УДК 004.853

ИЗВЛЕЧЕНИЕ И ИСПОЛЬЗОВАНИЕ ОЦЕНОЧНЫХ СЛОВ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОТЗЫВОВ НА ТРИ КЛАССА

Н. В. Лукашевич[1], И. И. Четверкин[2]

Предлагается подход к автоматическому извлечению оценочных слов для заданной предметной области на основе порождения признаков из нескольких текстовых коллекций. Полученные оценочные слова применяются в задаче классификации отзывов на три класса, в которой необходимо разделить отзывы на классы: “понравилось”, “понравилось, но есть замечания”, “не понравилось”. В задаче классификации исследуются разные виды весов для слов, учитываются знаки препинания и слова-операторы, которые могут менять тональность следующих за ними слов. Работа частично поддержана грантом РФФИ № 11–07–00588-а.

Ключевые слова: извлечение знаний из текстов, извлечение оценочных слов, классификация отзывов, машинное обучение.

1. Введение. В настоящее время в Интернете пользователь может найти огромное количество различных отзывов и мнений о товарах и услугах. Автоматический сбор, обработка и аннотирование такого рода информации полезны как для отдельных потребителей, так и для различных компаний.

Мнения пользователей о продуктах и услугах зачастую выражаются с помощью оценочных слов и выражений, которые имеют определенную положительную или отрицательную окраску. Однако невозможно заранее собрать список таких слов и выражений, которые будут применимы для всех предметных областей, поскольку некоторые оценочные выражения употребляются только в конкретных предметных областях, другие являются оценочными в одной области и не являются оценочными в другой.

В настоящей статье мы рассмотрим метод автоматического извлечения оценочных слов на основе нескольких корпусов текстов, которые существуют для многих предметных областей, а именно: корпуса отзывов о сущностях с вручную проставленными потребителями оценками, корпус нейтральных описаний сущностей и нейтрального контрастного корпуса общезначимых новостей. Из указанных корпусов мы извлекаем списки слов, рассчитываем для каждого слова набор статистических характеристик и используем методы машинного обучения для получения качественного списка оценочных слов, характерных для данной предметной области. Эти списки мы используем в дальнейшем для улучшения качества классификации отзывов.

В задаче классификации отзывов можно выделить разные подзадачи. Наиболее простой подзадачей является классификация отзывов на два класса: *положительный* или *отрицательный*. Качество классификации отзывов на два класса с использованием подхода, применяемого для тематической классификации, превышает 80% [1]. В работе [2] указывается, что качество классификации отзывов, основанное на таксономии оценочных групп, составило 90.2%.

Постановка задачи классификации отзывов при количестве классов, большем чем два, существенно отличается. В этой задаче необходимо не просто определить, понравился ли объект или не понравился, а необходимо поставить оценку по некоторой шкале [3]. Однако уже при переходе к задаче разделения отзывов на три класса (“нравится”, “не нравится”, “средне”) качество автоматической классификации существенно падает. Это связано с субъективностью человеческих оценок. В статье [4] авторы проводят исследование о возможности человека отличать отзывы, которые по десятибалльной шкале имеют близкие оценки. Результаты приводятся следующие: если разница больше трех баллов, то точность составляет 100%, двух — 83%, одного балла — 69% и ноль баллов, соответственно, 55%. Таким образом, при разделении отзывов на большое количество классов даже человек показывает низкую точность классификации.

Кроме того, в работе [4] указывается на отличия между стилями оценки различных людей: то, что может соответствовать оценке 5 (по десятибалльной шкале) у одного человека, может выражать то же мнение, что и 7 у другого. Было показано, что при настройке на индивидуальный стиль автора качество

[1] Научно-исследовательский вычислительный центр, Московский государственный университет им. М. В. Ломоносова, Ленинские горы, д. 1, стр. 4, 119991, Москва; вед. науч. сотр., e-mail: louk_nat@mail.ru

[2] Московский государственный университет им. М. В. Ломоносова, факультет вычислительной математики и кибернетики, Ленинские горы, д. 1, стр. 52, 119991, Москва; аспирант, e-mail: ilia2010@yandex.ru

© Научно-исследовательский вычислительный центр МГУ им. М. В. Ломоносова

классификации отзывов на три класса существенно возрастает, достигая 75% правильно классифицированных отзывов. При классификации же коллекции 5394 отзывов от большого количества авторов (494) наилучший результат по правильности классификации составил 66.3%.

