

*На правах рукописи*



Мансур Али Махмуд

**МОДЕЛЬ, МЕТОД И АЛГОРИТМЫ DATA MINING ДЛЯ  
ИНТЕЛЛЕКТУАЛЬНОЙ ОБРАБОТКИ И АНАЛИЗА  
ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ**

Специальность – 1.2.1. Искусственный интеллект и машинное обучение

**Автореферат**  
диссертации на соискание ученой степени  
кандидата технических наук

Таганрог – 2025

Работа выполнена в ФГАОУ ВО «Южный федеральный университет» на кафедре систем автоматизированного проектирования Института компьютерных технологий и информационной безопасности

Научный руководитель:

доктор технических наук, доцент

**Кравченко Юрий Алексеевич,**

ФГАОУ ВО «Южный федеральный университет», профессор кафедры систем автоматизированного проектирования (г. Таганрог).

Официальные оппоненты:

доктор технических наук, профессор

**Кравец Алла Григорьевна**

ФГБОУ ВО «Волгоградский государственный технический университет» (ВолгГТУ), профессор кафедры «Системы автоматизированного проектирования и поискового конструирования» (г. Волгоград)

доктор технических наук, профессор

**Вишняков Юрий Муссович**

ФГБОУ ВО «Кубанский государственный университет» (КубГУ), профессор кафедры «вычислительных технологий» (г. Краснодар)

Ведущая организация:

ФГБОУ ВО «Воронежский государственный технический университет» (ВГТУ, г. Воронеж).

Защита диссертации состоится «26» июня 2025 г. В 14<sup>00</sup> на заседании объединенного совета по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук 99.2.107.02 на базе ФГАОУ ВО «Южный федеральный университет», ФГБОУ ВО «Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова» по адресу: 347922, г. Таганрог, пер. Некрасовский, 44, ауд. Г-439.

С диссертацией можно ознакомиться в зональной научной библиотеке ФГАОУ ВО «Южный федеральный университет» по адресу: 344015, г. Ростов-на-Дону, ул. Зорге, 21Ж, а также на портале электронных ресурсов ЮФУ: <https://hub.sfedu.ru/diss/show/1338090/>.

Отзыв на автореферат, заверенный гербовой печатью организации, просим направлять ученому секретарю объединенного диссертационного совета 99.2.107.02 по адресу: 347922, г. Таганрог, пер. Некрасовский, 44.

Автореферат разослан «\_\_\_» 2025 г.

Ученый секретарь объединенного  
диссертационного совета 99.2.107.02,  
доктор технических наук, доцент

Н. Е. Сергеев

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность темы исследования и степень её разработанности.** В эпоху цифрового прогресса и массового применения технологии искусственного интеллекта (ИИ) экспоненциальный рост объёмов текстовых данных представляет собой серьезную проблему. Значительный объем неструктурированных текстов на естественном языке, генерируемых в различных областях, требует создания эффективных методов их обработки и анализа для извлечения скрытых закономерностей, позволяющих повысить эффективность алгоритмов машинного обучения в данной области. Проблема необходимости повышения эффективности процессов обработки и анализа текстов на естественном языке подчеркивает значимость методов *Data Mining* при решении задач классификации и кластеризации для структурирования текстовых данных.

Векторное представление (векторизация текстов) является одной из основных моделей пространства решений при классификации и кластеризации текстовых документов в системах искусственного интеллекта. Векторизация документов позволяет использовать различные математические операции и алгоритмы машинного обучения для выявления закономерностей, связей и тенденций в них, тем самым способствуя развитию приложений искусственного интеллекта в различных областях науки и техники.

Традиционные модели представления текстов, такие как «мешок слов» (BoW, Bag-of-Words) и TF-IDF (term frequency invers document frequency), обладают простотой и сравнительно высокой точностью, что делает их эффективными для задач классификации и кластеризации. Векторы, построенные на основе этих моделей, являются интерпретируемыми, но имеют два ключевых недостатка: разреженность данных с высокой размерностью и отсутствие учёта семантических отношений между словами.

Методы векторизации на основе нейронных сетей, такие как встраивание слов (word embeddings), способны выявлять скрытые закономерности и создавать плотные и низкоразмерные представления, но их применение для кодирования целых документов неэффективно из-за потери информации при усреднении или суммировании векторов слов. Кроме того, такие векторы не интерпретируемые. Также, методы на основе трансформеров создают низкоразмерные векторные представления и показывают высокие результаты в задачах классификации и поиска. Однако их эффективность снижается при работе с длинными документами, так как требуется либо усечение текста (с потерей информации), либо модификация модели, что увеличивает вычислительные затраты.

Таким образом, существующие методы векторизации текстов не позволяют обеспечить семантические представления документов (векторов) с малыми размерностями, которые можно интерпретировать без негативного влияния на эффективность алгоритмов классификации и кластеризации. Векторы с высокой размерностью включают в себя большее количество признаков, что повышает эффективность интеллектуального анализа текстов, но приводит к нехватке вычислительных ресурсов, занимает больше памяти и

отрицательно влияет на масштабируемость. *Интерпретируемость* векторов признаков позволяет обнаружить ошибки, что увеличивает доверие пользователей к таким моделям.

В данном исследовании применено комплексное решение для построения низкоразмерных, интерпретируемых векторных представлений текстов на основе концептов. *Концепт* – это набор слов или фраз, имеющих общее семантическое значение. В этом подходе для представления текста вместо слов используются концепты, где каждый элемент вектора соответствует одному концепту, что позволяет снизить размерность пространства решений и сохранить интерпретируемость.

Таким образом, **актуальной научной задачей** для развития отрасли искусственного интеллекта и машинного обучения является разработка *моделей, методов и алгоритмов Data Mining для интеллектуальной обработки и анализа текстов на естественном языке*, позволяющих снизить частоту ошибок при классификации и кластеризации текстов.

Ряд работ посвящен развитию методов анализа текста и методов представления текста для целей классификации и автоматизированной кластеризации. Источниками для проведенного исследования послужили работы отечественных и зарубежных ученых по основам текстового анализа, взвешивания терминов и извлечения информации: Р. Муни; Х. Шютце и К. Д. Мэннинг; К. С. Джонс; М. А. Хёрст; А. Панченко; И. Д. Иванович и А. Кутузов. Работы Н. Красвелл и Б. Митра, Д. М. Блей, Д. Уэстон, Д. Юрафски, П. Сердюкова и И. Титова вносят значительный вклад в представление текста, векторизацию и модели семантического поиска.

**Целью диссертационной работы** является повышение эффективности моделей, методов и алгоритмов классификации и кластеризации текстов. Под эффективностью понимается минимизация частоты ошибок классификации и кластеризации текстов при условии снижения размерности векторного пространства признаков с сохранением его интерпретируемости.

Для достижения поставленной цели были решены следующие основные задачи:

1. Проведён аналитический обзор современных методов Data mining и методов векторизации текстов на естественном языке.
2. Построена модель векторизации текстов с использованием алгоритмов извлечения ключевых фраз и алгоритмов кластеризации.
3. Разработан модифицированный метод генерации векторных представлений документов на основе алгоритмов обработки и анализа текстов в системах искусственного интеллекта.
4. Разработан алгоритм извлечения и фильтрации ключевых фраз на основе парсера.
5. Разработан алгоритм построения концептов на основе алгоритмов извлечения ключевых фраз и кластеризации.
6. Разработано программное приложение для проведения вычислительного эксперимента и подтверждения достоверности и эффективности полученных основных результатов.

**Объект исследования** – тексты на естественном языке.

**Предмет исследования** – модели, методы и алгоритмы обработки и анализа текстов для решения задач классификации и кластеризации текстовых документов.

**Методология и методы диссертационного исследования.** При выполнении диссертационной работы использовались методы интеллектуального анализа данных, методы обработки и анализа текстов на естественном языке, методы системного анализа, теории информационных систем, формальной логики, машинного обучения, а также методы объектно-ориентированного программирования.

**Научная новизна и соответствие научной специальности:**

1. Построена математическая модель векторизации текстов на основе концептов, **отличающаяся** применением новых правил построения эталонных концептов и новых функций определения их весов, **позволяющая** снизить размерность векторного пространства и улучшить дискриминационную способность результирующих векторов признаков (пункт 4 паспорта специальности 1.2.1, страницы 57-64 диссертации).

