

STRUCTURAL ABSTRACT

Report 105 pages, 7 figures, 2 tables, 114 sources, 7 appendixes.

OPEN INFORMATION SOURCES, MASS MEDIA, INFORMATION SYSTEM, INFORMATIVE FEATURES, SENTIMENT, GENERALIZATION, TOPIC MODELING, CLUSTERING, NATURAL LANGUAGE PROCESSING, MACHINE LEARNING, BIG DATA, TEXT MINING, KEY PHRASE EXTRACTION, NEWSWORTHY

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’ impact on society based on the analysis of published textual information and their algorithmic implementation as a part of the corresponding information and analytical system.

To achieve the goal and solve research problems at each stage of the work, various methods and approaches have been used, such as: sociological survey, expert survey, method of theoretical analysis and generalization of scientific and analytical literature, additive method of index construction, general scientific methods of observation, systematization, generalization, statistical method, linguistic methods of contextual, discursive, interpretive, pragmatic and component analysis, introspection, as well as big data and machine learning technologies and etc.

Key indicators: methods of assessing the impact on society of open information sources on the basis of the analysis of text publications (in numerical form): on the basis of the topic model of the media corpus and on the basis of a comprehensive assessment of various informative features. The development of a methodology of calculating individual criteria for assessing the impact of open text information sources on society has been started: a new topic modeling algorithm based on the cluster approach, a method for automatically setting the tonality of texts by the method of their conceptual analysis, a method for automatically numerically evaluating the degree of informational content of publications, an approach for automatic classification of publications by type, genre and style have been developed and tested; the analysis of methods that ensures the detection of destructive messages in the Internet environment and an assessment of the severity of informative criteria based on the vocabulary approach have been carried out. The development of an information system of assessing the impact of open text information sources on society has been started: the purpose and objectives of the task are determined, the architecture has been developed, and the need and sufficiency of the information system functionality have been determined. The necessary technical and expert-analytical conditions have been created for the development of an information system: new methods of cluster analysis, a method of dense topic vectorization of texts contained in a large text corpus (BigData), an algorithm of grouping news publications in accordance with information occasions, a method of clustering news media messages based on their conceptual analysis, an algorithm of clustering large data of high dimension based on the decomposition method, methods, algorithms and tools of preprocessing text publications, a method for automatic summarization of text documents and optimization methods have been developed.

CONTENT

[1 DEVELOPMENT OF METHODS FOR ASSESSING THE INFLUENCE OF OPEN INFORMATION SOURCES ON SOCIETY ON THE BASIS OF PUBLISHED TEXT INFORMATION ANALYSIS 11](#_Toc22899157)

[1.1 Methodology of assessing the impact of open information sources on society on the basis of multimodal media assessment based on a thematic model of the corpus of publications 11](#_Toc22899158)

[1.1.1 NLP and MCDA 12](#_Toc22899159)

[1.1.2 Media Evaluation Process 14](#_Toc22899160)

[1.1.3 Multimodal Media Evaluation Algorithm 16](#_Toc22899161)

[1.1.4 Analysis of the social importance of thematic categories 18](#_Toc22899162)

[1.1.5 Methodology of forming dictionaries of socially significant topics 19](#_Toc22899163)

[1.2 Improvement of the methods of assessing the impact of open information sources on society on the basis of a comprehensive assessment of various informative features 20](#_Toc22899164)

[2 DEVELOPMENT OF METHODS OF CALCULATING INDIVIDUAL CRITERIA FOR ASSESSING THE IMPACT OF OPEN TEXT INFORMATION SOURCES ON SOCIETY 23](#_Toc22899165)

[2.1 Determination of the composition, classes and significance of informative features 23](#_Toc22899166)

[2.2 Topic modeling 24](#_Toc22899167)

[2.2.1 Topic modeling based on BigARTM and GenSim libraries 24](#_Toc22899168)

[2.2.2 The topic modeling algorithm based on clustering methods 24](#_Toc22899169)

[2.3. Automatic setting of the tonality of texts by the method of their conceptual analysis 30](#_Toc22899170)

[2.4. Generalization 30](#_Toc22899171)

[2.5. Degree of informative content of publications 31](#_Toc22899172)

[2.6. The classification of publications by type, genre and style 31](#_Toc22899173)

[2.7. The probable destructiveness 32](#_Toc22899174)

[2.8. Assessment of severity in the analyzed text of information criteria based on the dictionary approach 33](#_Toc22899175)

[3 DESIGN AND DEVELOPMENT OF AN INFORMATION SYSTEM OF EVALUATING THE IMPACT OF OPEN TEXT INFORMATION SOURCES ON SOCIETY 36](#_Toc22899176)

[3.1 Purpose, goals and objectives of the developed 36](#_Toc22899177)

[3.2 Architecture of the information system 37](#_Toc22899178)

[3.3 Definition of necessary and sufficient functionality of the developed information system for the qualitative solution of tasks of the project 40](#_Toc22899179)

[4 MONITORING AND CORRECTION OF DEVIATIONS IN THE WORK OF THE DEVELOPED INFORMATION SYSTEM OF EVALUATING THE IMPACT OF OPEN TEXT INFORMATION SOURCES ON SOCIETY 43](#_Toc22899180)

[5 CREATION OF NECESSARY TECHNICAL AND EXPERT-ANALYTICAL CONDITIONS FOR DEVELOPMENT OF THE INFORMATION SYSTEM OF EVALUATING OPEN TEXT INFORMATION SOURCES' INFLUENCE ON SOCIETY 44](#_Toc22899181)

[5.1 Clustering Analysis methods and their application in the analysis of media publications 44](#_Toc22899182)

[5.1.1 Dense text embeddings based on the co-occurrence of words in texts and topic modelling using “Big Data” 45](#_Toc22899183)

[5.1.2 Clusterization of news publications in accordance with reflected information events (newsbreaks) 49](#_Toc22899184)

[5.1.3 Method of clustering news media based on their conceptual analysis 50](#_Toc22899185)

[5.1.4 High Dimension Big Data Clustering Algorithm Based on Decomposition Method 50](#_Toc22899186)

[5.2 Methods, algorithms, and tools for preprocessing text publications 55](#_Toc22899187)

[5.2.1 Automatic summarization of text documents using word mover’s distance and extracted document keywords 56](#_Toc22899188)

[5.3 Development of optimization methods 58](#_Toc22899189)

[5.4 Creating the necessary technical conditions 58](#_Toc22899190)

[5.5 Creation of necessary expert and analytical conditions 60](#_Toc22899191)

[CONCLUSION 64](#_Toc22899192)

[REFERENCES 65](#_Toc22899193)

[APPENDIX A](#_Toc22899194) [Work schedule 76](#_Toc22899195)

[APPENDIX B](#_Toc22899196) [List of publications 85](#_Toc22899197)

[APPENDIX C](#_Toc22899198) [List of foreign information resources 91](#_Toc22899199)

[APPENDIX D](#_Toc22899200) [The main parameters of media texts. 92](#_Toc22899201)

[APPENDIX E](#_Toc22899202) [Experiments on the applicability of the multimodal media estimation algorithm 94](#_Toc22899203)

[APPENDIX F](#_Toc22899204) [The results of ranking topics by method 1.3 100](#_Toc22899205)

[APPENDIX G 104](#_Toc22899206)

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 [1-4]. 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 2019 reporting year:

1. Development of methods for evaluation of open information sources’ influence on society based on the analysis of published textual information.
2. Development of methods of calculating individual criteria for assessing the impact of open text information sources on society
3. Design and development of an information system of assessing the impact of open text information sources on society
4. Monitoring and correction of deviations in the work of the developed information system of assessing the impact of open text information sources on society
5. 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.

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

## Methodology of assessing the impact of open information sources on society on the basis of multimodal media assessment based on a thematic model of the corpus of publications

According to the Edelman Trust Barometer 2019 survey conducted in 27 countries, trust in government information and media channels remains low. There is a growing gap between the informed public and the bulk of the population (2018 – 13 points, 2019 – 16) [5]. In cases where the audience does not have reliable knowledge or experience regarding what is happening, it especially depends on the information provided by the media [6]. According to studies [7], [8], - to form, give a certain opinion or focus the audience on specific topics, the media use various techniques and mechanisms of manipulation. An additional factor affecting our perception is the increased availability of various news on the Internet, which can create confusion caused by personal, often subjective, sources of information such as personal television, blogs, and unverified news. [9]. At the same time, the media is a kind of feedback system between government bodies and society.

In this regard, it is important to understand how the media can use its influence in order to mitigate the negative impacts of the media and encourage positive effects [10].

Media content assessment is a focus of researchers because of its practical relevance to news agencies, advertising companies and the public sector. Based on the automatic analysis of media content, you can predict the likely popularity of news articles, provide suitable and high-quality articles for users, plan PR strategies for promoting goods or services (Alexandru Tatar et al. [[1]](#footnote-1), Roja Bandari et al.[[2]](#footnote-2)). The public sector receives a tool to promote and educate on innovations, build a media plan and identify negative content prohibited by law. Individual users can quickly and efficiently filter out a large amount of information.

Each paper published by the media has a certain set of parameters, often heterogeneous. For example, it can be both objective indicators (the number of links and posts), and subjective (emotional coloring or tonality). Essentially, we need to attribute the published text to a certain class, which allows to evaluate, for example, the degree of its impact on readers or the need for a more detailed analysis. For this, it is necessary to identify the indicated parameters (criteria), evaluate their mutual importance and the magnitude of the presence of this parameter in the analyzed text. Then it is required to aggregate the parameter values in order to obtain an assessment of the belonging of the text to a certain class (Figure 1). We can classify the media having received these estimates based on the totality of papers of a single media, and by combining these estimates in some way.

Note that these estimates for one paper and for one media may not be zero at once for several parameters and classes. In addition, it is advisable to receive media estimates, having only part of the body of texts they publish, since obtaining all published papers can be difficult.

## NLP and MCDA

Given the diversity of digital sources of information and the correspondingly large volumes of news reports, there is a growing need for their automatic analysis. Evaluation of the parameters of media texts can be based on natural language processing, where significant progress has been made in the development of models and methods of text analysis. In turn, aggregation of values of heterogeneous parameters is often performed using systems of multi-criteria analysis and decision making (multiple-criteria decision making (MCDM)).

One of the methods efficiently used in the field of NLP, is a thematic analysis, or topic modeling. Thematic modeling is a method based on statistical characteristics of document collections, which is used in the tasks of automatic abstracting, information retrieval, information retrieval and classification [11]. The meaning of this approach is to intuitively understand that the documents in the collection form groups in which the frequency of occurrence of words or word combinations varies.

Topic modeling is distinguished by a well-developed algorithms and methods based on a statistical model of the language, and the use of clusters of documents related to a set of topics allows to solve problems of synonymy and polysemy of terms [12]. Probabilistic topic models describe documents (M) by a discrete distribution on a variety of topics (T), and topics by a discrete distribution on a variety of terms [13]. In other words, the topic model defines each document relates to which topics and which words form each topic. Probabilistic latent semantic analysis (PLSA), ARTM (Additive regularization of topic models) [14] and, very popular, Latent Dirichlet Allocation(LDA) [15] are used to build a topic model of the corpus of documents. LDA can be expressed by the following equality:

(1)

representing the sum of mixed conditional distributions on all topics of the set T, where p (w | t) is the conditional distribution of words in topics, p (t | m) is the conditional distribution of topics in news. The transition from the conditional distribution p (w | t, m) to p (w | t) is carried out due to the conditional independence hypothesis, according to which the appearance of words in news m on a topic t depends on the topic, but does not depend on news m, and is common to all news. This ratio is valid, based on the assumptions that there is no need to preserve the order of documents (news) in the corpus and the order of words in the news, in addition, the LDA method assumes that the components are generated by a continuous multivariate Dirichlet-multinomial distribution. The goal of the algorithm is to find the parameters and, by maximizing the likelihood function with appropriate regularization

(2)

nmw – the number of occurrences of the word w, in the news m, R (φ, θ) is the logarithmic regularizer.

Each topic can be considered as a separate parameter of the text or the media as a whole, the value of which is determined by the mentioned conditional probabilities. In cases where it is necessary to make a decision on the basis of a variety of heterogeneous parameters and alternatives, a number of methods are used to take into account the mutual importance of the parameters and aggregate their values in the form of one or a small number of estimates. By their nature, such tasks relate to the field of multi-criteria decision support, widely used in decision support systems (DSS) [16,17,18]. Including in NLP [19]. MCDM uses a number of methods that provide solutions based on heterogeneous criteria, which include [20]: Weighted Linear Combination (WLC) and Ordered Weighted Averaging (OWA) [21], PAPRIKA (Potentially all pairwise rankings of all possible alternatives [22], ELECTRE [23], TOPSIS (technique for order performance by similarity to ideal solution) [24], MAUT, PROMETHEE (Preference Ranking Organization METHod for Enrichment of Evaluations [25], VIKOR, AHP (analytical hierarchy process) [26], as well as Bayesian networks [27,28], fuzzy logic [29] and their combination BaFAHP [45], etc.

An important step in the application of these methods is to obtain knowledge from experts (knowledge extraction) [30], which include methods of iterative discussion by a panel of experts (Delphi [31, 32]) and various ways of ranking decision criteria (parameters) based on Likert scales [33], Mokken [34] and others. A pairwise comparison of criteria is often used because of its psychological validity, for example, MaxDiff [35], the essence of which is to select two factors from a given list by the respondent — the most important and least important, Bradley-Terry [36], which uses the processing of a large number of pairwise comparisons by experts of the “more/less important” method of maximum likelihood to obtain a probability distribution function that estimates any pair of criteria.

Some of the methods listed above, such as PAPRIKA, PROMETHEE, TOPSIS, ELECTRE, and AHP include peer review mechanisms.

In this study, a topic model constructed using LDA is used to calculate conditional probability distributions of documents by parameters, topics and classes. Application of the topic model provides in our opinion reduction of the errors connected with calculating the correspondence of texts and parameter dictionaries.

To assess the weight of the contribution of parameters to the classification process, we use AHP, which has been tested on many practical problems [37].

The aggregation of the obtained values is carried out using the Bayesian method, which ultimately allows to calculate the probability of hypotheses about the nature of the media.

## Media Evaluation Process

We believe that materials that have a significant impact on the information environment of society should be considered more carefully, in contrast to materials of a private, domestic, humorous, etc. For this, it is proposed, firstly, to consider mainly resonant papers that cause a large number of responses, and secondly, to allocate a socially significant group of news in them, which should be analyzed more carefully, evaluating the reliability and objectivity of the information contained in the paper.

It should immediately be noted that the social significance of the topic is far from always directly related to its resonance. Thus, according to Yandex [38], in 2018, users from Kazakhstan most often made requests on the following topics: 1) The football world Cup. 2) Winter Olympics in Pyeongchang. 3) The battle of Nurmagomedov and McGregor 4) Fire in the shopping center "Zimnyaya vishnya". 5) UEFA Nations League. 6) Murder of Denis Ten. 7) Raising the retirement age in Russia. 8) Lunar Eclipse on July 27. 9) Meningitis outbreak in Kazakhstan. 10) Presidential elections in Russia. It is obvious that only the themes "Outbreak of meningitis in Kazakhstan" and "Murder of Denis ten" (since any citizen can face street crime) are socially significant for Kazakhstanis.

In view of this, there is practically no alternative to expert assessments in determining the social significance of the topics, and the authorities are also interested in obtaining such assessments. Thus, the President of Kazakhstan N.A.Nazarbayev at the meeting of the Security Council of the Republic of Kazakhstan on November 7, 2018 called the most pressing issues that concern Kazakhstanis [39]: "The Results of the sociological study showed that the first place out of the six problems that were highlighted is the high cost of utilities." Among other issues of particular concern to Kazakhstan people, the President pointed out expensive medical care and training, low quality of education.

A study conducted by the Center of Political Analysis and Strategic Studies of the party "Nur Otan" based on the results of 2017 showed that the 10 most acute problems for the population of Kazakhstan are as follows [40]: 1) Rising prices for food, essential goods. 2) Low income, lack of money. 3) High utility rates. 4) Repayment of the loan. 5) Poor quality of medical care. 6) Corruption. 7) Prices for fuel and lubricants 8) Fear of losing a job. 9) Lack of own housing. 10) Lack of work opportunities.

Summarizing the results of these two studies, we can make a list of topics of social significance for the residents of Kazakhstan: Utility prices, Food prices, Fuel prices, Housing prices, Income level of the population, Lending to the population, Corruption, Medical care, Education (the last two topics cover the whole range of issues related to medicine and education: both the cost and the quality of the relevant services).

Thus, to separate news and analytical information requiring a more thorough analysis (class “suspicious”) from the mass of papers that do not require such analysis (class “not suspicious”), the following scheme is proposed that includes several sequential classification steps (Figure 1).

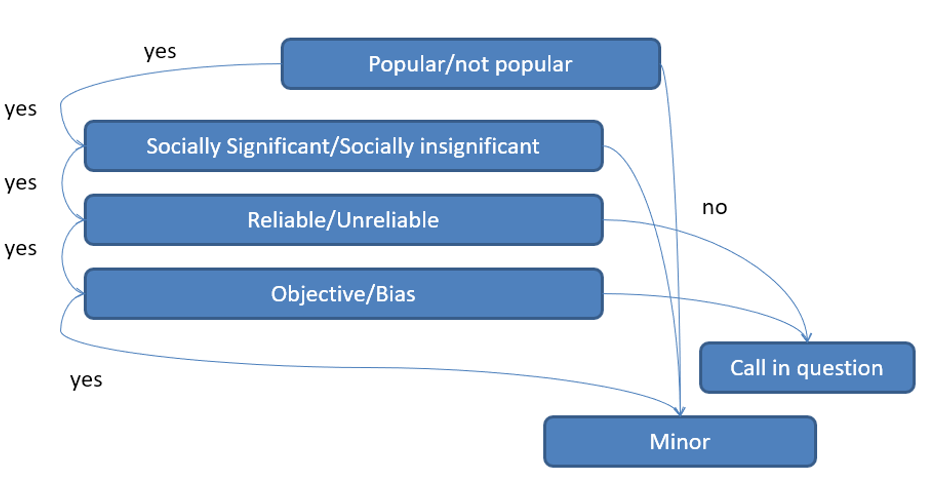


Figure 1 – Stages of classification of texts in order to identify "suspicious" publications [41].

In summary, the process of analyzing the text of the media is as follows:

* A list of parameters that determine whether the text belongs to one of the above classes is formed;
* Comparative importance of parameters is assessed;
* For each text to be considered, a parameter estimate is calculated;
* Estimates of parameters and their comparative importance are aggregated to obtain an estimate of the text's membership in a particular class;
* Using the received estimates of texts, estimates of media as a whole are formed.

To implement the described process, we determined the parameters of the texts and their mutual importance in the task of assigning the text to each of the above classes (Appendix D) and developed an algorithm for calculating parameter estimates based on the thematic model of the corpus of texts. The proposed approach is noteworthy in that the topic model created by cluster analysis (training without a teacher) is then applied in combination with expertly defined classes and features. Thus, the user (expert) sets the semantics of the required distribution, although the initial thematic analysis depends only on the corpus of documents.

The use of Bayesian approach makes it possible to form the probability of the corresponding hypothesis with incomplete information, having a part of the corpus of published texts. In other words, we can process only a part of the texts and get media estimates in the mentioned modalities, though reduced accuracy.

## Multimodal Media Evaluation Algorithm

The purpose of the algorithm is to use conditional probability distributions of documents to aggregate indicators of correspondence of papers to topics, topics to properties (dictionaries) and classes to obtain conformity assessments of the media in three modalities: topics, parameters and classes.

The initial data and the resulting conditional probability matrices are shown in figure 2, where MMS (mass media sources) is a set of text sources. Media (MMS) are the source of m papers, which are obtained through the application of data collection systems (process 2). The resulting document corpus M is divided into topic clusters T (process 1). Experts form classes C (process 4) and determine the properties or parameters of classes Q (process 3). The properties are described by dictionaries of words, expressions or procedures of their identification in the text (feature procedures).

Using a set of topics of the corpus, firstly, we obtain a discrete distribution of conditional probabilities of papers and topics - *p2(k | m)*, где , .

Secondly, we obtain a conditional distribution of dictionaries (parameters) and topics *p1(k | q)*, where, that is, we determine to what extent the parameter describes a particular topic.

Thirdly, using the analytical hierarchical process (AHP), we calculate the importance of parameters for classes (separately for each class) and obtain *p3(c | q)*, where . Then, using p1 and p3, we calculate the conditional distribution of topics by classes - *p4(k | c)*. Knowing the probability distribution of topics by classes (p4) and the probability distribution of the paper by topics (p2), we can calculate the distribution of the paper by classes *p5(m | c)*. In turn, the distribution of the paper by features or dictionaries- *p6(m | q)* depends on p1 and p2.

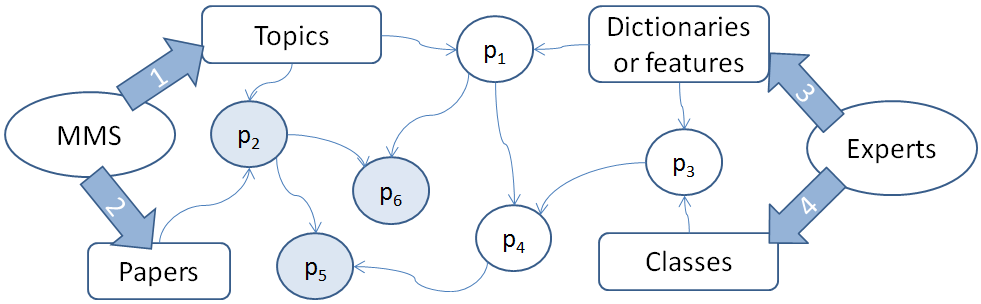
****

Figure 2 – Processes for determining conditional probabilities.

To calculate the probabilities p1 and p2, the Jacquard coefficient is used [42, 43].

The LDA algorithm is used to calculate the set of topics of the corpus.

After receiving these estimates for each paper, the media is evaluated using a chain of Bayesian rules. In this case, the conditional probability matrices p2, p5 and p6, highlighted in figure 2 by color, are used.

According to Bayes formula, we can express the conditional probability of the validity of hypothesis h when event e occurs in the form

(B1)

where *p(e│h)* is the conditional probability of the occurrence of event e at justice h, *p(h)* is the a priori probability of the hypothesis h, *p(e│∽h)* is the conditional probability e at non-justice h, *p(∽h)* is the probability that the event h is not true, which, in accordance with the formula for total probability, can be calculated as

(B2)

Thus, to calculate the conditional probability *p(h│e)*, it is sufficient to know the probabilities *p(e│h)*, *p(e│∽h)* and the a priori probability *p(h)*.

