Пермский филиал федерального государственного автономного
образовательного учреждения высшего образования
Национальный исследовательский университет
«Высшая школа экономики»

Факультет социально-экономических и компьютерных наук

Соломатин Роман Игоревич

РАЗРАБОТКА САЙТА ДЛЯ АВТОМАТИЧЕСКОГО СБОРА, АНАЛИЗА И ВИЗУАЛИЗАЦИИ ИНФОРМАЦИИ ПО ЭТИЧНОСТИ КОМПАНИЙ

Выпускная квалификационная работа

студента образовательной программы «Программная инженерия»
по направлению подготовки 09.03.04 Программная инженерия

Руководитель
к.т.н., доцент кафедры
информационных технологий в
бизнесе НИУ ВШЭ-Пермь

А. В. Бузмаков

Пермь, 2023 год

Аннотация

В данной работе проведен анализ этичности разных компаний.

В первой главе находится описание используемых алгоримов.

Во второй главе представлено проектирование системы.

В третьей главе представлена реализация системы.

В четвертой главе представлено тестирование работы системы.

Количество страниц – 33, количество иллюстраций – 12, количетсво таблиц – 18.

Оглавление

Введение

Этика компаний – это разделяемые всеми сотрудниками организации правила и нормы, ценности и убеждения, манера общения и другие факторы, которые регламентируют поведение и взаимодействии членов компании. Существует 3 уровня этики компаний[1]:

1. мировой – отвечает за увеличение общественного благосостояния, обеспечение рабочих мест, научно-технические инновации и модернизацию производственных процессов и т. д.
2. макроуровень – отвечает за принципы рыночной конкуренции, информационной прозрачность и равнодоступности для всех участников рынка и т. д.
3. микроуровне – отвечает за доверие и отсутствие дискриминации в отношениях между контрагентами, между сотрудниками и менеджерами, морально-нравственный климат в организации и т. д.

В данной работе будет рассматриваться этика на микроуровне.

Этичность компаний уже давно вызывает озабоченность, особенно их поведение в спорных ситуациях и предоставление услуг, ориентированных на клиента. В последние годы все большее внимание уделяется оценке этичности компаний[2, 3, 4], особенно в банковском секторе и через призму экологических, социальных и управленческих факторов (ESG). Необходимость в таких оценках становится все более острой по мере того, как общество продолжает бороться с последствиями неправомерных действий корпораций и более широким воздействием корпоративной деятельности на общество и окружающую среду.

В настоящее время существует несколько сервисов, которые призваны оценивать этику компании на основании финансовых показателей[1] и судебных дел[2]. Это привело к ситуации, когда отдельные лица должны проводить свои собственные исследования, чтобы определить насколько этична компания. Это часто включает в себя просмотр отзывов с различных веб-сайтов, что может занять много времени и не всегда может

[1] `https://kontur.ru/expert`, `https://www.esphere.ru/products/spk/financial`

[2] `https://proverki.gov.ru/portal/public-search`

дать исчерпывающую или точную картину, так как не включает в себя качество обслуживания.

Для решения этой проблемы реализована система, которая собирает и анализирует отзывы потребителей с различных веб-сайтов, чтобы дать более полную и точную оценку этической практики компании. Затем собранные данные анализируются с помощью различных методов, таких как обработка естественного языка и машинного обучения, для выявления закономерностей и тенденций, связанных с этической практикой компании. Полученный анализ может быть использован для разработки более надежной и достоверной системы оценки этичности компаний.

Объект исследования – взаимодействие компаний с клиентами.

Предмет исследования – программные средства для оценки этичности на основе взаимодействия компаний с клиентами.

Цель работы – создание системы для оценки этичности компаний.

Исходя из поставленной цели, необходимо:

1. Провести анализ предметной области и требований
2. Реализовать систему
3. Провести тестирование системы

Этап анализа должен:

1. Анализ предметной области
2. Анализ требований к системе
3. Анализ существующих алгоритмов

Этап проектирования должен включать:

1. Проектирование серверной части
2. Проектирование модели для определения этичности
3. Проектирование клиентской части приложения

Этап реализации должен включать:

1. Описание сбора данных
2. Реализации модели
3. Реализации серверной части
4. Реализации клиентской части

Этап тестирования должен включать:

1. Тестирование модели
2. Тестирование серверной части
3. Тестирование клиентской части

В ходе выполнения анализа, проектирования и реализации приложения используется объектно-ориентированный подход. Результаты анализа и решения задач проектирования формализуются с помощью диаграмм `UML`. При разработке базы данных используется реляционная СУБД `PostgreSQL`, а серверная часть приложения реализуется на языке python с помощью фреймворка `FastApi`, а алгоритмы анализы текста будут использовать методы машинного обучения.

Глава 1 Анализ предметной области

В данной главе представлен аналитический обзор оценок этичности компаний и алгоритмов машинного обучения, а также обзор существующих программных решений для поставленной проблемы.

Анализ предметной области следует разделить на следующие пункты:

1. анализ процесса определения этичности компаний сейчас позволяет понять, как этот процесс сейчас происходит и как его лучше всего автоматизировать;
2. анализ оценок этичности компаний для того, чтобы в дальнейшем определить этичность компаний;
3. анализ существующих решений выполняется с целью выделения их сильных и слабых сторон по отношению к решаемой проблеме и обоснования необходимости разработки нового средства, подходящего под регламент задач;
4. анализ алгоритмов позволяет понять с помощью каких алгоритмов можно найти полезную информацию в текстах;
5. анализ требований к системе позволит выделить функциональные и не функциональные требования.

1.1. Анализ определения этичности компании

Сейчас процесс поиска этичной компании выгладит следующим образом: сначала ищутся компании, которые предоставляют желаемые услуги. Далее они изучаются, чтобы определить их этичность. Этот процесс включает в себя:

1. просмотр отчетности компании
2. анализ ее финансовой деятельности
3. изучение информации о социальной ответственности

Для этого они обращаются к различным источникам информации, таким как веб-сайты компаний, рейтинговые агентства, исследовательские организации и другие источники. Потом, изучаются социальные сети компании или отзывы пользователей на разных сайтах, форумах и социальных сетях, чтобы получить дополнительную информацию и оценить общее мнение о компании. После изучения каждой компании люди

выбирают ту, которую они считают наиболее этичной и социально ответственной. Блок-схема данного поиска рис. 1.1. Важным фактором для определения этичности компании может быть ее социальная ответственность, устойчивость бизнеса и соблюдение норм и стандартов в области финансовой деятельности.

В целом, процесс поиска компаний и определения их этичности может быть длительным и требует серьезного подхода. Люди могут использовать различные источники информации, чтобы сделать осознанный выбор и инвестировать свои деньги в компанию, которая соответствует их ожиданиям и требованиям.

Рисунок 1.1 – Диаграмма того, как сейчас происходит поиск компании

1.2. Анализ оценок этичности компаний

Оценка этики компании – это не одноразовый процесс, а скорее непрерывная попытка понять и оценить действия, политику и практику компании с течением времени. Это включает в себя рассмотрение соблюдения компанией отраслевых этических стандартов и передовой практики, а также мониторинг любых изменений в этической

позиции компании с течением времени. Кроме того, участие в диалоге с компанией и консультации с организациями, специализирующимися на оценке корпоративной ответственности могут дать ценную информацию об этических практиках компании.

