RESEARCH OF LEXICAL APPROACH AND MACHINE LEARNING METHODS FOR SENTIMENT ANALYSIS

Blinov P. D. (blinoff.pavel@gmail.com),
Klekovkina M. V. (klekovkina.mv@gmail.com),
Kotelnikov E. V. (kotelnikov.ev@gmail.com),
Pestov O. A. (oleg.pestov@gmail.com)

Vyatka State Humanities University, Kirov, Russia

Methods and approaches used by the authors to solve the problem of sentiment analyses on the seminar ROMIP-2012 are described. The lexical approach is represented with the lexicon-based method which uses emotional dictionaries manually made for each domain with the addition of the words from the training collections.

The machine learning approach is represented with two methods: the maximum entropy method and support vector machine. Text representation for the maximum entropy method includes the information about the proportion of positive and negative words and collocations, the quantity of interrogation and exclamation marks, emoticons, obscene language. For the support vector machine binary vectors with cosine normalization are built on texts.

The test results of the described methods are compared with those of the other participants of the ROMIP seminar. The task of classification of reviews for movies, books and cameras is investigated. On the whole. The lexical approach demonstrates worse results than machine learning methods, but in some cases excels it. It is impossible to single out the best method of machine learning: on some collections maximum entropy method is preferable, on others the support vector machine shows better results.

Key words: sentiment analysis, lexical approach, machine learning, maximum entropy method, support vector machine, ROMIP

1. Introduction

Text sentiment analysis is an extensively researched area of computational linguistics in last ten years. The main problem of sentiment analysis is an identification of emotional attitude to some object in a text.

Obviously there are many practical applications for sentiment analysis. For example, opinion analysis of target audience helps to reveal strengths and weaknesses of a commercial product. Automatic rating of movie or book reviews enables to make support recommendations for choice of work. Sentiment analysis systems are also used in sociological and political researches, in human-computer interfaces and in other spheres [12, 15].

A majority of researches in sentiment analysis are made for English texts. For a variety of reasons such studies on Russian text collections were not as popular. Recently, however, the situation began to change for the better: for two years a seminar ROMIP [24] has proposed the sentiment analysis tracks including classification of user reviews into 2, 3, and 5 classes. At a seminar in 2012, two new tasks appeared: the classification of news fragments into 3 classes and opinions search on requests.

The purpose of this paper is to present the results of the participation of the team of authors at sentiment analysis tracks at ROMIP-2012. Two approaches were investigated: lexical approach and machine learning approach.

The reminder of this paper is structured as follows. Section 2 gives an overview of current approaches to the problem of sentiment analysis. In section 3 the method of the lexical approach is considered. Section 4 is devoted to the machine learning methods. Section 5 presents the results of experiments at the ROMIP-2012 and their analysis. We provide concluding remarks and findings in Section 6.

2. Existing approaches

There are two main approaches to the problem of sentiment analysis: lexical approach and machine learning approach [19]. In the lexical approach the definition of sentiment is based on the analysis of individual words and/or phrases; emotional dictionaries are often used: emotional lexical items from the dictionary are searched in the text, their sentiment weights are calculated, and some aggregated weight function is applied [5, 9, 19, 20].

In the machine learning approach the task of sentiment analysis is regarded as a common problem of text classification [17] and it can be solved by training the classifier on a labeled text collection [1, 7, 14, 16].

Each approach has its advantages and disadvantages. When using the lexical approach there is no need for labeled data and the procedure of learning, and the decisions taken by the classifier can be easily explained. However, this usually requires powerful linguistic resources (e.g., emotional dictionary), which is not always available, in addition it is difficult to take the context into account.

In the machine learning approach the dictionary is not required (although it can be used), and in practice the methods demonstrate the high accuracy of classification. However, this accuracy is achieved only with a representative collection of labeled training texts and by careful selection of features. At the same time the classifier trained on the texts in one domain in most cases does not work with other domains [8].

