World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Using Machine Learning Algorithms for Kyrgyz Sentiment Analysis

İbrahim BENLİ * and Bakyt SHARSHEMBAEV

Department of Computer Engineering, Faculty of Engineering, Kyrgyzstan-Turkey Manas University, Bishkek, Kyrgyzstan.

World Journal of Advanced Research and Reviews, 2024, 23(03), 554–561

Publication history: Received on 24 July 2024; revised on 01 September 2024; accepted on 03 September 2024

Article DOI: https://doi.org/10.30574/wjarr.2024.23.3.2681

Abstract

Nowadays, social media platforms, forums and online communities have become central arenas of public discourse where individuals can freely express their opinions, feelings and attitudes. The proliferation of these platforms has led to the generation of massive amounts of unstructured data in multiple languages, providing a unique opportunity for sentiment analysis – a technique for identifying and categorizing opinions expressed in text data. Despite the global reach of sentiment analysis, there remains a significant gap in research focusing on less-researched languages such as Kyrgyz.

This study fills this gap by conducting a comprehensive sentiment analysis of Kyrgyz comments on various online platforms. A range of machine learning algorithms have been used, including traditional methods such as K-Nearest Neighbors (KNN) and Naive Bayes (NB), but also more advanced techniques such as Long Short-Term Memory (LSTM) networks and Recurrent Neural Networks (RNNs). Among the evaluated models, logistic regression (LR) was found to be the most effective, achieving the highest accuracy (0.83) and the highest F1 measure (0.84). These results highlight the potential of LR in sentiment analysis tasks for the Kyrgyz language and provide valuable insights in the field of multilingual natural language processing.

Keywords: Kyrgyz; Natural language processing; Machine learning algorithms; Sentiment analysis

1. Introduction

In the era of digital communication; social media platforms; forums; and online communities have become central to public discourse; offering a space where opinions; emotions; and attitudes are expressed freely. The rise of these platforms has generated vast amounts of unstructured data in various languages; providing a rich resource for sentiment analysis a process of identifying and categorizing opinions expressed in text. However; sentiment analysis has predominantly focused on widely spoken languages; leaving less commonly used languages; such as Kyrgyz; relatively underexplored.

Turkic languages are spoken by around 200 million people worldwide. Among them; native speakers of Kazakh; Uzbek; Kyrgyz; and Turkmen; living both inside and outside their respective countries; make over 60 million of those speakers (1). The Kyrgyz language; spoken by over four million people primarily in Kyrgyzstan; represents a unique linguistic and cultural context. As the online presence of Kyrgyz speakers grows; understanding the sentiments conveyed in Kyrgyz texts becomes increasingly important for applications ranging from market research to sociopolitical analysis. Despite this; the development of natural language processing (NLP) tools for Kyrgyz remains in its nascent stages; posing challenges for accurate sentiment analysis.

This study aims to fill this gap by conducting a comprehensive sentiment analysis of Kyrgyz comments from various online platforms. To achieve this; the research will employ a range of machine learning algorithms; including both

* Corresponding author: İbrahim BENLİ

traditional methods like LR; KNN and NB; as well as more advanced techniques such as LSTM networks and RNNs. These algorithms will be applied and evaluated to determine their effectiveness in accurately classifying the sentiments expressed in Kyrgyz comments. The combination of various machine learning approaches and the consideration of linguistic nuances aims to provide a robust framework for sentiment analysis in the Kyrgyz language.

The findings of this research will not only contribute to the growing field of sentiment analysis in lesser-studied languages but also offer insights into the public sentiments of Kyrgyz-speaking communities; which could be valuable for policymakers; businesses; and social scientists alike.

2. Related Work

Sentiment analysis; a subfield of NLP; has received considerable attention in recent years; especially for widely spoken languages like English and Chinese. However; sentiment analysis for Turkic languages; which include languages such as Turkish; Azerbaijani; Uzbek; and Kyrgyz; has been relatively underexplored. Despite this; there have been significant strides in the development of sentiment analysis models tailored to these languages.

