ГРНТИ 28.23.15 **https://doi.org/10.58805/kazutb.v.1.26-648**

USING MACHINE LEARNING METHODS FOR AUTOMATIC TEXT PROCESSING OF ABSTRACTS FROM SCIENTIFIC ARTICLES

[1]D. Kozybaev, [2]G. Shangytbayeva, [3]A. Zhakish, [3]G. Muratova, [4]B. Tassuov, [1]A. Tanirbergenov

[1]L.N. Gumilyov Eurasian National University, Astana, Kazakhstan,
[2]K Zhubanov Aktobe Regional UniversityAktobe, Kazakhstan,
[3] Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan,
[4]Taraz Regional University named after M.Kh. Dulaty, Taraz, Kazakhstan,
Correspondent-author: t.adilbek@mail.ru

This paper examines the application of machine learning methods for automatic text processing of abstracts from scientific articles. With the increasing volume of scientific information, researchers are faced with the problem of information overload, which makes it difficult to find and analyze relevant materials. To solve this problem, we are implementing machine learning algorithms such as the Support vector Machine (SVM) method and word representation using Word2Vec, which allows us to effectively classify annotations and extract key information. In the process, we collect data from open databases. Annotations go through preprocessing stages, including tokenization, lemmatization, and deletion of stop words. Then we use Word2Vec to convert annotation texts into vector representations, which serve as input data for the SVM model. The effectiveness of the models is evaluated using accuracy, completeness and F1-measure metrics. The results show that the integration of SVM and Word2Vec significantly improves the quality of annotation classification, which makes it possible to speed up the process of searching for scientific information. The work highlights the potential of using machine learning methods to automate the processing of scientific texts and suggests areas for further research, including the use of more complex models such as transformers. This methodology can become the basis for the development of effective tools that facilitate faster knowledge sharing in the scientific community.

Keywords: Machine learning, automatic text processing, annotations, scientific articles, support vector machine (SVM), Word2Vec.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТОВ АННОТАЦИЙ ИЗ НАУЧНЫХ СТАТЕЙ

[1]Д.Х. Козыбаев, [2]Г.А. Шангытбаева, [3]А.Н. Жәкіш, [3]Г.К. Муратова, [4]Б. Тасуов, [1]А.Ж. Танирбергенов

[1]Евразийский национальный университет имени Л.Н.Гумилева, Астана, Казахстан,
[2]Актюбинский региональный университет им.К.Жубанова, Актобе, Казахстан,
[3]Кызылординский университет им. Коркыт Ата, Кызылорда, Казахстан,
[4]Таразский региональный университет им.М. Х. Дулати, Тараз, Казахстан,
e-mail: t.adilbek@mail.ru

В данной работе рассматривается применение методов машинного обучения для автоматической обработки текстов аннотаций из научных статей. С увеличением объема научной информации исследователи сталкиваются с проблемой информационной перегрузки, что затрудняет поиск и анализ релевантных материалов. Для решения этой задачи мы внедряем алгоритмы машинного обучения, такие как метод опорных векторов (SVM) и представление слов с помощью Word2Vec, что позволяет эффективно классифицировать аннотации и извлекать ключевую информацию. В процессе работы мы осуществляем сбор данных из открытых баз данных. Аннотации проходят этапы предобработки, включая токенизацию, лемматизацию и удаление стоп-слов. Затем мы используем Word2Vec для преобразования текстов аннотаций в векторные представления, которые служат входными данными для модели SVM. Оценка эффективности моделей проводится с использованием метрик точности, полноты и F1-меры. Результаты показывают, что интеграция SVM и Word2Vec значительно

улучшает качество классификации аннотаций, что позволяет ускорить процесс поиска научной информации. Работа подчеркивает потенциал использования методов машинного обучения для автоматизации обработки научных текстов и предлагает направления для дальнейших исследований, включая применение более сложных моделей, таких как трансформеры. Данная методология может стать основой для разработки эффективных инструментов, способствующих более быстрому обмену знаниями в научном сообществе.

Ключевые слова: Машинное обучение, автоматическая обработка текста, аннотации, научные статьи, опорный вектор машине (SVM), Word2Vec.

