На правах рукописи

Тутубалина Елена Викторовна

МЕТОДЫ ИЗВЛЕЧЕНИЯ И РЕЗЮМИРОВАНИЯ КРИТИЧЕСКИХ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ О ПРОДУКЦИИ

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

Автореферат

диссертации на соискание учёной степени

кандидата физико-математических наук

Казань — 2016

Работа выполнена в Высшей школе информационных технологий и информационных систем федерального государственного автономного учреждения высшего образования «Казанский (Приволжский) федеральный университет».

Научный руководитель: **Соловьев Валерий Дмитриевич,**
доктор физико-математических наук, ведущий научный сотрудник Научно-образовательного центра по лингвистике им. И. А. Бодуэна де Куртенэ ФГАОУ ВО «Казанский (Приволжский) федеральный университет»

Официальные оппоненты: **Ильин Вячеслав Анатольевич,**
доктор физико-математических наук, начальник Отдела информационных технологий и математического моделирования Курчатовского комплекса НБИКС-технологий НИЦ «Курчатовский институт»

Поляков Владимир Николаевич,
кандидат технических наук, доцент кафедры автоматизированных систем управления ФГАОУ ВО «Национальный исследовательский технологический университет «МИСиС»

Ведущая организация: Федеральное государственное бюджетное учреждение науки Санкт-Петербургское отделение Математического института им. В. А. Стеклова Российской академии наук

Защита состоится 16 июня 2016 г. в 17 часов на заседании диссертационного совета Д 002.087.01 при Институте системного программирования Российской академии наук по адресу: 109004, г. Москва, ул. А. Солженицына, дом 25.

С диссертацией можно ознакомиться в библиотеке и на сайте федерального государственного бюджетного учреждения науки Институт системного программирования Российской академии наук.

Автореферат разослан 6 мая 2016 года.

Ученый секретарь
диссертационного совета Д 002.087.01,
кандидат физико-математических наук — С. В. Зеленов

Общая характеристика работы

Актуальность темы. Диссертация посвящена разработке моделей и методов извлечения информации о высказываниях пользователей, содержащих указания на трудности в использовании продуктов (сервисов, товаров) и требующих устранения причин претензий от компаний. Рассмотрены наиболее распространенные задачи анализа мнений – классификация текстовых документов, извлечение высказываний относительно объектов мнений определенной предметной области, а также выделение объектов мнений по тематическим категориям.

В настоящее время одним из приоритетных направлений деятельности любой компании является улучшение качества продукции на основе изучения запросов пользователей в интернете: социальных сетях, блогах, сайтах интернет-сервисов[1]. Это связано, прежде всего, с развитием технологий, с широким распространением интернет-торговли и с возможностью пользователей сети обмениваться мнениями о товарах и услугах компаний. Пользователи публикуют свои мнения в открытом доступе на онлайн-ресурсах, позволяя компаниям и потенциальным покупателям продуктов учитывать информацию от потребителей. Неудовлетворенность продукцией может повлечь отрицательную рекламу для компании.

В последние десятилетия на рынке потребительских товаров появилась резкая динамика увеличения количества технически сложных товаров. Это связано, прежде всего, с развитием технологических инноваций, что приводит к постоянному увеличению конкретных видов компьютерных продуктов, и с концепцией соединения разной функциональности в едином устройстве. В связи с этим у покупателей возникают претензии по поводу удобства использования продукта наряду с ненадлежащим техническим качеством.

Анализ текстовых документов и отзывов пользователей с помощью методов машинного обучения и лингвистического анализа исследовались в трудах российских и зарубежных учёных, таких как Лиу Б., Тёрни П., Лукашевич Н. В., Вибе Дж., Блай Д., Джордан М., Воронцов К. В., Насукава Т., Дэйв К., Карди К., Эстер М., Гупта Н., Котов А. и других авторов. В

[1]Browning V., So K. K. F., Sparks B. The influence of online reviews on consumers' attributions of service quality and control for service standards in hotels // Journal of Travel & Tourism Marketing. — 2013. — Т. 30, 1-2. — С. 23—40

работе Хана и др.[2] приводится обзор исследований, связанных с феноменом "ошибка не найдена" как с классом проблем с продуктами, которые не могут быть легко диагностированы и воспроизведены в режиме тестирования. Перечисленными авторами разработаны основные теоретические аспекты анализа текстов на естественном языке с целью идентификации затруднений в использовании продуктов. Задача анализа мнений как задача анализа тональности текстов является общепринятой и достаточно хорошо изучена. Работа Бинга Лиу[3] дает развернутый обзор многих существующих автоматических методов классификации текстов, извлечения составных компонент продуктов с последующей категоризацией слов по тематикам. Однако, несмотря на это, в настоящее время задача автоматического извлечения высказываний, связанных с неисправностями и нарушением функциональности продуктов, выполняется, как правило, лишь с помощью лингвистических правил на основе ключевых слов, названных в данной работе *проблемными индикаторами*, базовых тематических моделей и методов машинного обучения на небольшом наборе признаков.

Таким образом, задача анализа высказываний, связанных с неисправностями и нарушением функциональности продуктов, на основании отзывов пользователей является актуальной и необходимой прикладной задачей.

Целью диссертационной работы является разработка методов и программных средств извлечения высказываний, составных компонент и функций продуктов, связанных с проблемными ситуациями и учитывающих особенности неструктурированных текстов пользователей в коллекции отзывов предметной области. Разрабатываемые методы и программные средства должны удовлетворять следующим требованиям:

- Более высокое по сравнению с существующими моделями качество предложенных методов;
- Переносимость методов на тексты различных языков; в данной диссертационной работе рассматриваются тексты пользователей на русском и английском языках;

[2]No Fault Found events in maintenance engineering Part 1: Current trends, implications and organizational practices / S. Khan [и др.] // Reliability Engineering & System Safety. — 2014. — Т. 123. — С. 183—195

[3]Liu B. Sentiment analysis and opinion mining // Synthesis lectures on human language technologies. — 2012. — Т. 5, № 1. — С. 1—167

- Переносимость методов на тексты отзывов о широкой группе товаров различной длины.

  В данной работе рассматриваются тексты пользователей (короткие тексты, отзывы) о продуктах из пяти предметных областей.

Объект и предмет исследования. Объектом исследования являются мнения пользователей о продуктах и сервисах компаний, представленные в виде неструктурированных текстов на естественном языке и доступные через Интернет. Мнения пользователей представлены в виде отзывов $D = \{d_1, d_2, \ldots, d_n\}$. В данной диссертационной работе для разработки более робастных методов автоматического извлечения информации используется синтаксическая сегментация отзывов на предложения: предложение s_{ij} отзыва $d_i = \{s_{i1}, \ldots, s_{i|d_i|}\}$ рассматривается как единичный элемент отзыва, поскольку данный элемент обладает определенным семантическим значением. Предметом исследования выступают задачи извлечения информации о высказываниях пользователей, содержащих указания на трудности в использовании продуктов, невозможность использования вследствие ошибок или недостатков продукта.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Провести классификацию отзывов пользователей о различных видах проблем с продуктами;
2. Создать словари проблемных индикаторов и оценочных слов;
3. Разработать следующие методы классификации: метод, основанный на правилах и словарях; метод, основанный на грамматической структуре сложных предложений относительно союзов;
4. Разработать на основе общедоступного тезауруса метод извлечения проблемных фраз по отношению к объектам (далее целевые объекты), связанным с предметной областью и относительно которых высказываются проблемные фразы;
5. Разработать методы резюмирования мнений для выделения тематически сгруппированных объектов мнений, указывающих на проблемные ситуации в использовании продуктов;
6. Реализовать предложенные методы в виде программного средства и провести экспериментальные исследования с целью определения

качества работы методов и моделей с использованием созданных коллекций текстовых документов.

Методы исследований. В данной диссертационной работе применялись методы обработки естественного языка, основанные на правилах, словарях и существующих лингвистических ресурсах, и вероятностные тематические модели, основанные на комплексе методов машинного обучения.

