*Yernar Akhmetbek*[1]
[1]Suleyman Demirel University, Kaskelen, Kazakhstan

# Analysis, Visualization, and Predictive Modeling for News Viewer Count on “Khabar.kz”

***Abstract***. As the world becomes increasingly interconnected, keeping up with news from around the globe has become more important than ever. For those interested in the latest developments in Kazakhstan, the website https://khabar.kz/ru/ provides a wealth of information about the country's current events, politics, economy, and culture. As a data collection analysis, we will delve into the content of this website and examine the topics covered, the frequency of updates, and other relevant metrics to gain a better understanding of the nature and scope of the news being reported. By analyzing the data collected, we can identify patterns, trends, and insights that will help us gain a more nuanced understanding of the current state of affairs in Kazakhstan.

***Keywords***: news, articles, machine learning, data visualization, statistics, khabar

***Аннотация***. Поскольку мир становится все более взаимосвязанным, быть в курсе новостей со всего земного шара стало важнее, чем когда-либо. Для тех, кто интересуется последними событиями в Казахстане, веб-сайт https://khabar.kz/ru / предоставляет обширную информацию о текущих событиях в стране, политике, экономике и культуре. В рамках анализа сбора данных мы углубимся в содержание этого веб-сайта и изучим освещаемые темы, частоту обновлений и другие соответствующие показатели, чтобы лучше понять характер и масштаб сообщаемых новостей. Анализируя собранные данные, мы можем выявить закономерности, тенденции и инсайты, которые помогут нам получить более детальное представление о текущем состоянии дел в Казахстане.

***Ключевые слова***: новости, статьи, машинное обучение, визуализация данных, статистика, хабар

***Аңдатпа.*** Әлем барған сайын бір-бірімен байланысты болғандықтан, бүкіл әлем жаңалықтарынан хабардар болу бұрынғыдан да маңыздырақ. Қазақстандағы соңғы оқиғаларға қызығушылық танытқандар үшін веб-сайт https://khabar.kz/ru / елдегі ағымдағы оқиғалар, саясат, экономика және мәдениет туралы кең ақпарат береді. Деректерді жинауды талдау аясында біз осы веб-сайттың мазмұнына тереңірек үңіліп, хабарланған жаңалықтардың сипаты мен ауқымын жақсырақ түсіну үшін тақырыптарды, жаңарту жиілігін және басқа да тиісті көрсеткіштерді қарастырамыз. Жиналған деректерді талдай отырып, біз Қазақстандағы істердің ағымдағы жай-күйі туралы неғұрлым егжей-тегжейлі түсінік алуға көмектесетін заңдылықтарды, үрдістер мен инсайттарды анықтай аламыз.

***Түйін сөздер***: жаңалықтар, мақалалар, Машиналық оқыту, деректерді визуализациялау, статистика, хабар

## Introduction

Kazakhstan is a country of great significance, both for its people and for the world at large. As the largest landlocked country in the world, Kazakhstan is located at the crossroads of Europe and Asia, and its strategic location makes it a key player in the geopolitics of Central Asia. Kazakhstan has a rich history and culture that is reflected in its architecture, art, music, and literature. The country is also home to vast natural resources, including oil, gas, and minerals, which make it an important player in the global economy.

In order to understand the current state of affairs in Kazakhstan, it is important to keep up with the latest news and developments. The news media plays a critical role in informing the public about events, issues, and trends, and in shaping public opinion. News can also serve as a watchdog, holding those in power accountable for their actions and shining a light on corruption and wrongdoing.

One of the most prominent news websites covering current events in Kazakhstan is https://khabar.kz/ru/. This website provides a wealth of information about politics, economy, culture, and society in Kazakhstan. By analyzing the content of this website, we can gain valuable insights into the state of affairs in Kazakhstan and the issues that matter most to its people.

