3. Хохолов Ю. А., Мамонов А. Ф., Зубков В. П. Оптимизация формирования льдопородного массива в горных выработках // Горный информационно-аналитический бюллетень. – 2004. - № 10 – С. 103-107.
4. Бураков А.М., Ермаков С.А., Каймонов М.В., Курилко А.С., Хохолов Ю.А. Технологические особенности строительства дражных полигонов в условиях россыпного месторождения криолитозоны // Горный информационно-аналитический бюллетень. – 2006. - № 8 – С. 271-276.
5. Курилко А.С., Ермаков С.А., Хохолов Ю.А., Каймонов М.В., Бураков А.М. Моделирование тепловых процессов в горном массиве при открытой разработке россыпей криолитозоны. - Новосибирск: Академическое изд-во «Гео», 2011. – 140 с.
6. Патент RUS 2134459, МПК6 G 21 F 9/24. Способ захоронения твердых радиоактивных отходов в зоне многолетней мерзлоты / В.В. Киселев, Ю.А. Хохолов; заявитель и патентообладатель ИГДС СО РАН. - №97113864/25; заявл. 12.08.97; опубл. 10.08.99, Бюл. № 18.
7. Патент RUS 2263985, МПК6 G 21 F 9/34. Подземный бесконтейнерный способ захоронения твердых источников радиоактивного излучения в отработанных подземных горных выработках криолитозоны / В.В. Киселев, Ю.А. Хохолов; М.В. Каймонов; заявитель и патентообладатель ИГДС СО РАН. - №2003125163/06; заявл. 11.08.03; опубл. 10.11.05, Бюл. № 31.
8. Киселев В.В., Хохолов Ю.А., Каймонов М.В. Использование горных технологий и выработанного пространства рудников криолитозоны для захоронения твердых радиоактивных отходов // Горный информационно-аналитический бюллетень. - 2006. - № 3. - С. 129–134.

Dmitrienko N.A., Avdeeva T.G., Alekseenko J.V.
The research of methods and development of the system
for Twitter messages sentiment analysis

Don State Technical University
(Россия, Shakhty)

doi: 10.18411/lj-31-01-2018-58

idsp: 000001:lj-31-01-2018-58

Abstract

The article describes the development of a program for twitter sentiment analysis based on machine learning methods. The approaches used to determine the key of the text are considered. The analysis of current methods for determining sentiment analysis and vector space models of text presentation were provided. Three algorithms for the classification of texts by tonality were implemented: naive Bayesian algorithm, random forests, and logistic regression. During the testing of algorithms using the ROC curve, the most effective one was determined.

Keywords: machine learning, vector space model, naive Bayesian classifier, random forests, logistic regression.

Recently, social networks have grown rapidly on the Internet, such websites as Facebook and Twitter have become extremely popular since their first appearance. Nowadays millions of people share their thoughts about events happening in the world with their friends and followers on social networks.

One of the features of such social networks is the search for the necessary contacts and the establishment of connections between people. Sentiment Analysis can be used to analyze the history of user messages.

Recognizing texts marked with emotional labels is one of the classification tasks. This task can be applied not only for the classification of text, but also for tasks of artificial intelligence. Sentiment analysis is a field of computer linguistics, which deals with the allocation of emotionally labeled vocabulary or author's emotional evaluation from texts. Applying sentiment analysis commercial companies can find out the opinion of buyers about a product. This data can be used to improve the quality of the product, to determine the key audience, as well as to find the main advantages and disadvantages of competitors.

At the moment, there are two commonly used approaches to the sentiment analysis problem: the first approach is based on machine learning methods; the second one is called sentiment vocabulary-based approach

Figure 1. Approaches used for Sentiment Analysis

The approach based on the application of sentiment dictionaries is analyzing the tonality of individual words (terms) in text and subsequent recognizing the tonality of the entire text according to the estimates of the individual words used in this text.

In the approach based on machine learning, analyzing the tonality is a task of classifying texts that can be solved by a classifier training on a pre-labeled collection of texts.

Each of these two approaches has its advantages and disadvantages. Now there are methods based on the use of sentiment dictionaries that do not require a training collection, that is, manual text marking is no longer needed. Moreover, these methods do not need to compile a learning function, also, "solutions" accepted by a classifier can be easily explained. In this case, pre-labeled sentiment dictionaries are required for these methods. In their turn, the dictionaries also have to take into account the subject area of the text being studied. These sentiment dictionaries are not used in the methods of machine learning. In practice, classifiers demonstrate high quality of classification.

One of the most significant problems in the field of computer algorithms is the problem of extracting the "sense" from the text written in a natural language, and then presenting it in the form that would be convenient for computer processing. The most famous and simple way of representing knowledge is vector model.

Vector model is representing the collection of documents in an information search by vectors from one vector space that is common to the whole collection.

