Methodology for Text Classification using Manually Created Corpora-based Sentiment Dictionary

Nina Rizun[1] and Wojciech Waloszek[2]
[1]*Department of Applied Informatics in Management, Gdansk University of Technology, Gdansk, Poland*
[2]*Department of Software Engineering, Gdansk University of Technology, Gdansk, Poland*

Keywords: Textual Content Classification, Hierarchical Sentiment Dictionary, Text Tonality, Evaluation the Quality, Bigrams, Polarity Scores.

Abstract: This paper presents the methodology of Textual Content Classification, which is based on a combination of algorithms: preliminary formation of a contextual framework for the texts in particular problem area; manual creation of the Hierarchical Sentiment Dictionary (HSD) on the basis of a topically-oriented Corpus; tonality texts recognition via using HSD for analysing the documents as a collection of topically completed fragments (paragraphs). For verification of the proposed methodology a case study of Polish-language film reviews Corpora was used. The main scientific contributions of this research are: writing style of the analyzed text determines the possibility of adaptation of the Texts Classification algorithms; Hierarchically-oriented Structure of the HSD allows customizing the classification process to qualitative recognition of text tonality in the context of individual paragraphs topics; texts of Persuasive style most often are initially empowered by authors with a certain tonality. The tone, expressed in the author's opinion, effects the qualitative indicators of sentiment recognition. Negative emotions of the author usually reduce the level of vocabulary variability as well as the variety of topics raised in the document, but simultaneously increase the level of unpredictability of words contextually used with both positive and negative emotional coloring.

1 INTRODUCTION

Since the early 2000s, the proliferation of computer technology has contributed to the fact that the procedure for creating information content has become available to almost everyone. The content of such information resources as social networks, feedback collection services, web forums and blogs, is actively formed by the users themselves.

Consolidated subjective experience of individual users is a source of valuable information (Hu, 2004). A special section of computer linguistics is devoted to extraction of such information – automatic analysis of text tonality (Sentiment Analysis or Opinion Mining) (Liu, 2012).

The initial goal of Sentiment analysis methods was classification of documents, and later of sentences, according to a given scale of tonality, usually a two-point (positive-negative) or three-point (positive-negative-neutral). However, instead of a general assessment of tonality, a more detailed study of the expressed views on specific aspects (contexts) is required. Therefore, over time, the initial formulation of the task of tonality analysis has acquired a more detailed formulation and has emerged as a separate problem of contextually-oriented sentiment analysis, which is to automatically determine the views of the user, expressed in the text, with respect to specific aspects of the entity being examined.

This study is devoted to finding ways of increasing the quality of the *Textual Content Classification* via effective implementation of the *Topics Modeling* approach for creating and further using *Hierarchical Contextually-Oriented Sentiment Dictionary*.

2 THEORETICAL BACKGROUND

Methods of contextually-emotional analysis of the text are developed within the framework of two machine learning approaches: supervised and unsupervised machine learning (Liu, 2012). In the approach based on supervised machine learning, a marked collection of documents is needed, which lists examples of emotional expressions and aspect terms.

Rizun, N. and Waloszek, W.
Methodology for Text Classification using Manually Created Corpora-based Sentiment Dictionary.
DOI: 10.5220/0006932602120220
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 212-220
ISBN: 978-989-758-330-8

The methods of unsupervised machine learning allow to avoid dependence on training data. For their work, one also needs a Corpus of documents, but preliminary markup is not required. Within the framework of this approach, the probabilistic-statistical regularities of the text are found and, on their basis, the key subtasks of the aspect-emotional analysis are solved: identification of aspect terms and determination of their tonality. However, such methods require complex tuning to a given domain. For example, the method based on Latent Dirichlet Allocation (LDA) in its original form is not able to effectively detect topics, therefore, its additional adaptation and adjustment of correspondence of identified topics to the target set of contexts is required (Titov, 2008).

The methods of Text Classification, considered above, requires the presence of Sentiment Dictionary of text tonality evaluation. There are three basic approaches to such Dictionary (Liu, 2012): expert; based on dictionaries / thesaurus; and on the basis of text collections.

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

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

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

In (Klekovkina and Kotelnikov, 2012), the dictionary of emotional vocabulary, compiled by experts manually, was used to determine the tone of individual words. In the dictionary, each word and phrase are associated with orientation of the key (positive / negative) and with strength (in points).