2. Разработан модифицированный метод генерации векторных представлений документов на основе построенной модели векторизации, **отличающийся** применением интерпретируемых признаков при векторизации, **позволяющий** снизить частоту ошибок алгоритмов классификации и кластеризации документов (пункт 4 паспорта специальности 1.2.1, страницы 53-67 диссертации).

3. Разработан алгоритм извлечения и фильтрации ключевых фраз на основе частоты их появления, **отличающийся** применением функции парсера для разметки частей речи, **что позволяет** извлекать ключевые фразы с правильной грамматической структурой (пункт 5 паспорта специальности 1.2.1, страницы 69-74 диссертации).

4. Разработан алгоритм построения концептов из семантически близких фраз, **отличающийся** решением задачи кластеризации фраз с учетом контекстуальной семантической близости, **что позволяет** повысить однородность кластеров, представляющих концепты (пункт 5 паспорта специальности 1.2.1, страницы 74-80 диссертации).

**Теоретическая значимость работы.** Полученные научные результаты развивают аппарат искусственного интеллекта и машинного обучения в области решения важной научной проблемы увеличения информационного объема семантически обработанных текстов в информационном пространстве; разработка методов и алгоритмов машинного обучения для обработки и анализа текстов на естественном языке, в том числе, методов векторизации, классификации и кластеризации текстов; исследования и разработки средств представления текстов.

**Практическая значимость работы** заключается в создании программного приложения, позволяющего использовать разработанные модель, метод и алгоритмы обработки и анализа текстов на естественном языке в системах искусственного интеллекта для минимизации частоты

появления ошибок при решении задач классификации и кластеризации с учётом условий снижения размерности векторного пространства и сохранения его интерпретируемости.

**Степень достоверности результатов.** Достоверность научных результатов работы подтверждается непротиворечивостью и согласованностью с известными фактами и исследованиями в рассматриваемой области, высокой степенью сходимости теоретических результатов с данными экспериментов, и определяется применением теоретических и методологических основ разработок ведущих ученых в области создания интеллектуальных систем, корректным и обоснованным использованием математического аппарата, экспериментальными исследованиями разработанных моделей и методов.

**Реализация и внедрение результатов работы.** Теоретические и практические результаты работы внедрены в информационные процессы ИТ-компании ООО «Ит-Эффект» (г. Москва). Полученные в работе научные результаты позволили повысить эффективность решения задач классификации, кластеризации и извлечения ключевых фраз в рекомендательной системе, реализующей технологию «look-alike» (поиск целевой аудитории для эффективного масштабирования деловой активности предприятия). Результаты работы также используются в учебном процессе института компьютерных технологий и информационной безопасности Южного федерального университета.

**Апробация результатов диссертации.** Основные положения и отдельные результаты исследования докладывались и обсуждались на следующих конференциях: VI International Conference on Information Technologies in Engineering Education (Inforino 2022), (Россия, Москва, апрель 2022); VI Всероссийская научно-техническая конференция «Фундаментальные и прикладные аспекты компьютерных технологий и информационной безопасности», (Россия, Таганрог, 2020); XVIII, XIX и XX Всероссийская конференция молодых ученых аспирантов и студентов «Информационные технологии, системный анализ и управление ИТСАУ» (Россия, Таганрог, 2019-2022); II научно-методическая конференция НПР «Современные компьютерные технологии» (Россия, Таганрог, 2021-2022); XII международная научно-техническая конференция «технологии разработки информационных систем (ТРИС-2022)» (Россия, Таганрог, 2022); «5th International Scientific Convention UCIENCIA» (Куба, сентябрь 2023); International Russian Automation Conference RusAutoCon, (Россия, Сочи, 2023);

**Публикации.** По теме диссертации опубликовано 17 научных работ, из которых: 3 статьи опубликованы в издании из перечня рекомендованных ВАК (К2), в т.ч. 1 статья опубликована без соавторов; 2 статьи – в изданиях из международных баз данных Scopus и/или Web of Science. Получены 2 свидетельства об государственной регистрации программ для ЭВМ. В трудах всероссийских и международных конгрессов и конференций опубликовано 9 работ.

**Личный вклад автора.** Все выносимые на защиту результаты и положения, составляющие основное содержание диссертационной работы, разработаны и получены лично автором или при его непосредственном участии. В работах, опубликованных в соавторстве, соискателю принадлежит определяющая роль в развитии информационных процессов моделей и методов обработки и анализа текстов на естественном языке.

**Структура и объем работы.** Диссертация состоит из введения, 4 разделов, заключения, списка литературы, содержащего 146 наименований, и 2 приложений. Основная часть работы содержит 150 страниц, включая 31 рисунок и 12 таблиц.

**Область исследования.** Новые результаты, полученные в ходе выполнения диссертационного исследования, соответствуют пунктам 4 и 5 паспорта научной специальности 1.2.1. Искусственный интеллект и машинное обучение.

#### **Положения, выносимые на защиту:**

1. Математическая модель векторного представления текстовых документов 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 документов на основе построенной модели векторизации **позволяет** снизить частоту ошибок алгоритмов классификации и кластеризации 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.

#### **КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ**

**Во введении** сформулирована цель работы, обоснована актуальность темы диссертации, описаны основные научные положения, выносимые на защиту, научная новизна, теоретическая и практическая ценность, апробация диссертационной работы, реализация и внедрение, а также структура диссертации.

**В первом разделе** представлен аналитический обзор научных исследований в области обработки и анализа текстов на основе методов искусственного интеллекта и машинного обучения. Даны определения основных терминов необходимых для понимания предметной области диссертационного исследования. Проведён аналитический обзор основных задач и методов интеллектуального анализа данных. Изучены этапы процесса интеллектуального анализа данных и перспективы использования этих этапов для обработки и анализа текстов на естественном языке.

Проведённый анализ исследований показал, что методы представления текста, в частности векторного, становятся актуальными при экспоненциальном росте объёмов текстов и являются наиболее значимыми

при решении задач обработки и анализа текстов на естественном языке. При анализе методов представления текста основное внимание уделялось двум важным свойствам векторов, а именно, размерности и интерпретируемости.

Кроме того, обоснована необходимость использования *концептов* для векторизации документов, при этом каждый элемент вектора соответствует одному концепту. Концепты извлекаются из самого текста без применения внешних источников, таких как базы знаний и онтологии. Для этой цели проанализированы методы векторизации документов на основе концептов и выявлены недостатки, препятствующие достижению цели диссертации. К таким недостаткам относятся: низкая эффективность алгоритмов построения концептов из свободного текста; недооценка значимости редких концептов, а также неадекватное представление концепта вектором центроида. В результате, сформулирована следующая *постановка задачи исследования*.

Дан набор текстовых документов  $D$ , представленных в виде последовательностей слов. Требуется построить модель для классификации и кластеризации документов в заранее определенные классы (для задачи классификации) или в кластеры (для задачи кластеризации). Таким образом, необходимо найти преобразование  $y = f(x): R^{\text{слово}} \rightarrow R^{\text{концепт}}$  из пространства слов в пространство концептов, такое, что преобразованный вектор признаков  $y_i \in R^{\text{концепт}}$  сохраняет большую часть структуры в  $R^{\text{слово}}$ .

Разрабатываемая модель должна обеспечивать *интерпретируемые* векторные представления текста, что позволяет понять причины отнесения документа к тому или иному классу. Сгенерированные векторы должны быть *малоразмерными*, чтобы алгоритм машинного обучения  $\alpha$  не требовал больших вычислительных затрат. В итоге, применение этих векторов должно привести к *снижению частоты ошибок алгоритмов классификации и кластеризации* ( $E_\alpha \rightarrow \min$ ).

**Во втором разделе** построена математическая модель векторизации текстов на основе концептов, включающая комплекс алгоритмов и функций обработки и анализа текстов, которые в совокупности образуют метод построения семантического векторного представления документов. Для каждого документа строится семантический вектор, состоящий из  $N$  признаков. Каждый признак представляет степень присутствия концепта в словаре эталонных концептов документа. Такое представление считается линейным преобразованием пространства слов в пространство концептов, позволяющее управлять размерностью создаваемых векторов. На рис. 1 показаны основные операции предлагаемого метода. Синие блоки обозначают предлагаемые новые алгоритмы и функции, компенсирующие недостатки канонических методов. Внутренняя логика предлагаемого метода состоит из двух этапов, на первом из которых создается словарь эталонных концептов, используемый в качестве основы для построения векторов. Второй этап заключается в анализе каждого документа и извлечении содержащихся в нем концептов с последующим их взвешиванием путем сопоставления со словарем эталонных концептов.

