COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 1 P. 185–195
DOI: 10.20537/2076-7633-2023-15-1-185-195

Ки&М

ENGINEERING AND TELECOMMUNICATIONS

UDC: 519.8

Development of and research on machine learning algorithms for solving the classification problem in Twitter publications

I. S. Makarov[a], E. R. Bagantsova[b], P. A. Iashin[c], M. D. Kovaleva[d], R. A. Gorbachev[e]

Moscow Institute of Physics and Technology,
9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russia

E-mail: [a] i.s.m.mipt@yandex.ru, [b] bagantsova.er@phystech.edu, [c] iashin.pa@phystech.edu,
[d] kovaleva.md@phystech.edu, [e] roman.gorbachev@phystech.edu

Received 01.11.2022.
Accepted for publication 23.12.2022.

Posts on social networks can both predict the movement of the financial market, and in some cases even determine its direction. The analysis of posts on Twitter contributes to the prediction of cryptocurrency prices. The specificity of the community is represented in a special vocabulary. Thus, slang expressions and abbreviations are used in posts, the presence of which makes it difficult to vectorize text data, as a result of which preprocessing methods such as Stanza lemmatization and the use of regular expressions are considered. This paper describes created simplest machine learning models, which may work despite such problems as lack of data and short prediction timeframe. A word is considered as an element of a binary vector of a data unit in the course of the problem of binary classification solving. Basic words are determined according to the frequency analysis of mentions of a word. The markup is based on Binance candlesticks with variable parameters for a more accurate description of the trend of price changes. The paper introduces metrics that reflect the distribution of words depending on their belonging to a positive or negative classes. To solve the classification problem, we used a dense model with parameters selected by Keras Tuner, logistic regression, a random forest classifier, a naive Bayesian classifier capable of working with a small sample, which is very important for our task, and the k-nearest neighbors method. The constructed models were compared based on the accuracy metric of the predicted labels. During the investigation we recognized that the best approach is to use models which predict price movements of a single coin. Our model deals with posts that mention LUNA project, which no longer exist. This approach to solving binary classification of text data is widely used to predict the price of an asset, the trend of its movement, which is often used in automated trading.

Keywords: cryptocurrency, Twitter, machine learning, natural language processing, vectorization, dense model, logistic regression, random fores classifier, KNN, naive Bayes classifier

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 1, pp. 185–195.

КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 1 С. 185–195
DOI: 10.20537/2076-7633-2023-15-1-185-195

Ки&М

ИНЖИНИРИНГ И ТЕЛЕКОММУНИКАЦИИ

УДК: 519.8

Разработка и исследование алгоритмов машинного обучения для решения задачи классификации в публикациях Twitter

И. С. Макаров[a], Е. Р. Баганцова[b], П. А. Яшин[c], М. Д. Ковалёва[d], Р. А. Горбачёв[e]

Московский физико-технический институт,
Россия, 141701, Московская область, г. Долгопрудный, Институтский пер., 9

E-mail: [a] i.s.m.mipt@yandex.ru, [b] bagantsova.er@phystech.edu, [c] iashin.pa@phystech.edu,
[d] kovaleva.md@phystech.edu, [e] roman.gorbachev@phystech.edu

Получено 01.11.2022.
Принято к публикации 23.12.2022.

Посты в социальных сетях способны как предсказывать движение финансового рынка, так и в некоторых случаях даже определять его направление. Анализ постов в Twitter способствует прогнозированию цен на криптовалюту. Специфика рассматриваемого сообщества заключается в особенной лексике. Так, в постах используются сленговые выражения, аббревиатуры и сокращения, наличие которых затрудняет векторизацию текстовых данных, в следствие чего рассматриваются методы предобработки такие, как лемматизация Stanza и применение регулярных выражений. В этой статье описываются простейшие модели машинного обучения, которые могут работать, несмотря на такие проблемы, как нехватка данных и короткие сроки прогнозирования. Решается задача бинарной текстовой классификации, в условиях которой слово рассматривается как элемент бинарного вектора единицы данных. Базисные слова определяются на основе частотного анализа упоминаний того или иного слова. Разметка составляется на основе свечей Binance с варьируемыми параметрами для более точного описания тренда изменения цены. В работе вводятся метрики, отражающие распределение слов в зависимости от их принадлежности к положительному или отрицательному классам. Для решения задачи классификации использовались dense-модель с подобранными при помощи Keras Tuner параметрами, логистическая регрессия, классификатор случайного леса, наивный байесовский классификатор, способный работать с малочисленной выборкой, что весьма актуально для нашей задачи, и метод k-ближайших соседей. Было проведено сравнение построенных моделей на основе метрики точности предсказанных меток. В ходе исследования было выяснено, что наилучшим подходом является использование моделей, которые предсказывают ценовые движения одной монеты. Наши модели имеют дело с постами, содержащими упоминания проекта LUNA, которого на данный момент уже не существует. Данный подход к решению бинарной классификации текстовых данных широко применяется для предсказания цены актива, тренда ее движения, что часто используется в автоматизированной торговле.

