Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2016"

Moscow, June 1–4, 2016

MANUALLY CREATED SENTIMENT LEXICONS: RESEARCH AND DEVELOPMENT

Kotelnikov E. V. (kotelnikov.ev@gmail.com),
Bushmeleva N. A. (bushmeleva_na@list.ru),
Razova E. V. (razova.ev@gmail.com),
Peskisheva T. A. (peskisheva.t@mail.ru),
Pletneva M. V. (pletneva.mv.kirov@gmail.com)

Vyatka State University, Kirov, Russia

The sentiment lexicons are an important part of many sentiment analysis systems. There are many automatic ways to build such lexicons, but often they are too large and contain errors.
The paper presents the algorithm of sentiment lexicons creation for a given domain based on hybrid—manual and corpus-based—approach. This algorithm is used for the development of the sentiment lexicons by means of four human annotators each for five domains—user reviews of restaurants, cars, movies, books and digital cameras. Created sentiment lexicons are analyzed for inter-annotator agreement, parts of speech distribution and correlation with automatic lexicons.
The performance of the sentiment analysis based on the created sentiment lexicons is researched and compared with the performance of the existing sentiment lexicons. The experiments with text corpora on various domains based on SVM show high quality and compactness of the human-built lexicons.

Key words: sentiment analysis, sentiment lexicons, manual approach, corpus-based approach, SVM

СЛОВАРИ ОЦЕНОЧНОЙ ЛЕКСИКИ, СОЗДАННЫЕ ВРУЧНУЮ: РАЗРАБОТКА И ИССЛЕДОВАНИЕ

Котельников Е. В. (kotelnikov.ev@gmail.com),
Бушмелева Н. А. (bushmeleva_na@list.ru),
Разова Е. В. (razova.ev@gmail.com),
Пескишева Т. А. (peskisheva.t@mail.ru),
Плетнева М. В. (pletneva.mv.kirov@gmail.com)

Вятский государственный университет, Киров, Россия

Ключевые слова: анализ тональности, словари оценочной лексики, экспертный подход, подход на основе корпусов, метод опорных векторов

1. Introduction

In recent years the sentiment analysis is one of the hottest research areas in natural language processing (Liu, 2012). 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, e. g., the analysis of consumer products and services reviews, the monitoring of social networks, the political studies (Feldman, 2013).

There are two main approaches to the sentiment analysis (Taboada et al., 2011): 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 creating. They generally use three main approaches (Liu, 2012): 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 (Fellbaum, 1998). 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 lexicons creation is very important, little attention is paid to the evaluation of the quality and in-depth analysis of the generated lexicons, especially for Russian.

In this paper, firstly, we propose a procedure of creating the sentiment lexicon for a given domain, secondly, we analyze the sentiment lexicon that is constructed by several annotators for various domains, thirdly, we research the performance of these sentiment lexicons in comparison with existing lexicons.

The rest of the paper considers the related work (Section 2) and the used text corpora (Section 3) are considered. At first the corpus-based approach to sentiment words extraction is applied to generate the sentiment lexicons, then their manual annotation is carried out by several annotators (Section 4). The generated lexicons are jointly analyzed (Section 5). Performance of the sentiment analysis based on the sentiment lexicons and Support Vector Machine (SVM) is evaluated (Section 6).

2. Related work

2.1. The creation of sentiment lexicons

Two stages of lexicons creation can be distinguished: 1) the generation of the sentiment-bearing words list, containing the candidates to sentiment lexicon, and 2) the assignment of sentiment labels to these words, e. g. positive/negative/neutral. Both stages are performed either manually or automatically.

Most of the studies on concerning sentiment lexicons creation are carried out on the material of English. For example, Taboada et al. (2011) both stages fulfilled manually. Mohammad and Turney (2013) used the crowdsourcing for the creation of word-emotion and word-polarity association lexicon.

There are also studies for other languages. For example, Amiri et al. (2015) formed word list manually, then this list was annotated by several human annotators by means of web interface.

