МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ
Кафедра информационных систем управления

КОРДИЯК Татьяна Дмитриевна

КЛАССИФИКАЦИЯ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА И ЕЁ ПРИЛОЖЕНИЯ

Дипломная работа

Научный руководитель:
старший преподаватель
Н.К. Рубашко

Допущена к защите
«____»_____20___г.
Заведующий кафедрой информационных систем управления, доктор технических наук, доцент А.М. Недзьведь

Минск, 2025

ОГЛАВЛЕНИЕ

ПЕРЕЧЕНЬ УСЛОВНЫХ ОБОЗНАЧЕНИЙ, СИМВОЛОВ И ТЕРМИНОВ

BERT (Bidirectional Encoder Representations from Transformers)	– нейронная сеть, разработанная Google для улучшения понимания ЕЯ
LDA (Latent Dirichlet Allocation)	– метод тематического моделирования
MAE	– средняя абсолютная ошибка
NLP (Natural Language Processing)	– обработка текстов естественного языка
REST API (Representational State Transfer Application Programming Interface)	– архитектурный стиль взаимодействия между клиентом и сервером через HTTP
RMSE	– среднеквадратичная ошибка
RoBERTa	– вариант модели BERT, которая была разработана исследователями Facebook AI
TF-IDF (Term Frequency-Inverse Document Frequency)	– статистическая мера, используемая для оценки важности слова в контексте документа или набора документов
ЕЯ	– естественный язык
Лемматизация	– нахождение основной формы (леммы) каждого слова в предложении
Мешок слов (Bag-of-Words, BoW)	– стратегия обработки ЕЯ для преобразования текстового документа в числа, которые могут быть использованы компьютерной программой. BoW часто реализуется в виде словаря, каждый ключ которого задается словом, а каждое значение - количеством повторений этого слова
ПО	– программное обеспечение
Токенизация	– процесс подготовки текстовых данных в NLP, включающий в себя разбиение фрагмента текста на более мелкие, значимые единицы, называемые токенами
Эмбеддинг (от англ. embedding)	– способ представления слов или фраз в виде числовых векторов, которые используются компьютерными моделями для обработки ЕЯ

РЕФЕРАТ

Структура и объём дипломной работы

62 страниц, 10 рисунков, 1 приложение, 9 источников

Ключевые слова: АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА ТЕКСТОВ, КЛАСТЕРИЗАЦИЯ ОТВЕТОВ, МАШИННОЕ ОБУЧЕНИЕ, СПАМ-ФИЛЬТРАЦИЯ, СЕМАНТИЧЕСКИЙ АНАЛИЗ, DBSCAN, ОЦЕНКА ИНФОРМАТИВНОСТИ, NLP.

Текст реферата

Объект исследования — текстовые данные, получаемые в процессе проведения массовых анкетных и опросных исследований в образовательной, социологической и маркетинговой областях.

Предмет исследования — методы и алгоритмы автоматизированной обработки открытых текстовых ответов в результатах опросов, включая фильтрацию спамовых сообщений, тематическую кластеризацию и автоматическую оценку качества ответов.

Цели исследования — разработка, программная реализация и экспериментальная оценка эффективности системы, позволяющей автоматически фильтровать нерелевантные и спамовые ответы, группировать сообщения по смысловой близости и производить предварительную автоматическую оценку их качества.

Методы исследования — сравнительно-аналитический обзор современных подходов, проектирование модульной архитектуры системы, реализация и тестирование алгоритмов фильтрации, кластеризации, автоматической оценки, а также проведение экспериментальных исследований с использованием размеченных тестовых данных и вычисление ключевых метрик качества.

Результатами являются разработанный программный комплекс, осуществляющий полную цепочку автоматической обработки анкетных текстовых данных; подтверждение высокой точности и эффективности автоматической фильтрации, тематической кластеризации и оценки качества текстов с использованием ансамблей алгоритмов и семантического анализа. Новизна работы заключается в интеграции современных методов машинного обучения, гибридных архитектур фильтрации и кластеризации, поддержке русского и белорусского языков, а также многоуровневой подготовке размеченной тестовой выборки для объективного сравнения методов. По результатам тестирования показано, что применение автоматических подходов позволяет снизить трудозатраты и существенно повысить качество и скорость анализа текстовой информации.

Достоверность материалов и результатов работы подтверждена экспериментальным тестированием на реальных и разнообразных по содержанию

корпусах открытых текстовых ответов, с многоуровневой ручной разметкой и объективной проверкой ключевых метрик качества (точность, полнота, F1, метрики кластеризации и автоматической оценки). Отдельные решения апробированы в нескольких сценариях применения, результаты воспроизводимы и подтверждают заявленные свойства системы.

Область возможного практического применения — система может быть внедрена в корпоративные и образовательные платформы, центры аналитики, HR-службы, исследовательские организации и любые структуры, обрабатывающие большие массивы открытых ответов респондентов в опросах, позволяя существенно повысить качество анализа и снизить трудозатраты на обработку текстовой информации.

РЭФЕРАТ

Структура і аб’ём дыпломнай працы

62 старонак, 10 малюнкаў, 1 дадатак, 9 крыніц

Ключавыя словы: АЎТАМАТЫЗАВАНАЯ АПРАЦОЎКА ТЭКСТАЎ, КЛАСТАРЫЗАЦЫЯ АДКАЗАЎ, МАШЫННАЕ НАВУЧАННЕ, СПАМ-ФІЛЬТРАВАННЕ, СЕМАНТЫЧНЫ АНАЛІЗ, DBSCAN, АЦЭНКА ІНФАРМАТЫЎНАСЦІ, NLP.

Змест працы

Аб'ект даследавання — тэкставыя дадзеныя, якія атрымлівае ў працэсе правядзення масавых анкетных і апытальных даследаванняў у адукацыйнай, сацыялагічнай і маркетынгавай абласцях.

Прадмет даследавання — метады і алгарытмы аўтаматызаванай апрацоўкі адкрытых тэкставых адказаў у выніках апытанняў, уключаючы фільтрацыю спамавых паведамленняў, тэматычную кластарызацыю і аўтаматычную ацэнку якасці адказаў.

Мэты даследавання — распрацоўка, праграмная рэалізацыя і эксперыментальная ацэнка эфектыўнасці сістэмы, якая дазваляе аўтаматычна фільтраваць нерэлевантныя і спамавыя адказы, групаваць паведамленні па сэнсавай блізкасці і вырабляць папярэднюю аўтаматычную ацэнку іх якасці.

Метады даследавання — параўнальна-аналітычны агляд сучасных падыходаў, праектаванне модульнай архітэктуры сістэмы, рэалізацыя і тэставанне алгарытмаў фільтрацыі, кластарызацыі, аўтаматычнай ацэнкі, а таксама правядзенне эксперыментальных даследаванняў з выкарыстаннем размечаных тэставых дадзеных і вылічэнне ключавых метрык якасці.

Вынікамі з'яўляюцца распрацаваны праграмны комплекс, які ажыццяўляе поўны ланцужок аўтаматычнай апрацоўкі анкетных тэкставых дадзеных; пацвярджэнне высокай дакладнасці і эфектыўнасці аўтаматычнай фільтрацыі, тэматычнай кластарызацыі і ацэнкі якасці тэкстаў з выкарыстаннем ансамбляў алгарытмаў і семантычнага аналізу. **Навізна** работы заключаецца ў інтэграцыі сучасных метадаў машыннага навучання, гібрыдных архітэктур фільтрацыі і кластарызацыі, падтрымцы рускай і беларускай моў, а таксама шматузроўневай падрыхтоўцы размечанай тэставай выбаркі для аб'ектыўнага параўнання метадаў. Па выніках тэставання паказана, што ўжыванне аўтаматычных падыходаў дазваляе знізіць працавыдаткі і істотна павысіць якасць і хуткасць аналізу тэкставай інфармацыі.

Дакладнасць матэрыялаў і вынікаў працы пацверджана эксперыментальным тэставаннем на рэальных і разнастайных па змесце карпусах адкрытых тэкставых адказаў, з шматузроўневай ручной разметкай і

аб'ектыўнай праверкай ключавых метрык якасці (дакладнасць, паўната, F1, метрыкі кластарызацыі і аўтаматычнай ацэнкі). Асобныя рашэнні апрабаваны ў некалькіх сцэнарах прымянення, вынікі Прайграваныя і пацвярджаюць заяўленыя ўласцівасці сістэмы.

Вобласць магчымага практычнага прымянення — сістэма можа быць укаранёна ў карпаратыўныя і адукацыйныя платформы, цэнтры аналітыкі, HR-службы, даследчыя арганізацыі і любыя структуры, якія апрацоўваюць вялікія масівы адкрытых адказаў рэспандэнтаў у апытаннях, дазваляючы істотна павысіць якасць аналізу і знізіць працавыдаткі на апрацоўку тэкставай інфармацыі.

SUMMARY

Structure and scope of the diploma work
62 pages, 10 figures, 1 appendix, 9 references
Keywords: AUTOMATED TEXT PROCESSING, RESPONSE CLUSTERING, MACHINE LEARNING, SPAM FILTERING, SEMANTIC ANALYSIS, DBSCAN, INFORMATIVENESS EVALUATION, NLP.
Summary text

The object of the study is text data obtained in the process of conducting mass questionnaire and survey research in the educational, sociological and marketing fields.

The subject of the study is methods and algorithms for automated processing of open text responses in survey results, including filtering of spam messages, thematic clustering and automatic assessment of the quality of responses.

The objectives of the study are the development, software implementation and experimental evaluation of the effectiveness of a system that allows for automatic filtering of irrelevant and spam responses, grouping messages by semantic similarity and performing a preliminary automatic assessment of their quality.

The methods of the study are a comparative analytical review of modern approaches, design of a modular system architecture, implementation and testing of filtering, clustering, automatic assessment algorithms, as well as conducting experimental studies using labeled test data and calculating key quality metrics.

The results are a developed software package that implements a full chain of automatic processing of questionnaire text data; confirmation of high accuracy and efficiency of automatic filtering, thematic clustering and text quality assessment using ensembles of algorithms and semantic analysis. **The novelty** of the work lies in the integration of modern methods of machine learning, hybrid filtering and clustering architectures, support for the Russian and Belarusian languages, as well as multilevel preparation of a marked test sample for objective comparison of methods. Based on the testing results, it is shown that the use of automatic approaches can reduce labor costs and significantly improve the quality and speed of text information analysis.

The reliability of the materials and results of the work is confirmed by experimental testing on real and diverse corpora of open text answers, with multilevel manual tagging and objective verification of key quality metrics (accuracy, completeness, F1, clustering metrics and automatic assessment). Individual solutions have been tested in several application scenarios, the results are reproducible and confirm the declared properties of the system.

Area of possible practical application: the system can be implemented in corporate and educational platforms, analytics centers, HR services, research

organizations and any structures that process large arrays of open responses from respondents in surveys, allowing for a significant improvement in the quality of analysis and a reduction in labor costs for processing text information.

ВВЕДЕНИЕ

Стремительное развитие информационных технологий и широкое внедрение цифровых платформ способствуют увеличению объёмов текстовой информации, требующей автоматизированной обработки. Особое место в этом контексте занимают тексты, генерируемые в процессе прохождения опросов, анкет и других инструментов обратной связи. Комплексная обработка таких данных, в особенности свободно сформулированных (открытых) ответов, представляет собой актуальную и практически значимую задачу для социологических, образовательных и маркетинговых исследований. Однако высокая вариативность, неоднородность и наличие неинформативных или спамовых сообщений существенно затрудняют анализ и интерпретацию результатов опросов вручную.

Актуальность выбранной темы обусловлена необходимостью создания эффективных автоматизированных решений, позволяющих значительно снизить трудозатраты на обработку текстовой информации, повысить объективность и ускорить получение итоговых результатов. В современных условиях острой конкуренции на рынке информационных услуг и стремления к повышению качества принимаемых управленческих решений возрастает потребность в инструментах, способных автоматически фильтровать нерелевантные ответы, группировать схожие по содержанию сообщения и давать первичную оценку их информативности.

Целью данной работы является разработка программного обеспечения для автоматизированной обработки результатов опросов, включающего следующие функциональные компоненты: выявление и фильтрация спамовых и бессмысленных ответов, кластеризация открытых текстовых сообщений по смысловой близости, а также автоматическая оценка качества и информативности ответов, где это возможно. В рамках работы предполагается решение следующих задач:

1. анализ существующих методов тестирования и автоматизированной обработки текстовых данных;
2. исследование современных подходов к фильтрации, кластеризации и оценке открытых ответов;
3. проектирование архитектуры и алгоритмов работы разрабатываемой системы;
4. экспериментальная проверка эффективности разработанного программного обеспечения на тестовых данных.

Хронологические рамки исследования определяются необходимостью анализа и обобщения современных отечественных и зарубежных разработок в области обработки естественного языка, а также актуальных на сегодняшний

день методов машинного обучения и их внедрения в прикладные решения, реализуемые с 2020 по 2024 годы. Такое ограничение позволит сконцентрироваться на наиболее актуальных, релевантных и практически применимых инструментах.

Необходимость проведения исследования по выбранной теме обоснована текущими проблемами заказчиков, осуществляющих массовые опросы населения, пользователей образовательных систем, компаний, собирающих обратную связь от клиентов, а также всеми, кто сталкивается с обработкой больших массивов открытых текстовых ответов. Разработка и внедрение эффективных автоматизированных инструментов позволит существенно оптимизировать бизнес-процессы, снизить издержки и повысить качество принимаемых решений.

Структура работы включает четыре главы. В первой главе дается теоретический обзор существующих подходов к классификации тестов естественного языка и обработки текстовых данных. Во второй главе анализируются основные задачи автоматизированной обработки текстовых ответов и методы их решения. В третьей главе рассматриваются вопросы проектирования и разработки программного обеспечения, в том числе архитектурные решения, используемые алгоритмы и инструменты. В четвертой главе приведены результаты экспериментальных исследований, произведена оценка эффективности внедрения разработанного программного продукта, рассчитаны основные экономические показатели.

ГЛАВА 1
ТЕОРЕТИЧЕСКИЕ ОСНОВЫ КЛАССИФИКАЦИИ ТЕКСТОВ ЕСТЕСТВЕННОГО ЯЗЫКА

1.1 Теоретико-методологические основы анализа и обработки текстов естественного языка

Анализ и обработка текстов ЕЯ — это междисциплинарная область, объединяющая лингвистику, психологию, информатику и педагогику. Эволюция методов в данной сфере отражает не только утилитарные задачи (оценка компетенций, обработка больших данных, автоматизация проверки), но и изменения в общем понимании природы языка и моделей его функционирования.

Зарубежная наука закладывает фундамент теоретических и методологических подходов к обработке языка в трудах Н. Хомского, который впервые описывает формальные принципы работы языковых механизмов, а также в работах Р. Ладо, где тест как инструмент контроля языковых знаний трактуется как часть педагогической парадигмы. В эти десятилетия акцент делался на структурную и формальную составляющую, то есть на фактологическую проверку знаний и навыков с помощью тестовых заданий.

С конца XX века методологическая база существенно расширяется за счёт психолингвистических и когнитивных направлений. Анализ текста начинает рассматриваться как диагностика не только лексико-грамматических, но и смысловых, коммуникативных и межкультурных компетенций. Это отражается в появлении комплексных моделей анализа, предусматривающих оценку способности к интерпретации, аргументации, порождению связного текста, что особенно важно при работе с открытыми текстовыми данными.

Стремительное развитие компьютерной лингвистики в 1990-е—2000-е годы стало толчком к автоматизации анализа текстов. За рубежом появляются специализированные метрики для машинного перевода и систем обработки естественного языка [6]. В это время возникает парадигма NLP (Natural Language Processing), где анализ текстов включает задачи морфологического, синтаксического, семантического и прагматического уровня. Автоматизация охватывает не только формальные тесты, но и интерпретацию свободных ответов, их группировку, поиск смысловых связей.

В Республике Беларусь уже в конце XX — начале XXI века возникает собственная школа исследований анализа текстов, отличающаяся вниманием к билингвизму, социокультурным и языковым особенностям. Белорусские исследователи отмечают уникальные вызовы, связанные с необходимостью двуязычной оценки, а также с адаптацией мировых методик к национальной языковой и образовательной среде. Современные работы отражают интерес к

автоматизации контроля и обработке текстов с использованием ИТ-решений, включая интеллектуальные системы для анализа свободных ответов.

В международной практике основной тренд заключается в стандартизации процедур анализа, развитии автоматизированных платформ для проверки письменных работ (например, TOEFL, IELTS, automated essay scoring). Зарубежные исследования подчеркивают значение объективности, масштабируемости и валидации результатов. При этом в Беларуси актуальна интеграция мировых стандартов с требованиями к сохранению экспертной оценки, гибкости тестовых заданий, учёта контекста и глубины анализа.

Характерной особенностью белорусской школы является развитие смешанных (гибридных) методик, сочетающих автоматический и ручной анализ, что позволяет учитывать нюансы языковых проявлений, ментальные и культурные особенности респондентов.

Анализ литературы свидетельствует о значительных достижениях в зарубежной автоматизации обработки текстовых данных, однако в белорусском контексте сохраняется ряд нерешённых вопросов. К ним относятся недостаточное распространение отечественных платформ, ограниченность полнофункциональных решений для автоматического анализа открытых ответов на белорусском и русском языках, а также низкая степень интеграции современных методов машинного обучения в практику массового лингвистического анализа.

1.2 Подходы и критерии классификации текстов естественного языка

Одним из ключевых вопросов в исследовании текстов естественного языка выступает задача их научной классификации — систематизации на основе целого ряда теоретических и практических критериев. С накоплением лингвистического, педагогического, культурологического и компьютерного опыта классификации текстов приобретали всё большую сложность и многоуровневость, что обусловило развитие различных подходов и школ.

Исторически становление теории классификации текстов связано с жанрово-стилевыми и функциональными традициями, отражёнными в трудах российских, белорусских и зарубежных исследователей. В отечественной лингвистике особое внимание уделялось вопросам жанрового членения, стилистического диапазона, а также тематических и структурных особенностей текстовых массивов.

Среди основных лингвистических классификаций традиционно выделяют:

– жанрово-стилевую классификацию, предусматривающую деление текстов на художественные, научные, официально-деловые, публицистические, разговорные и др.;

– тематическую классификацию, основанную на содержательном признаке (например, образовательные, технические, культурологические, юридические тексты);

– структурно-композиционную классификацию, охватывающую такие параметры, как наличие элементов повествования, описания, рассуждения, диалога, монолога.

В современных исследованиях данные классификации уточняются и расширяются с учётом прагматических, коммуникативных и семантических характеристик. Важным становится анализ текстов на уровне выразительных средств, структуре аргументации, способах реализации адресности и намерения говорящего. В белорусской лингвистике, учитывая специфику двуязычной среды и культурных контекстов, широко обсуждается вопрос пересечений между жанровыми и функциональными признаками, а также проблематика классификации смешанных и переходных форм текста.

