Post-print of: N. Rizun and W. Waloszek, "The algorithm of building the hierarchical contextual framework of textual corpora," 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, 2017, pp. 366-372, doi: 10.1109/INTELCIS.2017.8260064.

The Algorithm of Building the Hierarchical Contextual Framework of Textual Corpora

Nina Rizun
Gdansk University of Technology
Department of Applied Informatics in Management
Faculty of Management and Economics
Gdansk, Poland
nina.rizun@zie.pg.gda.pl

Wojciech Waloszek
Gdansk University of Technology,
Department of Software Engineering,
Faculty of Electronics, Telecommunications and Informatics, Gdansk, Poland
wowal@eti.pg.gda.pl

***Abstract* — This paper presents an approach for Modeling the Latent Semantic Relations. The approach is based on advantages of two computational approaches: Latent Semantic Analysis and Latent Dirichlet Allocation. The scientific question about the possibility of reducing the influence of these Methods limitation on the Quality of the Latent Semantic Relations Analysis Results is raised. The case study for building the Two-level Hierarchical Contextual Framework of Textual Corpora was performed. The main scientific contributions of this research are: using the paragraphs as a topically completed textual messages can guarantee that it will be centered on a single topic; collecting the topics within the Corpora via its identification in each document separately is the instrument for preventing the model size increasing; film's review as a specific type of textual document have the approximately similar writing style only within the Corpora with the same semantic tonality.**

Keywords — Latent Semantic Analysis; Latent Dirichlet Allocation; Corpora; Latent Semantic Relations; Topics

I. INTRODUCTION

The analysis of Latent Semantic Relations (LSR) in textual documents is, on the one hand, very developed, and on the other – still an open question in the field for improving every day. A lot of different styles of texts writing, the various content of these documents, their differing size and the specificity of a language – all this makes the relevance and topicality of the issues of finding algorithms and methods of recognizing semantic relations and the topical structure of textual Corpora.

A lot of scientific disputes and discussions arise on the basis of finding the optimal and universal way of the text-mining. However, in our opinion, the specifics of the analysed documents (article, review, essay, etc.), as well as the language of writing imposes certain limitations and makes specific requirements for the implementation of particular algorithms and techniques (both methodical and technical issues).

This paper is devoted to the development of the problem of finding an effective tool for analysis and formation of a topical framework of the text corpora, taking into account the type (style) of documents and language in which the text is written.

II. THEORETICAL JUSTIFICATION

The aim of the LSR analysis is to extract "semantic structure" of the collection of information flow and automatically expands term into the underlying topic. Significant progress on the problem of presenting and analysing the data have been made by researchers in the field of information retrieval (IR) [1, 11-12]. The basic methodology proposed by IR researchers for text collection – reduces each document in the corpora to a vector of real numbers, each of which represents ratios of counts.

The vector model (Vector Space Model, VSM) [9-13] of text representation is one of the first methods used to solve latent semantic relations revealing the topic modeling problems. Initially, this model was used in topic detection tasks by extracting events from the information flow [10, 14]. The representation of the corpora in this case realized with the help of vectors models form, in which each word is weighted according to the chosen weight function [15, 16]. To fully define the vector model it is necessary to specify exactly how will determine the weight of the word in the document. Various methods are used for this: a statistical approach (Boolean weight, TF-IDF, logarithm of word entry into text.), the place where the word appears, word processing, etc. [7, 9-10, 17].

Latent Semantic Analysis (LSA) is a Discriminant theory and method for extracting context-dependent word meanings by statistical processing of large sets of text data [15, 18, 20]. It uses the "bag-of-words" for modelling, begin by transforming text corpora into term-document frequency matrices, reduce the high dimensional term spaces of textual data to a user-defined number of dimensions by singular value decomposition (SVD), produce weighted term lists for each concept or topic, produce concept or topic content weights for each document, and produce outputs that can be used to compute document relationship measures [25].

