Peer-reviewed version available at *Information* **2018**, *9*, 307; doi:10.3390/info9120307

Article

Improving the Accuracy in Text Classification Methodology in Light of Modelling the Latent Semantic Relations

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

[1] Gdansk University of Technology; nina.rizun@zie.pg.gda.pl
[2] Alfred Nobel University, Dnipro; taranenkonew@gmail.com
[3] Gdansk University of Technology; wowal@eti.pg.gda.pl

* Correspondence: nina.rizun@zie.pg.gda.pl; Tel.: +48-575-434-778

† This manuscript is an extended version of our paper "The Algorithm of Modelling and Analysis of Latent Semantic Relations: Linear Algebra vs. Probabilistic Topic Models" published in the proceedings of Knowledge Engineering and Semantic Web, Szczecin, Poland, 8–10 November 2017

Abstract: The research presents the Methodology of Improving the Accuracy in Text Classification in Light of Modelling the Latent Semantic Relations (LSR). The aim of this Methodology is to find the ways of eliminating the Limitations of Discriminant and Probabilistic methods for LSR revealing and customizing the Text Classification Process to the more accurate recognition of the text tonality. This aim should be achieved by using the knowledge about the text's Hierarchical Semantic Context in the form of Corpora-based Hierarchical Sentiment Dictionary. The main scientific contribution of this research is the following set of approaches to improve the qualitative characteristics of Text Classification process: combination of the Discriminant and Probabilistic methods allowing to decrease the influences of the Limitations of these methods on the LSR revealing process; considering each document as a complex structure allowing to estimate documents integrally by separated classification of topically completed textual component (paragraphs); taking into account the features of Argumentative type of documents (Reviews) allowing to use the author's subjective evaluation of text tonality for development the Text Classification methodology. Tonality, expressed by the Review's author, has a significant, but not critical, effect on the qualitative indicators of Sentiment Recognition.

Keywords: Text Classification; Topic Modelling; Latent Semantic Analysis; Latent Dirichlet Allocation; Hierarchical Sentiment Dictionary, Contextually-Oriented Hierarchical Corpus; Text Tonality; Evaluation

1. Introduction

The rapid development of computer technology and the Internet space in recent decades has led to the fact that the procedures for creating and accessing the information content of many web resources have become an integral part of private and professional activities of a person. The content of information resources such as social networks, feedback services, web forums and blogs, is actively formed by the users themselves and is publicly available.

On the other hand, this content, as well as some more official information (for example, financial statements of enterprises, scientific and news articles) forms a large array of unstructured text information containing a huge amount of Explicit and Hidden knowledge.

One of these types of knowledge is the Latent Semantic Relations (LSR), hidden both inside the documents and between them in order to identify the context of the analyzed document, as well as to classify a group of documents based on their semantic proximity. Modelling and Analysis of Latent Semantic Relations (LSR) – is the approach of constructing a model, 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 [1, 2, 3].

In addition to the Latent Semantic analysis, an important task of the classification of texts is to identify their emotional coloring. A special section of computer linguistics is devoted to extraction of such information – automatic analysis of text tonality (Sentiment Analysis or Opinion Mining) [4]. The initial goal of Sentiment analysis methods was classification of documents, and later of sentences, according to a given scale of tonality, usually a two-point (positive-negative) or three-point (positive-negative-neutral). However, instead of a general assessment of tonality, a more detailed study of the expressed views on specific aspects (contexts) is required. Therefore, over time, the initial formulation of the task of tonality analysis has acquired a more detailed formulation and has emerged as a separate problem of contextually-oriented sentiment analysis, which is to automatically determine the views of the user, expressed in the text, with respect to specific aspects being examined.

This research is devoted to finding ways of Improving the Accuracy in Text Classification via effective implementation of Modelling the Latent Semantic Relations approaches for extracting the knowledge about the documents Semantic Context and further representation, and using this knowledge in the form of Corpora-based Hierarchical Contextually-Oriented Sentiment Dictionary.

This paper is an *extended* version of [5], and the following section were added:

- extended version of Case Study Results and Discussion section for Latent Semantic Relations Revealing Phase;
- representation of the new stage of research, based on the results obtained in the paper [5]. In this regard, a new section has been added describing the methodological and experimental parts of the Phase of Text Classification Based on the Contextually-Oriented Sentiment Dictionary.

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 analyzing the data has been made by researchers in the field of information retrieval (IR) [6-8]. The basic methodology proposed by IR researchers for text collection reduces each document in the corpus to a vector of real numbers, each of them representing ratios of counts.

In the popular $TF \times IDF$ scheme [9-13], 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 wth word occurrence in document t:

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

$k(w,L_t)$ – the number of wth word occurrences in the text t; df – 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 their relation to the same topic, different metric can be applied. The most appropriate metric is cosine measure of the edge between the vectors:

$$dist_{t_i} = \cos\theta = \frac{x \cdot y}{\|x\| \cdot \|y\|}, \quad (3)$$

where $x \cdot y$ – scalar product of the vectors, $\|x\|$ and $\|y\|$ – quota of the vectors, which are calculated by the formulas:

$$\|x\| = \sqrt{\sum_{i=1i}^{n} x_i^2}, \|y\| = \sqrt{\sum_{i=1i}^{n} y_i^2}, \quad (4)$$

Further part of the algorithm is to divide the source data into groups corresponding to the events, as well as to determine whether a text document describes a set of any topic. The main idea of the solution is the use of clustering algorithms [3, 14, 9-13].

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 [3, 7, 15].

2.2. Latent Semantic Indexing

In 1988, Dumais et al. [16] proposed a method of Latent Semantic Indexing (LSI), most frequently referred to as LSA. Deerwester et al. in 1990 [17], designed to improve the efficiency of IR algorithms and search engines by 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 [7]. As a result of the SVD of the matrix A we have three matrices:

$$X_{t\times d} \approx X_{K_{t\times d}} = U_{K_{t\times d}} \Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T, \quad (5)$$

$\Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T$ – 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 [9, 10] 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 [16, 17, 18]. 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, 24].

2.3.1. Latent Dirichlet Allocation

LDA – generative probabilistic graphical model proposed by David Blei [1, 2, 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 word in a document; randomly select a topic from the distribution θ_d, obtained in the first step; randomly

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

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.

2.3.2. Methods of Evaluating the Quality of Results

The most common method of evaluating the quality of probabilistic topic models is calculation of the Perplexity index on the test data set D_{test} [1, 2, 20, 21]. In information theory, perplexity is a measurement of how well a probability model predicts a sample. Low perplexity indicates that the probability distribution is good at predicting the sample:

$$Perplexity(D_{test} = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\}, \quad (6)$$

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) [1, 2, 22].

