

На правах рукописи



Секерин Алексей Викторович

**МЕТОДЫ И АЛГОРИТМЫ АВТОМАТИЧЕСКОЙ
ИДЕНТИФИКАЦИИ И КЛАССИФИКАЦИИ ИДЕОЛОГЕМ
В РУССКОЯЗЫЧНЫХ ОБЩЕСТВЕННО-ПОЛИТИЧЕСКИХ
ТЕКСТАХ**

2.3.8. Информатика и информационные процессы

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата технических наук

Курск – 2026

Работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Курский государственный университет».

Научный
руководитель

Кудинов Виталий Алексеевич,
доктор педагогических наук, профессор

Официальные
оппоненты:

Григорьев Сергей Георгиевич,
доктор технических наук, профессор, профессор
департамента информатики, управления и
технологий ГАОУ ВО «Московский городской
педагогический университет»

Романов Александр Сергеевич,
доктор технических наук, доцент, доцент
кафедры комплексной информационной
безопасности электронно-вычислительных систем
ФГАОУ ВО «Томский государственный
университет систем управления и
радиоэлектроники»

Ведущая
организация:

ФГАОУ ВО «Национальный исследовательский
университет «Высшая школа экономики»,
г. Москва

Защита состоится «24» сентября 2026 г. в 15 час. 15 мин. на заседании диссертационного совета 24.2.415.06, созданного на базе ТУСУРа, по адресу: 634050, г. Томск, пр. Ленина, 40, ауд. 201.

С диссертацией можно ознакомиться в библиотеке ТУСУРа по адресу: 634045, г. Томск, ул. Красноармейская, 146 и на официальном сайте ТУСУРа: <https://postgraduate.tusur.ru/urls/rq297kts>

Автореферат разослан «__» _____ 2026 г.

Ученый секретарь
диссертационного совета



Костюченко Евгений Юрьевич

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы. Современное информационное противоборство характеризуется активным использованием массовых коммуникационных практик через специфическое применение языковых единиц, направленных на формирование у определенных групп людей (географических, конфессиональных, политических, экономических) заданного мировоззрения, которое включает в себя систему взглядов на мир, роль человека в нем, а также соответствующие идеалы и убеждения. Одной из разновидностей таких единиц, действующих на эмоционально-подсознательном уровне, являются идеологемы, отражающие идеологии как системы идей и ценностей в когнитивном пространстве. Вопрос их функционирования служит предметом исследования в сфере лингвистики, политических и компьютерных наук, теории информации и информационных процессов.

Лингвистическому анализу функционирования идеологем русскоязычного дискурса в настоящее время посвящены исследования А. В. Таскаевой, Т. В. Романовой, М. Н. Ляшевой, С. Е. Тимофеева, И. Н. Зупинниковой, Д. В. Морозова, Н. К. Радиной. Применяемые учеными для идентификации и классификации идеологем инструменты компьютерной обработки данных носят вспомогательный характер. Обеспечивается автоматический сбор наборов текстовых данных по заданным тематикам на основе ключевых слов, методы машинного обучения используются на этапах предварительной обработки анализируемых текстов (токенизация и лемматизация), статистические методы – для выявления контекстуальных идеологем на ограниченном наборе текстов. Основные этапы процесса идентификации идеологем русскоязычных текстов осуществляются экспертами, что накладывает ресурсные и временные ограничения.

Аспекты разработки интеллектуальных информационных систем классификации информационных сообщений по идеологическому принципу рассмотрены в работах И. В. Машечкина, М. И. Петровского, R. Baly, E. Crothers, W. Chen, M. A. Walker. Однако использование нейросетевых технологий сводится к решению задач классификации данных на уровне текста для ограниченного домена и числа классов: определение экстремистской направленности текста или партийной принадлежности автора текста в четко

разделимом информационно-политическом поле таких государств, как США и Великобритания. Данный подход не может быть адаптирован для классификации идеологием постсоветского пространства, что подтверждается в работах российских исследователей.

Необходимость анализа растущего ежедневно количества публикуемых новостных материалов в сети Интернет в потоковом режиме для противодействия деструктивным ценностным установкам и обеспечения информационно-психологической безопасности населения и недостаточная автоматизация процесса обнаружения идеологием как элементов, отражающих политические идеологии и религиозные учения, свидетельствуют о необходимости разработки методов и алгоритмов автоматической идентификации и классификации идеологием русскоязычных общественно-политических текстов, что определяет актуальность и востребованность диссертационного исследования.

Объектом исследования являются процессы идентификации и классификации идеологием в русскоязычных общественно-политических текстах.

Предмет исследования – методы и алгоритмы автоматической идентификации и классификации идеологием в русскоязычных общественно-политических текстах.

Цель исследования – автоматизация процесса идентификации и классификации идеологием в русскоязычных общественно-политических текстах за счёт применения нейросетевых технологий.

Научная задача исследования – совершенствование методов и разработка алгоритмов для автоматической идентификации и классификации идеологием в русскоязычных общественно-политических текстах с применением нейросетевых технологий.

Для достижения поставленной цели сформулированы следующие **задачи исследования**:

1. Провести анализ методов идентификации и классификации идеологием русскоязычных новостных текстов и на его основе определить этапы, подлежащие автоматизации.

2. Проанализировать нейросетевые методы обработки текста и модифицировать их для генерации религиозно-политических ассоциаций и идентификации идеологием текста.

3. Разработать алгоритм классификации религиозно-политических понятий русскоязычного дискурса.

4. Разработать интеллектуальную систему идентификации и классификации идеологем, основанную на предложенных методах обработки информации.

5. Провести оценку эффективности применения разработанного математического и программного обеспечения.

