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

Eureka Journal of Computing Science & Digital Innovation (EJCSDI)

ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

# COMPARATIVE ANALYSIS OF THE EFFICIENCY OF CLASSIC MACHINE LEARNING MODELS USING DISTILBERT-BASED TEXT VECTORIZATION METHODS

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**

In this research work, the effectiveness of embeddings generated using the DistilBERT model based on deep learning and classical machine learning algorithms in the process of automatic text classification was comparatively studied. Within the framework of the research, Logistic Regression, Ridge Classifier, Linear SVC, SGD Classifier and Random Forest models were tested on selected categories of the 20 Newsgroups dataset. The texts were converted into contextual vectors using the transformer model and then transferred to classical classification algorithms. The analysis was carried out based on the accuracy level of the models, the F1 index and the training and testing times. The results of the research showed that transformer-based embeddings increase the effectiveness of classical machine learning models.

**Keywords:** Text classification, DistilBERT, transformer model, embedding, machine learning, classification, ensemble methods, artificial intelligence.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

## Introduction

Currently, the issue of automatic analysis and classification of large amounts of textual data is one of the important areas of artificial intelligence and machine learning. As a result of the rapid development of Internet networks, social platforms and scientific databases, the volume of textual data is increasing sharply. Effective processing of this data, extraction of important information from it and its use in automatic decision-making systems is an urgent scientific and practical issue. In recent years, deep learning models based on transformer architecture have shown high efficiency in natural language processing. At the same time, classical machine learning algorithms are also distinguished by their simplicity and computational efficiency. Therefore, the issue of integrating contextual embeddings generated using transformer models with classical classification algorithms is of particular scientific interest[1-3].

In the context of the increasing volume of textual data, there is a need to develop methods for their rapid and accurate classification. Although transformer models provide high accuracy, their computational complexity can be high. The use of classical machine learning algorithms in combination with transformer-based vectorization allows achieving high accuracy while saving computational resources. In this regard, the topic of this research is relevant [4-7].

The aim of the study is to compare the effectiveness of various classical machine learning algorithms based on text embeddings generated using the transformer-based DistilBERT model. The study considered the issues of forming and pre-processing a textual data set, generating contextual embeddings using the Distil BERT model, training and testing classical classification algorithms, evaluating the accuracy and speed of the models, and improving the results using the ensemble approach [8-10].

The study substantiates the possibility of increasing the efficiency of text classification by using transformer-based DistilBERT embeddings in combination with classical machine learning algorithms. A comparative analysis

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

based on the accuracy of the models, as well as their training and testing times, was carried out. A method for improving classification results using an ensemble approach is proposed. The results of the study can be used to create automatic classification systems for large amounts of text data. The proposed approach allows for effective use in information search systems, electronic document management, social network monitoring, and scientific database analysis. The integration of transformer and classical machine learning models helps to achieve high accuracy while saving computing resources [11-12].

## 2. Methodology

In this study, the process of automatic text classification consists of two main stages. In the first stage, contextual embeddings were generated using the transformer-based DistilBERT model, and in the second stage, these embeddings were transferred to classical machine learning algorithms and the classification process was performed.

### Vectorization of texts

A given set of texts

$D = \{(x_i, y_i)\}_{i=1}^{N}$

is viewed in the form of, where $x_i - i -$ text, $y_i \in \{1,2,\ldots,K\} -$ moss inf tag, $N -$ number of samples, $K -$ number of classes. Each text is mapped to a contextual vector using the DistilBERT model:

$h_i = f_\theta(x_i) \in \square^{d}$,

where $f_\theta -$ transformer model parameter set, $d = 768 -$ embedding size.

In the output layer of the model, the CLS token embedding is accepted as a generic text vector:

$z_i = CLS(h_i)$.

As a result, a matrix of vectors is formed for all texts:

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

$Z = [z_1, z_2, \ldots, z_N]^T \in \mathbb{R}^{N \times d}$.

**Data Standardization**

Some classical models are standardized to make their embeddings work efficiently. The mean and variance of each feature are calculated as follows::

$$\mu_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij}, \qquad \sigma_j = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_{ij} - \mu_j)^2}.$$

Then the normalized vectors:

$$\tilde{z}_{ij} = \frac{z_{ij} - \mu_j}{\sigma_j}$$

is obtained in the form of.

