

STRUCTURAL ABSTRACT

Report 94 pages, 23 figures, 19 tables, 76 sources, 7 appendixes.

INFORMATION SYSTEM, INFORMATION SOURCES, INFORMATIVE FEATURES, DEVELOPMENT OF ANALYTICAL TOOLS, MONITORING, SOCIETY, DATA VALIDATION, DESIGN, ANALYTICS

Object of the research is open electronic textual news information sources and their content.

The purpose of this work is study and develop various methods of evaluating open information sources’ influence on society based on the analysis of published textual information and their algorithmic implementation as a part of the corresponding information and analytical system.

During the work a sociological research has been done to evaluate the impact of open information sources (electronic media) on the society of Kazakhstan. The survey has covered 3,200 respondents from 16 regions of Kazakhstan. Informative features have been worked out, methodological approaches of work package have been developed. Integrated assessment of the used informative features’ significance and their experimental confirmation has been carried out. Research and approbation of existing hardware and software capabilities of a text classification system implementation, including cloud platforms of text processing in a natural language.

As a result of the study, a balanced, representative and marked corpus of journalistic texts with a total volume of 5,400 texts out of 5 sources has been formed for the first time. Linguistic markup of the corpus has been carried out for 2000 texts and according to its results dictionaries have been compiled and the rules according to 4 features have been formalized.

Main indicators: STS structure and its models have been created, a model has been created based on TOFI technologies in order to collect data on media publications and to make calculations by the method of obtaining an integrated evaluation of publications’ reliability degree, a dictionary-based conceptual model has been developed and a method of calculating text generalization based on program and expert approach has been developed, a clustering algorithm and its software implementation have been developed, vector and cluster representations of news information have been developed. A text classification method based on aggregating subjective probabilities has been developed, server equipment for the computational core of the information system has been set up, a virtual infrastructure and communication channel with the computational core has been set up.

CONTENT

[INTRODUCTION 8](#_Toc528232840)

[1. MODELS OF SOCIAL TRUST SYSTEM FOR STIMULATION OF PERSONALITY’S SUSTAINABLE DEVELOPMENT 10](#_Toc528232841)

[2 DEVELOPMENT OF METHODS FOR ASSESSING THE INFLUENCE OF OPEN INFORMATION SOURCES ON SOCIETY ON THE BASIS OF PUBLISHED TEXT INFORMATION ANALYSIS 13](#_Toc528232842)

[2.1 Evaluation of Media impact on society 13](#_Toc528232843)

[2.2 Methods for assessing the impact of open information sources on society based on the analysis of published textual information 14](#_Toc528232844)

[2.3 The method of obtaining integral estimations of the reliability of the analyzed publication and its influence on society 16](#_Toc528232845)

[3 STUDY OF EXISTING INFORMATIVE FEATURES TO DETERMINE ACCEPTABLE CRITERIA FOR ASSESSING THE IMPACT OF OPEN TEXTUAL INFORMATION SOURCES ON SOCIETY 20](#_Toc528232846)

[3.1 Informative features for determining impact assessment criteria 20](#_Toc528232847)

[3.2 Description of generalization study stages as one of the features of bias in media publications 25](#_Toc528232848)

[3.3 Extracting named entities from text 28](#_Toc528232849)

[3.4 Analysis and approbation methods for classifying texts in the task of publications sentiment evaluation 29](#_Toc528232850)

[4 STUDY OF PECULIARITIES OF ASSESSING OPEN TEXTUAL INFORMATION SOURCES’ IMPACT ON SOCIETY IN THE KAZAKHSTANI CONTEXT. 31](#_Toc528232851)

[4.1 Legal regulation in the field of media monitoring 31](#_Toc528232852)

[4.2 Experience of monitoring of open textual information sources in Kazakhstan 31](#_Toc528232853)

[4.3 Sociological study on evaluation of open information sources' influence (electronic media) on the society of Kazakhstan 32](#_Toc528232854)

[5 CREATION OF NECESSARY TECHNICAL AND EXPERT-ANALYTICAL CONDITIONS FOR DEVELOPMENT OF THE INFORMATION SYSTEM OF EVALUATING OPEN TEXT INFORMATION SOURCES' INFLUENCE ON SOCIETY 45](#_Toc528232855)

[5.1 Analysis of existing hardware and software capabilities implementing text classification systems, including the analysis of word processing cloud platforms in natural language, as well as analysis and evaluation of cloud services’ quality in a natural language text processing. 45](#_Toc528232856)

[5.2 Research and development of methods for clustering textual information based on its semantic content. 48](#_Toc528232857)

[5.3 Development of clustering algorithms 59](#_Toc528232858)

[5.4 Creation of an algorithm that maps multidimensional points to one dimensional single line by means of preserving their distances between them 62](#_Toc528232859)

[5.5 Formation of text boxes and their markup 63](#_Toc528232860)

[CONCLUSION 67](#_Toc528232861)

[APPENDIX A Work schedule 75](#_Toc528232862)

[APPENDIX B List of publications 84](#_Toc528232866)

[APPENDIX C List of foreign information resources 87](#_Toc528232868)

[APPENDIX D Model of methods for obtaining an integrated assessment of the media publication's reliability degree on the basis of TOFI technologies 88](#_Toc528232870)

[APPENDIX E Informative features 90](#_Toc528232872)

[APPENDIX F Frequency of the generalization quantifier “all” in the Wikipedia corpus 92](#_Toc528232874)

[APPENDIX G Software application for recognition of generalization in texts (screenshots) 93](#_Toc528232876)

# INTRODUCTION

At the present stage of society’s technological development digital information obtained from various network information sources is becoming increasingly socially important. There is a continuous increase in the volume of publicly available information. Individuals and their groups are both sources of information posted on the network and its consumers.

Information presented in the network sources may have a different degree of social significance, may have different effects, both on individual social groups and on the whole society in general. In view of this, an important scientific and practical task arises in regards to the integrated assessment of various information’s potential impact on various social groups or on society as a whole.

The purpose of this work is study and develop various methods of evaluating open information sources’ influence on society based on the analysis of published textual information and their algorithmic implementation as a part of the corresponding information and analytical system.

Development of these methods is carried out considering that in the future they can be used as part of a more generalized system of social trust. In this case, social trust system (STS) refers to the way a society is organized based on numerical evaluation of the trust relationship between its constituent subjects (individuals, groups and organizations). Building the complete social trust systems requires the use of a complex analysis of a various informative feature expressed in numerical form.

Each of these features may relate to different socio-economic spheres, reflecting the relationship of individuals and their groups with society. Interaction in information space of separate individuals and social groups is also an important aspect in STS creation. A multi-criteria evaluation of the information sources’ influence on society can be used as one of the significant factors within the STS. Individuals, their groups and entire organizations can act as information sources.

Nowadays society is increasingly acquiring features of a global information society. Nevertheless, one of the most important contemporary global challenges is the development of an individual culture of information consumption. The technologies, in multi-criteria evaluation of open information sources’ influence on society, developed within this project can take on the role of the tools in developing individual information and consumer culture. In this paradigm, an individual acts not only as a consumer of information, but he/she gets access to information about the information he uses in the form of a multi-criteria numerical evaluation. As well as the society gets information about the socially significant aspects of his/her information space, the information sources available in it and their multi-critical numerical evaluation. For example, the availability of information about alternative points of view on the issue under study in an information search in the network can have a significant positive effect on the information consumer culture of an individual and thus contribute to its sustainable development.

In accordance with the work schedule of this project, the following tasks are planned for the 2018 reporting year:

1. Development of methods for evaluation of open information sources’ influence on society based on the analysis of published textual information.
2. Study of existing informative features in order to determine acceptable criterion of evaluating influence of open text information sources on society.
3. Study of evaluation features of open information sources’ influence on society in the Kazakhstan context.
4. Creation of the necessary technical and expert-analytical conditions to develop an information system of evaluating open information sources’ influence on society.

All the tasks have been fully performed and their current status is reflected in the given interim report.

# MODELS OF SOCIAL TRUST SYSTEM FOR STIMULATION OF PERSONALITY’S SUSTAINABLE DEVELOPMENT

The general conceptual scheme of the social trust system is shown in Figure 1.

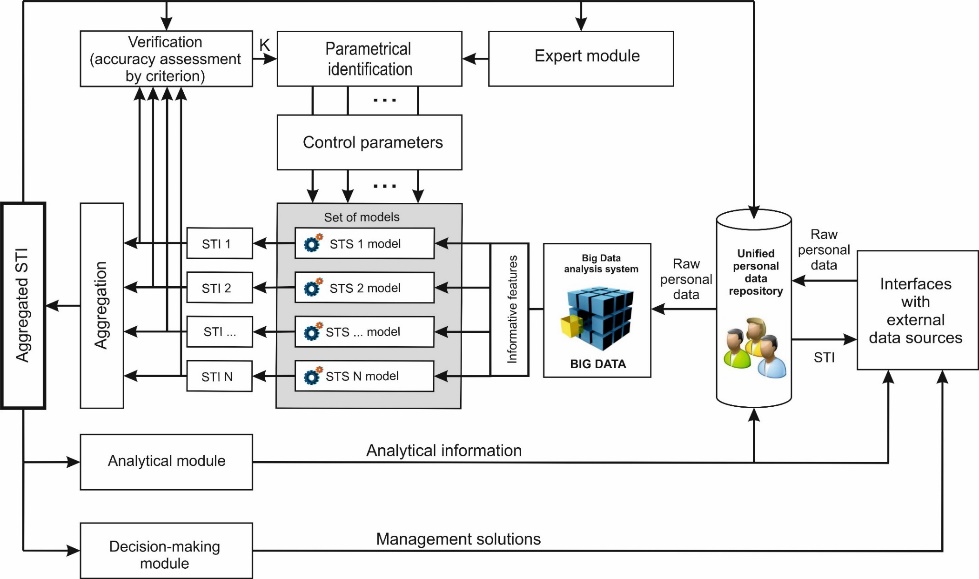


Figure 1 - STS structure

Mathematical models used within STS can be continuously refined.

The life cycle of the STS system is shown in Figure 2.

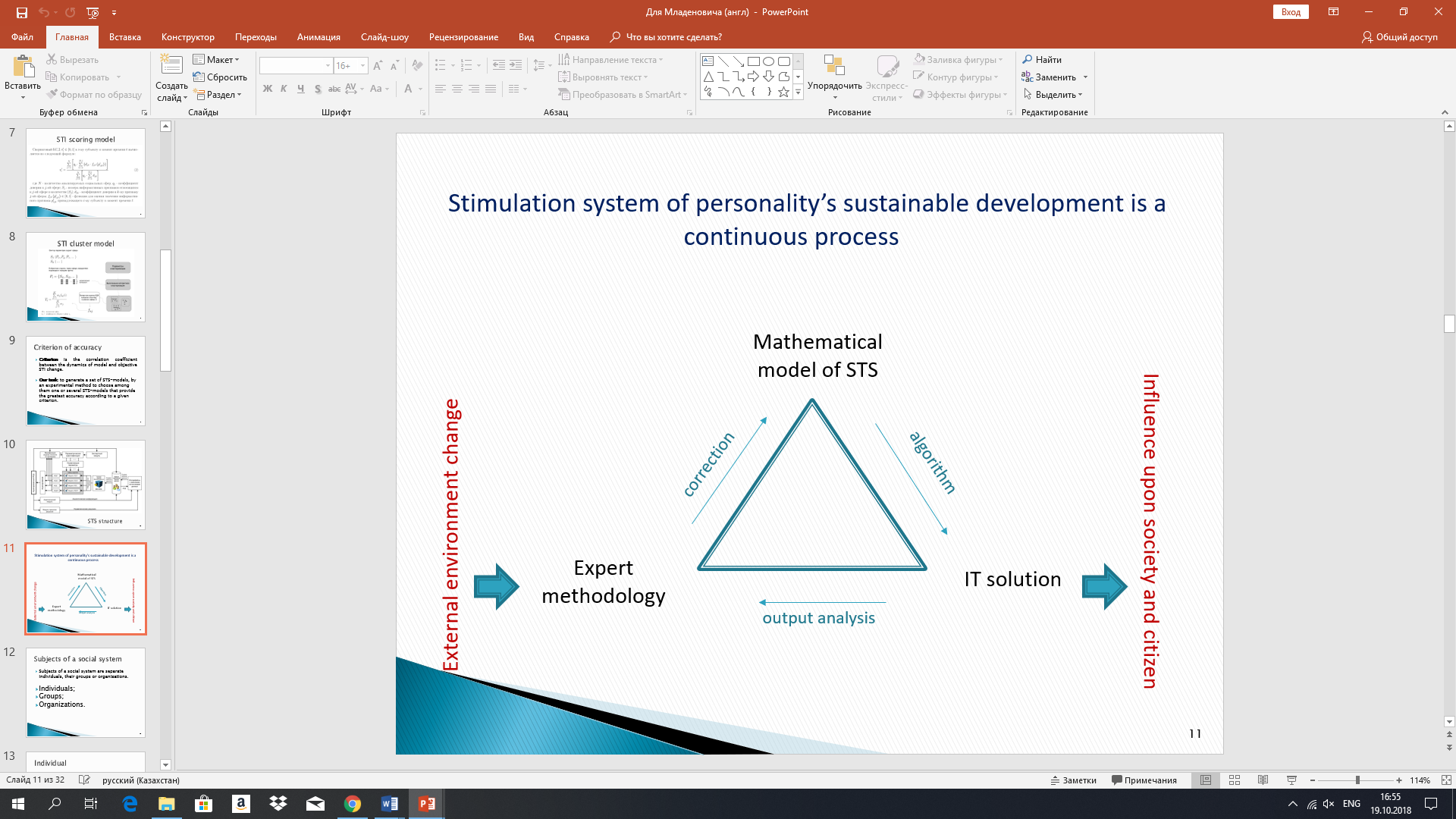


Figure 2 – Stimulation system of personality’s sustainable development is a continuous process

The indexation mechanism of a citizen’s social accountability considering all spheres of his/her life activity will be based on STS (Figure 3) [1].

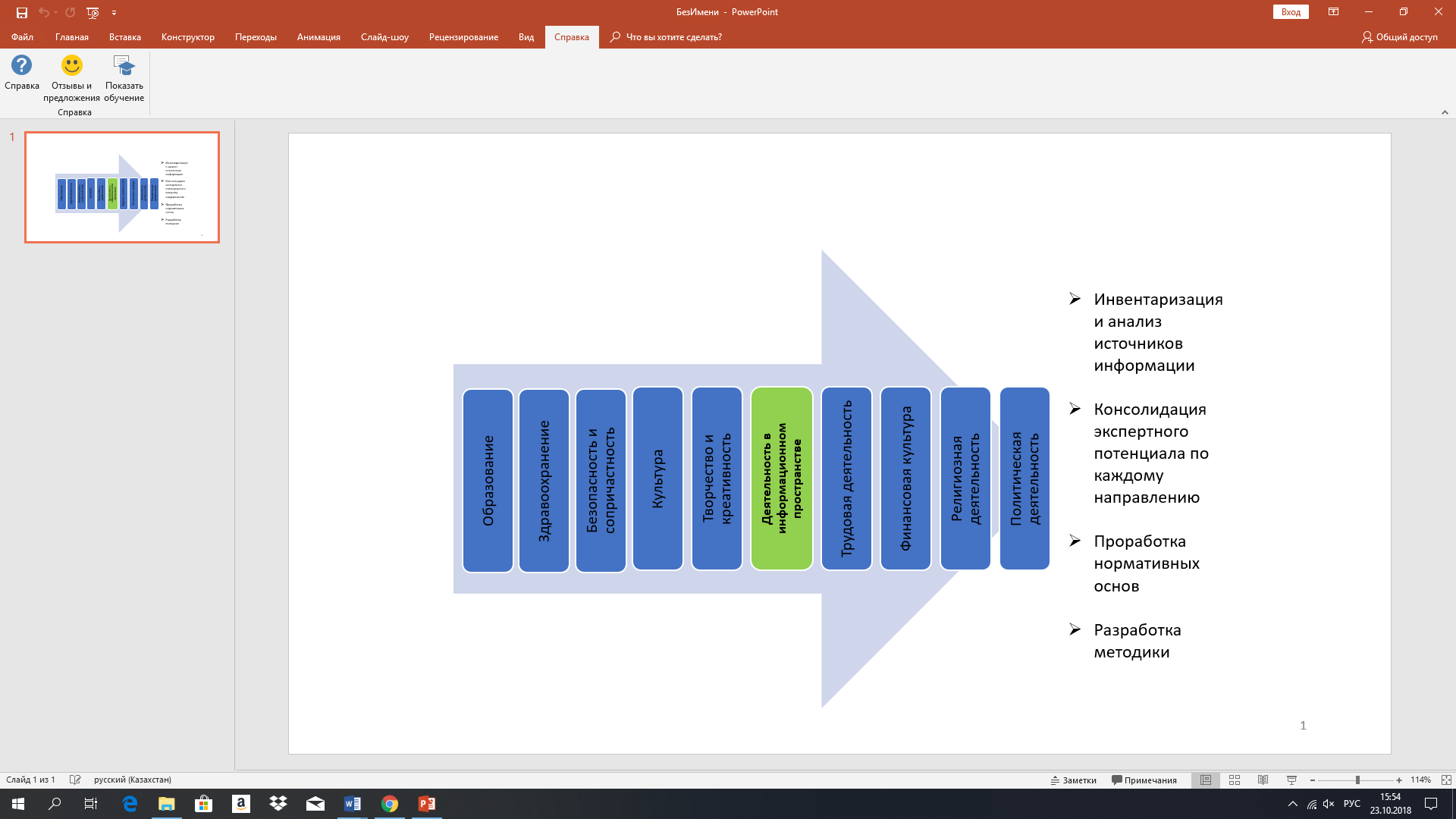


Figure 3 – The indexation mechanism of a citizen’s social accountability

*Social Trust Index.* Trusting the people in general is located between trust in close and unfamiliar people/ strangers [2]. Social trust is when a person trusts the state and society, i.e. according to the rules established in the society, provides the necessary personal, financial information, openly expresses his/her opinion, without being afraid that these rules will be violated by the interacting parties and being sure that all participants adhere to these rules.

The objective is to form a set of STI-models, among which, by an experimental method, select one or more STI-models providing the highest accuracy according to a given criterion.

*Objective social trust index.*Objective social trust index to the i-th subject at time t is calculated by the following formula:

(1)

where – a set containing the numbers of all subjects with which the i-th subject interacts at time t; – number of elements in the set , – trust index of the j-th subject to the i-th subject at time t. All trust indexes can be set at a certain date t in the process of surveys, social network and Web-resources analysis. The objective STI reflects the trust situation that actually exists in the social system and can be used for comparison with artificial model STI.

*STI Scoring Model.* STI scoring model to the i-th subject in the timepoint t is calculated by the following formula:

(2)

where N – a number of analyzed social spheres; – trust index to the j-th sphere; – numbers of informative features relating to the j-th sphere in the amount of ; – trust index to the k-th feature of the j-th sphere; – a function for evaluating the value of the informative feature belonging to the i-th subject at time t [2].

*STI cluster model.* The parameter vector sets the spheres . In the opposite direction: through spheres we define individuals and find the groups . On the other hand, clustering parameters are defined, and the clustering algorithm is carried out using normalization validation. After the clusters are determined using the algorithm, the experts estimate the cluster trust index (CTI) of each cluster within the sphere of S. Thus, the social trust index to the i-th individual within the sphere of is calculated by the formula:

(3)

where M – number of spheres, – trust index to the sphere [3].

*Criterion of accuracy.* As a criterion for determining the accuracy of various STI-models (1) - (3), we can use the correlation index between the corresponding time sequences of objective and model coefficients:

(4)

Development of scientific and methodological base and the subsequent introduction of social trust system with high criterion of accuracy at the state level will give a unique information system in which the main key criteria of development and formation of the personality are integrated.

The issues of integrating numerical evaluation of the open information sources’ influence on society within various models of STS are being investigated in the framework of this project.

# DEVELOPMENT OF METHODS FOR ASSESSING THE INFLUENCE OF OPEN INFORMATION SOURCES ON SOCIETY ON THE BASIS OF PUBLISHED TEXT INFORMATION ANALYSIS

## Evaluation of Media impact on society

Since the importance of the media in building an individual's worldview and their role in shaping public consciousness is undeniable, the assessment of media influence on society is one of the most popular areas of applied research. To date, the focus of such research has shifted from assessing the impact of television to assessing the impact of social networks due to their growing popularity. The findings of the research conducted by the MediaKix marketing company in 2016 aimed at assessing how much time users spend on social networks showed that electronic media and social networks took second place after the television. At the same time, in 2015 in the USA, users spent more time on social applications than on television. For the majority of Kazakhstani population (59.8%), main sources for obtaining news information are electronic sources: social networks/bloggers (30.3%), as well as news websites (29.5%) [11]. Despite the continuing interest of the scientific community in assessing media influence, this area is featured by the presence of many different approaches and theories. There are no unified commonly acceptable methods. In addition, theoretical developments in assessing media impact on society are numerous, unstructured, and significantly different from each other.

Methodology for assessing the media influence on society includes widely used methods of sociological research, content analysis and some quantitative indicators. The most common methods include sociological surveys, expert opinions, discourse analysis, content analysis, the case study method, graphematic and syntactic analysis.

*Manipulative mechanisms of media influence.* Mass media use different types of psychological means of influence and often their combinations to broadcast a certain position on a particular situation. According to a sociological survey of Kazakhstan population, the majority of respondents believe that Kazakhstan’s informational Internet sites are used as tools to discredit certain individuals (43.4%) and to form a positive image (55.1%). At the same time, 52.7% of the population believe that Kazakhstan's online media sources are used to highlight events and (or) activities in a particular area from a certain angle that benefits a certain circle of people [11]. These data allow concluding that, in the opinion of Kazakhstani people, the media are a tool for the formation and manipulation of public opinion.

Studied mechanisms of media influence on society are not a complete set of tools that the media use to form a picture of the world to be perceived by the public. The presence of manipulative techniques can be considered one of the main informative features for assessing the impact in this aspect, which includes generalisation, politicisation, distortion of information, and other.

## Methods for assessing the impact of open information sources on society based on the analysis of published textual information

Within the Project, the first version of the Methodology has been developed for assessing the impact of open information sources on a society based on published textual information analysis [57]. Main advantages of the assessment methodology being developed are as follows:

* group (aggregate) and individual indices vary in the range from 0 to 1;
* a holistic and balanced system of indicators of the impact of open textual information sources on society is proposed;
* a multi-level assessment system (for news web-sites as a whole, for a separate news web-site, for a separate publication) and ranking depending on the set analysis tasks and level of detail is provided;
* the assessment considers the dynamics of the quality of textual information within a certain period and compares different values of the indicators;
* the assessment allows obtaining both the integral value of the influence of open textual information sources, and group values for individual evaluation criteria.

General provisions of the developed methodology.

*Assessment goal*: determining the level of influence of open information sources on the society based on measured indicators using an automated information system.

*Key tasks:*

* determining the level of influence of the media-texts of Kazakhstani media based on criteria characterising the factors of media influence;
* determining the degree of reliability of news in national media;
* generation of analytics, statistical data based on assessment;
* monitoring of media consumption.

*Assessment object:* media-texts of Kazakhstani news web-sources.

*Assessment subject:* reliability, tone, public response, unbiasedness of the media-texts of news web-sites and media involvement of the audience.

Stages of the Methodology Development. Development of a methodology for assessing the impact of open textual information sources on a society based on an analysis of published textual information includes several stages, where at each stage a particular method and approach is used. *The first stage* is the study of existing informative features in order to determine acceptable criteria for assessing the impact of open textual information sources on society; *The second stage* is to determine the nature of the influence of open information sources on various social groups of the Kazakhstani population. *The third stage* is the definition of criteria and indicators for assessing the impact of open textual information sources on a society developed from informative features. *The fourth stage* is to conduct a comprehensive assessment of the significance of used informative assessment criteria and their indicators. *The fifth stage* is an expert assessment (marking) of the formed corpus of journalistic texts with the use of informative features and indicators. *The sixth stage* is the formalization of rules and formation of vocabularies according to informative features that can be determined with the help of linguistic practices.

Determining the assessment criteria. The results of the investigation of existing informative features of the influence of open textual information sources on the society allowed to determine existing informative features of assessment of media-texts [67]. Selected informative features were grouped into five informative features most appropriate for the assessment purpose: reliability, public response, tone, unbiasedness and media-involvement. 22 indicators were developed to assess the impact of open textual information sources according to proposed criteria. A comprehensive assessment of significance was carried out to determine the significance of selected informative features and assessment indicators. Based on the results of a comprehensive assessment of the significance of such informative publication criteria as reliability, unbiasedness and tone have a statistically significant effect on the number of views. [68-69].

