LANGUAGE INDEPENDENT APPROACH TO SENTIMENT ANALYSIS (LIMSI PARTICIPATION IN ROMIP'11)

Pak A. (alexpak@limsi.fr),
Paroubek P. (pap@limsi.fr)

Université Paris-Sud, Lab. LIMSI-CNRS, Bâtiment 508, F-91405
Orsay Cedex, France

Sentiment analysis is a challenging task for computational linguistics. It poses a difficult problem of identifying user opinion in a given text. In this paper, we describe participation of LIMSI in the sentiment analysis track of the Russian annual evaluation campaign (ROMIP'11). The goal of the track was classification of opinions expressed in blog posts into two, three, and five classes. Our system based on SVM with dependency graph and n-gram features was placed 1st in 5-class task on all three datasets (movies, books, cameras), 3rd in the 2-class task on the movies dataset, and 4th in the 3-class task on the cameras dataset, according to the official results.

Key words: sentiment analysis, polarity classification, SVM, dependency parsing

1. Introduction

Sentiment analysis is a recent field of computational linguistics which emerged due to the growing demand of analysis of social media and user generated content in the Internet. Hence, to encourage the research in this field and to discover the current state of the art, sentiment analysis tasks have been included in a set of traditional evaluation campaigns tracks in information retrieval (IR) and natural language processing (NLP). TREC[1] 2006 added a blog opinion mining track, SemEval[2] 2010 organized a task on polarity disambiguation of Chinese adjectives, I2B2[3] 2011 dedicated one of the tasks to sentiment classification in suicide notes. In this paper, we describe our participation in ROMIP[4] 2011 sentiment analysis track.

[1] Text Retrieval Conference: http://trec.nist.gov/

[2] Evaluation Exercises on Semantic Evaluation: http://semeval2.fbk.eu/

[3] Informatics for Integrating Biology and the Bedside: http://www.i2b2.org/NLP/

[4] Russian Information Retrieval Evaluation Seminar: http://romip.ru

1.1. Task description

ROMIP is an annual evaluation campaign in information retrieval launched in 2002 [3]. In ROMIP 2011, the organizers added the sentiment analysis track which aimed at classification of opinions in user generated content. A dataset composed ofproduct reviews collected from a recommendation service Imhonet[5] and product aggregator service Yandex.Market[6] was provided to participants for training their systems. The dataset contained reviews about three topics: digital cameras, books, and movies. Table 1 shows the characteristics of the dataset.

Table 1. Characteristics of the training dataset

Topic	Source	# of reviews
Books	Imhonet	24,159
Movies	Imhonet	15,718
Cameras	Yandex.Market	10,370

Each review consists of the text of the review and meta information. Meta information contains the rating score assigned to the product, the product ID, reviewer ID, and the review ID. Reviews from Yandex.Market also contain review creation time, usefulness ofthe review (assigned by other users), pros and cons of the product given by the review author. Inour work, we used only the review text, the score, and pros/cons if available. The score is given on 1–5 scale for Imhonet reviews, and 1–10 scale for Yandex.Market reviews, where a higher value represents more positive opinion. Figure 1 shows an example of a digital camera review.

The evaluation dataset was not provided until the evaluation phase at the end of the campaign. The organizers have collected 16 861 posts from LiveJournal[7] blogging platform that mention books, movies, or cameras out of which 874 posts were annotated by two human experts. What makes this track different from other evaluation campaigns, is that the evaluation dataset was not of the same nature as the training data. First, the texts had different genres (product reviews vs. blogposts), and secondly the annotations were produced differently: the training data was composed automatically, while the testdata was annotated manually. Figure 2 shows an example of a test document.