Методология и методы исследования. В диссертационной работе применяются методы машинного обучения, включая методы обработки естественных языков и компьютерной лингвистики, а также положения теории множеств и графов. При реализации разработанных методов и алгоритмов используются библиотеки на языках Python и JavaScript.

В ходе работы получены результаты, обладающие **научной новизной**:

1. Предложен способ использования нейросетевого метода преобразования текстовых последовательностей для построения религиозно-политических ассоциаций русскоязычного дискурса, и алгоритм, реализующий его моделью машинного обучения преобразования последовательности.

2. Усовершенствован метод классификации идеологем на основе анализа ассоциативных обобщающих цепочек и реализован алгоритм, обеспечивающий формирование ассоциативных религиозно-политических словарей с использованием векторных представлений слов.

3. Разработана интеллектуальная система идентификации и классификации идеологем, построенная на основе предложенных методов и алгоритмов, отличающаяся от существующих интеграцией совокупности русскоязычных моделей машинного обучения для выявления прямых и контекстуальных идеологем и обеспечивающая определение целевых фрагментов текста без необходимости формирования тематических наборов документов.

Теоретическая значимость исследования заключается в дальнейшем развитии нейросетевых методов обработки русскоязычных политических текстов, а также представлении процесса идентификации и классификации идеологем в виде алгоритма, реализуемого ансамблем моделей машинного обучения.

Практическая значимость работы заключается в разработке программного обеспечения, реализующего предлагаемые методы и алгоритмы обработки информации для автоматизации процесса идентификации и классификации идеологем русскоязычных

общественно-политических текстов для снижения зависимости решения данной задачи от экспертной оценки.

Результаты диссертационного исследования приняты к использованию в АНО «Институт развития цифровых и социокультурных компетенций», КРОО «Объединенный центр «МОНОЛИТ» и АНО «Ресурсный центр добровольчества Курской области», что подтверждается актом и справками о внедрении.

Положения, выносимые на защиту. На защиту выносятся следующие новые научные результаты:

1. Использование нейросетевого метода преобразования текстовых последовательностей для генерации религиозно-политических ассоциаций текста обеспечивает идентификацию русскоязычных новостных публикаций, содержащих фрагменты текста, выступающие в роли идеологом, а также классификацию текстов в открытом множестве классов, в отличие от существующих подходов к классификации. Реализующая его модель машинного обучения достигла значения 0,634 по метрике METEOR.

2. Применение метода классификации идеологом на основе анализа ассоциативных обобщающих цепочек, отличающегося от существующих использованием моделей векторных представлений слов, обеспечивает с точностью 76,8% отнесение идеологемы (набора идеологом) к идеологии или религии, исходя из контекста употребления и общей идеологической направленности текста.

3. Применение нейросетевых моделей выявления целей мнений, целевого анализа настроений и модифицированного структурного предсказания обеспечивают идентификацию контекстуальных идеологом, которые не могут быть обнаружены моделями выявления именованных сущностей категорий «идеология» и «религия», повышая значение F1-меры идентификации идеологом русскоязычного общественно-политического текста до 76,4%.

Соответствие диссертации паспорту научной специальности. Диссертационная работа соответствует паспорту специальности 2.3.8. «Информатика и информационные процессы» по следующим областям исследования:

п. 5 «Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации для представления в

базах данных и организации интерфейсов информационных систем с пользователями»;

п. 13 «Разработка и применение методов распознавания образов, кластерного анализа, нейросетевых и нечетких технологий, решающих правил, мягких вычислений при анализе разнородной информации в базах данных»;

п. 16 «Автоматизированные информационные системы, ресурсы и технологии по областям применения (научные, технические, экономические, образовательные, гуманитарные сферы деятельности), форматам обрабатываемой, хранимой информации. Системы принятия групповых решений, системы проектирования объектов и процессов, экспертные системы и др».

Достоверность результатов исследования подтверждается корректным применением алгоритмов, программного обеспечения, экспериментальной базой, а также результатами апробации в рамках отечественных и международных конференций.

Апробация результатов диссертации. Материалы диссертационного исследования обсуждались на следующих научно-технических конференциях:

1. III Всероссийская научно-техническая конференция «Интеллектуальные информационные системы: теория и практика» (г. Курск, 2022 г.);

2. III Региональная научная конференция молодых ученых «Актуальные проблемы современной науки: исторические, философские, методологические аспекты» (г. Курск, 2023 г.);

3. Научно-аналитический форум «Цифровое право: технологическая, юридическая и этическая нормативность в условиях функционирования цифровой среды» (г. Новосибирск, 2023 г.);

4. IV Всероссийская научно-техническая конференция «Интеллектуальные информационные системы: теория и практика» (г. Курск, 2023 г.);

5. X Всероссийская научная конференция молодых ученых «Наука и инновации XXI века» (г. Сургут, 2023 г.);

6. V Международная научно-техническая конференция «Интеллектуальные информационные системы: теория и практика» (г. Курск, 2024 г.);

7. XXII Всероссийская конференция молодых ученых, аспирантов и студентов «Информационные технологии, системный анализ и управление» ИТСАУ-2024 (г. Таганрог, 2024 г.).

Публикации. По теме исследования опубликовано 13 работ, в числе которых 3 статьи в рецензируемых изданиях из перечня ВАК РФ, получено 3 свидетельства о государственной регистрации программы ЭВМ, 1 свидетельство о государственной регистрации базы данных, 6 работ опубликовано в материалах научных конференций.

Личный вклад соискателя. Все основные научные результаты исследования получены автором лично или при его непосредственном участии. В работах, написанных совместно с научным руководителем, автором конкретизирована постановка целей и задач работ, осуществлен поиск и формирование обучающих наборов данных, разработаны алгоритмы и реализующее их программное обеспечение, проведены экспериментальные исследования и сформулированы выводы.

