

# 100,000+ Movie Reviews from Kazakhstan: Russian, Kazakh, and Code-Switched Texts

**Rustem Yeshpanov**

Independent Researcher / Astana, Kazakhstan

yeshpanov.rustem@gmail.com

## Abstract

We present a new publicly available corpus of 100,502 movie reviews from Kazakhstan collected from `kino.kz`, spanning 2001–2025 and covering 4,943 unique titles. The dataset is multilingual, consisting mainly of Russian reviews alongside Kazakh and code-switched texts. Reviews are manually annotated for language and sentiment polarity, and 11,309 reviews additionally contain explicit user-provided ratings. We define two sentiment tasks—three-way polarity classification and five-class score classification—and benchmark classical BoW/TF-IDF baselines against multilingual transformer models (mBERT, XLM-RoBERTa, RemBERT). Experimental results show that transformer models consistently outperform classical baselines on polarity classification, while score classification remains challenging under leakage-controlled evaluation due to severe class imbalance and subtle distinctions between adjacent rating levels.

## 1 Introduction

Movie reviews are widely used in sentiment analysis because they contain naturally occurring, explicitly evaluative language and typically provide more context than short social media posts. However, publicly available datasets of movie reviews in Kazakh remain scarce, limiting reproducible research on sentiment modelling in this under-resourced language. In addition, Kazakhstan provides a practically important multilingual setting in which user-generated reviews are predominantly written in Russian, while Kazakh reviews and code-switching also occur.

We introduce a new publicly available corpus of 100,502 movie reviews collected from `kino.kz`, spanning 25 years (2001–2025) and covering 4,943 unique titles. The dataset includes Russian, Kazakh, and code-switched texts, and is manually annotated for review language and sentiment

polarity. A subset of 11,309 reviews additionally contains explicit user-provided ratings, enabling fine-grained score prediction.

We define two supervised sentiment classification tasks: polarity classification with three labels (negative/neutral/positive) and score classification based on user ratings. We report benchmark results for classical BoW/TF-IDF baselines and multilingual transformer models, including per-language evaluation to characterise performance under data imbalance and code-switching. The dataset, accompanying documentation, and trained models are released to support future work on multilingual sentiment analysis and culturally grounded user-generated text in Kazakhstan and comparable contexts.

While sentiment classification is not the only possible use of this corpus, it provides a widely understood and reproducible probe task for characterising dataset difficulty and establishing baselines on Kazakhstan-specific review discourse. Beyond aggregate scores, our experiments surface two dataset-specific issues that are easy to miss in cleaner benchmarks: (i) neutral polarity is rare and often expressed implicitly, which leads to systematic confusions, and (ii) fine-grained score prediction is highly susceptible to label leakage because users frequently state ratings verbatim in the text, motivating leakage-controlled evaluation. These baselines therefore serve as reference points for future work on multilingual modelling, code-switching, and robust sentiment inference in real-world review text from Kazakhstan.

## 2 Related Work

Movie reviews are a longstanding benchmark for supervised sentiment analysis, dating back to early polarity-classification work on review corpora (Pang et al., 2002). For English, widely used resources such as the IMDb Large Movie Review

Dataset (Maas et al., 2011; Maas, 2011) and the Stanford Sentiment Treebank (Socher et al., 2013) have enabled extensive comparison of both classical and neural approaches across binary and fine-grained sentiment settings.

For Russian, sentiment datasets exist, but fewer have become standard movie-review benchmarks. A closely related resource is the Kinopoisk movie review corpus (Blinov et al., 2013). Other widely used Russian benchmarks focus on different domains, such as social media (e.g., RuSentiment (Rogers et al., 2018), and therefore differ from long-form reviews in length, register, and discourse structure.

For Kazakh, publicly available sentiment resources remain comparatively limited. KazSANDRA (Yeshpanov and Varol, 2024) provides a large-scale Kazakh review dataset (180,064 items) with 1–5 star ratings from four domains (mapping/navigation, e-commerce marketplace, online bookstore, and Android app store). The dataset reflects naturally occurring Kazakh online text, including Kazakh–Russian code-switching and mixed Cyrillic/Latin writing practices, and the accompanying baselines report competitive performance for polarity classification ( $F_1 = 0.81$ ) and substantially lower performance for fine-grained score prediction ( $F_1 = 0.39$ ).

Finally, code-switched sentiment analysis has been studied primarily in short-form social media via shared tasks such as SemEval SentiMix (Patwa et al., 2020). In contrast, our corpus targets long-form movie-review discourse from Kazakhstan and provides a Kazakhstan-specific multilingual setting; while code-switched reviews form a small subset, they still allow targeted evaluation on naturally occurring Kazakh–Russian mixed-language reviews, within the broader review corpus.

