УТВЕРЖДАЮ:

Проректор по науке и инновациям

ФГБОУ ВО «ВГТУ»

доктор технических наук, доцент

Башкиров Алексей Викторович

«12» мая 2025 г.

Отзыв ведущей организации

федерального государственного бюджетного образовательного учреждения высшего образования «Воронежский государственный технический университет» (ВГТУ) о диссертационной работе Мансур Али Махмуд на тему: «Модель, метод и алгоритмы Data Mining для интеллектуальной обработки и анализа текстов на естественном языке», представленной на соискание ученой степени кандидата технических наук по специальности 1.2.1. Искусственный интеллект и машинное обучение.

1. **Актуальность темы исследования**

В настоящее время во всех областях развития информационных технологий наблюдается стремительный рост объёмов текстовых данных, генерируемых в цифровой среде. Это связано с широким внедрением искусственного интеллекта (ИИ) и необходимостью обработки неструктурированной информации на естественном языке. Однако увеличение объёмов данных и усложнение процессов их анализа приводят к дефициту вычислительных ресурсов, что затрудняет

эффективное извлечение знаний и выявление скрытых закономерностей.

Особую значимость приобретают методы Data Mining, направленные на классификацию и кластеризацию текстовых документов. Векторизация документов играет ключевую роль в задачах машинного обучения, поскольку позволяет представлять тексты в числовой форме, пригодной для математического анализа. Однако существующие подходы сталкиваются с проблемами высокой размерности признакового пространства и недостаточной интерпретируемости векторов, что негативно влияет на вычислительную эффективность и доверие пользователей к моделям ИИ.

Проведенное Мансур А.М. диссертационное исследование направленно на разработку модифицированных модели, метода и алгоритмов Data Mining для интеллектуальной обработки и анализ текстов на естественном языке. Предложенный метод векторизации текстовых документов не только снижает размерность векторных представлений при сохранении их интуитивной интерпретируемости, но и позволяет минимизировать частоту ошибок при классификации и кластеризации текстовых данных. Данный метод в сочетании с разработанными алгоритмами способствует более точному извлечению знаний из текстов и расширению возможностей систем искусственного интеллекта при работе с возрастающими объемами информации. Полученные результаты обладают значительным потенциалом для практического применения и вносят существенный вклад в развитие методов обработки естественного языка в условиях экспоненциального

роста объемов текстовых данных, что подтверждает **актуальность** и важность темы исследования.

2. Достоверность и научная новизна результатов работы

Достоверность и обоснованность полученных соискателем теоретических и практических результатов обеспечиваются комплексным методологическим подходом, включающим строгую формулировку исследовательских целей, теоретическое обоснование ключевых положений с опорой на современные достижения в области искусственного интеллекта и обработки естественного языка, а также глубокий анализ отечественных и зарубежных научных работ. Применение методов интеллектуального анализа данных, современного математического аппарата и принципов объектно-ориентированного программирования позволило достичь высокой степени соответствия между теоретическими выводами и экспериментальными результатами.

Практическая ценность исследования подтверждена его успешной апробацией через публикации в рецензируемых научных изданиях, выступления на профильных конференциях, а также внедрением разработок в деятельность компании ООО "ИТ-Эффект" (г. Москва). Дополнительным свидетельством практической значимости работы является акт о внедрении результатов исследования в учебный процесс Южного федерального университета. Важным показателем научно-технической ценности проведённой работы служат два полученных свидетельства о государственной регистрации программ для ЭВМ, разработанных в рамках диссертационного исследования.

Настоящее исследование вносит существенный вклад в область обработки текстовых данных за счет разработки инновационной модели,

метода и алгоритмов, обеспечивающих эффективную векторизацию и семантический анализ текстовых документов. Ключевая **научная новизна** работы заключается в создании комплексного решения, которое объединяет:

1. Математическую модель векторизации текстов на основе концептов, которая отличается применением новых правил построения эталонных концептов и оригинальных функций определения их весов, что позволяет одновременно снизить размерность векторного пространства и повысить дискриминационную способность результирующих векторов признаков;

2. Модифицированный метод генерации векторных представлений документов на основе данной модели, который вводит использование интерпретируемых признаков при векторизации, что существенно снижает частоту ошибок алгоритмов классификации и кластеризации текстовых документов.