Структура и объем диссертации. Диссертация состоит из введения, трёх глав, заключения, списка сокращений и условных обозначений, списка литературы и приложений. Работа изложена на 148 страницах основного текста, включающего 44 рисунка, 24 таблицы, список литературных источников из 148 наименований, а также 3 приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертационной работы, охарактеризована степень её разработанности, сформулированы цели и задачи исследования, научная новизна, теоретическая и практическая значимость, методы диссертационного исследования, выносимые на защиту положения и степень достоверности результатов.

В **первой главе** проведен подробный обзор научной литературы и исследованы существующие подходы к определению понятия «идеологема», осуществлен анализ методов и алгоритмов идентификации и классификации идеологем и определены этапы процесса, подлежащие автоматизации.

В рамках настоящего диссертационного исследования под идеологемой понимается лексико-фразеологическая единица (слово или словосочетание), отражающая политическую или религиозную идеологию и имеющая высокую степень оценочности.

Проведенный анализ показал, что в настоящий момент существуют различные подходы к применению ЭВМ для решения задачи классификации идеологических категорий в текстах. В отечественной лингвистике компьютерный анализ, главным образом применение статистических методов, является этапом обработки данных, предвещающим экспертный анализ. Действия исследователя, выполняющего вручную сравнение словарного лексического значения с контекстом его употребления в тексте и определяющего эмоциональную оценку слов, могут быть автоматизированы, в том числе при помощи методов машинного обучения.

Получивший широкое распространение за рубежом подход к классификации политических текстов на уровне документов на базе рекуррентных и сверточных нейронных сетей требует адаптации ввиду трудности определения строгого перечня идеологий и их ценностей в русскоязычном политическом пространстве. Существующие в настоящий момент подходы к формированию тезаурусных отношений и классификации текстов на основе графовых представлений могут быть адаптированы для решения задачи классификации идеологием.

Обнаружение контекстуальных идеологием может быть осуществлено путем адаптации нейросетевых подходов к выявлению авторской позиции и целевого анализа настроений в русскоязычных общественно-политических текстах, а также разрешения корелляции.

Таким образом, процесс идентификации и классификации идеологием может быть рассмотрен как совокупность применения нейросетевых методов обработки текста, что требует подготовки наборов текстовых данных, содержащих разметку фрагментов, отражающих политические идеологии и религиозные учения, а также разработки методов оценки экспериментальных результатов.

Во **второй главе** представлен процесс идентификации и классификации идеологием, предложены реализующие его этапы нейросетевые методы и метрики оценки их качества.

Определены следующие критерии, по которым понятие может быть идентифицировано как идеологема в русскоязычном тексте, ассоциирующемся читателем с политическими идеологиями и религиозными течениями:

1. Именная группа, содержащая прямое упоминание идеологии или религии.

2. Именная группа – цель мнения на уровне предложения, авторская позиция по отношению к которой позитивна или негативна.

3. Именная группа – наименования организаций и географических объектов и упоминания личностей, имеющие ассоциативную связь с идеологическими ассоциациями к тексту.

4. Именные группы, квазисинонимичные (корреферентные) понятиям из пунктов 1-3.

Процесс классификации идеологом базируется на двух подходах:

1. Поиск разбиения идеологом по принципу отношения к ним автора текста («сторонник» и «противник» ценностей идеологии).

2. Поиск общих поддеревьев в ассоциативном иерархическом дереве идеологических понятий, где более общая идеология является корневой для множества других.

Алгоритм идентификации и классификации идеологом, представленный на схеме (рисунок 1), включает следующую последовательность действий:

1. Предварительная обработка текста (сегментация и токенизация).

2. Генерация религиозно-политических ассоциаций текста. Если список идеологом пуст, выполнение алгоритма прекращается.

3. Выявление прямых упоминаний идеологом (термины типа «идеология», «религия» и «события»).

4. Поиск наименований – стимулов религиозно-политических ассоциаций, обнаруженных на этапе 2.

5. Поиск позитивных и негативных целей мнений.

6. Выявление корреферентов, найденных на этапах 2-5.

7. Определение авторской позиции по отношению к идеологам – эмотивная классификация.

8. Классификация идеологом на основе анализа ассоциативных обобщающих цепочек.

Для каждой модели машинного обучения, реализующей шаги алгоритма, с учетом итогов экспертной оценки должно быть определено пороговое значение вероятности предсказания $P_M(pred)$, учитываемое при формировании вывода модели:

$$P_M(pred) = \max(P(pred)) \forall P(pred), S_{exp} \rightarrow 1 \wedge S_M \rightarrow s_0, \quad (1)$$

где $P(pred)$ – пороговое значение вероятности предсказания, S_{exp} – значение метрики качества интеллектуальной системы по итогам экспертной оценки, S_M – значение метрики качества модели

машинного обучения на валидационном (тестовом) наборе данных, S_0 – оптимальное значение метрики качества модели машинного обучения.

