UDC 004.032.26

Koybagarov K.Ch.[1], Mansurova M.Ye.[2]*

[1]Institute of Information and Computational Technologies of KS MES of the RK,
Republic of Kazakhstan, Almaty
[2]Al-Farabi Kazakh National University, Republic of Kazakhstan, Almaty
* E-mail: mansurova01@mail.ru

Automatic classification of reviews based on machine learning

Currently, there is strong interest in the problem of automatic analysis of reviews of Internet users on various issues. One of the main problems in the analysis of reviews is a tone classification of the texts. This article is about different approaches to the problem of tone classification in 3 classes using the machine learning methods on the example of three collections. The main objective that was set in this work is the comparison of different approaches to the text view within the frame of the vector model, several machine learning methods, and various combinations of statistical and linguistic features. To build the model of tone classification the follow set of statistical and linguistic features is identified: Building word vectors, accounting N-gramm, accounting emoticons, counting of exclamation and question marks, accounting parts of speech, replacing the long repetition of vowel to one vowel, accounting negations, accounting the review length. In this work we used the following machine learning methods: support vector machines, logistic regression and naive Bayesian classifier. The computing experiments were conducted with different variants of word vector models, N-grams and text description features. The experimental results allow us to make recommendations on the selection of the most effective features for tone classification.

Key words: tone classification, machine learning, support vector machines, logistic regression, naive Bayesian classifier.

Қойбағаров К.Ч., Мансұрова М.Е.

Машиналық оқыту алгоритмдер негізінде кіпірлерді автоматты түрде классификациялау

Қазіргі таңда әртүрлі сұрақтар бойынша Интернет қолданушыларының пікіріне сүйенсек автоматты талдау есебіне үлкен қызығушылық танылып жатыр. Талдау пікірінің негізгі мәселесінің бірі болып тон бойынша тексттердің классификациясы болып табылады. Жұмыста коллекциялар мысалында машиналық оқыту әдістерін қолдану арқылы класқа тондық классификациялау есебінің түрлі әдістері берілген. Осы жұмыста қойылған негізгі есептер, векторлы моделдің аясында текстті ұсыну түрлі тәсілдерді салыстыру болып табылады, бірнеше машиналық оқыту тәсілдерін қолдану, статистикалық және лингвистикалық белгілерінің түрлі бірігуі және алынған нәтижелердің талдауы. Тондық классификация моделін құру үшін келесідей статистикалық және лингвистикалық белгілерінің келесідей жиыны анықталған: сөз векторларының құрылуы, N-gramm есебі, эмитикондар есебі, сұрақ белгісі мен тыныс белгілерінің есептелуі, сөйлеу бөлігінің есебі, ұзақ қайталанатын дауысты дыбысты бір дауысты дыбысқа алмастыру, терістеулердің есебі, жауап беру ұзындығының есебі. Машиналық оқытудыө келесі әдістерін қолдану арқылы тәжірибе жүргізілді: тіреу векторлар машинасы, логистикалық регрессия және аңғалдық Байес классификаторы. Тәжірибе әртүрлі векторлық сөздер моделінің нұсқаларымен, n-граммалар және мәтінді сипаттаудың белгілерімен жүргізілді. Тәжірибе нәтижесі тондық классификация үшін ең тиімді сипаттамаларды таңдау бойынша ұсыныс жасауға мүмкіндік береді.

Түйінді сөздер: тондық классификациялау, машиналық оқыту, тіреу векторлар машинасы, логистикалық регрессия, аңғалдық Байес классификаторы.

Койбагаров К.Ч., Мансурова М.Е.

