USING DISTRIBUTED REPRESENTATIONS FOR ASPECT-BASED SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com),
Kotelnikov E. V. (kotelnikov.ev@gmail.com)

Vyatka State Humanities University, Kirov, Russia

The article is focused on aspect-based sentiment analysis, which is a specific version of the general sentiment analysis task. Its goal is to detect the opinions expressed in the text on the level of significant aspects of the specified entity. An overview of the existing approaches and previous work is presented.

The main result of our work is a new method of aspect-based sentiment analysis based on the distributed representations of words. Such representations are obtained by using deep learning algorithms. The method includes the well-known algorithm of training distributed representations of words, two new techniques for constructing the aspect and sentiment lexicons, and an algorithm for calculating aspect scores.

Examples of aspect and sentiment terms are given. The vectors of resulting terms are visualized using the t-SNE method. The article presents the results of experiments on a test corpus for three aspects—“food”, “interior” and “service”, which yield aF1-measure increase of 11 to 16% as compared to the baseline.

Key words: aspect-based sentiment analysis, machine learning, deep learning, distributed representations

1. Introduction

The area of sentiment analysis is actively developing recently. The sentiment analysis is the problem of finding user opinions and sentiments in a text [10]. The problem is evidently still far from its final solution therefore it is interesting in the academia. The new methods of computational linguistics and machine learning are being developed to solve the problem of sentiment analysis. The business community is interested in the commercial applications of such analysis, for example, the sentiment analysis may be useful in the study of opinions and preferences of the target audience of consumers.

The aspect-based sentiment analysis is a relatively new task in this area [10, p. 58]. Its appearance can be explained by the fact that the sentiment analysis on the level of a whole text or even on the sentence level is not able to detect the expressed opinion on the certain aspects of the studied entity. Such formulation of the problem saves the common sense of the sentiment analysis and at the same time it is more detailed and researches the opinions expressed in the text on the level of the meaningful aspects of an entity. For example, the sentence "In general, the food is great, but the service is terrible!" presents different opinions on aspects "food" and "service" of a single object "restaurant". Along with the sentiment terms detection task such version of sentiment analysis is extended with the aspect extraction task [10, p. 67].

An aspect term can be defined as a word or a collocation that explicitly determines an attribute of a target entity. A sentiment term is a word or a collocation that expresses the user's subjective opinion. Both types of terms, aspect and sentiment, vary from one domain to another, therefore the development of the effective methods of automatic selection of aspect and sentiment terms with minimal time costs and human labor is very important.

A continuous vector space of distributed representations [17] of words as the source of lexicons constructing is investigated in the article. The new techniques for the aspect and sentiment lexicons constructing from small initial sets of words are proposed.

The remainder of the article is as follows. The overview of the previous approaches and papers is given in section 2. Section 3 describes the used corpus of documents. The techniques for aspect and sentiment terms detection are presented in section 4. The results of the experiments and the conclusions are given in sections 5 and 6 respectively.

2. Related Works

The main two subtasks which must be solved to perform the aspect-based sentiment analysis are the aspect extraction and the sentiment terms detection.

The aspect extraction task can be solved within three main approaches [10, pp. 67–78]:

1. the frequent-based approach;
2. the supervised machine learning approach;
3. the unsupervised machine learning approach.

The core idea of the first approach is to select the most frequent nouns and collocations as the aspect terms [5, 15]. Despite its relative simplicity the approach can show not a bad quality of aspect extraction, however it has some shortcomings: it gives too many false aspect terms and tends to miss infrequent terms. Besides that, the clustering by aspect categories has to be done for the obtained terms.

The aspect extraction task can be expressed in the terms of information extraction task which, as it is known, can be solved by the supervised machine learning methods [6, 7]. The main shortcoming of such approach is a high complexity of obtaining the labeled train data. The result of this shortcoming is the problem of resetting of the methods for the new domains.

The method proposed in the article belongs to the third approach—unsupervised machine learning, which overcomes the mentioned disadvantages of the two previous approaches. The main methods of this approach are the methods of topic modeling for example Latent Dirichlet Allocation (LDA) [1].

