ОТЗЫВ

официального оппонента о диссертации

Тутубалиной Елены Викторовны

«Методы извлечения и резюмирования критических отзывов пользователей о продукции»

представленной к защите на соискание степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

1. Актуальность темы диссертации

Диссертационная работа Е. В. Тутубалиной посвящена разработке моделей и методов извлечения информации о высказываниях пользователей, содержащих указания на трудности в использовании продуктов (сервисов, товаров). В настоящее время задача анализа мнений авторов текстов по отношению к продуктам, обсуждаемым в отзывах и комментариях пользователей на страницах web-ресурсов, является актуальной и важной задачей. Иллюстрацией актуальности извлечения высказываний, связанных с неисправностями и нарушением функциональности продуктов, является то, что на рынке потребительских товаров наблюдается резкая динамика увеличения количества технически сложных товаров. Покупатели публикуют в сети Интернет отзывы о продукции, описывая возникшие затруднения с использованием продукта дополнительно к инцидентам о ненадлежащем техническом качестве. Неудовлетворенность продукцией может повлечь отрицательную рекламу и требует от компаний устранения инцидентов. Несомненным подтверждением актуальности является то, что в последнее время участились случаи выпуска продуктов питания, содержащих пальмовое масло и ГМО, влияющих на здоровье и жизнь населения, без указания этого на этикетке или с указанием этого в недоступных местах (на общей упаковке), мелким шрифтом. Парадоксом является то, что такие

продукты выпускаются под маркой существующих, хорошо зарекомендовавших себя брендов, продаются по высокой цене, и единственным источником информации для населения остается сеть Интернет.

Таким образом, диссертация Е. В. Тутубалиной посвящена актуальной теме.

2. Общая характеристика диссертационной работы

Диссертация состоит из введения, четырех глав, заключения, списка литературы, содержащего 150 наименований, и одного приложения. Общий объем работы составляет 145 страниц.

Во **введении** обосновывается актуальность диссертационной работы, формулируются цели и задачи представляемой работы, описывается методика исследования, апробация работы, практическая значимость работы и представляются выносимые на защиту основные положения.

В первой главе приводится описание существующих методов и подходов, применяемых в задачах анализа мнений пользователей. Проанализированы основные достоинства и недостатки существующих методов. На основании проведенного анализа делается заключение о том, что исследования на основе существующих методов машинного обучения чаще всего сводятся к созданию обучающей выборке и векторов признаков. Одной из ключевых задач остается задача создания словарей, использование которых позволяет повысить качество моделей, и разработка методов, не зависящих от предметной области и не требующих размеченных ресурсов.

Вызывает уважение тот факт, что Е. В. Тутубалина выполнила большой объем работ при подготовке обзора литературы, собрав обширный список публикаций, касающихся различных задач анализа мнений, включая классификацию текстов, извлечение аспектных терминов, идентификацию оценочных слов, резюмирование мнений по тематическим категориям. Обзор читается с большим интересом, многие выводы точно подмечены. Проведена четкая граница между задачей анализа тональности текста в целом и задачей

оценки мнений по различным аспектам продукта (качеству, функциям, составу, сервису и т.д.) (стр. 17).

Вторая глава посвящена задаче автоматического извлечения из текстов пользователей высказываний, указывающих на проблемные ситуации в использовании продуктов. Предложен метод классификации предложений, основанный на знаниях в виде словарей и правилах, учитывающих грамматическую структуру сложных предложений относительно союзов. Для достижения целей созданы набор словарей и коллекция из текстов пяти предметных областей на русском и английском языках. Представленные в данной главе эксперименты показывают улучшение качества классификации согласно значениям F-меры, полученной макроусреднением, относительно ряда методов машинного обучения с учителем (см. табл. 4-6, стр. 56-57). Статистическая значимость результатов подтверждена с помощью непараметрического статистического критерия знаковых рангов Вилкоксона, из которой следует, что предложенный в диссертации метод классификации улучшает качество классификации по сравнению с существующими моделями. Также подтверждается вклад ряда признаков, основанных на созданных словарях, для улучшения классификации с помощью существующих методов машинного обучения.

