SENTIRUEVAL: ТЕСТИРОВАНИЕ СИСТЕМ АНАЛИЗА ТОНАЛЬНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ ПО ОТНОШЕНИЮ К ЗАДАННОМУ ОБЪЕКТУ

Лукашевич Н. В. (louk_nat@mail.ru)[1],
Блинов П. Д. (blinoff.pavel@gmail.com)[2],
Котельников Е. В. (kotelnikov.ev@gmail.com)[2],
Рубцова Ю. В. (yu.rubtsova@gmail.com)[3],
Иванов В. В. (nomemm@gmail.com)[4],
Тутубалина Е. (tlenusik@gmail.com)[4]

[1]МГУ им. М. В. Ломоносова, Москва, Россия;
[2]Вятский государственный гуманитарный университет, Киров, Россия;
[3]Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск, Россия;
[4]Казанский федеральный университет, Казань, Россия

Статья описывает данные, правила и результаты SentiRuEval — тестирования систем автоматического анализа тональности русскоязычных текстов по отношению к заданному объекту или его свойствам. Участникам были предложены два задания. Первое задание было аспектно-ориентированный анализ отзывов о ресторанах и автомобилях; основная цель этого задания была найти слова и выражения, обозначающие важные характеристики сущности (аспектные термины), и классифицировать их по тональности и обобщенным категориям. Второе задание заключалось в анализе влияния твитов на репутацию заданных компаний. Такие твиты могут либо выражать мнение пользователя о компании, ее продукции или услугах, или содержать негативные или позитивные факты, которые стали известны об этой компании.

Ключевые слова: анализ тональности текстов, оценка качества, разметка коллекций, оценочные слова

SENTIRUEVAL: TESTING OBJECT-ORIENTED SENTIMENT ANALYSIS SYSTEMS IN RUSSIAN

Loukachevitch N. V. (louk_nat@mail.ru)[1],
Blinov P. D. (blinoff.pavel@gmail.com)[2],
Kotelnikov E. V. (kotelnikov.ev@gmail.com)[2],
Rubtsova Y. V. (yu.rubtsova@gmail.com)[3],
Ivanov V. V. (nomemm@gmail.com)[4],
Tutubalina E. (tlenusik@gmail.com)[4]

[1]Lomonosov Moscow State University, Moscow, Russia;
[2]Vyatka State Humanities University, Kirov, Russia;
[3]A. P. Ershov Institute of Informatics Systems, Novosibirsk, Russia;
[4]Kazan Federal University, Kazan, Russia

The paper describes the data, rules and results of SentiRuEval, evaluation of Russian object-oriented sentiment analysis systems. Two tasks were proposed to participants. The first task was aspect-oriented analysis of reviews about restaurants and automobiles, that is the primary goal was to find word and expressions indicating important characteristics of an entity (aspect terms) and then classify them into polarity classes and aspect categories. The second task was the reputation-oriented analysis of tweets concerning banks and telecommunications companies. The goal of this analysis was to classify tweets in dependence of their influence on the reputation of the mentioned company. Such tweets could express the user's opinion or a positive or negative fact about the organization.

Keywords: sentiment analysis, users review, collection labeling, aspect words, evaluation

1. Introduction

During last years the task of automatic sentiment analysis of natural language texts, that automatic extraction of opinions expressed in texts, attracts a lot of attention of researchers and practitioners. This is due to the fact that this task has a lot of useful applications. So the analysis and representation of users' opinions about products and services are of interest to their producers and competitors as well as to new users. Social opinion processing is important for authorities for better government.

The initial approaches to automatic sentiment analysis tried to determine the overall sentiment of the whole texts or sentences (Pang et al., 2002). This level of analysis presupposes that each document expresses opinions on a single entity (for example, a single product). Later, the task of object-oriented sentiment analysis appeared, when the system should reveal sentiment towards a specific entity mentioned in the text (Amigo et al., 2012; Jiang et al., 2011).