Компаниям важно оставаться этичными, так как на долгосрочной перспективе это приносит большую прибыль и улучшает показатели бизнеса, чем неэтичный способ ведение бизнеса[5, 2]. Насколько этична компания можно рассматривать с двух сторон, самой компании и их клиентов. Со стороны компаний можно выделить факторы, которые можно получить из их отчетности:

- количество капитала, чтобы они не могли обанкротиться;
- какое влияние они вносят на окружающую среду;
- куда идут инвестиции[6].

Для пользователей одними из ключевых факторов можно выделить:

- качество пользовательского сервиса[7], как правило пользователи оставляют отзывы на сайтах по 5-ти бальной шкале;
- насколько навязчивые услуги компании[8], как правило пользователи оставляют отзывы на сайтах по 5-ти бальной шкале.

В данной работе этичность компаний будет определяться по отзывам клиентов, которые освещают проблемы качества услуг и качество сервиса, и на основе отчетности компаний, что позволит полностью осветить проблему. Для анализа текстов будут использоваться алгоритмы машинного обучения.

1.3. Анализ существующих решений

Существует несколько индексов, предназначенных для измерения этичности – индекс Доу Джонса (DJSI)[9] и FTSE4GOOD[10].

DJSI оценивает показатели устойчивости компаний различных секторов на основе экономических, экологических и социальных критериев. Компании отбираются на основе их показателей по сравнению с аналогичными компаниями в том же секторе. Процесс оценки включает в себя тщательную оценку компаний по различным критериям, включая корпоративное управление, экологический менеджмент, трудовую практику, права человека и социальные вопросы.

Аналогичным образом, индекс FTSE4GOOD предназначен для оценки деятельности компаний, которые демонстрируют эффективную практику экологического, социального и управленческого менеджмента (ESG). Компании отбираются на основе их практики ESG и оцениваются по различным критериям, включая изменение климата, права человека и корпоративное управление.

Индексы DJSI и FTSE4GOOD разработаны для того, чтобы помочь инвесторам определить компании, которые привержены этической практике. Эти индексы предоставляют инвесторам стандартизированный способ сравнения компаний на основе их показателей. Это помогает инвесторам принимать более обоснованные инвестиционные решения и побуждает компании внедрять устойчивую практику для привлечения инвестиций.

Для российских компаний нет аналогичных индексов. Сейчас данные об этичности компаний можно получить из агрегаторов отзывов и отчётности. Агрегаторы позволяют собрать информацию о клиентском обслуживании, а отчетность компаний о положении дел в целом. Но сейчас не существует способов, как можно оценить все вместе.

1.4. Алгоритмы для анализа текста

Алгоритмы машинного обучения для анализа текста получили широкое распространение для извлечения информации из неструктурированных данных с помощью больших помеченных наборов данных. Среди различных используемых методов несколько алгоритмов оказались особенно эффективными в этой области. К ним относятся мешок слов[11], TF-IDF[12], Word2Vec[13], ELMO[14], GPT[15] и BERT[16]. Каждый из этих алгоритмов обладает уникальными характеристиками, которые делают их хорошо подходящими для определенных приложений.

Модель «Мешок слов» представляет текстовые данные путем присвоения уникального номера каждому слову в документе. Этот метод прост в реализации, но не учитывает порядок слов в предложении. С другой стороны, модель TF-IDF представляет текстовые данные, учитывая как частоту слова в документе (TF), так и его редкость во всех документах корпуса (IDF). Этот подход может быть использован для опреде-

ления важности слова в данном документе и обычно используется в задачах поиска информации и обработки естественного языка, но он не понимает контекста слов.

Word2Vec использует векторное представление слов, что позволяет алгоритму улавливать значение слов в сходных контекстах. Это позволяет более точно и изощренно представлять взаимосвязи между словами, что приводит к повышению производительности в таких задачах, как классификация текста и анализ настроений.

ELMO, GPT и BERT, с другой стороны, основаны на архитектуре трансформеров, в которой каждое предложение представлено вектором чисел, обычно известным как вложение. Такое представление позволяет получить более полное и целостное понимание текста, поскольку оно учитывает контекст всего предложения или текста.

Из этих алгоритмов BERT считается наиболее продвинутым и мощным, поскольку он способен учитывать контекст всего предложения или текста, в то время как GPT и ELMO рассматривают только односторонний контекст. Это позволяет BERT достигать самых современных результатов в широком спектре задач анализа естественного языка.

Таблица результата сравнения моделей 1.1.

Таблица 1.1 – Сравнение моделей

Модель	Вектор слов	Контекст
Мешок слов	зависит от количества слов	нет
TF-IDF	зависит от количества слов	очень слабо
Word2Vec	не зависит от количества слов	слабо
ELMO	не зависит от количества слов	однонаправленный
GPT	не зависит от количества слов	однонаправленный
BERT	не зависит от количества слов	двунаправленный

1.4.1. BERT

BERT [16] (Bidirectional Encoder Representations from Transformers) – это нейросетевая языковая модель, которая относится к классу трансформеров. Она состоит из 12 «базовых блоков» (слоев), а на каждом слое 768 параметров.

На вход модели подается предложение или пара предложений. Затем разделяется на отдельные слова (токены). Потом в начало последовательности токенов вставляется специальный токен `[CLS]`, обозначающий начало предложения или начало последовательности предложений. Пары предложений группируются в одну последовательность и разделяются с помощью специального токена `[SEP]`, затем к каждому токену добавляется эмбеддинг, показывающий к какому предложению относится токен. Потом все токены превращаются в эмбеддинги 1.2 по механизму описаному в работе [17].

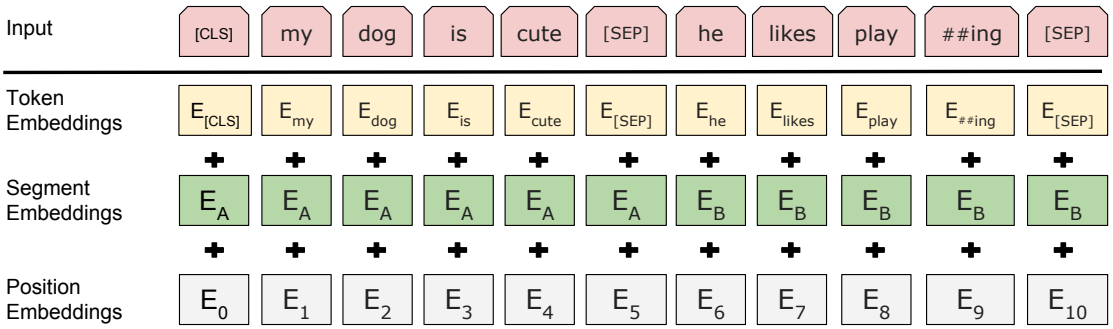

Рисунок 1.2 – Пример ввода текста в модель

При обучении модель выполняет на 2 задания:

1. Предсказание слова в предложении

   Поскольку стандартные языковые модели либо смотрят текст слева направо или справа налево 1.3, как ELMo[14] и GPT[15], они не подходят под некоторые типы заданий. Так как BERT двунаправленный, у каждого слова можно посмотреть его контекст, что позволит предсказать замаскированное слово.

