UDC 004.912
DOI: 10.15587/2706-5448.2021.237767
Article type «Reports on Research Projects»

Olena Kryshchenko,
Vadym Yaremenko

A COMPARATIVE ANALYSIS OF TEXT DATA CLASSIFICATION ACCURACY AND SPEED USING NEURAL NETWORKS, BLOOM FILTER AND NAIVE BAYES

The object of research is the methods of fast classification for solving text data classification problems. The need for this study is due to the rapid growth of textual data, both in digital and printed forms. Thus, there is a need to process such data using software, since human resources are not able to process such an amount of data in full.

A large number of data classification approaches have been developed. The conducted research is based on the application of the following methods of classification of text data: Bloom filter, naive Bayesian classifier and neural networks to a set of text data in order to classify them into categories. Each method has both disadvantages and advantages. This paper will reflect the strengths and weaknesses of each method on a specific example. These algorithms were comparatively among themselves in terms of speed and efficiency, that is, the accuracy of determining the belonging of a text to a certain class of classification. The work of each method was considered on the same data sets with a change in the amount of training and test data, as well as with a change in the number of classification groups. The dataset used contains the following classes: world, business, sports, and science and technology. In real conditions of the classification of such data, the number of categories is much larger than that considered in the work, and may have subcategories in its composition.

In the course of this study, each method was analyzed using different parameter values to obtain the best result. Analyzing the results obtained, the best results for the classification of text data were obtained using a neural network.

Keywords: *text data classification, Bloom filter, naive Bayes, neural network, classification time and accuracy.*

Received date: 29.04.2021
Accepted date: 02.06.2021
Published date: 29.07.2021

How to cite
Hryshchenko, O., Yaremenko, V. (2021). A comparative analysis of text data classification accuracy and speed using neural networks, Bloom filter and naive Bayes. Technology Audit and Production Reserves, 5 (2 (61)), 6–8. doi: http://doi.org/10.15587/2706-5448.2021.237767

1. Introduction

The amount of textual information is growing every day. In most cases, these works are digital, but there are also printed versions. The e-mails that appear can be attributed to many areas: social networks, web pages, e-mails, articles, messages, books, customer support, telephone conversations, and more. With the advent of large amounts of information, manual processing of large amounts of data becomes unrealistic, there is a need for its processing.

For such tasks methods of classification and clustering of texts are used. To date, a large number of methods and their various variations for the classification of texts have been developed. Each group of methods has its advantages and disadvantages, areas of application, features and limitations [1]. *The object of research* is the methods of classification of text data. *The aim of research* is a comparative analysis of classification methods and determining the best to solve the problem of classification of text data.

Therefore, it is important to conduct a study to assess the accuracy and time of texts classification by different methods depending on the amount of input data.

2. Methods of research

The study is based on the use of the following classification methods: Bloom filter [2, 3], naive Bayesian classifier [4] and neural network [5–7].

The following steps were defined for the classification of text data:

– text preprocessing;
– model training;
– data set classification;
– assessment of the accuracy of the results.

During this research the total amount of text data is as follows: 30,000 texts per each of the 4 categories for training and 1,900 texts for testing. The data set result overview is presented in Table 1.

Data preprocessing includes removal of punctuation marks, cleaning of stop words, lemmatization, stamming and cutting the words [8]. Then the data can be used to train or test the classification model, preprocessing results are shown in Table 2 [9].

If such data processing is sufficient to use the Bloom filter and the naive Bayesian classifier, it is not enough for the neural network. In the case of using neural networks,

the input text must be converted into numerical form to be able to use mathematical and statistical calculations [10].

Table 1

Input data

Class Index	Title	Description
1	Ailing Arafat Rushed to Paris Hospital	«Palestinian leader Yasser Arafat, suffering from a serious but mystery illness, was flown to France and rushed to a military hospital for treatment Friday – ending»
2	Montgomery, Gaines doping cases post-poned	«A hearing in the doping cases of US sprinters Tim Montgomery and Chryste Gaines that had been set for next week was postponed at the request of all parties, the Court of Arbitration for Sport said Friday»
3	Wall St. Bears Claw Back Into the Black (Reuters)	«Reuters – Short-sellers, Wall Street's dwindling\band of ultra-cynics, are seeing green again»
4	Group to Propose New High-Speed Wireless For-mat (Reuters)	«Reuters – A group of technology companies\including Texas Instruments Inc. (TXN.N), STMicroelectronics\(STM.PA) and Broadcom Corp. (BRCM.O), on Thursday said they\will pro-pose a new wireless networking standard up to 10 times\the speed of the current generation»

