A Sentiment Analysis Approach for the Cryptocurrency Market and Blockchain Technology Using Naïve Bayes, Support Vector Machine and Random Forest

Denisa Elena Bălă[1] and Stelian Stancu[2]
[1)2)] Bucharest University of Economic Studies, Bucharest, Romania.
E-mail: baladenisa16@stud.ase.ro; E-mail: stelian_stancu@yahoo.com

Please cite this paper as:
Bălă, D.E. and Stancu, S., 2022. A Sentiment Analysis Approach for the Cryptocurrency Market and Blockchain Technology Using Naïve Bayes, Support Vector Machine and Random Forest. In: R. Pamfilie, V. Dinu, C. Vasiliu, D. Pleșea, L. Tăchiciu eds. 2022. *8th BASIQ International Conference on New Trends in Sustainable Business and Consumption.* Graz, Austria, 25-27 May 2022. Bucharest: ASE, pp.209-214.
DOI: 10.24818/BASIQ/2022/08/027

Abstract

Virtual currencies or cryptocurrencies are based on Blockchain technology, also known as distributed ledger technology. As of March 2022, there are already over 10k virtual coins, their number being continuously growing since 2013. This paper aims to extract the public sentiment expressed towards the cryptocurrency market and Blockchain technology, two topics widely debated in the last decade. Our research was based on the use of Twitter data, collected with the help of an API in the RStudio environment. To identify the sentiment associated with the 5,000 tweets collected, we used the Bing lexicon approach and three supervised learning algorithms. The three classifiers are the Naive Bayes classifier, a Support Vector Machine and Random Forest algorithm. The accuracy of these algorithms was analyzed through four metrics, finding that the Random Forest classifier proved to be the most accurate, while the SVM algorithm offers the weakest results in terms of classification. The sentiment analysis conducted with the help of the Bing lexicon indicated a predominantly positive sentiment of online users regarding the cryptocurrency market and Blockchain technology. The present paper is structured as follows. The first part highlights a brief introduction to the issue of text mining analysis, as well as the area of supervised learning. Subsequently, a revision of the specialized literature in the approached subject is highlighted, by referring to some pertinent studies in this direction. The methods used as well as the data involved in this study are described in the chapter dedicated to research methodology. The paper continues with the presentation of the main results of the research, as well as with the highlighting of the conclusions and of some future research directions. The study concludes with an exposition of bibliographic sources.

Keywords

sentiment analysis, supervised learning, Blockchain technology, cryptocurrency market, Twitter data.

DOI: 10.24818/BASIQ/2022/08/027

Introduction

An increasing number of scientific researches have been undertaken using sentiment analysis in a wide variety of fields (health, commerce, business, social sciences etc), given that human emotions, feelings and opinions are determinants of the individuals' behavior and psychology. Recognized in the scientific literature under various names (such as sentiment analysis, opinion mining, emotion analysis or opinion extraction) this technique has developed rapidly in the last decade, with the growing popularity of social media platforms (from simple opinion blogs to trending platforms such as Twitter, Facebook or Youtube) and with the interactions manifested through these platforms: comments, reviews, forums etc. Thereby user-generated content is the foundation of this analysis technique, with researchers trying to use that content to discover and extract useful knowledge. Twitter is one of the most popular social media platforms and the opinions and emotions expressed through this platform are key elements for big brands and businesses that are trying to better understand their audience.

A subcategory of machine learning and artificial intelligence, supervised learning is characterized by learning various models to provide the desired outputs. Supervised learning algorithms use labelled data which is a dataset where the observations have already been assigned to different classes. One of the most popular supervised learning algorithms is the SVM (Support Vector Machine) algorithm, used in both classification and regression problems. In the classification problems based on SVM, the concept of hyperplan is used, the algorithm trying to identify a decisional boundary according to which the analyzed entities can be correctly distributed into classes. Another supervised learning algorithm is the Naïve Bayes classifier, which is known to be very effective. Used in classification problems, this algorithm is based on Bayes' theorem and seeks to identify each time the probability of an object or entity or the probability that a particular object or entity falls in a predefined class. One of the most popular and flexible supervised learning methods is the Random Forest algorithm. It works on the principle of combining several decision trees, each being trained on a set of observations. Finally, the prediction is made taking into account the average of the predictions made by each tree.

