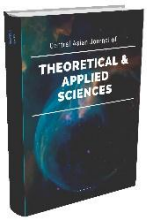

CENTRAL ASIAN JOURNAL OF THEORETICAL AND APPLIED SCIENCES

Volume: 04 Issue: 10 | Oct 2023 ISSN: 2660-5317
https://cajotas.centralasianstudies.org

Approach to Textual Data Analysis

O. Babomuradov
Executive director of the Kazan Federal University branch in Jizzakh, Jizzakh, Uzbekistan

O. Turakulov, Sh. Karaxanova
Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan

Received 25th Aug 2023, Accepted 26th Sep 2023, Online 27th Oct 2023

Abstract: *In this manuscript, approaches to processing textual data, based on which models and algorithms for classification and analysis of textual data are proposed. Developed algorithms serve to improve the efficiency of classification and analysis of textual data. A core algorithm for analyzing textual documents, a modification of a dictionary search algorithm, and algorithms A1, A2, and A3 for classification and analysis have been developed. The software developed on the basis of these algorithms is based on experimental research. A study sample of 2000 words was used in the experimental researches. The knowledge base is dynamic and expands during the training process.*

Introduction.

Currently, the amount of information belonging to different categories and types is increasing rapidly in the data ocean. Since the amount of data is very large, it becomes more difficult for the user to extract the information he/she needs from it. In order to search for the necessary information and extract it, humanity has to process the data, analyze it, or more precisely, extract the necessary pieces from the data. This statement of the problem shows that it is more appropriate to use intellectual analysis to identify the data structure, previously unknown relationships, and regularities between the data, rather than the traditional methods of data analysis, which are mainly focused on testing pre-existing hypotheses about the data [1-4].

Data collection varies according to the purpose of its use and the type of storage. Different types of data require different approaches. A homogeneous approach may produce different processing results in one species and another in another category. Especially nowadays, having very large volumes of data causes difficulties in their processing [5-7].

The Big Data phenomenon has a significant impact on data processing technology [8, 9]. The results of the research of the leading research institutions stated that in 2020, the world's data demand will exceed 40 Zettabytes (40 trln. GB) [10, 11].

Before the information is consumed in the form of knowledge (metadata), it is intended to be processed in the form of simple information, but the increase in the information flow requires the improvement of

recording and storage technologies [12]. Initial data are not always complete or given with errors that satisfy the mathematical apparatus, in such cases, the accuracy of the solution of the problem of the traditional mathematical apparatus, in particular, the methods of mathematical statistics, is observed to decrease significantly. Human consumption of processed data has led to a rapid increase in the volume and flow of information. A voluntary organization (commercial, manufacturing, medical, scientific, etc.) depends on the correct organization of calculations and records, covering the entire process of its activities. The question arises as to what to do with the created information array. Proper processing of data and information array simplifies its appearance and structure, making it easier to use. Mathematical statistics, which was considered the main tool of data analysis many times, lost its leadership role due to the complexity of the data structure. The main reason is the concept of approximation by choice, which leads to operations on spurious quantities (operations such as the average temperature of a patient in a hospital, the average height of a building). In such cases, it can be seen that the methods of mathematical statistics are useful for testing pre-specified hypotheses and for rough intelligence analysis based on rapid data analysis [13-15].

Intelligent Analysis of Data (IAD) technology combines rigorously formalized methods and unformalized analysis of methods. IAD's methods and algorithms are related to followings: artificial neural networks, decision tree, symbolic rules, basis vector method, Bayes networks, linear regression, correlation-regression analysis; hierarchical methods in cluster analysis, non-hierarchical methods in cluster analysis; The Aprior algorithm; bounded search method, evolutionary programming and genetic algorithms, various types of data visualization methods and other sets of methods [16-18].

The continuous growth of the data volume and information, limitation of primary data, complexity of their structure, uncertainty, and non-stationarity of parameters require the development of data analysis methods and algorithms. IAD develops technological indicators and complexes based on the conceptual principles of using new and previously unknown knowledge, their hidden properties and laws, their interdependence, and features of random and temporary processes that represent non-stationary objects in technical, economic, social, and monitoring systems [19, 20].

Data preprocessing serves to efficiently implement the problem of classification, which is one of the main problems of IAD. This problem requires determining whether the incoming object belongs to one of the specified classes (S_i) on the basis of the objects of the educational sample given by *Xi* symbols. Various models have been proposed for data classification. A decision tree forms a hierarchical model of training data. An effective path in the tree is used to classify each incoming object. It is reasonable to think of each path in the tree as a rule used to classify an incoming entity. Rule-based classifiers can be thought of as generalized decision trees where data need not be represented hierarchically. Therefore, multiple conflicting rules can be used to cover the same training or test case. Probabilistic classifiers assign probabilistic values to the features of the training sample. A simple Bayes rule or Boolean function is used to efficiently estimate probabilities. When using Support Vector Machine (SVM) and neural networks, the effectiveness of the objective functions is increased in different ways. In SVM, the maximum threshold principle is used, and for neural networks, the efficiency is increased by the least square error of the probability. A classifier based on a learning sample is a classifier that depends on the time of learning. A simple, uncomplicated form of sample-based learning is the nearest-neighbor classification algorithm. Many complex transformations can be made by applying different distance functions and center point-based models [21-27].

