UDC 004.75:004.8:004.42
DOI: 10.15587/2706-5448.2024.309472

Oleksii Kuznietsov,
Gennadiy Kyselov

AN OVERVIEW OF CURRENT ISSUES IN AUTOMATIC TEXT SUMMARIZATION OF NATURAL LANGUAGE USING ARTIFICIAL INTELLIGENCE METHODS

The object of the research is the task of automatic abstracting of natural language texts. The importance of these tasks is determined by the existing problem of creating essays that would adequately reflect the content of the original text and highlight key information. This task requires the ability of models to deeply analyze the context of the text, which complicates the abstracting process.

Results are presented that demonstrate the effectiveness of using generative models based on neural networks, text semantic analysis methods, and deep learning for automatic creation of abstracts. The use of models showed a high level of adequacy and informativeness of abstracts. GPT (Generative Pre-trained Transformer) generates text that looks like it was written by a human, which makes it useful for automatic essay generation.

For example, the GPT model generates abbreviated summaries based on input text, while the BERT model is used for summarizing texts in many areas, including search engines and natural language processing. This allows for short but informative abstracts that retain the essential content of the original and provides the ability to produce high-quality abstracts that can be used for abstracting web pages, emails, social media, and other content. Compared to traditional abstracting methods, artificial intelligence provides such advantages as greater accuracy, informativeness and the ability to process large volumes of text more efficiently, which facilitates access to information and improves productivity in text processing.

Automatic abstracting of texts using artificial intelligence models allows to significantly reduce the time required for the analysis of large volumes of textual information. This is especially important in today's information environment, where the amount of available data is constantly growing. The use of these models promotes efficient use of resources and increases overall productivity in a variety of fields, including scientific research, education, business and media.

Keywords: *automatic abstracting, natural language processing, artificial intelligence, generative models, neural networks, deep learning.*

Received date: 24.05.2024
Accepted date: 26.07.2024
Published date: 31.07.2024

How to cite
Kuznietsov, O., Kyselov, G. (2024). An overview of current issues in automatic text summarization of natural language using artificial intelligence methods. Technology Audit and Production Reserves, 4 (2 (78)), 12–19. https://doi.org/10.15587/2706-5448.2024.309472

1. Introduction

Every day, the amount of information stored in the form of texts is growing rapidly, and the process of analyzing and understanding them is becoming more and more difficult for people. In this regard, there is a need for effective methods of automatically creating a concise content of texts, which allows to highlight key information and create short, meaningful conclusions from large volumes of data.

Automatic abstracting of natural language texts (Automatic Text Summarization) is an active area of research in the field of natural language processing and artificial intelligence. The purpose of automatic abstracting is to create a compact representation of the content of the text, which reflects its main essence and important details. This may include highlighting key facts, events or ideas, and avoiding repetition and non-essential details.

Thus, the development of effective automatic text referencing systems is an urgent task that can have a significant impact on the development of natural language processing technologies and facilitate the analysis of large volumes of information.

Automatic abstracting of natural language texts using artificial intelligence methods remains a very relevant topic in the modern world. Due to the growing amount of information on the Internet, the importance of fast and efficient processing of text information is becoming more and more noticeable.

Automatic abstracting of texts is of great importance for various industries, including news analysis, scientific research, medicine, law, and many others. With the help of artificial intelligence methods, such as natural language processing (NLP), machine learning and deep learning, it is possible to develop systems that automatically create abstracts from large volumes of text [1].

The use of such systems can significantly facilitate the process of information analysis, reducing the time required to read and understand texts. They can also be useful for filtering out unnecessary or irrelevant information, keeping only the key points for the user.

Therefore, *the aim of research* is to study existing models and methods of automatic abstracting of natural language texts based on neural networks, in particular GPT, BERT, TextTeaser, SMMRY or SummarizeBot. To do this, it is necessary to identify regularities in text processing, determine the impact of different models on the accuracy and informativeness of abstracts, and also develop techniques for applying these methods to improve the quality of abstracting. The practical part of the study will make it possible to facilitate access to information, improve productivity in word processing and provide better referencing of web pages, e-mails, social media and other content. The use of various models such as GPT, BERT, TextTeaser, SMMRY, SummarizeBot, etc. will ensure the creation of abstracts that have higher adequacy and informativeness compared to traditional methods. This will significantly reduce the time required to analyze large volumes of text, which, in turn, will facilitate the work of researchers, analysts and other users who need quick access to key information from large text arrays.

2. Materials and Methods

The work considers the relevance of the task of automatic abstracting of texts, what solutions already exist, and also considers what new approaches can be used to solve the task.

The task of automatic referencing is caused by the following factors:

1. *Information overload:* Texts overloaded with irrelevant information complicates a person's ability to effectively navigate the text and understand its essence.
2. *Time shortage:* In a world where a large amount of new information is generated every minute, people often do not have enough time to understand large texts.
3. *Heterogeneity of sources:* Texts can be written by different authors at different times and have a different structure, which complicates their analysis.
4. *Searching for relevant information:* When users are looking for specific information from large volumes of text, its automatic abstracting helps them find what they need faster.

At the moment, these problems are actively investigated and solved using the following methods:

1. *Development of automatic referencing algorithms*: Researchers are working on creating algorithms that can automatically identify the key points of the text and concisely display them.
2. *Using deep learning:* Deep learning-based abstracting systems learn to understand the context and semantics of text to create more accurate and informative abstracts.
3. *Development of text generation models:* Text generation models based on neural networks allow systems to generate more natural and understandable abstracts.
4. *Experiments with different sources and types of texts*: Researchers test abstracting algorithms on different types of texts, including news, scientific articles, blogs, etc., in order to understand how they perform in different conditions.

In modern developments, two main methods of abstracting texts are used:

1) Extractive summarization;
2) Abstractive summarization.

