information

MDPI

Article

Improving the Accuracy in Sentiment Classification in the Light of Modelling the Latent Semantic Relations †

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

[1] Department of Applied Informatics in Management, Faculty of Management and Economics, Gdansk University of Technology, 80-233 Gdańsk, Poland
[2] Department of Computer Integrated Techniques and Metrology, Ukrainian State Chemical Technology University, 49000 Dnipro, Ukraine; taranenkonew@gmail.com
[3] Department of Software Engineering, Faculty of Electronics, Telecommunications and Informatics Gdansk University of Technology, 80-233 Gdańsk, Poland; wowal@eti.pg.edu.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. In Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web, Szczecin, Poland, 8–10 November 2017; pp. 53–68.

Received: 15 October 2018; Accepted: 28 November 2018; Published: 4 December 2018

Abstract: The research presents the methodology of improving the accuracy in sentiment classification in the light of modelling the latent semantic relations (LSR). The objective of this methodology is to find ways of eliminating the limitations of the discriminant and probabilistic methods for LSR revealing and customizing the sentiment classification process (SCP) to the more accurate recognition of text tonality. This objective was achieved by providing the possibility of the joint usage of the following methods: (1) retrieval and recognition of the hierarchical semantic structure of the text and (2) development of the hierarchical contextually-oriented sentiment dictionary in order to perform the context-sensitive SCP. The main scientific contribution of this research is the set of the following approaches: at the phase of LSR revealing (1) combination of the discriminant and probabilistic models while applying the rules of adjustments to obtain the final joint result; at all SCP phases (2) considering document as a complex structure of topically completed textual components (paragraphs) and (3) taking into account the features of persuasive documents' type. The experimental results have demonstrated the enhancement of the SCP accuracy, namely significant increase of average values of recall and precision indicators and guarantee of sufficient accuracy level.

Keywords: sentiment classification; topic modelling; Latent Semantic Analysis; Latent Dirichlet Allocation; hierarchical sentiment dictionary; contextually-oriented hierarchical corpus; text tonality; accuracy

1. Introduction

The rapid development of computer technology and the Internet space in recent decades has led to the fact that the process 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 populated by the users themselves and publicly available.

This content and some other official information (for example, financial statements of enterprises, scientific and news articles) form 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), which are hidden both inside the documents and between them and are used to identify the document's context as the set of topics and to group the documents based on their semantic proximity correspondingly. Closely related to the topical structure identification of open unstructured content is the problem of retrieving the knowledge about emotional colouring of such texts. A special section of computer linguistics is devoted to the extraction of such knowledge—automatic analysis of text tonality, more known as sentiment analysis, sentiment classification or opinion mining. The close connection between tasks of latent semantic relations and sentiment classification appeared due to the following reasons: the initial goal of sentiment classification methods is the classification of documents (paragraphs, sentences) according to a given scale of tonality, usually a two-point (positive-negative) or three-point (positive-negative-neutral). However, general assessment of tonality does not allow consideration of the specifics of the semantics and sentiment of the connection of words used by authors in certain topical contexts. That is why 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 context-sensitive (or contextually-oriented) sentiment classification, which is to automatically determine the views of the user, expressed in the text, with respect to previously detected topics being examined.

In recent years, a number of methods for detecting topics and sentiment analysis have been proposed. But, firstly, many of these methods are devoted to the improvement of the theory of Latent Semantic Analysis and its use to identify hidden contextual links between documents [1–4] or improving the recognition of topics while using the knowledge about words probability distributions within the text collection [5–17], for example Bayesian nonparametric topic models [5–7]. The results of these studies do represent a significant contribution to the development of science. However:

- firstly, they suggest using only one of the listed methods (belonging to a group of discriminant or probabilistic methods respectively) with characterizing its shortcomings and limitations;
- secondly, the results obtained during the research are not associated by authors with the possibility of their further use for conducting not only context-semantic but also context-sentiment classification.

Another part of the research is devoted to finding a solution to the sentiment of classification based on various methods (naive bayesian classifier, maximum entropy classifier, support vector machines, sentiment lexicons based classification) but not taking into account improvement of the methods' results via simultaneous identification of the topics on the different document's levels [18–30]. However, recently some research work appears which aims to combine the two research directions into one, substantiating the goal with the resulting effects of the research (namely, the quality of the classification).

Some recent scientific work [31–34] has considered this limitation and has performed both—recognizing the sentiment and at the same time detecting the topics. For example, the work [31] presented a joint sentiment/topic model that can simultaneously recognize document-level sentiment and detect the set of topics from the text. Also the work [32] proposes two sentiment topic models to find the relation between the latent topics and readers' emotions. Considering the significance of the research results of the above-mentioned works, it should be noted that:

- firstly, they use (and improve) only one method of modifying topics (for example, it is the LDA model in reference [31] or probabilistic latent semantic indexing in reference [33]);
- secondly, they do not focus their scientific interests on the creation of a multi-level sentiment dictionary that reflects the contextual dependencies of the words tonality on different hierarchical levels of document contextual structure (topic/subtopic).

In this research, we focus on the aim of finding the ways of improving the accuracy in sentiment classification based on proposed joint latent semantic relations and sentiment analysis. The state-of-the-art of our methodology is that it can detect both: hierarchical topical structure of the document and then its context-sensitive sentiment. Accuracy in sentiment classification improvement is achieved due to the following:

- combining linear algebra and probabilistic topic models methods for LSR revealing allowed to eliminate their limitations;
- retrieving the knowledge about the hierarchical topical structure of the analysed text allowed (1) development of the hierarchical contextually-oriented sentiment dictionary and (2) performance of the context-sensitive sentiment classification on the paragraph- and document-level.

The rest of the paper is organized as follows. Section 2 introduces the theoretical background of the research. Section 3 presents the methodology of sentiment classification, contains (1) the latent semantic relations revealing and (2) sentiment classification based on the corpora-based sentiment dictionary (CBSD) phases. For testing and evaluating the adequacy and quality of proposed methodology additionally for each phase. In Section 3 we discuss the experimental results based on the Polish-language film reviews dataset. Finally, Section 4 concludes the paper and poses some questions for discussion.

This paper is an extended version of [35]. The following sections were added:

- an extended version of experimental results and discussion section for latent semantic relations revealing phase;
- representation of the new stage of research based on the results obtained in the original paper (the section describes the methodological and experimental parts of the sentiment classification phase based on the CBSD).

