UDC 005.2

Olha O. Mezentseva[1], Candidate of Economic Sciences, Assistant of the Technology Management Department, E-mail: olga.mezentseva.fit@gmail.com, Scopus ID: 57210290327, ORCID: https://orcid.org/0000-0002-8430-4022
Anna S. Kolomiiets[1], Candidate of Economic Sciences, Assistant of the Technology Management Department, E-mail: anna.tsesliv@gmail.com, Scopus ID: 57200182743,
ORCID: https://orcid.org/0000-0003-4252-5975
[1]Taras Shevchenko National University of Kyiv, Volodymyrska Street, 60, Kyiv, Ukraine, 01033

OPTIMIZATION OF ANALYSIS AND MINIMIZATION OF INFORMATION LOSSES IN TEXT MINING

Annotation. *Information is one of the most important resources of today's business environment. It is difficult for any company to succeed without having sufficient information about its customers, employees and other key stakeholders. Every day, companies receive unstructured and structured text from a variety of sources, such as survey results, tweets, call center notes, phone emails, online customer reviews, recorded interactions, emails and other documents. These sources provide raw text that is difficult to understand without using the right text analysis tool. You can do text analytics manually, but the manual process is inefficient. Traditional systems use keywords and cannot read and understand language in emails, tweets, web pages, and text documents. For this reason, companies use text analysis software to analyze large amounts of text data. The software helps users retrieve textual information to act accordingly The most common manual annotation is currently the most common, which can be attributed to the high quality of annotation and its "meaningfulness". Typical disadvantages of manual annotation systems, textual information analysis systems are the high material costs and the inherent low speed of work. Therefore, the topic of this article is to explore the methods by which you can effectively annotate reviews of various products from the largest marketplace in Ukraine. The following tasks should be solved: to analyze modern approaches to data analysis and processing; to study basic algorithms for data analysis and processing; build a program that will collect data, design the program architecture for more efficient use, based on the use of the latest technologies; clear data using minimize information loss techniques; analyze the data collected, using data analysis and processing approaches; to draw conclusions from the results of all the above works. There are quite a number of varieties of the listed tasks, as well as methods of solving them. This again confirms the importance and relevance of the topic we choose. The purpose of the study is the methods and means by which information losses can be minimized when analyzing and processing textual data. The object of the study is the process of minimizing information losses in the analysis and processing of textual data. In the course of the study, recent research on the analysis and processing of textual information was analyzed; methods of textual information processing and Data Mining algorithms are analyzed.*

Keywords: *text analysis; annotation; text mining; software; algorithm; text data; natural language*

Introduction. The relevance of text mining today is determined by the need to process large amounts of textual information accumulated by mankind in the global information space. Plain text written by a human being in natural language is poorly structured information, so processing it is not an easy task beyond the traditional algorithmic processing of structured data. In order to receive the target information from the texts, it is necessary to structure them, arrange them, organize them, to provide search of texts at the user's request. We live in a time when the volume of human information produced is greater than ever and the amount of textual data is increasing day by day. However, significant benefit from this information can only be obtained by properly processing and analyzing this data.

Tasks of word processing – unstructured documentation, medical records, patents and dissertations, etc. – can be divided into two conditional categories.

The first is the tasks that every user faces every day: spell check, spam filtering, automatic translation of small text fragments (several sentences). From the point of view of automatic word processing (AWP) researchers, all these tasks are almost solved, and today more relevant tasks in the second category, which require processing large text arrays: finding relevant answers to the questions (question-answer), full-fledged machine translation of complete texts, analysis of responses, construction of recommendation systems that work with large arrays of unstructured data. A distinctive feature of such tasks is their complexity and lack of formalization, which leads to the fact that they do not yet have a complete set of solutions, and are used auxiliary methods of highlighting keywords and phrases, summarization (automatic abstracting) of texts and classification of texts [1].

Analysis of the structured information stored in databases requires preliminary processing: database design, input of information on the selected rules, placement of it in special structures (eg relational tables), etc. Thus, additional effort is needed to analyze this information and gain new knowledge directly. However, they are not always related to the analysis and do not necessarily lead to the desired result.

© Mezentseva, O. O., Kolomiiets, A. S., 2020
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/deed.uk)

Because of this, the Structural Information Efficiency (SEF) is reduced. In addition, not all types of data can be structured without losing useful information.

Analysis methods in unstructured texts lie at the intersection of several areas: Data Mining, Natural Language Processing, Information Search, Information Retrieval and Knowledge Management.

Every day, companies receive unstructured and different-quality structured text from a variety of sources, such as survey results, tweets, call center notes, telephone emails, online customer reviews, recorded interactions, emails and other documents. These sources provide raw text that is difficult to understand without using the right text analysis tool. But traditional manual text analysis methods to identify major topics and trends in large amounts of data are ineffective today.

Literature review. Traditional text mining systems use keywords and cannot read and understand the language in emails, tweets, web pages, and text documents. For these reasons, companies use text analysis software to analyze large amounts of textual data. Let's take a brief look at existing software that helps users retrieve information from text data.

Most software products process textual information on user requests using commercial database management systems. The incorporation of analytical capabilities into commercial DBMSs is a natural trend and has great potential. Today, data analysis functionality is more productively implemented in the following commercial databases: Oracle, Microsoft SQL Server, and IBM DB2. Each of the given DBMS allows to solve the main problems related to data analysis. However, only Oracle can be considered a truly analytical platform. In addition to implementing data mining, the platform has powerful tools for analyzing unstructured text information – Oracle Text. The platform allows you to use the SQL language for querying. This approach is used in the latest version of Oracle 11g.

