# Extraction of Russian Sentiment Lexicon for Product Meta-Domain

*Ilia Chetviorkin*[1] *Natalia Loukachevitch*[2]
(1) Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University,
Moscow, Leninskiye Gory 1, Building 52
(2) Research Computing Center,
Lomonosov Moscow State University,
Moscow, Leninskiye Gory 1, Building 4
ilia.chetviorkin@gmail.com, louk_nat@mail.ru

ABSTRACT

In this paper we consider a new approach for domain-specific sentiment lexicon extraction in Russian. We propose a set of statistical features and algorithm combination that can discriminate sentiment words in a specific domain. The extraction model is trained in the movie domain and then utilized to other domains. We evaluate the quality of obtained sentiment vocabularies intrinsically. Finally we combine the sentiment lexicons from five domains to obtain one general lexicon for the product meta-domain. We demonstrate the robustness of the extracted lexicon in the cross-domain sentiment classification in Russian.

TITLE AND ABSTRACT IN RUSSIAN

## Извлечение Словаря Оценочной Лексики на Русском Языке для Мета-Области Товаров

В данной работе рассматривается новый подход к извлечению предметно-ориентированного словаря оценочной лексики на русском языке. Мы предлагаем использовать совокупность статистических и лингвистических признаков, позволяющих выявлять оценочные слова, и комбинировать эти признаки с помощью алгоритмов машинного обучения. Модель извлечения создается для предметной области фильмов, а затем применяется в других предметных областях. Мы оцениваем качество полученных словарей оценочных слов посредством ручной разметки. Наконец, мы собираем из отдельных словарей общий словарь оценочных слов, рассматривая его как оценочный словарь в широкой области товаров. Мы демонстрируем полезность полученного общего лексикона в задаче переноса модели анализа тональности с одной области на другую для отзывов пользователей на русском языке.

KEYWORDS : Sentiment Analysis, Sentiment Lexicon, Domain Adaptation.
KEYWORDS IN RUSSIAN: Анализ Тональности, Оценочные слова, Настройка на Предметную Область

*Proceedings of COLING 2012: Technical Papers*, pages 593–610,
COLING 2012, Mumbai, December 2012.

В последнее время большие усилия были направлены на решение задачи анализа мнений в различных предметных областях. Автоматизированные подходы к анализу тональности могут быть полезны для государственных органов и политиков, компаний и простых пользователей. Одной из важнейших задач, являющейся основой для анализа мнений в текстах, написанных на различных языках, является создание словарей оценочных слов.

Многие исследователи создают словари общеупотребительных оценочных слов для своих языков. Вместе с тем известно, что в разных предметных областях могут применяться достаточно разные наборы оценочных выражений. Наконец, предметные области могут иметь сходство между собой в используемой оценочной лексике. Так, такие оценочные слова как *негодяй* или *зло* одинаково неприменимы ко всем областям оценки качества товаров.

В данной работе мы исследуем новую идею разработки русского словаря оценочной лексики для широкой области товаров. При этом важно подчеркнуть, что в настоящее время нет общественно доступного русскоязычного словаря оценочной лексики. Наш метод базируется на обучении алгоритма извлечения русской оценочной лексики в одной предметной области, и затем переносе обученной модели на другие предметные области. Мы показываем, что модель извлечения оценочной лексики может быть перенесена на другие предметные области, если имеются все необходимые для работы системы данные. Мы применяем нашу модель к нескольким предметным областям и затем из оценочных словарей отдельных предметных областей собираем единый словарь оценочной лексики, рассматривая его как словарь оценочной лексики в широкой области товаров.

Извлечение оценочных слов в заданной предметной области основано на нескольких текстовых коллекциях: коллекции отзывов о продуктах с оценками пользователей, коллекции описаний продуктов и контрастной коллекции (например, новостная коллекция). Такие коллекции могут быть автоматически сформированы для разных предметных областей. Кроме того, мы предположили, что можно выделить некоторые части корпуса мнений (например, о фильмах), в которых концентрация оценочных слов выше: предложения, заканчивающиеся на «!» или «…»; короткие предложения не более чем из 7 слов; предложения, содержащие слово «фильм» без других существительных. Условно назовем этот корпус – малый корпус.

Для каждого слова в коллекции отзывов мы вычисляем набор статистических и лингвистических признаков.

Для обучения алгоритмов нам необходимо размеченное множество слов. Для этого мы вручную разметили множество всех слов с частотой выше трех из предметной области о фильмах (18362 слова). Мы относили слово к категории оценочных в случае если могли представить его в каком-либо оценочном контексте.

Мы решали задачу классификации на два класса: разделение всех слов на оценочные и неоценочные. Для этих целей использовались следующие алгоритмы: *Logistic Regression*, *LogitBoost* и *Random Forest*. Все параметры алгоритмов были выставлены в соответствии с их значениями по умолчанию.