There are few such studies for Russian. Chetviorkin and Loukachevitch (2012) extracted and weighted sentiment words automatically on the base of machine learning. Manual annotation was performed only for evaluation. Ulanov and Sapozhnikov (2013) built up the lexicons by means of automatic translation of English dictionaries. Blinov and Kotelnikov (2014) created the sentiment lexicon based on the distributed representations of words and used it for aspect-based sentiment analysis. Ivanov et al. (2015) applied the corpus-based approach in the user review domain as well as for aspect-based sentiment analysis.

At present the following sentiment lexicons are publicly available:

- Russian Sentiment Lexicon for Product Meta-Domain (ProductSentiRus)—5,000 words (Chetviorkin, Loukachevitch, 2012)[1];
- NRC Emotion Lexicon translated in Russian via Google Translate (NRC)—4,590 words (Mohammad, Turney, 2013)[2];
- Russian sentiment lexicon—2,914 words (Chen, Skiena, 2014)[3];
- Sentiment lexicon for restaurants domain—7,312 words (including bigrams and trigrams) (Blinov, Kotelnikov, 2014)[4].

---

[1] http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

[2] http://www.saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

[3] https://sites.google.com/site/datascienceslab/projects/multilingualsentiment

[4] http://goo.gl/NhEvWu

These lexicons (except the latter, containing the large part of collocations) are used in our study to compare with the manual lexicons (see Section 6).

2.2. Analysis of lexicons

One of the main purposes of our study is a joint analysis of the word list sentiment labeling. The word list was made by several human annotators for various domains. To our knowledge such in-depth analysis of Russian sentiment lexicons hasn't been performed yet.

Andreevskaia and Bergler (2006) conducted simultaneous labeling of two sentiment lexicons by two teams, which resulted in the high degree of disagreement.

Taboada et al. (2011) compared manual lexicons with dictionaries built using Amazon Mechanical Turk. In addition, a comparison with SentiWordNet was drawn, but only at the level of performance test.

As well several sentiment lexicons are compared by the quality of sentiment analysis in English (Musto et al., 2014; Ozdemir, Bergler, 2015) and in Portuguese (Freitas, Vieira, 2013).

Within the context of our study we should mention the work (Kiselev et al., 2015) in which the thorough analysis of 12 existing lexical-semantic resources (printed explanatory dictionaries, dictionaries of synonyms, electronic thesauri) is performed.

3. Text corpora

In our work the reviews of restaurants, cars, movies, books and digital cameras are researched. The reviews of restaurants were collected from the site Restoclub[5], the reviews of cars—from the site Cars@mail.ru[6]. For the rest domains the text corpora of seminar ROMIP2011 and 2012 are used (Chetviorkin et al., 2012; Chetviorkin, Loukachevitch, 2013).

The initial score scales (movies, books, restaurants—ten-point, cameras, cars—fivepoint) were converted to binary scale by the following schemes: for ten-point scale—{1...4} → *neg*, {6...10} → *pos*; for five-point scale—{1...2} → *neg*, {4...5} → *pos*.

As a training set the random chosen ten thousand reviews are used for each domain. For the ROMIP's domains these reviews are chosen from train corpora of ROMIP2011, for remaining domains—from entire corpora. Test sets for ROMIP's domains are equal to the test corpora union of ROMIP2011 and 2012 for each domain separately. As test sets for restaurants and cars all reviews are used except for training reviews.

The characteristics of training and test corpora are given in Table 1.

---

[5] http://www.restoclub.ru

[6] https://cars.mail.ru/reviews

Table 1. Text corpora (N_{av}—an average number of words per review)

Domain	Train corpora				Test corpora				Total reviews
	Pos	Neg	Total	N_{av}	Pos	Neg	Total	N_{av}	
Restaurants	7,982	2,018	10,000	87	15,353	1,544	16,897	162	26,897
Cars	7,900	2,100	10,000	104	38,148	1,286	39,434	71	49,434
Movies	7,330	2,670	10,000	80	594	126	720	212	10,720
Books	7,888	2,112	10,000	31	356	39	395	235	10,395
Cameras	8,921	1,079	10,000	94	612	54	666	226	10,666

It should be noted that the corpora are highly imbalanced: the part of positive reviews is ranging from 73.3% for the movie training corpus to 96.7% for the car test corpus.

