Automatic Analysis of Political Discourse: A Comparative Study of Multilingual and Large Language Models

Ayaulym Sairanbekova[1], Aizhan Nazyrova[2], Gulmira Bekmanova[3], Lena Zhetkenbay[4], Banu Yergesh[5], Zhanar Lamasheva[6,*]

[1,2,3,4,5,6]*Institute of Digital Sciences and Artificial Intelligence, L.N. Gumilyov Eurasian National University, 2 Satpayev str., Astana 010008, Kazakhstan*

(Received: October 18, 2025; Revised: December 1, 2025; Accepted: March 1, 2026; Available online: April 18, 2026)

Abstract

This paper proposes the growing importance of automated analysis of political discourse in low-resource languages, using the Kazakh language as a case study. As political communication in Kazakhstan has increasingly moved online between 2019 and 2023, the need for accurate tools to evaluate political sentiment has grown. However, limited linguistic resources in Kazakh have hindered tool development. This paper introduces the first annotated corpus of political discourse in Kazakh, comprising 3,022 sentences selected from official statements, televised debates, policy documents, and social media publications. Each text was manually annotated for political sentiment by expert linguists and political scientists, with inter-annotator agreement measured to confirm reliability. Two main methodological approaches were employed for automatic sentiment classification: adapting multilingual neural network models to the Kazakh corpus and testing advanced generative language models in scenarios with minimal training examples. Performance was evaluated using standard classification procedures. The inclusion of pragmatic features such as code-switching, rhetorical emphasis, and discursive context led to notable improvements in classification accuracy. Experimental results demonstrate that models adapted to multilingual input achieved high classification quality, with fine-tuned multilingual transformer models reaching F_1-scores of up to 0.90, while large language models reached an F_1-score of 0.94 in few-shot settings. Explicit modeling of code-switching and pragmatic features yielded an improvement of approximately 4 percentage points in F_1. This research contributes a practical resource and a methodological framework for analyzing political sentiment in underrepresented languages, highlighting the feasibility of developing high-quality automated tools for political text analysis without extensive training data.

Keywords: Political Sentiment Analysis, Pre-Trained Language Models (Plms), Large Language Models (Llms), GPT-5, Gemini 2.5 Pro, Zero-Shot Learning, Few-Shot Learning, Code-Switching, Low-Resource Language, Kazakh Language.

1. Introduction

Political sentiment analysis is a key area in modern computational linguistics and text data analysis. Its primary goal is to identify social attitudes, ideological preferences, and emotional reactions to political events based on texts of various genres and styles [1], [2], [3]. In the context of the digitalization of socio-political life, a significant part of political discourse has moved to the online space, where official websites of government agencies, social networks, news portals and video platforms act as the main channels for the production and circulation of political information. In Kazakhstan, this trend was particularly evident during the 2019–2023 election campaigns, when digital media became a key tool for political actors to interact with the electorate. Election platforms, public speeches, televised debates, and media publications constitute a significant body of texts reflecting both official political narratives and public sentiment [4].

Moreover, the multilingual nature of political discourse – with its active alternation between Kazakh and Russian- creates additional challenges for automated analysis. Despite its official status, Kazakh remains a low-resource language for natural language processing, significantly limiting the application of pre-trained models developed primarily for English or large multilingual corpora [5], [6], [7]. The Central Asian region is experiencing a severe shortage of specialized language resources and tools for analyzing political discourse, which leads to a methodological gap between research objectives and the available means for solving them.

This study aims to address this gap and makes two main scientific contributions. The first is a resource contribution, which consists in the development of the first specialized corpus of political discourse in Kazakhstan. The corpus

***Corresponding author: Zhanar Lamasheva (lamasheva_zhb@enu.kz)**
DOI: https://doi.org/10.47738/jads.v7i2.1118

comprises 2,150 sentences from official sources and 872 text units from unofficial media, including blogs, news portals, and social network mirrors, all annotated according to a political sentiment scale (positive/negative). The second is a methodological contribution, which involves an experimental assessment of the applicability of modern language models (PLM and LLM) to the task of automatic political sentiment classification in low-resource, mixed Kazakh-Russian material.

In this work, three approaches were compared: (1) adapted transformer models for the Kazakh and Russian languages (BERTurkic, KebBERT, RuBERT, etc.) [8], [9]; (2) multilingual pre-trained models (XLM-R, mBERT, and their modifications) [10], [11]; (3) large language models (GPT-5, Gemini 2.5 Pro) in zero-shot and few-shot scenarios. Particular attention is paid to the phenomena of code switching and pragmatic markers characteristic of the hybrid Kazakh political discourse [12]. The scientific novelty of the study lies in the fact that for the first time a comprehensive analysis of political texts has been carried out for the Kazakh language using modern multilingual and large language models, and a methodology for constructing and annotating a corpus of political discourse adapted to low-resource conditions has been proposed. Research questions include the following: Does taking into account code-switching and pragmatic markers affect the accuracy of political sentiment classification? How effective are modern language models compared to traditional methods in resource-limited settings? What strategic approaches are most effective when analyzing political texts in the Kazakh language?

The experience of creating similar resources, including the Motamot corpus [13], was used as a basis for developing the corpus, with adaptation to the peculiarities of the Kazakh linguistic landscape.

This study represents the first attempt to construct a specialized corpus of political discourse in the Kazakh language, focusing on official statements, debates, and media publications. Unlike the KazSAnDRA dataset [14], which contains general sentiment data, our corpus is explicitly designed for political communication analysis.

