Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2016”

Moscow, June 1–4, 2016

SENTIRUEVAL-2016: ПРЕОДОЛЕНИЕ ВРЕМЕННЫХ РАЗЛИЧИЙ И РАЗРЕЖЕННОСТИ ДАННЫХ ДЛЯ ЗАДАЧИ АНАЛИЗА РЕПУТАЦИИ ПО СООБЩЕНИЯМ ТВИТТЕРА

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

Рубцова Ю. В. (yu.rubtsova@gmail.com)
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия

Ключевые слова: анализ тональности текстов, классификация текстов по тональности, социальные сети, разметка коллекций

SENTIRUEVAL-2016: OVERCOMING TIME GAP AND DATA SPARSITY IN TWEET SENTIMENT ANALYSIS

Loukachevitch N. V. (louk_nat@mail.ru)
Lomonosov Moscow State University, Moscow, Russia

Rubtsova Y. V. (yu.rubtsova@gmail.com)
A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

In this paper we present the Russian sentiment analysis evaluation SentiRuEval-2016 devoted to reputation monitoring of banks and telecom companies in Twitter. We describe the task, data, the procedure of data preparation, and participants’ results. At the previous evaluation

SentiRuEval-2015, it was noticed that the presented machine-learning approaches significantly depended on the training collection, which was not enough for qualitative classification of the test collection because of data sparsity and time gap. The current results of the participants at SentiRuEval-2016 showed that they have made successful steps to overcome the above-mentioned problems by combining machine-learning approaches and additional manual and automatically generated lexical resources.

Keywords: sentiment analysis, sentiment classification, social network, collection labeling, evaluation

1. Introduction

One of the important directions in automatic sentiment analysis is the analysis of social network messages, especially Twitter posts (tweets). Twitter messages convey a lot of opinions on various topics written by people of different origin, education, employment, etc, which can be interesting to governments, sociologists, companies, and ordinary people.

Twitter messages have several specific features. They are short (140 symbols), and their content is dynamic, often very dependent on current events. For this reason, automatic sentiment classifiers, trained in a restricted set of tweets, significantly lose in their quality if applied to tweet collections of other time intervals.

In [1, 2] the analysis of participants' results in the Russian tweet task of SentiRuEval-2015 was presented. Comparing results in two subtasks: sentiment analysis (reputation monitoring) towards telecommunication companies and towards banks, it was shown that best achieved levels of results significantly correlated with the differences between training and test collection. In that competition, the training and test collections were divided with the half-year interval, during which dramatic Ukraine events happened and partially changed the topics of the tweets. The analysis of the most problematic tweets for the participants showed that such tweets (30% in the bank domain) included sentiment words absent in the training set.

During this year, the second evaluation of tweet-oriented sentiment analysis systems was organized at SentiRuEval-2016. In this paper, we describe the task, the principles of data annotation, the achieved results and present the best approaches, which tried to overcome time-related problems of the tweet sentiment analysis.

2. Related Work

In past years several shared tasks were devoted to sentiment analysis and reputation monitoring of opinionated tweets.

In 2012–2013 RepLab, online reputation management evaluation, was held within the CLEF conference [3, 4]. The task was to determine if the tweet content has positive or negative implications for the company's reputation. The RepLab organizers

emphasize that the RepLab task is substantially different from standard sentiment analysis that should differentiate subjective from objective information. When analyzing polarity for reputation, both facts and opinions have to be considered to determine what implications a piece of information might have on the reputation of a given entity. The training and test collections were temporally divided with at least several month intervals.

To overcome the difference between the training and test collections, the participants combined supervised approaches with unsupervised approaches or lexicon-based approaches. Some runs incorporated external information by using provided links to Wikipedia, entities' official web sites, and external vocabularies.

The highest F-measure and accuracy values among RepLab 2013 were achieved by the system SZTE NLP [5]. The team utilized the external vocabularies: the SentiWordNet sentiment lexicon [6] and the acronym lexicon (from www.internetslang.com). Beyond the supervised steps, they experimented with unsupervised clustering using Latent Dirichlet Allocation (LDA) for detecting topics in the training and test collections. Then they used the topic distributions over each tweet as features.