The author's methods proposed in (Taboada et al., 2011; Boiy, 2007) are based on a dictionary approach: to determine the tonality of texts, a dictionary of estimated words is used, where each word has a numerical weight that determines the degree of word significance. In the method of working with the dictionary closest to the paper (Boucher and Osgood, 1969), however: the dictionary firstly is created on the basis of a statistical analysis of training collection; secondly, the weight of words is determined with the help of a genetic algorithm.

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

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

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

To classify texts according to the linear function:

$$f\left(W_T^{pos}, W_T^{neg}\right) = W_T^{pos} + k_{neg} \bullet W_T^{neg} \quad (2)$$

where W_T^{pos} is the positive weight of the text T; W_T^{neg} is the negative weight of the text T; k_{neg} – coefficient, compensating the fact of preponderance of positive vocabulary in text (Pang, 2008). If the value of the function f is greater than zero, the text is positive, otherwise – negative.

3 METHODOLOGY OF WEBSITES CONTENT SENTIMENT CLASSIFICATION

The *objective* of this research is testing and evaluation of *Text Classification Methodology grounded on the Manually Created Corpora-based Sentiment Dictionary* (SC- methodology).

The developed Methodology assumes realization of three main practical stages:

1) Manual Creation of Corpora-based Sentiment Dictionary (CBSD).

2) Carrying out Texts Classification based on created CBSD.

3) Evaluation of the adequacy of Texts Classification results.

As a *case study* for testing the basic workability and proposed Methodology quality the Polish-language Film Reviews Corpora will be used.

3.1 Novelty and Motivation

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

RQ_1: *Does the structure of the Sentiment Dictionary influence the quality of classification?*

RQ_2: *Does the writing style of the analyzed text influence the quality of Classification?*

RQ_3: *Does the tone, defined in the text by its author, influence the quality of Classification?*

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

A1: Taking into account the specificity of the chosen case study and presence of the nonofficial requirements of film's review writing style and structure (Rizun et al. 2017a; Rizun et al. 2017b), assume that each paragraph could be interpreted as a topically completed textual component (TCTC).

A2: Each document has a complex structure and can be estimated integrally based on separated classification of TCTC as elements of their structure.

A3. Classified texts are characterized by their initially known subjective (author's) evaluations of their tonality.

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

H1. *Quality of the Text Classification is increased thanks to an integral evaluation of its individual topically-oriented fragments, using the Contextually Structured Sentiment Dictionary.*

H2. *Quality of the classification does not depend on tonality, subjectively assigned to texts by the author.*

3.2 Algorithm of Manually Creating the Corpora-based Sentiment Dictionary

At the first stage of development of the SC-methodology, the authors consider specificity of the chosen case study and the results of previous research (Rizun et al. 2017b; Rizun and Taranenko, 2018). These results suggest the possibility of preforming the Hierarchical structure of the Contextually-Oriented Corpus (HC) via application of unsupervised machine learning Discriminant and Probabilistic Methods of the Topic Modelling and Latent Semantic Relations Analysis.

HC 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 Corpus. Such analysis presupposes division of the texts within the Corpus into Truly Subjectively Positive (TSP) and Truly Subjectively Negative (TSN) Corpora Samples.

For realization of such Corpora Samples construction, the following Assumptions were adopted:

A4: 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).

A5: As elements of the Sentiment Dictionary, it is suggested to use bigrams, which allow to implement contextual structure of the created dictionary. One of the two components of bigram must be the keywords, characterizing each Element of the Contextually-Oriented Corpus.

Based on the HC, it becomes possible to implement the following version of the *Algorithm of Manually Creating Hierarchical (Topically oriented) Corpora-based Sentiment Dictionary* (CBSD) (Figure 1).

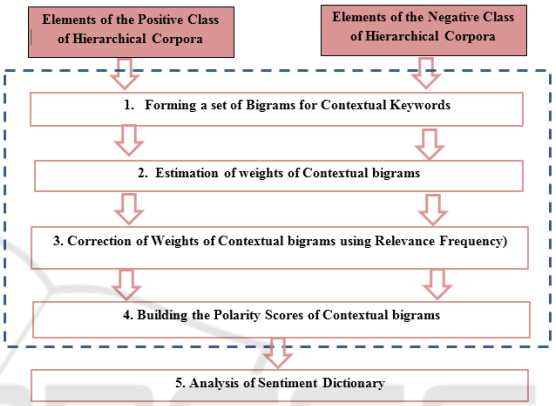

Figure 1: Algorithm of Manually Creating the Corpora-based Sentiment Dictionary

Step 1 is implemented using functions в NLTK **nltk.bigrams(...)** for Contextual Keywords for each Topic of Positive/Negative Elements.