There are two basic models of text representation:

- Bag of Words is a model often used in text processing, which is an unordered set of words that are part of the text being processed.
- TF-IDF is a statistical measure used to evaluate the importance of a word in the document context that is a part of the collection of documents or a corpus.

The analysis of classification problems was performed, and the most popular classifiers were identified:

- Logit Model is a statistical model used to predict the probability of occurrence of an event by matching the data to the logistic curve. [1, p. 72]
- RF (Random Forest) is the set of decision trees. In the regression problem their responses are averaged, in the classification problem the decision is taken by majority vote. [2, p. 109]
- A Bayesian Approach to classification is based on the theorem, which states that if distribution densities of each of classes are known, the wanted algorithm can be written in an explicit analytical form [3, p. 56]

On the basis of the data obtained during the study, some experiments were conducted, which helped to compare two types of classification algorithms: binary and hierarchical classification of Twitter text messages.

Twitter is a social network for public message sharing with the use of web interface, SMS, instant messaging tools or third-party software. [4]

To use text messages (tweets) in other external software products we need to use Twitter API, which is the set of pre-built classes and methods provided by Tweeker.

Python was used as the language for development. Python is a high-level general purpose programming language, which is focused on improving developer productivity and code readability. [5, p. 302]

Jupyter Notebook was chosen as a development environment. Jupyter Notebook is a web-based environment that allows users to perform interactive calculations in special documents for laptops.

The two algorithms were developed to determine the most efficient machine learning algorithms for automatic analysis of the tonality of Twitter text messages. The first one is an algorithm for binary classification of text messages, which was taken as a standard for comparison. This algorithm assumes that the sample of messages is divided into only two classes: "positive" and "negative".

The general view of the binary classification algorithm is presented below:

- Load the text of a message and break it into words, which will be turned later into lowercase
- Discard punctuation
- Discard numbers
- Use TF-IDF for the resulting word vector
- Use the logistic regression, naive Bayes classifier and random forest for the resulting feature vector is used

We used scikit-learn, which is a machine learning library, for preprocessing and producing features used.

To test the accuracy of the classification algorithm cross-validation method (cross-validation, cross-validation) was used. As one step of this procedure a set of partitions of the original sample is fixed for a set of partitions of the original sample into two groups: learning and controlling. [3, p. 5] For each partition setting up an algorithm for the training subgroup is executed, and then we evaluate its efficacy on vectors of the control subgroup. A cross-validation estimate is the average for all partitions of the accuracy and completeness values on control subgroups. If the sampling objects are independent, the average values of the evaluations of the cross-validation effectiveness will give unbiased performance estimates.

Figure 2 shows the classification accuracy of the three algorithms on the test data.

Figure 2. The Classification Accuracy

As we can see in Figure 2, the logistic regression has the greatest accuracy.

Emotional tone of tweets may not only be positive or negative. There are always tweets that do not contain any particular emotions, thus they are neutral and irrelevant or simply provide some information. For performing classification of text messages into three classes we applied a hierarchical approach. This is about separating "emotional" messages from neutral ones on the

first step, and then separating positive and negative ones among only emotional messages on the second step. This approach is presented in Figure 3.

Figure 3. Hierarchical Classification Algorithm

In the present hierarchical classification algorithm, the search of neutral tweets is based on the body of neutral and emotion-toned words Liu and Hu opinion lexicon.

Figure 4 shows the classification accuracy of the three types of classifiers on the test data.

Figure 4. The Classification Accuracy

As we can see in Figure 4, the logistic regression has the highest classification accuracy.

Comparing the two types of classification (Figure 5), we can conclude that the hierarchical classification is on average 5% better than the binary classification.

Figure 5. The Comparison of Classification Types

To sum up, the paper presents the most popular classification algorithms for the analysis of a text message tone. The authors considered the implementation of the most popular classification algorithms, as well as determined the most effective machine learning algorithm to

analyze the tone of Twitter text messages. Also, the environment for development and programming language are described.

1. Boyarsky K. K. Introduction to Computer Linguistics: Textbook. allowance - SPb: NIU ITMO, 2013. - 72 p.
2. Hastie T., Tibshirani R., Friedman J. Chapter 15. Random Forests // The Elements of Statistical Learning: Data Mining, Inference, and Prediction. - 2nd ed. - Springer-Verlag, 2009 -109 p.
3. Subbotin SV, Bolshakov D. Yu. Application of the Bayesian classifier for the recognition of classes of goals. // "Journal of Radioelectronics", Izd. 4e - 2006. - 56 with.
4. Khramov D. Using the Twitter API for data collection. URL: http://dkhramov.dp.ua/images/edu/Stu.WebMining/ch17_twitter.pdf (circulation date is 30/03/2017).
5. Koelo LP, Richard V. Construction of machine learning systems in Python. // "Machine learning", Izd. 2e - 2016 -302 s.