The second best system according F-measure and third one according Accuracy was POPSTAR [7]. The team used sentiment lexicons to extract features based on the prior polarity of words. Some tweet-oriented features were included to capture particular aspects of tweets (e.g. presence of emoticons). The participant claims that delta-tf.idf weight scheme for word features shows the best results for this task. To solve the problem of feature vector sparseness and unseen words, they implemented the Brown cluster algorithm that clusters words to maximize the mutual information of bigrams.

The approach of the UAMCLYR [8] was based on distributional term representations (DTRs) [9], which are a way to represent terms by means of contextual information, given by term-co-occurrence statistics. The participant demonstrated that the proposed approach shows better result in comparison to the traditional Bag-of-Words representation.

In 2013–2015, the Twitter-oriented sentiment evaluation was held within the SemEval conference. Two subtasks were given to participants in 2013–2014: to detect sentiment expressed by a phrase in the context of a tweet and to detect overall sentiment of a tweet [10, 11]. In 2015 organizers included three new subtasks asking to predict the sentiment towards a topic in a single tweet, the overall sentiment towards a topic in a set of tweets, and the degree of prior polarity of a phrase [12].

In SemEval 2013 and 2014, the best result by a large margin was shown by NRC-Canada system [13]. The sentiment lexicon features (both manually created and automatically generated) along with n-gram features (both word and character n-grams) led to the most achievement in performance. Additionally, they generated two large sentiment association lexicons, one from tweets with sentiment-word hash tags, and another one from tweets with emoticons.

The best system for Subtask C (prediction sentiment towards a given topic) at SemEval 2015 was TwitterHawk [14]. The team focused on identifying and incorporating the strongest features used by the best systems in previous years, most notably, sentiment lexicons that showed good performance in earlier studies. Their system used two kinds of features: basic text features and lexicon features. They have

incorporated eight external lexicons. To increase the classifier quality, they extended the training collection with the data from subtask B.

In 2015 the first SentiRuEval evaluation of Russian sentiment analysis systems was held [1, 2]. The aim of the tweet analysis was to classify messages according to their influence on the reputation of the mentioned company. The analysis of the participants' results showed that the best achieved performance in the reputation oriented-task for a specific domain was correlated with the difference between word probability distributions over the training and test collections in this domain. From the description of the approaches, it became clear that no additional data (word clusters or lexicons) were not used by the participants in their supervised machine-learning methods.

3. SentiRuEval-2016 Twitter Task

Similar to the previous SentiRuEval-2015 evaluation, the goal of the Twitter sentiment analysis at SentiRuEval-2016 was to find tweets influencing the reputation of a company in two domains: banks and telecom companies. Such tweets may contain sentiment-oriented opinions or positive and negative facts about the company.

Such a task is quite similar to the reputation polarity task at RepLab evaluation [3, 4] and sub-task C in SemEval 2015. The difference from RepLab evaluation is that at SentiRuEval, tweets from only two domains were taken, and the systems were evaluated for these domains separately, which gives the possibility to compare the results obtained in the domains. The task for participants was to define the reputation-oriented attitude of a tweet in relation to a given company: positive, negative, or neutral.

In the training and test collections, the fields with the list of all companies of the chosen domain were denoted. By default, the field of the company mentioned in the tweet obtained "0" (neutral attitude) value. The participants should either replace "0" with "1" (positive attitude), or "–1" (negative attitude), or remain "0", if the tweet attitude to a company mentioned in the message is neutral.

3.1. Text Collections

The SentiRuEval collections comprise tweets about seven entities from the telecom domain and eight entities from the bank domain. The datasets were collected with Streaming API Twitter (https://dev.twitter.com/streaming/overview). The previous SentiRueval-2015 training and test collections (December 2013—January 2014; July-August 2014) were utilized as training collections of the current evaluation. The current test collections were gathered in two parts: during July 2015 and November 2015. The distribution of messages in the training and test collections according to sentiment classes is shown in Table 1. The number of tweets is not equal to the sum of neutral, positive and negative messages, as user may mention more than one company in a message. As it can be observed, the collections are unbalanced and we did not artificially boost the number of sentiment tweets—just how classifiers would face the sentiment classification task in the real life.