Автоматическая классификация отзывов, основанная на алгоритмах машинного обучения

В настоящее время проявляется большой интерес к задаче автоматического анализа мнений пользователей Интернета по различным вопросам. Одной из основных проблем при анализе мнений является классификация текстов по тональности. В работе даны различные подходы к задаче тоновой классификации на 3 класса с использованием методов машинного обучения на примере трех коллекций. Основными задачами, которые ставились в данной работе, являются сравнение различных подходов к представлению текста в рамках векторной модели, применение нескольких методов машинного обучения, различное сочетание статистических и лингвистических признаков, а также анализ полученных результатов. Для построения модели тоновой классификации был выявлен следующий набор статистических и лингвистических признаков: построение векторов слов, учет N-граммов, учет эмотиконов, подсчет восклицательных и вопросительных знаков, учет частей речи, замена долгого повторения гласного на одну гласную, учет отрицаний, учет длины отзывов. В работе были использованы следующие методы машинного обучения: машина опорных векторов, логистическая регрессия и наивный байесовский классификатор. Вычислительные эксперименты проводились с различными вариантами векторной модели слов, N-граммов и признаков описания текста. Результаты экспериментов позволяют сделать рекомендации по выбору наиболее эффективных признаков для тоновой классификации.

Ключевые слова: тоновая классификация, машинное обучение, машина опорных векторов, логистическая регрессия, наивный байесовский классификатор.

1 Introduction

Analyzing the behavior of people around the world, we can discover the vast influence the opinions of those around them for their views and their choice. People are very often guided by the people opinion before they make a decision. The opinions, the tone of the statements is important not only for individuals but also for commercial organizations, political parties and other organizations.

The problem of text classification by the tone has become relevant quite recently from 2007. Dozens of articles on the subject have been recently published, detailed overviews can be found in [1, 2]. The heightened interest in this range of problems facilitates by the presence of a vast array of Internet sites, social networks, blogs, forums where people express their opinions.

This paper investigates a method of automatic classification of tone using the machine learning techniques (sentiment analysis) [3, 4]. By tone is meant (user review) author's subjective evaluation relative to some object [5]. In this paper we solve the problem of the analysis of comments on the publication on the Internet based on machine learning methods into three classes, denoting positive, neutral and negative evaluations. All experiments were performed on three collections of movies, books, games from the site imhonet.ru where every comment has a ten-point scale.

2 Problem description

There are 3 approaches to determine the text tone automatically [6,7]: approach based on dictionaries of tonal vocabulary on the basis of the machine learning method, hybrid approach and a combination of both approaches.

Table 1 – Characteristics of collections

Collection name	Source	Number of reviews
Books	Imhonet	39,120
Movies	Imhonet	43,200
Games	Imhonet	17,629

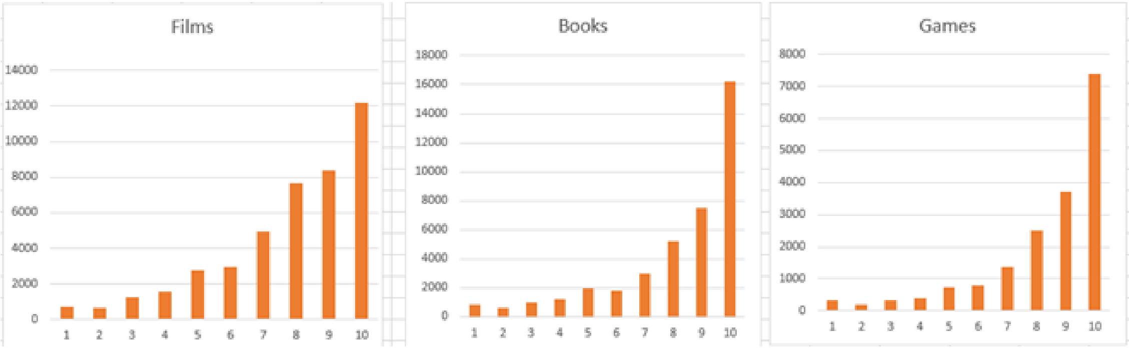

Figure 1 – The numerical distribution of authors assessments for the three collections logical inference

We use the second approach, a machine learning method in this work. The main objectives that we set in this work is the comparison of different approaches to the text view within the frame of the vector model, several machine learning methods, and various combination of statistical and linguistic features. We chose the method of support vector machines (Support vector machine, SVM), naive Bayesian classifier and logistic regression analysis [8, 9] for testing.

We collected 3 text collections from the site *www.imhonet.ru* for our experiments. Collection of movies, books and games, where each review was evaluated on a ten-point scale. Collection about movies contains 43,200 reviews, collection about books contains 39,120 reviews and collection about games contains 17,629 reviews. Each review has the author's assessment from 0 to 10 points. Quantitative distribution of assessments of all collections is shown in Fig. 1.