The extractive method is annotation, the basis of which is the selection of key phrases, sentence fragments from the initial document, which are then added to the final abstract without any changes in order. This approach is considered quite reliable because the final essay uses phrases that are taken directly from the original source. However, the method lacks flexibility because it cannot add new words, phrases or paraphrased phrases to the final essay.

The abstract method is an annotation, the basis of which is the selection of the most essential information from the original source, as well as the generation of new texts that are similar in content to the original ones, but summarize them. The method makes it possible to use those words that did not exist in the original source, so such annotations are more similar to those created by a person manually. Therefore, this task is considered more difficult, since this method requires solving the problems of semantic representation of the original source, as well as the problems of generating texts in natural language.

When using Extractive summarization, automatic abstracting of texts is performed in three stages [2]:

1. Construction of an intermediate representation of the input text in the form of a thematic or indicator representation. When using the thematic representation, the text is presented as an interpretation of the themes that are in the primary text. The indicator representation interprets the text as a list of formal features, i. e. indicators. These features include sentence length, place in the text, presence of keywords, etc.
2. Evaluation of text sentences based on intermediate representation. At the same time, each sentence is assigned an assessment of the importance of this sentence for the annotation.
3. Formation of the summary. The system selects a few of the most important sentences to create the resulting annotation.

Thematic representation of the text is based on finding words that describe the topic of the input text. Topic words can be defined in quite different ways, for example, work [3] was the first to use this method. The search for thematic words in this work was carried out using the values of the frequency of use of these words in the text. In work [4], a logarithmic algorithm of thematic word plausibility was used.

The most common frequency-oriented models for searching thematic words are TF-IDF model, log-likelihood ratio and their modifications. The thematic presentation of the text also includes the method based on the centering of sentences.

TF-IDF (Term Frequency – Inverse Document Frequency) is a statistical measure of the importance of a word within one document or a collection of documents. The score is calculated as the product of the function of the number of occurrences of the word in the document and the inverse function of the number of documents in the collection in which the word is present.

The log-likelihood ratio is calculated by the logarithm of the ratio of the probability of finding a word with the same probability in the collection of documents and the collection of summaries corresponding to the documents, to the probability of the appearance of words with different probability in this collection. The formula for calculating the log-likelihood ratio is proposed in [5].

The idea of the sentence centering method is based on the assumption that the most interesting information for annotation is not only in one sentence. Therefore, the main idea of the method is to calculate the «distances» between the sentences in the text and to select those that have an average value of «distances closer» to the others. Bag-of-words models are used to determine the proximity of sentences. For example, the model of Scott Deerwester [6] and the model of Y. Gong and X. Liu [7].

The methods of indicator presentation provide a presentation of texts in the form of a set of indicators (signs) and the subsequent use of these indicators for the classification of sentences, instead of a banal retelling of the topics of the input sentence.

These approaches include methods based on graphical representation, as well as methods that use machine learning to identify sentences to include in the final essay. The main indicators used in these methods are:

- sentences at the beginning or at the end of the text are more informative;
- too short or too long sentences are uninformative;
- presence of keywords;
- use of words from the title of the text in a sentence;
- use of punctuation marks that are emotionally colored (question mark, exclamation mark, three dots, etc.);
- other statistical classifiers that do not require training.

Graph methods represent the text in the form of a graph with connections based, as a rule, on the use of the PageRank algorithm [8]. The vertices in this graph are the sentences, and the connection between the sentences is shown with the help of edges. Edge weights determine the similarity of sentences and the «strength» of the connection between them. The graphic interpretation of the text has two main results. First, each subgraph is a separate section of the document. Secondly, the definition of important topics is based on the assumption that if a sentence has many connections with other sentences and is also the center of one of the subgraphs, then most likely it is important and should be included in the final annotation. Since the method does not require any special language processing, it can be applied to any language. Also, the method is suitable for abstracting both a single document and a multi-document collection.

Referencing texts using neural networks is based on solving the classification problem. The paper [9] shows a very early attempt by researchers to abstract text using machine learning. The researchers in this work classify the sentences based on a naive Bayesian classifier, and then divide the sentences into resulting sentences (those that make it to the final annotation) and non-resulting sentences that were created by the researchers while applying the annotations and that were created by the «mining method» as educational material.

Machine learning methods such as Naive Bayes, Decision trees, Support vector machines, Hidden Markov models and Conditional random fields are widely used in attempts to create new automatic abstracting methods [10–14].

The use of artificial intelligence methods in the field of automatic abstracting of texts makes it possible to create more efficient systems that can quickly analyze and summarize large volumes of text to obtain short informative abstracts. Today, there are several ready-made applications that use artificial intelligence methods for automatic abstracting of natural language texts. Some are used for general abstracting tasks, while others specialize in specific areas, such as journalism or scientific research. The list of the most famous applications of this type includes:

1. *SummarizeBot*: This application uses artificial intelligence to automatically generate short summaries from English text. It can be used for referencing web pages, emails, social media and other content [15].

2. *SMMRY*: This is another tool for automatic concise abstracting of texts. It uses NLP algorithms to highlight the main points of the text and generate a brief overview [16].

3. *GPT (Generative Pre-trained Transformer)*: This is a large-scale deep learning model developed by OpenAI, which can be used for automatic abstracting of texts. By applying GPT, it is possible to create a system that generates abbreviated abstracts based on the entered text [17].

4. *BERT (Bidirectional Encoder Representations from Transformers)*: This is another deep learning model developed by Google that can be used for automatic abstracting of texts. It is used for abstracting texts in many areas, including search engines and natural language processing [18].

5. T*extTeaser*: This application uses machine learning algorithms to create concise abstracts from texts. It takes into account the semantics and structure of the text to create informative summaries [19].

The use of the listed applications demonstrates the potential of artificial intelligence methods in facilitating access to information and improving productivity in word processing.