Ключевые слова: криптовалюты, Twitter, машинное обучение, обработка естественного языка, векторизация, dense модель, логистическая регрессия, случайный лес, KNN, наивный байесовский классификатор

Introduction

Cryptocurrencies are a new type of digital currency, which is very popular nowadays. Bitcoin [Bitcoin, 2022] has the largest market capitalization, and it was the first cryptocurrency ever. Its white paper was published on the 31st of October 2008. After that, a large number of other coins called *altcoins* were created. LUNA(Terra) [Terra, 2022], is one of the altcoins and it is widely discussed in the article.

The success of these projects caused other discussions on social media networks. People usually share their opinions and predictions on cryptocurrency. The most popular social media platform which engaged in crypto discussions is Twitter. Many researches deal with information this platform provides. The research includes machine learning, sentiment analysis etc. Unfortunately, historical data obtained from Twitter is limited due to Twitter API restrictions.

Twitter is a powerful source of expression social activity, sharing ideas and opinions. Evidently, global cryptocurrency community exposes market to movement this way, especially speaking about motile tokens or assets. A conspicuous example of this influence is Elon Musk's tweets affecting the Bitcoin price [itZone, 2022]. Phenomenon manifests existing the whole paradigma of tokens called *memecoins*. An illustration provided by the same ifluencer Elon Mask represents a rapidly growing Dogecoin [Decrypt, 2022].

Related research shows the perks of the approach of analysing Twitter activity [Otabek, Choi, 2022]. The number of followers of the poster, the number of comments on a tweet, the number of likes, and the number of retweets are used there. Our paper doesn't enable to gather such data in a correct way due to the Twitter restrictions problem.

In the P.Kaur, M.Edalati research [Kaur, Edalati, 2022] Twitter data were collected based on prices specific to energy. The markup was formed according to the sentiment of tweets. The approaches explored included such algorithms as Naive Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, and Logistic Regression. Analysing sentiment requires massive manual markup for dataset. In our case there is no need to explore the mood of Twitter users, so we can capture market prices connection with lemmas directly. However, the approaches were studied in this paper as well.

The article describes machine learning and NLP approaches for analysing Twitter messages. The processing includes several steps such as lemmatization, stop words removal, vectorization, model training and testing.

Speaking of approaches of text preprocessing, A. Balahur's study [Balahur, 2013] contemplates such relevant methods as word normalization, lower casing, and tokenization, which are supported in our research. Among other things, user and topic labeling are considered too. Thus, the users mentioned in the tweet, marked with "@" go for "PERSON" and the topics marked with "#" are replaced with "TOPIC". Here we eliminate these elements due to the frequency approach for vectorizing. In another case, these words could be in the top of most frequent lemmas, but truthfully they cannot be the feature to set tweets apart according to their labels. We used Stanza [CoreNLP, 2022; Stanza, 2022] for lemmatization. It is a Python natural language processing package for the linguistic analysis of many human languages. Lemmatization provides an opportunity to decrease overall variability of the text in order to reduce the number of words with the same meanings and the time of further text processing. Stop words removal is an essential step in text processing. Here the set of such words includes articles and pronouns which occur frequently, but ineffective for market movements recognition. For example, the words "is", "are", "and", "or", "we", "I" etc. are removed.

The last step of data preparation is vectorization. In a similar study [Ahmad, AlQurashi, Mehmood, 2022] word2vec [Word2vec, 2022] is utilized to vectorize the data. We decided to move away from the concept of using interpretation of words meaningless in terms of linguistic sense in order to concentrate on the statistical method of text data analysis. In this way we could pay more attention to subcultural features of the community and focus on the specific lexicon. Our methodology includes selecting valuable features which may represent each message as a binary vector. The features are represented by words that were found in the text. Since each message consists of many different words, specific metrics were defined in order to select valuable and useful words. Those metrics are calculated using the frequency analysis approach.

