"УТВЕРЖДАЮ"

Проректор

МГУ имени М.В.Ломоносова,

Федянин А.А.

«10» апреля 2019 г.

# ОТЗЫВ

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

**Актуальность темы**

Одной из существенных проблем автоматической обработки текстов является то, что в анализируемых текстах содержатся ошибки (например, опечатки), в результате чего применяемые алгоритмы дают результаты со значительно меньшим качеством. Это относится и к последним подходам к обработке текстов, а именно методам на основе нейронных сетей и представлением слов в виде векторов. Поэтому развитие подходов, которые позволили бы нейронным сетям быть более устойчивым к опечаткам анализируемого текста являются актуальными.

В диссертационной работе Малых В.А. предлагается подход по повышению устойчивости векторных представлений слов в условиях появления значительно объема опечаток (до 30% слов), что характерно для результатов оптического распознавания, некоторых жанров текстов в социальных медиа. Показано, что решаемые задачи классификации текстов, определения парафраз, извлечения

аспектов на основе предложенных представлений сохраняют более высокие оценки качества при внесении ошибок в текст.

Предложенный подход RoVe («устойчивых векторов слов») тестируется для нескольких задач (классификации текстов, извлечения именованных, сущностей, извлечения аспектов, определения парафраз, текстового следования) и трех языков (русский, английский, турецкий), реализован для различных архитектур нейронных сетей. Для сравнения используются известные векторные представления на основе подходов word2vec и fastText.

**Научная новизна и значимость** диссертационной работы

Следующие результаты диссертационной работы являются новыми:

1. Разработаны новые методы сравнения качества программных систем относительно их устойчивости к шуму для задач векторных представлений слов, классификации текстов, распознавания именованных сущностей и извлечения аспектов.

2. Разработаны новые методы построения программных систем на основе устойчивых к шуму векторных представлений слов, решающих задачи классификации текстов и извлечения аспектов в задаче оценки тональности. Разработанные методы применены в описанных задачах и показали лучшие результаты в большинстве проведенных экспериментов.

3. Создан, апробирован и внедрен программный комплекс, реализующий разработанные методы.

**Значимость для науки и практики полученных автором результатов**

**Теоретическая значимость** исследования заключается в разработке специализированной архитектуры нейронной сети для порождения векторных представлений слов, адаптированных к условиям применения к шумным данным в задачах автоматической обработки текстов.

**Практическая значимость работы** заключается в разработанных программных комплексах, реализующих: сравнение качества программных систем

по устойчивости к шуму; построение устойчивых к шуму векторных представлений слов; построение устойчивых к шуму методов классификации текстов, распознавания именованных сущностей, извлечения аспектов.

**Достоверность полученных результатов.** Обоснованность научных и практических результатов диссертации определяется проведенным автором подробным анализом подходов в разрабатываемой теме. Все полученные результаты подтверждаются экспериментами, проведенными в соответствии с общепринятыми стандартами.

**Конкретные рекомендации по использованию результатов и выводов диссертации**

Подход, предложенный в диссертации, может быть использован при построении векторных представлений слов в разнообразных задачах автоматической обработки текстов, включая задачи определения семантической близости предложения, автоматической классификации текстов, извлечения именованных сущностей и многие другие, которые решаются на основе зашумленных текстов, содержащих опечатки.

**Оценка содержания диссертации, ее завершенность.**

Диссертация написана ясным и четким языком и состоит из введения, шести глав и заключения. Библиография включает 112 наименований.

В **первой главе** рассматриваются основные подходы к сравнению систем на предмет устойчивости к шуму для различных задач автоматической обработки текстов.

**Вторая глава** представляет предложенный метод построения систем устойчивых к шуму под названием RoVe, генерирующей устойчивые к шуму вектора слов. Данная система сравнивается с известными системами word2vec и fastText, демонстрируя их меньшую устойчивость к шуму в различных прикладных задачах на нескольких языках.

В **третьей главе** рассмотрены результаты работы систем относительно устойчивости к шуму в задаче классификации текстов на примере задачи анализа тональности (sentiment analysis).

