ТЕСТИРОВАНИЕ ПОДХОДА К КЛАССИФИКАЦИИ ОТЗЫВОВ ОБ ОБЪЕКТАХ ИЗ РАЗЛИЧНЫХ ПРЕДМЕТНЫХ ОБЛАСТЕЙ — РОМИП 2011

Четверкин И. И. (ilia2010@yandex.ru)

Факультет Вычислительной Математики и Кибернетики,
МГУ им. М. В. Ломоносова

Ключевые слова: РОМИП, анализ тональности текстов, оценочные слова, классификация отзывов

TESTING THE SENTIMENT CLASSIFICATION APPROACH IN VARIOUS DOMAINS — ROMIP 2011

Chetviorkin I. I. (ilia2010@yandex.ru)

Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University

We offer a review of sentiment classification experiments in various domains using different training sets. In the movie domain we studied the impact of opinion word weights on the quality of classification. We selected the best feature set and ran them on each task-domain pair. In several tasks our algorithm achieved high quality of the classification.

Key words: ROMIP, sentiment classification, opinion words, domain adaptation

1. Introduction

This year within Russian Information Retrieval Seminar a new sentiment analysis track was offered to the participants. This track had three tasks related to the classification of documents by sentiment expressed in them:

- two-class classification task,
- three-class classification task,

- five-class classification task.

In addition the documents (blog posts) from the test collection were about entities from various domains: books, movies and digital cameras. Each domain requires extra tuning of the algorithms and it can be difficult to achieve a good performance in all domains.

The easiest task is to classify reviews into two classes: *positive* and *negative* [Pang and Lee, 2008]. Quality of two-way classification using the topic-based categorization approach for reviews exceeds 80% [Pang et al., 2002]. In [Whitelaw et al., 2005] the quality of review classification, based on the so-called appraisal taxonomy, is described as 90.2%.

However, when we turn to the problem of review division into three classes, the quality of automatic classification decreases to 75% after an adjustment to an individual author's style, and 66.3% in a case of author independent test collection [Pang and Lee, 2005].

In rating-inference problem with four classes reported accuracy is 54.6% using metric labeling formulation [Pang and Lee, 2005] and 59.2% using graph-based semi-supervised learning algorithm with adjustment to an author style [Goldberg et al., 2006].

Recently we had conducted the similar research for the three-way classification problem in the movie domain [Chetviorkin and Loukachevitch, 2011a]. It was interesting to compare our results with other participants and to try to utilize our approach in the two-class and five-class tasks in various domains.

In the current paper we describe our classification approach using such features as word weights, opinion words and polarity influencers. We have submitted five runs for the three-way classification task in the movie domain and one run (with complete set of features) for all other combinations of tasks and domains.

The reminder of this paper is structured as follows. Section 2 provides a short description of the training collections. Section 3 briefly describes our approach to the sentiment classification. Section 4 gives an overview of our submission results. We provide concluding remarks in Section 5.

2. Data Collections

All participants were granted three train collections, one for each domain (for score distribution in these collections see [Chetviorkin et al., 2012]). But we had created our own collections from the same sources earlier. It was more convenient for us to use our collections in the experiments.

Our movie and book collections (28,773 and 15,113 reviews accordingly) were collected from the online recommendation service *www.imhonet.ru*. Each review in these collections had user's score on a ten-point scale. The digital camera review collection (8,181 reviews) was collected from the Yandex.Market service and had user's score on a five-point scale. Score distributions in there three collections can be found in Fig.1–3.

