

IRSTI 81.93.29

## Creating the dataset of keywords for detecting an extremist orientation in web-resources in the Kazakh language

Bolatbek M.A., al-Farabi Kazakh National University,  
Almaty, Kazakhstan, +77056664007, e-mail: bolatbek.milana@gmail.com  
Mussiraliyeva Sh.Zh., al-Farabi Kazakh National University,  
Almaty, Kazakhstan, +77059011283, e-mail: mussiraliyevash@gmail.com  
Tukeyev U.A., al-Farabi Kazakh National University,  
Almaty, Kazakhstan, +77017106351, e-mail: ualsher.tukeyev@gmail.com

This paper is part of the research on creating semantic analysis models in web resources for defining an extremist orientation in the text. To solve this task a model was created, which consists of five stages: identifying websites of an extremist groups, preparing for data extraction, data extraction, data analysis and classification. This work presents the results of data analysis stage of above mentioned model. The purpose of this study is to identify keywords, often used by extremists, which will later be used to classify texts to "extremist" and "neutral" categories using machine learning methods. There is no such database for the Kazakh language. As the result of this study an experimental corpus and list of keywords in Kazakh language was created. The keywords were added to the database with various morphological variants. The program was built that checks for the presence of extremist keywords in the given input text and displays the words found.

**Key words:** extremist texts, term frequency, text classification, emotional scores.

### Создание базы данных ключевых слов, для определения экстремистской направленности в веб-контенте на казахском языке

Болатбек М.А., Казахский национальный университет имени аль-Фараби,  
г. Алматы, Республика Казахстан, +77056664007, e-mail: bolatbek.milana@gmail.com  
Мусиралиева Ш.Ж., Казахский национальный университет имени аль-Фараби,  
г. Алматы, Республика Казахстан, +77059011283, e-mail: mussiraliyevash@gmail.com  
Тукеев У.А., Казахский национальный университет имени аль-Фараби,  
г. Алматы, Республика Казахстан +77017106351, e-mail: ualsher.tukeyev@gmail.com

Данная работа является частью исследования создания моделей семантического анализа для определения экстремистской направленности в тексте. Для решения данной задачи была построена модель, которая состоит из пяти этапов: определение веб-сайтов экстремистских групп, подготовка к извлечению данных, извлечение данных, анализ данных и классификация. Данная работа представляет результаты этапа анализа данных указанной модели. Целью исследования является определение ключевых слов, часто используемых экстремистами, которые в дальнейшем будут использоваться для классификации текстов на «экстремистские» и «нейтральные» категорий с использованием методов машинного обучения. Для казахского языка не существует такой базы данных. В результате этого исследования был создан экспериментальный корпус и список ключевых слов на казахском языке. Ключевые слова были добавлены в базу данных с различными морфологическими вариантами. Была разработана программа, которая проверяет входной текст на наличие экстремистских ключевых слов и возвращает найденные слова.

**Ключевые слова:** экстремистские тексты, частота терминов, классификация текста, эмоциональные оценки

### Қазақ тілді веб-контенттегі экстремистік бағытты анықтауға арналған түйінсөздер мәліметтер қорын құру

Болатбек М.А., әл-Фараби атындағы Қазақ ұлттық университеті,  
Алматы қ., Қазақстан Республикасы, +77056664007, e-mail: bolatbek.milana@gmail.com

Мусылралиева Ш.Ж., әл-Фараби атындағы Қазақ ұлттық университеті,  
Алматы қ., Қазақстан Республикасы, +77059011283, e-mail: mussiraliyevash@gmail.com  
Тукеев У.А., әл-Фараби атындағы Қазақ ұлттық университеті,  
Алматы қ., Қазақстан Республикасы, +77017106351, e-mail: ualsher.tukeyev@gmail.com