In this article, we explore the importance of news and the benefits of analyzing news content from https://khabar.kz/ru/. We will answer three key questions: Why Kazakhstan? Why do we need news? And what can we gain from analyzing news content? By addressing these questions, we hope to shed light on the value of news and the insights that can be gleaned from analyzing it. Whether you are a student, a business person, a policymaker, or simply someone with an interest in the world around you, understanding the news from Kazakhstan can help you make more informed decisions and gain a deeper appreciation for this fascinating country.

## Aims and objectives of the research

The aim of this research is to analyze the types of news that are most popular among readers on the https://khabar.kz/ru/ website, which covers current events in Kazakhstan. Furthermore, this study aims to examine how the rate of news coverage changes over time, by comparing the frequency of updates during different periods.

Objectives:

1. To identify the most common topics and categories of news articles on the website, such as politics, economy, society, culture, and sports.
2. To determine which types of news articles receive the highest levels of user engagement, such as views, shares, and comments.
3. To compare the rate of news coverage across different time periods, such as months, quarters, or years, to identify any patterns or trends in the website's news cycle.
4. Research will also employ machine learning techniques to predict the viewership of news articles on the https://khabar.kz/ru/ website. By leveraging historical data and

various features associated with news articles, such as topic, category, and publication time, a predictive model will be developed.

## Literature review

News plays a crucial role in informing individuals about the world around them and shaping their perceptions of current events. In recent years, the rise of digital media has transformed the way news is produced, distributed, and consumed, leading to new challenges and opportunities for the news industry[1]. Analyzing the patterns and trends of news consumption is a critical task for researchers, as it can provide insights into the preferences and behaviors of news audiences.

Several studies have examined the factors that influence news consumption and engagement on digital platforms. For example, [2] found that personal interest and relevance are the most significant factors in predicting news engagement, while source credibility and social media sharing also play important roles. Similarly, [3] found that audience characteristics, such as age, gender, and education, influence news consumption patterns on mobile devices.

Other studies have focused on the role of news framing and presentation in shaping audience perceptions and behaviors. For instance, [4] found that negative news framing can increase engagement and discussion on social media, while positive news framing can enhance trust and credibility. Moreover, the placement and prominence of news stories on a website can also affect their visibility and impact [5].

In the context of Kazakhstan, several studies have examined the state of media and news consumption in the country. For example, [6] analyzed the use of social media by Kazakhstani journalists and found that Facebook and Twitter are the most popular platforms for news dissemination and engagement. Moreover, they found that the majority of journalists prioritize political and economic news, while social and cultural issues receive less coverage.

Another study by [7] examined the attitudes and perceptions of Kazakhstani youth towards news media and found that while traditional media such as television and newspapers are still popular, social media platforms are increasingly used for news consumption and engagement. They also found that the youth tend to prefer news that is locally relevant and personalized.

In terms of news websites in Kazakhstan, "*Khabar.kz*" stands out as one of the most prominent and widely-read platforms. [8] conducted a content analysis of 'Khabar.kz' and found that political news and events related to the President are the most frequently covered topics. They also found that news about sports, culture, and science receives less coverage[9].

## Methods and Materials

This study employed a quantitative research approach to analyze the news content on the "Khabar.kz" website. The data collection method used for this research was web scraping. Python programming language was used along with Beautiful Soup, a Python library for web scraping, to extract data from the "Khabar.kz" website[10].

The entire content of the website was scraped using a Python script to extract data on news articles and their respective publication dates, headlines, subheadings, and body text. The data was then cleaned and analyzed using Python packages such as pandas, numpy, and matplotlib.

To determine the types of news that are most interesting to the website's readers, the study analyzed the distribution of news articles across various categories such as politics, business, sports, and entertainment. The frequency of news articles published in each category was calculated and plotted using histograms.

The research was conducted using publicly available data from the “Khabar.kz” website. The data used for the analysis did not contain any personally identifiable information or violate any ethical guidelines.

