The combined method of textual information analysis for the content of destructive indicators

V A Gostyunina¹,³ and N V Davidyuk²

¹ Institute of Information Technologies and Communications, Federal State Budgetary Educational Institution of Higher Education "Astrakhan State Technical University", 414056, Astrakhan, Tatishchev St., 16
² Institute of Information Technologies and Communications, Federal State Budgetary Educational Institution of Higher Education "Astrakhan State Technical University", 414056, Astrakhan, Tatishchev St., 16
³ E-mail: lerusya93@list.ru

Abstract. Modern society is characterized by increased progress in science and technology, accompanied by rapid uncontrolled growth and a variety of textual information, including of destructive orientation, in the information space of the Internet. In this regard, as well as in the context of the potential creation of a “closed” Internet space in the Russian Federation, solving the problem of analyzing text content by searching and automatically identifying destructive information is a priority trend. Within the framework of the paper task of identifying destructive information is reduced to the task of classifying the analyzed textual data based on the presence or absence of destructive indicators. The article describes the universal three-stage combined method, including blocks for normalizing input data, the modified dictionary search of destructive indicators and the Bayesian classifier, which allows staged search of various indicators in the text and automatically classifies it as a “dangerous”.

1. Introduction. Formulation of the problem

The high level of penetration into the daily life of information and communication technologies, the development and accessibility of their mobile options, the growth of activity of all segments of the population in the online environment, initiate new serious risks and threats to information security of society, especially of children and adolescents. One of the key elements of the information sphere is the Internet, where giant volumes of information, which cannot be taken into account and controlled, are created, stored and transmitted. The information space of the Internet network contains a mass of resources that carry information of various nature, including destructive orientation [1], most of which is presented in a dynamic text form.

Destructive information is means such one that affects to the consciousness and behavior of a mass audience destructively and negatively. Social networks and the Internet generally are a “convenient” environment for organizing destructive informational and psychological influence, including for the purpose of manipulating individuals, social groups and society as a whole [2], [3]. Informative indicators of destructive data are the presence of profanity, aggressive appeals, offensive statements, information that contradicts the norms of the law and is prohibited for distribution, etc.
Considering the existing rate of growth and updating of information, it is almost impossible and ineffective to fully control its content only with the help of experts (or content moderators — in practice of common social networks).

Therefore, not only the task of analyzing publicly available information is urgent, but there is also a need to develop approaches to automating the solution of this problem.

In the framework of the paper, this task, which consists in analyzing the input stream of textual information by searching and identifying destructive indicators in it, reduces to the task of classifying information products to the corresponding categories based on search results by the attribute of presence or absence of these indicators [4].

Thus, there is a need for a method that will reduce time-consuming costs, minimize subjectivity and the probability of errors due to the human factor when analyzing and assigning textual information to the particular class.

2. The study of data processing methods for the possibility of solving the problem

To solve the problem of classifying textual information, the following methods which are theoretically applicable for finding various destructive indicators were analyzed [5]:

- methods based on expert information processing;
- dictionary search;
- machine learning methods.

Note that the systematization of the desired informative indicators and their conversion into a form suitable for use in automated text analysis procedures is a separate important task and is considered in other works [5].

The result of applying methods based on expert information processing is the creation of «black lists» and registries. The quality of this kind of assessment depends entirely on the experience of the expert.

Dictionary search, which consists in character-by-character enumeration of words from a dictionary and its indication in the text, is one of the most common types of destructive information detection. However, to solve the classification task, the use of thematic search in a “pure” form is inexpedient, because there is a problem of the impracticability of forming a dictionary that contains keywords indicating the presence of prohibited content in the text, since they can have ambiguous meaning [6].

Machine learning allows to implement not a direct solution to a problem, but training in the process of applying solutions to many similar problems [7]. The classifier correlates with a specific text and returns the corresponding class containing destructive or non-destructive information. Set of texts which will be used as a training sample is required to conduct information analysis using machine learning methods.

To date, a large number of machine learning methods and their variations in the classification of textual information have been developed. Among the most common, the following methods for classifier training can be distinguished [8]:

- Bayesian classifier;
- support vector method (SVM);
- neural networks;
- Rocchio method;
- k-nearest neighbor method (k-NN);
- decision trees;
- random forest.
According to the results of some studies [4], [5], [6], [7], [8], [9], the Bayes method is promising from the point of view of the task in question.

