Московский государственный университет имени М.В.Ломоносова

На правах рукописи

**Четвёркин Илья Игоревич**

# Автоматизированное формирование базы знаний для задачи анализа мнений

Специальность 05.13.11 — математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей

Автореферат

диссертации на соискание учёной степени

кандидата физико-математических наук

Москва — 2013

Работа выполнена на кафедре алгоритмических языков факультета вычислительной математики и кибернетики Московского государственного университета имени М.В.Ломоносова.

| | |
|---|---|
| Научный руководитель: | доктор физико-математических наук, профессор,<br>зав. каф. алгоритмических языков<br>ВМиК МГУ имени Ломоносова,<br>Мальковский Михаил Георгиевич |
| Официальные оппоненты: | доктор технических наук, профессор,<br>директор НИИ «Прикладная семиотика»<br>Академии наук Республики Татарстан,<br>зав. каф. информационных систем КФУ,<br>Сулейманов Джавдет Шевкетович |
| | кандидат физико-математических наук,<br>доцент кафедры математических методов<br>прогнозирования ВМиК МГУ имени Ломоносова,<br>Майсурадзе Арчил Ивериевич |
| Ведущая организация: | Институт системного анализа РАН |

Защита состоится 1 ноября 2013 г. в 11 часов на заседании диссертационного совета Д.501.001.44 при Московском государственном университете имени М.В.Ломоносова, по адресу: 119991, ГСП-1, Москва, Ленинские горы, МГУ, 2-й учебный корпус, факультет вычислительной математики и кибернетики, аудитория 685. Желающие присутствовать на заседании диссертационного совета должны сообщить об этом за 2 дня до указанной даты по тел. (495) 939-30-10 (для оформления заявки на пропуск).

С диссертацией можно ознакомиться в Фундаментальной библиотеке МГУ имени М.В.Ломоносова. С текстом автореферата можно ознакомиться на официальном сайте ВМК МГУ имени М.В. Ломоносова http://www.cmc.msu.ru в разделе «Наука» - «Работа диссертационных советов» - «Д 501.001.44».

Автореферат разослан «__» сентября 2013 года.

Ученый секретарь
диссертационного совета Д 501.001.44
к.т.н., в.н.с. Костенко В.А.

## Общая характеристика работы

**Актуальность темы.** С развитием сети Интернет неуклонно растет объем хранимой неструктурированной информации, представленной текстами на естественных языках. Условно такую информацию можно разделить на два типа: «фактическая информация» и «субъективные мнения пользователей». За прошедшее десятилетие наиболее активному исследованию подвергались алгоритмы и программные системы связанные с обработкой фактической информации (например, поисковые машины).

С появлением Веб 2.0 и построенных на его основе социальных сетей, платформ для блогов и микроблогов, объём информации второго типа стал существенно увеличиваться. «Блогосфера» стала наиболее динамичной частью всемирной паутины, которая развивается, реагируя на события в реальной жизни. Вследствие этого растет научный и практический интерес к задачам обработки субъективной информации.

В рамках решения таких задач важными понятиями являются субъективность и тональность текста. Субъективность текста определяется наличием в нём точки зрения автора и выражением его мнения, а не констатацией фактов. Тональность — это эмоциональное отношение автора высказывания к некоторому объекту (объекту реального мира, событию, процессу, или их свойствам/атрибутам).

Для решения задачи анализа мнений необходимо обладать большим количеством вспомогательных знаний об объектах и их атрибутах, оценочных словах и выражениях, а также владельцах мнений – в виде специализированной базы знаний. Одним из наиболее важных видов знаний являются словари оценочных слов и проставленные оценки тональности для каждого слова. Однако построение универсального словаря оценочных слов является невыполнимой задачей, так как оценочность суждения во многих случаях зависит от предметной области.

Каждая предметная область может иметь свойственную только ей оценочную лексику, либо значения оценочных слов могут меняться в разных областях. Например, «нужно увидеть» является сильным оценочным выражением в предметной области о фильмах, но нейтральным в общественно-политической области. На практике это означает, что необходимо повторять трудоёмкую процедуру по составлению словарей оценочных слов для каждой предметной области, чтобы качество обработки субъективной информации оставалось на приемлемом уровне.

Таким образом, актуальной является проблема автоматического извлечения словарей оценочных слов из коллекций текстов для различных предметных

областей. Такие словари могут быть полезны при адаптации алгоритмов анализа мнений к конкретной области, а также для улучшения качества работы в различных задачах, например в поиске оценочной информации. Кроме того, извлечение оценочных слов непосредственно из текстовых коллекций позволяет найти сленг и другие несловарные слова, которые могут быть важными факторами при обработке мнений.

**Цель диссертационной работы.** Целью диссертационной работы является разработка методов и программных средств построения базы знаний для задачи анализа мнений. Разрабатываемые программные средства должны удовлетворять следующим требованиям: высокая точность извлеченных словарей оценочных слов; применимость к различным предметным областям; возможность комбинации словарей из различных предметных областей; определение тональности извлеченных оценочных слов.