Table 1. The distribution of messages in the collections according to polarity classes in the SentiRuEval datasets

		Neutral	Positive	Negative	Total number of tweets
Telecom	Training collection	4,870	1,354	2,550	**8,643**
	Gold standard test collection	1,016	226	1,054	**2,247**
Banks	Training collection	6,977	704	1,734	**9,392**
	Gold standard test collection	2,240	312	722	**3,313**

The Twitter task of SentiRuEval-2016 was to determine the reputation-oriented attitude of a tweet in relation to a given company. Some tweets could contain more than one entity. Table 2 displays the number of tweets that contain more than one company and the number of tweets with different polarity labels.

To prevent manual labeling by participants, additional messages have been added to the test collections. The size of the collections sent to the participants was equal to 19,673 tweets for the Telecom domain and 19,586 tweets for the Bank domain.

Table 2. Number of tweets that contain more than one company

		Number of tweets containing more than one entity	Number of tweets containing different polarity labels
Telecom	Training collection	435	131
	Gold standard test collection	193	49
Banks	Training collection	857	23
	Gold standard test collection	101	11

3.2. Data Annotation and Quality Measures

A high-quality gold standard collection is essential for supervised machine learning. Traditionally the gold standard is created by expert annotators. However, traditional annotation is expensive and time-consuming. To reduce the cost of expert-based annotation, linguistic projects have turned to the crowdsourcing approach, which involves submitting smaller subtasks to a coordinated platform on the Internet and solving these smaller tasks with a large amount of people. Nowadays crowdsourcing is becoming an increasingly popular and rather practical approach for creation and annotation of linguistic resources [15,16]. Crowdsourcing can employ both paid workers and volunteers.

In the framework of SentiRuEval-2016, the online tool (http://sentimeter.ru/assess/texts/) for tweet labeling was created where one could mark tweets according to their attitude in relation to a given company. 8,509 tweets in total were loaded into the system and labeled by assessors: 3,970 tweets about telecommunication companies and 4,539 tweets about banks. The labeling process lasted since 1 September 2015 to 31 January 2016. The interface of the online crowd source platform for sentiment labeling is shown on Fig. 1.

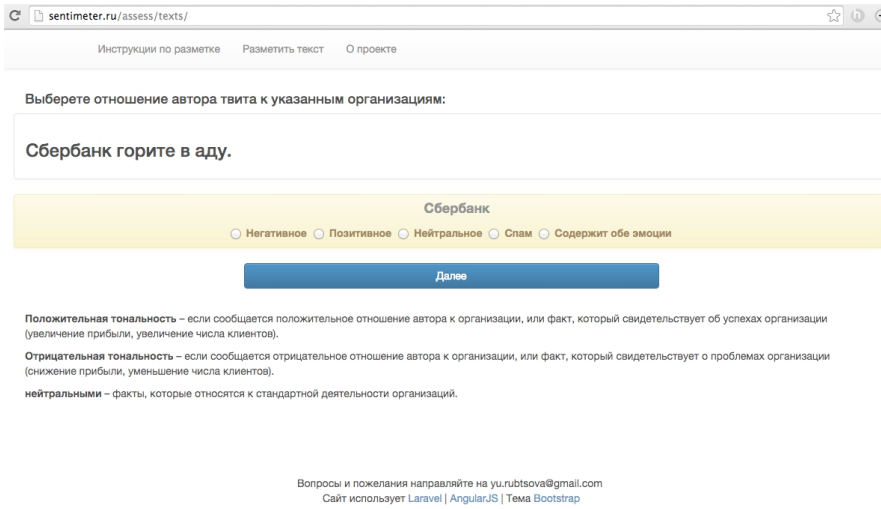


Figure 1. The interface of the online crowdsource platform for tweet sentiment labeling

For the four-month period of assessment, total 112 people from 25 cities and 7 countries took part in labeling. The annotators can be subdivided into three different groups: organizers—two persons, paid assessors—four persons, and volunteers—106 persons. All together they marked 45,450 companies. The organizers marked 10,322 companies, the paid assessors labeled 29,435 companies, and the volunteers labeled 5,693 companies (approximately 54 companies per volunteer).

To reduce the subjectivity, each tweet from the test collection was marked by at least four different persons as a person may feel preference or antipathy to some brand or company and mark tweets prejudiced. For instance, if a person is a "brand advocate" then he or she can label tweets as "neutral" if there is the slightest possibility not to label it as "negative".

