**DOI: 10.37943/14DKRN4681**

**Shynar Mussiraliyeva**
Candidate of Physical and Mathematical Sciences, Head of the department "Information systems"
mussiraliyevash@gmail.com, orcid.org/0000-0001-5794-3649
Al-Farabi Kazakh National University, Kazakhstan

**Milana Bolatbek**
PhD., Senior Lecturer of the department "Information systems"
bolatbek.milana@gmail.com, orcid.org/0000-0002-2153-180X
Al-Farabi Kazakh National University, Kazakhstan

**Aigerim Zhumakhanova**
Master of Technical Sciences, Lecturer of the department "Information systems"
aygerim129@gmail.com, orcid.org/0009-0008-0210-4037
Al-Farabi Kazakh National University, Kazakhstan

**Zhanar Medetbek**
Master of Military Affairs and Security, Lecturer of the department "Information systems"
medetbek.zhanar@gmail.com, orcid.org/0000-0001-7536-5889
Al-Farabi Kazakh National University, Kazakhstan

**Moldir Sagynay**
Master of Technical Sciences, Lecturer of the department "Information systems"
sagynaymoldir11@gmail.com, orcid.org/0009-0004-1377-5742
Al-Farabi Kazakh National University, Kazakhstan

# COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORTMS TO IDENTIFY EXTREMIST TEXTS IN THE KAZAKH LANGUAGE

**Abstract**. The article explores various models and methods employed in classifying text content with the aim of identifying destructive information within social networks. The study focuses on utilizing machine learning techniques, such as support vector machines, naive Bayes classifiers, random tree methods, decision tree, k-Nearest Neighbors algorithm, logistic regression, gradient boosting to identify extremist texts. The research findings showcase the effectiveness of these methodologies in the identification process.

The article also offers an overview of existing research, methodologies, and software products in the analysis of extremist texts, emphasizing the importance of case-based learning, deductive learning models, and automated data collection and analysis. Additionally, the article provides an overview of existing research, methods, and software products within the field of analyzing extremist texts. It highlights the significance of case-based learning and the use of deductive learning models, as well as automated data collection and analysis techniques. These approaches contribute to the overall understanding and detection of extremist content.

The article further discusses the relevance and future prospects of the presented research. It emphasizes the need to expand the corpus of documents studied, enabling a more comprehensive analysis of texts, including those in photo, audio, and video formats. The development of complex models for recognizing hidden extremist propaganda is also identified as a key direction for future work.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

By addressing these areas of focus, the research presented in the article aims to advance the field of identifying and combating extremist content within social networks. The incorporation of advanced techniques and technologies is crucial to effectively detect and address the presence of such content in various forms and formats.

**Keywords**: machine learning model, classification, extremist text.

**Introduction**

Machine learning (from the English machine learning, ML) is a class of artificial intelligence methods, the characteristic feature of which is learning using solutions to many similar problems, rather than directly solving the problem. Mathematical statistics, numerical methods, mathematical analysis, optimization methods, probability theory, graph theory, various methods of working with data in numerical form are used to construct such methods [1].

Machine learning technology based on data analysis dates back to 1950, when the first programs for checkers games were developed. With the growth of computing power of computers, the patterns and forecasts that they create have become many times more complicated, and the range of problems and tasks solved with the help of machine learning has expanded.

There is such a type of training:

1) case-based learning or inductive learning is based on the identification of empirical patterns in the data.
2) deductive learning includes the formalization of experts' knowledge and their transfer to a computer in the form of a knowledge base.

Supervised learning is the most common disease. Each case is a pair of "object, answer". It will be necessary to find the functional dependence of the responses on the characteristics of the objects and create an algorithm that will accept the description of the object as an input signal and respond at the output. The quality functional is usually defined as the average error of the responses generated by the algorithm for all sample objects.

1) The classification calculation differs in that the set of possible answers is finite. They are called class signs. A class is a set of all objects containing information about a sign.
2) The regression calculation differs in that the answer is a real number or a numeric vector.
3) The learning to rank calculation differs in that the answers are taken immediately in a set of objects and then sorted by the values of the answers. Can be reduced to classification or regression problems. It is often used in information retrieval and text analysis.
4) Forecasting accounting (forecasting) is distinguished by the fact that objects are segments of the time series in which it is necessary to make a forecast for the future. Regression or classification methods can be adapted to solve the prediction problem [2].

Uncontrolled learning. In this case, the answers will not be given and you will have to look for dependencies between objects.

The clustering problem involves grouping objects into clusters using data on their pairwise similarity. Quality functionality can be defined in various ways, for example, as the ratio of the average inter-cluster and intra-cluster distances.

Association rules learning report. Initial data is presented in the form of descriptions.

Learning through attachment. As an object, pairs of "situation, decision made" are considered, answers – functional quality values that characterize the correctness of the decisions made (the reaction of the environment). As in the forecasting report, the time factor plays an important role here. Examples of Applied Problems are: the formation of investment strategies, automatic control of technological processes, self-training of robots, etc. [3]

**Literature review.**

This review literature aims to fill this gap by conducting a comparative analysis of machine learning algorithms used for identifying extremist texts in the Kazakh language. By synthe-

sizing and critically evaluating existing studies, this review aims to provide insights into the strengths, limitations, and potential avenues for future research in this domain, ultimately contributing to the development of robust tools to counter the influence of extremism in the Kazakh online landscape.

Article [4] addresses the topic of the impact of online social media platforms on the process of radicalization of extremist groups. The study aims to identify radical social media texts and makes several contributions, including the creation of a new dataset for radicalization detection and analysis of variation in extremist group narratives. The proposed approach involves training the classifier using both religious and radical features to improve detection accuracy. The study also examines the use of aggressive and bad words in radical, neutral, and casual groups and finds that these words contribute to differentiation between radical and casual users. However, the neutral (anti-ISIS) group requires further investigation. The study provides insight into radical text detection and the impact of various factors on classifier performance.

This article [5] notes the importance of identifying and classifying extremist-related tweets as extremist groups use social media platforms to spread their ideologies and recruit people. The article proposes a system for analyzing content related to terrorism, specifically focused on classifying tweets into extremist and non-extremist categories. This approach uses deep learning based sentiment analysis techniques to develop a tweet classification system. The experimental results are described as encouraging, indicating the potential effectiveness of the proposed structure. The article addresses the urgent need for effective methods to detect extremist content on popular social networks such as Facebook and Twitter. It contributes to the field of extremism studies by providing a framework that can help monitor and combat the spread of extremist ideologies online. The work also offers future researchers a framework for developing and developing the field of identifying and classifying extremist content on social media platforms.

Article [6] describes the importance of online extremism research in monitoring the impact and spread of hate on social media platforms. He points out the limitations of existing research, which tends to be ideologically biased and lacks detailed classification beyond binary or tertiary classes. The research presented in the article aims to develop a balanced set of data on extremist texts, taking into account various ideologies, in particular, focused on extremist tweets. The dataset, named Merged ISIS/Jihadist-White Supremacist (MIWS), is evaluated using pretrained BERT and its variants (RoBERTa and DistilBERT), achieving a high f1 of 0.72. The study highlights the growing focus on natural language processing with deep learning in extremism detection research.

This article [7] highlights the importance of identifying extremism in social networks due to their influence on the spread of extremist ideologies. The existing literature on identifying extremism is limited to specific ideologies, subjective testing methods, and binary or tertiary classification. The review conducted a comprehensive review of datasets, classification methods, and validation methods using the PRISMA methodology, collecting 64 studies from various sources. The findings highlight the lack of publicly available, class-balanced, and unbiased datasets to effectively detect and classify social media extremism. In addition, the review shows a lack of methods to validate user datasets without human intervention and a bias towards ISIS ideology in current research. It concludes that deep learning-based automated extremism detection methods are superior to other methods and proposes research opportunities to develop an online extremism data collection and detection tool. Finally, the review proposes a conceptualization of an architecture for constructing a multi-ideological extremism text dataset with robust data validation methods for multi-class classification.