Taken together, prior datasets provide limited coverage for Kazakhstan-specific movie-review sentiment with long temporal span and naturally occurring multilingual (Russian/Kazakh) review text, motivating the dataset release and the use of sentiment baselines as a diagnostic benchmark in this work.

### 3 Dataset Development

#### 3.1 Source Data

Movie reviews were collected from kino.kz<sup>1</sup>, a major Kazakh online ticketing and entertainment

<sup>1</sup><https://kino.kz/>

portal launched in 2000, using BeautifulSoup<sup>2</sup>. The platform allows users to browse showtimes, view trailers, access film information, leave reviews, and purchase e-tickets for films, concerts, theatre performances, sports events, and other cultural activities via both its website and mobile applications (Android and iOS). After removing duplicates, the data collected comprised 100,567 reviews, including review text, review date and author, movie title in Russian/Kazakh and English, screening year, genre, director, duration, age restriction, and production country.

Production-country labels are available for 600 of 4,943 titles (12.1%). Among titles with known country labels, the most frequent countries (by number of unique titles) are the United States (182), Kazakhstan (110), the United Kingdom (68), Russia (58), and France (48). Kazakhstan is listed as a production country for 18.3% of titles with known labels. Kazakh-language reviews are more common for these Kazakhstan-produced titles: the median share of Kazakh reviews per title is 0.10, compared to 0.00 for all other titles (mean shares: 0.19 vs 0.009).

#### 3.2 Review Language Identification and Annotation

Since the language of the extracted movie reviews was not provided, the author manually identified the language of each review. Unlike Yeshpanov and Varol (2024), where reviews containing Kazakh-Russian words or grammar were labelled as Kazakh, in this study we aimed to annotate more granularly, labelling reviews as *kk* for Kazakh, *ru* for Russian, *en* for English, *cs* for instances of code-switching, and *ot* for all other languages. While *en* and *ot* reviews were found, they were extremely rare (65 in total) and were therefore excluded from subsequent analyses. Code-switched reviews include two or more languages within a single text, most commonly Kazakh–Russian, occasionally involving English or other languages.

We distinguish code-switching from loanword usage: the *cs* label is applied when a review contains a multiword segment from another language (e.g., an inserted phrase or clause), typically including the function words or grammatical marking of that language (i.e., an extended span in the other language). In contrast, isolated conventional

<sup>2</sup><https://www.crummy.com/software/BeautifulSoup>

borrowings that are integrated into the surrounding language are treated as loanwords and do not, by themselves, warrant cs. Consider the following Kazakh–Russian code-switched review:

*Уақыт аз болмаса, тема фильм. Звуктарды жақсы пайдаланған. Сюжет жақсы, но қысқа. Барып көруге стоит.*

*Uaqyt az bolmasa, tema fil'm. Zvuktardy zhaqsy paidalanğan. Syuzhet zhaksy, no qysqa. Baryp körüge stoit.*

“If you have some time, the film is solid. The sounds are used well. The plot is good, but short. It is worth going to see it.”

As Table 1 indicates, Russian-language reviews constitute the vast majority of the corpus, with smaller subsets of Kazakh and code-switched texts. By whitespace-delimited word count, Russian reviews have a median length of 30 words (95th percentile: 108), Kazakh reviews 24 (95th percentile: 65), and code-switched reviews 33 (95th percentile: 73). The table also shows a strong skew towards positive reviews, a pattern reported for many review platforms (Pang and Lee, 2008); neutral labels are comparatively rare.

Language	Negative	Neutral	Positive	Total
cs	201	62	851	<b>1,114</b>
kk	178	127	2,334	<b>2,639</b>
ru	26,936	4,140	65,673	<b>96,749</b>
<b>Total</b>	<b>27,315</b>	<b>4,329</b>	<b>68,858</b>	<b>100,502</b>

Table 1: Distribution of movie reviews by language and sentiment polarity

Moreover, although the platform allows users to rate movies with stars (from one to ten), these ratings are not publicly displayed, complicating the assignment of polarity scores (positive, neutral, negative). Accordingly, the author manually labelled reviews following guidelines specifically devised for this purpose.

In the absence of additional human annotators, we employed gpt-4.1-nano-2025-04-14 as a compensatory measure to support annotation reliability, which was considered a practical solution under the circumstances. The model was instructed as follows:

*You are a sentiment classifier for movie reviews (Russian or Kazakh). Return one digit only: 2 – clearly positive / recommends the movie; 1 – neutral, mixed, or unclear; 0 – clearly negative / does not recommend. Always choose one digit; never output anything else.*

GPT-generated labels achieved 89.54% accuracy relative to the single-annotator labels over the full corpus, with substantial agreement (Cohen’s  $\kappa = 0.78$ ), indicating strong consistency beyond chance (Landis and Koch, 1977). We report these figures to quantify label stability under single-annotator constraints; the released dataset uses the human annotations as the primary labels.

