Evaluating Sentiment Analysis Systems in Russian

Ilia Chetviorkin
Faculty of Computational
Mathematics and Cybernetics
Lomonosov Moscow State University
Moscow, Leninskiye Gory 1, Building 52
ilia2010@yandex.ru

Natalia Loukachevitch
Research Computing Center
Lomonosov Moscow State University
Moscow, Leninskiye Gory 1, Building 4
louk_nat@mail.ru

Abstract

In this paper we describe our experience in conducting the first open sentiment analysis evaluations in Russian in 2011-2012. These initiatives took part within Russian Information Retrieval Seminar (ROMIP), which is an annual TREC-like competition in Russian. Several test and train collections were created for such tasks as sentiment classification in blogs and newswire, opinion retrieval. The paper describes the state of the art in sentiment analysis in Russian, collection characteristics, track tasks and evaluation metrics.

1 Introduction

Sentiment analysis of natural language texts is one of the fast-developing technologies of natural language processing. Many lexical resources and tools were created for sentiment analysis in English. But lately a lot of research work was initiated for sentiment analysis in other languages (Mihalcea et al., 2007; Abdul-Mageed et al., 2011; Pérez-Rosas et al., 2012).

The development of sentiment analysis in Russian previously did not attract a lot of attention at international conferences. Besides, until recently, the interest to sentiment analysis within Russia was connected only with election campaigns. But now there is a considerable interest to sentiment analysis within Russia both from the research community and from the industry.

Therefore during the last years, two workshops on the evaluation of sentiment analysis systems were organized within the framework of Russian Information Retrieval Seminar ROMIP[1] . 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. Besides, there are various shared tasks connected to the sentiment analysis like (Morante and Blanco, 2012; Pestian et al., 2012; Wu and Jin, 2010; Amigó et al., 2012).

In this paper we partly overview the sentiment analysis tasks proposed at ROMIP-2011 (Chetviorkin et al., 2012) and ROMIP-2012 (Chetviorkin and Loukachevich, 2013), the data prepared for evaluation (and therefore available for other interested researchers), and the results obtained by participants. In addition we summarize the results of two initiatives, compare them with the state of the art in English and describe some interesting issues connected to news-based sentiment analysis. We justify all our decisions about the conducted tracks based on the experience of the other researchers, who made the similar initiatives in English. ROMIP-2011 and ROMIP-2012 are unique events for Slavic languages and other European languages different from English.

The paper is structured as follows. In section 2 we review papers on Russian sentiment analysis, not related to the ROMIP evaluations. In section 3 we consider sentiment analysis evaluation tasks proposed during ROMIP-2011, 2012 and consider the main results obtained by participants.

2 Sentiment Analysis in Russian

In Russia studies devoted to sentiment analysis in Russian before 2011 are not very numerous.

In (Ermakov, 2009) a sentiment analysis system extracting opinions about cars from a Russian blog community (http://avto-ru.livejournal.com/) is presented. The approach is based on the detailed description of knowledge about car trade marks, their details and characteristics, semantic patterns of sentiment expressions. This paper is the first, to our knowledge, paper in Russia that reports evaluation results of the proposed approach: precision 84%, recall 20% (self-evaluation).

[1] http://romip.ru/en/index.html

Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, pages 12–17,
Sofia, Bulgaria, 8-9 August 2013. ©2010 Association for Computational Linguistics

In international research Russian sentiment analysis appears mainly in multilingual experiments.

In (Zagibalov et al., 2010) comparable corpora of reviews related to the same books in English and in Russian are described. These corpora allowed authors to study specific ways of sentiment expression in Russian and English.

In (Steinberger et al., 2011) construction of general sentiment vocabularies for several languages is described. They create two source sentiment vocabularies: English (2400 entries) and Spanish (1737 entries). Both lists are translated by Google translator to the target language. Only the overlapping entries from each translation are taken into further consideration. The set of target languages comprises six languages including Russian. The extracted Russian list of sentiment words contained 966 entries with accuracy of 94.9%.