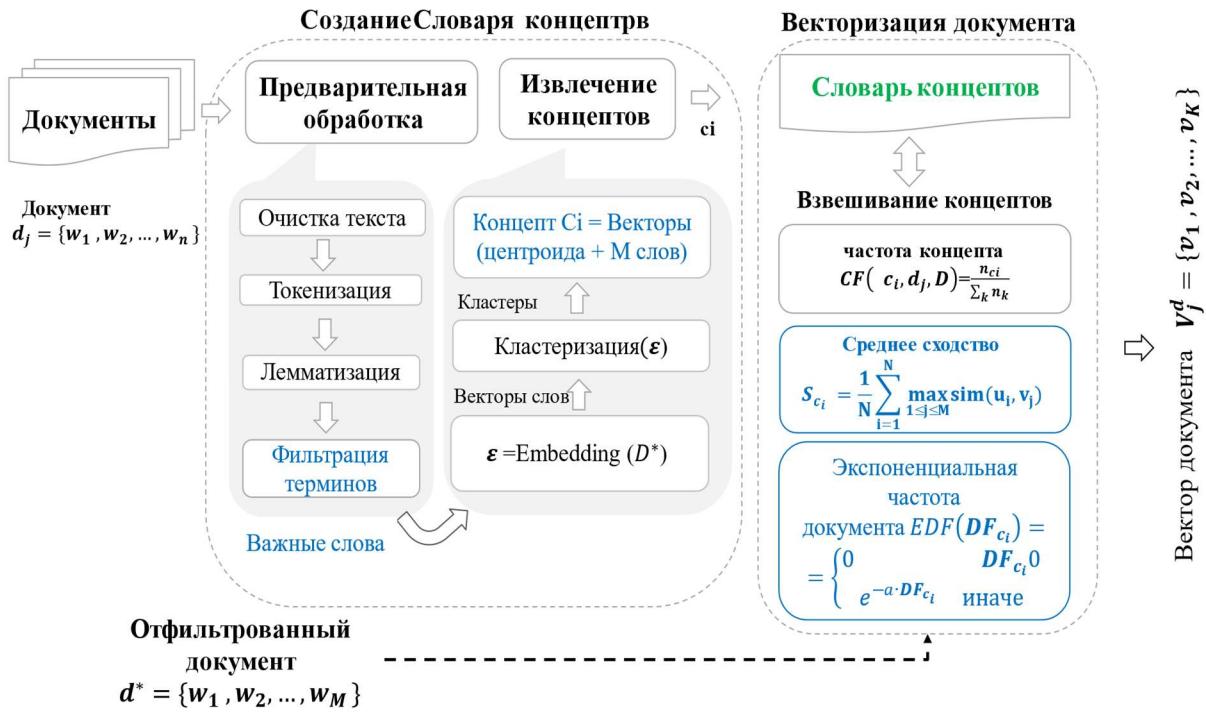


Рисунок 1 – Функциональная схема метода BoWC

Для построения концептов, все векторы слов кластеризуются в  $N_k$  кластеров с использованием алгоритма кластеризации сферических К-средних, который использует косинусную меру сходства для измерения расстояния между векторами слов. После этого слова распределяются между кластерами в соответствии с мерой близости к центроиду кластера, затем выбираются  $M$  слов, наиболее близких к центроиду кластера (рис. 2). Пусть  $C_1 = (w_1^1, w_2^1, \dots, w_M^1)$  – это набор слов в кластере, а  $S = (s_1^1, s_2^1, \dots, s_M^1)$  – это набор оценок близости слов с центроидом, тогда  $[w'_1, w'_2, \dots, w'_M]$  – множество слов из  $C$ , отсортированных в порядке убывания их оценок близости.

В итоге получается набор концептов (понятий), каждое из которых представлено  $M$  словами, соответствующими одному общему понятию. Полученный набор кластеров образует *словарь эталонных концептов*, который представляется следующим образом:

$$C = clustering(\epsilon) = (w_1^1, w_2^1, \dots, w_M^1, w_1^2, w_2^2, \dots, w_M^2, \dots, w_1^K, w_2^K, \dots, w_M^K), \quad (1)$$

где  $w_i^j$  –  $i$ -е слово  $j$ -го кластера.

Фильтрация нерелевантных слов снижает шум от получаемых кластеров, поскольку позволяет избавиться от схожих концептов. Это приводит к формированию более однородных кластеров концептов.

**Векторизация документов.** Документ кодируется посредством процесса сопоставления между документом и словарем с использованием мер оценки семантической близости. Для рассматриваемого документа  $d$  создается вектор признаков  $V^d$  размера  $K$ , равного количеству концептов,

$$V^d = vectorization(C, d^*, D^*, \theta) = \{c_1^d, c_2^d, \dots, c_K^d\}, \quad (2)$$

где  $c_i^d$  выражает степень значимости  $i$ -го концепта (его вес) в  $d$ , которая определяется аналогично методу BoC (Bag of concepts). Однако вместо того, чтобы сопоставить слово документа  $w_d$  с центроидом кластера, оно

сопоставляет ближайшими  $M$  словами к центроиду. Максимальное значение меры близости используется в качестве показателя сходства этого слова с концептом, как это показано в выражении (3):

$$S_{c,w_d} = \max_{1 \leq j \leq M} \text{sim}(w_d, w_j^c), \quad (3)$$

где  $w_d$  – слово документа  $d$ ;  $w_j$  –  $j$ -е слово концепта  $c$ . Функция  $\text{Max}$  возвращает максимальное значение оценки сходства слова документа  $w_d$ , соответствующего концепту  $c$ . Итоговая оценка сходства концепта с документом  $S_c$  вычисляется как среднее значение меры косинусного сходства между концептом и словами документа, которые соответствуют этому концепту:

$$S_c = \frac{1}{N} \sum_{i=1}^N S_{c,w_i} = \frac{1}{N} \sum_{i=1}^N \max_{1 \leq j \leq M} \text{sim}(w_i, v_j), \quad (4)$$

где  $w_i$  –  $i$ -е слово документа, принадлежащее концепту, а  $\text{sim}$  это мера косинусного сходства между векторами слов, которая вычисляется по следующему выражению:

$$\text{sim}(X, Y) = \cos(\theta) = \frac{X \cdot Y}{\|X\| \cdot \|Y\|} = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2} \sqrt{\sum_{i=1}^n Y_i^2}}, \quad (5)$$

где  $X_i$  и  $Y_i$  – компоненты вектора слов  $X$  и  $Y$  соответственно. На основе вычисленного значения сходства  $S_c$  между концептом и документом определяется частота концепта в документе:

$$CF(c_i, d_j, D) = \frac{n_c}{\sum_k n_k}, \quad (6)$$

где  $n_k$  – общее количество концептов в документе, а  $n_c$  – количество вхождений концепта  $c$  в документ. В работе принято, что концепт содержится в документе, если значение сходства  $S_c$  концепта со словом документа превышает заданный порог ( $\Theta$ ) который определяется экспертым путем.

$$g(s) = \begin{cases} 1, & S_c > \Theta \\ 0, & \text{иначе} \end{cases}. \quad (7)$$

На рисунке 2 показан процесс определения наличия концепта  $C_1$  в документе. Концепт  $C_1$  присутствует в документе, так как сходство слова документа со словами концепта высокое ( $S=0.9$ ).