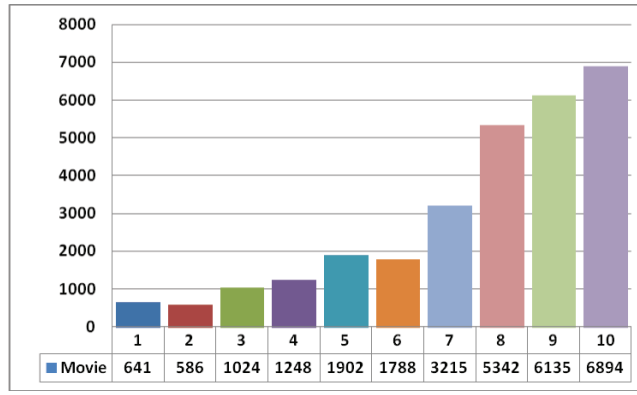

Figure 1. Score distribution in the movie review collection

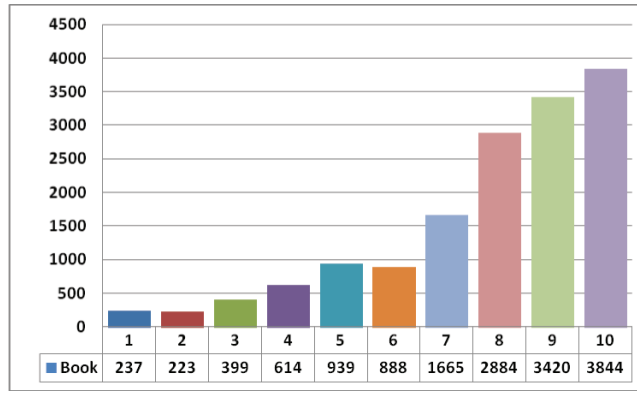

Figure 2. Score distribution in the book review collection

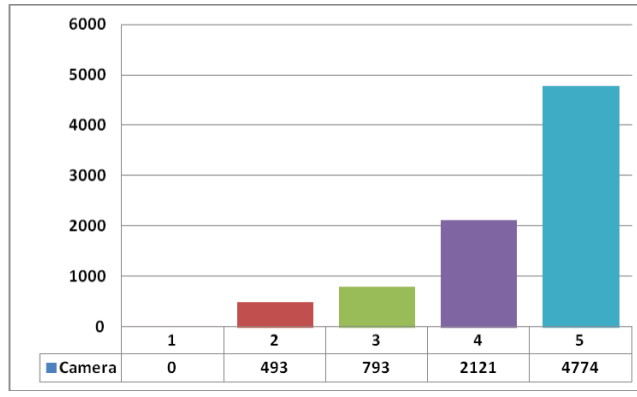

Figure 3. Score distribution in the camera review collection

In addition all participants gained the test collection with 16,821 blog posts about various entities.

3. Sentiment Classification Algorithm

In the sentiment classification track we used the same approach as provided in [Chetviorkin and Loukachevitch, 2011a]. We will shortly describe the main points of our algorithm and major changes, which were applied to it in correspondence with the various tasks and domains.

3.1. Features for review classification

In this research we utilized the best feature combinations which were obtained during the three-way classification experiments in the movie domain [Chetviorkin and Loukachevitch, 2011a]. To improve the quality of the review classification we analyzed the following features:

- word weights based on different collections,
- opinion words,
- use of polarity influencers: they may reverse or enhance (*not, very*) polarity of other words,
- length and structure of reviews,
- use of punctuation marks

The best results were achieved using the bag of words (all words from the train collection with frequencies higher than four), TFIDF word weights, polarity influencers and opinion word weights.

TFIDF

The main elements of our feature set were lemmas, which appeared in the train collection more than three times. The simplest approach for document classification was to create feature vectors using binary weights of words, but not the most effective.

To improve the quality of classification we used TFIDF weights [Ageev et al., 2004] for lemmas with inversed document frequency calculated using the news collection with one million documents.

$$TFIDF(l) = \beta + (1-\beta) \cdot tf(l) \cdot idf(l)$$

$$tf_D(l) = \frac{freq_D(l)}{freq_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{avg\_dl}} \qquad idf(l) = \frac{\log(\frac{|c| + 0.5}{df(l)})}{\log(|c| + 1)}$$

- $freq_D(l)$ — number of occurrences of l in a document D,
- $dl_D(l)$ — length measure of a document D, in our case, it is number of terms in a review,
- *avg_dl* — average length of a document,
- *df(l)* — number of documents in a collection (e. g. description or news collection) where term l appears,
- $\beta = 0.4$,
- $|c|$ — total number of documents in a collection.