Finally, an author of a text can have different opinions relative to specific properties (or aspects) of an entity. To reveal these opinions, so called aspect-based sentiment analysis should be fulfilled (Liu, 2012; Bagheri et al., 2013; Glavaš et al., 2013; Popescu, Etzioni, 2005; Zhang, Liu, 2014). Aspects are expressed in texts with aspect terms and usually can be classified into categories. For example, "Service" aspect category in restaurant reviews can be expressed such terms as *staff*, *waiter*, *waitress*, *server*.

Automatic sentiment analysis is a complex problem of natural language processing. Several evaluation initiatives were devoted to study the best methods in sentiment analysis and related applications. These initiatives include Blog Track within TREC conference (Macdonald et al., 2010), TAC Opinion QA Tasks (Dang, Owczarzak, 2008), opinion tracks at NTCIR conferences (Seki et al., 2008), reputation management tracks at CLEF conference (Amigo et al., 2012), Twitter and review sentiment analysis tasks within SemEval initiative (Nakov et al., 2013; Rosenthal et al., 2014), etc.

In this paper we present results of SentiRuEval evaluation focusing on entity-oriented sentiment analysis of Twitter and aspect-oriented analysis of users' reviews in Russian. This evaluation is the second Russian sentiment analysis evaluation event in Russian after ROMIP sentiment analysis tracks in 2011–2013. This year in SentiRuEval we had two types of tasks. The first task is aspect-oriented sentiment analysis of users' reviews. The data included reviews about restaurants and automobiles. The second task was object-oriented sentiment analysis of Russian tweets concerning two varieties of organizations: banks and telecommunications companies.

The structure of this paper is as follows. In Section 2 we consider related evaluation initiatives in sentiment analysis. Section 3 describes tasks, data and principles of labeling in aspect-based review analysis. Section 4 describes the data and the task in the entity-oriented sentiment analysis of Twitter. Section 5 discusses results obtained by participants.

2. Related work

Several evaluation initiatives were devoted to sentiment analysis tasks similar to current SentiRuEval evaluation.

Last years in the framework of SemEval conference two types of sentiment analysis evaluations have been organized: sentiment analysis in Twitter and aspect-based sentiment analysis of reviews. In the Twitter task one of the subtasks was a message-level task, that is participating systems should classify if the message has positive, negative, or neutral sentiment (Nakov et al., 2013; Rosenthal et al., 2014). The task is directed to reveal, namely, the author opinion in contrast to neutral or objective information.

In the framework of CLEF initiative (http://www.clef-initiative.eu/) in 2012–2014 RebLab evaluations devoted to monitoring of reputation-oriented tweets were organized. The tasks included the definition of the polarity for reputation classification. The goal was to decide if the tweet content has positive or negative implications for the company's reputation. The organizers stress that the polarity for reputation 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 (Amigo et al., 2012; Amigo et al., 2013).

Evaluation of aspect-based review analysis at SemEval was organized in 2014 for the first time (Pontiki et al., 2014). The dataset included isolated, out of context sentences (not full reviews) in two domains: restaurants and laptops. 3K sentences were prepared for training in each domain. Set of aspect categories for restaurants included: *food, service, price, ambience, anecdotes/miscellaneous*.

In 2015 SemEval evaluations the aspect-based sentiment analysis of reviews (http://alt.qcri.org/semeval2015/task12/) is focused on entire reviews. Aspect categories of terms became more complicated and now consist of Entity-Attribute pairs (E#A). The E#A inventories for the restaurants domain contains 6 Entity types (RESTAURANT, FOOD, DRINKS, SERVICE, AMBIENCE, LOCATION) and 5 Attribute labels (GENERAL, PRICES, QUALITY, STYLE_OPTIONS, MISCELLANEOUS). The Laptops domain contains 22 Entity types and 9 Attribute labels.

In 2011–2013 two evaluation events of Russian sentiment analysis systems were organized. The first evaluation was devoted to extraction of overall sentiment of users' reviews in three domains: movies, books and digital cameras. For training, reviews from recommendation services were granted to participants. The evaluation was fulfilled on blog posts extracted with the help of the Yandex blog service (Chetviorkin et al., 2012). The second evaluation offered two new tasks for participants, namely: extraction of the overall sentiment of quotation (direct or indirect speech) from news articles and sentiment-oriented information retrieval in blogs when for a query (from the abovementioned domains) user opinions in blog posts should be found (Chetviorkin, Loukachevitch, 2013).

