Н.Л. Русначенко, Н.В. Лукашевич

Методы интеграции лексиконов в машинное обучение для систем анализа тональности

Аннотация. Рассмотрена задача анализа тональности русскоязычных сообщений сети Твиттер в банковской и телекоммуникационной сферах. На основе машинного обучения изложены различные способы представления содержания сообщений. Показано, что использование дополнительных признаков сообщений на основе существующих или заранее порожденных словарей оценочных слов позволяет повысить качество классификации сообщений. Исследовано влияние различных типов обучающих коллекций (сбалансированных/не сбалансированных), их объемов, а также преимущества применения нескольких признаков на основе лексиконов на качество классификации. Подход тестировался на данных открытого тестирования систем анализа тональности SentiRuEval-2015 и SentiRuEval-2016. В итоге были получены результаты, превышающие лучшие результаты SentiRuEval-2015, близкие к результатам победителя SentiRuEval-2016.

Ключевые слова: машинное обучение, SVM, анализ тональности сообщений, лексиконы, SentiRuEval.

Введение

Одним из важных направлений в сфере автоматического анализа тональности текстов является анализ мнений и оценок, высказываемых в социальных сетях. В частности, большое число исследований посвящено анализу мнений, выраженных в сообщениях сети Твиттер, поскольку в ней общаются пользователи различного статуса, образования, профессии. Эти мнения интересны другим пользователям; компаниям, которые ищут отзывы на свои продукты или услуги; социологам, исследующим общественное мнение; государственным организациям для исследования процессов, происходящих в обществе.

Сообщения Твиттера (твиты) имеют несколько особенностей: они короткие, длиной не более 140 символов. Их тематика и сопутствующие оценки очень динамичны, тесно связаны с происходящими событиями, что усложняет многие процедуры автоматического анализа сообщений Твиттера, включая анализ тональности [1, 2].

Традиционно рассматриваются два подхода к анализу тональности, и у обоих есть проблемы с качественным анализом тональности твитов. Первый подход, так называемый лингвистико-инженерный, основан на том, что эксперты составляют вручную словари и правила для выявления позитивных или негативных мнений. Однако лексика, используемая в Твиттере, очень разнообразна, динамично меняется, и очень трудно собрать хоть сколько-нибудь полный словарь оценочной лексики.

Второй подход основан на машинном обучении с учителем, когда необходимо разметить обучающую выборку твитов, проставив экспертную оценку тональности. Но размеченная выборка всегда ограничена по объему, и поэтому по мере прохождения времени качество построенного классификатора существенно ухудшается.

В данной работе для русского языка будут исследованы возможности комбинированного подхода к анализу тональности твитов. Исследуемый подход заключается в том, что при построении представления твитов для автоматической классификации будут порождаться специализированные словарные признаки на основе нескольких словарей, включая автоматически порожденные и ручной. Таким образом, классификатор, построенный на основе

машинного обучения, будет обучаться не только вкладу отдельных слов и других признаков в позитивную или негативную тональность, но и использованию конкретных словарей в целом. Это повышает обобщающие способности классификатора, поскольку в обучающих данных может встретиться одно подмножество слов из конкретного словаря, а в тестовых данных другое подмножество слов из этого же словаря. В этом случае лексические признаки, использующие конкретные слова, не смогут помочь классификации твитов, а словарные признаки могут оказаться достаточно полезными.

Подход исследуется на данных открытых тестирований по автоматическому анализу тональности твитов на русском языке *SentiRuEval*, проведенных в 2015-2016 гг. [2, 3]. В рамках данных тестирований были извлечены сообщения Твиттера о банках и телекоммуникационных компаниях. Участникам было необходимо определить, является ли данное сообщение позитивным, негативным или нейтральным относительно репутации упоминаемой компании.

1. Обзор близких подходов

На протяжении нескольких лет в рамках зарубежной конференции *SemEval* проводятся соревнования по тональной классификации сообщений в сети Твиттер. Начиная с *SemEval-2013*, в проводимых тестированиях рассматривалось два типа задач: *A* – классификация на уровне фраз и *B* – на уровне всего сообщения [4, 5].

Авторы подхода с наилучшим результатом *SemEval-2013* [6] построили дополнительный корпус на основе данных сети Твиттер и автоматической разметки твитов по тональности. Сообщения в Твиттере содержат большое количество метаинформации (#хэштеги, эмотиконы), пользуясь которой можно принять решение о тональности сообщений. Полученная таким образом размеченная коллекция используется в целях обучения классификатора, а также для составления *лексикона оценочных слов и выражений*.

Лексиконы представляют собой словарь пар $\langle w, s \rangle$, в котором для каждого входящего в него слова *w* определена позитивная или негативная тональность $s \in R$. Это позволяет отобразить фразы сообщения в численную оценку тональности и, как следствие, получить приближенную тональную окраску сообщения в целом. Как построенный лексикон (на основе точечной взаимной информации *PMI* [6]), так и лексиконы на основе других корпусов[1] использовались авторами [7] для составления набора признаков сообщений для машинного обучения. Авторы также приводят ряд признаков, которые позволяют отделить тональные сообщения от нейтральных[2], что положительно сказывается на работе классификатора.

Схожий подход [8], примененный в 2014 г., демонстрирует прирост качества при использовании автоматической тональной разметки сообщений с целью обучения классификатора и создания дополнительных лексиконов. Результатом таких исследований стало второе место на соревнованиях *SemEval-2014*. Авторы подхода обучают SVM-классификатор на большой коллекции твитов, размеченных положительными и отрицательными хэштегами и эмотиконами, представляя твиты как набор униграмм и биграмм. Обучение производится для задачи классификации по тональности на два класса: положительный и отрицательный. После того как классификатор обучен, из полученной модели извлекаются веса слов и биграмм. Эти веса отражают общую тональность этих униграмм и биграмм, поскольку классификатор настраивает веса для решения задачи определения тональности. Подобно [6] на основе извлеченного словаря порождаются признаки для классификации твитов на данных *SemEval-2013*. Показано, что предложенный подход к извлечению лексикона позволяет улучшить качество классификации твитов.