Balanced, representative corpora of media texts have been formed, marked-up and classified in parameters for recognition of informative features in publications and development of methods for calculating them. Double-stage systematic cluster sampling was carried out to form corpora. For corpora marking experts applied two approaches. *The first approach* is to assign a publication to a specific class and/or parameter. *The second* is an in-depth markup of publications for the formalization of linguistic rules and formation of vocabularies by the parameter.

Assessment of the influence of open information sources on society. The developed system for assessing the influence of open textual information sources is a set of evaluation procedures and a multi-level system of indicators based on the principles of decomposition and aggregation.

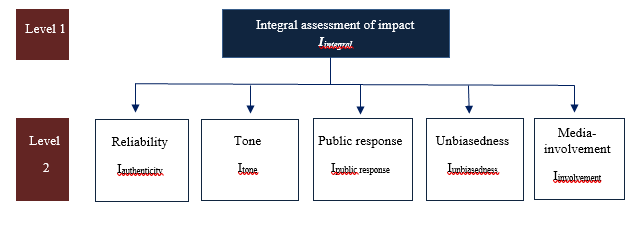


Figure 4 - Decomposition of indicators for assessing the impact of open textual information sources on society

As seen from Figure 4, a general (integral) impact assessment is performed at the highest (first) level. Decomposition of an integral indicator of impact is carried out according to five comprehensive assessment criteria: reliability, tone, public response, unbiasedness, media involvement.

For each assessment criteria, the corresponding group of indices are aggregated into a standard (integral) impact index. An essential task in the calculation of comprehensive assessments is the justification of weights that reflect the significance of a particular indicator (or group of indicators). Weights are set from expert assessments. The proposed system of assessment procedures and indicators includes the following sections:

* by the evaluation criteria of all news portals: reliability, tone, public response, unbiasedness, media involvement;
* by the evaluation criteria of a separate news portal;
* by the evaluation criteria of a separate publication.

# The method of obtaining integral estimations of the reliability of the analyzed publication and its influence on society

The obtaining task of the degree reliability integral estimates of the analyzed publication belongs to the so-called class of decision-making problems under uncertainty. Indeed, the reliability evaluation of information is establishing the process of the degree of conformity of our knowledge about an object (within the framework of the information model tasks used in solving problems) to the real state of the object, i.e. determination of the degree of adequacy of representations to the real state of the original object (object, phenomenon). (If available), which should reflect the complex structure of the system, to decision-making, operate, if necessary, the subjective evaluation of experts, consider the ambiguity, inaccuracy of these models, etc. It is known that in evaluating the objects of an arbitrary nature approaches have worked well in one way or another based on the analysis method of hierarchy developed by American scientist known T. Saaty [12]. Therefore, the possibility of applying this method to the development of integrated evaluation techniques of the reliability of the analyzed publications and its impact on society as a part of a single body of publications were investigated. After analyzing the various conceptual approaches, a large number of primary sources, analyzing advantages and disadvantages of other approaches previously defined contours of the method of obtaining integral estimations of the reliability of the analyzed publication and its influence on society. Calculations according to the Methodology are easily automated in the modeling environment TOFI [59-61] after the formation of the data set of the analyzed publication. Based on the analysis domain suggest the following, a preliminary list of factors and evaluation of relevant indicators integrated text publications:

1. *Group of factors 1.* Relevance F1
   1. The frequency of mentioning the main theses and conclusions
   2. The total relevance in terms of the well-known search engines
   3. Reliability of sources
2. *Group of factors 2.* Entropy F2
   1. Shannon Entropy
   2. Informational entropy
3. *Group of factors 3.* Impact on society F3
   1. Visits frequency
   2. Key reviews

This list of factors and indicators of evaluation will be further refined and supplemented in the course of subsequent research. The principle and the procedure for various calculations and integral evaluation index does not depend on the number of factors and a list of indicators.



Figure 5 – Objectives hierarchy for assessing the reliability of texts

For each factor, we obtained the following ready-to-use table of normalized indicators.

Table 1 – Normalized indicators for factors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Factor indicators Fs | Value indicators by publications | | | |
| Т1 | Т2 | … Тj … | Тm |
| П1 | Y11 | Y12 | …Y1j … | Y1m |
| П2 | Y21 | Y22 | …Y2j … | Y2m |
| П3 | Y31 | Y32 | …Y3j … | Y3m |
| …. |  |  |  |  |
| Пi | Yi1 | Yi2 | …Yij … | Yim |
| … |  |  |  |  |
| Пns | Yns,1 | Yns,2 | …Yns,j … | Yns,m |

In this table all values normalized figures purely numeric and dimensionless.

Moreover, for all values, the relation is:

0 < < 1 (8)

Indicator Пк is relevant to the factor Fs, determined coefficient significance αкs, wherein:

α1s + α2s +…. + α ns =1 **(**9)

The advantage of the approach is the following. In fact, we determine the value index of the factor by the formula:

IFs ( Tm) =α1 s Y1m + α2 s Y2m + ….+ α ns Ynm , (10)

Fixed the values of the normalized indices in (10) and determine the unknown coefficients of significance so that the functional IFs (Tm) had a local maximum. Then the coefficients significance are determined uniquely.

Factor Fs is relevant to the integral index IH (Тк), determined by its coefficient significance αsк. Increasing the index all factors values (factors selected like that) leads to an increase the integral index. Such an approach has already been used in the development of «Methods of integrated assessment of Kazakhstan territory» [13]

Table 2 – Indices primary factors used for calculating the integral evaluation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Factor  indices | Evaluated texts | | | | | |
| Т1 | Т2 | Т3 | Т4 | … | Тm |
| IF1 | Z1,1 | Z1,2 | Z1,3 | Z1,4 | …. | Z1,m |
| IF2 | Z2,1 | Z2,2 | Z2,3 | Z2,4 | …. | Z2,m |
| IF3 | Z3,1 | Z3,2 | Z3,3 | Z3,4 | …. | Z3,m |

Total (integrated) evaluation of the reliability Tk text (Publication) is calculated by the formula:

IН ( Tk) = (Z 21,к +Z 22,к + Z 23,к) / (Z 1,к + Z 2,к + Z3,к), к=1, m.

To collect data on publications in the media and carry out calculations using the proposed methodology, a model was created based on the technologies of TOFI (Appendix D).

# STUDY OF EXISTING INFORMATIVE FEATURES TO DETERMINE ACCEPTABLE CRITERIA FOR ASSESSING THE IMPACT OF OPEN TEXTUAL INFORMATION SOURCES ON SOCIETY

# Informative features for determining impact assessment criteria

In this research, we use various informative features to analyse the impact of open textual information sources on society, based on which assessment indicators are being developed. Various media sites and researchers use their indicators and methods for calculating them, and, accordingly, there is a significant range of variations of informative features for assessment.

Considering the direct relationship between indicators and informative features, the authors defined informative features based on the analysis of assessment indicators used in practice as a part of the international experience of assessing media impact. They studied widely used indicators for assessing the impact such as media coverage, involvement index, refusal index, audience growth rate, unlike rate, the tone of references towards the publication and comments.

A comprehensive analysis of theoretical calculations and practical developments in the field of media impact on society showed that today the media have an extensive arsenal of various methods and techniques of influencing the audience and actively use them. Effective handling of such techniques determines the role of the media as one of the most significant factors in shaping a person’s worldview.

Despite a large number of indicators that allow evaluating certain aspects of media impact, the general structure of impact analysis, as well as the nature, character and methods of calculating quantitative indicators, and definition of informative features depend on the objectives of the research. Whereas the same indicator can be calculated differently, depending on which aspect of the media impact or media effect is the object of the assessment, and also depending on which way and which primary data were collected for assessment.

According to the results of the analysis, the authors identified two groups of informative features that would be appropriate to use when assessing the impact of open information sources on society. Informative features of the first group characterising the “*Public response*” assessment criterion include the following features:

* availability of a certain number of readers;
* readers' interest:
* readers' reaction.

It is proposed to use the indicators below based on the defined informative features to assess media impact on society using an automated information system. These indicators are a quantitative expression of informative features and enable a more detailed assessment of publications on the following grounds.

1. *Media-coverage.* This indicator allows estimating how many people have viewed the publication and, accordingly, are aware of the published news. It characterises the scale of its distribution. The indicator is calculated as the sum of the number of views of a publication from a single IP address.
2. *Number of shares.* This indicator allows estimating how many people shared this publication in social networks or their accounts on the web resource pages. The indicator enables measuring the fact that the reader agrees with the author's point of view and believes that the news is important and should be disseminated. The indicator can be calculated as the sum of the number of publication shares to the total number of views.
3. *Number of comments.* This indicator measures the audience's interest in the news and encourages readers to comment. The indicator can be calculated as the sum of the number of comments to one publication to the total number of views of this publication.
4. *Indicator "tone of the comment"* is defined in the following range: negative, positive, neutral. It is calculated using the method of linguistic analysis when the program registers the words-markers of a negative, positive or neutral attitude of an individual reader. Another option for determining the tone by an information system is the use of machine learning technology "Deep Learning".

The indicators developed from informative features of the Public Response criterion characterise the specificity of media coverage assessment, determine the scale of distribution of media-texts, determine the nature of the audience's attitude towards the publication.

The next group of informative features refers to the “Media Involvement" criterion, which describes the reader's behaviour on a particular web resource. This group of informative features cannot be used to evaluate a separate publication. The reason for such kind of separation is the impossibility of applying media involvement metrics to the “publication” analysis unit because of their nature. However, this group of informative features allows assessing the loyalty and interest of the audience in the web resource itself and acts as a conditional indicator of the audience’s trust in it and its potential influence on the audience. This aspect of influence cannot be excluded from a comprehensive assessment, because according to a sociological survey conducted by the authors, the most significant number of respondents (32%) in determining their level of trust in news reports consider the source of information and its reputation to be the most important factor [11].

Thus, informative features related to the criterion of media involvement are defined as follows:

* loyalty of audience to a web-resource;
* the audience is interested in publications of a specific web resource, preferences of its readers.

Quantitative expression of these informative features can be indicators that measure:

1. Visit depth. The indicator is calculated as the ratio of the number of sessions with viewing more than N pages to the total number of sessions.
2. Duration of a visit. The indicator is calculated as the ratio of the number of sessions lasting more than N minutes to the total number of sessions.
3. The frequency of returns to the website. The indicator is calculated as the ratio of the number of sessions with viewing more than N pages over the N period to the total number of visits to the site.

*Informative features to determine the reliability of information.* Data from the analysis of various literary sources for determining the reliability of news show that the approaches and features for determining the reliability of publications are the same. However, most features of determining the reliability of information, data, as well as publications are based on expert assessments. Reliability of text-based news is determined using a combined approach — expert analysis of publications and application of information systems.

According to the results of the analysis of approaches to determine the reliability of new, the following informative features of the reliability of an individual publication has been determined:

1. Availability of references in the text to the competent source:

* information of state authority of Kazakhstan;
* information of an official Internet resource of a Kazakhstan state authority;
* information published by the entities of quasi-state sector of Kazakhstan;
* information published by Kazakhstani national companies;
* official information of news agencies, electronic media and other similar companies with a high level of reputation;
* data from analytical reports, publications, studies prepared by authoritative and recognized international organizations (UN, WHO, UNESCO, IMF, the World Bank, the European Bank for Reconstruction and Development, Interpol and other resources of government departments of other countries);
* recognized international statistical databases (UNESCO Institute for Statistics, Eurostat, OECD, the World Bank statistical information, and other);
* information on professional expert platforms (advisory, rating platforms) recognised in their area, on which representatives of authoritative international organizations rely on;
* results of research prepared by scientific organizations, universities, and other;
* response to a request made to government agencies and international organisations.

1. Reference to the primary source of information.
2. An indication of the author's name in the publication.
3. Rating/reputation of the media (news websites and information agencies, electronic media, other similar sites), which released the publication.
4. Coverage of the same event in different sources: crosschecking, discourse with other publications.
5. The authenticity of photos and videos that confirm the event.
6. Compliance of the title of the publication with its content.

Informative features to determine the tone and unbiasedness of the publication. Determining the tone of the publication is one of the mechanisms for monitoring media texts in order to assess the level of influence on readers' opinions becomes one of the critical elements when conducting a content analysis. All three variations of the sentiment-analysis - statistical one, the use of dictionaries and rules, as well as the mixed type of analysis, involve identification of certain words and word combinations in the text, which is considered as informative features for classifying publications into a particular tone category. A significant obstacle to an adequate analysis of the tone of the text is the presence of irony and sarcasm in the text; its correct determination can significantly improve the accuracy of the results. When using the elements of irony and sarcasm in the text, the meaning and general tone of the text can acquire the opposite meaning to the direct interpretation of the text. According to a study held by the Higher School of Economics, some signs of the presence of irony/sarcasm in the text may be reinforced exclamations such as excessive use of exclamation marks or question marks, the presence of interjections, irrealis markers, and their combination [14].

The results of studying the methods of content- and sentiment-analysis of media texts made it possible to determine the following informative features for the criterion "Tone":

* the presence of words/phrases with a pronounced negative colouring in the text;
* the presence of profanity in the text;
* the presence of irony and sarcasm;
* the presence of words/phrases with a pronounced positive colouring in the text;
* the presence of praise.

Unbiasedness of information is generally viewed as a lack of prejudice, with reference to the media unbiasedness is one of the basic principles of professional journalistic ethics assessment of which includes four components (according to D. Brewer).

Unbiasedness of information is one of the integral elements of the ethical code of journalists and the Law of the RK “On Mass Media” and provides an opportunity to assess the impartiality of the author of the publication to the considered events. Unbiasedness is a socially significant property of journalism, in which the author accurately and comprehensively presents the facts, confirms the information with competent opinions and does not manipulate the information, thereby imposing the opinion of any of the parties to the event. As the survey data showed, the majority (65%) of the respondents partially or fully agreed with the opinion that, using the information only from Kazakhstan’s internet sources, one can get an unbiased view of the events taking place in the country.

Based on the information studied, the following informative features were developed for working with the classification of media texts in terms of unbiasedness:

1. *Availability of expressed personal opinion of the author of the publication in the media-text.* At the same time, the presence of expressed opinion is not in itself a basis for classifying the text as “non-unbiased”, however, too frequent use of sentences with a vivid expression of the author’s personal opinion in combination with the absence/small number of facts, as well as the presence of signs of sarcasm/irony in the text, agitation/anti-agitation, may indicate bias in the media-text.
2. *The presence of manipulative techniques in the media-text such as:*

* the degree of generalisation (mild, strong, absent);
* politicisation;
* the presence of a call for action;
* distortion of information;
* presence of facts in the article.

The first four indicate bias, the last indicator is an indicator of unbiasedness.

Selected informative features for assessing the impact of open textual information sources on society can be combined into a group of five informative criteria that are most appropriate for the research: *reliability, tone, public response, unbiasedness, media involvement* (Appendix D).

Thus, according to the results of the study, the following was obtained:

1. The role of the media in shaping public opinion had been reviewed.
2. International and Kazakhstani laws on media regulation had been analysed to develop the basis for assessment criteria.
3. The tools, techniques and mechanisms of impact on society used by the media had been defined.
4. The existing informative features of media texts had been identified and investigated.
5. Criteria and indicators for evaluating the impact of open text information sources on society in line with informative features for subsequent use in an automated information system for assessing the influence of media on the society had been developed [66].

Within the study of the existing informative features, a comprehensive evaluation of the used informative criteria's significance [67] and its experimental confirmation of the adequacy [68] was carried out.

## Description of generalization study stages as one of the features of bias in media publications

Generalization problem has been considered within the task of informative features identification and research, used for preliminary classification and evaluation of media texts’ unreliability degree.

The strategy of generalization in media publications is able to form readers’ opinion, persuasion or prejudice to the subject or event. It is important to study generalization as way to influence society from this position. Generalization recognition is also associated with the identification of reliable articles, generalized publications have a low degree of accuracy. The developed method of generalization extraction can be used for other informative features such as politicization, subjectivity in the text. In our research we consider G. as an authorial technique, by means of which one can express prejudices, bias or stereotypes of the author, by transferring the conclusion formed on the basis of individual facts, events or observations to the entire set of objects or subjects. One of these methods is *“logical transition from the particular to the general, subordination of the particular phenomena to the general principle”,*1 which is conditioned by the property of human thinking to make general conclusion based on his/her own experience and observations. This technique or strategy is the subject of this study. From now forward such technique will be called generalization (hereinafter G.). In this article G. strategy is studied as the means of expressing bias in texts. This paper provides an analysis of scientific work in the field of text processing, research methods, experiments, research results.

Preliminary research. Natural language processing is an interdisciplinary area, because the study intersects with different scientific fields. As part of the research, a preliminary study of the subject area has been done, the works of foreign specialists have been studied, experiments have been performed, hypotheses have been formulated and tested on a marked corpus, and methods of publications annotation on generalization have been developed. The works of psychologists, linguists, political scientists, and developers have been investigated in the course of the studying the subject of extracting bias in the text and generalization. In the works of political scientists T.A. Van Dijk [15] and E. Ajiboye [16] G. is presented as one of the components of the ideological strategy. N.S. Dankova [17] tells about the close connection of G. with subjectivism and sentiment. I.F. Frolova in her work [18] describes G. as a particular problem of bias degree evaluation. The results of bias studies can be found in the works of M. Recasens [19] and F. Morstatter [20]. Researchers agree that G. is readers manipulation instrument. Thus, G. has a close connection with bias, subjectivism [21], sentiment. Researchers agree that G. is readers manipulation instrument [15] - [18]. In the works of political scientists T.A. Van Dijk [15] and E. Ajiboye [16] G. is presented as one of the components of the ideological strategy. N.S. Dankova [17] tells about the close connection of G. with subjectivism and sentiment. I.F. Frolova in her work [18] describes G. as a particular problem of bias degree evaluation. The results of bias studies can be found in the works of M. Recasens [19] and F. Morstatter [20]. Thus, G. has a close connection with bias [17], subjectivism [18], [21], sentiment [18], namely, bias combines subjectivism and sentiment. In our research we consider G. as an authorial technique, by means of which one can express prejudices, bias or stereotypes of the author, by transferring the conclusion formed on the basis of individual facts, events or observations to the entire set of objects or subjects.

Researchers identify three areas of text processing for automatic detection of informative features: lexical approach [22], approach based on machine learning and hybrid approach [23]. In this study, the lexical approach is applied, in the future a carefully developed lexical model and the power of machine learning algorithms will be combined.

Methods/Experiments. One of the first approaches to evaluate G. degree was an approach based on lexical-semantic features. There was an attempt to classify texts according to their degree of G. into five subtypes: very strongly expressed G., strongly expressed G., not identified in the text, weakly expressed G., very weakly expressed G. Researchers of applied linguistics [24] claim that the authors of articles that do not provide facts often make generalizations to form readers’ stereotypes. On this basis, a hypothesis on the influence of cognitive concepts of fact and opinion on the degree of generalization in the text has been formulated, and tested on the marked test corpus. In order to implement the text processing tasks, the research has been done from manual processing of news publications and evaluation of informative features based on expert evaluations to automatic processing methods. The expert approach to features recognition consisted in the articles analysis for generalization and words detection, bigrams (phrases) of the text expressing generalization and sentiment. A conceptual model based on dictionaries has been developed in order to determine the numerical evaluation and a method for calculating text generalization based on software expert approach has been developed. Numerical evaluation of G. was calculated as the sum of G. weights of each generalized sentence to the total number of sentences S in the text:

(11)

where is a set of sentences in the text, is a set of elements (words and phrases) of the sentence that coincide with the elements of the KEY-dictionary, is a set of elements in the sentence that express the fact, is a set of elements in the sentence expressing the opinion of the publication’s author, – sentences that have G., where , , is a number of sentences in the text. The weight of the G. sentence was calculated as the number of elements' matches from the KEY-dictionary with the elements of the sentence. G. sentences, that have only one element from the KEY-dictionary, . This method of calculation also has determined the classification of 2 news genres: analytical articles, information and news articles. Automatic recognition of information and news texts with an accuracy of 83% has been identified (Figure 6).

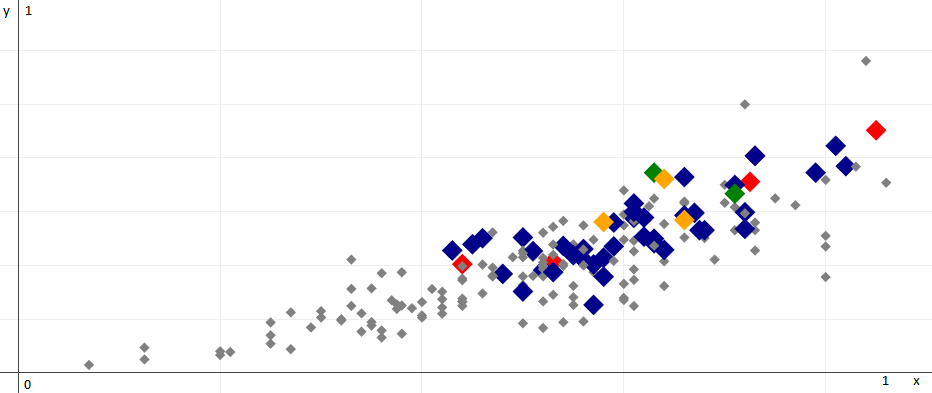


Figure 6 – Recognition of information and news texts

After preliminary cleaning of the corpus, 192 annotated articles of the media are displayed by the degree of their generalization. On the X-axis, the percentage of words and phrases from the KEY-dictionary in the articles is displayed. On the Y-axis (in Figure 6), the estimate of generalization, calculated by the formula (1) is shown. Gray points - generalization in the article is not found; blues and greens have weak generalization, yellows and reds have generalization which is expressed clearly.

Later on, it was decided to develop an annotation scheme for G. recognition in texts and to propose rules for automatic recognition of G.

Corpus. A small corpus from 192 articles has been taken to study properties of G. All publications have been taken from the newsportals and websites containing analytical publications such as tengrinews.kz, camonitor.kz, ratel.kz, time.kz. The articles have been read by linguist annotators and marked as G. or not G. Further, a corpus of analytical articles has been additionally built according to this scheme.

Results obtained. In the course of scientific work, research experiments have been done to study the subject area. Frequency of the generalization quantifier “all” in the Wikipedia corpus has been calculated and the results (Appendix F) have been analyzed.

A dictionary of generalized words, that consists of 600 words and phrases, has been manually formed for the lexical approach of generalization recognition. The dictionary has been improved using RegExp regular expressions [76]. The RuSentiLex dictionary [23] has been used in sentiment analysis and adapted to the program. The dictionary consists of 16000 words and phrases.

A software application with multiple functions on publications processing has been developed. It is designed to check the quality of text recognition for generalization and takes into account the sentiment on the basis of selected informative features. The program is developed in C # with SQLite connection, Regular Expressions (Appendix G).

Within the further researches, markers for linguistic rules have been identified by formalization of generalized words and phrases, syntactic links between them by means of texts annotation. Linguistic rules form a model that is connected to automate the detection of generalization in the text.

For further manual text analysis and in order to detect generalization, author's position, sentiment in the texts of media, a method and instructions have been developed for IAC experts dealing with publications.

More detailed description of the developed algorithm is given in the corresponding publications prepared and published within this project [25-28].

## Extracting named entities from text

There is need to extract entities from the text of an interesting article for analysing frequency of a person, an organization we are interested in, or for detecting location of the article.

We explored an open-source library DeepPavlov, which allows us to extract named entities from a text. The AI library is working on deep neural networks (128 layers). Basically, DeepPavlov is a library of conversational systems which was developed in a laboratory of MIPT. The library contains modules for processing Russian and English languages. It is also implemented as a service ipavlov.lab, where you can also analyze any text you sre interested in. All the source is publicly available. The library is supported only on Linux kernel, which can be an obstacle for inexperienced users of this OS. We took an instance on Google Cloud Platform, where we have done all needed computations.

All the tested data was mined by the Institute from tengrinews.kz: An example of the result is shown in Figure 7.