В данной работе мы предлагаем метод автоматического извлечения оценочных слов для заданной предметной области и исследуем применение полученного списка оценочных слов наряду с другими факторами для улучшения классификации отзывов на три класса на материале русского языка. В качестве таких факторов используются:

— различные способы установления весов слов,

— использование информации об оценочности слова,

— использование слов-операторов, т.е. таких слов, которые могут менять или усиливать (*не, самый*) оценочную направленность других слов (см., например, [2]),

— длина и структура отзыва,

— использование знаков препинания — так, в работе [5] авторы использовали знаки пунктуации для определения саркастических предложений.

Дальнейшее изложение статьи будет организовано следующим образом: в разделе 2 мы опишем коллекции для экспериментов, в разделе 3 будет представлен наш метод извлечения оценочных слов, раздел 4 описывает признаки для улучшения качества классификации отзывов на три класса, раздел 5 представляет результаты экспериментов и раздел 6 описывает ручную оценку отзывов и сравнение с лучшим алгоритмом.

2. Данные для экспериментов. Для экспериментов мы выбрали предметную область отзывов о фильмах. С рекомендательного портала www.imhonet.ru было собрано 28 773 мнения по различным фильмам. Каждому мнению соответствует оценка от одного до десяти и название фильма, по которому высказано мнение. Этот корпус является основным для работы, назовем его *корпус мнений* (MainSet). Кроме того, мы использовали дополнительные коллекции для извлечения оценочных слов (п. 2.1) и несколько коллекций для тестирования качества классификации (п. 2.2).

Пример мнения: *“Неплохой фильм, главное не выключить его в начале, где он напоминает просто ужасную пародию на Адреналин. Ну а в целом в фильме есть как и положительные (адреналиновые, захватывающие и интересные сцены), так и отрицательные (неоднозначный финал, не везде удачная режиссура) качества”.*

2.1. Дополнительные коллекции для извлечения оценочных слов. Для извлечения оценочных слов было сформировано несколько дополнительных коллекций с более низкой и более высокой концентрацией оценочных слов.

Для формирования контрастной коллекции, в которой концентрация мнений значительно меньше, были собраны 17 680 описаний фильмов. Назовем этот корпус *корпусом описаний.*

В качестве еще одного контрастного корпуса использовался список слов, содержащий количество документов в новостной коллекции (размер 1 миллион документов), в которых встречается данное слово. Условно этот список назовем *новостным корпусом.*

Кроме того, было высказано предположение, что можно выделить некоторые части корпуса мнений, в которых концентрация оценочных слов больше:

1) предложения, заканчивающиеся на “!”;

2) предложения, заканчивающиеся на “. . .”;

3) короткие предложения не более чем из 7 слов;

4) предложения, содержащие слово “фильм” без других существительных.

Условно назовем этот корпус — *малый корпус.*

2.2. Коллекции для тестирования качества классификации отзывов. Тестирование методов автоматической классификации отзывов на три класса проводилось на трех коллекциях.

1. Вышеупомянутый корпус мнений MainSet. В него вошли отзывы на фильмы за 2009 г. или ранее, а также частично за 2010 г.

2. Коллекция из 2353 отзывов на фильмы за 2010 г., исключая те, которые вошли в первый набор (ControlSet).

3. Коллекция из 1214 отзывов, выделенных из второй коллекции для того, чтобы получить распределение по классам, аналогичное большой коллекции (ControlCutSet).

3. Извлечение оценочных слов. В отзывах оценка пользователей содержится прежде всего в оценочных выражениях.

Существуют два основных подхода к автоматическому выделению оценочных слов из текстов. Первый подход базируется на описаниях слов в толковых словарях и тезаурусах. В данном подходе обычно

выбирается небольшое начальное множество слов, которое формируется вручную, и с помощью словарей и/или тезаурусов обогащается и дополняется. Основной принцип заключается в том, что если слово оценочное, то и его синонимы и антонимы будут оценочными (возможна только смена ориентированности). Поэтому, имея слова из начального множества, можно с помощью этих связей составить новое множество, которое будет более богатым и полным [6]. В [7] с помощью словаря толкований терминов выясняется их ориентированность (положительная или отрицательная). Основная идея заключается в том, что слова с одинаковой ориентированностью имеют "похожие" описания в глоссарии.

Второй подход (корпусный) основан на поиске правил и закономерностей употребления оценочных выражений в текстах. В работе [8] описывается метод извлечения оценочных слов и вычисления их полярности на основе частоты их взаимной встречаемости со словами "отлично" и "плохо".

