КОМПЬЮТЕРНЫЕ НАУКИ, ПРИБОСТРОЕНИЕ И АВТОМАТИЗАЦИЯ
КОМПЬЮТЕРЛІК ҒЫЛЫМДАР, АСПАП ЖАСАУ ЖӘНЕ АВТОМАТТАНДЫРУ
COMPUTER SCIENCE, INSTRUMENTATION AND AUTOMATION

UDC 519.6
IRSTI 50.05, 50.41
https://doi.org/10.53364/24138614_2025_38_3_6

D. Rakhimzhanov[1*], S. Belginova[2]

[1]Astana IT University, Astana, Kazakhstan
[2] University Turan, Almaty, Kazakhstan

[*]*E-mail: d.rahimzhanov@astanait.edu.kz*

TRANSFORMER MODELS FOR PASSENGER REVIEWS CLASSIFICATION: A STUDY USING RUBERT AND XLM-ROBERTA

Abstract. *This study investigates the development and performance evaluation of transformer-based models for the automatic classification of public transportation passenger reviews, aiming to enhance feedback processing while optimizing issue resolution. Efficient handling of passenger feedback is crucial for improving public transportation services, as unresolved complaints or operational inefficiencies can decrease passenger satisfaction and create logistical challenges. Traditional text classification approaches, such as keyword-based methods or classical Machine Learning (ML) algorithms, struggle with multilingual and heterogeneous textual data, particularly in low-resource languages. This study addresses this gap by systematically comparing transformer-based architectures for review classification in Russian and Kazakh, demonstrating their effectiveness in real-world applications. A key contribution of this research lies in evaluating both language-specific and multilingual transformers on passenger-generated feedback, offering insights into their generalization capabilities. Unlike previous studies, which predominantly focus on English-language datasets, this work introduces a newly created, manually labeled dataset covering diverse real-world scenarios in Russian and Kazakh, enabling an objective comparative analysis. Three transformer models DeepPavlov/rubert-base-cased, XLM-RoBERTa-base, and XLM-RoBERTa-large were trained and tested to assess their ability to process complex multilingual input. Experimental results indicate that XLM-RoBERTa-large achieves the highest classification accuracy (90%), particularly for code-mixed and multilingual reviews, whereas rubert-base-cased performs consistently well for Russian-language feedback (87.667%), reinforcing its suitability for monolingual classification tasks. XLM-RoBERTa-base exhibits a balanced trade-off between accuracy and robustness, making it a viable option for heterogeneous review processing (89.5%). Despite their effectiveness, transformer-based models still encounter challenges related to data balancing and the handling of underrepresented classes, particularly in scenarios with uneven language distributions or domain-specific terminology. These findings confirm that transformer models significantly enhance the automation of passenger feedback classification, providing a scalable solution for public transportation providers.*

Keywords: *Natural language processing, text classification, Transformers, BERT, DeepPavlov, XLM-RoBERTa, passenger feedback review, multilanguage modeling.*

Introduction.

As the number of public transport passengers increases, so does the volume of passenger appeals require prompt processing by support services. Until recently, in many vehicle fleets in

Astana, Kazakhstan, passenger reviews were manually analyzed by employees who read the text of the appeal or complaint, assessed its content, and forwarded it to the relevant department. This process demanded significant time and human resources, particularly during peak periods, leading to delays in processing and reduced passenger satisfaction. Advances in Artificial Intelligence (AI) and Natural Language Processing (NLP) have significantly enhanced the automation of passenger feedback analysis. Social media-based sentiment analysis has proven effective for assessing public transportation satisfaction, offering real-time insights that supplement traditional survey methods [1-2]. Additionally, data-driven approaches such as big data analytics and systematic complaint management have demonstrated the potential to optimize public sector responsiveness and service efficiency, particularly in bus transportation networks [3-4].

Previous research has applied traditional ML methods, such as SVM, Naïve Bayes, and Decision Trees, for text classification tasks in transport-related NLP [5-10]. For instance, [9] utilized an SVM classifier to process sentiment analysis of public transportation feedback from Twitter, achieving high accuracy on well-structured data. However, these models struggle with informal and code-mixed language, which is common in passenger complaints. Similarly, [6] demonstrated that Naïve Bayes is effective for classifying short text reviews, such as product reviews from Amazon, but struggles with generalization when applied to multilingual corpora, limiting its real-world applicability. Moreover, traditional ML models often require extensive feature engineering, making them less adaptable to real-world datasets that exhibit high linguistic variability.

Despite their widespread use, these methods exhibit significant limitations. Manual feature selection in traditional machine learning models relies on predefined linguistic patterns, failing to capture deep semantic relationships within the text, which limits their adaptability to complex language structures. Additionally, these models struggle to generalize across different datasets, particularly in multilingual settings, where linguistic variations and code-mixing introduce additional challenges. Furthermore, classical ML approaches are inefficient in handling noisy text, as passenger feedback is often short, informal, and contains abbreviations, leading to reduced classification accuracy and reliability in real-world applications.

To address the limitations of traditional machine learning approaches, modern NLP techniques employ deep learning architectures, particularly neural networks that automatically extract feature representations from text [11-13]. Recent advancements in deep learning have led to the development of transformer-based architectures, such as BERT and its multilingual adaptations, which have demonstrated superior performance in various text classification tasks [11-13]. Unlike classical methods, transformers utilize self-attention mechanisms that enable them to capture long-range dependencies and contextual nuances. Research indicates that BERT-based models generally surpass traditional machine learning techniques in multilingual text classification, particularly in tasks requiring nuanced contextual comprehension [12]. However, transformer models still face challenges in transport-related multilingual NLP applications. For instance, while mBERT exhibits strong performance in resource-rich languages, its effectiveness declines in low-resource and code-mixed settings due to limited training data and increased linguistic variability [14]. Similarly, while XLM-RoBERTa has been shown to enhance multilingual adaptability, it relies heavily on extensive labeled data for fine-tuning, which remains scarce for Kazakh [15]. DeepPavlov's ruBERT, optimized for Russian-language processing, achieves high accuracy in monolingual tasks; however, its applicability in handling mixed Russian-Kazakh text remains uncertain [14]. These challenges underscore the necessity of systematically evaluating transformer models to determine their suitability for real-world multilingual transport feedback classification.