In addition to the analysis of news categories, this study aimed to predict the number of views for news articles using machine learning techniques. A linear regression model was trained on a dataset(*Figure-2*) consisting of news article text and their corresponding views. To prepare the textual data for modeling, the TfidfVectorizer from the sci-kit-learn library was used to convert the text into numerical features.

Hyperparameter tuning was performed on the linear regression model to optimize its performance. The parameters considered for tuning included “learning rate”, and “momentum”. GridSearchCV from sci-kit-learn was employed to search for the best combination of hyperparameters using cross-validation.

The trained linear regression model was evaluated using the root square error (MSE) metric to assess its predictive performance. The MSE value provides an estimate of the average difference between the predicted and actual number of views for news articles.

The research was conducted using publicly available data from the Khabar.kz website. The data used for the analysis did not contain any personally identifiable information and adhere to ethical guidelines.

Overall, this study utilized web scraping, data cleaning, exploratory data analysis, machine learning, and hyperparameter tuning techniques to analyze the news content of the “Khabar.kz” website and predict the number of views for news articles.

## Data and Results

All data has been scrapped from the website https://khabar.kz/ru/news, where we took up all available data inside(*Figure-1*).

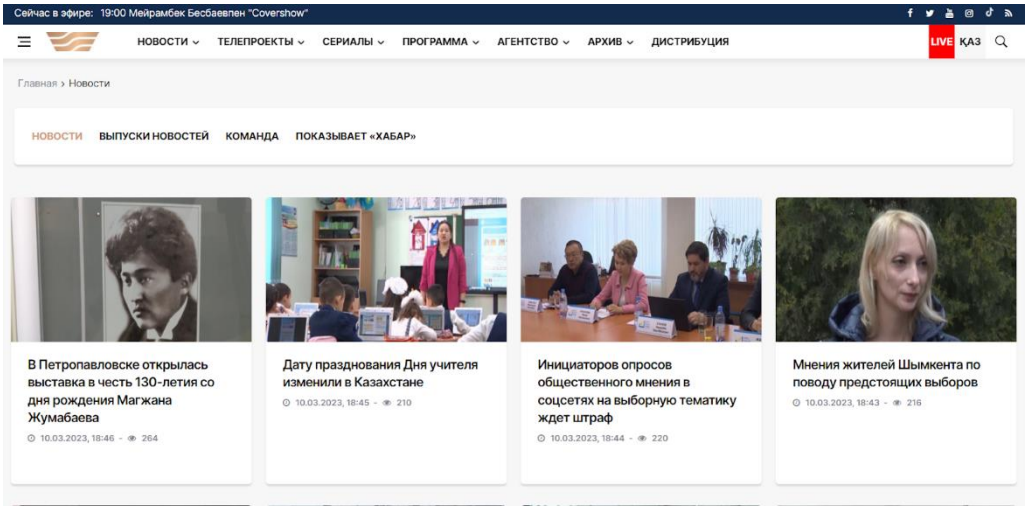


***Figure-1*:** “khabar.kz” website home page

The news publication dates were categorized by year, month, day, and time to determine the most frequent publishing times. This analysis aimed to identify the predominant patterns of news consumption in Kazakhstan, specifically highlighting the periods during which news actions were most prevalent. The results revealed that news articles were primarily published on festive occasions.(*Figure-3,4,5,6*)