Theoretical structure of textual data processing. In general, it is appropriate to consider the issue of automatic or learning classification of natural language texts based on classification symbols through direct preliminary processing and classification of textual data processing. The problem of classification implies the creation of some form of meta-data, the emergence of knowledge by revealing the hidden laws

of the data. Text analysis is basically an intellectual analysis of data by extracting useful concepts from the text or determining whether they belong to a class based on various algorithms. The expansion of the text data segment can be seen in the rapid development of various relatively new areas, including text data on the web, social networks, e-mail, digital libraries, and communication sites. In these areas, the issues of generating metadata are effectively solved by means of intelligent data analysis. Over the time, various methods of data processing are being developed. The main reason for this can be seen in the fact that the array of data that needs to be processed incorporates the characteristics of different types and large mass. Information resources, which have a complicated structure, reduce the effectiveness of individual approaches. In this case, it would be appropriate to propose an approach to distinguishing the internal structure of the problem and using different methods. Classification of text documents with a unified approach is carried out in 3 stages:

In step 1, normalization of incoming data is carried out. In this, the steamer algorithm is used, exactly the Ripple down vules approach is used. The step is used to generate a table of keywords based on the rules. The resulting table forms the basis for the analysis of texts. Unstructured text is converted to structured view: $X^{`} \rightarrow X$. The classification (analysis) process works exactly with *X*.

In the 2nd step, the structured text *X* is analyzed in order to eliminate the destructured views according to the dictionary. The structure of organizing a "topical" search in the dictionary is not always available in the information that appears to be broken. In this case, it is desirable to use a modified approach of dictionary search, with the help of which the dictionary is automatically filled with broken words, the approach serves to increase the accuracy of the final results. Based on the completion of step 2, by finding such a word or some modification of it in the dictionary, the information processing is terminated and the text structure is classified as corrupted. Otherwise, text analysis is performed in Step 3 using machine learning-based methods.

The text classification issue on the basis of the three-stage combined method is based on the accuracy of the matching of the words determined based on the broken indicators depending on the condition of the issues or some probability of matching one or more indicators.

Algorithms based on the text processing model. Any information that does not have a fixed structure will need a preliminary processing. The main complication of text analysis is the large number of words in the analyzed text, not all of which may obey some natural language laws. Especially, problematic is the fact that the incoming data flow is not in a certain pattern, which increases the time resource of the algorithm and the resulting errors. For this, a preliminary processing step is carried out, that is, the process of making the incoming data stream look normal (Fig. 1).

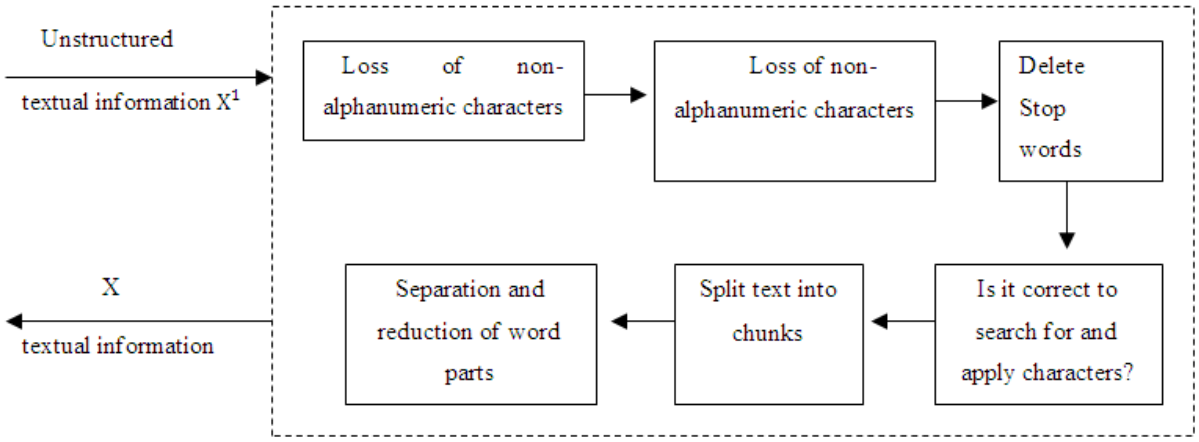

Figure 1. Steps of the coring algorithm.