1. Review of the scientific literature

Cui et al. (2006) study the performance of three classifiers in terms of product evaluation using 100k reviews posted online. The three algorithms analyzed are Passive-Aggressive (PA) Algorithm Based Classifier, Language Modeling (LM) Based Classifier and Winnow, being observed an increased performance of the discriminant algorithm (Passive-Aggressive (PA) Algorithm Based Classifier).

A comparative study of four classification algorithms is performed by Vinodhini and Chandrasekaran (2013). The purpose of the analysis is to compare the performance associated with the KNN classifiers, Decision Trees, Naïve Bayes and Support Vector Machine. Three distinct sampling methods are used, namely the linear sampling, the bootstrap sampling and the random sampling. They conclude that the SVM classifier with bootstrap sampling proves to be the one with the highest accuracy.

Mihalcea et al. (2013) make a comparison regarding the accuracy of the Naïve Bayes classifier, respectively SVM using a number of 140 records representing false and true statements with the help of unigrams from the word model. The two techniques indicate accuracy levels in the range of 52-73%, taking into account or not the stop words.

Abkenar et al. (2021) addresses the issue of detecting spam messages on the Twitter social media platform. They use a hybrid technique based on two strategies, namely Synthetic Minority Over-sampling Technique (SMOTE) and Differential Evolution (DE) which aims to improve spam detection. The technique proves to be useful and effective in increasing the performance of spam detection in the case of unbalanced samples.

Musleh et al. (2022) develop a model with the aim of analyzing the sentiment expressed by Arab users of the Twitter platform. Specifically, they seek to identify depression among users and classify tweet messages into three categories: "depressed", "non-depressed" or "neutral". They use and compare the accuracy of six classifiers, namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), AdaBoost and Naïve Bayes (NB), finding that the highest performance in the classification process corresponds to the Random Forest classifier, of approximately 82.39%.

2. Research methodology

The purpose of this paper is two-fold. On the one hand we will try to capture the sentiment expressed by Twitter users regarding Blockchain technology and the cryptocurrency market and on the other hand we will make a series of comparisons regarding the accuracy of some classification algorithms. In the present research, when we refer to the cryptocurrency market, we actually focus on the top ten cryptocurrencies at the time of writing this article, the ranking being made according to the market capitalization associated with each digital asset. These cryptocurrencies are: Bitcoin (BTC), Ethereum (ETH), Tether (USDT), BNB (BNB), USD Coin (USDC), XRP (XRP), Terra (LUNA), Cardano (ADA), Solana (SOL) and Avalanche (AVAX).

We will therefore use as a starting point a collection of 5000 tweets accessed in March 2022 using an API provided by Twitter platform for the RStudio environment. In order to identify the short tweet messages relevant to our study, we included a number of keywords in the data query section that we provided as parameters for the tweet extraction function.

The primary stage of the research was data preprocessing. Data cleaning involves completing steps such as converting all characters into lowercase, removing punctuation marks, removing extra blanks, deleting numbers or various symbols and removing the stop words. Once the pre-processing stage is completed, the sentiment analysis will be continued, more precisely with the extraction of the public sentiment expressed by Twitter users with the help of the Bing lexicon. Lexicons are the basis for sentiment analysis and represent dictionaries or collections of words, in which each word is associated with a particular feeling. In the case of the Bing lexicon, each term can be associated with either a positive feeling or a negative feeling.

Each analyzed tweet will be associated with a feeling, so each message will be later associated with a label denoting the public feeling: positive or negative. Three models of supervised learning will be built later, in order to make comparisons on the accuracy of classifying messages into positive tweets, respectively negative tweets. The three proposed classification algorithms are Naïve Bayes (NB), Support Vector Machine (SVM) and Random Forest (RF).

