Федеральное государственное автономное образовательное учреждение
высшего образования «Казанский (Приволжский) федеральный университет»

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»

Диссертация на соискание учёной степени
кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук, профессор
Соловьев Валерий Дмитриевич

Казань — 2016

## Оглавление

Стр.

# Введение

Диссертация посвящена разработке моделей и методов извлечения информации о высказываниях пользователей, содержащих указания на трудности в использовании продуктов (сервисов, товаров) и требующие устранения причин претензий от компаний. Рассмотрены наиболее распространенные задачи анализа мнений – классификация текстовых документов, извлечение высказываний относительно объектов мнений определенной предметной области, а так же 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, сайтов 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 [2; 3].

В последние десятилетия на рынке потребительских товаров появилась резкая динамика увеличения количества технически сложных товаров [4]. 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. Многие покупатели осуществляют возврат товаров компаниям даже, если товар работает исправно согласно государственным стандартам и техническим отчетам компаний, что негативно сказывается на доверии покупателей и имидже компании.

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 исследованиях [3—9] исследуется феномен “ошибка не найдена”как класс 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. Работы [10—14] дают 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* [15—17], базовых тематических моделей [17; 18] и методов машинного обучения на небольшом наборе признаков [19—21].

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;
- 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 метод извлечения проблемных фраз по отношению к объектам, относительно которых высказывается проблемная фраза (далее целевые объекты) и связанных с предметной областью, на основе общедоступного тезауруса;
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 су-

ществующих 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, связанных 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 ряд скрытых переменных для описания тем и проблемных индикаторов совместно.
4. Разработано программное обеспечение и проведено экспериментальное исследование, обосновывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами.

**Степень достоверности** подтверждается корректностью разработанных методов и моделей, 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, описанными в работах [16; 19; 22].

**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 к извлечению 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 и т.п.

**Научная новизна.** Задачи извлечения информации о высказываниях пользователей, указывающих на проблемные ситуации с продуктами, являют-

ся 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: перечисление основных конференций, симпозиумов:

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

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 ИТИС КФУ.

**Публикации.** Основные результаты по теме диссертации изложены в 10 печатных изданиях, 2 из которых изданы в журналах, рекомендованных ВАК [23; 24], 6 из которых изданы в журналах, входящих в базу SCOPUS [25—30], 2 — в тезисах докладов [31; 32].

**Личный вклад.** Автором проведено исследование предметной области, выполнен основной объём теоретических и экспериментальных исследований, изложенных в диссертационной работе, разработана программная система на основе созданных методов. В работе [25] Иванову В.В. принадлежит постановка задачи и привлечение разметчиков для получения экспертных оценок контрольной выборки. В работах [29; 30] вклад группы соавторов ограничен обсуждением результатов и тестированием классификаторов на различных наборах признаков. В работе [28] Николенко С.И. предложил формулу для расчёта гиперпараметров.

**Объем и структура работы.** Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации составляет 145 страниц с 5 рисунками и 36 таблицами. Список литературы содержит 150 наименований.

# Глава 1. Современное состояние исследований

В настоящее время в связи с быстрым развитием методик проектирования систем Web 2.0 и увеличением количества пользовательского контента на онлайн-ресурсах (блогах, форумах, сайтах, социальных сетях, сервисах электронной коммерции) анализ мнений (англ. opinion mining) и анализ тональности (англ. sentiment analysis) стали полезными инструментами для извлечения сложно-структурированной информации из web-ресурсов. Подобный анализ информации позволяет выявить общественные мнения в текстах различных тематик. В рамках задач используются коллекции текстовых документов, где каждый документ содержит отзыв конкретного пользователя онлайн-ресурса о различных объектах ресурса. В последние десятилетия в мировой науке было предложено много различных подходов в рамках задачи анализа мнений пользователей и задачи анализа тональности отзывов, подробный обзор которых описан в работах [10—13; 33]. Термин “opinion mining” появился в статье [34] в рамках задачи определения признаков продуктов, таких как качество и функциональность, используя результаты информационного поиска с агрегированием мнений относительно каждого признака (полезный, сложный, неисправный, проч.). Термин “sentiment analysis” появился в нескольких работах в рамках различных задач автоматической обработки текстов (анализ настроений рынка продаж, классификация отзывов как отличных или плохих), используя статистические и лингвистические подходы к определению эмоциональной окраски текста [35—38]. Ранние работы по анализу тональности были сосредоточены на бинарной классификации текстов пользователей (негативный или позитивный классы), в то время как работы по анализу мнений заключались в извлечении субъективных и объективных суждений пользователей о продуктах, фильмах и прочих объектах [12]. В настоящий момент термины “sentiment analysis” и “opinion mining” определяют более схожие понятия в исследованиях и означают статистическое оценивание мнений, тональности, субъективности, достоверности и другой информации в текстах пользователей. Анализ мнений пользователей, как эффективный инструмент мониторинга и оценивания конкретных групп пользователей, используется в различных приложениях социальных сетей [39], включая рейтинги в опросах общественного мнения [40], рынок цен-

ных бумаг [41], анализ событий [42], рекламные механизмы (поведенческий таргетинг, рекомендательные сервисы) [43; 44], техническая поддержка клиентов крупных компаний [18; 21] и т.д.

За последние годы было выделено несколько ключевых задач анализа мнений пользователей:

– классификация текстов на уровне документов и предложений (англ. document sentiment classification, sentence sentiment classification);
– анализ мнений по отношению к аспектных терминам, относительно которых высказывания были сделаны (англ. aspect-based sentiment analysis);
– идентификация оценочных слов (англ. sentiment lexicon identification).

## 1.1 Классификация текстов пользователей на уровне документов и предложений

За последнее десятилетие было проведено большое количество исследований по классификации текстов пользователей на два или более классов, которые разделяют мнения:

1. На 2 класса (положительные и отрицательные) и 3 класса (положительные, отрицательные и нейтральные)[22; 38; 45—57];
2. На объективные и субъективные [58—65];
3. На подлинные и фальшивые [66];
4. На отличные (thumbs up) и плохие (thumbs down) [5; 37; 38; 67];
5. На мнения, содержащие определенный вид информации (сарказм, спам, ирония, указание на дефект или улучшение продукта) [18; 21; 68—70].

Задача классификации текстов пользователей относится к традиционным задачам автоматической обработки естественного языка, в которых документы традиционно классифицировались по темам, например: спорт, политика, наука [11; 14]. В качестве критериев оценки качества методов используются стандартные метрики анализа текстов: достоверность (англ. accuracy), точность (англ. precision), полнота (англ. recall) и F-мера (англ. $F_1$-measure). Все предложен-

ные методы для автоматической классификации текстов пользователей можно разделить на следующие группы:

1. Методы, основанные на лингвистическом анализе, синтаксических правилах и шаблонах [37; 51; 55; 71—73];
2. Машинное обучение без учителя (англ. unsupervised methods) [60; 74—79];
3. Машинное обучение с учителем (англ. supervised methods) [22; 38; 45—50; 53; 56; 57; 59; 80].

Традиционные работы по анализу тональности, как одной из задач обработки текста, используют подходы, основанные на словарях оценочных слов и статистических мерах. В исследовании [37] предложен лингвистический подход анализа текста для извлечения тональных фраз в предложении: используются шаблоны на основе частей речи, учитывающие синтаксические отношения слов в предложении. Затем метод определяет тональность фраз, подсчитывая точечную взаимную информацию (англ. pointwise mutual information) между фразой и оценочными словами на основе данных выборки поискового запроса. Последующие исследования определяют тональность слова как разницу поточечной взаимной информации между словом в корпусах позитивных и негативных текстов [22; 74]. Многие подходы подсчитывают суммарную тональность текста на основе словарей оценочной лексики, содержащие слова с числовым значением априорной тональности [72; 73]. Методы учитывают отрицания и частицы, усиливающие тональность слова в тексте. Однако большинство работ, использующие лингвистические подходы, отмечают необходимость создания дополнительных предметно-ориентированных словарей оценочных слов для точной классификации текстов в соответствии с тематикой документов или предложений.

В настоящий момент многие исследования по задаче обработки текста чаще всего сводятся к задачам машинного обучения, где требуется сформировать вектор признаков и создать обучающую выборку. Затем статистический или вероятностный классификатор [81; 82] обучается по выборке и проверяется качество классификации на коллекции текстов определенной предметной области. В рамках задачи анализа мнений большинство работ исследуют эффективность различных векторов признаков для классификации отзывов или отдельных предложений отзывов, учитывая тональность. В первом исследовании [38] используются наивный байесовский классификатор (англ. Naïve Bayes)

и метод опорных векторов (англ. support vector machine, SVM) на основе мешка слов (англ. bag of words) для задачи бинарной классификации отзывов о фильмах. В последующих работах исследуются более сложные вектора признаков для улучшения результатов классификации методами машинного обучения. В работе [83] используется метод опорных векторов и применяется метод активного обучения (англ. active learning) для уменьшения размера обучающей выборки. Большинство предложенных в работах признаков можно разделить на следующие группы:

- признаки, основанные на частотности всех слов в тексте;
- признаки, учитывающие синтаксические зависимости слов в тексте и части речи слов;
- признаки, построенные на словарях оценочных слов;
- признаки, основанные на правилах и вхождениях отрицаний в текст;
- структурные признаки, использующие синтаксис сообщений из микроблогов социальных сетей.

В работах [45; 48] анализируется эффективность синтаксических признаков. В работе [46] анализируется добавление лингвистических признаков в вектор признаков для классификатора. В работах [47; 49] используются признаки, учитывающие изменение тональности слов за счет отрицаний в тексте. В работах [50; 57; 83] анализируются различия между векторами признаков для эффективной классификации отзывов, текстов форумов и сообщений из микроблога. В работе [22] анализируется эффективность признаков нескольких типов (синтаксические; признаки, построенные на нескольких словарях оценочных слов; структурные признаки) в рамках задачи анализа коротких сообщений в социальной сети Twitter.

С точки зрения классификации отзывов пользователей на русском языке по тональности интерес представляют несколько исследований, выполненных в рамках Российского семинара по оценке методов информационного поиска (РОМИП). В работе [84] приводится описание коллекций о различных сущностях (фильмы, книги, цифровые фотокамеры) на русском языке, в [80] приводится обзор методов классификации отзывов пользователей на русском языке. Приведены оценки эффективности алгоритмов на описанных корпусах отзывов пользователей. Статистически лучшие результаты показали методы машинного обучения, основанные на методе опорных векторов (SVM) и модели мак-

симальной энтропии, где в качестве классификационных признаков использовались оценочные слова. Исследование [55] посвящено задаче автоматической классификации отзывов о книгах по материалам семинара РОМИП. В качестве базовых классификационных признаков для методов машинного обучения рассматриваются все слова документа за исключением служебных частей речи, числительных и дат, а также простые именные группы. Для увеличения количества признаков авторы предлагают лингвистический подход, расширяя список атрибутов книг за счет синонимов и гипонимов с использованием словарей оценочной лексики. В работе [54] исследуется метод расширения классификационных признаков для автоматической классификации отзывов о книгах. Авторы используют лингвистический подход, применяя семантические фильтры для объединения нескольких фактов в один класс. Семантические фильтры автоматически пополнялись системой. Авторы приводят оценки эффективности метода по двух классификаторам: SVM и модели линейной регрессии (англ. linear regression). По результатам тестирования выявлено, что метод опорных векторов, основанный на леммах (отдельных словах) и не использующий дополнительные лингвистические признаки, дает лучшие оценки. Авторы полагают, что это связано с невозможностью удалить шумовую лексику с помощью семантических фильтров. В работе [56] было показано, что методы машинного обучения не являются универсальными, поскольку каждый классификатор показал наилучшие результаты лишь в одной из предметных областей. В целом следует отметить, что задача классификации мнений для русского языка изучена в меньшей степени, чем для английского языка. Отсутствуют в отрытом доступе хорошо проработанные словари позитивной и негативной лексики.

Исследования по анализу мнений на английском и русском языках подтверждают, что классификаторы, обученные на текстах определенной предметной области, показывают сравнительно низкие результаты классификации на новых текстах других предметных областей, в то время как создание обучающей выборки для переобучения классификации трудозатратно по времени и требует качественной ручной разметки.

Основным инструментом задач анализа тональности мнений является словарь оценочных слов. Подобные словари используются во многих прикладных задачах. В качестве слов чаще всего выступают прилагательные и наречия. Однако не существует одного универсального словаря, который подходит для каж-

дой предметной области или тематической категории, поскольку тональность слов является предметно-зависимой [85]. Безусловно, использование предметно-ориентированных словарей оценочных слов показывают улучшение результатов во многих задачах, включая классификацию текстов [86] и информационный поиск [87]. Существует несколько основных подходов к автоматическому извлечению оценочных слов из текстов:

1. Подходы, использующие экспертные знания и лингвистические ресурсы (тезаурусы, словари) [88; 89];
2. Подходы, основанные на правилах и частотности слов-кандидатов в тестовых коллекциях [90; 91];
3. Подходы, использующие методы машинного обучения [92—94].

Работы, использующие методы из первых двух групп, полагают, что существует небольшой список позитивных и негативных слов. Работы [90; 91] описывают лингвистические правила для извлечения новых оценочных слов. В исследованиях показано, что (i) два связанных слова с помощью союза *но* (*but*) содержат противоположную тональность; (ii) два связанных слова с помощью союза *и* (*and*) содержат одинаковую тональность. Методы, описанные в [88; 89], используют семантические отношения между существующими лексическими единицами в структуре электронных ресурсов, полагая, что тональность синонимичных слов совпадают в то время, как антонимы обладают противоположной тональностью. В работе [37] тональность слова определяется как разница поточечной взаимной информации между словом в корпусах позитивных и негативных текстов или между словом и двумя оценочными словами *бедный* (*poor*) и *превосходный* (*excellent*). Данный метод используется в работах [22; 74; 95]. В работе [94] используют коллекцию коротких сообщений, размеченную автоматическим способом за счет комбинаций принятых символов для передачи позитивных и негативных эмоций (смайликов) для обучения классификатора на мешке слов. Конечный вес признака классификатора, основанного на методе опорных векторов, используется как тональный вес для слов из текстовой коллекции.

Поскольку мнения пользователей содержат предметно-зависимую тональность для различных тематических категорий аспектных терминов, многие работы используют тематические вероятностные модели для создания лексикона [75—78]. В настоящий момент доминирующими методами создания словаря оценочных слов являются алгоритмы на основе векторного представления слов

(англ. word embeddings) и нейронных сетей (англ. neutral network) и алгоритмы на основе модели латентного размещения Дирихле. В работах [92; 93] описаны алгоритмы добавления информации о тональности предложения в векторное представление слов на основе нейронных сетей, показывающие наилучшие результаты классификации коротких сообщений по сравнению с популярными методами машинного обучения, использующие популярные словари оценочных слов MPQA и NRC-Emotion для английского языка.

Существует небольшое количество работ, посвященных созданию словаря оценочных слов для русского языка. Работа [84] посвящена извлечению предметно-ориентированного словаря оценочных слов на русском языке. В работе используется метод опорных векторов, использующий набор статистических и лингвистических признаков. Ряд признаков используют коллекцию отзывов заданной предметной области (фильмы, книги, телефоны, камеры) и контрастную коллекцию новостей. В исследовании [52] предложен метод подсчитывающий веса оценочных слов, используя пять статистических мер на коллекции коротких сообщений. В настоящее время не существует доступного русскоязычного словаря оценочной лексики, основанного на подходах первых двух групп. В данный момент не известны работы по применении модификаций тематических моделей для задачи анализа мнений на русском языке. В исследовании [96] проведен анализ применений тематических моделей к задаче извлечения однословных терминов. Результаты показывают, что использование тематической информации значительно улучшает качество автоматического извлечения терминов.

## 1.2 Анализ мнений по отношению к аспектным терминам

Задачу классификации текстов в целом по тональности отделяют от задачи анализа мнений относительно аспектных терминов, о которых высказывание было сделано (англ. aspect-based sentiment analysis). Аспектным термином (далее аспектом) называется слово или словосочетание, определяющее конкретный признак или составную часть продукта или сервиса. В данной задаче анализа естественного языка отзыв пользователя рассматривается как документ, описы-

вающий иерархическую структуру связного текста следующим образом: общая тема документа может быть описана посредством более конкретных тем текста, которые так же могут быть охарактеризованы более точными темами [97]. Отзывы о конкретном продукте с одинаковой тональностью могут содержать мнения о различных аспектах (признаках) продукта. Например, в предложении вида "мне нравится дисплей телефона, но батарея разряжается слишком быстро" пользователь описывает два аспектных термина (дисплей, батарея) с разной тональностью. Отзыв о продукте может быть описан посредством тематических категорий (например, качество, надежность, внешний вид, батарея, удобство), поскольку пользователь может быть недоволен медленной ответной реакцией дисплея, низкой эффективностью батареи, слишком громким или тихим звуком динамика, избыточным количеством функций, короткими проводами, недостаточно ярким цветом корпуса и т.д. Подобная структура связного текста объясняет необходимость анализа мнений. Задачи, исследуемые в рамках анализа мнений по отношению к аспектных терминам (далее аспектам), можно подразделить на следующие группы:

- *идентификация аспектов* (англ. *aspect identification*). Данная группа задач направлена на извлечение множества целевых объектов (компонентов, составных частей), относящихся к продукту, описанному в документе;
- *анализ тональности относительно аспектов* (англ. *aspect-based sentiment classification*). Одной из наиболее характерных задач данной группы является классификация тональных высказываний относительно аспектов документа;
- *резюмирование мнений по тематическим категориям* (англ. *multi-document opinion summarization*). Данная группа задач направлена на идентификацию $k$ основных тематических групп аспектов продукта и определение тональности.

### 1.2.1 Идентификация аспектных терминов

Задача идентификации аспектов из текстов определенной предметной области является хорошо изученной задачей и может быть рассмотрена как задача извлечения информации из отзыва с предположением, что каждое мнение выражается относительно целевых объектов (признаков продукта). В ряде работ аспект классифицируется на следующие типы: (i) явный аспект (англ. explicit aspect); (ii) неявный aspect (англ. implicit aspect); (iii) тональный факт (англ. fact) [11; 98; 99]. Явный аспект является конкретным признаком или составляющей продукта (например, батарея, экран), неявный аспект содержит в себе тональность и указание на тематическую категорию текста. Например, в предложении “телефон является медленным и дорогим” касается качества и цены продукта. Большинство предложенных в работах методов можно разделить на следующие группы:

1. Методы, извлекающие наиболее частотные существительные и именные группы [74; 100; 101];
2. Методы, исследующие возможные синтаксические связи в предложении между оценочными словами и целевыми объектами [98; 102];
3. Методы, использующие алгоритмы машинного обучения или тематического моделирования [22; 76—79; 83; 103—106].

Существует несколько наиболее популярных методов, решающих задачу извлечения аспектов как бинарную задачу классификации [107], как задачу классификации последовательностей (англ. sequential classification) [22; 103; 105], как задачу тематического моделирования или традиционную задачу кластеризации [76—79; 106]. В рамках задачи классификации требуется классифицировать извлеченное из текста существительное или словосочетание как аспект определенного типа. В работе [107] описан метод извлечения именных словосочетаний с частотностью выше установленного значения с последующим подсчетом точечной взаимной информации между словосочетанием и фразами, которые связаны отношением *часть-целое* с классом продукта. В работе [100] используется подход, основанный на частотности однословных и многословных именных выражений, которые извлекаются с помощью синтаксических шаблонов из предложений с оценочными словами. В работе [101] так же используется

подход к извлечению аспектов, основанный на частотности слов-кандидатов с применением дополнительных шаблонов для сокращения множества. По аналогии с работой [102], в исследовании [98] описан подход, основанный на правилах, для извлечения явных и неявных аспектов. Метод использует словарь оценочных слов, состоящий из 30,000 многословных выражений с заданными тональностями, и деревья зависимостей для определения связей между оценочными словами и словами-кандидатами. Авторы отмечают, что предложенный подход зависит от качества инструментов анализа зависимостей и от словаря оценочных слов.

В данный момент доминирующими методами являются методы классификации последовательностей (англ. sequential classifiers), основанные на обучении с учителем и часто использующиеся в задачах извлечения информации: скрытая марковская модель (англ. Hidden Markov Model, HMM) и условные случайные поля (англ. Conditional Random Fields, CRF). В работе [108] описана модификация HMM для совместного извлечения мнений наряду с их явными аспектами. В работах [22; 103—105] авторы используют модель CRF для извлечения явных аспектов, чтобы присвоить каждому предложению последовательность оценочных слов с соответствующими полярностями, определенными в зависимости от принадлежности к сущности мнения.

Статистические тематические модели так же используются для идентификации аспектов в рамках более комплексной задачи выделения тематически сгруппированных целевых объектов продуктов и тональных высказываний в коллекции документов. Методы [76—79; 106], использующие модель латентного размещения Дирихле, описаны в следующем подразделах. С точки зрения идентификации аспектов на русском языке интерес представляют несколько исследований, выполненных в серии тестирований в области систем анализа тональности SentiRuEval. Наилучшие результаты в задаче извлечения аспектов показали следующие методы: метод, основанный на правилах; метод, основанный на рекуррентных нейронных сетях; и метод классификации последовательностей, использующий метод опорных векторов на наборе следующих признаков: морфологических, синтаксических и семантических признаках [99]. Методы, использующие CRF, показали средний результат для извлечения аспектов на русском языке, в отличие от доминирующих методов для английского языка [109]. Это может быть связано со структурой предложений в русском языке: сво-

бодный порядок слов уменьшает вероятность вхождения размеченной цепочки слов в коллекции текстов, что усложняет обучение условных случайных полей.

### 1.2.2 Анализ тональности относительно аспектов

Задача тональности относительно аспектов подразумевает определение тонального класса, к которому относится извлеченный из текста аспект. По аналогии с задачей классификации текста, все предложенные методы можно разделить на две группы: методы машинного обучения с учителем и методы, основанные на словарях оценочных слов. Многие работы применяют методы машинного обучения, описанные в задаче классификации текстов на уровне документов и предложений. Специфика задачи заключается в том, чтобы определить тональность определенного в тексте высказывания относительно аспекта. Для учета специфики этих задач многие исследования используют синтаксический анализ для определения зависимостей и релевантной к аспекту информации. В работе [107] для выделения фраз, содержащих мнений об аспекте, используются синтаксические шаблоны на основе деревьев зависимостей. Тональность слов в выделенной фразе определяется на основе метода машинного обучения без учителя. В работах [22; 83; 110] используется синтаксический анализатор на основе грамматики зависимостей для извлечения слов в дереве зависимостей, которые связаны с целевым объектом. В работе [83] описан подход присваивания веса словам на основе дерева зависимостей и удаленности узлов, соответствующим слову и аспекту в дереве, учитывающий разницу в глубине дерева между узлами, длину пути между узлами с помощью поиска в ширину, удаленность слов в предложении. В работе [111] предложен подход присваивания веса словам в зависимости от удаленности в тексте относительно аспекта, учитывающий длину отзыва и позиции слова и упомянутого аспекта.

Существует несколько работ об анализе тональности относительно аспектов в отзывах пользователей о машинах и ресторанах на русском языке, выполненных в серии тестирований SentiRuEval. Метод машинного обучения, использующий градиентный бустинг (англ. gradient boosting classifier), показал наилучшие результаты классификации [95]. В качестве набора признаков ис-

пользовался метод мешка слов, полученный из контекста аспекта в тексте в 2 вариантах: три ближайших слова в тексте к аспекту и шесть слов к аспекту. В работе [22] используется словарь оценочных слов, созданный автоматическим образом как разница точечной взаимной информации между словом и коллекциями позитивных и негативных высказываний. Участники тестирования не использовали синтаксический анализатор на основе грамматики зависимостей в силу отсутствия открытых программных модулей, предназначенных для синтаксического анализа текстов, что не позволяет использовать подходы анализа тональности относительно аспекта, разработанные для английского языка.

### 1.2.3 Выделение тематически сгруппированных объектов мнений продуктов и тональных высказываний

Благодаря популярности сайтов, агрегирующих мнения пользователей (например, `tripadvisor.com`, `yelp.com`, `otzovik.com`), количество отзывов о продуктах и сервисах компаний резко возросло. Отдельный отзыв о продукте не содержит достаточное представление о качестве, поэтому пользователи предпочитают анализировать ряд мнений с целью принятия решения о покупке продукта. Однако увеличение числа пользовательского контента привело к усложнению процесса поиска требуемой информации на основе текстов, прочитанных на сайтах [112].

На текущий момент традиционно используемые методы классификации отзывов на два или более класса (позитивные или негативные; несколько классов, эквивалентные пользовательскому рейтингу) никак не отражают тот факт, что отзыв пользователя как связный текст посвящен раскрытию подтем основной темы, и каждое слово текста непосредственно относится к некоторой подтеме с определенным классом тональности. Данная группа задач направлена на идентификацию $k$ основных тематических групп аспектов продукта. Под темой в компьютерной лингвистике понимается множество общих слов в текстах, которые имеют тенденцию встречаться совместно в одних и тех же текстах.

Традиционным методом резюмирования мнений по категориям является кластеризация — выделение близких по содержанию кластеров предложений

[113]. В работе [114] предложен метод, состоящий из следующих шагов: идентификация аспектов продукта из отзывов, используя метод из работы [71]; присвоение аспекта к аспектной категории на основе иерархической структуры онтологии; кластеризация предложений, содержащих упоминание об аспекте. Все аспекты в структуре онтологии ранжируются по тональности. Предложения присваиваются категориям в зависимости от тональности и положении аспекта в онтологии. Работы [71; 115] так же используют существующие лингвистические онтологии для организации аспектов.

В настоящий момент доминирующими методами являются алгоритмы на основе модели латентного размещения Дирихле [116] (англ. latent Dirichlet allocation, LDA) для задачи выделения тематически сгруппированных целевых объектов продуктов и тональных высказываний в коллекции документов. Это связано с тем, что создание предметно-ориентированной обучающей выборки достаточно большого размера для классификатора требует больших затрат времени в то время, как вероятностные модели позволяют использовать коллекции неразмеченных документов, содержащиеся на онлайн-ресурсах, для нахождения скрытых переменных. Определение темы в LDA позволяет создать лингвистическую модель генерации контента документов и разработать алгоритм выявления распределения слов и документов по темам. В тематических моделях для задач анализа мнений, как правило, используется две различные модели:

- модель мешка слов, в которой каждый документ рассматривается как набор встречающихся в нём слов [75; 77; 78; 106];
- модель мешка тональных фраз, состоящих из аспекта и оценочного слова, в которой каждый документ рассматривается как набор встречающихся в нём фраз [76; 79].

В работах [75—78] представлены тематические модели, направленные на объединение задач идентификации аспектных терминов в отзыве и определения тональности для этих аспектов. Эти модели используют словарь позитивных и негативных слов для задания гиперпараметра $\beta$ (априорное распределение Дирихле на мультиномиальном распределении $\phi$ в пространстве слов для темы). В работе [76] фразы, содержащие комбинацию аспектных терминов и оценочных слов, являются наблюдаемыми переменными с двумя скрытыми переменными (тематической и тональной). В данной работе понятие «аспект» как множество эквивалентно понятию “тема” в тематическом моделировании (или

тематической категории). Авторы проводят сравнительный анализ четырех модификаций LDA, оценивая необходимость взаимосвязи скрытой тональной и тематической переменных в моделях мешка слов и тональных фраз. Результаты оценивания качества построенных тематических моделей показывают, что наилучшие результаты достигаются для модификации LDA, основанной на мешке тональных фраз. Выбор темы и тональности являются зависимыми событиями, то есть тональность слова зависит при тематической категории данного слова в отзыве.

В работе [77] авторы описывают модель *тональность-тема* (англ. joint sentiment-topic model, JST) и модель *тема-тональность* (англ. Reverse-JST), добавляя скрытую тональную переменную для моделей. Авторы предполагают, что в JST распределение тем в каждом документе зависит от тонального распределения, в Reverse-JST верно обратное. В моделях предполагается, что слово в документе порождено некоторой латентной темой и некоторой латентной тональной меткой. Таким образом, каждой тональной метке соответствует мультиномиальное распределение в пространстве тем, парам (тональная метка, тема) соответствует мультиномиальное распределение в пространстве слов. Эксперименты показали, что модели показывают наилучший результат классификации в нескольких доменах (книги, фильмы, электроника). В работе [75] описана объединенная модель *аспект-тональность* (англ. aspect and sentiment unification model, ASUM). Под аспектом понимается тема в отзывах пользователей. Авторы полагают, что каждое предложение отзыва принадлежит одной теме и тональности. ASUM моделирует аспекты из мультиномиального распределения в пространстве тональности для предложения, слово порождено некоторым аспектом и тональностью предложения. Эксперименты показывают, что ASUM показывает лучшие результаты тональной классификации по сравнению с JST. Однако, в работах [75; 77] авторы добавляют скрытую тональную переменную, фиксируя статичные гиперпараметры для слов, входящих в словари эмоционально-окрашенной лексики, что является недостатком предложенных подходов по добавлению знаний о тональности слов. В работе [106] авторы встраивают в модификацию тематической модели компоненту дифференциальной максимальной энтропии, чтобы отделить темы целевых объектов и эмоционально-окрашенную лексику.

В работе [117] описывается статистическая модель для извлечения и категоризации аспектных терминов на основе списка тематических слов для каждой категории, в которой заинтересован пользователь. В работе [118] описывается вероятностная тематическая модель для совместного извлечения свойств и признаков продуктов. К качестве коллекции текстов в работе используется коллекция фрагментов (фраз) из социальных сетей.

В исследованиях [78; 79] описываются частично маркированные модификации LDA (англ. partially labeled topic models) [119], где, кроме слов как наблюдаемых переменных, используются тональный признак слова на основе известной оценки продукта или метаданные о пользователе. В работе [79] описана модификация LDA, включающая 2 типа знаний: рейтинг продукта по пятибалльной шкале в нескольких категориях и словарь эмоционально-окрашенной лексики для идентификации тональности отзывов. В работе [78] представлена модификация LDA, названная User-aware Sentiment Topic Models (USTM), включающая в распределения метаданные профайлов пользователя и словари эмоционально-окрашенной лексики для определения связи между тематическими наборами аспектов и категориями пользователей. Авторы полагают, что существует взаимосвязь информации о пользователе (пол, возраст, место жительства) с темами, которые пользователь комментирует в отзывах. Модель показывает наилучшие результаты классификации отзывов о машинах и ресторанах по сравнению с популярными вероятностными моделями JST и ASUM. Таким образом, модель позволяет не только выяснить мнение конкретного пользователя относительно разных аспектов, но и определять общественные мнения для группы лиц.

## 1.3 Анализ конструктивных фраз пользователей

В последние годы необходимость оценки качества текстов, сгенерированных пользователями, привлекла пристальное внимание многих научных групп. С технологических инноваций на рынке потребительских товаров появилась резкая динамика увеличения количества технически сложных товаров. Компании стремятся обеспечить высокое качество продукции и оперативно устранять

сбои, влияющих на работу продуктов. Однако количество пользовательского контента не позволяет оперативно выявить различные проблемы с продуктами и сервисами компании из отзывов пользователей, используя чтение текстов как средство получения сведений. Это объясняет необходимость создания автоматических методов анализа мнений с целью идентифицировать фразы пользователей, содержащих явное или косвенное указание на неполадки с продуктами.

### 1.3.1 Анализ высказываний, содержащих проблемную ситуацию

Высокий уровень качества продукции является характеристикой продукта (товара, услуги), которая влияет на конкурентоспособность компании в условиях рыночной экономики [120]. В последнее десятилетие в литературе появились работы, направленные на изучение трудностей в использовании сложных технических продуктов, что ведет к неудовлетворенности потребителей. В работах [3—9] исследуются претензии потребителя к качеству продуктов. Работа [8] описывает феномен “ошибка не найдена” и содержит подробный обзор работ за десятилетие, рассматривая класс ошибок, которые не могут быть легко найдены, диагностированы и воспроизведены в режиме тестирования. В работе [3] анализируется взаимосвязь между удовлетворенностью потребителя с продуктами и сервисами компании и позитивными общественными мнениями. В работе [4] анализируется качество и надежность электронных продуктов с точки зрения потребителей. Авторы приводят классификацию неисправностей на трудные поломки (связанные со сбоем работы продуктов) и гибкие отказы (функционирующие согласно спецификации, но не ожиданию потребителя). Таким образом, традиционная техническая служба поддержки клиентов компании направлена лишь на устранение технических проблем, классифицируя второй вид ошибок как “отсутствие неисправностей”. Однако неудовлетворение потребностей потребителей вследствие второго вида проблем влечет возврат продукта. Авторы полагают, что компании должны учитывать доступную обратную информацию от пользователей, чтобы избежать подобных ситуаций на уровне технической поддержки. В исследовании [5] анализируется новый вид потребительских претензий, названные *мягкими проблемами* (англ. *soft problems*). Данный

вид претензий не связан с техническими дефектами электронного продукта. Авторы полагают, что тенденция объединения разной функциональности в едином устройстве (например, функции телефона, фотоаппарата, плеера в смартфоне) привела к резкому увеличению претензий пользователей, связанных с удобством использования. В работе выделяется несколько видов категорий мягких проблем: функциональность, производительность, восприятие, здоровье, тренд, механизм, техническое обслуживание, ограничения. Половину претензий занимают проблемы из категорий функциональности, производительности и ограничений. К категории ограничений относятся претензии на недостаток функции или улучшений. Затруднения с пониманием и поиском функций относятся к функциональности. В работе [6] проведен анализ информации от пользователей с целью улучшения дизайна электронных продуктов. В работе [7] описан методический разбор интеграции данных из текстов пользователей о ситуациях использования продуктов в проектирование систем технической поддержки компаний.