В работе [9] выделение оценочных слов и определение их семантической направленности основано на синтаксических шаблонах и союзах между словами. Основное значение имели союзы И, ИЛИ и НО. Предполагается, что если два прилагательных связаны союзами И или ИЛИ, то они оба являются или не являются оценочными, а также одинаково семантически направленными. В случае союза НО семантическое направление различается. На основе этого принципа был построен классификатор, определяющий семантическую направленность множеств прилагательных с точностью 92%.

Наша работа относится ко второму типу подходов. Предлагаемый метод извлечения оценочных слов основывается на извлечении признаков оценочных слов из нескольких текстовых корпусов и их комбинировании с помощью методов машинного обучения.

3.1. Предлагаемые характеристики извлечения оценочных слов. Оценочные выражения, применяемые пользователями, в значительной степени зависят от предметной области, поэтому необходимо иметь возможность автоматического извлечения таких выражений для заданной предметной области.

Для извлечения оценочных слов используются четыре корпуса текстов: корпус мнений, корпус описаний, корпус новостей и малый корпус с более высокой концентрацией оценочных слов (см. п. 2.1). На основе этих корпусов были вычислены следующие характеристики слов:

— частотность;
— количество документов, в которых встречается слово;
— странность;
— TFIDF (Term Frequency–Inverse Document Frequency);
— отклонение от средней оценки;
— частотность слов, употребляемых с заглавной буквы (в корпусе мнений).

Остановимся более подробно на некоторых из них.

Странность. Для подсчета странности необходимы два корпуса: один — содержащий мнения, другой — контрастный. Идея в том, что слова, которые несут оценки, будут "странными" в контексте контрастного корпуса. Сама характеристика вычисляется так: $\text{Weirdness} = \dfrac{w_s}{t_s}\left(\dfrac{w_g}{t_g}\right)^{-1}$, где w_s — частотность слова в исследуемой коллекции; t_s — число словоупотреблений во всей исследуемой коллекции; w_g — частотность слова в контрастной коллекции; t_g — число словоупотреблений в контрастной коллекции. Вместо частотности можно использовать количество документов, в котором встретилось слово.

TFIDF. Вес TFIDF является известным в информационном поиске методом взвешивания слов [10]. Вес TFIDF для леммы в некотором тексте представляет собой произведение двух множителей: TF, который характеризует частотность употребления этой леммы в данном тексте, и IDF, который характеризует встречаемость данной леммы в документах текстовой коллекции, причем чем больше такая встречаемость, тем меньше множитель IDF [10]. Множители TF и IDF могут задаваться различными формулами.

В данной работе использовался вариант подсчета TFIDF, который предложен в [11]:

$$\text{TFIDF} = \beta + (1-\beta)\cdot \text{tf}\cdot \text{idf}, \quad \text{tf}_D(l) = \frac{\text{freq}_D(l)}{\text{freq}_D(l) + 0.5 + 1.5\,\dfrac{dl_D}{avg\_dl}}, \quad \text{idf}(l) = \frac{\log\left(\dfrac{|c|+0.5}{df(l)}\right)}{\log\bigl(|c|+1\bigr)}. \qquad (1)$$

Здесь $\text{freq}(l)$ — частота леммы l в отзыве о фильме; $dl(l)$ — мера длины документа (в нашем случае количество лемм в отзыве); avg_dl — средняя длина отзыва; $df(l)$ — количество документов в коллекции (мнений, описаний фильмов или новостей), в которой встречалась лемма l; $\beta = 0.4$; $|c|$ — количество документов в коллекции.

Отклонение от средней оценки. Как уже упоминалось выше, для каждого собранного текста мнения сохранялась еще и числовая оценка (от одного до десяти), поставленная пользователем. Суть данной

характеристики в том, что положительные (отрицательные) слова более часто встречаются в положительных (отрицательных) отзывах и будут иметь большое отклонение от среднего значения. Для вычисления этой характеристики для каждого слова подсчитывается его средняя оценка (т.е. берутся оценки тех мнений, где оно встретилось, и делятся на их количество) и вычисляется модуль разности со средней оценкой всех отзывов корпуса: $\mathrm{Dev}(l) = \left|\frac{1}{k}\sum_{i=1}^{n} m_i k_i - \frac{1}{n}\sum_{i=1}^{n} m_i\right|$, $\sum_{i=1}^{n} k_i = k$, где l — рассматриваемая лемма, n — общее количество отзывов, m_i — оценка i-го отзыва, k_i — число словоупотреблений леммы в i-м отзыве (если не употребляется, тогда 0).

Частотность слов, употребляемых с заглавной буквы. Суть этой характеристики в том, что имена собственные обычно не являются оценочными словами. Поэтому мы подсчитываем, сколько раз каждое слово употреблялось с заглавной буквы и при этом не находилось в начале текста или в начале предложения.