The Naïve Bayes classifier is based on Bayes' theorem, which gives the conditional probability of an event A, given another event B, that is:

$$P(A|B) = \frac{\mathrm{P(B|A)} * P(A)}{P(B)} \quad (1)$$

Where:

$P(A|B)$- represents the conditional probability of A, given B;
$P(B|A)$ - represents the conditional probability of B, given A;
$P(A)$- represents the probability of event A;
$P(B)$- represents the probability of event B;
In the case of SVM the goal is to identify a hyperplane in an N-dimensional space that classifies entities in a distinct manner. In the separation of these entities there are many possible hyperplanes that could be chosen, the objective being to identify a hyperplane that corresponds to a maximum distance between the points of the two classes. The example of a SVM classifier with RBF kernel function can be represented as follows:

$$K(A,B) = \exp\left(-\frac{||A-B||^2}{2\sigma^2}\right) \quad (2)$$

Where:

σ is the variance and the hyperparameter
$||A-B||$ represents the Euclidean distance between points A and B.
Used in both regression and classification problems, the Random Forest algorithm is based on the concept of ensemble learning by which several classifiers are combined in order to solve a complex problem and at the same time to improve the accuracy of the model.

3. Results and discussion

This section will highlight on the one hand the results of the sentiment analysis undertaken using the 5,000 tweets collected with reference to the cryptocurrency market and Blockchain technology and on the other hand will present a comparison of three classification algorithms in terms of accuracy.

The dataset consisting of the 5,000 tweet messages was subjected to some preprocessing techniques, so that the URLs, punctuation marks and special characters were removed from the text. At the same time, the content of the text was converted to lowercase and additional spaces were removed. Stop words have also been removed, representing terms that have no meaning in the context of our analysis, so deleting them does not impact this research. Another procedure was tokenization, through which sentences were broken down into separate words called tokens. In the context of NLP, the tokenization procedure proves useful in understanding the meaning of the text by simply analyzing the words that make up the text. Also, the stemming procedure was applied, through which all the words are brought to their root form. Practically, the suffix and the prefix are eliminated, this technique of reducing the dimensionality in the NLP context helping to reduce the associated computations.

The figure below (Figure no. 1.) is a word cloud representation, a key element in the context of sentiment analysis. The word cloud is actually a common visual representation of text data. The highlighted words differ depending on the color and font size, based on how often they are used in the collected tweets. There is a growing interest in digital assets Solana, Tether and Ethereum, given their high frequency of use. The words "solana", "tether" and "ethereum" stand out in the graphic representation below. At the base of this

graphic representation is the concept of term document matrix, a common approach in Natural Language Processing. The term document matrix indicates the relationship between documents and terms and allows the determination of the frequencies of all terms in the analyzed dataset.

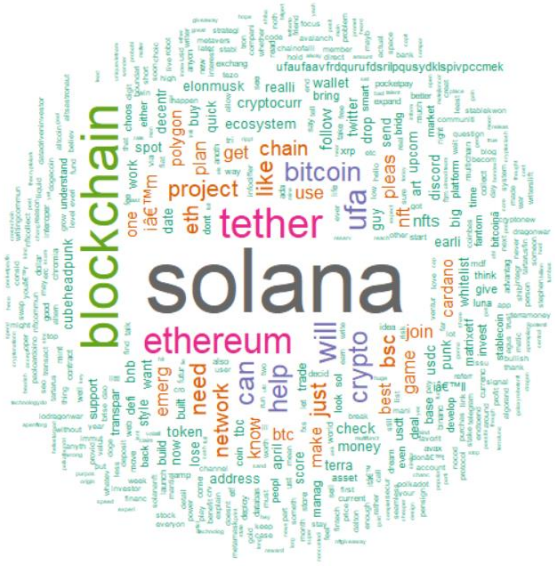

Figure no. 2. Word cloud of the most frequent terms

Source: Authors' processings

The public sentiment expressed about Blockchain technology and cryptocurrencies will be analyzed using the Bing lexicon, which consists of a collection of words associated in a simple way with a positive or a negative sentiment. The contribution of the terms to the negative or positive feeling is highlighted in the following figure. Terms such as "emergency", "lose" or "risk" contribute most to the negative feeling. In terms of positive sentiments, the major contribution is made by terms such as "like", "best", "support" and "work".