Для достижения этой цели были решены следующие **задачи**:

1. Исследование и разработка модели извлечения оценочных слов для заданной предметной области и перенос обученной модели извлечения оценочных слов на другие предметные области без дополнительной разметки;
2. Разработка методов автоматического определения тональности извлеченных слов;
3. Построение обобщенного словаря оценочных слов, не зависящего от предметной области, на основе извлеченных знаний;
4. Разработка программного модуля извлечения оценочных слов для заданной предметной области и комбинации знаний из разных предметных областей.

**Основные положения, выносимые на защиту:**

1. Предложен новый метод автоматического извлечения оценочной лексики заданной предметной области, основанный на использовании нескольких текстовых коллекций и совокупности признаков. Показана переносимость предложенной модели извлечения оценочных слов на разные предметные области;
2. Предложен и реализован новый метод автоматического определения тональности извлеченных оценочных слов. Метод основан на применении марковских случайных полей;
3. На основе предложенного метода извлечения оценочной лексики был создан словарь оценочных слов, независящих от предметной области. Данный

словарь был опубликован и является первым общедоступным словарем оценочной лексики для русского языка. Показана полезность данного ресурса в нескольких задачах анализа тональности текстов;

**Научная новизна:** Новизна настоящей диссертационной работы заключается в том, что предложен новый метод построения базы знаний для задачи анализа мнений на основе ряда текстовых коллекций и комбинации статистических признаков. Применимость метода обоснована теоретически, на основе анализа полезности ряда признаков для классификации оценочных слов, а также численно, для чего были проведены эксперименты с использованием извлеченных знаний. Разработанный метод может применяться в различных предметных областях для извлечения знаний и построения автоматических алгоритмов анализа мнений на их основе.

**Практическая значимость.** На основе предложенного метода спроектирована и реализована программная система для извлечения оценочных слов в заданной предметной области. Разработанная система также имеет возможность комбинирования списков оценочных слов для формирования общих, независящих от предметной области словарей. Такой словарь был создан для метаобласти товаров и общедоступен для исследовательских целей[1]. Таким образом, разработанная система может быть использована для создания баз знаний оценочных выражений в различных предметных областях без какого-либо дополнительного участия человека.

Результаты научных исследований, представленных в диссертации, частично использовались в рамках гранта РФФИ № 11-07-00588-а под руководством Лукашевич Н.В.

**Апробация работы.** Основные результаты работы докладывались на следующих конференциях и семинарах:

- На международной конференции «Диалог» (2010г.);
- На международной конференции «Ломоносов» (2010г.);
- На 12-й национальной конференция по искусственному интеллекту с международным участием (2010г.)
- На международной конференции «Диалог» (2011г.);
- На семинаре по поиску концептов в неструктурированной информации (CDUD), проходящему совместно с конференцией RSFDGrC (2011г.);
- На семинаре по поиску информации и извлечению знаний (IEKA), проходящему совместно с конференцией RANLP (2011г.);

[1] http://www.cir.ru/SentiLexicon/ProductSentiRus.txt

- На международной конференции «Диалог» (2012г.);
- На летней школе по информационному поиску RUSSIR (2012г.);
- На международной конференции COLING (2012г.);
- На международной конференции «Ломоносов» (2013г.);
- На международной конференции «Диалог» (2013г.);

Кроме того, результаты обсуждались на семинаре лаборатории анализа информационных ресурсов НИВЦ МГУ, на семинаре в НИУ ВШЭ и на регулярном семинаре ACM SIGMOD в Москве.

**Личный вклад** заключается в выполнении основного объема теоретических и экспериментальных исследований, изложенных в диссертационной работе, включая разработку теоретических моделей, методик экспериментальных исследований, проведение исследований, анализ и оформление результатов в виде публикаций и научных докладов.

Результаты, связанные с извлечением оценочных слов, частично использовались в рамках гранта РФФИ № 11-07-00588-а под руководством Лукашевич Н.В.

**Публикации.** По теме диссертации опубликовано 18 работ, основные результаты изложены в 10 печатных работах, в том числе в 2 статьях в журналах из списка ВАК [1, 2], 1 статье, входящей в базу SCOPUS [3], и в 7 других изданиях [4–10].

**Объем и структура работы.** Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объем диссертации составляет 131 страницу с 8 рисунками и 17 таблицами, объем приложений – 15 страниц. Список литературы содержит 103 наименования.

## Содержание работы

Во **введении** обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана их практическая значимость, представлены выносимые на защиту научные положения.

**Первая глава** посвящена описанию основных задач, связанных с обработкой субъективных мнений пользователей, и в частности одной из наиболее важных и востребованных задач — классификации текстов по тональности. Также в данной главе проводится обзор основных методов классификации текстов по тональности на базе методов машинного обучения и инженерии знаний с

использованием словарей оценочных слов. Особое внимание уделяется задаче формирования словарей оценочных слов и проблемам переносимости алгоритмов на различные предметные области. Целью данной главы является анализ достоинств и недостатков существующих подходов к классификации текстов по тональности и обоснование важности учёта оценочных слов, характерных для заданной предметной области.

