COMPUTER RESEARCH AND MODELING
2023 VOL. 15 NO. 1 P. 171–183
DOI: 10.20537/2076-7633-2023-15-1-171-183

UDC: 519.8

Development of and research on an algorithm for distinguishing features in Twitter publications for a classification problem with known markup

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.

Social media posts play an important role in demonstration of financial market state, and their analysis is a powerful tool for trading. The article describes the result of a study of the impact of social media activities on the movement of the financial market. The top authoritative influencers are selected. Twitter posts are used as data. Such texts usually include slang and abbreviations, so methods for preparing primary text data, including Stanza, regular expressions are presented. Two approaches to the representation of a point in time in the format of text data are considered. The difference of the influence of a single tweet or a whole package consisting of tweets collected over a certain period of time is investigated. A statistical approach in the form of frequency analysis is also considered, metrics defined by the significance of a particular word when identifying the relationship between price changes and Twitter posts are introduced. Frequency analysis involves the study of the occurrence distributions of various words and bigrams in the text for positive, negative or general trends. To build the markup, changes in the market are processed into a binary vector using various parameters, thus setting the task of binary classification. The parameters for Binance candlesticks are sorted out for better description of the movement of the cryptocurrency market, their variability is also explored in this article. Sentiment is studied using Stanford Core NLP. The result of statistical analysis is relevant to feature selection for further binary or multiclass classification tasks. The presented methods of text analysis contribute to the increase of the accuracy of models designed to solve natural language processing problems by selecting words, improving the quality of vectorization. Such algorithms are often used in automated trading strategies to predict the price of an asset, the trend of its movement.

Keywords: text analysis, natural language processing, Twitter activity, frequency analysis, feature selection, classification problem, financial markets

Citation: *Computer Research and Modeling*, 2023, vol. 15, no. 1, pp. 171–183.

**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2023 Т. 15 № 1 С. 171–183**
DOI: 10.20537/2076-7633-2023-15-1-171-183

Ки&М

УДК: 519.8

Разработка и исследование алгоритма выделения признаков в публикациях Twitter для задачи классификации с известной разметкой

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 социальных сетей играют важную роль в отражении ситуации на финансовом рынке, а их анализ является мощным инструментом ведения торговли. В статье описан результат исследования влияния деятельности социальных медиа на движение финансового рынка. Сначала отбирается топ инфлюенсеров, активность которых считается авторитетной в криптовалютном сообществе. Сообщения в Twitter используются в качестве данных. Подобные тексты обычно сильно зашумлены, так как включают сленг и сокращения, поэтому представлены методы подготовки первичных текстовых данных, включающих в себя обработку Stanza, регулярными выражениями. Рассмотрено два подхода представления момента времени в формате текстовых данных. Так исследуется влияние либо одного твита, либо целого пакета, состоящего из твитов, собранных за определенный период времени. Также рассмотрен статистический подход в виде частотного анализа, введены метрики, способные отразить значимость того или иного слова при выявлении зависимости между изменением цены и постами в Twitter. Частотный анализ подразумевает исследование распределений встречаемости различных слов и биграмм в тексте для положительного, отрицательного либо общего трендов. Для построения разметки изменения на рынке перерабатываются в бинарный вектор с помощью различных параметров, задавая таким образом задачу бинарной классификации. Параметры для свечей Binance подбираются для лучшего описания движения рынка криптовалюты, их вариативность также исследуется в данной статье. Оценка эмоционального окраса текстовых данных изучается с помощью Stanford Core NLP. Результат статистического анализа представляет непосредственно практический интерес, так как предполагает выбор признаков для дальнейшей бинарной или мультиклассовой задач классификации. Представленные методы анализа текста способствуют повышению точности моделей, решающих задачи обработки естественного языка, с помощью отбора слов, улучшения качества векторизации. Такие алгоритмы зачастую используются в автоматизированных торговых стратегиях для предсказания цены актива, тренда ее движения.

Ключевые слова: анализ текста, обработка естественного языка, активность в Twitter, частотный анализ, отбор признаков, задача классификации, финансовые рынки

Introduction

Global community creates the field to express their emotions and feelings about recent news in cryptoworld. It has been suggested that the sentiments expressed on Twitter may help in predicting price changes for cryptocurrencies. Moreover, coins became flexible depending on some people's opinion. The most vivid example of it is Elon Musk's (Tesla CEO) tweets affecting the Bitcoin price [itZone, 2022].

The main trend in conducting research on financial markets, evaluating and predicting changes in it is the use of social activity analysis. Many people share their opinions about the upcoming trend that could be tracked by lexical features and activity statistics. Related research demonstrates the potential 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 utilised there. Our study doesn't allow one to gather such data in a correct way because of a problem of data download mentioned below. A. Ibrahim research [Ibrahim, 2021] provides methods of gaining sentiment from VADER. We decided to attempt another corpus by Stanford CoreNLP. Also in the paper there is an approach of using sentiment as a feature in machine learning algorithms afterwards.

