ОТЗЫВ

на автореферат диссертации "Модель, метод и алгоритмы Data Mining для интеллектуальной обработки и анализа текстов на естественном языке", представленной Мансур Али Махмуд на соискание ученой степени кандидата технических наук по специальности 1.2.1 Искусственный интеллект и машинное обучение

Представленное исследование затрагивает важную и востребованную проблему в области искусственного интеллекта – эффективную обработку и анализ текстовых данных в условиях их стремительного роста. Актуальность работы не вызывает сомнений, поскольку современные ИИ-системы требуют более совершенных методов векторизации для преобразования текстов в структурированные числовые форматы, пригодные для машинного обучения.

Хотя традиционные подходы (BoW) отличаются простотой и интерпретируемостью, они страдают от проблем высокой размерности и не учитывают семантические связи между словами. Нейросетевые методы создают плотные низкоразмерные представления, но либо теряют информацию при обработке документов, либо требуют значительных вычислительных ресурсов для длинных текстов. Существующие методы не позволяют одновременно достичь трех ключевых характеристик: низкой размерности, высокой семантической выразительности и интерпретируемости векторов. При этом высокая размерность представлений, хотя и повышает точность анализа, создает проблемы с вычислительной эффективностью и масштабируемостью систем. Интерпретируемость же векторов остается важным фактором, повышающим доверие пользователей и позволяющим выявлять ошибки в работе моделей. Таким образом, актуальной остается задача разработки методов векторизации, сочетающих преимущества существующих подходов без их недостатков.

Основной вклад работы заключается в предложенном концепт-ориентированном решении для представления текстовых данных, которое отличается от традиционных word-based и topic-based моделей. В частности:

- Разработана новая модель и модифицированный метод построения текстовых векторов, где каждый элемент вектора отражает семантический концепт, а не отдельное слово. Это позволяет контролировать размеры векторов и их интерпретацию. Модель отличается применением новых критериев для более точного определения весовых коэффициентов концептов на основе мер семантической близости.
- Разработан алгоритм автоматического построения концептов путем кластеризации схожих ключевых фраз в однородные группы. Это повышает выразительную силу векторов, построенных на основе данных концептов, и, как следствие, увеличивает эффективность алгоритмов классификации и кластеризации.
- Для извлечения ключевых фраз разработан алгоритм на основе парсера, который идентифицирует фразы с корректной лингвистической структурой и исключает избыточные слова.

Экспериментальные результаты демонстрируют снижение ошибок классификации и кластеризации по сравнению с традиционными методами (например, TF-IDF, word2vec), что подтверждает практическую полезность разработки.

Несмотря на несомненные достоинства работы, в автореферате остались не полностью раскрытыми следующие аспекты:

1. Не указано, как обрабатывались многозначные слова (полисемия) при построении концептов.

2. Недостаточный анализ временной сложности. Утверждается, что сложность метода равна O(n), но нет экспериментального подтверждения этого на реальных данных разного объема.

Несмотря на эти замечания, работа соответствует уровню кандидатской диссертации, обладает значимым теоретическим и прикладным вкладом в область искусственного интеллекта и машинного обучения. Диссертация соответствует требованиям Положения о присуждении учёных степеней, утверждённого Постановлением Правительства РФ от 24 сентября 2013 г. № 842 (с изменениями и дополнениями), установленным для кандидатских диссертаций, а её автор, Мансур Али Махмуд, заслуживает присуждения учёной степени кандидата технических наук по специальности 1.2.1. Искусственный интеллект и машинное обучение.

Профессор кафедры сетей связи и передачи данных
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Санкт-Петербургский государственный университет
телекоммуникаций им. проф. М.А. Бонч-Бруевича».
доктор технических наук, профессор

Мутханна Аммар Салех Али
15.05.2025

Докторская диссертация защищена по специальности 2.2.15 – «Системы, сети и устройства телекоммуникаций»

Почтовый адрес: 193232, Санкт-Петербург, пр. Большевиков, д. 22, корп. 1
Тел.: +7 952 210-44-86
E-mail: muthanna.asa@sut.ru
Даю согласие на обработку своих персональных данных.

Подпись руки Мутханна А.С.А.
Удостоверяю
([illegible])
15.05.2025