According to the theorem on singular decomposition, any real rectangular matrix can be decomposed into a product of three matrices:

$$X_{t\times d} \approx X_{K_{t\times d}} = U_{K_{t\times d}} \Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T \quad (1)$$

$\Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T$ – represents terms in *k-d* latent space;

$U_{K_{t\times d}} \Sigma_{K_{t\times d}}$ – represents documents in *k-d* latent space;

$U_{K_{t\times d}}$, $V_{K_{t\times d}}$ – retain term–topic, document–topic relations for top *k* topics.

But, as [17, 18] proved, there are three *limitations* to apply LSA: documents having the same writing style (L1); each document being centered on a single topic (L2); a word having a high probability of belonging to one topic but low probability of belonging to other topics (L3). The limitations of LSA is based on orthogonal characteristics of dimension factors as well as the fact, that the probabilities for each topic and the document distributed uniformly, which does not correspond to the actual characteristics of the collections of documents [7, 8, 23]. That is why, LSA tends to prevents multiple occurrences of a word in different topics and thus LSA cannot be used effectively to resolve polysemy issues.

The next text mining technique that was developed to improve upon LSA was the Probabilistic topic modeling techniques. Probabilistic topic modeling as a set of algorithms that allow analyzing words in textual corpora and extract from them topics, links between topics [3-6]. Latent Dirichlet Allocation (LDA) is a generative model that explains the results of observations using implicit groups, which allows one to explain why some parts of the data are similar. It was proposed by David Blei [3, 4] and it use a Bayesian model that treats each document as a mixture of latent underlying topics, where each topic is modeled as a mixture of word probabilities from a vocabulary.

The algorithm of the method is following: Each document is generated independently: randomly select for document its distribution on topics θ_d for each document's word; randomly select a topic from the distribution θ_d, obtained in the first step; randomly select a word from the distribution of words in the chosen topic φ_k (distribution of words in the topic *k*). In the classical model of LDA, the number of topics initially fixed and specifies the explicit parameter *k*. In the process of the assigning the topics to documents usually LDA use the maximal form possible (not always very high) level of probability of a documents belonging to the topic.

According to [16] – words in a topic from LDA (as an extended LDA method) are more closely related than words in a topic from LSA. For polysemy, words in a topic from LDA can be appeared in other topics simultaneously: topics are dirichlet multinomial random variables, each word is generated by a single topic, and different words may be generated from different topics. The *limitation* of LDA is that there is no probability distribution model at the level of documents. Thus, the larger the number of documents, the larger the LDA model (L4).

The most common method of evaluating the quality of probabilistic topic models is the calculation *Perplexity* index on the test data set D_{test} [2, 3-4]. In information theory, perplexity is a measurement of how well a probability model predicts a sample. A low perplexity indicates the probability distribution is good at predicting the sample:

$$Perplexity\,(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \qquad (2)$$

There is some research, which performed the comparison of these two methods [18, 24]. But there are mostly dedicated only to find the differences (in compare with humans in classifying or in small/Large Scale Test Results). The major research gap is to find the algorithms to increase the quality of textual documents classification via synergy of usage the advantages and taking into account the limitation of determinant (LSA), and probabilistic (LDA**)** approaches. To address this gap, this paper focuses on:

– *developing* the complex Algorithm of Modeling and Analysis of the Latent Semantic Relations bases on advantages of both computational approaches;

– in the process of development *taking into account* the document's type and language;

– conducting a *case study* for building the Hierarchical Structure of topics in Polish-language Film Reviews Corpora as a demonstration of the basic workability of proposed Algorithm.