Современные зарубежные школы чаще прибегают к универсализации: для практических задач автоматизации различают повествовательные, описательные, экспозиционные, аргументативные тексты, что облегчает разработку алгоритмов автоматической категоризации [5]. Тем не менее белорусская традиция нередко оказывается ближе к глубинному качественному анализу, где важны не только формальные признаки, но и социально-культурные, прагматические параметры.

Для повышения эффективности анализа и обеспечения применимости результатов к реальным задачам, активно разрабатываются классификации, ориентированные на прикладные потребности современной обработки текстов. В образовательных, социологических и информационных системах особенно востребованы критерии, позволяющие учитывать форму представления, формат ответа, степень структурированности и сложность текстового материала.

К числу наиболее значимых классификаций по целям обработки и анализа относятся:

– деление на тексты с закрытой и открытой структурой (ответы с жёстко заданной или свободной формулировкой);

– стратификация по длине (краткие реплики, развёрнутые ответы, эссе);

– дифференциация по семантической насыщенности и функциональной направленности (фактическая, экспрессивная, апеллятивная и др.);

– выделение текстов в зависимости от предметной сферы, жанровой гибридности и степени проявления интертекстуальности;

– классификация по языковому и культурному признаку.

В последние десятилетия в белорусских исследованиях фиксируется возрастание интереса к проблемам построения гибридных классификационных систем. Здесь проводится интеграция классических лингвистических, функциональных и прагматических основ с современными требованиями автоматизации, что находит отражение в педагогических, образовательных и ИТ-проектах. Такой синтез позволяет не только обеспечить точную структуру корпусов текстов, но и выстраивать эффективные алгоритмы для актуальных задач массовых опросов, электронного обучения, педагогической диагностики, интеллектуальных систем анализа текстовой информации.

Вместе с тем, в научном сообществе высказывается необходимость продолжения исследований по систематизации смешанных, маргинальных и переходных типов текстов, кросс-языковых и мультикультурных корпусов, а также дальнейшего уточнения критериев для динамично изменяющихся форм письменной и устной коммуникации. Развитие подходов к классификации текстов естественного языка способно обеспечить методическую основу для инновационных разработок в области автоматизированной обработки, формируя теоретическую и прикладную базу для реализации современных систем анализа текстовых данных в лингвистике, образовании, науке и практике управления.

1.3 Особенности классификации текстов в условиях многоязычия и социокультурного многообразия

Переходя от вопросов типологизации текстов в классическом лингвистическом дискурсе к рассмотрению современных реалий, трудно игнорировать роль многоязычия и социокультурных различий как значимых факторов дифференциации текстовых массивов. Именно взаимодействие языков и культурных традиций формирует уникальные ситуации коммуникации, отражающиеся в структуре, содержании и прагматике текстов, что обогащает арсенал критериев их классификации [5,6].

Многоязычное пространство создает особые условия для функционирования и развития текстов, как это особо заметно на примере Республики Беларусь, где формируется сложная языковая и культурная ткань повседневного и официального общения. В таких условиях традиционные жанрово-стилистические и функциональные критерии становятся лишь базой, на которую накладываются специфические параметры билингвизма и социокультурной идентичности. Присутствие сразу нескольких языковых кодов в одной коммуникативной среде неизбежно приводит к широкому распространению явлений смешения языков, кодовых переключений, калькирования, заимствований, появлению гибридных форм текстов и новых семантических нюансов.

Практика анализа текстов, созданных в условиях языкового взаимодействия, требует от исследователя не только владения средствами классической классификации, но и понимания лингвистических и экстралингвистических факторов, формирующих новые параметры для отнесения текстов к определённому виду или типу. Параллельно с этим развивается и корпусная лингвистика, правильно отражающая в своих подходах принципы репрезентативности, охвата различных языковых групп и учёт межъязыковых трансформаций текстов. Например, при формировании национальных корпусов белорусских текстов отмечается особое внимание к идентификации многоязычных контактов, гибких жанровых границ, а также отражению культурно-региональных и профессиональных различий в структурах текстовых данных.

Среди таких особенностей часто выделяют интеграцию этнических, региональных и субкультурных элементов, которые становятся значимыми признаками в процессе дифференциации текстов. Это может проявляться в лексическом наполнении, фразеологических оборотах, способах выражения эмоций, оценочных суждениях, структуре аргументации. Отдельного исследования заслуживают тексты, создаваемые в интернет-пространстве: блоги, социальные сети, мессенджеры — здесь коммуникация проходит на стыке языков и культурных кодов, а используемые форматы и жанры часто не имеют прямых аналогов в традиционных типологиях [5, 6].

В национальных и многоязычных моделях классификации текстов неизбежно возникает вопрос о балансе между формальным анализом структуры и содержательным осмыслением культурных и социальных функций текста. Белорусский лингвистический опыт подчеркивает значимость междисциплинарного подхода, сочетающего лингвистику с социологией, культурологией и психологией, что в теории позволит более тонко улавливать динамику жанровых сдвигов и прагматических изменений в условиях языковой и культурной гибридности.

Классификация в аналогичных условиях всё чаще базируется на включении параметров, отражающих специфику межъязыкового взаимодействия и социокультурных контекстов, будь то критерии доминантного языка, уровней смешения кодов, культурных аллюзий, жанровых инноваций или прагматических различий в целевых аудиториях. Важно отметить, что именно в таких ситуациях возрастает роль не только формальных признаков, но и содержательных, этнокультурных, дискурсивных характеристик. Необходимость использования описанного подхода приводит к пересмотру устоявшихся схематических классификаций и потребности в обновлении понятийного аппарата, способного адекватно описывать новые явления и смысловые константы в текстах, созданных на стыке языков и культур.

Такое расширение методологических рамок становится одной из ступеней на пути разработки эффективных систем анализа текстов, обеспечения репрезентативности текстовых коллекций в образовательных, медиакоммуникационных, управленческих и научных задачах. Развитие комплексных моделей классификации позволит установить соответствие результатам обработки специфики современной языковой и культурной среды, усилить практическое значение теоретических основ для всего спектра дисциплин, связанных с исследованиями текстов естественного языка.

1.4 Значение и роль теоретических основ классификации текстов для развития современных лингвистических и информационных технологий

Применение разработанных теоретических подходов к классификации текстов естественного языка приобретает особое значение на фоне стремительного роста объемов текстовых данных и усложнения задач их анализа. Сегодня успешность цифровых лингвистических платформ, образовательных и экспертных систем напрямую зависит от того, насколько последовательно и глубоко реализованы базовые идеи лингвистической классификации. В основе эффективных IT-решений для массовой обработки текстовой информации лежит способность точно дифференцировать тексты по жанрово-стилевым, тематическим, структурным, прагматическим и семантическим основаниям, а также учитывать эксплицитные и имплицитные признаки, отражающие специфику конкретных языковых и культурных сред.

В современных автоматизированных системах обработки больших текстовых массивов обращают внимание на вопрос точности фильтрации нерелевантной, вторичной или заведомо низкокачественной информации. Практика разработки алгоритмов для электронных средств массовой информации, образовательных платформ, социальных сетей, опросных приложений и мессенджеров диктует необходимость создания гибких и адаптивных классификационных моделей. Отработанные на уровне теории подходы к выявлению структуры, намерения и наполнения текста позволяют формализовать целый ряд сценариев программной фильтрации: например, разграничивать информационные и экспрессивные сообщения, отделять содержательные отклики от бессмысленного или явно вредоносного контента.

При интеграции теоретических основ в информационные технологии стоит обратить внимание на выявление тематических и содержательных связей между текстами различной длины, жанра и функции. Это особенно важно при необходимости выявлять скрытые паттерны в совокупности открытых ответов в массовых опросах, проводить кластеризацию пользовательских сообщений в образовательных или экспертных предметных областях, а также

вычленять группы схожих по смысловой нагрузке текстов из разнородных по форме и происхождению массивов. Для достижения убедительных результатов лингвистические и вычислительные методы постепенно сливаются в комплексные алгоритмы, опирающиеся на лексико-семантические, синтаксические, онтологические и прагматические признаки, а также на современные архитектуры машинного обучения [5].

Параллельно с развитием методов тематической и смысловой группировки текстов перед исследователями и разработчиками встаёт задача объективной и воспроизводимой оценки качества текстовых данных. Важность этой задачи проявляется не только в образовательных тестовых платформах, но и в аналитике обратной связи, HR-сервисах, репутационных исследованиях, модерации интернет-дискуссий и мониторинге пользовательских отзывов. Критерии информативности, аргументированности, связности и логики построения текста, выработанные еще в классических работах по теории текста, сейчас закладываются в основу сложных метрик автоматической оценки, от которых зависит доверие к автоматизированным системам и адекватность их рекомендаций.

Разнообразие жанровых форм, наличие креолизованных и мультимодальных текстов, специфика билингвальной и мультикультурной коммуникации обусловливают необходимость постоянной доработки классификационных схем. Границы между функциональными классами текстов размываются, появляются гибридные типы сообщений и доселе неизученные речевые форматы. Постоянные изменения языка расширяют горизонты применения теоретических моделей и требует привлечения мультидисциплинарных методологий. Методология классификации становится не только лингвистическим инструментом, но и управленческой стратегией для больших информационных систем, научно-образовательных платформ и инструментов искусственного интеллекта.

В такой интерпретации классификация выходит за рамки традиционной лингвистики и становится основой для построения гибких программных сред, где автоматизация работы с текстовыми данными включает фильтрацию вредных или бесполезных сообщений, группировку по смыслу, оценивание новизны, оригинальности и качества представленных ответов, а также осуществление мониторинга содержательных и тематических тенденций внутри больших массивов. Особое значение приобретает внедрение алгоритмов, учитывающих не только количественные, но и качественные параметры текстов, способных работать с языковыми данными на уровне семантики, прагматики и дискурса — что требует углубленной теоретической базы.

Несмотря на значительный прогресс, сохраняется ряд вызовов, требующих дальнейшего научного и прикладного исследования. Среди них

продолжают оставаться открытыми вопросы уточнения критериев релевантности и взаимной семантической близости в многоязычных и гибридных корпусах, развитие научно обоснованных моделей для фильтрации спама и низкокачественного контента, совершенствование автоматических процедур оценки информативности и глубины смыслового наполнения открытых текстовых сообщений, а также внедрение комплексных архитектур, отражающих многоуровневую структуру современной речевой коммуникации.

В рамках рассмотрения роли и значения классификационных основ определяются перспективные подходы к разработке специализированного программного обеспечения для анализа текстов в массовых опросах и иных цифровых средах. Это предполагает не только интеграцию передовых методов автоматического анализа, но и учет национальных лингвистических традиций, адаптацию алгоритмов под специфику белорусско-русской билингвы, гибкость при работе с различными формами письменных откликов, а также создание достоверных автоматизированных инструментов фильтрации, кластеризации и оценки текстовой информации. Тем самым предмет данного исследования фокусируется на синтезе теоретических основ классификации, передовых технологических решений и современных алгоритмов анализа данных с целью повышения эффективности интеллектуальной обработки ответов на естественном языке, их лаконичного представления и объективного оценивания в контексте цифровых опросных систем.

ВЫВОДЫ

1. Анализ текстов развивается на стыке лингвистики, психологии, информатики и педагогики. Эволюция идет от формальных методов к учету лексико-грамматических, смысловых и межкультурных аспектов, особенно в цифровую эпоху.

2. Современные методы классификации текстов используют жанрово-стилевые, тематические, структурные и прагматические критерии. Применяются как классические, так и гибридные схемы, что позволяет создавать эффективные системы анализа для образовательных и прикладных задач.

3. Учет многоязычия и социокультурного разнообразия расширяет методы классификации, включая билингвизм и гибридные жанры. Белорусский опыт показывает важность адаптации моделей под национальные особенности.

4. Классификация текстов лежит в основе цифровых и интеллектуальных платформ, обеспечивая фильтрацию, ранжирование и кластеризацию текстов. Это требует сочетания формальных и содержательных признаков.

5. Остаются нерешенные вопросы: автоматизация фильтрации спама, релевантность в многоязычных коллекциях, критерии оценки качества текстов

и комплексные решения для анализа естественного языка. Эти направления требуют дальнейших исследований.

ГЛАВА 2
АНАЛИЗ ЗАДАЧИ АВТОМАТИЗИРОВАННОЙ ОБРАБОТКИ ТЕКСТОВЫХ ОТВЕТОВ В ОПРОСАХ

2.1 Виды текстовых ответов в опросах

В социологических, маркетинговых и образовательных областях часто возникает потребность в знакомстве со своей аудиторией, её исследовании и анализе для выстраивания моделей взаимодействия (различного рода промокампаний, акций, предложений и другого) и дальнейшего сотрудничества. Правильная организация формы и структуры вопросов определяет качество и полноту получаемых ответов, а их последующая автоматизированная обработка обеспечивает высокую скорость анализа и минимизирует трудозатраты на привлечение человеческих ресурсов. Однако, чтобы достигнуть максимальной эффективности в данном подходе, требуется разработать и внедрить универсальную систему типологии и обработки текстовых ответов респондентов.

Традиционно в анкетах рассматриваются два класса текстовых ответов: открытые и закрытые. Закрытые ответы обычно представляют собой краткие, заранее ограниченные по выбору формулировки, такие как варианты «да/нет», перечисления или шкальные оценки. В случае подобных типов вопросов автоматизация обработки сводится к распознаванию кодировок, статистическому анализу, алгоритмическому агрегированию и простейшей валидации. Систематика ответов на вопросы такого формата не вызывает сложностей и позволяет производить быструю количественную оценку мнений или опыта опрашиваемых.

Иная ситуация наблюдается при работе с открытыми текстовыми ответами, которые не ограничены фиксированными вариантами. Здесь респонденту предоставляется пространство для свободного высказывания — от простых фраз до развернутых пояснений, комментариев, рассуждений и предложений. Именно эта форма позволяет получить наиболее полное представление о мотивах, ценностях, личном опыте участников, выявить нюансы и контекстные детали, не отраженные в закрытых вопросах. С другой стороны широта и отсутствие ограничений в написании открытых текстовых ответов существенно усложняют их анализ, предоставляют возможности для осуществления фишинговых атак и других вредоносных действий в отношении системы или компании, а также требуют дополнительного времени на обработку. В частности, в многоязычной и поликультурной среде открытые ответы часто представляют собой билингвальные сообщения с элементами профессионального, регионального и молодёжного жаргона, что значительно усложняет их автоматизированную обработку.

Крупные компании регулярно проводят опросы и интервью с целью улучшения своих продуктов и лучшего понимания потребностей аудитории. Живое общение остается ключевым методом достижения поставленных целей, так как помогает выявить нюансы восприятия и получить более полное представление о респонденте. Стоит заметить, что довольно часто из-за высокой загруженности, нехватки времени, географической удалённости или других причин люди не всегда могут присутствовать на такого рода собеседованиях. В таких случаях открытые ответы могли бы стать удобным способом для респондентов высказать своё мнение в подходящее для них время. На практике компании не спешат использовать такой метод, так как затраты на обработку большого объёма данных значительно превышают выгоду, полученную он этих данных.

Бурное развитие нейросетевых технологий и чат-ботов в последние годы позволило многим организациям значительно улучшить анализ открытых ответов и уменьшить трудозатратность данного процесса. Тем не менее, значительная часть времени и ресурсов всё ещё тратится на ручную обработку данных, что делает проблему обработки открытых ответов актуальной в наше время.

Отдельно отмечается распространенность смешанных форм ответов, когда традиционные закрытые опросные блоки сочетаются с запросами на дополнительное пояснение, обоснование или расширенное описание причин выбора. Такой подход позволяет уточнять структуру мотивов, обогащать количественные данные глубинными качественными смыслами и предоставляет новые возможности для автоматизации поиска паттернов в поведении и взглядах респондентов.

2.2 Методы обнаружения спамовых ответов

В условиях массового сбора текстовых данных особое внимание уделяется достоверности и качеству полученных от респондентов ответов. Значительный объем обратной связи, поступающей при проведении социологических, маркетинговых, образовательных опросов, нередко сопровождается присутствием нерелевантных, бессмысленных, случайных либо намеренно спамовых сообщений. Подобные ответы формируют информационный шум, существенно искажают результаты анализа, усложняют задачу поиска значимых паттернов и требуют дополнительной нагрузки на работу исследователя или автоматических аналитических систем. Вопрос фильтрации спама — один из центральных в области автоматизированной обработки текстовых данных в прикладных и исследовательских задачах.

На ранних этапах развития этой проблематики основной упор делался на ручные и полуавтоматические проверки. В частности, распространённой

практикой был двойной пересмотр данных несколькими специалистами, использование контрольных вопросов, встроенных ловушек (trap questions), а также ограничений на объем и формат ввода. Однако в условиях увеличения числа участников опросов и масштабирования информационных систем ручные методы оказываются экономически и организационно неэффективными. Это стимулировало развитие методов автоматического выявления и фильтрации спамовых ответов, нашедших отражение как в зарубежных, так и в отечественных исследованиях.

Исследования последних двух десятилетий показали, что задачи по обнаружению «мусорных» ответов перекликаются с задачами антиспам-фильтрации в электронной почте, модерации веб-форумов и социальных сетей. Зарубежные авторы акцентируют внимание на статистических, вероятностных и обучаемых моделях идентификации спама. Одним из первых эффективных инструментов стала фильтрация на основе ключевых слов, шаблонов и частотных словарей, где анализировалось наличие неинформативных последовательностей (например, бессмысленных наборов символов, автозаполненных фраз, копипасты однотипных выражений). Со временем эти методы были дополнены более гибкими способами анализа, в том числе оценкой разнообразия лексики, повторяемости шаблонов, а также фиксированием необычно высокой или низкой длины ответа.

Современные зарубежные и российские работы демонстрируют эффективность интеграции методов машинного обучения: в качестве признаков для построения классификационных моделей используются частотные характеристики, метрики разнообразия, оценка специфичности и оригинальности, наличие повторяющихся ответов, совпадения с фабулой открытого вопроса. Также популярными становятся методы семантического анализа, позволяющие выявлять бессмысленные или искусственно сгенерированные тексты. Применение алгоритмов кластеризации и выявления выбросов обеспечивает дополнительную фильтрацию нетипичных или девиантных ответов, подчеркивая их отклонение от основного смыслового массива.

Отдельную группу методик составляет применение стилей и особенности письма. В ряде исследований отмечается, что автоматическое выделение «однотипных» или «аномальных» по авторскому почерку откликов позволяет существенно повысить качество отбора валидных текстов. Технологии анализа синтаксической структуры и морфологических особенностей текста, разрабатываемые в ряде белорусских и российских работ, позволяют учитывать особенности использования устойчивых оборотов, частоту грамматических ошибок, а также искажения и опечатки, получаемые при машинной генерации или незаинтересованном ответе респондента.

