Active learning enhanced document annotation for sentiment analysis

Peter Koncz[1], Ján Paralič[1]

[1]Dept. of Cybernetics and Artificial Intelligence,
Technical University of Košice, Slovak Republic
{peter.koncz, jan.paralic}@tuke.sk

Abstract. Sentiment analysis is a popular research area devoted to methods allowing automatic analysis of the subjectivity in textual content. Many of these methods are based on the using of machine learning and they usually depend on manually annotated training corpora. However, the creation of corpora is a time-consuming task, which leads to necessity of methods facilitating this process. Methods of active learning, aimed at the selection of the most informative examples according to the given classification task, can be utilized in order to increase the effectiveness of the annotation. Currently it is a lack of systematical research devoted to the application of active learning in the creation of corpora for sentiment analysis. Hence, the aim of this work is to survey some of the active learning strategies applicable in annotation tools used in the context of sentiment analysis. We evaluated compared strategies on the domain of product reviews. The results of experiments confirmed the increase of the corpus quality in terms of higher classification accuracy achieved on the test set for most of the evaluated strategies (more than 20% higher accuracy in comparison to the random strategy).

Keywords: sentiment analysis, active learning, semi-automatic annotation, text mining.

1 Introduction

Sentiment analysis, also called opinion mining, is a research area devoted to the methods of automatic quantification of the subjective content expressed in the form of natural language [1]. It aims to detect the presence, orientation or even the intensity of the opinion related to the object of the evaluation or its features/aspects in case of aspect-based sentiment analysis. Within the methods of sentiment analysis, one of the main streams is represented by methods based on using of machine learning algorithms, which deals with the task of sentiment analysis as with a text categorization task. However, these methods are dependent on manually annotated corpora for the training of the classifiers. Moreover these classifiers have been shown domain dependent, i.e. the classifier created for one domain is hardly portable to other domain. Hence, in real application scenario it is usually necessary to build separate corpora for different domains. In order to make the annotation process more effective methods of

adfa, p. 1, 2011.
© Springer-Verlag Berlin Heidelberg 2011

active learning can be utilized. However active learning is a common strategy used to increase the efficiency of classifiers creation, currently it is a lack of systematical research devoted to its application in the context of sentiment analysis. The need of the integration of active learning to annotation tools for sentiment analysis led us to the comparative evaluation of six active learning strategies. The rest of the work is divided as follows. The second chapter is devoted to related works. In the third chapter the evaluated active learning strategies are described. Consequently the fourth chapter describes their experimental evaluation. Finally, the last part is devoted to conclusions.

2 Related work

In the last years it was published a huge amount of works devoted to sentiment analysis. Their comprehensive overview can be found in the work of Liu [2]. The methods of sentiment analysis are typically divided into two groups.

The first group is represented by so-called lexicon-based methods. These methods are usually based on sentiment dictionaries and rules for working with them [3]. Examples of sentiment dictionaries include the *MPQA subjectivity lexicon*[1], used also in our work, *Appraisal lexicon*[2], *National Taiwan University Sentiment Dictionary*[3] or the *SentiWordNet*[4]. There are also some works devoted to the possibilities of automation of dictionaries creation, which usually try to utilize the existing structured knowledge resources. Kamps et al. [4] described a method for identification of adjectives orientation on the basis of their distance from the words *good* and *bad* in a WordNet graph. In the work of Kim and Hovy [5] another solution using WordNet was proposed, based on the extension of the set of emotional words using their synonyms and antonyms. Another solution to populate sentiment dictionaries is to use seed sets of polarity words and extend it by analysis of their coincidences with other words in the corpus of documents as in the work of Turney [6]. These dictionaries, as it will be described in the following section, can be utilized in order to compute the sentiment classification uncertainty. Unlike the above mentioned works, the method for generation of sentiment dictionaries used in this work is based only on the annotated corpora of documents.