Таким образом, в настоящий момент существует потребность в методах автоматической идентификации затруднений в использовании продуктов, которые будут непосредственно связаны с прикладной задачей для компаний. Задача автоматического анализа высказываний пользователей, на предмет обнаруженных ими проблем в использовании тех или иных устройств, является менее изученной по сравнению с задачами анализа мнений. Существует несколько работ, направленные на извлечение информации о проблемах с продуктами на основе коротких сообщений из социальных сетей [19; 21], отзывов пользователей о продуктах [20; 121; 122], отзывах о мобильных приложениях [16—18; 123; 124], web-документах [15; 125] и проч. В работах ставятся задачи идентификации (i) предложений, описывающих проблемные ситуации; (ii) целевых объектов, которых эти проблемы касаются; (iii) категоризации предложений на основе слов (в т.ч. целевых объектов и проблемных индикаторов для выявления наиболее проблемных компонент продукта. В большинстве работ под проблемным индикатором понимается многословная конструкция, указывающая на технические ошибки. Однако в данный момент общепринятого определения нет.

Первая группа работ исследует задачу анализа фраз пользователей мобильных приложений. В работах [16; 17] рассматривается задача классификации текстов пользователей, отражающих существование ошибок, запросов и

требований о функциях мобильных приложений. В исследовании [17] на основе списка из 24 ключевых слов (например, *missing* (*отсутствующий*), *request* (*запрос*), *lacks* (*не хватать*)) были выбраны предложения для извлечения шаблонов. Метод идентифицировал 237 лингвистических шаблона запросов таких как “wish <request> instead of <existing feature>” (“хотелось бы <запрос> вместо <существующая функция>”), «the only thing missing <request>” (“единственно отсутствует <запрос>”), “please include <request>” (“пожалуйста, добавьте <запрос>”). Для резюмирования найденных запросов используется стандартная модификация LDA, извлекая новые ключевые слова для темы об обновлении, поддержке и приложении в целом. В работе [16] авторы приводят классификацию ошибок с приложениями по скорости устранения: серьезные, средние и второстепенные. Вследствие серьезных ошибок приложение не подлежит использованию, средние ошибки касаются проблем с конкретной функцией, второстепенные ошибки связаны с внешней функциональностью (клавиатурой, камерой) и не затрудняют работу. В исследование составлен список из 37 проблемных индикаторов для серьезных ошибок, такие как *problems with* (*проблемы с*), *restart* (*рестарт*), *horrible* (*ужасно*), *malfunction* (*несправность*). Метод идентифицировал 74 лингвистических шаблона ошибок таких как “stopped downloading, running, syncing” (“перестало загружаться, запускаться, синхронизоваться”), “impossible to <action>” (“невозможно <действие>”). В обеих работах авторы рассматривают задачу как задачу классификации и не сравнивают полученные результаты с другими методами классификации. В работе [17] исследуется задача идентификации упоминаний пользователей о дефектах и запросах на улучшение приложения eBay в магазинах приложений App store и Google Play. Авторы отмечают, что лишь 20% отзывов о приложении содержат требуемую информацию. В исследование используется классификатор на основе метода опорных векторов, обученный на двух выборках по отдельности: (i) небольшой коллекции отзывов, размеченной вручную; (ii) автоматически созданной коллекции отзывов, размеченной с помощью метода отдаленного наблюдения (англ. distant supervision) на основе шаблонов. Если отзыв удовлетворял одному из 13 лингвистических шаблонов, то тексту присваивался класс упоминаний дефектах/улучшениях, в противном случае, негативный класс. Результаты экспериментов показали, что в обоих задачах наилучшие результаты показал классификатор, обученный на первой выборке. Классификатор, осно-

ванный на выборке с применением правил, показывает сопоставимые результаты полноты с первым классификатором, для которого требуется размеченная обучающая коллекция. Для категоризации слов автор использует стандартную модель LDA, показывая, что темы из словосочетаний и глагольных групп более информативны, чем отдельные слова.

Подход, описанный в работе [123], состоит из 3 методов: (i) идентифицировать информативные отзывы о приложении, удалив нерелевантные и бессмысленные тексты; (ii) категоризовать отзывы с помощью тематической модели; (iii) использовать ранжирующую схему и присвоить отзывам вес. Авторы отмечают, что лишь 35,1% отзывов в магазине приложений Google Play содержат конструктивную информацию о багах или требованиях, чтобы помочь разработчикам улучшить продукт. Первый метод основан на наивном байесовском классификаторе с EM алгоритмом (Expectation Maximization). В качестве алгоритма кластеризации использовались тематические модели LDA и ASUM [75], описанные ранее. Предложенная ранжирующая функция учитывает временной интервал между релизами приложений, средний рейтинг и плотность группы. Результаты показывают, что подход с LDA показывает лучшие результаты классификации, чем подход с ASUM.

Исследование [20] посвящено задаче извлечения отзывов о программах и компьютерных играх, содержащих описание опыта пользователя и удобства использования. Метод машинного обучения с моделью мешка слов и частотной схемой TF-IDF был использован для классификации на уровне предложений для различных тематических категорий аспектов таких как запоминаемость, обучаемость, выгодность, эффективность, ошибки, удовлетворение, комфортность, поддержка. Авторы не приводят результаты других методов классификации на тестовой коллекции для сравнения. Создание обучающей выборки показало, что из лишь 13%–49% предложений 3492 отзывов содержат информацию, релевантную к описанным категориям. Таким образом, необходимость разметки выборки большого размера так же является существенным недостатком данной работы. Дополнительно, в работе приводятся наиболее информативные слова для категорий на основе вектора весов классификатора.

Другая группа работ исследует классификацию фраз пользователей об электронных устройствах из текстов отзывов и коротких сообщений на английском языке [19; 21; 122]. В работах [19; 21] используется классификатор на

коллекции коротких сообщений (твитов) о компании AT&T в социальной сети Twitter. В работах [19] используется классификатор, основанный на методе максимальной энтропии, на наборе тональных и синтаксических признаков. В качестве синтаксических признаков используются бинарные признаки вхождения слов из списка проблемных индикаторов в текст. Список проблемных индикаторов включает несколько глаголов действия, таких как *fail* (*упасть*) и *crash* (*рухнуть*); 10 глаголов работы в комбинации с лексемами, выражающие действие, не достигшее результата, такие как *stop work* (*перестать работать*), *refuse connect* (*отказаться соединяться*); 6 проблемных индикаторов-существительных, таких как *problem* (*проблема*) и *trouble* (*неисправность*) и несколько многословных выражений. Автор использует набор признаков, основанных на словарях позитивных и негативных слов, для обучения классификатора, однако не приводит анализ важности созданных признаков в задаче обнаружения проблем. Для идентификации проблемной фразы предложен набор синтаксических шаблонов. Классификатор на предложенном наборе признаков достигает наилучшее значение F-меры (0.742) по сравнению с классификатором на модели мешка слов (F-мера равна 0.66). Исследование [21] посвящено извлечению целевых объектов (аспектов в задачах тональности), связанных с проблемным индикатором (триггером в статье) при синтаксическом анализе. Задача рассматривается как задача классификации именных групп с помощью метода максимальной энтропии. Классификатор достигает наилучшего значения F-меры (0.75) по сравнению методом, использующий синтаксические шаблоны (F-мера равна 0.64). Исследование [122] посвящено задаче классификации проблемных высказываний о продуктах компании Hewlett-Packard на форуме компании. В работе предложен метод извлечения фраз, основанный на правилах и словаре проблемных индикаторов ProblemWord. Данный словарь построен вручную и схож со словами из работы [19]. Дополнительно, в работе предложен способ расширения словаря автоматическим способом на основе многословных выражений в корпусе текстов Google Books NGram. Google Books NGram содержит корпуса размеченных текстов книг на 9 языках. Многословные выражения (N-граммы) представляют собой последовательность от одного до пяти слов. В работе используется правило вхождения негативных слов в предложение для идентификации проблемных высказываний, не анализируя влияние нейтральной или позитивной тональности. Метод учитывает вхождения глаголов с отрицаниями

как индикатор невыполнения действия в процессе эксплуатации электронного продукта. Метод, использующий словарь Problemord, показывает наилучшее значение F-меры (0.74) по сравнению с методом, основанным на расширенном словаре (0.69).

В работе [15] анализируются тексты на японском языке во всем Web и ищутся упоминания о трудностях и проблемах использования любых артефактов. Автоматически создается словарь проблемных индикаторов: для слова trouble (затруднение) в словарь добавляется все его синонимы и гипонимы по шаблонам “X similar Y”, “X called Y”, “X like Y” и др. Для извлечения целевых объектов подсчитывается точечная взаимная информация между проблемным индикатором и существительными. Затем объекты-кандидаты отбираются на основе синтаксических шаблонов “глагол-существительное”, составленные на основе статистики в web-корпусе. Создание обучающей выборки показало, что лишь 7% web-документов содержат информацию о проблемах. На тестовом корпусе из 200 документов значение точности составляет 85.5%.

Исследование [14] посвящено задаче определения тональности глаголов как оценочных слов в задачах тональности. В работе используются марковские сети (англ. markov networks) для моделирования лингвистических признаков и синтаксической зависимости между глагольными выражениями в тексте. Предложенный алгоритм сравнивается с классификаторами: методом опорных векторов и наивным байесовским классификатором. В работе отмечается необходимость идентификации глагольных выражений для прикладных задач устранения неполадок с продуктами и сервисами. Обучающий корпус построен автоматическим образом: глагольные выражения в заголовках отзывов с сайта `amazon.com` с минимальной оценкой пользователя рассматриваются как негативные, с максимальной оценкой как не негативные.

Автору диссертации неизвестны исследования, посвященные задаче автоматического извлечения информации о существовании различных проблем с продуктами и сервисами на русском языке.

### 1.3.2 Анализ объективных и информативных мнений

Ряд исследований направлен на задачу определения объективных мнений и фактов из коллекции отзывов пользователей [58—65; 126]. Используя небольшой список объективных слов, размеченных вручную, в работе [126] предложен метод кластеризации слов для создания словаря субъективных прилагательных. Исследование [60] описывает типы синтаксических шаблонов для определения фраз. В работах [59; 63] используются методы машинного обучения для задачи классификации текстов на отзывы и факты на уровне документов и грамматических основ предложений. В работе [61] определяется субъективность слов в текстах корпуса новостных документов, используя словари, составленные автоматическим способом и зависящие от корпуса.

Задача классификации текстов на объективные и субъективные часто исследуется как аспект анализа тональности, где методы идентифицируют субъективность текста с последующим определением позитивной и негативной окраски. В работе [83] предложен каскадный классификатор, где сначала определяется наличие субъективности в виде тональности в тексте, а затем определяется класс тональности. Обзор работ подробно описан в статьях [64; 65].

Ряд исследований направлен на задачу определения качества пользовательского текста (полезность, достоверность, целесообразность) [5; 67; 127; 128]. В работах утверждается, что многие позитивные и негативные отзывы могут быть бессодержательны [67]. Поэтому многие интернет-ресурсы предлагают систему голосования, рекомендуя отзыв к прочтению другим пользователям (например, «полезен ли отзыв?» на `ozon.ru`). Данные работы рассматривают задачу анализа качества как задачу классификации или задачу регрессии, используя результаты голосований как размеченные данные. В исследовании [67] рассматривается задача идентификации полезности отзыва как ортогональную задачу к анализу тональности. Предложенный метод машинного обучения показал, что наиболее эффективными признаками являются синтаксические признаки (количество собственных имен, цифр, модальных глаголов, прилагательных и наречий в сравнительной форме). В дополнение, пользовательские оценки продукта являются эффективными признаками при обучении классификатора [5]. Однако работы данной группы, используя систему голосования сайтов,

определяют качества пользовательского текста для других пользователей, а не для разработчиков данного продукта. В исследовании [127] анализируется зависимость между голосами пользователей для рекомендации отзывов и независимой разметкой экспертов в задаче классификации. Авторы отмечают закономерность рекомендации отзыва на «отлично», если отзыв содержит описание продукта, или содержит схожие оценки рекомендации, что так же подтверждено в исследовании [128]. В работе описан подход к разметке отзывов на четыре класса: лучший отзыв (содержащий детальную информацию о множестве аспектов), хороший отзыв (отзыв содержит рекомендацию без описания использования), честный отзыв (содержащий краткое описание о нескольких аспектах), плохой отзыв (содержит недостоверную информацию). Результаты разметки отзывов показали, что ручная разметка совпала с разметкой пользователей в системе голосований в 15% случаев. Таким образом, подтверждается необходимость создания тестовой выборки, размеченной вручную без использования оценок пользователей на сайте, и необходимость создания автоматических методов анализа качества существующих отзывов в дополнение к определению тональности.

## 1.4 Выводы к первой главе

В данной главе проведен обзор основных методов и подходов, применяемых в задачах анализа мнений пользователей. Данная группа задач востребована на практике.

Анализ предметной области показал, что 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 каче-

ства 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 (напр., 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 вектор 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.

# Глава 2. Извлечение высказываний, указывающих на проблемные ситуации с продуктами, на основании отзывов пользователей

## 2.1 Постановка задачи

Пусть $P = \{P_1, P_2, \ldots, P_m\}$ - множество продуктов (сервисов, товаров), выпускаемое компаниями на потребительском рынке. Каждая текстовая коллекция состоит из отзывов пользователей о продуктах определенной предметной области (например, электроника, автомобили, приложения). Для каждого продукта $P_i \in P$ 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. В некоторых отзывах пользователи сообщают о дефектах продуктов, недостаточной удовлетворенности в использовании или нехватке определенного функционала. Каждый продукт $P_i \in P$ состоит из множества целевых объектов (компонентов, составных частей) $T_i = \{t_1, t_2, \ldots, t_k\}$.

*Замечание.* В данной работе для разработки более робастных методов автоматического извлечения информации не используется последующая синтаксическая сегментация предложений: предложение $s_{ij}$ отзыва $d_i = \{s_{i1}, \ldots, s_{i|di|}\}$ рассматривается как единичный элемент отзыва, поскольку данный элемент обладает определенным семантическим значением.

Формальное описание поставленной задачи дано в последующем разделе.

### 2.1.1 Формальное описание задачи

**Определение**. *Мнение* – это суждение или точка зрения, выражающее оценку или взгляд на какой-то объект (информацию). Точка зрения пользователя онлайн-ресурса может быть не объективно мотивированна и описывать субъективную информацию.

**Определение**. Под *отзывом* пользователя подразумевается грамматически организованная последовательность слов, описывающая мнения автора относительно объектов мнения (например, продукта или сервиса).

**Определение**. *Объект мнения* — это конкретный или абстрактный продукт, событие или сервис, по поводу которого складывается мнение.

**Определение**. *Пользователь* – это человек, обладающий доступом к онлайн-ресурсу с возможностью пользоваться функциональностью ресурса (например, покупать продукты, использовать сервисы, просматривать страницы, оценивать продукты).

**Определение**. *Высказыванием, указывающим на проблемную ситуацию с продуктом*, или проблемным высказыванием называется 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 проблемных высказываний.

Примеры высказываний с $r(s_{**}) = 1$:

– “Недорогая и бесшумная, *трудность* в том, чтобы *найти комплектующие при их порче* (ведерко, мешалка)”.
– “*Подключиться* через IM-4G никак *не получается* - пишет “*ошибка подключения к сети*”.
– “После пробега 25000 км. появилось: сдох активатор центрального замка, дефект панельки, плохо дует вентилятор в стекло”.
– “Кондиционер работал, *но* из-за него вся машина *дребезжала* и *трясло* руль”.
– “*Невозможно проверить* баланс, зависает телефон и *приходится перегружать*”.

Каждое предложение представляет из себя множество слов $s_{ij} = \{f(w)|f \in s_{ij}\}$, состоящее из множества целевых объектов отзыва, оценочных слов, проблемных индикаторов и общеупотребительных слов.

**Определение**. Проблемный индикатор – это однословная или многословная конструкция, выражающая явное или косвенное указание на проблему с

продуктом. Примеры проблемных индикаторов: *трудность*, *отказывается работать*, *слишком яркий*.

**Определение**. *Целевым объектом* (также называющийся признаком продукта [34], аспектом [11; 75] или аспектным термином [99]) называется элемент отзыва, относительно которого высказывается некоторое мнение, представленный в виде однословной или многословной конструкции, характеризующей тему документа в определенной предметной области. Целевой объект чаще всего описывает компоненту или атрибут продукта $P_i \in P$. Примеры аспектов: "подсветка экрана", "дверь багажника", "бортовой компьютер", "приложение".

**Постановка задачи**. Требуется извлечь преобладающие высказывания, указывающие на проблемные ситуации с продуктом $P_i \in P$ и его целевыми объектами $T_i$, используя множество пользовательских отзывов $D$.

В рамках диссертации исходная задача подразделяется на следующие подзадачи, соответствующие особенностям задач анализа мнений [10]:

1. Идентификация высказываний, указывающих на проблемные ситуации в использовании продуктов, из текстов пользователей;
2. Извлечение высказываний о проблемных ситуациях по отношению к целевым объектам, зависящих от предметной области, в отзыве пользователя;
3. Выделение тематически сгруппированных целевых объектов продуктов для извлечения преобладающих проблемных ситуаций в коллекции отзывов определённой предметной области.

В главе 2 рассматривается задача идентификации проблемных высказываний как задача бинарной классификации предложений из текстов пользователей. Целью задачи является определение класса высказывания по числовому значению $r(s_{ij})$ для всех предложений документов контрольной выборки $s_{ij} \in d_i, j \in \{1, \ldots, |d_i|\}$, $i \in \{1, \ldots, |D|\}$.

## 2.2 Классификация пользовательских высказываний для описания проблем с продуктами

В данном разделе описывается предложенная классификация высказываний, используемых пользователями, для описания проблем, обнаруженных в ходе эксплуатации продуктов.

В литературе не существует общепринятой классификации высказываний пользователей о неполадках продуктов. Поэтому постановка задачи идентификации проблемных высказываний с целью классификации текста предполагает предварительный анализ отзывов пользователей, проведенный вручную, и анализ текущих схем классификации, встречающийся в литературе, в качестве дополнительного знания о задаче.

В литературе существует различная классификация на два и более классов. В работе [124] приводится классификация отзывов на: отчет об ошибке, запрос о функционале, отзывы об опыте использования, отзыв как отражение рейтинга (например, похвала, осуждение, отвлеченная критика, дискуссия). В работе [5] выделяется два класса проблемных высказываний: технические ошибки и мягкие проблемы, связанные с удобством использования. Данные работы классифицируют проблемные высказывания по разновидности последующих тестирований. В работе [17] авторы приводят классификацию ошибок с приложениями по скорости устранения на серьезные (продукт не подлежит использованию), средние (связанные с конкретной функцией продукта) и второстепенные (не затрудняют работы). В отличие от работы [124], исследования [16; 18] рассматривают нехватку функционала и пожелания как отдельную задачу определения запросов пользователей. В данной работе нехватка функционала рассматривается как вид проблемного высказывания, когда пользователь ощущает потребность в изменении продукта или сервиса.

После анализа высказываний из отзывов пользователей на русском и английском языках мы выделили фразы четырех типов:

– **Явное упоминание о проблемах с продуктом**. Данный тип фраз содержит явное указание на дефекты и технические неполадки в процессе использования продукта. Примеры: “не открывается дверь”, “возник-

ли проблемы с компьютером", "машине требуется ремонт", "программа не работает должным образом".

– **Косвенное упоминание о проблемах с продуктом**. Данный тип этих фраз не содержит явного упоминания о проблеме, но содержит вспомогательные слова и подразумевает проблему, следствием которой является неудовлетворенность пользователя продуктом. Примеры: "плохой дизайн", "отвратительный багажник", "слабый сигнал сети", "запутанное приложение".
– **Отрицание затруднений при использовании продукта**. Используя данный тип фраз, пользователь отрицает ранее упомянутую неполадку или ожидаемые проблемы. Примеры: "нет никаких претензий к качеству сборки", "вождение не приносит дискомфорта", "ходовая часть несколько жестковата, но зато это придает стабильность машине".
– **Отсутствие информации о проблемах с продуктом**. Мнение пользователя не содержит упоминаний ожидаемых или фактических затруднений. Примеры: "ездил на таком автомобиле около года назад", "идеальное сочетание цены и качества", "большой плюс то, что она достаточно грузоподъемная".

## 2.3 Создание словаря оценочной лексики на русском и английском языках

Описанные типы проблемных фраз содержат информацию о существовании тех или иных проблем с продуктами на основе индикативных конструкций. Одной из ключевых задач, являющейся основой при разработке методов для анализа мнений в текстах, является создание словарей индикативных слов. В задаче извлечения проблемных высказываний под индикативными словами и выражениями понимаются *проблемные индикаторы*, выражающие явное или косвенное указание на проблему с продуктом. Описанные типы проблемных фраз содержат информацию о существовании тех или иных проблем с продуктами на основе явных проблемных индикаторов (например: *трудность*, *проблема*, *трудноуправляемый*, *испортиться*, *гниение*, *ржавчина*, *оставляет желать*

*лучшего*), негативных слов (например, *дохлый*, *идиотский*, *некачественный*, *кошмарный*, *утомительный*), вспомогательных слов (например: *слишком*, *изрядно*, *избыточно*, *чрезмерно*) и отрицаний действий (например: *не едет*, *не работает*, *не ускоряется*, *не чиниться*).

Предыдущие работы по анализу мнений показали, что применение не зависищих от конкретной предметной области словарей индикативных конструкций, созданных автоматическим способом на основе больших текстовых коллекций, показывает схожие или меньшие результаты классификации относительно применения словарей, построенных вручную [22]. Таким образом, для достижения целей исследования в работе используются словари, составленные вручную.

Для создания словарей индикативных конструкций для каждого языка был использован описанный далее алгоритм. Небольшой список индикаторов ProblemWord, содержащий такие слова, как *проблема*, *ошибка*, *сгнить*, *протухнуть*, *ущерб*, *претензия*, *жалоба* для русского языка и *problem*, *error*, *failure*, *malfunction*, *fault* для английского языка, был составлен вручную на основе анализа различных типов проблемных фраз. Лексические единицы из словаря ProblemWord можно разделить на три основные группы: явные проблемные индикаторы (например: *дефектный*, *вмятина*), обозначенные далее как DirectPW; оценочные слова с негативной тональностью, обозначенные далее как NegativePW и связанные с удобством использованием продукта (например: *непонятный*, *дискомфорт*, *плохой*); глаголы, обозначенные далее как VerbPW и указывающие на проблемные ситуации в ходе эксплуатации продукта (например: *ломаться*, *промокать*, *разграбить*). Исходный список содержал около 200 слов для каждого языка. Затем был применен метод расширения списка слов путем добавления всех синонимов лексических единиц мультиязычного универсального онлайн-словаря Викисловарь[1], соответствующих словам из списка. Наконец, расширенный список был проверен вручную для удаления ошибочно добавленных слов. Таблица 1 содержит примеры проблемных индикаторов для двух языков. Слова из небольшого списка выделены жирным шрифтом. Словарь NotProblemWord, составленный вручную, включает слова, указывающие на работу, положительную ситуацию или исправление недостатков (например: *наладить*, *удобно*, *несущественно*, *комфортно*, *устойчиво*, *легковыполнимый*). Проблемные высказывания, содержащие явное указание на не выполне-

[1] `https://ru.wiktionary.org/`

Таблица 1 — Примеры проблемных индикаторов, добавленные в словари

| Примеры проблемных слов и выражений | | | | | |
|---|---|---|---|---|---|
| для русского языка | | | для английского языка | | |
| **проблема** | нарушение | неловко | error | **death** | unhappy |
| препятствие | **портить** | некомфортно | defect | **garbage** | distressed |
| затруднение | вредить | дискомфорт | fault | waste | **useless** |
| **авария** | повреждать | неуместно | mistake | refuse | unskillful |
| крушение | ухудшать | **поломка** | **difficult** | trash | **return** |
| ущерб | **ненадежный** | **хамство** | cumbersome | **failure** | **tech support** |
| **ошибка** | сомнительный | грубость | hard | **screw up** | **tech service** |
| неправильность | подозрительный | беспардонность | noise | fuck up | technical support |
| **сбой** | **испорченный** | **ломаться** | unresponsive | **scratch** | **send back** |
| перебой | **неудобно** | портиться | nonresponsive | **upset** | **refuse** |

ние какой-либо функции или отрицание действия, выражаются конструкциями $neg\_verb_A$, где $verb_A$ – глагол действия, neg указывает на отрицание частицей (например, “не”, “нет”) или наречием (например, “невозможно”). Словарь Action для русского языка был собран следующим образом: все глаголы были извлечены из Викисловаря, затем глаголы действия были выбраны вручную. Словарь Action для английского языка был заимствован из работы [122].

В качестве словарей NegativeWord и PositiveWord для английского языка был выбран общественно доступный словарь оценочный лексики MPQA[2], часто использующийся в работах по анализу мнений для английского языка. Поскольку в настоящее время нет русскоязычного словаря оценочных слов, эффективность которого показана в рамках соревнований классификации SentiRuEval, словари NegativeWord и PositiveWord для русского языка были созданы вручную на основе отзывов пользователей об автомобилях. Для создания словарей были собраны отзывы пользователей с сайта `otzovik.com` (4951 отзывов). Для точности, в позитивный корпус вошли только тексты *Преимуществ*, с оценкой 5, а в негативный корпус только *Недостатки*, с оценкой 1 или 2. *Преимущества* и *Недостатки* – это специальные фрагменты отзывов, в которых поль-

[2] `http://mpqa.cs.pitt.edu/`

Таблица 2 — Статистика размеров словарей, сгенерированных вручную

| Словарь | Размер словаря | |
|---|---|---|
| | для русского языка | для английского языка |
| Action | 7863 | 7886 |
| ProblemWord | 942 | 190 |
| NotProblemWord | 69 | 45 |
| NegativeWord | 1476 | 4169 |
| PositiveWord | 1078 | 2323 |
| AddWord | 30 | 15 |
| ImperativePhrases | 26 | 6 |
| Слова-отрицания | 14 | 22 |

зователь перечисляет только то, что он оценил позитивно в продукте, либо, соответственно, только то, что не понравилось. Затем были выбраны высокочастотные глаголы, существительные, прилагательные и наречия и удалены целевые объекты отзывов (например: *машина*, *двигатель*, *цена*). Полученный список был расширен путем добавления всех синонимов лексических единиц Викисловаря, соответствующих словам из списка.

Для выделения отрицаний действий и индикаторов создан словарь грамматических конструкций, таких как *нет*, *никогда*, *никакой*, *нисколечко*, *невозможно*. Для идентификации ситуаций, выходящих за рамки допустимого, по мнению пользователя, создан словарь дополнительных слов AddWord, таких как *чересчур*, *излишне*, *чрезмерно*, *немного*, *слишком*. Для идентификации ситуаций, в которых пользователь вследствие неудовлетворенности желает исправить продукт, создан словарь ImperativePhrases, содержащий глаголы в повелительной форме (например: *сделайте*, *откорректируйте*, *почините*) и фразы, указывающие на запрос пользователя к изменению, такие как *could have*, *wish*. Статистика словарей содержится в Таблице 2, полные словари описаны в приложении А. Таким образом, в данной диссертационной работе созданы (i) словарь оценочных слов для русского языка, (ii) словарь глаголов действий и (iii) англоязычный и русскоязычный словарь проблемных индикаторов для широкой области продуктов и не зависящие от определенной предметной области. Полезность данных словарей показана в задаче классификации для методов, описанных в следующих разделах.

## 2.4 Предложенный подход и методы классификации

Для достижения целей задачи в диссертационной работе предложен подход, основанный на знаниях. Данный подход предполагает использование дополнительных экспертных ресурсов в виде словарей индикативных слов и выражений, составленных вручную или автоматически, и написание правил, которые отражают структуру фрагментов текста относительно рассматриваемой информации. Ранние работы по анализу мнений с целью определения проблемных высказываний подтверждают эффективность подходов, основанных на знаниях [16; 17; 122]. Преимуществом этого подхода является способность обеспечить эффективность классификации текстов для широкой области продуктов: без серьезных потерь качества работы для различных предметных областей. В рамках подхода предложены два метода классификации проблемных высказываний:

1. Метод, основанный на условиях вхождения лексических единиц из словарей;
2. Метод, учитывающий грамматическую структуру сложных предложений относительно союзов.

Целью задачи классификации является определение класса предложений контрольной выборки. В данной работе выделяется класс проблемных высказываний (problem класс) и класс высказываний без проблем (no-problem класс). Согласно формулировке задачи, $r(s_{ij}) = 1$ для $s_{ij}, j \in \{1, \ldots, |d_i|\}, i \in \{1, \ldots, |D|\}$, принадлежащих к классу проблемных высказываний, в противном случае, $r(s_{ij}) = 0$.

### 2.4.1 Метод, проверяющий последовательность условий

Алгоритм извлечения проблемных фраз состоит из последовательности нескольких условий, учитывающие вхождения слов из созданных словарей в предложение $s_{ij}$:

1. Если найдено вхождение глагола из словаря Action вместе со связанным отрицанием, то алгоритм выделяет все предложение $s_{ij}$ как проблемное и присваивает $r(s_{ij}) = 1$.
2. Если найдено вхождение проблемного индикатора из словаря ProblemWord без связанного отрицания или найдено вхождение слова из словаря NotProblemWord вместе со связанным отрицанием, то алгоритм выделяет все предложение $s_{ij}$ как проблемное и присваивает $r(s_{ij}) = 1$.
3. Если найдено словосочетание, где первым словом является отрицание *нет*, *никакой*, *отсутствие*, *отсутствовать*, *нету*, вторым словом является существительное, а само словосочетание не выделено на шаге 1 и 2 и не найдено вхождение слова из словаря NotProblemWord вместе со связанным отрицанием, то алгоритм выделяет все предложение $s_{ij}$ как проблемное и присваивает $r(s_{ij}) = 1$.
4. Если найдено вхождение лексической единицы из словаря ImperativePhrases, то алгоритм выделяет все предложение $s_{ij}$ как проблемное и присваивает $r(s_{ij}) = 1$.
5. Если на шагах 1 – 4 предложение $s_{ij}$ не классифицировано как проблемное, оно выделяется как предложение, не содержащее проблемную ситуацию, $r(s_{ij}) = 0$.

Связанным отрицанием относительно глагола из словаря Action считается отрицание из словаря Negation, если в предложении существует вхождение отрицания на расстоянии не больше, чем $N_{act}$ слов слева от глагола. Связанным отрицанием относительно русскоязычного индикатора из словарей ProblemWord и NotProblemWord считается отрицание из словаря Negation, если в предложении существует вхождение отрицания на расстоянии не больше, чем $N_{pw}$ слов слева или справа (перед знаком препинания) от индикатора. Например, во фразе “не могу полноценно пользоваться”, отрицание находится на расстоянии двух слов от глагола *пользоваться*. Дополнительно, связанным отрицанием для проблемного индикатора считается словосочетание глагола с отрицанием из словаря Action, если словосочетание следует после индикатора в предложении. После ряда экспериментов, расстояние установлено как $N_{pw} = 2$ и $N_{act} = 2$ для текстов на русском языке; $N_{pw} = 2$ и $N_{act} = 1$ для текстов на английском языке.

