KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes

Rustem Yeshpanov, Huseyin Atakan Varol
Institute of Smart Systems and Artificial Intelligence
Nazarbayev University, Astana, Kazakhstan
{rustem.yeshpanov, ahvarol}@nu.edu.kz

Abstract

This paper presents KazSAnDRA, a dataset developed for Kazakh sentiment analysis that is the first and largest publicly available dataset of its kind. KazSAnDRA comprises an extensive collection of 180,064 reviews obtained from various sources and includes numerical ratings ranging from 1 to 5, providing a quantitative representation of customer attitudes. The study also pursued the automation of Kazakh sentiment classification through the development and evaluation of four machine learning models trained for both polarity classification and score classification. Experimental analysis included evaluation of the results considering both balanced and imbalanced scenarios. The most successful model attained an F_1-score of 0.81 for polarity classification and 0.39 for score classification on the test sets. The dataset and fine-tuned models are open access and available for download under the Creative Commons Attribution 4.0 International License (CC BY 4.0) through our GitHub repository.

Keywords: BERT, dataset, Kazakh, KazSAnDRA, polarity, review, sentiment analysis, text classification

1. Introduction

In natural language processing, sentiment analysis is a widely employed text classification task that involves extracting the sentiment expressed by individuals towards a variety of entities that include products, services, organisations, individuals, issues, events, and topics together with their respective attributes (Liu, 2012). In this context, sentiment represents the positive, negative, or neutral attitude of individuals conveyed through the extracted textual content (Jurafsky and Martin, 2009). Sentiment analysis demonstrates broad applicability across various domains, including marketing (Fang and Zhan, 2015), social media (Go et al., 2009), healthcare (Greaves et al., 2013), finance (Abraham et al., 2018), and politics (Abercrombie and Batista-Navarro, 2020), among others.

Although research efforts in sentiment analysis for lower-resourced languages are gradually gaining momentum (Mamta et al., 2022; Le et al., 2016; Gangula and Mamidi, 2018), the English language continues to dominate as the primary focus of current research in this area (Zhang et al., 2018). This preference can be attributed to the abundant availability of linguistic resources, such as lexica, corpora, and dictionaries specifically tailored to English (Medhat et al., 2014).

With respect to Kazakh, an agglutinative Turkic language generally considered lower-resourced, research in the field of sentiment analysis has only recently come to the fore (Narynov and Zharmagambetov, 2016). Despite its importance, the literature dealing with sentiment analysis in Kazakh remains limited and includes only a few academic papers published within eight years. Furthermore, there is a complete absence of publicly accessible Kazakh sentiment analysis datasets, whether small or large, further underscoring the challenges in this field.

Our study aims to address the existing gaps in this field and contribute to its further advancement. Specifically, we present a dataset consisting of customer reviews in Kazakh, accompanied by corresponding attitude scores. The dataset comprises a total of 180,064 reviews collected from four domains.

In the context of Kazakhstan, it is crucial to acknowledge the prevalent practice of code-switching between the Kazakh and Russian languages, as well as the ongoing shift from the Cyrillic to the Latin script. Consequently, Kazakh reviews may exhibit a combination of Cyrillic and Latin characters, incorporate a mixture of Russian and Kazakh vocabulary, or be solely recorded in the Cyrillic script with Russian characters substituting Kazakh ones. The dataset we present includes reviews containing both exclusive Kazakh vocabulary and words from other languages (Russian, English, and Arabic), making it the largest dataset available for Kazakh sentiment analysis.

We also developed and evaluated four machine learning models to automate the classification of Kazakh sentiments. The highest F_1-score on the test sets was 0.81 for polarity classification and 0.39 for score classification.

The subsequent sections of this paper are structured as follows: Section 2 presents a review of existing research in Kazakh sentiment analysis. Section 3 is devoted to the detailing the process of developing the dataset. Section 4 delves into the aspects of data pre-processing and partitioning, the score resampling techniques, the sentiment classification tasks, the models employed, the experimental design, and the metrics used for evaluation, and the corresponding results. Section 5 focuses on a thorough discussion of the results.

Section 6 provides a conclusive summary and final remarks for the paper.

2. Related Work

In recent years, remarkable progress has been made in addressing the limited resources available for the Kazakh language. Mussakhojayeva et al. have made significant contributions to this endeavour by presenting a text-to-speech synthesis corpus comprising a substantial 271 hours of speech data (Mussakhojayeva et al., 2022b), as well as introducing the first industrial-scale open-source Kazakh speech corpus for automatic speech recognition (Mussakhojayeva et al., 2022a). The latter corpus consists of 1,128 hours of transcribed audio data, comprising over 520 thousand utterances. In a separate study, Yeshpanov et al. (2022) made a noteworthy contribution to the field of Kazakh natural language processing by introducing the largest dataset for Kazakh named entity recognition. This dataset comprises 112,702 sentences and 136,333 annotations for 25 distinct entity classes. In addition, Toiganbayeva et al. (2022) proposed an extensive dataset specifically tailored for handwritten text recognition in Kazakh. This dataset includes 3,000 handwritten exam papers, with 140,335 segmented images and 922,010 symbols. While the collective contributions to various speech and language processing tasks for Kazakh have undeniably enriched the available resources, thus creating new opportunities for research and development, progress in Kazakh sentiment analysis research has been comparatively slower. This discrepancy can be attributed to the limited availability of dedicated resources in this area.