Finally, several models were trained and tested on historical data. The results for the most interesting models are presented in the article.

Training sample data preparation

Data selection

The initial data were gathered using Twitter API v2. It contains over 250,000 elements from 24 February 2021 for users taken from the list of inluencers granted by LunarCrush [LunarCrush, 2022]. The total number of users is 138. Each element of data consists of the following information: the text of a tweet, posting time in the format of ISO 8601 (UTC), author id, tweet id. Twitter restrictions limit the ability of downloading the full package of data, so the number of tweets collected is distributed unevenly in time.

Data preparation

One of the most noticeable features that Twitter could be marked with is special language used. This is the reason to eliminate several types of words to make text data more convenient to work with. We utilize regular expressions to get rid of emoji, hashtags, links, non-English language words, and slang abbreviations of words. Lemmatization is provided by a Python NLP package Stanza. All uppercase letters in a string are also transformed to lowercase.

Data markup

For markup the market data from Binance [Binance, 2022] is taken. In the paper two types of binary classification is explored. Above all, the definition of relevant terms is presented below:

- *hallway* — candlestick [Neeson, 2020] area with determined width and length;
- *hallway length* (N) — a number of candlesticks between the time of tweet creation and the time of price detection;
- *hallway width* (EPS) — half-width of range, in which a tweet is considered as neutral (0) in the in-out binary classification.

In the paper two types of binary classification are explored. One consists of positive (+1) and negative (−1) classes, the another, of neutral (0) and positive (1).

The target for *simple* binary classification is marked according to the difference between two close prices of the candlesticks: a price close of the candlestick, which contains the current tweet creation time, and a price close of the last candlestick in the hallway. Tweet is marked as belonging to negative (−1) class, if the difference is less than zero, and to positive (+1) class otherwise.

In *in-out* binary classification it is checked if the absolute value of the difference divided by price close of the first candlestick in the hallway is more than hallway width. If the value is larger, so we mark it as 1, 0 otherwise.

Vectorization

In order to create models for text processing, the messages obtained are represented as a numerical vector. The vector belongs to *n*-dimensional word space. A basis word is a single word which was chosen by specific rules. The structure of vector used in the article represents a sequence of numbers 0 and 1. Each position in the vector is strictly connected with a basis word. The message is scanned to check the existence of a basis word in the text. If the word is used in the processing text at least once, an appropriate number in the vector equals 1. If the message does not contain the basis word, an appropriate number in the vector is 0.

The dimension of words' space depends on the number of words used. Theoretically, it is possible to create a space consisting of all words that were found in tweets. The method of vectorization provides a full numeric representation of existing messages. However, the dimension of the space becomes extremely large due to the variety of unique words in tweets. Moreover, the system built probably could face a problem if it received a message containing a completely new word.

To diminish the space dimension, a specific part of the set of words can be chosen. Firstly, the lemmatization process is used to transform words in a specific way. The method provides a decrease in the number of words in the entire set of words found in the text. Secondly, specific rules were defined to select the words.

The article encompasses results obtained using different metrics in order to select better lemmas.

Feature selection

The simplest metric to describe the efficiency of lemmas is frequency. It is defined as follows:

$$F = \frac{M}{S},$$

where F is the frequency of lemma, M is the number of times a word was used among all tweets, S is the number of all words used.

The metric seems to be good at describing tweets. The most frequent words are included in almost every message in the sample, therefore appropriate vectors are non-zero. For example, if the lemma space contained less frequent lemmas, many vectors would contain only 0 values, which make it impossible to train a high-quality model using the data.

However, the set of words contains many frequent and insignificant lemmas such as articles and pronouns. Usually, we focus on verbs and nouns. In order to filter the text, as it was already mentioned, Stanza and regular expressions were used.

The frequency values obtained are pictured in Figure 1. The bar chart shows the frequency for each word in the top-50.

During the study of the frequency metric it was recognized that there were more meaningless elements among all lemmas than just articles and pronouns. Messages contain punctuation symbols and words, such as proper names. For instance, according to Figure 1, the most frequent lemma is a name of the investigated coin “luna”. Moreover, some derivatives, such as “lunatic”, are in the top-50 as well.

