

INNOVATION SCIENCE AND TECHNOLOGY

ISSUE 10, PART I / OCTOBER 2026

 **Acceptance of papers**



**Acceptance of
papers**

Published monthly



Topics

economics,
technology, social
sciences



EDITOR-IN-CHIEF:

Mirzaliyev Sanjar Makhamatjon ugli

DEPUTY EDITOR-IN-CHIEF:

Makhmudov Nosir Makhmudovich
DSc., Prof., Academician

DEPUTY EDITOR-IN-CHIEF:

Ochilov Bobur Bakhtiyor ugli – Senior
lecturer at TSUI

THE SCIENTIFIC-POPULAR ELECTRONIC
JOURNAL **"INNOVATION SCIENCE AND
TECHNOLOGY"** HAS BEEN REGISTERED
UNDER THE NUMBER **C-5669633** BY THE
AGENCY FOR INFORMATION AND MASS
COMMUNICATIONS (AOKA) OF THE
REPUBLIC OF UZBEKISTAN, EFFECTIVE
FROM OCTOBER 9, 2024.

CONTACTS

Phone: **+998 50 737 87 88**

Website: <https://ist-journal.uz>

Email: innovationist2025@gmail.com

The scientific electronic journal "Innovation Science and Technology" has been included in the list of scientific publications recommended for the publication of main scientific results of dissertations for the award of PhD and DSc degrees in economics and technical sciences, in accordance with the Resolution No. 370 of the Presidium of the Higher Attestation Commission of the Republic of Uzbekistan, dated May 8, 2025.

Editorial board:



Sharipov Kongiratbay Avezimbetovich,
Doctor of Technical Sciences (DSc), Professor



Abdurakhmanova Gulnora Kalandarovna, Doctor of
Economic Sciences (DSc), Professor



Cham Tat Huei,
Doctor of Philosophy (PhD), Professor (Malaysia)



Muhammad Imran Sadiq
Doctor of Philosophy in Economics (PhD), Professor,
Malaysia



Ahmed Aziz Ismail
Doctor of Technical Sciences (DSc),
Professor (Egypt)



Lee Chin
Doctor of Philosophy in Economics (PhD), (Malaysia)



Asongu Simplicie
Doctor of Philosophy in Economics (PhD), Cameroon



Rui Dang
Doctor of Chemistry (DSc), Professor, China



Zahoor Ahmed
Doctor of Philosophy in Economics (PhD), Turkey



Shujaat Abbas
Doctor of Philosophy in Economics (PhD), Russia



Tina A Coffelt
Doctor of Philosophy in Educational Sciences (PhD),
USA



Abdikarimova Dinara Rustamxanovna
Doctor of Economic Sciences (DSc), Professor

Kurbonbekova Mohichehra Turobjonovna
Doctor of Economic Sciences (DSc), Professor

Alimardonov Ilkhom Muzrabshokovich
Doctor of Economic Sciences (DSc), Professor



Razakova Barno Sayfieva
Doctor of Philosophy in Economics (PhD)



Khasanov Sarvar Ulugbek ugli
Doctor of Philosophy in Economics (PhD)



Kholikova Rukhsora Sanjarovna
Associate Professor (PhD)



Xoliyorova Shoxista Qahramon qizi
PhD in Economics, Assistant Lecturer at the Department TSUE