Existing research on public transport feedback classification has predominantly focused on English-language corpora, with studies utilizing sentiment analysis and social media data to evaluate passenger satisfaction and service quality [1]. Some research has extended this approach to other languages, such as Spanish, as seen in studies conducted in Santiago, Chile, where Twitter

data was analyzed to infer user sentiment regarding public transportation [2]. However, these methods primarily rely on structured datasets and high-resource languages, limiting their applicability in multilingual or low-resource settings. Meanwhile, other studies have explored data-driven techniques to optimize complaint processing in public transport systems [3-4], focusing on operational efficiency rather than addressing the linguistic complexities associated with multilingual passenger feedback. Studies addressing transport-related NLP in underrepresented languages are virtually nonexistent, despite the increasing demand for multilingual AI applications in public services. While general-purpose multilingual NLP research has progressed [14-15], there are no established benchmarks or large-scale annotated datasets for transport-related multilingual classification, particularly in Kazakh and Russian-Kazakh mixed text. This lack of dedicated datasets and evaluations makes it difficult to determine which transformer models are most suitable for processing real-world passenger feedback in multilingual settings.

This study addresses this gap by introducing a novel dataset of real-world passenger reviews in Russian, Kazakh, and mixed-language text, which has been manually annotated for supervised classification. Unlike previous datasets used for transport-related NLP, which predominantly focus on English-language complaints [1], this dataset offers the first systematically annotated multilingual corpus for transport-related feedback in a low-resource setting. It captures linguistic challenges such as code-mixing, informal phrasing, and orthographic variations, ensuring a more realistic evaluation of transformer-based architectures for real-world applications in multilingual public transport services.

Unlike previous research, which primarily focuses on monolingual text classification, our study provides a systematic comparison of multilingual and Cyrillic-specific transformers, highlighting their strengths and limitations for passenger feedback categorization. The findings of this study serve as a foundation for further research in multilingual NLP and contribute to the development of real-world applications for automatic feedback routing in transport companies.

Materials and methods. The automation of text classification processes is an important area of applied research, particularly in the domain of feedback analysis. One of the key factors influencing the success of such research is the availability of real-world datasets, which are essential for training robust ML models. This study utilizes a unique dataset of passenger reviews provided by a public transportation company in Astana, Kazakhstan, covering user-generated feedback from the year 2023. The dataset was compiled from passenger appeals submitted through the official communication channels of the transport company, including mobile applications, customer support hotlines, and online complaint forms. These appeals contain a diverse range of feedback types, reflecting real-world passenger experiences and concerns. However, as the raw data was not pre-labeled, it required manual annotation to create a structured dataset suitable for model training, ensuring high-quality labeling and reliable classification. To ensure linguistic and contextual accuracy, the annotation was performed by researchers specializing in NLP and transportation systems. This process involves carefully assigning reviews to predefined categories while considering linguistic features, contextual meaning, and variations in multilingual content. To improve consistency, annotations were systematically reviewed and refined, minimizing potential discrepancies. This structured approach ensures that passenger reviews are accurately categorized, facilitating efficient processing and automatic redirection to relevant departments.

The dataset includes passenger complaints of varying lengths, from brief statements to detailed descriptions. To improve processing efficiency, complaints were categorized by word count: short (≤5 words), medium (6–20 words), and detailed (>20 words). Figure 1 shows that detailed complaints dominate, followed by medium-length ones, while short complaints are rare. This classification aids optimization, allowing automated handling of short complaints and manual review of detailed cases requiring escalation.