4. Sentiment lexicons creating

The proposed procedure of sentiment lexicon creation consists of three main stages: 1) word weighting and selection; 2) collaborative manual word annotation; 3) consolidation of sentiment lexicons.

At the first stage the morphological analysis of training corpus is performed (we used mystem[7]), then full dictionary of training corpus is formed and stop words are removed. All the words are weighted using the supervised term weighting scheme, e.g., RF (Relevance Frequency), which demonstrated good performance in the text categorization task (Lan et al., 2009). In this scheme the weight of a given word to the sentiment category S is calculated by formula:

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

where a—a number of documents related to category S and containing this word, b—a number of documents not related to category S and containing this word as well.

For each word two weights are calculated: the first weight RF_{pos} towards $S = positive$ and second weight RF_{neg} towards $S = negative$. Two identical word lists, which contain all words from full dictionary, are generated. Lists are sorted, the first—in the order of weights RF_{pos}, and the second—in the order of weights RF_{neg}. First P words from each list are chosen so that $2P = N$, where N—a number of words for manual annotation (at the top of both lists the same words may occur). Thus the dictionary for manual labeling containing N hypothetical sentiment words is made for the second stage.

Table 2 shows the characteristics of full dictionaries and dictionaries for labeling.

[7] https://tech.yandex.ru/mystem

Table 2. Size of dictionaries

Domain	Size of full dictionary	Size of labelled dictionary
Restaurants	21,454	10,000
Cars	17,810	10,000
Movies	28,955	10,000
Books	15,328	10,000
Cameras	13,974	10,000

At the second stage M annotators independently label dictionary. In our study $M = 4$ annotators take part in the annotation process. $N = 10{,}000$ is the compromise between the laboriousness and the completeness. The annotators labelled 50,000 words (5 domains) altogether. The dictionary was shuffled before annotation.

Each word can be assigned one of four labels: positive, negative, neutral and unclear. Further neutral words are not used. The unclear word lists are of interest to further studies.

The desktop application that shows the current word, its context and possible labels is used for the labeling process (Fig. 1).

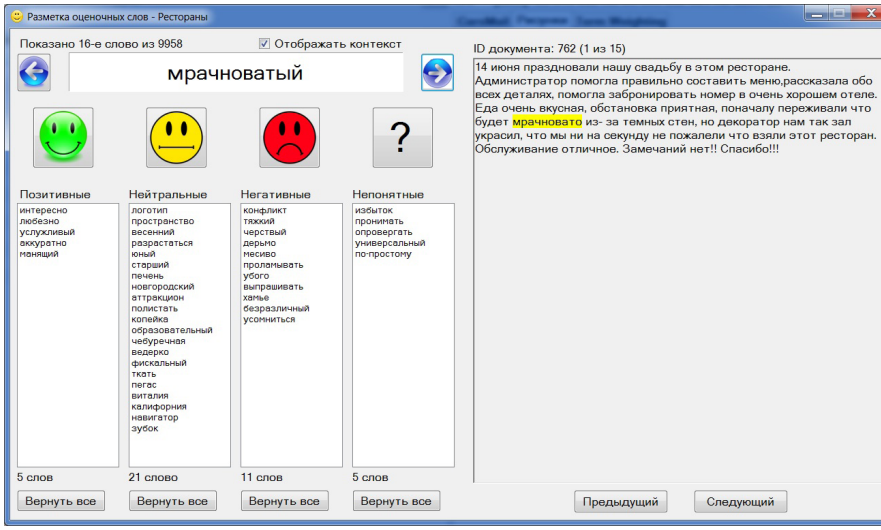

Fig. 1. Annotation tool

The annotators labelled the word as positive or negative in case they could imagine it in any sentiment context of current domain. If the annotator had some doubt the word was labelled as unclear, otherwise as neutral. The average time of labeling of a thousand of words was 90 minutes, overall labeling time was about 300 man-hours.