### 2.4.2 Метод, основанный на правилах и грамматической структуре предложений

Пользователь может описывать ситуацию использования продукта с помощью нескольких грамматических конструкций с разными типами проблемных фраз в одном сложном предложении. Выделим следующие группы предложений, чтобы показать важность идентификации проблемных фраз относительно соединительных союзов в предложении:

- Первая грамматическая часть предложения (до союза) обладает позитивной тональностью, в то время как вторая часть (после союза) отличается по тональной оценке. Например: "*хорошо* проходит неровности и канавы, хотя и *весьма жесткая*", "салон *теплый* и *большой*, но спартанский", "*приятная* стандартная комплектация, но нет кондиционера".
- Первая грамматическая часть предложения (до соединительного союза) подтверждает дефект или затруднение в использовании, однако вторая часть предложения (после союза) отрицает проблемную или негативную ситуацию. Например, "за такие деньги *хорошее не купить*, поэтому рено логан являться довольно *приемлемым* вариантом", "салон, конечно, простоватый, но это *не самое главное* в машине", "пользуюсь этой жестянкой давно, но явных *недостатков не обнаружил*".
- Все грамматические части предложения содержат схожую информацию о существовании тех или иных проблем в использовании. Например: "тормоз *оставляет желать лучшего*, поэтому хорошо на ней *не гоняю*".
- Первая грамматическая часть предложения содержит условие возникновения проблемы, в то время как вторая часть не указывает на затруднительную ситуацию. Например: "если *не ездить* по грязи и щебню, то *разгоняется* быстро".

Формальное описание предложенного метода представлено в виде контекстно-свободной грамматики (англ. context-free grammar) - 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:

- правила вывода нетерминальных символов, основанные на словарях;
- вспомогательные правила объединения слов и нетерминальных символов;
- 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, $\neg PS$ – не проблемное предложение, $WD$ – множество словосочетаний с отрицанием (шаг 3); $X$ – множество слов с неизвестной информацией о тональности и дефектах (не содержащее слов из $N$, $P$, $IP$, $DP$).

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 DirectPW из словаря ProblemWord, индикаторов с негативной тональностью NegativePW, индикаторов действия VerbPW 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 [117; 129]. Опишем правила классификации в 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. $clause_1 \to AW - IP$, $conj \to$ но; $clause_2 \to \neg DP$; $S \to \neg PS$
2. $clause_1 \to P - IP - DP$, $conj \to$ но; $clause_2 \to A - DP$; $S \to PS$
3. $clause_1 \to DP - IP$, $conj \to$ но; $clause_2 \to \neg DP | \neg IP$; $S \to \neg PS$
4. $clause_1 \to DP - IP$, $conj \to$ но; $clause_2 \to A - DP$; $S \to \neg PS$

5. $clause_1 \to \neg DP$, $conj \to$ но; $clause_2 \to DP|IP - \neg DP$; $S \to PS$
6. $clause_1 \to -DP - N$, $conj \to$ но; $clause_2 \to DW - P$; $S \to PS$
7. $clause_1 \to -DP - N$, $conj \to$ но; $clause_2 \to DW + P$; $S \to \neg PS$
8. $clause_1 \to IP|P - DP$, $conj \to$ но; $clause_2 \to P - DP$; $S \to \neg PS$
9. $clause_1 \to P - IP - DP$, $conj \to$ но; $clause_2 \to P - DP$; $S \to \neg PS$
10. $clause_1 \to P - IP - DP$, $conj \to$ но; $clause_2 \to -P|\neg DP$; $S \to PS$
11. $clause_1 \to X$, $conj \to$ а; $clause_2 \to X$; $S \to PS$
12. $clause_1 \to X$, $conj \to$ а; $clause_2 \to P - DP$; $S \to \neg PS$
13. $clause_1 \to P + \neg DP$, $conj \to$ а; $clause_2 \to NDP$; $S \to \neg PS$
14. $clause_1 \to IP|AW - DDP - VDP$, $conj \to$ а; $clause_2 \to P - DP$; $S \to \neg PS$
15. $clause_1 \to X$, $conj \to$ а; $clause_2 \to P - DP$; $S \to \neg PS$
16. $clause_1 \to IP|DP$, $conj \to$ хотя; $clause_2 \to \neg DP$; $S \to PS$
17. $clause_1 \to IP|DP$, $conj \to$ хотя; $clause_2 \to X$; $S \to PS$
18. $clause_1 \to IP|DP$, $conj \to$ хотя; $clause_2 \to IP - N$; $S \to \neg PS$
19. $clause_1 \to P - IP - DP$, $conj \to$ пока; $clause_2 \to AW - IP$; $S \to \neg PS$
20. $clause_1 \to P-IP-DP$, $conj \to$ если; $clause_2 \to A-IP-DP$; $S \to \neg PS$
21. $conj \to$ если; $clause_1 \to P$, $clause_2 \to -DP$; $S \to \neg PS$
22. $clause_1 \to P$; $conj \to$ если; $clause_2 \to IP|DP$; $S \to \neg PS$
23. $clause_1 \to IP-DP$, $conj \to$ поэтому; $clause_2 \to A-IP-DP$; $S \to \neg PS$
24. $clause_1 \to P$, $conj \to$ теперь; $clause_2 \to \neg DP$; $S \to \neg PS$
25. $clause_1 \to P - DP$, $conj \to$ правда; $clause_2 \to N - \neg DP$; $S \to PS$
26. $clause_1 \to P - DP$, $conj \to$ правда; $clause_2 \to AW - DP$; $S \to \neg PS$
27. $clause_1 \to -P - N$, $conj \to$ правда; $clause_2 - P|N$; $S \to \neg PS$
28. $S \to P + P + P - IP - DP$; $S \to \neg PS$

Опишем правила классификации для английского языка:

1. $clause_1 \to AW|DW$, $conj \to$ despite; $clause_2 \to P|\neg DP - DP - IP$; $S \to PS$
2. $conj \to$ despite; $clause_1 \to IP|DP$, $clause_2 \to P|\neg DP - DP - IP$; $S \to \neg PS$
3. $clause_1 \to P|\neg DP|A - DP - IP - N$, $conj \to$ because; $clause_2 \to IP - DP$; $S \to \neg PS$
4. $clause_1 \to N - \neg DP$, $conj \to$ but; $clause_2 \to N - \neg DP$; $S \to PS$

5. $clause_1 \rightarrow IP - \neg DP$, $conj \rightarrow$ but; $clause_2 \rightarrow AW + P - DP - IP$; $S \rightarrow \neg PS$
6. $clause_1 \rightarrow DW|NDP - DDP - VDP$, $conj \rightarrow$ but; $clause_2 \rightarrow P|\neg DP - DP - IP$; $S \rightarrow \neg PS$
7. $clause_1 \rightarrow \neg DP$, $conj \rightarrow$ but; $clause_2 \rightarrow AW|NDP$; $S \rightarrow PS$
8. $clause_1 \rightarrow \neg DP$, $conj \rightarrow$ but; $clause_2 \rightarrow IP - DP$; $S \rightarrow \neg PS$
9. $clause_1 \rightarrow P - DP - IP$, $conj \rightarrow$ but; $clause_2 \rightarrow AW$; $S \rightarrow PS$
10. $clause_1 \rightarrow P - DP - IP$, $conj \rightarrow$ but; $clause_2 \rightarrow -P - \neg DP$; $S \rightarrow PS$
11. $clause_1 \rightarrow P - DP - IP$, $conj \rightarrow$ but; $clause_2 \rightarrow \neg DP$; $S \rightarrow \neg PS$

Оператор "$-$" перед недетерминированным символом обозначает отсутствие данного символа во фрагменте $clause_1$ или $clause_2$. Оператор "$+$" между двумя символами обозначение присутствия обоих во фрагменте $clause_1$ или $clause_2$. Оператор "$|$" между символами обозначает бинарную операцию *или*. Алгоритм предложенного метода состоит из нескольких шагов для предложения $s_{ij}$ :

1. Применение правил вывода нетерминальных символов, основанные на словарях;
2. Применение объединения слов и нетерминальных символов;
3. Применение правил классификации;
4. Если предложение $s_{ij}$ было идентифицировано как $\neg PS$, то алгоритм выделяет предложение как не проблемное (no-problem класс) и присваивает $r(s_{ij}) = 0$. Если предложение ($s_{ij}$ идентифицировано как PS, то алгоритм выделяет предложение как проблемное (problem класс) и присваивает $r(s_{ij}) = 1$. В противном случае предложение классифицируется согласно результатам метода, проверяющего ряд условий.

Ниже приведены примеры классификации. В результате выполнения правил выражению "мелкие *недочёты* конечно имеются, но в целом приложение рабочее *пользоваться* можно" присваивается символ $\neg PS$ (правило 4), выражению "*не хватает* функции просмотра полных выписок, а в целом *удобно* и *функционально*" присваивается символ $\neg PS$ (правило 14), выражению "я *не могу перевести* деньги со счета на счет, хотя раньше с этим *не было проблем*" присваивается $PS$ (правило 16).

## 2.5 Экспериментальное исследование

В данном разделе проводится экспериментальная оценка предложенных методов и сравнение с существующими в литературе аналогами на нескольких наборах данных для русского и английского языка. Основная цель экспериментов – показать эффективность предложенных методов в задаче автоматической классификации проблемных высказываний с продуктами на уровне предложений отзывов и провести сравнительный анализ лексических единиц из словарей проблемных индикаторов. В качестве базовых методов используются существующие методы, основанные на правилах для английского языка, и методы машинного обучения, часто применимые в задачах анализа мнений: метод максимальной энтропии, классификатор на основе деревьев решений, метод опорных векторов. Также показано, что, используя предметно-независимые словари и правила, предложенные методы показывают стабильность результатов в нескольких предметных областях.

На каждом из наборов данных определенной предметной области проводятся две серии экспериментов. В первой используются все составленные словари и предложенные правила для демонстрации эффективности методов по отношению к существующим методам на корпусах текстовых данных. Во второй серии экспериментов рассматривается эффективность отдельных условий, использующие разные типы словарей для задачи классификации.

### 2.5.1 Наборы данных и архитектура программного компонента

Программный компонент, идентифицирующий в тексте лексические единицы из словарей и применяющий правила, реализован как часть системы обработки текстовых данных Apache UIMA Ruta (версия 2.3.1) на языке java и выложен в открытый доступ[3]. Написанные правила не являются вложенными. Результатом работы UIMA-аннотатора, реализующего описанные правила, являются аннотации требуемых типов (например, для найденных

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

в тексте вхождений из словаря ProblemWord формируется аннотация типа ru.kpfu.itis.cll.ProblemWord). Данный комплекс выложен в открытый доступ16

Для экспериментов были собраны и размечены отзывы о продуктах на русском и английском языках. С целью анализа отзывов различной специфики, составлена коллекция из текстов следующих предметных областей:

- отзывы об автомобилях на русском языке, опубликованные в рамках дорожки анализа тональности соревнования SentiRuEval-2015 [99];
- короткие сообщения о мобильных приложениях банков (длиной более 60 символов) категории «Финансы» на русском языке в онлайн-магазине приложений Google Play[4];
- сообщения пользователей на английском языке[5] об электронных продуктах с форума поддержки продуктов компании Hewlett-Packard;
- отзывы пользователей на английском языке[6] компании Amazon в трех следующих предметных областях: автомобили, детские товары и инструменты для дома.

Создана контрольная выборка из выше описанной коллекции путем аннотации вручную на основе следующей схемы: для предложений из отзывов, случайно выбранных из текстовой коллекции для каждой предметной области, разметчику предлагалось выбрать один из трех отметок: “явная проблема”, “неявная проблема”, “проблема отрицается”, “спам, реклама”. Отметка “явная проблема” указывает, что продукт/сервис имеет технические неполадки в процессе использования. Отметка “неявная проблема” подразумевает проблему: продукт функционирует согласно техническим спецификациям, но не ожиданию потребителя, вызывая неудовлетворенность продукцией. Если разметчик указывал предложение как “спам”, то оно исключалось из выборки. Учитывая сложность формализации понятий “проблема” и “упоминание проблемы”, предложения, отнесенные к первым двум отметкам, объединялись в одно множество: высказываний, указывающих на неполадки с продуктом. Каждая выборка отзывов пользователей была проаннотирована 3 экспертами. Отметка с максимальным числом голосов была выбрана как результирующий класс для задачи классификации. Для проверки качества разметки используется коэффициент Коэна (Cohen’s kappa, k), вычисленный после объединения меток. Коэффици-

[4] https://play.google.com/store/apps/category/FINANCE

[5] http://www8.hp.com/ru/ru/supportforums.html

[6] https://snap.stanford.edu/data/web-Amazon.html

Таблица 3 — Статистика размеченной коллекции пользовательских текстов

| Предметная область | Количество высказываний | | | | средняя длина в словах |
|---|---|---|---|---|---|
| | всего | | с союзами для правил | | |
| | problem | no-problem | problem | no-problem | |
| Электроника (англ. с форума HP) | 603 | 747 | 360 | 226 | 21 |
| Детские товары (англ. с Amazon) | 780 | 399 | 175 | 95 | 21 |
| Инструменты (англ. с Amazon) | 611 | 239 | 121 | 50 | 20 |
| Машины (англ. с Amazon) | 828 | 171 | 179 | 28 | 20 |
| Машины (рус. с SentiRuEval-2015) | 534 | 2285 | 157 | 422 | 12 |
| Приложения (рус. с Google play) | 1653 | 1216 | 566 | 182 | 13 |

ент Коэна равен 0.59 для текстов на английском языке, 0.61 для текстов об автомобилях для русского языка и 0.57 для текстов о мобильных приложениях, что указывает на средний коэффициент согласия. Статистика полученных коллекций приведена в Таблице 3. Как видно из таблицы, 26% предложений на английском языке содержат союзы *but, because* и предлог *despite*; 21% предложений на русском языке содержат слова *но, а, хотя, пока, если, поэтому, теперь, правда*. 67% предложений на английском языке и 40% предложений на русском языке содержат упоминание проблемной ситуации. Пользователи описывают проблему в отзывах о машинах только в 27% предложений на русском языке в то время, как 44.6% и 57.6% размеченных комментариев с Google Play и форума Hewlett-Packard содержат полезную информацию для разработчика или компании. Это подтверждает факт, что пользователи пишут комментарии о продукте в магазине приложений или на официальном форуме с целью получения отклика от представителей компании. Морфологическая обработка текста осуществлялась с помощью библиотеки Mystem[7] для русского языка: на этапе предварительной обработки текстов была выполнена лемматизация всех слов на русском языке.

[7] `https://tech.yandex.ru/mystem`

### 2.5.2 Критерии качества

В работе используются стандартные метрики качества систем анализа текстов на естественном языке: достоверность (англ. accuracy, Acc.), точность (англ. precision, P), полнота (англ. recall, R) и F-мера (англ. $F_1$-measure, F). В качестве положительного класса классификации рассматривается класс проблемных высказываний. Пусть $TP$, $FP$, $TN$, $FN$ — число истинно-положительных, ложноположительных, истинно-отрицательных и ложноотрицательных документов для положительного класса задачи, соответственно. Полнота вычисляется как отношение истинно-положительных документов к общему количеству известных положительных документов по формуле:

$$R = \frac{TP}{TP + FN} \tag{2.1}$$

Точность вычисляется как отношение истинно-положительных релевантных документов к общему количеству определенных системой положительных документов:

$$P = \frac{TP}{TP + FP} \tag{2.2}$$

Достоверность вычисляется как отношение истинно найденных документов к общему числу документов:

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

Полнота характеризует способность системы находить нужные пользователю документы, точность характеризует способность системы идентифицировать только положительные документы среди извлеченных документов. F-мера вычисляется как среднее гармоническое полноты и точности:

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{2.4}$$

Дополнительно, используется метод усреднения результатов бинарной классификации, называемый *макроусреднением* (англ. macro-averaged). Данный способ характерен для оценки задач анализа мнений пользователей, в которых важен результат в среднем по коллекции, независимо от важности извле-

каемых классов. Метрика качества, подсчитанная макроусреднением, является средним арифметическим значений, подсчитанных для каждого класса (например, положительного и отрицательного). F-мера, полученная макроусреднением (далее макро F-мера) представляет собой единую метрику, объединяющую метрики полноты и точности для двух классов, поэтому в данной главе макро F-мера используется как основной критерий качества.

### 2.5.3 Эксперименты и обсуждение

Для сравнения результатов классификации были выбраны следующие методы машинного обучения, популярные в задачах анализа мнений: метод максимальной энтропии (MaxEnt), классификатор на основе деревьев решений (DecisionTrees, DTs), метод опорных векторов (SVM). Метод опорных векторов и метод максимальной энтропи показывают наилучшие результаты для задачи классификации текстов по тональности относительно аспектов [109; 130]. Для обучения классификаторов были использована библиотека Weka[8] и метод «мешка слов» (англ. bag of words, BoW), состоящий из отдельных слов (англ. unigrams) и словосочетаний (англ. bigrams). Параметры классификаторов выставлены в соответствии с настройками по умолчанию библиотеки и получены с помощью кросс-валидации на 10 блоках (англ. 10-folds validation).

В качестве базовых моделей выбраны следующие классификаторы:

– классификатор на основе методов MaxEnt, SVM и DecisionTrees, обученные на словах и словосочетаниях;
– наивный байесовский классификатор (NaiveBayes), описанный в работе [124];
– классификаторы NRC-Canada, GU-MLT-LT и KLUE, показавшие наилучшие результаты классификации тональности коротких сообщений [22; 131—133].

В работе [124] описан наивный байесовский классификатор (NaiveBayes), показавший наилучшие результаты по сравнению с MaxEnt и DecisionTrees на

[8] http://www.cs.waikato.ac.nz/ml/weka/

корпусе текстов о мобильных приложений на английском языке. Для классификатора были использованы следующие признаки:

– слова и словосочетания;
– рейтинг продукта в отзыве;
– количество глаголов в различном времени;
– количество позитивных слов, количество негативных слов.

Для английского языка глаголы в прошедшем, настоящем и будущем временах были определены с помощью системы Stanford CoreNLP[9]. Для русского языка глаголы в прошедшем и непрошедшем временах были определены с помощью Mystem.

Детальные описания NRC-Canada, GU-MLT-LT и KLUE изложены в работах [22; 131—133], далее представлено краткое описание признаков классификации[10]. В работах [22] описывается классификатор NRC-Canada на основе линейного SVM для английского языка, использующий следующий набор признаков:

– многословные выражения, содержащие от 1 до 4 слов; выражения, содержащие от 6 до 4 символов; количество повторяющихся символов;
– число слов, набранных полностью заглавными буквами; числа вхождений слов с одинаковыми частями речи (для каждой части речи);
– признаки, использующие словарь позитивных и негативных слов: количество позитивных и негативных слов в тексте; количество слов со значением веса принадлежности к классу тональности, большим нуля; количество слов с весом, меньшим нуля; максимальный тональный вес слов; минимальный тональный вес слов; сумма тональных весов; вес последнего слова в тексте;
– число подряд идущих знаков пунктуации;
– число позитивных смайликов, число негативных смайликов, является ли последний токен текста позитивным смайликов, является ли негативным;
– признаки принадлежности к кластерам, основанные на кластеризации Брауна;

[9] `http://nlp.stanford.edu/`

[10] Реализация методов заимствована из работы [133]; `https://github.com/webis-de/ECIR-2015-and-SEMEVAL-2015`

– признаки, учитывающие отрицания слов, рассматривающиеся между частицей отрицания и последующим знаком пунктуации: количество слов с отрицанием

Классификатор GU-MLT-LT, основанный на методе стохастического градиентного спуска, использует следующий набор признаков [131]:

– слова, использующие статистическую меру TF-IDF вместо частотности слов; основы слов, полученные с помощью стеммера Портера;
– признаки принадлежности к кластерам, схожие с NRC-Canada;
– признаки, использующие словарь: сумма весов позитивных и негативных слов;
– признаки, учитывающие отрицания слов, схожие с NRC-Canada.

Классификатор KLUE, основанный на методе максимальной энтропии, использует следующий набор признаков [132]:

– слова и словосочетания;
– длина текста в токенах;
– признаки, использующие словарь: количество позитивных слов, количество негативных слов, среднее арифметическое весов слов в тексте;
– число позитивных смайликов, число негативных смайликов;
– отрицания слов, рассматривающиеся на расстоянии до 3 слов.

В качестве словаря негативных и позитивных слов использовались построенные словари PositiveWord и NegativeWord для русского языка и MPQA Subjectivity Lexicon для английского языка. Для текстов на английском языке классификаторы KLUE и NRC-Canada (далее NRC) используют MPQA, для GU-MLT-LT (далее GU) – лексикон SentiWordNet[11], указанный в [132]. Вес оценочных слов, отнесенных к позитивном и негативному классам, равен 1.0 и -1.0, соответственно.

Дополнительно, для классификаторов NRC, GU и KLUE были использованы признаки, основанные на словарях ProblemWord, NotProblemWord, Action, ImperativePhrases, AddWord. Для каждого вида словаря в зависимости от присутствия отрицания были подсчитаны следующие признаки, схожие с описанными для классификатора NRC признаками: количество слов из словаря с положительным и отрицательными весами в тексте, максимальный вес слов, сумма положительных весов, сумма отрицательных весов, вес последнего слова в

[11] http://sentiwordnet.isti.cnr.it

Таблица 4 — 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 |
| DTs 1gr. | .807 | .476 | .206 | .288 | .656 | .576 | .587 | .746 | .778 | .782 | .780 | .739 | .739 | .739 |
| MaxEnt 1gr. | .697 | .299 | .446 | .358 | .576 | .600 | .579 | .699 | .745 | .725 | .735 | .692 | .694 | .693 |
| SVM 1gr. | .817 | .521 | .418 | .464 | .695 | .663 | .676 | .803 | .826 | .833 | .829 | .797 | .798 | .798 |
| DTs 2-gr. | .810 | .493 | .212 | .296 | .665 | .580 | .593 | .747 | .791 | .764 | .777 | .742 | .744 | .743 |
| 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 |

тексте. Для лексических единиц из словаря ProblemWord без связанного отрицания и со связанным отрицанием веса равны 1.0 и -1.0, соответственно. Для слов из словарей Action и NotProblemWord со связанным отрицанием и без отрицания веса равны 1.0 и -1.0, соответственно. Лексическим единицам из словаря ImperativePhrases присвоен вес, равный 1.0. Классификаторы с расширенным набором признаков обозначены с “Dicts”.

Результаты классификации представлены в Таблицах 4, 5, 6. Метод, основанный на ряде условий, обозначен как **DbA** (англ. dictionary-based approach); метод, основанный на анализе сложных предложений, обозначен как **CbA** (англ. clause-based approach). Классификаторы на основе “мешка слов” обозначены с “1gr.”; классификаторы, обученные на словах и словосочетаниях, обозначены “2gr.”.

Результаты классификации позволяют сделать ряд следующих наблюдений. Во-первых, в зависимости от предметной области, среди базовых моделей (DecisionTrees, MaxEnt, SVM) наилучшие результаты по макро F-мере показывают различные классификаторы: SVM для текстов об электронике, детских товарах, приложениях, инструментах и машинах (рус.); MaxEnt для текстов о машинах (англ.). В рамках моделей, показывающих наилучшие результаты

Таблица 5 — 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 |
| DTs 1gr. | .804 | .829 | .961 | .890 | .504 | .501 | .478 | .646 | .621 | .537 | .576 | .641 | .635 | .635 |
| MaxEnt 1gr. | .706 | .848 | .786 | .816 | .540 | .551 | .542 | .585 | .537 | .541 | .539 | .581 | .581 | .581 |
| SVM 1gr. | .747 | .852 | .841 | .846 | .563 | .566 | .564 | .713 | .687 | .662 | .674 | .710 | .708 | .709 |
| DTs 2-gr. | .818 | .833 | .976 | .899 | .571 | .514 | .494 | .652 | .624 | .564 | .592 | .647 | .643 | .644 |
| 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 6 — 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 |
| DTs 1gr. | .631 | .746 | .736 | .741 | .547 | .548 | .547 | .601 | .693 | .549 | .714 | .547 | .703 | .548 |
| MaxEnt 1gr. | .632 | .722 | .794 | .756 | .506 | .505 | .502 | .577 | .715 | .559 | .600 | .565 | .652 | .555 |
| SVM 1gr. | .624 | .767 | .684 | .723 | .567 | .576 | .567 | .656 | .741 | .616 | .740 | .616 | **.740** | .616 |
| DTs 2-gr. | .635 | .750 | .740 | .745 | .552 | .553 | .553 | .607 | .694 | .552 | .728 | .549 | .710 | .550 |
| 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 |

в задаче анализа тональности (NRC, GU, KLUE), наилучшие результаты показывают: NRC для текстов о машинах (рус. и англ.), электронике, инстру-

ментах; GU и KLUE для текстов о приложениях (рус.); KLUE для текстов о детских товарах. Это подтверждает, что методы машинного обучения не являются универсальными для текстов всех предметных областей, что согласуется с выводами работы [56]. Во-вторых, предложенный в работе метод **CbA** показал наилучшие значения F-меры, полученной макроусреднением для бинарной задачи классификации. Классификаторы, показывающие наилучшие результаты в задаче анализа тональности (NRC, GU, KLUE), достигают меньшие значения макро F-меры по сравнению с **CbA** в 5 из 6 предметных областях, что подтверждает необходимость создания словарей проблемных индикаторов. В-третьих, использование признаков Dicts улучшает результаты NRC, GU, KLUE по макро F-мере в 15 из 18 экспериментах (исключ. значения NRC+Dicts на текстах о машинах (англ.) и электронике, KLUE+Dicts на текстах об инструментах). Методы показывают наилучшие значения классификации среди всех методов машинного обучения: GU+Dicts для текстов о приложениях (рус.), электронике; NRC+Dicts для текстов о детских товарах, инструментах, машинах (рус. и англ.). В-четвертых, значения F-меры, полученной макроусреднением, в результате классификации с помощью **CbA** отзывов о высокотехнологичных и многофункциональных продуктах (электроника, приложения) оказались на 11% выше, чем значения F-меры для текстов о механических продуктах (машины). Это подтверждает факт, что тенденция объединения разной функциональности в продукте приводит к увеличению общих (менее индивидуальных) претензий пользователей, связанных с удобством использования.

Для проверки статистической значимости результатов классификации на нескольких предметных областях использован непараметрический статистический критерий знаковых рангов Вилксоксона [134]. Нулевая гипотеза заключается в том, что два метода классифицируют документы одинаково. Две и четыре предметные области были объединены в русскоязычный и англоязычный корпуса, соответственно. Результаты классификации были проверены для двух выборок на разных языках. Разница попарного сравнения результатов классификации предложенным методом **CbA** и результатами каждого из классификаторов NRC, GU, KLUE статистически значима ($p<0.01$). 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 методом машин-

ного обучения. Разница между результатами классификации методом, который учитывает структуру предложения, и каждым из классификаторов (SVM 2gr., DecisionTrees 2gr., MaxEnt 2gr., NaiveBayes) статистически значима в большей степени ($p<0.001$), согласно тесту Вилксоксона.

Дополнительно для текстов на русском языке, в качестве систем для сравнения был адаптирован метод, основанный на правилах для английского языка без применения лексикона (Patterns), описанный в статье [18]. Следующие шаблоны фраз были использованы для русского языка: NEG (позволить|разрешать|допускать|давать|разрешить); NEG (выбор|возможность); NEG (нравиться|любить); NEG (мочь|уметь); (баг|глюк|сбой|ошибка|недочёт|неполадка). Метод Patterns, состоящий из 5 шаблонов на английском языке, которые были переведены на русский язык, показал среднее значение точности для автомобилей и приложений (.476 и .875, соответственно) и наименьшее значение полноты ($< .25$). Это связано с тем, что тексты на русском языке характеризуются относительно свободным порядком слов, что ухудшает результаты классификации методами, которые разработаны для английского языка и основаны на порядке слов в правилах для предложений.

Для проверки эффективности метода, основанного на анализе сложных предложений относительно союзов, был собран уменьшенный корпус с союзами для правил, размерность которого указана в Таблице 3. Таблицы 7 и 8 содержат результаты классификации на уменьшенном корпусе.

Методы машинного обучения показали ухудшение значений F-меры, полученной относительно класса проблемных высказываний и макроусреднением, относительно Таблиц 4, 5, 6 вследствие трех причин: (i) классификатор обучился на коллекции значительно меньшего размера; (ii) сложные предложения обладают большим количеством слов, что увеличивает сложность классификации; (iii) для методов машинного обучения существует вероятность возникновения проблемы переобучения (англ. overfitting), где все высказывания будут отнесены к классу проблемных (напр., GU+Dicts для машин (англ.)). Предложенный метод **CbA**, основанные на обучении без учителя и использующий словари и правила, показал наилучшие результаты, схожие с результатами по всей контрольной выборке. Это показывает эффективность предложенного ал-

Таблица 7 — Результаты классификации предложений отзывов пользователей относительно класса проблемных высказываний на уменьшенном корпусе

| Метод | Электроника (англ.) | | | | Машины (рус.) | | | | Приложения (рус.) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ |
| DTs 1gr. | .589 | .671 | .655 | .663 | .643 | .198 | .100 | .133 | .698 | .762 | .763 | .762 |
| MaxEnt 1gr. | .641 | .731 | .660 | .694 | .634 | .358 | .425 | .389 | .719 | .790 | .759 | .774 |
| SVM 1gr. | .647 | .711 | .721 | .716 | .651 | .351 | .325 | .338 | .757 | .800 | .822 | .811 |
| DTs 2gr. | .613 | .697 | .660 | .678 | .653 | .213 | .100 | .136 | .701 | .758 | .776 | .767 |
| MaxEnt 2gr. | .612 | .691 | .669 | .680 | .605 | .359 | .563 | .438 | .681 | .755 | .736 | .745 |
| SVM 2gr. | .656 | .725 | .713 | .719 | .677 | .399 | .356 | .376 | .758 | .813 | .803 | .808 |
| NaiveBayes | .656 | **.741** | .680 | .709 | .675 | .405 | .400 | .420 | .766 | .825 | .801 | .813 |
| NRC+Dicts | .691 | .728 | .795 | .760 | .727 | .495 | .318 | .388 | .775 | .801 | .862 | .830 |
| GU+Dicts | .693 | .722 | **.814** | .766 | **.743** | **.548** | .293 | .382 | .797 | .807 | **.895** | .849 |
| KLUE+Dicts | .686 | .722 | .798 | .758 | .737 | .526 | .318 | .397 | .800 | .815 | .889 | .851 |
| DbA | .666 | .701 | .798 | .746 | .717 | .484 | .684 | .566 | .763 | .811 | .820 | .815 |
| CbA | **.700** | .730 | **.814** | **.770** | .721 | .491 | **.713** | **.582** | **.814** | **.839** | .876 | **.857** |

Таблица 8 — Результаты классификации предложений отзывов пользователей относительно класса проблемных высказываний на уменьшенном корпусе

| Метод | Инструменты | | | | Детские товары | | | | Машины (англ.) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ |
| DTs 1gr. | .614 | .708 | .780 | .742 | .599 | .681 | .717 | .528 | .848 | .865 | .976 | .917 |
| MaxEnt 1gr. | .602 | **.771** | .627 | .692 | .618 | **.763** | .595 | .551 | .690 | .892 | .729 | .803 |
| SVM 1gr. | .669 | .723 | .864 | .788 | .581 | .667 | .705 | .621 | .853 | .881 | .959 | .918 |
| DTs 2gr. | .602 | .713 | .737 | .725 | .566 | .646 | .728 | .548 | .832 | .862 | .959 | .908 |
| MaxEnt 2gr. | .620 | .748 | .703 | .725 | .592 | .742 | .566 | .496 | .741 | .861 | .835 | .848 |
| SVM 2gr. | .693 | .711 | **.958** | .816 | .599 | .679 | .723 | .597 | .838 | .879 | .941 | .909 |
| NaiveBayes | .627 | .726 | .763 | .744 | .599 | .685 | .705 | .555 | .797 | .887 | .876 | .882 |
| NRC+Dicts | .707 | .748 | .884 | .811 | .618 | .675 | .794 | .730 | .850 | .874 | .966 | .918 |
| GU+Dicts | .655 | .694 | .917 | .790 | .588 | .647 | .806 | .718 | .859 | .864 | **.994** | .925 |
| KLUE+Dicts | .661 | .709 | .884 | .787 | .581 | .646 | .783 | .708 | **.865** | .868 | **.994** | **.927** |
| DbA | .661 | .733 | .818 | .773 | .641 | .695 | .794 | .741 | .744 | .870 | .827 | .848 |
| CbA | **.737** | **.771** | .896 | **.828** | **.656** | .701 | **.818** | **.755** | .806 | **.906** | .866 | .885 |

горитма и необходимость семантического анализа для обнаружения проблем на русском и английских языках.

