**REPUBLIC OF AZERBAIJAN**

*On the rights of the manuscript*

# ABSTRACT

of the dissertation for the degree of Doctor of Philosophy

# DEVELOPMENT OF METHODS AND ALGORITHMS FOR INTELLECTUAL ANALYSİS OF DATA ON ELECTRONIC GOVERNMENT-CITIZEN PLATFORM

Speciality: 3338.01 – “System analysis, control and information processing”

Field of science: Technical Sciences

Applicant: **Günay Yavar gizi Iskandarli**

**Baku – 2021**

The work was performed at at the Institute of Information Technology of the Azerbaijan National Academy of Sciences (ANAS).

Scientific supervisor: Corresponding Member of ANAS, Doctor of Technical sciences **Ramiz Mahammad oglu Aliguliyev**

Official opponents: Doctor of Technical sciences, professor
**Nadir Bafadin oglu Aghayev**

PhD in Technical sciences
**Lala Hekayat gizi Karimova**

PhD in Technical sciences
**Vugar Yadulla oglu Musayev**

Dissertation council ED 1.35 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Information Technology of the ANAS.

Chairman of the Dissertation council:
Full member of ANAS
Doctor of Technical sciences, Prof.
**Rasim Mahammad oglu Alguliev**

Scientific secretary of the Dissertation council:
Doctor of Philosophy in Technical sciences,
Assoc. prof.
**Farhad Firudin oglu Yusifov**

Chairman of the scientific seminar:
Doctor of Technical sciences
Mütəllim **Mutallim Mirzaahmed oglu Mutallimov**

Azərbaycan Milli Elmlər Akademiyası
İnformasiya Texnologiyaları İnstitutu
Institute of Information Technology
Azerbaijan National Academy of Science

## GENERAL CHARACTERISTICS OF WORK

**The relevance of the study.** The main purpose of creating an e-government, which is one of the key elements of the information society, is to increase the level of services provided by government agencies to citizens. Thus, it is possible to simplify access to government information resources, ensure the active participation of all segments of society in public administration through e-government. The study of this environment is important for effective decision-making and ensuring national security in the e-government environment.

It should be noted that e-government is a complex socio-technological environment, and the main target in this environment is the relationship between government and citizens. There are problems in various segments of the e-government environment. However, the main issues covered in this dissertation are the analysis of the information space in government-citizen relations. Timely identification of issues of interest to people in the e-government environment can help government agencies to improve the quality of services and increase citizen satisfaction. To increase the availability and effectiveness of e-government services, regular user-oriented assessments are needed.

One of the main functions of e-government is to protect citizens from possible harm and violence. Experience shows that criminal groups also "take advantage" of this favorable environment, and they use this opportunity to become a major threat to the state and society. So, one of the important tasks of the government is to detect and analyze the activities of criminal networks operating secretly in the virtual environment - the Internet and e-government. This environment has a wide range of opportunities to communicate and coordinate activities. Criminal group members can communicate via websites, e-mail, blogs, online chats, etc. In most cases, transmitted information consists of text Therefore, analysis of the text transmitted through the virtual environment is essential to preventing possible terrorist acts and ensuring the security of e-government.

As mentioned above, security is a very important issue, and there are different approaches and views to ensure security. Unfortunately, the comments of citizens in the e-government environment have not been sufficiently analyzed. Currently, text mining is considered one of the most advanced and effective technologies in knowledge management, intellectual analysis of texts collected from various sources. with this in mind in the dissertation, the texts collected in e-government-citizen relations has been analyzed using social networking and text mining technologies and new approaches, methods and algorithms has been proposed for the systems that support effective decision-making in this environment.

**The aim of the work** is to study e-government with the help of social networking and text analysis technologies in order to increase the level of security of e-government and the quality of e-government services, to propose new approaches, methods and algorithms for systems that support effective decision making.

**Research methods** are based on natural language processing, data mining, text mining, topic modeling, graph theory, probability theory, social network analysis technologies.

**The main provisions of the defense:**

- a hybrid classification method for detecting terrorism-related text in e-government;
- a method based on sentiment analysis and Bayes clasifier for filtering texts that promote terrorism in e-government;
- method and algorithm based on sentiment analysis technology for extraction and analysis of hidden social networks in e-government;
- a method for automatic assessment of citizen satisfaction with e-government services;
- a method for identifying hot topics that the citizens (including the regions) cared in e-government.

**The scientific novelty of the dissertation** is determined by the following **results**:

- a hybrid classification method has been proposed to detect terrorism-related text in e-government;

- a method based on sentiment analysis and Bayes classifier has been proposed to filter texts related to terrorism in e-government;
- a method and algorithm based on sentiment analysis technology has been proposed for the detecting and analysis of hidden social networks in e-government;
- a method has been proposed for automatic assessment of citizen satisfaction from e-government services;
- a method based on clustering and topic modeling technologies has been proposed for identifying hot topics that the citizens (including the regions) cared in e-government.

**Practical significance of the work.** The obtained scientific-theoretical and practical results can be used for the detecting and analysis of the different social networks in online environments, for the filtering of terrorism-related texts, improving the quality of e-services, analysing of the citizens comments, identyfing of the hot topics that the citizens and regions cared, and etc.

**Approbation of the work.** The main scientific-theoretical and practical results were presented and discussed at the following conferences: "İnformasiya təhlükəsizliyinin multidissiplinar problemləri" II respublika elmi-praktiki konfransı (Bakı, 14 may 2015-ci il); "Big data: imkanları, multidissiplinar problemləri və perspektivləri" I respublika elmi-praktiki konfransı (Bakı, 25 fevral 2016-cı il); 10th International Conference on Application of Information and Communication Technologies – AICT 2016 (Baku, 12-14 October 2016); "İnformasiya təhlükəsizliyinin aktual multidissiplinar problemləri" IV respublika elmi-praktiki konfransı, (Bakı, 14 dekabr 2018-ci il); 2nd International Symposium on Applied Sciences and Engineering (Turkey, 7-9 April 2021).

**Scientific publications:** 14 scientific works on the topic of the dissertation were published. 6 of them were published in peer-reviewed journals, 7 theses in conference materials and 1 express-information. 3 articles from these scientific works were published in journals indexed in the Web of Science database.

**The structure and volume of the work:** The dissertation consists of an introduction, 4 chapters, a conclusion, a bibliography of 179 titles and one appendix, 11 figures and 14 tables.

*The applicant expresses its deep gratitude to the full member of ANAS, doctor of technical sciences, professor Rasim Alguliyev and scientific adviser, corresponding member of ANAS, doctor of technical sciences Ramiz Aliguliyev for their valuable advice, constant attention and comprehensive support in the implementation of the dissertation.*

## BRIEF OVERVIEW OF THE WORK

**In the introduction** the relevance of the dissertation has been justified, the purpose of the research and the issues has been detected. The scientific novelty and practical significance of the obtained results has been showed.

**In the first chapter ("E-government analysis technologies: text mining and social networks")** the concept of e-government, its maturity models has been analyzed, the role of text mining and social network analysis technologies in the analysis of e-government has been studied, and the current state of problems in its analyis has been studied [4-6, 8].

**In the second chapter ("Methods for detecting texts that promote terrorism in e-government")** a new method and approach for identifying terrorism-related documents in the e-government environment has been proposed [2, 3].

A hybrid classification method consisting of a linear combination of kNN, Bayes and the newly proposed RG methods to identify terrorism-related documents has been proposed in the **first section of the second chapter.**

**The proposed method:** Let's suppose that a terror-related vocabulary database (VBase) has been created in a certain language, and a semantic network of words (WordNet) has been developed.

Each phase of the proposed approach is explained in detail below:

**1)Initial filtering of documents:** First, the terms are extracted from the document and analyzed morphologically and described as a set of document words (terms), $d = (\mathrm{t}_1, \mathrm{t}_2, ..., \mathrm{t}_m)$. Then, using the Simkevic-Simpson measure, the similarity between VBase and the set $d = (t_1, t_2, ..., t_m)$ is calculated:

$$\mathrm{sim}_{\mathrm{S-S}}(d, \mathrm{VBase}) = \frac{|d \cap \mathrm{VBase}|}{|d|} \tag{1}$$

where $|A|-$ denotes the number of elements in set $A$.

