Evaluating the effectiveness of machine learning methods for keyword coverage using semantic data analysis

Anargul Shaushenova[1], Aigulim Bayegizova[2], Gulnaz Baidrakhmanova[3], Zhanargul Abuova[4], Akmaral Kassymova[5], Dana Bakirova[6], Yekaterina Golenko[7]

[1]Department of Information System, S. Seifullin Kazakh Agro Technical Research University, Astana, Republic of Kazakhstan
[2]Department of Radio Engineering, Electronics and Telecommunications, L. N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan
[3]Department of Computer Science and Information Technologies, K. Zhubanov Aktobe Regional University, Aktobe, Republic of Kazakhstan
[4]Higher School of Information Technology, Zhangir Khan West Kazakhstan Agrarian and Technical University, Uralsk, Republic of Kazakhstan
[5]Department of Information Technology, Zhangir Khan University, Uralsk, Republic of Kazakhstan
[6]Department of Construction, Institute of Architecture and Construction, L. N. Gumilyov Eurasian National University, Astana, Republic of Kazakhstan
[7]Department of Information Systems, S. Seifullin Kazakh Agrotechnical Research University, Astana, Republic of Kazakhstan

Article Info

Article history:

Received Jun 18, 2024
Revised Aug 18, 2024
Accepted Sep 3, 2024

Keywords:

Bidirectional encoder representations from transformers
Hybrid methods
Inverse document frequency
Keyword extraction
Semantic data analysis
Term frequency

ABSTRACT

This article presents a comprehensive comparative analysis of two advanced hybrid machine learning approaches for keyword extraction: bidirectional encoder representations from transformers (BERT) combined with autoencoder (AE) and term frequency-inverse document frequency (TF-IDF) combined with autoencoder. The research targets the task of semantic analysis in text data to evaluate the effectiveness of these methods in ensuring adequate keyword coverage across diverse text corpora. The study delves into the architecture and operational principles of each method, with a particular focus on the integration with autoencoders to enhance the semantic integrity and relevance of the extracted keywords. The experimental section provides a detailed performance analysis of both methods on various text datasets, highlighting how the structure and semantic richness of the source data influence the outcomes. The evaluation methodology includes precision, recall, and F1-score metrics. The paper discusses the advantages and disadvantages of each approach and their suitability for specific keyword extraction tasks. The findings offer valuable insights for the scientific community, aiding in the selection of the most appropriate text processing method for applications requiring deep semantic understanding and high accuracy in information extraction.

Corresponding Author:

Aigulim Bayegizova
Department of Radio Engineering, Electronics and Telecommunications, L. N. Gumilyov Eurasian National University
010000 Astana, Republic of Kazakhstan
Email: baegiz_a@mail.ru

1. INTRODUCTION

In modern text processing [1]–[3], the task of keyword extraction [4]–[6] has become increasingly significant, playing a crucial role in the organization, search, and analysis of information [7], [8]. Effective keyword extraction [9], [10] enhances the performance of search engines, recommendation systems, as well as analytical and educational tools. Therefore, the development and comparison of machine learning methods capable of automating and optimizing this process is a relevant research task. This work conducts an in-depth analysis of two hybrid approaches: bidirectional encoder representations from transformers (BERT)+autoencoder (AE) and term frequency-inverse document frequency (TF-IDF)+autoencoder. Both methods leverage modern advances in machine learning [11], [12] and natural language processing to handle large volumes of textual data [13]. The BERT model [14], [15], developed by Google, represents a cutting-edge technology for understanding the semantics of words in a text, while TF-IDF [16], [17] is a traditional statistical method for assessing the importance of a word in a document and collection. Integrating these approaches with autoencoders, which efficiently compress and reconstruct data, enhances the quality and semantic richness of the extracted keywords.

The purpose of our study is to compare these two methods in keyword extraction to evaluate their ability to adequately cover keywords in different text corpora. We analyze the architecture, operating principles, and integration features of each method with autoencoders, and conduct experimental comparisons on several text datasets. This allows us to identify differences in their effectiveness, depending on the structure and semantics of the source data. Particular attention is paid to the methodology for assessing the quality of keyword extraction, including accuracy, recall and F1-measure. The results of our study provide important input to the scientific community and can help in selecting the most suitable method for specific applications that require deep semantic analysis and high accuracy in text processing.

Yang *et al.* [18] discusses a theoretical framework for analyzing semantic perception performance using a stochastic geometry tool. The method helps to understand the performance of semantic networks from a macroscopic perspective. First, contextual space is transformed into semantic space during the process of semantic perception. Secondly, for text data of different styles (literary and scientific), two typical semantic reflection models are created using stochastic geometry theory, and keyword covering expressions are derived. Thirdly, using various parameters, the correctness and efficiency of the proposed models are verified through simulation results and analysis. Bhuyan *et al.* [19] examines the use of bibliometric analysis to explore new topic areas in the field of urban mobility. The authors improved the traditional method of keyword co-occurrence analysis by using the rapid keyword extraction (RAKE) algorithm to automatically extract keywords from document annotations and create a semantic similarity matrix between these keywords. The weighted co-occurrence matrix, combining frequency and semantic relatedness of keywords, demonstrated higher modularity and better quality of clusters compared to the unweighted one. This improvement has enabled the identification of more meaningful and relevant topic areas, which contributes to the further development of research in urban mobility. Goz and Multu [20] presents a keyword extraction technique called SkyWords, which combines supervised and non-supervised approaches. SkyWords uses the skyline operator and majority voting to select high-quality candidate keywords, and ranks them based on semantic similarity to the document using the MPNet sentence transformer. Experiments on six benchmark datasets showed that SkyWords significantly reduced the number of candidate keywords and provided improvements in precision, recall, and F1-score compared to baseline methods.