Для обеспечения качественной семантической обработки текстов создан алгоритм построения концептов из семантически близких фраз, решающий задачу кластеризации с учетом контекстуальной семантической близости и тем самым повышающий однородность формируемых концептуальных кластеров. Дополняет его алгоритм извлечения и фильтрации ключевых фраз, который благодаря применению функции парсера для разметки частей речи обеспечивает выделение фраз с корректной грамматической структурой. Предложенный комплекс алгоритмов и модели обеспечивает эффективную обработку и семантический анализ текстов на естественном языке. Ключевыми преимуществами разработки являются: интерпретируемость векторных представлений текстов;

улучшенная дискриминационная способность признаков; а также снижение частоты ошибок при классификации и кластеризации документов. Эффективность предложенных решений подтверждена результатами экспериментальных исследований.

3. Наиболее существенные результаты исследований и ценность для практического использования полученных соискателем результатов

Научная и практическая значимость диссертационного исследования подтверждается следующими ключевыми результатами

1. Математическая модель векторизации текстов на основе применения новых правил построения эталонных концептов и новых функций определения их весов позволяет снизить размерность векторного пространства и улучшить дискриминационную способность результирующих векторов признаков;

2. Модифицированный метод генерации векторных представлений текстов на основе построенной модели векторизации позволяет снизить частоту ошибок алгоритмов классификации и кластеризации текстовых документов;

3. Алгоритм извлечения и фильтрации ключевых фраз на основе применения функций парсера для разметки частей речи позволяет извлекать ключевые фразы с правильной грамматической структурой;

4. Алгоритм построения концептов из семантически близких фраз позволяет повысить однородность кластеров, представляющих концепты.

Практическая ценность исследования подтверждается успешной реализацией разработанных модели, метода и алгоритмов обработки и

анализа текстов на естественном языке в виде программного обеспечения, нашедшего применение в коммерческой и образовательной сферах. Созданное программное приложение, интегрирующее авторские модели и алгоритмы обработки текстов, обеспечивает снижение погрешности алгоритмов классификации и кластеризации документов за счет оптимизированной векторизации с сохранением семантической интерпретируемости.

Ключевым достижением стало внедрение этих разработок в ООО "ИТ-Эффект" (Москва), где они позволили усовершенствовать рекомендательную систему на основе технологии look-a-like, повысив эффективность поиска целевой аудитории. Подтверждением практической значимости работы является наличие у соискателя 2-х свидетельств о государственной регистрации программ для ЭВМ, а также использование результатов диссертационного исследования в учебном процессе ФГБОУ ВО «Южный федеральный университет».

4. Соответствие требованиям по выполнению, оформлению и апробации диссертационной работы

Диссертационная работа объемом 150 страниц основного текста (157 страниц с приложениями) представляет собой законченное научное исследование, структурированное в соответствии с требованиями к кандидатским диссертациям. Работа состоит из введения, четырех содержательных разделов, заключения и двух приложений, дополненных 12 таблицами и 31 рисунком, которые наглядно иллюстрируют ключевые аспекты исследования. Список литературы включает 146 научных источников, что свидетельствует о глубине проработки темы.

Во введении обоснована актуальность темы диссертации и степень ее разработанности, сформулирована цель и задачи работы, а также методология проведенных исследований, описаны научная новизна, теоретическая и практическая значимость и основные научные положения, выносимые на защиту, представлены сведения о достоверности и апробации результатов, внедрении, публикациях автора, объеме и структуре работы.

В первой главе представлен аналитический обзор научных исследований в области обработки и анализа текстов на основе методов искусственного интеллекта и машинного обучения. Также рассмотрены существующие методы представления текста. Основное внимание уделено методам векторизации текстов. Проанализированы публикации, напрямую связанные с темой диссертации. Даны формализованные постановки основных задач исследования и сделаны выводы по разделу в целом.

Во второй главе представлена разработанная модель и метод семантической векторизации документов, основанные на технологиях Data Mining. Ключевой особенностью предложенного метода является построение словаря эталонных концептов с последующим сопоставлением терминов документа с этими концептами через оценку семантической близости. Разработанный метод позволяет получать низкоразмерные, но информативные векторные представления, сохраняющие смысловое содержание текстов любой длины. В отличие от существующих аналогов, данный метод обеспечивает целостность обработки документов без потери значимых фрагментов текста. В главе подробно описаны: поэтапная схема работы метода; его математическая модель; отличия и преимущества по сравнению с другими методами.