As the base LDA model can find only global topics of a document's collection, various modifications of this model which can find the distinct aspects were proposed [9, 19]. The results of the work of such models are the probability distributions on the words, which correspond to the aspects, that is the separation between aspect and sentiment terms is not performed. In our work such separation is performed explicitly what gives a user more interpretability over the result of the analysis.

The LDA method was also used in [2]: the aspect terms were found first and then the sentiment terms which can only be the adjectives were detected. Our method preserves the sequence of the actions, but as the sentiment terms beside adjectives it also takes into account the complex phrases, which are good indicators of sentiment and make analysis more precise.

The paper [12] investigated the method of aspect term generating based on the semi-supervised modeling. For each aspect the initial set of words was specified and then it was replenished with the new terms by using the LDA model. In our work the source of the new terms is the space of continuous vector representations of words, which is obtained by using the deep learning. This approach is more flexible: comparing with the LDA it gives the intermediate representation for each term. The vector space brings the notion of similarities between words, which is useful for solving natural language processing tasks [4, 18, 21].

Another important subtask—the sentiment terms detection—is often solved with the help of sentiment lexicons. Such lexicons list emotionally-colored lexical units and their weights. The main obstacle in using lexicons is the complexity of their creation. They are constructed either manually by the experts or automatically from the initial set of words with their sentiment weights [20, 23]. In [23] the authors used only one initial word ("*good*") and 6 negations for the lexicon creation. In [20] the initial sets consisted only of two words ("*excellent*" and "*poor*"), the sentiments of another phrases were calculated on the base of mutual information measure. Similarly our method of lexicon creation uses two initial words "*отличный*" (*great*) and "*ужасный*" (*terrible*) however the cosine similarity is used to detect the sentiment terms and to calculate their weights. In [20] the author used the search queries as the source of statistical information about terms co-occurrence. In our work such source is the large corpus of documents.

3. The Text Corpus

Unfortunately, there is no available text corpus in Russian because the task of aspect-based sentiment analysis is relatively new so the new corpus was created, it includes the user reviews of restaurants. 33,243 reviews were collected from *restoclub.ru*. For each review the user specifies the numeric score for the following aspects: *food*, *interior* and *service*. Initially the scores were presented in ten-point scale, we cast the scores to the binary scale by the following mapping scheme: {*1..5*}→*negative*, {*6..10*}→*positive*. 15,285 reviews[1] were selected as a test data set, for each review in this set at least one

[1] Test corpus and dictionaries are available at: http://goo.gl/NhEvWu.

of the aspect scores is less than 8. Such selection is made to reduce the imbalance of the collection to the positive scores. The distribution of the test data set in aspects and scores is shown in Table 1 (a single review can have positive score for one aspect and negative score for another).

Table 1. Some statistics of the test data set

Aspect	Positive reviews	Negative reviews	Total
Food	10,063	5,222	15,285
Interior	11,296	3,989	15,285
Service	8,707	6,578	15,285

To build the high quality distributed representations of words we need only the texts of reviews. The quality of the received vectors depends on the quantity of texts, so 14,058 reviews without any aspect scores were additionally collected from *restoran.ru*. Thus, the corpus of 47,301 reviews in total was used to build the distributed representations of words. Note that the aspect scores were not used in this process.

The text of each review was preprocessed with the segmentation, the tokenization and the morphological analysis. The procedures were performed using such tools as Mystem [13] and FreeLing [14].

4. The Aspect-Based Sentiment Analysis

In our work the aspect-based sentiment analysis includes four stages. On the first stage the vector space of distributed representations is built. On the second and the third stages the aspect and sentiment terms are determined respectively. The scores for each aspect are calculated on the final stage.

4.1. The Vector Space

The unsupervised deep learning algorithms were used to build the vector space. The common idea of such algorithms is to automatically find the "good" set of features to represent in high quality the target object (image, audio signal, text, etc.).

In case of textual information each lexical unit (word) is represented by the vector of real numbers called *distributed representation* [17]. The peculiarity of such representations is that they encode a set of degrees of linguistic similarity between words. In other words, semantically and syntactically related words appear together in the vector space.