Step 2: definition of the absolute bigrams weight (WB), estimated by the frequency of occurrence of this bigram in the elements of Corpora.

Step 3: increase of the degree of accuracy of the WB estimation using parameter to reverse the frequency – RF (Relevance Frequency) (Ivanov et al., 2015):

$$RF_S = \log_2\left(2 + \frac{a}{\max(1,b)}\right) \quad (3)$$

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

Step 4: scaling of the Relevance Frequency obtained at the previous stage to 0.

Step 5: analysis of the structure of received elements of the Manually Creating Corpora-based Sentiment Dictionary.

3.3 Algorithm of Sentiment Classification of the Textual Corpora

As the second stage of development of the SC-methodology, the basic version of proposed *Algorithm of Sentiment Classification of the Texts based on the CBSD* is presented in Figure 2.

The main steps for implementing this algorithm are the following:

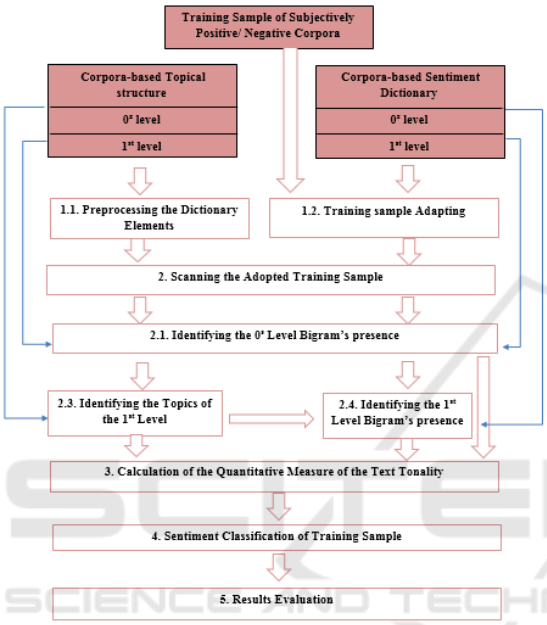

Figure 2: Algorithm of Sentiment Classification of the Texts based on the Manually Created CBSD.

Step 1. Preparing to Perform the Sentiment Classification Procedure

The stage of Sentiment Dictionary preprocessing is performance of decapitalization and lemmatization of the dictionary elements.

In the step of Training Sample preparing, taking into account the specificity of case study, as well as limited number of algorithmic implementations for the analysis of texts in Polish (Rizun and Taranenko, 2017), in addition to standard procedures for text pre-processing, the authors have provided text *adaptation* procedure (Rizun et al. 2017b).

Step 2. Scanning the Corpora Sample to Identify the Presence of Sentiment Dictionary Elements

With the purpose of acceptance / rejection of the Hypothesis 1, this step of algorithm involves implementation of the following procedures of scanning the Subjectively Positive/Negative Corpora Samples (SPCS/SNCS).

1) using CBSD without considering their Topical structure – simple classification;

2) using CBSD with considering their Contextual Framework – one-level classification.

As it was accepted in this study as A1 and A2, scanning and recognition of topics for One Level classification will be performed by paragraphs of the review. The importance of this approach is confirmed in (Rizun et al. 2017a), where it is proved that the accuracy of recognition of topics in the document is much higher with each document (paragraph) being focused on a single topic.

In addition to evaluation of words sentiment, word-modifiers, which shift the tonality weight of the neighboring words, were used.

Step 3. Formation of the Basic Quantitative Measures of Text Tonality Evaluation

To determine the quantitative measure of tonality estimated for the entire text of document T from Subjectively Corpora Samples, the number of *positive*, *neutral* and *negative* bigrams from the corresponding CBSD, found in Texts in accordance with the rules in Table 1, is calculated.