The figure 1 shows that website visitors leave more positive reviews than negative. For 3-class task we made the following distribution of estimates of 0-5 as negative, 6-8 as neutral and 9-10 as positive review.

Accounting the review length and its structural features. Reviews about movies can be long and short. If the review is long, it is often the author can touch upon various aspects of the described object, which in turn can have different scores. This was the basis for separate consideration of short and long reviews. The threshold on the review length was chosen equal to 50 words.

We have received two collections main and small to account the length of the reviews, the collection with length of messages to 50 words, we called a small collection.

Table 2 – Characteristics of the training dataset

Collection	Positive review	Neutral review	Negative review
Main collection Movies	28205	7921	6918
Small collection Movies	24267	6363	5565
Main collection Books	28879	4769	5490
Small collection Books	24139	3707	4294
Main collection Games	13595	2119	1914
Small collection Games	11758	1696	1568

3 Building feature description

We need to identify a set of statistical and linguistic features to build the model of tone classification. List them below:

- Building word vectors. We used the TF-IDF vectorizer in our work. Compared to simple word count (CountVectorizer), TF-IDF showed the best results.
- Accounting N-gramm. Unigrams, bigrams and trigrams.
- Accounting smiles (emoticon). Set of symbols depicting positive or negative emotions. For example: emoticons ":-) or =) or :) "show a smile, joy, ":-(or =(or :("indicate sadness, sorrow.
- Accounting exclamation and question marks. Based on the assumption that sentences expressing emotions contain a large number of exclamation and positive marks.
- Accounting parts of speech (noun, verb, adverb and adjective). Emotional speech is mostly expressed with adjectives, verbs and adverbs and rarely nouns.
- Replacing the long repetition of vowel to one vowel. "Wooonderful".
- Accounting negations (not, no). Words with the prefix "not"change the next word to the opposite meaning.
- Accounting the review length. The reviews with large number of words contain a large number of positive, negative and neutral words that complicate the task of tone classification.

The experiments were conducted of each feature to take account of their impact on classification accuracy.

4 Experiments

In the experiment we used three classification methods: SVM - support vector machine, Log - logistic regression, NB - naive Bayes classifier. We used cross validation to 5 parts for all experiments. There are task lists in which we alternately conducted experiments on three collections of movies, books and games above.

Table 3 – The accuracy for different collections of TF-IDF against CountVectorizer

	SVM	**Log**	**NB**
TF-ID Main collection Movies bigrams 2-gramm	73.6	72.1	71.2
CountVectorizer Main collection Books bigrams 2-gramm	71.2	72.8	70.0

Table 4 – The accuracy on different collections with regard to long and short reviews

	SVM	**Log**	**NB**
Main collection Movies	73.6	72.1	71.2
Small collection Movies	74.5	73.0	72.5
Main collection Books	80.2	78.1	78.1
Small collection Books	80.5	79.0	78.7
Main collection Games	78.5	77.3	77.1
Small collection Games	80.5	79.6	78.7

Building word vectors. We will test two vectorizers TF-IDF and CountVectorizer, building word vectors with a counting words.

According to the test results TF-ID 73.6 showed the best result compared to the CountVectorizer 72.8 at 0.8

Accounting review length. Reviews about movies can be long and short. If the review is long, it is often the author can touch upon various aspects of the described object, which in turn can have different scores. This was the basis for separate consideration of short and long reviews. The threshold on the review length was chosen equal to 60 words, which we called the main and small collections.