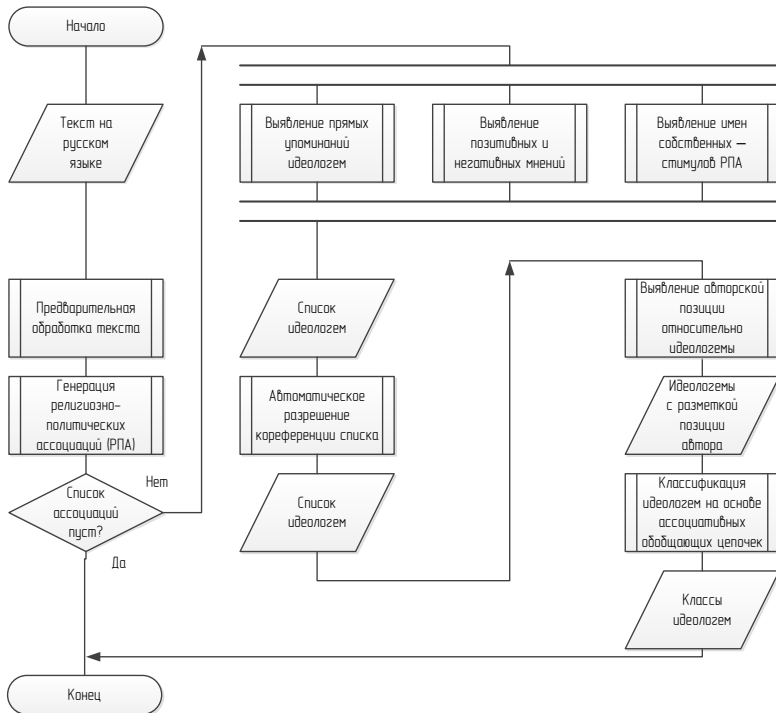


Рисунок 1 – Блок-схема алгоритма идентификации и классификации идеологем

Решение задачи поиска прямых упоминаний идеологем (понятий категорий «идеология» и «религия») и именных объектов может быть обеспечено моделями машинного обучения выявления именованных сущностей архитектуры трансформер, обученных на существующих наборах данных.

Автоматическое определение религиозно-политических ассоциаций русскоязычного текста можно рассматривать как задачу машинного перевода, где входной язык T является подмножеством русского языка, не содержащим слова и словосочетания, одним из лексических значений которых являются идеологические и религиозные категории, а выходной язык:

$$A = \{ R \setminus T \}, \quad (2)$$

где R – множество слов и словосочетаний русского языка.

Тогда решение задачи может быть обеспечено посредством модификации нейросетевых методов преобразования текстовых последовательностей и реализовано моделями архитектуры трансформер. При формировании обучающего набора данных для каждого текста, разбитого на группы по 5, в выходную последовательность изымаются лемматизированные прямые упоминания идеологий и религий, а также порождаемые ими высокочастотные словарные ассоциации.

Задача автоматического разрешения кореференции может также быть решена моделями преобразования последовательностей с использованием структурного предсказания. Введем символы дополненного русского языка

$$\aleph = \{ [**, **], ||, \& \& \}, \quad (3)$$

где $[**$ – начало ссылки на референт, $**]$ – окончание ссылки, $||$ – начало референта, $\&\&$ – объединение референтов.

Введем следующие действия, порождающие модифицированное структурное предсказание для разрешения кореференции:

$$A = \{ stc, endc, copy \}, W = \begin{cases} w_1 \dots \& w_n, n > 1 \\ w_1, n = 1 \end{cases}, \quad (4)$$

$$stc := w \rightarrow [**w, endc := w \rightarrow w || W**, copy := w \rightarrow w,$$

где w – слово текста, W – список референтов слова w , stc – операция создания начала ссылки на референт, $endc$ – операция создания конца ссылки на референт, $copy$ – операция копирования слова.

Выявление мнений в общественно-политическом тексте может быть реализовано как двухэтапный алгоритм, включающий извлечение именных сущностей – целей мнений – и объектно-ориентированный анализ настроений, заключающийся в определении авторской позиции по отношению к целевой сущности. Решающий данную задачу для русскоязычного текста алгоритм включает следующие этапы:

1. Разбиение текста на предложения.
2. Извлечение целей мнения посредством модели машинного обучения, обеспечивающего пословную разметку.

3. Для каждой цели формирование цель-ориентированных предложений, формат которых выражается следующими формулами:

$$Sent_{TEL} = (l_0, \dots, l_n \mid l \in \{O, B-TARGET, I-TARGET\}), \quad (5)$$

$$\{Sent, Sent_{TEL}\} \rightarrow [TE_0 <ots> Sent, \dots, TE_k <ots> Sent], \quad (6)$$

где $Sent$ – исследуемое предложение, $Sent_{TEL}$ – кортеж меток токенов предложения, TE – цель мнения, $<ots>$ – специальный токен-разделитель.

4. Для каждого предложения массива модель машинного обучения целевого анализа осуществляет его классификацию одной из трех меток, соответствующих позитивной, негативной или нейтральной оценке относительно цели мнения.

В иерархический идеологический словарь $DictH$ включаются понятия, отобранные с помощью нейросетевой модели выявления именованных сущностей религиозно-политического характера, в открытых онтологиях, имеющих иерархическую разметку отношений для понятий категорий «политическая идеология» или «религия». Ввиду неоднозначности экспертной оценки анализу подлежат только связи между нарицательными понятиями. Для обогащения новыми связями предлагается алгоритм обнаружения идеологических гиперонимов в словаре векторной модели представления слов.

Введем следующие определения: w – исследуемое понятие, s_w^n – n -ый сосед понятия w в векторном пространстве слов языковой модели, S_1 – множество соседей первого порядка понятия w ,

S_2 – множество соседей второго порядка понятия w , такие, что $S_1 \cap S_2 \neq \{\}$, $Mprox(w_1, w_2)$ – мера сходства векторов двух понятий, $POS(w)$ – часть речи понятия ($noun$ – имя существительное, $propn$ – имя собственное), $Idea(w)$ – понятие является идеологемой по данным нейросетевой модели, реализующей выделения именованных сущностей, или словарю.