При решении задачи анализа мнений можно выделить несколько ключевых этапов анализа:

1. Определение субъективности/объективности текста;
2. Классификация заданного текста, предложения или словосочетания по тональности;
3. Выявление ключевых объектов, по которым высказано мнение, и построение автоматических аннотаций/рефератов отзывов;

Наиболее распространённой и востребованной на практике является вторая задача, поскольку существует большое количество порталов в Интернете, которые нацелены на сбор именно мнений и отзывов пользователей, заведомо содержащих какую-либо оценку.

Каждый субъективный фрагмент текста характеризуется объектом высказывания, субъектом высказывания и оценкой тональности. Под субъектом высказывания понимается автор текста (человек или организация), под объектом высказывания — некоторый объект реального мира, либо его части или свойства.

Определим тональность некоторого фрагмента текста по отношению к объекту, его частям или свойствам, как функцию от слов, входящих в данный фрагмент. Пусть фрагмент текста представлен последовательностью слов $w_1, ..., w_s$, тогда тональность данного фрагмента определяется функцией $F(w_1, ..., w_s) \rightarrow \{0, +1, -1\}$, где $0$ обозначает нейтральную тональность, $+1$ положительную, а $-1$ отрицательную. Можно рассматривать различные более сложные отображения, например с учётом смешанной тональности (частично положительной, частично отрицательной), или принимать во внимание силу высказываний (сильно положительно, слабо отрицательно).

Таким образом, при моделировании функции тональности, оценочные слова играют ключевую роль. На текущий момент существуют два основных подхода к построению решающей функции, приближающей целевую функцию $F$, задающую тональность фрагмента текста:

- Построение решающей функции на основе методов машинного обучения;

- Построение агрегирующей функции для оценочных слов и выражений, входящих в текст.

К достоинствам алгоритмов с использованием методов машинного обучения можно отнести высокое качество работы. Недостатками данного подхода являются: необходимость создания обучающей выборки, которая зачастую требует большого количества ручной разметки; падение качества классификации при переносе на другие предметные области; сложности при интерпретации результатов классификации. Результаты оптимизации весов признаков не всегда понятны человеку.

Достоинствами методов на основе словарей оценочных слов и правил являются: легкость интерпретирования результатов работы; хорошая переносимость на другие предметные области.

К недостаткам данного подхода можно отнести: необходимость в формировании словарей оценочных слов для заданной предметной области; более низкое качество классификации, чем при использовании методов машинного обучения.

Каждый из вышеперечисленных подходов может быть использован в зависимости от наличия размеченной коллекции или словарей оценочных слов в заданной предметной области, но как показывают результаты тестирования [10], наилучшие результаты получаются при комбинировании этих двух подходов.

Таким образом, для качественного решения задачи классификации текстов по тональности, актуальным является автоматическое формирование словарей оценочных слов для заданной предметной области. Такие словари могут быть использованы в системах, основанных на правилах, либо для формирования признакового описания объектов при использовании алгоритмов машинного обучения.

Во **второй главе** вводится формальная модель «мнения» пользователя и описываются основные типы знаний, необходимые для решения задачи анализа мнений. Одним из ключевых видов знаний являются оценочные слова и выражения, с помощью которых выражается отношение автора к объекту. Для автоматического построения словарей оценочных слов предлагается новый метод, основанный на использовании нескольких текстовых коллекций и набора статистических признаков. Для наиболее значимого признака «Странность» предлагается теоретическая модель, объясняющая зависимость качества признака от коллекций, с помощью которых он вычисляется. Словари оценочных слов, извлеченные с использованием разработанного алгоритма, применяются в задаче классификации отзывов по тональности.

Основной вид информации в рамках исследования — субъективные мнения пользователей о тех или иных объектах и их атрибутах. Мнение об атрибуте

*f* — это общая эмоция, суждение или оценка по поводу *f*, высказанная владельцем мнения. У каждого мнения может присутствовать тональность или эмоциональная окраска: положительная, отрицательная, смешанная или нейтральная. Наиболее простым случаем является положительная тональность, например «Я в диком восторге!!!!» или отрицательная тональность «Это какой-то ужас».

На основе введенных понятий можно определить формальную модель объекта и на основании этой модели — формальную модель мнения.

**Формальная модель объекта**: каждый объект *o* представляется в виде конечного набора *атрибутов* $F = \{f_1, ..., f_n\}$, который включает в себя и сам объект в виде специального атрибута. Каждый атрибут $f_i \in F$ может быть выражен с помощью одной фразы из конечного набора $W_i = \{w_{i1}, ..., w_{im}\}$, которые являются синонимами данного атрибута.