When participating in ROMIP 2012, our team set itself the aim to research the capabilities of both approaches for the classification of user reviews.

3. Lexicon-based method

Within the lexical approach in the seminar ROMIP 2012 the lexicon-based method proposed in [22] was used. This method is based on emotional dictionaries for each domains. The creation of dictionaries was as follows: first of all 60 most impressive emotional Russian words (*хорошо — good, превосходно — great, плохо — bad, отвратительно — disgusting,* etc.) were put in each dictionary and were assigned weights in the range [−5…+5]. Next, each domain dictionary was replenished with appraisal words of appropriate training collection that have the highest weight, calculated by the method of RF (Relevance Frequency) [10]. The weight of a word in the dictionary was also appointed manually from the range [−5…+5]. The quantity of words in the dictionaries varied from 245 to 260.

In addition the dictionaries include word-modifiers (all in all 19, for example, *очень — very, самый — most, несколько — somewhat*, etc.) and the wordnegations (*не, ни, ничего*). The word-modifier changes (increases or decreases) the weight of the following appraisal word by a certain percentage. Word-negation shifts the weight of the following appraisal word by a certain offset: for positive words to decrease, for negative — to increase. Concrete percentages for word-modifiers and the offsets for the word-negations in every dictionary were automatically selected on the basis of cross-validation for the appropriate training collection.

The procedure of the text sentiment classification was carried out as follows. First we calculated the weights of all training texts and of the classified text. The weight of text was defined as the average of the weights of emotional words from the dictionary presented in the text, taking into account the changes made by word-modifiers and word-negations. Thus, all the texts are placed into a one-dimensional emotional space. To improve the accuracy of classification the texts that were too close to the texts of another sentiment class were excluded from the consideration. The proportion of deletions was determined by the cross-validation method.

Then the average weights of training texts for each sentiment class were found. The classified text was referred to the class which was located closer in the one-dimensional emotional space.

4. Machine learning methods

For the research at the seminar ROMIP2012 we chose two machine learning methods, well proved in solving various problems of computational linguistics: Maximum Entropy method (MaxEnt) [2] and Support Vector Machines (SVM) [21].

Both methods use a vector model of the text; to obtain the vector model the only one emotional dictionary (different from the dictionaries in lexical approach) is used.

In this section first the training collections are considered, then the procedure for the building of the dictionary is given, after that the features used in the construction of the vector model of texts are listed, in the conclusion the peculiarities of machine learning methods are shown.

4.1. Training collections

The organizers of the seminar ROMIP2012 granted the following training collections: user reviews of books and movies from advisory service Imhonet[1] and user reviews of cameras from service Yandex.Market[2]. In addition in our research we used a collection of user reviews of movies from ratings "Top 250" and "100 worst" of site Kinopoisk[3] (36 000 reviews). Final training collection contained more than 83 000 reviews.

Preliminarily documents with unknown ratings were removed from the training collection. Ratings of reviews were transferred to the 5-point scale, URL addresses were removed from review contents. Then, for each review such procedures were performed: tokenization, sentence segmentation and morphological analysis; linguistic instruments FreeLing [6] and Mystem [13] were used.

4.2. Dictionary creation

For machine learning methods common emotional dictionary for the three domains: books, movies, cameras was created. Russian sentiment lexicon for product meta-domain [4] was taken as the basis. Subset of words most clearly expressing positive (969 words) and negative (1138 words) emotions were manually selected from it. Next, each word was supplemented with synonyms and antonyms, obtained from Wiktionary[4], after which the number of positive words was 1864, negative — 2215. A similar approach to the completion of the dictionary, only using WordNet, was used in [9].