From table 4 we can conclude that the truncated reviews of small collection in contrast to the main collection gain accuracy 0.3 - 2The highest increase of accuracy up to 2

Table 5 – Classification results on unigrams, bigrams and trigrams

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies unigrams	69.4	67.8	64.9
Main collection Books unigrams	75.8	73.1	71.3
Main collection Games unigrams	74.0	70.9	70.7
Main collection Movies bigrams	71.3	68.5	67.8
Main collection Books bigrams	77.0	73.7	73.8
Main collection Games bigrams	74.0	71.1	71.7
Mail collection Movies trigrams	71.6	68.5	68.0
Mail collection Books trigrams	77.0	73.7	74.2
Mail collection Games trigrams	74.1	71.4	72.0

Table 6 – Examples of emoticons and their descriptions

Notation	**Emotion or state**
:-) or =) or :)	smile, joy
hline :-(or =(or :(	sadness, sorrow
hline :-D or :D	laughter
hline :-C or :C	strong chagrin

Table 7 – Classification results taking into account emoticons on bigrams

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies	71.4	68.5	67.8
Main collection Books	77.0	73.8	73.8
Main collection Games	74.1	71.2	71.8

Accounting N-grams. To answer the question what N-grams affect the classification accuracy. We have conducted experiments on unigrams, bigrams and trigrams.

Having analyzed results of Table 5 it is possible to draw a conclusion that bigrams and trigrams have better result on F1 measure than unigrams. The advantage of trigrams to bigrams is slightly 0-0.3% in all three collections. It can be concluded that is better to use bigrams for classification considering the accuracy and speed of obtaining results. In further experiments we will use bigrams and also compare all experiment results with Table 4.

Accounting smiles. Emoticons are widely used in popular culture, the word "smile"itself is also often used as a generic term for any emoticon (image of emotion not by graphics, but with punctuation marks) [Wikipedia]. Emoticons are quite often used in responses to expression their attitude (emotion) towards a subject of discussion.

For account of emoticons in collections we will replace positive emoticon to the word "excellent negative emoticon to the word "bad"in responses. We have conducted experiments taking into account replacement of emoticons.

Subsequent to the results of Table 7 it can be concluded that accounting of emoticons slightly increases the quality on average by 0.1%.

Accounting exclamation and question marks. Emotional speech is replete with exclamation points, question marks, and dots and asterisks. We have carried out an experiment for hypotheses about the impact of these signs on accuracy of classification.

According to the results of Table 8 classification quality has increased by 0.1-0.2%. Naive Bayes shows even better results than method of logistic regression.

Accounting parts of speech. As it is known the emotional speech is transmitted

Table 8 – Classification results taking into account emoticons on bigrams

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies	71.4	68.6	68.6
Main collection Books	77.1	73.5	74.4
Main collection Games	74.2	71.0	72.1

Table 9 – Classification results in consideration of parts of speech

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies	71.4	68.7	67.7
Main collection Books	77.0	73.6	73.7
Main collection Games	74.1	71.1	72.0

Table 10 – Classification results in consideration of parts of speech

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies	71.3	68.5	67.8
Main collection Books	77.0	73.7	73.7
Main collection Games	74.0	71.1	71.7

through four parts of speech: adjective, verb, adverb and rarely noun. We will make calculation of above mentioned parts of speech in each response to perform an experiment on influence of parts of speech on accuracy of classification.

According to the results of Table 9, we can conclude that addition of accounting features of parts of speech significantly increases classification accuracy by 0.1%. On a book collection method of basic vectors addition of accounting features of parts of speech haven't influenced an accuracy. From here it is possible to draw a conclusion that addition of accounting features of parts of speech can be neglected.

Replacing the long repetition of vowel to one vowel. Sometimes emotions are expressed by repeating long vowel "wooonderful "Aaaaa etc. in reviews. We have decided to check how replacement of several vowels to one vowel in a word will influence classification accuracy.

As a result of experiment in Table 10 we can conclude that substitution of several vowel one does not affect the accuracy of the classification. This feature can be excluded from consideration.

Accounting negation operator words "no, no, nothing". Negation words invert meaning of emotional word following it to the opposite meaning. For example, "not good"changes the meaning to "bad etc. To solve this problem, we decided to stick particles "no "not"with followed by the operator the word "not good"to "notgood".