In the earliest work found on Kazakh sentiment analysis (Narynov and Zharmagambetov, 2016), the researchers curated a dataset of 30,000 Russian news articles that were manually labelled. In addition, they labelled 10,000 Kazakh news articles, of which 3,021 were positive, 2,548 negative, and 4,431 neutral, to train a sentiment classifier. While the performance of the classifier for Russian was 86.3%, it yielded relatively lower results for Kazakh, with an accuracy of 72.8%. This result could possibly be attributed to the limited size of the training dataset and the absence of lemmatisation, which may have affected the overall performance.

In Abdullin and Ivanov (2017), the research aimed to analyse opinions in short texts written in several languages, with a specific focus on English, Russian, and Kazakh (Go et al., 2009; Rubtsova and Zagorulko, 2014). The authors presented an approach that utilised a deep recurrent neural network and bilingual word embeddings to effectively capture semantic features between words across these languages. By conducting sentiment analysis experiments on language pairs such as English-Russian and Russian-Kazakh, the authors achieved noteworthy performance, with a competitive accuracy rate of 72% for Russian and a comparatively lower accuracy of 58% for Kazakh. While the authors made the development codes available on their GitHub page for result reproducibility, it is worth noting that the repository does not include the Kazakh training data that were utilised.

In Yergesh et al. (2017), reviews of three hotels were collected from online travel platforms. The authors employed fuzzy logic (Zadeh, 1996) for sentiment analysis. However, it is important to acknowledge that the study does not provide precise information about the number of reviews collected and the accessibility of the data collected. In their later study (Yergesh et al., 2019), the researchers presented an overview of the rule-based methods employed in sentiment analysis and the approaches utilised to determine the sentiment of Kazakh sentences by formalising morphological rules. To facilitate the determination of text polarity, a dictionary of approximately 11,000 emotional Kazakh words and phrases was manually compiled and annotated on a five-point scale [-2, 2]. In addition, semantic hypergraphs were used to describe ontological models of the morphological rules of the Kazakh language. As a result of this research, a morphological analyser was developed, enabling the extraction of morphological information from texts. The authors reported that they achieved a result of 83% using their rule-based method, which they describe as "good", although it is interesting to note that this result was peculiarly compared to studies dedicated to sentiment analysis in Russian.

Bekmanova et al. (2019) developed a method for analysing Kazakh texts related to terrorist threats. The research involved selecting social networks along with specific foreign Internet resources that disseminate terrorist content. Through this selection, a database was developed that comprised 1,200 entries. This database facilitated the detection of more than 50 similar sites. It is important to note, however, that access to the database is not possible.

In Mutanov et al. (2021), a dataset of news posts from the Kazakhstani media space was collected, comprising texts labelled across three sentiment classes: positive, negative, and neutral. The dataset includes 80,873 sentiment-labelled texts in Russian and 15,933 in Kazakh, revised by graduate students specialising in political science. While the paper describes in detail the steps of data pre-processing and the classification methods employed, it does not delve into the approach taken for news posts exhibiting multiple polarities (e.g., "Controversial Policy Sparks Outrage and Support Among Citizens"), nor does it shed light on the provision of classification guidelines or the inter-annotator agreement. Additionally, while detailed classification metrics and confusion matrices are presented for Russian texts, analysis for Kazakh texts is notably absent.

In Gimadi (2021), the aim of the study was to collect a dataset of 3,000 Kazakh reviews from the Google Play

Store. However, the rationale behind the researcher's decision to manually label each of the collected reviews and subsequently compare the assigned scores with the original scores remains unclear.

More recently, Rakhymzhanov (2022) attempted to develop a Kazakh slang dictionary using a website[1] and its associated Instagram page. The researcher utilised BeautifulSoup[2] to collect slang word information, but due to the recent creation of the website, data availability was limited. This study breaks new ground in Kazakh sentiment analysis, and while the referenced slang dictionary provides amusing expressions, its practical usefulness as a reliable reference source can be questioned due to their infrequent use among native Kazakh speakers.

Nurlybayeva et al. (2022) presented the construction of a bag-of-words model (Zhang et al., 2010) for sentiment classification of restaurant reviews into positive or negative categories. The researchers collected a dataset of 2,000 restaurant reviews from the 2GIS application website[3] for analysis and model development. However, it should be noted that the dataset used in this study is not publicly accessible.

