Application of SAS Text Miner for the analysis of citizens' appeals in the system of social protection and social security*

Józef Korbicz[1] , Oleksii Sholokhov[2,*], Roman Koval[3], Oleksii Zarudnyi[3]

[1] *University of Zielona Góra, 9 Licealna Street, Zielona Góra, 65-417, Republic of Poland*

[2] *Taras Shevchenko National University of Kyiv, 64/13 Volodymyrska Street, Kyiv, 01601, Ukraine*

[3] *Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, 13 Chokolovsky Blvd., Kyiv, 03186, Ukraine*

Abstract

Issues of social protection and social security have always been among the most urgent for all, without exception, social strata. In the conditions of the war, this sphere acquired special importance. After all, the effectiveness of the state policy of social protection and social security depends not only on the well-being of citizens and the balanced development of society, but also on ensuring national security. During the war, the amount of spending on social protection and social security increased significantly and will continue to increase, despite the limited budgetary funding. Therefore, special attention needs to be paid to the targeting of funds for social protection and social security, as well as control over the targeting of state assistance. In the conditions of war, conducting sociological research, surveys, and personal reception of citizens becomes much more difficult. Taking into account the fact that a significant number of the population uses various social networks, digital platforms of state institutions and organizations, etc., the research of the online environment becomes a promising direction of work with citizens' appeals. Therefore, having information from Internet sources, it is possible to investigate problems that are significant for different social groups, to analyze the moods and expectations of the population. But at present, there are practically no software products in the social security system designed to analyze textual information presented in citizens' appeals.

The work proposes a method of building an analytical model for the study of social protection and social security problems that require special attention from the state, using means of analyzing textual information from Internet sources and building classification models.

Keywords

Text clustering, linguistic rules, intelligent data analysis, social protection and social security, information technology

1. Problems of automation and processing of citizens' appeals in the social sphere

Information and analytical activity in the conditions of deepening digitalization of society is becoming an increasingly important component of the system of social protection and social security, which in turn, as noted by domestic and foreign experts [14-16], requires its constant modernization, introduction of modern models, methods and information technology. The introduction of the "Unified Information System of the Social Sphere" [17] was a new step towards the end-to-end digitalization of the pension system and social protection of the population. The purpose of the introduction of the System is to "ensure integral automation of processes in the social

8th International Scientific and Practical Conference Applied Information Systems and Technologies in the Digital Society AISTDS'2024, 2024, October 1, Kyiv, Ukraine

* Corresponding author.

J.Korbicz@issi.uz.zgora.pl (J. Korbicz) gyroalex@knu.ua (O. Sholokhov); roman.koval.science@gmail.com (R. Koval); oleksii.zarudnyi@gmail.com (O. Zarudnyi)

2338-9598-800 (J. Korbicz); 0000-0002-8676-3724 (O. Sholokhov); 0009-0003-3821-3378 (R. Koval); 0009-0008-7462-3899 (O. Zarudnyi)

sphere by optimizing and developing electronic information interaction of the subjects of the Unified System aimed at ensuring transparency of the social sphere, digitalization of the social support market and increasing the level of its availability for persons who need it" [17].

The development of the Unified Information System of the Social Sphere [1] involves the creation of a unified information and reference environment for recipients of social support. An important place is occupied by the subsystem of working with citizens' appeals, because only in January-September 2024, the Pension Fund of Ukraine registered 504,856 appeals from citizens on issues, of which 229,537 (or 45.5 percent) were electronic appeals [2].

Therefore, the issue of developing methods, models, information technologies for the analysis of textual information from citizens' electronic appeals to institutions of social protection and social security, Internet sources, identifying issues that are most important for those who need state support, is urgent and of practical importance. [18-20].

2. Statement of the research problem

The paper proposes a method of using text analytics tools to build an analytical model for the classification of text information in the task of analyzing citizens' appeals to the Pension Fund of Ukraine.

3. Methods and results

In the course of the study, the practical task of determining the need for social protection and social security of residents of different regions of Ukraine and refugees was considered. SAS Text Miner tools [21-23] were used to analyze text information.

Incoming information is electronic appeals from citizens that have arrived at the web portal of electronic services of the Pension Fund of Ukraine and the state institution "Government Contact Center [2]. The materials of Internet publications, different in subject matter and audience, both state and non-state, were also examined, from which 162 were selected (names of sources and references to them are presented in Table 1.