3.2. Комбинации характеристик и корпусов. Для экспериментов из корпуса мнений были взяты первые десять тысяч слов, упорядоченных по частотности, и вся дальнейшая работа проводилась с ними. Слова были разделены на прилагательные и неприлагательные. Смысл такого разделения состоит в том, что многие исследователи указывали, что большинство оценочных слов являются прилагательными, а потому оценка качества нашего подхода в этом случае представляет отдельный интерес. В число неприлагательных входят существительные, глаголы и наречия. Все характеристики считались отдельно по этим двум категориям. Таким образом, получаются такие комбинации характеристик и корпусов:

— TFIDF по парам корпусов: малый-новости, малый-описания, мнения-новости, мнения-описания;

— странность по парам корпусов: мнения-новости и мнения-описания по количеству документов; малый-описания и мнения-описания по частотности;

— отклонение от средней оценки;

— частота по корпусу мнений и малому корпусу;

— количество документов, в которых встречается слово в корпусе мнений;

— частотность слов, употребляемых с заглавной буквы в корпусе мнений.

Кроме того, отдельно для корпуса описаний были посчитаны следующие характеристики: частотность, количество документов, странность описания-новости по количеству документов и TFIDF по корпусам описания-новости. Таким образом, для каждой леммы получается 17 признаков.

3.3. Алгоритмы и оценка качества. Для оценки качества работы алгоритмов необходимо эталонное множество оценочных слов, поэтому было решено взять исходный десятитысячный список слов и вручную разметить в нем оценочные слова.

При разметке оценочных слов выяснилось, что во многих случаях нельзя сделать однозначный вывод о том, является ли слово оценочным, поскольку иногда слово является оценочным в другой области, но не является оценочным в рабочей области. Многие слова могли употребляться как в оценочном, так и в неоценочном смысле при обсуждении фильмов. Поэтому было принято правило, что слово размечалось как оценочное, если можно представить какой-либо оценочный контекст (с участием этого слова) по отношениям к фильмам или их атрибутам. Кроме того, разметка делалась обоими авторами работы.

В результате разметки получился список оценочных слов размером три тысячи двести слов (1262 прилагательных, 296 наречия, 857 существительных, 785 глаголов).

Имея для каждого слова набор характеристик, можно построить классификатор для автоматического разделения слов на оценочные и неоценочные. Для классификации использовалась свободно распространяемая система Rapid Miner. В данной работе использовались следующие алгоритмы:

1) метод k ближайших соседей (kNN);

2) “наивный” байесовский классификатор (Naive Bayes);

3) персептрон (Perceptron);

4) нейронная сеть (двух- и трехслойная);

5) логистическая регрессия (Logistic Regression);

6) метод опорных векторов (SVM стандартный и с радиальной ядровой функцией).

Лучшие результаты по F-мере показали алгоритмы логистической регрессии для прилагательных (68.1%) и нейронных сетей для неприлагательных (50.9%, несбалансированные данные). Используя эти алгоритмы, мы получили списки прилагательных и неприлагательных, упорядоченных по вероятности их принадлежности к оценочным словам. В качестве примера приведем первые десять слов из двух списков.

Прилагательные: *позитивный, отличный, интересный, замечательный, затянутый, смешной, добрый, обалденный, предсказуемый, потрясающий.*

Неприлагательные: *пересматривать, простой, тягомотина, высосанный, хавать, плоско, наигран-*

но, фигня, блин, отвратительно.

В исследовании качества классификации отзывов на три класса мы применяли как списки наиболее вероятных оценочных слов, так и вес оценочности лемм, который был получен в результате применения методов машинного обучения.

4. Пространство признаков для классификации отзывов. В данном разделе будут описаны дополнительные признаки, которые могут улучшить качество классификации отзывов о фильмах на три класса: веса слов, слова-операторы, учет длины и структуры отзыва, знаки пунктуации.

4.1. Веса слов. В качестве основных элементов признакового пространства брался набор лемм (слов в нормальной форме), упоминавшихся в отзывах. Веса слов рассматривались бинарные (встречалось ли слово или нет) и вида TFIDF. Мы использовали два варианта вычисления TFIDF.

Во-первых, использовался наиболее простой вид TFIDF [10]:

$$\mathrm{TF} = \frac{n_i}{\sum\limits_k n_k}, \quad \mathrm{IDF} = \log \frac{|D|}{\left|d_i \supset t_i\right|}. \tag{2}$$

Здесь n_i — число вхождений слова в документ; в знаменателе TF — общее число слов в данном документе; $|D|$ — количество документов в корпусе; $|d_i \supset t_i|$ — количество документов, в которых встречается t_i (когда $n_i \neq 0$).