В статье рассмотрено построение классификатора для русскоязычных сообщений сети Твиттер с использованием идей подходов [7, 8]. Тестирование подхода осуществляется на данных соревнований *SentiRuEval-2015/2016* [2, 3].

2. Описание подхода

Предлагаемый подход на основе методов машинного обучения состоит в следующем.

[1] Словари оценочных слов *MPQA, Bing Liu, NRC-Emoticon, Sentiment-140*, использовались для добавления признаков в работах [7, 8].

[2] Учет количества слов в верхнем регистре; учет слов с большим числом повторения букв; признаки на основе пунктуации и т.д [7].

Предварительно собираются несколько словарей оценочной лексики, которые включают в себя как автоматически порожденные словари, так и доступные ручные словари оценочной лексики. Такие словари обычно содержат совокупность оценочных слов и выражений, которым присвоен числовой вес их тональности так, что негативная тональность соответствует отрицательным весам, а позитивная тональность – положительным. При этом, чем больше модуль веса, тем больше выраженность тональности слова. Таким образом, слова с весами, близкими к 0, предполагаются практически нейтральными. Эти словари будут использоваться при формировании словарных признаков для машинного обучения, т.е. в качестве признака будет использоваться идентификатор словаря и некоторая числовая величина, полученная на основе весов слов из этого словаря, упомянутых в сообщении.

В качестве метода машинного обучения будет использоваться Метод опорных векторов SVM, поскольку во многих работах было показано, что это лучший метод классификации текстов вообще и классификации текстов по тональности [8, 9]. Используется реализация метода SVM в библиотеке LibSVM [10].

В данной работе используются автоматические словари оценочных слов, а также вручную сделанный и опубликованный словарь *RuSentiLex* [11].

2.1. Автоматическое порождение лексиконов

Для автоматического порождения лексиконов используются большие коллекции твитов. Сообщения в Твиттере часто содержат эмотиконы (смайлики) или хэштеги, проставляемые пользователями и которые можно расклассифицировать как положительные или отрицательные по тональности, например:

Написала русский без единой ошибочки
#счастье #радость

Хотя далеко не всегда тональность твита соответствует тональности проставленного эмотикона и/или хэштега, например, из-за высказываемой иронии, в целом можно предположить, что позитивные слова чаще встречаются с позитивными эмотиконами и хэштегами, а негативные слова – с негативными. Таким образом, можно автоматически разметить твиты на основе эмотиконов и хэштегов на позитивные и негативные [8].

Построение лексикона производится на основе *меры точечной взаимной информации* [6]:

$$PMI(t_1, t_2) = log_2 \frac{P(t_1 \cap t_2)}{P(t_1) \cdot P(t_2)} \quad (1)$$

Такая мера оценивает, насколько t_1 и t_2 встречаются вместе чаще, чем, если бы они использовались независимо друг от друга. В качестве второго аргумента t_2 рассматривается метка, которая будет соответствовать одному из тональных классов, проставленная на основе соответствующих эмотиконов и хэштегов:

- *Excellent* – положительная тональность;
- *Poor* – негативная тональность.

Таким образом, относительно каждого t_1 можно установить его числовую тональную ориентацию:

$$SO\ (t) = PMI\ (t, Excellent) - PMI\ (t, Poor). \quad (2)$$

В результате лексикон строится следующим образом:

$$S : \{\langle t, SO(t)\rangle \mid t \in K\}. \quad (3)$$

На основе описанного подхода были составлены два автоматических лексикона представленных в Табл. 1, где T_+, T_- – количество термов с положительной и негативной оценкой соответственно, $\sum$ – общее число термов в лексиконе.

Лексикон l_1 составлен на основе корпуса сообщений сети Твиттер, подготовленный Ю. Рубцовой [12] (далее Twitter-corpus). Корпус собирался в период с ноября 2013 до конца февраля 2014 г. с помощью API сети Твиттер. За это время было собрано 15 млн сообщений, которые затем были отфильтрованы (например, были удалены дублирующиеся сообщения), а также классифицированы на положительные и отрицательные на основе содержащихся в сообщениях эмотиконов. В настоящее время коллекция содержит 100 тыс. позитивных и столько же негативных сообщений. В Табл. 2 приведены примеры слов с наибольшими значениями позитивной и негативной тональности составленного лексикона l_1.

Табл. 1. Параметры созданных лексиконов

L	T_+	T_-	$\sum$
l_1	67441	54049	121490
l_2	7370	228721	236091
l_3	2774	7148	10668

Табл. 2. Пример наиболее тональных термов лексикона l_1

Положительные	Оценка	Отрицательные	Оценка
#улыбнуло	+4.71	теракт	-6.88
позаимствовать	+4.42	некролог	-6.20
крутотень	+4.35	погибший	-6.16
Еееее	+4.35	#сми	-5.12
бесподобный	+4.21	траур	-5.08
Ржач	+4.21	критический	-5.03
Позитив	+4.21	поч	-4.89

Лексикон l_2 был составлен на основе сообщений сети Твиттер, собранных в течение января 2016 г. Он был собран для того, чтобы отразить новую лексику, возможно появившуюся в Твиттере. Сообщения сети извлекались с помощью *Streaming Twitter API*[3]. Определение тонального класса каждого из сообщений производилось на основе содержащихся в нем эмотиконов. В Табл. 3 приведены примеры самых тональных термов составленного лексикона l_2.