Article [8] describes the role of uncertainty in political, religious and social issues in inciting extremism among people, which is expressed in their moods on social networks. Although

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

English is the dominant language for social media exchanges, this study recognizes the importance of taking into account opinions expressed in other local languages in order to gain a better understanding of the data. The study focuses on sentimental analysis of multilingual textual data from social networks to determine the intensity of extremist sentiment. A multilingual dictionary with intensity weights is created and checked with 88% accuracy. Polynomial Naive Bayes and Linear Support Vector Classifiers are used for classification, and the accuracy of the Linear Support Vector Classifier reaches 82% on a multilingual dataset. The study contributes to the understanding of extremist sentiments expressed in several languages on social networks, gives an idea of the levels of extremism and demonstrates the effectiveness of the classification algorithms used.

The following article [9, 10] highlights the threat posed by online extremists on social media platforms and the limitations of suspending their accounts as they can easily create new ones. The study presents operational options to address this threat, focusing on the development of behavioral patterns for Twitter accounts associated with the "Islamic State of Iraq and Syria" (ISIS). These models are used to track existing extremist users by identifying pairs of accounts belonging to the same user. The study also presents a search model based on customizing the Polya urn to efficiently search for new blocked user accounts. The study contributes to this area by offering a new approach to tracking and identifying extremist online users that could also be applied to other areas.

Overall, these articles contribute to the field of detecting extremist texts using machine learning and deep learning techniques. They cover various aspects such as sentiment analysis, language models, social network analysis, and deep learning architectures. By reading these articles, you can get a complete understanding of the topic, as well as the various approaches and algorithms used in this area.

**Problem statement**

A set of data used in training $A = \{A_1, A_2, ..., A_{|A|}\}$ is described by a set of attributes, where $/ a /$ represents the number of attributes or the size of a set. In addition, the data set contains a special target c attribute called a class attribute. The Class C attribute has a discrete set of values, i.e. $\{c_1, c_2, ..., c_{|C|}\}$, where $|C|$ is the number of classes and $/ C / \geq 2$. The Class value is also called the class symbol.

A set of data for training is a relational table. Each data record is called an example, example, event, or vector in machine learning language.

The data collection consists mainly of examples. The purpose of the training is to create a classification/prediction function for matching the values of attributes in $A$ and classes in $C$, taking into account the data set $D$. This function is used to predict the values of classes of new data that will be encountered in the future. This function can also be called a classification model or a classifiactor [4].

The set of data used for training is called training data (or training set). Once the model has been studied or constructed based on learning data through a learning algorithm, it is evaluated with a set of test data (or invisible data) to assess the accuracy of the model. Test data are not used in the study of the classification model.

Examples in test data usually contain class labels. The data available for training and testing (with classes) is usually divided into two non-intersecting subsets: a training set (for training) and a test set (for testing).

In accordance with the purpose of the dissertation work, experiments were carried out using traditional machine learning methods. To identify extremist texts on web resources, traditional methods such as decision tree, multinomial Naive Bayes, random forest, linear regression and reference vector machine were used and their classification results were compared [5].

DOI: 10.37943/14DKRN4681



## Methods and results

1. Support vectors machine

The support vector machine is a set of algorithms necessary to solve the problems of classification and regression analysis. Based on the fact that an object in n - dimensional space belongs to one of two classes, the reference vector machine constructs a hyperspace with Dimension (N-1) so that all objects are in one of the two groups. In the process of using the support vector machine, the following two constraints must be satisfied:

$$wx_i - b \geq +1, \ if \ y_i = +1 \quad (1)$$

$$wx_i - b \leq -1, \ if \ y_i = -1 \quad (2)$$

In order for the hyperspace to move away from the nearest data of each class by an equal distance, the value of $||w||$ must be minimized, the minimization of which is equivalent to the minimization of $\frac{1}{2}\|w\|^2$. The optimization problem on the support vector machine is of the following nature [4]

$$min\frac{1}{2}\|w\|^2, y_i(x_i w - b) - 1 \geq 0, i = 1, \ldots, N \quad (3)$$

In the process of applying the given algorithm for the experiment, the following results were obtained (Table 1, Figure 1):

Table 1. Classification result using the reference vector machine method

| | |
|---|---|
| Accuracy | 0,73 |
| Precision | 0,99 |
| Recall | 0,36 |
| F1 | 0,53 |
| AUC-ROC | 0,68 |

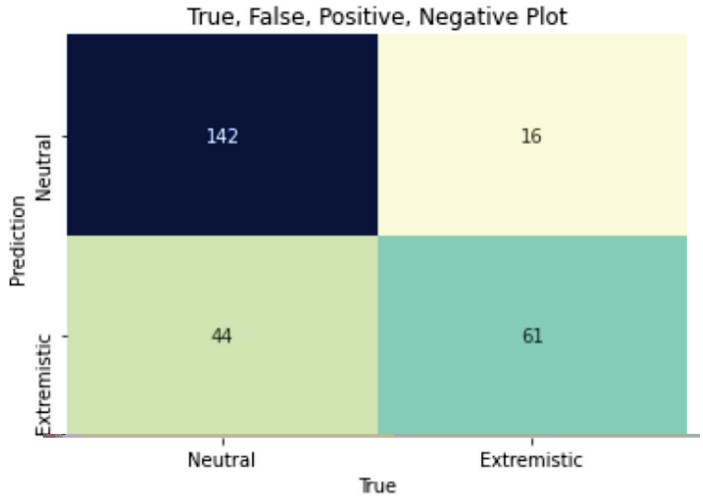


Figure 1. The result of classification using the support vector machine

Actual positive (TP) = 142; data of positive class 142 are correctly classified by the model.
Actual negative (TN) = 61; 61 negative class data is correct according to the classified model.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

False Positive (FP) = 16; data 16 of the negative class is incorrectly classified as belonging to the positive class by the model.

False negative (FN) = 44; data 44 of the positive class are erroneously classified as belonging to the negative class by the model.

The result of classification according to these characteristics in the form of a ROC curve is shown in the figure below (Fig. 2):

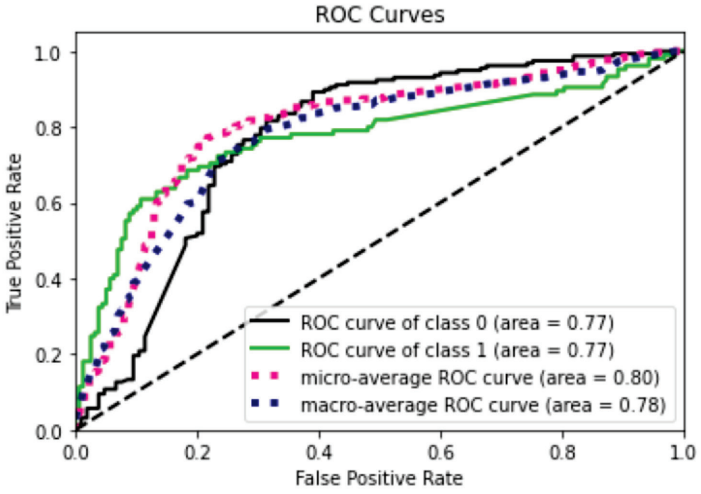


Figure 2. Classification result in the form of a ROC curve using a support vector machine

We see that the ROC value is 0.77, which indicates that the classifier has a good chance of distinguishing positive class values from negative class values. The AUC-ROC value is shown in the following Figure 3:

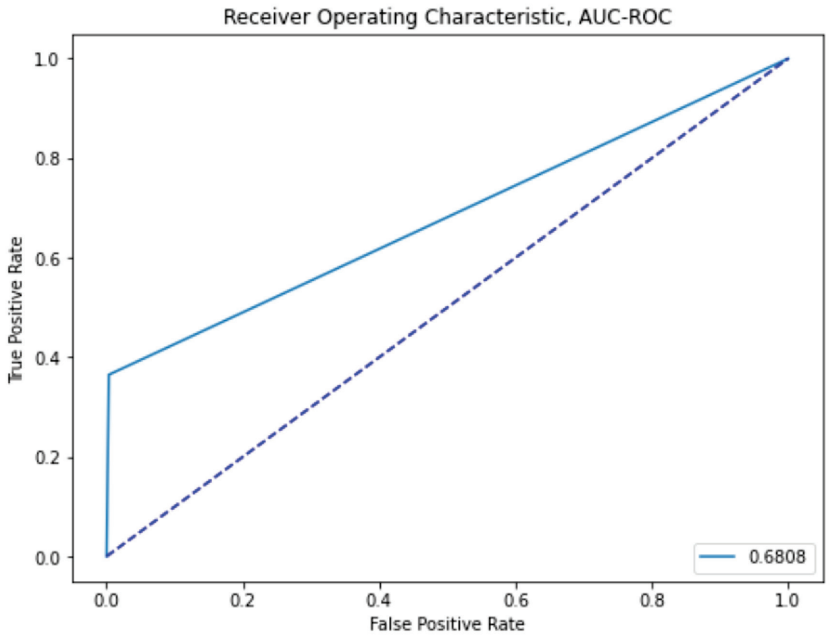


Figure 3. AUC-ROC classification value using a support vector machine

DOI: 10.37943/14DKRN4681



2. Decision tree

A decision tree (also known as a classification tree or regression tree) is a graphclic graph used for decision making. In each branching node of the graph, the j-th sign in the vector of signs is examined. If the value of the attribute is below the specified limit, the left branch is selected, otherwise the right branch is selected. When a leaf node is reached, a decision is made about the class in which this instance is appropriate.

In the process of data analysis, the decision tree can be used to describe, classify and generalize data sets as mathematical and computational methods and can be written as follows [6]:

$$(x, Y) = (x_1, x_2, x_3 \dots, x_{k,} Y) \tag{4}$$

The dependent variable Y is the target variable that needs to be analyzed and classified. Vector X consists of input variables such as $x_1$, $x_2$, $x_3$, etc.

In decision tree analysis, a visual and analytical decision support tool is used to calculate the expected values (or expected benefits) of competing alternatives.

The criterion of optimization in a given algorithm is an analogy with the average logarithmic reality:

$$\frac{1}{N}\sum_{i=1}^{N}[y_i ln f_{ID3}(x_i) + (1 - y_i)\ln(1 - f_{ID3}(x_i))] \tag{5}$$

where $f_{\text{ID3}}$ – decision tree.

Let S be a fixed data set. Initially, the decision tree will have a single node containing all the data: $S \stackrel{\text{def}}{=} \{(x_i, y_i)\}$

The constant model $f_{ID3}^{S}$ is defined as:

$$f_{ID3}^{S} = \frac{1}{|S|}\sum_{(x,y)\in S} y \tag{6}$$

The model $f_{ID3}^{S}(x)$ returns the same assumption for any X. Next, the signs j=1,,, D and the limits of t are searched, and the set S is divided into two subsets:

$$S_{-} \stackrel{\text{def}}{=} \{(x, y)|x, y) \in S, x^{(j)} < t\} \tag{7}$$

$$S_{+} \stackrel{\text{def}}{=} \{(x, y)|x, y) \in S, x^{(j)} \geq t\} \tag{8}$$

Two given subsets form new leaf nodes. The correctness of the model result is estimated using entropy. Entropy is the amount of uncertainty of a random variable.

Entropy reaches its maximum when the probabilities of all values of a random variable are equal, and entropy reaches its minimum when a random variable has only one value. The entropy of the set S is defined as [7]:

$$H(S) \stackrel{\text{def}}{=} -f_{ID3}^{S} ln f_{ID3}^{S} - (1 - f_{ID3}^{S})\ln(1 - f_{ID3}^{S}) \tag{9}$$

When using the decision tree method for the task of dividing texts into extremist and neutral classes, the following results were obtained (Table 2, Figure 4):

Table 2. The result of classification by the decision tree method

| | |
|---|---|
| Accuracy | 0,77 |
| Precision | 0,95 |
| Recall | 0,49 |
| F1 | 0,64 |
| AUC-ROC | 0,73 |

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

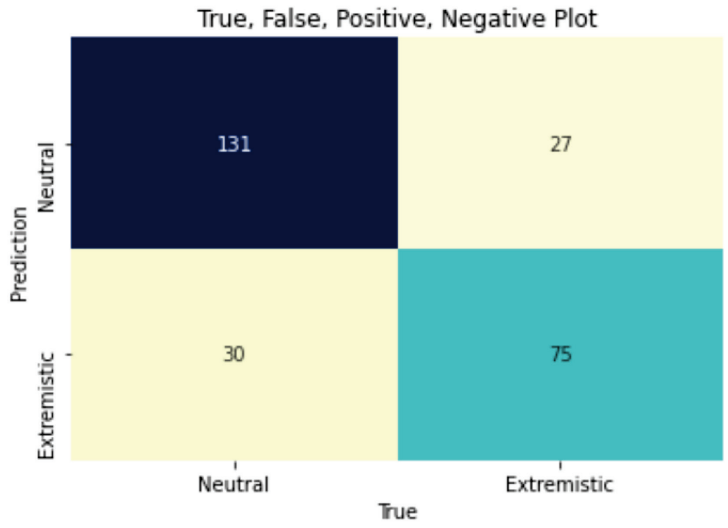


Figure 4. The result of classification by the decision tree method

Actual positive (TP) = 131; data of positive class 131 are correctly classified by the model.

Actual negative (TN) = 75; 75 the negative class data is correctly classified by the model.

False Positive (FP) = 27; data 27 of the negative class is incorrectly classified as belonging to the positive class by the model.

False negative (FN) = 30; Data 30 of the positive class are mistakenly classified as belonging to the negative class by the model.

We see that the ROC value is 0.77 (Fig. 5):

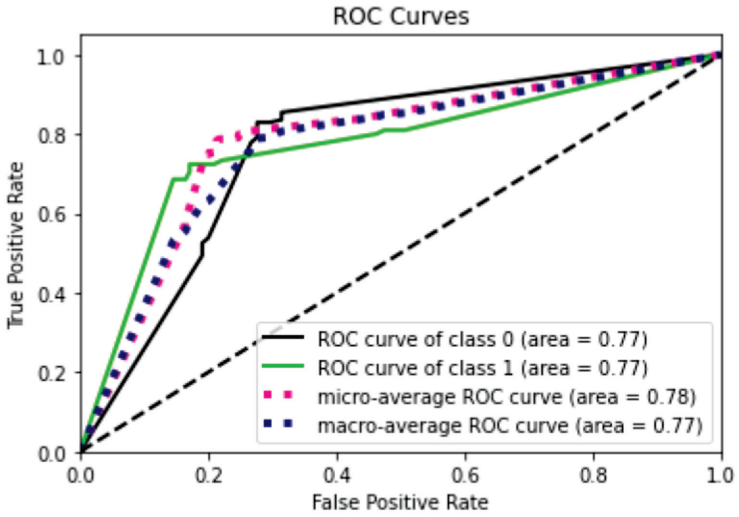


Figure 5. The result of classification by the decision tree method in the form of a ROC curve

The AUC-ROC value is shown in Figure 6 below.

