Proceedings of the Federated Conference on Computer Science and Information Systems pp. 505–514 DOI: 10.15439/2018F69 ISSN 2300-5963 ACSIS, Vol. 15

Methodology of Constructing and Analyzing the Hierarchical Contextually-Oriented Corpora

Nina Rizun
Gdansk University of Technology
Gabriela Narutowicza 11/12, 80-233
Gdansk, Poland
Email: nina.rizun@zie.pg.gda.pl

Yurii Taranenko
Alfred Nobel University, Dnipro
Naberezhna Lenina Str., 18, Dnipro,
49000, Ukraine
Email: taranenkonew@gmail.com

Abstract — **Methodology of Constructing and Analyzing the Hierarchical structure of the Contextually-Oriented Corpora was developed. The methodology contains the following steps: Contextual Component of the Corpora's Structure Building; Text Analysis of the Contextually-Oriented Hierarchical Corpus. Main contribution of this study is the following: hierarchical structure of the Corpus provides advanced possibilities for identification of the Morphological and Structural features of texts of different tonalities; Contextual, Morphological and Structural specificity of texts with tonality, originally assigned by the authors, has significant differences; exist the certain thought and writing style Templates, under the influence of which the formation of texts of various tonalities takes place. As basic features of such templates for the texts of the two basic (positive/negative) tonalities could be used: Contextual Structure, Morphological Types, Emotional Features, Writing Style and Vocabulary Richness. For verification of the proposed methodology, a case study of Polish-language film reviews Dataset was used.**

INTRODUCTION

In recent years the sentiment analysis has been one of the hottest research areas in natural language processing [1]. The challenges to the researchers are both theoretical aspects, such as the objective laws of the sentiment expressions in the natural language, and the practical aspects, for example, the analysis of consumer opinions and reviews [2].

There are two main approaches to the sentiment analysis [3]: *lexicon-based* and *machine learning*. The first of them determines the text sentiment by means of individual words polarity in the text. The latter considers the task of sentiment analysis as the problem of text categorization. Both approaches require high quality sentiment lexicons: even in the text categorization methods the word weights are often proportional to word polarity and strength.

There are many studies on the problem of Sentiment Lexicons creation. They generally use three main approaches [4]: manual approach, dictionary-based approach, and corpus-based approach. In the manual approach the sentiment lexicons are constructed by human annotators. In the dictionary-based approach the sentiment lexicons are created with the help of the universal dictionaries and thesauri, e. g., WordNet [5].

In the corpus-based approach the sentiment lexicons are built based on the analysis of text corpora. Also, the various hybrid combinations of these approaches are used. Though the problem of Sentiment Lexicon creation is very important, little attention is paid to the evaluation of the quality and in-depth analysis of the generated lexicons, especially for Polish language.

Main direction of this research is to design the methodology for constructing and analyzing the Hierarchical Corpora, which allows improving the quality of the algorithms for the Sentiment Lexicon building by offering the additional tools for determining the tonality based on the availability of data about the semantic properties of the text. As the main research tool, text mining methods and algorithms will be used.

THEORETICAL BACKGROUND

Under the notion of texts mining we understand the application of methods of texts computer analysis and presentation in order to achieve the quality, which corresponds to the "manual" processing for further usage in various tasks and applications. One of the actual tasks of automatic texts mining is their clustering (definition of groups of the similar documents). More and more often statistical topical methods are being applied [6].

The topics are presented as discrete distributions on a number of words, and the documents – as discrete distribution on a number of topics [6]. Topical methods perform a "non-precise" clustering of words and documents, which means that a word or a document can be referred to a few topics with different probabilities simultaneously. The synonyms with higher probability will appear in the same topics since they are frequently used in the same documents. At the same time, the homonyms (words different in meaning, but similar in writing) will be placed in different topics because they are used in different contexts [7].

A. Preprocessing Procedure

Topical methods, as a rule, apply the method of a "bag-of-words", where each document is considered as a set of words not connected to each other. Before the topics are defined, the text is preprocessed – its Graphematic and Morphologic

This work was not supported by any organization

analysis is conducted with the objective to define the initial form of words and their meanings in the speech context.

Graphematic Analysis

To start the preprocessing procedure of a text it is necessary to divide the original unstructured text into sentences and words. At first sight, it is a very simple task, but it has its own specificities and plays an important role in the further analysis of a text.

Graphematic analysis includes:

– division of the original text into elements (words, separators);

– elimination of non-text elements (tags, meta-information);

– extraction and formalization of non-standard elements: structural elements: headlines, paragraphs, notes; numbers, dates, complexes of letters and numbers; names, patronymics, surnames; extraction of e-mail addresses;

– extraction of files' names;

– extraction of sustained phrases, words that are not used separately from each other.

In English sources, we can meet the definition of tokenization, which, by its content, is similar to the graphematic analysis. Tokenization – is a process of dividing the text stream into tokens: words, collocations and sentences [8]. Thus, the graphematic analysis is the initial analysis of an unstructured text, presented as a chain of symbols in any coding, elaborating information, which is necessary for further text processing.

There are almost no tools specializing exceptionally on graphematic analysis. Basically, graphematic is included into integrated packages of text analysis: NLTK, Stanford CoreNLP, Apache NLP, AOT, MBSP etc. The function of division into tokens is also included into programs of text markup, for instance into the part-of-speech taggers.