Turkish; the most widely spoken Turkic language; has been the focus of numerous studies in sentiment analysis. Demirtas et.al explored sentiment classification using machine learning techniques; demonstrating that Support Vector Machines (SVM) and NB classifiers perform well on Turkish datasets (2). Similarly; Vural et al. examined the effectiveness of various machine learning algorithms; including KNN and Decision Trees; for sentiment analysis in Turkish; highlighting the challenges posed by the agglutinative nature of the language (3). Recent advancements have also seen the application of deep learning models. A study by Bilen and Horasan employed LSTM networks for sentiment analysis in Turkish; achieving superior results compared to traditional machine learning methods (4).

Azerbaijani; another prominent Turkic language; has seen fewer studies in sentiment analysis. Nonetheless; research by Hasanli and Rustamov applied SVM and NB classifiers to Azerbaijani social media data; demonstrating the potential for these methods despite the limited availability of annotated datasets (5). Guliyev et al. used a Bidirectional LSTM model to analyze sentiment in Azerbaijani texts; achieving promising results (6).

Sentiment analysis in Uzbek has been relatively underdeveloped; partly due to the scarcity of NLP resources for the language. However; recent efforts have begun to address this gap. A study by Rabbimov et al. utilized a combination of rule-based methods and machine learning algorithms, such as Decision Trees; for sentiment classification in Uzbek text (7). The authors highlighted the challenges posed by the language's complex morphology and the lack of a comprehensive lexicon.

Kazakh and Kyrgyz; while less studied than Turkish and Azerbaijani; have also seen emerging research in sentiment analysis. Gimadi et al. conducted one of the first studies on sentiment analysis in Kazakh; applying SVM and Random Forest classifiers to a dataset of social media posts. Their research underscored the need for more extensive annotated corpora and advanced NLP tools for Kazakh (8). In Kyrgyz Choi and Abdieva studied the process; which encompassed recording news from Kyrgyzstan; collecting and refining data; developing a sentiment lexicon; constructing a deep learning model; and processing data in detail(9) .

Some studies have explored cross-linguistic and multilingual approaches to sentiment analysis in Turkic languages. Çöltekin et al. investigated multilingual sentiment analysis using transfer learning techniques; applying models trained on Turkish to other Turkic languages like Azerbaijani and Uzbek. The study demonstrated the potential for leveraging linguistic similarities across Turkic languages to improve sentiment analysis performance (10).

Another promising approach is the development of multilingual models Acikalin et al. conducted a study using multilingual BERT (Bidirectional Encoder Representations from Transformers) and Turkish transformer models; including MBERT; XLM-Roberta; and BERTurk; to analyze a dataset of Turkish tweets. Their findings demonstrated significant improvements in sentiment analysis tasks for the Turkish language (11). Also Guven trained and tested pre-trained models such as BERT; ALBERT; ELECTRA; and DistilBERT on Turkish hotel and movie reviews. This approach underscores the potential benefits of shared linguistic features among Turkic languages in building more robust NLP models.

3. Methodology

This study investigates the effectiveness of various machine learning and deep learning models in performing sentiment analysis on a dataset of Kyrgyz text. The dataset was derived from the IMDB review dataset available on Kaggle; consisting of 500 reviews (250 positive and 250 negative); which were translated from English to Kyrgyz using Google Translate. The methodology encompasses several stages: initial model training; text preprocessing; feature extraction; and further model evaluation; culminating in the use of a pre-trained multilingual BERT model.

3.1. Dataset Preparation

3.1.1. Translation and Labeling

The IMDB review dataset; a widely used benchmark for sentiment analysis; was translated into Kyrgyz using Google Translate. The dataset was carefully balanced; containing an equal number of positive and negative reviews. Although machine translation may introduce some noise; it provides a practical approach for creating a Kyrgyz dataset; given the scarcity of native language resources.

