Intelligent information systems

UDC 004.048

doi: https://doi.org/10.20998/2522-9052.2023.4.09

Oleh Poliarush, Svitlana Krepych, Iryna Spivak

West Ukrainian National University, Ternopil, Ukraine

HYBRID APPROACH FOR DATA FILTERING AND MACHINE LEARNING INSIDE CONTENT MANAGEMENT SYSTEM

Abstract. The object of research is the processes of data filtering and machine learning in content management systems. **The subject** of research is developing a hybrid approach to data filtering based on a combination of supervised and unsupervised machine learning. The article explores machine learning approaches to content management and how they can change the way we organize, categorize, and derive value from vast amounts of data. **The main goal** is to develop and use a hybrid approach for data filtering and training that will help optimize resource consumption and perform supervised training for better categorization in the future. This approach includes elements of supervised and unsupervised learning using the BERT architecture that uses this kind of flow that help reduce resource usage and adjust the algorithm to perform better in a specific area. As a **result**, thanks to this approach, the intelligent system was able to independently optimize for a specific field of use and help to reduce the costs of using resources. **Conclusion**. After applying a hybrid approach of data filtering and machine learning to existing data streams, we obtain a performance increase of up to 5%, and this percentage increases depending on the running time of the application.

Keywords: software system; supervised learning; unsupervised learning; data streaming.

Introduction

In today's digital age, content has become a vital component for businesses and individuals. With the exponential growth of data, the need for efficient content management has never been more crucial. This article explores the approaches of machine learning in content management and how it can revolutionize the way we organize, categorize, and extract value from vast amounts of data. We will focus on the use of supervised and unsupervised learning algorithms for content management.

Usually data filtering goes with on model training steps and leads to funneling huge amount of data from captured source to almost the latest step. It is not very effective way to deal with data and model training because it requires additional costs from IO and communication modules, also in grid architecture efforts will grows significantly due to network issues. The best way to avoid this bottleneck is to introduce an additional data filtering step which use linear regression for data filtering [1–3].

The main innovation is using hybrid model for data filtering and learning which helps us to optimize resource consumption and do guided learning for better categorization in future. Supervised learning algorithms require labeled data, where human experts provide annotations or tags for different content trained to classify and categorize content accurately based on the provided labels. However, in many cases, obtaining labeled data can be time-consuming and expensive.

This is where unsupervised learning comes into play. Unlike supervised learning, unsupervised learning algorithms do not rely on labeled data. Instead, they uncover patterns, structures, and relationships within the data without any prior knowledge or guidance. In the context of content management, unsupervised learning algorithms can be employed to discover hidden patterns and group similar content together, even without explicit labels or categories (Fig. 1).

By utilizing unsupervised learning techniques such as clustering or size reduction, content management systems can automatically group similar documents, images, or videos together based on their inherent similarities. This can be particularly useful when dealing with large volumes of unstructured data where manually labeling each piece of content is impractical or infeasible [4–6].

Unsupervised learning algorithms can also assist in identifying anomalies or outliers within the content, helping to identify potentially valuable or suspicious data points that require further investigation.

By leveraging unsupervised learning, content management systems can gain valuable insights into the structure, distribution, and characteristics of the content, allowing for more effective organization and management [7–9].

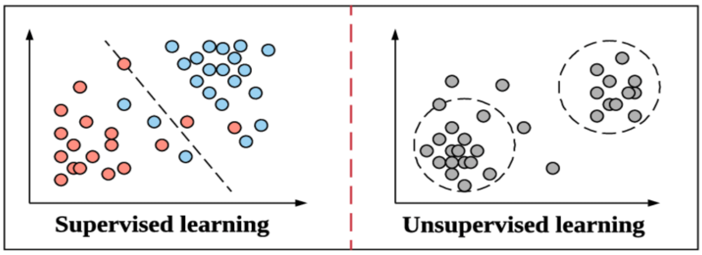


Fig. 1. Comprising between supervised and unsupervised learning approaches

In combination with supervised learning, unsupervised learning techniques can enhance the overall content management process. Unsupervised learning can be used as a preprocessing step to identify clusters or groups of content, which can then be further labeled or categorized using supervised learning algorithms. This hybrid approach leverages the strengths of both supervised and unsupervised learning to create a comprehensive and efficient content management system [10].