Sharma and Kumar [21] presents a new hybrid semantic document indexing system that uses machine learning techniques and domain ontology to improve information retrieval systems. The proposed method applies a skip-gram machine learning model with negative sampling and a domain ontology to identify concepts for annotating unstructured documents, and introduces an algorithm for ranking concepts based on multiple features, including statistical, semantic and scientific named entities. Experiments on five computer science benchmark datasets show that the proposed method outperforms state-of-the-art techniques, improving average accuracy by 29% and F-measure by 25%. Improved metrics confirm the system's ability to accurately extract document concepts even when the same concept is labeled with different terms, and to find similar concepts when terms are missing from the domain ontology. Breit *et al.* [22] explores a new sub-field of artificial intelligence, semantic web machine learning (SWeML), which combines machine learning components with techniques developed by the semantic web community. The authors conducted a systematic study and analyzed nearly 500 papers published over the past decade to evaluate the architectural and application features of SWeML systems. The analysis showed a rapidly growing interest in SWeML systems and their significant impact on various application areas. The main catalysts for this growth are the widespread use of deep learning and knowledge graph technologies. An important contribution of the article is the development and publication of a classification system for SWeML systems in the form of an ontology.

Additionally, this work becomes particularly relevant in the classification of documents with a large amount of data. The analyzed methods can be utilized to significantly improve the accuracy and efficiency of processing large-scale text sets. This improvement is crucial given the constant growth of information flows and the increasing need for rapid processing. By employing these hybrid approaches, organizations can better manage and analyze extensive collections of textual data, leading to more effective information retrieval and knowledge discovery. This relevance underscores the importance of continuous advancements in machine learning and natural language processing to keep pace with the expanding volume of data.

2. METHOD

In this study, we focused on two hybrid machine learning methods: BERT+autoencoder and TF-IDF+autoencoder, to evaluate their effectiveness in the task of extracting keywords from large text corpora using semantic data mining. Both methods integrate the concepts of deep learning and autoencoders [23]–[25], which allows not only to analyze texts at a surface level, but also to penetrate deep semantic and contextual relationships between words and phrases, enriching the process of keyword extraction. The first step in our research was to use the TF-IDF method, which allows us to evaluate the importance of each word in a document in relation to the entire text corpus. TF-IDF calculates word weights based on their frequency in a document and the inverse frequency of the documents in the corpus where they occur. This approach helps reduce the influence of frequently used but uninformative words by highlighting those that are unique to specific texts. The resulting vector representations of words served as input data for the subsequent stage of processing by the autoencoder. The autoencoder used in this work consists of two main components: an encoder and a decoder. The encoder compresses the vector representation of text obtained with TF-IDF into a denser and more informative internal representation. The decoder then works to reconstruct the original vector from this compressed representation, trying to minimize information loss. The goal of this stage is to train the model so that it can extract and retain the most significant semantic features from the source text, which improves the quality and accuracy of extracted keywords.

The use of autoencoders in combination with TF-IDF in Figure 1 significantly improved the semantic integrity of the selected keywords. This technique provides deeper analysis of text data, revealing hidden connections and meanings of words in a broader sense. Thus, the integration of these approaches contributes to the creation of more powerful and accurate text mining tools, which is an important step towards automating keyword extraction and improving information retrieval and big data processing.