Тогда множество потенциальных идеологических гиперонимов H_w понятия w может быть определено методом голосования посредством выбора соседей второго порядка на основе меры сходства векторов (косинусное сходство):

$$S_1 = \{s_w^k \mid Mprox(w, s_w^k) \rightarrow 1, k \in \{1, 2, \dots, 50\}\}, \quad (7)$$

$$S_2 = \bigcup_{k=1}^{50} \left\{ s_{S_w^k}^n \mid Mprox(s_w^k, s_{S_w^k}^n) \rightarrow 1 \wedge w \neq s_{S_w^k}^n, n \in \{1, 2, \dots, 150\} \right\}, \quad (8)$$

$$Vote(s_2^m) = Vote(s_{S_w^k}^n) = \sum_{s_{S_w^k}^n \in S_2, Idea(s_w^k) = True, POS(s_w^k) \in \{propn, noun\}} (Mprox(s_w^k, s_{S_w^k}^n)), \quad (9)$$

$$Trh(Vote(s_2^m), w) = \begin{cases} 1, & \text{если } (Vote(s_2^m) \geq 1 \wedge POS(w) = noun) \vee \\ & (Vote(s_2^m) > 0 \wedge POS(w) = propn) \\ 0, & \text{иначе} \end{cases}, \quad (10)$$

$$Votes(w) = \left\{ \begin{array}{l} V_p \mid V_t \geq V_{t+1}, V_t \geq \frac{V_1}{2}, p \in \{1, 2, 3\}, \\ V \in \left\{ Vote(s_2^m) \mid Trh(Vote(s_2^m), w), \right. \\ \left. m \in \{1, 2, \dots, x\}, 150 \leq x \leq 7500 \right\}, t \subset m \end{array} \right\}, \quad (11)$$

$$H_w = \{ s_2^m \mid Vote(s_2^m) \in Votes(w) \}, \quad (12)$$

где $Vote(s_2^m)$ – функция подсчета голосов соседей первого порядка исследуемого понятия w в пользу соседа второго порядка $s_2^m \in S_2$ с целью поиска элементов множества S_2 , ближайших для наибольшего числа элементов множества соседей первого порядка S_1 ; $Trh(Vote(s_2^m), w)$ – функция порога голосования, отсеивающая наиболее удаленные элементы S_2 ; $Votes(w)$ – множество результатов голосования за 3 победителей, удовлетворяющих условиям: имя существительное, не являющееся именем собственным, входящее в словарный список идеологом, значение $Vote(s_2^m)$ не менее половины максимального.

Гипо-гиперонимические отношения, построенные с помощью алгоритма на основе векторных моделей, а также найденные в онтологических базах знаний, могут быть представлены в виде ориентированного графа G , вершинами которого являются термины словаря векторной модели, а ребра имеют вид:

$$edge(w_1, w_2) = \begin{cases} (w_1, h_{w_1}) & \text{если } w_1 \notin DictH \\ (w_1, h_{w_1}^{DictH}) & \text{если } w_1 \in DictH \end{cases}, \quad (13)$$

где $DictH$ – иерархический словарь идеологом, h_{w_1} – значение гиперонима для понятия w_1 , полученное по итогам голосования, $h_{w_1}^{DictH}$ – словарное значение гиперонима для понятия w_1 .

Тогда под ассоциативной обобщающей цепочкой идеологемы w будем понимать любой путь l графа G , начинающийся от вершины w , удовлетворяющий следующему условию:

$$l \in L(w_0) = (w_0, \dots, w_k \mid 0 < k < 5, \exists w_k \in DictH), \quad (14)$$

Наименование идеологии является гиперонимом идеологемы, поэтому классификация множества идеологом W сводится к поиску такой вершины $h_U \in H$ графа G , что количество ассоциативных путей множества L из вершин $\{w_0, \dots, w_z\}$, проходящих через вершину w_U , максимально, при этом длина ассоциативной цепочки l минимальна.

При построении ассоциативных цепочек для предотвращения построения синонимичных рядов приоритет отдается обобщению из онтологий.

При классификации идеологом множества реализуется следующий алгоритм.

1. Осуществляется поиск понятия в графе, а в случае отсутствия среди 10 его ближайших соседей – в векторной модели.

2. Для каждого отобранного понятия осуществляется формирование не более 35 ассоциативных обобщающих цепочек. На данном этапе могут быть наложены дополнительные условия, если требуется классифицировать понятия относительно отдельной идеологии, а затем осуществляется поиск общей вершины на кластеры, где меткой кластера является общая вершина.

Поскольку алгоритм стремится найти общее понятие для всех элементов подаваемого на вход множества, в ходе его работы обеспечивается разделение понятий, относящихся как минимум в одном из своих значений к разным идеологическим системам. Однако разделение близких понятий кластера может быть достигнуто при помощи поиска общих подцепочек, при этом меткой подкластера может выступать первый элемент общей обобщающей ассоциативной подцепочки.

Третья глава посвящена разработке интеллектуальной системы, реализующей алгоритм автоматической идентификации и

классификации идеологем на основе модифицированных нейросетевых методов.

Клиент-серверное приложение обеспечивает обработку текстовых данных ансамблем моделей машинного обучения на высокопроизводительной ЭВМ и взаимодействие пользователя (рисунок 2) с помощью браузерного расширения, в том числе выбор данных и отображение аналитического отчёта (рисунок 3).