3. Ways to express opinions about aspects

Aspect terms also can be subdivided into several categories. They can be classified into three subtypes: **explicit aspects**, **implicit aspects** and **sentiment facts**.

Explicit aspects denote some part or characteristics of a described object such as *staff, pasta, music* in restaurant reviews. Explicit aspects are usually nouns or noun groups, but in some aspect categories we can meet explicit aspects expressed as verbs. For example, in restaurants the important characteristics of the service quality is time of order waiting, so this characteristic can be mentioned with verb *wait* (*ждать*): *ждали больше часа—waited for more than an hour.*

Implicit aspects are single words or single words with sentiment operators that contain within themselves as specific sentiments as the clear indication to the aspect category. In restaurant reviews the frequent implicit aspects are such words as *tasty* (*positive+food*), *comfortable* (*positive+interior), not comfortable* (*negative+interior*). The importance of these words for automatic systems consists in that fact that implicit aspects allow a sentiment system to reveal user's opinion about entity characteristics even if an explicit aspect term is unknown, written with an error or referred in a complicated way.

Sentiment facts do not mention the user sentiment directly, formally they inform us only about a real fact, however, this fact conveys us a user's sentiment as well as the aspect category it related to. For example, sentiment fact *отвечала на все вопросы* (*answered all questions*) means positive characterization of the restaurant service; this expression is enough frequent in restaurant reviews.

In the SentiRuEval labeling we annotated these three subtypes of aspect terms and our tasks for participants were not only to extract explicit aspect terms but also to extract all aspect terms (see Section 4).

An opinion about aspects can be expressed in several ways.

The **direct way of conveying the opinion** is through using opinion words such as *good*, *bad*, *excellent*, *awful*, *like*, *hate*, etc.

Opinions can be formulated as **comparisons** with other entities, previous cases or opinions of other people (Liu, 2012; Jindal, Liu, 2006). The problem of automatic analysis in these cases arise because used positive or negative words can be not relevant to the current review. In addition, comparison can be delivered in various ways not only using comparative constructions. For example, in the following extract from a restaurant review the comparison is marked with word *another,* and positive words *enjoyed* and *wonderful* characterize a restaurant distinct from the restaurant under review:

> *We decided not to have dessert and coffee there, but instead went to another restaurant where we **enjoyed** a **wonderful** end to our evening.*

We can formulate our opinion as **recommendation** (the constructive or suggestive opinion—see (Arora, Srinivasa, 2014)) or description of a **desirable situation** or characteristics of an entity, so called *irrealis factors* (Taboada et al., 2011; Kusnetsova et al., 2013). In these cases mentioned positive words can conceal the negative opinion.

At last, the opinion can be expressed with means of **irony or sarcasm** (Barbieri, Saggion, 2014; Riloff et al., 2013). In such cases the opinion can look like positive or at least medium one, but in fact it is strongly negative as in the following example: "***Excellent** translation, I don't understand anything*".

In the SentiRuEval labeling we marked these subtypes of opinions for further research (see Section 4).

4. Labeling and tasks of aspect-based analysis of reviews at SentiRuEval

For evaluation of aspect-oriented sentiment analysis systems we chose two domains: restaurant reviews and automobile reviews. In restaurant reviews aspect categories include: FOOD, SERVICE, INTERIOR (including atmosphere), PRICE, GENERAL. For automobiles aspect categories are: DRIVEABILITY, RELIABILITY, SAFETY, APPEARANCE, COMFORT, COSTS, GENERAL.

The length of reviews can vary drastically from one brief sentence to a long narrative. There can be also shifts to one or the other particular aspect. As an experiment, for labeling in the restaurant domain we tried to extract the most typical reviews

from our collection. To achieve it, the following procedure was performed. We represented each review as a bag-of-word vector and calculated the global collection's vector by averaging all the individual vectors. Then we imposed restrictions on min and max review length and chose most similar reviews according to the cosine similarity between global vector and single review vectors. As a result, most typical review representatives were selected for the labeling.

