

УЛУЧШЕНИЕ КАЧЕСТВА ТОНАЛЬНОЙ КЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ ЛЕКСИКОНОВ

Русначенко Н. Л. (kolyarus@yandex.ru)
МГТУ им. Н.Э. Баумана, Москва, Россия

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. Before SentiRuEval-2016, the classifier was tested on the previous year collection of the same competition (SentiRuEval-2015) to obtain a better settings set. 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. Введение

В настоящее время одним из наиболее популярных сервисов распространения коротких новостей является социальная сеть Twitter. Большинство пользователей сети часто выражают свое мнение о том, что им понравилось или не понравилось в определенной сфере услуг. Доступность данных сети извне дает возможность обработки и анализа высказанных мнений.

В этой работе рассматривается построение модели на основе SVM классификатора для определения тональности сообщений сети Twitter заданной тематики. Подразумевается построение моделей применительно к следующим тематикам: отзывы в банковской и телекоммуникационных сферах. Каждое сообщение может быть отнесено к одному из трех тональных классов: негативному, нейтральному, и положительному.

В ходе построения и настройки модели исследовались различные признаки для представления содержания сообщений. Особое внимание уделялось применению словарей оценочных слов для повышения качества классификации.

2. Построение лексиконов оценочных слов

Под термином «лексикон» понимается словарь S , состоящий из пар (t, v) , где t – терм, $v \in \mathbb{R}$ – параметр, знак которого определяет тональную окраску слова w (положительную или негативную), а $|v|$ – степень окраски.

Для построения лексикона применяется подход [8], основанный на определении семантической ориентации словосочетаний, которая, в свою очередь, определяется метрикой **точечной взаимной информации** (англ. PMI, Pointwise Mutual Information):

$$PMI(t_1, t_2) = \log_2 \frac{P(t_1 \wedge t_2)}{P(t_1) \cdot P(t_2)}$$

Поскольку для каждого термина t , содержащегося в лексиконе необходимо сопоставить оценку тональности, то в качестве одного из аргументов метрики PMI можно рассмотреть один из двух «маркеров»:

- «**Excellent**» («отличный») – положительный оттенок;
- «**Poor**» («плохой») – негативный оттенок.

Введение маркеров в качестве одного из параметров метрики PMI позволяет установить степень принадлежности слова соответствующему маркеру. Степень

принадлежности терма двум маркерам называется его **семантической ориентацией**, и определяется формулой:

$$SO(t) = PMI(t, "Excellent") - PMI(t, "Poor")$$

Пусть K – произвольная коллекция сообщений сети Twitter. Тогда, на основе коллекции K может быть составлен лексикон S следующим образом [6], [7]:

$$S: \{ \langle t, SO(t) \rangle \mid t \in K_{excellent} \vee K_{poor} \}$$

Где $K_{excellent}$ и K_{poor} – разделение исходной коллекции K на непересекающиеся тональные классы сообщений с положительным и негативным оттенками соответственно. Для построения тональных классов, в работах [6] и [7] предлагается анализировать сообщение на наличие положительных (негативных) эмотиконов, а также на наличие положительных (негативных) хэштегов.

3. Задачи и данные

В рамках соревнования SentiRuEval-2016 одна из предложенных задач посвящена теме анализа репутации по сообщениям сети Twitter. Необходимо было определить тональность сообщения по отношению к упомянутым в них организациям. В качестве организаций используются банки и телекоммуникационные компании (ТКК).

В каждой области была предоставлена обучающая и тестовая коллекции. Все сообщения описаны в XML формате. От участников требовалось для каждой организации предоставить преобразованную тестовую коллекцию, в которой каждой упомянутой организации в сообщениях проставлена одна из следующих оценок:

- 1 – положительное сообщение;
- 0 – нейтральное сообщение;
- 1 – негативное сообщение.

4. Предложенный подход

4.1 Обработка сообщений тестовой и обучающей коллекций

В области классификации сообщений методами машинного обучения, использование SVM классификатора (в сравнении с Naive Bayes) обусловлено результатами тестирования в [5], которые показывают преимущество SVM на униграммной модели обработки сообщений.¹ Для построения обучающей модели и предсказания тональности на ее основе, используется библиотека LibSVM² [1].