Современная практика построения автоматизированных опросных платформ указывает на тенденцию к интеграции ряда универсальных подходов: от простых евристик по длине и частоте слов до применения нейросетевых архитектур (BERT, RoBERTa) для семантической фильтрации и автоматического выделения аномалий [4]. Показатели успешности таких решений во многом зависят как от объема и разметки исходных данных, так и от адаптации алгоритмов под специфику языка, жанра, функциональной направленности вопроса.

В процессе совершенствования методов фильтрации спамовых откликов остаются открытыми важнейшие вопросы: повышение точности и надежности автоматической фильтрации, минимизация количества ложноположительных и ложноотрицательных срабатываний, возможность настройки гибких критериев для разных типов опросов и языковых сред. Отмечается рост потребности в эффективных, масштабируемых инструментах модерации обратной связи, что повышает значимость дальнейших исследований в области автоматизации фильтрации нерелевантной информации и совершенствования соответствующих алгоритмов.

Включение в работу передовых зарубежных и проверенных отечественных подходов, а также адаптация методов к работе с многоязычными и смешанными наборами данных — одна из ключевых целей на пути к разработке современной интеллектуальной системы обработки текста.

2.3 Методы кластеризации и группировки свободных ответов

Ни одно исследование в рамках задачи обработки свободных текстовых ответов, направленное на максимально эффективную аналитику собранных массивов данных и извлечение из них репрезентативной, практически значимой и статистически обоснованной информации, не представляется возможным без использования методов кластеризации и группировки.

Истоки методов группировки текстовых данных можно проследить в лингвистических и социологических традициях анализа смысловых полей и тематических группировок. На первоначальных этапах массовых опросов распространенной практикой выступала ручная кодировка и категоризация откликов по разработанным исследователем рубрикаторам. Операторы анализировали массивы записанных текстов, выделяли концептуальные и тематические единицы, группировали их в так называемые "ответные кластеры" – тематические или проблемные группы, обладающие близким содержанием. Несмотря на относительную надежность и глубину анализа, такие методы крайне трудоемки и плохо масштабируются при увеличении объемов данных, а также подвержены субъективности кодировщика.

Активное развитие методов машинного анализа текста, начавшееся в последние два десятилетия, радикально изменило подходы к обработке свободных текстов. В международной практике для решения задачи группировки обычно применяется широкая гамма алгоритмов машинного обучения, статистических и семантических методов, ориентированных на автоматизацию выделения смысловых связей и схожести между текстами. К числу базовых относятся классические алгоритмы кластеризации, такие как k-means, агломеративная и иерархическая кластеризация, DBSCAN, спектральные методы и другие [5]. Их применение предполагает представление текстовых данных в виде векторов — векторов признаков, получаемых с помощью так называемых "мешков слов", TF-IDF, либо современных эмбеддингов (например, Word2Vec, GloVe, FastText, BERT).

TF-IDF оказывает огромное влияние на традиционную автоматическую обработку небольших текстовых корпусов. Преобразовав ответы участников опроса в векторное пространство, алгоритмы типа k-means способны объединять близкие по смыслу тексты на основе вычисленных расстояний между векторами. Однако при обработке более сложных, емких и насыщенных комплексных данных всё большее распространение получают нейросетевые методы семантического представления текста, что позволяет лучше учитывать синонимию, многозначность, контекстуальную окраску и идиоматические особенности языка. С этой точки зрения выдающиеся результаты демонстрируют трансформерные архитектуры, такие как BERT, RoBERTa, ALBERT и их аналоги, возникшие в результате бурного развития глубокого обучения в последние 5 - 7 лет [4].

Важным этапом перед проведением кластеризации является препроцессинг: нормализация текста, очистка от стоп-слов, лемматизация, устранение пунктуации и цифровых шумов, а также отброшенные нерелевантные сообщения с помощью алгоритмов. Следующий этап — выделение признаков: элементарные (мешок слов, биграммы), лексико-семантические (TF-IDF, тематические модели) или контекстные (BERT, Doc2Vec). После этого применяются алгоритмы группировки, формирующие кластеры на основе схожести эмбеддингов.

Особое место занимает тематическое моделирование — автоматическое определение доминирующих тем и смысловых направлений, встречающихся в свободных ответах. Одним из наиболее часто применяемых методов является вероятностная тематическая модель LDA, позволяющая выявлять латентные (скрытые) темы в больших корпусах без предварительно заданных рубрикаторов [2]. Комбинация тематического моделирования с обычной кластеризацией значительно расширяет спектр аналитических задач: исследователь может не только получить группы схожих по содержанию ответов, но и

конкретно формализовать обсуждаемые темы для дальнейшей детализации или распознавания новых, ранее не идентифицированных проблем.

Для стран с несколькими национальными языками актуальна проблема мультилингвальности и смешения языковых кодов в одном массиве текстовых ответов. Это требует адаптации известных алгоритмов либо специального препроцессинга с учётом морфологических и семантических особенностей одновременно всех необходимых языков.

Современные программные пакеты и библиотеки (например, scikit-learn, spaCy, gensim, HuggingFace Transformers) обеспечивают исследователей набором готовых инструментов для подготовки, представления и кластеризации текстовых данных, что значительно ускоряет процедуру внедрения интеллектуального анализа текстов в практические опросные системы. Тем не менее ряд вопросов, связанных с интерпретируемостью кластеров (explainability), выбором числа групп, адаптацией методов к коротким пользовательским высказываниям и смешанным или специфическим жанрам (например, молодежная речь, профессиональные сленги, ответы в формате мемов), остаётся открытым и требует комплексных эмпирических и теоретических исследований.

Дополнительные проблемы возникают при анализе очень больших массивов текстов — необходимость масштабируемости алгоритмов, правильного учета редких тем и устойчивое определение границ смысловых кластеров. Особенно сложной является задача «чистки» кластеров от спама, выявления смысловых шумов, фильтрации редких, но потенциально важных паттернов поведения респондентов.

Наблюдается тренд к интеграции кластеризации с последующим автоматическим или полуавтоматическим аннотированием, что облегчает не только машинное, но и экспертное обобщение данных. В этом контексте важную роль начинают играть интерфейсы визуализации кластеров, облегчающие представление сложных многомерных группировок для конечного пользователя-аналитика.

В последнее время усиливается интерес к гибридным моделям – ручная и автоматизированная кластеризация, объединение классификаторов на основе традиционных статистических признаков и современных нейросетевых векторных представлений, что обеспечивает лучшую устойчивость алгоритмов к шумам и неявным ассоциациям. Особую перспективу представляют архитектуры с возможностью дообучения на малых, но тщательно размеченных частях корпуса, что соотносится с вызовами мультилингвальных и региональных опросных систем.

В итоге задачи кластеризации и группировки свободных текстовых ответов выступают стратегическим звеном в автоматизации анализа опросной информации. Их решение требует всестороннего учета как языкового

разнообразия, связанности смыслов, наличия спама и некорректных ответов, так и использования самых современных методов машинного обучения, натренированных на релевантных национальных корпусах. Только на базе глубокой интеграции теоретических, прикладных и технологических достижений возможно создание интеллектуальных систем, эффективно работающих с большими, разнообразными, тематически насыщенными текстовыми массивами.

2.4 Автоматическая оценка качества и содержательности текстовых ответов

Современные аналитические задачи при обработке результатов опросов всё чаще выходят за рамки простой фильтрации и тематической группировки текстовых данных. Одним из ключевых направлений становится автоматическая оценка качества и содержательности текстовых ответов респондентов — задача, с одной стороны, глубоко лингвистическая, а с другой — тесно связанная с развитием средств искусственного интеллекта и машинного обучения.

Первые попытки формализовать оценку качества текстов основывались на ручном экспертном анализе, когда опытные модераторы рассматривали полноту, связность, логическую аргументированность, соответствие смыслу вопроса, лексическое и грамматическое разнообразие каждого ответа. Несмотря на высокую надёжность, субъективность и большая трудоёмкость подобных методов стали стимулом к поиску автоматизированных и воспроизводимых решений. Оценка информативности и качества, как правило, становится особенно сложной при большом количестве данных и доминировании коротких, лаконичных, а зачастую фрагментарных высказываний, характерных для массовых онлайн-опросов.

Зарубежные исследования предлагают использовать комбинацию лингвистических признаков и вычислительных метрик для построения автоматизированных моделей. Классические подходы предполагают вычисление таких показателей, как длина ответа, разнообразие словарного запаса (type-token ratio), средняя длина слова/предложения, количество уникальных лексем, частотность определённых грамматических структур, а также доля тематических (ключевых) слов, встречающихся в формулировке вопроса. Исключительно количественные метрики оказываются недостаточными для полноценной оценки содержательности, так как "длинные" и формально сложные тексты не всегда бывают по-настоящему информативными или релевантными.

Заметен переход к более сложным комплексным подходам, включающим семантический анализ, вычисление тематического и смыслового соответствия между вопросом и ответом, распознавание плагиата, использование LDA для сравнения тематической направленности, а также современных моделей эмбеддингов. Семантические методы позволяют сравнивать не только

лексические, но и скрытые смысловые аспекты текста, вычисляя похожесть между векторным представлением вопроса и высказывания. Особенно эффективны такие методы при оценке коротких ответов, где классические статистические показатели дают неоднозначные результаты.

Для получения воспроизводимых оценок внедряются supervised-модели, предварительно обученные на размеченных конкурсах или корпусах, где каждому тексту присваивается балл по шкале понятности, логичности, информативности, креативности или полноты. Такие подходы широко апробированы на примере автоматической проверки письменных работ в образовательных системах (например, Automated Essay Scoring [7], сокращённо AES, используемый в TOEFL и аналогичных платформах). В последние годы замечено распространение "гибридных" систем, объединивших экспертные (человеческие) оценки и алгоритмические метрики, что позволяет применять поправку на культурно-языковую специфику, а также отслеживать смещение или упрощение критериев при автоматической обработке в многоязычных коллекциях.

В белорусской и российской практике вопросы автоматической оценки качества текстовых ответов только начинают получать массовое развитие. Основное внимание уделяется выявлению релевантности содержания, оригинальности высказывания, отсутствия повторов, орфографических и грамматических ошибок, а также обработке нестандартных языковых конструкций, свойственных двуязычной или региональной среде. Понимание того, что информативность определяется не только лексико-грамматическим, но и прагматическим компонентом, стало основой для новых моделей оценки ответов, способных учитывать цель вопроса, социальный или профессиональный контекст.

Ключевой вызов здесь заключается в необходимости балансировать между универсальностью алгоритмов и их адаптацией к конкретным языкам, жанрам, ожиданиям экспертов и тематике опроса. Семантические модели часто требуют локального дообучения на собственных, тщательно размеченных корпусах — в противном случае падает точность оценки, а система склонна либо "переоценивать" формальные параметры, либо ошибочно занижать информативность тематически узких или креативных ответов.

Перспективные направления развития связаны с интеграцией внешних источников знаний (например, онтологий, терминологических словарей), что позволяет повысить релевантность анализа для профессиональных сообществ, а также с гибкой настройкой метрик для разных типов задач: от единых оценки качества до многоаспектных шкал, учитывающих связь с контекстом, аналитическую насыщенность и коммуникативный замысел автора ответа. По-прежнему остаётся открытым вопрос корректной и автоматической обработки иноязычных и диалектных проявлений, смешанных текстов, эмоциональной

лексики, а также оценка оригинальности в массовых платформах, куда респонденты нередко копируют стандартные ответы.

ВЫВОДЫ

1. Характер и структура текстовых ответов в опросах напрямую влияют на подходы к их автоматизированной обработке.
2. Одна из важнейших проблем обработки текстов — наличие спама и нерелевантной информации.
3. Кластеризация и группировка свободных текстовых ответов обеспечивают структурирование больших массивов данных, выявление скрытых паттернов и тем самым сокращают трудоемкость аналитики для исследователя.
4. Наиболее перспективными являются модели, объединяющие лингвистические, семантические и прагматические критерии, а также учитывающие специфику языка, жанра и контекста опроса.

ГЛАВА 3
ПРОЕКТИРОВАНИЕ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ОБРАБОТКИ ТЕКСТОВЫХ ОТВЕТОВ

3.1 Требование к функционалу системы

Проектирование программного обеспечения для автоматизации анализа текстовых ответов должно опираться на строгий набор требований, обусловленных реальными исследовательскими и практическими задачами, а также современными принципами построения интеллектуальных и устойчивых ИТ-систем. Разработка Survey Analyzer — специализированного программного комплекса для автоматизированной обработки анкетных данных — подчинена принципам модульности, масштабируемости и расширяемости, что позволит учесть специфику работы с реальными анкетными данными.

Система должна удовлетворять следующим ключевым требованиям:

1. Поддержка импорта данных из различных источников. Необходимо реализовать поддержку загрузки данных из CSV, Excel, JSON-файлов. Также следует обеспечить легкую интеграцию с существующими внешними анкетными и корпоративными информационными системами, используя отдельные модули-загрузчики (loaders), способные автоматически определять формат входных данных и обеспечивать совместимость между источниками.

2. Мультиязычная предобработка текстовых данных. Система должна корректно обрабатывать тексты на русском, белорусском и других используемых языках. Функционал предобработки должен включать очистку (удаление пробелов, выбросов, дубликатов), токенизацию, лемматизацию, удаление стоп-слов, нормализацию символов. Должна быть предусмотрена возможность масштабируемой и конфигурируемой предобработки для разных типов опросов и языковых сред.

3. Интеграция автоматизированной фильтрации спамовых и бессмысленных ответов. Система обязана выявлять и удалять дублирующиеся, бессмысленные, неинформативные или случайно сгенерированные отклики. В ядро должны входить модифицируемые фильтры для качественной "очистки" в многообразных условиях.

4. Интеллектуальная кластеризация и тематическое моделирование. Возможность группировать свободные ответы по смыслу и структуре, формировать тематические кластеры, выявлять скрытые смыслы и паттерны. Модули кластеризации следует реализовать с помощью различных алгоритмов (k-means, DBSCAN, иерархическая кластеризация) и векторизации текста (TF-IDF, Word2Vec, BERT и др.), что позволит обеспечить адаптацию под характер данных и размер выборки. Система должна позволять визуализировать и анализировать результаты кластеризации для оценки качества работы.

5. Автоматическая оценка текстовых ответов. В функционал системы следует добавить многоуровневую оценка ответов: как на основе формальных признаков (длина, уникальность лексики, доля тематических слов), так и с помощью моделей машинного обучения или семантического анализа.

6. Современный и удобный пользовательский веб-интерфейс. Пользовательский интерфейс должен строиться на принципах простоты и интуитивности, включать в себя инструменты для загрузки и просмотра данных, настройки обработки, запуска анализа, просмотра и экспорта результатов (в том числе через REST API), а также административные функции управления пользователями и системными настройками.

7. Гибкая аналитика и отчетность. Для поддержки принимающих решений пользователей необходимо реализовать модуль анализа результатов: формирование графиков, диаграмм, тепловых карт, генерация комплексных отчетов в различных форматах (PDF, Excel, HTML, DOCX). Должна быть поддержка расширенного статистического анализа.

8. Быстродействие и масштабируемость. Предполагается возможность горизонтального масштабирования обработки (распределенные очереди задач, запуск воркеров на нескольких машинах), поддержка одновременной работы нескольких пользователей и обработка больших объемов анкетных данных без потери производительности.

9. Ролевая модель и безопасность данных. Поддержка разграничения прав доступа (администратор, аналитик, пользователь), надежная аутентификация, хранение пользовательских данных в безопасных хранилищах, логирование и аудит всех ключевых операций.

10. Интеграция, расширяемость и контейнеризация. Система должна быть легко масштабируемой и интегрируемой с внешними решениями — поддержка REST API, возможность добавления новых модулей (фильтров, кластеризаторов, моделей), развертывание в контейнерах Docker/Docker Compose для удобства эксплуатации и развертывания в различных инфраструктурах.

Функциональные требования отражают как необходимость решения практических исследовательских задач (фильтрация, кластеризация, оценка), так и современные представления о надежной архитектуре интеллектуальных аналитических систем, способных эволюционировать в соответствии с потребностями науки, образования и бизнеса.

3.2 Архитектура программного комплекса

Реализация программного комплекса Survey Analyzer предполагает четкое разграничение логических компонентов, каждый из которых отвечает за одну из ключевых задач обработки текстовых ответов в опросах, а их взаимодействие составляет единую модульную архитектуру, соответствующую

современным требованиям к интеллектуальным системам анализа данных. Важной особенностью архитектуры выступает прозрачность всех этапов прохождения данных — от момента загрузки до аналитической интерпретации и визуализации итогов. Сама архитектура системы представлена на рисунке 3.1.

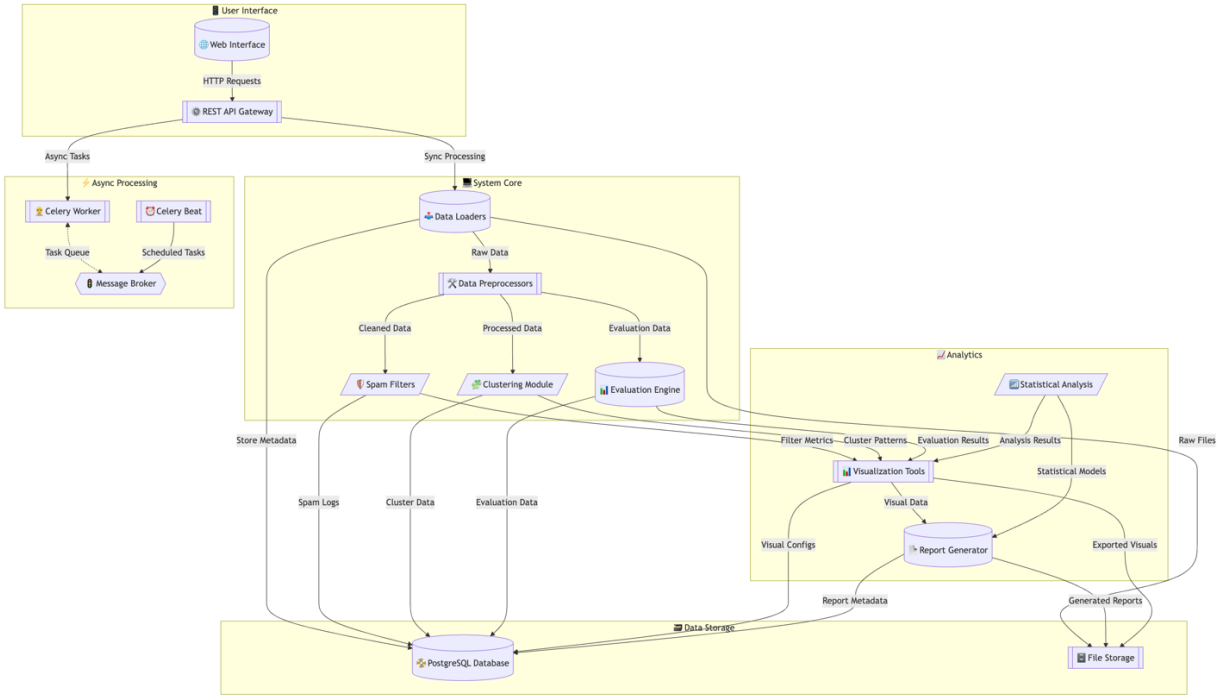


Рисунок 3.1 – Подробная архитектура системы

На верхнем уровне архитектоники выделяется пользовательский слой, реализованный средствами современного веб-интерфейса. Данный компонент служит точкой входа в систему для конечного пользователя и специалистов, участвующих в анализе данных: здесь осуществляется загрузка файлов опросов, настройка параметров и алгоритмов, запуск процедур анализа и просмотр результатов. Взаимодействие пользователя с системой поддерживается как через веб-формы, так и посредством REST API, что позволяет интегрировать решение с внешними цифровыми платформами. Подобная гибкость и открытость интерфейса определяет основу для будущей масштабируемости—добавление новых интерфейсных и аналитических модулей возможно без существенных переделок базовой логики.