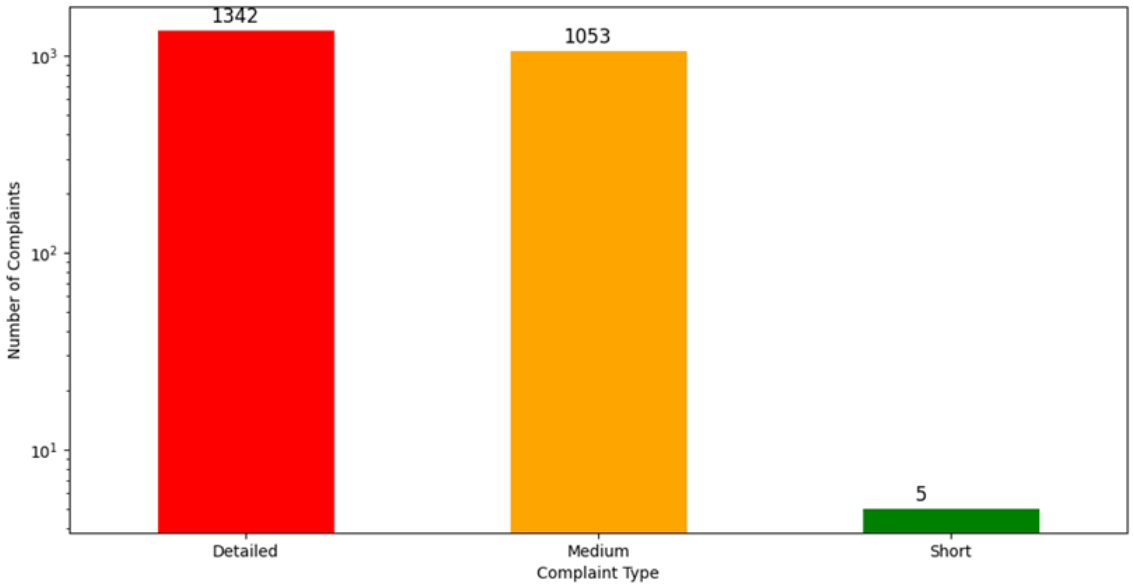

Figure 1 – Distribution of complaints by length

A distinctive feature of this dataset is its multilingual nature, as the reviews are written in both Russian and Kazakh, introducing additional challenges in text processing. Multilingual content variations complicate classification, as passengers may switch between languages within a single review, necessitating models capable of handling code-mixed text. Additionally, the dataset contains syntactic errors and informal language, including abbreviations, typos, and non-standard phrases, further increasing the complexity of text analysis. In some cases, reviews lack sufficient context, making it difficult to accurately determine user intent without context-aware classification methods. Addressing these challenges requires advanced NLP techniques, particularly transformer-based architectures, to improve classification accuracy and enhance the efficiency of automated complaint handling systems.

At table 1 provides an overview of selected complaints that illustrate varying degrees of literacy and informativeness in both Kazakh and Russian. These examples further highlight the linguistic diversity and challenges in processing such a dataset for automated classification purposes.

Table 1 – Selected reviews by literacy and informativeness levels

Literacy & Informativeness Level	Language	Example complaint
High literacy, highly informative	Kazakh	"N маршрут, 07:15-08:15, ЖК name аялдамасында күттік, шара қабылдаңыз."
	Russian	"Маршрут №N. В вечернее время ожидание маршрута более двух часов, нет освещения. Просим наладить график движения!"
Medium literacy, moderately informative	Kazakh	"N маршрут, name университет аялдамасынан name багытта 18:15-18:46 жоқ. Автобус неге кешігіп жүр?"
	Russian	"Автобус №N, 2 женщинам стало плохо из-за давки. Просим пересмотреть количество маршрутов в часы пик!"
Low literacy, low informativeness	Kazakh	"Аялдамада күттік, автобус жоқ, неге?"
	Russian	"Ждал, ждал N-й автобус, не приехал. Что за беспорядок?!"

Code-mixed (Kazakh & Russian)	Mixed	"Жүргізуші не остановился на остановке, хотя біз күткен 20 минут. Это нормально?"
	Mixed	"N м/а Доброе утро! Запустите дополнительные маршруты N автобуса. Бұл мүмкін емес, адамдар сыймайды, есіктер жабылмайды. То двери выламываются в автобусах из-за большого количества людей. Потом приходится ждать автобус и опаздывать на работу!"
Literacy & Informativeness Level	Language	Example complaint
High literacy, highly informative	Kazakh	"N маршрут, 07:15-08:15, ЖК name аялдамасында күттік, шара қабылдаңыз."
	Russian	"Маршрут №N. В вечернее время ожидание маршрута более двух часов, нет освещения. Просим наладить график движения!"

The classification categories were determined through content analysis of the dataset and evaluation of the most frequently occurring themes. This structured categorization ensures that feedback is systematically forwarded to the relevant operational units. Seven distinct categories were identified:

1) Transport problems – Complaints related to route delays, vehicle overcrowding, breakdowns, and bus shortages. These issues constitute a significant proportion of all feedback, requiring immediate action to maintain the stability of transportation services.

2) Personnel problems – Reports concerning driver misconduct, reckless driving, or improper behavior of inspectors. These complaints highlight the need for improved staff training and adherence to service quality standards.

3) Bus stop infrastructure – Issues related to damaged or missing shelters, broken information boards, and the absence of heated waiting areas. Addressing these concerns is essential for enhancing passenger comfort and accessibility.

4) Equipment faults – Reports on technical failures of onboard systems, such as non-functional validators, heating units, and digital information displays. These reviews emphasize the importance of regular vehicle maintenance and technical inspections.

5) Acknowledgments, Praise – Positive feedback regarding driver professionalism, improved route efficiency, and overall service enhancements. This category helps identify successful operational strategies and recognize high-performing staff.

6) Organizational and technical problems – Complaints about errors in route maps, issues with public transport card top-ups, and a lack of real-time scheduling information. These reviews suggest a need for better organizational planning and improved passenger communication.

7) Other complaints – Feedback that does not fit neatly into the defined categories or overlaps multiple issues. This category highlights the diversity of passenger concerns and the need for flexibility in automated classification systems.

While the dataset is highly informative, several limitations must be considered. Data imbalance presents a challenge, as certain categories, such as “Other complaints”, contain significantly fewer samples than major categories like “Transport problems”, potentially affecting model performance. Additionally, noise in the data, including spelling errors, slang, and informal expressions, complicates classification and may necessitate data preprocessing techniques to improve accuracy. Contextual ambiguity is another concern, as some reviews lack explicit details, requiring context-aware NLP models to accurately infer user intent. Finally, the dataset exhibits domain specificity, as it is sourced from a single transport company in Astana, meaning findings may not be directly generalizable to other cities or transportation networks without further

validation.

Unlike most publicly available transport-related review datasets, which are predominantly monolingual and focus on English-language feedback, this dataset presents a unique multilingual perspective, covering passenger-generated content in Russian and Kazakh. This makes it particularly valuable for developing NLP models tailored to low-resource languages and improving multilingual feedback classification in the public transportation domain.

Table 2 presents the distribution of annotated reviews across the training and test datasets. The dataset consists of 2,400 manually labeled reviews, with 1,800 allocated for training and 600 for testing.

Table 2 – Distribution of annotated reviews by category and their division into training and test datasets

No.	Name of class	Number of annotated reviews		
		Total	Training dataset	Test dataset
0	Transport problems	1214	910	304
1	Personnel problems	605	454	151
2	Bus stop infrastructure	175	131	44
3	Equipment faults	157	118	39
4	Praise	117	88	29
5	Organizational and technical problems	104	78	26
6	Other complaints	28	21	7
	TOTAL	2400	1800	600

Models. Modern language models based on the Transformer architecture demonstrate high efficiency in NLP tasks. In this study, we explore the following transformer-based language models: DeepPavlov/rubert-base-cased, XLM-RoBERTa-base, and XLM-RoBERTa-large.