Furthermore, when available, user ratings were extracted from reviews (e.g., 3 out of 10). For reviews where ratings were provided on a 1–5 scale, scores were multiplied by 2 to align with the standard 1–10 scale. In some cases, users explicitly indicated that a movie was so unsatisfactory that it deserved a score of 0, rather than the minimum 1; these instances were accordingly assigned a rating of 0. Consequently, the final rating scale spans from 0 to 10.

In a small number of cases, the rating format was ambiguous (e.g., a user stating a score of “3” without specifying the scale), which could correspond either to 3/10 or to 3/5 (i.e., 6/10 after normalisation). To resolve such cases, we manually inspected the surrounding review content and inferred the most plausible interpretation based on the expressed sentiment. While we applied this procedure consistently and aimed to minimise errors, a limited number of borderline instances may remain, and the extracted scores should therefore be treated as approximate in rare ambiguous cases.

Overall, 11,309 reviews (approximately 11% of the dataset) contained an explicit user-provided score (e.g., “10/10”, “9 из 10” [*devyat’ iz desyati*, “9 out of 10”], “твердая семерка” [*tvoryodaya semyorka*, “a solid seven”]). Table 2 presents the distribution of explicit user-provided scores across review languages.

In addition, during the review inspection, several recurring themes were occasionally noted, such as unmet expectations, whether the movie was a one-time watch, movie sections perceived as unsatisfactory, and cases where the overall impression was negative but the movie was still recommended for niche audiences. While these observa-

Language	Score											Total
	0	1	2	3	4	5	6	7	8	9	10	
cs	10	2	3	3	2	7	2	8	2	5	106	150
kk	3	4	1	3	3	5	4	7	10	14	212	266
ru	357	317	365	568	520	559	1,077	699	1,244	930	4,257	10,893
<b>Total</b>	<b>370</b>	<b>323</b>	<b>369</b>	<b>574</b>	<b>525</b>	<b>571</b>	<b>1,083</b>	<b>714</b>	<b>1,256</b>	<b>949</b>	<b>4,575</b>	<b>11,309</b>

Table 2: Distribution of user-provided ratings (0–10) by review language

tions were recorded, they are not the focus of the present analysis.

The language identification and annotation process, carried out single-handedly, spanned 110 days, from August 2025 to January 2026.

### 3.3 Collected Data Significance

We argue that the collected movie reviews are of substantial value to the natural language processing (NLP) community for several reasons. First, the dataset spans a period of 25 years, with the earliest reviews dating back to 2001 and the most recent to 2025. Such long-term temporal coverage makes it possible to trace changes in audience preferences and attitudes towards social phenomena and issues (e.g., traditions, domestic violence) over time, ranging from initial denial or avoidance to increased openness and willingness to engage with these topics. Changes in the role and use of the Kazakh language are also clearly observable. In particular, during manual language annotation, we found that although the earliest Kazakh-language review is associated with a film released in 2002, review creation timestamps indicate that the first Kazakh review in our data was authored in 2011, approximately a decade after the launch of the platform (Figure 1).

This likely reflects the initial predominance of Russian-language reviews and the gradual adoption of Kazakh for user-generated content on the platform. Earlier reviews frequently contain criticism of the quality of Kazakh dubbing and translations, or even explicit requests for permission to express opinions in Kazakh (e.g., *можно я на казахском* “May I speak in Kazakh?”), whereas later reviews increasingly express positive attitudes towards Kazakh-language film production and show greater confidence in using Kazakh to articulate opinions. Notably, the five films with the highest numbers of reviews were all produced in Kazakhstan.

Second, the dataset comprises reviews of 4,943 unique movie titles authored by 31,453 publicly visible reviewer identifiers, reflecting a large and diverse pool of contributors. While many identifiers correspond to self-selected usernames, 6,273 reviews (approximately 19%) are associated with a generic, platform-assigned label (e.g., “Kino.kz user”, Russian: “Пользователь kino.kz”), indicating anonymous or non-registered reviewers. Although such entries cannot be distinguished at the individual level, they constitute a substantial portion of the dataset and further contribute to its overall diversity. For release, reviewer identifiers are anonymised by replacing each unique user string with a stable pseudonymous identifier, preserving within-user consistency while removing direct identifiers; reviews associated with the platform-generic label remain indistinguishable, consistent with the source platform.

