an adaptation based on the specific terms of the cryptocurrency domain that influence the sentiment in an undetected way.

The proposed solution for capturing the additional information about the sentiment and modifying scores accordingly has led to much more accurate results than the standard approach that exclusively uses the Stanford CoreNLP library.

From the perspective of the origin of the data used, comparing the source with verified and moderated data (Reddit) to the one with uncontrolled data (Twitter), shows that the approached methodology provides close results. This fact proves that the proposed solution has low sensitivity to the data source. If a comparative analysis is made on the obtained results based on the amount of data used for each case study, it is found that the proposed method is predominantly invariant.

As a future extension of the project, for a better adaptation of the language processing results to the cryptocurrency jargon, the list of special terms influencing the sentiment needs to be completed and updated. Also, the results could be integrated in a machine learning system as an additional parameter for identifying the potential growths of the main currencies and the investment options.

References

[1] Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S. and McClosky, D. (2014), *The Stanford CoreNLP Natural Language Processing Toolkit*, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, (ACL 2014), pp. 55-60.

[2] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C., Ng, A. and Potts, C. (2013), *Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank*, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), vol. 1631, pp. 1631-1642.

[3] High, R. (2012), *The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works*, Redbooks and IBM Corporation, vol. 1.

[4] Ferrucci, D. (2012), *Introduction to "This is Watson"*, IBM Journal of Research and Development, vol. 56, pp. 1-15.

[5] Kushwanth, R., Sachin, A., Shambhavi, B. and Shobha, G. (2014), *Sentiment Analysis of Twitter Data*, International Journal of Advanced Research in Computer Engineering and Technology (IJARCET 2014), vol. 3, no. 12, pp. 4337-4342.

[6] Pak, A. and Paroubek, P. (2010), *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*, Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10), vol. 10, pp. 1320-1326.

[7] Dalmia, A., Gupta, M. and Varma, V. (2015), *Twitter Sentiment Analysis. The good, the bad and the neutral!*, Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pp. 520-526.

[8] Agarwal, A., Xie, B., Vovsha, I., Rambow, O. and Passonneau, R. (2011), *Sentiment Analysis of Twitter Data*, Proceedings of the Workshop on Languages in Social Media (LSM'11), pp. 30-38.

[9] Kouloumpis, E., Wilson, T. and Moore, J. (2011), *Twitter Sentiment Analysis: The Good, the Bad and the Omg!*, Proceedings of the 5th International AAAI Conference on Weblogs and Social Media, pp. 538-541.

[10] Younggue, B. and Hongchul, L. (2012), *Sentiment Analysis of Twitter Audiences: Measuring the positive or negative in influence of popular twitterers*, Journal of the American Society for Information Science and Technology, vol. 63, no. 12, pp. 2521-2535.

[11] Colianni, S., Rosales, S. C. and Signorotti, M. (2015), *Algorithmic Trading of Cryptocurrency Based on Twitter Sentiment Analysis*, CS229 Project, pp. 1-5.

[12] Kim, Y. B., Kim, J. G., Kim, W., Im, J. H., Kim, T. H., Kang, S. J. and Kim, C. H. (2016), *Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies*, PLoS One, 11(8), e0161197.

[13] Stanford CoreNLP (2018) [Online]. Available: https://stanfordnlp.github.io/CoreNLP/

[14] Stanford Sentiment Analysis (2018) [Online]. Available: https://nlp.stanford.edu/sentiment/

[15] Natural Language Understanding (2018) [Online]. Available: https://www.ibm.com/

watson/developercloud/natural-language-understanding/api/v1/

[16] Search Tweets (2018) [Online]. Available: https://developer.twitter.com/en/docs/tweets/search

[17] Reddit Api (2018) [Online]. Available: https://www.reddit.com/dev/api

[18] CoinMarketCap (2018) [Online]. Available: https://coinmarketcap.com

[19] BitInfoCharts (2018) [Online]. Available: https://bitinfocharts.com