**Формальная модель мнения**: В общем виде некоторый документ $d$ содержит мнения о наборе объектов $o_1, ..., o_q$ от набора владельцев мнений $h_1, ..., h_q$. Мнение по каждому объекту $o_j$ выражено в отношении подмножества его атрибутов $F_j$. *Мнение* — это пятёрка $(o_j, f_{jk}, oo_{ijkl}, h_i, t_l)$, где $o_j$ это некоторый объект, $f_{jk}$ это атрибут объекта $o_j$, $oo_{ijkl}$ это тональность мнения по отношению к атрибуту $f_{jk}$, $h_i$ это владелец мнения, а $t_l$ это время в которое было высказано мнение. Для каждого атрибута $f_{jk}$ владелец мнения выбирает соответствующее слово или фразу из набора $W_{jk}$ и выражает мнение с помощью оценочных слов и выражений из некоторого набора $W_D^{OP}$, которые зависят от предметной области $D$ и, зачастую, от контекста. Здесь под оценочными словами и выражениями понимаются все слова и словосочетания, которые выражают оценку напрямую, либо неявно, с использованием скрытого смысла, иронии или сарказма.

Таким образом, можно выделить три основных типа знаний необходимых для эффективного решения задачи анализа мнений:

- Объекты и их атрибуты;
- Оценочные слова и выражения;
- Владельцы мнений

В рамках текущей работы основное внимание будет уделяться второму типу знаний — оценочным словам и выражениям.

Оценочные слова обладают рядом особенностей: концентрация оценочных слов в отзывах пользователей существенно выше, чем, например, в новостных текстах; каждое оценочное слово обладает некоторой тональностью и за счёт этого чаще употребляется в текстах с определённой эмоциональной окраской; оценочные слова чаще являются прилагательными или наречиями, реже

существительными или глаголами. На основании данных наблюдений предлагается и исследуется модель оценочных слов для их качественного извлечения.

Модель извлечения оценочных слов базируется на нескольких текстовых коллекциях: коллекции отзывов пользователей с оценками, коллекции описаний объектов и общей новостной коллекции. Чтобы построить модель, которая качественно отличает оценочные слова от неоценочных, для каждой леммы (слово в начальной форме) из корпуса мнений, вычисляется набор статистических и лингвистических признаков:

- **Частотные характеристики**: частота слова во всей коллекции и подокументная частота; частота слов с большой буквы; частота слов после операторов; признак «Странность»; признак TFIDF;
- **Характеристики на основе оценки пользователя**: отклонение от средней оценки; дисперсия оценки слова; вероятность встретить заданное слово с каждой из оценок;
- **Лингвистические признаки**: набор признаков учитывающих морфологию и неоднозначность словоупотреблений.

**Странность.** Для вычисления признака «Странность» необходимо два корпуса, один — с высокой концентрацией оценочных слов, другой — контрастный (с низкой концентрацией). Идея в том, что слова, которые несут оценки, будут «странными» в контексте контрастного корпуса. Сама характеристика вычисляется так:

$$Weirdness(w) = \frac{P_s(w)}{P_g(w)}$$

где $P_s(w)$ — вероятность появления слова в исследуемой коллекции; $P_g(w)$ — вероятность появления слова в контрастной коллекции.

**TFIDF** является известным в информационном поиске методом взвешивания слов. В данной работе использовался следующий вариант подсчета TFIDF:

$$TFIDF(w) = \beta + (1 - \beta) \cdot tf(w) \cdot idf(w) \tag{1}$$

$$tf(w) = \frac{f(w)}{f(w) + 2} \quad idf(w) = \frac{log\left(\frac{|c| + 0.5}{df(w)}\right)}{log(|c| + 1)}$$

где $f(w)$ — частота леммы $w$ в коллекции, $df(w)$ — количество документов в коллекции (описаний или новостей), где встречалась лемма $w$, $\beta = 0.4$, $|c|$ — количество документов в коллекции.

**Характеристики на основе оценок текстов.** Пусть $C = \{1...10\}$ будет множество возможных оценок авторов в коллекции отзывов. Тогда имеют место следующие определения.

Условная вероятность каждой категории в зависимости от слова:

$$P(c|w) = \frac{f(w,c)}{\sum_{c_i \in C} f(w,c_i)}$$

Условная вероятность для каждого слова в зависимости от категории:

$$P(w|c) = \frac{f(w,c)}{\sum_{w_i \in c} f(w_i,c)}$$

Условное математическое ожидание для каждой категории в зависимости от слова:

$$E(c|w) = \sum_{c_i \in C} c_i \cdot P(c_i|w)$$

Математическое ожидание каждой категории в коллекции отзывов:

$$E(c) = \sum_{c_i \in C} c_i \cdot P(c_i)$$

С использованием данных определений, вычисляется набор характеристик.