Table 1
List of Internet sources, information from which was used for analysis

N	Name of the source	Resource address	Texts number
1	UkrInform	https://www.ukrinform.ua/rubric-society	50
2	Public. News	https://suspilne.media	25
3	Website of the international scientific publication "Financial and credit activity: problems of theory and practice"	https://fkd.net.ua	7
4	The newspaper "Government Courier" is the official printed publication of the Cabinet of Ministers of Ukraine.	https://ukurier.gov.ua/uk/articles	30
5	The official website of the Kyiv Regional Council of Professional Unions	http://korps.com.ua	5
6	Official website of the National Bank of Ukraine	https://knpf.bank.gov.ua	10
7	The official site of the magazine "Forbes Ukraine"	https://forbes.ua	15
8	Website of the electronic publication "Sudovo-yuridychna Gazeta"	https://sud.ua	20

Based on the analysis of texts related to issues of social protection and social security posted on the specified Internet resources and in electronic applications, six clusters were obtained.

The first cluster includes texts that contain issues related to the pension reform. The most characteristic words and phrases for this cluster were: "reform", "insurance payments", "insurance experience", "mandatory pension savings".

The second cluster includes words and phrases describing the issue of accrual and payment of pensions and social benefits by the Pension Fund of Ukraine: "timely payment of pensions", "voluntary contributions to pension insurance", "minimum pension", "indexation of pensions", "increase of pensions", "housing subsidy", "financing of current payments", "recalculation of pensions for working pensioners".

The third cluster summarizes the problems of social protection of internally displaced persons. The most characteristic are such words and phrases as "IDPs", "identification", "liberated territories", "payments to displaced persons", "inhabitants of the occupied Crimea", "UN World Food Program", "temporarily uncontrolled territories".

The fourth cluster includes words and phrases describing problems related to losses due to military conflict: "military serviceman", "policeman", "combat zone", "missing person", "loss of breadwinner", "family members of the deceased" ".

For the fifth cluster, the issues of social protection and social security of refugees are "relevant", in particular, "pension abroad", "work outside Ukraine", "proportional calculation of insurance experience", "insurance experience received in other countries".

The sixth cluster summarizes issues related to the victims of the accident at the Chernobyl NPP: "accident", "ChNPP", "Chernobyl".

Based on the preliminary analysis of the texts of the appeals, a corpus of texts was formed, a fragment of which is given in the table. 2.

Table 2
Frequency matrix of terms for the corpus of texts, built on the basis of the corpus of texts formed from electronic appeals of citizens

Marking	Term	Number of mentions in the document:									
		d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
t1	court	1	0	0	0	0	0	1	2	0	0
t2	allowances	1	0	1	1	0	0	1	0	2	0
t3	military	0	1	0	0	2	1	0	0	0	0
t4	monetary support	0	1	0	0	1	0	0	0	2	0
t5	pension	0	1	0	1	2	2	1	0	1	1
t6	law enforcement officers	0	1	0	0	1	0	0	0	0	0
t7	the former	0	1	0	0	1	0	0	0	0	0
t8	accident	0	0	1	1	0	0	0	0	0	1
t9	Chernobyl Nuclear Power Plant	0	0	1	2	0	0	0	0	0	1
t10	Ukraine	0	0	1	0	0	1	0	0	0	0
t11	received	0	0	0	1	1	0	0	0	0	0
t12	service	0	0	0	0	1	1	0	0	0	0

To solve the problem of reducing the dimensionality and sparsity of the frequency matrix of the corpus of texts, the method of singular distribution (SVD) was used [3-5]. After all, documents usually use a fairly small set of terms that describe a certain subject area. Therefore, if in the diagonal matrix of singular values (S) we leave exactly k of the first diagonal elements, and assign the value zero to the rest, then the use of the SVD method gives an optimal approximation. In the diagonal matrix of singular values S, the values are ordered, namely, $s_1 \geq s_2 \geq \dots \geq s_k$, that is, if you leave the first two values, then assign the value zero to the others. On the basis of the obtained matrix S, it is possible to calculate the percentage contribution of the dimension described by the corresponding singular value to the explanation of the data.

On the basis of the obtained matrix S, it is possible to calculate the amount in percent that the corresponding dimension, which is described by the corresponding singular value, contributes to the explanation of the data (table 3). The value of the column "Percentage of value contribution to the explanation of data variability" is calculated as the value of "Square of the singular value" divided by the sum of the values of the squares of the singular values, multiplied by 100%.