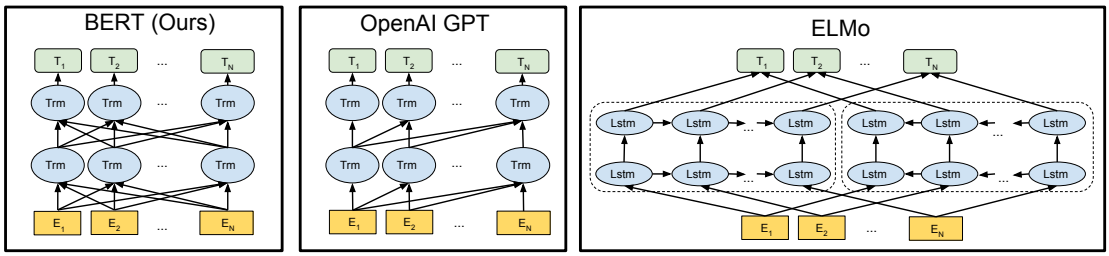

Рисунок 1.3 – Сравнение принципов работы BERT, ELMo, GPT

   Это задание обучается следующим образом – 15% случайных слов заменяются в каждом предложении на специальный токен `[MASK]`, а затем предсказываются на основании контекста. Однако иногда слова заменяются не

на специальны токена, в 10% заменяются на случайный токен и еще в 10% заменяются на случайное слово.

2. Предсказание следующего предложения

   Для того чтобы обучить модель, которая понимает отношения предложений, она предсказывает, идут ли предложения друг за другом. Для этого с 50% вероятностью выбирают предложения, которые находятся рядом и наоборот. Пример ввода пары предложений в модель 1.4.

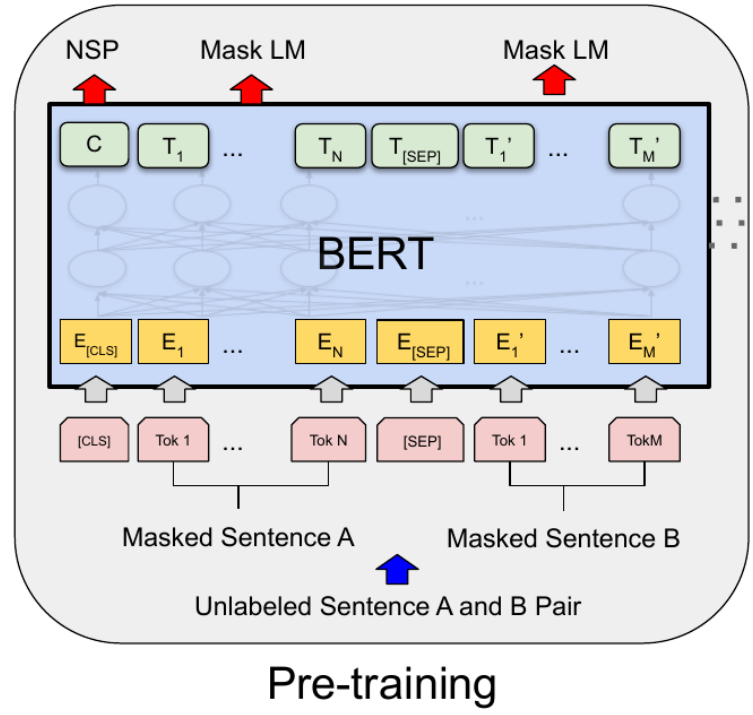

Рисунок 1.4 – Схемам работы BERT

1.4.2. Sentence BERT

Sentense BERT [18] – это модификация предобученных моделей BERT, которая использует 2 модели BERT, затем усреднят их выходы, а после с помощью функции ошибки выдаёт результат. Схема работы модели 1.5. Основное преимущество данной модели над классическим BERT: эмбеддинги предложений можно сравнивать друг с другом независимо и не пересчитывать их пару каждый раз. Например, если для поиска похожих предложений из 10000 для обычного BERT потребуется 50 миллионов вычислений различных пар предложений, и это займёт 50 часов, то Sentense BERT

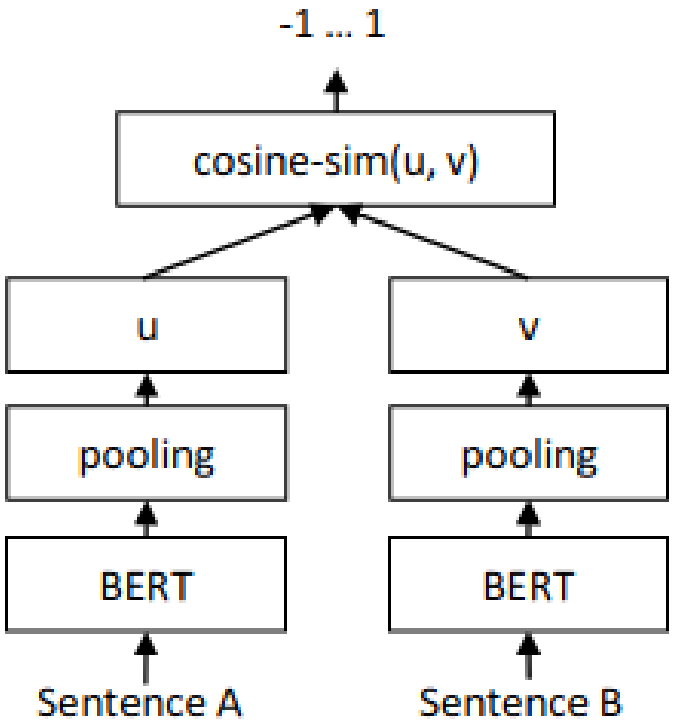

Рисунок 1.5 – Схема работы SBERT

рассчитает эмбеддинг каждого предложения отдельно, потом их сравнит. Такой способ рассчета ускоряет работу программы до 5 секунд.

1.5. Анализ требований к системе

Исходя из интервью с пользователями система должна уметь:

1. Показывать историю изменений индекса с возможностью фильтровать по:
    1. годам;
    2. отраслям компаний, с возможностью множественного выбора;
    3. компаниям, с возможностью множественного выбора;
    4. моделям, с возможностью множественного выбора;
    5. источникам, с возможностью множественного выбора.
2. Агрегировать значения индекса по годам и кварталам;
3. Анализировать тексты для построения индекса этичности;
4. Иметь возможность добавления анализа текста несколькими вариантами;
5. Сохранять тексты для последующего анализа другими методами;

6. Система должна собирать данные с сайтов banki.ru, sravni.ru и комментарии из групп «вконтаке»;
7. На сайте должен быть график, который показывать изменение индекса этичности компаний и количества собранных отзывов по разным источникам.
8. Для расчета индекса этичности компаний на основании рецензий должна использоваться формулаА.1:

$$\begin{aligned} \text{Base index} &= \frac{\text{positive} - \text{negative}}{\text{positive} + \text{negative}} \\ \text{Std index} &= \sqrt{\frac{\text{positive}}{\text{negative} \cdot (\text{positive} + \text{negative})^3} + \frac{\text{negative}}{\text{positive} \cdot (\text{positive} + \text{negative})^3}} \\ \text{Index} &= (2 \cdot (\text{Base index} - \text{Mean index} > 0) - 1) \cdot \\ & \quad max\left(|\text{Base index} - \text{Mean index}| - \text{Std index}, 0\right) \end{aligned} \tag{1.1}$$

positive – количество позитивных предложений,

negative – количество негативных предложений,

Mean index – среднее значения для пар источник сбора данных и модели, которая обрабатывала предложения.