If $\text{sim}_{\text{S-S}}(d, \text{VBase}) \geq \varepsilon$, then document $d$ is added to the set of suspicious documents and shifted to the next stage for identification. Where, $\varepsilon$ is a value of degree which defined in an experimental way.

**2) Semantic similarity of the words:** First, the informative content of the word is determined using $\text{IC(t)}$ to calculate the similarity between the words in WordNet:

$$\text{IC(t)} = 1 - \frac{\log(\text{synset}(t) + 1)}{\log(t_{\max})} \tag{2}$$

Then, semantic similarity between words is calculated using formula (2):

$$\text{sim}_{\text{IC}}(t_1, t_2) = \begin{cases} \frac{2 * \text{IC}(\text{LCS}(t_1, t_2))}{\text{IC}(t_1) + \text{IC}(t_2)}, & \text{if} \quad t_1 \neq t_2 \\ 1, & \text{if} \quad t_1 = t_2 \end{cases} \tag{3}$$

where, $\text{LCS}(t_1, t_2)$ – denotes the most similar common word with the words $t_1$ and $t_2$ in WordNet, $t_{\max}$ – denotes the total number of the words in WordNet, $\text{synset}(t)$ – the number of synonyms of the word $t$ .

Semantic similarity between the words is also calculated with the use of the WUP metric:

$$\text{sim}_{\text{WUP}}(t_1, t_2) = \frac{2 * \text{depth}(t)}{\text{depth}(t_1) + \text{depth}(t_2) + 2 * \text{depth}(t)} \tag{4}$$

where, $depth(t_1)$– denotes the number of nodes from $t_1$ to $t$ in WordNet (tree); $depth(t_2)$– is the number of nodes from $t_2$ to $t$; $depth(t)$ – the number of nodes from $t$ to the network roots.

The semantic similarity between the words is defined as a linear combination of the metrics given by the formulas (3) and (4):

$$\text{sim}(t_1, t_2) = \alpha \times \text{sim}_{\text{IC}}(t_1, t_2) + (1 - \alpha) \times \text{sim}_{\text{WUP}}(t_1, t_2) \tag{5}$$

where, $0 \leq \alpha \leq 1$ – denotes the weight coefficient.

**3)Similarity measure of the sentences:** Three metrics are used to calculate the similarity between the sentences: semantic, syntactic, and cosine.

***Semantic similarity.*** Semantic similarity between the sentences is calculated with the use of semantic similarity between the words (5):

$$\mathrm{sim}_{\mathrm{semantic}}(s_1, s_2) = \frac{\sum_{\mathrm{t}_1 \in s_1, \mathrm{t}_2 \in s_2} \mathrm{sim}(\mathrm{t}_1, \mathrm{t}_2)}{m_1 + m_2} \tag{6}$$

where $m_1$ and $m_2$ is the number of words in the sentences $s_1$ and $s_2$ respectively.

***Cosine metric.*** Using the cosine metric, the similarity between the two vectors is calculated as follows:

$$\mathrm{sim}_{\cos}(s_1, s_2) = \frac{\sum_{j=1}^{m}(w_{1j} \times w_{2j})}{\sqrt{\sum_{j=1}^{m} w_{1j}^2} \times \sqrt{\sum_{j=1}^{m} w_{2j}^2}}$$

where $s_1 = (w_{11}, w_{12}, \ldots, w_{1m})$ və $s_2 = (w_{21}, w_{22}, \ldots, w_{2m})$ – are the vectors corresponding to the sentences $s_1$ and $s_2$; $w_{pj}$ – denotes the weight of the word $\mathrm{t}_j$ in vector $s_p$; $m$ is the total number of words.

***Syntactic similarity.*** The following formula is used to calculate the similarity of a sentence based on the position of words in a sentence:

$$\mathrm{sim}_{\mathrm{wordorder}}(s_1, s_2) = 1 - \frac{\|\mathrm{o}_1 - \mathrm{o}_2\|}{\|\mathrm{o}_1 + \mathrm{o}_2\|}$$

where, $\mathrm{o}_1 = (w_{11}, w_{12}, \ldots, w_{1m})$ and $\mathrm{o}_2 = (w_{21}, w_{22}, \ldots, w_{2m})$ – are syntactic vectors of sentences $s_1$ and $s_2$; $w_{pj}$ denotes the weight of the word $\mathrm{t}_j$ in vector $\mathrm{o}_p$, $\|\cdot\|$ - is the Euclidean norm.

**Liner combination.** Linear combination of semantic cosine and syntactic measures is used to calculate the similarity between the sentences:

$$\begin{aligned}\mathrm{sim}_{\mathrm{sentences}}(s_1, s_2) = \beta_1 \cdot \mathrm{sim}_{\mathrm{semantic}}(s_1, s_2) + \beta_2 \cdot \mathrm{sim}_{\mathrm{wordorder}}(s_1, s_2) \\ + \beta_3 \cdot \mathrm{sim}_{\cos}(s_1, s_2)\end{aligned} \tag{7}$$

where, $\beta_i$ $(0 \le \beta_i \le 1,\ i = 1,2,3)$ are the weight parameters and provide the following condition: $\beta_1 + \beta_2 + \beta_3 = 1$.

**4) Similarity measure of the documents.** The similarity between the sentences (7) is used to define the similarity between the documents:

$$\mathrm{sim}_{\mathrm{documents}}(d_1, d_2) = \frac{\sum_{s_1 \in d_1, s_2 \in d_2} \mathrm{sim}_{\mathrm{sentences}}(s_1, s_2)}{n_1 + n_2}$$

where, $n_1$ and $n_2$ are the numbers of the sentences in the documents $d_1$ və $d_2$ respectively.

For simplicity, $\mathrm{sim}_{\mathrm{documents}}(d_1, d_2)$ is used instead of $\mathrm{sim}(d_1, d_2)$ below.

**5)Classification of documents:** Suppose that a set of classes $\mathbf{C} = (C_1, ..., C_k)$ is known. kNN ($k$-Nearest Neighbor), Bayes and the proposed RG method are used to determine the extent to which document $d_i$ belongs to class $C_q$.

***kNN method.*** According to this method, the extent to which the document $d_i$ belongs to class $C_q$ is defined by the value found by the following formula:

$$\mathrm{score}_{k\mathrm{NN}}(d_i \mid C_q) = \sum_{d \in k\mathrm{NN}_q(d_i)} \mathrm{sim}(d_i, d), \quad i = 1, 2, ..., N; \quad q = 1, 2, ..., k \quad (8)$$

where, $k\mathrm{NN}_q(d_i)$ – is $k$ number of documents, which is closest to the document $d_i$ to class $C_q$.

The document $d_i$ belongs to the class with the highest value $\mathrm{score}_{k\mathrm{NN}}(d_i \mid C_q)$, in other words, $d_i \in C_{k^*}$, if $k^* = \arg\max_q \mathrm{score}_{k\mathrm{NN}}(d_i \mid C_q)$.

***Modified Bayes method.*** According to the Bayes method, the degree to which document $d_i$ belongs to class $C_q$ is defined by the value of the following conditional probability:

$$\mathrm{score}_{\mathrm{M\,Bayes}}(C_q|d_i) = \mathrm{P}(C_q|d_i) = \frac{\log \mathrm{P}(C_q)}{w_i} + \sum_{j=1}^{m} \mathrm{P}(\mathrm{t}_j, d_i) \log \mathrm{P}(t_j|C_q) \tag{9}$$

where $P(t_j, d_i) = w_{ij} / w_i$ – is the probability of the word $\mathrm{t}_j$ to be used in the document $d_i$, $w_i = \sum_{j=1}^{m} w_{ij}$, $i = 1,...,n$; $q = 1,...,k$. $w_{ij}$ – is the weight of the word $\mathrm{t}_j$ in the document $d_i$, $\mathrm{P}(\mathrm{t}_j|C_q)$ – is the probability of the word $\mathrm{t}_j$ in the class $C_q$, $m$ – is the number of words in document set $\mathbf{D}$, $\mathrm{P}(C_q)$ – is the probability of the documents being in class $C_q$.