Morphologic Analysis

Morphologic analysis provides definition of the normal form, from which the word-form was created, and of the set of parameters, assigned to this word-form [9].

Stemming has been the most widely applied morphological technique for information retrieval. With stemming, the searcher does not need to worry about the correct truncation point of search keys. Stemming also reduces the total number of distinct index entries. With short queries and short documents, a derivational stemmer is most useful, but with longer ones the derivational stemmer brings in more non-relevant documents. Stemming increases search key ambiguity. Stemming may, however, is a non-optimal approach to the clustering of documents in agglutinative languages. Firstly, stemmers do not conflate compounds whenever the first components do not match exactly. Secondly, they are unable to split compounds, which typically have the head-modifier structure and the headword is the last and more important component for clustering [10]. The most widely-spread algorithm of stemming is the Porter's algorithm. Except for that algorithm there exists the Lancaster's algorithm (for English language) and the algorithms, working by the principle of a "snowball" (snowball stemmers) for other languages.

Lemmatization is another normalization technique: for each inflected word form in a document or request, its basic form, the lemma, is identified. The benefits of lemmatization are the same as in stemming. In addition, when basic word forms are used, the searcher may match an exact search key to an exact index key. Such accuracy is not possible with truncated, ambiguous stems. Homographic word forms cause ambiguity (and precision) problems – this may also occur with inflectional word forms [11]. Another problem is that words cannot be lemmatized, because the lemmatizer's dictionary does not contain them.

B. Vector Space Models of the Semantic Relations Analysis

The method of processing words in a machine-readable natural language, as a rule, is based on the vector-space method of data description (Vector Space Model) [12], suggested by [13]. Within the framework of the method each word in a document has its particular weight. Thus, each document is presented as a vector and its dimension is equal to the total number of words in the document.

Similarity of a document and a topic is evaluated as a scalar product of a few information vectors. The weight of separate words (terms) can be calculated both applying the absolute frequency of a term appearing in the text and the relative (normalized) frequency:

$$F_{w_i} = TF \times IDF = tf(w,t) \cdot \frac{\log_{10} D}{df} \tag{1}$$

$tf(w,t)$ – relative frequency of the *w*-th term occurrence in document *t*:

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

$k(w,L_t)$ – the number of *w*-th term occurrences in the document *t*; *df* – the number of documents in the collection that contain the *w*-th term; *D* – total number of documents in the collection.

Then, for solving the problem of finding the similarity of documents (terms) from the point of view of the relation to the same topic, the different metric can be applied, for example:

– *Euclid's* measure:

$$dist_{t_i} = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}\ , \tag{3}$$

where *x* – vector of the document, *y* – point of reference words vector;

– *Cosine* of the edge between the vectors:

$$dist_{t_i} = \cos\theta = \frac{x \cdot y}{\|x\| \cdot \|y\|}, \tag{4}$$

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

$$\|x\| = \sqrt{\sum_{i=1i}^{n} x_i^2}\ ,\ \|y\| = \sqrt{\sum_{i=1i}^{n} y_i^2} \qquad (5)$$

A further algorithm is to divide the source data into groups corresponding to the events, as well as in determining whether a text document describes a set of any topic. The main idea of the solution is the use of clustering algorithms [14] (e.g., k-means method, etc.). It is assumed that each cluster contains documents that describe an event.

Latent Semantic Analysis (LSA) is a Discriminant theory and method for extracting context-dependent word meanings by statistical processing of large sets of text data [15-17]. It uses the "bag-of-words" for modelling, begins with transforming text corpora into term-document frequency matrices, reduces the high dimensional term spaces of textual data to a user-defined number of dimensions by singular value decomposition (SVD), produces: weighted term lists for each concept or topic; concept or topic content weights for each document; outputs that can be used to compute document relationship measures [18].

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

$$X_{t\times d} \approx X_{K_{t\times d}} = U_{K_{t\times d}} \Sigma_{K_{t\times d}} \left(V_{K_{t\times d}}\right)^T \qquad (6)$$

where $\Sigma_{K_{t\times d}}\left(V_{K_{t\times d}}\right)^T$ – represents terms in *k-d* latent space; $U_{K_{t\times d}}\Sigma_{K_{t\times d}}$ – represents documents in *k-d* latent space; $U_{K_{t\times d}}$, $V_{K_{t\times d}}$ – retain term–topic, document–topic relations for top *k* topics.

But, as [19, 20] proved, there are three *limitations* to apply LSA: documents having the same writing style; each document being centered on a single topic; a word having a high probability of belonging to one topic but low probability of belonging to other topics. The limitations of LSA is based on orthogonal characteristics of dimension factors as well as on the fact, that probabilities for each topic and the document distributed uniformly, which does not correspond to the actual characteristics of the collections of documents [13, 21, 22]. That is why, LSA tends to prevents multiple occurrences of a word in different topics and thus LSA cannot be used effectively to resolve polysemy issues.

