Methodology of Mediametric Information Analysis with the Use of Machine Learning Algorithms

I. E. Gorbachev^a, A. A. Kriulin^a, *, and I. T. Latypov^a, **

^a MIREA—Russian Technological University, Moscow, 119454 Russia

*e-mail: kriulin@mirea.ru

**e-mail: lit@inbox.ru

Received August 20, 2020; revised August 20, 2020; accepted September 20, 2020

Abstract—The qualitative analysis of information messages and the assessment of publications on the Internet are becoming more urgent than ever. A large number of materials are published on the Internet on various events in the world; the nature of these publications can affect the political and social life of society. In order to ensure the safety of the population of the Russian Federation and meet the requirements of regulatory documents, a methodology for mediametric information analysis with the use of machine learning algorithms is proposed. Based on the results of research in this area, the main approaches to mediametric information analysis are determined. An approach is proposed for determining the sentiment of publications using the Word2Vec model and machine learning algorithms for natural language processing. A methodology is formulated that takes into account the technical features of a publication source and the existing methods of mediametric information analysis. On the basis of real information publications, the results of the implementation of the methodology of mediametric analysis and determination of the sentiment of messages are presented.

Keywords: mediametric analysis of publications, word cloud, sentiment of information messages, data mining, frequency analysis of words, machine learning libraries for the Python language

DOI: 10.3103/S0146411621080125

INTRODUCTION

In the context of the expanding spectrum of forms of confrontation in the political, economic, informational, diplomatic and other spheres, traditional views on armed military and local conflicts are changing. As noted in the Doctrine of Information Security of the Russian Federation [1], approved by the decree of the President of the Russian Federation of December 5, 2016, “The scale of the use of information and psychological influence by special services of individual states, aimed at destabilizing the internal political and social situation, is expanding... The informational influence on the population of Russia is increasing...”.

In order to counter information aggression against the Russian Federation, the mediametric assessment of information messages in communications of all scales is of particular importance. The relevance of mediametric research of information in the global information network (GIN) Internet is due to the contradiction between the need for an adequate assessment of publications for making informed decisions and the imperfection of methods for quantitative assessment of various qualitative properties of an information message and its source.

The results of the analysis of the works of scientists of the Russian Institute for Strategic Studies [2], Russian media researchers [3, 4] showed that, at present, a technical groundwork has been created for the possibility of quantitatively assessing information publications, economic aspects in the work of the media have been considered, a number of assessed indicators of political coloration of news messages has been proposed, and proposals for their application have been developed. However, it should be noted that there is no systematic approach to mediametrics of information to ensure the security of the Russian Federation. Assessment is carried out on separate indicators, which allows analyzing information messages only from one side, outside the framework of a systematic approach. The complex mediametric criterion is not defined and/or justified. Single-type approaches are used to assess information from various communication environments without taking into account the technical features of their functioning.

The scientific novelty of the research results is determined by a new approach to mediametric assessment of information messages using modern methods of data mining and a developed methodology that takes into account the existing methods of mediametric analysis.

The practical significance of the research lies in the development of a comprehensive methodology for the quantitative assessment of information publications in the GIN Internet, which can be used to provide psychological protection from information influences on the population of Russia.

1. ANALYSIS OF THE POSSIBILITY OF APPLICATION OF EXISTING MEDIAMETRIC METHODS WHEN ASSESING INFORMATION IN THE GIN INTERNET

The historical definition of mediametry was presumably formed in the middle of the 20th century and by the beginning of the 1990s was reduced to practical activities for the study and collection of quantitative and qualitative data of the media channel and audience in order to study the effectiveness of the media [5].

With the current development of modern communication media, the concept of mediametry has expanded significantly and is at the intersection of disciplines such as political science, sociology, economics, history, and applied mathematics. For each of these disciplines, the concept of mediametry reflects its own specifics within the framework of solving specific applied problems.

As the author of the study [3] points out, in the field of media consumption and mediametry, there is no uniform terminology and standards that take into account the dissemination of information using innovative technologies and services. There are also no quantitative indicators of media consumption, official data on the total volume and time spent on individual media resources, information on the distribution of the audience across different platforms, devices, and consumption methods.