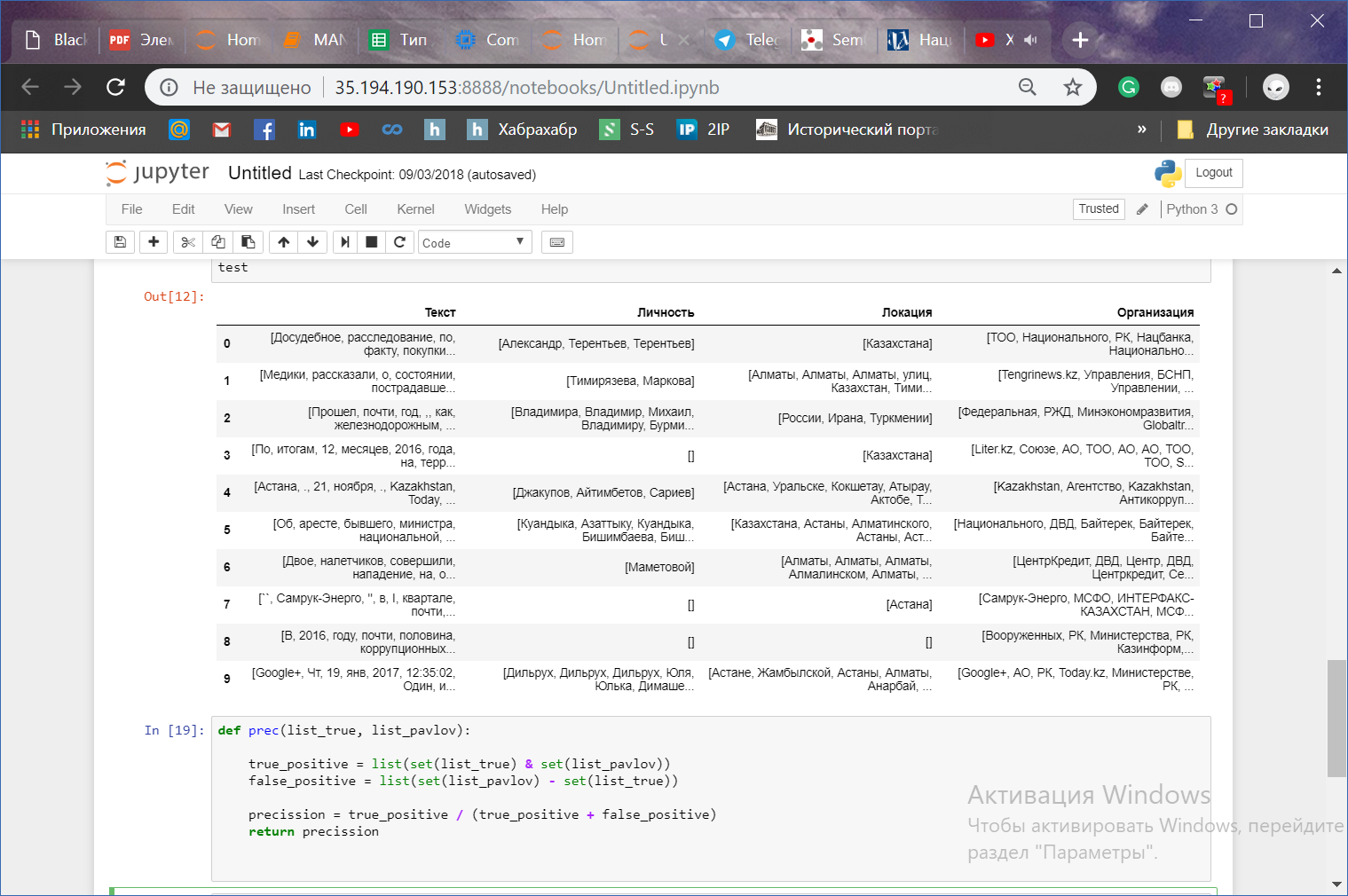


Figure 7 - Test results from tengrinews.kz

We did not have marked data, so we evaluated part of the sample manually. The results are shown in Table 3

Table 3 - Metrics of accuracy and completeness

|  |  |  |
| --- | --- | --- |
| Precission | Recall | Fscore |
| 88% | 75% | 82% |

## Analysis and approbation methods for classifying texts in the task of publications sentiment evaluation

The tone is one of the important parameters characterizing the text. The analysis allowed to assess the possibilities of software tools and services in the conduct of the automatic analysis of texts tone.

A comparative analysis and testing of text classification methods in the evaluation task tonality of publication using classification algorithms for language libraries Python and Google Cloud Platform. The tonality of the text is an important parameter, it to a certain extent determines the text's effect on the reader, its potential danger, and it can also be a sign of the bias in the presentation of information. During the study, the following classifiers were studied: Naive-Bayes, SVM, Linear Regression, Decision Tree and Random Forest. As the dataset was used corpus text collected by company Alem research, containing 8000 news, containing 1500 negative, 4000 neutral and 2500 positive news. For vectorization of texts used TF-IDF. Application results of algorithms are given in Table 4.

Table 4 - Results of applying TF-IDF algorithms

|  |  |
| --- | --- |
| Algorithm | Precision |
| DecisionTreeClassifier | 0.554761 |
| LinearSVC | 0.698417 |
| LogisticRegression | 0.697085 |
| MultinomialNB | 0.637540 |
| RandomForestClassifier | 0.488200 |

It can be seen that the SVM and Logistic Regression showed the best results in this corpus.

As an alternative, a platform was used to analyze the texts tonality Google Cloud Platform. The system displays an tonality evaluation from -1 to 1, while texts rated from -0.25 to 0.25 are considered neutral, as well as magnitude, in the range from 0 to infinity, which shows the emotionality of the text. Since it works only in English, the texts had to be pre-translated using Translate API. The length of the text for translation was limited to 80 words. It is possible that this restriction distorted the results presented in the Table 5.

Table 5 – Classification results using Google Cloud Platform, divided into three classes: negative, neutral and positive

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Precision | Recall | F-1 score |
| negative | 0.51 | 0.41 | 0.45 |
| positive | 0.32 | 0.56 | 0.41 |
| neutral | 0.50 | 0.43 | 0.46 |
| The average | 0.47 | 0.44 | 0.45 |

If we divide only into two classes, the result is much higher (Table 6)

Table 6 - the results of the classification using Google Cloud Platform, divided into two classes: negative and positive.

|  |  |  |  |
| --- | --- | --- | --- |
| Class | Precision | Recall | F-1 score |
| negative | 0.80 | 0.56 | 0.60 |
| positive | 0.63 | 0.56 | 0.60 |
| The average | 0.74 | 0.75 | 0.75 |

In general, it can be concluded that classical algorithms, using TF-IDF, do a good job with analyzing tonality, and the best results are shown by SVM. For use for this purpose Google Cloud Platform It is necessary to solve the problem of limiting the size of the text when translating and adjust the evaluation, since with the default range, most texts are rated as neutral. The results of the study are published in works [29-31].

# STUDY OF PECULIARITIES OF ASSESSING OPEN TEXTUAL INFORMATION SOURCES’ IMPACT ON SOCIETY IN THE KAZAKHSTANI CONTEXT.

# Legal regulation in the field of media monitoring

Kazakhstan legislation bounds the media to follow the principles of unbiased judgment, legality, reliability and respect of private live, honour and dignity of a person and a citizen.

Ministry of Information and Communications of the Republic of Kazakhstan (MIC RK) is an authorized body in the field of mass media. If some kind of information is found in the media space that can cause a negative reaction in the social environment, the MIC is obliged to inform CEO of the state body in a timely manner [32].

According to the Law of the Republic of Kazakhstan “On Mass Media”, state regulation of mass media is carried out by means of legal support, state control over the observance of specified Kazakhstan legislation. [33] MIC RK competence includes media monitoring. The media is monitored for compliance with Kazakhstan legislation. Monitoring is performed pursuant to the Media Monitoring Rules. These Rules set up media monitoring procedures. [34] Media monitoring is carried out with the use of Automated national information space monitoring information system, if the latter is available.

Monitoring of accessibility and compilation of information prohibited for distribution via telecommunications networks in Kazakhstan area is performed by an organization providing technical and methodical maintenance of monitoring, using the hardware and software infrastructure.

# Experience of monitoring of open textual information sources in Kazakhstan

Existing practice and conducted research in the media field cover only certain directions of media activities. In particular, state monitoring is carried out to determine compliance with media law. Republican State Enterprise Center for Information Analysis performs systemic media monitoring and impact assessment for compliance with Kazakhstan legislation. Major work to monitor and evaluate media influence is usually carried out across one specific aspect, for example, an assessment of the effectiveness of a certain political communication campaign, operations of a certain organisation, and other. Purpose of such assessment is mostly promotion of innovations and popularisation of governmental policy, and other.

Consulting services on the conduct of media monitoring and content-analysis are aimed only at assessing attitudes towards the company, persons and (or) brands, and are intended, to a greater extent, to promote innovations, services and products.

According to the database of National Center for State Scientific Technical Expertise JSC, 27 research works were registered within the period from 1995 to 2016 which are proximately connected to the topic thereunder. However, conducted researches are insufficient in number, cover only certain dimensions in studying the media role and have the only indirect reflection of the comprehensive assessment of media and open text information sources impact on society.

An expert approach is used as the main methodology for assessing media influence. Although organisations involved in media influence monitoring and evaluation have defined a set of supporting tools, such as the program to transcribe text from video and receiving data in text format, however, there are no automated tools for systematic monitoring and evaluation, particularly regarding the content analysis.

Automated media monitoring tools are widely used at the data collection stage, but automated tools are rarely used for data analysis or commercial purposes. Concerning media impact assessment, mainly the experts make data coding and analysis manually. In the context of the constant growth of the bulk of information, the issue of automation will become increasingly important.

## Sociological study on evaluation of open information sources' influence (electronic media) on the society of Kazakhstan

In order to determine the nature of the influence of open information sources (electronic media) on society, a sociological study was conducted to assess the impact of open information sources on Kazakhstani population. 3,200 respondents from 16 regions of Kazakhstan (14 regions and two cities of republican significance) participated in the survey. The survey covered all social groups in the country by gender (52.5% of men and 47.5% of women), age groups (15-19, 20-24, 25-34, 35-44, 45-54, 55 -64, 65 years and older), nationalities (61.8% Kazakhs, 26.8% Russians and 11.4% other nationalities) , language of communication, marital status, type of residence (56.7% city and 43.3% village), type of employment (28.8% are employees of the private sector, 17.5% are public sector employees, 13.1% are pensioners, 4.6% are entrepreneurs, and other) and well-being.

Social survey methods

*Goal:* to determine the nature of open information sources (electronic media) influence on society.

*Survey tasks:*

* 1. To define the main channels of news received by the audience;
  2. To define the social portrait of a typical reader of electronic media;
  3. To reveal the level of media-consumption by people, including electronic media;
  4. To assess the level of transparency in electronic media from the public point of view;
  5. To determine public trust level in media including electronic media;
  6. To reveal the level of reliability from the public point of view in covering the events by national mass media including electronic media;
  7. To define the level of e-media influence on public opinion.

*Survey method:* face-to-face questioning.

*Targeted audience:* Kazakhstani population ageing from 15 and older.

*Survey hypothesis applied:*

1. Economically active population prevails in general structure of consumers of electronic news media in Kazakhstan.
2. Population believes that media influences their vision of the situation, events, and phenomena.
3. Population believes that media objectively reflect the Kazakhstan reality.
4. The majority of the population believes that mass media are used as a tool for covering the events, areas of life, and other from a certain angle, beneficial to a particular circle of persons.
5. The majority of the population trusts the news web portals.
6. Operability and availability of news is not always supported by its reliability, which affects the credibility of internet sources.
7. Most electronic media often sacrifice of the accuracy in order to be the first to highlight an event to attract the audience.
8. Most of the population cannot distinguish qualitatively prepared journalistic material from journalistic material based on rumours and false information.

*Survey parameters:*

* Survey geography: 14 regions and two cities of republican significance.
* Survey sampling: representative to the population of Kazakhstan (15 years and older), equally filled across the regions and two cities of republican significance, the quota for gender and age for specific categories of population (without the quota for gender and age).
* Sampling scope: 3 200 respondents if estimated by 200 respondents per a region (Table 7).

Sampling scope n = 3200 satisfies the conditions of confidence coefficient to 95% with a confidence interval of ± 1.73%.

Sampling scope by regions n = 200 satisfies the conditions of confidence coefficient to 95% with a confidence interval of ± 6,93%.

Table 7 – Allocation of sampling across the regions and residence

| **Regions** | **Entire population** | | | **Sampling** | | |
| --- | --- | --- | --- | --- | --- | --- |
| **15 years and older** | **city** | **village** | **15 years and older** | **city** | **village** |
| Akmola region | 562 724 | 265 960 | 296 764 | 200 | 96 | 104 |
| Aktobe region | 614 002 | 387 463 | 226 539 | 200 | 126 | 74 |
| Almaty region | 1 388 296 | 332 703 | 1 055 593 | 200 | 48 | 152 |
| Atyrau region | 410 424 | 195 893 | 214 531 | 200 | 95 | 105 |
| West Kazakhstan region | 480 360 | 244 471 | 235 889 | 200 | 102 | 98 |
| Zhambyl region | 747 626 | 306 666 | 440 960 | 200 | 82 | 118 |
| Karaganda region | 1 060 927 | 847 986 | 212 941 | 200 | 160 | 40 |
| Kostanay region | 702 907 | 376 482 | 326 425 | 200 | 107 | 93 |
| Kyzylorda region | 519 813 | 227 410 | 292 403 | 200 | 87 | 113 |
| Mangystau region | 423 170 | 184 148 | 239 022 | 200 | 87 | 113 |
| South Kazakhstan region | 1 850 560 | 846 206 | 1 004 354 | 200 | 91 | 109 |
| Pavlodar region | 588 825 | 420 565 | 168 260 | 200 | 143 | 57 |
| North Kazakhstan region | 446 705 | 200 477 | 246 228 | 200 | 90 | 110 |
| East Kazakhstan region | 1 080 259 | 660 210 | 420 049 | 200 | 122 | 78 |
| Astana city | 704 246 | 704 246 | 0 | 200 | 200 | - |
| Almaty city | 1 374 944 | 1 374 944 | 0 | 200 | 200 | - |
| **TOTAL** | **12 955 788** | **7 575 830** | **5 379 958** | **3 200** | **1 843** | **1 836** |

Social and demographic structure of survey participants. Quality composition of respondents basing on social and demographic features shows that the study covered all social groups in the country by gender (*52.5% of men and 47.5% of women*), age groups (*15-19, 20-24, 25-34, 35-44, 45-54, 55 -64, 65 years and older*), nationalities (*61.8% Kazakhs, 26.8% Russians and 11.4% other nationalities*), language of communication, marital status, type of residence (*56.7% city and 43.3% village*), by type of employment (*28.8% are employees of the private sector, 17.5% are public sector employees, 13.1% are pensioners, 4.6% are entrepreneurs, and other*) and well-being.

Analysis of sources of news information consumption and the social portrait of a typical reader.

The results of the sociological study showed that the preferences of the survey participants when choosing the primary source of information with almost the same indicators were divided into three leading segments: social networks, bloggers (*30.3%*), television (*29.8%*) and news websites (*29.5%*). The share of respondents who prefer receiving information from printed publications was 6.2%. Other channels of information are less important.

For young people, social networks and bloggers are the primary sources of information if compared with other age categories. The survey revealed the dependence of social network use on the age of a reader: the younger the respondent, the more often social networks act as a primary source of information. Conversely, the older the respondent, the less often he/she uses network platforms (*social networks, blogs*) as a source of news. Least of all, rural residents spend their time on social media as well as the respondents of older age, most often are the students (*15-19 years old*) and city residents.

The opposite situation occurs regarding television. The frequency of watching TV is directly proportional to the age of a respondent, i.e. the older a person, the more often he/she watches TV, and, conversely, the younger, the less.

Internet websites are mostly used by respondents aged 25-34 years and 35-44 years old with higher education.

Tentative survey hypothesis concerning the structure of consumers of electronic media of news is partially confirmed with the exception for social networks. Economically active population dominates among consumers of almost all types of electronic media. Consumers of social networks are primarily young people under the age of 24 and schoolchildren.

More than half of the respondents (*59%*) emphasise that when they are consuming the news, they only want to be aware and are limited to one, maximum two news about the event. Only every fifth respondent seeks to use an extended list of information sources while consuming news in order to learn in detail about an event in progress.

According to the survey results, the leaders in both popularity and share of the regular audience are the news portal nur.kz and the Internet portal mail.ru. 61.8% of Internet users indicated that they know the site nur.kz. The share of the regular audience of this site (those who regularly use the site) is 51.3%. Tengrinews.kz and Zakon.kz took third and fourth place. These sites are known to every third or fourth internet user (*32.8% and 28%*). The share of the constant audience of these sites (*those who regularly use the site*) is 24.1% and 16%, respectively. The popularity of 365info.kz, Azattyk radiosy and Caravan varies at level 10 percent. The share of the daily audience of these sites is 4-5% (Table 8).

Table 8 - Tell us, what Kazakhstani information/news/analytical websites do you know/Name all Internet sites that you have used at least once/Name those Internet sites from the listed ones that you use regularly? (% of the total number of respondents)

| **Source name** | **Awareness** | | **Using at least once** | **Using regularly** |
| --- | --- | --- | --- | --- |
| **Unaided awareness** | **Aided awareness** |
| nur.kz | **61,8%** | **27,9%** | **54,7%** | **51,3%** |
| tengrinews.kz | **32,8%** | **29,4%** | **33,2%** | **24,1%** |
| camonitor.kz | 5,3% | 10,7% | 5,8% | 2,6% |
| mail.ru | **57,3%** | **30,3%** | **52,8%** | **52,3%** |
| zakon.kz | **28,0%** | **36,3%** | **34,8%** | **16,0%** |
| vlast.kz | 6,3% | 16,4% | 7,8% | 3,1% |
| forbes.kz | 8,2% | 15,8% | 8,3% | 4,6% |
| azzatyq.org | 6,8% | 10,7% | 5,0% | 3,7% |
| caravan.kz | 9,0% | 15,9% | 9,1% | 4,0% |
| bnews.kz | 6,8% | 16,7% | 8,9% | 2,9% |
| [kapital.kz](https://twitter.com/intent/user?screen_name=Kapitalkz) | 4,9% | 12,4% | 5,2% | 1,5% |
| [365info.kz](https://365info.kz/) | 9,9% | 18,5% | 12,6% | 4,8% |
| time.kz | 6,0% | 18,7% | 8,7% | 2,8% |
| rikatv.kz | 2,5% | 2,0% | 2,3% | 2,3% |
| Not sure | 4,3% | 8,7% | 5,1% | 6,0% |

The socio-demographic portrait of active users of news sites is given in Table 9.

Table 9 - Socio-demographic portrait of active users of information sites (% of the number of using the site regularly)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Mail.ru (52%)** | **Nur.kz (51%)** | **Tengrinews.kz (24%)** | **Zakon.kz (16%)** | **365info.kz (5%)** | **Azattyk radiosy (5%)** |
| **Gender** | Female | Of no importance | Of no importance | Male | Of no importance | Of no importance |
| **Age** | 15-34 | 20-34 | 25-34 | 35-44;  45-54 | 55-64;  25-34 | 45-54;  20-24 |
| **Ethnicity** | Othetr ethnic groups | Kazakhs | Kazakhs | Kazakhs and Russians | Kazakhs and other ethnic groups | Kazakhs |
| **Education** | Incomplete secondary  and higher | Higher | Higher | Higher | Higher | Secondary general education |
| **Income** | Average and below average income | Higher than average income | High level of income | Average and higher than average income | High level of income | Higher than average income |
| **Language of communication** | Russian | Kazakh, Kazakh  and Russian | Kazakh and Russian | Kazakh and Russian | Kazakh and Russian | Kazakh |
| **Place of residence** | Village | Of no importance | City | City | City | Village |

Reliability in covering events and the level of public confidence in Kazakhstan's news information sources. The survey results show that television continues to be the most trusted channel for information dissemination. Moreover, this source was most often chosen as the most reliable and truthful (61%).

Also, according to the survey, the hypothesis was confirmed that, in general, the majority of Kazakhstanis trust the Internet news portals as information channels: 30% said they trust, and 53% rather trust. 12% of respondents do not trust Internet portals. Nevertheless, in terms of confidence in the population's perception, Internet sites rank second (35%) after television. Private conversations and social networks cause the least trust. To some extent, 20% and 30% of the total number of respondents do not trust these channels, respectively.

For a comparative analysis of the level of trust in information distribution channels, an index of trust channels has been calculated. When calculating the Index, an index calculation approach was used to equalize the groups [20]. The index of the level of trust in each distribution channel is calculated according to the following formula:

IRсR= (nR+R+0,5\*nR1±R-0,5\* nR2±R-nR-R)/(nR+R+nR1±R+ nR2±R+ nR-R+nR0R), (12)

where: nR+ R- the number of trusting people, nR+ R- the number of rather trusting people, nR+ R- the number of rather not trusting people, nR+ R- the number of not trusting people, nR+ R- the number of those who found it difficult to answer.

According to the calculation results, the indices can take values from 1 to -1, where 1 is absolute trust, and -1 is absolute distrust.

The highest trust index is noted in television, the least trust is observed in social networks (Figure 4).

Figure 4 – Trust Indices for Information Distribution Channels

A socio-demographic analysis of affirmative answers (“yes, I trust”) to the question of trust in the information dissemination channels showed that youth groups under 24 years of age trust more in new types of media. The older generation trusts traditional media: television, newspapers and radio (Table 10).

Table 10 - The main socio-demographic characteristics of the respondents who chose the option “yes, I trust” on the question of confidence in the channels of dissemination of information

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Features | Social networks | Internet sites | Newspapers | Television | Radio | Chats |
| Gender | everybody | everybody | everybody | female | everybody | everybody |
| Age | 15-19;  20-24 | 15-19;  20-24 | 65 and older | 55-64; 65 and older | 20-24; 65 and older | 15-19 |
| Ethnicity | Kazakhs | Kazakhs | Kazakhs | Kazakhs and other ethnicities | Kazakhs | everybody |
| Education | Incomplete secondary;  secondary | everybody | everybody | everybody | everybody | Incomplete secondary |
| Income | Low; below the average; high | Below the average | Above average | Below the average; above average | Above average | Low |
| Place of residence | village | village | village | village | village | village |
| Region | Kostanay  South Kazakhstan  Almaty city  Almaty region  Mangistau and Kyzylorda | South Kazakhstan  Kostanay and Akmola  Almaty region  West Kazakhstan  Astana city | Atyrau  Kostanay  West Kazakhstan  Akmola  South Kazakhstan  Almaty region | South Kazakhstan  Atyrau  Almaty  Akmola  Kostanay | Atyrau, Almaty, Akmola, Kostanay | Akmola  South Kazakhstan  EKR |

The attitude takes an important role in the formation of an open attitude to the information sources that disseminate it. The main factors of trust respondents attributed to:

* information distribution channel (TV, newspapers, Internet sites);
* status of information source (national/oblast/city);
* presence of reasoned arguments and peer review.

According to the survey results, it was found that the urgency and availability of the transmitted information is not a guarantee of trust in the received messages. This confirms the assumption that the confidence in information is primarily based on an assessment of its reliability, and not of promptness. 39% believe the message from an official at the scene; one of four (29%) will believe the information that will be distributed via television or radio, and not by instant messengers or social networks.

For a comparative analysis of the level of trust in information distribution channels, an index of trust channels has been calculated. The Internet portal nur.kz takes the highest place in reliability among the population (Figure 5).

Figure5 - Indexes of trust to information Internet sites

Internet news portals are more trusted by a wide audience than websites that position themselves as informational and analytical. There is obvious difference in the audience of these sites. News portals, such as mail.ru or nur.kz, are more often trusted by young, not yet socially adapted respondents. The same sites, like Vlast or Forbes, in Kazakhstan are more trusted by established respondents, with education and higher level of income (Table 11).

Table 11 - The main socio-demographic characteristics of the respondents who chose the option “yes, I trust” on the question of confidence in the channels of dissemination of information, 1% of the respondents

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Features | mail.ru | nur.kz | tengrinews.kz | zakon.kz | Vlast | Forbes in Kazakhstan |
| Gender | everybody | everybody | everybody | everybody | male | everybody |
| Age | 15-19;  20-24 | 15-19;  20-24 | 20-24, 25-34 | 20-24 | 35-44 | 20-24 |
| Ethnicity | Other ethnicities | everybody | Kazakhs | everybody | everybody | everybody |
| Education | Incomplete secondary;  secondary | Higher, incomplete higher, secondary vocational | Higher, incomplete higher | Higher, incomplete higher | Higher, incomplete higher | Higher, incomplete higher |
| Income | Below the average; average | Below the average | Below the average; above average | Average; high | High | High |
| Place of residence | village | village | everybody | everybody | village | village |
| Region | Kostanay  South Kazakhstan  Akmola region | South Kazakhstan  Kostanay and Akmola  Almaty region  Astana city | Mangistau  Kostanay  Akmola  Astana city  Almaty region | Almaty region  Astana  Almaty  Akmola  Kostanay region | Akmola,  Almaty city | Akmola  Mangistau  Kostanay |

Internet sites, as an information channel, enjoy the unequivocal trust of 30% of respondents. And, nevertheless, every second (53%) said that he trusts Internet portals with the proviso “rather”, and 12% said that in one degree or another they do not trust.