Основные положения, выносимые на защиту:

1. Предложен и реализован метод классификации предложений, основанный на знаниях в виде созданных словарей и правилах, учитывающих грамматическую структуру сложных предложений относительно союзов.
2. Предложен и реализован метод классификации предложений отзывов пользователей по отношению к целевым объектам, связанным с предметной областью, на основе синтаксических связей слов и мер семантической связанности.
3. Предложены и реализованы две вероятностные модели для задачи выделения тематически сгруппированных объектов мнений, учитывающие скрытые переменные для описания тем и проблемных индикаторов совместно.
4. Разработано программное обеспечение и проведено экспериментальное исследование, подтверждающее улучшение качества предложенных методов по сравнению с существующими алгоритмами.

Достоверность результатов диссертационной работы подтверждается взаимосвязью данных экспериментов и научных выводов, сделанных в работе, результатами апробации алгоритмов и разработанного программного прототипа систем. Результаты экспериментальных исследований согласуются с результатами классификаций отзывов в задачах анализа мнений.

Теоретическая и практическая значимость. Разработаны методы и модели извлечения информации о высказываниях пользователей о неполадках с продуктами, основанные на анализе структуры текстовых фрагментов мнений как связного текста. Предложенные методы извлечения высказываний из коллекции отзывов предметной области могут быть использованы при решении прикладных задач анализа мнений: классификации текстовых

документов, извлечения информации, кластеризации информации на основе тематических моделей.

Научная новизна. Задачи извлечения информации о высказываниях пользователей, указывающих на проблемные ситуации с продуктами, являются недостаточно изученными в литературе. В настоящей работе предложены новые методы извлечения высказываний в задачах анализа мнений пользователей различных предметных областей, основанные на алгоритмах машинного обучения без учителя, словарях и использовании структурной информации лингвистического тезауруса.

Улучшение качества разработанных методов по сравнению с существующими методами подтверждено экспериментально с помощью стандартных метрик качества систем анализа текстов на естественном языке. Экспериментально показано, что разработанные методы применимы к широкому классу продуктов различных областей коммерческой деятельности.

Апробация работы. Основные результаты работы докладывались на следующих конференций и научных семинарах:

1. Летней школе по информационному поиску RuSSIR (Казань, 16–20 сентября 2013 г.);
2. Международной конференции по анализу изображений, социальных сетей и текстов АИСТ (Екатеринбург, 10–12 апреля 2014 г.);
3. Семинаре по интеллектуальному обнаружению информации предметной области из текстов AHA!-Workshop на конференции "International Conference on Computational Linguistics" (Дублин, 23–29 августа 2014 г.);
4. Европейской конференции "European Conference on Information Retrieval" (Вена, 29 марта – 2 апреля 2015 г.);
5. Международной конференции "International Conference on Text, Speech and Dialogue" (Пльзень, 14–17 сентября 2015 г.);
6. Мексиканской конференции "Mexican International Conference on Artificial Intelligence" (Куэрнавака, 25–31 октября 2015 г.);
7. Международной конференции "International Conference on Web Search and Data Mining" (Сан-Франциско, 22–25 февраля 2016 г.).

Кроме того, результаты обсуждались на республиканском научном семинаре АН РТ "Методы моделирования" (05.05.2015) и на регулярном семинаре

кафедры интеллектуальных технологий поиска Высшей школы информационных технологий и информационных систем ФГАОУ ВО КФУ.

Публикации. Основные результаты по теме диссертации изложены в 10 печатных работах, 2 из которых опубликованы в журналах, рекомендованных ВАК [1; 2]; 6 работ опубликованы в журналах, входящих в базу SCOPUS [3—8]; 2 — в тезисах докладов [9; 10].

Личный вклад. Автором проведено исследование предметной области, выполнены теоретические и экспериментальные исследования, изложенные в диссертационной работе, разработана программная система на основе созданных методов. В работе [3] Иванову В.В. принадлежит постановка задачи и привлечение разметчиков для получения экспертных оценок контрольной выборки. В работах [7; 8] группа соавторов участвовала в обсуждении результатов и тестировании классификаторов на различных наборах признаков. В работе [6] Николенко С.И. предложил формулу для расчёта гиперпараметров.

Объём и структура диссертации. Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации составляет 145 страниц с 7 рисунками и 36 таблицами. Список литературы содержит 150 наименований.

Содержание работы

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

Первая глава посвящена обзору основных методов и подходов, применяемых в задачах анализа мнений пользователей. Целью данный главы является анализ эффективности существующих методов автоматического извлечения информации из мнений пользователей. В большинстве работ по анализу мнений выделяют несколько постановок основных задач:

– для каждого отзыва идентифицировать класс отзыва в целом, его предложений или отдельных фрагментов, где в большинстве работ под классом отзыва понимается тональность теста (положительная,

отрицательная или нейтральная) или определенный тип информации (например: факт, сарказм, спам, ирония, указание на дефект);
- для каждого аспектного термина, относительно которого высказывание были сделано, определить класс высказывания;
- идентифицировать тематические группы аспектных терминов продукта и их классы высказываний в коллекции документов.

В данной диссертационной работе рассматриваются описанные выше постановки задач применительно к проблеме извлечения информации о высказываниях, содержащие указания на трудности в использовании продуктов и связанных с предметной областью аспектов. Постановки задач подразумевают следующую методологию оценки эффективности: для коллекции текстов предметной области создается контрольная выборка фрагментов текстов отзывов, классифицируемые на два или более классов. Результат сравнивается с помощью стандартных метрик качества: достоверности (англ. accuracy), точности (англ. precision), полноты (англ. recall) и F-меры (англ. F_1-measure).

На данный момент существуют три основные группы методов для автоматического извлечения информации из мнений: (i) методы, основанные на лингвистическом анализе, синтаксических правилах и шаблонах; (ii) машинное обучение с учителем (англ. supervised methods); (ii) машинное обучение без учителя (англ. unsupervised methods). К достоинствам первых методов относится лингвистическое обоснование методов. К недостаткам можно отнести необходимость создания словарей оценочных слов и правил. К достоинствам вторых методов относится комбинирование большого количества различных признаков с помощью машинного обучения для повышения качества решаемой задачи. К недостаткам можно отнести значительное ухудшение результатов классификации на новых текстах других предметных областей и процесс создания обучающей выборки, который трудозатратен по времени и требует качественной ручной разметки. В качестве достоинств методов третьей группы можно выделить то, что модели позволяют использовать коллекции неразмеченных документов, для нахождения скрытых переменных (тематической, тональной) с небольшим количеством изменений алгоритмов оценивания. К недостаткам можно отнести параметризацию моделей.

В настоящий момент многие исследования чаще всего сводятся к использованию методов машинного обучения, где требуется сформировать век-

тор признаков и создать обучающую выборку. Однако одной из ключевых задач, являющейся основой при разработке методов для анализа мнений в текстах, остается задача создания словарей оценочных слов. На данный момент многие работы показывают, что не существует универсального словаря, который подходит для каждой предметной области или тематической категории. Поэтому актуальными являются создание новых словарей, использование которых позволяет повысить качество моделей и разработка методов, не зависящих от предметной области и не требующих размеченных ресурсов.

Вторая глава посвящена задаче автоматического извлечения из текстов пользователей высказываний, указывающих на проблемные ситуации в использовании продуктов. Для этого изучена классификация фраз, используемых пользователями для выражения отношения к тем аспектам продуктов, которые не работают должным образом. Целью задачи классификации является определение класса высказываний контрольной выборки. В данной главе рассматривается классификация высказываний на два класса: класс высказываний, указывающий на проблемные ситуации с продуктом, и класс высказываний, не содержащий или отрицающий упоминания о неполадках с продуктами.

Множество продуктов (сервисов, товаров), выпускаемое компаниями на потребительском рынке, задается как $P = \{P_1, P_2, \ldots, P_m\}$. Для $P_i \in P$ задано множество отзывов пользователей $D = \{d_1, d_2, \ldots, d_n\}$, где $d_i = \{s_{i1}, \ldots, s_{i|di|}\}$ и s_{ij} является предложением отзыва. Под **высказыванием, указывающим на проблемную ситуацию с продуктом**, или проблемным высказыванием понимается текстовый отрывок в отзыве пользователя, содержащий явное указание на сложности в использовании тех или иных продуктов, невозможность использования продуктов вследствие ошибки (бага, дефекта). Формально, обозначим проблемным высказыванием конструкцию $pphrase_{ij} = (r(s_{ij}), s_{ij})$, где $r(s_{ij}) \in [0,1]$ обозначает численное значение принадлежности предложения s_{ij} к классу высказываний с неполадками. Целью задачи является определение класса высказывания для всех предложений документов контрольной выборки $s_{ij} \in d_i, j \in \{1, \ldots, |d_i|\}$, $i \in \{1, \ldots, |D|\}$.

