ТЕСТИРОВАНИЕ СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ НА СЕМИНАРЕ РОМИП-2012

Четверкин И. И. (ilia2010@yandex.ru),
Лукашевич Н. В. (louk_nat@mail.ru)

МГУ им. М. В. Ломоносова, Москва, Россия

Ключевые слова: РОМИП, извлечение отзывов из блогов, классификация отзывов, анализ тональности новостей

SENTIMENT ANALYSIS TRACK AT ROMIP 2012

Chetviorkin I. I. (ilia2010@yandex.ru),
Loukachevitch N. V. (louk_nat@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

In 2012, Russian Information Retrieval Seminar (ROMIP) continued the investigation of sentiment analysis issues. Along with the last year's tasks on sentiment classification of user reviews we proposed two new tasks on sentiment classification of news-based opinions and query-based extraction of opinionated blog posts. For all tasks new test collections were prepared. The paper describes the characteristics of the collections, track tasks, the labeling process, and evaluation metrics. We summarize the participants' results and describe our simple approach for sentiment extraction task.

Keywords: ROMIP, sentiment classification, news-based sentiment analysis, opinion mining

1. Introduction

Recently, the sentiment analysis task received a considerable interest from the research community and industry due to the large amount of sentiment-oriented data in social media and user-generated content. The increased interest in solving the problem of sentiment analysis in social media has led to the rapid development of online reputation management systems, where political parties or companies follow the user comments to reveal the opinion trends and the trends of positive and negative comments. Other applications based on social media analytics intend to reveal new social trends in a region or a social group.

Applications dealing with sentiment analysis for social media require a combination of many different techniques for processing unstructured text data [Bing, 2011; Taboada et al., 2011], e.g. sentiment analysis (including sarcasm detection), opinion mining, information retrieval, classification, summarization, etc.

During the Russian Information Retrieval Seminar (ROMIP, http://romip.ru) cycle in 2012, the second open evaluation of sentiment analysis systems takes place. The tasks of the ROMIP 2012 were closely connected to the social media analytics and consist of:

1. Query-based extraction of opinionated blog posts,
2. Sentiment classification of news-based opinions. News-based opinions are fragments of direct or indirect speech extracted from news articles.
3. Sentiment classification of user reviews.

The first task was very similar to the TREC Blog Track 2006 [Ounis et al., 2007]. Here participants had to find all relevant opinionated posts from the blog collection according to a specific query.

The second and the third tasks had the same objective: to classify texts according to sentiment expressed in them. The main difference was in the domains of texts. Sentiment classification of the news-based opinions differs significantly from classification of user reviews and can be considered as the first step to deep sentiment analysis of news articles.

The last task concerned the sentiment classification of blog posts about different products. There were three different scales in this task:

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

The rest of this paper is structured as follows. In Sections 2, 3 and 4 we provide a short description of each task and newly created collections used for training and evaluation. Section 5 provides an overview of runs submitted by participants. Concluding remarks can be found in Section 5.

2. Query-Based Sentiment Extraction

This task was a new one for social media analytics in Russian. The main objective was to find opinionated blog posts relevant to a specific query. Figure 1 depicts query results for the digital camera *Canon EOS 6D* with highlighted relevant posts.