The DeepPavlov/rubert-base-cased model is a Russian-language adaptation of BERT, developed as part of the DeepPavlov project [14]. It has been trained on a large corpus of Russian-language texts, including Wikipedia, news articles, and other open sources. The primary advantage of this model is its ability to process text while considering morphological and syntactic features of the Russian language, making it a strong choice for monolingual tasks [14]. An additional advantage is its support for pretraining on specialized datasets, allowing for adaptation to specific domains. In this study, rubert-base-cased was further trained on a manually labeled passenger feedback dataset containing texts in both Russian and Kazakh. This adaptation improved its performance in a multilingual environment, enhancing classification accuracy.

In contrast, XLM-RoBERTa-base is a multilingual model developed by Meta AI, designed to process texts in over 100 languages, including Russian and Kazakh. Its architecture is based on a multi-headed attention mechanism, enabling it to capture complex dependencies between words. The model employs the Byte-Pair Encoding (BPE) tokenization method, which enhances its robustness against spelling errors and syntactic variability, a particularly valuable feature when dealing with heterogeneous data [15].

The extended version of this model, XLM-RoBERTa-large, has a greater number of parameters and a deeper network, leading to higher accuracy in processing complex linguistic structures. Due to its increased number of layers and attention heads, the model demonstrates improved context understanding, particularly in tasks involving multilingual text analysis [15].

Table 3 summarizes the key architectural features of these models. These parameters directly impact computational efficiency and the model’s ability to process texts of varying length and complexity. While a greater number of layers and parameters enhances performance, it also increases computational requirements.

Table 3 – Key features of the models

Model Name	DeepPavlov/rubert-base-cased	XLM-RoBERTa-base	XLM-RoBERTa-large
Number of layers	12	12	24
Size of hidden layer	768	768	1024
Number of attention heads	12	12	16
Number of parameters	≈ 110 million	≈ 270 million	≈ 550 million
Max. length of input sequence	512	512	512

Experiments. In our experiment, the three selected models were fine-tuned on the dataset described above. The data was split with 75% allocated for training and 25% for testing, while maintaining class balance across all categories. This distribution ensured that each feedback category was equally represented in both the training and test sets, allowing for an unbiased evaluation of the models' performance across the seven predefined categories.

Training was conducted over three epochs using the fine-tuning method on pre-trained transformer models. The optimizer used was AdamW with a learning rate of 2e-5. For training the models, a maximum input sequence length of 256 tokens was used, ensuring a balance between computational efficiency and the quality of text processing. The optimal batch size during the experiments was determined to be 16, which accelerated the training process while maintaining stable gradient descent parameters. During training, the data was shuffled to prevent sequential dependencies, while the test set remained fixed to ensure reproducibility of results.

To assess the quality of classification, several evaluation metrics were utilized, including Accuracy, Macro Average F1-score, Weighted Average F1-score, Precision, and Recall. Accuracy serves as an indicator of the overall performance of the classification model across all categories. The Macro Average F1-score reflects the unweighted mean of the F1-scores across all categories, making it particularly relevant for handling class imbalance and evaluating the model's performance on underrepresented categories. In contrast, the Weighted Average F1-score provides a more comprehensive measure of overall performance by accounting for the proportion of samples in each category, ensuring that the evaluation reflects the real-world distribution of data. Precision and Recall were analyzed for key categories to assess the model's effectiveness in minimizing false positives and capturing true positives. The experimental results, summarized in Table 3, present a comparative analysis of the models for review classification in a multilingual environment, highlighting their respective strengths and limitations.

Results and their discussion.

As shown in Fig. 2, DeepPavlov/rubert-base-cased model testing demonstrated high accuracy for categories with a large amount of data, such as ‘Transport problems’ and ‘Personnel problems’. However, for sparse categories such as ‘Praise’, the F1-score remains low due to the insufficient number of examples in the training set. This result aligns with expectations, as models trained on imbalanced datasets tend to prioritize high-frequency classes, leading to suboptimal performance in underrepresented categories.

	precision	recall	f1-score	support
Transport problems	0.98258	0.92763	0.95431	304
Personnel problems	0.80899	0.95364	0.87538	151
Bus stop infrastructure	0.82759	0.92308	0.87273	26
Equipment faults	0.64286	0.62069	0.63158	29
Praise	1.00000	0.28571	0.44444	7
Organizational and technical	0.91176	0.70455	0.79487	44
Other complaints	0.59524	0.64103	0.61728	39
accuracy			0.87667	600
macro avg	0.82414	0.72233	0.74151	600
weighted avg	0.88559	0.87667	0.87577	600

Figure 2 – Model results: DeepPavlov/rubert-base-cased

The confusion matrix shown in Fig. 3 provides a detailed breakdown of prediction errors and confirms the high performance in 'Transport problems' and 'Personnel problems' categories while highlighting misclassifications in the 'Praise' and 'Other complaints' categories. These misclassifications suggest that DeepPavlov/rubert-base-cased struggles with classes that exhibit greater linguistic variation and fewer training examples.