Statement for the task

The primary goal of content management is to facilitate the storage, retrieval, and manipulation of digital content. Traditionally, this process has been labor-intensive, requiring human manual work to organize and tag content. However, with advancements in machine learning, we can automate many of these tasks, reducing human error and significantly improving efficiency. In this article, we will focus on the implementation of both supervised and unsupervised learning algorithms for content management. Specifically, for supervised learning, the task at hand is to develop an optimized model that can accurately classify and categorize content based on provided labels. The challenges associated with this task include:

1. **Data Quality and Quantity.** Obtaining high-quality labeled data is essential for training a robust supervised learning model. However, acquiring a large and diverse dataset with accurate labels can be challenging. The task involves exploring strategies to ensure data quality, such as data cleaning, addressing label noise, and augmenting the dataset through techniques like data synthesis or transfer learning [11–13].

2. **Feature Engineering.** Effectively representing the content data plays a crucial role in the performance of the supervised learning model. The task includes identifying relevant features that capture the essence of the content, as well as employing techniques like dimensionality reduction to handle high-dimensional data. Feature engineering techniques, such as text normalization, word embeddings, or image feature extraction, need to be carefully considered to improve the model's accuracy and efficiency [14, 15].

3. **Model Selection and Optimization.** Choosing the appropriate supervised learning algorithm is essential for achieving accurate content classification and categorization. The task involves evaluating various algorithms such as support vector machines (SVM), random forests, or deep learning models like convolutional neural networks (CNNs) or transformer-based architectures (e.g., BERT, GPT) [4–6]. Additionally, model optimization techniques such as hyperparameter tuning, cross-validation, and regularization methods need to be applied to enhance the model's performance [16, 17].

4. **Evaluation Metrics.** To assess the effectiveness of the supervised learning model, appropriate evaluation metrics need to be selected. The task includes determining metrics such as accuracy, precision, recall, F1 score, or area under the receiver operating characteristic curve (AUC-ROC) based on the specific content management requirements. The chosen metrics should align with the desired outcomes of the content classification and categorization tasks. By addressing these challenges and implementing an optimized supervised learning model, we aim to enhance the content management process by automating the classification and categorization of content, thereby improving efficiency, reducing manual effort, and enabling more effective utilization of digital assets.

Main part

Transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) have gained significant attention in natural language processing tasks due to their ability to capture contextual information effectively [4–6]. In the context of content management, these architectures can be leveraged to enhance the accuracy and efficiency of content classification and categorization. Here is a step-by-step implementation guide (Fig. 2):