### Анализ определения отрицания индикативных конструкций

Для анализа извлечения связанного отрицания из текста был проведен ряд экспериментов с определением значений $N_{act}$ и $N_{pw}$ для текстов на русском языке, которые характеризуются относительно свободным порядком слов по сравнению с текстами на английском языке. Таблица 9 содержит результаты экспериментов, полученных с помощью метода **DbA**. Результаты подсчитаны для $N_{act} = N_{pw}$, что объяснимо тем, что словарь ProblemWord и словарь Action содержат лексические единицы одинаковых частей речи. Наилучшие значения полноты и точности показывает метод, определяющий связанное отрицание на расстоянии не больше, чем два слова слева от глагола из словаря Action и два слова слева или справа от индикатора из ProblemWord или NotProblemWord. Дополнительно, мы проанализировали необходимость поиска связанного отрицания с помощью грамматики зависимостей на основе прямых и косвенных связей, описанных в 3.2.1. Как видно из результатов экспериментов, применение отношений зависимостей показывает сравнимые результаты с поиском отрицания в контексте высказывания около слова.

Таблица 9 — Результаты классификации предложений отзывов с различным определением связанных отрицаний

| Метод DbA | Машины (рус.) | | | | Приложения (рус.) | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ |
| $N_{act} = 0, N_{pw} = 0$ | .821 | **.531** | .528 | .529 | .799 | **.831** | .818 | .825 |
| $N_{act} = 1, N_{pw} = 1$ | **.814** | .508 | .632 | .563 | **.806** | .829 | .833 | .832 |
| $N_{act} = 2, N_{pw} = 2$ | **.814** | .507 | .636 | **.564** | **.806** | .829 | **.837** | **.833** |
| $N_{act} = 3, N_{pw} = 3$ | .800 | .500 | **.640** | .562 | .805 | .826 | **.837** | .832 |
| $N_{act} = 4, N_{pw} = 4$ | .809 | .498 | **.640** | .560 | .689 | .812 | .599 | .689 |
| С помощью грамматики зависимостей | .810 | .516 | .617 | .562 | .795 | .820 | .826 | .823 |

**Сравнительный анализ эффективности словарей**

Таблицы 10 и 11 показывают сравнительный анализ эффективности лексических единиц из разных словарей в задаче автоматического извлечения информации о существовании различных проблем. В качестве метода для анализа выбран метод, проверяющий ряд условий относительно лексических единиц. Каждая строка таблицы означает, что результаты классификации подсчитаны без одного из четырех условий, которые описаны в 2.4.1. Исключаемое из метода условие не указано в строке таблицы.

Таблица 10 — Результаты классификации предложений отзывов с различной комбинацией условий

| Методы | Электроника | | | | Машины (рус.) | | | | Приложения (рус.) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ |
| Шаги 2-4 | .728 | **.812** | .509 | .626 | **.836** | **.570** | .544 | .557 | .707 | **.848** | .599 | .702 |
| Шаги 1,3-4 | .655 | .634 | .539 | .582 | .792 | .434 | .312 | .363 | .699 | .811 | .621 | .704 |
| Шаги 1-2, 4 | .750 | .698 | .778 | .735 | .813 | .505 | .606 | .551 | .790 | .831 | .796 | .813 |
| Шаги 1-3 | .748 | .696 | .778 | .734 | .810 | .500 | **.640** | .562 | .788 | .844 | .776 | .808 |
| Метод DbA (шаги 1-4) | **.754** | .693 | **.809** | **.746** | .814 | .507 | .636 | **.564** | **.806** | .829 | **.837** | **.833** |

Таблица 11 — Результаты классификации предложений отзывов с различной комбинацией условий

| Методы | Инструменты | | | | Детские товары | | | | Машины (анг.) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ | Acc. | $P_{pos}$ | $R_{pos}$ | $F_{pos}$ |
| Шаги 2-4 | .555 | .775 | .537 | .634 | .557 | **.764** | .478 | .588 | .522 | .871 | .496 | .632 |
| Шаги 1, 3-4 | .515 | **.788** | .445 | .569 | .542 | .713 | .525 | .598 | .554 | **.880** | .534 | .665 |
| Шаги 1-2, 4 | .668 | .777 | .754 | .765 | .671 | .741 | .772 | .756 | .732 | **.880** | .783 | .829 |
| Шаги 1-3 | .675 | .776 | .769 | .773 | .659 | .734 | .759 | .746 | **.788** | .879 | .780 | .826 |
| Метод DbA (шаги 1-4) | **.702** | .778 | **.818** | **.798** | **.705** | .742 | **.848** | **.791** | .738 | .876 | **.795** | **.834** |

Наиболее эффективными условиями оказались условия вхождения слов из словарей ProblemWord и NotProblemWord для отзывов о инструментах, электронике и машинах (на рус. языке). Условия отрицания глаголов действий для текстов о приложениях, детских товарах и машинах (на анг. языке) оказались немного более эффективными признаками (значение F-меры уменьшилось на

0.01-0.04) по сравнению с проблемными индикаторами. Условия, учитывающие глаголы в повелительной форме и запросы на изменение функционала, оказались наиболее эффективными для коротких текстов о мобильных приложениях (значение F-меры уменьшилось на 0.02), поскольку на сайте магазина приложений пользователи адресуют свои замечания непосредственно компании-разработчику без участия третьих сторон. Таким образом, как видно из таблицы, исключение каждого из правил и словарей ухудшает значение полноты и F-меры, что говорит о необходимости проверки всех описанных условий.

### 2.5.4 Качественный анализ результатов классификации

Для изучения развития предложенных методов и улучшения результатов классификации был проведен анализ ошибочно классифицированных текстов. Случайным образом было выбрано 400 высказываний о мобильных приложениях и 200 высказываний о машинах на русском языке. На Рисунке 2.1 представлены результаты анализа ошибок классификации, где определены следующие типы наиболее частых ошибок:

- ошибка, связанная с определением связанных отрицаний, условий и правил, описанных в разделе 2.4.1;
- недостаточная полнота покрытия текстов с помощью созданных словарей и правил;
- избыточные словари;
- проблемная ситуации возникла при специфических (определённых) условиях;
- запрос/требование функционала или рекомендация к изменению;
- вопрос к разработчикам по использованию;
- орфографические ошибки;
- бессодержательные для разработчиков высказывания или высказывания о другом продукте;
- ошибки, связанные с индивидуальными предпочтениями пользователя.

На основе диаграмм получен ряд наблюдений относительно различий предметных областей. Во-первых, пользователи публикуют большее количество

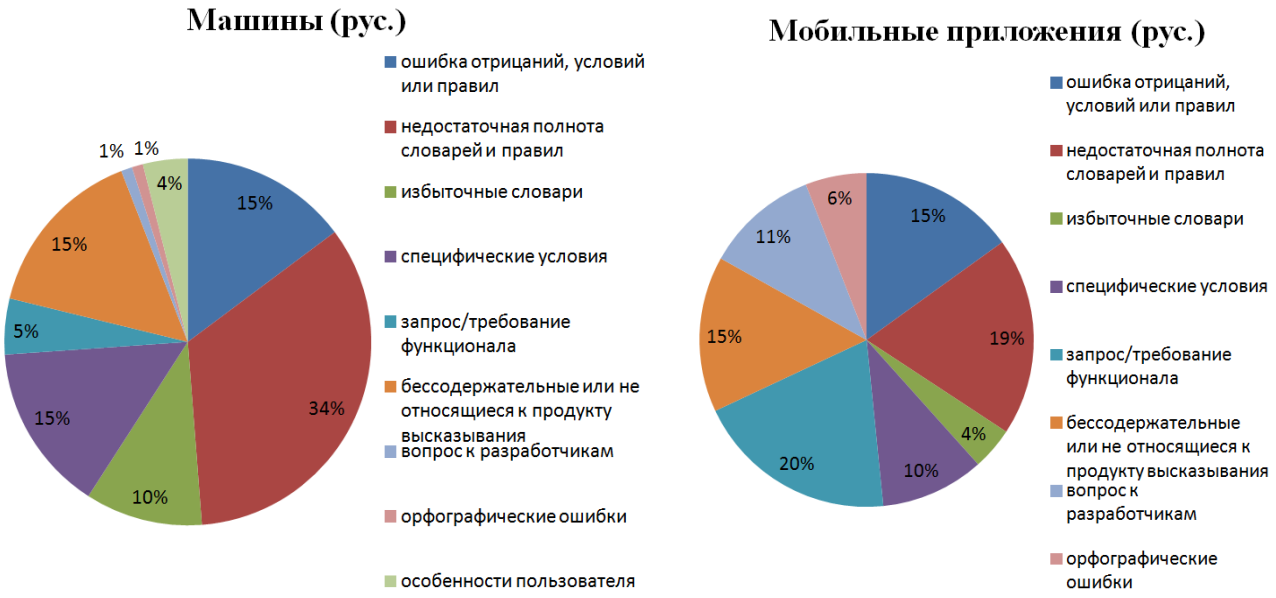



Рисунок 2.1 — Анализ ошибок классификации

требований о новом функционале и вопросов для разъяснения с помощью официальных магазинов приложений (20% и 11% ошибок), чем на сайтах-агрегаторах отзывов (5% и 1%), что подтверждаются количественными результатами в Таблице 10. Во-вторых, потребность в создании предметно-ориентированных словарях проблемных индикаторов для области механических товаров (34% ошибок) выше, чем для области приложений (19%). Это может быть объяснимо различием между количеством специфичных ситуаций, в которых происходят неполадки, для двух областей (15% для машин, 10% для приложений). В-третьих, орфографические ошибки преобладают в коротких текстах по сравнению с текстами отзывов пользователей, что согласуется с результатами, описанными в работе [22; 135]. Под избыточным подразумевается словарь, содержащий лексические единицы, идентификация которых в тексте приводит к ошибкам классификации. Результаты анализа показывают, что большая часть ошибок данного типа возникла по причине существования в тексте вхождений глаголов из словаря Action со связанным отрицанием. В высказывании “До июня отличное приложение для перевода денег с карты на карту, т.к. *не берут* комиссию” проблемная ситуация отсутствует вследствие отрицания факта сбора комиссии, которая не является проблемным индикатором.

Первый тип ошибок, связанный с определением связанных отрицаний, условий и правил, указывает на сложные случаи отрицания проблемной ситуациии пользователем, например “*не без кочек*, а иначе и *быть-то не может!*, “все быстро, качественно, в случае *проблем - техподдержка просто умницы*”,

“еще *не было такого момента* когда машина меня *подводила*”. Данный тип ошибок и проблемные ситуации, возникшие при определённых условиях, требуют глубокого семантического разбора мнения пользователя. В высказываниях “без регистрации ничего не сделать, даже карту офисов-банкоматов не посмотреть” и “автомобиль уже *не выпускается* и *не часто увидишь* его на дорогах” пользователь констатирует фактическую ситуацию, на которую не может повлиять техническая поддержка компании.

Четвертый тип ошибок, связанный с требованиями функционала и рекомендациями к изменению, возникают вследствие правил, основанных на словаре $ImperativePhrases$. В контрольной выборке предложения “*сделайте* в программе возможность увеличения лимита виртуальной карты. Спасибо” и “*сделайте* возможность увеличения лимита на снятие наличных в банкомате” относятся к двум разным классам высказываний, что затрудняет автоматический анализ предложений.

Вопросительные высказывания о приложениях указывают на сложности в использовании продукта, однако не содержат явных индикативных конструкций (“почему при попытке погасить кредит открывается диалог перевода на валютный счёт?”).

Восьмой тип ошибок связан с высказываниями, которые не содержат полезной информации для разработчиков, и с высказываниями о другом продукте. Данный тип ошибок классифицирован разметчиками как класс, не содержащий проблемные высказывания. Высказывание “отрежте руки програмистам которые *угробили* хорошую версию и выложили это Г... ” ошибочно классифицировано, поскольку содержит вхождение слова *угробить* из словаря ProblemWord.

Ошибки, зависящие от индивидуальных предпочтений пользователя, преобладают в отзывах о машинах и вызывают затруднения классификации. К примеру, фразы “подвеска в меру жесткая, прощает езду по неровностям” и “подвеска, о которой много раз упоминалось, довольно жесткая, и нервно относится к любым неровностям на дороге” относятся к разным классам, несмотря на то, что касаются одной составляющей продукта: жесткой подвески.

**Словари предметно-ориентированных проблемных индикаторов**

Второй тип ошибок, указывающий на недостаточную полноту созданных словарей, независящих от предметной области, показывает необходимость создания предметно-ориентированных словарей (англ. domain-specific lexicons). Создано два словаря $DomainPW$ (i) для автомобилей, содержащий такие слова и словосочетания, как *отклеиваться*, *заглохнуть*, *заносить*, *бренчать*, *дерганье* (32 лексических единиц); (ii) для мобильных, содержащий такие слова и словосочетания, как *вылетать*, *жрать батарея*, *вернуть функционал*, *просить исправлять*, *зависание*, *коверкаться*, *перегружать*, *пустой экран* (91 лексических единиц). Таблица 12 содержит результаты классификации с помощью метода **CbA** из Таблицы 4 и **CbA**, где словарь $ProblemWord = ProblemWord \bigcup DomainPW$. Результаты подтверждают эффективность создания предметно-ориентированных словарей, что согласуется с результатами классификации отзывов, описанными в работе [22; 136; 137].

Таблица 12 — 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 |
| CbA (правила) | .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** |

## 2.6 Выводы ко второй главе

В данной главе рассматривается задача идентификации проблемных высказываний на русском и английских языках как задача бинарной классификации предложений из текстов пользователей. Целью задачи является определение класса высказывания для всех единичных элементов документов кон-

трольной выборки $s_{ij} \in d_i, j \in \{1, \ldots, |d_i|\}, i \in \{1, \ldots, |D|\}$. Для достижения целей исследования в статье приводится классификация фраз пользователей, на основе которой построены словари индикативных слов и словосочетаний. Предложен подход, основанный на знаниях, представленных в виде правил и словарей. В работе созданы англоязычный и русскоязычный словари проблемных индикаторов не зависящие от определенной предметной области. В рамках подхода предлагается два метода извлечения фраз: (i) метод извлечения фраз, основанный на ряде условий о вхождении слов из словарей для простых предложений; (ii) метод анализа грамматической структуры сложного предложения относительно союзов. Для проверки эффективности предложенных методов созданы и размечены контрольные выборки сообщений пользователей, собранные с онлайн сайтов о высокотехнологичных, низкотехнологичных и механических продуктах компаний. Качество методов оценивается с помощью стандартных критериев задач классификации текстов: точность, полнота и F-мера, посчитанные относительно класса проблемных высказываний и макроусреднением. Представленные в статье результаты анализируются в сравнении с несколькими методами машинного обучения. Экспериментальное исследование показало, что наилучшие результаты классификации фраз о проблемах в использовании продуктов показывает предложенный метод, основанный на знаниях в виде словарей и анализе структуры предложений. Это подтверждает необходимость семантического анализа предложений для обнаружения проблем с продуктами. Анализ результатов классификации подтвердил, что дальнейшее улучшение результатов возможно за счет создания узкоспециализированных словарей и разработки условий вхождения лексических единиц в зависимости от тематический категории выбранного фрагмента текста.

# Глава 3. Извлечение высказываний, указывающих на проблемные ситуации, относительно предметно-ориентированных целевых объектов мнений

В данной главе рассматривается задача извлечения высказываний из отзыва пользователя о проблемных ситуациях с целевыми объектами, которые зависят от предметной области.

## 3.1 Описание задачи

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 класса высказывания 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. Подобная информация должна быть учтена системами технической поддержки клиентов компании, позволяя выявить проблемные компоненты продукции. Например, во фразах "*багажник* тоже *не отличается просторностью*" или "постоянно пишет про *плохое соединение, зайти практически невозможно*" важными для компании целевыми объектами являются багажник автомобиля и соединение телефона. Во фразах "основная *проблема* оказалась *с доставкой* - привезли всё кроме фурнитуры" и "от бортсети автомобиля *видиорегистратор* практически *не заряжается*" основные проблемные ситуации вызваны внешними для компании факторами: доставкой и дополнительной сторонней электроникой. Рассматриваемая в данной главе задача состоит из трех подзадач:

1. Задача идентификации высказываний, указывающих на проблемные ситуации;
2. Задача извлечения предметно-ориентированных целевых объектов, упомянутых в высказывании;

3. Задача идентификации высказываний, указывающих на проблемные ситуации, относительно предметно-ориентированных целевых объектов.

В рамках задачи идентификации высказываний используется подход, основанный на знаниях и описанный в предыдущей главе, и метод классификации, основанный на ряде условий. В рамках задачи извлечения целевых объектов предложен метод, основанный на синтаксических связях между проблемными индикаторами и существительными в предложении и описанный в разделе 3.2.1. Для идентификации предметно-ориентированных целевых объектов предложен метод, использующий меру семантической связанности целевых объектов к предметной области в лингвистическом ресурсе и описанный в разделе 3.2.2.

## 3.2 Метод извлечения предметно-ориентированных целевых объектов

Метод извлечения целевых объектов состоит из 3 компонентов:

- определения множества возможных целевых объектов, используя синтаксические связи между проблемными индикаторами и существительными в предложении;
- идентификации предметно-ориентированных целевых объектов, используя подсчет семантической связанности между объектом и терминами предметной области;
- классификации высказывания к классу высказываний, указывающих на проблемные ситуации, если метод извлекает хотя бы одну комбинацию (проблемный индикатор, предметно-ориентированный целевой объект).

### 3.2.1 Синтаксические зависимости в высказывании

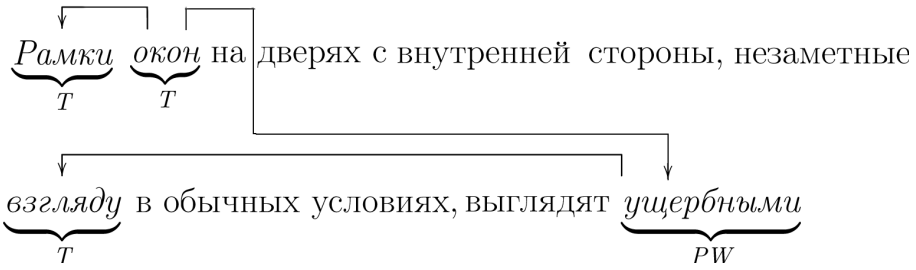
В рамках данного метода предполагается, что высказывание пользователя содержит информацию о составляющей продукта, которого касается проблемная ситуация. Однако некоторая доля терминов в документе может объясняться не определенной предметной областью, а индивидуальными особенностями ситуации (шумом) или общими особенностями всего текста (фоном). Поэтому целесообразным является метод определения необходимой информации, основанный на зависимостях между проблемными индикаторами, характеризующими описанную ситуацию, и связанными словами в предложении. Подобный метод позволит извлечь только синтаксически связанные упоминания о возможных предметно-ориентированных целевых объектах, уменьшив количество шумовых и фоновых терминов в извлеченном множестве объектов.

Для идентификации связей между проблемным индикатором и целевым объектом фразы метод использует два типа зависимости между словами, определенные в работах [102; 138]. *Прямая зависимость* (англ. direct dependency) между словами указывает на то, что одно слово напрямую зависит от другого слова. *Косвенная зависимость* (англ. indirect dependency) указывает, что одно слово зависит от другого слова с помощью слова-посредника. *Слово-посредник* – дополнительное слово, связанное с проблемным индикатором и целевым объектом. Пример 3.1 содержит фразу с прямой зависимостью между выбранными словами. Пример 3.2 содержит косвенную зависимость между словами. Обозначение $PW$ указывает на проблемный индикатор или отрицание действия, $T$ обозначает целевой объект фразы, $S$ указывает на слово-посредник.

**Пример 3.1** $\underbrace{\textit{Приложение}}_{T}$ даже $\underbrace{\textit{не запускается}}_{PW}$, вылетает мгновенно.

**Пример 3.2** Вы когда $\underbrace{\textit{ошибку}}_{PW}$ с $\underbrace{\textit{размещением}}_{S}$ $\underbrace{\textit{вкладов}}_{T}$, исправите?

**Пример 3.3**




### 3.2.2 Расчет семантической связанности целевых объектов к предметной области

Продукт $P_i \in P$ может быть представлен в отзыве как множество составных компонент $T_i = \{t_1, t_2, \ldots, t_k\}$. Предложенный метод использует семантическую информацию о (i) терминах из лингвистического ресурса, являющимися текстовыми представлениями понятий предметной области; (ii) о векторах распределённых представлений слов. Предметно-ориентированные целевые объекты – это связанные с продуктом понятия, существенные в определенной предметной области. Мера семантической связанности понятий предметной области представляет собой числовую оценку степени их смысловой связанности.

**Определение.** 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 целевых объект $t_k$ является предметно-ориентированным.

В данной работе рассматриваются меры семантической связанности (англ. semantic relatedness) и меры семантической близости (англ. semantic similarity) нескольких типов:

1. Мера WUP (анг. Wu & Palmer's measure), основанная на расстоянии (англ. path-based);
2. Мера RES (англ. Resnik's measure), основанная на информационном содержании (англ. information content-based);
3. Мера LESK (англ. Lesk's measure), основанная на определениях (англ. gloss-based);
4. Косинусная мера COS (англ. cosine similarity), использующая вектора распределённых представлений слов.

В качестве меры, основанной на определениях, используется мера LESK (анг. Lesk's measure), рассматривающая связанность терминов понятий по ко-

личеству слов, входящих одновременно в определения этих понятий. В рамках работы используется модификация этого подхода, где для подсчета мер используются термин понятия и единицы тезауруса, связанные с ним несколькими типами отношений [139]. Данная модификация учитывает семантические связи слов на уровне следующих типов отношений: родовидовые отношения (гипонимия/гиперонимия (анг. hyponymy/ hypernymy)), отношения *часть-целое* (меронимия/голонимия (анг. meronymy/holonymy)). Мера LESK вычисляется следующим образом:

$$lesk(d_i,d_j) = \sum \frac{overlap(r(c_i),r(c_j))}{len(r(c_i)) + len(r(c_j))} \quad r \in \{gloss, hypo, hype, mero, holo\}, \tag{3.1}$$

где $overlap(r(c_i),r(c_j))$ обозначает количество общих слов в определениях терминов тезауруса $c_i$ и $c_j$, $r(c_i)$ – связанный элемент с термином $c_i$ с помощью родовидовых отношений ($hypo$,$hype$), отношений *часть-целое* ($mero$,$holo$).

В качестве меры, основанной на расстоянии терминов, используются мера длины кратчайшего пути между понятиями WUP (анг. Wu & Palmer's measure). Мера WUP учитывает позиции терминов понятий относительно позиции наиболее специфичного общего узла графа иерархии:

$$wup(c_i,c_j) = \frac{2 * min_{p \in pths(mscs(c_i,c_j),rt)}(len_e(p))}{min_{c \in pths(c_i,c_j)}(len_e(p)) + 2 * min_{p \in pths(mscs(c_i,c_j),rt)}(len_e(p))} \tag{3.2}$$

где $mscs(c_i,c_j)$ – наиболее специфичный общий узел терминов $c_i$ и $c_j$, $len_e(p)$ – длина пути $p$ (количество ребер). Например, в тезаурусе WordNet[1] наиболее специфичным общим узлом для терминов "computer" и "monitor" является термин "device". Таким образом, WUP использует только отношения гиперонимии и гипонимии в иерархии терминов тезауруса.

В качестве меры, основанной на информационном содержании (англ. information Content, IC) используется RES (англ. Resnik's measure). Мера RES вычисляется по следующей формуле:

$$res(c_i,c_j) = -log(P(mscs(c_i,c_j))), \tag{3.3}$$

[1] `https://wordnet.princeton.edu`

где $mscs(c_i,c_j)$ – наиболее специфичный общий узел терминов $c_i$ и $c_j$, $P(mscs(c_i,c_j))$ – вероятность появления узла в ресурсе (например, корпусе текстов, тезаурусе).

Косинусная мера COS (англ. cosine similarity) вычисляется по следующей формуле:

$$cos(a,b) = \frac{a \cdot b}{\|a\| \, \|b\|} = \frac{\sum_{k=1}^{n} a_k b_k}{\sqrt{\sum_{k=1}^{n} a_k^2} \sqrt{\sum_{k=1}^{n} b_k^2}} \tag{3.4}$$

где $a, b$ – векторные представления слов $w_i, w_j$, соответственно; n - размерность вектора.

В настоящее время методы, использующие косинусную меру, показывает наилучшие результаты в задачах оценки семантической близости слов на русском языке [140].

Для анализа эффективности использования мер семантической связанности LESK, WU, RES в качестве ресурсов используются лингвистический тезаурус WordNet на английском языке и энциклопедия Википедия[2] (англ. Wikipedia) на русском языке. WordNet представляет собой структурированный лингвистический ресурс, разработанный в Принстонском университете США и содержащий более 155 тысяч слов и словосочетаний на английском языке. Википедия является многоязычной энциклопедией, содержащий более 1,2 млн. статей на русском языке. Для анализа эффективности использования косинусной меры использованы: (i) массив векторов распределённых представлений слов на английском языке (3 млн. слов и фраз), обученный по методу word2vec на корпусе новостей Google News dataset[3]; (ii) массив векторов распределённых представлений слов на русском языке, обученный по методу word2vec на корпусе текстов русскоязычной части Википедии (238 миллионов токенов) и корпусе текстов онлайн-библиотеки lib.rus.ec (12.9 миллиардов токенов) [140; 141].

[2] `https://ru.wikipedia.org/`

[3] Массив векторов доступен на странице word2vec `https://code.google.com/archive/p/word2vec/`

### 3.2.3 Алгоритм извлечения предметно-ориентированных проблемных высказываний и целевых объектов

Алгоритм использует результаты анализа текстового высказывания предложенными ранее методами: методом, основанным на ряде условий, и методом, основанным на анализе сложных предложений. Общее описание алгоритма состоит из нескольких шагов:

1. Извлечь из высказывания $s_{ij}$ вхождения индикаторов $\{pw_{i1}, pw_{i2}, \ldots, pw_{in}\}, n \leq |s_{ij}|$ в зависимости от связанных отрицаний из словарей Action, ProblemWord, NegativeWord, AddWord, ImperativePhrases, используя метод, основанный на ряде условий;
2. Для каждого $pw_{ij}$ определить множество возможных целевых объектов $\{t_1, t_2, \ldots, t_k\}$, если целевой объект $t_k$ синтаксически связан с $w_{ij}$, то есть существует прямая или косвенная зависимость между $t_k$ и $pw_{ij}$ в высказывании $s_{ij}$; если множество объектов пусто, то $w_{ij}$ исключается из множества индикаторов (см. Алгоритм 1);
3. Для каждого $t_k$ определить является ли объект предметно-ориентированным на основе мер связанности терминов понятия $t_k$ и терминов предметной области в лингвистическом ресурсе (Алгоритм 1);
4. Классифицировать высказывание $s_{ij}$ как высказывание, указывающее на проблемную ситуацию о предметно-ориентированном целевом объекте, если существует хотя бы одна комбинация $(pw_{ij}, t_k)$ и $r(s_{ij}) \neq 0$ согласно результату анализа методом, основанным на анализе сложных предложений; в противном случае, классифицировать высказывание $s_{ij}$ как не содержащее проблему.

## 3.3 Экспериментальное исследование

В данном разделе проводится экспериментальная оценка предложенного метода на текстовых данных для русского и английского языка, описанных в предыдущей главе. Основной целью экспериментов является выявление про-

**Algorithm 1:** Алгоритм извлечения предметно-ориентированных проблемных высказываний и целевых объектов

```
1  Function lookupForRelatedTargets(pw, DRs)
       Input: pw – найденный проблемный индикатор, DRs – множество
              зависимостей между словами
       Output: Ts – а множество целевых объектов
2      Ts ← ∅
3      foreach dr in DRs do
4          if dr.contains(pw) then
               /* поиск целевых объектов, прямо зависящих от
               индикатора                                    */
5              if dr.matches(direct_type_of_relations) then
6                  target=getTargetFromDep(dr)
7                  target=getAddWordsForTarget(target, DRs)
8                  Ts = Ts ∪ {target}
9              else
                   /* поиск целевых объектов, прямо зависящих от
                   слова-посредника                          */
10                 successor = theOtherWordFromDep(dr, pw)
11                 Ts = Ts ∪ lookupForRelatedTargets(successor, DRs\{dr})
12     return Ts

1  Function lookupForProblemsWithTargets(s, domain_terms,
   common_terms)
       Input: s – исходное предложение sentence, 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 терминами области domain_terms и широкой
               группы товаров 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
```

блемных фраз по отношению к целевым объектам и анализ принадлежности целевых объектов высказываний к предметной области продуктов, описанных в отзыве пользователя. На каждом из наборов данных определенной предметной области проводятся две серии экспериментов. В первой серии экспериментов анализируется эффективность двух разновидностей синтаксических зависимостях слов в тексте в задаче классификации высказываний пользователей по отношению к целевым объектам. Во второй серии экспериментов рассматривается эффективность определения семантической связанности предметно-ориентированным целевым объектам в задаче классификации высказываний по отношению к предметно-ориентированным целевым объектам.

### 3.3.1 Детали реализации и архитектура программного комплекса

Учитывая сайты-источники отзывов, в качестве терминов предметной области использовались названия основных разделов продуктов с официального сайта компании Hewlett-Packard, сайтов `amazon.com` и `otzovik.com`. Таблица 13 содержит определенные термины предметных областей корпусов отзывов. В качестве фоновых терминов, определяющих широкую группу товаров, использовались слова {*product*, *process*} и {*продукт*, *товар*} для английского и русского языка, соответственно.

Для подсчета семантической связанности целевого объекта и терминов предметной области использовалась фреймворк DKPro Similarity 2.1.0 на базе UIMA [142] (классы алгоритмов wikipedia.WikipediaBasedComparator.WikipediaBasedRelatednessMeasure и lsr.LexSemResourceComparator), WordNet 3.0, Википедия (версия от 27 октября 2014г.). Синтаксический разбор предложения на английском языке осуществляется библиотекой Stanford Parser (тип связей collapsed). Для синтаксического разбора предложения на русском языке выбран MST Parser [143], основанный на нахождении минимального остовного дерева и обученный на синтаксически размеченном корпусе русского языка СинТагРус [144], который содержит около 66 тысяч предложений. Согласно работе [145] MST Parser, обученный на 8,000 предложений корпуса СинТагРус, показывает качество разметки в 85%.

Таблица 13 — Термины предметной области

| Предметная область | Термины предметной области |
|---|---|
| Электроника (англ.) | computer, laptop, desktop, printer, scanner, ink, toner |
| Машины (англ.) | automotive, truck, equipment, car, vehicle, GPS, tire, wheel, motorcycle, powersport, garage |
| Детские товары (англ.) | toy, backpack, carrier, bedding, feeding, gear, stroller, swing, jumper |
| Инструменты (англ.) | tool, bedding, furniture, heating, cooling, kitchen, iron, steamer, floor, vacuum |
| Машины (рус.) | запчасть, машина, автомобиль, масло, кузов, двигатель, мотоцикл, шина, транспорт |
| Мобильные приложения (рус.) | финансы, приложение, сотовый телефон, программное обеспечение, компьютерная программа, смартфон, дисплей, подпрограмма |

**Программный компонент**

Программный компонент, извлекающий проблемные индикаторы и целевые объекты, используя синтаксические связи, реализован как часть общего конвейера обработки текстовых данных, построенного на базе платформы Apache UIMA и выложен в открытый доступ[4]. Исходный код UIMA-аннотаторов OpinionPhraseExtractor и PhraseReducer, реализующих описанный выше метод, написан на языке java. На входе аннотатор использует данные полученные от предыдущих этапов обработки текста: (i) результаты синтаксического разбора предложения, результаты определения частей речи, грамматики зависимостей и т.д. (Tokenizer, SentenceSplittter, Lemmatizer, PosTagger, DependencyParser); (ii) результаты работы компонента скрипта **DbA** на базе UIMA Ruta, сопоставляющего входной текст со словарями и рядом условий, который описан в предыдущем разделе. На выходе работы аннотаторов создаются аннотации типа ru.kpfu.itis.cll.general.OpinionPhrase, содержащие аннотацию индикатора ситуации и целевого объекта ru.kpfu.itis.cll.general.OpinionTarget. Рисунок 3.1 содержит общую архитектурную схему разработанного программного комплекса.

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

Финальная классификация высказываний осуществляется после выполнения скрипта **CbA** на базе UIMA Ruta согласно алгоритму из раздела 3.2.3. Общий разбор предложения в рамках платформы Apache UIMA осуществляется с помощью следующих библиотек с открытым исходным кодом на Java: DKPro[5] для английского языка и Textokit[6] для русского языка, представляющие собой технологический стек базовых функций обработки текста.

