ISSAI
NAZARBAYEV UNIVERSITY | Institute of Smart Systems and Artificial Intelligence

LREC COLING 2024

KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes

Rustem Yeshpanov
rustem.yeshpanov@nu.edu.kz

Huseyin Atakan Varol
ahvarol@nu.edu.kz

issai.nu.edu.kz

- **Foundation**:
  - Kazakhstan, 2019
- **Focus**:
  - digital domain and AI research
- **Collaboration**:
  - national and international partners in education, industry and government
- **Projects**:
  - control of robotic systems, speech and text corpora, AI- and AR-based technologies etc.
- **Access**:
  - publicly available, free, https://github.com/IS2AI
- **Team**:
  - 11 regular employees, 5 professors, 84 research assistants, gender-balanced staff body

Sentiment Analysis

- Analysing text to understand attitudes, opinions, and emotions.
- Categorising sentiment as positive, negative, or neutral.
- Used in market research, customer service, brand management, and more.

product review	*This smartphone exceeded my expectations!* *The camera quality is outstanding, and the battery life is impressive.*	
movie review	*The dialogue felt forced, and the storyline was predictable, leaving me disappointed.*	
customer service review	*The customer service representatives were polite and responsive,* *but the resolution process for my issue took longer than expected.* *While I appreciate their efforts to assist me, the overall experience was average.*	

KazSAnDRA

- **Dataset**:
  - first & largest dataset for Kazakh
  - 180,064 reviews from 4 domains
  - 2 tasks: polarity & score classification
  - 2 scenarios: balanced & imbalanced
  - publicly available and free for use
- **Models**:
  - 4 models (mBERT, XLM-R, RemBERT, mBART-50)
  - RemBERT :
    - polarity classification: F_1-score = 0.81
    - score classification: F_1-score = 0.39
  - publicly available and free for use

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.

Kazakh Sentiment Analysis

Paper	Domain	Dataset size	Task	Access	Method	Accuracy (%)	F1 (%)
Narynov & Zharmagambetov, 2016 [1]	news comments	10,000 articles	PC	?	LSTM	72.8	?
Abdullin & Ivanov, 2017 [2]	news comments	1,400 documents	PC	?	LSTM	58	?
Yergesh et al., 2017 [3]	customer reviews	?	PC	?	Fuzzy Logic	86	?
Yergesh et al., 2019 [4]	customer reviews	?	PC	?	Rule-based	83	?
Bekmanova et al., 2019 [5]	social media comments	1,200 entries	PC	?	?	?	?
Mutanov et al., 2021 [6]	news comments	15,933 texts	PC	?	NB, SVM, LR, k-NN, DT, RF, XGBoost	89	89
Gimadi et al., 2021 [7]	customer reviews	3,000 reviews	SC	?	?	?	?
Rakhymzhanov, 2022 [8]	slang	?	?	?	?	?	?
Nurlybayeva et al., 2022 [9]	customer reviews	2,000 reviews	PC	?	BoW	93	?
Nugumanova et al., 2022 [10]	social media comments customer reviews	30 samples	PC	?	mBERT	73	?
Yeshpanov & Varol, 2024 (ours)	customer reviews	180,064 reviews	PC | SC	✓	mBERT XLM-R RemBERT mBART-50	89 | 77 89 | 77 89 | 76 89 | 77	80 | 37 81 | 39 81 | 39 80 | 38

Data Collection

- **Domains:**
  - Mapping and navigation services (Mapping)
  - Marketplaces (Market)
  - E-shop of Kazakh books (Bookstore)
  - Store of Android applications (Appstore)
- **Period:**
  - September 2022 – September 2023
- **Means:**
  - Manual: Mapping & Market
  - Automatic: Appstore & Bookstore
- **Moderators**:
  - 6 native Kazakh speakers

Domain	1 ☆	2 ☆	3 ☆	4 ☆	5 ☆	Total
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**

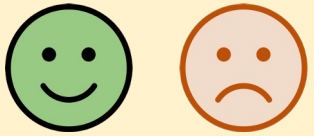
Review Variations

Actual review	Correct form (KK)	Correct form (EN)
керемет кітап	*керемет кітап*	*a wonderful book*
keremet	*керемет*	*wonderful*
jok кітап	*кітап жоқ*	*no books*
Керемет все!	*Барлығы керемет!*	*Everything's wonderful!*
Кушти!	*Күшті!*	*Great!*

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

Classification Tasks

Polarity classification



reviews with a score of 1 or 2	→	negative	→	0
reviews with a score of 3	→	neutral	→	x
reviews with a score of 4 or 5	→	positive	→	1

Score classification **[1, 5]**

Data Pre-Processing

Күштііі. 😃
Маған ұнадыыыы!