3.2. Initial Model Training and Evaluation

3.2.1. Deep Learning Models

- Initially; several machine learning and deep learning algorithms were applied to the translated dataset to evaluate their effectiveness in sentiment classification. The models included:
- LSTM: An RNN variant designed to capture long-term dependencies in text sequences; which is crucial for handling the complex sentence structures in Kyrgyz.
- RF: An ensemble learning method that builds multiple decision trees and merges their results to improve prediction accuracy. It is particularly robust against overfitting.
- KNN: A simple; instance-based learning algorithm that classifies a review based on the sentiment of its nearest neighbors in the feature space.
- RNN: A neural network architecture that processes sequences of text by maintaining a hidden state; capturing temporal dependencies.
- NB: A probabilistic classifier based on Bayes' theorem; commonly used for text classification tasks.
- LR: A linear model that predicts the probability of a review being positive or negative.

3.2.2. Initial Results

The performance of these models was evaluated using accuracy as the primary metric. Among the models; LR consistently outperformed the others; achieving the highest accuracy.

3.3. Text Preprocessing and Feature Extraction

To enhance the accuracy of the sentiment analysis models; further preprocessing was applied to the text data.

Stopword Removal: A list of 50 Kyrgyz stopwords was compiled and removed from the text. Stopwords; which are common words that do not contribute to the sentiment; can often introduce noise into the model. Their removal typically helps in improving model performance. Table 1 and Figure 1 illustrate the stopword removal process.

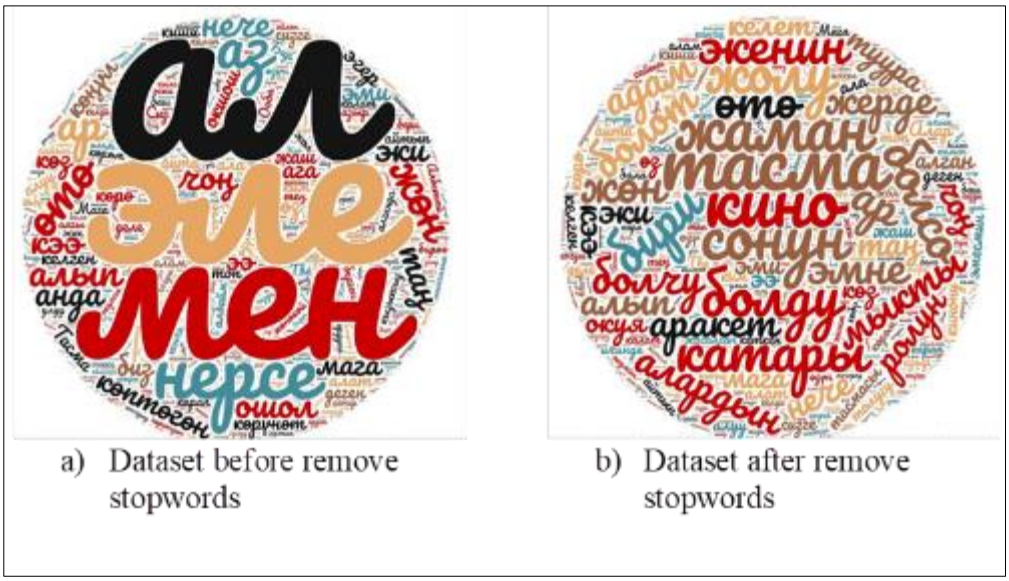

Figure 1 Word clouds for before and after stopwords

Table 1 Removal of Stopwords

Stage	**Text**
Normal Text	EN: This movie was a complete waste of time. The soundtrack is poor, the story is lame and predictable, and the acting is terrible. One of the 25 worst movies I've ever seen.
	KY: Бул кино толугу менен убакытты текке кетирди. Саундтрек начар, окуя аксап, алдын ала айтууга болот, ал эми актёрдук чеберчилиги коркунучтуу. Мен көргөн эң начар 25 тасманын бири.
After Remove Stopwords	EN: movie total waste time bad soundtrack bad story lame acting terrible worst movies
	KY: кино толугу убакытты текке кетирди Саундтрек начар окуя аксап айтууга болот актёрдук чеберчилиги коркунучтуу көргөн начар тасманын бири

3.3.1. Lemmatization

To reduce inflectional forms and related forms of a word to a common base form, lemmatization was applied using the UD Kyrgyz-KTMU model. Lemmatization is particularly important for agglutinative languages like Kyrgyz, where words can take on many different forms depending on their grammatical usage.