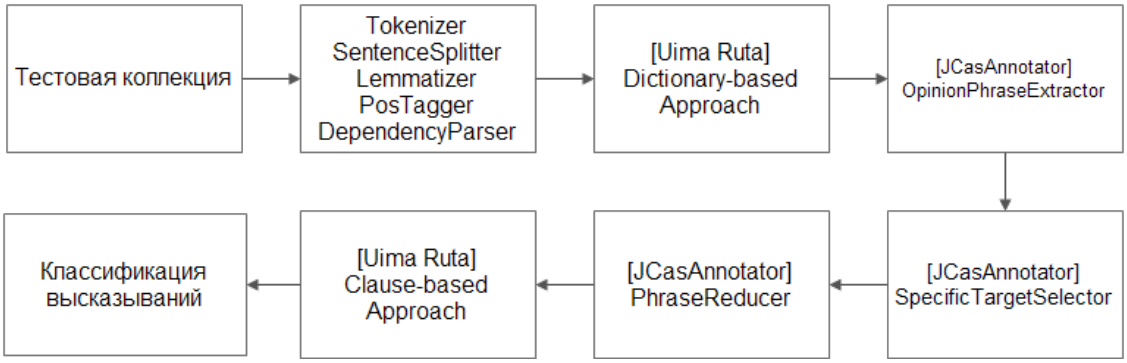



Рисунок 3.1 — Архитектурная схема программного компонента

### 3.3.2 Эксперименты и результаты

Эксперименты оцениваются с помощью следующих метрик качества: достоверность (обозн. Acc.), точность (обозн. P), полнота (обозн. R) и F-мера (обозн. F). В качестве базового метода классификации для сравнения результатов используются метод (обозн. CbA), основанный на анализе сложных предложений и описанный в главе 2.4. Применение алгоритма идентификации целевых объектов с помощью прямых и косвенных зависимостей без учета семантической связанности с предметной областью обозначено как DD и DD+ID, соответственно. Применение метода идентификации предметно-ориентированных целевых объектов с помощью семантической меры обозначен как DD+ID+CbA+мера (LESK, WU, RES или COS). Результаты классификации представлены в Таблицах 14, 15 и 16. Метод идентификации целевых объектов с последующей классификацией показывает наилучшие результаты по F-мере для текстов трех предметных областей (машины - .863, инструменты - .821,

[5] `https://dkpro.github.io/`

[6] `http://textocat.ru/textokit.html`

Таблица 14 — Результаты классификации высказываний

| Метод | Инструменты (англ.) | | | | Детские товары (англ.) | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | P | R | F1 | Acc. | P | R | F1 |
| 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 15 — Результаты классификации высказываний

| Метод | Машины (англ.) | | | | Электроника (англ.) | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | P | R | F1 | Acc. | P | R | F1 |
| 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 |

детские товары - .805). Для текстов о мобильных приложениях и электронике наилучшие результаты показывает метод классификации **CbA** без учета целевых объектов, основанный на анализе структуры предложений. Данные результаты можно объяснить сложной технической архитектурой двух предметных областей: пользователи не могут выявить конкретную проблемную компоненту продукта и упоминают лишь факт некорректной работы (например: “долго грузится, ничего не показывает весь день” не содержит указания на явный элемент приложения).

Наилучшие значения точности и полноты для английского и русского языка показывают методы проверки семантической связанности с помощью меры LESK, основанной на определениях в тезаурусе, и с помощью косинусной меры векторов распределённых представлений слов. Алгоритм DD+ID+CbA+LESK, подсчитывающий меру LESK на основе Википедии для текстов на русском языке, показывает более близкий результат к DD+ID+CbA, нежели алгоритм DD+ID+CbA+LESK на основе определений на английском языке в WordNet. Это объяснимо тем, что Википедия является открытой энциклопедией, что отражается в более значительном объеме статей.

Таблица 16 — Результаты классификации высказываний

| Метод | Машины (рус.) | | | | Мобильные приложения | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | P | R | F1 | Acc. | P | R | F1 |
| 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 |

Таблица 17 — Примеры проблемных индикаторов с извлеченными целевыми объектами

| Тип объекта | Примеры комбинаций индикаторов и объектов |
|---|---|
| предметно-ориентированный целевой объект | (неудобство, коробка), (недостатках, лампочек), (слабый, двигатель), (не открывающихся, дверей), (плохо, передач), (не удобны, сиденья), (низковатая, посадка), (разваливаться, спидометр), (не люксовая, машина) |
| не предметно-ориентированный целевой объект | (убиваемый, мнением), (слабо, обороты), (плохо, коэффициенты), (избавиться, хозяева), (разваливаться, эксплуатации), (плохом, климат), (не могут долго сидеть, ростом), (не светит словить, халявы), (нет, экстерьер), (отвратительное, отношение), (минусы, солярис), (нареканий, патриот), (отрицательный, отзыв), (неудобно, дизайн), (ошибки, расходники) |

**Анализ ошибок**

Таблица 17 содержит пример комбинаций $(pw_{ij}, t_k)$, в которой проверка связанности целевого объекта осуществлялась с помощью косинусной меры. Как видно из примеров, метод неверно определяет целевые объекты, упомянутые с помощью имен собственных и обозначающие марки машины (*солярис*, *патриот*), как не связанные с предметной областью.

После анализа ошибок выделено четыре типа высказываний о проблемных ситуациях с продуктами, зависящие от того, как именно пользователь указывает предметно-ориентированный целевой объект. Высказывание пользователя о продукте включает целевой объект в паре с проблемным индикатором в следующей форме:

1. Целевой объект, который явно указан в высказывании пользователя (например: “ремень ГРМ рвался очень часто”);
2. Целевой объект - ключевая компонента продукта, отвечающая за функциональность продукта и указанная с помощью частного термина (например: “на 4 андроиде не поддерживается на моем девайсе, сделайте пожалуйста, что бы работало” указывает на *операционную систему*);
3. Целевой объект указан в тексте с помощью проблемного индикатора (например: “я не смог печатать после месяца использования” указывает на проблемы с *принтером*);
4. Целевой объект не указан как конкретный объект, то есть является не определенным (например: “вообще ничего не работает: все сделал, вообще все, а не работает ничего”; “даже при том, что наш ребенок был очень маленьким, было все равно трудно поместить его”).

Третий тип целевых объектов требует создания методов идентификации объектов, основанных на знаниях в виде словарей составляющих продуктов и их возможных функций (например: *печатать* как *функция принтера*). Выражения пользователя с четвертым типом целевых объектов являются следствием того, что (i) продукт является сложным техническим и пользователь не может установить причину неполадки; (ii) отзыв пользователя носит повествовательный характер, поэтому целевой объект был упомянут ранее в тексте. Данный тип целевых объектов требует разработки методов, обеспечивающих более глубокий семантический анализ, рассматривая отзыв как структуру представления дискурса (англ. discourse-based methods) [117].

## 3.4 Выводы к третьей главе

В данной главе рассматривается задача извлечения информации о существовании тех или иных проблем с целевыми объектами, связанными с предметной областью продуктов компании. Для задачи извлечения высказываний, указывающих на проблемные ситуации с целевыми объектами предметной области, разработан метод извлечения проблемных индикаторов по отношению к предметно-ориентированным целевым объектам, основанный на синтаксиче-

ских зависимостях слов в высказывании. Для проверки принадлежности целевого объекта к предметной области используется мера семантической связанности. В работе исследуется эффективность различных мер семантической связанности на основе лингвистического тезауруса и с помощью векторов распределённых представлений слов. Результаты показали, что метод определения целевых объектов, использующий синтаксические связи между проблемными индикаторами и существительными, показывает наилучшие результаты классификации для отзывов о механических и низкотехнологичных продуктах на русском и английском языках (машины, инструменты, детские товары). Согласно результатам классификации, наиболее эффективными мерами подсчета семантической связанности между найденными целевыми объектами и терминами предметной области являются мера, основанная на определениях в лингвистическом ресурсе, и косинусная мера, использующая вектора представлений объектов и терминов.

# Глава 4. Выделение тематически сгруппированных объектов мнений, указывающих на проблемные ситуации в использовании продуктов, на основании коллекции отзывов предметной области

В данной главе рассматривается задача выделения тематически сгруппированных объектов мнений пользователей о проблемных ситуациях с продуктами на основании коллекции отзывов определенной предметной области.

## 4.1 Описание задачи

В рамках анализа естественного языка отзыв рассматривается как текстовый документ, описывающий иерархическую структуру связного текста следующим образом: общая тема документа может быть описана посредством более конкретных тем текста, которые так же могут быть охарактеризованы более точными темами [97]. Таким образом, каждое предложение связного текста посвящено раскрытию подтем основной темы текста [146]. Отзывы о конкретном продукте могут содержать мнения о различных аспектах продукта, которые связаны по тематике. Потенциальные покупатели продуктов могут больше учитывать непрерывную и качественную работу ходовых компонентов автомобиля (например, двигателя и ходовой автомобиля), несмотря на слишком высокие цены на оригинальные запчасти для автомобиля. С другой стороны, компании и разработчики могут быть заинтересованы в первостепенном устранении дефектов приложений, вызывающие наибольший негативный отклик, нежели мнениям о нехватке функционала или о размерах кнопок. Однако с каждым годом анализ требуемых отзывов вручную становится более затруднительным в связи с экспоненциальным ростом пользовательских текстов в сети. Это объясняет необходимость создания методов 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 тен-

денцию встречаться совместно с определенной тональностью в отзывах пользователей.

Рассмотрим два отзыва пользователей об автомобилях с сайта `otzovik.com` в качестве примеров. Отзывы no. 984595 и no. 67092 содержат описание автомобилей по следующим тематическим категориям: *безопасность*$_1$, *комфорт*$_2$, *ходовые качества*$_3$, *надежность*$_4$, *внешний вид*$_5$, *цена*$_6$, *машина в целом*$_7$ (согласно схеме сайта и разметки отзывов в рамках соревнования SentiRuEval-2015). Термины тем обозначены с помощью порядковых номеров, соответствующие номерам тематических категорий. Проблемы с целевыми объектами отзыва в тексте выделены жирным, отрицание проблемы пользователем подчеркнуто. Термины тем выделены в цвета, соответствующие названиям тематических категорий. Проблемы с целевыми объектами отзыва в тексте выделены жирным, отрицание проблемы пользователем подчеркнуто.

> Отзыв no. 984595. “Купил автомобиль$_7$ в 2011 году Ford$_7$ Focus$_7$ 2$_7$ хэтчбек$_7$ рестайлинг в комплектации Titanium с бензиновым$_3$ двигателем$_3$ 1.8 литра. Удивила плавность$_3$ хода$_3$. Абсолютно все нравилось **за исключением шумоизоляции**$_2$. **“Шумели”**$_2$ **задние**$_2$ **арки**$_2$. **Сделал шумку**$_2$ (2 слоя виброизоляции$_2$ и слой шумоизоляции$_2$) и проблема исчезла. Заменил штатную$_2$ магнитолу$_2$ на магнитолу$_2$ Pioneer$_2$ для лучшего качества$_7$ аудио$_7$ системы$_7$. Когда на одометре набежало 80 000 км начались **проблемы с ходовой**$_4$ (**сломалась**$_4$ **пружина**$_4$ **амортизатора**$_4$, **появился шум**$_2$) поэтому заменил$_4$ полностью$_4$ ходовую$_4$, не стал ждать последующих поломок$_4$. В общем можно сказать, что за 3 года пользования автомобилем$_7$ не было серьезных проблем. Общее впечатление: Хороший, практичный автомобиль$_7$.” (общая оценка: 4 из 5, url: `http://otzovik.com/review_984595.html`)

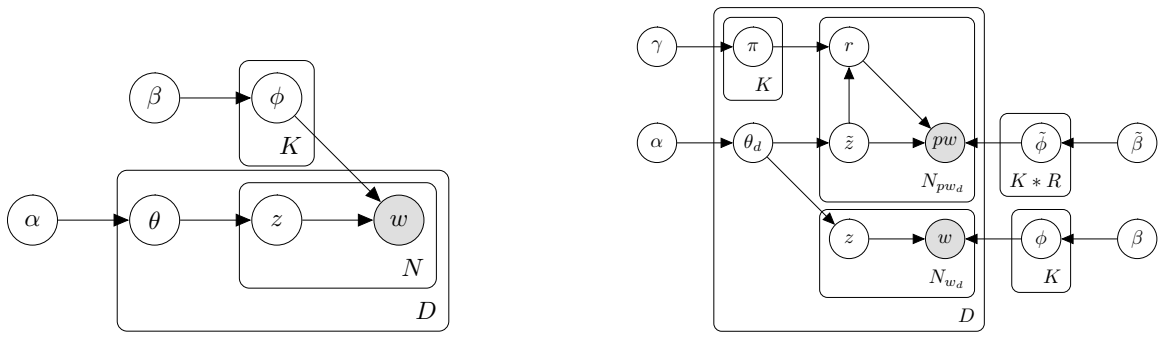
> Отзыв no. 67092. “Российский автопром$_7$ упорно не хочет радовать своим производством. Не хотят выпускать нормальные машины. А “двенашка$_7$” , купе$_7$ или хэч$_7$, она и есть - “двенашка$_7$” . Не доделанное произведение$_7$. Что то в ней есть. Молодёжная$_7$ модель$_7$. Смотрится$_5$ привычно. Движка$_3$ родная, не доведённая до ума. **Веч-**

**ные проблемы с электроникой**$_4$. **Шумоизоляция**$_2$ **- лишь пустое слово**. **Подвеска**$_3$ **звонкая**$_2$ **и жёсткая**$_3$. **Отзывчивость**$_3$ **педалей**$_3$ **тормоза**$_3$ **и газа**$_3$ **посредственная**. **Тяжёлый**$_3$ **руль**$_3$. **Комфорт**$_5$ **от вождения нулевой**. **Места**$_2$ **для**$_2$ **задних**$_2$ **пассажиров**$_2$ **мало**. **Багажник**$_2$ **средний**. **Интерьер**$_2$ **пустоватый и дешёвый**$_5$ с твёрдым$_2$ пластиком$_2$. **О безопасности**$_1$ **я молчу**. Но... За такие деньги$_6$ новый хороший авто **не купишь**. Я говорю: “Новый автомобиль$_6$” . Всегда есть альтернатива$_7$ - подержанная$_7$ иномарка$_7$. Всегда же стремимся к лучшему. Общее впечатление: **Не авто**$_7$.” (общая оценка: 2 звезды из 5; url: `http://otzovik.com/review_67092.html`)

Отзывы обладают противоположной тональностью: автор первого отзыва доволен автомобилем, поставив 4 звезды в то время, как второй автор не удовлетворён продуктом. Первый автор пишет о неудобствах в процессе использования, названные в литературе мягкими проблемами сложных функциональных продуктов [5]. Текст обладает негативной тональностью о комфорте и надежности, позитивной информацией о ходовых качествах и нейтральной тональностью о машине в целом. Второй автор пишет о проблемных ситуациях, связанные с ходовыми качества (более техническими целевыми объектами) и комфортом (составными частями автомобиля), и выражает больше негативных моментов в каждой теме (“автопром не хочет радовать”, “хороший авто не купишь”). Данные примеры подтверждают необходимость разработки методов резюмирования целевых объектов продуктов и высказываний в коллекции документов.

В настоящий момент доминирующими методами являются алгоритмы на основе вероятностной модели латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) [116]. Это связано с тем, что традиционные методы классификации текстов по категориям требуют больших затрат времени на создание размеченной коллекции для обучения. Вероятностные методы позволяют использовать коллекции неразмеченных документов, содержащиеся на онлайн-ресурсах, и сформулировать модель генерации текстов отзывов с помощью скрытых переменных. На основании разработанной модели вычисляется распределения документов и слов по темам.

В рамках данной задачи требуется решить две подзадачи анализа:




а) LDA б) TPrPhModel

Рисунок 4.1 — Вероятностные модели а) LDA и б) TPrPhModel

1. Определение проблемных индикаторов относительно тематических категорий отзывов;
2. Извлечение проблемных высказываний относительно тематических категорий с выявлением эмоционально-окрашенного мнения пользователей.

Для достижения целей исследования предложены две совместные тематические модели:

1. Модель *тематических высказываний, указывающих на проблемную ситуацию* (problem phrase topic model, TPrPhModel);
2. Модель *тема-тональность-проблема* (topic-sentiment-problem model, TSPM).

## 4.2 Совместная вероятностная тематическая модель для извлечения тем и высказываний, указывающих на проблемую ситуацию

Предложенная модель TPrPhModel является модификацией LDA для извлечения проблемных индикаторов и целевых объектов по отношению к темам. В вероятностных моделях на основе LDA используется допущение, что появление слов в документе $d$, относящихся к теме $t$, не зависит от документа $d$ и моделируется распределением $p(w|t)$ для всей текстовой коллекции.

Целевые объекты как составные компоненты продукта чаще всего являются существительными. Проблемные индикаторы, согласно анализу и составленным словарям в главе 1, чаще всего могут быть глаголами (например: *заме-*

*нить*, *сломать*), существительными (например: *проблема*, *ошибка*), наречиями (например: *слишком*, *только*) и прилагательными (например: *грязный*, *отвратительный*). Учитывая данное знание при моделировании распределений слов, целевые объекты не обладают признаком проблемности, то есть не могут указывать на существование проблемных высказываний и играют роль фактической информации о составных компонентах продуктов Графические представления моделей LDA и TPrPhModel приведены на Рисунке 4.1.

Пусть дано множество отзывов пользователей $D = d_1,d_2,\ldots,d_n$, каждый документ в коллекции представим как множество $N_d$ слов, состоящее из проблемных индикаторов, целевых слов и прочего контекста, $|N_d| = |N_{w_d}| + |N_{pw_d}|$. Каждое слово отзыва является лексической единицей словаря, содержащей индексы слов от 1 до $|V|$. В Таблице 18 приводится список основных обозначений, используемых в модели. В рамках модели TPrPhModel каждой теме соответствует мультиномиальное распределение в пространстве слов. Слово $pw_d$, которое указывает на проблемную ситуацию или ее отсутствие, называется проблемным словом, в противном случае слово называется контекстным. Для каждого слова в документе темы $z$ и $\tilde{z}$ выбирается из мультиномиального распределения $\theta$. Проблемному слову, соответствует тема $\tilde{z}$, контекстному слову в документе, относящиеся к целевым объектам или описываемой ситуации в целом, соответствует тема $z$. Затем выбирается проблемная метка $r$ из мультиномиального распределения $\pi$, соответствующего теме $\tilde{z}$. Затем слово $pw_d$ выбирается из распределения $\tilde{\phi}$, соответствующего теме $\tilde{z}$ и проблемной метке $r$. Контекстное слово $w_d$ выбирается из распределения $\phi$, соответствующего теме $z$. Таким образом, все слова в документах порождаются в зависимости от некоторой латентной темы или в зависимости от некоторой латентной темы и проблемной метки.

TPrPhModel характерен следующий порождающий процесс:

- для каждой пары (тема $\tilde{z}$, проблемная метка $r$) выбирается распределение проблемных слов в каждой теме $\tilde{\phi}_{\tilde{z},r} \sim Dir(\beta_r)$ $(r \in pr,no-pr)$;
- для каждой темы z выбирается распределение контекстных слов в каждой теме $\phi_{z,r} \sim Dir(\beta)$;
- для каждого документа (отзыва) $d$:
  - выбирается случайный вектор $\theta_d \sim Dir(\alpha)$;

Таблица 18 — Основные обозначения в вероятностных моделях

| Символ | Описание |
|---|---|
| $D$ | коллекция документов |
| $V$ | словарь коллекции (множество уникальных слов в коллекции D) |
| $N$ | число слов в коллекции |
| $K$ | число тем |
| $S$ | число тональных классов |
| $R$ | число проблемных классов |
| $w_d$ | вектор слов документа $d$ |
| $N_{w_d}$ | число слов в документе $d$ |
| $\theta_d$ | мультиномиальное распределение в пространстве тем с параметром $\alpha$, $\Theta = \{\{\theta_z\}_{z=1}^K\}_{d=1}^D$ |
| $\pi_z$ | мультиномиальное распределение в пространстве меток для темы $z$ с параметром $\gamma$ |
| $\xi_{z,l}$ | мультиномиальное распределение моделей TSPM в пространстве проблемных меток для пары (тема $z$, тональная метка $l$) с параметром $\eta$ |
| $\phi_{z,l,r}$ | мультиномиальное распределение моделей TSPM в пространстве слов для троек (тема $z$, тональная метка $l$, проблемная метка $r$) с параметром $\beta$, $\Phi = \{\{\{\{\phi_{z,l,r}\}_{r=1}^R\}_{l=1}^L\}_{z=1}^K\}_{v=1}^V$ |
| $\phi_z$ | мультиномиальное распределение модели TPrPhModel в пространстве слов для темы $z$ с параметром $\beta$, $\Phi = \{\{\phi_z\}_{z=1}^K\}_{v=1}^V$ |
| $\tilde{\phi}_{z,r}$ | мультиномиальное распределение модели TPrPhModel в пространстве слов для (тема $z$, проблемная метка $r$) с параметром $\tilde{\beta}$, $\tilde{\Phi} = \{\{\{\tilde{\phi}_{z,r}\}_{r=1}^R\}_{z=1}^K\}_{v=1}^V$ |
| $z_{di}$ | множество тем, присвоенных $i$-му слову в документе $d$ |
| $l_{di}$ | множество тональных меток, присвоенных $i$-му слову в документе d |
| $r_{di}$ | множество проблемных меток, присвоенных $i$-му слову в документе d |
| $\alpha$ | априорное распределение Дирихле на параметры $\theta$ |
| $\beta$ | априорное распределение Дирихле на параметры $\phi$ |
| $\eta$ | априорное распределение Дирихле на параметры $\xi$ |
| $\gamma$ | априорное распределение Дирихле на параметры $\pi$ |

– для каждой темы $\tilde{z}$ выбирается вектор тональных меток $\pi_d^{\tilde{z}} \sim Dir(\gamma)$;
– для каждого контекстного слова $w_i$ в документе $d$:
    * выбирается тема $z_{d,w_i} \sim Mult(\theta_d)$;

– для каждой темы выбирается слово $w_i$ из распределения в пространстве слов с параметром $\beta$, зависящее от переменной $z$
– для каждого проблемного слова $pw_i$ в документе $d$:
    * выбирается тема $\tilde{z}_{d,pw_i} \sim Mult(\theta_d)$;
– для каждой темы выбирается проблемная метка $r_{d,pw_i} \sim Mult(\pi_d^{\tilde{z}})$;
– для каждой пары (тема, проблемная метка) выбирается слово $pw_i$ из распределения в пространстве слов с параметром $\beta$, зависящее от комбинации $\tilde{z}_{d,pw_i}$,$r_{d,pw_i}$.

Таким образом, требуется оценить три множества параметров модели: отношение тем $\theta_d$ и отношение проблем $\pi_{d,\tilde{z},r}$, определенные для документа; отношение тем $\phi_z$ и совместное отношение тема-проблема $\phi_{\tilde{z},r}$, определенные для корпуса текстовых документов.

### 4.2.1 Статистическое оценивание модели

Для решения задачи статистического оценивания применительно к тематической модели LDA существует несколько алгоритмов: сэмплирование Гиббса (англ. Gibbs sampling), вариационный вывод (англ. variational inference), распространение математического ожидания (англ. expectation propagation) [116; 147; 148]. В данной диссертационной работе применяется сэмплирование Гиббса для оценивания параметров модели, поскольку такой подход позволяет эффективно находить скрытые темы в корпусах текстов [149].

Рассмотрим оценивание моделирования проблемных слов. Оценивание моделирования контекстных слов совпадает с оцениванием в стандартной LDA. Для $i$-го слова в документе d определим индекс $t = (d, i)$; для проблемных слов метод Гиббса выбирает скрытые переменные из условного распределения $P(\tilde{z}_t = k, r_t = r|\mathbf{w}, \tilde{\mathbf{z}}_{\neg t}, \mathbf{r}_{\neg t}, \alpha, \tilde{\beta}, \gamma)$, которое может быть посчитано из совместного распределения $P(\mathbf{w}, \tilde{\mathbf{z}}, \mathbf{r})$. Опуская из распределения гиперпараметры и используя $\neg t$ для обозначения соответствующего документа, темы и проблем-

ной метки, получаем:

$$P(\tilde{z}_t = k, r_t = r|\mathbf{w}, \tilde{\mathbf{z}}_{\neg t}, \mathbf{r}_{\neg t}) = \frac{P(\mathbf{w}, \tilde{\mathbf{z}}, \mathbf{r})}{P(\mathbf{w}, \tilde{\mathbf{z}}_{\neg t}, \mathbf{r}_{\neg t})} = \frac{P(\mathbf{w}|\mathbf{r}, \tilde{\mathbf{z}})P(\mathbf{r}, \tilde{\mathbf{z}})}{P(\mathbf{w}_{\neg t}|\mathbf{r}_{\neg t}, \tilde{\mathbf{z}}_{\neg t})P(w_t)P(\mathbf{r}_{\neg t}, \tilde{\mathbf{z}}_{\neg t})} \propto \frac{P(\mathbf{w}|\mathbf{r}, \tilde{\mathbf{z}})P(\mathbf{r}, \tilde{\mathbf{z}})}{P(\mathbf{w}_{\neg t}|\mathbf{r}_{\neg t}, \tilde{\mathbf{z}}_{\neg t})P(\mathbf{r}_{\neg t}, \tilde{\mathbf{z}}_{\neg t})} = \frac{P(\mathbf{w}|\mathbf{r}, \tilde{\mathbf{z}})}{P(\mathbf{w}_{\neg t}|\mathbf{r}_{\neg t}, \tilde{\mathbf{z}}_{\neg t})} \cdot \frac{P(\mathbf{r}|\tilde{\mathbf{z}})}{P(\mathbf{r}_{\neg t}|\tilde{\mathbf{z}}_{\neg t})} \cdot \frac{P(\tilde{\mathbf{z}})}{P(\tilde{\mathbf{z}}_{\neg t})} \tag{4.1}$$

Предложенная модель является модификацией латентного размещения Дирихле, поэтому векторы документов и совместные векторы тем и проблемных меток порождаются распределениями Дирихле [116]. Распределение Дирихле определяется следующим уравнением для вектора $\mathbf{p}$ и гиперпараметра $\alpha$:

$$Dir(\mathbf{p}, \alpha) = \frac{1}{B(\alpha)} \prod_{v=1}^{|\alpha|} p_t^{\alpha_t - 1}$$
$$B(\alpha) = \frac{\prod_{i=1}^{|\alpha|} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{|\alpha|} \alpha_i)}, \Gamma(n) = (n-1)! \tag{4.2}$$

Если вектор $\mathbf{x}$ имеет мультиномиальное распределение, то верно следующее (подробное описание можно найти в [150]):

$$P(\mathbf{p}|\mathbf{x}, \alpha) = Dir(\mathbf{p}; \mathbf{x} + \alpha) = \frac{1}{B(\mathbf{x} + \alpha)} \prod_{v=1}^{|\alpha|} p_t^{x_t + \alpha_t - 1}$$
$$B(\mathbf{x} + \alpha) = \int \prod_{v=1}^{|\alpha|} p_t^{x_t + \alpha_t - 1} dp \tag{4.3}$$

Совместная вероятность проблемных слов, тем и проблемных меток может быть разложена на следующие множители:

$$P(\mathbf{w}, \mathbf{r}, \tilde{\mathbf{z}}) = P(\mathbf{w}|\mathbf{r}, \tilde{\mathbf{z}})P(\mathbf{r}, \tilde{\mathbf{z}}) = P(\mathbf{w}|\mathbf{r}, \tilde{\mathbf{z}})P(\mathbf{r}|\tilde{\mathbf{z}})P(\tilde{\mathbf{z}}) = \int P(\mathbf{w}|\mathbf{r}, \tilde{\mathbf{z}}, \Phi)P(\Phi|\tilde{\beta})d\Phi \cdot \int P(\mathbf{r}|\tilde{\mathbf{z}}, \Pi)P(\Pi|\gamma)d\Pi \cdot \int P(\tilde{\mathbf{z}}|\Theta)P(\Theta|\alpha)d\Theta \tag{4.4}$$

Обозначим, для краткости записей, $\tilde{\phi} = \phi$ и $\tilde{\beta} = \beta$. Каждый множитель из уравнения 4.4 может быть подсчитан независимо. Интегрируя первый множитель

по Φ с помощью интегралов Дирихле, получаем:

$$P(\Phi|\beta) = \prod_{k=1}^{K}\prod_{j=1}^{R} P(\phi_{k,i}|\beta) = \prod_{k=1}^{K}\prod_{j=1}^{R} \frac{1}{B(\beta)} \prod_{i=1}^{V} \phi_{k,j,i}^{\beta_{j,i}-1} \tag{4.5}$$

$$\begin{gathered} P(\mathbf{w}|\mathbf{r},\tilde{\mathbf{z}}) = \int P(\mathbf{w}|\mathbf{r},\tilde{\mathbf{z}},\Phi)P(\Phi|\beta)d\Phi = \\ \int \prod_{k=1}^{K}\prod_{j=1}^{R}\prod_{i=1}^{V} \phi_{k,j,i}^{n_{k,j,i}} \frac{\Gamma(\sum_{i=1}^{V}\beta_{j,i})}{\prod_{i=1}^{V}\Gamma(\beta_{j,i})} \prod_{i=1}^{V} \phi_{k,j,i}^{\beta_{j,i}-1} d\phi_{k,j} = \\ \prod_{k=1}^{K}\prod_{j=1}^{R} \frac{\Gamma(\sum_{i=1}^{V}\beta_{j,i})}{\prod_{i=1}^{V}\Gamma(\beta_{j,i})} \cdot \int \prod_{i=1}^{V} \phi_{k,j,i}^{n_{k,j,i}+\beta_{j,i}-1} d\phi_{k,j} = \\ \prod_{k=1}^{K}\prod_{j=1}^{R} \frac{\Gamma(\sum_{i=1}^{V}\beta_{j,i}) \prod_{i=1}^{V}\Gamma(n_{k,j,i}+\beta_{j,i})}{\prod_{i=1}^{V}\Gamma(\beta_{j,i})\Gamma(n_{k,j}+\sum_{i=1}^{V}\beta_{j,i})}, \end{gathered} \tag{4.6}$$

где $n_{k,j,i}$ означает количество раз, когда слову $j$ присвоена тема $k$ и проблемная метка $j$ в коллекции документов, $n_{k,j}$ определяет общее количество слов, которым присвоена пара $(k,j)$, Γ - гамма-функция. Интегрируя второй множитель по Π, получаем:

$$\begin{gathered} P(\mathbf{r}|\tilde{\mathbf{z}}) = \int P(\mathbf{r}|\tilde{\mathbf{z}},\Pi)P(\Pi|\pi)d\Pi = \\ \int \prod_{d=1}^{D}\prod_{k=1}^{K}\prod_{j=1}^{R} \pi_{d,k,j}^{n_{d,k,j}} \frac{\Gamma(\sum_{j=1}^{R}\gamma_{k,j})}{\prod_{j=1}^{R}\Gamma(\gamma_{k,j})} \prod_{j=1}^{R} \pi_{d,k,j}^{\gamma_{k,j}-1} d\pi_{d,k} = \\ \prod_{d=1}^{D}\prod_{k=1}^{K} \frac{\Gamma(\sum_{j=1}^{R}\gamma_{k,j}) \prod_{j=1}^{R}\Gamma(n_{d,k,j}+\gamma_{k,j})}{\prod_{j=1}^{R}\Gamma(\gamma_{k,j})\Gamma(n_{d,k}+\sum_{j=1}^{R}\gamma_{k,j})}, \end{gathered} \tag{4.7}$$

где число $n_{d,k,j}$ обозначает количество слов в документе $d$, присвоенных теме $k$ и проблемной метке $j$, $n_{d,k}$ определяет количество слов документа $d$, присвоенных теме $k$. Интегрируя третий множитель по Θ, получаем:

$$\begin{gathered} P(\tilde{\mathbf{z}}) = \int P(\tilde{\mathbf{z}}|\Theta)P(\Theta|\alpha)d\Theta = \\ \int \prod_{d=1}^{D}\prod_{k=1}^{K} \theta_{d,k}^{n_{d,k}} \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k-1} d\theta_d = \\ \prod_{d=1}^{D} \frac{\Gamma(\sum_{k=1}^{K}\alpha_k) \prod_{k=1}^{K}\Gamma(n_{d,k}+\alpha_k)}{\prod_{k=1}^{K}\Gamma(\alpha_k)\Gamma(n_d+\sum_{k=1}^{K}\alpha_k)}, \end{gathered} \tag{4.8}$$

где общее число $n_{d,k}$ обозначает количество слов в документе $d$, присвоенных теме $k$, $n_d$ обозначает общее число слов в документе $d$.