In terms of the media evaluation problem, conditional probabilities p(e│h) can be interpreted as probabilities of occurrence of some paper e^∈ E (where E is the set of papers of a particular media) provided that three kinds of hypotheses are true (each hypothesis for its modality):

1. h1: Media works in one of the topics from the set of topics T : H1={h1[1],...,h1[k]}
2. h2: Media publishes papers related to a class from the set C: H2={h2[1],...,h2[с]}
3. h3: Media is distinguished by features from Q: H3={h3[1],...,h3[q]}

In general, assuming that we have G modalities and accordingly G hypotheses H= {H1, H2,..., HG}, we can write a generalized algorithm for calculating their probabilities for some media

For all , ,

As a result of the algorithm we get a set of probabilities

for media . Here

Where Ng is the number of discrete elements describing the modality g.

The experiments are described in Appendix E and [48].

## Analysis of the social importance of thematic categories

Topics of social significance are usually determined by interviewing experts who are aware of the state of public opinion, which, in turn, can be obtained through sociological surveys of the population. [50].

From the point of view of definition of socially significant topics the most effective method is expert evaluation. Experts with knowledge of the specifics of the media and the peculiarities of social relationships and beliefs will be able to make a significant contribution to the evaluation of topics covered in the media. The results of topic modelling do not provide the name of the topic. The use of expert analysis will allow: (1) naming of the topics obtained; (2) defining topics of social significance.

To analyse the interest of different social groups in different topics, the most effective method is a randomized sociological study in the form of a sociological survey. The main goal of the sociological survey is to determine the interest of different social groups in different topics. The sampling should be carried out in accordance with commonly accepted procedures. The sociological survey will link the topics covered by the media to the social groups interested in them, to study the impact of textual information in the context of age groups, in the context of stratification on the principle of "city-village", regions, gender stratification, level of education and professional occupation of the respondents.

With regard to evaluation of the impact of media on society the most optimal research approach is experimental research, as this research design allows establishing and proving causal relationships. The main purpose of the experiment will be to establish the fact of the impact of the informative features (or a part thereof) selected in the course of the research as the features to be analysed within the information system. Presumably, such features will include: sentiment, manipulativeness, objectivity. This method will allow to establish the fact of impact of mass media on society, to confirm the necessity to consider the selected informative features during machine analysis of electronic mass media content, to measure the degree of influence of such informative features on society. A more detailed experimental research design will be developed and discussed during the further works on the research.

Within the given parameter, 2000 texts were studied to identify features of socially significant publications, i.e. publications related to socially important issues that can cause a wide public response, affect interests of society and/or cause destabilisation of society.

Among the language units marking socially significant publications, names of social infrastructure – a complex of objects, enterprises, organisations that provide functional life activity of the population, the formation of an intellectually developed individual and society were singled out. Also, this group of language units includes job positions and personalities names engaged in activities in socially important organizations. Language units marking socially significant topics were distinguished into a separate group.

As a result of the analysis, words and phrases signalling the social significance of the text are recorded in the texts and an appropriate dictionary including 2598 language units is developed. Upon completion of the works on refining and correction of the dictionary, a number of language units amounted to 2426 words and word combinations. The largest number of words and word combinations was identified in the group "Topics" – 1196, further, in descending order "Organisations" – 455, "Persons" – 418, "Job positions" – 325, "Objects" - 32.

In order to increase the potential of automated search of socially significant publications, 213 synonymic series to socially marked language units were compiled. Study results showed that news affecting socially important topics can be successfully identified with the help of language tools. As a result of work, language units were distributed among the proposed classes of words and quite clearly differentiated.

## Methodology of forming dictionaries of socially significant topics

In [51, 52], a method for the formation of topic dictionaries based on the analysis of semantic similarity using the co-occurrence matrix and the Word2Vec tool is presented. In this study, we tried to show different ways of automatically forming topic dictionaries. For the beginning we formed topic dictionaries of common topics: sports, politics.

The following algorithm of actions to identify socially significant news is proposed (Figure 3).

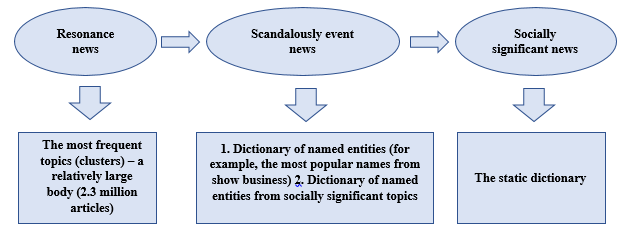


Figure 3 – Algorithm of actions to identify socially significant news

The next stage was to identify socially significant concepts and form their dictionaries. For accuracy of the concept of social significance, we relied on the definition and results of expert opinions. Three methods were used to form the dictionary. Dictionaries will be used to cluster text using machine learning algorithms.

## Improvement of the methods of assessing the impact of open information sources on society on the basis of a comprehensive assessment of various informative features

Based on the results of the analysis of the proposed methodology for evaluation of the impact of open information sources on society on the basis of published text information, a set of works on the refinement and improvement of the methodology was carried out [50].

The set of works includes: (1) formalization of rules and compilation of dictionaries; (2) experimental study of the completeness and correctness of the formed dictionaries and linguistic rules; (3) formation of the corpus of news text information and (4) its markup, (5) analysis and identification of patterns in the news text corpora; (6) the solution to the problem of imbalance in the marked corpora; (7) topic modeling, (8) development of algorithms for recognition of informative features and (9) development of methods for their calculation.

The Methodology is a complex of interrelated procedures for (1) collection of text information, (2) topic modeling, (3) analysis of text data for the recognition of informative features and (4) the issuance of reports based on the results of the analysis.

Figure 4 presents a general conceptual scheme for news texts evaluation.

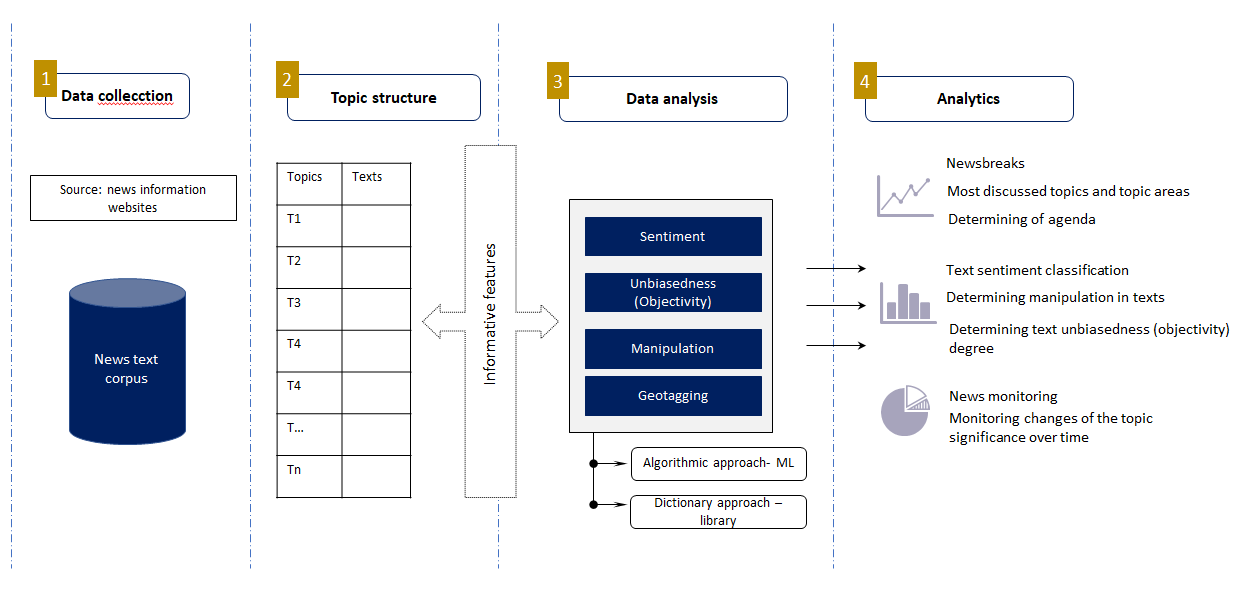


Figure 4 – Conceptual scheme for news texts evaluation

At the first stage, news texts are collected from Kazakhstani open text information web sources using automatic methods of text extraction or by downloading ready data collections (corpora).

At the next stage *topic modeling* is carried out. Most models are developed on the basis of Latent Dirichlet Allocation (LDA). The Latent Dirichlet Allocation model was proposed by David Blei in 2003[53]. This is a modern actively developing probabilistic tool used for data analysis tasks [54]. Topic models proved to be useful tools for solving this problem [55].

In the context of news media topic modeling, there are two main typical application options of it: analysis of media agenda and news frames analysis. Topic modeling itself, as a process, is one of the tools for automated analysis of a news texts corpus and evaluation of their impact on society [56].

Topic modelling in the analysis of news information can be applied to various tasks, both in research analysis in the field of Social Sciences, and in solving practical problems in the public and private sectors. The results of topic modelling may be useful in determining the media agenda, newsworthiness, issues of interest, changes in the significance of topics over time [55].

At the next stage, the recognition of informative features of sentiment, unbiasedness, manipulativeness in news texts is carried out. Recognition is performed using (1) algorithmic and (2) dictionary approaches. In particular, such machine learning algorithms as Random Forest, SVM, XGboost and KNN are used. The second approach uses dictionaries on the informative features compiled by experts in Linguistics.

At the final stage, data analysis results are visualized with issuance of analytic reports/visualization. Based on the results of the analysis, it is possible to determine the newsworthiness for various topic areas, the most discussed topics, the negative news content, manipulativeness and the degree of objectivity in the texts.

# DEVELOPMENT OF METHODS OF CALCULATING INDIVIDUAL CRITERIA FOR ASSESSING THE IMPACT OF OPEN TEXT INFORMATION SOURCES ON SOCIETY

Previously, the project was carried out a set of works to determine the composition of informative features.

## Determination of the composition, classes and significance of informative features

Within the framework of this project, a set of works was carried out to determine the composition, classes and significance of informative features for subsequent algorithmic and / or expert analysis in order to build appropriate methods of classification and evaluation of the potential impact of the analyzed text information on both the individual and society as a whole [57, 58, 59, 60].

The review of scientific and analytical literature allowed to determine indirect informative features that can be applied in the analysis of news texts. 16 informative features combined into 4 informative criteria were proposed: (1) reliability; (2) objectivity or bias; (3) sentiment; (4) public response (popularity) of a publication. In particular, (1) a references to the competent source in the publication; (2) reference to the primary source; (3) correspondence of the publication title to the content of the publication; (4) coverage of the same event by different publications: cross-checking, the presence of discourse with other articles; (5) the presence of the author in the article; (6) availability of verifiable facts in the article; (7) the reputation of the media which posted the information; (8) the presence of manipulative techniques; (9) the Politicisation of the publication; (10) the presence of the expressed personal opinion of the author of the publication in the media text ; (11) generalizations in the text; (12) the sentiment (positive/negative); (13) the number of views (14), shares (15), comments on the publication; (16) the sentiment of comments (positive/negative).

To determine the significance of the proposed informative features and criteria, a comprehensive assessment of their significance was carried out. The procedure of comprehensive assessment of significance includes: (1) sampling the corpus of publications, (2) building indexes to measure informative features, (3) building a statistical model, and (4) building a machine learning model. Double-stage systematic cluster sampling was carried out to form corpora. Expert markup of the selected texts was carried out to determine the significance of informative features. The markup was carried out on 16 informative features. An additive approach was used to construct the indices. This approach was chosen due to the fact that the data structure and the availability of theoretical concepts allow using an additive approach to the construction of indices, the purpose of which is to adequately measure the proposed criteria for assessing the impact of publications on society. During this procedure, three indices were built: unbiasedness index, reliability index and sentiment index. The obtained results of statistical modelling showed that the proposed informative criteria of publication can further serve as a framework for creating an automated machine learning model. These results were confirmed experimentally by bootstrapping, which showed a slight deviation from the model results.

Thus, the proposed Methodology defines informative features and criteria, assesses their significance for public response, develops a scheme for evaluating textual information and proposes an approach for recognizing informative features in texts.

## Topic modeling

## Topic modeling based on BigARTM and GenSim libraries

Topic modeling algorithms make it possible to identify existing topics in the processed publication corpus [61]. Subsequently, information on the topic structure of the analyzed publications can be used as the main or additional features in high-level processes of classification, clustering, machine learning and expert analysis.

To date, the most popular topic modeling algorithms are:

* 1. BigARTM is a library of algorithms for efficient stream parallel implementation of probabilistic topic modeling based on additive regularization [62].
  2. GenSim is a topic modeling library based on latent semantic analysis (LSA, LSI, SVD) and Direchle latent allocation (LDA) algorithms [63].

Both of these libraries were actively used by us to decompose a large analyzed body of media publications into topics. BigARTM as a topic modeling option was used in the information system being developed. Each library has its advantages and disadvantages. Their main drawback is the high computational costs and long time in the analysis of corpuses containing more than several million documents. To overcome the above disadvantages, our group has developed its own algorithm of topic modeling based on the cluster approach.

## The topic modeling algorithm based on clustering methods

Within the realization of this project, an own topic modeling algorithm based on clustering methods was developed. The developed algorithm is based on the following assumption: the context determines the thematic and (or) semantic proximity of words. That is, the closer the contextual environment of the words being compared, the greater their thematic / semantic proximity. Indeed, it is obvious that synonyms should have the same contextual environment. And vice versa, the less semantic proximity of words, the more their context differs. The contextual environment is defined as a “bag of words” constituting a fragment of text containing the analyzed word. A “precise” assessment of the proximity of contexts provides an assessment of semantic (meaning) proximity, and relatively rough - a thematic proximity. Thus, the topic can be understood as a relatively rough assessment of the proximity of the analyzed contexts. Adjusting the degree of roughness results in change in the number of topics.

Consider corpus of analyzed documents. This corpus is used to form its target dictionary V of size n, which is used for subsequent analysis. If the dictionary is too large, then it can be shortened by the decreasing the level of significance of the included words. The target dictionary can be formed both from words and from phrases. Based on the corpus, a symmetric square matrix C of cooccurrences of words from target dictionary in the documents can be evaluated. We assume that every i-th row vector of the matrix C is a sparse vector representation of the i-th word from the target dictionary V. A square symmetric matrix D of size n x n is calculated using cosine distance metrics. Each element of the matrix D with coordinates (i, j) contains the value of the cosine distance between the vector representations of the i-th and j-th words.

When considering the problem of topic modeling, words in the target dictionary *V* are objects to which cluster methods are applied. To clarify the essence of the algorithm, lets introduce the following definition: A clot (protocluster) is a subset of objects from *V* where all pairwise distances between them do not exceed a certain threshold *d1*. From the clustering point of view, clots should include as many objects as possible meanwhile the sum of the distances between all pairwise combinations of objects included in it should be as small as possible. Moreover, the same object, if it satisfies the restriction on *d1*, can be included in several clots at the same time. In other words, the process of forming one clot does not affect the process of forming any other clots. If an object is included in one clot, nothing prevents from subsequently including it in another clot if the restriction on *d1* is satisfied.

Since the same objects can be assigned to different clots, then different clots can have all or partial common objects. By estimating the degree of similarity of two different clots, conclusions can be drawn about the degree of their relatedness. The degree of relatednesss between two arbitrary clots is conveniently to estimate by the modified Jaccard index. By introducing the threshold value of the Jaccard index (*d3*), decision about the relatedness or unrelatedness of two arbitrary clots can be made.

A top-level cluster is a sequence (graph) of related clots.

The essence of the proposed algorithm is as follows: the algorithm splits the entire set of objects (words) into mutually intersected clots (protoclusters) with a restriction on the maximum pairwise distance d1. When clots are formed, the neighborhood is defined by the radius *d2*. Then, the evaluated clots are merged based on their level of relatedness where the threshold of relatedness is set by *d3*.

The core is the procedure for finding a clot in a given neighborhood of the object.

The neighborhood of the object is set by *dm* - the square matrix of pairwise distances of the analyzed object to all objects in its neighbourhood. The maximum intracluster distance is set by parameter *d1*. In the formed clusters, the distance between any two objects does not exceed *d1*. The start\_ind parameter is the index of the object that is used as the starting point of process of forming clot. As a result, the procedure returns the indices of all objects included in the clot:

def single\_clot(dm, d1, start\_ind):

n = dm.shape[0]

if start\_ind < 0 or start\_ind > n-1:

raise ValueError('start\_ind is out of bounds')

R = np.array([start\_ind])

C = np.delete(np.arange(n),start\_ind)

while len(C) > 0:

C = C[np.sum(dm[R][:,C] <= d1, axis=0) == len(R)]

if len(C) > 0:

dist\_sum = np.sum(dm[R][:,C], axis=0)

best\_ind = np.argsort(dist\_sum)[0]

R = np.append(R,C[best\_ind])

C = np.delete(C,best\_ind)

return R

The following procedure starts the process of forming of clots for each of the objects in its neighbourhood - circle centered on the object and with a radius d2. The maximum allowable pairwise distance inside the clot is set by parameter d1. Input data are: D - matrix of pairwise interobjects distances between objects; use\_medoid - if True, then medoids are used as points of clots growth. Otherwise, the centers of the circles. Each clot is specified by a set of indices of objects included in it. As a result, the procedure returns an array in which each i-th element is a clot in the neighbourhood of the i-th object.

def all\_clots(D, d1, d2, use\_medoid = True):

n = D.shape[0]

global\_inds = np.arange(n)

clots = [np.array([0])] \* n

for i in nb.prange(n):

local\_inds = global\_inds[D[i] <= d2]

if len(local\_inds) > 0:

dm = D[local\_inds][:,local\_inds]

if use\_medoid:

start\_ind = np.argmin(np.sum(dm, axis=0))

else:

start\_ind = np.where(local\_inds == i)[0][0]

clot = single\_clot(dm, d1, start\_ind)

if len(clot) > 0:

clots[i] = local\_inds[clot]

else:

clots[i] = np.empty(0,dtype=nb.int64)

else:

clots[i] = np.empty(0,dtype=nb.int64)

return clots

Further, the procedure for searching for unique clots is performed. Only clots of certain size are included in the result – the size of clot must not be less than *min\_size*.  
 A set of unique clots is represented as *list()* of *set(),* where *set()* is the set of indices that constitute each clot. The same object can be included in several clots:

def unique\_clots(clots, min\_size = 1):

R = set()

for clot in tqdm(clots):

if (len(clot) > 0) and (len(clot) >= min\_size):

R.add(frozenset(clot))

return np.array([set(v) for v in R])

At the final stage, the procedure of merging the clots in a single top-level cluster is performed based on the level of their relatedness. The parameters of the procedure are *clots* - an array of clots in which the *i-th* element is the set of indices of the objects that constitute this clot; *n\_jobs* - the number of parallel processes that are used to calculate the matrix of distances between clots. If *n\_jobs* = -1, then all available processes are used. At the same time, clots are merged to the cluster if their modified Jaccard index value at least *1-d3*. Clots can be related either directly or indirectly. In the case of direct relatedness, the clots directly have a sufficient fraction of common objects. An indirect relatedness between two clots is established if there is a chain of directly related clots between them (i.e., they indirectly relate through a chain of directly related clots). Thus, all sets of directly or indirectly related clots form a single cluster. Clots can be considered as the first level fuzzy clusters, which, as they cross, are combined into higher-level clusters with a relatively more complex spatial structure. The point of the clustering method is to divide the entire set of objects into clusters - simple mutually intersecting subsets of objects with the subsequent specification of relatedness (level of intersection) between them, which allows to approximate clusters having complex spatial structures by simple mutually intersecting clots. As a result, the function returns an array, each element of which is a cluster consisting of indices of objects included in it:

def clots\_binding(clots, d3, n\_jobs = -1):

# a, b – Clot indices, between which the distance is calculated by the modified Jacquard measure

def jaccard\_dist(a, b):

return 1 - len(clots[np.int(a[0])].intersection(clots[np.int(b[0])])) / min(len(clots[np.int(a[0])]),len(clots[np.int(b[0])]))

m = len(clots)

inds = np.reshape(np.arange(len(clots),dtype = np.int), (-1, 1))

dm = pairwise\_distances(inds, metric=jaccard\_dist, n\_jobs = n\_jobs)

labels = np.arange(m,dtype = np.int)

inds = np.arange(m,dtype = np.int)

for i in tqdm(range(m)):

old\_labels = labels[dm[i] < d3]

for j in old\_labels:

np.place(labels, labels==j, labels[i])

R = np.array([np.array(list(set().union(\*clots[labels == i]))) for i in np.unique(labels)])

return R

In addition, various methods of weighing words in estimated topics, that are not inside estimated topics (the level of their relatedness to topics), as well as topics themselves were implemented:

1. Weighing objects in the cluster (method 1.1): the object is assigned the greater weight the smaller its total distance to other objects in the cluster.
2. Weighing objects in the cluster (method 2.1): the object is assigned the greater weight the greater its total joint frequency with each object in the cluster.
3. Weighing the total set of objects in dependence of level of matching to estimated clusters (method 1.2). An object is assigned more weight the less its total distance to the objects of the cluster being compared to. Returns a matrix of size *NxM*, where *N* is the number of clusters, *M* is the number of objects, matrix elements are the level of matching (weight) coefficients between the *i-th* cluster and the *j-th* object.
4. Weighing the total set of objects as they fit existing clusters (method 2.2). An object is assigned more weight the greater its total joint frequency with each of object of the cluster being compared to. Returns a matrix of size *NxM*, where *N* is the number of clusters, *M* is the number of objects, matrix elements are the level of matching (weight) coefficients between the *i-th* cluster and the *j-th* object.
5. Weighing each cluster on the level of its correspondence to the total set of objects (method 1.3). The cluster is assigned more weight the smaller the average distance to all objects of the initial set. The pairwise distances between the objects are represented by the matrix *D*.
6. Weighing each cluster on the level of its correspondence to the total set of objects (method 2.3). The cluster is assigned the greater weight the greater the average frequency of occurrence of objects of the cluster with all objects of the initial set. The pairwise frequencies of occurrence between objects are presented in the matrix.

Two support procedures have been implemented:

1. Selects the first n keywords for each of the topic-cluster from the words included in the cluster based on the maximum of their weight assigned by methods 1.1 and 2.1.
2. Selects for each cluster the first n additional objects (words - candidates for inclusion in the cluster) from all objects not included in the cluster under consideration. The selection is made according to the assigned weights by methods 1.2 and 2.2.

As an experiment, a topic modeling of a corpus of news publications produced by one of media recourse - the tengrinews.kz media was carried out. The analyzed corpus contains 199 362 publications over a ten-year period from 2008 to 2018. The following values of the algorithm parameters were used: d1 = d2 = 0.75, d3 = 0.5, min\_size = 20, use\_medoid = False.

The performed topic modeling identified 204 topics in the analyzed case. Both the topics themselves and the words included in them were weighted by the above described methods. In addition, for each topic, a heading (name) was automatically generated consisting of the three most significant words in the topic. Also, for each topic, three candidates for inclusion were selected automatically among the words that were not included in the topic (according to methods 1.2 and 2.2). The results of topic modeling can be downloaded at the following link [64].

Based on the assigned weights to topics (methods 1.3 and 2.3), the topics were ranked in descending order by the level of their compatibility to the corpus of the analyzed media. The topics that are at the beginning of the lists are "central" to the analyzed media (in this case, tengrinews.kz) and then they are arranged in descending order. The most “uncharacteristic” topics for the media are placed at the very end. Under the "central" topic is specified the topics that are best compatible with the full range of topics. That is, they are core to other topics.

