ИССЛЕДОВАНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ ОТЗЫВОВ НА ТРИ КЛАССА

И. И. Четверкин (ilia2010@yandex.ru)
МГУ, Москва, Россия

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

Ключевые слова: отзывы, классы, классификация, обзор.

THREE-WAY MOVIE REVIEW CLASSIFICATION

I. I. Chetverkin (ilia2010@yandex.ru)
Faculty of Computational Mathematics and Cybernetics
Lomonosov Moscow State University, Moscow, Russian Federation

N. V. Loukachevitch (louk_nat@mail.ru)
Research Computing Center, Lomonosov Moscow State University, Moscow, Russian Federation

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 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.

Key words: classification, review, review classification, movie, movie review.

1. Introduction

Actually, users can find any type of information in the Internet. Tentatively, it can be divided into two classes: factual information and user opinions. Most of current information processing techniques (e. g., search engines) work with facts and have satisfactory quality. Processing of user's opinions is a more complicated problem. Ranking of the reviews according to their sentiment is a very difficult and urgent task.

The easiest subtask is to classify reviews into two classes: *positive* and *negative*. Quality of two-way classification using topic-based categorization approach for reviews exceeds 80 % [9]. In [12] 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 [7]. This is partly due to the subjectivity of human evaluation. In [8] 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 to 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 [8]. 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:

- word weights based on different sources,
- single word polarity,
- use of polarity influencers: they may reverse or enhance (*not, very*) polarity of other words,
- length and structure of reviews,
- usage of punctuation marks — as for example in [11] authors used punctuation to reveal sarcastic sentences.

2. Features for review classification

For our experiments, we chose movies' domain. We collected 28 773 film reviews of various genres from online recommendation service *www.imhonet.ru*. For each review, we extracted user's score on a ten-point scale.

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.

2.1. 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 [6]. 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 [6]. 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 [6]:

$$\mathrm{TF} = \frac{n_i}{\sum_k n_k} \quad \mathrm{IDF} = \log \frac{|D|}{|(d_i \supset t_i)|} \tag{1}$$

- 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 described in [1] (based on BM25 function [6]):

$$\boldsymbol{TFIDF\ (l) = \beta + (1 - \beta)\cdot tf(l)\cdot idf(l)}$$

$$tf_D(l) = \frac{\mathrm{freq}_D(l)}{\mathrm{freq}_D(l) + 0.5 + 1.5 \cdot \dfrac{dl_D}{\mathrm{avg\_dl}}} \quad idf(l) = \frac{\log\left(\dfrac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)} \tag{2}$$

- *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, in our case β = 0,
- |c| — total number of documents in a collection.

2.2. Opinion words

We considered opinion words as an important type of features for review classification.

We use the automatically extracted list of opinion words [3]. To generate this list, we exploited four text collections: a movie review collection (review corpus), a collection of film descriptions (description corpus), a special small corpus and a collection of general news. On the basis of these collections we calculated a set of statistical features for words mentioned in reviews. All features were calculated separately for adjectives and not adjectives (verbs, adverbs, nouns). At the next step, we used machine learning to classify terms' feature vectors. As a result we obtained term lists (adjectives and not adjectives), ordered by 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.

In our study of three-way review classification, we used the most probable opinion words and automatically obtained opinion probability weights. In addition, we manually labeled a set of opinion words [3].

2.3. 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 [3]. From the review corpus (see section 2.2), 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.

2.4. 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).

2.5. Punctuation marks

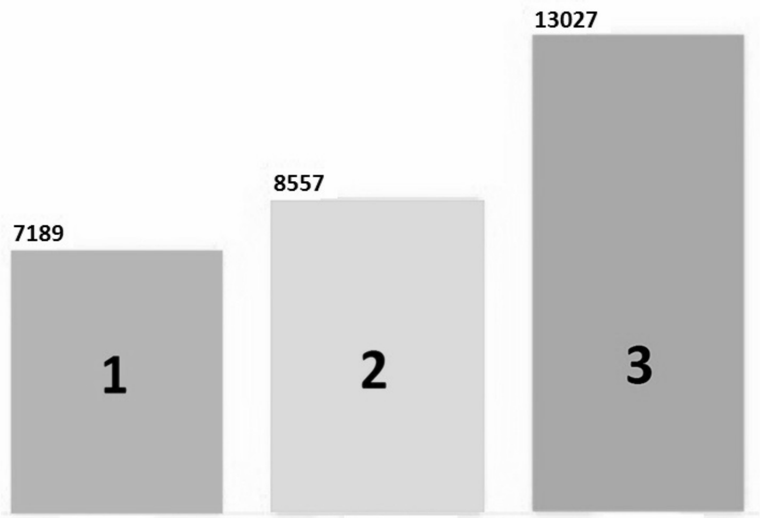
In addition we included punctuation marks «**!**», «**?**», «**...**» as elements of the feature set.