The labeling of training and test data was conducted with BRAT annotating tool (Stenetorp et al., 2012). Annotators had access to review collections through web interface. To unify and agree the annotation procedure, an assessor manual was prepared[1]. It is based on the SemEval-2014 (Pontiki et al., 2014) annotation guidelines.

The annotation task was to mark up two main types of tokens: aspect terms within a review and aspect categories attached to whole reviews. The aspect categories were labeled with the overall score of sentiment expressed in the text: positive, negative, both or absent.

According to the above-described categorization of opinions and aspect terms, the annotation of aspect terms within a text included several dimensions:

1. At first annotators should indicate explicit aspects, implicit aspects or sentiment facts in review texts and assign them their relevant type (explicit, implicit or fact).
2. All aspects terms should be assigned to aspect categories of the target entity.
3. Annotators marked the polarity of the aspect term: positive, negative, neutral, or both.
4. Annotators marked the relevance of the term to the review:
    a. *Rel—relevant* (to the current review),
    b. *Cmpr—comparison*, that is the term concerns another entity,
    c. *Prev—previous*, that is the term is related to previous opinions,
    d. *Irr—irrealis*, that is the term is the part of a recommendation or description of a desirable situation,
    e. *Irn—irony.*

So, for example, the annotation of word *девушка* (*girl*) in context *милая девушка* (*nice girl*) in a restaurant review includes sentiment orientation—*positive*, aspect category—*service*, aspect mark—*relevant,* aspect type—*explicit.*

Such detailed annotation process is very labor consuming. Therefore, each review was labeled only by a single assessor. However, to check the quality of aspect labeling two procedures were fulfilled after the labeling was finished. First, all labeled aspect terms were extracted from the markup according to their types and categories and were looked through; so some accidental mistakes were found and corrected. Second, we compared the aspect sentiment assigned to the review as a whole and sentiments of specific terms within this review. In cases of the differences between these two types of labeling the markup of the review was additionally verified.

During the annotation procedure, no balancing according to sentiment or aspect terms was performed; we tried to keep natural distributions specific for reviews in a given domain. Some statistics about relevant terms (Rel) are shown in Table 1.

---

[1] The manual is available at http://goo.gl/Wqsqit.

Table 1. Corpus statistics

		Restaurants		Automobiles	
		Train	Test	Train	Test
Number of reviews		201	203	217	201
Number of terms which are	explicit	2,822	3,506	3,152	3,109
	implicit	636	657	638	576
	fact	523	656	668	685
Number of terms which are	positive	2,530	3,424	2,330	2,499
	negative	684	865	1,337	1,300
	neutral	714	445	691	456
	both	53	85	100	115

The labeled data allowed us to offer the following tasks to the participants:

- **Task A**: automatic extraction of explicit aspects,
- **Task B**: automatic extraction of all aspects including sentiment facts,
- **Task C**: extraction of sentiments towards explicit aspects,
- **Task D**: automatic categorization of explicit aspects into aspect categories,
- **Task E**: sentiment analysis of the whole review on aspect categories.

To evaluate automatic systems the following quality measures were utilized.

For task A and B we applied macro F1-measure in two variants: exact matching and partial matching. Macro F1-measure means in this case calculating F1-measure for every review and averaging the obtained values.

To measure partial matching for every gold standard aspect term t we calculate precision and recall in the following way:

$$\text{Precision}_t = \frac{|t \cap t_s|}{|t_s|},$$

$$\text{Recall}_t = \frac{|t \cap t_s|}{|t|},$$

where t_s is an extracted aspect term that intersects with term t, $t \cap t_s$ is the intersection between terms t and t_s, $|t|$ is the length of the term in tokens. So F1-measure is calculated for every term and then we average the values for all gold standard terms.

For sentiment classification of aspect terms (task C) both variants of F1-measure (macro- and micro-) were utilized. Calculation of macro F1-measure is based on separate calculation of precision, recall, and F-measure for every category under consideration, then the obtained values are averaged. This allows us to evaluate the quality of categorization equally for every category. Micro F1-measure is calculated on the global confusion matrix, this measure greatly depends on the disbalance in the class distribution.

