Раздел II.
Доклады, представленные участниками тестирования систем анализа тональности

В данном разделе публикуется итоговая статья организаторов тестирования систем анализа тональности и отдельные статьи участников тестирования. Полностью с комментирующими сообщениями участников можно ознакомиться на сайте конференции «Диалог».

ДОРОЖКИ ПО АНАЛИЗУ МНЕНИЙ НА РОМИП 2011

Четверкин И. И. (ilia2010@yandex.ru),
МГУ им. М. В. Ломоносова

Браславский П. И. (pbraslavski@acm.org),
СКБ Контур, Уральский Федеральный Университет

Лукашевич Н. В. (louk_nat@mail.ru),
НИВЦ МГУ им. М. В. Ломоносова

Ключевые слова: РОМИП, анализ тональности текстов, классификация отзывов, данные из блогов

SENTIMENT ANALYSIS TRACK AT ROMIP 2011

Chetviorkin I. I. (ilia2010@yandex.ru)
Lomonosov Moscow State University

Braslavski P. I. (pbraslavski@acm.org)
Kontur Labs, Ural Federal University

Loukachevitch N. V. (louk_nat@mail.ru)
Research Computing Center of Lomonosov
Moscow State University

Russian Information Retrieval Seminar (ROMIP) is a Russian TREC-like IR evaluation initiative. In 2011 ROMIP launched a new track on sentiment analysis. Within the track we prepared a training collection of user reviews along with ratings for movies, books, and digital cameras. Additionally, we compiled a test collection of blog posts with reviews in the same domains and labeled them according to expressed sentiment. The paper describes the collections' characteristics, track tasks, the labeling process, and evaluation metrics. We summarize the participants' results and make suggestion for future editions of the track.

Key words: ROMIP, sentiment classification, sentiment analysis, opinion mining, blog data

1. Introduction

With the development of internet technologies an increasingly large number of people have got an opportunity to express their opinions on the web. Journal-like web pages (weblogs) allows internet users to share their feelings, emotions and attitudes about various products, services, and real-life events with other people. This information can be very useful both for other web users and for service providers or product manufacturers.

Extremely accessible blog software has facilitated blogging for a wide audience, and, as a result, boosted the growth rate of information available online. Thus, the blogosphere has become a highly dynamic subset of the World Wide Web that evolves responding to real-world events and offers several new research areas.

Today, sentiment analysis research attracts a lot of interest as a tool for opinion processing and company reputation management. Sentiment analysis has a lot of different subtasks [Pang&Lee2008]. The most well-known of them are:

- subjectivity/objectivity identification;
- polarity classification of a given text at the document, sentence, or feature/aspect level;

- advanced, "beyond polarity", sentiment classification that looks, for instance, at emotional states such as "angry," "sad," and "happy";
- recognition of sarcastic sentences (phrases);
- feature/aspect-based sentiment analysis;
- sentiment summarization.

Russian Information Retrieval Seminar (ROMIP, http://romip.ru) is a Russian TREC-like information retrieval evaluation initiative. It was launched in 2002 to increase communication and support research community (both academia and industry) in the area of IR in Russian by providing a basis for independent evaluation of IR methods. Since its start, ROMIP has organized a number of different tracks, e.g. ad hoc retrieval, snippet generation, document classification, question answering (QA), and image retrieval. ROMIP prepared and made available for researchers a number of data collections.

In many respects ROMIP seminars are similar to other international information retrieval events such as TREC and NTCIR, which have already conducted different sentiment analysis tracks (see Section 2). We decided to start with sentiment classification of reviews in Russian because it was quite simple to find data, but the good quality of classification was rather difficult to achieve. On the other hand, we were interested in the state of the art in this research area.

The task of the ROMIP 2011 sentiment analysis track was to classify blog posts about different products according to sentiment expressed in documents. It was reported in the literature that the more classes there are, the harder it is to classify a text by sentiment. Thus, in the first pilot run of the track in 2011 we had three tasks:

- two-class classification task,
- three-class classification task,
- five-class classification task.

