The Algorithm of Modelling and Analysis of Latent Semantic Relations: Linear Algebra vs. Probabilistic Topic Models

Nina Rizun[1], Yurii Taranenko[2], Wojciech Waloszek[3]

[1] Gdansk University of Technology, Department of Applied Informatics in Management, Faculty of Management and Economics, nina.rizun@zie.pg.gda.pl
[2] Alfred Nobel University, Dnipro, Department of Applied Linguistics and Methods of Teaching Foreign Languages, taranen@rambler.ru
[3] Gdansk University of Technology, Department of Software Engineering, Faculty of Electronics, Telecommunications and Informatics, wowal@eti.pg.gda.pl

Abstract: This paper presents the algorithm of modelling and analysis of Latent Semantic Relations inside the argumentative type of documents collection. The novelty of the algorithm consists in using a systematic approach: in the combination of the probabilistic Latent Dirichlet Allocation (LDA) and Linear Algebra based Latent Semantic Analysis (LSA) methods; in considering each document as a complex of topics, defined on the basis of separate analysis of the particular paragraphs. The algorithm contains the following stages: modelling and analysis of Latent Semantic Relations consistently on LDA- and LSA-based levels; rules-based adjustment of the results of the two levels of analysis. The verification of the proposed algorithm for subjectively positive and negative Polish-language film reviews corpuses was conducted. The level of the recall rate and precision indicator, as a result of case study, allowed to draw the conclusions about the effectiveness of the proposed algorithm.

Key words: Latent Semantic Analysis; Latent Dirichlet Allocation; Rules of Adjustment; Corpus; Linear Algebra; Probability.

1 Introduction

Modelling and Analysis of Latent Semantic Relations (LSR) – the approach of constructing a model of the corpus, reflecting the transition from a set of documents and set of words in the documents to a set of topics, describing the contents of documents. We can say that in the mathematical model of text collection, describing the words or documents is associated with a family of probability distributions on a variety of topics [4, 6, 13].

Construction of the mathematical model can be considered as a problem of simultaneous clustering of documents and words for the same set of clusters, known as topics. In terms of the cluster analysis the topic is the result of bi-clustering, i.e. the simultaneous clustering of words and documents in accordance with their semantic closeness. Thus, compressed semantic description of words or of a document is a

probability distribution on a variety of hidden variables (topics). The process of finding these distributions is called the topic model [18-20].

Those hidden variables (topics) allow presenting the document as a vector in the space of latent topics instead of submitting in the space of words. As a result, the document has a lower number of components, allowing faster and more efficient handling. Thus, the topic model is closely related to another class of problems known as a reduction of data dimension [14, 17-20].

The basic algorithms for modelling topics, on which we concentrate in this paper, are: determinant Latent Semantic Analysis (LSA), and probabilistic Latent Dirichlet Allocation (LDA**).** And although all of them share the fundamental assumption about latent semantic (topical) structure of the documents, they use different mathematical frameworks – Linear algebra (LSA) vs. Probabilistic Topic Modelling (LDA) [3-4, 15].

With the aim of improving the quality of Topic Modelling Process (TMP), this paper focuses on:

- *analysing* the advantages and disadvantages of Latent Semantic Relations, revealing algorithms inside the textual collection, using two different mathematical frameworks;
- *developing* the complex Algorithm of Modelling and Analysis of Latent Semantic Relations, based on advantages of two different mathematical frameworks;
- demonstration of the *effectiveness* of proposed Algorithm implementation for specific, Argumentative, type of documents, via conducting a *case study* for the Polish-language Film Reviews Corpora.

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".

2 Theoretical Background of the Research

2.1 Vector Space Models of the Semantic Relations Analysis

The aim of the LSR analysis is to extract "semantic structure" of the collection of information flow and automatically expands them into the underlying topic. Significant progress on the problem of presenting and analysing the data has been made by researchers in the field of information retrieval (IR) [1, 10-11]. The basic methodology proposed by IR researchers for text collection reduces each document in the corpus to a vector of real numbers, each of which represents ratios of counts.

In the popular $TF \times IDF$ scheme [17-21], on the basis of vocabulary of "bag of words" the $A(m \times n)$ terms-document matrix is built, which contains as elements the counts of absolute frequency of words occurrence. After suitable normalization, this term frequency count is compared to an inverse document frequency count, which measures the number of occurrences of a word in the entire corpus:

$$F_{w_i} = TF \times IDF = tf(w,t) \cdot \log_2 \cdot D/df \quad (1)$$

where, $tf(w,t)$ – relative frequency of the w^{th} word occurrence in document *t*:

$$tf(w,t) = \frac{k(w,t)}{df} \quad (2)$$

$k(w,L_t)$ – the number of w^{th} word occurrences in the text *t*; df – the total number of words in the text of *t*; *D* – total number of documents in the collection.

