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 is publicly available.
This content, as well as some more 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 identifying the topical structure of open unstructured content is the problem of retrieving the knowledge about the 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 a 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 to take into account the specifics of semantic 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 with usage 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 Discriminant or Probabilistic methods respectively) while noting characterizing their 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 (Naïve Bayesian Classifier, Maximum Entropy Classifier, Support Vector Machines, Sentiment Lexicons Based Classification), but no take into account improving the methods result via simultaneous identification the topics on the different document’s levels [18-30]. However, recently, works have begun to appear that aim to combine the two directions of research into one, substantiating the goal with the resulting effects of the research (namely, the quality of the classification).

Some of the recent scientific works [31-34] have been aware of this limitation and tried the both – to recognize the sentiments and at the same time to detect the topics. For example, the work [31] presented a joint sentiment/topic model which can detect document-level sentiment and extract a mixture of topics from text simultaneously. Or, work [32] proposes two sentiment topic models to associate latent topics with evoked emotions of readers. Considering the significance of the research results of the above works, it should, however, be noted that:

- firstly, use (and improve) only one method of modifying topics (for example, it is LDA model in [31] or Probabilistic Latent Semantic Indexing in [33]);
- secondly, do not concentrate their scientific interests on creating 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 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 the both: Hierarchical Topical structure of the
Document and then its Context-Sensitive Sentiment. Accuracy in Sentiment Classification improvement is achieved due to the fact that:

- combining the 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 analyzed text allowed (1) to develop the Hierarchical Contextually-oriented Sentiment Dictionary; (2) use it to perform the context-sensitive Sentiment Classification in a paragraph- and the 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 opens a window for discussion.

Due to this paper is an extended version of [35], and 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 (section described the methodological and experimental parts of the Phase of Sentiment Classification Based on the CBSD).

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" from the collection of information flow and automatically expand them into the underlying topic. Significant progress on the problem of presenting and analyzing the data was by the researchers in the field of information retrieval (IR) [35-38]. The basic methodology proposed by the 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 [2-4, 14, 39], based in the vocabulary of "bag of words" the $A(m \times n)$ term-document matrix is built. It contains the counts of the absolute frequency of words occurrence as elements. 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 [35]:

$$F_w = TF \times IDF = tf(w, t) \cdot \log \frac{D}{df},$$

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

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

$k(w, L_t)$ – the number of $w^{th}$ 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 similarity of documents (terms) from the point of view of their relation to the same topic, different metrics can be applied. The most appropriate metric is the cosine measure of the edge between the vectors:

$$\text{dist}_c = \cos \theta = \frac{x \cdot y}{\|x\| \|y\|},$$

where $x \cdot y$ – the scalar product of the vectors, $\|x\|$ and $\|y\|$ – quota of the vectors, which are calculated by the formulas:
Further part of the algorithm divides the source data into groups corresponding to the events as and determines whether a text document describes a set of any topic. The main idea of the solution is the use of clustering algorithms [2-4, 11, 14, 35, 39, 40]. This method’s 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

In 1988, Dumais et al. [41] proposed a method of Latent Semantic Indexing (LSI), most frequently referred to as LSA. In 1990 [42] Deerwester et al. designed a method to improve the efficiency of IR algorithms and search engines by the projection of documents and terms in the space of lower dimension, which includes semantic concepts of the original set of documents. LSA is a matrix algebra process. The most common version of LSA is based on the singular value decomposition (SVD) of a term-document matrix [41]. As a result of the SVD of the matrix $A$ we get three matrices:

$$X_{rd} \approx X_{Krd} = U_{Krd} \Sigma_{Krd} \left(V_{Krd}\right)^\top,$$  

where $\Sigma_{Krd}$ represents terms in $k$-dimensional latent space; $U_{Krd}$ represents documents in $k$-dimensional latent space; $U_{Krd}$, $V_{Krd}$ retain term–topic, document–topic relations for top $k$ topics.

However, as [2, 14] proved, there are three limitations to be consider while applying LSA: documents having the same writing style (Lim#1); each document being centred 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 and on the fact that the probabilities for each topic and document are distributed uniformly, which does not correspond to the actual characteristics of the collections of documents [35, 41-43]. That is why LSA tends to prevent multiple occurrences of a word in different topics and thus cannot be used effectively to resolve polysemy issues (Lim#4).