На основе описания функциональных требований была создана диаграмма вариантов использования, которая представлена на рисунке 1.6.

Рисунок 1.6** – **Диаграмма вариантов использования

Также были получены нефункциональные требования:

1. построение графика не должно занимать больше секунды;
2. данные должны собираться автоматически;
3. данные должны обрабатываться автоматически;
4. система должны способна работать с большим объемом информации;

5. система должна быть стабильна.

1.6. Выбор технологий для разработки

Для реализации этой системы будет использоваться язык Python. Для этого языка разработано много библиотек, которые позволят быстро реализовать нейротропные алгоритмы обработки естественного языка, в частности в этом проекте будет использоваться Pytorch[19] и HuggingFace[20], и собирать данные с сайтов. Для реализации API будет использоваться FastAPI, что позволит разрабатывать API с автоматической документацией.

Хранение данных будет использоваться объектно-реляционная система управления базами данных PostgreSQL, что позволит обрабатывать большие объемы данных. Для работы с ней будет использоваться Code first подход, с помощью Python библиотек Sqlalchemy и Alembic для изменения схемы данных (миграций).

Для клиентской части приложения будет использоваться библиотека React.

1.7. Выводы главы

По итогам анализа предметной области, можно сделать вывод о том, что определение этичности компаний является важной задачей, которую можно автоматизировать с помощью алгоритмов машинного обучения. Анализ оценок этичности компаний позволяет понять, какие факторы необходимо учитывать при разработке алгоритмов. Обзор существующих решений показал, что некоторые из них имеют свои преимущества и недостатки, и может потребоваться разработка нового средства, учитывающего особенности задачи. Анализ алгоритмов помогает выбрать наиболее подходящие алгоритмы для поиска полезной информации в текстах. Наконец, анализ требований к системе позволяет определить необходимые функциональные и нефункциональные требования, которые будут учитываться при разработке решения. В целом, эти аналитические пункты помогут определить оптимальный подход к решению задачи определения этичности компаний.

Глава 2 Проектирование системы

В данной главе определена общая архитектура системы и каждого микросервиса, осуществлено проектирование баз данных, API микросервисов для модуля анализа для универсальной рекомендательной системы.

2.1. Проектирование архитектуры системы

Система будет разделена на отдельные независимые компоненты (микросервисы), что позволит ей быть надежной, если в какой-то части системы будут сбои, то остальная часть системы продолжит работать, и масштабируемой, легко добавлять новые компоненты. Каждый микросервис системы будет представлять собой docker container, которые будут управляться с помощью docker compose. Каждый сервис будет реализовывать отдельный компонент бизнес-логики и коммуницировать с другими компонентами через HTTP API.

Было выделено 5 главных компонента бизнес логики:

1. Работа с базой данных – это HTTP API, который обеспечивает возможность сохранения и получения данных из базы данных. Данный компонент принимает запросы на сохранение данных, получение информации из базы данных и возвращает результаты обработки этих запросов.
2. Сбор данных – компонент, который отвечает за сбор информации с нескольких источников. Для этого используется несколько независимых сборщиков данных, которые работают с различными сайтами и другими источниками.
3. Обработка данных – данный компонент содержит несколько моделей, которые используются для анализа данных. Эти модели производят различные виды анализа, от простой фильтрации и сортировки до более сложных операций анализа и прогнозирования.
4. Агрегирование данных – этот компонент отвечает за агрегацию обработанных данных в единый индекс. Данный индекс может быть использован для удобного представления полученных результатов в виде отчетов и графиков.
5. Сайт – этот компонент будет отображать агрегированную информацию.

Результат архитектуры системы на рис. 2.1.

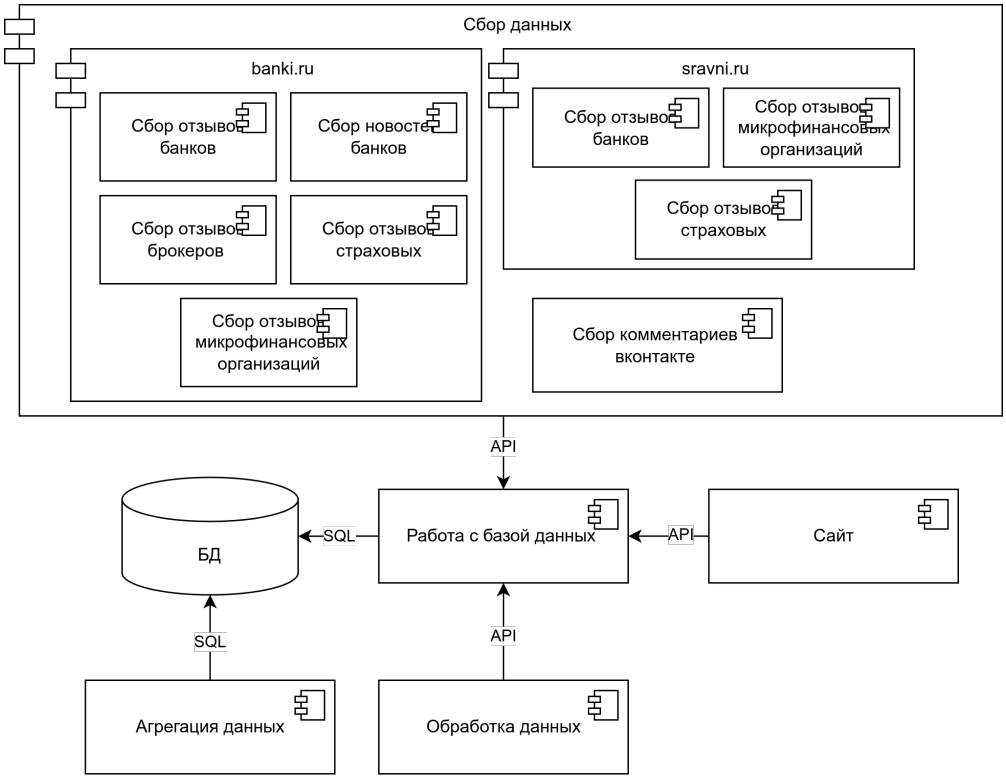

Рисунок 2.1 – Диаграмма архитектуры системы

Сервис для работы с базой данных, который будет обеспечивать сохранение и получение информации из различных сервисов сбора и обработки данных, а также сайтов. Для этого будет предоставлен API, который будет использоваться для отправки и получения данных.

Сервисы сбора данных будут отправлять собранные тексты в формате JSON на сервис работы с базой данных с помощью HTTP запросов. Кроме того, информация, необходимая для сбора данных, будет храниться в базах данных соответствующих сервисов.

Сервис агрегации данных будет периодически обновлять базу данных один раз в день для обеспечения актуальности данных.

Сервис сбора данных будет включать несколько моделей машинного обучения, которые будут использоваться для анализа данных, полученных из сервиса сбора данных. После обработки данных, результаты будут отправляться обратно в сервис сбора данных.

Сайт будет получать данные из сервиса работы с базой данных.

2.2. Проектирование базы данных

2.2.1. Проектирование основной базы данных

На основании требований была разработана следующая схема базы данных:

Таблица сфер компаний, чтобы можно было фильтровать различные сферы компаний и смотреть как меняется этичность сферы в целом.