DOI: 10.37943/14DKRN4681



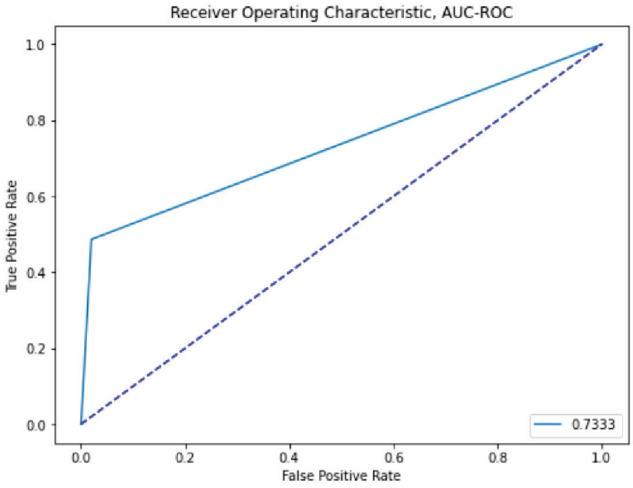


Figure 6. The value of AUC-ROC classification by the decision tree method

3. Random forest

A random forest is a machine learning algorithm proposed by Leo Breiman and Adele Cutler, while this algorithm uses an ensemble of a decision tree. The algorithm includes two main ideas: the Breiman bagging method and the method of random subspaces proposed by TinKam Ho. The main idea of the algorithm is to use a very large ensemble of decision trees, each of which gives very little classification accuracy, but affects a good result due to a large number.

In the case of using the random forest method, the following results were obtained (Table 3, Fig. 7):

Table 3. Classification results using the random forest method

| | |
|---|---|
| Accuracy | 0,73 |
| Precision | 1,0 |
| Recall | 0,36 |
| F1 | 0,52 |
| AUC-ROC | 0,68 |

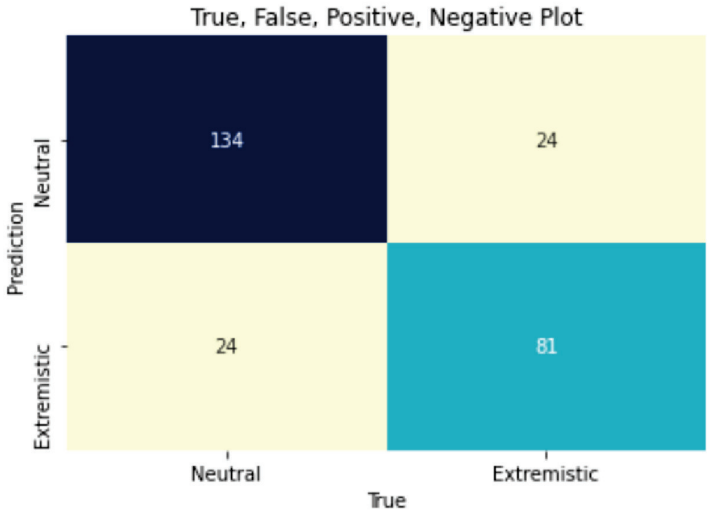


Figure 7. The result of classification by the random forest method

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

Actual positive (TP) = 134; data of positive class 134 are correctly classified by the model.

Actual negative (TN) = 81; 81 negative class data is correct according to the classified model.

False Positive (FP) = 24; data 24 of a negative class is incorrectly classified as belonging to a positive class by the model.

False negative (FN) = 24; data 24 of a positive class are erroneously classified as belonging to a negative class by the model.

The result of classification according to these characteristics in the form of a ROC curve is shown in Fig. 8:

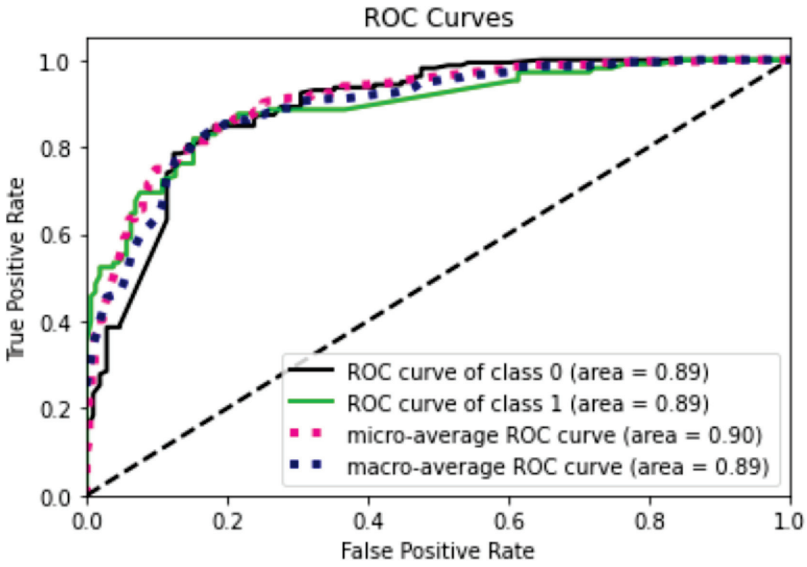


Figure 8. The result of classification by the random forest method in the form of a ROC curve

The AUC-ROC value is shown in Figure 9 below.

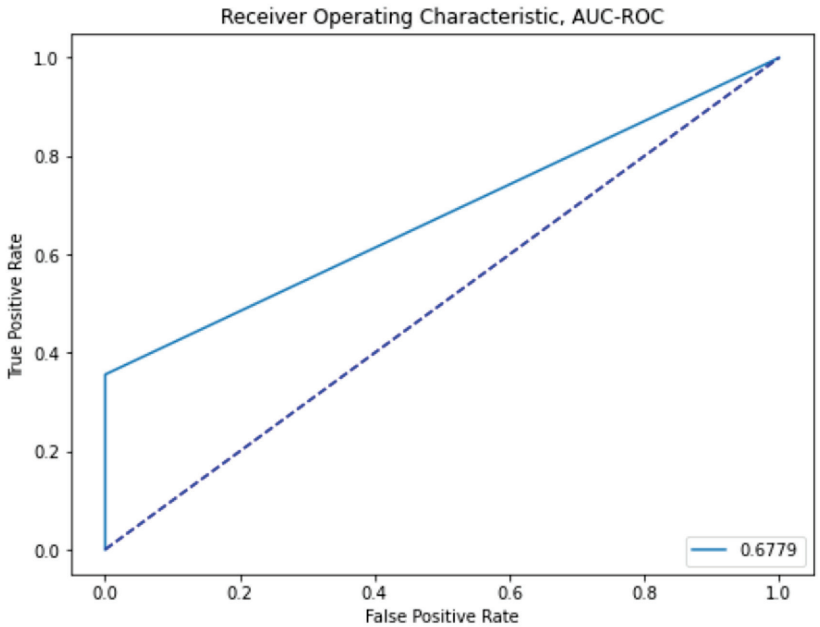


Figure 9. Value of AUC-ROC classification by random forest method

DOI: 10.37943/14DKRN4681



4. K-Nearest Neighbors algorithm

The nearest neighbor algorithm k (k-Nearest Neighbors, k-NN) is a nonparametric learning algorithm. The algorithm, presented as an exception from other machine learning algorithms, stores all training examples in memory. When a new instance of X that has not been encountered before appears, the k-NN algorithm finds the training data closest to X and returns the most frequent label. The proximity of two data is determined by the distance function. The Euclidean distance or a similar negative cosine distance is usually used. Cosine similarity is calculated by the following formula [7]:

$$s(x_i, x_k) \stackrel{\text{def}}{=} \cos\left(\angle(x_i, x_k)\right) = \frac{\sum_{j=1}^{D} x_i^{(j)} x_k^{(j)}}{\sqrt{\sum_{j=1}^{D} (x_i^{(j)})^2} \sqrt{\sum_{j=1}^{D} (x_k^{(j)})^2}} \quad (10)$$

If the angle between the vectors is 0 degrees, it means that they are oriented, and the cosine similarity is 1. If the vectors are orthogonal, the cosine similarity is 0. The cosine similarity of the inverse vectors is 1 [7].