2.3. Probabilistic Topic Models

In contrast with so-called discriminative approaches (LSI, LSA), in a probabilistic approach, the topics are given by the model, and the term-document matrix is used to estimate its hidden parameters, which can be used to generate the simulated distributions [18, 35, 36, 42].

2.3.1. Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a generative probabilistic graphical model proposed by David Blei [9, 10, 30]. It is a three-level hierarchical Bayesian model. The algorithm of the method is as follows: (i) each document is generated independently: randomly select its distribution for document on topics $\theta_d$ for each word in a document; (ii) randomly select a topic from the distribution $\theta_d$, obtained in the first step; (iii) randomly select a word from the distribution of words in the chosen topic $\phi_k$ (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 the calculation of the Perplexity index on the test dataset $D_{test}$ [9, 10, 16, 44]. 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:

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

(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 \(\varphi_k\) and the hyperparameter \(\alpha\) for topic-distribution \(\theta_d\) of documents (likelihood of unseen documents can be used to compare models; the higher likelihood implying the better model); \(\sum_{d=1}^{M} N_d\) is the count of tokens within the set of unseen documents \(w_d\).

The limitation of LDA method is: it is possible to choose the optimum value of the \(k\) but, even under the 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) [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 avoiding dependence on training the data. For these methods to work, one also needs a Corpus of documents, but the 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: i.e. 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 considered above methods of Sentiment Classification, considered above, require the presence of Sentiment Dictionary of text tonality evaluation. There are three basic approaches to such a Dictionary [29]: (i) expert; thesaurus based on dictionaries; (ii) 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, 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 author’s methods proposed in [25, 26] are based on the dictionary approach, i.e.: 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:
where $W^C_T$ – the weight of text $T$ for tonality $C$; $w_i$ – the weight of the evaluated word $i$; $N^C$ – the number of estimated bigrams of tonality $C$ in the text $T$.

Texts are classified according to the linear function:

$$f(W^+_{T}, W^-_{T}) = W^+_{T} \cdot k_{neg} \cdot W^-_{T},$$

where $W^+_{T}$ is the positive weight of the text $T$; $W^-_{T}$ 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 positive, otherwise – negative.

In addition, a separate group of studies [31-34] is occupied by 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, as well as taking advantage of the synergistic effect of join application the context-semantic and context-sentiment aspect 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. 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 keywords that describe the 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 Context-Sensitive Sentiment Classification Process to the more accurate recognition of the text tonality in the light of Semantic Context of the topic.

As an object of Sentiment Classification in this methodology, the group of *Persuasive* Documents Type 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 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 author’s 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 author’s 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 document will be considered simultaneously as limitations in on the scope of our research findings.

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

RQ_1. Whether taking into account the specific features of the Documents 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 using the combination of the Discriminant and Probabilistic Methods?

RQ_3. Whether taking into account 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 Document by its author effect on the qualitative indicators of Sentiment Recognition?

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

A1. Taking into account the specificity of Documents Type, chosen in this methodology, assume that the writing style of each review is approximately the same (eliminating the Lim#1).

A2. Taking into account the specificity of Documents Type, chosen in this methodology, assume that each document has a complex structure and can be estimated integrally by separated classification of topically completed textual component (paragraphs), centred on a single Topic, as elements of their structure (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 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 recognize via an increased 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 the 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 Sentiment 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.

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 methodology application results, the Polish-language film reviews dataset from the filmbi.pl was used.
This choice was due to the following factors:

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

The experimental part of author's 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).

#### 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.

![Figure 1](https://ssrn.com/abstract=3425530)

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

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

For preliminary evaluation, the quality level of the Latent Semantic Relations revealing phase, the *training dataset* (only one, randomly chosen, Polish-language film review, which contains seven CF) was used.

Table 1 demonstrates the experiments results base on the *training dataset* of the main parameters of the 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 the author's algorithm, obtained an 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. 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 the information on the most probable (significant) terms, described this topic, is presented in Table 2.