The track was divided into three subtracks:

- Opinion classification into two classes: negative/positive
- Opinion classification into three classes: negative/mixed/positive

[5] http://imhonet.ru

[6] http://market.yandex.ru

[7] http://livejournal.com

- Opinion classification into five classes: a score on the scale 1–5, where 1 represents an exclusively negative opinion, and 5 represents an exclusively positive opinion

In its turn, each subtrack had 3 runs by the number of topics: classification in each topic was evaluated separately, resulting in total 9 separate evaluations.

```
<row rowNumber="0">
  <value columnNumber="0">1328131</value>    <!-- review ID     -->
  <value columnNumber="1">926707</value>     <!-- product ID    -->
  <value columnNumber="2">48983640</value>   <!-- author ID     -->
  <value columnNumber="3">2009-05-03</value> <!-- creation time -->
  <value columnNumber="4">4</value>          <!-- rating        -->
  <value columnNumber="5">                   <!-- text          -->
    Хороший выбор для опытного фотолюбителя.
    <!-- A good choice for an experienced amateur photographer. -->
  </value>
  <value columnNumber="6">                   <!-- pros          -->
    Большой выбор режимов съемки,12-кратный оптический зум,
    естественная цветопередача,большой ЖК-экран.
    <!-- Large selection of shooting modes,12-times optical zoom,
    natural color, large LCD screen. -->
  </value>
  <value columnNumber="7">                   <!-- cons          -->
    Невысокая скорость подзарядки фотовспышки.
    <!-- The low speed of flash recharge. -->
  </value>
  <value columnNumber="8">0.59375</value>    <!-- usefulness    -->
</row>
```

Fig. 1. An example of a review from the training dataset. Russian text has been translated into English only for this example

1.2. Task challenge

Sentiment analysis is a difficult task even for resource-rich languages (read, English). Along with simple language processing, such as part-of-speech (POS) tagging, more sophisticated NLP tools such as discourse parsers and lexical resources may be required by existing approaches. Thus, it is quite difficult to adapt methods that were developed in other languages (read, English) to Russian.

The ROMIP track poses additional challenges other than the difficulty of analysing sentiments in general. As mentioned before, the evaluation set was not constructed the same way as the training data. That makes it more difficult for statistical based approaches as the language model differs in two datasets. Moreover, the distribution of classes is also different. The training set contained more positive reviews, however the way the reviews were picked for annotation was unknown. Finally, the interpretation of rating also varies, as there were different conventions when assigning scoring products and when annotating the test set. In other words, a user of Yandex.Market may have a different interpretation of 3 stars assigned to a camera from a human annotator who rates a review.Multiclass classification was another challenge, since most of research on polarity classification consider it a binary problem, i. e. classifying a document into positive/negative classes.

```xml
<?xml version="1.0" encoding="windows-1251"?>
<document>
  <ID>11347</ID>
  <link>http://vikilt.livejournal.com/12619.html</link>
  <date>2011-02-06T20:59:15Z</date>
  <object>
    Плохая училка
    <!-- Bad teacher -->
  </object>
  <text>
    Недавно посмотрел фильм "Очень плохая училка" и наконец,
    увидел этого самого Джастина Тимберлейка о котором так много
    было звона и сильно удивился. В фильме персонаж Кэмерон Диос
    как только видит этого Джастина начинает млеть и интенсивно
    намокать, хотя сам персонаж никаких эротический эмоций кроме смеха
    и недоумения не вызывает. Дальше он там, в фильме поёт песенку,
    которая тоже оставляет желать лучшего. Девушки, неужели вам
    действительно нравятся такие чахлые додики сомнительной наружности?
    <!-- Recently, I have watched a movie "Bad teacher" and finally,
    I've seen this Justin Timberlake about whom there have been
    so much buzz and I was surprised a lot. In the movie,
    the character of Cameron Diaz becomes excited as soon as
    she sees this Justin, although his character does not invoke
    any feelings except laughing. Next, he there, in the movie,
    sings a song, which is poor also. Girls, do you really
    like such doubtful looking nerds? -->
  </text>
</document>
```

Fig. 2. An example of a document from the evaluation set. Russian text has been translated into English only for this example

Therefore, to tackle the problem, we have decided to use a language independent approach that is not dependent on sophisticated NLP tools or lexical resources (e. g. affective lexicons) that are not available in Russian. We used an SVM based system with features based on n-grams, part-of-speech tags, and dependency parsing. For that we have trained a dependency parser on the Russian National Corpus[8]. Additionally, a study on terms weighting and corpus composition has been performed in order

[8] http://www.ruscorpora.ru/en/

to optimize the performance of our system. The detailed description of our system is presented in Section 3 right after the overview of the current state of the art in Section 2. We report our experimental evaluation along with official results in Section 4. Finally, we draw conclusions in Section 5.