The second group of sentiment analysis methods is represented by machine learning based or corpus based methods [3]. These methods use manually annotated corpora, on the basis of which classifiers are trained for identification of the sentiment. In the frame of this group of methods support vector machines (SVM) [1], [7–11] and naïve Bayes classifier (NBC) [10], [11] have been widely used. SVM was also shown to be the most advantageous method in the scope of our own works [1], [3]. Besides the learning algorithms feature selection methods have been also intensively studied [1], [7], [8], where information gain has been shown as one of the most effective

[1] www.cs.pitt.edu/mpqa
[2] lingcog.iit.edu/arc/appraisal_lexicon_2007b.tar.gz
[3] nlg18.csie.ntu.edu.tw:8080/lwku
[4] sentiwordnet.isti.cnr.it

methods. One of the main drawbacks of this group of sentiment analysis methods is their dependency on manually annotated corpora. For the robustness of the created solutions it is necessary to take into account the differences between the analyzed texts. Depending on the source as well as other characteristics the particular texts differ in language, size, formality or in the usage of nonlinguistic expression forms like emoticons [12]. Moreover the created classifiers are domain dependent [13], [14]. In praxis this leads to necessity of separate corpora for particular objects of evaluation, source types or languages. Hence, methods which try to increase the effectiveness of the annotation tools, like active learning, are becoming interesting.

The basic idea of active learning is the selection of unlabeled examples for manual annotation on the basis of their informativeness for the classification task [15]. The particular active learning methods differ in the way how this informativeness is calculated. A comprehensive overview of active learning methods in the context of natural language processing can be found in the work of Olson [15]. Despite the popularity of sentiment analysis in last years, the number of works related to the possibilities of their improvement by utilization of active learning is small. Boiy and Moens [16] in their work used three active learning strategies for sentiment analysis of blogs, reviews and forum contributions. The first of them was the classical uncertainty sampling based on the uncertainty of the classification of unlabeled data, which performed similarly as the random strategy. The second strategy was based on relevance sampling which uses examples most likely to be low-represented class members in order to increase the number of examples from class where there are too few examples. The third method was the *kernel farthest first* which uses examples farthest from the already labeled examples; however this strategy performed worse than the random sampling.

The problem of imbalanced samples is addressed by Li et al. [17], where an active learning approach for imbalanced sentiment classification tasks was proposed. They used complementary classifiers where the first one was used to get most certain examples from both classes and the second to get most uncertain examples from the minority class for manual annotation, while the classifiers were trained with two disjoint feature subspaces. Dasgupta and Ng [14] used active learning based on SVM where the classifier was trained on automatically labeled unambiguous reviews to identify ambiguous reviews for manual labeling and they confirmed the increase of annotation effectiveness. Zhou et al. [13] used active learning based on semi-supervised learning algorithm called active deep network. The above mentioned works computed the informativeness of examples using the classifiers uncertainty. However, there are some specifics of sentiment analysis which should be considered in the context of active learning. One of them is the possibility to use the above mentioned sentiment dictionaries, which represents an alternative to the informativeness computation methods based on the classifiers uncertainty. Therefore the aim of this work is to verify the possibilities of the use of methods from both mentioned groups of active learning methods considering their applicability in annotation tools.

3 Active learning strategies for sentiment analysis

On the basis of previous research we selected six active learning strategies applicable in the context of sentiment analysis, which were evaluated in a series of experiments. The first group of methods is represented by methods which use the confidence estimates for particular classes based on results of classification models, which is a common method also for other application domains of active learning. This group is represented by the first three of the bellow mentioned strategies, where the informativeness was computed according to the equation 1. The equation is based on the information entropy using the posterior probability of the positive class $\hat{P}(C_a|X)$ and its supplement to one for the negative class for the feature vector X of the classified document.

$$Inf = -\hat{P}(C_a|X)\, log_2\, \hat{P}(C_a|X) - (1 - \hat{P}(C_a|X))\, log_2(1 - \hat{P}(C_a|X)) \quad (1)$$