2. Theoretical Background of the Research

2.1. Vector Space Model Concept

The objective of the LSR analysis is to receive "semantic pattern" from the textual data and automatically extend them into the main latent semantic topic. The substantial contribution to this problem exploring and analysing the data came from researchers in the information retrieval scientific area (IR) [35–38]. The basic approach proposed by the IR researchers for the text corpus analysis is presenting each document as a vector, which contains the frequencies of particular words occurrence.

In the concept of $TF \times IDF$ weight [2–4,14,39], firstly the $A(m \times n)$ term-document matrix is built. As elements, it contains the values of absolute frequency of words occurrence. As the next step, this term frequency values are weighted by inverse document frequency indicator, which evaluates the number of occurrences of a word in the whole corpus [35]:

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

where $tf(w,t)$—relative frequency of the wth word occurrence in the document t:

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

$k(w, L_t)$—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.

Further on, for solving the problem of finding the distance between the documents (terms) from the point of view of their relation to the same topic, different metrics can be applied. The most popular and accurate measure is the indicator of cosine similarity between the vectors:

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

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

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

In the next step, the main task of the researchers is to find (develop) the most accurate clustering algorithms for documents (words) with the goal to present their semantic (topical, context) closeness [2–4,11,14,35,39,40].

This method limitation is that calculations measure only the "surface" usage of words as patterns of letters. Hereby, polysemy and synonymy are not captured [1,11,30,35].

2.2. Latent Semantic Indexing

The method of Latent Semantic Indexing (LSI), better known under the name of Latent Semantic Analysis (LSA) [41,42], represents the concept of determining the degree of closeness of documents (terms) and visualizing it in a space of a lower dimension by identifying and interpreting hidden semantic relations existing between them.

The most well-known version of LSA is based on the algorithm of singular value decomposition (SVD) of a term-document matrix [41]. As a result of the SVD of the term-document matrix A, we get 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.

However, despite the obvious advantages of the proposed method, according to [2,14], there are some significant limitations for effective LSA application: (1) documents should have the same writing style (Lim#1); (2) each document should be focused on only one topic (Lim#2); (3) a word should have a high probability of belonging to one topic but low probability of belonging to other topics (Lim#3). Such limitations are based on the assumption that each topic's or document's probability is distributed uniformly, which does not correspond to the real distribution in the document collections [35,41–43]. That is why LSA aims to eliminate multiple occurrences of a word in different topics and thus cannot effectively solve the polysemy problems (Lim#4).

2.3. Probabilistic Topic Models

In contrast to discriminant algorithms (LSI, LSA), in a probabilistic approach, the topics are described by the model but the term-document matrix is used to evaluate its hidden parameters [18,35,36,42].

2.3.1. Latent Dirichlet Allocation Model

Latent Dirichlet Allocation (LDA) is a generative probabilistic graphical model based on a three-level hierarchical Bayesian modelling approach [9,10,30]. In the LDA model, each text is generated independently, according to the following scheme [21]:

1. Randomly select its text distribution by topic θ_d.
2. For each word in the text: (a) randomly select a topic from the θ_d distribution obtained at the 1st step; (b) randomly select a word from the distribution of words in the selected topic φ_t.

In the considered set of texts D, each text consists of n_d words. Observable variables are words in the text—w_{dn}. All other variables—hidden. For each text d, the variable θ_d is the distribution topics in this text. In the classic LDA model, the number of topics is fixed and initially set by the parameter T.

In the LDA model, it is assumed that the parameters θ_d and φ_w are distributed as follows: $\theta \sim Dir(\alpha)$, $\theta \sim Dir(\beta)$), where α and β are defined as vector parameters (the so-called hyper parameters) Dirichlet distributions.

2.3.2. Results Quality Evaluation

The most well-known method of evaluating the quality of probabilistic topic models on the test dataset D_{test} is the perplexity index [9,10,16,44]. Perplexity is used as a measure of how well a probability model predicts a sample. The lower perplexity indicates the better probability distribution for a sample prediction:

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

where $\sum_{d=1}^{M} \log p(w_d)$ is the log-likelihood of a set of unseen documents w_d given the topics φ_k and the hyper parameter α for topic-distribution θ_d of documents (likelihood of unseen documents can be used to compare models, while the higher likelihood implies the better model); $\sum_{d=1}^{M} N_d$ is the count of tokens within the set of unseen documents w_d.

The main limitation of LDA method is the following: the result of finding the number of topics that provide optimal values of the perplexity indicator does not provide the maximum level of probability with which particular document belongs to a specific topic (Lim#5) [9,10,17,35].

2.4. Text Sentiment Classification

Methods of sentiment classification analysis of the text are developed within the framework of two machine learning approaches—supervised and unsupervised machine learning [29]. 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 the dependence on training the data. For these methods to work, one also needs a corpus of documents but the preliminary mark-up 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: that is, identification of aspect terms and determination of their tonality. However, such methods require complex tuning for a given domain. For example, the method based on the LDA-based method 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 [12].

The above-considered methods of sentiment classification require the presence of a sentiment dictionary of text tonality evaluation. There are three basic approaches to such a dictionary [29]: (1) expert thesaurus based on dictionaries; (2) dictionary based on text collections (corpus).

Within the expert approach, the dictionary is compiled by the experts. This approach can be distinguished by the complexity and high probability of the absence of domain-specific words in the dictionary on the one hand, on the other—by the high quality of the dictionary in the sense of adequacy of the assigned key.

In the dictionaries approach, an initial small list of evaluation words is expanded by various dictionaries, for example, explanatory or synonyms/antonyms. However, this approach 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 a dictionary.

In [18], the dictionary of emotional vocabulary, compiled by experts manually, was used to determine the tone of individual words. In this dictionary, each word and phrase is associated with the orientation of the key (positive/negative) and with strength (in points).

The authors' methods proposed in [25,26] are based on the dictionary approach, that is: 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 the word significance. In the method of working with the dictionary closest to the paper [27], the following needs to be considered: firstly, the dictionary is created based on a statistical analysis of a training collection; secondly, the weight of the words is determined with the help of a genetic algorithm.

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

$$W_T^S = \sum_{i=1}^{N_C} |w_i| \tag{7}$$

where W_T^S—the weight of text T for tonality C; w_i—the word's weight; N_C—the number of estimated bigrams of tonality C in the text T.

Texts are classified according to the linear function:

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

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 the preponderance of positive vocabulary in the text [28]. If the value of the function f is greater than zero, the text is recognized as positive, otherwise—as negative.