According to the results of Table 11 we have come to conclusion that the operators of negation n̈o, not, nothing̈have improved accuracy of the classification F1-least 0.1%. In the collections of movies, books, support vector has shown improvement by 0.1%, and there has

Table 11 – Classification results in consideration of parts of speech

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies	71.4	68.4	67.2
Main collection Books	77.1	73.6	74.1
Main collection Games	74.0	71.1	71.9

Table 12 – Classification results on unigrams, bigrams and trigrams

	SVM F1-macro	**Log F1-macro**	**NB F1-macro**
Main collection Movies bigrams	71.7	68.5	68.4
Main collection Books bigrams	77.2	73.7	74.4
Main collection Games bigrams	74.1	71.4	72.6
Mail collection Movies trigrams	71.9	68.5	68.7
Mail collection Books trigrams	77.2	73.8	74.5
Mail collection Games trigrams	74.0	71.4	72.8

not been any improvement at the games.

Now we will make experiment if we include all above listed characteristics.

After inclusion of all characteristics, classification accuracy has improved by 0.1-0.4%. Trigrams do not have much advantage over bigrams. In the collection of movies trigrams result are better by 0.2% than bigrams results are, but in the collection of games results are worse by 0.1% than bigrams.

5 Conclusion

After carrying out all the experiments in dealing with text automatic classification task of tonality on 3 classes using machine learning techniques following conclusions can be made:

- TF-IDF words vectorizer shows better result in comparison with CountVectorizer.
- The method of support vectors shows better results compared to the methods of logistic regression and naive Bayes.
- Bigrams and a trigrams show better results at F1 measure than unigrams. Trigrams have no big advantage before bigram on average by 0.1%. Thus trigrams can be neglected.
- The experiments performed with each of the features listed above showed an improvement in accuracy by 0.1% on average tonality classification. In the collections of books and games not all features shown improvement in accuracy. And characteristics such as substitution of long vowel repetition to one vowel can be neglected.
- Experiments including all three features in collections showed an increase in the accuracy of classification into 3 classes on average by 0.25%.
- Small or truncated collection limited to 50 words in each response gives a gain of classification quality to 2%.

We want to perform the following experiments using hybrid approach by means of machine learning methods and tone dictionaries. This work is supported by the Science Committee of RK, under grants grants 0115PK00552, 0115PK02741 and 0115PK00779.

References

[1] *Pang B., Lee L.* Opinion Mining and Sentiment Analysis. // J. Foundations and Trends in Information Retrieval. – 2008. Vol. 2. – P. 1–135.

[2] *Feng, V. W., Hirst G.* Detecting deceptive opinions with profile compatibility. // In: Proceedings of the 6th international joint conference on natural language processing. – 2013. – P. 338–346.

[3] *Liu B.* Sentiment Analysis and Opinion Mining. Morgan and Claypool Publ. – 2012.

[4] *Kotelnikov Y.V.* Combined method of automatic determination of the text tonality. // J. Software products and systems. – 2012. – Vol 3. – P. 189–195.

[5] *Prabowo R., Thelwall M.* Sentiment analysis: A combined approach. // Journal of Informetrics. – Vol. 3, issue 2. – 2009. – P. 143-157.

[6] *Kevin P. Murphy.* Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series). The MIT Press. – 2012.

[7] *Jindal N., Liu B., Lim E.* DFinding unusual review patterns using unexpected rules. // In: CIKM '10, Proceedings of the 19th ACM international conference on information and knowledge management. – 2010. – P. 219–230.

[8] *Montoyo A., Martinez-Barco P., Balahur A.* (2012). Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. // J. Decision Support Systems. – Vol. 53, issue 4. – P. 675–679.

[9] *Panicheva P., Cardiff J., Rosso P.* Identifying subjective statements in news titles using a personal sense annotation framework. // Journal of the American Society for Information Science and Technology. – 2013. – Vol. 64, issue 7. – P. 1411–1422.

[10] *Severyn A., Moschitti A., Uryupina O., Plank B., Filippova K.* Opinion mining on YouTube. // In: Proceedings of the Conference ACL. – 2014.

[11] *Uryupina O., Plank B., Severyn A., Rotondi A., Moschitti A.* SenTube: A corpus for sentiment analysis on YouTube social media. // In: Proceedings of the International Conference on Language Resources and Evaluation LREC. – 2014.

[12] *Basile V., Nissim M.* Sentiment analysis on Italian tweets. // In: Proceedings of the 4th Workshop on computational approaches to subjectivity, sentiment and social media analysis. – 2013.