Впечатлили работы по составлению двуязычного словаря (русский, английский) проблемных слов, индикаторов действий и т.д. с использованием внешних источников (с. 40-42, табл. 2). Русский словарь не уступает по полноте английскому, что свидетельствует о большой проведенной работе и новизне методики. Словари приведены в Приложении А диссертационной работы.

Также на стр. 45-48 приведено формальное описание применяемой грамматической модели в виде контекстно-свободной грамматики.

На стр. 56-57 в таблицах 4-6 на примере «сухих цифр» продемонстрирован весь драматизм этой работы. Автор применил разработанные словари и формальные грамматики, а они не дали

абсолютного превосходства предложенных моделей. По моим оценкам новые методы превзошли существующие примерно в 50-ти процентах случаев. Но это, кстати, не является недостатком диссертации. Просто работы в этой области ведутся довольно давно, они востребованы, постоянно совершенствуются, и надо совершить «чудо», чтобы получился даже такой результат. Кстати, автор не учел такой фактор, как стоимость обучения модели, который в ее случаях оказался бы гораздо ниже, чем в сравниваемых методах.

После этого автор пошел в статистику. См. 4 главу.

В **третьей главе** описывается новый алгоритм извлечения высказываний, указывающих на проблемные ситуации, по отношению к предметно-ориентированным целевым объектам на основе общедоступного тезауруса. Под предметно-ориентированными целевыми объектами понимаются связанные с продуктом понятия, существенные в определенной предметной области. Для определения множества возможных целевых объектов используются синтаксические связи между проблемными индикаторами и существительными с помощью зависимостей слов в предложении. Для проверки принадлежности целевого объекта к предметной области изучается возможность применения мер семантической связанности. В главе рассматриваются меры нескольких типов. Представленные в главе эксперименты показывают улучшение качества согласно значениям F-меры для текстов трех предметных областей (машины, инструменты, детские товары) с помощью предложенного алгоритма извлечения проблемных высказываний по отношению к целевым объектам (см. табл. 14-15, стр. 79). Наилучшие значения точности показывают методы проверки семантической связанности с помощью меры Lesk, основанной на определениях понятий в тезаурусе, и косинусной меры, использующей вектора распределённых представлений слов.

В **четвертой главе** предлагаются две вероятностные модели для задачи выделения тематически сгруппированных объектов мнений. Модель

тематических высказываний, указывающих на проблемные ситуации (TPrPhModel), рассматривает распределения проблемных индикаторов и целевых объектов как независимые мультиномиальные распределения в пространстве слов. Модель тема-тональность-проблема (TSPM) учитывает скрытые тематическую, тональную и проблемную переменные о словах совместно для анализа взаимосвязи между информацией о проблемных ситуациях и тональности высказываний относительно тематических категорий. В главе приводится статистическое оценивание моделей с помощью сэмплирования Гиббса (см. стр. 89-92, 96-98). Представленные в данной главе эксперименты показывают значительное улучшение качества тематических моделей относительно существующих модификаций моделей латентного размещения Дирихле для задач анализа мнений. Модель TPrPhModel показывает улучшение качества тематических моделей согласно значениям перплексии независимо от коллекции (см. табл. 20, стр. 104). Показан вклад скрытых переменных, которые улучшают качество классификации с помощью предложенных моделей в рамках задачи классификации (см. табл. 21-22, стр. 105; рис. 4.3, стр. 107). Также проведенный анализ тем для качественной оценки тематических распределений показывает, что предложенные модели корректно ассоциируют слова в зависимости от проблемной ситуации с продукцией.

3. Основные результаты диссертационной работы

В качестве основных результатов диссертации следует выделить:

1) Предложен и реализован метод классификации предложений, основанный на знаниях в виде созданных словарей и правилах, учитывающих грамматическую структуру сложных предложений относительно союзов.
2) Предложен и реализован метод классификации предложений отзывов пользователей по отношению к целевым объектам, связанным с предметной областью, на основе синтаксических связей слов и мер семантической связанности.

3) Предложена и реализована совместная вероятностная модель для задачи резюмирования высказываний с целевыми объектами и проблемными индикаторами, основанная на двух независимых мультиномиальных распределениях во множестве слов.
4) Предложена и реализована вероятностная модель, учитывая скрытые тематическую, тональную и проблемную переменные о словах совместно.
5) Разработано программное обеспечение и проведено экспериментальное исследование, подтверждающее улучшение качества предложенных методов по сравнению с существующими алгоритмами.