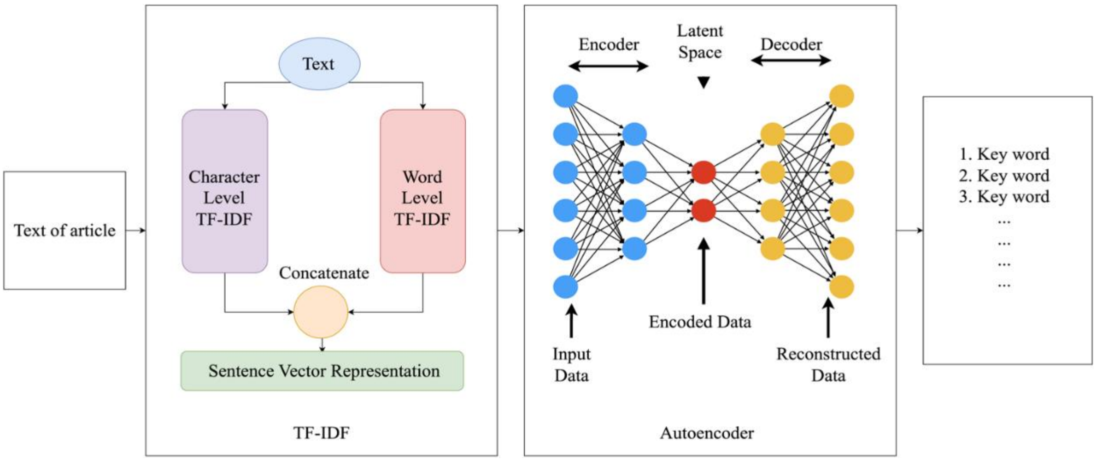


Figure 1. Architecture of the hybrid method TF-IDF+autoencoder

To improve text data analysis and more accurately extract keywords, this study used the BERT model in Figure 2. BERT is one of the advanced methods in the field of natural language processing that uses transformer attention mechanisms to analyze texts. A special feature of BERT is its ability to process text bidirectionally, which allows the model to simultaneously analyze the meaning of words both from left to right and from right to left. This bidirectional text understanding greatly enhances the model's ability to capture semantic relationships between words, providing deep insight into contextual nuances and improving the quality of keyword extraction. After processing the text using BERT, the next step was to use an autoencoder to compress the resulting vector representations. Autoencoders, consisting of an encoder and a decoder, work on the principle of minimizing information loss when moving from the original vector to the compressed representation and back. In our case, the encoder compressed the multidimensional BERT vectors into denser vectors, which the decoder then attempted to reconstruct. The purpose of this procedure was to highlight and retain the most significant information contained in the text, thereby ensuring high accuracy and quality in keyword extraction. This compression helped improve the manageability and analysis of large volumes of data, identifying the most important elements of text for later use in various information processing applications.

The use of combined BERT and TF-IDF methods with autoencoders in our study allowed us to more fully understand the importance of contextual text processing for keyword extraction. Particularly important was the bidirectional understanding of meaning that BERT provides. This allows you not only to analyze the sequence of the text, but also to take into account many contextual dependencies, which significantly enriches the perception of the text. Combining these technologies with autoencoders, which effectively compress information, enhances this effect, highlighting the most significant semantic attributes and improving the quality of extracted keywords. The experimental results confirmed that such an integrated application of methods not only increases the accuracy of keyword identification, but also contributes to a deeper analysis of texts. Using autoencoders to compress information helps avoid data overload and makes data processing more manageable and efficient. This is especially important when working with large text corpora, where each word and its meaning can have a significant impact on the outcome of the analysis. Ultimately, this approach not only improves the accuracy of keyword extraction, but also enriches the understanding of text structure and meaning, which is critical in many fields, including scientific research, knowledge management and information retrieval.

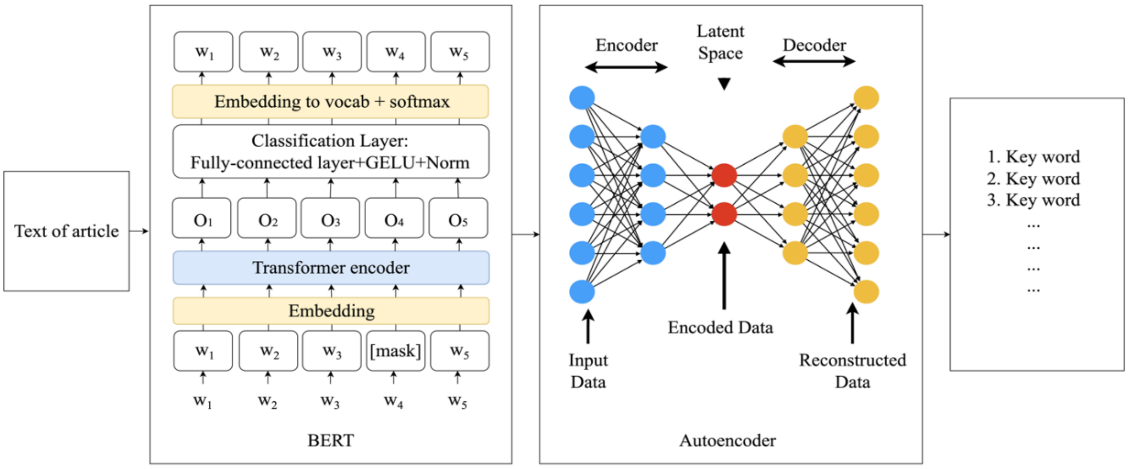


Figure 2. Architecture of the hybrid method BERT+autoencoder

3. RESULTS AND DISCUSSION

For the study, 182 scientific articles were used, covering various scientific fields. These areas include biology, computer science, physics, chemistry, psychology and linguistics. Each category represents a specific area of knowledge and includes approximately 20 to 25 documents. This variety of categories provides a comprehensive analysis and allows you to evaluate the effectiveness of keyword extraction methods in different scientific disciplines. The articles were selected in such a way as to evenly represent each scientific area, which allows us to draw more generalized conclusions about the applicability of the selected methods in various areas. To conduct the experiments, we used a variety of text datasets covering a variety of topics and writing styles. Evaluation methods included analysis of precision, recall, and F1-measures to evaluate the quality of keyword extraction. Keywords extracted using both methods were visualized using k-means and principal component analysis (PCA) methods to analyze the clustering and relative positions of words in vector space. These methods helped evaluate how keywords are grouped based on their semantic proximity and topic relevance.