The implementation of the coring algorithm performs a process tailored to the problem:

Step 1. Eliminating non-alphanumeric characters is done by deleting numbers, punctuation marks and other special characters. This allows us to create a vector or matrix that can be manipulated.

Step 2. Homogeneity means that all letters are represented in either upper or lower case (uppercase or lowercase). For example, texts in the form of "Text"; "TEXT"; "teXt" are normalized by being expressed in lower case: "text".

Step 3. Stop words are removed by removing auxiliary words that do not affect the content of the text. They can include the following parts of working words: particles, adverbs, adjectives, conjunctions, pronouns, etc. For this, a list of auxiliary words is formed. After that, cuts will be made based on this list.

Step 4. Character search and replacement replaces some letters with close case, for example the word "h" with "x" to reduce the time resource in words.

Step 5. It implies extracting texts at certain sizes. For example by character or word counting.

Step 6. Extraction of word parts is done by splitting the word suffixes present in the list and shortening by comparison.

Step 7. The result is presented in natural language form or graphically.

The following forms of coring are used in text preprocessing:

- search algorithm (full selection);
- reduction of suffixes (formation of the word core based on the rules);
- lemming (making words appear in the original dictionary);
- stochastic algorithms that determine the core of words;
- statistical algorithms such as N-gram or comparison

Normalization allows us to reduce the size of the space (characters). Because of this, it is important to reduce characters in text analysis to leave only words of significant value. Reducing the size increases the precision in the process.

In this section, we can convert $X^{`}$, which does not have a fixed structure, into a structured X representation. After the normalization step is performed, analysis is performed for the presence of indicators that do not have a structure:

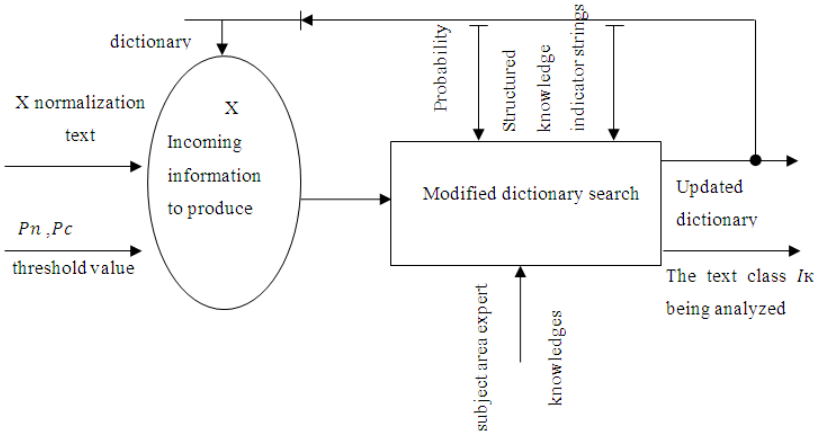

Figure 2. Search chart modified according to the dictionary.

Taking into account the above, a simple classification model can be implemented in the form of the following scheme (Fig. 3):

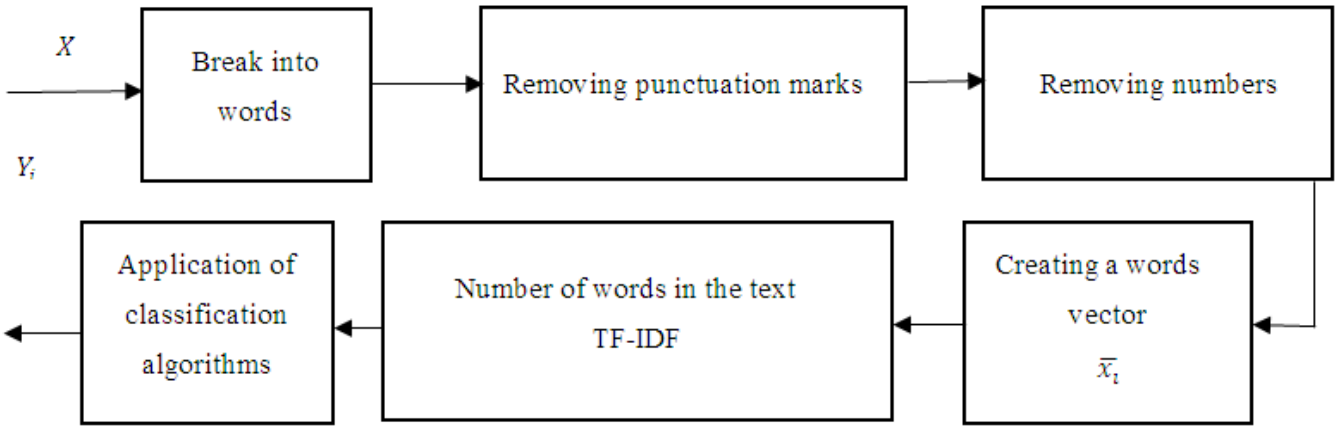

Figure 3. Implementation of the classification model.