Ручной лексикон оценочных слов. Словари оценочных слов, созданные экспертами, обычно включают информацию о тональности слов, которая ассоциируется с данным словом по умолчанию. Такая информация может быть полезна в тех случаях, когда слово отсутствовало в обучающей выборке.

Поэтому в качестве еще одного словаря (l_3) в исследовании используется опубликованный в 2016 г. словарь оценочных слов *RuSentiLex*, порожденный автоматически на основе извлечения информации из нескольких источников, а затем проверенный вручную экспертами. Словарь содержит: оценочные слова и словосочетания из тезауруса РуТез; сленговые слова, извлеченные из сообщений Твиттера; слова с негативными или позитивными ассоциациями, извлеченные из новостей (*безработица, инфляция*). Каждая единица словаря (т.е. слово или словосочетание) состоит из набора атрибутов, представленного в Табл. 4.

В данной работе для построения лексикона были рассмотрены только те единицы словаря, которые описывали слово (т.е. словосочетания игнорировались). Из всех атрибутов, представленных в Табл. 4, использовались только *лематизированная форма* и *тональность*. Атрибут тональности преобразовывался в *числовую тональность* (-1 – negative, 0 – neutral, 1 – positive). Поскольку некоторые слова могут быть представлены в нескольких различных тональностях, то будем рассматривать только те из них, тональность которых определена однозначно в рамках всего словаря.

Табл. 3. Пример наиболее тональных термов лексикона l_2

Положительные	Оценка	Отрицательные	Оценка
#badoo	+7.23	злостный	-7.62
#happynewyear	+4.39	ухудшение	-6.18
выздоровление	+3.79	хнык	-6.10
бодрый	+3.74	ужест	-6.03
активный	+3.73	противоречить	-5.85
хахах	+3.54	нерешенный	-5.81

Табл. 4. Формат представления единицы словаря RuSentiLex

Атрибут	Возможные значения	Пример
Слово или фраза	Строка	пресный
Часть речи	Adj, Noun	Adj
Лемматизированная форма	Строка	пресный
Тональность	positive, negative, neutral	negative
Источник тональности	opinion, emotion, fact	emotion
Отсылки к понятиям тезауруса РуТез	Строка (возможно пустая)	Невкусный

В результате, список пар ⟨лемма, числовая тональность⟩ образует результирующий лексикон l_3, объем которого составляет около 10 тыс. слов (Табл. 1*)*.

2.2. Обработка сообщений

Процесс обработки сообщений коллекции сообщений состоит из следующих этапов:

1. Лемматизация слов сообщений с целью получения списка термов[4];

2. Удаление из сообщения следующих термов: символы «Ретвита» (термы со значением «RT»), имена пользователей сети Твиттер (термы с префиксом «@»), URL-адреса. Таким образом, помимо слов естественных языков в сообщении остаются только #хэштеги;

[3] Из всего потока сообщений использовались только русскоязычные сообщения.

[4] Используется пакет *Yandex Mystem*: http://tech.yandex.ru/mystem/

3. Замена некоторых биграмм и унограмм на тональные префиксы. Для выполнения этого этапа, используется предварительно составленный список пар $D_{tone} = \langle t, s \rangle$, где t – терм, а s – тональная оценка («+» или «–»). На этом этапе для каждого терма t_i сообщения m такого, что $t_i \in D_{tone}$ выполняется замена на соответствующую оценку s, которая становится префиксом следующего терма t_{i+1}. Пример преобразования (список используемых слов и словосочетаний приведен в Табл. 5):

Рост числа клиентов может привести
к снижению качества обслуживания
+числа клиентов может привести
к -качества обслуживания

4. Каждое сообщение представляется как вектор термов; в качестве меры весовых коэффициентов используется $tf\text{-}idf$:

$$tf - idf(t, d, D) = tf(t, d) \cdot idf(t, D), \quad (4)$$

где t – терм; d – документ элемент коллекции всех документов D; функция tf определяет *частоту встречаемости* терма t в документе d:

$$tf(t, d) = \frac{n_i}{|d|}. \quad (5)$$

Под idf понимается *инвертированная частота терма* в коллекции документов. Чем больше значение $idf(t, d)$, тем выше уровень «уникальности» терма t в коллекции документов D:

$$idf(t, D) = log \frac{|D|}{|\{d_i \in D | t_i \in d_i\}|}. \quad (6)$$

Помимо термов, составляющих вектор сообщения, вычисляются еще следующие признаки:

- на основе числа тональных префиксов: $E_+ - E_-$, где E_+ – число префиксов «+»; E_- – число префиксов «–»;
- подсчет количества термов, записанных в ВЕРХНЕМ РЕГИСТРЕ [7];
- подсчет числа знаков препинания: «?», «...», «!» [7].

2.3. Словарные признаки представления сообщения

На основе имеющихся лексиконов оценочных слов вычисляются несколько признаков сообщений. Пусть L – множество составленных лексиконов, тогда относительно каждого лексикона $l_j \in L$ для сообщения m, вычисляются признаки:

$$x_\Sigma := \sum_{i=1}^{N} l_j(t_i),\ x_{min} := \min_{i=1..N} l_j(t_i),$$
$$x_{max} := \max_{i=1..N} l_j(t_i), \quad \text{где } t_i \in m \quad (7)$$

Если терм t_i отсутствует в лексиконе, то в качестве коэффициента рассматривается $l_j(t_i) = 0$. Если у терма t_i присутствует тональный префикс «–», то рассматривается значение оценки тональности $-l_j(t_i)$.