In addition, a separate group of studies [31–34] is devoted to the work on the search for a combination of approaches of sentiment classification and latent semantic topics detection. Such scientific works, thanks to the improvement of methods for topics identifying and taking advantage of the synergistic effects of joint application of the context-semantic and context-sentiment aspects of text classification, provide sufficiently stable and high quality indicators (accuracy) of sentiment recognition (in the works [21–25,30,31] has been declared level of accuracy between the 82.9% and 87.2%). As we can notice, the level of accuracy varies and use as one of the quality indicators in addition to the measure of sentiment classification and topics recognition efficiency depending on the supervised/unsupervised type of model used.

3. Research Methodology

In this research, the following authors' definitions will be used:

1. Term is a basic unit of discrete data.
2. Contextual fragment (CF) is an indivisible, topically completed set of terms, located within a document's paragraph.
3. Document is a set of CF.
4. Topic is the label (one term) that defines the main semantic context of the text.
5. Contextual dictionary (CD) is a set of keywords that describe the semantic context of the topic.
6. Semantic cluster (SC) is the set of CF that characterized by high hidden semantic closeness.
7. Contextually-oriented corpus (HC) is a hierarchical structure of semantically close CF, built via application of unsupervised machine learning discriminant and probabilistic methods of the topic modelling and latent semantic relations analysis.
8. 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

The motivation for this research concerns the analysed document type specificity and finding ways to completely or partially:

- eliminate the existing limitations of the discriminant and probabilistic approaches by their joint use and adjustment for latent semantic relations revealing;
- customize the context-sensitive sentiment classification process to the more accurate recognition of the text tonality in the light of the semantic context of the topic.

As an object of sentiment classification in this methodology, the group of persuasive document types have been chosen (reviews, newspaper editorials, letters to the editor, opinion articles, speeches, monologues). This choice is justified by the fact that persuasive documents are characterized by clear or clear enough defined rules and structure of writing style [45]. It follows that:

- such a document will have a wide palette of topics and sub-topics, which allows guaranteeing a high accuracy of the formation of the hierarchical contextual structure of the document;
- the completeness of the authors' vocabulary should be sufficiently broad and meaningful to create a sentiment dictionary adequate to the general context of the dataset to be examined;
- the need to express the authors' own opinion in such type of document will allow to carry out a qualitative evaluation of the sentiment classification results on a guaranteed relevant dataset.

The choice of this type of the document will be at the same time considered as limitations in the scope of our research findings.

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

RQ_1. Does taking into account of the specific features of the analysed document type affect the quality of the topic modelling process results?

RQ_2. Is it possible to increase the level of quality of the topic modelling process results by joint usage and adjustment of the discriminant and probabilistic methods?

RQ_3. Does taking into account of the hierarchical structure of latent semantic relations within the corpus affects the accuracy of the sentiment classification results?

RQ_4. Is it possible to increase the sentiment classification process accuracy via building and using the contextually-oriented and semantically structured sentiment dictionary?

RQ_5. Is the tone, expressed in the document by its author effect on the qualitative indicators of sentiment recognition?

In order to answer these questions, the following main Assumptions (A) were formulated:

A1. *Taking into account the specificity of document type, chosen in this methodology, assume that each document has approximately the same writing style (eliminating the Lim#1).*

A2. *Taking into account the specificity of the document type chosen in this methodology, assume that each document has a complex structure and can be estimated integrally by separated classification centred on only one topic paragraph (eliminating the Lim#2).*

Based on 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 topic modelling veracity. This effect is expected to be achieved via increasing [35]:*

- *quality of LDA-method of topics recognition via an increased level of probability of assigning the topic to a particular CF by considering the hidden LSR phenomena (eliminating the Lim#5);*
- *quality of LSA-method of LSR recognition via adjusting the consequences of the influence of the uniform distribution of the topics within the document by considering probabilistic approaches (eliminating the Lim#3 and Lim#4).*

H2. *Identifying and using the hierarchical structure of latent semantic relations within the corpus effects and improves the sentiment classification process accuracy. This effect is expected to be achieved via increasing:*

- *adequacy of tonality assessment instruments via building and manually creating a 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 considering the tonality subjectively assigned to texts by the author.*

The proposed methodology for improving the accuracy of sentiment classification based on revealing and using the knowledge about latent semantic relations includes two main phases:

- latent semantic relations revealing phase;
- sentiment classification based on the CBSD phase.

As a dataset for a demonstration of the basic workability and evaluation of the quality of the methodology application results, the Polish-language film reviews dataset from the filmweb.pl was used.

This choice was due to the following factors:

- film reviews are a bright representative of a persuasive type of the group of documents;
- the choice of Polish-language texts makes it possible to simultaneously fill in the existing gap in such a study direction for a given language.

The experimental part of authors' methodology has been implemented in Python 3.4.1.

3.2. Latent Semantic Relations Revealing Phase

The basic version of latent semantic relations revealing phase includes seven steps (Figure 1).