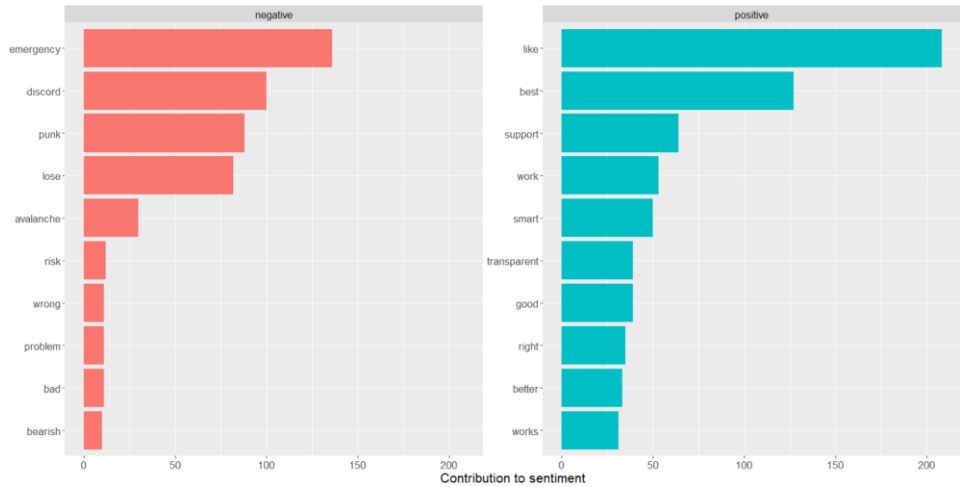

Figure no. 2. Sentiment contribution using the Bing lexicon

Source: Authors' processings

With the help of the Bing lexicon, the collected tweets were classified into negative tweets and positive tweets, a label being assigned to each message. The figure below (Figure no. 3) shows the distribution of the five thousand tweets. It is noticeable that the predominant feeling is positive, so the attitude of online users towards Blockchain technology and the cryptocurrency market is a positive one. Next, using the associated labels, we will proceed to build the supervised learning algorithms.

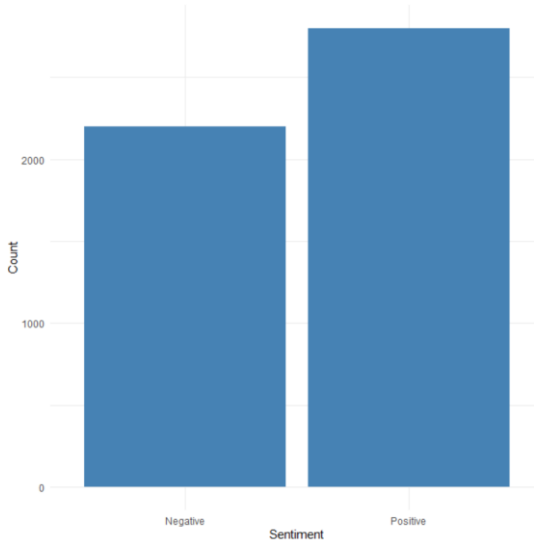

Figure no. 3. The number of positive and negative tweets
Source: Authors' processings

The quality of the classification will be evaluated in terms of Accuracy, Precision, Recall and F1 Score. Accuracy is calculated as the ratio between the number of correctly classified entities and the total number of entities considered.

$$Accuracy = \frac{TP+TN}{TN+FP+TP+FN} \quad (4)$$

Precision of a class is the number of positive entities relative to the total number of items labeled as belonging to the positive class.

$$Precision = \frac{TP}{FP+TP} \quad (5)$$

Another important measure of the classification is represented by Recall, also recognized as true positive rate or sensitivity.

$$Recall = \frac{TP}{FN+TP} \quad (6)$$

The fourth indicator we will focus on is the F1 Score, calculated as the harmonic average between Precision and Recall.

$$F1\ Score = 2 * \frac{Precision*Recall}{Precision+Recall} \quad (7)$$

In order to build the classification algorithms, the dataset was divided into a training set consisting of 70% of the observations and a test set, comprising the remaining 30% of the collected observations. The performance of the classifiers can be assessed based on the metrics highlighted in the table below.