Таблица 2.1 – Таблица сфера компании

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Сфера компании	Строка	

Таблица 2.2 – Таблица компании

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Название компании	Строка	
Описание компании	Строка	Дополнительное поле для сохранения вспомогательной информации о компании
Лицензия компании	Строка	По лицензии компаний может будет сопоставлять компании на разных сайтах
Код сферы компании	Целое	Внешний ключ из таблицы Сфера компании

Таблица 2.3 – Таблица тип источников

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Название типа источника	Строка	

Таблица 2.4 – Таблица источники

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Сайт	Строка	Сайт источника
Код типа источника	Целое	Внешний ключ из таблицы тип источника
Состояние сборщика данных	JSON	Данные о текущем состояние сборщика данных, если возникнет сбой
Дата последнего сбора	DateTime	Точка когда сбор данных закончился, для дальнейшего сбора данных

Таблица 2.5 – Таблица тип модели

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Название модели	Строка	

Таблица 2.6 – Таблица модели

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Название модели	Строка	
Код типа модели	Целое	Внешний ключ на таблицу тип модели

Таблица 2.7 – Таблицы текст

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Ссылка	Строка	Ссылка на текст
Код источника	Целое	Внешний ключ из таблицы источники
Дата текста	DateTime	Время публикации текста
Заголовок	Строка	Заголовок текста
Код компании	Целое	Внешний ключ на компанию
Количество комментариев	Целое	

Так как Bert на вход принимает отдельные предложения, было решено сделать для них отдельную таблицу.

Таблица 2.8 – Таблица предложений

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Код текста	Целое	Внешний ключ из таблицы тексты
Предложение	Строка	
Номер предложения	Целое	Порядковый номер предложения в тексте

Таблица 2.9 – Таблица результатов анализа текстов

Название	**Тип**	**Назначение**
Идентификатор	Целое	Уникальный идентификатор
Код предложения	Целое	Внешний ключ из таблицы предложения
Код модели	Целое	Внешний ключ из таблицы модели
Результат	Вещественный массив	Результат работы модели
Обработано	Логическое	Показатель, обработано ли предложение или нет

Диаграмма полученной схемы базы данных рис. Б.1.

2.2.2. Проектирование базы данных для агрегации

При сборе функциональных требований было выявлено, что надо быстро показывать количество собранных отзывов и индекс компаний.

Обработанные данные из таблицы~2.9 агрегируются для каждого квартала и для расчета индекса по формуле А.1 будут добавлены дополнительные колонки.

Таблица 2.10 – Таблица для расчета и показа индекса

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Год	Целое	Год за который был агрегирован индекс
Квартал	Целое	Квартал за который был агрегирован индекс
Название модели	Строка	
Сайт источника	Строка	
Тип источника	Строка	
Название банка	Строка	
Код банка	Целое	Для запросов через API
Нейтральный	Целое	Количество нейтральных предложений за период
Позитивный	Целое	Количество позитивных предложений за период
Негативный	Целое	Количество негативных предложений за период
Базовый индекс	Вещественное	Индекс для расчета итогового индекса
Средний индекс	Вещественное	Индекс для расчета итогового индекса
Std индекс	Вещественное	Индекс для расчета итогового индекса
Индекс	Вещественное	Рассчитанный индекс

Собранные отзывы из таблицы~2.7 агрегируются для каждого месяца и будет собираться количество собранных отзывов за месяц.

Таблица 2.11 – Таблица для расчета и показа индекса

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Дата	DateTime	
Квартал	Целое	Квартал за который был агрегирован индекс
Тип источника	Строка	
Сайт	Строка	
Количество отзывов	Целое	

Диаграмма полученной схемы базы данных рис. Б.2.

2.3. Проектирование серверной части

2.3.1. Модуль работы с данными

Модуль будет представлять собой HTTP API для работой с базой данных.

При первом старте приложение будет получаться список компаний (банки, брокеры, микрокредитные организации и страховые). Для этого будут выгружаться данные с сайта «Центрального банка России» и помещаться в базу данных. При последующих стартах приложение будет проверяться, что в каждом списке есть компании и новые компании не будут выгружаться.

Для работы с источниками информации будет разработано 4 запроса:

Таблица 2.12 – Таблица для запросв текста

Метод	**Ссылка**	**Тело**	**Описание**
GET	source		Получение списка источников

Продолжение на следующей странице

Таблица 2.12 – Таблица для запросв текста (Продолжение)

Метод	**Ссылка**	**Тело**	**Описание**
POST	source	Название сайта – строка Тип источника – строка	Будет принимать новые источники данных. Также, если получен новый тип источника, то он будет добавлять в базу данных
GET	source/item/{id}		Получение источника данных
GET	source/type		Получение списка типов источников

2.3.2. Модуль агрегации данных

Для построения индекса этичности компаний будет ежедневно агрегироваться база данных и перестраиваться индексы.

2.3.3. Модуль обработки данных

2.3.4. Проектирование сбора данных с banki.ru

Для получения на сайт banki.ru будут отправляться запросы на их внутренний API. Для этого предварительно будут собранны данные о всех организациях, которые у них представлены на сайте и перемещены в базу данных2.3.4.

Таблица 2.13 – Таблица для сайта banki.ru

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Идентификатор банка	Целое	Идентификатор банка в основной базе данных
Имя банка	Строка	

Продолжение на следующей странице

Таблица 2.13 – Таблица для сайта banki.ru (Продолжение)

Название	**Тип**	**Описание**
Код банка	Строка	Код банка для запросов по API

Диаграмма полученной схемы базы данных рис. Б.3.

2.3.5. Проектирование сбора данных с sravni.ru

Для получения на сайт sravni.ru будут отправляться запросы на их внутренний API. Для этого предварительно будут собранны данные о всех организациях, которые у них представлены на сайте и перемещены в базу данных2.14.

Таблица 2.14 – Таблица для сайта sravni.ru

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Идентификатор банка	Целое	Идентификатор банка в основной базе данных
Код банка в sravni.ru	Целое	
Старый код банка в sravni.ru	Целое	
Псевдоним компании	Строка	
Название банка	Строка	

Диаграмма полученной схемы базы данных рис. Б.4

2.3.6. Проектирование сбора данных с vk.com

Для получения на сайт vk.com будут отправляться запросы на их внутренний API. Для этого предварительно будут собранны данные о всех организациях, которые у них представлены на сайте и перемещены в базу данных2.15.

Таблица 2.15 – Таблица для сайта vk.com

Название	**Тип**	**Описание**
Идентификатор	Целое	Уникальный идентификатор
Идентификатор на vk.com	Строка	
Имя компании	Строка	
Домен компании на vk.com	Строка	

Диаграмма полученной схемы базы данных рис. Б.5.

2.4. Проектирование клиентской части

Глава 3 Реализация системы

3.1. Реализация серверной части

3.1.1. Реализация API

3.1.2. Реализация парсера banki.ru

3.1.3. Реализация парсера sravni.ru

3.1.4. Реализация модуля обработки текста

3.1.5. Дообучение модели

3.2. Реализация клиентской части

Глава 4 Тестирование системы

Заключение

Библиографический список

1. *Смирнова, И. Л.* Бизнес-Этика Как Приоритетный Вектор Современного Развития Организаций / И. Л. Смирнова, М. В. Соловьева // Вестник Волжского Университета Им. В.н. Татищева. — 2021. — Т. 2, 1 (47).