"Oracle Text" provides the following tasks: search for documents by their content; classification and clustering of documents; removal of key concepts; automatic annotation; search associative link documents and more.

The main tasks that Oracle Text tools aim at are the task of finding documents by their content - by vocabulary or phrase, which are combined with Boolean operations if necessary. Search results begin by relevance, taking into account the frequency of occurrence of search words in the documents found. To improve search quality, Oracle Text provides several tools that extend search engine capabilities:

1. Expansion of the query words of all morphological forms realized by the attraction to the knowledge of morphology.

2. Expansion of query words related to the content of the word by connecting a thesaurus – a semantic dictionary.

3. Expanding the search for words close to the words and sounds – fuzzy search and search for consonant words. Fuzzy Search is intended for use when searching for misspelled words, as well as when there are doubts about the correct spelling of the name, organization name, etc.

All of the tools described can be used together, which supports language query in traditional SQL and PL / SQL syntax for document search. Oracle Text system provides the opportunity to work with modern advanced tools in the context of complex multiple search and analysis of text data [6].

The ability to process textual information in Russian in Oracle Text is quite limited. A Russian Contextual Optimizer (RCO) module designed to share InterMedia (or Oracle Text) text was used to solve this problem. In addition to Russian language support, RCO includes tools for fuzzy search, case analysis, and document abstracting [17].

IBM Tool – Intelligent Miner for Text. IBM Intelligent Miner for Text is a product of more than 100 individual utilities running from the console or from scripts independently. The system contains the following basic utilities for solving textual information analysis tasks: Language Identification Tool – an automatic definition of the language in which the document is compiled; classification utility (Categorization tool) – automatic classification of text into a certain category (input information about the work of clustering of the tool); Clusterisation Tool – splitting a large number of documents into groups with similar style, forms, different frequency characteristics, which are expressed by keywords; Feature Extraction Tool - detection in keywords documents (proper names, names, abbreviations) based on the analysis of a given word; Annotation Tool – Annotations to the original texts.

IBM Intelligent Miner for Text combines powerful resource support for tools that use basic information retrieval (retrieval) tools. Text development: Search engine – information retrieval system; Web crawler – Web space scanning utility; Net Question Solution – a solution for searching on a local website or on several intranets / Internet servers; Java Sample GUI – A set of Java Beans interfaces to administer and organize text search engine searches.

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

IBM Intelligent Miner for text product Included in the Content Information Integrator for DB2 DBMS as a means of analyzing information [7].

Software application from SAS Institute – Text Miner. The American company SAS Institute has released this system for comparing certain grammatical and verbal lines in writing. Text Miner is a versatile system because it can handle various text documents in databases, file systems and even the Web [3].

The author [18] shows that an example of successful use of the Text Miner capabilities is the activity of Compaq Computer Corp., which is currently testing Text Miner, analyzing over 2.5 GB of text documents received by e-mail and collected by company representatives. It was practically impossible to process such data before.

Text Miner software allows you to determine how true a text document is [2]. Detecting "lies" in documents is done by analyzing text and finding out changes in message style that may be displayed when requesting or viewing information. At the same time, Text Miner includes a large set of documents that differ from the true ones, the structure of which is used as templates. A document that is being "examined" for accuracy is compared to these samples. After the analysis, the software application assigns the document a certain index of correctness. The Text Miner system can be especially useful for organizations that handle large volumes of electronic correspondence or for law enforcement agencies, in the case of analyzing text documents based on observation of a person's emotional state [2].

An analysis of existing text data analysis systems has led to the following conclusions: the potential for the development of such systems is still present, but most of them have certain disadvantages associated with information losses. Therefore, it is important to solve the problem of optimizing the analysis of text information and minimizing information loss by the following criteria:

– Increasing adaptability of analysis from various sources: emails, news and other multimedia sources.

– Reduction of time and memory used to code

– increase the level of text structure.

The purpose of the article is to investigate the methods by which the information losses in the analysis and processing of textual data can be minimized in order to find the most efficient algorithm of analysis. Object – The process of minimizing information loss when analyzing and processing text data. The subject of the study is methods, algorithms and tools for analyzing and processing textual data.

Main part. Machine learning was first named "an area of learning that enables computers to learn without explicit programming" by Arthur Samuel in 1959. A more formal definition is given by T. Mitchell: "A computer program is said to be learning from the experience of E with respect to some class of problems T and the productivity index P, if its performance at tasks in T as measured by P improves with the experience of E" [9].

Regardless of the platform used, machine learning is implemented by a relatively simple mechanism (Fig. 1).

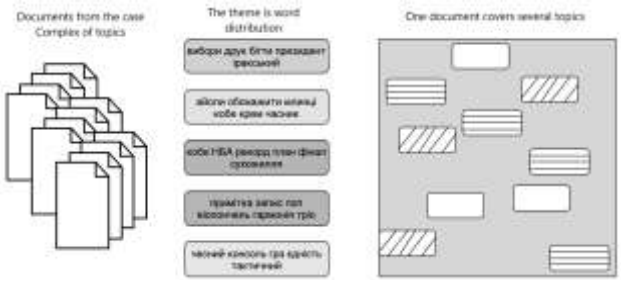

Fig. 1. The mechanism of artificial intelligence for Text Mining [8]

Natural-language processing (NLP) – a general algorithm of artificial intelligence and mathematical linguistics, designed for word processing and so-called natural language and studies the problems of computer analysis and synthesis of natural language.

