Совместная вероятностная тематическая модель для идентификации проблемных высказываний, связанных нарушением функциональности продуктов

Е.В. Тутубалина <elvtutubalina@kpfu.ru>
Казанский (Приволжский) федеральный университет,
420008, Россия, г. Казань, ул. Кремлевская, дом 18.

Аннотация. В статье исследуется задача автоматического извлечения информации о существовании различных проблем с продуктами из отзывов пользователей. В последние десятилетия на рынке потребительских товаров появилась резкая динамика увеличения количества технически сложных товаров. У покупателей возникают претензии по поводу удобства использования продукта наряду с ненадлежащим техническим качеством. Пользователи публикуют свои мнения о сложностях в использовании продуктов, что может оказывать влияние на процесс принятия решения о покупке продуктов потенциальными потребителями. Для достижения целей исследования предложены две тематические модели на основе латентного размещения Дирихле, позволяющие совместно учитывать несколько типов информации для идентификации проблемных высказываний. Предложенные алгоритмы моделируют распределение слов в документе, учитывая взаимосвязь между скрытыми тематической, тональной и проблемной переменными. Результаты экспериментального исследования анализируются в статье в сравнении с результатами популярных вероятностных моделей для задач анализа мнений, в качестве критериев оценки используются стандартные метрики качества систем анализа текстов и перплексия контрольных данных (perplexity). Для качественной оценки тематических распределений моделей был проведен анализ тем, подтверждающий целесообразность определения тональности для критических высказываний пользователей. Эксперименты показали, что наилучшие результаты классификации фраз о проблемах в использовании продуктов показывают предложенные модели, использующие совместную информацию из отзывов пользователей на русском и английском языках.

Ключевые слова: отзывы пользователей, латентное размещение Дирихле, Latent Dirichlet Allocation, совместная вероятностная модель, извлечение проблемных высказываний

DOI: 10.15514/ISPRAS-2015-27(4)-6

Для цитирования: Тутубалина Е.В. Совместная вероятностная тематическая модель для идентификации проблемных высказываний, связанных нарушением функциональности продуктов. Труды ИСП РАН, том 27, вып. 4, 2015 г., стр. 87-96. DOI: 10.15514/ISPRAS-2015-27(4)-6.

1. Введение

В последние десятилетия на рынке потребительских товаров появляется все больше технически сложных товаров. Это связано, прежде всего, с развитием технологических инноваций, что приводит к постоянному увеличению конкретных видов компьютерных продуктов, и с концепцией соединения разной функциональности в едином устройстве. В связи с этим у покупателей возникают претензии по поводу удобства использования продукта наряду с ненадлежащим техническим качеством. Многие покупатели осуществляют возврат товаров компаниям даже, если товар работает исправно согласно государственным стандартам и техническим отчетам компаний. Описанные примеры иллюстрируют необходимость извлечения информации из отзывов пользователей о существовании проблем с теми или иными продуктами для обеспечения высокого качества продукции и устранения затруднений, влияющих на работу продуктов. Высокое качество продукции и услуг является свойством, которое определяет их конкурентоспособность в условиях рыночной экономики [1]. В данной работе исследуется корреляция между проблемными индикаторами, эмоционально-окрашенными (тональными) словами пользователя и темами внутри отзывов о товарах.

Описание проблемной ситуации с продуктами может сопровождаться негативными, позитивными или нейтральными высказываниями относительно аспектных терминов (целевых объектов), о которых высказывание было сделано. Описания технических проблем, связанных с нарушением функциональности продуктов, не содержат в себе явных тональных слов, сообщая нейтральную информацию о действительном событии. В отличие от этого, описание проблем с ухудшением эффективности, продуктивности и удобства использования продуктов характерно тем, что содержит некую тональную оценку относительно разных категорий целевых объектов. Например, пользователи электронных устройств могут быть недовольны медленной ответной реакцией дисплея, низкой эффективностью батареи, слишком громким или тихим звуком динамика, избыточным количеством функций, короткими проводами, недостаточно ярким цветом корпуса. Это объясняет необходимость агрегировать совместную информацию, влияющую на распределение слов в отзыве пользователя, о тематической категории целевого объекта, тональном контексте (позитивном, нейтральном, негативном) и знаниях о проблемных высказываниях.

В мировой науке существует небольшое количество работ, посвященных задаче анализа высказываний пользователей, на предмет обнаруженных ими проблем в использовании тех или иных устройств [2-5]. В основном эти

исследования больше сосредоточены над созданием методов, основанных на словарях индикативных конструкций [2, 3], частичном обучении с применением правил или синтаксических конструкциях [4, 5], не анализируя возможность внедрения дополнительных знаний о тональности слов, кроме определения негативных высказываний. В данной работе представлены тематические модели на основе модели латентного размещения Дирихле (Latent Dirichlet Allocation, LDA) [6], направленные на объединение задач идентификации категорий аспектных терминов, определения тональности и идентификации проблем с продуктами в отзывах.