As can be seen from the obtained results, table 3, if only the two basic dimensions are left, a total of 66.16% of the data variability will be explained.

Table 3
Analysis of the obtained singular values

Measurement number	Singular value	Singular value square	The percentage of value contribution to the explanation of data variability	Cumulative value of deposit interest
1	5.1435	26.45	45.61	45.61
2	3.4526	11.92	20.55	66.16
3	2.7696	7.67	13.23	79.38
4	2.3736	5.63	9.71	89,11
5	1.7711	3.13	5.41	94.51
6	1.2251	1.5008	2.58	97.09
7	1,029	1.0588	1.82	98.92
8	0.684	0.4678	0.81	99.73
9	0.371	0.1376	0.23	99.96
10	0.1352	0.0182	0.03	100

In this case, all documents can be located in two-dimensional space and determine the clusters that they form according to the degree of similarity and belonging to a certain topic (Fig. 1).

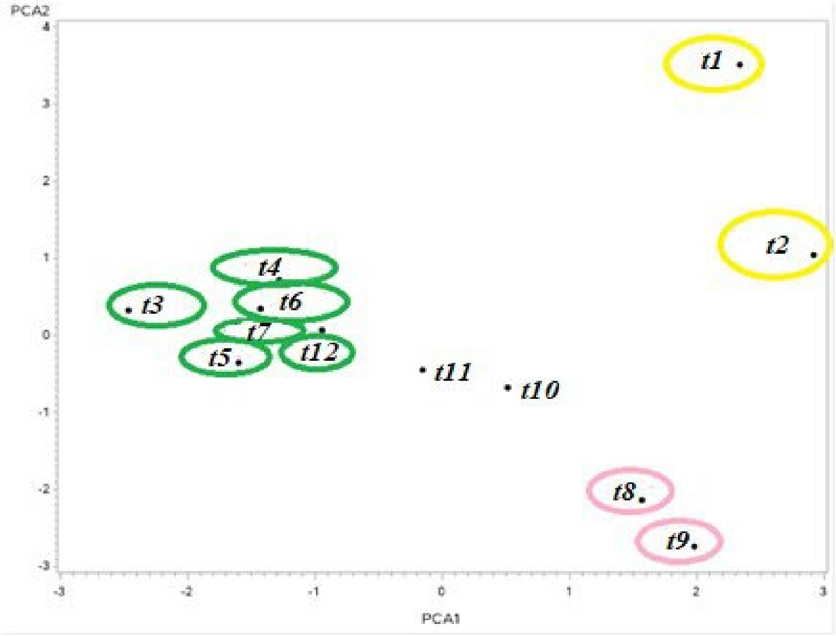


Figure 1: Location of terms in two-dimensional space.

As can be seen from fig. 1, the first dimension explains 45.61% of the data variability; the second dimension explains 20.55% of the data variability. As a result, three thematic clusters were formed, which included documents based on the similarity of the use of terms [6-9].

The SAS Text Miner system was used in this study. When using the SAS Text Miner software, a technological project is built in which the following steps are performed:

1. Loading data.
2. Text parsing.
3. Text filtering.
4. Text clustering.

The technological process of analyzing the corpus of texts for the purpose of their clustering is presented in fig. 2.

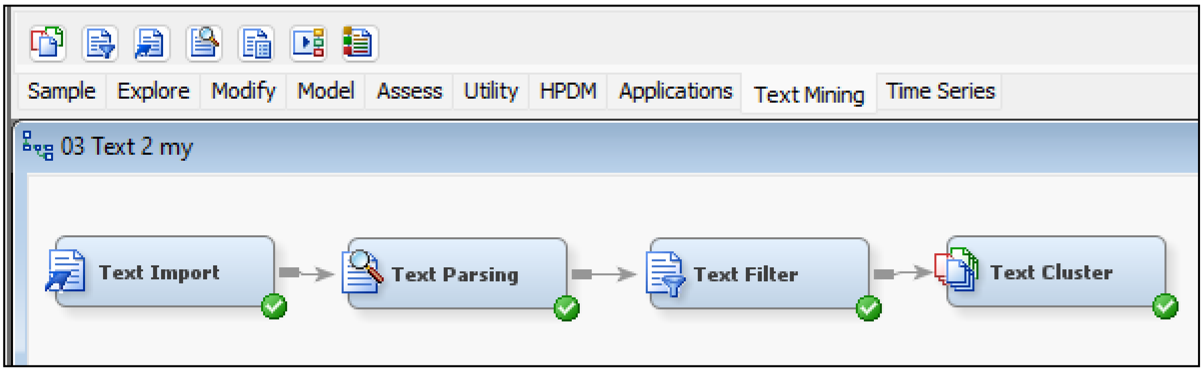


Figure 2: Technological process of text corpus analysis in the SAS Text Miner system.