Currently, many applications are addressed in the literature, which are solved by analyzing text documents. These are the classic Data Mining tasks: classification, clustering, and specific only to text documents of the task: automatic annotation, removal of key concepts and more. The primary purpose of feature extraction is to identify facts and relationships in the text. In most cases, such terms are nouns and generic names: names of people, names of organizations, and so on. Concept retrieval algorithms can use dictionaries to identify some terms and linguistic patterns to define others.

Text-base navigation allows users to navigate through topic documents and meaningful terms. This is done by identifying key concepts and some relationships between them.

Trend analysis allows you to identify trends in document sets over a period of time. The trend can be used, for example, to identify changes in the interests of a company from one market segment to another.

Searching for associations is also one of the main tasks of Data Mining. To solve it in a given set

of documents, associative relations between key concepts are identified.

Summarization allows you to reduce text while maintaining its meaning [8]. The solution to this task is usually governed by the user by determining the number of proposals extracted or the percentage of text extracted relative to the entire text. The result includes the most significant sentences in the text. According to Article [29], the annotation process consists of three steps:

1. Analysis of the source text.
2. Determination of its characteristic fragments.
3. Formation of appropriate conclusion.

There are two main approaches to the automatic annotation of text documents: extraction - involves the selection of the most important fragments (often sentences) from the source text and merging them into annotation; generalization - involves the use of previously developed natural language grammars, thesauri, ontological reference books, etc. On the basis of which the reformulation of the source text and its generalization is performed [4].

In a template-based approach to extracting fragments, the most lexically and statistically significant parts are distinguished. As a result, the annotation in this case is created by simply connecting the selected fragments.

But when using large text datasets, using simple statistics is not effective. With 30 or 40 million detailed shopping records, it's not enough to know that two million of them are made in the same place. To better meet the needs of buyers, you need to understand whether these two millions belong to a certain age group or know the average salary of buyers.

The complexity of business requirements has led to the fact that classical statistical analysis is not sufficient to handle such arrays of information. To solve these problems, you need to use text-based data mining tools that will not only create a model for describing information, but also produce a resulting report. (Fig. 2). We can use thematic modeling tools to do this.

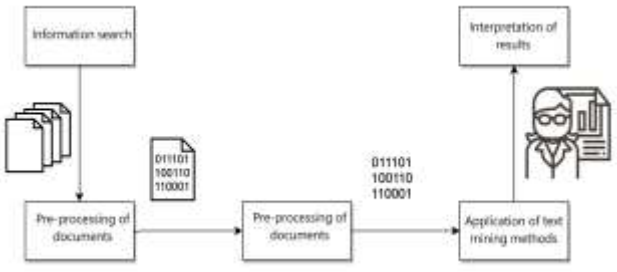

Fig. 2. Text Mining Stages (Dirichlet Scheme) [5]

Thematic modeling is a way of constructing a text document collection model that defines which topics each document belongs to. Moving from a space of terms to a space of found topics helps you to solve tasks such as thematic search, classification, summarization and annotation of document collections and news feeds more effectively. To use thematic modeling methods, you need to create a custom pipeline that will extrapolate topics from unstructured text data and develop an algorithm to save the best model so that it can then be used to analyze new data [11].

The word2vec algorithm was considered, the output of which generates vector representations of words. Word vectors underlie many natural language processing (NLP) systems. Vector views are used by Amazon, Google translate and others [6].

Word vectors are numerical representations of words that retain a semantic link between them. For example, for the vector “cat” one of the closest will be the word “dog”. However, the vector representation of the word “pencil” will be quite different from the vector “cat”. This similarity is caused by the frequency of occurrence of incorrect translation of two words (ie [cat, dog] or [cat, pencil]) in the same context (Fig. 3).

Fig. 3. Vector representation of words

Word2vec algorithms use context to form numerical representations of words, so words used in the same context have similar vectors.

To implement the algorithm, the authors propose to use neural network technologies.
Preparation of data for training.

We construct the training data set for the word2vec neural network as follows [3]:

1. Clear the input text T from unnecessary characters (punctuation marks, etc.).

2. From the purified text T collect the dictionary W.

For each word $w \epsilon T$ we construct a context, that is, a set of words $C_i \epsilon T$, that are distant from w_i o more than s positions in the sequence of words T.

$$C_i = \{w_j \epsilon T: (i - s)\} \leq j \leq (i + s), j \neq I,$$

in other words – C_i is the word $w_j \epsilon T$ from the s-neighborhood of the word $w_j \epsilon T$

1) Perform unitary encoding [5] (one-hot encoding) of the dictionary W, that is, to match each

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

word $w_j \epsilon W$ put to the vector $u_i \epsilon U$ of zeros and one unit, the length of the vector u_i is equal to the size of the dictionary W, the position of the unit in the vector u_i corresponds the word number in the *W*.

2) Replace the words in the *T* text and C contexts with the corresponding P and *Q* codes from *U*.

Thus, obtain two sets – coded text *P* and sets of coded words of context *Q* (Fig. 4).