В работе предлагаются две вероятностные модели: (i) совместная модель *тема-тональность-проблема* (topic-sentiment-problem model, TSPM), моделирующая слова в документе в зависимости от нескольких скрытых переменных; (ii) совместная модель *оценка пользователя-тема-тональность-проблема* (rating-aware topic-sentiment-problem model, RTSPM), являющейся модификацией модели TSPM с добавлением известной оценки продукта пользователем. В качестве критериев оценки качества моделей используется перплексия (perplexity). В качестве критериев оценки качества моделей в задаче классификации использованы популярные метрики задач автоматической обработки естественного языка такие, как аккуратность (accuracy), точность (precision), полнота (recall) и F-мера (F-measure). Результаты работы предложенных методов оценены на корпусах текстов с отзывами пользователей о компьютерах, машинах, инструментам для дома и детских товарах.

Статья состоит из следующих разделов: в разделе 2 обсуждается современное состояние исследований по задаче анализа мнений, в разделе 3 приводятся описания вероятностных моделей. Раздел 4 посвящен статистическому оцениванию предложенных моделей с помощью сэмплирования Гиббса. В разделе 5 анализируются результаты экспериментов в сравнении с популярными вероятностными моделями. В разделе 6 обсуждаются выводы, сделанные на основе экспериментов, и направления дальнейшей работы.

2. Современное состояние исследований

В мировой науке подавляющее большинство работ по анализу мнений пользователей посвящено английскому языку и методам определения эмоциональной окраски текста относительно аспектных терминов (aspect-based sentiment analysis), подробный обзор которых описан в работах [7, 8]. Популярными подходами к задаче анализа тональности являются подходы, основанные на знаниях в виде словарей и использующие статистические метрики или машинное обучение для задачи классификации мнений.

В настоящий момент доминирующими методами являются алгоритмы на основе модели латентного размещения Дирихле [6] для задачи определения аспектных терминов, предметно-ориентированных тональных индикаторов и тематической категоризации аспектов продукта. Это связано с тем, что

создание предметно-ориентированной обучающей выборки для классификатора требует больших затрат времени в то время, как вероятностные модели позволяют использовать коллекции неразмеченных документов для нахождения скрытых переменных. В работах [9-12] представлены тематические модели, направленные на объединение задач идентификации аспектных терминов в отзыве и определения тональности для этих объектов. Эти модели используют словарь позитивных и негативных слов для задания гиперпараметра β (априорное распределение Дирихле на мультиномиальном распределении Ф в пространстве слов для темы).

В работе [10] авторы описывают модель *тональность-тема* (joint sentiment-topic model, JST) и модель *тема-тональность* (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 [11] описана объединенная модель *аспект-тональность* (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. В работе (Yang et al., 2015) представлена модификация LDA, названная User-aware Sentiment Topic Models (USTM), которая включает в распределения метаданные профайлов пользователя (геолокацию, возраст, пол) и словари эмоционально-окрашенной лексики для определения связи между тематическими наборами аспектов и категориями пользователей. Модель показывает наилучшие результаты классификации отзывов о машинах и ресторанах по сравнению с популярными вероятностными моделями JST и ASUM.

Задача анализа высказываний пользователей, на предмет обнаруженных ими проблем в использовании тех или иных устройств, является менее изученной. Существует несколько работ по классификации фраз отзывов об электронных устройствах [2-5]. В авторе [2] использует метод машинного обучения (метод максимальной энтропии) для извлечения аспектов из коротких сообщений о компании AT&T. 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. В работе [3] используются методы, основанные на правилах, для идентификации проблем в предложениях отзывов пользователей. Авторы используют словарь негативных слов для идентификации проблемных высказываний, не анализируя влияние нейтральной или позитивной тональности. В работе [5] авторы описывают модификацию LDA для идентификации целевых объектов и проблемных высказываний: модификация моделирует два распределения в пространстве аспектных терминов и в пространстве проблемных индикаторов. В качестве целевых объектов рассматриваются существительные отзывов, проблемными индикаторами считаются прочие слова (глаголы, наречия, прилагательные и пр.). Каждому документу соответствует два распределение в пространстве тем, где проблемные индикаторы зависят от темы и проблемной метки. В качестве критериев оценки метода использовалась перплексия, эффективность модели в задаче классификации не анализировалась. В работе [4] результаты экспериментов показали, что метод, основанный на частичном обучении с применением правил, может показать сопоставимые результаты относительно методов машинного обучения, для которых 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. Авторам статьи неизвестны работы, посвященные задаче автоматического извлечения информации о существовании различных проблем с товарами на русском языке, использующие вероятностные тематические модели.

3. Совместные вероятностные тематические модели для задачи идентификации проблем

Определим точную формулировку описанной задачи. Пусть $P = \{P_1, P_2, \dots, P_m\}$ – множество продуктов (сервисов, товаров), выпускаемое компаниями на потребительском рынке. Для каждого продукта $P_i \in P$ задано множество отзывов пользователей $D = \{d_1, d_2, \dots, d_n\}$. Каждый продукт $P_i \in P$ состоит из множества целевых объектов (компонентов, составных частей) $T = \{t_1, t_2, \dots, t_k\}$. В некоторых отзывах пользователи сообщают о дефектах или нарушениях функционирования продуктов. Проблемным высказыванием называются фраза пользователя в отзыве, содержащая явное указание на трудность в использовании тех или иных продуктов, невозможность использования продуктов вследствие ошибки (бага, дефекта) или сложности в использования продукта. Например,

– «Когда машину только купили, *дверь багажника закрывалась плохо*, приходилось очень сильно хлопать».
– «Один минус в салоне, что *нет бортового компьютера*».