Corresponding to the found bigrams, *Polarity scores* w_i^{pos}, w_i^{neu} and w_i^{neg} are summed up.

$$W_T^{pos} = \sum_{i=1}^{N_C^{pos}} w_i^{pos}, W_T^{neu} = \sum_{i=1}^{N_C^{neu}} w_i^{neu}, W_T^{neg} = \sum_{i=1}^{N_C^{neg}} w_i^{neg} \quad (4)$$

where W_T – 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. We can find the final *basic* estimator of the texts tonality according to the linear function:

$$f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right) = W_T^{pos} + W_T^{neu} + W_T^{neg} \quad (5)$$

Step 4. Sentiment Classification of Training Sample

Implementation of this step involves usage of the following provisions:

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

– for SPCS a text has: *High* positive tonality (HP). *Quite* positive tonality (QP). *Reasonably* positive tonality (RP).

– for SNCS a text has: *Rather* negative tonality (RN). *Clearly* negative tonality (CN). *Absolutely* negative tonality (AN).

Rule 2. To implement the training procedure for the algorithm being developed, the Sentiment Classification of texts is suggested using the following *Rating Formats*:

1. Basic quantitative measure of the text tonality:

$$R_1 = f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right) \quad (6)$$

2. Text tonality, normalized considering the *size* of the original *document* S_T:

$$R_2 = f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right)/S_T \quad (7)$$

Introduction of this indicator is due to the ability to estimate the *percentage of the text's tonality* R_1 in the total number of words in the document.

3. Text tonality, normalized taking into account the *number* of *bigrams* found in the original document N_C:

$$R_3 = f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right)/_{N_C^T} \quad (8)$$

Introduction of this indicator is due to the ability to estimate the *average weight of the text's tonality* in one bigram, recognized in the document.

4. Text tonality, normalized taking into account the *ratio* of the *number* of *bigrams* found in the document and the *size* of this *document*:

$$R_3 = f\left(W_T^{pos}, W_T^{neu}, W_T^{neg}\right)/P_T = R_2 \bullet N_C \quad (9)$$

Introduction of this indicator is due to the possibility to *approximate* the *total evaluation* of the entire text on the basis of average tonality weight of one bigram (provided that the author expresses his opinion in a balanced and logical manner).

Step 5. Evaluation of the Sentiment Classification Quality

The metrics used for the reviews classification evaluation are: *precision*, *recall* and *accuracy,* separately for each Subjective Corpora Sample.

To give definition to these metrics, we will use Table 1.

Table 1: Classifier Output Types.

PC	Actual class		
	AC1	AC2	AC3
PC1	CC1: Correctly classified as AC1	QC2: Quite correctly classified as AC2	NC 3: Not correctly classified as AC3
	TB1: Truly belong to AC1	PB1: Possibly belong to AC1	FB1: Falsely belong to AC1
PC2	QC 1: Quite correctly classified as AC1	CC2: Correctly classified as AC2	QC 3: Quite correctly classified as AC3
	PB2: Possibly belong to AC2	TB2: Truly belong to AC2	FB2: Falsely belong to AC2
PC3	NC 1: Not correctly classified as AC1	NC 2: Not correctly classified as AC2	CC 3: Correctly classified as AC3
	FB3: Falsely belong to AC3	PB3: Possibly belong to AC3	TB1: Truly belong to AC3

In this table and further, *Actual classes* assume the rating according to subjective (according to the filmweb.pl rating) evaluations of texts tonality.

In this case, *precision* is the proportion of reviews classified as X that truly belong to class X.

$$percision = \frac{TB_X}{TB_X + PB_X + FB_X} \quad (10)$$

Recall is the proportion of all reviews of class X that is classified by the algorithm as X.

$$recall = \frac{CC_X}{CC_X + QC_X + NC_X} \quad (11)$$

Accuracy is proportion of correctly classified objects in all objects processed by the algorithm:

$$\text{accuracy} = \frac{\sum_{X=1}^{3} CC_X}{\sum_{X=1}^{3} CC_X + \sum_{X=1}^{3} QC_X + \sum_{X=1}^{3} NC_X} \quad (12)$$

4 ANALYSIS OF RESULTS OF SC-METHODOLOGY APPLICATION

To test and evaluate the adequacy of the author's Methodology realization, as a *case study* were used the following training samples:

for the first stage (CBSD Creation Algorithm) – 5000 Polish-language films reviews (2500 TSP and 2500 TSN); for the second stage (Sentiment Classification Algorithm) – 3000 Polish-language films reviews (1500 SPCS and 1500 SNCS) from filmweb.pl. We consider the SPCS films reviews if the subjective review's assessment is more than 5 points, and SNCS – if it is equal or less than 5 points.