C. Probabilistic Topic Models

To get rid of the above-mentioned disadvantages the probability LSA is conducted, based on the multinomial distribution – in particular, on the algorithm of Latent Dirichlet Allocation (LDA) [23, 24]. The *probabilistic topic modelling* – a set of algorithms to analyze the words in large sets of documents and from the retrieve the threads that connect into topics [25, 26]. In this case document is regarded as a set of words, the order of which does not matter. For each document to determine the distribution $\boldsymbol{\theta_d}$ of its words on topics, that probability for each topic meets it herein. This topic is presented in the form of distributions $\boldsymbol{\varphi_t}$ of words from a fixed vocabulary, i.e. each word included in the subject with a certain probability

The next text mining technique that was developed to improve upon LSA was the Probabilistic topic modeling techniques. Probabilistic topic modeling is a set of algorithms that allow analyzing words in textual corpora and extract from them topics, links between topics [23, 24, 27]. Latent Dirichlet Allocation (LDA) is a generative model that explains the results of observations using implicit groups, which allows one to explain why some parts of the data are similar. It was proposed by David Blei [23, 24] and it uses a Bayesian model that treats each document as a mixture of latent underlying topics, where each topic is modeled as a mixture of word probabilities from a vocabulary.

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

According to [28] – words in a topic from LDA (as an extended LDA method) are more closely related than words in a topic from LSA. For polysemy, words in a topic from LDA can appear in other topics simultaneously: topics are Dirichlet multinomial random variables, each word is generated by a single topic, and different words may be generated from different topics. The *limitation* of LDA is that there is no probability distribution model at the level of documents. Thus, the larger the number of documents, the larger the LDA model.

METHODOLOGY FOR CONSTRUCTING AND ANALYZING THE HIERARCHICAL CORPORA

A. Novelty and Motivation

The *purpose* of this research is development of the methodology of constructing and analyzing the *Hierarchical Corpora* intended for subsequent use in the creation of the Sentiment Lexicon using text mining tools.

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

RQ: *Using what methods and algorithms it is possible to increase the quality of the formation of the Corpus intended for the analysis of text tonality*?

RQ_1: *Does creation of the Contextual Structure of the Corpus provides advanced possibilities for identification of the morphological and tonal specifics of analyzed texts?*

RQ_2: *Does the preliminary Morphological and Structural analysis of the Corpus allow to reveal specific characteristics of Corpus content in the light of improving the quality of texts Sentiment recognition?*

For finding the answers for these questions, the following assumptions (A) were formulated:

A1: Taking into account the specificity of the chosen case study [16, 17, 29], assume that each paragraph could be interpreted as a topically completed textual component (TCTC).

A2. Classified texts are characterized by their initially known subjective (author's) evaluations of their tonality.

On the basis of the research questions and assumptions, the following scientific hypotheses (H) were formulated:

H1. *Contextual structure of the certain type of texts writing does not depend on its tonality, initially assigned to it by the author.*

H2. *Morphological structure of the certain type of texts writing does not depend on its tonality, initially assigned to it by the author.*

H3. *Writing style of certain type of texts does not depend on the tonality, initially assigned to it by the author.*

H4. *Vocabulary richness of certain type of texts does not depend on the tonality, initially assigned to it by the author.*

As a *case study* for testing the basic workability and proposed Methodology quality, the Polish-language Film Reviews Dataset was used.

B. Contextual Component of the Corpora's Contextually-Oriented Hierarchical Structure Building

At the first stage of the methodology development, the authors take into account specificity of the chosen case study and the results of previous research results [29]. These results suggest the possibility of building the Hierarchical structure of the Contextually-Oriented Corpus (COHC) via application of the Discriminant and Probabilistic Methods of the Latent Dirichlet Allocation (LDA) and Latent Semantic Relations Analysis (LSA) [28, 30]. In our case COHC is the two-point (Positive/Negative Classes) structure of the sets of paragraphs, semantically close to Topics, identified as the main Contextual Framework of the analyzed initial Dataset. The process of Hierarchical structure of COHC *Building* involves the following stages (figure 1):

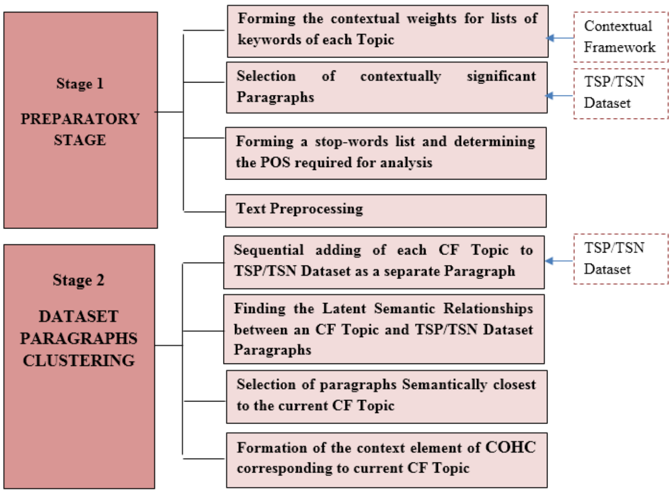
Stage 1
PREPARATORY STAGE
Forming the contextual weights for lists of keywords of each Topic
Contextual Framework
Selection of contextually significant Paragraphs
TSP/TSN Dataset
Forming a stop-words list and determining the POS required for analysis
Text Preprocessing
Stage 2
DATASET PARAGRAPHS CLUSTERING
Sequential adding of each CF Topic to TSP/TSN Dataset as a separate Paragraph
TSP/TSN Dataset
Finding the Latent Semantic Relationships between an CF Topic and TSP/TSN Dataset Paragraphs
Selection of paragraphs Semantically closest to the current CF Topic
Formation of the context element of COHC corresponding to current CF Topic

Fig 1. Steps of the Corpora's Contextually-Oriented Hierarchical Structure Building

Formation and analysis of the Contextual components of the COHC are based on the following concepts:

Concept A. Obtaining an adequate sentiment description of positive and negative texts tonality is possible only via formation of the Corpus on the basis of Truly Subjectively Positive (TSP, the subjective evaluation by reviewes is more than 8 points) and Truly Subjectively Negative (TSN, the subjective evaluation is less than 4 points) Dataset.