In the course of applying this method to classify extremist and neutral texts, the following results were obtained (Table 4, Fig. 10):

Table 4. The result of classification by the nearest neighbor k method

| | |
|---|---|
| Accuracy | 0,78 |
| Precision | 0,96 |
| Recall | 0,49 |
| F1 | 0,66 |
| AUC-ROC | 0,5 |

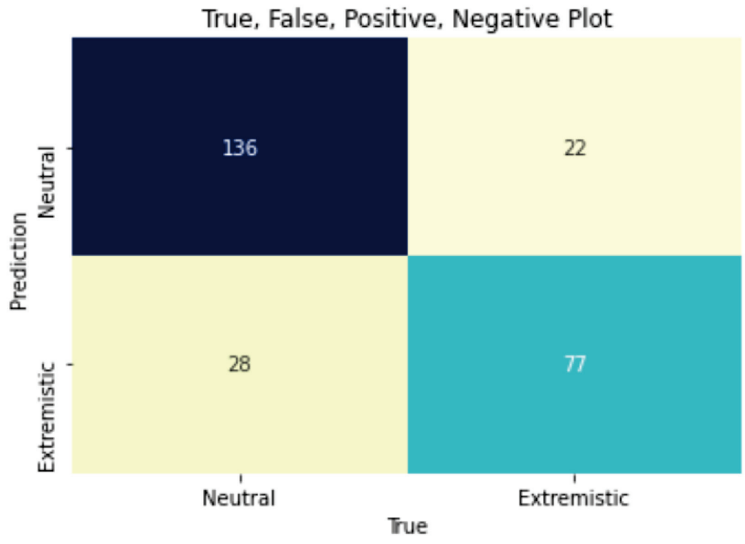


Figure 10. K matrix of inaccuracies of classification by the Near neighborhood method

Actual positive (TP) = 136; data of positive class 136 are correctly classified by the model.

Actual negative (TN) = 77; 77 the negative class data is correct according to the classified model.

False Positive (FP) = 22; data 22 of the negative class is incorrectly classified as belonging to the positive class by the model.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

False negative (FN) = 28; data 28 of the positive class are erroneously classified as belonging to the negative class by the model.

The result of classification according to these characteristics in the form of a ROC curve is shown in Fig. 11:

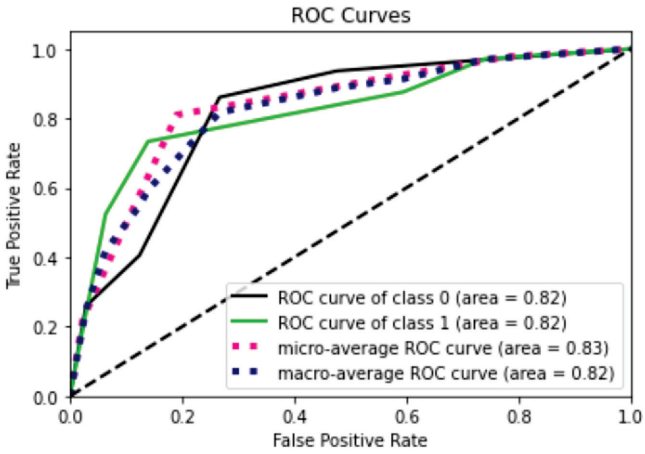


Figure 11. Classification results by the K-Nearest Neighbors method in the form of a ROC curve

The AUC-ROC value is shown in Fig.12.

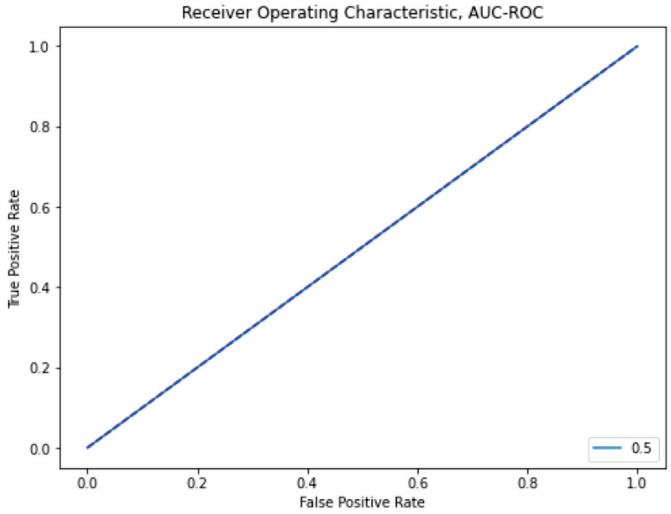


Figure 12. the value of AUC-ROC classification by the nearest neighbor knn method

5. Naive Bayes classifier

A naive Bayes classifier is a simple probabilistic classifier based on the application of Bayes' theorem along with strict independence guidelines. The advantage of this algorithm is the small amount of data required for training, evaluation and classification of parameters. The probabilistic model of the classifier is the following conditional model,

$$p(C \mid F_1, \ldots F_n) \tag{11}$$

where $C$-class, $F_1$, … $F_n$-variables. Using Bayes' theorem, we can obtain the following equation:

$$p(C|F_1, \ldots F_n) = \frac{p(C)p(F_1,\ldots F_n \mid C)}{p(F_1,\ldots F_n)} \tag{12}$$

DOI: 10.37943/14DKRN4681



In practice, the numerator of the fraction is interesting because the denominator does not depend on C and the values of $F_i$ are set, so the denominator of the fraction is a constant value [8].

The numerator of the fraction $p(C|F_1, \dots F_n)$ is equivalent to a combined sample of probabilities and can be written as follows, using the definitions of conditional probability:

$$\begin{aligned} &p(C|F_1, \dots F_n) = p(C)p(F_1, \dots, F_n \mid C) = p(C)p(F_1 \mid C)\, p(F_2, \dots, F_n \mid C, F_1) = \\ &p(C)p(F_1|C)p(F_2|C, F_1)p(F_3, \dots, F_n \mid C, F_1, F_2) = \\ &p(C)p(F_1|C)p(F_2|C, F_1) * \dots * p(F_n|C, F_1, F_2, F_3, \dots, F_{n-1}) \end{aligned} \quad (13)$$

Next, we can use «naive» conditional probability propositions: assume that each property $F_i$ is conditional independently of any other property $F_j$, $i \neq j$, it means $p(F_i|C, F_j) = p(F_i|C)$. Thus, the combined pattern can be expressed as follows:

$$p(C, F_1, \dots, F_n) = p(C)p(F_1|C)p(F_2|C)p(F_3|C) * \dots * p(F_n|C) = p(C) \prod_{i=1}^{n} p(F_i|C) \quad (14)$$

This means that the independence clauses, conditional distribution over a class C variable can be expressed as follows:

$$p(C|F_1, \dots, F_n) = \frac{1}{Z} p(C) \prod_{i=1}^{n} p(F_i|C) \quad (15)$$

where $Z = p(F_1, \dots, F_n)$ – this is a scale factor that depends only on $F_1, \dots F_n$, it is a constant value if the values of the variables are known. According to the probabilistic model, a classifier is created

$$classify(f_1, \dots, f_n) = \text{argmax}_c\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i|C = c) \quad (16)$$

When classifying texts by naive Bayes classifier into extremist and neutral categories, the following results were obtained (Table 5, Fig. 13):

Table 5. Naive Bayes classification result

| | |
|---|---|
| Accuracy | 0,86 |
| Precision | 0,80 |
| Recall | 0,89 |
| F1 | 0,84 |
| AUC-ROC | 0,86 |

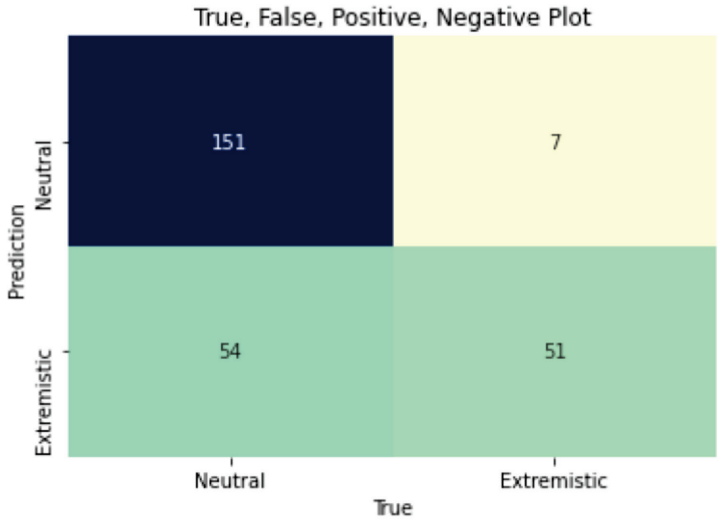


Figure 13. Matrix of classification inaccuracies by the naive Bayes method

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

Actual positive (TP) = 151; data of positive class 151 are correctly classified by the model.