P_i:

0 0 1 0 0

Q_i:

0 1 0 0 0

0 0 0 0 1

0 0 0 1 0

1 0 0 0 0

0 0 0 1 0

Fig. 4. Code examples (word and context)

The result of Word2Vec is a set of vectors (matrix) – word codes that are formed by training a particular neural network on some text (an ordered set of words). For the research and training of the network we used a set of reviews on the product area prom.ua for the period 2015-2020.

The practical implementation of the algorithm is performed using the Pandas library of the Python language. The library considers the values of None and NaN as interchangeable means of indicating missing or empty values.

The several methods to identify and clear data from gaps (Fig. 5) were used:

Dropping the NaN / none values from the dataset

```
df = df.replace('none', np.nan)
```

```
df.isnull().sum()
```

```
Review    2551
Name         0
Date         0
Rating       0
dtype: int64
```

```
df = df.dropna()
```

```
df = df.reset_index(drop=True)
```

```
df.head()
```

	Review	Name	Date	Rating
0	все клас, швидка доставка!)	Илья	2015-04-06	Хорошо
1	Куплечкою дуже задоволена. Дуже дякую! Все пр...	Ірина	2015-03-20	Отлично
2	Всем довольна сумочка пришла бомбанная!Спасибо	Яна	2015-02-27	Отлично
3	я заказывала на сайте. но оказалось что данно...	Ирина	2015-02-22	Хорошо
4	Работают на 100% и доставлять людям счастье ум...	Наталья	2015-02-11	Отлично

Fig. 5. The result of the performed functions

The reviews under review are very diverse. First of all, you need to clear the data of various characters, extra spaces, and emoji. Let's create a function that will receive the input in the original form, and the output to generate a response, cleared of unnecessary characters and stops.

Stopwords are words that are manually excluded from the text because they are often found in corpus documents (Fig. 6).

We use the CountVectorizer module in the sklearn library, which allows you to convert typing to a token matrix.

That is, let's create a Bag-of-Words or “bag” of words – a matrix that will be fed to the entrance to the LDA. For convenience, we will use the built-in library of Russian scikit-learn stop words using stop_words = StopWords.

Set the maximum document frequency for words by 10 % (max_df = .1) to exclude words that are commonly found in documents.

The following code example demonstrates how to adjust the LatentDirichletAllocation estimator to the Bag-of-Words word matrix and remove 10 different topics from the documents (Fig. 6).

```
from sklearn.decomposition import LatentDirichletAllocation

lda = LatentDirichletAllocation(n_topics=10,
                                random_state=123,
                                learning_method='batch')
X_topics = lda.fit_transform(X)
```

Fig. 6. Using the LatentDirichletAllocation module

By setting learning_method = 'batch', we allow the LDA assessment to perform an assessment based on all available learning data (Bag-of-Words matrix) for one iteration, which is slower than the alternative online learning method. But the results will be more accurate.

Results were visualized using the pyLDAvis library. This library is designed to help users interpret topics in a thematic model that matches the text data corpus.

The algorithm was adapted to highlight five different thematic contexts (DocumentTermMatrix). The number of thematic contexts can be varied depending on the level of granularity that is required in the simulation. The results are shown in Fig. 7. Topics 3 and 4 appear to be closely related, topic 5 is partially related to the general context, and topics 1 and 2 are not related to each other. Topics 1 and 5 cover relatively different topics in the review.

The algorithm developed allows you to analyze what words are used to write reviews and see the overall emotional picture.

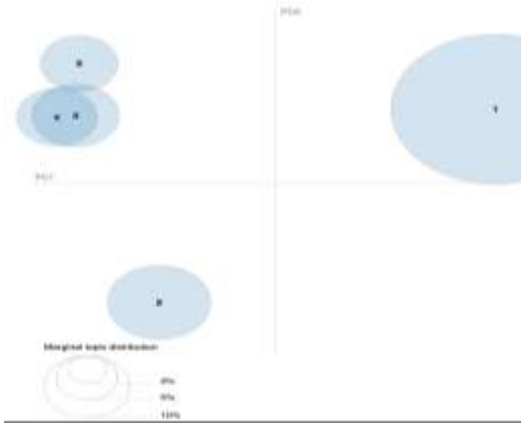

Fig. 7. A map of topics from a sample of words

Fig. 8 shows the plurality of terms a cloud is formed of the most common words / phrases used in reviews.

Annotation (summarization) is based on the ranking of each individual sentence using a variation of the TextRank algorithm.

TextRank is a general automatic annotation algorithm that performs chart-based ranking for natural language processing.

Graph-based ranking algorithms are a means of determining the importance of a vertex within a graph, and works well with responses that contain more than 10 lines of text (Fig. 9). If the reviews are less than ten lines long, then using the library is not effective.

The next step is to examine the semantic similarities between words in reviews. Let's find out the next words to the left “thank you” and “in time”(Fig. 10).