The step of preprocessing plays one of the crucial roles in our paper. In the study [Nosrati et al., 2022] the approach of removing commonly used words is suggested. There is the demonstration of method to solve the spam classification problem using Naive Bayesian classifier as well. Another research [Kumar, Harish, 2018] supports the idea of removing syntax from the text, however, we consider that can influence the sentiment of the sentences, so we keep it there for finer score of Stanford CoreNLP.

This article describes methods designed to process the specific text information linked with the community that shares news, ideas, and opinions. The paper presents the main stages of preparing data, markup, highlighting significant textual information using statistical methods, and analysis of sentiment provided by Stanford CoreNLP technology [CoreNLP, 2022]. Various approaches to the task in terms of the scale of this data are considered as well.

Text data preprocessing methods

Data selection

Detailed research includes studying sources exposed to rapidly altering moods of social media, therefore, text information is obtained from Twitter API v2. The initial data contain more than 250,000 tweets dated 24 February 2021 and present the following information: text of a tweet, posting time in the format of ISO 8601 (UTC), author id, tweet id. Due to the inner Twitter limits, it is not allowed to download data distributed in regular intervals in time. Users are selected from the list of influencers taken from LunarCrush [LunarCrush, 2022]. Lunar Crush scans Twitter to calculate influencers across various social channels and news sites for content related to coins, exchanges, and non-fungible token (NFT). The total number of users is 138.

Data preparation

Because of the specifics of Twitter source the raw text is needed to be prepared. To eliminate such unprocessable elements as emoji, hashtags, links, non-English language words, and slang abbreviations of words, regular expressions are utilized. Stanza [Stanza, 2022], a Python NLP package for many human languages, is used for lemmatization, and dependency relations search. All uppercase letters in a string also should be converted to lowercase [Qi et al., 2020].

Tweets research

Data analysis

The number of tweets is unevenly distributed in time. This behavior is due to Twitter's restrictions on the amount of data received. The frequency of usage of a word F is defined as follows:

$$F = \frac{M}{S}, \tag{1}$$

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

During the investigation the most frequent words of the sample were obtained. It is pictured in Figure 1. Bar charts show the frequency for each word.

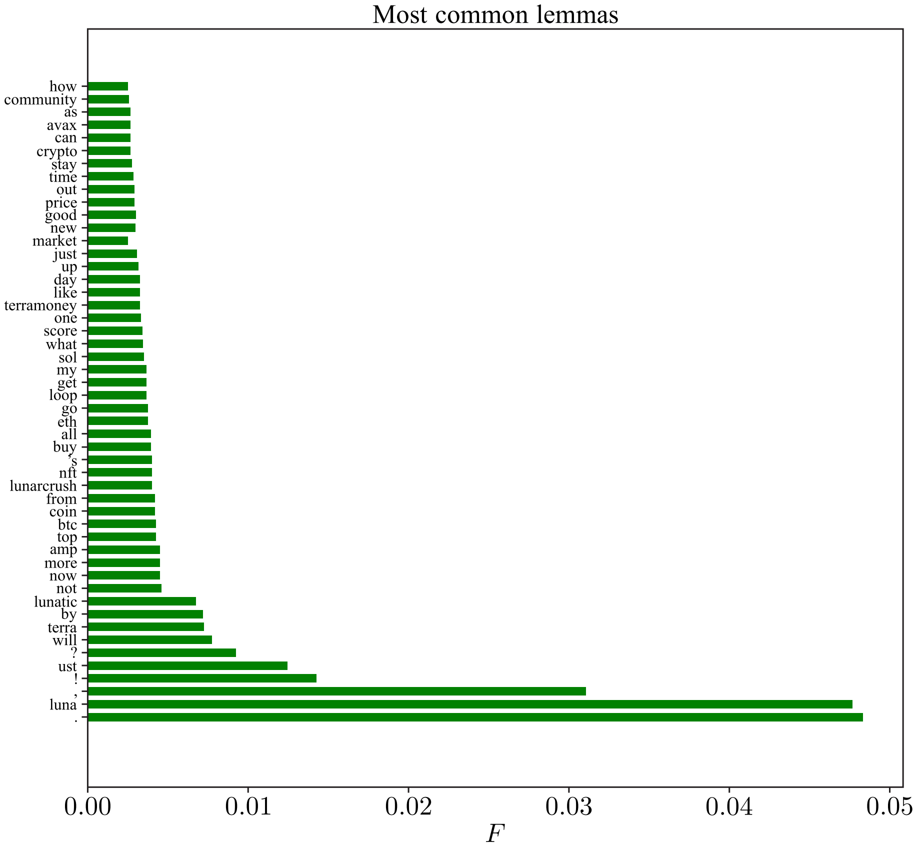

Figure 1. Top-50 of the most frequent words among downloaded tweets of selected users

In the process of exploring the frequency of usage of words in tweets, it was established that there are common senseless words and elements of punctuation that do not describe samples. Furthermore, in the top of histogram the name of a coin, that was leveraged to gather the list of influencers, and its derivatives are mentioned. This feature demonstrates how precisely frequency can describe the sample. Thus, exploring more words in the top, “not”, “can”, “top”, “buy” are used.

The frequency of word combinations was studied as well. In Figure 2 the bigrams with “not” are given. As we can see, some powerful expressions such as “not wait”, “not sell”, “not buy”, “not worry” are presented in the top.