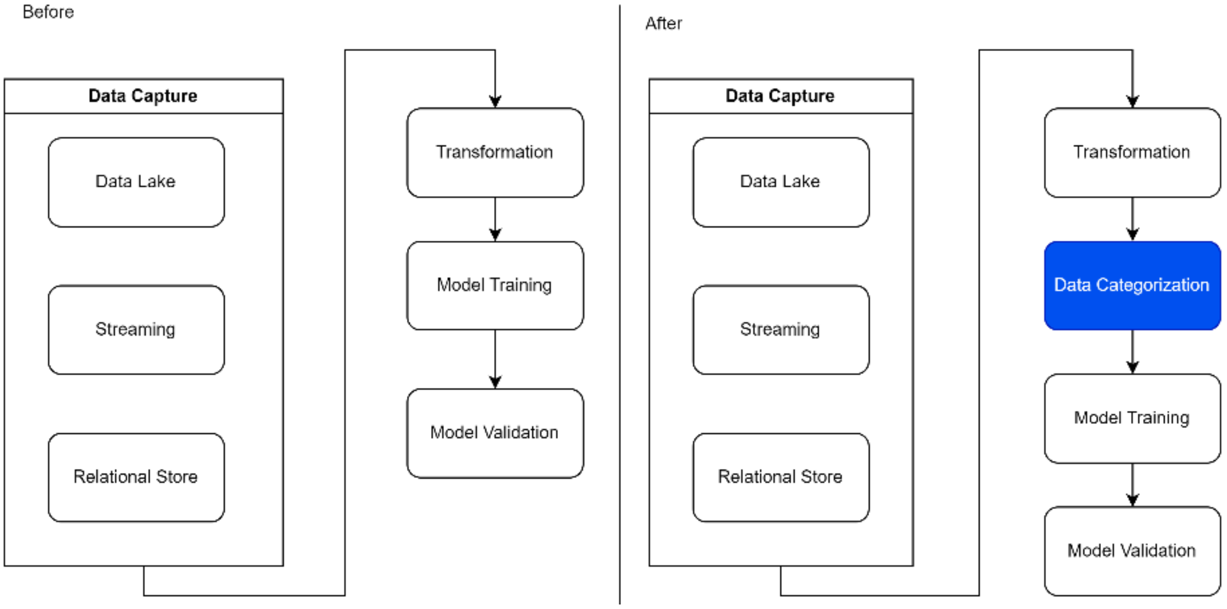


Fig. 2. Data flow chart

1. Data Preprocessing. Start by collecting and preprocessing the content data. Perform standard text preprocessing techniques such as lowercasing, tokenization, and removing stop words or special characters. Additionally, ensure that the text data is transformed into a format compatible with transformer-based architectures. For example, BERT requires tokenizing text into sub word units using a tokenizer specific to the pre-trained model.

2. Fine-tuning the Transformer-Based Architecture. Next, fine-tune the pre-trained transformer-based architecture for the content classification task. Fine-tuning involves training the transformer model on a labeled dataset specific to the content management task at hand. This dataset should consist of content samples labeled with their respective categories. The process involves feeding the preprocessed text data as input to the transformer model and optimizing the model's parameters using a suitable optimizer (e.g., Adam) and a loss function (e.g., cross-entropy loss).

3. Training and Validation. Split the labeled dataset into training and validation sets. Train the fine-tuned transformer-based architecture using the training data, adjusting the model's weights based on the loss calculated during each iteration. Monitor the model's performance on the validation set to prevent overfitting and determine when to stop training.

4. Evaluation. Evaluate the performance of the trained model using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, or AUC-ROC. Calculate these metrics by comparing the model's predicted categories with the ground truth labels from the validation set. Adjust the model and hyperparameters if necessary to improve performance.

5. Inference and Content Categorization. Once the model is trained and evaluated, it can be used for content categorization on new, unseen data. Feed the preprocessed text data into the fine-tuned transformer-based architecture and obtain predictions for the content's category (Fig. 3). These predictions can then be used to automatically categorize and organize the content within a content management system.

6. Iterative Refinement. Content management is an iterative process, and it is essential to continually monitor and refine the model's performance. Periodically retrain the model with updated labeled data or fine-tune it with additional tasks specific to the content management domain. This iterative refinement process helps to ensure the model remains accurate and up to date with evolving content categories.