In one of the recent papers not related to the ROMIP evaluations (Chetviorkin and Loukachevitch, 2012), the generation of the Russian sentiment vocabulary for the generalized domain of products and services is described. Authors constructed a new model based on multiple features for domain-specific sentiment vocabulary extraction, then applied this model to several domains, and at last combined these domain-specific vocabularies to generate Russian sentiment vocabulary for products and services – ProductSentiRus. Now the extracted list is publicly available[2].

3 Sentiment analysis tasks

The tasks of two Russian sentiment analysis evaluations ROMIP-2011 and ROMIP-2012 included:

- Sentiment classification of user reviews in three domains (movies, books, digital cameras) using several different sentiment scales,
- Sentiment classification of news-based opinions, which are fragments of direct or indirect speech extracted from news articles,
- Query-based retrieval of opinionated blog posts in three domains (movies, books, digital cameras).

In ROMIP-2011 sentiment evaluation there were 12 participants with more than 200 runs. In ROMIP-2012 17 teams sent more than 150 runs. The presentations describing approaches were organized as a section of International Conference on Computational Linguistics and Information Technologies ”Dialog” (www.dialog-21.ru/en/).

3.1 Sentiment classification of reviews

The only task of ROMIP-2011 and one of the tasks of ROMIP-2012 was sentiment classification of users reviews in three domains: movies, books and digital cameras.

The training data for this task included movie and book collections with 15,718 and 24,159 reviews respectively from Imhonet service (imhonet.ru) and the digital camera review collection with 10,370 reviews from Yandex Market service (http://market.yandex.ru/). All reviews have the authors score on the ten-point scale or the five-point scale.

For testing, another collection of reviews without any authors' scores was created. The testing collection contained blog posts about the above-mentioned entities found with Yandex's Blog Search Engine (http://blog.yandex.ru). So in this track we tried to model a real-word task, when a classifier should be trained on available data, which can be quite different from the task data. The participants stressed that our track is more difficult than training and testing on the similar data, but agreed that this task setting is more realistic.

For each domain a list of search queries was manually compiled and for each query a set of blog posts was extracted. Finally, results obtained for all queries were merged and sent to the participants.

For the evaluation, annotators selected subjective posts related to three target domains, assessed the polarity of these posts and labeled them with three scores corresponding to different sentiment scales (two-class, three-class and five-class scales).

The participants systems had to classify the reviews to two, three or five classes according to sentiment. The primary measures for evaluation of two and three class tasks were accuracy and macro-F1 measure. Macro-measures (Manning et al., 2008) were used because the majority of user reviews in blogs are positive (more than 80%). Macro-averaging means a simple average over classes. The five-class task was additionally evaluated with Euclidean distance measure, which is the quadratic mean between the scores of the al-

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

Domains	2-class		3-class		5-class	
	F1	Acc.	F1	Acc.	F1	Acc.
Movies	0.786	0.881	0.592	0.754	0.286	0.602
Books	0.747	0.938	0.577	0.771	0.291	0.622
Cameras	0.929	0.959	0.663	0.841	0.342	0.626

Table 1: Best results of blog review classification in ROMIP-2011

Domains	2-class		3-class		5-class	
	F1	Acc.	F1	Acc.	F1	Acc.
Movies	0.707	0.831	0.520	0.694	0.377	0.407
Books	0.715	0.884	0.560	0.752	0.402	0.480
Cameras	0.669	0.961	0.480	0.742	0.336	0.480

Table 2: Best results of blog review classification in ROMIP-2012

gorithm and the assessor scores.

Practically all the best approaches in the review classification tasks used SVM machine learning method (Kotelnikov and Klekovkina, 2012; Pak and Paroubek, 2012; Polyakov et al., 2012). Besides, the best methods usually combined SVM with other approaches including manual or automatic dictionaries or rule-based systems. The best achieved results according to macro-F1 measure and Accuracy within ROMIP 2011 are presented in Table 1 and within ROMIP 2012 in Table 2.