Similar to the *k*NN method, $\mathrm{score}_{\mathrm{MBayes}}(C_q|d_i) = \mathrm{P}(C_q|d_i)$ is adopted in the formula (9). According to the model, $d_i$ belongs to a class for which the probability $\mathrm{P}(C_q|d_i)$ has the highest value, $d_i \in C_{k*}$, where, $k^* = \arg \max_{1 \le q \le k} \mathrm{score}_{\mathrm{M\,Bayes}}(C_q|d_i)$

***RG method.*** With the help of this method, the degree to which document $d_i$ belongs to class $C_q$ is defined by the following formula:

$$\mathrm{score}_{\mathrm{RG}}(d_i \mid C_q) = \lambda \times \frac{\mathrm{sim}(\mathrm{O}_{d_i}, \mathrm{O}_{C_q})}{\sum_{p=1}^{k} \mathrm{sim}(\mathrm{O}_{d_i}, \mathrm{O}_{C_p})} + $$
$$+ (1-\lambda) \times \frac{\sum_{v \in C_q} \mathrm{sim}(\mathrm{O}_{d_i}, \mathrm{O}_v)}{\sum_{p=1}^{k} \sum_{d \in C_p} \mathrm{sim}(\mathrm{O}_{d_i}, \mathrm{O}_d)} \tag{10}$$

where, $\mathrm{sim}(\mathrm{O}_{d_i}, \mathrm{O}_{C_q})$ – is a similarity measure between the image $\mathrm{O}_{d_i}$ of the document $d_i$ and the image $\mathrm{O}_{C_q}$ of the class $C_q$ sinfinin obrazı; $\mathrm{sim}(\mathrm{O}_d, \mathrm{O}_v)$ – is a similarity measure between the images $\mathrm{O}_d$ and $\mathrm{O}_v$ of the documents $d$ və $v$; $\lambda$ with the weight coefficient, $0 \le \lambda \le 1$ .

Image $\mathrm{O}_{C_q}$ is defined as the centre of the class $C_q$, $\mathrm{O}_{C_q} = (w_1^q, w_2^q, ..., w_m^q)$:

$$w_j^q = \frac{1}{|C_q|} \sum_{d \in C_q} w_j^{q,d} \text{ , } q = 1,...,k \text{ , } j = 1,...,m$$

where, $|C_q|$ – is the number of documents in the class $C_q$, $w_j^{q,d}$ and the weight of $j$ –th words in the document $d$ included in the class $C_q$.

Analogically, Image $\mathrm{O}_d$ is defined as the center of the document $d$, $\mathrm{O}_d = (w_1^d, w_2^d, ..., w_m^d)$:

$$w_j^d = \frac{1}{|d|} \sum_{s \in d} w_j^{d,s} \text{ , } j = 1,...,m$$

where, $|d|$ – is the number of sentences in the document $d$, $w_j^{d,s}$ as the weight of $j$ –th words in the sentence $s$ included in the document $d$.

**Hybrid method.** As a final classification method, a linear combination of the results obtained by means of formulas (8), (9) and (10) is used:

$$\text{score } (d^{\text{new}} \mid C_q) = \gamma_1 \cdot \text{score}_{k\text{NN}}(d^{\text{new}} \mid C_q) + \gamma_2 \cdot \text{score}_{\text{Bayes}}(d^{\text{new}} \mid C_q)$$
$$+ \gamma_3 \cdot \text{score}_{\text{RG}}(d^{\text{new}} \mid C_q)$$

where, $0 \leq \gamma_i \leq 1$, $(i = 1,2,3)$ denote weight coefficients and provid the following condition: $\gamma_1 + \gamma_2 + \gamma_3 = 1$.

Thus, the document $d^{\text{new}}$ belongs to a class $C_{k^*}$ such that it has the highest $\text{score}(d^{\text{new}} \mid C_q)$ and may have the new value $d^{\text{new}} \in C_{k^*}$, for the class, where $k^* = \arg\max_q \text{score}(d^{\text{new}} \mid C_q)$.

**6)Evaluation:** Accuracy, precision, recall, and the F-measure are used to evaluate the classification:

$$\text{Accuracy} = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$\text{Precision} = \frac{T_p}{T_p + F_p}$$

$$\text{Recall} = \frac{T_p}{T_p + F_n}$$

$$\text{F} - \text{measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where, $T_p$– denotes the number of precisely classified terror-related documents; $F_p$ – is the number of incorrectly classified terror related documents; $T_n$ – is the number of precisely classified non-terror-related documents; $F_p$ – is the number of incorrectly classified non-terror-related documents.

As is known e-government is an interactive environment and everyone can benefit from this environment. Thus, members of terrorist groups can write threatening messages to the state at the same time to frighten the state, send a political message, and then post a trail. Taking this into account, in the **second section of the second chapter** proposes a new approach to filtering texts related to terrorism [9,10].

The main purpose of the proposed method is to define suspicious comments based on the analysis of them.

Each stage is explained in detail below:

**Pre-processing:** During the pre-processing, each text (comment) is divided into sentences by "." and "!". The text is cleared from common words. All words are returned to the original version because of accepting various forms of suffixes. Then, the synonyms of each word are detected and described as a set.

**Determination of Opinions Polarity:** First, the comment are described as a set of sentences, $T = \{S_1, S_2, ..., S_N\}$. Where, $N$ – is the number of sentences. Then, using the following equation, the polarity of each comment is determined:

$$score\,(T) = sign\,(\sum_{S \in T} score\,(S))$$

where, the $score\,(S)$– is the polarity degree of the $S$ sentence included in the comment. The function $sign\,(x)$ is defined as follows:

$$sign\,(x) = \begin{cases} 1, & \text{if} \quad x > 0 \\ 0, & \text{if} \quad x = 0 \\ -1, & \text{if} \quad x < 0 \end{cases}$$

If the overall polarity of sentences in the comment greater than zero, then the comment can be included to positive, if equals to zero, then included to neutral else negative class. The polarity degree $score(S)$ is calculated using the following formula:

$$score\,(S) = \sum_{w \in S} score\,(w)$$

The polarity of the word $w$ is calculated in the order shown in [10].

**Selection of Terrorism-Related Opinions:** In the next stage, in order to determine whether the selected negative comments are related to terrorism, their preliminary comparison is made with the words in the pre-created vocabulary database. If the similarity is greater than the defined threshold, then a more detailed comparison between the review and vocabulary database is made.

Let, $V_{terror}$– is a vocabulary database that contains words about terror and their extension. Here, the extension means the synonym set of words. Selected negative comments are described as $T = \{w_1, w_2, ..., w_m\}$. Each word is described as $w_i \rightarrow synset\,(w_i),\ i = 1, 2, ..., m$ word sets. To calculate the similarity between the comments and words in the vocabulary database the following formula is proposed:

$$SW_i = synset\,(w_i) \cap V_{terror} = \{w_{i1}, w_{i2}, ..., w_{im_i}\},\ i = 1, ..., m \quad (11)$$

Then we determine how similar the words to each other obtained by the formula (11):

$$\theta_i = \frac{2}{m_i(m_i-1)} \sum_{j=1}^{m_i-1} \sum_{k=j+1}^{m_i} sim(w_{ij}, w_{ik}),\ i = 1,...,m \qquad (12)$$

For estimation of semantic similarity between words, the method proposed in [3] is used.

Then, the polarity degree of the words determined in formula (12) is summed up. If the amount exceeds the predetermined threshold $t$, then the commented user is considered to be suspicious of terrorism and is gained control:

$$P(T) = \begin{cases} 1, & \text{if } \ \frac{1}{m}\sum_{i=1}^{m}\theta_i \geq t \\ 0, & \text{else} \end{cases}$$

where “1” indicates that the statement is related to terrorism and, ”0” is not.

The determination of relation probability of users' comments with terrorism is possible. For this purpose applying of the Naive Bayes classification method is recommended. In order to estimate the probability of suspicious comments being terrorism-related, we have to determine the probability of every word included in comments being terrorism-related:

$$P(T) = P(w_1, w_2, ..., w_n)$$

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i)$$

where $P(T)$ - means the probability of the user's comment and, $P(w_i)$ indicates the probability of each word being terrorism-related included in this comment.

Note that the extended synonym sets of each word are also taken into account. In order to estimate the probability of the comments being terrorism-related, the usage frequency of each word and its synonyms in the precreated $V_{terror}$ vocabulary is regarded. The probability of the suspicious comments being terrorism-related is estimated on this basis. According to the Bayes method, the use of following formula is suggested:

$$P(V_{terror}|T)=\frac{P(T|V_{terror})P(V_{terror})}{P(T)}$$

$$P(T|\,V_{terror})=P(w_1,w_2,...,w_n|V_{terror})=\prod\nolimits_{i=1}^{n}P(w_i|\,V_{terror}) \qquad (13)$$

$$P(w_i|\,V_{terror})=\frac{P(w_i\cap V_{terror})}{P(V_{terror})}$$

where $P(V_{terror}|T)$- indicates the probability of the suspicious comments being from $V_{terror}$ class.