Используя данные алгоритмы, мы получили списки слов, упорядоченные по вероятности оценочности слов. Для оценки качества этих списков использовалась мера *Precision@n*. Для сравнения качества работы системы в разных предметных областях мы использовали значение n = 1000.

Мы заметили, что извлеченные списки оценочных слов существенно различаются в зависимости от алгоритма. Поэтому мы решили вычислить среднее от значений вероятностей в каждом из списков. В результате качество автоматического извлечения оценочных слов в области фильмов *Precision@1000* составило 81.5%.

Для использования системы в новой предметной области необходимо собрать аналогичный набор коллекций, как и предметной области о фильмах. Мы применили модель извлечения оценочных слов в таких областях, как книги, игры, цифровые камеры, мобильные телефоны.

Для того чтобы собрать обобщенный список оценочной лексики в области товаров, мы применили формулу, поощряющую нахождение оценочного слова в начале наибольшего количества полученных списков оценочной лексики в разных предметных областях. Качество полученного списка составило P@1000 = 91.4%.

Для проверки полезности полученного обобщенного списка оценочных слов в мета-области товаров мы протестировали его в задаче переноса системы анализа тональности с одной области на другую.

Для тестирования мы взяли по 1000 положительных и 1000 отрицательных отзывов в четырех предметных областях. Мы обучали классификатор тональности в одной области на трех разных наборах признаков: всех словах, извлеченному списку оценочных слов этой предметной области и обобщенному списку оценочных слов. Далее мы применяли обученный классификатор на другой предметной области. Всего было рассмотрено 9 пар предметных областей. Было показано, что в среднем классификатор, обученный на обобщенном списке предметных областей, лучше переносится на новую предметную область.

Таким образом, в нашей работе мы создали русскоязычный список оценочных слов для широкой области товаров и показали его полезность в задачах, связанных с настройкой систем анализа тональности на новую предметную область. Мы планируем опубликовать полученный список оценочных слов, и это будет первый общественно доступный список оценочной лексики для русского языка.

## 1 Introduction

Over the last few years a lot of efforts were made to solve sentiment analysis tasks in different domains. Automated approaches to sentiment analysis can be useful for state bodies and politicians, companies, and ordinary users. Most of these efforts concern English, where a lot of resources and tools for natural language processing and especially for sentiment analysis exist.

One of the important tasks, considered as a basis for sentiment analysis of documents written in a specific language, is a creation of its sentiment lexicon (Abdul-Mageed et al., 2011; Peres-Rosas et al., 2012).

Usually authors try to gather general sentiment lexicons for their languages. However a lot of researchers stress the differences between sentiment lexicons in specific domains. For example, "must-see" is a strongly opinionated word in the movie domain, but neutral in the digital camera domain (Blitzer et al., 2007). For these reasons, supervised learning algorithms trained in one domain and applied to other domains demonstrate considerable decrease in the performance (Ponomareva & Thelwall, 2012; Read & Carroll, 2009; Taboada et al., 2011).

To overcome this issue various adaptation methods are proposed, like ensembles of classifiers (Aue & Gamon, 2005) or graph-based approaches (Wu et al., 2009). Nevertheless such approaches usually do not work well for domains whose lexicons differ significantly and recent studies are focused on bridging the gap between domain-specific words (Pan et al, 2010). Indeed, sentiment lexicons adapted to a particular domain or topic have been shown to improve task performance in a number of applications, including opinion retrieval (Jijkoun et al., 2010), and expression-level sentiment classification (Choi & Cardie, 2009). In addition sentiment word extraction from a text collection enables to find slang and non-vocabulary words, which can be strong sentiment predictors.

Stressing the differences in sentiment lexicons between domains, one should understand that domains can form clusters of similar domains. So a lot of sentiment words relevant to various product domains are not relevant to the political domain or the general news domain and vice versa. For example, such words as *evil* or *villain* are not applicable to all product domains. Therefore we suppose that gathering a specialized sentiment lexicon for the product meta-domain can be useful for researchers and practitioners.

In the current study we focus on the novel idea of construction of Russian sentiment lexicon for the product meta-domain. At this moment we should also emphasize that no publicly available Russian sentiment lexicon exists. Our method is based on training of the supervised algorithm for sentiment lexicon extraction in one domain and further transfer of the model to other domains. We show that in comparison with supervised sentiment classifiers, our sentiment lexicon extractor can be transferred to other domains if all necessary data are available. The trained sentiment lexicon extraction model is applied to an extensive number of domains and then extracted lexicons are summed up to the single list of sentiment words. So we obtain the generalized sentiment lexicon for the group of domains.

We opt to focus on recognizing sentiment words without any polarity scores. It is pointed in the research papers that the two-stage approach is often beneficial, in which on the first stage we determine main sentiment bearers in a text and on the second stage classify them according to the polarity (Pang and Lee, 2008). Thus such sentiment lexicons can be very useful for more accurate processing of user opinions.