Third, although the dataset is dominated by Russian-language reviews, the variety of Russian observed is of particular relevance. Specifically, the reviews frequently employ features of Kazakhstani Russian, a regional variety shaped by sustained contact with Kazakh and by local sociocultural context. This includes references to culturally specific events, institutions, and named entities, as well as lexical items and expressions uncommon or opaque to speakers of Russian outside Kazakhstan. Examples include *ажека, агашика, бастык, токалка, болашаковицы, шапалак, уят, Наурыз, Бауржан Шоу, Sulpak, Керуен, Kcell, Otau Cinema*, referring to kinship terms, social roles, cultural concepts, holidays, media productions, and local organisations specific to the Kazakhstani context, as well as regionally marked constructions such as *чѐп-чѐрный* (“pitch black”), which illustrates calquing of Kazakh reduplicative intensification patterns into Russian; *не уятыте* (“do not shame [someone]”), an example of contact-induced verb formation combining a

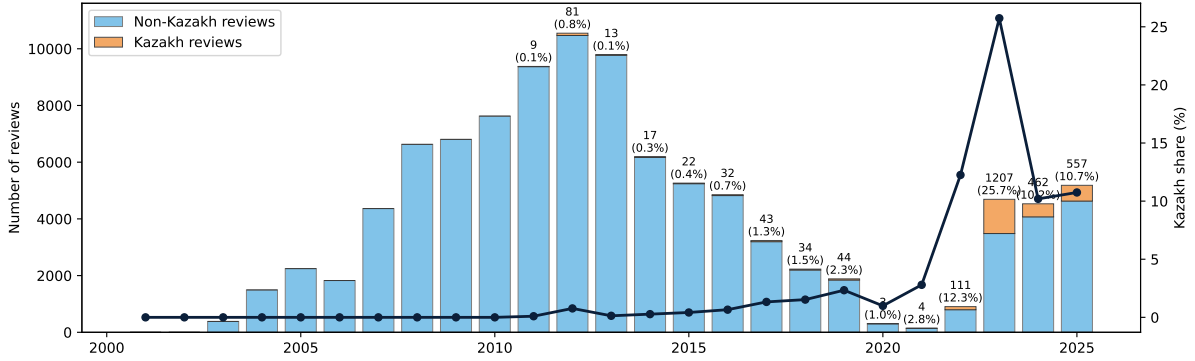


Figure 1: Kazakh-language review share over time

Kazakh lexical root with Russian negation and imperative morphology; and *еркеку* (“men”), formed using a Kazakh lexical root combined with a Russian plural inflection. Such phenomena make the data collected particularly valuable for studying regional language variation, code-switching, and culturally grounded named entity usage in real-world user-generated text.

### 3.4 Sentiment Classification Tasks

Following the design of prior work on Kazakh sentiment analysis, particularly KazSAnDRA, we formulate two primary sentiment classification tasks for our dataset. First, we define a polarity classification (PC) task, in which reviews are categorised into three broad sentiment categories: positive, neutral, and negative. Second, we consider a score classification (SC) task based on explicit user-provided ratings extracted from reviews.

During dataset construction, user ratings were normalised to a unified 0–10 scale. Accordingly, the initial formulation of the SC task involved predicting 11 discrete score labels (0–10). However, as shown in Table 2, the distribution of scores is highly imbalanced, with a substantial concentration of reviews assigned the maximum score and relatively few instances in lower score categories. Preliminary experiments with the 11-class setting resulted in unstable training and near-random macro-averaged  $F_1$  scores, indicating that the fine-grained formulation is severely affected by data sparsity and long-tailed label distribution.

To obtain more reliable and statistically meaningful results, we therefore adopt a collapsed 5-class SC setting, where adjacent score ranges are grouped into broader ordinal bins (0–2, 3–4, 5–6, 7–8, 9–10). Furthermore, due to the highly imbal-

anced language distribution in the scored subset and the very limited number of Kazakh and code-switched reviews with explicit ratings, the SC task is restricted to Russian-language reviews.

### 3.5 Data Partitioning

The data for both tasks were divided into training (Train), validation (Valid), and testing (Test) splits in an 80/10/10 ratio. To reduce topical leakage, splitting was performed at the movie level, so that all reviews of a given film appear in exactly one split. Table 3 reports the distribution of reviews across splits by sentiment label and language for the PC task. Table 4 reports the distribution of Russian reviews across splits by score bin for the SC task.

