

ОТЗЫВ

председателя диссертационного совета Тулупьева Александра Львовича на диссертацию Али Ноаман Мухаммад Абоалязид Мухаммад на тему «Аналитика больших текстовых данных», представленную на соискание ученой степени кандидата технических наук по специальности 2.3.8. Информатика и информационные процессы

Представленное на рассмотрение диссертационное исследование относится к пересечению двух междисциплинарных сфер: больших данных и обработки естественного языка (ОЕЯ). Соискатель выделяет одну из соответствующих масштабных проблематик — обработка больших массивов текстов — и внутри указанной проблематики ставит общую научно-техническую задачу собственного исследования, а именно формирование «модели предварительной обработки текстовых данных с использованием комбинации методов ОЕЯ», а также эмпирическое раскрытие ее свойств на основе сформированных наборов данных (или, следуя формулировкам диссертанта, «экспериментального поля для разработанных алгоритмов»), причем само исследование в этой части опирается на известные методы и алгоритмы, а продвижение осуществлено, в первую очередь, за счет экстенсивного компаративного анализа.

Составными частями комплексного результата, достигнутого в диссертационном исследовании, являются

1. Формирование упомянутого выше «экспериментального поля»;
2. Новая модель для извлечения значимых аспектных терминов из текста с использованием методов машинного и глубокого обучения;
3. Алгоритма гендерной классификации на основе динамически обрезанной модели n-грамм и предложенного конвейера для обработки данных;
4. Рекомендательная модель на основе предложенного алгоритма извлечения аспектных терминов и анализ тональности;
5. Формирование экспериментального прототипа комплекса программ на основе предложенных моделей с целью эмпирической проверки их эффективности.

Все составные части комплексного результата представляются важными и существенными для предметной области, а также четко соответствующими заявленной специальности, однако выделил бы особо две из них.

Определение пола автора текста является критически значимым в анализе тональности, поскольку «языковое поведение» женщин и мужчин существенно различаются.

Формирование *работоспособного* прототипа комплекса программ, воплотившего частные результаты теоретической и первичной эмпирической составляющих исследования, является ярким свидетельством достоверности и практической применимости научных достижений диссертанта.

В целом, диссертация соискателя отличается от классического образа диссертаций в сфере математики и информатики, сложившегося за предшествующие десятилетия в СПбГУ: в ней не представлен обширный компонент со строгими математическими утверждениями и их доказательствами, эмпирические подходы в целом, на мой взгляд, преобладают над классическими теоретическими. В связи с этим особо отметил бы, что такая особенность диссертации не может рассматриваться как ее недостаток, она обусловлена все еще складывающимися научными подходами к решению исследовательских задач в сферах больших данных и анализа естественного языка. Кроме того, предмет, объект, совокупность методов исследования, особенности представления выводов (положений) диссертации вполне сочетаются с заявленной научной специальностью и в значительной степени обусловлены ее особенностями и ее отличиями от других достаточно давно институционализированных научных специальностей.

Отмечу небольшое число несообразностей, которые, на мой взгляд, содержатся в диссертации и обращают на себя внимание либо вызывают вопросы.

1. Диссертант объявил основной целью исследования «*изучение* теоретических, методологических и практических вопросов аналитики больших данных, направленных на разработку вычислительных алгоритмов и реализацию соответствующего программного обеспечения, работающего в режиме реального времени, для обеспечения эффективной обработки текстового контента онлайн с использованием структуры больших данных.» Такая формулировка цели («изучение») вызывает сомнение, поскольку, скорее, соотносится либо с планом освоения некоторой образовательной программы, либо с одним из подготовительных этапов некоторого обширного исследования, которое при такой формулировке цели останется «за скобками», иными словами, за пределами объявленного фокуса диссертации. Более того, при оценке результатов диссертации встанет вопрос «хорошо или плохо диссертант изучил» области, указанные в

формулировке цели, что тоже не может стать ключевым моментом процедуры определения достоин или не достоин соискатель степени кандидата наук. Возможно, под истинной целью исследования соискатель имел в виду «обеспечение эффективной обработки текстового контента онлайн с использованием структуры больших данных» с соответствующими акцентами и уточнениями.

2. При анализе выводов всякого исследования, использовавшего или опиравшегося на эмпирические подходы, всегда встает вопрос о генерализуемости таких выводов и, одновременно, о границах их применимости, о границах применимости сформированных комбинаций методов, алгоритмов, разработанного прототипа комплекса программ. Насколько было можно понять из текста диссертации, реальным объектом эмпирической части исследования и источником данных для синтеза «экспериментального поля для разработанных алгоритмов» были веб-сайты, «содержащие отзывы покупателей об указанных продуктах» (стр. 212 объединенного текста диссертации); причем к «указанным продуктам» отнесены персональные компьютеры и ноутбуки. Иными словами, речь идет о весьма узком спектре товаров, имеющем свою специфику и достаточно четко выраженные характеристики/параметры. Хотелось бы получить от соискателя разъяснения по степени указанных генерализуемости и границах применимости комплексного результата диссертации (совокупности результатов диссертации).
3. Стержневым результатом диссертации является модель предварительной обработки данных (раздел 2.4. и иные части текста диссертации). Хотелось бы получить обозримые и более формальные пояснения, что именно диссертант рассматривает как модель (какой формальный объект или какую совокупность формальных объектов), что формально является «входом» для данной модели, что является «выходом» и в силу каких особенностей «выход» оказывается «лучше», чем «вход».

Отмеченные несообразности не оказывают существенного влияния на оценку достоверности представленных в диссертации выводов и, полагаю, обусловлены тем, что большой объем материала соискателю пришлось вписать в достаточно ограниченные рамки.

Диссертация Али Ноаман Мухаммад Абоалязид Мухаммад на тему: «Аналитика больших текстовых данных» соответствует основным требованиям, установленным Приказом от 19.11.2021 № 11181/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Али Ноаман Мухаммад Абоалязид Мухаммад

заслуживает присуждение ученой степени кандидата технических наук по специальности 2.3.8. Информатика и информационные процессы. Нарушения пунктов 9 и 11 указанного Порядка в диссертации не обнаружены.

Председатель диссертационного совета

Д-р физ.-мат. наук, профессор,
профессор кафедры информатики СПбГУ



А.Л. Тулупьев

28 июля 2022 г.