2.4. Textual Content Classification

Methods of contextually-emotional analysis of the text are developed within the framework of two machine learning approaches: supervised and unsupervised machine learning [4]. In the approach based on supervised machine learning, a marked collection of documents is needed, which lists examples of emotional expressions and aspect terms.

The methods of unsupervised machine learning allow to avoid dependence on training data. For their work, one also needs a Corpus of documents, but preliminary markup is not required. Within the framework of this approach, the probabilistic-statistical regularities of the text are found and, on their basis, the key subtasks of the aspect-emotional analysis are solved: identification of aspect terms and determination of their tonality. However, such methods require complex tuning to a given domain. For example, the method based on Latent Dirichlet Allocation (LDA) in its original form is not able to effectively detect topics, therefore, its additional adaptation and adjustment of correspondence of identified topics to the target set of contexts is required [23].

The methods of Text Classification, considered above, require the presence of Sentiment Dictionary of text tonality evaluation. There are three basic approaches to such Dictionary [4]: expert; based on dictionaries / thesaurus; and on the basis of text collections.

With the expert approach, the dictionary is compiled by experts. The approach differs, on the one hand, by complexity and high probability of the absence of domain-specific words in the dictionary, on the other – by high quality of the dictionary in sense of adequacy of the assigned key.

In the dictionaries / thesaurus approach, the initial small list of evaluation words is expanded by various dictionaries, for example, explanatory or synonyms / antonyms. This also does not take into account the subject area.

In the approach based on text collections, statistical analysis of the marked texts, as a rule, belonging to the subject domain in question, is used to compile the Dictionary.

In [24], the dictionary of emotional vocabulary, compiled by experts manually, was used to determine the tone of individual words. In the dictionary, each word and phrase are associated with orientation of the key (positive / negative) and with strength (in points).

The author's methods proposed in [25, 26] are based on a dictionary approach: to determine the tonality of texts, a dictionary of estimated words is used, where each word has a numerical weight that determines the degree of word significance. In the method of working with the dictionary closest to the paper [27]), however: the dictionary firstly is created on the basis of a statistical analysis of training collection; secondly, the weight of words is determined with the help of a genetic algorithm.

In most studies, tone of the text is determined on the basis of calculation of weights of the appraisal words included in it:

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

$$W_T^C = \sum_{i=1}^{N_C} |w_i|, \quad (7)$$

where W_T^C – weight of text T for tonality C; w_i – weight of the evaluated word i; N_C – number of estimated bigrams of tonality *C* in the text *T*.

To classify texts according to the linear function:

$$f\left(W_T^{pos}, W_T^{neg}\right) = W_T^{pos} + k_{neg} \bullet W_T^{neg}, \quad (8)$$

where W_T^{pos} is the positive weight of the text *T*; W_T^{neg} is the negative weight of the text *T*; k_{neg} is the coefficient compensating the fact of preponderance of positive vocabulary in text [28]. If the value of the function *f* is greater than zero, the text is positive, otherwise – negative.

3. Methodology

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

1. *Term* is a basic unit of discrete data.
2. *Contextual Fragment* (CF) is an indivisible, topically completed sequence of terms, located within a document's paragraph.
3. *Document* is a set of CF.
4. *Corpus* (films reviews corpus, FRC) is a collection of Documents.
5. *Topic* is the Label (one term) that defines the main semantic context of the CF.
6. *Contextual Dictionary* (CD) is a set of key words that describe semantic context of the Topic.
7. *Semantic Cluster* (SC) is the set of CF that have hidden semantic closeness (HSC).
8. *Contextually-Oriented Corpus* (HC) is a Hierarchical structure of semantically closes CF, built via application of unsupervised machine learning Discriminant and Probabilistic Methods of the Topic Modelling and Latent Semantic Relations Analysis.
9. *Corpora-based Sentiment Dictionary* (CBSD) is a Manually Created Dictionary, which has Semantic and Hierarchical Structure thanks to using the Contextually-Oriented Corpus for its building.

3.1. Novelty and Motivation

Motivation scenario of this research presupposes taking into account the *Specificity* of the analyzed Document Type and concerns finding the ways to completely or partially:

- Eliminate the Limitations characterizing the Discriminant and Probabilistic approaches for Latent Semantic Relations revealing;
- Customize the Text Classification Process to the more accurate recognition of the text tonality in light of Semantic Context of the topic.

In this regard the following scientific research questions (RQ) were raised:

RQ_1. Whether taking into account the specific features of Argumentative/ Persuasive type of document allows to affect Quality of the Topic Modelling Process Results.

RQ_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?

RQ_3. Whether taking into account Hierarchical structure of Latent Semantic Relations within the Corpus allows to affect Accuracy of the Text Classification Results.

RQ_4. Is it possible to increase the Text Classification Process Accuracy via building and using the Contextually-Oriented and Semantically Structured Sentiment Dictionary?

For finding the answers to these questions the following main Assumptions (A) were formulated:

A1. Taking into account the specificity of Type of Documents, chosen for this study, and presence of the nonofficial requirements of Film's Review structure and writing rules [29], assume that the writing style of each review is approximately the same (eliminating the Lim#1).

A2. Taking into account the chosen Document Type Specificity, assume that each document has a complex structure and can be estimated integrally by separated classification of topically completed textual component (paragraphs), centered on a single Topic, as elements of their structure (eliminating the Lim#2).

On the basis of the research questions and proposals raised, the following scientific Hypotheses (H) were formulated:

H1. Combination of the unsupervised machine learning Discriminant and Probabilistic methods has 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:

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

H2. Identifying and taking into account the Hierarchical structure of Latent Semantic Relations within the Corpus effect to improve the Text Classification Process Accuracy. This effect is expected to be achieved via increasing:

- Adequacy of Tonality Assessment Instruments via building the Manually Creating Hierarchical Contextually-Oriented and Semantically Structured Corpora-based Sentiment Dictionary;
- Quality of the Sentiment Analysis results via adjusting the Algorithms of using the Tonality Assessment Instruments by applying integral evaluation of its individual topically-oriented fragments using the CBSD and taking into account the tonality, subjectively assigned to texts by the author.

Proposed Methodology of Improving the Accuracy of Text Classification based on revealing and using the knowledge about Latent Semantic Relations includes 2 main phases:

- Latent Semantic Relations revealing Phase;
- Text Classification based on the CBSD Phase.

As a sample for case study experiments the Polish-language film reviews dataset from the filmweb.pl was used. The experimental part of author's Methodology has been implemented in Python 3.4.1.

3.2. Latent Semantic Relations Revealing Phase

Basic version of Latent Semantic Relations Revealing Phase includes 7 steps (figure 1).

3.1.1. LDA-based Analyzing of Latent Semantic Relations Layer

Step I. Identifying the Topics

LDA-based Analysis of LSR is the layer, 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 probabilistic LSR within the paragraphs.