– «Телефоном доволен, даже не смотря на то, что *пластиковый корпус потускнел*».
– «В целом все было отлично, единственное, что не понравилось так это то, что в зимние дни *батарея разряжается быстрее* из-за металлической крышки».
– «Когда едешь по не асфальтированной дороге, возникает *шум плюс скрип в салоне*».

В целом, задача автоматического извлечения информации о существовании различных проблем с товарами состоит из трех основных подзадач:

1. идентификация проблемных высказываний из текстов пользователей;
2. извлечение проблемных фраз по отношению к целевым объектам, зависящих от конкретной предметной области;
3. резюмирование целевых объектов продуктов и проблемных индикаторов по тематическим категориям.

В данной работе рассматривается задача резюмирования целевых объектов продуктов по тематическим категориям с учетом информации о тональных и проблемных индикаторах.

После анализа высказываний из отзывов пользователей мы выделили несколько типов фраз с различной тональной окраской. 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 тональностью, 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 с различной тональной степенью, влияющие на общее представление о проблемной ситуации.

Для достижения целей исследования в статье предложены два совместные тематические модели:

1. модель *тема-тональность-проблема* (topic-sentiment-problem model, TSPM)
2. модель *оценка пользователя-тема-тональность-проблема* (rating-aware topic-sentiment-problem Model, RTSPM).

Графические представления TSPM и RTSPM моделей приведены на рис. 1.

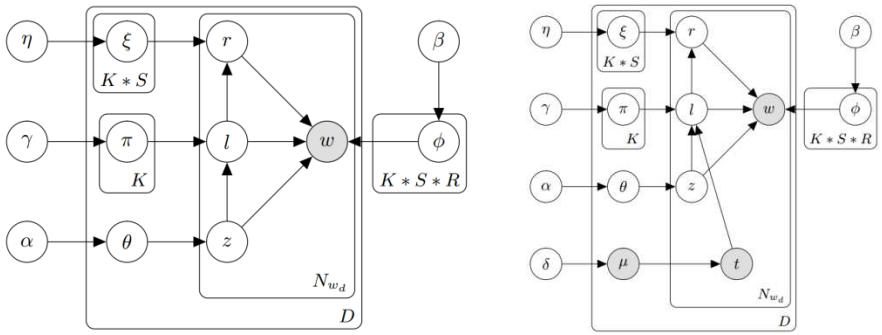

Рис. 1. Совместные вероятностные модели: (1) TSPM и (2) RTSPM

3.1 Вероятностная модель тема-тональность-проблема

В рамках TSPM модели каждой теме соответствует мультиномиальное распределение в пространстве слов. Для каждого слова в документе тема *z* выбирается из мультиномиального распределения θ, затем выбирается тональная метка *l* из мультиномиального распределения π, соответствующего теме *z*, затем выбирается проблемная метка *r* из мультиномиального распределения ε, соответствующего паре (тема *z*, тональная метка *l).* Наконец, слово *w* выбирается из распределения Φ 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. Графическое представление RTSPM приведено на рис.1 (1). В таблице 1 приводится список основных обозначений, используемых в моделях.

Табл. 1. Основные обозначения в TSPM и RTSPM моделях.

Символ	Описание
D	число документов
V	размер словаря слов в коллекции
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*
θ_d	мультиномиальное распределение в пространстве тем с параметром α
π_z	мультиномиальное распределение в пространстве тональных меток для темы *z* с параметром β

$\varepsilon_{z,l}$	мультиномиальное распределение в пространстве проблемных меток для пары (тема *z,* тональная метка l) с параметром β
$\Phi_{z,l,r}$	мультиномиальное распределение в пространстве слов для троек (тема *z*, тональная метка *l*, проблемная метка *r*) с параметром β
t_{di}	множество оценочных тэгов, присвоенных *i*-му слову в документе *d*
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*
α	априорное распределение Дирихле на параметры θ
β	априорное распределение Дирихле на параметры Φ
η	априорное распределение Дирихле на параметры ξ
γ	априорное распределение Дирихле на параметры π
δ	априорное распределение Дирихле на параметры ψ

В данной работе число тональных классов равно 3 (позитивный, нейтральный, негативный), число проблемных классов равно 2 (проблемный класс и класс с отсутствием указаний на проблемы). Совместная вероятность слов, тем, тональных и проблемных меток для TSPM могут быть посчитана следующим образом:

$$p(w,z,r,l) = p(w|z,l,r)\, p(r|z,l)\, p(l|z)\, p(z)$$

TSPM характерен следующий порождающий процесс:

- для каждой тройки (тема *z*, тональная метка *l,* проблемная метка *r*) выбирается распределение слов в каждой теме $\Phi_{z,l,r} \sim Dir\,(\beta_{zlk})$ $(l \in \{neu, pos, neg\}, r \in \{pr, no-pr\})$
- 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 β, зависящее от комбинации $(z_{d,w_i}, l_{d,w_i}, r_{d,w_i})$

3.2 Вероятностная модель оценка пользователя-тема-тональность-проблема

Модель *оценка пользователя-тема-тональность-проблема* (rating-aware topic-sentiment-problem Model, RTSPM) является модификацией предложенной модели TSPM. Графическое представление RTSPM приведено на рис. 1 (2). В рамках RTSPM добавляется переменная, связанная с оценкой пользователя по отношению к продукту, который описывается в отзыве. На многих онлайн-ресурсах пользователи оценивают продукт с точки зрения того, как хорошо он работает по пятизвездочному рейтингу качества. Затем пользователи выражают свои мнения, используя эмоционально-окрашенные слова, которые коррелируют с выставленным значением.