In this regard, the analysis of the application of existing mediametric methods is carried out based on the characteristics of various applied disciplines.

Modern media economics, as one of the varieties of mediametric knowledge of information processes, distinguishes a number of special indicators [4].

(1) Coverage of a unique audience—quantitative characteristics of a particular information carrier relative to the number of contacts with recipients that it can provide. For printed publications, this is the circulation and the number of readers for each copy. For radio or TV channels, this is the number of viewers or listeners in a given period of time.

(2) Technical coverage of the audience—quantitative indicators of the technical capabilities of information media in relation to different regions.

The study of the audience is also accompanied by demographic research (age composition of the audience, education, income level, etc.) and psychographic research (lifestyle, values, opinions, etc.).

(3) Relevance to the target group—an indicator that takes into account the percentage of client audience for a particular communication channel of the total audience size from the coverage indicator.

(4) Rating—the ratio of the number of people watching, listening or reading certain information content to the total number of the audience, which has the appropriate technical capabilities.

(5) Cumulative rating number—the sum of all presentations of a particular content, expressed as a percentage of the audience. Also, in [4], more specific media economic indicators are considered in detail.

It should be noted that the above indicators have a number of disadvantages due to the technical features of the implementation of specific information sources. Counters of Internet resources do not record real people, but various device identifiers (IP-address, browser, etc.) of the GIN Internet users. Accordingly, the quantitative audience will grow when one user has several devices, which will distort the real assessment.

Within the framework of political mediametry, two main characteristics are used [5]: the event coverage intensity and the aggressiveness of the event being covered. The coverage intensity indicator is defined as the average number of significant materials that appeared in the mass media of a particular country during a calendar month. The degree of aggressiveness of an event is determined using the aggressiveness index, which is calculated as the quotient of dividing the number of negative materials by the number of neutral publications. With the help of indicators of political mediametry, it is possible to determine some characteristics of an information attack, for example, the moment of its beginning, intensity, and degree of pressure.

From the analysis of the possibility of using existing mediametric methods, it follows that, firstly, there is no official regulatory framework for quantitative and qualitative assessment of information messages, which significantly complicates research in the field of mediametric analysis. Secondly, the economic methods of mediametrics are quite specific and are used to assess the profitability of the media. Thirdly,

the indicators of political mediometry make it possible to determine some quantitative indicators of information messages, but they are quite subjective and are mainly determined by an expert assessment of the sentiment of a publication.

2. APPROACH TO QUANTITATIVE ASSESSMENT OF INFORMATION IN THE GIN INTERNET BASED ON INTELLECTUAL DATA ANALYSIS

Information on the Internet and in the media is big data, and, in order to be able to assess it qualitatively and quantitatively, it is necessary to use data science as the core for modern information and analytical systems.

As part of the approach to the quantitative assessment of information, it is proposed to systematize the most popular methods of data analysis for their further application to specific events.

(1) Determination of the publication sentiment: positive, negative, or neutral. Solving this problem will allow assessing the opinion of the media regarding various events, the reactions of famous people to the event, and so on. The purpose of sentiment analysis using machine learning is to determine the properties of text information and find opinions and emotions in the text. Depending on the nature of the tasks being solved, sentiment analysis comes down to determining the topic of publication and the author's opinion and position. The use of machine learning methods is due to a large number, which show its effective use in information security problems [6–9].

To determine the sentiment using machine learning algorithms, the text of publications is represented as number vectors. There are two main ways to do this. The first method, bag-of-words, involves compiling one large dictionary of words that occur in the text and assigning an index to each word. The original words in the text are replaced with the corresponding indices, and, using metrics (Euclidean distance, Chebyshev distance, etc.), the vectors are compared. The second way is to use the word2vec model, the task of which is to maximize the values of vectors of words that are nearby and minimize the values of vectors of words that do not appear with each other. Formally, such a model can be represented as follows:

$$M = \frac{(w_v \times w_c)}{\sum (w_v \times w_{cl})}, \quad (1)$$

where w_v is the target word, w_c is the context word, and w_{cl} is all the context words in the document.