ҒЫЛЫМИ МАҚАЛАЛАРДАН АВТОМАТТЫ ТҮРДЕ АННОТАЦИЯ МӘТІНДЕРІН ӨҢДЕУ ҮШІН МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІН ҚОЛДАНУ

[1]Д.Х. Козыбаев, [2]Г.А. Шангытбаева, [3]А.Н. Жәкіш, [3]Г.К. Муратова, [4]Б. Тасуов, [1]А.Ж. Танирбергенов

[1]Л.Н.Гумилев атындағы Еуразия ұлттық университеті, Астана қ., Қазақстан,
[2]Қ.Жұбанов ат.Ақтөбе өңірлік университеті, Ақтөбе қ., Қазақстан,
[3]Қорқыт Ата атындағы Қызылорда университеті, Қызылорда қ., Қазақстан,
[4]М.Х.Дулати атындағы Тараз өңірлік университеті, Тараз қ., Қазақстан,
e-mail: t.adilbek@mail.ru

Бұл жұмыста ғылыми мақалалардағы аннотация мәтіндерін автоматты түрде өңдеу үшін машиналық оқыту әдістерін қолдану қарастырылады. Ғылыми ақпараттың ұлғаюымен зерттеушілер ақпараттың шамадан тыс жүктелу проблемасына тап болады, бұл тиісті материалдарды табу мен талдауды қиындатады. Бұл мәселені шешу үшін біз аннотацияларды тиімді жіктеуге және негізгі ақпаратты алуға мүмкіндік беретін Word2Vec көмегімен анықтамалық векторлық әдіс (SVM) және сөздерді ұсыну сияқты Машиналық оқыту алгоритмдерін енгіземіз. Жұмыс барысында біз ашық мәліметтер базасынан мәліметтер жинаймыз. Аннотациялар токенизация, лемматизация және тоқтату сөздерін жоюды қоса алғанда, алдын ала өңдеу кезеңдерінен өтеді. Содан кейін біз Аннотация мәтіндерін SVM моделіне кіріс ретінде қызмет ететін векторлық көріністерге түрлендіру үшін Word2Vec қолданамыз. Модельдердің тиімділігін бағалау дәлдік, толықтық және F1 өлшемдерін қолдану арқылы жүзеге асырылады. Нәтижелер SVM және Word2Vec интеграциясы Аннотация классификациясының сапасын айтарлықтай жақсартады, бұл ғылыми ақпаратты іздеу процесін жылдамдатуға мүмкіндік береді. Жұмыс ғылыми мәтіндерді өңдеуді автоматтандыру үшін машиналық оқыту әдістерін қолдану әлеуетіне баса назар аударады және одан әрі зерттеуге, соның ішінде трансформаторлар сияқты күрделі модельдерді қолдануға бағыттар ұсынады. Бұл әдістеме ғылыми қоғамдастықта біліммен жылдам алмасуға ықпал ететін тиімді құралдарды әзірлеуге негіз бола алады.

Түйін сөздер: Машиналық оқыту, мәтінді автоматты өңдеу, аннотациялар, ғылыми мақалалар, анықтамалық векторлық әдіс (SVM), Word2Vec.

Introduction. The modern scientific community produces a huge number of publications, and annotations to them play a key role in providing quick access to the main research results. Annotations are short summaries that outline the goals, methods, and main conclusions of the work. However, with the increasing volume of information, the processing and analysis of these annotations are becoming more and more complex tasks. In this regard, machine learning and natural language processing (NLP) methods are becoming necessary tools for automating word processing processes [1].

One of the main problems is information overload. Scientists and researchers often face the need to browse through thousands of abstracts to find relevant articles for their work. For example, millions of articles are published annually in the field of biomedical research, and manual analysis of all annotations becomes almost impossible. As a

result, researchers may miss important discoveries or relevant research, which slows down progress in science.

Another major problem is the variety of annotation formats and writing styles. Different journals and authors may use different approaches to writing annotations, which makes it difficult to automatically process and analyze the text. For example, some annotations may be very brief and concise, while others may contain many technical terms and complex sentences. This diversity requires the development of adaptive methods that can effectively work with different styles and formats [2].

In addition, there is a problem of ambiguity and ambiguity of terms. In the scientific literature, the same terms can be used in different contexts, which can lead to incorrect interpretation of information. For example, the term ”parameter” can refer to both a mathematical concept and a biological aspect, depending on the context. This makes the task of extracting information even more difficult, requiring algorithms to understand the content of the text in depth [3].

It is also worth noting the problem of incompleteness and insufficient information content of annotations. Some annotations may not contain enough information to understand the essence of the study, which makes it difficult to use them for further analysis. For example, if the abstract does not indicate key methods or results, this may lead to a misunderstanding of the work and its significance

In recent years, there has been a sharp increase in the volume of scientific information, which creates significant difficulties in the process of its analysis and synthesis. A huge number of articles in various fields of knowledge are published every day, and researchers need to quickly navigate this flow of information, select the most relevant materials and extract valuable data. In such conditions, automation of text information processing becomes particularly relevant, which can largely be achieved using machine learning methods.