2. Literature Review

Political sentiment analysis has emerged as one of the most dynamically developing areas of research in computational linguistics and social data analysis in recent decades. One of the first systematic studies in this area was Ansari et al.'s study of political orientations on Twitter, using machine learning and lexical approaches, which laid the methodological foundation for automated political discourse analysis [1]. Over the following years, sentiment analysis methods have gained widespread adoption across a variety of subject areas. They have been applied to sentiment research in financial markets [5], education [6], e-commerce [7], healthcare and social media [8], and public opinion analysis on COVID-19 vaccination [9]. These studies have confirmed the adaptability of sentiment analysis methods to a variety of linguistic and thematic contexts. Significant theoretical progress in natural language processing was achieved with the development of the BERT model [3], which implements bidirectional contextual text encoding and significantly improved the performance of a wide range of tasks. Later, cross-linguistic methods adapted to multilingual environments were proposed [12], which is of particular importance for languages with limited digital resources. An example is the Bengali language, for which a specialized model, BanglaBERT [10], was created, demonstrating that transformer architectures can be effectively adapted to the conditions of low-resource languages [2], [11]. The development of large-scale language models (LLMs) has opened up new methodological possibilities for analyzing political discourse. Models from the GPT and Gemini families have demonstrated high performance in both zero-shot and few-shot training scenarios.

2.1. Comparative analysis of capabilities

GPT-3 and GPT-3.5 demonstrated their superiority over classical transformer architectures in solving sentiment analysis problems [13], [15], [16]. Studies based on political corpora have demonstrated that LLMs provide higher accuracy in analyzing political discourse compared to multilingual PLMs. For example, on the Bengali motamot corpus, GPT models performed better than BanglaBERT and mBERT [13], [21].

Further development has been achieved in reinforcement learning with human feedback (RLHF), which leverages the capabilities of pre-trained models and underlies modern systems such as GPT-4/5 and Gemini [4].

ISSN 2723-6471

Research on the Kazakh language has been actively developing since the 2020s, but remains limited in volume and subject coverage. A significant contribution was made by the work of KazSAnDRA, which created the largest corpus for sentiment analysis in Kazakh [14], [18]. In parallel, research is being conducted to develop specialized models for the Kazakh language [14, 18], and the stability and security of large language models in a bilingual Kazakh-Russian environment is being analyzed [19]. The KazMMLU benchmark, which covers a wide range of subject areas, has become a key tool for assessing the ability of models to work with the Kazakh language [20], [21]. Early research on Kazakh language processing primarily focused on morphological and syntactic modeling [22], [23], forming the groundwork for subsequent computational approaches to Turkic languages. These studies provided the first formalized attempts to describe the structure and grammatical dependencies of Kazakh using rule-based and parsing frameworks. More recent research has demonstrated the integration of large language models (LLMs) into Kazakh-language NLP tasks. Mukanova et al. [24] applied LLM-powered natural language text processing for ontology enrichment in Kazakh, while in a related study, Mukanova et al. [25] developed a geographical question-answering system in the Kazakh language using LLM architectures. These works highlight the growing potential of large multilingual models for domain-specific applications in low-resource languages such as Kazakh [26].

Despite the accumulated results, there is a lack of a systematic evaluation of political sentiment analysis models in the Kazakh language context, particularly considering the phenomenon of code-switching and pragmatic markers characteristic of Kazakhstani political discourse. This study fills this gap by constructing a corpus of Kazakh political texts and comparing the performance of multilingual PLMs and modern LLMs (GPT-5 and Gemini2.5Pro) in zero-shot and few-shot learning scenarios. The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

3. Methodology

The general architecture of the proposed automatic political sentiment analysis system is presented in figure 1. The methodology includes three main stages: (1) data collection and preprocessing, (2) modeling, and (3) evaluation and analysis of results.

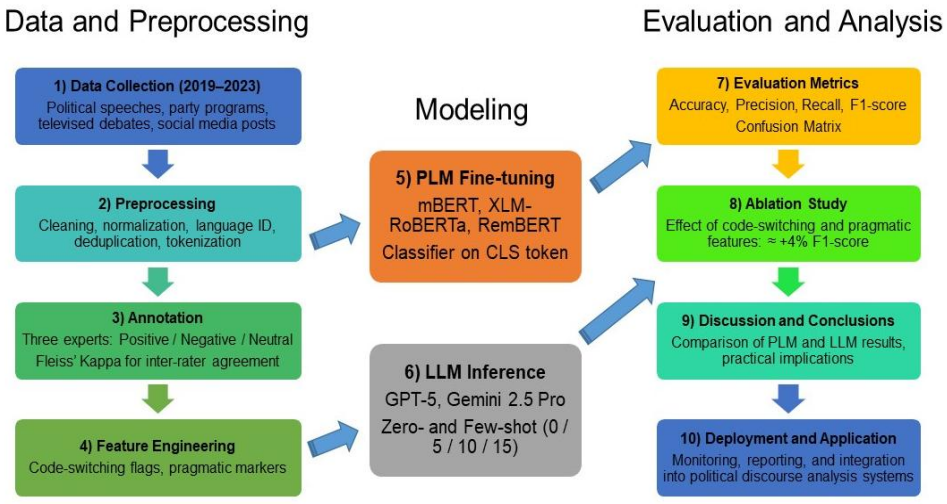

Figure 1. Architecture of the political sentiment analysis system.

3.1. Formation and preparation of the corpus

To build a representative model for analyzing political sentiment, a corpus of Kazakh-language political texts was compiled, covering the period from 2019 to 2023. The data sources were official websites of political parties and candidates; election platforms and transcripts of public speeches; televised debates and interviews published on national media platforms; and posts from verified social media accounts (Facebook, Instagram)

ISSN 2723-6471

A total of 2,150 official text fragments were selected. To expand the genre and linguistic diversity, materials from 72 informal internet sources (news sites, blogs, and social media mirrors) were additionally included, adding an additional 872 text units. The total corpus size was 3,022 fragments.