Под **проблемным индикатором** понимается однословная или многословная конструкция, выражающая явное или косвенное указание на проблему с продуктом (например: *трудность*, *отказывается работать*).

Для достижения целей задачи предложен подход, основанный на знаниях. Данный подход использует ресурсы в виде словарей, составленных вручную, и правила, отражающие структуру фрагментов текста относительно проблемных ситуаций. В данной работе созданы следующие словари:

- **ProblemWord**, содержащий проблемные индикаторы для широкой области продуктов, не зависящие от определенной предметной области ($\sim$ 940 и 190 лексических единиц для русского и английского языка, соответственно);
- **NotProblemWord**, включающий слова, указывающие на корректную работу, положительную ситуацию или исправление недостатков (например: *наладить*, *удобно*, *комфортно*) (69 и 45 лекс. единиц).
- **PositiveWord** (1078 лекс. единиц) и **NegativeWord** (1476 лекс. единиц), содержащие позитивные и негативные слова для русского языка;
- **Action** с глаголами действий ($\sim$ 7800 единиц для русского языка);
- **AddWord** (30 и 15 лекс. единиц) для ситуаций, выходящих за рамки допустимого пользователем (например: *чересчур*, *излишне*);
- словарь отрицаний **Negation** (14 и 22 лекс. единиц);
- **ImperativePhrases** (26 и 6 лекс. единиц), содержащий глаголы в повелительной форме (например: *сделайте*, *почините*) и фразы, указывающие на запрос пользователя к изменению;
- предметно-ориентированные словари проблемных индикаторов **DomainPW** для машин ($\sim$ 30 лекс. единиц) и приложений ($\sim$ 90 лекс. единиц) на русском языке.

В рамках подхода предложены два метода классификации проблемных высказываний: (i) метод, основанный на условиях вхождения лексических единиц из словарей (обозн. dictionary-based approach, **DbA**); (ii) метод, учитывающий грамматическую структуру сложных предложений относительно союзов (обозн. clause-based approach, **CbA**).

Метод **DbA** состоит из последовательной проверки условий вхождения лексических единиц из словарей Action, ProblemWord, NotProblemWord, ImperativePhrases с учетом отрицаний слов и вхождений словосочетаний, где первым словом является отрицание *нет*, *никакой*, *отсутствие*, *нету*, вто-

рым словом является существительное. Если вхождение найдено, то s_{ij} выделяется как проблемное, $r(s_{ij}) = 1$; в противном случае $r(s_{ij}) = 0$.

Формальное описание предложенного метода **CbA** представлено в виде контекстно-свободной грамматики – системы $G =< V, \sum, S, R >$, заданной следующими элементами: V – множество нетерминальных (вспомогательных) символов, $\sum$ - множество терминальных символов, $S \in V$ – начальный символ грамматики, R – множество правил вывода вида $A \to c$, где $A \in V, c \in (V \bigcup \sum)$. Правила вывода разделяются на несколько типов: (i) правила вывода нетерминальных символов, основанные на словарях; (ii) вспомогательные правила объединения слов и нетерминальных символов; (iii) правила классификации.

Множество терминальных символов определено как $\sum$ - алфавит системы. Множество нетерминальных символов определено как $V = \{Z, WD, X, S, PS, \neg PS, clause_1, clause_2, conj\}$, где S - предложение, PS – проблемное предложение, $\neg PS$ – предложение без проблемных ситуаций, WD – множество словосочетаний с отрицанием; X – множество слов с неизвестной информацией (не содержащиеся в N, P, IP, DP).

Опишем правила вывода нетерминальных символов $Z \to w_0^k$, $Z \in \{N, P, AW, A, IP, DP, NDP, DDP, VDP\}$, $w_0^k = w_0 \dots w_{k-1}, w_0^k \in \sum$, основанные на вхождении слов из словарей NegativeWord, PositiveWord, AddWord, Action (со связанным отрицанием), ProblemWord (без связанного отрицания) или NotProblemWord (со связанным отрицанием), вхождении явного индикатора, индикаторов с негативной тональностью, индикаторов действия ошибочной или некорректной ситуации, соответственно. Конструкция вида $\neg Z$ ($Z \in V$) обозначает отрицание Z (напр., $DP \to$ “проблема”, $\neg DP \to$ “без проблем”). Вспомогательные правила объединения слов предложений и нетерминальных символов представляют собой правила вида $I \to w_0^k$, где $w_0^k \in \sum$, $Z \to IZ_i$, $Z \to Z_i I$.

Метод содержит правила относительно слов *but*, *because*, *despite*, *но*, *а*, *хотя*, *пока*, *если*, *поэтому*, *теперь*, *правда* как наиболее значимых операторов, влияющих на семантическую связь между фрагментами текста согласно дискурсивному анализу[4]. Примеры правил классификации представлены в

[4]Wolf F., Gibson E. Representing discourse coherence: A corpus-based study // Computational Linguistics. — 2005. — Т. 31, № 2. — С. 249—287.

виде $S \to clause_1, conj\ clause_2$; $S \to conj\ clause_1, clause_2$, в которых $clause_1$, $clause_2$ обозначают фрагменты предложения, разделенные союзом $conj$:

1. $clause_1 \to P - IP - DP$, $conj \to$ но; $clause_2 \to A - DP$; $S \to PS$
2. $clause_1 \to IP|DP$, $conj \to$ хотя; $clause_2 \to \neg DP$; $S \to PS$
3. $clause_1 \to P$; $conj \to$ если; $clause_2 \to IP|DP$; $S \to \neg PS$
4. $clause_1 \to IP - DP$, $conj \to$ поэтому; $clause_2 \to A - PD - DP$; $S \to \neg PS$
5. $clause_1 \to P - DP$, $conj \to$ правда; $clause_2 \to N - \neg DP$; $S \to PS$
6. $clause_1 \to AW|DW$, $conj \to$ despite; $clause_2 \to P|\neg DP - DP - IP$; $S \to PS$
7. $clause_1 \to DW|NDP - DDP - VDP$, $conj \to$ but; $clause_2 \to P|\neg DP - DP - IP$; $S \to \neg PS$

Оператор "$-$" перед недетерминированном символом обозначает отсутствие данного символа во фрагменте $clause_1$ или $clause_2$. Оператор "$|$" между символами обозначает бинарную операцию *или*. Метод использует 28 правил для русского языка и 11 правил для английского языка. Алгоритм метода **CbA** состоит из нескольких шагов для предложения s_{ij}:

1. Применение правил вывода нетерминальных символов;
2. Применение объединения слов и нетерминальных символов;
3. Применение правил классификации;
4. Если s_{ij} было идентифицировано как $\neg PS$, то s_{ij} не содержит упоминаний проблем. Если предложение s_{ij} идентифицировано как PS, то алгоритм выделяет s_{ij} как проблемное. В противном случае s_{ij} классифицируется согласно результатам метода **DbA**.

В качестве контрольной выборки были собраны и размечены отзывы о продуктах из пяти предметных областей: 5,688 на русском языке (из них 2,187 относятся к классу выражений с проблемными ситуациями) и 4,378 предложений на английском языке (из них 2,822 относятся к классу выражений с проблемными ситуациями).

Для определения лучшего метода автоматического извлечения высказываний о проблемных ситуациях предложенные методы сравнивались с базовыми методами машинного обучения с учителем на основе модели "мешок слов" (англ. bag of words), обозначенными с "2-gr." и обученными на словах и словосочетаниях: метод максимальной энтропии (MaxEnt), метод

опорных векторов (SVM). Так же методы сравнивались с классификаторами NRC-Canada[5] (далее NRC), GU-MLT-LT[6] (далее GU) и KLUE[7], показавшие наилучшие результаты классификации тональности коротких сообщений на английском языке и использующие базовые методы, и классификатором NaiveBayes[8]. Результаты классификации представлены в Таблицах 1, 2, 3. В качестве основного критерия качества используется F-мера, полученная макроусреднением (далее макро F-мера), как единая метрика, объединяющая метрики полноты и точности для двух классов. Экспериментально было показано, что лучшее качество согласно значениям макро F-меры независимо от предметной области для английского языка и для текстов о машинах на русском языке показывает предложенный метод **CbA**. На коротких текстах о приложениях **CbA** показывает сравнимые результаты с разницей значений макро F-меры в .002 относительно GU и KLUE.