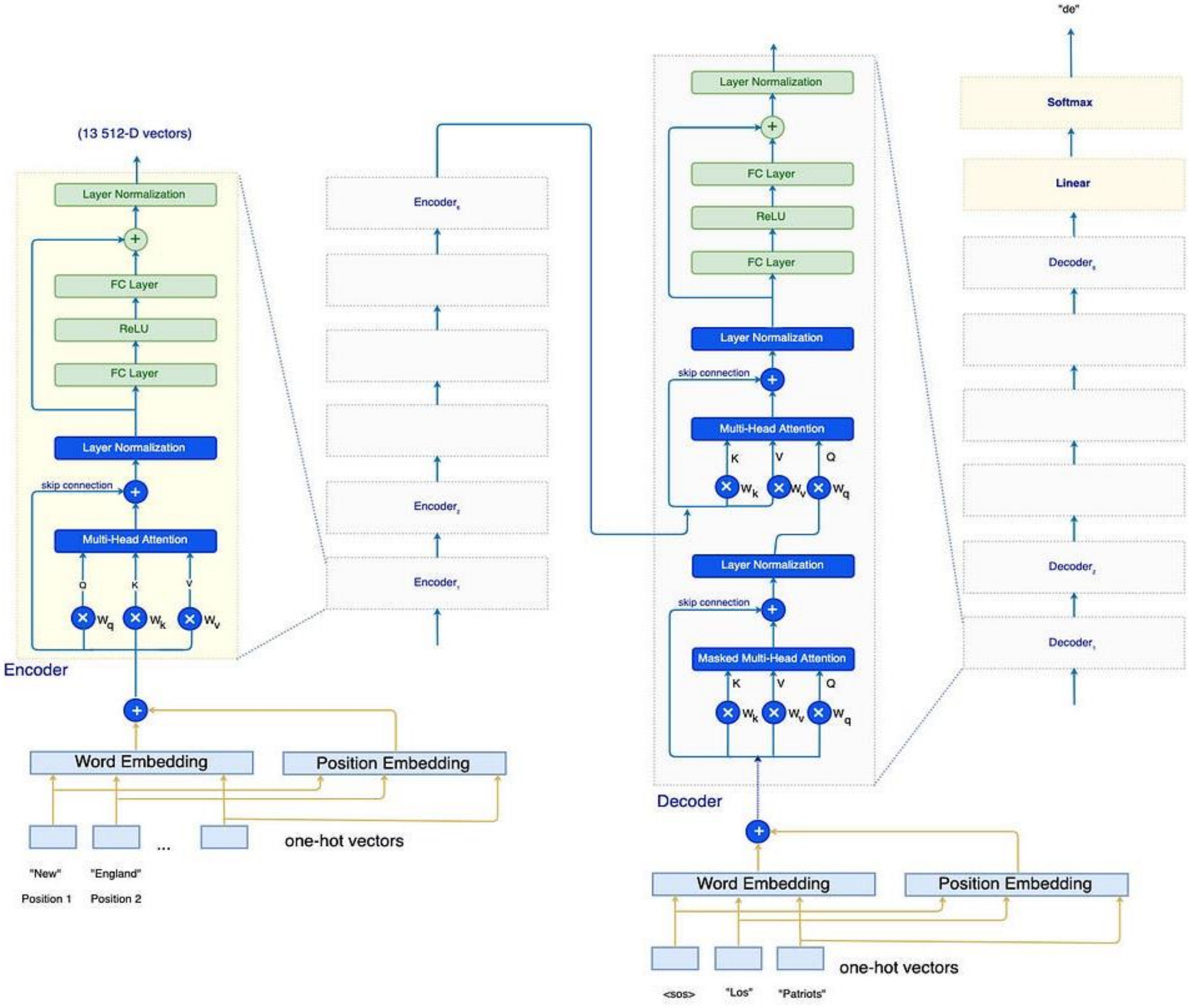
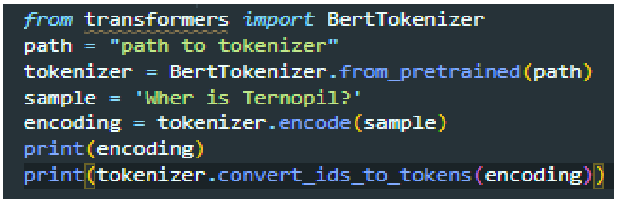


Fig. 3. Transformer overview

Using Filtering before training helps in reducing noise and outliers, leading to a cleaner and more reliable dataset. This enhances the quality of the training data and improves the accuracy and generalization of the machine learning models. Additionally, filtering reduces the computational burden by removing unnecessary data, making the training process more efficient.