Observing the results of the open evaluation of sentiment analysis systems in Russian during two years we can make some conclusions about the state of the art performance and specific characteristics of the track.

The average level in 2-class classification task according to Accuracy is near 90%, near 75% for 3-class classification task and near 50% for 5-class task. Such results are consistent with the state of the art performance in English. However these figures are slightly overestimated due to the skewness of the testing collections. This fact is the consequence of using blogs as a test set. The majority of blog opinions about various objects is positive, but such a collection is a priori unlabeled, which leads to fair evaluation results.

3.2 Sentiment classification of opinionated quotations

The next task of ROMIP-2012 concerned sentiment classification of short (1-2 sentences on average) fragments of direct or indirect speech automatically extracted from news articles (further quotations). The somewhat similar task was conducted within the NTCIR-6, where one of the main tasks was extraction of opinion sentences from the news articles in three languages: English, Chinese and Japanese (Seki et al., 2007).

The topics of quotations could be quite different: from politics and economy to sports and arts. Therefore this task should be difficult enough for both knowledge-based and machine-learning approaches.

Assessors annotated quotations as positive, neutral, negative, or mixed. After the annotation the quotations with mixed sentiment were removed from the evaluation. So the participating systems should classify quotations to three classes. This task is similar to sentiment classification of political quotations (Awadallah et al., 2012; Balasubramanyan et al., 2012) to pro and contra positions. In (Awadallah et al., 2012) authors state that short quotations are difficult for classification because useful linguistic features tend to be sparse and the same quotation can have different polarities for different topics. In our case the task was even more difficult because of unlimited topics and three-class classification.

In ROMIP-2012 evaluation 4,260 quotations were prepared for training. For testing more than 120 thousand quotes were sent to participants, but real evaluation was made on the basis of 5,500 quotations randomly sampled and annotated from the testing set. An example of the quotation is as follows: Patriarch Kirill, says feminism is a ”very dangerous” phenomenon offering an illusion of freedom to women, who he says should focus on their families and children.

In this task class distribution was rather balanced in comparison with the review classification task: 41% of quotes were negative, 32% of

RunID	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

Table 3: Best results for the news quotation classification task in ROMIP 2012

RunID	Domain	P@1	P@5	P@10	NDCG@10
xxx-0	book	0.3	0.32	0.286	0.305
xxx-8	book	0.25	0.31	0.332	0.298
yyy-9	camera	0.402	0.313	0.302	0.305
yyy-1	camera	0.402	0.328	0.325	0.226
zzz-3	film	0.494	0.449	0.438	0.338
zzz-8	film	0. 494	0.448	0.444	0.332

Table 4: Best results in the task of retrieval of opinionated blog posts

quotes were positive and 27% of quotes were neutral. For evaluation again macro-measures and accuracy were applied.

The results of the participants are presented in Table 3. The baseline results correspond to classification of quotations according to the major class. In opposite to the review classification task, the leaders in the news-based classification 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 authors of the best approach in this task report that their knowledge-based system has a considerable vocabulary including 15 thousand negative expressions, 7 thousand positive expressions, around 120 so-called operators (intensifiers and invertors) and around 200 neutral stop expressions including sentiment words as their components. The system has a small number of rules for aggregating scores of sentiment word and operator sequences (Kuznetsova et al., 2013). The second and third results in this task were obtained by a rule-based system with comparably small sentiment dictionaries but a rich rule set based on syntactic analysis (Panicheva, 2013).

An interesting conclusion is that the size of sentiment dictionaries can be compensated with various syntactic rules, which allows handling the variety of situations in expressing sentiment.