To build such space the skip-gram model was used [11]. Formally, the model tries to maximize the following function for the given train sequence of words $w_1, \ldots, w_T$ [11]:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c,\, j\neq 0}\log p(w_{t+j} \mid w_t) \rightarrow \max, \qquad (1)$$

where c—the size of the training context (window size), T—the length of the train sequence of words.

The probability $p(w_{t+j} \mid w_t)$ is defined as [11]:

$$p(w_O \mid w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W} \exp({v'_w}^T v_w)}, \tag{2}$$

where v_w and v'_w are the input and output vector representations of w; w_I and w_O are the current and predicted words, W—the number of words in vocabulary.

For the experiments we used the Gensim [16] implementation of the skip-gram model. All texts of the corpus (47,301 reviews) presented as a single sequence of sentences were used to build the vector representations of words. On the base of this corpus we construct the lexicon with the words which frequency is more than 5. Next, the dimensionality of the space is chosen (in our case 150). The greater number of dimensions allows to capture more language regularities but leads to more computational complexity of the learning. Each word from the lexicon is associated with the real numbers vector of the selected dimensionality. Originally all the vectors are initialized with random numbers close to zero. During the learning procedure the algorithm "slides" with the fixed size window (in our case 5) along the words of the sequence and calculates the probability (2) of context words appearance within the window on the base of its central word under review (or more precisely, its vector representation). The ultimate goal of the described process is to get such vectors for each word, which allow to predict its probable context. This goal is achieved by maximizing the function (1).

4.2. The Aspect Lexicon Construction

The idea of our method is to extend automatically the initially specified sets of terms. Five initial terms were selected for each aspect (Table 2). The assumption was made that the aspect terms can only be the single words.

For each term in the vector space of distributed representations we can find its nearest neighbors. The cosine similarity was used as a measure of similarity between vectors. Formally, the similarity between two vectors $\vec{a} = (a_1,...,a_n)$ and $\vec{b} = (b_1,...,b_n)$ is given by:

$$similarity(\vec{a},\vec{b}) = \cos(\theta) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \cdot \sqrt{\sum_{i=1}^{n} b_i^2}}, \tag{3}$$

where Θ—the angle between the vectors, n—the dimensionality of the space.

Thus a list of several closest terms can be formed for each term. In our experiments we took 10 of such nearest terms. By joining all the lists and excluding the repeated terms a set of new terms emerges. We call such new set of terms a *generation*. The initial set of words can be considered as the zero generation. The repeating of the same procedure for the new generations is the iterative process which generates the aspect terms. The noise words can appear in a generation, especially when the number of iteration is getting large. To keep the thematic coherence of terms under control the additional constraint was injected: at least three of five terms from the zero generation must be close (*similarity* > 0) to the new term. After some of such iterations the set of possible terms runs out and the whole process terminates. Finally, the aspect terms vocabulary is formed by joining all the generations. In our experiments after 10 iterations the lexicon of 3,080 terms was formed. It included 1,749 terms for aspect "*food*", 996 terms for "*interior*" and 335 for "*service*". Table 2 shows the terms of the zero and the first generations. Note that the terms are given in its original spelling. The capability to find such low-frequency terms appears due to the specifics of the vector space of distributed representations.

Table 2. The aspect terms of the zero and the first generations

Food	Interior	Service
Generation 0 (initial sets)		
закуска, суп, десерт, салат, плов	интерьер, атмосфера, музыка, дизайн, бар	обслуживание, персонал, официант, менеджер, сервис
Generation 1		
оливье, солянка, шашлык, похлебка, штрудель, салатик, сметанник, манта, фрикаделька, блюдо, соление, закусочка, явство, ассорти, хачипури, люля, щи, морепродукт, хинкали, хачапури, чизкейк, баранина, нарезка, цезарь, тортик, мороженое, медовик, эклер, уха, супчик, кебаб, сациви, …	саксофон, убранство, стилистика, комфортность, музыкант, беззаботность, гитара, интерьр, уют, обстановка, lounge, продуманность, вокал, репертуар, отмосфера, клуб, интереьер, dj, комфорт, джаз, исполнитель, диджей, времяпровождение, звук, кабак, оформление, …	заместитель, обслужевание, бармен, управлять, девушка, коллектив, сотрудник, офицант, официатка, администратор, тамада, внимательность, официантка, директор, обсуживание, команда, девочка, обслуга, отзывчивость, еда, офицантка, дирекция, …

Using the t-SNE (t-Distributed Stochastic Neighbor Embedding) [22] algorithm the results of the process can be visualized on the plain. Figure 1 shows the aspect terms vectors for the first four generations. One can trace the cluster structure according to the aspects.