We evaluate the extracted general lexicon intrinsically, by manually labelling of word lists, and extrinsically, by transferring of sentiment classifiers based on our general lexicon to domains without any labelled data. The results demonstrate the effectiveness of our constructed general sentiment lexicon.

The reminder of this article is organized as follows. In Section 2 we observe state-of-the-art methods for the sentiment lexicon generation, Section 3 describes the data collections and features involved in the model, in Section 4 we utilize our approach for four other domains and combine sentiment word vocabularies from all of them in Section 5. Finally, in Section 6 we conduct the experiments on the cross-domain sentiment classification involving extracted sentiment words.

## 2 Related work

The related works can be divided into two categories: the creation of a sentiment lexicon for a specific language, and the creation of a sentiment lexicon for a specific domain.

### 2.1 Creation of sentiment lexicons for specific languages

There are four main methods that are exploited by researchers to develop the sentiment lexicons for their languages: use of translated English sentiment resources, use of language-specific wordnets aligned to Princeton WordNet, use of corpora-based techniques similar to the techniques proposed for English sentiment lexicon extraction, use of electronic dictionaries of specific languages.

In (Mihalcea et al., 2007) two methods for translating sentiment lexicons to Romanian are proposed. The first method uses bilingual dictionaries to translate an English sentiment lexicon gathered using OpinionFinder (Wiebe & Riloff, 2005) and obtain 4,983 Romanian sentiment words. The evaluation of randomly chosen units shows the percentage of the sentiment words in the list is around 50%; besides, the low coverage of existing Romanian sentiment expressions is revealed. The second method is based on parallel corpora. The corpus on the source language is annotated with sentiment information, and the information is then projected to the target language. The problems arise due to mistranslations, e.g. because irony is not recognized.

Researchers in (Banea et al., 2008) propose to use a monolingual dictionary to acquire a sentiment lexicon from 60 manually selected seeds, equally sampled from verbs, nouns, adjectives and adverbs. To filter erroneous entries the LSA similarity measure is used.

In (Perez-Rosas et al., 2012) a method to derive Spanish lexicons by using manually or automatically annotated data available in English is presented. The multilingual sense-level aligned WordNet structure is used to generate a highly accurate (90%) polarity lexicon comprising 1,347 entries, and one with accuracy (74%) encompassing 2,496 words.

(Clematide & Klenner, 2010) begin their work with German polarity lexicon from 8000 polarity words obtained from GermaNet, a WordNet-like lexical database. Revealing rather low coverage of German novels by polarity-bearing adjectives from this list, they expand the set of 2899 German sentiment adjectives extracting coordinated adjectives pairs similar to (Hatzivassiloglou & McKeown, 1997).

To enhance the quality of dictionary-based methods for the general sentiment vocabulary generation in other languages, (Steinberger et al., 2011) create two source sentiment vocabularies: English (2400 entries) and Spanish (1737 entries). Both lists are translated by Google translator to the target language. Only 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 comparison with these approaches we create a Russian lexicon for a very broad domain - meta-domain of products and services, for which we do not use any dictionaries - only users' reviews, and in this paper we show usefulness of this general lexicon.

## 2.2 Development of sentiment lexicons for specific domains

In many studies domain-specific sentiment lexicons are created using various types of propagation from a seed set of words, usually a general sentiment lexicon (Kanayama & Nasukawa, 2007; Lau et al., 2011; Qiu et al., 2011). In such approaches an important problem is to determine an appropriate seed lexicon for propagation, which can heavily influence the quality of the results. Besides, the propagation often lead to unclear for a human sentiment lists. So, for example, in (Lau et al., 2011) only 100 first obtained sentiment words were evaluated by experts, *precision@100* was around 80%, what means that the intrinsic quality of the extracted 4000 lexicon (as announced in the paper) can be quite low.

Another approaches apply statistical measures based on domain-specific corpora to extract domain-specific sentiment words: $\chi 2$ (Jijkoun et el., 2010), divergence from randomness (DFR), which measures the divergence between a term's probability distribution in a set of relevant and opinionated documents and its probability distribution in a set of relevant documents (He et al., 2009) etc.

The sentiment lexicon extraction method proposed in this paper exploits a set of statistical and linguistic measures, which can characterize domain-specific sentiment words from different sides. We combine these features into a single model using machine learning methods. Then we train it on one domain and show that such a model can be effectively transferred to other domains for extraction of their sentiment lexicons.

# 3 Extraction of sentiment lexicon in a specific domain

In the current study a new supervised method for domain-specific sentiment lexicon extraction is presented. We train our model in one domain and then apply it to several others. Finally, we combine the extracted word lists to construct a general lexicon of sentiment words typical for products and services.