2. Related work

Polarity classification is one of the basic problems of sentiment analysis and probably the most studied. The existing approaches fall into two large categories: lexicon based and machine learning based methods.

Lexicon based methods make use of existing lexical resources that vary in their complexity starting from simple lists of positive and negative words to more sophisticated semantic maps. For English, one may use resources developed specifically for sentiment analysis and affective science such as SentiWordNet [4], WordNet-Affect [19], ANEW [2], General Inquirer [18], and also general purpose resources, such as WordNet [6]. However, to our knowledge no similar publicly available resource exists in Russian, therefore a lexicon based approach would require to create a lexicon from scratch which is a costly process. More over, the quality of the system would strongly depend on the quality of the developed resource. As the lexicon should cover well the analysed language model.

Machine learning based approaches in the majority are based on a classical framework for text classification. The most commonly used one is support vector machines with n-gram features trained on a large set of text with known polarities (usually positive or negative) [14][13].Other systems add on top of this basic framework additional text preprocessing, feature selection, and NLP.

The amount and the complexity of NLP varies in different approaches. We have previously reported the usefulness of POS tags for opinion mining [10]. Dependency parsing has been also widely used in the sentiment analysis domain for extracting additional features [1][8], determining opinion subject [21], and additional text analysis. A recent work by Zirn et al. [22], used discourse parsing to take into account relation between phrases for fine-grained polarity classification. One of few works on sentiment analysis in Russian by Pazelskaya and Solovyev [15] used a manually constructed affective lexicon along with POS-tagging and lexical parsing information for a rule based polarity classifier. However, many of these approaches are difficult to reproduce for the ROMIP track as there are few NLP tools for Russian that are publicly available.

3. Our approach

To overcome the difficulties of the task, thus to create a sentiment analysis system for Russian that would be robust in different topics without overfitting the training model, we developed an SVM based system using the LIBLINEAR package developed by Fan et al. [5]. For the 2-class track we trained SVM in binary classification mode, for the 3 and 5-class tracks, we used a multiclass and regression modes.

3.1. Training dataset composition

The distribution of opinion scores in the training data set was highly unbalanced, which caused difficulties for training the model. Figure 3 shows distributions of reviews by scores in different topics. In general, positive reviews are prevailing in the training dataset which creates a bias towards a positive class. For the 2-class problem, we have decided to balance the training dataset by using an equal number of reviews of negative and positive opinions. Thus we considered books and movies reviews with scores 1–4 as negative and 9–10 as positive, and in the cameras collection,we considered reviews with scores 1–2 as negative and 5 as positive. The rest of the reviews were not included in the training. For 3-class and 5-class problems we left the dataset as is, because there would not be enough data to represent each class.

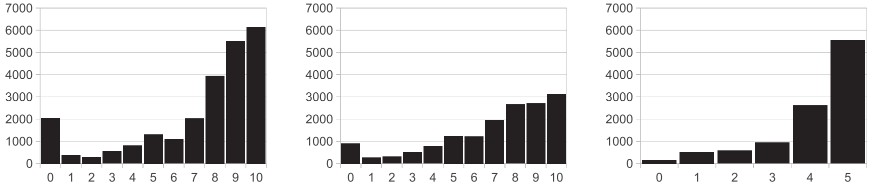

Fig. 3. Score distribution in books (left), movies (center), and cameras (right) datasets

Another decision which had to be made, was whether to train three separate modelsfor each topic orto combine all the data and to train one general model to classify reviews from each topic. We have experimented with both settings, and report the results in Section 4.

