Ontological Resources for Representing Security Domain in Information-Analytical System*

Natalia Loukachevitch
Research Computing Center
Lomonosov Moscow State University
Moscow, Russia
louk_nat@mail.ru

Boris Dobrov
Research Computing Center
Lomonosov Moscow State University
Moscow, Russia
dobrov_bv@mail.ru

***Abstract*—The paper presents the approach to the description of the broad domain of national security as a thesaurus for automatic document processing. The created Security thesaurus has the representation model of the RuThes thesaurus.The Security thesaurus includes terminology related to social, national and religious conflicts, extremism and terrorism, information security. It is used in a specialized information-analytical system and for automatic text categorization according to several categorization schemes. The information-retrieval system provides several search instruments including word, phrase and concept search, category and facet search. It also supports the creation of analytical reports.**

***Keywords*—Security domain, thesaurus, text categorization, automatic document processing**

I. Introduction

National security is the important concept for describing the social phenomenon of the protection of the vital interests of the individual, society and the state against dangers and threats. The National Security Strategy of the Russian Federation ensures the implementation of constitutional rights and freedoms of citizens of the Russian Federation, decent quality and standard of living, sovereignty, independence, state and territorial integrity, sustainable social and economic development of the Russian Federation. National security includes the defense of the country and all types of security provided for by the Constitution of the Russian Federation and the legislation of the Russian Federation, primarily state, public, information, environmental, economic, transport, energy and personal security [22].

In accordance with this, the priorities and prospects of scientific and technological development include the counteraction to technogenic, biogenic, sociocultural threats, terrorism and ideological extremism, as well as cyber threats and other sources of danger for society, economy and the state.

One of important sources of information for forming counteraction against existing threats is the analysis of textual data including news reports, analytical papers, messages in social networks. In this paper we consider the approach to description of the broad domain of national security as a thesaurus for automatic document processing. The created Security thesaurus has the representation model of RuThes thesaurus [9]. We use the Security thesaurus in a specialized information-analytical system and for automatic text categorization of documents according to several categorization schemes, including Threats categories, Values categories, Regional problem categories and others. The information-retrieval system provides several search instruments including word, phrase and concept search, category and facet search. It also supports the creation of analytical reports.

*The work is supported by the Russian Foundation for Basic Research (project 16-29-09606) and the Ministry of Education and Science of the Russian Federation (project N 14.601.21.0018)

II. Related Work

In literature, several directions of natural language processing for national and international security issues have been studied.

Many works are devoted to the analysis of extremist messages in social networks. The part of such research is devoted to the activities of ISIL in Twitter. It is known that this organization actively works with a variety of social media [16], [20]. In particular, ISIL and associated organizations support a large number of Twitter accounts in order to spread their ideas in several languages. In the work [6], an analysis of the posts of people who joined or attempted to join ISIL is given. In most cases, they expressed strong anti-American and anti-Western attitudes long before joining ISIL.

The paper [2] describes a corpus containing 100 texts written by Islamists. In particular, fragments concerning war, non-believers, and punishments (64 texts) have been extracted from the collections of hadith (Islamic religious texts). Also, the corpus contains messages from Islamist blogs, as well as articles from the Islamist magazine Inspire. These texts have been labeled at several levels, including syntactic, time, and referential annotations.

Since one of the prerequisites of supporting terrorism is a sharply negative attitude toward certain phenomena or groups of people, so-called "hate" statements require special attention. In the work [15], blog posts are classified not only into classes

of expressed emotions (positive, negative, neutral), but also according to the type of action discussed (negative - *attack, attack, bloodshed, cruelty*, or positive - *help, support*). Kwok and Wang [7] describe the collected balanced set of 24.5 thousand tweets, which are classified into racist and normal ones. Waseem and Hovy [19] created a corpus of 16,000 tweets, in which 3.3 thousand tweets are labeled as sexist, and 1.9 thousand as racist.

In work [13], it is pointed out that the automatic detection of statements that incite hatred towards some groups of the population (which is one of the signs of extremism) is complicated by the following factors: such a message cannot always be detected by a simple set of keywords, because some words are intentionally distorted (to avoid detection); any fixed lists of offensive words constantly require adding; hate statements can be written in a completely fine literary language; offense and hatred may cross the boundaries of the sentence, when the object of the sentence is in another sentence; use of sarcasm. Schmidt and Wiegand [17] provide an overview of current approaches to detecting hate messages. It is pointed out that such approaches are based on the application of supervised machine learning methods based on several groups of features.