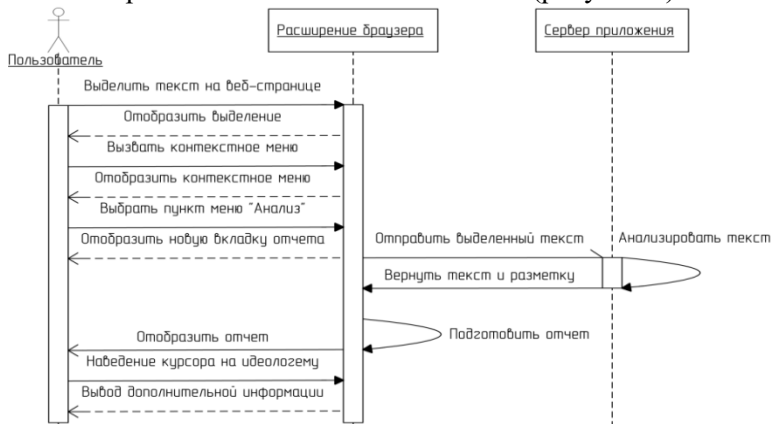


Рисунок 2 – Схема взаимодействия пользователя с приложением

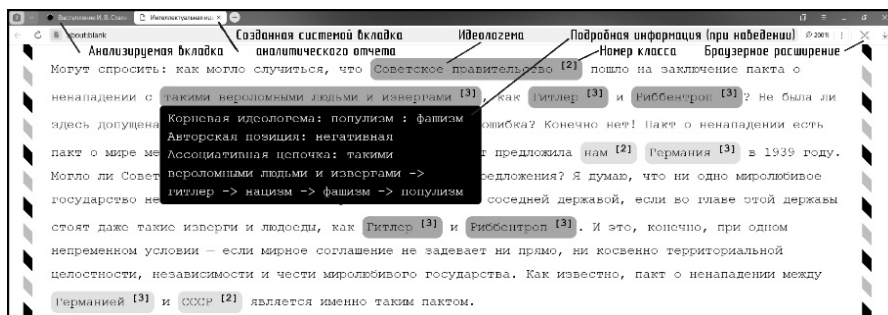


Рисунок 3 – Внешний вид отчета браузерного расширения

Проведено экспериментальное исследование характеристик разработанных моделей машинного обучения выявления понятий категорий «идеология» и «религия» (таблица 1), генерации религиозно-политических ассоциаций (таблица 2), разрешения кореференции (таблица 3), определения целей мнений (таблица 4) и целевого анализа настроений (таблица 5).

Таблица 1 – Параметры и оценка моделей выявления категорий «идеология» и «религия» (размер батча – 16)

Базовая модель	Шаг обучения	Ошибка обучения		Precision	Recall	F1 / seqEva l-F1
google-bert/bert-base-multilingual-uncased	$2 \cdot 10^{-5}$	Трансформер	5,17*	0,985	0,985	0,985
		CNN-NER	30,99*			
deepvk/DeBERTa-base	$4 \cdot 10^{-5}$	Трансформер	0,0186	0,795	0,729	0,780

* Используется логистическая функция ошибки (log_loss)

Таблица 2 – Параметры и оценка моделей генерации религиозно-политических ассоциаций (размер батча – 16)

Базовая модель	Параметры модели и токенизатора					Шаг обучения	Ошибка обучения	METEOR
	Словарь		Слоев кодировщика / декодировщика / размерность скрытого слоя	Функция активации	LORA			
	Токенизатор	N						
t5-small-ru	Byte-pair encoder (BPE)	20100	6 / 6 / 512	ReLU	-	3·10 ⁻⁵	0,3281	0,5897
ruT5-base		32101	12 / 12 / 768	ReLU	-	1·10 ⁻⁴	0,181	0,6344
ruT5-large		32101	24 / 24 / 1024	ReLU	α = 128 R = 64 d = 0,05	1·10 ⁻⁴	0,357	0,5865
FRED-T5-1.7B	Byte-level BPE	50364	24 / 24 / 1536	Gated-GeLU	α = 64 R = 16 d = 0,1	9·10 ⁻⁵	0,3	0,2391

Таблица 3 – Параметры и оценка моделей разрешения корефенции (размер батча – 1, число входных токенов – 1024, выходных – 1300)

Модель	Базовая модель	N словаря	Квантизация	LORA	Шаг обучения	LEA
ruT5-base-SP	ruT5-base	32 105	нет	-	$2 \cdot 10^{-5}$	0,687
FRED-T5-SP	FRED-T5-1.7B	50 368	8 бит	$\alpha = 32$ $R = 16$ $d = 0,05$	$5 \cdot 10^{-4}$	0,776

Таблица 4 – Параметры и оценка моделей извлечения целей мнений (длина входного текста – 512 токенов)

Базовая модель	Шаг обучения	Размер пакета	Число эпох	Ошибка обучения	F1-seqeval
DeBERTa-base	$2 \cdot 10^{-6}$	16	13	0,179	0,607
ruRoPEBert-e5-base-512	$4 \cdot 10^{-6}$	8	16	0,221	0,249
ruBert-large	$3 \cdot 10^{-6}$	16	5	0,219	0,551
ruElectra-large	$1 \cdot 10^{-6}$	16	6	0,159	0,539
ruRoBERTa-large	$1 \cdot 10^{-6}$	16	9	0,165	0,599
XLNet-V-base	$4 \cdot 10^{-6}$	8	6	0,158	0,607

Таблица 5 – Параметры и оценка моделей целевого анализа настроений (длина входного текста – 512 токенов)

Базовая модель	Шаг обучения	Размер пакета	Число эпох	Ошибка обучения	F1-macro	F1-micro
DeBERTa-base	$8 \cdot 10^{-6}$	16	3	0,588	0,741	0,789
ruRoPEBert-e5-base-512	$1 \cdot 10^{-5}$	16	3	0,641	0,726	0,779
ruBert-large	$2 \cdot 10^{-6}$	4	4	1,361	0,759	0,793
ruElectra-large	$8 \cdot 10^{-6}$	4	1	0,707	0,676	0,712
ruRoBERTa-large	$8 \cdot 10^{-6}$	4	1	0,613	0,773	0,804
XLNet-V-base	$8 \cdot 10^{-6}$	8	3	0,640	0,735	0,785