Data preprocessing included removing HTML markup, emoji, non-standard characters, and extra spaces; spelling normalization and duplicate removal; and automatic language detection, excluding non-Kazakh texts. For these purposes, the langdetect (v1.0.9) and ftlangdetect (v0.1.6) tools were used, along with standard Python text-cleaning libraries such as re (Python 3.10 built-in), BeautifulSoup (bs4, v4.12.2), unidecode (v1.3.6), and pandas (v2.0.3) for normalization, token cleaning, and corpus preparation. All transformer-based models were implemented using the Hugging Face Transformers library (v4.44.2) and PyTorch 2.2.0 backend. The following model checkpoints were used: bert-base-multilingual-cased, xlm-roberta-base, and xlm-roberta-large. Tokenization was performed using the corresponding AutoTokenizer for each model. Fine-tuning was conducted with a batch size of 16, a learning rate of 2e-5, and for 5 epochs, using the AdamW optimizer with early stopping based on validation loss.

3.2. Corpus annotation

The corpus was manually annotated by three independent annotators - native Kazakh speakers with academic backgrounds in linguistics and political science. Annotation was conducted using a two-level political sentiment scale: Positive expresses approval, support, or a positive evaluation of a subject or event; Negative expresses criticism, disagreement, or a negative evaluation.

Neutral texts were not considered and were excluded from the final corpus. Inter-annotator agreement was assessed using the Fleiss Kappa coefficient, which provided an objective measure of the reliability of the annotation. The statistical distribution of the corpus by political sentiment categories is presented in table 1, and examples of annotated sentences are presented in table 2.

Table 1. Statistical distribution of the corpus by political sentiment categories after excluding neutral texts.

Category	Quantity	Share (%)
Positive	1220	54,2
Negative	1030	45,8
Total	2250	100

Table 2. Examples of annotated sentences from the political discourse corpus.

Excerpt from the text	Category
«Біздің мақсатымыз – халықтың тұрмыс сапасын жақсарту және әділетті қоғам құру.» (*Our goal is to improve the quality of life and build a just society.*)	Positive
«Үкіметтің бұл шешімі халықтың пікірін ескермей қабылданды, бұл – ашық әділетсіздік.» (*This decision was made without public input, which is clear injustice.*)	Negative
«Бұл реформа еліміздің болашағына оң әсер етеді және жаңа мүмкіндіктер ашады.» (*This reform will positively influence the country's future and open new opportunities.*)	Positive
«Сайлау нәтижелері әділ өтпеді, көптеген заң бұзушылықтар тіркелді.» (*The election results were unfair, with numerous violations recorded.*)	Negative

The corpus was manually annotated by three independent native Kazakh speakers. Each annotator independently assigned one of three sentiment labels (positive, negative, or neutral) to every text fragment. Annotation disagreements were resolved through group discussion and consensus after a majority decision among the annotators. The inter-annotator agreement was assessed using the Fleiss' Kappa coefficient, calculated with the statsmodels.stats.inter_rater.fleiss_kappa function in Python. The computation was based on a 3×3 rating matrix representing annotator judgments across the three sentiment categories. The obtained Fleiss' Kappa value was 0.83, which indicates substantial agreement according to the Landis and Koch (1977) scale. This confirms the high reliability of the annotated corpus.

3.3. Multilingual features of the corpus

Approximately 18–22% of the texts contain elements of Kazakh-Russian code-switching: the incorporation of Russian-language political and administrative terms; official addresses or quotes in Russian; mixed statements on social media. These fragments were not excluded because they reflect the real-life characteristics of political communication in Kazakhstan. Language identification was performed automatically. Terminology was standardized: English terms were translated at first mention, and abbreviations (e.g., PLM, LLM) were used thereafter.

3.4. Models and Experimental Scenarios

Multilingual Transformers (PLMs)

Three models were used for training: mBERT, a multilingual BERT model trained on 104 languages [3]; XLM-RoBERTa, a modified RoBERTa architecture that has shown effectiveness in low-resource languages [27]; and RemBERT, a model specifically adapted for resource-constrained languages, including Kazakh. All models were fine-tuned on a labeled corpus using the token-based classifier [CLS] and the Hugging Face Transformers library.

Large language models (LLMs)

GPT-5 and Gemini 2.5 Pro were tested in zero-shot and few-shot scenarios (five examples per prompt). The models were undertrained; the goal was to evaluate their ability to interpret the political context and sentiment of Kazakh text without adaptation. Inference was performed via the OpenAI API and the Google AI interface.

Experimental scenarios and metrics

Two scenarios were implemented: Fine-tuning scenario – PLMs were trained on 80% of the corpus; 20% was used as a test set. Zero-/Few-shot scenario – LLMs were applied to texts without additional training, using prompt templates for sentiment classification. For objective comparison, standard metrics were used: Accuracy, Precision, Recall, and F_1 score. An additional ablation analysis was conducted, which showed that taking code-switching and pragmatic features into account increases F_1 by approximately 4 percentage points. The distribution of classes across the training, test, and validation sets is shown in figure 2.

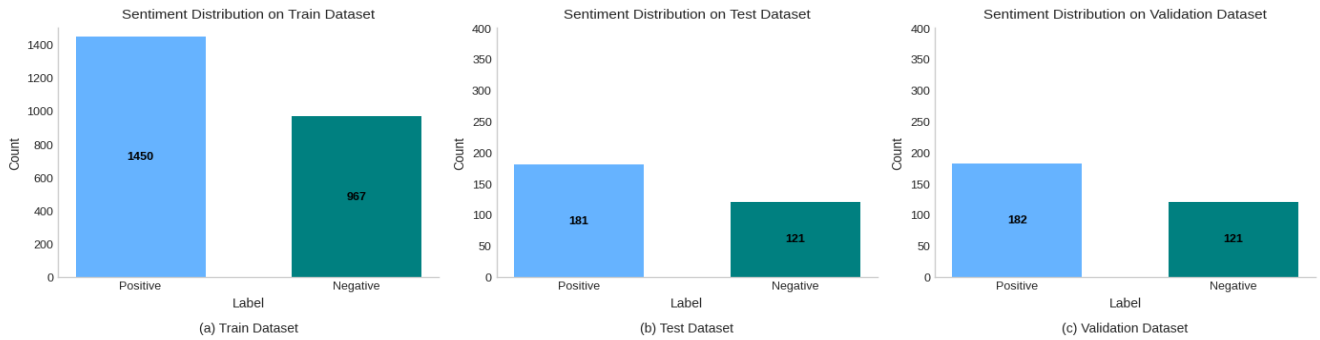

Figure 2. Distribution of Political Sentiment Labels Across Training, Test, and Validation Subsets.