As a technical support, for the implementation this phase the LDA Genism Python package (https://radimrehurek.com/gensim/ models/ldamodel.html) was used. For demonstration of the basic workability of the Latent Semantic Relations revealing phase, as a *preliminary case study* (PCS) was used (the data set of only one, randomly chosen, Polish-language film review, which contains 7 CF).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

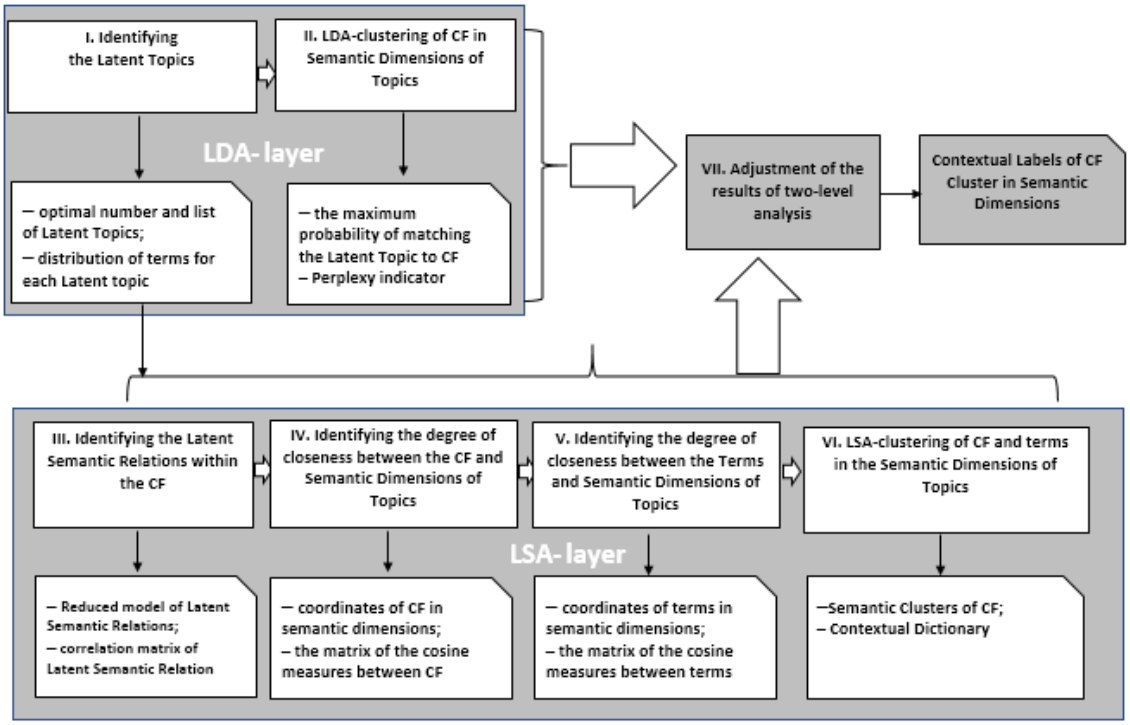

Figure 1. The steps of the Latent Semantic Relations Revealing Phase. *Source*: own research results

Table 1 demonstrates the pretesting experiments results of preliminary case study 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	0.102	0.057
633	7	7	100	1.50	1.00	0.605	0.177
202	5	5	100	1.50	1.00	0.713	0.167
64	3	5	100	1.50	1.00	0.841	0.132
63	3	7	100	1.50	1.00	0.841	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	Probability	Terms	Probability	Terms	Probability
Topic #0		Topic #1		Topic #2	
story	0.080	cinema	0.109	character	0.166
action	0.062	creator	0.066	playing	0.140
effect	0.050	woman	0.062	good	0.130
character	0.047	cast	0.052	character	0.090
book	0.046	stage	0.051	role	0.040
image	0.044	main	0.050	typical	0.030
history	0.042	director	0.049	intrigue	0.029

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

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 assumption A2 about providing the analysis the Corpora by paragraphs. 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).

3.1.2. The LSA-based Analysis of Latent Semantic Relations Layer

LSA-based Analysis of LSR is the layer, 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 phenomena of LSR:

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

Terms	CF_0	CF_1	CF_2	CF_3	CF_4	CF_5	CF_6	Sum
character	1	1	4	5	2	2	1	16
movie	**0**	2	1	**0**	**0**	1	1	5
good	0	1	0	2	1	3	2	9
main	1	3	0	2	1	0	2	9
cinema	**0**	3	0	**0**	1	0	**0**	4
woman	1	2	1	0	0	0	0	4

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

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

Term "Movie" seems to have the presence in all CF where the word "Character" appears

Term "Woman" seems to have the presence in the CF where the word "Cinema" appears

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

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
Character. Movie	-0.333	0.985
Cinema. Woman	0.641	0.984

Steps IV-VI. Identifying the Degree of Closeness Between the CF / Terms in the Semantic Dimensions of Topics. LSA Clustering of 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. 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 [18, 30] algorithm for CF and terms (in the condition of LDA-based number of SC) is presented in the Tables 7-8.

Table 7. 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

3.1.3. 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 9). 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 9. 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 analyzed 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 10).

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,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

when CF coordinates located on the cluster's boundary. Suggested instrument – the probabilistic characteristics of the LDA method.

Table 10. Rules of Adjustments of CF Clustering Results

Rule	LSA-analysis Result	Comparison Result	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

The PCS results of the implementation of Rules of Adjustments are presented in Table 11.

Table 11. 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

3.1.4. Case Study Results and Discussion

For the process of verification of the author's Methodology in this phase 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 evaluations of their tonality (SE) of films by the reviewers (measured by 10-point scale). We consider the SPCS films reviews if the subjective review's assessment is more than 5 points, and SNCS – if it is equal or less than 5 points.

During the verification, the 5000 Polish-language films reviews (2500 SNCS and 2500 SNSC) reviews were analyzed. As a result, two-level Contextual Hierarchical structure of Topics (CHST) was defined (Table 11). The recommended number of clusters (identified in LDA-level of analysis):

- at the 1st level of hierarchy is equal to 5 for SNCS and is equal to 4 for SNSC;
- at the 2nd level of hierarchy is equal to 4 for SNCS and is equal to 3 for SNSC.