**Отклонение от средней оценки.**

$$Dev(w) = |E(c|w) - E(c)|$$

Данный признак позволяет выделять слова, которые употребляются в широком спектре оценочных категорий. Как следствие, вероятность принадлежности таких слов к оценочным ниже, чем у слов с более детерминированным поведением.

**Дисперсия оценки слова.** Еще одной важной характеристикой является дисперсия оценки слова. Если у оценки слова маленькая дисперсия, это значит, что данное слово употребляется в отзывах с близкими оценками. Такие слова более вероятно являются оценочными.

$$Var(w) = E(c^2|w) - E(c|w)^2$$

**Вероятность встретить заданное слово с каждой из категорий.** Чтобы формализовать информацию о встречаемости слов в различных категориях, вводится логарифм нормированной условной вероятности для каждого слова, в зависимости от категории.

$$Lhc(w) = \log \frac{P(w|c)}{P(w)}$$

Нормировка необходима для сравнения значений данной функции у различных слов.

**Лингвистические признаки.**

– Четыре бинарных признака частей речи (существительное, глагол, прилагательное и наречие);

– Два бинарных признака, отражающие неоднозначность употребления леммы в разных частях речи (т.е. лемма может быть разными частями речи, в зависимости от контекста) и нахождение данной леммы в словаре морфологического анализатора;

– Нахождение в слове заранее заданного списка приставок. Эта характеристика является важным индикатором слов, начинающихся с отрицания (например, *несмешной*);

На основе разработанного набора признаков предлагается формальное представление слов в коллекции текстов из некоторой предметной области. Для проверки качества определения оценочных слов на основе предложенной модели, проводятся эксперименты в предметной области о фильмах (объем коллекции 28773 отзыва с общедоступного интернет ресурса).

Для каждого слова в коллекции отзывов строится его признаковое представление и решается задача классификации на два класса: оценочные слова и неоценочные. Для настройки алгоритмов обучения с учителем необходимо сформировать размеченное множество слов. Чтобы его получить, все слова из коллекции отзывов о фильмах с частотой выше трёх (18362 слова) были размечены вручную. Слово считалось оценочным, если можно было представить какой-либо оценочный контекст с его участием в предметной области о фильмах. Каждое слово было размечено двумя экспертами. В результате данной процедуры было получено множество из 4079 оценочных слов.

Наилучший результат показала комбинация трёх различных алгоритмов классификации: Logistic Regression, LogitBoost и Random Forest. Для всех экспериментов применялась кросс-валидация на 10 частей. Выбор алгоритмов был обусловлен предварительными экспериментами и высоким результирующим качеством предложенных методов [9].

Данные алгоритмы применялись для формирования списков слов, упорядоченных по вероятности принадлежности каждого слова к классу оценочных. Для оценки качества извлеченных оценочных слов использовалась мера качества *Precision@n*. Эта мера хорошо подходит для оценки качества комбинаций списков, а также может быть использована с различными порогами. Для сравнения качества работы алгоритмов в различных предметных областях был выбран

| Фильмы | Книги | Игры | Цифровые камеры | Мобильные телефоны |
|---|---|---|---|---|
| 81.5% | 86.0% | 72.2 % | 65.8% | 73.2% |

Таблица 1: Результаты классификации в различных предметных областях

порог $n = 1000$. Этот порог не слишком велик для ручной разметки и достаточен для демонстрации качества работы модели. Результаты классификации в различных предметных областях можно найти в Таблице 1. Необходимо заметить, что модель обучалась в предметной области о фильмах и применялась без какой-либо дополнительной разметки в других областях.

Далее в работе исследуется вклад в качество классификации всех наиболее значимых групп признаков. В результате удаётся найти небольшую группу признаков (9 признаков), которая позволяет получить результат, сопоставимый по качеству с результатом исходного алгоритма. Стоит отметить, что все корпуса данных оказались задействованы в отобранном множестве признаков.

После проведения экспериментов по классификации слов и вычисления качественных оценок для каждого признака, было проведено теоретическое исследование признака «Странность», как одного из самых значимых признаков. В результате была установлена связь между параметрами распределения слов в коллекциях и мерой взаимной информации между признаком и меткой класса (оценочное или неоценочное) для каждого слова.

Данное исследование строилось на предположении, что распределение слов в коллекции текстов подчиняется закону Ципфа-Мандельброта. В этом случае можно найти аналитическое представление для плотности распределения признака «Странность» $z$:

$$p(z) = \begin{cases} \frac{(1-\alpha_g)(1-\alpha_s)}{2-\alpha_g-\alpha_s}\left(\frac{B_g}{B_s}\right)^{1-\alpha_s} z^{-\alpha_s}, z \leqslant B_s/B_g \\ \frac{(1-\alpha_g)(1-\alpha_s)}{2-\alpha_g-\alpha_s}\left(\frac{B_s}{B_g}\right)^{1-\alpha_g} z^{\alpha_g-2}, z > B_s/B_g \end{cases} \tag{2}$$

где $\alpha_s, B_s$ параметры специальной коллекции, а $\alpha_g, B_g$ параметры общей коллекции.