Обозначим x_* — значение любого признака на основе лексикона (1). Для получения окончательных значений признаков, применяется нормализация полученных значений в диапазоне $[-1, 1]$ на основе преобразования:

$$\begin{cases} s = 1 - e^{|x|}, & x_* > 0 \\ s = -\left(1 - e^{-|x|}\right), & x_* < 0 \end{cases} \quad (8)$$

Табл. 5. Список используемых слов и словосочетаний, используемых для замены на тональные префиксы «+»/«–»

Положительные	Отрицательные
Абсолютно, абсолютный, безумно, безусловный, весьма, все время, вырасти, гораздо, дикий, жутко, жуть как, заметный рост, запредельный, запросто, значительный, избыток, колоссально, крайне, масштабность, масштабный, много, намного, нарастание, настолько, не просто, невероятно, невообразимый, немыслимый, необычайно, нет никакой, особенно, острый, очень, по-настоящему, повышать, повышение, полный, порядочно, просто, рост, сильнейший, сильно, совершенно, совсем, увеличить, усилить, явный	*Без, даже если, запрет, защита от, избавиться, избавление, имитировать, конец, ликвидация, ликвидировать, лишать, лишаться, лишить, не, недостаточно, нельзя назвать, нет, никакой, ничего, ослабить, ослаблять, острый, отсутствие, отсутствовать, падение, пересекать, перестать, потеря, преодоление, противодействовать, разрушать, разрушение, разрушить, снижаться, снижение, снимать, снять, спад, терять, уменьшать, уменьшение, уменьшить, утрата, утратить, утрачивать, якобы*

2.4. Текстовые коллекции

Для обучения классификатора используются соответствующие коллекции данных соревнований *SentiRuEval* 2015 и 2016 г., которые содержат твиты, относящиеся к банкам ($bank$) и телекоммуникационным компаниям (tcc). В Табл. 6 N_+, N_0, N_- – число сообщений положительного, нейтрального и негативного классов соответственно; $\sum$ – общее число сообщений в коллекции.

Поскольку в предоставляемых данных число тональных сообщений существенно уступает объему класса нейтральных сообщений, то дополнительно была создана *сбалансированная обучающая коллекция*. В работе [9], применительно к классификаторам *Наивного Байеса* и *SVM*, отмечается существенный прирост качества при использовании коллекций сбалансированного типа.

Для решения подобной задачи воспользуемся упомянутым ранее корпусом Twitter-corpus[5], в котором каждое сообщение автоматически распределено в одну из тональных групп: *positive* и *negative*. Для построения сбалансированной коллекции требуется существенно меньшее число сообщений, чем предлагается в тональном корпусе. В связи с этим, выберем небольшой набор наиболее эмоциональных сообщений:

- пусть имеется лексикон *l, построенный* на основе Twitter-corpus для определения списка наиболее эмоциональных термов;
- сообщение m будем считать *наиболее эмоциональным*, если для него выполнено следующее условие:

$$max_{i=1..N}|l(t_i)| > B, \quad (9)$$

где B –пороговое значение; t_i – термы сообщения m; N – общее количество термов в сообщении m.

Таким образом, были сбалансированы коллекции, описанные в Табл. 6. Параметры дополнительно составленных коллекций представлены в Табл. 7, где N_+, N_0, N_- – размер класса коллекции (положительного, нейтрального, негативного соответственно); $\sum$ – общее число сообщений в коллекции.

[5] Корпус коротких текстов на русском языке на основе «постов» сети *Twitter*: study.mokoron.com

Табл. 6. Обучающие коллекции

Название	N_+	N_0	N_-	$\sum$
I^{15}_{bank}	356	3482	1077	4915
I^{15}_{tcc}	956	2269	1634	4859
I^{16}_{bank}	1354	4870	2550	8783
I^{16}_{tcc}	704	6756	1741	9102

Табл. 7. Сбалансированные обучающие коллекции

Название	N_+	N_0	N_-	$\sum$
B^{15}_{bank}	3400	3400	3400	10400
B^{15}_{tcc}	2269	2269	2269	6888
B^{16}_{bank}	6756	6756	6756	20268
B^{16}_{tcc}	4870	4870	4870	14610

Табл. 8. Эталонные тестовые коллекции, предоставленные организаторами

Название	N_+	N_0	N_-	$\sum$
$BANK_{15}$	348	3548	668	4564
TCC_{15}	399	2601	916	3916
$BANK_{16}$	312	2240	772	3324
TCC_{16}	226	1016	1054	2296

Обученная на таких коллекциях классификационная модель применяется к *тестовым данным*. Тестовые коллекции содержат неразмеченные сообщения, отнесенные по умолчанию к нейтральному классу. Таким образом, классификационной модели требуется выделить из этих сообщений тональные сообщения и проставить соответствующий класс.

Проверка модели осуществляется с помощью *эталонных коллекций*. Они находятся в открытом доступе сразу после окончания проведения соревнований. В Табл. 8 приводятся характеристики эталонных коллекций.

3. Тестирование подхода на данных SentiRuEval

Качество работы подготовленных моделей оценивается на основе F_1 меры:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \quad (10)$$

Такая мера позволяет одновременно учитывать результаты следующих параметров относительно некоторого класса:

- точность (*Precision*) – количество сообщений, которые классификатор правильно

Табл. 9. Настройки векторизации сообщений
Знак "X" показывает используемые признаки в каждой модели

Модель	Термы	Доп. признаки	l_1			l_2			l_3		
			$\sum$	$\max_{1..N}$	$\min_{1..N}$	$\sum$	$\max_{1..N}$	$\min_{1..N}$	$\sum$	$\max_{1..N}$	$\min_{1..N}$
1	X										
2	X	X									
3	X	X	X								
4	X	X	X			X					
5	X	X	X			X			X		
6	X	X	X	X	X	X	X	X	X	X	X

отнес к эталонному классу, по отношению ко всему объему сообщений, определенных системой в этот класс;

- полнота (*Recall*) – количество сообщений, которые классификатор правильно отнес к эталонному классу, по отношению к количеству всех сообщений эталонного класса.