One of the reasons for low confidence in information portals is the lack of practices using these sites, based on which positive or negative experience of using news websites can be formed. Under these conditions, it becomes important to understand how the official media have a counterweight in relation to social networks.

Survey results show that the overwhelming majority of Kazakhstanis (76.3%) trust more to the messages posted on websites of officially registered media than messages that different people and bloggers write on social networks (Facebook, Twitter, VKontakte, and other).

The same trend line can be traced in the level of trust in Kazakhstan's journalism as in relation to social networks. The difference in the level of trust in journalism is only 9 points higher than the level of trust in social networks (a total of 52.2% against 43.3%).

According to the survey, one of the reasons for distrust of information lies in the dissemination of false news, which, according to respondents' estimates, is a frequent phenomenon not only in social networks but also on sites of officially registered media. At the same time, the share of those who believe that on social networks rumours and false information is published 1.5 times more than the same indicator for registered media (38% and 21%, respectively)).

The hypothesis was confirmed that the majority of the population cannot distinguish qualitatively prepared journalistic material from journalistic material based on rumours and false information (59%). Moreover, the criteria that the audience uses to recognize fake news are extremely vague, and they are not always consistent.

On the issue of the need for state regulation of published information in social networks, there is no sustained opinion: 44% support, 25% oppose, 31% not sure, i.e. the scales can at any time swing in the other direction.

The main arguments for the need to introduce state regulation on the part of survey participants were the following:

* guarantee of receiving true, reliable information (31%);
* provision of state control over the dissemination of information (27%).

And arguments against are: violation of citizens' rights to freedom of speech (29%), violation of democratic principles, threat of dictatorship (23%) and reduced access to information (8%).

It should be noted that if state regulation causes ambiguous assessments, then the majority of respondents (67%) would like to get a tool to assess the reliability of the publication. The idea of introducing a measure to assess the confidence of journalists and editors caused less support (63%).

About 50% of respondents made suggestions on the reliability criteria, but their quality, as well as the percentage, are insignificant. The most frequently mentioned criteria are the presence of argumentation, factual information (13%), comparison with information from other sources (7%), presence of reaction, resonance in society (7%).

Unbiasedness of events presented by Kazakhstan's news information sources

In the course of the survey, respondents were asked to assess how accurately Kazakhstani news websites cover events in the country. The survey results show that none of the areas received a rating above 4 points (on a 5-point scale). The overall assessment of the level of information coverage is 3.36 points. It is noticeable that positive ratings are more typical of rural residents (3.46 points); respondents with low incomes (3.45 points) and Kazakhs (3.4 points).

The assessment of the quality of media coverage of the most important areas of society showed that the highest priority was given to activities in the “emergency situations” (3.56 points). The topic of labour relations and employment received the lowest mark among the other - 3.17 points.

According to the study results, the majority of respondents satisfied with the amount of information about the activities of state bodies that they receive from the media. Every fourth respondent complained about the insufficiency of such information. 67% of respondents believe that information on the activities of state bodies is displayed reliably and truthfully. Every fourth respondent noted that the information on the activities of government bodies is biased.

The survey results revealed an insufficiently high level of trust in media. Only 46% of respondents reported that the activity of Kazakhstan's Internet sources is transparent. One-third of the respondents do not consider Kazakhstanis internet sources to be transparent.

The majority of respondents (68%) believe that Kazakhstan's Internet sources provide accurate and truthful information when covering events in the country. At the same time, 23% of respondents expressed no confidence in Kazakhstan's Internet sources.

As the analysis of the data shows, the main shortcomings of the domestic Internet media include an excess of information on topics of no interest (28.8%), hiding specific facts and events (27.8%) and a lack of information on topics of interest (27.6%). In addition, 26.1% of respondents are concerned about the lack of comprehensive information on certain events. Every fifth respondent drew attention to the fact that in the news from Internet sources there is a biased interpretation of the facts and unreliable information. Additional analysis of the results shows that the Kazakh-speaking internet audience suffers from a lack of information on topics of interest, a lack of comprehensive information on events in the country and a lack of efficiency in obtaining information from internet sources.

During the survey, the majority (65%) of the respondents partially or fully agreed with the opinion that, using the information only from Kazakhstan internet sources, one can get an unbiased view of the events taking place in the country. Thus, it can be said that the population believes that the media accurately reflect the Kazakhstani reality.

The survey results partially confirmed the hypothesis that most electronic media often sacrifice accuracy in order to be the first to highlight an event to attract an audience. Almost half of the survey participants agree with the statement that, in practice, the speed of presenting information prevails over its accuracy and reliability. Respondents support the idea that Internet media inform the public as well as they are concerned with attracting a broader audience.

A partial confirmation received the hypothesis that the majority of Kazakhstanis tend to think that informational Internet sites are used as a tool for presenting events and some spheres at an angle that is beneficial for achieving certain goals. The share of those who support this thesis varies from 40 to 50%, which indicates the prevalence of opinion, but not its dominance. More often, respondents agree that Internet sites are used to present events from an angle that benefits a certain circle of people (53%), as well as to form a positive image of a particular subject (55%). There is less support for claims that Internet sites are engaged in distracting the audience from serious problems (47%) or imposing a certain opinion on readers (45%). More rarely, information that discredits a subject (43%) is disseminated through informational Internet sites. The complementary position is most often taken by respondents aged 20-24 years and 25-34 years who have a higher education. At the same time, almost a significant part – from 25 to 35% – believes that information websites are not tools for manipulating public opinion.

*The influence of internet information sources on society.* Some features were identified as part of studying the reflective communication of respondents in electronic media. Thus, an almost equal number of respondents (48% and 50%) read and do not read comments on publications respectively. The interest in comments is influenced mainly by two characteristics of the respondents: age and education. Thus, in young groups, the number of people reading comments is higher than in older ones, and vice versa. The higher the educational level of the respondents, the more often they read the comments. The majority of respondents almost do not write comments on texts in internet sources (57% never, 22% rarely). Only one-fifth of the respondents admitted that they carried out such activity (5% - often, 15% - sometimes). The study did not reveal the key reason why respondents write comments on publications. At the same time, it can be noted that negative reasons (disagreement with the author of the publication and comments) are more common as a reason for commenting than positive. The majority of respondents (57%) called the most attractive aspect of information dissemination that the internet has the opportunity not only to get acquainted with the news but also to find out how other users perceive certain events. Somewhat less attractive, but weighty enough (44%) is such an aspect as the opportunity through the Internet to learn information from the direct participants of the events.

The ability of the Internet to help avoid responsibility and ensure anonymity for users and even certain impunity is an attractive side for 16% and 14% of respondents, respectively. Most often, these aspects were noted by respondents aged 15-19 years; men; Kazakhs; with secondary and vocational education.

The majority of respondents (60%) almost never and rarely share news, opinions of people or other informational messages, send them through social networks. Only 15% of respondents can be attributed to active internet users who share or send various kinds of information through social networks. More often it is respondents between the ages of 15 and 34; having an incomplete secondary education (schoolchildren); Kazakhs; representatives of the extreme income, the least and most wealthy, groups.

The study did not reveal the essential motive, because of which the main part of the respondents would have made a certain response to publication on the internet. At the same time, a third of the respondents (37%) would like to share information that they considered important and interesting with a wide range of readers; a quarter of respondents (25%) were concerned about the content of the publication and were active in response; one fifth (20%) did not agree with the point of view of the publication's author and therefore shared information. Only 7% of respondents would follow the calls of authors of publications.

The hypothesis was confirmed that, in the population's opinion, the media influence their view of the situation, events, and phenomena. According to the survey, the majority of respondents (58%) acknowledged the influence of the media on their vision of the situation in the country. At the same time, every third (33%) denies such influence. Thus, we can distinguish three conditional groups of respondents: “Influenced by the media”, “not influenced by the media”, “difficult to assess the influence of the media”, differing in their socio-demographic characteristics.

60% of respondents agree with the statement that Kazakhstan's internet sources are a tool for shaping public opinion. Every fourth (27%) disagrees with this statement. At the same time, the number of respondents who agreed with the opinion about the presence of influence of domestic internet sources is influenced to a greater degree by the social characteristics of respondents (financial position, the language of communication, level of education), and not demographic (age).

However, the nature and intensity of this influence from the point of view of consumption of Internet sources, as the survey has shown, are not apparent. So, only 35% of respondents often and very often discuss what they see/hear/read in Internet sources with their surroundings. Other 35% of respondents do this rarely and very rarely. At the same time, 1/3 of the respondents found it difficult to answer.

The respondents from Kazakhstan are not inclined to perform any actions following the results of publications in the media: 53% of respondents said that very rarely and rarely they will take certain actions, another 33% found it difficult to answer the question. Only 14% of respondents said that they often and very often take any action caused by publications in the media. The most important feature of this group is age: Respondents to the group aged 15-24 responded positively to this question. At the same time, the place of residence, gender, ethnicity of the respondents do not have a significant influence on the answers.

The majority of the surveyed population (52%) could not or did not want to answer the question about the consequences of a biased reflection of information in Internet sources. At the same time, the responses received from 48% of respondents show that they most often predict direct (at the call-response level) negative-emotional and protest reactions of the population (loss of confidence, rallies and protests, chaos and panic, unpredictable and sad, scandals). At the same time, rational, emotionally unstained answers are relatively few.

Thus, according to the study results, the goals and objectives set have fully been achieved. During the sociological research the following results were obtained:

1. The main channels for the public to receive news information were identified.
2. The socio-demographic portrait of the average reader of information portals has been defined.
3. The behavioural features of the population at media consumption of information have been determined.
4. The results of assessing the public opinion on the level of unbiased and reliability of information in covering events in the country and the level of trust in Kazakhstan news information sources have been obtained.
5. Significant aspects of the influence of electronic media on public opinion have been considered.
6. According to the sociological study, confidence indices for information distribution channels, an index of trust for information Internet sites, an index of objectivity of Internet information sources, and an index of transparency of information Internet sources have been identified.

# CREATION OF NECESSARY TECHNICAL AND EXPERT-ANALYTICAL CONDITIONS FOR DEVELOPMENT OF THE INFORMATION SYSTEM OF EVALUATING OPEN TEXT INFORMATION SOURCES' INFLUENCE ON SOCIETY

## Analysis of existing hardware and software capabilities implementing text classification systems, including the analysis of word processing cloud platforms in natural language, as well as analysis and evaluation of cloud services’ quality in a natural language text processing

PaaS is a service that provides remote access to the application development platform using cloud computing. The study analyzed the capabilities of the following platforms: PaperSpace, FloydHub, Colab Research Google, Microsoft Azure Machine learning Studio, Google Cloud Platform.

Studies have led to the conclusion that compared to other platforms PaperSpace and FloydHub have a large number of pre-installed software and tools. While large platforms from Microsoft and Google requires some effort when setting up the development environment, platforms PaperSpace and FloydHub already include a pre-installed set of tools and services.

Table 12 –Compare cloud platforms in terms of parameters

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Services (PaaS) | Notebook  (Jupyter) | Pyhton  libraries | Pricing model | Documentation | Virtual  Machine | TPU  usage | Auto API  Deployment |
| PaperSpace | yes | yes | hourly or monthly fee  + flat monthly storage fee  + optional monthly subscription | yes | yes | yes | yes |
| FloydHub | yes | yes | hourly fee for cpu, gpu.  monthly fee for storage,  subscription. | yes | no | no | no |
| Colab Research Google | yes | yes | free | yes | no | no | no |
| Microsoft Azure Machine learning Studio | yes | yes | subscription | yes | no | no | yes |
| Google Cloud Platform | yes | no | hourly or monthly fee | yes | yes | yes | yes |

Table 13 –Pre-installed software and tools in cloud platforms

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | PaperSpace | FloydHub | Colab Research  Google | Microsoft Azure Machine learning  Studio | Google  Cloud  Platform |
| Notebook Jupyter | yes | yes | yes | yes | no |
| Python libraries (pip) | yes | yes | yes | yes | yes |
| Virtual machines | yes | no | no | no | no |
| Language R | yes | no | no | yes | no |
| Scikit-Learn | no | yes | yes | no | no |
| Tensorflow | yes | yes | yes | no | yes |
| Keras | no | yes | no | no | no |

SaaS, provide customized software deployed on servers. Customers using SaaS do not need to maintain software and allocate their computing power.

The study examined the possibility of the following systems*:* TextRazor, Rosette Text Analytics, Eureka Engine, Cloud Natural Language, Texterra, Pullenti, NER-ru, UDPipe, AOT.

Table 14 – Service Characteristics

|  |  |  |  |
| --- | --- | --- | --- |
|  | TextRazor | Rosette Text Analytics | Eureka Engine |
| 1 | 2 | 3 | 4 |
| Part-of-speech tagging | For Russian is not available, for English available | correctly markered words in 3 text  152 from 157  179 from 188  173 from 180 | - |
| Tokenization | Available | Available | Available |
| Stemming | Available | - | - |
| Lemmatization | For Russian is not available, for English available | correct lemmatization  152 from 157  182 from 188  176 from 180 | - |
| Sentiment analysis | Not available | Not available | Negative 9/10 Positive 10/10 |
| Categorization | 5 from 5 | For Russian is not, for English - 0 from 5 | 3 from 5 |
| Price policy | Demo, 6000 API requests - 200$ in month | Demo, 40 000 API requests - 100$ in month | Demo, 50 000 documents - 800 euro in month |
| Availability of demo | Available | Available | Available |
| Number of languages supported | 11 | 20 recognized names. noun 32 for part of markup | 1  (rus) |
| Uniqueness | Entities with links to FreeBase and Wikipedia | Stable Part-of-speech tagging marking and lemmatization | The presence of tonal markup text |

Continuation of Table 14

|  |  |  |  |
| --- | --- | --- | --- |
|  | TextRazor | Rosette Text Analytics | Eureka Engine |
| 1 | 5 | 6 | 7 |
| Part-of-speech tagging | Available  rus - syntagrus  eng. - Penn tree bank | Available  Has: word options, number, gender, case | - |
| Tokenization | Available | Available | - |
| Stemming | - | - | - |
| Lemmatization | Available | Available | - |
| Opinion mining | Available | - | - |
| Categorization / topic highlighting | - | - | - |
| Price policy | Free | Free for non-commercial use. | - |
| Availability of demo | Available | Available | Available |
| Number of languages supported | 2  (rus, eng) | 3  (rus, ukr, eng) | 1  (rus) |
| Uniqueness | Corpus-based morphological analysis SinTagRus | Building semantic links in the form of object graphs | Minimalism |

Continuation of Table 14

|  |  |  |
| --- | --- | --- |
|  | UDPipe | AOT |
| 1 | 8 | 9 |
| Part-of-speech tagging | Available  Based on UD Treebanks | Available  Based on dictionary by A.A. Zaliznyak |
| Tokenization | Available | Available |
| Stemming | - | - |
| Lemmatization | Available | Available |
| Opinion mining | Not available | Not available |
| Categorization | - | - |
| Price policy | Free for non-commercial use. | Free |
| Availability of demo | Not available | Not available |
| Number of languages supported | 74 | 3 (rus, en, ger) |
| Uniqueness | Building parsing sentences | Parsing sentences |

Accuracy, Precision, Recall and harmonic mean F1, which are commonly used to evaluate data classification algorithms and programs, are widely used to quantify the recognition of named entities. The results of the analysis of services for the recognition of named entities are shown in Table 15 and Figure 8:

Table 15 - Results of the services analysis on recognition of named entities

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Services | The average | | | |
| A | P | R | F1 |
| TextRazor | 0,48 | 0,59 | 0,75 | 0,64 |
| Rosette Text Analytics | 0,66 | 0,74 | 0,86 | 0,78 |
| Eureka Engine | 0,86 | 0,98 | 0,88 | 0,92 |
| Texterra | 0,65 | 0,85 | 0,73 | 0,77 |
| Pullenti | 0,63 | 0,73 | 0,75 | 0,74 |
| NER-ru | 0,67 | 0,85 | 0,79 | 0,80 |

Figure 8 - Comparative characteristics of cloud services by the quality of entity recognition

Therefore, it can be concluded that when solving problems of OJETT in Russian, it is better to use the Eureka Engine to recognize named entities and determine the tonality of the text and for part-of-speech tagging text markup is best to use Rosette Text Analytics.

Preliminary results of a study published in the [36-37].

## Research and development of methods for clustering textual information based on its semantic content.

News articles published by the media in Russian are chosen as a scope. In the future, the developed approaches will be applied to news articles in the Kazakh language. A comparative analysis of the developed approaches application results to both languages was carried out to determine the similarity of results in both cases. The developed methodology is implemented using software tools.

Methodology development. The review of modern advanced methods of grouping textual information in natural languages (mostly the Russian and English languages) on the basis of its semantic contents by means of clustering methods was carried out. For example, the review of text clustering algorithms by 2012 is provided in chapter 4 [2]. These algorithms are based on various representation of text information, for example, by means of SVD decomposition, n-gramm, on use of special corpora [3] or the electronic semantic networks as WordNet and different types of clustering algorithms. Since 2013 model based on neural networks got wide spread as the showing best results at comparable work expenses. These models were the named as of distributive semantics models [4], [5], [6]. Since 1998 with a different time frequency (in recent years annually) the seminar competition on computational linguistics of Semantic Evaluation (SemEval) is conducted where often one of tasks put before participants is development of methods evaluating texts similarity. The efficiency of newly developed methods increases each time, but mostly all of researches is given only to application of the developed methods to English as to the most widespread. The Russian and Kazakh languages differ in a big variety and richness of a form, not the strict (fixed) word order that brings additional difficulties. Perhaps, methods, successful in case of English, will not show the same result in respect to the Russian and Kazakh languages. The methodology which directly solves the above described problem concerning publications in Russian (or Kazakh) and which not only would classify publications by subjects, but also would be capable to distinguish the changes in texts made by addition or replacement of words, modifying sense of publications, is absent (at least in open sources). Thus, the task comes down to independent development on the basis of the existing well-known research and development in the field of natural language processing of the method of grouping of textual information which is quantitatively defining semantic similarity of text documents in the Russian and Kazakh languages. At the moment a number of potentially perspective approaches regarding their applicability to the above task is considered. In case of successful application of several approaches the best approach will be determined. Efficient approach has to detect similarity of texts, similar in semantic contents, separate from each other texts which, on the contrary, differ in sense and to reveal unique information (outliers). In the considered approaches various ways of text information representation are used as various clustering algorithms with variously chosen metrics.

As a scope of approaches news articles are chosen as they are characterized by formal official style, the steady fixed dictionary, a briefness and succinctness of the provided information, are subject to less level of authors subjectivity, the coreference problem is less relevant. As approaches in regard to the Kazakhstan media space are considered, articles are chosen in Russian language because of availability of data, news articles in Kazakh language will be also considered later (if necessary base will be available). The analysis about similarity of results of application of the developed approaches to both languages will be carried out. Subsequently, provided the proved applicability of the developed approaches concerning news articles, it is possible to consider a question of success of application them to more general view of text information.

Textual information representation. News article may contain from one to several tens sentences depending on media resource and editorial policy to which they adhere concerning a type of the published information. In turn each sentence can consist of several words to several tens of words where sentences can be simple, compound or complex. Despite the dynamic volume of words, various stylistics, various ways of information delivering, it is required to present news articles in the transformed form which preserves information containing in article, that is, keeps sense of contents and would allow to apply to it clustering methods to group articles according to contents. At the same time representation has to be sensitive to change of contents of the text which modifies or corrects sense of contents. For example, in case of addition of negotiation part "not" to a predicate, or the predicate is replaced with word opposite in sense, then representation before and after has to be various enough for change detection. Also, if the text changes by addition or removal of additional information or additional facts, in particular, of dates, persons, the place of an event, then representation before and after has to reflect the made changes and extent of the happened changes.

Text representation when applying clustering methods has to support formation of clusters with similar semantic meaning, at the same time texts representations arranged in the same cluster which are more remote from the center will be characterized by more distinguished information content in comparison with average cluster information content; thereby, the further from the center of a cluster the less characteristic information would be in comparison with cluster average. In addition, successful text representation for a clustering will provide an opportunity to find unique information (outliers) by means of determination of remoteness of texts location from centroid of formed clusters.

Text information representation: vector representation. Models of distributive semantics correspond to each word its vector representation which differs from counting methods where length of a vector of a word is equal to the size of the dictionary and can reach more than one million, makes value in practice of not surpassing 1000 (generally in practice the dimension is equal to 300 or 500) that considerably simplifies computing processes. The neural network which layer is used as vector representation of words is the cornerstone of these models. These models bade on the distributional hypothesis [8] claiming that the sense of a word is defined by contexts in which this word is used. Models of distributive semantics provide one-orientation of vector representations of semantic similar words, thereby allow to define vectors of semantic similar words to a vector of the chosen word through distance function – cosine similarity. The predicative model of the distributive semantics which was widely adopted since 2013 is called Word2Vec and includes two algorithms: Skip-Gram and CBOW (Continuous Bag of Words) [9], Later at the Stanford university in 2014 modification of this model called GloVe (Global Vectors for Word Representation) [10] was developed. It differs in principle of construction, based on the frequency of words in the dictionary. The disadvantage of these models is that they do not provide vector representation for all words (as in case of rare words or absent in the training corpus) and that vector representation of words is contextually independent that is if words have several semantic values and are respectively used in several various contexts, then vector representation is result of averaging of these contexts. Due to these shortcomings advanced models were offered: fastText (2017) [11] and ELMo model (Embeddings from Language models, 2018) [12]. The fastText model uses while training neural network all words n-gram instead of only a single whole word as in a case with Word2Vec, thus, this model provides vector representation of not only word, but also its derivatives. The ELMo model is capable to provide vector representation of words depending on the used context. In difference from Word2Vec, it is based on other type of neural network - bidirectional LSTM.

Besides models of vector representation of words, models of vector representation of the whole texts (documents) were also developed. One of such model is the modified Word2Vec called by Doc2Vec where vector representation of the document is the additional unique vector for each document and constructed by adding vector to each document in addition to vector to each word while training neural network. There exists two approaches: PV-DM (Distributed Memory version of Paragraph Vector) and PV-DBOW (Distributed Bag of Words version of Paragraph Vector) [4].

As it was already noted, the above-stated models of distributive semantics present words or the entire documents in the form of vectors which have property of same orientation of semantic similarities to each other words or documents. Thereby, the attempt of development of methods of classification of news information on the basis of these models is potentially perspective.

Text information representation: “factual cube”. Other potentially perspective way of news information representation to which clustering methods can be applied is the extraction of a certain structure from sentences which in the best way reflects semantic contents of them. At such representation news article will consist of a set of such structures, at the same time cardinality of structure directly depends on number of sentences in article. As structure the following set is considered: the part modifying a verb (adverb) + a predicate, a subject, object, the place, time (or the part modifying a verb + a predicate, a subject, object, all remained nouns which describe the place, time, faces, etc. and named entities). We will call this structure the fact, and a set of structures - the facts. Each separate fact or each point corresponding to a set of the facts of article is possible to arrange in multidimensional space where each axis will correspond to a certain projection of the fact, for example, to a subject. The axis assumes existence of one to one correspondence between values of an axis and possible values of a projection of the fact. All facts or points of the considered corpus arranged in multidimensional space will be called "factual cube". We will apply clustering methods to the constructed "factual cube". As a result provided success of this text information representation, articles with similar contents will be close to each other and are arranged in one cluster, articles located on a distance depending on degree of remoteness will be characterized by certain level of differences in contents.

Approaches to text grouping based on its content.

The first approach. The first approach to represent textual information and the subsequent its clustering is based on model of the distributive semantics which determines to each word its vector representation. Use a set of the vectors representing words constituting news article as vector representation of the text is the natural idea. For example, it is possible to take the average value of vector representations of words as vector representation of the text. At the same time, it is possible to consider several options of that of what words of article vector representation to use in the average calculation and what weights are more appropriate, that is, what selection of a set of words of the text will be the most optimal for the above-stated purposes.

Options of words:

* 1. all words of the text;
  2. all words of the text except for words, not bearing semantic value, so-called stopwords which include conjunctions, preзpositions, all types of pronouns, etc.;
  3. nouns and named entities;
  4. the facts of "factual cube" (the part modifying a verb + a predicate, a subject, object, all remained names of nouns which describe the place, time, faces and named entities).