Далее в логике обработки данных следует ядро системы или Core, которое в архитектуре разделено на ряд независимых функциональных модулей. Каждый модуль реализует определённую технологическую цепочку, а их взаимодействие обеспечивает полноту и последовательность анализа.

Формирование любого аналитического процесса начинается с модуля загрузки данных. Система поддерживает работу с несколькими источниками: Excel, CSV, JSON. Логика загрузки основана на абстрактном базовом интерфейсе, благодаря чему можно не только легко добавлять новые типы

загрузчиков, но и обеспечивать последовательную обработку форматов, что особенно важно при использовании решения в организациях с разнородными системами хранения и экспорта данных. На этом этапе важно подчеркнуть, что схема взаимодействия между компонентами предусматривает автоматическое определение формата и дальнейшую передачу данных на последующую обработку.

Ключевой стадией для дальнейшей аналитики становится модуль предобработки (препроцессинга). Здесь происходит очистка данных от служебной информации, удаление дубликатов, пропусков, лемматизация и токенизация текста, а также его нормализация. Система реализует отдельные подмодули для каждого из этапов обработки, что создает условия для замены или доработки механизмов обработки без переделки всей архитектуры, позволяет гибко работать с задачами мультиязычности или специфичной терминологии. Логика прохождения данных по модулям строго последовательная — каждый из них обрабатывает входной поток и передает результаты далее.

После очистки и стандартизации текстовая информация поступает к цепочке фильтрации спамовых и нерелевантных ответов, которая представлена на рисунке 3.2. В состав этого логического блока входят: фильтры дубликатов, эвристические и ML-фильтры бессмысленных сообщений, а также модули для расширяемого моделирования критериев релевантности. Каждый из фильтров работает в рамках универсального абстрактного интерфейса и может быть активирован или деактивирован в зависимости от типа опроса или специфических требований к чистоте данных. Программные интерфейсы позволяют гибко добавлять новые методы фильтрации, не затрагивая функционирование других слоев системы.

После фильтрации данные поступают на обработку модулями кластеризации и тематического моделирования. Каждая группа реализует свой алгоритм — k-means, DBSCAN, иерархическое моделирование, тематические эмбеддинговые модели — и интегрирована с модулем преобразования текста в векторное пространство, что позволяет осуществлять содержательное объединение схожих ответов, выявление скрытых тем и паттернов. Результаты кластеризации и тематического анализа записываются для последующего отображения пользователю как в виде списков, так и в виде графических визуализаций.

Важной частью архитектуры является модуль оценки текстовых ответов. Предусмотрена реализация универсального интерфейса для оценщиков, который поддерживает работу с простыми rule-based решениями, расширенными моделями на основе машинного обучения и компонентами семантической оценки. Такая реализация позволяет осуществлять сквозную настройку как для образовательных, так и для маркетинговых или научных задач, а также

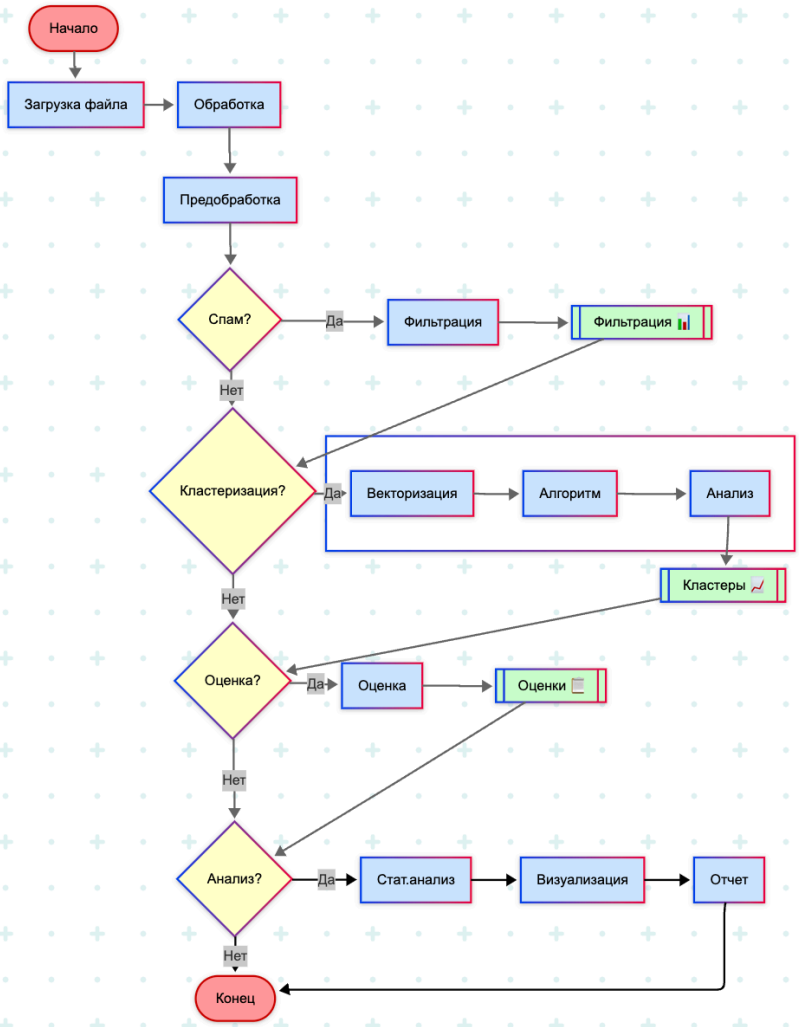


Рисунок 3.2 – Схема процесса анализа ответов

легко расширять систему за счет новых критериев или метрик качества.

В отдельном логическом слое организованы модули аналитики и визуализации. Здесь происходит построение и публикация отчетов, формирование графических и табличных представлений, генерация сводок, тепловых карт, иных форм анализа для интерпретации результатов как специалистами, так и конечными пользователями. Выходные данные можно экспортировать в стандартные форматы для дальнейшей работы.

Важно подчеркнуть, что в рамках всей архитектуры предусмотрен централизованный обмен данными между компонентами — результаты работы каждого слоя сохраняются в надежных хранилищах (PostgreSQL, Redis) и могут быть переиспользованы или выгружены по требованию пользователя. Это обеспечивает не только безопасность и целостность данных, но и масштабируемость за счет поддержки распределенных вычислений и контейнеризации.

С точки зрения безопасности и администрирования предусматривается гибкая ролевая модель доступа: разделение прав между администраторами, аналитиками и пользователями, защищенные протоколы взаимодействия, а также надёжная аутентификация.

3.3 Выбор инструментов и технологий

Подбор инструментов и технологий для реализации Survey Analyzer осуществлялся в полном соответствии с поставленными задачами интеллектуальной обработки текстовых ответов в опросах, учетом требований к масштабируемости, скорости, удобству поддержки и расширяемости решения. Особое внимание уделено технологической совместимости между компонентами, поддержке русского и белорусского языков, а также прозрачности интеграции с современными инфраструктурными решениями.

Основным языком программирования системы выбран Python — как индустриальный стандарт в области анализа данных и машинного обучения. Использование Python предоставляет богатый выбор профильных библиотек для обработки текста, построения моделей машинного обучения, автоматизации аналитических процессов и создания современных веб-сервисов.

Работа с текстами естественного языка реализуется с помощью таких библиотек, как NLTK и spaCy, которые отвечают за синтаксический и морфологический анализ, токенизацию и лемматизацию, а также обеспечивают устойчивую поддержку русского языка и богатые возможности предобработки. Для тематического анализа, построения векторных представлений слов и кластеризации в системе интегрированы Gensim (векторизация Word2Vec, тематическое моделирование), TextBlob (анализ тональности и работы с короткими ответами), а также модуль langdetect для автоматического определения языка каждой записи. Ключевую роль в обеспечении точных и гибких эмбеддингов играет пакет Transformers с использованием предобученной модели BERT для русского языка (DeepPavlov/rubert-base-cased), что особенно важно при автоматической оценке качества и смысловой близости текстов.

Методы векторизации охватывают полный спектр — от классических TF-IDF, позволяющих оперативно анализировать короткие и средние высказывания, до современных моделей Word2Vec и BERT, которые раскрывают глубокие смысловые и контекстуальные связи в больших и разнородных корпусах данных. Интеграция этих инструментов даёт возможность эффективно переключаться между быстрыми и интерпретируемыми метриками и более мощными, но ресурсоёмкими нейросетевыми представлениями.

Для машинного обучения и построения всех ключевых функциональных блоков — фильтрации спама, кластеризации, оценки качества и информативности ответов — используются Scikit-learn (библиотека алгоритмов для анализа данных и моделирования), TensorFlow и PyTorch (фреймворки для построения глубоких нейронных сетей). Поддержка разнообразных методов (k-means, DBSCAN, Agglomerative Clustering, Ridge Regression, Random Forest)

позволяет эффективно решать задачи на выборках разной природы и объема, а также подобрать оптимальные алгоритмы через быстрый эксперимент.

Веб-интерфейс и REST API реализованы на основании Django и Django REST Framework, что гарантирует масштабируемость, устойчивость к ошибкам и широкую функциональность для интеграции внешних приложений или аналитических интерфейсов. Важные технологии для обслуживания запросов, фильтрации данных, управления API-документацией и обслуживания статики (gunicorn, Whitenoise, django-filter, drf-yasg) делают систему готовой для эксплуатации в корпоративной и исследовательской среде.

В качестве баз данных используется связка PostgreSQL (основная реляционная база для хранения структурированных данных) и Redis (брокер сообщений, поддержка асинхронности и кэширования). Для работы с неструктурированными документами предусмотрена интеграция с MongoDB. Разработка и тестирование системы выполняются на SQLite и локально встроенных БД для ускорения итераций.

Асинхронная обработка реализована посредством Celery, обеспечивающего распределение длительных вычислений, параллельную обработку задач и эффективную работу с большими наборами данных. Мониторинг очередей и рабочих процессов организован через интерфейс Flower.

Для визуализации и генерации детализированных аналитических отчетов в Survey Analyzer интегрированы ведущие библиотеки: Matplotlib, Seaborn, Plotly, Bokeh, Altair для практически любого типа графиков, WordCloud для анализа ключевых слов и тематических облаков. Генерация и форматирование отчетов выполняются с помощью Jinja2, pdfkit, WeasyPrint и python-docx, что позволяет экспортировать результаты в популярные офисные форматы и PDF.

Вспомогательные библиотеки, такие как NumPy и Pandas, использованы для работы с табличными и массивными данными, а joblib и tqdm используются для повышения производительности вычислительных процессов. Для работы с файлами различных типов применяются openpyxl, xlrd, PyPDF2 и pdfminer.six, покрывая потребности в загрузке и выгрузке отчётных материалов, обмене с внешними платформами и корпоративными сервисами.

Внимание к поддержке морфологических и синтаксических особенностей русского языка выражается через использование специализированных моделей BERT и фукнционала NLTK/spaCy, что делает систему релевантной для задач в национальной языковой среде.

Архитектурно вся система поддерживает модульность и простое масштабирование: логирование и мониторинг организованы на всех ключевых этапах (Celery, Django, API, обучение моделей), а Docker и Docker Compose обеспечивают быструю развертку в любых инфраструктурах. Для

тестирования и контроля качества кода используются фреймворки pytest/pytest-django, инструмент покрытия coverage.

3.4 Описание используемых алгоритмов

3.4.1 Фильтрация спамовых и нерелевантных ответов

В разрабатываемом ПО автоматизация анализа текстовых ответов реализуется с помощью тщательно продуманных, модульных алгоритмов, предназначенных для трёх ключевых задач: фильтрации спамовых/нерелевантных ответов, кластеризации свободных откликов и автоматической оценки качества и информативности полученных данных. Каждый алгоритмический модуль строится на расширяемой базе и может быть настроен или доработан под специфику конкретного опроса.

Фильтрация спама в Survey Analyzer реализована как последовательность специализированных этапов, организованных в гибкую цепочку обработки с возможностью конфигурирования порядка и принципов взаимодействия между модулями. Такая организация процесса обеспечивает эффективное выделение и исключение из анализа некачественных, повторяющихся и бессмысленных текстовых ответов, что напрямую влияет на надёжность последующих стадий кластеризации и оценки данных.

Первую ступень обработки текстовых откликов занимает DuplicateFilter, который выявляет дублирующиеся ответы. Алгоритм начинается с поиска абсолютно идентичных текстов посредством стандартных инструментов поиска дубликатов в pandas. Этот базовый шаг позволяет мгновенно удалить наиболее очевидные повторы, не нагружая систему дальнейшими вычислениями. Там, где простого сравнения недостаточно (например, при наличии случайных опечаток или незначительных изменений), применяется метод нечеткого совпадения, построенный на расстоянии Левенштейна, что обеспечивает высокий уровень чувствительности к схожести текстовых формулировок. Для массовых опросов, где количество текстов велико, для быстрой предварительной фильтрации дополнительно используется сравнение по n-граммам и коэффициенту Жаккара, что сокращает количество попарных сравнений и ускоряет обработку. На завершающем этапе DuplicateFilter способен вычислять семантическую схожесть между ответами, опираясь на TF-IDF векторизацию текстов и вычисление косинусного расстояния между векторами – это даёт возможность обнаруживать перефразированные, но схожие по смыслу отклики.

Второй этап связан с ликвидацией бессмысленных и неинформативных ответов посредством NonsenseFilter. Работа этого фильтра построена на многоуровневой оценке текста: контролируется диапазон длины ответа (как слишком короткие, так и чрезмерно длинные тексты могут быть

низкокачественными), анализируется уровень лексического разнообразия (типичная проблема простых повторов и бессвязных «разговорных» сообщений), а также выявляются нетипичные паттерны, такие как заметное повторение отдельных символов. Важной частью анализа становится пересечение словаря ответа с набором осмысленных слов, что позволяет отсеивать заведомо несущественные и фрагментарные высказывания, не дающие информации для дальнейшего анализа.

Реализация третьего этапа предусматривает использование MLFilter — модуля, основанного на методах машинного обучения для детекции спама. Здесь текстовые отклики предварительно преобразуются к векторному виду с помощью TF-IDF, после чего обученные модели (логистическая регрессия, случайный лес, нейронная сеть) определяют вероятность принадлежности ответа к категории спама на основе изученных паттернов из размеченного корпуса. Такой подход позволяет системе быть адаптивной, а точность классификации улучшается по мере накопления новых данных или актуализации моделей.

Ключевая особенностью архитектуры фильтрации является единый базовый интерфейс, от которого наследуются все фильтры. Это позволяет:

– унифицировать методы запуска фильтрации;

– сохранять на выходе общие метки ('is_spam', 'spam_score') и взаимосвязанные оценки для каждого этапа;

– комбинировать фильтры в разных последовательностях в зависимости от специфики задачи;

– наращивать новые методы фильтрации без перестройки существующего кода.

Алгоритмы взаимодействуют по принципу передачи промежуточных результатов: дубликаты удаляются или маркируются первыми, бессмысленные ответы фильтруются на следующей стадии, а заключительное решение о спаме уточняется моделью машинного обучения. Совокупная оценка учитывает максимальную из «spam-score» всех фильтров, позволяя оперативно выявлять даже сложные случаи нежелательных откликов. Каждый фильтр логирует свои решения, что облегчает анализ качества фильтрации для последующей калибровки параметров или поиска аномалий в массивах опросных данных.

В результате формируется отзывчивая и расширяемая экосистема очистки данных, необходимая для полноценной работы интеллектуальных систем анализа открытых текстовых ответов.

3.4.2 Кластеризация открытых ответов

Кластеризация текстовых ответов в системе реализована как многоэтапный процесс, который направлен на автоматическое выделение групп схожих по содержанию или структуре текстов. Такой подход позволяет анализировать большие объемы разнородных данных открытых вопросов, выявлять скрытые темы, основные направления мнений и паттерны поведения респондентов. Благодаря использованию нескольких алгоритмов и модульной архитектуре, система может быть гибко настроена под задачи конкретного исследования.

Ключевой ступенью подготовки данных для кластеризации является этап векторизации: каждый текстовый ответ преобразуется в числовое представление (вектор) с помощью такого механизма, как класс TextEmbedder. Методы векторизации (TF-IDF, Word2Vec, BERT) подбираются исходя из требуемой точности, особенностей корпуса и размера выборки. На выходе формируется матрица признаков, где каждая строка соответствует вектору отдельного ответа. Вся логика подготовки унифицирована: независимо от типа выбранного алгоритма дальнейшая обработка оперирует этим универсальным представлением.

На следующем этапе выбор алгоритма кластеризации осуществляется с учетом структуры данных, целевых требований к точности и интерпретируемости результатов. В качестве базовых решений реализованы три подхода:

K-means позволяет разбить ответы на фиксированное число групп, минимизируя внутригрупповую разницу. В рамках работы алгоритма происходит инициализация центроидов, к которым ответы "притягиваются" по мере итераций. Центроиды обновляются с каждым циклом, обеспечивая постепенное уточнение структуры кластеров. Применение этого метода оправдано при наличии представления о количестве тематических групп или необходимости строгой сегментации данных.

DBSCAN строит кластеры на основе плотности, что особенно полезно при неравномерном распределении мнений и наличии "редких" тем. Он выделяет группы, где ответы расположены достаточно близко друг к другу в векторном пространстве, и способен автоматически отличать "шум" или выбросы. Такой подход незаменим, когда количество групп заранее неизвестно, а структура данных предполагает возможные выбросы или неявные формы кластеров.

Иерархическая кластеризация создает дерево вложенных группировок, позволяя исследовать данные на различных уровнях детализации. Основываясь на выбранном методе связи (ward, complete, average, single), на каждом этапе объединяются самые близкие кластеры вплоть до образования единой иерархии. Визуализация результатов с помощью дендрограммы облегчает

интерпретацию, делает доступным ручной выбор оптимального уровня разбиения.

В основе системы лежит абстрактный класс BaseClusterer, задающий общий программный интерфейс для всех видов кластеризации. Такой подход обеспечивает строгую унификацию и простую замену одного алгоритма другим без необходимости переработки сопутствующих модулей. Методы fit, predict, find_optimal_clusters и _create_model реализуются для всех типов кластеризаторов, что способствует стандартизации работы на всех этапах.