In order to reflect the nearest context of words, instead of using a list of word-modifiers we included all word collocations of training collection that have the following patterns: *<particle>* + *<dictionary word>*, *<adverb>* + *<dictionary word>*, *<particle>* + *<adverb>* + *<dictionary word>*, etc. For example, an incomplete list of the resulting fragments with a verb *понравиться (like)* is: {*невероятно понравиться, понравиться, понравиться безумно, не понравиться, очень не понравиться, не очень понравиться, …*}. As a result, the final dictionary, created by the method described above contained about 19 000 words and collocations.

For each lexical unit of the dictionary conditional probabilities were computed by means of training collection:

$$p(w|score) = \frac{|M_w|}{|N_{score}|}, \qquad (1)$$

[1] URL: http://imhonet.ru.

[2] URL: http://market.yandex.ru.

[3] URL: http://www.kinopoisk.ru.

[4] URL: http://www.wiktionary.org.

where w — lexical unit of the dictionary, $score \in \{-,+\}$ — review rating (correspondence between the scales: $\{3,4,5\} \to +$, $\{1,2\} \to -$); N_{score} — set of reviews with the rating *score*; $M_w \subseteq N_{score}$ — set of reviews containing lexical unit w.

4.3. Features

To obtain a vector model of a text for the Maximum Entropy method we used vectors containing 7 components:

1. a component, which takes into account the sentiment of lexical units of the text by means of likelihood ratio; it will be discussed later in details;
2. a component, reflecting the ratio of positive and negative lexical units in the text reduced to the following scale: {*much more negative*, *more negative*, *equally*, *more positive*, *much more positive*};
3. the average number of exclamation marks in the text; concrete numerical values were reduced to a scale: {*absence*, *little*, *middle*, *many*, *very many*};
4. the average number of interrogative marks in the text — it was considered in the same way as the average number of exclamation marks;
5. the ratio of positive emoticons in the text to negative emoticons; numerical value is transferred to the scale: {*less*, *equally*, *a little more*, *more*, *much more*}; emoticons are detected using regular expressions;
6. the ratio of negative emoticons in the text to positive emoticons; it was taken into account similarly to the previous component;
7. the binary feature of presence of obscene language in the text; this feature was granted with the morphological analyzer FreeLing [6].

Let’s consider the algorithm of calculating the value of the first component of the feature vector for Maximum Entropy method in details. This component uses the log-likelihood ratio. For each sentence *s* the expression is calculated:

$$L_s = \sum_{i=1}^{m} \ln \frac{p(w_i|-)}{p(w_i|+)}, \tag{2}$$ [*]

where m is the number of words and collocations included in the dictionary, which are found in the sentence s.

The resulting likelihood ratio L for review r then will be:

$$L_r = \frac{\sum_{i=1}^{m} L_i}{n}, \tag{3}$$

where n is the number of sentences of review r.

[*] В бумажном варианте сборника опечатка: в формуле вместо *ln* (натуральный логарифм) напечатано *h*.

The values of likelihood ratios derived from the formula (3) can be considered as the values of a continuous random variable having a normal distribution $N(\mu,\sigma^2)$. As a component of the vector its values can be represented, if they are discretized. According to the three sigma rule at least 95.4% of all the values of a normal random variable fall within the range $(\mu - 2\sigma,\ \mu + 2\sigma)$. Reviews having values L_r that lie outside of this range can be attributed to the boundary values of a five-point scale of 1 and 5. Let's divide this interval into some quantity of parts. In this case the first component of the feature vector for the review will be a number of interval in which the value L_r falls.

For the Support Vectors Machines in accordance with the results of [23] the reviews were represented as binary vectors with cosine normalization. The dimension of these vectors coincides with the dimension of machine learning dictionary. In this case the i^{th} component of the vector representing the review is equal to one if the i^{th} element of the dictionary is present in a review.

