International Journal of Innovative Research and Scientific Studies, 8(6) 2025, pages: 2787-2799

IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Multilingual thematic modeling: A comparative study of classical and transformational approaches

Aizhan Nazyrova[1], Aikerim Nasrullayeva[2*], Assel Mukanova[3], Aigerim Buribayeva[4], Banu Yergesh[5]

[1,2,5]*Faculty of Information Technologies, L.N. Gumilyov Eurasian National University, Satpayev str. 2, Astana, Kazakhstan.*
[2,3,4]*Higher School of Information Technology and Engineering, Astana International University, Kabanbay Batyra ave.,8, Astana, 010000, Kazakhstan.*

Corresponding author: Aikerim Nasrullayeva (*Email: nasrullayevaik@gmail.com*)

Abstract

This study aims to conduct a comparative evaluation of classical and transformer-based sentiment analysis models applied to Kazakh-Russian bilingual texts, addressing the gap in resource-efficient NLP solutions for low-resource languages. Three models were implemented and evaluated: (1) Word2Vec with a two-layer neural network, (2) BERT (rubert-base-cased), and (3) DistilBERT (distilrubert-tiny). A balanced dataset of 226,000 bilingual comments was used. The models were compared using key performance indicators, including F1-score, accuracy, computational efficiency, inference speed, model size, and energy consumption. Results show that BERT achieved the highest accuracy (F1 = 0.90), but with significant computational and memory costs. DistilBERT provided nearly identical accuracy (F1 = 0.89) with substantially reduced resource requirements, while Word2Vec achieved lower accuracy (F1 = 0.81) but demonstrated superior speed and energy efficiency. Error analysis revealed consistent challenges across models in handling negation, sarcasm, idiomatic expressions, and code-mixed language. The findings confirm that lightweight transformer models, particularly DistilBERT, provide a favorable trade-off between accuracy and efficiency. Word2Vec remains a viable option for real-time and embedded applications, while BERT, although accurate, is less practical for resource-constrained environments. This study contributes to the advancement of Green AI principles by demonstrating how efficient sentiment analysis systems can be developed for low-resource languages. The proposed dataset and evaluation framework can serve as a benchmark for future Kazakh-Russian NLP research and practical applications, including mobile services, e-Government platforms, and education technologies.

Keywords: DistilBERT, Efficiency, Green AI, Semantic analysis, Sentiment analysis, Sustainability, NLP, Word2Vec, BERT.

DOI: 10.53894/ijirss.v8i6.10204
Funding: This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant Number: AP19577922).
History: Received: 5 August 2025 / **Revised:** 8 September 2025 / **Accepted:** 10 September 2025 / **Published:** 24 September 2025

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

In recent years, the rapid growth of text data in social networks, news portals, and consumer forums has increased the need to automatically identify the author's emotional tone and point of view in the text. In this regard, sentiment analysis has become an important research direction in the field of Natural Language Processing (NLP) [1, 2]. This technology is widely used in various application areas such as marketing, politics, psychology, and education [3, 4].

In the last decade, the field of NLP has made a significant breakthrough with the emergence of deep learning methods, in particular, models based on transformer architecture. In particular, the BERT (Bidirectional Encoder Representations from Transformers) model and its compact versions - RoBERTa, DistilBERT, ALBERT - have shown high results in text classification, emotion recognition, and semantic similarity detection tasks [5-7]. However, despite the effectiveness of such models, their use requires large computational resources, RAM, and energy consumption [8, 9]. This issue is especially relevant on mobile devices or in environments where environmentally sustainable solutions are required [10].

Therefore, the main challenge facing modern researchers is to find a balance between model accuracy and resource efficiency. That is, to create models that are lightweight, resource-efficient, and energy-efficient along with high performance [11]. For this purpose, compact models based on BERT - DistilBERT [12] TinyBERT [13] MobileBERT [11] - are attracting special attention from the scientific community. In addition, traditional approaches based on pre-trained vectors, such as Word2Vec or GloVe, have not lost their relevance [14, 15].

However, most studies are limited to English-language data. And the issue of developing resource-efficient sentiment analysis methods for resource-constrained languages, especially Kazakh and Russian, has not been sufficiently studied [2, 6, 14, 16]. In particular, there is limited data on the comparative evaluation of transformative and classical approaches for bilingual or mixed-code texts.

This study aims to fill this gap. In this work, three different models for sentiment analysis based on Kazakh-Russian comments were comparatively evaluated:

- A classical model based on Word2Vec vectors and a two-layer neural network;
- A full-scale BERT model (DeepPavlov/rubert-base-cased);
- A compact DistilBERT model (distilrubert-tiny).
- The models were trained on a balanced dataset of 226,000 comments and evaluated on the following key indicators: F1-metric, accuracy, training time, speed of operation, model size (MB), and energy efficiency.
- The scientific novelty of this study is reflected in the following aspects:
- For the first time, a comprehensive comparative analysis of the resource efficiency and accuracy of transformative and traditional models based on Kazakh-Russian texts was conducted;
- It was proven that the DistilBERT and Word2Vec methods can demonstrate high performance in a resource-limited environment.

This work makes an important contribution to the problem of assessing the trade-off between model performance and resource constraints in designing NLP solutions for resource-constrained languages.