For aspect categorization of terms (task D) and the sentiment analysis of whole reviews (task E) macro F1-measure was used.

Table 2. Results in aspect-oriented review analysis (Restaurant domain)

Task	Measure	Baseline	Participants' results	Participant identifier
A	Exact matching, Macro F	0.608	**0.632**	**2**
			0.627	1
A	Partial matching, Macro F	0.665	**0.728**	**4**
			0.719	1
B	Exact matching, Macro F	0.587	**0.600**	**1**
			0.596	2
B	Partial matching, Macro F	0.619	**0.668**	**1**
			0.645	1
C	Macro F	0.267	**0.554**	**4**
			0.269	3
C	Micro F	0.710	**0.824**	**4**
			0.670	3
D	Macro F	0.800	**0.865**	**8**
			0.810	4
E	Macro F	0.272	**0.458**	**4**
			0.372	10

For all tasks we prepared baseline runs. The baseline system for tasks A and B extracts the list of labeled terms from the training collection, lemmatizes them and apply them to the lemmatized representation of the test collection. If more than one term matches the same word sequence, then a longer term is preferred.

The task C and D baseline systems attribute an aspect term to its most frequent category in the training collection. If a term is absent in the training collection then the most frequent aspect category is applied. The task E baseline is the most frequent sentiment category for the given aspect category (positive in all cases).

Altogether 12 participants with 21 runs were participated in the review sentiment analysis tasks. Due space limitations here we represent only two best results in each task and only primary F-measure, the full results are available at http://goo.gl/Wqsqit. Table 2 presents the participants' results for restaurant reviews, Table 3 contains the results for automobile reviews. Automobile reviews obtained much less attention from participants.

From the Tables 2, 3 it can be seen that the baselines for extracting aspect terms (tasks A and B) are quite high, which means the considerable agreement between annotation of training and testing collections. The best methods in these tasks were based on distributional approaches augmented with a set of rules (participant 4) and recurrent neural nets (participant 1). For the exact aspect matching, the best results were achieved by sequence labeling with SVM on the rich set of morphological, syntactic and semantic features (participant 2).

Table 3. Results in aspect-oriented review analysis (Automobile domain)

Task	Measure	Baseline	Participants' results	Participant identifier
A	Exact matching, Macro F	0.594	**0.676**	**2**
			0.651	1
A	Partial matching, Macro F	0.697	**0.748**	**1**
			0.730	2
B	Exact matching, Macro F	0.589	**0.636**	**2**
			0.630	1
B	Partial matching, Macro F	0.674	**0.714**	**1**
			0.704	1
C	Macro F	0.264	**0.568**	**4**
			0.342	1
C	Micro F	0.619	**0.742**	**4**
			0.647	1
D	Macro F	0.564	**0.652**	**8**
			0.607	4
E	Macro F	0.237	0.439	4

The best result in the analysis of sentiment towards aspect terms (task C) was obtained with Gradient Boosting Classifier (participant 4). The features were based on the skip-gram model exploiting word contexts for learning better vector representations and pointwise mutual information. In the task of categorization of explicit aspect terms (task D) the best results were obtained by SVM with features based on pointwise mutual information (participant 8). The second-place result is obtained by the method relying on the term similarity in the space of distributed representations of words (participant 4). For task E the best results were achieved by integration of the results obtained in tasks A, C and D (participant 4).

5. Object-oriented sentiment analysis of tweets

The goal of Twitter sentiment analysis at SentiRuEval was to find sentiment-oriented opinions or positive and negative facts about two types of organizations: banks and telecom companies. This task is quite similar to the reputation polarity task at RepLab evaluation (Amigo et al., 2013).

The training and test tweet collections were provided with fields corresponding all possible organizations for that tweets were extracted. A concrete organization mentioned in a given tweet was indicated with "0" label, denoting "neutral" as a default value. Annotators and participating systems should to leave this value unchanged if the tweet was considered as neutral or replace the value with "1" (positive) or "−1" (negative). The annotators also could label tweets with "−−", which means =meaningless=, or with "+−", which means positive and negative sentiments in the same tweet. Both latter cases were excluded from evaluation.