Nasiba G'anjon qizi Sattorova
PhD, Senior Lecturer at the Department of Banking

CONTENTS

IMPLEMENTATION OF THE IMPROVED METHODOLOGY IN PRACTICE AND ASSESSMENT OF ITS ECONOMIC EFFICIENCY	8
Eshniyazova Yulduz Yuldashbayevna	
EXPERIMENTAL OPTIMIZATION OF DESIGN AND ENERGY PARAMETERS OF A GARLAND-TURBINE MOBILE MICRO HYDROPOWER PLANT FOR FREE-FLOWING WATER	13
Sultonov Sanjarbe, Zokirov Islomjon, Nurillaev Ahadullo, Abdullaev Shavkat	
METHODOLOGY FOR INCREASING THE EFFICIENCY OF CREDIT OPERATIONS IN COMMERCIAL BANKS	19
Djumaniyazov Ikram Karimbaevich	
COMPARATIVE ANALYSIS OF A COMPACT RESIDUAL CONVOLUTIONAL MODEL AND LINEAR REGRESSION FOR IMAGE SUPER-RESOLUTION	27
Husan Arziqulov Normurod o'g'li	
COMPARATIVE EVALUATION OF MACHINE LEARNING ALGORITHMS AND CLASS IMBALANCE HANDLING STRATEGIES FOR CREDIT CARD FRAUD DETECTION	33
Mahmudova Dilnoza Avaz qizi	
DIGITAL TRANSFORMATION OF FINANCIAL RESOURCE MANAGEMENT IN PUBLIC SOCIAL SECTOR INSTITUTIONS	39
Khasanov Tulkin Axmatovich	
WAYS TO ENSURE THE STABILITY OF THE CAPITAL BASE OF COMMERCIAL BANKS	45
Isakov Isobek Janabay's son	
CROSS-DAY EXPERIMENTAL EVIDENCE OF THERMAL-LOAD SHIFTING IN A SOLAR-DRIVEN HYDRONIC RADIANT FLOOR WITH MACROENCAPSULATED RT52 PHASE CHANGE MATERIAL	50
Jurabek Kamolov, Sokhib Temirov, Marjona Makhmudova	
STUDY OF THE TRACTION DYNAMICS OF AN AGGREGATE CONSISTING OF A FRUIT-HARVESTING MACHINE AND ITS TWO-AXLE CONTAINER-CARRYING TRAILER PLATFORM	59
Norchayev Davron, Sardor Nigmatjonov, Norchayev Jaloliddin, Norchayev Rustam	
PRIORITY DIRECTIONS FOR STRENGTHENING THE FINANCIAL STABILITY OF AGRICULTURAL ENTERPRISES	67
Kulmatov Zafar Jumanazarovich	
METHODOLOGICAL APPROACHES TO DEVELOPING UNIVERSITY STUDENTS' DIGITAL COMMUNICATIVE COMPETENCE BASED ON EMOTIONAL INTELLIGENCE	73
Malikakhon Kamiljanova	
SIMULTANEOUS REMOVAL CHARACTERISTICS OF OIL AND HEAVY METAL IONS BY BIOFLOCCULANTS BASED ON THE CELLULOSE-DMAEMA/PEPA COMPOSITE	80
Xayitboeva Dilshoda, Ibragimov Nasrilla, Jumaev Anvarbek	
ECONOMIC ANALYSIS AND DEVELOPMENT STRATEGY OF THE WOOD-POLYMER COMPOSITES MARKET	84
Abdumalikova Zarnigor, Ayubova Indira, Jumaev Anvarbek	
ENERGY UTILIZATION ANALYSIS OF AGRICULTURAL DRYING BASED ON INTEGRATED SOLAR THERMAL AND PHOTOVOLTAIC ENERGY SUPPLY	89
Ibragimov Umidjon, Mirzayarova Sevara, Khujayorova Aziza	
RECENT ADVANCES IN THE SYNTHESIS OF NANOSTRUCTURED MATERIALS FOR ZINC-ION BATTERIES	96
Diyorbek Absattorov, Dilnoza Erimbatova, Murodjon Samadiy	
ENTHALPY-MODEL SELECTION AND PRACTICAL IDENTIFIABILITY OF A MACROENCAPSULATED RT52 RADIANT FLOOR UNDER UNCERTAIN BRANCH FLOW	101
Jurabek Kamolov, Sokhib Temirov, Marjona Makhmudova	
ENTHALPY-MODEL SELECTION AND PRACTICAL IDENTIFIABILITY OF A MACROENCAPSULATED RT52 RADIANT FLOOR UNDER UNCERTAIN BRANCH FLOW	101
Jurabek Kamolov, Sokhib Temirov, Marjona Makhmudova	