Table 11. PCS results of the of Final Version of the Labels of the CF's Semantic Clusters

Topics of the 1st level	Topics of the 2nd level	LSA&LDA, %	Topics of the 1st level	Topics of the 2nd level	LSA&LDA, %
Hero	Actor / Play	24	Hero	Action / History	49
	History / Film	43		Director / Cinema	21
	Picture / Scene	30		Scene / Actor	31
	Director / Creator	3	Actor	Hero / Image	24
Director	Film / Director	30		Role / Scene	58
	Scene / Story	10		Script / History	18
	Style	6	Creator	Hero / Scene	23
	Creator / Author	54		Film / Script	60
Script	Film / Director	8		Picture / Actor	18
	Story / Hero	58	Plot	Story / Hero	39
	Author / Creator	13		Director / Image	18
	Role / Actors	21		Creator / Film	43
Plot	Film / Effects	5			
	Portrait / Image	31			
	Director / Production	24			

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

Spectator	Script / History	40
	Hero / Fan	40
	Film / Aspects	20
	Role / Formulation	16
	Scene / Director	24

The Hierarchical structure of the Contextually-Oriented Corpus (HC), created as a two-point (Positive/Negative Classes) structure of the sets of Paragraphs, semantically close to revealed Topics with Contextual Dictionaries (for each separate layer and after adjustment – on the 1st level of Topics) is presented in Table 12 [30, 31]. The Contextual labels (CL) of the Topics were assigned automatically on the bases of the terms with the highest frequency in each topic.

Table 12. The Hierarchical structure of the Contextually-Oriented Corpus

CL of the 1st level Topics	SPSC				SNSC		
	LSA, %	LDA, %	LSA&LDA, %	Topics of the 1st level	LSA, %	LDA, %	LSA&LDA, %
Hero	29.05	23.50	32.50	Hero	35.10	38.40	37.30
Director	15.80	12.70	10.30	Actor	19.30	20.30	18.30
Script	30.11	26.19	30.94	Creator	28.10	29.10	29.20
Plot	9.50	12.40	15.11	Plot	17.50	12.20	15.20
Spectator	15.54	25.21	11.15				

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) the 1st level of Topics are given in Table 13.

Table 13. The Quality of the of LSR Analysis Results

Topics	SPSC		Topics	SNSC	
	Indicator 1	Indicator 2		Indicator 1	Indicator 2
Hero	7.70	8.56	Hero	9.23	4.18
Director	3.84	3.44	Actor	5.30	9.42
Script	4.19	16.60	Creator	2.45	12.10
Plot	6.11	7.30	Plot	6.47	4.11
Spectator	7.19	2.55			
Recall rate		95.30		Recall rate	93.60

In this phase we can conclude that the combination of the Discriminant and Probabilistic Methods (Hypothesis H1) gave the opportunity:

- to improve the following qualitative characteristics of LSR Analysis: *recall rate* (as a ratio of the number of semantically clustered/recognized paragraphs to the total number of paragraphs in the corpora) to 90-95%; precision indicator (as the average probability of significantly clustered/recognized paragraphs) from 62 to 70-75%;
- to increase the depth of recognition of Latent Semantic Relations by providing the a mathematical and methodological basis for building the Contextual Hierarchical structure of Semantic Topics.

3.2. Text Classification Based on the Contextually-Oriented Sentiment Dictionary Phase

Basic version of Text Classification phase includes 9 steps (figure 2).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

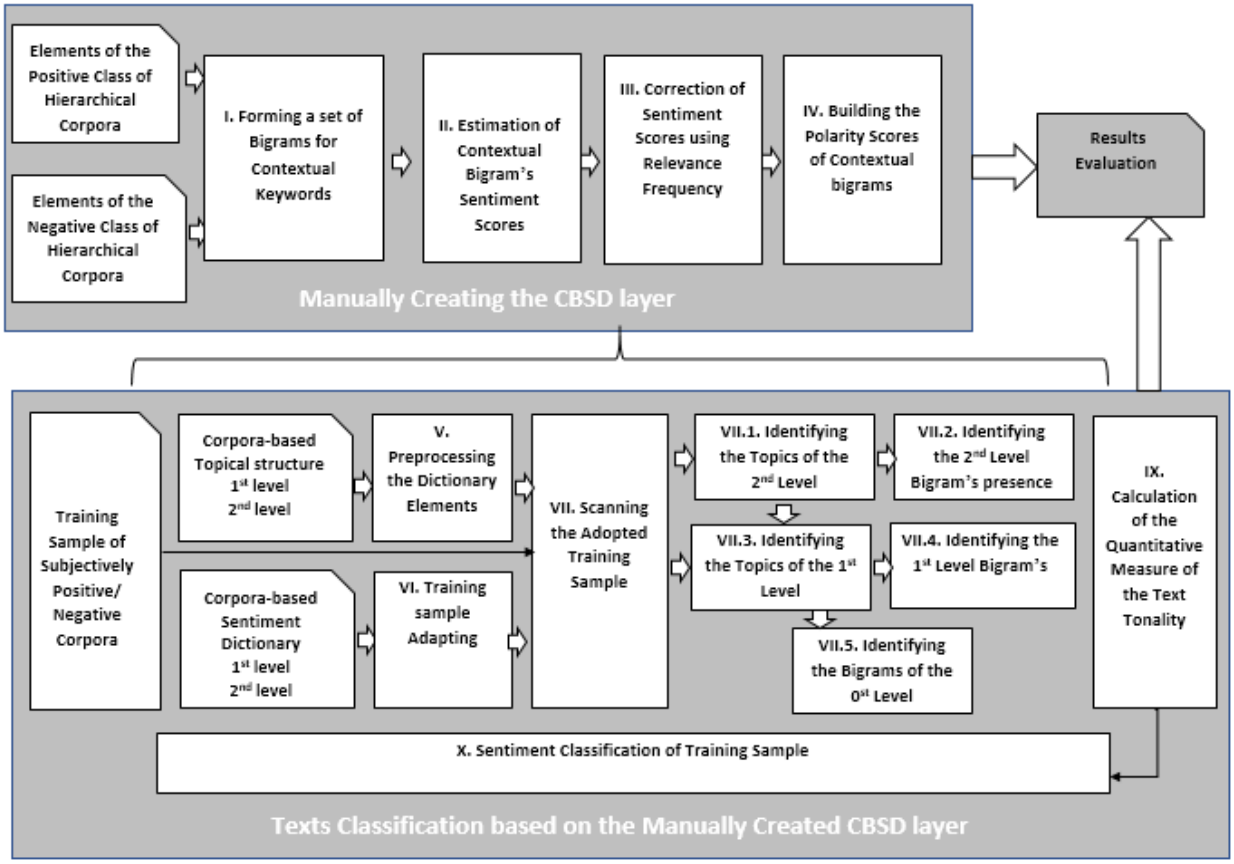

Figure 2. The steps of the Text Classification based on the CBSD phase. *Source*: own research results

3.2.1. Creating the Corpora-based Sentiment Dictionary Layer

Creating the Corpora-based Sentiment Dictionary is the layer, which *aims* to identify the Contextually Oriented Hierarchically structured set of Dictionary items (bigrams) and their Sentiment Scores, allowing to measure and evaluate the tonality of the analyzed texts with the high accuracy. One of the two components of bigram must be the elements of Contextual Dictionary of Semantic Clusters (Phase 1). CBSD should have three levels [32]:

- 0s level is the set of Dictionary items without taking into account the Contextual Hierarchical structure of Topics;
- 1st level is the set of Dictionary items with taking into account the 1st level of CHST;
- 2nd level is the set of Dictionary items with taking into account the 2nd level of CHST.

