

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

191023 Санкт-Петербург, наб.р.Фонтанки,27
тел. (812) 312-40-58, факс (812) 310-53-77
e-mail: admin@pdmi.ras.ru

ИНН/КПП 7825351570 / 784101001

11.05.2016 № 11102/33/02-2181

ГС

«Утверждаю»
директор Федерального государственного бюджетного учреждения науки Санкт-Петербургского отделения Математического института им. В. А. Стеклова Российской академии наук
Чл.-корр. РАН
Кисляков С.В.

**Отзыв ведущей организации о диссертации
Тутубалиной Елены Викторовны**

**«МЕТОДЫ ИЗВЛЕЧЕНИЯ И РЕЗЮМИРОВАНИЯ
КРИТИЧЕСКИХ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ О ПРОДУКЦИИ»,
представленную на соискание**

**ученой степени кандидата физико-математических наук по специальности
05.13.11 – Математическое и программное обеспечение вычислительных
машин, комплексов и компьютерных сетей**

Диссертация Е.В. Тутубалиной посвящена методам и моделям для извлечения информации о различных недостатках, трудностях в использовании и других причинах претензий, высказанных потребителями товаров и услуг в свободной форме в виде текстов отзывов и рецензий, опубликованных в сети Интернет. Тематика диссертации относится к области интеллектуального анализа текстов, и актуальность её несомненна: извлечение информации из неструктурированного текста является одной из важнейших задач современного интеллектуального анализа текстов и имеет прямые практически важные применения как в анализе собственно пользовательских отзывов, так и в других задачах. Работа полностью соответствует заявленной специальности 05.13.11.

Текст диссертации построен следующим образом. В Главе 1 даётся обзор литературы по теме диссертации (анализ тональности, классификация текстов, анализ мнений, идентификация аспектных терминов и аспектный анализ тональности), а также даётся весьма полный обзор литературы по основным проблемам диссертации – анализу проблемных высказываний пользователей, автоматическому резюмированию целевых объектов продуктов и тональных высказываний.

Глава 2 формально вводит основные задачи: извлечение проблемных высказываний, классификация высказываний, создание словарей тональной лексики, а затем анализирует проблемные фразы с помощью метода, основанного на правилах и учитывающего структуру сложных предложений относительно союзов. Предложенный метод описан с помощью контекстно-свободной грамматики, покрывающей большую долю реально встречающихся проблемных высказываний. Экспериментальные результаты показывают высокое качество полученной классификации. Дополнительно, автор анализирует вклад созданных словарей в виде признаков для ряда стандартных классификаторов, показывая улучшение качества классификации.

В Главе 3 задача извлечения проблемных высказываний уточняется далее: теперь задача состоит в том, чтобы извлечь проблемные высказывания о конкретных целевых объектах (продуктах, частях продуктов, их свойствах), а также установить соответствие между высказываниями и целевыми объектами. Предложен метод, основанный на синтаксических связях слов в предложении и семантической близости между терминами понятия (описывающими объект) и терминами предметной области; полученные экспериментальные результаты указывают на высокое качество выделения терминов и высказываний, а также основанной на них классификации.

Глава 4 ставит следующую, ещё более амбициозную задачу: автоматически резюмировать проблемные высказывания; практическая мотивация задачи состоит в том, чтобы из большого массива отзывов выделить конкретные претензии, которые компании-производителю следует устранить в первую очередь. Для этого вводятся две новые модификации тематической модели латентного размещения Дирихле со специальными переменными для тональности и проблемных слов. В диссертации проводится вывод алгоритмов обучения предложенных моделей на основе сэмплирования по Гиббсу. Экспериментальные результаты показывают, что предложенные модификации работают лучше ранее известных тематических моделей с учётом тональности как с точки зрения перплексии (т.е. моделирования корпуса текстов), так и с точки зрения результатов классификации отзывов пользователей.

