# Extraction and Use of Opinion Words for Three-Way Review Classification Task

Ilia Chetviorkin[1], Natalia Loukachevitch[2],

[1] Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University
ilia2010@yandex.ru
[2] Research Computing Center Lomonosov Moscow State University
louk_nat@mail.ru

**Abstract.** In this paper, we consider a three-way classification approach for Russian movie reviews. All reviews are divided into groups: “thumbs up”, “so-so” and “thumbs down”. To solve this problem we use various sets of words together with such features as opinion words, word weights, punctuation marks and polarity influencers that can affect the polarity of the following words. Besides, we estimate the maximum upper limit of automatic classification quality in this task.

**Keywords:** Opinion words, rating-inference problem

## 1 Introduction

The web is full of customers’ opinions on various products. Automatic collection, processing and summarization of such opinions are very useful for future users. Opinions about the products are often expressed using evaluative words and phrases that have a certain positive or negative sentiment. Therefore, important features in the qualitative classification of opinions about a particular entity are opinion words and expressions used in this domain. The problem is that it is impossible to compile a list of opinion expressions, which will be applicable to all domains, as some opinion phrases are used only in a specific domain; while the others are domain-oriented.

There are two main approaches to the automatic identification of opinion words in texts. The first approach is based on information from a dictionary or a thesaurus. In this approach a small initial set of words is usually chosen manually, and then expanded with the help of dictionaries and thesauruses entries. The basic principle of this approach is that if a word has sentiment polarity, then its synonyms and antonyms have polarity too (orientation may change). Therefore, from the initial set of words, a new, more complete set of opinion words can be constructed [5]. In [4], dictionary definitions are used for opinion words extraction. The basic idea is that words with the same orientation have "similar" glosses.

The second approach – corpus based training. This approach is based on finding rules and patterns in the texts. In [14] word polarity is calculated by comparing the co-occurrence statistics of various words with words “excellent” and “poor”. Authors assume that words with similar semantic orientation tend to co-occur. The resulting opinion orientation of the words is used to classify reviews to positive and negative.

In this article we will use our method for automatic opinion words extraction based on several text collections, which can be automatically built for many domains. The set of text collections includes: a collection of products reviews with author evaluation scores, a text collection of product descriptions and a contrast corpus (for example, general news collection).

The task of review ranking according to their sentiment has different subtasks. The easiest subtask is to classify reviews into two classes: positive and negative. The quality of two-way classification using topic-based categorization approach for reviews exceeds 80% [11]. In [15] the quality of review classification, based on the so-called appraisal taxonomy, was described as 90.2%.

However, when we turn to the problem of review division into three classes («thumbs up», «thumbs down», «so-so»), the quality of automatic classification decreases significantly [9]. This is partly due to the subjectivity of human evaluation. In [10] the authors conducted a study on the possibility of a human to distinguish reviews rated on a ten-point scale. They describe that if the difference between review scores is more than three points, the accuracy is 100%, two – 83%, one point – 69% and zero points, correspondingly, 55%. Thus, if we classify reviews into a large number of classes, even a human will show low classification accuracy.

In addition, in that paper the difference between evaluation styles of various people was indicated: a review estimated in 5 points (on a ten-point scale) by one person, may express the same opinion and be estimated as 7 points by the other [10]. It was shown that after adjustment to an individual author's style, the quality of the classification increased significantly and reached 75%. But in the classification of 5394 reviews from a large number of authors (494), the achieved accuracy was 66.3%.

In this paper, we analyze various features to improve three-way classification of movie reviews in Russian. For Russian language, studies of this task practically *do not exist*.

We used the following classification features:

- extracted opinion words,
- word weights based on different sources,
- use of polarity influencers: they may reverse or enhance (not, very) polarity of other words,
- length and structure of reviews,
- punctuation marks – as for example in [13] authors used punctuation to reveal sarcastic sentences.

## 2 Extraction of Opinion Words

For our experiments, we chose movie domain. We collected 28773 film reviews of various genres from 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 corpus***.