The resulting numeric word vectors are a sample of data for machine learning algorithms. The use of machine learning algorithms allows determining the sentiment of a publication, that is, classifying it into one of three groups: positive, negative, or neutral. To solve the classification problem, it is necessary to form a training sample of publications, which should equally contain positive, negative, and neutral texts. In the absence of a trained sample, it is advisable to start the analysis of publications with the clustering problem according to the same principles as for the classification problem.

(2) Use of oriented graphs in the analysis of social networks [10]. This method is based on the determination of the centrality index, which characterizes the most important vertices of the graph. Applications using this metric can be used to identify influencers on a social network. Google has also developed an influencer-based PageRank algorithm for the link ranking of web pages.

(3) Use of a word cloud to visualize words in a publication [10]. This method belongs to natural language processing (NLP) techniques and boils down to placing words on the graph in accordance with the sizes proportional to their frequencies. Most of the problems solved by the NLP can be fully used for information mediometric analysis when making decisions on specific events.

The disadvantages of this approach include the need for a high level of knowledge in the subject area of the assessed publications to create high-quality (representative) samples when training algorithms. Also, with a significant increase in the volume of processed texts, the requirements for the computing power of machines will proportionally increase.

3. METHODOLOGY FOR ASSESSING INFORMATION PUBLICATIONS IN THE GIN INTERNET

Taking into account the results of the analysis of the application of mediometry methods, as well as the approach to the quantitative assessment of information using machine learning algorithms, a methodology has been developed for assessing text publications. The general scheme of the methodology is shown in Fig. 1.

1	Determining the sources of publications
2	Defining time ranges
3	Calculating economic indicators of the source of publications
4	Preprocessing the entire array of text data
5	Calculating statistical indicators
6	Visualizing statistical features and characteristics
7	Constructing graphs of links and dependencies
8	Determining the sentiment of the source and publications

Fig. 1. General outline of the publication assessment methodology.

The methodology consists of eight main stages. At the first stage, all publication methods are determined for each source. It is worth noting that publications on a topic from one source may differ structurally, technically, and substantively, depending on the information platform (microblog, website, page on a social network, etc.); therefore, the methods of posting publications are grouped by web content. This step is also required to prepare automated content requests. Not every website provides or has a programming interface, so requests can be made through standard protocols. Also, some websites may have RSS feeds for describing news feeds. Major social media and microblogging services provide application programming interfaces (APIs) for accessing content. After identifying all the publication methods, the next step is to determine the time range of the event of interest. Using the time range, one can track the dynamics of publications, calculate intensity intervals, build a time series of an event, and analyze it.

To perform the stages of the methodology, the following information messages were taken:

- using the Twitter API for the Python programming language, a sample of US President Donald Trump's tweets from 2010 to March 2020 was received. The sample contains 41122 unique records;
- from the platform for research, data processing, and machine learning Kaggle, a sample of open data was received on the topic of the coronavirus pandemic from January 2020 to March 2020. The sample contains 48421 unique records [11];
- from the service for developers GitHub [12], a sample of H. Clinton's mail messages was received during her function as US Secretary of State. The sample contains 7945 unique records.

At the third stage, the calculation of the main media economic indicators for the source of publications is performed. Formalized methods for calculating media economic indicators are presented in [4]. Media economic indicators may reflect the desire of a source to increase its level of profitability by attracting an additional audience to high-profile events.

Text data received from requests to web pages, as a rule, come in hypertext format, which significantly reduces their quality. For this, at the fourth stage, preprocessing of text publications is performed, which includes:

- removal of html tags, punctuation, symbols, and other service data. The most popular automated software implementation for hypertext processing is the Beautiful Soup HTML and XML parsing library;
- removal of common stop words such as prepositions, suffixes, participles, and the like. Stop word removal is accomplished using the Python Natural Language Toolkit (NLTK) library for symbolic and aggregate natural language processing.