Adding “Data Categorization” step can boost app performance and up to 20% removing data which is not valid or incorrect for further training. Data Categorization step use linear regressions for data filtering this is the most suitable approach for simple data analysis, it is less expensive than other methods. Also, “Model Training” and “Model validation” could have submodules which is responsible for data validation, the most suitable place to have it on “Model validation” step doing on this step we can encapsulate logic which is suitable for model validation and data correctness check.

By implementing transformer-based architectures for content management, we can benefit from the powerful contextual understanding capabilities of these models, allowing for more accurate and efficient content classification and categorization. The ability to leverage pre-trained transformer models like BERT [4–6] (Fig. 4) and GPT saves time and computational resources while achieving state-of-the-art performance in content management tasks (Tabl. 1).



```
from transformers import BertTokenizer
path = "path to tokenizer"
tokenizer = BertTokenizer.from_pretrained(path)
sample = 'Wher is Ternopil?'
encoding = tokenizer.encode(sample)
print(encoding)
print(tokenizer.convert_ids_to_tokens(encoding))
```

Fig. 4. Example of usage BERT Tokenizer.

Table 1 – **Performance comparison of using hybrid approach**

Total requests	AVG time request exec time (ms)	AVG time request exec time (Hybrid) (ms)
1000	1432	3312
5000	1331	3112
10000	1375	2902
50000	1573	2123
100000	1425	1523
200000	1497	1274
500000	1456	1029

Conclusion

Machine learning, particularly the integration of supervised and unsupervised learning techniques, has revolutionized content management, offering significant improvements in efficiency, accuracy, and user experience. Through the implementation of advanced machine learning models, such as transformer-based architectures like BERT, so now we can streamline our content management processes and extract greater value from their digital assets. Supervised learning models allow us to do automated content classification and categorization, reducing manual effort and enabling efficient organization of vast amounts of data.

With careful data preprocessing, feature engineering, model selection, and optimization, these models can achieve high accuracy in identifying and labeling content according to predefined categories. Unsupervised learning techniques, on the other hand, empower content management systems to automatically discover patterns, group similar content, and identify anomalies without the need for explicit labels.

Clustering algorithms and dimensionality reduction methods help to organize and make sense of unstructured data, allowing for more effective content retrieval and recommendation. By combining the strengths of supervised and unsupervised learning, we can enhance our content management systems further. Supervised learning models can leverage the insights gained from unsupervised learning to improve accuracy and efficiency in content classification and categorization tasks. This hybrid approach ensures the system benefits from both labeled data and the ability to discover hidden patterns within the content.

Transformer-based architectures have demonstrated their effectiveness in capturing contextual information and have emerged as powerful tools for content management. Fine-tuning these pre-trained models on specific content management datasets gives us the chance to leverage our state-of-the-art capabilities without the need for training from scratch. The adoption of machine learning in content management not only improves internal processes but also enhances user experiences. By automating content categorization and recommendation, organizations can provide personalized and relevant content to their users, fostering engagement and satisfaction.

After adopting this approach to existing data streaming flows, we get boos in productivity up to 5% and this percentage is growing based on amount of application running time. As you can see in a table that in the beginning of hybrid approach has some degradation compared to normal flow it happens due to additional calculation costs for model training, but some period of time when the model has trained enough average request execution time became much faster than normal.

