On Classification Hidden Concepts Language in Specialized Texts Based on Methods of the Intellectual Data Processing

Iurii Krak[a,b], Valentina Petrovych[a], Vladislav Kuznetsov[a], Eduard Manziuk[c], Olexander Barmak[c], and Anatoliy Kulias[a]

[a] *Glushkov Cybernetics Institute, Kyiv, 40, Glushkov ave., 03187, Ukraine*
[b] *Taras Shevchenko National University of Kyiv, Kyiv, 64/13, Volodymyrska str., 01601, Ukraine*
[c] *National University of Khmelnytskyi, Khmelnytskyi, 11, Institytska str., 29000, Ukraine*

Abstract

In this article discussed and solved the problems of comparing language concepts in specialized texts, in particular, scientific texts in the Ukrainian language. A corpus of scientific texts and dictionaries as well as stop words and affixes has been formed for processing specialized texts. The resulting texts were analyzed and converted into text frequency-inverse document frequency (TF-IDF) feature representation. To transform the original vector of features, it is proposed to use an algorithm for the synthesis of linear systems, in combination with the T-stochastic neighbor embedding (T-SNE). A series of experiments were performed on test examples for the determination of informational density in the text and classification by keywords in specialized texts using the method of random samples consensus (RANSAC). A method of classification of hidden language concepts was proposed, use of clustering methods (K-means). As a result of the experiment, the structure of the classifier of hidden language concepts was obtained in structured texts. The stability of the proposed method is investigated by using the perturbation of the original data by a variational autoencoder. The obtained structure allowed to achieve a relatively high recognition accuracy (97%-99%) using decision trees and machines of extreme gradient amplification.

Keywords 1

Texts analysis, language concepts, pseudo inversion, clastering, feature extraction

1. Introduction and problems statement

One of the important problems in scientific research in the specialized texts are to compare texts according to a certain criterion, namely the inclusion of a certain phrase or set of phrases in a given scientific text to obtain results that contain the desired set of phrases [1-3]. This method is currently used successfully in searching for information by keywords, but it has disadvantages - first, the need to search all the text and search for each keyword by a given criterion, which causes search results that are not relevant to the searched text [2,4]. Note that among the many known methods of textual information research are the most popular: the method of frequency analysis of the term, taking into account the inversion of frequency to other documents - TF- IDF [4], the linear method of reference vectors (LSVM) [5] and the method of the Bag of Words [6]. The advantages of TF-IDF and Bag of Words include good speed and great applications; the main advantage of the LSVM method is accuracy. The disadvantages are the slowness of execution, "rejection" of stop words, which

CMIS-2021: The Fourth International Workshop on Computer Modeling and Intelligent Systems, April 27, 2021, Zaporizhzhia, Ukraine
EMAIL: yuri.krak@gmail.com (I. Krak); filonval63@gmail.com (V. Petrovych); kuznetsowwlad@gmail.com (V. Kuznetsov); eduard.em.km@gmail.com(E. Manziuk); alexander.barmak@gmail.com(O. Barmak); anatoly016@gmail.com(A. Kulias)
ORCID: 0000-0002-8043-0785 (I. Krak); 0000-0002-5982-8983(V. Petrovych); 0000-0002-1068-769X(V. Kuznetsov); 0000-0002-7310-2126(E. Manziuk); 0000-0003-0739-9678(O. Barmak); 0000-0003-3715-1454(A. Kulias)

inevitably leads to loss of meaning; lack of consideration of the position of each word in the text, which can provoke difficulties in finding content in a particular text.

Based on the analysis of the subject area and the selection of research issues, the following problems are formulated:

- to form a sample of scientific texts on various topics;
- get an idea of the vectors of the features of individual sentences in the text;
- analyze the affinity of texts from the annotations of texts with the source texts;
- to investigate the representation of feature vectors with the involvement of different methods of data dimension reduction, clustering and grouping of features;
- assess the information concentration of the content of different samples of text on the representation of the vectors of the characteristics of individual sentences of the text;
- to classify individual sentences of texts by subject;
- to investigate the selection of hidden concepts by clustering methods;
- evaluate the stability of classification algorithms to transformations of dimensionality of features.