In the application two files are attached [64]. They have the same set of topics, but their weighing and ranking was carried out by two different methods:

1. Based on the distances between the vector representations (method 1.3);
2. Based on the frequencies of сo-occurrences (method 2.3).

The ranking results for the topics according to method 1.3 are illustrated in Appendix F.

As can be seen in a subjective visual assessment, the results of both the topic modeling itself and the ranking of words and topics are quite interpretable, coherent and meaningful.

Currently a complex of works has begun on an objective comparative assessment of the quality of the developed topic modeling algorithm based on standard criteria that most proven themselves in the field. Preliminary results show that the developed algorithm has a significant advantage both in terms of time and quality of the results. The final results will be published and presented in a scientific article.

## Automatic setting of the tonality of texts by the method of their conceptual analysis

One of the significant informative features for the classification of textual information is its tonality. Within the framework of this project, various approaches for the numerical assessment of the tonality of the analyzed texts were considered, analyzed and developed [65, 66, 67]. In particular, the method of automatic establishment of tonality of texts by the method of their conceptual analysis was developed and tested [67].

The article [67] describes the solution to the problem of creating linguistic tools and methods of automatically determining the tonality of news messages related to the quality of life of an ordinary citizen. An approach to solving the problem was defined, software and methods for automated creation of object dictionaries and dictionaries of evaluation predicates, as well as dictionaries of evaluation measure modifiers were developed. The experiment confirmed the correctness of the proposed methodology of assessing events covered in news reports and the performance of the software complex.

This technique, with an appropriate selection of event assessment objects, can be used to create tonal portraits of specific authors from the totality of their publications, as well as tonal portraits of various news aggregates on the totality of events covered by them in a specific time interval.

The proposed method, based on the methods of automated compilation of dictionaries based on document texts, methods of semantic-syntactic analysis and use of ontological resources, is universal because it allows for detailed semantic analysis of document collections according to various multidimensional criteria. A complex of tonal dictionaries of large volume and means of formalizing texts will allow this analysis to be carried out with fairly high accuracy.

## Generalization

One of the important signs of manipulativeness in evaluating text publications is the degree of generalization (generalization of utterances - deduction). Within the framework of this project, a set of works was carried out aimed at developing methods of numerically assessing the degree of generalization, their algorithmic implementation and testing on the formed body [68,69,70].

In particular, a definition of the concept of generalization is given, the phenomenon of “generalization” in the texts of the media is analyzed, and an attempt is made to synthesize a linguistic model for its recognition and subsequent numerical evaluation. There are studies that view generalization as a sign of bias or as an ideological strategy that affects readers' opinions. However, no studies on automatic recognition of generalization in texts have been revealed. An annotation scheme is proposed, on the basis of which linguistic rules for manual definition of generalization and its subsequent automatic recognition in texts are determined. The experiment to detect generalization in the texts of the official media was conducted using the rules.

## Degree of informative content of publications

The next no less important feature is the degree of information saturation of the analyzed text publications. Within the framework of this project, a method for automatic numerical evaluation of the degree of informative content of publications was proposed and developed [71].

## The classification of publications by type, genre and style

## 

The publications in the media have different goals. The purpose of ones is only to inform the reader about the event: who did, what, where and when it happened. The others not only inform about the event, but also add the background information to the details, summarize the information, make the conclusions. Such publications contain both facts and opinion of the author. The others only start from a specific fact and try to interpret events in the way they want, perhaps by manipulating the reader's opinion.

An approach for automatic recognition of such types of publications based on the selection of features is proposed [72, 73]. We investigate and identify a number of features inherent in three types of publications. In addition, on a small text corpus we investigate their applicability for automatic classification of texts by their types.

The small corpus was formed, which is sufficient to investigate the features and to conduct an experiment. The sources have a fundamentally different focus: sports news site, conspiracy site, a site with popular science articles, and a site about business. We have selected the articles from these sources in which the typification of publications is most visible: informative publication, analytical or artistic.

Our approach can be used for different kinds of tasks. For example, to classify the fake news or search for propaganda, or to search for the latest news, the financial market news or other information.

In our work, we have collected a corpus from open sources of official and semi-official media and divided it by type; we investigated and identified the features; and investigated with what accuracy the features automatically recognize types.

We suppose that the conspiracy and analytical articles may differ from informative ones by such statistical features as the average number of words in a sentence, the number of words in a document, the number of sentences, the average number of paragraphs. The news articles, in turn, are more informative, contain more numbers and named entities. We also assumed that the quotes, as a sign of direct speech, are found most often in analytical and conspiracy texts. The lexical features include the frequency of verbs, the frequency of adjectives and nouns in a document. We expected that a large number of verbs in the text would be associated with more informative news texts. Earlier in our work, we assumed that the plural nouns and the presence of quantifier-words *like everything, everyone, anyone, always, forever, never, constantly, no one, nothing* are the signs of generalized texts and it was interesting to analyze their distribution in genres. Finally, we selected twelve of the stylistic and lexical characteristics.

Then we experimented with different feature variations to see which feature set showed the highest accuracy, we trained a model for all feature sets. The combination of four features: the number of quotes, the figures and the numbers, the average number of words in a sentence, the frequency of verbs and adjectives showed the highest accuracy of 85%. For the aggregate of all features, the algorithm showed 67%. Thus, the combination of four features in general gives a higher recognition accuracy than one feature or all features.

The additional experiments on the recognition of text sources showed that the algorithm recognizes the news articles with an accuracy of 88.5%, the conspiracy articles with a result of 100%, the business articles with a result of 67.7% and the popular science articles with a result of 75%. The higher results of automatic recognition of news and conspiracy articles may be the result of the fact that the syntactic features available in them are more expressed, in our opinion. While the popular science and business articles are differ more in lexical composition. This requires the additional study and the introduction of new features, such as the frequency of occurrence of unique words.

## The probable destructiveness

Within the framework of this project the analysis of methods providing the detection of destructive messages in the Internet environment was carried out [74, 75]. It is shown that the approaches based on the analysis of syntactic regularities; the analysis of semantic information contained in the text and its correlation with the text corpus; the crowdsourcing methods; the identification of regularities of users' behavior in social networks; the consideration of additional information, etc. are used to identify them. Good results can be achieved under certain conditions: the access to traffic social networks and other online news resources, the opportunities to organize or to get results from crowdsourcing, etc., With restrictions on the fulfillment of the listed conditions the detection of destructive messages can be performed based on indirect signs. The results of identifying destructive news in the corpus of news messages posted on the websites of the Republic of Kazakhstan are presented.

The results of the identification of destructive news in the corpus of news messages in Russian, posted on the websites of the Republic of Kazakhstan using pre-compiled dictionaries containing words indicating the following properties of the text are presented: the presence of verifiable facts, the politicization,the recall to action, negative or positive tonality, the manipulativeness.

The results of the algorithm show that in some cases it gives a good degree of coincidence with expert estimates. On the other hand, it allows to rank the texts in situations where the texts are very similar in their characteristics. The above texts are taken from one fairly reliable source, so the reliability and the objectivity of the texts are close. However, this approach requires a very careful compilation of dictionaries and the usage of additional algorithms to determine a wider range of parameters of texts.

## Assessment of severity in the analyzed text of information criteria based on the dictionary approach

One of the methods for assessing the severity of informative criteria (signs) in the analyzed text is the dictionary approach [50].

The preparatory stage which determines the solution of tasks for automatic evaluation of publications in the media and recognition of selected informative features is a set of works for the creation of a dictionary array, selection of language units and methods (rules) of their use in news texts. In this regard, this work aims to develop methods to formalise rules and dictionaries for recognition of informative features in marked-up corpora and determination of the degree of their expression.

12 informative features for the recognition of which it is necessary to form dictionaries and formalize linguistic patterns and rules were identified.

2,000 texts from 5 informational Internet news sources became an object of analysis: “KazakhSTAN 2.0”, “Central Asia Monitor”, “zakon.kz”, “Radio Azattyk”, “Tengrinews.kz”.

The following tasks were set and solved to achieve the goal:

1. Conducting an in-depth analysis of the formed corpus of media publications on 9 informative features (3 features in 2018, 9 - in 2019).
2. Markup of text corpus according to outlined informative features.
3. the degree of generalisation (mild, strong, absent); Formation and systematisation of dictionaries by certain informative features from analysed texts.
4. Identification of patterns to formalise the rules of separate informative features recognition.
5. Obtaining expert opinions on the feasibility of identified patterns in the project.
6. Description of methods for rules formalisation based on linguistic analysis of texts.

Thus, the following results were obtained on the basis of summary analysis and correction of results of linguists' work on the study and a detailed marking of 2000 media texts:

1. A dictionary consisting of 8961 language units (3091 words and 5870 word combinations) for the "sentiment" informative feature.
2. A dictionary consisting of 166 language units (30 words and 136 word combinations) for the "politicisation" informative feature.
3. A dictionary consisting of 1543 language units (652 words and 891 word combinations) for the "presence of manipulative techniques" informative feature.
4. A dictionary consisting of 2639 language units (6 parameters) for the "social significance" informative feature.
5. A dictionary consisting of 1398 words and 5501 word combinations for the "Kazakhstani content" informative feature.
6. A dictionary of competent sources (named entities) is formed: organisations, job positions and personalities), consisting of 2378 words and word-combinations, 88 linking words used in the reference to the source.
7. A dictionary consisting of 1202 words and 4221word combinations for the "Presence of verifiable facts in the context of subject/object – action – place – time" informative feature.
8. A dictionary consisting of 4001 words and 6409 word combinations for the "keywords" informative feature.
9. A dictionary for the "bias" informative feature– 32 words and 423 word combinations.
10. A dictionary consisting of 89 words and 1427word combinations for the "call to action" informative feature.
11. The patterns for formalising rules to identify informative features of sentiment, politicisation, presence of manipulative techniques, references to competent sources, presence of verifiable facts in the context of subject/object – action – place – time, correspondence of the title with the content of the publication, bias and a call to action were identified.

Refining, systematisation and additions to the above dictionaries were performed.

Also, during the analysis of the corpus, the tests were marked with the selection of text fragments, including markers of informative features.

According to the results of the patterns identified during the linguistic analysis of the corpus, expert opinions on their feasibility within the framework of the project were obtained from Russian and Kazakhstani experts.

Methods of formalising rules based on the patterns revealed during the linguistic analysis of the corpus were also described.

Results of the study generally confirmed the necessity and possibility of identifying specified informative features at the language level for automated recognition of the desired text parameters.

The number of informative features, uniquely determined by linguistic means, includes sentiment, social significance, Kazakhstani content, references to competent sources and call to action.

At the same time, as shown by the results of work at this stage of the study, a part of the informative features selected for the evaluation of media publications should be determined and recognised using a combination of linguistic methods and machine algorithms. Such features include politicisation, presence of manipulative techniques, keywords, bias. As part of this work, it is necessary to artificially simulate structures built according to linguistic rules and patterns on the basis of models and word usage defined in the course of in-depth analysis of news texts. For example, analysis of the word "development" in combination with other words, analysis of particle "not" in combination with verbs with positive and negative sentiment, with adjectives, pronouns, nouns, the particle "not" in combination with other parts of speech, repetition, etc.

For more effective operation on determining the presence of the Kazakhstani content in the text, in addition to 2000 items for Kazakhstani persons, place names, household items, etc. due to the limited volume of the analysed material it is necessary to include a universal list of names of geographical objects, famous persons, officials, names of Kazakhstani state bodies, institutions, organisations and business structures into the material for automated recognition. Such dictionaries will significantly increase the "recognition" of the Kazakhstani content in the text. The results of the research were published in the [76, 77, 78].

# DESIGN AND DEVELOPMENT OF AN INFORMATION SYSTEM OF EVALUATING THE IMPACT OF OPEN TEXT INFORMATION SOURCES ON SOCIETY

## Purpose, goals and objectives of the developed

A system of assessing the impact of open text information sources on society or the Information Trends Analysis System (NLPMonitor) is intended for collecting, classifying, analyzing text publications of the Kazakhstani segment of electronic media for making socially significant management decisions in stimulating sustainable personal development by departmental government organizations, such as Ministry of Education and Science, Ministry of Labor and Social Protection, etc., and/or commercial organizations.

The purpose of creating the system: to develop a system to stimulate the sustainable development of the personality, which in real time shows the user relevant socially significant topics in a given social sphere or field with an assessment of the impact of open text information sources on society using big data technologies.

Criteria of assessing the impact of open text information sources on society:

* degree of positive orientation of the publication;
* degree of negative orientation of the publication;
* degree of generalization;
* degree of politicization;
* degree of subjectivity (degree of expression of the author’s opinion on the subject of publication);
* degree of bias (bias, partiality);
* degree of relevance ( degree of relevance of information to date);
* degree of originality (uniqueness);
* degree of attractiveness of the publication (as far as the publication can attract the attention of potential readers);
* degree of Kazakhstan content;
* degree of provocation (manipulativeness);
* degree of connection with other publications on the issues discussed;
* degree of contradiction to other publications.

System tasks:

1. Collection, storage and selection of unique publications from the Internet space to the database system
2. Distribution of publications by topics: clustering, classification, definition of topic combinations, ranking and filtering (by social spheres, areas, industries, etc.)
3. Definition of newsworthy occurrence
4. Calculation of degrees of informative features of a publication, such as: veracity, resonance, tonality, objectivity, media involvement
5. Definition of information trends
6. Formation of all kinds of analytical graphs, infographics, statistics for further management decisions

## Architecture of the information system

As part of the tasks, an information system architecture was developed (Figure 5) for monitoring, collecting, storing, processing, visualizing and analyzing large volumes of text data [80], in particular, data from the media space of Kazakhstan NLPMonitor, as well as software implementation of the developed architecture, deployed on existing hardware and put into operation.

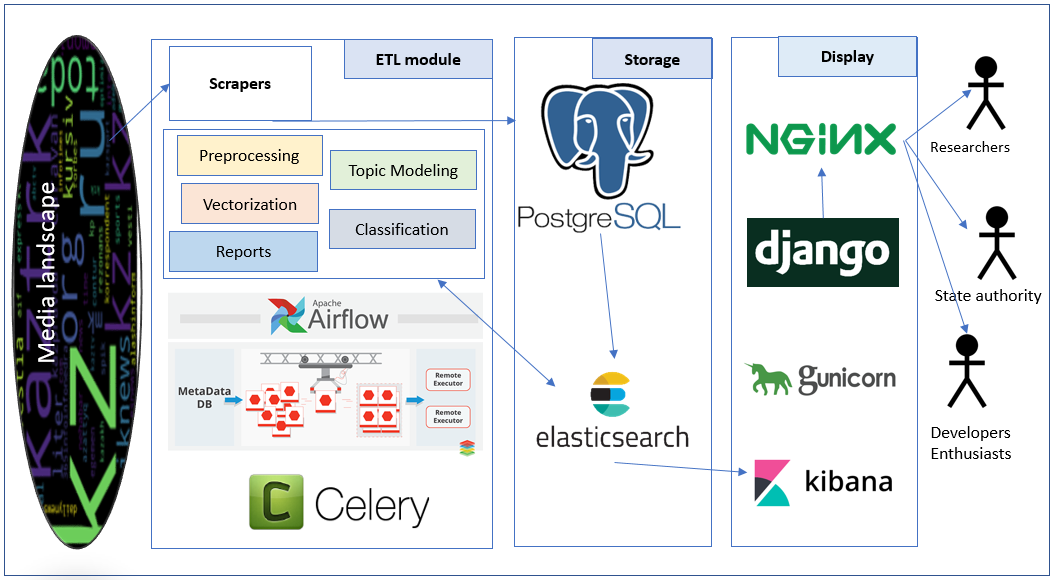


Figure 5 - Architecture of the NLPMonitor information system

Due to the need to store and process large amounts of data, as well as the use of resource-intensive algorithms of machine learning (including deep learning), the architecture was developed taking into account the need for scalability and distribution. We consider the main components of this system, shown in Figure 5:

* *Storage*

The system provides three types of storage:

1. *PostgreSQL - acts as a persistent storage for structured data. The main data types stored in this database are:*
   1. News and metadata;
   2. Processed data at the level of different basic units of analysis (token / word / phrase/sentence/text), including vectorization, lemmatization results, cleaning, etc .;
   3. The results of topic modeling;
   4. News classification results for various features (tonality, politicization, social significance, etc.) [81].
2. *ElasticSearch - in-memory NoSQL storage designed for storing unstructured or weakly structured data, as well as quick search (including full-text), filtering and streaming access. ElasticSearch performs several functions:*
   1. Primary storage for end-user access, search, and filtering;
   2. Primary storage for ETL (Extract-Transform-Load) data processing – including recording any intermediate results in free form;
   3. Storage for caching certain calculation results needed to build dashboards and reports in the system.

ElasticSearch duplicates data stored in PostgreSQL as a persistent storage, since ElasticSearch is an in-memory database without guarantees regarding data persistence and integrity.

1. *Redis is a fast key-value storage used for caching individual pages and elements, as well as for caching authorization sessions. Redis stores service data, as well as a cache of pages and elements that are frequently accessed.*

All three main storage systems can be easily scaled to several separate machines, both sharding and replication are supported, while ElasticSearch and Redis show a near-linear increase in performance with horizontal scaling.

* *Data Processing*

In developing the data processing architecture (ETL), the following key needs were identified:

1. possibility of parallelization of calculations, including on several machines;
2. flexible scheduling of various data processing tasks;
3. ability to monitor the performance of tasks in real time, including operational reporting of exceptions;
4. flexibility in the tools and technologies used.

During the analysis, Apache Airflow, an open source software platform, was chosen to meet all these needs. Let's consider the main components of this platform:

1. *Airflow-worker* – the main component that performs data processing. It can be scaled horizontally, including to separate servers/cloud VMs. In the current version of the architecture, the necessary dependencies are built into the Airflow-worker container image in advance, however, in principle, the process of dependency injection can occur in various ways, including by dynamically obtaining Docker containers from public or private repositories.
2. *Airflow-scheduler* is a component responsible for assigning tasks to Airflow-workers in the order defined by Airflow DAGs. Airflow DAG is a non-cyclic directed graph that describes the procedure for performing certain tasks, and also contains information about the schedule, priorities, behavior in case of exceptions, etc.
3. *Airflow web server* – a web interface that allows to monitor and control the progress of tasks.

A separate PostgreSQL cluster is used to store service data, such as task execution states. Celery+Redis is used to start and track the progress of tasks.

The primary component for task execution in Airflow-DAG-Directed Acyclic Graph is a directed non-cyclic graph that describes dependencies, execution order, schedule, and priority of various tasks Unlike similar systems, Airflow DAGs are fully described in the standard Python language, which allows, among other things, creating dynamic tasks and task groups, for example, for parallelizing computations, meta-optimization, and tasks requiring adjustment of the execution structure depending on external data ( e.g. scraping). To add DAGs to the battle server, research members can push to a separate git repository (https://github.com/KindYAK/NLPMonitor-DAGs), after which the main server will automatically receive a new code, with the possibility of manual or automatic launch on a schedule.

* *Interface*

The system interface is a standard HTML + CSS + JS website with access via HTTP. The web application is implemented in the Python Django framework; Gunicorn, the Nginx reverse proxy acts as the web server. The web application has access to both PostgreSQL persistent storage and ElasticSearch. The web application implements a number of pages for filtering, searching, accessing various dashboards and reports.

* *System deployment*

NLPMonitor is implemented as a modular system based on the principle of containerization. Containerization (virtualization at the OS level), in contrast to classical virtualization (at the hardware level), allows isolation and scalability of individual system components with minimal loss in performance and the possibility of full use of all hardware elements of the system.

NLPMonitor is developed on the basis of Docker CE Linux containers, local deployment for development purposes takes place using docker-compose, the main server uses Docker stack. In the official system repository (https://github.com/KindYAK/NLPMonitor) in the configs folder there are configuration files for deployment using docker-compos /docker stack, and in the folders airflow-worker, visartm and web / docker there are Dockerfile and files necessary for assembling custom containers. At the moment, all versions of pre-assembled custom container images are stored in a private Docker Registry deployed on a separate server.

One of the advantages of containerization is the ability to easily scale system elements, including storage sharding. It is also planned to use the Kubernetes container orchestrator to implement the ability to easily scale the system to a computing cluster with automatic load balancing between servers in the cluster.

The work on the formation of a database of media articles (corpus) has been carried out. Algorithms of automatic collection of articles from 54 news sites of the Kazakh media segment have been developed and implemented. The work on aggregation and subsequent processing of the collected articles has been carried out. As a result, 2 387 974 unique articles have been collected and processed. The news corpus of the Kazakhstani media segment was formed with the help of specially developed software. A total of 54 agencies (news sites) were used, of which: news agencies - 29, print - 15, television - 8 and Internet - 2.

In total, articles were collected: 2 388 534 unique on sites, 2 387 974 unique on the whole corpus. The data volume was 9.5 GB.

## Definition of necessary and sufficient functionality of the developed information system for the qualitative solution of tasks of the project

When designing the information system, the necessary and sufficient functionality of the developed information system was determined for the qualitative and sufficient solution of the project tasks. Although necessary and sufficient functionality is defined here, but there will always be "redundant" functionality in the system, since in order to add new and expand existing functionality, it is necessary to test different methods/algorithms and compare them in terms of efficiency, and only then implement the most effective ones, thereby reducing the redundancy of the system functionality.

Below is a list of necessary and sufficient functionality of an information system for the qualitative decision of tasks of the project (in the future it can be extended, but it is important to determine the functionality with the lowest cost decides the project's objectives):

1. Automatic collection (scraping) of publications from open text information sources on the Internet and the formation of the analyzed corpus on their basis. Scraping should be carried out according to a predetermined list of sources.
2. Formation of a dictionary of basic units of analysis (BUA), which may include: individual words, stable phrases, named entities and should not include uninformative vocabulary (prepositions, introductory words and their combinations, etc.). For the purpose of optimization, the BUA dictionary can be filtered by lexical and morphological features (for example, include only certain parts of speech in the dictionary; add / remove named entities, impose restrictions on the frequency of included elements, etc.).
3. Classic Topic Modeling based on the previously formed BUA dictionary.
4. Formation of the dictionary of thematic combinations (DTC): Combinatorial analysis of mutual compatibility of topics in publications of the analyzed corpus. Since each individual publication in the corpus can contain several topics, it is possible to analyze what are the combinations of topics. Each individual topic is assigned a unique identifier. According to the results of the analysis, we get a list of possible combinations of identifiers indicating the number of publications corresponding to this combination of topics.
5. Break the formed DTC on applied classes. For example, you can select the following as application classes:

* social significance;
* promote sustainable development of the individual;
* social tension;
* policy;
* Kazakhstan content;
* education;
* development, science, innovation, technology;
* crisis phenomena and incidents;
* positive;
* negativity;
* etc.

Partitioning by application classes can be carried out expertly, on the basis of pre-trained models or by dictionaries. Also, when splitting, you can specify / calculate the weight of this thematic combination in this application class.