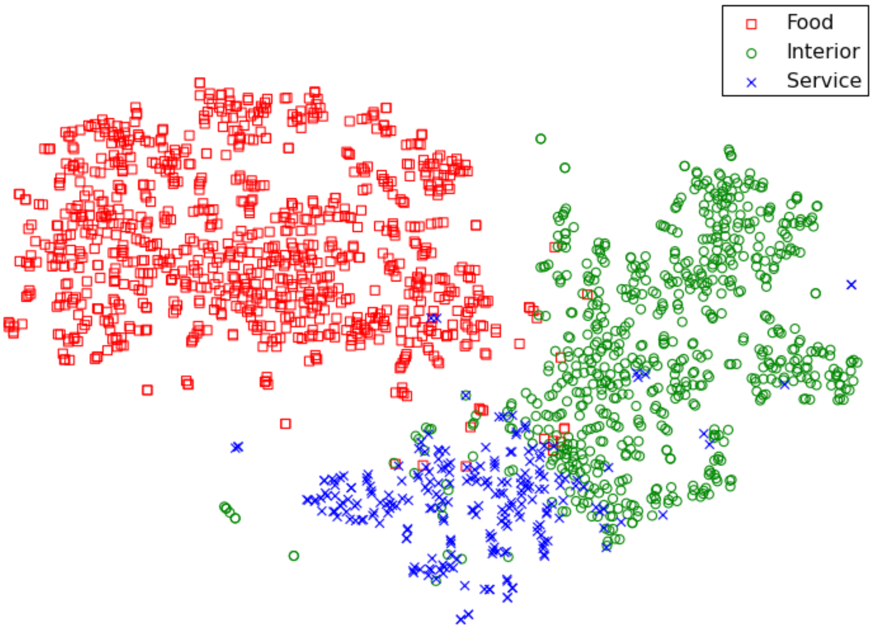

Fig. 1. The aspect terms for the first four generations

4.3. The Sentiment Lexicon Construction

The method of sentiment lexicon construction includes two steps: candidates' selection and its weighting.

As the sentiment phrases can consist of more than one word the additional preprocessing is required. It is known that the modifiers and the negations have the significant meaning in sentiment phrases. The possible set of such words for Russian is proposed in [8]. By our estimates the adverb "*очень*" (*very*) bears the most amplifications and the particle "*не*" (*not*) covers the most negations. By simple pattern

$$<very \mid not> + <very \mid not> + <adjective \mid verb \mid adverb>$$

complex lexical units were formed, for example, *не_готовый, очень_сытный, очень_не_приятный, не_очень_опрятный,* etc. Of course such a way doesn't take into account the whole variety of sentiment phrases, but definitely covers the essential part of it.

We took the single adjectives and the set of complex lexical units as the candidates to the sentiment lexicon. It was the list of $N = 7312$ such candidates, about 34% of them were complex lexical units.

Besides the thematic similarity in the vector space of distributed representations the emotional similarity between the terms can also be traced. So the space can be used not only for sentiment terms detection, but also as a source of sentiment terms weighting.

For the initial setting of sentiment values the etalon terms were determined: *great* was used for the positive sentiment and *terrible* was used for the negative sentiment. For each candidate from the list two values of similarity (3) with the etalon terms were considered as its weights.

Some examples of the most positive and negative terms obtained in such a way are shown in Table 3 (in their original spelling).