In the sphere of natural language processing for information security, Lim et al. [8] discuss the construction of a database for annotated malware texts. The annotation framework is based on the MAEC vocabulary for defining malware characteristics [5], along with a database consisting of annotated APT reports. The authors plan to use the database to construct models that can potentially help cybersecurity researchers in their data collection and analytics efforts. Gorokhov et al. [3] study application of convolutional neural network for anomaly detection in e-mail data.

Several projects are devoted to global monitoring of events occurring in the world to understand and respond to global problems, for example, International Crisis Early Warning System (ICEWS), maintained by Lockheed Martin, and Global Data on Events Language and Tone (GDELT) [14], [18]. Hand-coded and automated event data have been used to anticipate conflict escalation [14]. When combined with statistical and agent-based models, ICEWS claims a forecasting accuracy of 80%. GDELT has been used to track, e.g., wildlife crime and the rise of hate speech following the U.K. Brexit vote [18].

III. RuThes Model of Representation

We use the model of the RuThes thesaurus for knowledge representation in the security domain. The structure of RuThes is based on three traditions of developing computer resources for document procssing: information-retrieval thesauri, WordNet-like thesauri, and formal ontologies [11]. The RuThes thesaurus is created in form of a linguistic ontology, which concepts are based on senses of really existing words and phrases. There exist several large Russian thesauri presented in the same format:

- RuThes thesaurus comprising words and phrases of literary Russian together with terms of so-called sociopolitical domain (see below) [9];
- RuThes-lite, a published version of RuThes[1], can be obtained for non-commercial purposes [10];
- Sociopolitical Thesaurus comprising lexical items and terms from the sociopolitical domain. The sociopolitical domain is a broad domain describing everyday life of modern society and uniting many professionals domains, such as politics, law, economy, international relations, finances, military affairs, arts and others. Terms of this domain are usually known not only professional, but also ordinary people [11]. The Sociopolitical thesaurus can exist and be used separately. At the same time it is included as a part into three larger thesauri: RuThes, OENT ontology, and the Security Thesaurus;
- Ontology on Natural Sciences and Technology (OENT) includes terms of mathematics, physics, chemistry, geology, astronomy etc., terms of technological domains (oil and gas, power stations, cosmic technologies, aircrafts, etc.). It also contains the Sociopolitical thesaurus as a part because scientific and technological problems can be discussed together with political, economical, industrial and other issues [1];
- Security Thesaurus is an extension of the RuThes thesaurus and includes terminology related to social, national and religious conflicts, extremism and terrorism, information security.

The Table 1 contains quantitative characteristics of the abovementioned resources.

Table I
RuThes-like Thesauri

Thesaurus	Number of concepts	Number of Text Entries	Number of Conc. Relations
RuThes	55,275	170,130	226,743
RuThes-lite	31,540	111,559	128,866
Sociopolitical Thesaurus	41,426	121,292	161,523
OENT	94,103	262,955	376,223
Security Thesaurus	66,810	236,321	271,297

The Security thesaurus is created in form of a linguistic ontology, which concepts are based on senses of really existing words and phrases. Each concept has a unique name and is associated with a set of text entries, the senses of which correspond to the concept. Text entries of a specific concept can comprise single words of different parts of speech, including ambiguous ones, and phrases that can be either idiomatic or compositional groups. Large rows of synonyms and term variants are collected to provide better recognition of concepts in texts.

Fig. 1-3 show the interface of thesaurus developing. The upper left form contains list of concepts, the lower left form shows text entries for the highlighted concept. The right upper

[1] http://www.labinform.ru/pub/ruthes/index.htm

form presents the relations of the highlighted concept, and the lower right form shows text entries for a related concept. It can be seen (Fig. 1-3) that concepts are provided with numerous text variants extracted from real texts, for example, *import dependence* concept (Fig.1) can be expressed also as *dependence on import* or *dependence on import goods*. Fig. 3 shows different text variants to express *zero-day vulnerability* and *vulnerability attack* concepts.