Reviews from Yandex.Market on cameras contain product prosand cons. To benefit from this additional information, we decided to include it in the text of there view. Thus, if a review is considered to be positive (using the criteria as mentioned above) then we add pros as the last phrase of the text. Otherwise, if a review is negative, we use cons. We have discovered that by doing this, we improved the accuracy of binary polarity classification up to 13.7%.

3.2. Feature vector construction

We have experimented with two types of features to build the model: traditional n-grams and our proposed d-grams features that are based on dependency tree of text sentences [12].

N-grams In the n-gram model, text is represented as a bag of words subsequences of a fixed size. We have experimented with unigrams and bigrams. Any non alphanumeric character was considered as a word boundary. Negations has been handled by attaching a negation particle (*не* — no, *ни* — neither, нет — not) to a preceding and a following word when constructing n-grams [10][20].

D-grams D-grams are similar to n-grams, however, while n-grams are constructed by splitting a text into subsequences of consecutive words, d-grams are constructed from a dependency parse tree, where words are linked by syntactic relations.

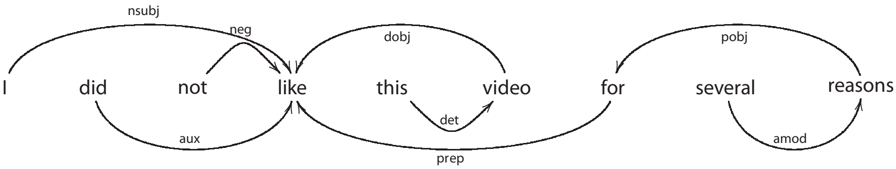

Fig. 4. Dependency graph of a sentence “The soundtrack was awful”

Figure 4 depicts an example of dependency parse tree of a sentence “The soundtrack was awful”. The dependency relations that we obtain are as follows:

{(I, nsubj, like),
(did, aux, like),
(not, neg, like),
(this, det, video),
(video, dobj, like),
(for, prep, like),
(several, amod, reasons),
(reasons, pobj, for)}

They are served as features in our d-gram model replacing the traditional n-gram model.To obtain dependency parse trees, we first applied TreeTagger [16][17] for tokenization and POS-tagging. Next, we fed the tagged output to the MaltParser [9] that we had trained on the Russian National Corpora.

Weighting scheme We consider two weighting schemes which are used in sentiment analysis.

Binary weights were used in first experiments by Pang et al. [14] and proven to yield better results than traditional information retrieval weighting such as TF-IDF. It assigns equal importance to all the terms presented in a document:

$$w(g_i) = 1, \text{if} g_i \in d, \text{otherwise} = 0 \tag{1}$$

where g_i is aterm(n-gram), d is a document. **Delta TF-IDF** was proposed by Martineau et al. [7] and proven to be efficient by Paltoglou et al. [13], assigns more importance to terms that appear primarily in one set (positive or negative):

$$w(g_i) = \mathrm{tf}(g_i) \cdot \log \frac{\mathrm{df_p}(g_i) + 0.5}{\mathrm{df_n}(g_i) + 0.5} \tag{2}$$

where tf(g_i) is term-frequency of a term (number of times g_i appears in document D), df$_\mathrm{p}$(g_i) is positive document frequency (number of times g_i appears in documents with positive polarity), dfp(g_i) is negative document frequency.

We augment Delta TF-IDF formula with our proposed average term-frequency normalization that lowers importance of words that are frequently used in a document [11]:

$$\mathrm{avg.tf}(g_i) = \frac{\sum_{\forall T, g_i \in T} \mathrm{tf(g_i)}}{\{T | g_i \in T\}} \tag{3}$$

where $\{T|g_i \in T\}$ is a set of documents containing term g_i . Thus, we modify Delta TF-IDF weight as follows:

$$w(g_i) = \frac{\mathrm{tf}(g_i)}{\mathrm{avg.tf}(g_i)} \cdot \log \frac{\mathrm{df_p}(g_i) + 0.5}{\mathrm{df_n}(g_i) + 0.5} \tag{4}$$

4. Experiments and results

In this section, we report results obtained during the system development phase and the offocial results provided by the organizers of ROMIP.All the development results were obtained after performing 10-fold cross validation.

