Comparative analysis of machine learning methods for news categorization in Russian

Sergey Vychegzhanin[1[0000-0001-6456-7856]], Vladimir Milov[2[0000-0002-8746-0100]] and Evgeny Kotelnikov[1,3[0000-0001-9745-1489]]

[1] Vyatka State University, 36, Moskovskaya st., Kirov, 610000, Russian Federation
[2] Nizhny Novgorod State Technical University n.a. R.E. Alekseev, 24, Minin st., Nizhny Novgorod, 603950, Russian Federation
[3] ITMO University, 49A, Kronverksky Pr., St. Petersburg, 197101, Russian Federation
{vychegzhaninsv, vladimir.milov, kotelnikov.ev}@gmail.com

Abstract. Text categorization is one of the important areas of research in the field of natural language processing and machine learning. The relevance of the topic is due to the demand for automatic categorization methods for the operational processing of the growing volume of news content published in online publications and social networks. The article investigates the influence of the feature selection procedure on the performance of machine learning methods for solving the problem of categorizing news articles: Logistic Regression, Light Gradient Boosted Machine, k-Nearest Neighbors, Random Forest, Naïve Bayes, Support Vector Machine and RuBERT. The research was carried out on the Russian corpus of documents containing texts from six topics: incidents, culture, economics, politics, society, sports. According to the results of experiments, for most of the considered methods, a positive effect of the feature selection procedure on the quality of categorization, the speed of analysis and the amount of memory consumed was noted. Of the considered classifiers, the RuBERT model made it possible to obtain the best average classification quality on a test corpus, reaching F1=0.882.

Keywords: Text categorization, machine learning, deep learning, feature selection.

1 Introduction

The Internet contains a huge amount of text data, which is growing rapidly. Every day, a large number of text news is published on various web resources by the media and users, which requires systematization, therefore, an important area of research in the field of natural language processing is the development of effective systems for automatic categorization of text documents. Text categorization is the comparison of texts with predefined labels (classes). This paper provides a comparative analysis of

popular machine learning methods in relation to solving the problem of categorizing news articles in Russian. The problem to be solved is the problem of multi-class classification of text documents.

2 Related work

There are many studies devoted to solving the problem of categorizing news articles in different languages using machine learning methods. The paper [1] evaluates the performance of real-time machine learning methods for classifying English news from the BBC website into five topics: business, entertainment, politics, sports and tech. Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT) and Random Forest (RF) are used as classifiers. The authors perform feature selection using TF-IDF. The highest value of the accuracy was obtained using LR and is equal to A=95.5%.

The article [2] also evaluates the performance of news texts from the BBC corpus. In this case, the classifiers NB, SVM, Multilayer Perceptron Neural Network, RF and DT are used. TF-IDF is used for feature selection. In this work, NB was the best in quality, which made it possible to obtain an accuracy equal to A=96.8%.

Sreedevi et al. [3] investigate bag-of-words and bag-of-n-gram text representation models, as well as four machine learning methods: SVM, NB, k-Nearest Neighbors (kNN) and Convolutional Neural Network. Testing of methods is performed on 20 NewsGroup and AG's News corpora. According to the results of the experiments, the highest value of the accuracy was obtained using the SVM with bag-of-words model and is equal to A=90.8% for the 20 NewsGroup corpus and is equal to A=85.14% for the AG's News corpus. Also in the article, the authors provide estimates of the training time and prediction time for algorithms.

Luo [4] in his research applies the technique of selecting text features based on a cross-validation procedure. SVM, NB and LR are used to classify news. Testing of methods is performed on three text corpora: 1) Data1 is categorized into *women*, *sports*, *literature*, *campus*; 2) Data2 is categorized into *sport*, *constellation*, *game*, *entertainment*; 3) Data3 is categorized into *science and technology*, *fashion*, *current event*. For the Data1 and Data2 corpora, the best results in terms of the classification quality were obtained using SVM and are equal to F1=0.86 and F1=0.71, respectively. For the Data3 corpus the best estimate was obtained using LR and is equal to F1=0.63.

There are a number of works in which the problem of classification of news articles is solved for the Arabic language. The article [5] uses a corpus from the BBC website, containing news from 7 topics, and a corpus from CNN website, containing news from 6 topics. The authors investigate the influence of preprocessing on the quality of classification. Three stemming techniques and twelve methods of weighting terms are explored. C4.5, NB и Discriminative parameter learning for Bayesian networks for text (DMNBtext) are used as classifiers. Experimental results showed that the DMNBtext algorithm achieves higher performance compared to other machine learning algorithms.