2. Related works

A large number of publications are devoted to the study of specialized and scientific texts, among which the following should be noted: in [7] the amplitude and phase characteristics of terms in the text were investigated to assess the concentration of categories of terms in the text, in [8], experimental information technology was developed analysis of frequency characteristics of semantic terms and word combinations in the text and built a model Sub-Verb-Sub [8,9]. Despite the availability of these studies on this topic, a number of problems have not been resolved, in particular, the analysis of the frequencies of individual sentences in the text, the relationship of the meaning of the abstracted text (annotations, abstracts) with the text, statistical proximity [10,11] different authors on one topic and the study of ways to improve the results of text classification in the context of the problem of analysis of hidden concepts of language, in particular using the methods of grouping and reducing the dimensionality of the vector of characteristics [12,13,14].

3. Implementation

3.1. Getting data

The study formed three samples of scientific texts for modeling and recognition of communicative information on three topics: facial expressions, Ukrainian sign language and texts in Ukrainian (about 1500 sentences) [15]. After detailed processing these sets will be presented as open dataset.

Thus, using the proximity of themes, it is possible to determine how the representation of these texts in the space of characteristic features changes, which will allow to determine the common and different concepts of language in these texts. To solve this problem, scientific texts were presented in the form of a matrix in which each line corresponds to a separate sentence, including the title, annotation, captions, conclusions, and other textual elements that contain the text.

3.2. Data analysis environment

To analyze this data, a module for intelligent analysis of scientific texts in Python was created with the involvement of the scikit-learn data processing library in the Jupiter development environment [16, 17]. This module implements the following operations:

- parsing of text data;
- deletion of uninformative data and word endings;
- obtaining statistical characteristics of the text;
- transformation of text elements into a vector of characteristic features;
- transformation of the vector of features into a vector of reduced dimension;
- training and testing of methods of classification of text data by a set of features;

• visualization of the vector of characteristic features.

The resulting module was tested on a hardware platform with Windows and the following characteristics: Intel Core i5-6600k processor, 8 GB of DDR4 RAM.

3.3. Data organisation

To solve this problem, we present scientific texts in the form of a matrix, in which each line corresponds to a separate sentence, including title, abstract, captions, conclusions, and other elements that characterize this text, presented as a vector of features, where each of the features corresponds to a particular term and the frequency of its appearance in the text. To compare several texts with different topics, we will form a body of scientific texts that limits the area of interest (for example, scientific articles that contain the required information).

Texts (body of texts and compared text) are analyzed by a parser, which eliminates all words that do not carry significant meaning (stop words) and cuts off affixes (suffixes and endings) of words. On the basis of the received set of terms frequency characteristics are formed. Then both texts (or corpora of texts) are compared and all terms and, accordingly, features that are not included in at least one of the texts are cut off.

3.4. Features extraction

Each row of the matrix was presented in the representation TF-IDF (text frequency inverse document frequency) [8,9,14], where each of the elements corresponded to a separate term and frequency of its occurrence. The obtained frequency characteristics were compared and all terms and, accordingly, features that are not included in at least one of the texts were cut off from the vector of features. This allowed to preserve the dimensionality of the data and to take into account only the ratio of the number of scientific terms common to texts with different topics.

In addition, when using some methods to reduce the dimensionality of the data (in particular, Karunen-Loeve [9]), this allowed to reduce the number of zero elements and the degree of sparseness of the sample matrix.

3.5. Statistic relationship of abstracts in specialized texts

To estimate the statistical affinity in the first group of experiments, it was proposed to use the following indicators: Cartesian distance, Pearson's test and standard deviation. These values indirectly indicate the heterogeneity of sentences and therefore at the preliminary stage of the process allow you to indirectly check the validity of the sample [18,19]. As a result of tests when comparing individual sentences from a single scientific text and annotation from this text, it was noted that sentences that had different meanings differed in the proposed parameters.

The example of the 1st sentence from the annotation shows the similarity and difference of sentences on three parameters are given in Table. 1.

Table 1

Statistic features of the studied sentences

№	Abstract			Article		
n/n	Dist	Pearson	St. Dev	Dist	Pearson	St. Dev
1	0	1	0.490698	11	0.425414	0.347540
2	10	0.618363	0.529891	4	0.854699	0.325396
3	11	0.409801	0.300327	12	0.501629	0.490698
4	17	0.002661	0.300327	14	0.334416	0.418213
5	12	0.541444	0.529891	11	0.425414	0.34754

The calculated values in Table 1. show that for sentences with common concepts, there is a pattern of similarity of statistical indicators for texts with similar content, for example, in sentences that had the same Cartesian distance, but for the 1st sentence of the annotation and the 2nd of the main content, which contained close in content sentences, these values did not allow to argue about the unambiguous similarity of these sentences in content.