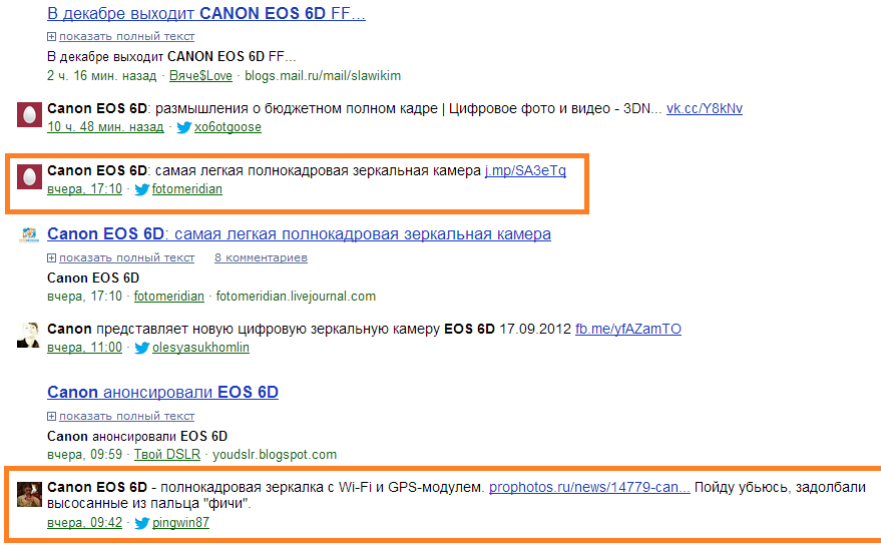

Fig. 1. Query results with highlighted opinion posts

There were three domains: books, digital cameras and movies. For the purposes of query-based extraction two new datasets were released.

The training dataset consists of 874 blog posts about various products (movies, books, digital cameras) with sentiment scores and the list of objects mentioned in this post in some opinionated context. This collection was created from the test set of sentiment classification task during the ROMIP 2011.

To evaluate the quality of sentiment classification and extraction algorithms, we needed additional collections without any authors' scores. We decided to collect blog posts about various entities in three domains (as in ROMIP 2011). 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 2,713 book queries, 1,412 camera queries, and 281 movie queries. Each query was about only one entity (or related objects) from selected domains.

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 60,737 posts for entities from various domains.

From this test collection we selected a set of blog posts for human evaluation, which corresponds to randomly selected set of queries: 221 book queries, 235 movie queries and 301 queries about digital cameras.

The task for assessor was the following: for each document-query pair to decide if the document is relevant to a specific query and what sentiment is expressed about the object in the query. In situations where a blog post describes several different objects or some object which is not mentioned in the query, the assessor should mark this document as relevant to the mentioned objects.

In addition assessor was asked to put score on 2, 3 and 5 point scale for each document containing sentiment. Such document would be used in sentiment classification

task. The resulting markup for each document consists of objects mentioned in this document and sentiment scores (on three scales) associated with each object. This year we have only one assessor, but in general framework for sentiment classification is the same as in [Chetviorkin et al., 2012] where the level of annotators' agreement can be found.

The example of the evaluated blog post: *"Девушка с татуировкой дракона" — фильм крутой, вы чего. Недавно американосами был экранизирован, правда шведские книга и фильм круче..*

```
<object main="+">
  Девушка с татуировкой дракона
  <type>F</type>
  <evaluation-2>2 </evaluation-2>
  <evaluation-3>3</evaluation-3>
  <evaluation-5>5</evaluation-5>
</object>
```

3. Sentiment Classification of User Reviews

This task was similar to one from ROMIP 2011. Here the aim was to classify blog posts about different products according to sentiment expressed in documents. We consider different number of classes for classification: two, three and five.

For the sentiment classification tasks we used the same train collections as in the ROMIP 2011 sentiment analysis track [Chetviorkin et al., 2012]. There were three collections: movie and book collections with 15,718 and 24,159 reviews respectively and the digital camera review collection with 10,370 reviews. All reviews have an author's score on a ten-point scale or a five-point scale.

For testing purposes we selected all opinionated blog posts (see Section 2) from the markup which were annotated during the preparation to query-based sentiment extraction task. We obtained 408 sentiment posts about movies, 129 posts about books and 411 posts about digital cameras.

The class distribution for each task was highly skewed. For example, in the two-class task we had 96% of positive reviews for cameras, 87% of positive reviews for books and 81% of positive reviews for movies.

4. News-Based Opinions Classification

This task was new for ROMIP, and it served as the first step for sentiment analysis of whole news articles. Participants should provide sentiment classification of opinions in form of direct or indirect speech extracted from news articles. For each fragment a participant's system should classify it to one of three classes:

1. Opinion expressed in the news fragment is explicitly negative,
2. Opinion expressed in the news fragment is explicitly positive,
3. The news fragment does not contain any opinion.