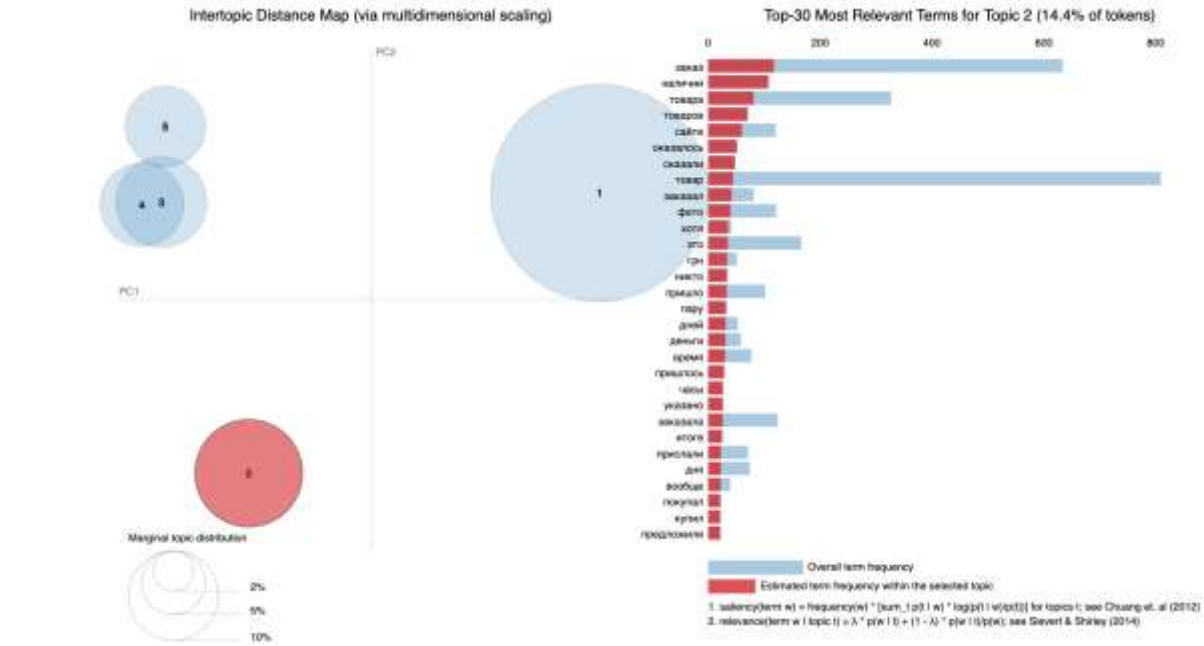

Fig. 8. Visualization of the words used in the second topic
(using wordcloud library)

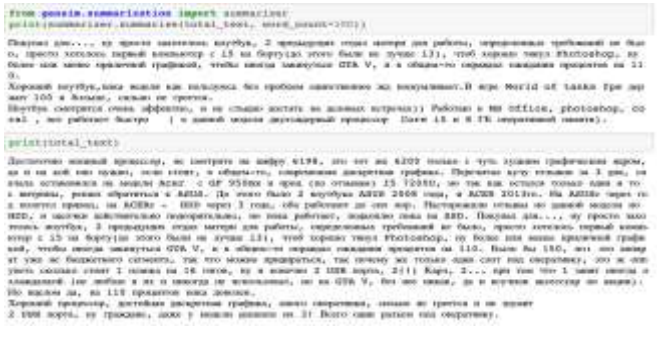

Fig. 9. Results of the TextRank algorithm

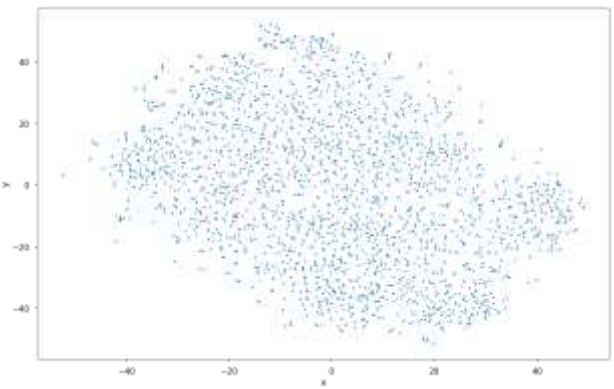

Fig. 10. Visualize word vectors from reviews

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

The last step is to create a direct propagation neural network using Skip-gram or CBOW models.

Using the CBOW model, the neural network learns to predict which word fits into the word sequence, and in which case the word is deliberately deleted. For example, given the sequence "Good quality goods, _____ very fast" the network will predict that the word "delivered" is missing. However, a neural network can only work with one word.

Therefore, it is necessary to turn the sentence into several training (input, target) pairs: (goods, delivered), (quality, delivered), (very, delivered) and (quickly, delivered).

The Skip-gram model is the complete opposite of the CBOW model. Given the input word, it is a neural network that predicts what words surround it [10]. For example, the word "delivered" would imply the words "Good, very, very fast." As with CBOW, it is necessary to present the phrase in the form of several pairs (delivered, goods), (delivered, quality), (delivered, very), (delivered, quickly)

In addition to Word2Vec, there are other word embedding models. It is worth noting the model proposed by Stanford University's GloVe Computational Linguistics Laboratory, which combines the features of SVD decomposition and Word2Vec. This algorithm is a machine learning algorithm without a teacher. Learning is done by creating a co-occurrence matrix (words X context) that basically calculates how often a word appears in a particular context [15]. That is, we have performed all the preparation steps to implement the Word2Vec algorithm, similar to the GloVe algorithm.

But the results of both algorithms are different. Practice shows that the best results are demonstrated by the GloVe algorithm (Fig. 11). At the fundamental level, the two classes of methods are not different, but the efficiency with which counting methods capture global statistics may differ [33]. GloVe achieves better results faster and also has the best results regardless of speed compared to the Word2Vec algorithm (Fig. 11)

Data preprocessing and refinement significantly improves results, making more efficient use of the data set for machine learning. Using raw data may result in poor results or may not produce the expected results at all.