Active learning strategy based on SVM. This active learning strategy uses the probability estimates of target classes based on SVM, which is well performing method for sentiment analysis and it is commonly used for active learning tasks. In this strategy the model trained on the corpus of annotated documents was used to classify the unlabeled documents. The informativeness of the unlabeled examples is then computed on the basis of difference between the probabilities of target classes. The probability of the positive class for the equation 1 is computed according to the equation 2, where $\hat{P}(C_a|X)$ is the posterior probability for the positive class for vector X and $f(X)$ is the output of the SVM. The probability of the negative class is computed as its supplement.

$$\hat{P}(C_a|X) = 1/(1 + exp(-f(X))) \quad (2)$$

The problem of this strategy is the relatively large amount of documents which don't contain any of the words from the model trained on corpus. In this case the computed probability equals for both classes; however these documents are not considered to be the best candidates for annotation. Hence these documents wouldn't be selected for annotation.

Active learning strategy based on Naïve Bayes classifier. Within the methods of machine learning NBC is another commonly used method for sentiment analysis. Also here is the informativeness of the document evaluated on the basis of equation 1. The probability of the positive class is computed on the basis of equation 3.

$$\hat{P}(C_a|X) = \hat{P}(X|C_a).\hat{P}(C_a)/\hat{P}(X) \quad (3)$$

Active learning strategy based on external model. The above mentioned strategies use classifiers trained on the corpora of annotated documents created during prior iterations. This strategy uses available corpora of annotated documents for sentiment analysis. An example of such a corpus, commonly used for evaluation of sentiment analysis methods, is the movie review corpus from the work of Pang and Lee [10], described in more details in the next section. The created model is then used to compute the classification uncertainty as in the previous strategies. The main issue in the

utilization of these corpora is the mentioned domain dependency. On the other hand the whole pool of unlabeled documents is evaluated only once.

Active learning strategy based on external dictionaries. This and the following strategies are based on using of dictionaries of positive and negative words. An example sentiment dictionary is the commonly used MPQA Subjectivity Lexicon [18], which was used also in this work. From this dictionary we extracted a list of positive and negative words. The number of occurrences of positive words N_p and negative words N_n in each document was used to compute the values of informativeness according to the equation 4. This strategy as well as the following strategies is based on assumption that documents containing words with opposite sentiment are more informative.

$$Inf = -\frac{N_p}{N_p+N_n} log_2 \frac{N_p}{N_p+N_n} - \frac{N_n}{N_p+N_n} log_2 \frac{N_n}{N_p+N_n} \quad (4)$$

Active learning strategy based on offline generated dictionaries. As it was mentioned in the related works, there are many methods which can be used to get dictionaries of positive and negative words, while the used method does this using the annotated documents. In our case information gain was used, which is a well performing method in feature extraction tasks. Features are in the case of text classification represented by vectors indicating the presence or absence of words or n-grams of words. Information gain was used to get the list of words relevant for sentiment analysis. The orientation of words was extracted as a simple sign of the difference between the number of occurrences of word in positive and negative evaluations. As a source of evaluations we used the corpus from the work of Pang and Lee [10]. Thus the sentiment dictionary was created before the active learning, hence the name offline.

Active learning strategy based on online generated dictionaries. This strategy is analogous with the previous one, while the difference is in the continuous (online) generation of dictionaries of positive and negative words on the basis of actual corpora of annotated evaluations in each iteration.

4 Experiments

The aim of the series of experiments was to evaluate the performance of the described active learning strategies in context of their application in annotation tools. For the comparison of evaluated strategies we used the experimental design depicted in Figure 1. At the beginning of the annotation we have a pool of unlabeled documents, from which we select examples for annotation. From this pool are in each iteration selected documents for annotation. The selection of documents is realized on the basis of compared active learning strategies, while in the first iteration the documents are selected randomly. Annotated documents are subsequently added to the corpus. As it will be described in the next chapter, the pool was in reality created by documents with known polarity, but the documents' polarity was not taken into account until adding them to the corpus. This solution was used to simulate the process of the real annotation. The number of documents added in each iteration to the corpus was 20