The anotators had the following main problems:

1) the ambiguity, e. g. «*нашли кусок пластика*» — «*прекрасная пластика танца*» (“we found a piece of plastic”—“a great plastic of dance”);
2) the reviews often have two parts—descriptive and evaluative. The words that are sentiment-bearing for descriptive part are not those for evaluative and vice versa. In (Taboada et al., 2009) the solution of the problem of the descriptive noise is proposed;
3) the author of review’s was afraid of something but his or her fear was not confirmed;
4) for many words a number of reviews containing such words exceeds several tens (and even hundreds)—for the annotators it was hard to see all reviews in such cases;
5) the morphological errors, e.g. word «*отстой*» (“*bullshit*”) is recognized as «*отстоять*» (“*to stand*”);
6) typos, e.g. «*комплимент — комплемент*» (“*compliment—complement*”).

At the third stage positive and negative labelled word lists are joined, domain-dependent and universal sentiment lexicons are formed[8].

5. Analysis of lexicons

5.1. Description

As a result of the proposed procedure each annotator created four lexicons for each of five domains (80 lexicons altogether). The characteristics of lexicon for restaurant domain are shown in Fig. 2.

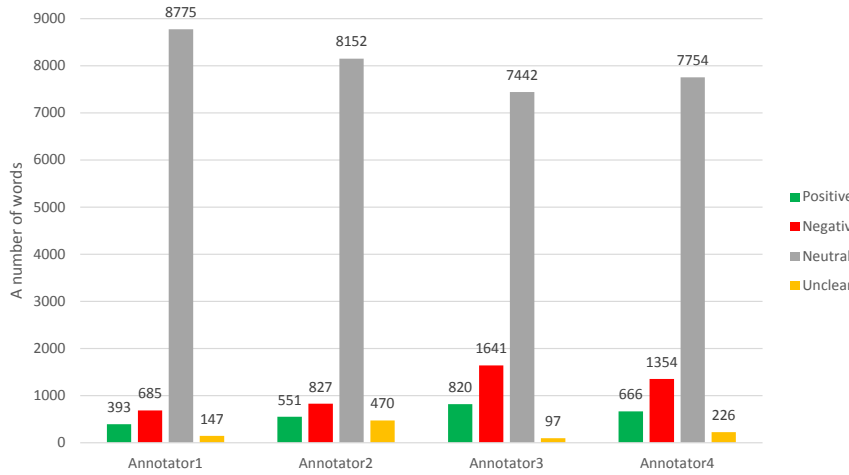

Fig. 2. The distribution of sentiment words for restaurant reviews

[8] Created sentiment lexicons are available at: https://goo.gl/KRWo5X.

The analysis of created lexicons allows us to draw following conclusions. Firstly, negative lexicon is more diverse: on average the size of negative lexicons is 1.63 times more than of positive ones, despite the fact that the positive words prevail in texts (Boucher, Osgood, 1969). Secondly, the annotators differ in the degrees of confidence in their labels: the average rate of unclear words varies from 0.5% to 3.6%. At the same time the intersections of all or the most part of manual lexicons give good results of sentiment analysis comparable with automatic dictionaries (see Section 6).

Thirdly, the rate of sentiment lexicon ranges from 8.4% to 17.3% on average for various domains (Table 3). It should be noted that this rate is in specially collected dictionary of candidate words. For the full dictionary this rate is likely to be even lower.

Table 3. Average sizes of lexicons

Domain	Positive	Negative	Neutral	Unclear	Total	(Pos+Neg)/Total
Restaurants	608	1,127	8,031	235	10,000	17.3%
Cars	429	975	8,444	152	10,000	14.0%
Movies	389	451	9,026	134	10,000	8.4%
Books	491	623	8,754	132	10,000	11.1%
Cameras	535	965	8,382	119	10,000	15.0%

5.2. Intersections and unions

We built the intersection of two types of lexicons: for which all 4 annotators agree and for which at least 3 of 4 annotators agree. The characteristics of these lexicons are shown in Tables 4 and 5.