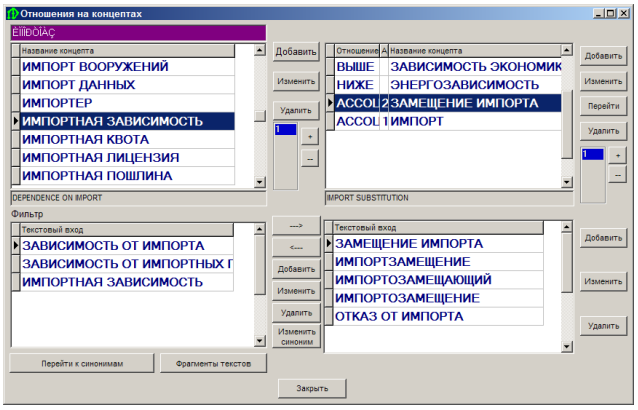

Figure 1. Example of economic threat description.

There are four basic types of relationships between concepts. The first type of relation is the class-subclass relationship, has the properties of transitivity and inheritance.

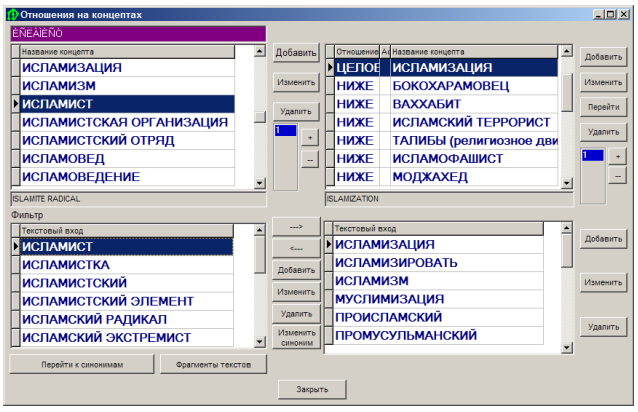

Figure 2. Example of state threat description.

The second type of relations is the part-whole relation. It is used not only to describe the physical parts, but also to other internal attributes, such as properties or roles for situations. An important condition for establishing this relationship is that the concept-parts must be rigidly connected with their whole, that is, each example of the concept-part must, throughout its entire existence, be part of the concept of the whole, and not relate to anything else. Under these conditions, it is possible to rely on the transitivity property of the part-whole relation, which is very important for automatic logical inference in the process of automatic text processing [11].

For example, concept *halal* (sanctioned in Islam) is described as a part of concept *Shariat*, and concept *Shariat* is presented as a part *Islam* concept. *Islamist radical* is presented as a part of *Islamization* (Fig. 2).

The third type of relations, called the asymmetric association $asc_1 - asc_2$, connects two concepts that cannot be related by the relationships discussed above, but when one of which does not exist without the existence of another, for example, the *import dependence* concept can exist if only the *import* concept exists (Fig.1). The *vulnerability attack* concept can appear if only the *computer vulnerability* concept exists (Fig.3)

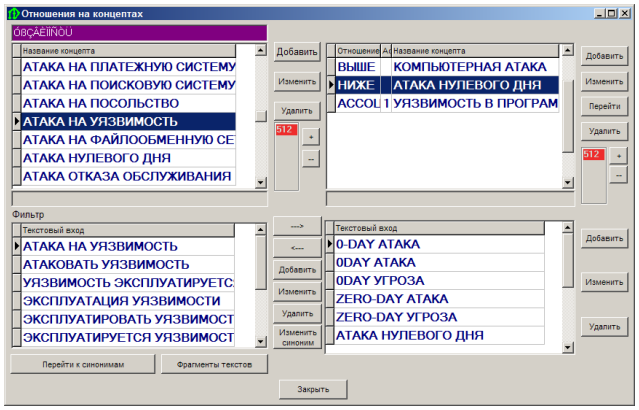

Figure 3. Example of information threat description.

The last type of relationships is the symmetrical association links concepts that are very similar in meaning, but which seems difficult to combine in one concept.

Thus, the system of the thesaurus relations describes the most significant relationship of concepts. Fig 4. presents the hierarchy of concepts under the *threat* concept. There can be seen such threats as military confrontation, cyber threats, radiation threats (nuclear war, radioactive contamination, nuclear terrorism), terroristic threat, etc.

IV. Thesaurus-Based Text Categorization

The mainstream technology in automatic text categorization is the machine learning approach. This approach presupposes that consistent training data of sufficient volume is available for training algorithms. However, in a new complicated text categorization task, even a system of categories can be absent and should be created.