and in experiments were realized 100 iterations, i.e. the maximal size of corpora was 2000 documents. This corpora were used in each iteration to train the classifiers, which accuracy was evaluated on dedicated test set. The achieved accuracies correspond with the quality of the created corpus and the active learned strategy used to its creation. For the creation of classifiers we used SVM with linear kernel and C equal to 0. In parallel the informativeness of documents in pool was evaluated in case of strategies where this is done in each iteration. These values are then used to select new documents for annotation in the next iteration. The whole process finishes by achieving the required size of corpus. Depending on the strategy sentiment dictionaries and external classifiers should be used. It should be also mentioned that we used binary vector representation of documents and each word was stemmed using the Porter stemmer. Besides the compared active learning strategies we used the baseline strategy, which is a simple random selection of documents from the pool of unlabeled documents.

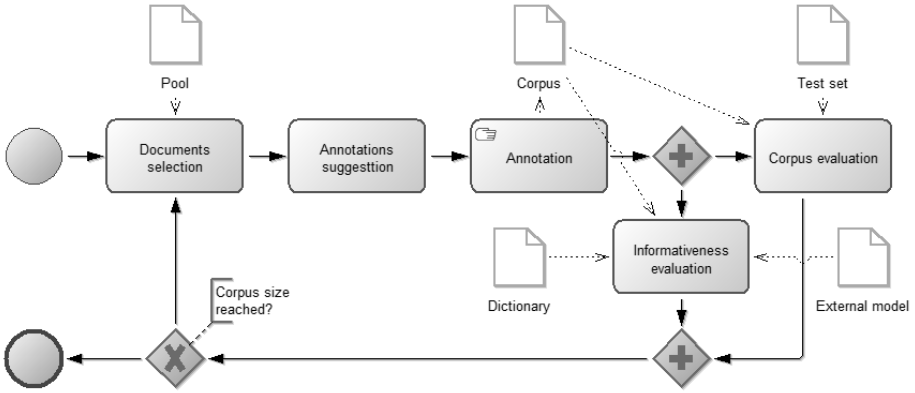


Fig. 1. Experimental design

4.1 Samples

For evaluation of possibilities of particular active learning strategies it was necessary to simulate the process of real annotation. Possible resources of evaluations with corresponding numerical evaluation representing the annotation are product review sites. For this purpose we crawled and parsed reviews from Reviewcenter[5]. From the original set of reviews a total of 17,240 were selected, with half made up by negative reviews (an evaluation from the interval 0 to 1) and half by positive reviews (evaluations equal to 5). The sample was created by random selection of reviews from different fields, with the aim of to create a balanced sample in respect to reviews orientation. From this sample was by stratified random sampling created a test set of evaluations with 1000 documents. The rest of the evaluations were used as a pool of unlabeled documents. The corpora used for training of classifiers were built on the fly by

[5] http://www.reviewcentre.com/

adding new documents in each iteration according to the informativeness of documents evaluated according to active learning strategies.

The data sample used for the active learning strategy based on external model and offline generated dictionaries was created from the corpus used by Pang and Lee [10]. It is a corpus made up of 1000 positive and 1000 negative reviews of films. This data sample is one of the most commonly used samples for sentiment analysis.

4.2 Results

The results achieved by using of particular active learning strategies are depicted in graphs on figures 2 and 3. Axis x represents the sizes of corpora, whereas axis y represents the achieved classification accuracy on the test set. Besides the accuracy values achieved by compared strategies they are depicted also the accuracy values for the baseline strategy as well as the accuracy achieved by using the whole pool of documents for training, which equals 93,6%. The maximal accuracy achieved by the baseline strategy was 74% by using the maximal size of corpus.