Example of the review:

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

We also needed a contrast collection of texts for our experiments. In this collection concentration of opinions should be as little as possible. For this purpose, we had collected 17680 movie descriptions. This collection was named ***description corpus***.

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

### 2.1 Collection with higher concentration of opinion words

We suggested that it was possible to extract some fragments of the reviews from **review corpus**, which had higher concentration of opinion words. These fragments include:

- Sentences ending with a «!»;
- Sentences ending with a «…»;
- Short sentences, no more than 7 word length;
- Sentences containing the word «movie» without any other nouns.

We call this corpus – ***small corpus***.

### 2.2 Proposed features

The main task was automatic creation of the opinion word list based on the calculation of various features. Further we exploit the following set of features.

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

$$\text{Weirdness} = \frac{w_s / t_s}{w_g / t_g}\text{, where}$$

$w_s$ – frequency of the word in special corpus, $t_s$ – total count of words in special corpus, $w_g$ – frequency of the word in general corpus, $t_g$ – total count of words in general corpus. Instead of frequency one can use the number of documents where the word occurs.

**TFIDF.** There are many varieties of this feature. We used TFIDF variant described in [1] (based on BM25 function [8]):

$$\text{TFIDF} = \beta + (1-\beta)\cdot \text{tf}\cdot \text{idf} \qquad (1)$$

$$tf_D(l) = \frac{\text{freq}_D(l)}{\text{freq}_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{\text{avg\_dl}}} \quad idf(l) = \frac{\log\left(\frac{|c|+0.5}{df(l)}\right)}{\log(|c|+1)}$$

- freq(l) – number of occurrences of l in a document,
- dl(l) – length measure of a document, in our case, it is number of terms in a review,
- avg_dl – average length of a document,
- df(l) – number of documents in a collection (e.g. movie descriptions, news collection) where term l appears,
- β = 0.4 by default,
- |c| – total number of documents in a collection.

**Deviation from the average score.** As we mentioned above we had collected user’s numerical score (on a ten point scale) for each review. The main idea of this feature is to calculate average value for each word (sum of review ratings where this word occurs divided into their count) in the collection and then subtract average score of all reviews in the collection from it. Absolute value of this variable is what we need.

$$dev(l) = |\frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n}|$$

$$\sum_{i=1}^{n} k_i = k$$

where l – considered lemma, n – total count of the reviews in the collection, mi – i-th review score, $k_i$ – frequency of the lemma in the i-th review (may be 0).

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

### 2.3 Feature and collection combinations

For our experiments we took top ten thousand words ordered by frequency from the review corpus. We were not interested in words, which occur in the collection only a few times.

Then we divided these words into two groups: adjectives and not-adjectives. The sense of such division is that the majority of opinion words are adjectives and quality evaluation of our approach on them was our special interest. The not-adjective group of words contains nouns, verbs and adverbs. All features were calculated for two above mentioned groups independently.

Thus, we had the following combinations of features and collections:

- TFIDF calculation using the pairs of collections: *small-news*, *small-description*, *opinion-news*, *opinion-description*;
- Weirdness calculation using the pairs of collections: *opinion-news and opinion-description* with document count and *small-description*, *opinion-description* with frequency;
- Deviation from the average score;
- Word frequency in *opinion* and *small collections*;
- Total number of documents in the *opinion corpus*, where the word occurs;
- Frequency of capitalized words.

In addition, separately for description corpus we calculated the following features: frequency, document count, weirdness using *description-news* collections with document count and TFIDF using the same pair. Thus, each lemma had 17 features.

### 2.4 Data preparation and algorithms

To train supervised machine learning algorithms we needed a set of labeled opinion words. Firstly we tried to extract a list of domain-independent opinion words (near 500) from RuThes thesaurus [7]. The results were unsatisfactory and did not meet expectations as far as this list of general opinion words did not contain domain-specific words, slang words et al. Therefore, we decided to label the full list of ten thousand words manually and then use cross-validation. We marked up word as opinion one in case we could imagine it in any opinion context in the movie domain. All words were tagged by two authors.