На основании распределения признака «Странность» можно вычислять взаимную информацию между данным признаком и классом слова. Взаимная информация может быть записана в терминах дивергенции Кульбака-Лейблера в следующем виде:

$$I(Z,T) = p(t=1)D_{KL}\Big(p(z \mid t=1)\|p(z)\Big) + p(t=0)D_{KL}\Big(p(z \mid t=0)\|p(z)\Big)$$

В работе рассматривается первый член суммы, так как обе дивергенции имеют схожий вид.

Было найдено аналитическое представление для дивергенции $D_{KL}\Big(p(z \mid t=1) \| p(z)\Big)$ в предположении, что распределение $p(z \mid t = 1), z \in \mathbb{R}^+$ подчиняется тем же законам, что и $p(z), z \in \mathbb{R}^+$ только с другими значениями параметров $\alpha_{s1}, B_{s1}$.

Если $B_s > B_{s1}$ тогда:

$$log\frac{1-\alpha_{s1}}{2-\alpha_{s1}-\alpha_g}\frac{2-\alpha_s-\alpha_g}{1-\alpha_s} - \frac{(\alpha_s-\alpha_{s1})(1-\alpha_g)}{(2-\alpha_{s1}-\alpha_g)(1-\alpha_{s1})} + \\ + \log\Big(\frac{B_s}{B_{s1}}\Big)^{1-\alpha_s} - \frac{(1-\alpha_{s1})(2-\alpha_g-\alpha_s)}{(1-\alpha_g)(2-\alpha_{s1}-\alpha_g)}\Big(1-\Big(\frac{B_{s1}}{B_s}\Big)^{1-\alpha_g}\Big)$$

.

Если $B_s < B_{s1}$ тогда:

$$log\frac{1-\alpha_{s1}}{2-\alpha_{s1}-\alpha_g}\frac{2-\alpha_s-\alpha_g}{1-\alpha_s} - \frac{(\alpha_s-\alpha_{s1})(1-\alpha_g)}{(2-\alpha_{s1}-\alpha_g)(1-\alpha_{s1})} + \\ + \log\Big(\frac{B_{s1}}{B_s}\Big)^{1-\alpha_g} - \frac{(1-\alpha_g)(2-\alpha_g-\alpha_s)}{(1-\alpha_{s1})(2-\alpha_{s1}-\alpha_g)}\Big(1-\Big(\frac{B_s}{B_{s1}}\Big)^{1-\alpha_{s1}}\Big)$$

.

В случае $B_{s1} = B_s$ выражение существенно упрощается:

$$log\frac{1-\alpha_{s1}}{2-\alpha_{s1}-\alpha_g}\frac{2-\alpha_s-\alpha_g}{1-\alpha_s} - \frac{\alpha_s-\alpha_{s1}}{1-\alpha_{s1}}\frac{1-\alpha_g}{2-\alpha_{s1}-\alpha_g}$$

и можно доказать следующую теорему:

**Теорема 1.** *Функция* $D_{KL}\Big(p(z \mid t=1) \| p(z)\Big)$ *является монотонно убывающей от параметра* $\alpha_{s1}$.

Доказанная теорема позволяет сравнивать между собой коллекции данных и давать оценку качества работы признака «Странность» с каждой из них.

В конце главы описывается исследование задачи классификации отзывов по тональности в различных предметных областях. Для формирования признакового представления текста используются различные наборы признаков, включающие в себя оценочные слова, слова-операторы, знаки препинания. В результате проведённых экспериментов можно сделать вывод, что извлеченные оценочные слова позволяют улучшить качество классификации текстов по тональности.

**Третья глава** посвящена исследованию методов формирования обобщенного списка оценочной лексики в области товаров. В результате исследования была предложена формула, поощряющая нахождение оценочного слова в начале наибольшего количества извлеченных списков оценочной лексики в разных предметных областях. Чтобы оценить полезность построенного ресурса, в работе приводятся результаты широкомасштабных экспериментов с его использованием.

Для поиска наилучшего способа комбинирования весов слов в различных областях были проведены эксперименты с тремя вариантами вычисления итогового веса, которые основаны на наиболее распространённых и простых функциях среднего или максимума:

- Взвешенное среднее по всем областям;
- Максимальный вес слова из всех областей, умноженный на долю областей в которых данное слово встречается;
- Произведение максимального веса из всех областей на среднюю позицию данного слова во всех областях.

По результатам исследования, наилучший результат был получен с помощью последнего варианта. Формула веса слова в результирующем списке:

$$R(w) = \max_{d \in D}(prob_d(w)) \cdot \sum_{d \in D} \frac{1}{|D|} \cdot \left(1 - \frac{pos_d(w)}{|d|}\right)$$

где $D$ — это множество, состоящее из пяти предметных областей, $d$ — это список оценочных слов в заданной предметной области, а $|d|$ это общее количество в слов в данном списке. Функции $prob_d(w)$ и $pos_d(w)$ возвращают значение вероятности и позицию заданного слова $w$ в списке $d$.