Through experimental studies, LDA, TF-IDF, and GloVe were found to be more effective methods for summarizing responses (Fig. 11).

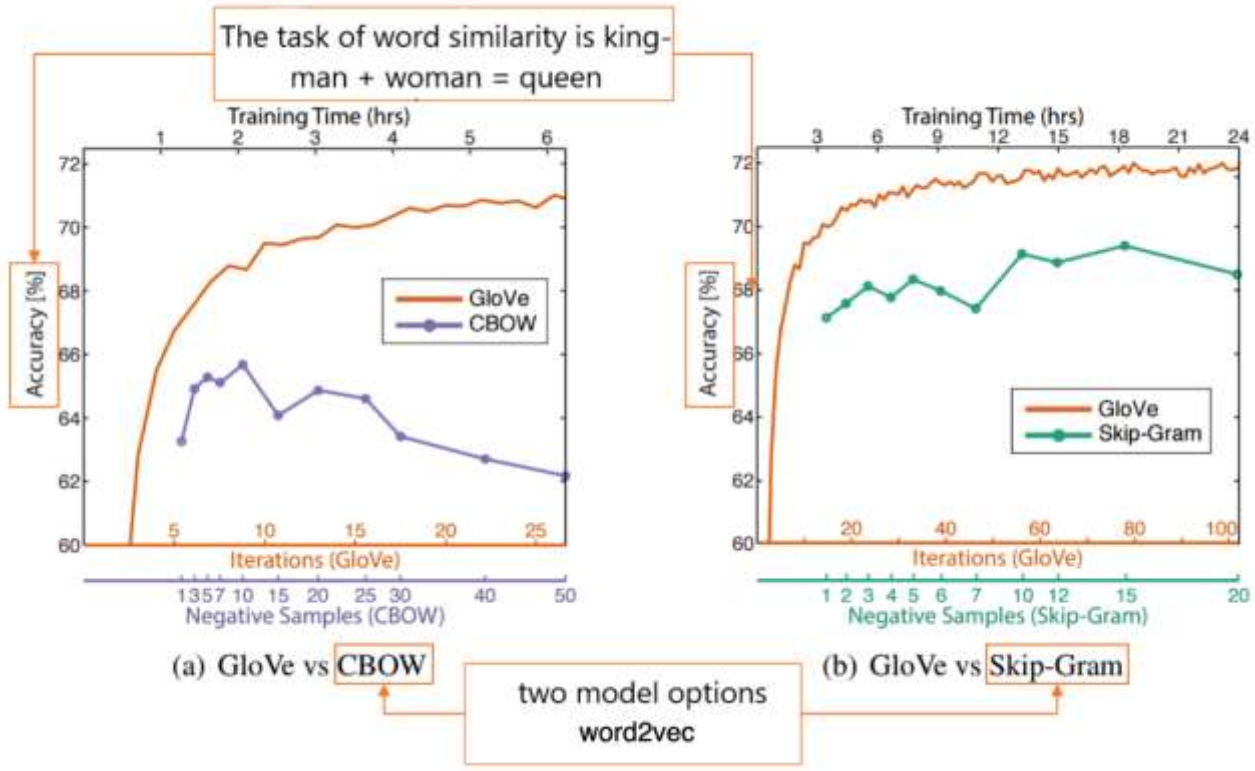

Fig. 11. Comparison of two algorithms of models of processing of responses

Conclusions. The main problems that exist in the field of analysis and processing of large unstructured arrays of text data are investigated. The main software products used for analyzing text information are examined and their effectiveness and capabilities remain rather limited. The role of machine learning and artificial intelligence in natural language processing has been determined. It is to accelerate and automate the basic functions of textual information processing, to transform huge arrays of unstructured data into new knowledge without loss of information. The methods and algorithms of data mining that have been applied to acquire new knowledge from raw arrays are considered. A software application was developed to automate the process of collecting and storing the information received in the database. The software application is implemented in Python programming language.

The main areas of natural language analysis were discussed, as well as modern tools and methods used to process textual data. This made it possible to choose the most convenient algorithm for their analysis.

Methods and algorithms implemented using Python programming language and libraries. Since the feedback is short from the start, the minimum response length is 5 words. Short word processing is not a trivial task; it is difficult to make annotation shorter. Therefore, the software implementation of the TextRank algorithm in the gensim library for word processing did not produce significant results.

Through experimental studies, it has been found that more effective methods for summarizing responses are LDA, TF-IDF, GloVe.

The GloVe algorithm performed better than the Word2Vec algorithm. With each iteration of the GloVe algorithm, the accuracy (results) of the obtained results increases by an average of 3-5 %.

The proposed algorithm can be used to annotate reviews in any industry. For example, when a user needs to quickly and accurately identify and remove major ideas or events from emails, news, and other multimedia sources.

References

1. Aronovich, E. (2012). “TF-IDF”. – Available at: https://www.cs.tau.ac.il/~nin/Courses/Workshop13a/tf-idf.pdf. – Access date: 12.01.2020.

2. Barzilay, R. (2011). “Using Lexical Chains for Text Summarization”. – Available at: https://www.aclweb.org/anthology/W97-0703. – Access date: 12.12.2019.

3. Borgman, C. L. (2018). “Text Data Mining from the Author’s Perspective: Who’s Text, who’s mining, and to who’s Benefit?” – Available at: https://arxiv.org/pdf/1803.04552.pdf. – Access date: 24.12.2019.