The results of this task can be compared with one of the recent studies on lexicon-based methods for sentiment analysis in English (Taboada et al., 2011). The text fragments in the paper and in ROMIP evaluation are rather equal by style (news quotes versus opinionated news sentences). We cannot directly compare the results of analogous systems in Russian and English, because we worked with 3 class classification problem (positive, negative, neutral) versus 2 class task in the paper, but available figures are the following: the accuracy of sentiment analysis systems in Russian is near 61.6% in the three-class task versus 71.57% for the two-class task in English.

3.3 Query-based retrieval of opinionated blog posts

For several years TREC Blog tracks were connected with opinion finding and processing of blog data (Ounis et al., 2007; Macdonald et al., 2008; Ounis et al., 2008; Macdonald et al., 2010; Ounis et al., 2011). During the research cycles within these initiatives, the following sentiment analysis tasks were considered:

- Opinion finding (blog post) retrieval task,
- Polarised opinion finding (blog post) retrieval task.

The query-based retrieval of opinions from blogs was one of the basic tasks for the TREC Blog Track. Thus, we also decided to start with the similar task for Russian language. Here the participants had to find all relevant opinionated posts from the blog collection according to a specific query. Examples of queries include (translation from Russian):

- movie domain: *The Girl with the Dragon Tattoo; film “The dictator”*;
- book domain: *Agatha Cristie ”Ten little niggers”; Dan Brown “The Code da Vinci”*;
- digital camera domain: *Canon EOS 1100D Kit; Canon PowerShot G12*.

Only one group participated in this task and therefore organizers implemented a simple approach to conduct the track. The approach to the sentiment post retrieval was based on computation of weighted sum of three components: TFIDF similarity of a query to the title of a blog post, TFIDF of a query to the text of the post and the share of sentiment words in the post. For computation of the latter component, aforementioned Russian sentiment list ProductSentiRus (see section 2) was used:

$$Weight = \alpha \cdot (\sum_{w \in q} tfidf + \sum_{w \in q} tfidf^{header}) +$$

$$+(1-\alpha) \cdot (SentiWeight)$$

The organizers experimented with different values of $\alpha = 0.2, 0.4, 0.5, 0.6, 0.8$. The best performance was obtained with $\alpha = 0.6$ for all subdomains of this task. To avoid underestimation of participant results, the evaluation was made only on the basis of labeled documents. For this task we used two measures: $P@n$ and $NDGN@n$. $Precision@n$ indicates the number of correct (relevant) objects in the first n objects in the result set and $NDCG@n$ measures the usefulness, or gain, of a document based on its position in the result list (Manning et al., 2008). The main measures of the performance in this task were $NDCG@10$ and $Precision@10$ (Table 4).

4 Conclusion

In this paper we reported the state of the art of Russian sentiment analysis. Our report is based on the results of two evaluations of sentiment analysis systems organized in 2011–2012 within the framework of Russian seminar on information retrieval ROMIP. We proposed user review classification tasks in a practical setting, when available data should be used for training a classifier intended for similar, but another data. Besides, one of the interesting and complicated tasks of ROMIP-2012 was sentiment classification of opinions extracted from news articles.

Acknowledgments

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

References

Muhammad Abdul-Mageed, Mona Diab, and Mohammed Korayem. 2011. Subjectivity and sentiment analysis of modern standard arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 2, pages 587–591.

Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij, and Md Rijke. 2012. Overview of replab 2012: Evaluating online reputation management systems. In *CLEF 2012 Labs and Workshop Notebook Papers*, pages 1–24.

Rawia Awadallah, Maya Ramanath, and Gerhard Weikum. 2012. Polaricq: polarity classification of political quotations. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1945–1949.

Ramnath Balasubramanyan, William W Cohen, Doug Pierce, and David P Redlawsk. 2012. Modeling polarizing topics: When do different political communities respond differently to the same news? In *the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Ilia Chetviorkin and Natalia Loukachevich. 2013. Sentiment analysis track at romip 2012. In *Proceedings of International Conference Dialog*, volume 2, pages 40–50.