The Bayesian method is a simple classification mechanism that is based on the probability of fulfilling the conditions and shows a high speed of operation and a fairly high quality of classification [9].

According to the results of the analysis, it can be concluded that the heterogeneous nature of the texts and the use of various identification bases lead to the need to develop some kind of combined method that allows to level out the limitations of the listed methods, at the same time, possesses their advantages.

It should also be borne in mind that the method to be developed should involve a phased search of various destructive indicators in order to reduce the probability of errors when sorting a text to the class of destructive ones.

3. **The combined three-stage method of textual information analysis**

The essence of the text information classification procedure based on the detection of destructive indicators is reduced to the using of the three-stage combined method, the process of which is divided into three modules (figure 1).

![Figure 1. Stages of the combined text analysis method.](image)

This method allows to analyze the input text information stream according to one or more destructive indicators, depending on the task, using blocks 2 and 3.

In this case, the first stage, which consists the normalization of the input data, is a prerequisite for the preliminary processing of the text.

When loading the investigated content, the checked text is presented in the form of unstructured information. Due to the fact that not all words can carry the desired semantics, this problem is eliminated by means of preliminary processing of the input data. The following types of stemming algorithms are used in text preprocessing [10]:

- search algorithms (exhaustive search of the word basis in the list);
- truncation of endings (there is a small list of rules by which the foundations of the word are found);
- lemmatization (reduction of a word to a canonical or initial dictionary form - a lemma);
- stochastic algorithms (probabilistic determination of the root form of a word);
- static algorithms (analysis of N-grams, matching algorithms).

The second block is intended for the initial stage of the analysis of the text on the content of destructive indicators in the dictionary. Most often, character-by-character exhaustive search is used for these purposes. Dictionary search, which consists in searching by the system for an exact match of a word from the dictionary and its indication in the text, is one of the most common types of destructive information detection. To solve the task of classifying texts based on the detection of destructive indicators, the use of thematic search in a “pure” form is impractical due to the following reasons:
• the need for constant replenishment of the dictionary due to the emergence of new forms of word formation;
• the high probability of missing a destructive word on account of the using in an investigated text of its word form which is unknown to the system.

In this regard, the application of the modified dictionary search is proposed.

To implement the third stage of the analysis of textual information, it is necessary to use the machine learning method. However, let’s note that if at least one word or its modification from the dictionary is found (with a certain probability), the system will notify the user of the presence of dangerous content and stop the further process, otherwise the procedure will be continued.

Thus, the proposed three-stage combined method makes it possible to classify texts depending on the task both by the exact coincidence of the words found with the given destructive indicators, and by the certain probability of the coincidence of one or more indicators in the aggregate. Let’s consider the second and third stages more detail.

3.1. The modified dictionary search
After passing through the normalization stage, the text is ready for direct analysis on the content of destructive indicators. The proposed modified search method is based on calculating the ratio of the number of characters of the analyzed word corresponding to the characters of the word from the dictionary to the total number of characters in the word and assigning the given text to destructive / non-destructive.

Input data for the analysis unit:
• the dictionary whose elements are destructive indicators;
• normalized text to be analyzed.

The output of the second stage is the correlation of the analyzed text content to the classes of destructive D / non-destructive nD.

For this, it is necessary to determine the probability that the word in question belongs to the class of destructive ones.

Let’s describe the probability calculation mechanism.

Consider the part of the text $X_i$, consisting of the certain set of words $\{X_{i1}, \ldots, X_{iq}, \ldots, X_{im}\}$, $q=1 \ldots m$, where $X_i$ is the vector representation of the text; $n$ is the number of words in the text. Due to the fact that the proposed algorithm uses character-by-character comparison, it is necessary to process each character of the string separately. Accordingly, we introduce a notation for each word of the text $X_i$, which has a finite number of characters:

$$\overline{X}_i = \{X_{i1}, \ldots, X_{iq}, \ldots, X_{im}\}, \ q=1 \ldots m,$$

where $m$ is the number of letters in the word.

Then the vector representation of the words from the dictionary:

$$\overline{C} = \{\overline{C}_1, \ldots, \overline{C}_j, \ldots, \overline{C}_k\}, \ j=1 \ldots k,$$

where $k$ is the number of words in the dictionary

The symbol vector will have the following form for each word from the dictionary $\overline{C}_j$:

$$\overline{C}_j = \{c_{j1}, \ldots, c_{jw}, \ldots, c_{jt}\}, \ w=1 \ldots t,$$

where $t$ is the number of letters in the dictionary word.
To increase the accuracy of assigning the word to the destructive one, during symbolically comparing of two words it is necessary to prevent the length of the analyzed word from exceeding the length of the dictionary one. In this case, excess characters are subject to clipping.