In the issue of our mark up we had the list of 3200 opinion words (1262 adjectives, 296 adverbs, 857 nouns, 785 verbs).

Our aim in this part of work was to classify words into two classes: opinion or not opinion. For this purpose we used Rapid Miner[1] data mining tool. We considered the following algorithms: K nearest neighbors (kNN), Naïve Bayes, Perceptron, Neural Network (2 and 3 layers), Logistic Regression and SVM (with linear and radial kernels). For all experiments we used 10 fold cross-validation.

As a result the best algorithms were Logistic Regression for adjectives (F-measure: 68.1%) and Neural Net with 2 layers for not-adjectives (F-measure: 50.9%, unbalanced data). Using them we obtained term lists (adjectives and not adjectives), ordered by the predicted probability of their opinion orientation.

Let us look at some examples of opinion words with high probability value:

**Adjectives:** *dobryj (kind), zamechatel'nyj (wonderful), velikolepnyj (gorgeous), potrjasajushij (stunning), krasivyj (beautiful), smeshnoj (funny), ljubimyj (love)* etc.,

**Not-adjectives***: fuflo (trash), naigranno (unnaturally), fignja (junk), fil'm-shedevr (masterpiece film), tufta (rubbish)* etc.

Obtained opinion word lists with probabilities we use in our review classification task.

[1] http://rapid-i.com/

# 3 Features for review classification.

In this section we will consider some additional features, which can help us to solve three-way classification problem.

## 3.1 Word weights

As the main elements of a feature set we used lemmas (words in the normal form) mentioned in the reviews. Word weights can be binary and reflect only word presence in a review or TFIDF formula can be used.

TFIDF is the most popular method of word weighting in information retrieval [8]. For each term in a text, its TFIDF weight can be represented by multiplication of two factors: TF that defines the frequency of this term in the text and IDF specifying occurrence of the term in documents of a text collection. The more frequently such occurrences are, the smaller resulting IDF will be [8]. TF and IDF factors can be defined by various formulas. We used two variants of TFIDF for calculation.

First, we used the simplest form of TFIDF [8]:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k} \qquad \mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|} \quad (2)$$

- $n_i$ is the number of occurrences of a term in a document, and the denominator is the sum of occurrence number of all terms in the document,
- |D| – total number of documents in a collection,
- $|(d_i \supset t_i)|$ – number of documents where term $t_i$ appears (that is $n_i \neq 0$).

In addition, we used TFIDF variant, which was described in Section 2.2.

## 3.2 Polarity influencers

Intuitive is the fact that there are some words, which can affect polarity of other words – *polarity influencers*. To find them the manually compiled set of opinion words (3200 units) was used (see Section 2.4). From the review corpus, we automatically extracted words directly preceding the manually labeled opinion words and ordered them by decreasing frequency of their occurrence.

Then from the first thousand of words from this list, potential polarity influencers were manually chosen (74 words). To assess how significant the effect of these polarity influencers can be, the following procedure was made: we calculated the average score of opinion words in two cases, when they follow the potential polarity influencers and when they occur without them. The average score of a word is the average value of numerical scores of reviews where this word occurs.

After comparison of these average scores, two significant groups of polarity influencers were discriminated. If an opinion word had the high average score (>8) and changed it to the lower when used after a given polarity influencer, and an opinion word with the low average score (<6.7) changed it to the higher one, it means that this polarity influencer *reverses* word polarity (operator –).

If after a polarity influencer, an opinion word with the high score increased its average score, and an opinion word with the low average score decreased its score, it means that this polarity influencer *magnifies* polarity of other words (operator +).