Figure 3 shows the relationship between points can be interpreted as the degree of semantic or closeness between keywords. Clusters with closely grouped points may indicate more clearly defined themes or concepts in the data. The color scale on the right shows the weight or metric associated with the keywords, but its nature is difficult to determine without additional words. This could be the importance or frequency of keywords in the data set. Keyword clustering is presented in two graphs using k-means and PCA methods. The BERT model visualization shows a clearer and more explicit distribution of clusters, indicating a more expressive and differentiated vector representation of words. This suggests that the vectors produced by BERT better reflect the semantic similarities and differences between keywords, which is important for good clustering. The presented graph shows the results of keyword clustering using a combination of TF-IDF and autoencoders. For dimensionality reduction and subsequent visualization, k-means and PCA methods are used. The results show the formation of clusters with varying degrees of concentration and dispersion in two-dimensional space, which indicates the main trends in the semantic proximity of keywords in the test data set. The color scale can display additional metrics associated with each keyword.

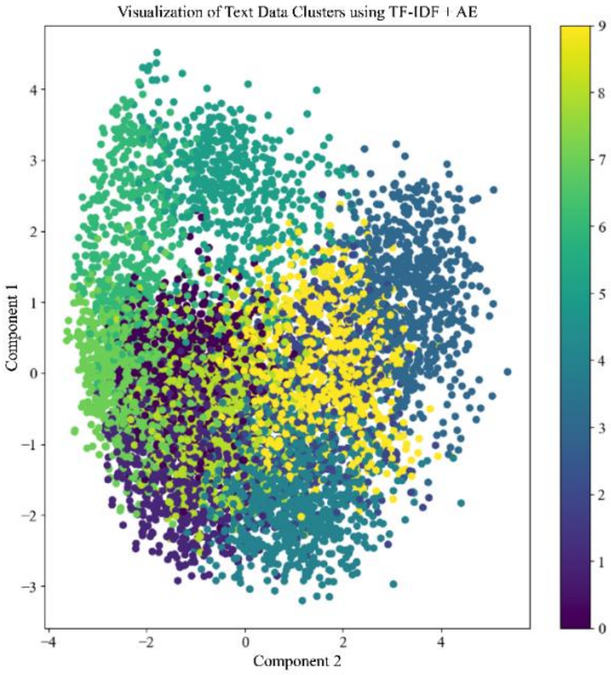
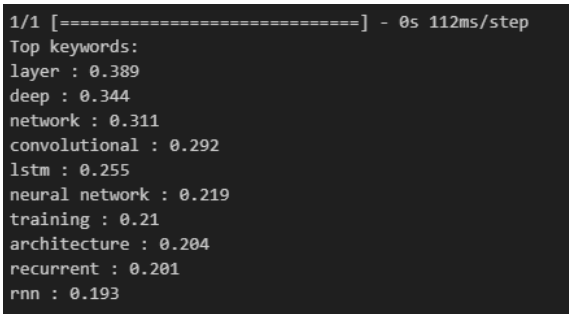


Figure 3. The result of keyword clustering using the TF-IDF+autoencoder method

Figure 4 shows the text result with a list of keywords and their corresponding percentages. This data is extracted from the text of scientific articles using the TF-IDF method in combination with an autoencoder, as described. TF-IDF, which stands for “document inverse term frequency,” is a statistical measure used to evaluate the importance of a word in a document that is part of a collection or corpus. Autoencoders can be used to reduce data size, helping to identify the most important details.



```
1/1 [==============================] - 0s 112ms/step
Top keywords:
layer : 0.389
deep : 0.344
network : 0.311
convolutional : 0.292
lstm : 0.255
neural network : 0.219
training : 0.21
architecture : 0.204
recurrent : 0.201
rnn : 0.193
```

Figure 4. The result of keyword classification using the TF-IDF+autoencoder method

Figure 5 shows the keyword clustering results obtained from the test data set using the Burt model adapted by the autoencoder. K-means and PCA methods were used to visualize the results in two-dimensional space. On the graph, each point corresponds to a keyword, the colors of the points indicate the cluster to which the keyword belongs, and their position is determined by the values of the first two principal components obtained using PCA.

In the figures, the metrics of the decision tree classifier model for the training and test data sets remained unchanged. This is because the tree returns predictions not as probabilities, but as integers. The accuracy of the models is high in the original tables with a threshold of 0.5, since in the data under study the number of one class significantly exceeds the number of another. Models are good at predicting bad customer data, but bad at predicting good ones. Therefore, it is advisable to rely on other indicators. After adjusting the threshold, the main metrics increased significantly for all models except the decision tree, indicating the positive impact of choosing the right thresholds.