Обработка сообщений состоит из выполнения следующих этапов:

- Лемматизация слов сообщений³ для получения списка термов;
- Очистка списка термов от символов ретвита ('RT'), имен пользователей (термы с префиксом '@') и URL-адресов;
- Применение списка стоп слов⁴. Список составлен из термов множества всех лексиконов, и включает в себя те термы, модуль параметра SO для которых был меньше порогового значения K ($K = 0.05$):

$$L_{stopwords} = \{ t \mid |SO(t)| < K, t \in S \}$$

- Замена некоторых биграмм и униграмм на тональные префиксы. Предварительно составлен список пар⁴ $L_{tone} = \{ \langle t, s \rangle \}$, где t – терм, s – тональная оценка ('+' или '-'). На этом этапе для каждого терма $t_i \in L_{tone}$ выполняется замена на соответствующую оценку s , которая становится префиксом следующего терма t_{i+1} . Пример:

Сейчас хорошо работать не то что раньше
Сейчас +работать –то что раньше.

При преобразовании списка термов в вектор, весовые коэффициенты терма

¹ Использование униграммной модели упрощает процесс обработки сообщения с точки зрения добавления метаинформации, в том числе и на основе лексиконов. В текущем подходе все термы, содержащиеся во всех лексиконах, являются униграммами.

² LIBSVM: A library For Support Vector Machines: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

³ Mystem – морфологический анализ текста: <https://tech.yandex.ru/mystem/>

⁴ https://github.com/nicolay-r/tone-classifier/tree/2016_jan_contest/test/default/msg.conf

определялись с помощью меры TF-IDF. Дополнительно в векторизацию добавлялись следующие признаки:

- На основе «эмотиконов» подсчет $\sum e_i$: $e_i = 1$, если $e_i \in E_{pos}$, и $e_i = -1$, при $e_i \in E_{neg}$:
 - E_{pos} : { ':)', ':*', ':P', ':D', ':-', ':-D', '=)', 'x)', 'xD', 'xD' };
 - E_{neg} : { ':(', 'D:', ':(', ':/', ':-((', 'D-:', ':-'((', '=(', '=('(', 'x(', 'Dx' }.
- Количество слов написанных в верхнем регистре [6];
- Учет числа подряд идущих знаков: {'?', '...', '!'}.
- Вычисление $x = \sum SO(t)$, $t \in S$ термов t , составляющих сообщение и в входящих в лексикон S [6], [7]. Сумма вычисляется для каждого лексикона, и нормализуется:

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

Лексиконы были составлены⁵ на основе следующих данных (параметры представлены в таблице 1):

1. Корпуса коротких текстов на русском языке⁶;
2. Сообщений сети Twitter за январь 2016 года (подключение к трансляции сообщений на русском языке с помощью Streaming API Twitter);
3. Обучающая коллекция SentiRuEval-2015 года [2];
4. Тональный словарь созданный вручную экспертами [3].⁷

Номер	Задачи	$K_{\text{excellent}}$ термов	K_{poor} термов	Всего термов
1	Для всех	62637 (55.5%)	50177 (44.5%)	112814
2	Для всех	7370 (3.12%)	228721 (96.8%)	236091
3	BANK	1748 (41.51%)	2466 (58.56%)	4211
	ТКК	2460 (38.47%)	3934 (61.53%)	6394
4	Для всех	2774 (26.0%)	7148 (67.0%)	10668

Таблица 1. Параметры созданных лексиконов.

4.2 Составление тестовых коллекций

Одно из последних соревнований в этой области проводилось в 2015 году (SentiRuEval-2015) [2], данные которого находятся в открытом доступе и содержат эталонную коллекцию. Поэтому можно использовать коллекции SentiRuEval-2015 для предварительного тестирования.

Обучающие коллекции не являются сбалансированными, и содержат преобладающий по объему класс нейтральных сообщений. В связи с этим, дополнительно была произведена балансировка сообщениями (твитами), содержащих термины t с высокими по модулю значениями $SO(t)$ лексикона №1. Параметры коллекций для предварительного тестирования представлены в таблице 2.

Несбалансированная обучающая коллекция SentiRuEval-2015				
Коллекция	positive	neutral	negative	всего
BANK	356 (7.2%)	3482 (70.84%)	1077 (21.29%)	4915
ТКК	956 (19.67%)	2269 (46.69%)	1634 (33.62%)	4859
Сбалансированная обучающая коллекция				
Коллекция	Объем класса			всего
BANK	3482			10446
ТКК	2296			6888

Таблица 2. Параметры обучающих коллекций для предварительного тестирования.