2. *Murè, P.* ESG and Reputation: The Case of Sanctioned Italian Banks / P. Murè [et al.] // Corporate Social Responsibility and Environmental Management. — 2021. — Vol. 28, no. 1. — P. 265–277.

3. *Семенко, И. Е.* Корпоративная Социальная Ответственность И Бизнес-Этика Компании / И. Е. Семенко // Экономические науки: актуальные вопросы теории и практики. — Наука и Просвещение, 2022. — С. 43—45.

4. *Кудрявцева, Ю. А.* Корпоративно-Социальная Ответственность В Контексте Этики Банковского Дела / Ю. А. Кудрявцева, Г. Г. Чахкиев. — 2016.

5. *Climent, F.* Ethical Versus Conventional Banking: A Case Study / F. Climent // Sustainability. — 2018. — July. — Vol. 10, issue 7, no. 7. — P. 2152.

6. *Harvey, B.* Ethical Banking: The Case of the Co-operative Bank / B. Harvey // Journal of Business Ethics. — 1995. — Dec. 1. — Vol. 14, no. 12. — P. 1005–1013.

7. *Brunk, K. H.* Exploring Origins of Ethical Company/Brand Perceptions — A Consumer Perspective of Corporate Ethics / K. H. Brunk // Journal of Business Research. — 2010. — Mar. 1. — Vol. 63, no. 3. — P. 255–262.

8. *Mitchell, W. J.* Bank Ethics: An Exploratory Study of Ethical Behaviors and Perceptions in Small, Local Banks / W. J. Mitchell, P. V. Lewis, N. L. Reinsch // Journal of Business Ethics. — 1992. — Mar. 1. — Vol. 11, no. 3. — P. 197–205.

9. *López, M. V.* Sustainable Development and Corporate Performance: A Study Based on the Dow Jones Sustainability Index / M. V. López, A. Garcia, L. Rodriguez // Journal of Business Ethics. — 2007. — Oct. 1. — Vol. 75, no. 3. — P. 285–300.

10. *Collison, D. J.* The Financial Performance of the FTSE4Good Indices / D. J. Collison [et al.] // Corporate Social Responsibility and Environmental Management. — 2008. — Vol. 15, no. 1. — P. 14–28.

11. *Harris, Z. S.* Distributional Structure / Z. S. Harris // WORD. — 1954. — Aug. 1. — Vol. 10, no. 2/3. — P. 146–162.

12. *Jones, Karen Sparck.* A Statistical Interpretation of Term Specificity and Its Application in Retrieval / Jones, Karen Sparck // Journal of Documentation. — 1972. — Jan. 1. — Vol. 28, no. 1. — P. 11–21.

13. *Mikolov, T.* Distributed Representations of Words and Phrases and Their Compositionality / T. Mikolov [et al.] // Advances in Neural Information Processing Systems. Vol. 26. — Curran Associates, Inc., 2013.

14. *Peters, M. E.* Deep Contextualized Word Representations / M. E. Peters [et al.]. — 03/22/2018.

15. *Radford, A.* Language Models Are Unsupervised Multitask Learners / A. Radford [et al.]. — 2019.

16. *Devlin, J.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin [et al.]. — 05/24/2019.

17. *Vaswani, A.* Attention Is All You Need / A. Vaswani [et al.] // Advances in Neural Information Processing Systems. Vol. 30. — 2017.

18. *Reimers, N.* Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks / N. Reimers, I. Gurevych. — 08/27/2019.

19. *Paszke, A.* PyTorch: An Imperative Style, High-Performance Deep Learning Library / A. Paszke [et al.] // Advances in Neural Information Processing Systems. Vol. 32. — Curran Associates, Inc., 2019.

20. *Wolf, T.* Transformers: State-of-the-Art Natural Language Processing / T. Wolf [et al.] // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. — Online : Association for Computational Linguistics, 10/2020. — P. 38–45.

ПРИЛОЖЕНИЕ А Техническое задание на разрабатываемую систему

УТВЕРЖДЕНО

А.В.00001-01 ТЗ 01

СИСТЕМА ДЛЯ АВТОМАТИЧЕСКОГО СБОРА, АНАЛИЗА И ВИЗУАЛИЗАЦИИ ИНФОРМАЦИИ ПО ЭТИЧНОСТИ КОМПАНИЙ

Техническое задание

Лист утверждения

Подпись и дата	
Инв. № дубл.	
Взам. инв. №	
Подпись и дата	
Инв. № подл.	

Руководитель разработки

_____ Бузмаков А.В.

«___»_____ 2023

Исполнитель

_____ Соломатин Р.И.

«___»_____ 2023

1. Общие сведения

Наименование программы – «Система для автоматического сбора, анализа и визуализации информации по этичности компаний» (далее – «Система»). Основная функция системы - сбор и анализ данных из различных источников, включая новостные сайты, социальные сети, отзывы о компаниях и другие открытые источники данных. Система использует алгоритмы машинного обучения и обработки естественного языка для автоматической обработки данных и определения этичности компаний.

Система также предоставляет визуализацию данных в виде графиков и диаграмм, позволяя пользователям легко понять и сравнивать данные по разным компаниям. Кроме того, система может предоставлять аналитические отчеты и рекомендации по улучшению этичности компаний на основе собранных данных.

Система разрабатывается в рамках выполнения выпускной квалификационной работы. Основанием для разработки являются:

- Положение о курсовой и выпускной квалификационной работе студентов, обучающихся по программам бакалавриата, специалитета и магистратуры в Национальном исследовательском университете «Высшая школа экономики», утвержденным ученым советом НИУ ВШЭ (протокол от 28.11.2014 № 08), с изменениями от 29.03.2016;
- Правила подготовки выпускной квалификационной работы студентов основной образовательной программы бакалавриата «Программная инженерия» по направлению подготовки 09.03.04. Программная инженерия, утвержденные протоколом ученого совета НИУ ВШЭ – Пермь от 19.11.2020 № 8.2.1.7-10/10.

2. Цели и назначение создания автоматизированной системы

2.1. Цели создания АС

Целью создания системы является получение инструмента который позволит анализировать компании на основании их этичности, соответствующего следующим требованиям:

- Показывать историю изменений индекса с возможностью фильтровать по

  - годам;
  - отраслям компаний, с возможностью множественного выбора;
  - компаниям, с возможностью множественного выбора;
  - моделям, с возможностью множественного выбора;
  - источникам, с возможностью множественного выбора.
- Агрегировать значения индекса по годам и кварталам;
- Анализировать тексты для построения индекса этичности;
- Иметь возможность добавления анализа текста несколькими вариантами;
- Сохранять тексты для последующего анализа другими методами;
- Система должна собирать данные с сайтов banki.ru, sravni.ru и комментарии из групп «вконтаке»;
- На сайте должен быть график, который показывать изменение индекса этичности компаний и количества собранных отзывов по разным источникам.
- Для расчета индекса этичности компаний на основании рецензий должна использоваться формулаА.1:

$$\begin{gathered}\text{Base index} = \frac{\text{positive} - \text{negative}}{\text{positive} + \text{negative}} \\ \text{Std index} = \sqrt{\frac{\text{positive}}{\text{negative} \cdot (\text{positive} + \text{negative})^3} + \frac{\text{negative}}{\text{positive} \cdot (\text{positive} + \text{negative})^3}} \\ \text{Index} = (2 \cdot (\text{Base index} - \text{Mean index} > 0) - 1) \cdot \\ max\left(|\text{Base index} - \text{Mean index}| - \text{Std index}, 0\right)\end{gathered} \tag{A.1}$$

positive – количество позитивных предложений,

negative – количество негативных предложений,

Mean index – среднее значения для пар источник сбора данных и модели, которая обрабатывала предложения.