The last study on Kazakh sentiment analysis that we discuss in this paper is by Nugumanova et al. (2022), who applied pre-trained BERT (Devlin et al., 2019) models, originally developed for multilingual and Turkish sentiment analysis, to Kazakh due to the lack of large labelled datasets in this language. The Kazakh dataset was collected from Facebook groups, a consumer complaints website[4], and the 2GIS website, with the reviews manually labelled and transliterated. The training dataset was very small (only 30 samples), but the experimental results showed that the multilingual BERT model outperformed the Turkish BERT model.

From the literature reviewed, it follows that while substantial efforts have been made in the fields of automatic speech recognition, text-to-speech synthesis, image-to-text conversion, and named entity recognition for Kazakh, resulting in extensive, high-quality and publicly available datasets, the same level of generosity seems to be lacking when it comes to research specifically focused on Kazakh sentiment analysis. In other words, the availability of sentiment analysis datasets for Kazakh is currently non-existent or significantly limited.

3. Dataset Development

3.1. Domains

The source data for our dataset came from four domains: (1) digital mapping and navigation services (hereafter Mapping), (2) online marketplaces (hereafter Market), (3) an online library that serves as a source of books and audiobooks in Kazakh (hereafter Bookstore), and (4) an online store for Android devices that offers a diverse range of applications (hereafter Appstore).

3.2. Data Collection

The dataset was collected over a one-year period, from September 2022 to September 2023. Reviews from Mapping and Market were collected through manual means, while a *BeautifulSoup* script was employed for the acquisition of reviews from Bookstore. The use of the Python package *google-play-scraper*[5] facilitated the collection of reviews from Appstore.

All reviews were manually checked by a group of native Kazakh speakers. During the review process, it came to light that reviews contained recurrent instances of inappropriate content. However, in order to preserve the authenticity and integrity of the reviews, no alterations or removals were made to this content.

As a result, 8,897 reviews were obtained from Mapping, covering 407 institutions. Market provided a considerable portion of 30,289 reviews, encompassing 8,418 unique items. Bookstore provided 5,805 reviews, comprising 3,792 audiobook reviews and 2,013 book reviews, resulting in a total of 1,026 unique audiobooks and books. Finally, Appstore provided 135,073 unique reviews of 1,759 Android applications and games. Of the users contributing to these reviews, 47,887 had a unique username, while the remaining 31,490 users remained anonymous.

Each review was accompanied by a numerical rating from 1 to 5, providing a measurable representation of individuals' attitudes. Consequently, we named the dataset KazSAnDRA /kəˈsændrə/, an acronym for the **Kaz**akh **S**entiment **An**alysis **D**ataset of **R**eviews and **A**ttitudes, reflecting its purpose and content. The total number of reviews collected was 180,064. Table 1 provides information about the distribution of reviews across different scores and domains.

Domain	Score					Total
	1	2	3	4	5	
Appstore	22,547	4,202	5,758	7,949	94,617	**135,073**
Bookstore	686	107	222	368	4,422	**5,805**
Mapping	959	270	369	525	6,774	**8,897**
Market	1,043	350	913	2,775	25,208	**30,289**
Total	**25,235**	**4,929**	**7,262**	**11,617**	**131,021**	**180,064**

Table 1: Domain and score statistics

3.3. Variations of Kazakh Reviews

In Kazakhstan, code-switching between the Kazakh and Russian languages has been observed (Pavlenko, 2008), alongside an ongoing shift from the Cyrillic to the Latin script. Consequently, reviews regarded as Kazakh can

[1] https://janasozdik.kz

[2] https://www.crummy.com/software/BeautifulSoup

[3] https://www.2gis.kz

[4] https://zhalobikz.com

[5] https://pypi.org/project/google-play-scraper

take different forms: (a) solely Kazakh words written in the Kazakh Cyrillic script, (b) Kazakh words written in Latin script, (c) a combination of Cyrillic and Latin characters, (d) a mixture of Russian and Kazakh words, or (e) a text entirely in the Cyrillic script with Russian characters replacing Kazakh characters, among other possible variants. Table 3 provides examples of actual reviews, demonstrating their appropriate representation in accordance with Kazakh spelling rules and the use of the Kazakh Cyrillic script, accompanied by their correct form in English. Table 2 shows the number of reviews with percentage of Cyrillic and Latin characters per review.

Character	0–25%	26–50%	51–75%	76–100%
Cyrillic	67	399	1,694	170,233
Latin	5,374	1,114	246	2,617

Table 2: Review counts by Cyrillic and Latin character percentages

3.4. Sentiment Classification Tasks

To evaluate the effectiveness of KazSAnDRA, we utilised the dataset for two tasks: (a) polarity classification (PC), which involves predicting whether a review is positive or negative, and (b) score classification (SC), which involves predicting the score of a review on a scale of 1 to 5. In the PC task only, reviews with an original score of 1 or 2 were designated as negative and subsequently assigned a new score of 0. In contrast, reviews with an original score of 4 and 5 were classified as positive and assigned a new score of 1. Reviews with an original score of 3 were categorised as neutral and excluded from the task.