В третьей главе исследования представлены два ключевых алгоритма обработки текстовых данных. Первый алгоритм решает задачу формирования концептов, используя в качестве основы не отдельные слова (униграммы), а семантически значимые ключевые фразы (n-граммы). Такой подход позволяет преодолеть проблему лексической многозначности при построении словаря эталонных концептов. Второй алгоритм реализует процесс извлечения и отбора ключевых фраз с применением синтаксического анализа для идентификации их принадлежности к определенной части речи. Особое внимание уделено весовым функциям, обеспечивающим оценку релевантности выделяемых фраз. Важным преимуществом данного решения является возможность автоматического отсева фраз с некорректной грамматической структурой, что в конечном итоге способствует повышению качества формируемых концептов за счет снижения уровня шума и увеличения семантической однородности элементов словаря.

В четвертом разделе диссертационной работы представлены результаты практического применения разработанных решений. Проведена разработка программного приложения и серия вычислительных экспериментов для оценки эффективности предложенных модели, метода и алгоритмов. Экспериментальные результаты подтвердили преимущества разработанного метода векторизации текстов, демонстрирующего снижение частоты ошибок классификации и кластеризации документов при одновременном уменьшении размерности признакового пространства по сравнению с существующими аналогами. Приведенные автором результаты вычислительного эксперимента подтвердили непротиворечивость и

высокую эффективность разработанных модели, метода и алгоритмов обработки и анализа текстов на естественном языке

В заключении изложены итоги выполненного исследования, рекомендации, перспективы дальнейшей разработки темы.

В приложениях приведены свидетельства об официальной регистрации программ для ЭВМ и копии актов внедрения.

Научные положения и выводы диссертационного исследования получили широкое отражение в научных публикациях и профессиональном обсуждении. Автором опубликовано 17 работ по теме диссертации, включая 3 статьи в рецензируемых журналах перечня ВАК (из которых 1 одна без соавторов) и 2 публикации в международных изданиях, индексируемых в базах Scopus и Web of Science. Дополнительно опубликовано 12 статей в других изданиях. Практическая значимость работы подтверждена двумя зарегистрированными программами для ЭВМ, а также представлением основных результатов на 9 научных конференциях международного и всероссийского уровня, где они получили профессиональную оценку научного сообщества.

Материал диссертационной работы представлен в виде четко структурированного и логически связанного изложения. Текст работы написан ясным научным языком, соответствующим требованиям академического стиля. Все структурные элементы диссертации – от содержания до списка литературы аппарата – оформлены в строгом соответствии с актуальными нормативными требованиями Высшей аттестационной комиссии (ВАК) Российской Федерации. Автореферат диссертации представляет собой полноценное и систематизированное изложение основных положений исследования, в котором четко

отражены содержание диссертационного исследования, методологическая база работы, полученные научные результаты, значимые выводы и практические рекомендации.

Рекомендации по использованию результатов диссертационной работы. Разработанные в диссертации Мансур Али Махмуд метод и алгоритмы рекомендуется использовать в организациях научного и прикладного профиля, работающих с задачами обработки естественно-языковых текстов и анализа данных. Результаты исследования особенно актуальны для сферы искусственного интеллекта и машинного обучения, где они могут быть применены при создании моделей текстовой аналитики, выявлении семантических закономерностей и разработке новых алгоритмов обработки и анализа текстовых данных. Теоретические выводы и практические наработки работы соискателя открывают перспективы для совершенствования существующих подходов к автоматизированному анализу текстовых данных. К числу потенциальных пользователей результатов исследования относятся ведущие научные центры, включая институты Российской академии наук: Московский физико-технический институт (МФТИ); Институт проблем передачи информации им. А.А. Харкевича РАН (ИППИ РАН); Санкт-Петербургский федеральный исследовательский центр РАН; Кабардино-Балкарский научный центр РАН в г. Нальчике. Полученные результаты могут быть внедрены в высокотехнологичных отраслях, требующих интеллектуальной обработки текстов, и использованы в учебном процессе программ бакалавриата, магистратуры и аспирантуры по направлению "Информатика и вычислительная техника" для формирования продвинутой теоретической и практической базы в области обработки естественного языка и искусственного интеллекта.

5. Замечания и недостатки

Хотя работа в целом соответствует академическим стандартам и заслуживает высокой оценки, в ней прослеживаются отдельные аспекты, которые могли бы быть усовершенствованы:

1. В работе автор анализирует два подхода к извлечению концептов и останавливается на применении текстового подхода, однако не проводит сравнение с методами, основанными на знаниях, а это могло бы усилить обоснованность выбора и повысить объективность результатов;

2. Для кластеризации фраз используется алгоритм сферических k-средних с мерой косинусного сходства, однако в исследовании отсутствует анализ альтернативных метрик.

3. Не совсем ясно, что именно подразумевается под термином «эталонные концепты». Было бы полезно дать более точное определение или пояснение этого понятия, чтобы повысить ясность изложения.