Параметры коллекций SentiRuEval-2016 [4] представлены в таблице 3.

Несбалансированная обучающая коллекция SentiRuEval-2016				
Коллекция	positive	neutral	negative	всего
BANK	1354 (15.41%)	4870 (55.4%)	2550 (29.03%)	8783
ТКК	704 (7.7%)	6756 (74.22%)	1741 (19.12%)	9102

Таблица 3. Параметры обучающих коллекций SentiRuEval-2016.

⁵ https://github.com/nicolay-r/tone-classifier/tree/2016_jan_contest/data/lexicons

⁶ Корпус коротких текстов на основе постов Twitter: <http://study.mokoron.com/>

⁷ Ручной словарь опубликован: <http://www.labinform.ru/pub/rusentilex/index.htm>

5. Предварительное тестирование

Предварительное тестирование классификатора производилось на данных соревнований 2015 года. В таблице 4-5 приведены оценки качества работы классификаторов в зависимости от настроек.⁸

BANK				
№	Не сбалансированная коллекция		Сбалансированная коллекция	
	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
1	0.3659	0.4	0.4206 (+15.0%)	0.458 (+14.5%)
2	0.3933	0.4128	0.4305 (+9.4%)	0.4718 (+14.2%)
3	0.4119	0.4394	0.4349 (+5.5%)	0.4792 (+9.0%)

Таблица 4. Предварительные результаты тестирования (задача BANK, SentiRuEval-2015).

ТКК				
№	Не сбалансированная коллекция		Сбалансированная коллекция	
	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
1	0.4608 (+0.5%)	0.5172 (2.5%)	0.4583	0.5045
2	0.4701 (+0.26%)	0.5207 (2.0%)	0.4689	0.5104
3	0.4925 (+3.3%)	0.5378 (3.7%)	0.4767	0.5184

Таблица 5. Предварительные результаты тестирования (задача ТКК, SentiRuEval-2015).

Настройки векторизации сообщений в предварительных прогонах следующие:

№1. Использование русскоязычных термов и хэштегов;

№2. Прогон №1 + применение тональных префиксов, использование лексиконов №1 и №2, а также учет всех признаков;

№3. Прогон №2 + использование всех лексиконов (кроме №3)⁹.

На основе полученных результатов было принято решение о создании **расширенной сбалансированной коллекции**: дополнение положительных и негативных классов коллекции 2016 года соответствующими классами коллекции 2015 года, и дальнейшая балансировка твитами. Параметры расширенной сбалансированной коллекции (см. таблицу 6).

Расширенная сбалансированная обучающая коллекция		
Коллекция	Объем класса	total
BANK	6765	20295
ТКК	4894	14682

Таблица 6. Расширенная обучающая сбалансированная коллекция.

6. Результаты соревнований SentiRuEval-2016

В таблице 5 приведены оценки качества работы классификатора для тестовой коллекции SentiRuEval-2016 [4] при использовании настроек предварительного тестирования. Прогоны с такими настройками показали лучшие результаты среди других вариаций настроек предложенного подхода (см. таблицы 7-8).

BANK				
№	Сбалансированная коллекция (2015 год)		Расширенная сбалансированная коллекция	
	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
1	0.384	0.4203	0.4536 (+18.1%)	0.4982 (+18.53%)
2	0.3849	0.415	0.4672 (+20.9%)	0.5029 (+21.10%)
3	0.3862	0.4218	0.4683 (+21.25%)	0.5022 (+19.06%)

Таблица 7. Результаты прогонов соревнования (задача BANK, SentiRuEval-2016).

ТКК				
№	Несбалансированная коллекция		Расширенная сбалансированная коллекция	
	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
1	0.4849	0.641	0.5103 (+5.2%)	0.6509 (+1.5%)
2	0.4832	0.6473	0.5231 (+8.2%)	0.6508 (+0.5%)
3	0.5099	0.677 (+2.0%)	0.5286 (+3.6%)	0.6632

Таблица 8. Результаты прогонов соревнования (задача ТКК, SentiRuEval-2016).