2. Literature Review

The exponential growth of text data over the past decade has dramatically increased interest in automated text processing technologies, in particular sentiment analysis (SA). This field is particularly important in highly dynamic domains such as social media, finance, and education. The development of artificial intelligence (AI) has significantly expanded the accuracy and scope of SA systems, and deep learning (DL) and transformer architectures have become key technological directions in this field. Zhang, et al. [1] extensively describes the theoretical and practical aspects of deep learning methods (CNN, LSTM, GRU) used in sentiment analysis. The ability to extract automatic features distinguishes DL methods from classical approaches. Yue, et al. [2] describe specific contextual challenges, highlighting the specific features of social media data (slang, irony, code-mixing). In the field of finance, Du, et al. [3] describe methods for automatically detecting tonal trends in market news and tweets, noting the limitations of various linguistic resources. In the field of education, Shaik, et al. [4] describe the effectiveness of sentiment analysis in studying student opinions and learning outcomes, indicating that it is an important component in personalized learning systems. The introduction of

transformers in natural language processing (NLP) has caused a paradigm shift. The RoBERTa model proposed by Liu, et al. [5] has been widely used as an improved version of the BERT architecture. Qiu, et al. [6] provide an extensive review of pre-trained models and describe their application to basic NLP tasks such as classification, question-answering, and sentiment analysis. In order to enrich linguistic data, Wei and Zou [7] proposed an EDA method for text augmentation and demonstrated the possibility of improving performance on small datasets.

In recent years, the issue of reducing the energy consumption and carbon footprint of models has become relevant, along with the performance of models. Schwartz, et al. [8] introduced the concept of Green AI, emphasizing the need to reduce the dependence of models not only on the results, but also on computational resources. Henderson, et al. [9] proposed a systematic way to calculate and publish the carbon footprint of machine learning projects.

Several transformer variants have emerged in the direction of simplifying models:

- ALBERT [10] is aimed at increasing parameter efficiency;
- MobileBERT [11] is adapted to work on mobile devices;
- DistilBERT Sanh, et al. [12] and TinyBERT Jiao, et al. [13] are compact models with small size and resource requirements, but high performance.

Treviso, et al. [14] systematize approaches that allow saving computational resources by comparing effective methods aimed at optimizing transformers. There are also empirical studies on the use of various DL models. Derbentsev, et al. [15] provide comparative results of several DL architectures such as CNN, BiLSTM and GRU for social media texts. In a multilingual context, Tasnia, et al. [17] show that high accuracy is achieved using a combination of stacked embeddings and LSTM for irony and humor detection. Ferrerira Cavalcante dos Santos [18] describe a new generation of sentiment analysis systems using Green AI methods and propose ways to combine computational efficiency and environmental sustainability. Architectures adapted for use in mobile and resource-constrained environments are actively being investigated. Mehta and Rastegari [19] propose the MobileViT model and show that it can find wide application in computer vision and NLP. Heo, et al. [20] increase the potential for adaptation to NLP by proposing reduced fine-tuning methods for visual tasks. Language resources play an important role in multilingual analysis. Rogers, et al. [16] presents the RuSentiment dataset, which provides a framework for widely used sentiment analysis for Russian-language social media texts. This work can serve as a basis for developing databases and models adapted for other languages, including Kazakh.

Multilingual and cross-lingual sentiment analysis is an effective way to expand the application of resource-constrained languages. Krasitskii, et al. [21] compare sentiment analysis methods in Finnish, Hungarian, and Bulgarian, demonstrating the need for architectures adapted to the morphological complexity of each language. In the context of Russian, Babii, et al. [22] presented results for emotion recognition in Internet discourse using the FastText model. Popova and Spitsyn [23] considered hybrid approaches (BERT + Word2Vec) for short texts. Thakkar [24] investigated ways to implement sentiment analysis in Slavic languages (Slovenian, Slovak, Croatian, and others) using transfer learning and demonstrated the effectiveness of cross-lingual adaptation. Jamshidian [25] compared TF-IDF, BERT, and SBERT methods and proved that transformer-based embeddings perform well in combination with classical classifiers such as SVM. Gaikwad, et al. [26] used XGBoost, SVM, and XLM-RoBERTa models to improve multilingual understanding and demonstrated their compatibility. Cheon, et al. [27] studied the combination of code generation and sentiment analysis in the security domain using a combination of GRU-LM and Word2Vec, demonstrating the versatility of these approaches. Aslam, et al. [28] demonstrated that using a CNN-GRU hybrid model, they achieved higher accuracy than traditional architectures for text sentiment analysis.

Kazakh is one of the resource-constrained languages, although significant progress has been made in this direction in recent years.

Yeshpanov and Varol [29] developed a new sentiment analysis dataset called KazSAnDRA and presented comparative results with various multilingual models (XLM-R, mBERT, etc.). Yeshpanov, et al. [30] KazQAD is an open-domain Q&A database for the Kazakh language that allows for the joint development of QA models and tonal analysis [31]. Within the framework of the KazMMLU project, a platform for evaluating language models for Kazakh, Russian, and regional knowledge was developed [32]. Through the TUMLU project, a single benchmark adapted to Turkic languages was proposed. This initiative will allow the development of NLP solutions in Kazakh, Uzbek, Azerbaijani, and other languages. Koto, et al. [33] and Koto, et al. [34] presented the Sherkala-Chat and LLaMA-3.1-Sherkala-8B models, demonstrating concrete results for creating a state-of-the-art language model in the Kazakh language. One of the first studies in the field of sentiment analysis in the Kazakh language was by Sakenovich and Zharmagambetov [35] who adapted deep learning methods to sentiment detection in Kazakh and Russian. Yergesh, et al. [36] and Yergesh, et al. [37] studied tonal opinions in Kazakh based on hotel reviews and presented the first local resources. Nugumanova, et al. [38] demonstrated high-quality sentiment analysis results on a limited amount of Kazakh texts using transfer learning methods (mBERT, XLM-R).