Бұл жұмыс мәтіндегі экстремистік бағытты анықтау үшін веб-ресурстарды семантикалық талдау үлгілерін құру зерттеуінің бөлімі болып табылады. Аталған есепті шешу үшін бес кезеңнен тұратын үлгі құрылды: экстремистік топтардың веб-сайттарын анықтау, мәліметтерді алуға дайындық жүргізу, мәліметтерді алу, мәліметтерді талдау және жіктеу. Берілген жұмыс жоғарыда аталған үлгінің мәліметтерді талдау кезеңінің нәтижелерін көрсетеді. Жұмыстың мақсаты экстремистер жиі қолданатын және келесі кезеңдерде мәтінді “экстремистік” және “бейтарап” санаттарға машиналық оқыту әдістері көмегімен жіктеуде пайдаланылатын түйінсөздерді анықтау болып табылады. Қазақ тілі үшін мұндай сөздік жоқ. Зерттеу нәтижесі ретінде қазақ тіліндегі эксперименталдық корпус пен мәліметтер қоры құрылды. Түйінсөздер мәліметтер қорына бірнеше морфологиялық нұсқаларымен бірге енгізілді. Кіріс мәтінді экстремистік түйінсөздердің болуына тексеретін және табылған сөздерді қайтаратын бағдарлама құрылды.

**Түйін сөздер:** экстремистік мәтіндер, термин жиілігі, мәтінді жіктеу, эмоциялық ұпайлар

## 1 Introduction

The rapid development of the Internet and information technologies poses new challenges in the field of national security, as in the last decade terrorist organizations are oriented on the Internet space, where today there are up to 10000 extremist electronic platforms (Zhavoronkova, 2015: 30). Extremist organizations use social networks, blogs, forums, etc. for propaganda, attraction of new members, conducting ideological works, collecting money for the implementation of terrorist acts. Unfortunately, in recent years, Kazakhstanis have also joined the ranks of extremist organizations. According to the Foreign Ministry’s reports, to October 2017, about 500 Kazakhstanis left the country and joined the ISIS (Islamic State of Iraq and Syria), this number includes both the militants themselves and their wives and children (Information Portal of Kazakhstan, 2017).

In this regard, it becomes urgent to automatically monitor Internet resources in order to identify text messages of an extremist orientation. This problem can be presented in the form of a binary classification problem in which the texts of messages in social networks, blogs and other resources will play the role of the analyzed objects and solved with the help of machine learning methods (Ananyeva, 2016: 210). Such an approach requires the presence of a labeled corpus of texts (Cohen, 2014: 246-253) (Finlayson, 2014: 896–902) and a predefined set of analyzed characteristics, such as the results of a complete linguistic analysis, list of keywords, etc.

There is no database of extremist keywords for the Kazakh language. This paper is part of the research on creating semantic analysis models in web resources for defining an extremist orientation in the text. To solve this task a model was created, which consists 5 stages: identifying the websites of extremist groups, preparing to data extraction. data extraction, data analysis and data classifying. This work presents the results of data analysis stage of above mentioned model. The purpose of this study is to identify keywords, often used by extremists, which will later be used to classify texts to “extremist” and “neutral” categories using machine learning methods.

## 2 Literature review

Researchers proposed many methods for solving the problem of determining extremist orientation in the texts. Some of mentioned works are listed below.

The research (Scanlon, 2014: 10–25) presents methods for identifying and forecasting the recruitment activities of violent groups within extremist social media websites. Authors used naive Bayes models, logistic regression, classification trees, boosting, and support vector machines (SVM) to classify the forum posts in a 10-fold cross-validation experimental setup. They used data from the western jihadist website Ansar ALJihad Network, where authors employed a bag-of-words feature space by parsing each forum post in the corpus into a term-by-document matrix. This matrix of term frequency (tf) features was created using the RTextTools and tm text mining packages. The number of features was further reduced through stemming using the Porter Stemming Algorithm.

In the work (Enghin, 2015: 17–33) authors made an attempt to automatically detect radical content on Twitter. They used a machine learning approach that classifies a tweet as radical or non-radical. Features are based on the polarity of words, which is determined by using several dictionaries like Dictionary of Affect in Language (DAL) or WordNet which assigns each word a pleasantness score between 1 (negative) and 3 (positive). The experiments were conducted using a tool called Weka which is a suite of machine learning software written in Java. For experiments they used three different classifiers: Support Vector Machine (SVM), Naive Bayes and AdaBoost.