| | Unnamed: 0 | Title | Date | Viewers | Type | Day | Month | Year | Time |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | Мәжіліс пен мәслихаттар сайлауында 12 451 канд... | 25.02.2023, 21:03 | 166 | News | 25 | 02 | 2023 | 21:03 |
| **1** | 1 | Бейбарыс Сұлтанның туғанына 800 жыл | 24.02.2023, 20:13 | 252 | News | 24 | 02 | 2023 | 20:13 |
| **2** | 2 | Тіқұшақ апаты: ТЖМ өкілдері қаза тапқан 4 адам... | 24.02.2023, 19:57 | 202 | News | 24 | 02 | 2023 | 19:57 |
| **3** | 3 | Польша-Беларусь шекарасы жабылды | 24.02.2023, 19:55 | 215 | News | 24 | 02 | 2023 | 19:55 |
| **4** | 4 | Түркияның Хатай провинциясында тағы жер сілкін... | 24.02.2023, 19:42 | 204 | News | 24 | 02 | 2023 | 19:42 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **10314** | 10314 | Ардагерлер азайып бара жатыр... | 26.12.2014, 18:22 | 6295 | News | 26 | 12 | 2014 | 18:22 |
| **10315** | 10315 | Жақия Мұсатаев: Өскелең ұрпақ бүгінгі бейбіт к... | 21.12.2014, 04:18 | 4890 | News | 21 | 12 | 2014 | 04:18 |
| **10316** | 10316 | Таразда Ардагерлер үйі ашылды | 16.12.2014, 23:50 | 7138 | News | 16 | 12 | 2014 | 23:50 |
| **10317** | 10317 | ҰОС ардагері А.Кияткин ұрпақтарымен бақуатты ө... | 04.12.2014, 11:20 | 2796 | News | 04 | 12 | 2014 | 11:20 |
| **10318** | 10318 | Ардагерлер болашақ ел қорғаушыларды отансүйгіш... | 25.11.2014, 12:25 | 2728 | News | 25 | 11 | 2014 | 12:25 |

***Figure-2*:** dataset took from webpage

All the data has been visualized with the time-number of new relation.