For training and testing collections assessors labeled 5,000 tweets in each domains (20000 tweets were labeled altogether). It is important to stress, that the training and testing collections were issued during different time intervals. The tweets of the training collection were written in 2014, the tweets of the testing collection were published in 2013.

Table 4. Results of the voting procedure in labeling of the tweet testing collection

Domain	The number of tweets with the same labels from at least 2 assessors	Full coincidence of labeling	The final number of tweets in the testing collection
Banks	4,915 (98.30%)	3,816 (76.36%)	4,549
Telecom companies	4,503 (90.06%)	2,233 (44.66%)	3,845

Analyzing the markup of the training collection we found that the estimation of some tweets can arise considerable discussion on their sentiment. To lessen the subjectivity of labeling and also accidental mistakes the testing collection was labeled by three assessors, and the voting scheme was applied to obtain the results of manual labeling. Finally, from the collection irrelevant tweets were removed. Results of the preparing the collection are presented in Table 4.

The participating systems were required to perform a three-way classification of tweets: positive, negative or neutral. As the main quality measure we used macro-average F-measure calculated as the average value between F-measure of the positive class and F-measure of the negative class. So we ignored Fneutral because this category is usually not interesting to anybody. But this does not reduce the task to the two-class prediction because erroneous labeling of neutral tweets negatively influences on Fpos and Fneg. Additionally micro-average F-measures were calculated for two sentiment classes.

Table 5. Results of participants in tweet classification tasks. The identifiers of participants in review and Twitter tasks are different

Domain	Measure	Baseline	Participant results	Participant identifier
Telecom	Macro F	0.182	**0.488**	**2**
			0.483	2
			0.480	3
Telecom	Micro F	0.337	**0.536**	**2**
			0.528	10
			0.510	3
Banks	Macro F	0.127	**0.360**	**4**
			0.352	10
			0.335	2
Banks	Micro F	0.238	**0.366**	**2**
			0.364	2
			0.343	8

In Table 5 we present the best results of tweet sentiment analysis for each domain and measure. Most best approaches in this task utilized SVM classification method. The features of the participant 2 comprised syntactic links presented as triples (head word, dependent word, type of relation). Participant 3 applied a rule-based method accounting syntactic relations between sentiment words and the target entities without any machine learning.

Additionally, one of participants fulfilled independent expert labeling of telecom tweets and obtained Macro-F—0.703, and Micro F—0.749, which can be considered as the maximum possible performance of automated systems.

The analysis of the obtained results showed that the most participants solved the general (not entity-oriented) task of tweet classification; entity-oriented approaches did not achieve better results in comparison with general approaches on tweets mentioned several entities.

6. Conclusion

In this paper we described the data, rules and results of SentiRuEval, evaluation of Russian object-oriented sentiment analysis systems. We offered two tasks to participants. The first task was aspect-oriented analysis of reviews about restaurants and automobiles, that is the primary goal was to find word and expressions indicating important characteristics of an entity (aspect terms) and then classify them into polarity classes and aspect categories.

The second task was the reputation-oriented analysis of tweets concerning banks and telecommunications companies. The goal of this analysis was to classify tweets in dependence of their influence on the reputation of the mentioned company. Such tweets could express the user's opinion or a positive or negative fact about the organization.

In each task about ten participants from universities and the industry took part. They have applied various machine-learning approaches including SVM, gradient boosting, CRF, recurrent neural networks and others. Given the participants' results, it can be concluded that the object-oriented sentiment analysis is poorly addressed by the applied methods. And most systems and methods need to be significantly improved to perform better on such tasks.

In the review collections interesting linguistic phenomena were also marked up. In particular, we have labeled comparisons with other entities or with previous opinions, desirable but not existing situations, irony. So the study of the markup can be useful also for linguists. All prepared materials are accessible for research purposes (reviews: http://goo.gl/Wqsqit and tweets: http://goo.gl/qHeAVo).

Acknowledgements

This work is partially supported by RFBR grants No. 14-07-00682, No. 15-07-09306 and by the Russian Ministry of Education and Science, research project No. 586.