Table 2

Preprocessing results

Class Index	Title	Description
1	Arafat rush pari hospit	Palestinian leader yasser arafat suffer serious mysteri flown franc rush militari hospit treat-ment friday
2	Montgomeri gain dope case postpon	Hear dope case sprinter montgomeri chryst gain next week postpon request parti court arbitr sport said friday
3	Wall bear claw back black reuter	Reuter short-sel wall street dwindl ultra-cyn green
4	Group propos high-spe wire-less format reuter	Reuter group technolog compani includ texa in-strument stmicroelectron broadcom corp brcm thursday said propos wireless network standard time speed current generat

The operation of each method was investigated using different parameter values. For example, for a neural network, a different number of epochs, neurons, and different activation functions were considered. For each method, the optimal values of the parameters were determined, i. e. the values that gave the best results with different amounts of data, their completeness and the number of categories. Comparative analysis was performed by analyzing the obtained operating time and accuracy, the faster and more accurate the algorithm, the better it is. In this study, the following activation functions for the neural network were considered: rectified linear unit, sigmoid, tan h, exponential linear unit. The best result with optimal parameter values was obtained using the last activation function. Also, it is needed to say that this experiment is a continuation of research started in [7] where neural networks and naive Bayesian classifiers were compared.

3. Research results and discussion

Using the data from Table 1 and the obtained optimal values of variables for each of the methods showed the dependence of the accuracy of data classification on the number of data and the number of classification groups (Fig. 1–3). For the neural network the parameters are as follows: the activation function of the exponential linear unit was chosen, 10 neurons, 10 epochs and 3,000 unique words were used. For the Bloom filter, the probability of false positive case was chosen to be 0.6. A personal computer with an Intel® Core™ i7-8550U CPU @ 1.8 GHz 1.99 GHz processor, 8 GB RAM and 256 GB SSD was used to perform the testing.

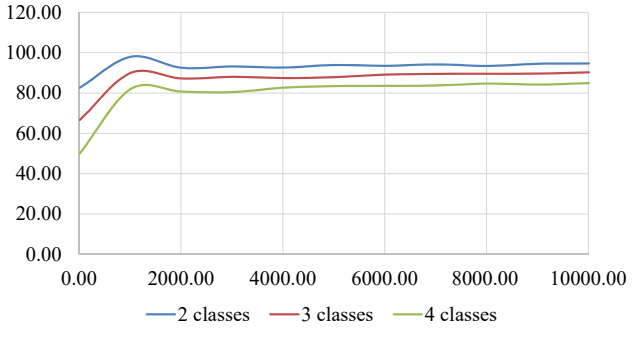

Fig. 1. Dependency of the accuracy of data classification depending on the amount of data and the number of categories using a naive Bayesian classifier

Fig. 2. Dependency of the accuracy of data classification depending on the amount of data and the number of categories using the Bloom filter

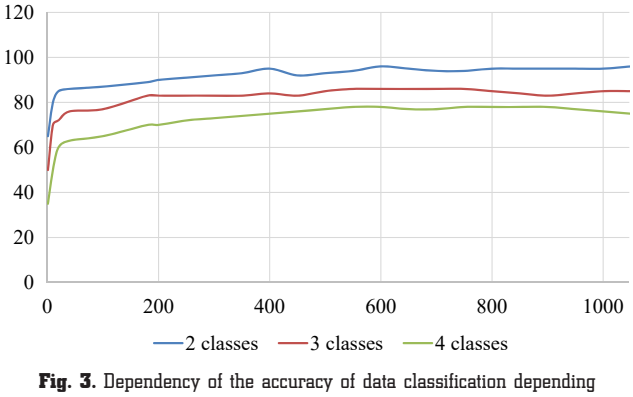

Fig. 3. Dependency of the accuracy of data classification depending on the amount of data and the number of categories using the neural network

Before presenting the results, the following research limitations need to be mentioned: the classification was done only for 2–4 classes; the testing was performed only on 1 PC thus

ISSN 2664-9969

the timing might be different in another environment; words in the texts were converted into the numeric format but considering synonyms. The last limitation means that it is possible to improve the classification accuracy by using methods for finding the similarity of the words.

Also, the speed of data classification depending on the amount of text data and the number of categories was investigated. The results are presented in Table 3.