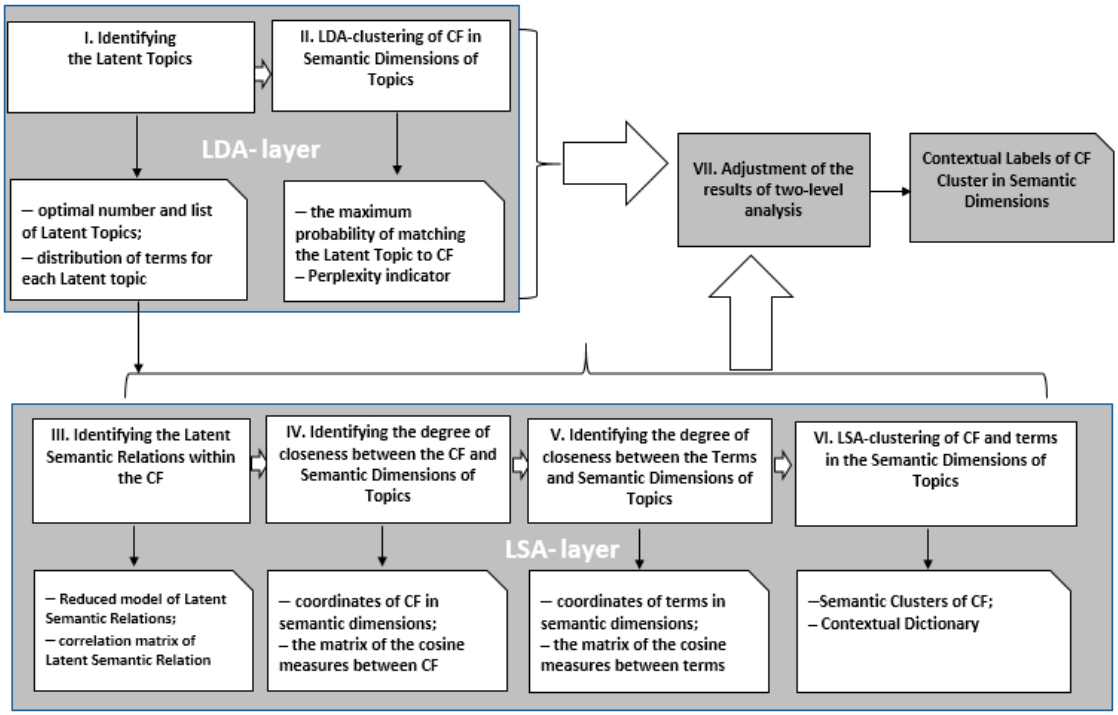

Figure 1. The steps of the latent semantic relations revealing phase. Source: own research results.

3.2.1. LDA-Based Analysis of Latent Semantic Relations Layer

Step I. Topics Identifying

LDA-based Analysis of LSR is the layer, which aims:

1. to reveal the optimal number of latent probabilistic topics that describe the main semantic context of the analysed document;
2. to assign these revealed topics to the CFs based on the discovered probabilistic Latent Semantic Relations within the paragraphs.

As a technical implementation support, the LDA Gensim Python package (https://radimrehurek.com/gensim/models/ldamodel.html) was used.

For preliminary evaluation of the latent semantic relations revealing phase quality, the training dataset (only one randomly chosen Polish-language film reviews that contained seven CF) was used.

Table 1 demonstrates the experimental results of such evaluation. The optimum value of the perplexity index is achieved at the moment when further parameter tuning changes do not lead to its significant decrease. In accordance with the authors' algorithm, obtained optimal number $t = 3$ (marked in red in the table) of latent probabilistic topics will be used as a recommended number of semantic clusters in the LSA-based layer of LSR analysis.

Table 1. Results of the Study of the LDA Model Parameters.

Perplexity	Number of Topics (t)	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 set of received latent probabilistic topics with the list of their most probable (significant) descriptive terms is presented in Table 2.

Table 2. 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

Step II. LDA-clustering

Based on the information about the maximum probability of assigning the latent probabilistic topics to the individual CF, in this step, the semantic (topical) clustering of CF could be performed. The results of this process based on the training dataset are presented in Table 3.

Table 3. 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 Table 1 prove the validity of the assumption A2 about the possibility and expediency of providing the analysis of the documents by paragraphs. However, we can note that the level of probability of belonging the individual CF to a particular topic/cluster is not very significant for all CF (for example, for CF_6 it is lower than 0.5, marked red in Table 3).

3.2.2. The LSA-Based Analysis of Latent Semantic Relations Layer

LSA-based Analysis of LSR is the layer, which *aims* to identify the hidden relationships between the terms and latent semantic topics. For revealing such information, we need the following: (1) to evaluate the degree of semantic correlation relationship between CF/terms via building the reduced model of LSR; (2) to form the semantic clusters of CF via determining the cosine distance between the CF; (3) to form the contextual dictionary of semantic clusters of CF via determining the cosine distances between the terms.

Step III. Identifying the Hidden Semantic Relations

Mathematically the reduced model, as the instrument of preliminary identification of the LSR presence, is the process of multiplying of the results of SVD transformation with chosen k-dimension $X_{K_{t\times d}} = U_{K_{t\times d}}\Sigma_{K_{t\times d}}\left(V_{K_{t\times d}}\right)^T$ [35]. The fragment of this step results based on the training dataset is presented in Table 4.

Table 4. Fragment of the Reduced Model.

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 [a]	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 [b]	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

[a] Term "Movie" seems to have the presence in all CF where the word "Character" appears; [b] Term "Woman" seems to have the presence in the CF where the word "Cinema" appears.

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

Table 5. Fragment 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

At the same time, we can note the fact of increasing values of the correlation coefficient (CC) (marked red in Table 6) between terms when compared to the results of Tables 4 and 5 (Table 6):

Table 6. Example of Results of the Comparison of the CC between Terms.

Terms \ Source	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. LSA Clustering of CF/Terms

For measuring the level of LSR, identified in the previous step, the matrix of *cosine distance* between the vectors of CF / terms (steps IV-V) should be built. Based on such matrices, in this step, the semantic clustering (step VI) process could be performed. An example of the implementation of k-means clustering [18,30] algorithm for CF and terms (using an LDA-based number of semantic clusters t = 3) is presented in Table 7.

Table 7. 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

3.2.3. Adjustments of the Results of the Two Layers of Analysis

In this step (step VII), it is supposed to combine the results of the implementation of LSA and LDA layers, namely:

1. Building the table of the comparison of the labels of latent semantic clusters of a set of CF, obtained in two layers of research (Table 8). As we can notice, the results of clustering for CF_4 and CF_6, obtained in LSA- and LDA-analysis layers, do not match (marked in red in Table 8).
2. Development and implementation of the rules of adjustments of the results obtained in the LSA- and LDA-analysis layers.

Table 8. 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

As stated above, LDA method implementation presupposes the assignment of the corresponding topics to CF based on the largest (from existing) probability (P) of the degree of their compliance with the analysed CF. In this connection, the authors' concept of rules of adjustments (RA) of the results of semantic clustering of the LSA- and LDA-analysis layers for each particular CF is proposed (Table 9).

Table 9. 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	$p \leq 0.3$	Cluster is Not recognized
3	LSA Cluster	$\neq$	LDA Cluster	$p \leq 0.3$	LSA Cluster
4	LSA Cluster	$\neq$	LDA Cluster	$0.3 < p \leq 0.7$	LSA Cluster/Re-clustering
5	LSA Cluster	$\neq$	LDA Cluster	$p > 0.7$	LDA Cluster

These rules allow:

- to improve the quality of LDA-method recognition of the CF's topics (rules 3, 4) due to the possibility of correcting those clustering results, which are characterized by the low level of

probability of a CF belonging to a particular topic. Suggested instrument—the specificity of the LSA method, consisting of the ability to extract knowledge of latent semantic relationships;

- to improve the quality of LSA-method recognition of hidden relations between the CF (rules 2, 5) due to the possibility of correcting those clustering results, which are characterized by the situations, when the particular CF coordinates are located on the cluster's boundary. Suggested instrument—the specificity of LDA method, consisting in the ability to extract the knowledge from latent topics probabilistic characteristics.