Взаимодействие между компонентами построено на едином стандарте передачи данных: результаты каждого шага (например, метки кластеров после работы модели) имеют одинаковый формат, что позволяет свободно анализировать, визуализировать, сопоставлять или дообучать выбранный алгоритм. Анализ произведенных группировок осуществляется классом ClusterAnalyzer, который извлекает для каждой группы характерные слова, рассчитывает их статистики, визуализирует структуру данных, а также подбирает репрезентативные примеры для детального разбора содержимого кластера.

Все это в совокупности создает гибкую среду для поиска и анализа скрытых тем, паттернов и особенностей поведения респондентов в рамках открытых текстовых опросов. Модульность и расширяемость архитектуры позволяют не только варьировать методы векторизации и оптимизации параметров, но также внедрять новые алгоритмы для работы с особыми типами данных или специфическими корпусами. Такая организация обеспечивает высокий уровень адаптивности системы под требования современных исследований.

3.4.3 Оценка открытых ответов

Автоматическая оценка открытых ответов строится на принципах модульности, расширяемости и возможности использования различных подходов в зависимости от структуры данных, целей опроса и доступных ресурсов. Взаимодействие компонентов устроено так, чтобы аналитик мог легко комбинировать эвристические, машинно-обучаемые и семантические методы для оперативного и достоверного выявления качественных или релевантных откликов.

Базовым уровнем в системе выступает RuleEvaluator, реализующий оценку по ряду предопределённых критериев. Каждый критерий (например, длина ответа, разнообразие словаря, грамматическая корректность, наличие ключевых слов) оценивается отдельно и взвешивается в общей формуле. Такой подход позволяет гибко настраивать параметры для различных типов опросов: где-то акцент делается на лаконичности и структуре, где-то на оригинальности и лексическом богатстве. Оценка строится как сумма баллов по отдельным правилам с дальнейшей нормализацией: результат всегда

укладывается в заданный масштаб (например, от 0 до 10). Это удобно для быстрой первичной проверки массива ответов и служит надёжной защитой от откровенно слабых откликов еще до запуска более сложных схем оценки.

На следующем уровне реализован подход, опирающийся на машинное обучение (MLEvaluator). В этой архитектуре для каждого ответа строится числовой признак на основе TF-IDF, а далее регрессионная или ансамблевая модель (линейная регрессия, регрессионный случайный лес, многослойная нейронная сеть) предсказывает количественную оценку на основании заранее размеченных данных. Оценка на подобной выборке "учится" у экспертов, получая способность выявлять сложные скрытые зависимости и интегрировать в итоговую шкалу сразу большое количество признаков (длину, структуру, оригинальность и др.). Такой подход оправдан при наличии исторических разметок или необходимости регулярного масштабного анализа.

Третий подход заключён в модуле SemanticEvaluator. Здесь реализуется сравнение смысловой близости ответа респондента с эталонными образцами (при их наличии) через вычисление косинусного сходства между векторами, полученными с помощью методов векторизации текста. Если эталонов нет, в дело вступают эвристики определения семантической глубины: объединение показателей длины, лексического разнообразия и структуры текста. Данный компонент хорошо подходит для анализа предметных, тематических опросов либо для задач, где важна логическая и понятийная целостность ответа.

Все оценщики реализуют иерархию наследования от абстрактного класса BaseEvaluator, обеспечивающего единообразие методов (evaluate, score, нормализация). Такой дизайн облегчает интеграцию новых алгоритмов и позволяет на практике комбинировать методы — например, запускать ансамблевую оценку, в которой каждый модуль вносит свой вклад в итоговую шкалу. Если какой-либо из подходов оказывается временно недоступен или некорректен (например, отсутствует обученная модель для MLEvaluator), система автоматически переключается на альтернативную стратегию (обычно RuleEvaluator).

Архитектурное преимущество системы заключается во взаимосвязи модулей оценки с остальными частями комплекса: например, фильтрация спама и группировка ответов могут применяться перед оценкой, улучшая качество входных данных, а результаты кластеризации — использоваться для контекстной настройки критериев или анализа срезов. Кроме того, обратная связь полезна и для дообучения моделей спама или донастройки критериев группировки.

Особое внимание уделено расширяемости: каждый оценщик допускает настройку веса отдельных критериев, импорт/экспорт обученных моделей, добавление новых правил и методов обработки. Результаты приводятся к

унифицированному формату, что облегчает их анализ, сравнение и использование в последующей аналитике либо отчетности. Модуль оценки спроектирован как взаимозаменяемый и масштабируемый инструмент, способный адаптироваться под специфику любых опросных кампаний и исследовательских задач. В такой схеме каждый алгоритм – от эвристического до глубоких моделей – дополняет другие, повышая общую объективность, надежность и информативность аналитики по открытым текстовым данным.

3.5 Модели данных и пользовательский интерфейс

3.5.1 Модели данных

В разрабатываемом ПО особое внимание уделено грамотному проектированию моделей данных и построению удобного, интуитивно понятного пользовательского интерфейса. Такой подход позволит обеспечить устойчивость, масштабируемость и логическую целостность всей платформы, включая корректную работу с исследовательскими сценариями и высокую адаптивность к задачам обработки и аналитики.

Основу хранения информации в системе составляет реляционная база данных с использованием ORM Django, организующая единое информационное пространство для работы с опросами, файлами, задачами обработки и результатами аналитики. Архитектура моделей построена на связях "один-ко-многим" и "многие-ко-многим", что позволяет эффективно отражать реальную логику работы исследовательских проектов.

Центральной сущностью выступает модель Survey, хранящая метаданные опроса, его владельца, дату создания и актуальные состояния. Все полученные данные загружаются в виде файлов (SurveyFile), что позволяет гибко работать с наборами анкет в различных форматах и отслеживать статус их обработки. Для декомпозиции и отслеживания последовательности действий реализована модель ProcessingTask, агрегирующая отдельные этапы — фильтрацию, кластеризацию, оценку, генерацию отчетов. Это дает возможность не только детально анализировать процесс, но и управлять долгосрочными или асинхронными задачами.

Результаты анализа (отдельные объекты ClusterResult, SpamFilterResult, EvaluationResult) сохраняют всю ключевую информацию: используемые методы, параметры и метрики, а также тесно связаны с задачей обработки (ProcessingTask) и опросом. Такая структура облегчает последующее сравнение, повторный анализ и аудиторскую проверку. Для хранения визуализаций, отчетов, детализации кластеров и отслеживания прогресса созданы специализированные модели: Visualization, Report, Cluster, ProgressInfo.

3.5.2 Пользовательский интерфейс

Фронтенд системы Survey Analyzer разработан в соответствии с парадигмой Model-View-Template (вариант MVC в Django), что обеспечивает четкое разделение ответственности, хороший пользовательский опыт и простоту поддержки. Веб-интерфейс служит основным инструментом для загрузки, управления и анализа данных, а также работы с результатами машинного анализа.

Пользователь проходит аутентификацию, что гарантирует безопасность и защиту персональных данных. На главной странице предоставляется обзор всех доступных функций, опросов и быстрый доступ к основным операциям. Для каждого опроса реализована детальная страница, отражающая все связанные с ним файлы, задачи обработки, отчеты и визуализации. Это дает возможность контролировать полный жизненный цикл анализа от загрузки данных до их комплексной аналитики.

Загрузка новых данных интуитивно понятна: пользователь выбирает формат файла, видит статус его обработки и может следить за прогрессом операции в реальном времени. Запуск задач обработки (фильтрация, кластеризация, оценка) сопровождается простыми формами выбора параметров, а результаты отображаются на специализированных страницах с развернутыми визуализациями и деталями.

Для больших и длительных операций реализовано асинхронное выполнение, динамическое отображение состояний и сообщений об ошибках, а также страницы для отслеживания выполнения и просмотра истории. Отчеты и визуализации доступны для загрузки в распространённых форматах, что важно для последующей работы с данными (например, представления результатов заказчику или для публикаций).

Дизайн интерфейса построен на Bootstrap, что гарантирует адаптивность, корректное отображение как на десктопных, так и на мобильных устройствах, а современные компоненты (графики, отчеты, интерактивные элементы) делают аналитический процесс максимально эффективным и наглядным.

3.5.3 Интеграция моделей данных и пользовательского интерфейса

Ключевым принципом интеграции является использование сериализации и четкое разделение логики между слоями приложения. Все модели сериализуются для передачи во внутренние шаблоны, API и формы, что обеспечивает модульность и простоту расширения функциональности. Через Django-формы и REST Framework реализован удобный ввод и редактирование данных, автоматизирована валидация форм и бизнес-логика для критически важных сценариев.

Асинхронная обработка задач с помощью Celery и AJAX позволяет пользователю работать с системой без ожидания завершения длительных задач. Хранилище прогресса ProgressInfo отображает актуальное состояние любой обработки, а пользователь сразу информируется о завершении или ошибке операции.

Механизмы интеграции моделей и представлений обеспечивают согласованность, безопасность и масштабируемость решения. Любая бизнес-логика — от загрузки файла до генерации сложного отчета — четко контролируется цепочкой, представленной на рисунке 3.3. Такой паттерн позволяет быстро дорабатывать новые функции, наращивать аналитику, добавлять типы отчетов и визуализаций без переписывания уже отлаженных компонентов.

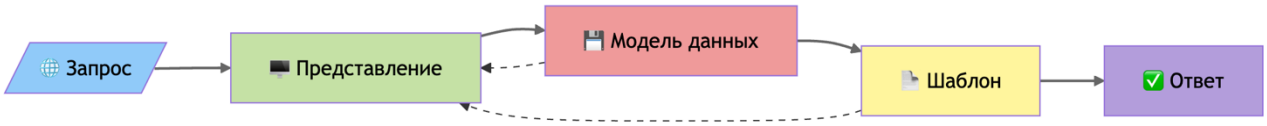


Рисунок 3.3 – Паттерн интеграций

В результате реализованная архитектура моделей и пользовательского интерфейса формирует надежный, гибкий и интуитивно понятный инструмент для команд исследователей и аналитиков, полностью отвечающий задачам интеллектуального анализа текстовых данных в опросах различного масштаба и тематики.

ВЫВОДЫ

1. Проектирование программного обеспечения для обработки текстовых ответов в опросах требует учёта специфики задач анализа естественного языка и построения функционала, удовлетворяющего требованиям к гибкости, расширяемости, безопасности и поддержке многоязычных данных.
2. Архитектура комплекса реализована в виде модульной, масштабируемой системы с чётким логическим разделением компонентов по этапам обработки данных: от загрузки и предобработки до фильтрации спама, кластеризации, оценки и визуализации результатов, что обеспечивает прозрачность и устойчивость работы.
3. Выбор инструментов и технологий выстроен с опорой на современные, устойчивые решения для обработки текста и данных (Python, Django, Scikit-learn, spaCy, Gensim, Transformers, Celery, PostgreSQL, Redis и др.), что позволяет эффективно реализовать масштабируемые вычисления, асинхронные задачи, работу с русским языком и визуализацию данных.

4. Для автоматизации анализа текста применены алгоритмы нескольких классов: фильтрация дубликатов, бессмысленных и спамовых ответов; кластеризация (k-means, DBSCAN, иерархическая); количественная и качественная оценка с использованием как формальных правил, так и моделей машинного обучения и методов семантической оценки.

5. Модели данных, реализованные на базе ORM, формируют устойчивую и масштабируемую структуру хранения информации об опросах, задачах, результатах и визуализациях.

6. Пользовательский интерфейс спроектирован по принципам адаптивности и интерактивности, чтобы обеспечивать удобную навигацию по задачам анализа, визуализировать прогресс и результаты, поддерживать интеграцию с отчетами и визуализациями, контролировать безопасность доступа и персонализацию пользовательского опыта.

7. Интеграция всех уровней — от моделей данных и логики обработки до визуальных решений — создает единую цифровую среду, готовую к дальнейшему развитию, расширению наборов задач и масштабированию на разные исследовательские и отраслевые сценарии.

8. Предложенная архитектура и комплекс решений закладывают основу для создания эффективной, интегрируемой и масштабируемой среды интеллектуального анализа текстовых данных.

ГЛАВА 4
РЕАЛИЗАЦИЯ И ЭКСПЕРИМЕНТАЛЬНАЯ ОЦЕНКА ЭФФЕКТИВНОСТИ СИСТЕМЫ

4.1 Реализация прототипа программы для обработки открытых ответов в опросах

В процессе создания прототипа системы автоматизированного анализа текстовых ответов особое внимание уделялось практической организации архитектурных и прикладных решений, а также быстрому запуску ключевых функций для оценки работоспособности комплекса в реальных условиях. Внешний вид ПО представлен в Приложении А.

Основную блок работы заключался в разработке на языке Python с использованием Django серверной логики и пользовательского веб-интерфейса. На ранних этапах фокус был смещён на реализацию минимально жизнеспособного продукта (MVP): пользовательская панель входа и регистрации, базовая страница опросов, а также механизмы загрузки файлов с результатами опросов в поддерживаемых форматах (CSV и Excel). Структура базы согласно модели Survey строилась сразу с учётом будущего расширения — все объекты опросов, задач, файлов и пользователей были спроектированы в виде связанных таблиц через Django ORM, а статусы каждой задачи мониторились через отдельную структуру (ProcessingTask).

Первым из ключевых обработчиков стал функционал фильтрации. В качестве прототипа реализованы три класса-фильтра: DuplicateFilter для обработки явных дубликатов (с использованием методов pandas), NonsenseFilter для удаления откровенно бессмысленных или слишком коротких/длинных сообщений (через простую валидацию длины и разнообразия словаря), а также простейший MLFilter, использующий обученную на заранее размеченной выборке логистическую регрессию (через scikit-learn и TF-IDF-векторизацию текста). Все фильтры реализовывались таким образом, чтобы сохранять промежуточные результаты — маски и вероятностные оценки — для последующего анализа и настройки параметров.

В прототипе кластеризации фокус был сделан на интеграции алгоритмов k-means и DBSCAN, оптимальных для начальных экспериментов с группировкой ответов. Векторизация текстов выполнялась с помощью TfidfVectorizer из scikit-learn. Для ручных запусков и проверки гипотез пользователю предоставлялась возможность выбирать число кластеров для k-means или задавать параметры плотности для DBSCAN прямо через веб-интерфейс. Результаты кластеризации отображались в виде списков (по группам), а позднее подключались простейшие визуализации через matplotlib/seaborn.

Оценка качества ответов на первом этапе строилась на эвристиках — длина сообщения, разнообразие слов и базовая проверка на ключевые формулировки, прописанные администратором через интерфейс создания параметров оценки. Для качественной проверки правильности работы модулей на реальных примерах использовалась валидация функций оценки на тестовых опросниках с заранее подготовленной экспертной разметкой.

Следующим основополагающим пластом выступает интерфейс мониторинга прогресса выполнения задач. Для длительных операций (например, кластеризации больших выборок) был разработан отдельный модуль ProgressInfo с индикацией статуса задачи и возможностью асинхронного оповещения пользователя о завершении обработки.

Пользовательский интерфейс создавался максимально адаптивным, чтобы даже на этапе MVP обеспечить понятную навигацию: отчетливо отделялись этапы работы с анкетами, запуски различных задач, детальный просмотр результатов обработки и скачивание итоговых файлов. Для демонстрационных целей отчеты и результаты кластеризации формировались в виде HTML-страниц и PDF, обеспечив совместимость с требованиями документирования и предоставления аналитики для внешних пользователей.

После реализации базовой связки фильтрация — кластеризация — оценка проводилось тестирование на нескольких корпусах реальных опросов: проводились проверки качества фильтрации, анализ распределения ответов по кластерам, сопоставление автоматических и экспертных оценок информативности и релевантности, а функционал визуализации позволял контролировать точность итогового разбиения.

В целом прототип разрабатывался по итерационному сценарию: после реализации каждой функции проводился ручной анализ ошибок, корректировка интерфейсных и бизнес-правил, доработка моделей данных и связь новых сущностей с системой отображения в интерфейсе. Такой динамичный подход обеспечивал не только быстрое получение рабочей версии, но и возможность гибко реагировать на изменения в требованиях или методах анализа в ответ на реальные эксперименты с корпусами текстов.

4.2 Подготовка и разметка тестовых данных

Для объективной оценки работы системы автоматизированного анализа текстовых ответов критически важно использовать тщательно подготовленные и размеченные тестовые данные. На этапе подготовки корпуса для экспериментальной части реализации прототипа большое значение придавалось разнообразию источников, качеству исходного материала и стандартизации процедур разметки. В основу тестового набора были включены данные из реальных анкет и опросов, проводившихся в образовательных, социологических

и HR-проектах — это позволило имитировать естественные условия применения системы.

Первым шагом в подготовке тестовых данных стала анонимизация исходных анкет с целью исключения персональной информации и соблюдения этических стандартов работы с пользовательскими сведениями. Оригинальные массивы были приведены к единым форматам CSV и Excel для быстрой загрузки и дальнейшей обработки средствами прототипа. Каждый опрос содержал как закрытые, так и открытые вопросы, но для работы системы фокус ставился только на открытых текстовых ответах.

На последующем этапе данные прошли полную процедуру предобработки: очистку от технических строк, дубликатов, специальных символов и очевидных бессмысленных сообщений. Для повышения качества тестирования отфильтровывались ответы длиной менее двух символов, тексты, не содержащие кириллических или латинских букв, а также заведомо пустые или состоящие только из пунктуации строки.

Разметка данных проводилась вручную для обеспечения объективности итоговых меток. Для задач фильтрации каждый ответ получал ярлык по категориям: «валидный», «дубликат», «бессмысленный», «спам». Особое внимание уделялось пограничным случаям — коротким неформальным репликам, повторяющимся по смыслу (но не по формулировке) ответам, а также разновидностям "бессмысленности" (например, сочетаниям хаотичных символов и автозаполнения).

Отдельный слой разметки предназначался для проверки алгоритмов оценки качества: выставляли баллы по шкале (например, от 0 до 10), учитывая полноту, содержательность, релевантность вопросу, структуру, аргументированность и степень проявления креативности в ответах. Для семантических алгоритмов также формировался набор эталонных ответов на каждый вопрос с максимальной оценкой — этот эталон служил ориентиром для сравнения смысловой близости автоматически обработанных текстов.

Для подготовки корпусов, используемых в задаче кластеризации, вручную определялись ключевые темы, к которым могли быть отнесены открытые ответы. Учитывались синонимы, парафразы, варианты употребления профессиональной или молодежной лексики. Таким образом, создавалась «золотая разметка» для объективной проверки структуры и состава автоматически найденных кластеров.

Отдельно прорабатывались вопросы согласованности и разрешения конфликтных разметок. Этот подход повышал достоверность результирующего набора и позволял строить корректные метрики точности, полноты, F1-меры для оценки работы фильтров, кластеризаторов и оценщиков.

Подготовленный и размеченный корпус данных стал не только базой для функционального тестирования реализованных модулей системы, но и критерием для повторяемой объективной проверки результатов дальнейших экспериментов и калибровки алгоритмов, что способствовало повышению надежности экспериментальной части работы.