To enable the implementation of steps I-IV of the 2nd phase of author's Methodology (figure 2), only Truly Subjectively Positive (TSP) and Truly Subjectively Negative (TSN) Corpora Samples from Hierarchically structured Contextually-Oriented Corpus (Table 12-13) should be used. To consider the text if the element of TSP, if the subjective text's assessment is truly positive (more than 8), and the element of TSN – if it is truly negative (assessment less than 4 points) [31].

Definition of the Sentiment Scores of the bigrams are estimated by the frequency of occurrence of this bigram in the elements of Corpora. To increase of the degree of accuracy of the Sentiment Scores estimation the parameter to reverse the frequency – RF (Relevance Frequency) is used [33]:

$$RF_S = \log_2\left(2 + \frac{a}{\max(1,b)}\right), \quad (9)$$

where a – number of documents related to category S (positive, negative) and containing this bigram, b – number of documents not related to category S and containing this bigram as well.

The *purpose* of this layer is to evaluate the adequacy and prove the effectiveness of using hierarchical to improve accuracy of Text Classification process.

The main tasks of this layer are:

- to teach the developed Text Classification algorithm to classify the texts, based on the quantitative measures of the tonality (Sentiment Scores) and with taking into account the one and two-level Hierarchical structure of the Corpora-based Sentiment Dictionary
- to evaluate the quality of the conducted classification for the purpose of modification / improvement the applied Algorithm via comparing the results of the Text classification.

3.2.1. Texts Classification Based on the Manually Created CBSD Layer

Steps V-VI. Preparing to Perform the Sentiment Classification Procedure

In the step of Training Sample preparing, 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 [29], in addition to standard procedures for text pre-processing, the authors have provided text adaptation procedure [30].

Step VII. Scanning the Corpora Sample to Identify the Presence of Sentiment Dictionary Elements

With the purpose of acceptance / rejection of the Hypothesis 2, this step of algorithm involves the implementation of the following three procedures of scanning the Subjectively Positive/Negative Corpora Samples (SPCS/SNCS).

Procedure 1. Using CBSD without taking into account their Topical structure – Simple Classification (step VII.5);

Procedure 2. Using CBSD with taking into account their CHST – One-level classification (steps VII.3-5).

Procedure 3. Using CBSD with taking into account their CHST – One- and Two-level classification (steps VII.1-5)

As was accepted in this study as an Assumption 2, scanning and recognition of topics for One- and Two- Level Classification will be performed by paragraphs (elements of document) [5].

For realizing the Procedures 3 (with the deepest Topics Identification process) the following algorithm is developed):

Step VII.1. This step realized via scanning the Adopted Training Sample texts and identifying the topics at the 2nd Level of the CHST for each Review paragraph. This procedure is implemented by adding to Training Sample the Topic (Contextual Dictionary elements) from CHST as one of its Paragraphs and then using the LSA method to find paragraphs that have a Latent Semantic Relationships.

Step VII.2. This step realized via scanning the part of the Training Sample for which Topics at the 2nd Level were identified, with the aim to find the **bigrams** form 2nd level of CBSD which correspond to the Topic identified for each Paragraph.

Steps VII.3-4. For paragraphs for which topics not been defined in the step VII.1, these steps realized via scanning this part of Adopted Training Sample texts for identifying the topics at the 1st Level of the CHST and subsequent search the **bigrams** form 1st level of CBSD which correspond to the Topic identified for each Paragraph.

Step V. For paragraphs for which topics not been defined in the steps VII.1 and VII 3, these step realized via search the **bigrams** form 0s level of CBSD.

The rules for determining the presence of the elements of the Sentiment Dictionaries and word-modifiers in the text are presented in Table 14.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

Table 14. Rules for Detecting the Presence of Elements of the Sentiment Dictionary in the Text

Rules No	Rule	Execution result
1	Presence the elements of the bigram at a distance of no more than 3 words from each other	True
2	Presence the elements of the bigram within one sentence	True
3	Presence the elements of the bigram within one phrase, not separated by commas	True
4	The presence of word-modifiers in the immediate vicinity of the elements of the bigram	True

Step IX. Calculation of the Quantitative Measure of the Text Tonality

To determine the quantitative measure of the tonality estimate for the entire text of document T from Subjectively Corpora Samples, the number of positive N_C^{pos}, neutral N_C^{neu} and negative N_C^{neg} bigrams from the corresponding CBSD, found in Texts in accordance with the rules in Table 15, is calculated.

Corresponding to the found bigrams Polarity scores w_i^{pos}, w_i^{neu} and w_i^{neg} are corrected (if necessary) taking into account the rules for Words-modifiers and are summed up.

$$W_T^{pos} = \sum_{i=1}^{N_C^{pos}} w_i^{pos}, W_T^{neu} = \sum_{i=1}^{N_C^{neu}} w_i^{neu}, W_T^{neg} = \sum_{i=1}^{N_C^{neg}} w_i^{neg}, \quad (10)$$

where W_T – weight of text T for particular tonality; wi – Polarity score of bigram i; N_C – the number of estimated bigrams of particular tonality in the text T.

Each texts are placed in a three-dimensional estimated space (positive–neutral–negative tonality) in accordance with their scales W_T. To find the final basic estimator of the texts tonality we can according to the linear function:

$$f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right) = W_T^{pos} + W_T^{neu} + k_{neg} \bullet W_T^{neg}, \quad (11)$$

where k_{neg} is the coefficient, compensating fact of preponderance of positive vocabulary in texts [32].

Step IX. Sentiment Text Classification

The implementation of this step involves the use of the following rules:

Rule 1. Classification for each training sample will be performed in three classes respectively:

- for Subjectively Positive Corpora Sample (SPCS):

C1. Text have the High Positive tonality (HP).
C2. Text have the Quite Positive tonality (QP).
C3. Text have the Reasonably Positive tonality (RP).

- for Subjectively Negative Corpora Sample (SNCS):

C4. Text have the Rather Negative tonality (RN).
C5. Text have the Clearly Negative tonality (CN).
C6. Text have the Absolutely Negative tonality (AN).