3. Results and Discussion

3.1. Problems of automatic abstracting of texts. There are several directions that can be investigated or developed to solve the problem of automatic abstracting of natural language texts using artificial intelligence methods:

1. *Using deep learning to improve the quality of essays*: Applying modern deep learning architectures, such as Transformer-based models, to create more accurate and informative essays.

2. *Development of systems capable of abstracting multimedia content*: Expanding the possibilities of abstracting systems to include audio, video and graphic content, which will make abstracts more complex.

3. *Improvement of algorithms for taking into account the context*: Development of algorithms to better understand the context of the text and take it into account when creating abstracts.

4. *Development of personalized reporting systems*: Creation of systems that can create reports taking into account the individual needs of the user, its interests and the context of use.

5. *Automatic determination of the importance of information*: Development of algorithms that can automatically determine the importance of different parts of the text and give preference to key points when generating abstracts.

6. *Using semantic graphs for abstracting texts*: Using semantic graphs to present textual information and automatically create abstracts taking into account the semantic relationships between different concepts.

To solve the problem of automatic abstracting of natural language texts using artificial intelligence methods, the capabilities of the Python language, which contains appropriate libraries and tools, are actively used. Example:

1. *NLTK (Natural Language Toolkit)*: This is a popular Python library for natural language processing. It contains various tools for tokenization, stemming, lemmatization, and others, which can be used for preprocessing the text before abstracting it [20].

2. *Gensim*: This is a library for thematic modeling, which also contains tools for automatic abstracting of texts. It allows to create models of the distribution of topics in the text and use them to highlight key points [21].

3. *Sumy*: This is a Python library that specializes in automatic text referencing. It contains implementations of various abstracting algorithms, such as LSA (Latent Semantic Analysis), TextRank, Luhn, and others [22].

4. *BERTSUM*: This is a model that uses the BERT architecture for automatic abstracting of texts. It is based on the use of a pretrained BERT model for vector representation of text and generation of abstracts [23].

These tools provide possible ways of development and improvement of systems of automatic abstracting of texts using methods of artificial intelligence. The development of these tools contributes to the improvement of the quality and efficiency of referencing systems and makes them more useful for various industries and fields of application.

The target groups of the project on automatic abstracting of natural language texts using artificial intelligence methods include:

1. *Natural Language Processing (NLP) research communities*:

– This group consists of scientists, researchers and students engaged in research in the field of natural language processing. They research various aspects of NLP, including automatic text abstracting, with the aim of improving methods and algorithms.

– Their work consists in the development of new approaches, experimentation with algorithms and machine learning models, as well as in the study of problems that arise in the process of automatic abstracting of texts.

2. *Specialists in reporting and data analysis*:

– This group may include data analysts, data scientists, business intelligence professionals, and other professionals who work with large amounts of information.

– They are interested in using automatic text abstracting to analyze structured or unstructured information to identify key points, trends, anomalies, etc.

3. *Programmers and software developers*:

– This group includes software developers who are interested in using automatic text referencing in their applications or systems.

– They can look for ready-made libraries, tools or APIs that will allow them to build automatic text referencing functionality into their applications.

4. *Enterprises and companies from the field of media and information technologies*:

– These organizations can use automatic abstracting of texts to automate the process of creating summaries of news or information materials.

– They may also be interested in using these technologies to analyze and process large volumes of textual information for the purpose of identifying trends, analyzing global opinion, etc.

5. *End product users*:

– This group includes all people who are looking for quick and efficient ways to get a concise summary of large amounts of text.

– Users can be students, academics, journalists, professionals from various fields who search for information on the Internet, or anyone who wants to efficiently browse large amounts of text.

3.2. Examples of using existing word processing tools

3.2.1. Analysis of tonality of texts. To analyze the tonality of natural language texts, two tools from the NLTK library were studied:

1) VADER sentiment analysis tool;

2) «word_tokenize» tokenizer designed to break large texts into smaller linguistic units such as sentences or words.

The tonality of the headlines of the online publication «Ukrainian Pravda» [24], which has a convenient RSS tool for receiving the latest news, was analyzed. Each of the headings is numbered. The result is presented in Fig. 1.

The VADER algorithm provides four mood scores:

1) neg: negative assessment;

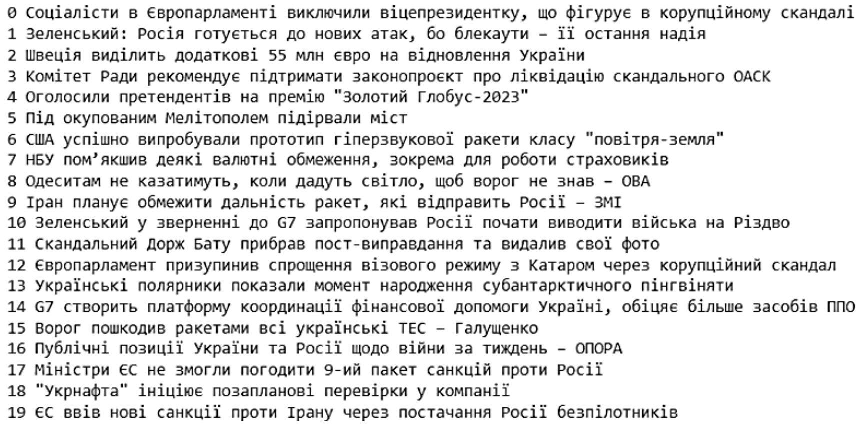
2) neu: neutral assessment;

3) pos: positive assessment;

4) compound: a cumulative assessment that combines the three previous ones.