3.5. Data Pre-Processing

Irrespective of the task for which the dataset was intended, the data pre-processing stage involved several essential steps aimed at preserving the integrity and uniformity of the dataset. First, all emojis were systematically removed from the text to eliminate potential noise. Subsequently, to ensure consistency and ease of analysis, all reviews were lowercased. The elimination of punctuation helped to streamline the text for further processing. In addition, the characters for line break (\n), tab (\t), and carriage return (\r) were removed to avoid interference with subsequent computations. To enhance readability and minimise unnecessary mismatches, multiple spaces were uniformly replaced with a single space. Furthermore, it is important to note that in the Kazakh language, consecutive characters are allowed to occur in pairs (e.g., *айттым* "I said", *құжаттар* "documents", *қосса* "if s/he adds") but not in larger clusters. Hence, to conform to this linguistic feature, any consecutive characters that appeared repeatedly were reduced to two instances (e.g., *кеерррееммееетт* to *ке-ррееммееетт*). Lastly, all duplicate entries, defined as reviews sharing identical text and scores, were removed. The total numbers of reviews following pre-processing were 167,961 for PC and 175,158 for SC.

3.6. Data Partitioning

To ensure consistency and reproducibility of our experimental results across different research groups, KaZSAnDRA was divided into training (Train), validation (Valid), and test (Test) sets, maintaining a ratio of 80/10/10. The division ensured a balanced and proportional distribution of review scores and domains across the sets.

Tables 4 and 5 present the distribution of reviews across the three sets based on the domains for the PC and SC tasks, respectively, after pre-processing, including the total number of reviews as well as the respective counts for each set pertaining to both tasks. Tables 6 and 7 present the distribution of reviews across the three sets in terms of their scores for the PC and SC tasks, in turn.

Furthermore, an analysis of the KazSAnDRA dataset was conducted to extract unique reviews (i.e., reviews with distinct textual content) per classification task. The numbers of unique reviews in the training, validation, and test sets for PC were 132,152, 16,739, and 16,757, respectively. For SC, the corresponding counts were 137,365, 17,464, and 17,445. The intersection between the training, validation, and test sets was then computed, as depicted in Figure 1. Significantly, over 96% of the unique reviews present in the test sets for both PC and SC tasks did not occur in either the training or validation sets. This substantial discrepancy corroborated the appropriateness of utilising the test sets to effectively evaluate the generalisation capabilities of sentiment classification models.

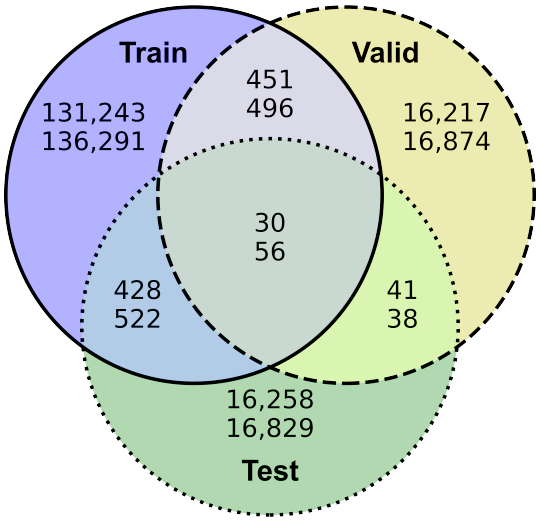

Figure 1: Unique reviews across the sets for PC (top) and SC (bottom)

	Actual review	Correct form (Kazakh)	Correct form (English)
a	*керемет кітап*	*керемет кітап*	*a wonderful book*
b	*keremet*	*керемет*	*wonderful*
c	*jok кітап*	*кітап жоқ*	*no books*
d	*Осы приложениеге көп рахмет!*	*Осы қолданбаға көп рақмет!*	*Many thanks to this app!*
e	*Кушти!*	*Күшті!*	*Great!*

Table 3: Kazakh review variations

Domain	Train		Valid		Test	
	#	%	#	%	#	%
Appstore	101,477	75.52	12,685	75.52	12,685	75.52
Market	22,561	16.79	2,820	16.79	2,820	16.79
Mapping	6,509	4.84	813	4.84	814	4.85
Bookstore	3,821	2.84	478	2.85	478	2.85
Total	**134,368**	**100**	**16,796**	**100**	**16,797**	**100**

Table 4: Domains across the sets for PC

Domain	Train		Valid		Test	
	#	%	#	%	#	%
Appstore	106,058	75.69	13,258	75.69	13,257	75.69
Market	23,278	16.61	2,909	16.61	2,910	16.61
Mapping	6,794	4.85	849	4.85	849	4.85
Bookstore	3,996	2.85	500	2.85	500	2.85
Total	**140,126**	**100**	**17,516**	**100**	**17,516**	**100**