Our approach is based on several text collections, which can be automatically formed for many domains, such as: a collection of product reviews with authors' evaluation scores, a text collection of product descriptions and a contrast corpus (for example, a general news collection). For each word in the review collection we calculate a set of linguistic and statistical features using the aforementioned collections and then apply machine learning algorithms for term classification.

Our method does not require any seed words, and is rather language-independent, however, lemmatization (or stemming) and part-of speech tagging are desirable. Working with Russian language, we use a dictionary-based morphological processor, including unknown word processing. Below in the text we will speak only about lemmatized words.

## 3.1 Data preparation

We collected 28, 773 movie reviews of various genres from the online recommendation service *www.imhonet.ru*. For each review, user's score on a ten-point scale was extracted. We called this collection the **review collection**.

Example of the movie review:

*Nice and light comedy. There is something to laugh - exactly over the humour, rather than over the stupidity… Allows you to relax and gives rest to your head.*

We also required a contrast collection of texts for our experiments. In this collection the concentration of opinions should be as little as possible. For this purpose, we collected 17, 680 movie descriptions. This collection was named the **description collection**.

One more contrast corpus was a collection of two million news documents. We had calculated a document frequency of each word in this collection and used only this frequency list further. This list was named the **news corpus**.

## 3.2 Collections with higher concentration of opinions

We suggested that it was possible to extract some fragments of reviews from the review collection that had higher concentration of sentiment words. These fragments may include:

- Sentences ending with a "!";
- Sentences ending with a "…";
- Short sentences, no more than seven word length;
- Sentences containing the word «movie» without any other nouns.

We called this collection the **small collection**.

## 3.3 Statistical features

Our aim is to create a high quality list of sentiment words based on the combination of various discriminative features. We propose the following set of features for each word:

- Frequency-based

- Collection frequency $f(w)$ (i.e. number of occurrences in all documents in the collection)
- Document frequency
- Frequency of capitalized words
- Weirdness
- TFIDF

- Rating-based
  - Deviation from the average score
  - Word score variance
  - Sentiment category likelihood for each *(word, category)* pair

We will consider some of them in more detail.

**Frequency of capitalized words**. The meaning of this feature is the frequency (in the review corpus) of each word starting with the capital letter and not located at the beginning of the sentence. With this feature we are trying to identify potential proper names, which are always neutral.

**Weirdness.** To calculate this feature two collections are required: one with high concentration of sentiment words and the other – contrast one. The main idea of this feature is that sentiment words will be «strange» in the contexts of the contrast collection. This feature is calculated as follows (Ahmad et al., 1999):

$$Weirdness = \frac{P_s(w)}{P_g(w)}$$

where $P_s(w)$ – probability of the word in a special corpus, $P_g(w)$ – probability of the word in a general corpus. Here and further we consider maximum likelihood estimation of the probabilities. Instead of the collection frequency one can use the document frequency for the probability calculation.

Weirdness was calculated using the following collection pairs: *opinion-news*, *opinion-description, description-news* with document frequency and *small-description*, *opinion-description* with collection frequency.

**TFIDF**. We use TFIDF variant described in (Callan et al., 1992), based on BM25 function. We calculate TFIDF using the collection pairs: *small-news*, *small-description*, *opinion-news*, *opinion-description, description-news.*

## 3.4 Rating-based features

As we mentioned above we had collected user's numerical score (on a ten point scale) for each review. Let $C = \{1...10\}$ to be the set of rating categories in the review collection. First, we want to give some definitions, which we will use further.

**Definition 1.**

i. The probability of a rating category **c** given a word ***w***:

$$P(c \mid w) = \frac{f(w,c)}{\sum_{c_i \in C} f(w,c_i)}$$

ii. The probability of a word ***w*** given a rating category **c**:

$$P(w \mid c) = \frac{f(w,c)}{\sum_{w_i \in C} f(w_i,c)}$$

**Definition 2.**

i. An expected category for a given word:

$$E(c \mid w) = \sum_{c_i \in C} c_i \cdot P(c_i \mid w)$$

ii. An expected category in the review collection:

$$E(c) = \sum_{c_i \in C} c_i \cdot P(c_i)$$

Using our definitions we suggest the following features:

**Deviation from the average score.**

$$Dev(w) = \left| E(c \mid w) - E(c) \right|$$

This feature can discriminate words appearing in a wide range of rating categories.

**Word score variance.** One more useful predictor is word score variance. If a word has small variance then it might be used in reviews with similar scores and has high probability to be a sentiment word.

$$Var(w) = E(c^2 \mid w) - E(c \mid w)^2$$

**Scaled likelihood.** To get some intuition about how likely a word is to appear in each sentiment class we define a scaled log-likelihood:

$$Lhc(w) = \log \frac{P(w \mid c)}{P(w)}$$

Scalability is required to be comparable between words. We have also added some features aggregating *Lhc values* like maximum and average.

## 3.5 Morphological Features

Some linguistic features were also added to our system because they can play crucial role in improving the sentiment lexicon extraction.