Выражение условного распределения $P(\tilde{z}_t = k, r_t = r|\mathbf{w}, \tilde{\mathbf{z}}_{\neg t}, \mathbf{r}_{\neg t}, \alpha, \beta, \gamma)$ вычисляется из уравнения 4.1 с помощью подставления уравнений 4.6, 4.7 и 4.8 и удаления множителей, которые не содержат слова $w_t$. В данной работе $\alpha_k = \alpha$ и $\gamma_{k,r} = gamma$ для всех $k, r$. Таким образом, скрытые параметры темы $\tilde{z}$ и проблемной метки $r$ в модели TPrPhModel могут быть выбраны по следующей формуле для всех проблемных слов $pw$:

$$P(\tilde{\mathbf{z}}_t = z, \mathbf{r}_t = r|\mathbf{pw}_t = w, \tilde{\mathbf{z}}_{-t}, \mathbf{r}_{-t}, \mathbf{pw}_{-t}, \alpha, \tilde{\beta}, \gamma) \propto \frac{n^{-j}_{k,r,w} + \tilde{\beta}^w_r}{n^{-t}_{k,r} + \sum_{i'=1}^{V} \tilde{\beta}^{i'}_r} \frac{n^{-t}_{d,k,r} + \gamma}{n^{-t}_{d,k} + R * \gamma} \frac{n^{-t}_{d,k} + \alpha}{n^{-t}_{d} + K * \alpha} \tag{4.9}$$

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

$$P(\mathbf{z}_t = z|\mathbf{w}_t = w, \mathbf{z}_{-t}, \mathbf{w}_{-t}, \alpha, \beta) \propto \frac{n^{-j}_{k,w} + \beta^w}{n^{-t}_{k} + \sum_{i'=1}^{V} \beta^{i'}} \frac{n^{-t}_{d,k} + \alpha}{n^{-t}_{d} + K * \alpha} \tag{4.10}$$

Приближенные распределения $\phi_{k,j}$ и $\tilde{\phi}_{k,r,j}$, определенные для корпуса, вычисляются по формулам:

$$\begin{aligned} \tilde{\phi}_{k,r,j} &= \frac{n^{-j}_{k,r,w} + \tilde{\beta}^w_r}{n^{-t}_{k,r} + \sum_{i'=1}^{V} \tilde{\beta}^{i'}_r} \\ \phi_{k,j} &= \frac{n^{-j}_{k,w} + \beta^w}{n^{-t}_{k} + \sum_{i'=1}^{V} \beta^{i'}} \end{aligned} \tag{4.11}$$

Приближенное распределение тем $\theta_{d,k}$, определенное для документа, вычисляется по формуле:

$$\theta_{d,k} = \frac{n^{-t}_{d,k} + \alpha}{n^{-t}_{d} + K * \alpha} \tag{4.12}$$

Приближенное совместное распределение тем и проблемных меток $\pi_{d,k,r}$, определенное для документа, вычисляется по формуле:

$$\pi_{d,k,r} = \frac{n^{-t}_{d,k,r} + \gamma}{n^{-t}_{d,k} + R * \gamma} \tag{4.13}$$

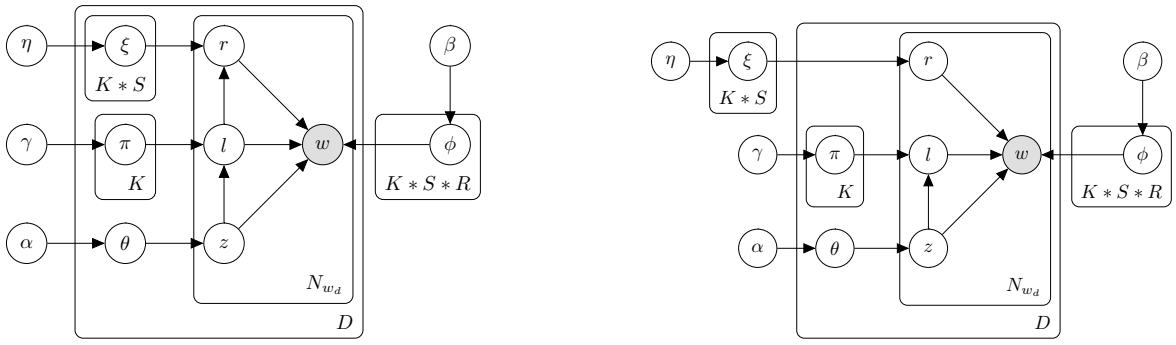
Используя полученные оценки модели, порождение слов в документах происходит согласно Алгоритму 2.

**Algorithm 2:** Алгоритм порождения слов с помощью модели TPrPhModel.

**Function** `sampling`(*корпус документов,* $\alpha$, $\beta$, $\tilde{\beta}$, $\gamma$)
  **Input**: гиперпараметры $\alpha$, $\beta$, $\tilde{\beta}$, $\gamma$, 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** $i = 1$ *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 $\tilde{z}$ и проблемной метке $r$, из счетчиков $n_{k,j,i}, n_{k,j}, n_{d,k,j}, n_{d,k}$ и $n_d$
        Сэмплировать новую пару $(\tilde{z}', r')$ из уравнения 4.9
        Обновить счетчики $n_{k,j,i}, n_{k,j}, n_{d,k,j}, n_{d,k}$ и $n_d$ для новой темы $z'$ и новой проблемной метки $r'$
      **foreach** *контекстное слово* $i \in 1, \ldots, N_{w_d}$ **do**
        Исключить слово $i$, присвоенное теме $z$ из $n_{k,i}, n_k, n_{d,k}$ и $n_d$
        Сэмплировать новую тему $z'$ из уравнения 4.10
        Обновить $n_{k,j,i}, n_{k,j}, n_{d,k,j}, n_{d,k}$ и $n_d$ для новой темы $z'$;

## 4.3 Совместная вероятностная тематическая модель для извлечения тем, тональных и проблемных высказываний

Описание проблемной ситуации с продуктами может сопровождаться эмоционально-окрашенными высказываниями относительно аспектных терминов (целевых объектов), о которых высказывание было сделано. 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”). Проблемное высказывание может сопровождаться негативной или позитивной тонально-




а) TSPM(DP) б) TSPM(GP)

Рисунок 4.2 — Вероятностные модели а) TSPM(DP) и б) TSPM(GP)

стью, если пользователь описывает затруднения с комфортным использованием продукта относительно разных категорий целевых объектов (например, “в ресторане отвратительное обслуживание”, “не слишком чувствительный сенсор”, “батарея держится долго, но меньше, чем заявлено в инструкции”, “было бы лучше, если бы не было шумов от двигателя”). Таким образом, в зависимости от темы (категории) целевого объекта, пользователь может использовать слова различной тональности.

Для анализа взаимосвязи между информацией о проблемных ситуациях и тональности высказываний в рамках диссертационной работы предложена модель *тема-тональность-проблема* (topic-sentiment-problem model, TSPM) в 2 модификациях:

1. Модель TSPM(DP), в которой проблемные переменные слов зависят от локального контекста: тональной и тематической переменных 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).

Графические представления моделей TSPM(DP) и TSPM(GP) приведены на рисунке 4.2. В рамках TSPM моделей каждой теме соответствует мультиномиальное распределение в пространстве слов. Для каждого слова в документе тема $z$ выбирается из мультиномиального распределения $\theta$, затем выбирается тональная метка $l$ из мультиномиального распределения $\pi$, соответствующего теме $z$. Затем в рамках TSPM(DP) выбирается проблемная метка $r$ из мультиномиального распределения $\xi$, соответствующего тройке (документ $d$, тема $z$,

тональная метка $l$). В рамках TSPM(DP) выбирается проблемная метка $r$ из мультиномиального распределения $\xi$, соответствующего паре (тема $z$, тональная метка $l$), агрегируя информацию по всем документам коллекции. Наконец, слово $w$ выбирается из распределения $\Phi$, соответствующего теме $z$, тональной метке $l$, проблемной метке $r$. Таким образом, слова в документах порождаются в зависимости от некоторой латентной темы, латентной тональной и проблемной меток. Совместная вероятность слов, тем, тональных и проблемных меток для TSPM(DP) может быть посчитана следующим образом:

$$P(w, z, l, r) = P(w|z, l, r) \cdot P(r|z, l) \cdot P(l|z) \cdot P(z) \tag{4.14}$$

Для TSPM(DP) характерен следующий порождающий процесс:

- для каждой тройки (тема $z$, тональная метка $l$, проблемная метка $r$) выбирается распределение слов в каждой теме $\Phi_{z,l,r} \sim Dir(\beta_{l,r})$ ($l \in \{neu,pos,neg\}, r \in \{pr,no-pr\}$)
- для каждого документа (отзыва) $d$:
  - выбирается случайный вектор $\theta_d \sim Dir(\alpha)$;
  - для каждой темы $z$ выбирается вектор тональных меток $\pi_d^z \sim Dir(\gamma)$;
  - для каждой пары $(z, l)$ выбирается вектор проблемных меток $\xi_d^{z,l} \sim Dir(\eta)$;
  - для каждого слова $w_i$ в документе $d$:
    - выбирается тема $z_{d,w_i} \sim Mult(\theta_d)$;
    - для каждой темы выбирается тональная метка $l_{d,w_i} \sim Mult(\pi_d^z)$;
    - для каждой пары (тема, тональная метка) выбирается проблемная метка $r_{d,w_i} \sim Mult(\xi_d^{z,l})$;
    - для каждой тройки (тема, тональная метка, проблемная метка) выбирается слово $w_i$ из распределения в пространстве слов с параметром $\beta$, зависящее от комбинации $(z_{d,w_i},l_{d,w_i},r_{d,w_i})$.

Порождающий процесс для TSPM(GP) является аналогичным описанному процессу для TSPM(DP), где проблемная метка выбирается из мультиномиального распределения $r_{d,w_i} \sim Mult(\xi^{z,l})$, независящего от документа.

### 4.3.1 Статистическое оценивание предложенной модели

Для статистического оценивания параметров предложенных моделей применяется сэмплирование Гиббса. Для слова $i$-го слова в документе $d$ определим индекс $t = (d,i)$; для слов выбранный метод Гиббса выбирает скрытые переменные из условного распределения $P(z_t = k, l_t = l, r_t = r|w_{d,i} = w, z_{\neg t}, l_{\neg t}, r_{\neg t}, \alpha, \beta, \gamma, \eta)$ которое может быть посчитано из совместного распределения $P(w,z,l,r)$. Опуская из распределения гиперпараметры и используя $\neg t$ для обозначения соответствующего документа, темы и проблемной метки получаем:

$$\begin{gathered} P(\tilde{z}_t = k, l_t = l, r_t = r|\mathbf{w}, \mathbf{z}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{r}_{\neg t}) = \frac{P(\mathbf{w}, \mathbf{z}, \mathbf{l}, \mathbf{r})}{P(\mathbf{w}, \mathbf{z}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{r}_{\neg t})} = \\ \frac{P(\mathbf{w}|\mathbf{r}, \mathbf{l}, \mathbf{z})P(\mathbf{r}, \mathbf{l}, \mathbf{z})}{P(\mathbf{w}_{\neg t}|\mathbf{r}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{z}_{\neg t})P(w_t)P(\mathbf{r}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{z}_{\neg t})} \propto \\ \frac{P(\mathbf{w}|\mathbf{r}, \mathbf{l}, \mathbf{z})P(\mathbf{r}, \mathbf{l}, \mathbf{z})}{P(\mathbf{w}_{\neg t}|\mathbf{r}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{z}_{\neg t})P(\mathbf{r}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{z}_{\neg t})} = \\ \frac{P(\mathbf{w}|\mathbf{r}, \mathbf{l}, \mathbf{z})}{P(\mathbf{w}_{\neg t}|\mathbf{r}_{\neg t}, \mathbf{l}_{\neg t}, \mathbf{z}_{\neg t})} \cdot \frac{P(\mathbf{r}|\mathbf{l}, \mathbf{z})}{P(\mathbf{r}_{\neg t}|\mathbf{l}_{\neg t}, \mathbf{z}_{\neg t})} \cdot \frac{P(\mathbf{l}|\mathbf{z})}{P(\mathbf{l}_{\neg t}|\mathbf{z}_{\neg t})} \cdot \frac{P(\mathbf{z})}{P(\mathbf{z}_{\neg t})} \end{gathered} \tag{4.15}$$

Предложенная модель является модификацией LDA, поэтому векторы документов и совместные векторы тем, тональных и проблемных меток порождаются распределениями Дирихле. Используя схожий способ сэмплирования переменных но основе распределений Дирихле, описанный в 4.2.1, скрытые параметры в модели TSPM(DP) могут быть выбраны по следующим формулам:

$$\begin{gathered} P(\mathbf{z}_t = k, \mathbf{l}_t = l, \mathbf{r}_t = r|\mathbf{w}_t = w, \mathbf{z}_{-t}, \mathbf{l}_{-t}, \mathbf{r}_{-t}, \alpha, \beta, \gamma, \eta) \propto \\ \frac{n^{-t}_{k,l,r,w} + \beta^w_{l,r}}{n^{-t}_{k,l,r} + \sum_{i'=1}^{V} \beta^{i'}_{l,r}} \frac{n^{-t}_{d,k,l,r} + \eta}{n^{-t}_{d,k,l} + R * \eta} \frac{n^{-t}_{d,k,l} + \gamma}{n^{-t}_{d,k} + L * \gamma} \frac{n^{-tj}_{d,k} + \alpha}{n^{-t}_{d} + K * \alpha}, \end{gathered} \tag{4.16}$$

где $n_{k,l,r,w}$ означает количество раз, когда слову $w$ присвоена тема $k$, тональная метка $l$ и проблемная метка $r$ в коллекции документов, $n_{k,l,r}$ определяет общее количество слов, которым присвоена тройка $(k,l,r)$. Общее число $n_{d,k,l,r}$ обозначает количество слов в документе $d$, присвоенных теме $k$, тональной метке $l$ и проблемной метке $r$, $n_{d,k,l}$ определяет количество слов документа $d$, присвоенных теме $k$ и тональной метке $l$. Общее число $n_{d,k}$ обозначает количество слов

в документе $d$, присвоенных теме $k$, $n_d$ обозначает общее число слов в документе $d$. Индекс $t = -(d,i)$ обозначает количество элементов, исключая текущие значения слова $w$, в документе $d$. Используя схожие обозначения в сэмплировании Гиббса, присвоенные скрытые параметры могут быть выбраны в модели TSPM(GP) по следующей формуле:

$$P(\mathbf{z}_t = k, \mathbf{l}_t = l, \mathbf{r}_t = r | \mathbf{w}_t = w, \mathbf{z}_{-(t)}, \mathbf{l}_{-(t)}, \mathbf{r}_{-(t)}, \alpha, \beta, \gamma, \eta) \propto \frac{n^{-t}_{k,l,r,w} + \beta^w_{l,r}}{n^{-(t}_{k,l,r} + \sum_{j=1}^{V} * \beta^j_{l,r}} \frac{n^{-(t)}_{k,l,r} + \eta}{n^{-(t}_{k,l} + R * \eta} \frac{n^{-(t)}_{d,k,l} + \gamma}{n^{-(t)}_{d,k} + L * \gamma} \frac{n^{-(t)}_{d,k} + \alpha}{n^{-(t)}_{d} + K * \alpha}, \tag{4.17}$$

где число $n_{k,l,r}$ обозначает количество слов в коллекции, присвоенных теме $k$, тональной метке $l$ и проблемной метке $r$, $n_{k,l}$ определяет количество слов в коллекции, присвоенных теме $k$ и тональной метке $l$.

Приближенное распределение тема-тональность-проблема $\phi_{k,l,r,j}$, определенное для корпуса, вычисляется по формуле:

$$\phi_{k,l,r,j} = \frac{n^{-t}_{k,l,r,w} + \beta^w_{l,r}}{n^{-(t}_{k,l,r} + \sum_{j=1}^{V} * \beta^j_{l,r}} \tag{4.18}$$

Приближенное распределение тем $\theta_{d,k}$, определенное для документа, вычисляется по формуле:

$$\theta_{d,k} = \frac{n^{-(t)}_{d,k} + \alpha}{n^{-(t)}_{d} + K * \alpha} \tag{4.19}$$

Приближенное совместное распределение тем и тональных меток $\pi_{d,k,l}$, определенное для документа, вычисляется по формуле:

$$\pi_{d,k,l} = \frac{n^{-(t)}_{d,k,l} + \gamma}{n^{-(t)}_{d,k} + L * \gamma} \tag{4.20}$$

Приближенное совместное распределение тем, тональных и проблемных меток $\xi_{d,k,l,r}$ в модели TSPM(DP), определенное для документа, вычисляется по формуле:

$$\xi_{d,k,l,r} = \frac{n^{-t}_{d,k,l,r} + \eta}{n^{-t}_{d,k,l} + R * \eta} \tag{4.21}$$

Приближенное совместное распределение тем и тональных меток $\xi_{k,l,r}$ в модели TSPM(GP), определенное для коллекции документа, вычисляется по формуле:

$$\xi_{k,l,r} = \frac{n_{k,l,r}^{-(t)} + \eta}{n_{k,l}^{-(t} + R * \eta} \tag{4.22}$$

Используя полученные оценки модели TSPM(DP), порождение слов в докумен-

**Algorithm 3:** Алгоритм порождения слов с помощью модели TSPM(DP).

```
1 Function sampling(корпус документов, α, β, γ, η)
      Input: гиперпараметры α, β, η, γ, корпус документов
      Output: присвоенные темы, тональные и проблемные метки для всех
              слов в корпусе
2     Инициализировать присвоение тем и меток для всех слов случайным
      образом
3     foreach i = 1 to максимальное количество итераций do
4         foreach документ d ∈ 1, ..., D do
5             foreach слово i ∈ 1, ..., N_{w_d} do
6                 Исключить слово i, присвоенное теме z, тональной метке l
                  и проблемной метке r, из счетчиков
                  n_{k,j,i,t}, n_{k,j,i}, n_{d,k,j,i}, n_{d,k,j}, n_{d,k} и n_d
7                 Сэмплировать новую пару (z', l', r') из уравнения 4.16
8                 Обновить счетчики n_{k,j,i,t}, n_{k,j,i}, n_{d,k,j,i}, n_{d,k,j}, n_{d,k} и n_d для
                  новой параметров (z', l', r')
```

тах происходит согласно Алгоритму 3.

## 4.4 Экспериментальное исследование

В данном разделе проводится экспериментальная оценка предложенных методов и сравнение с существующими вероятностными моделями для задач анализа мнений на шести наборах данных для русского и английского языка. Основная цель экспериментов – показать эффективность предложенных методов в задаче автоматической классификации проблемных высказываний с продуктами на уровне предложений отзывов и провести анализ качества по-

лученных тематических распределений слов. В качестве базовых методов используются существующие вероятностные методы, учитывающие информацию о тематической категории и тональности слов. На каждом из наборов данных определенной предметной области проводятся две серии экспериментов. В первой серии качество построенных тематических моделей оценивается с помощью стандартной меры качества тематических моделей - перплексии контрольных данных, которая связана с правдоподобием. Во второй серии экспериментов демонстрирует эффективность предложенный моделей для задачи классификации текстов на уровне предложений.

### 4.4.1 Наборы данных и критерии качества

Морфологическая обработка текста осуществлялась с помощью библиотеки NLTK для английского языка и Mystem для русского языка: на этапе предварительной обработки текстов была выполнена лемматизация и выделение корней слов (стемминг), удалены стоп-слова. Необходимость удаления стоп-слов объясняется следующим образом: стоп-слова как наиболее высокочастотные (фоновые) слова имеют высокую вероятность в большинстве тем, снижая релевантность тематической информации. Списки стоп-слов были взяты из пакета $stop - words$ для языка Python, затем из каждого списка были удалены слова из словаря Negation. Программный комплекс по построению тематических моделей написан на языке Java и выложен в открытый доступ[1]. Суммарное количество строк написанного кода – 14,000. В состав комплекса входят следующие модули: (i) модуль предварительной обработки данных (удаление стоп-слов, определение гиперпараметров на основе словарей); (ii) модуль тематического моделирования; (ii) модуль анализа результатов классификации с помощью моделей.

Обучающая коллекция для английского языка составлена следующим образом: из коллекции корпусов Amazon были выбраны отзывы с учетом тридцати наиболее часто встречающихся местоположений, указанных в профилях ста тысяч наиболее активных пользователей, для каждой предметной области.

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

Чтобы избежать разреженности тематик из отзывов были удалены слова, встречающиеся в корпусе менее пяти раз, и выбраны несколько тысяч отзывов для обучения случайным образом. К словам рядом с отрицанием поставлен префикс NEG. Таблица 19 содержит статистику по обучающей коллекции. В качестве контрольной выборки использовалась коллекция из главы 2.5.1.

Таблица 19 — Статистика обучающей коллекции

| Предметная область отзывов | Всего собрано отзывов | Количество отзывов с оценкой $r$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | r=1 | r=2 | r=3 | r=4 | r=5 | \|V\| |
| Электроника (англ. с Amazon) | 82127 | 1098 | 656 | 885 | 2170 | 5192 | 32853 |
| Детские товары (Amazon) | 3794 | 470 | 285 | 327 | 723 | 1989 | 7127 |
| Инструменты для дома (Amazon) | 32488 | 1040 | 602 | 857 | 1953 | 5539 | 19709 |
| Машины (Amazon) | 24720 | 915 | 491 | 682 | 1886 | 6027 | 14576 |
| Машины (рус. с SentiRuEval-2015) | 8271 | 40 | 159 | 671 | 3146 | 4254 | 29025 |
| Мобильные приложения (рус.) | 61182 | 2741 | 1308 | 1862 | 10597 | 10597 | 18768 |

Качество предложенных моделей оценивается с помощью нескольких критериев, предъявляемых к тематическим моделям для задач анализа мнений, согласно работам [77; 78]. Качество тематического моделирования слов оценивается с помощью перплексии контрольных данных, которая часто используется в задачах компьютерной лингвистики. Чем меньше эта величина, тем лучше модель предсказывает появление слов $w$ в документе $d$ и описывает распределения, скрытые в коллекции текстов. Она определяется следующим образом и тесно связана с правдоподобием модели [116]:

$$perplexity(D_{test}) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{D} N_d}\right) \tag{4.23}$$

После обучения модели вектора Ф, связанные с темами и множеством слов, фиксируются, а вектора $\theta_d$, непосредственно связанные с коллекцией документов, оцениваются по каждому документу контрольной выборки и вычисляется перплексия.

Для оценки вклада добавленных скрытых переменных, которые используются для описания взаимосвязей между темой и проблемными индикаторами в документах, модели оцениваются в рамках задачи классификации с помощью стандартных метрик качества, которые описаны в главе 2.

### 4.4.2 Детали реализации моделей

Предложенные модели используют словари следующих типов: словарь оценочных слов (SL) и словарь проблемных индикаторов (PL). В качестве тонального лексикона (SL) используется словарь MPQA , часто использующийся в работах по анализу мнений для английского языка; для русского языка используется словарь позитивных и негативных слов, расширенный синонимами и родственными словами из многофункциональный электронного словаря Викисловарь и описанный в главе 2.3. В качестве словаря проблемных индикаторов (PL) используются слова, описанные в главе 2.3 для русского и английского языка. На основе выбранных словарей задаются асимметричные априорные распределения Дирихле с гиперпараметром $\beta$. Начальные значения гиперпараметров $\beta_w$ для всех слов равны 0.01.

**Определение гиперпараметров предложенных моделей**

Определение гиперпараметров $\beta_{*w}$ заимствовано из работ [75; 77; 78]. Значения гиперпараметров $\beta_{lw}$ для эмоционально-окрашенных слов определяются следующим образом: если существует вхождение слова в лексикон SL с позитивной пометкой, то $\beta_{*w}$=(1,0.01,0.001) (1 для позитивных меток (pos), 0.01 для нейтральных (neu), 0.001 для негативных (neg)); для слов с негативной пометкой $\beta_{*w}$=(0.001,0.01,1). Аналогично определяются гиперпараметры $\beta_{rw}$ для проблемных индикаторов на английском языке: если существует вхождение слова в словарь PL с проблемной пометкой, то $\beta_{*w}$=(1,0.001) (1 для проблемных меток (pr), 0.001 для меток, указывающие на отсутствие проблемы (no-pr)); проблемный ин-

дикатор с префиксом отрицания получает значения $\beta_{*w} = (0.001,1)$. Для всех моделей при решении задачи статистического оценивания использовалось сэмплирование Гиббса, число итераций равно 1000. Все эксперименты проводились при следующих параметрах: $\alpha = 50/K, \beta = 0.01, \gamma = 0.01 \cdot \frac{AvgLen}{S}, \eta = 0.01 \cdot \frac{AvgLen}{R}$, где $AvgLen$ обозначает среднее количество слов в отзыве, $R = 2, S = 3, K = 5$.

**Идентификация высказываний, указывающих на проблемную ситуацию**

Для автоматической идентификации проблемных высказываний применяется вероятностный подход и оценивается $P(r|d)$, распределение вычисленных проблемных меток для документа $d$. Формально, сравниваются вероятность присвоения метки проблемных высказываний $P(r = pr|d)$ и вероятность присвоения метки с отсутствием проблемы $P(r = no - pr|d)$ для документа $d$. Документ классифицируется как проблемное высказывание, если $P(r = pr|d) > P(r = no - pr|d)$. Поскольку предложенные модели TSPM(DP) и TSPM(GP) не подсчитывают $P(r|d)$ в рамках оценивания, вероятности класса проблемного высказывания могут быть подсчитаны из распределения $\Phi$ следующим образом:

$$P(r|d) \propto P(d|r) = \prod_{w \in w_d} P(w|r) = \prod_{w \in w_d} \sum_{z=1}^{K} \sum_{s=1}^{S} P(w|r,s,z) \tag{4.24}$$

Вероятности $P(r|d)$ могут быть подсчитаны для модели TPrPhModel следующим образом:

$$P(r|d) \propto P(d|r) = \prod_{w \in w_d} P(w|r) = \prod_{w \in pw_d} \sum_{z=1}^{K} P(w|r,z) \tag{4.25}$$

Таким образом, численное значение $r(s_{ij}) = P(r|d), r \in \{pr, no - pr\})$, равное вероятности принадлежности предложения $s_{ij}, i \in \{1, \cdots, |D|\}$, j $\in \{1, \cdots, |d_i|\}$.

### 4.4.3 Эксперименты и результаты

В качестве базовых алгоритмов для сравнения с предложенными моделями в экспериментах были выбраны следующие тематические модели, наиболее распространенные в задачах анализа мнений пользователей и моделирующие слова или предложения в документе в зависимости от тематической и тональной переменных:

- модель *joint sentiment-topic model* (JST), в которой каждой тональной переменной слова соответствует мультиномиальное распределение в пространстве тем [77];
- модель *reverse joint sentiment-topic model* (Reverse-JST), в которой каждой теме слова соответствует мультиномиальное распределение в пространстве тональных переменных [77];
- модель *aspect and sentiment unification model* (ASUM), в которой каждой тональной переменной на уровне предложения соответствует мультиномиальное распределение в пространстве тем [75];
- модель *user-aware sentiment topic model* (USTM), включающая в распределения метаданные профайлов пользователей, где каждой комбинации тэга соответствует мультиномиальное распределение в пространстве тем, а каждой теме соответствует мультиномиальное распределение в пространстве тональностей [78].

Для каждой JST, Reverse-JST, ASUM, USTM мы использовали оба словаря SL и PL независимо: префикс '+SL' свидетельствует, что модель учитывает только тональные метки слов и задает гиперпараметры $\beta_{lw}$ ($S = 3$) на основе словаря SL; префикс '+PL' указывает на то, что модель учитывает только проблемные метки слов и задает гиперпараметры $\beta_{lw}$ ($R = 2$) на основе словаря PL. Для моделей, учитывающих только тональные метки слов (с префиксом '+SL'), используется следующее предположение: высказывание считается проблемным, если вероятность негативного класса $P(l = neg|d)$ выше, чем вероятность позитивного и нейтрального классов: $P(l = pos|d)$ и $P(l = neu|d)$; аналогично высказывание не содержит проблем с продуктами, если $P(l = pos|d)$ выше $P(l = neg|d)$ и $P(l = neu|d)$. Вероятности тональных классов для JST, Reverse-JST, ASUM, USTM вычисляются на основе мультиномиального распределения

в пространстве слов $\Phi_{z,l}$ по схожей формуле, описанной выше для предложенных моделей. Для USTM число различных пользовательских метаданных о географическом местоположении пользователя (T) равно 25.

В качестве критерия качества построенных моделей используется перплексия контрольных данных: 90% отзывов использовано в качестве обучающей выборки для вероятностных моделей, 10% отзывов использованы для тестирования. Результаты экспериментов по оценке качества модели представлены в Таблице 20. Поскольку мета-данные об авторе отзывов отсутствуют для русского языка, результаты USTM для отзывов для русского языка не описаны. Модели TSPM(DP) и TSPM(GP) показываются наименьшие значения перплексии по сравнению с моделями JST и Reverse-JST, где каждое слово документа выбирается для пары (тема, тональность) без дополнительных параметров. Таким образом, добавление скрытой проблемной переменной, условно зависимой от темы и тональности слова, не ухудшает качество моделей. Предложенная модель TPrPhModel показывает наименьшие значения перплексии среди всех тематических моделей, что характеризует лучшую способность TPrPhModel предсказывать появление проблемных индикаторов $pw$ и контекстных слов $w$ в документах коллекции в зависимости от темы.

Таблица 20 — 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 |
| USTM+SL | 1764.293 | 880.766 | 420.289 | 547.784 | - | - |
| 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 |
| USTM+PL | 1840.842 | 814.009 | 361.743 | **504.590** | - | - |
| 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 |

Результаты классификации представлены в Таблицах 21, 22 и 23; в качестве критериев использовались следующие метрики: достоверность, точность,

Таблица 21 — Результаты классификации предложений отзывов об инструментах и детских товарах на английском языке

| Метод | Инструменты (англ.) | | | | Детские товары (англ.) | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | P | R | F | Acc. | P | R | F |
| JST+SL | .436 | .758 | .317 | .448 | .461 | .675 | .386 | .468 |
| R.-JST+SL | .396 | .674 | .311 | .426 | .431 | .669 | .276 | .391 |
| ASUM+SL | .459 | .757 | .363 | .491 | .511 | .716 | .439 | .544 |
| USTM+SL | .536 | **.766** | .511 | .612 | .500 | **.797** | .328 | .465 |
| JST+PL | .649 | .762 | .745 | .753 | .606 | .737 | .628 | .678 |
| R.-JST+PL | .615 | .738 | .720 | .729 | .537 | .676 | .577 | .676 |
| ASUM+PL | .536 | .746 | .538 | .625 | .590 | .722 | .625 | .669 |
| USTM+PL | .635 | .722 | .798 | .759 | .559 | .702 | .579 | .635 |
| TPrPhModel | .607 | .649 | .768 | .704 | **.616** | .734 | .656 | **.693** |
| TSPM(DP) | .620 | .763 | .684 | .721 | .587 | .700 | **.659** | .679 |
| TSPM(GP) | **.659** | .730 | **.831** | **.778** | .508 | .749 | .386 | .509 |

Таблица 22 — Результаты классификации предложений отзывов пользователей об электронике и машинах на английском языке

| Метод | Электроника (англ.) | | | | Машины (англ.) | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | P | R | F | Acc. | P | R | F |
| JST+SL | .557 | .511 | .187 | .274 | .394 | .823 | .342 | .483 |
| R.-JST+SL | .516 | .450 | .383 | .414 | .375 | .847 | .300 | .443 |
| ASUM+SL | .419 | .293 | .198 | .236 | .407 | **.856** | .343 | .490 |
| USTM+SL | .511 | .433 | .303 | .356 | .506 | .819 | .218 | .344 |
| JST+PL | .521 | .469 | .544 | .503 | .648 | .844 | .706 | .769 |
| R.-JST+PL | .521 | .471 | .577 | .518 | .647 | .834 | .717 | .771 |
| ASUM+PL | .554 | .509 | .580 | .542 | .609 | .734 | .647 | .687 |
| USTM+PL | .477 | .459 | **.961** | **.621** | .669 | .819 | **.771** | .794 |
| TPrPhModel | **.568** | **.515** | .570 | .541 | .612 | .846 | .651 | .736 |
| TSPM(DP) | .566 | .511 | .726 | .599 | .539 | .824 | .565 | .671 |
| TSPM(GP) | .466 | .441 | .738 | .552 | **.680** | .839 | .760 | **.798** |

полнота, посчитанные как среднее арифметическое значение после 5 прогонов алгоритма. Модели JST+SL, Reverse-JST+SL, ASUM+SL, USTM+SL показали наименьшие значения F-меры и достоверности классификации по сравнению с JST+PL, Reverse-JST+PL, ASUM+PL, USTM+PL, соответственно, что опровергает взаимно-однозначное соответствие негативного класса и класса проблемных высказываний. Наилучшие результаты по F-мере достигают предложенные модели TSPM(DP) и TSPM(GP) по сравнению с базовыми алгоритмами на корпусе отзывов на английском языке, что показывает эффективность

Таблица 23 — Результаты классификации предложений отзывов о машинах и мобильных приложениях на русском языке языке

| Метод | Машины (рус.) | | | | Мобильные приложения | | | |
|---|---|---|---|---|---|---|---|---|
| | Acc. | P | R | F | Acc. | P | R | F |
| JST+SL | .614 | .272 | .472 | .345 | .651 | .508 | .359 | .421 |
| R.-JST+SL | .617 | .277 | .488 | .354 | .656 | .327 | .651 | .436 |
| ASUM+SL | **.666** | **.302** | .421 | .352 | .608 | .748 | .481 | .586 |
| JST+PL | .547 | .249 | .550 | .343 | .528 | .588 | .605 | .596 |
| R.-JST+PL | .546 | .262 | .615 | .368 | .552 | .436 | **.916** | .590 |
| ASUM+PL | .498 | .232 | .579 | .331 | .656 | .669 | .798 | .728 |
| TPrPhModel | .577 | .291 | **.767** | **.422** | **.718** | .746 | .756 | **.751** |
| TSPM(DP) | .517 | .260 | .674 | .376 | .675 | .729 | .695 | .711 |
| TSPM(GP) | .571 | .257 | .661 | .370 | .665 | **.755** | .686 | .719 |

порождения слова в документах в зависимости от некоторой скрытой темы, тональной и проблемной информации. Наилучшие результаты по F-мере достигают модель TPrPhModel по сравнению с другими алгоритмами на корпусе отзывов на русском языке, что подтверждает эффективность разделения проблемных и контекстных слов для задачи классификации.