RTSPM модель рассматривает документ с известной оценкой рейтинга: каждое слово документа взаимосвязано с оценочным тэгом *t*, эквивалентного пятизвездочному рейтингу. Для каждого слова в документе тема *z* выбирается из мультиномиального распределения θ, затем для пары (тэг *t*, тема *z*) выбирается тональная метка *l* из мультиномиального распределения π, затем выбирается проблемная метка *r*, соответствующая паре (тэг *t*, тема *z*, тональная метка *l*). Слово *w* выбирается из распределения Φ соответствующего тэгу *t*, теме *z*, тональной метке *l,* проблемной метке *r*.

RTSPM характерен следующий порождающий процесс:

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

4. Статистическое оценивание предложенных моделей

Данный раздел описывает алгоритм статистического оценивания предложенных TSPM и RTSPM моделей. Для решения задачи статистического оценивания применительно к тематической модели LDA существует несколько алгоритмов: сэмплирование Гиббса (Gibbs sampling), вариационный вывод (variational inference), распространение математического ожидания (expectation propagation) [6, 13, 14]. В данной работе применяется сэмплирование Гиббса для оценивания параметров модели, поскольку такой подход позволяет эффективно находить скрытые темы в корпусах текстов [15].

Используя сэмплирование Гиббса, присвоенные скрытые параметры в модели TSPM могут быть выбраны по следующей формуле:

$$P(\mathbf{z}_{d,i} = k, \mathbf{l}_{d,i} = l, \mathbf{r}_{d,i} = r | \mathbf{w}_{d,i} = w, \mathbf{z}_{-(d,i)}, \mathbf{l}_{-(d,i)}, \mathbf{r}_{-(d,i)}, \alpha, \beta, \gamma, \eta) \propto \frac{n_{k,l,r,w}^{-(d,i)} + \beta_{l,r}^{w}}{n_{k,l,r}^{-(d,i)} + \sum_{j=1}^{V} * \beta_{l,r}^{j}} \frac{n_{d,k,l,r}^{-(d,i)} + \eta}{n_{d,k,l}^{-(d,i)} + R * \eta} \frac{n_{d,k,l}^{-(d,i)} + \gamma}{n_{d,k}^{-(d,i)} + L * \gamma} \frac{n_{d,k}^{-(d,i)} + \alpha}{n_{d}^{-(d,i)} + K * \alpha} \quad (1)$$

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 *–(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 RTSPM по следующей формуле:

$$P(\mathbf{z}_{d,i} = k, \mathbf{l}_{d,i} = l, \mathbf{r}_{d,i} = r | \mathbf{w}_{d,i} = w, \mathbf{t}_{d,i} = t, \mathbf{z}_{-(d,i)}, \mathbf{l}_{-(d,i)}, \mathbf{r}_{-(d,i)}, \alpha, \beta, \gamma, \eta) \propto \frac{n_{t,k,l,z,w}^{-(d,i)} + \beta_{l,r}^{w}}{n_{t,k,l,r}^{-(d,i)} + \sum_{j=1}^{V} * \beta_{l,r}^{j}} \frac{n_{d,t,k,l,z}^{-(d,i)} + \eta}{n_{d,t,k,l}^{-(d,i)} + R * \eta} \frac{n_{d,t,k,l}^{-(d,i)} + \gamma}{n_{d,t,k}^{-(d,i)} + L * \gamma} \frac{n_{d,k}^{-(d,i)} + \alpha}{n_{d}^{-(d,i)} + K * \alpha} \quad (2)$$

где $n_{t,k,l,r,w}$ означает количество раз, когда слову *w* с оценочным тэгом *t* присвоены тема *k*, тональная метка l и проблемная метка *r* в коллекции документов, $n_{d,t,k,l,r,w}$ означает число раз, когда слову *w* с оценкой *t* присвоены тема *k*, тональная метка l и проблемная метка *r* в документе *d,* $\beta_{l,r}^{w}$ обозначает априорное распределение Дирихле на Φ для слова *w* с тональной меткой l и проблемной меткой *r.*

5. Эксперименты и результаты

Для наших экспериментов мы использовали отзывы об автомобилях на русском языке, опубликованные в рамках дорожки анализа тональности соревнования SentiRuEval-2015 [16], и отзывы пользователей на английском языке компании Amazon[1] в четырех различных предметных областях. В качестве тестового корпуса использовались размеченные предложения пользователей из работ [3, 17]. Тестовый корпус содержит бинарную классификацию: предложения без упоминания проблем (no-problem класс) и с проблемными высказываниями (problem класс). Морфологическая обработка текста осуществлялась с помощью библиотеки NLTK[2] для английского языка и Mystem[3] для русского языка: на этапе предварительной обработки текстов была выполнена лемматизация и стемминг всех слов, удалены стоп-слова. Дополнительно из отзывов на русском языке были удалены слова, встречающиеся в корпусе менее двух раз. К словам рядом с отрицанием поставлен префикс *neg*. Таблица 2 содержит статистику по обучающей и тестовой коллекции.