Table 5: Domains across the sets for SC

Score	Train		Valid		Test	
	#	%	#	%	#	%
1	110,417	82.18	13,801	82.17	13,804	82.18
0	23,951	17.82	2,995	17.83	2,993	17.82
Total	**134,368**	**100**	**16,796**	**100**	**16,797**	**100**

Table 6: Scores across the sets for PC

Score	Train		Valid		Test	
	#	%	#	%	#	%
5	101,302	72.29	12,663	72.29	12,663	72.29
1	20,031	14.29	2,504	14.30	2,504	14.30
4	9,115	6.50	1,140	6.51	1,139	6.50
3	5,758	4.11	719	4.10	720	4.11
2	3,920	2.80	490	2.80	490	2.80
Total	**140,126**	**100**	**17,516**	**100**	**17,516**	**100**

Table 7: Scores across the sets for SC

3.7. Score Resampling

Table 1 shows the score distribution in KazSAnDRA, indicating a significant imbalance. This raises concerns about biassed model performance, favouring the majority scores and neglecting underrepresented scores. Therefore, our study included analysis of results obtained from both balanced and imbalanced data, which had previously been employed with varying degrees of success (see Burns et al., 2011; Mutanov et al., 2021).

In response to the data imbalance in our training data, we employed random oversampling (ROS) and random undersampling (RUS) to balance the representation of classes by creating new samples for the smaller class to align with the majority class count and eliminating samples from the larger class to match the minority class count, respectively (Ramentol et al., 2012). In this study, we deferred the investigation of alternative approaches (e.g., data augmentation through back-translation) for future research. The resulting training sets for both tasks are detailed in Tables 8 and 9.

Score	Balanced		Imbalanced
	ROS	RUS	
0	110,417	23,951	23,951
1	110,417	23,951	110,417

Table 8: Training sets for PC

Score	Balanced		Imbalanced
	ROS	RUS	
1	101,302	3,920	20,031
2	101,302	3,920	3,920
3	101,302	3,920	5,758
4	101,302	3,920	9,115
5	101,302	3,920	101,302

Table 9: Training sets for SC

3.8. Dataset Organisation

The dataset comprises ten CSV files. Files "01" to "05" are associated with PC, while files "06" to "10" are related to SC. Different training set variations are indicated by the suffixes "ib" for imbalanced data, "ros" for random oversampling, and "rus" for random undersampling. Each file includes records containing a custom review identifier (`custom_id`), the original review text (`text`), the pre-processed review text (`text_cleaned`), the corresponding review score

(`label`), and the domain information (`domain`). The dataset can be conveniently downloaded from our GitHub repository.[6]

4. Experiment

4.1. Sentiment Classification Models

For the evaluation of KazSAnDRA, we utilised four multilingual machine learning models, all incorporating the Kazakh language and accessible through the Hugging Face Transformers framework (Wolf et al., 2020). The framework was chosen for its state-of-the-art pre-trained models, user-friendly interface, and collaborative ecosystem. Details on the implementation of the sentiment classification model are available on our GitHub repository.[6]

mBERT is a case-insensitive variant of the multilingual BERT (Devlin et al., 2019) model. This model comprises about 168 million parameters and has been pre-trained on a corpus of 102 languages.

XLM-R We leveraged the XLM-RoBERTa model (Conneau et al., 2020), a multilingual variant of RoBERTa (Liu et al., 2019). The rationale for selecting this model stems from its substantial parameter count, exceeding that of BERT by over fivefold (561M), and its pre-training on the CommonCrawl corpus encompassing 100 languages.

RemBERT The rebalanced multilingual BERT model (Chung et al., 2021) is a multilingual encoder pre-trained on Wikipedia over 104 languages. RemBERT exhibits superior performance compared to the similarly sized XLM-R, despite being trained on 3.5 times fewer tokens.

mBART-50 (Tang et al., 2020) is a multilingual sequence-to-sequence model built on the foundation of the original mBART model (Liu et al., 2020). This extended version was thoughtfully augmented with an additional 25 languages, bringing the total number of languages supported to 50.

4.2. Experimental Setup

All four models were fine-tuned using both the balanced and imbalanced training sets, while the hyperparameters were refined using the validation set. The final and most optimal models were evaluated on the test sets. The fine-tuning of the models was executed on a single A100 GPU hosted on an NVIDIA DGX A100 machine. The initial learning rate was set at 10^{-5}; the weight decay rate was set at 10^{-3}. Early stopping was employed, executed when the F_1-score exhibited no improvement for three consecutive epochs. We set the batch size to 32 (mBERT, XLM-R, RemBERT) or 16 (mBART-50) and applied 800 warm-up steps.