В случае, когда необходимо оценить качество работы по метрике F_1 относительно нескольких классов, то применяются усреднения меры. Различают микро- и макроусреднения. Для вычисления усредненной F_1 меры определяются параметры полноты и точности с соответствующим усреднением относительно интересующих нас классов.

Макроусреднение придает одинаковый вес каждому из усредняемых классов, в то время как при микроусреднении вес учитывается на основе числа документов в классе. При использовании F_1-макро смещение среднего значения будет производиться в сторону того класса, для которого классификатор сработал лучше. В тоже время, при использовании F_1-микро, смещение будет произведено в сторону наибольшего класса [13].

В данной задаче нас интересует качество определения тональных твитов, т.е. сообщений соответствующих положительному (*Positive*) и отрицательному (*Negative*) классам. Будем рассматривать результаты с *макроусреднением* F_1 *меры* (7):

$$F_{1-macro}^{PN} = 2 \cdot \frac{P_{macro}^{PN} \cdot R_{macro}^{PN}}{P_{macro}^{PN} + R_{macro}^{PN}} \qquad (11)$$

В тестировании участвуют проходы с настройками, представленными в Табл. 9. Все проходы будут одновременно протестированы в двух областях: BANK – отзывы о банках, TCC – отзывы о телекоммуникационных компаниях.

3.1. Тестирование на данных SentiRuEval-2015

Рассмотрим качество работы классификационных моделей каждого из проходов для коллекций $BANK_{15}$ и TCC_{15} данных *SentiRuEval-2015* в зависимости от типа обучающей коллекции (*I* – несбалансированной, *B* – сбалансированной). Результаты проведенного тестирования зафиксированы в Табл. 10, где жирным шрифтом отмечены лучшие результаты относительно каждой задачи. Результаты *I* получены на исходных обучающих коллекциях, результаты *B* – на сбалансированных.

При применении представленного прохода можно видеть, что, независимо от типа тестируемой коллекции, добавление признаков на основе лексиконов стабильно повышает качество классификации. Так, начиная с прохода №3, наблюдается прирост качества.

Наибольший результат достигается в проходе №6, где по каждому из лексиконов Табл. 1 вычисляются все признаки: сумма, минимум и максимум тональной оценки слов в сообщении. Добавление последних двух признаков в векторизацию сообщений изменило результат

Табл. 10. Результаты $F_{1-macro}^{PN}$ тестирования на коллекции *SentiRuEval-2015*

№	*BANK*		*TCC*	
	I	*B*	*I*	*B*
1	36.70	41.93	46.10	45.67
2	38.04	41.78	46.91	45.30
3	40.32	42.80	47.82	46.97
4	41.41	42.75	49.12	47.56
5	41.28	43.31	49.10	47.41
6	42.21	**45.56**	**50.12**	49.11

качества на +2.25 для коллекции $BANK$, и на +1.02 для коллекции TCC, если сравнивать с проходом №5.

Если проанализировать результат с точки зрения влияния балансировки, то здесь она улучшила результаты для коллекции *BANK*. Средний прирост по каждому проходу, при сравнении результатов I с B, составил $+3.02$.

Твиты коллекции TCC классифицируются несколько лучше. Эта особенность отмечается в [1], что объясняется ухудшением ситуации на Украине в период сбора сообщений для тестовой коллекции *BANK*, которая вызвала проблемы в работе банков, что проявилось в соответствующем потоке твитов. Так, например слово «санкции» часто несет негативный характер в тестовой выборке, в то время как в обучающей коллекции аналогичное слово является нейтральным.

Если сравнить лучший полученный в данной статье результат с результатами участников соревнований *SentiRuEval-2015*[6] (Табл. 11), то наилучший результат среди участников по метрике $F^{PN}_{1-macro}$ на коллекции $BANK_{15}$ составляет 35.98 (участник №4), а для коллекции TCC_{15} – 48.04 (участник №3). Таким образом, текущий проход достиг лучшего результата по сравнению с результатами участников *SentiRuEval-2015*. Сравнивая лучший результат соревнования с проходом №6, можно видеть, что отрыв предложенного прохода относительно победителя *SentiRuEval-2015* по метрике $F^{PN}_{1-macro}$ составляет $+9.58$ для задачи $BANK$, и $+2.08$ для задачи TCC.

3.2. Участие в соревнованиях SentiRuEval-2016

В январе 2016 г. соревнования по тональной классификации сообщений сети Твиттер были продолжены. В качестве областей также были выбраны сообщения о банках и телекоммуникационных компаниях. Тестовые коллекции $BANK_{16}$ и TCC_{16} составлялись в период с сентября 2015 и до начала 2016 г. (Табл. 8). Таким образом, имея весь набор необходимых коллекций для тестирования, рассмотрим результаты (Табл. 12), которые были получены при применении разных проходов Табл. 9. Для обучения классификатора были рассмотрены несбалансированные (*I*) и сбалансированные (*B*) версии 2016 г., представленные в Табл. 6 и 7.

Табл. 11. Сравнение лучших проходов участников соревнований SentiRuEval-2015 с описанным в статье проходом

Участник	*BANK*		*TCC*	
	$F^{PN}_{1-macro}$	$F^{PN}_{1-micro}$	$F^{PN}_{1-macro}$	$F^{PN}_{1-micro}$
1	29.86	32.26	34.19	38.0
2	33.54	36.56	48.29	53.62
3	–	–	48.04	50.94
4	35.98	34.30	46.70	50.60
8	32.76	34.32	38.43	42.83
9	–	–	35.27	37.65
10	35.20	33.70	44.77	52.82
статья	**45.56**	**51.36**	**50.12**	**54.41**

В результате можно наблюдать схожую картину, что и в проходе 4. Во всех задачах наблюдается рост результатов, если увеличивать число признаков на основе лексиконов (проходы 3 – 6).