In the work (Johansson, 2016: 374–390) authors discussed the possibility of detecting violent extremism by identifying signs of warning behaviours in written text, such as "leakage", "fixation" and "identification". To detect linguistic markers signalling a leakage, it is proposed that predefined word lists of violent actions are used, and to extend such a predefined list of words using lexical databases such as WordNet. For linguistic markers of fixation, they proposed to simply count the relative frequency of key terms relating to named entities such as persons, organisations, etc. The proposed approach is based on simple lists of keywords, where a keyword can consist not only of single words, but also of multi-word units (e.g., the bigram "Al-Shabab"). The monitoring tool performs a keyword search in the incoming data for occurrences of the keywords included in the markers that already have been described. For each marker, the monitoring tool outputs a list of Uniform Resource Identifiers (URIs) and the frequency of occurrence of the marker in the document. These lists of URIs are then input to an analysis script where it is possible to define how many or which markers that have to be triggered in order for an URI to be shown as potentially interesting to the user of the system.

The work (Azizan, 2017: 691–698) presents the approach to sense user's act leading to terrorism based on the tweets they shared at the Tweeter platform. In this research the data will be collected through Twitter streaming API. The data extracted based on the user criteria, for instance by matching a keyword "terrorist". In the next stage data gathered is cleaned by filtering and removing all tags, hashtags, spelling errors, non-english words. Next, data is tagged with the input and output labels. Then the data is classified into sentiment polarity which are positive, negative and neutral class by using machine learning methods. In mapping sentiwordnet is used. It consists of english words which have been attributed to a positive or negative score. Tweet sentence is compared and calculated the score by referring

to sentiwordnet dictionary. In sentiment classification sentence is categorized into the classes of positive, negative and neutral using Naive bayes algorithm.

In the study (Devyatkin, 2017:188–190) authors analyzed two basic groups of features that distinguish extremist texts, which are lexical features, and psycholinguistic and semantic features. Authors created a text corpus for the research. The corpus includes 493 manually collected texts (650 000 words), 368 of them are extremist texts. In experiments they used three approaches for lexical feature representation: bag of words and collocations, frequency dictionary and word embedding. For creation of word embeddings the fastText was used. Psycholinguistic and semantic features were extracted from Russian text corpus. The values of psycholinguistic features are estimated on the basis of morphology of lexical units of the analyzed texts. Values of semantic features are calculated as frequencies of semantic roles in the corpus. Authors divided dataset into two parts: texts represented using by "bag of words and collocations" model, and texts represented using word embedding, both of them included extremist and neutral texts. Authors trained and tested classifiers using these parts of the dataset. Classification quality was estimated by calculating F1-score during 5-fold cross-validation. For classification authors applied multinomial naive Bayes, logistic regression, linear SVM, random forest, gradient boosting. For this experiment authors used open source library of machine learning methods - Scikit-learn - where all above methods are realized.

The purpose of the study (Targeir, 2013: 1–27) is to explore which words and expressions are typical for extremist forums. Authors analysed vocabularies of forums in English, Norwegian, and German. In this thesis, authors found frequent and characteristic words by means of Global Term Frequency (GTF) and pairs of co-occurring words by means of odds ratio in different extremist forums. They compared normalized GTF (NGTF) of words in two forums to find out where they are used most. Words used in only one of two forums are found as well. They found the GTFs for words written by five of the ten most active authors in each forum, and they found words that one author writes, while the other of ten most active authors does not write.

Authors of the paper (Chen, 2012: 3–105) have developed various multilingual data mining, text mining, and web mining techniques to perform link analysis, content analysis, web metrics analysis, sentiment analysis, authorship analysis, and video analysis in our research. Dark Web Forum Portal system contains three components: data acquisition, data preparation, and system functionality. At the first stage spidering programs are developed to collect the web pages from online forums that contain jihadist-related content identified by domain experts. At the Data Preparation stage forum parsing programs are developed to extract the detailed forum data from the raw HTML web pages and store it in a local database. The Dark Web Forum Portal is implemented using Apache Tomcat, and the database is implemented using Microsoft SQL Server 2008. For forum statistics analysis, Java applet-based charts are created to show the trends based on the numbers of messages produced over time.