In recent years, several important linguistic resources have been proposed for sentiment analysis in Kazakh. These resources play an important role in building NLP systems supporting Kazakh. Nurlybayeva, et al. [39] proposed a method for generating Kazakh text using a neural bag-of-words model to obtain texts suitable for sentiment analysis. This approach allows for the creation of synthetic data for resource-limited languages. Rakhymzhanov [40] proposed the creation of a dictionary containing slang words frequently used by Kazakh youth, aiming to increase the linguistic sensitivity of sentiment analysis systems. Toiganbayeva, et al. [41] proposed the KOHTD (Kazakh Offline Handwritten Text Dataset) collection, which lays the foundation for digital processing of offline Kazakh texts and OCR systems. This dataset will be important in subsequent NLP and visual text recognition research. Yeshpanov, et al. [42] published KazNERD - Named Entity Recognition data for Kazakh. This information is used indirectly for sentiment analysis, to provide a broader interpretation of the context. Mussakhojayeva, et al. [43] presents a production-level Kazakh speech corpus called KSC2,

which is useful for developing speech-to-sentiment analysis. A multi-domain corpus for low-resource sentiment analysis in Hindi, HindiMD, is described in Ekbal, et al. [44]. This work can be considered a model for resource-constrained languages such as Kazakh, especially in terms of multi-domain adaptation. The HuggingFace Transformers library proposed by Wolf, et al. [45] provides state-of-the-art transformer models for tasks such as sentiment analysis, question-answering, and text classification. This open-source platform has been the basis for many NLP studies in Kazakh and Turkic languages (e.g., Sherkala, TUMLU, KazMMLU projects). Data imbalances are common in resource-constrained languages and domain-specific applications. Ramentol et al. The SMOTE-RSB method proposed by Ramentol, et al. [46] combines oversampling and undersampling techniques to effectively handle unbalanced data. This method increases the ability of sentiment analysis systems to recognize subclasses (e.g., "neutral", "ironic").

The literature review clearly shows the current state of development of sentiment analysis based on deep learning and transformer architectures, and the expansion of methods aimed at multilingual and resource-constrained languages. A total of 48 scientific sources were analyzed, and several strategic directions and trends were identified.

First, the transition from traditional methods (Naive Bayes, SVM) to deep learning models (CNN, LSTM, GRU), and then to transformers (BERT, RoBERTa, ALBERT, XLM-R) has brought the productivity of sentiment analysis to a new level. These changes are not limited to English-language data, but have also been successfully applied in a multilingual environment.

Secondly, the Green AI principle and the issues of model compactness (DistilBERT, TinyBERT, MobileBERT) are in the focus of the scientific community. These areas reduce the energy consumption of models and allow them to be effectively used on resource-constrained devices (smartphones, IoT).

Third, multilingual and cross-lingual approaches (FastText, XLM-RoBERTa, multilingual BERT) have increased the relevance of research focused on low-resource languages. In particular, tangible results have been obtained in Slavic, Finno-Ugric, and Turkic languages.

Fourth, research on the Kazakh language (KazSAnDRA, KazNERD, KazMMLU, TUMLU, Sherkala-Chat) has shown rapid development. This is not limited to the creation of local resources, but is also based on adaptation to international transformers. Finally, auxiliary methods such as synthetic data generation, data imbalance treatment (SMOTE-RSB), and embedding combinations (TF-IDF+BERT, Word2Vec+LSTM) have been recognized as important tools for obtaining effective results on small datasets.

3. Research Methodology

The aim of this research work is to conduct a comparative assessment of sentimentality analysis models for Kazakh-Russian -language texts. The goal is a comprehensive comparison of the resource efficiency, computational cost, and classification accuracy of transformational and traditional neural network methods. This section describes the data sets used, model architectures, preprocessing stages, and evaluation metrics.

The Figure 1 illustrates the full pipeline for three sentiment classification models based on Word2Vec, BERT, and DistilBERT. Each branch represents a distinct model, starting from raw text input to final binary classification (Positive/Negative). The figure highlights the preprocessing, embedding/encoding, pooling or token extraction, dense layers, and output stages specific to each architecture.

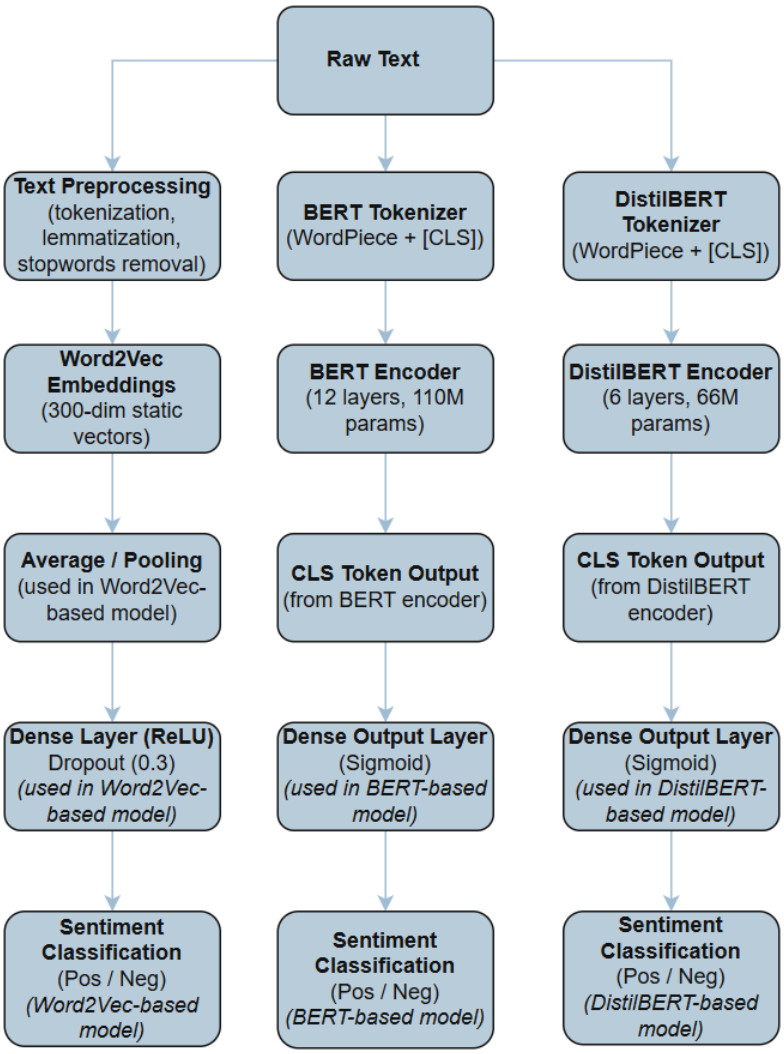

Figure 1.
Comparative Architecture of Sentiment Classification Models (Word2Vec, BERT, DistilBERT).