Then, for solving the problem of finding the similarity of documents (terms) from the point of view of the relation to the same topic, the different metric can be applied. The most appropriate metric is cosine measure of the edge between the vectors [14, 20-22].

A further part of the algorithm is to divide the source data into groups corresponding to the events, as well as in determining whether a text document describes a set of any topic. The main idea of the solution is the use of clustering algorithms [12, 14, 17-21].

The *limitations* of this method are: the calculations measure the "surface" usage of words as patterns of letters; they can't distinguish such phenomena as polysemy and synonymy [10, 13, 16].

2.2 Latent Semantic Indexing

In 1988, Dumais et al. [7] proposed a method of Latent Semantic Indexing (LSI), most frequently referred to as LSA. Deerwester et al., 1990 [8], designed to improve the efficiency of IR algorithms and search engines by the projection of documents and terms in the space of lower dimension, which includes semantic concepts of the original set of documents.

LSA is a matrix algebra process. The most common version of LSA is based on the singular value decomposition (SVD) of a term-document matrix [10]. As a result of the SVD of the matrix *A* we have three matrices:

$$X_{t\times d} \approx X_{K_{t\times d}} = U_{K_{t\times d}} \Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T \quad (3)$$

$\Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T$ – 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 [18, 19] proved, there are three *limitations* to apply LSA: documents having the same writing style (Lim#1); each document being centered on a single topic (Lim#2); a word having a high probability of belonging to one topic but low probability of belonging to other topics (Lim#3). The limitations of LSA are based on orthogonal characteristics of dimension factors as well as on the fact, that the probabilities for each topic and the document are 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 (Lim#4).

2.3 Probabilistic Topic Models

In contrast to the so-called *discriminative* approaches (LSI, LSA), in a *probabilistic* approach the topics are given by the model, and then term-document matrix is used to estimate its hidden parameters, which can then be used to generate the simulated distributions [4, 6, 17, 25].

Latent Dirichlet Allocation

LDA – generative probabilistic graphical model proposed by David Blei [3-4, 15]. LDA is a three-level hierarchical Bayesian model. The algorithm of the method is as follows: Each document is generated independently: randomly select its distribution for document 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 is initially fixed and specifies the explicit parameter *k*.

Methods of Evaluating the Quality of Results

The most common method of evaluating the quality of probabilistic topic models is the calculation of the *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 that the probability distribution is good at predicting the sample:

$$Perplexity\,(D_{test} = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \tag{4}$$

The *limitation* of LDA method is: it is possible to choose the optimum value of the *k*, but, even under condition of finding the optimal value of the *k*, the level of probability of a document belonging to a particular topic could be insignificant (Lim#5) [3-4, 15].

3 Methodology

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

1. *Term* is a basic unit of discrete data.
2. *Latent Semantic/Probabilistic topic* (*topics*) is a basic unit of Latent Semantic Relations, received by *LSA/LDA* approach.
3. *Context Fragment* (CF) is indivisible, topically completed, sequence of terms, located within a document's paragraph.

4. *Document* is a set of CF.
5. *Corpus* (films reviews corpus, FRC) is a collection of the Documents.
6. *Semantic Cluster* (SC) is the set of CF that have hidden semantic closeness (HSC).
7. *Contextual Dictionary* (CD) is a set of terms that have HSC.
8. *Subjective Sentiment Corpus* (SSC) is a collection of Documents that have common sentiment closeness.

3.1 Novelty and Motivation

Motivation scenario of this research presupposes taking into account the *Specificity* of the Document Type (SDT) and concerns finding the ways to completely or partially eliminate the *Limitations* characterizing the Discriminant and Probabilistic approaches for Latent Semantic Relations revealing. In this regard the following scientific research questions were raised:

1. Whether the taking into account of specific features of Argumentative type of document allows to affect Quality of the Topic Modelling Process Results.

2. Is it possible to increase the Level of Quality of the Topic Modelling Process Results via using the combination of the Discriminant and Probabilistic Methods?