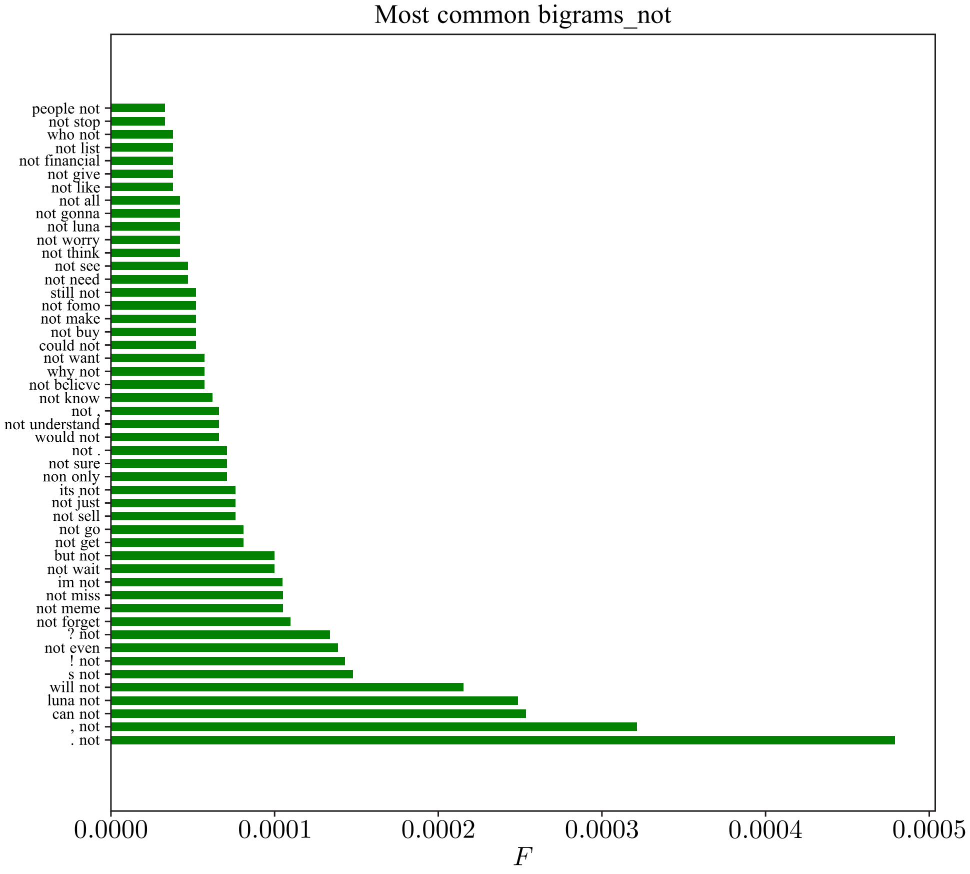

Figure 2. Top-50 of the most frequent bigrams with "not" among downloaded tweets of selected users

Data markup

The market data about price is taken for the same time period from Binance [Binance, 2022]. To start with, it is necessary to define several terms:

- *hallway* — candlestick [Neeson, 2020] area with determined width and length;
- *hallway length* (N) — the 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 ternary classification.

Two different classifications are used: binary classification, which consists of positive (+1) and negative (−1) classes, and ternary classification, which consists of positive (+1), negative (−1) and neutral (0) classes.

In binary classification to markup the single data item a price close of the candlestick, which contains the current tweet creation time, and a price close of the last candlestick in the hallway are checked. Tweet belongs to negative (−1) class if the difference between those two prices is negative and it belongs to positive (+1) class otherwise.

In ternary classification firstly it should be checked if an absolute value of the difference divided by price close of the first candlestick in the hallway is more than hallway width. It means that the tweet leads to an exit of the hallway and then binary classification markup rules might be used to classify the data item (Figure 3).

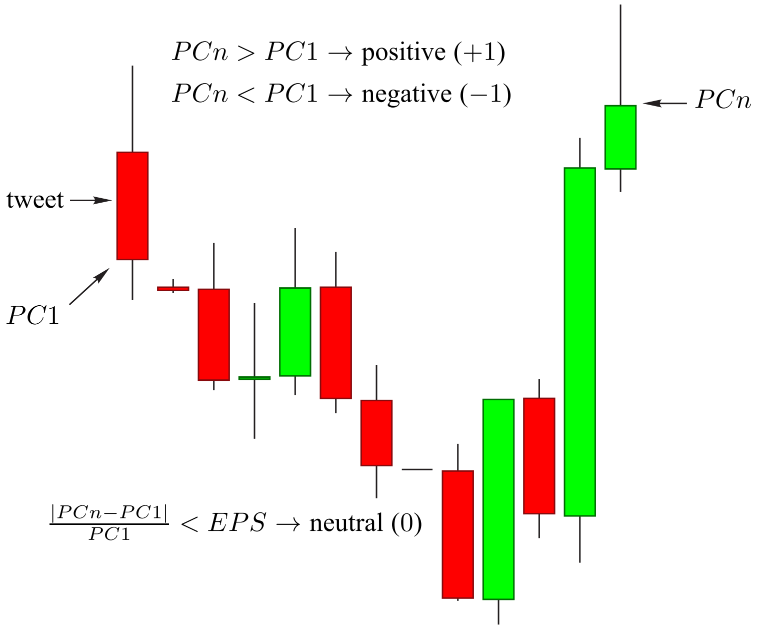
$PCn > PC1 \to$ positive (+1)
$PCn < PC1 \to$ negative (−1)
PCn
tweet
$PC1$
$\frac{|PCn-PC1|}{PC1} < EPS \to$ neutral (0)