3.6. Feature vector dimensionality reduction

To reduce the dimensionality of the original vector of characteristic features of individual sentences, the Karunen-Loeve transformation was applied. Before performing the decomposition of the original data into eigenvectors, insignificant features were discarded, in particular, features with a low frequency of occurrence in the text to reduce the sparseness of the matrix and to exclude zero rows or columns in the covariance matrix.

As a result of this experiment, feature sets were obtained for sampling, which were analyzed further. This schedule of text sampling for the first body of texts showed that 95% of the energy of eigenvectors was within first three vectors, which allowed to visualize the spatial arrangement of feature vectors of data elements in the form of a three-dimensional diagram, shown below in Fig. 1.

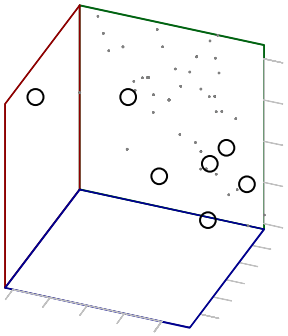


Figure 1. First three eigenvalues of the studied sentences in Karhunen-Loeve representation

As can be seen from the diagram, first, the first three eigenvalues form separate clusters of points, which allows us to argue about the difference of concepts in these sentences. Second, the distance between the elements of the test sample (circled) and the training sample (other points) allows you to visually assess the proximity of data samples to each other, and, therefore, used to identify hidden text parameters.

3.7. Dimensionality reduction using grouping of features

Since the three-dimensional representation is not convenient for visualization, it was proposed to perform feature grouping for visualization in the form of a flat image, which additionally used T-stochastic neighbor embedding [20] to reduce the dimensional vector of features to two. For clarity, Fig. 2 below shows the representation of the feature vector for the second body of texts by T-stochastic neighbor embedding and clustiring of points by the K-means.

Figure 2. Representation of a feature vectors for the second body of texts by T-stochastic neighbor embedding, grouping of features and clustering of points by the method of K-means

At the Fig. 3 shows histograms of the distribution of feature values for the first and second corpora of texts, respectively.