III. METHODOLOGY

On this paper the following author's definitions will be used:

1. Corpora (films reviews corpora sample, FRCS) is a collection of the textual Documents.
2. Term is a word after preprocessing.
3. Context Fragment (F) is indivisible, topically completed, sequence of terms unit (nor less than 150 terms), located within a document's paragraph.
4. Latent Semantic/Probabilistic topics is a basic unit of Latent Semantic Relations, received by LSA/LDA approach.
5. Subjectively Positive (CFSP) and Subjectively Negative (CFSN) Corpora Samples is a result of the classification of the FRCS on the basis of information on the subjective assessment of film's semantic tonality by the reviewers (measured by 10-point scale).
6. Contextual Summary (CS) is a set of Latent Probabilistic Topics (LPT), described the topical context of each document of the Corpora.
7. Contextual Framework (CF) is a set of Latent Semantic Topics (LST), described the main semantic context the of the whole Corpora.
8. Hierarchical Semantic Corpora (HSC) is a structure of the clustered paragraphs of the FRCS, which relate to a particular Topic from Contextual Framework.
9. Contextual Dictionary (CD) is a set of terms, described the particular Topic.

A. Novelty and Motivation

Taking into account noted above strengths and weaknesses of Discriminant and Probabilistic approaches of Latent Semantic Relations analysis, the following scientific research question was raised:

Using what approaches is possible to reduce the influence of Discriminant and Probabilistic Methods limitation on the Level of Quality of the Latent Semantic Relations Analysis Results?

For finding the answers for this question the following main hypothesis was formulated:

Hypothesis H1. Taking into account the specificity of chosen case study and presence the nonofficial requirements of film's review structure and writing rules [21, 22] (future – case study specificity), assume that the writing style of each review is the approximately the same (reducing the influence of L1).

Hypothesis H2. Taking into account the case study specificity, assume, that each paragraph (F) centered on a single Latent topic and should be analyzed separately (reducing the influence of L2).

Hypothesis H3. Taking into account that paragraph interpreted as topically completed textual messages, assume, that each document is the sub-corpora, characterize by particular set of topics, and it should be analyzed separately (reducing the influence of L3 and L4).

Hypotheses H4. The synergy effect of using the advantages of LSA and LDA methods may consist in the applying them for realization the decomposition and synthesis of solved problem as a steps of Systems Analysis approach.

Basic version of proposed *Algorithm of Bilayer Modelling and Analysis of the LSR* (further – *author's Algorithm*) for receiving (as a case study) the Two-level Hierarchical Contextual Framework of Textual Corpora includes the 5 steps (fig. 1).

For demonstration the of basic workability of the author's Algorithm realization, as a *case study* the 3000 Polish-language films reviews (1500 *Subjectively Positive* and 1500 *Subjectively Negative*) from the filmweb.pl were used.

All words/terms of film reviews in this paper will be presented in English languages. The experimental part of all steps of author's Algorithm technically realized in Python 3.4.1.