Табл. 2. Статистика обучающей и тестовой коллекций.

Предметная область отзывов	**Количество отзывов с оценкой *r***						**Тестовый корпус**	
	r=1	*r=2*	*r=3*	*r=4*	*r=5*	размер словаря V	# problem	# no-problem
Электроника	820	598	922	2211	5450	35610	498	222
Детские товары	327	195	258	570	1498	9452	780	363
Инструменты для дома	873	503	965	19115	5745	23706	611	239
Машины (анг.)	399	239	361	1135	3760	13314	827	171
Машины (рус.)	40	150	634	3041	4061	22958	774	2824

Для сравнения результатов классификации были выбраны популярные вероятностные модели: модель тональность-тема (joint sentiment-topic model, JST) [10], модель тема-тональность (Reverse-JST) [10], модель аспект-тональность (aspect and sentiment unification model, ASUM) [11], модель User-aware Sentiment Topic Models (USTM) [12]. Эти модели объединяют тематическую и тональную переменные, USTM так же включает в распределения метаданные профайлов пользователей. Для обучения USTM в качестве метаданных используется информация о месте жительства пользователей из их личных профилей Amazon. Чтобы избежать

[1] Корпус отзывов доступен по ссылке https://snap.stanford.edu/data/web-Amazon.html.
[2] http://www.nltk.org/
[3] https://tech.yandex.ru/mystem/

разреженности тематик в USTM, текстовая коллекция собрана с учетом двадцати пяти наиболее часто встречающихся местоположений, указанных в профилях пользователей, для каждой предметной области. Поскольку корпус об автомобилях на русском языке не содержит информацию о пользователях, модель USTM не применима.

В качестве тонального лексикона (SL) используется словарь MPQA[4], 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 Викисловаря[5] и описанный в статье [18]. В качестве словаря проблемных индикаторов (PL) используются слова, описанные в [3, 17]. На основе выбранных словарей задаются асимметричные априорные распределения Дирихле с гипепараметром β. Начальные значения гиперпараметров β_w для всех слов равны 0.1. Затем значения гиперпараметров β_{lw} для эмоционально-окрашенных слов определяются следующим, характерным для современных работ, образом: если существует вхождение слова в лексикон SL с позитивной пометкой, то $\beta_{*w} = (1{,}0.1, 0.01)$ (1 для позитивных меток (pos), 0.1 для нейтральных (neu), 0.01 для негативных (neg)); для слов с негативной пометкой $\beta_{*w} = (0.01{,}0.1{,}1)$. Аналогично определяются гиперпараметры β_{rw} для проблемных индикаторов на английском языке: если существует вхождение слова в словарь PL с проблемной пометкой, то $\beta_{*w} = (1, 0.01)$ (1 для проблемных меток (pr), 0.01 для меток, указывающие на отсутствие проблемы (no-pr)); проблемный индикатор с префиксом отрицания получает значения β_{*w}= $(0.01, 1)$. Для русского языка гиперпараметры β_{rw} задаются следующим образом: если существует вхождение слова в словарь PL с проблемной пометкой, то $\beta_{*w} = (0.001, 0.0001)$, для проблемного индикатора с префиксом отрицания $\beta_{*w} = (0.00001, 0.001)$. Гиперпараметр β_{lrw} определяется как сумма β_{lw} и β_{rw} ($l \in \{\text{neu}, \text{pos}, \text{neg}\}, r \in \{\text{pr}, \text{no} - \text{pr}\}$).

В качестве критерия качества построенных моделей используется перплексия контрольных данных (perplexity) [19]. 90% отзывов использовано в качестве обучающей выборки для вероятностных моделей, 10% отзывов использованы для тестирования. Для всех моделей при решении задачи статистического оценивания использовалось сэмплирование Гиббса, число итераций равно 1000. Все эксперименты проводились при следующих параметрах: α = 50/K, β=0.01, γ = 0.01 ∗ AvgLen/S, η = 0.01 ∗ AvgLen/R, где AvgLen обозначает среднее количество слов в отзыве, R=2, S=3, K=5. Для USTPM число различных пользовательских метаданных о географическом местоположении пользователя (T) равно 25. Результаты экспериментов представлены в таблице 3. Модель TSPM наименьшие значения перплексии среди моделей JST, Reverse-JST и USTM в коллекциях о детских товарах и инструментах, где

[4] http://mpqa.cs.pitt.edu/
[5] https://ru.wiktionary.org/wiki

любое слово документа зависит от скрытой тональной переменной. Предложенная модель RTSPM показывает наименьшие значения перплексии среди всех тематических моделей, что характеризует лучшую способность RTSPM предсказывать появление слов w в документах коллекции в зависимости от оценочного тэга, темы, тонального и проблемного контекста.

Табл. 3. Перплексия вероятностных моделей.