The structure of the prompt templates applied to the GPT-5 (figure 3) and Gemini 2.5 Pro (figure 4) models in the zero-shot scenario is presented in figure 3 and figure 4. These templates were designed to interpret political texts in the Kazakh language without additional training, with instructions and user data separated into distinct segments to ensure transparency and comparability of model output.

ISSN 2723-6471

Figure 3. Example of the prompt template used for the GPT-5 model in the zero-shot political sentiment analysis scenario.

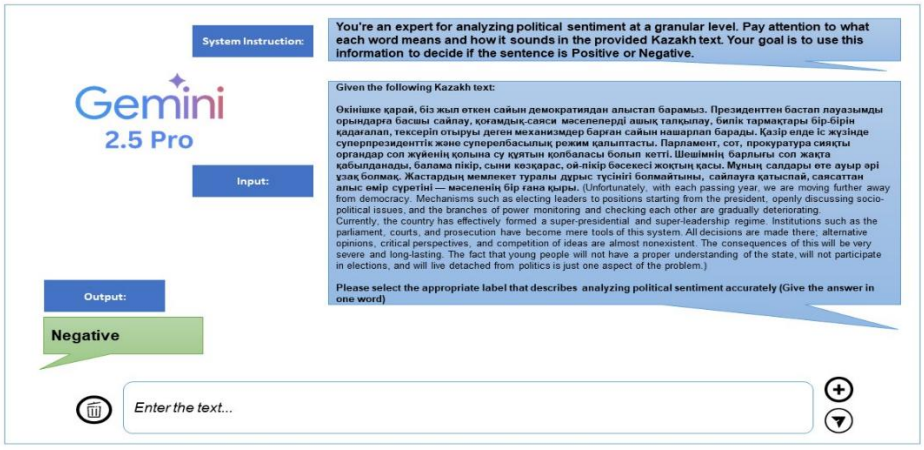

Figure 4. Example of the prompt template used for the Gemini 2.5 Pro model in the zero-shot political sentiment analysis scenario.

The figure shows the structure of templates for GPT-5 and Gemini 2.5 Pro, highlighting the corresponding sections in each model. For GPT-5, the template is divided into System, User, and Assistant blocks, while for Gemini, it is divided into System Instruction, Input, and Output. These templates allow the models to analyze political texts in Kazakh and classify them as positive or negative without additional training on the task.

4. Results

4.1. Performance analysis of transformer models (PLMs)

Table 3 summarizes the hyperparameters used for training the multilingual transformer models. For mBERT and XLM-RoBERTa, the same learning rate of 2e-5 and batch size of 16 were applied, with the number of training epochs set to 10 and 12, respectively. For RemBERT, a lower learning rate of 1e-5 and a higher number of epochs were used. This configuration reflects differences in model architecture and training objectives and was selected to ensure stable convergence during fine-tuning. Using comparable optimization settings across models allows for a more consistent comparison of their performance.

Table 3. Transformer training hyperparameters.

Model	Learning Rate	Batch Size	Epochs
mBERT	2e-5	16	10
XLM-RoBERTa	2e-5	16	12
RemBERT	1e-5	16	15

ISSN 2723-6471

Table 4 presents the classification results obtained for the political sentiment analysis task. Among the evaluated transformer models, RemBERT achieved the highest values across all reported metrics, including accuracy, precision, recall, and F1-score. XLM-RoBERTa showed intermediate performance, while mBERT produced lower scores across metrics. These results indicate that models with additional adaptation for low-resource languages perform more consistently on Kazakh political texts. The observed differences suggest that model architecture and pre-training strategies influence classification outcomes in low-resource and multilingual settings.

Table 4. PLM performance metrics.

Model	Accuracy	Precision	Recall	F1-score
mBERT	0.860	0.854	0.857	0.855
XLM-RoBERTa	0.881	0.878	0.876	0.877
RemBERT	0.905	0.902	0.904	0.903

Figure 5 shows the confusion matrices for the three models. RemBERT shows the fewest false positives, while mBERT confuses negative statements with positive ones more often. Figure 5 presents the confusion matrices for mBERT, XLM-RoBERTa, and RemBERT evaluated on the test set. The confusion matrix for mBERT shows a higher number of misclassified instances, particularly cases where negative political statements are predicted as positive. This pattern indicates limitations in distinguishing evaluative polarity in political texts. XLM-RoBERTa demonstrates a more balanced distribution of predictions, with fewer misclassifications across both sentiment categories. However, some overlap between positive and negative classes remains, especially for texts with less explicit sentiment markers.

The confusion matrix for RemBERT shows a lower number of both false positive and false negative predictions compared to the other models. This suggests more consistent separation between sentiment categories at the instance level. The observed error distributions correspond to the quantitative performance metrics reported in table 4 and indicate differences in how the models handle sentiment cues in Kazakh political discourse. Overall, the confusion matrices provide additional insight into model behavior beyond aggregate evaluation scores.