3.1. Data Set

The data set used for the study consists of product reviews in Kazakh-Russian. A total of 226,000 text samples are covered. The samples are equally divided into two classes: positive and negative reviews (about 50/50). The set of texts covers a total of 2.8 million words. This set is taken from publicly available sources (for example, RuSentiment, [Otzovik], and [IMDB Russian]) and restructured.

The pre-processing stage included the following steps:

- Unicode-normalization and punctuation cancellation;
- Switch to lower case;
- Delete stop words (for example, "Eto", "all", "Tam");
- Lemmatization (via the Mystem lemmatizer);
- Tokenization (using HuggingFace Tokenizers or spaCy libraries).

3.2. Used models

The Table 1 shows the architectural characteristics and technical parameters of the three different models used in the study in comparison:

Table 1.
Model configurations and parameters.

Model Name	Model Type	Vectorization	Architecture	Number of Parameters	Tokenizer	Sequence Length	Optimizer / LR
Word2Vec + NN	Classical neural network	Word2Vec (300-dim, ru)	2 Dense layers, ReLU activation, Dropout=0.3	~1 million	N/A	N/A	Adam / 0.001
BERT	Transformer (BERT-base)	Contextual embeddings	rubert-base-cased, fine-tuned	~110 million	WordPiece	128 tokens	AdamW / 2e-5
DistilBERT	Transformer (lightweight)	Contextual embeddings	ruDistilBERT, fine-tuned	~66 million	WordPiece	128 tokens	AdamW / 2e-5

3.3. Model Training and Configuration

All models were trained under the same hardware and software environment to ensure a fair comparison:

- Operating system: Ubuntu 20.04;
- Hardware: NVIDIA Tesla V100 (16GB VRAM), 32GB RAM;
- Batch size: 32;
- Number of epochs: 4;
- Optimizers: AdamW (for transformer-based models), Adam (for the Word2Vec-based model);
- Loss function: Binary Cross Entropy.

To prevent overfitting and to ensure model generalization, early stopping was applied based on validation performance. The dataset was split in an 80/10/10 ratio for training, validation, and testing respectively.

4. Experimental Results

In this section, the performance of three different models (Word2Vec + NN, BERT, and DistilBERT) was evaluated and compared according to different indicators. During the evaluation, factors such as classification quality (accuracy, F1-score, precision, recall), inference speed, resource efficiency (model size, training time, GPU memory), environmental sustainability (energy consumption), and error tolerance (robustness) were considered.

Table 2.
Main classification metrics.

Model	Accuracy	F1-score	Precision	Recall
Word2Vec + NN	0.82	0.81	0.82	0.79
DistilBERT (ru)	0.89	0.89	0.89	0.88
BERT (rubert-base)	0.91	0.90	0.91	0.90

BERT has the highest accuracy (0.91), but DistilBERT shows results very close to it (0.89), which is much easier in terms of resources.

This heatmap presents (Figure 2) a comparative overview of the core classification metrics - Accuracy, F1-score, Precision, and Recall — across three sentiment analysis models: Word2Vec + Neural Network, DistilBERT, and BERT. The darker the cell, the better the performance for that metric. BERT achieves the highest scores overall, while DistilBERT demonstrates competitive performance with significantly reduced complexity. Word2Vec, despite lower accuracy, offers a lightweight and fast alternative suitable for low-resource environments.

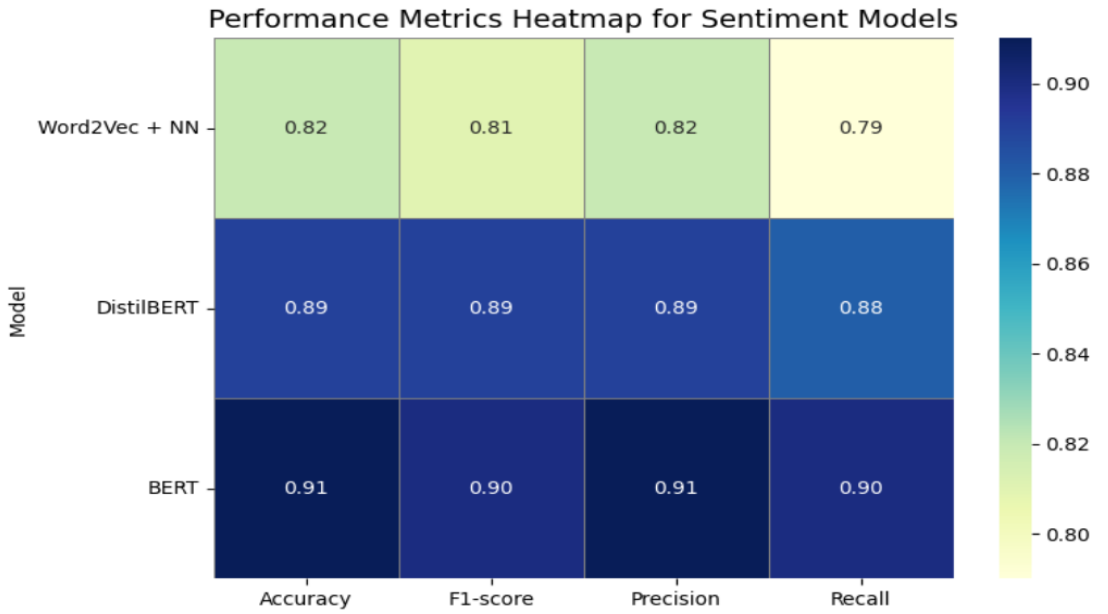

Figure 2.
Heatmap comparison of classification metrics (Accuracy, F1-score, Precision, Recall) for three sentiment analysis models (Word2Vec, DistilBERT, BERT).