Opinion words

Opinion words are the main polarity carriers in a text. We tried to utilize them in various ways in a combination with a bag of words [Chetviorkin and Loukachevitch, 2011a]. Only one useful variant was found: to modify word weights accordingly to opinion word weights in the extraction model.

We used our algorithm [Chetviorkin and Loukachevitch, 2011b] to extract high quality domain dependent opinion words. To generate the list of such words, four text collections were exploited: the review collection about entities from a specific domain, the collection of entity descriptions, the special small corpus and the collection of general news. On the basis of these collections a set of statistical features for words mentioned in reviews was calculated. We trained our model using word feature vectors in the movie domain and then utilized this model in two other domains. As a result we obtained a list of sentiment words for each domain, ordered by the predicted probability of their opinion orientation (opinion weight).

There are examples of opinion words with high probability value in the movie domain:

- *Trogatel'nyi (affecting), otstoi (trash), fignia (crap), otvratitel'no (disgustingly), posredstvenniy (satisfactory), predskazuemyi (predictable), ljubimyj (love) etc.*

In the review classification tasks we modified the weight of each word in the feature vectors as follows:

$$wordweight(x) = TFIDF(x) \cdot e^{opinweight(x)-0.5}$$

Thus, we increased weights of words with high opinion weight, and decreased weights of other words.

Polarity influencers

We used the same set of polarity influencers in all domains:

- operator (–): *net (no), ne (not)*;
- operator (+): *polnyj (full), ochen' (very), sil'no (strongly), takoj (such), prosto (simply), absoljutno (absolutely), nastol'ko (so), samyj (the most)*.

On the basis of this polarity shifter list we substituted sequences “polarity influencer word” using special operator symbols (“+” or “–“) depending on an polarity shifter, for example:

NE HOROSHIJ (NOT GOOD) → –HOROSHIJ (— GOOD)
SAMYJ KRASIVYJ (THE MOST BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)
NASTOL'KO KRASIVYJ (SO BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)

Thus we added to the review vector representation only the operator phrases but not both words. It allowed us to take into account the impact of the polarity influencers.

3.2. Classification algorithm

Authors of previous studies almost unanimously agreed that Support Vector Machine algorithm works better for text classification tasks (and review classification in particular) [Pang and Lee, 2008]. In view of the fact that we had a large amount of data and features (bag of words), library LIBLINEAR was chosen [Fan et al., 2008]. All parameters of the algorithm were left in accordance with their default values.

3.3. Scale mapping

To train our algorithm for classification in a certain scale, we need to map scores from the train collection scale to the task scale. We used the following mapping functions:

- **Two-class task:** {1–7} → “1” (thumbs down), {8–10} → “3” (thumbs up)
- **Three-class task:** {1–6} → “1” (thumbs down), {7–8} → “2” (so-so), {9–10} → “3” (thumbs up)
- **Five-class task:** {1–3} → “1”, {4–5} → “2”, {6–7} → “3”, {8} → “4”, {9–10} → “5”

For the digital camera collection we firstly multiplied each user's score by two and then used aforementioned mapping schemes.

It is rather important to choose a correct mapping function. We investigated the best mapping functions for the three-way classification problem in previous studies [Loukachevitch and Chetviorkin, 2011]. For the two other tasks we used our insights to define the mapping functions.

4. Results Overview

We have submitted five runs for the three-class task in the movie domain:

- Bag of words with TFIDF word weights (**BoW+tfidf**)
- Bag of words with opinion word weights (**BoW+opweight**)
- Bag of words with combination of TFIDF and opinion weights. We took only the first thousand of the most probable opinion words (**BoW+tfidf+opweigh1000**).