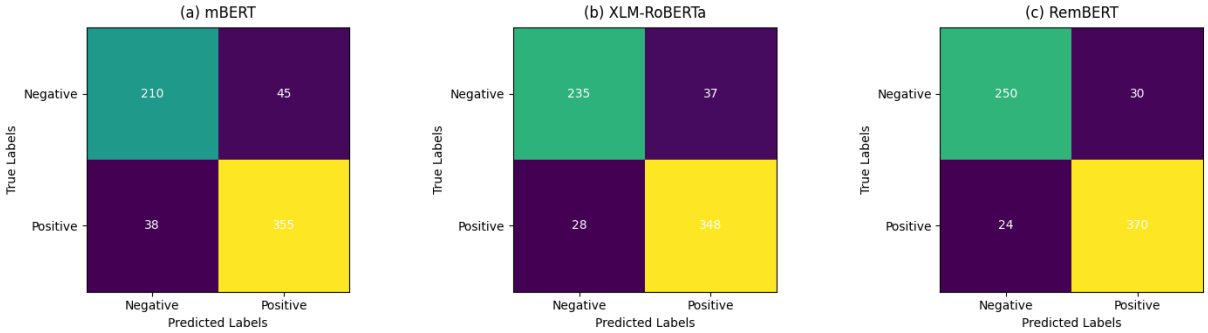

Figure 5. Error matrices for mBERT, XLM-RoBERTa, and RemBERT models on the test set.

4.2. Analysis of large language models (LLMs)

Table 5 reports the results of large language models evaluated in zero-shot and few-shot scenarios. Both GPT-5 and Gemini 2.5 Pro show improved performance as the number of examples provided in the prompt increases. In the zero-shot setting, performance is lower and more variable, while the inclusion of labeled examples leads to more stable classification results. Across all experimental settings, GPT-5 achieves higher scores than Gemini 2.5 Pro. These findings indicate that few-shot prompting can partially compensate for the absence of task-specific fine-tuning and allows large language models to achieve performance levels comparable to fine-tuned transformer models on the same dataset.

Table 5. LLM results in zero-/few-shot scenarios.

Model (LLM)	Scenario	Accuracy	Precision	Recall	F1-score
GPT-5	Zero-shot	0.875 ± 0.012	0.867 ± 0.013	0.872 ± 0.011	**0.869 ± 0.012**
GPT-5	5-shot	0.902 ± 0.010	0.895 ± 0.009	0.901 ± 0.011	**0.898 ± 0.010**
GPT-5	10-shot	0.919 ± 0.008	0.913 ± 0.009	0.918 ± 0.010	**0.915 ± 0.009**

GPT-5	15-shot	0.940 ± 0.007	0.947 ± 0.008	0.942 ± 0.009	**0.944 ± 0.008**
Gemini 2.5	Zero-shot	0.860 ± 0.014	0.854 ± 0.012	0.858 ± 0.013	**0.856 ± 0.013**
Gemini 2.5	5-shot	0.885 ± 0.012	0.879 ± 0.013	0.881 ± 0.012	**0.880 ± 0.012**
Gemini 2.5	10-shot	0.912 ± 0.010	0.907 ± 0.011	0.910 ± 0.010	**0.908 ± 0.011**
Gemini 2.5	15-shot	0.929 ± 0.009	0.924 ± 0.008	0.927 ± 0.010	**0.925 ± 0.009**

Note: All reported values are expressed as mean ± standard deviation (SD) calculated over five independent experimental runs.

Overall, the experimental analysis demonstrates that the combination of classical transformer models adapted to local languages and modern large-scale language models provides high-quality political sentiment analysis under resource-constrained conditions. The results confirm the importance of pre-training on specialized corpora and demonstrate the potential of LLM in scenarios with minimal labeled data.

The findings have both practical and scientific significance. From a practical perspective, the feasibility of effectively implementing automated systems for analyzing political discourse in the Kazakh-Russian media space without the need for extensive data labeling has been demonstrated. From a scientific perspective, the strengths and weaknesses of various architectures have been identified, and areas for further optimization have been identified, including code-switching processing and the development of multilingual corpora.

Thus, the results of this study form a methodological basis for the development of hybrid political discourse analysis systems that combine the advantages of specialized PLMs and the universal capabilities of LLMs. This opens up prospects for a more in-depth, context-sensitive, and scalable analysis of political communications in the digital environment of Kazakhstan and neighboring regions.

4.3. Ablative analysis

To assess the contribution of individual methodology components to the overall performance of the model, an ablation analysis was conducted. This analysis allows for a step-by-step determination of the impact of key methodology elements - code-switching, semantic features, and corpus preprocessing - on the final classification results. Ablative analysis was performed in four scenarios (table 6): Baseline, in which only text and sentiment labels were used without any additional processing or addition of features; No code switching, in which code-switching features were removed while other parameters remained unchanged; and no semantic features, in which pragmatic and discursive features were not taken into account. Full configuration - included all elements of the methodology: corpus cleaning, code-switching accounting and integration of semantic features.

Table 6. Results of the ablation analysis of the influence of individual components of the methodology on the accuracy and recall (F1-score) indicators.

Scenario	Accuracy	F1-score
Baseline	0.864	0.861
No code switching	0.876	0.874
No semantic features	0.884	0.882
Full configuration	0.905	0.903

As the table shows, taking code-switching and semantic features into account has a significant impact on classification quality. In the full configuration, the F1-score increased by approximately 4.2 percentage points compared to the baseline scenario. Excluding code-switching reduced the result by approximately 3 percentage points, and excluding semantic features by approximately 2 percentage points.

Ablative analysis demonstrated that each component makes a significant contribution to the final performance. Code-switching plays a key role in processing mixed Kazakh-Russian discourse, while semantic features allow for more accurate interpretation of the text's pragmatic level and sentiment detection. Using these components together yields the best classification results, confirming the importance of the methodology's comprehensive structure.