Category	Train		Valid		Test	
	#	%	#	%	#	%
<b>Sentiment</b>						
Negative	21,767	27.21	2,890	26.69	2,658	27.50
Neutral	3,388	4.23	502	4.64	439	4.54
Positive	54,854	68.56	7,437	68.68	6,567	67.95
<b>Language</b>						
cs	809	1.01	196	1.81	109	1.13
kk	2,178	2.72	230	2.12	231	2.39
ru	77,022	96.27	10,403	96.07	9,324	96.48
<b>Total</b>	<b>80,009</b>	<b>100</b>	<b>10,829</b>	<b>100</b>	<b>9,664</b>	<b>100</b>

Table 3: Distribution of polarity labels and languages across the sets for PC

## 4 Experiment

### 4.1 Sentiment Classification Models

For the evaluation of sentiment classification tasks, we employed a set of multilingual transformer-based models that support both Kazakh and Rus-

Score	Train		Valid		Test	
	#	%	#	%	#	%
1	839	9.56	83	8.09	117	10.71
2	877	9.99	116	11.31	95	8.70
3	1,349	15.37	135	13.16	152	13.92
4	1,560	17.78	197	19.20	186	17.03
5	4,150	47.29	495	48.25	542	49.63
<b>Total</b>	<b>8,775</b>	<b>100</b>	<b>1,026</b>	<b>100</b>	<b>1,092</b>	<b>100</b>

Table 4: Scores across the sets for SC

sian and are readily available through the Hugging Face Transformers framework (Wolf et al., 2019).

**mBERT**<sup>3</sup> is a multilingual BERT model (Devlin et al., 2019) pre-trained on Wikipedia in 100+ languages, including Kazakh and Russian, with a shared WordPiece vocabulary (168M parameters).

**XLM-RoBERTa**<sup>4</sup> is a multilingual RoBERTa model (Conneau et al., 2020; Liu et al., 2019) pre-trained on CC-100 CommonCrawl data (100 languages; 270M parameters).

**RemBERT**<sup>5</sup> is a rebalanced multilingual BERT variant (Chung et al., 2021) trained on Wikipedia and web data in 110+ languages, designed to improve performance on underrepresented languages.

Evaluating instruction-tuned generative models and stronger multilingual encoders is a natural next step. Here we focus on widely used pre-trained multilingual transformers to provide a stable, reproducible supervised baseline that is less sensitive to prompting and instruction-tuning choices.

In addition to transformer-based models, we evaluated several classical machine learning baselines, including linear support vector machine (SVM), logistic regression (LR), and multinomial naïve Bayes (MNB), using bag-of-words (BoW) and Term Frequency–Inverse Document Frequency (TF–IDF) feature representations, which remain strong and widely adopted baselines for text classification tasks (Salton and Buckley, 1988; Joachims, 1999).

## 4.2 Experimental Setup

### 4.2.1 Transformer-based models

All three transformer models were fine-tuned separately for the PC and SC tasks using the corresponding training splits, while hyperparameters were se-

<sup>3</sup><https://huggingface.co/google-bert/bert-base-multilingual-cased>

<sup>4</sup><https://huggingface.co/FacebookAI/xlm-roberta-base>

<sup>5</sup><https://huggingface.co/google/rembert>

lected based on performance on the respective validation sets. The final model configurations yielding the best validation macro-averaged  $F_1$  were subsequently evaluated on the held-out test sets.

Fine-tuning was conducted on Vast.AI<sup>6</sup> using a single NVIDIA RTX 3090 GPU (24 GB VRAM). The total computational cost of all fine-tuning runs was approximately \$2. The initial learning rate was set to  $2 \times 10^{-5}$  and the weight decay to 0.01. Training was terminated early when the validation loss showed a consistent increase across epochs, even in cases where the macro-averaged  $F_1$  exhibited marginal fluctuations, in order to mitigate potential overfitting and promote stable generalisation.

Across fine-tuning runs, we used a maximum of three epochs with early stopping based on validation loss. For PC, mBERT and RemBERT were trained for two epochs, while XLM-RoBERTa was trained for three epochs. For SC, mBERT and XLM-RoBERTa were trained for three epochs, and RemBERT for one epoch. We used a batch size of 150 for mBERT and XLM-RoBERTa on both tasks. Due to GPU memory constraints, RemBERT was fine-tuned with smaller batch sizes (20 for PC and 16 for SC).

### 4.2.2 Classical Baselines

For all classical models, text was represented using either sparse BoW or TF–IDF features with a maximum vocabulary size of 50,000 and n-gram ranges determined via validation-based hyperparameter tuning.

For the PC task, the best-performing configuration relied on BoW features with 1–3-gram representations and lowercasing enabled. Linear SVM was trained with  $C = 0.01$  and squared hinge loss, LR with  $C = 2.0$ ,  $\ell_2$  regularisation, and the saga solver, and MNB with  $\alpha = 0.01$  and disabled prior fitting. Class imbalance was addressed using class weights for SVM and LR.

For the SC task, different optimal configurations were observed. Linear SVM performed best with TF–IDF features and 1–3-gram representations ( $C = 0.1$ ), whereas LR and MNB achieved stronger results with BoW features using 1–2-gram and 1–3-gram representations, respectively. The optimal LR setup used  $C = 0.1$  with  $\ell_2$  regularisation and the saga solver, while MNB was configured with  $\alpha = 0.5$  and enabled prior fitting. As in the PC task, class weights were applied to discriminative models to mitigate class imbalance.