Table 2. List of Latent Probabilistic Topics with Distribution of Terms

| Terms  | Probability | Terms  | Probability | Terms  | Probability |
|--------|-------------|--------|-------------|--------|-------------|
| 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 of CF in Semantic Dimensions of Corpus

Based on information about the maximum probability of matching the obtained Latent Probabilistic Topics to the CF, in this step, the process of Semantic (topical) clustering of CF could be performed. The results of this process with 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    | 0.841 | 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 providing the analysis of the Corpora by paragraphs. However, 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, the 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 the following: (i) to assess the degree of semantic correlation relationship between CF/terms via building the reduced model of LSR; (ii) 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; (iii) 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 the SVD transformation results with chosen k-dimension

\[ X_k = U_k \Sigma_k V_k^T \]. The fragment of this step results with training dataset of the Reduced model is presented in Table 5.

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

Table 4. 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    |

Table 5. Fragment 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|

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

Table 6. Example of Results of the Comparison of the CC Between 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 in the Semantic Dimensions of Topics. LSA Clustering of CF / Terms in the Semantic Dimensions of Topics

For measuring the level of LSR, identified in 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 (under the condition of an LDA-based number of SC) is presented in the Tables 7-8.

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    |

3.1.3. Adjustments of the Results of the Two Levels of Analysis

In the VII step of the author’s Algorithm, it is supposed to combine the results of the implementation of LSA and LDA levels for the analysis, namely:
1. Forming the table of the Comparison of the numerical labels of Latent Semantic Clusters of a set of CF, obtained in 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. Results of the Comparison of the Semantic Clusters as a set of CF Labels

| CF    | # Topic (Cluster) | LDA-level Probability | LSA-level 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 of 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 the 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 is characterized by the situations, when CF coordinates are 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 | LDA-analysis Result | LDA Probability (P) | Assignable Cluster |
|------|---------------------|---------------------|--------------------|-------------------|
| 1    | LSA Cluster =       | LDA Cluster         | >0.3               | LSA Cluster = LDA Cluster |
| 2    | LSA Cluster =       | LDA Cluster         | ≤0.3               | Cluster is Not recognized |
| 3    | LSA Cluster ≠       | LDA Cluster         | ≤0.3               | LSA Cluster |
| 4    | LSA Cluster ≠       | LDA Cluster         | 0.3<P≤0.7          | LSA Cluster / Re-clustering |
| 5    | LSA Cluster ≠       | LDA Cluster         | >0.7               | LDA Cluster |

The results of the implementation of Rules of Adjustments for training dataset are presented in Table 11.

Table 11. The 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. Experimental Results and Discussion

For the process of verification of the author’s methodology in this phase, the sentimental structure of FRC via classification of the reviews collection on the Subjectively Positive (SPSC) and Subjectively Negative Sentiment Corpora (SNSC) was formed. This procedure is realized based on the subjective evaluations of their tonality (SE) of films expressed by the reviewers (measured by a
10-point scale). We consider the SPCS films reviews if the subjective review’s assessment is more than 4.5 points, and SNCS – if it is equal or less than 5 points.

During the methodology verification, test dataset of 5000 Polish-language films reviews (2500 SNCS and 2500 SNSC) were analyzed. 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                | 23         |
|                        | Creator / Author       | 54         | Film / Script          | 60         |
| Script                 | Film / Director        | 8          | Picture / Actor        | 18         |
|                        | Story / Hero           | 58         | Plot                   | 39         |
|                        | Author / Creator       | 13         | Director / Image       | 18         |
|                        | Role / Actors          | 21         | Creator / Film         | 43         |
| Plot                   | Film / Effects         | 5          | Director / Production  | 24         |
|                        | Portrait / Image       | 31         | Script / History       | 40         |
| Spectator              | Hero / Fan             | 40         | Role / Formulation     | 16         |
|                        | Film / Aspects         | 20         | 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 | 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      |                        |        |        |            |

Electronic copy available at: https://ssrn.com/abstract=3425530
The quantitative indicators of the adjustments process of the Latent Semantic Relations Analysis results are the following: (i) percentage of not recognized CF inside the Topic (Indicator 1); (ii) percentage of CF, which changed the Cluster (Indicator 2); (iii) final qualitative characteristic of research (Recall rate) the 1st level of Topics are given in Table 13.