1. Analysis of the existing corpus in the context of applied classes:

* the volume of publications corresponding to different application classes and the dynamics of its change over time in the context of the entire corpus
* the volume of publications corresponding to different applied classes and the dynamics of its change over time in the context of a particular media (source)
* ranking the media as appropriate to different application classes
* ranking of publications on the measure of concordance of the various application classes
* clustering of publications on the measure of compliance [application class]/[theme combinations]
* detection of anomalies (non-standard combination of subjects / classes or non-standard combinations and dynamics of the analyzed features)
* identification of topics, their combinations and classes are in trend, gaining popularity.

# MONITORING AND CORRECTION OF DEVIATIONS IN THE WORK OF THE DEVELOPED INFORMATION SYSTEM OF EVALUATING THE IMPACT OF OPEN TEXT INFORMATION SOURCES ON SOCIETY

Monitoring of the state of computing processes is carried out using standard methods of the Airflow web-based administrative panel, as well as using the Sentry.io system integrated into all system components, including Airflow and the system’s web interface. Sentry.io allows to automate the collection of information about errors and exceptions, including logging data and informative memory dumps at the time of the error. It is possible to track statistics on the frequency of certain errors, as well as flexible search and filtering. An automatic system for warning of exceptional situations to the email addresses of responsible persons has also been implemented.

Monitoring and analysis of the developed information system was carried out, on the basis of which it was recommended to consider the possibility of writing own codes for using dashboards, the user interface and for the possibility of configuring the functionality for the specifics of the project; inclusion of the LDA topic modeling algorithms developed during the project, machine learning of recognition of tonality, determining the polarity of tonality and the degree of objectivity of news texts with a lexical vocabulary approach and creating own library with the placement of dictionaries and experimental cases formed as part of the study.

Further directions of NLP Monitor system development

The main directions for the further development of the system are:

1. Development of a Kubernetes-based cluster for automatic load balancing between multiple physical servers;
2. 2) Full integration of CUDA-compatible computations into most algorithms and computational models;
3. Development of an advanced user interface for accessing and visualizing information;
4. Development of a system for detecting anomalies and novelty (novelty detection) with automatic notification of the relevant responsible persons;
5. Automatic generation of analytical reports and recommendations to support management decisions.

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

## Clustering Analysis methods and their application in the analysis of media publications

Since the tasks associated with the processing of several million text documents, such as publications in the media, are being solved within the framework of this project, then it is obvious that the most appropriate approach is the group analysis of publications. As analyzing / classifying each publication separately in the context of many informative features is a very labor intensive task, especially for experts. In contrast, in the group analysis the publications automatically, by means of clustering algorithms, split into groups / clusters as the proximity measure of their sets of informative features increases. There are significantly fewer such cluster groups than the original objects – publications. Thus, their group analysis can be analyzed, interpreted and visualized by experts in the information system.

Within the confines of this project the clustering algorithms are used for solving the following tasks:

1. Clustering the mass media publications on information guides, topics, sources, existing objects / subjects and others [83, 84].
2. Clustering the base units of analysis (words, stable phrases, n-grams and others) based on its vector representations gained by means of distributive semantics [81, 82].
3. Agglomerative clustering of topics and their combinations in order to build thematic ontologies.
4. Media clustering by their aggregated vector representations.
5. Clustering of individuals (consumers of media information) based on the analysis of their “vectors of information preferences” and “vectors of individual features”.

In the process of clustering analysis, one has to face the needs of solving the following main problems caused by the specifics of subject area:

* + 1. Clustered data collection;
    2. Data preprocessing, normalization and parametrization in the form of a vector of numerical features (various applied tasks require the use of different composition of informative features);
    3. Determination of optimal compositions of vectors of informative features;
    4. Selection or development of an algorithm with the required characteristics;
    5. Development of criteria for a numerical assessment of the quality of clustering results.

The variability of the mentioned sections above is directly related the specifics of the defined applied problem.

In the present time there are a large number of diverse news media resources in the world. Over time, the number of resources only grows. The spread of new technologies provides a greater accessibility to information, simplification in its dissemination and realization. Every day media resources produce a huge amount of news information related to various events occurring in the world at the present or in the past, expressed opinions and ideas. It may relate to different things or materials depending on the area of interest, specialization or editorial policy.

Often consumers are required to find certain information regarding an interesting news area, for this they turn to news media trying to find a needed information in news publications. He / she needs to develop an objective understanding, opinion, his / her point of view based on the matter of observed interest. For this aim, in order to fully understand and obtain a comprehensive picture, the user needs to find publications from as many possible media resources as possible, in which various opinions and views regarding his / her informational interests can be formed. Whereas, when the information is received from similar resources, one can acquire a biased opinion. It is also possible to be manipulated and imposed by media resources with some one-sided opinion favorable to them, if they prefer to undergo with an unethical policy.

When accessing a search query, even if it is formulated successfully, the search system often provides a large amount of information including the information which is unrelated or indirect to the information interest. The user can easily get lost in this amount of information, in order to process the results a lot of effort and time is required. For more efficient work, it would be convenient for the user to have at hand an intellectual system that would automatically find the queried information produced by different media resources in the Internet space.

## Dense text embeddings based on the co-occurrence of words in texts and topic modelling using “Big Data”

Within the realization of the given project, the approach of dense topical vectorizations of texts that constitute “Big Data” corpus was developed. The basis of the approach is the use of the vector representation of words based on the co-occurrence of words in texts and the subsequent application of the fuzzy clustering algorithm to them to obtain groups of thematically similar words. The resulting groups of words constructed on a basis of “Big Data” corpus are used then to obtain a dense vector representations of texts (text embeddings). The application of the approach is considered in relation to corpus of news publications. Also it was demonstrated that the approach is practically realizable with respect to the “Big Data”.

The developed approach of text representation has three advantages. First, it provides the text as a dense vector or text embedding, which is convenient for computational purposes. Secondly, it can represent thematically related texts in the same way even when texts are composed using words of different forms in them, provided that they are thematically related. Third, the components of evaluated vectors are interpretable.

When applying clustering algorithms to the developed vector representation, if the developed approach assigns similar (in some metric) vector representation to texts that are thematically or semantically similar, and assigns distinct vector representations to texts thematically different, obtained texts clusters are to include group of texts that are similar in theme and maybe even in meaning. The experiments showed that the developed vector representation had the indicated property. The developed approach to obtain text embeddings consists of four stages.

*The first stage:* construction of vocabulary

The very first task is to identify all the words from which the texts of corpus are composed. For this purpose, all nouns (including named entities) used in the texts are selected. The vocabulary includes all nouns in the texts with the exception of high-frequency words that serve only functional purposes, the so-called stopwords. Also, very rare words should be excluded from consideration. The resulting set of words forms vocabulary of corpus (basic units of analysis).

*The second stage:* grouping of words according to their topic similarity

At this stage, it is required to group all the vocabulary words according to their topic relatedness. A topic in Ozhegov’s dictionary is defined as an object, the main content of reasoning, narration, creativity, etc. Here, the topic refers to a set of word combinations that are often used in the same context or are constantly used in the same texts with a repeated frequency of cooccurrence. This step will, as will be presented later, provide denseness of vector representation and also represent similar in topic or even in meaning texts in similar way. Therefore, the natural idea is to initially determine groups of words that are topically or possibly even semantically close.

To implement the grouping of words by topic, a matrix X of the cooccurrence of words in texts is initially formed, with a dimension equal to the number of texts per vocabulary length. The elements of this matrix X are either equal to the frequency of occurrence of the corresponding word in the texts or to their values by statistical measure *tf-idf* , and it is considered that two words occur if they are used in the same text regardless of their location relative to each other in text. To obtain a vector representation of the words of the vocabulary, the transposed matrix X is multiplied by itself (3).

(3)

As a result, square symmetric matrix V with a dimension equal to the length of the vocabulary is evaluated, where each row or column is a vector representation of the word. The vector representation of the word characterizes the coherence of the given word with other words regarding the occurrence in the texts of the corpus. The vector representation of words having the same level of coherence with other words and thus most likely related to one topic are similar and their cosine proximity value is higher relative to others. Thus, calculation of the matrix of cosine distances of the word vectors (rows or columns of V matrix) of the vocabulary (4)

(4)

and application of fuzzy clustering algorithm to it provides clusters of words that are topically related. The fuzziness of the clustering algorithm ensures that the words are placed not exclusively in one cluster.

*Fuzzy clustering algorithm*

As “fuzzy” clustering algorithm an algorithm was selected which has the minimum number of parameters, constructs the clusters based on the distances between the elements, which does not require initial setting of the total number of clusters as a parameter, since their number is not known in advance, which also can be implemented on parallel processes and provides relatively fast exucution of the algorithm when applied to “big data”. The selected clustering algorithm is described in section 2.2.3 “Topic modeling algorithm based on the cluster approach”. The fuzziness of the algorithm is provided because the algorithm can attribute one object to several different clusters.

*The third stage:* aggregation of the resulting clusters to obtain thematic categories

Since the number of clusters obtained in the second stage can be very high (may exceed the size of the dictionary), in order to ensure a dense vector representation, it is necessary to combine thematically similar word clusters to obtain vectors with a dimension not exceeding the selected threshold. Further the obtained clusters of words are called thematic categories and their number determines the dimension of the vector representation (it equals exactly to it).

*The fourth stage:* constructing text embeddings

A text embedding is formed by determining the degree of relatedness of the text with each thematic category identified on the third stage. Those texts that are similar in topic have a similar level of relatedness with each individual category. The degree of relatedness of the text with each thematic category can be calculated through the modified Jaccard index (5).

(5)

Another variation of the modified Jacquard index, which takes into account the *tf-idf* values of the words, is represented by the formula (6).

(6)

Formula (6), if a text and a category have a common word, adds not only one in the numerator (as in (5)), but multiplies it by the *tf-idf* value of the word, therefore in cases where the common word (for both category and text) which often is met throughout number of texts makes a smaller contribution compared to word, which are less common in an entire corpus, but relatively often is met in a text under consideration.

The high value of the Jaccard index between a category and a text indicates a high level of coherence between a category and a text. As a result, the number of different topic relationships of a text with thematic categories is equal to the number of categories and all the estimated values can be considered as one dense vector or text embedding. Thus, the dimension of the resulting text embedding is equal to the number of thematic categories.

The methodology was developed to assess the quality of clustering of news publications. The assessment is based on comparing the results of texts publications clustering with labeling results performed by experts.

The quality of clustering results evaluated using the developed vector representation far exceeds the clustering results obtained by random construction of clusters. The clustering results in comparison with the results obtained using vectors constructed based on distributive semantics models and sparse vectors with a dimension equal to the size of the dictionary turned out between them, where the best result was showed using sparse vector representation. It is also noted that more thematic categories are used (thereby the dimension of the vector is higher), the results become better.

Further, application of developed vector representation to other specific tasks of the natural language processing problems can be studied; analyze of results with those that the best advanced approaches currently show can be performed.

Currently the article is being prepared for publication, which presents in detail the descriptions of the developed approach, as well as the experimental results.

## Clusterization of news publications in accordance with reflected information events (newsbreaks)

Within the realisation of the given project, the algorithm for grouping news publications in accordance with news events they reflect has been developed. It is based on clustering methods. It's detailed description, estimated results and conclusions were presented in [83]. The application of the developed approach to corpus of news publications lets automatically group news publications printed by various news media in accordance with information event they reflect. Depending on the type of the information event, estimated groups of publications might contain publications where different points of views with respect to given information event are presented, including an opposite point of view, a more or less detailed information or complementary information or correction of initially presented information is stated.

The development of this approach included three stages: the first stage - the development or selection of a type of representation of news publications (headings and publications texts itself) suitable for the task, that is, the determination of appropriate features; the second stage - the development of a measure of proximity (one minus distance function) that allows to quantitatively determine the similarity of news publications in accordance with information event they present. The third stage is the development or selection of a clustering algorithm to provide grouping publications into clusters. In the given approach the developed measure of proximity (one minus distance function) is the key element that provides grouping of news publications in accordance with information events.

In the process of development the developed approaches were tested on the data set prepared by experts consisting of 822 news publications to evaluate performance, efficiency of approaches, comparison of results and identifying the optimal ones. An assessment of the working time of the approaches was performed as well.

Based on the results of application of various approaches to the given data set, the approach that showed the best results was chosen. As stated above a key invention of this approach is the developed measure of proximity (one minus distance function) for determining the semantic proximity of publications texts in dependence of given information events. The measure consists of two components: the first one is a Jaccard's index applied to publications headings that are presented as a bag of words, and the second one determines the proximity of texts publications using word embeddings (dense vectors) constructed using distributional semantic models. The proximity can be evaluated via Word Mover’s Distance distance function or via euclidean distance of text embeddings evaluated as weighted average of word embeddings with equal weights. In both cases all words from publications were used and both showed close to each other results.

As was presented in article [83] , the developed approach for grouping news publications in accordance with news events was successfully applied to the data set of news publications marked out by experts. On the next stages of work it is supposed to study the applicability of the developed approach to the “big data” corpora, both in relation to the overall performance and its computational adequacy. Further, in addition, the developed approach can be studied in terms of finding of more optimal components by considering other clustering algorithms, other variations of proximity measures, developing other representations in order to improve results according to selected criteria.

## Method of clustering news media based on their conceptual analysis

Within the framework of this project, a method for clustering news media data based on their conceptual analysis published in [84].

The article describes a solution to the problem of clustering media messages based on a technique developed by the authors to automatically calculate the measure of the semantic significance of the names of document concepts using their statistical, syntactic and semantic features, and technologies for creating automatically declarative tools for clustering documents based on the methods of their semantic/syntactic and conceptual analysis.

When solving the problem of clustering media messages, it is shown that it can be implemented on the basis of a technique developed by the authors to automatically calculate the semantic value of the names of document concepts and technologies for automatically creating declarative tools for clustering documents based on methods of their semantic-syntactic and conceptual analysis.

Based on the proposed methodology for calculating the measure of the semantic significance of the concepts names and declarative software tools created during this study, an experiment was conducted to process a representative array of media messages. The analysis of the results shows that when automatically establishing the semantic significance of textual names of concepts, the use of semantic correlating coefficients of concepts improves the accuracy of establishing semantic similarity between documents.

In addition, we note that the limits in respect to computational resources affect the quantitative characteristics of the text processing results insignificantly.

## High Dimension Big Data Clustering Algorithm Based on Decomposition Method

## 

Since, as part of the implementation of the objectives of this project, we were faced with the need to cluster “big data” of high dimension, we developed and implemented the corresponding algorithm published in [82], [85].

Clustering of large datasets is widely discussed and currently a topic of high demand. The quality of data clustering is commonly estimated by Sum of Square Distance (SSD) criteria. In this study we show that it is possible to achieve high quality results in the terms of SSD criteria by applying iteratively the clustering procedure on subsets of the input dataset. Also, we demonstrate that the clustering time can be used as an additional parameter that allows to rig the quality of the clustering.

Clustering methods receive a lot of attention recently as effective tools in Machine Learning that allow discovering phenomena in the raw data. Another motivation comes from the necessity to process Big datasets in order to obtain natural grouping in the data. Hence, one of the main aspects of the clustering methods is scalability.

*Existing methods for solving the problem of clustering big data*

There is a number studies that are dirrected on improvement of clustering quality in cost of time complexity [86, 87]. These type of methods usally have an essential drawback: it is practically impossible to cluster medium and large datasets (about 105 elements and more). These methods cannot work on huge data bases becouse their time/space complexity growths (so, polynomially) very fast. Hence, it is reasonable to search for an algorithms having a property of a balance between scalability and clustering quality [88-91]. One of the well-known methods in data clustering is *k*-means algorithms which is widely used due to its simplicity and good characteristics. [92]. A number of techniques are dirrected on partinional clustering methods [93] of *k*-means.. There are algorithms which use data decomposition approach, e.g., minibatch *k*-means [94]. A weighted version of the *k*-means algorithm is applied in [90]. Meta-heuristics may help essentially when the exact solution is hard or expensive in terms of time/space. There are several heuristics implemented in *k*-means that speed up compuation:

* removing at each iteration patterns that are unlikely to change their membership thereafter as in [91];
* using triangle inequality in [95];
* mixture of various techniques [96, 97, 98].

*Informal statement of the problem*

We use a modification of *k*-means++ in order to build meta-heuristics that is executed on some subsets of the enire dataset. The goal of the study is to build a dataset decomposition method for centroid initialization of *k*-means clustering in order to produce competing results regarding MSSD (Minimum Sum of Square Distance) criteria. I.e., we produce a method for finding *k*-means initialization such that it is close to the optimal one, while having fast computation speed. Here we use meta-heuristics over *k*-means clustering by processing obtained data into secondary (high-level) clustering procedure. Another purpose of our research is to determine the efficiency of such heuristics and in different behaviours regarding meta-parameters.

By our approach, first, we decompose (split) the entire dataset shuffled randomly on subsets of fixed size. We are taking either all elements, or some portion of the dataset such that it remains representative. Next step is to do k-means clustering on some of these subsets (=packets). By estimating SSD on their corresponding clusterings we make a heuristic for a good initialization of the algorithm on the entire dataset.

*Model parameters:*

* is number of required clusters;
* is number of objects in the entire data set;
* is size of data packet (number of objects in one packet). The sizes of the packets are chosen in proportion to the entire dataset*. E.g.,* taking 5 decomposition means taking the size of the packets equal to .
* is number of packets used for independent initialization of *k*-means during Phase 1;
* *m* (≥*n*) is total number of packets used for the clustering. The union of packets may (or may not) cover the entire dataset.

We considered the following two modes for the packets generation:

1. Segmentation of the entire data set on packets: A random permutation of objects in the data set is created. The data set is segmented into successive packets of size . We refer to is as uniform packet decomposition mode.
2. For each packet, random objects are selected from the entire set. By repeating this, the required number of packets is generated (objects may be picked repeatedly in different packets). We refer to is as random packet generation mode.

*Big Data Clustering Algorithm*

An independent application of *k*-means++ algorithm on a fixed number of packets in order to find aggregated set of centroids The scheme for the algorithms is shown in figure 6. For this, the centroids of all packets are joined into one new data set.

Each object is assigned the weight corresponding to the normalized SSD value for the packet in which it is calculated as centroid. The weight of -th object is calculated as follows:

(5)

where is value for packet from which the -th centroid is taken as object. Then, using -means, the new dataset is divided into -clusters, taking into account the weights of the objects. In the case of degeneration, -means is reinitialized. The resulting centroids are used for:

1. Initialization of -means on the entire data set
2. Evaluation of the on the entire data set
3. Initialization Phase 2 of the algorithm described in the following sections.

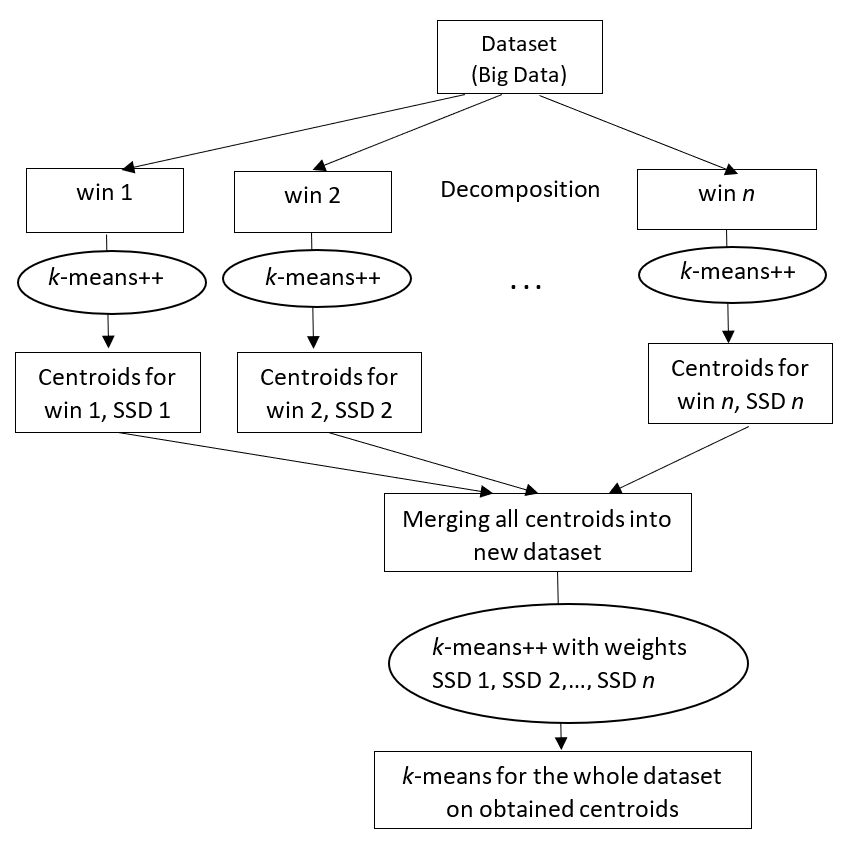


Figure 6 - Scheme for parallel decomposition clustering on packets (wins) .

During processing subsequent packets, , ... we have considered two options: parallel (2.1) and sequential (2.2).

*(2.1)* *Parallel option*

The centroids obtained in the previous Phase 1 are used to initialize *k*-means on each subsequent packet , , ..., . The resulting centroids and *SSD* are stored if there is no centroid degeneration. The stopping condition is the specified limit either on the computation time or on the number of packets being processed. Similar to Phase 1 we do the clustering on the joined set of centroids. Subsequent use of its results is similar to clauses 1.1 and 1.2 of Phase 1.

*(2.2) Sequential option*

The sequential version of the algorithm is schematically represented in Figure 7.

Centroids from phase 1 are used for k-means clustering on the packet m+l, m+l+1,… . If there is no degeneration during clustering then memorize the resulting centroids and the corresponding SSD values. In order to obtain new centroids, we carry out clustering with weights on the united set of centroids (similarly to Phase 1). If the time limit has not been exhausted then use centroids obtained in step 4, and repeat the procedure. Finally, when time limit is reached then the usage of obtained results on the last calculated centroids is similar to clauses 1.1 and 1.2 of Phase 1.

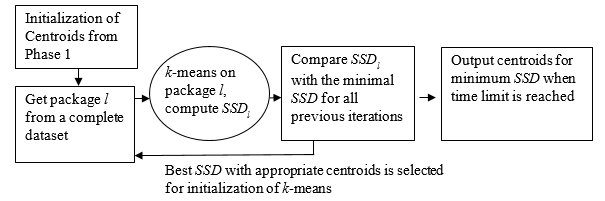


Figure 7 – Scheme of sequential cluster decomposition for packets *n*+1, *n*+2, ..., *n+l*, *..*.

*Used tools and material base*

Implementation of this algorithm is performed in the programming language Python, using libraries that are in the public domain: numpy, scikit-learn, math, multiprocessing. The material base consists of desktop computers Intel core i7-4790K 4.0 GHz with 6 cores, a laptop Intel core i5 7200U 2.5GHz, as well as a remote high-speed server, 40 cores, 600 GB of RAM.

*Results*

In Table 1 we show the testing result of our algorithm on datasets of various sizes. Each dataset is decomposed on packets of sizes , ,..., .