Чтобы максимально использовать рассчитанную статистику и повысить дискриминационную способность концептов, автор предлагает включить функцию  $S_c$  в формулу определения веса концепта, которая выглядит следующим образом:

$$BoWC_{c_i} = \frac{n_{c_i}}{\sum_k n_k} \cdot e^{S_i^j}. \quad (8)$$

Использование функции  $IDF$  (Inversed Document Frequency) для уменьшения веса часто встречающихся концептов не подходит для концептуального представления, поскольку приводит к неверным выводам об объеме общих концептов. Поэтому автор предложил новую функцию взвешивания, основанную на обратной частоте документа, ранее установленной функцией монотонного убывания, цель использования которой состоит в том, чтобы придать большее значение характерным общим концептам. Функция задана следующим выражением:

$$f(F) = e^{-\alpha \cdot DF}, \quad (9)$$

где  $\alpha$  – константа (по умолчанию  $\alpha=1$ ), а  $DF = \frac{|\{d \in D | c_i \in d\}|}{|D|}$  – частота документа. Окончательная формулировка выражения для определения весовой функции концепта принимает следующий вид:

$$CF - EDF = \frac{n_{c_i}}{\sum_k n_k} \cdot e^{-\frac{|\{d \in D | c_i \in d\}|}{|D|}} \cdot e^{S_{c_i}}. \quad (10)$$

В результате метод генерирует вектор признаков  $V^d = \{v_1^d, v_2^d, \dots, v_K^d\}$  для каждого заданного документа  $d$ , где  $v_i^j$  отражает важность (вес)  $i$ -го концепта, а  $k$  – количество концептов. Полученное представление сохраняет возможность интерпретации. Каждый элемент вектора документа содержит вес концепта в словаре эталонных концептов. Сопоставив словарь с вектором и весами, вектор можно легко интерпретировать.

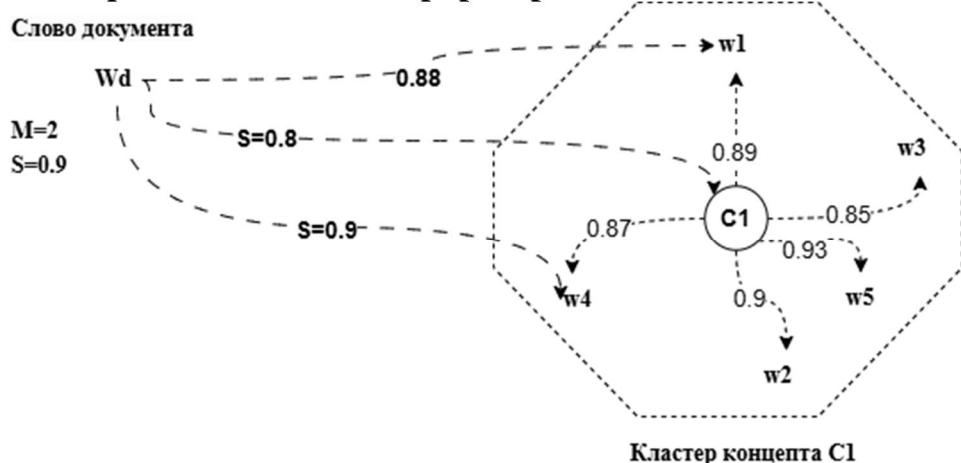


Рисунок 2 – Иллюстрация процесса сопоставления одного слова из документа с концептом

**Третий раздел** диссертационной работы посвящен разработке двух алгоритмов. Первый – это *алгоритм построения концептов* на основе ключевых фраз (*n*-грамм) вместо униграмм. Использование ключевой фразы обусловлено тем, что она подтверждает значение слова и раскрывает многозначность его смысла, расширяя соседними словами имеющийся контекст. Это повышает *однородность* концептов, что отражается в дискриминационной способности векторов, построенных на их основе. Высокая оценка *однородности* означает, что ключевые фразы, входящие в кластер, соответствуют одному конкретному концепту.

Алгоритм построения концептов (Алгоритм 1) создан на основе алгоритма *FBKE* (Frequency and Bert-based Keyword Extraction), который определяет веса ключевых фраз, учитывая их частоту встречаемости в документе и семантическую близость к его контексту. Алгоритм начинается с предварительной обработки документов. Далее следует этап извлечения ключевых фраз-кандидатов на основе частоты их встречаемости в документе с помощью следующей формулы:

$$TF_{n-gra} = \frac{n_t^i}{\sum_k n_k^i}, \quad (11)$$

где  $n_k$  относится к общему количеству  $n$ -грамм определенного типа (с элементами  $i$ ), а  $n_t$  представляет количество вхождений термина  $t$  в документ. Не все полученные кандидаты подчиняются грамматическим правилам, регулирующим границы фразы. Поэтому применяется *второй разработанный алгоритм фильтрации ключевых фраз на основе применения лингвистического анализатора парсера* (Алгоритм 2), который отбирает фразы с правильной грамматической структурой. Если ключевая фраза-кандидат соответствует одной из именных фраз, она сохраняется. Если именная фраза является частью фразы-кандидата, в окончательном списке сохраняется именная фраза, так как она имеет более полную структуру.

---

Алгоритм 1 – алгоритм построения концептов на основе ключевых фраз

---

- 1:   **Ввод**  $D = \{d_1, d_2, \dots, d_N\}$  //множество  $N$  документов
- 1:   **Вывод**  $CD$  //словарь эталонных концептов
- 2:    $D^* = preprocessing(D)$
- 3:    $CKW = extract\_candidates\_FBKE(D^*)$ // извлечения ключевых фраз-кандидатов на основе частоты их встречаемости в документе
- 4:    $KW\_FNP = Filtering(D^*)$  //алгоритм извлечения ключевых фраз на основе парсера
- 5:   **Foreach** candidate in CKW **do**:
- 6:     | **If** candidate in KW\_FNP:
- 7:       | | KWS [] **←** candidate // добавить кандидата в список
- 8:       | | **Else** // проверка наличия общих терминов между кандидатом и именной фразой
- 9:       | | **Foreach** np in FNP **do**
- 10:       | | | **If** np  $\cap$  candidate  $\neq \emptyset$  **do**:
- 11:       | | | | KWS [] **←** np // добавить np в список
- 12:       | | | | **End**
- 13:       | | **End**
- 14:     | **End**
- 15:    | **CKW = sim(D^\*, KW\_FNP)**//Оценка семантической близости фраз к документу
- 16:    | **CKW = ranking(D^\*, KW\_FNP)**// ранжирование выбранных
- 17:    | | ключевых фраз на основе их частоты и семантической близости по формуле (12)
- 18:    | |  $UKWS = set(KWS)$  // Выбор уникальных фраз для кластеризации
- 19:    | |  $concepts = Clustering(UKWS)$  // кластеризация ключевых фраз
- 20:    | | Сохранить кластеры в виде словаря  $C$ .

Семантическая близость фразы к контексту документа вычисляется с помощью меры косинусного сходства между SBERT-векторами этих фраз. Окончательный вес фразы-кандидата определяется следующей формулой:

$$rel(d, c^j) = 2 \cdot \frac{tf_{c^j} \cdot s_{norm}^{j,d}}{tf_{c^j} + s_{norm}^{j,d}}, \quad (12)$$

где  $S_{norm}^{j,d} = S^j \cdot e^{S^j / |c^j|}$  – нормализованное значение сходства кандидата  $c^j$  документа  $d$ .  $S^j$  – косинусное сходство  $j$ -го кандидата с документом.  $tf$  – частота ключевых фраз. Результирующие ключевые фразы и их векторы используются для создания словаря эталонных концептов в следующем формате:

$$C = (kp_1^1, kp_2^1, \dots, kp_M^1, kp_1^2, kp_2^2, \dots, kp_M^2, \dots, kp_1^K, kp_2^K, \dots, kp_M^K), \quad (13)$$

где  $kp_i^j$  –  $i$ -я ключевая фраза  $j$ -го кластера.

*Алгоритм фильтрации ключевых фраз на основе применения парсера*, является универсальным и применяется в качестве фильтра с любым алгоритмом извлечения ключевых фраз.