It was the first shared task evaluation of document sentiment classification in Russian.

The rest of this paper is structured as follows. In Section 2, we make a brief overview of similar evaluation campaigns and available datasets. Section 3 provides a short description of the newly created collections used for training and evaluation. Section 4 describes the sentiment classification task. Section 5 provides an overview of runs the submitted by participants. Concluding remarks can be found in Section 6.

2. Related evaluation campaigns and datasets

In this section we briefly overview cognate evaluation campaigns within TREC (http://trec.nist.gov) and NTCIR (http://research.nii.ac.jp/ntcir/index-en.html), as well as provide a list of datasets available for research. [Pang&Lee2008] gives a good overview of evaluation initiatives, available data and resources in opinion mining and

sentiment analysis. However, some new datasets and shared tasks emerged after the book had been published.

2.1. TREC

Blog track was organized in 2006–2010 within TREC initiative [Macdonald2010, Ounis2008]. In 2006–2008 the track investigated an opinion-finding task, complemented with a polarity subtask in 2007–2008.

In the opinion-finding task, participating systems had to retrieve opinionated posts about a given target such as person, location or organization, concept (such as type of technology), product name or event. Both *relevance* and *opinionatedness* of retrieved posts were judged. Additionally, polarity of the opinion expressed in relevant posts was labeled as *positive*, *negative*, or *mixed*. This labeling led to a supplemental polarity subtask in two subsequent years. In 2007 the task was formulated as a classification task, i. e. for each retrieved post participants should have predicted its polarity. For TREC 2008, this task was reformulated as a ranking task: only posts expressing polarity should have been retrieved and ranked by the degree of positivity or negativity respectively.

The aforementioned experiments within TREC were performed on the TREC Blogs06 collection. Blogs06 is a collection of over 3.2 million permalinks (i. e. a single blog post and all associated comments) from over 100,000 blogs that had been crawled during an 11-week period from 6th December 2005 until 21st February 2006. To make settings more realistic, a sample of spam blogs, news feeds, as well as non-English documents was injected. (This collection was also used within TAC 2008 Opinion QA Task, http://www.nist.gov/tac/data/past/2008/OpSummQA08.html)

URL: http://ir.dcs.gla.ac.uk/wiki/TREC-BLOG

2.2. NTCIR

NTCIR, a Japanese counterpart of TREC, launched a pilot opinion track in 2006. The dataset was compiled from news articles in Japanese, Chinese, and English. Participants had to solve the following tasks on the sentence level: 1) detection of opinionated sentences, 2) detection of opinion holders, 3) sentence relevance to the topic, and 4) polarity labeling as *positive*, *negative*, or *neutral* [Seki2007]. In NTCR-7 the track evolved into Multilingual Opinion Analysis Track (MOAT); documents in Simplified Chinese and the opinion target identification subtask were added. Moreover, some tasks were performed with finer granularity, i. e. identification was applied to sentence fragments [Seki2008]. In NTCIR-8 the subtasks were extended towards cross-language analysis and question answering: opinionated answers in different languages had to be extracted in response to questions in English [Seki2010].

URL: http://research.nii.ac.jp/ntcir/permission/ntcir-6/perm-en-OPINION.html
http://research.nii.ac.jp/ntcir/permission/ntcir-7/perm-en-MOAT.html

2.3. Data collections

What follows is a non-exhaustive list of datasets not associated with established evaluation campaigns, which can be used for sentiment and opinion analysis. Some of the datasets are no longer available and are mentioned here for reference only. The terms and conditions under which the data are released may vary, so please consult provided URLs.

Cornell Movie Review Datasets contains reviews from IMDb (http://imdb.com). There are 1,000 'polarity reviews' tagged positive or negative, as well as a larger amount of original reviews along with users' star ratings.

URL: http://www.cs.cornell.edu/people/pabo/movie-review-data/

Bing Liu and colleagues compiled several dataset and made them available for researchers in opinion mining and sentiment analysis. The most notable is probably the **Amazon Product Review Dataset** containing 5.8M+ reviews on books, music, DVDs and consumer electronics.