4.3. Performance Metrics

Several conventional metrics were used to evaluate the performance of the models, including accuracy (A), precision (P), recall (R), and F_1-score (F_1). Given the imbalanced nature of the dataset, where all classes carry equal importance, we opted for macro-averaging, calculated from the arithmetic (i.e., unweighted) mean of all F_1-scores per class, and thus ensuring equal treatment of all classes during the evaluation, resulting in a stronger penalty if the model performs worse on minority classes (Jurafsky and Martin, 2009; Yang, 2001).

4.4. Experiment Results

Table 10 presents the performance of the four models on KazSAnDRA test sets for the PC and SC tasks. XLM-R and RemBERT consistently outperformed mBERT and mBART-50 across various training scenarios. The highest F_1-scores of 0.81 (PC) and 0.39 (SC) were achieved by both XLM-R and RemBERT in the imbalanced training scenario. In five out of six training scenarios, RemBERT achieved the highest F_1-scores, while XLM-R led in four. Table 11 presents data on the number of epochs required to train models for the PC and SC tasks, considering both balanced (ROS, RUS) and imbalanced (IB) training data scenarios. Tables 12–14 summarise the performance of RemBERT in the imbalanced training context, with a detailed analysis following in the subsequent section.

5. Discussion

Scores in the SC task were lower than in the PC task for all models, possibly due to its increased complexity involving five-way classification. Table 12 shows that, in the PC task, the RemBERT model had stronger performance in identifying positive reviews, but had a notable drawback misclassifying 1,036 positive reviews as negative, indicating a relatively high number of false negatives.

In the SC task (Table 13), it appears that the model had higher accuracy in identifying reviews with scores at the extremes (1 and 5) compared to the middle scores (2, 3, and 4). The model had particularly low accuracy in identifying reviews with a score of 2, with only 55 true positives. The reason for this is most likely that the training data contained many more reviews with scores of 1 and 5 than the middle scores (see Table 1).

In addition, the model exhibited a tendency for misclassifying reviews with scores other than 5 as if they had a rating of 5. This also seems to be related to the preponderance of reviews with a score of 5, causing the model to have a bias towards this score. The substantial disparity between the number of positive and negative reviews can be attributed to the fact that the reviewed items predominantly represent highly popular or top-rated selections. It is therefore to be expected that such items naturally receive a significantly higher number of reviews than less popular or lower-rated items (Aly and Atiya, 2013).

In Table 14, an interesting observation is that the model exhibited more accurate classification in both tasks for

[6] https://github.com/IS2AI/KazSAnDRA

Model	Polarity Classification												Score Classification											
	Balanced (ROS)				Balanced (RUS)				Imbalanced				Balanced (ROS)				Balanced (RUS)				Imbalanced			
	A	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A	P	R	F_1	A	P	R	F_1
mBERT	0.84	0.74	0.83	0.77	0.85	0.76	0.82	0.78	0.89	0.82	0.79	0.80	0.67	0.34	0.36	0.35	0.63	0.35	0.39	0.36	0.77	0.44	0.36	0.37
XLM-R	0.86	0.76	0.83	0.79	0.85	0.75	0.83	0.78	**0.89**	**0.81**	**0.81**	**0.81**	0.58	0.36	0.42	0.36	0.66	0.36	0.41	0.37	**0.77**	**0.42**	**0.37**	**0.39**
RemBERT	0.88	0.79	0.82	0.81	0.87	0.78	0.82	0.80	**0.89**	**0.81**	**0.82**	**0.81**	0.73	0.37	0.36	0.36	0.62	0.35	0.40	0.35	**0.76**	**0.41**	**0.38**	**0.39**
mBART-50	0.87	0.77	0.79	0.78	0.81	0.72	0.81	0.74	0.89	0.82	0.78	0.80	0.74	0.36	0.34	0.35	0.55	0.36	0.41	0.34	0.77	0.42	0.37	0.38

Table 10: PC and SC results on the test sets

Model	PC			SC		
	ROS	RUS	IB	ROS	RUS	IB
mBERT	4	7	6	8	10	11
XLM-R	5	7	5	4	9	16
RemBERT	4	5	5	6	6	9
mBART-50	5	7	5	8	7	5

Table 11: Number of training epochs for models

Polarity Classification			
predicted → actual ↓	**0**	**1**	**Total**
0	2,155	838	2,993
1	1,036	12,768	13,804

Table 12: RemBERT PC results

Score Classification						
predicted → actual ↓	**1**	**2**	**3**	**4**	**5**	**Total**
1	1,379	145	132	64	784	2,504
2	182	55	56	25	172	490
3	173	54	118	65	310	720
4	110	39	90	169	731	1,139
5	564	59	165	297	11,578	12,663

Table 13: RemBERT SC results

Domain	PC				SC			
	A	P	R	F_1	A	P	R	F_1
Appstore	0.87	0.80	0.81	0.80	0.74	0.41	0.37	0.38
Bookstore	0.86	0.75	0.80	0.77	0.73	0.34	0.32	0.32
Mapping	0.92	0.84	0.88	0.86	0.80	0.42	0.41	0.41
Market	0.97	0.84	0.91	0.87	0.82	0.43	0.41	0.42

Table 14: RemBERT results by domain

reviews from the Mapping and Market domains, which were manually collected, as opposed to reviews from the other two domains acquired through automated means. This observation suggests that the moderators may have selectively collected reviews with higher readability, fewer spelling and grammar errors, and reduced instances of code-switching and inappropriate content during the review collection process. The availability of cleaner, less noisy data could have positively influenced the performance of the model in classifying Kazakh reviews.