Table 3. The examples of the sentiment terms

Positive	Negative
хорошоя, замечательный, великолепный, превосходный, очень_гостеприимный, прекрасный, велеколепный, дружественный, очень_веселый, очень_хороший, шикарный, доброжилательный, очень_душевный, чудесный, суперский, хороший, приятный, не_пошлый, професиональный, очень_теплый, классный, очень_доброжелательный, супер, очень_дружелюбный, тактичный, безупречный, …	отвратительный, безобразный, очень_плохой, отвратный, ужастный, плохой, хамский, невнимательный, не_заслуживать, очень_обидно, отстойный, не_довольный, ужасный, не_вкусный, нулевой, откровенный, не_очень_позитивный, бездушный, не_ровный, не_способный, безответственный, недопустимо, очень_не_понравиться, дурной, очень_разочаровывать, пренеприятный, ужасно, гадкий, не_выдаваться, …

Similarly to the aspect terms the vectors of sentiment phrases can be plot with the t-SNE method. Figure 2 shows the subset of the most positive and the most negative phrases.

4.4. The Aspect Score Calculation

On the final stage it is necessary to get sentiment scores for each aspect. Every sentence is segmented by the following set of punctuation marks: {*? ! , . : ;*}. For each segment the aspect and sentiment terms from the lexicons are found. For every aspect term the summarized similarity (3) with the zero generation terms is calculated and the maximum value $similarity_{\max}$ is chosen. Then the summarized score sum_{em} of the sentiment terms from the current, the couple of the previous and the couple of the next segments is calculated. The final score of a sentence S is found for each aspect a as follows:

$$S_a = \sum_{a \in A} similarity_{\max}^{a} \cdot sum_{em}, \qquad (4)$$

where A—the set of the aspect terms for the sentence.

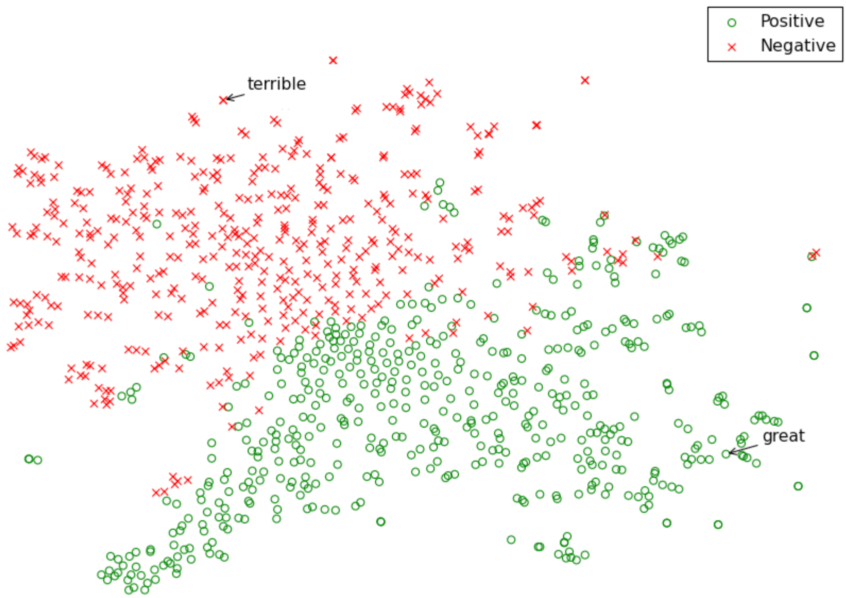

Fig. 2. Some sentiment phrases

The review's emotional score for each aspect a is the sum of scores S_a for every sentence. A sign of this score defines the aspect sentiment—positive or negative.

Table 4 lists some sentences and shows the sentiment calculation for them. The aspect terms are in bold, the sentiment terms are in italics. The numeric values represent either the summarized similarity $similarity_{max}$ with the zero generation aspect terms or the sentiment score of the term.