Options of weights:

* 1. equal weights are assigned to each word, thus, weight is proportional to word frequency in sentence;
  2. equal weights, but the word frequency excluded as the factor which influence value of weight because high-frequency words might not bear the unique semantic value characterizing text contents;
  3. values of scales are taken inversely proportional to word frequency in the dictionary of the case, to those more rare words as bearing unique information, the greater influence in calculation is given.

Option of models:

1. Word2Vec;
2. GloVe;
3. FastText;
4. ELMo

After vector representation of news articles is obtained, it is possible to consider application of clustering methods, where as a metric let’s consider cosine similarity or Euclidean distance. As a clustering algorithm let’s choose an algorithm which can determine number of clusters. After number of clusters is defined, the same clustering algorithm or more effective clustering algorithm but which requires number of clusters as a parameter needs can be applied.

Vector representation of entire article is a process of averaging of a large number of information, at the same time the uniqueness of some content which is required when determining uniqueness (outliers) is perhaps lost. Therefore after successful top level classification of articles by its contents, for example, to thematic, is achieved, it is necessary to consider process of clustering of a set of simple sentences (simple sentences – the most minimum atomic unit which bears certain unique information) of all articles of a concrete cluster (in a case with compound sentences it is necessary to decompose them into set of simple sentences with coreference resolution; in a case of news a similar problem is not very relevant). This likely can provide sensitivity to content modification. Cluster centroids will indicate the general average information in a given cluster, remote from cluster centroids sentences will be those simple ones that are not typical and unique for a cluster.

As well as at the first stage, a clustering algorithm will be applied to identify number of clusters and then, having defined its, the same clustering algorithm or another one which needs as a parameter number of clusters in advance, but being more effective, will be applied. It will be necessary to carry out the analysis of application of several clustering algorithms to find out optimal one.

It would be desirable if vector representation of antonyms (words opposite on sense, including verbs) would satisfy to property of orthogonality. In that case the cosine similarity of simple sentences in which content of one sentence is changed to opposite (for example, by negotiating of a verb), would be equal to zero. However, methods of distributive semantics do not provide this desirable property as antonyms usually meet in the same context and the cosine similarity is not close to zero. It will be required to analyze the received clustering results on how the lack of this property affects clustering process of news articles.

The second approach. The second approach also is based on distributed semantics, but it uses Doc2Vec model, namely its PV-DM variant. The reason to choose this approach is that context is important for problem and PV-DM uses surroundings of each word to build numerical representation of document. The model represents each article as a vector which will be used in clustering where cosine similarity or Euclidian distance will be used as a metric. To succeed Doc2Vec model should group similar articles by their content. Then it is necessary to analyze sensitivity of numerical representation of document to changes in its content. If it happens that model is not sensitive to changes, additional stage of clustering will be considered. Based on Doc2Vec model every alone sentence will be assigned its vector representation and clustering algorithms will be applied to the vectors that constitute same cluster. Consequently, analysis of how well Doc2Vec representations facilitate clustering process on content similarity and of how its representation on sensitivity it to changes.

The third approach. This approach is based on the concept of "factual cube". On the basis of the corpus, a “factual cube” is created in n-dimensional space (in our case, n is five), where each axis corresponds to a certain coordinate of the fact (for example, predicate). A point in space represents one fact. To place them, it is necessary to observe the order of words along each axis while preserving the semantic interrelation of words, that is, closer words on the axis should be semantically closer than those that are distant from each other. For this task Word2Vec model is used, which provides the property that the vector representation of semantically related words is unidirectional and present the words as an m-dimensional vector (in practice, usually 300-dimensional). A vector representation of words related to a certain coordinate of a fact (for example, a predicate) will allow applying clustering methods in m-dimensional vector space, for example, applying the k-means algorithm. As a result of clustering we obtain a set of clusters and define center of clusters as . Then the clusters are placed on the axis depending on their distance from the center . Further, all words are placed on the axis. After presenting all facts in n dimensional space, we apply clustering methods to identify similar groups. Each article is characterized by a set of facts, then the article will correspond to several educated clusters. Thus, it will be necessary to develop a metric that determines the closeness of articles depending on the clusters associated with them and on their distance from cluster centers. Finally, we apply the final clustering process.

The fourth approach. This approach is also based on concept of “factual cube”. Because of possibility that clustering based only on factual cube will be not enough successful and the lack of one-to-one correspondence between point in factual cube and article, other features as named entities, numbers and dates presented in article are considered in this approach. We used modified Jaccard similarity coefficient:

(13)

This metric often used in searching similar documents, checking for plagiarism, etc.

Density-based spatial clustering of applications with noise (DBSCAN) and its modifications ((DBSCAN\*, HDBSCAN, K-DBSCAN, RDBSCAN) are chosen as base algorithm [50]. It is said that it has below described advantages and limitations [51]:

Advantages:

* DBSCAN does not require to specify number of clusters.
* DBSCAN can find arbitrarily shaped enough isolated clusters;
* (it is expected) DBSCAN can find small sized cluster comparatively good.
* DBSCAN has a notion of noise, and is robust to outliers.
* requires only two parameters;
* in most cases insensitive to the ordering of the points in the database.
* parallel realization is balanced by quantity and a type of the made operations (calculation of distance, comparison with ε).

Limitations:

* DBSCAN cannot cluster well data sets with large differences in densities, since parameters cannot then be chosen separately for each cluster density.
* border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data are processed.
* The quality of DBSCAN depends on the distance measure

While applying this type of algorithm to the set task, analysis of above stated properties of algorithm will be verified. Other clustering algorithms will be considered as well to compare their efficiency with DBSCAN.

Results. Syntactic/semantic analyzer. SyntaxNet - the syntactic/semantic analyzer of Russian-language texts was used for creation of "factual cube" [52]. SyntaxNet is the open public library founded on Tensorflow. SyntaxNet is based on the neural network trained on corpus (SynTagRus is called) of 1 107 135 tokens and 61 899 marked sentences (Universal Dependencies marking). This analyzer provides decomposition of sentence in words in the form of a text format – json. The realization of the program on Python for processing of json data structure to extract the facts from sentences in a desirable representation and named entities was implemented. The main disadvantage of SyntaxNet is relatively low speed of work processing.

*Data preprocessing.* Application of distributive models and also work with text data demands preliminary preprocessing of texts. For training of distributive models the following options of preprocessing of texts were performed (generally for Word2Vec and GloVe) :

* + tokenization and lemmatization of texts, with removal of high-frequency functional words.
  + the first option was complemented with assignment to a lemmatized form PoS-tag (part of speech) word labels - tag specifying a part of speech. Assignment of PoS-tag allows to receive separate vectors for words which are used both in the form of a verb and in the form of a noun (for example, to bank a noun and a verb).
  + the second option is complemented with merging of the modifying words/particles with a verb (adverbs). merging of the modifying parts/words with a verb allows to receive vector representation not only for a verb, but also, for example, for its negotiation.
  + the third option is complemented with application of an algorithm of coreference resolution as the pronouns which are deleted as stopwords bear in themselves information on contexts of objects to which pronouns refer.

For processing purpose programs on Python were realized. Several Python libraries were used: nltk, pymorphy2, syntaxNet, string.

Training of distributive semantics models. Models of distributive semantics were realized in Python (Word2Vec, Doc2Vec, fastText, ELMo). Training of these models was performed on the following corpora:

* + the Kazakhstan and Russian news (about 900 000 articles, in a proportion 2/9 and 7/9).
  + National corpus of Russian language.

Training (Word2Vec and GloVe) was carried out with a window size equal 10. Length of a vector was equal 300. The minimum frequency of occurrence is chosen to equal 3. The number of iterations (epochs) was taken equal 100.

Vector representation of named entities taken from the models trained on news corpus. Vector representation of other words is taken from the models trained at the National corpus of Russian language as it was revealed that more balanced corpora as the National corpora of Russian language provide more exact semantic identification in comparison with larger corpora, but less balanced and qualitative.

Realization of the first approach. Based on trained Word2Vec model vector representations for each news article from the test sample of 30 news was obtained. All options were considered. Two clustering stages were carried out: at first a clustering of vector representation of news was performed, then a clustering of the vectors represented each sentence in each cluster was performed. The algorithm Affirmation Propagation was chosen [15]. This algorithm detected the correct number of clusters on the sample test, then it was also applied in determination of clusters contents. As a result only two news from 30 were incorrectly assigned to clusters, the others were correctly.

This approach is applied to the data set of 900 000 news. Results of application of this approach to the given data set will be analyzed.

Realization of the second approach. After training of Doc2Vec model on normalized documents, we got numerical representation for each document. On the testing phase, one article from different resources but on the same topic.

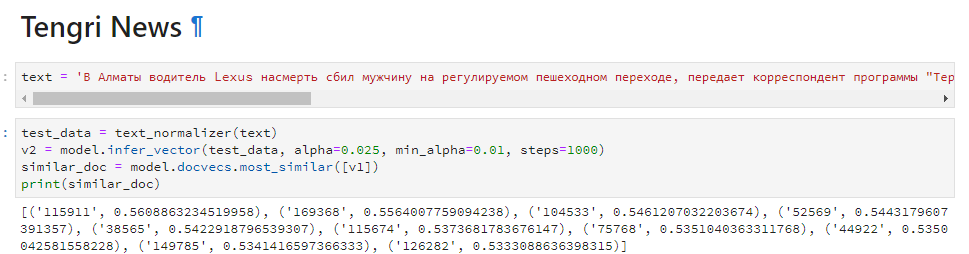
News from nur.kz:

*«The man who crossed the road on the crosswalk on the section of the road where Satpayeva passes into Shalyapin was hit by the Lexus ES300 car. The ambulancemen who arrived at the scene of the tragedy verified the death of the pedestrian. Police officers investigate the accident.»*

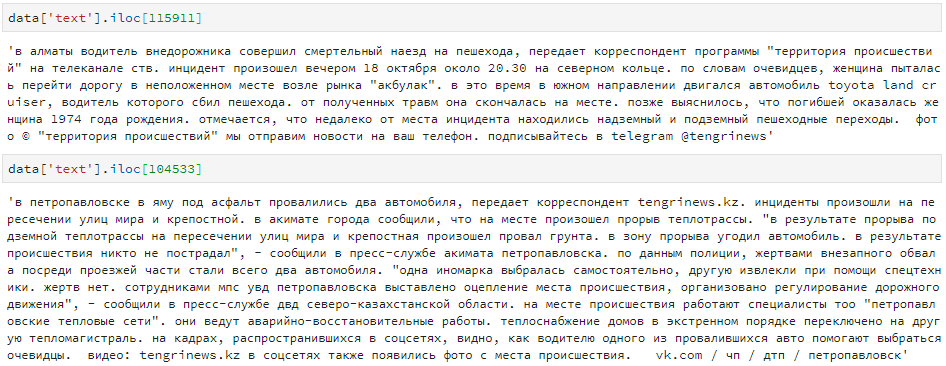
For this news, we discovered the most similar articles from corpus via cosine similarity (Figure 9).

Figure 9 - The most similar articles for «nur.kz»

The same algorithm we did for news from tengrinews.kz (Figure 10):

Figure 10 - The most similar articles for «tengrinews.kz»

It is obvious that both articles have the same set of the most similar articles. The articles is shown on figure 11.

Figure 11 - The most similar articles

For this experiment we changed a few facts in the article and then used cosine similarity to compute distance between numerical representation of changed and original articles. As a result, we discovered that changed article is closer to original than original to articles with the same meaning.

In this approach, representation of documents as a vector is not sensitive to changes in its content. Thus, we suggest to include stage where each sentence will be represented as a vector based on trained Doc2Vec model and then the vectors will be used for clustering. Here we will also analyze sensitivity and suitability to its content. We will use this approach for corpus of 900 000 news. Moreover, we will compare received results with their tags.

Realization of the third approach. For current stage of the work, a 2-dimensional “factual cube” (predicate and subject) has been analyzed. In figure 4, all facts from the corpus www.tengrinews.kz consisting of 200 thousand articles are shown in the form of points.

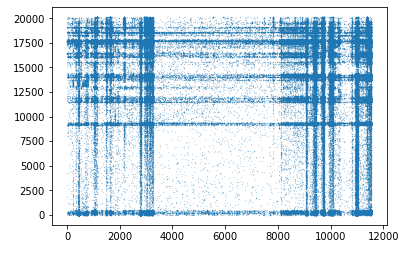


Figure 12 - A 2-dimensional “factual cube” (predicate and subject)

One can observe clumps of facts in the form of vertical and horizontal lines in Figure 12, which appear to represent frequently used predicates and subjects. This may indicate that in the articles, some clusters of words are used more often than others.

Figures 13 and 14 show the “factual cube” generated from articles with named objects “Golovkin” or “Alvarez”. In this case, the distribution of points in the fact cube is uniform.

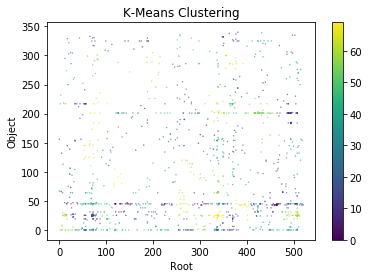


Figure 13 - Clusters of facts collected from articles where the names “Golovkin” and “Alvarez” are found

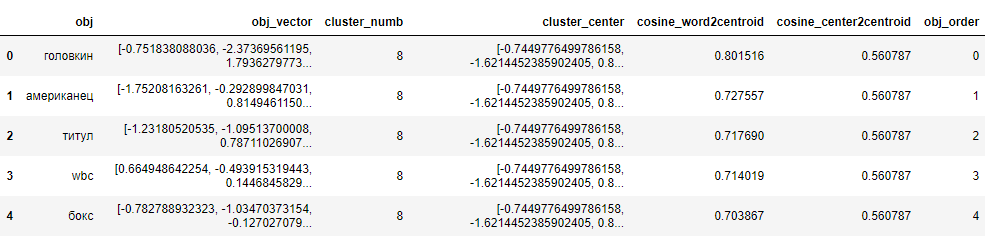


Figure 14 - An example of obtaining the axis “subject” of a factual cube from words that have a connection with the entities “Golovkin” and “Alvarez”

Future work will include construction of 5th dimensional factual cube of the news corps (900,000 news) and application of the discussed approach to it. The results will be analyzed in terms of the efficiency of this approach in relation to grouping articles according to their content.

Realization of the fourth approach. This approach on test sample with the size of 40 articles were tested. Density based algorithm DBSCAN with Jaccard metric shows satisfactory result (2/40 mistakes). Application of this algorithm requires manual parameters tuning (depending on density of topical (eventful) space), which might happen to be time consuming.

This approach will be applied on a data set with 900 000 articles in order to evaluate its performance on large dataset. Also, results of other algorithms and comparative analysis will be provided.

Conclusion. The complex of works on approbation of the offered algorithms on test samples of the average size is carried out. All offered algorithms showed acceptable qualitative results. The complex of works on the analysis of productivity of each of approaches on the case from 900 000 news articles is now carried out. The component analysis of each of approaches regarding improvements/optimization of work of its stages for ensuring excellent results will be in addition carried out. In the long term the possibility of development of the hybrid use of approaches combining the best of each approach which can show more optimal work will be possibly worked out. All program tools are implemented on Python

## Development of clustering algorithms

This algorithm will be included in the library of clustering algorithms for the corpus of media publications. Clustered objects can be both individual publications and the entire media, which are defined by a set of their numerical attributes. The resulting distribution of publications by clusters will be used to further multi-criteria group classification of publications and the subsequent rating formation of information sources, as well as in the task of social significance assessment.

A new clustering algorithm was developed and implemented in order to separate sets of objects into groups in according to similarity of their attributes based on the approach of maximizing the energy criterion. This algorithm belongs to the class of energy algorithms [53] or density [57] clustering. The clustering algorithm consists of two separate sub-algorithms. The first part is a parametric algorithm for determining the initial points (centroids) of future clusters. This algorithm has a number of adjustable parameters that indirectly affect the number of centroids obtained and their spatial distribution in . The algorithm was constructed in such a way that the number of obtained centroids was not specified initially, yet it was determined in the course of calculations. This is convenient when analyzing data, where the number of centroids is not known a priori, as, for example, when analyzing large arrays of textual publications.

The second part of the algorithm uses the centroids obtained in the first stage for the concurrent clustering growth. The principle of competitive growth of clusters due to absorption of boundary objects is used. The object is absorbed by the cluster or transmitted from the neighboring one if the maximum energy criterion of the clusters is satisfied. The cluster propagation stops when there is no object that is not distributed among the clusters, or an object which is transferred from one cluster to another leads to an increase in the total energy of the clusters. Objects that have not been grouped are classified as noise. The second part of the algorithm also has a set of parameters that affect the course of the clustering propagation process. In view of this, the parameters of the second part of the algorithm are also involved in the process of parametric identification, in which the accuracy criterion is maximized.

In the first part, we implement the algorithm that, for given points in the data set, finds a subset corresponding to points *sj* with the highest local energy:

(14)

where *d*(*i*,*j*) is the Euclidean distance between points *si* and *sj*; *R1* and *R2* customizable real parameters of the algorithm. The nonlinear heuristic function of the energy is determined by the parameter In this case, the centroid corresponding to *sj* is the most representative element for some cluster and the most suitable point to start the cluster growth process.  
As one of the variants of function, the following heuristic function was chosen:

(15)

where has a fixed value .

This approach requires the definition of a number of parameters: *R1*, *R2, R3, q.* We solve the problem of parameter identification by iterative search for the best input parameters: it turns to the problem of optimizing clustering accuracy. Accuracy is calculated by comparing clusters, which we compute using a known distribution across clusters from the corresponding source data set. In order to solve this problem, we encapsulate the clustering algorithm in a “black box”, so that its parameters are used as input data, and we obtain the resulting accuracy as an evaluation of quality criterion.

|  |  |  |
| --- | --- | --- |
| image  а) Initial clustering | image  b) clustering at the 50th step | image  c) stabilized clustering after step 125 |

Figure 15 - Example of clustering on the Iris data set

In the Figure 15 as an example on some steps of the clustering process are visualized. Table 1 shows the results of experimental testing of the developed algorithm on commonly used UCI data sets [53]. Table 17 shows the results of an experimental comparison of the accuracy of our algorithm with the algorithms of k-means, ELM k-means, and the Kernel algorithm [55-58].

Table 16 - The achieved clustering accuracy for different data sets from UCI [1] and the corresponding values of the algorithm parameters

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | Precision () |  |  |  |  |
| Iris | 0.9574 | 0.5300 | 1.0193 | 0.2388 | 0.9862 |
| WPBC | 0.9140 | 0.6989 | 1.6511 | 0.2192 | 0.9836 |
| Wine | 0.9575 | 0.4724 | 0.9173 | 0.2378 | 0.9907 |

Table 17 - Results of an experimental comparison of the accuracy of our algorithm with the k-means, ELM k-means, and Kernel algorithms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Our algorithm | k-means | ELM k-means | Kernel |
| Iris | 0.9574 | 0.891 | 0.960 | 0.947 |
| WPBC | 0.9140 | 0.961 | 0.979 | 0.970 |
| Wine | 0.9575 | 0.693 | 0.702 | 0.696 |

As can be seen from the above tables, the clustering accuracy of the developed algorithm on some data sets exceeds known algorithms or is comparable to them.

A more detailed description of the developed algorithm is given in the relevant publications prepared and published in the framework of this project [4,59-61]. Currently, work is underway on the application of this algorithm to the clustering of large arrays of textual information, as well as the analyzed media.

## Creation of an algorithm that maps multidimensional points to one dimensional single line by means of preserving their distances between them

The algorithm that maps multidimensional points to one dimensional single line mentioned will be used in different algorithm called “Semantic cube” where each dimension of the cube will have multidimensional points mapped on them. And the “Semantic cube” algorithm is going to be a clusterization algorithm for media publications.

We conducted test studies on applying some of the known algorithms of Traveling salesman and clustering algorithms to map multidimensional points on a single line.

The next is the description of our problem. If we take a set of points:

P = {p1, p2 … pn}

in N - dimensional space

P ∈ Rn

where we need to spread them across one dimensional line so that it would satisfy the next criteria

*(a)* Best permutation = Min( TotalDist(P) | P ∈ S(P) ) (16)

where P – is permutation of points, S(P) – is the set of all possible permutations of points, TotalDist – is the sum of distances between points in permutation:

TotalDist(P) = (17)

and norm\_dist – is the normalized distance between two points, n – is the count of points and arg\_max{norm\_dist} – is the possible maximum distance between two points.

norm\_dist() = (18)

norm\_dist ∈ [0, 1], dist – is the distance between two points.

And in order to measure how close is any given permutation of points to the best permutation of points, we used the following metric formula:

Error = (19)

Error ∈ [0,1], – is the permutation of points accessed and Error – is the error rate that shows the quality of (e.g. how close is) permutation to the best permutation. So, Error = 0 means that our permutation is the best permutation that meets the criteria in section (a). And correspondingly, Error =1 means that our permutation is the worst permutation.

In order to solve the problem, we check all permutations and pick the best one that would meet our criteria. But this approach is the brute-force approach and not the efficient one in terms of time complexity, as the time required to check all permutations raises exponentially as the amount of points increases. But, of course this approach guarantees the best result that meets the criteria in scope of our problem.

For research purposes, we decided to try different algorithms and compare their results with results of brute-force approach mentioned above in terms of accuracy and time complexity. As part of our research to solve the problem, we took a test set of 300-dimensional word vectors and applied the following approaches: (1a) all permutations сheck; (1b) all permutations сheck (parallelized computation); (2) traveling salesman algorithm; (3) hierarchical clustering algorithm.

## Formation of text boxes and their markup

## 

In order to determine the necessary and sufficient composition of the corps being formed, an expert survey was conducted, we received the recommendations of leading Kazakh and Russian experts was carried out and described their results. As a result of a comprehensive study, the composition and structure of the hull being formed, the sources for hull formation were determined, the requirements for hull formation were determined, the levels of required markings, their parameters and text classification, and methods of data collection for hull formation were determined [70]. Formed a balanced, representative corpus of journalistic texts, marked up and classified in terms of parameters. The corpus included 5,210 publications, 2,000 of which produced additional linguistic markup. Sources for the formation of the corpus were informational news Internet portals: «KazakhSTAN 2.0», «Central Asia Monitor», «Zakon.kz», «Радио Азаттык», «Tengrinews.kz». For the formation of the body was carried out 2-stage systematic cluster sampling.

A methodological approach has been developed and measures have been taken to organize work with experts and linguists involved in markup to ensure the reliability of the data obtained (data triangulation, development of instructions and explanatory meetings of experts) [71]. The corpus marking parameters corresponding to this stage of the study were determined:

* 1. Primary markup of 5210 publications on metadata (17 data categories)
  2. Expert marking for 5210 publication (9 categories: authenticity, objectivity, social significance, potential resonance, tonality, availability of manipulative techniques, politicization, Kazakhastan content, belonging to areas according to the subject)
  3. Deep linguistic markup with the text fragments selection, the dictionaries formation and the rules formalization for determining parameters in the 2000 publications text (3 categories: tonality, politicization, the present manipulator techniques).

The analysis of the actual characteristics of the hull in the context of sources according to the results of meta-marking by 17 parameters. According to the results of the analysis, it established that the texts volume included in the corpus differs from different sources in the number characters, which is explain by the editorial resource policy. Meta markup data is heterogeneous and some of the required data is missing. This is due to the fact that the architecture of the sites selected in the sample does not always provide for the possibility of collecting and having all the selected parameters of the meta-markup (for example, there is no possibility on the site to comment on the publication, or no data on the number of views) [72].

Work was done on in-depth analysis and linguistic markup of the formed body. An instruction has been developed to conduct in-depth analysis of textual information in the media in the context of the research objective. Linguistic markers were drawn to linguistic experts.

A deep analysis of the media publications formed corpus on 3 signs out of 12 indicated was carried out. Marked text corpus on the indicated informative features.

Formed Dictionary of 2892 words (15,076 word usage) on the basis of informative tonality. Dictionary formed of 37 words (156 word usage) on the informative politicization basis. Dictionary formed of 430 words (3797 word usage) on the basis of whether informative manipulative techniques. The rules of formalization are formulated by definition of informative signs of tonality, politicization, availability of manipulative techniques [73].

A set of preparatory works carried out to identify and identify patterns in the formed corpus of journalistic texts in order to formalize the rules. The complex of preparatory works included the following stages:

* 1. Determination of the necessary and sufficient composition of the body of texts being formed.
  2. Determination of parameters and markup levels of publications and their classification.
  3. Balanced formation and representative corpus of journalistic texts.
  4. Carrying out a marking of the body of texts.

According to the results of the analysis of the corpus, regularities are shown in two ways: tonality and the presence of manipulative techniques [74].