- Bag of words with combination of TFIDF and opinion word weights. We took only first ten thousand of the most probable opinion words (**BoW+tfidf+opweight10000**).
- Bag of words with combination of TFIDF and opinion word weights. We took opinion weights for all words from the bag of words (**BoW+tfidf+opweight**).

For all the other pairs of tasks and domains we submitted only one run with **BoW+tfidf+opweight** set of features.

Besides we continued our study of the proposed tasks after the ROMIP deadlines and present our unofficial runs (in italic) in the same tables.

To obtain our first unofficial run 1,393 review duplicates were excluded from our book review collection. On the basis of such collection we obtained slightly better results. We marked such runs with "*nodupl*" postfix in the result tables.

Further we were interested to compare the results of our algorithm trained on the ROMIP data collections with the results of the algorithm trained on our data collections. In this way we retrained the classification model in each domain and evaluated it. These results were marked with *"romip"* postfix in corresponding tables.

4.1. Official metrics

There were a large amount of available metrics for evaluation [Chetviorkin et al., 2012]. To evaluate the performance of our algorithm we used *macro_precision, macro_recall, macro_F-measure, accuracy and average Euclidian distance*.

In addition two evaluation schemes were offered:

- **AND**, in evaluation involved only those reviews, which had the same score from both assessors (only for two-class classification)
- **OR,** we considered the answer of the algorithm as the right one if it matched with at least one of the assessors.

4.2. Three-class task

We started our study of sentiment classification with the three-class classification task. We had the best results in the classification of reviews about digital cameras and movies accordingly to accuracy and macro_F measures. In the book domain our algorithm was the second one accordingly to macro_F and fifth accordingly to the accuracy. The results can be found in Table 3. Our submissions are underlined; the best official results are in bold.

Four out of five of our runs in the movie domain had no statistically significant differences (Wilcoxon signed-rank test/Two-tailed test, $\alpha = 0.05$), and the result of one of them was considerably worse. Thus TFIDF word weights were very important for the quality of the classification but the amount of opinion words had no crucial meaning.

The exclusion of book review duplicates had improved all primary measures. In this case our macro_F result was the best in the book domain. Training on ROMIP collections gave roughly the same results in book and camera domains, but worse results in the movie domain. We discuss these differences in Section 4.5.

Table 1. Three-class classification results (OR)

Run_ID	*Object*	*Macro_Prec*	*Macro_Rec*	*Macro_F*	*Accuracy*
xxx-3	book	0.677	0.532	**0.577**	0.756
xxx-43 tfidf_op	book	0.671	0.517	0.570	0.756
xxx-11	book	0.658	0.475	0.488	**0.771**
Baseline	book	0.227	0.333	0.270	0.68
tfidf_op nodupl	*book*	*0.679*	*0.525*	*0.578*	*0.76*
tfidf_op romip	*book*	*0.664*	*0.510*	*0.571*	*0.76*
yyy-3 tfidf_op	camera	0.843	0.594	**0.663**	**0.841**
yyy-11	camera	0.797	0.596	0.661	0.815
Baseline	camera	0.216	0.333	0.262	0.648
tfidf_op romip	*camera*	*0.804*	*0.598*	*0.658*	*0.837*
zzz-10 tfidf_op	film	0.671	0.535	**0.592**	**0.754**
zzz-19 tfidf_op1000	film	0.657	0.526	0.583	0.754
zzz-9 tfidf_op10000	film	0.660	0.524	0.582	0.751
zzz-1 tfidf	film	0.661	0.524	0.584	0.751
zzz-18 op_weight	film	0.585	0.431	0.494	0.635
Baseline	film	0.235	0.333	0.276	0.705
tfidf_op romip	*film*	*0.582*	*0.425*	*0.487*	*0.629*

4.3. Two-class task

In this task our results were the second by two primary measures in the camera domain (and first after training on the ROMIP collection) and second by macro_F in the movie domain (after training on the ROMIP collection we have lower results, see Section 4.5). In the book domain the results were rather low, but after training on the ROMIP collection the best macro_F result was obtained. The removal of duplicate reviews from the book collection had no effect in this task.