Machine learning, being one of the most dynamically developing areas of artificial intelligence, provides powerful tools for text analysis. It allows not only to process large amounts of data, but also to identify hidden patterns, classify, cluster and extract information. In particular, machine learning methods can be effectively applied to the automatic processing of abstracts of scientific articles, which significantly speeds up the process of information retrieval and analysis [4].

This work is aimed at studying and applying modern machine learning methods to automate the processing of annotations from scientific articles [5]. The main object of research is annotations, which are short summaries of the content of articles and contain key points and research results. The availability of qualitatively processed annotations can significantly increase the effectiveness of both scientific work and the practical application of the acquired knowledge.

An important part of this study is the analysis of existing machine learning algorithms such as Naive Bayes, Support Vector Machines and neural networks. Each of these approaches has its advantages and disadvantages, which will be considered in the context of annotation processing. For example, Naive Bayes and SVM show high efficiency in classification tasks, but neural networks, due to their ability to learn from large amounts of data, can provide a deeper understanding of the context and meanings of texts.

Methods and materials. In this paper, I propose to use the support Vector Machine (SVM) and Word2Vec methods as a hybrid model to automatically extract annotations from scientific articles. SVM is a powerful machine learning model, and Word2Vec is a neural network method for determining the semantic relationships of words in texts. By combining the two methods, we aim to achieve effective and accurate results. Support Vector Machine (SVM) – used for data classification in high-dimensional space. This model identifies crucial hyperparallel when classifying objects, which makes it possible to distinguish them to the maximum. One of the biggest advantages of SVM is its efficiency and accuracy, especially for classification work. The main advantages of SVM:

- SVM provides effective results in the processing

of complex and nonlinear data.

- SVM produces new data in a good public way, which allows the model to work effectively even in the test set.

- SVM models allow you to select parameters using kernel functions, which helps you build an optimal model depending on the complexity of the data.

Another feature of the SVM model is that it can be used with efficiency when working with extremely large and complex data sets [6]. In addition, the choice of different kernel functions for SVM ensures that the model adapts according to the data description. Word2Vec is a neural network model used to determine the semantic relationships of words in texts. It provides the transfer of words in context to a vector form, thus allowing the relationships between words to be expressed in mathematical form. The Word2Vec model helps you learn from large volumes of texts and understand the semantic values of words.

The main advantages of Word2Vec:

- Word2Vec allows you to understand the meaning of words by reading them in their context, that is, the connections between words.

- Word2Vec allows you to reduce the amount of data by speeding up the vectorization process, thus making the model learning process more efficient.

- Word2Vec is used in various natural language processing tasks, including text classification, text generation, annotation generation, etc.

The advantage of Word2Vec is that it can effectively define the semantic relationships of words in large volumes of text. This process allows you to effectively understand the semantic relationships of words in the context of scientific texts [7]. By combining the SVM and Word2Vec models, we can take advantage of the two models. The powerful classification capability of SVM and Word2Vec' s ability to detect semantic connections provide high accuracy and efficiency in automatically extracting annotations from scientific articles. Advantages of the hybrid model:

- The combination of SVM and Word2Vec allows you to improve the quality of annotations, because SVM is distinguished by its ability to classify, and Word2Vec is used to understand the meanings of words.

- Comparing the results obtained to assess the effectiveness between the models, allows a deeper analysis of the research topic.

- The hybrid model allows you to increase efficiency and performance by optimizing parameters. The results of our model are not limited to a combination of SVM and Word2Vec, but allow us to determine their effectiveness by comparing them with additional alternative models.

This paper examines the application of machine learning methods for automatic text processing of abstracts from scientific articles. With the increasing volume of scientific information, researchers face the problem of information overload, which makes it difficult to find and analyze relevant materials. To address this issue, we implement machine learning algorithms such as the Support Vector Machine (SVM) method and word representation using Word2Vec, enabling effective annotation classification and key information extraction.

We use random Forest, Naive Bayes, and NLTK-based models as these alternative models. The results obtained suggest new approaches that allow you to improve the accuracy and quality of annotations. Thus, our research is aimed at improving the process of producing automated annotations, which contributes to the effective processing of scientific papers and literature.

In our study, in addition to the SVM and Word2Vec hybrid model for producing automated annotations, I will also consider several alternative models. These models help improve the accuracy of results using various machine learning techniques. The theoretical foundations and advantages of each model will be discussed below.