Table 2 Lemmatization Process in Texts

Stage	Text
Normal Text	EN: movie total waste time bad soundtrack bad story lame acting terrible worst movies
	KY: кино толугу убакытты текке кетирди Саундтрек начар окуя аксап айтууга болот актёрдук чеберчилиги коркунучтуу көргөн начар тасманын бири
Lemmatization	EN: movie total waste time bad soundtrack bad story lame act terrible bad movie KY: кино толук убакыт тек кетир Саундтрек начар окуя акса бол актёр чеберчилик коркунуч көр начар тасма бири

3.4. Multilingual BERT Model

To further enhance the performance of the sentiment analysis, a pre-trained multilingual BERT model was fine-tuned on the Kyrgyz dataset. BERT is a transformer-based model that captures context more effectively than traditional models, particularly for languages with complex grammar and morphology.

Fine-Tuning and Evaluation: The multilingual BERT model was fine-tuned using the translated Kyrgyz reviews, following standard practices for text classification tasks. The fine-tuned model's performance was evaluated using the same test set, and while it showed strong results, the Random Forest model still outperformed it in terms of accuracy, underscoring its robustness for this specific task.

4. Results and Discussion

As demonstrated in Table 3, the pre-processing steps applied to the dataset led to a notable reduction in both total and unique word counts. Specifically, there was a 21% decrease in the total word count and a 35% reduction in the number of unique words. This significant reduction in unique words is primarily due to the lemmatization process, where words are reduced to their base forms, thereby standardizing different inflected forms of a word into a single representation. Such findings align with previous research in the field, which has shown that lemmatization and stopword removal can significantly reduce the vocabulary size without losing essential information for sentiment classification tasks (12).

Table 3 Dataset statistics

Process		**Words Count**
Dataset	Total	75641
	Unique	16632
After Preprocessing	Total	59226
	Unique	10339

In the study, the dataset was split into 80% for training and 20% for validation, a standard practice in machine learning to ensure model generalizability and to reduce the risk of overfitting (13). The training set was used to train various machine learning models, and their performance was evaluated using the validation set.

The performance of various machine learning models applied to the sentiment analysis of Kyrgyz film reviews was systematically evaluated, with results presented in Table 4. Default parameters were used in all methods during evaluations. Among the models, LR emerged as the most effective, achieving the highest accuracy (0.83) and F1-measure (0.84). This superior performance aligns with existing literature, where LR is frequently highlighted for its robustness in text classification tasks (14).

The RF model also demonstrated strong performance, with an accuracy of 0.77 and an F1-measure of 0.78, further supporting its reputation for being a reliable and versatile classifier in NLP tasks (15). Other models such as LSTM, RNN, KNN, and NB exhibited moderate performance, with accuracies ranging from 0.64 to 0.70 and F1-measures between 0.69 and 0.76. Additionally, the confusion matrix in Figure 4 provides a detailed analysis of the model's performance on individual labels, further illustrating the effectiveness of LR in this context.