The results of the implementation of the rules of adjustments for the training dataset are presented in Table 10.

Table 10. Results of the 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

3.2.4. Experimental Results and Discussion

For the verification of the authors' methodology in this phase, the sentimental structure of the test dataset via classification of the film review dataset on the subjectively positive (SPSC) and subjectively negative sentiment corpora (SNSC) was created. This procedure is realized based on the information about subjective (provided by its authors) evaluations of the review tonality (measured by a 10-point scale). We consider that review refers to SPSC if the subjective review assessment is more than 5 points and refers to SNCS—if its assessment is equal or less than 5 points.

During the methodology verification, test dataset of 5000 polish-language film reviews (2500 SNCS and 2500 SNSC) were analysed. As a result, the two-level contextual hierarchical structure of topics (CHST) was defined (Table 11). The recommended number of clusters (identified in LDA-level of analysis):

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

Table 11. Results of the 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			
	Script/History	40			
Spectator	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 [8,38]. The contextual labels (CL) of the topics were assigned automatically based on the terms with the highest frequency in each topic.

Table 12. Hierarchical Structure of the Contextually-Oriented Corpus.

CL of the 1st Level Topics	SPSC			Topics of the 1st Level	SNSC		
	LSA, %	LDA, %	LSA&LDA, %		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 are the following: (1) percentage of not recognized CF inside the topic (indicator 1); (2) percentage of CF, which changed the cluster (indicator 2); (3) final qualitative characteristics of the research (recall rate) for the 1st level of topics are given in Table 13.

Table 13. Qualitative Indicators 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 1) gave the opportunity:

- to improve the 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 dataset) 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 mathematical and methodological basis for building the contextual hierarchical structure of semantic topics.

3.3. Sentiment Classification Based on the Contextually-Oriented Sentiment Dictionary Phase

The basic version of sentiment classification phase includes nine steps (Figure 2).

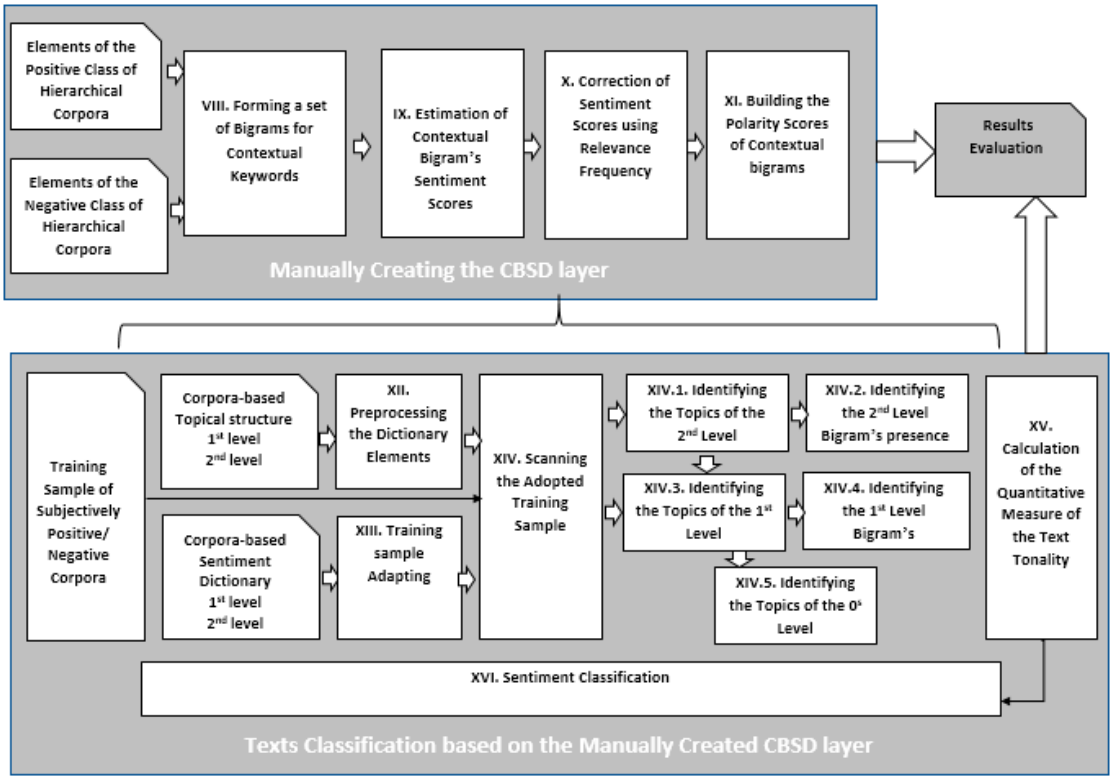

Figure 2. The steps of the sentiment classification based on the CBSD phase. Source: own research results.

3.3.1. Creating the Corpora-Based Sentiment Dictionary Layer

Creating the corpora-based sentiment dictionary (steps VIII-XI) 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 analysed texts with the high accuracy. One of the two components of bigram must be an element from a contextual dictionary of semantic clusters (Phase 1). CBSD should have three levels [46]:

- 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 taking into account the 1st level of CHST;
- 2nd level is the set of dictionary items taking into account the 2nd level of CHST.

Definition of the sentiment scores of the bigrams is estimated by the frequency of occurrence of this bigram in the elements of corpora. To increase the degree of accuracy of the sentiment scores, an estimation parameter to reverse the frequency—RF (relevance frequency) is used [47]:

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

where a—number of documents related to category S (positive, negative) and containing this bigram, b—the 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 this hierarchical approach to improve the accuracy of the sentiment classification process.

The main tasks of this layer are:

- to teach the developed sentiment classification algorithm to classify the texts, based on the quantitative measures of the tonality (sentiment scores) and considering one-level 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 of the applied algorithm via comparing the results of the sentiment classification.

3.3.2. Sentiment Classification Based on the Manually Created CBSD Layer

Steps XII–XIII. Preparing to Perform the Sentiment Classification Procedure

In this step, considering the specificity of the chosen dataset language as well as limited number of existing methods for the analysis of the text in Polish [45], in addition to the standard text pre-processing procedures, the authors have proposed the specific text adaptation procedure [8].