It is also important to recognise that the poorer performance on the SC task may not have solely stemmed from the increased complexity and challenges inherent in multi-class classification tasks. It could also be an indicator of the quality of the reviews present within the dataset. Recall that the main objective of this study was to develop a dataset that includes a diverse array of Kazakh reviews of different products and services, which, in turn, would hopefully facilitate in-depth research in Kazakh sentiment analysis. Nevertheless, we frankly admit that certain aspects, such as the correction of spelling errors, the effective handling of frequent code-switching between Kazakh and Russian, and the application of lemmatisation techniques, were not explicitly addressed and may have resulted in the lower performance of the models. These specific challenges offer promising opportunities for future investigations to improve the quality and linguistic processing capabilities of the dataset.

Upon addressing the aforementioned aspects, data augmentation techniques, such as back-translation, could be considered as possible alternatives to ROS and RUS, which were used in our study to address the data imbalance issue. The experimental findings suggest that of the four models trained on data balanced using the mentioned techniques only RemBERT exhibited improvement, albeit solely in the PC task (Burns et al., 2011).

Another challenge that may have caused the low performance on the SC task lies in the pronounced dependence of classification on the discretion of the author of a review (Smetanin and Komarov, 2021). The potential introduction of inaccuracies during the process of assigning ratings by the author could engender misclassification of the final labels within the dataset. For instance, an ostensibly positive review may paradoxically carry a score of 1; conversely, a review strongly critical of a product may be concomitantly associated with a high rating of 5. The absence of standardised criteria for sentiment labelling leads to a subjective, intuitive approach by individual authors and thus to considerable variability in the assignment of ratings between authors. This underscores the exigency of formulating sentiment annotation guidelines in and,

more importantly, for the Kazakh language, which can serve as a framework for future research in this area.

6. Conclusion

The aim of this study was to create an extensive dataset for Kazakh sentiment analysis. The result is KazSAnDRA, the first and largest publicly available dataset for Kazakh. Comprising 180,064 reviews from four domains, KazSAnDRA includes numerical ratings from 1 to 5 that quantitatively represent customers' attitudes. To automate Kazakh sentiment classification, we developed and evaluated four machine learning models for both polarity and score classification. The experimental analysis involved examining the results obtained with both balanced and imbalanced training data. The most successful model achieved an F_1-score of 0.81 for the polarity classification task and 0.39 for the score classification task on the test sets. In the future, we plan to improve KazSAnDRA by addressing spelling errors and effectively handling code-switching phenomena. These improvements will facilitate the use of advanced data augmentation techniques to cope with data imbalance challenges.

The dataset and fine-tuned models are available for unrestricted access and can be freely downloaded under the Creative Commons Attribution 4.0 International License (CC BY 4.0) from our GitHub repository.[6]

7. Acknowledgements

We sincerely thank Alma Murzagulova, Aizhan Seipanova, Meiramgul Akanova, Almas Aitzhan, Aigerim Boranbayeva, and Assel Kospabayeva, who acted as moderators during the review collection process. Their tireless efforts, diligence, and remarkable patience contributed significantly to the successful completion of this endeavour.

8. Bibliographical References

Y. B. Abdullin and V. V. Ivanov. 2017. Deep Learning Model for Bilingual Sentiment Classification of Short Texts. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 17(1):129–136.

Gavin Abercrombie and Riza Batista-Navarro. 2020. ParlVote: A Corpus for Sentiment Analysis of Political Debates. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5073–5078, Marseille, France. European Language Resources Association.

Jethin Abraham, Daniel Higdon, John Nelson, and Juan Ibarra. 2018. Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. *SMU Data Science Review*, 1(3):1–21.

Mohamed Aly and Amir Atiya. 2013. LABR: A Large Scale Arabic Book Reviews Dataset. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 494–498, Sofia, Bulgaria. Association for Computational Linguistics.

Gulmira Bekmanova, Gaziza Yelibayeva, Saltanat Aubakirova, Nurgul Dyussupova, Altynbek Sharipbay, and Rozamgul Nyazova. 2019. Methods for Analyzing Polarity of the Kazakh Texts Related to the Terrorist Threats. In *Computational Science and Its Applications – ICCSA 2019*, pages 717–730, Cham. Springer International Publishing.