In this case several approaches were studied:

- lemma space consisted of *n* most frequent lemmas including inappropriate elements;
- lemma space consisted of *n* most frequent lemmas excluding inappropriate elements;
- lemma space consisted of *n* the most frequent lemmas with bias of top beginning;

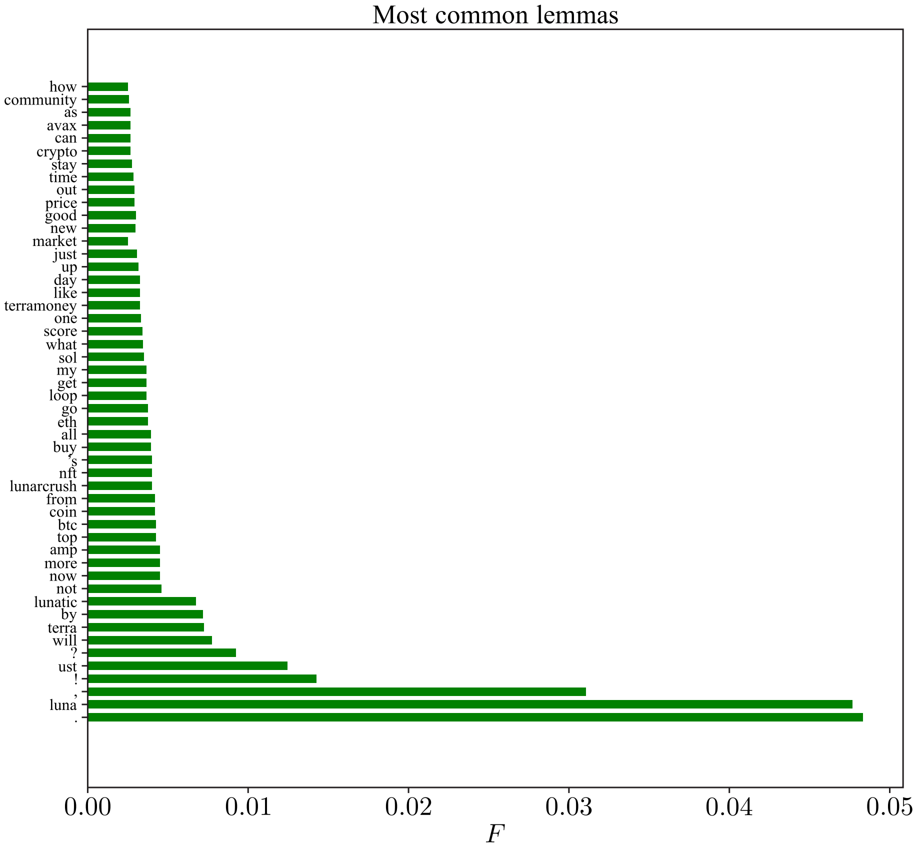

Figure 1. Top-50 of the most frequent words among downloaded tweets of selected users. The bar chart demonstrates the frequency for each word in the top-50

where n is the number of basis lemmas, dimension of space. Spaces with n equal to 50, 100, 150, 200, 250 and 300 were studied.

The next suggested approach in feature selection is the usage of two parameters for each lemma. Since it is undesirable for the lemma space to contain less frequent basis words, the frequency defined above is considered to be the first parameter. The second value describes the quality of a word in predicting the tweet mark. Metrics for simple binary classification are defined as follows:

$$\nu_{pos} = \frac{M_{pos}}{M},$$
$$\nu_{neg} = \frac{M_{neg}}{M},$$

where M is the number of messages in the training sample, M_{pos} is the number of positive (+1) class messages, M_{neg} is the number of negative (−1) class messages. The same metrics may be defined for in-out classification:

$$\nu_{in} = \frac{M_{in}}{M},$$
$$\nu_{out} = \frac{M_{out}}{M},$$

where M_{in} is the number of neutral (0) class messages and M_{out} is the number of positive (1) class messages. Moreover, it is obvious that $M = M_{pos} + M_{neg} = M_{in} + M_{out}$.

To select features, a limit of frequency was set up. The limit was varied during the investigation and lemmas were selected from the elements that satisfied the restrictions. Using the v_i value it is possible to recognize if the maximization of the parameter creates a better vector.