References

1. Shen, Q. (2022), “A machine learning approach to predict the result of Leaque of Legends”, *International Conference on Machine Learning and Knowledge engineering* (MLKE), Guilin, China, doi: https://doi.org/10.1109/MLKE55170.2022.00013
2. Goyushova, U. (2023), “Algorithms for finding non-intersecting roads on images”, *Advanced Information Systems*, Vol. 7, no. 2, pp. 5–8, Jun. 2023, doi: https://doi.org/10.20998/2522-9052.2023.2.01
3. Subramanian, S., Tseng, B., Barbieri, R. and Browm E.N. (2021), “Unsupervised Machine Learning Methods for Artifact Removal in Electrodermal Activity”, in *43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (*EMBC), pp.399–402, doi: https://doi.org/10.1109/EMBC46164.2021.9630535
4. Devlin, J., Chang, K., Lee, M.-W. and Toutanova, K. (2019), “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL-HLT 2019), pp. 4171–4186, doi: https://doi.org/ 10.18653/v1/N19-1423

5. Dun, B., Zakovorotnyi, O. and Kuchuk, N. (2023), “Generating currency exchange rate data based on Quant-Gan model”, *Advanced Information Systems*, Vol. 7, no. 2, pp. 68–74, doi: https://doi.org/10.20998/2522-9052.2023.2.10
6. Guven, Z.A. (2021), “Comparison of BERT models and nachine learning methods for sentiment analysis on Turkish Tweets”, *6th Int. Conf. on ComputerScience and Eng.* (UBMK), pp. 98–101, doi: https://doi.org/10.1109/UBMK52708.2021.9559014
7. Radford., A., Wu, J., Child, R., Luan., D., Amodei, D. and Sutskever, I. (2019), “Language Models are Unsupervised Multitask Learners. OpenAI Blog”, available at: https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf
8. Kuchuk, N., Mozhaiev, O., Mozhaiev, M. and Kuchuk, H. (2017), “Method for calculating of R-learning traffic peakedness”, *2017 4th International Scientific-Practical Conference Problems of Infocommunications Science and Technology*, PIC S and T 2017 – Proceedings, pp. 359–362, doi: https://doi.org/10.1109/INFOCOMMST.2017.8246416
9. Tegen, A., Davidssom, P. and Persson, J.A. (2021), “Active Learning and Machine Teaching for Online Learning: A Study of Attention and Labelling Cost”, *20th IEEE International Conference on Machine Learning and Applications* (ICMLA), pp. 1215–1220, doi: https://doi.org/10.1109/ICMLA52953.2021.00197
10. Darouich, A., Khoukhi, F. and Douzi, K. (2015), “A dynamic learning content pattern for adaptive learning environment”, *10th Int. Conf. on Intelligent Systems: Theories and Applications (SITA)*, pp. 1–6, doi: https://doi.org/10.1109/SITA.2015.7358428
11. Bobalo, Yu., Dyvak, M., Krepych, S. and Stakhiv, P. (2014), “Evaluation of functional device suitability, with considering of random technological deviations of the parameters from the nominal and process of component aging”, *Przegld Elektrotechniczny, Warszawa, Poland*, Vol. 2014, No. 4. pp. 224–228, doi: https://doi.org/10.12915/pe.2014.04.54
12. Kovalenko, A., Kuchuk, H., Kuchuk, N. and Kostolny, J. (2021), “Horizontal scaling method for a hyperconverged network”, *2021 International Conference on Information and Digital Technologies* (IDT), Zilina, Slovakia, doi: https://doi.org/10.1109/IDT52577.2021.9497534
13. Spivak, I., Krepych, S., Litvynchuk, M. and Spivak, S. (2021), “Validation and data processing in JSON format”, *19th IEEE Int. Conf. on ST, Proc.* (EUROCON2021), pp. 326–330, doi: https://doi.org/10.1109/EUROCON52738.2021.9535582
14. Kovalenko, A. and Kuchuk, H. (2022), “Methods to Manage Data in Self-healing Systems”, Studies in Systems, Decision and Control, Vol. 425, pp. 113–171, doi: https://doi.org/10.1007/978-3-030-96546-4_3
15. Sanzharovskyi, A. and Yurchyshyn, V. (2023), “A modified method of detecting fake news based on machine learning algorithms”, *Bulletin of the Cherkasy State Technological University*, Vol. 2, pp. 58–70, (in Ukrainian), doi: https://doi.org/10.24025/ 2306-4412.2.2023.279984
16. Yaloveha, V., Hlavcheva, D., Podorozhniak, A. and Kuchuk, H. (2019), “Fire hazard research of forest areas based on the use of convolutional and capsule neural networks”, 2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering, UKRCON 2019 – Proceedings, DOI: http://dx.doi.org/10.1109/UKRCON.2019.8879867
17. Koriashkina, L. S. and Symonets, H. V. (2021), “Application of machine learning algorithms for processing comments from the youtube video hosting under training videos”, Science and Transport Progress, (6(90), pp. 33–42 (in Ukrainian), doi: https://doi.org/10.15802/stp2020/225264