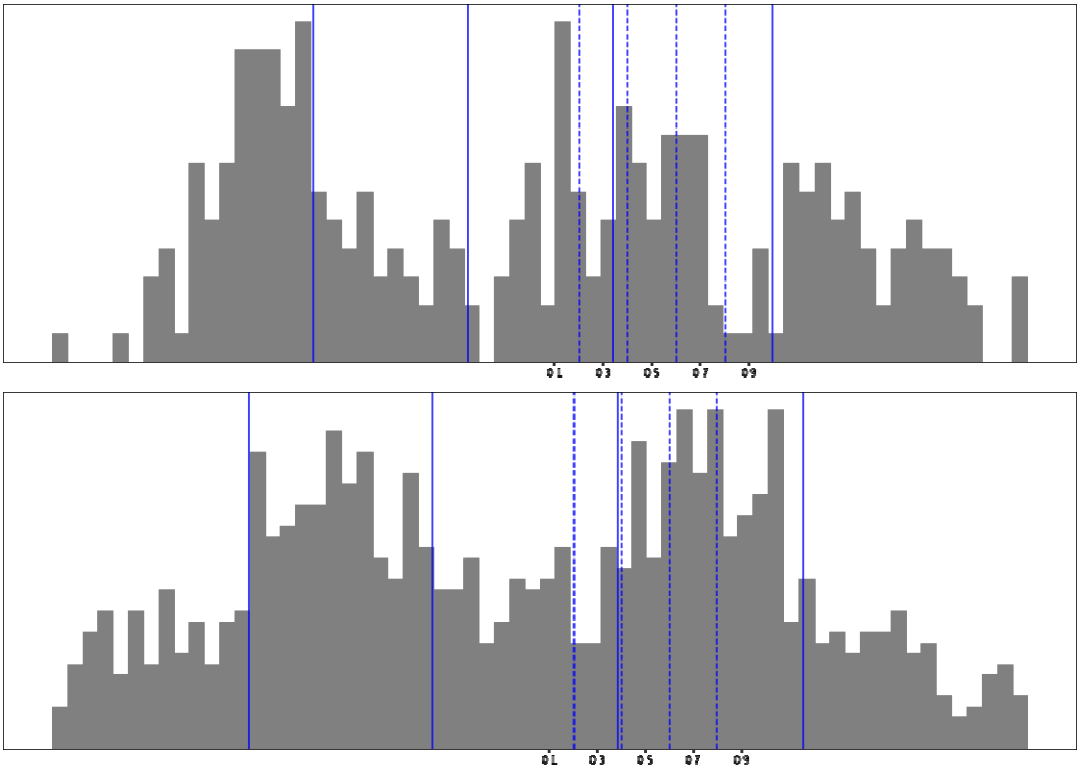


Figure 3. Histograms of the distribution of feature values for the body of texts 1 and 2

3.8. Informational density of sentences in text

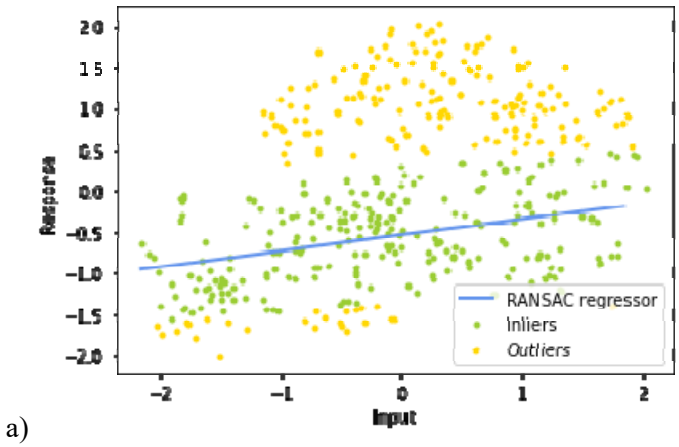
With a relatively large number of samples in the body of texts and, accordingly, the educational sample, the sentences and their location in the feature space should come to the fore [21]. Accordingly, texts that differ significantly in content will have a large number of non-occurrences of sentence elements in clusters, which indicates the difference between the subject matter of the text and hidden concepts.

To test this hypothesis, it was proposed to use a regression function - the method of consensus of random samples (RANSAC) [22,23]. In Fig. 4 shows the regressions on three samples of scientific texts in the representation of T-SNE features.

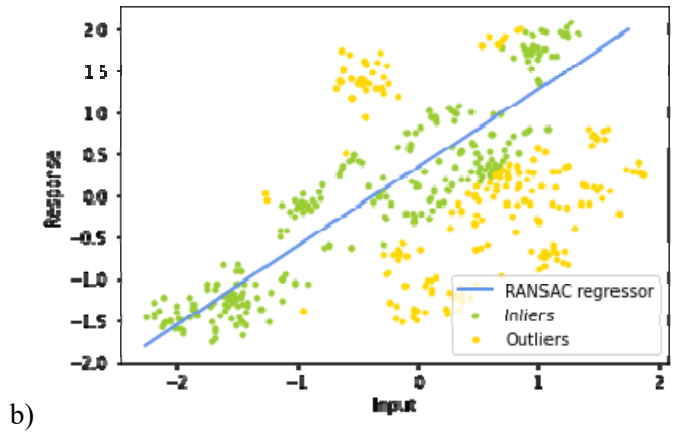
Dark gray in Fig. 3 shows which points are accepted by the method of the most informative and used to build a regression. Accordingly, comparing the graphs, we can indicate that the texts have a common theme, because the location of the points intersect. In addition, the dark gray dots in Fig. 3 also indicate the areas with the greatest information density of the content of each of the corpora.