Concept B. As a Contextual Framework (CF) for COHC building the hierarchical structure of Topics (with list of keywords) for TSP and TSN dataset are used. Applied methods for CF creating – combination of LSA and LDA methods [16, 17, 28].

Concept C. As a quantitative measure of the degree of influence of each keyword from the CF Topics on the process of COHC building the contextual keyword weights (CKW) are used. Applied methods for CKW creation – combination of LSA and LDA methods, measure – the probability of occurrence of each word in the topic [16, 17, 28].

Concept D. As a tool for determining the belonging of each paragraph to a CF topic, the LSA is used [16, 17, 28].

Concept E. As a TCTC a paragraph of at least 100 characters should be used. The possibility to determine the topic of such paragraph with sufficient accuracy is experimentally proved [16, 17, 28].

C. Text Analysis of the Contextually-Oriented Hierarchical Corpus

The process of Text Analysis of COHC involves the following steps (figure 2).

Step 1. Morphological Analysis of the Contextually-Oriented Hierarchical Corpus

The purpose of this first stage is to conduct the COHC analysis and create the Hierarchical Morphological Framework (HMF) for each element of each COHC level.

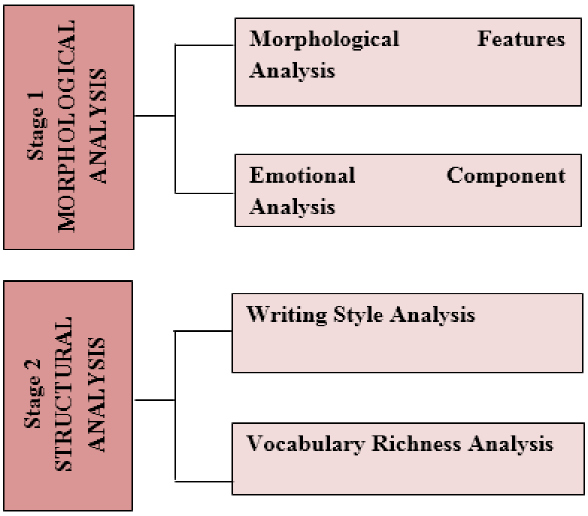
Stage 1
MORPHOLOGICAL ANALYSIS
Morphological Features Analysis
Emotional Component Analysis
Stage 2
STRUCTURAL ANALYSIS
Writing Style Analysis
Vocabulary Richness Analysis

Fig 2. Steps of the Text Analysis of Contextually-Oriented Hierarchical Corpus

The objective of creating this HMF is to accumulate the Hierarchy of specific morphological types and emotional features of the COHC texts to identify their differences depending on: the text tonality, initially assigned to it by the author, and belonging to a particular CF Topic. As main indicators to carry out this stage of analysis, the measures with following interpretations are proposed:

– the part of speech (POS) distribution for each COHC Element (*M1*) – determining the authors morphological types of the expression of positive or negative judgments;

– the percentage of new (unique) words in each COHC Element (*M2*) – characterizing the emotional component specificity of positive or negative judgments expression.

Step 2. Structural Analysis of the Contextually-Oriented Hierarchical Corpus

The purpose of this first stage is to conduct the COHC analysis and create the Hierarchical Structural Framework (HSF) for each COHC element of each level. The objective of creating this HSF is to accumulate the Hierarchy of specific writing styles and vocabulary richness features of the COHC texts to identify their differences depending on: the text tonality, initially assigned to it by the author, and belonging to a particular CF Topic. As main indicators to carry out this stage of analysis, the measures with following interpretations are proposed:

– the specificity of first Zipf's "rank-frequency" law for each element of each COHC level (*M3*), determining the authors *writing styles* of the expression of positive or negative judgments and classically characterized by a constant value of C as the ratio [31, 32]:

$$C= F*R, \qquad (7)$$

where: F – frequency of occurrence of a term in the text; R – Rank of the word (the most commonly used word gets rank 1, the next – 2, etc.); C – constant;

– the specificity of the second Zipf's "quantity-frequency" law for each element of each COHC level (*M4*), determining the authors *vocabulary richness* of the expression of positive or negative judgments [5,6].

Case Study Results and Discussion

For testing and evaluating the adequacy of the author's Methodology realization, as a *case study* were used the training samples: 3000 Polish-language films reviews (1500 TSP and 1500 TSN) from the filmweb.pl.

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

D. Contextual Component of the Corpora's Contextually-Oriented Hierarchical Structure Building

In the process of implementing this stage, about 30% of contextually insignificant paragraphs, and about 20% of paragraphs, for which topic could not be identified, were separated.

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

As a method of evaluating the quality of probabilistic topic models the calculation of *Perplexity* index on the test data set [2-4] is used. In information theory, perplexity is a measurement of how well a probability model predicts a sample. A low perplexity indicates that the probability distribution is good for predicting the sample.

Table I.
Structure of the Contextual Summary

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

As a result, the following structure of the two-level two-point *Contextually-Oriented Hierarchical Corpora of Polish-Language Film Reviews* [29] was obtained (tables II- III).

