Отзыв
научного руководителя
на диссертацию Али Ноамана Мухаммада Абоалязида Мухаммада
на тему
«Аналитика больших текстовых данных»,
представленную на соискание ученой степени кандидата технических наук по специальности
2.3.8 Информатика и информационные процессы.

Обработка больших данных – направление, к которому относится работа Али Ноамана Мухаммада Абоалязида Мухаммада, является одним из доминирующих направлений в области компьютерных наук и технологий в последние полтора десятилетия. Не существует точного определения понятия «большие данные». Обычно говорят о трех V-свойствах больших данных: Volume, Velocity, Variety (объем, скорость, многообразие). В контексте диссертации наиболее важным из них является многообразие, понимаемое как необходимость обработки слабоструктурированных и неструктурированных данных.

Необходимость рассмотрения неструктурированных данных, представленных обычно как тексты на естественном (или близком к естественному) языке, связана с тем, что значительная часть данных, доступных в глобальной сети, представлена именно в неструктурированном виде. Необходимость автоматической обработки текстов на естественных языках привела к появлению многочисленных методов компьютерной лингвистики, в том числе средств для извлечения информации в более структурированном виде с помощью методов машинного обучения.

Несмотря на большое разнообразие известных методов обработки текстов, их применение требует исследований в каждом конкретном случае. Это вызвано тем, что эти методы основаны на статистических моделях, и, следовательно, не могут быть точными. При этом качество результатов зависит от особенностей данных.

В диссертации рассматривается и решается ряд задач обработки текстов и предложена обобщенная модель, в рамках которой возможно применение различных известных методов для выполнения основных операций подготовки неструктурированных данных для последующего получения аналитических результатов. Хорошо известно, что подавляющая часть ресурсов, необходимых для анализа больших данных, используется именно для предварительной подготовки данных, что определяет практическую значимость этого результата диссертации. Такая предварительная обработка включает очистку и улучшение качества обрабатываемых данных, а также извлечение признаков (характеристик), необходимых для последующего анализа.

Далее в диссертации демонстрируется применение предложенной обобщенной схемы для решения ряда конкретных задач анализа. В частности, решаются задачи выделения признаков, определяющих тональность текста, необходимые для анализа комментариев пользователей, Рассматриваются также задачи выделения характеристик, необходимых для предсказания наиболее вероятного перехода при навигации в глобальной сети интернет.

Одним из наиболее интересных результатов работы является методы выделения признаков из идентификаторов учетных записей пользователей приложений, работающих в сети интернет. Такие идентификаторы, вообще говоря, не являются текстами на естественном языке, так как могут строиться как комбинация различных частей полного имени, а также включать сокращения и числовые вставки. Кроме этого, особая сложность задачи вызвана малым размером идентификаторов. В диссертации предложен метод предварительной обработки с учетом особенностей этого типа данных, после применения которого удается получать полезные признаки с довольно высокой точностью.

Предложенные в этой части диссертации методы применяются для извлечения признака пола пользователя. Выбор этого признака, по-видимому, связан тем, что такая задача всегда вызывает оживленное обсуждение работы в любой аудитории. Конечно, что предложенные в диссертации методы применимы и для предсказания других признаков.

Основательность работы, проделанной диссертантом, подтверждается и списком цитируемой литераторы, содержащим около 200 наименований. Подготовленный на основе анализа этой литературы обзор имеет самостоятельную ценность и был опубликован в виде журнальной статьи.

Необходимо отметить, что в течение всего периода обучения в аспирантуре и проведения исследований работа выполнялась на английском языке и текст диссертации также был подготовлен на английском языке. Русский перевод текста диссертации содержит большое количество стилистических погрешностей, которые, с учетом того, что русский язык не является родным для диссертанта, нельзя рассматривать как серьезный недостаток работы.

Али Ноаман Мухаммад Абоалязид Мухаммад успешно выполнил большой объем работы, требующей как теоретических изысканий, так и практических навыков. В ходе исследований он продемонстрировал умение самостоятельно ставить и на достаточно высоком уровне решать сложные практические задачи, проявил аналитические способности, позволяющие работать с теоретическим материалом и находить новые решения. Работая над диссертацией, он проявил себя активным и добросовестным исследователем, готовым к проведению самостоятельных исследований.

Считаю, что диссертация удовлетворяет всем требованиям, предъявляемым к таким работам, а ее автор, Али Ноаман Мухаммад Абоалязид Мухаммад, заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.8 — Информатика и информационные процессы.

Доктор физико-математических наук, доцент,
профессор кафедры информатики
Санкт-Петербургского государственного университета

Т.М. Косовская

16 февраля 2022 г.

Личную подпись Косовской Т.М.
заверяю
Заместитель начальника
Управления кадров О.С. Суворова