При тестировании на задаче $BANK$, максимальный результат наблюдается в проходе 6, и составляет $F^{PN}_{1-macro} = 51.73$. Как и в тестировании прохода 4, результат последнего дает средний прирост в $+1.6$ на коллекции I и в $+3.0$ при обучении на B при сравнении с проходами 3 – 5.

Для задачи *TCC* ситуация в целом аналогична банковской коллекции, за исключением того, что чуть лучший результат достигается в проходе №5, где $F^{PN}_{1-macro} = 52.90$, что на $+0.27$ лучше результата последнего прохода. В итоге, результат прохода №6 в очередной раз показывает полезность применения нескольких разных признаков на основе лексиконов.

Табл. 12. Результаты $F^{PN}_{1-macro}$ тестирования на коллекции SentiRuEval-2016

№ I	*BANK*		*TCC*	
	B	*I*	*B*	
1	48.77	45.53	48.32	50.90
2	50.24	47.36	49.48	50.69
3	50.28	47.53	48.90	51.95
4	50.45	47.13	49.31	52.72
5	49.72	46.99	50.55	**52.90**
6	**51.73**	50.25	51.66	52.63

[6] Результаты участников соревнований *SentiRuEval-2015*: https://docs.google.com/spreadsheets/d/1IxGFhGO4zS5t356FePlMdJQ5IU6-tkpetoOzoXpGLPs/edit#gid=0.

Сравнивая результаты при использовании разных типов обучающих коллекций, прирост можно наблюдать на задаче *TCC*. При использовании версии B он составляет $+2.26$ в среднем для каждого прохода при сравнении с I. Для задачи $BANK$ ситуация обратная, поскольку разность в среднем по каждому из проходов качество при обучении на версиях B и I составляет 2.73.

В следующем разделе рассмотрим, насколько изменится эта разница после настройки SVM классификатора.

3.3. Изменение настроек SVM классификатора

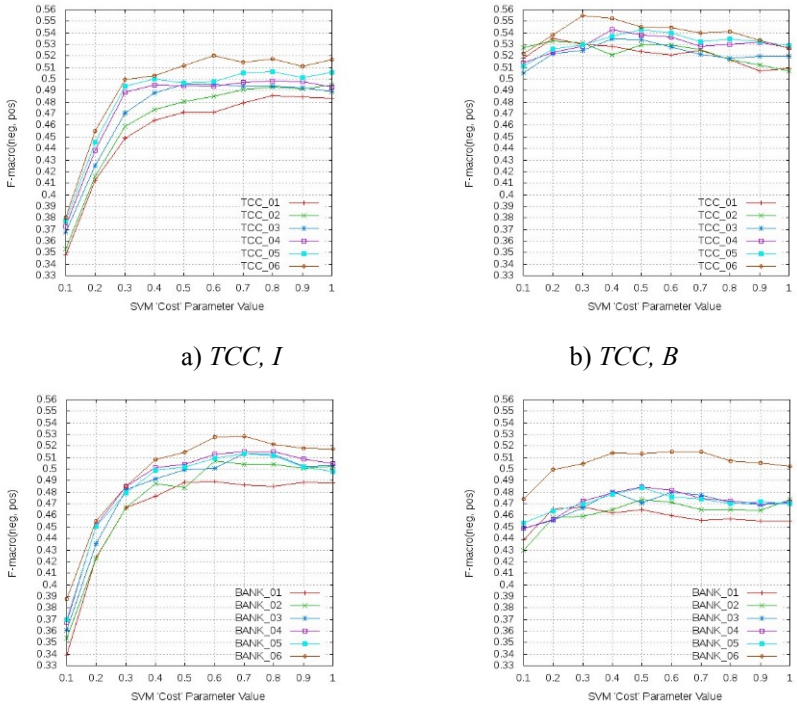
Рассмотрим динамику изменения результатов, в зависимости от величины отступа для разделения классов SVM классификатором. В пакете LibSVM отступ изменяется на основе параметра C.

На Рис. 1 показаны результаты прогонов с изменением значения параметра, отвечающего за величину отступа в диапазоне $[0.1, 1]$ с шагом 0.1. Ранее, результаты Табл. 12 были получены при максимальном значении параметра в рассматриваемом диапазоне, т.е. при $C = 1$. Лучшие результаты по каждому были вынесены в Табл. 13.

Сравнивая результаты относительно использования разных типов обучающих коллекций, для несбалансированных коллекций (Рис. 1, a, c) можно наблюдать резкий спад при использовании малого значения параметра ($C < 0.4$). Что касается всех результатов, то положительный эффект от добавления признаков на основе лексиконов сохраняется. В большинстве случаев лучший результат достигается при подходе №6, что можно было наблюдать в Табл. 10 и 12.

Табл. 13. Наилучшие результаты $F^{PN}_{1-macro}$ для каждой задачи по каждому из прогонов. Кроме того фиксируется значение параметра Cost, при котором достигается такой результат

№	*BANK*		*TCC*	
	I /cost	B /cost	I/cost	B/cost
1	$48.93_{/0.6}$	$46.71_{/0.3}$	$48.57_{/0.7}$	$53.50_{/0.2}$
2	$50.69_{/0.6}$	$47.33_{/0.5}$	$49.48_{/1.0}$	$53.31_{/0.3}$
3	$51.25_{/0.8}$	$48.00_{/0.6}$	$49.56_{/0.5}$	$53.46_{/0.4}$
4	$51.52_{/0.8}$	$48.43_{/0.5}$	$49.84_{/0.8}$	$54.25_{/0.4}$
5	$51.33_{/0.7}$	$48.39_{/0.5}$	$50.55_{/1.0}$	$54.26_{/0.5}$
6	$\mathbf{52.82}_{/0.7}$	$51.50_{/0.7}$	$52.02_{/0.6}$	$\mathbf{55.46}_{/0.3}$



a) *TCC, I* b) *TCC, B*

c) *BANK, I* d) *BANK, B*

Рис. 1. Влияние параметра штрафной функции SVM классификатора (C) на результаты прогонов для задач TCC и $BANK$ SentiRuEval-2016 на разных обучающих коллекциях (I/B)

Наибольший отрыв такого подхода отмечается на Рис. 1, d. Кривыми на графиках обозначены проходы, номера которых соответствуют настройкам Табл. 9. Значение параметра измерялось в пределах [0.1,1] с шагом 0.1.