Held a set of preparatory works on the experimental work on the completeness research and correctness of the formed set of rules and dictionaries. Proposed an approach to the experimental research. For an experimental research of the completeness and correctness of the formed set of rules and dictionaries, an in-depth analysis of 2000 texts and the formation of dictionaries are carried out with 12 parameters. On the basis of the results texts in-depth analysis on 12 informative features, methods for formalizing rules will be developed [75]. Developed Lemmatizer (Lemmatization — the process of bringing the word form to the lemma - its normal (vocabulary) form) in the package Pymorphy2 (http://isa1.pythonanywhere.com/). Tests made on the allocation of subjects using an algorithm BigARTM.

Experiments on the determination of similar or paraphrased articles based on the metric of their cosine similarity were made. Started crawling (software search of Internet resource pages in order to build interconnections graphs and enter information into the database.) on the cloud server KazTeleport. Scraping added (popular method of getting content. The method consists in the fact that trained the algorithm goes to the home page and starts to pass over all the internal links, carefully collecting the insides of these you div) tags and links to sources of information in the publication. Produced reskraping all articles. Prepared database design to store the information obtained in step skrapinga. The database is running in test mode..

For successful and uninterrupted testing of algorithms, a group of engineering and technical workers installed and configured server equipment for the computational core of the IS (NVIDIA Tesla), operating system installed Centos 7.5, configured network installed and configured TensorFlow, Keras, server resources are divided between workgroups, remote server operation is configured, security is configured, uninterruptible power supply is configured and tested, libraries are installed cuda 9.0, cudnn7.3, neural networks tensorflow-gpu and their correct operation with the accelerator.

Configured virtual infrastructure for information system subsystems (crawler, database, front-end application). Created and configured virtual machine, network, remote access, work environment Anaconda Python. Configured communication channel between the virtual infrastructure and the computing core of the IS, that allows you to get the best results on the algorithms of automatic parameterization systems, classification and estimation of unreliability, as well as the results of the software of the hierarchical clustering algorithm.

# CONCLUSION

Within the framework of the project-targeted funding project for 2018, in accordance with the calendar plan, the following main results were obtained:

1. A preliminary version of the methodology was obtained for assessing the impact of open information sources on society based on the analysis of published textual information (it is planned to refine the methodology in 2019);
2. The research results of the of existing informative features were obtained in order to determine acceptable criteria for assessing the impact of open textual information sources on society;
3. The results of the study of the features of the assessment of the impact of open textual information sources on society in the conditions of the Republic of Kazakhstan
4. For the current stage, the necessary technical and expert-analytical conditions for the development of information system of assessing the impact of open source on the text information society.

All tasks have been completed in full and in accordance with the schedule (Appendix A).

The research results will be use in the development of information and analytical system for assessing the impact of open textual information sources on society.

The results obtained correspond to the current level of scientific and technological development and based on the latest achievements in the field of data processing in natural language, machine and in-depth training, pattern recognition, etc.

Total for the project published 29 publications. Of these, 3 articles in peer-reviewed foreign scientific journals with non-zero impact factor; 5 articles in domestic scientific journals with non-zero impact factor; 20 articles in the proceedings of international conferences; 1 chapter in the monograph of the publisher Springer.

REFERENCES

1. Musabayev R.R. Development of social trust information system in the education field using the "Big Data" technologies//Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 79-80. (in English)
2. K.Veltsel The birth the freedom / translation from English A.V. Lisovsky. – M, 2018. - 404p. (In Russian)
3. A.S. Sorokin, Building scoring cards using the logistic regression model // Naukovedenie online journal. - Moscow: Institute of Public Administration, Law and Innovation Technologies (IGUIT), 2014. - Vol. 2 (In Russian)
4. Krasovitsky A.M., Musabayev R.R. Energy-based centroid identification and cluster propagation with noise detection// Proceedings of 10th International Conference on Computational Collective Intelligence. - Bristol, Great Britain, 2018. – P. 523-533. (in English)
5. Morozova A. A. Reliability of information in social networks and criteria for its verification (on the example of “VKontakte”) // Bulletin of Chelyabinsk State University. -2017. -No 6 (402). - P. 75-83 // <https://cyberleninka.ru/article/n/dostovernost-informatsii-v-sotsialnyh-setyah-i-kriterii-ee-verifikatsii-na-primere-vkontakte>: 06.06.2018. (In Russian)
6. Musabayev R.R., Krasovitsky A.M., Ualiyeva I.M., Meyrambekkyzy Zh., Kassymzhanov B.K., Ibrayeva V.M. Monitoring and verification system of media publications’ reliability // Proceedings of the Scientific Conference IICT CS MES RK “Modern problems of computer science and computing technology”. - Almaty, 2018. – P. 178-184. (In Russian)
7. Flegel, R. C., Chaffee, S. H. Influences of editors, readers, and personal opinions on reporters // Journalism Quarterly. – 1971. - № 48(4). - Р. 645–651. (in English)
8. Vosoughi, S., Roy, D., Aral, S. The spread of true and false news online // Science. – 2018. – Vol. 359. - Р. 1146-1151. (in English)
9. Global research "Edelman Trust Barometer 2018" // URL: <http://cms.edelman.com/sites/default/files/2018-02/2018_Edelman_Trust_Barometer_Global_Report_FEB.pdf> : 29.07.2018 (In Russian)
10. Analyzing time devoted to online and traditional forms of media at a global level, as well as by age and across countries. // Digital vs. Traditional Media Consumption. INSIGHT REPORT SUMMARY | Q1. – 2017 // <https://cdn2.hubspot.net/hubfs/304927/Downloads/Digital%20vs%20Traditional%20Summary%20-%20Q1%202017.pdf> : 29.07.2018 (in English)
11. Results of a sociological survey on assessing the impact of open information sources (electronic media) on society. - Astana: Information and Analytical Center JSC MES RK, 2018. - 109 p. (In Russian)
12. Thomas L. Saaty, Luis G. Vargas, Models, Methods, Concepts & Applications of the Analytic Hierarchy Process: Second edition. – NY: Springer Science-Business Media, 2012. – P. 330. (in English)
13. Methods of complex assessment of territories of RK: within the General scheme of the organization of the territory of RK till 2030. – Astana, 2011. (In Russian)
14. Ryzhakova E.V. Automatic detection of irony and sarcasm in the text. – M.: Higher School of Economics, 2014. – P. 100. (In Russian)
15. Van Dijk T. A., Discourse, Ideology and Context // Folia Ling. - 2001. - №35(1–2). - Р. 11–40. (in English)
16. Ajiboye E., Ideological Discourse Analysis of the Functions of Feedback Comments on Online Reports of Socio-political Crises in Nigeria // Covenant J. Lang. Stud. – 2013. - № 1(12). - Р. 128-147. (in English)
17. N. Dankova Generalization strategy as a means of representation of the judiciary (on the example of Russian and English print media) // Political linguistics. - 2016. - Vol. 1, section 2. - P. 73–81. (In Russian)
18. I.V. Frolova On the expression of subjectivity in analytical articles of high-quality British and Russian press // Political linguistics. - 2015. - № 1 (51). - P. 138–145. (In Russian)
19. Recasens M., Danescu-Niculescu-Mizil C., Jurafsky D. Linguistic Models for Analyzing and Detecting Biased Language // Proc. 51st Annu. Meet. Assoc. Comput. Linguist. – 2013. - P. 1650–1659. (in English)
20. Morstatter F., Wu L., Yavanoglu U., Corman S. R., Liu H. Identifying Framing Bias in Online News //ACM Trans. Soc. Comput. – 2018. - № 1(2). - Р. 1–18. (in English)
21. Orlova O. G. Compiling a dictionary of stereotypes about Russians: content and form // Political linguistics. – 2013. - № 3(45). - P. 175–182. (in English)
22. Taboada M., Brooke J., Tofiloski M., Voll K., Stede M. Lexicon-Based Methods for Sentiment Analysis // Comput. Linguist. – 2011. - № 37(2). - Р. 267–307. (in English)
23. D’Andrea A., Ferri F., Grifoni P., Guzzo T. Approaches, Tools and Applications for Sentiment Analysis Implementation // Int. J. Comput. Appl. – 2015. - № 125(3). - Р. 26–33. (in English)
24. Lukashevich N. V., Levchik A. V. Creating the vocabulary of assessment words of the Russian language RUSENTILEX // OSTIS-2016, Open Semant. Technol. Intell. Syst. Minsk, BSUIR. - 2014. - P. 375-380. (In Russian)
25. Ualiyeva I.M., Krasovitsky A.M., Musabayev R.R. Generalization rate in open publication materials in Russian Language// Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 81-82. (In Russian)
26. Ualiyeva I.M., Krasovitsky A.M., Musabayev R.R., Meyrambekkyzy Zh. Numerical evaluation of generalization in Russian media texts taking into account sentiment sources: lexical approach // Proceedings of the Scientific Conference IICT CS MES RK «Modern problems of computer science and computing technology». - Almaty, 2018. – P. 289- 297 (In Russian)
27. Ualiyeva I.M., Krasovitsky A.M., Meyrambekkyzy Zh., Musabayev R.R. Allocation of lexico-semantic features and the formation of a key vocabulary based on a lexical software and expert approach// Proceedings of the XIV International Asian School Seminar “Complex Systems Optimization Problems”. - Cholpan-Ata, Kyrgyz Republic, 2018. - P. 280-286. (In Russian)
28. Ualiyeva I.M., Krasovitsky A.M., Meyrambekkyzy Zh., Mussabayev R.R. Recognition of generalization in the texts of the media on the basis of the program and expert approach // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 130-136. (In Russian)
29. Yusupov Madi. Analysis of short text messages’ sentiment in Russian using machine learning algorithms // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 334-339. (In Russian)
30. Shalkarbayuli A., Kairbekov A., Amangeldi Y. Comparison of traditional machine learning methods and google services in identifying tonality on Russian texts // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Алматы, 2018. - P. 28-37. (in English)
31. Shalkarbayuli A., Kairbekov A., Amangeldi Y. Comparison of traditional machine learning methods and Google services in identifying tonality on Russian texts // Journal of Physics: Conference Series. - 2018. – P. 9. (in English)
32. Order of the Minister of Information and Communications of the Republic of Kazakhstan "On approval of the Standard provision on the authorized officer (division) on interaction with mass media": approved. February 23, 2018, No. 50. (In Russian)
33. Law of the Republic of Kazakhstan "On Mass Media": approved. July 23, 1999, № 451-I.
34. Order of the Acting Minister of Investment and Development of the Republic of Kazakhstan: approved. January 26, 2016, No. 73. (In Russian)
35. Tatarova G. Methodology of data analysis in sociology. - M .: NOTA BENE, 1999. –224 p. (In Russian)
36. Ergaliyev E.N., Muhamediev R.I., Yakunin K., Symagulov A., Kairbekov A., Dusenbayeva A. Using cloud platforms to solve machine learning problems // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 87-95. (In Russian)
37. Ergaliyev E., Kayrbekov A., Kuchin Ya., Mukhamediyev R., Symagulov A. NLP as SAAS // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 220-230. (In Russian)
38. Zhai C., Aggarwal C. Mining text data: chapter 4: A survey of text clustering algorithms. Springer Science+Business Media, LLC. - NY: Springer-Verlag, 2012. – 524 p. (in English)
39. Jurafsky, Martin J., Speech and Language Processing (2nd Edition). – USA: Prentice-Hall, Inc. Upper Saddle River, NJ. – 2009. –1032 p. // <https://web.stanford.edu/~jurafsky/slp3/> : 29.09.2018 (in English)
40. Mikolov T., Le Q. Distributed representations of sentences and documents // ICML'14 Proceedings of the 31st International Conference on International Conference on Machine Learning. - Beijing, China, 2014. – Р. 1188-1196 // <https://arxiv.org/abs/1405.4053> : 01.10.2018 (in English)
41. Zhu Z., Hu J. Context Aware Document Embedding, CoRR. – 2017. // <https://arxiv.org/abs/1707.01521> : 01.10.2018 (in English)
42. Kusner M. From word embeddings to document distances // Proceeding of the 32nd International Conference on International Conference on Machine Learning. - Lille. France, 2015. – Vol. 37. – Р. 957-966. (in English)
43. SemEval // <https://en.wikipedia.org/wiki/SemEval#Semantic_evaluation_tasks>: 01.10.2018
44. Firth J. R. A synopsis of linguistic theory 1930–1955 // In Studies in Linguistic Analysis. – Blackwell, Oxford, 1957. - Р. 1-32. (in English)
45. Mikolov T., Chen K., Corrado G. Dean J. Efficient Estimation of Word Representations in Vector Space. – 2013. – Р. 12. // <https://arxiv.org/abs/1301.3781> : 01.10.2018 (in English)
46. Pennington J, Socher R, Manning C. GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). - Doha, Qatar, 2014. – Р. 1532–1543. (in English)
47. Mikolov T., Grave E., Bojanowski P., Puhrsch P., Joulin Advances in pre-training distributed word representations. – 2017. – Р. 4. // <https://arxiv.org/abs/1712.09405> : 01.10.2018 (in English)
48. Peters M., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. Deep contextualized word representations. – 2018. – Р. 15. // <https://arxiv.org/abs/1802.05365> : 01.10.2018 (in English)
49. Jaccard index // <https://ru.wikipedia.org/wiki/Коэффициент_Жаккара> : 01.10.2018 (In Russian)
50. DBSCAN // <https://en.wikipedia.org/wiki/DBSCAN> : 01.10.2018 (in English)
51. DBSCAN pros and cons// <http://algowiki-project.org/ru/Участник:Dadrik/Плотностный_алгоритм_кластеризации> :01.10.2018 (in English)
52. SyntaxNet // <https://opensource.google.com/projects/syntaxnet> : 01.10.2018 (in English)
53. Albertini M.K., de Mello R.F. Energy-based function to evaluate data stream clustering // Advances in Data Analysis and Classification. -2013. -Vol 7(4). –P. 435-464. (in English)
54. Online clustering data sets UCI // <https://archive.ics.uci.edu/ml/datasets.html>: 01.10.2018 (in English)
55. Bagirov A.M. Modified global k-means algorithm for minimum sum-of-squares clustering problems // Pattern Recognition. -2008. - № 41(10). –P. 3192 - 3199. (in English)
56. Carrizosa E., Alguwaizani A., Hansen P., Mladenovic N. New heuristic for harmonic means clustering // Journal of Global Optimization. - 2015. – № 63(3). -P.427-443. (in English)
57. Ester M., Kriegel P., Sander H., Xu J. A density-based algorithm for discovering clusters in large spatial databases with noise // AAAI Press. -1996. - P. 226-231. (in English)
58. Girolami M. Mercer kernel-based clustering in feature space // IEEE Transactions on Neural Networks. -2002. – № 13. –P. 780-784. (in English)
59. Krasovitsky A.M., Musabayev R.R. Energy maximization approach for solving clastering problem//Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 77-78. (In Russian)
60. R.R.Musabayev, A.M.Krasovitsky, Hierarchical clustering algorithm based on the energy criterion // Proceedings of the Scientific Conference IICT CS MES RK «Modern problems of computer science and computing technology. - Almaty, 2018. – P. 185-188 (In Russian)
61. A.M.Krasovitsky, R.R.Musabayev, Algorithm of centroid identification and formation of clusters on the basis of maximizing energy criterion // Proceedings of the XIV International Asian School Seminar “Complex Systems Optimization Problems”. - Cholpan-Ata, Kyrgyz Republic, 2018. - P. 46-49. (In Russian)
62. Development of a methodology for evaluating open information sources's influence on a society based on the analysis of published textual information in relation to the system of social trust: the annual report of the research work of AO // Information and Analytical Center. – Astana,2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018 (In Russian)
63. Gabbassov М.B. TOFI technology capabilities for data processing and visualization. // Collection conference materials «Application of Information and Communication Technologies-AICT2014». – Astana, 2014. - P. 276 – 277. (in English)
64. Abilkaeva Zh.N., Gabbasov M.B., Ismagulova F.E., Kuanov T.D., Shpekbaev M.B. The concept of implementation of the TOFI synchronizer. // VII Int. scientific Conf. "Problems of differential equations, analysis and algebra". - Aktobe, 2015. - P. 75-84. (In Russian)
65. Gabbasov M.B., Kuanov T.D., Abilkaeva Zh.N., Yermagambetov T.K. Synchronization of metadata using the namespace TOFI. // Mater. Int. scientific Conf. “Computer science and applied mathematics” (“Computer science and Applied Mathematics”), dedicated to the 60th anniversary of Academician M.N.Kalimoldayev. - Almaty, 2017. - P. 152-161. (In Russian)
66. Study of existing informative features to determine acceptable criteria for assessing the impact of open textual information sources on society in relation to STS: annual research report. JSC "Information and Analytical Center". - Astana, 2018. // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018 (In Russian)
67. Study of the complex assessment of the used informative criteria's significance: annual research report. JSC "Information and Analytical Center". - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018 (In Russian)
68. Conducting a set of works aimed at experimentally confirming the adequacy of the obtained comprehensive assessment of the used informative criteria's significance for assessing the impact of open information sources on a society based on the analysis of published textual information: annual research report. JSC «Information and Analytical Center». - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018 (In Russian)
69. Study of the evaluating features of the open textual information sources' influence on society in the conditions of the Republic of Kazakhstan in relation to STS: annual research report. JSC "Information and Analytical Center". - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018 (In Russian)
70. Complex research and also the expert analysis of necessary and sufficient structure of the formed text corpus, its structure, presentability, balance demanded levels of a marking and classifications, their unambiguity; ways of collecting, storage and representation of the text case; methods of the work and mutual control organization of the experts involved at his formation: annual report of research. JSC «Information and Analytical Center». - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018 (In Russian)
71. Conducting a complex works on the formation of a balanced, representative, marked-up and classified class of journalistic texts: texts of official media, journal publications, news portals, etc .: annual research report. JSC «Information and Analytical Center». - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC : 01.10.2018
72. Conducting a complex works on the analysis of the actual characteristics of the formed corpus (composition, structure, representativeness, balance, levels of marking and classification, their correctness and completeness; etc.): annual research report JSC «Information and Analytical Center». - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC: 01.10.2018 (In Russian)
73. Development of methods for formalizing rules and dictionaries for recognizing informative features in the marked corpus and determining their expressiveness degree: annual research report JSC «Information and Analytical Center». - Astana, 2018 // https://drive.google.com/open?id=17NTO4bzEy7sfHiOaxVZxu5okb-YT4IaC: 01.10.2018 (In Russian)
74. Conducting a complex works on the analysis and identification of patterns that are present in the formed corpus, with the aim of formalizing the rules and forming dictionaries for recognizing informative features in the marked corpus and determining their expressiveness degree based on the methods developed: annual research report JSC «Information and Analytical Center». - Astana, 2018. - P. 29. (In Russian)
75. Conducting a complex works on experimental study of the completeness and correctness of the formed set of rules and dictionaries for recognizing informative features in the marked corpus and determining their expressiveness degree: annual research report JSC «Information and Analytical Center». - Astana, 2018. - P .31. (In Russian)
76. Ualiyeva I.M., Meyrambekkyzy Zh., Krasovitsky A.M., Musabayev R.R. Dictionary of generalized words and phrases // <https://drive.google.com/open?id=0B84afv_CAE9jdXlYcDVUaTg1WDkwYVN2elNpRUNEdEhXX3hN> : 01.10.2018. (In Russian)

# APPENDIX A

# Work schedule

Appendix 1

to the Additional agreement No. 1 dated \_\_\_\_\_\_\_\_\_ 2018

under the contract No. 319 dated March 30, 2018 on

program and target financing

TECHNICAL SPECIFICATION AND WORK SCHEDULE CALENDAR

Under the Additional agreement No. 1 dated 2018

to the contract No. 319 dated March 30, 2018

1. NAME OF THE CONTRACTOR

1. By priority: 3. Information, telecommunication and space technologies, scientific research in the field of natural sciences.
2. By subpriority: 3.1 Intelligent Information Technologies.
3. By the program subject: No. BR05236839 «Development of information technologies and systems for stimulation of personality’s sustainable development as one of the bases of development of digital Kazakhstan».
4. Overall budget of the program 945 000 000 (nine hundred and forty-five million) tenge, including breakdown by years, for work performance according to paragraph 3:

- for 2018 - in the amount of 305 000 000 (three hundred and five million) tenge;

- for 2019 - in the amount of 325 000 000 (three hundred and twenty-five million) tenge;

- for 2020 - in the amount of 315 million (three hundred and fifteen million) tenge.

2. Characteristics of scientific and technical products by qualification and economic indicators

1. Direction: Information technology.
2. Scope: Departmental analytical information systems.
3. Final outcome:

* 2018: research results of the existing informative features to determine acceptable criteria for evaluating the open textual information sources' influence on society will be obtained: results of the research on peculiarities of evaluating open textual information sources' influence on society in the Kazakhstani context will be obtained; an article in a peer-reviewed foreign or domestic scientific publication with a non-zero impact factor will be published.
* 2019: a method of evaluating open information sources' influence on society based on textual

publications analysis will be created: development of a calculation method for particular criteria

of evaluating open information sources' influence on society will be initiated; an article in a peer-

reviewed foreign or domestic scientific publication with a non-zero impact factor will be published.

* 2020: an information system evaluating open information sources' influence on society will be

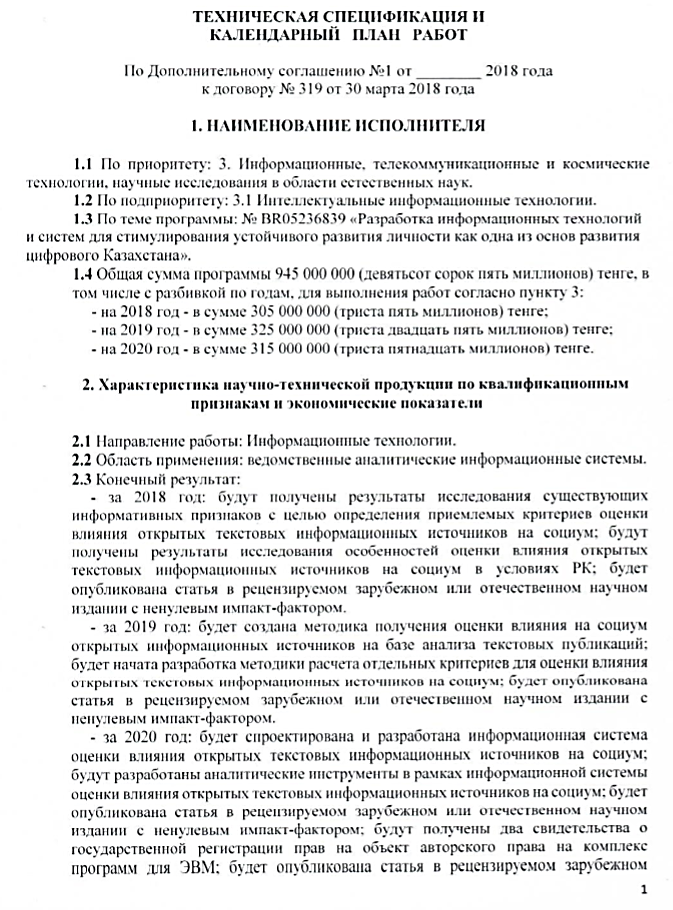
designed and developed; analytical tools within the information system evaluating open information sources' influence on society will be developed: two certificates of state registration of rights for subject of copyright to a set of computer programs will be obtained; an article in a peer-reviewed foreign publication indexed in the Web of Science or Scopus database with a non-zero impact factor will be published.

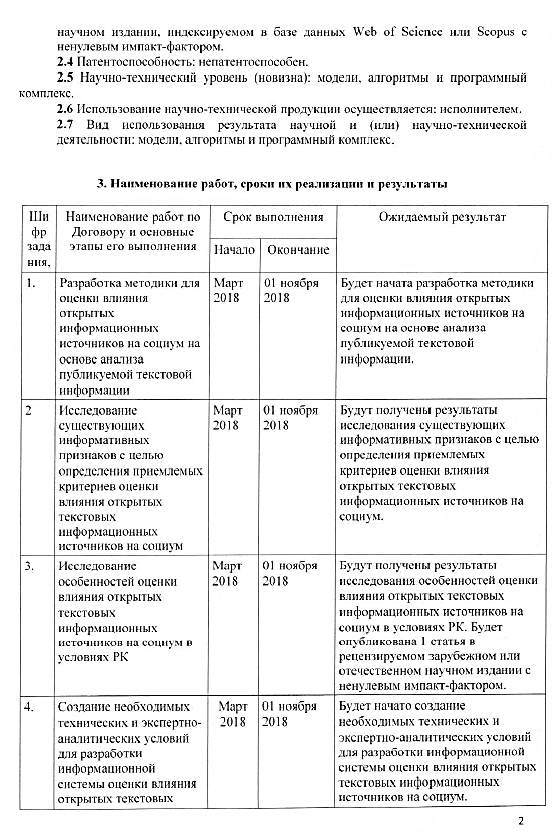
1. Patentability: nonpatentable.
2. Scientific and technical level (novelty): models, algorithms and software complex.
3. The use of scientific and technical products is carried out: by the contractor.
4. Type of the scientific and (or) scientific and technical activities result's use: models, algorithms and software complex.