4.1. Development results

Table 2. Macro-averaged accuracy over different training and test data. Rows correspond to a dataset on which the model has been trained, columns correspond to test data. *Combined* is a combination of all three topics

		Train data			
		books	movies	cameras	combined
Test data	books	76.0	74.0	65.5	73.4
	movies	77.3	76.4	66.4	74.5
	cameras	63.2	62.0	76.0	65.5
	combined	78.4	78.9	77.1	78.6

For the development phase, we present results only on binary classification as all the system parameters were tuned according to the results of these experiments.

Table 2 shows results of n-gram based model with binary weights across different topics. According top revious research on domain-adaptation for sentiment analysis a model trained on the same topics as the test set performs better than one trained on another topic. However, we were interested whether combining all the training data thus increasing the size of the available training data set improves the model. As we can see from the results, the model trained on the combined data performs better than a model trained only on one topic and the model trained on the same topic as the test set performs better than a model trained on another topic. However, we will see that it would change once we add additional information.

Table 3. Performance gain when adding class balancing and including pros/cons

	Books		**Movies**		**Cameras**	
	div	com	div	com	div	com
default	76.0	78.4	76.4	78.9	76.0	77.1
+ balanced	78.1 +1.9	79.5 +0.9	76.3 −0.1	78.2 −0.7	77.4 +1.4	77.5 +0.4
+ pros/cons	78.1	79.6 +0.1	76.3	78.6 +0.4	91.8 +13.7	87.9 +10.4

Table 3 shows show the performance changes after balancing the training data, and after adding pros and cons. Balancing the training set improves accuracy when classifying books and cameras and slightly degrades the performance on the movies collection. Adding pros and cons drastically improves the performance over the cameras test set (up to 13.7 % of gain). Notice, also that the model trained only on the cameras collection performs much better than the one trained on combined data (91.8 % vs. 87.9 %). Thus, for the following experiments we keep these settings: balancing training set and including pros and cons.

Table 4. Classification accuracy across different topics. For each topic, we evaluated a model trained on the same topic (div) and a model trained on all the reviews (com)

	Books		**Movies**		**Cameras**	
	div	com	div	com	div	com
ngrams + binary	78.1	79.6	**76.3**	**78.6**	91.8	87.9
ngrams + Δtfidf	77.4	78.8	76.2	76.5	93.1	90.4
dgrams + binary	78.0	79.8	74.9	77.8	91.3	88.2
dgrams + Δtfidf	**78.4**	**80.2**	76.1	77.3	**93.6**	**91.3**

Table 4 shows the comparison of the model using different features and weighting schemes. Here we have compared the traditional n-grams model with our proposed d-grams features using the same weighting schemes (binary and Delta TF-IDF). As we observe from the results, d-grams with Delta TF-IDF yields better accuracy on books and cameras test sets, while n-grams with binary weights perform better on the movies collection. However the difference is not very big.

4.2. Official results

According to the results we have obtained during the development phase, we have submitted the official runs on the unseen data. For 2-class track we have submitted 6 systems. For 3-class and 5-class tracks, we trained only systems based on n-grams due to time and resource constrains. For each of these tracks, we have submitted 4 systems. The summary of the submitted systems is presented in Table 6. The overall standings are depicted in Figures 5–7.