We prepared a new training set for sentiment classification of direct and indirect speech from news articles, containing 4,260 text fragments. The test collection for the news-based opinion classification task has the same structure as the training set. The main difference between these collections is that test dataset was collected during the other period of time. It contains 124,647 direct and indirect speech fragments from news articles. From whole bunch of text fragments there were evaluated 5,500 quotes for testing purposes.

The example of direct speech is: *"Посредством этих структур десяткам тысяч избирателей предлагают деньги в обмен на паспортные данные и подписи за какого-либо кандидата", — сказал Черненко.*

5. 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. In query-based sentiment extraction we used two additional measures *Precision@n*, *NDCG@n*.

To give definition to the first part of these metrics, we will use Table 1.

Table 1. Classifier output types

	actual class	
predicted class	**tp_x** (true positive) Correct result	**fp_x** (false positive) Unexpected result
	fn_x (false negative) Missing result	**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).

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

In the query-based sentiment extraction we have the ordered list of answers for each query, and the objective was to place all relevant blog posts as close to the beginning of the answer list as possible. Because of different from sentiment classification objective function, we used the other metrics for this task.

Precision@n indicates the number of correct (relevant) objects in the first ***n*** objects in the result set. We assume that *rel(i)* is above zero (e.g. equals to one) in case of relevance of document in position *i* to the query and zero otherwise.

$$P@n = \sum_{i=1}^{n} rel(i)$$

NDCG@n measures the usefulness, or gain, of a document based on its position in the result list, where *IDCG@n* is *DCG@n* of perfect ranking algorithm.

$$NDCG@n = \frac{DCG@n}{IDCG@n} \qquad DCG@n = rel(1) + \sum_{i=2}^{n} \frac{rel(i)}{\log_2(i)}$$

6. Results Overview

In all, sixteen groups took part in five tasks. In the review classification task there were 94 submitted runs in the two-class task, 46 runs in the three-class task, and 15 runs in the five-class task. In news-based opinion classification there were 16 runs and only two participants were in the query-based sentiment extraction with 33 runs.

For each classification 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.

6.1. Review classification task

Primary measures for evaluating performance in review classification were macro-F1 and accuracy. Table 2–4 shows the best two runs for all tasks. Due to skewness of class distribution in the test collection in some tasks it was difficult to beat the baselines.

Table 2. Two-class classification results

Run_ID	Object	Macro_P	Macro_R	Macro_F1	Accuracy
xxx-17	book	0.749	0.684	0.715	0.884
xxx-1	book	0.666	0.748	0.705	0.821
Baseline	book	0.434	0.500	0.465	0.868
yyy-12	camera	0.589	0.734	0.669	0.895
yyy-13	camera	0.688	0.635	0.660	0.961
Baseline	camera	0.483	0.500	0.491	0.966
zzz-19	film	0.695	0.719	0.707	0.806
zzz-23	film	0.731	0.641	0.683	0.831
zzz-12	film	0.759	0.586	0.661	0.828
Baseline	film	0.404	0.500	0.447	0.809

Table 3. Three-class classification results

Run_ID	Object	Macro_P	Macro_R	Macro_F1	Accuracy
xxx-10	book	0.532	0.591	0.560	0.659
xxx-17	book	0.544	0.554	0.550	0.698
xxx-13	book	0.505	0.532	0.518	0.752
xxx-7	book	0.471	0.501	0.486	0.729
Baseline	book	0.258	0.333	0.291	0.775
yyy-12	camera	0.399	0.602	0.480	0.742
yyy-1	camera	0.440	0.498	0.467	0.523
Baseline	camera	0.285	0.333	0.307	0.854
zzz-11	film	0.569	0.479	0.520	0.694
zzz-6	film	0.486	0.521	0.503	0.596
zzz-1	film	0.487	0.451	0.468	0.650
Baseline	film	0.217	0.333	0.263	0.651