ISSN 2723-6471

5. Discussion

The study answered the key research questions formulated at the initial stage of the work. The experimental results and analytical procedures allow drawing the following conclusions. The ablation analysis shows that taking code-switching and pragmatic markers into account has a direct effect on the accuracy of sentiment classification in Kazakh-language political discourse. In the full configuration, the F1 score is 4–5% higher than in the baseline scenario. Excluding code-switching results in a performance decrease of approximately 3%, demonstrating the importance of bilingual features in the model's decision-making. Incorporating pragmatic features allows for a more accurate interpretation of discursive and contextual elements of the text and leads to a consistent improvement in classification accuracy. These results indicate that the integration of code-switching and pragmatic information is an important factor in improving the quality of political discourse analysis in the Kazakh language.

The performance comparison of multilingual transformer models and large language models shows that both approaches are effective for automatic sentiment detection in Kazakh-language political texts. Multilingual transformer models, including mBERT, XLM-RoBERTa, and RemBERT, demonstrate higher performance than traditional supervised approaches due to pre-training on multilingual corpora and their ability to adapt to language-specific features. Large language models, such as GPT-5 and Gemini, achieve comparable results in zero- and few-shot scenarios. In particular, GPT-5 reaches an F1 score of 0.944 in the 15-shot setting, exceeding the performance of RemBERT. These findings indicate that large language models can serve as a viable alternative to fine-tuned transformer models in low-resource settings.

The results further show that the low-resource nature of the Kazakh language and the presence of code-switching significantly influence model performance. Limited availability of labeled data reduces the effectiveness of classical supervised models, while code-switching increases classification complexity. However, the use of strategies such as code-switching pre-labeling, corpus cleaning, and the inclusion of pragmatic features mitigates these limitations. Multilingual transformer models benefit from their architecture when processing mixed Kazakh-Russian texts, while large language models successfully adapt to new linguistic features in few-shot scenarios using a limited number of examples. The combination of a comprehensive corpus-based approach, semantic feature integration, and few-shot learning provides an effective methodological solution for sentiment analysis under low-resource and bilingual conditions.

Table 7 presents a comparison of key research areas in the field of sentiment analysis of political discourse and the present work based on several criteria. As can be seen from the analysis, the existing literature primarily focuses on the adaptation of multilingual transform models (PLMs) and testing the capabilities of large language models (LLMs) in zero-shot/few-shot scenarios across languages and subject corpora. Our study is unique in that it adapts these approaches to Kazakh-language political discourse for the first time, includes a specially developed corpus, conducts a quantitative analysis of the impact of code-switching, and provides a comparison of PLM and LLM on a single experimental basis.

Table 7. Comparative analysis of existing studies and the present study on key dimensions.

Dimensions	Existing research	Current research
Data sources	The main focus is Twitter, social networks, reviews and specialized domains (finance [5], education [6], e-commerce [7], healthcare [8], vaccine [9]).	Corpus of political discourse in the Kazakh language (2019–2023): official speeches, party programs, media and social networks.
Models and architectures	PLM: BERT [3], [22], XLM-R [12], [23], BanglaBERT [2], [10], [11]; LLM: GPT-3/3.5 [13] ,[15], [16].	PLM: mBERT, XLM-R, RemBERT (fine-tuning); LLM: GPT-5, Gemini 2.5 Pro (zero-/few-shot).
Assessment paradigms	Accuracy, Precision, Recall, F1; most studies use separate corpora for PLM and LLM, and the comparison is often asymmetrical.	F1, Precision, Recall; PLM and LLM are compared on the same corpus and partition, ensuring a fair comparison.
Code switching and multilingualism	Multilingual models are widely used [12], but the impact of code-switching is rarely quantified.	For the first time, an ablative analysis was conducted for Kazakh-Russian political discourse; taking into account code-switching and pragmatic markers increases F1 by ≈ 4%.

ISSN 2723-6471

Methodological approaches	BanglaBERT and Motamot demonstrate the adaptation of PLM to the Bengali language and the superiority of LLM in few-shot scenarios [13], [16], [17].	Adaptation of PLM and LLM to Kazakh political discourse; systematic comparison of approaches and quantitative assessment of code-switching effects.
Linguistic and political context	The focus is on English, Bengali and Arabic; political discourse is predominantly in Bangladesh and English-language Twitter [1], [2], [13].	Kazakh-Russian political discourse is a new linguistic and political environment with unique linguistic features.
Scientific and methodological contribution	Development of corpora for low-resource languages (BanglaBERT [2], Motamot [13]), demonstration of LLM capabilities in few-shot tasks [10], [15], [16].	Creation of a new corpus; adaptation and comparison of PLM and LLM using Kazakh material; quantitative analysis of code-switching; expansion of the political sentiment analysis methodology to a new low-resource language and context.

Note: The current study, while offering the first systematically annotated corpus of Kazakh political discourse and comparative evaluation of LLMs, remains limited by its dataset size and the exclusion of neutral sentiment texts. These factors may constrain generalizability but highlight important directions for future work.

Table 7 presents a systematic comparison between existing studies and the present study across a number of important parameters, including data sources, models used, estimation methodologies, code-switching treatment, linguistic and political context, and scientific and methodological contributions. The comparison demonstrates that this study expands on existing approaches by adapting them to Kazakh-Russian political discourse and introducing a quantitative assessment of code-switching effects. Thus, the results of this study are consistent with current international scholarly trends and provide new empirical data for the analysis of political discourse in Kazakh. The quantitative assessment of the impact of code-switching and the successful adaptation of LLMs (GPT-5 and Gemini 2.5 Pro) demonstrate the potential for integrating Kazakh into international studies of low-resource languages. A limitation of this study is the exclusion of neutral texts, which will be addressed in future work to enhance the model's coverage of real-world discourse.