Table II.
Contextual Structure of the Subjectively Positive (SP) Elements of the COHC of Polish-Language Film Reviews

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

Table III.
Contextual Structure of the Subjectively Negative (SN) Elements of the COHC of Polish-Language Film Reviews

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

The results obtained at this stage of the experiment indicate that the tonality, initially assigned by the authors specificity of texts with Persuasive writing type, affects the Contextual Structure of the analyzed content. As can be seen from the tables II-III, the Contextual Structure of Subjectively Positive and Subjectively Negative elements of COHC differs both in content and in the variety (amount) of topics covered in the texts (H2 is rejected).

E. Text Analysis of the Contextually-Oriented Hierarchical Corpus

Step 1. Morphological Analysis of the Contextually-Oriented Hierarchical Corpus

As a result of the specific morphological types and emotional features of the COHC texts identification, the following initial statistics were obtained (Table IV)

TABLE IV.
FREQUENCY CHARACTERISTICS OF THE PARTS OF SPEECH DISTRIBUTION IN SP / SN HIERARCHICAL CORPORA'S ELEMENTS

Negative Hierarchical Corpora's Elements				Positive Hierarchical Corpora's Elements			
Morphological Types (M1)		Emotional Features (M2)		Morphological Types (M1)		Emotional Features (M2)	
% of adjectives	**Frequency**	**% of Unique adjectives**	**Frequency**	**% of adjectives**	**Frequency**	**% of Unique adjectives**	**Frequency**
18,09003	1	46	1	12.50	2	38	1
19,61974	1	53.5	2	15.02	0	50.4	2
21,14945	4	61	3	17.55	0	62.8	2
22,67915	9	68.5	4	20.07	0	75.2	9
More	1	More	6	22.59	12	87.6	6
				More	11	More	5
% of nouns	**Frequency**	**% of Unique nouns**	**Frequency**	**% of nouns**	**Frequency**	**% of Unique nouns**	**Frequency**
0.906111	1	38	1	52.88136	1	31	1
15.52333	0	46.5	2	56.16872	5	42.6	1
30.14056	0	55	5	59.45609	12	54.2	3
44.75778	0	63.5	4	62.74345	4	65.8	8
More	15	More	4	66.03082	1	77.4	7
				More	2	More	5
% of verbs	**Frequency**	**% Unique verbs**	**Frequency**	**% verbs**	**Frequency**	**% of Unique verbs**	**Frequency**
20.02801	1	51	1	13.43874	1	43	1
21.01446	4	60.75	2	15.22556	0	54.4	1
22.00092	6	70.5	3	17.01239	2	65.8	1
22.98737	2	80.25	6	18.79922	3	77.2	4
More	3	More	4	20.58605	5	88.6	9
				More	14	More	9

In table IV the "% of part of speech" is the border of % of these types of words in the whole number of words in particular Corpora's Element; "Frequency" – the number of COHC elements in which this "% of part of speech" occurs.

The results of comparative analysis of differences in the part of speech distribution and percentage of new words in the different COHC elements could be interpreted in the following way:

1. The law distribution of adjectives in the positive and negative COHC *Elements* indicates that:

– % of adjectives used in *positive* elements of the Hierarchical Corpora is slightly higher than this percentage in *negative* Elements.

This result can be interpreted as the presence of a general tendency to make reviews more intonational in expressing positive emotions;

– % of new (unique) adjectives used in positive elements of the Hierarchical Corpora is significantly higher (by 20%) compared to this indicator in negative Elements.

Thus, the need and realization of the emotional component of the authors' positive judgments through the use of different adjectives (characteristics) is much higher than in the expression of negative emotions.

2. The distribution of nouns in the positive and negative COHC Elements indicates that:

– % of the nouns, used in positive Hierarchical Corpora's Elements, obeys the classical normal distribution law and indicates the average weightiness of the judgments expressed. Negative judgments are characterized by extremes – either many, or very few nouns;

– % of new (unique) nouns used in positive Hierarchical Corpora's Elements is higher (about 10%) compared to this indicator in negative reviews.

Since the nouns primarily serve to ascertain the facts, the existence of objects (entities, etc.), on the whole these facts can indicate that negative judgments are more based on emotions rather than facts.

3. The distribution law of verbs in the positive and negative COHC Elements indicates that:

– % of verbs used in *positive* elements of the Hierarchical Corpora is insignificant, and even lower than this percentage in *negative* Elements.

This can be interpreted as the desire of reviewers to characterize the actions that caused the particular emotions, in a greater degree;

– % of new (unique) verbs used in *positive* elements of the Hierarchical Corpora is higher (about 10%) compared to this indicator in negative reviews.

This again testifies to a more creative approach to writing reviews by authors of positive reviews.

In general, these facts may indicate that the expression of negative emotions is characterized by greater stinginess of emotional coloring (from the point of view of linguistic evaluation).

When generalizing the results obtained at this stage of the experiment, it can be argued that the tonality, initially assigned by the authors specificity of texts with Persuasive writing type, affects the Morphological Structure of the analyzed content. As it can be seen from table V, the Morphological Structure of Subjectively Positive and Subjectively Negative elements of COHC differs both in the morphological types and emotional features (H2 is rejected).