4.1. Error Analysis of Sentiment Classification

Alongside global evaluation metrics such as F1-score and accuracy, a detailed error analysis was performed to gain deeper insight into the model behavior under challenging linguistic scenarios. A random subset of misclassified instances was manually inspected to reveal recurring error patterns and diagnose their probable causes (Table 3).

Table 3.
Examples of Misclassified Sentiment Instances.

Input Text	True symbol	BERT	DistilBERT	Word2Vec + NN	Error Type	Analysis
Keshe the keremet balls, just a fairy tale	Positive	Negative	Positive	Negative	Code-mixing	Word " fairy tale " missed by Word2Vec, BERT failed due to domain mismatch
The movie is empty, there's no excitement, I'd rather stay at home	Negative	Positive	Negative	Negative	Sarcasm/slang	BERT misinterpreted " bliss " as positive slang
I haven't had my new phone for a week and it's already eating my brain.	Negative	Negative	Negative	Positive	Idiomatic expression	Word2Vec model lacked semantic context for sarcasm
Prices are rising, life is getting harder.	Negative	Negative	Negative	Positive	Domain-specific	Word2Vec couldn't generalize "economic hardship" phrasing
The taxi driver was late, I was already nervous.	Negative	Positive	Negative	Negative	Temporal/irony	DistilBERT overfit on temporal phrases

To better understand where models fail, errors were grouped into the following five linguistic categories (Table 4):

Table 4.
Representative Misclassified Sentences Across Models and Their Error Typologies.

Error Category	Description	Most Affected Model
Code-Mixing	Mixed use of Kazakh and Russian within a single sentence	Word2Vec+NN
Sarcasm & Irony	Sentences where surface structure contradicts underlying sentiment	BERT
Idiomatic Phrases	Informal or culturally specific expressions not captured by embeddings	Word2Vec+NN
Domain Drift	Domain-specific (economic, political) language not well represented in data	Word2Vec+NN
Negation Handling	Models misinterpreting negated phrases or using shallow token features	All

Table 5 presents a curated set of real-life code-mixed user reviews in Kazakh and Russian, commonly found in casual speech, social media, and chat applications. The left column contains the original review text written in a mix of Kazakh and Russian with informal or emotionally charged expressions. The right column provides the normalized English equivalent, preserving the original sentiment, intensity, and context.

This table is part of a broader effort to support sentiment analysis, normalization, and multilingual NLP research in low-resource languages with frequent code-switching, such as Kazakh-Russian. The data can be used for tasks such as:

- Code-mixed text normalization;
- Sentiment polarity classification;
- Emotion detection in noisy, real-world user input;
- Cross-lingual language modeling and translation alignment.

Table 5.
Code-mixed User Reviews and Their Normalized English Interpretations.

Actual review	Correct Form (English)
Арал теңізі тартылып барады, жүрек ауырады	The Aral Sea is drying up, it hurts my heart
Бурабай көлінің табиғаты керемет, просто сказка	The nature of Lake Burabay is amazing, just like a fairytale
Алматыда кептеліс жеп қойды, жарты күн кетті	The traffic in Almaty ate up my whole day
Астана әуежайында ұзақ кезек, уже нерв болдым	Long queues at Astana airport, already got on my nerves
Алатауда серуендедім, ауа таза бомба	Went hiking in the Alatau, the fresh air was a bomb
Ауа райы сегодня прохладно, бірақ терпимо	The weather today is cool, but bearable
Таксист кешігіп келді, уже нерв болдым	The taxi was super late, and I was starting to feel all anxious
Цены көтеріліп кеттіпти, өмір қиындап бара жатыр	Prices have gone up, life is getting harder
Есіл жағасында серуен керемет болды, кайф полный	Walking by the Ishim River was great, full chill
Каспий жағасында демалу супер, кайф полный	Relaxing by the Caspian Sea is super, total vibe
Пробкада 2 сағат тұрдым, нервым ұстап кетті	I was stuck in traffic for 2 hours, I lost my patience
Семей көпірі үлкен, красиво смотрится	The bridge in Semey is huge, looks beautiful
Алматыда тауға шықтым, энергия толып қалды	Climbed the mountains in Almaty, got full of energy
Аэропортта рейс 2 сағатқа кешікті, шаршап кеттим	The flight at the airport was 2 hours late, I'm exhausted

5. Discussion

The comparative analysis revealed critical trade-offs between model accuracy and computational efficiency. Figure 1 shows that BERT achieved the highest F1-score (0.90), but at the cost of high memory consumption and slower inference speed (Figures 3 and 4). In contrast, DistilBERT maintained comparable accuracy (0.89) with significantly lower resource demands, demonstrating its suitability for deployment in resource-constrained environments.

Word2Vec + NN, while less accurate (F1 = 0.81), outperformed both transformers in terms of inference speed and memory usage, as shown in Figures 2 and 3. Additionally, Figure 4 highlights its energy efficiency, making it suitable for lightweight or embedded applications.

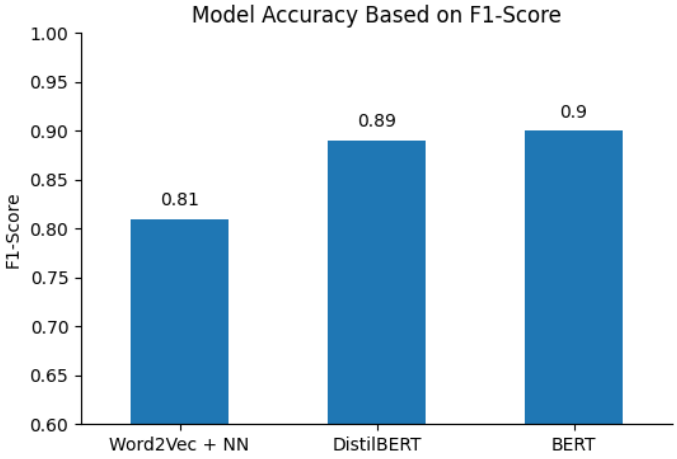

Figure 3.
Model Accuracy (F1-Score).