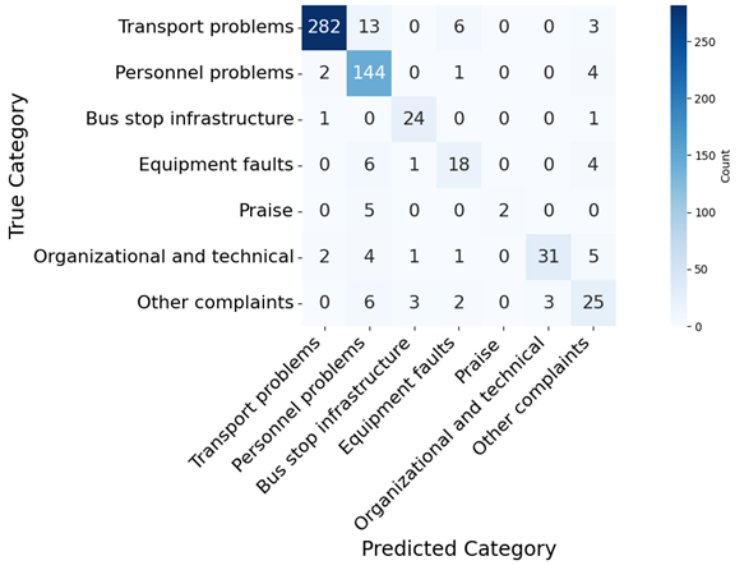

Figure 3 – Confusion Matrix for DeepPavlov/rubert-base-cased

According to the results shown in Fig. 4, the XLM-RoBERTa-base model also showed high accuracy for categories with a large amount of data: ‘Transport problems’ and ‘Personnel problems’. However, for low-resource categories such as ‘Praise’, the F1-score remains suboptimal due to the limited availability of training examples. Notably, XLM-RoBERTa-base exhibited a slight advantage over the DeepPavlov/rubert-base-cased model in the category 'Bus stop infrastructure', suggesting that it is more effective in capturing context in multilingual texts. In the category ‘Equipment faults’ the difference was insignificant, indicating that both models face similar difficulties in handling such complaints. The category 'Other complaints' also showed a slight advantage for XLM-RoBERTa-base, which may indicate its improved ability to handle heterogeneous and multi-thematic reviews. In general, both models perform well in processing large categories of reviews, but XLM-RoBERTa-base demonstrates superior performance in

handling complex and heterogeneous texts, while DeepPavlov/rubert-base-cased is more consistent in Russian-language contexts.

	precision	recall	f1-score	support
Transport problems	0.95161	0.97039	0.96091	304
Personnel problems	0.88344	0.95364	0.91720	151
Bus stop infrastructure	0.88889	0.92308	0.90566	26
Equipment faults	0.70000	0.72414	0.71186	29
Praise	0.00000	0.00000	0.00000	7
Organizational and technical	0.80952	0.77273	0.79070	44
Other complaints	0.67857	0.48718	0.56716	39
accuracy			0.89500	600
macro avg	0.70172	0.69017	0.69336	600
weighted avg	0.88031	0.89500	0.88619	600

Figure 4 – Model results: XLM-RoBERTa-base

The confusion matrix in Fig. 5 provides a detailed visualization of classification performance, indicating the number of correctly classified reviews and misclassifications. This analysis allows for a deeper understanding of the model's strengths and areas for improvement, particularly in handling underrepresented classes.

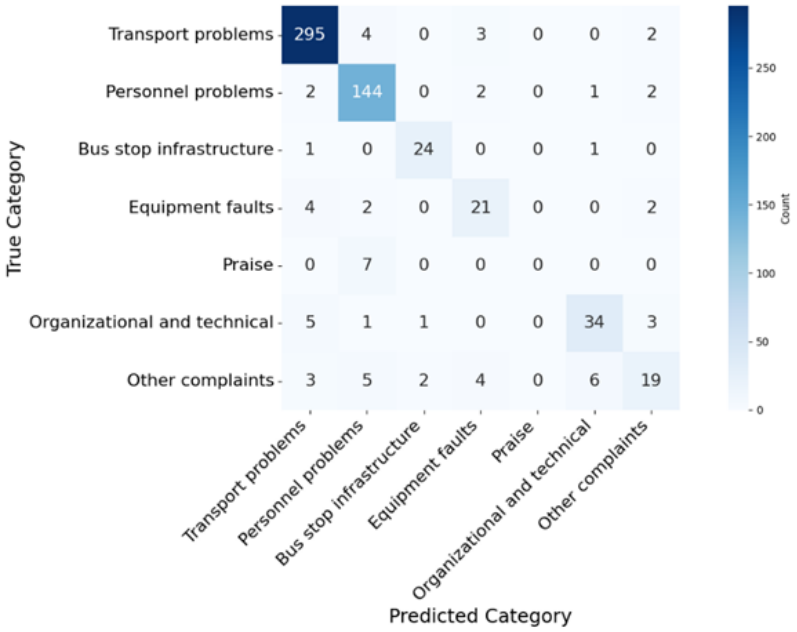

Figure 5 – Confusion Matrix for the XLM-RoBERTa-base model

The test results of the XLM-RoBERTa-large model, shown in Fig. 6, demonstrated high accuracy for most categories, especially for classes with large amounts of data. The categories ‘Transport problems’ and ‘Personnel problems’ showed strong results, indicating the model’s high ability to handle common types of reviews. Additionally, XLM-RoBERTa-large showed significant improvements in the classification of categories such as ‘Bus stop infrastructure’ and ‘Organizational and technical problems’, suggesting that it better captures contextual nuances in multilingual data.

	precision	recall	f1-score	support
Transport problems	0.95146	0.96711	0.95922	304
Personnel problems	0.89744	0.92715	0.91205	151
Bus stop infrastructure	0.92000	0.88462	0.90196	26
Equipment faults	0.72414	0.72414	0.72414	29
Praise	0.66667	0.85714	0.75000	7
Organizational and technical	0.80952	0.77273	0.79070	44
Other complaints	0.73333	0.56410	0.63768	39
accuracy			0.90000	600
macro avg	0.81465	0.81385	0.81082	600
weighted avg	0.89760	0.90000	0.89781	600

Figure 6 – Model results: XLM-RoBERTa-large

Despite the overall improvement in accuracy, the model displayed varying results when processing certain categories. For instance, the ‘Praise’ category achieved an F1-score of 75.00%, which is significantly better than previous models but still indicates room for improvement due to the small dataset size. Meanwhile, the ‘Organizational and technical problems’ category showed notable progress, achieving an F1-score of 79.07%, reflecting the model’s enhanced ability to capture context in this category. Additionally, the F1-score for ‘Bus stop infrastructure’ reached 90.00%, highlighting the model’s ability to process multilingual data with complex context. The confusion matrix in Fig. 7 provides a comprehensive view of classification performance across all categories, highlighting both the model's strengths and areas requiring further optimization.