Here, the incoming text data is parsed into words to form a vector of words, from which the punctuation marks are extracted and discarded. Various digital data are also discarded. A vector of words is generated and occurrences of words in the text are determined. After that, one of the classification algorithms (such as linear regression, Bayesian classifier, random forest, logic computing tools) is applied and the result is obtained.

The implementation of the considered process serves to implement the following issue: Let the piece of text is given and it is $\{X_1, \dots, X_i, \dots, X_n\}, i = \overline{1 \dots n}$, where X is the number of n-words in the text with the text vector highlighted. For each X_i, $X_i = \{X_1^i, \dots, X_q^i, \dots, X_m^i\}$, $q = \overline{1 \dots m}$, is an equation of letters in words. The vector of words in the dictionary is $\exists C = \{C_1, \dots, C_j, \dots, C_k\}$, $j = \overline{1 \dots x}$, where x is the number of words in the dictionary that have no structure. For each C_j, we have $C_j = \{C_1^j, \dots, C_w^j, \dots, C_t^j\}$, $w = \overline{1 \dots t}$, number of letters. In order to reduce the uncertainty of the results, the threshold value P_n- is introduced, the presence of unstructured information in the text. $P_n \in [0{,}5; 1]$; P_c- is an indicator of the loss of some abstract words, which are determined by the addition of word-formers from the words in the dictionary, and P_s∈ is defined in the interval [0.5; 0.75].

So, from unstructured text information, X text part is mapped into a vector, creating a dictionary vector. Thresholds P_n and P_c are set for the accuracy of the result.

Куйидаги A_1алгоритм таклиф этилади:

Step 1. Comparison of the amount of word symbols that do not have the structure X_i and C_j under analysis. X_i → C_j s.e.b. to increase the accuracy of $X_i \to C_j$ in m>t. we reduce X_i to the number of t characters in the word C_j s.e.b. in m<t

Step 2. We compare the Hamming distance $\bar{R}_j$ letters X_i with symbols of letters C_j in a row:

$$\bar{R}_j = \sum_{j=1}^{t} |C_w^j - X_a^i|, \qquad \overline{R_j^i} \epsilon [0, t]$$

Step 3. C_j and X_i words to determine the number of overlaps of letters:

$$R_j^i = \begin{cases} t - \overline{R_j^i},\ at\ m > t\ , \\ \text{in the remaining cases}, m - R_j^i \end{cases}$$

Step 4. P_j^i- calculating the indicator of the degree of relevance of the analyzed words to the unstructured:

$$P_j^i = \begin{cases} \dfrac{R_j^i}{m}, \text{if } m \leq 1 \\ \text{in the remaining cases}, \dfrac{R_j^i}{t} \end{cases}$$

Step 5. Determine the exponent $X \rightarrow I$ of R:

$$\mathrm{P} = max\left\{P_j^i\right\}$$

Step 6. Comparing the index *P* with the threshold value P_n and classifying the text *X*.

If $P_j^i \in P_n$, then $X \rightarrow I$.

Such a classification is carried out when $\mathrm{X} \cap \mathrm{C} \neq \emptyset$.

Step 7. Forming a set of words Δ t.e.b. After setting this threshold value, it is identified in the text that the analysis is: $X_i \epsilon \Delta | P_j^i \epsilon P_n$, $i = \overline{1 \ldots e}$, where *e* is found t.e.b. words. In some cases, for example, when $\mathrm{X} \cap \mathrm{C} \neq \emptyset$ is not sufficient, $P = \frac{e}{z}$, $z = \overline{1, n}$, where *z* are unique words in the text *X*.

Step 8. $X_i \rightarrow \mathrm{C}$ found t.e.b. words determine the compatibility of words and fill in the dictionary as necessary:

$$\begin{cases} X_i \epsilon\ C, \text{агар } P_j^i \epsilon P_c \\ X_i \notin C, \text{акс ҳолда} \end{cases}$$

If $P_j^i \in P_c$ then X_i is a new t.e.b word is added to the indicator dictionary of t.e.b..

Step 9. $X \rightarrow I_k$ set whether the text belongs to a class.

Step 10. If even at least 1 $X_i \rightarrow C$ is found, it is considered to belong to class $X \rightarrow I$. Otherwise, work continues in step 3, which is based on machine learning. We introduce a Bayesian classifier to implement machine learning. This means that the automatic determination of the presence of the indicator I_k in X based on the training sample is carried out using the a posteriori maximum estimate.