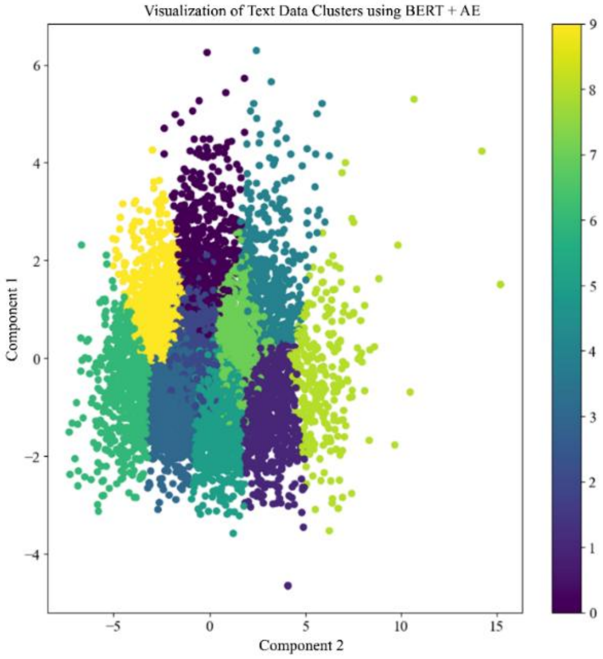
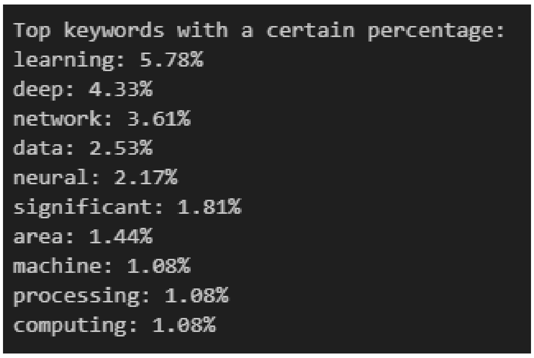


Figure 5. The result of keyword clustering using the BERT+autoencoder method

Figure 6 shows a list of keywords extracted from the text of a scientific article using a combination of the Burt model and an autoencoder, where each keyword is assigned a certain percentage. It shows the contribution of each keyword in the processed text. Words related to machine learning and artificial intelligence are at the top of the list. The presence of the word "learning" in first place with a share of 5.78%, followed by the word "deep" with a share of 4.33% shows that the topic of deep learning is given special attention in the analyzed text. The following terms "network," "data," and "neural" support this by focusing on neural networks and data. The keywords "significant," "area," "machine," "processing," and "computing" can also play a role in the content of the article, but indicate fewer additional aspects or characteristics compared to the main topic.



```
Top keywords with a certain percentage:
learning: 5.78%
deep: 4.33%
network: 3.61%
data: 2.53%
neural: 2.17%
significant: 1.81%
area: 1.44%
machine: 1.08%
processing: 1.08%
computing: 1.08%
```

Figure 6. Classification result using the BERT+autoencoder method

The results of the study show that the BERT method combined with an autoencoder demonstrates better results compared to the TF-IDF method combined with an Autoencoder in the task of keyword extraction. This demonstrates the ability of the BERT+autoencoder method to gain a deeper understanding of natural language. It also adapts better to complex text data. The training history of the models is represented by graphs showing the change in error during training and testing. The model plots show that the error in both the training and testing phases decreased steadily and consistently across all epochs. This demonstrates the high generalization ability of the BERT model, which is able to reliably learn information without obvious signs of overfitting. In contrast, no data is presented for the TF-IDF model, but based on general trends in machine learning, it can be assumed that TF-IDF may perform well on certain types of data, but may be inferior to the BERT model in complex natural language processing tasks. Figure 7 is a graph showing the comparative results of two different methods for extracting keywords from scientific articles using a combination of TF-IDF and autoencoders. The graph shows two lines: “read error” and “read error validation,” where the x-axis is labeled “epochs,” indicating the number of iterations, or passes, of the learning algorithm through the data set. The graph displays the training error and validation error measured and recorded at specific stages. Judging by the table, the training error decreases sharply at the initial stage, indicating that the model quickly improves in reproducing the target keywords from the training data. However, the validation error remains relatively constant throughout the process, which may indicate overfitting or the model's inability to improve performance on new, unknown data. This highlights the importance of monitoring both errors, as significant improvements in training error without improvements in testing error may indicate a need to refine the model to improve its generalizability and perhaps adjust or obtain other data sets for testing. It is also important to note that the graph does not provide information about error values, since the Y-axis is labeled “quantitative error” and has a scale of 0 to 0.2 with no additional labels indicating absolute error values, meaning that to accurately explain errors and performance methods requires additional text or data.

In Figure 8, the vertical axis (Y) is labeled “catalyst mass” and represents the error value represented on a negative logarithmic scale, as seen by the prefix “1e13” in the upper left corner, making the error distribution more visual, especially over a large range of values. Both curves have the same shape, indicating that the training error and the testing error decrease in close correlation during the training process. This is a positive sign that the model not only learns the training data well, but also adequately generalizes the acquired knowledge to data not included in the training set. The similarity of the trajectories of these curves may indicate that the model's training process was constant and there was no overfitting, which is a common problem in deep learning where the model performs well on training data but poorly on unprecedented data.