In our review corpus, we found the following polarity influencers:

operator (–): *net (no), ne (not);*

operator (+): *polnyj (full), ochen' (very), sil'no (strongly), takoj (such), prosto (simply), absoljutno (absolutely), nastol'ko (so), samyj (the most).*

On the basis of this list of polarity influencers we substituted sequences *"polarity influencer_word"* using special operator symbols («+» or «–») depending on an influencer, for example:

*NE HOROSHIJ (NOT GOOD) → –HOROSHIJ (– GOOD)*

*SAMYJ KRASIVYJ (THE MOST BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)*

*NASTOL'KO KRASIVYJ (SO BEAUTIFUL) → + KRASIVYJ (+ BEAUTIFUL)*

Modified lemmas were added to the feature set. Now if in a text a word with a polarity influencer occurs, then only the corresponding modified lemma would be added to the review's vector representation, but not both words. This allows us to take into account the impact of polarity influencers.

### 3.3 Review length and structural features

Movie reviews can be long or short. We chose a threshold on the review length to be 50 words. If a review is long, it often contains overall assessment for a movie at the beginning or at the end. This was the basis for separate consideration of short and long reviews and dividing long reviews into three parts: the beginning (first sentences of a review with total length less than 25 words), the end (last sentences of a review with total length less than 25 words) and the middle (all that is left). We classified each part separately and then aggregated obtained scores in various ways (voting, average).

### 3.4 Punctuation marks

In addition we included punctuation marks «**!**», «**?**», «**…**» as elements of the feature set.

## 4 Review classification experiments

Reviews in the working dataset (***opinion collection***) are provided with authors' scores from 1 to 10 points. To map from the ten-point scale to the three-point scale we used the following function: {1-6} → «**1**» (thumbs down), {7-8} → «**2**» (so-so), {9-10} → «**3**» (thumbs up). Thus, the number of reviews belongs to class «3» is approximately 45% of the total.

All reviews from the collection were preprocessed by a morphological analyzer and lemmas with part of speech tagging were extracted.

Authors of previous studies almost unanimously agreed that Support Vector Machine (SVM) algorithm works better for text classification tasks (and review classification task in particular). We also decided to use this algorithm. In view of the fact that we had a large amount of data and features, library LIBLINEAR was chosen [12]. This library had sufficient performance for our experiments. To obtain statistically significant results five fold cross-validation was used. All other parameters of the algorithm were left in accordance with their default values.

We used the following word sets in our classification experiments:

1. An optimal set of opinion words produced by the method described in Section 2. From the list of adjectives and not adjectives (ordered by the probability of their opinion orientation – ***opinweight***) we selected the optimal adjectives and not-adjectives combination. We iterated over the words in these lists and compared quality of classification. We denote this experimental set ***OpinCycle***,
2. Set of words, which was used in [3] to achieve the best results (***OpinContrast***). This set contains near 500 the most frequent words with high opinion probability weight and 400 words with the highest TFIDF score calculated using review and news collections (see Section 2.2),
3. Set of opinion words (3200 units), obtained by manual labeling by two experts (see Section 2.4) (***OpinIdeal***),
4. Set of all words occurring in the review corpus four or more times (***BoW***). The set includes prepositions, conjunctions and particles as well.

From all these word sets, we chose one set, which yields the best classification accuracy, and analyzed the effect of other features: word weights (***tfidf***), opinion weights (***opinweight***), punctuation marks (***punctuation***), polarity influencers (***operators***), review length (***long*** and ***short***).

TFIDF word weights were calculated relying on two formulas: the most well known formula (2) (***tfidf simple***) and formula (1) (***tfidf)*** (see Section 3.1). IDF factor was calculated on the basis not only the *review corpus*, but also two other collections: the *news corpus* (***tfidf news)*** and the *description corpus* (***tfidf descr)***.

To assess the quality of classification we used *Accuracy measure*. It is calculated as the ratio of correct decisions taken by the system to the total number of decisions [2].

The results of the algorithm using different sets of words and features are listed in Table 1. It is worth mentioning that different sets have different coverage area. All reviews without any features from the set were considered as strongly positive ("thumbs up") in accordance with the review distribution between classes. The basic weight of each word is its presence in a review.