Figure 3. Illustration of the principle of binary and ternary tweets markup

In Figure 4 the vertical axis shows the part of tweets belonging to positive (+1), neutral (0), and negative (−1) classes responsible for going beyond the boundaries of the hallway, while the horizontal one presents EPS for different values of N. It is easy to notice that for each N with changing EPS we have equal shares of tweets of classes +1 and −1, which means data stay balanced. With the growth of hallway length the intersection of the lines is deviating to the right and the part of neutral tweets for each EPS is diminishing as well.

Feature selection

As an important parameter to describe a sample of text messages it was chosen to use sentiment. Stanford CoreNLP is the tool that allows one to provide mood of the particular paragraph. The output of the core represents one of the following sentiments: “Positive”, “Negative”, “Neutral”. There is no correlation between sentiment and mark on the small timeframes less than 4 hours with EPS of 0.55 %. The primary explanation lies in the specifics of the influencers’ posts: there are lots of slang lexicon with special meaning among the auditory. For instance, tweets that include “bullish” are particularly pointing to a growing trend of price, whereas Stanford CoreNLP classified them as “Negative”, which is incorrect.

The next method to characterize a tweet is the existence in a tweet of a word from a set of the selected strong words. To select the important words let’s define the following metrics:

$$C_{bin} = \frac{n_{pos} - n_{neg}}{n_{tweets}}, \tag{2}$$

where n_{pos} is the number of tweets belonging to the binary positive class that include the word, n_{neg} is the number of tweets belonging to the binary negative class that include the word, n_{tweets} is the number tweets in the sample overall;

$$C_{bin\_hallway} = \frac{n_{inside} - n_{outside}}{n_{tweets}}, \tag{3}$$

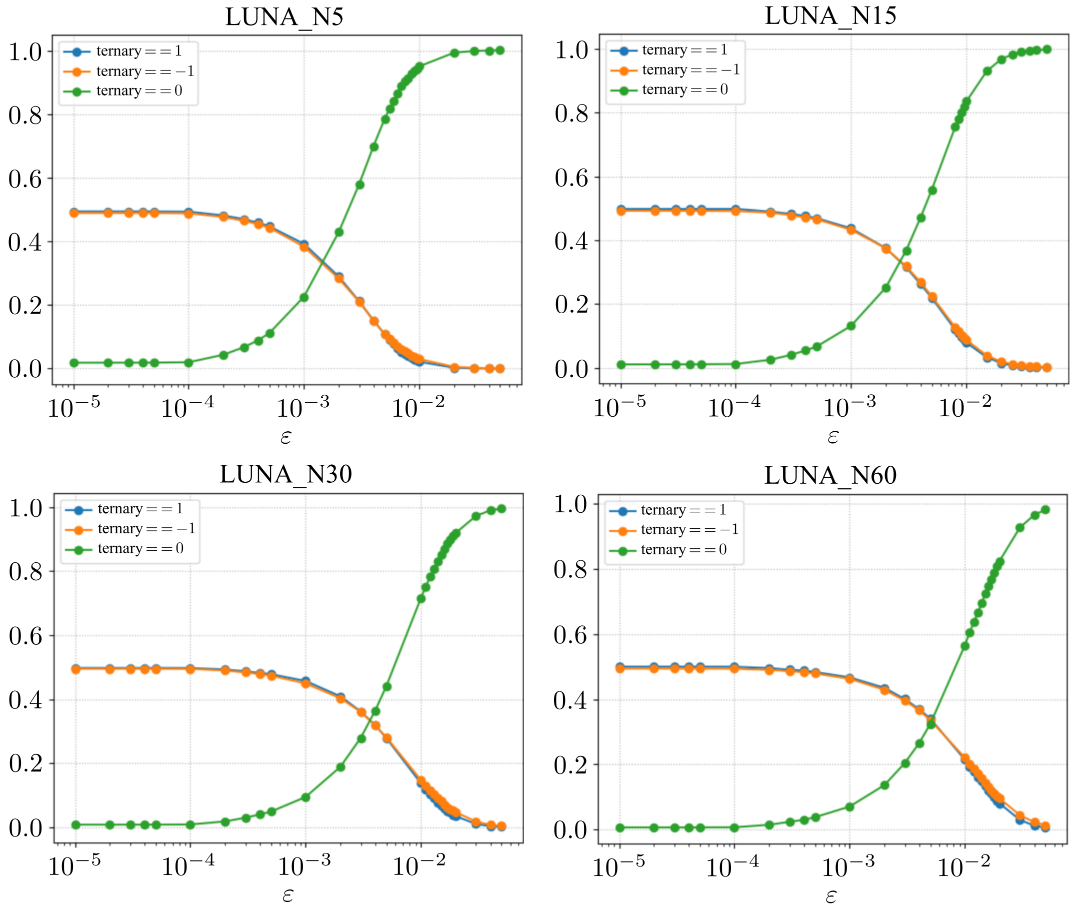

Figure 4. Comparison of connections between the part of tweets belonging to positive, negative and neutral classes and hallway width for different hallway lengths. Color versions of the figures are available in the online version of this article

