USE OF LEXICONS TO IMPROVE QUALITY OF SENTIMENT CLASSIFICATION

Rusnachenko N. L. (kolyarus@yandex.ru)
BMSTU, Moscow, Russia

Abstract

This paper describes the application of SVM classifier for sentiment classification of Russian Twitter messages in the banking and telecommunications domains of SentiRuEval-2016 competition. A variety of features were implemented to improve the quality of message classification, especially sentiment score features based on a set of sentiment lexicons. We compare the result differences between train collection types (balanced/imbalanced) and its volumes, and advantages of applying lexicon-based features to each type of the training classifier modification. The created system achieved the third place at SentiRuEval-2016 in both tasks. The experiments performed after the SentiRuEval-2016 evaluation allowed us to improve our results by searching for a better 'Cost' parameter value of SVM classifier and extracting more information from lexicons into new features. The final classifier achieved results close to the top results of the competition.

Key words: Machine Learning, SVM, Sentiment Analysis, Lexicons, SentiRuEval 2016

1. Problem and Data

In the SentiRuEval-2016 competition one of the suggested tasks is devoted to the reputation analysis of Twitter messages. Being participant you had to determine a sentiment class, which shows the relationship between the message and a company mentioned in it. The organizers offered two domains of companies: bank companies (BANK) and telecommunication companies (TCC).

In each domain, the organizers provided the participants with the training and test collections. The latter collection contains unlabeled messages. The participants were required to label each message of the test collection with one of the following scores:

1 – *positive*, **0** – *neutral*, **-1** – *negative*.

2. Approach

Among the variety of classifiers of the area of machine learning, we used SVM with linear kernel method due to the results [5] which shows advantages (in comparison with NB) in case of the unigram message processing model. The sentiment classification model has been built by means of LibSVM[1] library [1].

The message processing algorithm consists of the following steps:

1. Lemmatize message words to produce a list of message terms;
2. Removing retweet symbols ('RT'), user names (term with '@' prefixes), and URL-links;
3. Applying the list of the *stop words*.
4. Replace pattern bigrams and unigrams with sentiment prefixes:

I am prepared for the worst part, but hope for the best part of the day
I am prepared for the –part, but hope for the +part of the day

As the measure of weight coefficients we use TF-IDF. We also added the following extra features:

- Emoticons (calculating the *sum* of positive and negative emoticons);
- Amount of UPPERCASE WORDS;
- Amount of signs: {'?', '…', '!'};
- Calculating *sum* $x = \sum L(t),\ t \in L$, those terms t, which presented in the message and *sentiment lexicon* L [6], [7]. Sentiment lexicons are lists of words and phrases with sentiment scores. The calculated *sum* was normalized by the formula:

$$\begin{cases} s = 1 - e^{-|x|}, x > 0 \\ s = -\left(1 - e^{-|x|}\right), x < 0 \end{cases}$$

The sentiment scores of automatically created lexicons were composed by *Semantic Orientation* (SO) which based on *Pointwise mutual information* (PMI) [8]:

$$\boldsymbol{PMI}(t_1,\ t_2) = \log_2 \frac{P(t_1 \wedge t_2)}{P(t_1) \cdot P(t_2)}\ ,\quad \boldsymbol{SO}(t) = PMI(t,\ \boldsymbol{Positive}) - PMI(t,\ \boldsymbol{Negative})$$

Table 1 presents all used lexicons. The following list shows data used to build lexicons:

1. Corpora of short Twitter posts in Russian language[2];
2. Twitter messages through the January, 2016 (1% of all stream, Russian locale, *Streaming Twitter API*);
3. Sentiment lexicon, created manually by experts [3]. [3]

[1] LIBSVM: A library For Support Vector Machines: http://www.csie.ntu.edu.tw/~cjlin/libsvm
[2] Corpora of short Twitter posts in Russian language Twitter: http://study.mokoron.com/
[3] http://www.labinform.ru/pub/rusentilex/index.htm

№	Positive	Negative	Total
1	62 637 **(55.5%)**	50 177 (44.5%)	112 814
2	7 370 (3.12%)	228 721 **(96.8%)**	236 091
4	2 774 (26.0%)	7 148 **(67.0%)**	10 668

Table 1. Created sentiment lexicons (amount of terms)

3. Train Collections

Provided by the SentiRuEval organizers train collections are not balanced. The most part of each collection is belongs to the neutral class (more than 46%, Table 2). To balance collections, all provided collections has been merged separately in each domain, and then balanced by sentiment messages from an external source. This source represents messages from the corpora of short Twitter posts in Russian language.[2] The messages have been filtered by means of L – lexicon №1 and added in *Extended balanced collection* (see Table 2). The Filtering rule was as follows: a message must contain terms with the highest absolute value $L(t)$. The sign of the absolute value determines the sentiment class (positive or negative).