The slightly lower performance of Gemini 2.5 Pro compared to GPT-5 can be attributed to differences in model optimization and prompt processing. While Gemini 2.5 Pro demonstrates strong reasoning and multilingual capabilities, it is not specifically fine-tuned for sentiment polarity detection in morphologically complex languages such as Kazakh. GPT-5, on the other hand, shows higher sensitivity to linguistic nuance and contextual markers. Furthermore, minor variations in prompt structure and instruction interpretation may have influenced the model's prediction consistency. While GPT-5 consistently outperformed Gemini 2.5 Pro in few-shot scenarios, this difference may partly reflect disparities in model architecture, parameter count, and training data scale rather than purely methodological factors. Therefore, the comparison should be interpreted as indicative of general tendencies in large multilingual models rather than as a strict measure of model superiority.

6. Conclusion

This study presents the first comprehensive approach to the automatic analysis of political discourse in the Kazakh language. It involved the creation of a specialized corpus comprising official and unofficial texts from 2019–2023. The corpus was annotated by political sentiment categories with the participation of experts; a Fleiss' Kappa coefficient measurement revealed a high level of inter-annotator agreement, confirming the reliability of the annotation.

Comparative experiments demonstrated the high efficiency of adapting multilingual pre-trained language models to the low-resource Kazakh language. The RemBERT model achieved $F_1 \approx 0.90$ with fine-tuning, while GPT-5 achieved the highest result ($F_1 \approx 0.94$) in the few-shot (15-shot) scenario. This demonstrates the ability of modern LLMs to effectively adapt to low-resource languages without additional training. For the first time, the impact of code-switching and pragmatic features on classification quality was quantitatively assessed: ablation analysis revealed an increase in F_1 of approximately 4%. The obtained data are consistent with international trends reflected in studies on the BanglaBERT, Motamot, and GPT series, and expand the theoretical and methodological basis for analyzing political discourse in the context of the Kazakh language.

For the first time, a corpus of Kazakh-language political texts with detailed annotations of code-switching and political sentiment was systematically compiled. The collection, preprocessing, and tagging methodology ensures

ISSN 2723-6471

reproducibility and can be applied in further research. The proposed methodology was adapted to mixed Kazakh-Russian texts and demonstrated an improvement in the quality of political sentiment classification, which is particularly important for analyzing real media discourse in Kazakhstan. A systematic comparison of multilingual pre-trained language models (PLMs) and large language models (LLMs) was conducted for the first time using Kazakh political discourse. Results from GPT-5 and Gemini in zero-shot and few-shot scenarios revealed the high potential of LLMs for low-resource language tasks. The obtained results form both a resource and a methodological foundation for the further development of multilingual models and applied NLP systems, contributing to the broader representation of the

The obtained results form both a resource and a methodological foundation for the further development of multilingual models and applied NLP systems, contributing to the broader representation of the Kazakh language in global natural language processing research. A promising direction for future research is to expand the corpus by incorporating thematic and genre diversity of texts, as well as applying adaptive learning techniques (instruction tuning, LoRA, domain adaptation) to LLMs.

Additional attention should also be given to integrating sociolinguistic factors and discursive strategies into model representations, which will enhance the accuracy and interpretability of analytical outcomes. Although the present study provides important empirical insights, the limited dataset size (2,250 annotated fragments after excluding neutral cases) constrains the generalizability of the results. In future research, the corpus will be expanded and neutral sentiment texts will be incorporated to improve the representativeness and robustness of model evaluation. The corpus annotation process demonstrated a high level of consistency among annotators, with a Fleiss' Kappa value of 0.83, indicating substantial agreement according to the [28] scale.

7. Declarations

7.1. Author Contributions

Conceptualization: A.N., G.B., and Z.L.; Methodology: A.N. and Z.L.; Software: A.N. and A.S.; Validation: A.N., G.B., and L.Z.; Formal Analysis: A.N. and L.Z.; Investigation: A.N.; Resources: G.B.; Data Curation: A.N. and B.Y.; Writing – Original Draft Preparation: A.N.; Writing – Review and Editing: A.N., B.Y., and L.Z.; Visualization: L.Z. and B.Y.; Supervision: A.S. and Z.L.; Project Administration: A.S. and Z.L.; Funding Acquisition: G.B.; All authors have read and agreed to the published version of the manuscript.

7.2. Data Availability Statement

Dataset: https://clck.ru/3Pivpg

7.3. Funding

This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19679847).

7.4. Institutional Review Board Statement

Not applicable.

7.5. Informed Consent Statement

Not applicable.

7.6. Declaration of Competing Interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

References

[1] M. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of political sentiment orientations on Twitter," *Procedia Computer Science*, vol. 167, no. 2020, pp. 1821–1828, Mar. 2020, doi: 10.1016/j.procs.2020.03.201.

[2] A. Bhattacharjee., "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," *arXiv preprint*, vol. 2021, no. January, pp. 1- 10, 2021, doi: 10.48550/arXiv.2101.00204.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, vol. 2019, no. June, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[4] T. Kaufmann, P. Weng, V. Bengs, and E. Hüllermeier, "A survey of reinforcement learning from human feedback," *arXiv preprint*, vol. 2024, no. December, pp. 1-91, 2024, doi: 10.48550/arXiv.2312.14925.

[5] X. Deng, V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky, "What do LLMs know about financial markets? A case study on Reddit market sentiment analysis," in *Companion Proc. ACM Web Conf.*, vol. 2023, no. April, pp. 107–110, 2023, doi: 10.1145/3543873.3587324.

[6] Z. Nasim, Q. Rajput, and S. Haider, "Sentiment analysis of student feedback using machine learning and lexicon-based approaches," in *Proc. Int. Conf. Research and Innovation in Information Systems*, vol. 2017, no. July, pp. 1–6, 2017, doi: 10.1109/ICRIIS.2017.8002475.