Note that the absence of one of the words from (13) in the vocabulary database directly, affects the result (thus, the probability of the user's opinion being terrorism-related is 0 if at least one word is not in $V_{terror}$ vocabulary). In this case, to estimate the probability of the text being terrorism-related, it would be better to use the following formula instead of the formula (13):

$$P(T|\,V_{terror})=\frac{1}{n}\sum_{i=1}^{n}P(w_i|\,V_{terror})$$

Thus, according to the user's comments, the terrorism-related users are identified and their activities are monitored by relevant authorities.

**The third chapter ("Method and algorithm for extracting hidden social networks in e-government")** has been devoted for the extracting and analysing of the hidden social networks[1,11].

Here the methods for detecting hidden social networks in various environments has been explored, and a method for detecting hidden social networks via text mining and social network analysis technologies using user comments in e-government environment has been proposed.

The proposed approach can be divided into 4 principal steps:

1. Data collection and preprocessing;
2. Classification;
3. Social network extraction;
4. Social network analysis.

Each stage of the proposed approach is explained in detail below.

**1) Data collection and preprocessing:** Suppose, $n$ number information $T_i$, $(i = 1,2,...,n)$ are placed in e-government environment. The comments written to the $i$ - th information are denoted as follows:

$$C_i = \{c_i^j\},\ i = 1,2,...,n,\ j = 1,2,...,m$$

where, $c_i^j$ –is the set of comments written by $j$ - th user to the $i$ –th information, $m$ – is the number of users.

After collecting comments, the preprocessing process is carried out on them.

**2) Classification:** At the next stage, the set of comments written to each information are grouped into 3 classes: positive class is denoted as $C_i^+$ , negative $C_i^-$ and neutral $C_i^0$:

$$C_i = C_i^+ \cup C_i^- \cup C_i^0,\ i = 1,...,n$$

In this study, sentiment analysis is utilized to group the comments into three classes. Sentiment analysis is one of the most advanced technologies for the analysis of texts.

To group the written comments into these three classes, the polarity of each comment is determined using the method proposed above:

$$C_i^- = \{c_i^j \mid score(c_i^j) = -1\}$$

$$C_i^+ = \{c_i^j \mid score(c_i^j) = 1\}$$

$$C_i^0 = \{c_i^j \mid score(c_i^j) = 0\}$$

**3) Social network extraction:** At this stage, the social network actors and the relationships between them are determined. Firstly, users gathering around the negative class are defined. We consider that each user can be identified (either by registering, or by IP address). Users writting negative comments at least to one information are defined as follows:

$$U_\Sigma^- = \bigcup_{i=1}^n U_i^-$$

where $U_i^-$ – are the users writting negative comments to the $i$ - th information.

Users writting negative comments to all information are defined as follows:

$$U_{\Pi}^{-} = \bigcap_{i=1}^{n} U_i^{-}$$

where $U_{\Pi}^{-}$ - is the core (key actors) of social network.

Two types of approaches are used to determine the relationship between the social network actors:

***In the first approach,*** the relationships between social network actors are determined through the number of information written negative comments by the users:

$$w_1^{j_1 j_2} = \frac{n^{j_1 j_2}}{n^{j_1} + n^{j_2}}$$

where $n^{j_1 j_2}$ – is the number of information written negative comments by $j_1$ and $j_2$ th users, $n^{j_1}$– is the number of information written negative comments by $j_1$-th user, $n^{j_2}$– is the number of information written negative comments by $j_2$ - th user:

$$n^{j_1 j_2} = \sum_{i=1}^{n} I(c_i^{j_1}) \cdot I(c_i^{j_2})$$

$$n^{j_1} = \sum_{i=1}^{n} I(c_i^{j_1})$$

$$n^{j_2} = \sum_{i=1}^{n} I(c_i^{j_2})$$

where $I(c_i^j)$ is defined as follows:

$$I(c_i^j) = \begin{cases} 1, & \text{if} \quad c_i^j \neq \varnothing \\ 0, & \text{else} \end{cases}$$

If the $j$ - th user comments at least one time to the $i$ - th information then $I(c_i^j)$ function is defined as 1, else 0.

Here the number of negative comments written to the same information by users can be considered too. In this case, the weight of relationship between the users can be defined by the following formula:

$$\tilde{w}_1^{j_1 j_2} = \frac{\sum_{i=1}^{n}\left(m_i^{j_1} + m_i^{j_2}\right) * I(c_i^{j_1}) \cdot I(c_i^{j_2})}{M^{j_1} + M^{j_2}}$$

$$M^{j} = \sum_{i=1}^{n} m_i^{j}$$

where $m_i^{j_1}$ – is the number of comments written by the $j_1$- th user to the $i$ - th information, $m_i^{j_2}$ – is the number of comments written by the $j_2$ - th user to the $i$ - th information. $\sum_{i=1}^{n}\left(m_i^{j_1} + m_i^{j_2}\right) * I(c_i^{j_1}) \cdot I(c_i^{j_2})$ – is the overall number of comments written by the $j_1$ and $j_2$ th users to the same information, $M^{j}$ – is the overall number of comments written by the $j$ - th user.

***In the second approach***, the relationships between the social network actors are determined through the semantic similarity between comments written by users in the negative class. Here, Jaccard measure is used to calculate the similarity between comments:

$$w_2^{j_1 j_2} = sim(c^{j_1}, c^{j_2}) = \frac{\left|c^{j_1} \cap c^{j_2}\right|}{\left|c^{j_1} \cup c^{j_2}\right|}$$

where $sim\,(c^{j_1}, c^{j_2})$ – is the semantic similarity between comments written by the $j_1$ and $j_2$ th users.

So, the relationships between the actors of hidden social network are determined through the linear combination of above-proposed weights:

$$w^{j_1 j_2} = \alpha \cdot \widehat{w}_1^{j_1 j_2} + (1-\alpha) \cdot w_2^{j_1 j_2}$$

where $\alpha\left(0 \le \alpha \le 1\right)$ denotes weight coefficients.

**4) Social network analysis:** To determine the key actors in the social network, it is necessary to show compactness of the core. Therefore, using the number of users and relations between them in the social network is proposed.

After determining the number of users in the social network, the number of relations between them is defined as follows:

$$M_{\Sigma}^{-} = \sum_{j_1, j_2 \in U_{\Sigma}^{-}} I_1(w^{j_1 j_2})$$

$$M_{\Pi}^{-} = \sum_{j_1, j_2 \in U_{\Pi}^{-}} I_1(w^{j_1 j_2})$$

where $M_{\Sigma}^{-}$ – is the number of relations between users in the whole social network, $M_{\Pi}^{-}$ – is the number of relations between users in the core of the social network. The function $I_1(w^{j_1 j_2})$ is defined as follows:

$$I_1(x) = \begin{cases} 1, & \text{if} \quad x > 0 \\ 0, & \text{if} \quad x = 0 \end{cases}$$

Then the density coefficient of the whole network is determined using the folllowing formula:

$$\sigma_{\Sigma}^{-} = \frac{M_{\Sigma}^{-}}{\dfrac{N_{\Sigma}^{-}(N_{\Sigma}^{-} - 1)}{2}} \tag{14}$$

where $N_{\Sigma}^{-} = \left|U_{\Sigma}^{-}\right|$ – is the number of users in the whole social network, $\dfrac{N_{\Sigma}^{-}(N_{\Sigma}^{-} - 1)}{2}$ – is the number of possible relations between the social network actors.

Similarly, the density coefficient of the core is determined as follows:

$$\sigma_{\Pi}^{-} = \frac{M_{\Pi}^{-}}{\dfrac{N_{\Pi}^{-}(N_{\Pi}^{-} - 1)}{2}} \tag{15}$$

where $N_{\Pi}^{-} = \left|U_{\Pi}^{-}\right|$ – is the number of users in the core of the social network, $\dfrac{N_{\Pi}^{-}(N_{\Pi}^{-} - 1)}{2}$ – is the number of all possible relations between the core's actors.