Table 4. Lexicons with 4 agreed annotators

Domain	Positive	Negative	Neutral	Unclear	Total	Part of labelled dictionary
Restaurants	200	410	6,673	0	7,283	72.8%
Cars	87	159	7,183	0	7,429	74.3%
Movies	87	109	8,123	0	8,319	83.2%
Books	109	155	7,786	1	8,051	80.5%
Cameras	79	89	6,969	0	7,137	71.4%
Average	112	184	7,347	0	7,644	76.4%

Table 5. Lexicons with the minimum 3 agreed annotators

Domain	Positive	Negative	Neutral	Unclear	Total	Part of labelled dictionary
Restaurants	483	857	7,740	14	9,094	90.9%
Cars	342	780	8,091	2	9,215	92.2%
Movies	251	317	8,873	2	9,443	94.4%
Books	359	477	8,507	1	9,344	93.4%
Cameras	396	739	7,974	3	9,112	91.1%
Average	366	634	8,237	4	9,242	92.4%

The study of Tables 4 and 5 shows the decrease in scattering of labelled lexicons parts in the transition from the agreement of all annotators to an agreement of at least three of them: from [71.4%...83.2%] to [90.9%...94.4%]. Thus, the degree of agreement of the majority is higher than 90%.

Also the universal dictionaries were created—the unions of dictionaries for all domains with different minimum number of agreed annotators (Table 6).

Table 6. The characteristics of universal lexicons

A minimum number of agreed annotators	Positive	Negative	Positive ∪ Negative	Neutral	Unclear
1	2,731	4,978	7,526	25,688	2,324
2	1,614	3,338	4,927	24,260	260
3	1,047	2,210	3,247	23,026	22
4	388	724	1,111	21,145	1

It may be noticed that the size of positive and negative lexicons union is less than the sum of positive and negative lexicons sizes separately. The reason is that some words occur in positive and negative lexicons simultaneously. For example in Table 7 there are 10 such words for the minimum three agreed annotators.

Table 7. Words belonging to both universal lexicons

Word	Positive lexicon		Negative lexicon	
	Domain	Examples	Domain	Examples
засасывать	books	*сюжет засасывает*	cameras	*засасывает пыль*
предсказуемость	cars	*предсказуемость в поворотах*	movies, books	*предсказуемость интриги*
непредсказуемость	books	*сюжет нравится непредсказуемостью*	cars, cameras	*непредсказуемость результата съемки*
предсказуемый	cars, cameras	*предсказуемо ведет себя*	books	*конец предсказуем*
непредсказуемый	movies, books	*непредсказуемые реакции героев*	cars, cameras	*непредсказуемые отказы*

Word	Positive lexicon		Negative lexicon	
	Domain	Examples	Domain	Examples
простенько	cameras	*все простенько и со вкусом*	books	*слишком простенько*
цеплять	books	*книга цепляет за живое*	cars	*цепляет днищем землю*
затрепывать	books	*книга уже затрепана*	restaurants	*инвентарь затрепан*
реветь	books	*ревела в три ручья*	cars	*мотор ревет*
разжевывать	cameras	*разжевано для «тормозов»*	books	*разжеванный автором до неприличия*

5.3. Inter-annotator agreement

We compute inter-annotator agreement by means of Fleiss' kappa statistical measure (Fleiss, 1971). It is calculated as the ratio of degree of annotators agreement actually attained above what would be predicted by chance and the degree of agreement attainable above chance:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e},$$

where $\overline{P}$—the mean of the proportions of agreeing annotator-annotator pairs for each word; $\overline{P}_e$—the degree of agreement expected by chance.

If the annotators are in complete agreement then $\kappa = 1$. If there is chance agreement then $\kappa = 0$.

Also we compute inter-annotator agreement for each category—positive, negative, neutral and unclear. The results are shown in Table 8.

Table 8. Inter-annotator agreement

Domain	Positive	Negative	Neutral	Unclear	Fleiss' kappa
Restaurants	0.353	0.364	0.790	0.027	0.535
Cars	0.317	0.306	0.796	0.017	0.471
Movies	0.248	0.284	0.877	0.011	0.462
Books	0.297	0.322	0.849	0.019	0.504
Cameras	0.262	0.274	0.775	0.017	0.432
Average	0.295	0.310	0.817	0.018	0.481

The obtained values of Fleiss' kappa (from 0.432 for cameras to 0.535 for restaurants) on the scale from paper (Landis, Koch, 1977) refer to "the moderate agreement" (0.4...0.6). Although (Artstein, Poesio, 2008) indicate, that only values above 0.8 ensured an annotation of reasonable quality, our experiments show that the created lexicons are of sufficient quality for sentiment analysis (see Section 6).