4.3 Эксперименты по фильтрации спама: сравнение с ручной модерацией

Целью эксперимента является оценка эффективности различных способов фильтрации спама, результаты сопоставлялись с работой человека-модератора. В рамках эксперимента сравниваются пять подходов: фильтр дубликатов, фильтр бессмысленных ответов, фильтр на основе машинного обучения, комбинированная фильтрация и непосредственно ручная модерация. Для получения максимально объективного результата оценивания сравнение построено на ряде метрик качества, эффективности и ошибок. Графики сравнения метрик для экспериментов приведены на рисунках 4.1 – 4.3.

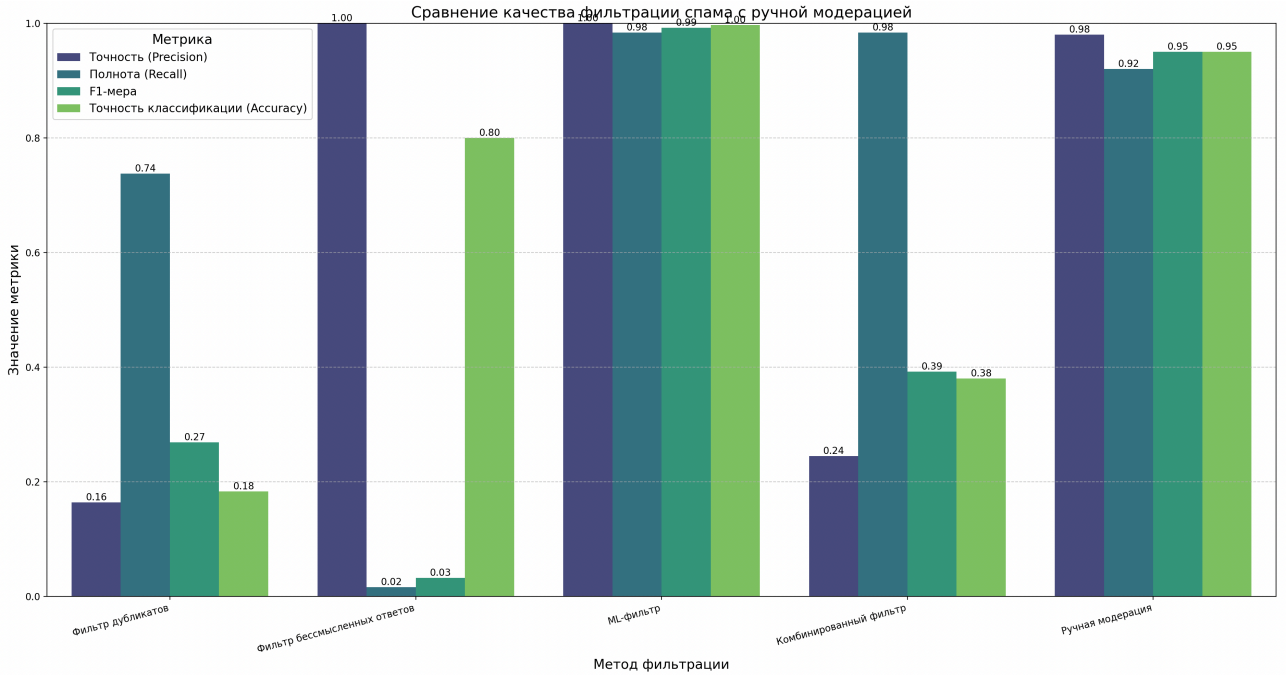


Рисунок 4.1 – График сравнения метрик качества

Как видно из графиков, фильтр дубликатов показал весьма низкую точность при достаточно высокой полноте, значение F1-меры указывает на значительный дисбаланс между количеством правильно отфильтрованного спама и ошибочно отмеченных сообщений, а низкий accuracy свидетельствует о слабой способности данного метода отделять спам от не спама в общем объеме сообщений. Фильтр бессмысленных ответов ближе всего к идеалу по точности, при этом полнота и F1-мера крайне низки, что указывает на крайне огра-

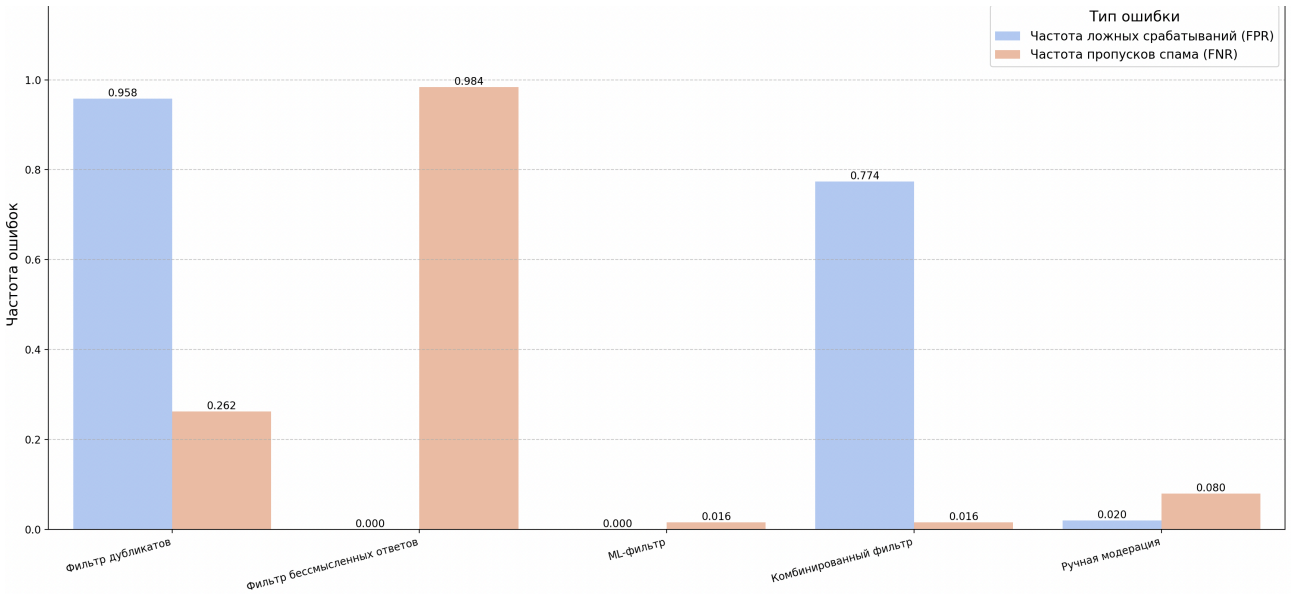


Рисунок 4.2 – График сравнения частоты ошибок

ниченный охват. Наибольший интерес представляет метод на основе машинного обучения. Здесь достигаются максимальные значения всех ключевых метрик эффективности. Минимальное число ложноположительных результатов и минимальные пропуски спама подтверждают, что модель практически не ошибается как в ту, так и в другую сторону. Комбинированный фильтр соединяет возможности предыдущих методов, что приводит к некоторому росту полноты при существенном падении точности. Высокое количество ложных срабатываний говорит о том, что часть не спам-сообщений ошибочно блокируется. Тем не менее, F1-мера здесь выше, чем у большинства базовых подходов.

Ручная модерация показывает наиболее сбалансированные показатели из всех методов. Модератор способен достаточно полно учитывать контекст, новизну и сложные случаи. Тем не менее, такие действия требовали больших затрат по времени. Количество ложных и пропущенных срабатываний здесь минимально, что подчеркивает преимущество "человеческого фактора", когда рассматривается весь спектр, включая пограничные случаи.

Если обратить внимание на метрики производительности, становится очевидным, что автоматические подходы принципиально превосходят ручной труд по времени обработки: выполнение каждого из фильтров требовало доли секунды при экономии времени на уровне 99.98–100%. Экономия распространена и на финансовые затраты соответственно, так как стоимость ручной модерации остается на относительно постоянном уровне, а автоматизация практически устраняет эти расходы при кратном росте объема обрабатываемых данных.

Можно заключить, что использование фильтра дубликатов позволяет быстро исключать повторяющиеся сообщения, однако почти не захватывает другие виды спама, а высокая частота ложных срабатываний ограничивает его

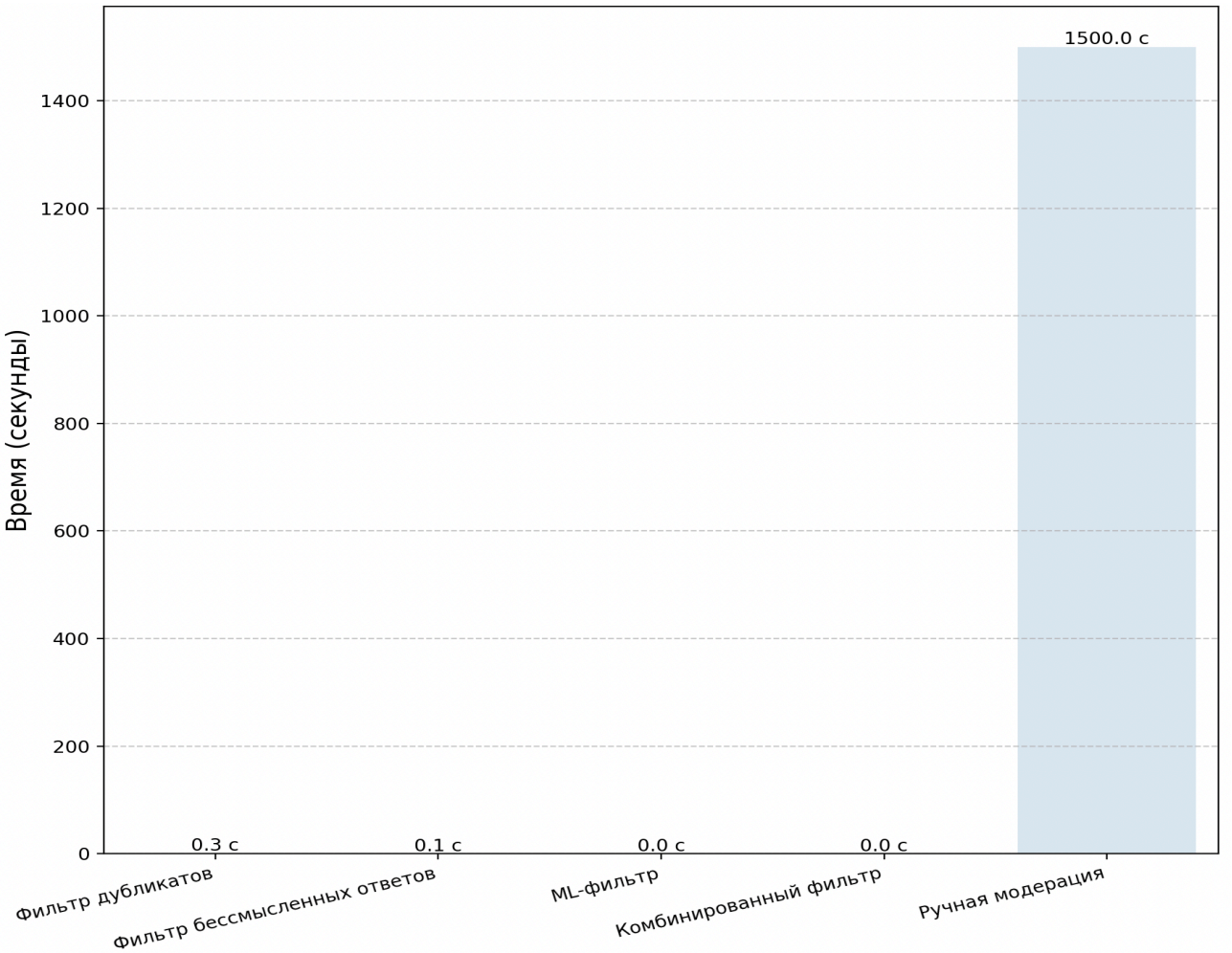


Рисунок 4.3 – График затраченного времени на обработку данных

самостоятельную ценность для задачи фильтрации. Фильтр бессмысленных ответов крайне строг, и его следует использовать для систем с критическим уровнем осторожности к спаму, так как высока вероятность ложноположительных срабатываний. По метрикам можно заключить, что комбинированный фильтр является компромиссным решением для большинства сценариев.

Наилучший баланс достигается для ML-фильтра: почти полное отсутствие как ложных срабатываний, так и пропусков. Несмотря на то, что ручная модерация позволяет наиболее полно учесть нюансы языка и контекста, она почти не может конкурировать с машинным подходом по времени и затратам.

По итогам анализа видно, что использование фильтра на основе машинного обучения позволяет получить наилучший баланс между качеством фильтрации и экономией ресурсов — этот подход минимизирует необходимость ручного труда и допускает обработку больших потоков данных без существенной потери точности. Если фильтрация требует не только масштабируемости,

но и надежности принятия решений, можно выносить сомнительные результаты работы ML-фильтра на ручную проверку.

4.4 Эксперименты по кластеризации: качество группировки, примеры

Данный эксперимент ставил целью выявление оптимального метода для интеграции в разрабатываемое программное обеспечение, способного автоматически группировать ответы респондентов по смысловым кластерам. В качестве объектов сравнения выступили три алгоритма: k-means, DBSCAN и иерархическая кластеризация, настроенные на формирование пяти кластеров для обеспечения сопоставимости результатов.

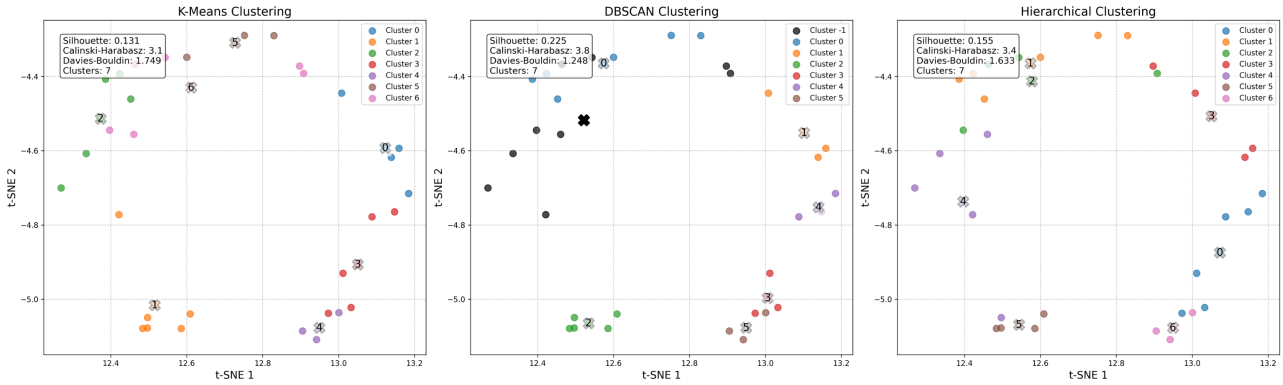


Рисунок 4.4 – Результаты работы алгоритмов на плоскости

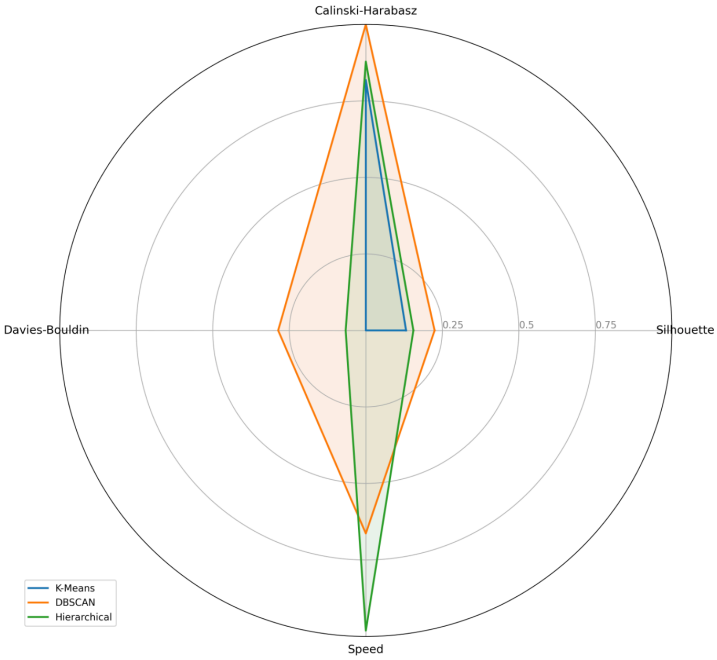


Рисунок 4.5 – Радар сравнения алгоритмов

По графику на рисунке 4.5 видно, что наилучших значений по базовым метрикам добился алгоритм DBSCAN, хотя и был медленнее иерархического алгоритма.

В k-means кластеры получились сравнительно равномерными по размеру, самые крупные из них содержат по 6 текстов, наименьшие — по 3 текста. Тематический анализ слов внутри кластеров показывает значительное тематическое разнообразие.

DBSCAN также выделил 7 кластеров, однако значительно отличается характер работы с "шумовыми" элементами: кластер -1 содержит почти 30% отзывов, которые не были отнесены ни к одной из чётких групп — метод детектирует и оставляет в стороне разнородные, неструктурированные комментарии. Остальные кластеры довольно сбалансированы и тематически раскрыты. Внутри кластеров сохраняется высокая тематическая однородность, а лишние, не объединяемые по смыслу тексты, корректно выводятся за скобки группировки.

Иерархическая кластеризация также выявила 7 сравнительно равномерных по размеру групп. Тематики, найденные в каждом кластере, во многом повторяют находки предыдущих методов.

Результаты эксперимента подтвердили, что DBSCAN сохраняет устойчивость к разнородности текстовых ответов, автоматически выделяя смысловые группы без предварительного задания числа кластеров. Что критично для автоматизированной системы, где ручная настройка параметров нежелательна.

Рассмотренные примеры отзывов из разных кластеров наглядно показывают, что качественная тематическая группировка достигается всеми тремя подходами — при этом наилучшее разделение и минимальная внутрикластерная размытость фиксируется у DBSCAN.

4.5 Эксперименты автоматической оценки: метрики качества

В ходе экспериментов была проведена сравнительная оценка различных методов автоматической проверки текстовых ответов на опросы. Для анализа были выбраны четыре подхода: регрессионный метод с L2-регуляризацией (MLEvaluator_Ridge), метод на основе случайного леса (MLEvaluator_RF), правило-ориентированный метод (RuleEvaluator) и метод семантической оценки (SemanticEvaluator). Качество всех методов оценивалось по ряду метрик, отражающих, как точность приближения к истинным баллам, так и степень корреляции с реальными оценками человека.