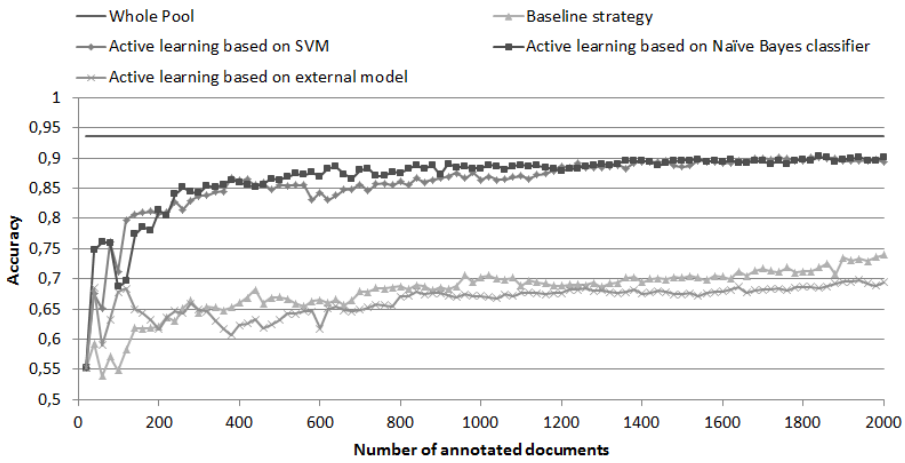


Fig. 2. Results for methods using classifier uncertainty

In Figure 2 are depicted the achieved accuracies for methods using the class probability estimates based on classifiers, i.e. active learning strategies based on SVM, NBC and external model. The best results were achieved by using active learning strategy based on SVM. The maximal achieved accuracy for this strategy was 90,2%. Similar accuracies were achieved also by using the active learning strategy based on NBC. The maximal achieved accuracy for this strategy was 90,3%. For both strategies were the values of accuracy significantly higher in comparison to baseline strategy. In case of active learning based on external model wasn't achieved increase of accuracy.

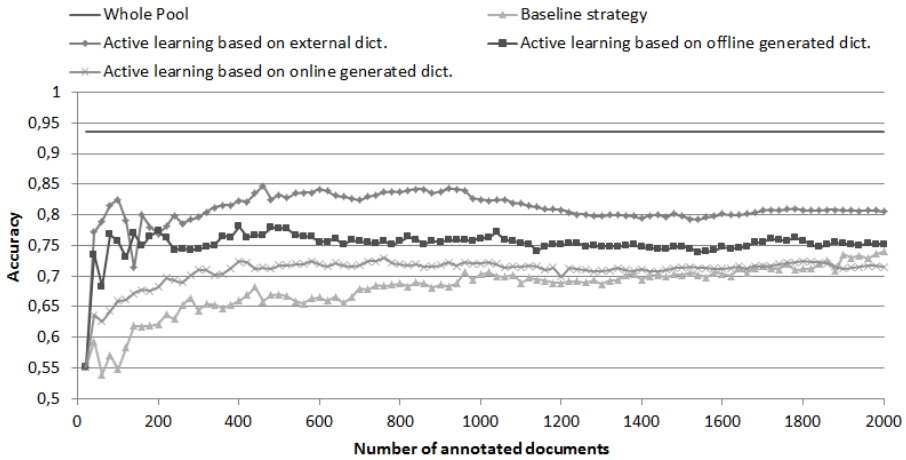


Fig. 3. Results for methods using sentiment dictionaries

In Figure 3 are depicted the achieved accuracies for methods using sentiment dictionaries, i.e. active learning strategy based on external dictionaries and offline and online generated dictionaries. The best results were achieved by using active learning based on external dictionaries. The maximal achieved accuracy for this strategy was 84,7%. Active learning based on offline generated dictionaries achieved worse results. The maximal achieved accuracy for this strategy was 78,1%. Lowest accuracy was achieved by active learning strategy based on online generated dictionaries. Also in case of these methods were the achieved accuracies better than in case of baseline strategy, however the increase of accuracy for these methods was not so significant. It should be also mentioned that the accuracy increase for these strategies was more significant in first iterations and in some cases the accuracy even decreased after adding new documents.