Table 3

Dependence of classification speed on the amount of text data and the number of categories

Amount of data	Classification time, s					
	Training sample – 500 per category, testing sample – 50			Training sample – 1000 per category, testing sample – 100		
Number of categories	2	3	4	2	3	4
Naive Bayesian classifier	13.8	18.4	41.1	20	53.4	109.8
Bloom filter	0.23	0.31	0.44	0.45	0.7	0.79
Neural network	4	4.5	5.7	5.6	6.6	8.2

Based on result the further research should be done on the next directions: improving the accuracy of neural networks by changing its architectures and parameters, improving the Bloom filter to increase the quality of classification. Also, one of the possible ways to improve the result is to combine the two classification methods.

4. Conclusions

In this paper the methods of fast classification of text data are considered and their comparative analysis is prepared. Analysis of the methods was developed for two important parameters: speed and efficiency. Depending on the number of classification groups, the accuracy and time of classification change. As the number of categories increases, so does the amount of data to be taught to achieve the accuracy that is achieved for a small number of classes.

The neural network has the best accuracy – 97.3 % for 2 classes, 85.21 % for 3 classes and 75.54 % for 2 classes. At that time, the results for naive Bayesian classifier is 94.66 %, 90.28 % and 84.94 % accordingly, and for the Bloom filter – 42 %, 23.1 % and 19.12 % accordingly.

When comparing the time, the best results are shown with the Bloom filter as shown in Table 3. In problems where the problem of accuracy and speed is solved, the neural network is best suited. The naive Bayesian classifier also gives a fairly high accuracy, but it is very slow compared to other methods. However, the Bloom filter has a great advantage, this method is quite fast. If to solve the problem of rapid classification of data, this method is better than others considered in this paper. Before choosing a method, it is necessary to determine the problems to be solved by a particular method and based on the selected tasks to choose the algorithms accordingly.

References

1. Khatun, A., Mafiul Hasan, M., Miah, A.-A., Miah, R. (2020). *Comparative Study on Text Classification*. Available at: https://www.researchgate.net/publication/344199138_Comparative_Study_on_Text_Classification
2. Yaremenko, V., Budonnyi, D. (2019). Approach of the bloom filter application for real time text data multi-class classification. *Computer-integrated technologies: education, science, production, 36*, 153–159. doi: http://doi.org/10.36910/6775-2524-0560-2019-36-24
3. Leskovec, J., Rajaraman, A., Ullman, J. D. (2014). *Mining Data Streams. Mining of Massive Datasets*. Cambridge: Cambridge University Press, 123–153. doi: http://doi.org/10.1017/cbo9781139924801.005
4. Parsian, M. (2015). *Data Algorithms: Recipes for Scaling Up with Hadoop and Spark*. O'Reilly Media, Inc.
5. Lakshmi Prasanna, P., D. Rajeswara Rao, D. (2017). Text classification using artificial neural networks. *International Journal of Engineering & Technology, 7 (1.1)*, 603–606. doi: http://doi.org/10.14419/ijet.v7i1.1.10785
6. Aggarwal, C. (2014). *Data Classification Algorithms and Applications*. New York: CRC Press, 707.
7. Yaremenko, V., Rogoza, W., Spitkovskyi, V. (2021). Application of neural network algorithms and naïve bayes for text classification. *Journal of Theoretical and Applied Information Technology, 99 (1)*, 125–134.
8. Vander Plas, J. (2016). *Python data science handbook: essential tools for working with data*. Sebastopol: O'Reilly Media, Inc.
9. Mowafy, M., Rezk, A., El-bakry, H. M. (2018). An Efficient Classification Model for Unstructured Text Document. *American Journal of Computer Science and Information Technology, 6 (1)*. doi: http://doi.org/10.21767/2349-3917.100016
10. Antons, D., Grünwald, E., Cichy, P., Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management, 50 (3)*, 329–351. doi: http://doi.org/10.1111/radm.12408

Olena Hryshchenko, *Department of System Design, Institute for Applied Systems Analysis, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine, ORCID: https://orcid.org/0000-0001-6888-8665*

✉ ***Vadym Yaremenko,*** *Postgraduate Student, Assistant, Department of System Design,Institute for Applied Systems Analysis, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine, e-mail: yaremenko.v.s@gmail.com, ORCID: https://orcid.org/0000-0001-8557-6938*

✉ *Corresponding author*