Another suggested method is to define the following metric and to use it as the second parameter:

- For *simple* classification

$$C_{bin} = \frac{M_{pos} - M_{neg}}{M},$$

- For *in-out* classification

$$C_{bin\_hallway} = \frac{M_{in} - M_{out}}{M}.$$

The visual representation of the metric for in-out classification is given in Figure 2.

Blue lemmas predominate in tweets which lead to the exit of the hallway, while red lemmas predominate in tweets that do the opposite. It is suggested to select frequent enough lemmas with maximization of absolute value of C_{bin} to create the vector.

The C_{bin} metric is completely different from the previous one. For example, if v_{pos} is utilized, it is possible to create a vector with a larger value of v_{neg}. Otherwise, C_{bin} does not provide an opportunity to do this, especially if the space dimension is not huge.

In order to use those metrics the limit of frequency is set for features and then it is sorted using values defined above. To select basis lemmas several items were gained from the top of the sorted sequence. Then some lemmas may be added from the other side of the sequence.

Models

The experiment consisted in solving simple and in-out classification problems based on Binance candlestick data with predetermined parameters of N and EPS. We used different types of vectorization mentioned above in order to find the best one according to its accuracy. All values were sorted through just in the way of full-search.

In the course of the research, several types of machine learning methods were studied. The training of models was carried out on 80 % of overall data, the other 20 % were utilized as a test sample, respectively. Vectorized data was randomly shuffled, balance was contemplated as well by altering parameters of N and EPS, so we avoid suffering from Twitter restrictions.

The model consists of 2 fully connected layers. Most customizable options were established by Keras Tuner to achieve the highest value of accuracy. Hence, the activation function of the first layer was the hyperbolic tangent function (tanh) with the dropout coefficient equal to 0.2. The number of the layer's neurons is 144. The second layer implies the sigmoid activation function. Model's summary is given in Table 1.

Table 1. Architecture of a dense model

Layers	Type	Structure of the output tensor	Number of parameters
dense_1	Dense	(None, 144)	11,664
dropout_1	Dropout	(None, 144)	0
dense_2	Dense	(None, 1)	145

For compilation we utilized the RMSprop optimizer and binary crossentropy loss. The total number of parameters is equal to 11,809. The best result was obtained on the data vectorized based on a wordset constructed by the rule of LUNA mentioning with the following parameters:

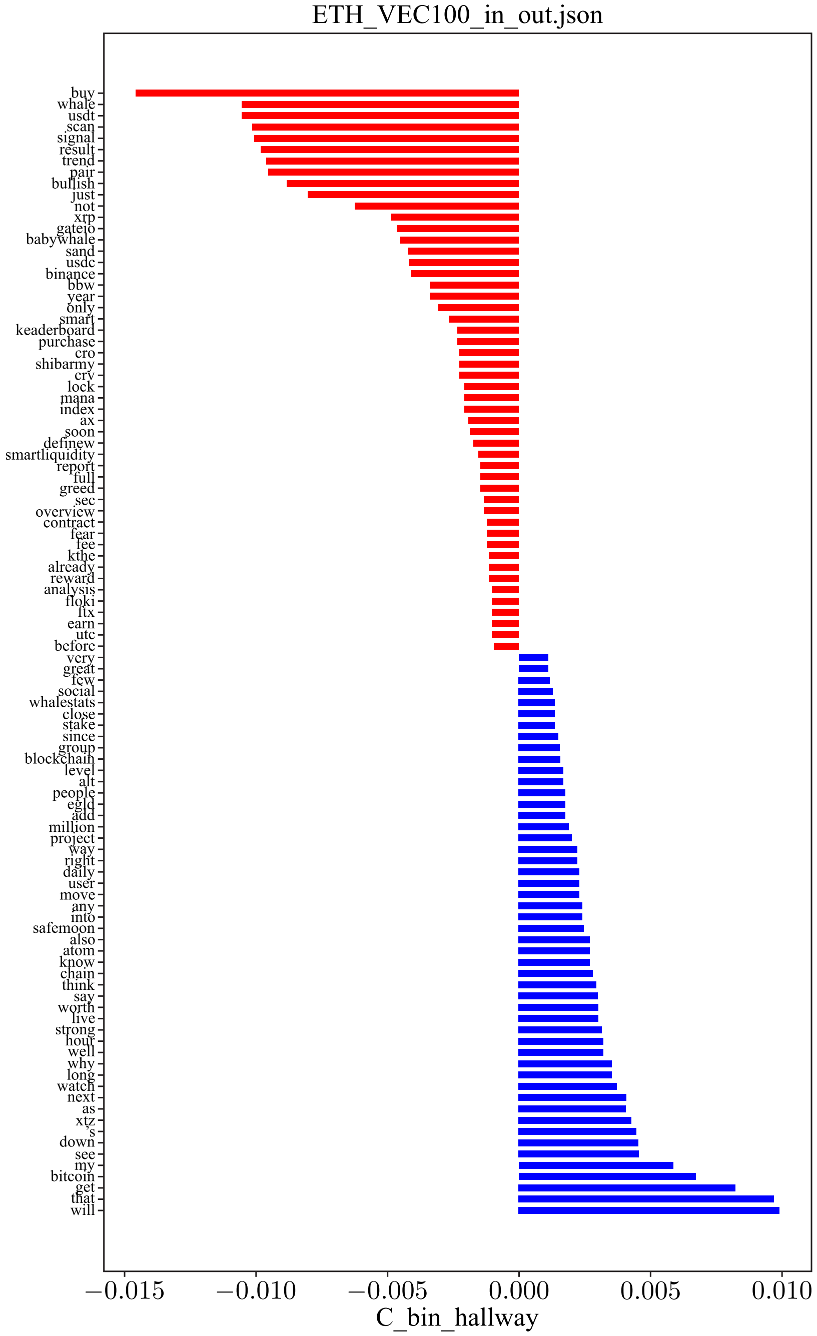

Figure 2. An example of $C_{bin\_hallway}$ metric. The bar chart demonstrates $C_{bin\_hallway}$ metric for classes in the in-out classification problem

- hallway length equals to 5 ($N = 5$);
- word space dimension, i. e., the number of basis words equal to 150 ($n = 150$);
- sequence of lemmas sorted by v_{pos};
- the ratio of lemmas from the top of the sorted sequence to lemmas from the other side is 2:1 ($M_{top} : M_{end} = 2 : 1$).

The distribution of training sample labels is 0.51, and 0.49 for the testing, so the dataset was balanced enough to use accuracy as a metric. The accuracy for simple classification of 0.55 in training and 0.54 on validation corresponds to this configuration. The result given is demonstrated in Figure 3.