**Classification models**

Several classical machine learning algorithms have been used for text classification.

The logistic regression model is expressed by the likelihood function:

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

The model parameters are determined based on the maximum likelihood principle. In the Ridge classifier, the following regularized loss function is minimized:

$$L(w) = \sum_{i=1}^{N} (y_i - w^T z_i)^2 + \alpha \| w \|_2^2,$$

where $\alpha -$ is the regulation coefficient.

A linear support vector machine is based on solving the following optimization problem:

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

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

$y_i(w^T z_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0,$

where $C$ – penalty parameter.

The random forest algorithm consists of an ensemble of decision trees, and the final decision is determined based on majority vote.:

$\hat{y} = mode\{h_t(z)\}_{t=1}^{T},$

here $T$ – is the number of trees.

**Ensemble approach**

The results of several classifiers are combined using a hard voting method. The final class is determined as follows:

$$\hat{y} = \arg\max_k \sum_{m=1}^{M} I(y_m = k),$$

where $M$ – is the total number of models, $I(\cdot)$ – is the indicator function..

**Evaluation criterion**

The effectiveness of models is evaluated through the accuracy indicator:

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

Precision, completeness, and F1-measure are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN},$$

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

## 3. Results and Analysis

The study compared the performance of several classical machine learning algorithms based on contextual embeddings generated using the DistilBERT

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

model. The results of the models based on precision, recall, F1 score, and computation times are presented in Table 1 below.

Table 1 Model performance results

| Model | Accuracy | Precision | Recall | F1 | Train time (s) | Test time (s) |
|---|---|---|---|---|---|---|
| Logistic Regression | 0.825 | 0.834 | 0.825 | 0.825 | 0.491 | 0.002 |
| Ridge Classifier | 0.788 | 0.790 | 0.788 | 0.783 | 0.155 | 0.000 |
| Linear SVC | 0.788 | 0.785 | 0.788 | 0.785 | 0.303 | 0.000 |
| SGD Classifier | 0.775 | 0.785 | 0.775 | 0.778 | 0.048 | 0.000 |
| Random Forest | 0.813 | 0.813 | 0.813 | 0.807 | 0.280 | 0.008 |
| Voting Ensemble | **0.825** | **0.839** | **0.825** | **0.819** | 0.250 | 0.000 |

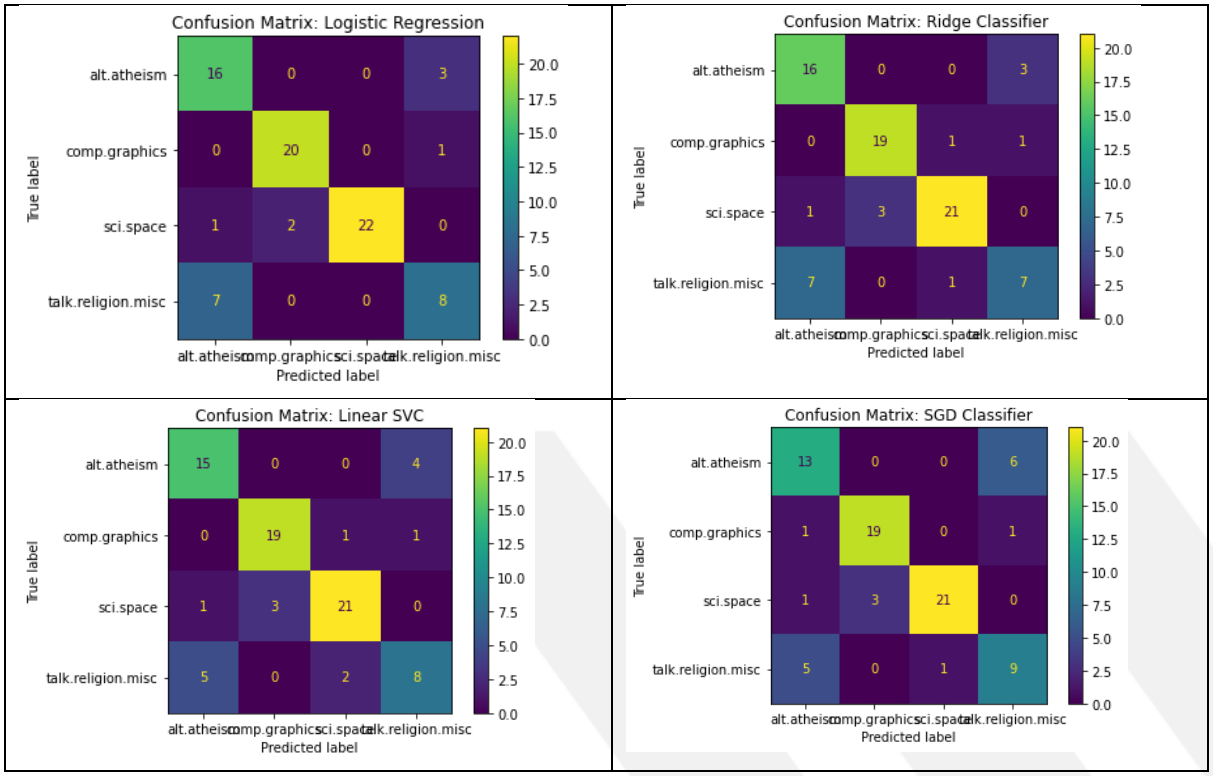
The results show that the transformer-based embeddings provide a sufficiently informative set of features for all classical machine learning models. The highest accuracy among the models was observed for the Logistic Regression and Voting Ensemble methods, which were equal to 0.825. This indicates that the models with a linear separating boundary work effectively in high-dimensional embedding spaces.