Table 5. Summary of the submitted systems

System ID	Mode	Features	Weights	Training set
		2-class track		
2-class track	binary	d-grams	Δtfidf	divided
2-dgram-delta-com	binary	d-grams	Δtfidf	combined
2-ngram-delta-div	binary	n-grams	Δtfidf	divided
2-ngram-delta-com	binary	n-grams	Δtfidf	combined
2-ngram-bin-div	binary	n-grams	binary	divided
2-ngram-bin-com	binary	n-grams	binary	combined
		3-class track		
3-ngram-bin-div	multiclass	n-grams	binary	divided
3-ngram-bin-com	multiclass	n-grams	binary	combined
3-regr-ngram-bin-div	regression	n-grams	binary	divided
3-regr-ngram-bin-com	regression	n-grams	binary	combined
		5-class track		
5-ngram-bin-div	multiclass	n-grams	binary	divided
5-ngram-bin-com	multiclass	n-grams	binary	combined
5-regr-ngram-bin-div	regression	n-grams	binary	divided
5-regr-ngram-bin-com	regression	n-grams	binary	combined

5. Conclusions

Sentiment analysis is a challenging task for computational linguistics. It becomes especially difficult for resource-poor languages. In this paper, we have described our participation in Russian sentiment analysis evaluation campaign

Table 6. Official ranking of the submitted systems

System ID	Books		Movies		Cameras	
	score	rank	score	rank	score	rank
			2-class track			
2-dgram-delta-div	65.1	24/53	70.3	5/27	**81.7**	**11/25**
2-dgram-delta-com	**66.1**	**23/53**	**70.9**	**3/27**	76.6	17/25
2-ngram-delta-div	61.8	31/53	70.0	7/27	77.8	15/25
2-ngram-delta-com	63.0	27/53	67.7	8/27	80.6	12/25
2-ngram-bin-div	57.9	36/53	63.7	10/27	79.2	13/25
2-ngram-bin-com	58.8	35/53	65.3	9/27	78.8	14/25
			3-class track			
3-ngram-bin-div	**48.4**	**12/52**	47.7	9/21	55.7	8/15
3-ngram-bin-com	49.9	18/52	**50.4**	**5/21**	**62.6**	**4/15**
3-regr-ngram-bin-div	47.6	21/52	48.4	8/21	50.0	9/15
3-regr-ngram-bin-com	48.8	16/52	49.8	6/21	57.4	7/15
			5-class track			
5-ngram-bin-div	27.0	4/10	24.6	5/10	**34.2**	**1/10**
5-ngram-bin-com	**29.1**	**1/10**	**28.6**	**1/10**	28.3	7/10
5-regr-ngram-bin-div	28.5	3/10	26.6	3/10	31.1	4/10
5-regr-ngram-bin-com	**29.1**	**1/10**	**28.6**	**1/10**	28.3	7/10

ROMIP 2011. We have tested our language independent framework for polarity classification that is based on SVM with the traditional n-grams model and our proposed features based on dependency parse trees. The developed system was ranked 1st in the 5-class track in all topics, 3rd in the 3-class track in movies domain, and 4th in the binary classification track in cameras domain according to the official evaluation metrics.

References

1. *S. Arora, E. Mayfield, C. Penstein-Ros´e, and E. Nyberg*. Sentiment classification using automatically extracted subgraph features. In Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, CAAGET '10, pages 131–139, Morristown, NJ, USA, 2010. Association for Computational Linguistics.
2. *M. M. Bradley and P. J. Lang*. Affective norms for English words (ANEW). Gainesville, FL. The NIMH Center for the Study of Emotion and Attention. University of Florida, 1999.
3. *B. Dobrov, I. Kuralenok, N. Loukachevitch, I. Nekrestyanov, and I. Segalovich*. Russian Information Retrieval Evaluation Seminar. In Proceedings of the Fourth International Conference on Language Resources and Evaluation,Lisbon,Portugal, May 2004.