TABLE V.
MORPHOLOGICAL TYPES AND EMOTIONAL FEATURES IN SP / SN HIERARCHICAL CORPORA'S ELEMENTS

Part of speech	**SP Elements**		**SN Elements**	
	M1	**M2**	**M1**	**M2**
Adjectives	High	High	High	High
Nouns	Average	High	Polar	High and Average
Verbs	High	High	Average	Average

Step 2. Structural Analysis of the Contextually-Oriented Hierarchical Corpus

a) First Zipf's "Rank-Frequency" Law

As a result of comparative analysis of writing styles of the expression of positive or negative judgments, the following types of internal structure of the COHC Elements from the point of view of the specific of the Rank-Frequency distribution (*writing styles*) of the words usage were identified:

– the *Classical* structure of the COHC Elements, in which the proportion of terms with a high frequency of usage (in the authors' algorithm it is intended to remove the often-used and not load-bearing stop-words of the Polish language at the stage of preprocessing) account for no more than 0.25% of all terms (C≈0.06-0.07);

– the *Medium normalized* structure of the COHC Elements, in which terms with a high frequency of use account for about 10% of all terms (C≈0.02-0.05);

– the *non-standard* COHC Elements structure, in which about 90% of terms have a frequency of use no more than 1 time (C≈0.03-0.04).

TABLE VI.
THE RANK-FREQUENCY DISTRIBUTION STRUCTURE OF THE 1ST LEVEL OF HIERARCHICAL CORPORA'S ELEMENTS

Element of 1st level CF	**Classical**				**Medium normalized**				**Non-standard**			
	All POS	**Adjective**	**Nouns**	**Verbs**	**All POS**	**Adjective**	**Nouns**	**Verbs**	**All POS**	**Adjective**	**Nouns**	**Verbs**
Positive Hierarchical Corpora's Elements												
"Hero"	18.75%	12.50%	25.00%		6.25%	6.25%		18.75%	6.25%			6.25%
"Spectator"					18.75%	6.25%	25.00%		6.25%	18.75%		25.00%
"Script"	25.00%	12.50%	25.00%					12.50%		12.50%		12.50%
"Director"	12.50%		6.25%		6.25%	12.50%	12.50%	12.50%	6.25%	12.50%	6.25%	12.50%
"Plot"	6.25%		6.25%		12.50%	6.25%	12.50%	6.25%	6.25%	18.75%	6.25%	18.75%
Negative Hierarchical Corpora's Elements												
"Hero"	25.00%		25.00%	8.33%		25.00%		16.67%				
"Actor"	16.67%		25.00%		8.33%	16.67%		8.33%		8.33%		16.67%
"Creator"	25.00%		25.00%			25.00%		16.67%				8.33%
"Plot"	16.67%		16.67%		8.33%	25.00%	8.33%					25.00%

Characteristics of Rank-Frequency distribution, obtained during this stage of experiment for the 1st level of COHC, are presented in table VI.

Interpretation of these results as a characteristic of specific Writing Style, could be the following:

– negative reviews are characterized by a small part of opinions, expressed with the help of unique words. And the most non-standard from the point of view of the use of unique words and the brevity of presentation are the paragraphs characterizing the topic Plot of the film with the help verbs.

Adjectives used in negative reviews mainly refer to the second type of Hierarchical Corpora structure, which can be interpreted as a fairly high percentage of frequently repeated definitions (terms).

The predominant type of COHC Elements structure are standard reviews, in which the most commonly used words account for 0.25% of all COHC Element vocabulary;

– structure of the *positive* part of the COHC is fairly uniform – all the texts represented in it are equally structured. However, from the point of view of distinctive features, it is necessary to note the percentage of non-standard (unique) adjectives.

The structure of the COHC that characterizes the Spectator is especially different – in this part of the COHC there is no standard Corpora Elements structure, and there is a high percentage of both repeating and unique verbs.

Generalizing the results obtained at this stage of the experiment (table VII) it again can be argued that the tonality, initially assigned by the authors specificity of texts, affects the Writing Style of the analyzed content (H3 is rejected).

TABLE VII.
WRITING STYLES STRUCTURE OF THE HIERARCHICAL CORPORA OF POLISH-LANGUAGE FILM REVIEWS

Writing Style (M3)	SP Hierarchical Corpora's Element	SN Hierarchical Corpora's Element
Classical	30.00%	45.83%
Medium normalized	35.00%	39.58%
Non-standard	35.00%	14.58%

b) Second Zipf's "Quantity-Frequency" Law

As a result of the experiments, the initial statistics, which describes the specificity of second Zipf's "quantity-frequency" law for Polish-Language Film Reviews COHC were obtained.

The basic coefficients for the analysis were the coefficients of the approximation in the equation for the second Zipf's law:

$$y = a + \frac{b}{x} \qquad (8),$$

where, a – determining the average frequency of occurrence of most part of the terms in the COHC Element; b – determining the average speed of appearance of new words in the text – the *Vocabulary Richness* (M4) of text's author.

The *lower* the value of the coefficient b, the higher the richness of the vocabulary of the COHC Element, since the curve of the dependence of the occurrence frequency of each word in the number of these words decreases more quickly, accordingly a smaller number of terms appears frequently (that is, the same words are used more often).

In order to ensure the adequate comparability of the conducted studies results, the corrected (taking into account the number of unique words in the COHC Element) coefficients of equations a and b were used.

The characteristics of Quantity-Frequency distribution (Zipf's law coefficients), obtained during this stage of experiment, are presented in the figures 3-4.

