Assessment of Society Based on the Analysis of the Behavior of Citizens on the Platform of Electronic Demography

2023 5th International Conference on Problems of Cybernetics and Informatics (PCI) | 979-8-3503-1906-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/PCI60110.2023.10326017

Irada Alakbarova
Institute of Information Technology
Baku,Azerbaijan
airada.09@gmail.com
0000-0002-9876-3035

***Abstract*—Due to the fact that documents collected in various registries (medical certificates, court documents, fines, tax payments, etc.) as well as demographic indicators create big data, it is impossible today to determine human behavior by conventional statistical or empirical methods. The proposed approach to assessing the behavior of citizens on the e-demography platform consists of five steps: a collection of documents, data pre-processing, description of the data, clustering of documents, and assessment of the behavior of citizens by conducting sentiment analysis. Such an approach can be critical in studying society, identifying social and economic problems, and ensuring transparency of relations between citizens and authorities in the e-government environment.**

Keywords—e-demography, human behavior analysis, sentiment analysis, cluster analysis

I. Introduction

The concept of electronicdemography (e-demography) considers the impact of information technology on demographic situations and allows the use of new sources of information for a deeper study of demographic processes. In this regard, the issue of assessing the behavior of citizens on the platform of e-demography and studying and predicting the processes taking place in society based on the knowledge gained is very relevant.

Human behavior is a key factor influencing both his financial situation and his reputation and trust (social capital) in society. Human behavior in the social sense is understood as a system of activity that reflects the implementation of human moral principles [1].

There are many different approaches to the analysis of human behavior [1, 2]. Documents and personal data reflecting the behavior of citizens can be obtained from various sources. Such sources include government registries, tracking devices, reference systems, mobile devices, social networks, the Internet of things, etc. To analyze human behavior, registries must first be identified, in which all demographic data and personal documents of everyone are collected [3].

II. Related Works

Text mining is one of the most common approaches in behavior analysis. By analyzing documents, letters, comments, and other textual information, one can judge a person's behavior, his mood, and vision of events. However, today text data is big data and is often stored in an unstructured form in various sources, which complicates their analysis [4].

The work [4] analyzes the behavior of users of social networks in order to identify terrorist groups. In this approach, user behavior is determined based on the comments they write. The authors propose a method for abstracting documents by clustering sentences. In the article [5], by clustering unstructured data of a text type, hidden social networks on Wikipedia that are engaged in information warfare and disinformation are revealed. To solve the problem, the algorithm of fuzzy c-means was used.

Independent Component Analysis (ICA) has a special place in behavior analysis. ICA is a systematic analysis of two or more independent variables (components), a statistical and computational model used to detect latent factors [6]. Using the ICA method, it is possible to reveal hidden factors in human behavior by analyzing personal data, which in the modern world constitutes big data. This method is widely used in medicine to analyze the behavior of patients.

Based on the sentiment analysis of comments written on social networks, it is possible to determine the behavior of users and their relationship with others. For example, in [7], they try to study human behavior based on the analysis of texts written by him. The purpose of the study is to determine the psychological attitude and understand and predict behavior through the intellectual analysis of text data. The study used ISI Web of Science, Engineering Village Compendex, ProQuest Dissertations, and Google Scholar databases as data sources. Analysis of unstructured texts on various topics is carried out by searching these databases by keywords.

III. Analysis of the Behavior of Citizens using Clustering and Sentiment Analysis

As we can see, there are many different approaches to analyzing human behavior. However, due to the expansion of cyber-physical systems and the emergence of the concept of "big data", there is a need to continue these studies to determine the individual's role in society. The proposed methodology for analyzing the behavior of citizens based on documents collected in various state registries consists of the following steps:

1) Collection of documents;

2) Preliminary processing;

3) Description of documents;

4) Clustering a set of documents;

5) Sentiment analysis of documents.

A. Collection of documents

To analyze the behavior of citizens on the platform of e-demography, first of all, you need to collect all possible information about citizens stored in different state registries under one identification number. To do this, first of all, personal data registries should be identified, in which all demographic data and personal documents of everyone are collected [8].

Registers of personal data exist in medicine, education, social insurance, banking systems and many other areas. These registries are an effective tool for collecting information about the population.

B. Preliminary processing

Words in a set of documents are analyzed morphologically. Common words are removed from the text. The main terms in the set of documents are defined. Word lemmatization can facilitate the parsing of text-type documents. This method more accurately determines the meaning of words and avoids discarding a large number of words during filtering [9].