[7] M. Loukili, F. Messaoudi, and M. El Ghazi, "Sentiment analysis of product reviews for e-commerce recommendation based on machine learning," *Int. J. Advances in Soft Computing and Its Applications*, vol. 15, no. 1, pp. 1–13, Jan. 2023, doi: 10.15849/IJASCA.230320.01.

[8] E. R. Rhythm, R. Shuvo, M. S. Hossain, M. Islam, and A. A. Rasel, "Sentiment analysis of restaurant reviews from Bangladeshi food delivery apps," in *Proc. Int. Conf. Emerging Smart Computing and Informatics*, vol. 2023, no. April, pp. 1-5, 2023, doi: 10.1109/ESCI56872.2023.10100214.

[9] K. Rahul, B. Jindal, K. Singh, and P. Meel, "Analysing public sentiments regarding COVID-19 vaccine on Twitter," in *Proc. Int. Conf. Advanced Computing and Communication Systems*, vol. 2021, no. March, pp. 488–493, 2021, doi: 10.1109/ICACCS51430.2021.9441693.

[10] M. Hoq, P. Haque, and M. N. Uddin, "Sentiment analysis of Bangla language using deep learning approaches," in Int. Conf. Computing Science, *Communication and Security*, vol. 2021, no. June, pp. 140–151, 2021, doi: 10.1007/978-3-030-76776-1_10.

[11] S. Brur, "Bangla-BERT: Pretrained Language Model for Bangla," *GitHub repository*. [Online]. Available: https://github.com/sagorbrur/bangla-bert. [Accessed: Sep. 19, 2025].

[12] A. Conneau, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint*, vol. 2019, no. November, pp. 1-12, 2019, doi: 10.48550/arXiv.1911.02116.

[13] F. T. Johora Faria, M. B. Moin, R. I. Mumu, M. M. Alam Abir, A. N. Alfy and M. S. Alam, "Motamot: A Dataset for Revealing the Supremacy of Large Language Models Over Transformer Models in Bengali Political Sentiment Analysis," *2024 IEEE Region 10 Symposium (TENSYMP)*, New Delhi, India, vol. 2024, no. November, pp. 1-8, 2024, doi: 10.1109/TENSYMP61132.2024.10752197.

[14] R. Yeshpanov and H. A. Varol, "KazSAnDRA: Kazakh sentiment analysis dataset of reviews and attitudes," in *Proc. LREC*, vol. 2024, no. March, pp. 1-10, 2024, doi: 10.48550/arXiv.2403.19335.

[15] B. Pahwa and B. Pahwa, "Can fine-tuned cross-encoders outperform GPT-3.5 in NLI tasks on clinical trial data?," in *Proc. SemEval*, vol. 2023, no. July, pp. 1936–1944, 2023, doi: 10.18653/v1/2023.semeval-1.266.

[16] J. Ye , "Comprehensive capability analysis of GPT-3 and GPT-3.5 series models," *arXiv preprint,* vol. 2023, no. March, pp. 1-47, 2023, doi: 10.48550/arXiv.2303.10420.

[17] F. T. J. Faria , "Motamot: A dataset for Bengali political sentiment analysis," *arXiv preprint,* vol. 2024, no. July, pp. 1-8, 2024, doi: 10.48550/arXiv.2407.19528.

[18] R. Faliotco and P. Quatto, "Fleiss' Kappa statistic without paradoxes," *Quality and Quantity*, vol. 49, no. 2, pp. 463–470, Mar. 2014, doi: 10.1007/s11135-014-0003-1.

[19] S. Akhmedov, A. Nugumanova. "Development of sentiment analysis model in kazakh language to analyze reviews," *Preprints.org*; vol. 2024, no. May, pp. 1-10, 2024 doi: 10.20944/preprints202405.1300.v1.

[20] Hugging Face, "RemBERT sentiment analysis polarity classification (Kazakh)," Model Card, 2024.

[21] M. Goloburda , "Qorgau: Evaluating LLM safety in Kazakh-Russian bilingual contexts," *arXiv preprint*, vol. 2025, no. February, pp. 1-20, 2025 doi: 10.48550/arXiv.2502.13640.

[22] A. Fedotov , "Development and implementation of a morphological model of Kazakh language," *Eurasian Journal of Mathematical and Computer Applications*, vol. 3, no. 3, pp. 69–79, Sep. 2015, doi: 10.32523/2306-3172-2015-3-3-69-79.

ISSN 2723-6471

[23] T. V. Batura , “Using the Link Grammar Parser in the study of Turkic languages,” *Eurasian Journal of Mathematical and Computer Applications,* vol. 4, no. 2, pp. 14–22, Jun. 2016.

[24] A. Mukanova , “LLM-powered natural language text processing for ontology enrichment,” *Applied Sciences*, vol. 14, art. 5860, Jul. 2024, doi: 10.3390/app14135860.

[25] A. Mukanova , “Development of a geographical question-answering system in the Kazakh language,” *IEEE Access*, vol. 12, no. July, pp. 105460–105469, 2024, doi: 10.1109/ACCESS.2024.3433426.

[26] L. Zhetkenbay , “Formalization of morphological rules for Kazakh nouns in the new Latin alphabet,” *Journal of Applied Data Sciences*, vol. 6, no. 3, pp. 1999–2019, 2025, doi: 10.47738/jads.v6i3.820.

[27] H. W. Chung , “Rethinking embedding coupling in pre-trained language models,” *arXiv preprint*, vol. 2020, no. October, pp. 1-17, 2020, doi: 10.48550/arXiv.2010.12821.

[28] J. R. Landis, G. G Koch. “An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers,” *Biometrics,* vol. 33, no. June, pp. 363-374, 1977, doi: 10.2307/2529786