In such conditions, knowledge-based methods of automatic text categorization based on manual rules of category assignment are more appropriate. Working in a broad domain, it is necessary to use thesaurus support in rule description because the thesaurus allows for operating not with single words and expressions but with concepts and thesaurus substructures [11] to describe inference of a specific category from a text.

In the security domain, the following categorization schemes have been created:

- Threat categories, describing existing national security threats (188 categories, 5-level hierarchy);

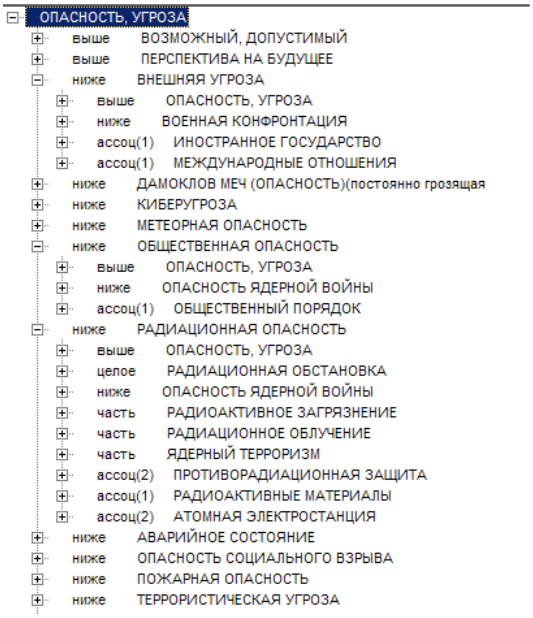

Figure 4. Hierarchy of the threat concept.

- Values categories presenting individual, social, and state values, for example, freedom, democracy, human rights, family values, etc. (109 categories, 4-level hierarchy);
- Categories of ethno-confessional relations (94 categories, 4-level hierarchy);
- Regional problems categories (84 categories, 2-level hierarchy);
- Region facets (325 facets: subjects of the Russian Federation and foreign states).

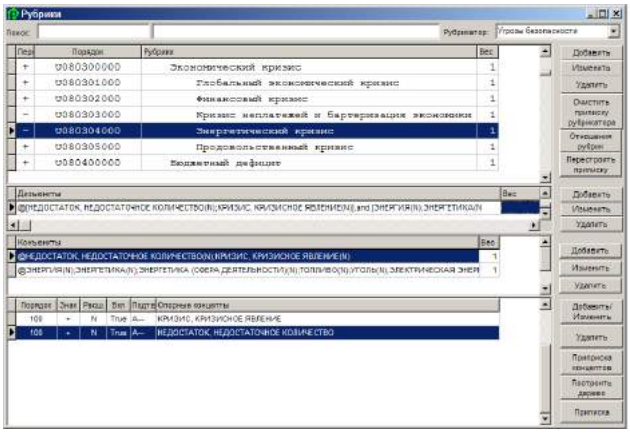

Figure 5. Example of description of a threat for text categorization system.

A. Using Thesaurus to Describe the Category Contents

In our case, for each category, a Boolean expression over thesaurus concepts is created. The syntax of the expression allows using the thesaurus hierarchies to provide appropriate coverage of the description. Here the transitivity of class-subclass and part-whole relations is significantly exploited.

In a special interface, an expert can edit the description of a category, delete extra concepts, add new concepts, or change their expansion scope. The expansion scope determines what related concepts to an initial concept should be used in the category profile:

- no expansion, only concept text entries are used (N);
- all lower concepts, including subclasses, parts and dependent concepts (Y) obtained from the concept hierarchy;
- only subclasses from the concept hierarchy (L);
- all directly related lower concepts (W);
- all directly related lower concepts without subclasses (V).

Fig. 5 presents the description of the category "Energy crisis" from the Threats subject heading scheme. This is conjunction of two disjunctions. The first one contains crisis-related concepts, including *crisis, lack, deficiency*. The second one comprises energy-related concepts.