The experimental part of all steps of authors' Algorithm is technically realized in Python 3.4.1.

4.1 CBSD Creation Algorithm

As a result of the first stage of the developed SC-methodology realization, Hierarchical CBSD was created (Figure 3).

The main specificities of the received CBSD:

– for Positive Class of CBSD: Almost equal numbers of bigrams of neutral (46.2%) and positive (43.7%) polarity. This suggests that half of the adjectives and verbs used to characterize the reviewer's opinion without having a positive coloring, formally confirm (ascertain) the existing facts.

– 10% of the negatively colored bigram, indicating that, despite the truly positive tonality of reviews, the reviewer doubts about the positivity of certain shades (elements) of the film.

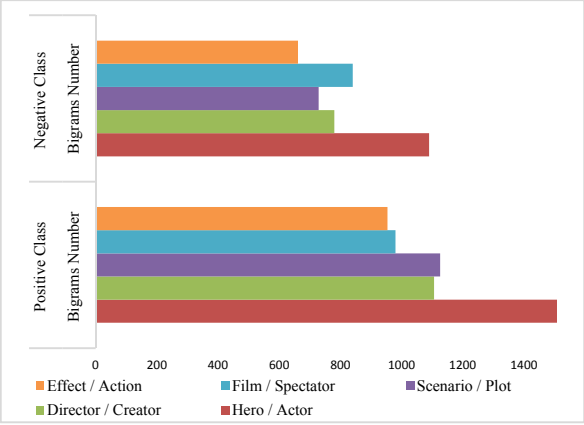

Figure 3: Structure of the 1st level of CBSD (case study results).

– for Negative Class of CBSD: more bigrams have negative (41.7%) polarity and only 37.5% have neutral polarity. Negative reviews are characterized, in turn, by a large number of oppositely painted bigrams – 20.7%. Perhaps, some of these positive emotions are introduced by the authors for comparison or contrast.

4.2 Sentiment Classification Algorithm

4.2.1 Simple Sentiment Classification

At the second stage of the developed SC-methodology firstly the algorithm of Sentiment Classification was realized using CBSD without considering their Topical structure (0^s level).

The main goal of this step was to teach this algorithm to guarantee a higher quality of sentiment recognition by:

– choosing the most accurate measures of the text tonality (R1-R4);

– finding the set of main specificities of Corpora Sample, which allows to formulate recommendations for qualitative conduction of the research process of acceptance / rejection of Hypothesis 1 (Table 2).

1. The highest quality indicators for the SPCS, as far as for SNCS, are observed when using the R_1 and R_2 *tonality measures*.

Explanation of this phenomenon may be the fact that a large amount of text *does not have an explicit tonal coloring* as such.

This is evidenced by the proportion of words, the tonality of which is recognized. Quite often, a review in half consists of retelling the content (script) of the film. It is practically impossible to evaluate the tonality of such a text.

2. Quality indicators for the SPCS and SNCS when using R_3 and R_4 *Rating Formats* are slightly lower. This trend is again explained by *the low density of distribution of words with a clearly recognized tonality inside the text*. And since the rating of the film allows to estimate the average weight of tonality in one bigram present in the document (R_3) and approximate the total score of the entire text based on the average weight of tonality of one bigram (R_4), the unevenness of textual coloring of the text distorts the quantitative assessment of tonality of the whole document.

For these films rating formats, the density of word distribution with an explicitly recognized tonality inside the text is on average 17.5% (10.65% for SNCS), but the spread of values is much smaller (ranging from 13.8% to 23.2% for SPCS / from 7.5% to 14.2% for SNCS). This fact indicates that the use

Table 2: Evaluation of the Quality of Sentiment Classification of the Films Reviews Results (Simple Classification, in %).

