ОТЗЫВ
ОФИЦИАЛЬНОГО ОППОНЕНТА
на диссертационную работу 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»

Диссертационная работа Тутубалиной Е.В. посвящена исследованиям в области применения современных математических методов в экономической сфере, конкретно, для автоматизации выделения значимых критических формулировок из отзывов потребителей (пользователей) продуктов массового спроса. Тематика диссертационной работы имеет несомненную практическую мотивацию. При этом понятно, что условие автоматизации выделения, классификации и тематической привязки формулировок в отзывах требует исследований с целью выяснения какие математические методы окажутся более эффективными. Тут необходимо отметить, что, в отличие от естественных наук, в которых (*априори*) известно, что в основе исследуемых процессов и явлений лежат законы, базовые принципы которых известны, или, по крайне мере, для которых математический аппарат применим в известной основе, в области экономических процессов ситуация принципиально сложнее – применимость тех или иных математических методов заранее не известна и она должна определяться на практике предметных разработок при решении конкретной проблемы. В этом состоит, по нашему мнению, научная новизна проведенных исследований, так как по проблеме, которая вынесена в название диссертации, нет пока полного комплекса математических методов и программных разработок, позволяющих перевести задачу выделения значимых критических формулировок из массовых отзывов потребителей в автоматический режим с достаточным (для практики) статистическим обоснованием.

Актуальность темы диссертационной работы Е.В. Тутубалиной определяется, прежде всего тем, что в настоящее время, в результате взрывного характера развития компьютерной техники и программного обеспечения, а также появлением новых мощных математических методов, стало возможным ставить вопрос о разработке инструментов для полной автоматизации анализа больших массивов данных (в нашем случае – отзывов) по различным типам запросов от реальных участников экономической жизни. Диссертант выбрал конкретную тему в этом направлении – автоматическое выделение критических формулировок в отзывах потребителей массовых продуктов, и их тематический анализ. Отметим высокую востребованность в коммерческом мире в таких инструментах, что определяет актуальность проведенных в диссертации исследований.

Мы отмечаем следующие идеи диссертанта, которые, на наш взгляд, хорошо представляют научную значимость полученных им результатов:

1) предложение по учету в задаче классификации критических формулировок потребителей, конкретно в словарях и правилах их анализа, грамматической структуры сложных предложений относительно союзов. Подчеркнем, что это предложение хорошо соотносится с особенностями русского языка, в котором сложные предложения являются не исключением, а скорее правилом;

2) предложение по использованию мер семантической связности (близости) для классификации формулировок отзывов с привязкой к целевым объектам в предметной области;

3) предложение использовать скрытые переменные для совместного описания тем и проблемных индикаторов.

Отметим, что результаты, полученные в диссертационной работе Е.В. Тутубалиной, являются научно обоснованными, как в отношении постановки задач исследований, так и в части используемых математических методов. Положения, которые вынесены диссертантом на защиту, представляют значимый интерес для исследований конкретных проблем экономики на основе применения современных математических методов. Полученные в диссертации результаты имеют хорошие прикладные перспективы.

Диссертация состоит из 145 страниц, она включает Введение, четыре главы, Заключение, список литературы (150 источников) и два Приложения. В ней содержится 7 рисунков и 36 таблиц.

Во Введении дается обоснование актуальности темы исследований, формулируются цель и задачи диссертационной работы.

В первой главе приводится обзор и сравнительный анализ существующих методов и подходов к задачам, связанным с автоматическим выделением критических формулировок из текстовых отзывов пользователей, среди которых определены две основные группы – лингвистические методы и методы, основанные на машинном обучении. Отмечены как сильные, так и слабые стороны этих методов. Сделан вывод, что как для первой, так и второй группы методов критическим является создание новых словарей оценочных слов, в том числе и специализированных под конкретную предметную область.

Во второй главе обсуждается решение задачи автоматического выделения формулировок, которые разделяются на два класса – указывающих на проблемы, связанные непосредственно с продуктом, и проблемы (любых) других типов. Для решения этой задачи предложен подход, основанный на словарях и правилах, составленных экспертами. Конкретно, для исследований предложены два метода: метод, основанный на поиске вхождений лексических элементов из словарей (обозначенный в диссертации как *DbA*), и метод, учитывающий грамматическую структуру сложных предложений (*CbA*). Во втором методе используется контекстно-свободная грамматика. Оба метода сравнивались с методом машинного обучения с учителем по модели "bag of words" и методами обучения на словах и словосочетаниях с использованием метода максимальной энтропии и метода опорных векторов. Также были проведены сравнения с известными в литературе классификаторами NRC-Canada, NRC-Canada, KLUE и NaiveBayes. Сравнения проводились на двух русскоязычных предметных выборках и четырех английских. Для определения близости использовалась мера F_1-measure. Проведенное сравнение показало, что предложенные методы классификации дают лучшее качество по сравнению со сравниваемыми методами: порядка 1% по используемой мере F_1 для метода DbA, и улучшение от 0.005 до 0.37 по мере F_1 для метода CbA для различных групп продуктов. Анализ статистической значимости (выполненный с помощью непараметрического статистического критерия знаковых рангов Вилкоксона) проведенных сравнений показывает значимость добавления в задачу классификации правил анализа сложных предложений.