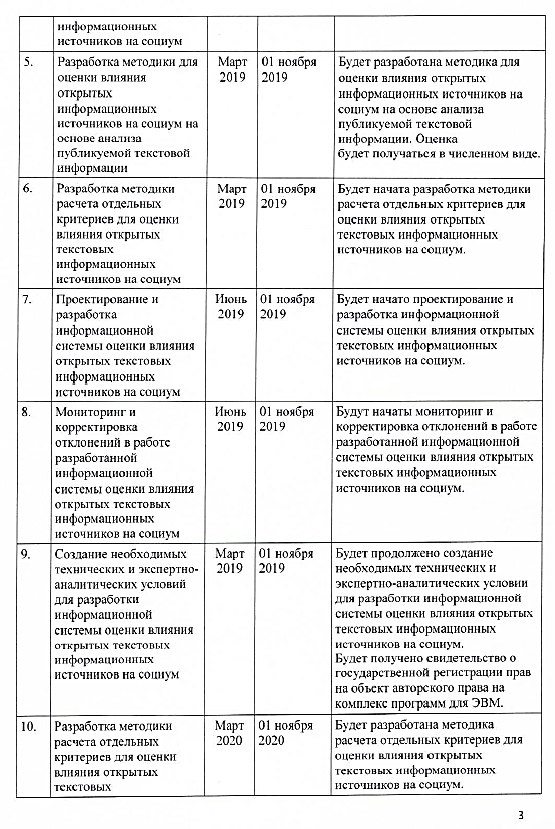
3. Activities, implementation dates and results

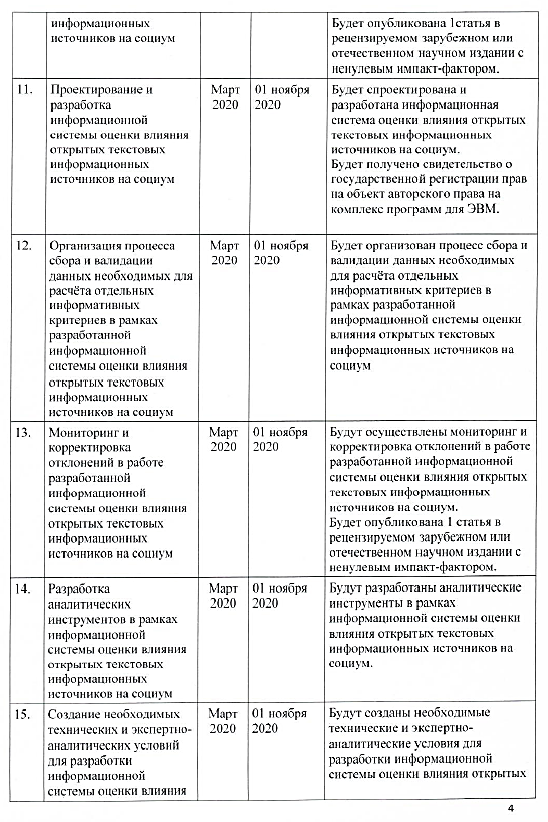
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task cipher | Activities under the Contract and the main stages of its implementation | Implementation date | | Expected result |
| Start | End |
|  | Development of methods evaluating the open textual information sources' influence on society on the basis of published textual information analysis | March 2018 | November 01, 2018 | Development of methods evaluating the open textual information sources' influence on society on the basis of published textual information analysis will be initiated. |
|  | Study of the existing informative features in order to determine acceptable criteria for evaluating the open textual information sources' influence on society | March 2018 | November 01, 2018 | Research results of the existing informative features in order to determine acceptable criteria for evaluating the open textual information sources' influence on society will be obtained. |
|  | Research on peculiarities of evaluating open textual information sources' influence on society in the Kazakhstani context | March 2018 | November 01, 2018 | Results of the research on peculiarities of evaluating open textual information sources' influence on society in the Kazakhstani context will be obtained.  An article in a peer-reviewed foreign or domestic scientific publication with a non-zero impact factor will be published. |
|  | Creation of the necessary technical and expert-analytical conditions for the development of an information system evaluating open textual information sources' influence on society | March 2018 | November 01, 2018 | Creation of the necessary technical and expert-analytical conditions for the development of an information system evaluating open textual information sources' influence on society will be initiated. |
|  | Creation of a method of evaluating open information sources' influence on society based on textual publications analysis | March 2019 | November 01, 2019 | A method of evaluating open information sources' influence on society based on textual publications analysis will be created. Evaluation will be obtained in numerical form. |
|  | Development of a calculation method for particular criteria of  evaluating open textual information sources' influence on society | March 2019 | November 01, 2019 | Development of a calculation method for particular criteria of evaluating open textual information sources' influence on society will be initiated. |
|  | Design and Development  information system evaluating open textual information sources' influence on society | June  2019 | November 01, 2019 | Design and Development  information system evaluating open textual information sources' influence on society will be initiated. |
|  | Monitoring and correction of deviations in the work of the developed information system for evaluating open textual information sources' influence on society | June  2019 | November 01, 2019 | Monitoring and correction of deviations in the work of the developed information system for evaluating open textual information sources' influence on society will be initiated. |
|  | Creation of the necessary technical and expert-analytical conditions for the development of an information system evaluating open textual information sources' influence on society | March  2019 | November 01, 2019 | Creation of the necessary technical and expert-analytical conditions for the development of an information system evaluating open textual information sources' influence on society it will be continued.  A certificate of state registration of rights for subject of copyright to a set of computer programs will be obtained. |
|  | Development of a calculation method for particular criteria of evaluating open textual information sources' influence on society | March  2020 | November 01, 2020 | Calculation method for particular criteria of evaluating open textual information sources' influence on society will be developed.  An article in a peer-reviewed foreign or domestic scientific publication with a non-zero impact factor will be published. |
|  | Design and development of  information system evaluating open textual information sources' influence on society | March  2020 | November 01, 2020 | An information system evaluating open textual information sources' influence on society will be designed and developed.  A certificate of state registration of rights for subject of copyright to a set of computer programs will be obtained. |
|  | Organization of the collecting process and validating data necessary for calculating particular informative criteria within the developed information system for evaluating open textual information sources' influence on society | March  2020 | November 01, 2020 | Process of collecting and validating data necessary for calculating particular informative criteria within the developed information system for evaluating open textual information sources' influence on society will be organized. |
|  | Monitoring and correction of deviations in the work of the developed information system for evaluating open textual information sources' influence on society | March  2020 | November 01, 2020 | Deviations in the work of the developed information system for evaluating open textual information sources' influence on society will be monitored and corrected.  An article in a peer-reviewed foreign or domestic scientific publication with a non-zero impact factor will be published. |
|  | Development of analytical tools within the information system evaluating open textual information sources' influence on society | March  2020 | November 01, 2020 | Analytical tools within the information system evaluating open textual information sources' influence on society will be developed. |
|  | Creation of the necessary technical and expert-analytical conditions for the development of an information system evaluating open textual information sources' influence on society | March  2020 | November 01, 2020 | Necessary technical and expert-analytical conditions for the development of an information system evaluating open textual information sources' influence on society will be created.  An article in a peer-reviewed foreign publication indexed in the Web of Science or Scopus database with a non-zero impact factor will be published. |

|  |  |
| --- | --- |
| **Requester:**  **Acting Chairman of SA**  **“The Committee of Science of the Ministry of Education and Science of Republic of Kazakhstan”**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R.S.Nurseitov** | **Сontractor:**  **Director General of RSE on the basis of economic control rights “IICT” IICT CS MES RK**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ M.N. Kalimoldayev** |
|  | **Agreed and accepted by**  **Scientific supervisor of the project**  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ R.R. Mussabayev** |











# APPENDIX B

# List of publications

* 1. Musabayev R.R. Development of social trust information system in the education field using the "Big Data" technologies//Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 79-80. (In English)
  2. I.M.Ualiyeva, A.M.Krasovitsky, R.R.Musabayev Generalization rate in open publication materials in Russian Language// Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 81-82 (In English)
  3. A.M.Krasovitsky, R.R.Musabayev, Energy maximization approach for solving clastering problem//Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 77-78 (In English)
  4. B.Bisarinov, R.R.Musabayev, A.Bisarinova, Collective method in solving the Big Data clastering problems// Proceedings of the 16th International Scientific Conference "Information Technologies and Management 2018". - Riga, Latvia, 2018. - P. 75-76 (In English)
  5. R.R.Musabayev, A.M.Krasovitsky, I.M.Ualiyeva, Zh.Meyrambekkyzy, B.K.Kassymzhanov, V.M.Ibrayeva, Monitoring and verification system of media publications’ reliability // Proceedings of the Scientific Conference IICT CS MES RK “Modern problems of computer science and computing technology”. - Almaty, 2018. – P. 178-184 (In Russian)
  6. R.R.Musabayev, A.M.Krasovitsky, Hierarchical clustering algorithm based on the energy criterion // Proceedings of the Scientific Conference IICT CS MES RK «Modern problems of computer science and computing technology. - Almaty, 2018. – P. 185-188 (In Russian)
  7. I.M.Ualiyeva, A.M.Krasovitsky, R.R.Musabayev, Zh.Meyrambekkyzy, Numerical evaluation of generalization in Russian media texts taking into account sentiment sources: lexical approach // Proceedings of the Scientific Conference IICT CS MES RK «Modern problems of computer science and computing technology». - Almaty, 2018. – P. 289- 297 (In Russian)
  8. R.R.Musabayev, T.Turdalykyzy, Сluster analysis: value, methods and approaches // Bulletin of KazNRTU. - Almaty, 2018. - No 3(127). - P. 173-177. (In Kazakh)
  9. N.Tasbolatuly, B.K.Kassymzhanov, T.T.Duzbaev, overview of textual parameterization algorithms // Proceedings of the XIV International Asian School Seminar “Complex Systems Optimization Problems”. - Cholpan-Ata, Kyrgyz Republic, 2018. - P. 64-68. (In Kazakh)
  10. R.R.Musabayev, B.K. Kassymzhanov, V.M.Ibrayeva, Development of information technologies and systems for stimulation of personality’s sustainable development as one of the bases of development of digital Kazakhstan // Proceedings of the XIV International Asian School Seminar “Complex Systems Optimization Problems”. - Cholpan-Ata, Kyrgyz Republic, 2018. - P. 39-46. (In Russian)
  11. A.M.Krasovitsky, R.R.Musabayev, Algorithm of centroid identification and formation of clusters on the basis of maximizing energy criterion // Proceedings of the XIV International Asian School Seminar “Complex Systems Optimization Problems”. - Cholpan-Ata, Kyrgyz Republic, 2018. - P. 46-49. (In Russian)
  12. I.M.Ualiyeva, A.M.Krasovitsky, Zh.Meyrambekkyzy, R.R.Musabayev, Allocation of lexico-semantic features and the formation of a key vocabulary based on a lexical software and expert approach// Proceedings of the XIV International Asian School Seminar “Complex Systems Optimization Problems”. - Cholpan-Ata, Kyrgyz Republic, 2018. - P. 280-286. (In Russian)
  13. Krassovitskiy A., Mussabayev R. (2018) Energy-Based Centroid Identification and Cluster Propagation with Noise Detection. In: Nguyen N., Pimenidis E., Khan Z., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2018. Lecture Notes in Computer Science, Springer, Cham, 2018. vol 11055. – Р. 523-533. <https://doi.org/10.1007/978-3-319-98443-8_48>: 01.10.2018 (In English)
  14. E.Ergaliyev, A.Kayrbekov, Ya.Kuchin, R.Mukhamediyev, A.Symagulov, NLP as SAAS // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 220-230. (In Russian)
  15. Yusupov Madi. Analysis of short text messages’ sentiment in Russian using machine learning algorithms // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 334-339. (In Russian)
  16. Shalkarbayuli A., Kairbekov A., Amangeldi Y. Comparison of traditional machine learning methods and google services in identifying tonality on Russian texts // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Алматы, 2018. - P. 28-37. (In English)
  17. S.A.Abdullayeva, R.R. Musabayev, The problem of extracting named entities. Review of modern services // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 47-54. (In Russian)
  18. V.B. Barakhnin, E.N.Ergaliyev, O.Yu. Kozhemyakina, R.I.Muhamediev, Automatic processing of texts in natural language. Some bibliometric indicators of the current state of the research area // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 60-70. (In Russian)
  19. E.N.Ergaliyev, R.I.Muhamediev, K.Yakunin, A.Symagulov, A.Kairbekov, Dusenbayeva A. Using cloud platforms to solve machine learning problems // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 87-95. (In Russian)
  20. I.M.Ualiyeva, A.M.Krasovitsky, Zh.Meyrambekkyzy, R.R.Mussabayev, Recognition of generalization in the texts of the media on the basis of the program and expert approach // Proceedings of the III International Scientific **and Practical** Conference **“Сomputer Science and Applied Mathematics”,** dedicated to the 80th anniversary of professor R.G. Biyashev. and the 70th anniversary of professor M.B. Aidarkhanov. - Almaty, 2018. - P. 130-136. (In Russian)
  21. Plotnikov Roman, Erzin Adil, Mladenovic Nenad. VNDS for the min-power symmetric connectivity problem // Optimization Letters (Impact Factor – 1,013). - Springer Berlin Heidelberg, 2018 // <https://doi.org/10.1007/s11590-018-1324-0> : 01.10.2018 (In English)
  22. Mukhamedyev R., Kiseleva S., Mustakaev R., Yakunin K., Gopejenko V. Multi-criteria decision support system for RES evaluation //Application of Information and Communication Technologies (AICT) 2018. 11th International Conference on. – IEEE. – 2018. – P. 13-18. (In English)
  23. U.A. Ospanova, M.K. Atanayeva, T. K. Buldybayev, I.G.Akoyeva, Study of informative features of reliability and objectivity of publications in electronic media from the point of view of the population of Kazakhstan // «Bulletin of Kazakh State Women's Teacher Training University». - 2018. -No 3(75). – P.139-146 (In Russian)
  24. M.K. Atanayeva, T. K. Buldybayev, U.A. Ospanova, I.G.Akoyeva, The instruction for the expert analysis and classification of reliability of suspicious publications of news character // «Bulletin of Kazakh State Women's Teacher Training University». - 2018. – No 3(75). – P. 148-153 (In Russian)

# APPENDIX C

# List of foreign information resources

* 1. Springer Link // <https://link.springer.com/>
  2. Scopus // <https://www.scopus/>
  3. Web of Science // <https://apps.webofknowledge.com/>
  4. Cornell University Library // <https://arxiv.org/>
  5. РМЭБ // <http://rmebrk.kz/>
  6. Mendeley // <https://www.mendeley.com/>
  7. Cyberleninka // <https://cyberleninka.ru/>
  8. Google Scholar // <https://scholar.google.com/>
  9. ResearchGate // <https://www.researchgate.net/>

# APPENDIX D

# Model of methods for obtaining an integrated assessment of the media publication's reliability degree on the basis of TOFI technologies