THE IMPORTANCE OF THERMAL ENERGY STORAGE SYSTEMS IN ENHANCING THE EFFICIENCY OF SOLAR DRYERS	110
B. A. Hikmatov, M. S. Mirzaev,	
METHODOLOGY AND PRINCIPLES FOR THE APPLICATION OF ARTIFICIAL INTELLIGENCE IN ADAPTIVE ECONOMIC MODELS OF HIGHER EDUCATION.....	117
O'lmasxo'jayev Zoirxo'ja Abbos o'g'li	
FORECASTING ECONOMIC EFFECTS OF GREEN INNOVATION IN MINING THROUGH COST REDUCTION, ENERGY SAVINGS AND LONG-TERM COMPETITIVENESS.....	122
Kurbanova Mehriniso Nematjanovna	
AN INTELLECTUAL HYBRID MODEL AND FEATURE SPACE FOR NAMED ENTITY RECOGNITION IN KARAKALPAK TEXTS	127
Kalmuratov Bekbosin Kushkinbaevich, Uzakbaev Jangeldi Sabit-uli, Rzambetova Suliwxan Baxtiyarovna	
COMPARATIVE ANALYSIS OF EXISTING METHODOLOGIES FOR ASSESSING THE EFFICIENCY OF LENDING PRACTICE	133
Turdiyev Eralijon Muxtarovich	
STRATEGIC MANAGEMENT OF SERVICE QUALITY: ADAPTING THE GAP MODEL TO SERVICE ENTERPRISES.....	140
Tuxliyev Baxtiyor Maxmaziyyevich	
ORGANISATIONAL AND LEGAL FOUNDATIONS FOR THE DEVELOPMENT OF ISLAMIC FINANCE IN UZBEKISTAN	145
Khodjaev Aziz Shovkatovich	
REGIONAL TRANSMISSION OF MACROECONOMIC SHOCKS AND THE DYNAMICS OF ECONOMIC CONVERGENCE	153
Elyor Nazarov	
PRIORITY DIRECTIONS FOR INCREASING SECTORAL DIVERSIFICATION IN THE ECONOMY OF UZBEKISTAN'S REGIONS	159
Mamarsulov Nozimjon Mamajonovich	
THE ROLE OF STATE STRATEGIC RESERVES IN ENSURING FOOD SECURITY: POLICY EVOLUTION AND REASSESSMENT OF THE MILITARY SUPPLY SYSTEM IN THE CASE OF UZBEKISTAN.....	165
Bazarov Nazirjon Sobirovich	
LOCAL ADAPTATION OF MRA-FW FOR CARDIOVASCULAR RISK ASSESSMENT IN ARAL SEA REGION INPATIENTS.....	170
Uteuliev Niyetbay, Saytov Aralbek, Saytov Kamiljan	
GREEN FINANCE IN UZBEKISTAN: FROM SOVEREIGN BONDS TO FOREIGN DIRECT INVESTMENT IN RENEWABLE ENERGY (2021–2026).....	176
Temirova Nozima Shuhrat qizi	
IMPROVING THE ACCOUNTING OF REVENUE, EXPENSES AND FINANCIAL RESULTS IN PRIVATE HIGHER EDUCATION INSTITUTIONS	182
Saidova Kamolat Farhodovna	
THE IMPACT OF DIGITAL TRANSFORMATION ON SERVICE DELIVERY EFFICIENCY IN LOCAL HIGHER EDUCATION INSTITUTIONS.....	187
Masaidova Marguba Muratovna	
INVESTIGATION OF METAL REFINING DURING ELECTRIC ARC WELDING.....	195
Ermatov Ziyadulla, Rakhmatullaeva Nozima, Galperin Leonid	
STATE AND INSTITUTIONAL SUPPORT FOR THE EXPORT ACTIVITIES OF CONSTRUCTION ENTERPRISES.....	200
Mirzayeva Khulkar Abduladjon kizi	
ECONOMIC DIPLOMACY AS THE EXTERNAL TIER OF INVESTMENT GOVERNANCE: INSTITUTIONAL DESIGN AND IMPLEMENTATION GAPS IN UZBEKISTAN.....	207
Imamov Javokhir	
METHODS FOR IMPROVING THE EFFICIENCY OF PRODUCTION RESOURCE UTILIZATION IN AGRICULTURE	213
Nurxonov Nurkomil Qahramon ugli	