Table 4. The examples of the aspect sentiment detection

Предложение	**Оценка**
+1.000 1.482 *Отличный* имбирный **лимонад** с достаточным количеством льда и палочками тимьяна.	$1.482 \cdot 1.000 =$ $= +1.482 \Rightarrow pos$
−0.195 1.199 Его мариновали с *какими-то* **травами**, −0,274 которые абсолютно *не понравились*.	$1.199 \cdot (-0.195 -$ $- 0.274) = -0.562 \Rightarrow neg$
+0.790 2.046 +0.435 +0.591 А теперь о *прекрасном* =) **Бургер** был *просто чудесным*!	$2.046 \cdot (0.790 + 0.435 +$ $+ 0.591) = +3.716 \Rightarrow pos$
+0.229 1.737 +0.142 В *общем*, **оформление** *симпатичное*, но я люблю другое.	$1.737 \cdot (0.229 + 0.142) =$ $= +0.644 \Rightarrow pos$
−0.431 2.064 Пришёл совершенно *неквалифицированный* **сотрудник**!	$2.064 \cdot (-0.431) =$ $= -0.890 \Rightarrow neg$

5. Experimental Results

The proposed method was evaluated according to the precision, the recall and the F1-measure [3]. Figure 3 shows the dependence of the values of the F1-measure for each aspect and the number of the aspect terms (the logarithmic scale). The start values in the baseline correspond to the case when all the scores for the aspect are randomly assigned.

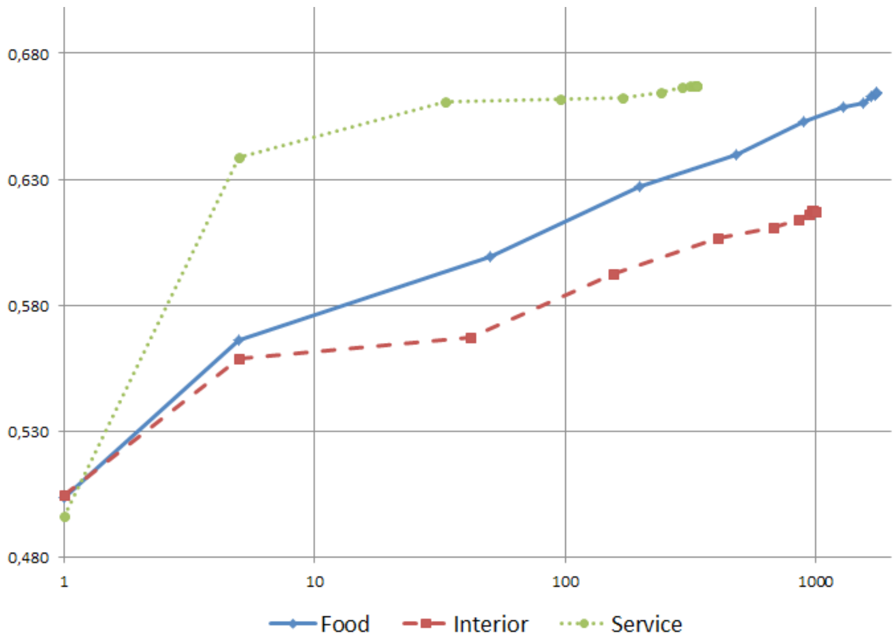

Fig. 3. The F1-measure for each generation

The baseline metrics and the best results of our method (in bold) are shown in Table 5.

Table 5. The evaluation results for the aspects

Aspect	Number of terms	Precision		Recall		F1-measure	
Food	1,749	0.503	**0.686**	0.504	**0.644**	0.503	**0.664**
Interior	996	0.504	**0.629**	0.505	**0.606**	0.504	**0.617**
Service	335	0.497	**0.692**	0.496	**0.644**	0.496	**0.667**

The greatest number of terms is in the "*food*" aspect, because there is a large variety of dish names. When the vocabulary of such terms grows, the value of the F1-measure only increases. In contrast, the aspect "*service*" contains not so many terms and quite a small vocabulary is already sufficient to get almost maximum values of metrics that were achieved. Low metrics for the aspect “*interior*” can probably be explained by the significant imbalance of the test collection to the positive scores.

6. Conclusion

The article studies the aspect-based sentiment analysis task. The new method of the aspect-based sentiment analysis based on the continuous vector space of the distributed representations is proposed. The suggested method allows to conduct the sentiment analysis with the use of minimal additional information and with minimal dependency from a domain.