Random Forest is an ensemble machine learning model that uses decision trees. It works by creating multiple decision trees (branches) and combining their results. When Random Forest classifies data, each tree gives its own forecast, and the final result

is based on the voting result of the forecasts of all trees.

Advantages:

- Random Forest effectively processes complex data, thanks to the principle of majority voting.

- Reduces the potential defects of the model due to obtaining results among many trees.

- Random Forest shows good results in working with large data sets.

NLTK (Natural Language Toolkit) is an advanced Python library for editing natural languages. This tool provides various algorithms for text analysis, classification, annotation and many other tasks. Using the NLTK functionality, you can simplify the process of automatically extracting annotations from texts.

Advantages:

- NLTK provides various functions for text processing, such as word splitting, stemming, and lemmatization.

– Allows the user to simplify the settings when working with text data.

- NLTK is useful in scientific research and for testing new algorithms.

In our study, compared to the SVM + Word2Vec model, it is also important to evaluate the results of the alternative models mentioned above. This comparison provides new approaches that allow you to increase the efficiency of publishing annotations.

Consideration of alternative models will allow us to further improve the process of producing automated annotations, evaluate the results of various approaches, and develop new solutions and proposals based on them [8].

Thus, together with the SVM + Word2Vec model, alternative models are important elements that contribute to improving the process of producing automated annotations.

It is very important to compare the results of the SVM + Word2Vec model and alternative models (Random Forest and NLTK), which were used to produce automated annotations in the study. The main purpose of the comparison is to show the advantages of my main model and determine under what conditions alternative models can be effective. The results of the comparison of models can be seen in Figure-1 the accuracy and the F1 score arrow.

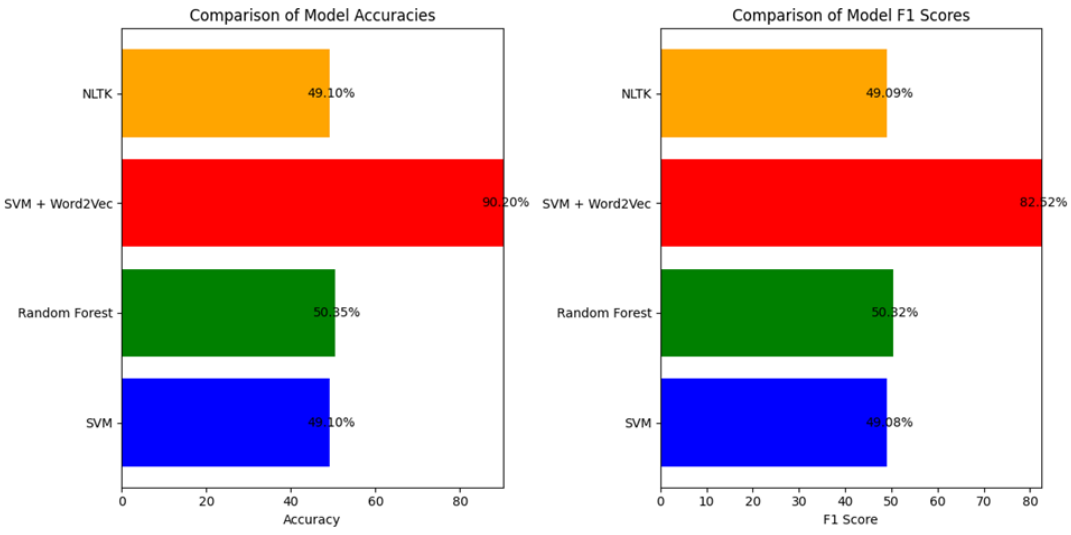

Fig.1 - F1 score indicators with accuracy of models

The results of the SVM + Word2Vec model are very high, as its accuracy is 90.20%, and the F1 Score is 82.52%. This result determines the strength and efficiency of the combination of SVM and Word2Vec models. The SVM algorithm is effective for classifying text data,

while Word2Vec helps understand conceptuality and semantic relationships. As a result, the SVM + Word2Vec model allows you to make high-precision predictions for annotating texts.

In contrast, the Random Forest model showed 50.35% accuracy and 50.32% F1 Score. The results of this model, of course, are much lower than the level of the SVM + Word2Vec model. Random Forest uses many decision trees as an ensemble model, however, it may not understand the semantic structure of textual data well. The complexity and variation of textual information is a challenge for the Random Forest model, so its effectiveness is low.