Алгоритм идентификации и классификации идеологием реализуется с помощью конвейера интеллектуальных агентов, решающих отдельные задачи анализа с помощью следующих ИНС:

- ruT5-base-idea-v21-finetuned-970000 – модель архитектуры T5 преобразования текстовых последовательность для выявления религиозно-политических ассоциаций к тексту;

- FRED-T5-1.7B-SP-merged – модель на базе FRED-T5, реализующая модифицированное структурное предсказание для разрешения кореференции;

- OTE-DeBERTa-finetuned-6994 – модель архитектуры DeBERTa, осуществляющая поиск целей мнений на уровне предложений;

- IDEA-DeBERTa-finetuned-1034 – реализует поиск прямых упоминаний фрагментов текста, относящихся к категориям «идеология» и «религия»;

- TSA-ruRoberta-large-finetuned-9858 – модель на базе большой модели ruRoberta для решения задачи определения авторской позиции относительно фрагмента текста на уровне предложений;

- Natasha (Slovnet) – дистиллированная в WordCNN-CRF версия модели BERT, используемая для морфологического разбора слов, синтаксического разбора предложений, а также выявления имен, наименований географических объектов и организаций;

- geowac lemmas_none fasttextskipgram_300_5_2020 – векторная языковая модель, обеспечивающая поиск ближайших соседей для понятий в эвристических алгоритмах.

Таблица 6 содержит экспериментально определенные пороговые значения для моделей машинного обучения (формула 1).

Для сегментации предложений и токенизации слов применяется подход, основанный на правилах, реализованный в библиотеке Razdel.

Для построения ассоциативных обобщающих цепочек используется сгенерированный словарь `wiki_idea_linear.lin`, включающий ассоциативные обобщающие идеологические цепочки длиной до 10 термов для 25914 понятий.

Таблица 6 – Пороговые значения вероятности предсказания моделей

Наименование модели	Пороговое значение
IDEA-DeBERTa-finetuned-1034	0,8
OTE-DeBERTa-finetuned-6994	0,7
TSA-ruRoberta-large-finetuned-9858	0,9
ruT5-base-idea-v21-finetuned-970000	-*
FRED-T5-1.7B-SP-merged	-*

* Значение не определяется из-за особенностей работы text-to-text моделей в конвейере библиотеки Huggingface

Определены метрики и проведена экспертная оценка качества интеллектуальной системы (таблица 7).

Таблица 7 – Результаты экспертной оценки (N=100 текстов, r=3 эксперта)

Целевая задача		Метрика	Значение
Идентификация идеологем		<i>F1</i>	0,764
Классификация идеологем	Поиск корневых идеологем	<i>precision</i>	0,785
	Определение авторской позиции	<i>precision</i>	0,933

Для определения возможности применения полученных экспертных оценок рассматривается статистическая мера согласия между оценщиками $AC2$, основанная на α -Криппендорфа.

Полученные результаты свидетельствуют о значительном согласии экспертов при оценке интеллектуальной системы по вопросу определения корневой (базовой) идеологемы ($AC2 = 0,630375$) и авторской позиции по отношению к идеологеме ($AC2 = 0,7999$). Средний уровень согласованности мнений оценщиков при определении того, является ли представленный фрагмент текста идеологемой ($AC2 = 0,458437$), обусловлен необходимостью оценки контекстуальных идеологем.

ЗАКЛЮЧЕНИЕ

Представленные в диссертационной работе методы и алгоритмы, которые основаны на применении нейросетевой настройки, позволяют автоматизировать работу экспертов в рамках процесса идентификации и классификации идеологем русскоязычных общественно-

политических текстов. В работе решены задачи исследования и получены следующие научные результаты:

1. Проведен анализ методов идентификации и классификации идеологем русскоязычных новостных текстов и определены этапы, подлежащие автоматизации. Существующие подходы на основе статистических методов требуют формирования набора публикаций узкой тематики, посвященных событию или явлению, для идентификации контекстуальных идеологем. Модели машинного обучения тематической классификации текстов из конечного набора идеологий и поиск токенов-идеологем, повлиявших на предсказание моделей, не могут быть применены с достаточной достоверностью в русскоязычном политическом домене из-за разницы между фактическими ценностями политических объединений и декларируемыми в программных документах идеологиями. Нейросетевые методы могут быть модифицированы для идентификации прямых и контекстуальных идеологем, а также построения ассоциативных цепочек. Определено понятие «идеологема», согласующееся с предложенными отечественными исследователями, и критерии для обнаружения посредством ЭВМ.

2. Проанализированы нейросетевые методы обработки текста и проведена их модификация для генерации религиозно-политических ассоциаций и идентификации идеологем текста, обеспечивающая с достаточной достоверностью выявление идеологической направленности текста за счет формирования списка идеологических понятий – ассоциаций как выходов модели преобразования последовательности, моделей выявления именованных сущностей категорий «идеология» и «религия», а также моделей извлечения мнений (обнаружения цели мнения и целевого анализа настроений) и преобразования последовательности для разрешения кореференции, реализующих поиск контекстуальных идеологем текста.

3. Разработан алгоритм классификации религиозно-политических понятий русскоязычного дискурса как их группировки на основе общности ассоциативных обобщающих цепочек религиозно-политических ассоциаций, сформированных на базе тезауруса общественно-политических понятий, и алгоритма поиска гиперонимов в модели векторных представлений слов и поиска общего рефлекс. Алгоритм обеспечивает лучшее качество формирования цепочек и требует меньшей вычислительной мощности по сравнению с большими языковыми моделями.

4. Разработана интеллектуальная система идентификации и классификации идеологом, основанная на предложенных методах обработки информации.