Removal of emojis	*Күштііі.* *Маған ұнадыыыы!*
Lowercasing all reviews	*күштііі.* *маған ұнадыыыы!*
Removal of punctuation marks	*күштііі* *маған ұнадыыыы*
Replacement of newline (\n), tab (\t), carriage return (\r) characters, and multiple spaces with a single space	*күштііі маған ұнадыыыы*
Reduction of consecutive recurring characters to two single instances	*күштіі маған ұнадыы*

Sets

Polarity Classification

Domain	Training		Validation		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**

Score	Training		Validation		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**

Score Classification

Domain	Training		Validation		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**

Score	Training		Validation		Test	
	#	%	#	%	#	%
5	101,302	72.29	12,663	72.29	12,663	72.29
1	20,031	14.29	2,504	14.3	2,504	14.3
4	9,115	6.5	1,140	6.51	1,139	6.5
3	5,758	4.11	719	4.1	720	4.11
2	3,920	2.8	490	2.8	490	2.8
Total	**140,126**	**100**	**17,516**	**100**	**17,516**	**100**

Score Resampling

Polarity Classification

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

Score Classification

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

- Random oversampling (ROS):
  - creating new samples for the smaller class to align with the majority class count.
- Random undersampling (RUS):
  - eliminating samples from the larger class to match the minority class count.

Experiment

Training:
- training set

Hyperparameter tuning:
- validation set

Evaluation:
- test set

Hardware:
- a single GPU on an NVIDIA DGX A100 machine

Model	mBERT[11]	XLM-R[12]	RemBERT[13]	mBART-50[14]
Support of Kazakh	✓			
Warmup steps	800			
Learning rate	10^{-5}			
Weight decay rate	10^{-3}			
Early stopping	no improvement in F_1 for 3 consecutive epochs			
Batch size	32			16

Experiment Results

Model	Polarity Classification												Score Classification											
	Balanced (ROS)				Balanced (RUS)				Imbalanced (IB)				Balanced (ROS)				Balanced (RUS)				Imbalanced (IB)			
	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	84	74	83	77	85	76	82	78	89	82	79	80	67	34	36	35	63	35	39	36	77	44	36	37
XLM-R	86	76	83	79	85	75	83	78	89	81	81	81	58	36	42	36	66	36	41	37	77	42	37	39
RemBERT	88	79	82	81	87	78	82	80	89	81	82	81	73	37	36	36	62	35	40	35	76	41	38	39
mBART-50	87	77	79	78	81	72	81	74	89	82	78	80	74	36	34	35	55	36	41	34	77	42	37	38

Model	Polarity Classification			Score Classification		
	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

Domain	Polarity Classification				Score classification			
	A	P	R	F_1	A	P	R	F_1
Appstore	87	80	81	80	74	41	37	38
Bookstore	86	75	80	77	73	34	32	32
Mapping	92	84	88	86	80	42	41	41
Market	97	84	91	87	82	43	41	42

Experiment Results (continued)

Polarity Classification

Predicted → / Actual ↓	0	1	Total
0	2,155	838	2,993
1	1,036	12,768	13,804

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

Challenges

- Spelling errors
- Frequent code-switching
- Transliteration
- Mixed script
- Homoglyphs
- Review score idiosyncrasy