Ilia Chetviorkin and Natalia Loukachevitch. 2012. Extraction of russian sentiment lexicon for product meta-domain. In *Proceedings of COLING 2012*, pages 593–610.

Ilia Chetviorkin, P Braslavskiy, and Natalia Loukachevich. 2012. Sentiment analysis track at romip 2011. In *Proceedings of International Conference Dialog*, volume 2, pages 1–14.

Alexander Ermakov. 2009. Knowledge extraction from text and its processing: Current state and prospects. *Information technologies*, (7):50–55.

Evgeniy Kotelnikov and Marina Klekovkina. 2012. Sentiment analysis of texts based on machine learning methods. In *Proceedings of International Conference Dialog*, volume 2, pages 27–36.

Ekaterina Kuznetsova, Natalia Loukachevitch, and Ilia Chetviorkin. 2013. Testing rules for sentiment analysis system. In *Proceedings of International Conference Dialog*, volume 2, pages 71–80.

Craig Macdonald, Iadh Ounis, and Ian Soboroff. 2008. Overview of the trec 2007 blog track. In *Proceedings of TREC*, volume 7.

Craig Macdonald, Iadh Ounis, and Ian Soboroff. 2010. Overview of the trec 2009 blog track. In *Proceedings of TREC*, volume 9.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Rada Mihalcea, Carmen Banea, and Janyce Wiebe. 2007. Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, volume 45, pages 976–983.

Roser Morante and Eduardo Blanco. 2012. * sem 2012 shared task: Resolving the scope and focus of negation. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 265–274.

Iadh Ounis, Maarten de Rijke, Craig Macdonald, Gilad Mishne, and Ian Soboroff. 2007. Overview of the trec 2006 blog track. In *Proceedings of TREC*, volume 6.

Iadh Ounis, Craig Macdonald, and Ian Soboroff. 2008. Overview of the trec-2008 blog track. Technical report, DTIC Document.

Iadh Ounis, Craig Macdonald, and Ian Soboroff. 2011. Overview of the trec 2010 blog track. In *Proceedings of TREC*, volume 10.

Alexander Pak and Patrick Paroubek. 2012. Language independent approach to sentiment analysis (limsi participation in romip11. In *Proceedings of International Conference Dialog*, volume 2, pages 37–50.

Polina Panicheva. 2013. Atex. a rule-based sentiment analysis system. processing texts in various topics. In *Proceedings of International Conference Dialog*, volume 2, pages 101–113.

Verónica Pérez-Rosas, Carmen Banea, and Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, K Bretonnel Cohen, John Hurdle, and Christopher Brew. 2012. Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(Suppl 1):3–16.

Pavel Polyakov, Maria Kalinina, and Vladimir Pleshko. 2012. Research on applicability of thematic classification methods to the problem of book review classification. In *Proceedings of International Conference Dialog*, volume 2, pages 51–59.

Yohei Seki, David Kirk Evans, Lun-Wei Ku, Hsin-Hsi Chen, Noriko Kando, and Chin-Yew Lin. 2007. Overview of opinion analysis pilot task at ntcir-6. In *Proceedings of NTCIR-6 Workshop Meeting*, pages 265–278.

Josef Steinberger, Mohamed Ebrahim, Maud Ehrmann, Ali Hurriyetoglu, Mijail Kabadjov, Polina Lenkova, Ralf Steinberger, Hristo Tanev, Silvia Vázquez, and Vanni Zavarella. 2011. Creating sentiment dictionaries via triangulation. *Decision Support Systems*, pages 28–36.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Yunfang Wu and Peng Jin. 2010. Semeval-2010 task 18: Disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 81–85.

Taras Zagibalov, Katerina Belyatskaya, and John Carroll. 2010. Comparable english-russian book review corpora for sentiment analysis. In *Computational Approaches to Subjectivity and Sentiment Analysis*, pages 67–72.