2.2. Назначение АС

Система предназначена для собора и анализа отзывы потребителей с различных веб-сайтов, с помощью алгоритмов обработки естественного языка.

3. Характеристика объекта автоматизации

Система автоматизирует процесс анализа этичности компаний.

4. Требования к автоматизированной системе

Требования к АС:

1. построение графика не должно занимать больше секунды;
2. данные должны собираться автоматически;
3. данные должны обрабатываться автоматически;
4. система должны способна работать с большим объемом информации;
5. система должна быть стабильна.

4.1. Требования к структуре АС в целом

Должно быть несколько модулей, которые общаются между собой с помощью HTTP API:

- сборщики данных;
- взаимодействие с базой данных(API);
- сайт;
- модели для обработки данных.

Все подсистемы должны быть в Docker контейнерах.

4.2. Требования к функциям (задачам), выполняемым АС

4.2.1. Требования к API

Система взаимодействия с базой данных должна выполнять функции:

- хранить информацию о компаниях;
- модель должна отдавать информацию о компаниях из разных сфер;
- хранить информацию о разных источниках;
- добавлять различные источники;
- получать отзывы и разбивать их на предложения;
- иметь возможность отдавать необработанные предложения в зависимости от модели;

- сохранять информацию о моделях;
- сохранять результат обработки предложений;
- агрегировать результат обработки моделей для каждого квартала;
- хранить информацию о состоянии сборщиков данных, если у них возникнут проблемы;
- отдавать информацию об индексе менее, чем за минуту;
- рассчитывать индекс на основе полученных результатов обработки предложений.

4.2.2. Требования к сборщикам данных

Сборщики данных должны выполнять функции:

- собирать отзывы пользователей ежедневно;
- собранные отзывы отправлять на API.

4.2.3. Требования к моделям

Модели должны выполнять функции:

- обрабатывать отзывы пользователей ежедневно;
- обработанные отзывы отправлять на API.

4.3. Требования к видам обеспечения АС

4.3.1. Требования к лингвистическому обеспечению

Система должна соответствовать следующим требованиям:

- Программный код должен быть реализован на языке Python;
- Документация к программе должна быть на русском языке. Других языков не планируется.

4.3.2. Требования к программного обеспечению

Система должна использовать:

- Для разработки API следует использовать библиотеку FastAPI;
- Для взаимодействия с базой данных должна использоваться библиотека SQLAlchemy, а для миграций Alembic;

– Сборщики данных должны собирать информацию с помощью библиотеки запросов requests и для работы с HTML Beautiful Soup;
– Для нейросетевых моделей должен использоваться Pytorch;
– Для клиентской части будет использоваться библиотека React.

4.3.3. Требования к техническому обеспечению

Для работы приложения необходим сервер, который обладает следующими параметрами:

– Процессор с тактовой частотой не ниже 2,5 ГГц, при количестве ядер не менее 4;
– Графическая карта с объемом памяти не менее 4 Гб;
– ОЗУ не менее 16 Гб;
– Не менее 100 Гб свободного места на жестком диске для хранения собранных данных;
– Скорость интернета не менее 100 Мб/с;
– ОС Ubuntu 20.04 и выше;
– Docker 20.10.23 и выше.

4.3.4. Требования к информационному обеспечению

Система должна соответствовать следующим требованиям:

– Система должна использовать PostgreSQL;
– Сервисы между собой должны взаимодействовать при помощи HTTP;
– Для базы данных должен всегда быть резервная копия данных.

4.4. Общие технические требования к АС

4.4.1. Требования к численности и квалификации персонала

Для разработки системы требуется программист со средней квалификацией. Для работы с конечной системой (сайтом), не требуется высокой квалификации, поэтому пользователь с ней справится пользователь, который пользуется сайтами.

4.4.2. Требования к надежности

Надежность системы зависит от надежности функционирования сервера. Устойчивое функционирование программы будет обеспечено с помощью:

- Бесперебойное питание сервера;
- Использованием лицензионного программного обеспечения, необходимого для запуска приложения, включая лицензионную операционную систему;
- Регулярным выполнением рекомендаций Министерства труда и социального развития РФ, изложенных в Постановлении от 23 июля 1998 г. «Об утверждении межотраслевых типовых норм времени на работы по сервисному обслуживанию ПЭВМ и оргтехники и сопровождению программных средств»;
- Регулярным выполнением требований ГОСТ 51188-98 «Защита информации. Испытания программных средств на наличие компьютерных вирусов»;

Время восстановление системы будет до 10 минут с момента сбоя.

4.4.3. Требования по сохранности информации при авариях

Отказы как самой системы, так и ее отдельных функций, могут привести к аварийному завершению работы программы, однако при перезапуске программы ее функциональность не должна пострадать. При таких сбоях программы база данных не должна пострадать. Дополнительно все данные будут резервно копироваться на дополнительную базу данных.

5. Состав и содержание работ по созданию автоматизированной системы

Таблица А.1 – Этапы реализации, контрольные точки проекта

Основной этап	Подэтап	**Крайний срок**
Анализ	Литературный обзор	26.12.2022
Анализ	Сравнительный анализ существующих решений	08.01.2023

Продолжение на следующей странице

Таблица А.1 – Этапы реализации, контрольные точки проекта (Продолжение)

Основной этап	Подэтап	**Крайний срок**
Анализ	Анализ сценариев использования	14.01.2023
Анализ	Написание тех задания	21.01.2023
Проектирование	Проектирование архитектуры приложения	14.02.2023
Проектирование	Проектирование базы данных	20.03.2023
Проектирование	Проектирование графического интерфейса	20.03.2023
Проектирование	Проектирование алгоритмов машинного обучения	20.03.2023
Разработка	Разработка алгоритмов для анализа текста	01.05.2023
Разработка	Реализация серверной части	01.05.2023
Разработка	Реализация клиентской части	01.05.2023
Тестирование	Подготовка тестовых сценариев	15.05.2023
Тестирование	Функциональное тестирование	15.05.2023
Тестирование	Системное тестирование	15.05.2023
Завершение	Сдача проекта	22.05.2023

6. Порядок разработки автоматизированной системы

В разделе «Порядок разработки автоматизированной системы» приводят следующее:

- порядок организации разработки АС;
- перечень документов и исходных данных для разработки АС;
- перечень документов, предъявляемых по окончании соответствующих этапов работ;
- порядок проведения экспертизы технической документации;
- перечень макетов (при необходимости), порядок их разработки, изготовления, испытаний, необходимость разработки на них документации, программы и методик испытаний;
- порядок разработки, согласования и утверждения плана совместных работ по разработке АС;
- порядок разработки, согласования и утверждения программы работ по стандартизации;
- требования к гарантийным обязательствам разработчика;
- порядок проведения технико-экономической оценки разработки АС;
- порядок разработки, согласования и утверждения программы метрологического обеспечения, программы обеспечения надежности, программы эргономического обеспечения.

7. Порядок контроля и приемки автоматизированной системы

Осуществление приемо-сдаточных испытаний для всей системы осуществляется на основе Программы и методики испытаний и включает:

- Функциональное тестирование;
- Тестирование удобства эксплуатации;
- Оценка сгенерированных уровней.