5 Conclusions

The achieved results verified the efficiency of active learning methods in process of documents annotation for the needs of sentiment analysis. From the active learning strategies based on classifiers uncertainty active learning based on SVM has been shown as the best performing strategy. One of its advantages is the independency on quality of external resources. Its disadvantage is the necessity of model creation after each iteration. From the active learning strategies based on dictionaries the strategy using an external dictionary achieved the best performance. The advantage of this strategy is in the simplicity of its application. Moreover the dictionaries can be used for highlighting of words in annotated document and improve the efficiency of annotation. From the point of view of the application of these strategies in annotation tools the combination of both types of strategies should be useful, where in the initial phase of annotation the dictionary based methods can be used which will be then replaced

by classifiers with uncertainty based methods. In our future work we will design such kind of combined annotation supporting active learning method and adjust it also for the purpose of aspect-based sentiment analysis.

Acknowledgment. This work was partially supported by the Slovak Grant Agency of the Ministry of Education and Academy of Science of the Slovak Republic under grant No. 1/1147/12 and partially by the Slovak Research and Development Agency under the contract No. APVV-0208-10.

References

1. P. Koncz and J. Paralic, "An approach to feature selection for sentiment analysis," in *15th IEEE International Conference on Intelligent Engineering Systems (INES 2011)*, 2011, pp. 357–362.
2. B. Liu, *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.
3. P. Koncz and J. Paralič, "Automated creation of corpora for the needs of sentiment analysis," *3rd RapidMiner Community Meeting and Conference (RCOMM 2012)*. Aachen: Shaker Verlag, Budapest, pp. 107–113, 2012.
4. J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke, "Using WordNet to Measure Semantic Orientations of Adjectives," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 2004, pp. 1115–1118.
5. S.-M. Kim and E. Hovy, "Automatic Detection of Opinion Bearing Words and Sentences," in *Proceedings of the Second International Joint Conference on Natural Language Processing (JCNLP 2005)*, 2005, pp. 61–66.
6. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP 2002)*, 2002, pp. 79–86.
7. A. Abbasi, H. Chen, and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 1–34, 2008.
8. A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting Attributes for Sentiment Classification Using Feature Relation Networks," *Ieee Transactions on Knowledge and Data Engineering*, vol. 23, no. 3, pp. 447–462, 2011.
9. R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *Journal of Informetrics*, no. Volume 3, Issue 2, pp. 143–157, 2009.
10. B. Pang and L. Lee, "A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts," *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, Barcelona, Spain, p. 271, 2004.
11. R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, no. Volume 181, Issue 6, pp. 1138–1152, 2011.
12. G. Petz, M. Karpowicz, H. Fürschuß, A. Auinger, S. Winkler, S. Schaller, and A. Holzinger, "On Text Preprocessing for Opinion Mining Outside of Laboratory Environments," in *Active Media Technology SE - 62*, vol. 7669, R. Huang, A. Ghorbani, G. Pasi, T. Yamaguchi, N. Yen, and B. Jin, Eds. Springer Berlin Heidelberg, 2012, pp. 618–629.

13. S. Zhou, Q. Chen, and X. Wang, “Active deep learning method for semi-supervised sentiment classification,” *Neurocomputing*, vol. null, no. null, May 2013.
14. S. Dasgupta and V. Ng, “Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification,” in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language*, 2009, pp. 701–709.
15. F. Olsson, “A literature survey of active machine learning in the context of natural language processing SE - SICS Technical Report,” Swedish Institute of Computer Science, Box 1263, SE-164 29 Kista, Sweden, 2009.
16. E. Boiy and M.-F. Moens, “A machine learning approach to sentiment analysis in multilingual Web texts,” *Information Retrieval*, vol. 12, no. 5, pp. 526–558, Sep. 2008.
17. S. Li, S. Ju, G. Zhou, and X. Li, “Active learning for imbalanced sentiment classification,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 139–148.
18. T. Wilson, J. Wiebe, and P. Hoffmann, “Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis,” in *Proceeding of the conference on empirical methods in natural language processing (EMNLP 2005)*, 2005, pp. 347–354.