The constructed rules for the corresponding clusters are generated in the form of the following program code:

```
F_TextCluster_cluster_ =1 ::
(OR
, "reform"
, "insurance"
, (AND, (OR, "payments", "seniority") )
, "accumulation"
, (AND, (OR, "pensionable", "mandatory") )

F_TextCluster_cluster_ =2 ::
(OR
, "voluntary"
, (AND, (OR, "payments" , "pension"))
, "timely"
, (AND, (OR, "contributions" , "pension" , "insurance", "recalculation"))
, "pension"
, (AND, (OR, "minimum" , "index" , "increment"))
, "subsidy"
, (AND, (OR, "residential"))
, "current"
, (AND, (OR, "payment" , "funding"))

F_TextCluster_cluster_ =3 ::
(OR
, "identification"
, (AND, (OR, "refugee" , "displaced person". "payments"))
, "resident"
, (AND, (OR, "Crimea" , "uncontrolled" , "territory" , "temporary"))
, "UN"
, (AND, (OR, "global" , "food" , "program"))

F_TextCluster_cluster_ =4 ::
(OR
, (AND, (OR, "serviceman" , "military", "policeman"))
, "zone"
, (AND, (OR, "combat" , "actions"))
, (AND, (OR, "missing" , "missing"))
, "deceased"
, (AND, (OR, "loss" , "breadwinner" , "members" , "family"))

F_TextCluster_cluster_ =5 ::
(OR
, "pension"
, (AND, (OR, "border", "borders", "others", "countries"))
, "experience"
, (AND, (OR, "calculation" , "insurance" , "proportional"))

F_TextCluster_cluster_ =6 ::
(OR
, "accident"
, (AND, (OR, "CHAES" , "nuclear" , "power plant"))
, "Chernobyl"))))
```

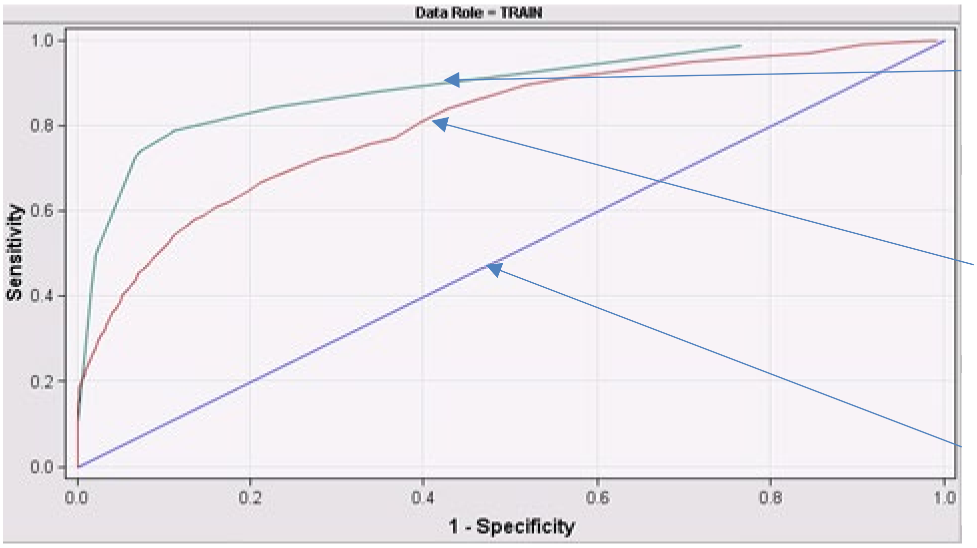
The statistical characteristics of the built classification model based on linguistic rules were calculated separately for the training and test data sets: the ratio is 70% for training and 30% for testing, i.e. 114 and 48 texts, respectively.

The results are summarized in Table 3.

Table 3
Statistical characteristics of the classification model of the studied texts

Statistics	Data set	
	training	Test
TP (True Positive)	30	11
TN (True Negative)	67	26
FP (false positive)	10	6
FN (false negative)	7	5
MISC,% (proportion of incorrectly classified values)	15	23
Ginny	0.82	0.71
ROC	0.79	0.67

The image of the ROC curve for the text information classification model based on linguistic rules is presented in Fig. 3.