For finding the answers to these questions the following main heuristics and hypotheses were formulated:

Heuristic H1.1. Taking into account the specificity of chosen for this study Type of Documents and presence the nonofficial requirements of Film's Review structure and writing rules [22], assume that the writing style of each review is approximately the same (eliminating the Lim#1).

Hypothesis H1.2. Taking into account the chosen Document Type Specificity, assume, that each paragraph (CF) is centered on a single Topic and should be analyzed separately (eliminating the Lim#2).

Hypothesis H2.1. The combination of the Discriminant and Probabilistic methods have a synergistic effect to improve the recall rate and precision indicator of Topic Modelling Process realization. This effect is expected to be achieved via increasing:

– the quality of LDA-method of topics recognizing via increasing the level of probability of assigning the topic to particular CF by taking into account the hidden LSR phenomena (eliminating the Lim#5);

– the quality of LSA-method of LSR recognition via adjusting the consequences of influence the uniform distribution of the topics within the document by taking into account the probabilistic approaches (eliminating the Lim#3 and #4).

Basic version of *analysing* the part of proposed Algorithm of Two-Level Modelling and Analysis of the LSR includes 7 steps (figure 1). Each level additionally assumes a preliminary *modelling* stage (are not included to the figure 1).

As a sample for case study experiments the Polish-language film reviews from the filmweb.pl are used. For demonstration of the basic workability of the author's Algorithm, as a *preliminary case study* was used (the data set of only one, randomly chosen, Polish-language film review, which contains 7 CF). All words/terms of film reviews in this paper will be presented in Polish and English languages (separated by

symbol "/"). The experimental part of all steps of author's Algorithm has been implemented in Python 3.4.1.

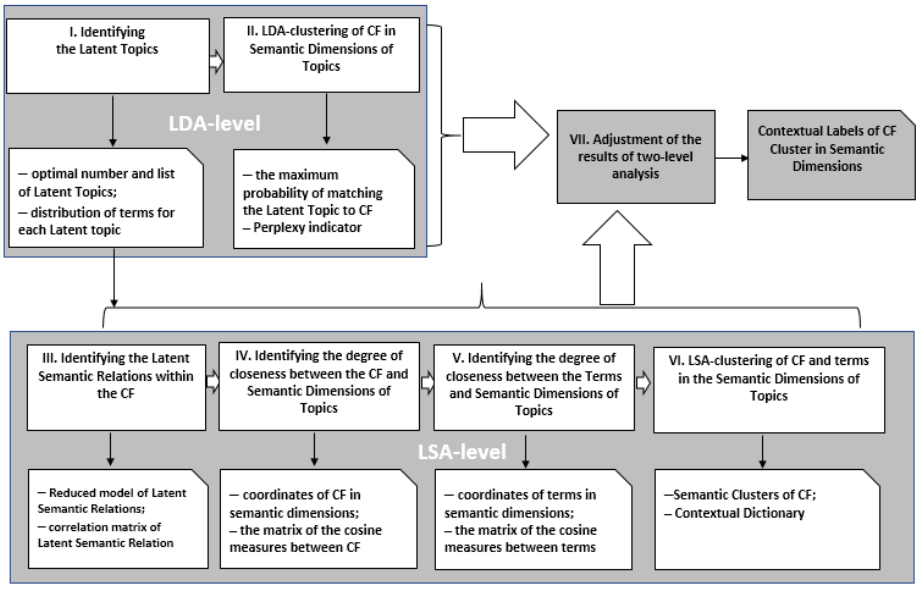

Fig.1 The Steps of the Algorithm of Two-Level Analysis of Latent Semantic Relations
Source: own research results

3.2 The Level of LDA-based Modelling and Analysis of Latent Semantic Relations

LDA-based Modelling of LSR

LDA-based Modelling of LSR is the stage, which *aims* to ensure the implementation of the level of LDA-based Topic Analysis, presupposed the Forming the "Bag of Words" (preprocessing) step.

Taking into account the Specificity of chosen Document Type, as well as the case study language peculiarities (limited number of existing algorithms and software implementations for the analysis of texts in Polish) [22], in addition to standard procedures for text preprocessing, the authors have provided:

– text *adaptation* procedure, based on the specificity of the structure of reviews document layout (Figure 2). This procedure is to implement the replacement of the Film's Titles, the Names/Surnames of Director/Actors/Characters into the corresponding position of the descriptive part of review (for example, the Title of the film is replaced by "*Film*", Name and Surname of the actor – by "*Actor*" etc.);

– expanding by authors the list of *stop words* (near 400 Polish words) for improving the process of lemmatization (based on the dictionary *pyMorfologik* [13, 16, 22]);

– *part-of-speech* (adjective, nouns, verbs) morphological tagging and filtering procedures performing, allowed to increase the resolution of the of LSR analysis.