Сформированный список был оценен двумя экспертами. Качество результирующего ресурса составило 91.4% согласно мере Precision@1000. Согласие ответов экспертов составляло 84% ($k = 0.63$).

Для проверки полезности полученного обобщенного списка оценочных слов в мета-области товаров, он был протестирован в двух задачах:

- Задаче переноса системы анализа тональности с одной области на другую;
- Задаче извлечения мнений пользователей по запросу (эксперименты проводились в рамках Российского семинара по методом оценки информционного поиска РОМИП 2012).

В задаче переноса классификатора, алгоритмы на основе наборов признаков, состоящих из извлеченных оценочных слов для каждой предметной области или обобщенного списка оценочных слов, показывают существенный прирост в качестве классификации. Средний прирост качества при использовании обобщенного списка слов составил $1.76\%$, что доказывает полезность построенного ресурса для решения реальных задач.

В задаче поиска оценочных текстов по коллекции блогов с использованием построенного списка, были получены лучшие результаты по всем официальным мерам РОМИП 2012.

Таблицы, отражающие результаты экспериментов по каждой из задач, приведены в тексте третьей главы.

В **четвертой главе** описан новый метод определения тональности оценочных слов для заданной предметной области. Ключевые требования к данному методу были следующие:

- Использование оценочных слов, извлеченных методом, предложенным в главе 2;
- Независимость от предметной области и внешних лингвистических ресурсов;
- Использование только информации о словах из коллекции текстов, без дополнительной ручной разметки.

Исходя из данных требований, необходимо было составить признаковое описание, максимально использующее информацию, содержащуюся в коллекции. Основные источники информации о тональности оценочных слов были следующие:

- Средняя оценка слов в коллекции;
- Информация о совместной встречаемости оценочных слов;
- Учет отрицания, встречающегося рядом с оценочным словом.

В качестве алгоритма классификации предпочтительно было использовать алгоритмы обучения, где не требуется дополнительная ручная разметка, так как данная процедура требует существенных трудозатрат.

Наиболее естественным подходом к данной задаче является представление множества оценочных слов и связей между ними в виде структурированного набора, состоящего из нефиксированного заранее числа «элементарных» объектов, которые в свою очередь кодируются конечномерными векторами. То есть решается задача классификации взаимосвязанного массива данных.

Таким образом, для оценочных слов в заданной предметной области, строится неориентированный граф, где каждая вершина представляет случайную величину, обозначающую тональность некоторого слова из вышеупомянутого списка. Также высказывается предположение, что тональность оценочного слова зависит только от слов, встречающихся с ними в непосредственной близости. То есть тональность слова является локальной характеристикой, или формально, выполняется локальное марковское свойство для вершин в графе. Тогда построенный граф представляет собой марковскую сеть, в которой заданы

попарные и унарные потенциальные функции. Реализуемая модель схожа по построению с моделью Изинга, математической моделью статистической физики, предназначенной для описания намагничивания материала.

**Адаптация модели Изинга**. Пусть задан набор из $N$ слов, каждое из которых имеет свою тональность (равную +1 или -1). По аналогии с моделью Изинга и её энергетической функцией системы, в данном случае полная энергия может быть представлена в следующем виде:

$$E(x, S, H) = -\sum_{ij} s_{ij} x_i x_j - \sum_i h_i x_i$$

Где $x_i$ задаёт тональность слова $i$, $s_{ij}$ – попарный потенциал взаимодействия между двумя словами, $h_i$ унарный потенциал каждого слова (априорная тональность). Вероятность принять то или иное состояние для системы задается распределением Больцмана (частный случай распределения Гиббса):

$$P(x|S, H) = \frac{1}{Z(S, H)} \exp(-\beta \cdot E(x, S, H))$$

где коэффициент $\beta$ пропорционален обратной температуре в модели Изинга, а в данном случае является параметром модели. В традиционной модели Изинга от температуры зависит фазовое состояние системы:

- При высокой температуре частицы имеют случайные спины (парамагнетики)
- При низкой температуре большинство спинов направлены в одну сторону (ферромагнетики)
- Также известно, что при некоторой промежуточной температуре ферромагнетики становятся парамагнетиками. Этот процесс называется фазовым переходом.

Непосредственно перед фазовым переходом спины всех частиц локально согласованы. Именно это состояние сети представляет наибольший интерес, так как есть основания полагать, что тональности оценочных слов локально согласованы. Для поиска данного состояния были проведены эксперименты при различных значениях $\beta$.