Table 4 shows our results and best two runs for each entity for evaluation schema OR in terms of macro f-measure and accuracy, our runs are underlined and unofficial runs are in italic.

Table 2. Two-class classification results (OR)

Run_ID	*Object*	*Macro_Prec*	*Macro_Rec*	*Macro_F*	*Accuracy*
xxx-40	book	0.714	0.804	**0.747**	0.895
xxx-0	book	0.751	0.721	0.735	0.924
xxx-24 (46)	book	0.968	0.630	0.690	**0.938**
xxx-19	book	0.790	0.651	0.694	0.931
xxx-35 tfidf_op	book	0.682	0.851	0.720	0.851
Baseline	book	0.46	0.5	0.479	0.92
tfidf_op nodupl	*book*	*0.682*	*0.851*	*0.720*	*0.851*
tfidf_op romip	*book*	*0.710*	*0.852*	*0.751*	*0.876*
yyy-24	camera	0.918	0.940	**0.929**	**0.959**
yyy-16 tfidf_op	camera	0.944	0.898	0.919	0.956
Baseline	camera	0.426	0.5	0.46	0.852
tfidf_op romip	*camera*	*0.931*	*0.945*	*0.938*	*0.963*
zzz-23	film	0.776	0.797	**0.786**	**0.881**
zzz-9 tfidf_op	film	0.706	0.794	0.730	0.812
zzz-14	film	0.743	0.597	0.623	0.860
Baseline	film	0.427	0.5	0.461	0.854
tfidf_op romip	*film*	*0.682*	*0.790*	*0.685*	*0.742*

4.4. Five-class task

The five class evaluation scheme is very widespread in the Internet (five stars system), but a five-class sentiment classification is a rather difficult problem because we need not only to determine a text sentiment, but also to show its strength (the rating-inference problem).

Primary measures here were the accuracy and the average Euclidian distance. We achieved the best result accordingly to the accuracy measure in the movie domain and the second result in the book domain. After training on the book collection

without duplicate reviews our algorithm gained the best accuracy result. On the ROMIP book collection the quality dropped significantly (see Section 4.5).

In the digital camera domain our results were quite low. Partly it could be explained by utilization of pros and cons by the other participants and differences in training collections. In our collection there was no strictly negative class (see Section 2).

Table 3. Five-class classification results (OR)

Run_ID	*Object*	*Avg_Eucl_Distance*	*Macro_F*	*Accuracy*
xxx-7	book	**0.872**	0.284	**0.622**
xxx-4 (9)	book	0.892	**0.291**	0.622
xxx-5 tfidf_op	book	0.972	0.270	0.615
Baseline	book	0.909	0.123	0.48
tfidf_op nodupl	*book*	*0.953*	*0.281*	*0.629*
tfidf_op romip	*book*	*1.04*	*0.201*	*0.542*
yyy-1	camera	**0.928**	0.298	0.567
yyy-3	camera	0.940	0.287	0.570
yyy-4	camera	0.971	**0.342**	**0.626**
yyy-2	camera	1.215	0.332	0.626
yyy-9 tfidf_op	camera	1.203	0.193	0.485
Baseline	camera	1.165	0.144	0.563
tfidf_op romip	*camera*	*1.125*	*0.234*	*0.530*
zzz-1 (5)	film	**1.026**	**0.286**	0.599
zzz-1	film	1.071	0.266	0.559
zzz-6 tfidf_op	film	1.133	0.247	**0.602**
Baseline	film	1.460	0.135	0.506
tfidf_op romip	*film*	*1.107*	*0.268*	*0.593*

4.5. The differences between collections

To substantiate the differences between the results obtained by our algorithm trained on different collections in one domain we decided to conduct some additional statistical research.