Table 1 – Experiments on different datasets. SSD criteria and computation time are considered.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| dataset  size | num. of  experiments | improves  k-means++ |  | num. of  overtime |  |
| 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 1590 | 33 (2.1%) | 0.964 | 0 | 0.330 |
|  | 1186 | 35 (2.9%) | 0.989 | 17 | 0.433 |
|  | 312 | 14 (4.5%) | 0.992 | 15 | 0.051 |
|  | 1140 | 345 (30.3%) | 1.000 | 10 | 0.386 |
|  | 300 | 160 (53.3%) | 1.000 | 12 | 0.394 |

Continuation of Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 3 | 4 | 5 | 6 |
|  | 82 | 40 (48.8%) | 0.996 | 41 | 0.971 |
| \* | 30 | 26 (86.7%) | 1.000 | 30 | 1.650 |
| \*\* | 260 | 165 (63.5%) | 1.000 | 0 | 0.160 |

*Description of datasets:*

* Our first experiments are conducted on syntactic datasets from 10 elements to 2 and number of attributes from 20 to 100. Also different sizes of packets are considered. The structure of the data: some number of synthetic blobs of Gaussian distribution.
* For size 50 dataset the number of clusters was considered in the range from 5 to 70 with the step 2. The number of blobs was picket in the range from 30 to 140. The number of chunks was in the range from 5 to 100.
* For size 5 dataset SUSY is obtained from open UCI database [99].
* \* Dataset is preprocessed by normalization.
* \*\* 2-means clustering problem, decomposition of dataset on packets .

*The direction of further work*

This clustering approach involves the introduction of a search methodology with alternating neighborhoods for solving computational problems on large data of high dimension and its further testing on a vectorized text corpus. It is also planned to integrate this technique into a complex with alternative cluster methods for extracting information from texts.clustering algorithms and provide estimates to corresponding time/SSD quality reduction.

## Methods, algorithms, and tools for preprocessing text publications

To solve most of the tasks of this project, it is necessary to process and further analyze text data from open information sources. Text data preprocessing is a basic step. The quality of the pre-processing and the efficiency of the developed information system largely depend on the quality of the pre-processing. Within this project the following tasks of text data preprocessing were solved:

* Lemmatization [100];
* The issues of applicability of existing cloud services for processing natural language texts were investigated. [101];
* The problem of automatic selection of phrases from text cases was solved [102];
* Automation of the process of extracting keywords from analyzed publications [103, 104];
* Algorithms for automatic summarization of text documents were developed [103, 104];
* Morphological disambiguation [105];
* The issues of applicability of existing methods and algorithms for parameterization and vector representation of text data, in particular, using methods of distributive semantics, were investigated. [106].
* Automatic parsing of dependencies at the syntactic level (article in the process of publication);
* Recognized named entities (article in the process of publication);
* Automatic markup of parts of speech in the text (article in the process of publication).

## Automatic summarization of text documents using word mover’s distance and extracted document keywords

The project has developed a method of automatic summarization test documents, published in [103, 104]. The main idea behind the method is in selecting sentences based on Word Mover's Distance Similarity between each sentence and set of centroid keywords. This approach leverages both compositional property of word embeddings and advantages of recently discovered powerful text to text distance metric. ROUGE results on DUC 2002 data set showed that quality of produced summaries can compete with well-known state of the art systems. In this work we also discuss limitations of gold summaries in evaluating quality of summarization systems.

In this work we aim to present single document extractive summarization system that can leverage advantages from both compositional properties of the word embeddings and WMD metric. Firstly, we use word embeddings to extract centroid keywords. Secondly, WMD will be used to calculate similarity between sentences and centroid keywords of the document. In order to evaluate competitiveness of the proposed system we will compare it with other state of the art methods based on ROUGE results on DUC-02 dataset.

We consider the developed method for ranking sentences for the summation of documents. The main idea is to highlight those sentences that are close to the keywords of the documents. Keywords are defined as words that are closest to the center of mass of the presentation of the document (the document centroid). Word Mover's Distance (WMD) is used as a similarity metric. WMD is one of the best metrics for determining the semantic similarity of texts, but the process of calculating it is time-consuming. The main reason for this is that WMD is based on a solution to the transport problem. Next, an automatic summarization system will be presented, which takes advantage of the WMD metric, while avoiding pairwise computations, and applies it to measuring the distance between sentences and document keywords.

To test this method for representing words in vector space, we used the pre-trained word2vec model on GoogleNews2. Preliminary processing of documents was performed in such a way as to maximize the number of words that can be found in the dictionary of this model. The first step was to split the documents into sentences, then tokenize and delete stopwords. The lowering of the register of all words was carried out taking into account the fact that there are words that are stored in the model only in upper case.

The method of extractive summarization based on centroids was proposed in [107]. The centroid of a document is a kind of abstract document that represents the only vector of the most significant information of the original document. Consider a document with a dictionary *V* of all words of size *N*, construct a matrix *E ∈* , where the i-th row of the matrix is the vector representation of the i-th word from V. The centroid of the document will be the sum of the vector representations of all unique words in document.

We deﬁne keywords as a set of words of a limited amount that can describe original document satisfying both coverage and diversity of its content. We make assumption that centroid embedding from previous section represents central topic of the document and words that have embeddings close to this center can form set of centroid keywords. For this purpose we simply rank all unique words in a document by calculating cosine similarity between each word embedding and centroid embedding. Based on similarity scores we select top 25% of words from all unique words of the document which will form centroid keywords.

To determine the document sentences, which will be the summation of the document, we use the WMD similarity metric described in section 2.2:

(6)

Where bag of words for i-th sentence and K is is set of keywords.

Choosing the closest sentences so that they are equal in number to 100 words (if the sentences do not fit exactly one hundred words, then the last sentence is cut off), we get a summary of the document.

Choosing the closest sentences so that they are equal in number to 100 words (if the sentences do not fit exactly one hundred words, then the last sentence is cut off), we get a summary of the document.

Based on the results of comparative experiments [103], the developed method generates a fairly successful summary of the text with an acceptable ROUGE score. Moreover, it can successfully compete with existing systems.

## Development of optimization methods

Most of the algorithms developed in the framework of this project have sets of adjustable input parameters. In addition, when developing each algorithm, it is necessary to determine proper criterion that estimates numerically the quality of its results. Since most of the algorithms have several configurable parameters, a wide range of problems arises related to their optimization based on special parametric optimization algorithms.

Within the framework of this project, various optimization methods have been developed and applied [108-114]. Optimization has been widely used in solving problems of cluster analysis [108, 110]. Optimization is considered for the Minimum Sum-of-Squares Clustering Problem class of tasks, which include k-means, k-means ++, j-means. In most cases, k-means class algorithms stop in one of local minimums according to the Minimum Sum-of-Squared Distances criterion. However, the use of Variable Neighborhood Search techniques makes it possible to approach a global minimum without stopping at local ones [113, 114].

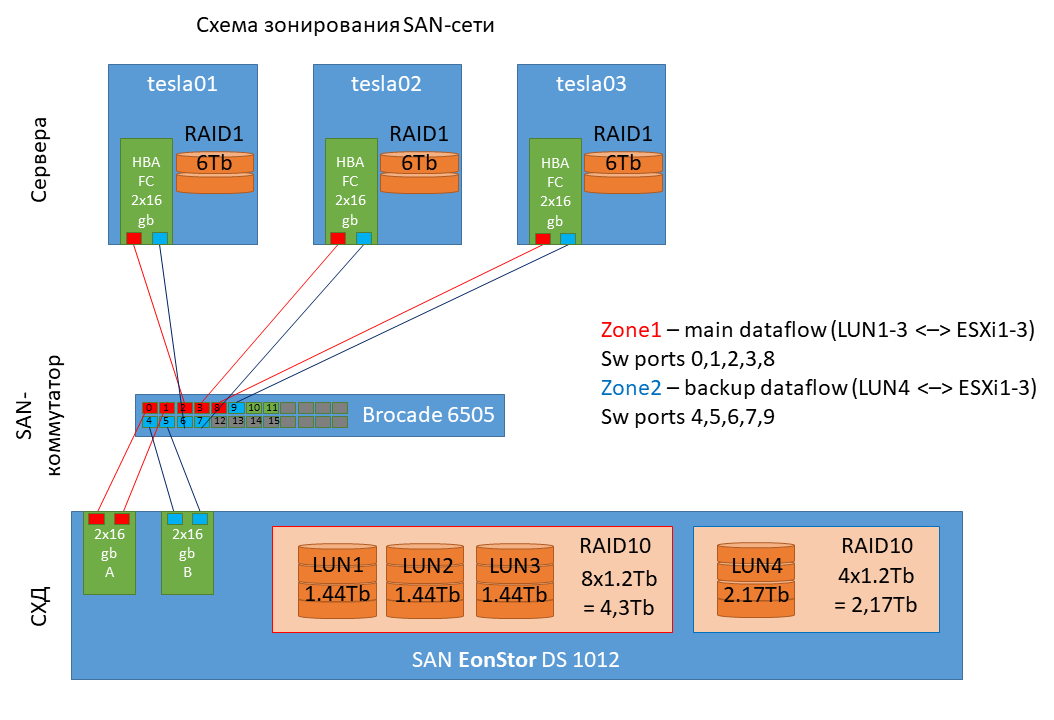
Since for various tasks of the project it is necessary to process large amount of information, including parametric identification tasks, it is often necessary to use methods of parallelization of processing algorithms and data analysis algorithms. In the parallelization, additional problems of optimal planning arise for multiprocessor computing [110].

During the clustering of both text documents and various lexical units, one can construct different graph structures, including ontologies. For example, based on the methods of distributive semantics, one can construct various graph representations of dictionaries where the connections between words are established on the basis of measures of their contextual and semantic proximity. Having constructed such lexical graph structures, it becomes necessary to solve the Hub Location Problem [111].

In order to solve the problems of approximating experimental data, the variable neighborhood search (VNS) method has been used [113, 114].

## Creating the necessary technical conditions

A group of engineering and technical workers improved technical conditions for the successful and smooth operation of the developed information system, the collection, analysis of text data and testing algorithms, [115]: data storage system installed (DSS) and server virtualization VMware software (hypervisor ESXi and vCenter); upgraded existing server: HBA cards are installed and configured to connect the server to the storage system. Two disk arrays RAID10, 4.3Тб and 2.2Тб created in DSS, LUN is allocated for each server and in each array, as well as zoning of data streams for fault tolerance (Figure 8). All servers and storage equipment moved to the data center to the CO-LOCATION and are configured remote access to each unit from the local network by the IICT.



DSS SAN-switch Servers

Figure 8 – SAN Network Zoning Scheme

Created virtual machines: working virtual servers for calculations on CPU and GPU (configured forwarding of video cards in Passthru mode); for storage and data exchange; for virtualization management; for remote access to servers; virtual router to provide internet traffic. On working servers with video cards installed and configured: ОС Ubuntu 16.04, cuda 10.0, cudnn 7.3, Anaconda 3, machine learning libraries, JupyterHub, nbextensions. On working servers with video cards, access from the Internet via HTTPS is configured, domain names for each production server and personal accounts for all project developers, configured server and data access security policy. For storage and data exchange, all user home directories were transferred to a server for data storage and exchange. Configured data exchange between servers via NFS protocol.

For the development and implementation of the information system, according to the project, resources were allocated and virtual servers created:

1. *Vm-app* – the main application server. System for monitoring, storing and processing data. It contains database (Postgresql), noSQL repository (ElasticSearch+Kibana), Django web application, stack Airflow+Celery and related service components (nginx, redis). Everything is deployed in docker containers.
2. *Vm-registry* - Private Docker registry for storing ready-made docker images.

The resources of the existing virtual servers were reconfigured due to the changing needs for computing power. Most of the virtual CPU and RAM were removed from one of the servers for mathematical modeling and transferred to the application server.

## Creation of necessary expert and analytical conditions

This work aims to form a balanced, representative, marked and classified corpus for automated assessment of the impact of open text electronic information sources on society on the basis of published information [50].

As part of the set of works on the formation of a corpus of journalistic texts: texts of the official media, magazine publications, news portals, etc.; the following tasks were defined for:

* primary source selection and data collection;
* definition of requirements for the corpus: provision of balance, representativeness, main categories of meta-markup;
* formation of main, small and test corpora;
* refining corpora and systematisation of data, marking of the test corpus.

In the course of the activities on developing the additional balanced, representative, annotated and classified corpus of journalistic texts of electronic news portals, there was formed 3 corpora:

1. the main corpus, which includes 1 982 433 articles from 22 sources with meta-marking by 14 categories;
2. the small corpus of texts, including 245 352 articles from 5 sources with meta-marking by 7 categories;
3. the test corpus, including 35 643 articles from 5 sources, with a markup by 39 categories and classification by features.

Methodology of formation of the main body. The sources were selected from the total list of domains consisting of 2613 media, with indication of the domain, the type of media (print, TV, radio, news agencies, Internet-media), and scale (regional, central) on the basis of the specificity of the machine learning methods in the analysis of textual information [85]. When reviewing and selecting sources the following requirements were taken into account:

* exclusion of regional media;
* exclusion of field-specific sources;
* ensuring inclusion of publications in Russian to the corpus;
* adding the information portals specialising in providing daily news about developments in the country to the corpus;
* adding the publications of Mass Media information portals of various levels of popularity among the population.

Based on these requirements, regional media, field-specific media, publications in other languages, suspended mass media were excluded from the sampling frame.

Thus, the main corpus of 1 982 433 articles was formed, the general characteristics of which are presented in table G1 of Appendix G.

Based on the results of work on the formation of an additional vast corpus, the following categories of meta-markup were identified:

1. Publication author – name of the author;
2. Category/topic of publication – thematic section of the site to which the publication relates.
3. Publication release date;
4. Publication download date;
5. A publication identification number is the identification number that is automatically assigned to the article to search for it and avoid double counting or omission of the article in the analysis process;
6. Links - the availability of hyperlinks to other articles or pages in the text of the publication;
7. Number of comments;
8. Number of views (as of download date);
9. Title;
10. Subtitle;
11. Tags (tags and hashtags assigned to the publication);
12. The text of the publication;
13. Publication text in HTML format;
14. Publication URL.

So in our case, the representativeness in relation to the analysed area of the large corpus is undeniable, since the sample of the supplementary corpus included exclusively the texts representing the research area (all types of journalistic texts in available information sources that influence the society of Kazakhstan). Since the main purpose of our project is to study the impact of open information text sources on society, this problem was solved completely: the additional corpus is a collection of texts exclusively from sources of this type and covers almost the entire period of existence of such sources, and almost the entire range of sources that have an impact on society.

As part of our study, based on the fact that the collected corpus reflects the totality of Kazakhstan's open text media sources with information of a news nature that affect society, the requirement to balance the entire corpus should not be perceived as critically important, since the most important condition - the reflection of objective reality, which is fully complied (Figure G1 of Appendix G).

It is also necessary to clarify that for the purposes of machine learning and testing algorithms designed in the framework of the project information system, there will be a sample from the general corpus, where the principle of balance in relation to the studied features will be also observed properly.

It should also be noted that in many cases data for the above meta-marking categories are not available. This is primarily due to the existing data architecture. For example, only a small part of the sites are designed in such a way that it is possible to collect data on the number of views and the number of comments to the publications placed on them. Many sites do not provide the option to comment on publications. This greatly complicates the analysis of data, since, from the point of view of assessing the impact on society, the number of comments and views of publications is almost the only direct indicator of the interest of the population.

The preparatory stage which determines the solution of tasks for automatic evaluation of publications in the media and recognition of selected informative features is a set of works for the creation of a dictionary array, selection of language units and methods (rules) of their use in news texts. In this regard, this work aims to develop methods to formalise rules and dictionaries for recognition of informative features in marked-up corpora and determination of the degree of their expression.

12 informative features for the recognition of which it is necessary to form dictionaries and formalize linguistic patterns and rules were identified.

2,000 texts from 5 informational Internet news sources became an object of analysis: “KazakhSTAN 2.0”, “Central Asia Monitor”, “zakon.kz”, “Radio Azattyk”, “Tengrinews.kz”.

The following tasks were set and solved to achieve the goal:

1. Conducting an in-depth analysis of the formed corpus of media publications on 9 informative features (3 features in 2018, 9 - in 2019).
2. Markup of text corpus according to outlined informative features.
3. the degree of generalisation (mild, strong, absent); Formation and systematisation of dictionaries by certain informative features from analysed texts.
4. Identification of patterns to formalise the rules of separate informative features recognition.
5. Obtaining expert opinions on the feasibility of identified patterns in the project.
6. Description of methods for rules formalisation based on linguistic analysis of texts.

Upon completion of the work of the linguists group, the revealed patterns on 8 informative features out of 12 were submitted for discussion. During the expert discussions on the revealed patterns of the experts made comments and conclusions concerning applicability and feasibility of results of linguists' work, where the conclusion "feasible" means the availability of software tools that enable implementation of the identified patterns within the developed product; conclusion "difficult to implement" means that implementation of the identified pattern requires construction of complex multi-stage algorithms and time-consuming and perhaps impractical to implement from the point of view of the objectives and budget of a project; conclusion "not feasible" means the absence of software tools that enable implementation of the identified patterns within the developed product; conclusion "requires discussion" means need for more extensive consultation with experts for a definitive conclusion.

# CONCLUSION

Within the framework of the project-targeted funding project for 2019, in accordance with the calendar plan, the following main results were obtained:

1. Methods for assessing the impact of open information sources on society based on the analysis of published text information have been obtained;
2. The development of a methodology of calculating individual criteria to assess the impact of open text information sources on society has been started;
3. The design and development of an information system of assessing the impact of open text information sources on society has been started;
4. Monitoring and correcting deviations in the work of the developed information system of assessing the impact of open text information sources on society have been started.
5. For the current stage, the necessary technical and expert-analytical conditions have been created for the development of an information system of assessing the impact of open text information sources on 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.

According to clause 9 of the project schedule, in 2019, 15 articles have been published in foreign peer-reviewed publications with a non-zero impact factor and 3 author's certificates have been received for a computer program, as well as 5 articles in domestic scientific publications recommended by the Committee on control in education and science sphere and 20 articles in proceedings of international conferences have been published.