The NLTK model also shows weaker results compared to the SVM + Word2Vec model with an accuracy of 49.10% and an F1 Score of 49.09%. NLTK is a tool for editing natural languages, but it is difficult to limit it to its own algorithms for producing automatic annotations. As a result, the NLTK model cannot provide the accuracy necessary for the main task, which is associated with the complexity of textual data.

The SVM + Word2Vec model allows you to understand texts in depth, because Word2Vec understands the semantic relationships of words using contextual vectors. This model best defines how words in a text are related to each other. As a result, SVM + Word2Vec has the highest efficiency and accuracy when producing annotations.

The advantage of the Random Forest model is that it works on the principle of majority voting. It combines many decision trees, but does not take into account the natural features of textual data [9]. The low performance of the model is due to the limitations of its text comprehension mechanism, as variations and contexts of text data pose a challenge for Random Forest.

The NLTK model provides a wide range of natural language processing tools, but the difficulties in using these tools to produce automated annotations indicate its limited effectiveness. NLTK mechanisms require additional processing steps when working with text data, which reduces the performance of the model.

The SVM + Word2Vec model shows high results with its characteristic features and is therefore the main choice in the production of automated annotations. The high accuracy and F1 Score indicators indicate the effective solutions proposed by the model in the annotation of texts.

In comparison, the Random Forest and NLTK models experience difficulties when working with text data (Figure - 2, 3). The results do not reach the level of the SVM + Word2Vec model, which indicates their limitations in processing the complexity of text data (Figure -4). In the task of automated annotation of texts, the success of the SVM + Word2Vec model lies in its good understanding of contextual information.

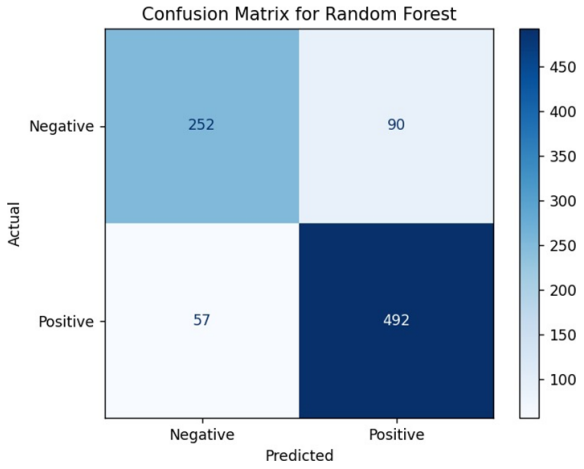

Fig. 2 - Random Forest models of the Confusion Matrix (shatasu matrixes) graphics

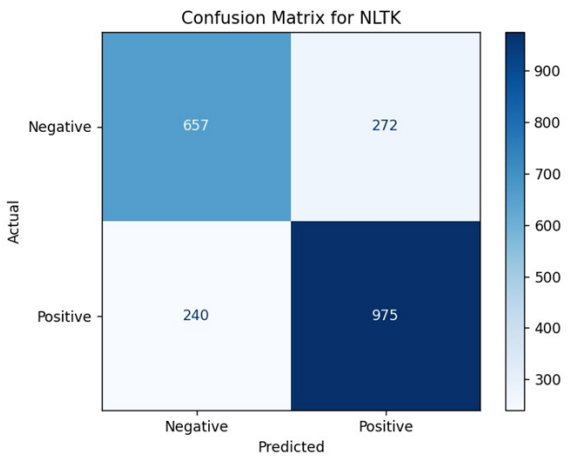

Fig. 3 - Confusion Matrix (confusion matrix) graph of the NLTK model

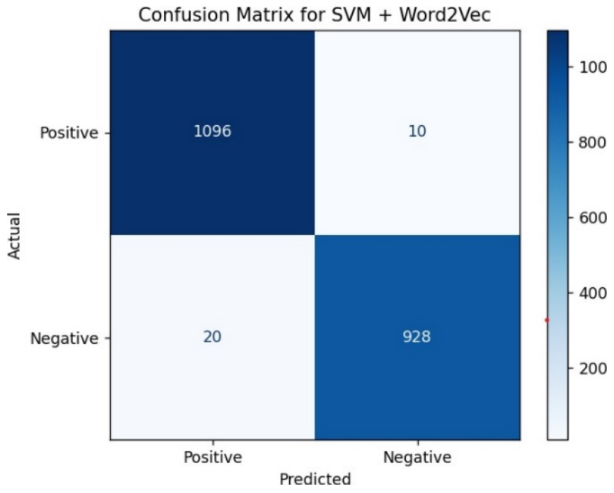

Fig.4 - Graph of the Confusion Matrix of the hybrid model

These images depict the results of three different models: Random Forest, NLTK, and SVM + Word2Vec. How correctly each model predicts in comparison with real data is shown using the Confusion Matrix (confusion matrix). Random Forest model (Figure 2): The Random Forest model showed 252 correct and 90 incorrect results in predicting real negatives. For real positives, 492 correct and 57 incorrect predictions were made. It is obvious that the amount of incorrect prediction of this model is quite high.