---

**Алгоритм 2 – алгоритм фильтрации ключевых слов с помощью парсера**

---

```

1: Ввод текст
2: Вывод список ключевых фраз KW_FNP
3:  $T^* = preprocessing(T)$  //удаление ненужных символов, знаков препинания, стоп-слов и перевод текста в нижний регистр.
4:  $S = parsing(T^*)$  // Разбор текста на предложения  $S$  и их атрибуты
5:  $NP \leftarrow get\_NP\_NE(S)$ //сохранение именной фразы и сущности в списке
6: // преобразования именных фраз в ключевые фразы
7:  $KW\_FNP = []$  //Список для хранения отфильтрованных именных фраз
8: Fforeach np in NP do
9:   if len(np) == 1:
10:    if np[0].pos_ in ['NOUN', 'PROPN', 'ADJ']:
11:      // первое слово pr – существительное, имя собственное или прилагательное
12:      KW_FNP.append(np.text.strip())
13:    elseif 1 < len(np) < 6: // если длина pr не превышает 6 слов
14:      if np[0].pos_ in ['DET']: //первый термин – это артикль
15:        KW_FNP.append(np[1:].text) // удалить артикль
16:      Else
17:        KW_FNP.append(np.text.strip())
18:      Else // длина np превышает 6 слов
19:        root_deps = np.root.dep_
compound = [t.text for t in np.children if "compound" in root_deps]
KW_FNP.append(np.root.text + " " + compound) //Сохранить корень фразы или составные существительные
End if
End for
19: Return KW_FNP

```

Алгоритм 2 сначала обрабатывает текст для удаления ненужных символов, знаков препинания, стоп-слов и переводит текст в нижний регистр. После этого текст анализируется с помощью парсера, который выделяет отдельные предложения и определяет их составные части: глаголы; именные фразы; именованные сущности; части речи для каждого слова и связи между ними.

Это позволяет отфильтровать бессмысленные слова, которые не являются частью составного существительного. Длинные словосочетания и фразы, которые не начинаются с существительного или прилагательного, исключаются из рассмотрения. Таким образом, алгоритм гарантирует, что полученные ключевые фразы представляют собой не просто последовательность слов, а устоявшиеся фразы с правильной грамматической структурой.

Применение алгоритма извлечения ключевых фраз на этапе построения концептов позволяет модифицировать весовую функцию *CF-EDF* метода *BoWC*, выраженную формулой (11). Так как ключевые фразы документа выбираются на основе частоты их встречаемости в тексте и их семантической близости с контекстом документа, автор предлагает заменить терм  $S_{c_i}$  формулы (12) на средние веса ключевых фраз документа, которые относятся к концепту. Данный терм задаётся следующей формулой:

$$S_i^j = \frac{1}{N} \sum_{n=1}^N \text{rel}(c_i, kw_n^j), \quad (14)$$

где  $N$  – количество ключевых фраз документа ( $j$ -го), появившихся в текущем концепте  $c_i$ . Это снижает сложность алгоритма, поскольку нет необходимости повторно вычислять семантическую близость между концептами и документом.

Окончательная формулировка весовой функции концептов согласно *BoWC* принимает следующий вид:

$$\text{BoWC}^*_{c_i} = \frac{n_{c_i}}{\sum_k n_k} \cdot e^{\frac{|\{d \in D | c_i \in d\}|}{|D|}} \cdot e^{S_i^j}. \quad (15)$$

В результате этого процесса, метод *BoWC* генерирует вектор признаков  $V$  для каждого заданного документа  $d$ , где  $v_i^j$  отражает важность (вес)  $i$ -го концепта, а  $k$  – количество концептов.

$$V^d = \text{vectorization}(C, d^*, D^*) = \{v_1^d, v_2^d, \dots, v_K^d\}. \quad (16)$$

Полученный вектор признаков используется при проведении вычислительных экспериментов, результаты которых подтверждают эффективность предложенных автором решений.

**Четвертый раздел** диссертационной работы посвящён описанию разработки программного приложения, а также проведения серии вычислительных экспериментов по оценке эффективности, разработанных модели, алгоритмов и метода векторизации текстов. Рисунок 3 иллюстрирует компонентную архитектуру программного приложения, реализующего метода векторизации текстов, алгоритмов извлечения ключевых фраз и построения концептов.

Для оценки качества результатов классификации используется *мера F1*, которая представляет собой среднее значение точности (precision) и полноты (recall). Точность показывает, сколько из предсказанных положительных классов верны, а полнота – сколько реальных положительных классов найдено. Для оценки качества результатов кластеризации используется *мера VI*, которая представляет собой среднее гармоническое для различных оценок

однородности и полноты. *Однородность (Homogeneity)* показывает, насколько документы в кластере принадлежат одной группе, а *полнота (completeness)* – насколько все документы группы попали в один кластер.

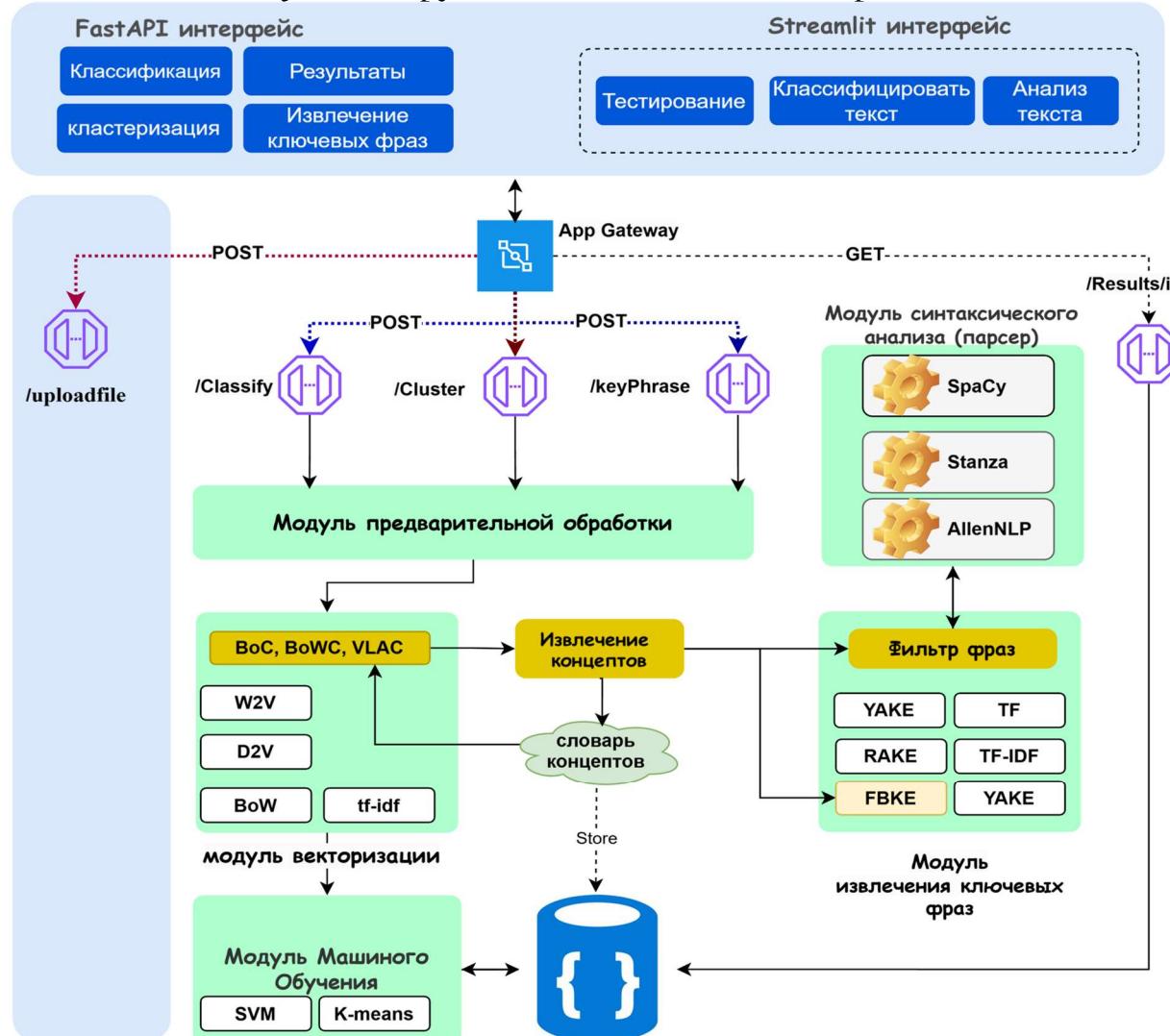


Рисунок 3 – Компонентная архитектура программного приложения

В таблицах 1-2 представлены результаты сравнения предложенного метода *BoWC* с каноническими методами в задачах классификации и кластеризации документов из пяти англоязычных стандартных наборов данных. Эксперименты показали, что метод *BoWC* снижает частоту ошибок алгоритмов классификации и кластеризации текстовых документов на 0,2-1% и 2-6% соответственно, а также позволяет уменьшить число признаков по сравнению с конкурирующими методами.