3. Experiments

Reviews in the working dataset 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). The resulting distribution of reviews by grade is shown on Picture 1. 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 [10]. 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.



Pic. 1. The distribution of reviews into three groups by sentiment: "thumbs down"(1),"so-so" (2),"thumbs up"(3)

We used the following word sets in our classification experiments:

- Finding an optimal set of opinion words produced by the method described in Section 2.2. From the list of adjectives and not adjectives (ordered by the probability of their opinion orientation — ***opinweight***) we selected the optimal opinion word combination. We iterated over words in these lists and compared quality of classification. We denote this experiment set ***OpinCycle***,
- set of words, which was used in [4] to achieve the best results (***OpinContrast***). This set contains near 500 the most frequent words with high opinion probability weight [3] and 400 words with the highest TFIDF score calculated using review and news collections (see Section 2.2),
- set of opinion words (3200 units), obtained by manual labeling by two experts (see Section 2.2) (***OpinIdeal***),
- 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 (1) (***tfidf simple***) and formula (2) (***tfidf)*** (see Section 2.1). 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 algorithms 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 review distribution between classes. The basic weight of each word is its presence in a review.

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 (2) usage gives slightly better quality than the first one (1). The choice of the news corpus for IDF calculation in (2) draws better results than using the description corpus (***BoW + tfidf descr***) and the review corpus (***BoW + tfidf***).

Table 1. The classification results using various features

Feature set	**Feature number**	**Accuracy %**
OpinCycle	1000 *adj* + 1000 *not adj*	58.00
OpinContrast	884	60.33
OpinIdeal	3 200	57.62
BoW	19 214	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	19 214	*62.52*
BoW + tfidf	19 214	61.71
BoW + tfidf descr	19 214	61.74
BoW + tfidf news	19 214	62.90
BoW + tfidf news + operators	22 218	63.46
BoW + tfidf news + punctuation + operators	22 221	63.17
BoW + tfidf news + opinweight + operators	22 218	**64.48**

Feature set	Feature number	Accuracy %
BoW + tfidf news+ opinweight + operators + short	22 218	63.56
BoW + tfidf news + opinweight + operators + long	22 218	62.37
BoW + tfidf news + opinweight + operators + avg	22 218	63.14

To increase weights of opinion word in contrast with the other words we used the list of opinion words with probability weights from 0 to 1 (see Section 2.2). 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 ***opinweight*** 0. We modified the weight of each word in the feature vectors in the following manner:

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

Thus, we want to increase weights of the words with high ***opinweight***, and decrese 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 %.**

4. 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 [5]. 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 2. The last row of the table indicates the agreement in scores between two assessors.

Table 2. The results of humans' estimating

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 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.

5. Conclusion

In this paper, we investigated influence of various factors on the quality of 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.* 2004. Experimental Algorithms vs. Basic Line for Web Ad Hoc, Legal Ad Hoc, and Legal Categorization in RIRES2004. RIRES.
2. *Ageev M., Kuralenok I. Nekrest'ianov I.* 2010. Official RIRES Metrics. Kazan: Russian Information Retrieval Evaluation Seminar (RIRES 2010).
3. *Chetverkin I., Loukachevitch N.* 2010. Automatic Extraction of Domain-specific Opinion Words. Dialogue.
4. *Chetverkin I., Loukachevitch N.* 2019. Automatic Review Classification Based on Opinion Words. Tver: Conference on Artificial Intelligence.
5. *Kilgarriff A., Rosenzweig J.* 2000. Framework and Results for English Senseval. Computers and Humanities, Special Issue on SENSEVAL : 15–48
6. *Manning C., Raghavan P., Schütze H.* 2008. Introduction to Information Retrieval.
7. *Pang B., Lee L.* 2008. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval.
8. *Pang B., Lee L.* 2005. Seeing stars: Exploiting Class Relationships for Sentiment Categorization with respect of Rating Scales. Proceedings of the ACL.
9. *Pang B., Lee L.* 2002. Thumbs Up? Sentiment Classification using Machine Learning Techniques. EMNLP.
10. *R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, C.-J. Lin.* 2008. LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research, 9, 2008 : 1871–1874. Software available at http://www.csie.ntu.edu.tw/~cjlin/liblinear
11. *Tsur O., Davidov D., Rappoport A.* 2010. ICWCM — a Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. International AAAI Conference on Weblogs and Social Media.
12. *Whitelaw C., Garg N., Argamon S.* 2005. Using Appraisal Taxonomies for Sentiment Analysis. CIKM.