Во-вторых, использовалась формула (1) для TFIDF.

4.2. Слова-операторы. Интуитивно понятным является тот факт, что существуют слова, которые могут влиять на оценочность других слов, — слова-операторы. Для нахождения таких слов был использован набор из 3200 оценочных слов, размеченных вручную (см. п. 3.3). Из корпуса мнений (см. раздел 2) были выделены все слова, идущие непосредственно перед словами из вышеупомянутого списка оценочных слов (без каких-либо слов и знаков препинания между ними), и упорядочены по частоте встречаемости.

Из первой по частотности тысячи таких слов были вручную извлечены кандидаты на роль слов-операторов (74 слова). Чтобы оценить, насколько значимо влияние этих слов-операторов на изменение оценочной направленности слов, была выполнена следующая процедура: на основе корпуса мнений были вычислены средние оценки оценочных слов в тех случаях, когда они следуют за предполагаемым словом-оператором и когда предполагаемое слово-оператор перед ними не стоит. Средняя оценка в данном случае — это усредненная оценка, поставленная авторами отзывов, в которых встречался тот или иной вариант вхождения слова.

На основе сравнения такого рода средних оценок было выделено два значимых класса слов-операторов.

Если оценочные слова с высокой оценкой (> 8) при употреблении после слова-оператора меняли свою среднюю оценку на более низкую, а оценочные слова с низкой оценкой (< 6.7) меняли свою среднюю оценку на более высокую, то это означало, что слово-оператор меняет оценочность слова (оператор –).

Если после слова-оператора оценочные слова с высокой оценкой повышали свою среднюю оценку, а оценочные слова с низкой оценкой снижали свою среднюю оценку, то это означало, что слово-оператор усиливает оценочность слова (оператор +).

На рабочей коллекции частотно значимыми представителями каждого класса операторов оказались следующие слова:

а) оператор –: *не, нет*;

б) оператор +: *полный, очень, сильно, такой, просто, абсолютно, настолько, самый.*

Используя этот список и список слов из признакового пространства, мы получали все словосочетания операторов и признаковых слов. Эти словосочетания заменялись на мнемонические обозначения (“+” или “–”) в зависимости от оператора, например:

$$\text{НЕХОРОШИЙ} \rightarrow -\,\text{ХОРОШИЙ}; \quad \text{САМЫЙ КРАСИВЫЙ} \rightarrow +\,\text{КРАСИВЫЙ};$$
$$\text{НАСТОЛЬКО КРАСИВЫЙ} \rightarrow +\,\text{КРАСИВЫЙ}.$$

Полученные леммы с модификаторами добавлялись к множеству слов, используемому для составления признакового описания отзыва. Теперь, если в тексте отзыва встречается слово с оператором, то в описание попадает только слово с модификатором, а не оба слова. Это позволяет учесть влияние одних слов на другие.

4.3. Учет длины и структурных особенностей отзыва. Отзывы о фильмах могут быть длинными и короткими. Если отзыв длинный, то, зачастую, в начале или конце автор делает выводы. Это послужило основанием для рассмотрения отдельно коротких и длинных отзывов, а также разбиения отзывов на три части: начало, середину и конец. Порог по длине отзыва был выбран равным 50 словам.

Длинные отзывы делились на три части: начало (целые предложения, которые попадают в первые 25 слов отзыва), конец (целые предложения, попадающие в последние 25 слов) и середина (все, что осталось). К каждой части применялся алгоритм классификации, и полученные оценки агрегировались различными способами (голосование, усреднение).

4.4. Знаки препинания. Помимо лемм в пространство признаков для описания отзывов включались знаки препинания: “!”, “?”, “. . . ”.

5. Эксперименты. Поскольку отзывы во всех коллекциях имели оценки их авторов от 1 до 10 баллов, для отображения из десятибалльной шкалы в трехбалльную шкалу использовалась функция:

{1–6} → “1” (не понравился),

{7–8} → “2” (понравился, но есть недочеты или средне),

{9–10} → “3” (понравился).

Обоснование выбора именно такого отображения приводится в разделе 6.

Результирующее распределение отзывов по классам для каждой коллекции показано в табл. 1. Таким образом, количество отзывов класса “3” в главной коллекции составляет $\sim 45\%$ от общего числа.

Все отзывы были обработаны программой морфологического анализа, и получены списки лемм с информацией о частях речи.