Actual negative (TN) = 51; 51 negative class data is correct according to the classified model.

False Positive (FP) = 7; data 7 of the negative class is incorrectly classified as belonging to the positive class by the model.

False negative (FN) = 54; data 54 of the positive class are erroneously classified as belonging to the negative class according to the model.

In the course of classification by the naive Bayes algorithm, the ROC curve is obtained, as shown below (Fig. 14):

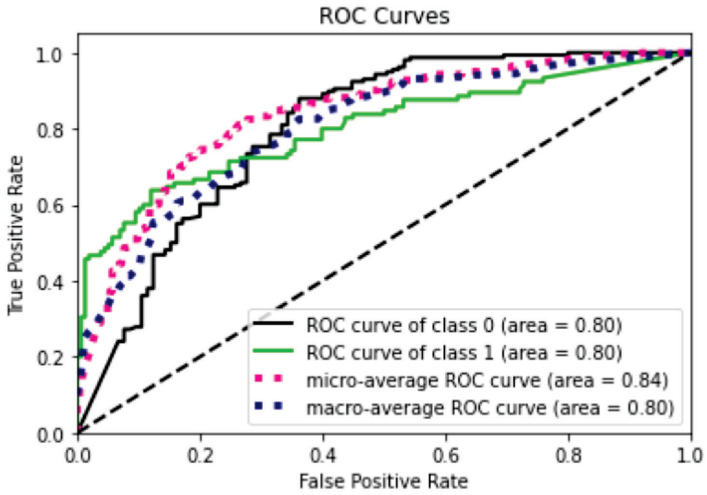


Figure 14. The result of classification by the naive Bayes method in the form of a ROC curve

The AUC-ROC value is shown in Fig.15.

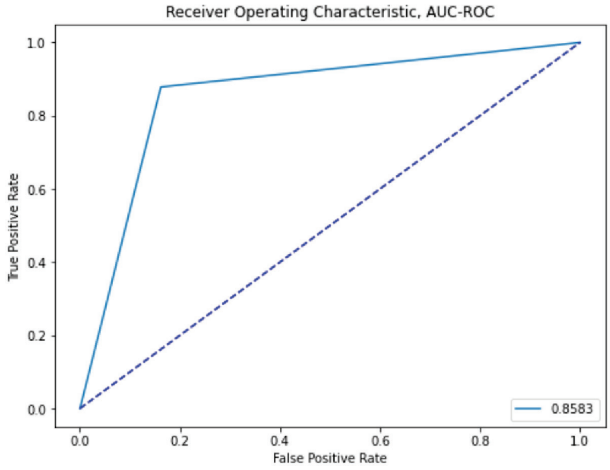


Figure 15. The value of AUC-ROC classification by naive Bayes method

6. Logistic regression

Logistic regression is a machine learning method that allows you to model $y_i$ in the form of a linear function that depends on $x_i$. A linear combination of signs like $wx_i+b$ is a function that extends from negative infinity to positive infinity, whereas $y_i$ can have only one of two values. If the return value of the Model for instance x is 0, a negative sign is assigned to it; otherwise,

DOI: 10.37943/14DKRN4681



a positive sign is assigned to this instance. One of the functions with such properties is the standard logistic function:

$$f(x) = \frac{1}{1+e^{-x}} \tag{17}$$

where e – is the base of the natural logarithm. The logistic regression model will look like this:

$$f_{w,b}(x) = \frac{1}{1+e^{-(wx+b)}} \tag{18}$$

If the values of w and b are optimized respectively, then the result of f(x) can be represented as the probability that $y_i$ will have a positive value.

The optimization criterion in the logistic regression model is called similarity with maximum reality (maximum likelihood). The analogy of the teacher's data with reality is maximized in accordance with the model [9]:

$$L_{w,b} \overset{\text{def}}{=} \prod_{i=1\ldots N} f_{w,b}(x_i)^{y_i} (1 - f_{w,b}(x_i))^{(1-y_i)} \tag{19}$$

The results of using the logistic regression method in experiments to identify extremist data in the Kazakh language on Web resources are presented in Table. 6 and Fig. 16:

Table 6. The result of classification by logistic regression

| | |
|---|---|
| Accuracy | 0,77 |
| Precision | 0,98 |
| Recall | 0,46 |
| F1 | 0,63 |
| AUC-ROC | 0,86 |

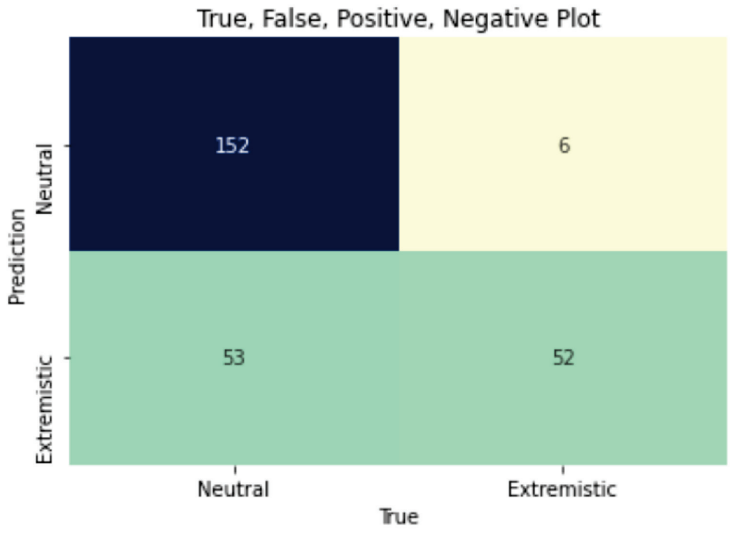


Figure 16. Matrix of classification inaccuracies by logistic regression method

Actual positive (TP) = 152; data of positive class 152 are correctly classified by the model.

Actual negative (TN) = 52; 52 negative class data is correct according to the classified model.

False Positive (FP) = 6; data 6 of the negative class is incorrectly classified as belonging to the positive class by the model.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

False negative (FN) = 53; data 53 of the positive class are erroneously classified as belonging to the negative class according to the model.