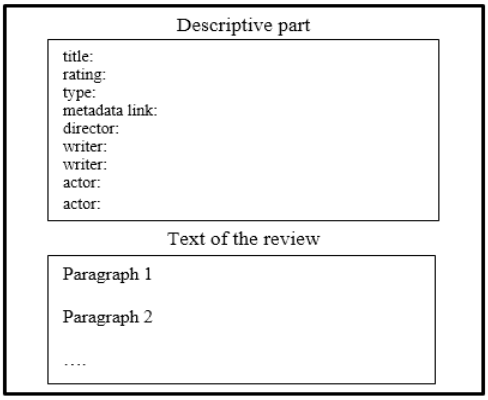

Fig 2. Structure of Text Reviews Layout
Source: own research results

LDA-based Analysing of Latent Semantic Relations

Step I. Identifying the Topics

LDA-based Analysis is the stage, which *aims*: 1) to reveal the optimal number of latent probabilistic topics that describe the main content of the analyzed document; 2) to assign them to the CFs based on the probalistic LSR within the paragraphs. As a technical support, for the implementation this phase the LDA Gensim Python package (https://radimrehurek.com/gensim/ models/ldamodel.html) was used.

Table 1 demonstrates the pretesting experiments results of preliminary case study (further – *PCS results*) of the main parameters of LDA model. The optimum value of the Perplexity index is achieved in the point, when further changes in the parameters do not lead to its significant decrease. In accordance with author's algorithm, obtained optimal number of latent probabilistic topics will be used as a recommended number of semantic clusters in the LSA-based level of SLR analysis.

Table 1. PCS results of the Studying of the of LDA Model Parameters

Perplexity	Number of Topics	Number of Terms	Number of Passes	Alpha Parameter	Eta Parameter	Max Probability Topic	Max Probability of Terms in the Topics
3336	10	10	100	1.70	1.00	0.1025	0.057
633	7	7	100	1.50	1.00	0.6050	0.177
202	5	5	100	1.50	1.00	0.7134	0.167
64	3	5	100	1.50	1.00	0.8417	0.132
63	**3**	**7**	**100**	**1.50**	**1.00**	**0.8411**	**0.166**

The list of obtained latent probabilistic topics with information about most probable (significant) terms, described this topic, is presented in the Table 2.

Table 2. PCS results of the List of Latent Probabilistic Topics with Distribution of Terms

Terms (Polish/ English)	Probability	Terms (Polish/ English)	Probability	Terms (Polish/ English)	Probability
Topic #0		**Topic #1**		**Topic #2**	
fabuła / story	0.080	kino / cinema	0.109	bohater / character	0.166
akcja / action	0.062	twórca / creator	0.066	gra / playing	0.140
efekt / effect	0.050	kobieta / woman	0.062	dobry / good	0.130
bohater / character	0.047	obsada / cast	0.052	postać / character	0.090
ksiazka / book	0.046	scena / stage	0.051	rola / role	0.040
obraz / image	0.044	glowny / main	0.050	typowy / tipical	0.030
historia / history	0.042	reżyser / director	0.049	intryga / intrigue	0.029

Step II. LDA-clustering of CF in Semantic Dimensions of Corpus

Based on information about the maximum probability of matching the obtained Latent Probabilistic Topics to the CF, on this step the process of Semantic (topical) clustering of CF could be performed. The PCS results of this process are presented in Table 3.

Table 3. PCS results of the Semantic Clustering of CF

CF	CF_5	CF_0	CF_1	CF_4	CF_6	CF_2	CF_3
# topic (cluster)	0	1	1	1	1	2	2
Probability	0.8411	0.6228	0.8022	0.7039	0.4800	0.7957	0.6603

The values of the Perplexity in the Table 1 proves the validity of the assumptions about providing the analysis the Corpora by paragraphs (*Hypothesis H1.2*). But, on the other hand, we can note, that the level of probability of a CF belonging to a particular topic/cluster is not significant for all CF (for example, for CF_6 it is lower than 0.5).