After labeling was finished, the "strong agreement" voting scheme was applied to form the test collections. The labeling of a tweet was considered to be in strong agreement if the number of votes for a specific sentiment label exceeded votes for other labels with the margin 2. So, a tweet was filtered out from the gold standard if three assessors voted for one mark and two ones for another one—it was assumed as disagreement. Only tweets with strong agreement among assessors have formed the gold standard. Irrelevant, unclear, or spam messages were removed from the test sets.

As the main quality measure, macro-average F-measure was used. Macro F-measure is calculated as the average value between F-measure of the positive class and F-measure of the negative class ignoring the neutral class. But similar to SentiRuEval-2015, this does not reduce the task to the two-class prediction because erroneous labeling of neutral tweets negatively influences Fpos and Fneg. Additionally, micro-average F-measures were calculated for two sentiment classes.

4. Results and Description of Approaches

This year ten participants have submitted 58 runs to the Twitter sentiment analysis task at SentiRuEval-2016. The best runs according to macro-F for each participant are presented in Table 3 for telecom tweets and Table 4 for bank tweets.

In the evaluation we calculated two baselines. The first baseline is based on the major reputation-oriented category—negative one in both cases. The best runs of all participants show results above the F-macro majority baseline, however, some systems could not surpass the F-micro baseline.

The second baseline is obtained with the use of SVM to Boolean representation of tweet wordforms (if a wordform is presented in a tweet then the feature is equal to 1, otherwise 0). Six of ten participants could beat this baseline. If compared to SentiRuEval-2015, the considerable improvement can be seen because at the previous evaluation the best approaches in the bank domain were at the level of the SVM baseline (F-macro=0.3578, F-micro=0.3736). In the telecom domain, the best results were better than the SVM baseline, but the current margin between the SVM baseline and the best result is bigger (baseline: F-macro=0.4396, F-micro=0.48; the best result: F-macro=0.488, F-micro=0.536).

Two most popular machine-learning approaches among participants were SVM and neural networks. To overcome the differences between the training and test collections, five best approaches used machine learning in conjunction with external resources. Two participants (1 and 10) tried to increase the classification results by balancing the train collections. Three participants (1, 9, and 10) incorporated external sentiment vocabularies into supervised machine learning algorithms.

Table 3. The best run from each participant for telecom tweets according Macro F

	F-macro	F-micro
Majority Baseline	0.3146	0.5895
SVM baseline	0.4640	0.5728
1_4	0.5286	0.6632
2_k	**0.5594**	0.6569
3_1	0.3634	0.3994
4_5	0.4955	0.6252
5_1	0.3499	0.4044
6_con	0.3545	0.5263
7_5_a	0.4842	0.6374
8_533_2	0.4871	0.5745
9_hand_ext_tri	0.5493	**0.6813**
10_10	0.5055	0.6254

Table 4. The best run from each participant for banks tweets according Macro F

	F-macro	F-micro
Majority Baseline	0.1885	0.3503
SVM-baseline	0.4555	0.4952
1_4	0.4683	0.5022
2_k	**0.5517**	**0.5881**
3_1	0.3423	0.3524
4_1	0.376	0.4108
5_1	0.3859	0.464
6_con	0.2398	0.3127
7_5_a	0.471	0.5128
8_533_2	0.4492	0.4705
9_auto_ext_tri	0.5245	0.5653
10_5	0.4659	0.5053

These sentiment lexicons were as follows: the manual lexicon SentiRuLex[1] [17], the automatically generated lexicon study.mokoron.com [18], and the crowdsourced lexicon—Linis crowd[2] [19]. Participant 2 generated word clusters on a large collection of social network posts and comments and utilized them in tweet classification.

Below we briefly describe the best (compared to the baselines) approaches employed by the SentiRuEval participants. Participant 1 used words in uppercase, bigrams and punctuation marks as features for the linear-kernel SVM. The participant also integrated extra lexicons based on the following collections: study.mokoron.com [18], the collection of tweets (January 2016), and manual lexicon RuSentiLex [17]. The training collections were extended with other tweets in order to balance them. The telecom balanced collection consisted of 4,894 tweets, the banking balanced collection consisted of 6,980 tweets.

Participant 2 employed the recurrent neural network, and the long short-term memory (LSTM) model in particular. As features, Participant 2 used word2vec trained on the external collection of social network posts and comments.