Table 13. Quality of the of LSR Analysis Results

| Topics   | SPSC Indicator 1 | SPSC Indicator 2 | SNSC Indicator 1 | SNSC Indicator 2 |
|----------|------------------|------------------|------------------|------------------|
| Hero     | 7.70             | 8.56             | Hero             | 9.23             |
| Director | 3.84             | 3.44             | Actor            | 5.30             |
| Script   | 4.19             | 16.60            | Creator          | 2.45             |
| Plot     | 6.11             | 7.30             | Plot             | 6.47             |
| 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 mathematical and methodological basis for building the Contextual Hierarchical structure of Semantic Topics.

3.2. 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
3.2.1. 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 [46]:

- 0th 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 as the element of TSP, if the subjective text’s assessment is truly positive (more than 8), and as the element of TSN – if it is truly negative (assessment less than 4 points) [38].

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 estimation the parameter to reverse the frequency – RF (Relevance Frequency) is used [47]:

\[
RF_x = \log_2 \left(2 + \frac{a}{\max(1,b)} \right),
\]

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 hierarchical to improve the accuracy of 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 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 of the Applied Algorithm via comparing the results of the Sentiment Classification.

3.2.1. Sentiment Classification Based on the Manually Created CBSD Layer

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

In this step, taking into account the specificity of the dataset language, as well as a limited number of existing software and algorithmic implementations for the analysis of texts in Polish [45], in addition to standard procedures for text pre-processing, the authors have provided text adaptation procedure [8].

Step VII. 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 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 the document) [35].

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

Step VII.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 Relations.

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

Steps VII.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 the bigrams form the 1st level of CBSD which correspond
to the Topic identified for each Paragraph.

Step V. For paragraphs for which topics had not been defined in the steps VII.1 and VII.3, this
step is realized via search the bigrams form 0-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 IX. Calculation of the Quantitative Measure of the Text Tonality

To determine the quantitative measure of the tonality estimation for the entire text of document
T from Subjectively Corpora Samples, the number of positive $N^{pos}_c$, neutral $N^{neu}_c$ and negative $N^{neg}_c$
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 $w^{pos}_i$, $w^{neu}_i$ and $w^{neg}_i$ are corrected (if necessary)
taking into account the rules for Words-modifiers and are summed up.

$$W^{pos}_{t} = \sum_{i=1}^{N^{pos}_c} w^{pos}_i, \quad W^{neu}_{t} = \sum_{i=1}^{N^{neu}_c} w^{neu}_i, \quad W^{neg}_{t} = \sum_{i=1}^{N^{neg}_c} w^{neg}_i,$$

(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:

$$f(W^{pos}_T, W^{neu}_T, W^{neg}_T) = W^{pos}_T + W^{neu}_T + k_{neg} \cdot W^{neg}_T,$$

(11)
where $k_{neg}$ is the coefficient, compensating the fact of preponderance of positive vocabulary in the texts [46].

Step IX. Sentiment Classification

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

1. 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).

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

   $$ R = f(W_T^{pos}, W_T^{neg}, W_T^{neu}). $$

3. Rule 3. Taking into account 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 15-16):

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 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 Belonging the Text to a Certain Class (Predicted Classes)

| Positive                                      | Left Border | Right Border |
|------------------------------------------------|-------------|--------------|
| Review expressed is High Positive opinion     | LB_1 = RB_1 | Max($R^{pos}$) |
| Review expressed is Quite a Positive opinion  | LB_2 = RB_2 | $R_B = LB_2 + k_2 \cdot \Delta^{pos}$ |
| Review expressed is reasonably positive opinion| LB_3 = Min($R^{pos}$) | $R_B = LB_3 + k_3 \cdot \Delta^{pos}$ |
| k_2, k_3 - adjusters                          |             |              |

$$ \Delta_{pos} = \max(R^{pos}) - \min(R^{pos}) \quad \frac{3}{3} $$

| Negative                                      | Left border | Right border |
|------------------------------------------------|-------------|--------------|
| Review expressed is Rather a Negative opinion  | LB_1 = RB_1 | Max($R^{neg}$) |
| Review expressed is Clearly Negative opinion   | LB_2 = RB_2 | $R_B = LB_2 + k_2 \cdot \Delta^{neg}$ |
| Review expressed is Absolutely Negative opinion| LB_3 = Min($R^{neg}$) | $R_B = LB_3 + k_3 \cdot \Delta^{neg}$ |

$$ \Delta_{neg} = \max(R^{neg}) - \min(R^{neg}) \quad \frac{3}{3} $$
3.2.3. 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 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).