URL: http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

One of the first sizeable blog datasets available for research was **BlogPulse 2005 dataset** released to participants in the Workshop on Weblogging Ecosystem (WWE) in 2006. The dataset contained 10M posts from 1M weblogs collected during three weeks in July 2005.

URL (as preserved in Web Archive): http://web.archive.org/web/20090615025713/http://www.blogpulse.com/www2006-workshop/cfp.html

Several datasets were made available through International Conference on Weblogs and Social Media (http://www.icwsm.org), which continued the tradition from the WWE2006 workshop.

Nielsen BuzzMetrics 2006 Dataset contains 14M weblog posts in XML format from 3M weblogs published in May 2006. The dataset contains posts in different languages (e. g. about 6% of posts are reported to be in Russian).

URL: http://www.icwsm.org/data.html

In the following years much bigger datasets were compiled and released. **ICWSM 2009 Spinn3r Blog Dataset** contains posts made between August 1st and October 1st, 2008 along with some metadata, 44 million blog posts in total. **ICWSM 2011 Spinn3r Dataset** is one magnitude bigger and much more versatile — it covers blog posts, news articles, classifieds, forum posts, and social media content created between January 13th and February 14th 2011, resulting in 386 million items.

URL: http://www.icwsm.org/data/

Content Analysis in Web 2.0 (CAW 2.0) is a dataset associated with a workshop of the same name at the WWW2009 conference. The dataset comprises tweets, forum discussions, comments on news, movie reviews, and on-line chats that total to 680K messages. Workshop organizers offered a number of shared tasks on these data, including opinion and sentiment analysis. The sentiment analysis task was to assign a message to categories *neutral*, *happy*, *angry* or *sad* (fuzzy assignments were allowed); whereas opinion tasks dealt with three categories: *factual*, *opinionated-positive* and *opinionated-negative*.

URL: http://caw2.barcelonamedia.org/node/7

The main task of the **TREC Microblog** track is *ad hoc* retrieval in tweets. However, we envision that the track data collection — 16 million tweets sampled between

January 23rd and February 8th, 2011 — might be employed for sentiment analysis and opinion mining research.

URL: http://trec.nist.gov/data/tweets/

CyberEmotions is an integrating, ongoing, large-scale European research project focusing on the role of collective emotions in creating, forming and breaking-up eCommunities. One of the project outcomes is the creation of a corpus that consists of three parts: 1) 2,5M+ comments from BBC News forum, including 1K+ labeled items; 2) Digg post comments (1.6M+ comments, including 1K+ labeled items); and 3) MySpace comments exchanged between pairs of friends from a total of 100K+ social network members (including 1K+ labeled items).

URL: http://www.cyberemotions.eu/data.html

The **MPQA Opinion Corpus** contains news articles from a wide variety of news sources manually annotated for opinions and other private states (i. e. beliefs, emotions, sentiments, speculations, etc.).

URL: http://www.cs.pitt.edu/mpqa/

The **Multi-Domain Sentiment Dataset** consists of product reviews taken from Amazon.com with many product types (domains). Some domains (books and DVDs) have hundreds of thousands of reviews. Others (musical instruments) have only a few hundred.

URL: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/

3. ROMIP Data Collections

For the sentiment classification tasks we chose three different domains: movies, books, and digital cameras. Movie and book collections (15,718 and 24,159 reviews, respectively) were obtained from online recommendation service IMHONET (http://www.imhonet.ru). Each review in these collections had user's score on a ten-point scale (zero means unmarked). The digital camera review collection (10,370 reviews) was provided by Yandex. Reviews for cameras were collected from the Yandex.Market comparison shopping service (http://market.yandex.ru) and had users' scores on a five-point scale.

The average review length in the movie domain was 72 words, 49 words in the book domain, and 101 words in the camera domain. Score distributions can be found in Fig. 1–3.

These three collections were presented to participants for training their algorithms. No additional information was provided.