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

With the purpose of acceptance/rejection of Hypothesis 2, this step of the 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, taking into account their CHST—one-level classification (steps VII.3–5).

Procedure 3. Using CBSD, taking into account their CHST—one-level 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 the document) [35].

For realizing the procedures 3 (with the deepest topics identification process) the following algorithm is developed:

Step XIV.1. This step is realized via scanning the adopted training sample texts and identified the topics on the 2nd level of the CHST for each review paragraph. This procedure is implemented by adding the topic (contextual dictionary elements) from CHST to training sample as one of its paragraphs and then using the LSA method to find paragraphs that have a latent semantic relation.

Step XIV.2. This step is realized via scanning the part of the training sample for which topics on the 2nd level were identified, with the goal to find the bigrams from the 2nd level of CBSD which correspond to the topic identified for each paragraph.

Steps XIV.3–4. For paragraphs for which topics had not been defined in the step VII.1, these steps are realized via scanning this part of adopted training sample texts for identifying the topics on the 1st level of the CHST and subsequent search of the bigrams from the 1st level of CBSD which correspond to the topic identified for each paragraph.

Step XIV.5. For paragraphs for which topics had not been defined in the steps VII.1 and VII 3, this step is realized via search of the bigrams from 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.

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	Presence of word-modifiers in the immediate vicinity of the elements of the bigram	True

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

To determine the quantitative measure of the tonality estimation for the entire text of the 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.

Based on 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 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—the weight of text T for particular tonality; w_i—polarity score of bigram i; N_C—the number of estimated bigrams of particular tonality in the text T.

Each text is 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 use the following linear function:

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

where k_{neg} is the coefficient, compensating the fact of the preponderance of positive vocabulary in the texts [46].

Step XVI. Sentiment 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. The text has high positive tonality (HP).
  C2. The text has quite positive tonality (QP).
  C3. The text has reasonably positive tonality (RP).

- For subjectively negative corpora sample (SNCS):

  C4. The text has rather negative tonality (RN).
  C5. The text has clearly negative tonality (CN).
  C6. The text has absolutely negative tonality (AN).

Rule 2. To implement the training procedure for the algorithm being developed, the sentiment classification of texts is suggested using a basic quantitative measure of the text tonality [46]:

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

Rule 3. Considering the specificity of chosen document type and in order to implement the training procedure for the algorithm being developed, the sentiment classification is suggested using the following empirical rules for determining the text belonging to a certain class (Tables 15 and 16):

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

Positive	Left Border	Right Border
Review expressed is high positive opinion	8	10
Review expressed is quite a positive opinion	6	7
Review expressed is reasonably positive opinion		5
Negative	**Left Border**	**Right Border**
Review expressed is rather a 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 the Text Belonging to a Certain Class (Predicted Classes).

Positive	Left Border	Right Border
Review expressed is high positive opinion	$LB_1 = RB_2$	$Max(R^{pos})$
Review expressed is quite a positive opinion	$LB_2 = RB_3$	$RB_2 = LB_2 + k_2 \cdot \Delta^{pos}$
Review expressed is reasonably positive opinion	$LB_3 = Min(R^{pos})$	$RB_3 = LB_3 + k_3 \cdot \Delta^{pos}$
k_2, k_3—adjusters	$\Delta_{pos} = \frac{\max(R^{pos}) - \min(R^{pos})}{3}$	
Negative	**Left Border**	**Right Border**
Review expressed is rather a negative opinion	$LB_1 = RB_2$	$Max(R^{neg})$
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	$LB_3 = Min(R^{neg})$	$RB_3 = LB_3 + k_3 \cdot \Delta^{neg}$
	$\Delta_{neg} = \frac{\max(R^{neg}) - \min(R^{neg})}{3}$	

3.3.3. Experimental Results and Discussion

For testing and evaluating the adequacy of the sentiment classification based on the CBSD phase, the following test dataset was used: for the first layer (CBSD creation algorithm)—5000 Polish-language film reviews (2500 TSP and 2500 TSN); for the second layer (sentiment classification algorithm)—3000 Polish-language film reviews (1500 SPCS and 1500 SNCS) from the filmweb.pl. To consider the SPCS film reviews, if the subjective (provided by its authors) evaluations of review tonality are more than 5 points and SNCS—if it is equal or less 5 points.

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).

Table 17. 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 [46]:

- 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 colouring, formally confirm (ascertain) the existing facts. 10% of negatively coloured 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 coloured Bigram is related to the to the topics: Role/Actors and Script/History.
- for the negative class of CBSD: more bigrams are negative and, less are 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 coloured bigram refers to the topics: Scene/Actor and Role/Scene.

Sentiment Classification Algorithm

A. 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 [38]:

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

B. 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).
2. Recognize the sentiment of texts paragraphs taking into account the 2nd level topics of CBSD (Table 20).
3. To compare the quality of simple, one-level and two-level sentiment classification of the film reviews results (Figure 3).

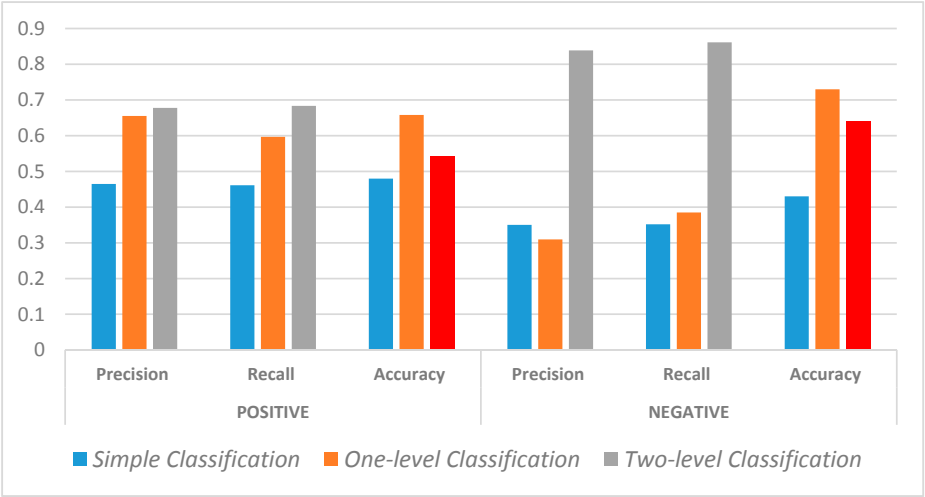

Figure 3. The difference between the average values of quality evaluation of the one-level and simple sentiment classifying. Source: own research results.