- Four binary features indicating the word part of speech (noun, verb, adjective and adverb)

- Two binary features reflecting POS ambiguity (i.e. word can have various parts of speech depending on a context) and the feature indicating if this word is recognized by the POS tagger.
- Predefined list of prefixes of a word (for example, Russian prefixes "*ne*", "*bes*", "*bez*" etc. similar to English "*un*", "*in*", "*im*" etc.)

The last feature is a strong predictor for words starting with negation.

## 3.6 Algorithms and evaluation

To train supervised machine learning algorithms we needed a set of labeled sentiment words. For our experiments we manually labeled words with the frequency greater than three in the movie review collection (18362 words). We marked up a word as a sentiment one in case we could imagine it in any opinion context in the movie domain. All words were tagged by two assessors. If there was a disagreement about the sentiment of a specific word, the collective judgment after discussion was used as a final ground truth. As a result of our assessment procedure we had obtained the list of 4079 sentiment words in the movie domain.

We solved the two class classification problem: to separate all words into sentiment and neutral categories. For this purpose Weka[1] data mining tool was used. We considered the following algorithms: *Logistic Regression*, *LogitBoost* and *Random Forest*. All parameters in the algorithms were set to their default values. For each experiment 10 fold cross-validation was used.

Using this algorithms we obtained word lists, ordered by the predicted probability of their opinion orientation. To measure the quality of these lists the *Precision@n* metric was used. This metric was very convenient for measuring the quality of list combinations and it could be used with different thresholds. To compare quality of the algorithms in different domains we chose n = 1000. This level was not too large for the manual labeling and demonstrated the quality in an appropriate way.

The results of classification are in Table 1.

| **Logistic Regression** | **LogitBoost** | **Random Forest** | **Average** |
|---|---|---|---|
| 75.7% | 75.3% | 72.4% | 81.5% |

TABLE 1 – Precision@1000 of word classification

We noticed that the lists of sentiment words extracted by the algorithms differ significantly. So we decided to average word probability values in these three lists. The result of this summation can be found in the last column of the Table 1.

As the baseline for our experiments we used the lists ordered by frequency in the review collection and deviation from the average score. *Precision@1000* in these lists was 26.9% and 35.5% accordingly. Thus our algorithms gave significant improvements over the baselines. All the other features can be found in Table 2.

[1] http://www.cs.waikato.ac.nz/ml/weka/

Let us look at some examples of sentiment words with the high probability value in the sum list: *Trogatel'nyi (affective), otstoi (trash), fignia (crap), otvratitel'no (disgustingly), posredstvenniy (satisfactory), predskazuemyi (predictable), ljubimyj (love) etc.*

| Feature | Collection | Precision @1000 |
|---|---|---|
| TFIDF | small – news | 38.5% |
| TFIDF | small – descr | 36.4% |
| TFIDF | review – news | 30.5% |
| TFIDF | review – descr | 39.8% |
| Weirdness | review – news (doc. count) | 31.7% |
| Weirdness | review – descr (doc. count) | 48.1% |
| Weirdness | small – descr (frequency) | 49.1% |
| Weirdness | review – descr (frequency) | 46.6% |
| Dev | review | 35.5% |
| Var | review | 21.5% |
| Lhc | review | 33.0% |
| Frequency | review | 26.9% |
| Frequency | small | 31.9% |
| Document Frequency | review | 27.8% |

TABLE 2 – Precision@1000 for different features

## 4 Model adaptation

In the previous section we described the construction of the sentiment lexicon extraction model for the movie domain. The next step of the current research is utilizing this model in four other domains and combining obtained results to form a general sentiment lexicon for the product meta-domain.

| | Review Collection | Description Collection | Source |
|---|---|---|---|
| **Books** | 23, 883 | 22, 321 | Imhonet |
| **Games** | 7, 928 | 1, 853 | Imhonet |
| **Digital Cameras** | 10, 208 | 920 | Yandex Market |
| **Mobile Phones** | 30, 620 | 890 | Yandex Market |

TABLE 3 – The characteristics of the data collections

### 4.1 Additional datasets

We collected[2] data in the four domains: books, computer games, mobile phones and digital cameras. The structure of the datasets is the same as for movie domain. Data collection characteristics for each domain can be found in Table 3.

In further experiments we use the same **news corpus** as for movie domain.

### 4.2 Model utilization and evaluation

For all words in a particular field (excluding low frequent ones) we computed feature vectors (see Sections 3.3-3.5) and constructed a domain word-feature matrix. We applied our classification model, which was trained in the movie domain, to these word-feature matrixes and manually evaluated the first thousand of the most probable sentiment words in each domain. The results of the evaluation are in Table 4.

| | **Average** |
|---|---|
| **Books** | 86.0% |
| **Games** | 72.2% |
| **Digital Cameras** | 62.0% |
| **Mobile Phones** | 73.2% |

TABLE 4 – The results of domain adaptation

Despite the drop in some other domains the quality of sentiment word extraction continues to be much higher than the quality level of single features (Table 2). So we can conclude that the sentiment lexicon extraction model is robust enough to be transferred to other domains.