Rule 2. To implement the training procedure for the algorithm being developed, the Sentiment Classifying of texts is suggested using basic quantitative measure of the text tonality [32]:

$$R = f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right), \quad (12)$$

Rule 3. Take into account the specificity of chosen case study, to implement the training procedure for the algorithm being developed, the Sentiment Classification is suggested using the following empirical rules for determining belonging the Text to a certain class (Tables 15-16):

Table 15. Rules for Determining Belonging the Text to a Certain Class (Actual Classes)

Positive	Left border	Right border
Review expressed is High Positive opinion	8	10
Review expressed is Quite Positive opinion	6	7
Review expressed is Reasonably Positive opinion		5
Negative	**Left border**	**Right border**
Review expressed is Rather Negative opinion	3	4
Review expressed is Obviously Negative opinion	2	3
Review expressed is Absolutely Negative opinion	0	1

Table 16. Empirical Rules for Determining Belonging the Text to a Certain Class (Predicted Classes)

Positive	Left border	Right border
Review expressed is High Positive opinion	$LB_1 = RB_2$	$\mathrm{Max}(R^{pos})$
Review expressed is Quite Positive opinion	$LB_2 = RB_3$	$RB_2 = LB_2 + k_2 \cdot \Delta^{pos}$
Review expressed is reasonably positive opinion	$LB_3 = \mathrm{Min}(R^{pos})$	$RB_3 = LB_3 + k_3 \cdot \Delta^{pos}$
k_2, k_3 – adjustors		$\Delta_{pos} = \frac{\max(R^{pos}) - \min(R^{pos})}{3}$
Negative	**Left border**	**Right border**
Review expressed is Rather Negative opinion	$LB_1 = RB_2$	$\mathrm{Max}(R^{neg})$
Review expressed is Clearly Negative opinion	$LB_2 = RB_3$	$RB_2 = LB_{23} + k_2 \cdot \Delta^{neg}$
Review expressed is Absolutely Negative opinion	$LB_3 = \mathrm{Min}(R^{neg})$	$RB_3 = LB_3 + k_3 \cdot \Delta^{neg}$
		$\Delta_{neg} = \frac{\max(R^{neg}) - \min(R^{neg})}{3}$

3.2.3. Case Study Results and Discussion

For testing and evaluating the adequacy of the Text Classification based on the CBSD phase, as a case study were used the following training samples: for the first layer (CBSD Creation Algorithm) – 5000 Polish-language films reviews (2500 TSP and 2500 TSN); for the second layer (Sentiment Classification Algorithm) – 3000 Polish-language films reviews (1500 SPCS and 1500 SNCS) from the filmweb.pl. To consider the SPCS films reviews, if the subjective review's assessment is more than 5 points, and SNCS – if it is equal or less 5 points.

3.2.3.1 CBSD Creation Algorithm

As a result of the first layer of the developed methodology the Hierarchical Topically Oriented Corpora-based Sentiment Dictionary was created (Table 17).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2018 doi:10.20944/preprints201810.0338.v1

Peer-reviewed version available at Information 2018, 9, 307; doi:10.3390/info9120307

Table 17. The Semantic Structure of CBSD (%)

Polarity	Positive Bigrams	Neutral Bigrams	Negative Bigrams
2nd level of CBSD Positive Class	43.70	46.30	9.91
2nd level of CBSD Negative Class	20.75	37.53	41.72

The main specificities of the received CBSD [32]:

- for Positive Class of CBSD: Almost equal numbers of bigrams of neutral and positive polarity. This suggests that half of the adjectives and verbs used to characterize the reviewer's opinion without having a positive coloring, formally confirm (ascertain) the existing facts. 10% of negatively colored bigram, indicating that, despite the truly positive tonality of reviews, the reviewer doubts about the positivity of certain shades (elements) of the film. The greatest number of positively colored Bigram is related to the to the Topics: Role / Actors and Script / History.
- for Negative Class of CBSD: Almost more bigrams are negative and, are less, neutral polarity. Negative reviews are characterized, in turn, by a large number of oppositely painted bigrams. Perhaps some of these positive emotions are introduced by the authors for comparison or contrast. most of the negatively colored bigram refers to the Topics: Scene / Actor and Role / Scene.

3.2.3.2 Sentiment Classification Algorithm

Simple Sentiment Classification

At the step VII.5 of the developed methodology the algorithm of Sentiment Classification using CBSD 0s level of CBSD (without taking into account their Contextual Hierarchical structure of Topics) was realized (Table 18).

Table 18. Evaluation of the Quality of Sentiment Classification of the Films Reviews Results (Simple Classification, in %)

SPCS					SNCS				
Class	%	Precision	Recall	Accuracy	Class	%	Precision	Recall	Accuracy
HP	28.57	53.57	51.72		RN	33.00	33.33	29.73	
QP	47.96	51.06	53.33	47.96	CN	56.00	53.57	57.69	43.00
RP	23.47	34.78	33.33		AN	11.00	18.18	18.18	

Additionally, results of comparing the quality of the recognition of the reviews of the films SPCS / SNCS allowed to draw the following conclusions:

1. A large part of reviews is characterized by an average degree of density of the distribution of words with recognizable tonality. This fact complicates the process of an assessment of the rating of the film.
2. The morphological analysis of Training Sample testifies that [31]:

- the positive reviews characterized by highly semantic structured opinion, expressed in a carefully and balanced manner. In this connection, they have a more even (in comparison with negative) distribution of words that have the explicit tonality color.
- the negative reviews characterized by average level of semantic structure of the opinion, expressed more spontaneously and under the influence of emotions. On the other hand, this spontaneity causes less variability of the words used, and, as a consequence, greater probability of their precise recognition and classification.

One- and Two-Level Sentiment Classification

Realizing the algorithm of Sentiment Classification using the 1st level of CBSD, taking into account the recommendations formulated at the previous stage, allowed:

1. Recognize the Sentiment of texts Paragraphs taking into account the 1st level Topics of CBSD (Table 19).

Table 19. The Contextual Framework of 1st level of Films Reviews Corpora (% to the total number of paragraphs)

Class	Hero	Director	Script	Plot	Spectator	Unrecognized
HP	19.28	57.45	46.38	17.39	45.45	
QP	37.35	34.04	37.68	26.09	31.82	9.29
RP	43.37	8.51	15.94	56.52	22.73	
Class	**Hero**	**Actor**	**Creator**	**Plot**		**Unrecognized**
RN	57.14	-	44.12	37.84		
CN	28.57	-	47.06	45.95		14.50
AN	14.29	-	8.82	16.22		

2. Recognize the Sentiment of texts Paragraphs taking into account the 2nd level Topics of CBSD (Table 20).

Table 20. The Contextual Framework of 2nd level of Films Reviews Corpora (% to the total number of paragraphs)