The relatively low value of Fleiss' kappa = 0.432 for the cameras, is possibly due to a lesser awareness of annotators in this domain than in others.

Note that Fleiss' kappa was lower for movies regarding restaurants (despite the high degree of agreement in the Table 4), due to the high values of the degree of agreement P_e expected by chance.

5.4. Parts of speech

We analyzed parts of speech distribution in the unions of positive and negative lexicons for different domains (see Table 5), formed by at least 3 agreed annotators (Table 9).

Table 9. The distribution of parts of speech

	Nouns		Verbs		Adjectives		Adverbs		Others		Total	
Domain	#	%	#	%	#	%	#	%	#	%	#	%
Restaurants	336	25.1	276	20.6	512	38.2	215	16.0	1	0.1	1,340	100
Cars	281	25.0	338	30.1	377	33.6	125	11.1	1	0.1	1,122	100
Movies	146	25.7	72	12.7	226	39.8	121	21.3	3	0.5	568	100
Books	189	22.6	141	16.9	334	40.0	171	20.5	1	0.1	836	100
Cameras	255	22.5	294	25.9	437	38.5	148	13.0	1	0.1	1,135	100
Universal	865	26.6	834	25.7	1,118	34.4	428	13.2	2	0.1	3,247	100
Average	241	24.6	224	22.0	377	37.4	156	15.9	1.5	0.2		

As a result of the analysis it was found that adjectives occupy the largest part in the sentiment dictionaries (on average 37.4%). Adverbs have the smallest part (15.9%), except for *Others*. Nouns and verbs have approximately the same proportion (24.6% and 22%, respectively).

Verbs have the highest variation of proportions in the domains: from 12.7% for movies to 30.1% for cars. This is probably due to the predominance of actions description in the reviews of the goods (cameras, cars), than in the reviews of the works of art (movies, books).

5.5. Interconnection between manual and automatic lexicons

We compared the sentiment lexicons created by annotators (minimum three agreed) and automatically generated based on the weight RF. If the size of manual lexicon is equal to *N*, we take *N* first words with maximal RF-weights (Table 10).

You may notice that in general, the coincidence is low—on average 17.1% in all lexicons and domains. At the same time the scattering is very large: for the positive—from 6.0% to 33.3%, for the negative—from 11.0% to 31.3%.

Therefore, you should not rely only on automatic methods for sentiment lexicon creating. For example, top-100 RF-weighted positive words for the books domain

contains such neutral words as *подход (an approach)*, *окружающий (surrounding)*, *сестра (a sister)*, *вставать (to stand up)*, *держать (to hold)*, etc. In our opinion, the used hybrid approach where human annotators mark up a subset of the words selected by automatic methods is more promising.

Table 10. A comparison of manual and automatic lexicons

Domain	Positive		Negative	
	Size	Coincidence	Size	Coincidence
Restaurants	483	33.3%	857	31,3%
Cars	342	15.5%	780	20.1%
Movies	251	6.0%	317	11.0%
Books	359	19.2%	477	19.1%
Cameras	396	15.9%	739	14.2%
Average	366	18.0%	634	16.1%

6. Comparison of lexicons in automatic sentiment analysis

We researched the performance of the sentiment analysis for different domains using prepared sentiment lexicons and compared with the dictionaries automatically formed on the basis of train collections, as well as with the existing lexicons (see Section 2).

A vector space model of text representation was used. Automatically created dictionaries based on the training collection were weighted using an RF scheme (Lan et al, 2009). Also a feature selection was applied for the dictionaries—the first p% of words with the highest weights were selected. The ratio p ranged from 10% to 100% with 10% step. For the other dictionaries the binary weights were used.

The method SVM from scikit-learn package (Pedregosa et al., 2011) was used for classification. The kernel (linear, polynomial, RBF), SVM parameters and parameter p in the feature selection through grid search and 3-fold cross-validation were selected. The best results were achieved with a linear kernel.

We included in testing the formed by annotators domain-dependent sentiment lexicons, which contained only the words about which agree all annotators (denoted "Domain, $n = 4$") and most annotators ("Domain, $n = 3$"). In addition, we used universal sentiment lexicons ($n = 3$ and $n = 4$).