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR AUTOMATIC TOPIC CLASSIFICATION OF KARAKALPAK NEWS TEXTS.....	218
Asenbayev Nurbek Adilbayevich	

UDC: 004.85

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR AUTOMATIC TOPIC CLASSIFICATION OF KARAKALPAK NEWS TEXTS

Asenbayev Nurbek Adilbayevich

Nukus State Technical University, Assistant Lecturer

E-mail: asenbaevnurbek1993@gmail.com

Abstract: This paper investigates automatic topic classification of Karakalpak news texts. The experiment was conducted using the KKNEWS dataset, which initially contained 13,828 records. After analyzing the distribution of topic categories, seven categories containing at least 200 texts each were selected. Following script unification and duplicate removal, the final experimental corpus consisted of 13,442 Karakalpak news texts written in Latin script. TF-IDF was used for numerical text representation, while Logistic Regression, Linear Support Vector Machine, and Multinomial Naive Bayes were evaluated as classifiers. Linear SVM achieved the best performance with an accuracy of 0.8702 and a macro-F1 score of 0.7485. The results indicate that the combination of TF-IDF and Linear SVM can serve as an effective baseline approach for automatic topic classification of Karakalpak news texts.

Keywords: Karakalpak language, natural language processing, text classification, machine learning, TF-IDF, Linear SVM, Logistic Regression, Naive Bayes.

Аннотация: В статье рассматривается задача автоматической тематической классификации новостных текстов на каракалпакском языке. Для эксперимента использован набор данных KKNEWS, первоначально включавший 13 828 записей. После анализа распределения тематических категорий были отобраны семь категорий, содержащих не менее 200 текстов каждая. После унификации графики и удаления дублирующихся записей итоговый экспериментальный корпус составил 13 442 новостных текста на каракалпакском языке в латинской графике. Для числового представления текстов использован метод TF-IDF, а для классификации применены алгоритмы Logistic Regression, Linear Support Vector Machine и Multinomial Naive Bayes. Наилучший результат показал Linear SVM: accuracy 0,8702 и macro-F1 0,7485. Полученные результаты показывают, что комбинация TF-IDF и Linear SVM может служить эффективным базовым подходом для тематической классификации новостных текстов на каракалпакском языке.

Ключевые слова: каракалпакский язык, обработка естественного языка, классификация текста, машинное обучение, TF-IDF, Linear SVM, Logistic Regression, Naive Bayes.

INTRODUCTION

The rapid growth in the volume of digital information sources has intensified the need for the automated processing of large amounts of textual data. Every day, a substantial volume of new textual content is generated on news websites, in electronic libraries, on social media platforms, and across official information resources. Manually categorizing such a continuous flow of information by topic requires considerable time and effort. Therefore, automatic classification systems based on Natural Language Processing (NLP) methods are becoming increasingly important in practical applications.

In automatic text classification, a document is assigned to one of the predefined classes based on its content. This approach is widely applied to tasks such as categorizing news into politics, economics, sports, culture, and other topics, sorting electronic documents, improving information retrieval systems, and enabling users to quickly identify relevant content. In low-resource languages, however, solving such tasks is additionally complicated by the limited availability of annotated corpora, lexical resources, and pre-trained models.

The Karakalpak language is also among the languages with relatively limited digital resources. Madatov, Bekchanov, and Vičić [1], in developing a stop-word set for the Karakalpak language, constructed a corpus based on school textbooks and applied TF-IDF-based unigram, bigram, and collocation methods. Subsequent studies proposed the use of multilingual transformer models, such as mBERT, for classifying Karakalpak texts

according to subject domains [2]. However, before evaluating complex neural models, it is important to establish reliable baseline results using classical algorithms that are computationally inexpensive, easily reproducible, and highly interpretable.