Рисунок 4.3 содержит результаты классификации предложенных алгоритмов и базовых моделей JST-PL и Reverse-JST-PL для различного количества тем в моделях. Графики во всех доменах свидетельствуют, что предложенные модели показывают лучшие значения достоверности на 5 темах по сравнению с моделями из 1 темы, что согласуется с результатами классификации, описанными в работе [77]. Это подтверждает, отзыв пользователя по структуре относится к типу связного текста из нескольких подтем и совместное моделирование темы и проблемных переменных помогает улучшить классификации предложений. Модели с увеличенным количеством тем (до 25ти) не показывают значимый прирост результатов в отзывах о детских продуктах и инструментах. В области электроники результаты достоверности возрастают на 7.8% для модели TSPM(DP). Для отзывов о машинах на английском языке результаты достоверности возрастают незначительно (на 1.2%) для модели TSPM(GP) с $K = 25$, по сравнению с TSPM(GP) с $K = 10$. Для модели TSPM(GP) наблюдается прирост значений достоверности в 5-9% на 10 темах по сравнению с 5 темами для отзывов об автомобилях и о домашних инструментов на английском языке. Модель TSPM(GP) показывает лучшее значение достоверности класси-

фикации по сравнению с TSPM(DP) для отзывов о функциональных (механических) продуктах (автомобили, инструменты) с более общими проблемными высказываниями для всех пользователей. TSPM(DP) показывает лучшее значение достоверности классификации по сравнению с TSPM(GP) для отзывов о высокотехнологичных продуктах (электроника, приложения), где проблемные ситуации более специфичны и в большей мере зависят от пользователя.

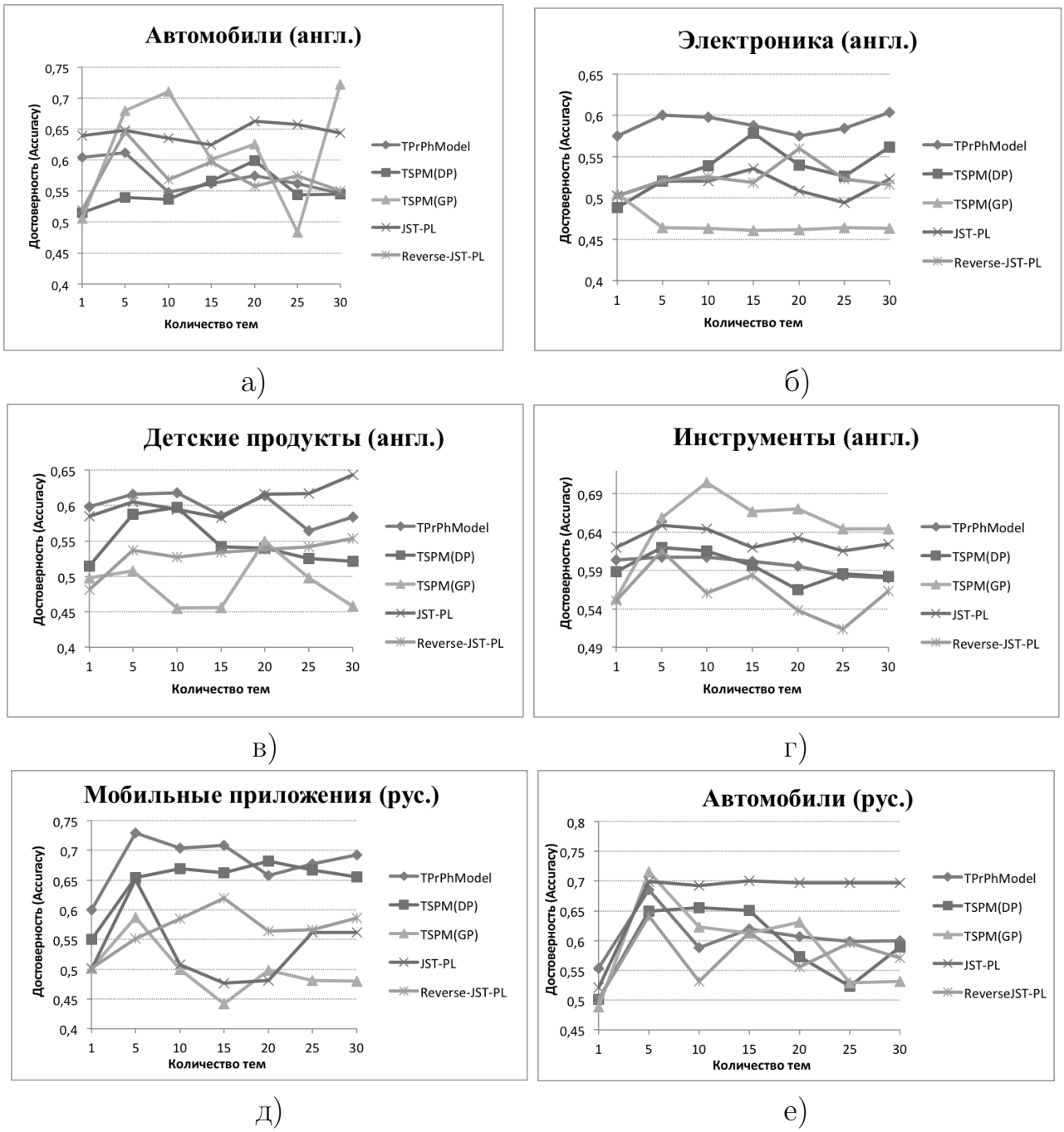



Рисунок 4.3 — Результаты классификации текстов пользователей для вероятностных моделях, обученных на разном количестве тем

Таблица 24 — Примеры тем слов в TPrPhModel для отзывов об автомобилях на русском языке.

| пример темы | | | пример темы | | | пример темы | | |
|---|---|---|---|---|---|---|---|---|
| объект | слово с меткой | | объект | слово с меткой | | объект | слово с меткой | |
| | no-pr. | probl. | | no-pr. | probl. | | no-pr. | probl. |
| впечатление | общий | общий | авто | хороший | NEG машина | машина | ездить | плохой |
| салон | отличный | минус | запчасть | купить | NEG авто | машинка | купить | ремонт |
| дорога | хороший | плохой | ваз | новый | владеть | семья | маленький | ходовой |
| город | приобретать | маленький | деньги | отечественный | дешевый | муж | большой | ездить |
| расход | советовать | слабый | иномарка | российский | общий | вариант | нравиться | простой |
| багажник | брать | недостаток | автопром | думать | мягкий | ребенок | удобный | дешевый |
| проходимость | просторный | тяжело | ничто | сделать | подводить | друг | устраивать | ломаться |
| поездка | пожалеть | проехать | цена | взять | работать | дача | выбирать | советский |
| топливо | вместительный | шумный | калина | покупать | NEG лишний | рыбалка | спокойно | старый |
| бензин | удобный | греметь | универсал | русский | NEG всякий | коляска | супер | убивать |

Дополнительно к результатам эффективности построенных моделей, проведен качественный анализ полученных тематических распределений моделей TPrPhModel, TSPM(DP), TSPM(GP) для отзывов пользователей о машинах на русском языке. Для модели TPrPhModel, использующей информацию о частях речи слов, приводится примеры полученных тем, состоящих из слов, в Таблице 24. Первая тема описывает характеристики машины (“расход”, “салон”, “багажник”), где у пользователя возникают претензии по поводу комфортности вождения (“шумный”, “слабый”) в то время, как авторы отзывов удовлетворены выбранными габаритами машины и салоном (“вместительный”, “удобный”). Вторая тема описывает ситуации с заменой запчастей и возможными затратами (“запчасть”, “деньги”). У пользователей возникают проблемные ситуации с необходимостью замены деталей (“подводить”, “NEG машина”), однако пользо-

ватели отрицают проблемы из-за качества автомобиля (“отечественный”, “купить”). Третья тема описывает ситуацию использования (“машина”, “дача”), где у владельцев возникают проблемы из-за технических неисправностей (“ремонт”, “убивать”), но не с самой ситуацией (“ездить”, “супер”).

Таблица 25 — Примеры тем, основанных на словосочетаниях в TSPM(DP), для отзывов об автомобилях на русском языке

| Примеры тем словосочетаний об автомобилях | | | | | |
|---|---|---|---|---|---|
| нейтральный | | позитивный | | негативный | |
| no-problem | problem | no-problem | problem | no-problem | problem |
| NEG проблема | чудо техника | двигатель резвый | существенный недостаток | начинать греметь | предел тыс |
| общий впечатление | проблема запчасть | разница почувствовать | сидение удобный | авто плохой | ремонт обходиться |
| новый машина | владелец машина | плавный линия | NEG дребезжать | плохой качество | ездить проблема |
| масло фильтр | высоко сидеть | появляться возможность | становиться резвый | предпродажный подготовка | быстро изнашиваться |
| купить новый | zaz vida | двигатель надежный | выглядеть привлекательный | ремонт двигатель | план ремонт |
| тысяча рубль | NEG греметь | становиться тихо | отличный рулевой | неподходящий момент | поломка машина |
| ремень грм | ремонт ходовой | дорожный ситуация | слышный шум | холодный зима | немного маловат |
| менять масло | жалко продавать | литр хороший | подводить салон | смотреть машинка | тяговитый двигатель |
| тысяча километр | деньги ремонт | машинка понравиться | большой вместимость | вести машина | серьезный поломка |
| отечественный автопром | приходиться искать | отличный обзор | заводиться отлично | NEG проблема | плохо работать |

Таблица 26 — Примеры тем, основанных на словосочетаниях в TSPM(GP), для отзывов об автомобилях на русском языке

| Примеры тем словосочетаний об автомобилях | | | | | |
|---|---|---|---|---|---|
| нейтральный | | позитивный | | негативный | |
| no-problem | problem | no-problem | problem | no-problem | problem |
| общий впечатление | пожалеть выбор | деталь двигатель | возможность купить | китайский машина | капля жалеть |
| хороший автомобиль | возникать проблема | красивый удобный | компактный габарит | правда говорить | серьезный ремонт |
| цена качество | давать сбой | легко быстро | разумный предел | прежде покупать | минус передний |
| автомобиль ваз | мелкий царапина | нужный рабочий | качественный материал | грязь снег | мороз минус |
| вместительный багажник | мелкий ремонт | комфортный место | вместимый багажник | капитальный ремонт | слабый задний |
| просторный салон | практически ломаться | экономия бензин | NEG хватать | дальний растояние | шумоизоляция слабоватый |
| купить автомобиль | зима проблема | мощный двигатель | машина тепло | автомобиль забывать | машина тяжелый |
| рабочий лошадка | плохой бюджетный | удобно ребенок | внушительный габарит | средство передвижение | NEG садиться |
| внешний вид | ремонт практически | ходовой качество | выносливый двигатель | NEG дизайн | NEG брать |
| автомобиль общий | NEG чувствовать | удобный ездить | волга ужас | пробивать колесо | условие подводить |

Основная цель объединения информации о темах отзыва, тональности слов и проблемных индикаторов – суммировать мнения с различной тональностью для лучшего понимания причин снижения удовлетворенности потребителя с связи с проблемными высказываниями о продукте. Для более детального анализа моделей TSPM(DP), TSPM(GP) приводятся примеры словосочетаний

(пар последовательных слов) как более информативных лексических единиц текста. Например, по фрагменту текста “зато у этого автомобиля отличный салон”, который может быть представлен как мешок слов “автомобиль”, “отличный”, “салон”, трудно сказать, пользователь считает отличным сам автомобиль в целом или его салон. Таблицы 25 и 26 содержат примеры распределения словосочетаний в обнаруженных моделями темах. Чтобы избежать разреженности тематик из отзывов были удалены словосочетания, встречающиеся в корпусе менее пяти раз.

Ряд наблюдений может быть сделан на основе представленных тем. Как показано в Таблице 25 для модели TSPM(DP), общее мнение покупателя может быть позитивным, и он может быть доволен внешним видом автомобиля и его ходовыми качествами (“заводиться отлично”, “выглядеть привлекательный”), однако испытывать дискомфорт в использовании продукта (“слышный шум”, “становиться резвый”). Напротив, покупатель с негативным мнением испытывает технические трудности в использовании автомобиля (“быстро изнашиваться”, “поломка машина”), однако фокус негативного мнения может быть связан с самой ситуацией, нежели с проблемами автомобиля (“холодный зима”, “неподходящий момент”). Нейтральное мнение пользователь высказывает о покупке автомобиля или плановой замене деталей (“купить новый”, “менять масло”), однако проблемные высказывания могут быть сделаны о продаже автомобиля с предварительным ремонтом (“ремонт ходовой”, “жалко продавать”).

Как показано в Таблице 26, модель TSPM(GP), агрегирует проблемные высказывания по коллекции документов, с целью выявить наиболее общие технические неполадки и проблемы с удобством использования. Примеры тем значительно отличаются от тем модели TSPM(DP), описанных ранее. Распределение словосочетаний по темам становится более общим, частные ситуации с продуктами не описываются. Пользователи с позитивным мнением описывают автомобиль как удобный, качественный, вместительный и экономный, однако, некоторые покупатели наблюдают нехватку функционала (“машина тепло”, “NEG хватать”). Пользователи с негативным мнением описывают проблемы с ходовыми качествами и с комфортом поездки в то время, как пользователи отрицают проблемные ситуации из-за внешних причин: производителя автомобиля (“китайский машина”), погоды (“грязь снег”, “дальний расстояние”). Нейтральные мнения связаны с покупкой автомобиля и характеристиками продук-

та (“рабочий лошадка”, “вместительный багажник”) в то время, как проблемные высказывания связаны с объективными неполадками (“мелкий царапина”, “плохой бюджетный”). Примеры словосочетаний, смоделированные в TSPM(GP), могут быть использованы для расширения словаря проблемных индикаторов в определенной предметной области.

## 4.5 Выводы к четвертой главе

Целью главы является резюмирование слов в отзывах пользователей по тематическим группам, то есть идентификация $k$ основных тем отзывов, где тема определена как множество слов в текстах, которые имеют тенденцию встречаться в совокупности с проблемными индикаторами, эмоционально-окрашенными (тональными) словами пользователя и аспектами продуктов. Для достижения целей в диссертации представлены тематические модели на основе модели латентного размещения Дирихле: (i) модель тематических проблемных высказываний (problem phrase topic model, TPrPhModel); (ii) модель тема-тональность-проблема (topic-sentiment-problem model, TSPM). Предложенные модели используют знания о проблемных индикаторах в качестве асимметричных гиперпараметров для всех слов документов. Анализ результатов модели TPrPhModel в качественном отношении полученных тематических распределений отражает различия между проблемными индикаторами применительно к различных целевым объектам продукта. Анализ результатов модели TSPM подтверждает необходимость идентификации тональности высказывания для более качественного определения проблемных высказываний о продукте: ряд экспериментов показал, что проблемные высказывания о технических поломках или сложность функциональности устройств носят негативный оттенок в то время, как или отсутствие функций, или нехватка носят нейтральный или позитивный суммарный оттенок и не вызывают у пользователя резкой неудовлетворенности продукцией. Оценка качества классификации с помощью предложенных методов анализируются в сравнении с результатами популярных модификаций латентного размещения Дирихле для задач анализа мнений. Предложенные модели

достигают наилучшие результаты F-меры и значения перплексии в сравнении с другими вероятностными моделями.

# Заключение

Основные результаты работы заключаются в следующем.

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, связанных 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 ряд скрытых переменных для описания тем и проблемных индикаторов совместно.
4. Разработано программное обеспечение и проведено экспериментальное исследование, доказывающее улучшение качества предложенных методов по сравнению с существующими алгоритмами.

Дальнейшие перспективы развития исследований могут быть связаны (i) с задачей улучшения качества предложенных методов классификации с помощью узкоспециализированных словарей и глубокого семантического анализа сложных предложений; (ii) с задачей улучшения качества предложенных тематических моделей с помощью оптимизации гиперпараметров для семантически связанных слов в процессе обучения. Предложенные методы и модели могут быть использованы для более качественного решения задач анализа мнений, включая кластеризацию и классификацию высказываний по тематическим категориям, определение рейтинга продукта в отзыве, определения показателей прироста продаж продукта на основе коллекции отзывов.

## Список литературы


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.
2. *Anderson E. W.* Customer satisfaction and word of mouth // Journal of service research. — 1998. — Т. 1, № 1. — С. 5—17.
3. Spreading the word: Investigating antecedents of consumers' positive word-of-mouth intentions and behaviors in a retailing context / T. J. Brown [и др.] // Journal of the Academy of Marketing Science. — 2005. — Т. 33, № 2. — С. 123—138.
4. Quality and reliability problems from a consumerś perspective: an increasing problem overlooked by businesses? / E. den Ouden [и др.] // Quality and Reliability Engineering International. — 2006. — Т. 22, № 7. — С. 821—838.
5. Automatically assessing review helpfulness / S.-M. Kim [и др.] // Proceedings of the 2006 Conference on empirical methods in natural language processing. — Association for Computational Linguistics. 2006. — С. 423—430.
6. Usability of consumer-related information sources for design improvement / G. Thiruvenkadam [и др.] // Professional Communication Conference, 2008. IPCC 2008. IEEE International. — IEEE. 2008. — С. 1—7.
7. Improving product quality and reliability with customer experience data / A. Brombacher [и др.] // Quality and Reliability Engineering International. — 2012. — Т. 28, № 8. — С. 873—886.
8. 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.
9. No Fault Found events in maintenance engineering Part 2: Root causes, technical developments and future research / S. Khan [и др.] // Reliability Engineering & System Safety. — 2014. — Т. 123. — С. 196—208.

10. *Pang B.*, *Lee L.* Opinion mining and sentiment analysis // Foundations and trends in information retrieval. — 2008. — Т. 2, 1-2. — С. 1—135.

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

12. *Tsytsarau M.*, *Palpanas T.* Survey on mining subjective data on the web // Data Mining and Knowledge Discovery. — 2012. — Т. 24, № 3. — С. 478—514.

13. Sentiment analysis in Twitter / E. Martinez-Cámara [и др.] // Natural Language Engineering. — 2014. — Т. 20, № 01. — С. 1—28.

14. Extracting Verb Expressions Implying Negative Opinions. / H. Li [и др.] // AAAI. — 2015. — С. 2411—2417.

15. *De Saeger S.*, *Torisawa K.*, *Kazama J.* Looking for trouble // Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. — Association for Computational Linguistics. 2008. — С. 185—192.

16. *Iacob C.*, *Harrison R.*, *Faily S.* Online reviews as first class artifacts in mobile app development // Mobile Computing, Applications, and Services. — Springer, 2013. — С. 47—53.

17. *Iacob C.*, *Harrison R.* Retrieving and analyzing mobile apps feature requests from online reviews // Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on. — IEEE. 2013. — С. 41—44.

18. *Moghaddam S.* Beyond Sentiment Analysis: Mining Defects and Improvements from Customer Feedback // Advances in Information Retrieval. — Springer, 2015. — С. 400—410.

19. *Gupta N. K.* Extracting descriptions of problems with product and services from twitter data // Proceedings of the 3rd Workshop on Social Web Search and Mining (SWSM2011). Beijing, China. — 2011.

20. *Hedegaard S.*, *Simonsen J. G.* Extracting usability and user experience information from online user reviews // Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. — ACM. 2013. — С. 2089—2098.

21. *Gupta N. K.* Extracting phrases describing problems with products and services from twitter messages // Computación y Sistemas. — 2013. — Т. 17, № 2. — С. 197—206.

22. *Kiritchenko S.*, *Zhu X.*, *Mohammad S. M.* Sentiment analysis of short informal texts // Journal of Artificial Intelligence Research. — 2014. — С. 723—762.

23. *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.

24. *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.

25. *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.

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

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

28. *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.

29. 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.

30. 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.

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

32. *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.

33. *Liu B.* Sentiment analysis: Mining opinions, sentiments, and emotions. — Cambridge University Press, 2015.

34. *Dave K.*, *Lawrence S.*, *Pennock D. M.* Mining the peanut gallery: Opinion extraction and semantic classification of product reviews // Proceedings of the 12th international conference on World Wide Web. — ACM. 2003. — С. 519—528.

35. *Das S.*, *Chen M.* Yahoo! for Amazon: Extracting market sentiment from stock message boards // Proceedings of the Asia Pacific finance association annual conference (APFA). Т. 35. — Bangkok, Thailand. 2001. — С. 43.

36. *Tong R. M.* An operational system for detecting and tracking opinions in on-line discussion // Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification. Т. 1. — 2001. — С. 6.

37. *Turney P. D.* Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews // Proceedings of the 40th annual meeting on association for computational linguistics. — Association for Computational Linguistics. 2002. — С. 417—424.

38. *Pang B.*, *Lee L.*, *Vaithyanathan S.* Thumbs up?: sentiment classification using machine learning techniques // Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. — Association for Computational Linguistics. 2002. — С. 79—86.

39. *Prentice S.*, *Huffman E.* Social medias new role in emergency management // Idaho National Laboratory. — 2008. — С. 1—5.

40. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. / B. OĆonnor [и др.] // ICWSM. — 2010. — Т. 11, 122-129. — С. 1—2.

41. *Nguyen T. H.*, *Shirai K.* Topic modeling based sentiment analysis on social media for stock market prediction // Proceedings of the 53rd Annural Meeting of the Association for Computational Linguistics. — 2015.

42. Detecting sadness in 140 characters: Sentiment analysis of mourning michael jackson on twitter / E. Kim [и др.] // Web Ecology. — 2009. — Т. 3. — С. 1—15.

43. *Fan T.-K.*, *Chang C.-H.* Sentiment-oriented contextual advertising // Knowledge and Information Systems. — 2010. — Т. 23, № 3. — С. 321—344.

44. Graphical modeling of macro behavioral targeting in social networks / Y. Xie [и др.] // Proceedings of SDM. — SIAM. 2013.

45. *Mullen T.*, *Collier N.* Sentiment Analysis using Support Vector Machines with Diverse Information Sources. // EMNLP. Vol. 4. — 2004. — Pp. 412–418.

46. *Ng V.*, *Dasgupta S.*, *Arifin S.* Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews // Proceedings of the COLING/ACL. — Association for Computational Linguistics. 2006. — Pp. 611–618.

47. *Kennedy A.*, *Inkpen D.* Sentiment classification of movie reviews using contextual valence shifters // Computational intelligence. — 2006. — Vol. 22, no. 2. — Pp. 110–125.

48. *Xia R.*, *Zong C.* Exploring the use of word relation features for sentiment classification // Proceedings of the 23rd International Conference on Computational Linguistics: Posters. — Association for Computational Linguistics. 2010. — Pp. 1336–1344.

49. Structure-aware review mining and summarization / F. Li [et al.] // Proceedings of the 23rd international conference on computational linguistics. — Association for Computational Linguistics. 2010. — Pp. 653–661.

50. Improving blog polarity classification via topic analysis and adaptive methods / F. Liu [et al.] // Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computa-

tional Linguistics. — Association for Computational Linguistics. 2010. — Pp. 309–312.

51. *Васильев В. Г.*, *Худякова М. В.*, *С. Д.* Классификация отзывов пользователей с использованием фрагментных правил // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции “Диалог”. — 2012. — Т. 11, № 18. — С. 66—76.

52. *Рубцова Ю.* Метод построения и анализа корпуса коротких текстов для задачи классификации отзывов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XV Всероссийской научной конференции RCDL. — 2013. — С. 269—275.

53. *Клековкина М.*, *Котельников Е.* Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды XIV Всероссийской научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” (RCDL-2012, Переславль-Залесский, 15-18 октября 2012 г.) — 2012. — С. 118—123.

54. *Фролов А.*, *Поляков П.Ю.and Плешко В.* Использование семантических категорий в задаче классификации отзывов о книгах // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции “Диалог”. — 2013. — Т. 12, № 19.

55. *Поляков П. Ю.*, *Калинина М. В.*, *Плешко В. В.* Исследование применимости методов тематичес- кой классификации в задаче классификации отзывов о книгах // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции “Диалог”. — 2012. — Т. 11, № 18. — С. 51—59.

56. Research of lexical approach and machine learning methods for sentiment analysis / P. Blinov [и др.] // Computational Linguistics and Intellectual Technologies. — 2013. — Т. 2, № 12. — С. 48—58.

57. An Empirical Study on the Effect of Negation Words on Sentiment. / X. Zhu [и др.] // ACL (1). — 2014. — С. 304—313.

58. *Hatzivassiloglou V.*, *Wiebe J. M.* Effects of adjective orientation and gradability on sentence subjectivity // Proceedings of the 18th conference on Computational linguistics-Volume 1. — Association for Computational Linguistics. 2000. — Pp. 299–305.

59. *Yu H.*, *Hatzivassiloglou V.* Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences // Proceedings of the 2003 conference on Empirical methods in natural language processing. — Association for Computational Linguistics. 2003. — Pp. 129–136.

60. *Riloff E.*, *Wiebe J.* Learning extraction patterns for subjective expressions // Proceedings of the 2003 conference on Empirical methods in natural language processing. — Association for Computational Linguistics. 2003. — Pp. 105–112.

61. Learning subjective language / J. Wiebe [et al.] // Computational linguistics. — 2004. — Vol. 30, no. 3. — Pp. 277–308.

62. *Riloff E.*, *Patwardhan S.*, *Wiebe J.* Feature subsumption for opinion analysis // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics. 2006. — Pp. 440–448.

63. *Wilson T.*, *Wiebe J.*, *Hwa R.* Recognizing strong and weak opinion clauses // Computational Intelligence. — 2006. — Vol. 22, no. 2. — Pp. 73–99.

64. *Montoyo A.*, *Martinez-Barco P.*, *Balahur A.* Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments // Decision Support Systems. — 2012. — Vol. 53, no. 4. — Pp. 675–679.

65. *Balahur A.*, *Mihalcea R.*, *Montoyo A.* Computational approaches to subjectivity and sentiment analysis: Present and envisaged methods and applications // Computer Speech & Language. — 2014. — Vol. 28, no. 1. — Pp. 1–6.

66. Finding deceptive opinion spam by any stretch of the imagination / M. Ott [et al.] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. — Association for Computational Linguistics. 2011. — Pp. 309–319.

67. *Zhang Z.*, *Varadarajan B.* Utility scoring of product reviews // Proceedings of the 15th ACM international conference on Information and knowledge management. — ACM. 2006. — С. 51—57.

68. *Jindal N.*, *Liu B.* Opinion spam and analysis // Proceedings of the 2008 International Conference on Web Search and Data Mining. — ACM. 2008. — С. 219—230.

69. *Barbieri F.*, *Saggion H.* Modelling Irony in Twitter. // EACL. — 2014. — С. 56—64.

70. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. / E. Riloff [и др.] // EMNLP. — 2013. — С. 704—714.

71. *Hu M.*, *Liu B.* Mining and summarizing customer reviews // Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2004. — Pp. 168–177.

72. *Polanyi L.*, *Zaenen A.* Contextual valence shifters // Computing attitude and affect in text: Theory and applications. — Springer, 2006. — Pp. 1–10.

73. Lexicon-based methods for sentiment analysis / M. Taboada [и др.] // Computational linguistics. — 2011. — Т. 37, № 2. — С. 267—307.

74. *Popescu A.-M.*, *Nguyen B.*, *Etzioni O.* OPINE: Extracting product features and opinions from reviews // Proceedings of HLT/EMNLP on interactive demonstrations. — Association for Computational Linguistics. 2005. — Pp. 32–33.

75. *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. — Hong Kong, China : ACM, 2011. — С. 815—824. — (WSDḾ11). — ISBN 978-1-4503-0493-1. — DOI: 10.1145/1935826.1935932. — URL: http://doi.acm.org/10.1145/1935826.1935932.

76. *Moghaddam S.*, *Ester M.* On the design of LDA models for aspect-based opinion mining // Proceedings of the 21st ACM international conference on Information and knowledge management. — ACM. 2012. — Pp. 803–812.

77. 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. — DOI: 10.1109/TKDE.2011.48.

78. Parametric and non-parametric user-aware sentiment topic models / Z. Yang [и др.] // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. — ACM. 2015. — С. 413—422.

79. *Wang H.* Sentiment-aligned Topic Models for Product Aspect Rating Prediction: дис. ... канд. / Wang Hao. — Applied Sciences: School of Computing Science, 2015.

80. *Chetviorkin I.*, *Loukachevich N.* Research of lexical approach and machine learning methods for sentiment analysis // Proceedings of International Conference Dialog. — 2013. — Т. 2. — С. 40—50.

81. *Joachims T.* Making large scale SVM learning practical: tech. rep. / Universität Dortmund. — 1999.

82. *Cristianini N.*, *Shawe-Taylor J.* An introduction to support vector machines and other kernel-based learning methods. — Cambridge university press, 2000.

83. *Boiy E.*, *Moens M.-F.* A machine learning approach to sentiment analysis in multilingual Web texts // Information retrieval. — 2009. — Т. 12, № 5. — С. 526—558.

84. *Chetviorkin I.*, *Braslavskiy P.*, *Loukachevich N.* Research of lexical approach and machine learning methods for sentiment analysis // Computational Linguistics and Intellectual Technologies. — 2012. — Т. 2. — С. 1—14.

85. *Turney P. D.*, *Littman M. L.* Measuring praise and criticism: Inference of semantic orientation from association // ACM Transactions on Information Systems (TOIS). — 2003. — Vol. 21, no. 4. — Pp. 315–346.

86. *Choi Y.*, *Cardie C.* Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. — Association for Computational Linguistics. 2009. — Pp. 590–598.

87. Improving Opinion Retrieval Based on Query-Specific Sentiment Lexicon. / S.-H. Na [et al.] // ECIR. Vol. 9. — Springer. 2009. — Pp. 734–738.

88. *Neviarouskaya A.*, *Prendinger H.*, *Ishizuka M.* Sentiful: Generating a reliable lexicon for sentiment analysis // Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. — IEEE. 2009. — Pp. 1–6.

89. *Mohammad S.*, *Dunne C.*, *Dorr B.* Generating high-coverage semantic orientation lexicons from overtly marked words and a thesaurus // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2. — Association for Computational Linguistics. 2009. — Pp. 599–608.

90. *Hatzivassiloglou V.*, *McKeown K. R.* Predicting the semantic orientation of adjectives // Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics. — Association for Computational Linguistics. 1997. — Pp. 174–181.

91. *Kanayama H.*, *Nasukawa T.* Fully automatic lexicon expansion for domain-oriented sentiment analysis // Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics. 2006. — Pp. 355–363.

92. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. / D. Tang [и др.] // COLING. — 2014. — С. 172—182.

93. Coooolll: A deep learning system for Twitter sentiment classification / D. Tang [и др.] //. — 2014.

94. *Severyn A.*, *Moschitti A.* On the automatic learning of sentiment lexicons // Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2015). — 2015.