Received (Надійшла) 16.09.2023
Accepted for publication (Прийнята до друку) 29.11.2023

ВІДОМОСТІ ПРО АВТОРІВ/ ABOUT THE AUTHORS

Поляруш Олег Віталійович – аспірант кафедри комп’ютерних наук, Західноукраїнський національний університет, Тернопіль, Україна;
Oleh Poliarush – graduate student of Computer Science Department, Western Ukrainian National University, Ternopil, Ukraine;
e-mail**:** ovpoliar@gmail.com; ORCID ID: https://orcid.org/0009-0002-2254-0123.

Крепич Світлана Ярославівна – кандидат технічних наук, доцент, доцент кафедри комп’ютерних наук, Західноукраїнський національний університет, Тернопіль, Україна;
Svitlana Krepych – Candidate of Technical Sciences, Associate Professor, Associate Professor of Computer Science Department, Western Ukrainian National University, Ternopil, Ukraine;
e-mail: msya220189@gmail.com; ORCID ID: https://orcid.org/0000-0001-7700-8367.

Співак Ірина Ярославівна – кандидат технічних наук, доцент, доцент кафедри комп’ютерних наук, Західноукраїнський національний університет, Тернопіль, Україна;
Iryna Spivak – Candidate of Technical Sciences, Associate Professor, Associate Professor of Computer Science Department, Western Ukrainian National University, Ternopil, Ukraine;
e-mail: spivak.iruna@gmail.com: ORCID ID: https://orcid.org/0000-0003-4831-0780.

Гібридний підхід до фільтрації даних та машинного навчання в системі управління контентом

О. В. Поляруш, С. Я. Крепич, І. Я. Співак

Анотація. Об’єктом дослідження є процеси фільтрації даних та машинного навчання в системах управління контентом. **Предметом дослідження** є розробка гібридного підходу до фільтрації даних на основі поєднання контрольованого та неконтрольованого машинного навчання. У статті досліджуються підходи машинного навчання до керування вмістом і те, як вони можуть змінити спосіб організації, категоризації та отримання цінності від величезних обсягів даних. **Основною метою** є розробка та використання гібридного підходу для фільтрації даних і навчання, який допоможе оптимізувати споживання ресурсів і проводити навчання під наглядом для кращої категоризації в майбутньому. Цей підхід включає елементи контрольованого та неконтрольованого навчання з використанням архітектури BERT, яка використовує цей вид потоку, що допомагає зменшити використання ресурсів і налаштувати алгоритм для кращої роботи в певній області. В результаті завдяки такому підходу інтелектуальна система змогла самостійно оптимізуватись під конкретну сферу використання та допомогти знизити витрати на використання ресурсів. **Висновок:** після застосування гібридного підходу фільтрації даних і машинного навчання до існуючих потоків даних ми отримуємо збільшення продуктивності до 5%, і цей відсоток збільшується залежно від часу роботи програми.

Ключові слова: програмна система; контрольоване навчання; неконтрольоване навчання; потокове передавання даних.