We also compared the quality of the analysis with the results of publicly available Russian sentiment lexicons: ProductSentiRus (Chetviorkin, Loukachevitch, 2012), NRC (Mohammad, Turney, 2013) and Chen-Skiena (Chen, Skiena, 2014). The sizes of all the lexicons are listed in Table 11.

As a baseline we used dummy classifier, which categorized all the objects as positive.

For evaluation we used F1-measure, for which macro-averaging was carried out due to the strong imbalance of test corpora. The test results are shown in Table 12.

Table 11. Size of lexicons (for the lexicons of train collection in the brackets it shows the part *p* of the full lexicon—the result of feature selection)

Lexicon	Restaurants	Cars	Movies	Books	Cameras
Dictionaries of train corpus (RF)	21,454 (1.0)	12,467 (0.7)	23,164 (0.8)	15,328 (1.0)	9,781 (0.7)
Domain ($n = 4$)	610	246	196	264	168
Domain ($n = 3$)	1,340	1,122	568	836	1,135
Universal ($n = 4$)	1,111				
Universal ($n = 3$)	3,247				
ProductSentiRus	5,000				
NRC	4,590				
Chen-Skiena	2,914				

Table 12. The results of experiments—F1-measure, %

Lexicon	Restaurants	Cars	Movies	Books	Cameras	Average F1
Baseline	47.6	49.2	45.2	47.4	47.9	47.5
Dictionaries of train corpus (RF)	74.4	63.6	64.4	61.2	80.2	68.8
Domain ($n = 4$)	74.9	62.3	65.2	**64.0**	76.0	68.5
Domain ($n = 3$)	75.0	65.2	62.0	60.5	73.9	67.3
Universal ($n = 4$)	74.3	63.3	61.4	63.1	76.8	67.8
Universal ($n = 3$)	75.3	**65.8**	**65.7**	60.2	78.9	**69.2**
ProductSentiRus	**76.2**	63.6	61.7	59.2	82.6	68.7
NRC	71.8	62.2	58.6	53.6	**82.9**	65.8
Chen-Skiena	71.2	59.6	58.7	56.6	80.2	65.2

From Table 12 it can be seen that the created sentiment lexicons allow to perform the sentiment analysis with high quality, comparable or superior the auto-generated dictionaries. At the same time the size of manual lexicons is much smaller than of automatic lexicons: for example, lexicon *books* (Domain, $n = 4$) comprises a total of 264 words and shows the quality that surpasses all other lexicons (64%).

Also the universal lexicons demonstrate the high quality, for example, the universal lexicon ($n = 3$) shows the best results in two areas of the five (cars and movies), as well as on the average.

Due to the high degree of imbalance of corpora (see Table 1) and the use of macro-averaging scheme, the quality of the analysis highly depends on the F1-measure for negative texts. Almost all relatively low results in Table 12 (e.g., Chen-Skiena for cars, dictionary of train corpus for books, NRC for movies) are closely related with poor recognition of negative texts. Low results of manual lexicons for cameras also depend on it. The reason is the insufficient size of the negative lexicon for cameras (89 words, $n = 4$). Perhaps it was the result of poor awareness of annotators in a given domain.

We note also that ProductSentiRus performed well in the analysis of product reviews (cars and cameras), as well as the restaurants. Lexicons, received by automatic translation into Russian (NRC and Chen-Skiena) tend to show relatively low quality (except cameras).

7. Conclusion

Thus, the proposed in the article procedure allows creating a compact and domain-dependent sentiment lexicon, which is very effective in sentiment analysis. The laboriousness of lexicon creation is reduced through the use of automated methods of terms weighting to generate a set of words to labeling process. It is also important for annotators to be familiar with the domain.

The universal lexicons created by union of manual lexicons also show good results comparable or superior to automatic dictionaries.

We see the following directions for future research: to expand the set of domains (news, social networks, policy) to increase the reliability of research; to investigate the influence of collocations and parts of speech on the effectiveness of lexicons; to test the lexicons with lexical-based method of sentiment analysis (Taboada et al., 2011).

Acknowledgements

The reported study was funded by RFBR according to the research project No. 16-07-00342 a.