The participant 9's best result for telecom companies is based on SVM over unigram, bigrams, and trigrams. Additionally, two vocabularies were implemented into the classifier: RuSentiLex [17] and automatic connotation vocabularies generated from a news collection. The best approach of this participant for the bank tweets also was based on SVM with the same features as it was used for the telecom domain but only the connotation lexicon was used and the consideration of the part-of-speech ambiguity was added.

The best runs of Participant 10 for the telecom tweets and banks also differ, but the only difference is that the classifier for telecom tweets worked better with a stop-word list, which showed the poor results for the bank domain. The participant used linear SVM with tweet-specific normalizations and integrated the RuSentiLex lexicon. The tweet-specific normalizations mean that the participle "not" plus a word was considered as one feature; multiple characters were replaced by a two-fold repetition; links, replies, dates, numbers were replaced with patterns.

The distribution of results of all 58 runs for the telecom test collection can be observed on graph 1. Graph 2 shows the distribution among all runs for the bank test collection.

We analyzed tweets that were incorrectly classified by all participants and found that most such tweets mention several entities with different sentiments, for example:

> *"А я вам всегда говорил, что лучший сотовый оператор это Билайн. Мегафон вас не уважает".* [I always said to you that the best operator is Beeline. Megaphone does not respect you].

Another found problem concerns phrase sentiment. The following tweet (and its several variants) was considered by all systems as negative:

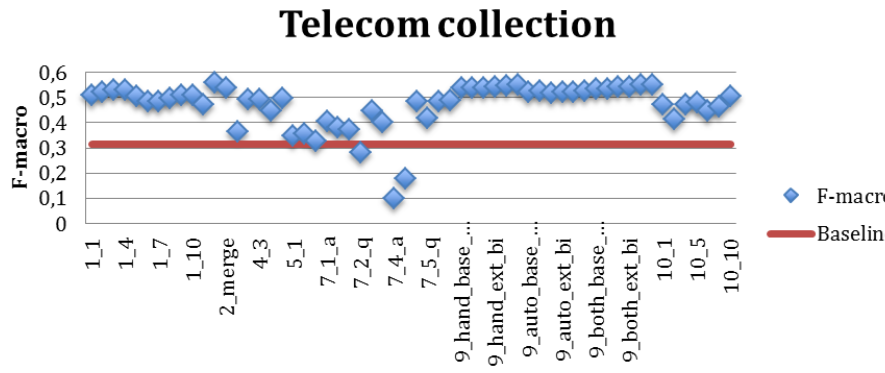
---

[1] http://www.labinform.ru/pub/rusentilex/index.htm

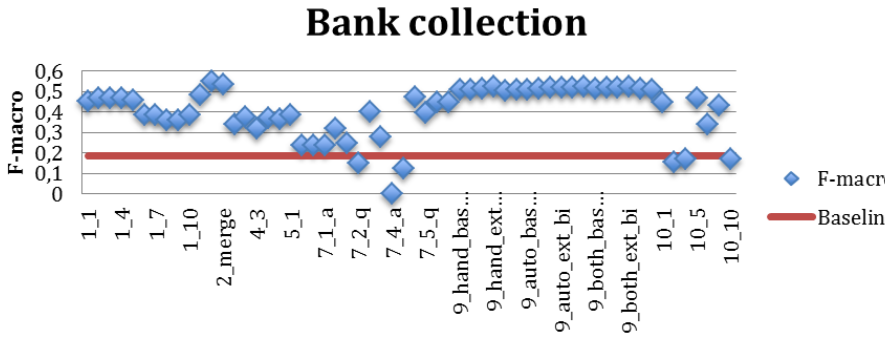
[2] http://linis-crowd.org/

"ВТБ 24 ***сократил убыток*** *вдвое во II квартале"* [VTB-24 **reduced losses** in II quarter].

Thus, it seems that dependence of the best systems from a training collection decreased, the systems now can use a lot of additional information. But they also should extract additional knowledge about phrase sentiment and try to find better ways to analyze different attitudes in the same tweet.