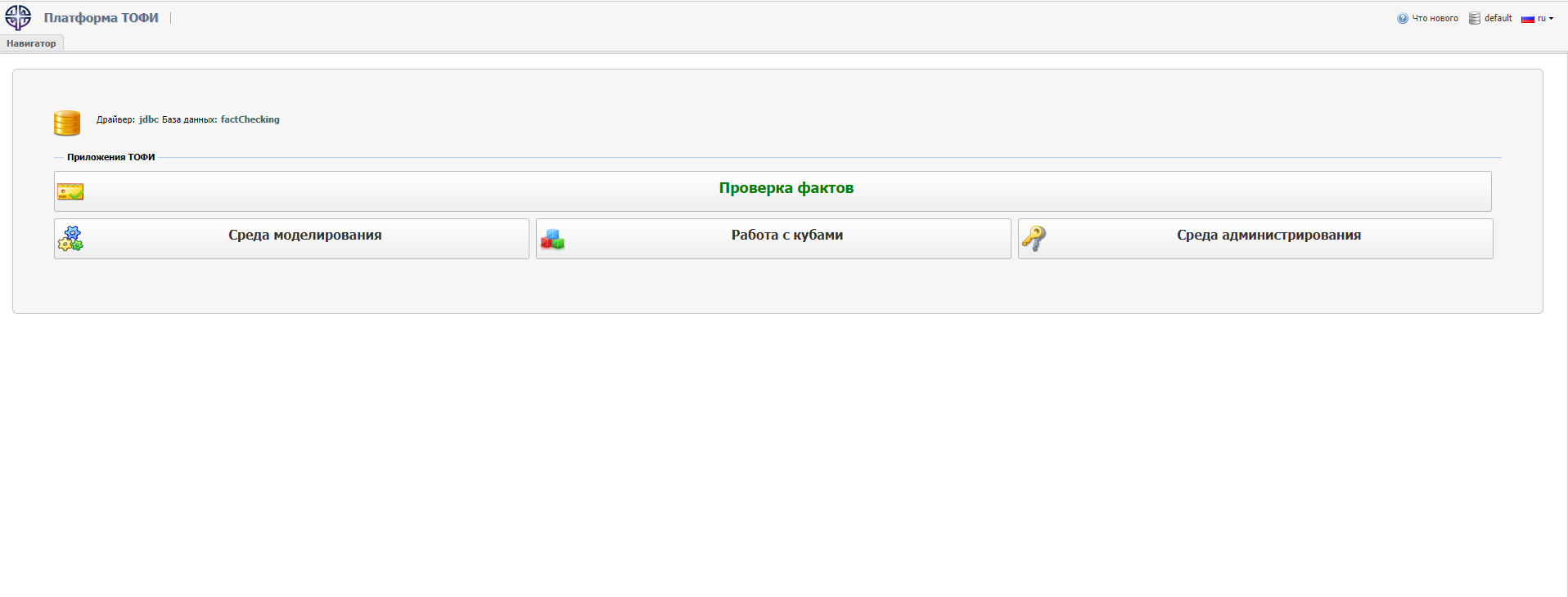


Figure D.1 - Verification of facts on the TOFI platform

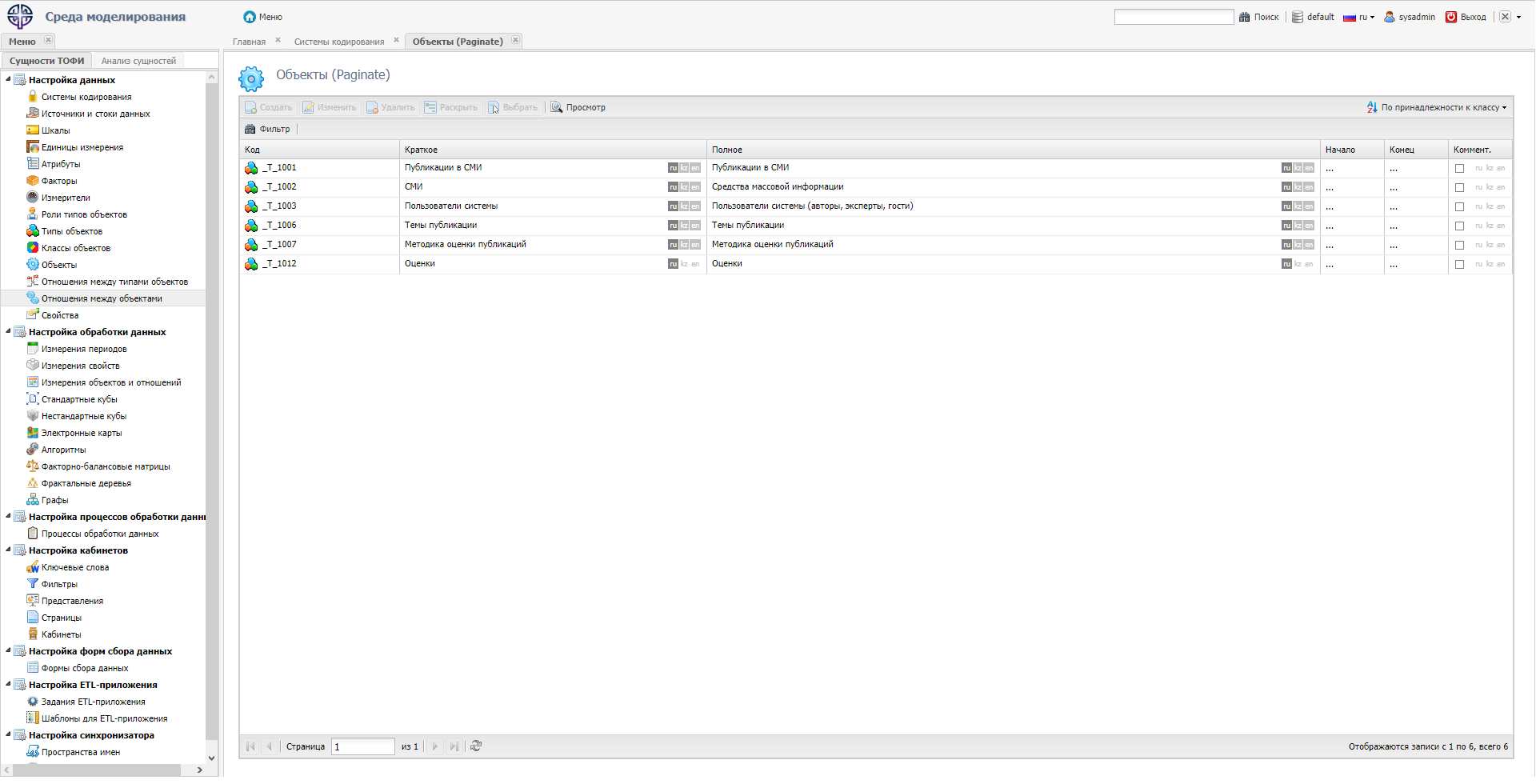


Figure D.2 – Modeling environment. Objects

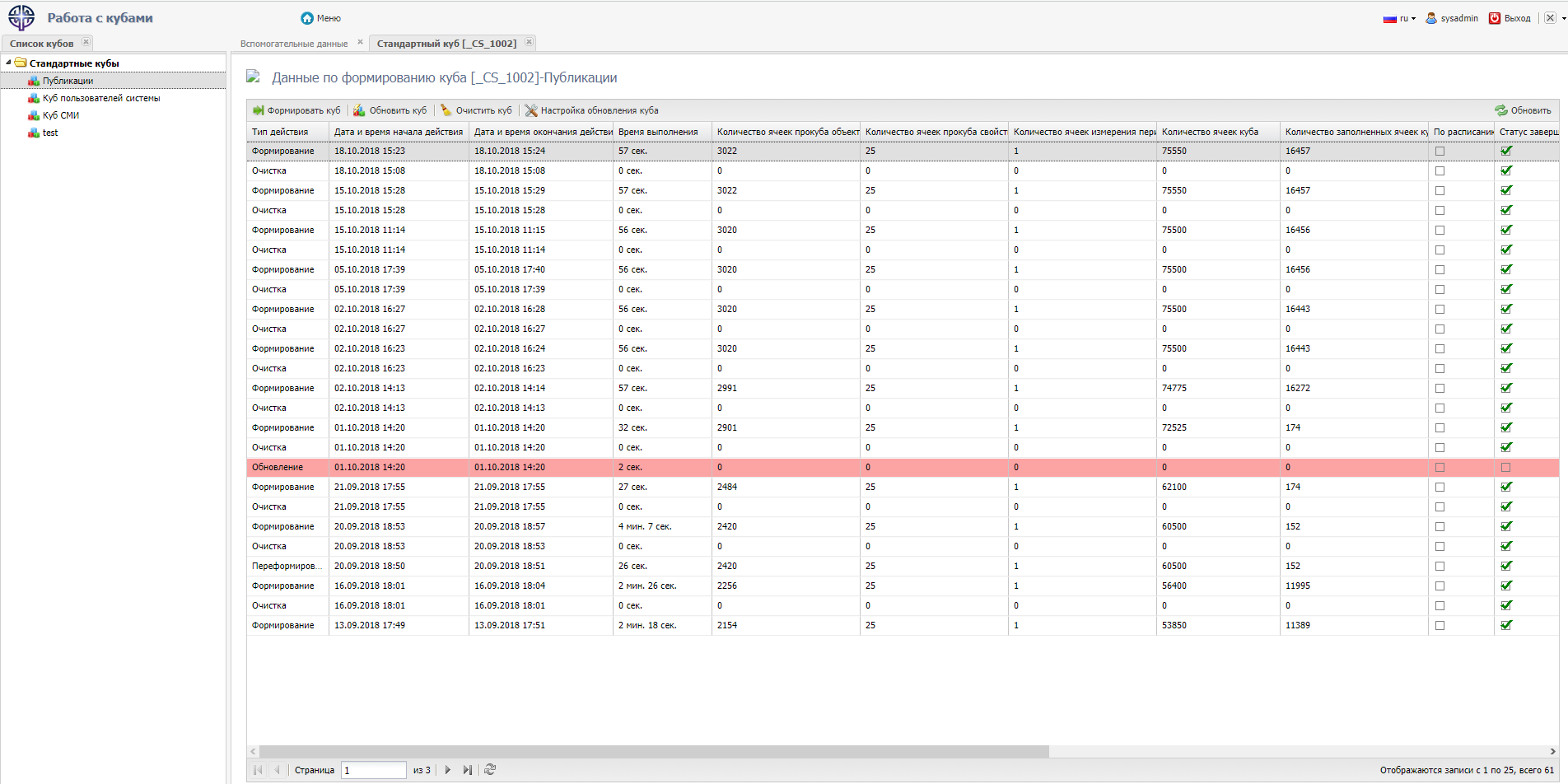


Figure D. 3 – Work with cubes. Data on formation of a cube

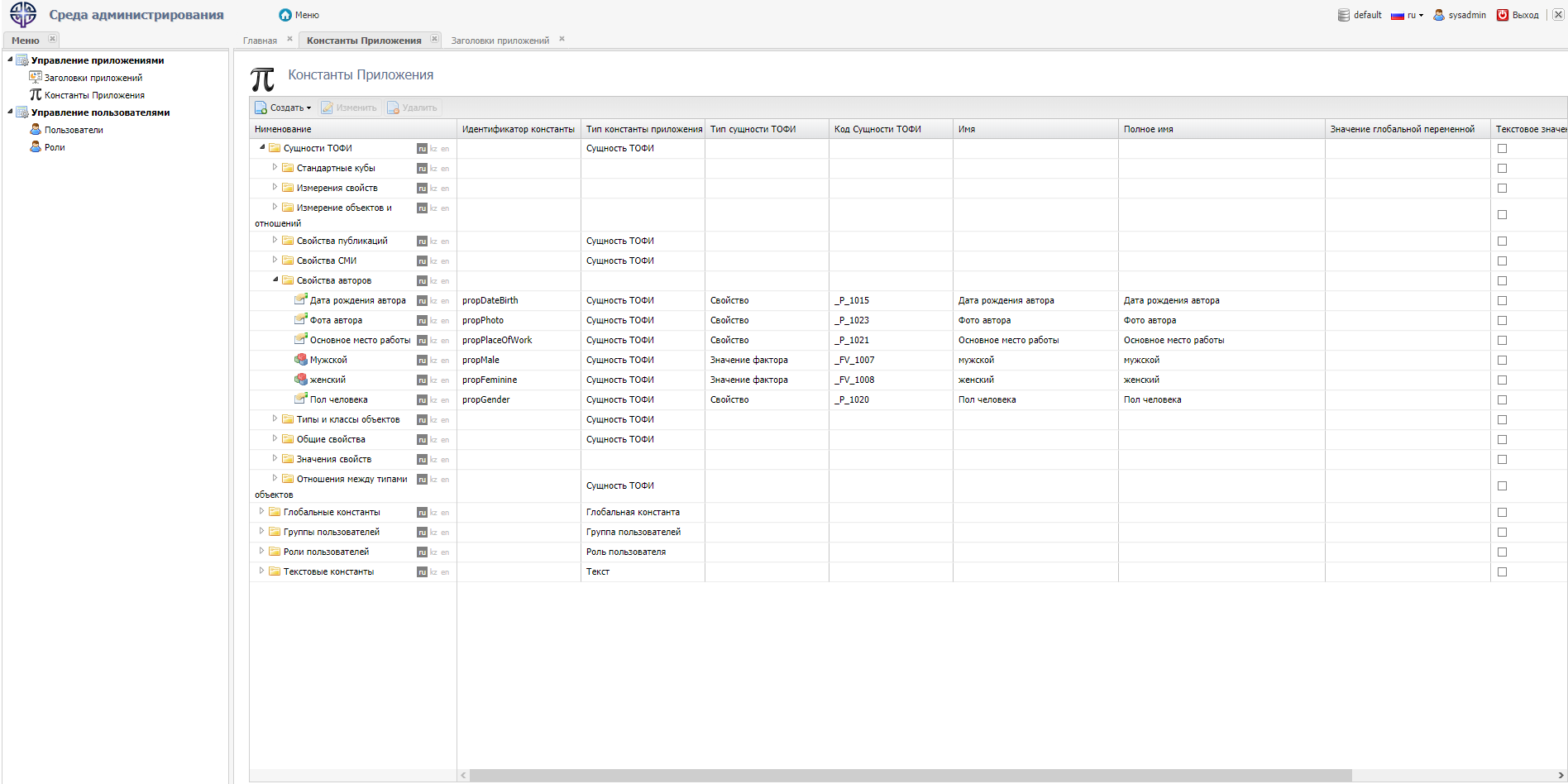


Figure D.4 – Administration environment. Application constants

# APPENDIX E

# Informative features

Table E.1 - Informative features

|  |  |
| --- | --- |
| Informative criteria | Performance indicators |
| *Reliability –* allows to assess the accuracy of published information based on the availability of links in publications to the source, the credibility of the publication on which publications are posted, coverage of the same event by different editions, checking the authenticity of photos and videos that confirms the event, the publication title matches the media text. | 1. 1. The presence of links in the publication to a competent source |
| * 1. Information of the state authorities of the Republic of Kazakhstan |
| * 1. Information from official internet resource of state authorities of the Republic of Kazakhstan; |
| * 1. Information published by organizations of the quasi-public sector of the Republic of Kazakhstan |
| * 1. Information published by nation companies of the Republic of Kazakhstan |
| * 1. Official information of news agencies, electronic media and other similar companies with a high level of the reputation |
| * 1. Data from analytical reports, publications, studies prepared by authoritative and recognized international organizations (UN, WHO, UNESCO, IMF, World Bank, European Bank for Reconstruction and Development, Interpol and other resources of government departments of other countries) |
| * 1. Recognized international statistical databases (UNESCO Institute for Statistics, Eurostat, OECD statistics, World Bank etc.) |
| * 1. Professional expert platforms (consulting, rating platform, information) recognized in their field, which is supported by representatives of reputable international organizations |
| * 1. Research results prepared by recognized scientific organizations, universities, etc. |
| * 1. Responses to requests of government agencies and international organizations |
| 1. Link to the source of information. |
| 1. Coverage of the same event in different editions crosschecking, discourse with other articles. (Registered Ip parameters: Subject (Discussion Subject)/Location /Action/Time) |
| 1. The authenticity of the photo, which confirms the event |
| 1. The authenticity of the video that confirms the event |
| 1. The fact of subsequent changes in the already published article |
| 1. Correspondence of the publication title to the content of the media |
| 1. Presence/ absence of the author |
| 1. The presence/ absence of verifiable facts in the article |
| 1. Reputation of the publication (websites of news and information agencies electronic media, other similar sites) on which information is published |
| *Resonance* - allows to assess the scale of distribution of a news publication and determine the main character of the audience’s attitude to the publication based on media coverage, number of reposts, shares, comments and comment sentiments. | 1. Media coverage (number of views) |
| 1. 2. Number of information shares |
| 1. Number of publication comments |
| 1. Comments sentiment  * Negative * Neutral * Positive |
| *Sentiment* - allows you to assess the emotional tone of the presentation of information, facts in the publication, which the author broadcasts to the audience. There are two basic methods for automatically determining the tonality of the statistical method and the method based on dictionaries and rules. | 1. 1. Publication sentiment  * Negative * Neutral * Positive |
| *Objectivity* - allows to evaluate objectivity of the facts presentation, information in the publication.  In itself, the presence of an expressed opinion is not a basis for classifying the text as “biased”, however, too frequent use of sentences with a vivid expression of the author’s personal opinion in combination with the absence/small number of facts, as well as signs of sarcasm/irony, agitation/anti-agitation in the text - may indicate the bias of the information provided. | 1. The presence in the media text, the expression of author’s personal opinions regarding a publication |
| 1. The presence of manipulator techniques in the media text |
| 1. Politicization Degree of publication |
| 1. The degree of generalization:  * Weakly expressed * Strongly expressed * no |
| *Media Involvement* - allows to evaluate the behavior of the reader on a particular web resource based on the analysis of the depth and duration of viewing the publication and the frequency of returning to the site. | 1. Viewing depth of the publication. |
| 1. Duration of viewing the publication by the audience |
| 1. Frequency of returning the audience to the Internet media portal |

# APPENDIX F

# Frequency of the generalization quantifier “all” in the Wikipedia corpus

Table F.1 - Frequency of the generalization quantifier “all” in the Wikipedia corpus

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Bigrams with the word “all” | years | know | want | can | others | these | they | Residents | Citizens |
| tengrinews.kz | 356 | 3160 | 1330 | 657 | 2910 | 9300 | 4540 | 777 | 503 |
| % | 0,002 | 0,019 | 0,008 | 0,004 | 0,018 | 0,057 | 0,028 | 0,005 | 0,003 |
| Camonitor.kz | 171,000 | 271,000 | 69,000 | 49,000 | 592,000 | 1620,000 | 1120,000 | 65,000 | 203,000 |
| % | 0,009 | 0,015 | 0,004 | 0,003 | 0,032 | 0,089 | 0,061 | 0,004 | 0,011 |
| occurrence frequency of the word “all” in Wikipedia articles | 0,132501 | | | | | | | | |

Continuation of the table F.1

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| Bigrams with the word “all” | Copy-books | We | Books | Politi-cians | Mini-sters | Offi-cials | Civil servants | Others | Number of pages |
| tengrinews.kz | 1 | 4070 | 110 | 71 | 98 | 438 | 90 | 511 | 163000 |
| % | 0,000 | 0,025 | 0,001 | 0,000 | 0,001 | 0,003 | 0,001 | 0,003 | 1,000 |
| Camonitor.kz | 0,000 | 454,000 |  | 11,000 | 16,000 | 26,000 | 1,000 |  | 18300,000 |
| % | 0,000 | 0,025 | 0,000 | 0,001 | 0,001 | 0,001 | 0,000 | 0,000 | 1,000 |
|  |  | | | | | | | | |

# APPENDIX G

# Software application for recognition of generalization in texts (screenshots)

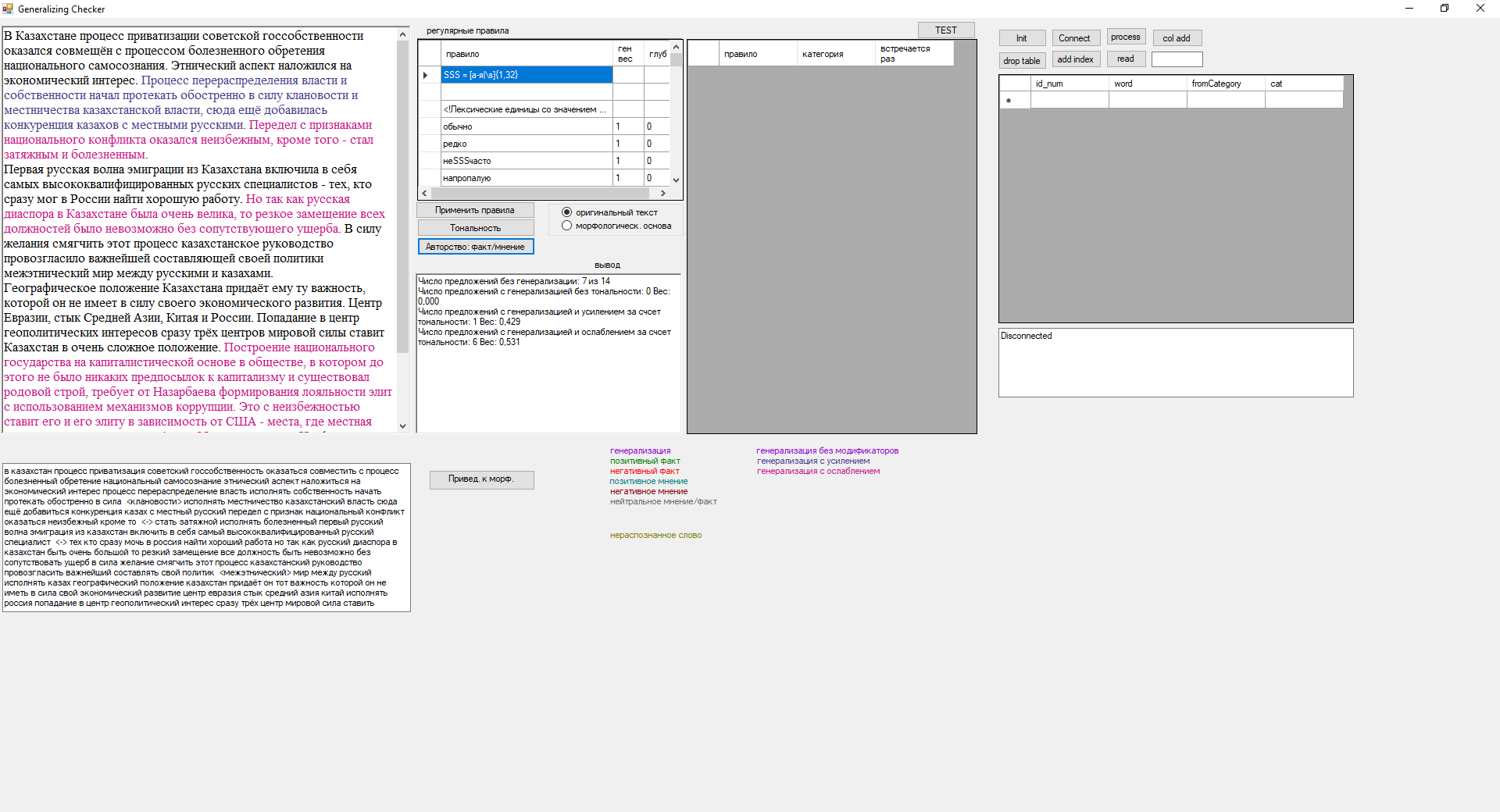


Figure G.1 – Recognition of generalization in a sentence

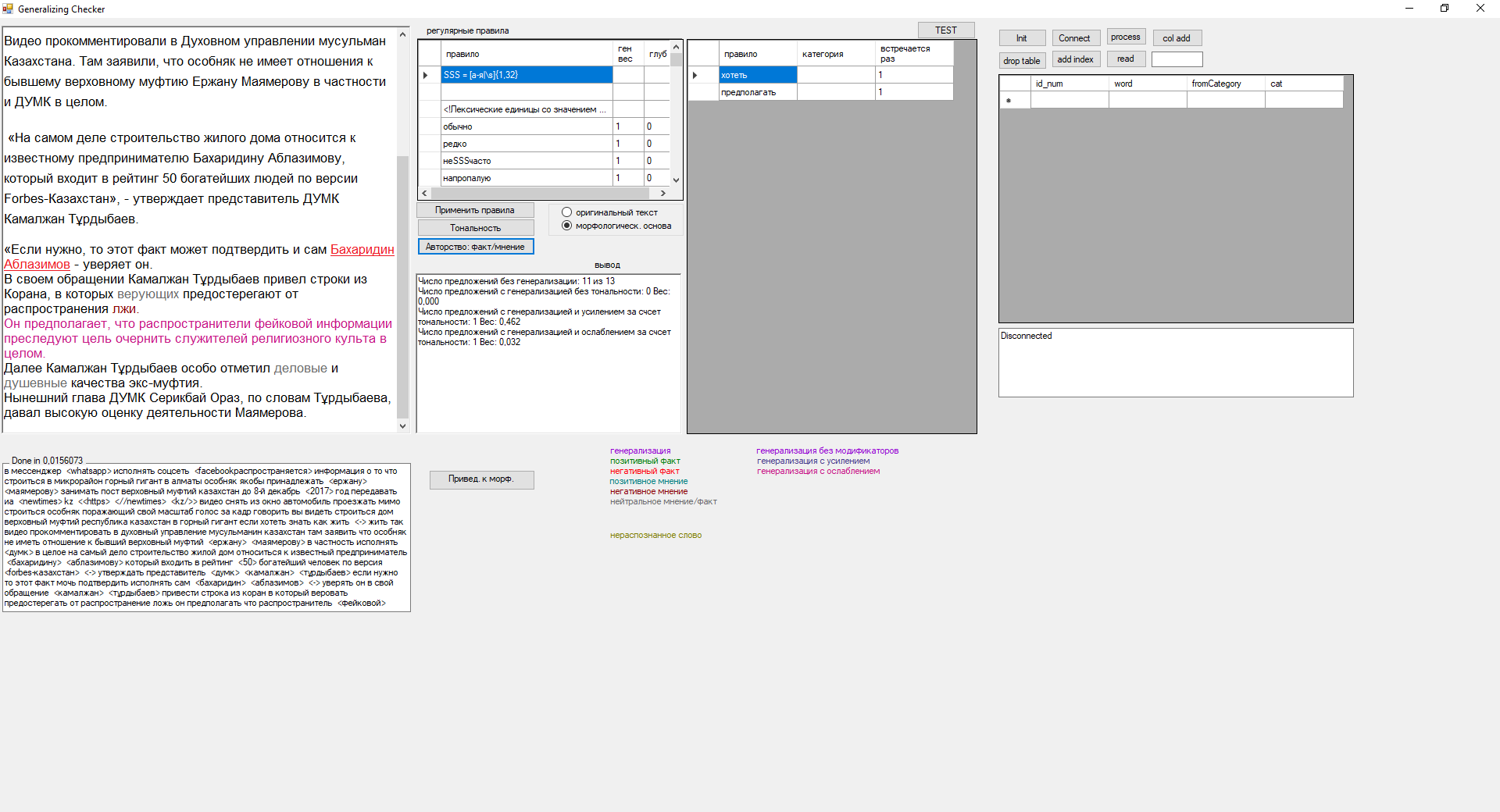


Figure G.2 – Fact and generalization

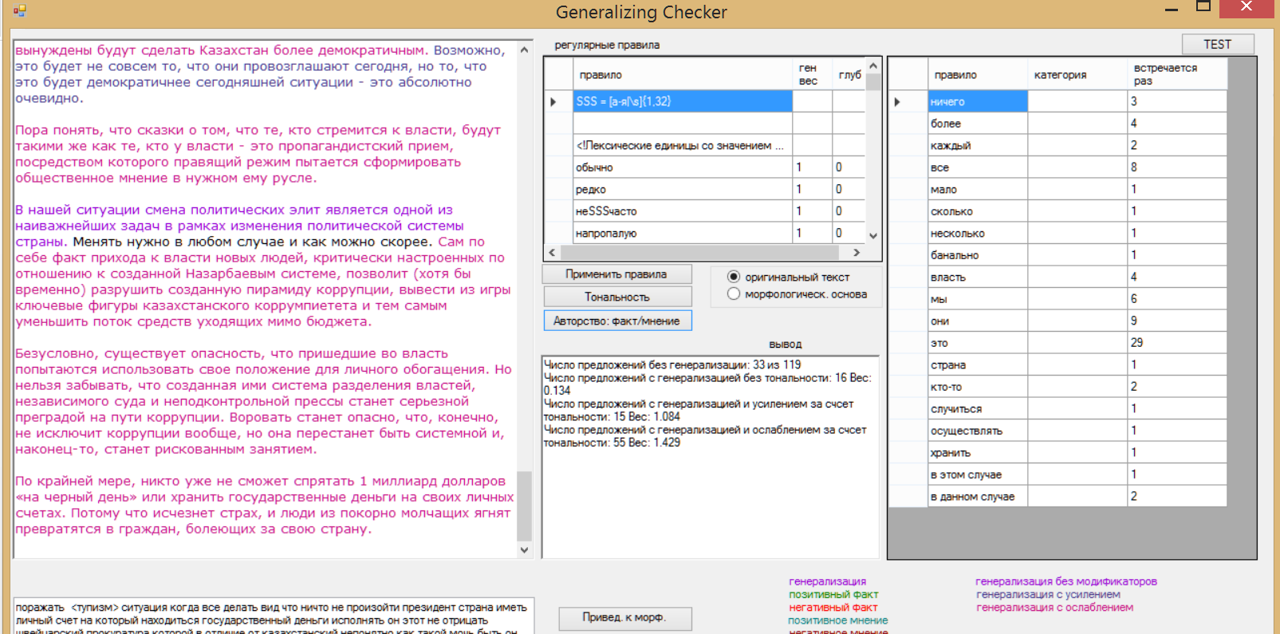


Figure G.3 – Generalization degree

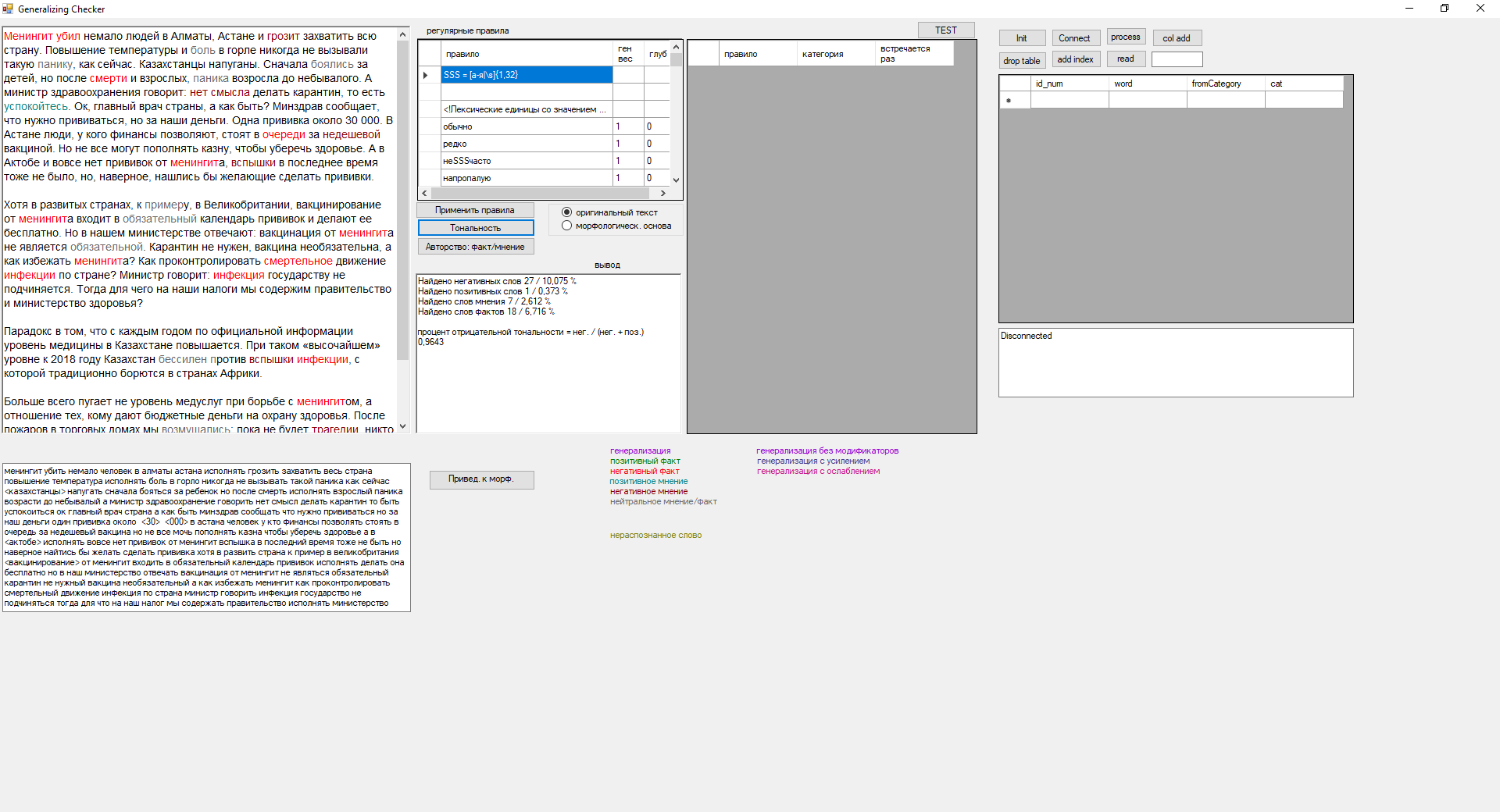


Figure G.4 – Sentiment recognition