Метод	**Коллекции отзывов пользователей о продуктах разных категорий**				
	Электроника	Инструменты	Детские товары	Машины (анг.)	Машины (рус.)
JST+SL	6139.432	2934.682	1899.184	4114.005	4293.087
Reverse-JST+SL	5833.777	3345.068	2322.585	4405.159	4691.698
ASUM+SL	2544.348	2321.319	1327.732	1837.371	2047.021
USTM+SL	5543.127	4963.680	2525.952	3434.839	-
JST+PL	4191.591	2768.168	1878.961	3934.514	4203.726
Reverse-JST+PL	4643.836	2913.586	1902.013	4661.959	4470.233
ASUM+PL	2346.913	2580.232	1368.797	2294.031	2044.301
USTM+PL	5346.548	4603.271	1959.309	2919.317	-
TSPM	4006.628	2524.141	1572.788	4314.725	4296.813
RTSPM	2081.009	1433.861	759.591	1354.284	1824.547

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 β_{lw} (S=3) на основе словаря SL; префикс '+PL' указывает на то, что модель учитывает только проблемные метки слов и задает гиперпараметры β_{rw} (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 $p(\mathrm{l}_{neg}|d)$ выше, чем вероятность позитивного и нейтрального классов: $p(\mathrm{l}_{pos}|d)$ и $p(\mathrm{l}_{neu}|d)$; аналогично высказывание не содержит проблем с продуктами, если $p(\mathrm{l}_{pos}|d)$ выше $p(\mathrm{l}_{neg}|d)$ и $p(\mathrm{l}_{neu}|d)$. Вероятности проблемного и тонального классов вычисляются на основе мультиномиального распределения в пространстве слов $\Phi_{z,l,r}$ по схожей формуле, описанной в [12].

Результаты классификации представлены в табл. 4 и табл. 5, в качестве критериев использовались стандартные метрики качества систем анализа текстов: аккуратность (accuracy), точность (precision) и полнота (recall) и F-мера (F1-measure). Поскольку мета-данные об авторе отзывов отсутствуют для русского языка, результаты USTM для отзывов для русского языка не описаны. Модели JST+SL, Reverse-JST+SL, ASUM+SL, USTM+SL показали наименьшие значения F-меры и аккуратности классификации по сравнению с JST+PL, Reverse-JST+PL, ASUM+PL, USTM+PL, что опровергает взаимно-однозначное соответствие негативного класса и класса проблемных высказываний. Наилучшие результаты F-меры достигают предложенные

модели TSPM и RTSPM на корпусе отзывов на английском языке, что показывает эффективность порождения слова в документах в зависимости от некоторой скрытой темы, тональной и проблемной информации. TSPM и RTSPM достигают сравнимые с другими вероятностными моделями результаты на корпусе отзывов на русском языке, что может быть связано с размером словарей позитивных и негативных слов (словарь SL для английского содержит 7629 слов, словарь SL для русского языка содержит около 3800 слов) или с необходимостью более сложной линейной комбинации гиперпараметров (полученной с помощью EM алгоритма (expectation-maximization) или линейного классификатора).

Табл. 4. Результаты классификации предложений отзывов пользователей о детских товарах и инструментах для дома.

Метод	**Инструменты для дома**				**Детские товары**			
	A	P	R	F1	A	P	R	F1
JST+SL	.442	.718	.368	.487	.569	.788	.481	.597
Reverse-JST+SL	.526	.747	.513	.609	.508	**.797**	.348	.485
ASUM+SL	.481	.741	.427	.542	.581	.794	.499	.613
USTM+SL	.536	.766	.511	.612	.500	.797	.328	.465
JST+PL	.587	.728	.677	.702	.497	.665	.490	.564
Reverse-JST+PL	.647	**.763**	.738	.750	.533	.665	.598	.630
ASUM+PL	.572	.731	.641	.684	**.621**	.742	.658	.698
USTM+PL	.514	.678	.617	.646	.480	.616	.574	.595
TSPM	**.664**	.724	**.857**	**.786**	.577	.698	635	.665
RTSPM	.622	.714	.792	.751	.618	.691	**.765**	**.726**

Табл. 5. Результаты классификации предложений отзывов пользователей об электронике и машинах.

Метод	**Электроника**				**Машины (анг)**				**Машины (рус.)**			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
JST+SL	.408	.71	.41	.523	.244	**.92**	.10	.176	.614	.27	.47	.345
Reverse-JST+SL	.457	.70	.38	.494	.294	.89	.17	.285	.617	**.28**	.49	.354
ASUM+SL	.523	.71	.53	.606	.457	.85	.42	.563	.666	.30	.42	.352
USTM+SL	.564	.72	.61	.657	.373	.88	.28	.429	-	-	-	-
JST+PL	.689	.71	.93	.804	.671	.84	.74	.789	**.547**	.25	.55	.343
Reverse-JST+PL	.689	.72	.91	.802	.678	.83	.77	.798	.546	.26	.62	.368
ASUM+PL	.588	.74	.63	.679	.604	.84	.65	.729	.498	.23	.58	.331
USTM+PL	.575	.67	.77	.715	.529	.81	.57	.667	-	-	-	-
TSPM	**.693**	.71	.95	**.811**	.699	.83	**.80**	**.814**	.517	.26	.67	**.376**
RTSPM	.654	.71	.83	.769	**.701**	.83	**.80**	**.815**	.380	.23	**.78**	.351

В таблице 6 приведены примеры тем с различными проблемными метками, полученные с помощью моделей JST и TSPM. Слова из словарей PL и SL

выделены курсивом. По каждой теме представлены наиболее вероятностные слова. Согласно тональным меткам, TSPM по сравнению с JST+PL различает темы, где возможные неполадки или дискомфорт в использовании автомобилей связаны с внешними факторами (*зима, покупка, впечатление*) в отличие от тем, где существующие проблемные фразы связаны с составными элементами или функциями автомобилей (*крыло, ездить, ремонт двигателя*).