Then the desired probability is the ratio of the number of characters of the analyzed word $\vec{X}_i$, corresponding to the characters of the word from the dictionary $\vec{C}_j$ to the number of letters in the word:

$$P'_v = \frac{\sum_{j=1}^{m} X_i \cap C_j}{m} = \begin{cases} j=1 \ldots m, & \text{if } m<t \\ j=1 \ldots t, & \text{otherwise} \end{cases}, \quad i=1 \ldots n,$$

(1)

where $v$ is the index of the calculated sum of every two words; $m$ is the number of letters in the word; $t$ is the number of letters in the dictionary word; $n$ is the number of words in the text.

We take as the cut-off point a certain threshold value $P_n$ based on a comparison of the calculated probability $P'_v$, with which the system will establish the presence or absence of destructive information in the text. The critical value for this indicator, established empirically, is proposed to be accepted in the range [0.5; 1].

Then the probability of assigning the text $\vec{X}_o$ the class of destructive $D$ or non-destructive $nD$:

$$P = \max_i P'_v = \begin{cases} \vec{X} \in D, & \text{if } P \in P_n \\ \vec{X} \in nD, & \text{otherwise} \end{cases}$$

(2)

Thus, the system identifies the text as destructive or non-destructive having made a comparison between the found probability $P$ and the threshold value $P_n$.

The integral step is also the process of replenishing the dictionary. The word found in the text, belonging to the class of destructive, will be new if the value of its probability of assignment to this class will satisfy the threshold value $P_c$. This value, identified empirically, is necessary to cut off those words that differ from existing dictionary words only by case ending or plural. For the threshold value, we take the range "[0.5; 0.75)". Thus, the dictionary will be updated only with new words under the following conditions:

$$\begin{cases} \vec{X}_i \in \vec{C}, & \text{if } P'_v \in P_c \\ \vec{X}_i \notin \vec{C}, & \text{otherwise} \end{cases}$$

(3)

Based on calculations of the probabilities of classifying words as destructive with $P'_v \in P_c$ the system will automatically replenish the dictionary of destructive indicators. In order to optimize this process, it is advisable to periodically check the correctness of the dictionary update by an expert by considering only newly added words.

3.2. Application of the Bayesian classifier to the task solving

The purpose of the Bayesian classifier is to automatically determine the presence or absence of destructive indicators in the text based on the training sample [9]. Therefore, this method reduces to determining the most probable class (to which the analyzed text belongs) when using the estimate of the posterior maximum. To do this, it is necessary to calculate the probability for all classes and select the class for which we will have the maximum probability value.

According to Bayes' theorem, the probability that an information resource $x$ belongs to the class $K_j$ is calculated using formula:

$$P(K_j|x) = \frac{P(x|K_j)P(K_j)}{P(x)},$$

(4)

Based on calculations of the probabilities of classifying words as destructive with $P'_v \in P_c$ the system will automatically replenish the dictionary of destructive indicators. In order to optimize this process, it is advisable to periodically check the correctness of the dictionary update by an expert by considering only newly added words.

3.2. Application of the Bayesian classifier to the task solving

The purpose of the Bayesian classifier is to automatically determine the presence or absence of destructive indicators in the text based on the training sample [9]. Therefore, this method reduces to determining the most probable class (to which the analyzed text belongs) when using the estimate of the posterior maximum. To do this, it is necessary to calculate the probability for all classes and select the class for which we will have the maximum probability value.

According to Bayes' theorem, the probability that an information resource $x$ belongs to the class $K_j$ is calculated using formula:

$$P(K_j|x) = \frac{P(x|K_j)P(K_j)}{P(x)},$$

(4)
where $P(K_j|x)$ is the probability that the information object $x$ belongs to the class $K_j$; $P(x|K_j)$ is a likelihood, that is the probability of a given meaning of a word in a given class; $P(x)$ is the a priori probability of the appearance of a word in the text; $P(K_j)$ is the a priori probability of a given class $K_j$.

For the Bayesian classifier, the text is the set of words with conditionally independent of each other probabilities. It is believed that the position of a word in a sentence is not important.