### 3 Materials and methods

To solve the problem of detecting extremist orientation in texts a model was constructed, which consists five stages: identifying websites of an extremist groups, preparing for data extraction, data extraction, data analysis and classification. According to this model, the first stage determines the web resources which people often use to exchange extremist messages

using search engines by terrorist vocabulary (group names, leader names, special keywords, etc.). In our case they are social networks like Youtube, vk, blogs, forums and news sites. The second stage is preparing for the extraction of data (registration on forums, select suitable proxy servers). As noted above, there is no database of extremist keywords in the Kazakh language. For this reason, this research firstly needs the corpus of texts written in Kazakh. In order to create the corpus, comments and messages from above mentioned web-resources were downloaded. The words, which in foreign languages, have been translated using the Google Translator (Google Online Translator, 2018). Data downloaded from web resources were collected to a text file. At the moment, there are 150 texts in this file, 80 of them are texts with extremist orientation, the remaining texts refer to the "neutral" category, which contain a comment condemning extremism and news texts.

To determine most frequent words researchers use different methods. For example, the method of Part of Speech Tagging is used in (Scrivens, 2016: 104-107), this method highlights frequently occurring words and divides the text into groups by parts of speech, for example, nouns, verbs, adverbs, etc. This method is convenient to use when researchers focus on certain parts of speech, for example, in the above work, researchers only consider nouns, because according to the results of the study, extremist keywords in English are in most cases are nouns. In our case, this method is not suitable, since visual inspection showed that most keywords in the Kazakh language are divided by parts of speech. For this reason, in this research the TF-IDF method was used. This method is used to evaluate the importance of the word. For finding words that are typical to a document, which in this case corresponds to a forum message, one finds the term frequency (TF) of a term (in the entire document inverse document frequency (IDF). The product of TF and IDF we call TF-IDF (Targeir, 2013: 23).

The term frequency  $tf(t, d)$ , the simplest choice is to use the raw count of a term in a document, i.e. the number of times that term  $t$  occurs in document  $d$ .

The inverse document frequency is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is the logarithmically scaled inverse fraction of the documents that contain the word, obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient. IDF defines using (1) (Wikipedia, 2017).

$$idf(t, D) = \log \frac{|D|}{|\{d_i \in D | t \in d_i\}|} \quad (1)$$

where  $|D|$  – total number of documents in the corpus;

$|\{d_i \in D | t \in d_i\}|$  – the number of documents where the term  $t$  appears.

Then TF-IDF is calculated as:

$$tf-idf(t, d, D) = tf(t, d) * idf(t, D) \quad (2)$$

Using (2) the values of TF-IDF of keywords were calculated. Results are given in the decreasing order in Table 2. In the stage of analyzing most frequent words, the following interesting facts were revealed:

1. Kazakh letters are often replaced by Cyrillic, for example, "sogys" (соғыс) instead of "sog'ys" (соғыс), "tozak" (тозақ) instead of "tozaq" (тозақ);

2. Using the terms in Arabic written by Cyrillic, for example, kafir (non-believer), muzhahid (participant in jihad), martyr (Muslim who died for religion), hijra (migration from one locality to another);

3. Frequent use of bigrams, for example, hijra қулу (хижра қылу), fard kefair (фард кефайр), daulatul Islam (даулатуль ислам), jihad jasau (жихад жасау), etc .;

4. One word can be written in several variants, for example, jihad (жихад), jihad (жихад), jihat (жихат), djihat (джихат).

Once the keywords have been defined, their basics have been written into the SQLite Expert Personal 3.5.46.2466 database (SQLite Expert Database, 2018), we are interested only in the basics not in the endings, because basics reduce the efficiency of the program (reduces the time for searching for possible word variants), and we will consider different versions of the word with different endings as one word, for example, the words "jihad" (жихад), "jihadtyn'" (жихадтың), "jihadqa" (жихадқа), "jihadtan" (жихадтан), "jihadta" (жихадта) will be considered as the same word. As mentioned above, one word may have several variants of writing, so words were entered into the database with all possible variants. Possible morphological variants of words were determined by studying the content of web forums, blogs.

Next, a program was built that checks for the presence of extremist keywords in the entered text and displays the words found. The program was developed in the integrated development environment Visual C# (Integrated development environment Visual C, 2017). At the first stage, morphological analysis is performed to incoming text, where for each word morphological labels such as base and ending are determined. Since we are no longer interested in endings, we will only consider the basics. Next, a query to the database is made, where each word of the incoming text is searched in the database. If the base is found in the database, it displays as an output text, otherwise it is skipped and the next word is searched.