References

1. *Amiri F., Scerri S., Khodashahi M.* (2015), Lexicon-based Sentiment Analysis for Persian Text, Proceedings of Recent Advances in Natural Language Processing, Hissar, pp. 9–16.
2. *Andreevskaia A., Bergler S.* (2006), Mining WordNet for Fuzzy Sentiment: Sentiment Tag Extraction from WordNet Glosses, Proceedings EACL-06, the 11rd Conference of the European Chapter of the Association for Computational Linguistics, Trento, pp. 209–216.
3. *Artstein R., Poesio M.* (2008), Inter-Coder Agreement for Computational Linguistics, Computational Linguistics, Vol. 34, No. 4, pp. 555–596.
4. *Blinov P., Kotelnikov E.* (2014), Using Distributed Representations for Aspect-Based Sentiment Analysis, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2014”, Bekasovo, No. 13 (20), Vol. 2., pp. 68–79.
5. *Boucher J. D., Osgood Ch. E.* (1969), The Pollyanna Hypothesis, Journal of Verbal Learning and Verbal Behavior, No. 8, pp. 1–8.

6. *Chen Y., Skiena S.* (2014), Building Sentiment Lexicons for All Major Languages, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, pp. 383–389.
7. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* (2012), Sentiment Analysis Track at ROMIP 2011, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog2012", No. 11 (18), Vol. 2., pp. 1–14.
8. *Chetviorkin I., Loukachevitch N.* (2012), Extraction of Russian Sentiment Lexicon for Product Meta-Domain, Proceedings of COLING 2012, Mumbai, pp. 593–610.
9. *Chetviorkin I., Loukachevitch N.* (2013), Sentiment Analysis Track at ROMIP 2012, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog2013", No. 12 (19), Vol. 2, pp. 40–50.
10. *Feldman R.* (2013), Techniques and Applications for Sentiment Analysis, Communications of ACM, Vol. 56, No. 4, pp. 82–89.
11. *Fellbaum C.* (1998), WordNet: An Electronic Lexical Database, Cambridge, MA: MIT Press.
12. *Fleiss J. L.* (1971), Measuring nominal scale agreement among many raters, Psychological Bulletin, Vol. 76, No. 5, pp. 378–382.
13. *Freitas L., Vieira R.* (2013), Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis, International Journal of Computational Linguistics and Applications, Vol. 4 (1), pp. 147–158.
14. *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.
15. *Kiselev Y., Braslavski P., Menshikov I., Mukhin M., Krizhanovskaya N.* (2015), Russian Lexicographic Landscape: a Tale of 12 Dictionaries, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog2015", No. 14 (21), Vol. 1, pp. 254–50271
16. *Lan M., Tan C. L., Su J., Lu Y.* (2009), Supervised and Traditional Term Weighting Methods for Automatic Text Categorization, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 31 (4), pp. 721–735.
17. *Landis J. R., Koch G. G.* (1977), The Measurement of Observer Agreement for Categorical Data, Biometrics, Vol. 33, pp. 159–174.
18. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5 (1).
19. *Mohammad S., Turney P.* (2013), Crowdsourcing a Word-Emotion Association Lexicon, Computational Intelligence, Vol. 29 (3), pp. 436–465.
20. *Musto C., Semeraro G., Polignano M.* (2014), A Comparison of Lexicon-based Approaches for Sentiment Analysis of Microblog Posts, DART 2014 8th International Workshop on Information Filtering and Retrieval, Pisa.
21. *Ozdemir C., Bergler S.* (2015), A Comparative Study of Different Sentiment Lexica for Sentiment Analysis of Tweets, Proceedings of Recent Advances in Natural Language Processing, Hissar, pp. 488–496.

22. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay É.* (2011), Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
23. *Taboada M., Brooke J., Stede M.* (2009), Genre-based Paragraph Classification for Sentiment Analysis, Proceedings of the 10th Annual SIGDIAL Meeting on Discourse and Dialogue, London, pp. 62–70.
24. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, Vol. 37 (2), pp. 267–307.
25. *Ulanov A., Sapozhnikov G.* (2013), Context-Dependent Opinion Lexicon Translation with the Use of a Parallel Corpus, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2013”, Bekasovo, pp. 165–174.