ROC-characteristics of the model on the training set

ROC-characteristics of the model on the test set

The reference line is 50 for 50 percent of the occurrence of the event

Figure 3: ROC curve for the built classification model based on linguistic rules.

The constructed linguistic rules were used to cluster news texts that were published on the Internet from September 2023 to September 2024. In general, about 10,000 tons were unloaded and processed. texts on social protection and social security of Ukrainians.

After clustering the texts, the number of texts belonging to contributors from a certain region was calculated for each cluster. The obtained values were normalized on a scale from 0 to 100 according to formula (1):

$$popularity_i = \frac{n_i}{\max(n_i \,|\forall\, i)}, \tag{1}$$

where $popularity_i$ – the popularity of the texts of the corresponding cluster for the i-th region, n_i – the number of texts by region, $\max(n_i \,|\forall\, i)$ – maximum number texts by all regions.

The results of the calculations are presented in Table 4.

Table 4
Results of cluster analysis of textual information on issues of social protection and social security by regions of Ukraine

Name of the region	Popularity of the texts of the corresponding cluster					
	Cluster 1 (pension reform)	Cluster 2 (accrual and payment of pensions and social benefits by the Pension Fund of Ukraine)	Cluster 3 (problems of social protection of internally displaced persons)	Cluster 4 (issues related to losses due to military conflict)	Cluster 5 (issues of social protection and social security of refugees)	Cluster 6 (issues related to victims of the accident at the Chernobyl nuclear power plant)
Vinnytsia region	94	65	24	72	79	
Volyn region	87	57	20	100	63	
the city of Kyiv	82	49	32	37	26	33
the city of Sevastopol	-	-	-	-	-	-
Dnipropetrovsk region	58	39	43	33	14	
Donetsk region	27	32	59	37		
Zhytomyr region	94	73	19	34	62	88
Transcarpathian region	67	45	29	40	75	
Zaporizhzhia region	58	39	90	30		
Ivano-Frankivsk region	87	66	24	63	72	
Kyiv region	84	42	28	37	37	100
Kirovohrad region	92	88	32	73	46	
Autonomous Republic of Crimea	-	1	1	-	-	-
Luhansk region		22	8			
Lviv region	73	45	20	60	57	
Mykolayiv region	76	70	64	47	18	
Odesa region	32	24	27	13	13	
Poltava region	75	63	32	73	42	77
Rivne region	100	64	17	81	100	
Sumy region	92	100	52	43	30	
Ternopil region	50	56	24		63	
Kharkiv region	47	35	100	15	9	
Kherson region	71	62	89			
Khmelnytskyi region	87	55	28	78	73	
Cherkasy region	87	47	30	50	55	74
Chernihiv region	81	58	24	50	29	
Chernivtsi region	83	31	25	1	61	

The results of the analysis presented in the table can be visualized using SAS tools Enterprise Guide 7.1 (fig. 4-9).

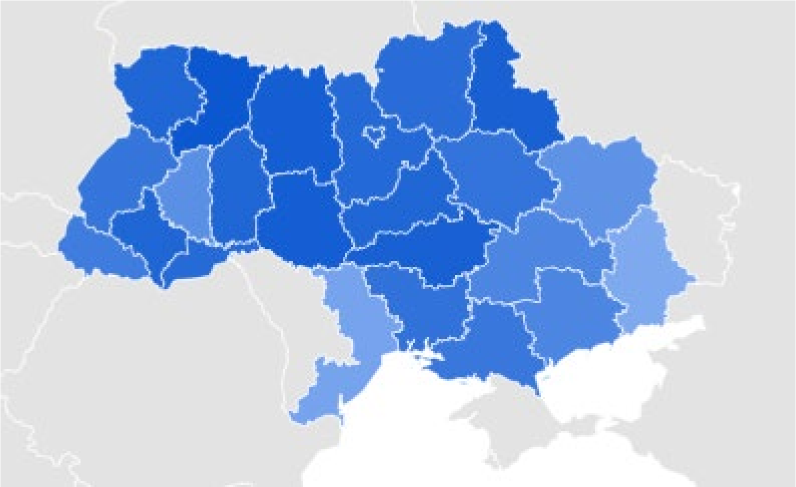


Figure 4: Cluster 1 - popularity of texts on " Pension reform " by regions of Ukraine.