4.4. Classification methods

Among the proposed formulations of the problem of classification at the ROMIP-2012 the most common is the 5-point classification task. If the solution of this task is known, the solution of 2-point and 3-point classification tasks can be automatically received by combining the reviews of different classes. For example, the conversion from 5 classes to 3 classes: $\{4,5\}\rightarrow 3$, $\{3\}\rightarrow 2$, $\{1,2\}\rightarrow 1$; from 5 classes to 2 classes: $\{3,4,5\}\rightarrow 2$, $\{1,2\}\rightarrow 1$. Guided by these considerations, it was decided to implement the classification with machine learning methods only for 5point scale. Classification decisions for 2-point and 3-point scales were obtained by combining the reviews as described above.

The Maximum Entropy method is implemented using the library SharpEntropy [18]. Conditional probability distribution $p(y \mid x)$, $y \in Y$, $x \in X$ is modeled in the method, where $Y = \{1,2,3,4,5\}$ is the set of ratings, X — the set of input vectors. Such distribution must be consistent with the training data, but also be as even as possible. Mathematical measure of the uniformity of the distribution is the *entropy* [2]:

$$H(y|\,x) \;=\; -\sum\nolimits_{(x,y)\,\in Z} p\;(y,x)\,\log p(y|\,x), \tag{4}$$

where $Z = X \times Y$ is the Cartesian product of the sets X and Y.

From the set of all possible distributions the one that maximizes the entropy is chosen (4):

$$p(y|\,x) \;=\; \underset{p(y|\,x)\,\in Z}{\arg\max}\, H(y|\,x), \tag{5}$$

To implement the Support Vectors Machines the library LIBSVM [11] was used. The selection of the kernel and optimal parameters was conducted. As in [23] the best results a linear kernel with regulating parameter $C = 1$ produced.

5. Experimental results

Let's consider the results of the classification of user reviews at the seminar ROMIP2012. In this section our methods are identified as follows: *Dict* — the lexicon-based method, *MaxEnt* — the Maximum Entropy method; *Svm* — the Support Vectors Machines; *yyy-N* — the code of our results, *xxx-N* — the code of the results of other participants.

Tables 1-3 show the results of the classification of user reviews to 2, 3 and 5point scales (for technical reasons the lexicon-based method for a 5-point scale was not used; instead of it the variant of the Maximum Entropy method was used, which takes into account the uneven distribution of reviews in classes — in Table 3 this method is identified as *MaxEntT*). The values of *precision* (P), *recall* (R), *F1measure* (F1), computed by *macro-averaged* variant, and value of *accuracy* [3] are shown.

Table 1. Two-class classification results

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
MaxEnt (yyy17)	1 from 43	book	0,749	0,684	0,715	0,884
xxx1	2 from 43	book	0,667	0,748	0,705	0,822
Dict (yyy31)	5 from 43	book	0,627	0,684	0,655	0,798
Svm (yyy7)	13 from 43	book	0,593	0,593	0,593	0,814
Svm (yyy12)	1 from 25	camera	0,589	0,774	0,669	0,895
xxx13	2 from 25	camera	0,688	0,635	0,660	0,961
Dict (yyy6)	9 from 25	camera	0,541	0,626	0,580	0,876
MaxEnt (yyy17)	10 from 25	camera	0,569	0,588	0,579	0,937
xxx19	1 from 26	movie	0,695	0,719	0,707	0,806
MaxEnt (yyy23)	2 from 26	movie	0,731	0,641	0,683	0,831
Svm (yyy13)	5 from 26	movie	0,680	0,642	0,660	0,809
Dict (yyy7)	6 from 26	movie	0,659	0,659	0,659	0,789

Table 2. Three-class classification results

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
Dict (yyy10)	1 from 18	book	0,532	0,591	0,560	0,659
xxx17	2 from 18	book	0,544	0,554	0,549	0,698