Table 4. Five-class classification results

Run_ID	Object	Avg_Eucl_Distance	Macro_F1	Accuracy
xxx-1	book	1.341	0.402	0.480
xxx-4	book	1.121	0.384	0.473
Baseline	book	1.180	0.131	0.488
yyy-3	camera	1.163	0.336	0.457
yyy-1	camera	1.127	0.288	0.489
yyy-4	camera	1.068	0.207	0.513
yyy-0	camera	1.005	0.245	0.494
Baseline	camera	0.992	0.134	0.504
zzz-2	film	1.388	0.377	0.407
zzz-1	film	1.387	0.323	0.385
Baseline	film	1.720	0.097	0.319

In the review classification task practically all the best results were obtained with machine learning approaches. The best results in the sentiment classification according to F1-measure were obtained by [Blinov et al., 2013] using machine learning approaches on base of SVM and MaxEnt classifiers. The features for classification were semi-automatically crafted on base of the sentiment lexicon from [Chetviorkin & Loukachevitch, 2012] and augmented by collocations with particles and adverbs. Additionally, authors took into account the weighting scheme, the fraction of positive and negative words in texts, exclamation and question marks, emoticons and obscene language. Finally, only the five class classification was conducted and then simple mapping scheme was applied to obtain two or three classes depending on the task.

In [Frolov et al., 2013] the authors use the semantics graph to complement the feature representation for machine learning and make extensive analysis of difficulties occurred during the sentiment classification of book reviews.

In paper [Panicheva, 2013] the rule-based approach using the syntactic structure and an opinion word dictionary is described. The authors obtained the best result according to F1-measure in the two-class movie review classification task. The other rule-based approach is described in [Mavljutov & Ostapuk, 2013]. The authors used the syntactic parser based on context-free grammar and text mining techniques for dictionary construction including objects, proper names, object parts and opinion expressions.

6.2. News-based opinion classification

In this task class distribution was rather balanced in comparison with the review classification task: 41% of quotes were negative, 32% of quotes were positive and 27% of quotes were neutral. Thus the majority of participants performed better than the baseline but the overall quality is still mediocre. The best results according to accuracy and F1-measure could be found in Table 5.

Table 5. News-based opinion classification results

Run_ID	Macro_P	Macro_R	Macro_F1	Accuracy
xxx-4	0.626	0.616	0.621	0.616
xxx-11	0.606	0.579	0.592	0.571
xxx-15	0.563	0.560	0.562	0.582
Baseline	0.138	0.333	0.195	0.413

In opposite to review classification the leaders in the news-based task were knowledge-based approaches. It is due to the absence of a large training collection appropriate for this task because of the broad scope of quotation topics.

The best results in this task were obtained using the lexicon-based system described in [Kuznetsova et al. 2013]. The system has an extensive dictionary of opinion words and expressions obtained using various text mining techniques and manual refinement. Several rules taking into account intensifiers, negation and consequent opinion words were also applied.

The rule-based approach is described in [Panicheva 2013]. The authors used the same system both for sentiment review classification and news-based opinion classification. The system has an extensive rule set and manually crafted sentiment lexicon. The results of this system were second and third in news-based opinion classification.

6.3. Query-based sentiment extraction

In the query-based sentiment extraction task only one participant submitted his result before the deadline. To conduct the track we built our own very simple approach on base of TFIDF measure from [Ageev et al., 2004], which performs at the high level on the standard ad-hoc search task and the five-thousand opinion word list presented in [Chetviorkin & Loukachevitch, 2012].

This sentiment lexicon was constructed in several stages by building the supervised algorithm for sentiment lexicon extraction in the movie domain and further transfer of the model to other domains. The trained sentiment lexicon extraction model was applied to an extensive number of domains and then extracted lexicons were summed up to the single list of sentiment words. This lexicon is proved to be rather clean (P@1000 = 91.4%) to be used is various sentiment analysis tasks and is freely available on the ROMIP web site[1].