The weight of the core in the whole social network is defined using the (14) and (15) formulas as follows:

$$\sigma^{-} = \frac{\sigma_{\Pi}^{-}}{\sigma_{\Sigma}^{-}}$$

where $\sigma^{-}$ – is the weight of the core in the whole social network. Based on this, compactness of the core is defined.

After identifying the compactness of the core, the importance score of core actors is calculated by using the number and weight of relations between users. For this purpose, the following formula is proposed:

$$c_j^{wa} = k_j^{(1-\alpha)} \cdot s_j^{\alpha}, \quad 0 \le \alpha \le 1$$

$$k_j = \sum_{l \in U_{\Sigma}^{-}} I_1(w^{jl})$$

$$s_j = \sum_{l \in U_{\Sigma}^{-}} w^{jl}$$

where $c_j^{w\alpha}$ – is the centrality degree, $k_j$ – is the total number and $s_j$ is the total weight of relations between the $j$ th actor of the core and other actors in the network, respectively.

So, hidden social networks suspected of anti-government propaganda, the key actors of this network and their importance degree have been defined through the analyzing comments of citizens in e-government environment.

**In the fourth chapter ("Method and algorithm for feedback mechanisms in e-government")** the methods for identifying hotspot services in e-government, determining regional interests and citizen satisfaction with services, as well as the hot topics of citizens' comments on e-government, and improving the quality of e-services has been suggested [7,13].

The method has been proposed in **the first section of the fourth chapter** is as follows:

Let, the number of services proposed by e-government portal is $m$ the number of citizens using these services is $n$. Let's point them as $(EG_1, EG_2, ..., EG_m)$ and $(U_1, U_2, ..., U_n)$. We should note that each service can be evaluated within a certain period of time. The number of citizens using the services and their satisfaction score can be used for this purpose. In the proposed method we will first establish a service usage vector for each user within a specific $T$ – time period to determine the satisfaction degree of citizens for each service and to find hotspot services:

$$U_i = \{ u_{i1}, u_{i2}, ..., u_{im} \}, \quad i = 1,2,...,n \tag{16}$$

where, $u_{ij}$ $(i = 1,2,...,n, \; j = 1,2,...,m)$ – represents the usage of $i$ - th user to the $j$ - th servise:

$$u_{ij} = (u_{ij,1}, u_{ij,2}) \tag{17}$$

where, $u_{ij,1}$ – denotes the number of usages from the $j$ - th service by the $i$ - th user, and $u_{ij,2}$ is the satisfaction element from the service. It should be noted two variants are possible here: 1) A number of service usage is not taken into account; 2) A number of service usage is taken into account.

In the **first variant** a number of service usage is not taken into account. It means that the number of usages from the service is not considered, we only take into account if they use the service or not. If the citizen has used the service, it is evaluated by 1, if not then by 0:

$$u_{ij,1} = \begin{cases} 1 & \text{if the i - the user has used j - th service,} \\ 0 & \text{otherwise.} \end{cases}$$

The scale [1, 5] is recommended to evaluate the degree of citizens satisfaction from services:

$$u_{ij,2} = \begin{cases} 1 & \text{very poor} \\ 2 & \text{poor} \\ 3 & \text{normal} \\ 4 & \text{good} \\ 5 & \text{very good} \end{cases} \tag{18}$$

Let's note that if the user has not used the service, then the service vector is accepted as $u_{ij} = (0,0)$.

Considering (17), we can present a citizens usage matrix from services in the following way:

$$U = \begin{Vmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \dots & u_{nm} \end{Vmatrix} = \begin{Vmatrix} (u_{11,1}, u_{11,2}) & \dots & (u_{1m,1}, u_{1m,2}) \\ \vdots & \ddots & \vdots \\ (u_{n1,1}, u_{n1,2}) & \dots & (u_{nm,1}, u_{nm,2}) \end{Vmatrix} \qquad (19)$$

We can sum up the similar elements of these vectors on the rows after establishing of each users' vector within services. This can help us to determine how much the citizens are in need of every service and their satisfaction from these services. In this case, each e-service $EG_j \left( j = 1,2,...,m \right)$, can be expressed in the form of two-dimensional vectors:

$$EG_j = \sum_{i=1}^{n} u_{ij} = \left( \sum_{i=1}^{n} u_{ij,1}, u_{ij,2} \right) = \left( u_{j,1}, u_{j,2} \right) \quad j = 1,2,...,m \qquad (20)$$

where, the first element of the vector $u_{j,1}$ – is the total number of usages from the $j$-th service, $u_{j,2}$ – is the total satisfaction degree from it. By using (20), we can estimate the average satisfaction degree from the $j$- th service:

$$EG_j^{savg} = \frac{u_{j,2}}{u_{j,1}}$$

where, $EG_j^{savg}$ – represents the average satisfaction degree from the $j$-th service. If we make ranking according to $EG_j^{savg}$, we will obtain citizen satisfaction rating from services.

The **second variant** we consider the number of service usage. It means that the same citizen may use the service several times and regularly evaluate the service with different satisfaction scores. In this case, the average satisfaction rating of e-services will be performed in the following way.

Considering the number of usages from the service, we can present the usage vector as the following:

$$u_{ij}=\left(u_{ij,1},u_{ij,2}\right)=\left(N_{ij},\sum_{k=1}^{N_{ij}}u_{ij,2}^{k}\right)=\left(N_{ij},u_{ij,2}^{\Sigma}\right) \qquad (21)$$

where, $N_{ij}$ – is the number of usages of $i$ - th user from the $j$ - th service, $u_{ij,2}^{k}$ – is the satisfaction score using in $k$ - th time, and $u_{ij,2}^{\Sigma}$ – is the total satisfaction score given by $i$ - th user to the $j$ - th service. Using (21), we also can determine the average satisfaction degree of the $i$ - th user from $j$ - th service:

$$u_{ij}^{savg}=\frac{u_{ij,2}^{\Sigma}}{N_{ij}}$$

Then, the matrix (19) is represented as follows:

$$U=\begin{Vmatrix} u_{11} & \dots & u_{1m} \\ \vdots & \ddots & \vdots \\ u_{n1} & \dots & u_{nm} \end{Vmatrix}=\begin{Vmatrix} (N_{11},u_{11}^{savg}) & \dots & (N_{1m},u_{1m}^{savg}) \\ \vdots & \ddots & \vdots \\ (N_{n1},u_{n1}^{savg}) & \dots & (N_{nm},u_{nm}^{savg}) \end{Vmatrix}$$

According to it, we can express (20) as follows:

$$EG_j=\sum_{i=1}^{n}u_{ij}=\left(\sum_{i=1}^{n}N_{ij},\sum_{i=1}^{n}u_{ij,2}^{\Sigma}\right)=\left(N_j,u_{j,2}^{\Sigma}\right) \qquad (22)$$

Considering the number of usages as in (22), we can determine the average satisfaction degree from each service:

$$EG_j^{savg}=\frac{u_{j,2}^{\Sigma}}{N_j}$$

We will get citizens satisfaction rating from services if we conduct a ranking according to $EG_j^{savg}$.

We can also find the average satisfaction degree of each user from all services:

$$u_i^{savg}=\frac{1}{m}\sum_{j=1}^{m}u_{ij}^{savg}$$

where, $u_i^{savg}$ – denotes the satisfaction degree of the $i$ - th user from all services.

In this case, the average satisfaction degree of all users from e-services is expressed as follows:

$$U^{savg} = \frac{1}{n}\sum_{i=1}^{n} u_i^{savg} \tag{23}$$

where, $U^{savg}$ – defines the average satisfaction degree of users from e-services.

Through (23), the e-system is evaluated entirely.

We can also conduct a regional evaluation of each service. Accordingly, we can determine which regions are more using the services and do they satisfy from them. $u_{ij}^{savg}$- can be used for this purpose. So, if we conduct a ranking according to $u_{ij}^{savg}$- we will get the users satisfaction rating from $j$- th service. After ranking, we can divide the rating table into three sections according to (18): $U_j^-$ (not satisfied), $U_j^0$ (satisfied), $U_j^+$ (very satisfied). Here $U_j^-$ – is a users group which satisfaction scores from $j$- th service is in [1,2), $U_j^0$ – is in [2,4), $U_j^+$ – is in [4,5) interval.