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
MaxEnt (yyy13)	3 from 18	book	0,505	0,532	0,518	0,752
Svm (yyy11)	5 from 18	book	0,454	0,485	0,469	0,690
Svm (yyy12)	1 from 14	camera	0,399	0,602	0,480	0,742
xxx1	2 from 14	camera	0,440	0,498	0,467	0,523
MaxEnt (yyy2)	4 from 14	camera	0,419	0,481	0,448	0,805
Dict (yyy4)	10 from 14	camera	0,370	0,391	0,380	0,745
MaxEnt (yyy11)	1 from 14	movie	0,569	0,479	0,520	0,694
xxx6	2 from 14	movie	0,486	0,521	0,503	0,596
Dict (yyy0)	3 from 14	movie	0,505	0,477	0,491	0,627
Svm (yyy2)	7 from 14	movie	0,454	0,445	0,449	0,640

Таблица 3. Five-class classification results

Run_ID	Position	Object	Macro_P	Macro_R	Macro_F1	Accuracy
Svm (yyy1)	1 from 5	book	0,339	0,496	0,402	0,481
MaxEnt (yyy4)	2 from 5	book	0,330	0,460	0,384	0,473
MaxEntT (yyy2)	3 from 5	book	0,219	0,402	0,284	0,380
Svm (yyy3)	1 from 5	camera	0,272	0,441	0,336	0,457
MaxEntT (yyy1)	2 from 5	camera	0,258	0,326	0,288	0,489
MaxEnt (yyy2)	3 from 5	camera	0,246	0,315	0,276	0,470
MaxEnt (yyy2)	1 from 5	movie	0,401	0,352	0,375	0,407
Svm (yyy1)	2 from 5	movie	0,330	0,317	0,323	0,385
MaxEntT (yyy3)	3 from 5	movie	0,318	0,319	0,319	0,382

After analyzing the results the following theses can be concluded:

1. The lexical approach in our study showed significantly worse results than the methods of machine learning. Out of 6 tasks of 2-class and 3class classification in only one case (reviews of books, 3class) the lexicon-based method

was better than the other two methods. Perhaps this is related to the small size of the dictionary (no more than 300 words) and a lack of time to adjust the lexicon-based method.

2. We cannot make an unambiguous conclusion about the benefits of one machine learning method over the other: the Maximum Entropy method always shows the best results on a collection of movie reviews, while the Support Vectors Machines always exceed at a collection of camera reviews. In the case of book reviews for 5-class task the SVM show a slight advantage (2% by value F1), and in the other two problems — on the contrary, the results of the MaxEnt predominate (by 12% and 5%).

All in all, out of 9 tasks the Maximum Entropy method has shown results in 5 tasks higher than the Support Vectors Machines.

3. It is impossible also to summarize the effectiveness of different methods on the parameters of precision and recall: in different tasks all methods show different ratios of these important parameters — sometimes precision dominates, sometimes recall.
4. When the quantity of classes increases the results reduce, although not as dramatically as in the seminar ROMIP2011: for camera reviews the best result for binary classification is 67%, for 5-class task — 34% (43% down) while at the seminar ROMIP 2011 the decrease was 66% (from 92% to 26%). Thus we can conclude that the methods used in the seminar ROMIP2012 are more steady to the increase of the quantity of classes.

In addition to the problems of user reviews classification a new task of the classification of the fragments of direct and indirect speech from news articles was offered at the seminar. It was proposed to perform the classification for 3-point scale. The essential features of this problem should be noted: first, a small amount of the content of each fragment, secondly, the greater thematic variation. To solve this problem we used the MaxEnt and the SVM methods with emotional dictionary, created for reviews classification. Both methods showed low results because of the fact that the dictionary didn't sufficiently reflect the specific emotional terms of the news domain.

6. Conclusion

Thus this paper focuses on two main approaches to the problem of sentiment analysis — lexical approach and machine learning. In the first approach the lexicon-based method developed by the authors was used, which differs from the existing methods by the way of creating both emotional dictionaries for each domain and the algorithm which calculates the weight of texts. Machine learning approach was presented to the Maximum Entropy method and the Support Vectors Machines; it used the technique developed by the authors to create a dictionary and an algorithm for the construction of the feature vector for the Maximum Entropy method.