To find opinionated blog posts we build two inverted indexes with TFIDF values for all frequent lemmas using posts and headers from the full blog collection. IDF values for all words were calculated using full blog test collection. The third index was built using the aforementioned sentiment word list. For each post in the collection we calculated the fraction of opinion words in it. This fraction serves as opinion weight of each document in the third index.

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

Finally, for each query we calculated weights of all documents in the collection in accordance with the following formula:

$$Weight = \alpha \cdot (\sum_{w \in q} tfidf_w + \sum_{w \in q} tfidf_w^{header}) + (1 - \alpha) \cdot SentiWeight$$

We have experimented with different values of $\alpha = \{0.2, 0.4, 0.5, 0.6, 0.8\}$. The best result was obtained with $\alpha = 0.6$. This result shows the importance of sentiment words in the task of query-based sentiment extraction. All the best results in the resulting Table 6 were obtained using aforementioned approach.

We tried to evaluate the participant results dealing with unlabeled documents as with irrelevant, but it leaded to serious underestimation of the performance. Thus we decided to use only labeled documents, excluding all other documents from the results preserving the order of the remaining documents. The main measures of the performance in this task were NDCG@10 and P@10.

Table 6. Query-based sentiment extraction results

Run_ID	Object	P@1	P@5	P@10	NDCG@10
xxx-0	book	0.3	0.32	0.286	0.305
xxx-9	book	0.3	0.31	0.323	0.304
xxx-8	book	0.25	0.31	0.332	0.298
xxx-6	book	0.25	0.31	0.327	0.302
yyy-9	camera	0.402	0.313	0.302	0.305
yyy-7	camera	0.427	0.319	0.300	0.303
yyy-1	camera	0.402	0.328	0.325	0.226
yyy-2	camera	0.440	0.325	0.311	0.303
zzz-3	film	0.494	0.449	0.438	0.338
zzz-8	film	0. 494	0.448	0.444	0.332

7. Conclusions

ROMIP 2012 is the second seminar which is dedicated to the sentiment analysis problems. In this year we continued the investigation of sentiment analysis tasks, and the list of such tasks was substantially supplemented. Several new collections were created and made available for the research purposes.

The results of this year showed that the sentiment analysis task are still very challenging and attract a lot of researchers from industrial companies and academia.

We find that sentiment classification results are consistent with the results of ROMIP 2011. In query-based sentiment extraction task we found a big role of sentiment lexicons, which is comparable to the role of underlying topic relevance task.

Acknowledgements. We are grateful to Yandex and Anton Pavlov in particular for help with collecting data for research purposes of the seminar. 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 Proceedings of RIRES, 2004, (in Russian)
2. *Blinov P., Klekovkina M., Kotelnikov E, Pestov O.* Research of lexical approach and machine learning methods for sentiment analysis. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
3. *Bing L.* Sentiment Analysis Tutorial, AAAI, San Francisco, USA, 2011
4. *Chetviorkin I., Braslavskiy P., Loukachevitch N.* Sentiment Analysis Track at ROMIP 2011 In Proceedings of Dialog, Bekasovo, 2012, pp. 1–14.
5. *Chetviorkin I.* and Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain In Proceedings of COLING 2012*,* Mumbai, India, 2012, pp. 593–610
6. *Frolov A., Polyakov P., Pleshko V.* Using semantics categories in application to book reviews sentiment analysis. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
7. *Kuznetsova E. S., Loukachevitch N. V., Chetviorkin I. I.* Testing rules for sentiment analysis system. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
8. *Mavljutov R., Ostapuk N.* Using basic syntactic relations for sentiment analysis. Computational Linguistics and Intellectual Technologies. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
9. *Ounis I., de Rijke M., Macdonald C., Mishne G., Soboroff I.* Overview of TREC-2006 Blog track. In Proceedings of TREC-2006*,* Gaithersburg, USA, 2007.
10. *Panicheva P.* Atex. A rule-based sentiment analysis system. Processing texts in various topics. Computational Linguistics and Intellectual Technologies. In Proceedings of Dialog, Bekasovo, 2013, (In Russian)
11. *Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M.* Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), 2011, pp. 267–307.