Qadi et al. [6] categorize news articles into four topics: business, sports, technology and Middle East. Weights of terms during text vectorization are determined using TF-IDF. The paper explores 10 popular classical machine learning methods. According to the experimental results, the best result F1=97.9 belongs to SVM, and the worst result F1=87.7 belongs to Ada-Boost.

The work [7] explores 9 neural network models using corpora AR-5, KH-7, AB-7, RT-40, the numbers in the title of which correspond to the number of topics. On the AR-5 corpus the best accuracy is A=97.41% (Bidirectional Gated Recurrent Unit), on the KH-7 corpus the best accuracy is A=96.86% (Convolutional Gated Recurrent Unit), on the AB-7 corpus the best accuracy is A=94.00% (Convolutional Gated Recurrent Unit), on the RT-40 corpus the best accuracy is A=64.24% (Convolutional Neural Network).

This study has the following differences from the existing ones: 1) the problem of topic classification is solved for Russian; 2) the influence of the number of the most relevant features, selected on the basis of TF-IDF weights, on the quality of news classification by topics is investigated; 3) the comparison of traditional machine learning methods with the modern neural network model BERT, showing state-of-the-art results in many natural language processing problems, is made; 4) the training time of the models is estimated, as well as the amount of memory required to store the models.

3 Materials and methods

3.1 Method for solving the problem of topic classification

The solution to the problem of topic classification consists of the following stages:

1. Pre-processing of text corpus documents.

At the pre-processing stage, html tags and stop words are removed from the texts, and the tokenization of the texts is performed.

Separate word forms are used as features.

2. Feature selection.

When the procedure for feature selection is performed, it is required to determine their weights. As a method of weighting features, the statistical measure Term Frequency – Inverse Document Frequency (tfidf) is often used, which for term t and document d in collection D is calculated by the formula:

$$\text{tfidf}(t,d) = f_{t,d} \cdot \log \frac{|\text{D}|}{n_t}, \quad (1)$$

where $f_{t,d}$ – term frequency t in document d; $|\text{D}|$ – the total number of documents in the collection D; n_t – the number of documents in collection D, in which the term t occurs.

After calculating the tfidf, the features are ranked in descending order of weights. The first n features with the highest weight are selected as the most relevant ones.

3. Building classification models.

Popular models that have performed well in many research papers are Logistic Regression, Light Gradient Boosted Machine (LGBM), k-Nearest Neighbors, Random Forest, Naïve Bayes, Support Vector Machine, Bidirectional Encoder Representations from Transformers, pretrained on Russian Wikipedia and news articles (RuBERT). Note that the RuBERT neural network model differs from the rest of the listed machine learning models in that it itself forms a vector representation of the text, and then can be fine-tuned to solve a specific problem, in particular, the classification of texts. Other models work with external vector representations such as tfidf.

4. Calculation of performance measures.

Performance measures are used to assess the overall performance of classification models.

3.2 Text corpus

To solve the problem of multiclass topic classification, a text corpus was formed from news articles, each of which belongs to one of six large topics: accidents, culture, economics, politics, society, sports. The articles were taken from the Internet portals ”Gazeta.ru”, ”Lenta.ru”, ”Komsomolskaya Pravda”, ”RBK” and news agencies ”Interfax”, ”ITAR-TASS”, ”RIA Novosti” for the period from 2010 to 2020. The number of texts in each of the topics is presented in Table 1.

Table 1. Distribution of news articles by topics.

Topic	Number of texts
Incidents	10,008
Culture	18,706
Economics	38,423
Politics	13,765
Society	13,832
Sports	33,253

The created text corpus is unbalanced. The largest topic ”Economics“ contains 38,423 texts. The topic ”Incidents“ has the smallest size, which contains 10,008 texts.

The markup of news articles by topics was carried out on the basis of the topics indicated for these articles on the information resource from which they were taken.

3.3 Design of the experiments

The experiments were carried out on a computer with an Intel (R) Xeon (R) CPU @ 2.30GHz and a Tesla K80 video card. The experiments were carried out using the Python programming language. Seven machine learning methods were used to categorize texts, as described in subsection 3.1. The software implementation of the LR, RF, NB and SVM methods is taken from the *scikit-learn* library [8], the LGBM

method is taken from the *lightgbm* library [9], the RuBERT model is taken from the *DeepPavlov* library [1].

Word forms are features of the text. The number of features with the highest TF-IDF value taken into account in the text representation model was taken equal to 0.01*N*, 0.05*N*, 0.1*N*, 0.25*N*, 0.5*N* and *N*, where *N* is the total number of features in the training corpus.