Table no. 1. Accuracy assessment of the three classifiers

Classifier	Accuracy	Precision	Recall	F1 Score
Naïve Bayes	76.32	76.33	77.14	76.73
Support Vector Machine	62.81	62.81	63.01	62.90
Random Forest	81.42	81.43	82.21	81.81

Source: Authors' processings

It is noted according to the results (Table no. 1) that in terms of Accuracy, Precision, Recall and F-Score, Random Forest is the algorithm that offers the best classification solution. The lowest performance is associated with the Support Vector Machine classifier, recording the lowest values for the associated metrics.

Conclusions

Cryptocurrencies and the technology on which they are based, namely the Blockchain technology are innovative compounds of the Fintech industry. Their rapid infiltration on financial markets and the growing

public interest into this subject offer a revolutionizing prospect of these technologies. Sentiment analysis is a popular technique for extracting emotions and feelings expressed through written texts, applied in a variety of fields. This paper aimed to perform a sentiment analysis regarding the cryptocurrency market and Blockchain technology using the approach based on lexicons. Specifically, using Twitter data, we have extracted the public sentiment expressed in the online environment regarding the digital assets market and Blockchain technology. With the help of the Bing lexicon, it was found that the attitude of Twitter users towards the previous mentioned areas is a positive one. The text mining analysis reveals the contribution of terms such as "like", "best", "work", "smart" to the positive feeling, while the negative attitude is denoted by the use of terms such as "emergency", "lose" or "risk", a sign that the prospect of potential losses is discouraging for the Twitter public. Each tweet was later associated with a label denoting the corresponding feeling: negative or positive. Once the labeling was available, we proceeded to build three supervised learning algorithms, namely the Naïve Bayes classifier, Support Vector Machine and Random Forest. The classifiers were then compared in terms of accuracy and the best results were provided by the Random Forest classifier, while the Support Vector Machine was associated with the lowest accuracy. As future research directions, we intend to carry out an extension of the study, taking into account the following elements: the use of a higher number of observations in the dataset as well as the approach of additional classification algorithms. On the other hand, we will also consider performing an intertemporal sentiment analysis in order to identify how public sentiment is changing on the cryptocurrency market and Blockchain technology, identifying at the same time the determinants of these changes.

References

Bazzaz Abkenar, S., Mahdipour, E., Jameii, S.M. and Haghi Kashani, M., 2021. A hybrid classification method for Twitter spam detection based on differential evolution and random forest. Concurrency and Computation: Practice and Experience.

Cui, H., Mittal, V. & Datar, M., 2006. Comparative Experiments on Sentiment Classification for Online Product Reviews. *AAAI*, 6, pp.1265-1270.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1), pp.1–167.

Loughran, T. and McDonald, B., 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), pp.35–65.

Mihalcea, R., Perez-Rosas, V. & Mihai, B., 2013. Automatic detection of deceit in verbal communication, *Association for Computing Machinery*, pp.131-134.

Musleh, D., Alkhales, T., Almakki, R., Alnajim, S., Almarshad, S., Alhasaniah, R., Aljameel, S. and Almuqhim, A., 2022. Twitter Arabic Sentiment Analysis to Detect Depression Using Machine Learning. *Computers, Materials & Continua*, 71(2), pp.3463–3477.

Naeem, M.A., Mbarki, I., Suleman, M.T., Vo, X.V. and Shahzad, S.J.H., 2020. Does Twitter Happiness Sentiment predict cryptocurrency? *International Review of Finance*, 45(1), pp.67–77.

Vinodhini, G. & Chandrasekaran, R.M., 2013. Performance Evaluation of Machine Learning Classifiers in Sentiment Mining. *International Journal of Computer Trends and Technology* (IJCTT), 4(6).