Let it is $\exists Q = \{Q^0, Q^1, \ldots, Q^k, \ldots, Q^K\}$, $k = 0 \ldots K$, Q^k- structured word dictionary, in the condition of $1 \leq k \leq K$, Q^k t.e.b dictionary of words. $Q^k = \left\{Q_{1_k}^k, \ldots, Q_{P_k}^k, \ldots, Q_{S_k}^k\right\}$, $P_k = 1 \ldots S_k$, where $S_k - I_k$ is the number of unique words in the dictionary Q^k with indicators.

During the training process in the Bayesian classifier, the set Q^k is appropriately filled with structured words and so on.

The frequency of occurrence of each occurring word in the text is determined based on the training sample $\forall X_i \rightarrow X_{iI_k}$ (variable representing the frequency of occurrence of the word $X_{iI_k} - X_i$ in text *X* of class *I*). The frequency X_{iI_k} of the Q^k dictionary is filled up by one unit of value, which is the frequency of this text encounter. Through this function, the Bayesian classifier becomes a self-educator.

The analyzed texts are counted: $T_{I_k} \_ I_k$ is a variable representing the number of texts belonging to class; *T* is the total amount of texts in the study sample. During training, the allocator calculates the relevance

$X \to I_k$ with probability $P(I_k)$. X_{iI_k}, T_{I_k}, T values are adjusted in each text analysis session. This improves as the classifier works. The creation of new benchmarks and the increase of the educational sample increase the probability of qualitative classification of the incoming object.

We perform this process based on the A2 (Bayes classifier) algorithm.

Algorithm A2:

Step 1. Determine the degree of relevance of words $X_i \to X_{iI_k}$ in $X_i \in Q^k$:

$$P(X_i/I_k) = X_{iI_k}/\sum X_{I_k},$$

where $\sum X_{I_k}$ $Q^k_{P_k}$ is the total amount of non-important words of the texts belonging to the class $I_k{}^k$.

Step 2. Applying Laplace smoothing on $X_i \notin Q^k$ given $(X_{iI_k} + 1)$:

$$P(X_i/I_k) = (X_{iI_k} + 1)/(\sum X_{I_k} + \sum S_k)$$

where $\sum S_k$ is the amount of important (unique) words in the educational sample in X_i.

Step 3. Determination of $X \to I_k$ relevance index of the text:

$$P(I_k) = T_{I_k}/T$$

Step 4. Calculation of final indicators

$$K^* = \arg\max_k \left[P(I_k) \prod\nolimits_s P(X_i|I_k)\right].$$

The relation $X \to I_k$ is obtained at the maximum value K^*.

Step 5. Updating the variables $X_{i_{I_k}}, T_{I_k}, T$ related to the training process and filling the dictionaries Q^k, which are the results of the Bayesian classifier, with unknown words.

Step 6. $X \to I_k$ classification into the appropriate class.

Depending on the level of complexity, it will be possible to apply a number of approaches to the classification of texts consisting of social network correspondence. A Neuro-fuzzy model with a cascade structure is proposed for text analysis based on combined and different modifications, on the basis of which it will be possible to group texts (Fig. 4).

A proposed cascaded neuro-fuzzy model

- it is used in the above approaches as a preprocessing algorithm. As a result, in text data with *X* input structure. The generated result is $X_k = \left\{X_1^{(k)}, \dots, X_n^{(k)}, \dots, X_N^{(k)}\right\}, k = \overline{1, K}.$
- classify text data into groups, applying a relevance function or a weight coefficient in two processes: first, each group is compared to the corresponding indices in the weight coefficient base, and each comparison is against G_j. Second, weight coefficients are accumulated and normalized;
- a set of neuro-fuzzy models of assessment of belonging to separate groups is designed to form the degree of belonging of texts to a separate G_j group;
- the grouping model serves for the resulting classification of the analyzed textual information.

The algorithm being built using this model will look like the following algorithm A3:

Step 1. $G = \{G_1, \dots, G_j, \dots, G_J\}$ in the form of the initial data collection is formed. where $G_j = \left\{\left(\omega_1^{(j)}, g_1^{(j)}, fg_1^{(j)}, gr_1^i\right), \dots, \left(\omega_{m_j}^{(t)}, g_{m_j}^{(j)}, fg_{m_j}^{(j)}, pg_{m_j}^{(i)}\right), \dots, \left(\omega_{M_j}^{(j)}, g_{M_j}^{(j)}, fg^{(j)}, pg_{M_j}^{(j)}\right)\right\}$,

The word m_j in the group $j = 1, \dots, J$ $\omega_{m_j}^{(j)} - G_j$ is relevance level of word m_j in group $m_j = 1, \dots, M_j, g_{m_j}^{(i)} \in [0,1] - j$, frequency of occurrence of word m_j in group $fg_{m_j}^{(j)} - j$ threshold value of use of word m_j in group $pg_{m_j}^{(j)}$-j .