The Voting Ensemble model provided the highest precision value (0.839), which indicates that the number of false positive classifications can be reduced by combining the decisions of different classifiers. At the same time, the F1 indicator is also high, indicating the overall balanced performance of the model. Although the Random Forest model showed relatively high precision (0.813), its testing time was longer than other models. This is explained by the computational complexity of ensemble tree methods. The Ridge Classifier and Linear SVC models showed average performance. The precision level in these models was

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

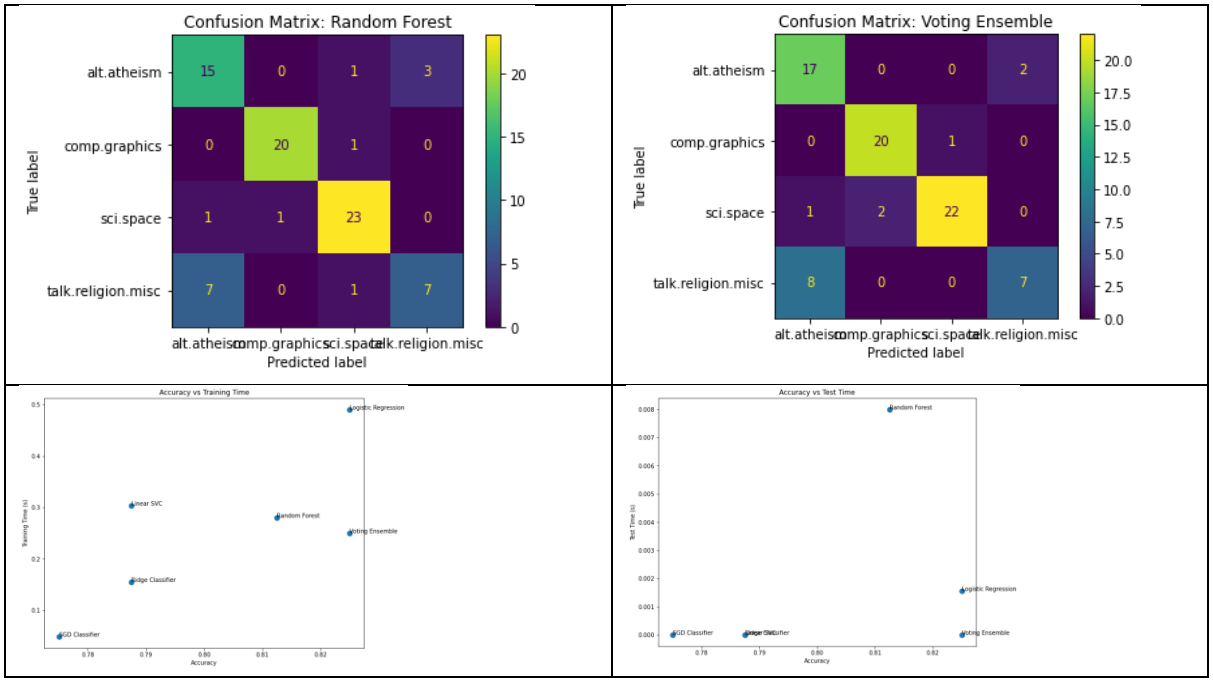
https://eurekaoa.com/index.php/10

around 0.788, indicating a partial linear separation of classes in the embedding space. The occurrence of a convergence warning in the Linear SVC model may be due to an insufficient number of iterations. The SGD Classifier was the fastest trained model, with a training time of only 0.048 seconds. However, the precision indicator was lower than other models. This indicates that the accuracy of gradient-based optimization algorithms is slightly reduced at the expense of speed. Overall, the experimental results show that linear classifiers and ensemble approaches, combined with transformer embeddings, provide the most effective results.



Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10




Based on the experimental results, the dependence of the accuracy of the models on the calculation time was analyzed using scatter plots. The Accuracy–Training Time graph allowed us to visually assess the balance between the efficiency of the models and the computational complexity. It can be seen from the graph that the Logistic Regression model requires a relatively large training time while providing high accuracy. This is explained by the complexity of the iterative optimization process of the model parameters. Although the SGD Classifier model showed the smallest training time, its accuracy was slightly lower. This indicates that the accuracy may decrease at the expense of the speed of gradient-based optimization methods. The Ridge Classifier and Linear SVC models demonstrated average accuracy and average training time, which confirms the stable operation of linear classifiers in high-dimensional embedding spaces. The Accuracy–Test Time graph allowed us to assess the speed of the models in practical applications. In this graph, it was observed that the test time of almost

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

all models was very small. The Random Forest model is characterized by a slightly larger testing time, which is explained by the additional computational costs of the forecasting process through many decision trees. The Voting Ensemble model, on the other hand, provided a very small testing time while maintaining high accuracy, which demonstrates the effectiveness of the ensemble approach. Scatter plots served as an important tool in determining the trade-off between the accuracy and computational speed of the models.