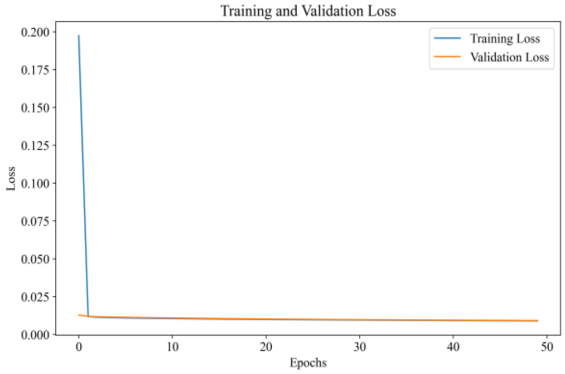


Figure 7. Training and validation result for TF-IDF+autoencoder, showing error dynamics during training

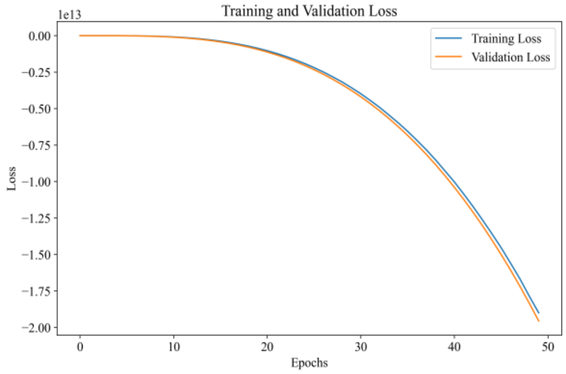


Figure 8. Training result for BERT+autoencoder, showing improvement in model quality as training progresses

The keyword extraction results from both methods show that both models identify the same terms as meaningful. However, with a deeper understanding of word semantics, BERT can identify more nuanced and contextually relevant keywords, making it suitable for keyword extraction tasks from complex scientific texts. BERT is well established for its ability to analyze texts, making it very effective at extracting keywords. Unlike TF-IDF, which considers each word in isolation, BERT takes into account bidirectional relationships between words, which allows you to more accurately understand their meaning in the text. This is especially important when analyzing scientific articles, where the same terms may have different meanings in different categories. In conclusion, based on the analysis, it can be concluded that the BERT method with autoencoder outperforms TF-IDF with autoencoder in the task of extracting keywords from scientific texts. BERT not only provides better clustering and keyword extraction, but also shows a consistent reduction in errors during training and testing. With its powerful word representation and deep semantic analysis capabilities, BERT is the preferred choice for processing complex scientific materials where the precise meaning of each term is critical.

4. CONCLUSION

This article provides a thorough comparative analysis of two modern hybrid approaches in machine learning for extracting keywords from scientific texts: BERT in combination with autoencoder and TF-IDF also in combination with an autoencoder. Based on the conducted research, the following conclusions can be drawn. First, the results showed that BERT+autoencoder performed better than TF-IDF+autoencoder in the keyword extraction task. This demonstrates the BERT model's ability to better understand natural language and better adapt to complex text data. BERT's bidirectional text processing allows the model to take into account relationships between words, which greatly improves the accuracy and relevance of extracted keywords.

Second, the training history of the models, represented by graphs, shows that the error in both the training and testing phases for the BERT model decreased steadily in all epochs, demonstrating its high generalization ability without signs of overfitting. In the case of the TF-IDF model, a similar trend was not revealed, which may indicate its lower efficiency in processing complex texts. Thirdly, visualization of keyword clustering results using K-means and PCA methods confirmed the advantage of the BERT model. Vector representations of words obtained using BERT better reflect the semantic similarities and differences between keywords, which facilitates more accurate clustering and in-depth text mining. Thus, we can conclude that the BERT method with autoencoder is superior to TF-IDF with autoencoder in the task of extracting keywords from scientific texts. BERT not only provides better clustering and keyword extraction, but also shows a consistent reduction in errors during training and testing. With its powerful word representation and deep semantic analysis capabilities, BERT is the preferred choice for processing complex scientific materials where the precise meaning of each term is critical.

REFERENCES

[1] Z. Sadirmekova *et al.*, “Ontology engineering of automatic text processing methods,” *International Journal of Electrical and Computer Engineering*, vol. 13, no. 6, pp. 6620–6628, 2023, doi: 10.11591/ijece.v13i6.pp6620-6628.

[2] I. Polo-Blanco, M. J. González López, A. Bruno, and J. González-Sánchez, “Teaching students with mild intellectual disability to solve word problems using schema-based instruction,” *Learning Disability Quarterly*, vol. 47, no. 1, pp. 3–15, 2024, doi: 10.1177/07319487211061421.

[3] Y. Suzuki, H. Jeong, H. Cui, K. Okamoto, R. Kawashima, and M. Sugiura, “An fMRI validation study of the word-monitoring task as a measure of implicit knowledge: Exploring the role of explicit and implicit aptitudes in behavioral and neural processing,” *Studies in Second Language Acquisition*, vol. 45, no. 1, pp. 109–136, 2023, doi: 10.1017/S0272263122000043.

[4] P. Tiwari, S. Chaudhary, D. Majhi, and B. Mukherjee, “Comparing research trends through author-provided keywords with machine extracted terms: a ML algorithm approach using publications data on neurological disorders,” *Iberoamerican Journal of Science Measurement and Communication*, vol. 3, no. 1, 2023, doi: 10.47909/ijsmc.36.