Посмотрев на работу классификационных моделей с точки зрения изменения настроек, можно рекомендовать использование лексиконов для составления дополнительных признаков.

3.4. Сравнение с результатами участников тестирования SentiRuEval-2016

На SentiRuEval-2016 качество классификации тестовых коллекций оценивалось по микро- и макроусреднению F_1^{PN}. Из всех полученных результатов[7] в Табл. 14 включен лучший результат каждого участника.

Проход, рассмотренный в данной статье, тестировался под №1. После окончания тестирования был проделан ряд улучшений: добавление признаков максимума/минимума на основе лексиконов, и произведена настройка классификатора. Результатом таких улучшений стал результат прохода №6 (Табл. 13). Этот результат с дополнительно вычисленной мерой $F_{1-micro}^{PN}$ был добавлен к сравнению с участниками в Табл. 14. Описание подходов участников соревнований представлено в Табл. 15.

Табл. 14. Сравнение лучших прогонов участников соревнований *SentiRuEval-2016* с описанным в статье проходом

Участник	*BANK*		*TCC*	
	$F_{1-macro}^{PN}$	$F_{1-micro}^{PN}$	$F_{1-macro}^{PN}$	$F_{1-micro}^{PN}$
1	46.83	50.22	52.86	66.32
2	**55.17**	**58.81**	**55.94**	65.69
3	34.23	35.24	39.94	39.94
4	37.30	39.67	49.55	62.52
5	38.59	46.40	34.99	40.44
6	23.98	31.27	35.45	52.63
7	47.10	51.28	48.42	63.74
8	44.92	47.05	48.71	57.45
9	51.95	55.95	54.89	68.22
10	46.59	50.53	50.55	62.54
статья	52.82_2	54.81_3	55.46_2	$\mathbf{70.47_1}$

[7] Таблица с результатами прогонов всех участников соревнований, а также настройками прогонов: https://docs.google.com/spreadsheets/d/1rCaklClawfnnSnyk4q8CW4zWuO3P38DSrLw_f2wyyjg/edit#gid=0.

Табл. 15. Сравнение лучших прогонов участников соревнований *SentiRuEval-2016* с описанным в статье проходом

Участник	Настройки лучших проходов
1	Настройки прохода №5 из Табл. 9. Для обучения использовались сбалансированные коллекции B_{bank}^{16} и B_{tcc}^{16} для задачи *BANK* и *TCC* соответственно [14].
2	Рекуррентная нейронная сетка (*LSTM*); в качестве признаков *Word2Vec*, обученный на внешней коллекции (посты и комментарии из социальных сетей) [15].
4	Словарные признаки + признаки метаклассификаторов (логистическая регрессия, ридж-регрессия, классификатор на основе градиентного бустинга и классификатор на основе нейронной сети) и линейный *SVM* в качестве классификатора.
8	Поиск эмоциональных слов по словарю (200 тыс. словоформ), правила их комбинирования на основе синтаксического анализа; применение онтологических правил, характерных для данной предметной области.
9	*SVM*: униграммы, биграммы, словарь *RuSentiLex*, учет частей речи, многозначных слов (автоматический словарь коннотаций по новостям для ТКК задачи).
10	*SVM*, в качестве признаков использовались униграммы, подвергшиеся преобразованиям (*не + слово = один признак*, множественные повторения символов заменяются двукратным; ссылки, ответы, даты, числа – заменяются паттернами и другие преобразования); подключение словаря *RuSentiLex*.

Нужно отметить, что все лучшие проходы на тестировании (1, 2, 9, 10) использовали дополнительную информацию, помимо обучающей выборки, что позволило им лучше преодолеть различие между обучающей и тестовой выборкой (Табл. 14, 15). Лучший результат участника №2 связан с использованием дополнительной информации, полученной на основе обработки большого объема комментариев в социальных сетях, представленных в виде сокращенного векторного пространства с помощью инструмента *Word2Vec*[8].

[8] https://radimrehurek.com/gensim/models/word2vec.html

Заключение

В статье был описан подход к решению задачи тональной классификации сообщений сети Твиттер с использованием лексиконов. Они нашли свое применение в качестве дополнительных признаков в векторе представления сообщений, а также для отбора сообщений с целью увеличения объема обучающих коллекций.

Качество работы классификатора было протестировано на данных тестирований систем анализа тональности русского языка *SentiRuEval-2015* и *SentiRuEval-2016*. Добавляя в сообщения признаки на основе лексиконов, а также применяя лексиконы для расширения коллекций наиболее тональными сообщениями, удалось добиться стабильного роста качества классификации.

После проведения настройки классификатора, рассматриваемый в статье подход можно считать одним из наиболее успешных на сегодняшний день для проведения тональной классификации сообщений различных областей русскоязычной сети Твиттер.