3.9. Sentences classification using different text corpora

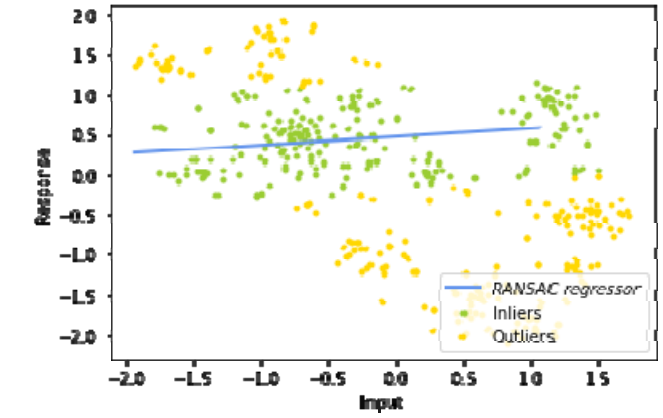
A series of tests was performed on T-SNE representations of texts involving methods of intelligent data processing [24,25]. Thus, the following methods were studied: random forest, decision tree, nearest neighbors method, reference vectors method, naive Bayesian classifier, single-layer neural network. As part of the experiment, the Bayesian classifier performed best on the obtained data set. The learning results of the algorithms are illustrated in Table 2.



a)



b)



c)

Figure 4. Sentences from 1 to 3 (a to c) and their informational density

Table 2

Precision, recall, score and support for Bayesian classifier on given datasets

Dataset	Precision	recall	f1-score	support
Emotion	0.94	0.87	0.90	138
Gesture	0.82	0.94	0.87	163
NLP	0.95	0.87	0.91	159
accuracy	0.90	0.89	0.89	460
macro avg	0.90	0.89	0.89	460
weighted avg	0.90	0.89	0.89	460

Based on the experiment, it was shown that the obtained set of TF-IDF features allows to classify texts with high reliability (87%). The presence of recognition errors of the 1st and 2nd kind (see Fig.

4) is explained by the great affinity of the texts and the closeness of the author's styles in scientific texts, which involves the use of common terms, and this can be seen from the similarity of representations in Fig. 3.

3.10. Classification of text using help of data clustering

In order to study the possibility of improving the results of classification, the following approach is proposed: since the studied texts include hidden concepts (subsets) from the category of sentences that are close in content, the selection of such hidden concepts allows to correctly set the task of classifying texts, namely to classify sentences according to their similarity to a particular concept of the text [26].

To do this, it was proposed to use one of the implementations of the method of K-means (mini-batch K-means), which is best suited for experimental data. The obtained data clusters were selected as data labels, which thus allowed to move from the classification by subject of texts to the classification by the method of clustering [27,28], taking into account the internal structure of the data, namely the hidden concepts of language.

The following methods were chosen as classification methods: the method of reference vectors with linear and nonlinear hypothesis, single-layer neural network, Bayesian classifier, decision trees and related methods, including adaptive amplification machine and extreme gradient amplification. As a result of the test it was shown that the obtained set of data from the studied methods of the highest accuracy of recognition of hidden concepts is achieved by the method of reference vectors with a linear hypothesis (97.4%), single-layer neural network (97.4%), random forest (99.1%) , decision tree with extreme gradient boost (99.1%).

3.11. Algorithm stability testing using perturbations of feature matrix

To verify the algorithm, an additional experiment was performed to study the stability of determining data classes when introducing perturbations into the vector of characteristic features. Perturbation was performed by converting the original representation of T-SNE (two-dimensional) into a latent feature space (two-dimensional), in which the data points from the source space corresponded to the location in the latent feature space. A variational autoencoder (VAE) was used for this purpose [28]. This method minimizes the rms error between the input (in the T-SNE representation) and the output data set (in the latent feature space), and generates additional data elements that have the same distribution as the sample of training samples in Fig. 5.

In addition, in order to achieve high accuracy of data representation, the hidden layer of the autocoder has a dimension much higher than in conventional applications.

Using the dimensional transformation obtained by the autocoder, it was noted that the representation of features in the latent space (see Fig. 1) allows us to build a linear hypothesis when classifying each other (one class versus another class). Below in Fig. 6 shows hypotheses and data labels for representation of features in the latent space.

Considering Fig. 5 and Fig. 6 it can be noticed that the separation band has decreased. Accordingly, this led to a decrease in the accuracy of recognition by classification methods by 12% - up to 85% for the method of reference vectors and single-layer neural network and by 2% - up to 97.1% for random forest and decision tree methods with extreme gradient enhancement. Despite the decrease in the accuracy of recognition, this experiment is interesting to study the effect of perturbation on the accuracy of recognition by different methods of classification.