Table 19. Contextual Framework of the 1st level of Film 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		

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

Topic	Classes						
	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				

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

However, a more detailed analysis of the obtained results allows us to identify the following strengths and weaknesses of the conducted stages of sentiment 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 a subjectively negative sample at the 2nd level step grow steeply. This can be explained by the following facts:
    - during the process of sentiment 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 analysed sample. This fact, on the one hand, affected the stepwise increase in the recognition of quality indicators at the two-level sentiment classification, on the other hand, it explains the decrease in the precision indicator for the one-level sentiment classification;
    - this phenomenon is also explained by the results of research conducted at the previous stages, indicating the 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 both samples 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 the first level of the hierarchy, accuracy values are much higher.

4. Conclusions and Discussion

In this paper, the authors presented the methodology of improving the accuracy in sentiment classification in the light of modelling the latent semantic relations. In contrast to most existing approaches in sentiment classification, our methodology uses the joint latent semantic relations and sentiment analysis based on:

1. Detection of the hierarchical (in this study—maximal two-level) topical structure of the document and its context-sensitive sentiment.
2. Combining linear algebra and probabilistic topic models methods for LSR revealing allowed to eliminate their limitations. Such an approach allowed to bring the average value of the topic recognition recall rate indicator close to 90–95% and increase the precision indicator from 62 to 70–75% (Hypothesis 1 is accepted).
3. Retrieving the hierarchical topical structure of the analysed text, which allowed (1) to develop the hierarchical (in this study—two-level) contextually-oriented sentiment dictionary; (2) to use it to perform the context-sensitive sentiment classification in a paragraph- and then full document-level.

Such an approach allowed to increase the recall and precision indicators in average to 78% and guarantee the accuracy level at about 75%. These indicators unfortunately do not exceed the indicators presented in the works [21–25,30,31] (Hypothesis 2 is partly accepted). This fact can be explained by the reason of the chosen language and its methodology quality verification and lack of any available basic tools as well as a publicly available list of words (vocabulary) with established polarities.

As a future research, authors plan to verify this methodology for the English-language persuasive type of documents. The choice of persuasive type of document is considered by authors as a limitation in the scope of our research findings.

The main contribution of the paper is finding the answers to the main scientific research questions of the author's study:

1. Taking into account the specific features of the document types affects the quality of the topic modelling process results. One of the identified manifestations of this effect is the possibility of flexible adaptation of topic modelling algorithms for the sentiment classification process. for example, persuasive type of the analysed texts allowed to consider each document as a collection of topically completed fragments (paragraphs), which positively affects the classification quality (*RQ_1*).
2. It is possible to increase the level of quality of the topic modelling process results by using the combination of the discriminant and probabilistic methods. One of the identified manifestations of this effect is the possibility to apply the rules of adjustments of the results obtained in levels of semantic clustering of the LSA- and LDA-analysis in order to obtain the final result, what allows to eliminate the individual limitations of the methods being combined (*RQ_2*).
3. Taking into account the hierarchical structure of latent semantic relations within the corpus affects the accuracy of the sentiment classification results. One of the identified manifestations of this effect is allowing to customize the sentiment classification process by hierarchical iterative recognition of the topics of each paragraph of the document (from the lower to the higher level of contextual hierarchical structure of topics) with the subsequent use of the elements of the CBSD corresponding to the identified topic (*RQ_3* and *RQ_4*).
4. The tone, expressed in the document by its author, has a significant but not critical effect on the qualitative indicators of document 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 colouring. At the same time, for authors' negative opinions, there is an increase in the quality indicators characterizing tonality recognition (recall and precision) but with a slight decrease in the indicator of the accuracy of the tonality recognition as a whole (*RQ_5*).

Author Contributions: Conceptualization, N.R. and W.W.; Data curation, N.R.; Methodology, N.R., W.W. and Y.T.; Resources, W.W.; Software, Y.T.; Validation, N.R.; Writing—original draft, N.R.

Funding: 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".

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 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 Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, 13–15 June 1998; pp. 465–480.
2. Anaya, L.H. *Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers*; Doctor of Philosophy (Management Science); ProQuest LLC: Ann Arbor, MI, USA, 2011; 226p.
3. Papadimitrious, C.H.; Raghavan, P.; Tamaki, H.; Vempala, S. Latent semantic indexing: A probabilistic analysis. *J. Comput. Syst. Sci.* **2000**, *61*, 217–235. [CrossRef]
4. Rizun, N.; Kapłanski, P.; Taranenko, Y. *Method of a Two-Level Text-Meaning Similarity Approximation of the Customers' Opinions*; Economic Studies—Scientific Papers; Nr. 296/2016; University of Economics in Katowice: Katowice, Poland, 2016; pp. 64–85.