8. Требования к составу и содержанию работ по подготовке объекта автоматизации к вводу автоматизированной системы в действие

Таблица А.2 – Требования к программной документации

Название документа	**Краткое содержание**
Текст программы (ГОСТ 19.401–78)	Программный код всех модулей программы с необходимыми комментариями.
Программа и методика испытаний (ГОСТ 19.301–79)	Требования, подлежащие проверке при испытании программы, а также порядок и методы их контроля.
Техническое задание (34.602-2020)	Назначение и область применения программы, технические, технико-экономические и специальные требования, предъявляемые к программе, необходимые стадии и сроки разработки, виды испытаний.

ПРИЛОЖЕНИЕ Б Схема базы данных

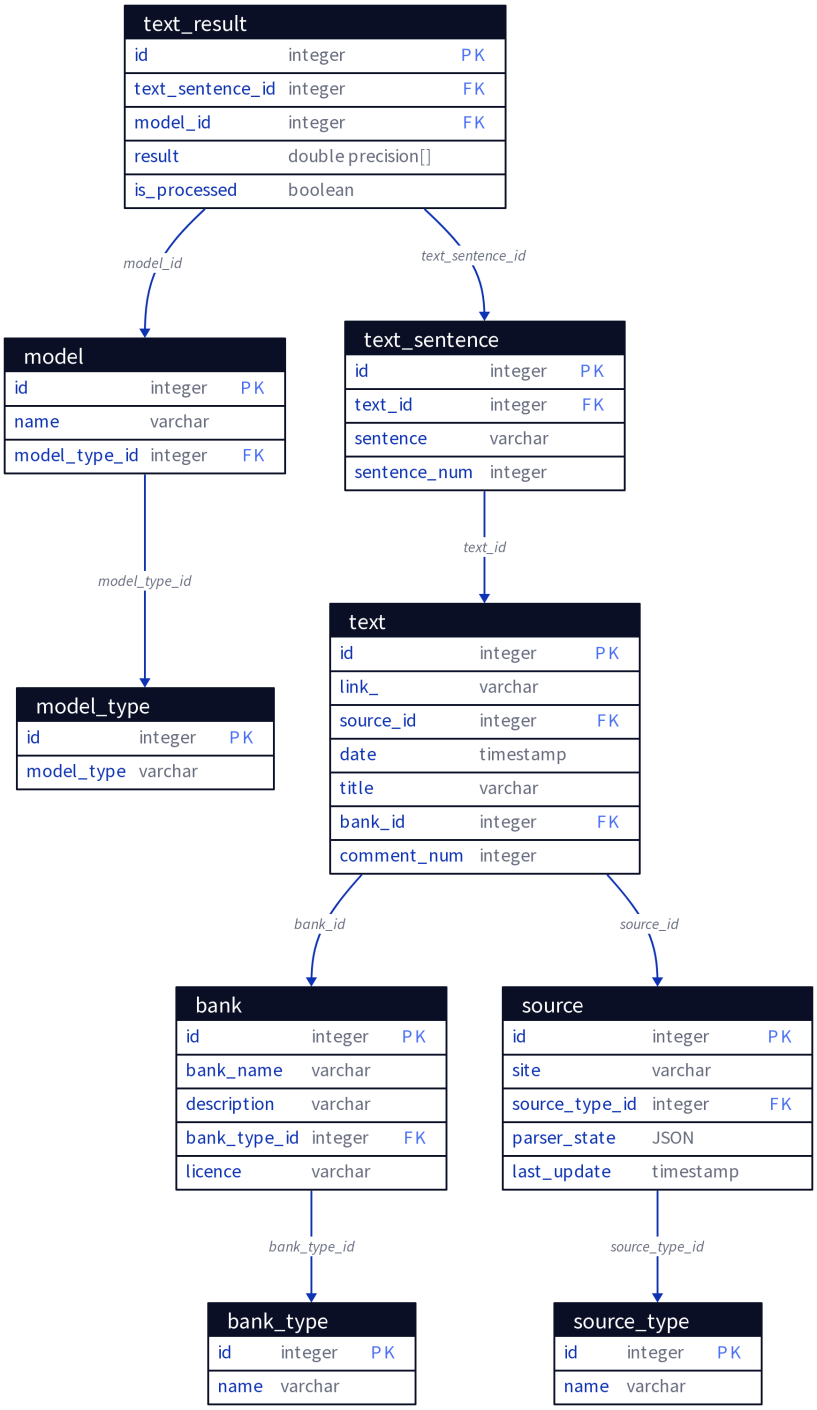

Рисунок Б.1 – Схема базы данных

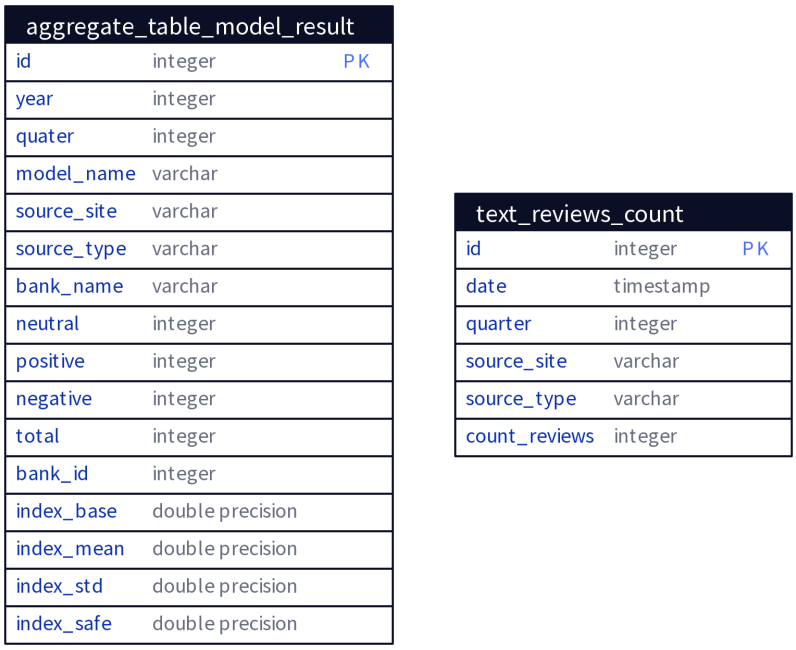

Рисунок Б.2 – Схема базы данных для агрегаций

Рисунок Б.3 – Схема базы данных сайта banki.ru

Рисунок Б.4 – Схема базы данных сайта sravni.ru

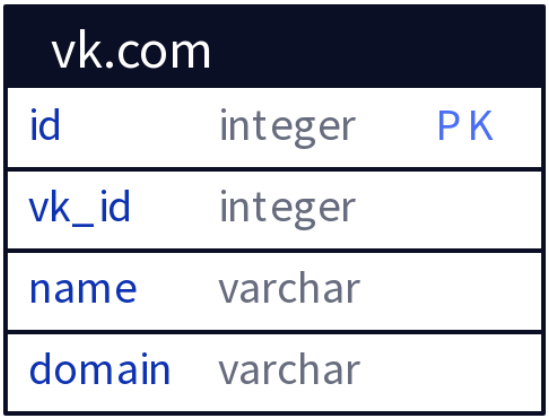

Рисунок Б.5 – Схема базы данных сайта vk.com