The performance of the categorization was determined by the F1-score calculated by the formula:

$$F1 = \frac{2 \cdot P \cdot R}{P + R}, \tag{1}$$

where P – precision; R – recall. Macro-averaging was applied to obtain the average value of the F1-score.

4 Results

The total number of features (word forms) in the training corpus was N=559,108. The average values of the F1-score, obtained using seven classifiers for a different number of features with the highest weight, are presented in Table 2 and Figure 1.

Table 2. Average values of F1-score.

Number of features	Classifiers						
	LR	LGBM	kNN	RF	NB	SVM	RuBERT
0.01*N*	0.821	0.850	0.779	**0.805**	0.765	0.865	0.809
0.05*N*	0.841	**0.857**	0.782	**0.805**	0.784	0.874	0.820
0.1*N*	0.844	0.856	**0.787**	0.802	**0.792**	**0.877**	0.819
0.25*N*	0.845	0.854	0.784	0.801	0.786	0.876	0.813
0.5*N*	**0.847**	0.856	0.782	0.796	0.759	0.875	0.808
N	**0.847**	0.854	0.779	0.793	0.701	0.874	**0.882**

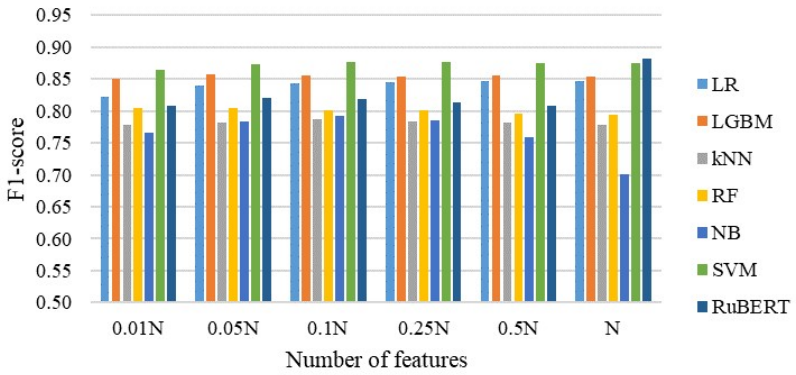

Fig. 1. F1-score for topic classification.

Tables 3 and 4 show the training time of the models and the amount of memory required for them, respectively.

Table 3. Model training time, hour.

Number of features	Classifiers						
	LR	LGBM	kNN	RF	NB	SVM	RuBERT
0.01*N*	0.023	0.529	0.007	0.411	0.010	2.142	3.028
0.05*N*	0.029	0.969	0.006	0.429	0.010	5.885	3.137
0.1*N*	0.034	1.101	0.007	0.428	0.010	7.548	3.115
0.25*N*	0.043	1.147	0.007	0.411	0.010	10.219	3.244
0.5*N*	0.054	1.106	0.007	0.476	0.010	9.411	3.225
N	0.059	1.072	0.007	0.705	0.010	8.798	3.397

Table 4. The amount of memory required for the models, Mb.

Number of features	Classifiers						
	LR	LGBM	kNN	RF	NB	SVM	RuBERT
0.01*N*	0.34	3.52	160.67	536.07	0.68	88.35	678.48
0.05*N*	1.71	4.39	215.55	642.30	3.41	175.99	678.48
0.1*N*	3.41	5.39	232.34	711.10	6.83	217.42	678.48
0.25*N*	8.53	7.25	246.38	826.35	17.06	250.21	678.48
0.5*N*	17.06	9.80	252.20	914.60	34.13	256.77	678.48
N	34.13	14.88	256.02	1013.12	34.13	265.68	678.48

The values of performance measures for the classification of news articles by topics using two leading models of the considered models – RuBERT with N features and SVM with $0.1N$ features – are presented in Table 5.

Table 5. Precision, recall and F1-score by topics for classifiers SVM and RuBERT.

Topic	SVM ($0.1N$ features)			RuBERT (N features)		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Incidents	0.858	0.811	0.834	0.885	0.767	0.822
Culture	0.965	0.950	0.958	0.968	0.959	0.964
Economics	0.963	0.941	0.952	0.958	0.962	0.960
Politics	0.840	0.827	0.833	0.789	0.870	0.827
Society	0.654	0.753	0.700	0.720	0.743	0.731
Sports	0.985	0.982	0.983	0.994	0.985	0.990
Average	0.878	0.877	0.877	0.886	0.881	0.882

5 Discussion

From Table 2 it follows that feature selection can improve the performance of classification of news articles by topics for most machine learning methods. For the LGBM method the best classification quality was obtained at $0.05N$ features, for RF – at $0.01N$ and $0.05N$ features, for kNN, NB and SVM – at $0.1N$ features. The feature selection for LR and RuBERT did not improve the quality of the classification. For these methods the highest F1-score is achieved with the full set of features.