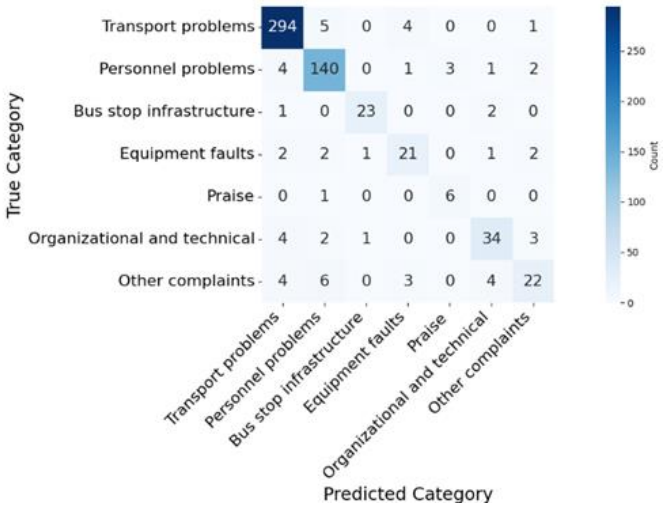

Figure 7 – Confusion Matrix for the XLM-RoBERTa-large model

Comparing XLM-RoBERTa-large with XLM-RoBERTa-base and DeepPavlov/rubert-base-cased, the former demonstrated superior performance in accuracy and F1-score across most categories. The improvement is particularly evident in the ‘Bus stop infrastructure’ category, where XLM-RoBERTa-large achieved an F1-score of 90.20%, emphasizing its stronger ability to capture contextual nuances in multilingual texts. This improvement can be attributed to XLM-RoBERTa’s cross-lingual training on diverse multilingual corpora, which enables better generalization across languages. Unlike DeepPavlov/rubert-base-cased, which is optimized

specifically for Russian, XLM-RoBERTa-large effectively processes both Russian and Kazakh text, reducing errors in code-mixed reviews.

The overall accuracy of XLM-RoBERTa-large was 90.00%, the highest among all tested models, which underscores its enhanced capacity to handle complex and multilingual data. The high weighted F1-score of 89.78% suggests that the model performs well on dominant categories, while the improved macro-average F1-score of 81.80% indicates better generalization across underrepresented classes. In contrast, XLM-RoBERTa-base achieved a lower macro-average F1-score, reflecting difficulties in handling imbalanced categories such as 'Praise' and 'Other complaints'. This highlights that while all models perform well on frequent categories, XLM-RoBERTa-large demonstrates a superior ability to balance classification across all categories, making it the most effective model in a multilingual setting.

As shown in Table 4, DeepPavlov/rubert-base-cased demonstrated consistent performance in processing Russian-language texts, particularly in high-frequency categories like 'Transport problems' and 'Personnel problems'. This can be attributed to the model's architecture, which is specifically fine-tuned for Russian, allowing it to capture linguistic nuances more effectively. However, its performance in rarer categories such as 'Praise' remained limited, likely due to overfitting on dominant classes and difficulties in generalizing to low-resource categories. Meanwhile, XLM-RoBERTa-base delivered comparable accuracy, with the added advantage of effectively handling Kazakh texts and complex reviews due to its multilingual architecture. Overall, XLM-RoBERTa-large, with its more advanced architecture, provided the best performance by achieving a macro-average F1-score of 81.80% and a weighted average F1-score of 89.78%, particularly excelling in the 'Organizational and technical problems' and 'Bus stop infrastructure' categories.

Table 4 – A comparative analysis of the models for the classification of reviews in a multilingual environment

Metrics	DeepPavlov RuBERT	XLM-RoBERTa Base	XLM-RoBERTa Large
Accuracy	87.667%	89.500%	90.000%
Macro Avg (F1-score)	74.151%	69.336%	81.802%
Weighted Avg (F1-score)	87.577%	88.619%	89.781%

Conclusion.

Experimental results demonstrated that leveraging Transformer-based models for automated classification of public transport passenger feedback yields high accuracy. However, analysis highlighted disparities in model performance due to variations in architectural design and multilingual data processing capabilities. A comparative evaluation of the tested models revealed that rubert-base-cased exhibits strong performance when processing Russian-language feedback, particularly in large categories. Following pre-training on Kazakh-language datasets, their effectiveness in multilingual contexts improved but still lagged behind XLM-RoBERTa models in handling sparse categories. XLM-RoBERTa-base outperformed in complex classification tasks due to its multilingual structure but encountered challenges with underrepresented categories. XLM-RoBERTa-large achieved the highest overall accuracy across categories but required significantly greater computational resources. Generally, XLM-RoBERTa-based models excelled in multilingual recall, whereas DeepPavlov/rubert-base-cased, despite enhanced Kazakh-language pre-training, maintained superior consistency in Russian-language classification but struggled with intricate multilingual contexts. The results confirmed that Transformer-based models provide an effective framework for automatic classification of multilingual passenger complaints. While all models demonstrated high accuracy in major categories, improvements in handling rare

complaint types remain necessary. Potential enhancements include expanding data representation for sparse categories, implementing data augmentation techniques, and applying ensemble methods to boost model resilience. Another promising avenue involves pre-training models on extensive domain-specific corpora of Kazakh texts to refine multilingual classification accuracy. Future research should prioritize advancing adaptation strategies for low-resource languages, optimizing data balancing techniques, and investigating hybrid architectures that integrate linguistic and domain-specific knowledge. Moreover, refining pre-training methodologies and enhancing model robustness against linguistic variability will be critical for improving classification efficiency in real-world multilingual feedback processing.

