Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas

Eureka Journal of Education & Learning Technologies (EJELT)

ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

COMPARISON OF THE EFFICIENCY OF VARIOUS MACHINE LEARNING ALGORITHMS IN MULTI-CLASS TEXT CLASSIFICATION

Muhamediyeva D. T.
Mamatov A. A.
National Research University "Tashkent Institute of Irrigation and Agricultural Mechanization Engineers", Namangan State University
email: dilnoz134@rambler.ru
email: abduvali_mamatov@mail.ru

Abstract

This article considers the issue of automatic classification of text data. In the study, documents were transferred to a numerical character space using TF-IDF vectorization technology based on Machine Learning and Natural Language Processing methods. The accuracy, training and testing time of Logistic Regression, Ridge Classifier, k-Nearest Neighbors, Random Forest, Linear Support Vector Machine, Stochastic Gradient Descent, Nearest Centroid and Complement Naive Bayes algorithms in the classification process were analyzed. The experiments were carried out on four categories of 20 Newsgroups text sets. The results showed that there are significant differences in accuracy and computational speed between the algorithms and proved the importance of choosing an optimal model for real-time text analysis systems.

Keywords: Text classification, TF-IDF, machine learning, NLP, Logistic Regression, Support Vector Machine, Naive Bayes, Random Forest, model efficiency, algorithm comparison.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

Introduction

In recent years, the sharp increase in the volume of digital information has made the problem of automatic processing and analysis of text data an urgent issue. Fast and accurate classification of large volumes of texts created in Internet news, scientific articles, social networks and forums is one of the important tasks of modern information systems. The main problem in the process of text classification is to convert natural language data into a numerical form that can be understood by a computer and to choose the most effective classification algorithm. Therefore, feature extraction and optimization of the classification model remain one of the central directions of scientific research [1-4]. In this work, a mathematical representation of documents is generated based on the TF-IDF vectorization method and the performance of various machine learning algorithms is compared. The results of the research are of great importance in creating automatic text classification systems, optimizing information retrieval processes, and developing intelligent information systems [5-9].

2. Methodology

In this study, the problem of multi-class text classification is solved based on statistical and optimization approaches. A collection of text documents

$$D = \{d_1, d_2, \ldots, d_N\}$$

Each document is given in the form of d_i a specific class label

$$y_i \in \{1, 2, \ldots, K\}$$

belongs to

Feature extraction based on TF-IDF

The TF-IDF (Term Frequency – Inverse Document Frequency) method is used to transform texts into numerical space.

In the document t frequency of the term:

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

$$TF(t,d)=\frac{f_{t,d}}{\sum f_{t,d}},$$

here $f_{t,d}-t$ of the term d number of occurrences in the document.

Frequency of reverse documents:

$$IDF(t)=\log\frac{N}{1+n_t},$$

here n_t- number of documents containing the term.

Resulting weight:

$$w_{t,d}=TF(t,d)\cdot IDF(t).$$

Documents are represented in the form of a high-dimensional sparse matrix.

$$X\in\square^{N\times M}$$

Logistic Regression model

The multi-class probabilistic model is defined as follows:

$$P(y=k|\ x)=\frac{\exp(w_k^T x+b_k)}{\sum_{j=1}^{K}\exp(w_j^T x+b_j)}.$$

The model parameters are determined by minimizing the log-loss function:

$$\min_w -\sum_{i=1}^{N}\log P(y_i\mid x_i)+\lambda\parallel w\parallel_2^2.$$

Ridge Classifier

Ridge classification is based on a quadratic loss function:

$$\min_w \parallel Xw-y\parallel_2^2+\alpha\parallel w\parallel_2^2,$$

here $\alpha-$ regularization parameter.

Support Vector Machine

In a linear SVM model, the optimal separating hyperplane is found through the following optimization problem:

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i ,$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i , \quad \xi_i \geq 0 .$$

Stochastic Gradient Descent

The parameters are updated iteratively:

$$w_{t+1} = w_t - \eta_t \nabla L(w_t) ,$$

here η_t – learning speed.

k-Nearest Neighbors

The classification decision is based on Euclidean distance:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{M} (x_{ik} - x_{jk})^2} .$$

Test point class:

$$\hat{y} = \arg\min_{c} \sum_{x_i \in N_k} I(y_i = c) .$$

Random Forest

Ensemble model decision:

$$\hat{y} = mode\{T_1(x), T_2(x), \ldots, T_B(x)\} ,$$

here T_b – separate decision tree.

Naive Bayes klassifikatori

Based on Bayes' rule:

$$P(y \mid x) = \frac{P(x \mid y) P(y)}{P(x)} .$$

Under the assumption of independence:

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

$$P(x \mid y) = \prod_{j=1}^{M} P(x_j \mid y) .$$

Evaluation criteria

Model accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} .$$

Training and testing time are also evaluated as performance criteria .