To evaluate the quality of sentiment classification algorithms, we needed additional collections without any authors' scores. We decided to collect blog posts about various entities in three domains. For this purpose we used Yandex's Blog Search Engine (http://blog.yandex.ru).

For each domain a list of search queries was manually compiled. There were 61 book queries, 922 camera queries, and 112 movie queries. Each query was about only one entity (or related objects) from selected domains. There is a query example from the book domain: [vpechatlenyia ot kniga “Victor Pelevin” -spisok] *[impression from the book “Victor Pelevin” -list]*.

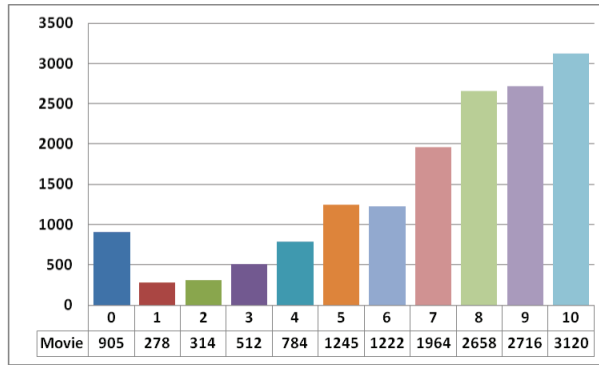

Figure 1. Score distribution in movie review collection

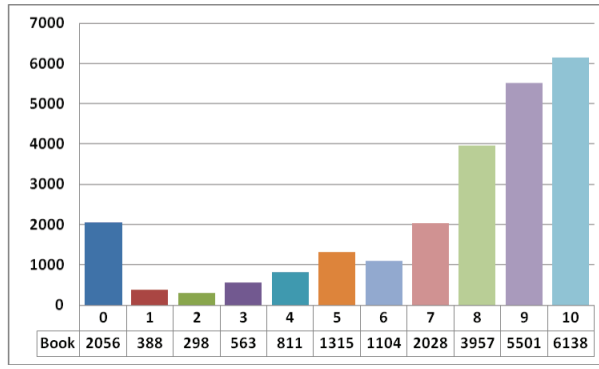

Figure 2. Score distribution in book review collection

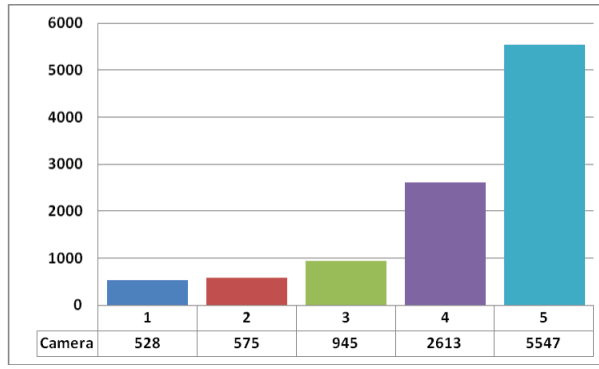

Figure 3. Score distribution in camera review collection

For each query we obtained a set of blog posts (both relevant and irrelevant). Finally results for all queries were merged. The resulting collection included 16,821

reviews for entities from various domains. The average review length in this collection was 1,146 words. Participating systems had to return sentiment labels for all these documents.

4. Assessment Procedure

Test collection included a lot of irrelevant texts, reviews containing sentiment about various topics or texts with both subjective and objective information. Since we wanted to solve only document sentiment classification task we had to select for evaluation only strongly subjective texts with one dominant topic related to entities in the target domains. As a result we selected 275 book reviews, 329 movie reviews, and 270 digital camera reviews for testing.

At the next step, all reviews were labeled by two assessors with three scores (at once) on different scales ***S***:

- ***S*** = ***{1, 2}*** for two-class classification task, where 1 — a negative review and 2 — a positive review;
- ***S*** = ***{1, 2, 3}*** for-three class classification task, where 1 — a generally negative review, 2 — a review has significant positive and negative aspects of the evaluated entity, 3 — a generally positive review;
- ***S*** = ***{1, 2, 3, 4, 5}*** for five-class classification task, where 1 — a generally negative review, 2 — a generally negative, but points to some positive aspects of the entity, 3 — a review has significant positive and negative aspects of the evaluated entity, 4 — a generally positive, but points to some negative aspects of the entity, 5 — a generally positive review.