**Table 1.** The classification results using various features

| Feature Set | Feature Count | Accuracy % |
|---|---|---|
| ***OpinCycle*** | 1000 *adj* + 1000 *not-adj* | 58.00 |
| ***OpinContrast*** | 884 | 60.33 |
| ***OpinIdeal*** | 3200 | 57.62 |
| ***BoW*** | 19214 | 57.37 |
| ***OpinCycle + tfidf simple*** | 1000 *adj* + 1000 *not-adj* | 59.13 |
| ***OpinContrast + tfidf simple*** | 884 | 59.43 |
| ***OpinIdeal + tfidf simple*** | 3200 | 59.72 |
| ***BoW + tfidf simple*** | 19214 | *62.52* |
| ***BoW + tfidf*** | 19214 | 61.71 |
| ***BoW + tfidf descr*** | 19214 | 61.74 |
| ***BoW + tfidf news*** | 19214 | 62.90 |
| ***BoW + tfidf news + operators*** | 22218 | 63.46 |
| ***BoW + tfidf news + punctuation + operators*** | 22221 | 63.17 |
| ***BoW + tfidf news + opinweight + operators*** | 22218 | **64.48** |
| ***BoW + tfidf news+ opinweight + operators + short*** | 22218 | 63.56 |
| ***BoW + tfidf news + opinweight + operators + long*** | 22218 | 62.37 |
| ***BoW + tfidf news + opinweight + operators + avg*** | 22218 | 63.14 |

The results obtained by using ***BoW + tfidf simple*** were taken as a *basic line*. The best results were obtained using bag of words (***BoW***) with TFIDF, opinion weights and polarity influencers. This is clear improvement over *62.52* where ***BoW + tfidf simple*** is applied; indeed the difference is highly statistical significant ($p < 0.001$, $\alpha = 0.05$, Wilcoxon signed-rank test/Two-tailed test). Punctuation marks did not give any quality improvement, although their usage gave slightly better coverage. Formula (1) usage gives slightly better quality than the second one (2). The choice of the news corpus for IDF calculation in (1) draws better results than using the description corpus (***BoW + tfidf descr***) and the review corpus (***BoW + tfidf***).

To increase weights of opinion words in contrast with the other words we used the list of opinion words with probability weights from 0 to 1 (see Section 2.4). We took 800 the most probable adjectives and 200 not-adjectives (we have tried another combinations also) as opinion words. All other words from the feature set were considered with zero ***opinweight***. We modified the weight of each word in the feature vectors in the following manner:

$$\text{wordweight}(x) = \text{TFIDF}(x)\cdot e^{(\text{opinweight}(x) - 0.5)} \qquad (3)$$

Thus, we want to increase weights of the words with high ***opinweight***, and decrease for the other words.

The classification accuracy for short reviews (***BoW + tfidf news + opinweight + operators + short***) is better than for long one (***BoW + tfidf news + opinweight + operators + long***). Although, in average (in accordance with review number in each part) the results were not improved (***BoW+ tfidf news + opinweight + operators + avg***).

For the method with the best results of classification ***BoW + tfidf news + opinweight + operators***, we made additional evaluation with so-called *soft borders*, that is if in the basic scale the author of a review puts a boundary score («8» or «6»), then classification of this review as either class «3» or «2» in case of basic «8», and class «2» or «1» in case of basic «6», was not considered as an error. Such weakening of conditions was made on the assumption that even a human distinguishes boundary classes unsatisfactory. The classification accuracy with *soft borders* reaches **76.48%.**

## 5 Evaluation of reviews by assessors

We also studied the human's ability in three-way review classification. We wanted to know what the maximal quality of classification we could expect from automatic classification algorithms. Significance of such quality upper bound evaluation is declared, for example, in [6]. For a benchmark, we selected one hundred short reviews (with length less than 50 words) and one hundred long reviews (with length more than 50 words) from the review corpus. Assessors did not know the initial score of a review set by its author. Reviews were extracted in such a manner, as to retain original class distribution. All explicit references to the initial score were removed.