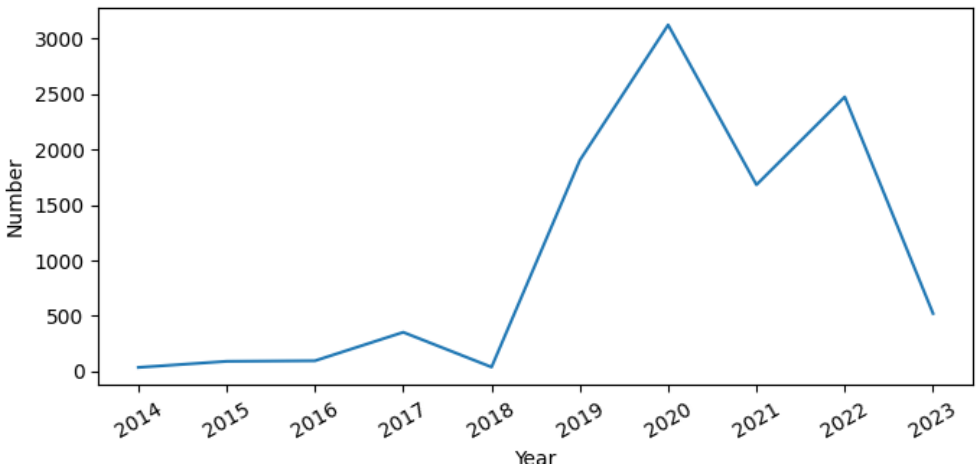


***Figure-3*:** news occurrences according to the year

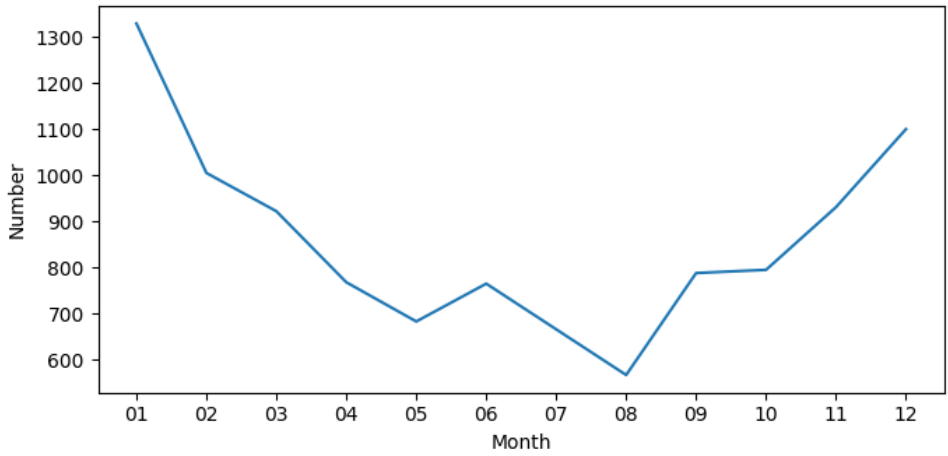


***Figure-4*:** news occurrences according to the month

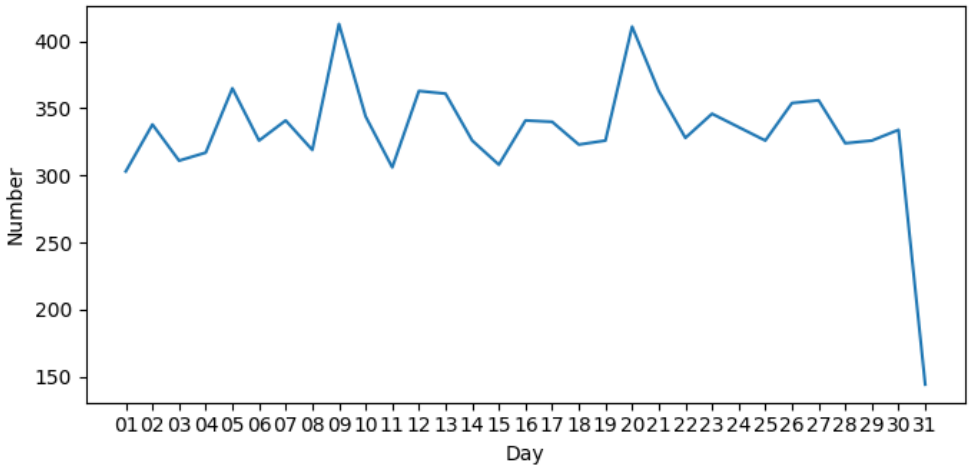


***Figure-5*:** news occurrences according to the days

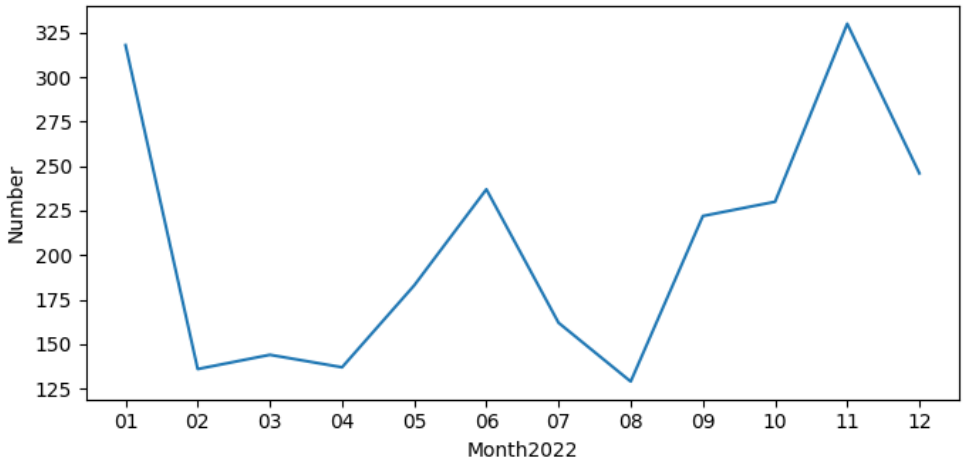


***Figure-6:*** news occurrences according to the month of 2022

A quantitative analysis was conducted on the news articles published on a particular website. Specifically, the occurrences of individual words were tallied and the top 20 most frequently used words were visualized. It is worth noting that these top 20 words are likely to be closely related to the dominant themes and topics covered on the website. For instance, the frequent appearance of the word 'облысында' suggests that regional affairs are a major focus of news coverage in Kazakhstan.(*Figure-7*)