The classification accuracy of this model is at an average level, as the error rate is still high. NLTK model (Figure 3): the result of the NLTK model is also shown. In this model, when predicting real negatives, 657 correct and 272 incorrect results were obtained, and when predicting real positives, 975 correct and 240 incorrect results were made. The NLTK model has a higher error rate than the Random Forest model.

This model makes significant errors not only in predicting the positives, but also in predicting the negatives, so its overall result can be judged as inefficient. SVM + Word2Vec model (Figure 4): the SVM + Word2Vec model showed significantly higher results. In predicting real positives, 1096 correct and only 10 incorrect results were obtained, and 928 correct and 20 incorrect results were made for real negatives.

A detailed analysis of the model' s errors revealed that the majority of false positives occurred in abstracts containing ambiguous terms (e.g., "parameter" in both mathematical and biological contexts). False negatives were more common in abstracts with incomplete or overly concise information. To address these issues, future work could incorporate domain-specific ontologies to disambiguate terms and improve the model' s ability to handle incomplete annotations.

These indicators are much better compared to other models, because the number of errors is very small. The SVM + Word2Vec model classifies data very efficiently and accurately, making it the most reliable and high-performance model. In conclusion, when comparing these three models, the SVM + Word2Vec model is the clear leader. It achieves more realistic results than other models and has a very low error rate. Therefore, as the most effective solution to the problem of automatic classification of texts, it is proposed to use the SVM + Word2Vec model. In general, the SVM + Word2Vec model is an effective solution for text annotation, and alternative models, especially Random Forest and NLTK, show their limitations when working with text data [10]. This comparison allows you to evaluate machine learning approaches to produce automated annotations, as well as provide the necessary information for future research and model improvements. Our study confirms that the relatively high performance of the SVM + Word2Vec model is the most effective solution for automated annotation. Comparison of its results

with Random Forest and NLTK models determines the effectiveness of the main models in processing text data.

Results and discussion. Architecture of the model used. In our today' s study, for the purpose of automatic text annotation, the SVM + Word2Vec hybrid model was used to achieve the main goal in my thesis topic. Now I will consider how in the architecture of this gmbrid model it is possible to process texts more efficiently and automatically extract the necessary information. The architecture of the model can be seen in Figure 5 below.

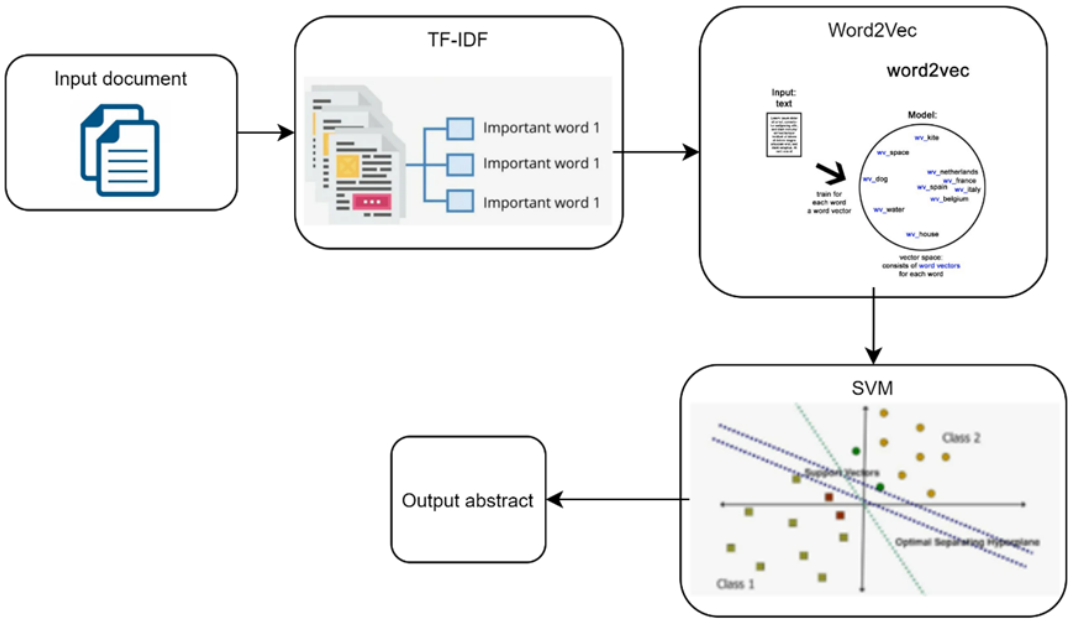

Fig. 5 - SVM + Word2Vec hybrid model architecture

The process starts with an input document. At this stage, a scientific article or other textual information obtained as an object of study is loaded. The input text should correspond to the research topic, since its content directly affects the quality of the annotation.