Метод *BoWC* не превосходит *TF-IDF* в задаче классификации наборов данных: *OHSUMED (OH)*; *20newsgroup (20NG)*; *Webkb*. Автор утверждает, что это связано с тем, что в некоторых текстах недостаточно слов для извлечения информации о каждом понятии, что снижает вероятность построения лучшего вектора концептов. Это приводит к выводу, что метод *BoWC* лучше работает с длинными документами. Производительность *BoWC* зависит от используемого метода встраивания слов (*Glove, W2V, FASTTEXT*) и пороговых значений сходства (рис. 4). Также векторные представления, полученные в

результате обучения метода встраивания на самих наборах данных (self-embedding), работают намного лучше, чем предварительно обученные представления (pre-trained).

Таблица 1 – Результаты классификации, измеренные с помощью F1

Методы (размер вектора)	BBC	RE	OH	20NG	WebKB
<i>BoWC</i> <sub>Разработанный</sub> (100)	<b>98.2</b>	94.23	71.25	76	87.3
<i>BoWC</i> <sub>Разработанный</sub> (200)	98	<b>94.75</b>	<b>72.0</b>	<b>85.43</b>	<b>87.93</b>
<i>BoC<sub>CF_IDF</sub></i> (200) (аналог)	95.48	90.9	48.8	69.44	78.26
TF-IDF (>20000)	97.89	94.7	<b>74.31</b>	<b>91.49</b>	<b>89.5</b>
BoW (>20000)	96.98	94.15	69.89	83.85	85.56
Averaged GloVe (300)	97.75	93.32	67.39	78.7	80.22
self-averaged w2v (300)	91.599	87.81	56.78	62.37	83.27
VLAC (9000)	<b>98.2</b>	94.06	69.37	86.4	85.54

Таблица 2 – Результаты кластеризации, измеренные с помощью V1

Методы (размер вектора)	Значения V1-мера (%)				
	BBC	RE	OH	20NG	WebKB
<i>BoWC</i> <sub>Разработанный</sub> (100)	<u>83.2</u>	<u>61.3</u>	<u>15.3</u>	<u>40.3</u>	<u>38.9</u>
<i>BoWC</i> <sub>Разработанный</sub> (200)	<b>82.7</b>	<b>64.3</b>	<b>15.54</b>	<b>42.5</b>	<b>39.0</b>
<i>BoC<sub>CF_IDF</sub></i> (200) (аналог)	74.3	52.7	10	39.4	8.8
TF-IDF (>20000)	66.3	51.3	12.2	36.2	31.3
BoW (>20000)	20.9	24.8	02.7	02.1	02.1
Averaged GloVe (300)	77.4	48.1	10.9	38.1	21.9
self-averaged w2v (300)	63.1	55.7	11.5	32.5	32.4
VLAC (9000)	80.8	45.6	11.8	—	28.6

### Результаты классификации набора данных Reuters со 100 концептами

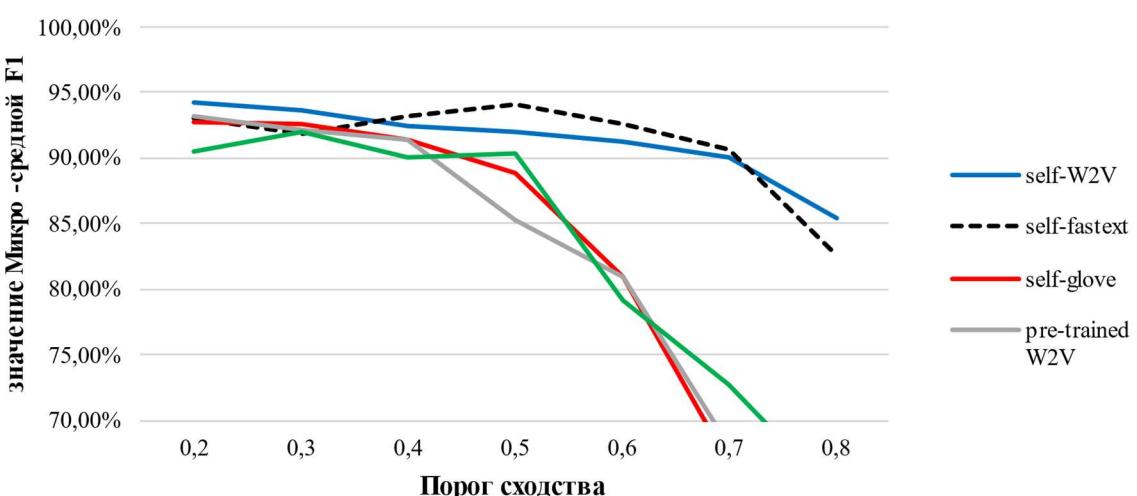


Рисунок 4 – зависимость точности классификации набора данных *Reuters* от метода встраивания слов и пороговых значений сходства

Предложенный алгоритм извлечения ключевых фраз был реализован с использованием трех различных парсеров (SpaCy, AllenNLP, Stanza).

Результаты работы данного алгоритма сравнивались с результатами алгоритмов *Yake* и *Rake*. Применялось точное совпадение (exact matching), при котором автоматически извлеченная ключевая фраза из текста должна полностью соответствовать ключевой фразе эталонного паттерна. Это объясняет низкие показатели точности всех рассмотренных алгоритмов. Результаты показали, что предложенный алгоритм превзошел другие алгоритмы в среднем на 1% по метрике MAP@K на стандартном наборе данных *Inspec* (табл. 3).

Таблица 3 – Точность алгоритмов, измеренная с помощью MAP@K

Метод	MAP@K (%)				
	K	@1	@5	@10	@20
TF-SpaCy	<b>36</b>	<b>15</b>	<b>9.5</b>	7.7	
TF-Stanza	30.3	13.1	8.3	6	
TF-AllenNLP	29.9	12.8	8.1	6.4	
Yake	26.7	12.7	8.4	<b>7.75</b>	
Rake	16.9	10.3	8.1	7.4	

Хотя статистические алгоритмы *Yake* и *Rake* работают быстрее, чем алгоритмы на основе парсера, поскольку их работа не требует анализа текста, предложенный алгоритм (*TF-SpaCy*) превзошёл все рассмотренные канонические алгоритмы кроме *Rake*. Это связано с тем, что SpaCy использует переходы для пошагового построения дерева зависимостей, что делает его более эффективным. В то время как *AllenNLP* и *Stanza* применяют графовые модели, которые медленнее из-за использования множества созданных вручную признаков.

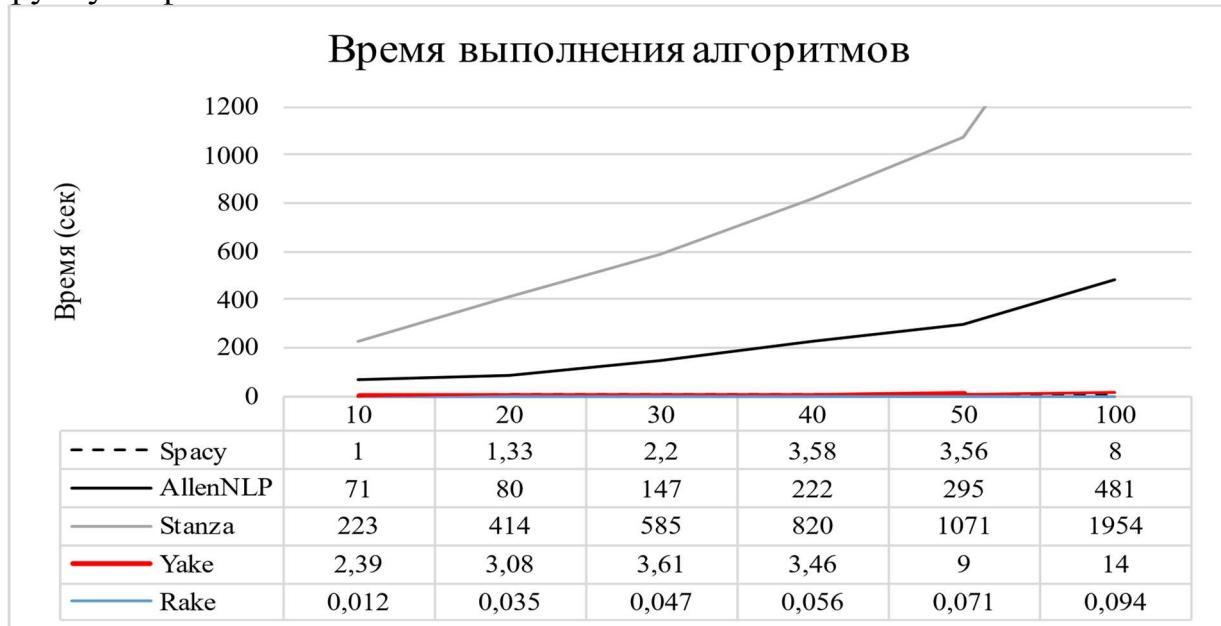


Рисунок 5 – Анализ времени выполнения алгоритмов

Результаты в таблице 4 показывают, что разработанный алгоритм построения концептов (n-граммы\_FBKE) превосходит другие аналоги и достигает значительного улучшения однородности кластеров, представляющих концепты. Это также повлияло на дискриминационную

способность векторов, созданных на основе этих кластеров, что привело к значительному повышению эффективности как метода *BoWC*, так и его аналога *BoC* при их применении для решения задачи кластеризации текстов. Это наглядно отразилось в снижении частоты ошибок алгоритма кластеризации на 2-6% по сравнению с исходной версией обоих методов (рис. 6, табл. 5).