3.3 The Level of LSA-based Modelling and Analysis of Latent Semantic Relations

LSA-based Modelling of Latent Semantic Relations

LSA-based Modelling of LSR is the stage, which *aims* to ensure the implementation of the level of LSA-based Analysis of Latent Semantic Relations. As well as LDA-based level, this stage presupposed the preprocessing procedure, which contain additionally to forming the "Bag of Words", the Creating the Term-Document Matrix (TDM) step [20-22]. The fragment of the PCS results of LSA initial data building is presented in Table 4.

As for results of TF-IDF transformation of this matrix, we can state the following facts: differences in absolute term frequencies were reduced; frequently appearing

terms are less relevant compared to infrequent terms; terms-CF matrix contains weighted term frequencies.

Table 4. The fragment of PCS results of the Absolute Frequency Terms-CF Matrix

Terms (Polish/ English)	CF_0	CF_1	CF_2	CF_3	CF_4	CF_5	CF_6	Sum
bohater / character	1	1	4	5	2	2	1	16
film / movie	0	2	1	0	0	1	1	5
akcja / action	0	1	0	2	1	3	2	9
kino / cinema	1	3	0	2	1	0	2	9
kobieta / woman	0	3	0	0	1	0	0	4
główny / main	1	2	1	0	0	0	0	4

However, according to [26], and during a number of author's experiments, the solutions were found: TF-IDF approach does not work well because when a CF contains only a 100-150 words, there are seldom terms that occur more than once within a document; but, the most common words occurred within one CF are the so-called key terms, which determine the topic's label of analysed CF in a large scale; it is more important to focus on the allocation of stop words and most significant part-of-speech, to maximise the weight of keywords of the CF by excluding consideration of the terms that have no semantic weight.

LSA-based Analyzing of Latent Semantic Relations

LSA-based Analysis of LSR is the stage, which *aims* to identify the patterns in the relationships between the terms and latent semantic topics. As we already stated, LSA method is based on the principle that terms that are used in the same contexts tend to have similar meanings. For revealing this information about LSR between topics and CF/terms, we need: to assess the degree of semantic correlation relationship between CF/terms via building the reduced model of LSR; to form the semantic clusters of CF *via* determining the cosine distance between the CF in order to identify the LSR between topics and CF; to form the contextual dictionary of semantic clusters of CF *via* determining the cosine distances between the terms in order to identify the LSR between *k* terms and topics.

Step III. Identifying the Hidden Semantic Connection Within the Documents

Mathematically the Reduced model, as the instrument of preliminary LSR presence identification, is the process of multiplying of SVD transformation results with chosen *k*-dimension $X_{K_{t\times d}} = U_{K_{t\times d}} \Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T$. The fragment of *PCS* results of Reduced model is presented in Table 5.

Via comparison of the red numbers in Table 5 with zero's values in the same places of Table 4 could be, as an example, identified the existence of the following phenomena of LSR:

– the term "*Film / Movie*" seems to have the presence in all CF where the word "*Bohater / Character*" appears;

– the term "*Kobieta / Woman*" seems to have the presence in the CF where the word "*Kino / Cinema*" appears.

Table 5. The fragment of PCS results of the Reduced Model for Identifying the LSR

Terms (Polish/ English)	CF_0	CF_1	CF_2	CF_3	CF_4	CF_5	CF_6
bohater / character	1.115	2.785	2.974	3.535	1.676	2.907	1.636
film / movie	0.384	0.964	0.888	1.071	0.537	0.626	0.508
dobry / good	0.162	0.406	0.401	0.481	0.234	0.338	0.225
główny / main	0.479	1.211	0.687	0.882	0.542	-0.369	0.459
kino / cinema	0.963	2.431	1.512	1.915	1.129	-0.384	0.978
kobieta / woman	0.569	1.440	0.725	0.950	0.617	-0.687	0.508

At the same time, we can observe the increasing of the values of the correlation coefficient (CC) between terms, compared the results of Tables 4 and 5 (Table 6):

Table 6. Example of PCS results of the Comparison of the CC Between Terms

Source / Terms	Absolute Frequency Terms-CF Matrix	Reduced Model for Identifying the Hidden Connection
Bohater. Film	-0.333	0.984754769
Kino. Kobieta	0.641	0.984405802

Steps IV-V. Identifying the Degree of Closeness Between the CF / Terms in the Semantic Dimensions of Topics

For measuring the level of LSR, identified on the previous step, the matrix of cosine distance between the vectors of CF and terms should be built. The *PCS* results of this estimation are presented in the Tables 7, 8.