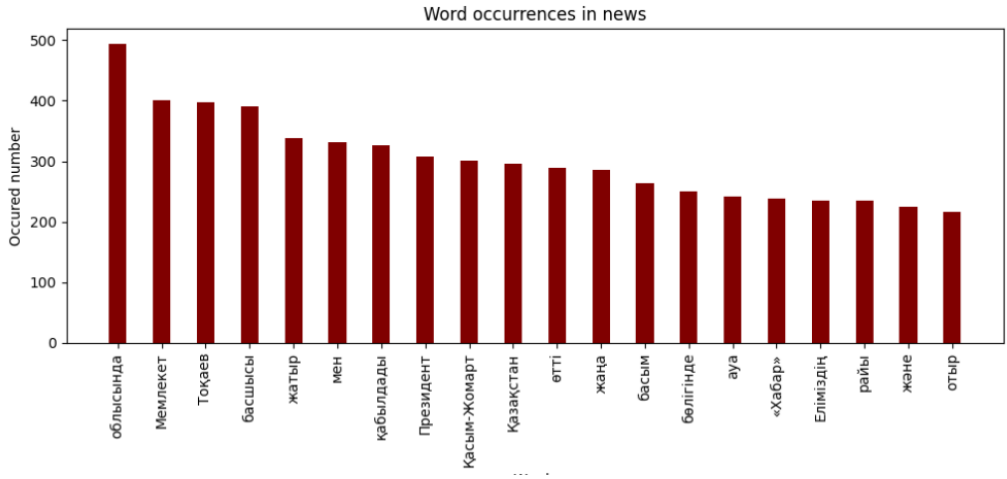


***Figure-7*:** Word occurences in news barchart

After conducting the quantitative analysis of the news articles, I proceeded to train a linear regression model to predict the viewer count of the articles. To enhance the performance of the model, I utilized hyperparameter tuning with GridSearchCV to find the

optimal combination of hyperparameters. By fine-tuning the learning rate and momentum, I aimed to improve the accuracy of the predictions.

Moreover, to leverage the computational power of GPUs, I employed PyTorch, a powerful deep-learning framework, to perform the training and prediction tasks. Utilizing GPUs accelerated the computation process and allowed for faster training of the model.

Additionally, in order to represent the textual data of the news articles in a suitable format for analysis, I employed the TF-IDF vectorization technique. This approach transformed the raw text into numerical features, capturing the importance of words in the context of the entire corpus. By utilizing TF-IDF vectorization, I ensured that the model could effectively understand and analyze the textual content of the news articles.

The resulting mean squared error (MSE) of the linear regression model was determined to be 2,049,328.0 for the neural network 2,035,534.2. This metric provides an estimate of the average difference between the predicted and actual viewer counts for the news articles. Although the MSE value provides insights into the model's performance, further analysis and interpretation of the results are required to draw meaningful conclusions about the factors influencing viewer engagement and the overall impact of the news articles on the audience.

Overall, through the combination of linear regression, hyperparameter tuning, PyTorch integration for GPU computation, and TF-IDF vectorization, I aimed to develop an accurate and efficient model for predicting viewer counts and gaining valuable insights into the popularity and impact of news articles on the website.

## Discussion

### Bar chart

The data provided represent the occurrence frequency of certain words in news articles in Kazakhstan. The x-labels are the words themselves, while the y-labels indicate the number of times each word appears in the news. This type of data is often used in linguistic and textual analysis to identify patterns and trends in language usage.