Таблица 1 — Результаты классификации относительно класса высказываний о проблемных ситуациях и результаты классификации, полученные макроусреднением

Метод	Машины (рус.)							Приложения (рус.)						
	Acc.	P_{pos}	R_{pos}	F_{pos}	макроуср.			Acc.	P_{pos}	R_{pos}	F_{pos}	макроуср.		
					P	R	F					P	R	F
MaxEnt 2-gr.	.704	.308	.451	.366	.581	.607	.586	.689	.745	.701	.723	.684	.687	.685
SVM 2-gr.	.817	.519	.434	.473	.695	.670	.680	.805	.834	.826	.830	.800	.801	.800
NaiveBayes	.754	.380	.470	.420	.624	.645	.634	.791	.809	.834	.821	.786	.783	.784
NRC	.840	.601	.474	.530	.742	.701	.720	.821	.830	.867	.848	.818	.812	.815
GU	.835	.701	.232	.349	.772	.604	.678	.829	.832	.877	.854	.827	.820	.824
KLUE	.849	**.730**	.330	.454	**.795**	.650	.715	.829	.830	**.884**	.856	.828	.819	.824
NRC+Dicts	.847	.621	.496	.552	.754	.712	.732	.831	.841	.874	.857	.829	.824	.826
GU+Dicts	.852	.694	.391	.501	.782	.675	.725	**.833**	**.843**	.874	**.858**	**.831**	**.826**	**.829**
KLUE+Dicts	**.853**	.715	.380	.496	.792	.672	.727	.832	**.843**	.870	.856	.829	.825	.827
DbA	.814	.507	.636	.564	.708	.746	.726	.806	.829	.837	.833	.802	.803	.802
CbA	.814	.508	.649	.571	.709	.751	.730	.820	.842	.846	.845	.816	.815	.816
CbA с DomainPW	.835	.550	**.721**	**.624**	.741	**.791**	**.765**	.827	.833	.880	.854	.825	.818	.822

[5]Mohammad S. M., Kiritchenko S., and Zhu X. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. Proceedings of SemEval 2013. — 2013. — С. 321–327.

[6]Günther T., Furrer L. GU-MLT-LT: Sentiment analysis of short messages using linguistic features and stochastic gradient descent // Proceedings of SemEval 2013. — 2013. — С. 328—332.

[7]KLUE: Simple and robust methods for polarity classification / Proisl T. [и др.] // Second Joint Conference on Lexical and Computational Semantics (* SEM). Т. 2. — 2013. — С. 395—401.

[8]Maalej W., Nabil H. Bug report, feature request, or simply praise? on automatically classifying app reviews // Requirements Engineering Conference (RE), 2015 IEEE 23rd International. — IEEE. 2015. — С. 116— 125.

Таблица 2 — Результаты классификации относительно класса высказываний о проблемных ситуациях и результаты классификации, полученные макроусреднением

Метод	Машины (анг.)							Электроника (анг.)						
	Acc.	P_{pos}	R_{pos}	F_{pos}	макроуср.			Acc.	P_{pos}	R_{pos}	F_{pos}	макроуср.		
					P	R	F					P	R	F
MaxEnt 2-gr.	.763	.851	.866	.858	.569	.564	.566	.615	.571	.564	.568	.610	.610	.610
SVM 2-gr.	.625	.843	.673	.748	.520	.532	.505	.715	.689	.665	.677	.712	.710	.711
NaiveBayes	.751	.868	.825	.846	.591	.608	.600	.701	.632	.796	.704	.710	.709	.710
NRC	**.831**	.847	.973	**.906**	**.689**	.559	.617	.767	.749	.718	.733	.764	.762	.763
GU	.813	.841	.957	.895	.605	.539	.570	.757	.742	.700	.720	.755	.751	.753
KLUE	.827	.833	**.990**	.905	.650	.515	.575	.760	.740	.713	.726	.757	.755	.756
NRC+Dicts	.817	.853	.941	.895	.642	.579	.609	.766	.745	.723	.730	.763	.762	.763
GU+Dicts	.821	.839	.970	.900	.622	.534	.575	**.779**	**.769**	.723	.745	**.777**	**.773**	**.775**
KLUE+Dicts	.821	.843	.963	.899	.634	.548	.588	.776	.759	.731	.745	.774	.772	.773
DbA	.738	.876	.795	.834	.596	.626	.611	.754	.693	.809	.746	.757	.760	.758
CbA	.751	**.885**	.803	.842	.614	**.650**	**.632**	.768	.710	**.814**	**.758**	.770	**.773**	.771

Таблица 3 — Результаты классификации относительно класса высказываний о проблемных ситуациях и результаты классификации, полученные макроусреднением

Метод	Инструменты (анг.)							Детские товары (анг.)						
	Acc.	P_{pos}	R_{pos}	F_{pos}	макроуср.			Acc.	P_{pos}	R_{pos}	F_{pos}	макроуср.		
					P	R	F					P	R	F
MaxEnt 2-gr.	.613	.707	.787	.745	.471	.477	.470	.556	.692	.534	.592	.538	.638	.531
SVM 2-gr.	.593	.753	.645	.695	.544	.552	.541	.635	.722	.591	.729	.590	**.726**	.590
NaiveBayes	.648	.768	.732	.749	.578	.583	.580	.670	**.758**	.736	.747	.635	.538	.582
NRC	.705	.748	.892	.813	.601	.561	.580	.705	**.758**	.815	.786	.653	.640	.646
GU	.698	.733	**.915**	**.814**	.567	.530	.548	.695	.729	**.859**	.789	.653	.617	.634
KLUE	.704	.745	.895	.813	.596	.556	.575	.698	.746	.826	.784	.657	.638	.648
NRC+Dicts	.711	.765	.863	.811	.592	.592	.606	**.708**	.756	.824	.789	**.670**	.652	**.661**
GU+Dicts	.701	.740	.900	.812	.585	.546	.565	.695	.733	.847	.786	.641	.622	.636
KLUE+Dicts	.695	.741	.885	.807	.578	.547	.563	.701	.752	.818	.784	.662	.646	.654
DbA	.702	.778	.818	.798	.622	.612	.617	.705	.742	.848	.791	.666	.636	.650
CbA	**.720**	**.790**	.831	.810	**.646**	**.633**	**.639**	**.708**	.744	.851	**.794**	**.670**	.640	.655

Для анализа вклада созданных словарей в классификаторы NRC, GU и KLUE, обозначенные с “Dicts”, был добавлен набор признаков, схожий с признаками NRC для тональных слов и подсчитанный на словарях ProblemWord, NotProblemWord, Action, ImperativePhrases, AddWord с учетом отрицаний слов. Экспериментально подтверждено, что использование признаков Dicts улучшает результаты NRC, GU, KLUE по макро F-мере в 15 из 18 экспериментах (в среднем на 1.1% макро F-меры). Предложенный метод **CbA** показывает лучшее качество относительно NRC+Dicts, GU+Dicts и KLUE+Dicts

в 12 из 18 экспериментах со средней разницей значений макро F-меры в .037 и сравнимые результаты со средней разницей значений макро F-меры в .005 на текстах о приложениях, электронике и детских товарах. Таким образом, результаты подтверждают улучшение качества классификации с помощью предложенного метода **CbA** и созданных вручную словарей.