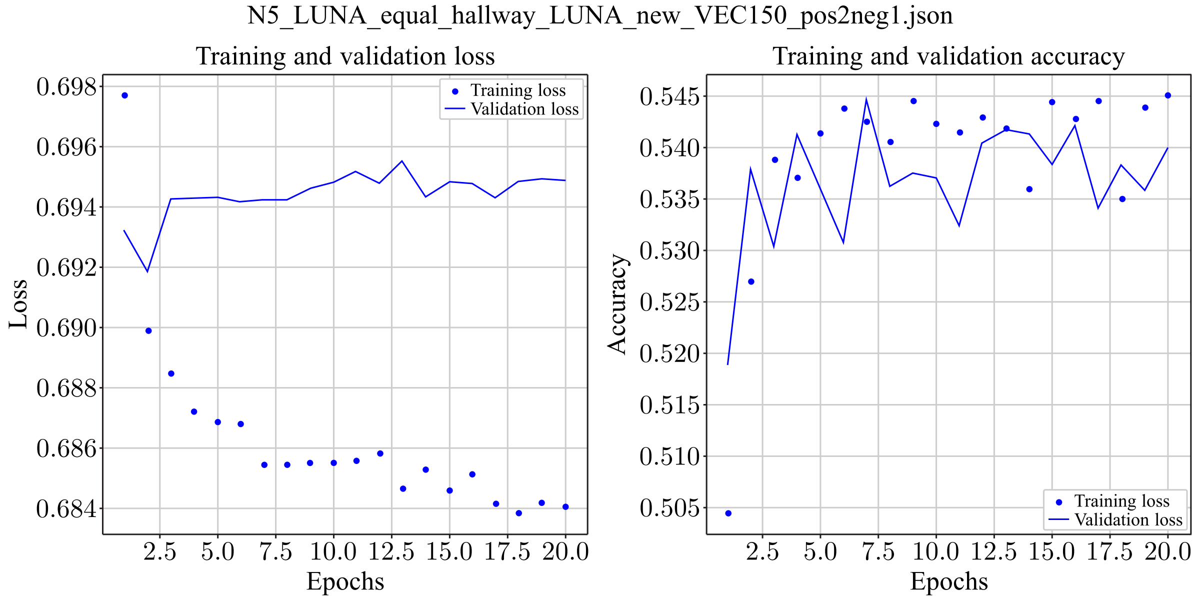

Figure 3. Result for $N = 5$, $n = 150$, sorted by v_{pos} with $M_{top} : M_{end} = 2 : 1$. The vertical axis represents loss and accuracy changing during training and validation processes. The horizontal axis shows epochs. The graph demonstrates the best result for simple classification

It is noticeable that 10 epochs are enough to fit the model. After this value loss undergoes stagnation with an increasing trend which may lead to overfitting.

Speaking of in-out classification, the best results were given using a vector based on dataset constructed by the rule of ETH mentioning with the following parameters:

- hallway length equals 5 ($N = 5$);
- word space dimension, i. e., the number of basis words equals 100 ($n = 100$);
- sequence of lemmas sorted by v_{in};
- the ratio of lemmas from the top of the sorted sequence to lemmas from the other side is 1:1 ($M_{top} : M_{end} = 1 : 1$).

The distribution of training sample labels is 0.53, and 0.46 for the testing. Training accuracy equals 0.56 and testing one to 0.543. that is shown in Figure 4.

We also implemented classic machine learning approaches. The best result taken by logistic regression for simple binary classification is 0.542 of accuracy for vectorizing with the following parameters:

- hallway length equals 5 ($N = 5$);
- word space dimension, i. e., the number of basis words equals 50 ($n = 50$);
- sequence of lemmas sorted by v_{pos};
- basis words selected were on the top of sorted sequence.

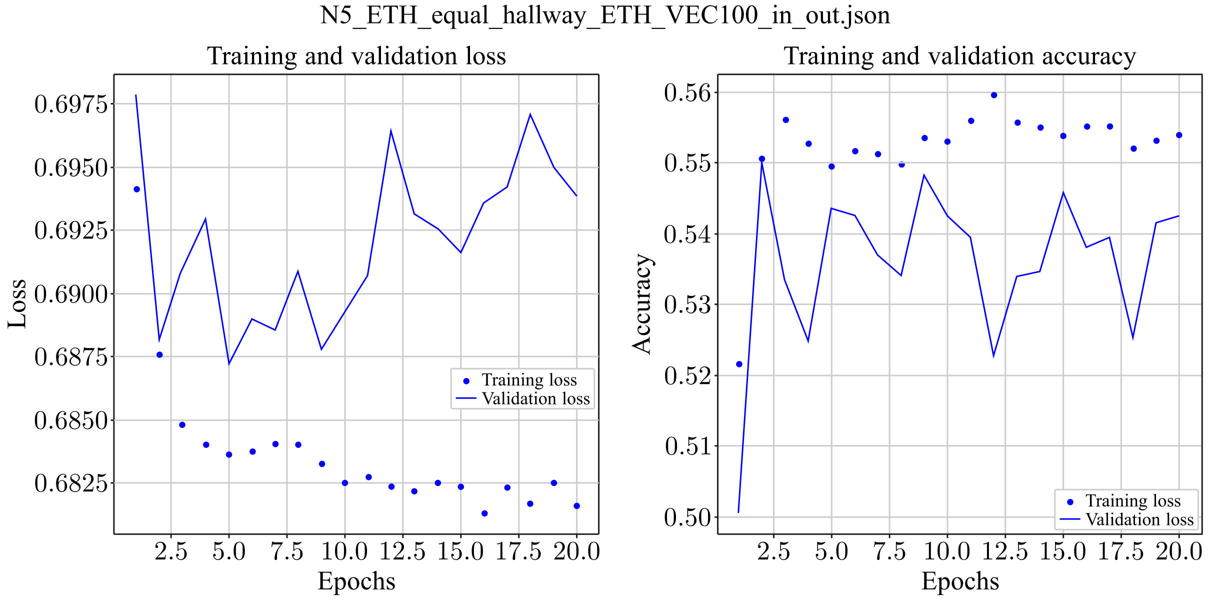

Figure 4. Result for $N = 5$, $n = 100$, sorted by v_{in} with $M_{top} : M_{end} = 1 : 1$. The vertical axis represents loss and accuracy changing during training and validation processes. The horizontal axis shows epochs. The graph demonstrates the best result for in-out classification