The aim of this study is to comparatively evaluate TF-IDF-based Logistic Regression, Linear Support Vector Machine (Linear SVM), and Multinomial Naive Bayes algorithms for the automatic topic classification of Karakalpak-language news texts under a unified experimental setting. Considering the class imbalance in the dataset, the study evaluates not only overall accuracy but also macro-precision, macro-recall, macro-F1, and weighted-F1 metrics.

LITERATURE REVIEW

In classical approaches to text classification, textual data are first transformed into numerical feature vectors, after which a classification model is trained on the basis of these features. One of the most widely used methods in information retrieval and text classification systems is TF-IDF (Term Frequency–Inverse Document Frequency), which jointly considers the frequency of a term within a document and its distribution across the entire corpus [3]. As a result, frequently occurring units that provide relatively little information for distinguishing topics receive lower weights, whereas topic-specific terms are assigned higher weights.

The Support Vector Machine (SVM) algorithm is well known for its effectiveness in handling high-dimensional and sparse textual features. Joachims [4] demonstrated that the high dimensionality of text data and the presence of a large number of relevant features create favorable conditions for the application of SVM. Logistic Regression is also a widely used linear model for multiclass text classification; it can be trained efficiently on TF-IDF vectors and is capable of providing strong baseline performance. Multinomial Naive Bayes, in turn, is based on a probabilistic approach and is distinguished by its computational efficiency, particularly when working with sparse features such as word-frequency representations or TF-IDF vectors.

In recent years, the development of transformer-based models such as BERT and XLM-R has significantly improved the performance of text classification systems. BERT is based on the principle of pre-training bidirectional contextual representations and subsequently fine-tuning them for a specific downstream task [5]. XLM-R, meanwhile, is trained on large-scale multilingual corpora and enhances the ability to transfer knowledge to low-resource languages [6]. For the Karakalpak language, Abdurakhmonova, Sayfullaeva, and Shirinova [2] combined rule-based preprocessing with mBERT and investigated the classification of a dataset consisting of 20,000 sentences into four subject domains. This study highlights the relevance and growing importance of text classification research for the Karakalpak language.

At the same time, transformer-based models require substantial computational resources and specialized fine-tuning procedures. Classical algorithms, by contrast, remain practically valuable because of their computational efficiency, ease of experimental reproducibility, and interpretability of results. Therefore, in this study, three TF-IDF-based classical algorithms were compared for Karakalpak-language news texts using the same corpus and an identical train/test split. Computationally efficient approaches to text classification have also been extensively investigated in the literature [8].

RESEARCH METHODOLOGY

The study used a Karakalpak-language news dataset compiled from the KKNEWS source. The initial dataset consisted of 13,828 records, with each record containing the article text, title, thematic category, language, script type, text length, and technical metadata related to the preprocessing procedure. A total of 14 categories were identified in the dataset; however, their distribution was highly imbalanced in terms of the number of samples.

Seven categories containing at least 200 samples were included in the experiment: Society, Politics, World, Sport, Economy, Documents, and Culture. After this filtering procedure, 13,461 records remained. An examination of script type showed that 13,445 texts were written in the Latin script, whereas 16 texts were written in the Cyrillic script. To ensure that the experiment was conducted using a uniform writing system, the 16 Cyrillic-script texts were excluded. In addition, three duplicate records were identified using dedup_hash, and the redundant copies were removed. As a result, the final corpus consisted of 13,442 texts.

The average text length was 258.2 words, while the median length was 172 words. The shortest text contained 30 words, whereas the longest consisted of 8,942 words. To construct the model input, the article title and the main body text were concatenated. This made it possible to utilize important topic-indicative terms appearing in news headlines as additional classification features.

The data were divided into training and test sets using a stratified split while preserving the original class proportions, with 80% of the data allocated to training and 20% to testing. Consequently, 10,753 texts were

used for model training and 2,689 texts for evaluation. In TF-IDF vectorization, both unigrams and bigrams were employed. Terms occurring in at least two documents were retained, a $\text{max_df}=0.95$ threshold was applied to highly frequent terms, and the total number of features was limited to 50,000. The resulting training matrix had dimensions of $10,753 \times 50,000$, while the test matrix had dimensions of $2,689 \times 50,000$.