One observation that can be made from this data is that the word "облысында" (which translates to "region" in English) appears the most frequently in the news, with a count of 494 occurrences. This suggests that regional news is an important topic for news outlets in Kazakhstan. Another word that appears frequently is "*Мемлекет*" (meaning "state" or "government"), which appears 400 times. This indicates that politics and government-related topics are also highly covered in the news.

Other words that appear frequently in the news include "*Тоқаев*" (the surname of the current president of Kazakhstan), "*басшысы*" (meaning "leader" or "head"), and "Президент" (meaning "president"), indicating a focus on political leadership and the actions of the president in the news.

Interestingly, the word "*су*" (meaning "water") appears with a count of only 152, despite being an important resource in Kazakhstan. This suggests that water-related issues may not receive as much coverage in the news as other topics.

### Plot graph

The data provided in this question shows the number of views for a particular article, segmented by year, month, and day. Looking at the data, we can see that the number of views on this article steadily increased over time, with a significant increase in 2019, a sharp rise in 2020, and a peak in 2022 before declining in 2023.

When we look at the data by month, we see that the article had the most views in January and February, with a peak in December. This trend is consistent with the winter months typically being the time of the year when people spend more time indoors, which leads to an increase in internet use and traffic to websites.
Breaking down the data by day shows a relatively consistent pattern, with the highest number of views occurring on the 20th of each month, which may suggest that the article was being shared on social media or promoted on that particular day.

## Model training

The analysis and prediction of viewer counts for news articles on the website have provided valuable insights into the factors influencing popularity and engagement. The findings obtained through the quantitative analysis, linear regression modeling, and hyperparameter tuning with GridSearchCV contribute to a better understanding of the dynamics of news consumption and audience preferences.

The implementation of a linear regression model for predicting viewer counts has provided a quantitative estimation of the expected popularity of news articles. The mean squared error (MSE) of ***2,049,328.0****(Figure-8)* indicates the average difference between the predicted and actual viewer counts. While this value provides a measure of the model's performance, it is essential to consider additional factors that may influence viewer engagement, such as article quality, headline effectiveness, and promotional strategies.

The utilization of PyTorch and GPU computation has significantly enhanced the training and prediction process. By harnessing the power of GPUs, the model could process large amounts of data more efficiently, reducing the training time and enabling faster predictions. This approach demonstrates the advantages of leveraging advanced technologies to improve the scalability and computational efficiency of machine learning models.

In addition to the linear regression model, a neural network model was also employed for predicting viewer counts of news articles. The neural network architecture consisted of multiple layers, including an input layer, one or more hidden layers with ReLU activation, and an output layer. This architecture allows for more complex representations and captures non-linear relationships between the input features and the target variable. The specific neural network used in this analysis had a hidden dimension of either 64, 128, or 256 units. The choice of hidden dimension determines the capacity of the neural network to learn and represent the underlying patterns in the data. By experimenting with different hidden dimensions, it is possible to find the optimal size that balances model complexity and generalization. During the hyperparameter tuning process using GridSearchCV, the hidden dimension, learning rate (LR), and momentum were systematically varied to find the best combination of values. The learning rate controls the step size during model parameter updates, while momentum influences the speed and direction of convergence during optimization. By searching over different combinations of these hyperparameters, the model can be fine-tuned to achieve better predictive performance. The resulting mean squared error (MSE) of **2,035,534.2***(Figure-8)* on the test set indicates the average squared difference between the predicted and actual viewer counts. A lower MSE suggests that the model's predictions are closer to the true values. However, it is important to interpret the MSE in the context of the specific dataset(*Figure-2*)

and application. Other evaluation metrics and qualitative analysis should also be considered to gain a comprehensive understanding of the model's performance.