## 5 Developing the Russian lexicon for product meta-domain

To construct the general sentiment lexicon for products and services we combine sentiment word lists from five domains. We want to boost words that occur in many different domains and have high weights in each of them. We propose the following function for the word weight in the resulting list:

$$R(w) = \max_{d \in D}(prob_d(w)) \cdot \sum_{d \in D} \frac{1}{|D|} \cdot \left(1 - \frac{pos_d(w)}{|d|}\right)$$

where D – is the domain set with five domains, $d$ is the sentiment word list for a particular domain and $|d|$ is the total number of words in this list. Functions $prob_d(w)$ and $pos_d(w)$ are the sentiment probability and position of the word in the list $d$ .

The *Precision@1000* of the obtained sentiment word list is **91.4%.** The inter-rater agreement between the two Russian annotators is measured at 0.84 (κ = 0.63).

[2] Review data collections in the book and digital camera domains are obtained from Russian Seminar of Information Retrieval Methods (www.romip.ru)

As a baseline for our method of construction of the general sentiment lexicon for product meta-domain, we take the combined *weirdness* list (review – descr) as rather simple, but high quality one. We construct it from weirdness lists in the same manner as described in the beginning of the section. The Precision@n plots of the extracted lexicon and weirdness list combination are depicted on Figure 1.

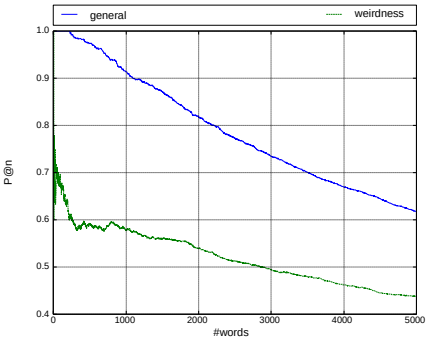


FIGURE 1 – Precision@n depending on #words

The first ten most probable sentiment words are: *bespodobniy (matchless), kleviy (cool), obaldenniy (astounding), neponiatniy (incomprehensible), neprivichniy (unusual), srednenkiy (mediocre), posredstvenniy (moderate), neploho (not bad), otlichneishiy (splendiferous), nenuzhniy (unnecessary).* This sentiment lexicon is clean enough to be used in various sentiment analysis tasks.

This meta-domain list of sentiment words consists of words really used in users' reviews and its creation does not require any dictionary resources. We plan to make it available for further research in sentiment analysis of Russian texts.

## 6 Lexicon evaluation on the cross-domain sentiment classification task

### 6.1 Experimental setup

To evaluate usefulness of our meta-domain sentiment list we test it in the cross-domain sentiment classification task as described for example in (Blitzer et al., 2007; Bollegala et al., 2011; Pan et al., 2010). In these studies the dataset consisting of Amazon product reviews for four different product types (books (B), DVDs (D), electronics (E) and kitchen appliances (K)) is used. There are 1000 positive and 1000 negative reviews selected randomly and labeled for each domain. Domain-adaptation algorithms are trained on the one domain (source domain) and tested on the other domain (target domain).

We do not compare our approach with these approaches because we do not make any efforts to adapt a classifier to a new domain. We use the similar setup to show the

generalization abilities of the sentiment word lists. In these experiments we try to demonstrate the influence of our meta-domain list on the sentiment classification quality in a new domain without any labeled data.

So we randomly take 1000 positive and 1000 negative labeled Russian reviews from four domains: movies (**M**), books (**B**), mobile phones (**P**) and digital cameras (**C**). The reviews with user's score 9-10 are considered as positive and reviews with authors' score 1-4 are considered as negative.

Taking pairs of the domains, we train a sentiment classifier in one domain (source domain) and then transfer the classifier to the other domain (target domain). We treat a review text as a bag-of-words and use the following features for classification:

- All frequent words of the source domain (**F**ull **L**ist),
- Sentiment words from the generated sentiment lexicon of the source domain (**S**ource **D**omain **L**exicon),
- Words from the meta-domain sentiment lexicon, excluding the sentiment vocabulary of the target domain during the extraction (**G**eneral **L**exicon).

In this task we utilize the LIBLINEAR realization of the support vector machine (SVM) classification algorithm with the default parameter values.

Additionally we include TFIDF weights for each feature, as it is pointed to give higher quality of the classification in comparison with the binary weights and we also take into account the polarity influencers, which can revert or magnify the polarity of the following words. The specific details can be found in (Chetviorkin & Loukachevitch, 2011).