4. *A. Esuli and F. Sebastiani.* SentiWordNet: A Publicly Available Lexical Resource forOpinionMining. InIn Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06), pages 417–422, 2006.
5. *R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin*. Liblinear: A library for large linear classification. J. Mach. Learn. Res., 9:1871–1874, June 2008.
6. *C. Fellbaum,* editor. WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press, illustrated edition edition, May 1998.
7. *J. Martineau and T. Finin*. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. In Proceedings of the Third AAAI Internatonal Conference on Weblogs and Social Media, San Jose, CA, May 2009. AAAI Press. (poster paper).
8. *T. Nakagawa, K. Inui, and S. Kurohashi.* Dependency tree-based sentiment classification using crfs with hidden variables. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pages 786–794, Morristown, NJ, USA, 2010. Association for Computational Linguistics.
9. *J. Nivre, J. Hall, and J. Nilsson*. MaltParser: A data-driven parser-generator for dependency parsing. In Proc. of LREC-2006, 2006.
10. *A. Pak and P. Paroubek*. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, may 2010. European Language Resources Association(ELRA).
11. *A. Pak and P. Paroubek*. Normalization of Term Weighting Scheme for Sentiment Analysis. In Proceedings of the 5th Language Technology Conference, Poznan, Poland, November 2011.
12. *A. Pak and P. Paroubek*. Text representation using dependency tree subgraphs for sentiment analysis. In Proceedings of the 16th international conference on Database systems for advanced applications, DASFAA'11, pages 323–332, Berlin,Heidelberg, 2011. Springer-Verlag.
13. *G. Paltoglou and M. Thelwall.* A study of information retrieval weighting schemes for sentimentanalysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL '10, pages 1386–1395, Morristown, NJ, USA, 2010. Association for Computational Linguistics.
14. *B. Pang, L. Lee, and S. Vaithyanathan*. Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing -Volume 10, EMNLP '02, pages 79–86, Morristown, NJ, USA, 2002. Association for Computational Linguistics.
15. *A. G. Pazelskaya and A. N. Solovyev.* A method of sentiment analysis in Russian texts. In Proceedings of the Dialog 2011 the 17th International Conference On Computational Linguistics, Moscow region, Russia, May 2011.
16. *H. Schmid.* Probabilistic Part-of-Speech Tagging Using Decision Trees. In Proceedings of the International Conference on New Methods in Language Processing, pages 44–49, 1994.
17. *S. Sharoff, M. Kopotev, T. Erjavec, A. Feldman, and D. Divjak*. Designing and evaluating a russian tagset. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may

2008 .European Language Resources Association(ELRA). http://www.lrecconf.org/proceedings/lrec2008/.

18. *P. J. Stone and E.B. Hunt*. A computer approach to content analysis: studies using the general inquirer system. In Proceedings of the May 21–23, 1963, spring joint computer conference, AFIPS'63 (Spring), pages 241–256, New York, NY, USA, 1963. ACM.
19. *C. V. Strapparava* and A. WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation , LREC, 2004.
20. *M. Wiegand, B. Roth, and D. Klakow.* A survey on the role of negation in sentiment analysis, 2010.
21. *L. Zhuang, F. Jing, and X.-Y. Zhu*. Movie review mining and summarization. In Proceedings of the 15th ACM international conference on Information and knowledge management, CIKM '06, pages 43–50, New York, NY, USA, 2006. ACM.
22. *C. Zirn, M. Niepert, H. Stuckenschmidt, and M. Strube.* Fine-grained sentiment analysis with structural features. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 336–344, Chiang Mai, Thailand, November 2011. Asian Federation of Natural Language Processing.