	Classes						
Topic	**HP**	**QP**	**RP**	**Topic**	**RN**	**CN**	**AN**
Hero				Hero			
Actor / Play	7.14	53.57	39.29	Action / History	67.86	28.57	3.57
History / Film	2.33	55.81	41.86	Director / Cinema	77.78	22.22	-
Picture / Scene	21.54	48.46	30.00	Scene / Actor	80.23	16.28	3.49
Director / Creator	-	28.57	71.43				
Director				Creator			
Film / Director	5.88	35.29	58.82	Hero / Scene	80.00	20.00	-
Scene / Story	-	100.00	-	Film / Script	-	100.00	-
Style	19.05	52.38	28.57	Picture / Actor	88.24	11.76	-
Creator / Author	18.52	55.56	25.93				
Script				Plot			
Film / Director	12.00	48.00	40.00	Story / Hero	67.74	25.81	6.45
Story / Hero	15.49	50.70	33.80	Director / Image	61.40	36.84	1.75
Author / Creator	12.00	60.00	28.00	Creator / Film	85.71	-	14.29
Role / Actors	-	64.71	35.29				
Plot				Actor			
Film / Effects	13.33	40.00	46.67	Hero / Image	73.68	15.79	10.53
Portrait / Image	0.00	66.67	33.33	Role / Scene	–	-	-
Director / Production	33.33	50.00	16.67	Script / History	63.64	27.27	9.09
Script / History	-	100.00	-				
Spectator							
Hero / Fan	13.33	66.67	20.00				
Film / Aspects	37.50	37.50	25.00				
Role / Formulation	-	-	-				
Scene / Director	-	66.67	33.33				

3. To compare the Quality of Simple, One- and Two-level Sentiment Classification of the Films Reviews Results (Figure 3).

Peer-reviewed version available at *Information* **2018**, *9*, 307; doi:10.3390/info9120307

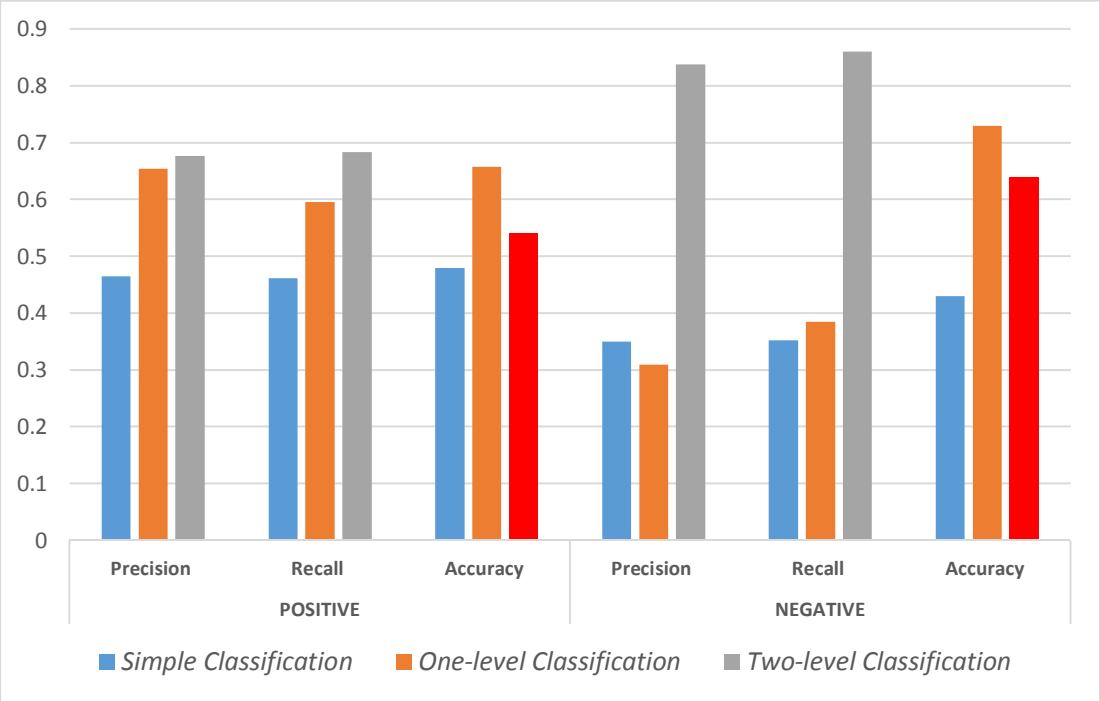

Figure 3. The difference between the Average values of Quality evaluation of the One-level and Simple Sentiment Classifying. *Source*: own research results

The general conclusions on the stage of classification can be the following: in comparison with the results of using 0ˢ, 1ˢᵗand 2ⁿᵈ level of CBSD, the Quality of Sentiment classification has increased significantly.

However, a more detailed analysis of the results obtained allows us to identify the following strengths and weaknesses of the conducted stages of text classification:

1. Indicators of *Precision* and *Recall* for *Subjectively Positive Sample* grow from 0ˢ Level step to 2ⁿᵈ Level step linearly and gradually. This confirms the previous conclusions that, in general, positive reviews have a higher level of semantic structure and orderliness in expressing emotions. In this regard, the process of recognizing the tonality of the text is better and more accurate even without using the Hierarchical Context Structure of the Sentiment Dictionary;
2. Indicators for indicators of *Precision* and *Recall* for *Subjectively Negative Sample* at the 2ⁿᵈ Level step grow steeply. This can be explained by the following facts:

- during the process of Text Classification using the 1ˢᵗ level of the CBSD, the topic Actor was not recognized for any paragraph of the SNCS. However, when using CBSD of the 2ⁿᵈ level, 2 of 3 subtopics of the topic Actor were recognized and assigned to paragraphs of the analyzed sample. This fact, on the one hand, it affected the stepwise increase in the recognition Quality Indicators at the Two-Level Text Classification, on the other hand, it explains the decrease in the Precision Indicator for the One-Level Text Classification;
- this phenomenon is also explained by the results of research conducted at the previous stages, indicating spontaneous, unstructured and sometimes illogical use of words of different tonality when writing Negative reviews under the influence of emotions.

3. A slight decrease in the average Accuracy indicator for the both Samples is could be caused by:

- too many Topics of the Second level of the Hierarchy used for Reviews analysis
- provided in the algorithm 6-class Tonality classification of each paragraph,

which makes the matrix of the results of the classification sufficiently sparse. For those first level of hierarchy, Accuracy values are much higher.