Class distribution for each task was highly skewed. For example, in the two-class task we had 84% of positive reviews for cameras, 92% of positive reviews for books and 85% of positive reviews for movies. In the three-class and the five-class tasks we had the same situation — the majority of reviews were positive.

In Table 1 one can find Cohen's kappa coefficient for measuring the inter-rater agreement.

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where Pr(a) is the relative observed agreement among raters, and Pr(e) is the hypothetical probability of chance agreement.

Table 1. Kappa coefficients for different tasks

Kappa	2 classes	3 classes	5 classes
Movies	0.818	0.615	0.429
Books	0.812	0.674	0.545
Digital Cameras	0.808	0.602	0.398

Proportion of reviews that were assigned the same score by both assessors for each task-domain pair can be found in Table 2.

Table 2. Proportion of reviews in AND evaluation scheme

	2 classes	3 classes	5 classes
Movies	0.948	0.799	0.590
Books	0.967	0.829	0.684
Digital Cameras	0.944	0.766	0.548

5. Results Overview

In all, twelve groups took part in the sentiment classification task. There were 105 submitted runs in the two-class task, 81 runs in the three-class task, and 30 runs in the five-class task. We used different metrics to evaluate the quality of classification algorithms.

5.1. Official metrics

The metrics used for the opinion classification task were *precision, recall, F1-measure, accuracy and average Euclidian distance*. For the first three measures we used traditional (separately for each category) and macro-averaged variants.

Macro metrics show classification quality for all classes, while traditional metrics evaluate the quality of algorithms only in relation to one specific class. Macro metrics are convenient for multiclass classification tasks to account for imbalanced test data. Since we had highly imbalanced test collection (see Section 4) we used macro-averaged metrics to evaluate the ability of algorithms to determine each of the classes.

To give definition to all these metrics, we assume that:

- $\mathbf{tp}_x$ is the number of objects correctly classified as class X by the algorithm,
- $\mathbf{fp}_x$ is the number of objects falsely classified as class X,
- $\mathbf{fn}_x$ is the number of objects belonging to class X, but classified as non-X by the algorithm,
- $\mathbf{tn}_x$ the number of objects classified to non-X and they actually belong to one of the non-X classes

Table 3. Classifier output types

	actual class	
predicted class	$\mathbf{tp}_x$ (true positive) Correct result	$\mathbf{fp}_x$ (false positive) Unexpected result
	$\mathbf{fn}_x$ (false negative) Missing result	$\mathbf{tn}_x$ (true negative) Correct absence of result

Precision is the proportion of objects classified as X that truly belong to class X. The macro variant of this feature averages all class precision values.

$$P = \frac{tp_x}{tp_x + fp_x}$$

$$Macro\_P = \frac{1}{|S|} \cdot \sum_{x \in S} \frac{tp_x}{tp_x + fp_x}$$

Recall is the proportion of all objects of class X that is classified by the algorithm as X. The macro variant of this feature averages all class recall values.

$$R = \frac{tp_x}{tp_x + fn_x}$$

$$Macro\_R = \frac{1}{|S|} \cdot \sum_{x \in S} \frac{tp_x}{tp_x + fn_x}$$

F1-measure is the harmonic mean of Precision and Recall. Macro_F1 is the average from all F1-measures of particular classes.

$$Fmeasure = \frac{2 \cdot P \cdot R}{P + R}$$

Accuracy is proportion of correctly classified objects in all objects processed by the algorithm.

$$Accuracy = \frac{tp_x + tn_x}{tp_x + tn_x + fp_x + fn_x}$$

Average Euclidean distance is the average from the quadratic difference between the scores of the algorithm and the assessor scores (average of the assessors' scores).

$$D = \sqrt{\frac{\sum_{i=1}^{n} (q_i - p_i)^2}{n}}$$

5.2. Participants' results

For each task we calculated baseline values for all measures. We took as the baseline a dummy classifier that assigns all reviews to the most frequent class. For this reason, the maximum value for all macro metrics was equal to one divided by the number of classes in the task, which was rather low in comparison with participants' runs. On the other hand, the accuracy and average Euclidian distance were very close to the best results.