After the determination of the user groups, we will look through the intersection of these groups within services:

$$\begin{aligned} U^- &= \bigcap U_j^- \\ U^0 &= \bigcap U_j^0 \\ U^+ &= \bigcap U_j^+ \end{aligned} \tag{24}$$

where $U^-$ – is dissatisfied, $U^0$ – is satisfied, $U^+$ – is very satisfied users group from all services. We can associate a citizen to the region who uses each service. Considering that each user has their own IP address, it is possible to determine automatically which regions the dissatisfied and satisfied citizens belong to according to the equation (24). Possibly, these citizens are gathered in one region or distributed over the regions.

We can also determine the common interest of the regions. So, we are able to determine the hotspot services for the regions by defining the services that each region uses and the rating of these services. For this purpose, we divide $(U_1, U_2, ..., U_n)$ users into regions through IP as mentioned above. By using (21), the services used by each user (region) are defined. Ranking the services requested by the users from the same region (in decreasing order), we will obtain usage rating of the service. Based on this rating, services most commonly addressed by users from the same region are identified. E-service interests of the regions are automatically assigned from here.

Table 1

Number of usages from services and their total scores

| ***Users*** | ***E-services*** | | | | |
|---|---|---|---|---|---|
| | ***EG₁*** | ***EG₂*** | ***EG₃*** | ***EG₄*** | ***EG₅*** |
| $U_1$ | (41, 82) | (33, 44) | (22,80) | (38,76) | (17,34) |
| $U_2$ | (46,153) | (1, 4) | (19,69) | (13,43) | (42,112) |
| $U_3$ | (6,16) | (43,129) | (39,104) | (25,100) | (29,106) |
| $U_4$ | (46,107) | (47,156) | (40,106) | (35,128) | (28,56) |
| $U_5$ | (32,64) | (34,124) | (9,30) | (45,45) | (46,214) |
| $U_6$ | (4,13) | (38,152) | (24,64) | (48,80) | (14,46) |
| $U_7$ | (14,56) | (37,86) | (22,58) | (27,90) | (38,126) |
| $U_8$ | (27,108) | (20,40) | (32,96) | (7,32) | (38,139) |
| $U_9$ | (48,192) | (33,88) | (36,72) | (7,11) | (19,76) |
| $U_{10}$ | (49,130) | (8,32) | (38,126) | (13,43) | (28,46) |
| $U_{11}$ | (8,24) | (36,96) | (14,46) | (42,98) | (3,11) |
| $U_{12}$ | (49,98) | (1,3) | (34,90) | (12,32) | (2,7) |
| $U_{13}$ | (48,160) | (14,28) | (33,132) | (41,123) | (27,117) |
| $U_{14}$ | (24,88) | (2,7) | (8,26) | (12,32) | (39,78) |
| $U_{15}$ | (40,146) | (4,5) | (6,20) | (47,125) | (47,141) |
| $U_{16}$ | (7,21) | (41,41) | (25,91) | (17,39) | (6,8) |
| $U_{17}$ | (21,56) | (35,105) | (48,112) | (10,40) | (29,106) |
| $U_{18}$ | (46,168) | (16,32) | (17,45) | (12,32) | (23,69) |
| $U_{19}$ | (40,53) | (48,144) | (29,87) | (31,103) | (0,0) |
| $U_{20}$ | (48,112) | (1,2) | (11,25) | (24,56) | (17,45) |

To evaluate the proposed method the calculations were performed in the Matlab 2018. Let, the scores presented the number of usages from e-services in a certain time interval and the satisfaction degree with these services is described in Table 1. The first number of each vector given in Table 1 is the total number of each user’s usages from the service, and the second one is the total satisfaction score given to it. Note that the regions that the users belong to are described in Table 2.

Table 2

Distribution of users by the regions

| **Regions** | **Users** |
|---|---|
| $R_1$ | $U_2, U_5, U_7, U_{15}, U_{14}$ |
| $R_2$ | $U_1, U_{20}, U_9, U_{19}$ |
| $R_3$ | $U_8, U_{11}, U_{12}$ |
| $R_4$ | $U_3, U_4, U_{13}, U_{10}$ |
| $R_5$ | $U_{16}, U_{17}, U_6, U_{18}$ |

Based on this data, calculations were made, e-services satisfaction rating and user rating were determined (figure 1 və 2). The e-system was then has been evaluated in general using these calculations:

$$U^{savg} = \frac{1}{20} * (2.1939 + 3.3864 + ... + 2.3173) = 2.8687$$

The system's performance can be considered "satisfying" because of the $U^{savg} = 2.8687$ total satisfaction degree from the system being in [2,4) interval.

Hotspot services for the regions have been defined by ranking after determining the services frequently used by the users from the same region (table 3). Note that the *U.num* indicates the number of usages in the table. When we say the request number we mean the maximum number of users request to each service.

Thus, the service satisfaction rating, the user total satisfaction rating from all services, only the regions satisfied and dissatisfied from all services, the interest of these regions, and the e-system was generally evaluated.

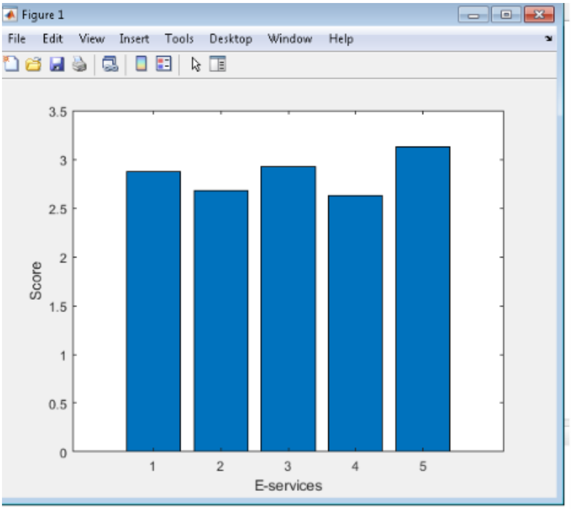

Figure 1. E-services satisfaction rating

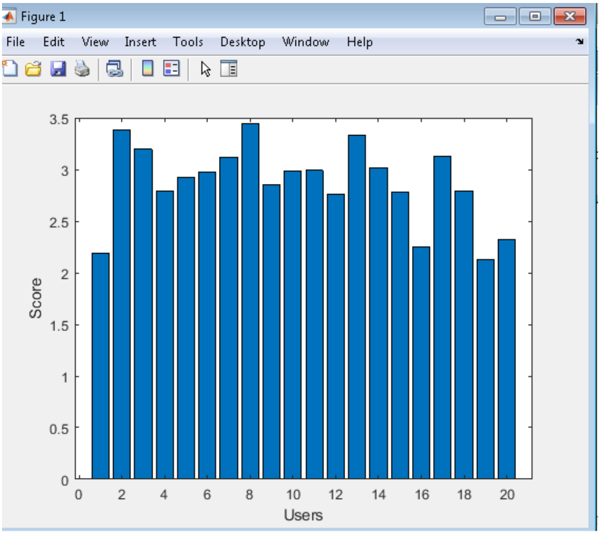

Figure 2. Users rating

Table 3

Hotspot services for regions

| Services | *Region 1* | | *Region 2* | | *Region 3* | | *Region 4* | | *Region 5* | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *U.num* | *Rank* | *U.num* | *Rank* | *U.num* | *Rank* | *U.num* | *Rank* | *U.num* | *Rank* |
| $EG_1$ | 46 | *3* | 48 | *1* | 49 | *1* | 49 | *1* | 46 | *3* |
| $EG_2$ | 37 | *4* | 48 | *2* | 36 | *4* | 47 | *2* | 41 | *4* |
| $EG_3$ | 22 | *5* | 36 | *4* | 34 | *5* | 40 | *4* | 48 | *1* |
| $EG_4$ | 47 | *2* | 38 | *3* | 42 | *2* | 41 | *3* | 48 | *2* |
| $EG_5$ | 47 | *1* | 19 | *5* | 38 | *3* | 29 | *5* | 29 | *5* |

As is known, citizens can show their attitude to any service by commenting on in the e-government environment. Analyzing these comments, it is possible to identify the main issues annoying them. It is known that as the number of comments increases, it becomes more difficult to analyze them. It is important to apply text mining methods to quickly identify the main points of the citizens is concerned. Taking this into account, an approach to determine the hot topics of citizens' in e-government through the analysing of citizens' comments has been proposed in the **second section of the fourth chapter**. Latent Dirichlet Allocation (LDA) and k-means algorithms has been used in this approach [12, 14]. The main steps of the proposed approach are illustrated in Fig. 3.