For the 150-dimensional vector with sequence sorted by C_{bin} and $M_{top} : M_{end} = 1 : 2$, $N = 5$ alternative methods were explored to solve the simple classification problem as well. Training target distribution here is 0.509 and for the validation it is 0.514. Bernoulli Naive Bayes is considered to be used in the binary classification for a small sample, so it is relevant to the problem we have. This method provides us the value 0.536 of accuracy. The worse result was gained using k-nearest neighbors (KNN) algorithm — only 0.49 of accuracy. The last method represented random forest classifier. We used the full-search to adjust parameters, and concluded the number of estimators equal to 50 and 10 as the maximal death of the trees, Gini index was used for splitting the vertex of the trees. The result given is 0.523 on test. The summary is given in the pivot Table 2.

Table 2. Results for alternative methods in solving the simple classification problem

Method	Accuracy
Bernoulli naive bayes	0.536
KNN	0.49
Random forest classifier	0.523

Conclusion

In this paper we have developed several machine learning models in order to predict cryptocurrencies price movements. Each model receives vectorized messages obtained from Twitter using its API v2. The vectorization was followed by tweets' preprocessing. It includes such steps as lemmatization and stop words removal.

Feature selection was based on frequency distribution of lemmas usage. Metrics are developed in a way of different types of binary classification created with the help of Binance candlesticks. Represented metrics can be used for measuring features weights in the related research. They also contribute to distinguishing classes when solving text classification problems.

The approaches applied are Dense model, Logistic Regression, Random Forest Classifier, Naive Bayes Classifier, k-nearest neighbor method. Bernoulli naive Bayes demonstrated, as expected, one of the best results for the in-out classification problem in comparison with others. The best created model showed 0.54 of validation accuracy and related to LUNA project, which does not exist nowadays.

References

Ahmad I., AlQurashi F., Mehmood R. Machine and Deep Learning Methods with Manual and Automatic Labelling for News Classification in Bangla Language // arXiv:2210.10903. — 2022.

Balahur A. Sentiment Analysis in Social Media Texts // Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. — Atlanta, Georgia, 14 June 2013. — P. 120–128.

Binance — a cryptocurrency exchange. — [Electronic resource]. — URL: https://www.binance.com/ (accessed: October 20, 2022).

Bitcoin — an innovative payment network and a new kind of currency. — [Electronic resource]. — URL: https://bitcoincore.org (accessed: November 13, 2022).

Decrypt. Dogecoin Pumps as Elon Musk Agrees (Again) to Buy Twitter. — [Electronic resource]. — URL: https://decrypt.co/111226/dogecoin-pumps-elon-musk-buy-twitter (accessed: 13.11.2022).

itZone. Elon Musk tweets alluding to "break up" with Bitcoin. — [Electronic resource]. — URL: https://itzone.com.vn/en/article/elon-musk-tweets-alluding-to-break-up-with-bitcoin/ (accessed: October 20, 2022).

Kaur P., Edalati M. Sentiment analysis on electricity Twitter posts // arXiv:2206.05042. — 2022.

LunarCrush — Social Intelligence for Crypto, NFTs and Stocks. — [Electronic resource]. — URL: https://lunarcrush.com/ (accessed: October 20, 2022).

Neeson S. Japanese candles. Graphical Analysis of Financial Markets. — Intellectual Literature, 2020. — 290 p. (in Russian)

Otabek S., Choi J. Twitter Attribute Classification with Q-Learning on Bitcoin Price Prediction // arXiv:2208.02610. — 2022.

Stanford CoreNLP. — [Electronic resource]. — URL: https://stanfordnlp.github.io/CoreNLP (accessed: November 13, 2022).

Stanza — A Python NLP Package for Many Human Languages. — [Electronic resource]. — URL: https://stanfordnlp.github.io/stanza/ (accessed: November 13, 2022).

Terra — a blockchain protocol and payment platform used for algorithmic stablecoins. — [Electronic resource]. — URL: https://www.terra.money (accessed: November 13, 2022).

Word2vec embeddings. — [Electronic resource]. — URL: https://radimrehurek.com/gensim/models/word2vec.html (accessed: November 13, 2022).