where n_{inside} is the number of tweets belonging to the ternary neutral class that include the word, $n_{outside}$ is the number of tweets belonging to the ternary positive or negative classes that include the word. The visual representation of the metric is given in Figure 5.

Blue lemmas predominate in tweets which leads to the exit of hallway, while red lemmas predominate in tweets which do the opposite.

The better the set of words, the more absolute the value of C_{bin} of its elements, because it makes the set more contrast and there is less possibility to meet two words that are common for different classes in one tweet. Furthermore, if there is a word that is considered to prefer one of the classes with some neutral words in a tweet, it is more possible to mark the tweet as this particular class.

Packages research

Data formation

The data for research here is the set of tweets taken for the interval of $t_1 - t_0$ with the period of $dt = t_2 - t_1 \leqslant t_1 - t_0$. In further references to the set of tweets we shall call them packages. All timeframes related to package structure are considered to be more than 4 hours which is relatively wide. There should be mentioned a problem of overlay: dt is less than $t_1 - t_0$, so some tweets become the same for different packages. The only obvious way when not suffering from the issue is to have

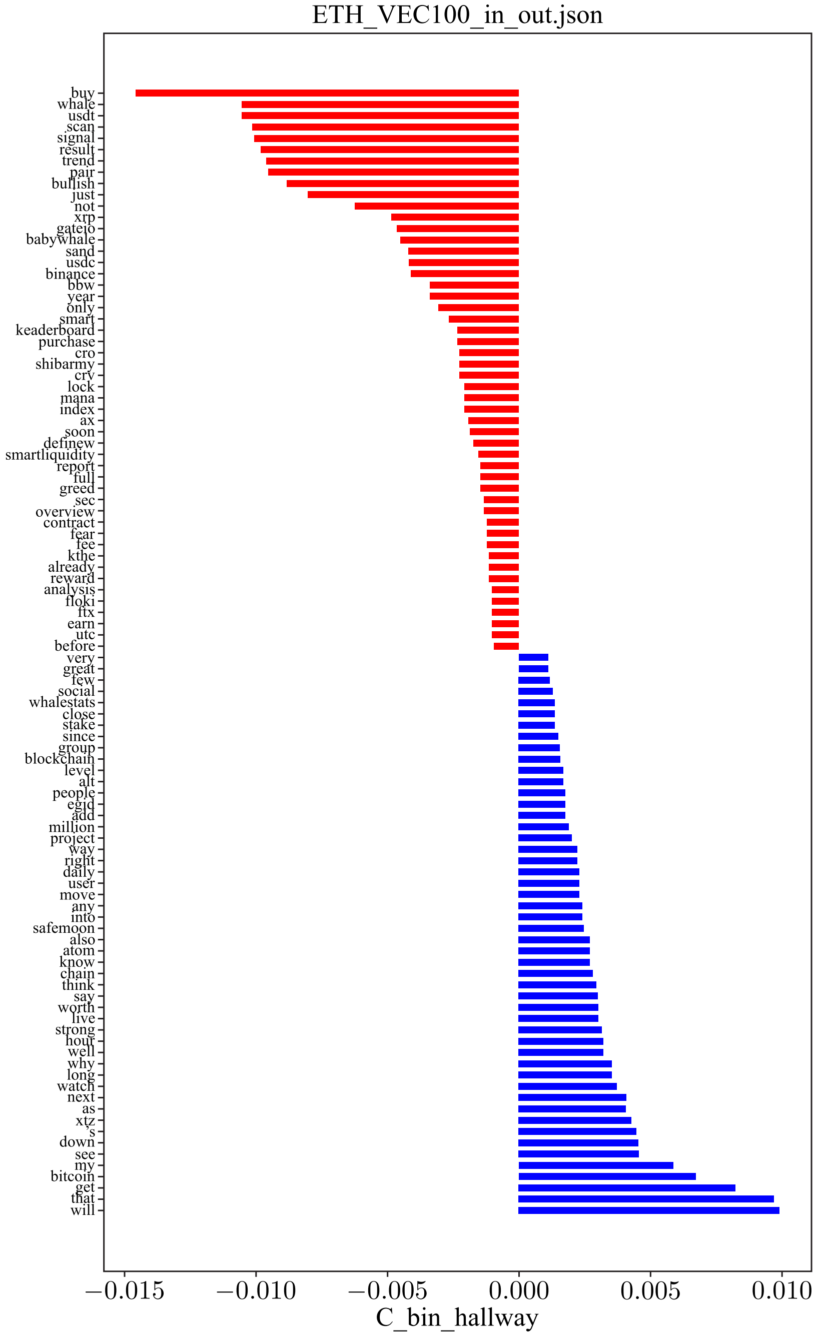

Figure 5. An example of $C_{bin\_hallway}$ metric. Color version of the figure is available in the online version of this article

equal values of dt and $t_1 - t_0$. Also due to the Twitter limits, packages are not equal in size, thus earlier packages consist of less tweets than the later ones, and the variety of set of words in a package based on specified periods changes in ascending order (Figure 6).