Rating formats	Training Subjectively POSITIVE Corpora Sample					Training Subjectively NEGATIVE Corpora Sample				
	Class	Class Proportion	Precision	Recall	Accuracy	Class	Class Proportion	Precision	Recall	Accuracy
R_1	*High* positive	28.57	53.57	51.72	47.96	*Rather* negative	33.00	33.33	29.73	43.00
	Quite positive	47.96	51.06	53.33		*Clearly* negative	56.00	53.57	57.69	
	Reasonably positive	23.47	34.78	33.33		*Absolutely* negative	11.00	18.18	18.18	
R_2	*High* positive	29.59	51.72	51.72	43.00	*Rather* negative	35.00	31.43	29.73	39.00
	Quite positive	44.90	47.73	46.67		*Clearly* negative	53.00	50.94	51.92	
	Reasonably positive	25.51	28.00	29.17		*Absolutely* negative	12.00	8.33	9.09	
R_3	*High* positive	30.61	43.33	44.83	39.00	*Rather* negative	38.00	31.58	32.43	38.00
	Quite positive	46.94	43.48	44.44		*Clearly* negative	51.00	49.02	48.08	
	Reasonably positive	22.45	27.27	25.00		*Absolutely* negative	11.00	9.09	9.09	
R_4	*High* positive	28.57	42.86	41.38	43.00	*Rather* negative	37.00	32.43	32.43	39.00
	Quite positive	46.94	52.17	53.33		*Clearly* negative	51.00	49.02	48.08	
	Reasonably positive	24.49	29.17	29.17		*Absolutely* negative	12.00	16.67	18.18	

of indicators R_3 and R_4 allows you to evaluate the rating of reviews in a smoothed standard scale.

In general, the results of comparing the quality of the recognition of reviews of films SPCS / SNCS allow us to draw the following conclusions:

1. A large part of reviews is characterized by an *average degree of density of the distribution* (AD) of words, with recognizable tonality.

2. Additionally, morphological analysis of Training Sample testifies that:

– *positive* reviews characterized by *highly semantic structured opinion*, expressed in a carefully and balanced manner;

– *negative* reviews characterized by *average level of semantic structure of the opinion*, expressed more spontaneously and under the influence of emotions. As a consequence, there is greater probability of their precise recognition and classification (Rizun and Taranenko, 2018).

In this regard, main *recommendations* for qualitative conduction of the next steps of the research process were formulated:

1. One of the analyzed tonality measures (R1) is *less influenced* by the factor of AD of words with recognizable tonality.

2. Division of the document into TCTC and usage of the topically oriented CBSD for reviews classification could give the chance to *increase the* AD of the explicit words within the text (and thereby *increase the quality* of text classification).

4.2.2 One-level Sentiment Classification

Realization of the algorithm of Sentiment Classification using the 1st level of CBSD, allowed:

1. To recognize the Topics of texts paragraphs (the Contextual Framework of Classified Reviews) (Table 3).

Table 3: The Contextual Framework of Films Reviews Corpora (% to the total number of paragraphs).

Class	Hero / Actor	Director / Creator	Scenario / Plot	Film / Spectator	Effect / Action	Unrecognized paragraphs
HP	19.28	57.45	46.38	17.39	45.45	9.29
QP	37.35	34.04	37.68	26.09	31.82	
RP	43.37	8.51	15.94	56.52	22.73	
RN	57.14	–	44.12	37.84	–	16.09
CN	28.57	–	47.06	45.95	–	
AN	14.29	–	8.82	16.22	–	

2. To compare the Quality of Simple and one-level Sentiment Classification of the Films Reviews Results (Figure 4).

General conclusions on the stage of classification can be the following:

1. Positive dynamics of the difference of average values of the main indicators of quality of recognition and classification of texts when using the One-level structure of the Semantic Dictionary is significant and varies from 13.53% till 72.97%. This fact is a

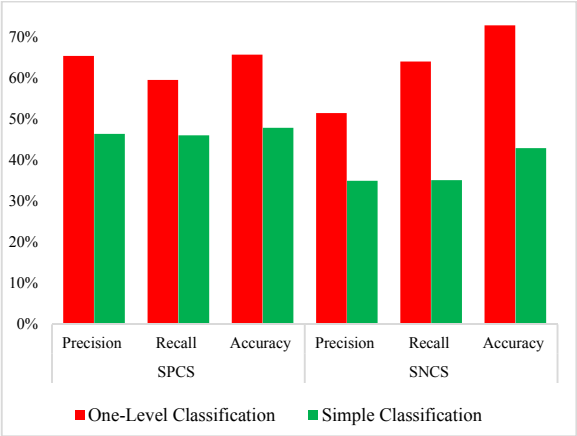

Figure 4. Difference between the Average values of Quality evaluation of the One-level and Simple Sentiment Classification.

confirmation of the *significance* of Hypothesis 1.

2. Reviews with Subjectively Negative tonality have pure Topical variety as well as lower (in 6.8%) percentage of paragraphs Topic recognition in comparison with Subjectively Positive sample. This fact can be considered a reason to *reject* Hypothesis 2 – that the quality of classification does not depend on the tonality, subjectively assigned by the author.