Table 4 Results of machine learning methods

Method	**Accuracy**	**F1-Measurement**
LSTM	0.66	0.70
RF	0.77	0.78
KNN	0.64	0.69
RNN	0.76	0.77
NB	0.70	0.76
LR	0.83	0.84

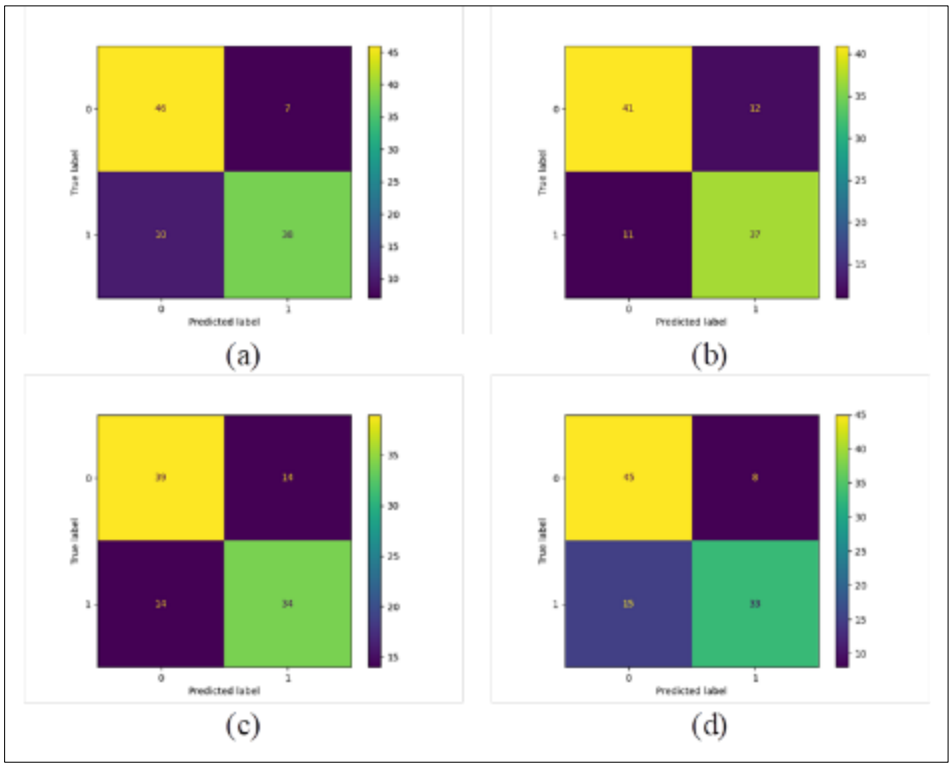

Figure 4 Confusion matrices of machine learning methods (a) LR (b) KNN (c) NB (d) RF

In this study, we directly imported BERT's preprocessor and the pre-trained BERT model from the TensorFlow Hub website. We imported both the preprocessor and the model by accessing them through the URL. Interestingly, the multilingual BERT model, which leverages transformer architecture, achieved an accuracy of 0.60. While BERT-based models are generally known for their superior performance in NLP tasks across various languages, the relatively lower accuracy observed here could be attributed to the specific characteristics of the Kyrgyz language or the domain-specific nature of the dataset. This outcome suggests that while transformer models offer significant advantages, their effectiveness may vary depending on the linguistic and contextual nuances of the dataset being analyzed (16,17).

Firstly, the development of a Kyrgyz-specific pre-trained BERT model is essential. The results from this study indicate that the multilingual BERT model, while useful, did not perform as well as traditional models such as LR. This outcome underscores the importance of having a BERT model specifically tailored to the Kyrgyz language. Previous studies have shown that language-specific BERT models, such as BERTurk for Turkish, significantly outperform their multilingual counterparts in various NLP tasks, including sentiment analysis (18,19). Therefore, a Kyrgyz-specific BERT model could potentially yield better performance by capturing the linguistic nuances of the language more effectively.

Additionally, there is a pressing need for more extensive datasets in the Kyrgyz language. The limited size of the dataset used in this study, comprising 500 translated IMDB reviews, constrained the ability to fully leverage the capabilities of deep learning models. Studies in other languages have demonstrated that larger datasets significantly improve model performance, particularly for complex models like LSTM and RNN (20). Expanding the availability of annotated Kyrgyz datasets will be crucial for advancing sentiment analysis and other NLP tasks in this language.