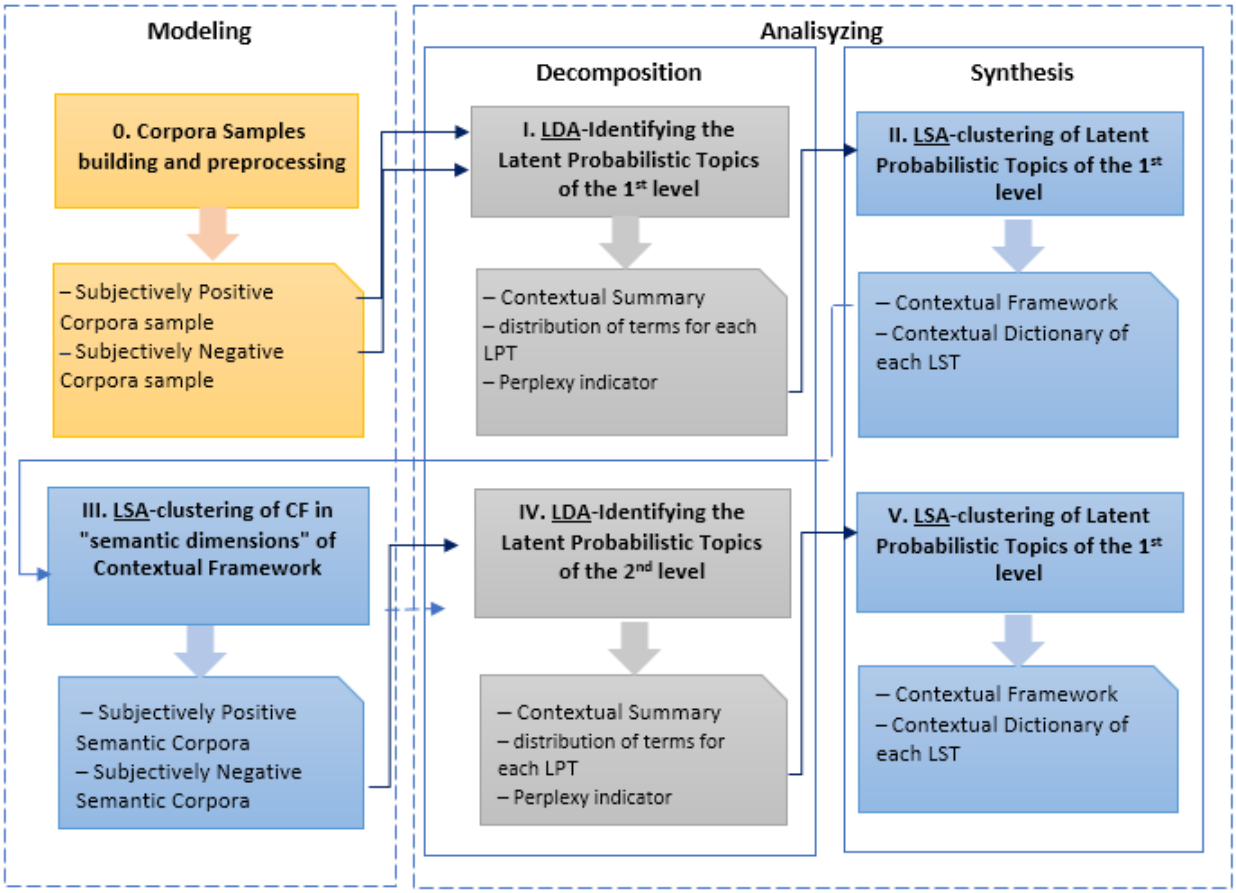

Figure 1. The Algorithm of the Two-level Hierarchical Contextual Framework of Textual Corpora building

B. Algorithm of the Two-level Hierarchical Contextual Framework of Textual Corpora building

1) Step 0. Corpora Samples Building and Preprocessing

The purpose of this step – to perform the *modeling* of the initial data for the 1st layer of LSR analysis, namely: CFSP and CFSN Corpora Samples building; text preprocessing; text preparation.

For realizing the Corpora Samples building, the following heuristic was adopted: to consider the CFSP, if the subjective review's assessment is more than 5 points, and CFSN – if it is equal or less 5 points.

Taking into account the specificity of the case study, as well as limited number of existing software and algorithmic implementations for the analysis of texts in Polish [21], in addition to standard procedures for text *preprocessing*, the authors have provided text *adaptation* procedure, based on replacement of the Film's Titles, the Names/Surnames of Creators/Actors/Hero of the film into the corresponding positions in the reviewed film (for example, the Title of the film is replaced by "Film", Name /Surname of the actor – by "Actor" etc.).

The Structure of Distribution of the Number of Terms Remaining after the Preprocessing, as a case study result, presented in Table 1. Even these facts may indicate that the Subjectively Positive reviews in compare with Subjectively Negative are characterized by a higher percentage of words to be deleted, namely low level of repetitive words. From the point of view of the morphological analysis, these results can presumably attest to the following: Subjectively Positive reviews characterized by highly semantic structured opinion, expressed in a carefully and balanced manner; the Subjectively Negative reviews characterized by average level of semantic structure of the opinion, expressed more spontaneously and under the influence of emotions.

This fact partly rejects the *Hypothesis* H3: reviews have the approximately same writing style only within the same Corpora sample.