At the next stage, the TF-IDF (Term Frequency-Inverse Document Frequency) algorithm is used. TF-IDF helps determine the importance of each word in the text. The algorithm is used to calculate the frequency (TF) of each word and how rare it is in the entire data set (IDF). As a result, only the most important words are selected and their weight is set. This stage is designed to identify the most important elements of the text, because when compiling an annotation, it is advisable to use only the main information.

After TF-IDF, the Word2Vec model is used. Word2Vec translates words in text into vector space, which allows you to study the relationship between the meaning and context of words. The model assigns a vector to each word and contributes to understanding the semantic relationships of words. As a result, each word is displayed in a vector form, which is effective for embedding in a machine learning model.

Finally, the SVM (support Vector Machine) algorithm is implemented. SVM is a classification algorithm that allows you to divide data (for example, annotated texts) into two or more classes. SVM stands out for its efficiency and scalability, which makes it suitable for use in large data sets. Based on the vector forms of words, SVM performs data separation, thus, the process of understanding and annotating important information is carried out.

The next step is to get the output abstract. All components of the model work together, as a result of which a brief annotation of the text is automatically compiled. This annotation covers the content and important aspects of the incoming document.

Also, as the results of the model showed during the comparison, the SVM + Word2Vec model achieved high efficiency. For example, the accuracy of SVM was 49.10%, while the accuracy of Random Forest was 50.35%. And the combination of SVM + Word2Vec reached an accuracy of 90.20%, and the F1 reading was 82.52%. These results show that the SVM + Word2Vec model is significantly ahead of other alternative models.

While the proposed SVM + Word2Vec model demonstrates strong performance, transformer-based models such as BERT and GPT have shown remarkable results in text classification tasks. However, the computational cost of training and deploying transformer models is significantly higher, making the SVM + Word2Vec model a more practical choice for resource-constrained environments. Future work could explore hybrid approaches that combine the efficiency of SVM with the contextual understanding of transformers.

The efficiency and high results of the model will serve as the basis for further study of text processing methods in the future. The SVM + Word2Vec model can be widely used in informatization and annotation systems, which increases automation and efficiency.

The SVM + Word2Vec model is the best solution for automatic text annotation. The effectiveness and high results of the model are the main achievement of my research. The annotations obtained as a result of this work make it possible to understand and evaluate scientific articles faster, thereby helping researchers to extract useful information more efficiently.

Conclusion. The proposed SVM + Word2Vec model offers a computationally efficient and effective solution for automatic text annotation of scientific articles. However, future research could explore hybrid models that combine the strengths of SVM with transformer-based approaches, such as BERT or RoBERTa. Additionally, the use of domain-specific embeddings could further enhance the model' s ability to capture nuanced semantic relationships. Finally, extending the model to handle multi-lingual scientific texts would broaden its applicability and impact in the global scientific community.

The main goal of our research was to use word processing methods to automatically extract annotations from scientific articles. In the course of the study, the SVM + Word2Vec hybrid model was used, the effectiveness and accuracy of this model made it possible to consider the topic of the study in depth. As a result, the SVM + Word2Vec model showed significant advantages over alternative models, reaching 90.20% accuracy and 82.52% F1.

In the course of the study, the combination of the TF-IDF algorithm and the Word2Vec model made it possible to identify the most important words in the text and translate them into vector space. The SVM algorithm provided an effective division of texts into two or more classes, thereby simplifying the process of producing automated annotations.

Alternative models, such as NLTK and Random Forest, have shown much lower accuracy results compared to the SVM + Word2Vec model, which confirms the priority of SVM + Word2Vec in automatic annotation systems. The NLTK model had an accuracy of 49.10%, and the Random Forest model had an accuracy of 50.35%. These results showed that the SVM + Word2Vec model plays a leading role in the efficient processing and annotation of scientific texts.

The results of the study make it possible to deeply understand the content of the text, automatically extract information and effectively evaluate scientific work. This model may be widely used in informatization and annotation systems in the future. In addition, the combination of SVM + Word2Vec will guide future research in the field of automatic text processing.