4. Christopher, M. D. (2014). “The Stanford CoreNLP Natural Language Processing Toolkit”. – Available at: https://www.aclweb.org/anthology/P14-5010. – Access date: 20.01.2020.

5. Kolesnikova, K., Lukianov, D., Gogunskii, V., Olekh, T. & Bespanskaya-Paulenka, K. (2017). “Communication management in social networks for the actualization of publications in the world scientific community on the example of the network researchgate”. *Eastern-European Journal of Enterprise Technologies* . Vol 4, No. 3 (88) , pp. 60-65. – Available at: http://journals.uran.ua/eejet/article/view/108589. – Access date: 10.12.2019.

6. Kolomiets, A. & Tsesliv, O. (2017). “Technologiya pobydovi ta upravlinnya bazami ta shovischami danih (textbook)”. *Publ. KPI*, 281 p. (in Ukranian).

7. Mezentseva, O. (2019). “Intellectualization of enterprise management using business intelligence instruments”. *Eastern-European Journal of Enterprise Technologies* . Vol. 4, No. 3 (88) , pp. 60-65. – Available at: http://journals.uran.ua/tarp/article/view/179264. – Access date: 14.12.2019.

8. Miller, G. A. (1956). “The magical number seven, plus or minus two: Some limits on our capacity for processing information”. *Psychological revie*w, 63(2), pp. 81-97.

9. Morozov, V., Kalnichenko, O., Proskurin, M. & Mezentseva, O. (2019). “Investigation of Forecasting Methods of the State of Complex IT-Projects with the Use of Deep Learning Neural Networks”, *Advances in Intelligent Systems and Computing.* – Available at: https://link.springer.com/chapter/10.1007/978-3-030-26474-1_19. – Access date: 24.01.2020.

10. Morozov, V., Steshenko, G. & Kolomiiets, A. (2017). “Learning through practice in IT management projects master program implementation approach”. *Proceedings of the 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications*. – Available at: https://ieeexplore.ieee.org/document/8095223. – Access date: 15.01.2020.

11. Qingyu, Z. (2018). “Neural Document Summarization by Jointly Learning to Score and Select Sentencesю”. – Available at:

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)

https://arxiv.org/pdf/1807.02305v1.pdf. – Access date: 29.12.2019.

12. Rakshith, V. (2017). "What is One Hot Encoding? Why And When do you have to use it?" – Available at: https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f. – Access date: 12.01.2020.

13. Redmore, S. (2019). "Machine Learning for Natural Language Processing". – Available at: https://www.lexalytics.com/lexablog/machine-learning-vs-natural-language-processing-part-1. – Access date: 17.01.2020.

14. Sinha, S. (2019). "Extractive Text Summarization using Neural Networks". – Available at: https://arxiv.org/pdf/1802.10137.pdf. – Access date: 12.12.2019.

15. Stilo, G. & Velardi, P. (2016). "Efficient temporal mining of micro-blog texts and its application to event discovery, Data Mining and Knowledge Discovery", 30(2), pp. 372-402. – Available at: https://link.springer.com/article/10.1007/s10618-015-0412-3. – Access date: 24.01.2020.

16. "SVM (Support Vector Machine) – Theory". – Available at: https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72. – Access date: 10.12.2019.

17. Wang, F. (2019). "Feature Learning Viewpoint of AdaBoost and a New Algorithm". – Available at: https://arxiv.org/pdf/1904.03953.pdf. – Access date: 17.01.2020.

18. Wong, K. (2008). "Extractive Summarization Using Supervised and Semi-supervised Learning". – Available at: https://www.aclweb.org/ anthology/C08-1124. – Access date: 20.01.2020.

Received 02.02.2020
Received after revision 16.02.2020
Accepted 19.02.2020

УДК 005.2

[1]**Мезенцева, Ольга Олексіївна,** кандидат економічних наук, асистент кафедри технології управління, E-mail: olga.mezentseva.fit@gmail.com, ORCID: https://orcid.org/0000-0002-8430-4022
[1]**Коломієць, Анна Степанівна**, кандидат економічних наук, асистент кафедри технології управління, E-mail: anna.tsesliv@gmail.com, ORCID: https://orcid.org/0000-0003-4252-5975
[1]Київський національний університет ім. Т. Г. Шевченка, вул. Володимирська, 60, Київ, Україна, 01033