Graph 1. The distribution of all runs for Telecom collection



Graph 2. The distribution of all runs for Bank collection

Conclusion

In this paper we presented the Russian sentiment analysis evaluation SentiRuEval-2016 devoted to reputation monitoring of banks and telecom companies in Twitter. We described the task, data, the procedure of data preparation, and participants' results. At the previous evaluation SentiRuEval-2015, it was noticed that the presented machine-learning approaches significantly depended on the training

collection, which was not enough for qualitative classification of the test collection because of data sparsity and time gap. The current results of the participants at SentiRuEval-2016 showed that they have made successful steps to overcome the above-mentioned problems by combining machine-learning approaches and additional manual and automatic lexical resources.

All prepared collections are available for research purpose https://goo.gl/GhX3vU.

Acknowledgments

This work is partially supported by RFBR grants No. 14-07-00682 and No. 15-07-09306.

References

1. *Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E.* (2015), SentiRuEval: testing object-oriented sentiment analysis systems in Russian. Proceedings of International Conference Dialog-2015, pp. 3–9.
2. *Loukachevitch N., Rubtsova Y. Entity-Oriented Sentiment Analysis of Tweets: Results and Problems* (2015), Proceedings of Text-Speech-Dialog-2015, LNAI, Springer, 9302, pp. 551–559.
3. *Amigo E., Corujo A., Gonzalo J., Meij E., de Rijke M.* (2012), Overview of RepLab 2012: Evaluating Online Reputation Management Systems, CLEF 2012 Evaluation Labs and Workshop Notebook Papers, Rome.
4. *Amigo E., Albornoz J. C., Chugur I., Corujo A., Gonzalo J., Martin T., Meij E., de Rijke M., Spina D.* (2013), Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems, CLEF 2013, Lecture Notes in Computer Science Volume 8138, pp. 333–352.
5. *Hangya V., Farkas R.* (2013), Filtering and Polarity Detection for Reputation Management on Tweets, CLEF-2013 Working Notes.
6. *Baccianella, S., Esuli, A., Sebastiani, F.* (2010), SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).
7. *Filgueiras J., Amir S.* (2013), POPSTAR at RepLab 2013: Polarity for Reputation Classification, CLEF-2013 Working Notes.
8. *Villatoro-Tello E., Rodríguez-Lucatero C., Sánchez-Sánchez C., López-Monroy A. P.* (2013), UAMCLyR at RepLab 2013: Profiling Task. In CLEF (Working Notes).
9. *Lavelli A., Sebastiani F., Zanoli, R.* (2004), Distributional Term Representations: An Experimental Comparison. In Italian Workshop on Advanced Database Systems.
10. *Nakov P., Kozareva Z., Ritter A., Rosenthal S., Stoyanov V., Wilson T.* (2013), Semeval-2013 task 2: Sentiment analysis in Twitter, Proceedings of the 7th International Workshop on Semantic Evaluation SemEval-2014.

11. *Rosenthal S., Ritter A., Nakov P., Stoyanov V.* (2014), SemEval-2014 Task 9: Sentiment Analysis in Twitter, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 73–80.
12. *Rosenthal S., Nakov P., Kiritchenko S., Mohammad S., Ritter A., Stoyanov V.* (2015), Semeval-2015 task 10: Sentiment analysis in twitter. Proceedings of SemEval-2015. Denver, Colorado, June 4–5, 2015. Association for Computational Linguistics, pp. 451–463.
13. *Mohammad S., Kiritchenko S., Zhu X.* (2013), NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR'13).
14. *Boag W., Potash P., Rumshisky A.* (2015), TwitterHawk: A Feature Bucket Approach to Sentiment Analysis. SemEval-2015, pp. 640–646.
15. *Bocharov V., Alexeeva S., Granovsky D., Protopopova E., Stepanova M., Surikov A.* (2013), Crowdsourcing morphological annotation. In Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue", RGGU, pp. 109–124.
16. *Braslavski P., Ustalov D., Mukhin M.* (2014), A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics.—Gothenburg, Sweden : Association for Computational Linguistics, pp. 101–104.
17. *Loukachevitch N., Levchik A.* (2016), Creating a General Russian Sentiment Lexicon. In Proceeding of LREC-2016.
18. *Rubtsova Y.* (2015), Constructing a corpus for sentiment classification training, "Programmnye produkty i sistemy" (Software & Systems), №1 (109), pp. 72–78.
19. *Alexeeva, S., Koltsov, S., Koltsova O.* (2015), Linis-crowd.org: A lexical resource for Russian sentiment analysis of social media, Computational linguistics and computational ontology, pp 25–34.