Two assessors evaluated the selected reviews. The results of their evaluation are given in Table 4. The last row of the table indicates the agreement in scores between two assessors.

**Table 2.** The results of humans‘ estimation

| Assessor | Assessors accuracy relative to the author of the review | Accuracy with soft borders | Accuracy of the best classification algorithm relative to the assessor |
|---|---|---|---|
| 1 | 72.5 | 86.5 | 69.5 |
| 2 | 72.5 | 78.5 | 63.5 |
| 1 AND 2 | 71.5 | – | – |

Thus, we see that human assessors can reproduce the original scores or be consistent with each other only at the level of 71-72%, which is the absolute upper limit to improve the quality of automatic algorithms. Note that the quality of the automatic classification with soft borders, taking into account the possible ambiguity of the border scores, is 76.48%, which is very close to the classification quality of the second assessor (78.5%).

The percentage of coincident scores between the best algorithm and assessor's scores confirms the results obtained by cross-validation.

## 6 Conclusion

In this paper, we described a method for opinion word extraction for any domain on the basis of several domain specific text collections. We studied the role of obtained opinion words in three-way classification of movie reviews in Russian. The most significant impact on the quality of classification had the choice of TFIDF formula, polarity influencers accounting and opinion words information usage. We estimated the upper limit of classification quality, which is very close to the results of the best automatic algorithm. This fact makes it difficult to reach further quality improvement of automatic three-way review classification.

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

## References

1. Ageev M., Dobrov B., Loukachevitch N., Sidorov A.: Experimental algorithms vs. basic line for web ad hoc, legal ad hoc, and legal categorization in RIRES2004 (in Russian). In: Proceedings of Russian Information Retrieval Evaluation Seminar (2004)
2. Ageev M., Kuralenok I. Nekrestyanov I.: Official RIRES Metrics (in Russian). In: Proceedings of Russian Information Retrieval Evaluation Seminar. Kazan (2010)

3. Chetviorkin I., Loukachevitch N.: Automatic review classification based on opinion words (in Russian). In: Proceedings of Conference on Artificial Intelligence. Tver (2010)
4. Esuli A., Sebastiani F.: Determining the Semantic Orientation of Terms through Gloss Classification. In: Conference of Information and Knowledge Management (2005)
5. Hu M., Liu B.: Mining and Summarizing Customer Reviews. In: KDD (2004)
6. Kilgarriff A., Rosenzweig J.: Framework and results for English Senseval Computers and Humanities. In: Special Issue on SENSEVAL. pp. 15-48 (2000).
7. Loukachevitch N.V., Dobrov B.V., Development and Use of Thesaurus of Russian Language RuThes. In: Proceedings of workshop on WordNet Structures and Standardization, and How These Affect WordNet Applications and Evaluation. (LREC2002) / Dimitris N. Christodoulakis – 2002. pp 65-70. Gran Canaria, Spain (2002).
8. Manning C., Raghavan P., Schütze H.: Introduction to Information Retrieval. Cambridge University Press (2008)
9. Pang B., Lee L.: Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval. Now Publishers (2008)
10. Pang B., Lee L.: Seeing stars: Exploiting class relationships for sentiment categorization with respect of rating scales. In: Proceedings of the ACL (2005)
11. Pang B., Lee L.: Thumbs Up? Sentiment Classification using Machine Learning Techniques. In: Proceedings of EMNLP (2002)
12. Fan R.-E. , Chang K.-W., Hsieh C.-J., Wang X.-R., and Lin C.-J.: LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research 9. pp. 1871-1874 (2008). Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear
13. Tsur O., Davidov D., Rappoport A.: ICWCM – a Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. In: International AAAI Conference on Weblogs and Social Media (2010)
14. Turney P.D.: Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of ACL. pp. 417-424. (2002)
15. Whitelaw C., Garg N., Argamon S.: Using Appraisal Taxonomies for Sentiment Analysis. In: Proceedings of CIKM (2005)