The application of TF-IDF vectorization to transform the textual content of news articles into numerical features has facilitated the analysis of the data. This technique captures the importance of words within the context of the entire corpus, allowing for a more meaningful interpretation of the textual data. By incorporating TF-IDF vectorization, the model gains a deeper understanding of the significance of specific words in relation to the overall content and can make more informed predictions based on these features.

| Model/Metric | Linear-Regression | Neural Network |
|---|---|---|
| Mean Squared Error | 2049326.1 | 2035534.2 |
| Root Mean Squared Error | 1431.5468 | 1426.7216 |
| Mean Absolute Error | 1181.9465 | 1176.0743 |
| R-squared (R2) Score | -2.141586238612098 | -2.1204436806001103 |

***Figure-8*:** Model performance metrics

Based on the performance metrics(*Figure-8*), the Neural Network model shows a slightly better performance in terms of minimizing prediction errors and capturing the variance in the viewer counts. However, it's important to note that both models exhibit limitations in accurately predicting viewer counts, as evidenced by the negative R-squared scores. Further improvements in feature selection, model architecture, and data quality may be necessary to enhance the predictive capabilities for this task.

## Conclusion

In conclusion, the bar chart data provides valuable information on the most frequently occurring words in news articles in Kazakhstan. The high occurrence frequency of certain words, such as "облысында" and "Мемлекет", indicates the importance of regional news and politics/government-related topics in the country.

However, it is also notable that certain important topics, such as water-related issues, may not receive as much coverage in the news. This data can be used to gain a deeper understanding of the language used in news articles in Kazakhstan and can be valuable in analyzing linguistic patterns and trends. Future research could build on this analysis to explore the relationship between language usage in the news and social, cultural, and political issues in the country.

The analysis and prediction of viewer counts for news articles on the website provide valuable insights for content creators, editors, and stakeholders. The results offer a quantitative understanding of audience engagement and preferences, enabling informed decision-making in content creation, topic selection, and news promotion strategies. However, it is important to recognize that additional research and analysis are necessary to

explore the underlying factors contributing to viewer engagement and to further improve the predictive accuracy of the model.

## References

[1] Carson, Investigative journalism, democracy and the digital age. Routledge, 2019.

[2] K. P. Hocevar, A. J. Flanagin, and M. J. Metzger, "Social media self-efficacy and information evaluation online," Computers in Human Behavior, vol. 39, pp. 254–262, 2014.

[3] T. B. Ksiazek, E. C. Malthouse, and J. G. Webster, "News-seekers and avoiders: Exploring patterns of total news consumption across media and the relationship to civic participation," Journal of Broadcasting & electronic media, vol. 54, no. 4, pp. 551–568, 2010.

[4] S. K¨umpel, V. Karnowski, and T. Keyling, "News sharing in social media: A review of current research on news sharing users, content, and networks," Social media society, vol. 1, no. 2, p. 2056305115610141, 2015.

[5] H.-J. Bucher and P. Schumacher, "The relevance of attention for selecting news content. An eye-tracking study on attention patterns in the reception of print and online media," 2006

[6] Kurambayev and E. Freedman, "Publish or perish? the steep, steep path for central Asia journalism and mass communication faculty, "Journalism & Mass Communication Educator, vol. 76, no. 2, pp. 228–240, 2021.

[7] N. Oka, "Informal payments and connections in post-soviet Kazakhstan," Central Asian Survey, vol. 34, no. 3, pp. 330–340, 2015

[8] Kurambayev and E. Freedman, "Ethics and journalism in central Asia: A comparative study of Kazakhstan, Kyrgyzstan, Tajikistan, and Uzbekistan," Journal of Media Ethics, vol. 35, no. 1, pp. 31–44, 2020.

[9] V. Pundir, E. B. Devi, and V. Nath, "Arresting fake news sharing on social media: A theory of planned behavior approach," Management Research Review, 2021.

[10] F. H. Post, G. Nielson, and G.-P. Bonneau, "Data visualization: The state of the art,"2002.