[5] L. P. Hung and S. Alias, “Beyond sentiment analysis: a review of recent trends in text based sentiment analysis and emotion detection,” *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 27, no. 1, pp. 84–95, 2023, doi: 10.20965/jaciii.2023.p0084.

[6] C. P. Chai, “Comparison of text preprocessing methods,” *Natural Language Engineering*, vol. 29, no. 3, pp. 509–553, 2023, doi: 10.1017/S1351324922000213.

[7] U. E. Chigbu, S. O. Atiku, and C. C. Du Plessis, “The science of literature reviews: searching, identifying, selecting, and synthesising,” *Publications*, vol. 11, no. 1, 2023, doi: 10.3390/publications11010002.

[8] F. Guo, C. M. Gallagher, T. Sun, S. Tavoosi, and H. Min, “Smarter people analytics with organizational text data: demonstrations using classic and advanced NLP models,” *Human Resource Management Journal*, vol. 34, no. 1, pp. 39–54, 2024, doi: 10.1111/1748-8583.12426.

[9] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, “Text preprocessing for text mining in organizational research: Review and recommendations,” *Organizational Research Methods*, vol. 25, no. 1, pp. 114–146, 2022, doi: 10.1177/1094428120971683.

[10] D. L. Cogburn, M. J. Hine, N. Peladeau, and V. Y. Yoon, “Text mining in big data analytics,” *Proceedings of the Annual Hawaii International Conference on System Sciences*, vol. 2018-January, pp. 584–586, 2018.

[11] A. M. H. Taha, S. B. Binti Ariffin, and S. S. Abu-Naser, “A systematic literature review of deep and machine learning algorithms in brain tumor and meta-analysis,” *Journal of Theoretical and Applied Information Technology*, vol. 101, no. 1, pp. 21–36, 2023.

[12] D. G. Murray, J. Simsa, A. Klimovic, and I. Indyk, “tf. data: a machine learning data processing framework,” *arXiv preprint arXiv:2101.12127*, 2021.

[13] M. Bayer, M. A. Kaufhold, B. Buchhold, M. Keller, J. Dallmeyer, and C. Reuter, “Data augmentation in natural language processing: a novel text generation approach for long and short text classifiers,” *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 1, pp. 135–150, 2023, doi: 10.1007/s13042-022-01553-3.

[14] A. T. G. Tapeh and M. Z. Naser, “Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices,” *Archives of Computational Methods in Engineering*, vol. 30, no. 1, pp. 115–159, 2023, doi: 10.1007/s11831-022-09793-w.

[15] J. A. Benítez-Andrades, J. M. Alija-Perez, M. E. Vidal, R. Pastor-Vargas, and M. T. García-Ordas, “Traditional machine learning models and bidirectional encoder representations from transformer (BERT)–based automatic classification of tweets about eating disorders: Algorithm development and validation study,” *JMIR Medical Informatics*, vol. 10, no. 2, 2022, doi: 10.2196/34492.

[16] M. Z. Naeem, F. Rustam, A. Mehmood, Mui-zzud-din, I. Ashraf, and G. S. Choi, “Classification of movie reviews using term frequency-inverse document frequency and optimized machine learning algorithms,” *PeerJ Computer Science*, vol. 8, 2022, doi: 10.7717/PEERJ-CS.914.

[17] R. K. Dey and A. K. Das, “Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis,” *Multimedia Tools and Applications*, vol. 82, no. 21, pp. 32967–32990, 2023, doi: 10.1007/s11042-023-14653-1.

[18] Y. Yang *et al.*, “Semantic sensing performance analysis: assessing keyword coverage in text data,” *IEEE Transactions on Vehicular Technology*, vol. 72, no. 11, pp. 15133–15137, 2023, doi: 10.1109/TVT.2023.3282742.

[19] A. Bhuyan, K. Sanguri, and H. Sharma, “Improving the keyword co-occurrence analysis: An integrated semantic similarity approach,” in *2021 IEEE International Conference on Industrial Engineering and Engineering Management*, 2021, pp. 482–487, doi: 10.1109/IEEM50564.2021.9673030.

[20] F. Goz and A. Mutlu, “SkyWords: an automatic keyword extraction system based on the skyline operator and semantic similarity,” *Engineering Applications of Artificial Intelligence*, vol. 123, 2023, doi: 10.1016/j.engappai.2023.106338.

[21] A. Sharma and S. Kumar, “Machine learning and ontology-based novel semantic document indexing for information retrieval,” *Computers and Industrial Engineering*, vol. 176, 2023, doi: 10.1016/j.cie.2022.108940.

[22] A. Breit *et al.*, “Combining machine learning and semantic web: a systematic mapping study,” *ACM Computing Surveys*, vol. 55, no. 14 S, 2023, doi: 10.1145/3586163.

[23] P. Li, Y. Pei, and J. Li, “A comprehensive survey on design and application of autoencoder in deep learning,” *Applied Soft Computing*, vol. 138, May 2023, doi: 10.1016/j.asoc.2023.110176.