# REFERENCES

1. Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O.Yu., Yergaliyev Y. N. and Muhamedyev R. I. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area// Journal of Physics: Conference Series. -2018 // <https://doi.org/10.1088/1742-6596/1117/1/012001>: 21.10.2019
2. Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O.Yu., Kuchin Ya.I., Yakunin К.О., Muhamedyev R. I. The studies of dinamics’ changes in publication activity in the field of natural language processing// Mater. 17th Int. scientific conf. "Information Technology and Management 2019". - Riga, Latvia, 2019. -P. 71-74
3. Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O.Yu., Kuchin Ya.I., Yakunin К.О., Muhamedyev R. I. Izmeneniye publikatsionnoy aktivnosti v oblasti obrabotki yestestvennogo yazyka // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty,- 2019. -P. 130-135 (In Russian)
4. A.c. № 4505. Programma rascheta dinamicheskikh bibliometricheskikh pokazateley D1 i D2 (Calculation of dynamic scientometric indicators D1, D2 – CalcDStMi)/ Yakunin K.O., Mukhamediyev R.I., Kuchin YA.I.; publ.12.07.2019. – 1 p. (In Russian)
5. Global research "2019 Edelman Trust Barometer"// <https://www.edelman.com/trust-barometer> : 18.07.2019
6. Miller D. Promotional strategies and media power // Briggs, A., Cobley, P. (eds.), THE MEDIA: AN INTRODUCTION. - London, United Kingdom: Longman, 1998. -P. 65-80
7. Bushman B., Whitaker J. Media Influence on Behavior // Encyclopedia of Human Behavior, -2012. -P. 571-575
8. Don W., Zongchao C., Cylor, S. Media Effects. In: James, D. INTERNATIONAL ENCYCLOPEDIA OF THE SOCIAL & BEHAVIORAL SCIENCES, Second Edition, - Elsevier, 2015. - P. 29-34
9. Hoon K., Jong Y., Sangheon K., Libor M., In S.: Human-machine interaction: A case study on fake news detection using a backtracking based on a cognitive system // Cognitive Systems Research, -2019. -Т. 55. –P. 77-81
10. Bushman B., Whitaker, J. Media Influence on Behavior // Reference Module in Neuroscience and Biobehavioral Psychology, -2017.
11. Mashechkin I.V., Petrovskiy M.I., Tsarov D.V. Metody vychisleniya relevantnosti fragmentov teksta na osnove tematicheskikh modeley v zadache avtomaticheskogo annotirovaniya // Vychislitel'nyye metody i programmirovaniye, -2013. - Vol. 14(1). –P.91-102 (In Russian)
12. Parkhomenko P., Grigor'yev A.A., Astrakhantsev A.A. Obzor i eksperimental'noye sravneniye metodov klasterizatsii tekstov // Trudy institut sistemnogo programmirovaniya RAN, -2017. –Vol. 29(2). –P. 161-200 (In Russian)
13. Vorontsov K., Anna, P.: Regulyarizatsiya, robastnost' i razrezhennost' veroyatnostnykh tematicheskikh modeley // Komp'yuternyye issledovaniya i modelirovaniye, - 2012. -Т. 4(4). –P. 693-706 (In Russian)
14. Vorontsov, K., Frei, O., Apishev, M., Romov, P., & Dudarenko, M. (2015). BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. Analysis of Images, Social Networks and Texts, -2015. –P. 370–381 // doi:10.1007/978-3-319-26123-2\_36  : 21.10.2019
15. Blei D., Ng A., Jordan M. Latent Dirichlet allocation// Journal of Machine Learning research. -2003. –Vol.3. –P. 993-1022
16. Scott J. A decision support system for supplier selection and order allocation in stochastic, multi-stakeholder and multi-criteria environments// International Journal of Production Economics, №166, -2015. –P. 226-237
17. Mardani A., Jusoh A., Zavadskas E., Cavallaro F., KhalifahZ. Sustainable and renewable energy: an overview of the application of multiple criteria decision-making techniques and approaches //Sustainability. -2015. -Vol. 7(10). –P. 13947-13984
18. Wanderer T., Stefan H. Creating a spatial multi-criteria decision support system for energy related integrated environmental impact assessment// Environmental impact assessment review. -2015. -№52. -P. 2-8
19. Hoceini Y., Mohamed C., Moncef A. Towards a new approach for disambiguation in NLP by multiple criterian decision-aid// The Prague Bulletin of Mathematical Linguistics. -2011. -№ 95. –P. 19-32
20. Kumar A. A review of multi criteria decision making (MCDM) towards sustainable renewable energy development// Renewable and Sustainable Energy Reviews. -2017. -№69. -P. 596-609
21. Yager R. On ordered weighted averaging aggregation operators in multicriteria decision making// IEEE Transactions on systems, Man, and Cybernetics. -1988. -№18 (1). –P. 183-190
22. Hansen, P., Franz O. A new method for scoring additive multi‐attribute value models using pairwise rankings of alternatives// Journal of Multi‐Criteria Decision Analysis. -2008. -№15(3‐4). –P. 87-107
23. Figueira J., Vincent M., Bernard R. Electre methods // Multiple Criteria Decision Analysis: State of the Art Surveys. Part of the International Series in Operations Research & Management Science book series, -2005. –Vol.78. –P. 133-153
24. Lai Y., Ting-Yun L., Ching-Lai H. Topsis for MODM // European journal of operational research. -1994. -№76(3). –P. 486-500
25. Brans J., Ph V. A Preference Ranking Organization Method (The PROMETHEE Method for Multiple Criteria Decision-Making)// Management science. -1985. -№ 31 (6). –P. 647-656
26. Saaty T. Group decision making and the AHP// The Analytic Hierarchy Process. Springer, Verlag. -1989. –P. 59-67// <https://doi.org/10.1007/978-3-642-50244-6_4>: 21.10.2019
27. Abaei M. Developing a novel risk-based methodology for multi-criteria decision making in marine renewable energy applications// Renewable energy. -2017. №102. –P. 341-348
28. Mukhamediyev R. Tools and Methods of Information Support for the Development of Renewable Energy// Renewable Energy and Power Supply Challenges for Rural Regions, IGI Global, -2019. –P. 39-65
29. Charabi Y., Adel G. PV site suitability analysis using GIS-based spatial fuzzy multi-criteria evaluation// Renewable Energy. -2011. -№36 (9). -P. 2554-2561
30. Jelassi T., Vladimir O. A framework for building an expert system for MCDM models selection// Improving Decision Making in Organizations, -1989. -P. 553-562
31. Dalkey N., Olaf, H. An experimental application of the Delphi method to the use of experts// Management science, -1963. -№ 9(3). -P. 458-467
32. Okoli C., Suzanne P. The Delphi method as a research tool: an example, design considerations and applications// Information & management, -2004. -№42 (1). -P. 15-29
33. Allen E., Christopher S. Likert scales and data analyses// Quality progress, -2007. -№40 (7). -P. 64-65
34. Mokken, R. A theory and procedure of scale analysis with applications in political research// De gruyter mouton, Series:Methods and Models in the Social Sciences. -2011
35. Orme B. Maxdiff analysis: Simple counting, individual-level logit, and HB // Sawtooth Software, Inc. -2009
36. Bradley R., Milton T. Rank analysis of incomplete block designs: I. The method of paired comparisons// Biometrika. -1952. -№ 39(3/4). -P. 324-345
37. Saati T., Andreychikova O. About the change of the intangible. The approach to relative changes based on the main eigenvector of the matrix of pairwise comparisons // Cloud of science. -2015. -№2(1). –P. 5-39
38. Yandeks nazval samyye populyarnyye zaprosy kazakhstantsev // <https://kapital.kz/lifestyle/74330/yandeks-nazval-samye-populyarnye-zaprosy-kazahstancev.html>: 20.09.2019
39. Baymanov D. Volnuyushchiye kazakhstantsev voprosy nazval Prezident RK. // <https://www.inform.kz/ru/volnuyuschie-kazahstancev-voprosy-nazval-prezident-rk_a3449476> : 20.09.2019
40. Zhulmukhametova ZH. Troyku samykh ostrykh voprosov, volnuyushchikh kazakhstantsev, nazvali sotsiologi. // <https://informburo.kz/novosti/troyku-samyh-ostryh-voprosov-volnuyushchih-kazahstancev-nazvali-sociologi.html> : 20.09.2019
41. Barakhnin V.B., Kuchin YA.I., Mukhamediyev R.I. K voprosu o postanovke zadachi vyyavleniya feykovykh novostey i algoritmakh ikh monitoringa // Mater. III Mezhdunar. nauch. konf. «Informatika i prikladnaya matematika», posv. 80-letnemu yubileyu prof. Biyasheva R.G. i 70-letiyu prof. Aydarkhanova M.B. - Almaty, 2018. -P. 113-118 (In Russian)
42. Jaccard P. Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines// Bull Soc Vaudoise Sci Nat. -1901. -№37 (140), -P. 241-272
43. Kravtsov G. Обучаемая модель вычислений на классификациях// Yelektronne mo-delyuvannya. -2018. -№40(3). -P. 63-76
44. Articles and term dictionaries in Russian language// <https://drive.google.com/open?id=1CDEs6ooXtCq6v_jg-xSUVuPS1AE1ZSOn> : 18.07.2019.
45. R. I. Mukhamediev, R. Mustakayev, K. Yakunin, S. Kiseleva and V. Gopejenko. Multi-Criteria Spatial Decision Making Support Ssystem for Renewable Energy Development in Kazakhstan // IEEE Access. -2019. - Vol. 7. - P. 122275-122288
46. Musabayev R., Mukhamediyev R., Kuchin YA., Symagulov A., Yakunin K., Murzakhmetov S. Metod mul'timodal'noy otsenki sredstv massovoy informatsii na osnove tematicheskoy modeli korpusa tekstov // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy». – Almaty. - 2019. -P. 239-247 (In Russian)
47. Musabayev R.R., Mukhamediyev R.I., Kuchin YA.I., Symagulov A., Yakunin K.O. On a method of multimedial media ranking using corpus based topic modeling// Mater. 17th Int. scientific conf. "Information Technology and Management 2019". - Riga, Latvia, 2019. -P. 64-67
48. Mukhamediyev R.I., Musabayev R.R., Buldybayev T., Kuchin YA., Symagulov A., Ospanova U., Yakunin K., Murzakhmetov S., Sagyndyk B. Eksperimenty po otsenke sredstv massovoy informatsii na osnove tematicheskoy modeli korpusa tekstov // Mater. IV Mezhdunar. nauch.-praktich. konf. «Informatika i prikladnaya matematika", posvyashchennoy 70-letnemu yubileyu prof. Biyarova T.N., V. Vuytsika i 60-letiyu prof. Amirgaliyeva Ye.N. - Almaty, Kazakhstan, 2019. -Vol. 2. -P. 419-431 (In Russian)
49. А.с. №4523. Programma mul'timodal'noy otsenki SMI na baze tematicheskoy modeli korpusa tekstov (Multi Modal Mass Media Assesment – M4A)/ Yakunin K.O., Mukhamediyev R.I., Kuchin YA.I., Murzakhmetov S.B., Symagulov A., Musabayev R.R.; publ. 15.07.2019. -1 p.
50. Atanaeva M., Buldybaev T. et al. Report on research work of IAC JSC for 2019 under the scientific and technical program "Development of information technologies and systems to stimulate the sustainable development of the individual as one of the foundations for the development of digital Kazakhstan." Nur-Sultan. – 2019. // <http://iict.kz/wp-content/uploads/2019/10/Report-on-research-work-of-IAC-JSC-for-2019.docx> : 23.10.2019
51. Meyrambekkyzy ZH., Amanbay A., Musabayev R. Avtomaticheskoye formirovaniye tematicheskikh slovarey sotsial'no-znachimykh ponyatiy // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty, - 2019. -P. 233-239 (In Russian)
52. Meirambekkyzy Zh., Mussabaev R.R. Automatic development of topic dictionaries of socially significant concepts // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part. 2. -P. 65-70
53. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation // Journal of Machine Learning Research. - 2003. -Vol. 3. -P. 993–1022
54. K.V. Vorontsov. Probabilistic thematic modeling // Electronic resource <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>: 14.08.2019
55. Korenčić D., Ristov S., Najder J.E. Document-based topic coherence measures for news media text // Expert Systems with Applications. – 2018. – Vol, 114. – P. 357-373.
56. Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G., Nurumov K.S., Baymakhanbetov M.A. Metodika dlya opredeleniya informativnykh priznakov novostnykh tekstov i proverka ikh znachimosti // Nauchnyy aspekt, – 2019. - Vol.3(3). –P. 277-295 (In Russian)
57. Buldybayev T., Baimakhanbetov M., Nurumov K., Akoyeva I., Ospanova U., Atanayeva M. Analysis of digital media: a comprehensive assessment of informative features significance // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part. 2. -P. 31-44
58. Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G., Nurumov K.S., Baymakhanbetov M.A. Opredeleniye tonal'nosti i ob"yektivnosti novostnykh tekstov slovarnym podkhodom // Nauchnyy aspekt. -2019. -Vol.3(3). -P.296-308 (In Russian)
59. Nurumov K.S., Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G. Kompleksnaya otsenka znachimosti informativnykh priznakov dlya otsenki vliyaniya na sotsium na osnove analiza publikuyemoy tekstovoy informatsii // Vestnik YENU im. L.N. Gumileva, Seriya Zhurnalistika. -2018. -№3 (124). -P. 44-52 (In Russian)
60. Atanayeva M.K., Ospanova U.A., Buldybayev T.K., Akoyeva I.G., Nurumov K.S. Ispol'zovaniye lingvisticheskikh podkhodov v opredelenii klassov i informativnykh priznakov v tekstakh // Vestnik KazNPU, Seriya filologicheskaya. -2018. -№2 (64). -P. 166-171 (In Russian)
61. Toleu A., Tolegen G., Mussabayev R. Initial exploration for Topic modeling with word sense // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty, - 2019. -P. 31-44
62. Vorontsov K., Potapenko A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // Ignatov D., Khachay M., Panchenko A., Konstantinova N., Yavorsky R. (eds) Analysis of Images, Social Networks and Texts. AIST 2014. Communications in Computer and Information Science. Springer, Cham. -2014. –Vol.436. –P.29-46
63. R. Rehurek, P.Sojka. Software framework for topic modelling with large corpora // The LREC 2010 workshop on new challenges for NLP frameworks, University of Malta, -2010. -P.45-50
64. Musabaev R. Results of thematic modeling for the tengrinews corpus. Almaty -2019. // <http://iict.kz/wp-content/uploads/2019/10/results-for-tengrinews.zip> : 23.10.2019
65. 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. –Vol.1117. -P.1-9
66. Nurumov K., Baimakhanbetov M., Buldybayev T., Akoyeva I., Ospanova U. An approach for detecting sentiments in mass media publications using content specific Informative features // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty, - 2019. -P. 21-31 (In Russian)
67. Khoroshilov A.A., Musabayev R.R., Krasovitskiy A.M., Ualiyeva I.M., Meyrambekkyzy ZH. Avtomaticheskoye ustanovleniye tonal'nosti tekstov metodom ikh kontseptual'nogo analiza // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty, - 2019. -P. 365-375 (In Russian)
68. A.M. Krassovitsky, I.M. Ualiyeva, Zh. Meirambekkyzy, R.R. Mussabayev. In the search of a linguistic model for automatic generalization recognition in media texts // International Journal of Advances in Electronics and Computer Science. -2019. -Vol. 6(3). -P.38-43
69. Krasovitskiy A.M., Ualiyeva I.M., Meyrambekkyzy ZH., Musabayev R.R. Osnovannyy na leksikone podkhod v otsenke obobshcheniy v russkoyazychnykh SMI // Mezhdunar. nauch. zhurnal «Sovremennyye informatsionnyye tekhnologii i IT obrazovaniye». -2018. -Vol.14 (3). -P. 563-567 (In Russian)
70. Krassovitsky A., Ualiyeva I., Meirambekkyzy Zh., Mussabayev R. Automatic generalization recognition in media texts /// Proceedings 20th International Conference on Computational Linguistics and Intelligent Text Processing. -La Rochelle, France, 2019.
71. Ualiyeva I.M., Mussabayev R.R. Features analysis for detecting information-dense texts in online news media // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part 2. -P. 74-79
72. Barakhnin V.B., Kozhemyakina O.YU., Rychkova Ye.V., Pastushkov I.S., Borzilova YU.S. Izvlecheniye leksicheskikh i metroritmicheskikh priznakov, kharakternykh dlya zhanra i stilya i ikh kombinatsiy v protsesse avtomatizirovannoy obrabotki tekstov na russkom yazyke // Mezhdunar. nauch. zhurnal «Sovremennyye informatsionnyye tekhnologii i IT obrazovaniye» – 2018. – Vol. 14(4). – P.876-883. – <http://dx.doi.org/10.25559/SITITO.14.201803.876-883>: 21.10.2019 (In Russian)
73. Ualiyeva I.M., Musabayev R.R. Issledovaniye priznakov dlya avtomaticheskoy identifikatsii zhanrov publikatsiy v SMI // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty, - 2019. -P. 331-336 (In Russian)
74. Barakhnin V.B., Mukhamediyev R.I., Musabayev R.R., Kozhemyakina O.YU., Isayeva A., Kuchin YA.I., Murzakhmetov S.B., Yakunin K.O. Metody vyyavleniya destruktivnykh novostey v mediaprostranstve // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part 2. -P. 205-219 (In Russian)
75. Barakhnin V.B., Yakunin K. Methods to identify the destructive information // Journal of Physics: Conference Series. - 2019.
76. Atanayeva M.K., Ospanova U.A., Buldybayev T.K., Akoyeva I.G., Nurumov K.S. Opredeleniye indikatorov otsenki dostovernosti publikatsiy v SMI dlya razrabotki metodiki opredeleniya stepeni veroyatnoy dostovernosti otdel'noy publikatsii // Vestnik KazNPU, Seriya filologicheskaya. -2018. -№2 (64). -P . 158-165 (In Russian)
77. Ospanova U.A., Atanayeva M.K., Akoyeva I.G., Buldybayev T.K. Identification of features of probable reliability of news // Modern teacher education. -Moscow, 2019. -№6. -P. 44-49
78. Atanayeva M.K., Buldybayev T.K., Ospanova U.O., Akoyeva I.G., Nurumov K.S., Baymakhanbetov M.A. Opredeleneiye zakonomernostey podozritel'nosti v novostnykh tekstakh // Vestnik YENU im. L.N. Gumileva, Seriya zhurnalistika. - Nur-Sultan, 2019. -№2 (127). -P 8-19 // DOI: <https://doi.org/10.32523/2616-7174-2019-127-2-8-20> : 21.10.2019 (In Russian)
79. Ospanova U.A., Atanayeva M.K., Akoyeva I.G., Buldybayev T.K. Informative features of bias and reliability of electronic Mass Media // Sociology. - Moscow, 2019. -№2. –P. 259-269
80. А.s. № 2894. NLP-Preprocessor / Yakunin K., Mukhamediyev R., Kuchin YA., Murzakhmetov S., Symagulov A., Mustakayev R., publ. 18.04.2019. -1 p.
81. Mladenovic N., Krassovitskiy A., Mussabayev R. Decomposition approach for clustering problem // Mater. scien. conf. IIVT MON RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 15-21
82. Kozbagarov O.B., Musabayev R.R. Klasterizatsiya novostnykh publikatsiy po informatsionnym povodam // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part 2. -P. 340-350 (In Russian)
83. Zakharov V.N., Musabayev R.R., Krasovitskiy A.M., Kozlovskaya YA.D., Khoroshilov A-r A., Khoroshilov A-y A. Metod klasterizatsii novostnykh soobshcheniy sredstv massovoy informatsii na osnove ikh kontseptual'nogo analiza // Sistemy i sredstva informatiki. -2019. -Vol.29 № 3. -P. 52-66 // DOI: 10.14357/08696527190305: 21.10.2019 (In Russian)
84. Mladenovich N., Krasovitskiy A., Musabayev R. Metod dekompozitsii v zadache klasterizatsii bol'shikh dannykh // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part 2. -P. 525-533 (In Russian)
85. Karlsson, C. (ed.): Handbook of Research on Cluster Theory // Edward Elgar Publishing, - 2010.
86. Krassovitskiy A., Mussabayev R. Energy-Based Centroid Identification and Cluster Propagation with Noise Detection.// Nguyen N., Pimenidis E., Khan Z., Trawiński B. (eds) Computational Collective Intelligence. ICCCI 2018. Lecture Notes in Computer Science. -Springer, Cham. -2018. -Vol. 11055. –P. 523-533
87. Bagirov, A.M.: Modified global k-means algorithm for minimum sum-of-squares clustering problems. Pattern Recognition. -2008. –Vol. 41 (10). -P. 3192—3199 // http://www.sciencedirect.com/science/article/pii /S0031320308001362 : 09.06.2019
88. Bahmani B., Moseley B., Vattani A., Kumar R., Vassilvitskii S., Scalable k-means++ // Proc. VLDB Endow. -2012. -№5(7). -P. 622–633
89. Capó M., Pérez A., Lozano J.A., An efficient approximation to the k-means clustering for massive data. Knowl.-Based Syst. -2017. –Т. 117. –P. 56–69
90. Chiang M.C., Tsai C.W., Yang C.S., A time-efficient pattern reduction algorithm for k-means clustering // Inform. Sci. -2011. -№181(4). -P. 716–731
91. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J. and Steinberg, D. Top 10 Algorithms in Data Mining // Knowl. Inf. Syst. -2007. -№14(37) // DOI 10.1007/s10115-007-0114-2, 2008: 21.10.2019
92. HajKacem M.A.B., N’Cir CE.B., Essoussi N. Overview of Scalable Partitional Methods for Big Data Clustering // Nasraoui O., Ben N'Cir CE. (eds) Clustering Methods for Big Data Analytics. Unsupervised and Semi-Supervised Learning. -Springer, Cham, 2019. –P. 1-23
93. Comparison of the K-Means and MiniBatchKMeans clustering algorithms// <https://scikit-learn.org/stable/auto_examples/cluster/plot_mini_batch_kmeans.html> : 09.06.2019
94. S.J. Redmond, C. Heneghan, A method for initialising the K-means clustering algorithm using kd-trees// Pattern Recogn. Lett. -2007. -№28(8). -P. 965–973
95. Comparing different clustering algorithms on toy datasets// URL: http:// scikit-learn.org/stable/auto\_examples/cluster/plot\_cluster\_comparison.html : 09.06.2019
96. X. Cui, P. Zhu, X. Yang, K. Li, C. Ji, Optimized big data K-means clustering using MapReduce. J.// Supercomput. -2014. -№70(3). –P. 1249–1259
97. T. Washio, K. Nakanishi, H. Motoda. Deriving Class Association Rules Based on Levelwise Subspace Clustering. // In: Jorge A.M., Torgo L., Brazdil P., Camacho R., Gama J. (eds) Knowledge Discovery in Databases.PKDD 2005. Lecture Notes in Computer Science. - Springer, Berlin, Heidelberg. 2005. -Vol. 3721. –P. 692-700
98. Online clustering data sets UCI// URL https://archive.ics.uci.edu/ml/datasets.html. : 09.06.2019
99. Akhmetov I.R., Krassovitsky A., Ualiyeva I., Myssabayev R.R. Lemmatization of Russian Language by Tree Regression Models // Proceedings 20th International Conference on Computational Linguistics and Intelligent Text Processing, -La Rochelle, France, 2019.
100. Mukhamediyev R.I., Symagulov A., Kuchin YA., Abdullayeva S., Abdoldina F.N. Oblachnyye servisy dlya obrabotki tekstov na yestestvennom yazyke // Mezhdunar. nauch. zhurnal «Sovremennyye informatsionnyye tekhnologii i IT obrazovaniye». – 2018. – Vol.14(4). –P.859-869 (In Russian)
101. Barakhnin V.B., Kozhemyakina O.Yu., Rychkova E.V., Borzilova Yu.S. Automatic extraction of phrases from texts of Slavic origin: a comparison of approaches // Cloud of Science,– 2018. –Vol.5(4). –P.713-728 (In Russian)
102. Seitkali D., Mussabayev R. Using Centroid Keywords and Word Mover’s Distance for Single Document Extractive Summarization // ACM International Conference Proceeding Series, 3rd International Conference on Natural Language Processing and Information Retrieval, NLPIR 2019, -Tokushima, Japan, 2019. –P. 149-152. // DOI:10.1145/3342827.3342852: 21.10.2019
103. Musabayev R.R., Seytkali D. Referirovaniye tekstovogo dokumenta s pomoshch'yu word mover’s distance i izvlechennykh klyuchevykh slov dokumenta // Mater. nauch. konf. IIVT MON RK «Sovremennyye problemy informatiki i vychislitel'nykh tekhnologiy», - Almaty, - 2019. -P. 247-253
104. Tolegen G., Toleu A., Mussabayev R. Мoted perceptron approach for morphological disambiguation // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part 2. -P. 68-74
105. Musabayev R.R., Kozbagarov O.B., Kasymzhanov B.K., Mukashev A.SH., Ibrayeva V.M., Merkebayev A. Obucheniya distributivnykh modeley dlya predobrabotki tekstov v WORD2VEC i GLOVE // Mater. IV International scientific and practical conf. “Informatics and Applied Mathematics” dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsyk and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019. -Part 2. -P. 405-414 (In Russian)
106. Dragomir R. Radev, Hongyan Jing, Malgorzata Stys, and Daniel Tam. Centroid-based summarization of multiple documents // Inf. Process. Manage. - 2004. – P. 919-938
107. Pereira T., Aloise D., Brimberg J., Mladenović N. Review of Basic Local Searches for Solving the Minimum Sum-of-Squares Clustering Problem // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. Springer Optimization and Its Applications, Springer, Cham. -2018. –Vol. 141. –P. 249-270
108. Mladenovic N., Alkandari A., Pei J., Todosijevic R., Pardalos P.M. Less is more approach: basic variable neighborhood search for the obnoxious p-median problem // International transactions in operational research, Intl. Trans. in Op. Res. -2019. -№00. –P.1–14
109. Cadi AA. El., Ratli M., Mladenović N. New MIP model for multiprocessor scheduling problem with communication delays // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. -Springer Optimization and Its Applications, -Springer, Cham. – 2019. -Т. 141. –P. 129-149
110. Brimberg J., Mladenovic N., Todosijevic R., Uroševic D. A non-triangular hub location problem // Optimization Letters, Springer Verlag, - 2019. –P. 1-20
111. Pei J., Drazic Z., Drazic M., Mladenovic N., Pardalos P.M. Continuous Variable Beighborhood Search (C-VNS) for Solving Systems of Nonlinear Equations // Informs journal on computing, 2019. -Vol.31(2). -P. 193-410
112. Ivanov S. V., Kibzun A. I., Mladenović, N. Variable Neighborhood Search for a Two-Stage Stochastic Programming Problem with a Quantile Criterion. Automation and Remote Control // Automation and Remote Control, 2019. -№ 80(1), -P. 43–52
113. Ivanov S. V., Kibzun A. I., Mladenović, N., Urosevic D. Variable neighborhood search for stochastic linear programming problem with quantile criterion // Journal of global optimization, --2019. Vol. 74(3). -P. 549-564
114. Mussabayev R., Kassymzhanov B., Mukashev A., Ibrayeva V., Merkebayev A. Creation of necessary technical and expert-analytical conditions for development of the information system of evaluating open text information sources' influence on society // Proceedings 15th International Asian School-Seminar "Optimization Problems of complex systems", - Novosibirsk, Russia, 2019.