The corpus of users’ reviews of restaurants is prepared for the experiments. The method is evaluated on this corpus for tree aspects. The result values of the F1-measure significantly outperform the chosen baseline: 66% versus 50% for the aspects “*food*” and “*interior*”, 62% versus 50% for the aspect “*service*”.

Sentiment lexicon decomposition by aspects seems to be a promising direction to boost the results of the aspect-based sentiment analysis.

References

1. *Blei M., Ng A., Jordan M.* (2003), Latent Dirichlet allocation, Journal of Machine Learning Research, Vol. 3, pp. 993–1022.
2. *Brody S., Elhadad N.* (2010), An Unsupervised Aspect-Sentiment Model for Online Reviews, Proceedings of The 2010 Annual Conference of the North American Chapter of the ACL.
3. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2012), Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue”, No. 11(18), pp. 739–746.
4. *Glorot X., Bordes A., Bengio Y.* (2011), Domain adaptation for large-scale sentiment classification: A deep learning approach, ICML, pp. 513–520.
5. *Hu M., Liu B.* (2004), Mining and summarizing customer reviews, International Conference on Knowledge Discovery and Data Mining (ICDM).
6. *Jakob N., Gurevych I.* (2010), Extracting Opinion Targets in a Single-and Cross-Domain Setting with Conditional Random Fields, Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-2010).
7. *Jin W., Ho H.* (2009), A novel lexicalized HMM-based learning framework for web opinion mining, Proceedings of International Conference on Machine Learning (ICML-2009).
8. *Klekovkina M. V., Kotelnikov E. V.* (2012), The automatic sentiment text classification method based on emotional vocabulary [Metod avtomaticheskoj klassifikatsii tekstov po tonalnosti osnovannyj na slovare èmotsionalnoj leksiki], Digital libraries: advanced methods and technologies, digital collections (RCDL-2012) [Èlektronnye biblioteki: perspektivnye metody i tehnologii, èlektronnye kollektsii], Pereslavl-Zalessky, pp. 118–123.

9. *Li F., Huang M., Zhu X.* (2010), Sentiment analysis with global topics and local dependency, Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010).
10. *Liu B.* (2012), Sentiment analysis and opinion mining, Morgan & Claypool Publishers.
11. *Mikolov T., Sutskever I., Chen K., Corrado G., Dean J.* (2013), Distributed Representations of Words and Phrases and their Compositionality, Proceedings of NIPS.
12. *Mukherjee A., Liu B.* (2012), Aspect Extraction through Semi-Supervised Modeling, Proceedings of 50th Annual Meeting of Association for Computational Linguistics (ACL-2012).
13. Mystem, available at: http://company.yandex.ru/technology/mystem.
14. *Padró L., Stanilovsky E.* (2012), FreeLing 3.0: Towards Wider Multilinguality, Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.
15. *Popescu A., Etzioni O.* (2005), Extracting product features and opinions from reviews, Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).
16. *Řehůřek R., Sojka P.* (2010), Software Framework for Topic Modelling with Large Corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pp. 45–50.
17. *Rumelhart D., Hintont G., Williams R.* (1986), Learning representations by back-propagating errors, Nature, pp. 533–536.
18. *Socher R., Lin C., Ng A., Manning C.* (2011), Parsing natural scenes and natural language with recursive neural networks, Proceedings of the 26th International Conference on Machine Learning (ICML).
19. *Titov I., McDonald R.* (2008), A joint model of text and aspect ratings for sentiment summarization, Proceedings of Annual Meeting of the Association for Computational Linguistics (ACL-2008).
20. *Turney P.* (2002), Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews, Proceedings of the Association for Computational Linguistics (ACL), pp. 417–424.
21. *Turney P.* (2013), Distributional semantics beyond words: Supervised learning of analogy and paraphrase, Transactions of the Association for Computational Linguistics (TACL), pp. 353–366.
22. *Van der Maaten L., Hinton G.* (2008), Visualizing High-Dimensional Data Using t-SNE, Journal of Machine Learning Research, pp. 2579–2605.
23. *Zagibalov T., Carroll J.* (2008), Automatic seed word selection for unsupervised sentiment classification of chinese text, Proceedings of the 22nd International Conference on Computational Linguistics, Morristown, pp. 1073–1080.