3. Results

A text set consisting of 2034 training documents and 1353 test documents was used in the experiment. As a result of TF-IDF vectorization, a high-dimensional sparse space with 7831 features was generated.

Table 1 Comparison of classification algorithms in terms of accuracy and time

№	Model	Training time (s)	Test time (s)
1	Logistic Regression	0.376	0.0024
2	Ridge Classifier	0.034	0.000
3	k-Nearest Neighbors	0.002	0.124
4	Random Forest	1.120	0.037
5	Linear SVC	0.057	0.0010
6	SGD (log-loss)	0.022	0.0010
7	Nearest Centroid	0.003	0.232
8	Complement Naive Bayes	0.002	0.000

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
OPEN ACCESS JOURNALS

Eureka Journal of Education & Learning Technologies (EJELT)

ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

Table 2 Computational efficiency indicators of models

Model	Complexity (Training)	Complexity (Testing)	Practical adjustment
Logistic Regression	O(NM)	O(M)	High accuracy and fast testing
Ridge Classifier	O(NM)	O(M)	Very fast training
kNN	O(1)	O(NM)	Testing very slow
Random Forest	O(BNlogN)	O(BlogN)	High resource requirement
Linear SVC	O(NM)	O(M)	Optimal model
SGD	O(kNM)	O(M)	Fastest linear model
Nearest Centroid	O(NM)	O(KM)	Testing slow
Complement NB	O(NM))	O(M)	Most efficient model

Table 3 Overall rating of models

Rating	**Model**	**Evaluation Criteria**
1	Complement Naive Bayes	Highest accuracy + fastest
2	SGD Classifier	High accuracy + very fast training
3	Ridge Classifier	Balanced model
4	Logistic Regression	Stable result
5	Linear SVC	High quality classifier
6	kNN	Large testing time
7	Nearest Centroid	Lower accuracy
8	Random Forest	Slowest training

The results show the following:

$Accuracy_{\text{CNB}} > Accuracy_{\text{SGD}} \approx Accuracy_{\text{Ridge}}$

and computational efficiency:

$Time_{\text{test}}^{kNN}$ □ $Time_{\text{test}}^{\text{LinearModels}}$.

This confirms the superiority of probabilistic models in high-dimensional text space

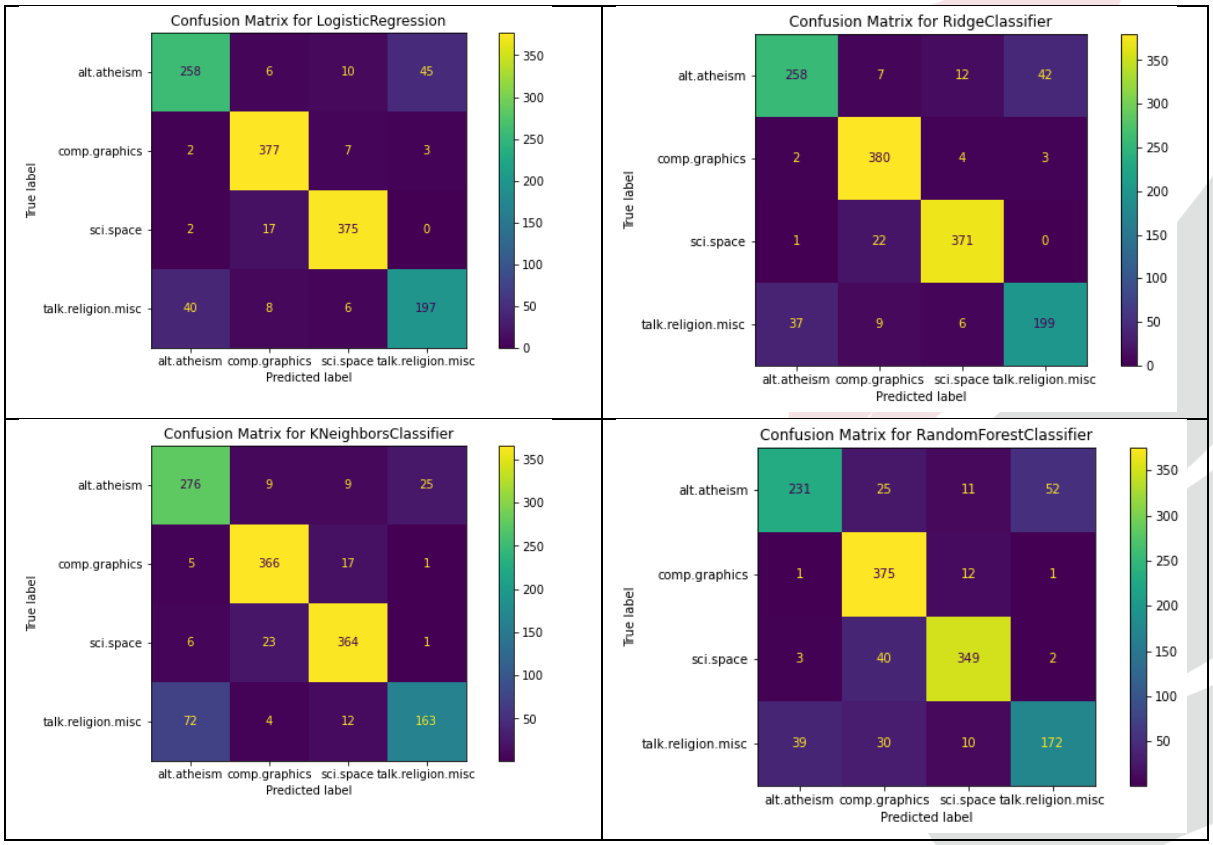
The interclass confusion for each model is represented by the following matrix:

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

$$C_{ij} = \left|\{x : y = i, \hat{y} = j\}\right|.$$

The analysis showed that:
Documents on religious topics are often confused
Technical topics are more clearly separated
This is due to semantic proximity.

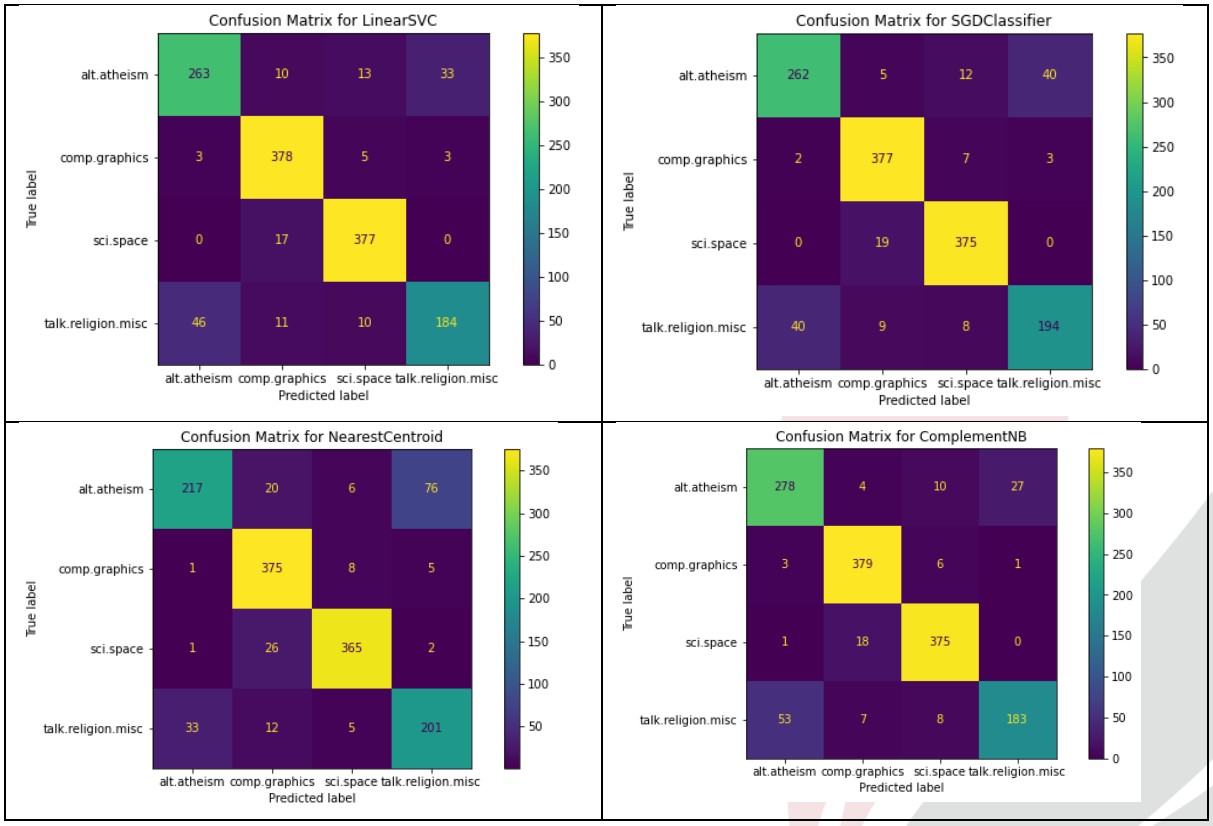


Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
OPEN ACCESS JOURNALS

Eureka Journal of Education & Learning Technologies (EJELT)

ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2



The Accuracy – Training Time graph shows the relationship between the accuracy of classification models and the time spent on their training. It can be seen from the graph that most linear classification models provide high accuracy with a relatively small training time. In particular, the points of the SGD, Ridge Classifier and Linear SVC models are located in the optimal region of the graph, which provides a balance between computational efficiency and classification quality.