5 CONCLUSION

In this paper, authors present the Methodology for Web Content Sentiment Classification. Main contribution of the authors' study is finding the answers to the main scientific research questions:

– the style of the analyzed text determines the possibility of flexible adaptation of the algorithms for texts classification. For example, in the case study in this paper, the style/type of the analyzed texts (Review / Persuasive) allowed each document to be considered as a collection of TCTC (paragraphs), which positively affects the classification quality;

– the Hierarchically-oriented Structure of the Sentimental Dictionary allows customizing the classification process to more accurate recognition of the text tonality in the context of topic;

– texts were written in the Persuasive style, most often initially empowered by authors with a certain tonality. Such tone has a significant, but not critical, effect on the qualitative indicators of sentiment recognition. Negative emotions of the author usually, on the one hand, reduce the level of variability of the words used and the variety of topics raised in the document; on the other – increase the level of unpredictability of contextual use of words with both positive and negative emotional coloring.

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

Bijuraj, L. V., 2013. Clustering and its Applications. *Proceedings of National Conference on New Horizons in IT* (NCNHIT). pp. 169–172.

Boiy, E., 2007. Automatic Sentiment Analysis in On-line Text. *Proceedings of the 11th International Conference on Electronic Publishing* (ELPUB 2007). pp. 349–360.

Boucher J.D., Osgood Ch.E., 1969. The Pollyanna hypothesis. *Journ. of Verbal Learning and Verbal Behaviour*, no. 8, pp. 1–8.

Hu, M., 2004. Mining and Summarizing Customer Reviews. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '04), pp. 168–177.

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, pp. 22–33.

Klekovkina MV, Kotelnikov EV., 2012. 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): tr. XIV Vseros. sci. Conf. Pereslavl-Zalessky: pp. 118-123.

König A.C., Brill E., 2006. Reducing the human overhead in text categorization. *Proc. 12th ACM SIGKDD conf. on knowledge discovery and data mining*, pp. 598–603.

Liu, B., 2012. Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*. Vol. 5(1).

Manning, Ch., 2009. Introduction to Information Retrieval. *Cambridge University Press*, p. 544, p. 222.

Pang, B., 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval.* Vol. 2. pp. 18-22.

Popovic, M., 2006. Statistical Machine Translation with a Small Amount of Training Data. In *Proceedings of the 5th LREC SALTMIL Workshop on Minority Languages*. pp. 25–29.

Rizun N., Ossowska K., Taranenko Y., 2018. 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* – ISAT 2017, pp.364-373.

Rizun N., Taranenko Y., Waloszek W., 2017a. The Algorithm of Modelling and Analysis of Latent Semantic Relations: Linear Algebra vs. Probabilistic Topic Models. *Knowledge Engineering and Semantic Web. 8th International Conference*, KESW 2017, pp.53-68.

Rizun N., Taranenko Y. Methodology of Constructing and Analyzing the Hierarchical Contextually-Oriented Corpora. *Proceeding of Federated Conference on Computer Science and Information Systems* – FedCSIS 2018.

Rizun N., Taranenko Y., Waloszek W., 2017b. 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, pp.366-372..

Rizun, N., Taranenko, Y., 2017. Development of the Algorithm of Polish Language Film Reviews Preprocessing. *Research Yearbook Faculty of Management in Ciechanów* WSM, 1-4 (IX), pp. 168-188.

Salton, G., 1988. Term-weighting Approaches in Automatic Text Retrieval. *Information Processing and Management*. № 5, Vol. 24. pp. 513–523.

Salton, G. 1989. Automatic Text Processing: the Transformation, Analysis, and Retrieval of Information by Compute. *Addison-Wesley Longman Publishing*, 543 p.

Taboada M., Brooke J., Tofiloski M., Voll K., Stede M., 2011. Lexicon-based methods for sentiment analysis, *Computational Linguistics*. no. 37 (2), pp. 267–307,

Titov, I., 2008. Modeling Online Reviews with Multi-grain Topic Models. *Proceedings of the 17th International Conference on World Wide Web* (WWW'08), pp. 111–120.

Ur-Rahman, N., 2012. Textual data mining for industrial knowledge management and text classification: A business oriented approach. *Expert Systems with Applications*. № 39. pp. 4729–4739.