ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

доцента Автономной некоммерческой организации
высшего образования "Университет Иннополис",
кандидата физико-математических наук Иванова Владимира
Владимировича

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"

Актуальность: в настоящее время в мире наблюдается так называемый информационный взрыв, когда скорость накопления знаний все время увеличивается. Знания, а в контексте данной диссертационной работы, тексты, появляются все быстрее. Но само качество этих знаний (текстов) подчас может быть неудовлетворительным. Качество текстов может быть оценено либо содержательно, либо по внешним признакам. Данная работа сосредоточена на исследовании методов и систем автоматического анализа текстов, содержащих шум в виде опечаток в словах. Тексты с опечатками представляют собой сложную и все более **актуальную** задачу для прикладных систем обработки текстов. Традиционно, проблема опечаток в обрабатываемых текстах решается с помощью систем проверки орфографии. Но в диссертационной работе Малых В.А. показано, что качество этих систем может быть недостаточным. В связи с этим автором диссертационной работы был предложен альтернативный подход, а именно создание систем изначально устойчивых к шуму.

В ходе решения актуальной задачи автором получены следующие **новые научные результаты**:

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

Первый научный результат 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 научный результат – предложенные методы построения устойчивых к опечаткам систем анализа текстов. Данные методы используют специально сконструированное представление текстов на основе входящих в них символов. В работе продемонстрировано, что представленные методы являются эффективными для различных языков, основывающихся на различных алфавитах. Данные методы не являются универсальными, т.к. есть системы письменности, не использующие алфавиты, но данная задача не ставилась автором, так что это не стоит считать недостатком работы.

Описанные научные результаты распределены по главам диссертационной работы следующим образом:

Во *второй главе* описан метод сравнения систем векторных представлений слов. Данный метод позволяет сравнивать произвольные системы

векторного представления слов, независимо от их внутреннего устройства. Он основан на контролируемом внесении шума во входные данные. Также во второй главе представлен метод построения систем, устойчивых к шуму, и описана система RoVe. Эта система в сравнении с базовыми системами показала лучшие результаты для большинства наборов данных. Результаты проверки приведены в таблицах 2-6. Дополнительно система была проверена на устойчивость к различным типам шума, результаты этой проверки отражены на рисунках 2.6-2.8.

В *третьей главе* представлен метод сравнения систем классификации текстов на примере задачи анализа тональности. Описанный метод позволяет сравнивать системы независимо от их внутреннего устройства. В третьей главе рассмотрены современные системы классификации текстов, в частности CharCNN, CharCNN-WordRNN. В этой главе также описан метод построения систем, устойчивых к шуму в задаче анализа тональности и представлена система RoVe, построенная по этому методу. Данная система продемонстрировала лучший результат для большинства проведенных экспериментов. Результаты проверки систем в описанной задаче анализа тональности приведены на графиках 3.3 - 3.6.

В *четвертой главе* описывается метод сравнения систем распознавания именованных сущностей. Метод построен на внесении искусственного шума в данные и может применяться к системам распознавания именованных сущностей независимо от их внутреннего устройства. В этой главе рассмотрена система biLSTM-CRF для трех языков: русского, английского и французского. Для русского и английского языка рассмотрены системы, которые уже были описаны в литературе. Для французского языка система применена Малых В. А.

впервые. В четвертой главе рассмотрены варианты системы biLSTM-CRF, в частности вариант EmbedMatrix-CNN, который показал лучшую устойчивость к шуму во всех проведенных экспериментах на большинстве уровней шума. Результаты сравнения систем приведены на рисунках 4.2-4.10.

В *пятой главе* представлен метод сравнения систем извлечения аспектов. Данный метод построен таким образом, чтобы не зависеть от внутреннего устройства проверяемой системы, используя контролируемое внесение шума в данные. В этой главе рассмотрены системы извлечения аспектов ABAE и LDA. Для системы ABAE представлены устойчивые к шуму расширения. В пятой главе также описан метод построения устойчивых к шуму систем извлечения аспектов. По этому методу сконструировано одно из расширений системы ABAE. Это расширение показало лучшее качество для всех уровней шума. Результаты проверки всех расширений представлены на рисунке 5.2, для системы LDA на рисунке 5.3.В *шестой главе* выведена алгоритмическая сложность использованных систем и указано на существование зависимости демонстрируемого результата от сложности системы, а именно положительная корреляция.

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".

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.

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.

Следует отметить, что в процессе рассмотрения работы были выявлены некоторые недостатки; в целом диссертация выполнена на высоком научном уровне, однако к данной работе необходимо высказать следующие **замечания**:

1) в работе используется термин "кодировщик", который описывается в каждой главе для отдельных систем; следовало изложить описание архитектуры кодировщика в главе 1 в соответствующем разделе, чтобы избежать повторного введения термина;

2) в главах 2-5 при использовании терминов "система Word2Vec" и "система FastText" не уточняется, какие именно, собственные или какие-то публично доступные, реализации программных систем используются;

3) графики в главах 2, 3 и 4 по стилю не соответствуют друг другу;

4) в разделе объем и структура работы указано, что в работе четыре главы, тогда как представленная работа имеет шесть глав;

5) также в тексте работы допущено несколько опечаток, например, на стр. 55 "последнюю системы", стр. 57 "слови".

Однако, эти замечания не носят принципиального характера и не влияют на общую положительную оценку работы.

Заключение: диссертационная работа Малых Валентина Андреевича является законченным исследованием, выполненным самостоятельно на актуальную тему и на высоком научном уровне. Автореферат диссертации правильно и полно отражает ее содержание и основные положения. Исследование обладает практической и научной значимостью, что подтверждается полученными результатами. Результаты достаточно полно отражены в опубликованных статьях, **достоверность** результатов подтверждается их апробацией на конференциях и семинарах различного уровня. По теме диссертации опубликовано одиннадцать научных работ, в том числе, шесть работ опубликовано в изданиях индексируемых международной базой цитирования Scopus, две из которых опубликованы в изданиях, рекомендованных ВАК РФ. Еще одна работа опубликована в издании из списка рекомендованных изданий ВАК РФ, таким образом всего опубликовано три таких работы.

Принимая во внимание актуальность темы диссертации, научную новизну и практическую значимость результатов, считаю, что представленная диссертационная работа полностью **отвечает всем критериям**, предъявляемым диссертациям на соискание ученой степени

кандидата технических наук ВАК РФ, а ее автор, Малых Валентин Андреевич, заслуживает присуждения ему ученой степени по специальности 05.13.11 “Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей”.

доцент Автономной некоммерческой организации
высшего образования “Университет Иннополис”
кандидат физико-математических наук Иванов В.В.

“29” апреля 2019 г.