Табл. 6. Примеры тематических слов из отзывов о машинах на русском языке.

JST+PL		TSPM					
no-pr.	**probl.**	**negative**		**positive**		**neutral**	
		no-pr.	**probl.**	**no-pr.**	**probl.**	**no-pr.**	**probl.**
автомобиль	машина	зима	кузов	автомобиль	Машина	машина	масло
хороший	купить	машина	бампер	**хороший**	ездить	купить	поменять
качество	новый	заводиться	дверь	общий	расход	новый	замена
цена	деньги	мороз	краска	впечатление	менять	решать	двигатель
класс	машинка	печка	порог	авто	трасса	месяц	менять
модель	продавать	лето	крыло	салон	город	продавать	**проблема**
кузов	сразу	**проблема**	покрытие	двигатель	купить	деньги	приходится
надежный	решать	быстро	удар	расход	литр	становиться	задний
дизайн	тысяча	ездить	коррозия	**отличный**	масло	покупка	**ремонт**

6. *Заключение*

Исследование в рамках статьи рассматривает задачу анализа мнений пользователей о продуктах на русском и английском языках. Целью исследования является резюмирование слов в отзывах пользователей по тематическим отзывам, используя взаимосвязи между проблемными индикаторами, эмоционально-окрашенными (тональными) словами пользователя и категориями аспектных терминов продуктов при моделировании слов в отзывах. Для достижения целей исследования в статье представлены тематические модели на основе модели латентного размещения Дирихле: (i) модель *тема-тональность-проблема (topic-sentiment-problem model)* и (ii) модель *оценка пользователя-тема-тональность-проблема* (rating-aware topic-sentiment-problem model). Предложенные модели объединяют знания на основе словарей тональных слов и проблемных индикаторов с помощью асимметричных гиперпараметров для всех слов документов. В статье оценка качества предложенных методов анализируются в сравнении с

результатами популярных модификаций латентного размещения Дирихле для задач анализа мнений. Предложенные модели достигают наилучшие результаты F-меры и сравнимые значения перплексии в сравнении с другими вероятностными моделями. В последующих исследованиях планируется улучшить предложенные модели за счет изменения линейной комбинации гиперпараметров с помощью EM алгоритма или методов машинного обучения.

Список литературы

[1]. Сабирова И.М. Качество–ключевой фактор обеспечения конкурентности продуктов и услуг в условиях рыночной экономики. Автоматизация и управление в технических системах, №1, 2015, С. 181-190.

[2]. 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.

[3]. Solovyev V., Ivanov V. Dictionary-Based Problem Phrase Extraction from User Reviews. Text, Speech and Dialogue, Springer International Publishing, 2014, P. 225-232.

[4]. Moghaddam S. Beyond Sentiment Analysis: Mining Defects and Improvements from Customer Feedback. Advances in Information Retrieval, Springer International Publishing, 2015, P. 400-410.

[5]. Tutubalina E. Target-Based Topic Model for Problem Phrase Extraction. Advances in Information Retrieval, 2015, P. 271-277.

[6]. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation. The Journal of machine Learning research., T. 3, 2003, P. 993-1022.

[7]. Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, T. 5. , 2012, P. 1-167.

[8]. Martínez-Cámara E, Martín-Valdivia M. T., Urena-López L. A., Montejo-Ráe, A. R. Sentiment analysis in twitter. Natural Language Engineering, T. 20(1), 2014, P. 1-28.

[9]. 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., P. 803-812.

[10]. Lin C., He Yu., Everson R., Ruger S. Weakly supervised joint sentiment-topic detection from text. Knowledge and Data Engineering, IEEE Transactions on, T. 24(6), 2012, P. 1134-1145.

[11]. Jo Y., Oh A. H. Aspect and sentiment unification model for online review analysis. Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, P. 815-824.

[12]. Z. Yang, A. Kotov, A. Mohan S. Lu. Parametric and Non-parametric User-aware Sentiment Topic Models. Proceedings of the 38th ACM SIGIR, 2015.

[13]. Heinrich G. Parameter estimation for text analysis. Technical report, 2005.

[14]. 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., P. 352-359.

[15]. Griffiths T. L., Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences, T. 101 (1), 2004, P. 5228-5235.

[16]. Loukachevitch N., Blinov P. , Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. Proceedings of International Conference Dialog-2015, Moscow, Russia, 2015.