Logistic Regression, Linear SVM, and Multinomial Naive Bayes algorithms were selected for classification. To reduce the effect of class imbalance, the `class_weight="balanced"` parameter was applied to the Logistic Regression and Linear SVM models. Model performance was evaluated using accuracy, macro-precision, macro-recall, macro-F1, and weighted-F1 scores. Macro-F1 was selected as the primary comparative metric because it assigns equal importance to the performance of each class. The experimental implementation of the models was carried out using the scikit-learn library [7].

Table 1. Distribution of Selected Categories in the Latin Script¹

No.	Category	Number of Texts
1	Society	8749
2	Politics	2339
3	World	747
4	Sport	709
5	Economy	463
6	Documents	225
7	Culture	213
	Total	13445

Note: After removing three duplicate records identified based on the `dedup_hash` field from the 13445 Latin-script records, the final experimental corpus consisted of 13442 texts.

ANALYSIS AND RESULTS

The three algorithms were evaluated using the same TF-IDF features and exactly the same train/test split. The results showed that the Linear SVM model outperformed the other models in terms of overall classification performance. Linear SVM achieved an accuracy of 0.8702, a macro-precision of 0.7843, a macro-recall of 0.7213, a macro-F1 score of 0.7485, and a weighted-F1 score of 0.8673. Logistic Regression achieved an accuracy of 0.8115 and a macro-F1 score of 0.7030. Although Multinomial Naive Bayes achieved an overall accuracy of 0.7772, its macro-F1 score was only 0.3570.

Table 2. Comparative Results of Machine Learning Algorithms²

Model	Accuracy	Macro Precision	Macro Recall	Macro F1	Weighted F1
Linear SVM	0,8702	0,7843	0,7213	0,7485	0,8673
Logistic Regression	0,8115	0,6468	0,8025	0,7030	0,8231
Multinomial Naive Bayes	0,7772	0,5171	0,3330	0,3570	0,7237

The accuracy advantage of Linear SVM over Logistic Regression was 5.88 percentage points, while its advantage over Multinomial Naive Bayes was 9.30 percentage points. The difference was even more pronounced in terms of macro-F1: Linear SVM achieved 0.7485, Logistic Regression 0.7030, and Naive Bayes 0.3570. Logistic Regression produced the highest macro-recall score of 0.8025; however, its lower macro-precision indicates a higher number of false-positive classifications.

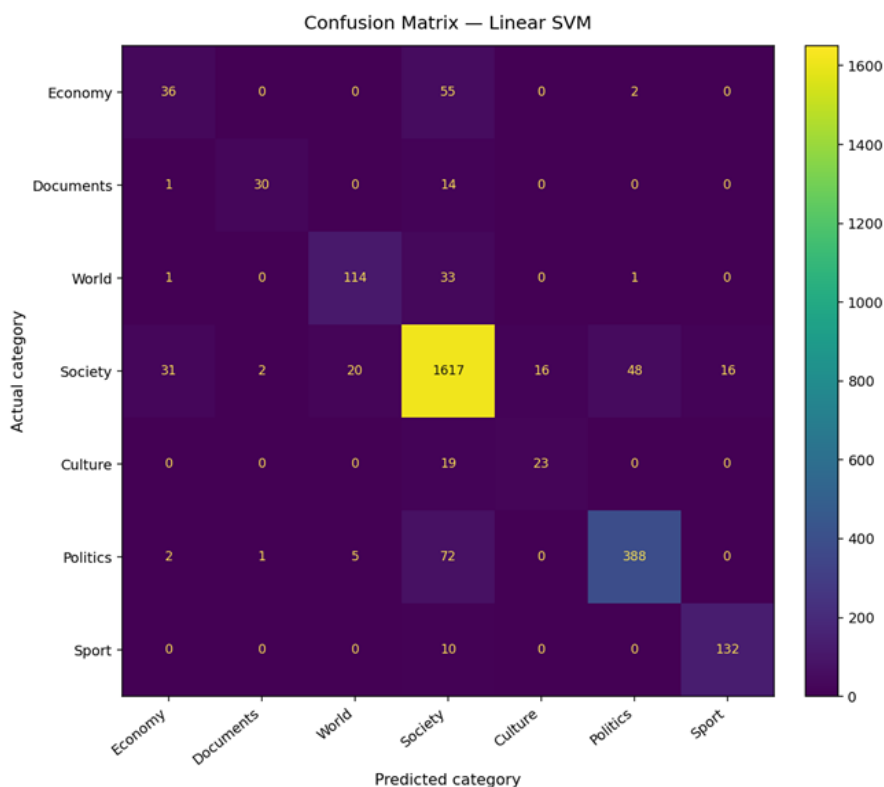
The category-level performance of the Linear SVM model revealed substantial variation across classes. The highest F1-score was achieved for the Sport category at 0.9103, followed by Society at 0.9059, Politics at 0.8556, and World at 0.7917. For the Documents category, precision reached 0.9091, whereas recall remained at 0.6667. The lowest performance was observed for the Economy category, with an F1-score of 0.4390.