Table 1. The Structure of Distribution of the Number of Terms Remaining After the Preprocessing

SPSC		SNSC	
% of Remaining Terms	% of Documents	% of Remaining Terms	% of Documents
4.16	15	4.16	12
17.96	46	17.96	22
31.75	17	31.75	37
45.55	10	45.55	10
59.35	6	59.35	7
73.14	5	73.14	8
92.00	1	92.00	4

2) Step I. LDA-Identifying the Latent Probabilistic Topics of the 1st level

The *Text Preparation* is the process of the initial data for LDA modeling forming: the files in specific format with list of terms by each paragraph with frequency its characteristics.

The purpose of this step – via implementation of the *decomposition* approach, present the Corpora in form of Contextual Summary. This involves the revealing the sets of Latent Probabilistic Topics for each Document with information about most probable (significant) words assigning to this topic. For obtaining the optimal combination – Number of topics / Number of terms in topic– the values of Perplexity were used. The optimum value of the Perplexity index achieved in the point, when further changes in the parameters do not lead to its significant decrease.

The studying of the Perplexity value depending on the size of the Corpora, proves the validity of the assumptions that providing the analysis the Corpora by paragraphs (*Hypothesis* H2) and by document as a sub-corpora give the possibility to increase the quality (*Hypothesis* H3) of textual data classification (fig. 2.).

The structure of the Contextual Summary, described the main context the of the Corpora Samples, as a *case study* result, is presented in Table 2.

The quality indicator – recall rate as the ratio of the number of topically recognized paragraphs (probability of belonging the paragraph of topic >0,7) to the total number of paragraphs – is within 90-95%.

Table 2. The Structure of the Contextual Summary

Corpora Samples	Number of paragraphs	Number of topics in Corpora	Average Number of topics in Document	Average Number of terms in Topic	Average Perplexity Value
CFSP	10239	36730.0	5.1	5.7	1 182 169
CFSN	10934	41015.0	4.2	6.1	1 342 155

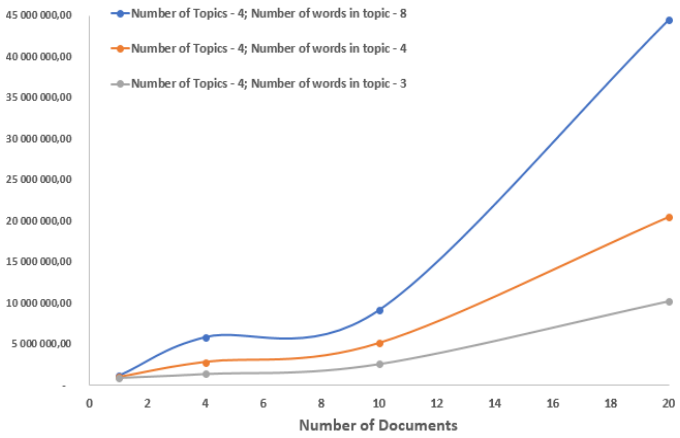

Figure 2. Perplexity Values Depending on the Size of the Corpora

3) Step II. LSA-clustering of Latent Probabilistic Topics of the 1st level

The purpose of this step – via implementation of the *synthesis* approach, to present the Corpora as a Contextual Framework. This involves the revealing the sets of Latent Semantic Topics for the whole Corpora Sample, based on the presence the Latent Semantic Relations between the elements of Contextual Summary.

This step includes: text preprocessing; creating the Term-Document Matrix; SVD process; identifying the hidden semantic connection between the Latent Probabilistic Topics; LSA clustering of topics / terms in the semantic dimension.

