

ОТЗЫВ

члена диссертационного совета Демьяновича Юрия Казимировича о диссертации 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.

Актуальность темы диссертационного исследования. В настоящее время одной из важнейших тенденций в социальных сетях и на электронных рынках стал анализ тональности поступающих сообщений. При этом наиболее трудоемким вычислительным этапом является преобразование и подготовка необработанных текстовых данных в форму, пригодную для применения упомянутого анализа. Таким образом, актуальность темы диссертации сомнений не вызывает.

Характеристика работы и ее результатов. В представленной к защите работе предлагается модель предварительной обработки текстовых данных с использованием комбинации методов обработки естественных языков (ОЕЯ).

Основной целью данной диссертации является изучение теоретических, методологических и практических вопросов аналитики больших данных, направленных на разработку вычислительных алгоритмов и реализацию соответствующего программного обеспечения, работающего в режиме реального времени.

В данной работе предложена модель Извлечение Аспектных Терминов. Модель основана на кластеризации векторов слов, сгенерированных с помощью предварительно обученной модели BERT. Для улучшения качества кластеров слов, полученных с помощью алгоритма кластеризации K-Means++, была использована техника снижения размерности "SOM". В данной работе предлагается динамическая обрезанная модель n-грамм для распознавания пола клиентов по их именам пользователей. Она использует доступность данных об отзывах на веб-сайтах и извлекает набор данных об именах пользователей. Для предварительной обработки естественного текста предложена ансамблевая модель, которая улучшает качество получаемого текста и результаты анализа. Предложена новая методика --- «Извлечение аспектных терминов», предназначенная для обработки естественного текста. Эта методика основывается на нейронных сетях и методах глубокого обучения. Разработан метод гендерной классификации и предложена динамическая модель n-грамм для извлечения признаков. Кроме того,

разработана модель рекомендаций на основе анализа тональности относительно аспектов.

Достоверность положений работы. Степень достоверности результатов определяют эксперименты с предложенным конвейером для предварительной обработки естественного текста. Они показывают улучшение качества обработанного текста, что впоследствии положительно сказывается на применяемой методике анализа. Алгоритм Извлечение Аспектных Терминов обеспечивает улучшение извлечения признаков из свободного текста. Алгоритм для определения пола дает многообещающие результаты, учитывая небольшой объем информации, необходимой для классификации. Модель для веб-рекомендаций показала приемлемую производительность при извлечении предпочтений пользователя. и составлении рекомендаций. Это дополняет решение проблемы холодного старта. Эксперимент с предложенными моделями подтверждает возможность их реализации и демонстрирует их эффективность. Реализация и эффективность предложенных моделей подтверждается проведенными экспериментами. Основные результаты работы были представлены на 6 российских и международных конференциях и в пяти научных статьях.

Методология и методы исследования. В диссертационной работе используется общая методология наук о данных и информации, основанная на сборе и предварительной обработке данных, представлении текста, на моделировании, анализе и обобщении теоретического и практического материала работы.

Научная новизна работы состоит в следующем.

1. Предложена ансамблевая модель для предварительной обработки естественного текста, которая улучшает качество получаемого текста и результаты анализа.
2. Разработана новая методика --- Извлечение Аспектных Терминов из естественного текста. Методика использует теорию нейронных сетей и методы глубокого обучения.
3. Разработана методика гендерной классификации на основе словаря. Предложена динамическая обрезанная модель n-грамм для извлечения признаков.
4. Разработана модель рекомендаций на основе анализа тональности относительно аспектов.

Теоретическая ценность данной работы для будущих исследований заключается в предложенном анализе проблем аналитики больших данных и разработке методов предварительной обработки естественного языка, гендерной классификации.

Практическая ценность работы заключается в предложенных новых алгоритмах.

Недостатки работы.

1. Не разработаны вопросы масштабируемости предложенных алгоритмов.
2. Степень проработки предлагаемых решений недостаточна с точки зрения точности, эффективности и сложности.
3. Имеются отдельные стилистические погрешности.

Перечисленные недостатки не снижают теоретической и практической ценности работы..

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

Член диссертационного совета

доктор физико-математических наук
профессор, заведующий кафедрой параллельных алгоритмов, профессор Санкт-Петербургского государственного университета

Демьянович Ю.К.

29.06.2022



Демьянович Ю.К.
29.06.2022