Data markup

We assign a label to each package using Binance according to the following rules.

- Look for a label for each coin we are interested in.

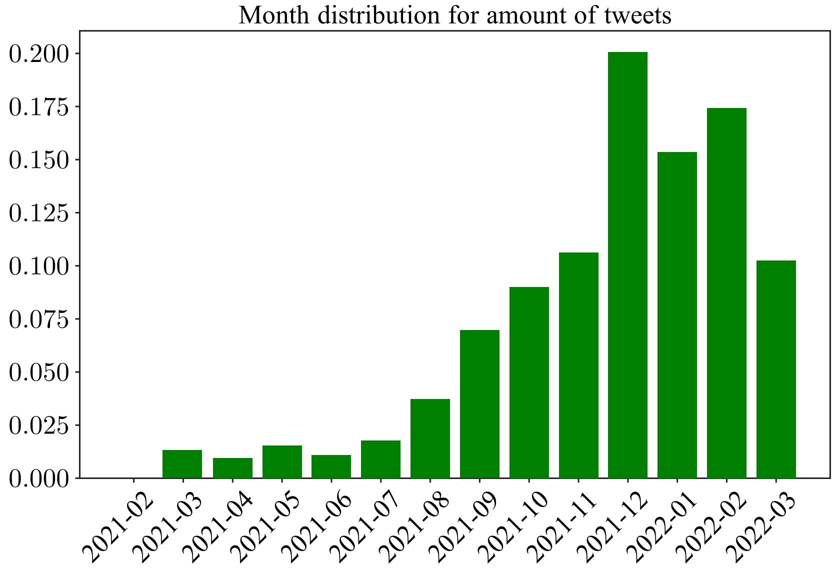

Figure 6. Uneven distribution of downloaded tweets for months since February 2021

- Consider the moment of time — the upper bound of the package.
- Find the candlestick corresponding to the upper bound.
- Find the candlestick that is t_3 away from the upper bound.
- Find the difference between close price of both candlesticks.
- If the difference found is less than zero, then the binary label is -1, otherwise $+1$.
- If the absolute value of difference found is not more that some EPS, then the ternary label is 0, otherwise it is equal to the binary one.
- Summarize all labels by each coin.
- If the sum of all binary labels is less than zero, the binary label of the package is -1, otherwise $+1$.
- If the absolute value of the sum of all ternary labels is less than V, the ternary label is 0, otherwise, if the sum is more than zero, it is 1, if the sum is less than zero, it is -1.

V has been assigned the half of the number of coins (Figure 7).

Then we take a look at how the labels are distributed depending on the EPS parameter. To do this, we build a graph pictured in Figure 8.

The number of packages that do not lead to the exit from the hallway has a certain non-zero limit for the width of the hallway tending to zero as well as packages that lead. There are also limits if EPS tends to infinity: 1.0 for ternary zero labelled packages and 0.0 for ternary non-zero labelled packages. Sigmoid shape can be recognized as well.

When changing V to a third of the number of coins distribution of labels altering in the following way (Figure 9).

To conclude, the distribution of labels now depends not only from EPS, but from V as well. The limit of the part of tweets that do not lead to the exit of the hallway is rising and the limit of

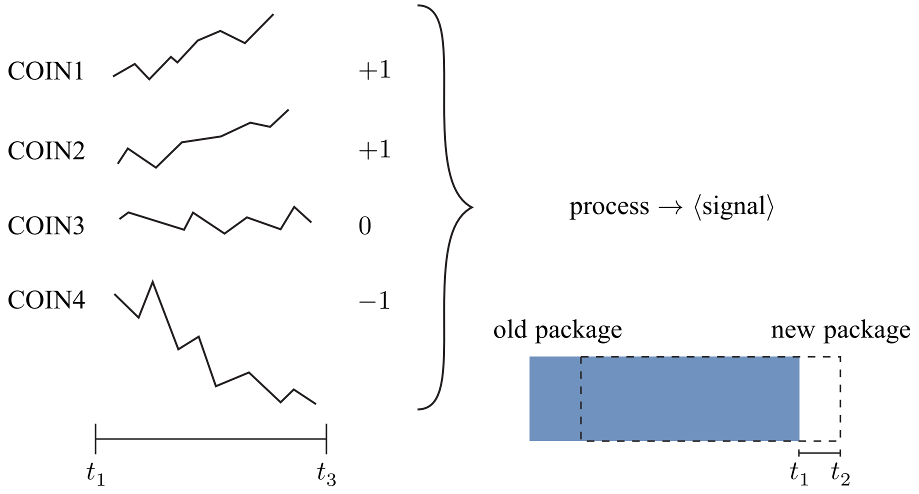

Figure 7. Illustration of the principle of packages creation and binary and ternary packages markup