In addition, two evaluation schemes were applied:

- **AND**, only those reviews that have the same score from both assessors were involved in evaluation (see Section 4)
- **OR,** we considered an answer of the algorithm to be the right one if it matched with the answer of at least one assessor

In addition, it was important to determine if the difference (according to task's primary measures) between the best runs was statistically significant. For this purpose the Wilcoxon signed-rank test/Two-tailed test ($\alpha = 0.05$) was used. We marked the top result with "*" in case of insignificant difference with the second result.

Two-class task

Primary measures for evaluating the two-class classification performance were macro-F1 and accuracy. Table 4 shows the best two runs for each type of entities for evaluation scheme OR in terms of macro F1-measure and accuracy. Table 5 shows similar results for evaluation scheme AND.

Table 4. Two-class classification results (OR)

Run_ID	*Object*	*Macro_P*	*Macro_R*	*Macro_F1*	*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
Baseline	book	0.460	0.500	0.479	0.920
yyy-24	camera	0.918	0.940	**0.929***	**0.959***
yyy-16	camera	0.944	0.898	0.919	0.956
Baseline	camera	0.426	0.500	0.460	0.852
zzz-23	film	0.776	0.797	**0.786**	**0.881**
zzz-9	film	0.706	0.794	0.730	0.812
zzz-14	film	0.743	0.597	0.623	0.860
Baseline	film	0.427	0.500	0.461	0.854

Table 5. Two-class classification results (AND)

Run_ID	*Object*	*Macro_P*	*Macro_R*	*Macro_F1*	*Accuracy*
xxx-34	book	0.698	0.761	**0.723**	0.902
xxx-0	book	0.739	0.709	0.723	0.921
xxx-24 (46)	book	0.967	0.614	0.668	**0.936***
xxx-19	book	0.789	0.651	0.693	0.929
Baseline	book	0.459	0.500	0.478	0.917

Run_ID	*Object*	*Macro_P*	*Macro_R*	*Macro_F1*	*Accuracy*
yyy-24	camera	0.909	0.934	**0.921***	**0.957***
yyy-16	camera	0.936	0.881	0.905	0.953
yyy-9	camera	0.890	0.929	0.908	0.949
Baseline	camera	0.422	0.500	0.457	0.843
zzz-23	film	0.760	0.781	**0.770**	**0.875**
zzz-9	film	0.680	0.772	0.702	0.801
zzz-14	film	0.715	0.580	0.600	0.853
Baseline	film	0.423	0.500	0.458	0.846

Results in these two evaluation scheme s are highly correlated. For schema AND, the results are slightly worse, because all reviews with ambiguous scores were excluded (any algorithm answer was correct in the OR scheme). For the three-class and the five-class tasks we give results only for OR.

According to the results, reviews in different domains have different complexity. Traditionally, [Turney2002] the movie domain is the most difficult one (in accordance with accuracy).

All best runs have outperformed the baseline, but not all participants did.

Three-class task

In this task, primary measures were the same as in the previous task: macro F1-measure and accuracy. Table 6 shows the two best results for each object. The results and baselines drop significantly in comparison with the two-class task.

Table 6. Three-class classification results (OR)

Run_ID	*Object*	*Macro_P*	*Macro_R*	*Macro_F1*	*Accuracy*
xxx-3	book	0.677	0.532	**0.577***	0.756
xxx-43	book	0.671	0.517	0.570	0.756
xxx-11	book	0.658	0.475	0.488	**0.771**
xxx-36	book	0.625	0.481	0.499	0.764
Baseline	book	0.227	0.333	0.270	0.68
yyy-3	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
zzz-10	film	0.671	0.535	**0.592***	**0.754***
zzz-1	film	0.661	0.524	0.584	0.751
zzz-19	film	0.657	0.526	0.582	0.754
Baseline	film	0.235	0.333	0.276	0.705

Classifying camera reviews seems to be easier than classifying reviews from the other domains.

