МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ САНКТ-ПЕТЕРБУРГСКИЙ ИНСТИТУТ ИНФОРМАТИКИ И АВТОМАТИЗАЦИИ РОССИЙСКОЙ АКАДЕМИИ НАУК (СПИИРАН)

199178 Санкт-Петербург, 14 линия, д.39. Тел.:(812)328-3311 Факс: (812) 328-4450;
E-mail:spiiran@iias.spb.su; http://www.spiiras.nw.ru
ОКПО 04683303, ОГРН 1027800514411 ИНН/КПП 7801003920/780101001

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Малых Валентина Андреевича "Методы сравнения и построения систем, устойчивых к шуму, в задачах обработки текстов", представленную к защите на соискание ученой степени кандидата технических наук по специальности 05.13.11 "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей"

В настоящее время проблема зашумленных текстов является широко распространенной и обнаруживающей себя во многих областях. Такое положение дел обусловлено тем, что в связи с ростом электронного документооборота значительно увеличилось количество создаваемых текстов, причем не все их авторы-пользователи компьютерных систем применяют системы проверки орфографии. В результате получаемые тексты могут содержать (и зачастую содержат) в себе опечатки, то есть шум. Такие тексты впоследствии используются в различных задачах обработки текстов, например классификации, однако их зашумленность влечет снижение качества конечных результатов, формируемых компьютерной системой обработки текстов.

Существующий подход к решению данной проблемы заключается в разработке и последовательном использовании систем автоматической проверки орфографии. Но, как показано в диссертационной работе В.А. Малых, качество результатов работы этих систем может быть недостаточным. Для разрешения этой **актуальной проблемы** автором диссертационной работы был предложен альтернативный подход, а именно создание систем обработки текстов, изначально устойчивых к шуму.

В ходе решения актуальной задачи автором получена система **результатов, новизна** которых характеризуется двумя аспектами:

1) разработаны методы сравнения систем анализа текстов по их устойчивости к шуму в различных задачах;
2) разработаны методы построения систем, устойчивых к шуму, для различных прикладных задач обработки текста.

Первый аспект позволяет осуществлять сравнение различных систем анализа текстов, независимо от их внутреннего устройства. Для сравнения используется контролируемое внесение шума в данные для обучения и/или тестирования систем, что позволяет получать оценки систем по их устойчивости к шуму для различных условий. В этой возможности заключается новизна и польза разработанных методов сравнения.

Второй аспект – предложенные методы построения устойчивых к опечаткам систем анализа текстов. Данные методы используют специально сконструированное представление текстов на основе входящих в них символов. В представленных ранее в литературе методах построения не учитывался аспект устойчивости к шуму, либо аспект устойчивости к шуму являлся вспомогательным и не рассматривался применительно к описанным задачам.

Система научных результатов состоит из 5 компонент:

1. Описан метод сравнения систем векторных представлений слов. Описанный метод позволяет сравнивать произвольные системы векторного представления слов, независимо от их внутреннего устройства с использованием контролируемого внесения шума во входные данные. Также описан метод построения систем, устойчивых к шуму, и представлена система RoVe. Данная программная система в сравнении с базовыми системами, показала лучшие результаты для большинства проверенных автором наборов данных. Результаты проверки приведены в таблицах 2-6 главы 2 и рисунках 2.6-2.8;
2. Описан метод сравнения систем классификации текстов на примере задачи систем анализа тональности. Приведенный метод позволяет сравнивать системы независимо от их внутреннего устройства. Рассмотрены несколько современных систем классификации текстов, в частности CharCNN, CharCNN-WordRNN. Также описан метод построения систем, устойчивых к шуму, в данной задаче и описана система RoVe, построенная по этому методу. Данная система показывает лучший результат для большинства поставленных экспериментов. Результаты проверки систем в задаче анализа тональности приведены на графиках 3.3 – 3.6;

3. Описан метод сравнения систем распознавания именованных сущностей. Описанный метод построен на внесении искусственного шума в данные и позволяет производить проверку систем распознавания именованных сущностей независимо от внутреннего устройства проверяемых систем. Рассмотрены различные варианты системы biLSTM-CRF для трех языков: русского, английского и французского. Для русского и английского языка рассмотрены системы, которые ранее были описаны в литературе. Для французского языка система применена Малых В.А. впервые. Полученные результаты позволяют заключить, что один из вариантов описываемой системы, а именно EmbedMatrix, оказывается наиболее устойчивым к шуму во всех проведенных экспериментах на большинстве уровней шума. Результаты сравнения систем приведены на рисунках 4.2-4.10;
4. Описан метод сравнения систем извлечения аспектов, то есть неявно заданных в тексте характеристик некоторого объекта. Метод построен таким образом, чтобы не зависеть от внутреннего устройства проверяемой системы. Он использует контролируемое внесение шума в данные. Рассмотрена система извлечения аспектов ABAE и представлены ее расширения. Также произведена проверка базовой системы LDA. В пятой главе также описан метод построения устойчивых к шуму систем извлечения аспектов, и представленное расширение RoVe, построенное по данному методу, показало лучшее качество для всех уровней шума. Результаты представлены на рисунке 5.2;
5. Выведены теоретические оценки алгоритмической (по числу операций) сложности способов решения задач во всех рассмотренных программных системах и выявлена тенденция более высокого качества для более сложных систем.