Для задания унарных потенциалов были использованы оценки, поставленные авторами отзывов. Для каждого оценочного слова его унарный потенциал вычислялся по формуле:

$$h_i = E(c|w) - E(c)$$

| $\beta$ | **BP** | **MF** | **Gibbs** |
|---|---|---|---|
| 0.1 | 83.2% | 82.8% | 82.1% |
| 0.2 | 83.3% | 83.6% | 83.1% |
| 0.3 | 83.3% | 83.6% | 83.6% |
| 0.4 | 83.8% | 85.2% | 83.7% |
| 0.5 | 83.6% | 84.5% | 82.0% |
| 0.6 | 85.0% | 83.1% | 79.4% |
| 0.7 | 84.4% | 82.8% | 78.2% |
| 0.8 | 82.8% | 82.6% | 77.5% |
| 0.9 | 80.9% | 82.5% | 77.1% |
| 1.0 | 80.8% | 81.8% | 75.5% |

Таблица 2: Зависимость качества классификации от параметра $\beta$

Исходя из эмпирических данных, для всех попарных связей, вес вычислялся по следующей формуле:

$$s_{ij} = \frac{f(w_i, w_j)}{3} \cdot \Big(0.5 - min\Big(\frac{d(w_i, w_j)}{d(w_i, w_j) + 4}, 0.5\Big)\Big)$$

где $f(w_i, w_j)$ частота совместной встречаемости слов $w_i, w_j$, $d(w_i, w_j)$ — среднее расстояние между словами $w_i, w_j$.

Для поиска наиболее вероятного, согласованного состояния поля использовались три алгоритма:

- Алгоритм распространения доверия (**BP**);
- Метод самосогласованного поля (**MF**);
- Сэмплирование Гиббса (**Gibbs**).

Задача состояла в поиске состояния сети, при котором тональности всех слов являются локально согласованными. В связи с этим эксперименты были проведены при различных значениях $\beta$, результаты которых можно найти в таблице 2.

В результате работы алгоритмов было найдено согласованное состояние сети, в котором наилучшее качество классификации (мера качества — правильность) составило $85.0\%$ для алгоритма **BP**, $85.2\%$ для метода **MF** и $83.7\%$ для **Gibbs**. В качестве базового алгоритма (baseline) было взято отклонение от средней оценки, такое же, как и вес $h_i$ в графе. Если отклонение от средней оценки было больше 0, то слово считалось положительным, иначе отрицательным. Качество разделения слов на два класса с учетом данного правила составило $82.2\%$.

Как и описывалось в модели Изинга, при возрастании значения параметра, тональности слов начинают принимать случайные значения (фазовый переход). В процессе перехода от низкого значения к высокому удалось найти состояние, когда тональности слов локально согласованы и при этом качество классификации достигает наивысшего значения $85.2\%$.

В **заключении** приведены основные результаты работы, которые состоят в следующем:

1. Предложен новый метод автоматического извлечения оценочной лексики заданной предметной области, основанный на использовании нескольких текстовых коллекций и совокупности признаков. Показана переносимость предложенной модели на разные предметные области;

2. Предложен и реализован новый метод автоматического определения тональности извлеченных оценочных слов. Метод основан на применении марковских случайных полей;

3. На основе предложенного метода извлечения оценочной лексики создан словарь оценочных слов, независящих от предметной области, который опубликован и является первым общедоступным словарем оценочной лексики для русского языка. Показана полезность данного ресурса в нескольких задачах анализа тональности текстов.

## Основные публикации автора по теме диссертации


1. Лукашевич Н.В., Четверкин И.И. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. 2011. Т. 12. С. 73–81.

2. Лукашевич Н.В., Четверкин И.И. Построение модели для извлечения оценочной лексики в различных предметных областях // Моделирование и анализ информационных систем. 2013. Т. 20, № 2. С. 70–79.

3. Chetviorkin I., Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // COLING 2012: Technical Papers. 2012. P. 593–610.

4. Четвёркин И. И. Метод извлечения оценочных слов из текстов // Международная молодежная научная олимпиада «Ломоносов-2010». 2010. С. 99–100.

5. Chetviorkin I. Testing the sentiment classification approach in various domains — ROMIP 2011 // International Conference on Computational Linguistics Dialog. 2012. Vol. 2. P. 15–26.

6. Четвёркин И. И. Кластеризация оценочных слов по тональности на основе марковских случайных полей // Международная молодежная научная олимпиада «Ломоносов-2013». 2013. С. 62–63.

7. Четвёркин И. И. Кластеризация оценочных слов по тональности на основе марковских случайных полей // Новые информационные технологии в автоматизированных системах. 2013. С. 245–252.

8. Четвёркин И. И. Анализ и применение признаков оценочных слов для формирования словаря оценочной лексики // Сборник статей молодых ученых факультета ВМК МГУ. 2013. Т. 10. С. 279–295.

9. Четверкин И.И., Лукашевич Н.В. Автоматическое извлечение оценочных слов для конкретной предметной области // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». 2010. С. 565–571.

10. Chetviorkin I., Braslavskiy P., Loukachevich N. Sentiment Analysis Track at ROMIP 2011 // International Conference on Computational Linguistics Dialog. 2012. Vol. 2. P. 1–14.