As a result, as in many other studies, the benefits of machine learning methods are demonstrated, but the lexical approach even with a small dictionary in some cases shows the best results among the others methods. So, perhaps, the lexical approach should not be rejected, rather, the combination of both approaches is promising.

The participation of our team in seminar ROMIP2012 was very productive: out of 9 reviews of classification task our methods took first place in 8 cases according to the metric of F1, and in one case — the second place.

We would like to thank the organizers for their considerable efforts and express hope for further development of ROMIP which has a major positive impact on research in computational linguistics and information retrieval in Russia.

References

1. *Agarwal A., Xie B., Vovsha I., Rambow O., Passonneau R.* (2011), Sentiment analysis of twitter data, Proceedings of the Workshop on Language in Social Media (LSM 2011), pp. 30–38.
2. *Berger A. L., Della Pietra S., Della Pietra V.* (1996), A maximum entropy approach to natural language processing, Journal Computational Linguistics, Vol. 22(1), pp. 39–71.
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. *Chetviorkin I. I., Loukachevitch N. V.* (2012), Extraction of Russian sentiment lexicon for product meta-domain, Proceedings of COLING 2012: Technical Papers, pp. 593–610.
5. *Ding X., Liu B., Yu P. S.* (2008), A holistic lexiconbased approach to opinion mining, Proceedings of the Conference on Web Search and Web Data Mining (WSDM), pp. 231–240.
6. *FreeLing 3.0* An open source suite of language analyzer, available at: http://nlp.lsi.upc.edu/freeling/.
7. *Go A., Bhayani R., Huang L.* (2009), Twitter sentiment classification using distant supervision, Association for Computational Linguistics, pp. 30–38.
8. *He Y.* (2012), Incorporating sentiment prior knowledge for weakly supervised sentiment analysis, ACM Transactions on Asian Language Information Processing, Vol. 11(2).
9. *Hu M., Liu B.* (2004), Mining and summarizing customer reviews, Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD2004), Seattle, pp. 168–177.
10. *Lan M., Tan C. L., Su J., Lu Y.* (2009), Supervised and traditional term weighting methods for automatic text categorization, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 31(4), pp. 721–735.
11. *LIBSVM* — A library for support vector machines, available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
12. *Liu B.* (2012), Sentiment analysis and opinion mining, Morgan & Claypool Publishers.
13. *Mystem,* available at: http://company.yandex.ru/technology/mystem.

14. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up? Sentiment classification using machine learning techniques, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86.
15. *Pang B., Lee L.* (2008), Opinion Mining and Sentiment Analysis, Foundations and Trends® in Information Retrieval, No. 2.
16. *Saif H., He Y., Alani H.* (2012), Alleviating data sparsity for twitter sentiment analysis, Workshop: The 2nd Workshop on Making Sense of Microposts (#MSM2012): Big things come in small packages at World Wide Web (WWW), Lyon, France.
17. *Sebastiani F.* (2002), Machine learning in automated text categorization, ACM Computing Surveys, Vol. 34, pp. 1–47.
18. *SharpEntropy*, available at: http://www.codeproject.com/Articles/11090/Maximum-Entropy-Modeling-Using-SharpEntropy.
19. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexiconbased methods for sentiment analysis, Computational Linguistics, Vol. 37(2), pp. 267–307.
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. *Vapnik V.* (1998), Statistical learning theory, New York, Wiley.
22. *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.
23. *Kotelnikov E. V., Klekovkina M. V.* (2012), Sentiment analysis of texts based on machine learning methods [Avtomaticheskij analiz tonalnosti tekstov na osnove metodov mashinnogo obuchenija], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012” [Komp’iuternaia Lingvistika i Intellektual’nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii “Dialog 2012”], Bekasovo, pp. 753–762.
24. *Russian* Information Retrieval Evaluation Seminar (ROMIP). URL: http://romip.ru/