Three-Way Movie Review Classification

Ilia Chetviorkin

CMC MSU

Natalia Loukachevitch

Research Computing Center MSU

Roadmap

- Task Definition
- Features for review classification
- Experiments
- Evaluation of reviews by assessors

Problem definition

- **Opinion** – evaluation, impression, or estimation of the value or worth of a person or thing

- **Review classification** – problem of determining the overall sentiment of the text (document, sentence, phrase …)
  - Binary
  - Ternary
  - Ordinal scale

- A lot of useful applications!

Movie review classification

Well as usual Keanu Reeves is nothing special–, but surprisingly, the very talented+ Laurence Fishbourne is not good– either, I hope they do not shoot a sequel.

Good+ film with an excellent+ sense of humor. For fans of Guy Ritchie. Only a picture is poor–.

The actors are first grade+ and it has a really well thought out story line. I've seen it 10 times and I'll watch it a few more. Enjoy!

Movie review classification

Well as usual Keanu Reeves is nothing special–, but surprisingly, the very talented+ Laurence Fishbourne is not good– either, I hope they do not shoot a sequel.

Good+ film with an excellent+ sense of humor. For fans of Guy Ritchie. Only a picture is poor–.

The actors are first grade+ and it has a really well thought out story line. I've seen it 10 times and I'll watch it a few more. Enjoy!

Movie review classification

Well as usual Keanu Reeves is nothing special–, but surprisingly, the very talented+ Laurence Fishbourne is not good– either, I hope they do not shoot a sequel.

Good+ film with an excellent+ sense of humor. For fans of Guy Ritchie. Only a picture is poor–.

The actors are first grade+ and it has a really well thought out story line. I've seen it 10 times and I'll watch it a few more. Enjoy!

Movie review classification

Well as usual Keanu Reeves is nothing special–, but surprisingly, the very talented+ Laurence Fishbourne is not good– either, I hope they do not shoot a sequel.

Good+ film with an excellent+ sense of humor. For fans of Guy Ritchie. Only a picture is poor–.

The actors are first grade+ and it has a really well thought out story line. I've seen it 10 times and I'll watch it a few more. Enjoy!

Roadmap

- Task Definition
- Features for review classification
- Experiments
- Evaluation of reviews by assessors