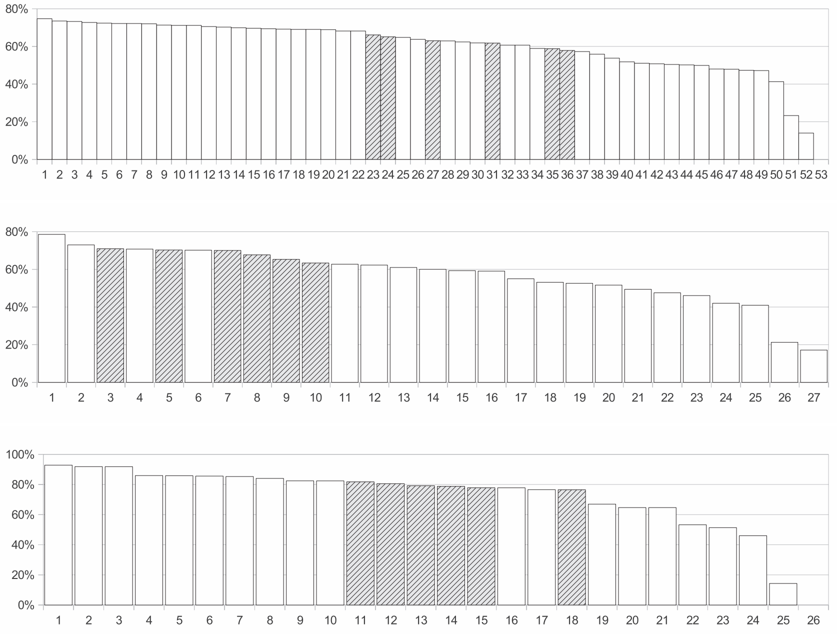

Fig. 5. Systems performance and ranking on the 2-class track on books (top), movies (middle), and cameras (bottom) collections. Our systems are highlighted

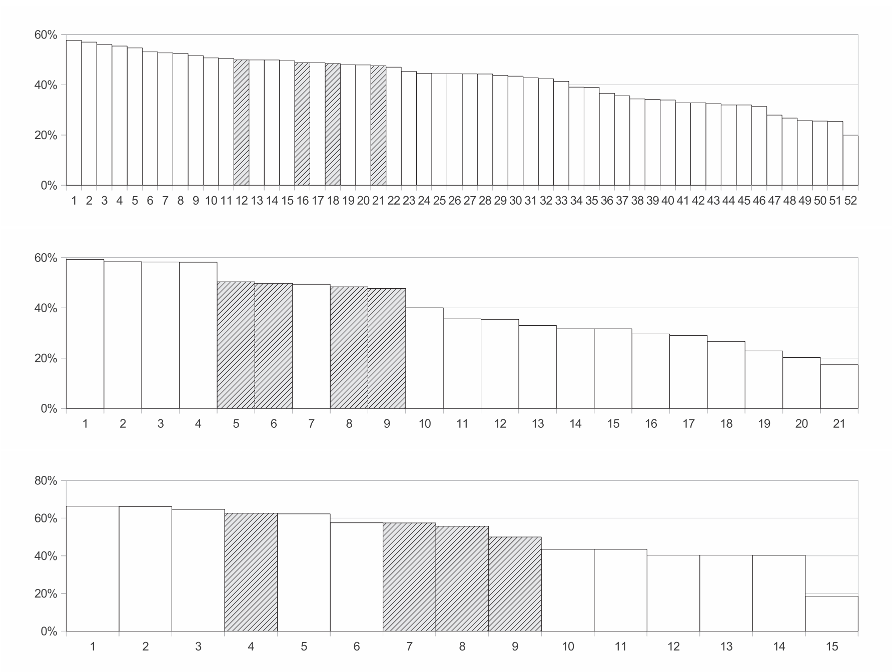

Fig. 6. Systems performance and ranking on the 3-class track on books (top), movies (middle), and cameras (bottom) collections. Our systems are highlighted

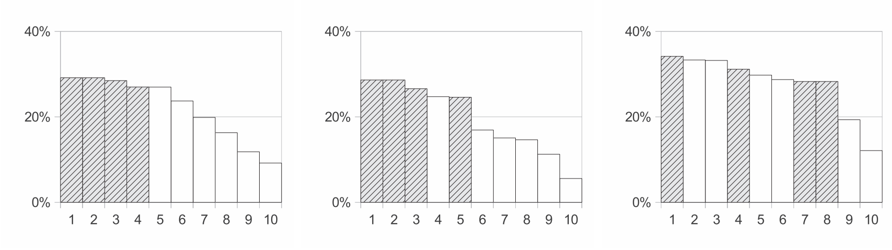

Fig. 7. Systems performance and ranking on the 5-class track on books (top), movies (middle), and cameras (bottom) collections. Our systems are highlighted