Литература

1. Лукашевич Н., Рубцова Ю. Объектно-ориентированный анализ твитов по тональности: результаты и проблемы // Труды XVII Международной конференции DAMDID/RCDL'2015 «Аналитика и управление данными в областях с интенсивным использованием данных». № 2015.
2. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog-2015, Vol. 2, 2015. pp. 3-13.
3. Loukachevitch N., Rubtsova. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis, Proceedings of International Conference Dialog-2016 // Proceedings of International Conference Dialog-2016, 2016.
4. Nakov P., Kozareva Z., Ritter A., Rosental S. SemEval-2013 Task 2: Sentiment Analysis in Twitter // Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, June 2013. pp. 312-320.
5. Rosental S., Nakov P., Ritter A., Stoyanov V. SemEval-2014 Task 2: Sentiment Analysis in Twitter // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), August 2014. pp. 73-80.
6. Turney P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002. pp. 417-424.
7. Saif M., Kiritchenko S., Xiaodan Z. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets // In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Vol. 2, June 2013. pp. 321-327.
8. Severyn A., Moschitti A. On the Automatic Learning of Sentiment Lexicons // Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, 2015. pp. 1397-1402.
9. Pang B., Lee L., Vaithyanathan S. 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, 2002. pp. 79-86.
10. Chih-Chung C., Chih-Jen L. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011. pp. 2(3):27:1–27:27.
11. Loukachevitch N., Levchik A. Создание лексикона оценочных слов русского языка РуСентиЛекс // Open Semantic Technologies for Intelligent Systems (OSTIS-2016), 2016. pp. 377-382.
12. Рубцова Ю. Построение корпуса текстов для настройки тонового классификатора // Программные продукты и системы. №1(109). 2015. pp.72-78.
13. Asch V.V. Macro- and micro-averaged evaluation measure [[basic draft]], 2013.
14. Rusnachenko N. Use of lexicons to improve quality of sentiment classification // Proceedings of International Conference Dialog-2016, June 1-4 2016.
15. Arhnipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of neural network architectures for sentiment analysis of russian tweets // Proceedings of the International Conference "Dialogue 2016", June 2016.

Русначенко Николай Леонидович. Специалист Московского государственного технического университета (МГТУ) им. Н.Э Баумана. Окончил МГТУ им. Н.Э Баумана в 2016 году. Автор одной печатной работы. Область научных интересов: компьютерная лингвистика, анализ тональности сообщений, информационный поиск. kolyarus@yandex.ru

Лукашевич Наталья Валентиновна. Ведущий научный сотрудник НИВЦ МГУ им. М.В. Ломоносова. Канд. физико-математических наук. Автор 180 печатных работ, в том числе двух монографий. Область научных интересов: компьютерная лингвистика, информационный поиск. louk_nat@mail.ru

Methods of lexicon integration with machine learning for sentiment analysis system

N.L. Rusnachenko N.V. Loukachevitch

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. Varieties of features were implemented to improve the quality of message classification, especially sentiment score features based on a set of sentiment lexicons. We study the impact of different training types (balanced/imbalanced) and its volumes, and advantages of applying several lexicon-based features. Before SentiRuEval-2016, the classifier was tuned 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

References

1. Loukachevitch N., Rubtsova Yu. Entity-Oriented Sentiment Analysis of Tweets: Results and Problem // XVII International Conference DAMDID/RCDL'2015 ««Data Analytics and Management in Data Intensive Domains»., № 2015.
2. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: testing object-oriented sentiment analysis systems in Russian // Proceedings of International Conference Dialog-2015, Vol. 2, 2015. pp. 3–13.
3. Loukachevitch N., Rubtsova. SentiRuEval-2016: Overcoming Time Gap and Data Sparsity in Tweet Sentiment Analysis, Proceedings of International Conference Dialog-2016 // Proceedings of International Conference Dialog-2016, 2016.
4. Nakov P., Kozareva Z., Ritter A., Rosental S. SemEval-2013 Task 2: Sentiment Analysis in Twitter // Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, June 2013. pp. 312–320.
5. Rosental S., Nakov P., Ritter A., Stoyanov V. SemEval-2014 Task 2: Sentiment Analysis in Twitter // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), August 2014. pp. 73–80.
6. Turney P. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002. pp. 417–424.
7. Saif M., Kiritchenko S., Xiaodan Z. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets // In Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013), Vol. 2, June 2013. pp. 321–327.
8. Severyn A., Moschitti A. On the Automatic Learning of Sentiment Lexicons // Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL, 2015. pp. 1397--1402.
9. Pang B., Lee L., Vaithyanathan S. 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, 2002. pp. 79–86.
10. Chih-Chung C., Chih-Jen L. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011. pp. 2(3):27:1–27:27.
11. Loukachevitch N., Levchik A. Creating a General Russian Sentiment Lexicon // Open Semantic Technologies for Intelligent Systems (OSTIS-2016), 2016. pp. 377–382.
12. Rubtsova Yu. Constructing a corpus for sentiment classification training // Software & Systems, No. №1(109), 2015. pp. 72-78.
13. Asch V.V. Macro- and micro-averaged evaluation measure [[basic draft]], 2013.
14. Rusnachenko N. Use of lexicons to improve quality of sentiment classification // Proceedings of International Conference Dialog-2016, June 1-4 2016.
15. Arhnipenko K., Kozlov I., Trofimovich J., Skorniakov K., Gomzin A., Turdakov D. Comparison of neural network architectures for sentiment analysis of russian tweets // Proceedings of the International Conference "Dialogue 2016", June 2016.

Rusnachenko N.L. Graduate of Bauman Moscow State Technical University (Moscow, Russia), Computer Design and Technology faculty, master degree. Amount of printed papers and monographs: 1. Scientific interests: computational linguistics, sentiment analysis, information retrieval. kolyarus@yandex.ru

Loukachevitch N.V. PhD, leading researcher of Research Computer Center of Lomonosov Moscow State University, Amount of printed works: 180. Scientific interests: computational linguistics, information retrieval. louk_nat@mail.ru