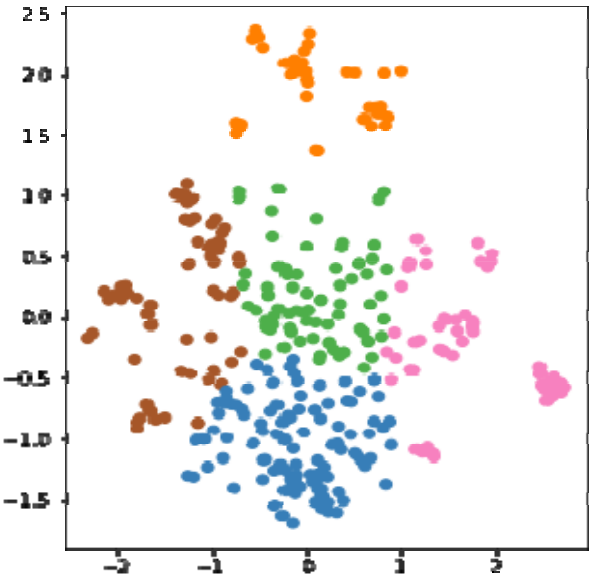
3.12. Determining the nature of perturbations in a sample matrix using pseudo inversion

Perturbation of input data affects the convergence of the classifier learning algorithm. The measure of perturbation in the original and transformed matrix in the latent space of features is the

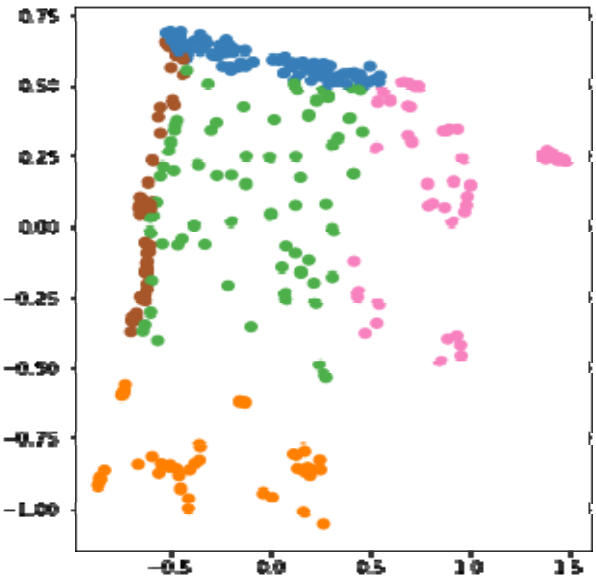
value of entropy. Entropy can be determined indirectly by the decrease in energy of its eigen mean square error vectors of the sample matrix.

A variational auto encoder minimizes the average quadratic error between data elements (MSE). MSE is also an expression of the magnitude of data. Thus, it is possible to associate the value of variance change with the nature of the decline in energy of its own numbers of the source and transformed matrix.

Knowing the magnitude of the variance change, we can estimate the number of vectors responsible for the informative part of the data under study. The advantage of latent representation is the reduction of the distance between data classes (and separation band), which ultimately affects the number of iterations of the optimization algorithm and its convergence of the classification algorithm.



a)



b)