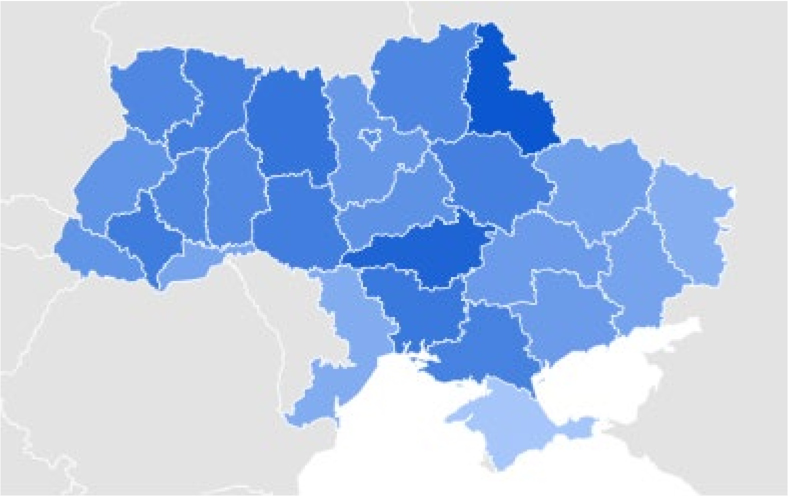


Figure 5: Cluster 2 - the popularity of texts on the topic "Questions related to the pension fund in general" by regions of Ukraine.

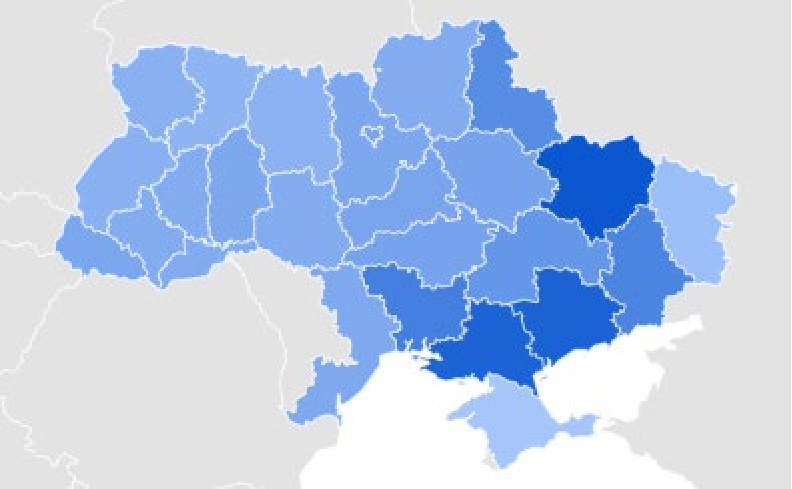
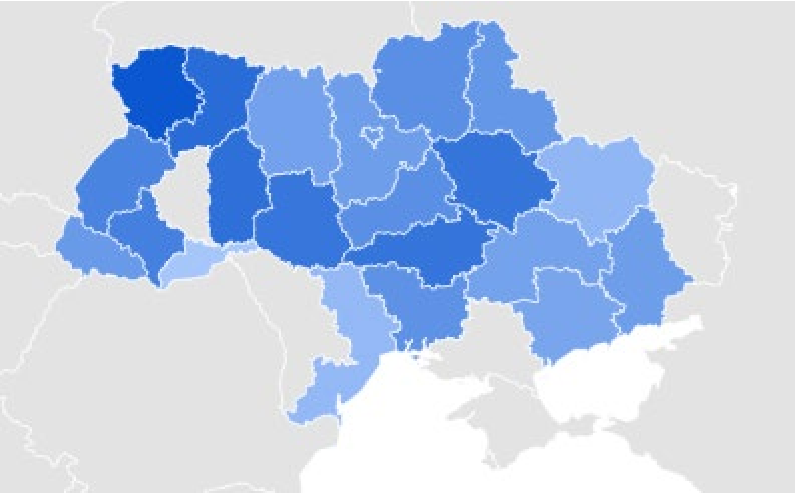


Figure 6: Cluster 3 - popularity of texts on the topic "Problems related to IDPs" by regions of Ukraine.



.

Figure 7: Cluster 4 - the popularity of texts on the topic "Issues related to the military and police" by regions of Ukraine.