In studies of Turkish sentiment analysis, for instance, transformer-based models like BERTurk and mBERT have shown strong results, with accuracy scores ranging between 0.74 and 0.80 (11). However, the relatively lower performance of the multilingual BERT model in this study (accuracy 0.60) suggests that traditional models like LR may offer more robust performance for languages like Kyrgyz, which have fewer linguistic resources and annotated datasets available.

Research on Uzbek sentiment analysis by Kuriyozov et al. found that RNN was the most effective model, achieving an accuracy of 0.89 (21) This finding is consistent with the RNN performance observed in this study (accuracy 0.76, F1-measure 0.77), supporting the idea that RNN is a versatile and reliable model across different Turkic languages, particularly for smaller datasets.

Similarly, in a study by Rasul, the NB model achieved one of high performance in the sentiment analysis of Azerbaijani text, with an accuracy of 0.74, which is comparable to the NB results in this study (accuracy 0.70, F1-measure 0.76) (22). This suggests that NB continues to be a strong contender for sentiment analysis in Turkic languages, despite the increasing popularity of deep learning models.

Moreover, the moderate performance of deep learning models like LSTM and KNN in this study, with accuracy scores of 0.66 and 0.64 respectively, contrasts with the higher accuracy typically reported in Turkish sentiment analysis studies, where LSTM models often exceed 0.90 (4). This discrepancy may be due to the limited availability of large annotated datasets and language-specific resources for Kyrgyz, which are crucial for training deep learning models effectively (23).

5. Conclusion

The comparative analysis of machine learning models for sentiment analysis of Kyrgyz comments reveals LR achieved the highest performance, with an accuracy of 0.83 and an F1-measure of 0.84. This performance surpasses that reported in similar studies on other Turkic languages, indicating that LR is particularly effective for sentiment analysis in Kyrgyz.

These findings highlight the importance of context and language-specific characteristics in determining the optimal approach for sentiment analysis. While advanced transformer models have shown promise, traditional methods such as LR and RF continue to offer reliable performance in under-resourced languages like Kyrgyz.

While this study has provided valuable insights into the application of machine learning models for sentiment analysis of Kyrgyz text, several areas warrant further investigation to enhance the accuracy and reliability of sentiment analysis in this language. Another critical area for future research is the development of more effective lemmatization models. The UD Kyrgyz-KTMU model, trained on just 7.4K words, provided a foundational resource for this study, but its limited scope likely hindered the full potential of the sentiment analysis. Expanding this model to include a lexicon of at least 50K words could significantly enhance the accuracy of lemmatization, thereby improving the overall performance of sentiment analysis models.

In conclusion, future work should prioritize the creation of a Kyrgyz-specific BERT model, the expansion of Kyrgyz language datasets, and the development of more comprehensive lemmatization tools. These advancements will not only improve sentiment analysis in Kyrgyz but also contribute to the broader field of natural language processing in underrepresented languages.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Veitsman Y. Recent Advancements and Challenges of Turkic Central Asian Language Processing [Internet]. 2024. Available from: https://arxiv.org/abs/2407.05006

[2] Demirtas E, Pechenizkiy M. Cross-lingual polarity detection with machine translation. Proc 2nd Int Work Issues Sentim Discov Opin Mining, WISDOM 2013 - Held Conjunction with SIGKDD 2013. 2013;

[3] Vural AG, Cambazoglu BB, Senkul P, Tokgoz Z. A Framework for Sentiment Analysis in Turkish Application to Polarity Detection of Movie Reviews in Turkish. In: Computer and information sciences III: 27th international symposium on computer and information sciences. Springer; 2013. p. 437–45.

[4] Bilen B, Horasan F. LSTM Network based Sentiment Analysis for Customer Reviews. Politek Derg. 2022;25(3):959–66.

[5] Hasanli H, Rustamov S. Sentiment Analysis of Azerbaijani twits Using Logistic Regression, Naive Bayes and SVM. In 2019. p. 1–7.