At the fifth stage, the main statistical characteristics of samples with text data of publications are calculated: average, maximum, and minimum number of words, standard deviation, and other indicators.

At the sixth stage, graphical visualization of statistical indicators, dependencies, and characteristics of text data is carried out. For plotting two-dimensional or three-dimensional graphs, a library in the Python

Fig. 2. Chart comparing the dynamics of coronavirus in different cases.

programming language, Matplotlib, is used. An example of using visualization tools from the Matplotlib package to sample open data on the coronavirus pandemic is shown in Fig. 2.

An example of constructing a word cloud for the tweets of US President D. Trump is shown in Fig. 3. If the sample of publications is long in time, it is advisable to divide the word clouds into several intervals.

Figure 3 shows that the words in D. Trump's publications at different periods of his political career differ significantly.

At the seventh stage, the construction of network graphs for publications is carried out, which will allow visualizing the process of disseminating information from various sources. To construct network graphs, it is most efficient to use special programs such as Gephi.

The final stage of the methodology is to determine the sentiment of text publications. Using the statistical indicator TF-IDF [10], which is used to assess the importance of a word in the context of a document, a diagram of the most frequently used words of D. Trump and H. Clinton has been constructed (Fig. 4). TF (2) word frequency is the ratio of the number of occurrences of a word to the total number of words in one message:

$$TF(t, d) = \frac{n_t}{\sum_k n_k}, \quad (2)$$

where t is the word in document d , n_t is the number of occurrences of a word t , and n_k is the number of all words in the document.

The inverse frequency of an IDF document (3) is the inverse of the frequency with which a word occurs in all messages:

$$IDF(t, D) = \log \frac{|D|}{|\{d_i \in D \mid t \in d_i\}|}, \quad (3)$$

where $|D|$ is the total number of messages and $|\{d_i \in D \mid t \in d_i\}|$ is the number of messages in which t occurs.

The result of applying the TF-IDF word frequency index to samples with D. Trump's tweets and H. Clinton's mail messages is illustrated in Fig. 4. Based on the statements of D. Trump and H. Clinton shown in the figure, their priorities in a certain period of time can be assessed. As seen, they can differ significantly.

At this stage, we also used the word2vec model based on distribution semantics. The input of the model was text data from a sample of D. Trump's tweets; at the output, the model returned a weighted vector for each word.

Using the random forest algorithm, all publications from D. Trump's twitter were classified into three categories: positive, negative, and neutral. Figure 5 shows the result of classification by the random forest algorithm of D. Trump's publications on Twitter by sentiment using the word2vec model.

[illegible][illegible]

H. Clinton

Word	tf-idf
Information	0.00095
Justice	0.00075
Proposed	0.00075
Everyone	0.00065
National	0.00065
Part	0.00065
Both	0.00055
Criminal	0.00055
Future	0.00055
Hope	0.00055
Incomes	0.00055
Kinds	0.00055
Plans	0.00055

D. Trump

Word	tf-idf
Clinton	0.00185
Leaving	0.00120
Agree	0.00115
Wrong	0.00105
i'll	0.00100
she's	0.00100
Tremendous	0.00095
Politicians	0.00085
Chicago	0.00065
Company	0.00065
Hillary	0.00065
Mean	0.00065
Sean	0.00065

AUTOMATIC CONTROL AND COMPUTER SCIENCES Vol. 55 No. 8 2021

Fig. 5. Sentiment of all D. Trump's tweets.

Fig. 6. Sentiment of messages relative to countries-recipients.

As can be seen from the diagrams in Fig. 5, most of D. Trump's publications are positive. It is worth noting that the algorithm could not classify about 40 percent of messages.

To analyze the sentiment of H. Clinton's mail messages using the word2vec model, a Bayesian classifier was used, the result is shown in Fig. 6.

As is seen in Fig. 6, the most negative mail messages were sent towards Iraq and the UK.