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 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 coloured 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 0th 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.

One- and Two-Level Sentiment Classification

Realizing the algorithm of Sentiment Classification using the 1<sup>st</sup> 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 1<sup>st</sup> level Topics of CBSD (Table 19).

| 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     |              |

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

| 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 19. Contextual Framework of the 1<sup>st</sup> level of Films Reviews Corpora (% to the total number of paragraphs)

| Classes | Topic  | HP | QP | RP |
|---------|--------|----|----|----|
| Hero    | Hero   |    |    |    |
| Actor / Play | 7.14 | 53.57 | 39.29 |    |
| History / Film | 2.33 | 55.81 | 41.86 |    |
| Picture / Scene | 21.54 | 48.46 | 30.00 |    |
| Director / Creator | - | 28.57 | 71.43 |    |
| Film / Director | 5.88 | 35.29 | 58.82 |    |
| Scene / Story | - | 100.00 | - | - |
| Style | 19.05 | 52.38 | 28.57 |    |
| Creator / Author Script | 18.52 | 55.56 | 25.93 |    |

Table 20. Contextual Framework of the 2<sup>nd</sup> level of Films Reviews Corpora (% to the total number of paragraphs)
3. To compare the Quality of Simple, One- and Two-level Sentiment Classification of the Films Reviews Results (Figure 3).

The general conclusions on the stage of classification can be the following: in comparison with the results of using 0\textsuperscript{th}, 1\textsuperscript{st} and 2\textsuperscript{nd} 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 Sentiment Classification:

1. Indicators of Precision and Recall for Subjectively Positive Sample grow from 0\textsuperscript{th} Level step to 2\textsuperscript{nd} 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 2\textsuperscript{nd} Level step grow steeply. This can be explained by the following facts:
   - during the process of Sentiment Classification using the 1\textsuperscript{st} level of the CBSD, the topic Actor was not recognized for any paragraph of the SNCS. However, when using CBSD of the 2\textsuperscript{nd} level, 2 of 3 subtopics of the topic Actor were recognized and assigned to paragraphs of the analyzed

Electronic copy available at: https://ssrn.com/abstract=3425530
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 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 that first level
of the hierarchy, accuracy values are much higher.

4. Conclusion and Discussion

In this paper, authors presented the methodology of Improving the Accuracy in Sentiment
Classification in Light of Modelling the Latent Semantic Relations. In contrast to most of the existing
approaches in Sentiment Classification, our methodology uses the joint Latent Semantic Relations
and Sentiment Analysis, based on:
1. Detection the Hierarchical (in this study – maximally two-level) Topical structure of the
Document and its Context-Sensitive Sentiment.
2. Combining the 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%; increase the precision indicator from 62 to 70-75% (Hypothesis 1 is accepted).
3. Retriving the Hierarchical Topical Structure of the analyzed text, which allowed (1) to develop
the Hierarchical (in this study – two-level) Contextually-oriented Sentiment Dictionary; (2) use
it to perform the context-sensitive Sentiment Classification in a paragraph- and then full
document-level.
Such an approach allowed to increase of recall and precision indicators in average till 78%, and
guarantee of accuracy level near 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 chosen language of 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 English-language Persuasive
Type of Documents. The choice of Persuasive type of document now is considered by authors as
limitations in on 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 Documents Type 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 analyzed 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 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 to obtain the final result, allowing to
eliminate 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 Document by its author, has a significant, but not critical, effect on the qualitative indicators of Documents 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 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 the accuracy of the tonality recognition as a whole (RQ_5).

Acknowledgements: 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. 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.

2. Anaya, Leticia H. Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers, Doctor of Philosophy (Management Science), 2011; 226 p.

3. Papadimitrion, 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.

4. Rizun, N.; Kaplanski, 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.

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, 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 Transactions on Knowledge and Data Engineering, 2017; 29(7), 1357-1369.