In the digital camera domain performance of the algorithm trained on our collection was worse than on ROMIP collection. We connect this gap with the differences in the review score distributions. (class "1" frequency, Section 2).

For the book and movie domains we had calculated the share of reviews in each class accordingly to the mapping scheme for a two-class task (for three class and five-class tasks results are the similar) and compared it with assessors' score distribution (OR evaluation scheme). We underlined the distribution that was more similar to the assessors.

Table 4–5. Score distribution in the train collections

Movie	1	2
Our	0.36	0.64
ROMIP	0.43	0.57
Eval	0.19	0.81

Book	1	2
Our	0.33	0.67
ROMIP	0.29	0.71
Eval	0.11	0.89

Thus the score distribution similarity between the train and test collections is highly correlated with the quality of review classification. The size of train collection has low influence on the quality of classification if the score distributions differ significantly.

5. Conclusions

This work is based on our previous research about influence of various features on the three-way review classification quality. In this study we describe the contribution of word weights to the quality of the three-class movie review classification. Then we apply the algorithm with the complete set of features to the other domains and tasks. Our approach demonstrates the good quality of classification in almost all domain-task pairs.

In addition we studied the dependence of the classification quality on the training collection. The similarity of the train and test collection score distributions played here a key role.

Acknowledgements. This work is partially supported by RFBR grant N11-07-00588-a.

References

1. *Ageev M., Dobrov B., Loukachevitch N., Sidorov A.* Experimental algorithms vs. basic line for web ad hoc, legal ad hoc, and legal categorization in RIRES2004 (in Russian). Proceedings of the Russian Information Retrieval Evaluation Seminar. Saint-Petersburg, 2004, pp. 62–89.
2. *Chetviorkin I., Braslavskiy P., Loukachevitch N.,* 2012 Sentiment Analysis Track at ROMIP 2011 (In this volume). Komp'iuternaia Lingvistika i Intellektual'nye

Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2012” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012”]. Bekasovo, 2012.

3. *Chetviorkin I. and Loukachevitch N.* Three-way movie review classification. Komp’iuternaia Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii “Dialog 2011” [Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2011”]. Bekasovo, 2011a, 168–177.
4. *Chetviorkin I. and Loukachevitch N.* Extraction of Domain-specific Opinion Words for Similar Domains. Proceedings of the Workshop on Information Extraction and Knowledge Acquisition (IEKA 2011). Hissar, Bulgaria. 2011b. 7–12.
5. *Goldberg A., Zhu X.* Seeing stars when there aren’t many stars: Graphbased semi-supervised learning for sentiment categorization. HLT-NAACL 2006 Workshop on Textgraphs: Graph-based Algorithms for Natural Language Processing. New York, 2006, pp. 45–52.
6. *Loukachevitch N. V., Chetviorkin I. I.* (2011) Extraction and use of opinion words for the three-way review classification problem (in Russian). Numerical Methods and Programming, Vol. 12, pp. 73–81
7. *Pang B., Lee L.* (2008) Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. Hanover, Massachusetts, Now Publishers.
8. *Pang B., Lee L.* Seeing stars: Exploiting class relationships for sentiment categorization with respect of rating scales. Proceedings of the ACL, 2005. pp. 115–124.
9. *Pang, B., Lee, L., and Vaithyanathan, S.,* Thumbs up? Sentiment Classification Using Machine Learning Techniques. In Proceedings of EMNLP, 2002,
10. *Fan R.-E. , Chang K.-W., Hsieh C.-J., Wang X.-R., and Lin C.-J.* (2008), LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research, Vol. 9. pp. 1871–1874. Software available at http://www.csie.ntu.edu.tw/∼cjlin/liblinear
11. *Whitelaw C., Garg N., Argamon S.*: Using Appraisal Taxonomies for Sentiment Analysis. In: Proceedings of CIKM, Bremen, 2005.