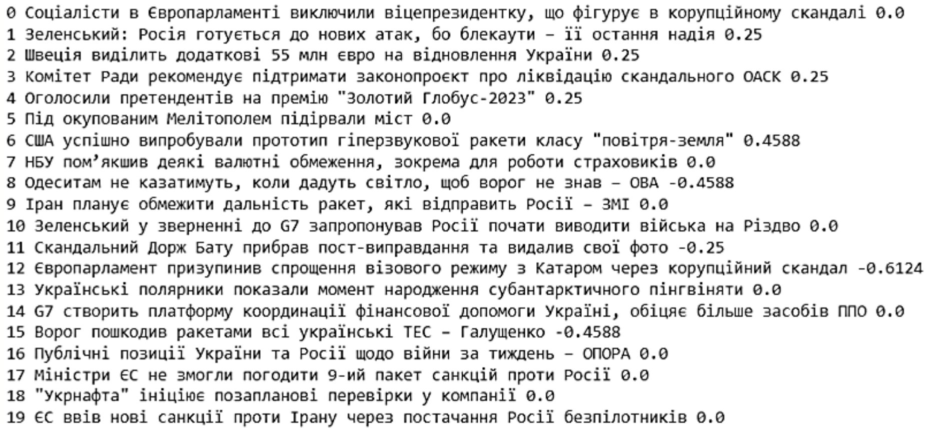
After that, it is possible to evaluate the mood of the collected headlines and evaluate the effectiveness of the selected tool (Fig. 2).

As can be seen from the initial processing results, many headlines have a score of 0.0, which indicates a neutral sentiment. This is due to the lack of prior training that could improve the performance of the algorithm. It is necessary to bring the words to their normal form and clean the text from stop words. Since the corpus of the Ukrainian language has not yet been added to NLTK, pymorphy2 was used for morphological analysis [25], Fig. 3.



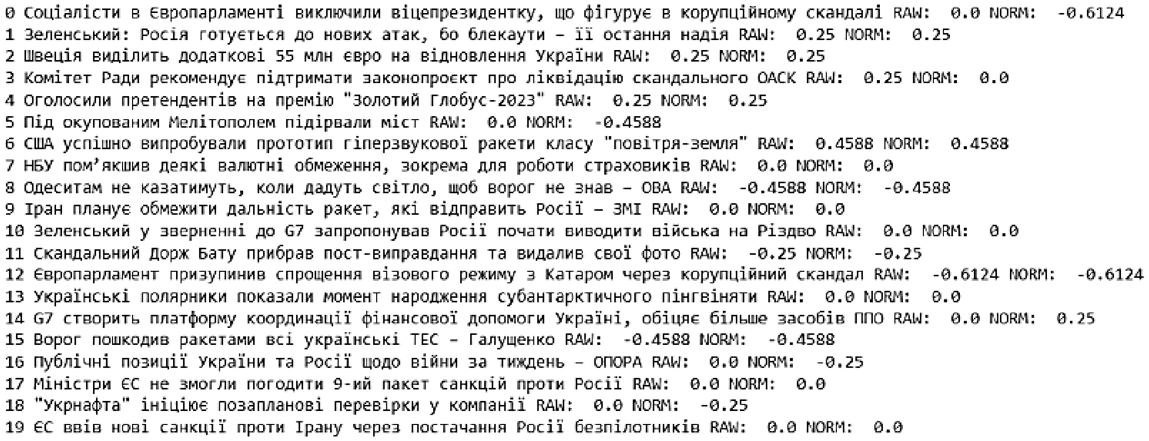
```
0 Соціалісти в Європарламенті виключили віцепрезидентку, що фігурує в корупційному скандалі
1 Зеленський: Росія готується до нових атак, бо блекаути – її остання надія
2 Швеція виділить додаткові 55 млн євро на відновлення України
3 Комітет Ради рекомендує підтримати законопроєкт про ліквідацію скандального ОАСК
4 Оголосили претендентів на премію "Золотий Глобус-2023"
5 Під окупованим Мелітополем підірвали міст
6 США успішно випробували прототип гіперзвукової ракети класу "повітря-земля"
7 НБУ пом’якшив деякі валютні обмеження, зокрема для роботи страховиків
8 Одеситам не казатимуть, коли дадуть світло, щоб ворог не знав – ОВА
9 Іран планує обмежити дальність ракет, які відправить Росії – ЗМІ
10 Зеленський у зверненні до G7 запропонував Росії почати виводити війська на Різдво
11 Скандальний Дорж Бату прибрав пост-виправдання та видалив свої фото
12 Європарламент призупинив спрощення візового режиму з Катаром через корупційний скандал
13 Українські полярники показали момент народження субантарктичного пінгвіняти
14 G7 створить платформу координації фінансової допомоги Україні, обіцяє більше засобів ППО
15 Ворог пошкодив ракетами всі українські ТЕС – Галущенко
16 Публічні позиції України та Росії щодо війни за тиждень – ОПОРА
17 Міністри ЄС не змогли погодити 9-ий пакет санкцій проти Росії
18 "Укрнафта" ініціює позапланові перевірки у компанії
19 ЄС ввів нові санкції проти Ірану через постачання Росії безпілотників
```