We performed experiments with the proposed feature sets on the 9 domain pairs: ***B→C, M→C, P→C, B→P, M→P, C→P, M→B, P→B, C→B*** where the letter before an arrow corresponds with the source domain and the letter after an arrow corresponds with the target domain. We do not consider cross-domain sentiment classification with the movie domain as a target one, because we manually labeled and trained the sentiment word extraction model in it, and the results of the classification can be unclear.

For domain specific and general sentiment lexicons we explored different word quantity thresholds: {1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000} and report the results with each of them (see Figure 2 and 3).

## 6.2 Metrics

We denote by $A(S,T,L)$ the accuracy obtained during the transfer from source domain ***S*** to target domain ***T*** of the sentiment classifier trained using the lexicon ***L***. The main point of comparison in the current research is the accuracy $A(S,T,FL)$, which corresponds to the accuracy obtained by the *baseline* lexicon, i.e. all frequent words from the source domain.

Thus we can define the main measure in the current experiment:

$$\Delta(S,T,L) = A(S,T,L) - A(S,T,FL)$$

This is the difference between the accuracy obtained with the lexicon L and baseline lexicon FL, during the transfer from source domain S to target domain T. We also use the averaged variant of this measure:

$$\overline{\Delta}(L) = \frac{1}{|D|} \sum_{(S,T) \in D} \Delta(S,T,L)$$

In our case $|D| = 9$.

### 6.3 Main results

We report all results in this section using first 4000 words in the general lexicon and domain specific lexicons. This is the maximum amount of words with rather reliable intrinsic precision values ~70% in the general lexicon (see Section 5). We also provide the results of cross-domain sentiment classification quality with the other threshold values in general and domain specific lexicons on the Figure 2 and 3.

On all tasks the general sentiment lexicon performs on bar or better than the other feature sets. In Table 5 and Table 6, we summarize the comparison results of cross-domain classification using different feature sets.

| **A** | **B->C** | **M->C** | **P->C** | **B->P** | **M->P** | **C->P** | **M->B** | **P->B** | **C->B** |
|---|---|---|---|---|---|---|---|---|---|
| FL | 74.0 | 72.55 | 78.65 | 70.05 | 70.5 | 79.9 | 78.15 | 65.1 | 66.5 |
| SDL | 76.1 | 75.2 | 75.55 | 73.45 | 71.15 | 78.85 | 79.0 | 64.3 | 66.9 |
| GL | 76.1 | 75.7 | 81.9 | 73.35 | 72.55 | 79.8 | 78.05 | 66.6 | 67.2 |

TABLE 5 – The accuracy of cross-domain classification

| $\Delta$ | **B->C** | **M->C** | **P->C** | **B->P** | **M->P** | **C->P** | **M->B** | **P->B** | **C->B** | $\overline{\Delta}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| SDL | 2.1 | 2.65 | -3.1 | 3.4 | 0.65 | -1.05 | 0.85 | -0.8 | 0.4 | 0.57 |
| GL | 2.1 | 3.15 | 3.25 | 3.3 | 2.05 | -0.1 | -0.1 | 1.5 | 0.7 | 1.76 |

TABLE 6 – The difference with baseline of cross-domain classification

The results demonstrate the effectiveness of the general meta-domain sentiment lexicon. In the Table 6 one can see that for some domain pairs our lexicons show significantly better results than the baseline. The average difference over all domain pairs between **FL** (baseline) and **GL** is 1.76%.

In some domain pairs the difference is very small or even negative. We connect this issue with the similarity of the domain lexicons in general (Ponomareva & Thelwall, 2012) and sentiment lexicons in particular. Sometimes sentiment words from one domain can be utilized in the other one, but not vice versa.

We suppose that such a general lexicon for the product meta-domain can serve as a good source of sentiment seed words to generate domain-specific vocabularies in a lot of specific domains.

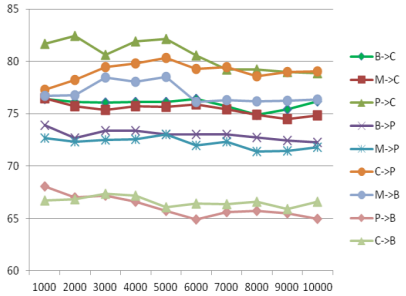


FIGURE 2 – The dependence of the classification quality on the threshold in the general lexicon

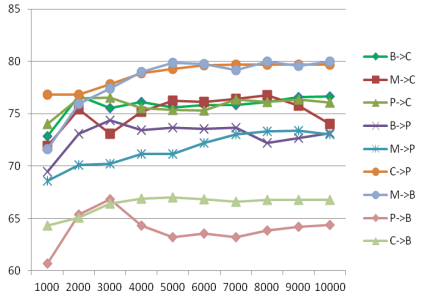


FIGURE 3 – The dependence of the classification quality on the threshold in the domain specific lexicons

## Conclusion and perspectives

In this paper, we described a method for sentiment lexicon extraction for any domain on the basis of several domain-specific text collections. We utilized our algorithm in different domains and showed that it had good generalization abilities. We combined sentiment lexicons from various domains and constructed the general meta-domain sentiment lexicon for products and services. This lexicon was evaluated intrinsically, with P@1000 = 91.4% and extrinsically in the cross-domain classification task. The sentiment classification algorithm based on the meta-domain sentiment lexicon outperformed all baselines and proved usefulness of the constructed resource. Besides, this meta-lexicon can be a useful source of sentiment seeds for sentiment lexicon extraction in new domains of products and services.