Таблица 4 – Однородность концептов, измеренная с помощью *Индекса Дэвиса-Боулдина*

<b>Алгоритм</b>	<b>Нормализованные значения (DBI) в диапазоне [0,1]</b>				<b>Среднее значение</b>
	Кластер 1	Кластер 2	Кластер 3	Кластер 4	
Униграммы	0.7	0.77	0.72	0.69	0.72
n-граммы_FBKE	<b>0.25</b>	<b>0.35</b>	<b>0.2</b>	<b>0.23</b>	<b>0.32</b>
n-граммы_Yake	0.48	0.46	0.45	0.49	0.5
n-граммы_Rake	0.65	0.62	0.6	0.64	0.636
n-граммы_TF-IDF	0.72	0.9	0.81	0.73	0.79

Таблица 5 – Результаты кластеризации, измеренные по V1-мере

<b>Методы (размер вектора)</b>	<b>Значения V1-мера (%)</b>				
	BBC	RE	ОН	20NG	WebKB
BoWC <sub>исходный</sub> (100)	76.8	<u>59.1</u>	12.9	31.6	31.8
BoWC <sub>исходный</sub> (200)	77.1	<u>54.5</u>	14.0	39.1	32.8
BoWC <sub>FBKE</sub> (100)	<u>83.0</u>	<u>62.3</u>	<u>15.4</u>	<u>41.4</u>	<u>39.0</u>
BoWC <sub>FBKE</sub> (200)	<b>83.2</b>	<b>64.0</b>	<b>15.5</b>	<b>42.6</b>	<b>39.1</b>
BoC (200) (аналог)	74.3	52.7	10	39.4	8.8
BoC <sub>FBKE</sub> (200) (аналог)	79.3	59	13	40.0	<b>20</b>
TF-IDF (>20000)	66.3	51.3	12.2	36.2	31.3
BoW (>20000)	20.9	24.8	2.7	02.1	02.1
Averaged FastText (300)	77.4	48.1	10.9	38.1	21.9

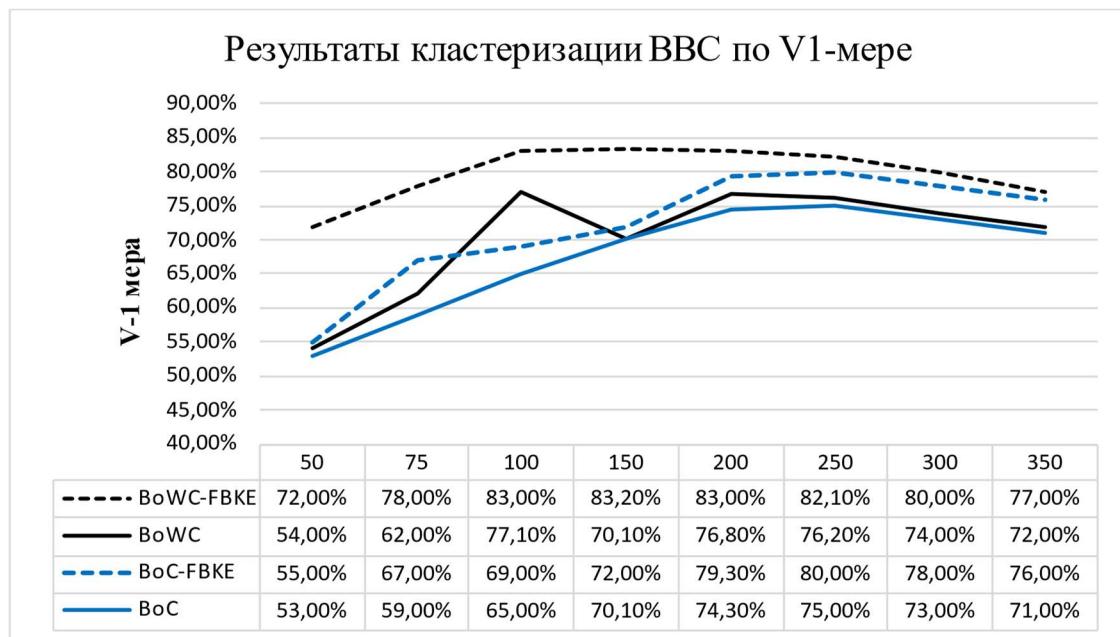


Рисунок 6 – Результаты кластеризации на наборе данных *BBC*

**Временная сложность.** Процесс векторизации документа состоит из двух вложенных циклов: один для прохождения всех концептов ( $k$  концептов), а второй для выполнения сравнения между каждым концептом со словами каждого документа ( $n$  слов). Таким образом, в зависимости от размера каждого документа и количества концептов временная сложность равна  $O(kn)$ . Так как в худшем случае  $k$  может быть равно  $n$ , тогда времененная сложность будет равной  $O(n^2)$ . В результате применения предложенных автором модели, метода и алгоритмов число концептов значительно сокращается и, согласно экспериментальным исследованиям, не превышает 200, поэтому значение переменной  $k$  всегда несоизмеримо меньше значения переменной  $n$ , тогда в общем случае временная сложность предложенного метода становится равной  $O(n)$ .

**В заключении** изложены итоги выполненного исследования, даны рекомендации, а также описаны перспективы дальнейшей разработки темы.

**В приложениях** приведены свидетельства об официальной регистрации программ для ЭВМ и копии актов о внедрении результатов работы.

## ЗАКЛЮЧЕНИЕ

Диссертация посвящена решению научной задачи создания моделей, алгоритмов и методов обработки и анализа текстов на естественном языке, в условиях экспоненциального роста их объёмов, что позволяет повысить эффективность средств и инструментов систем искусственного интеллекта и машинного обучения. Одной из основных моделей пространства решений при классификации и кластеризации текстовых документов в системах искусственного интеллекта является векторное представление (векторизация текстовых данных). Для решения поставленной задачи автором были разработаны модель, алгоритмы и метод Data mining для интеллектуальной обработки и анализа текстов на естественном языке, позволяющие снизить частоту ошибок при классификации и кластеризации текстов. Основные результаты диссертационной работы перечислены в следующих пунктах:

1. Проанализированы научные направления развития методов и алгоритмов векторизации текстов на естественном языке, выявлены основные недостатки, обоснована актуальность разработки. Сформулирована постановка задачи исследования;
2. Построена математическая модель векторизации текстов на основе концептов, отличающаяся применением новых правил построения эталонных концептов и новых функций определения их весов, позволяющая снизить размерность векторного пространства, и улучшить дискриминационную способность результирующих векторов признаков;
3. Разработан модифицированный метод генерации векторных представлений документов на основе построенной модели векторизации, отличающийся применением интерпретируемых признаков при векторизации, позволяющий снизить частоту ошибок алгоритмов классификации и кластеризации документов;
4. Разработан алгоритм извлечения и фильтрации ключевых фраз на основе частоты их появления в документе, отличающийся применением

функции парсера для разметки частей речи, что позволяет извлекать ключевые фразы с правильной грамматической структурой;

5. Разработан алгоритм построения концептов из семантически близких фраз, отличающийся решением задачи кластеризации фраз с учетом контекстуальной семантической близости, что позволяет повысить однородность кластеров, представляющих концепты;

6. Разработано программное приложение, позволяющее использовать предложенные автором модель, метод и алгоритмы обработки и анализа текстов на естественном языке в системах искусственного интеллекта для снижения частоты появления ошибок в алгоритмах классификации и кластеризации с учётом условий снижения размерности векторного представления текстов и сохранения его интерпретируемости;

7. Выполнена серия вычислительных экспериментов, которые подтвердили эффективность полученных решений, превосходящих результаты работы известных алгоритмов. В среднем предложенный автором метод векторизации текстов снижает частоту ошибок алгоритмов классификации и кластеризации документов на 0,2 – 1 % и 2 – 6 % соответственно, а также позволяет уменьшить число признаков по сравнению с конкурирующими методами.