Step 2. A collection of pre-separated groups of text is separated. This process rule can be defined as "applicable" or "not applicable", with a value of 1 or 0, respectively.

Step 3. For each X_j text $X_N^{(k)}$ word, a weight coefficient is determined, which indicates whether the word belongs to the j group. As a result, a training sample is formed, in our case a term pair.

Step 4. Weight coefficients are sought using the training sample. According to it, j – the importance is changed based on the level of relevance depending on the group belonging. Based on it, a dictionary of the G_j group is created

Step 5. Classification results are obtained using a neuro-fuzzy mechanism, based on expert evaluations of the performed grouping.

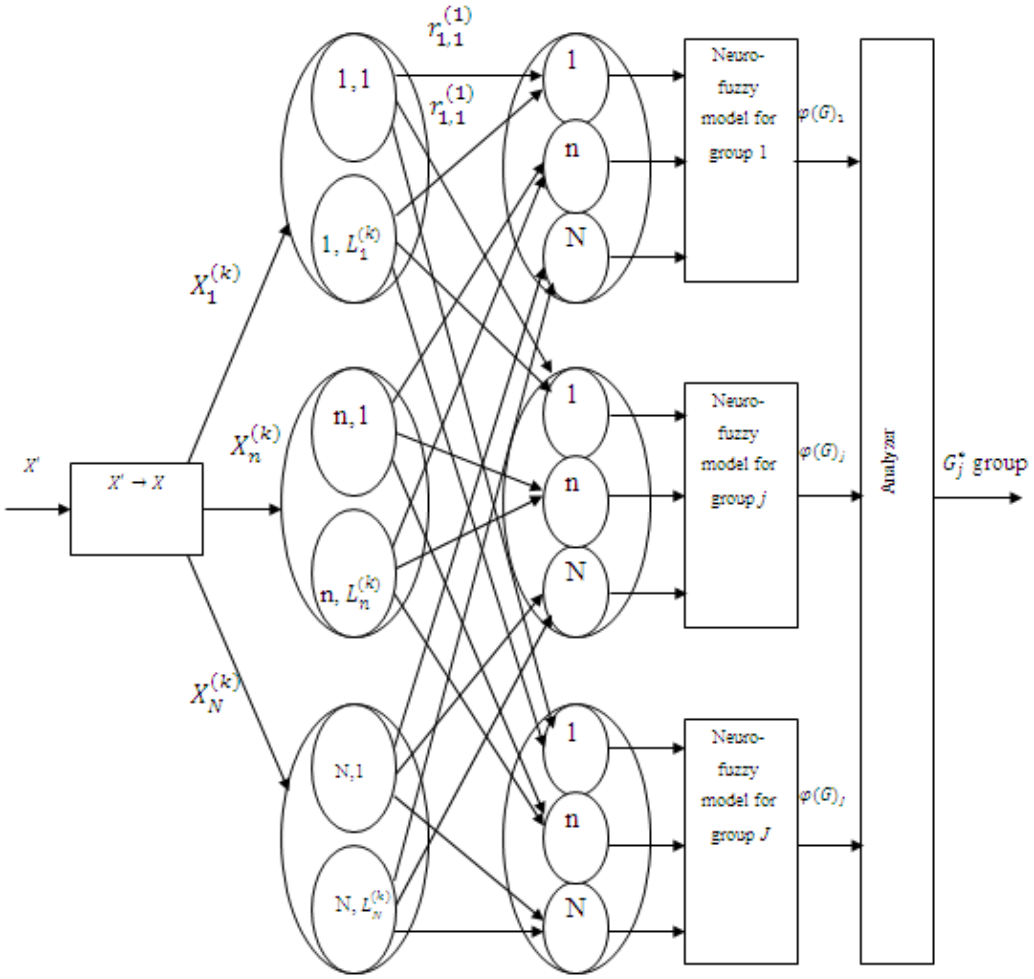

Figure 4. Grouping structure of a cascaded neuro-fuzzy classifier.

Results of experimental research. Based on the developed software tool, the incoming data was classified based on the study sample of 2000 words, and the analysis of the text data was carried out. As a comparison, the results obtained from several algorithms were studied.