Статистическая значимость результатов классификации текстов различных предметных областей с помощью метода **CbA** в попарном сравнении с методами машинного обучения была показана с помощью непараметрического статистического критерия знаковых рангов Вилкоксона[9]. Нулевая гипотеза заключается в том, что два метода классифицируют отзывы пользователей о различных продуктах одинаково. Для этого две и четыре предметные области были объединены в русскоязычный и англоязычный корпуса, соответственно. Результаты классификации были проверены для двух объединенных выборок. Результаты попарного сравнения **CbA** с MaxEnt, SVM ($p < 0.001$), NRC, GU-MLT-LT и KLUE ($p < 0.01$) подтверждают, что правила анализа сложных предложений вносят дополнительную информацию в процесс классификации высказываний. Разница сравнения результатов CbA и NRC+Dicts, GU+Dicts, KLUE+Dicts значима в меньшей степени ($p < 0.05$), что подтверждает вклад новых признаков, основанных на созданных словарях, для улучшения классификации с помощью существующих методов машинного обучения.

В **третьей главе** описывается новый алгоритм извлечения проблемных высказываний по отношению к предметно-ориентированным целевым объектам на основе общедоступного тезауруса. Для извлечения целевых объектов предложен метод, основанный на синтаксических связях между проблемными индикаторами и существительными в предложении. Для извлечения предметно-ориентированных целевых объектов изучается возможность применения мер семантической связанности.

Под **целевым объектом** понимается элемент отзыва, относительно которого высказывается некоторое мнение, представленный в виде однословной или многословной конструкции, характеризующей тему документа в определенной предметной области. Под предметно-ориентированными целевыми объектами понимаются связанные с продуктом понятия, существен-

[9]Demšar J. Statistical comparisons of classifiers over multiple data sets // The Journal of Machine Learning Research. — 2006. — Т. 7. — С. 1—30.

ные в определенной предметной области. Целью задачи является определение множества целевых объектов $T_i = \{t_1, t_2, \cdots, t_k\}$ для продуктов компании $P = \{P_1, P_2, \cdots, P_m\}$ и классов для всех предложений документов контрольной выборки $s_{i,j} \in d_i, j \in \{1, \cdots, |d_i|\}, i \in \{1, \cdots, |D|\}$. Актуальность поставленной задачи объясняется необходимостью извлечения информации о товарах, существенной для компании.

Для определения множества возможных целевых объектов использовались синтаксических связи между проблемными индикаторами (обозн. PW) и существительными (обозн. T) с помощью прямых и косвенных (через словопосредник S) зависимостей слов в предложении. Примеры 1 и 2 содержат косвенную и прямую зависимости между словами.

Пример 1. Вы когда *ошибку* (PW) с *размещением* (S) *вкладов* (T) исправите?

Пример 2. *Рамки* (T) *окон* (T) на дверях с внутренней стороны, незаметные *взгляду* (T) в обычных условиях, выглядят *ущербными* (PW)

Для идентификации предметно-ориентированных целевых объектов изучается возможность применения мер семантической связанности. Мера семантической связанности понятий предметной области представляет собой числовую оценку степени их смысловой связанности. Если семантическая связанность терминов целевого объектов t_k и терминов предметной области выше, чем семантическая связанность терминов понятия t_k и фоновых терминов, определяющих широкую группу товаров, то t_k является *предметно-ориентированным целевым объектом*. В данной работе рассматриваются меры нескольких типов: (i) мера WUP (анг. Wu & Palmer's measure), основанная на расстоянии в графе тезауруса; (ii) мера RES (англ. Resnik's measure), основанная на информационном содержании; (iii) мера LESK (англ. Lesk's measure), основанная на определениях понятий; (iv) косинусная мера COS (англ. cosine similarity), использующая вектора распределённых представлений слов.

Алгоритм 1 извлечения высказываний по отношению к предметно-ориентированным целевым объектам использует результаты анализа текстового высказывания предложенными методами **DbA** и **CbA**.

Algorithm 1: Алгоритм извлечения высказываний, указывающих на проблемные ситуации, по отношению к предметно-ориентированным целевым объектам

```
1 Function lookupForProblemsWithTargets(s, domain_terms,
  common_terms)
      Input: s – исходное предложение, domain_terms –
             предметно-ориентированные термины, common_terms -
             фоновые термины, определяющие широкую группу товаров
      Output: PWTs – множество пар (проблемный индикатор, объект)
2     PWTs ← ∅;
      /* поиск аннотаций из словарей в предложении        */;
3     PWs = lookupForPW(s);
      /* анализ предложения с помощью грамматики зависимостей */;
4     DRs = (getGrammStructure(s)).typedDependenciesCollapsed(true)
      foreach pw in PWs do
5         targets=lookupForRelatedTargets(pw, DRs);
6         foreach t_i in targets do
              /* подсчет семантической связанности между целевым
              объектом и domain_terms или common_terms       */;
7             if relScore(domain_terms, t_i) ≥ relScore(common_terms, t_i)
              then
8                 PWTs = PWTs ∪ {pair(pw, t_i)}
9     return PWTs;
```

Ниже приведено описание предложенного алгоритма извлечения высказываний, состоящего из ряда шагов:

1. Извлечь из s_{ij} вхождения индикаторов $\{pw_{i1}, pw_{i2}, \ldots, pw_{in}\}, n \leq |s_{ij}|$ из словарей Action, ProblemWord, NegativeWord, AddWord, ImperativePhrases в зависимости от связанных отрицаний, используя метод **DbA**;
2. Для каждого pw_{ij} определить множество возможных целевых объектов $\{t_1, t_2, \ldots, t_k\}$: если существует прямая или косвенная зависимость между t_k и pw_{ij} в высказывании s_{ij}, то целевой объект t_k синтаксически связан с w_{ij}; если множество объектов пусто, то w_{ij} исключается из множества индикаторов;
3. Для каждого t_k определить, является ли объект предметно-ориентированным на основе меры семантической связанности терминов понятия t_k и терминов предметной области;

4. Классифицировать s_{ij} как проблемное, если существует хотя бы одна комбинация (pw_{ij}, t_k) и s_{ij} не было идентифицировано как $\neg PS$ согласно результатам метода **CbA**; в противном случае $r(s_{ij}) = 0$.

Для оценки вклада предложенного метода извлечения целевых объектов, используя синтаксические связи в предложении, алгоритм извлечения проблемных высказываний по отношению к предметно-ориентированным целевым объектам (обозн. DD+ID) сравнивался с методом **CbA**. Результаты сравнения приведены в Таблицах 4, 5 и 6. Применение метода идентификации предметно-ориентированных целевых объектов на основе меры семантической связанности обозначено как DD+ID+CbA+*мера*. Экспериментальные оценки подтверждают улучшение качества согласно F-меры для текстов трех предметных областей (машины, инструменты, детские товары). Для текстов о мобильных приложениях и электронике наилучшие результаты показывает метод **CbA** без учета целевых объектов, что объясняется сложной технической архитектурой продуктов. Наилучшие значения точности для английского и русского языков показывают методы проверки семантической связанности с помощью мер LESK и COS.

Таблица 4 — Результаты классификации высказываний

Метод	Инструменты (англ.)				Детские товары (англ.)			
	Acc.	P	R	F	Acc.	P	R	F
CbA	.720	**.790**	.831	.810	**.708**	**.744**	.851	.794
DD+ID+CbA	**.721**	.780	**.868**	**.821**	**.708**	.732	**.895**	**.805**
DD+ID+CbA+LESK	.694	.769	.819	.793	.691	.725	.858	.786
DD+ID+CbA+WU	.685	.768	.805	.786	.688	.721	.859	.784
DD+ID+CbA+RES	.695	.765	.831	.796	.692	.723	.864	.787
DD+ID+CbA+COS	.689	.766	.810	.787	.695	.729	.854	.787

Таблица 5 — Результаты классификации высказываний

Метод	Машины (англ.)				Электроника (англ.)			
	Acc.	P	R	F	Acc.	P	R	F
CbA	.751	**.885**	.803	.842	**.768**	**.710**	.814	**.758**
DD+ID+CbA	**.776**	.876	**.850**	**.863**	.699	.623	**.824**	.710
DD+ID+CbA+LESK	.750	.871	.819	.844	.684	.614	.787	.690
DD+ID+CbA+WU	.704	.868	.758	.809	.680	.616	.753	.677
DD+ID+CbA+RES	.715	.870	.771	.817	.693	.622	.796	.698
DD+ID+CbA+COS	.739	.872	.804	.837	.683	.611	.799	.693