For realization of the clustering process, based on the matrices of cosine distances between the LPT-Vectors of Reduced dimension ($U_{K_{t\times d}}\Sigma_{K_{t\times d}}$), the k-means algorithm had chosen [19, 21]. As a recommended number of clusters, Average Number of topics for each Contextual Framework *(Table 2).*

For each cluster of both Contextual Summary: Contextual Dictionary as the list of keywords was identified (CD); a weights for keywords were formed (W); the Contextual Labels of each of the CF were specified. The Contextual labels (CL) of the Topics were assigned automatically on the bases of the terms with the highest frequency in each topic. The examples of Contextual Frameworks, as *a case study* result, are given in Tables 3-4.

Table 3. Contextual Framework for Subjectively Positive Corpora Sample

CD	W	CD	W	CD	W	CD	W	CD	W
"Hero"		**"Director"**		**"Script"**		**"Plot"**		**"Spectator"**	
hero	1.0	director	1.0	script	1.0	plot	1.0	spectator	1.0
playing	0.8	creator	0.8	history	0.8	character	0.8	fan	0.6
person	0.6	stage	0.6	writer	0.6	action	0.6	watch	0.8
actor	0.4	drama	0.4	picture	0.4	film	0.4	interest	0.4
main	0.2	effect	0.2	layer	0.2	history	0.2	movie	0.2

Table 4. Contextual Framework for Subjectively Negative Corpora Sample

CD	W	CD	W	CD	W	CD	W
"Hero"		**"Actor"**		**"Creator"**		**"Plot"**	
hero	1.0	actor	1.0	writer	1.0	plot	1.0
spectator	0.8	character	0.8	director	0.8	history	0.8
climate	0.6	picture	0.6	film	0.01	stage	0.6
person	0.4	role	0.4	cast	0.6	script	0.4
fan	0.2	layer	0.2	action	0.2	scenarist	0.2

4) Step III. LSA-clustering of CF in "semantic dimensions" of Contextual Framework

The purpose of this step – to perform the *modeling* of the initial data for the 2nd layer of LSR analysis, namely: build the Hierarchical Semantic Corpora as a structure of Subjectively Positive and Subjectively Negative Semantic Clusters.

In this step as a basic the LSA algorithm is used in the following interpretation: each element of Contextual Framework is added to the CFSP/CFSN as a separate paragraph of Corpora Sample and use in the process of clustering as a query.

The goal of clustering process – to receive the sets of paragraphs, semantically close to Topics from Contextual Framework. This method was compare with the Classical Method (CM) of frequency analysis of matching the terms from the paragraphs with weighted keywords form the Contextual Framework elements.

The example of the Structure of the Hierarchical Semantic Corpora (Table 5), as a case study results, show that the quality of CF semantic closeness recognized is higher with using the LSA method.

Table 5. The Structure of the Hierarchical Semantic Corpora

CFSP			CFSN		
Contextual Labels	**% of paragraphs (LSA)**	**% of paragraphs (CM)**	**Contextual Labels**	**% of paragraphs (LSA)**	**% of paragraphs (CM)**
"Hero"	32.93	27.19	**"Hero"**	27.08	32.17
"Director"	7.55	9.67	**"Actor"**	16.62	14.48
"Script"	32.93	25.08	**"Creator"**	36.46	38.34
"Plot"	15.11	15.41	**"Plot"**	19.84	15.01
"Spectator"	11.48	22.66			
% of not recognized paragraphs	11.10	21.5	**% of not recognized paragraphs**	15.6	25.7
Recall	85.5%	68.6%	**Recall**	82.3%	70.1%
Precision	80.2%	77.5%	**Precision**	79.1%	76.8%

5) Step IV-V. LDA-clustering of Latent Probabilistic Topics of the 2nd level + LSA-clustering of Latent Probabilistic Topics of the 2nd level

In accordance with the philosophy of implementing the step III, the following *Hypotheses* for realizing the steps IV and V is proposed:

Hypotheses 5. Each element of HSC contains the separate set of topically close paragraphs. In its turn this set could be characterized by their own set of Latent Semantic Topics.

In these case, the purpose of these steps – to repeat the steps II and III with Hierarchical Semantic Corpora as initial data.

