UDC 519.68; 681.513.7; 612.8.001.57; 007.51/.52

Olena Gavrilenko, Yuri Oliynik, Hanna Khanko
(NTUU «Igor Sikorsky Kyiv Polytechnic Institute»)

Comparative analysis of TEXT MINING algorithms for identifying agitation data

The work is devoted to the development of a general approach to methods of data mining for the classification of textual information. Algorithms TF-IDF, criterion χ^2, gradient boosting algorithm, logistic regression, support vector method and naive Bayesian classifier are considered as the proposed approaches to solve the problem.

The objects of the research are the content of electronic media news, users, and the interrelations between them.

Nowadays the problem of identifying elements of propaganda in text and metadata is of great importance. Obviously, one of the difficulties that we face in dealing with this global problem is the classification of data.

The purpose of this work is to improve the accuracy of the classification of textual information through the appropriate use of existing methods of data mining using the most effective methods of text preprocessing and powerful machine learning algorithms for problems classification.

The methods for solving the problem are considered to solve the problem of classifying text information for spam filtration tasks, contextual advertising, news categorization, creation of subject catalogues.

Today the importance of the intellectual analysis of text documents is growing rapidly. This is due to the large amount of textual information available on the Internet. Every day a huge amount of text content is generated, which requires analysis for various purposes - from monitoring users' opinions about the political situation and tracking the demand for manufactured goods. Thus, the Ukrainian monitoring project OKO monitors the references to Ukraine in foreign media and uses an analytical tool for the media, government, activists. The project has accumulated over 1 million articles in 13 languages. The project uses an algorithm that analyzes the collected articles on thematic content, the emotional coloring of topics and articles, the frequency of the mention of those articles classified in the manual mode. But still the problem of languages maintenance and their automatic classification remains.

There is a large number of content aggregators. But such aggregators contain simple algorithms for processing, do not automatically determine the emotional color and classify the content.

That is why it is necessary to automate the process of searching, filtering and structuring text data. To solve this problem, the automated classification of texts is used - the task of machine learning from the field of natural language processing. The task of text classification has practical application in many areas, for example, spam filtering, contextual advertising, news categorization, creation of thematic catalogs. Most methods of automatic classification of texts are based on the assumption that the texts of each thematic heading contain certain features, the presence or absence of which indicates the

belonging of the text of a rubric. The task of classification methods is to select the best following characteristics and formulate rules, will decide whether to refer the text to a certain category and conduct interactive drilling.

Existing means of text data analysis do not allow to obtain the desired results solving problems of classification of textual information, so there is a need to develop new algorithms, based on the accumulated information can effectively classify textual information.

Statement of problems. The task of classifying text documents can be formulated as an approximation problem for the unknown function Φ: D × C → {0,1} (how documents should be classified) via the function K: D × C → {0,1}}, which is a classifier, where $C = \{c_1, c_2, \ldots c_{|C|}\}$ - the set of possible categories, and $D = \{d_1, d_2, \ldots d_{|D|}\}$ - the set of documents (1).

$$\Phi(d_j, c_i) = \begin{cases} 1, d_j \in c_i, \\ 0, d_j \notin c_i. \end{cases} \quad (1)$$

A document d_j is called a positive example of the category c_i, if $\Phi(d_j, c_i) = 1$, and negative otherwise. If only one category can correspond to each document, then this is a task of unambiguous classification, and if several - of multivalued classification.

Finding of a classifier for a set of categories is considered as a search for binary classifiers.