4. Было бы более информативно проверить, как разработанный алгоритм конкурирует с алгоритмами тематического моделирования (например, LDA) в задачах извлечения семантических концептов, поскольку они используют схожий подход.

5. Для повышения воспроизводимости проводимых экспериментов было бы желательно разместить исходный код метода и отдельных алгоритмов в открытом доступе.

6. Хотя в работе приведен псевдокод алгоритма извлечения ключевых слов и концептов, добавление структурной схемы значительно улучшило бы его понимание.

7. В работе имеются стилистические неточности и опечатки, например на странице 64 отсутствует пробел между словом *Шаг* и его номером.

Отмеченные недостатки носят частный характер и не влияют на общую положительную оценку диссертационной работы Мансур Али Махмуд.

Выводы

С учетом изложенного, считаем, что диссертационное исследование Мансур Али Махмуд является самостоятельной, завершенной научно-квалификационной работой, в которой решена актуальная научная задача разработки модели, метода и алгоритмов Data Mining для интеллектуальной обработки и анализа текстов на естественном языке. Решение данной научной задачи имеет принципиальное значение для развития технологий искусственного интеллекта и машинного обучения, так как позволяет существенно снизить частоту ошибок при классификации и кластеризации текстовых данных в условиях их экспоненциального роста.

Диссертационная работа Мансур Али Махмуд на тему «Модель, метод и алгоритмы Data Mining для интеллектуальной обработки и анализа текстов на естественном языке», представленная на соискание ученой степени кандидата технических наук по специальности 1.2.1. Искусственный интеллект и машинное обучение, обладает научной новизной и сочетает в себе теоретическую ценность и практическую значимость. Научные результаты исследования опубликованы в ведущих научных изданиях: рецензируемых

журналах Перечня ВАК; а также международных изданиях, индексируемых в базах данных Scopus и Web of Science. Кроме того, основные положения работы были успешно представлены и получили положительную оценку на ряде авторитетных международных и всероссийских научных конференций.

Диссертационная работа в полной мере соответствует пунктам 4 и 5 паспорта научной специальности 1.2.1. Искусственный интеллект и машинное обучение (технические науки).

Хотя в работе имеются отдельные замечания, диссертация Мансур А.М. по всем ключевым параметрам: актуальности; научной новизне; объему исследований и практической значимости – полностью отвечает требованиям пп. 9-14 Положения о присуждении ученых степеней, утверждённого Постановлением Правительства РФ от 24 сентября 2013 г. № 842 "О порядке присуждения ученых степеней" (с изменениями и дополнениями) в редакции от 16 октября 2024 г., предъявляемым ВАК РФ к диссертациям на соискание ученой степени кандидата технических наук.

Автор работы, Мансур Али Махмуд, достоин присуждения ему ученой степени кандидата технических наук по специальности 1.2.1. Искусственный интеллект и машинное обучение.

Отзыв на диссертацию Мансур Али Махмуд на тему: «Модель, метод и алгоритмы Data Mining для интеллектуальной обработки и анализа текстов на естественном языке», представленной на соискание ученой степени кандидата технических наук по специальности 1.2.1. Искусственный интеллект и машинное обучение, рассмотрен и утвержден на заседании кафедры «Искусственный интеллект и

цифровых технологий, ФГБОУ ВО «Воронежский государственный технический университет».

По результатам обсуждения диссертации сформулировано положительное заключение. Присутствовало на заседании 19 человек, из них 5 докторов наук. Проголосовало за утверждение заключения 19 человек, против 0.

Протокол № 14 от «21» мая 2025 г.

Заведующий кафедрой «Искусственного интеллекта и цифровых технологий», к.т.н. (по специальности 05.13.18 Математическое моделирование, численные методы и комплексы программ), доцент — Гусев Павел Юрьевич

«21» мая 2025 г.

Профессор кафедры «Искусственного интеллекта и цифровых технологий», д.т.н. (по специальности 2.3.4. Управление в организационных системах), доцент — Рындин Никита Александрович

«21» мая 2025 г.

Федеральное государственное бюджетное образовательное учреждение высшего образования «Воронежский государственный технический университет» (ВГТУ).
394006, Воронежская область, город Воронеж, улица 20-летия Октября, дом 84.
Тел. +7(473) 271-59-05. Адрес эл. почты: rector@cchgeu.ru.

Личную подпись Гусева П.Ю., Рындина Н.А. УДОСТОВЕРЯЮ
«21» 05 2025 г.
Начальник отдела по работе с персоналом О.В. Кулова