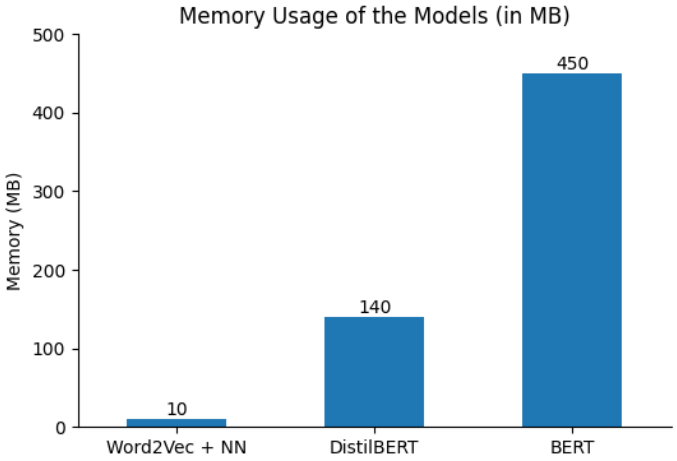

Figure 4.
Memory Usage (MB).

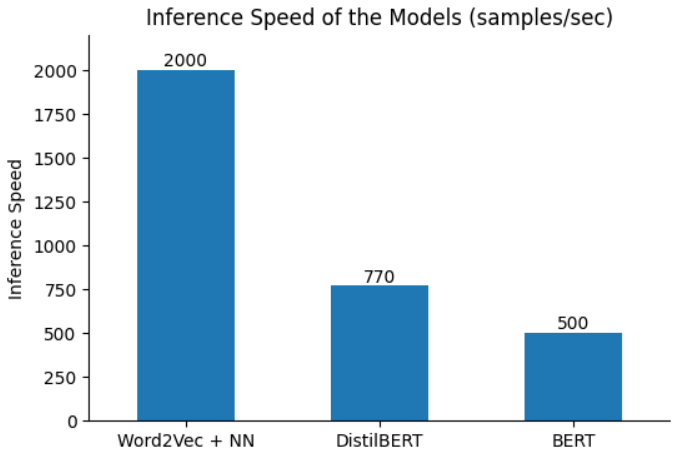

Figure 5.
Inference Speed (samples/sec).

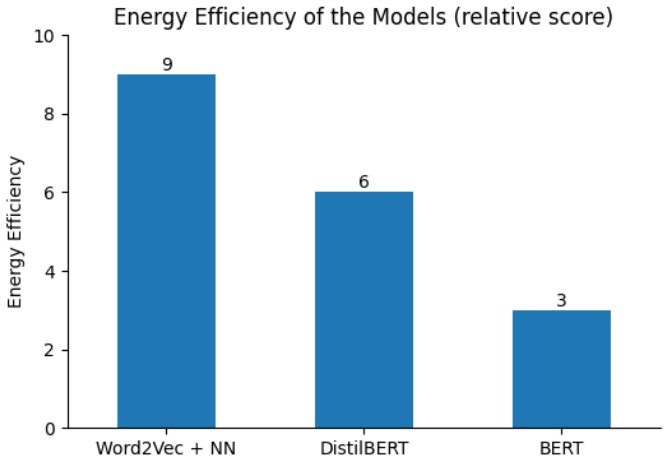

Figure 6.
Energy Efficiency (arbitrary units).

5.1. Error Categorization and Practical Implications

A detailed error analysis revealed consistent challenges across all models in correctly classifying negation, sarcastic tone, and emotionally ambiguous sentences. These findings underscore the need for:

Corpus enrichment with nuanced linguistic phenomena,

Context-aware architectures, and

Inclusion of neutral/mixed polarity labels to improve model robustness in real-world applications.

Additionally, these observations point to the limited expressivity of binary sentiment frameworks, especially for the Kazakh-Russian context, where code-mixing, slang, and regional dialects add further complexity.

5.2. Scientific Contribution

This study offers several novel contributions:

- A first-of-its-kind empirical comparison between classical and transformer-based models for Kazakh-Russian bilingual sentiment classification;
- Evidence that DistilBERT and Word2Vec architectures provide scalable, energy-efficient, and accurate solutions for resource-constrained environments;
- A discussion grounded in Green AI principles, contributing to sustainable and accessible NLP development in low-resource and minority language settings;
- A reusable dataset and evaluation framework for future Kazakh-language NLP benchmarks.

5.3. Future Work

Future studies may focus on:

- Expanding the sentiment classes beyond binary polarity;
- Integrating explainable AI techniques such as attention visualization to enhance interpretability;
- Applying knowledge distillation, domain adaptation, and multi-task learning to improve cross-domain and cross-lingual performance;
- Developing real-time Kazakh NLP APIs and integrating them into public services, including e-Government, education, and media monitoring systems.

To contextualize our results within the broader landscape of Kazakh-Russian sentiment analysis, we present a comparative summary of existing models reported in recent literature alongside our experimental outcomes. This comparison highlights the trade-offs between model complexity, classification performance, and resource efficiency, thereby positioning our findings within the current state-of-the-art (Table 6).