<sup>6</sup><https://vast.ai/>

Model	Valid					Test				
	A	P	R	F <sub>1</sub>	$\kappa$	A	P	R	F <sub>1</sub>	$\kappa$
<b>Polarity classification</b>										
LR	0.88	0.70	0.71	0.70	0.75	0.89	0.70	0.72	0.71	0.76
MNB	0.87	0.68	0.70	0.69	0.73	0.88	0.68	0.72	0.70	0.74
SVM	0.90	0.71	0.71	0.71	0.77	0.90	0.73	0.74	0.73	0.79
mBERT	0.92	0.78	0.70	0.72	0.82	0.93	0.81	0.72	0.74	0.83
RemBERT	0.94	0.82	0.79	0.81	0.87	0.95	0.84	0.81	0.82	0.88
XLM-RoBERTa	0.94	0.80	0.78	0.79	0.86	0.94	0.81	0.81	0.81	0.87
<b>Score classification</b>										
LR	0.63	0.51	0.52	0.52	0.47	0.65	0.54	0.54	0.54	0.50
MNB	0.67	0.55	0.54	0.54	0.52	0.68	0.55	0.54	0.54	0.52
SVM	0.66	0.54	0.54	0.53	0.51	0.69	0.56	0.55	0.55	0.53
mBERT	0.64	0.52	0.52	0.50	0.48	0.67	0.53	0.51	0.51	0.50
RemBERT	0.67	0.55	0.54	0.52	0.53	0.69	0.56	0.55	0.54	0.55
XLM-RoBERTa	0.66	0.51	0.55	0.52	0.52	0.68	0.55	0.56	0.54	0.54

Table 5: Performance of classical and transformer-based models on the PC and SC tasks

All hyperparameters were selected via grid search on the validation set, with macro-averaged  $F_1$  used as the primary optimisation criterion due to class imbalance.

#### 4.2.3 Score Masking

For the SC task, explicit rating expressions often appear in the review text and may cause label leakage. We therefore replaced all explicit score mentions (identified via manual inspection of the scored subset, focusing on direct numeric and conventional rating expressions) with a placeholder token (`scoretoken`). The placeholder is alphanumeric for BoW/TF-IDF compatibility; it was also added to transformer tokenisers (with resized embeddings) to prevent subword splitting.

#### 4.3 Sequence Length

To select an appropriate maximum input length for transformer models, we examined the token length distribution of the training, validation, and test splits after tokenisation with the respective model tokenisers (mBERT, XLM-RoBERTa, and RemBERT). The analysis showed that approximately 95–97% of reviews contain fewer than 256 tokens, depending on the tokeniser used. Only a small fraction of instances (about 2–5%) exceed this length and are consequently truncated when a maximum sequence length of 256 is applied.

Considering the relatively low proportion of longer reviews and the quadratic computational complexity of self-attention with respect to sequence length, we set the maximum input length to 256 tokens for all transformer-based models.

#### 4.4 Performance Metrics

We evaluate model performance using accuracy (A), precision (P), recall (R), macro-averaged  $F_1$ -score ( $F_1$ ), and Cohen’s kappa ( $\kappa$ ). Given the imbalanced class distributions in both PC and SC tasks, we treat macro-averaged  $F_1$ -score as the primary evaluation metric, as it assigns equal importance to all classes and offers a more balanced assessment than accuracy alone (Jurafsky and Martin, 2009; Yang, 2001). In addition, Cohen’s  $\kappa$  is reported to measure the agreement between model predictions and gold labels while accounting for chance agreement.

#### 4.5 Experiment Results

We report results for all models in Table 5. In general, transformer-based encoders still achieve the highest scores on PC, but the gap between them and classical methods narrows after masking explicit score mentions in the SC task. Since RemBERT performs consistently well across both PC and SC, we use it as the main reference system for per-class, per-language, and qualitative error analyses below.

##### 4.5.1 Polarity Classification

On the PC task, the transformer models continue to set the standard. The test set macro-averaged  $F_1$  scores for RemBERT and XLM-RoBERTa are 0.82 and 0.81, respectively. mBERT lags slightly behind with an  $F_1$  of 0.74. Among the classical baselines, SVM and LR attain  $F_1$  scores of 0.73 and 0.71, respectively, while MNB reaches 0.70. These results indicate that, although simple BoW

models remain strong baselines, contextualised encoders offer a consistent advantage for this three-way classification task.