4. Conclusion and Discussion

In this paper, authors present the Methodology of Improving the Accuracy in Text Classification in Light of Modelling the Latent Semantic Relations. The main contribution of the paper and the author's study is finding the answers to the main scientific research questions:

1. Combination of the unsupervised machine learning Discriminant and Probabilistic Methods taking into account the specific features of Argumentative/ Persuasive type of documents, gave the opportunity to minimize these methods Limitation and, as a result, improve the qualitative characteristics of Topic Recognition process: Recall rate (the ratio of the number of semantically clustered/recognized paragraphs to the total number of paragraphs in the corpora) and Precision indicator (the average probability of significantly clustered/recognized paragraphs) (*Hypothesis* 1 is accepted).
2. Style of the analyzed text determines the possibility of flexible adaptation of the algorithms for the Texts Classification. For example, in the case study in this paper, the style/type of the analyzed texts (Review / Persuasive) allowed each document to be considered as a collection of Topically completed fragments (paragraphs), which positively affects the classification quality;
3. Hierarchically-oriented Structure of the Sentimental Dictionary allows customizing the Text Classification process to more accurately recognize the tonality of the text in the context of topic (*Hypothesis 2* is accepted).
4. Texts were written in the style of Persuasive, most often initially empowered by authors with a certain tonality. The tone, expressed in the author's opinion, has a significant, but not critical, effect on the qualitative indicators of sentiment recognition. Negative emotions of the author usually, on the one hand, reduce the level of variability of the words used and the variety of topics raised in the document, on the other – increase the level of unpredictability of contextual use of words with both positive and negative emotional coloring. At the same time, for author's negative opinions, there is an increase in the quality indicators characterizing Tonality recognition (Recall and Precision), but a slight decrease in the indicator of accuracy of the tonality recognition as a whole

Acknowledgments: The research results, presented in the paper, are partly 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".

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest

References

1. Blei, D. Introduction to Probabilistic Topic Models. *Comm. ACM.*, 2012; *55* (4), pp. 77-84.
2. Blei, D. Topic modeling. http: //www.cs.princeton.edu/~blei/topicmodeling.html
3. Gramacki, J.; Gramacki A. Metody algebraiczne w zadaniach eksploracji danych na przykładzie automatycznego analizowania treści dokumentów. *XVI Konferencja PLOUG*, 2010; pp.227-249.
4. Liu, B. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 2012; Vol. 5(1).
5. Rizun, N.; Taranenko, Y.; Waloszek, W. The Algorithm of Modelling and Analysis of Latent Semantic Relations: Linear Algebra vs. Probabilistic Topic Models. *Knowledge Engineering and Semantic Web. 8th International Conference*, 2017; pp.53-68.
6. Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. *Addison-Wesley, Wokingham, UK*, 2011; Second edition.
7. Furnas, G.W.; Deerwester, S.; Dumais, S.T.; Landauer, T.K.; Harshman, R.A.; Streeter L.A.; Lochbaum, K.E. Information retrieval using a singular value decomposition model of latent semantic structure. *In Proc. ACM SIGIR Conf.*, ACM, New York, 1998; pp. 465-480.
8. Salton G.; Michael J. McGill Introduction to modern information retrieval. *New York McGraw-Hill - McGraw-Hill computer science series*, XV, 1983; 448 p.
9. Aggarwal, C.; Zhai, X. Mining Text Data, *Springer*, 2012.

10. Anaya, Leticia H. Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers, *Doctor of Philosophy (Management Science)*, 2011; 226 p
11. Papadimitrious, C.H.; Raghavan, P.; Tamaki, H.; Vempala, S. Latent semantic indexing: A probabilistic analysis. *Journal of Computer and System Sciences*, 2000; 61, pp. 217-235.
12. Rizun, N.; Kapłanski, P.; Taranenko, Y. Development and Research of the Text Messages Semantic Clustering Methodology. *Third European Network Intelligence Conference, Publisher*: ENIC, 2016; # 33, pp.180-187
13. Rizun, N.; Kapłanski, P.; Taranenko, Y. 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.
14. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data Clustering: A Review; *ACM Computing Surveys*,1999; Vol. 31, Nr. 3.
15. Tomanek, K.; Analiza sentymentu – metoda analizy danych jakościowych. Przykład zastosowania oraz ewaluacja słownika RID i metody klasyfikacji Bayesa w analizie danych jakościowych, *Przegląd Socjologii Jakościowej*, 2014; pp. 118-136, www.przegladsocjologiijakosciowej.org
16. Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; Deerwester, S. Using latent semantic analysis to improve information retrieval. *In Proceedings of CHI'88: Conference on Human Factors in Computing*, New York: ACM, 1988; pp. 281-285.
17. Deerweste,r S., Susan, T.; Dumais, Harshman R. Indexing by Latent Semantic Analysis. 1990. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf
18. Rui, X.; Donald, C.; Wunsch, II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*. 2005; 16(3): pp. 645-678.
19. Canini, K. R.; Shi, L.; Griffiths, T. Online Inference of Topics with Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 2009; Proceedings Track 5: pp. 65-72.
20. Bahl, L.; Baker, J.; Jelinek, E.; Mercer, R. Perplexity – a measure of the difficulty of speech recognition tasks. In Program, *94th Meeting of the Acoustical Society of America*, 1977; Volume 62, p. S63.
21. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2013; 3: pp.993–1022.
22. Asgari, E., and Bastani, K.The Utility of Hierarchical Dirichlet Process for Relationship Detectionof Latent Constructs. *In Academy of Management Proceedings*, 2017; 1, p. 16300.
23. Titov, I. Modeling Online Reviews with Multi-grain Topic Models. *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*, 2008; pp. 111–120.
24. Klekovkina, M.V.; Kotelnikov, E.V. The method of automatic classification of texts by tonality, based on the dictionary of emotional vocabulary. *Electronic libraries: promising methods and technologies, electronic collections* (RCDL-2012), 2012; pp.118-123.
25. Taboada, M.; Brook,e J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis, *Computational Linguistics*, 2011; No. 37 (2), pp. 267–307.
26. Boiy, E.. Automatic Sentiment Analysis in On-line Text. *Proceedings of the 11th International Conference on Electronic Publishing* (ELPUB 2007), 2007; pp. 349–360.
27. Boucher, J.D.; Osgood, Ch.E. The Pollyanna hypothesis. *Journ. of Verbal Learning and Verbal Behaviour*, 1969; No. 8, pp. 1–8.
28. Pang, B. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*. 2008; Vol. 2. pp. 18-22.
29. Rizun, N.; Taranenko, Y. Development of the Algorithm of Polish Language Film Reviews Preprocessing. *Research Yearbook Faculty of Management in Ciechanów* WSM, 1-4 (IX), 2017; pp. 168-188.
30. Rizun, N.; Taranenko, Y.; Waloszek W. The Algorithm of Building the Hierarchical Contextual Framework of Textual Corpora. *Eighth IEEE International Conference on Intelligent Computing and Information System*, ICICIS 2017, Cairo, Egypt, 2017; pp.366-372.
31. Rizun, N.; Taranenko, Y. Methodology of Constructing and Analyzing the Hierarchical Contextually-Oriented Corpora. *Proceeding of Federated Conference on Computer Science and Information Systems*, 2018; pp. 501-510.
32. Rizun, N.; Waloszek, W. Methodology for Text Classification using Manually Created Corpora-based Sentiment Dictionary. *In Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018)*, 2018; Volume 1: KDIR, pp. 212-220. ISBN: 978-989-758-330-8

33. 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,2015; pp. 22–33.