[24] A. Scheinker, F. Cropp, and D. Filippetto, “Adaptive autoencoder latent space tuning for more robust machine learning beyond the training set for six-dimensional phase space diagnostics of a time-varying ultrafast electron-diffraction compact accelerator,” *Physical Review E*, vol. 107, no. 4, 2023, doi: 10.1103/PhysRevE.107.045302.

[25] S. Barwey, V. Shankar, V. Viswanathan, and R. Maulik, “Multiscale graph neural network autoencoders for interpretable scientific machine learning,” *Journal of Computational Physics*, vol. 495, 2023, doi: 10.1016/j.jcp.2023.112537.

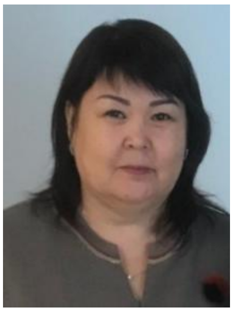
BIOGRAPHIES OF AUTHORS

Anargul Shaushenova in 1999, she graduated from the West Kazakhstan Humanitarian University with a degree in mathematics and computer science. In 2007, she graduated from the graduate school of Taraz State University with a degree in Geo-ecology. In 2010, she defended her Ph.D. thesis on the topic “Theoretical assessment of hydro-meteorological regimes and the ecological state of industrial cities (on the example of Balkhash)” (scientific supervisors T. Omarbekuly, B.B. Bakirbayev). Currently, she is a candidate of technical sciences, associate professor of the Department of Information Systems of the S. Seifullin Kazakh Agrotechnical Research University. She is the author of more than 40 scientific papers. Her research interests include mathematical modeling, geoinformation systems. She can be contacted at email: shaushenova_78@mail.ru.

Aigulim Bayegizova graduated from S.M. Kirov Kazakh State University in 1982 with a degree in Applied Mathematics. In 2010, she defended her Ph.D. thesis in the specialty “01.01.02 – differential equations and mathematical physics” and received the degree of Candidate of Physical and Mathematical Sciences. She began her career in 1982 as an assistant at the Department of Higher Mathematics of the Dzhezkazgan branch of the Karaganda Polytechnic Institute. Currently, he is a senior lecturer at the Department of Radio Engineering, Electronics and Telecommunications of the L.N. Gumilyov Eurasian National University. She is the author of more than 60 scientific papers, including 1 monograph, 6 articles in the Scopus database. Her research interests are programming, information security and information protection, artificial intelligence, and cloud technologies. She can be contacted at email: baegiz_a@mail.ru.

Gulnaz Baidrakhmanova in 2004, she graduated from Aktobe State University named after K. Zhubanov in 2004 with a degree in physics and computer science. In 2007 she received a master's degree in computer science. In 2018, she graduated from doctoral studies at the Abai Kazakh National Pedagogical University, majoring in 6D011100 – “Informatics”. From 2019 to the present, he is a Doctor of Philosophy Ph.D. in specialty 6D011100 - “Informatics” K. Zhubanov Aktobe Regional University. She is the author of more than 43 works. Her research interests include the development of mobile applications in the field of computer graphics, animated graphics and training of future computer science teachers in the context of fundamentalization of education, sensor networks, and mobile communication technologies. She can be contacted at email: gulnaztai83@mail.ru.

Zhanargul Abuova teaches at the Zhangir Khan West Kazakhstan Agrarian and Technical University, the main research interests of the master Zhanargul Abuova lie in the field of information technology, digitalization of the agro-industrial complex, remote sensing of the Earth and satellite engineering. Over the past 4 years, she has published 12 articles, including 5 in journals included in the Scopus database. She can be contacted at email: Zhanargul81@mail.ru.



Akmaral Kassymova candidate of Pedagogical Sciences, currently works at Zhangir Khan University at the Department of Information Technology, Professor. He has more than 39 years of scientific and pedagogical experience and more than 100 scientific papers, including 9 articles in the Scopus database, 20 teaching aids. She can be contacted at email: kasimova_ah@mail.ru.

Dana Bakirova graduated from Karagandy State University in 1990 with a major in “Construction” at Architecture and Construction Faculty. From 2005 until 2021 she worked as a Senior Lecturer at the department of “Construction materials and technology” at Karagandy State Technical University. In 2019, she obtained her Master’s Degree of Technical Sciences majoring 6M072900 Construction. From 2021, she is a Senior Lecturer at the Department of “Construction”, Faculty of Architecture and Civil Engineering at Eurasian National University named after L.N.Gumilev. She is an author of plethora of scientific and educational publications related to construction field. Her research interests are computational analysis of building constructions and structures. She can be contacted at email strelec6767@mail.ru.

Yekaterina Golenko in 2015, she graduated from the L.N. Gumilyov Eurasian National University with a degree in Engineering and Technology. In 2018, she received a Master's degree in Engineering. In 2022, she graduated from the doctoral program “S. Seifullin Kazakh Agrotechnical Research University”, educational program 8D06101 –big data analytics. Her research interests include bioinformatics, neural networks, deep learning and data analysis. She can be contacted at email: golenko.katerina@gmail.com.