Table 7. PCS results of the Matrix of Cosine Distance Between the Vectors of CF

	CF_0	CF_1	CF_2	CF_3	CF_4	CF_5	CF_6
CF_0	1	0.9998	0.8052	0.8403	0.9537	-0.3376	0.8764
CF_1	0.9998	1	0.8164	0.8505	0.9592	-0.3196	0.8855
CF_2	0.8052	0.8164	1	0.9981	0.9463	0.2863	0.9912
CF_3	0.8403	0.8505	0.9981	1	0.9645	0.2266	0.9975
CF_4	0.9537	0.9592	0.9463	0.9645	1	-0.0387	0.9807
CF_5	-0.3376	-0.3196	0.2863	0.2266	-0.0387	1	0.1573
CF_6	0.8764	0.8855	0.9912	0.9975	0.9807	0.1573	1

Table 8. The fragment of PCS results of the Matrix of Cosine Distance Between the Vectors of Terms

	akcent / accent	akcja / action	bohater / character	…	łatwo / easily	osiągać / reach
akcent	1	0.9938	0.6136	…	0.873	0.1269
akcja	0.9938	1	0.6978	…	0.8132	0.2367
bohater	0.6136	0.6978	1	…	0.1506	0.8611
…						
łatwo	0.873	0.8132	0.1506	…	1	-0.373
osiągać	0.1269	0.2367	0.8611	…	-0.373	1

Step VI. LSA Clustering of CF / Terms in the Semantic Dimensions of Topics

Based on the matrices of cosine distances between the vectors of CF and terms, in this step the Semantic clustering process should be realized. An example of the implementation of *k*-means clustering [12, 22] algorithm for CF and terms (in the condition of LDA-based number of SC) is presented in the Tables 9-10 and figure 3.

Table 9. PCS results of the Labels of Contextual Fragments' Clustering

CF	CF_0	CF_1	CF_5	CF_2	CF_3	CF_4	CF_6
Cluster	0	0	1	2	2	2	2

Table 10. PCS results of the Contextual Dictionary of Semantic Clusters

Terms (Polish/ English)	Cluster	Terms (Polish/ English)	Cluster	Terms (Polish/ English)	Cluster
fabuła / story	0	reżyser / director	1	bohater / caracter	2
akcent / accent	0	kino / cinema	1	dobry / good	2
scenariusz / script	0	kobieta / woman	1	film / movie	2
akcja / action	0	główny / main	1	intryga / intrigue	2
ksiazka / book	0	obsada / cast	1	sposób / method	2
scena / scene	0	efekt / effect	1	typowy / typical	2
obraz / image	0	schemat / scheme	1	gra / playing	2
historia / history	0	stworzyć / create	1	rola / role	2

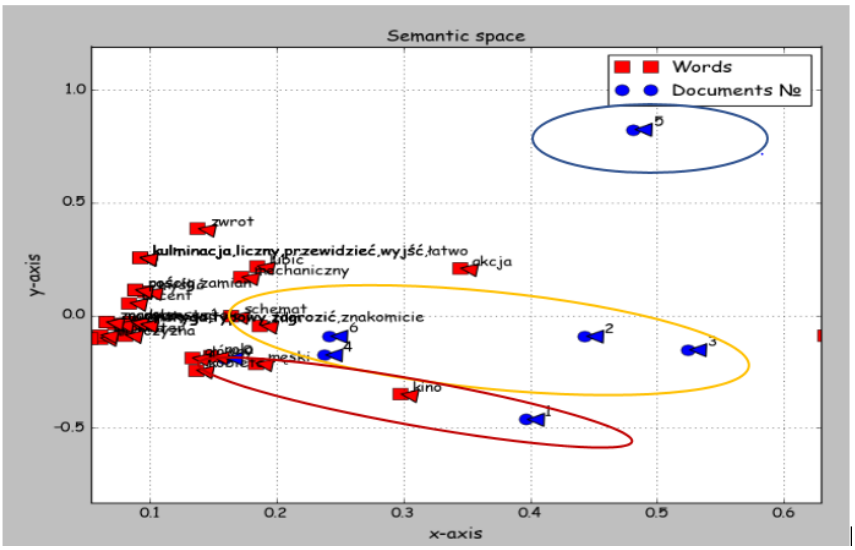

Fig.3. The Example of the Graphical Presentation of the Results of CF Semantic Clustering
Source: own research results

3.4 Adjustments of the Results of the Two Levels of Analysis

On the VII step of *Author's Algorithm*, it is supposed to combine the results of the implementation of LSA and LDA levels for analysis, namely:

1. Forming the table of the Comparison of the numerical labels of Latent Semantic Clusters of a set of CF, obtained on two levels of research (Table 11). As we can see, the results of clustering for CF_4 and CF_6, obtained in LSA- and LDA-analysis levels, do not match.