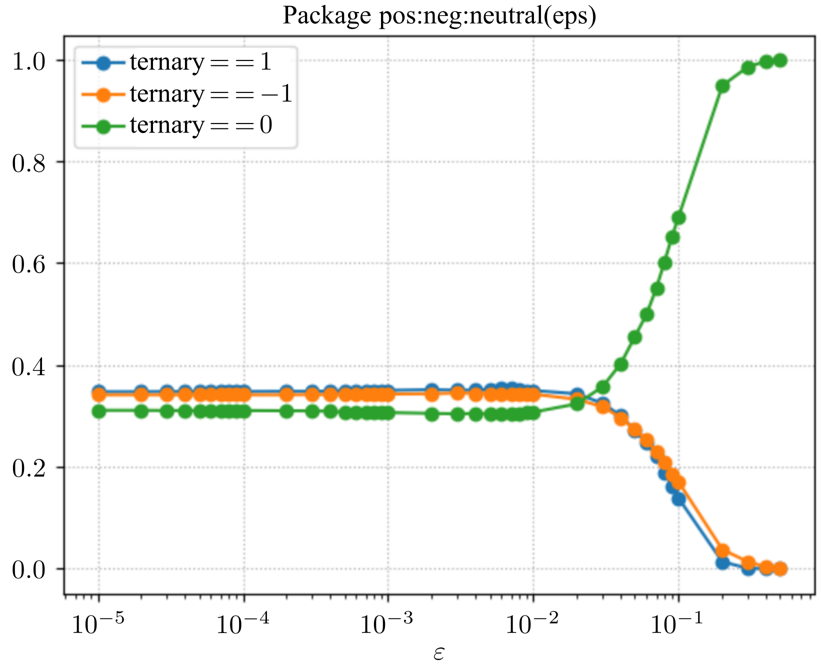

Figure 8. Part of differently labelled packages dependent on hallway width with V equalled to half of the number of coins. Color version of the figure is available in the online version of this article

the part of tweets that leads, otherwise, is falling down. But the shape of the figure does not change with V parameter.

In accordance with the Figure 8, the results in further work are given for EPS = 0.02 (2 %) to better attitude of dependence evaluation by the making our data balanced.

Feature analysis and results

To assign the final value of sentiment of a package it was decided to use the sum of sentiment gathered by each tweet from a package. If the sum is positive, the value of sentiment is 1, if the sum is negative, it is −1, 0 otherwise. The connection between sentiment of a package and its label has not been established as well. Moreover, the sentiment summed by days (Figure 10) distributed in time

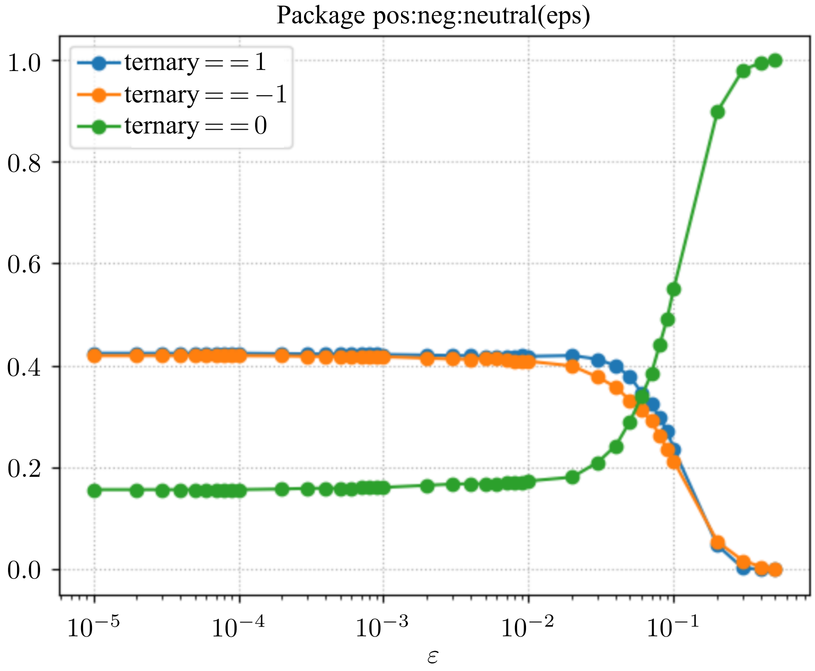

Figure 9. Part of differently labelled packages dependent on hallway width with *V* equalled to third of the number of coins. Color version of the figure is available in the online version of this article