Five-class task

The five-class classification task differs significantly from previous tasks. Even though such evaluation scheme is very common on the internet ("five stars" system), it is a quite difficult task because not only does one need to determine the text's sentiment, but it is also necessary to find its strength (rating-inference problem). Even assessors' agreement in five-class labeling is much lower than it is in other tasks.

Accuracy and average Euclidian distance were the primary measures for this task. Firstly, it was important to know what percentage of reviews was classified correctly, secondly, what was the average score deviation from assessors' scores.

Table 7. Five-class classification results (OR)

Run_ID	*Object*	*Avg_Eucl_Distance*	*Macro_F1*	*Accuracy*
xxx-7	book	**0.872***	0.284	**0.622***
xxx-4 (9)	book	0.892	**0.291***	0.622
xxx-5	book	0.972	0.270	0.615
Baseline	book	0.909	0.123	0.48
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
Baseline	camera	1.165	0.144	0.563
zzz-1 (5)	film	**1.026***	**0.286***	0.599
zzz-2	film	1.071	0.266	0.559
zzz-6	film	1.133	0.247	**0.602**
Baseline	film	1.460	0.135	0.506

In all domains F1-measure is very low. In comparison to the accuracy level it means that it is difficult for the algorithms to classify reviews from minority classes.

6. Conclusions

ROMIP 2011 was the first shared task evaluation of text sentiment classification in Russian. New collections in different domains (movies, books, digital cameras) were created and made available for research. We thought that sentiment classification was rather a challenging task and it was important to know the state of art for Russian language.

In each task/domain pair the best runs show quite high performance despite highly unbalanced test collection. Based on these results we can conclude that each domain has different complexity and each of them requires an additional adaptation of the algorithms.

We discovered that the interest in sentiment analysis of Russian texts was very high among researchers and specialists in natural language processing. Results in each task coincide with the results for other languages described in literature. At ROMIP 2012 we are planning to offer two new tasks: subjectivity\objectivity identification task and detection of review's domain.

Instructions of how to obtain any of ROMIP collections can be found at http://romip.ru/ru/participation.

Acknowledgements. We are grateful to Yandex and IMHONET for granting their review data collections for research purposes of the seminar. We thank Marina Nekrestyanova, Maxim Gubin and Boris Dobrov for many valuable comments and help with ROMIP organization. We also thank Alya Ageeva for proofreading the paper. This work is partially supported by RFBR grant N11-07-00588-a.

References

1. *Macdonald C., Santos R. L., Ounis I., and Soboroff I.* (2010) Blog track research at TREC. SIGIR Forum 44(1), pp 58–75.
2. *Ounis I., Macdonald C. and Soboroff I.* (2008) On the TREC Blog Track. In Proceedings of International Conference on Weblogs and Social Media (ICWSM 2008).
3. *Pang B., Lee L.*: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. Now Publishers, 2008.
4. *Seki Y., Evans D. K., Ku L. W., Chen H. H., Kando N. and Lin C. Y.* (2007) Overview of Opinion Analysis Pilot Task at NTCIR-6. In Proc. of the Sixth NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pp 265–278.
5. *Seki Y., Evans D. K., Ku L. W., Sun L., Chen H. H. and Kando N.* (2008) Overview of Multilingual Opinion Analysis Task at NTCIR-7. In Proc. of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pp 185–203.
6. *Seki Y., Ku L. W., Sun L., Chen H. H. and Kando N.* (2010) Overview of Multilingual Opinion Analysis Task at NTCIR-8 — A Step Toward Cross Lingual Opinion Analysis. In Proc. of the Eights NTCIR Workshop Meeting on Evaluation of Information Access Technologies, pp 209–220
7. *Turney P. D.* (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Procs. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02). pp. 417–424.