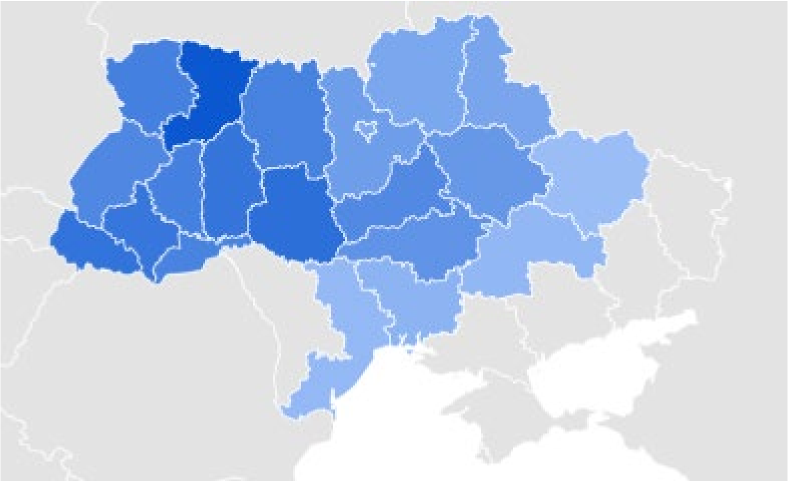


Figure 8: Cluster 5 - the popularity of texts on the topic "Questions regarding the payment of pensions abroad" by regions of Ukraine.

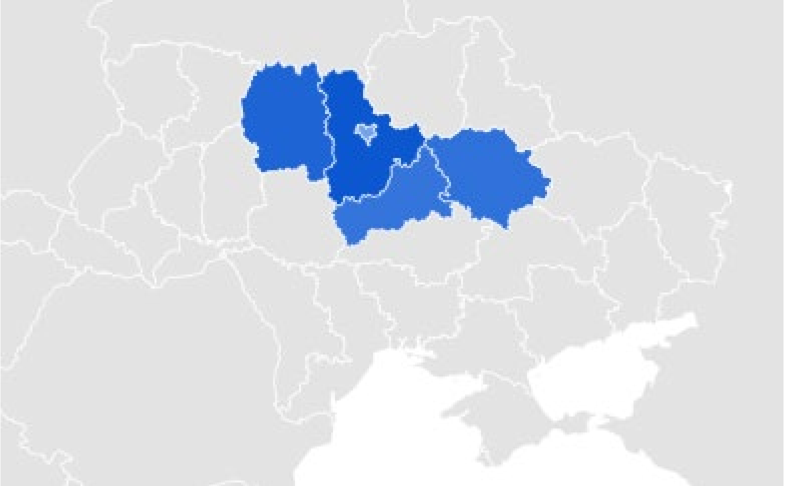


Figure 9: Cluster 6 - the popularity of texts on the topic "Issues related to pensions for victims of the accident at the ChAES" by regions of Ukraine.

4. Declaration on Generative AI

The authors have not employed any Generative AI tools.

5. Conclusion

The proposed method of textual information analysis using text tools mining designed for automated processing of large volumes of texts on a certain topic. The use of text analytics allows you to deepen your knowledge of the subject area by using unstructured data. In this study, the problem of dimensionality and sparsity of the frequency matrix of the corpus of texts is solved using the key theorem of linear algebra - the singular matrix decomposition (SVD) method. Pre-executed. frequency weighting operation, which helped to partially solve the problem of unevenness of high-frequency terms, making them less influential. This made it possible to obtain results of classification of textual information of high quality.

Therefore, the use of intellectual analysis of large volumes of textual data allows to identify the most important problems that require a priority solution, to find out for which categories of the population they are most relevant. The obtained results can be further used during the planning of social expenditures of budgets of different levels, in the model of actuarial calculations, during the planning of social expenditures of budgets of various levels. The proposed approach can improve the quality of forecasts in modern conditions, when there is no complete information about the investigated process or phenomenon or the information is distorted.