1. Оценка новизны полученных результатов, степень обоснованности и достоверности научных положений

Для **оценки новизны** следует отметить, что задача извлечения информации о высказываниях пользователей, указывающих на проблемные ситуации с продуктами, являются недостаточно изученными в литературе. В диссертационной 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.

Теоретическая и практическая значимость заключается в том, что разработаны методы и модели извлечения информации о высказываниях пользователей о неполадках с продуктами, основанные на анализе структуры текстовых фрагментов мнений как связного текста. Предложенные методы извлечения высказываний из коллекции отзывов предметной области могут быть использованы при решении прикладных задач анализа мнений: классификации текстовых документов, извлечения информации, кластеризации информации на основе тематических моделей.

В рамках диссертационной работы была разработана программная система на основе предложенных методов и моделей. Программный комплекс по извлечению высказываний пользователей и построению

тематических моделей выложен в открытый доступ, модули извлечения высказываний могут взаимодействовать друг с другом по принципу конвейера.

Достоверность полученных результатов подтверждается взаимосвязью данных экспериментов и научных выводов, апробацией работ на семинарах, российских и международных конференциях, а также научными статьями. Основные результаты по теме диссертации изложены в 10 печатных работах, 2 из которых опубликованы в журналах, рекомендованных ВАК, 6 из которых опубликованы в журналах, входящих в базу SCOPUS.

2. Замечания по работе

Вызывает само по себе уважение тщательный анализ литературы, сделанный в 1 главе диссертации, солидный список публикаций.

По диссертационной работе можно сделать следующие замечания:
Формальные замечания.

1) На стр. 9 в диссертации остался мусор: «перечисление основных конференций, симпозиумов».

2) В диссертации нет списка сокращений и условных обозначений, что явилось бы нелишним, учитывая, насколько сама диссертация и автореферат переполнены сокращенными названиями и идентификаторами различных мер и признаков.

3) Круговые диаграммы на стр. 64 представлены в цветном исполнении, что при выводе на печать дает плохое распознавание секторов и противоречит нормам ГОСТа по правилам оформления научных отчетов.

Неформальные замечания.

4) Замечание к таблице 3 (стр. 51). Данные в таблице приведены в абсолютном выражении, а в тексте далее на стр. 51 автор оперирует процентами, что затрудняет понимание.

5) На стр. 63 из рис.2.1 становится ясно, что большой процент ошибок в моделях (49 % и 34 % , соответственно для «машин» и для «мобильных приложений») связан с «недостатком словарей, ошибкой отрицаний, условий или правил». Что говорит о том, что потенциал этой модели еще далеко не исчерпан.

6) Автор не учел такой фактор, как стоимость обучения модели, который в ее случаях оказался бы гораздо ниже, чем в сравниваемых методах.

7) Результаты испытания альтернативных методов не взяты из научной печати, а моделируются автором, что понижает к ним уровень доверия.

3. Заключение о соответствии диссертации критериям, установленным Положением о порядке присуждения ученых степеней

Отмеченные замечания в целом не снижают качества проведенного диссертационного исследования. Личное участие диссертанта в выполнении теоретических и экспериментальных исследований, разработке программных средств на основе созданных методов и получении научных результатов подтверждается соответствующими публикациями. Результаты диссертации представлены в 10 статьях автора, докладывались на российских и международных научных конференциях. Автореферат диссертации правильно и полно отражает содержание работы и надлежащим образом оформлен.

Принимая во внимание актуальность темы диссертации, научную новизну и практическую значимость ее результатов, считаю, что диссертационная работа Тутубалиной Е. В. «Методы извлечения и резюмирования критических отзывов пользователей о продукции» полностью соответствует всем требованиям ВАК РФ, предъявляемым к диссертациям на соискание ученой степени кандидата физико-математических наук, а Тутубалина Елена Викторовна заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Официальный оппонент

кандидат технических наук, доцент кафедры автоматизированных систем управления федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский технологический университет (НИТУ «МИСиС»)

119991, г. Москва, Ленинский проспект, д. 4

Телефон: +7 (495) 236-41-03

Факс университета: +7 (499) 236-21-05

E-mail: pvn-65@mail.ru

В. Н. Поляков

16.05.2016