[17]. Тутубалина Е. В. Извлечение проблемных высказываний, связанных с неисправностями и нарушением функциональности продуктов, на основании отзывов пользователей. «Вестник КГТУ им. А.Н.Туполева», Т. 3, 2015.
[18]. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars. Proceedings of International Conference "Dialog-2015", Moscow, Russia, 2015.
[19]. Воронцов К. В., Потапенко А. А. Регуляризация, робастность и разреженность вероятностных тематических моделей. Компьютерные исследования и моделирование, Т. 4 (4), 2012, С. 693-706.

Sentiment-Based Topic Model for Mining Usability Issues and Failures with User Products

E.V. Tutubalina <ElVTutubalina@kpfu.ru>
Kazan (Volga Region) Federal University,
18 Kremlyovskaya Str., Kazan, 420008, Russian Federation

Abstract. This paper describes an approach to problem phrase extraction from texts that contain user experience with products. During the last decades, consumer products have grown in complexity, and consumer dissatisfaction is increasingly caused by usability problems, in addition to problems with technical failures. Moreover, user reviews from online resources, that describe actual difficulties with products experienced by users, affect on other people's purpose decisions. In this paper, we present two probabilistic graphical models which aim to extract problems with products. We modify Latent Dirichlet Allocation (LDA) to incorporate information about problem phrases with words' sentiment polarities (negative, neutral or positive). The proposed models learn a distribution over words, associated with topics, both sentiment and problem labels. Topic models were evaluated on reviews of different domains collected from online consumer review platforms. The algorithms achieve a better performance in comparison to several state-of-the-art models in terms of the likelihood of a held-out test and in terms of an accuracy of classification results. Qualitative analysis of the topics discovered using the proposed models indicates the utility of considering sentiment information in users' critical feedback. Our contribution is that incorporating sentiment and problem information about words with reviews' topics by the model's asymmetric priors gives an improvement for problem phrase extraction from user reviews published in English and Russian.

Keywords: opinion mining, problem phrases, user reviews, mining defects, topic modeling, LDA, problem phrase extraction.

DOI: 10.15514/ISPRAS-2015-27(4)-6

For citation: Tutubalina E.V. Sentiment-Based Topic Model for Mining Usability Issues and Failures with User Products. *Trudy ISP RAN/Proc. ISP RAS*, vol. 27, issue 4, 2015, pp. 111-128 (in Russian). DOI: 10.15514/ISPRAS-2015-27(4)-6.

References

[1]. Sabirova I.M. [Quality-key factor of ensuring competition of products and services in the conditions of market economy]. Avtomatizaciya i upravlenie v tehnicheskih sistemah [Automation and management in technical systems], vol. 1, 2015, pp. 181-190. (In Russian)
[2]. 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. Beijing, China, 2011.
[3]. Solovyev V., Ivanov V. Dictionary-Based Problem Phrase Extraction from User Reviews. Text, Speech and Dialogue, Springer International Publishing, 2014, vol. 225-232.
[4]. Moghaddam S. Beyond Sentiment Analysis: Mining Defects and Improvements from Customer Feedback. Advances in Information Retrieval, Springer International Publishing, 2015, PP. 400-410.
[5]. Tutubalina E. Target-Based Topic Model for Problem Phrase Extraction. Advances in Information Retrieval, 2015, pp. 271-277.
[6]. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation. The Journal of machine Learning research., vol. 3, 2003, pp. 993-1022.
[7]. Liu B. Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, T. 5. , 2012, pp. 1-167.
[8]. Martínez-Cámara E, Martín-Valdivia M. T., Urena-López L. A., Montejo-Ráe, A. R. Sentiment analysis in twitter. Natural Language Engineering, T. 20(1), 2014, PP. 1-28.
[9]. 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.
[10]. Lin C., He Yu., Everson R., Ruger S. Weakly supervised joint sentiment-topic detection from text. Knowledge and Data Engineering, IEEE Transactions on, vol. 24(6), 2012, pp. 1134-1145.
[11]. Jo Y., Oh A. H. Aspect and sentiment unification model for online review analysis. Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 815-824.
[12]. Z. Yang, A. Kotov, A. Mohan S. Lu. Parametric and Non-parametric User-aware Sentiment Topic Models. Proceedings of the 38th ACM SIGIR, 2015.
[13]. Heinrich G. Parameter estimation for text analysis. Technical report, 2005.
[14]. 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., pp. 352-359.
[15]. Griffiths T. L., Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences, vol. 101 (1), 2004, pp. 5228-5235.
[16]. Loukachevitch N., Blinov P., Kotelnikov E., Rubtsova Y., Ivanov V., Tutubalina E. SentiRuEval: testing object-oriented sentiment analysis systems in Russian. Proceedings of International Conference Dialog-2015, Moscow, Russia, 2015.
[17]. Tutubalina E.V. [Extracting problem phrases about product defects and malfunctions in user reviews about cars]. Vestnik KGTU im. A.N.Tupoleva [Proceeding of KGTU im. A.N.Tupoleva], vol. 3, 2015. (in Russian)
[18]. Ivanov V., Tutubalina E., Mingazov N., Alimova I. Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars. Proceedings of International Conference "Dialog-2015", Moscow, Russia, 2015.
[19]. Vorontsov K.V., Potapenko A.A. [Regularization, robustness and sparsity of probabilistic topic models]. Komp'yuternyye issledovaniya i modelirovaniye [Computer research and modeling], vol. 4 (4), 2012, pp. 693-706. (in Russian).