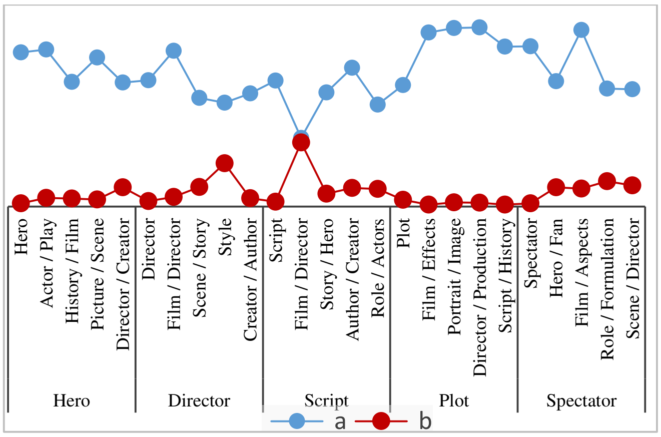

Fig 3. Second Zipf's law coefficients for Positive Hierarchical Corpora's Elements

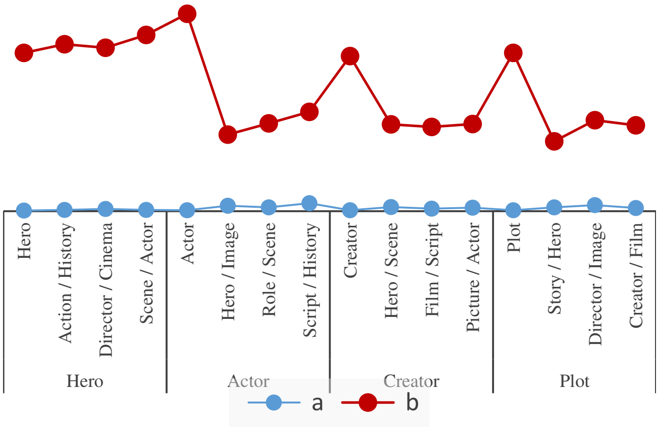

Fig 4. Second Zipf's law coefficients for Negative Hierarchical Corpora's Elements

Based on the obtained data, the comparative analysis of the specificity of second Zipf's "quantity-frequency" law coefficients distribution of the Positive and Negative Elements of Polish-Language Film Reviews COHC was carried out (table VIII), where "Frequency" – the number of COHC elements in which such value of coefficients occurs:

TABLE VIII.
STRUCTURE OF THE SECOND ZIPF'S LAW COEFFICIENTS DISTRIBUTION (M4)

Positive COHC Elements				Negative COHC Elements			
Coefficient a	Frequency	Coefficient b	Frequency	Coefficient a	Frequency	Coefficient b	Frequency
0.016	3	0.004	22	0.001	9	0.029	9
0.022	12	0.007	2	0.001	5	0.041	0
0.028	10	0.010	1	0.002	2	0.052	7

Interpretation of the specificity of Polish-Language Film Reviews COHC from the point of view of the vocabulary richness of the expression of positive or negative judgments (table VIII) could be the following:

– *positive* reviews are characterized by an initially high (in comparison with negative) values of the corrected coefficient ***a*** (a≈0.010-0.027), which indicates a high average level of frequency of most part of the terms in the COHC Element. At the same time, this part of the case is characterized by relatively low values of the corrected coefficient ***b*** (b≈0.0003-0.0009), on the one hand, testifying to a sufficiently rich (in comparison with negative reviews) vocabulary of the text. That is, in general, positive reviews characterized by a greater proportion of terms that are used *uniformly often* throughout the text. This, in turn, may indicate a rather *highly semantic structured opinion*, expressed in a carefully and balanced manner;

– the *negative* reviews are characterized by sufficiently low (in comparison with positive) values of the corrected coefficient ***a*** (a≈0.0002-0.0021), which indicates a lower (i.e., more unique) average level of frequency of most part of the terms in the COHC Element. In this case, this part of the COHC is characterized by sufficiently high values of the corrected coefficient ***b*** (b≈0.0183-0.0517), which may indicate that in the expression of the main negative emotions, authors use the same words quite often, and the rest of the words are used *randomly*, depending on the context of the film or the specific expression of the author's thoughts. This, in turn, can testify to the *average level of semantic structure of the opinion*, expressed more spontaneously and under the influence of emotions.

Generalizing the results obtained at this stage of the experiment (table IX) it again can be argued that the tonality, initially assigned by the authors specificity of texts, affects the Vocabulary Richness of the analyzed content (H4 is rejected).

TABLE IX.
VOCABULARY RICHNESS STRUCTURE OF THE HIERARCHICAL CORPORA OF POLISH-LANGUAGE FILM REVIEWS

Vocabulary Richness (M4)	**SP COHC Element**	**SN COHC Element**
Average frequency of words occurrence	High	Low
Average speed of new words appearance	Low	Polar
Vocabulary Richness	Sufficiently High	Random
Semantic Structuredness	Highly Semantic Structured	Medium Semantically Structured

CONCLUSIONS

In this paper, authors present the methodology of constructing and analyzing the *Hierarchical Corpora* for the Purpose of Further Forming and Training the Sentiment Lexicon. The main contribution of the paper and the authors' study is finding the answers to the main research questions:

1. The hierarchical structure of the Corpus allows for more flexible and clear identification of the Morphological and Structural features of texts of different tonalities and contexts, originally assigned by the authors.