Таблица 6 — Результаты классификации высказываний

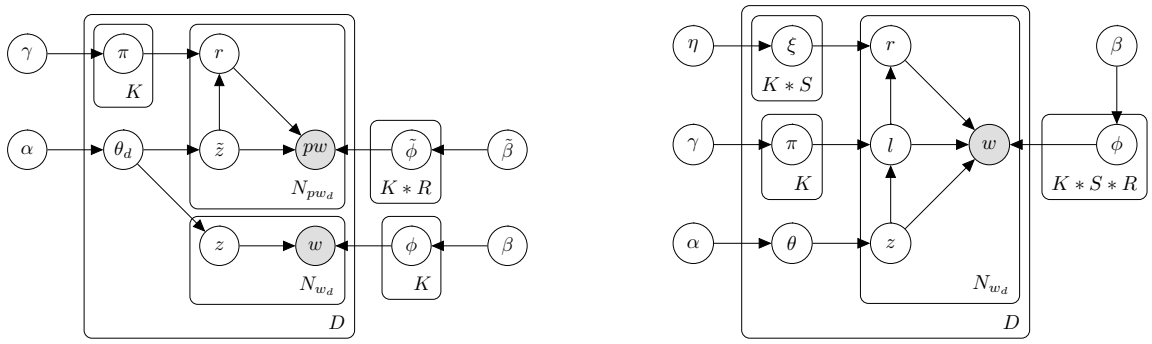
Метод	Машины (рус.)				Мобильные приложения			
	Acc.	P	R	F	Acc.	P	R	F
CbA	.814	.508	**.649**	.571	**.820**	**.842**	**.846**	**.845**
DD+ID+CbA	.829	.537	.640	**.584**	.789	.824	.806	.815
DD+ID+CbA+LESK	.827	.538	.622	.577	.785	.828	.791	.809
DD+ID+CbA+WU	.826	.536	.616	.573	.780	.829	.778	.803
DD+ID+CbA+RES	.825	.534	.616	.572	.779	.827	.781	.803
DD+ID+CbA+COS	**.830**	**.545**	.612	.576	.757	.826	.733	.777

В **четвертой главе** для задачи выделения тематически сгруппированных объектов мнений используются методы автоматического резюмирования мнений относительно тематических категорий. В данной группе задач под резюмированием мнений (англ. sentiment summarization, opinion summarization) понимают идентификацию k основных тематических групп аспектов продукта, где тематическая группа определена как множество слов в текстах, которые имеют тенденцию встречаться совместно с определенной тональностью в отзывах пользователей.

Компании и разработчики могут быть более заинтересованы в устранении сбоев приложений, вызывающих наибольший негативный отклик, нежели во мнениям о нехватке функционала или о цветовой гамме продукта. В свою очередь, с каждым годом анализ требуемых отзывов вручную становится более затруднительным в связи с ростом количества отзывов в сети. Это объясняет актуальность поставленной задачи.

В рамках задачи предложены новые модели, являющиеся модификацией модели латентного размещения Дирихле (latent Dirichlet allocation, LDA). Модель *тематических высказываний, указывающих на проблемную ситуацию* (**TPrPhModel**), используется для определения проблемных индикаторов относительно тематических категорий отзывов. Модель оценивает распределения проблемных индикаторов и целевых объектов как независимые распределения в пространстве слов. Также предлагается модель *тема-тональность-проблема* (**TSPM**) для анализа взаимосвязи между информацией о проблемных ситуациях и тональности высказываний относительно тематических категорий. Для решения задачи статистического оценивания применя-

ется сэмплирование Гиббса[10] (англ. Gibbs sampling), число итераций равно 1000. Графические представления моделей приведены на Рисунке 1.



а) TPrPhModel б) TSPM

Рис. 1 — Вероятностные модели а) TPrPhModel и б) TSPM

При описании **TPrPhModel** используются следующие обозначения:

- D – коллекция документов; w_d – вектор слов документа d;
- V – словарь коллекции (множество уникальных слов в D);
- K, S, R – число тем, тональных и проблемных классов, соответственно; $R = 2, S = 3$;
- $z_{di}, \tilde{z}_{di}$ – темы, присвоенные i-му контекстному слову w_i и i-му проблемному слову pw_i в d, соответственно;
- r_{di} – проблемная метка, присвоенная i-му слову в документе d, $r_{di} \in \{pr, no-pr\}$, где pr - метка слов, указывающая на проблемную ситуацию; $no - pr$ - метка, указывающая на отсутствие проблемы;
- α, β, γ – априорное распределение Дирихле;
- $n_{d,k,j,w}$ – количество раз, когда слову w в документе d присвоена тема k и проблемная метка j; далее "$*$" обозначает суммированное количество по индексу в $n_{*,*,*,*}$. $n_{*,k,j,*}$ – общее количество слов, которым присвоена пара (k, j);
- n_d – количество слов в документе d;
- ϕ_z – мультиномиальное распределение в пространстве слов для темы z с параметром β, $\Phi = \{\{\phi_z\}_{z=1}^{K}\}_{v=1}^{V}$;
- $\tilde{\phi}_{z,r}$ – мультиномиальное распределение в пространстве слов для (z, r) с параметром $\tilde{\beta}$, $\tilde{\Phi} = \{\{\{\tilde{\phi}_{z,r}\}_{r=1}^{R}\}_{z=1}^{K}\}_{v=1}^{V}$.

TPrPhModel учитывает, что целевые объекты не обладают признаком проблемности (в отличие от проблемных индикаторов), то есть не могут

[10]Griffiths T. L., Steyvers M. Finding scientific topics // Proceedings of the National Academy of Sciences. — 2004. — Т. 101, suppl 1. — С. 5228—5235

указывать на существование проблемных высказываний. Объекты играют роль фактической информации о составных компонентах продуктов. Согласно графической модели, слово w_d в документе d порождается в зависимости от некоторой латентной темы z; слово pw_d – в зависимости от некоторой латентной темы $\tilde{z}$ и проблемной метки r. Для слова i-го слова в документе d определен индекс $t = (d, i)$. Скрытые параметры темы $\tilde{z}$ и проблемной метки r могут быть оценены по следующей формуле для всех слов pw:

$$P(\tilde{\mathbf{z}}_t = k, \mathbf{r}_t = r | \mathbf{pw}_t = w, \tilde{\mathbf{z}}_{-t}, \mathbf{r}_{-t}, \mathbf{pw}_{-t}, \alpha, \tilde{\beta}, \gamma) \propto \frac{n^{-t}_{*,k,r,w} + \tilde{\beta}^w_r}{n^{-t}_{*,k,r,*} + \sum_{i'=1}^{V} \tilde{\beta}^{i'}_r} \frac{n^{-t}_{d,k,r,*} + \gamma}{n^{-t}_{d,k,*,*} + R * \gamma} \frac{n^{-t}_{d,k,*,*} + \alpha}{n^{-t}_{d,*,*,*} + K * \alpha} \quad (1)$$

Скрытые параметры темы z в модели **TPrPhModel** могут быть выбраны по формуле, совпадающей с формулой сэмплирования в LDA, для всех контекстных слов w:

$$P(\mathbf{z}_t = k | \mathbf{w}_t = w, \mathbf{z}_{-t}, \mathbf{w}_{-t}, \alpha, \beta) \propto \frac{n^{-t}_{*,k,w} + \beta^w}{n^{-t}_{*,k,*} + \sum_{i'=1}^{V} \beta^{i'}} \frac{n^{-t}_{d,k,*} + \alpha}{n^{-t}_{d,*,*} + K * \alpha} \quad (2)$$

Используя полученные оценки модели, порождение слов в документах происходит согласно Алгоритму 2.

Описание проблемной ситуации с продуктами сопровождается эмоционально-окрашенными фразами: пользователь может описывать технические дефекты и неполадки в процессе использования продукта в отзыве, который не содержит эмоционально-окрашенных слов (например, “не могу открыть флэшку”, “машине требуется ремонт”), или сопровождать отзыв негативной или позитивной тональностью, если описываются ситуации с комфортным использованием продукта. Для определения взаимосвязи различных типов информации предложена **TSPM** в 2 модификациях:

1. Модель **TSPM(DP)**, в которой проблемные переменные слов зависят от локального контекста: тональной и тематической переменных слов документов;
2. Модель **TSPM(GP)**, в которой проблемные переменные слов зависят от общего контекста коллекции и моделируются из распределения пар (тема, тональность), агрегируя информацию из троек (документ, тема, тональность).