Table 1. Accuracy of applying text analysis models

Linear vector based neural network model	0.8002
Logistic regression model (based on word sequences)	0.8547
Recurrent+convolutional neuro-fuzzy model	0.8653
A recurrent neuro-fuzzy model (based on a portfolio of words)	0.8782
Logistic regression (based on word portfolio)	0.8846
A recurrent neuro-fuzzy model	0.8868
Logistic regression model (based on n-gram model)	0.8868
Convolutional neuro-fuzzy model	0.8888
LSTM model	**0.8987**

At the same time, the evaluation of the location of the words according to the content was calculated, the evaluation of the analysis of the experimental data produced the following image in a five-point system:

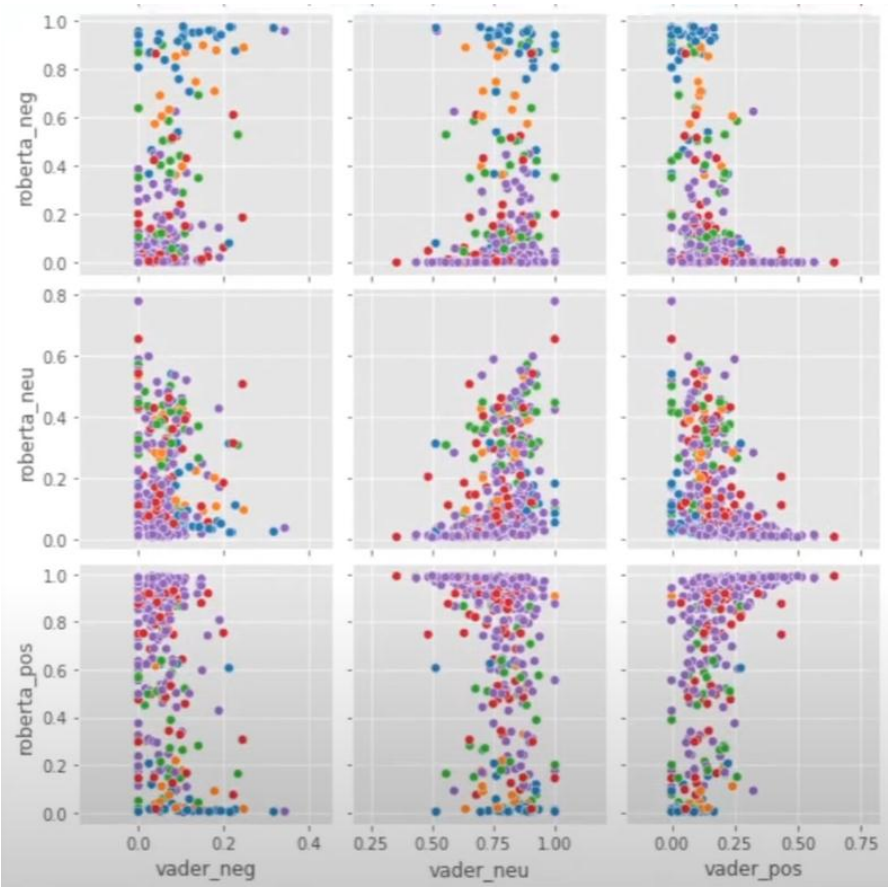

Figure 5. Placement of words by polarity

From this view, the color of words used in text correspondence can be extracted. This gives us an opportunity to form the emotional relevance of the textual information, the evaluation of correspondence (negative or positive).

More than 2,000 correspondences obtained as an experiment were obtained in this order, and the accuracy of the used A1-A3 algorithms was 88-89% based on different tools.

Conclusion. The conducted theoretical researches have shown that the formation of an adaptive mechanism in the models built for the analysis of textual data and the identification of emotional shades in them is of great importance. Depending on the conditions of the problem, the use of algorithms with a logical heuristic approach is appropriate. It can be seen in the high results obtained that the use of coordination of different approaches in the analysis of textual data (documents) in Uzbek language obtained for experimental research is effective.