Таблица 1

Распределение отзывов в коллекциях по трем классам оценок: “не понравился”, “средне”, “понравился”

Название множества	Не понравилось	Средне	Понравилось
MainSet	7189	8557	13 027
ControlSet	1013	795	545
ControlCutSet	304	365	545

Авторы предшествующих работ практически единогласно сходятся в том, что метод опорных векторов работает лучше для задач текстовой классификации (и классификации отзывов в частности). Мы также решили использовать этот метод. Ввиду большого количества данных и признаков была выбрана библиотека LIBLINEAR [12], которая обладает достаточной эффективностью для проведения наших экспериментов. Все параметры алгоритма были оставлены в соответствии со значениями по умолчанию.

На первом этапе мы проводили эксперименты, описанные ниже, только с коллекцией мнений MainSet. Для получения статистически значимых результатов использовалась кросс-валидация на пять частей. Затем для проверки результатов на новых данных и их переносимости алгоритмы обучались на большой коллекции и тестировались на двух других: ControlSet и ControlCutSet (см. п. 2.2).

В наших экспериментах мы использовали следующие множества слов для составления признакового описания отзывов.

1. Оптимальное множество оценочных слов, полученных по методу, описанному в разделе 3. Для его составления был запущен цикл с шагом в 100 слов по списку прилагательных и неприлагательных, упорядоченных по мере оценочности (вероятности в выдаче классификатора — opinweight). Эксперимент с использованием полученного множества обозначим OpinCycle.

2. Множество слов, на основе которого получился наилучший результат в работе [13] (OpinContrast). Это множество содержит 500 самых частотных слов с высоким оценочным весом (см. п. 3.3) и около 400 слов с самым высоким TFIDF, посчитанным по корпусам мнений и новостей (см. раздел 2).

3. Множество оценочных слов (всего 3200), полученных с помощью ручной разметки двух экспертов (см. п. 3.3) (OpinIdeal).

4. Множество всех слов, которые встречаются в корпусе 4 раза или больше (BoW). В это множество также вошли предлоги, союзы, частицы.

Из этих наборов слов выбирается тот, который позволяет получить наилучшие показатели классификации. На основе выбранного множества исследуется влияние других факторов: весов слов (tfidf), знаков препинания (punctuation), оценочных весов (opinweight), операторов (operators), длины отзывов (long и short).

Веса слов на основе TFIDF рассчитывались на основе двух формул: наиболее известной формулы (1) (tfidf simple) и формулы (2) (tfidf). Множитель IDF вычислялся не только по собственно корпусу отзывов (корпус мнений), но и на основе контрастных корпусов: корпусу новостей (tfidf news) и корпусу описаний фильмов (tfidf descr).

Для оценки качества классификации использовалась одна основная метрика — *правильность классификации* (accuracy). Она вычисляется как отношение правильно принятых системой решений к общему числу решений [14].

Результаты прогонов алгоритма с использованием различных наборов слов и факторов приведены в

табл. 2.

5.1. Обсуждение результатов. Поведение всех признаков при классификации оказалось похожим на каждой из коллекций: основной коллекции мнений MainSet, дополнительной коллекции ControlSet и сбалансированной дополнительной коллекции ControlCutSet (см. п. 2.2).

Следует заметить, что у разных наборов лемм не совпадают области покрытия, т.е. если в отзыве не встречается ни одного признака из набора, то он относится к классу “понравилось” в соответствии с распределением отзывов по классам в главной коллекции. Базовым весом каждого слова из набора считается его встречаемость в отзыве.

Таблица 2

Результаты классификации на основе различных признаков отзывов

Набор характеристик	Количество характеристик	Правильность классификации		
		MainSet с кросс-валидацией	ControlSet	ControlCutSet
OpinCycle	1000 adj + 1000 not adj	58.00	52.82	57.49
OpinContrast	884	60.33	55.50	58.64
OpinIdeal	3200	57.62	54.90	57.33
BoW	19 214	57.37	54.82	54.77
OpinCycle + tfidf simple	1000 adj + 1000 not adj	59.13	51.84	58.40
OpinContrast + tfidf simple	884	59.43	52.74	60.21
OpinIdeal + tfidf simple	3200	59.72	50.82	58.48
BoW + tfidf simple	19 214	62.52	57.50	61.86
BoW + tfidf	19 214	61.71	59.32	59.55
BoW + tfidf descr	19 214	61.74	59.07	59.80
BoW + tfidf news	19 214	62.90	60.26	61.44
BoW + tfidf news + operators	22 218	63.46	60.98	61.94
BoW + tfidf news + punctuation + + operators	22 221	63.17	61.11	61.69
BoW + tfidf news + opinweight + + operators	22 218	**64.48**	**62.04**	**64.00**
BoW + tfidf news+ opinweight + + operators + short	22 218	63.56	—	—
BoW + tfidf news + opinweight + + operators + long	22 218	62.37	—	—
BoW + tfidf news + opinweight + + operators + avg	22 218	63.14	—	—