Algorithm 2: Алгоритм порождения слов с помощью TPrPhModel

Function `sampling`(*корпус документов,* α, β, $\tilde{\beta}$, γ)
  Input: гиперпараметры α, β, $\tilde{\beta}$, γ, корпус документов
  Output: присвоенные темы всех слов и проблемные метки проблемных слов в корпусе
  Инициализировать присвоение тем для всех контекстных слов и пар (тема, проблемная метка) для всех проблемных слов случайным образом
  foreach $iter = 1$ *to максимальное количество итераций* **do**
    foreach *документ* $d \in 1, \ldots, D$ **do**
      foreach *проблемное слово* $i \in 1, \ldots, n_{pw_d}$ **do**
        Исключить слово i, присвоенное теме k и проблемной метке r, из счетчиков $n_{k,r,i}, n_{k,r}, n_{d,k,r}, n_{d,k}$ и n_d
        Сэмплировать новую пару (k', r') из уравнения 1
        Обновить счетчики $n_{k,r,i}, n_{k,}, n_{d,k,r}, n_{d,k}$ и n_d для k', r'
      foreach *контекстное слово* $i \in 1, \ldots, n_{w_d}$ **do**
        Исключить слово i, присвоенное теме k из $n_{k,i}, n_k, n_{d,k}$ и n_d
        Сэмплировать новую тему k' из уравнения 2
        Обновить $n_{k,i}, n_k, n_{d,k}$ и n_d для k';

При описании модели **TSPM** будут использоваться обозначения, указанные для модели **TPrPhModel**, и следующие обозначения:

- l_{di} – тональная метка, присвоенная i-му слову в d; $l_{di} \in \{neu, neg, pos\}$, где neu, neg, pos – классы тональности;
- η – априорное распределение Дирихле на параметры ξ;
- $\xi_{z,l}$ – мультиномиальное распределение в пространстве проблемных меток для пары (тема z, тональная метка l) с параметром η;
- $\phi_{z,l,r}$ – мультиномиальное распределение пространстве слов для троек (тема z, тональная метка l, проблемная метка r) с параметром β, $\Phi = \{\{\{\{\phi_{z,l,r}\}_{r=1}^{R}\}_{l=1}^{L}\}_{z=1}^{K}\}_{v=1}^{V}$;
- $n_{d,k,l,r,w}$ – количество раз, когда слову w в документе d присвоена тема k, тональная метка l и проблемная метка r.

В рамках **TSPM** каждой теме соответствует мультиномиальное распределение в пространстве слов. В **TSPM(DP)** выбирается проблемная метка r из мультиномиального распределения ξ, соответствующего тройке (доку-

мент d, тема z, тональная метка l). В **TSPM(GP)** проблемная метка выбирается из мультиномиального распределения $r_{d,w_i} \sim Mult(\xi^{z,l})$, независящего от документа. Используя сэмплирование Гиббса, скрытые параметры в модели **TSPM(DP)** могут быть выбраны по следующим формулам:

$$P(\mathbf{z}_t = k, \mathbf{l}_t = l, \mathbf{r}_t = r | \mathbf{w}_t = w, \mathbf{z}_{-t}, \mathbf{l}_{-t}, \mathbf{r}_{-t}, \alpha, \beta, \gamma, \eta) \propto$$
$$\frac{n^{-t}_{*,k,l,r,w} + \beta^w_{l,r}}{n^{-t}_{*,k,l,r,*} + \sum_{i'=1}^{V} \beta^{i'}_{l,r}} \frac{n^{-t}_{d,k,l,r,*} + \eta}{n^{-t}_{d,k,l,*,*} + R * \eta} \frac{n^{-t}_{d,k,l,*,*} + \gamma}{n^{-t}_{d,k,*,*,*} + L * \gamma} \frac{n^{-t}_{d,k,*,*,*} + \alpha}{n^{-t}_{d,*,*,*,*} + K * \alpha} \quad (3)$$

Качество предложенных моделей оценивается с помощью нескольких критериев, предъявляемых к тематическим моделям для задач анализа мнений. Качество тематического моделирования слов оценивается с помощью перплексии[11]. Для обучения моделей собрана коллекция текстов ($\sim$ 69 тыс. отзывов). 10% отзывов, выбранных случайным образом, используются для подсчета перплексии. Перплексия связана с правдоподобием модели:

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{|D_{test}|} \log p(\mathbf{w}_d)}{\sum_{d=1}^{|D_{test}|} n_d}\right) \quad (4)$$

Для оценки вклада предложенных скрытых переменных и распределений слов, которые используются для описания взаимосвязей между темой и проблемными индикаторами в документах, модели оцениваются в рамках задачи классификации с помощью стандартных метрик качества систем анализа текстов на естественном языке.

Для сравнения использовались базовые тематические модели, порождающие слова или предложения в документе в зависимости от тематической и тональной переменных: JST, Reverse-JST[12], ASUM[13]. Для определения гиперпараметров β используются словарь тональных слов (SL) и словарь проблемных индикаторов (PL), $K = 5$. Постфиксы "+SL" и "+PL" свидетельствуют, что базовая модель использует соответствующий вид словаря. **TPrPhModel** использует словарь PL. **TSPM** обучена на обоих видах словарей. Результаты экспериментов по оценке качества моделей представлены в Таблице 7.

[11]Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation //the Journal of machine Learning research. – 2003. – Т. 3. – С. 993-1022.

[12]Weakly Supervised Joint Sentiment-Topic Detection from Text / C. Lin [и др.] // IEEE Transactions on Knowledge and Data Engineering. — 2012. — Т. 24, № 6. — С. 1134—1145

[13]Yohan J., H. O. A. Aspect and Sentiment Unification Model for Online Review Analysis // Proceedings of the Fourth ACM International Conference on Web Search and Data Mining. — ACM, 2011. — С. 815—824.

Таблица 7 — Перплексия вероятностных моделей (чем меньше величина, тем лучше модель предсказывает появление слов w в документе d)

Метод	Коллекции отзывов пользователей					
	Электро-ника	Инстру-менты	Детские товары	Машины (анг.)	Машины (рус.)	Прило-жения
JST+SL	1799.108	1599.084	754.608	1322.535	1451.302	1502.473
R.-JST+SL	2282.377	2134.653	916.123	2013.413	1544.798	1850.458
ASUM+SL	1854.151	1528.275	1189.605	1321.351	1403.584	1501.21
JST+PL	1941.149	1740.916	796.324	1446.680	1573.653	1458.751
R.-JST+PL	2412.924	2165.745	1056.324	2031.738	1504.514	1625.145
ASUM+PL	1811.037	1551.750	1040.451	1357.859	1314.124	1489.360
TPrPhModel	**842.448**	**714.416**	**347.831**	**620.124**	**1049.288**	**604.203**
TSPM(DP)	1524.455	1382.600	663.128	1113.759	1136.421	1069.963
TSPM(GP)	1769.500	1495.423	761.635	1285.700	1274.758	1321.336

Экспериментально подтверждено, что модели **TSPM(DP)** и **TSPM(GP)** показывают наименьшие значения перплексии по сравнению с моделями JST, Reverse-JST, ASUM. Таким образом, скрытая проблемная переменная, зависимая от темы и тональности слова, вносит дополнительный вклад в процесс моделирования слов. Наименьшие значения перплексии модели **TPrPhModel** подтверждают улучшение качества тематического моделирования относительно базовых моделей независимо от коллекции.