В третьей главе описывается новый алгоритм выделения критических высказываний по отношению к предметно-ориентированным целевым объектам на основе общедоступного тезауруса. Диссертантом предложен метод, основанный на синтаксических связях между проблемными индикаторами и существительными в предложении, и предлагается использовать меры семантический близости для выделения предметно-ориентированных целевых объектов. Предложенный алгоритм использует результаты анализа текстовых высказываний с помощью предложенных диссертантом методов DbA и CbA. Полученные результаты говорят о том, что использование мер семантической близости приводит к улучшению качества классификации для групп продуктов с несложной для потребителей технической стороной их использования (например, для детских товаров, но не для электроники).

В четвертой главе изложены результаты исследований по выделению тематически сгруппированных объектов мнений пользователей, выполненных с помощью метода автоматического резюмирования мнений относительно тематических категорий. Актуальность этой задачи определяется таким важным фактором, как заинтересованность компаний-

производителей в устранении недостатков, выявляемых по отзывам потребителей, относящихся к наиболее массовым претензиям к продукции. Для решения этой задачи в диссертации предложены новые модели, являющиеся модификацией модели латентного размещения Дирихле: модель тематических высказываний для определения проблемных индикаторов по тематическим категориям отзывов, и модель *тема-тональность-проблема* для анализа взаимосвязи между проблемами и тональностью высказываний о них относительно тематических категорий. Для сравнения качества предложенных моделей использовались известные тематические модели JST, Reverse-JST и ASUM. Результаты исследований, выполненных в диссертации, говорят о том, что предложенная модель тематических высказываний для определения проблемных индикаторов по тематическим категориям отзывов показывает лучшее качество по сравнению с известными моделями, что говорит о значимости учета связи темы и тональности. Лучшие результаты показала модель *тема-тональность-проблема.*

В рамках диссертационной работы была разработана программная система, в которой реализованы модели и алгоритмы, изложенные во второй, третьей и четвертой главах, и которая выложена в Интернет для открытого доступа.

В Заключении приведены основные результаты выполненных исследований и разработок.

Результаты, выносимые диссертантом на защиту, опубликованы в двух статьях, опубликованных в рецензируемых научных журналах, а также в шести статьях, опубликованных в изданиях, индексируемых Scopus и в двух публикациях в трудах международных научных конференций.

Достоверность научных положений и выводов диссертационной работы Е.В. Тутубалиной обоснована корректностью постановки задач, использованием современных математических методов и проведением детального сравнения разработанных моделей, алгоритмов и программ с известными в научной литературе базовыми моделями и программами.

Диссертация Е.В. Тутубалиной выполнена на высоком научно-техническом уровне. Тем не менее необходимо сделать следующее замечание. В выполненной работе можно отметить близость проведенных фундаментальных исследований с актуальной проблемой в реальной практической жизни коммерческих предприятий, выпускающих продукцию массового спроса. Так вот, для характеристики реального значения результатов исследований и разработок, полученных в диссертационной работе, недостаточно ограничиваться только сравнением с известными в научной литературе моделями и программами. Например, что означает для реальной коммерческой практики улучшение качества классификации по мере F_1 на 1%, или на 0.37 в абсолютных значениях? Много это или мало для оценки практической деятельности коммерческих предприятий? Или же полученные количественные характеристики улучшения качества классификации имеют важное, но академическое, достижение, базируясь на котором можно выполнять прикладные исследования с перспективой реального коммерческого эффекта?

Это замечание, однако, не снижает качества проведенного Е.В. Тутубалиной диссертационного исследования и не изменяют общую ее положительную оценку.

В заключение отметим, что представленная диссертация Е.В. Тутубалина представляет собой завершенное исследование, проведенное на высоком научно-техническом уровне.

Основное содержание диссертации отражено в опубликованных диссертантом статьях, доложено на международных и российских научных конференциях. Автореферат диссертации в целом правильно и полно отражает ее содержание.

Приходим к заключению, что диссертационная работа Тутубалиной Е.В. по теме «Методы извлечения и резюмирования критических отзывов пользователей о продукции», представленная на соискание ученой степени кандидата физико-математических наук по специальности 05.13.11 «Математическое и программное обеспечение вычислительных

машин, комплексов и компьютерных сетей» соответствует Положению ВАК о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24.09. 2013 № 842 (ред. от 30.07.2014), а её автор Тутубалина Елена Викторовна заслуживает присуждения ей ученой степени кандидата физико-математических наук по специальности 05.13.11.

Доктор физико-математических наук,
начальник Отдела информационных технологий
и математического моделирования
Курчатовского комплекса НБИКС-технологий
НИЦ «Курчатовский институт»
123182 Россия, Москва,
пл. Академика Курчатова, д. 1
тел. +7 915 016-00-48
e-mail: ilyin0048@gmail.com

В.А. Ильин
30.05.2016

Подпись сотрудника НИЦ «Курчатовский институт» В.А. Ильина заверяю

Главный ученый секрет
НИЦ «Курчатовский ин
к.ф.м.н.

С.Ю. Стремоухов