95. *Blinov P. D.*, *Kotelnikov E. V.* Semantic Similarity for Aspect-Based Sentiment Analysis // Proceedings of International Conference "Dialog". — 2015. — С. 12—22.

96. *Nokel M.*, *Loukachevitch N.* Application of Topic Models to the Task of Single-Word Term Extraction // RCDL. — 2013. — С. 52—60.

97. *Дейк Т.* ван, Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. — 1988. — № 23. — С. 153—211.

98. A rule-based approach to aspect extraction from product reviews / S. Poria [et al.] // Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP). — 2014. — Pp. 28–37.

99. SentiRuEval: testing object-oriented sentiment analysis systems in russian / N. Loukachevitch [и др.] // Proceedings of International Conference Dialog. Т. 2. — 2015. — С. 12—24.

100. Building a sentiment summarizer for local service reviews / S. Blair-Goldensohn [и др.] // WWW Workshop on NLP in the Information Explosion Era. Т. 14. — 2008. — С. 339—348.

101. *Moghaddam S.*, *Ester M.* Opinion digger: an unsupervised opinion miner from unstructured product reviews // Proceedings of the 19th ACM international conference on Information and knowledge management. — ACM. 2010. — С. 1825—1828.

102. Opinion word expansion and target extraction through double propagation / G. Qiu [и др.] // Computational linguistics. — 2011. — Т. 37, № 1. — С. 9—27.

103. *Jakob N.*, *Gurevych I.* Extracting opinion targets in a single-and cross-domain setting with conditional random fields // Proceedings of the 2010 conference on empirical methods in natural language processing. — Association for Computational Linguistics. 2010. — С. 1035—1045.

104. *Choi Y.*, *Cardie C.* Hierarchical sequential learning for extracting opinions and their attributes // Proceedings of the ACL 2010 conference short papers. — Association for Computational Linguistics. 2010. — С. 269—274.

105. *Chernyshevich M.* IHS R&D Belarus: Cross-domain extraction of product features using conditional random fields // Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). — 2014. — С. 309—313.

106. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid / W. X. Zhao [et al.] // Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. — Association for Computational Linguistics. 2010. — Pp. 56–65.

107. *Popescu A.-M.*, *Etzioni O.* Extracting product features and opinions from reviews // Natural language processing and text mining. — Springer, 2007. — Pp. 9–28.

108. *Jin W.*, *Ho H. H.*, *Srihari R. K.* A novel lexicalized HMM-based learning framework for web opinion mining // Proceedings of the 26th Annual International Conference on Machine Learning. — Citeseer. 2009. — С. 465—472.

109. Semeval-2015 task 12: Aspect based sentiment analysis / M. Pontiki [и др.] // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Association for Computational Linguistics, Denver, Colorado. — 2015. — С. 486—495.

110. Target-dependent twitter sentiment classification / L. Jiang [и др.] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. — Association for Computational Linguistics. 2011. — С. 151—160.

111. SZTE-NLP: Aspect Level Opinion Mining Exploiting Syntactic Cues / V. Hangya [и др.] // SemEval 2014. — 2014. — С. 610.

112. Multi-aspect sentiment analysis with topic models / B. Lu [et al.] // Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on. — IEEE. 2011. — Pp. 81–88.

113. Centroid-based summarization of multiple documents / D. R. Radev [и др.] // Information Processing & Management. — 2004. — Т. 40, № 6. — С. 919—938.

114. *Carenini G.*, *Cheung J. C. K.*, *Pauls A.* Multi-Document Summarization of Evaluative Text // Computational Intelligence. — 2013. — Vol. 29, no. 4. — Pp. 545–576.

115. Exploiting structured ontology to organize scattered online opinions / Y. Lu [et al.] // Proceedings of the 23rd International Conference on Computational Linguistics. — Association for Computational Linguistics. 2010. — Pp. 734–742.

116. *Blei D. M.*, *Ng A. Y.*, *Jordan M. I.* Latent dirichlet allocation // the Journal of machine Learning research. — 2003. — Vol. 3. — Pp. 993–1022.

117. *Mukherjee A.*, *Liu B.* Aspect extraction through semi-supervised modeling // Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. — Association for Computational Linguistics. 2012. — С. 339—348.

118. *Sauper C.*, *Haghighi A.*, *Barzilay R.* Content models with attitude // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. — Association for Computational Linguistics. 2011. — Pp. 350–358.

119. *Ramage D.*, *Manning C. D.*, *Dumais S.* Partially labeled topic models for interpretable text mining // Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. — ACM. 2011. — Pp. 457–465.

120. *Сабирова И.* Качество–ключевой фактор обеспечения конкурентности продуктов и услуг в условиях рыночной экономики // Автоматизация и управление в технических системах. — 2015. — № 1. — С. 181—190.

121. *Zhang W.*, *Xu H.*, *Wan W.* Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis // Expert Systems with Applications. — 2012. — Т. 39, № 11. — С. 10283—10291.

122. *Solovyev V.*, *Ivanov V.* Dictionary-based problem phrase extraction from user reviews // Text, Speech and Dialogue. — Springer. 2014. — С. 225—232.

123. AR-Miner: mining informative reviews for developers from mobile app marketplace / N. Chen [и др.] // Proceedings of the 36th International Conference on Software Engineering. — ACM. 2014. — С. 767—778.

124. *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.

125. Extraction from the web of articles describing problems, their solutions, and their causes / M. Murata [и др.] // IEICE transactions on information and systems. — 2011. — Т. 94, № 3. — С. 734—737.

126. *Wiebe J.* Learning subjective adjectives from corpora // AAAI/IAAI. — 2000. — С. 735—740.

127. Low-Quality Product Review Detection in Opinion Summarization. / J. Liu [и др.] // EMNLP-CoNLL. — 2007. — С. 334—342.

128. How opinions are received by online communities: a case study on amazon. com helpfulness votes / C. Danescu-Niculescu-Mizil [и др.] // Proceedings of the 18th international conference on World wide web. — ACM. 2009. — С. 141—150.

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

130. Semeval-2014 task 4: Aspect based sentiment analysis / M. Pontiki [и др.] // Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014). — 2014. — С. 27—35.

131. *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.

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

133. Twitter sentiment detection via ensemble classification using averaged confidence scores / M. Hagen [и др.] // Advances in Information Retrieval. — Springer, 2015. — С. 741—754.

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

135. Semeval-2015 task 10: Sentiment analysis in twitter / S. Rosenthal [и др.] // Proceedings of SemEval-2015. — 2015.

136. *Fahrni A.*, *Klenner M.* Old wine or warm beer: Target-specific sentiment analysis of adjectives // Proc. of the Symposium on Affective Language in Human and Machine, AISB. — 2008. — С. 60—63.

137. *Thet T. T.*, *Na J.-C.*, *Khoo C. S.* Aspect-based sentiment analysis of movie reviews on discussion boards // Journal of Information Science. — 2010. — С. 0165551510388123.

138. *Hays D. G.* Dependency theory: A formalism and some observations // Language. — 1964. — Т. 40, № 4. — С. 511—525.

139. *Patwardhan S.*, *Banerjee S.*, *Pedersen T.* Using measures of semantic relatedness for word sense disambiguation // Computational linguistics and intelligent text processing. — Springer, 2003. — С. 241—257.

140. Russe: The first workshop on russian semantic similarity / A. Panchenko [и др.] // Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue). — 2015. — С. 89—105.

141. Evaluating three corpus-based semantic similarity systems for russian / N. Arefyev [и др.] // Proceedings of the International Conference on Computational Linguistics and Intellectual Technologies (Dialogue). — 2015. — С. 106—118.

142. *Bär D.*, *Zesch T.*, *Gurevych I.* DKPro Similarity: An Open Source Framework for Text Similarity. // ACL (Conference System Demonstrations). — 2013. — С. 121—126.

143. *McDonald R.*, *Lerman K.*, *Pereira F.* Multilingual dependency analysis with a two-stage discriminative parser // Proceedings of the Tenth Conference on Computational Natural Language Learning. — Association for Computational Linguistics. 2006. — С. 216—220.

144. Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы / Ю. Д. Апресян [и др.] // Национальный корпус русского языка 2003–2005 г. — “Индрик”. 2005. — С. 193—214.

145. *Solovyev V.*, *Ivanov V.* Knowledge-Driven Event Extraction in Russian: Corpus-Based Linguistic Resources // Computational Intelligence and Neuroscience. — 2016. — Т. 501. — С. 4183760.

146. *Лукашевич Н.*, *Добров Б.* Исследование тематической структуры текста на основе большого лингвистического ресурса // Материалы ежегодной международной конференции “Диалог”. — 2000.

147. *Heinrich G.* Parameter estimation for text analysis // University of Leipzig, Tech. Rep. — 2008.

148. *Minka T.*, *Lafferty J.* Expectation-propagation for the generative aspect model // Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence. — Morgan Kaufmann Publishers Inc. 2002. — С. 352—359.

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

150. *Wang Y.* Distributed gibbs sampling of latent topic models: The gritty details: тех. отч. / Tech. Rep. — 2008.

# Приложение А

## Словари ProblemWord, NotProblemWord, Negation, AddWord, ImperativePhrases

Таблица 27 — Список лексических единиц словаря ProblemWord, явно указывающих на проблемную ситуацию (DirectPW)

| | | |
|---|---|---|
| аварийный | недостоверность | трудноватый |
| авария | недостоверный | трудновоспитуемый |
| архисложный | недоступно | трудновыполнимый |
| баг | недоступность | труднодоступный |
| банкрот | недоступный | трудноизлечимый |
| банкротство | недочет | трудноисполнимый |
| беда | неисправность | труднообрабатываемый |
| бедственный | ненадежно | труднообъяснимый |
| бедствие | ненадежный | трудноопределимый |
| безрезультативно | неправильно | труднопреодолимый |
| безрезультативный | неправильный | труднопроизносимый |
| безрезультатность | неприятно | труднопроходимый |
| безуспешно | неприятность | трудноразрешимый |
| безуспешность | неприятный | труднорастворимый |
| безуспешный | непроизносимо | труднореализуемый |
| бесплодность | непроизносимый | трудность |
| бесплотно | неровность | труднотекущий |
| бесплотный | несовершенный | трудноуловимый |
| бесцельно | несовершенство | трудноуправляемый |
| бесцельность | несоответствие | трудночитаемый |
| бесцельный | носоответствующий | трудный |
| бремя | обвинение | трудоёмкость |
| вмятина | обида | тухлятина |
| вред | оскорбление | тухлый |
| глюк | осложнение | тяготение |
| очень сложный | отказ | тяжесть |
| убыток | ошибка | укор |
| трёхсложный | ошибочно | укоризна |
| двенадцатисложный | ошибочность | унижение |
| двусложный | ошибочный | упрёк |

| | | |
|---|---|---|
| двухсложный | повреждение | урон |
| десятисложный | поврежденный | ухудшение |
| дефект | подделка | ущерб |
| дефектный | поддельный | уязвимое место |
| дырявый | поломка | фальшивка |
| забоина | помеха | фальшивый |
| заболевание | порча | фиаско |
| загвоздка | препятствие | царапина |
| запутывание | проблема | недостаток |
| затруднение | проблематичность | бесценок |
| затруднительно | проблематичный | минус |
| затруднительный | проблемно | претензия |
| злоключение | проблемный | претензионный |
| изъян | провал | жалоба |
| катаклизм | промах | заморочка |
| катастрофа | промашка | вылет |
| крах | просчёт | хелп |
| косяк | пятисложный | переустановка |
| крушение | пятно | неверный |
| лажа | разорение | неверно |
| ломка | разрушение | недопустимый |
| ляп | разрушенный | недопустимо |
| многократносложный | рваный | некорректный |
| многосложный | ремонт | некорректно |
| многотрудный | ржавчина | невозможно |
| мошенник | ржавый | не возможно |
| мошенничество | сбой | пропадать |
| надувательство | слабая сторона | перезагружать |
| нарекание | слабина | замечание |
| нарушение | слабое место | неактивный |
| невзгода | слабость | невозможный |
| невзгодье | сложнейший | неоплаченный |
| недалёкость | сложность | неоплаченый |
| недоработка | сомнительность | неполадка |
| недостаточно | ссадина | непонятно |
| недостаточность | трагедия | неработающий |
| недостаточный | трудно | неудобный |
| недостоверно | труднобольной | перезагрузка |
| неудобство | болячка | недочёт |
| неприемлемый | вводить в заблуждение | jumpy |

| | | |
|---|---|---|
| aimperfect | annoyance | lack |
| ineffective | bother | defect |
| unavailable | afraid | defects |
| warranty | support only | defecting |
| wrong | attempted | defective |
| difficult | issue | failing |
| noise | issues | failingly |
| crash | problem | difficulty |
| crashes | problems | difficulties |
| crack | problematic | bug |
| malfunction | error | bugs |
| trouble | errors | faults |
| necessary | delay | complain |
| barely enough | minor | complains |
| fail | for years | frustrations |
| useless | troubleshoot | not the same |
| flaw | step backward | limiations |
| unresponsive | backward | limiation |
| infamous | rebuild | broken |
| death | sharp | broke |
| break | stuck | unhappy |
| breaking | remove | garbage |
| distract | over heat | garbages |
| strange | popping up | distressed |
| suddenly | down forcefully | waste |
| needless | goes blank | mistakable |
| unwanted | wipe | mistaking |
| disappointed | faster | mistakes |
| garbage | smear | mistake |
| fault | blotchy | refuse |
| loss | scratch | unskillful |
| stop | bloat ware | trash |
| hardly use | tinny | cumbersome |
| leak | treble | cumbersomes |
| leaks | to come and go | failure |
| failed | twist | failures |
| trying | detract | technical support |
| tinkering | gone | scratch |
| headed back | reconfigure | nonresponsive |
| send back | requiring | impossible to |

| return | fragile | probably |
|---|---|---|
| gimmick | misaligned | complaints |
| request assistance | downgrading | complaint |
| inadvertently | removing | replacement |
| tech service | snap | unfortunately |
| tech support | tight | dissapointed |
| fear | staining | dissapointments |
| frustration | locking | dissapointment |
| too large | loose | too small |
| too big | too long | too old |
| too short | too tall | have to spend the money |
| too low | too long | have to buy |
| too high | | |

Таблица 28 — Список лексических единиц словаря ProblemWord с негативной тональностью (NegativePW), связанные с удобством использованием продукта

| абсурд | отрицательный | мерзость |
|---|---|---|
| абсурдизм | отрицательный момент | мусор |
| абсурдность | паршивый | мусорный |
| ахинея | плохенький | мутный |
| бардак | плохо | наглупить |
| безголовый | плоховатый | надоедливый |
| безмозглый | плохое самочувствие | неблагозвучие |
| безобразный | плохой | невосполнимый |
| безрассудица | плохонький | негативный |
| безрезультатный | поганый | недалёкий |
| безумец | позорный | недомогание |
| безуспешный | попасть в просак | недостойный |
| белиберда | придурковатость | нежелательный |
| белибердень | придурковатый | нездоровье |
| бесплодный | примитивный | неимоверно |
| бессвязица | проигрыш | неимоверный |
| бессильный | пролёт | некачественный |
| бессмысленность | просроченный | некомфортно |
| бессмысленный | профукать | некомфортный |
| бессмыслица | пугающий | нелепица |
| бессодержательный | расхуячить | нелепость |
| бестолковщина | свалка | нелепый |
| бесчестный | сглаз | ненадёжный |

| | | |
|---|---|---|
| болезненный | сглупить | неоправданный |
| больной | скудоумие | непонятливый |
| бредовость | слабо | непонятно |
| бредовый | слабоватый | непонятный |
| вздор | слабоумие | непристойный |
| волноваться | слабый | несдержанный |
| вонючий | сложно | несмышлёный |
| враньё | сложный | несообразительный |
| втридорога | смерть | несообразность |
| гадкий | страшный | несуразица |
| гадость | стучать | несуразность |
| галиматья | сумасбродство | несусветица |
| галимый | трусливый | несусветность |
| глупить | трындец | несчастный |
| глупо | туповатый | несчастье |
| глуповатость | тупоголовый | неудобно |
| глуповатый | туполобый | неудобный |
| глупость | тупость | неумный |
| глупый | тупоумие | нечестивый |
| голимый | тупоумный | нечестный |
| грязный | тяжелобольной | низковато |
| дебилизм | тяжеловатый | низковатый |
| дисгармония | убогий | никчёмный |
| дискомфорт | убого | обидно |
| диссонанс | ужасающий | обидный |
| дохлый | ужасно | облом |
| жалеть | ужасный | обременительный |
| жалкий | уродливый | огорчение |
| жалко | утомительный | огорченный |
| жуткий | ущербный | огорчить |
| жутко | ущербно | однобокость |
| запачканный | хана | околесица |
| запредельный | хаос | омерзительный |
| засада | херовый | оправданно |
| затруднительное положение | хилый | оставлять желать лучший |
| зловещий | хитровыебанный | оставлять желать хороший |
| идиотический | хитрожопистый | отбросы |
| имбецильность | хитрожопый | отваливаться |

| какофония | хитрозадый | отвратительно |
|---|---|---|
| капут | хитромудрый | отвратительность |
| каюк | хитромудый | отвратительный |
| кончина | хитропопый | отвратный |
| кошмар | хитросделанный | destroyed |
| кошмарно | хлам | overused |
| кошмарный | хреновый | poorly |
| лажовый | чепуха | overworked |
| медленно | швах | stupid |
| медленный | юрод | negative |
| мерзкий | юродивый | terribly |
| bad | stupidly | negatively |
| badly | stupidness | negativeness |
| horrible | stupidest | shame |
| horribly | dead | upset |
| mess | ugly | upsetting |
| upset | ugliest | unhappy |
| slow | hard to | worry |
| slower | impossible to | worn |
| slowly | never able to | poorly |
| slowness | poor | |

Таблица 29 — Список лексических единиц словаря ProblemWord, указывающие на проблемные ситуации в ходе эксплуатации продукта (VerbPW)

| батрачить | ухудшаться | усложнять |
|---|---|---|
| бесславить | ухудшиться | затруднять |
| бесчестить | чернить | обострять |
| бычиться | шуметь | чихать |
| вздорить | щетиниться | дохать |
| вредить | трещать | кашлять |
| выгнать | тошнить | карать |
| выгнить | плющить | спорить |
| выдворить | подташнивать | пугаться |
| выломать | залихорадить | наводить страх |
| вымачиваться | разочаровываться | пугать |
| выпереть | отваливаться | терроризировать |
| вытеснить | треснуть | ужаснуть |
| вышвырнуть | отламываться | запугивать |
| гадить | ошибиться | устрашать |

| | | |
|---|---|---|
| громить | ошибаться | страшить |
| грязнить | подводить | запужать |
| губить | кусаться | застращивать |
| догнить | пожалеть | перепугаться |
| доломать | напрягать | напугаться |
| дребезжать | мерзнуть | опозориться |
| дрогнуть | замерзать | сдрейфить |
| ерихониться | огорчать | устрашиться |
| ерошиться | лопнуть | прибздеть |
| ершиться | жаловаться | облить |
| загнить | разваливаться | испугаться |
| загрязнить | деформировать | струсить |
| замусориваться | деформироваться | заочковать |
| замусорить | изгибать | убояться |
| заржаветь | изгибаться | обосраться |
| засорить | искривляться | усраться |
| засоряться | домять | осрамиться |
| застукать | гнуться | недосмотреть |
| злословить | перегружать | недоглядеть |
| зябнуть | висеть | набздеть |
| изводить | тупить | напердеть |
| извращать | выкидывать | напукать |
| изгнить | вылетать | шалить |
| изломать | зависать | кудесничать |
| изнашивать | врать | шкодить |
| изнашиваться | тормозить | бедокурить |
| изувечить | виснуть | проказничать |
| изуродовать | перезагружать | злобить |
| искажать | повреждаться | сердить |
| искажаться | медлить | возмущать |
| исказить | мешкать | озлоблять |
| искалечить | бедовать | дразнить |
| исковеркать | бедствовать | подбешивать |
| искривить | разграбить | злить |
| искривиться | расхитить | разъярять |
| искривлять | раздосадовать | гневать |
| иссушить | унизить | раздражать |
| истлеть | обидеть | гневить |
| калечить | оскорбить | выводить из себя |
| клеветать | хулить | требовать |

| | | |
|---|---|---|
| коверкаться | осрамить | преклонять |
| кокнуть | хаять | замерзнуть |
| кривляться | опозорить | вымерзать |
| крушить | пакостить | браконьерить |
| леденеть | гваздать | браконьерствовать |
| ликвидировать | пакостничать | поспорить |
| ломать | выругать | жалковать |
| ломаться | присочинять | горевать |
| марать | осквернять | соболезновать |
| мокнуть | охаять | унывать |
| мусорить | запачкать | обезобразить |
| мучить | запятнать | выпроводить |
| надгнить | испачкать | отбрасывать |
| надломать | замарать | откидывать |
| надломить | ехидничать | взгомониться |
| надрываться | лашать | переполошиться |
| накосячить | лгать | всполошиться |
| наломать | застудить | избавиться |
| намокать | простудить | отвлекать |
| намусорить | пасовать | растлевать |
| напортачить | грохотать | развращать |
| напортить | грохать | хряпнуть |
| нарушаться | греметь | изранить |
| нездоровиться | громыхать | ранить |
| обветшать | грюкать | ужалить |
| обвинять | хуякнуться | назюзюкаться |
| обгнить | уродовать | мусолить |
| оклеветать | коверкать | взбунтовать |
| околеть | грубиянить | остоёбнуть |
| окромсать | нагличать | взбудоражить |
| опрокинуть | грубить | всполошить |
| отвалиться | дерзить | набухаться |
| отказать | задираться | приесться |
| отломать | хамить | фордыбачить |
| отломить | промёрзнуть | фордыбачиться |
| отмокать | намёрзнуться | суетиться |
| отсыревать | назябнуться | хакнуть |
| паскудить | иззябнуться | взломать |
| паскудничать | прохлаждаться | изнемочь |
| паскудствовать | бездельничать | устать |

| | | |
|---|---|---|
| пачкать | лодырничать | утомиться |
| паясничать | нахлебничать | притомиться |
| перегнить | паразитничать | задолбаться |
| перегреть | паразитствовать | лишиться сил |
| переломать | тунеядствовать | обессилеть |
| перепортить | лентяйствовать | скрючиться |
| перержаветь | паразитировать | ожиревать |
| пнуть | лентяйничать | жиреть |
| повалить | лодарничать | напроказничать |
| повредить | лениться | нахулиганить |
| повреждать | изгрязнить | наделать дел |
| погнить | измарать | крысятничать |
| погубить | выпачкать | хомячить |
| подгнить | сбагрить | сжирать |
| подломать | шелестеть | скупердяйничать |
| подорвать | зудеть | скаредничать |
| подпортить | беднеть | скопидомничать |
| позорить | изощрять | скопидомствовать |
| покалечить | подтрунивать | крохоборствовать |
| пококать | мямлить | двоедушничать |
| покоцать | лепить | лицемерствовать |
| поломать | брюзжать | лицемерить |
| поломаться | зашхериться | притворствовать |
| полусгнить | наебнуться | лицемерничать |
| пообломать | навернуться | загрязняться |
| попортить | покрыть себя | гваздаться |
| поржаветь | пиздануться | грязниться |
| порочить | гуфнуться | пачкаться |
| портить | наводниться | мараться |
| портиться | ёбнуться | наводить |
| пробить | охренеть | отвратить |
| прогнать | вздуриться | отнять |
| прогнать со двора | ополоуметь | придуриваться |
| прогнить | обнаглеть | филонить |
| продолбать | обезуметь | отрицать |
| продолбить | сдуриться | отказывать |
| проломать | охереть | завинчиваться |
| промокать | сдуреть | вилять |
| пропахнуть | охуеть | отлынивать |
| противиться | обеспамятеть | гавкать |

| | | |
|---|---|---|
| протухнуть | опсихеть | казнить |
| разбить | офонареть | алкоголизировать |
| развалить | пыжиться | глумиться |
| разворотить | чваниться | ёрничать |
| разгромить | жилиться | насмехаться |
| раздраконить | кичиться | упрекать |
| разлагаться | потеть | злорадствовать |
| разламывать | ишачить | хохмить |
| разломать | охуждать | дурачиться |
| размозжить | осуждать | волноваться |
| разнести | порицать | хлопотать |
| разодрать | сердиться | обмарывать |
| разрушать | гневаться | спамить |
| разрушаться | серчать | мешать |
| разрушить | злиться | надругаться |
| расколоть | обматюгать | выругаться |
| рваться | обматерить | ругнуться |
| ржаветь | замарывать | поссориться |
| рушить | оговаривать | поругаться |
| сгнивать | зачернять | каверкнуть |
| сгнить | наговаривать | рехнуться |
| сгореть | обезображивать | охаметь |
| сдохнуть | убирать | одуреть |
| скоморошничать | перезагружаться | сойти с ума |
| скрипеть | перезапускать | хандрить |
| сломать | перезапускаться | хмуриться |
| сломаться | перемудрить | помешать |
| сокрушать | подтормаживать | просрочивать |
| сорить | переустанавливать | вымогать |
| сыреть | переустановить | обкрадывать |
| толкнуть | просрочить | впаривать |
| турнуть | сбоить | навязывать |
| убивать | дурить | высмеивать |
| убить | напрягать | неметь |
| увечить | мучаться | онемевать |
| угробить | мучиться | паниковать |
| ударить | испоганить | разорять |
| уничтожать | изничтожить | грабить |
| уничтожить | похерить | выблевать |
| уродоваться | сгубить | продырявиться |

| ухудшать | осложнять | прохудиться |
|---|---|---|
| намучаться | leaks | lost |
| rebuild | misaligned | fail |
| sharp | downgrading | failed |
| stuck | removing | failing |
| remove | snap | fails |
| over heat | tight | complain |
| popping up | staining | complained |
| down forcefully | locking | frustrating |
| goes blank | loose | frustrated |
| wipe | freeze | refuse |
| smear | freezed | gave up |
| blotchy | freezes | give up |
| scratch | freezing | disappointing |
| bloat ware | removed | disappoint |
| upset | removes | disappointed |
| tinny | crashed | risking |
| treble | crashes | bother |
| to come and go | crashing | bothered |
| slow | fragile | bothering |
| twist | force | screw up |
| detract | forcing | screws up |
| reconfigure | forced | fuck up |
| requiring | replace | send back |
| leak | replaced | leaks |
| leaked | | |

Таблица 30 — Предметно-ориентированный словарь ProblemWord для машин

| слишком маленький | большой расход | перегорать |
|---|---|---|
| слишком жесткий | гнить | дорогой в обслуживание |
| слишком низкий | гниение | дешевый качество |
| немного завышать | гнилой | косо |
| слабоватый | гниль | сырой |
| жестковатый | скрип | заносить |
| бренчать | шум | выбрация |
| сыпаться | шумный | рыжик |
| дерганье | отклеиваться | дыра |
| дёрганье | заглохнуть | приходиться менять |

| | | |
|---|---|---|
| расход большой | глохнуть | приходиться ремонтировать |

Таблица 31 — Предметно-ориентированный словарь ProblemWord для мобильных приложений

| | | |
|---|---|---|
| вернуть старый | батарея жрать | зависать |
| вернуть прежний | аккум жрать | зависание |
| вернуть отображение | аккумулятор жрать | не грузить |
| вернуть деньги | трафик жрать | не мочь |
| вернуть функционал | только возможность | не совпадать |
| вернуть функция | садить батарейка | не скачивать |
| вернуть обратно к | садить батарея | нельзя редактировать |
| вернуть все обратно | садить аккум | каверкаться |
| вернуть возможность | садить аккумулятор | каверкать |
| давать возможность | временно недоступна | пытаться |
| хотеться возможность | немного мелковатый | сбрасываться |
| обновление переставать | немного запутывать | сложно сделать |
| надеяться исправлять | немного не четкий | трудно сделать |
| приходиться вводить | немного сложный | неправильно отображаться |
| просить исправлять | немного сложноватый | вернуть обратно |
| белый экран | немного кривой | требовать четкий |
| пустой экран | немного надоедать | перезапускать |
| постоянно выдавать | почему только | перегружать |
| постоянно расходовать | почему нельзя | перегружено |
| постоянно искать | ограничивать | глюк |
| исправлять поставлять | вылетать | глякавый |
| перезагрузиться | удалять | глюкать |
| писать неверный | исправлять | глючный |
| постоянно выскакивать | выбрасывать | глючно |
| служба поддержка | выкидывать | глючать |
| устанавливать заново | тормозить | глючить |
| жрать батарея | подтормаживать | не работоспособный |
| жрать аккум | раздражать | неработоспособный |
| жрать аккумулятор | падать | при попытка |
| жрать трафик | пропадать | лимит |
| исчезать | | |

Таблица 32 — Список лексических единиц словаря NotProblemWord

| | | |
|---|---|---|
| работать | несущественно | выгодно |
| выходить | несущественный | верный |
| запускаться | малозначимый | верно |
| обновляться | пустячный | активный |
| показывать | нетрудный | корректно |
| загружаться | несложный | корректный |
| переключаться | пустяковый | безопасный |
| удаваться | незначительно | безопасно |
| устраивать | незначительный | понятно |
| помогать | неважный | понятный |
| починка | неважно | полезный |
| наладка | неважнецки | полезно |
| чинка | хорошо | доступный |
| исправление | хороший | доступно |
| наладить | удобно | радовать |
| чинить | удобный | устраивать |
| дочинить | комфортный | мелочь |
| чиниться | комфортно | неплохо |
| безошибочный | комфортабельный | неплохой |
| легковыполнимый | комфортабельно | актуальный |
| легкоисполнимый | устойчиво | выручать |
| советовать | устойчивый | выгодный |
| излечить | легкий | починить |
| вылечить | легко | runs |
| adjust | get | safety |
| approve | gets | send |
| approved | getting | sending |
| comfortable | got | sends |
| comfortably | perfect | sent |
| connect | properly | sufficiently |
| connected | resolve | suitable |
| connecting | resolved | support |
| connects | resolves | trust |
| easy | respond | work |
| enough | responded | workable |
| fast | responding | worked |
| faster | responds | working |
| fine | run | works |

| | | |
|---|---|---|
| function | running | worth |

Таблица 33 — Список лексических единиц словаря Negation для определения отрицаний действий

| | | |
|---|---|---|
| ни за что | редко | невозможно |
| ни хера | нечего | не мочь |
| нисколечко | ничто | нельзя |
| не | ни когда | не возможно |
| ни один | никогда | не хотеть |
| no | nobody | in no way |
| not | nowhere | stop |
| n't | cannot | stopped |
| neither | no one | not able to |
| nor | no way | solved |
| non | will never | refused to |
| none | not able to | |

Таблица 34 — Список лексических единиц словаря Negation для определения отрицаний индикаторов

| | | |
|---|---|---|
| ни за что | ни один | не мочь |
| ни хера | нету | нельзя |
| нисколечко | нет | не возможно |
| ничто | редко | не хотеть |
| в отсутствие | нечего | устранять |
| при отсутствии | ничего | исправить |
| за вычетом | ни один | исправляться |
| без | нету | исправлять |
| не | никогда | перестать |
| никакой | невозможно | переставать |
| не возникать | не видеть | не бывать |
| не наблюдать | не найти | не замечать |
| no | cannot | refuse |
| not | with out | cease |
| n't | without | solved |
| never | no one | stopped |
| neither | no way | breaking |
| nor | in no way | not able to |

| | | |
|---|---|---|
| none | stop | nowhere |
| nobody | nothing | |

Таблица 35 — Список лексических единиц словаря ImperativePhrases

| | | |
|---|---|---|
| сделайте | попробуйте | делайте |
| сделаете | реализуйте | наймите |
| обеспечьте | разработайте | верните |
| доделайте | откорректируйте | почините |
| введите | проверьте | допишите |
| исправьте | внесите | доделайте |
| поправьте | берите | исправляйте |
| добавьте | восстановите | дайте |
| would be very useful | wish | should |
| would be helpful | would be very helpful | could have |
| would be useful | | |

Таблица 36 — Список лексических единиц словаря AddWord

| | | |
|---|---|---|
| пришлось | больно | следует |
| приходиться | избыточно | надо |
| прийтись | чрезвычайно | требуется |
| слишком | крайне | понадобиться |
| немного | изрядно | правда |
| достаточно | необычайно | добавлять |
| чрезмерно | надлежит | сделать |
| чересчур | необходимо | вернуть |
| излишне | переставать | перестать |
| шибко | перестать | нельзя |
| весьма | anymore | only |
| no | something | too much |
| none | nothing to | too many |
| at least | still | after |
| sometimes | too | rather |
| not last | to much | |