Figure 5. T-SNE a) and b) representation of feature space of higher dimension

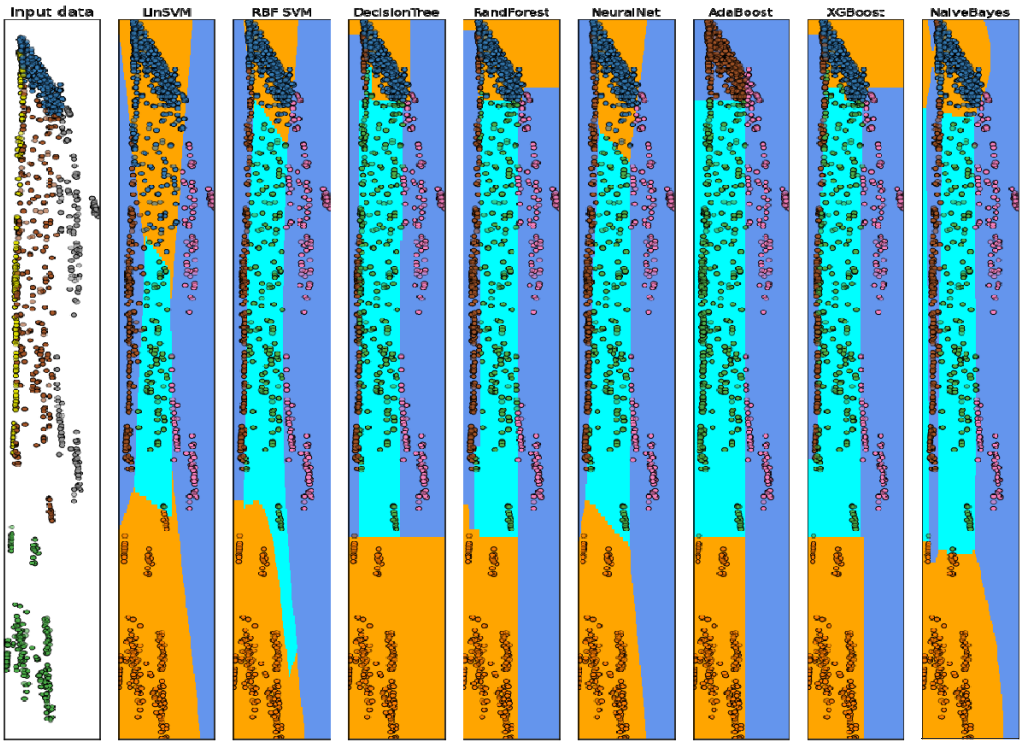


Figure 6. Hypotheses visualizing for different classifiers in latent feature space

4. Conclusion

The presented work proposes an approach to text analysis uning text mining methods, including TF-IDF feature extraction method, feature dimensionality reduction based on Karhunen-Loeve transform and T-SNE data representation in 2-dimensional space and classification of acquired data using decision tree models. Also the paper discusses how the stability of the methods is affected by perturbation of data using variational autoencoder.

In order to study proposed method an experimental dataset was obtained; this dataset consists of three text corpora on three different topics – facial expression recognition, gesture recognition and text mining; these datasets related to similar topics of research and thus giving a possibility to study proposed methods in condition where the datasets have some features in common.

The experimental results of algorithms is shown in tables and figures; most important results are depicted on the Fig. 4, Fig.5 and Table 2. Given data is separable and has alignment axes in feature vector in reduced dimensionality representation; moreof, application of variational autoencoder copes well with data noise and reduces innecessary variance but decreases separability of the data.

Feature vectors of individual sentences formed from the appearance of individual words are very effective for the classification of texts by subject, but the presence of common themes of scientific texts and typical sentences reduces the efficiency of recognition.

The use of dimensional reduction and grouping methods indicated the presence of data point concentrations that could be used to assess the author's style using typical sentence constructs. Two results should be pointed out separately: in the classification of hidden concepts (clusters of data points) with the use of clustering methods and the introduction of perturbations into the original data.

Thus, given results make it possible to determine the further direction of research, namely the study of a sample of scientific texts with other topics and authorial styles, the use of other methods of classification of structured texts, what will be the goal of our further investigation.

5. References

[1] N. Firoozeh, A. Nazarenko, F. Alizon, B. Daille. Keyword extraction: Issues and methods. Natural Language Engineering, 26(3) (2020): 259-291. doi:10.1017/S1351324919000457

[2] J. Ventura, J. Silva. (2007). New techniques for relevant word ranking and extraction. In: Neves J., Santos M.F., Machado J.M. (eds) Progress in Artificial Intelligence. EPIA 2007. Lecture Notes in Computer Science, 4874. Springer, (2007), pp.691–702. https://doi.org/10.1007/978-3-540-77002-2

[3] M. Ortuno, P. Carpena, P. Bernaola, E. Munoz, A.M. Somoza. Keyword detection in natural languages and DNA. Europhys. Lett, 57 (5) (2002): 759-764.

[4] B. Das, S. Chakraborty. An Improved Text Sentiment Classification Model Using TF-IDF and Next Word Negation. 2018. arXiv preprint arXiv:1806.06

[5] M. Labbé, L.I. Martínez-Merino, A.M. Rodríguez-Chía. Mixed Integer Linear Programming for Feature Selection in Support Vector Machine. Discrete Applied Mathematics, 261. Elsevier, (2019), pp.276-304. ff10.1016/j.dam.2018.10.025f.