Fig. 1. The result of parsing news headlines from the website of the online publication «Ukrainska Pravda»



```
0 Соціалісти в Європарламенті виключили віцепрезидентку, що фігурує в корупційному скандалі 0.0
1 Зеленський: Росія готується до нових атак, бо блекаути – її остання надія 0.25
2 Швеція виділить додаткові 55 млн євро на відновлення України 0.25
3 Комітет Ради рекомендує підтримати законопроєкт про ліквідацію скандального ОАСК 0.25
4 Оголосили претендентів на премію "Золотий Глобус-2023" 0.25
5 Під окупованим Мелітополем підірвали міст 0.0
6 США успішно випробували прототип гіперзвукової ракети класу "повітря-земля" 0.4588
7 НБУ пом’якшив деякі валютні обмеження, зокрема для роботи страховиків 0.0
8 Одеситам не казатимуть, коли дадуть світло, щоб ворог не знав – ОВА -0.4588
9 Іран планує обмежити дальність ракет, які відправить Росії – ЗМІ 0.0
10 Зеленський у зверненні до G7 запропонував Росії почати виводити війська на Різдво 0.0
11 Скандальний Дорж Бату прибрав пост-виправдання та видалив свої фото -0.25
12 Європарламент призупинив спрощення візового режиму з Катаром через корупційний скандал -0.6124
13 Українські полярники показали момент народження субантарктичного пінгвіняти 0.0
14 G7 створить платформу координації фінансової допомоги Україні, обіцяє більше засобів ППО 0.0
15 Ворог пошкодив ракетами всі українські ТЕС – Галущенко -0.4588
16 Публічні позиції України та Росії щодо війни за тиждень – ОПОРА 0.0
17 Міністри ЄС не змогли погодити 9-ий пакет санкцій проти Росії 0.0
18 "Укрнафта" ініціює позапланові перевірки у компанії 0.0
19 ЄС ввів нові санкції проти Ірану через постачання Росії безпілотників 0.0
```

Fig. 2. The initial result of the VADER algorithm



```
0 Соціалісти в Європарламенті виключили віцепрезидентку, що фігурує в корупційному скандалі RAW:  0.0 NORM:  -0.6124
1 Зеленський: Росія готується до нових атак, бо блекаути – її остання надія RAW:  0.25 NORM:  0.25
2 Швеція виділить додаткові 55 млн євро на відновлення України RAW:  0.25 NORM:  0.25
3 Комітет Ради рекомендує підтримати законопроєкт про ліквідацію скандального ОАСК RAW:  0.25 NORM:  0.0
4 Оголосили претендентів на премію "Золотий Глобус-2023" RAW:  0.25 NORM:  0.25
5 Під окупованим Мелітополем підірвали міст RAW:  0.0 NORM:  -0.4588
6 США успішно випробували прототип гіперзвукової ракети класу "повітря-земля" RAW:  0.4588 NORM:  0.4588
7 НБУ пом’якшив деякі валютні обмеження, зокрема для роботи страховиків RAW:  0.0 NORM:  0.0
8 Одеситам не казатимуть, коли дадуть світло, щоб ворог не знав – ОВА RAW:  -0.4588 NORM:  -0.4588
9 Іран планує обмежити дальність ракет, які відправить Росії – ЗМІ RAW:  0.0 NORM:  0.0
10 Зеленський у зверненні до G7 запропонував Росії почати виводити війська на Різдво RAW:  0.0 NORM:  0.0
11 Скандальний Дорж Бату прибрав пост-виправдання та видалив свої фото RAW:  -0.25 NORM:  -0.25
12 Європарламент призупинив спрощення візового режиму з Катаром через корупційний скандал RAW:  -0.6124 NORM:  -0.6124
13 Українські полярники показали момент народження субантарктичного пінгвіняти RAW:  0.0 NORM:  0.0
14 G7 створить платформу координації фінансової допомоги Україні, обіцяє більше засобів ППО RAW:  0.0 NORM:  0.25
15 Ворог пошкодив ракетами всі українські ТЕС – Галущенко RAW:  -0.4588 NORM:  -0.4588
16 Публічні позиції України та Росії щодо війни за тиждень – ОПОРА RAW:  0.0 NORM:  -0.25
17 Міністри ЄС не змогли погодити 9-ий пакет санкцій проти Росії RAW:  0.0 NORM:  0.0
18 "Укрнафта" ініціює позапланові перевірки у компанії RAW:  0.0 NORM:  -0.25
19 ЄС ввів нові санкції проти Ірану через постачання Росії безпілотників RAW:  0.0 NORM:  0.0
```

Fig. 3. The result of the VADER algorithm after word normalization

After these actions, some results became significantly better compared to the initial results of the analysis.

3.2.2. Classification of texts. Text classification is the process of dividing documents into certain categories. It is important not to confuse classification with clustering: in clustering, texts are grouped according to criteria that are not defined in advance. Classification can be done both manually and automatically, using a custom set of rules or machine learning methods to classify documents into one or more categories.

Text categorization has many applications such as sentiment analysis, topic tagging, news classification, language detection, intent detection, spam detection, customer routing, resume classification, and more.

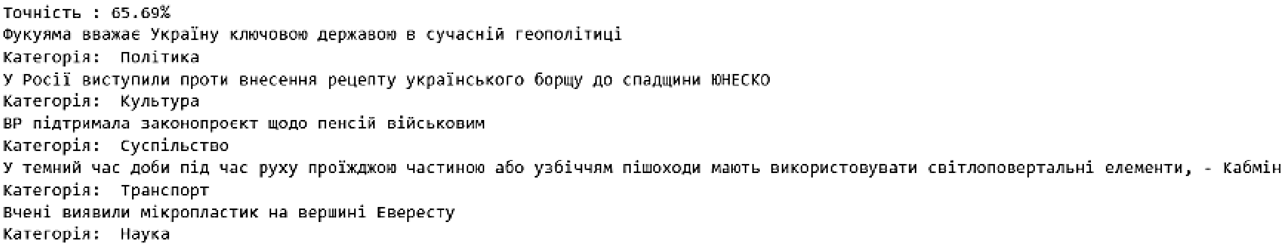
Text data cannot be directly used for machine learning, so it must be converted to numerical values. In order not to complicate the task, it is possible to use the TF-IDF conversion method, the principle of which has already been described in this paper.

To classify our marked news, it is possible to choose the LinearSVM method, which is based on the method of support vectors (SVM). This method belongs to linear classifiers and focuses on minimizing the empirical classification error [26].

The main idea of SVM is to translate the initial vectors into a higher-dimensional space and find the hyperplane with the largest «gap» in it, so this method is also called the classification method with the largest «gap». Two parallel hyperplanes are constructed on either side of the hyperplane separating the classes. The best hyperplane is the one that provides the largest distance between two parallel hyperplanes, which reduces the probability of misclassification.