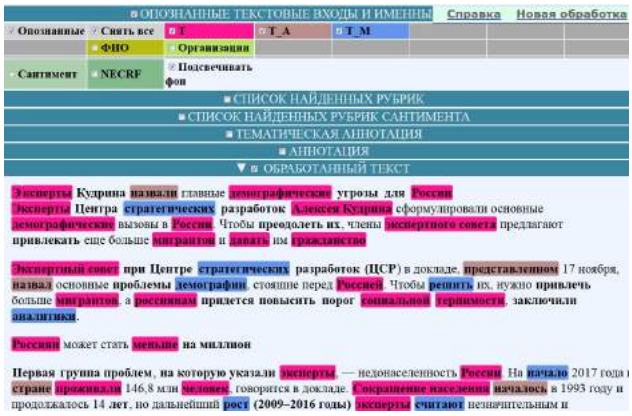

Figure 6. Thesaurus terms found in a text. Brown and blue boxes show ambiguous terms, which should be disambiguated

B. Categorization of Texts according to Subject Headings

The main stages of thesaurus-based text categorization include:

- Tokenization and lemmatization; that is, the transfer of word forms to lemmas (dictionary word forms);
- Matching with the thesaurus based on the lemma representation of the document. Fig. 6 shows the term coverage of news text " Kudrin's experts named the main demographic threats for Russia"[2]);
- Disambiguation of ambiguous text entries. Brown and blue boxes on Fig. 6 highlight ambiguous terms. For example, Russian word *demografiya (demography)* can mean *demographic situation* or *demographic science*;
- Grouping semantically related concepts into so called thematic nodes. This provides better determination of concept weights, which calculated on the basis of the concept frequency in the given document and the significance of the corresponding thematic node. Fig. 7 (in the

[2]https://www.rbc.ru/economics/17/11/2017/5a0eb1d39a79470f724250b4

center) demonstrate such thematic nodes for the above-mentioned document about the demographic threats. The important thematic node about the demographic situation (*demographic situation, life expectancy, natural population decline, age structure of the population, population aging*... etc.) ;

- Forming the conceptual index of the document. The conceptual index consists of concepts found in the document and their assigned weights. The weight of a concept accounts for the significance of the corresponding thematic node and the frequency of the concept in the document. In the example text, the important threat "population aging" was explicitly mentioned only once in the text, and it could obtain a too low frequency-based weight, but with the support of the main thematic node "demographic situation", its weight is considerably higher;
- Calculation of category weights in dependence of concepts included into the rules of the inference for this category. Fig. 6 (upper part) shows the categories found in the mentioned document including "Depopulation", "Population aging", "Fertility decline";
- The results of document processing, including the word index, the conceptual index, the calculated categories, etc. are loaded into an information-analytical system.

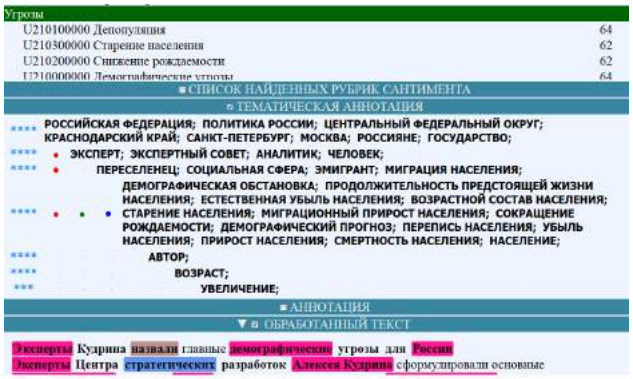

Figure 7. The upper part shows the threat categories found for a text. The central part presents thematic nodes of related concepts such as "demographic situation" node.

V. Document Searching and Categorization Schemes in Information-Analytical System

As a result of single document processing, a large number of entities of different nature (terms, named entities, sentiment expressions, etc) are extracted. The found entities are attached to word positions. The extracted entities, their weights and text positions are loaded into the search indexes of the information-analytical system.

Two types of search indexes are used: inverse and direct indexes, which are stored in special noSQL databases. The inverse index stores information organized from vocabulary elements (E_i) to documents and used for document searching:

$$[E_i[d_j, rank(E_i, d_j), [positions_{ijk}]]] \quad (1)$$

.

The direct index stores the same information, but is organized according to the documents:

$$[d_j[E_i, rank(E_i, d_j), [positions_{ijk}]]] \quad (2)$$

The direct index is used, for example, to highlight the relevant text fragments, which makes it possible to speed up the process of selecting the documents needed by a user. Also, the direct index is the main support tool for advanced search engine analysis instruments, when, for the initial query, a group of documents is received and it is required to select the most significant elements in this sample.

NoSQL database NearIdx 9.0 is used as a search engine, developed by LLC "Laboratory for Information Research" in cooperation with the Research Computing Center of Lomonosov Moscow State University. The search time for simple queries on the inverse index for a collection of 10 million documents on standard modern computers is 1-2 seconds.