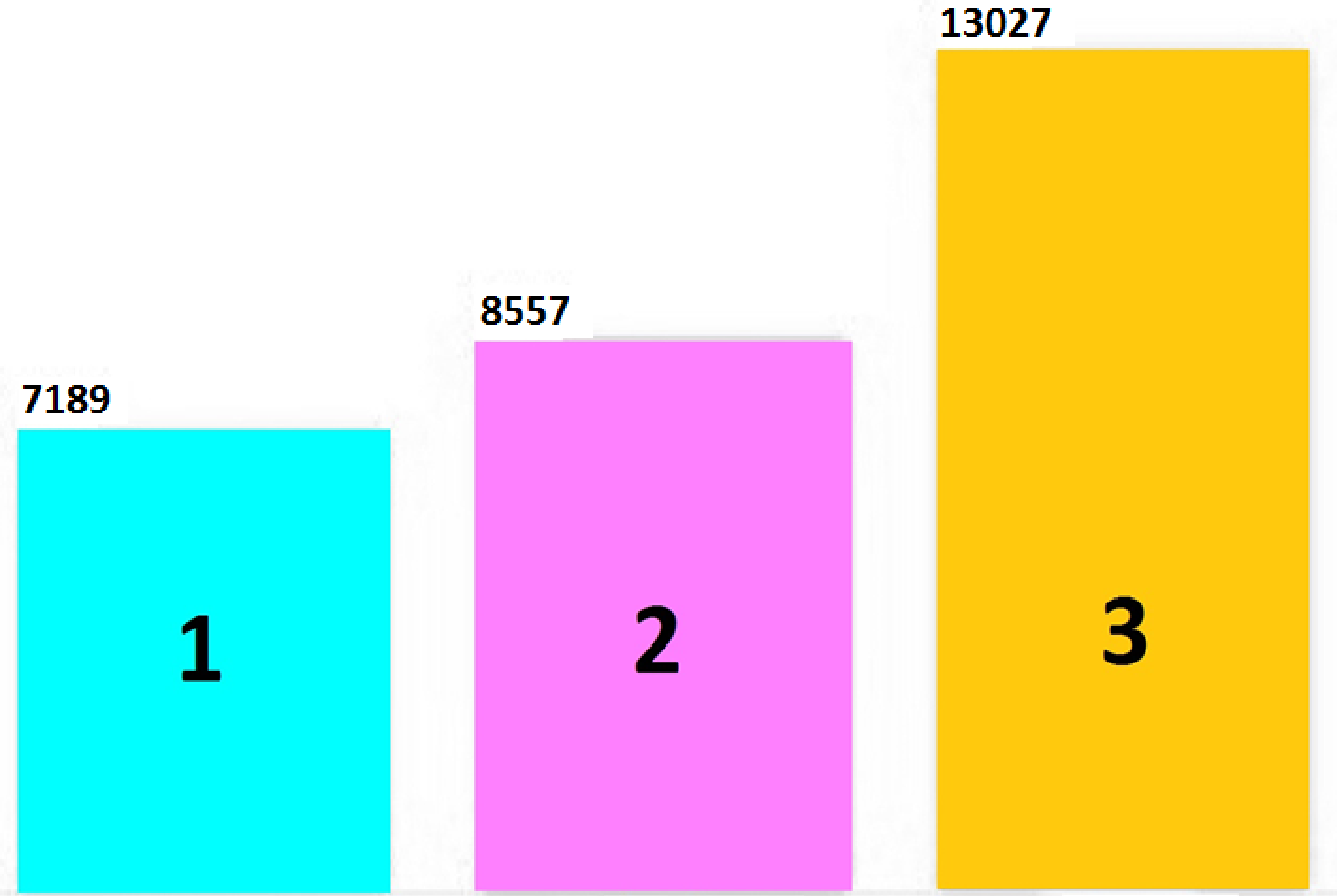
Data

- We collected 28773 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.

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.

Map to the three-point scale

- 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 number of reviews belonging to class «3» is approximately 45% of the total


7189
8557
13027
1
2
3

Features for review classification

- Word weights
- Opinion words
- Polarity influencers
- Review length and structural features
- Punctuation marks

Word weights

- Binary weight (reflect only word presence)
- The simplest form of TFIDF

Manning, Introduction to Information Retrieval

- More complex TFIDF variant based on BM25

Experimental algorithms vs. basic line for web ad hoc, legal ad hoc, and legal categorization in RIRES2004

  - IDF factor was calculated on the basis not only the *review corpus*, but also two other collections: the *news corpus* and the *description corpus*

TFIDF (1)

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

- n_i is the number of term occurrences 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;

- number of documents where term t_i appears.

TFIDF (2)

TFIDF (l) = β + (1- β)·tf(l)·idf(l)

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

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

Features for review classification

- Word weights
- Opinion words
- Polarity influencers
- Review length and structural features
- Punctuation marks

Opinion word extraction

[Chetviorkin, Loukachevitch 2010]

- Four text collections: a movie review collection, a collection of film descriptions , a special small corpus and a collection of general news
- For each word 17 statistical features
- Different machine learning algorithms (best: logistic regression)
- Lists of the most probable opinion words are used in this research

Opinion weights

- 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
- We modified the weight of each word in the feature vectors in the following manner:

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

Features for review classification

- Word weights
- Opinion words
- Polarity influencers
- Review length and structural features
- Punctuation marks

Polarity influencers

- There are some words, which can affect polarity of other words
- From the review corpus, we automatically extracted words directly preceding the manually labeled opinion words and chose candidates from them
- Statistical validation of the candidates
  - If an opinion word had the **high (low)** average score and changed it to the **lower (higher)** → Operator(-) *reverses*

  - If after a polarity influencer, an opinion word with the **high (low)** score **increased (decreased)** its average score→ Operator(+) *magnifies*

Polarity influencers

- 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 :

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

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

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

Features for review classification

- Word weights
- Opinion words
- Polarity influencers
- Review length and structural features
- Punctuation marks

Review length and structural features

- Movie reviews can be long or short
- 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, the end and the middle.
- We classified each part separately and then aggregated obtained scores in various ways

Features for review classification

- Word weights
- Opinion words
- Polarity influencers
- Review length and structural features
- Punctuation marks

Punctuation marks

- There are a lot of studies where researchers point to the importance of punctuation [Tsur O. et. al, *Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews*]
  - Example: *“Trees died for this book?”*
- We included punctuation marks «**!**», «**?**», «**…**» as elements of the feature set

Roadmap

- Task Definition
- Features for review classification
- Experiments
- Evaluation of reviews by assessors

Review vector representation

- We used two types of feature sets:
  - Using the set of all words occurring in the review corpus with high frequency
  - Using the set of opinion words
- For classification LIBLINEAR algorithm was utilized. It is based on SVM and has a great performance.
- 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

Feature sets

- The optimal set of opinion words produced by our method **OpinCycle**
- Set of words, which was used in previous works to achieve the best results **OpinContrast**
- Set of opinion words (3200 units), obtained by manual labeling by two experts **OpinIdeal**
- Set of all words occurring in the review corpus four or more times **BoW**

Experiments

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

Feature set	Feature number	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

Evaluation with soft-borders

- The assumption: even a human distinguishes boundary classes unsatisfactory
- 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
- The classification accuracy with *soft borders* reaches **76.48%**

Roadmap

- Task Definition
- Features for review classification
- Experiments
- Evaluation of reviews by assessors

Evaluation of reviews by assessors

- For a benchmark, we selected 100 short reviews and 100 long reviews from the review corpus
- Reviews were extracted in such a manner, as to retain original class distribution
- Two assessors evaluated the selected reviews
- All explicit references to the initial score were removed

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

Conclusion

- We investigated influence of various factors on the quality of three-way classification of movie reviews in Russian
- We obtained the quality of classification at level **64.48**%, which can be used as a starting point for further work
- We estimated the upper limit of classification quality, which is very close to the results of the best automatic algorithm
- The quality of classification with soft borders is very close to the quality of assessors

Thanks for your attention!

ilia2010@yandex.ru