Для увеличения веса оценочных слов по отношению к остальным словам мы использовали списки слов с оценочными весами от 0 до 1 (см. п. 3.3). Из этих списков были взяты 800 прилагательных и 200 неприлагательных с самыми высокими значениями вероятности (другие комбинации также проверялись). Значение веса opinweight для слов, не вошедших в указанную тысячу, полагалось равным нулю. Мы модифицировали вес каждого слова в векторных представлениях отзывов по следующей формуле:

$$\mathrm{WordWeight}(x) = \mathrm{TFIDF}(x)\exp\bigl\{\mathrm{opinweight}(x) - 0.5\bigr\}.$$

Результат, полученный с помощью формулы (BoW + tfidf simple), был принят за *basic line*. Наилучшие результаты были получены на основе набора всех слов вместе с весами TFIDF, оценочными весами и словами-операторами (BoW + tfidf news + opinweight + operators). Прирост, полученный наилучшим методом по отношению к *basic line 62.52*, является статистически значимым ($p < 0.001$, $\alpha = 0.05$, Wilcoxon signed-rank test/Two-tailed test).

Знаки пунктуации не дали существенного прироста, хотя их использование немного улучшает покрытие. Использование формулы (1) для TFIDF показывает практически такие же результаты, как и использование формулы (2). Выбор новостного корпуса в качестве контрастного показывает немного лучшие результаты, чем для корпуса описаний (BoW + tfidf descr) и корпуса мнений (BoW + tfidf).

Эксперименты, проводимые с длиной отзыва, показали, что качество классификации для коротких отзывов (BoW + tfidf news + operators + short) лучше, чем для длинных (BoW + tfidf news + operators + long). Хотя в среднем, с учетом количества отзывов в каждой части, результат не улучшается (BoW + tfidf news + operators + avg).

5.2. Оценка с нечеткими границами. Для метода с лучшим результатом классификации (BoW + tfidf news + opinweight + operators) мы провели еще дополнительную оценку с так называемыми мягкими границами. Оценка с мягкими границами заключается в следующем: если по базовой шкале автор отзыва поставил в отзыве граничную оценку (“8” или “6”), то не считается ошибкой отнесение алгоритмом отзыва к обоим из граничных классов. Такие ослабления условий сделаны из предположения о том, что даже человек очень плохо отличает граничные классы. Результаты классификации в этом случае для главной коллекции MainSet достигают 76.48% правильно классифицированных отзывов.

Таблица 3

Результаты проставления оценок людьми

Асессор	Правильность классификации		
	асессора по отношению к автору отзыва	с мягкими границами	лучшего алгоритма автоматической классификации по отношению к асессору
1	72.5	86.5	69.5
2	72.5	78.5	63.5
1 AND 2	71.5	—	—

6. Дополнительная оценка отзывов. Для понимания того, какого максимального качества классификации могут достичь автоматические алгоритмы, был проведен эксперимент по классификации отзывов на три класса людьми, которым не была известна исходная оценка отзыва, поставленная автором. О значимости такого рода верхних оценок качества автоматической классификации указывается, например, в работе [15]. Для эксперимента из исходного корпуса отзывов MainSet мы отобрали сто коротких отзывов (с длиной менее 50 слов) и сто длинных отзывов (с длиной более 50 слов). Отзывы извлекались так, чтобы сохранить исходное распределение по классам. Все явные указания на оценку, которую хотел поставить автор, были удалены.

Отобранные двести отзывов были даны на оценку двум асессорам (оценщикам). Результаты их оценки приводятся в табл. 3. Последняя строка таблицы указывает на соответствие в оценках между двумя асессорами.

Таким образом, мы видим, что по имеющимся текстам отзывов люди могут воспроизвести исходную оценку или оценку другого асессора всего лишь на уровне 71–72%, что является абсолютной верхней границей для улучшения качества автоматических алгоритмов. Отметим, что автоматическая классификация с мягкими границами, которая учитывает возможную неоднозначность пограничных между классами оценок, составляет 76.48%, что очень близко к качеству классификации с мягкими границами, получившемуся у одного из оценщиков (78.5%).

Таблица 4

Значения корреляции Спирмена для различных отображений

Асессор	Отображение		
	первое	второе	третье
1	0.761	0.753	0.759
2	0.725	0.702	0.669

Процент совпадения оценок лучшего алгоритма с оценками людей подтверждает показатели, полученные при кросс-валидации.