The ROC curve obtained as a result of the classification of texts by logistic regression is shown in Fig. 17:

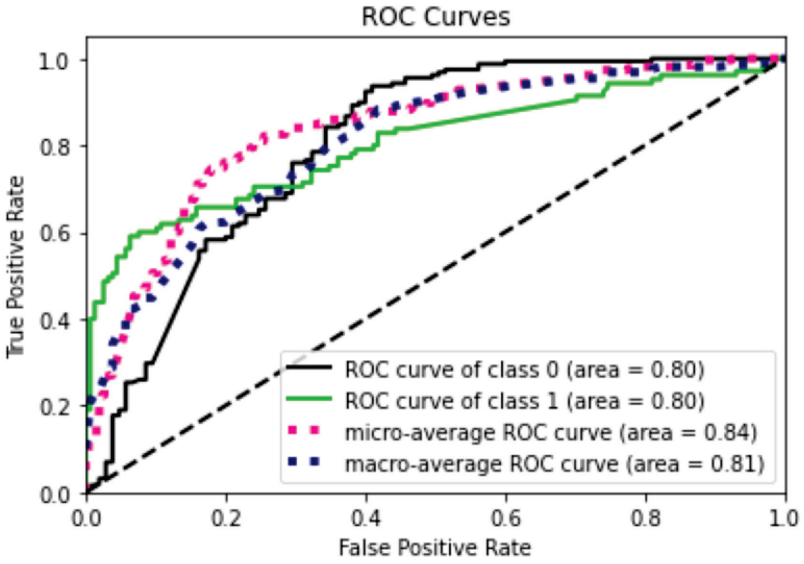


Figure 17. The result of classification by the logistic regression method in the form of a ROC curve

The AUC-ROC value is shown in Fig.18.

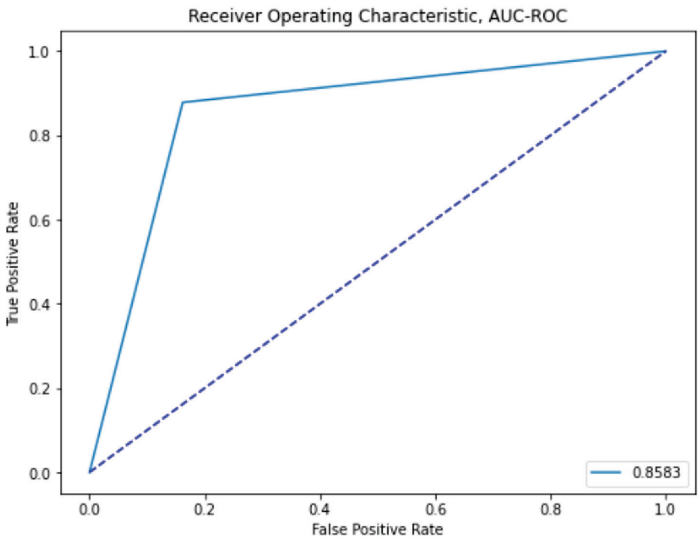


Figure 18. The value of AUC-ROC classification by logistic regression method

7. Gradient boosting

Gradient boosting is a machine learning method that creates a prediction model, usually consisting of an ensemble of weak prediction models, such as a decision tree, to solve classification and regression problems. The purpose of the algorithm, which is taught by any teacher, is to determine the loss function and minimize it. If the root-mean-square error is chosen as the loss function

DOI: 10.37943/14DKRN4681



$$Loss = \sum(y_i - y_i^p)^2 \quad (20)$$

where $y_i$ is the target value, $y_i^p$ is the assumption, $\mathrm{L}(y_i, y_i^p)$ is the loss function.

When updating values based on the learning rate using gradient shift, values are searched for at which MSE is minimal.

$$y_i^p = y_i^p + \alpha * \delta \sum \frac{(y_i - y_i^p)^2}{\delta y_i^p} \quad (21)$$

This expression has the form:

$$y_i^p = y_i^p - \alpha * 2 \sum(y_i - y_i^p) \quad (22)$$

where $\alpha$ is the learning rate, and $(y_i - y_i^p)$ is the residuals.

Thus, forecasts are updated until the sum of deviations tends to zero, and the predicted values are not close to the actual ones [9].

The result of applying this algorithm to the problem of classifying extremist and neutral texts is shown in Table.7 , Fig. 19.

Table 7. Classification result by gradient boosting method

| | |
|---|---|
| Accuracy | 0,75 |
| Precision | 0,99 |
| Recall | 0,41 |
| F1 | 0,58 |
| AUC-ROC | 0,71 |

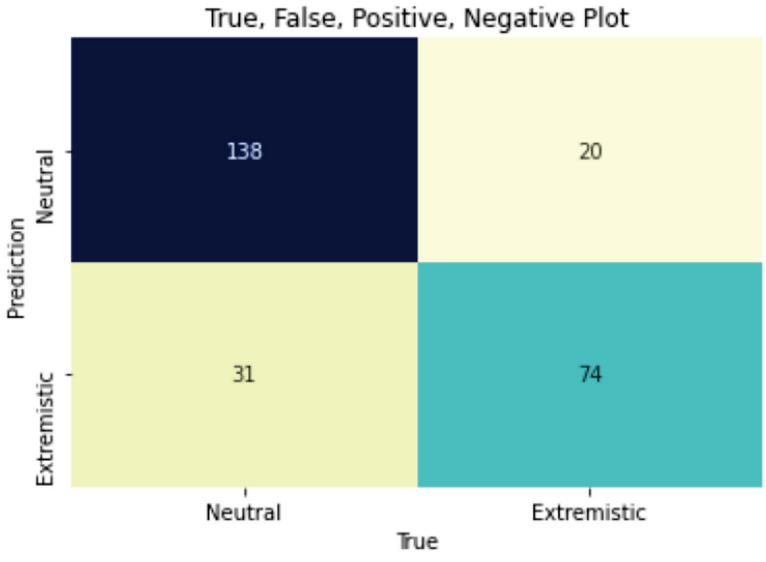


Figure 19. Matrix of classification inaccuracies by gradient boosting method

Actual positive (TP) = 138; data of positive class 138 are correctly classified by the model.

Actual negative (TN) = 74; 74 negative class data is correct according to the classified model.

False Positive (FP) = 20; data 20 of the negative class is incorrectly classified as belonging to the positive class by the model.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

False negative (FN) = 31; data of 31 positive classes are erroneously classified as belonging to a negative class by the model.

In the process of classification according to the gradient boosting algorithm, the ROC curve is obtained, as in Fig. 20:

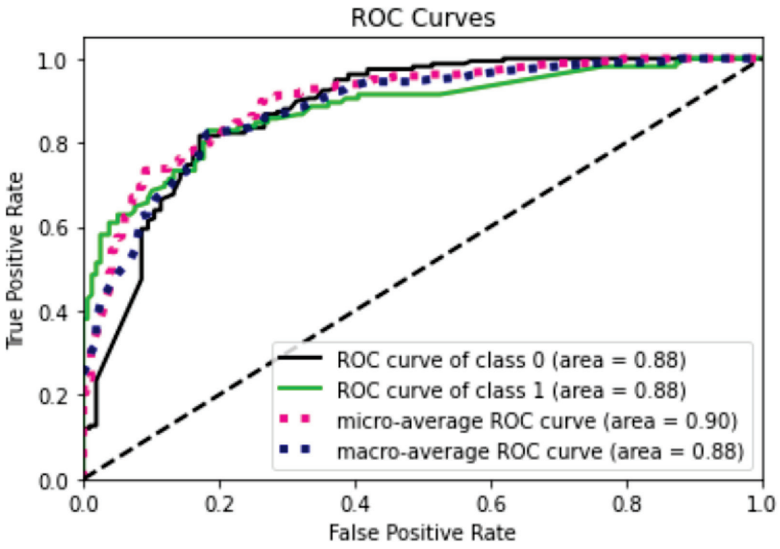


Figure 20. The result of classification by the gradient boosting method in the form of a ROC curve

The AUC-ROC value is shown in Figure 21 below.