Training collection SentiRuEval-2015 [2]				
Collection	Positive messages	Neutral messages	Negative messages	Total
BANK	356 (7,2%)	3482 **(70.84%)**	1077 (21.29%)	4915
TCC	956 (19.67%)	2269 **(46.69%)**	1634 (33.62%)	4859
Training collection SentiRuEval-2016 [4]				
Collection	Positive messages	Neutral messages	Negative messages	Total
BANK	1354 (15.41%)	4870 **(55.4%)**	2550 (29.03%)	8783
TCC	704 (7.7%)	6756 **(74.22%)**	1741 (19.12%)	9102
Extended balanced collection				
Collection	Messages in class		Total messages	
BANK	6765		20295	
TCC	4894		14682	

Table 2. Training collections

4. Results of SentiRuEval-2016

The following list shows classifier settings (used terms and features in a message vector):

№1. Using only Russian terms and #hashtags (sentiment prefixes disabled);

№2. Setting №1 + *use sentiment prefixes*, *use lexicons №1 and №2*, include *all features*;

№3. Setting №2 + *use lexicon №3*.

Table 3 presents results according to the settings. **Bold tagged** results shows the 3rd place among the all (10) participants of SentiRuEval-2016.[4]

№	BANK: Train collection SentiRuEval-2015 $F_{macro}(neg,pos)$	BANK: Train collection SentiRuEval-2015 $F_{micro}(neg,pos)$	BANK: Extended balanced train collection $F_{macro}(neg,pos)$	BANK: Extended balanced train collection $F_{micro}(neg,pos)$
1	38.40	42.03	45.36 (+6.96)	49.82 (+7.79)
2	38.49	41.50	46.72 (+8.23)	50.29 (+8.79)
3	38.62	42.18	**46.83** (+8.21)	**50.22** (+8.04)
№	**TCC: Train collection SentiRuEval-2016 $F_{macro}(neg,pos)$**	**TCC: Train collection SentiRuEval-2016 $F_{micro}(neg,pos)$**	**TCC: Extended balanced train collection $F_{macro}(neg,pos)$**	**TCC: Extended balanced train collection $F_{micro}(neg,pos)$**
1	48.49	64.10	51.03 (+2.54)	65.09 (+0.99)
2	48.32	64.73	52.31 (+3.99)	65.08 (+0.35)
3	50.99	67.70 (+1.38)	**52.86** (+1.87)	**66.32**

Table 3. Competition results[4], SentiRuEval-2016 (advantages of *Extended balanced train collection*)

5. Results improvements

Experiments performed after the SentiRuEval-2016 evaluation allowed us to improve results in following ways:

1. Searching for best *Cost* parameter of SVM classifier *penalty function*. (Cost = 0.5)
2. Introducing new lexicon-based features: calculating *max* and *min* values through all terms t_i of message $m = \{t_i\}_{i=1}^{N}$ for each lexicon L of Table 1.

Table 4 presents the results after applying all improvements. **Bold tagged** results shows 2nd place among the all participants.[4]

№	BANK $F_{macro}(neg,pos)$	BANK $F_{micro}(neg,pos)$	TCC $F_{macro}(neg,pos)$	TCC $F_{micro}(neg,pos)$
1	49.55	53.88	52.59	66.62
2	50.12	53.79	52.83	67.20
3	**52.39**	**55.14**	**54.53**	**69.70**

Table 4. Results improvements

[4] All results: https://docs.google.com/spreadsheets/d/1rCaklClawfnnSnyk4q8CW4zWuO3P38DSrLw_f2wyyjg/edit#gid=0

References

[1] Chang C.-C., Lin C.-J. (2011), LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2(3):27:1–27:27

[2] Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Yu., Ivanov V., Tutubalina E. (2015), SentiRuEval: testing object-oriented sentiment analysis systems in Russian, Proceedings of International Conference Dialog-2015, Vol. 2, pp. 3-13.

[3] Loukachevitch N., Levchik A. (2016), Building lexicon of valuable Russian words of RuSentileks language, [Sozdanie leksikona ocenochnyh slov russkogo jazyka RuSentileks], Proceedings of Conference OSTIS-2016, pp. 377-382.

[4] Loukachevitch N., Rubtsova Yu. (2016), SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis, Proceedings of International Conference Dialog-2016.

[5] Pang B., Lee L., Vaithyanathan S. (2002), Thumbs up: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, Vol. 10, pp. 79-86.

[6] Saif M., Kiritchenko S, Xiaodan Z. (2015), NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets, Second Joint Conference on Lexical and Computational Semantics, Vol. 2, pp. 321-327.

[7] Severyn A., Moschitti A. (2015), On the Automatic Learning of Sentiment Lexicons, Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, pp. 1397-1402.

[8] Turney P. (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, Proceeding ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424