Table 11. PCS results of the Comparison of the Semantic Clusters as a set of CF Labels

LDA-level			LSA-level	
CF	**# Topic (Cluster)**	**Probability**	**CF**	**Cluster**
CF_0	1	0.6228	CF_0	0
CF_1	1	0.8022	CF_1	0
CF_2	2	0.7957	CF_2	2
CF_3	2	0.6603	CF_3	2
CF_4	**1**	**0.7039**	**CF_4**	**2**
CF_5	0	0.8411	CF_5	1
CF_6	**1**	**0.4800**	**CF_6**	**2**

2. Formulation and implementation the *Rules of Adjustments* of the results obtained in the LSA- and LDA-analysis levels.

As stated above, LDA method implementation presupposes the assignment of the corresponding topics to CF based on the largest (from existing) *probability* (P) of degree of their compliance with the analysed CF. In this connection, the author's concept of *Rules of Adjustments* (RA) of the results of Semantic Clustering of the LSA- and LDA-analysis levels for each particular CF is proposed (Table 12).

Table 12. Rules of Adjustments of CF Clustering Results

# of rule	**LSA-analysis Result**	**Result of comparison**	**LDA-analysis Result**	**LDA Probability (P)**	**Assignable Cluster**
1	LSA Cluster	=	LDA Cluster	P>0.3	LSA Cluster = LDA Cluster
2	LSA Cluster	=	LDA Cluster	P≤0.3	Cluster is Not recognized
3	LSA Cluster	≠	LDA Cluster	P≤0.3	LSA Cluster
4	LSA Cluster	≠	LDA Cluster	0.3<P≤0.7	LSA Cluster / Re-clustering
5	LSA Cluster	≠	LDA Cluster	P>0.7	LDA Cluster

These rules allow:

– to improve the quality of LDA-method recognizing the CF's topics (rules 3, 4) due to the possibility of correcting the results of clustering, which are characterized by the low level of probability of a CF belonging to a particular topic. Suggested instrument – latent semantic specificity of the LSA method;

– to improve the quality of LSA-method recognition of hidden relations between the CF (rules 2, 5) due to the possibility of correcting the results of clustering, which characterize by situations, when CF coordinates located on the cluster's boundary. Suggested instrument – the probabilistic characteristics of the LDA method.

The *PCS* results of the implementation of RA are presented in Table 13.

Table 13. PCS results of the of Final Version of the Labels of the CF's Semantic Clusters

CF	CF_5	CF_0	CF_1	CF_4	CF_2	CF_3	CF_6
# topic	0	1	1	1	2	2	2

4 Case Study Results and Discussion

For the process of verification of the *author's Algorithm* was formed the sentimental structure of FRC via classification of the reviews collection on the Subjectively Positive (SPSC) and Subjectively Negative Sentiment Corpuses (SNSC). This procedure is realized on the basis of information on the subjective assessment (SA) of films by the reviewers (measured by 10-point scale).

As a condition of sentimental structure of FRC building, the following *Heuristic 1.3* was adopted: *to consider the SPSC, if the SA is more than 5 points, and SNSC – if it is equal or less than 5 points.*

During the verification, the 30 reviews from each reviews collection were analysed. Totally 208 paragraphs from SPSC and 260 paragraphs from SNSC were studied. The recommended number of clusters (identified in LDA-level of analysis) is equal to 4. The structure (percentage of paragraphs, belonging to the topic) of the Semantic Clusters in each separate level and after adjustment (LSA&LDA) is presented in Table 14. The Contextual labels (CL) of the Topics were assigned automatically on the bases of the terms with the highest frequency in each topic.