We extracted such a general lexicon for Russian language, for which sentiment analysis resources practically do not exist. We plan to make our general lexicon for the product meta-domain publicly available.

## Acknowledgments

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

## References

Abdul-Mageed M., Diab M., Korayem M. (2011). Subjectivity and Sentiment Analysis of Modern Standard Arabic. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, number 3, pp. 587-591.

Ahmad K., Gillam L., Tostevin L. (1999). University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval *In the Proceedings of Eigth Text Retrieval Conference (Trec-8)*.

Aue A. and Gamon M. (2005). Customizing sentiment classifiers to new domains: A case study. In *International Conference on Recent Advances in Natural Language Processing*, Borovets, BG.

Banea C., Mihalcea R., Wiebe J. and Hassan S. (2008). Multilingual subjectivity analysis using machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Blitzer J., Dredze M., Pereira F. (2007) Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of ACL 2007*, pp. 440–447.

Bollegala D., Weir D. and Carroll J. (2011) Using multiple sources to construct a sentiment sensitive thesaurus for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, Portland, Oregon. pp. 132-141.

Callan J.P., Croft W.B., Harding S.M. (1992). The INQUERY Retrieval System. In Proceedings of *3rd International Conference on Database and Expert Systems Applications* / A.M. Tjoa and I. Ramos (eds.). – Springer Verlag, New York, pp.78-93.

Chetviorkin I. and Loukachevitch N. (2011). Three-way movie review classification. In *Proceedings of the International Conference on Computational Linguistics Dialog*, pp 177-186.

Choi Y. and Cardie C. (2009). Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 590–598.

Clematide S., Klenner S. (2010) Evaluation and extension of a polarity lexicon for German. In *WASSA-workshop held in conjunction with ECAI-2010*, pp 7-13.

Hatzivassiloglou V. and McKeown K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of ACL-97*, pp. 174–181, Madrid, ES.

He B., Macdonald C., He J., and Ounis I. (2009). An effective statistical approach to blog post opinion retrieval. In *Proceedings of the 17th ACM CIKM*, pp. 1063-1072.

Jijkoun V., de Rijke M. and Weerkamp W. (2010). Generating focused topic-specific sentiment lexicons. In *Proceedings of ACL '10*, pp. 585–594.

Kanayama H. and Nasukawa T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *EMNLP '06*, pp. 355–363, Morristown, NJ, USA.

Lau R., Lai C., Bruza P. and Wong K. (2011). Pseudo Labeling for Scalable Semi-supervised Learning of Domain-specific Sentiment Lexicons. In *20th ACM Conference on Information and Knowledge Management*.

Mihalcea R., Banea C. and Wiebe J. (2007). Learning multilingual subjective language via cross-lingual projections. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 976–983, Prague, Czech Republic.

Pan S. J., Ni X., Sun J-T, Yang Q. and Chen Z. (2010). Cross-Domain Sentiment Classification via Spectral Feature Alignment. In *Proceedings of the World Wide Web*

*Conference*. pp. 751-760, New York, USA.

Pang B., Lee L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. *Now Publishers*.

Perez-Rosas V., Banea C. and Mihalcea R. (2012). Learning Sentiment Lexicons in Spanish. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Ponomareva N. and Thelwall M. (2012): Bibliographies or blenders: Which resource is best for cross-domain sentiment analysis? In *Proceedings of the 13th Conference on Intelligent Text Processing and Computational Linguistics*.

Qiu G., Liu B., Bu J. and Chen C. (2011). Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, 37(1).

Read J., Carroll J. (2009). Weakly Supervised techniques for domain independent sentiment classification. In *Proceedings of the first International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, pp. 45-52.

Steinberger J., Lenkova P., Ebrahim M., Ehrmann M., Hurriyetogly A., Kabadjov M., Steinberger R., Tanev H., Zavarella V. and Vazquez S. (2011). Creating Sentiment Dictionaries via Triangulation. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011*, pp. 28–36,

Taboada M., Brooke J., Tofiloski M., Voll K. and Stede M. (2011). Lexicon-based methods for Sentiment Analysis. *Computational linguistics*, 37(2).

Wiebe J. and Riloff E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of CICLing 2005*. pp. 486-497.

Wu Q., Tan S. and Cheng X. (2009). Graph ranking for sentiment transfer. In *Proceedings of ACL-IJCNLP 2009*, pp. 317–320.