References

1. Xin-She Yang Introduction to Algorithms for Data Mining and Machine Learning// Copyright © 2019 Elsevier Inc. All rights reserved. Academic Press, ISBN: 978-0-12-817216-2, 171p.
2. Hemlata Sahu, Shalini Shrma, Seema Gondhalakar A Brief Overview on Data Mining Survey, International Journal of Computer Technology and Electronics Engineering (IJCTEE), 2013, Volume 1, Issue 3; P. IndiraPriya, Dr. D.K. Ghosh A Survey on Different Clustering Algorithms in Data Mining Technique, International Journal of Modern Engineering Research (IJMER) www.ijmer.com Vol.3, Issue.1, Jan-Feb. 2013 pp-267-274.
3. M. A. Deshmukh, Prof. R. A. Gulhane Importance of Clustering in Data Mining, International Journal of Scientific & Engineering Research, Volume 7, Issue 2, February-2016
4. Jaro M. A. Advances in record linkage methodology as applied to the 1985 census of Tampa Florida // Journal of the American Statistical Association.1989. | 84 (406). | Pp. 414{420. | DOI: 10.1080/01621459. 989.10478785.
5. Rassel S. Iskusstvenniy intellekt. Sovremenniy podxod [Artificial intelligence. Modern approach] / S. Rassel, P. Norvig, 2-ye izd.: Per. s angl. – M.: Izdatelskiy dom «Vilyams», 2006. – 1408 s.
6. Feldman R. The text mining handbook: advanced approaches in analyzing unstructured data [Tekst] / R. Feldman, J. Sanger. – Cambridge University Press, 2007. – 410 p.
7. Moyotl-Hernandez E. An Analysis on Frequency of Terms for Text Categorization [Tekst] / E. Moyotl-Hernandez, H. Jimenez-Salazar // Procesamiento del lenguaje natural. – 2004. – Vol. 33. – P. 141-146.
8. Moyotl-Hernandez E. Some Tests in Text Categorization using Term Selection by DTP [Tekst] / E. Moyotl-Hernandez, H. Jimenez-Salazar // Proceedings of the Fifth Mexican International Conference on Computer Science ENC'04. – Colima. – 2004. – P. 161-167.
9. Bolshakova Ye., Lukashevich N., Nokel M. Izvlechenie odnoslovnix terminov iz tekstovix kolleksiy na osnove metodov mashinnogo obucheniya [Extracting single-word terms from text collections based on machine learning methods] // Informatsionnie texnologii. — 2013. — S. 31—37
10. Usama F., Smyth P., Piatetsky–Shapiro G. From Data Mining to Knowledge Discovery in Databases // Arti_cal intelligence Magazine. | 1996. |17(3). | Pp. 34-54.
11. Gmurman V. Ye. Teoriya veroyatnostey i matematicheskaya statistika [Theory of Probability and Mathematical Statistics]. — Moskva : Visshaya shkola, 2013. — 479 s.

12. Roussopoulos N. Conceptual Modeling: Past, Present and the Continuum of the Future // Conceptual Modeling: Foundations and Applications. 2009. | Pp. 139{152.
13. Hutchins J. ALPAC: The (In)Famous Report // Readings in machine translation. 2003. Vol. 14. P. 131–135.
14. Manning K. D., Ragxavan P., Shyutse X. Vvedenie v informatsionniy poisk [Introduction to Information Retrieval]. : Per. s angl. / Pod red. P. I. Braslavskogo, D. A. Klyushina, I. V. Segalovicha. M.: OOO «I.D. Vilyams», 2011. 528 s.
15. Lukashevich N. V. Tezaurusi v zadachax informatsionnogo poiska [Thesauruses in information retrieval tasks]. M.: Izd-vo Moskovskogo universiteta, 2011. 512 s.
16. Deliyanni A., Kowalski R. A. Logic and Semantic Networks // Communications of the ACM. 1979. Vol. 22, no. 3. P. 184–192.
17. Shapiro S. C. Encyclopedia of Artificial Intelligence. 2nd edition. New York, NY, USA: John Wiley & Sons, Inc., 1992. 1724 pp.
18. Gavrilova T. A., Xoroshevskiy V. F. Bazi znaniy intellektualnix sistem [Intelligent systems knowledge bases]. SPb: Piter, 2000. 384 s.
19. Apresyan Yu.D., BoguslovskiyI.M., IomdinL.L. i. dr. Lingvisticheskiy protsessor dlya slojnix informatsionnix sistem [Linguistic processor for complex information systems]. M.: Nauka 1992.-256s.
20. Osipov G.S. Metodi iskusstvennogo intellekta [Artificial Intelligence Methods].-FIZMATLIT, 2011.
21. Osipov G, Smirnov I., Tikhamirov I. Relation-situational method for text search and analysis and its applications// Seientific and Technical Information Processing. -2010.-vol.37, no b.-P.432-437.
22. O. J. Babomuradov, N. S. Mamatov, L. B. Boboev, B. I. Otaxonova, “Text documents classification in Uzbek language,” *International journal of recent technology and engineering*, vol. 8, no. 2, pp. 3787–3789, 2019.
23. Y. Du, J. Liu, W. Ke, and X. Gong, “Hierarchy construction and text classification based on the relaxation strategy and least information model,” *Expert Systems with Applications*, vol. 100, pp. 157–164, 2018
24. G. Vinodhini and R. M. Chandrasekaran, “A comparative performance evaluation of neural network based approach for sentiment classification of online reviews,” *Journal of King Saud University-Computer and Information Sciences*, vol. 28, no. 1, pp. 2–12, 2016.; A. Abbasi, H. Chen, and A. Salem, “Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums,” *ACM Transactions on Information Systems*, vol. 26, no. 3, p. 12, 2008.