## 4 Results

Examples of the revealed keywords are presented in Table 1.

**Table 1** – Examples of most frequent used words

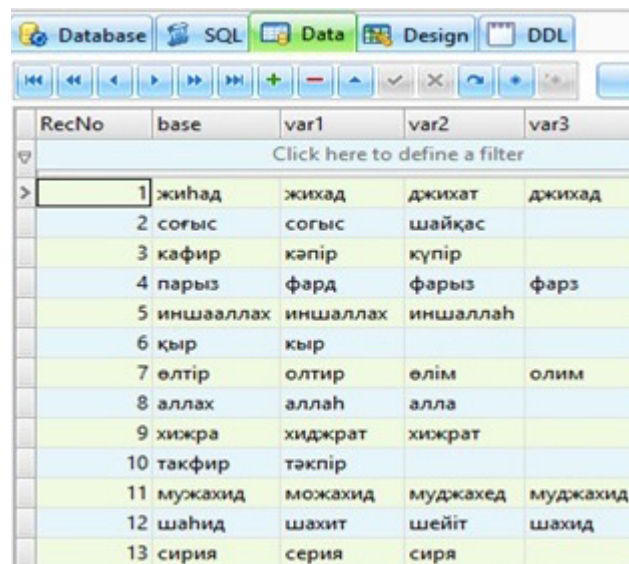
Variant 1	Frequency	Variant 2	Frequency	Variant 3	Frequency
allah (аллах)	21	alla (алла)	24	allah (аллах)	8
jihad (жихад)	29	djihad (джихад)	19	jihad (жихад)	9
sog'ys (соғыс)	13	sogys (соғыс)	6		
ka'pir (кәпір)	4	kafir (кафир)	2		
tozaq (тозақ)	4	tozak (тозак)	2		

F-IDF values of frequent words are given in Table 2. The table of keywords in the database is presented in Figure 1 These words can be used to improve the reliability of determining the extremist orientation in the text. In the future it is planned to assign emotional tones to the revealed words, which will later be used to create algorithms and software for analyzing the tonality of the text (sentiment analysis).

**Table 2**– TF-IDF values of used words

Keyword	TF-IDF value
allah (аллах)	25.62
jihad (жихад)	22.62
alla (алла)	19.92
djihad (джихад)	17.1
allah (аллах)	16.72
sog'ys (соғыс)	14.3
jihad (жихад)	11.43
sogys (соғыс)	8.4
ka'pir (кәпір)	7.98
tozaq (тозақ)	6.28
tozak (тозак)	5.88
kafir (кафир)	3.4

For the English language there are such dictionaries as, AFINN, General Inquirer, Senti-WordNet, SentiStrength, WordNet, which are used for sentiment analysis. For example, in this paper, the authors present the development of a bilingual Sentiment Analysis Lexicon (BiSAL) for the cyber security domain, which consists of a Sentiment Lexicon for Sensei (SentiLEN, 279 words) and a Sentiment Lexicon for ARabic (SentiLAR, 1019 words ) that can be used to develop mining and sentiment analysis systems for bilingual textual data from Dark Web forums (Al-Rowaily, 2015: 53-62).



RecNo	base	var1	var2	var3
1	жихад	жихад	джихат	джихад
2	соғыс	соғыс	шайқас	
3	кафир	кәпір	күпір	
4	парыз	фард	фарыз	фарз
5	иншааллах	иншаллах	иншаллах	
6	қыр	қыр		
7	өлтір	олтир	өлім	олим
8	аллах	аллах	алла	
9	хижра	хиджрат	хиджрат	
10	такфир	тәкпір		
11	мужахид	можахид	муджахед	муджахид
12	шаһид	шахит	шейіт	шахид
13	сирия	серия	сиря	

**Figure 1** - List of keywords in the database

The program constructed is working effectively, it finds all extremist keywords from database in the given input text. The program is being further development to classify given texts into "extremist" and "neutral" categories.