ОПТИМІЗАЦІЯ АНАЛІЗУ ТА МІНІМІЗАЦІЯ ІНФОРМАЦІЙНИХ ВТРАТ У TEXT MINING

***Анотація**. Стаття присвячена вирішенню таких завдань: провести аналіз сучасних підходів до аналізу та обробки даних; вивчити основні алгоритми для аналізу та обробки даних; на основі застосування новітніх технологій створити програму, яка буде збирати дані, спроектувати архітектуру програми для більш ефективного використання; очистити дані, застосовуючи методи мінімізації інформаційних втрат; проаналізувати отримані очищені дані застосовуючи підходи до аналізу та обробки текстових даних; зробити висновки за результатами усіх вищезгаданих робіт. Існує досить велика кількість різновидів перерахованих завдань, а також методів їх вирішення Інформація є одним з найважливіших ресурсів сучасного бізнес-середовища. Для будь-якої компанії важко досягти успіху, не маючи достатньої інформації про своїх клієнтів, співробітників та інших ключових зацікавлених сторін. Щодня компанії отримують неструктурований і структурований текст з різних джерел, таких як результати опитування, твіти, нотатки до колл-центру, телефонні розсилки, онлайн-відгуки клієнтів, записані взаємодії, листи та інші документи. Ці джерела надають необроблений текст, який нелегко зрозуміти без використання правильного інструменту аналізу тексту. Можна виконувати аналітику тексту вручну, але процес вручну неефективний. Традиційні системи використовують ключові слова і не можуть читати і розуміти мову в електронних листах, твітах, веб-сторінках і текстових документах. З цих причин компанії використовують програмне забезпечення для аналізу текстів для аналізу великих обсягів текстових даних. Програмне забезпечення допомагає користувачам отримувати інформацію з текстових даних, щоб діяти відповідно В даний час найбільш поширене ручне анотування, до переваг якого можна віднести, безумовно, високу якість складання анотації та її «осмисленість». Типові недоліки ручних систем анотування, систем аналізу текстової інформації - це високі матеріальні витрати і притаманна їм низька швидкість роботи. Тому тематика цієї статті – це дослідження методів за допомогою яких можна ефективно анотувати відгуки про різноманітні товари з найбільшого торгівельного майданчику України. Це ще раз підтверджує значущість і актуальність обраної нами теми. Метою дослідження є методи та засоби за допомогою яких можна мінімізувати інформаційні втрати при аналізі та обробці текстових даних. Об'єктом дослідження є процес мінімізації інформаційних втрат при аналізі та обробці текстових даних. В ході дослідження проведено аналіз останніх досліджень з аналізу та обробки текстової інформації; проаналізовано методи обробки текстової інформації та алгоритми Data Mining.*

***Ключові слова**: аналіз текстової інформації; анотування; інтелектуальний аналіз текстів; програмний продукт; алгоритм; текстові дані; природна мова*

УДК 005.2

[1]**Мезенцева, Ольга Алексеевна**, кандидат экономических наук, ассистент кафедры технологий управлення, E-mail: olga.mezentseva.fit@gmail.com, ORCID: https://orcid.org/0000-0002-8430-4022
[1]**Коломиец, Анна Степановна**, кандидат экономических наук, ассистент кафедры технологий управлення, E-mail: anna.tsesliv@gmail.com, ORCID: https://orcid.org/0000-0003-4252-5975
[1]Киевский национальный университет им. Т. Г. Шевченко, ул. Владимирская, 60, Киев, Украина, 01033

ОПТИМИЗАЦИЯ АНАЛИЗА И МИНИМИЗАЦИЯ ИНФОРМАЦИОННЫХ ПОТЕРЬ В TEXT MINING

***Аннотация.** Информация является одним из важнейших ресурсов современного бизнес-среды. Для любой компании трудно добиться успеха, не имея достаточной информации о своих клиентах, сотрудников и других ключевых заинтересованных сторон. Ежедневно компании получают неструктурированный и структурированный текст из различных источников, таких как результаты опроса, твиты, заметки в колл-центр, телефонные рассылки, онлайн-отзывы клиентов, записанные взаимодействия, письма и другие документы. Эти источники предоставляют необработанный текст, который нелегко понять без использования правильного инструмента анализа текста. Можно выполнять аналитику текста вручную, но процесс вручную неэффективен. Традиционные системы используют ключевые слова и не могут читать и понимать язык в электронных письмах, твитах, веб-страниц и текстовых документах. По этим причинам компании используют программное обеспечение для анализа текстов для анализа больших объемов текстовых данных. Программное обеспечение помогает пользователям получать информацию из текстовых данных, чтобы действовать в соответствии В настоящее время наиболее распространено ручное аннотирования, к преимуществам которого можно отнести, безусловно, высокое качество сборки аннотации и его «осмысленность». Типичные недостатки ручных систем аннотирования, систем анализа текстовой информации - это высокие материальные затраты и присущая им низкая скорость работы. Поэтому тематика этой статьи - это исследование методов с помощью которых можно эффективно аннотировать отзывы о различных товары из крупнейшего торгового площадке Украины. И решение следующих задач: провести анализ современных подходов к анализу и обработки данных; изучить основные алгоритмы для анализа и обработки данных; на основе применения новейших технологий создать программу, которая будет собирать данные, спроектировать архитектуру программы для более эффективного использования; очистить данные, применяя методы минимизации информационных потерь; проанализировать полученные очищенные данные применяя подходы к анализу и текстовых данных; сделать выводы по результатам всех вышеупомянутых работ. Существует достаточно большое количество разновидностей перечисленных задач, а также методов их решения. Это еще раз подтверждает значимость и актуальность выбранной нами темы. Цель исследования являются методы и средства с помощью которых можно минимизировать информационные потери при анализе и обработке текстовых данных. Объектом исследования является процесс минимизации информационных потерь при анализе и обработке текстовых данных. В ходе исследования проведен анализ последних исследований по анализу и обработки текстовой информации; проанализированы методы обработки текстовой информации и алгоритмы Data Mining.*

***Ключевые слова:** анализ текстовой информации; аннотирования; интеллектуальный анализ текстов; программный продукт; алгоритм; текстовые данные; естественный язык*

Olha O. Mezentseva, Candidate of Economic Sciences,
Assistant of the Technology Management Department
Research field: data science, text mining, data analytics

Anna S. Kolomiiets, Candidate of Economic Sciences,
Assistant of the Technology Management Department
Research field: Project management, strategic management, innovation management

ISSN 2663-0176 (Print)
ISSN 2663-7731 (Online)