Now it is possible to select several news items and see to which topic the proposed classifier based on the support vector method assigns them (Fig. 4).

As it is possible to see, the classifier was immediately able to show a very good result in news classification.



```
Точність : 65.69%
Фукуяма вважає Україну ключовою державою в сучасній геополітиці
Категорія:  Політика
У Росії виступили проти внесення рецепту українського борщу до спадщини ЮНЕСКО
Категорія:  Культура
ВР підтримала законопроєкт щодо пенсій військовим
Категорія:  Суспільство
У темний час доби під час руху проїжджою частиною або узбіччям пішоходи мають використовувати світлоповертальні елементи, - Кабмін
Категорія:  Транспорт
Вчені виявили мікропластик на вершині Евересту
Категорія:  Наука
```

Fig. 4. The result of the classifier work

3.2.3. Thematic modeling. The main tasks of thematic modeling are formulated as follows:

- Capturing semantic information not only at the level of individual words, but also beyond them.
- Detection of not only obvious topics in documents, but also hidden variations of topics.
- Abstract of documents.
- Using annotations to manage content summarization, search, and recommendation.

In natural language processing or machine learning, a topic model is a statistical model that allows to discover hidden «topics» in a collection of documents. Thematic modeling is often used in the intelligent analysis of texts to reveal hidden semantics, which allows efficient analysis of large volumes of texts by document clustering.

Large text arrays are usually under-recognized, which means that many of the previous learning methods discussed above cannot be applied. Texts may not have convenient labels (for example, positive or negative). However, they can have many different topics, as in newspaper articles, and only topic modeling can detect them.

To prepare the data, let's first define the parameters of the vectorizer:

- the max_df value determines what percentage of words or terms to ignore. If max_df=0.2, it is possible to ignore words/terms that occur in more than 20 % of documents;
- a value of min_df=3 means that if a word/term occurs in less than 3 documents, it is possible to ignore it;
- max_features is the size of the subset of features used in node splitting.

Such restrictions allow to reject words or terms that occur in many documents in the collection, and to remove random or misspelled words.

First, a training dataset (for 50 values) and a random dataset were created, LDA (Latent Dirichlet Allocation) was performed for them, and the resulting topics were analyzed. For this, a cloud of tags [27] is derived for each topic, which gives an opportunity to see how it looks for all the divisions under consideration. The intersection of topics and categories is presented in Fig. 5.

It is also possible to see how topics and categories intersect (Fig. 6).

As can be seen, categories such as «Coronavirus», «Science», «Politics», «Sports» and «Transport» are most clearly distinguished (news that were in the information field at the end of 2022 were taken for analysis).

Tools of the pyLDAvis library provide a clearer visualization (Fig. 7).