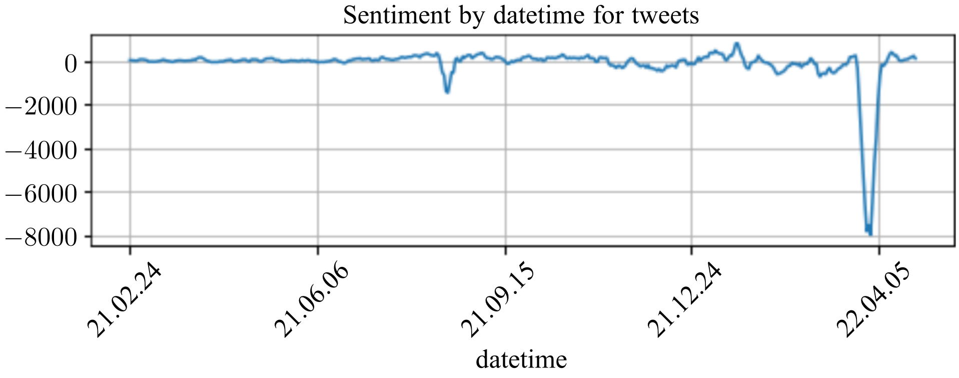

Figure 10. Accumulated packages sentiment summed by days since 24 February 2021

had several bright peaks, which literally means that some huge social events were held, but researching newsbreaks led to the thought that it was incorrect. The phenomenon might be described by the mistake (related to mislabeling a tweet with specific words) mentioned earlier. It led to the accumulation of a considerable error causing the wrong peaks.

Continuing research in describing package, frequencies of words are explored. To mitigate the impact of the data package disparity problem, we use not the number of uses of a word in a package but the ratio of this number to the total number of words in a package.

In Figure 11 it is shown that the histogram could be divided by zones based on the predominance of data with a particular label. As the result when predicting the mark of a package it could be checked whether the package falls into the zone by the frequency of specific words, so it again enhances the possibility of meeting as the particular class.

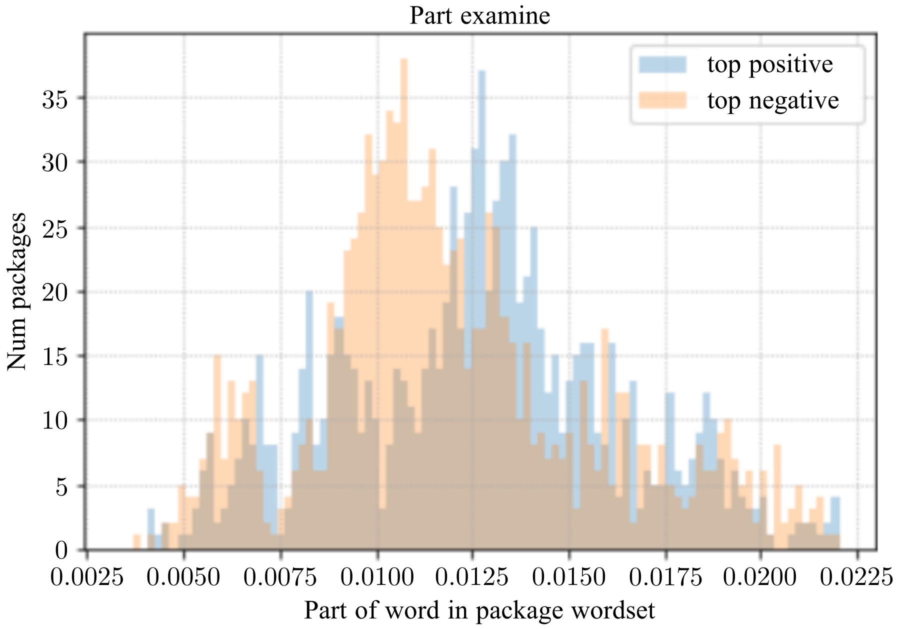

Figure 11. Example of possible histogram for "top". Color version of the figure is available in the online version of this article

Conclusion

The paper describes the results of preparing and markup data. Different methods of data processing and analysis are considered as well. The data represented by Twitter messages is called tweets which were processed using Stanza mostly and marked using cryptocurrency prices of 50 most popular coins. Data markup was built by the variation of the parameters which can be used to describe Binance candlesticks. It was also concluded that text can be classified according to on the words it consists of. Several metrics using words frequency were introduced and visualized as well. To characterize dataset sentiment provided by Stanford CoreNLP was contemplated and gave us the negative result with the idea. Two different approaches of data representation are given in the article: tweets themselves and groups of them. The obtained result might be applied to solve the classification problem by learning or benchmark algorithms when utilizing as tokenization and vectorization methods.

References

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

Ibrahim A. Forecasting the Early Market Movement in Bitcoin Using Twitter's Sentiment Analysis: An Ensemble-based Prediction Model // IEEE International IOT, Electronics and Mechatronics Conference. — 2021.

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).

Kumar H. M. Keerthi, Harish B. S. A New Feature Selection Method for Sentiment Analysis in Short Text // Journal of Intelligent Systems. — 2018. — December 4.

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).

Nosrati V., Rahmani M., Jolfaei A., Seifollahi S. A Weak-Region Enhanced Bayesian Classification for Spam Content-Based Filtering // ACM Transactions on Asian and Low-Resource Language Information Processing. — 2022. — DOI: https://dl.acm.org/doi/10.1145/3510420

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

Qi P., Yuhao Zhang Y., Yuhui Zhang Y., Jason Bolton J., Manning C. D. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages // Association for Computational Linguistics (ACL) System Demonstrations. — 2020.

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).