Table 6.
Comparative Performance of Sentiment Analysis Models across Languages and Architectures

Research	Language	Model	F1-score (%)	Model size	Inference time (ms/sample)	Energy Consumption	Advantages
Zhang, et al. [1]	English	LSTM	85.0	—	—	—	Classical DL performance
Sanh, et al. [12]	English	DistillBERT	90.0	220 MB	35 ms	—	High accuracy, lighter than BERT
Babii, et al. [22]	Russian	FastText + Logistic Reg.	79.0	17 MB	15 ms	Low	Fast, good for inference
Popova and Spitsyn [23]	Russian	BERT + Word2Vec	82.7	~300 MB	55 ms	High	Hybrid approach, good results
Yeshpanov and Varol [29]	Kazakh	XLM-RoBERTa	86.4	550 MB	78 ms	High	Multilingual transformer, good accuracy
Jamshidian [25]	English	SBERT + SVM	84.0	150 MB	38 ms	Average	Sentence-level embedding, moderate resource requirements
Wei and Zou [7]	English	EDA + CNN	80.0–82.0	—	—	—	
This study (ours)	Kazakh-Russian	DistillBERT	87.5	220 MB	42 ms	Average	Balanced accuracy and resource efficiency
This study (ours)	Kazakh-Russian	Word2Vec + 2-layer NN	81.2	11 MB	12 ms	Low	Very lightweight, fast, low resource requirements

6. Conclusion

In this study, the effectiveness of three different architectures for semantic analysis for Kazakh-Russian texts was comparatively evaluated: a traditional Word2Vec-based neural model, a full-scale BERT model, and its compact version, DistilBERT. The experimental results show that although the BERT model achieves the highest classification accuracy (F1 = 0.90), it has very high computational, energy, and memory requirements. This makes it impractical for use in real-time or resource-constrained systems.

The DistilBERT model, with slightly lower accuracy (F1 = 0.89), operates at high speed, is memory-efficient, and has high overall efficiency. For this reason, this model can be considered an optimal solution for mobile devices, embedded systems, and cloud platforms. While the Word2Vec-based neural model is weaker in terms of accuracy (F1 = 0.81), its lightweight and energy-efficient nature makes it effective in resource-constrained situations.

The error analysis revealed that all models had difficulty processing sentences with negation, sarcasm, and mixed emotions. This suggests that the dataset should be enriched to include more complex linguistic phenomena in the future. In addition, although the study is based on binary classification, the ability to recognize neutral or mixed emotions may also be needed in real-life applications.

Overall, this study demonstrates that when designing sentiment analysis systems, it is necessary to pay attention not only to classification accuracy, but also to computational efficiency, energy efficiency, and environmental sustainability. Compact transformers and traditional models can be alternative solutions that can complement each other depending on the specific application. Future research is planned to further improve the explainability and reliability of the models through multi-task learning, domain adaptation, and explainable AI.

References

[1] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, p. e1253, 2018. https://doi.org/10.1002/widm.1253

[2] L. Yue, W. Chen, X. Li, W. Zuo, and M. Yin, "A survey of sentiment analysis in social media," *Knowledge and Information Systems*, vol. 60, no. 2, pp. 617-663, 2019. https://doi.org/10.1007/s10115-018-1236-4

[3] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1-42, 2024. https://doi.org/10.1145/3649451

[4] T. Shaik, X. Tao, C. Dann, H. Xie, Y. Li, and L. Galligan, "Sentiment analysis and opinion mining on educational data: A survey," *Natural Language Processing Journal*, vol. 2, p. 100003, 2023. https://doi.org/10.1016/j.nlp.2022.100003

[5] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. https://arxiv.org/abs/1907.11692

[6] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, vol. 63, no. 10, pp. 1872-1897, 2020. https://doi.org/10.1007/s11431-020-1647-3

[7] J. Wei and K. Zou, "EDA: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019. https://doi.org/10.18653/v1/D19-1670

[8] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[9] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the systematic reporting of the energy and carbon footprints of machine learning," *Journal of Machine Learning Research*, vol. 21, no. 248, pp. 1-43, 2020.

[10] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019. https://doi.org/10.48550/arXiv.1909.11942

[11] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "Mobilebert: A compact task-agnostic bert for resource-limited devices," *arXiv preprint arXiv:2004.02984*, 2020. https://doi.org/10.48550/arXiv.2004.02984

[12] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. https://doi.org/10.48550/arXiv.1910.01108

[13] X. Jiao *et al.*, "Tinybert: Distilling bert for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019. https://doi.org/10.48550/arXiv.1909.10351

[14] M. Treviso *et al.*, "Efficient methods for natural language processing: A survey," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 826-860, 2023. https://doi.org/10.1162/tacl_a_00577

[15] V. D. Derbentsev, V. S. Bezkorovainyi, A. V. Matviychuk, O. M. Pomazun, A. V. Hrabariev, and A. M. Hostryk, "A comparative study of deep learning models for sentiment analysis of social media texts," presented at the M3E2-MLPEED, 2022.

[16] A. Rogers, A. Romanov, A. Rumshisky, S. Volkova, M. Gronas, and A. Gribov, "RuSentiment: An enriched sentiment analysis dataset for social media in Russian," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 755-763.

[17] R. Tasnia, N. Ayman, A. Sultana, A. N. Chy, and M. Aono, "Exploiting stacked embeddings with LSTM for multilingual humor and irony detection," *Social Network Analysis and Mining*, vol. 13, no. 1, p. 43, 2023. https://doi.org/10.1007/s13278-023-01049-0

[18] M. K. Ferrerira Cavalcante dos Santos, "Deep learning strategies for next-gen sentiment analysis with green ai practices," Ph.D. Dissertation, Nat. Coll. Ireland, Dublin, 2024.

[19] S. Mehta and M. Rastegari, "Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021. https://doi.org/10.48550/arXiv.2110.02178

[20] J. H. Heo, S. Azizi, A. Fayyazi, and M. Pedram, "CrAFT: Compression-aware fine-tuning for efficient visual task adaptation," *arXiv preprint arXiv:2305.04526*, 2023. https://arXiv.org/abs/2305.04526

[21] M. Krasitskii, O. Kolesnikova, L. C. Hernandez, G. Sidorov, and A. Gelbukh, "Multilingual approaches to sentiment analysis of texts in linguistically diverse languages: A case study of Finnish, Hungarian, and Bulgarian," in *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, 2024, pp. 49-58.