¹ Prepared by the author based on the research data.

² Author's elaboration.

Table 3. Category-Level Results of the Linear SVM Model³

Category	Precision	Recall	F1-score	Support
Economy	0,5070	0,3871	0,4390	93
Documents	0,9091	0,6667	0,7692	45
World	0,8201	0,7651	0,7917	149
Society	0,8885	0,9240	0,9059	1750
Culture	0,5897	0,5476	0,5679	42
Politics	0,8838	0,8291	0,8556	468
Sport	0,8919	0,9296	0,9103	142

Figure 1. Confusion Matrix for the Linear SVM Model⁴

The confusion matrix showed that the most frequent classification errors were associated with the Society category. Of the 93 test samples in the Economy class, 36 were classified correctly, while 55 were incorrectly classified as Society. In addition, 72 texts from Politics, 33 from World, and 19 from Culture were also incorrectly assigned to the Society category. One of the main reasons for this is the disproportionately large share of the Society category in the corpus and the partial overlap of its thematic boundaries with other socio-political categories.

The high performance of the Sport category can be explained by the more distinctive terminology associated with this topic. Terms such as athlete, tournament, record, championship, match, and competition occur less frequently in other categories and therefore form strong discriminative features in the TF-IDF space. In the Economy and Culture categories, by contrast, the topics overlap more substantially with Society and Politics in lexical and semantic terms.

CONCLUSIONS AND RECOMMENDATIONS

The obtained results are consistent with the classical findings regarding the effectiveness of the Linear SVM

³ Prepared by the author based on the research data.

⁴ Author's elaboration.

algorithm when applied to high-dimensional and sparse textual features [4]. In the present experiment, Linear SVM also produced a better-balanced performance than the other two classical models within a TF-IDF feature space consisting of 50,000 features. In particular, the macro-F1 score provided a more accurate reflection of the model's actual performance under conditions of class imbalance.

The results of Multinomial Naive Bayes clearly demonstrate that accuracy alone is not a sufficient evaluation metric. Although the model achieved an accuracy of 0.7772, the recall values for the Economy, Documents, and Culture classes were equal to zero. The dominance of the large Society class kept the overall accuracy relatively high, while the macro-F1 score decreased to 0.3570. Therefore, in imbalanced multiclass datasets, macro-F1 should be used as one of the primary evaluation metrics.

Since the Karakalpak language has an agglutinative structure, numerous affixal forms derived from the same root may be represented as separate TF-IDF features. No specialized stemming or lemmatization procedures were applied in this study. Nevertheless, the combination of unigrams and bigrams produced a strong baseline result. At the next stage, the application of morphological normalization tailored to the Karakalpak language, subword tokenization, and comparison with transformer-based models may further improve classification performance.

The mBERT-based study by Abdurakhmonova et al. also reported that the subject domain had a significant effect on classification performance [2]. Similarly, in our experiment, the semantic overlap of the Economy and Culture categories with other topics reduced classification performance, whereas higher F1-scores were observed for classes with more clearly defined terminological boundaries, such as Sport and Politics. These findings provide a basis for reconsidering the thematic category structure and evaluating hierarchical classification approaches in future research.