Methods for solving the problem. Like any machine learning algorithm, you need to have the original data. Warehouse functions related to what I would like, so I did not know what it is, but I do not know how to do it. The choice of characteristics is also a difficult stage in the solution of the problem. The whole process of classifying texts consists of the following stages:

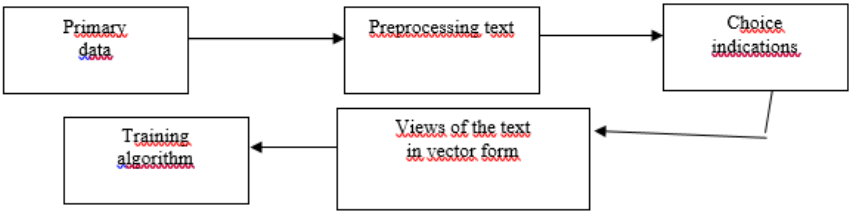

Figure 1 - Basic steps for data mining

Analysis of text mining algorithms: The paper describes the most effective algorithms for classification of textual information, such as the gradient boosting algorithm, logistic regression, support vector method and naive Bayesian classifier.

ROC-curves are constructed for selected algorithms (Fig. 2-5):

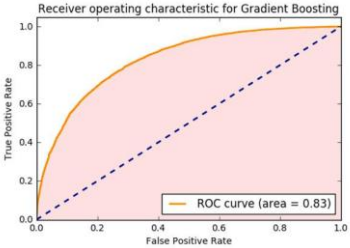

Figure 2 - ROC curve for gradient boost

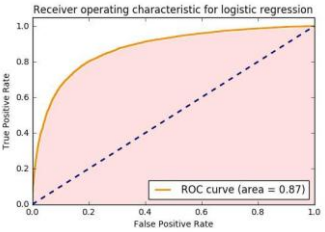

Figure 3 - ROC curve for logistic regression

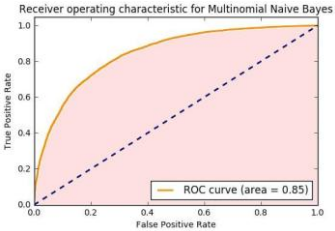

Figure 4 - ROC curve for a naive Bayesian classifier

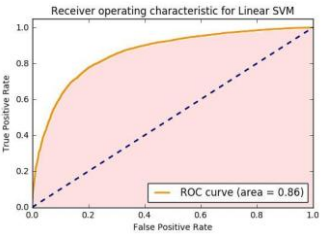

Figure 5 - ROC curve for the reference vector method

The following results are obtained on the selected algorithms (Table 1):

Table 1 - Test results

Algorithms	Accuracy	Average Precision	Average Recall	AUC
Logistic regression	0.848	0.878	0.811	0.87
Naive Bayesian classifier	0.827	0.842	0.773	0.85
The reference vector method	0.837	0.861	0.792	0.86
Gradient boost	0.783	0.801	0.739	0.83

It can be concluded that all selected algorithms are sufficiently effective and can be used to solve this problem. However, the most effective for the chosen task was the logistic regression algorithm. Its indicators of accuracy, completeness and area under the ROC curve were the highest. The worst results are in gradient booting. The main disadvantages of the algorithm are its propensity to retrain, sensitivity to emissions and low efficiency in working with sparse data.

References

1. T. S. Guzella & W. M. Caminhas. (2009). A review of machine learning approches to spam filtering. [Elsever, Expert System with Applications] [in English].

2. C.-H. Lee and H.-C. Yang. (2009). Construction of supervised and unsupervised learning systems for multilingualtext categorization. [ExpertSystems with Applications] [in English].

3. Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers. [A comparison of logistic regression and naive bayes. In NIPS 14] [in English].

4. A. Markov and M. Last, (2005). A simple, structure- sensitive approach forweb document classification. [Atlantic Web Intelligence Conference –AWIC] [in English].

5. Vinciarelli A. (2004) Noisy Text Categorization,Pattern Recognition. [17th International Conference on (ICPR'04)] [in English].

6. S. Chakrabarti, S. Roy & M. V. Soundalgekar (2003). Fast and accurate text classification via multiple linear discriminant projection. [TheInternational Journal on Very Large Data Bases (VLDB)] [in English].

7. Monitoring project of OKO. [Electronic resource]. Retrieved from http://www.ukroko.org/ [in English]

8. Gavrilenko Olena, Oliynik Yuri, Khanko Hanna (2018) Review and analysis of algorithms TEXT MINING [Project management, systems analysis and logistics] [in Ukrainian].