CONCLUSIONS

The analysis, systematization, and improvement of the scientific and methodological apparatus of mediametric information assessment in the GIN Internet in order to determine its nature have been carried out. The analysis showed that mediametric research of information messages is carried out for sepa-

rately taken applied industries; there is no comprehensive systematic approach to solving problems of quantitative assessment of publications.

An approach to the quantitative assessment of information in the GIN Internet based on data mining has been proposed. The approach involves the use of modern machine learning algorithms, which will automate the process of classifying publications in the media by sentiment. The disadvantages of this approach are the mathematical complexity in determining the sentiment of messages, as well as the need for representative samples marked up for training.

The proposed approach made it possible to develop a comprehensive methodology for assessing information publications in the GIN Internet, which takes into account the existing methods of mediametric analysis, and also uses methods of quantitative analysis based on machine learning algorithms. The use of the methodology allows assessing text messages from sources that differ in technical characteristics, as well as determining the sentiment level of the message or source.

Thus, the developed complex methodology for assessing information publications in the GIN Internet makes it possible to quantitatively determine the main statistical indicators of the source of messages and take into account the economic indicators of resources. The technique also allows calculating the sentiment level of text publications using machine learning algorithms when analyzing the nature of information and its source.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. RF Presidential Decree no. 646 of December 5, 2016 *On Enactment of Doctrine of Information Security of the Russian Federation*, Kremlin Official Site. <http://www.kremlin.ru/acts/bank/41460>.
2. Nikolaichuk, I.A., *Politicheskaya mediametriya. Zarubezhnye SMI i bezopasnost' Rossii* (Political Mediametry: Foreign Mass Media and Security of Russia), Moscow: RISI, 2015.
3. Poluekhova, I.A., The studies of audience and media consumption in the digital environment: Methodological and practical problems, *Mediascope*, 2016, no. 4, p. 6.
4. Vartanova, E.L., *Mediaekonomika zarubezhnykh stran. Uchebnoe posobie* (Economics of Media in Foreign Countries: Handbook), Moscow: Aspekt Press, 2003.
5. Bulanin, D., *Novye slova i znacheniya* (New Words and Meanings), vol. 3 of *Slovar'-spravochnik po materialam pressy i literatury 90-kh godov XX veka* (Dictionary and Reference Book on Materials of Mass Media and Literature of 1990th), St. Petersburg: Parkomat-Ya, 2014.
6. Zegzhda, P.D., Zegzhda, D.P., Nikolskiy, A.V., Using graph theory for cloud system security modeling, in *Computer Network Security. MMM-ACNS 2012*, Kotenko, I. and Skormin, V., Eds., Lecture Notes in Computer Science, vol. 7531, Berlin: Springer, 2012, pp. 309–318. https://doi.org/10.1007/978-3-642-33704-8_26
7. Zegzhda, P., Zegzhda, D., Pavlenko, E., and Ignatev, G., Applying deep learning techniques for Android malware detection, *Proc. 11th Int. Conf. on Security of Information and Networks*, Cardiff, 2018, New York: Association for Computing Machinery, 2018, p. 7. <https://doi.org/10.1145/3264437.3264476>
8. Zegzhda, P.D., Malyshev, E.V., and Pavlenko, E.Yu., The use of an artificial neural network to detect automatically managed accounts in social networks, *Autom. Control Comput. Sci.*, 2017, vol. 51, no. 8, pp. 874–880. <https://doi.org/10.3103/S0146411617080296>
9. Izergin, D.A., Ereemeev, M.A., Magomedov, S.G., and Smirnov, S.I., Information security evaluation for Android mobile operating system, *Russ. Technol. J.*, 2019, vol. 6, no. 6, pp. 44–55. <https://doi.org/10.32362/2500-316X-2019-7-6-44-55>
10. Grus, J., *Data Science from Scratch: First Principles with Python*, O'Reilly Media, 2015.
11. Data sample on corona virus pandemic in the world. www.kaggle.com/imdevskp/corona-virus-report. Cited April 5, 2020.
12. Data sample of emails of H. Clinton. <https://github.com/benhamner/hillary-clinton-emails>. Cited April 5, 2020.

Translated by S. Avodkova