В качестве основных критериев для анализа эффективности использовались RMSE и MAE. Также вычислялся коэффициент детерминации R^2,

показывающий степень объяснённости моделью изменчивости исходных данных; значение, близкое к единице, означает успешную модель, отрицательные значения — неудачное приближение. Для более детального анализа использовались коэффициенты корреляции Пирсона и Спирмена, которые позволяют судить о линейной и монотонной связи между истинными и предсказанными результатами, и процентное количество совпадений предсказанных баллов с эталонными с точностью до одного балла. Временные характеристики включали скорость обучения модели (для методов машинного обучения) и время непосредственно оценки.

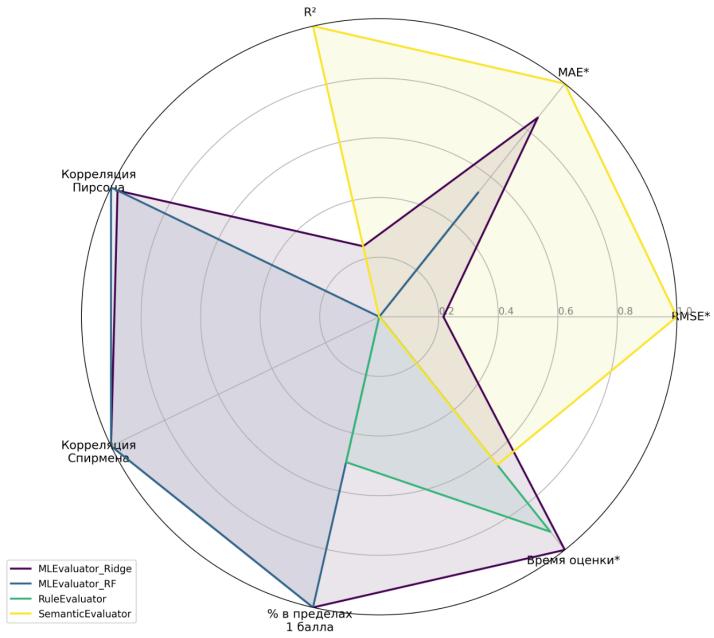


Рисунок 4.6 – Сравнение метрик оценщиков

По диаграмме на рисунке 4.6 можно увидеть, что наименьшие значения RMSE и MAE, а также наилучшее значение R^2 получены у метода RuleEvaluator. Это свидетельствует о более точном приближении данного метода к экспертным оценкам по сравнению с другими участниками эксперимента. В свою очередь, максимальные значения корреляции Пирсона и Спирмена отмечаются для MLEvaluator_RF и MLEvaluator_Ridge соответственно, что говорит о линейной и монотонной согласованности их предсказаний с реальными баллами. RuleEvaluator заметно уступает во всех показателях качества, а его уровень совпадений с правдой по одной балльной градации вдвое ниже, чем у методов на основе машинного обучения, несмотря на выигрыш в скорости и прозрачности принципа работы. SemanticEvaluator полностью опирается на сопоставление смысловой близости ответа с референсным, что подтверждает значимость учета семантики при автоматической проверке развернутых текстовых ответов, хотя, ввиду специфики данных тестирования, не все метрики корреляции для этого метода были вычислены.

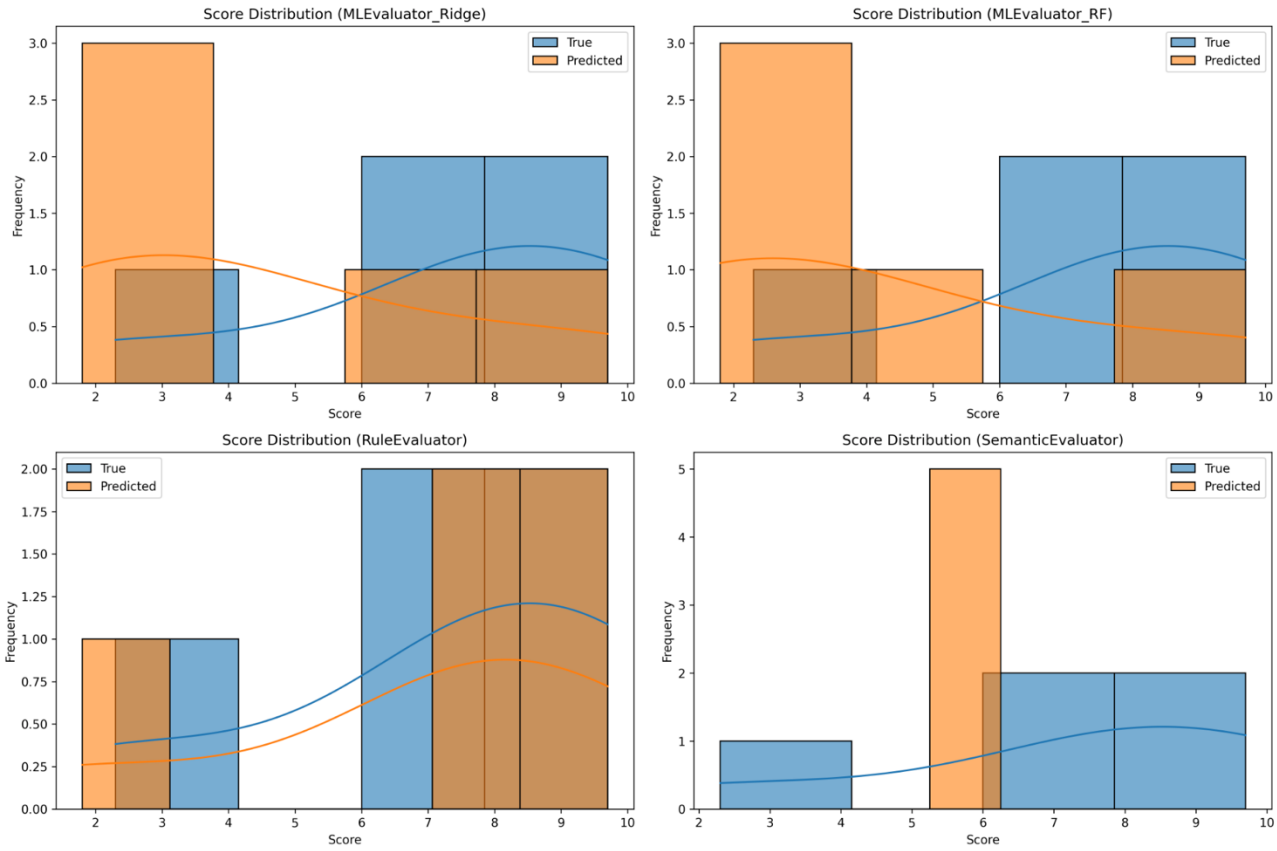


Рисунок 4.7 – Распределение предсказанных оценок

Графики распределения баллов, представленные на рисунке 4.7, для каждого метода наглядно демонстрируют различие типов и характера ошибок. Наибольшее совпадение в формах распределения наблюдается для методов на машинном обучении.

MLEvaluator_Ridge характеризуется хорошей скоростью, устойчивостью к переобучению, но меньшей гибкостью по сравнению с нелинейными моделями, и требует наличия разметки данных. MLEvaluator_RF обеспечивает высокую точность и умеет работать с более сложными зависимостями в текстах, но может быть медленнее, требует больше памяти и объяснимость результатов понижается. RuleEvaluator не нуждается в обучении и отличается прозрачными правилами работы, однако имеет ограниченную точность из-за невозможности учёта смысловой стороны. SemanticEvaluator улавливает именно семантическую близость и не требует больших обучающих выборок, хотя его качество зависит от хорошо подобранных референсных ответов.

По итогам эксперимента можно отметить, что семантический подход обеспечивает наилучшее приближение к истинным экспертным оценкам по большинству метрик точности. При наличии хорошо размеченных данных и необходимости массовой быстрой прогонки ответов оправдано применение методов машинного обучения, особенно если требуется объяснимый результат и быстрое время работы. Если интересует высокая интерпретируемость и

прозрачность процесса — можно использовать RuleEvaluator, однако при этом точность страдает.

ВЫВОДЫ

1. В ходе разработки реализован программный прототип системы, включающий модули фильтрации, кластеризации и автоматической оценки текстовых данных.

2. Для полноценного обучения и тестирования системы выполнена тщательная подготовка, очистка и разметка тестовой выборки.

3. Эксперименты продемонстрировали, что автоматизация процессов обработки текстовых данных существенно сокращает затраты времени и повышает воспроизводимость и объективность анализа.

4. В рамках задачи лучше всего использовать фильтры на основе машинного обучения, DBSCAN и комбинацию MLEvaluator_RF и RuleEvaluator оценщиков.

ЗАКЛЮЧЕНИЕ

Проведенное исследование позволило разработать программное обеспечение для автоматизированной обработки текстовых ответов в опросах, решающее ключевые задачи фильтрации спама, кластеризации и оценки информативности данных. Результаты работы подтвердили гипотезу о том, что комбинация современных методов машинного обучения и лингвистического анализа обеспечивает высокую точность и эффективность обработки неструктурированных текстовых данных. Было установлено, что гибридные подходы, сочетающие формальные и семантические критерии, наиболее адаптивны для задач массовых опросов. Теоретическая часть работы подчеркнула важность интеграции лингвистических моделей с алгоритмами машинного обучения, что стало методологической основой для проектирования системы.

Экспериментальная часть работы показала, что алгоритм DBSCAN, благодаря автоматическому определению числа кластеров и устойчивости к шуму, обеспечивает наилучшее качество группировки текстов. Машинные методы фильтрации спама на основе случайного леса сократили время обработки данных при минимальной потере точности (8%), а семантическая оценка ответов через сравнение с эталонами достигла минимальной RMSE (3.094), доказав важность учета контекстной близости.

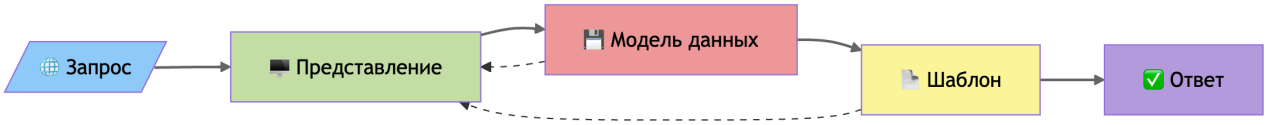
Практическая значимость работы заключается в сокращении временных затрат на анализ опросов и повышении объективности интерпретации результатов за счет минимизации человеческого фактора. Внедрение системы в образовательные и маркетинговые проекты позволит выявлять скрытые паттерны в ответах респондентов, оптимизировать процессы принятия решений и снизить операционные издержки. Однако текущие ограничения, такие как зависимость от ручной разметки обучающих данных и необходимость тонкой настройки эталонов для семантической оценки, требуют дальнейшей доработки.

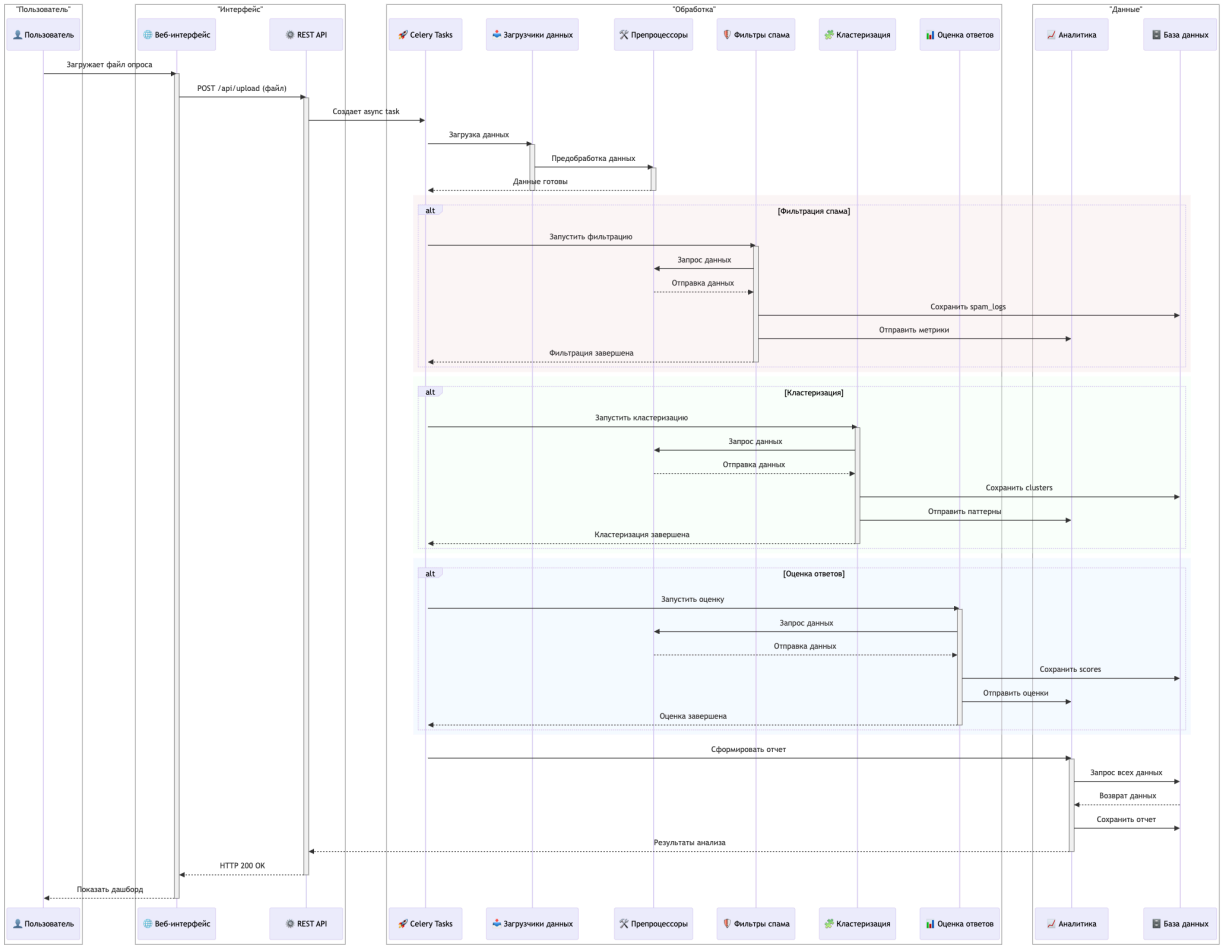
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

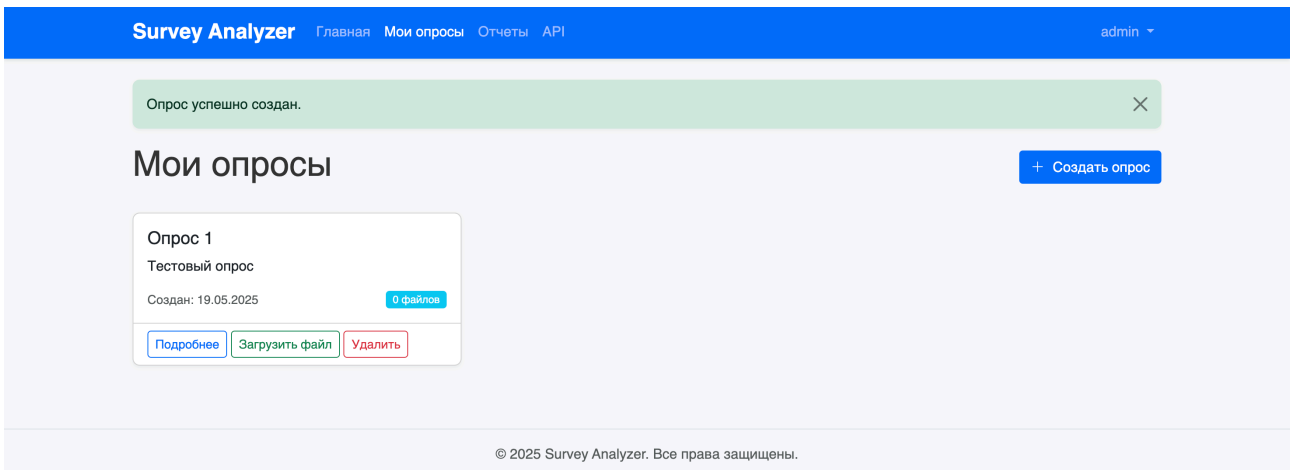
1. Хобсон Лейн. Обработка естественного языка в действии = Natural Language Processing in Action / Л. Хобсон, Х. Ханнес, Х. Коул; [пер. с англ. И. Пальти, С. Черников]. – Санкт-Петербург [и др.]: Питер, 2020. – 575 с.
2. Blei, D. Latent Dirichlet Allocation / D. M. Blei, A.Y. Ng, M. I. Jordan // J. of Mach. Learn. Research. – 2003. – Vol. 3, iss. 5. – P. 993 – 1022.
3. Chomsky, N. Syntactic structures / N. Chomsky. – Hague : Mouton & Co, 1957.
4. Devlin, J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [et al.] // Minneapolis : Association for Computational Linguistics. –2019. – DOI: 10.18653/v1/N19-1423.
5. Jurafsky, D. Speech and Language Processing / D. Jurafsky, J. H. Martin. – New Jersey : Upper Saddle River, 2000. – Vol. 20. – P. 640–644. – Mode of access: https://djvu.online/file/biEvmf6idHpa7. – Date of access: 20.05.2025.
6. Jurafsky, D. Speech and Language Processing (3rd ed. draft) / D. Jurafsky, J. H. Martin. // Stanford University, 2025. – Vol. 14. – P. 3–13. – Mode of access: https://web.stanford.edu/~jurafsky/slp3/14.pdf. – Date of access: 20.05.2025.
7. Nair, P. S. Automated Essay Scoring for IELTS using Classification and Regression Models / P. S. Nair; T. N. Malleswari // IEEE 2024 : 3rd World Conference on Applied Intelligence and Computing (AIC), Gwalior, 27–28 Jul. 2024. – DOI: 10.1109/AIC61668.2024.10731045.
8. Rethinking Attention with Performers: published as a conference paper at ICLR 2021, Vienna, 4–8 May / Univ. of Cambridge ; ed.: V. Likhosherstov [et al.]. – Cambridge : Univ. of Cambridge, 2021. – Mode of access: https://openreview.net/pdf?id=Ua6zuk0WRH. – Date of access: 20.05.2025.
9. Schuster, T. Cross-Lingual Alignment of Contextual Word Embeddings, with Applications to Zero-shot Dependency Parsing / T. Schuster [et al.] // Minneapolis : Association for Computational Linguistics. – 2019. – DOI: 10.18653/v1/N19-1162.