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 transactions on neural networks and learning systems, 2018; 29(5), 1835-1849.

8. 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.

9. Blei, D. Introduction to Probabilistic Topic Models. Comm. ACM., 2012; 55 (4), pp. 77-84.

10. Blei, D. Topic modeling. http://www.cs.princeton.edu/~blei/topicmodeling.html

11. Gramacki, J.; Gramacki A. Metody algebraicznych w zadaniach eksploracji danych na przykładzie automatycznego analizowania treści dokumentów. XVI Konferencja PLOUG, 2010; pp. 227-249.

12. 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.

13. Griffiths T.; Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences, 2005(10001): 2004, 5228–5235.

14. Aggarwal, C.; Zhai, X. Mining Text Data, Springer, 2012.

15. 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.

16. Blei, D.; Ng, A.; Jordan, M. Latent Dirichlet allocation. Journal of Machine Learning Research, 2013; 3: pp.993–1022.
17. Asgari, E.; Bastani, K. The Utility of Hierarchical Dirichlet Process for Relationship Detection of Latent Constructs. In Academy of Management Proceedings, 2017; 1, p. 16300.
18. 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.
19. Kim S.-M.; Hovy E. Determining the sentiment of opinions. In COLING ’04: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics. 2004; pp. 1367.
20. Choi Y.; Cardie C., E. Riloff; S. Patwardhan. Identifying sources of opinions with conditional random fields and extraction patterns. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing Vancouver, British Columbia, Canada. Association for Computational Linguistics, 2005; pp 355–362.
21. Whitelaw C.; Garg N.; Argamon S. Using appraisal groups for sentiment analysis. In CIKM ’05: Proceedings of the 14th ACM international conference on Information and knowledge management, New York, NY, USA, 2005. ACM. pp 625–631.
22. Pang B.; Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In ACL ’04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, 2004. Association for Computational Linguistics. pp. 271.
23. Pang B.; Lee L.; Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques. In EMNLP ’02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Morristown, NJ, USA, 2002. Association for Computational Linguistics. pp 79–86.
24. Turney P. D.; M. L. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. CoRR, cs.LG/0212012, 2002.
25. Taboada, M.; Brooke 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. Liu, B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 2012; Vol. 5(1).
30. Tomanek, K.; Analiza sentymantu – 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.przegladsocjologijakosciowej.org.
31. Lin, C.; He, Y. Joint sentiment/topic model for sentiment analysis. In Proceedings of the 18th ACM conference on Information and knowledge management. ACM. 2009; pp. 375-384
32. Rao, Y.; Li, Q.; Mao, X.; Wenyin, L. Sentiment topic models for social emotion mining. Information Sciences, 2014; No 266, pp. 90-100.
33. Mei Q.; Ling X.; Wondra M.; Su H.; C. Zhai. Topic sentiment mixture: modeling facets and opinions in weblogs. In WWW ’07: Proceedings of the 16th international conference on World Wide Web.
34. Titov I.; McDonald R. A joint model of text and aspect ratings for sentiment summarization. In Proceedings of ACL-08: HLT, Columbus, Ohio. Association for Computational Linguistics. 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. Knowledge Engineering and Semantic Web. 8th International Conference, 2017; pp.53-68.
36. Baeza-Yates, R.; Ribeiro-Neto, B. Modern Information Retrieval. Addison-Wesley, Wokingham, UK, 2011; Second edition.
37. Salton G.; Michael J. McGill Introduction to modern information retrieval. New York McGraw-Hill - McGraw-Hill computer science series, XV, 1983; 448 p.
38. 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.

39. Rizun, N.; Kaplanski, 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.

40. Jain, A.K.; Murty, M.N.; Flynn, P.J. Data Clustering: A Review; ACM Computing Surveys, 1999; Vol. 31, Nr. 3.

41. 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.

42. Deerwester, S., Susan, T.; Dumais, Harshman R. Indexing by Latent Semantic Analysis. 1990. http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

43. Rui, X.; Donald, C.; Wunsch, II. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005; 16(3): pp. 645-678.

44. 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.

45. 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.

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), 2018; Volume 1: KDIR, pp. 212-220. ISBN: 978-987-758-330-8

47. 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

© 2018 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).