References

[1] Shapovalova T. The concept and content of social protection and social security of the population in modern Ukraine. Economic analysis. 2022. Volume 32. No. 3. P. 123-130. https://doi.org/10.35774/econa2022.03.123 (ukr)

[2] Gren T. I. Peculiarities of implementation of the policy of social protection of territories in war conditions. Academic notes of TNU named after V.I. Vernadskyi. Series: Public management and administration. 2022. Volume 33 (72) No. 6. P. 81-84. https://doi.org/10.32782/TNU-2663-6468/2022.6/13 (ukr)

[3] Expenditures on social assistance. URL: https://mof.gov.ua/uk/expenditures_on_social_assistance (ukr)

[4] Smush-Kulesha M. Fedorova A., Moysa B. Social rights in Ukraine during the war. Report on needs assessment. Council of Europe. 2022, 64 p. URL : https://rm.coe.int/needs-assessment-ua-2/1680a9b408 (ukr)

[5] On the approval of the Regulation on the Unified Information System of the Social Sphere. Resolution of the Cabinet of Ministers of Ukraine dated April 14, 2021 No. 404. URL: https://zakon.rada.gov.ua/laws/show/404-2021-п#Text (ukr)

[6] Report on appeals of citizens for 9 months of 2024. URL: https://www.pfu.gov.ua/2167929-zvit-pro-zvernennya-gromadyan-za-9-misyatsiv-2024-roku/ (ukr)

[7] Sharma S., JainRole A. Role of sentiment analysis in social media security and analytics. WIREs Data Mining and Knowledge Discovery: Vol. 10, Issue 5. https://doi.org/10.1002/widm.1366

[8] Shkurko O. IN. Types of linguistic text analysis: teaching. manual Dnipro: Univ. Alfred Nobel, 2018. 119 p. (ukr)

[9] Perebijnis V. AND. Statistical methods for linguists: training. manual Vinnytsia: Nova Kniga, 2013. 176 p. (ukr)

[10] Lande D. IN. Elements of computer linguistics in legal informatics. Kyiv: NDIIP National Academy of Sciences of Ukraine, 2014. 168 p. (ukr)

[11] Find the information that matters using natural language processing (NLP). URL: https://www.sas.com/ru_ua/software/visual-text-analytics.html

[12] Survey of Text Mining I: Clustering, Classification, and Retrieval / Ed. by MW Berry. Springer, 2003. 261 p.

[13] Aggarwal CC, Zhai C. Mining Text Data. Springer, 2012. 527 p.

[14] Text Cluster Node Results. URL: https://documentation.sas.com/?docsetId=tmref&docsetTarget=n1d7r58qug6sefn162cu6cqx0nq4.htm&docsetVersion=14.3&locale=en

[15] Emerging Technologies of Text Mining: Techniques and Applications / Ed. by HA Do Prado, E. Ferneda. Idea Group Reference, 2007. 358 p.

[16] Valls Martínez, MdC, Santos-Jaén, JM, Amin, F.-u., Martín-Cervantes, PA Pensions, Aging and Social Security Research: Literature Review and Global Trends. Mathematics 2021, No. 9, 3258. https://doi.org/10.3390/math9243258

[17] Social Protection Systems. Ed. E. Schüring, M. Loewe. Elgar Publishing. 2021. 776 p. https://doi.org/10.4337/9781839109119

[18] Official website of the Ministry of Digital Transformation of Ukraine. URL : https://thedigital.gov.ua (ukr)

[19] On the approval of the Regulation on the Unified Information System of the Social Sphere. Resolution of the Cabinet of Ministers of Ukraine dated April 14, 2021 No. 404. URL : https://zakon.rada.gov.ua/laws/show/404-2021-п#Text (ukr)

[20] Gladun A. Ya., Rogushina Yu. IN. Data mining : searching for knowledge in data: a tutorial. Kyiv: ADEF-Ukraine, 2016. 451 p. (ukr)

[21] Lytvyn V.V., Pasichnyk V.V., Nikolskyi Yu.V. Analysis of data and knowledge: training. manual Lviv: Magnolia 2006, 2017. 276 p. (ukr)

[22] Analysis and processing of data flows by means of computational intelligence: monograph / Ye. IN. Bodyanskyi et al. Lviv: View of Lviv. polytechnics, 2016. 235 p. (ukr)

[23] Text analytics using SAS Text Miner: course notes. NC.: SAS Institute, 2014. 218 p.

[24] Getting Started with SAS® Text Miner 12.1 URL: https://support.sas.com/documentation/onlinedoc/txtminer/12.1/tmgs.pdf

[25] Matignon R. Data Mining Using SAS Enterprise Miner. URL: https://www.amazon.com/Data-Mining-Using-Enterprise-Miner/dp/0470149019

[26] Sharma S., JainRole A. Role of sentiment analysis in social media security and analytics. WIREs Data Mining and Knowledge Discovery: Vol. 10, Issue 5. https://doi.org/10.1002/widm.1366

[27] Find the information that matters using natural language processing (NLP). URL: https://www.sas.com/ru_ua/software/visual-text-analytics.html