References

1. *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.
2. *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.
3. *Arora R., Srinivasa S.* (2014), A Faceted Characterization of the Opinion Mining Landscape, COMSNETS Workshop on Science and Engineering of Social Networks, Bangalore, pp. 1–6.
4. *Bagheri A., Saraee M., de Jong F.* (2013), An Unsupervised Aspect Detection Model for Sentiment Analysis of Reviews, in Natural Language Processing and Information Systems, Springer, Berlin, Heidelberg, pp. 140–151.
5. *Barbieri F., Saggion H.* (2014), Modelling Irony in Twitter: Feature Analysis and Evaluation, Proceedings of LREC, pp. 4258–4264.
6. *Chetviorkin I. I., Braslavski P. I., Loukachevitch N. V.* (2012), Sentiment Analysis Track at ROMIP 2011, Proceedings of International Conference Dialog, pp. 739–746.
7. *Chetvirokin I., Loukachevitch N.* (2013), Sentiment Analysis Track at ROMIP 2012, Proceedings of International Conference Dialog, volume 2, pp. 40–50.
8. *Dang H. T., Owczarzak K.* (2008), Overview of the TAC 2008 Opinion Question Answering and Summarization Tasks, Proceedings of the First Text Analysis Conference.
9. *Glavaš G., Korencic D., Šnajder J.* (2013), Aspect-Oriented Opinion Mining from User Reviews in Croatian, Proceedings of the 4th Workshop on Balto-Slavonic Natural Language Processing, pp. 18–22.
10. *Jiang L., Yu M., Zhou M., Liu X., Zhao T.* (2011), Target-dependent Twitter Sentiment Classification, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp. 151–160.
11. *Jindal N., Liu B.* (2006), Mining Comparative Sentences and Relations, Proceedings of the 21st National Conference on Artificial Intelligence, Boston, pp. 1331–1336.
12. *Kusnetsova E., Loukachevitch N., Chetviorkin I.* (2013), Testing Rules for a Sentiment Analysis System, Proceedings of International Conference Dialog, pp. 71–80.
13. *Liu B.* (2012), Sentiment Analysis and Opinion Mining, Synthesis Lectures on Human Language Technologies, Vol. 5(1).
14. *Macdonald C., Santos R., Ounis I., Soboroff I.* (2010), Blog Track Research at TREC, ACM SIGIR Forum, Vol. 44(1), pp. 58–75.
15. *Mohammad S. M., Kiritchenko S., Zhu X.* (2013), NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, Proceedings of 7th International Workshop on Semantic Evaluation Exercises (SemEval-2013), Atlanta, pp. 321–327.
16. *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-2013), Atlanta, pp. 312–320.

17. *Pang B., Lee L., Vaithyanathan S.* (2002), Thumbs up? Sentiment Classification using Machine Learning Techniques, Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Vol. 10, pp. 79–86.
18. *Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.* (2014), SemEval-2014 Task 4: Aspect Based Sentiment Analysis, Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, pp. 27–35.
19. *Popescu A. M., Etzioni O.* (2005), Extracting Product Features and Opinions from Reviews, Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, pp. 339–346.
20. *Riloff E., Qadir A., Surve P., De Silva L., Gilbert N., Huang R.* (2013), Sarcasm as Contrast between a Positive Sentiment and Negative Situation, Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 704–714.
21. *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.
22. *Seki Y., Evans D. K., Ku L. W., Sun L., Chen H. H., Kando N.* (2008), Overview of Multilingual Opinion Analysis Task at NTCIR-7, Proceedings of NTCIR-7 Workshop Meeting, Tokyo, pp. 185–203.
23. *Stenetorp P., Pyysalo S., Topić G., Ohta T., Ananiadou S., Tsujii J.* (2012), BRAT: a Web-based Tool for NLP-assisted Text Annotation, Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon, pp. 102–107.
24. *Taboada M., Brooke J., Tofiloski M., Voll K., Stede M.* (2011), Lexicon-Based Methods for Sentiment Analysis, Computational Linguistics, Vol. 37(2), pp. 267–307.
25. *Zhang L., Liu, B.* (2014), Aspect and Entity Extraction for Opinion Mining, in Data Mining and Knowledge Discovery for Big Data, Springer, Berlin, Heidelberg, pp. 1–40.