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 – 23, количество иллюстраций – 6, количетсво таблиц – 1.

Оглавление

Введение 5
Глава 1 Анализ предметной области 8
1.1 Анализ определения этичности компании 8
1.2 Анализ оценок этичности компаний 9
1.3 Анализ существующих решений 10
1.4 Алгоритмы для анализа текста 11
1.4.1 BERT 12
1.4.2 Sentence BERT 14
1.5 Анализ требований к системе 15
1.6 Выбор технологий для разработки 16
1.7 Выводы главы 16
Глава 2 Проектирование системы 18
2.1 Проектирование архитектуры системы 18
2.2 Проектирование базы данных 18
2.3 Проектирование серверной части 18
2.3.1 Модуль сбора данных 18
2.3.2 Модуль обработки данных 18
2.3.3 Модуль агрегации данных 18
2.4 Проектирование клиентской части 18
Глава 3 Реализация системы 19
3.1 Реализация серверной части 19
3.1.1 Реализация API 19
3.1.2 Реализация парсера banki.ru 19
3.1.3 Реализация парсера sravni.ru 19
3.1.4 Реализация модуля обработки текста 19
3.1.5 Дообучение модели 19
3.2 Реализация клиентской части 19

Глава 4 Тестирование системы 20
Заключение 21
Библиографический список 22

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].

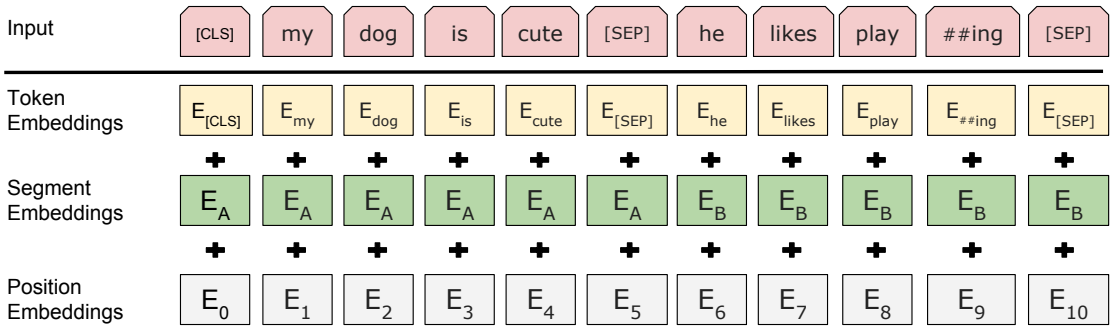

Рисунок 1.2 – Пример ввода текста в модель

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.

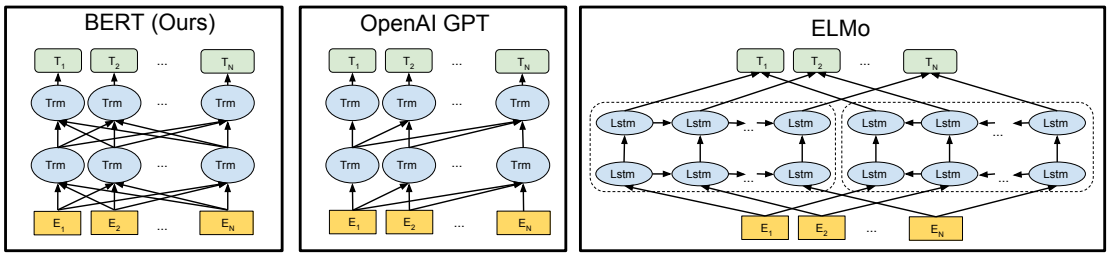

Рисунок 1.3 – Сравнение принципов работы BERT, ELMo, GPT

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.

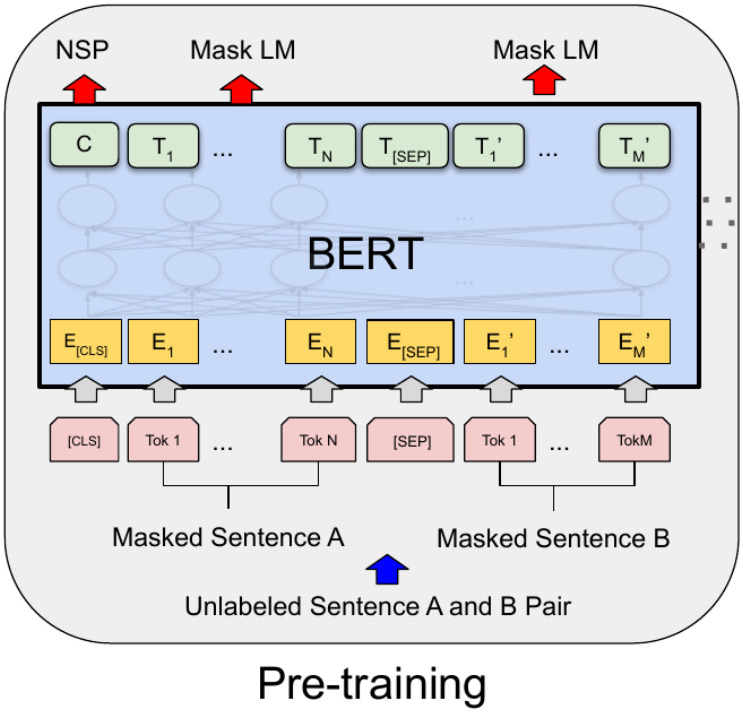

Рисунок 1.4 – Схемам работы BERT

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

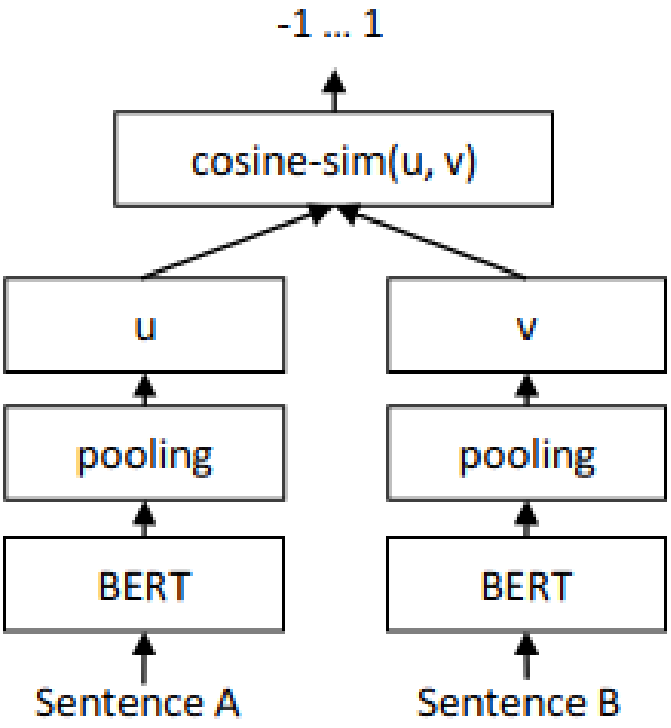

***Рисунок 1.5** – Схема работы SBERT*

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. Сохранять тексты для последующего анализа другими методами.

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).

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 позво-

ляет 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 иметь микросервисную архитектуру, что позволит ей быть надежной, если в какой-то части системы будут сбои, и масштабируемой, будет легко добавлять новые компоненты.

2.2. Проектирование базы данных

2.3. Проектирование серверной части

2.3.1. Модуль сбора данных

2.3.2. Модуль обработки данных

2.3.3. Модуль агрегации данных

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.