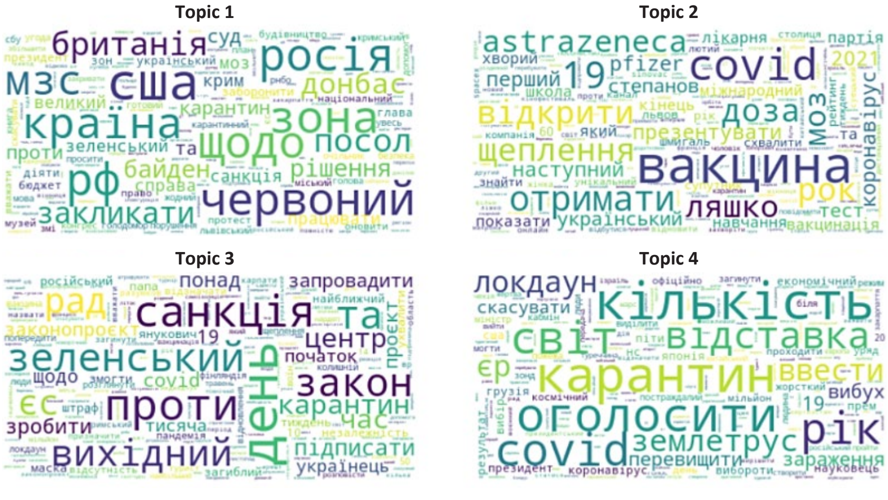


Fig. 5. The result of thematic modeling based on the tag cloud

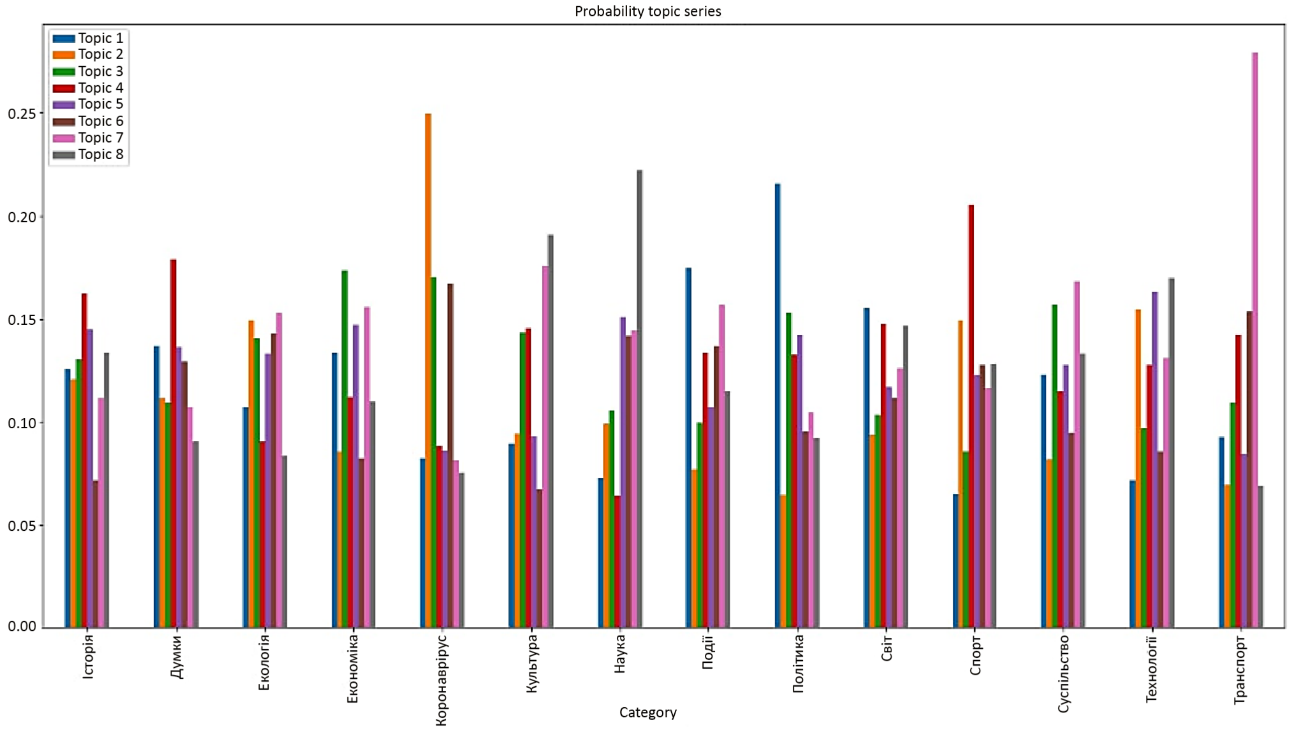


Fig. 6. The results of intersection of topics and categories

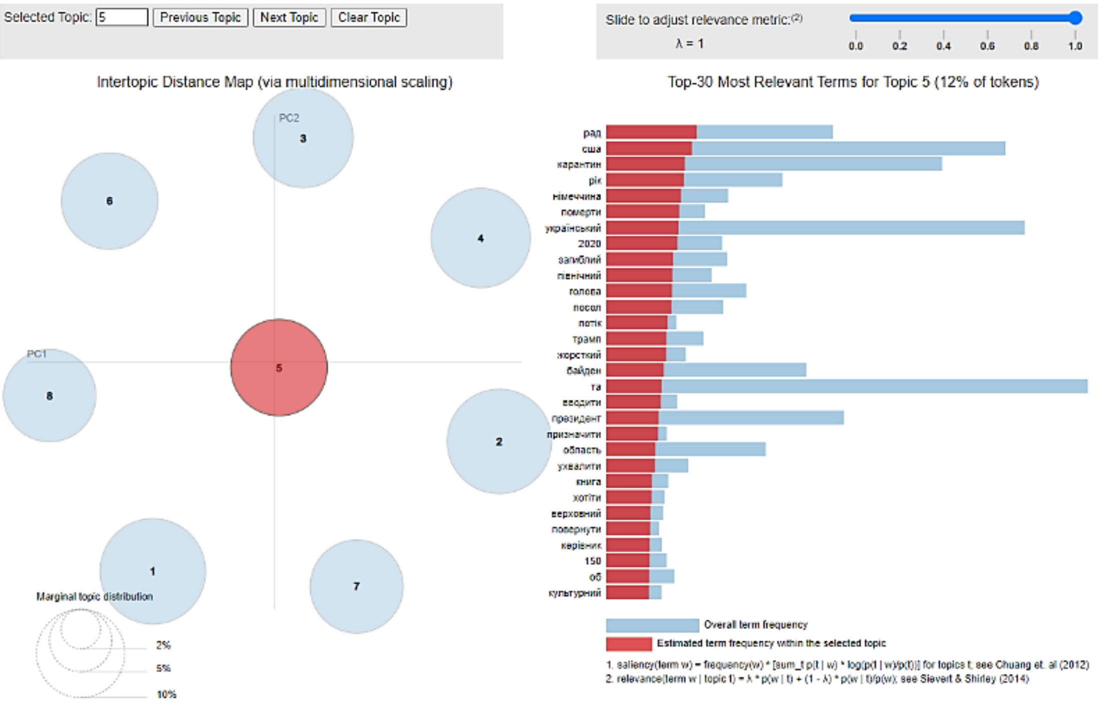

Selected Topic: 5
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
7
8
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 5 (12% of tokens)
рад
сша
карантин
рік
німеччина
померти
український
2020
загиблий
північний
голова
посол
потік
трамп
жорсткий
байден
та
вводити
президент
призначити
область
ухвалити
книга
хотіти
верховний
повернути
керівник
150
об
культурний
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 7. Visualization of the result using pyLDAvis tools

3.3. Discussion. The research results show that the use of models based on neural networks ensures the necessary informativeness of the obtained automatic abstracts. This can be explained by the ability of these models to take into account the context and structure of the text, which allows generating more relevant and meaningful abstracts.

Practical significance: The obtained results can be applied in many areas where it is necessary to quickly and efficiently obtain key information from large volumes of texts. For example, in media and journalism, automatic abstracting can be used to create short news reviews, making the work of journalists and editors easier. In the field of education, automatic referencing can help students and researchers quickly find the information they need from scientific articles and other sources. Also in business analytics, these methods can be used to analyze reports, market research and other text documents.

Limitations of the study: One of the main limitations of the study is the dependence of the results on the quality and quantity of training data. Generative models require large amounts of text for training, and the quality of the abstracts may deteriorate with insufficient or low-quality input data. Deep learning models can be resource-intensive, which can limit their use in systems with limited computing resources.

To implement the obtained results in practice, it is necessary to continue research in the direction of optimization of models and algorithms, which will reduce their computational complexity and increase the efficiency of working with large volumes of data.