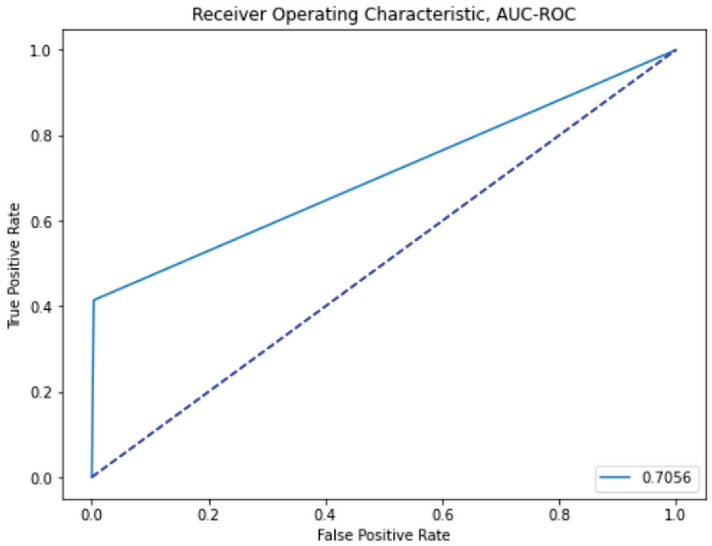


Figure 21. AUC-ROC value of classification by gradient busting method

**Discussion of results**

In conclusion, in this chapter, the problem of classifying the text into extremist and neutral categories was solved using several machine learning methods. The result of the experiment is presented in the following table 8, figure 22:

DOI: 10.37943/14DKRN4681



Table 8. The result of classifying text into extremist and neutral categories using machine learning methods

| Machine / deep learning method | Accuracy | F1-Score | AUC-ROC |
|---|---|---|---|
| Logistic regression | 0.77 | 0.63 | 0.86 |
| k-nearest neighbors | 0.78 | 0.66 | 0.5 |
| Decision Tree | 0.77 | 0.64 | 0.73 |
| Random Forest | 0.73 | 0.52 | 0.68 |
| Gradient Boosting | 0.75 | 0.58 | 0.71 |
| SVM | 0.73 | 0.53 | 0.68 |
| Naïve Bayes | 0.86 | 0.84 | 0.86 |
| Recommended model (TF-IDF_bigram_LSTM) | 0.9 | 0.88 | 0.89 |

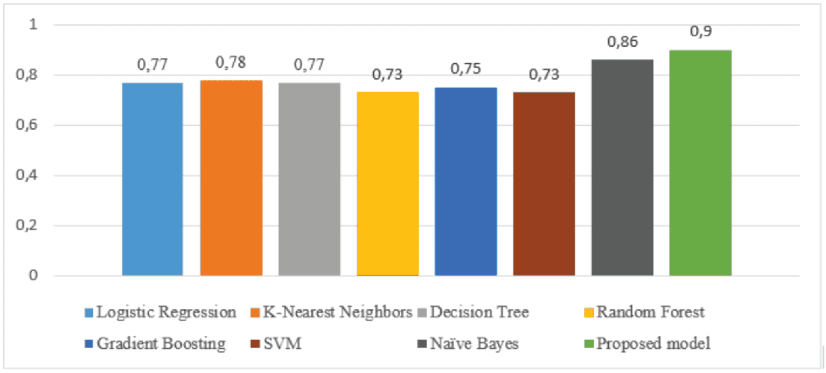


Figure 22. The result of classifying the text into extremist and neutral categories using machine learning methods

## Conclusion

In conclusion, this article sheds light on the significance of utilizing machine learning techniques for classifying text content and detecting destructive information within social networks. The research presented in this article offers valuable insights into the effectiveness of various machine learning models, such as support vector machines, naive base classifiers, and random tree methods, in identifying extremist texts. Additionally, the article provides an overview of existing research, methodologies, and software products in the field of analyzing extremist texts, emphasizing the importance of case-based learning, deductive learning models, and automated data collection and analysis. We conclude that the proposed model allows to identify extremist texts in the Kazakh language on web resources with high accuracy.

The findings underscore the ongoing relevance and importance of advancing this area of research. Future developments in this field are anticipated to involve the expansion of the corpus of documents studied, enabling a more comprehensive analysis of texts, including those presented in photo, audio, and video formats. Furthermore, the use of complex models for recognizing hidden extremist propaganda is expected to play a vital role in enhancing the accuracy and effectiveness of identifying extremist content.

Overall, this article contributes to the growing body of knowledge in the domain of text classification and highlights the potential for using machine learning techniques to combat the dissemination of destructive information and extremist ideologies within social networks. Further research and advancements in this field hold promising prospects for addressing the evolving challenges associated with detecting and countering online extremism.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 14, JUNE 2023

**Funding Statement**: This work was supported by the "Development of models and methods to identify youth extremism and ensure the safety of youth in the modern information space" funded by the grant of young scientists for scientific and (or) scientific and technical projects for 2023-2025 (Ministry of Science and Higher Education of the Republic of Kazakhstan). Grant No. IRN AP19576868. Supervisor of the project is Milana Bolatbek, email: bolatbek.milana@gmail.com.

## References

1. *Machine learning*. (2022, January 20). Retrieved from http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5
2. Arpinar, I.B., Kursuncu, U., & Achilov, D. (2016). Social media analytics to identify and counter Islamist extremism: Systematic detection, evaluation, and challenging of extremist narratives online. In *2016 International Conference on Collaboration Technologies and Systems* (pp. 611-612). IEEE. https://doi.org/10.1109/CTS.2016.0113
3. Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. Prentice Hall.
4. Ul Rehman, Z., Abbas, S., Khan, M. A., Mustafa, G., Fayyaz, H., Hanif, M., ... & Saeed, M. A. (2020). Understanding the language of ISIS: An empirical approach to detect radical content on Twitter using machine learning. *Computers, Materials & Continua, 66(2)*, 1075-1090. https://doi.org/10.32604/cmc.2020.012770
5. Ahmad, S., Asghar, M.Z., Alotaibi, F.M., & Awan, I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences, 9(24)*, 1-23. https://doi.org/10.1186/s13673-019-0185-6
6. Mayur, G., Swati, A., Ketan, K., & Ajith, A. (2022). Multi-ideology multi-class extremism classification using deep learning techniques. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.3205744
7. Mayur, G., Swati, A., Shraddha, P., & Ketan, K. (2021). Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools. *IEEE Access*. https://doi.org/10.1109/ACCESS.2021.3068313
8. Asif, M., Ishtiaq, A., Ahmad, H., Aljuaid, H., & Shah, J. (2020). Sentiment analysis of extremism in social media from textual information. *Telematics Informat, 48*, 101345. https://doi.org/10.1016/j.tele.2020.101345
9. Klausen, J., Marks, C.E., & Zaman, T. (2018). Finding extremists in online social networks. *European Journal of Operational Research, 66(4)*, 957-976. https://doi.org/10.1287/opre.2018.1719
10. Ul Rehman, Z., Abbas, S., Khan, M.A., Mustafa, G., Fayyaz, H., Hanif, M., & Saeed, M.A. (2020). Understanding the language of ISIS: An empirical approach to detect radical content on Twitter using machine learning. *Computers, Materials & Continua, 66*(2), 1075-1090. https://doi.org/10.32604/cmc.2020.012770
11. Burkov, A. (2020). *Machine learning without further ado*. Peter.
12. Swamy, M. N., Hanumanthappa, M., & Jyothi, N. M. (2014). Indian language text representation and categorization using supervised learning algorithm. In *2014 International Conference on Intelligent Computing Applications* (pp. 406-410). https://doi.org/10.1109/ICICA.2014.89
13. Mashechkin, I., Petrovskiy, M., Tsarev, D., & Chikunov, M. (2019). Machine learning methods for detecting and monitoring extremist information on the internet. *Programming and Computer Software, 45*, 99-115.
14. Ashraf, N., Rafiq, A., Butt, S., Shehzad, S.M.F., Sidorov, G., & Gelbukh, A. (2022). YouTube based religious hate speech and extremism detection dataset with machine learning baselines. *Journal of Intelligent and Fuzzy Systems, 42(5)*, 4769-4777.
15. Neurohive. (2022, February 18). Gradientnyj busting – prosto o slozhnom. [Gradient bousting - simple to complex]. Retrieved from https://neurohive.io/ru/osnovy-data-science/gradientyj-busting/