## 5 Conclusion

During the study, a corpus and key vocabulary were constructed for training and testing machine learning methods to determine the extremist orientation in the Kazakh texts. However, the size of the corpus is now small and work on expanding it is continuing. In the future it is planning to add a sentiment polarities between  $[-1;1]$ , which will be used during sentiment analysis. The next stage of the study is to classify the incoming texts using machine learning methods, such as the Bayesian method, support vector machine (SVM), random forest and logistic regression.

## References

- [1] Al-Rowaily Kh., Abulaish M., Haldar N., Al-Rubaian M., "BiSAL – A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security," *Digital Investigation: The International Journal of Digital Forensics & Incident Response archive* 14 (2015): 53–62.
- [2] Ananyeva M. I., Deviatkin D. A., Kobozeva M. V., Smirnov I. V., "Lingvostatisticheskii analiz tekstov ekstremistskoi napravlenosti [Lingua-statistic analysis of extremist texts]" (paper presented at the proceedings of International Conference on Situational Centers and Information-Analytical System 4i Class for Monitoring and Security Tasks, Russia, TsarGrad, November 21-24, 2015-2016).
- [3] Azizan S.A., Aziz I.A., "Terrorism detection based on sentiment analysis using machine learning," *Journal of Engineering and Applied Sciences* 12(3)(2017):691–698.
- [4] Chen Hsinchun, *Dark Web Exploring and Data Mining the Dark Side of the Web*(Springer, 2012), 3–105.
- [5] Cohen K., Johansson F., Kaati L. and Mork, Clausen J., "Detecting Linguistic Markers for Radical Violence in Social Media, Terrorism and Political Violence," *Terrorism and Political Violence* 26(2014): 246–253.
- [6] Devyatkin D.A., Smirnov V., Ananyeva V.I., Kobozeva M.V., "Exploring linguistic features for extremist texts detection (on the material of Russian-speaking illegal texts)" (paper presented at 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, July 22-24, 2017).
- [7] Enghin Omer, "Using machine learning to identify jihadist messages on Twitter" (Independent thesis, Uppsala University, 2015).
- [8] Finlayson M.A., Halverson J.R., Cormann S.R., "The N2 corpus: A semantically annotated collection of Islamist extremist stories" (paper presented at the Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Iceland, Reykjavik, May 26-31, 2014).
- [9] "Google Online Translator." Accessed January 12, 2018, <https://translate.google.com/#ru/kk/>.
- [10] Information Portal of Kazakhstan. "Okolo 500 kazakhstanstsev voiuuit v Sirii i Irake na storone IGIL. [About 500 Kazakhstanis are fighting in Syria and Iraq on the side of ISIS]." Accessed February 10, 2018, <http://today.kz/news/mir/2017-10-31/753466-okolo-500-kazahstantsev-voyuyut-v-sirii-i-irake-na-storone-igil/>.
- [11] "Integrated development environment Visual C." Accessed December 17, 2017, <https://www.visualstudio.com>.
- [12] Johansson F., Kaati L., Sahlgren M., "Detecting Linguistic Markers of Violent Extremism in Online Environments," in *Combating Violent Extremism and Radicalization in the Digital Era*, ed. Khader, Majeed (Hershey PA: Information Science Reference, 2016), 374–390.
- [13] Scanlon J.R., "Automatic Detection and Forecasting of Violent Extremist Cyber-Recruitment" (Master thesis, University of Virginia, 2014).
- [14] Scrivens R., Frank R., "Sentiment Based on the classification of radical text on the Web" (paper presented at the Proceedings of 2016 European Intelligence and Security Informatics Conference (EISIC), Uppsala, Sweden, August 17-19, 2016).
- [15] "SQLite Expert Database." Accessed January 15, 2018, [www.sqliteexpert.com](http://www.sqliteexpert.com).
- [16] Targeir A., Perera S., "Mapping Extremist Forums using Text Mining" (Master thesis, University of Agder, 2013).



- [17] Wikipedia. "tf-idf." Accessed November 12, 2017, <https://ru.wikipedia.org/wiki/TF-IDF>.
- [18] Zhavoronkova T.V., "Ispol'zovanie seti Internet terroristicheskimi i ekstremistskimi organizatsiiami [Usage of the Internet by terrorist and extremist organizations]." *Orenburg State University bulletin* 3 (178): 30.