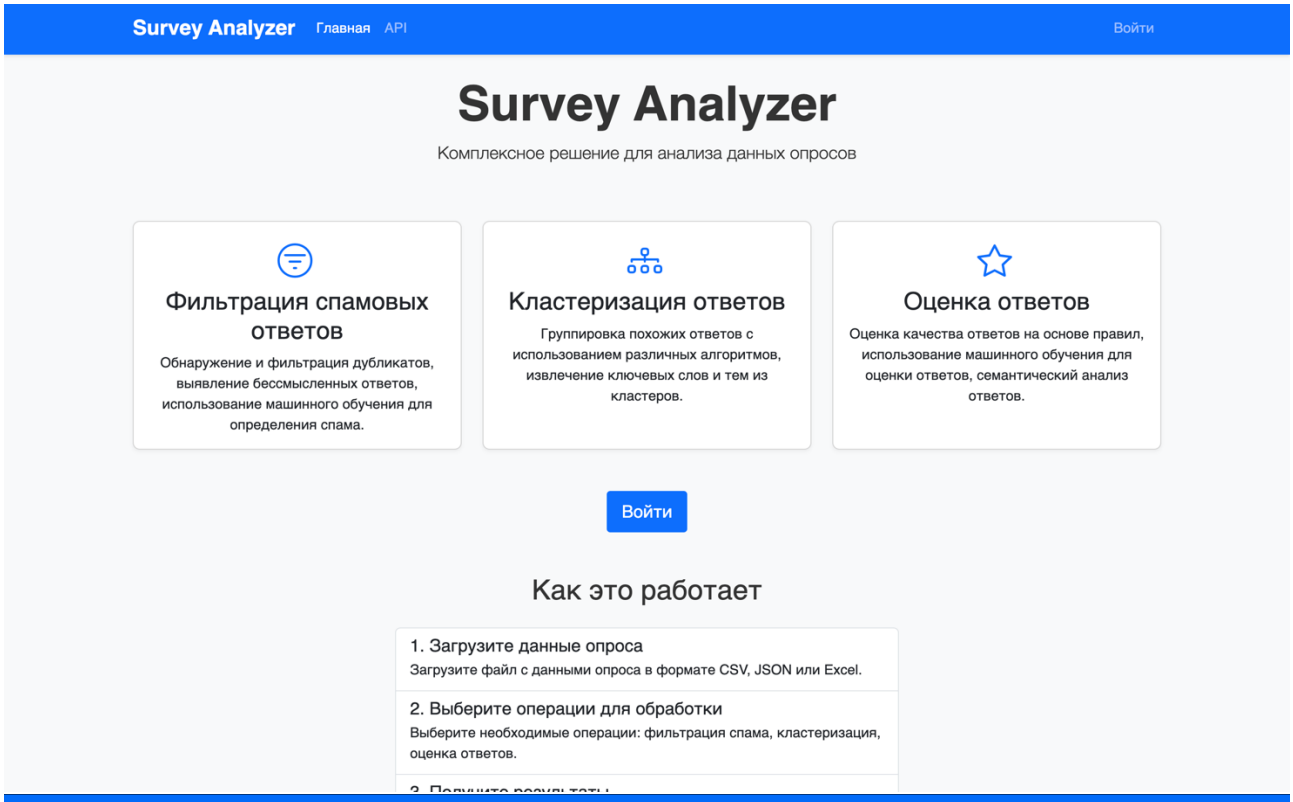
ПРИЛОЖЕНИЕ А

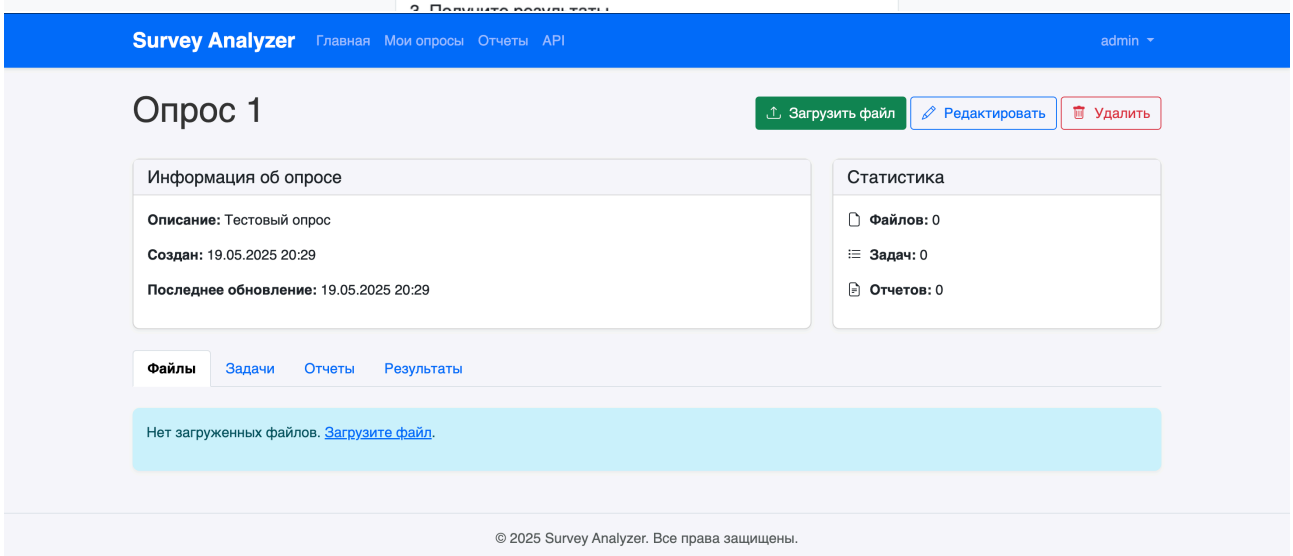
ПРОТОТИП ПРОГРАММЫ ДЛЯ ОБРАБОТКИ ОТКРЫТЫХ ОТВЕТОВ В ОПРОСАХ

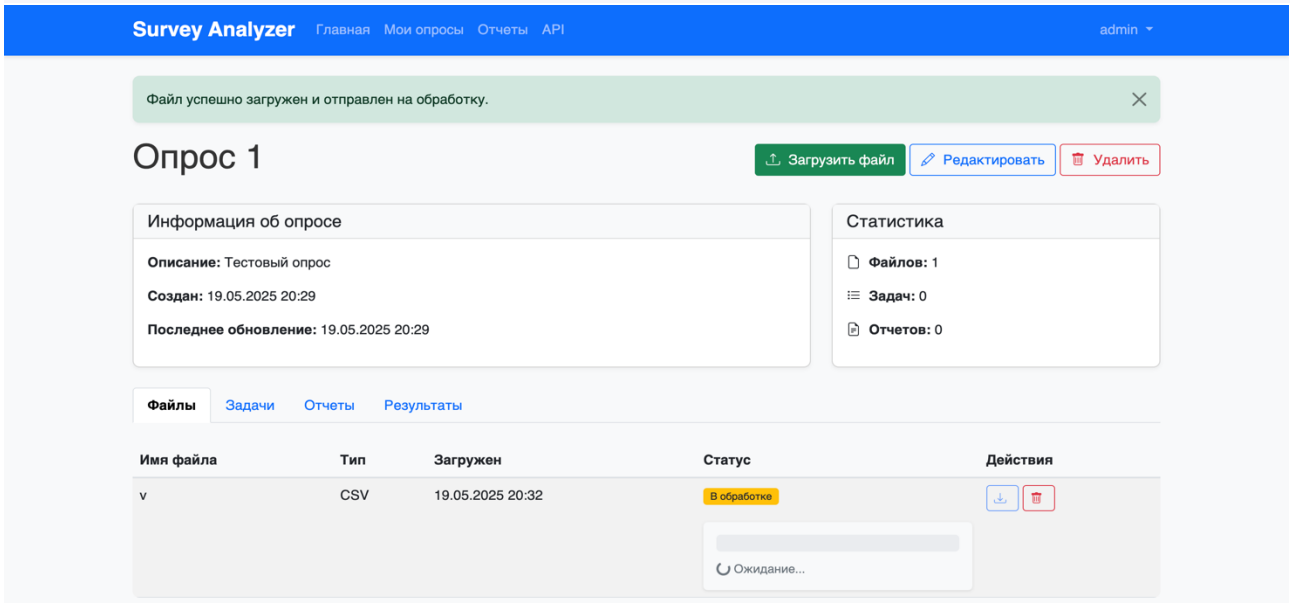


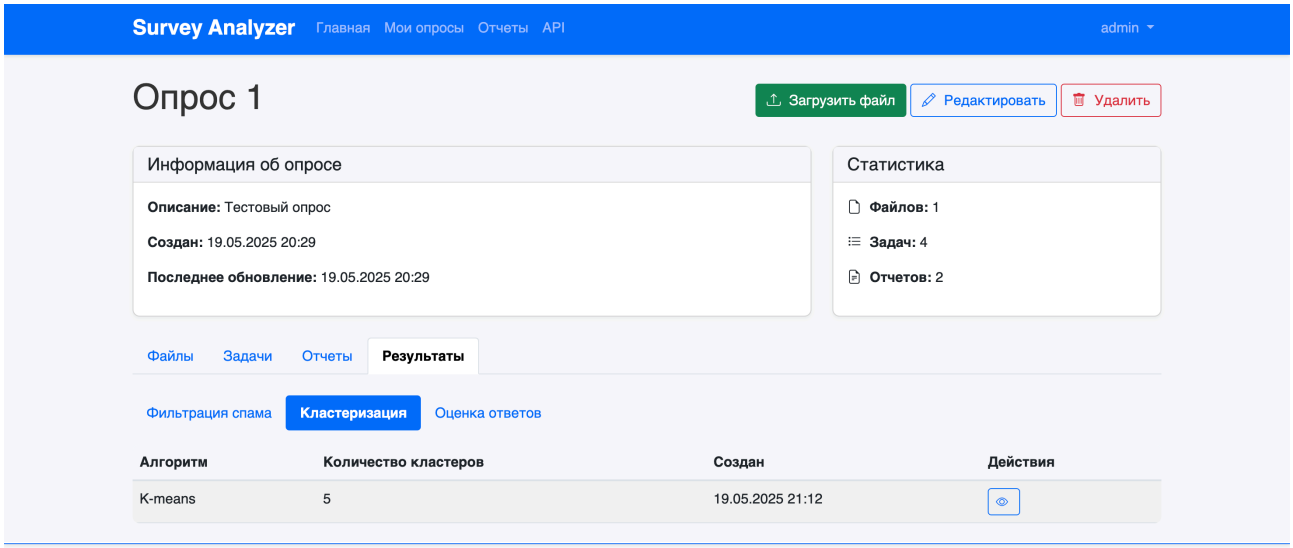


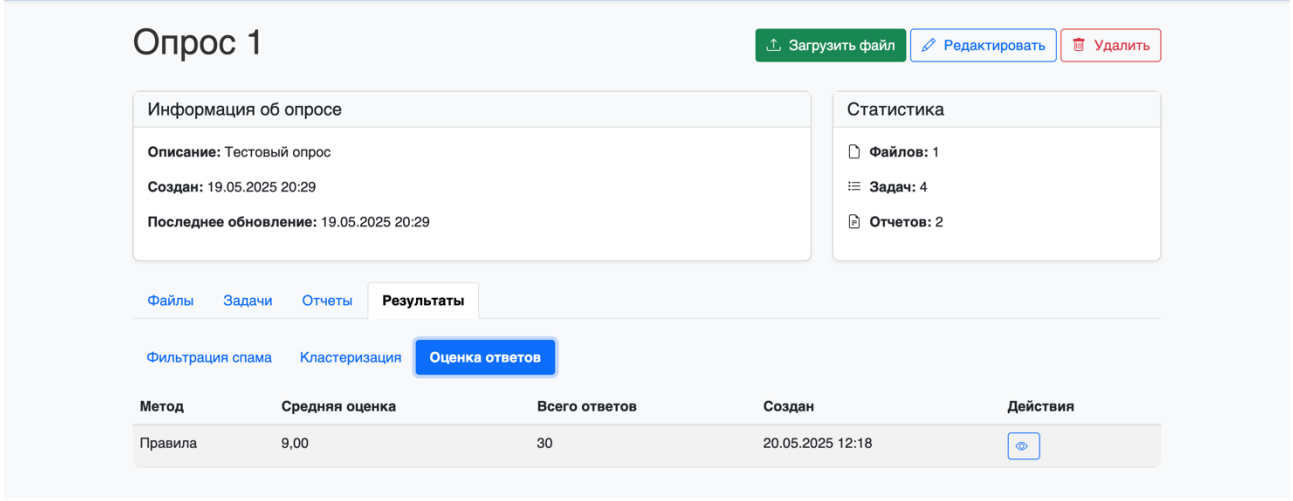


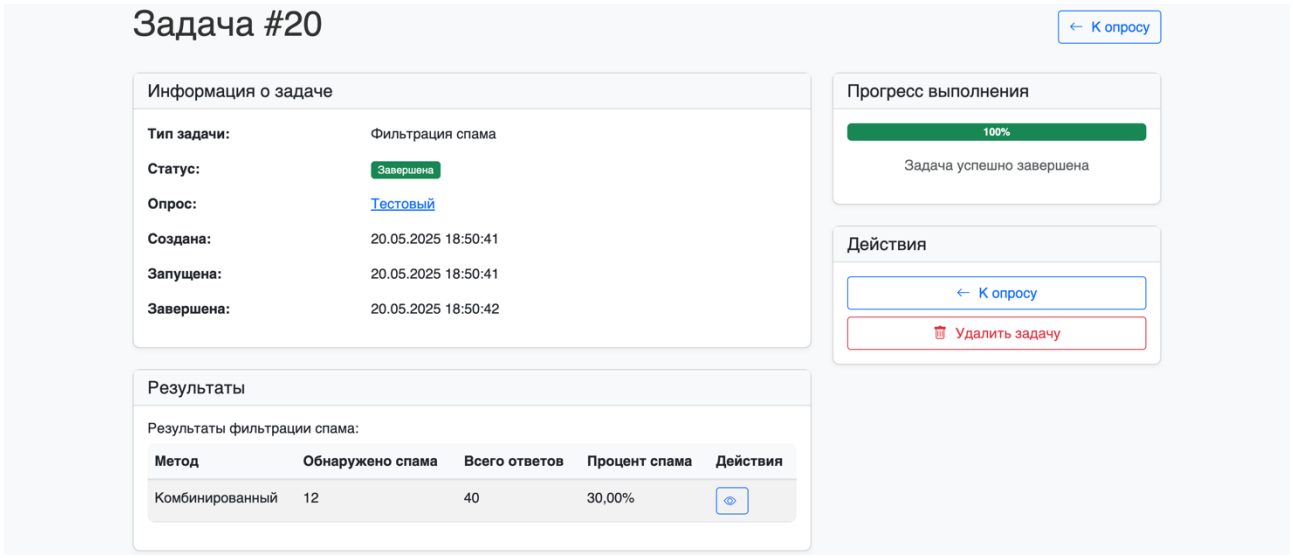

Survey Analyzer
Главная
API
Войти
Survey Analyzer
Комплексное решение для анализа данных опросов
Фильтрация спамовых ответов
Обнаружение и фильтрация дубликатов, выявление бессмысленных ответов, использование машинного обучения для определения спама.
Кластеризация ответов
Группировка похожих ответов с использованием различных алгоритмов, извлечение ключевых слов и тем из кластеров.
Оценка ответов
Оценка качества ответов на основе правил, использование машинного обучения для оценки ответов, семантический анализ ответов.
Войти
Как это работает
1. Загрузите данные опроса
Загрузите файл с данными опроса в формате CSV, JSON или Excel.
2. Выберите операции для обработки
Выберите необходимые операции: фильтрация спама, кластеризация, оценка ответов.

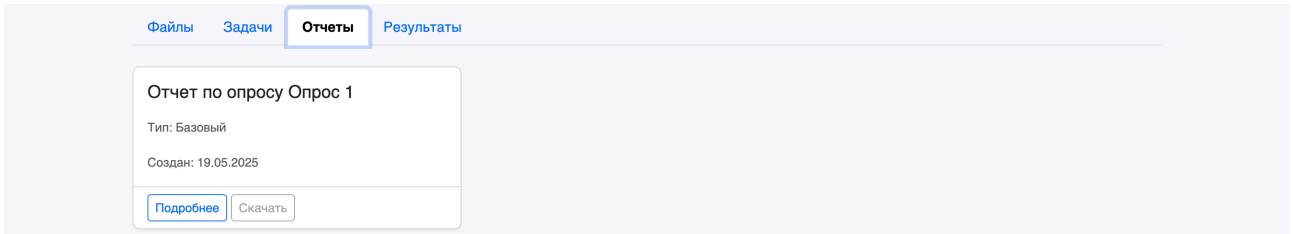

Survey Analyzer
Главная
Мои опросы
Отчеты
API
admin
Опрос 1
Загрузить файл
Редактировать
Удалить
Информация об опросе
Описание: Тестовый опрос
Создан: 19.05.2025 20:29
Последнее обновление: 19.05.2025 20:29
Статистика
Файлов: 0
Задач: 0
Отчетов: 0
Файлы
Задачи
Отчеты
Результаты
Нет загруженных файлов. Загрузите файл.
© 2025 Survey Analyzer. Все права защищены.


Survey Analyzer
Главная
Мои опросы
Отчеты
API
admin
Файл успешно загружен и отправлен на обработку.
Опрос 1
Загрузить файл
Редактировать
Удалить
Информация об опросе
Описание: Тестовый опрос
Создан: 19.05.2025 20:29
Последнее обновление: 19.05.2025 20:29
Статистика
Файлов: 1
Задач: 0
Отчетов: 0
Файлы
Задачи
Отчеты
Результаты
Имя файла
Тип
Загружен
Статус
Действия
v
CSV
19.05.2025 20:32
В обработке
Ожидание...


Survey Analyzer
Главная
Мои опросы
Отчеты
API
admin
Опрос 1
Загрузить файл
Редактировать
Удалить
Информация об опросе
Описание: Тестовый опрос
Создан: 19.05.2025 20:29
Последнее обновление: 19.05.2025 20:29
Статистика
Файлов: 1
Задач: 4
Отчетов: 2
Файлы
Задачи
Отчеты
Результаты
Фильтрация спама
Кластеризация
Оценка ответов
Алгоритм
Количество кластеров
Создан
Действия
K-means
5
19.05.2025 21:12


Опрос 1
Загрузить файл
Редактировать
Удалить
Информация об опросе
Описание: Тестовый опрос
Создан: 19.05.2025 20:29
Последнее обновление: 19.05.2025 20:29
Статистика
Файлов: 1
Задач: 4
Отчетов: 2
Файлы
Задачи
Отчеты
Результаты
Фильтрация спама
Кластеризация
Оценка ответов
Метод
Средняя оценка
Всего ответов
Создан
Действия
Правила
9,00
30
20.05.2025 12:18


Задача #20
К опросу
Информация о задаче
Тип задачи: Фильтрация спама
Статус: Завершена
Опрос: Тестовый
Создана: 20.05.2025 18:50:41
Запущена: 20.05.2025 18:50:41
Завершена: 20.05.2025 18:50:42
Прогресс выполнения
100%
Задача успешно завершена
Действия
К опросу
Удалить задачу
Результаты
Результаты фильтрации спама:
Метод
Обнаружено спама
Всего ответов
Процент спама
Действия
Комбинированный
12
40
30,00%


Файлы
Задачи
Отчеты
Результаты
Отчет по опросу Опрос 1
Тип: Базовый
Создан: 19.05.2025
Подробнее
Скачать