Для оценки вклада скрытых переменных в рамках задачи классификации применяется вероятностный подход и оценивается $P(r|d)$. Документ классифицируется как проблемное высказывание, если $P(r = pr|d) > P(r = no - pr|d)$. Вероятности классов подсчитаны из Φ (на примере TSPM(DP)):

$$P(r|d) \propto P(d|r) = \prod_{w \in w_d} P(w|r) = \prod_{w \in w_d} \sum_{z=1}^{K} \sum_{s=1}^{S} P(w|r, s, z) \tag{5}$$

Для изучения влияния количества тем и оценки вклада скрытых переменных результаты классификации моделей с различными K представлены на Рисунке 2. Экспериментально подтверждено, что отзыв пользователя по структуре относится к типу связного текста из нескольких подтем и предложенные модели **TPrPhModel** и **TSPM** достигают наилучшие значения достоверности и F-меры на небольшом количестве тем (5–10 тем) по сравнению с базовыми методами. Таким образом, совместное моделирование темы, тональных и про-

блемных переменных с помощью предложенных моделей помогает улучшить классификации предложений, что подтверждает эффективность добавления скрытых переменных в модель латентного размещения Дирихле.

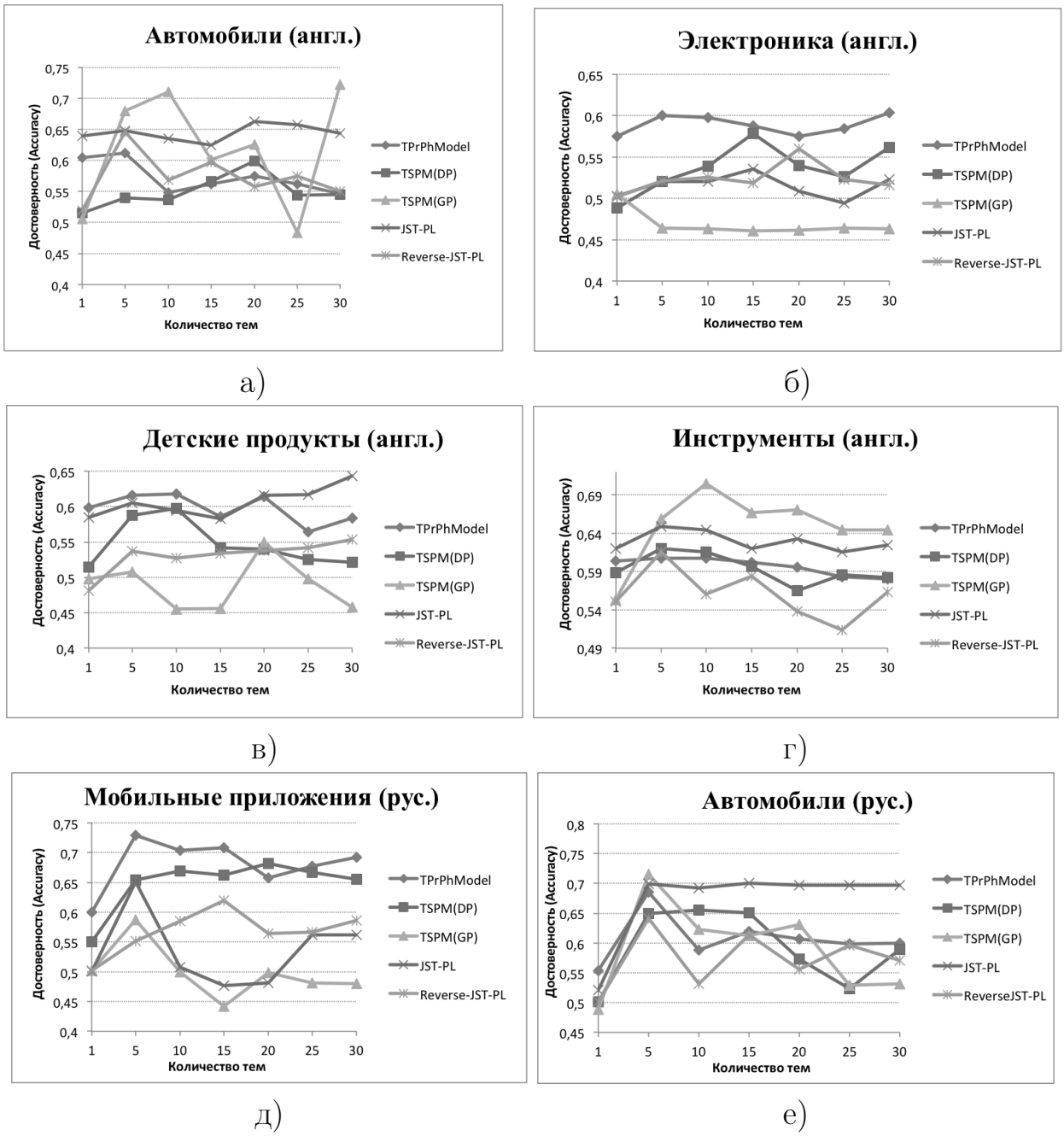


Рис. 2 — Результаты классификации текстов пользователей для вероятностных моделей, обученных на разном количестве тем

Для качественной оценки тематических распределений моделей был проведен анализ тем. Анализ порожденных тем подтверждает, что **TPrPhModel** определяет различия между проблемными словами применительно к различных целевым объектам продукта. **TSPM(DP)** и **TSPM(GP)** ассоциируют различные индивидуальные или общие аспекты

с претензиями пользователя об удобстве использования или о дефектах в зависимости от позитивного или негативного отношения пользователя к теме.

В рамках диссертационной работы была разработана программная система, описания модулей которой приводятся в соответствующих методам главах. Программный комплекс по извлечению высказываний пользователей и построению тематических моделей написан на языке Java и выложен в открытый доступ[14]. Модули извлечения высказываний могут взаимодействовать друг с другом по принципу конвейера.

В **заключении** приведены основные результаты работы:

1. Предложен и реализован метод классификации предложений, основанный на знаниях в виде созданных словарей и правилах, учитывающих грамматическую структуру сложных предложений относительно союзов.
2. Предложен и реализован метод классификации предложений отзывов пользователей по отношению к целевым объектам, связанным с предметной областью, на основе синтаксических связей слов и мер семантической связанности.
3. Предложены и реализованы две вероятностные модели для задачи выделения тематически сгруппированных объектов мнений, учитывающие скрытые переменные для описания тем и проблемных индикаторов совместно.
4. Разработано программное обеспечение и проведено экспериментальное исследование, подтверждающее улучшение качества предложенных методов по сравнению с существующими алгоритмами.

В **приложении А** приведен список лексических единиц из созданных словарей.

Публикации автора по теме диссертации

1. *Тутубалина Е. В.* Совместная вероятностная тематическая модель для идентификации проблемных высказываний, связанных нарушением функциональности продуктов // Труды Института системного программирования РАН. — 2015. — Т. 4, № 27. — С. 100—120.

[14] `https://bitbucket.org/tutubalinaev/dissertation/`

2. *Тутубалина Е. В.* Извлечение проблем, связанных с неисправностями и нарушением функциональности продуктов, на основании отзывов пользователей // "Вестник КГТУ им. А. Н. Туполева". — 2015. — Т. 3. — С. 139—146.

3. *Ivanov V.*, *Tutubalina E.* Clause-based approach to extracting problem phrases from user reviews of products // Analysis of Images, Social Networks and Texts. — Springer International Publishing, 2014. — С. 229—236.

4. *Tutubalina E.* Target-Based Topic Model for Problem Phrase Extraction // Advances in Information Retrieval. — Springer International Publishing, 2015. — С. 271—277.

5. *Tutubalina E.* Dependency-Based Problem Phrase Extraction from User Reviews of Products // Text, Speech, and Dialogue. — Springer International Publishing, 2015. — С. 199—206.

6. *Tutubalina E.*, *Nikolenko S.* Inferring Sentiment-Based Priors in Topic Models // Advances in Artificial Intelligence and Its Applications. — Springer International Publishing, 2015. — С. 92—104.

7. Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars / V. Ivanov [и др.] // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Т. 14. — 2015. — С. 22—34.

8. Supervised Approach for SentiRuEval Task on Sentiment Analysis of Tweets about Telecom and Financial Companies / E. Tutubalina [и др.] // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Т. 14. — 2015. — С. 65—75.

9. *Tutubalina E.*, *Ivanov V.* Unsupervised Approach to Extracting Problem Phrases from User Reviews of Products // COLING 2014. — 2014. — С. 48—53.

10. *Tutubalina E.* Mining Complaints to Improve a Product: a Study about Problem Phrase Extraction from User Reviews // Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. — ACM. 2016. — С. 699—699.