In this study, three classical TF-IDF-based machine learning algorithms were comparatively evaluated for the automatic topic classification of Karakalpak-language news texts. Logistic Regression, Linear SVM, and Multinomial Naive Bayes models were tested under identical experimental conditions using 13,442 deduplicated Latin-script texts selected from the KKNEWS dataset.

According to the experimental results, Linear SVM achieved the best performance, with an accuracy of 87.02% and a macro-F1 score of 0.7485. For Logistic Regression, these values were 81.15% and 0.7030, respectively, while Multinomial Naive Bayes achieved an accuracy of 77.72% and a macro-F1 score of 0.3570. Linear SVM demonstrated particularly strong performance in the Sport, Society, and Politics categories, whereas classification was more challenging for the Economy and Culture categories due to the greater semantic overlap between classes.

The findings indicate that the combination of TF-IDF and Linear SVM provides an effective and computationally efficient baseline approach for the automatic topic classification of Karakalpak-language news texts. Future research may focus on class balancing, morphological normalization, subword features, the application of transformer-based models such as mBERT and XLM-R, and systematic comparisons between transformer-based and classical approaches.

REFERENCES

1. Madatov K., Bekchanov Sh., Vičič J. Dataset of Karakalpak language stop words. *Data in Brief*. 2023. Vol. 48. Article 109111. DOI: 10.1016/j.dib.2023.109111.
2. Abdurakhmonova N.Z., Sayfullaeva R.R., Shirinova R.K. Leveraging mBERT with Rule-Driven Pre-Processing for Domain-Specific Karakalpak Text Classification. *International Conference on Computer Science and Engineering (UBMK)*, 2025. P. 311–314. DOI: 10.1109/UBMK67458.2025.11207028.
3. Manning C.D., Raghavan P., Schütze H. *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.
4. Joachims T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the European Conference on Machine Learning (ECML)*. Springer, 1998. P. 137–142.
5. 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. DOI: 10.18653/v1/N19-1423.
6. Conneau A., Khandelwal K., Goyal N. et al. Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020. P. 8440–8451.
7. Pedregosa F., Varoquaux G., Gramfort A. et al. *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*. 2011. Vol. 12. P. 2825–2830.
8. Joulin A., Grave E., Bojanowski P., Mikolov T. Bag of Tricks for Efficient Text Classification. *Proceedings of EACL*. 2017. P. 427–431.

Proofreader: Zokir ALIBEKOV

Layout and Designer: Oloviddin Sobir ugli

2026. № 10

© When materials are reproduced, the INNOVATION SCIENCE AND TECHNOLOGY journal must be cited as the source. Authors are responsible for the accuracy of the information in materials and advertisements published in the journal. Editorial opinions may not always align with those of the authors. Submitted materials will not be returned to the editorial office.

To publish articles in this journal, you may submit articles, advertisements, stories, and other creative materials through the following links. Materials and advertisements are published on a paid basis.

You may subscribe to the journal at any time using the following details. Once subscribed, please send a screenshot or photo of your payment confirmation to our Telegram page @iqtisodiyot_77. Based on this, we will send the latest issue of the journal to your address each month.

“The journal “INNOVATION SCIENCE AND TECHNOLOGY” has been registered by the Agency for Information and Mass Communications under the Administration of the President of the Republic of Uzbekistan from 09.10.2024 under the registration number №390637. License number: C-5669633. PNFL: 30407832680027

Our address: Tashkent city, Yunusobod district, 19th block,
House 17.



Acceptance of articles

Published every
monthly



Directions

Social, economic, political,
technological, scientific



Scopus || Scientific electronic journal specializing in Scopus

CERTIFICATE NUMBER: №390637

**ORDER NUMBER ACCORDING TO
THE LICENSE REGISTER: C-5669633**

CONTACT:



Contact us
+998 50 737 87 88



Telegram channel
t.me/scopus_IST2100



Journal official website
<https://ist-journal.uz/index.php/IST>