⁸ Процентный прирост качества вычисляется как отношение наибольшего значения оценки по соответствующей метрике ($F_{macro}(neg, pos)$ или $F_{micro}(neg, pos)$) к наименьшему.

⁹ Применение лексикона, составленного на обучающей коллекции SentiRuEval 2015 года не привело к повышению качества (ввиду малого объема).

После проведения соревнований, в целях повышения качества классификации, настройки прогонов изменялись в следующих направлениях:

1. Настройка параметра C (Cost) штрафной функции SVM классификатора.
 - а. По умолчанию $C=1$. Среди множества протестированных значений $\{1, 0.75, 0.5, 0.25\}$, наибольший прирост достигается при $C = 0.5$ (см. таблицу 9).

№	BANK (Расшир. Сбалансированная коллекция, $C=0.5$)		ТКК (Расшир. Сбалансированная коллекция, $C=0.5$)	
	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
1	0.4558 (+0.4%)	0.5037 (+1.1%)	0.5235 (+2.5%)	0.6612 (+1.5%)
2	0.4795 (+2.6%)	0.5167 (+2.7%)	0.5338 (+2.0%)	0.6610 (+1.5%)
3	0.4768 (+1.8%)	0.5135 (+2.2%)	0.5452 (+3.1%)	0.6733 (+1.5%)

Таблица 9. Влияние настройки параметра Cost ($C=0.5$) (SentiRuEval-2016).

2. Добавление новых признаков: вычисление максимальных и минимальных значений (с учетом нормализации) среди всех термов сообщения по каждому из лексиконов.

Комбинация рассмотренных выше улучшений привела к настройке финальных прогонов (результаты представлены в таблице 10). Во всех прогонах использовались русскоязычные термы и хэштеги, применялись тональные префиксы, а также учитывались все признаки. Изменения в настройках касались только числа используемых лексиконов, а также признаков построенных на их основе (настройки прогонов):

- №1. Вычисление суммы, минимума, максимума на основе лексикона №1 (см. таблицу 1).
- №2. Прогон №1 + признаки суммы, минимума, максимума на основе лексикона №2.
- №3. Прогон №2 + признаки суммы, минимума, максимума на основе лексикона №4.
- №4. Прогон №3 + признаки минимума и максимума на основе лексиконов №3.

№	BANK (Расшир. Сбалансированная коллекция, $C=0.5$)		ТКК (Расшир. Сбалансированная коллекция, $C=0.5$)	
	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
1	0.4955	0.5388	0.5259	0.6662
2	0.5012	0.5379	0.5283	0.6720
3	0.5239	0.5514	0.5453	0.6970
4	0.4818	0.5238	0.5356	0.6659

Таблица 10. Результаты финального тестирования (SentiRuEval-2016).

7. Вывод

Использование метаинформации на основе лексиконов стабильно повышает качество классификации. Наибольший прирост качества достигается в случае, если классификатор был обучен на коллекции несбалансированного типа (см. таблицу 11)¹⁰.

Соревнования (год)	Тип	BANK		ТКК	
		$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$	$F_{macro}(neg, pos)$	$F_{micro}(neg, pos)$
2015	Несбалансированная	+12.57%	+9.8%	+6.8%	+3.9%
	Сбалансированная	+3.3% (+19.0%)	+4.6% (+19.8%)	+4% (+3.4%)	+2.7% (+1.9%)
2016	Несбалансированная	-	-	+5.1%	+4.6%
	Сбалансированная	+0.5%	+0.03%	-	-
	Расширенная, сбалансированная	+4.6% (+21.95) ¹¹	+1.9% (+19.48%) ¹¹	+4.1% (+9.0%)	+1.8% (+3.4%)

Таблица 11. Рост качества при использовании признаков на основе лексиконов.

Увеличение числа признаков по каждому из лексиконов позволяет повысить показания таблицы 11. В совокупности с использованием сбалансированной обучающей коллекции и настройкой классификатора, в рамках этой статьи были получены максимальные результаты (см. таблицу 10, прогон №3).

¹⁰ В таблице рассматривается прирост качества 3-его прогона по отношению к 1-ому (согласно таблицам 4-5, и 7-8). В скобках указывается общий прирост качества с учетом балансировки.

¹¹ Общий прирост качества с учетом расширенной балансировки по отношению к обычной балансировке.

Список литературы

- [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