5. Проведена оценка эффективности применения разработанного математического и программного обеспечения. Применение совокупности моделей машинного обучения (генерации религиозно-политических ассоциаций, разрешения кореференции, выявления именованных сущностей категорий «идеология», «религия», «политические и религиозные события», обнаружения целей мнений и целевого анализа настроений) обеспечивает идентификацию более двух третей идеологом русскоязычного общественно-политического текста, в том числе контекстуальных идеологом, со скоростью, превышающей в 6,96 раз скорость работы эксперта-лингвиста, и с меньшими ресурсными затратами относительно больших языковых моделей. Модель целевого анализа настроений и алгоритм классификации понятий на основе их ассоциативных цепочек обеспечивают решение задачи классификации идеологом со средней долей правильных ответов на основе двух подходов более 85% .

Рекомендации. Результаты исследования рекомендуется использовать в организациях, занимающихся исследованием общественно-политических процессов, а также для разработки интеллектуальных систем изучения общественного мнения.

Перспективы дальнейшей разработки темы связаны с совершенствованием предложенных моделей и алгоритмов в соответствии с нарративным подходом к анализу политических процессов.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ИССЛЕДОВАНИЯ

Статьи в рецензируемых научных изданиях, рекомендованных ВАК РФ

1. Секерин, А. В. Алгоритм классификации идеологом на основе наборов цепочек религиозно-политических ассоциаций / А. В. Секерин, В. А. Кудинов // Вестник Донецкого национального университета. Серия Г: Технические науки. – 2024. – № 4. – С. 110-119.

2. Секерин, А. В. Двухэтапный алгоритм автоматизированного извлечения мнений в русскоязычных политических текстах /

А. В. Секерин, В. А. Кудинов // Вестник компьютерных и информационных технологий. – 2025. – Т. 22. – № 3 (249). – С. 50-56.

3. Секерин, А. В. Применение автоматизированных методов и алгоритмов для идентификации идеологием новостных текстов / А. В. Секерин, В. А. Кудинов // Экономика. Информатика. – 2024. – Т. 51. – № 1. – С. 201-210.

Публикации в сборниках научных трудов

4. Секерин, А. В. Идеологема как текстовая форма представления идеологии: анализ подходов / А. В. Секерин // Актуальные проблемы современной науки: исторические, философские, методологические аспекты: сборник статей 3-ей Региональной научной конференции молодых ученых, Курск, 15 мая 2023 года / Курский государственный университет. – Курск: Закрытое акционерное общество «Университетская книга», 2023. – С. 229-234.

5. Секерин, А. В. Интеллектуальный анализ идеологического поля: обзор решений и разработка теоретической модели / А. В. Секерин // Интеллектуальные информационные системы: Теория и практика: Сборник научных статей по материалам III Всероссийской конференции, Курск, 22–23 ноября 2022 года / Отв. редактор А.А. Халин. – Курск: Курский государственный университет, 2022. – С. 40-50.

6. Секерин, А. В. Исследование алгоритмов машинного обучения для выявления идеологических категорий текста / А. В. Секерин, В. А. Кудинов // Интеллектуальные информационные системы: теория и практика: Сборник научных статей по материалам IV Всероссийской с международным участием конференции, Курск, 21–23 ноября 2023 года. – Курск: Курский государственный университет, 2023. – С. 79-86.

7. Секерин, А. В. Модель и алгоритм формирования набора данных религиозно-идеологических ассоциаций русскоязычных новостных публикаций / А. В. Секерин // Наука и инновации XXI века: Материалы X Всероссийской конференции молодых ученых, Сургут, 15 декабря 2023 года. – Сургут: Сургутский государственный университет, 2024. – С. 88-92.

8. Секерин, А. В. Нейросетевое преобразование последовательностей для построения русскоязычных религиозно-политических ассоциаций / А. В. Секерин // Информационные технологии, системный анализ и управление (ИТСАУ-2024): сборник трудов XXII Всероссийской научной конференции молодых ученых,

аспирантов и студентов (Таганрог, 5–7 декабря 2024 г.). – Таганрог: ДиректСайнс (ИП Шкуркин Д.В.), 2024. – Т. 2. – С. 381-384.

9. Секерин, А. В. Модифицированное структурное предсказание для разрешения кореференции в русскоязычных текстах / А. В. Секерин, В. А. Кудинов // Интеллектуальные информационные системы: теория и практика : Сборник научных статей по материалам V Международной конференции, Курск, 19–21 ноября 2024 года. – Курск: Курский государственный университет, 2024. – С. 26-29.

Свидетельства о государственной регистрации программ для ЭВМ и баз данных

10. Свидетельство о государственной регистрации базы данных № 2024620009 Российская Федерация. Идеологические и религиозные ассоциации русскоязычных новостных текстов: № 2023625014: заявл. 22.12.2023: опубл. 09.01.2024 / А. В. Секерин.

11. Свидетельство о государственной регистрации программы для ЭВМ № 2024691104 Российская Федерация. Интеллектуальная идентификация идеологем. Браузерное расширение: № 2024690414: заявл. 14.12.2024: опубл. 19.12.2024 / А. В. Секерин.

12. Свидетельство о государственной регистрации программы для ЭВМ № 2025613511 Российская Федерация. Интеллектуальная идентификация идеологем. Серверное приложение: № 2025611924: заявл. 09.02.2025: опубл. 12.02.2025 / А. В. Секерин.

13. Свидетельство о государственной регистрации программы для ЭВМ № 2024666800 Российская Федерация. КИПАРИС: Классификатор идейно-политический ассоциативный русскоязычный интеллектуальный семантический: № 2024665401: заявл. 01.07.2024: опубл. 17.07.2024 / А. В. Секерин.