Предложенные в диссертации методы сравнения программных систем по их устойчивости к шуму и построения устойчивых к шуму систем соответствуют **паспорту специальности** 05.13.11 "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей", а именно пункту 10 "Оценка качества, стандартизация и сопровождение программных систем", а также отвечают формулировке научного и народнохозяйственного значения данной специальности, в которой отмечено повышение эффективности передачи данных и знаний в вычислительных машинах, комплексах и компьютерных сетях.

Достоверность и обоснованность полученных научных **результатов, выводов** и предлагаемых **рекомендаций** обеспечивается корректностью и последовательностью использования математического аппарата, логической непротиворечивостью рассуждений. Предложенные методы были реализованы в комплексе программ с последующим проведением экспериментов, результаты которых не противоречат закономерностям, известным специалистам в соответствующих областях знаний. Дополнительным обоснованием полученных научных результатов служат выводы, полученные на основе вычислительных экспериментов, а именно то, что существующие распространенные системы анализа текстов не являются устойчивыми к шуму в данных, а предложенные Малых В.А. методы создания устойчивых систем, реализованные в соответствующих задаче системах, позволяют создать более устойчивые к шуму программные системы.

Теоретическая значимость данной работы заключается в представленных методах сравнения систем по их устойчивости к шуму, которые используют методы теории вероятностей; а также методах построения устойчивых к шуму систем, использующих методы теории алгоритмов.

Практическая значимость рассматриваемой работы заключается в разработанном автором комплексе программ, решающем задачи сравнения и содержащем устойчивые к шуму системы для упомянутых задач.

В целом диссертация выполнена на высоком научном уровне, однако оказалась несвободна от **некоторых недостатков**:

1. Во второй главе при рассмотрении различных вариантов системы RoVe, построенной по предложенному автором методу, не указаны используемые параметры для обучения различных вариантов системы;

2. В разделе 2.9 при анализе результатов рассматриваются не все базовые системы, а только две из них, а именно word2vec и fasttext+speller. Мотивировка этого выбора не приведена. Можно предположить, что были выбраны системы, демонстрирующие наибольшую устойчивость, как следует из представленных выше по тексту результатов;

3. В главе 5 рассмотрена задача извлечения аспектов при помощи системы, решающей задачу тематического моделирования. Другие методы извлечения аспектов, например использование ключевых слов, не рассмотрены. Эти методы опираются на разметку людьми ключевых слов и требуют дополнительных размеченных данных. В силу этого они не могут быть применены для решения задачи извлечения аспектов на рассматриваемом корпусе, о чем следовало упомянуть в тексте диссертации;

4. В пятой главе приведенные результаты для базовой системы LDA находятся на отдельном графике, что не позволяет сравнивать их с результатами системы ABAE и предлагаемых автором расширений этой системы;

5. Стр. 68 диссертации: система fasttext в одной строке обозначена двумя разными способами, отличающимися от принятого обозначения в тексте диссертации;

6. Стр. 80 диссертации: в описании корпуса SentiRuEval указаны неверные размеры тестовой и обучающей выборок.

Указанные недостатки не носят принципиального характера и не влекут сомнений в отношении корректности полученных в диссертационной работе результатов.

Диссертационная работа Малых Валентина Андреевича является законченной научно-квалификационной работой, выполненной самостоятельно на актуальную тему и на высоком научном уровне, содержит решение важной задачи обработки шума в данных для последующего анализа текстов.

Исследование обладает научной и практической значимостью, которая обеспечивается комплексом полученных автором результатов. Указанные результаты отражены в публикациях, их достоверность подтверждается апробацией на семинарах и конференциях различного уровня.

По теме диссертации опубликовано девять научных работ: одна научная статья опубликована в рецензируемом журнале, входящем в перечень, сформированный по правилам, утвержденным Минобрнауки, пять работ опубликовано в изданиях, индексируемых международной базой цитирования Scopus.

Автореферат с достаточной полнотой отражает содержание и основные положения диссертации.

Принимая во внимание актуальность темы диссертации, научную новизну, практическую значимость результатов, считаю, что представленная диссертационная работа полностью соответствует всем критериям, установленным в действующей редакции «Положения о присуждении ученых степеней» для диссертаций на соискание ученой степени кандидата наук, а соискатель, Малых Валентин Андреевич, заслуживает присуждения ему искомой ученой степени кандидата технических наук по специальности 05.13.11 — "Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей".

Официальный оппонент —
доктор физико-математических наук, доцент, заведующий лабораторией теоретических и междисциплинарных проблем информатики Федерального государственного бюджетного учреждения науки Санкт-Петербургского института информатики и автоматизации Российской академии наук

А.Л. Тулупьев

09 апреля 2019 г.