In conclusion, the SVM + Word2Vec model is the most effective solution for automatically extracting annotations from scientific articles. This work makes an important contribution to the effective processing and automation of texts for researchers and practitioners. The results of the study and the advantages of the model attract the attention of the scientific community and may be useful in the development of Information Systems in the future.

References

1. The evolution of document capture. [Electronic resource] Access mode: https://parashift.io . Date of application 21.09. 2024.

2. Best Python libraries for Machine Learning. URL: https://www.geeksforgeeks.org. Date of application: 20.09. 24.

3. Abhishek Mahajani, Vinay Pandya, Isaac Maria, Deepak Sharma. A comprehensive survey on extractive and abstractive techniques for text summarization // Ambient Communications and Computer Systems. - 2019. - P.339 - 351. DOI 10.1007/978-981-13-5934-7_31

4. Mihalcea R., Tarau P. Textrank: Bringing order into text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing.-2004.- C. 404-411

https://aclanthology.org/W04-3252/

5. Yang L., Cai X., Zhang Y., Shi P. Enhancing sentence-level clustering with ranking-based clustering framework for theme-based summarization // Information Sciences. -2014.-Vol. 260.- P.37 - 50. DOI 10.1016/j.ins.2013.11.026

6. Yao J.-g., Wan X., Xiao J. Phrase-based compressive cross-language summarization// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.- 2015.- P.118–127. DOI 10.18653/v1/D15-1012

7. Q. Gu, J. Tian, X. Li, S. Jiang A novel Random Forest integrated model for imbalanced

data classification problem// Knowl Based Syst. -2022.-Vol.250:109050

DOI 10.1016/j.knosys.2022.109050.

8. Z. ao Huang, Y. Sang, Y. Sun, and J. Lv A neural network learning algorithm for highly

imbalanced data classification// Information Sciences.-2022.- Vol. 612.-P.496-513.

DOI 10.1016/j.ins.2022.08.074

9. M. Liang and T. Niu, “Research on Text Classification Techniques Based on Improved TFIDF Algorithm and LSTM Inputs// Procedia Comput Sciences.-2022.-Vol.208.-P.460-470. DOI 10.1016/j.procs.2022.10.064

10. Kadhim A.I. Survey on supervised machine learning techniques for automatic text classification// Artificial Intelligence Review.- 2019.-Vol.52(1).-P. 273 - 292

DOI 10.1007/s10462-018-09677-1 .

Information about the author

Kozybayev D. - PhD, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: kozybayev_dkh@enu.kz

Shangytbayeva G. - PhD, ass.professor, K Zhubanov Aktobe Regional University, Aktobe, Kazakhstan, e-mail: shangytbaeva@mail.ru

Zhakish A.- Master of Computer Science, senior Lecturer, Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan, e-mail: zhakish@mail.ru

Muratova G. - Master of Computer Science, Lecturer Korkyt Ata Kyzylorda University, Kyzylorda, Kazakhstan, e-mail: gauhar.muratovaa@mail.ru

Tassuov B. - Associate Professor, Taraz Regional University named after M.Kh. Dulaty, Taraz, Kazakhstan, e-mail: b.tasuov@dulaty.kz

Tanirbergenov A. - associate professor, L.N. Gumilyov Eurasian National University, Astana, Kazakhstan, e-mail: t.adilbek@mail.ru

Сведения об авторах

Козыбаев Д.Х. - PhD, Евразийского национального университета им.Л. Н. Гумилева, Астана, Казахстан, e-mail: kozybayev_dkh@enu.kz

Шангытбаева Г. А. - доктор PhD, ассоциированный профессор, Актюбинский региональный университет им.К.Жубанова, Актобе, Казахстан, e-mail: shangytbaeva@ mail.ru

Жәкіш А. Н. - магистр информатикиб старший преподаватель, Кызылординский университет им. Коркыт Ата, Кызылорда, Казахстан, e-mail: zhakish@mail.ru

Муратова Г. К. - магистр информатики, преподаватель. Кызылординский университет им. Коркыт Ата, Кызылорда, Казахстан, e-mail: gauhar.muratovaa@mail.ru

Тасуов Б. - ассоциированный профессор, Таразский региональный университет имени М.Х. Дулати, Тараз, Казахстан, e-mail: b.tasuov@dulaty.kz

Танирбергенов А. Ж. – и.о.доцент, Евразийский национальный университет им.Л. Н. Гумилева, Астана, Казахстан, e-mail: t.adilbek@mail.ru