The example of the Two-level Contextual Framework of the studied Corpora, as a case study results, presented in the Tables 6-7.

Table 6. Two-level Contextual Framework for CFSP

CF of the 1st level	CF of the 2nd level	% of CF
"Hero"	Actor / Game	24%
	History / Film	43%
	Picture / Scene	30%
	Director / Creator	3%
"Director"	Film / Director	30%
	Scene / Story	10%
	Style	6%
	Creator / Author	54%
"Script"	Film / Director	8%
	History / Hero	58%
	Author / Creator	13%
	Role / Actors	21%
"Plot"	Film / Effects	5%
	Portrait / Image	31%
	Director / Production	24%
	Script / History	40%
"Spectator"	Hero / Fan	40%
	Film / Aspects	20%
	Role / Formulation	16%
	Scene / Director	24%

Table 7. Two-level Contextual Framework for CFSN

CF of the 1st level	CF of the 2nd level	% of CF
"Hero"	Action / History	49%
	Director / Cinema	21%
	Scene / Actor	31%
"Actor"	Hero / Image	24%
	Role / Scene	58%
	Script / History	18%
"Creator"	Hero / Scene	23%
	Film / Script	60%
	Picture / Actor	18%
"Plot"	History / Hero	39%
	Director / Image	18%
	Creator / Film	43%

Such data give us the generalized Framework of Semantic (Contextual) Structure of the Corpora, which can be used for:

– increasing the efficiency of the search machine processing (recall rate and precision indicators) both within the Corpora and in cases of external accessing to find information about the films;

– preparing the information platform for creating a topically oriented Sentiment Dictionary;

– forming a platform for ontologies building.

IV. CONCLUSIONS

In this paper authors presented the complex Algorithm of Modelling and Analysis of the Latent Semantic Relations bases on advantages of both computational approaches. The main contribution of the paper and the author's studying results is the finding the answers on the main scientific research question:

– one of the manifestations of the synergy effect of using the advantages of Discriminant and Probabilistic Methods for can be the applying them for realization the decomposition and synthesis of solved problem of increasing the Quality of the Latent Semantic Relations Analysis Results;

– using the paragraphs as a separate topically completed textual messages can guarantee that each paragraph mainly being centered on a single topic and decrease the influence of the Discriminant Methods in the process of Latent Semantic Relations Analysis limitation;

– collecting the topics within the Corpora via its identification in each document separately is the instrument for preventing the model size increasing and improving the topic modelling quality;

– the set of semantically close paragraph with high level of significance could be associated with Hierarchical structure of own set of Latent Semantic Topics.

In the *future research*, these results planned be used these results for development the algorithms of Hierarchical Sentiment Dictionary building.

ACKNOWLEDGMENTS

The research results, presented in the paper, are supported by the Polish National Centre for Research and Development (NCBiR) under Grant No. PBS3/B3/35/2015, the project "Structuring and classification of Internet contents with the prediction of its dynamics" (Polish title: "Strukturyzacja i klasyfikacja treści internetowych wraz z predykcją ich dynamiki").

REFERENCES

[1] Baeza-Yates R., Ribeiro-Neto B. (2011) Modern Information Retrieval. Addison-Wesley, Wokingham, UK, 1999. Second edition

[2] Bahl L., Baker J., Jelinek E., & Mercer R. (1977) Perplexity – a measure of the difficulty of speech recognition tasks. In Program, 94th Meeting of the Acoustical Society of America, volume 62, page S63.

[3] Blei D., Ng A., Jordan M. (2003) Latent Dirichlet allocation. Journal of Machine Learning Research, 3: pp. 993–1022.

[4] Blei, David M. (2012) Introduction to Probabilistic Topic Models. Comm. ACM 55 (4), April, 2012: pp. 77-84

[5] Daud Ali, Li Juanzi, Zhou Lizhu, Muhammad Faqir (2010) Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China. pp. 280-301.