The documents can be searched according to different search criteria. Also analytical reports can be formed. An analytical report usually contains document fragments related to the given topic with additional subcategorization. Similar fragments are clustered. The screen interface allows for editing, removing etc. fragments to obtain a more meaningful report. Fig. 9 presents the analytical report on "Demographic threat" category with additional regional subcategorization. It can be seen that the document fragments do not contain words *demography* and *threat*.

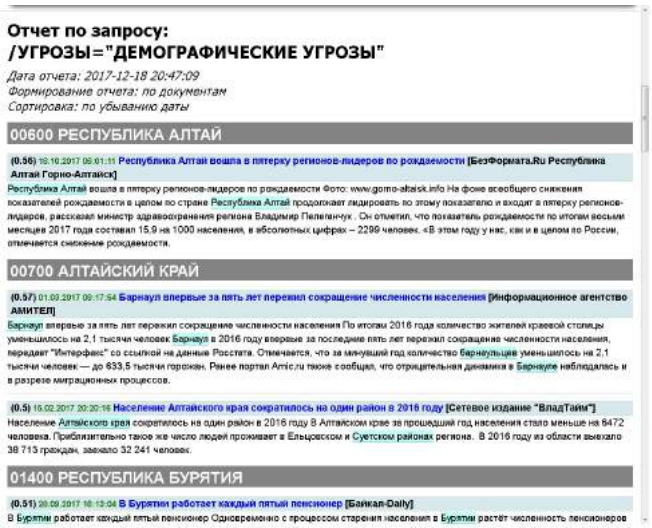

Figure 8. Analytical report created for "Demographic Threat" category

VI. Conclusion

In this paper we considered the approach to description of the broad domain of national security as a thesarus for automatic document processing. The created Security thesaurus includes terminology related to social, national and religious conflicts, extremism and terrorism, information security. The

Security thesaurus has the representation model of RuThes thesaurus. Its structure is based on three traditions of developing computer resources for document processing: information-retrieval thesauri, WordNet-like thesauri, and formal ontologies. We continue the development of the Security thesaurus.

We use the Security thesaurus in a specialized information-analytical system and for automatic text categorization of documents according to several categorization schemes, including Threats categories, Values categories, and Regional problem categories. The information-retrieval system provides several search instruments including word, phrase and concept search, category and facet search, useful for analytics. It also supports the creation of analytical reports.

References

[1] B.Dobrov and N. Loukachevitch, "Development of Linguistic Ontology on Natural Sciences and Technology", in Proc. of Linguistic Resources and Evaluation Conference LREC-2006, 2006.

[2] M. Finlayson, J. Halverson, and S. Corman, "The N2 corpus: A semantically annotated collection of Islamist extremist stories", in Proc. of Linguistic Resources and Evaluation Conference LREC-2014, pp. 896-902.

[3] O. Gorokhov O., M. Petrovskiy and I. Mashechkin, "Convolutional Neural Networks for Unsupervised Anomaly Detection in Text Data", in Int. Conf. on Intelligent Data Engineering and Automated Learning, Springer, Cham, 2017. pp. 500-507.

[4] R. Killam, P. Cook and N. Stakhanova, "Android Malware Classification through Analysis of String Literals", in First Workshop on Text Analytics for Cybersecurity and Online Safety TA-COS, 2016.

[5] I. Kirillov, D. Beck, P. Chase, R. Martin, " Malware Attribute Enumeration and Characterization", The MITRE Corporation, Tech. Rep., 2010

[6] E. Kohlmann, L. Alkhouri L,"Profiles of foreign fighters in syria and iraq", CTC Sentinel, September, 2014.

[7] I. Kwok, and Y. Wang, "Locate the hate: Detecting tweets against blacks", in Proc. of AAAI-2013, 2013.

[8] S. Lim, K. Swee, A.Muis, W., Lu, and C. Ong, "MalwareTextDB: A Database for Annotated Malware Articles", in Proc. of the 55th Annual Meeting of the Association for Computational Linguistics, Vol. 1, 2017.

[9] N. Loukachevitch, and B. Dobrov, “RuThes linguistic ontology vs. Russian wordnets“, in Proc. of Global WordNet Conf. GWC-2014, 2014.