The results of the study showed that the use of transformer-based DistilBERT embeddings significantly increases the efficiency of classical machine learning algorithms. Since the embeddings reflect the semantic and contextual features of the text with high accuracy, linear classifiers also provided high results. The results show that the Logistic Regression and Voting Ensemble models demonstrated the highest accuracy. This indicates a relatively linear separation of classes in a high-dimensional space. The high performance of the Random Forest model is also explained by the ability to detect nonlinear relationships. The fast performance of the SGD Classifier model makes it a promising option for working with large amounts of data. However, the slightly lower accuracy level indicates the need to further improve the optimization parameters. The convergence problem observed in the Linear SVC model can be eliminated by increasing the number of iterations or optimizing the regularization parameters. The high efficiency of the ensemble approach shows that it is possible to improve the overall classification quality by combining the advantages of different models. This confirms the importance of integrating multiple algorithms in real-world application systems..

## 4. Conclusion

In this study, the effectiveness of text embeddings generated using the transformer-based DistilBERT model and classical machine learning algorithms was comparatively analyzed. The experimental results showed that the

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

embeddings effectively represent the semantic structure of the text and increase the accuracy of various classification algorithms. According to the results obtained, the Logistic Regression and Voting Ensemble models provided the highest level of accuracy. The Random Forest model, while showing relatively high results, was found to be somewhat more complex in terms of computational costs. The SGD Classifier model was distinguished by its speed. The results of the study showed that the combined use of transformer embeddings and classical machine learning algorithms allows achieving high efficiency. The proposed approach is of practical importance in areas such as automatic classification of large amounts of text data, information search systems, social media monitoring, and scientific document analysis. In the future, the study can be further developed using a larger dataset, classifiers based on deep neural networks, and hybrid ensemble models.

## References

1. 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.
2. 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.
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.

Eureka Open Access Journals
Open Access | Peer Reviewed | International Journals | Multidisciplinary Areas
Eureka Journal of Computing Science & Digital Innovation (EJCSDI)
ISSN 2760-4993 (Online) Volume 2, Issue 3, March 2026

https://eurekaoa.com/index.php/10

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.
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. 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.
10. Kalchbrenner N., Grefenstette E., Blunsom P. A convolutional neural network for modelling sentences // Proceedings of ACL. – 2014. – P. 655–665.
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.