[6] David M. Blei. Topic modeling. http: //www.cs.princeton.edu/~blei/topicmodeling.html

[7] Deerwester S., Susan T. Dumais, Harshman R. (1990) Indexing by Latent Semantic Analysis. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

[8] Dumais, S. T., Furnas, G. W., Landauer, T. K. and Deerwester, S. (1988) Using latent semantic analysis to improve information retrieval. In Proceedings of CHI'88: Conference on Human Factors in Computing, New York: ACM, 281-285

[9] Ed. Charu Aggarwal, Cheng Xiang Zhai, (2012) Mining Text Data (Springer).

[10] Eden L. (2007) Matrix Methods in Data Mining and Pattern Recognition, SIAM.

[11] Furnas G.W., Deerwester, S., Dumais S.T., Landauer T.K., Harshman R.A., Streeter L.A., Lochbaum K.E. (1998) Information retrieval using a singular value decomposition model of latent semantic structure. In Proc. ACM SIGIR Conf., s. 465-480, ACM, New York

[12] Gerard Salton, Michael J. (1983) McGill Introduction to modern information retrieval. New York McGraw-Hill - McGraw-Hill computer science series, XV, 448 p

[13] Ivanov V., Tutubalina E., Mingazov N., Alimova I. (2015), Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2015", Moscow, pp. 22–33

[14] Jarosław Gramacki, Artur Gramacki (2010) Metody algebraiczne w zadaniach eksploracji danych na przykładzie automatycznego analizowania treści dokumentów. XVI Konferencja PLOUG, pp.227-249

[15] Kapłanski P., Rizun N., Taranenko Y., Seganti A. (2016) Text-mining Similarity Approximation Operators for Opinion Mining in BI tools. Chapter: Proceeding of the 11th Scientific Conference "Internet in the Information Society-2016", Publisher: University of Dąbrowa Górnicza, pp.121-141

[16] Lee, S., Song, J., and Kim, Y. (2010). An empirical comparison of four text mining methods. Journal of Computer Information Systems, Fall 2010.

[17] Leticia H. Anaya. (2011). Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers, Doctor of Philosophy (Management Science), 226 pp

[18] Papadimitrious, C.H., Raghavan, P., Tamaki, H., and Vempala, S. (2000). Latent semantic indexing: A probabilistic analysis. Journal of Computer and System Sciences, 61, 217-235.

[19] Rizun N., Kapłanski P., Taranenko Y. (2016) Development and Research of the Text Messages Semantic Clustering Methodology. 2016, Third European Network Intelligence Conference, Publisher: ENIC, # 33, pp.180-187

[20] Rizun N., Kapłanski P., Taranenko Y. (2016) Method of a Two-Level Text-Meaning Similarity Approximation of the Customers' Opinions. Economic Studies – Scientific Papers. University of Economics in Katowice, Nr. 296/2016, pp.64-85.

[21] Rizun N., Taranenko Y. (2017) Development of the Algorithm of Polish Language Film Reviews Preprocessing. Proceeding of the 2nd International Conference on Information Technologies in Management, Publisher: Rocznik Naukowy Wydziału Zarządzania WSM (in print).

[22] Rizun N., Ossowska K., Taranenko Y. (2017) Modeling the Customer's Contextual Expectations Based on Latent Semantic Analysis Algorithms. Information Systems Architecture and Technology: Proceedings of 38th International Conference on Information Systems Architecture and Technology, ISAT 2017, pp.364-373.

[23] Salton G., Wong A., Yang C. S. (1975) A Vector Space Model for Automatic Indexing, Communications of the ACM, Vol. 18, Nr. 11, s. 613-620.

[24] Toni Cvitanic, Bumsoo Lee, Hyeon Ik Song, Katherine Fu, and David Rosen (2016). LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents, Proceeding of the 2016 Workshop, Atlanta, Georgia, US.

[25] Patricia J. Crossno, Andrew T. Wilson and Timothy M. Shead, Daniel M. Dunlavy (2011). TopicView: Visually Comparing Topic Models of Text Collections