[6] Guliyev N, Rustamov Z, Rustamov S. Analysis of Public Sentiment in Azerbaijani News and Social Media: Exploring Public Opinion Trends in Social Media and News Articles for Azerbaijani Language. ACM Int Conf Proceeding Ser. 2024;

[7] Rabbimov IM, Yusupov OR, Kobilov SS. Algorithm of decision trees ensemble for sentiment analysis of Uzbek text. In: Artificial Intelligence, Blockchain, Computing and Security Volume 2. CRC Press; 2024. p. 686–93.

[8] Gimadi D, Evans R, Simov K. Web-sentiment Analysis Of Public Comments (Public Reviews) For Languages With Limited Resources Such As The Kazakh Language. Int Conf Recent Adv Nat Lang Process RANLP. 2021;2021-Septe:65–8.

[9] Choi Y-S, Abdieva S. a Study on the Sentiment Analysis (Positive, Negative) of Words Appearing in Kyrgyz News By Applying the Deep Learning-Based Nlp (Natural Language Processing) Techniques for Students Practice. Her KSUCTA n a N Isanov. 2021;3(3–2021):372–80.

[10] Zampieri M, Nakov P, Rosenthal S, Atanasova P, Karadzhov G, Çöltekin Ç, et al. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). arXiv Prepr arXiv200607235. 2020;

[11] Acikalin UU, Bardak B, Kutlu M. Turkish sentiment analysis using bert. In: 2020 28th Signal Processing and Communications Applications Conference (SIU). 2020. p. 1–4.

[12] Palomino MA, Aider F. Evaluating the Effectiveness of Text Pre-Processing in Sentiment Analysis. Appl Sci. 2022;12(8765):1–21.

[13] Hastie T, Tibshirani R, Wainwright M. Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press; 2021.

[14] Gudari S, G DV. Comparative Study of Logistic Regression and LSTM for Sentiment Classification Across Diverse Textual Dataset. 2023.

[15] Chen H, Wu L, Chen J, Lu W, Ding J. A comparative study of automated legal text classification using random forests and deep learning. Inf Process Manag. 2022;59(2):102798.

[16] Devlin J, Chang M-W, Lee K, Google KT, Language AI. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Naacl-Hlt 2019 [Internet]. 2018;(Mlm):4171–86. Available from: https://aclanthology.org/N19-1423.pdf

[17] Pires T, Schlinger E, Garrette D. How Language-Neutral is Multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics [Internet]. 2019. p. 4996–5001. Available from: http://arxiv.org/abs/1911.03310

[18] Schweter S. BERTurk - BERT models for Turkish [Internet]. Zenodo; 2020. Available from: https://doi.org/10.5281/zenodo.3770924

[19] Antoun W, Baly F, Hajj H. AraBERT: Transformer-based Model for Arabic Language Understanding. arXiv Prepr arXiv200300104 [Internet]. 2021; Available from: https://arxiv.org/abs/2003.00104

[20] Liu Y. Roberta: A robustly optimized bert pretraining approach. arXiv Prepr arXiv190711692. 2019

[21] Kuriyozov E, Matlatipov S, Alonso MA, Gómez-Rodríguez C. Deep Learning vs. Classic Models on a New Uzbek Sentiment Analysis Dataset. Proc 9th Lang Technol Conf Hum Lang Technol as a Chall Comput Sci Linguist [Internet]. 2019;(December):258–62. Available from: http://aboutworldlanguages.com/uzbek

[22] Rasul M. Comparative Analysis of Machine Learning Algorıthms for Sentiment Analysis of Text in Azerbaijani and English. Glob J Bus Integr Secur [Internet]. 2023;(2). Available from: https://gbis.ch/index.php/gbis/article/view/228

[23] Tohma K, Kutlu Y. Challenges Encountered in Turkish Natural Language Processing Studies. Nat Eng Sci. 2020;5(3):204–11.