Analyzing formula (4), we can neglect the quantity $P(x)$, since it is constant and does not affect the final result. Then, to calculate the probability that the text $x$ belongs to the class $K_j$, we need to know two quantities: $P(x|K_j)$ и $P(K_j)$. It’s useful to note that the class probabilities and word probabilities in the class are estimated based on the training sample.

The calculation of estimates of the probability of occurrence of a word can be represented by the formula:

$$P(x|K_j) = \frac{W_{ik}}{\sum W_k}$$  \hspace{1cm} (5)

where $W_{ik}$ – the number of times the $i$-th word occurs in the texts of the class $K_j$; $\sum W_k$ is the total number of words in the texts of class $K_j$.

Due to the fact that during the work of the Bayesian method the word may occur that did not fall into the list of destructive indicators, Laplace smoothing, which involves the artificial addition of a unit to the frequency of occurrence of each word, is used:

$$P(x|K_j) = \frac{W_{ik} + 1}{\sum W_k + V}$$  \hspace{1cm} (6)

where $V$ – the number of unique words in the training set.

The probability of assignment to the class $K_j$ can be calculated by the formula (7), using the values $D_k$ – the number of texts in the class $K_j$ and $D$ – the total number of texts in the training sample:

$$P(K_j) = \frac{D_k}{D}$$  \hspace{1cm} (7)

The sequence of actions for solving the problem by the Bayesian classifier is given below:

- highlighting each word in the text $W_i$ and determining the probability of classifying the word as a destructive class $K_1 P(W_i|K_1)$;
- likewise, the probability of assigning a word to the non-destructive class $K_2 P(W_i|K_2)$ is determined;
- these probabilities are estimated by the frequency of occurrence of the word $W_i$ in destructive / non-destructive classes, obtained as a result of the Bayes method on the training sample;
- calculating the final probability according to the following formula:

$$K^* = \arg \max_j \left[ P(K_j) \prod_{i=1}^{n} P(W_i|K_j) \right], i = 1 \ldots n, j = 1, 2$$  \hspace{1cm} (8)

where $n$ is the number of words in the texts of class $K_j$;

$\ j$ is the class identifier, in particular, 1 is a destructive class, 2 is a non-destructive class.

Thus, the analyzed text will be assigned to that class, the probability of correlation with which will be maximum.

4. Conclusion
To analyze textual information regarding the content of destructive indicators, it is proposed to use the combined three-stage method that sequentially uses the steps of normalizing input data, the modified dictionary search, and Bayesian classifier. The combined use of different approaches to identifying
destructive indicators significantly increases the efficiency of the analysis, reducing the probability of potential errors and increasing the accuracy of classifying the analyzed text content as non-destructive / destructive. The feature of the proposed procedure is the ability to automatically obtain new knowledge for subsequent use in the analysis process.

The developed method is the universal tool for analyzing textual information, lends itself to algorithmization and can be used to identify various informative indicators during the classification.

References
[1] Smirnov A 2015 Negative content: identification problems in context of legal regulation Information Law 2 18-25
[2] Talimonchik V 2008 Convention on Cybercrime and the unification of legislation Information Law 2 27-30
[3] Etzioni O 1996 The world-wide web: Quagmire or gold mine Communications of the ACM 65-69
[4] Davidyuk N, Gostyunina V and Baidulova D 2019 An intelligent algorithm for identifying destructive information in the text Bulletin of the Astrakhan State Technical University. Series: Control, Computing and Informatics 2 29-39
[5] Batura T 2017 Methods for automatic classification of texts Software products and systems 1 85-99
[6] Baidulova D, Gostyunina V and Davidyuk N 2019 Identification of destructive information using the Bayesian intellectual method International scientific and practical forum "The Caspian in the era of the digital economy" (Astrakhan: ASU) pp 151-4
[7] Gostyunina V and Savelyev A 2017 Use of Text Mining technology to classify Web resources by age categories Mathematical methods in engineering and technology 3 111-5
[8] Mangalova E and Agafonov E 2013 On the problem of highlighting informative features in the task of classifying text documents Bulletin of Tomsk State University. Series: Control, computer engineering and informatics 1(22) 96–103
[9] Tang B 2016 Toward optimal feature selection in naive Bayes for text categorization IEEE Transactions on Knowledge and Data Engineering 28(9) 2508-21
[10] Lazareva O and Bolomutova M 2016 Methods for highlighting keywords in the context of electronic learning systems Young scientist 26 143-6