Кроме того, размеченные данные использовались для выяснения правильности отображения десятибалльной шкалы в трехбалльную. Для этого мы использовали различные отображения и вычисляли корреляцию полученных оценок и оценок асессоров. Основными кандидатами на отображения были:

— первое отображение {1-6} → “1”, {7-8} → “2” , {9-10} → “3”;
— второе отображение {1-5} → “1”, {6-8} → “2” , {9-10} → “3”;
— третье отображение {1-5} → “1”, {6-7} → “2” , {8-10} → “3”.

Для данных отображений значения корреляции Спирмена приводятся в табл. 4.

Из таблицы видно, что оценки, полученные с помощью первого отображения, лучше коррелируют с ручными оценками асессоров. Таким образом, наше исходное разбиение шкалы является обоснованным.

7. Заключение. В данной работе мы описали метод извлечения оценочных слов для конкретной предметной области на основе нескольких специальных коллекций. Кроме того, мы исследовали роль оценочных слов в задаче классификации отзывов о фильмах на три класса. Наиболее существенное влияние на качество классификации оказало использование информации об оценочности слова и весов, вычисленных с помощью TFIDF, а также учет слов-операторов. Все наборы признаков продемонстрировали

схожее поведение на различных коллекциях отзывов, что дает основание считать полученные результаты достоверными.

Мы оценили верхнюю границу качества классификации, которая оказалась весьма близка к результатам лучшего автоматического алгоритма, что затрудняет дальнейшее улучшение качества автоматической классификации отзывов на три класса.

СПИСОК ЛИТЕРАТУРЫ

1. *Pang B., Lee L.* Thumbs up? Sentiment classification using machine learning techniques // Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Philadelphia: ACL, 2002. 79–86.
2. *Whitelaw C., Garg N., Argamon S.* Using appraisal taxonomies for sentiment analysis // Proc. of CIKM-05, 14th ACM Int. Conf. on Information and Knowledge Management. Bremen: ACM, 2005. 625–631.
3. *Pang B., Lee L.* Opinion mining and sentiment analysis. Foundations and trends in information retrieval. Hanover, Massachusetts: Now Publishers, 2008.
4. *Pang B., Lee L.* Seeing stars: exploiting class relationships for sentiment categorization with respect of rating scales // Proc. of ACL, 43rd Meeting of the Association for Computational Linguistics. Ann Arbor: ACM, 2005. 115–124.
5. *Tsur O., Davidov D., Rappoport A.* ICWSM — a great catchy name: semi-supervised recognition of sarcastic sentences in online product reviews // Int. AAAI Conference on Weblogs and Social Media. Washington, DC: AAAI, 2010.
6. *Hu M., Liu B.* Mining and summarizing customer reviews // Proc. of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Seattle: ACM, 2004. 168–177.
7. *Esuli A., Sebastiani F.* Determining the semantic orientation of terms through gloss classification // Proc. of the 14th ACM Int. Conf. on Information and Knowledge Management (CIKM-05). Bremen: ACM, 2005. 617–624.
8. *Turney P.D.* Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews // Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia: ACM, 2002. 417–424.
9. *Hatzivassiloglou V., McKeown K.* Predicting the semantic orientation of adjectives // Proc. of the 35th Annual Meeting of ACL. Madrid: ACM, 1997. 174–181.
10. *Manning C., Raghavan P., Schütze H.* Introduction to information retrieval. Cambridge: Cambridge Univ. Press, 2008.
11. *Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В.* Экспериментальные алгоритмы поиска/классификации и сравнение с “basic line” // Российский семинар по оценке методов информационного поиска. Санкт-Петербург: НИИ Химии СПбГУ, 2004. 62–89.
12. *Fan R.-E. , Chang K.-W., Hsieh C.-J., Wang X.-R., Lin C.-J.* LIBLINEAR: a Library for Large Linear Classification // J. of Machine Learning Research. 2008. **9**. 1871–1874.
13. *Четверкин И.И., Лукашевич Н.В.* Автоматическая классификация отзывов на основе оценочных слов // 12-я Национальная конференция по искусственному интеллекту с международным участием (КИИ-2010). Москва: Физматлит, 2010. Т. 1. 299–307.
14. *Маннинг К., Рагхаван П., Шютце Х.* Введение в информационный поиск. М.: Вильямс, 2011.
15. *Kilgarriff A., Rosenzweig J.* Framework and results for English Senseval // Computers and Humanities, Special Issue on SENSEVAL. 2000. **34**, N 1, 2. 15–48.

Поступила в редакцию
04.10.2011