Lemmatization translates a word into a basic form that expresses its main meaning, taking into account the context. It also groups words with the same root so that they can be analyzed as a single element. Lemmatization performs morphological analysis of words and is used to avoid unnecessary and time-consuming calculations during text processing [9]. There are various Python software packages for lemmatization. These include WordNet Lemmatizer, Spacy Lemmatizer, TextBlob, GensimLemmatizer and others.

C. Description of documents

At this stage, each document (in this case, each document d_n of the set D_i) is described as a vector. For this, the widely used Vector Space Model [10] is used. With this model, each document is displayed as a vector in n-dimensional Euclidean space. The set of documents related to personnel U_i can be formulated as follows:

$$D_i = (d_1, d_2, ..., d_n)$$

We need to group documents by topic (medical certificates, fines, requests, bank accounts, and others). Suppose all the different terms in the set D_i are represented as a set:

$$\mathbf{T} = \{t_1, ..., t_j\}$$

where j – is the number of terms. Each document is presented as: $d_i = [w_{i1}, \ldots, w_{ij}]$. w_{ij} – word weightt_j in documentD_i.

The *tf-idf* (*term frequency–inverse document frequency*) algorithm is used to calculate this weight. The tf-idf allows you to evaluate the importance of words in a document, determine the weight of words, etc. The algorithm is mainly used for parsing textual data. The word weight in document Diis defined as follows:

$$w_{ij} = \mathrm{IF}_{ij} \times \mathrm{IDF}_j \tag{1}$$

$i = 1,...,n$, $j = 1,...,m$, where IF_{ij} – frequency of use of the word t_j in the document:

$$\mathrm{IF}_{ij} = \frac{m_{ij}}{m_i}; \; i = 1,...,n, \; j = 1,...,m \tag{2}$$

where m_i – total number of documents. m_{ij} – number of terms t_j in document D_i.

The importance of terms is measured as follows:

$$\mathrm{IDF}_j = \log(n / n_j) \tag{3}$$

where n – total number of documents, n_j – number of documents containing the term t_j.

To determine the semantic proximity of documents, the Euclidean distance is used. Euclidean distance can be represented as the geometric distance between any two vectors in a multidimensional space [3].

According to this metric, the proximity of the vectors $D_i = [w_{i1}, \ldots, w_{im}]$ and $D_l = [w_{l1}, \ldots, w_{lm}]$ is calculated as follows:

$$d_{il} = \|D_i - D_l\| = \sqrt{\sum_{j=1}^{m} (w_{ij} - w_{lj})^2}, \; i,l = 1,...,n \tag{4}$$

In the next step, we cluster documents by topic, and then identify negative and positive documents in each cluster by performing a semantic analysis of the documents.

D. Clustering a set of documents

After pre-processing and description of documents, a clustering algorithm is applied. Clustering is a method of learning without a teacher and belongs to the methods of data mining [11]. In cluster analysis, a data set is divided into groups (clusters) depending on the characteristics of the data.

The model proposed in this work is based on the principle of precise clustering. As a result of clustering, documents are automatically divided into clusters by topic: tax invoices, insurance documents, bank accounts, traffic fines, medical, etc. By ranking the clusters, we can determine which topic has the most papers. After clustering, in order to identify problems in human behavior, an analysis of the tonality of these documents follows. Sentiment analysis is one of the most popular clustering methods [12]. By determining the sentiment of the documents in a cluster, we can determine whether these documents are more positive or negative.

E. Sentiment analysis of documents

There are many methods for semantic analysis [12]. The lexicon-based method is widely used in sentiment analysis and many researchers use approaches such as word embedding to bootstrap lexicons. The initial set of words is created using tools such as WordNet, HowNet, etc. [13]. Each document has a set of synsets:

$$w_i \in synset(w_i), \; i = 1,2,...,n \tag{5}$$

Documents can be submitted as follows:

$$D_i = D_i^+ \cup D_i^0 \cup D_i^-, i = 1,2,...,n \quad (6)$$

where D^- – documents that form a negative opinion about an individual (negative documents), D^+ – documents forming a positive opinion (positive documents), D^0–neutral documents.

In [14] a positive or negative document is determined by mathematical calculations of words using the Semantic Orientation CALculator. The vocabulary of subjectivity is used here, which means that the difference between weak and strong words expressing a thought is calculated. Weak words are rated on a scale of 2 (positive) or -2 (negative). Strong words are rated on a scale of 4 (positive) or -4 (negative). The approach uses the lexical database "Senti-WordNet".

The Senti-WordNet dictionary was created using WordNet and contains the "synset" structure. Using this experience, we determine the tone of each word based on the evaluation of the words. After determining the sentiment of the words in the document, we find the overall sentiment of the document using the following formula:

$$Score(T) = Sign(\sum_{w_i \in D} Score(w_i)) \quad (7)$$

where $Score(w_i)$ – the degree of tonality of the words found in document *D*.