Nicola Burns, Yaxin Bi, Hui Wang, and Terry Anderson. 2011. Sentiment Analysis of Customer Reviews: Balanced versus Unbalanced Datasets. In *Knowledge-Based and Intelligent Information and Engineering Systems: 15th International Conference, KES 2011, Kaiserslautern, Germany, September 12-14, 2011, Proceedings, Part I 15*, pages 161–170. Springer.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking 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 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.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):1–14.

Rama Rohit Reddy Gangula and Radhika Mamidi. 2018. Resource Creation Towards Automated Sentiment Analysis in Telugu (a low resource language) and Integrating Multiple Domain Sources to Enhance Sentiment Prediction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dinara Gimadi. 2021. Web-sentiment Analysis of Public Comments (Public Reviews) for Languages

with Limited Resources such as the Kazakh Language. *Proceedings of the Student Research Workshop Associated with RANLP 2021*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. *CS224N project report, Stanford*, 1(12):2009.

Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. 2013. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Quality & Safety*, 22(3):251–255.

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.

Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment Analysis for Low Resource Languages: A Study on Informal Indonesian Tweets. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 123–131, Osaka, Japan. The COLING 2016 Organizing Committee.

Bing Liu. 2012. *Sentiment Analysis: A Fascinating Problem*, pages 1–8. Springer International Publishing, Cham.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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*.

Mamta, Asif Ekbal, Pushpak Bhattacharyya, Tista Saha, Alka Kumar, and Shikha Srivastava. 2022. HindiMD: A Multi-domain Corpora for Low-resource Sentiment Analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7061–7070, Marseille, France. European Language Resources Association.

Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022a. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus. In *Proc. Interspeech 2022*, pages 1367–1371.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022b. "KazakhTTS2: Extending the open-source Kazakh TTS corpus with more data, speakers, and topics". In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5404–5411, Marseille, France. European Language Resources Association.

Galimkair Mutanov, Vladislav Karyukin, and Zhanl Mamykova. 2021. Multi-Class Sentiment Analysis of Social Media Data with Machine Learning Algorithms. *Computers, Materials & Continua*, 69(1):913–930.

Sergazy Sakenovich Narynov and Arman Serikuly Zharmagambetov. 2016. On One Approach of Solving Sentiment Analysis Task for Kazakh and Russian Languages Using Deep Learning. In *Computational Collective Intelligence: 8th International Conference, ICCCI 2016, Halkidiki, Greece, September 28-30, 2016. Proceedings, Part II 8*, pages 537–545. Springer.

Aliya Nugumanova, Yerzhan Baiburin, and Yermek Alimzhanov. 2022. Sentiment Analysis of Reviews in Kazakh With Transfer Learning Techniques. In *2022 International Conference on Smart Information Systems and Technologies (SIST)*, pages 1–6.

Assel Nurlybayeva, Ali Abd Almisreb, Syamimi Mohd Norzeli, and Musab AM Ali. 2022. Kazakh Text Generation using Neural Bag-of-Words Model for Sentiment Analysis. *Southeast Europe Journal of Soft Computing*, 11(2):29–39.

Aneta Pavlenko. 2008. Russian in post-Soviet countries. *Russian Linguistics*, 32(1):59–80.

Dauren Rakhymzhanov. 2022. An Approach to the Study of Implementation of Kazakh Slang Dictionary for Better Sentiment Analysis in Kazakh. *Prospects and Key Tendencies of Science in Contemporary World*.

Enislay Ramentol, Yailé Caballero, Rafael Bello, and Francisco Herrera. 2012. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowledge and Information Systems*, 33(2):245–265.

Yuliya Vladimirovna Rubtsova and Yury Alekseevich Zagorulko. 2014. An Approach to Construction and Analysis of a Corpus of Short Russian Texts Intended to Train a Sentiment Classifier. *The Bulletin of NCC*, 37:107–116.

Sergey Smetanin and Mikhail Komarov. 2021. Deep Transfer Learning Baselines for Sentiment Analysis in Russian. *Inf. Process. Manage.*, 58(3).

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.

Nazgul Toiganbayeva, Mahmoud Kasem, Galymzhan Abdimanap, Kairat Bostanbekov, Abdelrahman Abdallah, Anel Alimova, and Daniyar Nurseitov. 2022. KOHTD: Kazakh offline handwritten text dataset. *Signal Processing: Image Communication*, 108:116827.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Banu Yergesh, Gulmira Bekmanova, and Altynbek Sharipbay. 2017. Sentiment Analysis on the Hotel Reviews in the Kazakh Language. In *2017 International Conference on Computer Science and Engineering (UBMK)*, pages 790–794.

Banu Zh. Yergesh, Gulmira Bekmanova, and Altynbek Sharipbay. 2019. Sentiment Analysis of Kazakh Text and Their Polarity. *Web Intell.*, 17:9–15.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh Named Entity Recognition Dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 417–426, Marseille, France. European Language Resources Association.

Lotfi A. Zadeh. 1996. Fuzzy logic = computing with words. *IEEE Trans. Fuzzy Syst.*, 4:103–111.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(4):e1253.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1):43–52.