Таким образом, основными результатами диссертационной работы являются:

- новые алгоритмы выделения и анализа проблемных высказываний, основанные, в частности, на построенных в диссертации словарях и правилах, описанных в виде контекстно-свободной грамматики проблемных высказываний, учитывающей союзы в сложных предложениях;
- новый алгоритм выделения взаимосвязей между проблемными высказываниями и целевыми объектами;
- новые тематические модели, предназначенные для автоматического резюмирования проблемных высказываний, алгоритмы вывода в этих моделях;
- программные реализации всех разработанных алгоритмов.

Указанные результаты являются новыми, их достоверность не вызывает сомнений. Принятые в работе допущения и ограничения обоснованы и отражены в полном объёме. Ошибок в доказательствах, выводах и постановках экспериментов не обнаружено. В диссертации есть как теоретические, так и прикладные результаты; практическая применимость результатов не вызывает сомнений и доказана в самой работе.

В качестве сильной стороны работы хочется специально отметить детальный и очень интересный анализ результатов экспериментов. Так, в главах 2 и 3 диссертант не ограничивается таблицами численных результатов для точности и полноты различных классификаторов, а отдельно проводит качественный анализ ошибок, вручную выделяя разные типы ошибок предложенных алгоритмов и предлагая возможные их причины; в Главе 4 приводятся и анализируются примеры тем, выделенных при помощи предложенных тематических моделей.

Тем не менее, любопытно, что анализ результатов можно считать и одним из недостатков работы. Несмотря на то, что во всех трёх главах приводятся примеры конкретных результатов модели, это всегда конечный результат, что не всегда позволяет проследить работу предложенных алгоритмов. Так, в Главе 3 было бы очень интересно увидеть таблицу с примерами исходных предложений и извлечённых из них проблемных высказываний, объектов и связей между ними. Аналогично, в Главе 2 за представленным качественным анализом крайне трудно проследить, потому что в нём нет ссылок на конкретные применимые к приведённым примерам правила из грамматики (а правил в грамматике достаточно много). А в Главе 4 было бы интересно увидеть таблицу со сравнением тем двух предложенных моделей и ранее известных тематических моделей с учётом тональности (JST, ASUM, USTM): какие новые темы выделились по сравнению со старыми моделями, какие больше не выделяются. Кроме того, из недостатков можно отметить то, что в Главе 4 (4.4.2) выбор гиперпараметров для предложенных тематических моделей не мотивирован (кроме

β), в то время как существуют методы автоматического подбора гиперпараметров, которые могли бы улучшить полученные модели.

Ещё одно замечание – достаточно большое число опечаток, стилистических и пунктуационных ошибок в диссертации. Иногда опечатки и текст, оставшийся от копирования из работ диссертанта, затрудняют понимание: например, внизу с. 84 написано «в статье предложены», и приходится задуматься, о какой статье идёт речь (на самом деле, видимо, имелось в виду «в диссертации»). Однако в основном эти недочёты не принципиальны.

Тем не менее, указанные замечания не ставят под сомнение ценность основных результатов работы. Диссертация является законченной научно-квалификационной работой, выполненной автором самостоятельно на высоком научном уровне. Основные этапы работы, её выводы и результаты полностью отражены в автореферате. Результаты диссертационной работы рекомендуются к использованию в исследованиях по интеллектуальному анализу текстов в ПОМИ РАН, НИУ ВШЭ, ФГАОУ ВО КФУ и других научных и учебных организациях. В частности, прикладные результаты работы могут быть напрямую использованы в программных системах и продуктах для анализа мнений пользователей, анализа тональности, извлечения знаний из порождённых пользователями текстов.

Таким образом, диссертационная работа Тутубалиной Е.В. удовлетворяет всем требованиям ВАК, предъявляемым к кандидатским диссертациям, а её автор заслуживает присвоения ей учёной степени кандидата физико-математических наук по специальности 05.13.11 – математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей.

Отзыв обсуждён и утверждён на заседании лаборатории математической логики Санкт-Петербургского отделения Математического института им. В.А. Стеклова РАН 06.05.2016 г., протокол № 2.

Заведующий лабораторией
математической логики ПОМИ РАН,
д.ф.-м.н., акад. Матияевич Юрий Владимирович