Per-class analysis for RemBERT in Table 6 shows uniformly high precision and recall for positive and negative reviews ( $F_1 = 0.97$  and  $0.94$ ), but markedly lower scores for the neutral class ( $F_1 = 0.56$ ). The difficulty of detecting neutral sentiment is unsurprising given that neutral instances comprise only about 4–5% of the corpus and often contain ambiguous language. The per-language breakdown in Table 7 also confirms that performance correlates with data volume: Russian reviews dominate the dataset and yield the highest  $F_1$  ( $0.81$ ,  $\kappa = 0.88$ ), whereas Kazakh ( $0.77$ ,  $\kappa = 0.68$ ) and code-switched ( $0.76$ ,  $\kappa = 0.73$ ) reviews achieve slightly lower but still respectable results.

Sentiment	P	R	$F_1$
Negative	0.94	0.94	0.94
Neutral	0.60	0.52	0.56
Positive	0.97	0.98	0.97

Table 6: Per-class results of RemBERT on the PC test set

Language	A	P	R	$F_1$	$\kappa$
cs	0.87	0.87	0.74	0.76	0.73
kk	0.89	0.89	0.73	0.77	0.68
ru	0.94	0.82	0.80	0.81	0.88

Table 7: Per-language results of RemBERT on the PC test set

Qualitative inspection of model errors suggests that most confusions involve borderline or implicitly neutral statements. For instance, the Russian review “*Неожиданный фильм для меня. Совершенно не похож на воспоминания из детства об индийском кино. Куда делись танцы?!?*” [*Neozhidanny fil'm dlya menya. Sovershennno ne pokhozhd na vospominaniya iz detstva ob indiyском кино. Kuda delis' tantsy?!?*]; “The film was unexpected for me. It is completely unlike my childhood memories of Indian cinema. Where did the dances go?!”] was labelled `Neutral` but predicted as `Negative`. Here, the author primarily expresses surprise and a mismatch with genre expectations (the absence of song-and-dance elements typical of Indian cinema), which can be interpreted as either mild criticism or a neutral observation; the model appears to treat the rhetorical question as negative sentiment.

A complementary error appears in a Kazakh review that combines obligation framing with a balanced appraisal: “*Озимиздин Қазақтар тусиргесин қолдап баруға тура келеді, бірақ отиниш режиссёр сценаристы кишкене қарау керек ау, актерлер өте жақсы ойнап шықты.*” [*Ozimizding qazaqtar tüsirgesin qoldap baruға тура келеді, бірақ ötinish, rezhissior scenaristi kishkentai qarau kerek аu, aktiorler öte zhaqsy oinap shyqty*]; “We have to go and support films made by our Kazakhs, but the director and screenwriter really should improve a bit; the actors performed very well.”]. Although the gold label is `Neutral` due to the combination of critique and praise, the model predicts `Positive`, likely overweighting explicit positive cues such as “*өте жақсы*” (“very good”) while underweighting the qualifying criticism. These cases illustrate that neutral reviews in this corpus often encode evaluation indirectly via expectations, rhetorical framing, or mixed praise and critique, making them particularly prone to polarity drift in automatic classification.

#### 4.5.2 Score Classification

The SC task becomes considerably more challenging under the leakage-controlled setting with masked score mentions. Across all models, macro-averaged  $F_1$  values fall into the 0.50–0.55 range, suggesting that predicting 5-level ratings from text alone is a much harder problem than PC. RemBERT and SVM achieve the highest test-set  $F_1$  of 0.54–0.55, followed closely by XLM-RoBERTa and MNB (0.54). mBERT is the weakest among the transformers with an  $F_1$  of 0.51. The similar performance of classical and transformer models indicates that the masking procedure effectively removed many of the lexical cues that deep models previously exploited, forcing them to rely on more subtle sentiment features in the text.

Per-score analysis in Table 8 highlights the extreme class imbalance. RemBERT performs well on the highest rating (score 5) with  $P = 0.87$ ,  $R = 0.89$ , and  $F_1 = 0.88$ , but struggles on intermediate ratings; the  $F_1$  for score 2, for example, is just 0.09 due to very few examples. This skew explains why overall accuracy is relatively high (69%) while macro-averaged metrics remain modest.

## 5 Discussion

The experiments provide two main takeaways. First, multilingual transformer encoders offer a

Score	P	R	F <sub>1</sub>
1	0.72	0.65	0.68
2	0.24	0.05	0.09
3	0.44	0.67	0.53
4	0.53	0.51	0.52
5	0.87	0.89	0.88

Table 8: Per-class results of RemBERT on the SC test set

consistent advantage for PC on this corpus, but the margin over strong linear baselines is modest. The gap is most visible in the minority neutral class, where language is often ambiguous and underrepresented, while positive and negative sentiment are detected reliably by all models. This suggests that much of the polarity signal in movie reviews can be captured by surface lexical cues, yet contextual modelling remains beneficial for borderline cases and for improving robustness under class imbalance.