These differences should contribute to improving the quality of algorithms development for the Sentiment Lexicon creation, allowing the introduction of additional tools for determining the tonality based on the availability of data about semantic properties of the text being studied.

2. The Contextual, Morphological and Structural specificity of texts that have a tone, originally assigned by the authors, has significant differences. The basic features of the Sentiment Patterns for the texts of the two basic tonalities, which were obtained, are the following (table X):

Table X.
SENTIMENT PATTERNS OF THE TEXTS WITH THE DIFFERENT TONALITY

Features	**Subjectively Positive Texts**	**Subjectively Negative Texts**
Contextual Structure	Wide	Moderate
Morphological Types	More Adjectives and Verbs	More Adjectives and Partly Verbs
Emotional Features	Very Emotional	Restrained Emotionally
Writing Style	Colorful	Monochrome
Vocabulary Richness	Structured and rich	Medium Structured and Random

The frameworks obtained by the authors testify to the existence of certain thought and style templates, under the influence of which the formation of texts of various sentiments takes place.

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".

REFERENCES

[1] Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. *Proceedings of Workshop at International Conference on Learning Representations* (ICLP), 2013. http://arxiv.org/abs/1301.3781

[2] Feldman, R. Techniques and applications for sentiment analysis. *Communications of the ACM*, 2013. 56(4), pp. 82-89.

[3] Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Word and Phrases and their Compositionaly. *Proceedings of Workshop at The Twenty-seventh Annual Conference on Neural Information Processing Systems.* 2013 (NIPS) http://arxiv.org/abs/1310.4546

[4] Mikolov T., Le Q. Distributed Representations of Sentences and Documents. *Proceedings of Workshop at The 31st International Conference on Machine Learning* (ICML). 2014. http://jmlr.org/proceedings/papers/v32/le14.pdf.

[5] Elias P. Interval and recency rank source encoding: two on-line adaptive variable-length schemes. *IEEE Trans. Inform. Theory*. 1987. V. 33, N 1. P. 3–10.

[6] Popescu, I.-I., Altmann, G., Čech, R. The Lambda-structure of Texts. *Lüdenscheid: RAM-Verlag,* 2011.
(1) Vocabulary Richness Measure in Genres. Available from: https://www.researchgate.net/publication/258518594_Vocabulary_Richness_Measure_in_Genres [accessed Jul 10 2018].

[7] Dempster, A.P., Laird, N.M., & Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*. 1977. Series B., 39(1), 1-38.

[8] Feinerer, I., Hornik, K. & Meyer, D. Text mining infrastructure in R. *Journal of statistical software*, 2008. 25(5). American Statistical Association.

[9] Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications*. 2003.

[10] Koreniu T., Laurikkala J., Järvelin K., & Juhola M. Stemming and Lemmatization in the Clustering of Finnish Text Documents. *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004. Washington, DC, USA, 625-633.

[11] Alkula, R. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Information Retrieval*, 2001. 4, 195-208.

[12] Nokel, M. A. & Lukashevich, N.V. Thematic Models: Adding Bigrams and Accounting Similarities Between Unigrams and Bigrams. *Computational methods and programming*, 2015. 16, 215-217

[13] Salton G., Wong A., Yang C.S. (A vector space model for automatic indexing. *Communications of the ACM*. 1975. Volume 18. Issue 11, pp. 613-620

[14] Jain A.K., Murty M.N. & Flynn P.J. Data Clustering: A Review; *ACM Computing Surveys*, 1999. 31 (3), 264-323. http://dx.doi.org/10.1145/331499.331504

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

[16] Rizun N., Ossowska K., Taranenko Y. Modeling the Customer's Contextual Expectations Based on Latent Semantic Analysis Algorithms. *Information Systems Architecture and Technology*: 38th

International Conference on Information Systems Architecture and Technology. 2018, pp.364-373.

[17] 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.

[18] Patricia J. Crossno, Andrew T. Wilson and Timothy M. Shead, Daniel M. Dunlavy. Topic View: *Visually Comparing Topic Models of Text Collections*. 2011.

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

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

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

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

[23] Blei, D. M. Introduction to Probabilistic Topic Models. *Communications of the ACM*, 2012. 55 (4), 77-84.

[24] Blei, D. M., Ng, A., & Jordan, M. Latent Dirichlet Allocation. *International Journal of Advanced Computer Science and Applications (3):* 2003. 147-153.

[25] Anagha R Moosad, Aiswarya V., Subathra P and P.N. Kumar. Browsing Behavioural Analysis Using Topic Modelling. *International Joural of Computer Technology and Applications*. 2015. No.8. Issue No. :5. Pp. 1853-1861

[26] Alghamdi R. Alfalqi K. A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications* (IJACSA), 2015. Volume 6 Issue 1.

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

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

[29] 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,* 2017, pp.366-372.

[30] Rizun N., Kucharska W. Text Mining Algorithms for Extracting Brand Knowledge from Facebook. The Fashion Industry Case. *International Business Information Management Conference*. 2018.

[31] Mandelbrot B. On recurrent noise limiting coding. *Lab. d'Electronique et de physique appliquces*. 1954. Paris.

[32] Mandelbrot B. In the theory of word frequencies and on related markovian models of discourse. The structure of language and its mathematical aspects. Providence, RI: Amer. Math. Soc. 1961. pp. 190–219. *Proceeding Symposium on Applied Mathematics*; V. 12.