[10] N. Loukachevitch, B. Dobrov and I. Chetviorkin, ’ RuThes-Lite, a publicly available version of Thesaurus of Russian language RuThes", Computational Linguistics and Intellectual Technologies: Papers from the Annual Intern.l Conf. “Dialogue-2014”, 2014, pp. 340-349.

[11] N. Loukachevitch and B. Dobrov, "The Sociopolitical Thesaurus as a resource for automatic document processing in Russian", Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication 21.2, 2015, pp. 237-262.

[12] M.Saleem, KP. Dillon, S. Benesch, D. Ruths, "A Web of Hate: Tackling Hateful Speech in Online Social Spaces", in Proc. of 1st Workshop on Text Analytics for Cybersecurity and Online Safety TA-COS 2016, 2016.

[13] Ch. Nobata, J. Tetreault, A. Thomas, M. Yashar, and Yi Chang, "Abusive language detection in online user content", in Proc. of the 25th International Conference on World Wide Web, 2016, pp. 145–153.

[14] S. O’brien, "Crisis early warning and decision support: Contemporary approaches and thoughts on future research", International studies review, 12(1), 2010, pp. 87-104.

[15] S. Sela, T. Kuflik, G. Mesch. "Changes in the discourse of online hate blogs: The effect of Barack Obama’s election in 2008", First Monday, V. 17, N. 11, 2012.

[16] S. Shane, and B.Hubbard, "Isis displaying a deft command of varied media", New York Times 31.

[17] A. Schmidt, and M. Wiegand. 2017. "A survey on hate speech detection using natural language processing",in Proc. of SocialNLP 2017, 2017.

[18] W. Wang, R. Kennedy, D. Lazer, and N. Ramakrishnan, "Growing pains for global monitoring of societal events", Science, 353(6307), 2016, pp. 1502-1503.

[19] W. Zeerak and D. Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter". in Proc. of NAACL-HLT, 2016, pp. 88–93.

[20] G.Weimann, New Terrorism and New Media,Wilson Center Common Labs, 2014.

[21] J. Qiao, H. Jie, Y. Yuexiang, and J. Lin, "Analyzing Malware by Abstracting the Frequent Itemsets in API Call Sequencesm", in 12th IEEE Intern. Conf. on Trust, Security and Privacy in Computing and Communications. IEEE, 2013, pp. 265–270. https://doi.org/10.1109/TrustCom.2013.36.

[22] The National Security Strategy of the Russian Federation, approved by decree by the President of Russian Federation N 683, 31st December of 2015.

ПРЕДСТАВЛЕНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ "БЕЗОПАСНОСТЬ"В ФОРМЕ ТЕЗАУРУСА ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ДОКУМЕНТОВ

Лукашевич Н.В., Добров Б.В.

НИВЦ МГУ имени М.В. Ломоносова

Национальная безопасность является важной концепцией для описания социального феномена защиты жизненных интересов личности, общества и государства от опасностей и угроз. Стратегия национальной безопасности Российской Федерации определяет концепцию «национальной безопасности» как состояние защиты личности, общества и государства от внутренних и внешних угроз, обеспечивающих реализацию конституционных прав и свобод граждан Российской Федерации, достойного качества и уровня жизни, суверенитета, независимости, государственной и территориальной целостности, устойчивого социально-экономического развития Российской Федерации. Национальная безопасность включает защиту страны, а также обеспечение государственной, общественной, информационной, экологической, экономической, транспортной, энергетической и личной безопасности.

Один из важных источников информации для формирования противодействия существующим угрозам - анализ текстовых данных, включая новостные сообщения, аналитические документы, сообщения в социальных сетях. В этой статье мы рассматриваем подход к описанию широкой области национальной безопасности как тезаурус для автоматической обработки документов. Созданный Тезаурус по безопасности имеет модель представления тезауруса РуТез, фактически это расширение тезауруса РуТез и включает терминологию, связанную с социальными, национальными и религиозными конфликтами, экстремизмом и терроризмом, информационной безопасностью.

Мы используем Тезаурус по безопасности в специализированной информационно-аналитической системе и для автоматической текстовой классификации документов в соответствии с несколькими рубрикаторами, включая рубрикаторы угроз, ценностей и региональных проблем. Информационно-поисковая система имеет несколько инструментов поиска, включая поиск по словам, фразам и понятиям, поиск категорий и фасетов. Она также поддерживает создание аналитических отчетов.