[6] B. Heap, M. Bain, W. Wobcke, A. Krzywicki, S. Schmeidl. Word Vector Enrichment of Low Frequency Words in the Bag-of-Words Model for Short Text Multi-class Classification Problems, 2017. arXiv:1709.05778

[7] Y. Krak, O.Barmak, O. Mazurets. The practical implementation of the information technology for automated definition of semantic terms sets in the content of educational material. CEUR WS, Vol. 2139, (2018):245-254. DOI:10.15407/pp2018.02.245

[8] S. Robertson. Understanding inverse document frequency: On theoretical arguments for IDF, Journal of Documentation. 60 (5) (2004): 503–520.

[9] A. Aizawa. An information-theoretic perspective of tf-idf measures, Information Processing and Management. 39 (1) (2003): 45–65.

[10] M. Farouk. Measuring Sentences Similarity: A Survey. Indian Journal of Science and Technology, 12(25) (2019): 1-11. DOI: 10.17485/ijst/2019/v12i25/143977

[11] W.H. Gomaa, A. A. Fahmy. A survey of text similarity approaches. International Journal of Computer Applications, 68(13) (2013): 13-18.

[12] A.V. Barmak, Y.V. Krak, E.A. Manziuk, V.S. Kasianiuk. Information technology of separating hyperplanes synthesis for linear classifiers. Journal of Automation and Information Sciences, 51(5) (2019): 54-64. doi: 10.1615/JAutomatInfScien.v51.i5.50

[13] Iu.V. Krak, G.I. Kudin, A.I. Kulyas. Multidimensional scaling by means of pseudoinverse operations. Cybernetics and Systems Analysis, 55(1) (2019): 22-29. doi: 10.1007/s10559-019-00108-9

[14] E.L. Shimomoto, L.S. Souza, B.B. Gatto, K. Fukui. Text classification based on word subspace with term frequency. 2018. arXiv:1806.03125v1

[15] Iu.V. Krak, O.V. Barmak, S.O. Romanyshyn. The method of generalized grammar structure for text to gesture computer-aided translation, Cybernetics and Systems Analysis, 50(1) (2014): 116-123. doi: 10.1007/s10559-014-9598-4

[16] S. Bird, E. Klein, E. Loper. Natural Language Processing with Python. O'Reilly Media, 2009

[17] J. Perkins. Python Text Processing with NLTK 2.0 Cookbook. Packt Publishing, 2010.

[18] T. Mikolov, K. Chen, G. Corrado, J. Dean. Efficient Estimation of Word Representations in Vector Space. 2013. arXiv:1301.3781.

[19] A. Globerson, G. Chechik, F. Pereira, N. Tishby. Euclidean Embedding of Co-occurrence Data, Journal of Machine Learning Research, 8 (2007): 2265-2295.

[20] L. Van der Maaten, G. Hinton. Visualizing Data using t-SNE, Journal of Machine Learning Research, 9 (2008): 2579-2605.

[21] T. Mikolov. Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems. 2013. arXiv:1310.4546.

[22] M.A. Fischler, R.C. Bolles. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Comm. of the ACM, 24(6) (1981): 381-395. https://doi.org/10.1145/358669.358692.

[23] A. Hast, A. Nysjö, A. Marchetti. Optimal RANSAC – Towards a Repeatable Algorithm for Finding the Optimal Set, Journal of WSCG, 21(1)(2013): 21-30.
[24] Survey of Text Mining I: Clustering, Classification, and Retrieval. Ed. by M. W. Berry. 2004, . https://www.springer.com/gp/book/9780387955636.
[25] Emerging Technologies of Text Mining: Techniques and Applications. Ed. by H. A. Do Prado, E. Ferneda. IGI Global, 2007.
[26] G.E. Hinton, R.R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks, Science, 313(5786) (2006):504-507. doi: 10.1126/science.1127647.
[27] E.A. Manziuk, A.V. Barmak, Y.V. Krak, V.S. Kasianiuk. Definition of information core for documents classification, J. Autom. Inf. Sci. 50(4) (2018): 25-34.
[28] S. Ioffe, C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. 2015. arXiv:1502.03167[cs].