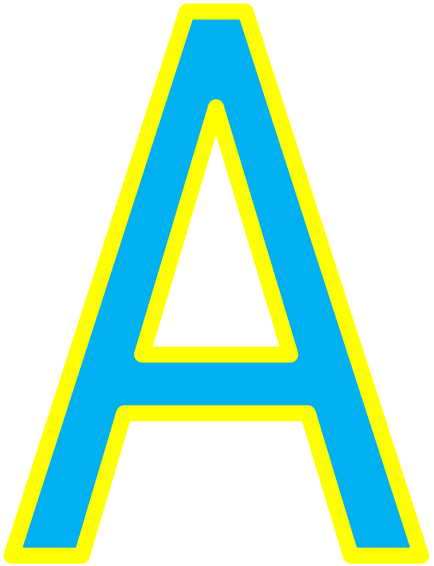
ИЯ, РАХМЕТ! → ***ИӘ, РАҚМЕТ!***

ЖАҚСЫ ИГРА. → ***ЖАҚСЫ ОЙЫН.***

ONSHA EMES. → ***ОНША ЕМЕС.***

KUSHTI БАНК! → ***КҮШТІ БАНК!***

ЕСКІ КІТАП. → ***ЕСКІ КІТАП.***



U + 0041 Latin A

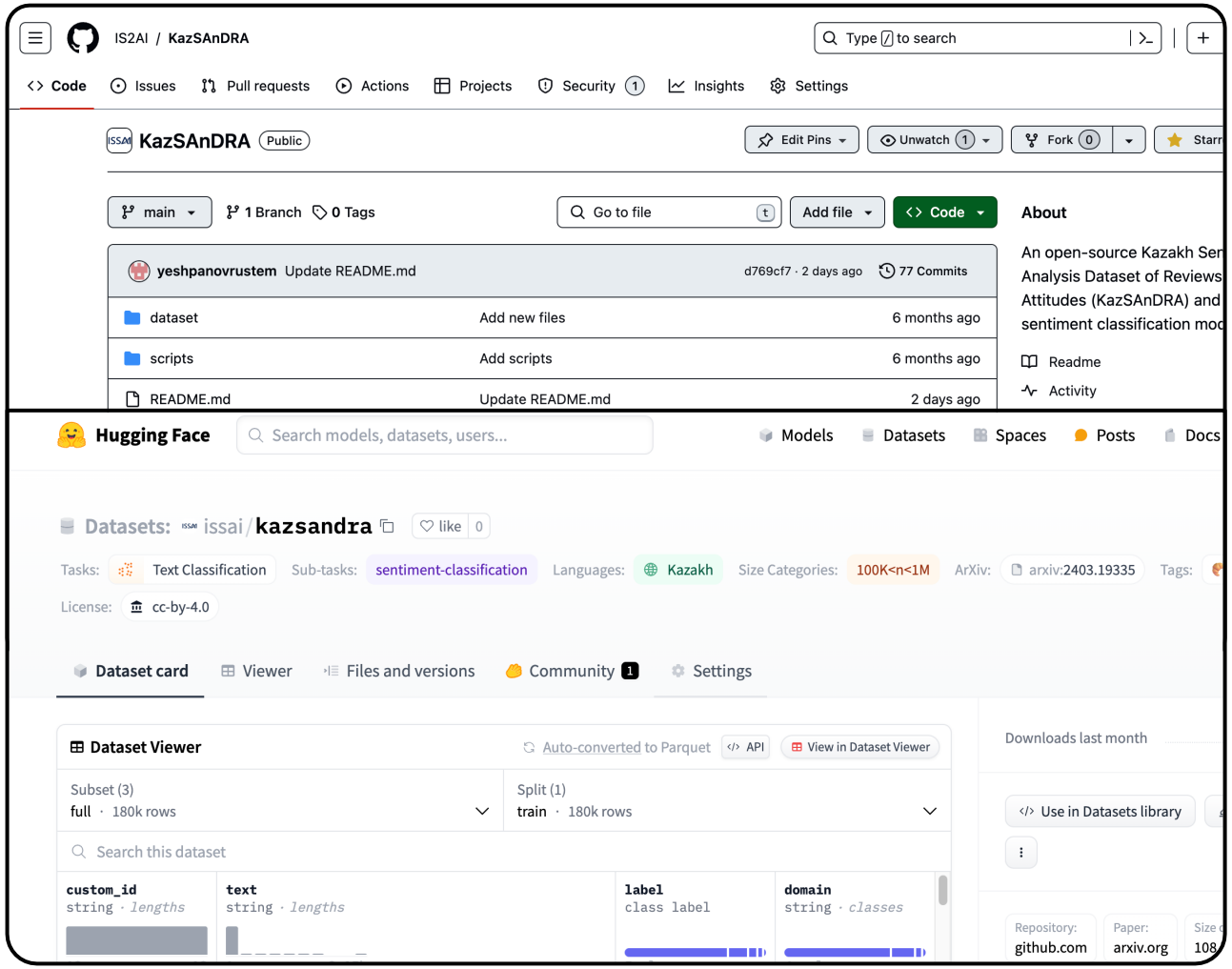
U + 0410 Cyrillic A

ЖАҚСЫ ОЙЫН. ☆ & ***ЖАМАН ОЙЫН.*** ☆☆☆☆☆

EN: GOOD GAME. *EN: BAD GAME.*

Conclusion

- **KazSAnDRA**:
  - first & largest dataset for Kazakh
  - 180,064 reviews from 4 domains
  - 2 tasks: polarity & score classification
  - 2 scenarios: balanced & imbalanced
  - publicly available and free for use
- **Models**:
  - 4 models (mBERT, XLM-R, RemBERT, mBART-50)
  - RemBERT :
    - polarity classification: F_1-score = 0.81
    - score classification: F_1-score = 0.39
  - publicly available and free for use



References

1. 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.
2. 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.
3. 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.
4. Banu Zh. Yergesh, Gulmira Bekmanova, and Altynbek Sharipbay. 2019. Sentiment Analysis of Kazakh Text and Their Polarity. Web Intell., 17:9–15.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, 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.
13. 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.
14. 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.