Acknowledgement

This research has been funded by the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number BR24992852 «Intelligent models and methods of Smart City digital ecosystem for sustainable development and the citizens' quality of life improvement».

References

1. Farzadnia, S., & Vanani, I. R. (2022). Identification of opinion trends using sentiment analysis of airlines passengers' reviews. Journal of Air Transport Management, 103, 102232.

2. Méndez J. T. et al. Using Twitter to infer user satisfaction with public transport: the case of Santiago, Chile //IEEE Access. – 2019. – T. 7. – C. 60255-60263.

3. Liu, W. K., & Yen, C. C. (2016). Optimizing bus passenger complaint service through big data analysis: Systematized analysis for improved public sector management. Sustainability, 8(12), 1319.

4. Ghazzawi, A., & Alharbi, B. (2019). Analysis of customer complaints data using data mining techniques. Procedia Computer Science, 163, 62-69.

5. Basu, A., Walters, C., & Shepherd, M. (2003, January). Support vector machines for text categorization. In 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the (pp. 7-pp). IEEE.

6. Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing, 5(2), 221.

7. Soliman, T. H. A., Elmasry, M. A., Hedar, A. R., & Doss, M. M. (2012, October). Utilizing support vector machines in mining online customer reviews. In 2012 22nd International Conference on Computer Theory and Applications (ICCTA) (pp. 192-197). IEEE.

8. Zheng, W., & Ye, Q. (2009, November). Sentiment classification of Chinese traveler reviews by support vector machine algorithm. In 2009 Third International Symposium on Intelligent Information Technology Application (Vol. 3, pp. 335-338). IEEE.

9. Effendy, V., Novantirani, A., & Sabariah, M. K. (2016). Sentiment analysis on Twitter about the use of city public transportation using support vector machine method. Intl. J. ICT, 2(1), 57-66.

10.Chaturvedi, N., Toshniwal, D., & Parida, M. (2019). Twitter to transport: geo-spatial sentiment analysis of traffic tweets to discover People's feelings for urban transportation issues. Journal of the Eastern Asia Society for Transportation Studies, 13, 210-220.

11. González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

12. Alimzhanov, Y. (2020). An academic assistant based on a pre-trained model for contextual answering questions. Вестник вычислительные технологии Национальной инженерной академии Республики Казахстан, Федеральный исследовательский центр информационных и вычислительных технологий, 54. Retrieved from https://acagor.kz/media/uploads/citech-2020/CITech_Part1_Final_rev2.pdf

13. Мукашев, А. Ш., Байзакова, С. М., Едилхан, Д., & Мукашева, А. К. (2022). DEVELOPMENT OF A SYSTEM FOR ANALYZING LARGE TEXT DATA ARRAYS USING DISTRIBUTION SEMANTICS METHODS. Вестник АУЭС, 2(57).

14. Savkin, M., Voznyuk, A., Ignatov, F., Korzanova, A., Karpov, D., Popov, A., & Konovalov, V. (2024, November). DeepPavlov 1.0: Your Gateway to Advanced NLP Models Backed by Transformers and Transfer Learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 465-474).

15. Conneau, A. (2019). Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

ТРАНСФОРМЕР МОДЕЛЬДЕРІ АРҚЫЛЫ ЖОЛАУШЫЛАР ПІКІРЛЕРІН ЖІКТЕУ: RUBERT ЖӘНЕ XLM-ROBERTA МОДЕЛЬДЕРІНІҢ ҚОЛДАНЫСЫ

Аңдатпа. *Бұл зерттеу қоғамдық көлік жолаушыларының пікірлерін автоматты түрде жіктеу үшін трансформер негізіндегі модельдерді әзірлеу және олардың өнімділігін бағалауға бағытталған. Мақсаты - пікірлерді өңдеуді жетілдіру және мәселелерді шешуді оңтайландыру. Жолаушылардың кері байланысын тиімді өңдеу - қоғамдық көлік қызметтерін жақсартудың маңызды бөлігі, себебі шешілмеген шағымдар мен операциялық тиімсіздік жолаушылардың қанағаттануын төмендетіп, логистикалық қиындықтарға әкелуі мүмкін. Кілтсөздерге негізделген әдістер немесе классикалық машиналық оқыту (ML) алгоритмдері сияқты дәстүрлі мәтін жіктеу тәсілдері көптілді және әртүрлі мәтіндік деректермен, әсіресе ресурсы шектеулі тілдерде жұмыс істеуде қиындықтарға тап болады. Бұл зерттеуде аталған мәселе орыс және қазақ тілдеріндегі пікірлерді жіктеудегі трансформаторлық архитектураларды жүйелі түрде салыстыру арқылы шешіледі, олардың нақты қолданыстағы тиімділігін көрсетеді. Зерттеудің басты үлесі - жолаушылардың пікірлері негізінде тілге бейімделген және көптілді трансформерлерді бағалап, олардың жалпылау мүмкіндіктеріне талдау жүргізу. Бұған дейінгі зерттеулер көбіне ағылшын тіліндегі деректер жиынтығына сүйенсе, бұл жұмыс орыс және қазақ тілдеріндегі әртүрлі нақты сценарийлерді қамтитын, қолмен таңбаланған жаңа деректер жиынтығын ұсынады. Зерттеу аясында үш трансформер моделі: DeepPavlov/rubert-base-cased, XLM-RoBERTa-base және XLM-RoBERTa-large жаттықтырылып, олардың күрделі көптілді мәтіндерді өңдеу қабілеті бағаланды. Эксперимент нәтижелері көрсеткендей, XLM-RoBERTa-large ең жоғары дәлдікке (90%) жетіп, аралас тілдер мен көптілді пікірлерді жіктеуде үздік нәтиже көрсетті. Ал rubert-base-cased моделі орыс тіліндегі пікірлерді жіктеуде тұрақты түрде жақсы нәтиже көрсетті (87.667%), бұл оны біртілді тапсырмаларға тиімді етеді. XLM-RoBERTa-base дәлдік пен тұрақтылықтың теңгерімді нұсқасын ұсынып, әртүрлі пікірлерді өңдеуде тиімді шешім ретінде ерекшеленді (89.5%). Алайда, трансформер модельдері тиімді болғанымен, олар деректердің теңгерімсіздігі мен сирек кездесетін санаттарды өңдеу сияқты мәселелерге тап болады, әсіресе тілдер тең бөлінбеген немесе арнайы терминология қолданылатын жағдайларда. Бұл нәтижелер трансформер модельдерінің жолаушылар пікірлерін автоматты түрде жіктеуді едәуір жақсартатынын және қоғамдық көлік қызметтері үшін ауқымды шешім ұсынатынын дәлелдейді.*