# 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

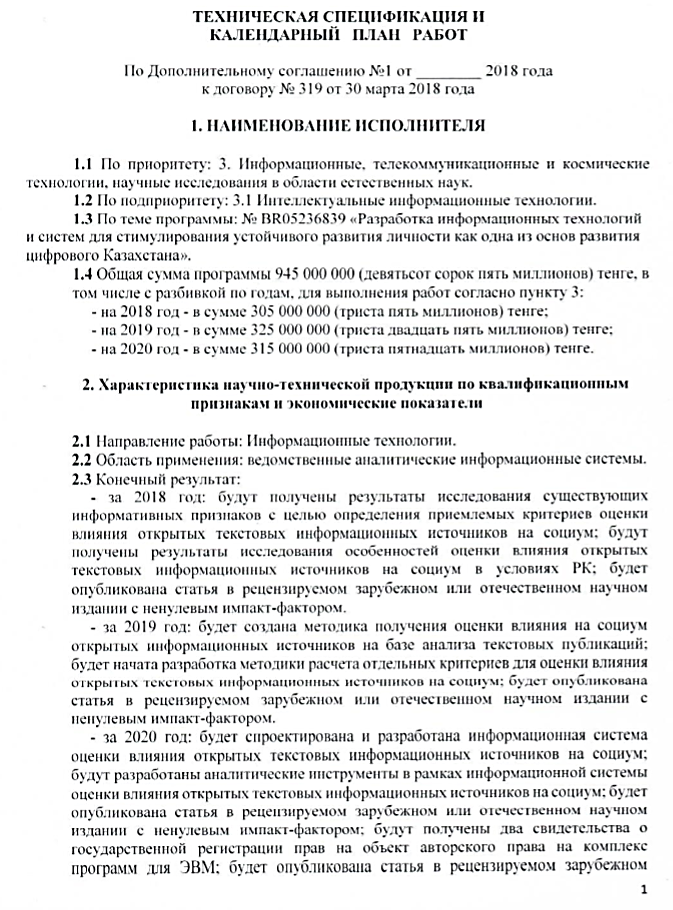
designed and developed; analytical tools within the information system evaluating open information sources' influence on society will be developed: two certificates of state registration of rights for subject of copyright to a set of computer programs will be obtained; an article in a peer-reviewed foreign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

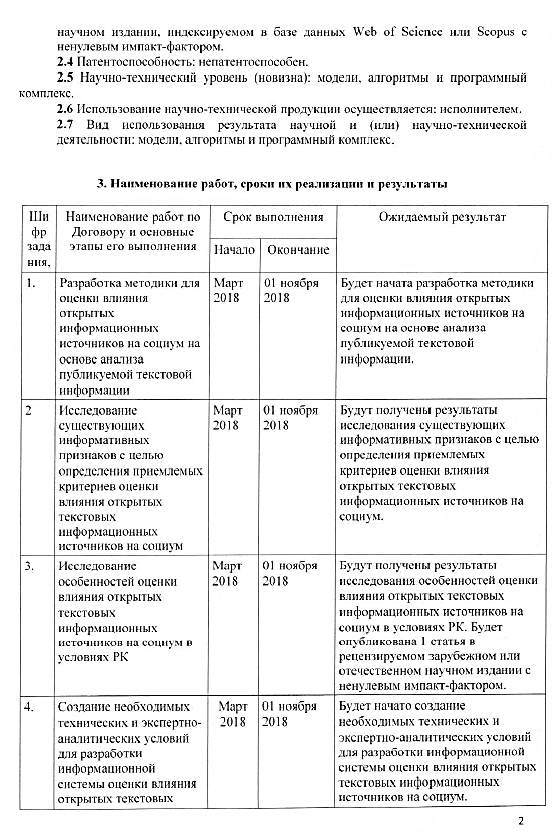
1. Patentability: nonpatentable.
2. Scientific and technical level (novelty): models, algorithms and software complex.
3. The use of scientific and technical products is carried out: by the contractor.
4. Type of the scientific and (or) scientific and technical activities result's use: models, algorithms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

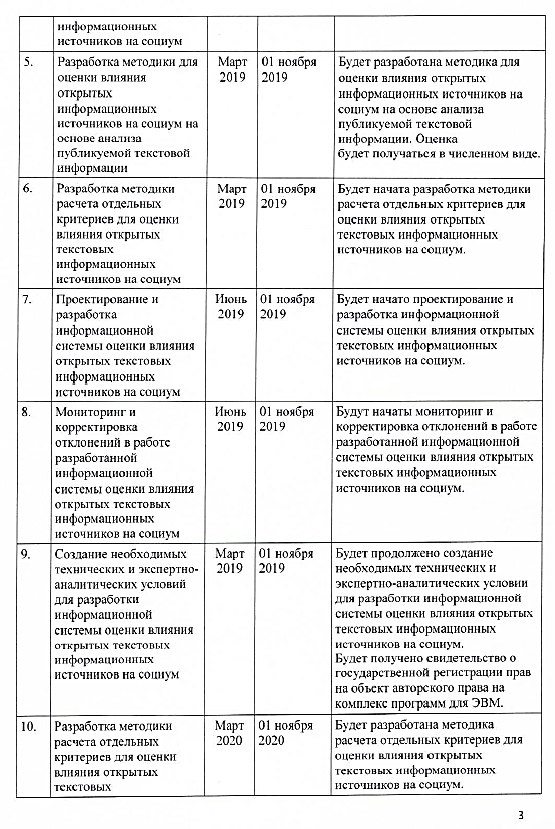
3. Activities, implementation dates and results

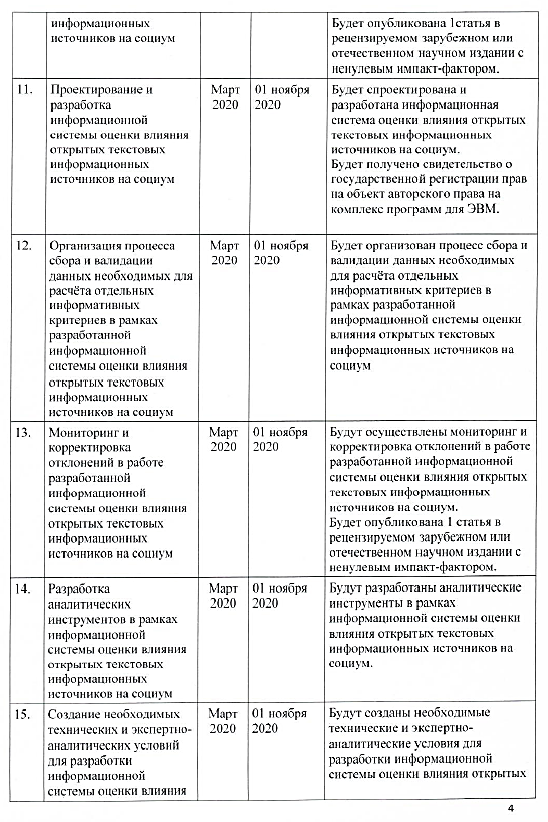
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task cipher | Activities under the Contract and the main stages of its implementation | Implementation date | | Expected result |
| Start | End |
|  | Development of methods evaluating the open textual information sources' influence on society on the 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. Toleu A., Tolegen G., Mussabayev R.: KeyVector Unsupervised Keyphrase Extraction Using Weighted Topic via Semantic Relatedness // Computación y Sistemas, 2019. -Vol. 23(3). -P. 861–869 // doi: 10.13053/CyS-23-3-3264 : 23.10.2019 (Impact Factor – 0.53)
2. 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. -Vol. 1117(1), -P.1-9 (Impact Factor = 0.48)
3. Barakhnin V. B., Duisenbayeva A. N., Kozhemyakina O.Yu., Yergaliyev Y. N. and Muhamedyev R. I. The automatic processing of the texts in natural language. Some bibliometric indicators of the current state of this research area//Journal of Physics: Conference Series. -2018. // <https://doi.org/10.1088/1742-6596/1117/1/012001> : 21.10.2019 (Impact factor = 0.48)
4. Pereira T., Aloise D., Brimberg J., Mladenović N. Review of Basic Local Searches for Solving the Minimum Sum-of-Squares Clustering Problem // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. Springer Optimization and Its Applications, Springer, Cham, 2018. -Vol 141. –P. 249-270 // <https://doi.org/10.1007/978-3-319-99142-9_13>: 21.10.2019 (Impact Factor – 0,31)
5. Cadi AA. El., Ratli M., Mladenović N. New MIP model for multiprocessor scheduling problem with communication delays // In: Pardalos P., Migdalas A. (eds) Open Problems in Optimization and Data Analysis. Springer Optimization and Its Applications. Springer, Cham, -2018. –Vol. 141. –P. 129-149 <https://doi.org/10.1007/978-3-319-99142-9_8> (Impact Factor – 0,31)
6. Brimberg J., Mladenovic N., Todosijevic R., Uroševic D. A non-triangular hub location problem // Optimization Letters, Springer Verlag, 2019. –P. 1-20 // <https://doi.org/10.1007/s11590-019-01392-2> : 21.10.2019 (Impact Factor – 1,399)
7. Pei J., Drazic Z., Drazic M., Mladenovic N., Pardalos P.M. Continuous Variable Beighborhood Search (C-VNS) for Solving Systems of Nonlinear Equations // Informs journal on computing. -2018. -Vol.31(2). -P. 193-210. // <https://doi.org/10.1287/ijoc.2018.0876> : 21.10.2019 (Impact Factor – 2.39)
8. Mladenovic N., Alkandari A., Pei J., Todosijevic R., Pardalos P.M. Less is more approach: basic variable neighborhood search for the obnoxious p-median problem // International transactions in operational research, Intl. Trans. in Op. Res. -2019. –Vol.00. –P. 1–14 // <https://doi.org/10.1111/itor.12646> :21.10.2019 (Impact Factor – 2.51)
9. Ivanov S. V., Kibzun A. I., Mladenović, N. Variable Neighborhood Search for a Two-Stage Stochastic Programming Problem with a Quantile Criterion. Automation and Remote Control // Automation and Remote Control. -2019. -Vol. 80(1), -P. 43–52. // <https://doi:10.1134/s0005117919010041> : 21.10.2019 (Impact Factor – 0.86)
10. Ivanov S. V., Kibzun A. I., Mladenović, N., Urosevic D. Variable neighborhood search for stochastic linear programming problem with quantile criterion // Journal of global optimization, 2019. -Vol. 74(3). –P.549-564 // <https://doi.org/10.1007/s10898-019-00773-2> : 21.10.2019 (Impact Factor – 1.631)
11. A.M. Krassovitsky, I.M. Ualiyeva, Zh. Meirambekkyzy, R.R. Mussabayev. In the search of a linguistic model for automatic generalization recognition in media texts // International Journal of Advances in Electronics and Computer Science, 2019. -Vol. 6(3). -P.38-43, (JIFACTOR – 2.68)
12. Krasovitsky A.M., Ualieva I.M., Meirambekkyzy J., Musabaev R.R. Lexicon-based approach in generalization evaluation in russian language media // International Scientific Journal "Modern Information Technologies and IT Education". -2018. -Vol. 14(3). -P. 563-567 (Impact factor RSCI - 0,229)
13. Mukhamediyev R.I., Symagulov A., Kuchin YA., Abdullayeva S., Abdoldina F.N. Oblachnyye servisy dlya obrabotki tekstov na yestestvennom yazyke // International Scientific Journal "Modern Information Technologies and IT Education". -2018. -Vol. 14(4). -P. 859-869 (Impact factor RSCI - 0,229) (In Russian)
14. Barakhnin V.B., Kozhemyakina O.YU., Rychkova Ye.V., Pastushkov I.S., Borzilova YU.S. Izvlecheniye leksicheskikh i metroritmicheskikh priznakov, kharakternykh dlya zhanra i stilya i ikh kombinatsiy v protsesse avtomatizirovannoy obrabotki tekstov na russkom yazyke // International Scientific Journal "Modern Information Technologies and IT Education". -2018. -Vol. 14(4). -P. 876-883. // <http://dx.doi.org/10.25559/SITITO.14.201803.876-883> : 21.10.2019 (Impact factor RSCI - 0,229) (In Russian)
15. Barakhnin V.B., Kozhemyakina O.YU., Rychkova Ye.V., Borzilova YU.S. Avtomaticheskoye vydeleniye slovosochetaniy iz tekstov slavyanskogo proiskhozhdeniya: sravneniye podkhodov // Cloud of Science. – 2018. – Vol.5(4). –  P.713-728 (Impact factor RSCI 0.533) (In Russian)
16. Nurumov K.S., Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G. Kompleksnaya otsenka znachimosti informativnykh priznakov dlya otsenki vliyaniya na sotsium na osnove analiza publikuyemoy tekstovoy informatsii // Vestnik YENU im. L.N. Gumileva, Seriya Zhurnalistika, -2018. -№3 (124). -P. 44-52 (In Russian)
17. Atanayeva M.K., Ospanova U.A., Buldybayev T.K., Akoyeva I.G., Nurumov K.S. Opredeleniye indikatorov otsenki dostovernosti publikatsiy v SMI dlya razrabotki metodiki opredeleniya stepeni veroyatnoy dostovernosti otdel'noy publikatsii // Vestnik KazNPU, Seriya filologicheskaya, -2018. -№2 (64). -P. 158-165 (In Russian)
18. Atanayeva M.K., Ospanova U.A., Buldybayev T.K., Akoyeva I.G., Nurumov K.S. Ispol'zovaniye lingvisticheskikh podkhodov v opredelenii klassov i informativnykh priznakov v tekstakh // Vestnik KazNPU, Seriya filologicheskaya, 2018. -№2 (64). -P. 166-171 (In Russian)
19. Ospanova U.A., Atanayeva M.K., Akoyeva I.G., Buldybayev T.K. Identification of features of probable reliability of news // Modern teacher education. - Moscow, 2019. -№6. -P. 44-49
20. Atanayeva M.K., Buldybayev T.K., Ospanova U.O., Akoyeva I.G., Nurumov K.S., Baymakhanbetov M.A. Opredeleniye zakonomernostey podozritel'nosti v novostnykh tekstakh // Vestnik YENU im. L.N. Gumileva, Seriya zhurnalistika. -Nur-Sultan, 2019. -№2 (127). -P 8-19 (In Russian)
21. Ospanova U.A., Atanayeva M.K., Akoyeva I.G., Buldybayev T.K. Informative features of bias and reliability of electronic Mass Media // Sociology. -Moscow, 2019. -№2. -P. 259-269
22. Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G., Nurumov K.S., Baymakhanbetov M.A. Metodika dlya opredeleniya informativnykh priznakov novostnykh tekstov i proverka ikh znachimosti // Nauchnyy aspekt. -2019. -Vol.3(3). -P. 277-296 (Impact factor RSCI 0,03) (In Russian)
23. Zakharov V.N., Musabayev R.R., Krasovitskiy A.M., Kozlovskaya YA.D., Khoroshilov A-r A., Khoroshilov A-y A. Metod klasterizatsii novostnykh soobshcheniy sredstv massovoy informatsii na osnove ikh kontseptual'nogo analiza // Sistemy i sredstva informatiki, 2019. -Vol.29 (3). -P. 52-66 (In Russian)
24. Seitkali D., Mussabayev R. Using Centroid Keywords and Word Mover’s Distance for Single Document Extractive Summarization // ACM Intern. Conf. Proc. Series, 3rd Inter. Conf. on NLP and Information Retrieval, NLPIR 2019. -Tokushima, Japan, 2019. –P. 149-152 (Scopus)
25. Barakhnin V.B., Duysenbaeva A., Kozhemyakina O.Yu., Kuchin Y.I., Yakunin K.O., Mukhamediev R.I. The studies of dinamics’ changes in publication activity in the field of natural language processing// Mater. 17th Int. scientific conf. "Information Technology and Management 2019". - Riga, Latvia, 2019. -P. 71-74
26. Musabayev R.R., Mukhamediyev R.I., Kuchin YA.I., Symagulov A., Yakunin K.O. On a method of multimedial media ranking using corpus based topic modeling// Mater. 17th Int. scientific conf. "Information Technology and Management 2019". - Riga, Latvia, 2019. -P. 64-67
27. Mladenovic N., Krassovitskiy A., Mussabayev R. Decomposition approach for clustering problem // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 15-21
28. Nurumov K., Baimakhanbetov M., Buldybayev T., Akoyeva I., Ospanova U. An approach for detecting sentiments in mass media publications using content specific Informative features // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 21-31
29. Toleu A., Tolegen G., Mussabayev R. Initial exploration for Topic modeling with word sense // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 31-44
30. Barakhnin V., Duysenbayeva A., Kozhemyakina O., Kuchin YA., Yakunin K., Mukhamediyev R. Izmeneniye publikatsionnoy aktivnosti v oblasti obrabotki yestestvennogo yazyka // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 130-135 (In Russian)
31. Meyrambekkyzy ZH., Amanbay A., Musabayev R. Avtomaticheskoye formirovaniye tematicheskikh slovarey sotsial'no-znachimykh ponyatiy // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 233-239 (In Russian)
32. Musabayev R., Mukhamediyev R., Kuchin YA., Symagulov A., Yakunin K., Murzakhmetov S. Metod mul'timodal'noy otsenki sredstv massovoy informatsii na osnove tematicheskoy modeli korpusa tekstov // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 239-247 (In Russian)
33. Musabayev R.R., Seytkali D. Referirovaniye tekstovogo dokumenta s pomoshch'yu word mover’s distance i izvlechennykh klyuchevykh slov dokumenta // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 247-253 (In Russian)
34. Ualiyeva I.M., Musabayev R.R. Issledovaniye priznakov dlya avtomaticheskoy identifikatsii zhanrov publikatsiy v SMI // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 331-336 (In Russian)
35. Khoroshilov A.A., Musabayev R.R., Krasovitskiy A.M., Ualiyeva I.M., Meyrambekkyzy ZH. Avtomaticheskoye ustanovleniye tonal'nosti tekstov metodom ikh kontseptual'nogo analiza // Mater. Scientific conf. IICT MES RK "Modern problems of informatics and computational technologies", - Almaty, - 2019. -P. 365-375 (In Russian)
36. Meirambekkyzy Zh., Mussabaev R.R. Automatic development of topic dictionaries of socially significant concepts // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 1. –P. 65-70
37. Buldybayev T., Baimakhanbetov M., Nurumov K., Akoyeva I., Ospanova U., Atanayeva M. Analysis of digital media: a comprehensive assessment of informative features significance // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 31-44
38. Tolegen G., Toleu A., Mussabayev R. Мoted perceptron approach for morphological disambiguation // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 68-74
39. Ualiyeva I.M., Mussabayev R.R. Features analysis for detecting information-dense texts in online news media // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 74-79
40. Barakhnin V.B., Mukhamediyev R.I., Musabayev R.R., Kozhemyakina O.YU., Isayeva A., Kuchin YA.I., Murzakhmetov S.B., Yakunin K.O. Metody vyyavleniya destruktivnykh novostey v mediaprostranstve // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 205-219 (In Russian)
41. Kozbagarov O.B., Musabayev R.R. Klasterizatsiya novostnykh publikatsiy po informatsionnym povodam // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 340-350 (In Russian)
42. Musabayev R.R., Kozbagarov O.B., Kasymzhanov B.K., Mukashev A.SH., Ibrayeva V.M., Merkebayev A. Obucheniya distributivnykh modeley dlya predobrabotki tekstov v WORD2VEC i GLOVE // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 405-414 (In Russian)
43. Mukhamediyev R.I., Musabayev R.R., Buldybayev T., Kuchin YA., Symagulov A., Ospanova U., Yakunin K., Murzakhmetov S., Sagyndyk B. Eksperimenty po otsenke sredstv massovoy informatsii na osnove tematicheskoy modeli korpusa tekstov // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 2. –P. 419-431 (In Russian)
44. Mladenovich N., Krasovitskiy A., Musabayev R. Metod dekompozitsii v zadache klasterizatsii bol'shikh dannykh // Mater. IV Intern. Scient. and prac. conf. “Informatics and Applied Mathematics”, dedicated to the 70th anniversary of Prof. T. Biyarov, V. Vuytsik and the 60th anniversary of Prof. E. Amirgaliev - Almaty, Kazakhstan, 2019.-Part 1. –P.525-533 (In Russian)
45. Baymakhanbetov M. A., Nurumov K. S., Ospanova U. A., Buldybayev T. K., Akoyeva I. G. Vliyaniye metoda ADASYN na shiroko rasprostranennyye metriki effektivnosti mashinnogo obucheniya // Sovremennyye informatsionnyye tekhnologii i IT-obrazovaniye, -2019. -Vol.15(2). –P. 290-297 (In Russian)
46. Atanayeva M.K., Buldybayev T.K., Ospanova U.A., Akoyeva I.G., Nurumov K.S., Baymakhanbetov M.A. Opredeleniye tonal'nosti i ob"yektivnosti novostnykh tekstov slovarnym podkhodom // Nauchnyy aspekt, -2019. -Vol.3(3). -P.296-308 (ИФ РИНЦ 0,03) (In Russian)
47. А.s. № 2894. NLP-Preprocessor / Yakunin K., Mukhamediyev R., Kuchin YA., Murzakhmetov S., Symagulov A., Mustakayev R., publ. 18.04.2019. -1 p.
48. А.с. №4523. The program of multimodal media assessment based on the thematic model of the body of texts (Multi Modal Mass Media Assesment – M4A)/ Yakunin K.O., Mukhamediyev R.I., Kuchin YA.I., Murzakhmetov S.B., Symagulov A., Musabayev R.R.; publ. 15.07.2019. -1 p.
49. А.с. № 4505. The program for calculating dynamic bibliometric indicators D1 and D2 (Calculation of dynamic scientometric indicators D1, D2 – CalcDStMi)/ Yakunin K.O., Mukhamediyev R.I., Kuchin YA.I.; publ. 12.07.2019. – 1p.