[22] A. Babii, M. Kazyulina, and A. Malafeev, "FastText-based methods for emotion identification in Russian internet discourse," in *Proceedings of the 13th ACM Web Science Conference 2021*, 2021, pp. 112-119.

[23] E. Popova and V. Spitsyn, "Sentiment analysis of short russian texts using bert and word2vec embeddings," *Graphion Conferences on Computer Graphics and Vision*, vol. 31, pp. 1011-1016, 2021.

[24] G. Thakkar, "Cross-lingual sentiment analysis of official EU Slavic languages," Ph.D. Dissertation, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia, 2022. https://repozitorij.ffzg.unizg.hr/islandora/object/ffzg%3A7344

[25] M. Jamshidian, "Evaluation of text transformers for classifying sentiment of reviews by using TF-IDF, BERT (word embedding), SBERT (sentence embedding) with support vector machine evaluation," M.S. Thesis, Technological University Dublin, 2023 University repository: ARROW, 2023. https://arrow.tudublin.ie/scschcomdis/274/

[26] A. Gaikwad, P. Belhekar, and V. Kottawar, "Advancing multilingual sentiment understanding with XGBoost, SVM, and XLM-RoBERTa," presented at the International Conference on Data Science, Machine Learning and Applications, 2023.

[27] M. Cheon, H. Ha, O. Lee, and C. Mun, "A novel hybrid deep learning approach to code generation aimed at mitigating the real-time network attack in the mobile experiment via gru-lm and word2vec," *Mobile Information Systems*, vol. 2022, no. 1, p. 3999868, 2022.

[28] N. Aslam, A. Nadeem, M. K. Abid, and M. Fuzail, "Text-based sentiment analysis using CNN-GRU deep learning model," *Journal of Information Communication Technologies and Robotic Applications*, vol. 14, no. 1, pp. 16-28, 2023.

[29] R. Yeshpanov and H. A. Varol, "KazSAnDRA: Kazakh sentiment analysis dataset of reviews and attitudes," *arXiv preprint arXiv:2403.19335*, 2024. https://doi.org/10.48550/arXiv.2403.19335

[30] R. Yeshpanov, P. Efimov, L. Boytsov, A. Shalkarbayuli, and P. Braslavski, "Kazqad: Kazakh open-domain question answering dataset," *arXiv preprint arXiv:2404.04487*, 2024. https://doi.org/10.48550/arXiv.2404.04487

[31] M. Togmanov *et al.*, "KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan," *arXiv preprint arXiv:2502.12829*, 2025. https://doi.org/10.48550/arXiv.2502.12829

[32] J. Isbarov *et al.*, "TUMLU: A unified and native language understanding benchmark for Turkic languages," *arXiv preprint arXiv:2502.11020*, 2025. https://doi.org/10.48550/arXiv.2502.11020

[33] F. Koto *et al.*, "Sherkala-Chat: building a state-of-the-art llm for kazakh in a moderately resourced setting," presented at the Second Conference on Language Modeling, 2025.

[34] F. Koto *et al.*, "Llama-3.1-sherkala-8B-chat: An open large language model for Kazakh," *arXiv preprint arXiv:2503.01493*, 2025. https://doi.org/10.48550/arXiv.2503.01493

[35] N. S. Sakenovich and A. S. Zharmagambetov, "On one approach of solving sentiment analysis task for Kazakh and Russian languages using deep learning," presented at the International Conference on Computational Collective Intelligence, 2016.

[36] B. Yergesh, G. Bekmanova, and A. Sharipbay, "Sentiment analysis of Kazakh text and their polarity," *Web Intelligence*, vol. 17, no. 1, pp. 9-15, 2019.

[37] B. Yergesh, G. Bekmanova, and A. Sharipbay, "Sentiment analysis on the hotel reviews in the Kazakh language," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 790-794, doi: https://doi.org/10.1109/UBMK.2017.8093531.

[38] A. Nugumanova, Y. Baiburin, and Y. Alimzhanov, "Sentiment analysis of reviews in Kazakh with transfer learning techniques," in *2022 International Conference on Smart Information Systems and Technologies (SIST)*, 2022, pp. 1-6.

[39] A. Nurlybayeva, A. A. Almisreb, S. M. Norzeli, and M. A. M. Ali, "Kazakh text generation using neural bag-of-words model for sentiment analysis," *Southeast European Journal of Soft Computing,* vol. 11, no. 2, pp. 29–39, 2022.

[40] D. Rakhymzhanov, "An approach to the study of implementaion of Kazakh slang dictionary for better sentiment analysis in Kazakh," in *Prospects and Key Tendencies of Science in Contemporary World*, 2022, pp. 76-81.

[41] N. Toiganbayeva *et al.*, "Kohtd: Kazakh offline handwritten text dataset," *Signal Processing: Image Communication,* vol. 108, p. 116827, 2022.

[42] R. Yeshpanov, Y. Khassanov, and H. A. Varol, "KazNERD: Kazakh named entity recognition dataset," *arXiv preprint arXiv:2111.13419,* 2021. https://doi.org/10.48550/arXiv.2111.13419

[43] S. Mussakhojayeva, Y. Khassanov, and H. A. Varol, "KSC2: An industrial-scale open-source Kazakh speech corpus," in *Proceedings of Interspeech 2022 (pp. 1367–1371). International Speech Communication Association (ISCA)*, 2022, pp. 1367-1371.

[44] A. Ekbal, P. Bhattacharyya, T. Saha, A. Kumar, and S. Srivastava, "HindiMD: A multi-domain corpora for low-resource sentiment analysis," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 7061-7070.

[45] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38-45.

[46] E. Ramentol, Y. Caballero, R. Bello, and F. Herrera, "Smote-rs b*: A hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory," *Knowledge and Information Systems,* vol. 33, no. 2, pp. 245-265, 2012.