В целом совокупность полученных в диссертации теоретических и практических результатов позволяет сделать вывод о том, что цель исследований достигнута, сформулированная научная задача решена. Перечисленные результаты получили высокую оценку научного сообщества при апробации и положительные рекомендации для внедрения в информационные процессы предприятий, учреждений и организаций различного профиля деятельности.

Дальнейшее развитие полученных результатов диссертационного исследования возможно в следующих основных направлениях:

1. Повышение эффективности метода векторизации документов при обработке коротких текстов за счет разработки алгоритмов и функций, которые обеспечат баланс между концептами коротких и длинных документов;

2. Применение разработанного алгоритма построения концептов в дополнительных областях, таких как разработка профилей пользователей на основе концептов, а также автоматизация процесса построения и расширения онтологии в системах ИИ для извлечения знаний из текстов;

3. Проведение дополнительных исследований алгоритма извлечения ключевых фраз для достижения баланса между точностью и скоростью парсера, в связи с важностью этой проблемы для повышения эффективности анализа текстов в реальном масштабе времени. Одним из возможных направлений улучшения предложенных решений является использование атомарных признаков, таких как униграммы слов и униграммы POS-тегов вместо использования большого количества признаков, созданных вручную;

4. Расширение словаря эталонных концептов за счёт интеграции концептов из онтологий или базы знаний;

5. Модификация разработанных метода и алгоритмов для их применения при обработке и анализе текстов на других языках.

## **ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИОННОЙ РАБОТЫ**

### **Публикации в изданиях, индексируемых в базах**

#### **«Scopus» и «Web of Science»:**

1. Mansour, A. Harnessing Key Phrases in Constructing a Concept-Based Semantic Representation of Text Using Clustering Techniques / A. Mansour, J. Mohammad, Y. Kravchenko, D. Kravchenko, N. Silega // Lecture Notes in Computer Science. – LNCS. – Vol. 14335. – 2023. – P. 190-201. (Scopus, Q2)

2. Mansour, A. Algorithm for Optimization of Keyword Extraction Based on the Application of a Linguistic Parser / A. Mansour, J. Mohammad, D. Kravchenko, Y. Kravchenko, N. Pavlov // Informatics and Automation. – Vol. 23. – 2024. – no. 2. – P. 467-494. (Scopus, Q4, RSCI)

#### **Публикации в изданиях, рекомендованных ВАК РФ:**

3. Мансур, А. М. Модифицированный метод устранения неоднозначности смысла слов, основанный на методах распределенного представления / А.М. Мансур, Ж.Х. Мохаммад, Ю.А. Кравченко // Известия ЮФУ. Технические науки. – 2021. – № 3 (220). – С. 92-101. (ВАК, К2)

4. Мансур, А. М. Векторизация текста с использованием методов интеллектуального анализа данных / Ю.А. Кравченко, А.М. Мансур, Ж.Х. Мохаммад // Известия ЮФУ. Технические науки. – 2021. – № 2 (219). – С. 154-167. (ВАК, К2)

5. Мансур А. М. Алгоритм на основе трансформеров для классификации длинных текстов/А. М. Мансур // Известия ЮФУ. Технические науки. – 2024. – №. 3. – С. 196-187. (ВАК, К2)

#### **Публикации в других изданиях, опубликовано 12 работ, основные из них следующие:**

6. Мансур, А. М. Метод извлечения ключевых фраз на основе новой функции ранжирования / А.М. Мансур, Ж.Х. Мохаммад, Ю.А. Кравченко, В.В. Бова // Информационные технологии. – 2022. – Том. 28. № 9. – С. 465-474. (ВАК, К1)

7. Mansour, A. Text vectorization method based on concept mining using clustering techniques / A. Mansour, J. Mohammad, Y. Kravchenko // VI International Conference on Information Technologies in Engineering Education, Inforino. – IEEE. – 2022. – P. 1-10. (Scopus)

8. Mansour, A. Generating Conceptual Semantic Vectors Based on Key Phrase Extraction Techniques / A. Mansour, J. Mohammad, Y. Kravchenko // 2023 International Russian Automation Conference (RusAutoCon). – IEEE. – 2023. – P. 374-379. (Scopus)

9. Мансур, А. М. Модифицированный метод построения семантического представления текста на основе методов кластеризации и взвешивания терминов / А.М. Мансур, Ж.Х. Мохаммад, Д.Ю. Кравченко, Ю.А. Кравченко // Труды XII международной научно-технической конференции «Технологии

разработки информационных систем (ТРИС-2022)». – Таганрог: 2022. – С. 94-100.

10. Мансур, А. М. Метод автоматического извлечения ключевых слов / А.М. Мансур, Ж.Х. Мохаммад, Д.Ю. Кравченко, Ю.А. Кравченко // Труды международного научно-технического конгресса «Интеллектуальные системы и информационные технологии – 2022» («ИС & ИТ-2022», «IS&IT'22»). Научное издание. – Таганрог: Изд-во Ступина С.А., Т.1. – 2022. – С. 90-97.

11. Мансур, А. М. Метод генерации векторов низкой размерности для представления текстовых документов / А.М. Мансур, Ж.Х. Мохаммад // Труды XIX всероссийской научной конференции молодых ученых, аспирантов и студентов «Информационные технологии, системный анализ и управление» (ИТСАУ-2021). – 2021. – С. 199-203.

12. Мансур, А. Развитие кластерного поиска документов на основе разработки методов векторизации текстов /А. Мансур, Ж. Мохаммад, Ю.А. Кравченко// Труды II научно-методической конференции НПР «Современные компьютерные технологии» (ИКТИБ ЮФУ). – 2021. – С. 28-31.

#### **Свидетельства о государственной регистрации программ для ЭВМ:**

1. Мансур, А. Программный модуль оптимизации работы классификатора при векторизации текста на основе биоэвристик / А. Мансур, Ж. Мохаммад, Ю.А. Кравченко, Д. Ю. Кравченко // Свидетельство регистрации программы для ЭВМ. – 12.12.2023. – № 2023687185.

2. Мансур, А. Программный модуль оптимизации извлечения ключевых слов при обработке лингвистической экспертной информации / А. Мансур, Ж. Мохаммад, Ю.А. Кравченко, К. Н. Владимировна // Свидетельство регистрации программы для ЭВМ. – 14.12.2023. – № 2023687372.

#### **Личный вклад автора.**

В опубликованных трудах:

[8, 1, 2, 4, 9, 10] – автором разработаны модель, методы и алгоритмы векторизации текстов для решения задач классификации и кластеризации. Также разработаны алгоритмы построения концептов на основе извлечения ключевых фраз при обработке и анализе текстов на естественном языке;

[7, 6, 11, 12] – автором разработаны методы и алгоритмы извлечения и фильтрации ключевых фраз на основе парсера при обработке и анализе текстов на естественном языке;

[3, 5, 14 - 17] – автором разработаны и реализованы различные методы представления текстов, методы оценки семантической близости для решения задачи устранения неоднозначности слов и концептов при обработке и анализе текстов на естественном языке.

При создании указанных программ для ЭВМ, использованы теоретические результаты, метод, модель и алгоритмы, полученные лично автором.

Автореферат  
Подписано в печать «17» апреля 2025 г.  
Бумага офсетная. Печать офсетная. Формат 60 × 84 1/16.  
Усл. печ. л. 1,375. Уч.-изд. л. 1,1. Заказ № 53. Тираж 100 экз.

---

Отпечатано в центре услуг «Караван»  
347922, Ростовская область, г. Таганрог, пер. Некрасовский, 63,  
тел. 8 (928) 600-80-00