The conditions of martial law in Ukraine increase the relevance of the task of automatic abstracting of texts. In particular, researchers and developers of new equipment and technologies need to quickly analyze large volumes of information for conducting experiments and creating new samples of hardware and software products. Changes in the education system, such as the transition to distance learning, also affect research processes, in particular, opportunities for collective work and interaction with other researchers.

Prospects for further research include the development and improvement of deep learning models for automatic abstracting of texts. In particular, one should focus on the development of methods that will reduce the computational complexity of models and increase their efficiency. Also, a promising direction is researching the possibilities of using automatic abstracting for multimedia content, which will allow creating complex abstracts, including text, audio and video. In addition, it is worth continuing research in the direction of creating personalized referencing systems that will take into account the individual needs of users and the context of their use.

4. Conclusions

Analysis of the relevance of work in the field of automatic referencing of natural language texts using artificial intelligence methods allows to understand the importance of this technology in the context of the growing volume of textual information and the need to quickly obtain meaningful consolidated information. Based on the analysis, the following conclusions can be drawn:

The need for new solutions: Existing methods of automatic abstracting of texts have limitations in terms of accuracy, scope of analysis and universality. This creates a need for the development of new, more efficient and intelligent approaches to compression of textual information.

Advantages of artificial intelligence techniques: The use of artificial intelligence techniques, including machine learning and natural language processing, allows to create models that can automatically identify key elements of text and generate condensed versions.

Development of a specific task: The main task of the work is the development of a specific algorithm or model for automatic abstracting of texts, which will combine advanced methods of artificial intelligence taking into account the requirements of efficiency and accuracy.

Development opportunities: The development of new approaches to automatic abstracting of texts can have wide practical applications, such as the automated generation of brief reviews of texts in journalism, scientific research, medicine and other fields.

Therefore, based on the analysis of the relevance of the work, it can be concluded that the development of new solutions in the field of automatic abstracting of texts is an important direction of research using artificial intelligence methods, which has the potential to improve the accessibility and quality of textual information analysis.

Conflict of interest

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

Financing

The study was performed without financial support.

Data availability

Manuscript has no associated data.

Use of artificial intelligence

The authors have used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

References

1. Pustejovsky, J., Stubbs, A. (2012). *Natural Language Annotation for Machine Learning*. Cambridge: Farnham, 343.
2. *Natural language processing*. Available at: https://en.wikipedia.org/wiki/Natural_language_processing
3. Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development, 2 (2),* 159–165. https://doi.org/10.1147/rd.22.0159
4. Guarino, N., Masolo, C., Vetere, G. (1999). Content-Based Access to the Web. *IEEE Intelligent Systems*, 70–80.
5. Lin, C.-Y., Hovy, E. H. (2000). The Automated acquisition of topic signatures for text summarization. *Proceedings of COLING-00.* Saarbrücken, 495–501. https://doi.org/10.3115/990820.990892
6. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41 (6),* 391–407. https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9
7. Gong, Y., Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 19–25. https://doi.org/10.1145/383952.383955
8. *PageRank*. Available at: https://en.wikipedia.org/wiki/PageRank
9. Kupiec, J., Pedersen, J., Chen, F. (1995). A trainable document summarizer. *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval – SIGIR'95.* ACM, 68–73. https://doi.org/10.1145/215206.215333
10. Ouyang, Y., Li, W., Li, S., Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing &amp; Management, 47 (2),* 227–237. https://doi.org/10.1016/j.ipm.2010.03.005
11. Wong, K.-F., Wu, M., Li, W. (2008). Extractive summarization using supervised and semi-supervised learning. *Proceedings of the 22nd International Conference on Computational Linguistics – COLING '08*, 985–992. https://doi.org/10.3115/1599081.1599205
12. Zhou, L., Hovy, E. (2003). A web-trained extraction summarization system. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology – NAACL '03*, 205–211. https://doi.org/10.3115/1073445.1073482
13. Conroy, J. M., O'leary, D. P. (2001). Text summarization via hidden Markov models. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM*, 406–407. https://doi.org/10.1145/383952.384042
14. Shen, D., Sun, J.-T., Li, H., Yang, Q., Chen, Z. (2007). Document Summarization Using Conditional Random Fields. *IJCAI, 7*, 2862–2867.
15. *SummarizeBot*. Available at: https://www.summarizebot.com/about.html
16. *SMMRY*. Available at: https://smmry.com/about
17. *Generative Pre-trained Transformer.* Available at: https://openai.com/chatgpt
18. Devlin, J., Chang, M.-W. (2018). *Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing.* Available at: https://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/
19. *TextTeaser*. Available at: https://pypi.org/project/textteaser/
20. *Answers to Frequently Asked Questions about NLTK* (2022). Available at: https://github.com/nltk/nltk/wiki/FAQ
21. *Gensim*. Available at: https://radimrehurek.com/gensim/intro.html#what-is-gensim
22. *SUMY*. Available at: https://github.com/miso-belica/sumy
23. *Bert-Extractive-Summarizer.* Available at: https://github.com/dmmiller612/bert-extractive-summarizer
24. *Ukrainska pravda*. Available at: https://www.pravda.com.ua/
25. *Pymorphy2 0.9.1.* Available at: https://pypi.org/project/pymorphy2/
26. *Support vector machine*. Available at: https://en.wikipedia.org/wiki/Support_vector_machine
27. *Tag Cloud*. Available at: https://en.wikipedia.org/wiki/Tag_cloud

✉***Oleksii Kuznietsov,*** *PhD Student, Department of System Design, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine, e-mail: oleksiy1908@gmail.com, ORCID: https://orcid.org/0000-0002-3537-9976*

Gennadiy Kyselov, *PhD, Senior Researcher, Associate Professor, Department of System Design, National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute», Kyiv, Ukraine, ORCID: https://orcid.org/0000-0003-2682-3593*

✉*Corresponding author*