Although the Logistic Regression model has high accuracy, it was observed that its training time was slightly higher than other linear models. The Random Forest model is located in the upper right part of the graph, which indicates that it has

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

the largest training time. This is explained by the complex structure of ensemble models and the need to build a large number of decision trees.

The Complement Naive Bayes model is in the most favorable position on the graph, providing high accuracy with minimal training time. This shows that probabilistic models work effectively in high-dimensional text space.

The Accuracy – Test Time graph represents the relationship between the speed and accuracy of classification models during testing. It can be seen from the graph that linear models work very quickly during testing and maintain high accuracy. This is explained by the fact that their decision-making process is based on a simple linear function.
The k-Nearest Neighbors model stands out as the model with the largest testing time on the graph. This model requires large computing resources due to the need to compare each new sample with all training samples during testing.

The Nearest Centroid model also showed relatively poor results in terms of testing speed. The Random Forest model, on the other hand, showed average results during testing, but its accuracy was lower than that of linear models. The Complement Naive Bayes model, on the other hand, provided the highest accuracy with minimal testing time, making it the most suitable model for real-time text classification systems.

4. Conclusion

In this study, the effectiveness of various machine learning algorithms in multi-class classification by vectorizing texts based on TF-IDF was comprehensively studied. Experimental results showed that linear classification models have an advantage in accuracy and computational speed in high-dimensional and sparse text space. The Complement Naive Bayes model was determined as the optimal

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

model, providing the highest accuracy and the smallest computational time. The high efficiency of the SGD and Ridge Classifier models also showed that they are of practical importance in working with large amounts of text data. It was found that ensemble and distance-based models may be less effective in real-time systems due to their high computational complexity. The results of the study confirm the importance of choosing an effective model in automatic text classification, optimizing information retrieval systems, and creating intelligent analysis platforms. In the future, the use of deep learning models and research on semantic vectorization methods will allow further improvement of classification accuracy.

References

1. Liu W., Quan X., Feng M., Qiu B. A short text modeling method combining semantic and statistical information // Information Sciences. – 2010. – Vol. 180, No. 20. – P. 4031–4041.
2. Kalchbrenner N., Grefenstette E., Blunsom P. A convolutional neural network for modelling sentences // Proceedings of ACL. – 2014. – P. 655–665.
3. Conneau A., Schwenk H., Barrault L., LeCun Y. Very deep convolutional networks for text classification // Proceedings of EACL. – 2017. – P. 1107 1116.
4. Lee J.Y., Dernoncourt F. Sequential short-text classification with recurrent and convolutional neural networks // Proceedings of NAACL-HLT. – 2016. – P. 515–520.
5. Zhang D., Tian L., Hong M., Han F., Ren Y., Chen Y. Combining convolution neural network and bidirectional gated recurrent unit for sentence semantic classification // IEEE Access. – 2018. – Vol. 6. – P. 73750–73759.
6. Qiu X.P., Sun T.X., Xu Y.G., Shao Y.F., Dai N., Huang X.J. Pre-trained models for natural language processing: a survey // Science China Technological Sciences. – 2020. – Vol. 63, No. 10. – P. 1872–1897.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Education & Learning Technologies (EJELT)
ISSN 2760-4918 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/2

7. Mikolov T., Chen K., Corrado G., Dean J. Efficient estimation of word representations in vector space // Proceedings of ICLR. – 2013. – P. 1–12.
8. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed representations of words and phrases and their compositionality // Advances in Neural Information Processing Systems (NIPS). – 2013. – P. 3111–3119.
9. Alsmadi I., Gan K.H. Review of short-text classification // International Journal of Web Information Systems. – 2019. – Vol. 15, No. 2. – P. 155–182. 16. Song G., Ye Y., Du X., Huang X., Bie S. Short text classification: a survey // Journal of Multimedia. – 2014. – Vol. 9, No. 5. – P. 635–643.
10. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems (NIPS). – 2017. – P. 5998–6008.
11. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
12. Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations // Proceedings of NAACL-HLT. – 2018. – P. 2227–2237.
13. Kim Y. Convolutional neural networks for sentence classification // Proceedings of EMNLP. – 2014. – P. 1746–1751.
14. Hochreiter S., Schmidhuber J. Long short-term memory // Neural Computation. – 1997. – Vol. 9, No. 8. – P. 1735–1780.