If most of the documents in any cluster are negative, then citizens have problems in a certain area. For example, if most of the documents in the traffic cluster are negative (fines, warnings, etc.), it can be concluded from this that people do not follow the rules of the road. But there may be another reason as well. For example, a road sign is installed incorrectly, repairs are underway, etc. This means that the approach presented by us not only determines the behavior of citizens but also allows us to identify different types of problems.

IV. Conclusion

It is impossible to completely eliminate the socio-economic problems that arise in society through special rules and measures applied by the state. The most convenient approach to determining the behavior of citizens is the intellectual analysis of documents stored in various state registries. The study showed that the use of an e-demographic platform in assessing society and studying the behavior of citizens has more opportunities to obtain more accurate results. The disadvantage of this approach is that the algorithms used depend on the quantity and quality of training. Also, this approach does not take into account the sequence of words in the text. But, despite the shortcomings, this approach provides an effective solution to the problem of document clustering and makes it possible to simplify the object of study as much as possible, especially when analyzing unstructured, big data.

If the documents about any person stored in state registers are mostly positive, then we can talk about the high confidence of the state in this person. Also, the fact that most of the documents in the registries are positive indicates the correct management on the e-government. The geographic relief of a society can be determined based on an analysis of citizen behavior on an electronic demographic platform. That is, by assessing citizens by their behavior and position in society, it is possible to create a visual portrait of society.

References

[1] Selvithi D. FPGA Based Human Fatigue and Drowsiness Detection System Using Deep Neural Network for Vehicle Drivers in Road Accident Avoidance System // Learning and Analytics in Intelligent Systems, 2019, vol. 6, pp. 69–91 (in Norway).

[2] Tripathy A., Agrawal A., Rath S.K. Classification of sentiment reviews using n-gram machine learning approach // Expert Systems with Applications, 2016,vol. 57,pp. 117–126 (in USA).

[3] Lyngstad T.H., Skardhamar T. Nordic register data and their untapped potential for criminological knowledge // Crime and Justice, 2011, vol. 40, no. 1, pp. 613–645 (in USA).

[4] 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. vol. 8,issue 4, pp. 35–48 (in USA).

[5] Alguliyev R.M., Aliguliyev R.M., Alakbarova I.Y. Extraction of hidden social networks from wiki-environment involved in information conflict. International // Journal of Intelligent Systems and Applications, 2016, vol. 8,issue 2, pp. 20–27 (in Hong Kong).

[6] Benjamin S.R., Markelz A.M., Ruiz S. The Nature and Extent of Component Analyses for Improving or Mitigating Behavior: A Systematic Review // Indexing & Metrics, 2020, vol. 46, issue 1,pp. 230–253 (in USA).

[7] Gutierrez E., Karwowski W., Fiok K., Davahli M.R., Liciaga T., Ahram T. Analysis of Human Behavior by Mining Textual Data: Current Research Topics and Analytical Techniques // Symmetry, 2021, vol.13. issue 1276, pp. 1–22 (in Spanish).

[8] Pikulík T. , Štarchoň P. Public registers with personal data under scrutiny of DPA regulators // Procedia Computer Science,2020, vol. 170,pp. 1174–1179 (in Holland).

[9] Khyani D., Siddhartha B.S. An Interpretation of Lemmatization and Stemming in Natural Language Processing // Journal of University of Shanghai for Science and Technology, 2021, vol. 22, issue 10, pp. 350–357 (in Shanghai).

[10] Zhang C., Huang W., Niu T., Liu Z., Li G., Cao D. Review of Clustering Technology and Its Application in Coordinating Vehicle Subsystems // Automotive Innovation,2023,vol. 6, pp. 89–115 (in China).

[11] Manikandan S., Lourdu C.A., Kanniamma D. The Study on Clustering Analysis in DataMining. International // Journal of Data Mining Techniques and Applications, 2018, vol. 07, issue 01, pp. 46–49 (in Chennai, India).

[12] Pang B., Lee L. Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval, 2008,vol. 2, no.1-2, pp. 1–135 (in USA).

[13] Mäntylä M.V., Novielli N., Lanubile F., Claes M., Kuutila M. Bootstrapping a lexicon for emotional arousal in software engineering / In: Proceedings of the 14th International Conference on Mining Software Repositories, Buenos Aires, Argentina, 20–21 May 2017, pp. 198–202. (in Argentina).

[14] Taboada M., Brooke J., Tofiloski M., Voll K.D., Stede M. Lexicon-Based Methods for Sentiment Analysis // Computational Linguistics, 2011, vol.37, issue 2, pp. 267–307 (in USA).