# 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

## The main parameters of media texts.

Table D1 - List of parameters for the classification of resonant / non-resonant publications.

|  |  |
| --- | --- |
| Parameter | Weight (CW) |
| Media coverage (number of views) for a certain period | 0.99 |
| The number of reposts of information for a certain period | 0.99 |
| The number of information shears for a certain period | 0.577505268 |
| Number of publication comments for a certain period | 0.806507958 |
| Resonance Coverage | 0.956947071 |
| Sentiment analysis of the comments (Negative / Neutral / Positive) | 0.653676928 |

Table D2 - List of parameters for the classification of socially significant publications

|  |  |
| --- | --- |
| Parameter | Weight (CW) |
| The presence / absence of a link in the publication to an proper source | 0.061593752 |
| Link to the primary source | 0.16683118 |
| Coverage of the same event by various publications: cross-checking, the presence of discourse with other articles | 0.21234652 |
| Authenticity of the photo that confirms the event | 0.142772968 |
| Authenticity of the video that confirms the event | 0.14244636 |
| The fact of subsequent editing in an already published article | 0.045305297 |
| Compliance of the publication title with the content of the media text | 0.108959251 |
| Presence / absence of the author (his name, if any) | 0.086821073 |
| The absence of verifiable facts in the article | 0.202623491 |
| Reputation of the publication (websites of news and information agencies, electronic media, other similar sites) on which information is published | 0.154870571 |
| Media coverage (number of views) for a certain period | 0.211239893 |
| The number of reposts of information for a certain period | 0.193185632 |
| The number of information shears for a certain period | 0.174316737 |
| Number of publication comments for a certain period | 0.130902894 |
| Publication tone (Negative / Neutral / Positive) | 0.069418546 |
| Hate comments (hate detection) | 0.083327509 |
| Politicization of the media text | 0.103780306 |
| Degree of generalization: strongly expressed, weakly expressed, no | 0.063146285 |
| Call to action | 0.140608592 |
| The tone of the comments (Negative / Neutral / Positive) | 0.087363666 |
| Significance for socially active groups | 0.137737477 |

Table D3 - List of parameters for classification as reliable / not reliable

|  |  |
| --- | --- |
| Parameter | Weight (CW) |
| The presence / absence of a link in the publication to an proper source | 0.363788658 |
| Link to the primary source | 0.359066664 |
| Coverage of the same event by various publications: cross-checking, the presence of discourse with other articles | 0.535835773 |
| Authenticity of the photo that confirms the event | 0.205189785 |
| Authenticity of the video that confirms the event | 0.196834318 |
| The fact of subsequent editing in an already published article | 0.1662592 |
| Compliance of the publication title with the content of the media text | 0.178352836 |
| Presence / absence of the author (his name, if any) | 0.1154985 |
| The absence of verifiable facts in the article | 0.340389563 |
| Reputation of the publication (*websites of news and information agencies, electronic media, other similar sites*) on which information is published | 0.463591703 |

Table D4 - List of parameters for classification as objective / not objective

|  |  |
| --- | --- |
| Parameter | Weight(CW) |
| The presence in the media text of the expressed personal opinion of the author of the publication | 0.72462313 |
| The presence of manipulative techniques in the media text | 0.99 |
| Degree of generalization: weakly expressed, strongly expressed, no | 0.478821031 |
| Politicization | 0.99 |
| The tone of publication (Negative / Neutral / Positive) | 0.867781263 |
| Tonality of comments (negative, neutral, positive) | 0.628555507 |

# APPENDIX E

# Experiments on the applicability of the multimodal media estimation algorithm

Experiment 1

The applicability of the described algorithm was tested on a small corpus of docu-ments, conditionally distributed between the two media sources.The matrices p1, ...,p6 were calculated using the Jacquard measure, which is often used as a measure of document proximity in clustering algorithms. Briefly, the initial data for the algo-rithm operation are as follows:

The corpus, consisting of 5 articles (m = 5), in the following areas:

1. Economy
2. Sports (boxing)
3. Politics
4. Show business
5. Education and science

The specified corpus of texts is manually divided into 4 thematic clusters (k = 4)

1. Politics
2. Sport
3. Show business
4. Economy and Finance

Consider classes (c = 2):

1. Socially significant
2. Objective

and two media sources.

There are 4 features (q = 4), each has its own glossary of terms [40].

1. Manipulative capability
2. Politicization
3. Negative tone
4. Positive tone

Conditional probability distributions of topics is calculated for articles presented in the form of a matrix p2 [1..k] [1..m], where k is the number of topics (line by line), m- articles (by columns)

p2 = [[0.03, 0.0, 0.03, 0.004, 0.005],

[0.0, 0.01, 0.0, 0.008, 0.005],

[0.0, 0.007, 0.004, 0.03, 0.005],

[0.04, 0.0, 0.007, 0.0, 0.005]]

Weights of each attribute are obtained for classes in the form of the matrix p3 [1..c] [1..q] using the analytical hierarchical process (AHP), where c is the classes, q are the features

p3 = [[0.55, 0.27, 0.18, 0.18],

[0.23, 0.43, 0.34, 0.34]]

Matrices calculation provided the following results p4 [1..k] [1..c] (conditional probabilities of the distribution of topics across classes), p5 [1..m] [1..c] (conditional probabilities of document distribution by classes), p6 [1..m] [1..q] (conditional proba-bilities of the distribution of documents by features):

p4= [[0.1967 0.2239]

[0.0371 0.0531]

[0.0316 0.0508]

[0.0427 0.0443]]

p5= [[0.007609 0.008489 ]

[0.0005922 0.0008866]

[0.0063263 0.0072303]

[0.0020316 0.0028444]

[0.0015405 0.0018605]]

p6= [[0.0058 0.0149 0.0014 0.0008 ]

[0.00027 0.00017 0.00055 0.00166]

[0.00452 0.01329 0.00094 0.00046]

[0.00102 0.0021 0.00174 0.00328]

[0.00105 0.0025 0.00055 0.00105]]

The processes of formation dictionaries of features, the choice of classes for fea-tures, the formation of AHP tables, etc. are not considered in this paper.

Based on these values, using the multimodal media assessment algorithm, mass media distribution by topics, classes, and characteristics was calculated under the assumption that the articles were provided by two mediasources.The first of which published articles number 1 and 3 (economic and political orientation), the second is the source of articles 2, 4,5 (“sport (boxing)”, “show business”, “education and sci-ence”).

Table E1 – The distribution of media by topics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Media\Topics: | Politics | Sports | Show Business | Economy and Finance | Articles |
| Mass Media0 | 0.971327 | 0.057346 | 0.431078 | 0.92300 | [economic, political] |
| Mass Media1 | 0.458623 | 0.927009 | 0.969495 | 0.06101 | [sports (boxing), show business, education and science] |

|  |  |  |  |
| --- | --- | --- | --- |
| Media\Classes: | Socially significant | Objective | Articles |
| Mass Media0 | 0.597028 | 0.701486 | [economic, political] |
| Mass Media1 | 0.460911 | 0.769545 | [sports (boxing), show business, education and science] |

Table E2 – The distribution of media to classes

Table E3 – Distribution of media by features

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Media/Features: | Manipulative capability | Politicization | Negative  Tone | Positive  Tone | Articles |
| Mass Media0 | **0.675538** | **0.917128** | **0.317990** | **0.165743** | [economic, political] |
| Mass Media1 | **0.273881** | **0.496645** | **0.392071** | **0.863059** | [sports (boxing), show business, education and science] |

The experiment on the "toy"sample showed logical results. Articles of the corre-sponding topics have determined the distribution of the probabilities of hypotheses H1, H2, H3 for two nominal media. It can be seen that Mass Media0 and Mass Me-dia1 differ in objectivity, manipulative capability, and emotional tone. For example, Mass Media0 publishes in the field of economics and politics;its news is much less positive, more politicized, as well as it uses manipulative techniques more than MassMedia1.

Experiment 2

The second experiment used a corpus of Tengrinews agency articles (https://tengrinews.kz/) consisting of 20,000 articles for the period from 2014 to 2019. The purpose of the experiment was a comparative analysis of publications over time. The articles were sorted by time, and 5 arrays of publications were formed: 2014 and 2015, 2015, 2017, 2018, 2019. The set of classes, parameters and diction-aries remained the same as in Experiment 1. The results of the experiment are shown in tables E4 and E5.

Table E4 – Distribution of publications by class

|  |  |  |
| --- | --- | --- |
| MassMedia: | Socially Significant | Objective |
| 2014, 2015 | **0.001396** | **0.987428** |
| 2016 | **0.000669** | **0.993975** |
| 2017 | **0.000542** | **0.995114** |
| 2018 | **0.000719** | **0.993520** |
| 2019 | **0.006352** | **0.942828** |

Table E5 – Distribution of media by features

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| MassMedia: | Manipulative capability | Politicization | Negative | Mass Media: |
| 2014, 2015 | 0.00040 | 0.31335 | 0.99633 | 0.61259 |
| 2016 | 0.00019 | 0.30883 | 0.99826 | 0.63196 |
| 2017 | 0.00015 | 0.31287 | 0.99856 | 0.66602 |
| 2018 | 0.00021 | 0.31293 | 0.99810 | 0.69037 |
| 2019 | 0.00193 | 0.29577 | 0.98261 | 0.70545 |

Despite a very small volume of dictionaries and a limited set of parameters, the experiment showed a gradual increase in the positive tone of publications over time and a slight increase in socially significant content in recent years.

Experiment 3

For the third experiment, we have used a set of articles by the Tengri News (https://tengrinews.kz/) news agency consisting of 1800 articles marked up by experts, which made it possible to compare the results of the algorithm with expert estimates.

The following dictionaries of features have been used: Facts (657 lexical units), Negative tonality (2344), Positive tonality (1207), Politicization (35), Call to action (145), Manipulation (1072). Using expert estimates, the p3 matrix is formed (the importance of parameters for classes):

[[0.23, 0.18, 0.18, 0.2 , 0.21, 0. ],

[0. , 0.22, 0.22, 0.26, 0. , 0.3 ],

[1. , 0. , 0. , 0. , 0. , 0. ]],

where the columns (from left to right) correspond to the features listed above, and the rows (from top to bottom) to the classes: “Socially significant”, “Objective”, “Reliable”. Note that this matrix is not complete. Not all features for the corresponding classes are defined. Zero means that this attribute is not used to determine the probability p5. For example, it can be seen that the social significance of the text is determined by five criteria, while authenticity using only one (facts).

Figure E1 shows the tables where 5 news with the highest and 5 with the least probability of matching to the parameter or class (Aggregator score). The expert ratings are given in the column Expert Score.

|  |
| --- |
|  |
|  |
|  |

Figure E1 – Assessment of politicization, objectivity and reliability of media texts

The results of the algorithm show that in some cases it gives a good degree of agreement with expert estimates. On the other hand, it allows ranking texts in situations where texts are very close in their characteristics. The above texts are taken from one sufficiently reliable source; therefore, the reliability and objectivity of the texts are close. However, this approach requires very careful compilation of dictionaries and the use of additional algorithms to determine wider range of text parameters.

# APPENDIX F

# The results of ranking topics by method 1.3

(specified numbers - identifiers topics that correspond to those specified in the file Cluster No. NNN):

167, 135, 195, 68, 188, 107, 178, 164, 144, 166, 86, 203, 192, 134, 11, 95, 184, 29, 190, 75, 176, 139, 111, 67, 45, 13, 146, 53, 66, 185, 43, 14, 78, 89, 34, 197, 104, 152, 101, 7, 202, 85, 159, 149, 96, 172, 74, 28, 57, 133, 171, 102, 162, 1, 55, 120, 64, 12, 147, 150, 4, 76, 3, 105, 108, 191, 50, 154, 187, 2, 126, 193, 168, 124, 40, 87, 163, 8, 119, 186, 56, 6, 194, 23, 17, 71, 79, 99, 42, 180, 25, 37, 93, 30, 112, 138, 125, 80, 109, 59, 92, 0, 110, 122, 35, 173, 77, 136, 116, 129, 198, 151, 19, 63, 114, 100, 118, 15, 182, 62, 137, 88, 24, 196, 44, 5, 73, 91, 200, 16, 27, 84, 174, 60, 36, 21, 98, 201, 128, 183, 148, 123, 20, 38, 127, 65, 83, 177, 81, 61, 103, 170, 157, 70, 94, 181, 48, 54, 51, 117, 132, 106, 49, 46, 39, 121, 9, 47, 31, 199, 160, 142, 97, 141, 161, 69, 189, 22, 26, 145, 130, 90, 10, 155, 18, 179, 33, 72, 32, 113, 143, 169, 115, 58, 131, 140, 175, 156, 158, 52, 41, 153, 165, 82

The results of topic ranking according to the method 2.3:

202, 164, 188, 203, 134, 192, 107, 184, 10, 153, 100, 198, 29, 187, 59, 171, 143, 64, 3, 199, 176, 157, 155, 138, 50, 92, 123, 167, 108, 96, 136, 173, 170, 126, 194, 69, 97, 135, 49, 65, 190, 162, 93, 104, 141, 196, 174, 148, 23, 83, 68, 12, 201, 44, 45, 189, 163, 13, 183, 105, 41, 177, 168, 0, 161, 160, 9, 118, 142, 175, 150, 11, 25, 73, 28, 2, 182, 129, 144, 180, 89, 109, 8, 26, 106, 66, 127, 120, 165, 18, 131, 85, 1, 116, 27, 16, 80, 57, 88, 140, 152, 63, 75, 169, 159, 14, 110, 34, 102, 156, 151, 17, 115, 53, 181, 42, 191, 40, 81, 99, 172, 137, 39, 166, 101, 112, 128, 35, 111, 52, 48, 186, 32, 72, 122, 76, 178, 125, 15, 19, 149, 95, 90, 91, 193, 185, 36, 70, 87, 103, 117, 98, 94, 7, 5, 51, 31, 33, 119, 21, 200, 79, 158, 67, 82, 60, 121, 147, 56, 43, 77, 62, 47, 30, 145, 130, 24, 71, 58, 139, 132, 74, 55, 78, 86, 133, 37, 46, 179, 61, 113, 38, 114, 154, 146, 84, 124, 6, 4, 195, 22, 197, 20, 54

The ranking was carried out by the weight of topics that are listed in the file immediately after the cluster number (topics), for example, Cluster number 28 0.44

In this case, 28 is the cluster identifier, and 0.44 is its weight (degree of consistency with the entire corpus).

As can be seen by the first method, the topic No.167 has the greatest weight:

Cluster No. 167 1.0

<< FRIENDSHIP 0.28, WELFARE 0.25, WISH 0.24>>

beginning 0.11, impulse 0.1, patience 0.1, compassion 0.05, creation 0.08, hospitality 0.08, joy 0.2, becoming 0.17, unity 0.04, building 0.07, prosperity 0.23, unity 0.19, nation leader 0.03, manifestation 0.09, care 0.15, aspiration 0.05, wisdom 0.18, great 0.12, needy 0.0, finding 0.11, neighborliness 0.1, deepening 0.13, patriotism 0.16, friendship 0.28, wealth 0.15, calm 0.13, appreciation 0.14, significance 0.14, consolidation 0.13, alliance 0.07, wish 0.24, will 0.18, ethnicity 0.07, accomplishment 0.13, congratulation 0.19, relationship 0.17, telegram 0.09, shoulder 0.12, service 0.05, aspiration 0.23, anniversary 0.15, context 0.11, well-being 0.25, Abish 0.15, aspect 0.15, affluence 0.09, intelligence 0.06, pride 0.19, vision 0.13

The next most important No.135:

Cluster No. 135 0.91

<< WALK 0.27, CORNER 0.27, SIGHTSEEING 0.27 >>

beach 0.21, island 0.15, valley 0.14, paradise 0.13, pattern 0.09, contemplation 0.04, surfing 0.12, palm 0.13, nursain 0.17, shower 0.04, century 0.17, shopping 0.04, sharmelsheikh 0.02, pool 0.13, phuket 0.06, hainan 0.12, rheumatism 0.0 , dominican republic 0.04, gate 0.11, dome 0.04, landscape 0.21, heritage 0.1, white 0.16, darkness 0.13, pastime 0.09, gentle 0.09, rarity 0.15, pass 0.07, hike 0.16, Chinese 0.13, selection 0.17, oasis 0.12, ruin 0.13, hill 0.1, photographer 0.12, lane 0.13, plain 0.09, expanse 0.18, pebble 0.02, excursion 0.22, sun lounger 0.06, massage 0.06, vacation 0.11, waterfall 0.2, trail 0.08, mind 0.15, bungalow 0.07, abundance 0.19, diving 0.17, tower 0.15, continent 0.12, bali 0.1, landmark 0.27, location 0.14, era 0.16, neighborhood 0.13, leisure 0.11, guide 0.13, canyon 0.11, legend 0.2, miracle 0.16, shop 0.08, relief 0.11, mainland 0.09, foot 0.22, story 0.15, color 0.14, sunbed 0.12, sunset 0.17, town 0.12, water Park 0.04, uniqueness 0.15, ridge 0.14, vegetation 0.09, ruin 0.08, ancient 0.12, entertainment 0.23, romance 0.09, rock 0.16, tourist 0.04, caravan 0.08, width 0.14, souvenir 0.12, acquaintance 0.14, bay 0.08, great 0.15, wealth 0.17, sensation 0.18, combination 0.14, corner 0.27, hamlet 0.11, vietnam 0.08, vanity 0.13, arch 0.15, interior 0.13, hurghad 0.06, imagination 0.1, umbrella 0.12, grass 0.12, thailand 0.04, charm 0.11, walk 0.27, jungle 0.17, steppe 0.13, ladder 0.14, hut 0.05, spectacle 0.13, pleasure 0.18, variety 0.2, summer 0.18, shop 0.08, solitude 0.06, sand 0.19, hermitage 0.14, desert 0.14, lagoon 0.14,once 0.19, traveller 0.2, landscape 0.25, camel 0.12, riviera 0.05

And the least significant is the topic No.82:

Cluster No.82 0.15

<< TURAGELDI 0.43, SELFNOMINEE 0.42, MELDESHOV 0.37 >>

guldana 0.02, 0.37 meldeshov, selfnominee 0.42, musagali 0.26 selfnomination 0.21, zhaksybay 0.08, koichieva 0.29, foos 0.26, state language 0.16, oten 0.2, liman 0.18, centralelectioncommission 0.33, kaysarov 0.14, tokbayev 0.29, 0.2 zholdasbek,duambek 0.25, turageldi 0.43, mukasha 0.07,

cec 0.13, masino 0.19, eszhan 0.0, amantaykazh 0.34, amantayakazh 0.21, maslikhat 0.2, bazilbay 0.23, suleimena 0.09, salim 0.04, turgankul 0.21, turgankulov 0.2, myrzatay 0.24, turgun 0.19, ksdp 0.08, nominee 0.24, nomination 0.14

According to the second weighing method, the most significant is 202 topics:

Cluster No.202 1.0

<< IMPROVEMENT 1.0, BAYBEK 0.84, LIGHTING 0.75 >>

lighting 0.75, appearance 0.45, arrangement 0.24, movement 0.17, esim 0.29, bypass 0.4, building 0.35, small 0.16, convenience 0.23, baybek 0.84, documentation 0.29, inconvenience 0.18, design 0.24, sidewalk 0.64, issekesheva 0.46, improvement 1.0, location 0.16, builder 0.31, architect 0.72, facade 0.69

Next in importance 164:

<< DISCOUNT 1.0, SET 0.95, THING 0.91 >>

box 0.67, assortment 0.44, nuance 0.28, accent 0.39, pocket 0.76, corner 0.33, shoes 0.89, counter 0.37, thing 0.91, discount 1.0, second 0.57, analog 0.53, concept 0.56, walk 0.51, quarter 0.37, convenience 0.5, low 0.42, variety 0.31, summer 0.6, furniture 0.74, set 0.95, arrangement 0.47, preference 0.59, leisure 0.3, penny 0.23

And the least significant is the topic No. 54:

Cluster No.54 0.08

<< FISH 1.0, CAVIAR 0.49, FISHER 0.24 >>

string 0.0, squid 0.03, GOST 0.02, trawler 0.02, pollen 0.0, belly 0.02, fish 0.02, carp 0.04, shelf life 0.03, faeces 0.03, beluga 0.05, dildahmet 0.01, spinning 0.01, pump 0.03, bioresources 0.01, appendage 0.02, amazon 0.01, farm 0.18, catfish 0.02, halfaton 0.01, preservative 0.06, catch 0.07, wheel 0.01, piranha 0.05, sarys 0.02, fry 0.05, tasmania0.0, process 0.01, algae 0.02, inhabitant 0.06, epidemiology 0.0, larva 0.04, calf 0.01, livestock 0.01, spawning 0.03, ammonia 0.03, juvenile 0.03, otter 0.01, boatswain 0.01, bering 0.01, zaysan 0.02, fishing 0.08, fisherman 0.24, fisher 0.12 , fish 1.0, fishing 0.06, perch 0.03, piscatology 0.01, angler 0.02, fish 0.05, catch 0.04, fishery 0.05, trouble 0.01, balyk 0.02, fin 0.03, poacher 0.08, uba 0.02, hold 0.01, poaching 0.04, ichthyologist 0.01, guillman 0.01, kigach 0.0, tackle 0.01, caviar 0.49, technologist 0.01, jar 0.06, pituitary 0.06, pikeperch 0.03, coral 0.02, eel 0.02, suffocation 0.01, pike 0.04, genby 0.01, ural 0.06, atoll 0.01, individual 0.04, crucian 0.04, fishing rod 0.02, tray 0.02, trawl 0.01, sulfate 0.01, bream 0.03, cutting 0.02, shark 0.15 reproduction 0.03

At the beginning of each topic in triangular brackets <<...> > capital letters indicate its title, formed from the first three most significant words for this topic.

# APPENDIX G

Table G1 – Main characteristics of the vast corpus

|  |  |  |  |
| --- | --- | --- | --- |
| No. | Source | Period | A number of publications |
| 1 | [https://365info.kz](https://365info.kz/) | 05.2014-02.2019 | 57 326 |
| 2 | [https://rus.azattyk.org](https://rus.azattyk.org/) | 08.2003-02.2019 | 52 583 |
| 3 | [https://bnews.kz](https://bnews.kz/) | 04.2012-02.2019 | 7 631 |
| 4 | [https://www.caravan.kz](https://www.caravan.kz/) | 07.2001-02.2019 | 246 125 |
| 5 | [http://dailynews.kz](http://dailynews.kz/) | 02.2012-02.2019 | 25 409 |
| 6 | [https://esquire.kz](https://esquire.kz/) | 05.2013-02.2019 | 1 510 |
| 7 | [http://www.exclusive.kz](http://www.exclusive.kz/) | 10.2006-02.2019 | 7 184 |
| 8 | [https://kz.expert](https://kz.expert/) | 01.2017-12.2018 | 671 |
| 9 | [https://forbes.kz](https://forbes.kz/) | 12.2011-02.2019 | 159 755 |
| 10 | [https://www.inform.kz](https://www.inform.kz/) | 01.2010-02.2019 | 341 232 |
| 11 | [https://kapital.kz](https://kapital.kz/) | 09.2007-02.2019 | 73 468 |
| 12 | [https://www.kazpravda.kz](https://www.kazpravda.kz/) | 07.2013-02.2019 | 432 |
| 13 | [http://www.kp.kz](http://www.kp.kz/) | 12.2011-02.2019 | 12 422 |
| 14 | [https://kursiv.kz](https://kursiv.kz/) | 10.2015-02.2019 | 4 253 |
| 15 | [https://www.nur.kz](https://www.nur.kz/) | 08.2018-02.2019 | 18 003 |
| 16 | [https://ru.sputniknews.kz](https://ru.sputniknews.kz/) | 05.2016-01.2019 | 23 406 |
| 17 | [https://tengrinews.kz](https://tengrinews.kz/) | 01.2009-02.2019 | 198 797 |
| 18 | [https://time.kz](https://time.kz/) | 05.2016-02.2019 | 46 557 |
| 19 | [http://today.kz](http://today.kz/) | 12.2011-05.2017 | 102 456 |
| 20 | [http://vesti.kz](http://vesti.kz/) | 10.2008-08.2018 | 27 740 |
| 21 | [https://vlast.kz](https://vlast.kz/) | 04.2012-02.2019 | 21 051 |
| 22 | [https://www.zakon.kz](https://www.zakon.kz/) | Information is not available | 554 422 |
|  | Total |  | 1 982 433 |

Figure G1 – Balance of the main supplementary corpus composition by sources, %

1. Alexandru Tatar, Panayotis Antoniadis, Marcelo Dias de Amorim, and Serge Fdida, Ranking news articles based on popularity prediction [↑](#footnote-ref-1)
2. Roja Bandari, Sitaram Asur, Bernardo A., HubermanThe Pulse of News in Social Media: Forecasting Popularity [↑](#footnote-ref-2)