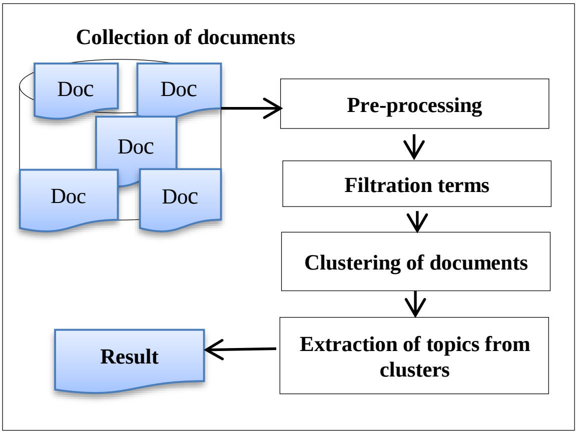

Figure 3. Schemes of the proposed approach

Each step of the proposed approach is described in detail below:

**Step 1.** Firstly, user comments are collected in the e-government environment. For simplicity, these comments are treated as documents and are signified as follows:

$$D = \{d_1, d_2, ..., d_n\}$$

where, $n$ – is the number of documents (comments).

**Step 2.** The collected comments are pre-processed. In pre-processing, common words, figures and punctuation marks are extracted from the documents. Each word is converted to its original (root) form as they take affixes in their different forms.

**Step 3.** The terms are extracted from the comments. Then, the sets of documents are described as a vector using the "Term Frequency-Inverse Document (TF-IDF)" .

Euclidean distance is used to calculate the distance between documents.

Known that, the number of terms in the set of documents is too high and this number is greater than the number of terms found in a single document. Then, most vector elements represented by the TF-IDF of documents will be "0". In other words, vectors will be sparse. This creates two important problems in document clustering:

- "Cursed" measurement problem;
- Quality of clustering.

Sparse terms are pre-removed from the vector to overcome these problems. After removing the sparse terms, another factor emerges that affects to the problems represented above. The reason for this is the existence of synonyms in the documents. If the set of documents contains a lot of synonyms, then documents with similar contents may fall into different groups in clustering. This leads to a decreasing in quality of clustering. To overcome such situations, it is suggested to find and extract semantic similar words from the sets of documents, to keep one of them and to exclude the others. The usage of extended sets of synonyms of each term is suggested to find the semantic similarity of words. For this purpose, we find the set of synonyms of each term using the WordNet and they are signified by $t_i \rightarrow synset\ (t_i)$ . Note that

WordNet is a network that provides you to determine semantic relationships between words. For example, synonyms, hypernyms, hyponyms, etc. can be easily detected via this network

After finding sets of the extended synonyms of each term, the semantic similarity between words is calculated using the following metric:

$$sim(t_g, t_s) = \frac{2\left|synset\,(t_g) \cap synset\,(t_s)\right|}{\left|synset\,(t_g) \cup synset\,(t_s)\right|} \geq \alpha, \quad g, s = 1,2,...,m$$

where $|synset\,(t)|$ – is the number of synonyms of the word $t$, $0 \leq \alpha \leq 1$ – is a managed parameter. If the similarity between words is greater than $\alpha$ , these words are considered as an unique term. Thus, only one of these words is maintained and the others are omitted. So, we reduce the measure of the vector $d$. In this case, the vector $d_i$ is transformed into the following vector:

$$d_i \rightarrow d_i^* = \{\overline{w}_{i1}, \overline{w}_{i2}, ..., \overline{w}_{im_0}\}, \quad m_0 \leq m$$

where $\overline{w}_{ij}$ – is TF-IDF weight of $j$- th word in the $i$- th document after removing the synonyms.

**Step 4.** Documents are clustered after displaying as vectors. Various methods exist for document clustering. In this paper, we propose the use of the k-means method for document clustering. k-means is one of the popular algorithms in big data analysis due to its short execution time and ease of application.

**Step 5.** After the clustering of documents, we can find topics for each cluster. For this purpose, the usage of LDA is proposed. The previous sections provide detailed information about LDA. The extraction of main topics from the documents for each cluster via LDA is implemented as follows.

Let's assume that clusters $\{C_1, C_2, ..., C_k\}$ are selected. The LDA algorithm is applied to each cluster and for each $C_q$ cluster, $T_q = \{T_{q1}, T_{q2}, ..., T_{qs}\}$ topics are assigned. Here $s$ – is the number of topics. Thus, we identify the main topic of citizens' comments.

The experiments were conducted in the R programming language. The BBC News database was used for the experiment. This dataset contains 2225 documents collected from the BBC news website in five associated areas: Business, Entertainment, Politics, Sports and Technology covering 2004-2005. In the experiment, a large number of documents from Business, Entertainment, and Sports were collected and analyzed. The criterion of "purity coefficient" was used to evaluate the clustering results.

Pre-processing is one of the key steps in text mining. Considering this, the documents collected during the experiment were pre-processed. In pre-processing, punctuation marks, figures, symbols, common words were removed from the sets of documents, and they were represented in vector form using the TF-IDF scheme. Then, the documents were cleared from sparse terms.

The number of words remaining in the sets of documents before and after the pre-processing is described in Table 4.

Table 4

Number of documents and words

| **Number of documents** | **Number of words** | |
|---|---|---|
| | **Before the pre-processing** | **After the pre-processing** |
| **100** | 8040 | 4421 |
| **300** | 18356 | 8851 |
| **500** | 25490 | 11766 |
| **800** | 33410 | 14750 |
| **1000** | 36346 | 15860 |

After pre-processing, the proposed method has been applied to the sets of documents. The semantic similarity between words was calculated at different values $(0.1, 0.2, 0.3, 0.4, 0.5)$ of $\alpha$. The number of terms remaining in the sets of documents after the method is described in Table 5. In the version where the number of documents is 100, we observe that more semantic similar words were found at $\alpha = 0.1$ value and the vector measure decreased significantly (26.42%) compared to the remaining words after removing the sparse terms. Since the $\alpha -$ th value is increased fewer words were omitted. Thus, at the value of $\alpha = 0.5$, we notice that the vector measure gets more less (1.29%). As the number of documents increases, the number of semantically similar words also increases accordingly, and the vector measure significantly decreases. For example, if we consider the number of documents with 800, we observe that the vector measure decreases by 31.67% at $\alpha = 0.1$ value.

Table 5

The number of words remaining in the documents after the removal of sparse and semantically similar words.

| $\alpha =$ | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
|---|---|---|---|---|---|---|
| **Number of documents** | **Number of words** | | | | | |
| | **After removing the sparse terms** | **After the removal of semantically similar words** | | | | |
| 100 | 772 | 568<br>(26.4%) | 691<br>(10.4%) | 733<br>(5.05%) | 756<br>(2.07%) | 762<br>(1.29%) |
| 300 | 878 | 633<br>(27.90%) | 777<br>(11.50%) | 838<br>(4.56%) | 856<br>(2.51%) | 865<br>(1.48%) |
| 500 | 827 | 589<br>(28.77%) | 725<br>(12.33%) | 789<br>(4.59%) | 809<br>(2.17%) | 816<br>(1.33%) |
| 800 | 821 | 561<br>(31.67%) | 716<br>(12.79%) | 778<br>(5.23%) | 802<br>(2.31%) | 810<br>(1.34%) |
| 1000 | 801 | 561<br>(29.96%) | 702<br>(12.36%) | 765<br>(4.49%) | 784<br>(2.12%) | 791<br>(1.25%) |

Then, the k-means clustering method was applied to sets of documents, and the clustering accuracy is illustrated in Table 6. Note that the value of $\alpha = 0$ indicates that sets of the remaining terms after removing the sparse terms. As seen from the table, the removal of semantically similar words did not negatively affect the quality of the clustering, but rather, the purity coefficient got sufficiently high value. As seen from the table, the purity coefficient gets significantly high value as the number of documents increases.