Table 14. The Structure of the Semantic Clusters

SPSC				SNSC			
CL of the Topics	LSA, %	LDA, %	LSA&LDA, %	CL of the Topics	LSA, %	LDA, %	LSA&LDA, %
Bohater / Character	19.71	18.75	19.23	Bohater / Character	11.54	13.46	12.31
Reżyser / Director	32.21	36.06	33.65	Aktor / Actor	30.00	30.38	29.23
Scenariusz / Scenario	17.79	12.50	16.83	Widz / Spectator	28.85	26.92	28.08
Fabuła / Story	30.29	32.69	30.29	Fabuła / Story	29.62	29.23	30.38

The quantitative indicators of the adjustments process of the Latent Semantic Relations Analysis results: percentage of not recognized CF inside the Topic (*Indicator 1*); percentage of CF, which changed the Cluster (*Indicator 2*) and as well as final qualitative characteristic of research (*Recall rate*) are given in Table 15.

Table 15. The Quality of the of LSR Analysis Results

SPSC			SNSC		
Labels of the Topics	**Indicator 1**	**Indicator 2**	**Labels of the Topics**	**Indicator 1**	**Indicator 2**
Bohater / Character	7.50	5.56	Bohater / Character	9.23	6.25
Reżyser / Director	2.82	5.48	Aktor / Actor	1.27	5.13
Scenariusz / Scenario	3.17	12.00	Widz / Spectator	5.52	9.09
Fabuła / Story	6.11	7.81	Fabuła / Story	2.61	2.70
Recall rate		95.19	**Recall rate**		96.15

5 Conclusions

In this paper authors presented the complex two-level Algorithm of Modelling and Analysis of TMP, aimed at elimination the *Limitations* characterizing the of two mathematical frameworks and taking into account the Document Type *Specificity*. The answers for the main scientific research question were found: the combination of the Discriminant and Probabilistic Methods (*Hypothesis* H2.1) as well as Specificity of the Argumentative Type Document oriented approach (*Hypothesis* H1.2), gave the opportunity to improve the following qualitative characteristics of LSR Analysis:

– recall rate (the ratio of the number of semantically clustered/recognized paragraphs to the total number of paragraphs in the corpora) to 90-95%;

– precision indicator (the average probability of significantly clustered/recognized paragraphs) from 62 to 70-75%.

In the *future research*, these results are planned to be used: to evaluate the Algorithm effectiveness for processing the English language Documents; to develop the algorithm of forming the hierarchical structure of the Latent topics of Corpora with taking into account the Sentiment specificity.

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 D. (2012) Introduction to Probabilistic Topic Models. Comm. ACM 55 (4), April, 2012: pp. 77-84
5. Ali D., Juanzi L., Lizhu Z., Faqir M. (2010) Knowledge discovery through directed probabilistic topic models: a survey. In Proceedings of Frontiers of Computer Science in China. pp. 280-301.
6. Blei D. Topic modeling. http: //www.cs.princeton.edu/~blei/topicmodeling.html
7. 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

8. Deerwester S., Susan T. Dumais, Harshman R. (1990) Indexing by Latent Semantic Analysis. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf
9. Eden L. (2007) Matrix Methods in Data Mining and Pattern Recognition, SIAM.
10. 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
11. Salton G., Michael J. (1983) McGill Introduction to modern information retrieval. New York McGraw-Hill - McGraw-Hill computer science series, XV, 448 p
12. Jain A.K., Murty M.N., Flynn P.J. (1999) Data Clustering: A Review; ACM Computing Surveys, Vol. 31, Nr. 3.
13. Gramacki J., Gramacki A. (2010) Metody algebraiczne w zadaniach eksploracji danych na przykładzie automatycznego analizowania treści dokumentów. XVI Konferencja PLOUG, pp.227-249.
14. 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.
15. Canini KR., Shi L., Griffiths T. (2009) Online Inference of Topics with Latent Dirichlet Allocation. Journal of Machine Learning Research. Proceedings Track 5: 65-72.
16. Tomanek K. (2014). Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych, Przegląd Socjologii Jakościowej, pp. 118-136, www.przegladsocjologiijakosciowej.org
17. Aggarwal C., Zhai X, (2012) Mining Text Data (Springer).
18. Leticia HA.(2011). Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers, Doctor of Philosophy (Management Science), 226 pp
19. 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.
20. 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
21. 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.
22. 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).
23. Rui X., Donald C. Wunsch II. (2005) Survey of clustering algorithms. IEEE Transactions on Neural Networks 16(3): pp. 645-678.
24. 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
25. Hofman T. (1999) Probabilistic Latent Semantic Analysis. UAI, 1999, 289-296; Thomas Hofmann. Probabilistic Latent Semantic Indexing. SIGIR, pp. 50-57.
26. Mika T. (2013) Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion. PhD Thesis, Series of Publications A, Report A-2013-1.