Among the considered classifiers, the RuBERT model showed the best results, reaching F1=0.882. The second result in the quality of classification belongs to the SVM method and is equal to F1=0.877.

Based on Tables 3 and 4, we can conclude that a decrease in the number of features has a positive effect on the performance of classifiers. As the number of features decreases, the training time for LR, LGBM, RF and SVM models decreases, and the amount of memory required for all models, except for RuBERT, decreases. The SVM classifier turned out to be the longest in terms of training time. It took about 7.5 hours to train it with $0.1N$ features. The best quality RuBERT method learned 2.2 times faster than SVM – in about 3.4 hours. The RF and RuBERT models turned out to be the most demanding in terms of the amount of memory, while LR, LGBM and NB required an order of magnitude less memory for storing models on average.

Analysis of Table 5 shows that the topics ”Culture“, ”Economics“ and ”Sports“ are recognized the best by classifiers (F1 varies from 0.952 to 0.990), the topic “Society” is recognized the worst of all (F1=0.700 for SVM and F1=0.731 for RuBERT) due to the fact that this topic may contain texts that also belong to five other topics. The largest gap in the F1-score for the SVM and RuBERT models (3.1 percentage points (p.p.)) is observed in the topic “Society” in favor of RuBERT due to the higher precision of this model (6.6 p.p. higher than SVM). However, SVM has 5.1 p.p. higher precision than RuBERT for ”Politics“. The SVM classifier provides higher recall in

the topic ”Incidents“ (4.4 p.p. higher than RuBERT), and RuBERT provides higher recall in the topic ”Politics“ (4.3 p.p. higher than SVM).

6 Conclusion

The problem of text categorization is of great practical importance and can be solved using machine learning methods. The efficiency of solving the problem is significantly influenced by data pre-processing, including the selection of the most relevant features. This study investigates the influence of the number of features selected at the feature selection stage on the performance of seven classifiers, among which there are both the classic well-proven SVM and LGBM, and the relatively new and popular BERT. It was found that the feature selection in most cases improves the quality of the classification, but not for all classifiers it gives a positive effect.

Among the considered machine learning methods, the best average classification quality for six topics was obtained using BERT and was equal to F1=0.882. On average for topics (Table 5), RuBERT slightly surpasses SVM in both precision and recall. The topics ”Culture“, ”Economics“ and ”Sports“ were most easily recognized by the classifiers, the topic ”Society” turned out to be the most difficult to recognize.

In future works, it is planned to investigate the effectiveness of machine learning methods for solving the problem of multi-label classification of news articles in Russian.

References

1. Patro, A. et al.: Real Time News Classification Using Machine Learning. In: International Journal of Advanced Science and Technology, 29(9), 620–630 (2020).
2. Deb, N. et al.: A Comparative Analysis of News Categorization Using Machine Learning Approaches. International journal of scientific and technology research, 9(1), 2469–2472 (2020).
3. Sreedevi, J., Bai, M.R., Reddy, M.C.: Newspaper Article Classification using Machine Learning Techniques. International Journal of Innovative Technology and Exploring Engineering, 9(5), 872–877 (2020).
4. Luo, X.: Efficient English text classification using selected Machine Learning Techniques. Alexandria Engineering Journal, 60(3), 3401–3409 (2021).
5. Alshammari, R.: Arabic Text Categorization using Machine Learning Approaches. International Journal of Advanced Computer Science and Applications, 9(3), 226-230 (2018).
6. Qadi, L.A. et al.: Arabic Text Classification of News Articles Using Classical Supervised Classifiers. In: Proceedings of the 2nd International Conference on new Trends in Computing Sciences (2019).
7. Elnagar, A., Al-Debsi, R., Einea, O.: Arabic text classification using deep learning models. Information Processing and Management, 57(1), 1–17 (2020).
8. Pedregosa, F. et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830 (2011).
9. Light Gradient Boosting Machine Homepage, https://github.com/microsoft/LightGBM, last accessed, 2021/06/19.

10. Burtsev, M. et al.: DeepPavlov: Open-source library for dialogue systems. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, 122–127 (2018).