Table 6

Evaluation of the purity coefficient of clustering at different values of $\alpha$

| $\alpha =$ | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
|---|---|---|---|---|---|---|
| **Number of documents** | **Purity** | | | | | |
| 100 | 0.95 | 0.88 | 0.81 | 0.82 | 0.82 | 0.88 |
| 300 | 0.996 | 0.98 | 0.98 | 0.98 | 0.98 | 0.98 |
| 500 | 0.996 | 0.93 | 0.94 | 0.99 | 0.99 | 0.99 |
| 800 | 0.98 | 0.81 | 0.89 | 0.99 | 0.99 | 0.99 |
| 1000 | 0.982 | 0.97 | 0.98 | 0.97 | 0.98 | 0.98 |

Table 7

The top 10 words on clusters ($\alpha = 0$)

| **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|
| music | ireland | cluster |
| award | england | growth |
| people | play | year |
| show | win | rate |
| year | wale | economi |
| won | side | bank |
| radio | game | econom |
| veto | beat | oil |
| years | nation | price |
| song | scotland | rise |

Table 8

The top 10 words on clusters ($\alpha = 0.3$)

| **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|
| film | england | cluster |
| star | play | team |
| year | cluster | year |
| award | year | rate |
| role | rugbi | economi |
| cluster | player | rise |
| includ | game | price |
| director | season | bank |
| play | cup | econom |
| bbc | week | month |

After the clustering of documents, topic modeling method was used to extract topics from each cluster . The top10 words that  extracted from each cluster has been described in tables 7 and 8.

As can be seen, the topics were extracted accurately, and we saved in time. Thus, the following table describes the time spent on clustering and the extraction of topics from each cluster and their comparative analysis (Table 9).

Thus, the experiment and the results show that the proposed method can significantly reduce the size of a large number of documents, save time spent on data analysis and improve the quality of clustering, GDP algorithm.

Table 9

The time spending on the application of LDA and k-means

| $\alpha$ = | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** |
|---|---|---|---|---|---|---|
| **Number of documents** | **Time used** | | | | | |
| 100 | 11.6 | 9.3<br>(19.82%) | 9.86<br>(15%) | 10.32<br>(11.04%) | 10.65<br>(8.19%) | 11.06<br>(4.65%) |
| 300 | 12.39 | 9.08<br>(26.71%) | 9.74<br>(21.39%) | 10.68<br>(13.80%) | 11.26<br>(9.12%) | 11.51<br>(7.10%) |
| 500 | 18.7 | 11.35<br>(39.30%) | 12.87<br>(31.18%) | 13.28<br>(28.98%) | 14.21<br>(24.01%) | 14.98<br>(19.89%) |
| 800 | 22.28 | 13.49<br>(39.45%) | 14.59<br>(34.52%) | 15.76<br>(29.26%) | 16.57<br>(25.63%) | 17.85<br>(20.33%) |
| 1000 | 21.06 | 12.78<br>(39.31%) | 14.09<br>(33.09%) | 15.01<br>(28.72%) | 16.01<br>(23.97%) | 17.25<br>(18.09%) |

## RESULTS

1. The role of social networking and intellectual analysis of texts in e-government analysis of texts in e-government analysis was studied, as well as the requirements for e-government and approaches to its development, the main problems of e-government development and analysis were identified [4-6, 8];
2. A hybrid classification method has been proposed for detecting of the terrorism-related texts in e-government [2, 3];
3. A method has been proposed for filtering of the terrorism-related texts in e-government [9, 10];
4. A method and algorithm based on sentiment analysis technology has been proposed for the detecting and analysis of hidden social networks in e-government [1, 11];
5. A method has been proposed for automatic assessment of citizen satisfaction from e-government services [7, 12];
6. A method based on clustering and topic modeling technologies has been proposed for identifying hot topics that the citizens (including the regions) cared  in e-government [13, 14].

## THE FOLLOWING SCIENTIFIC WORKS ON DISSERTATION MATERIALS HAVE BEEN PUBLISHED:

1. Alıquliyev, R.M., Niftəliyeva, G.Y. E-dövlət mühitində gizli sosial şəbəkələrin aşkarlanmasi üçün yanaşma // İnformasiya təhlükəsizliyinin multidissiplinar problemləri” üzrə II respublika elmi-praktiki konfransı, – Bakı, –14 may 2015, –s.116-118.
2. Alıquliyev, R.M., Niftəliyeva, G.Y. E-dövlət mühitində terrorizmlə əlaqəli mətnlərin aşkarlanması metodu // İnformasiya təhlükəsizliyinin multidissiplinar problemləri” üzrə II respublika elmi-praktiki konfransı, – Bakı, – 14 may 2015, – s.111-115.
3. Aliguliyev, R.M., Niftaliyeva, G.Y. Detecting terrorism-related articles on the e-government using text-mining techniques // Problems of Information Technology,–2015, 6 (2), –p. 36-46.
4. Alıquliyev, R.M., Niftəliyeva, G.Y. E-dövlət sisteminin analizində Data mining texnologiyalarının tətbiq imkanları // “Big Data: imkanları, multidissiplinar problemləri və perspektivləri” I respublika elmi praktiki konfransı, –Bakı, 25 fevral 2016, – s. 81-84.
5. Alıquliyev, R.M., Niftəliyeva, G.Y. E-Dövlətin Big Data Mənbələri // “Big Data: imkanları, multidissiplinar problemləri və perspektivləri” I respublika elmi praktiki konfransı, – Bakı, 25 fevral 2016, – s. 78-80.
6. Alıquliyev, R.M. E-Dövlətin analizi texnologiyları: text mining və sosial şəbəkələr. Ekspress-informasiya. “İnformasiya Texnologiyaları seriyası” / R.M. Alıquliyev, G.Y. Niftəliyeva – Bakı: –2016. –78 s.
7. Aliguliyev, R.M., Niftaliyeva, G.Y. Hotspot Information of Public Opinion in E-Government // 10th IEEE International Conference on Application of Information and Communication Technologies (AICT2016), –Baku, –12-14 october, 2016, – p.645-646.
8. Aliguliyev, R.M., Niftaliyeva, G.Y. The current state, problems and perspectives of e-government analysis technologies // Problems of Information Technology, – 2017, 8 (1), –p. 53-63.

9. İskəndərli, G.Y. E-dövlətə kiber hücumlar və onlarla mübarizə üsulları haqqında // "İnformasiya təhlükəsizliyinin aktual multidissiplinar problemləri" üzrə IV respublika elmi-praktiki konfransı, – Bakı, – 14 dekabr 2018, – s.158-160.
10. Alguliyev, R. M., Aliguliyev, R. M. Niftaliyeva, G. Y. Filtration of Terrorism-Related Texts in the E-government Environment // International Journal of Cyber Warfare and Terrorism, –2018, 8 (4), p.35-48. **(Web of Science)**
11. Alguliyev, R. M., Aliguliyev, R. M., Niftaliyeva, G. Y. A Method for Social Network Extraction From E-Government // International Journal of Information Systems in the Service Sector, –2019, 11 (3), p.37-55. **(Web of Science)**
12. Iskandarli G.Y. Using Hotspot Information to Evaluate Citizen Satisfaction in E-Government: Hotspot Information // International Journal of Public Administration in the Digital Age, –2020, 7 (1), –p. 47-62. **(Web of Science)**
13. Iskandarli G.Y. Applying Clustering and Topic Modeling to Automatic Analysis of Citizens' Comments in E-Government // International Journal of Information Technology and Computer Science, – 2020, 12 (6), –p.1-10.
14. Iskandarli G.Y. Detecting the Main Topics of Citizens' Comments in e-Government / 2nd International Symposium on Applied Sciences and Engineering,– Atatürk University, – Erzurum, –Turkey, –7-9 april, 2021, –pp.581-584.

The defense will be held on **24 September** at **$14^{00}$** at the meeting of the Dissertation council ED 1.35 of Supreme Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Information Technology of the Azerbaijan National Academy of Sciences.

Address: AZ1141, Azerbaijan Republic, Baku, B.Vahabzade str., 9A

Dissertation is accessible at the library of the Institute of Information Technology of the ANAS.

Electronic versions of dissertation and its abstract are available on the official website of the Institute of Information Technology of the ANAS.

Abstract was sent to the required addresses on **22 July 2021.**

Signed for print: 16.07.2021

Paper format: $60 \times 80^{1}/_{16}$

Volume: 37347 characters

Number of hard copies: 20