Түйін сөздер: *Табиғи тілді өңдеу, мәтінді жіктеу, Трансформер, BERT, DeepPavlov, XLM-RoBERTa, жолаушылар пікірлерін талдау, көптілді модельдеу.*

МОДЕЛИ ТРАНСФОРМЕРЫ ДЛЯ КЛАССИФИКАЦИИ ОТЗЫВОВ ПАССАЖИРОВ: ИССЛЕДОВАНИЕ С ИСПОЛЬЗОВАНИЕМ RUBERT И XLM-ROBERTA

Аннотация. *В этом исследовании изучается разработка и оценка производительности моделей на основе трансформеров для автоматической классификации отзывов пассажиров общественного транспорта с целью улучшения обработки обратной связи при оптимизации решения проблем. Эффективная обработка отзывов пассажиров имеет решающее значение для улучшения услуг общественного транспорта, поскольку нерешенные жалобы или эксплуатационная неэффективность могут снизить удовлетворенность пассажиров и создать логистические проблемы. Традиционные подходы к классификации текста, такие как методы на основе ключевых слов или классические алгоритмы машинного обучения (ML), испытывают трудности с многоязычными и неоднородными текстовыми данными, особенно на языках с низкими ресурсами. В этом исследовании этот пробел устраняется путем систематического сравнения архитектур на основе трансформаторов для классификации отзывов на русском и казахском языках, демонстрируя их эффективность в реальных приложениях. Ключевой вклад этого исследования заключается в оценке как языковых, так и многоязычных трансформеров на основе отзывов пассажиров, что дает представление об их возможностях обобщения. В отличие от предыдущих исследований, которые в основном фокусировались на наборах данных на английском языке, в этой работе представлен недавно созданный, вручную размеченный набор данных, охватывающий различные реальные сценарии на русском и казахском языках, что позволяет проводить объективный сравнительный анализ. Три модели трансформеры DeepPavlov/rubert-base-cased, XLM-RoBERTa-base и XLM-RoBERTa-large были обучены и протестированы для оценки их способности обрабатывать сложный многоязычный ввод. Экспериментальные результаты показывают, что XLM-RoBERTa-large достигает наивысшей точности классификации (90%), особенно для смешанных и многоязычных отзывов, тогда как DeepPavlov/rubert-base-cased работает стабильно хорошо для русскоязычных отзывов (87,667%), что подтверждает его пригодность для задач одноязычной классификации. XLM-RoBERTa-base демонстрирует сбалансированный компромисс между точностью и надежностью, что делает его жизнеспособным вариантом для обработки гетерогенных обзоров (89,5%). Несмотря на свою эффективность, модели на основе трансформера по-прежнему сталкиваются с проблемами, связанными с балансировкой данных и обработкой недостаточно представленных классов, особенно в сценариях с неравномерным распределением языков или терминологией, специфичной для предметной области. Эти результаты подтверждают, что модели трансформеры значительно улучшают автоматизацию классификации отзывов пассажиров, предоставляя масштабируемое решение для поставщиков общественного транспорта.*

Ключевые слова: *Обработка естественного языка, классификация текста, Transformers, BERT, DeepPavlov, XLM-RoBERTa, анализ отзывов пассажиров, многоязыковое моделирование.*

Information about the authors

Daniyar Rakhimzhanov	Master of Science, Junior Researcher of Science and Innovation Center “Big Data and Blockchain Technologies” E-mail: d.rahimzhanov@astanait.edu.kz, Astana IT University, Kazakhstan
Saule Belginova	PhD, Associate professor, E-mail: Sbelginova@gmail.com, University Turan, Kazakhstan

Сведения об авторах

Данияр Рахимжанов	Магистр наук, младший научный сотрудник Научно-инновационного центра “Большие данные и блокчейн-технологии”, Астанинский университет информационных технологий, Казахстан, E-mail: d.rahimzhanov@astanait.edu.kz
Сауле Бельгинова	Кандидат технических наук, доцент, Университет Туран, Казахстан, E-mail: Sbelginova@gmail.com

Авторлар туралы мәлімет

Данияр Рахимжанов	Ғылым магистрі, "Үлкен деректер және блокчейн-технологиялар" ғылыми-инновациялық орталығының кіші ғылыми қызметкері, Астана Ақпараттық технологиялар университеті, Қазақстан, e-mail: d.rahimzhanov@astanait.edu.kz
Сауле Бельгинова	Техника ғылымдарының кандидаты, доцент, Тұран университеті, Қазақстан, e-mail: Sbelginova@gmail.com