Second, leakage-controlled SC remains substantially more difficult than PC. After masking explicit rating mentions with `scoretoken`, performance converges across model families and drops to modest macro-averaged F<sub>1</sub> values, indicating that fine-grained rating inference depends on subtle and often implicit cues that are harder to learn than coarse polarity. The per-class behaviour further shows that the models perform well on the most frequent high-score bin but struggle on intermediate bins, reflecting both severe label imbalance and the inherently ordinal nature of ratings, where adjacent categories may be expressed with very similar language.

These findings highlight several limitations and directions for future work. The score distribution is strongly skewed toward favourable ratings, and even after collapsing to five bins, the mid-range classes remain sparse. More data in the lower and mid ranges, targeted rebalancing, or modelling approaches that explicitly account for ordinality (e.g., ordinal regression or regression-based formulations) may yield more stable improvements. In addition, fully eliminating score leakage is non-trivial because rating information can be expressed indirectly or idiomatically, and manual score extraction occasionally involves ambiguous cases; consequently, some residual noise in the score labels is likely. Finally, while the polarity results on Kazakh and code-switched reviews are encouraging, stronger conclusions about multilingual gener-

alisation will require more balanced language coverage or dedicated evaluation subsets.

## 6 Conclusion

We introduced a new publicly available corpus of 100,502 movie reviews from Kazakhstan collected from `kino.kz`, spanning 2001–2025 and covering 4,943 titles, with Russian, Kazakh, and code-switched texts. Reviews were manually annotated for language and sentiment polarity, and 11,309 reviews additionally contain explicit user-provided ratings, enabling fine-grained sentiment modelling.

We defined two supervised tasks: three-way polarity classification and five-class score classification. On polarity classification, multilingual transformer encoders achieved the best results, with RemBERT performing strongest (macro-averaged F<sub>1</sub> = 0.82,  $\kappa$  = 0.88 on the test set). For score classification, we evaluated a leakage-controlled setting by masking explicit score mentions; under this setup, all models achieved modest macro-averaged F<sub>1</sub> scores (0.51–0.55), highlighting the difficulty of inferring rating levels from text alone under severe class imbalance.

To support reproducibility, we release the dataset, annotation guidelines, and trained models on the Hugging Face Hub.<sup>7</sup>

## References

- PD Blinov, Maria Klekovkina, Eugeny Kotelnikov, and Oleg Pestov. 2013. Research of lexical approach and machine learning methods for sentiment analysis. *Computational Linguistics and Intellectual Technologies*, 2(12):48–58.
- Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. [Re-thinking Embedding Coupling in Pre-trained Language Models](#). In *International Conference on Learning Representations*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised Cross-lingual Representation Learning at Scale](#). In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 8440–8451. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

<sup>7</sup>[https://huggingface.co/datasets/yeshpanovrustem/100k\\_movie\\_reviews\\_from\\_kz](https://huggingface.co/datasets/yeshpanovrustem/100k_movie_reviews_from_kz)

- Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Thorsten Joachims. 1999. **Text Categorization with Support Vector Machines: Learning with Many Relevant Features**. In *European Conference on Machine Learning*.
- Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.
- Andrew Maas. 2011. Large Movie Review Dataset. <https://ai.stanford.edu/~amaas/data/sentiment/>. Accessed 2026-02-26.
- Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. **Learning Word Vectors for Sentiment Analysis**. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.
- Bo Pang and Lillian Lee. 2008. **Opinion Mining and Sentiment Analysis**. *Found. Trends Inf. Retr.*, 2(1–2):1–135.
- Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. **Thumbs up? Sentiment Classification using Machine Learning Techniques**. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 79–86. Association for Computational Linguistics.
- Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Björn Gambäck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. 2020. **SemEval-2020 Task 9: Overview of Sentiment Analysis of Code-Mixed Tweets**. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 774–790. Association for Computational Linguistics.
- Anna Rogers, Alexey Romanov, Anna Rumshisky, Svitlana Volkova, Mikhail Gronas, and Alex Gribov. 2018. **RuSentiment: An Enriched Sentiment Analysis Dataset for Social Media in Russian**. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 755–763. Association for Computational Linguistics.
- Gerard Salton and Chris Buckley. 1988. **Term-Weighting Approaches in Automatic Text Retrieval**. *Inf. Process. Manag.*, 24:513–523.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. **Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank**. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 1631–1642. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2019. **Transformers: State-of-the-Art Natural Language Processing**. In *Conference on Empirical Methods in Natural Language Processing*.
- Yiming Yang. 2001. A Study on Thresholding Strategies for Text Categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 137–145.
- Rustem Yeshpanov and Huseyin Atakan Varol. 2024. **KazSanDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes**. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9657–9667. ELRA Language Resource Association.