5. Teh, Y.W.; Jordan, M.I.; Beal, M.J.; Blei, D.M. Sharing clusters among related groups: Hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems: Vancouver, BC, Canada, 2005; pp. 1385–1392.
6. Xuan, J.; Lu, J.; Zhang, G.; Da Xu, R.Y.; Luo, X. Bayesian nonparametric relational topic model through dependent gamma processes. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 1357–1369. [CrossRef]
7. Xuan, J.; Lu, J.; Zhang, G.; Da Xu, R.Y.; Luo, X. Doubly nonparametric sparse nonnegative matrix factorization based on dependent Indian buffet processes. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 1835–1849. [CrossRef] [PubMed]
8. Rizun, N.; Taranenko, Y.; Waloszek, W. The Algorithm of Building the Hierarchical Contextual Framework of Textual Corpora. In Proceedings of the Eighth IEEE International Conference on Intelligent Computing and Information System, ICICIS 2017, Cairo, Egypt, 5–7 December 2017; pp. 366–372.
9. Blei, D. Introduction to Probabilistic Topic Models. *Commun. ACM* **2012**, *55*, 77–84. [CrossRef]
10. Blei, D.; Lafferty, J.D. Topic Modeling. Available online: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf (accessed on 4 December 2018).
11. Gramacki, J.; Gramacki, A. Metody algebraiczne w zadaniach eksploracji danych na przykładzie automatycznego analizowania treści dokumentów. In Proceedings of the XVI Konferencja PLOUG, Kościelisko, Poland, 19–20 October 2010; pp. 227–249.
12. Titov, I. Modeling Online Reviews with Multi-grain Topic Models. In Proceedings of the 17th International Conference on World Wide Web (WWW'08), Beijing, China, 21–25 April 2008; pp. 111–120.
13. Griffiths, T.; Steyvers, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 5228–5235. [CrossRef] [PubMed]
14. Aggarwal, C.; Zhai, X. *Mining Text Data*; Springer: New York, NY, USA, 2012.
15. Canini, K.R.; Shi, L.; Griffiths, T. Online Inference of Topics with Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **2009**, *5*, 65–72.
16. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **2013**, *3*, 993–1022.
17. Asgari, E.; Bastani, K. The Utility of Hierarchical Dirichlet Process for Relationship Detection of Latent Constructs. *Acad. Manag. Proc.* **2017**, *1*, 16300. [CrossRef]
18. Klekovkina, M.V.; Kotelnikov, E.V. The automatic sentiment text classification method based on emotional vocabulary. In Proceedings of the Digital libraries: Advanced Methods and Technologies, Digital Collections (RCDL-2012), Pereslavl-Zalessky, Moscow, 15–18 October 2012; pp. 118–123.
19. Kim, S.-M.; Hovy, E. Determining the sentiment of opinions. In Proceedings of the 20th International Conference on Computational Linguistics (COLING '04), Geneva, Switzerland, 23–27 August 2004; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; p. 1367.
20. Choi, Y.; Cardie, C.; Riloff, E.; Patwardhan, S. Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; Association for Computational Linguistics: Stroudsburg, PA, USA, 2005; pp. 355–362.
21. Whitelaw, C.; Garg, N.; Argamon, S. Using appraisal groups for sentiment analysis. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05), Bremen, Germany, 31 October–5 November 2005; pp. 625–631.
22. Pang, B.; Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL '04), Barcelona, Spain, 21–26 July 2004; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; p. 271.
23. Pang, B.; Lee, L.; Vaithyanathan, S. Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP '02), Philadelphia, PA, USA, 6–7 June 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 79–86.
24. Turney, P.D.; Littman, M.L. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *CoRR*, 2002; arXiv:cs/0212012.
25. Taboada, M.; Brooke, J.; Tofiloski, M.; Voll, K.; Stede, M. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **2011**, *37*, 267–307. [CrossRef]

26. Boiy, E. Automatic Sentiment Analysis in On-line Text. In Proceedings of the 11th International Conference on Electronic Publishing (ELPUB 2007), Vienna, Austria, 13–15 June 2007; pp. 349–360.
27. Boucher, J.D.; Osgood, C.E. The Pollyanna hypothesis. *J. Memory Lang.* **1969**, *8*, 1–8. [CrossRef]
28. Pang, B. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.* **2008**, *2*, 18–22. [CrossRef]
29. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–67. [CrossRef]
30. 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. Available online: www.przegladsocjologiijakosciowej.org (accessed on 30 November 2018).
31. Lin, C.; He, Y. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, Hong Kong, China, 2–6 November 2009; pp. 375–384.
32. Rao, Y.; Li, Q.; Mao, X.; Wenyin, L. Sentiment topic models for social emotion mining. *Inf. Sci.* **2014**, *266*, 90–100. [CrossRef]
33. Mei, Q.; Ling, X.; Wondra, M.; Su, H.; Zhai, C. Topic sentiment mixture: Modeling facets and opinions in weblogs. In Proceedings of the 16th international conference on World Wide Web (WWW '07), Banff, AB, Canada, 8–12 May 2007.
34. Titov, I.; McDonald, R. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of the ACL-08: HLT, Columbus, OH, USA, 19 June 2008; Association for Computational Linguistics: Stroudsburg, PA, USA, 2008; pp. 308–316.
35. Rizun, N.; Taranenko, Y.; Waloszek, W. The Algorithm of Modelling and Analysis of Latent Semantic Relations: Linear Algebra vs. Probabilistic Topic Models. In Proceedings of the 8th International Conference on Knowledge Engineering and Semantic Web, Szczecin, Poland, 8–10 November 2017; pp. 53–68.
36. Baeza-Yates, R.; Ribeiro-Neto, B. *Modern Information Retrieval*, 2nd ed.; Addison-Wesley: Wokingham, UK, 2011.
37. Salton, G.; Michael, J. *McGill Introduction to Modern Information Retrieval*; McGraw-Hill Computer Science Series, XV; McGraw-Hill: New York, NY, USA, 1983; 448p.
38. Rizun, N.; Taranenko, Y. Methodology of Constructing and Analyzing the Hierarchical Contextually-Oriented Corpora. In Proceedings of the Federated Conference on Computer Science and Information Systems, Poznań, Poland, 9–12 September 2018; pp. 501–510.
39. Rizun, N.; Kapłanski, P.; Taranenko, Y. Development and Research of the Text Messages Semantic Clustering Methodology. In Proceedings of the Third European Network Intelligence Conference, Wroclaw, Poland, 5–7 September 2016; pp. 180–187.
40. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data Clustering: A Review. *ACM Comput. Surv.* **1999**, *31*, 264–323. [CrossRef]
41. Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Deerwester, S. Using latent semantic analysis to improve information retrieval. In Proceedings of the CHI'88: Conference on Human Factors in Computing; ACM: New York, NY, USA, 1988; pp. 281–285.
42. Deerwester, S.; Dumais, S.T.; Furnas, G.W.; Landauer, T.K.; Harshman, R. Indexing by Latent Semantic Analysis. 1990. Available online: http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf (accessed on 30 November 2018).
43. Xu, R.; Wunsch, D. Survey of clustering algorithms. *IEEE Trans. Neural Netw.* **2005**, *16*, 645–678. [CrossRef] [PubMed]
44. Bahl, L.; Baker, J.; Jelinek, E.; Mercer, R. Perplexity—A measure of the difficulty of speech recognition tasks. *J. Acoust. Soc. Am.* **1977**, *62* (Suppl. 1), S63.
45. Rizun, N.; Taranenko, Y. Development of the Algorithm of Polish Language Film Reviews Preprocessing. *Rocznik Naukowy Wydzialu Zarzadzania w Ciechanowie* **2017**, *XI*, 168–188.

46. 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), Seville, Spain, 18–20 September 2018; Volume 1, KDIR. pp. 212–220, ISBN 978-989-758-330-8.
47. Ivanov, V.; Tutubalina, E.; Mingazov, N.; Alimova, I. Extracting Aspects, Sentiment and Categories of Aspects in User Reviews about Restaurants and Cars. In Proceedings of the International Conference "Dialogue 2015", Moscow, Russia, 27–30 May 2015; pp. 22–33.