**Четвертая глава** описывает сравнение систем по устойчивости к шуму в задаче распознавания именованных сущностей в зашумленных данных.

В **пятой главе** рассматривается применение методов построения систем устойчивых к шуму для задачи выделения аспектов для анализа тональности в зашумленных текстах. Исследуется модель извлечения аспектов ABAE.

В **шестой главе** приводятся теоретические оценки алгоритмической сложности для исследуемых моделей. Сделаны выводы относительно зависимости сложности от качества.

По представленной диссертации имеются следующие **замечания:**

1) в работе не указываются гиперпараметры, используемых в работе программных архитектур, важных при практическом применении, включая предложенную архитектуру RoVe: размеры контекстного окна в конкретных экспериментах, размерность порождаемого вектора, размер ячейки в нейронных сетях LSTM, и др. Это может затруднить применение предложенного подхода другими исследователями. Также не указывается, изменялись ли эти гиперпараметры при применении предложенного подхода в разных задачах;

2) архитектуру «кодировщика» (термин, используемый в работе для англ. «encoder») необходимо было пояснить в общем обзоре, в котором рассматривались различные архитектуры нейронных сетей. Иначе в данном тексте получается, что термин "кодировщик" появляется первый раз уже при изложении предложенного подхода;

3) в разделе **объем и структура работы** в диссертации указывается, что в диссертации четыре главы, хотя на самом деле диссертация содержит шесть глав,

4) замечено несколько опечаток: стр. 55 «последнюю системы», стр.57 «слови», пропущены некоторые запятые (стр. 65) и др.

Указанные недостатки не являются принципиальными и не умаляют достоинств диссертации.

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

Результаты диссертационной работы опубликованы в 11 печатных работах, указанных в автореферате, из которых 7 издано в ведущих рецензируемых изданиях, входящих в перечень ВАК, 6 опубликовано в изданиях, индексируемых в системе Scopus. В списке литературы диссертации указано 7 работ автора.

Материал диссертации изложен последовательно и логично. Структурные составляющие диссертационной работы (введение, главы, заключение, библиографический список, приложения) позволяют получить полное представление о проделанных исследованиях и полученных результатах.

Автореферат соответствует диссертации, отражает её содержание и дает представление об актуальности темы, целях, задачах, объекте и методах исследования, научной новизне, практической ценности, реализации, апробации, объеме, кратком содержании и результатах работы.

Исходя из вышеизложенного, можно утверждать, что диссертация Малых В.А. на соискание ученой степени кандидата технических наук является законченной научно-квалификационной работой, в которой содержится описание нового подхода к созданию векторных представлений слов, устойчивых к шумным данным в различных задачах автоматической обработки текстов.

Принимая во внимание актуальность темы диссертации, научную новизну и практическую значимость ее результатов, считаем, что диссертационная работа «Методы сравнения и построения устойчивых к шуму программных систем в задачах обработки текстов» удовлетворяет требованиям Положения о порядке присуждения ученых степеней (постановление Правительства Российской Федерации от 24 сентября 2013 г. № 842 «О порядке присуждения ученых степеней»), предъявляемым к кандидатским диссертациям, а ее автор, Малых Валентин Андреевич, безусловно, заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и

программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Отзыв был обсужден и одобрен на семинаре лаборатории анализа информационных ресурсов Научно-исследовательского вычислительного центра Московского государственного университета имени М.В. Ломоносова 19 апреля 2019 года, протокол № 2/2019.

Заведующий лабораторией
анализа информационных ресурсов
Научно-исследовательского
вычислительного центра
Московского государственного
университета имени М.В. Ломоносова,
кандидат физико-математических наук Добров Б.В.

«24» апреля 2019г.

Подпись Б.В. Доброва заверяю

Директор
Научно-исследовательского
вычислительного центра
Московского государственного
университета имени М.В. Ломоносова,
профессор,
доктор физико-математических наук Тихонравов А.В.

«25» апреля 2019 г.

Адрес ведущей организации:
119991, г. Москва, ул. Ленинские горы, д. 1
Тел.: (495) 939-10-00
http://www.msu.ru,
E-mail: info@rector.msu.ru